Coexpression Analysis of Human Genes Across Many Microarray Data Sets

Homin K. Lee,[1] Amy K. Hsu,[1,2] Jon Sajdak,[1] Jie Qin,[1] and Paul Pavlidis[1,3,4]

[1]*Columbia Genome Center,* [2]*College of Physicians and Surgeons, and* [3]*Department of Biomedical Informatics, Columbia University, New York, New York 10032, USA*

We present a large-scale analysis of mRNA coexpression based on 60 large human data sets containing a total of 3924 microarrays. We sought pairs of genes that were reliably coexpressed (based on the correlation of their expression profiles) in multiple data sets, establishing a high-confidence network of 8805 genes connected by 220,649 "coexpression links" that are observed in at least three data sets. Confirmed positive correlations between genes were much more common than confirmed negative correlations. We show that confirmation of coexpression in multiple data sets is correlated with functional relatedness, and show how cluster analysis of the network can reveal functionally coherent groups of genes. Our findings demonstrate how the large body of accumulated microarray data can be exploited to increase the reliability of inferences about gene function.

[Supplemental material is available online at www.genome.org and http://microarray.cpmc.columbia.edu/tmm.]

Gene expression microarray data is a form of high-throughput genomics data providing relative measurements of mRNA levels for thousands of genes in a biological sample. In the last few years, hundreds of laboratories have collected and analyzed microarray data, and the data are beginning to appear in public databases or on researchers' Web sites. These resources serve at least two purposes. One is as an archive of the data, which allows other researchers to confirm the results that have been published by the originator of the data. A second use is to permit novel analyses of the data, that go beyond what was envisioned or possible at the time of the original study. A novel analysis could involve just a single data set, or a meta-analysis of many data sets (where a "data set" is a group of microarrays that were collected together, and typically described as a group in a single publication). The combined analysis of multiple data sets forms the main topic of this paper.

Most existing studies that have analyzed multiple independently collected microarray data sets have focused on differential expression, comparing two or more similar data sets to look for genes that distinguish different sets of samples (Breitling et al. 2002; Rhodes et al. 2002; Yuen et al. 2002; Choi et al. 2003; Detours et al. 2003; Ramaswamy et al. 2003; Sorlie et al. 2003; Xin et al. 2003). Another type of comparison is exemplified by a study that examined the variability of expression for individual genes in several human and mouse data sets (Lee et al. 2002). These studies have generally been able to exploit the availability of multiple data sets to identify more robust sets of genes than would be found using a single data set.

Another way of using microarray data is to exploit gene coexpression instead of differential expression. In this approach, genes that have similar expression patterns across a set of samples are hypothesized to have a functional relationship. It has been shown in a number of studies that coexpression is correlated with functional relationships, such as physical interaction between the encoded proteins, though coexpression does not necessarily imply a causal relationship among transcript levels (Eisen et al. 1998; Ge et al. 2001; Jansen et al. 2002; Kemmeren et al. 2002). Because microarray data are noisy, there has been an interest in seeking supporting evidence for predictions made based on coexpression. Although several studies have combined microarray data with other data types (Marcotte et al. 1999; Greenbaum et al. 2001; Kemmeren et al. 2002; von Mering et al. 2002), the reproducibility of coexpression patterns between microarray data sets has not been studied in much detail. Graeber et al. (Graeber and Eisenberg 2001) identified a number of coexpression patterns found in several tumor data sets, but their analysis was focused on a small number of genes (receptors and their ligands). A recent study identified a subset of coexpression patterns that were common to multiple model organisms (Stuart et al. 2003). A direct comparison of two closely related mouse brain data sets showed a high degree of reproducibility of expression profiles between the studies as long as the data were stringently filtered prior to analysis (Dabrowski et al. 2003). Such an analysis requires that the samples in the two data sets be directly comparable, and Dabrowski et al. did not consider coexpression as such. In contrast to the positive findings of Dabrowski et al., a study comparing two data sets, both obtained from the National Cancer Institute reference tumor cell lines (NCI-60) but on two different microarray platforms, found that clustering results were not reproducible (Kuo et al. 2002).

In this paper we describe an analysis of gene coexpression in 60 large human microarray data sets, and we assess the functional relevance and reproducibility of the coexpression patterns we detected. We found that a substantial number of correlated expression patterns occur in multiple independent data sets. This confirmation of correlated expression provides a useful way to improve the confidence in any particular correlated expression pattern. Indeed, we show that coexpression patterns that are confirmed are more likely to be functionally relevant. The database and methods we describe can form the basis for further large-scale exploration of gene coexpression data.

RESULTS

We analyzed pairwise correlation of gene expression in a large corpus of microarray data of 60 diverse data sets (Table 1). This corpus contains a total of 62.2 million expression measurements distributed among 3924 microarrays; all of the data sets have at least 10 samples (microarrays), and the largest contains 255 samples. We analyzed correlation of gene expression profiles within each data set, selecting for further study the "coexpression

[4]Corresponding author.
E-MAIL pp175@columbia.edu; FAX (212) 851-5149.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1910904.

Table 1. Summary of the Microarray Data Sets Used[a]

Reference[b]	Samples[c]	Genes[d]	Raw links[e]	Reference	Samples	Genes	Raw links
(Alizadeh et al. 2000)	96	1759	25748	(Nielsen et al. 2002)	46	3359	25175
(Allander et al. 2001)	19	1205	1251	(Perou et al. 1999)	26	3027	42105
(Armstrong et al. 2002)	72	8242	213456	(Perou et al. 2000)	84	5701	167826
(Bhattacharjee et al. 2001)	203	8243	243303	(Pomeroy et al. 2002)	90	5418	85909
(Bittner et al. 2000)	38	4382	16141	(Ramaswamy et al. 2001)	255	9528	372500
(Butte et al. 2000)	68	4906	81755	(Rickman et al. 2001)	51	5418	60169
(Chang et al. 2002)	50	13079	328274	(Rosenwald et al. 2001)	102	3751	129814
(Chaussabel et al. 2003)	28	8243	80559	(Ross and Perou 2001)	24	12437	3409
(Chen et al. 2002)	207	9169	597313	(Ross et al. 2000)	68	5837	45468
(Cheok et al. 2003)	120	8243	258731	(Shipp et al. 2002)	77	5418	87486
(Dhanasekaran et al. 2001)	53	5613	161097	(Singh et al. 2002)	20	4119	19895
(Diehn et al. 2002)	68	10400	1635022	(Smith et al. 2003)	102	8242	248952
(Dyrskjot et al. 2003)	31	5418	60343	(Sorlie et al. 2001)	85	9132	1578
(Dyrskjot et al. 2003)	40	5418	58461	(Sorlie et al. 2003)	122	13121	14351
(Erraji-BenChekroun et al., in prep.)	75	12057	824563	(Staunton et al. 2001)	11	8257	4760
(Garber et al. 2001)	73	9171	258866	(Su et al. 2002)	12	8257	128303
(Golub et al. 1999)	72	5418	52283	(Tezak et al. 2002)	24	12057	12944
(Greenberg et al. 2002)	12	8243	5280	(Unpublished, GSE443)	60	5418	51286
(Gruvberger et al. 2001)	58	2756	24781	(Unpublished, GSE470)	85	8243	241088
(Hedenfalk et al. 2001)	22	2253	1265	(Unpublished, GSE474)	10	5418	7941
(Hedenfalk et al. 2003)	16	3312	673	(Vahey et al. 2002)	30	5418	60268
(Huang et al. 2003)	89	8257	137512	(van't Veer et al. 2002)	117	11312	752390
(Huang et al. 2001)	16	8243	9634	(Virtaneva et al. 2001)	21	4804	5923
(Jazaeri et al. 2002)	61	3644	46187	(Welle et al. 2001)	12	8243	2670
(Khan et al. 2001)	88	1952	19868	(Welsh et al. 2001)	49	5418	52459
(Khatua et al. 2003)	13	8257	10072	(Welsh et al. 2001)	55	8258	260155
(Leung et al. 2002)	126	12657	993195	(West et al. 2001)	49	5418	84842
(Luo et al. 2001)	25	4354	14873	(Whitfield et al. 2002)	114	12801	1547199
(Ma et al. 2003)	61	1569	10086	(Yeoh et al. 2002)	248	8257	257979
(MacDonald et al. 2001)	31	1309	3179	(Yoon et al. 2002)	12	5418	53305

[a]A version of this table with additional information is available as Supplemental data.
[b]In three cases where the data are not published, the Gene Expression Omnibus accession number is given.
[c]The number of samples (microarrays) in the data set.
[d]The number of unique RefSeq genes represented on the array that were included for analysis.
[e]The number of raw coexpression links between the RefSeq gene selected for inclusion in the database.

links" that were deemed to be statistically significant (see Methods). For the analysis presented in this paper, we considered a set of 16,511 human genes from RefSeq, of which 15,700 were detectably expressed in at least one data set.

