Title Page

Large-Scale Comparison of Publicly Available SAGE, cDNA Microarray, and Oligonucleotide Microarray Expression Data for Global Co-Expression Analyses

Authors

Obi L. Griffith[1], Erin D. Pleasance[1], Debra L. Fulton[2], Mehrdad Oveisi[1], Martin Ester[3], Asim Siddiqui[1] and Steven J.M. Jones[1]

Affiliations

1. Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, British Columbia, Canada, V5Z 4E6
2. Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada, V5A 1S6
3. 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
Email: sjones@bcgsc.ca

Keywords:

Co-expression, microarrays, SAGE, spotted cDNA, Affymetrix oligonucleotide, Gene Ontology, gene expression platforms.

Abstract

Background: Large amounts of gene expression data from several different platforms are being made available to the scientific community and increasingly used as tools for validation and integration of other studies. Several studies have compared two or three platforms to evaluate the consistency of expression profiles for a single tissue or sample series but few have determined if these translate into reliable gene co-expression patterns across many conditions.

Results: We have analyzed *Homo sapiens* data from 1202 cDNA microarray experiments, 242 SAGE libraries and 667 Affymetrix oligonucleotide microarray experiments. Using standard co-expression analysis methods, we have assessed each platform for internal consistency, performed inter-platform comparisons, and tested each platform's predictions against the Gene Ontology. An overall correlation of correlations (r_c) analysis showed that the platforms agree significantly better than random ($p<0.001$, 1000 randomizations) but with very low correlations of $r_c < 0.102$. A rank analysis also showed significant but poor agreement with only 3-8% better performance than randomized data. Comparison against the Gene Ontology (GO) revealed that all three platforms identify more co-expressed gene pairs with common biological processes than random data and as the Pearson correlation for a gene pair increased it was more likely to be confirmed by GO.

Conclusions: The three datasets compared demonstrate significant but low levels of global concordance. When evaluated for biological relevance, the Affymetrix dataset performed best with gene pairs of correlation 0.9-1.0 confirmed by GO in 74% of cases. However, our results suggest that all three datasets may provide some biologically relevant predictions of co-expression. Researchers are cautioned against using any one dataset exclusively for their analyses.

Background

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 co-expressed. Genome wide co-expression analyses in *C. elegans* have been used with some success to identify gene function or genes that are co-regulated [4]. This "guilt-by-association" approach has received criticism because of high levels of noise and other problems inherent to the methods [5] but still holds great interest for biologists. Additionally, co-expression data are increasingly used for validation and integration with other 'omic' data sources such as sequence conservation [6], yeast two-hybrid interactions [7, 8], RNA interference [9] and regulatory element predictions [10] to name only a few.

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 [11-13]. 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) [14] 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 [13, 15-21]. Other studies have reported 'poor' correlations [20, 22-25]. The correlations reported above were for expression levels or expression changes of individual genes, not co-expression of gene pairs. One study has examined the correlation of co-expression results from multiple platforms [26]. 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. Finally, 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). If matched samples can display reasonable levels of consistency between expression profiles generated by different platforms the question remains as to how effectively unmatched samples from many different sources can be combined for co-expression analyses. If two genes are co-regulated (i.e. controlled by an identical set of transcription factors) they should display similar expression patterns across many conditions and be identified as co-expressed. This is the basic premise of many gene function and regulation studies. If true, large datasets from different expression platforms should identify the same co-expressed gene pairs even if derived from different conditions and tissues.

To explore this, we have compared three large publicly available datasets for SAGE, cDNA microarray (cDNA), and Affymetrix oligonucleotide microarray (Affymetrix) (Figure 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. Pearson was chosen as a similarity metric because it is one of the most commonly used, with numerous published examples for Affymetrix [11, 27, 28], cDNA [4, 7, 29] and SAGE [30, 31]. Our motivation for this study was to explore the fecundity of large extant expression datasets to identify co-regulated genes and their utility as a resource for biological study.

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 co-expression. 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. 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 2D).

Increasing the MCE increased the internal consistency but decreased the number of gene pairs considered for both the random (Suppl. Figure 1) and pseudo-random (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 for the study of the cDNA microarray data [6] (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 2). 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).

