Editorial Manager(tm) for Genomics
Manuscript Draft

Manuscript Number:

Title: Assessment and Integration of Publicly Available SAGE, cDNA Microarray, and Oligonucleotide Microarray Expression Data for Global Coexpression Analyses

Article Type: Regular Article

Section/Category:

Keywords: Gene Expression; Gene Expression Profiling; Microarray Analysis; cDNA Microarray; Oligonucleotide Microarray; Coexpression; Serial Analysis of Gene Expression (SAGE); Gene Ontology

Corresponding Author: Dr. Steven John Mathias Jones, PhD

Corresponding Author's Institution: Genome Sciences Centre, British Columbia Cancer Agency

First Author: Obi L Griffith, B.Sc.

Order of Authors: Obi L Griffith, B.Sc.; Erin D Pleasance, B.Sc.; Debra L Fulton, B.Sc.; Mehrdad Oveisi, M.Sc.; Martin Ester, PhD; Asim S Siddiqui, PhD; Steven JM Jones, PhD

Manuscript Region of Origin:

Abstract: Large amounts of gene expression data from several different technologies are becoming available to the scientific community. A common practice is to use this data to calculate global gene coexpression for validation or integration of other 'omic data. To assess the utility of publicly available datasets for this purpose we have analyzed Homo sapiens data from 1202 cDNA microarray experiments, 242 SAGE libraries and 667 Affymetrix oligonucleotide microarray experiments. The three datasets compared demonstrate significant but low levels of global concordance (rc < 0.11). Assessment against the Gene Ontology (GO) revealed that all three platforms identify more coexpressed gene pairs with common biological processes than expected by chance. As the Pearson correlation for a gene pair increased it was more likely to be confirmed by GO. The Affymetrix dataset performed best individually with gene pairs of correlation 0.9-1.0 confirmed by GO in 74% of cases. However, in all cases, gene pairs confirmed by multiple platforms were more likely to be confirmed by GO. We show that combining results from different expression platforms increases reliability of coexpression. A comparison with other recently published coexpression studies found similar results in terms of performance against GO but with each method producing distinctly different gene pair lists.

Cover letter to Editor

Dear Editor,

Large publicly available gene expression datasets are increasingly being used to determine coexpression for validation and integration in genomic studies, but few methods to assess and combine these coexpression predictions are available. We have just completed an extensive comparison and evaluation of large-scale datasets that is the first to include Serial Analysis of Gene Expression along with spotted cDNA and oligonucleotide microarray platforms. Such datasets are currently being used to invoke genes into biological processes, establish genes under similar genetic regulation and indeed form the basis of numerous manuscripts and inferences. Thus, we believe it is important to evaluate how well these datasets, being used extensively around the world by a large number of computational scientists, actually compare and perform. To this end, we assessed coexpression predictions from each dataset for internal consistency, cross-platform concordance, and biological confirmation with the Gene Ontology. Furthermore, we present an approach for combining coexpression predictions from different datasets to produce a high-confidence list of coexpressed gene pairs.   This resource is being used for identification of regulatory elements and is available for other genomic integration studies. We believe that the appraisal and results we present would be of general interest to the readers of Genomics.

Best wishes,

Steven Jones

Title Page

Assessment and Integration of Publicly Available SAGE, cDNA Microarray, and Oligonucleotide Microarray Expression Data for Global Coexpression Analyses

Authors

Obi L. Griffith[a], Erin D. Pleasance[a], Debra L. Fulton[b], Mehrdad Oveisi[a], Martin Ester[c], Asim S. Siddiqui[a] and Steven J.M. Jones[a]

Affiliations

a. Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, British Columbia, Canada, V5Z 4E6
b. Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada, V5A 1S6
c. School of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada, V5A 1S6

Email addresses:

OG, obig@bcgsc.ca
EP, epleasance@bcgsc.ca
DF, dlfulton@sfu.ca
MO, moveisi@bcgsc.ca
ME, ester@cs.sfu.ca
AS, asims@bcgsc.ca
SJ, sjones@bcgsc.ca

Corresponding author

Dr. Steven Jones
Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, British Columbia, Canada, V5Z 4E6
Tel: (604) 877-6083
Fax: (604) 876-3561
Email: sjones@bcgsc.ca

Introduction

Large-scale expression profiling has become an important tool for the identification of gene functions and regulatory elements. The development of three such techniques, cDNA microarrays [1], oligonucleotide (oligo) microarrays [2] and serial analysis of gene expression (SAGE) [3] has resulted in a plethora of studies attempting to elucidate cellular processes by identifying groups of genes that appear to be coexpressed.

Our motivation for this study was to explore the fecundity of large extant expression datasets to identify coexpressed genes and their utility as a resource for biological study. Coexpression data are increasingly used for validation and integration with other 'omic' data sources such as sequence conservation [4], yeast two-hybrid interactions [5,6], RNA interference [7] and regulatory element predictions [8] to name only a few. If different platforms or datasets produce widely different measures of coexpression it could have significant impacts on the results of such studies. Furthermore, methods to assess these datasets and identify a coherent, consistent picture of coexpression will be needed.

As increasing amounts of expression data are published and deposited in public databases, the issue of data integration becomes more important. High degrees of consistency within a platform have been reported for cDNA microarrays and Affymetrix oligonucleotide microarrays [9,10,11]. The reproducibility of SAGE has not been demonstrated as clearly given the time and cost required to produce individual SAGE libraries. However, a recent study showed a high degree of reproducibility and accuracy for microSAGE (a modification of SAGE) [12] and preliminary analysis of SAGE replicates has demonstrated high levels of correlation, similar to those seen for Affymetrix platforms (A. Delaney, pers. comm.). Cross-platform comparisons of gene expression values have found 'reasonable' correlations for matched samples, especially for more highly expressed transcripts [11,13,14,15,16,17,18,19]. Other comparisons have reported 'poor' correlations [15,18,20,21,22,23,24].

The correlations reported above were for expression levels or expression changes of individual genes, not coexpression of gene pairs. To our knowledge, only one study has examined the correlation of coexpression results from multiple platforms [25]. The authors compared matched Affymetrix oligonucleotide chips and spotted cDNA microarrays for the NCI-60 cancer cell panel. For each platform, the calculation involved determining the Pearson correlation (r) between expression profiles (across 60 cell lines) for all pairwise gene combinations. Then, a correlation of correlations (r_c) between the two platforms was determined. When all gene pairs were considered a global concordance of r_c = 0.25 was reported. As the correlation cutoff was increased, r_c improved steadily to 0.92 at a correlation cutoff of r = 0.91 (but only 28 of 2061 genes remained). Thus, for most gene pairs there is poor correlation of correlations for global coexpression values.

Genome wide coexpression analyses in *C. elegans* and *S. cerevisiae* have been used with some success to identify gene function or genes that are coregulated [26,27,28]. This "guilt-by-association" approach has received criticism because of high levels of noise and other problems inherent to the methods [29] but still holds great interest for biologists. If matched samples display questionable levels of consistency between expression profiles generated by different platforms the question remains as to how effectively unmatched samples from many different sources will compare. If two genes are coregulated (i.e. controlled by an identical set of transcription factors) they should display similar expression patterns across many conditions and be identified as coexpressed. This is the basic premise of many gene function and regulation studies. If true, large datasets from different expression platforms should identify the same coexpressed gene pairs even if derived from different conditions and tissues. However, it may be that few genes are globally coregulated and thus datasets comprised of different samples will identify different sets of coregulated genes. Similarly, noise and biases inherent to the different methods may result in highly discordant measures of coexpression, even for genes with similar function or under similar regulatory control. If true, the choice of expression platform and dataset could have significant effects on the outcome of integration and validation studies that use coexpression predictions.

The purpose of this study was to assess the differences between publicly available expression data for global coexpression analyses and investigate the value of combining multiple platforms to decrease noise and improve confidence in coexpression predictions. To explore this, we have compared large publicly available datasets for SAGE, cDNA microarray (cDNA), and Affymetrix oligonucleotide microarray (Affymetrix) platforms (Suppl. Fig. 1). We calculated all gene-to-gene Pearson correlation coefficients and assessed the platforms for internal consistency, cross-platform concordance, and agreement with the Gene Ontology. The Pearson correlation was chosen as a similarity metric because it is one of the most commonly used, with numerous published examples for Affymetrix [9,30,31], cDNA [5,27,32] and SAGE [33,34]. Because the datasets represent unmatched samples, a direct comparison of platforms is challenging. However, given that these datasets are being used individually in numerous studies we believe a relative assessment of the available data for each platform is critical. Our results indicate that the three platforms identify very different measures of coexpression for most gene pairs with a very low correlation of correlations between platforms. However, coexpression predictions become more reproducible with larger datasets and each of the three platforms performs better (identifies more gene pairs with common GO terms) as the Pearson correlation increases. Furthermore, gene pairs confirmed by more than one platform (high 2-platform average Pearson) were much more likely to share a GO term than those identified by only a single platform. Other recently published coexpression methods (TMM, ArrayProspector) also performed well against GO at higher scores but identified very different gene pairs. By using the Gene Ontology to choose thresholds of high-confidence pairs for each we identify a set of coexpressed gene pairs that represents the best of each approach.

Results

Internal Consistency

Before performing cross-platform comparisons, it is relevant to evaluate each platform individually to determine how consistently different experiments from one technology identify the same levels of gene coexpression. To this end, internal consistency was determined by dividing each of the datasets in half and comparing the gene-to-gene Pearson correlations for each subset (Figure 1A-C). We first divided the data in a purely random fashion. To make the internal consistency calculation more comparable to the cross-platform comparisons, we also devised a pseudo-random division which takes into account the presence of experimental replicates and very similar experimental conditions in the datasets (see methods).

Internal consistency was found to be dependent on the minimum number of common experiments (MCE) between any two genes on which Pearson correlations are calculated. MCE was defined as follows:

MCE – The minimum required number of common or shared experiments for which any two genes actually have values available in their respective expression profiles (Figure 1D).

Increasing the MCE increased the internal consistency but decreased the number of gene pairs considered for both the pseudo-random (Figure 1) and random (Suppl. Figure 2) division methods. With the random division, and an MCE of 100, Affymetrix showed the highest average internal correlation of 0.925, then cDNA microarray with correlation of 0.889, and SAGE with correlation of 0.776. This MCE cutoff was used by the group that provided the cDNA microarray data [4] (E. Segal, pers. comm.). As expected, the pseudo-random division, which groups replicates and experimental datasets, reduced internal consistencies with values of 0.253 for Affymetrix, 0.273 for the cDNA microarray and 0.660 for SAGE with MCE of 100 (Figure 1). Unfortunately, as the SAGE dataset contains only 242 samples, division into two groups of approximately 120 results in relatively few gene pairs that meet the criteria of 100 MCE (only 1518 pairs on average). Although approximately 60% of these SAGE libraries are derived from cancer samples, we found no evidence of an effect on the coexpression results (Suppl. Figure 3) and therefore included them in subsequent analysis.

Internal consistency is a measure of the reproducibility or robustness of gene coexpression predictions similar to a cross-validation test. This is based on the assumption that if a gene pair is truly coexpressed based on an expression dataset, it should be predicted as coexpressed by random subsets of the data. The consistency increases with higher MCE but at different rates for the three datasets because of their different natures in terms of number of experiments and experiment composition. Thus, it would be unfair to compare the datasets with MCEs that resulted in different levels of reproducibility. Studies generally choose some cutoff for a minimum number of common experiments such as 5, 10 or 100 [4,30,35]. In an effort to produce an unbiased comparison of the three platforms, the pseudorandom division was used to determine an appropriate MCE which would generate the same internal consistency ($r_c = 0.25$) for each (Affymetrix MCE = 95; cDNA MCE = 28; SAGE MCE = 23) (Figure 1). All internal consistency correlations are summarized in Table 1.

Cross-Platform Correlation Analysis

Considering that the levels of consistency between subsets of data from a single platform were relatively low (when replicates and similar experiments were kept together) it is not surprising that datasets from different platforms compared poorly against each other. All comparisons were found to have significant but poor positive correlations when compared to randomly permuted data ($p < 0.001$, 1000 permutations). Affymetrix versus cDNA showed the best correlation of 0.102, then Affymetrix versus SAGE with 0.086, and finally cDNA versus SAGE with 0.041 (Suppl. Figure 4). A Pearson rank analysis also showed significant but poor agreement with only 3-8% better performance than randomly permuted data (Suppl. Figure 5).