This analysis yielded 9.7 million different "raw" coexpression links between genes. A total of 11 million occurrences of these links were found, indicating that some links occur in multiple data sets. Of the 9.7 million different links, 5.39 million (56%) had positive correlations, compared to 4.31 million negative correlations. This imbalance apparently occurs because negative correlations tended to be less common than positive correlations in the raw data, and fewer of them reach significance in our primary analysis. Between 673 and 1.5 million correlated gene pairs (raw coexpression links) were stored for each data set (median 56,000; Table 1). Of the genes tested, 15,458 (98%) had at least one coexpression link, with a median of 990 per gene. For the most part, the number of links a data set yielded was proportional to the number of genes represented on the array, but this was also affected by the number of samples in the data set (data not shown). This is because our criteria for link acceptance takes into account the number of samples in the calculation of statistical significance.

Coexpression Link Confirmation

For a variety of reasons, some of the coexpression links for a gene are likely to be artifacts or of questionable biological relevance. A primary goal of this work was to evaluate the reoccurrence of links in multiple data sets, with the expectation that this will improve the reliability of the inferences that might be made on the basis of coexpression. We refer to this as "coexpression link confirmation" (Fig. 1). A second type of link confirmation occurs within data sets, when genes are represented by multiple probes or probe sets on the same microarray. A preliminary analysis of such "intra-data set" link confirmation is presented as Supplemental data.

Figure 2A shows the number of times a link is confirmed in a given number of data sets. This figure shows that whereas most links are not confirmed in our database, many links are confirmed and some links are found in numerous data sets. The largest number of data sets a link was seen in was 31. Of the links in our original selected pool of 9.7 million, none were testable in all 60 data sets (the maximum was 57), because as mentioned none of the genes we considered occurred or were considered detectable in all 60 data sets. The wide variety of microarray platforms represented in our database lead to most links being tested in far fewer than 60 data sets, and the links in the original pool were tested in a mean of 18 data sets (median 15).

Although confirmation of coexpression suggests greater reliability, we expect some confirmations to occur purely by chance, due to the large number of data sets we tested. To estimate the statistical significance of link confirmation, we created randomized databases where the number of links per gene and per data set had the identical distributions as in our real data, but the links were created between genes within a data set at random. These randomized databases produce links confirmed in three or more data sets (hereafter denoted as "3+ confirmed") at a rate of $5.24 \pm 0.08\%$ (mean $\pm$ standard deviation) of that observed using the original data, and produce very few links confirmed in more

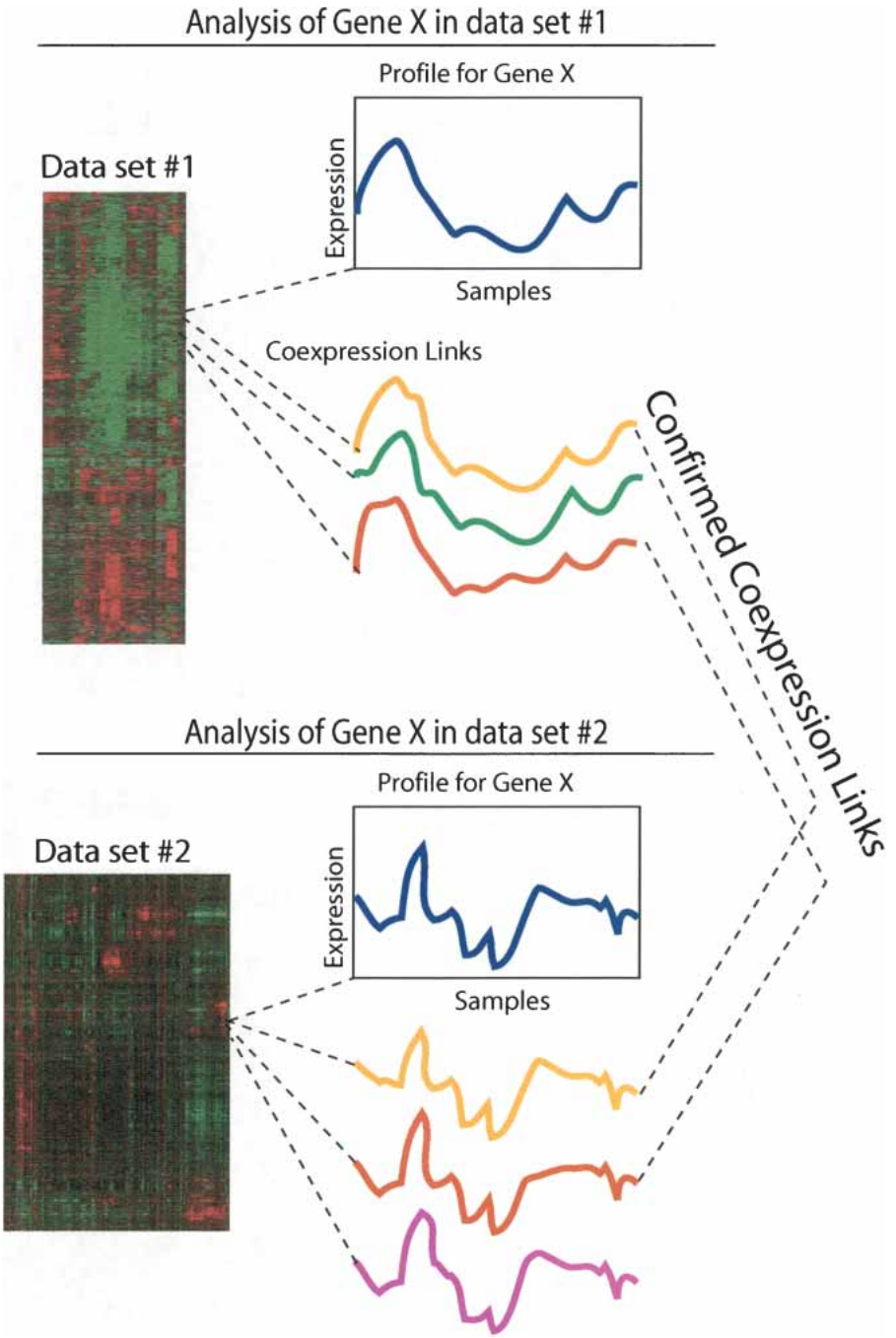

Figure 1 Schematic of the methodology. Only two data sets are shown here; our analysis made use of 60 data sets. The schematic outlines the analysis of a hypothetical "Gene X" in two data sets. First (*top*) in data set 1 we seek genes with expression profiles that are similar to that of Gene X, generating a set of raw "coexpression links." Only links that are deemed statistically significant in the context of data set 1 are stored. Then, we repeat this analysis in data set 2 (*bottom*). We then seek coexpression links that are common between the two data sets. This procedure is then repeated for each gene, and in more data sets. It is important to note that the profiles themselves need not be similar between data sets, nor do the profiles need to be "relevant" to any sample groups in the data sets. The data sets can also be from different microarray platforms, tissues, or species (though we present only human comparisons here). See Methods for details.

than four data sets (less than 0.5% of those found in the unshuffled data). However, for 2+ confirmed links the rate is 34%. We note that these tests examine the random occurrence of confirmed links, not the random occurrence of links within single data sets. When we instead shuffled the microarray expression profiles before raw link determination, we obtained almost no 3+ confirmed links (<10). For much of the remainder of our analysis we focus on 3+ confirmed links.