Internal consistency is a measure of the reproducibility or robustness of gene co-expression predictions. This is based on the assumption that if a gene pair is truly co-expressed based on an expression dataset, it should be predicted as co-expressed by two completely different 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. 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 2). All internal consistency correlations are summarized in Table 1.

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. 2). 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.

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 the platforms compared poorly against each other (Figure 3). All comparisons were found to have significant but poor positive correlations when compared to randomized data ($p<0.001$, 1000 randomizations). 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.

An analysis of correlation at different minimum Pearson cutoffs (r-cutoff) for gene pairs was performed as described previously [26] (Suppl. Fig. 3). 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.

Ranked Match Analysis

The ranked match analysis shows that different expression platforms can identify the same co-expressed genes (Figure 4). 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.

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 [32] was used to evaluate gene co-expression predictions for each platform. We hypothesized that genes which are co-expressed 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 (Figure 5). In general, the platforms all perform better than expected by chance. Affymetrix performed best, followed by cDNA microarray and SAGE which performed about equally better than random data. The analysis was extended up the GO hierarchy to 'most specific + parent term' and 'most specific + parent + grandparent terms' (Suppl. Fig. 4). The exact same trends and relationships were observed for the extended GO subspaces. As the parent and grandparent terms are included the random lines get closer to the platform lines, because the GO space expands to include increasingly general terms and there is a greater likelihood that any two genes will share the same GO term by chance.

A final GO 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 6). 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 randomized data (p<0.001, 1000 randomizations). The improvement over random data was very slight for the cDNA and SAGE datasets (2-4%). For the Affymetrix data however, the trend was striking. Gene pairs identified as co-expressed 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 co-expressed 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).

Discussion

It is important to bear in mind that the evaluation presented here may reflect the characteristics of the datasets more than the technologies used to generate the data. Intrinsic characteristics of the platforms will of course have an effect on the nature and quality of the data. But, to truly evaluate the three platforms for their relative ability to identify co-expressed genes we would have to perform the same experiments across many different tissues and conditions using each platform. Given the costs of these techniques this evaluation is not currently practical. Our intention was instead to evaluate the largest currently existing, publicly available datasets for these platforms to better understand their utility for integration with and validation of other data.

We have shown that the genes identified as co-expressed are highly dependent on the dataset used. We report measures of internal consistency ranging from 0.253 to 0.925 for Affymetrix oligonucleotide arrays, 0.254 to 0.889 for cDNA microarrays and 0.267 to 0.776 for SAGE depending on how the data is divided and the minimum number of common experiments (MCE) required for each gene-to-gene Pearson correlation calculation. 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 0.27 whereas the SAGE dataset continued to improve to above 0.6 as increasing MCE between genes was required. 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.

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). Comparison using Pearson rank instead of actual Pearson value confirmed the poor correlations, although the comparisons do agree significantly more often than random data. The fact that none of these data sets agree well raises some serious questions. There are several possible explanations to consider: (1) The data comprising these datasets are so noisy as to prevent reliable identification of many truly co-expressed genes; (2) Only one of the datasets is accurate and the others inaccurate; (3) The method of identifying co-expressed genes is inadequate; (4) The unmatched and non-overlapping nature of the samples that make up each dataset result in identification of different subsets of truly co-expressed genes; (5) The vast majority of genes are under such complex regulatory control that genes co-regulated in one cell-type or tissue behave in an entirely different manner in others.

The issue of noise is undoubtedly responsible for at least some of the non-concordance between datasets. Technical and biological sources of noise are always present to some degree and their effects are difficult to assess when using public data, especially for genes expressed at low levels. A related issue is that the platforms utilize intrinsically different methods and are prone to different biases. For SAGE, tag counts are in theory proportional to the actual amount of RNA transcript in the sample [3]. For Affymetrix and cDNA, each transcript may be measured with different sensitivity and saturation depending on the hybridization characteristics of the sequence in question. Cross-hybridization is also an issue for both microarray methods but may affect each differently depending on oligonucleotide design and cDNA selection. Oligonucleotide probes or cDNA clones can be annotated to the same gene but represent very different sequences. A recent study compared gene expression ratios and difference calls for Agilent cDNA and Affymetrix oligonucleotide microarrays [28]. The authors found that cross-platform consistency was significantly better for sequence-matched probes than gene-identifier matched probes (the standard method of comparison used in this study). SAGE does not have the problem of cross-hybridization but is prone to others such as PCR bias and the lack of tag-specificity inherent to the 14mer SAGE protocol [33, 34].