An analysis of correlation at different minimum Pearson cutoffs (r-cutoff) for gene pairs was performed as described previously [25] (Suppl. Fig. 6). Lee et al. (2003) observed a steady increase in global concordance (r_c = correlation of correlations) up to 0.92 at an r-cutoff of 0.91. Our data did not show such an obvious trend. Global concordance stayed close to zero (or even below) for all three pairwise platform comparisons up to 0.5-0.6 Pearson cutoff. The Affymetrix/cDNA correlation did show an improvement to $r_c = 0.163$ ($p = 0.003$, $n = 289$ gene pairs) at a r-cutoff = 0.65. Similarly the Affymetrix/SAGE comparison improved to $r_c = 0.290$ ($p = 0.028$, $n = 44$ gene pairs) at an r-cutoff = 0.7. After these cutoffs, both Affymetrix/cDNA and Affymetrix/SAGE comparisons returned to r_c values close to zero (or below) and were reduced to insignificant gene pair numbers. The cDNA/SAGE comparison showed no significant increases in r_c with any r-cutoff.

Gene Ontology Analysis

Since the datasets under study demonstrated little agreement, we attempted to determine which dataset was most 'biologically relevant'. GO biological process domain knowledge [36] was used to evaluate gene coexpression predictions for each platform. We hypothesized that genes which are coexpressed will be more likely to be involved in the same biological process. The number of gene pairs annotated to the same 'most specific' GO (Biological Process) term for each platform was determined (Suppl. Figure 7). In general, the datasets from all platforms perform better than expected by chance. Affymetrix performed best, followed by cDNA microarray and SAGE which performed about equally better than randomly permuted data. The analysis was also extended up the GO hierarchy to parent and grandparent terms, and identical trends and relationships were observed (Suppl. Fig. 8).

A second analysis looked at the relationship between the Pearson correlation and performance against GO. For each platform, the number of gene pairs annotated to the same 'most specific term' at different Pearson correlation ranges was determined (Figure 2). Generally, as Pearson correlation for a gene pair increases it is more likely to be confirmed by GO. With a Pearson value in the range of 0.3-0.4 or better the platforms always performed significantly better than randomly permuted data ($p < 0.001$, 1000 permutations). The improvement over randomly permuted data was very slight for the cDNA and SAGE datasets (2-4%). However, for the Affymetrix data, the trend was striking. Gene pairs identified as coexpressed with a Pearson correlation of 0.9-1.0 were confirmed by GO in 74% of cases. Gene pairs from this list include a large set of highly coexpressed protein biosynthesis genes as well as a few genes involved in translational elongation (a sub-process of protein biosynthesis) and muscle contraction. It should be noted that, in the case of the SAGE and cDNA datasets, only a few gene pairs had Pearson correlations > 0.9 (1 for cDNA, 5 for SAGE).

A third analysis examined the effect of averaging platform results and comparing to individual platforms using GO. Requiring coexpression evidence from multiple datasets may represent a method of reducing noise, and increase our confidence that coexpressed genes are actually coregulated. The percentage of gene pairs annotated to the same 'most specific term' at different average Pearson correlation ranges was determined as above. The results were again quite striking. With a 2-platform combined Pearson of 0.4 or greater the combined platforms all performed significantly better than randomly permuted data ($p < 0.005$, 1000 permutations). Furthermore, for any platform combination, a gene pair with an average Pearson correlation of $r > 0.6$ was much more likely to share a GO term than a gene pair with this level of correlation in only a single platform (Figure 3). For example, a gene pair with a two-platform average Pearson of 0.7-0.8 was found to share a common GO term 40-50% of the time. Pairs with this same Pearson range in individual datasets shared a common GO term only 5-10% of the time, only a few percent better than expected by chance. Gene pairs confirmed by multiple datasets ($r_{avg} > 0.6$ for any two-platforms) covered a wide range of GO categories (52 in total) (Suppl. Figure 9).

Comparison to other coexpression methods

Finally, an analysis was conducted to assess two other recent coexpression studies that were published while this analysis was in progress. The ArrayProspector method [37], the TMM method [35], and our 2-platform combination method (2PC) were each mapped to uniprot IDs and assessed using the same GO analysis as above. In all three cases, we observed significantly more gene pairs with common GO terms at higher scores (Figure 4). For our method (2PC), the percent of gene pairs with a common GO term rises sharply at a score of approximately 0.6-0.7. For, ArrayProspector this occurs at a score of approximately 0.7-0.8 and for TMM at a score of 5-6. At these cutoffs, each method represents 2,500 to 10,000 gene pairs. Each utilizes different genes and identifies different gene pairs as highly coexpressed. Thus, a comparison of the highest-scoring 2,500 gene pairs for each found only a minimal overlap of less than 10% (Figure 4D).

Discussion

We have shown that the genes identified as coexpressed are highly dependent on the dataset and expression platform used. In general, we find that the more data a correlation is based on, the more reproducible it is. When division of samples takes similar or replicate experiments into consideration, Affymetrix and cDNA internal consistencies level off at approximately $r_c = 0.25$ with MCE of about 90 and 30-40 respectively. The SAGE dataset continued to improve to nearly $r_c = 0.6$ with MCE of 80. This may reflect the diverse nature of the SAGE dataset for which libraries are rarely constructed from the same or similar tissue. In contrast, it is not uncommon for many Affymetrix or cDNA experiments to measure expression of a very similar series of samples. A recent yeast study found that the ability to correctly identify coregulated genes from coexpression analyses is highly dependent on the number of experiments with accuracy leveling off at 50 to 100 experiments [38]. Our results agree closely with this observation for human data and suggest that coexpression predictions will be most reproducible if based on 30 to 100 experiments. Furthermore, global coexpression analysis may benefit from a greater representation of tissues and conditions rather than greater numbers.

Given that different experimental subsets of the same platform show poor correlation it is perhaps not surprising that inter-platform comparisons show very poor correlations ($r < 0.11$). The fact that none of these data sets agree well raises some serious questions about their use for validation and integration with other data. There are several possible explanations for this observation: (1) The data comprising these datasets are so noisy as to prevent reliable identification of many truly coexpressed genes; (2) The method of identifying coexpressed genes is inadequate; (3) The unmatched and non-overlapping nature of the samples that make up each dataset result in identification of different subsets of truly coexpressed genes; and (4) Genes are under such complex regulatory control that genes coregulated in one cell-type or tissue behave in an entirely different manner in others and are therefore not globally coexpressed. It is likely that each of the explanations outlined above is to some degree responsible for the lack of concordance between coexpression analyses produced from different datasets and different platforms. It is not the purpose of this study to identify which is most important. Rather, we wish to make researchers aware that the choice of dataset or platform for integration or validation of other data could dramatically affect their results and methods that integrate or combine different platforms may be more appropriate.

The fact that intra-platform comparisons show some correlation and improve with number of data points suggests that some gene pairs identified are truly coexpressed. Furthermore, the GO analysis shows that gene pairs identified as highly coexpressed (higher Pearson correlation) are more likely to share the same biological process and thus actually be related. Similarly, gene pairs with lower Pearson correlations were as or less likely than random chance to share the same biological process. These results suggest that the Pearson correlation is a useful metric and that both high and low Pearson values have the meaning we expect. The GO analysis did not conclusively identify a single ‘correct’ platform or dataset but it did show that the Affymetrix dataset identified more biologically relevant gene pairs than the cDNA or SAGE datasets. However, gene pairs coexpressed in multiple expression platforms were much more likely to be confirmed by GO. Thus, combining platforms appears to act as a filter, producing high-confidence predictions from noisy datasets.

Recent investigations into the utility of combining expression data from different high-throughput platforms have identified highly variable levels of agreement. Based on an analysis of a small set of matched samples using oligonucleotide arrays, SAGE, and EST data, Haverty and colleagues [39] caution against the combination of platforms to confirm expression patterns for specific sets of genes. However, they do suggest that such methods can be used to extract high-confidence subsets of related genes. We agree that for many genes a poor level of agreement between datasets raises questions about their utility. However, our results do show that platform combination methods can be extended to large sets of unmatched publicly available expression data to produce biologically meaningful information.

As we were nearing completion of our analysis, a similar study using multiple microarray datasets (TMM) was published [35]. The authors examined 60 microarray datasets (cDNA and Affymetrix oligonucleotide) for gene pairs identified as coexpressed in multiple datasets. They report that even gene pairs confirmed by only a single dataset have better GO similarity scores than random pairs and GO score increases steadily with the number of confirmed links. Their method differs from ours in that experimental subsets are analyzed separately and a 'vote-counting' method was used to identify gene pairs that appear highly coexpressed (above some Pearson cutoff) in multiple sets. Our method combines all experimental subsets into a single dataset for each expression platform and then averages the global Pearson correlations between platforms. Our method is also the first to include SAGE data. A third recently published method (ArrayProspector), used a combination of singular value decomposition and kernel density estimation [37]. This method combines evidence from related arrays and weights the contribution of each array according to how well they correlate with functional annotation.

When attempting to infer function or coregulation from coexpression we should consider that it is likely that genes are biologically related in a number of different ways and therefore different methods will be required to identify each type of relationship. For example, one pair of genes might be 'tightly' coexpressed only under very specific conditions whereas another gene pair might be 'loosely' coexpressed across a broad range of conditions depending on the regulatory elements that they share. The three methods discussed above (TMM, ArrayProspector, and 2PC) represent three different approaches to the problem of identifying high-confidence coexpression for the purpose of inferring function or coregulation. Because the methods use different datasets, scoring methods, and comprise different gene sets, a direct comparison of the methods is difficult. Therefore, we chose to simply assess their respective predictions against GO independently. Thus, we do not identify the 'best' method but rather show that each method is at least partially effective based on performance against the Gene Ontology. Furthermore, because the highest-scoring pairs for each are almost completely non-overlapping we advocate combining the best results of each into a single set of high-confidence predictions. To this end we have chosen score thresholds for each method based on GO performance (2PC > 0.65; AP > 0.7; TMM > 7) and make available a list of 13,145 high-confidence coexpressed gene pairs (representing 2,979 unique genes) for use in regulatory element prediction or other integration studies.

Materials and methods

Data Sources

Human gene expression data for three major expression platforms were collected from public sources. We used a recently published data set of 1202 cDNA microarray experiments [4] representing 13595 genes, 242 SAGE libraries from the Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) representing 15,426 genes, and 667

Affymetrix HG-U133A oligonucleotide microarray experiments (889 were available but 667 had PMA detection calls) representing 8,106 genes, also from GEO (Suppl. Figure 1). cDNA microarray genes provided by Stuart *et al.* (2003) were identified by LocusLink ids [40]. Therefore this identifier was used for the other two platforms to allow the gene intersection of the three datasets to be determined and used for the subsequent analyses.

Data Filtering

cDNA microarray data for 13595 genes were used as provided by Stuart *et al.* (2003) except for minor formatting changes (see Suppl. Materials for our data). The 242 SAGE libraries ranged from 1,430 to 308,589 total tags in size with an average size of 52,723. SAGE data was first filtered to remove tags with less than one count in at least 10 libraries reducing the unique tags from 609,224 to 87,521 (and total tags from 12,758,981 to 11,219,373). Next, SAGE tags were mapped to genes by the lowest sense tag predicted from Refseq [40] or MGC [41] sequences and then mapped to LocusLink ids using the DiscoverySpace software package (Varhol et al., unpublished, http://www.bcgsc.ca/discoveryspace/) reducing the tag set further to 47,263 unique tags. In the event of discrepancy between Refseq and MGC, the former was taken as correct. If a tag mapped to more than one LocusLink or more than one tag mapped to the same LocusLink it was discarded resulting in a final set of 15,426 unique tags (2,762,500 total tags) confidently mapped to LocusLink ids. 22215 Affymetrix probe ids were mapped to 20577 LocusLink Ids using the most current Affymetrix annotation file for the HG-U133A chip (www.affymetrix.com, Suppl. Materials). As with the SAGE tags, probes with ambiguous mapping to LocusLink were discarded resulting in a final set of 8106 genes from the Affymetrix dataset. Once LocusLink ids were available for all three platforms, the intersection was determined. This subset of 5881 genes, present in all three platforms, was used for all subsequent analyses. The final 5881 unique SAGE tags represent 1,173,430 total tags sequenced.