Out of 9.7 million unique coexpression links, 220,649 (2.2%) are seen in at least three data sets (3+ confirmed). In addition, 8805 of the genes tested have at least one 3+ confirmed link, encompassing 60% of the 14,172 genes that were expressed in at least three data sets (and thus capable of having 3+ confirmable links). Not surprisingly, genes with many raw links tended to have more 3+ confirmed links (Spearman's rank correlation 0.81; Fig. 2B).

Figure 2C shows the number of 3+ confirmed links per gene. The distribution approximately obeys a power law distribution, as is observed for many biological as well as other types of networks (Barabasi and Albert 1999; Jeong et al. 2001; Bhan et al. 2002; Featherstone and Broadie 2002). Thus most genes have only a few confirmed links, whereas a small number of genes have many 3+ confirmed links (up to a maximum of 913). No gene had all of its raw links confirmed: the highest 3+ confirmation rate was 0.19, and the highest 2+ confirmation rate was 0.66.

Although the numbers of positive and negative correlations we selected were fairly similar, a much larger fraction (88.8%) of confirmed links were for genes that showed positive correlations (a positive correlation in one data set and a negative correlation in another data set was not considered a confirmation). The overall 3+ confirmation rate for negative correlations was 0.5%, over seven times lower than the rate for positive correlations of 3.6%. Very few negative correlations (694) were confirmed at higher levels than 4+, and none were confirmed in more than eight data sets.

Functional Relevance of Link Confirmation

We predicted that as the level of confirmation of a link increases, it is more likely that the link is between two genes that are already known to have a functional relationship. We evaluated this by examining the overlap of Gene Ontology (GO) annotations for each pair of linked genes. This semantic similarity metric is reasonable because it reflects both the extent to which each gene has a known function, and the extent to which they are similar. This method will fail to detect known functional associations for genes that have poor GO annotations.

As links are increasingly confirmed, the semantic similarity of the genes also tends to increase (Fig. 3). Importantly, the dis-

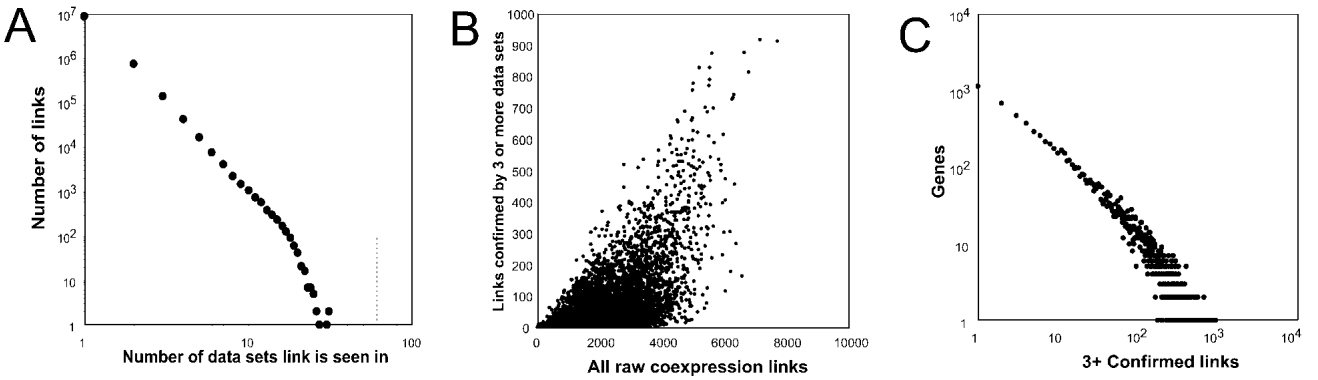

Figure 2 General properties of coexpression confirmation in the database. (*A*) Distribution of links at different levels of confirmation. The vertical dashed line marks the total number of data sets analyzed (60). Most links are not confirmed, but some links are confirmed in up to 31 data sets. (*B*) The number of "raw links" (those that are confirmed or not) plotted against the number of links that are confirmed in at least three data sets. Each point represents one gene. Genes with many raw links tend to have more confirmed links. (*C*) Degree distribution of links confirmed in at least three data sets.

tribution of GO term overlap for links that are seen only in a single data set is significantly different from randomly generated links (signed-rank test, $P < 10^{-15}$). This suggests that our initial link selection procedure is at least somewhat effective in selecting biologically relevant links, even if they are never confirmed in other data sets. Links that are confirmed two or more times have higher GO term overlaps than those seen only once ($P < 10^{-15}$), and those 3+ confirmed are significantly more similarly annotated than those at 2+ ($P < 10^{-15}$), each confirmation corresponding to about one additional GO term in common, on average. At high levels of confirmation, a high degree of known functional relatedness of the pairs is very likely, as shown by the curve for 15+ confirmations (Fig. 3). These findings were also confirmed using an alternative measure of semantic similarity (Lord et al. 2003). These results suggest that functional inferences based on confirmed coexpression have increased reliability. Interestingly, the effect of confirmation on increasing semantic similarity was weaker for negative correlations considered alone, and the genes in these pairs generally had lower semantic similarity scores (see Supplemental data).

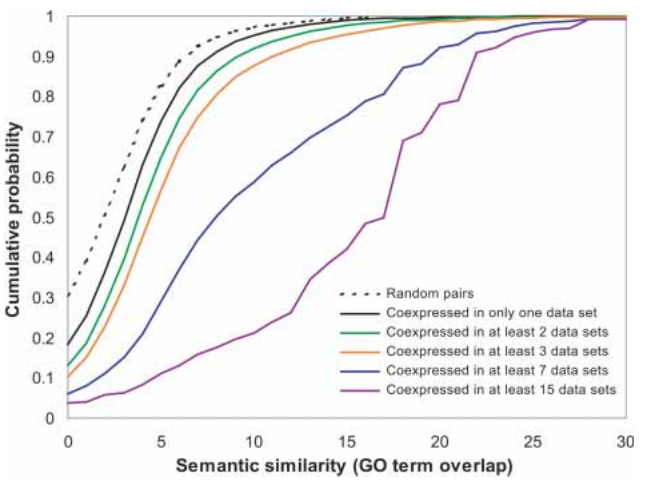

Figure 3 Relationship between link confirmation on semantic similarity of the selected genes. The *x*-axis indicates GO term overlap (see Methods). The cumulative distributions of semantic similarity scores for sets of links selected by different criteria are plotted. The dashed line indicates the distribution for randomly selected pairs of genes. Each solid curve is the cumulative probability distribution measured for pairs of genes identified by coexpression links at varying levels of confirmation (including both positive and negative correlations). The black curve is the distribution for coexpression pairs that are not confirmed. Confirmed links tend to have higher levels of GO term overlap. The *x*-axis is truncated at 30 (there are only 694 2+ pairs with more than 30 terms in common; the maximum is 95, for one pair).

Cluster Analysis of the Confirmed Coexpression Network

The set of coexpression links forms a network among the genes. The density of the 3+ network (the ratio of links between genes to the number of possible links) is 0.0057, with a diameter of 10 (the longest minimal path between two genes). The network can be broken into just 49 unconnected components, the largest of which contains almost all the genes (8705). The remaining 48 components contain only two or three genes each.