Thus, even if levels of noise are kept low, intrinsic biases and design differences in each method are likely to contribute significantly to non-concordance.

The fact that intra-platform comparisons show some correlation and improve with number of data points seems to indicate that at least some gene pairs identified are truly co-expressed. Furthermore, the GO analysis found that gene pairs identified as co-expressed are more likely to share the same biological process. Thus, we believe that at least a portion of the genes identified as co-expressed are real and can be validated biologically. The GO analysis did not conclusively identify a single 'correct' platform/dataset but it did show that the Affymetrix dataset identified more biologically relevant gene pairs than the cDNA or SAGE datasets.

The question of the best method of analysis is a difficult and contentious one. The Pearson correlation coefficient was chosen as a similarity metric simply because it is one of the most commonly used methods in the literature [35]. A recent study of SAGE mouse retina data found that Poisson-based distances are more appropriate and reliable for SAGE data than commonly used metrics such as Pearson or Euclidian [36]. A plethora of similar studies for cDNA and oligonucleotide microarray have been published outlining the strengths and weaknesses of different metrics, normalization methods, and so on, but a consensus has not been reached [37]. For now, the Pearson correlation seems a reasonable method of identifying co-expressed genes, as it focuses on patterns of changes in expression rather than absolute expression levels. Alternate methods could be attempted and evaluated for biological relevance using the GO analysis. We did not find strong evidence that gene pairs with higher Pearson correlations (r) show greater global concordance (r_c) between platforms as in the NCI-60 study [26]. However, we did find some evidence that gene pairs with higher Pearson values are more likely to be biologically relevant (according to GO), especially in the case of the Affymetrix dataset. This gives us some confidence in the Pearson correlation as a metric for expression analysis. In any case, our purpose in this study was simply to compare the results of a global co-expression analysis from different platforms using standard methods. A comparison of platforms using a wide variety of different similarity metrics would make an interesting follow-up study.

Given our results, we believe the final two explanations are most likely, wherein each of the platforms is correct to some degree but identifies different sets of co-expressed genes because most genes are under complex condition- or tissue-specific regulatory control. Many studies using each of these platforms have demonstrated real potential to dissect the biological activities of the cell. But, because each dataset is comprised of different experiment series and tissues, each identifies different subsets of co-expressed genes. Indeed, it may be that many genes perform different functions at different stages and are under multiple regulatory systems. Thus, gene A could be co-expressed with B under some set of conditions in one tissue and co-expressed with gene C in another set. If this is the case, a sub-space analysis, that looks for gene pairs highly correlated in subsets of tissues or conditions may be most effective. Shortly after the completion of our analysis, a study along these lines was published [38]. The authors examined 60 large microarray datasets (cDNA and Affymetrix oligonucleotide) for gene pairs identified as co-expressed in multiple datasets. Of the 9.7 million different co-expression pairs that passed their selection criteria, only 2.2% are seen in 3 or more data sets. Knowing this, it is perhaps not surprising that a comparison between pairs of large amalgamated datasets reveal poor overall correlations. The authors find 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. Thus, a gene pair co-expressed under only a few conditions in the dataset could be biologically relevant, but could easily be 'drowned out' by the 'noise' of samples from conditions in which the genes do not act together. However, many of the links confirmed by only a few datasets are also likely false positives. Genes found to be co-expressed across many conditions, as in our global analysis, are more likely to be truly co-regulated. A recent study of yeast cDNA microarray experiments found that the ability to correctly identify co-regulated genes depends strongly on the number of microarray experiments in the data set. But, even for large numbers of experiments (all available) the true-positive rate was only 28% (defined as genes in same cluster sharing at least one known transcription factor) [39]. Thus, the resolving power of co-expression studies should continue to improve as public datasets grow in size.