Distance Calculations

Ratio values for the cDNA microarray data were used as is for the Pearson calculation. Affymetrix probe intensities were converted to natural log values. All ln(intensity) values were normalized by subtracting the median and dividing by the inter-quartile range for the experiment [42]. Only Affymetrix probe intensities with a ‘P’ call were considered (p-value < 0.04). Intensities with ‘A’ or ‘M’ calls were set to null. To compensate for different library sizes SAGE tag counts were normalized to 10,000-tags/library and log-transformed as follows [34]:

Tag frequency = ln((tag count x 10000)/total tags in library).

SAGE tag counts of zero were converted to nulls. In all platforms, genes are represented by a vector of expression values for all the experiments in the data set. In each case, genes have null values if not represented on that array (cDNA), no tags observed (SAGE), or intensity not significantly detected (Affymetrix). Thus, when calculating Pearson correlations between gene pairs, the number of shared data points varied from zero to the total number of experiments. A minimum number of common experiments

(MCE) were required for each gene pair to provide some confidence in the value calculated (a Pearson correlation based on observations from only two experiments is meaningless). A range of MCEs was used for the internal consistency analysis (see below) and then one minimum chosen for subsequent analyses.

A Pearson correlation coefficient was calculated for all possible gene pairs for each platform as a measure of expression similarity. These calculations were performed by a modified version of the C clustering library [43] on 64-bit opteron linux machines with 8-32GB memory. Please see supplementary materials for modified C source code and explanation of changes.

Correlation of correlations analysis

Correlation of correlations (r_c) for internal consistencies and platform comparisons were performed as previously described [25] using the Pearson correlation function (cor) of the R statistical package (version 1.8.1). This correlation involves millions of data points and thus can not be graphed easily. Therefore, data were binned and density plots created using the Bioconductor hexbin (version 1.0.3) add-in function for R [44].

Internal consistency analysis

To evaluate the consistency of coexpression observed within each platform, we divided the experiments available and determined coexpression for each subset independently. If a platform consistently finds coexpressed genes regardless of the exact experiments involved, the r_c will be close to 1. To determine whether the observed r_c is significant, we repeat the procedure with randomly permuted gene expression values, expecting a r_c close to 0.

Pseudo-Random Division Method

Division was performed first randomly, and then pseudo-randomly. The pseudo-random division was necessary to prevent artificially high internal consistencies resulting from comparing mostly replicates (or very similar experiments) in the two subsets. In many cases (especially for the Affymetrix data) experimental replicates or very similar samples exist in the dataset. The purpose of coexpression analysis is to identify genes that behave similarly across many conditions. The internal consistency analysis is meant to measure how consistently a series of experiments across different conditions would identify the same coexpressed genes. If the two subsets of experiments contain replicates, they are more likely to identify the same coexpressed genes as the expression values of the replicates will be very similar. The cross-platform comparisons do not have this advantage because they consist of different experiments. Thus, to make the internal consistency calculation more comparable to the cross-platform comparisons, we used a pseudo-random division for subsequent analysis. Experiments were randomly divided into two subsets but experiments belonging to the same experimental series (Affymetrix), publication (cDNA), or tissue (SAGE) were required to fall into the same subset.

Minimum Common Experiments Analysis

Differences in the number of common experiments between any two genes result from missing values in the data matrices. In the case of the cDNA microarray data, different arrays were used in different experiments, and not all genes are present on all the arrays. For SAGE, a tag is often observed in one library but will have a zero tag count in other libraries. For Affymetrix oligonucleotide arrays, an intensity is always reported for every probe but in some cases the Affymetrix statistical software will determine that the probe was not reliably detected and assign an absent (A) or marginal (M) call instead of a present (P) call for that probe. As missing SAGE tags and probes not called Present represent genes expressed below the detection threshold of the SAGE and Affymetrix array experiments, we did not include these data in our analysis. Thus, for each dataset, there were gene pairs that were rarely represented in the same experiment and their Pearson correlations were based on very few data points. The effect of number of common experiments on internal consistency was determined by calculating the internal consistency for a series of datasets with different minimum common experiment (MCE) criteria. 100 different pseudo-random divisions were performed to get an average internal consistency for each MCE. An MCE was chosen for each such that the same internal consistency would result ($r = 0.25$) (Figure 1). Thus, all subsequent analyses were based on an MCE of 95 for Affymetrix, 28 for cDNA, and 23 for SAGE. Requiring an MCE removes gene pairs from the datasets. To maintain an unbiased comparison, only the 1,173,330 gene pairs common to all three platform datasets (after application of MCE criteria) were used in the subsequent platform comparisons.

Platform Comparisons

As with the internal consistency analysis, a correlation of gene correlations was calculated, but was determined for each of the three pairwise platform comparisons instead of between subsets of one platform. If the two platforms being compared report the same correlation between each gene pair, we expect the overall correlation between platforms would be near 1. The global concordance (r_c) was determined for increasing gene correlation cutoffs to compare to results obtained in the NCI-60 study [25].

Gene Ontology Analysis

The Gene Ontology (GO) is a controlled vocabulary that describes the roles of genes and proteins in all organisms [36]. GO is composed of three independent ontologies: biological process, molecular function, and cellular component. The GO descriptive terms are represented as nodes connected by directed edges that may have more than one parent node (directed acyclic graph). A gene is annotated to its most specific GO term description and all ancestor GO terms are implied.

The Gene Ontology (GO) MySQL database dump (release 200402 of assocdb) was downloaded from http://www.godatabase.org/dev/database. A GO MySQL database was built and a Perl script was developed to extract three GO information subspaces from the biological process ontology: 1) the most specific GO terms for each gene; 2) the most specific terms along with their associated parent terms; and 3) the most specific terms along with their associated parent and grandparent terms. Two categories of annotations were used for the evaluation of each GO information subspace: 1) gene annotations that did not include those derived from inferred electronic annotations (IEAs) (1007 genes found in common with our data set) and 2) gene annotations including IEAs (1426 genes found in common with our data set). Similar results were obtained for both non-IEA and IEA analyses. Only the IEA results are reviewed in the figures and text.

One potential issue with our analysis is that of a circular argument. It is possible that a coexpressed gene pair could be found to share a common GO term that was annotated for both genes by a coexpression analysis. Thus, coexpression data could be confirming coexpression data. To check for this problem we assessed the degree to which our dataset depends on annotations inferred from expression profiles (IEP evidence code). Only 93 of 32669 biological process annotations use IEP evidence, corresponding to only 73 genes with one or more IEP annotations. Of these, only 1 was present in our gene set and this gene also had non-IEP annotations. Therefore the potential for a circular argument is negligible.

Results shown in Suppl. Figure 7 were extracted from the gene pair correlation data, by enumerating the number of gene pairs found at common GO terms across a gene's expression similarity neighborhood for each GO information subspace. Results shown in Figure 2 were extracted by enumerating the number of gene pairs found at common GO terms for each range of Pearson correlations from 0 to 1 in increments of 0.1. The results summarized in Figure 3 were enumerated in a similar manner but used average Pearson correlations between two platforms instead of individual Pearson correlations. 1000 random permutations of the data were conducted to determine how often GO confirmation of a gene pair at each neighborhood or Pearson range would occur by chance. Scripts were written in Perl and are available at: http://www.bcgsc.ca/gc/bomge/coexpression/suppl_materials.

Comparison to other coexpression methods

Results shown in Figure 4 were generated using the GO analysis method described above for Figures 2 and 3. ArrayProspector (AP) data was obtained by request from the author [37]. Only pairs with scores above 0.150 were provided. TMM data was downloaded from the authors' supplemental webpage (see web references) [35]. Both negative and positive correlations were included and thus a gene pair can appear twice. Only pairs with scores of 1 or greater were provided. The 2-platform combination (2PC) method represents all 2-platform averages (Affymetrix/cDNA, Affymetrix/SAGE, and cDNA/SAGE). Thus, a gene pair can appear as many as three times if all three pairwise averages fall within the 0-1 range graphed. All datasets were converted from their respective identifiers to Uniprot [45] and the percent of gene pairs found at common GO terms for each range of scores determined. The top 2,500 pairs of each were examined to determine the overlap in results for high scoring pairs. Thresholds for a high-confidence set of coexpressed gene pairs were chosen for each method at the approximate respective score where performance was at least 3 to 4 times better than random chance (2PC > 0.65; AP > 0.7; TMM > 7).

List of abbreviations:

SAGE, Serial Analysis of Gene Expression; GEO, Gene Expression Omnibus; GO, Gene Ontology; IEA, Inferred Electronic Annotation; MGC, Mammalian Gene

Collection; MCE, Minimum number of Common Experiments; r, Pearson correlation; r_c, Correlation of Pearson correlations.

Acknowledgments
We thank Misha Bilenky, Jaswinder Khattra, Sheldon McKay, Gordon Robertson, Peter Ruzanov, Kim Wong, and Scott Zuyderduyn for invaluable help and advice, and Allen Delaney and Marco Marra for editorial suggestions. Eran Segal and Josh Stuart provided useful information regarding the published cDNA data. Lars Jensen (Peer Bork Lab) and Paul Pavlidis generously provided their coexpression datasets through personal communication or supplemental web pages. OG and EP are Michael Smith Foundation for Health Research Trainees and are supported by the Natural Sciences and Engineering Research Council of Canada. DF is supported by the CIHR/MSFHR Bioinformatics Training Program. SJ is a Michael Smith Foundation for Health Research Scholar. We also gratefully acknowledge the support of Genome BC, the BC Cancer Agency, and the BC Cancer Foundation.

SUPPLEMENTARY MATERIALS

All necessary data will be provided on a supplementary materials webpage hosted by the Genome Sciences Centre at:
http://www.bcgsc.ca/gc/bomge/coexpression/suppl_materials

WEBSITE REFERENCES

http://www.ncbi.nlm.nih.gov/geo/, The Gene Expression Omnibus.
http://www.r-project.org/, R Statistical Package Home Page
http://www.bioconductor.org/, Bioconductor Home Page
http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm, The C Clustering Library Home Page
http://www.affymetrix.com/, Affymetrix Home Page
http://cmgm.stanford.edu/~kimlab/multiplespecies/Supplement/, Stuart *et al.* Data Home Page
http://www.cytoscape.org/, Cytoscape Home Page
http://www.bork.embl.de/ArrayProspector, ArrayProspector
http://microarray.genomecenter.columbia.edu/tmm/, TMM data
http://www.bcgsc.ca/discoveryspace/, DiscoverySpace Software

References

[1] M. Schena, D. Shalon, R.W. Davis, P.O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science 270 (1995) 467-70.

[2] D.J. Lockhart, et al., Expression monitoring by hybridization to high-density oligonucleotide arrays, Nat Biotechnol 14 (1996) 1675-80.

[3] V.E. Velculescu, L. Zhang, B. Vogelstein, K.W. Kinzler, Serial analysis of gene expression, Science 270 (1995) 484-7.

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

[5] S. Li, et al., A map of the interactome network of the metazoan C. elegans, Science 303 (2004) 540-3.

[6] P. Kemmeren, et al., Protein interaction verification and functional annotation by integrated analysis of genome-scale data, Mol Cell 9 (2002) 1133-43.

[7] A.J. Walhout, et al., Integrating interactome, phenome, and transcriptome mapping data for the C. elegans germline, Curr Biol 12 (2002) 1952-8.

[8] E. Segal, et al., Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, Nat Genet 34 (2003) 166-76.

[9] E.J. Yeoh, et al., Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling, Cancer Cell 1 (2002) 133-43.

[10] A. Nimgaonkar, et al., Reproducibility of gene expression across generations of Affymetrix microarrays, BMC Bioinformatics 4 (2003) 27.

[11] P.K. Tan, et al., Evaluation of gene expression measurements from commercial microarray platforms, Nucleic Acids Res 31 (2003) 5676-84.

[12] S. Blackshaw, et al., MicroSAGE is highly representative and reproducible but reveals major differences in gene expression among samples obtained from similar tissues, Genome Biol 4 (2003) R17.

[13] L. Huminiecki, A.T. Lloyd, K.H. Wolfe, Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases, BMC Genomics 4 (2003) 31.

[14] V. Detours, J.E. Dumont, H. Bersini, C. Maenhaut, Integration and cross-validation of high-throughput gene expression data: comparing heterogeneous data sets, FEBS Lett 546 (2003) 98-102.