We used two clustering approaches to gain further insight into the structure of the gene interaction network predicted from confirmed coexpression. First, we used hierarchical clustering (Methods; Fig. 4). Because of the large size of the 3+ network, for this analysis we used the set of 7+ confirmed links, further limiting the analysis to those genes having at least six 7+ links (720 genes and 10,089 links). By applying hierarchical clustering to a matrix representation of the network, we identified a series of "core clusters" that appear along the diagonal of the matrix (left-hand side of Fig. 4). Interactions between genes in these core clusters appear as spots off the diagonal. The right-hand side of Figure 4 is a visualization of GO categories associated with each gene. The columns of the GO matrix were also clustered to put terms with similar patterns near each other.

A statistical analysis (see Methods) allowed us to associate many of the clusters with specific GO terms, illustrated by the color coding on the right-hand side of Figure 4. For example, a cluster of genes at the upper right is clearly associated with the GO terms related to protein translation including "cytosolic ribosome," and indeed includes many ribosomal proteins and translation initiation and elongation factors. A smaller identifiable cluster is represented by MHC II protein coding genes. The MHC II genes are associated with several other clusters containing many genes related to the immune response (in the lower left of the matrix, orange box). The middle of Figure 4 is dominated by a large, fairly diffuse cluster of about one-third of the genes (indicated by the light blue box) that contains within it several tighter groups of genes associated with GO terms related to RNA processing, DNA replication, and the cell cycle. The many links between these groups of genes (off the diagonal) may represent robust interactions between these processes. We stress that all of

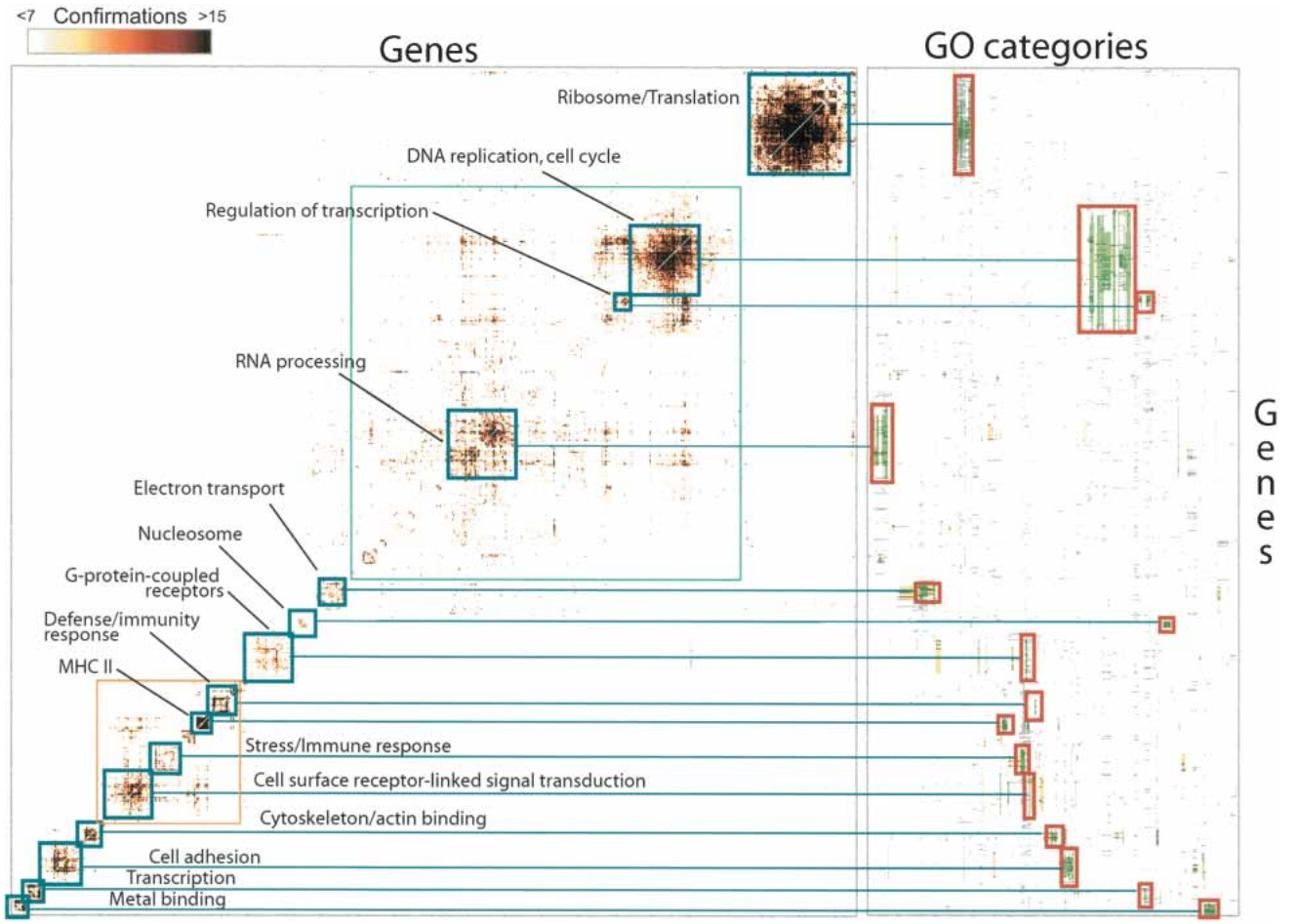

Figure 4 Hierarchical clustering of the coexpression network at a high level of confirmation. The *left-hand* side of the figure is the (diagonally symmetric) interaction matrix for 506 genes. Each color-coded entry is an interaction that is seen in seven or more data sets. The colored boxes indicate the main clusters, which are labeled according to their functional theme. A light blue box indicates a large diffuse cluster that dominates the *upper* half of the figure. A second box (orange) indicates several immune system-related clusters that are placed near each other. Blue lines connect many of the smaller clusters to the *right-hand* side of the figure, which depicts GO annotations for the same genes. On the *right-hand* side, each column represents a different GO term. The columns (495 GO terms) were arranged by hierarchical clustering, placing terms with similar annotation patterns together. The entries of the matrix are colored according to the status of the cluster-GO term association for the gene and term (see Methods). Green indicates term-cluster associations that were significant. Dark gray indicates the best GO term-gene cluster associations but that did not meet all criteria. Light gray points indicate GO terms-gene combinations that were not associated with a high-scoring cluster. These groups were used to define the cluster labels in the *left* half of the figure.

the coexpression events in Figure 4 were seen in at least seven different microarray data sets.

Although the hierarchical clustering approach yields a high-level overview, it is difficult to study individual genes in the network in this manner, and it was difficult to analyze larger networks. Therefore to analyze the network of 3+ confirmed genes, we used a second approach based on MCODE, an algorithm designed to identify groups of highly interconnected genes from networks (Bader and Hogue 2003). MCODE uses different criteria than hierarchical clustering to place genes in groups and can be used fruitfully on much larger networks. MCODE found between 29 and 363 clusters (depending on the input parameters for MCODE).

Two illustrative clusters are shown in Figure 5. Figure 5A shows a cluster of 15 genes, several of which are associated with the GO terms "cell junction" (CLDN3, CLDN4, CLDN7, CDH1) and "epidermal differentiation" (ELF3, CRABP2). Many of the other genes in this cluster have identified or suspected roles in the regulation of cell motility or tumor cell invasiveness (including DDR1, SPINT2, HRIHFB2122, TACSTD1, and WNT5A; Vogel et al. 1997; Seipel et al. 2001; Weeraratna et al. 2002). Our findings provide further evidence of a role for these genes in regulation of cell motility, and may be particularly useful in elucidating the functions of less clearly described genes in this cluster, such as MAL2 (Wilson et al. 2001). Another example is given in Figure 5B, a cluster of eight genes that includes seven genes known to play roles in sterol biosynthesis. The final gene, C14Orf1, has not been functionally characterized in mammalian cells and is poorly annotated in the public databases. However, it has been predicted to play a role in sterol biosynthesis based on the analysis of its yeast homolog, ERG28 (Gachotte et al. 2001). The coexpression pattern of C14Orf1 that we identified further suggests a role for this gene in sterol biosynthesis in humans. These examples serve to illustrate how the network of confirmed gene coexpression can be used to make new inferences or add support to existing hypotheses.