Conclusions

The three datasets compared demonstrate significant but low levels of global concordance. When evaluated for biological relevance, the Affymetrix dataset performed best with gene pairs of correlation 0.9-1.0 confirmed by GO in 74% of cases. However, our results suggest that all three datasets may provide some biologically relevant predictions of co-expression. The selection of co-expressed gene pairs for validation and integration of other data sets will likely be dictated by the goals of the study in question and the confidence the researcher desires. In any case, researchers should be cautioned against using any one of these data sets as representative of gene co-expression as each dataset may be telling different parts of the whole story. For now, a combination of the most reliable results from each method might be the best option. To this end, we are providing a co-expression database from which the most significant gene pairs from each dataset (including those published elsewhere) will be made available (http://www.bcgsc.ca/gc/bomge/coexpression/). In general, co-expression identified by larger sets of experiments will be most reliable (with more than 100 experiments preferable). SAGE in particular will benefit from the increasing amount of publicly available data, as currently its primary weakness is a lack of samples. The other platforms will benefit from a wider range of experimental conditions. Co-expressed gene pairs identified by more than one platform and/or sharing functional annotations may be of biological interest. Further analysis of these genes, using orthology and motif finding algorithms, can attempt to identify common transcription factor binding sites that may regulate the expression of these co-expression networks.

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 [6] 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 222 did not have PMA calls (detection calls)) representing 8,106 genes, also from GEO (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 (developed in house, Zuyderduyn et al., unpubl.) 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 [31]:

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 [26] 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 co-expression observed within each platform, we divided the experiments available and determined co-expression for each subset independently. The results were then compared by calculating a correlation of the gene correlations. If a platform consistently finds co-expressed genes regardless of the exact experiments involved, the correlation will be close to 1. To determine whether the observed correlation is significant, we repeat the procedure with randomized gene expression values, expecting a correlation 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 co-expression 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 co-expressed genes. If the two subsets of experiments contain replicates, they are more likely to identify the same co-expressed 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 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. In an effort to produce an unbiased comparison of the three platforms, an MCE was chosen for each such that the same internal consistency would result ($r=0.25$) (Figure 2). 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.

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.

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 [26].

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.

Gene Ontology Analysis

The Gene Ontology (GO) is a controlled vocabulary that describes the roles of genes and proteins in all organisms [32]. GO is composed of three independent ontologies: biological process, molecular function, and cellular component. The biological process ontology describes the biological objectives to which the gene or gene product contributes. The molecular function ontology describes the biochemical activities of a gene product. The cellular component ontology describes the locations where the genes can be active. 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. For simplicity sake, 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 co-expressed gene pair could be found to share a common GO term that was annotated for both genes by a co-expression analysis. Thus, co-expression data could be confirming co-expression 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 Figure 5 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 6 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. 1000 randomizations 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/.

List of abbreviations:

SAGE, Serial Analysis of Gene Expression; GEO, Gene Expression Omnibus; GO, Gene Ontology; IEA, Inferred Electronic Annotation; MGC, Mammalian Gene Collection; NCE, number of common experiments; r, Pearson correlation; r_c, Correlation of 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. 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.

References

1. **Schena M, Shalon D, Davis RW, Brown PO: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995, 270(5235):467-470.**
2. **Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H *et al*: Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996, 14(13):1675-1680.**
3. **Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: Serial analysis of gene expression. *Science* 1995, 270(5235):484-487.**
4. **Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS: A gene expression map for Caenorhabditis elegans. *Science* 2001, 293(5537):2087-2092.**
5. **Quackenbush J: Genomics. Microarrays--guilt by association. *Science* 2003, 302(5643):240-241.**
6. **Stuart JM, Segal E, Koller D, Kim SK: A gene-coexpression network for global discovery of conserved genetic modules. *Science* 2003, 302(5643):249-255.**
7. **Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T *et al*: A map of the interactome network of the metazoan C. elegans. *Science* 2004, 303(5657):540-543.**
8. **Kemmeren P, van Berkum NL, Vilo J, Bijma T, Donders R, Brazma A, Holstege FC: Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol Cell* 2002, 9(5):1133-1143.**
9. **Walhout AJ, Reboul J, Shtanko O, Bertin N, Vaglio P, Ge H, Lee H, Doucette-Stamm L, Gunsalus KC, Schetter AJ *et al*: Integrating interactome, phenome, and transcriptome mapping data for the C. elegans germline. *Curr Biol* 2002, 12(22):1952-1958.**
10. **Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 2003, 34(2):166-176.**
11. **Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A *et al*: Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* 2002, 1(2):133-143.**
12. **Nimgaonkar A, Sanoudou D, Butte AJ, Haslett JN, Kunkel LM, Beggs AH, Kohane IS: Reproducibility of gene expression across generations of Affymetrix microarrays. *BMC Bioinformatics* 2003, 4(1):27.**
13. **Tan PK, Downey TJ, Spitznagel EL, Jr., Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, Cam MC: Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* 2003, 31(19):5676-5684.**
14. **Blackshaw S, Kuo WP, Park PJ, Tsujikawa M, Gunnersen JM, Scott HS, Boon WM, Tan SS, Cepko CL: MicroSAGE is highly representative and**