[15] A.K. Jarvinen, et al., Are data from different gene expression microarray platforms comparable?, Genomics 83 (2004) 1164-8.

[16] C.A. Iacobuzio-Donahue, et al., Highly expressed genes in pancreatic ductal adenocarcinomas: a comprehensive characterization and comparison of the transcription profiles obtained from three major technologies, Cancer Res 63 (2003) 8614-22.

[17] H.L. Kim, Comparison of oligonucleotide-microarray and serial analysis of gene expression (SAGE) in transcript profiling analysis of megakaryocytes derived from CD34+ cells, Exp Mol Med 35 (2003) 460-6.

[18] A.T. Rogojina, W.E. Orr, B.K. Song, E.E. Geisert, Jr., Comparing the use of Affymetrix to spotted oligonucleotide microarrays using two retinal pigment epithelium cell lines, Mol Vis 9 (2003) 482-96.

[19] M. Ishii, et al., Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis, Genomics 68 (2000) 136-43.

[20] S.J. Evans, et al., Evaluation of Affymetrix Gene Chip sensitivity in rat hippocampal tissue using SAGE analysis. Serial Analysis of Gene Expression, Eur J Neurosci 16 (2002) 409-13.

[21] J. Li, M. Pankratz, J.A. Johnson, Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays, Toxicol Sci 69 (2002) 383-90.

[22] W.P. Kuo, T.K. Jenssen, A.J. Butte, L. Ohno-Machado, I.S. Kohane, Analysis of matched mRNA measurements from two different microarray technologies, Bioinformatics 18 (2002) 405-12.

[23] N. Mah, et al., A comparison of oligonucleotide and cDNA-based microarray systems, Physiol Genomics 16 (2004) 361-70.

[24] J. Lu, A. Lal, B. Merriman, S. Nelson, G. Riggins, A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips, Genomics 84 (2004) 631-6.

[25] J.K. Lee, et al., Comparing cDNA and oligonucleotide array data: concordance of gene expression across platforms for the NCI-60 cancer cells, Genome Biol 4 (2003) R82.

[26] D.J. Allocco, I.S. Kohane, A.J. Butte, Quantifying the relationship between co-expression, co-regulation and gene function, BMC Bioinformatics 5 (2004) 18.

[27] S.K. Kim, et al., A gene expression map for Caenorhabditis elegans, Science 293 (2001) 2087-92.

[28] S.A. Jelinsky, P. Estep, G.M. Church, L.D. Samson, Regulatory networks revealed by transcriptional profiling of damaged Saccharomyces cerevisiae cells: Rpn4 links base excision repair with proteasomes, Mol Cell Biol 20 (2000) 8157-67.

[29] J. Quackenbush, Genomics. Microarrays--guilt by association, Science 302 (2003) 240-1.

[30] E.J. Williams, D.J. Bowles, Coexpression of neighboring genes in the genome of Arabidopsis thaliana, Genome Res 14 (2004) 1060-7.

[31] B.H. Mecham, et al., Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements, Nucleic Acids Res 32 (2004) e74.

[32] D.T. Ross, et al., Systematic variation in gene expression patterns in human cancer cell lines, Nat Genet 24 (2000) 227-35.

[33] M. Nacht, et al., Molecular characteristics of non-small cell lung cancer, Proc Natl Acad Sci U S A 98 (2001) 15203-8.

[34] D.A. Porter, et al., A SAGE (serial analysis of gene expression) view of breast tumor progression, Cancer Res 61 (2001) 5697-702.

[35] H.K. Lee, A.K. Hsu, J. Sajdak, J. Qin, P. Pavlidis, Coexpression Analysis of Human Genes Across Many Microarray Data Sets, Genome Res. 14 (2004) 1085-1094.

[36] M. Ashburner, et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat Genet 25 (2000) 25-9.

[37] L.J. Jensen, J. Lagarde, C. von Mering, P. Bork, ArrayProspector: a web resource of functional associations inferred from microarray expression data, Nucleic Acids Res 32 (2004) W445-8.

[38] K. Yeung, M. Medvedovic, R. Bumgarner, From co-expression to co-regulation: how many microarray experiments do we need?, Genome Biology 5 (2004) R48.

[39] P.M. Haverty, L.L. Hsiao, S.R. Gullans, U. Hansen, Z. Weng, Limited agreement among three global gene expression methods highlights the requirement for non-global validation, Bioinformatics 20 (2004) 3431-41.

[40] K.D. Pruitt, K.S. Katz, H. Sicotte, D.R. Maglott, Introducing RefSeq and LocusLink: curated human genome resources at the NCBI, Trends in Genetics 16 (2000) 44-47.

[41] Mammalian Gene Collection Program Team*, et al., Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences, PNAS 99 (2002) 16899-16903.

[42] G.S. Davidson, B.N. Wylie, K.W. Boyack (2001). Cluster Stability and the Use of Noise in Interpretation of Clustering, pp. 23, IEEE Computer Society.

[43] M.J. de Hoon, S. Imoto, J. Nolan, S. Miyano, Open source clustering software, Bioinformatics 20 (2004) 1453-4.

[44] R. Ihaka, R. Gentleman, R: A language for data analysis and graphics., Journal of Computational & Graphical Statistics 5 (1996) 299.

[45] A. Bairoch, et al., The Universal Protein Resource (UniProt), Nucleic Acids Res 33 Database Issue (2005) D154-9.

Legends

Table 1. Summary of r_c values for internal consistency analysis using different sample division methods and MCE cutoffs.

Note that many different divisions are possible for each result below (except cancer/normal). Gene pair and r_c values represent mean values from 100 different random or pseudo-random divisions.

Figure 1. Internal consistency and minimum common experiments analysis using pseudo-random division method.

For each gene pair, the number of common experiments is determined as the number of experiments for which expression values are available for both genes. On the left axis, MCE is plotted against internal consistency. On the right axis, MCE is plotted against number of gene pairs. In general, as more MCE are required, less gene pairs meet the criteria but the internal consistency improves as the correlation is based on more expression data. Notice that the Affymetrix (A) and cDNA (B) datasets appear to level off at approximately $r_c = 0.3$ with 100 MCE. However, the SAGE correlation (C) continues to improve up to nearly $r_c = 0.7$ before zero genes meet the cutoff and a leveling is not observed. Data represent mean r_c value and gene pair number of 100 pseudo-random divisions at each MCE. Error bars indicate one standard deviation.

Figure 2. GO Correlation Range Analysis.

At higher Pearson correlations (in particular, $r > 0.8$) gene pairs are more likely to have similar GO biological processes, although very few gene pairs have high correlations in SAGE and cDNA datasets. 75% of gene pairs with correlation > 0.9 calculated from Affymetrix data have the same GO annotation. Interestingly, gene pairs with very low Pearson values are less likely to share a common GO term than randomly permuted data. Random lines represent mean values from 1000 random permutations. Error bars indicate one standard deviation.

Figure 3. GO Correlation Range Analysis for Multi-Platform Average

Comparison of two-platform average Pearson to individual platform indicates that gene pairs identified as coexpressed in multiple platforms (higher average Pearson) are much more likely to be confirmed by GO. Random line represents mean values from 1000 random permutations of all two-platform combinations. Error bars indicate one standard deviation.

Figure 4. Comparison of 2-Platform Combination Method to Other Recent Coexpression Methods. (A-C) For each method, gene pairs with higher scores are more likely to share a common GO term. Lines with hollow squares represent numbers of gene pairs (right axis). Lines with solid triangles represent % of gene pairs with a common GO term. Random lines represent mean values from 1000 random permutations. Error bars indicate one standard deviation. (D) Venn diagram indicates overlap between the 2,500 top scoring pairs for each method (not required to be in GO). Each method is comprised of different datasets and has different genes. Therefore a direct comparison of method performance is difficult.

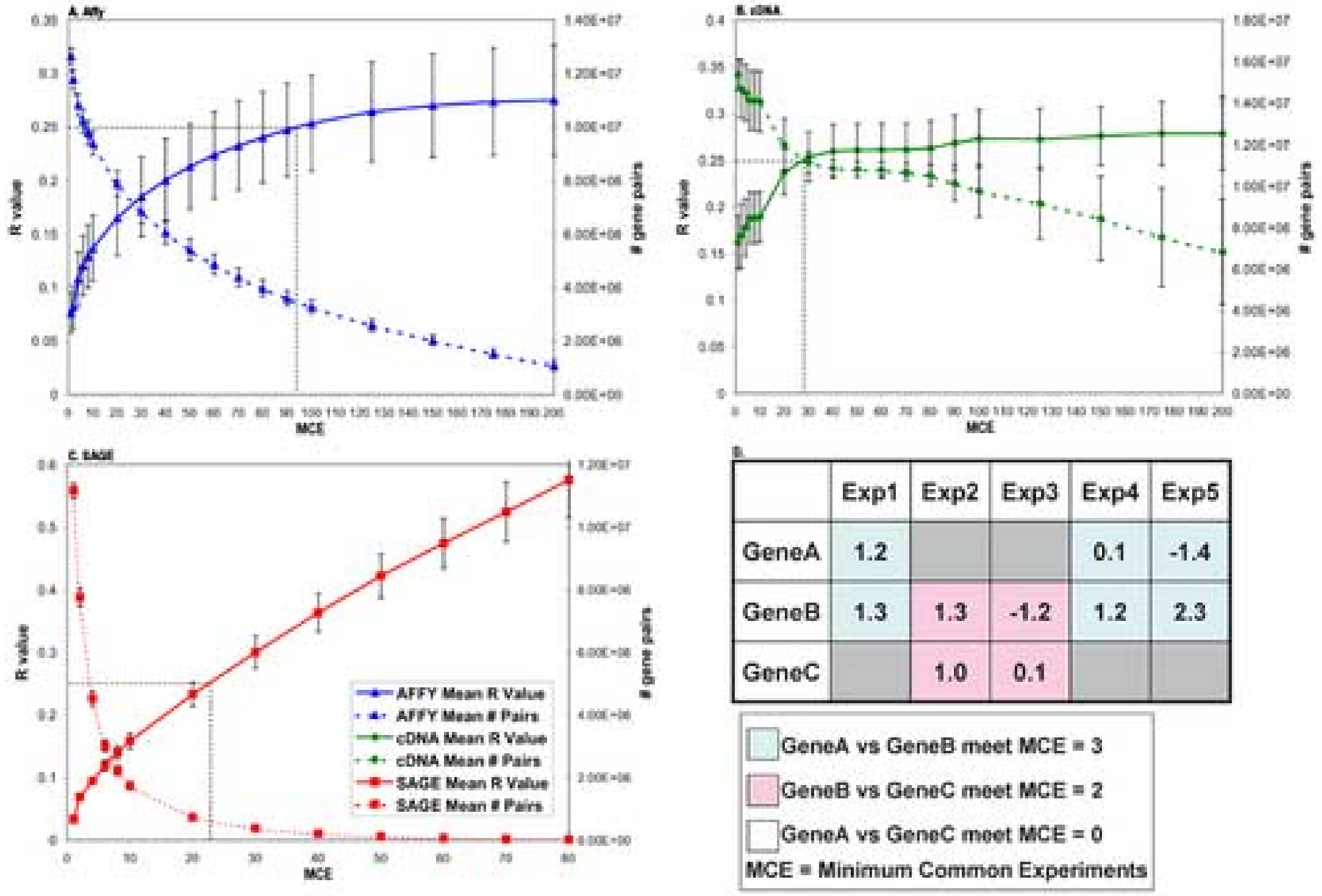
Instead, the graphs illustrate that each method is capable of identifying biologically relevant gene pair relationships and the Venn diagram indicates that they identify very different sets of relationships. Furthermore, the GO analysis provides a means of choosing reasonable score-thresholds for each method to generate lists of high-confidence coexpressed genes.