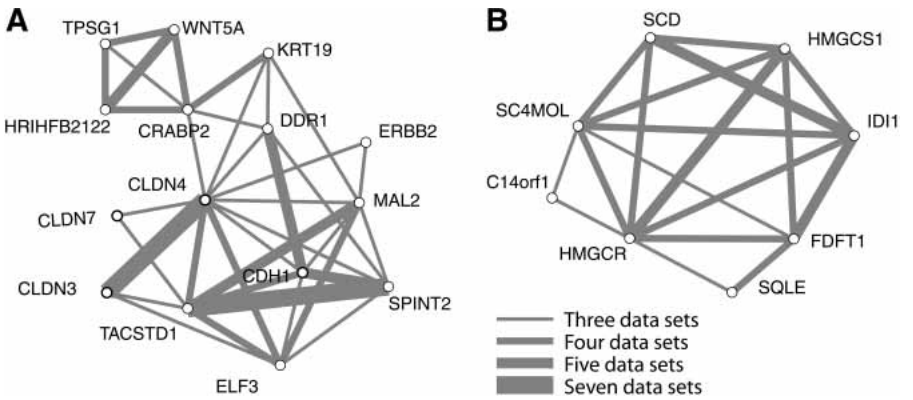

Figure 5 Examples of clusters extracted from the 3+ network with MCODE. See text for details. Increasing thickness of lines denotes increasing numbers of data sets in which the link was observed.

DISCUSSION

This study provides information on the structure of correlation-based links between genes in multiple microarray data sets. Our main goal was to establish whether comparing analyses across data sets is relevant to understanding gene function. The primary evidence that this is the case is that many genes show patterns of correlated expression that are reproducible across data sets, and that there is a clear relationship between confirmation of correlated expression and related gene function.

Reproducible coexpression links are found for numerous genes. This suggests that this type of analysis can be used rather broadly, and is not confined to use on a small set of genes. On the other hand, only a small fraction of all links were confirmed in at least three data sets. Though this suggests that many links seen only once may not be biologically relevant, our Gene Ontology analysis shows that even links that are never confirmed are substantially more informative than random data (Fig. 3). The obvious difficulty with using results that are never confirmed is identifying the meaningful novel relationships, and therefore focusing on confirmed coexpression seems preferable.

In order for a link to be confirmed, several criteria must be met. First, the pairs of genes must be present and detectably expressed in multiple data sets; a gene that is only represented in one data set will never have any confirmed links. In our database, not a single gene was considered detectable in all 60 data sets; the maximum was 57, for seven genes, and 5667 were detectable in 25 or more data sets. We also expect that confirmation of a link might be sample-type specific, even if the genes are expressed in all cases. Thus, two genes might be coexpressed only in leukemia data sets, even though they are expressed in other types of data sets. Because we used a fairly wide variety of data sets in our study, the lack of confirmation of many links could be due simply to lack of including appropriate data (there may also be a positive bias to the links discovered due to the particular data sets we studied). Finally, we may miss confirmations if our link selection criteria are too stringent.

When a link is seen in many data sets, it is increasingly likely that it represents a known functional relationship. This means that, to a certain extent, it is unlikely that many novel functional relationships will be found by seeking coexpression that is ubiquitous. We believe that confirmation near the 3+ level, or even 2+ for smaller data corpuses, will yield a higher fraction of novel relationships while still having a high enough degree of reliability. The exact level of confirmation required before one is motivated to seek additional evidence or perform follow-up studies is difficult to generalize, and our method provides a high degree of flexibility in how the results are interpreted. For some purposes, a higher level of confirmation may be worth the risks of losing information, whereas in other cases even links seen only a single time can be of value.

Most previous studies of gene networks have used data from unicellular organisms, primarily the budding yeast *Saccharomyces cerevisiae*. In yeast, it has been estimated that there are at least 30,000 interactions among the ~6000 protein gene products in the genome, based on a combined analysis of RNA microarray and protein–protein interaction data (von Mering et al. 2002). This yields a network density of about 0.0016 (the fraction of the possible interactions). In our network, which is based solely on microarray data, at the 3+ confirmation level we find over 220,000 coexpression relationships for 15,000 genes, or a density of about 0.002, although more than 6000 of these genes are "orphans" having no connections. When we consider only genes that are part of the network, the density is 0.0057, higher than the protein-protein interaction network density of 0.0013 found by Bader and Hogue (2003) for 4825 yeast genes. This apparent discrepancy might be explained by the difference in the types of interactions detectable by mRNA analysis compared to proteins, by species differences, and by the diversity of sample types and tissues our database contains. A commonality among the networks we obtain and those observed in previous work on yeast is that the link degree distribution follows a power law (Jeong et al. 2001; Bhan et al. 2002; Featherstone and Broadie 2002).

Negative correlations were much less likely to be confirmed in independent data sets. This was counter to our expectation because, in principle, negative correlations seem less likely to be the product of technically induced artifacts. Thus we expected the raw pool to be "cleaner" than for the positive correlations. There are several possible explanations for this result. One is that biologically meaningful negative correlations are harder to detect using microarrays, and our failure to detect them is due to experimental or analytical shortcomings. We may also not have appropriate data sets to confirm negative correlation links. A final explanation is that there may be biological reasons to favor positive coregulation of gene expression. We are unaware of any global analysis of this issue, though it may be relevant that active gene-specific transcriptional repression is a relatively uncommon regulatory mechanism in eukaryotes (Struhl 1999). Confirmed negative correlations were also less "biologically relevant" as measured by GO semantic similarity analysis. It is possible that this reflects limitations in available annotations.

We envision that databases of correlated expression will have many uses for biologists. One is to discover or confirm functional relationships that could only be made with low confidence from a single data set. Taken as a whole, the database represents a complex network of correlated expression that can be used for the analysis of large-scale properties of biological networks. It will also be of interest to integrate the information from correlated expression with other types of 'links,' including the GO approach we have taken thus far, as well information mined from literature databases and other experimental sources such as yeast two-hybrid data. Careful integration of heterogeneous data types will be essential to making full use of the accumulated expression data. Another topic of interest is coexpression that is conserved across species (Stuart et al. 2003). Our current database is also skewed towards tumor data, and we bear in mind that some of the interactions we observe may reflect a

disease state. Analysis of particular sample types or comparing sample types will be of interest.

To make our findings and database available for further evaluation and use by the scientific community, we have developed a simple Web interface to the database that can be accessed at http://microarray.cpmc.columbia.edu/tmm. The interface permits simple queries to extract the links for a gene at a desired degree of confirmation stringency. The interface also displays visualizations of the original microarray data that generated the coexpression links, and has hyperlinks to external databases for each set of linked genes to facilitate exploration of the results. We are also making available extracted tables of coexpression links from the entire database that can be used for further bioinformatic analysis.

METHODS

Data Preparation

Sixty human microarray data sets were included in this study, totaling 3924 arrays. All but one of the data sets is currently publicly available (the exception is the 'Sibille-pfc' data set). Major data sources were the Stanford Microarray Database (Sherlock et al. 2001) and the Gene Expression Omnibus (Edgar et al. 2002). Data sets were not subjected to any additional normalization, as all had been normalized when we obtained them. No imputation was used to replace missing data. For Affymetrix data sets, we downloaded "signal" or "average difference" data as supplied by the source; for ratiometric data, we obtained or computed log-transformed (base 2) ratios. These expression metrics were the inputs to the rest of the analysis. We filtered each data set to remove genes or data points with very low expression. Exceptions were two ratiometric data sets, where only unfiltered normalized ratios were available, and some data sets that had already been filtered by the originator. For data from Affymetrix GeneChips, the 30% of the probe sets with the lowest maximal expression across the samples in the data set were removed. Probe sets having all negative "average difference" values were excluded before applying this filter. For the 13 ratiometric data sets obtained from the Stanford database, measurements with signal to background ratios of less than 1.5 in both channels were removed. For all ratiometric data sets, genes missing more than 25% of the data were excluded from further analysis, up to the removal of 30% of the genes unless there were fewer than five data points present in a gene (to keep the filtering comparable to that used for the Affymetrix arrays). The identities of genes across microarray data sets was established using public annotations, primarily based on Unigene (Wheeler et al. 2001). Genes are referred to by their official names where known, based on information from RefSeq (Pruitt and Maglott 2001). Further details of the data sets used are available at http://microarray.cpmc.columbia.edu/tmm.

Coexpression Link Identification

After filtering, each gene expression profile was compared to all others using the standard Pearson correlation coefficient. Comparisons between genes involving fewer than five data points due to missing values were discarded. The significance of each correlation was assessed by assuming that the distribution of correlations under the null hypothesis of no correlation follows a t-distribution with $n - 2$ degrees of freedom, where n is the number of measurements in the expression profile (the number of samples). The assumptions inherent in this test were validated by comparing these P-values to those obtained by a permutation-based test on a subset of the data, indicating that large deviations from the assumptions were rare. P-values were corrected using Bonferroni correction for the number of genes tested such that the family-wise error rate was controlled at α=0.01 per data set (Westfall and Young 1993). In addition, pairs were only considered for further study if they were among the top 0.5% or lowest (most negative) 0.5% of correlations in the data set. This criterion was implemented to penalize data with many 'nonspecific' high correlations. The combination of criteria means that for all but the largest data sets, correlations with magnitudes below ~0.6–0.7 were rejected.

Correction for Multiply Represented Genes

An additional Bonferroni multiple test correction was applied to tests of genes that occurred multiple times on the array. For example, the required P-value threshold for a gene that occurred twice on an array was adjusted (multiplied) by a factor of two before comparison to the desired alpha of 0.01. When two such genes were compared, the adjustment was multiplicative. Due to this correction, 1.9 million raw links were rejected.

Link Confirmation

A coexpression link between two genes was termed "confirmed" if the link was observed in more than one data set. To measure the statistical significance of link confirmation, we created "shuffled" databases by generating random links between probes in each data set, maintaining the same degree distribution and number of links per data set as in the real data. The proportion of negative to positive correlations was also maintained. We created and analyzed 100 such databases to collect statistics on the occurrence of link confirmation by chance.

GO Similarity Metric

Each gene was characterized by the set of GO terms it is associated with (according to publicly available sources on the Gene Ontology Web site; Ashburner et al. 2000). Eighty-five percent of the genes analyzed had at least one GO term. Included in this set of terms are all parent terms in the GO hierarchy of the directly annotated terms. Genes that have many terms associated with them in this manner are described in greater detail than genes with only a few terms. The similarity k of a pair of genes A and B is measured simply by the number of terms they share, $|GO_A \cup GO_B|$ where GO_x denotes the set of GO terms for gene x. Pairs of genes where one or both genes have no terms are given scores of zero (i.e., $k = 0$ where $GO_A = \varnothing$ or $GO_B = \varnothing$). We also tested an alternative semantic similarity metric suggested by Lord et al. (2003). This metric appears to be highly correlated with our overlap metric (R ~0.7). We measured semantic similarity for all pairs of genes identified at each level of coexpression link confirmation as well as for 5,000,000 randomly selected pairs of genes. The resulting distributions were then compared.

Cluster Analysis

For hierarchical clustering, we express the network as an interaction matrix, which is a symmetric square matrix with entries indicating how many times a link was replicated (we provisionally consider the number of replications as a crude measure of the "strength" of a coexpression link). We applied hierarchical clustering to the rows and columns of the interaction matrix using "Xcluster" (http://genetics.stanford.edu/~sherlock/cluster.html) using the default parameters, combined with visualization using matrix2png (Pavlidis and Noble 2003). We also used a novel implementation of the MCODE algorithm (Bader and Hogue 2003), with Pajek visualization (Batagelj and Mrvar 1998). Pajek was also used to analyze global properties of the network. Individual MCODE runs used a vertex weight threshold of 0.05 or 0.0, with and without the 'fluff' procedure, run at a fluff threshold of 0.1 (Bader and Hogue 2003). The 3+ network was first filtered to remove 176 highly connected genes (those with more than 350 links) before applying MCODE.

To compare hierarchical clustering with GO annotations, we first identified all GO terms that were associated with at least five genes in the set under consideration, but that did not apply to more than 20% of the genes (to avoid overly general or specific terms). For each term we examined each branch of the hierarchical clustering tree to identify the branch with the highest overrepresentation of the term relative to the rest of the genes, flagging clusterings with $P < 0.05$ (cumulative hypergeometric distri-

bution and Bonferroni-corrected for the number of GO terms examined), which contained at least five genes, and had an average pairwise correlation of at least 0.5 (to avoid always detecting the entire data set as the optimal cluster). The GO annotations were then represented as a binary matrix, where each entry indicates whether a gene and GO term were associated. Note that this procedure only analyzes relative GO term enrichment within the genes used for clustering, not the entire database. We performed a similar analysis to help identify MCODE clusters that were enriched in particular GO terms.

ACKNOWLEDGMENTS

We sincerely thank the many groups who generously made their microarray data available, in some cases prior to publication, and to the organizers of the public microarray databases that facilitated data acquisition; Etienne Sibille, John Mann, and Victoria Arango for use of the human prefrontal cortex microarray data set; and Andrey Rzhetsky, Etienne Sibille, Gary Bader, Nick Socci, Agnes Viale, Alex Lash, and the anonymous reviewers for helpful suggestions. This work was supported in part by a pilot grant from the Avon Breast Cancer Foundation to P.P.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., et al. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403:** 503–511.

Allander, S.V., Nupponen, N.N., Ringner, M., Hostetter, G., Maher, G.W., Goldberger, N., Chen, Y., Carpten, J., Elkahloun, A.G., and Meltzer, P.S. 2001. Gastrointestinal stromal tumors with KIT mutations exhibit a remarkably homogeneous gene expression profile. *Cancer Res.* **61:** 8624–8628.

Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R., and Korsmeyer, S.J. 2002. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.* **30:** 41–47.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.

Bader, G.D. and Hogue, C.W. 2003. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4:** 2.

Barabasi, A.L. and Albert, R. 1999. Emergence of scaling in random networks. *Science* **286:** 509–512.

Batagelj, V. and Mrvar, A. 1998. Pajek: Program for large network analysis. *Connections* **21:** 47–57.

Bhan, A., Galas, D.J., and Dewey, T.G. 2002. A duplication growth model of gene expression networks. *Bioinformatics* **18:** 1486–1493.

Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., et al. 2001. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci.* **98:** 13790–13795.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., et al. 2000. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406:** 536–540.

Breitling, R., Sharif, O., Hartman, M.L., and Krisans, S.K. 2002. Loss of compartmentalization causes misregulation of lysine biosynthesis in peroxisome-deficient yeast cells. *Eukaryot. Cell* **1:** 978–986.

Butte, A.J., Tamayo, P., Slonim, D., Golub, T.R., and Kohane, I.S. 2000. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci.* **97:** 12182–12186.

Chang, H.Y., Chi, J.T., Dudoit, S., Bondre, C., van de Rijn, M., Botstein, D., and Brown, P.O. 2002. Diversity, topographic differentiation, and positional memory in human fibroblasts. *Proc. Natl. Acad. Sci.* **99:** 12877–12882.

Chaussabel, D., Semnani, R.T., McDowell, M.A., Sacks, D., Sher, A., and Nutman, T.B. 2003. Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites. *Blood* **102:** 672–681.

Chen, X., Cheung, S.T., So, S., Fan, S.T., Barry, C., Higgins, J., Lai, K.M., Ji, J., Dudoit, S., Ng, I.O., et al. 2002. Gene expression patterns in human liver cancers. *Mol. Biol. Cell* **13:** 1929–1939.