Table 1

Platform	**Division**	**MCE cutoff**	**Gene pairs**	**r_c value**
Affymetrix	Random	100	4,149,092	0.925
	Pseudo-random by GSE series	95	3,427,174	0.257
		100	3,260,557	0.253
cDNA Microarray	Random	100	10,429,219	0.889
	Pseudo-random by author	28	11,178,346	0.253
		100	9,747,169	0.273
SAGE	Random	100	2,635	0.776
	Pseudo-random by tissue	23	577,820	0.253
		100	1,518	0.660

Figure 1

Click here to download high resolution image



	Exp1	Exp2	Exp3	Exp4	Exp5
GeneA	1.2			0.1	-1.4
GeneB	1.3	1.3	-1.2	1.2	2.3
GeneC		1.0	0.1		

🟦 GeneA vs GeneB meet MCE = 3

🟪 GeneB vs GeneC meet MCE = 2

□ GeneA vs GeneC meet MCE = 0

MCE = Minimum Common Experiments

Figure 2

Click here to download high resolution image

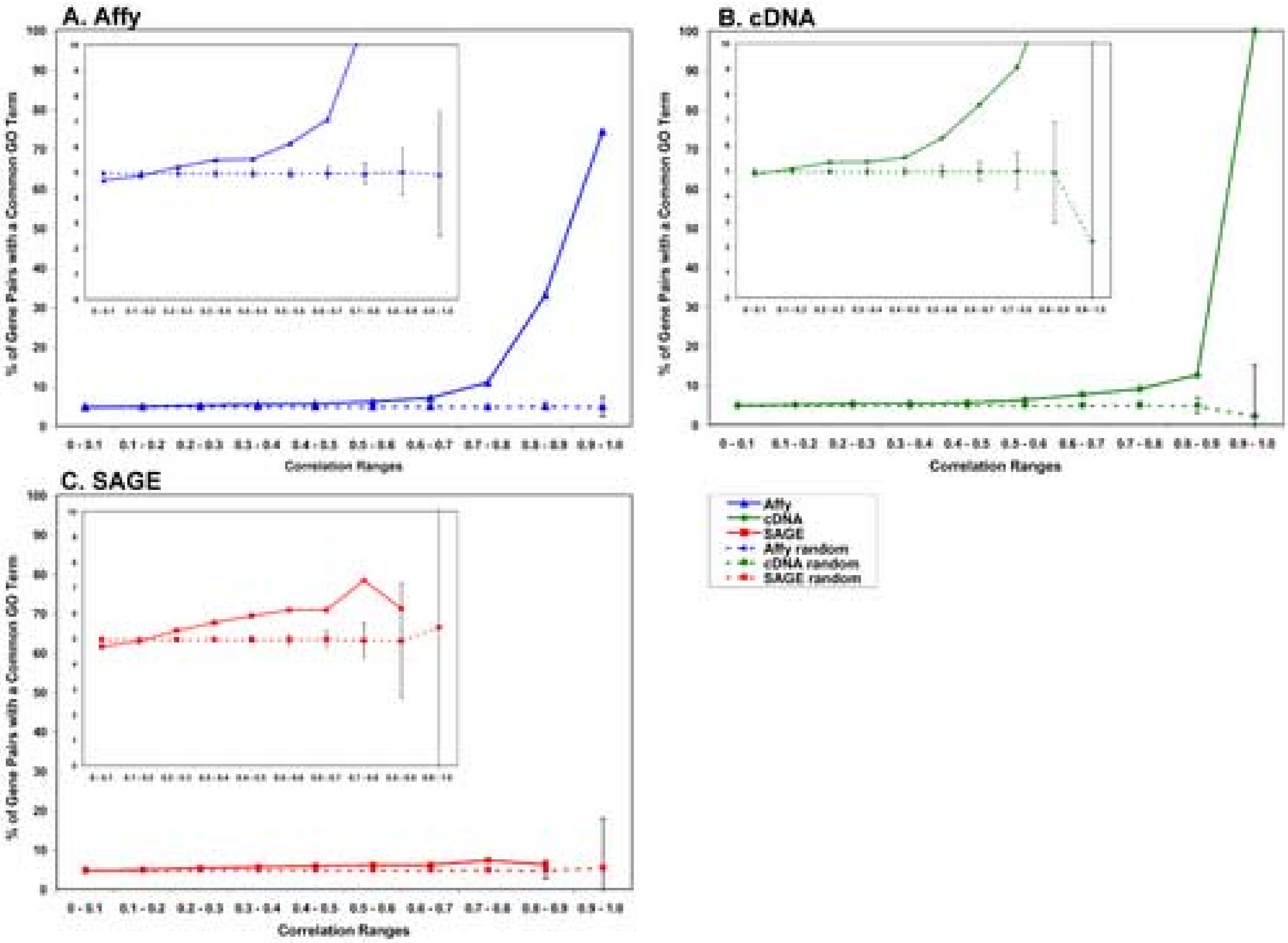


Figure 3

Click here to download high resolution image

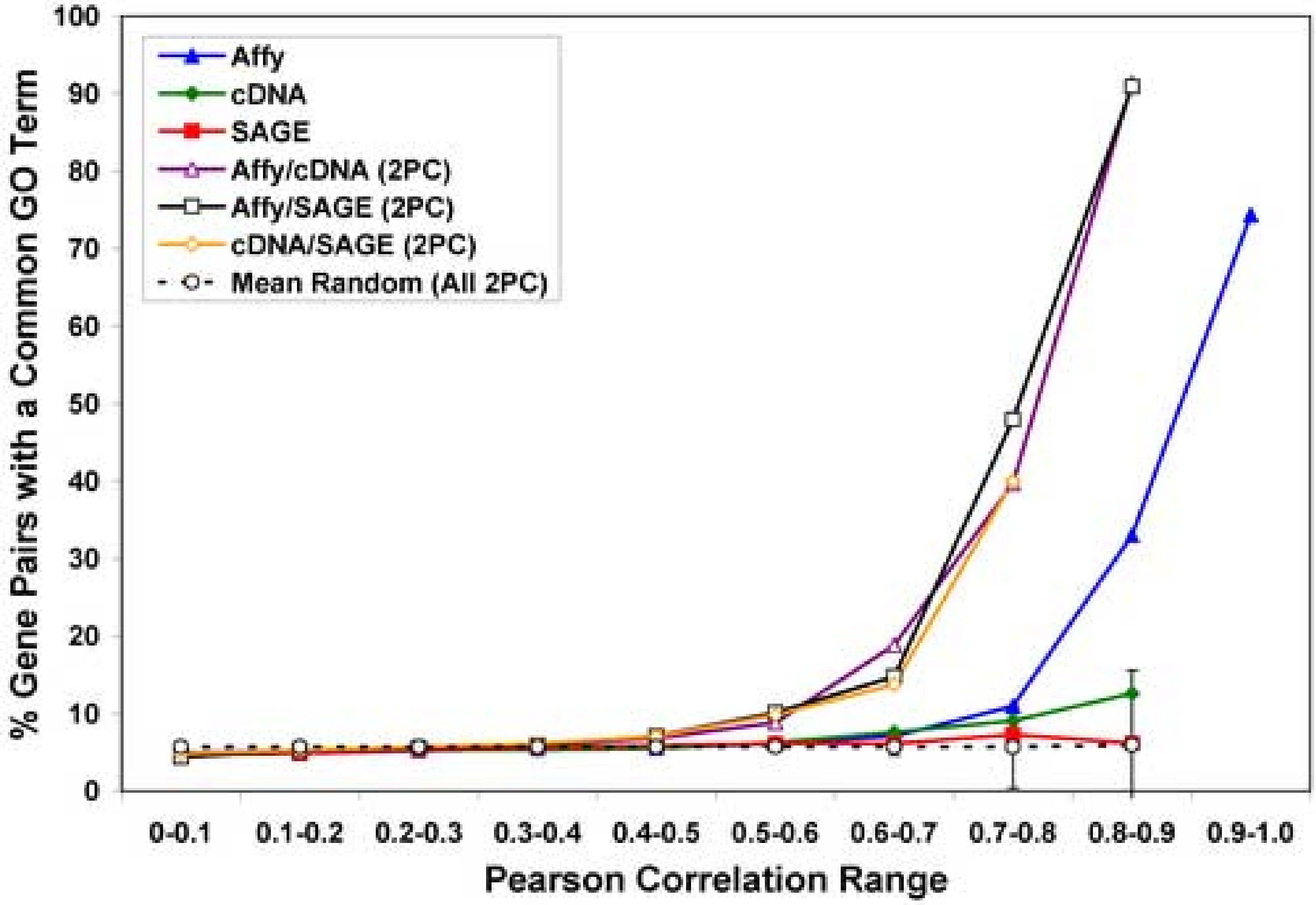
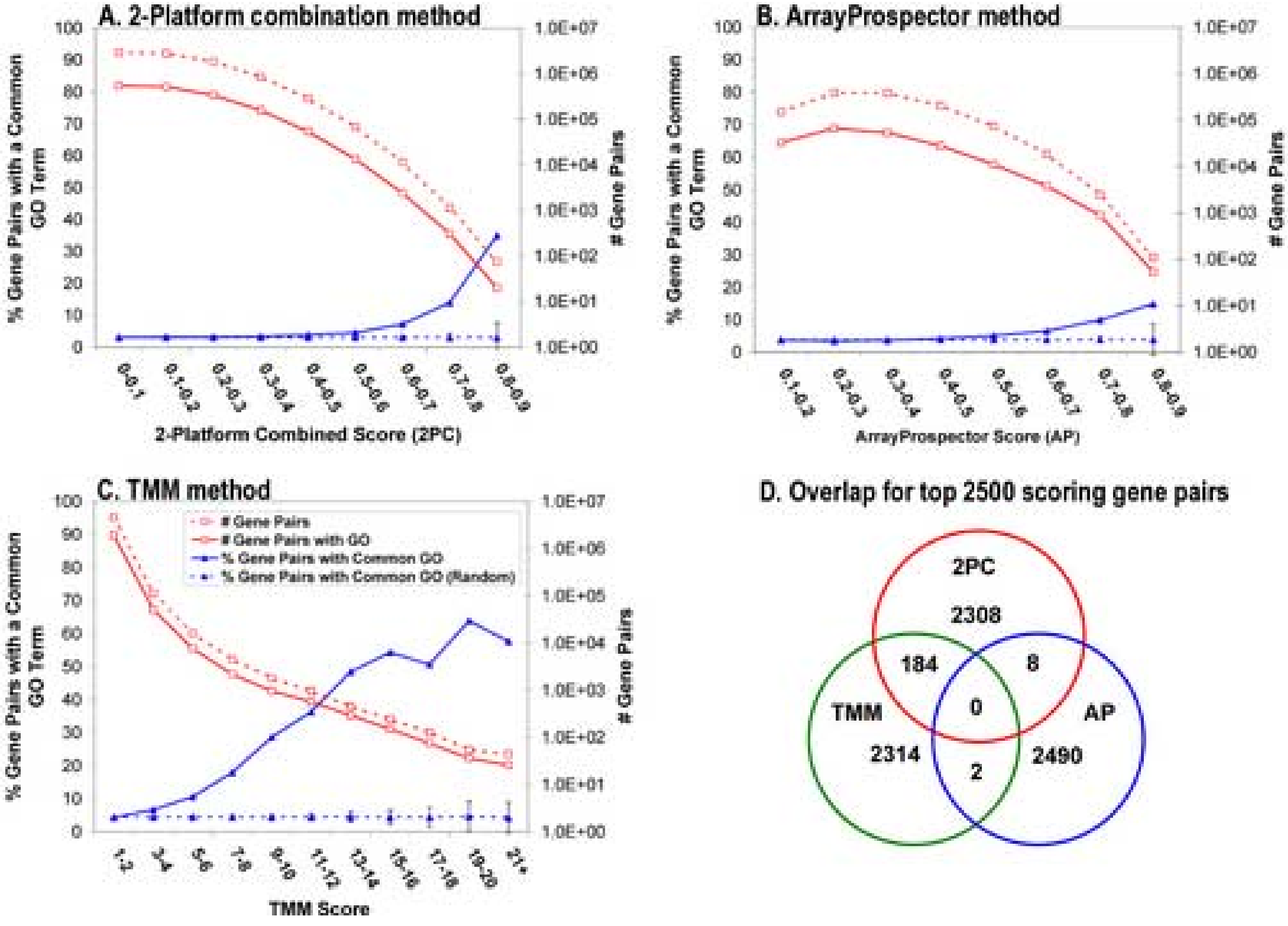


Figure 4

Click here to download high resolution image



Supplementary Materials

This file and additional supplementary data can be found at: http://www.bcgsc.ca/gc/bomge/coexpression/suppl_materials

Supplementary Results

Cancer sample analysis

Cancer samples were found to represent a substantial fraction in the cDNA (~29%), Affymetrix (~40% of the complete 889 samples) and SAGE (~61%) datasets. Cancer tissues are often characterized by changes in gene expression and thus could act as a confounding factor when trying to identify co-expressed genes. To investigate this issue the SAGE dataset was divided into cancer and normal subsets and consistency between these measured. The comparison of normal and cancer SAGE libraries resulted in a correlation of 0.324 for an MCE of 23 and 0.707 for an MCE of 80 (MCE of 100 could not be used because the normal tissue subset only contained 94 samples). These results are comparable to that seen for consistencies of SAGE when not taking cancer status into account (Suppl. Fig. 3). Thus, we cautiously concluded that the presence of cancer libraries was not seriously affecting the SAGE co-expression analysis and proceeded to subsequent analyses without removing the cancer libraries.

Ranked Match Analysis

The ranked match analysis shows that different expression platforms can identify the same co-expressed genes (Suppl. Fig. 5). It may be that for gene A, SAGE experiments identify its most similar gene (in terms of expression patterns) to be gene B with a Pearson correlation of 0.9. The cDNA microarray data might also identify gene B as the closest gene to A but with a Pearson value of 0.78. Thus, a comparison of Pearson ranks may be a more useful method for evaluating cross platform consistency than actual Pearson values. The Affymetrix/cDNA comparison found that 26.5% of genes have a co-expressed gene of Pearson rank 10 or better confirmed by both platforms compared to 18.9% for random data. Affymetrix versus SAGE agreed for 26.4% of genes compared to 18.9% for random, and cDNA versus SAGE for 21.8% compared to 18.8% for random. The high percentages of genes in agreement for random data are the result of our MCE criteria. Each gene pair must have at least 95, 28 or 23 MCE (for Affymetrix, cDNA and SAGE respectively). Some genes will have close to this number of experiments and thus realize the required MCE for only a few gene pair comparisons. Since we only consider gene pairs that are common in all three datasets, there will be some genes that only have a little more than 10 gene pairs. In these cases, a shared match within a rank of 10 for the two platforms will occur quite commonly by chance. Thus, it is the difference over random, rather than the actual percentage, that indicates a significant number of shared matches. In all three comparisons, the percentage of shared matches observed was significantly greater than that observed between randomized datasets ($p<0.001$, 1000 randomizations). We can conclude that the platform comparisons do identify more of the same co-expressed genes than expected by chance. However, in general the platforms show poor agreement.

Supplementary Methods

Cancer Sample Analysis

The proportion of cancer samples was determined from the literature for the cDNA dataset [6] and from GEO sample records for Affymetrix and SAGE. SAGE, having the highest percentage of cancer samples, was used for the analysis. The SAGE data set was manually divided into 94 normal and 148 cancer libraries based on sample descriptions from the GEO sample records. The consistency between these two subsets of the data was calculated as described above and compared to the other data sets.

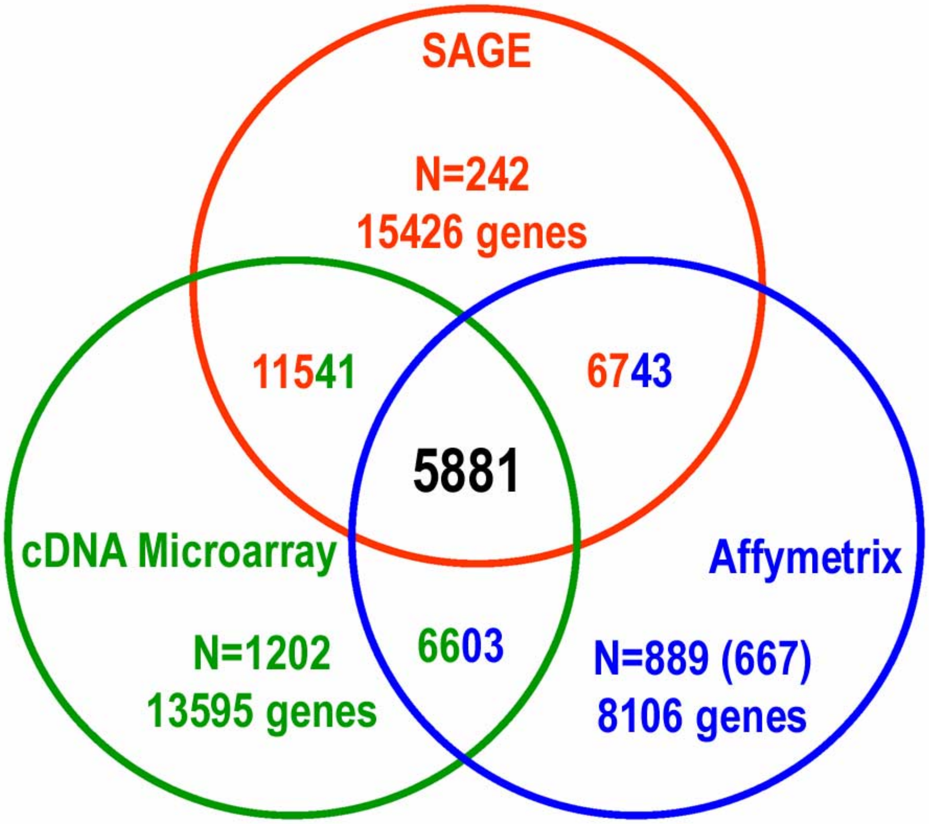
Ranked Match Analysis

In addition to considering the actual Pearson correlation between each gene pair and comparing between platforms, the correlation rank was considered. This analysis identifies shared co-expressed genes, or matches, between platforms. For instance, a shared match would be illustrated by the following: Gene A's 2nd most similar gene is gene B in the Affymetrix data. This is gene A's 3rd most similar gene in the SAGE data. This example would count as one shared 'match' for a neighborhood of k = 3 for the Affymetrix versus SAGE comparison. A Perl script was written to determine each gene's closest k neighbors from one dataset and compare to another dataset. Numbers of shared neighbors within each neighborhood size (k) were tallied and graphed. 1000 randomizations were conducted for each platform comparison to determine how often the level of agreement at each neighborhood would be observed by chance.

Supplementary Figures

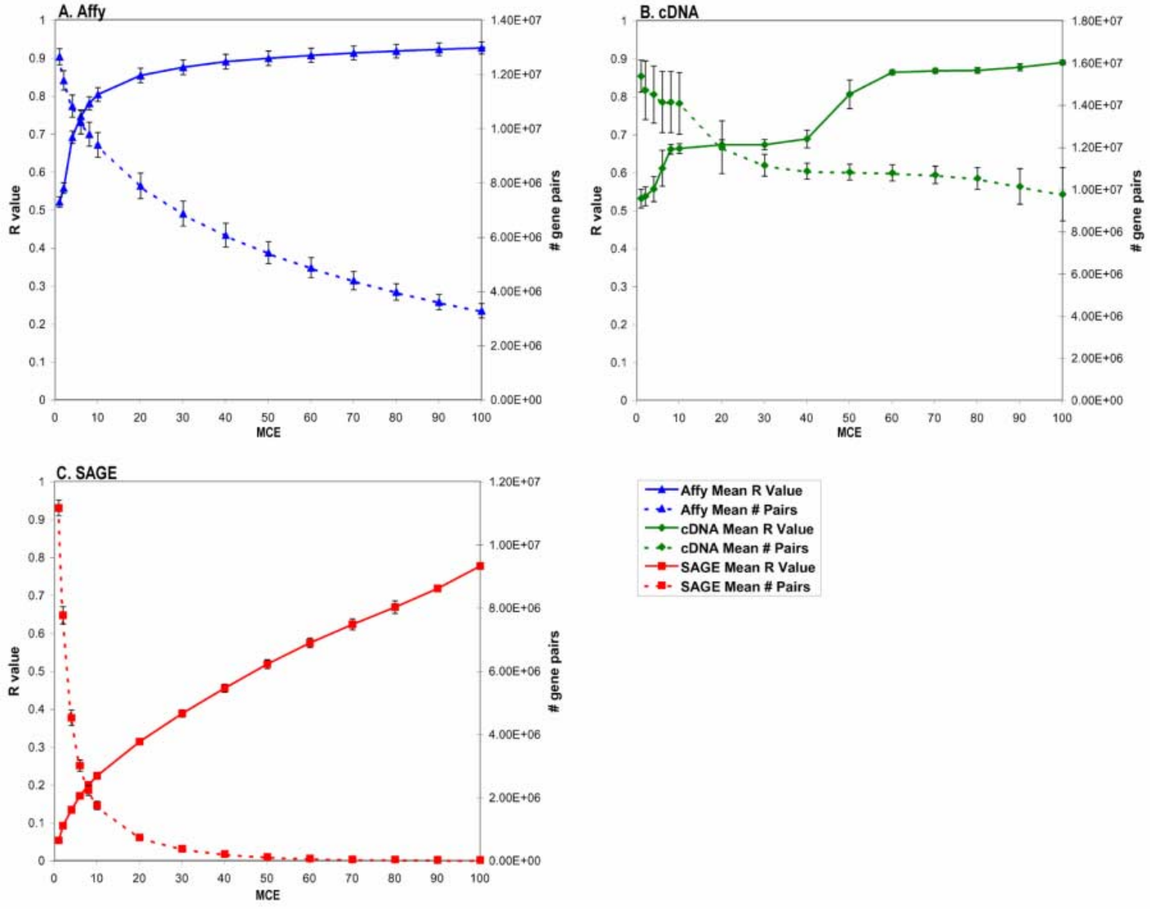
Suppl. Figure 1. Venn Diagram outlining datasets used in analysis.

N indicates the number of experiments available for the platform. For Affymetrix, the number in brackets indicates the subset of experiments providing detection (PMA) calls. The number of genes represents only those genes that could be unambiguously mapped to a LocusLink ID.



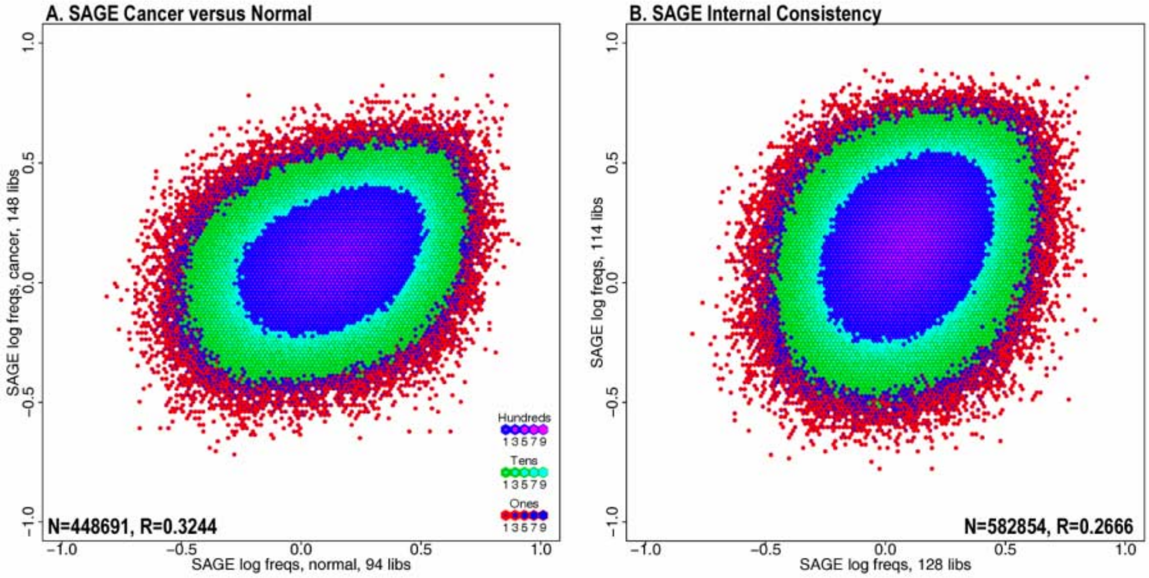
Suppl. Figure 2. Internal consistency analysis based on random division of experiments.

Analysis is identical to Figure 1, except division of libraries is random rather than by experiment, author, or tissue, resulting in much higher r_c values due to presence of replicates or very similar experiments. Data represent mean r_c value and gene pair number of 100 random divisions. Error bars indicate one standard deviation.