Cheok, M.H., Yang, W., Pui, C.H., Downing, J.R., Cheng, C., Naeve, C.W., Relling, M.V., and Evans, W.E. 2003. Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells. *Nat. Genet.* **34:** 85–90.

Choi, J.K., Yu, U., Kim, S., and Yoo, O.J. 2003. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* (Suppl.) **19:** I84–I90.

Dabrowski, M., Aerts, S., Van Hummelen, P., Craessaerts, K., De Moor, B., Annaert, W., Moreau, Y., and De Strooper, B. 2003. Gene profiling of hippocampal neuronal culture. *J. Neurochem.* **85:** 1279–1288.

Detours, V., Dumont, J.E., Bersini, H., and Maenhaut, C. 2003. Integration and cross-validation of high-throughput gene expression data: Comparing heterogeneous data sets. *FEBS Lett.* **546:** 98–102.

Dhanasekaran, S.M., Barrette, T.R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K.J., Rubin, M.A., and Chinnaiyan, A.M. 2001. Delineation of prognostic biomarkers in prostate cancer. *Nature* **412:** 822–826.

Diehn, M., Alizadeh, A.A., Rando, O.J., Liu, C.L., Stankunas, K., Botstein, D., Crabtree, G.R., and Brown, P.O. 2002. Genomic expression programs and the integration of the CD28 costimulatory signal in T cell activation. *Proc. Natl. Acad. Sci.* **99:** 11796–11801.

Dyrskjot, L., Thykjaer, T., Kruhoffer, M., Jensen, J.L., Marcussen, N., Hamilton-Dutoit, S., Wolf, H., and Orntoft, T.F. 2003. Identifying distinct classes of bladder carcinoma using microarrays. *Nat. Genet.* **33:** 90–96.

Edgar, R., Domrachev, M., and Lash, A.E. 2002. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30:** 207–210.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95:** 14863–14868.

Featherstone, D.E. and Broadie, K. 2002. Wrestling with pleiotropy: Genomic and topological analysis of the yeast gene expression network. *Bioessays* **24:** 267–274.

Gachotte, D., Eckstein, J., Barbuch, R., Hughes, T., Roberts, C., and Bard, M. 2001. A novel gene conserved from yeast to humans is involved in sterol biosynthesis. *J. Lipid Res.* **42:** 150–154.

Garber, M.E., Troyanskaya, O.G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G.D., Perou, C.M., Whyte, R.I., et al. 2001. Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl. Acad. Sci.* **98:** 13784–13789.

Ge, H., Liu, Z., Church, G.M., and Vidal, M. 2001. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat. Genet.* **29:** 482–486.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286:** 531–537.

Graeber, T.G. and Eisenberg, D. 2001. Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. *Nat. Genet.* **29:** 295–300.

Greenbaum, D., Luscombe, N.M., Jansen, R., Qian, J., and Gerstein, M. 2001. Interrelating different types of genomic data, from proteome to secretome: 'Oming in on function. *Genome Res.* **11:** 1463–1468.

Greenberg, S.A., Sanoudou, D., Haslett, J.N., Kohane, I.S., Kunkel, L.M., Beggs, A.H., and Amato, A.A. 2002. Molecular profiles of inflammatory myopathies. *Neurology* **59:** 1170–1182.

Gruvberger, S., Ringner, M., Chen, Y., Panavally, S., Saal, L.H., Borg, A., Ferno, M., Peterson, C., and Meltzer, P.S. 2001. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.* **61:** 5979–5984.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O.P., et al. 2001. Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* **344:** 539–548.

Hedenfalk, I., Ringner, M., Ben-Dor, A., Yakhini, Z., Chen, Y., Chebil, G., Ach, R., Loman, N., Olsson, H., Meltzer, P., et al. 2003. Molecular classification of familial non-BRCA1/BRCA2 breast cancer. *Proc. Natl. Acad. Sci.* **100:** 2532–2537.

Huang, Y., Prasad, M., Lemon, W.J., Hampel, H., Wright, F.A., Kornacker, K., LiVolsi, V., Frankel, W., Kloos, R.T., Eng, C., et al. 2001. Gene expression in papillary thyroid carcinoma reveals highly consistent profiles. *Proc. Natl. Acad. Sci.* **98:** 15044–15049.

Huang, E., Cheng, S.H., Dressman, H., Pittman, J., Tsou, M.H., Horng, C.F., Bild, A., Iversen, E.S., Liao, M., Chen, C.M., et al. 2003. Gene expression predictors of breast cancer outcomes. *Lancet* **361:** 1590–1596.

Jansen, R., Greenbaum, D., and Gerstein, M. 2002. Relating whole-genome expression data with protein–protein interactions. *Genome Res.* **12:** 37–46.
Jazaeri, A.A., Yee, C.J., Sotiriou, C., Brantley, K.R., Boyd, J., and Liu, E.T. 2002. Gene expression profiles of BRCA1-linked, BRCA2-linked, and sporadic ovarian cancers. *J. Natl. Cancer Inst.* **94:** 990–1000.
Jeong, H., Mason, S.P., Barabasi, A.L., and Oltvai, Z.N. 2001. Lethality and centrality in protein networks. *Nature* **411:** 41–42.
Kemmeren, P., van Berkum, N.L., Vilo, J., Bijma, T., Donders, R., Brazma, A., and Holstege, F.C. 2002. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell* **9:** 1133–1143.
Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., et al. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* **7:** 673–679.
Khatua, S., Peterson, K.M., Brown, K.M., Lawlor, C., Santi, M.R., LaFleur, B., Dressman, D., Stephan, D.A., and MacDonald, T.J. 2003. Overexpression of the EGFR/FKBP12/HIF-2α pathway identified in childhood astrocytomas by angiogenesis gene profiling. *Cancer Res.* **63:** 1865–1870.
Kuo, W.P., Jenssen, T.K., Butte, A.J., Ohno-Machado, L., and Kohane, I.S. 2002. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18:** 405–412.
Lee, P.D., Sladek, R., Greenwood, C.M., and Hudson, T.J. 2002. Control genes and variability: Absence of ubiquitous reference transcripts in diverse mammalian expression studies. *Genome Res.* **12:** 292–297.
Leung, S.Y., Chen, X., Chu, K.M., Yuen, S.T., Mathy, J., Ji, J., Chan, A.S., Li, R., Law, S., Troyanskaya, O.G., et al. 2002. Phospholipase A2 group IIA expression in gastric adenocarcinoma is associated with prolonged survival and less frequent metastasis. *Proc. Natl. Acad. Sci.* **99:** 16203–16208.
Lord, P.W., Stevens, R.D., Brass, A., and Goble, C.A. 2003. Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation. *Bioinformatics* **19:** 1275–1283.
Luo, J., Duggan, D.J., Chen, Y., Sauvageot, J., Ewing, C.M., Bittner, M.L., Trent, J.M., and Isaacs, W.B. 2001. Human prostate cancer and benign prostatic hyperplasia: Molecular dissection by gene expression profiling. *Cancer Res.* **61:** 4683–4688.
Ma, X.J., Salunga, R., Tuggle, J.T., Gaudet, J., Enright, E., McQuary, P., Payette, T., Pistone, M., Stecker, K., Zhang, B.M., et al. 2003. Gene expression profiles of human breast cancer progression. *Proc. Natl. Acad. Sci.* **100:** 5974–5979.
MacDonald, T.J., Brown, K.M., LaFleur, B., Peterson, K., Lawlor, C., Chen, Y., Packer, R.J., Cogen, P., and Stephan, D.A. 2001. Expression profiling of medulloblastoma: PDGFRA and the RAS/MAPK pathway as therapeutic targets for metastatic disease. *Nat. Genet.* **29:** 143–152.
Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., and Eisenberg, D. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature* **402:** 83–86.
Nielsen, T.O., West, R.B., Linn, S.C., Alter, O., Knowling, M.A., O'Connell, J.X., Zhu, S., Fero, M., Sherlock, G., Pollack, J.R., et al. 2002. Molecular characterisation of soft tissue tumours: A gene expression study. *Lancet* **359:** 1301–1307.
Pavlidis, P. and Noble, W.S. 2003. Matrix2png: A utility for creating matrix visualizations. *Bioinformatics* **19:** 295–296.
Perou, C.M., Jeffrey, S.S., van de Rijn, M., Rees, C.A., Eisen, M.B., Ross, D.T., Pergamenschikov, A., Williams, C.F., Zhu, S.X., Lee, J.C., et al. 1999. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci.* **96:** 9212–9217.
Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al. 2000. Molecular portraits of human breast tumours. *Nature* **406:** 747–752.
Pomeroy, S.L., Tamayo, P., Gaasenbeek, M., Sturla, L.M., Angelo, M., McLaughlin, M.E., Kim, J.Y., Goumnerova, L.C., Black, P.M., Lau, C., et al. 2002. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415:** 436–442.
Pruitt, K.D. and Maglott, D.R. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29:** 137–140.
Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., et al. 2001. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci.* **98:** 15149–15154.
Ramaswamy, S., Ross, K.N., Lander, E.S., and Golub, T.R. 2003. A molecular signature of metastasis in primary solid tumors. *Nat. Genet.* **33:** 49–54.
Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D., and Chinnaiyan, A.M. 2002. Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.* **62:** 4427–4433.
Rickman, D.S., Bobek, M.P., Misek, D.E., Kuick, R., Blaivas, M., Kurnit, D.M., Taylor, J., and Hanash, S.M. 2001. Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. *Cancer Res.* **61:** 6885–6891.
Rosenwald, A., Alizadeh, A.A., Widhopf, G., Simon, R., Davis, R.E., Yu, X., Yang, L., Pickeral, O.K., Rassenti, L.Z., Powell, J., et al. 2001. Relation of gene expression phenotype to immunoglobulin mutation genotype in B cell chronic lymphocytic leukemia. *J. Exp. Med.* **194:** 1639–1647.
Ross, D.T. and Perou, C.M. 2001. A comparison of gene expression signatures from breast tumors and breast tissue derived cell lines. *Dis. Markers* **17:** 99–109.
Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Van de Rijn, M., Waltham, M., et al. 2000. Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.* **24:** 227–235.
Seipel, K., O'Brien, S.P., Iannotti, E., Medley, Q.G., and Streuli, M. 2001. Tara, a novel F-actin binding protein, associates with the Trio guanine nucleotide exchange factor and regulates actin cytoskeletal organization. *J. Cell Sci.* **114:** 389–399.
Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J.C., Dwight, S.S., Kaloper, M., Weng, S., Jin, H., Ball, C.A., et al. 2001. The Stanford Microarray Database. *Nucleic Acids Res.* **29:** 152–155.
Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., et al. 2002. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* **8:** 68–74.
Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., et al. 2002. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1:** 203–209.
Smith, L.L., Coller, H.A., and Roberts, J.M. 2003. Telomerase modulates expression of growth-controlling genes and enhances cell proliferation. *Nat. Cell Biol.* **5:** 474–479.
Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., et al. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.* **98:** 10869–10874.
Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., et al. 2003. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci.* **100:** 8418–8423.
Staunton, J.E., Slonim, D.K., Coller, H.A., Tamayo, P., Angelo, M.J., Park, J., Scherf, U., Lee, J.K., Reinhold, W.O., Weinstein, J.N., et al. 2001. Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci.* **98:** 10787–10792.
Struhl, K. 1999. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell* **98:** 1–4.
Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302:** 249–255.
Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A., et al. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci.* **99:** 4465–4470.
Tezak, Z., Hoffman, E.P., Lutz, J.L., Fedczyna, T.O., Stephan, D., Bremer, E.G., Krasnoselska-Riz, I., Kumar, A., and Pachman, L.M. 2002. Gene expression profiling in DQA1*0501+ children with untreated dermatomyositis: A novel model of pathogenesis. *J. Immunol.* **168:** 4154–4163.
Vahey, M.T., Nau, M.E., Jagodzinski, L.L., Yalley-Ogunro, J., Taubman, M., Michael, N.L., and Lewis, M.G. 2002. Impact of viral infection on the gene expression profiles of proliferating normal human peripheral blood mononuclear cells infected with HIV type 1 RF. *AIDS Res. Hum. Retroviruses* **18:** 179–192.
van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415:** 530–536.
Virtaneva, K., Wright, F.A., Tanner, S.M., Yuan, B., Lemon, W.J., Caligiuri, M.A., Bloomfield, C.D., de La Chapelle, A., and Krahe, R. 2001. Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc. Natl. Acad. Sci.* **98:** 1124–1129.
Vogel, W., Gish, G.D., Alves, F., and Pawson, T. 1997. The discoidin domain receptor tyrosine kinases are activated by collagen. *Mol. Cell* **1:** 13–23.
von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S.,

and Bork, P. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417:** 399–403.
Weeraratna, A.T., Jiang, Y., Hostetter, G., Rosenblatt, K., Duray, P., Bittner, M., and Trent, J.M. 2002. Wnt5a signaling directly affects cell motility and invasion of metastatic melanoma. *Cancer Cell* **1:** 279–288.
Welle, S., Brooks, A., and Thornton, C.A. 2001. Senescence-related changes in gene expression in muscle: Similarities and differences between mice and men. *Physiol. Genomics* **5:** 67–73.
Welsh, J.B., Zarrinkar, P.P., Sapinoso, L.M., Kern, S.G., Behling, C.A., Monk, B.J., Lockhart, D.J., Burger, R.A., and Hampton, G.M. 2001. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc. Natl. Acad. Sci.* **98:** 1176–1181.
West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson Jr., J.A., Marks, J.R., and Nevins, J.R. 2001. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci.* **98:** 11462–11467.
Westfall, P.H. and Young, S.S. 1993. *Resampling-based multiple testing.* Wiley, New York.
Wheeler, D.L., Church, D.M., Lash, A.E., Leipe, D.D., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Tatusova, T.A., Wagner, L., et al. 2001. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **29:** 11–16.
Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O., et al. 2002. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13:** 1977–2000.
Wilson, S.H., Bailey, A.M., Nourse, C.R., Mattei, M.G., and Byrne, J.A. 2001. Identification of MAL2, a novel member of the mal proteolipid family, though interactions with TPD52-like proteins in the yeast two-hybrid system. *Genomics* **76:** 81–88.
Xin, W., Rhodes, D.R., Ingold, C., Chinnaiyan, A.M., and Rubin, M.A. 2003. Dysregulation of the annexin family protein family is associated with prostate cancer progression. *Am. J. Pathol.* **162:** 255–261.
Yeoh, E.J., Ross, M.E., Shurtleff, S.A., Williams, W.K., Patel, D., Mahfouz, R., Behm, F.G., Raimondi, S.C., Relling, M.V., Patel, A., et al. 2002. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* **1:** 133–143.
Yoon, H., Liyanarachchi, S., Wright, F.A., Davuluri, R., Lockman, J.C., de la Chapelle, A., and Pellegata, N.S. 2002. Gene expression profiling of isogenic cells with different TP53 gene dosage reveals numerous genes that are affected by TP53 dosage and identifies CSPG2 as a direct target of p53. *Proc. Natl. Acad. Sci.* **99:** 15632–15637.
Yuen, T., Wurmbach, E., Pfeffer, R.L., Ebersole, B.J., and Sealfon, S.C. 2002. Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res.* **30:** e48.

WEB SITE REFERENCES

http://microarray.cpmc.columbia.edu/tmm; Database and additional resources for analysis of coexpression across data sets.
http://genetics.stanford.edu/~sherlock/cluster.html; Clustering software.

Received August 26, 2003; accepted in revised form February 24, 2004.