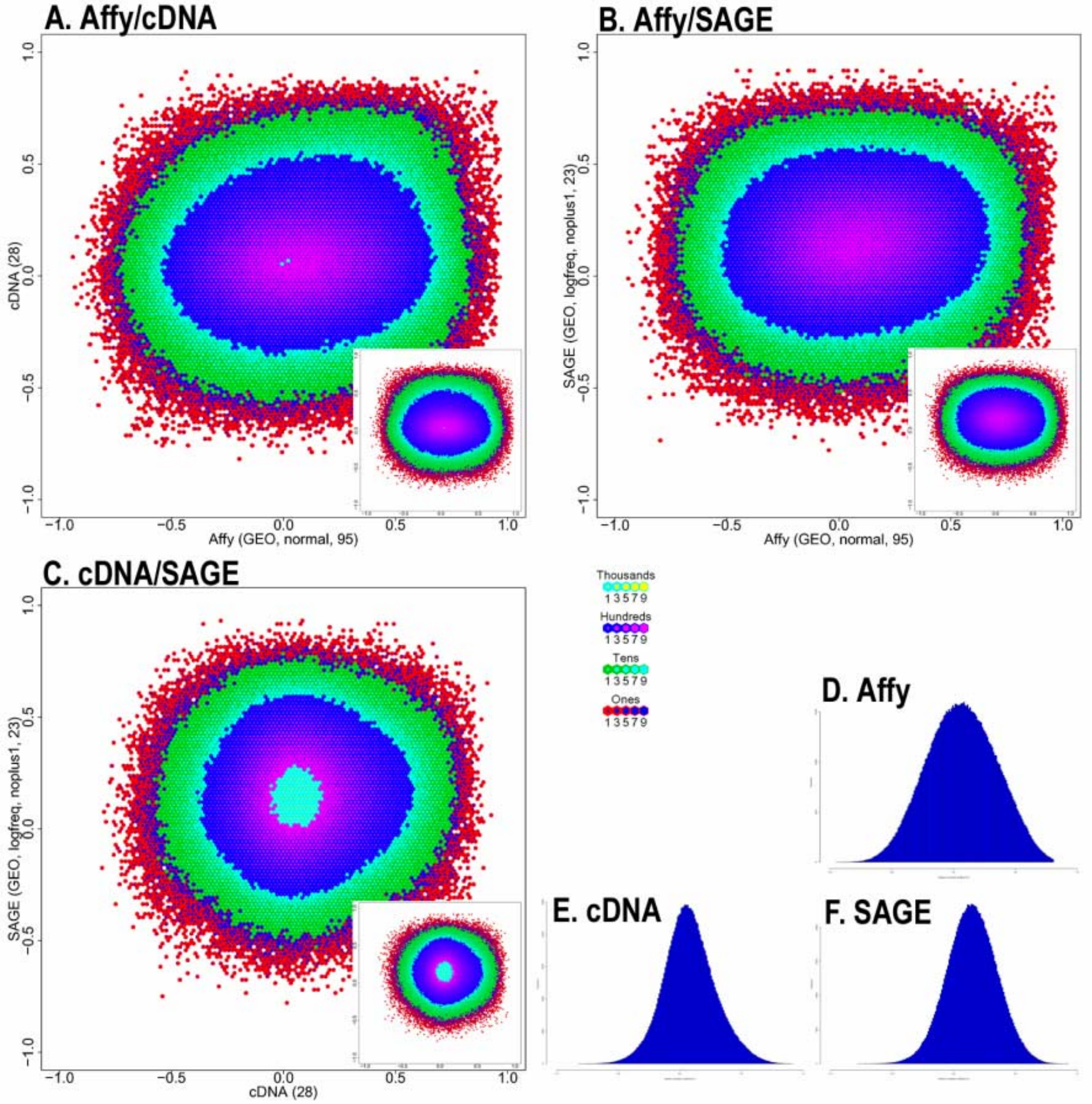
Suppl. Figure 3. SAGE cancer versus normal analysis.

Plots represent correlation of correlations for subsets of SAGE data. (A) Correlation between normal and cancer SAGE libraries, r_c=0.324 for 23 MCE; (B) Correlation between randomly divided subsets of SAGE data, r_c=0.267 for MCE of 23.



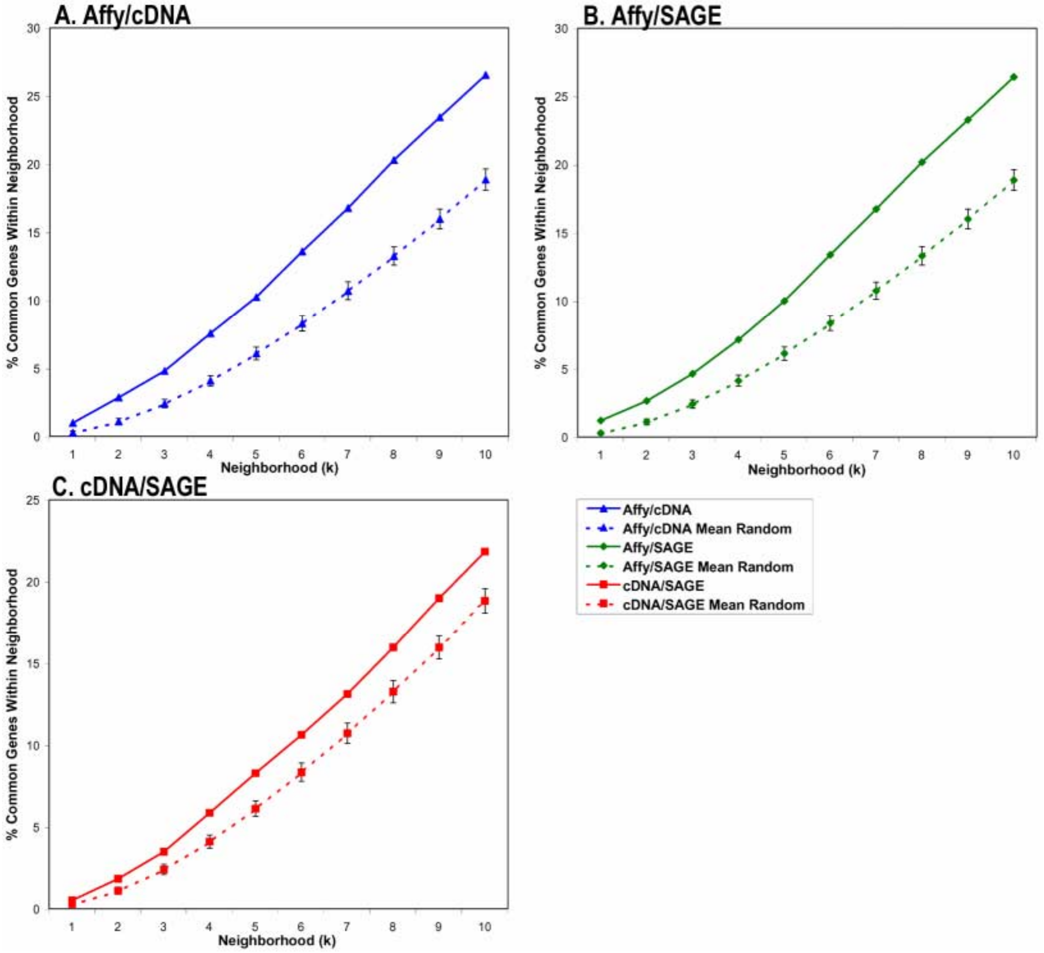
Suppl. Figure 4. Platform Comparisons.

Plots represent correlation of correlations (r_c) between each pairwise platform comparison. A. Affymetrix versus cDNA, r_c=0.102; B. Affymetrix versus SAGE, r_c=0.086; C. cDNA versus SAGE, r_c=0.041. 1,173,330 gene pairs are shown representing the intersection between Affymetrix, cDNA, and SAGE for which 95, 28, and 23 MCE were required respectively for each Pearson correlation calculation. Correlations observed in A-C were significant when compared to randomized data (p<0.001, 1000 randomizations). Small inset boxes show representative randomized data; D-E. Pearson correlation (r) frequency distributions for each platform. Notice that each displays a similar, approximately normal distribution with a slight skew towards positive correlations.

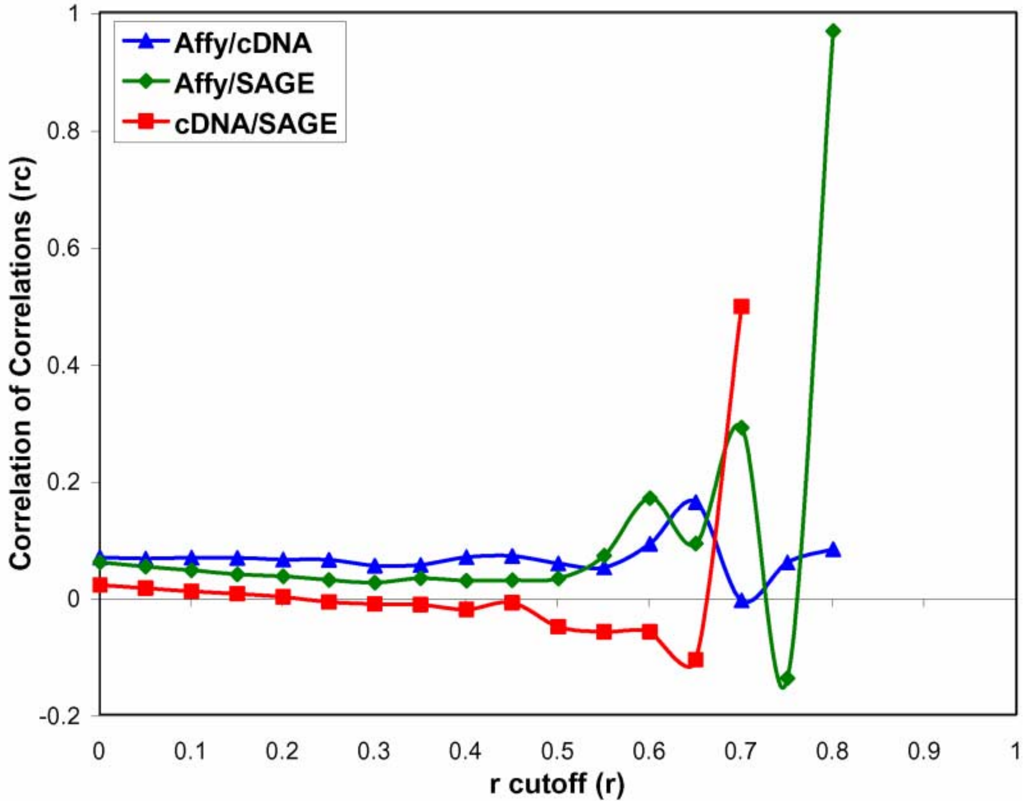


Suppl. Figure 5. Ranked Pearson Analysis.

Percentage of genes with a co-expressed gene identified by both platforms within a rank or neighborhood of k for each platform comparison. Random lines represent mean values from 1000 randomizations. Error bars indicate one standard deviation.

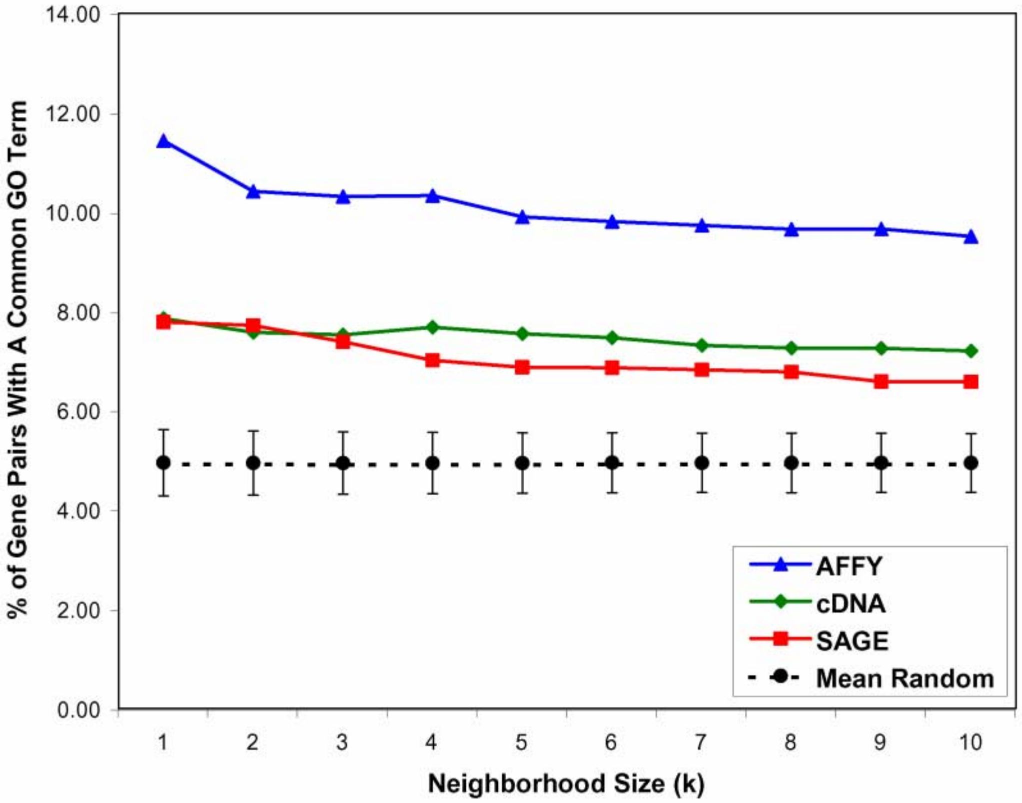


Suppl. Figure 6. Effect of correlation cutoff on r_c.
Platform comparisons (Suppl. Fig. 4) were repeated with subsets of gene pairs with correlations above cutoffs (0.1 increments). Only positive correlations were considered. Higher global concordance was observed for the Affymetrix/cDNA comparison at a Pearson cutoff (r-cutoff) of 0.65 and for the Affymetrix/SAGE comparison at r-cutoff of 0.6 and 0.7 ($p<0.05$). The cDNA/SAGE comparison did not show any increase that was significant. In any case, the steady trend of increasing r_c with more stringent r-cutoff was not observed as reported elsewhere (Lee *et al.*, 2003). Asterisks indicate increased r_c values which were also found to be significant ($p<0.05$).

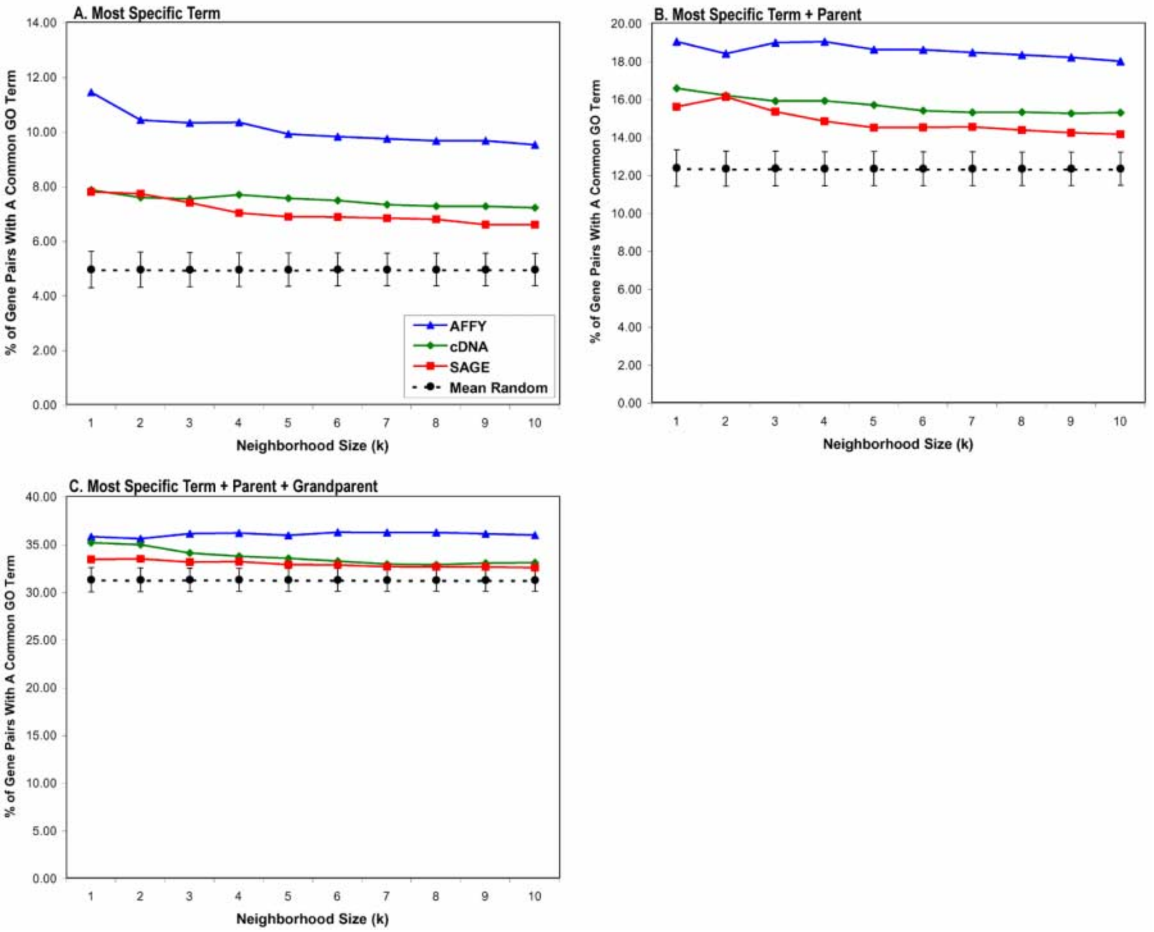


Suppl. Figure 7. GO Analysis.

Gene pairs for which both genes were annotated with Gene Ontology Biological Process terms were evaluated to determine the percentage of pairs within a neighborhood of k that are annotated with the same GO term. As the GO annotation is hierarchical, only the most specific GO terms for each gene were considered. Comparison of these percentages to results produced from randomizing gene pair correlations indicate that gene pairs found to be correlated by any platform are more likely to share the same function than randomly chosen gene pairs (p<0.001, 1000 randomizations). Affymetrix appears to predict the most biologically relevant gene pair correlations.



Suppl. Figure 8. Expanded Go Analysis including hierarchical relationships. Analysis performed as for Suppl. Figure 7, but in addition to considering only most specific GO term annotations (A), the percentage of gene pairs sharing parent terms (B) or parent and grandparent terms (C) were also determined. As higher levels in the GO hierarchical tree (parent and grandparent terms) are considered, there is a higher chance that randomly chosen gene pairs will share GO terms, resulting in less difference between random and actual data.

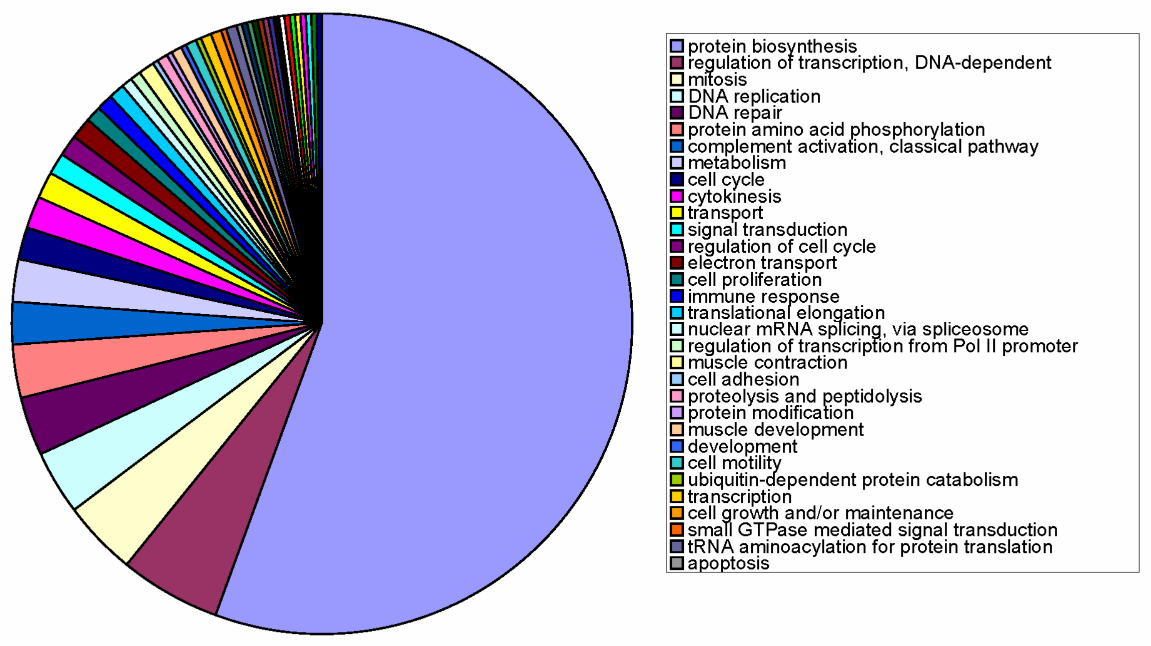


A. Most Specific Term

B. Most Specific Term + Parent

C. Most Specific Term + Parent + Grandparent

Suppl. Figure 9. GO categories for gene pairs confirmed by multiple datasets. The chart shows GO terms of gene pairs with an average Pearson correlation of r>0.6 for any two of three platform datasets (Affymetrix, cDNA microarray, SAGE). The legend only shows the 32 categories with more than one gene pair. However, another 20 categories are represented on the chart and are summarized in Suppl. Table 1.



Suppl. Table 1.

Gene Pairs	**Percent**	**Common Term**	**Go Term**
257	55.508	GO:0006412	protein biosynthesis
25	5.3996	GO:0006355	regulation of transcription, DNA-dependent
18	3.8877	GO:0007067	mitosis
15	3.2397	GO:0006260	DNA replication
14	3.0238	GO:0006281	DNA repair
13	2.8078	GO:0006468	protein amino acid phosphorylation
11	2.3758	GO:0006958	complement activation, classical pathway
10	2.1598	GO:0008152	metabolism
8	1.7279	GO:0007049	cell cycle
7	1.5119	GO:0000910	cytokinesis
6	1.2959	GO:0006810	transport
6	1.2959	GO:0007165	signal transduction
5	1.0799	GO:0000074	regulation of cell cycle
5	1.0799	GO:0006118	electron transport
4	0.8639	GO:0008283	cell proliferation
4	0.8639	GO:0006955	immune response
3	0.6479	GO:0006414	translational elongation
3	0.6479	GO:0000398	nuclear mRNA splicing, via spliceosome
3	0.6479	GO:0006357	regulation of transcription from Pol II promoter
3	0.6479	GO:0006936	muscle contraction
2	0.432	GO:0007155	cell adhesion
2	0.432	GO:0006508	proteolysis and peptidolysis
2	0.432	GO:0006464	protein modification
2	0.432	GO:0007517	muscle development
2	0.432	GO:0007275	development
2	0.432	GO:0006928	cell motility
2	0.432	GO:0006511	ubiquitin-dependent protein catabolism
2	0.432	GO:0006350	transcription
2	0.432	GO:0008151	cell growth and/or maintenance
2	0.432	GO:0007264	small GTPase mediated signal transduction
2	0.432	GO:0006418	tRNA aminoacylation for protein translation
2	0.432	GO:0006915	apoptosis
1	0.216	GO:0015031	protein transport
1	0.216	GO:0006461	protein complex assembly
1	0.216	GO:0008380	RNA splicing
1	0.216	GO:0006364	rRNA processing
1	0.216	GO:0006366	transcription from Pol II promoter
1	0.216	GO:0007242	intracellular signaling cascade
1	0.216	GO:0006091	energy pathways
1	0.216	GO:0006635	fatty acid beta-oxidation
1	0.216	GO:0006177	GMP biosynthesis
1	0.216	GO:0006096	glycolysis
1	0.216	GO:0006259	DNA metabolism
1	0.216	GO:0007186	G-protein coupled receptor protein signaling pathway
1	0.216	GO:0007267	cell-cell signaling
1	0.216	GO:0006098	pentose-phosphate shunt
1	0.216	GO:0006917	induction of apoptosis
1	0.216	GO:0016575	histone deacetylation
1	0.216	GO:0006954	inflammatory response
1	0.216	GO:0045786	negative regulation of cell cycle
1	0.216	GO:0009966	regulation of signal transduction
463	**100**	**52**	