reproducible but reveals major differences in gene expression among samples obtained from similar tissues. *Genome Biol* 2003, 4(3):R17.

15. **Huminiecki L, Lloyd AT, Wolfe KH: Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. *BMC Genomics* 2003, 4(1):31.**
16. **Detours V, Dumont JE, Bersini H, Maenhaut C: Integration and cross-validation of high-throughput gene expression data: comparing heterogeneous data sets. *FEBS Lett* 2003, 546(1):98-102.**
17. **Jarvinen AK, Hautaniemi S, Edgren H, Auvinen P, Saarela J, Kallioniemi OP, Monni O: Are data from different gene expression microarray platforms comparable? *Genomics* 2004, 83(6):1164-1168.**
18. **Iacobuzio-Donahue CA, Ashfaq R, Maitra A, Adsay NV, Shen-Ong GL, Berg K, Hollingsworth MA, Cameron JL, Yeo CJ, Kern SE *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* 2003, 63(24):8614-8622.**
19. **Kim HL: Comparison of oligonucleotide-microarray and serial analysis of gene expression (SAGE) in transcript profiling analysis of megakaryocytes derived from CD34+ cells. *Exp Mol Med* 2003, 35(5):460-466.**
20. **Rogojina AT, Orr WE, Song BK, Geisert EE, Jr.: Comparing the use of Affymetrix to spotted oligonucleotide microarrays using two retinal pigment epithelium cell lines. *Mol Vis* 2003, 9:482-496.**
21. **Ishii M, Hashimoto S, Tsutsumi S, Wada Y, Matsushima K, Kodama T, Aburatani H: Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics* 2000, 68(2):136-143.**
22. **Evans SJ, Datson NA, Kabbaj M, Thompson RC, Vreugdenhil E, De Kloet ER, Watson SJ, Akil H: Evaluation of Affymetrix Gene Chip sensitivity in rat hippocampal tissue using SAGE analysis. Serial Analysis of Gene Expression. *Eur J Neurosci* 2002, 16(3):409-413.**
23. **Li J, Pankratz M, Johnson JA: Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays. *Toxicol Sci* 2002, 69(2):383-390.**
24. **Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS: Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* 2002, 18(3):405-412.**
25. **Mah N, Thelin A, Lu T, Nikolaus S, Kuhbacher T, Gurbuz Y, Eickhoff H, Kloppel G, Lehrach H, Mellgard B *et al*: A comparison of oligonucleotide and cDNA-based microarray systems. *Physiol Genomics* 2004, 16(3):361-370.**
26. **Lee JK, Bussey KJ, Gwadry FG, Reinhold W, Riddick G, Pelletier SL, Nishizuka S, Szakacs G, Annereau JP, Shankavaram U *et al*: Comparing cDNA and oligonucleotide array data: concordance of gene expression across platforms for the NCI-60 cancer cells. *Genome Biol* 2003, 4(12):R82.**
27. **Williams EJ, Bowles DJ: Coexpression of neighboring genes in the genome of Arabidopsis thaliana. *Genome Res* 2004, 14(6):1060-1067.**
28. **Mecham BH, Klus GT, Strovel J, Augustus M, Byrne D, Bozso P, Wetmore DZ, Mariani TJ, Kohane IS, Szallasi Z: Sequence-matched probes produce increased cross-platform consistency and more reproducible biological**

results in microarray-based gene expression measurements. *Nucleic Acids Res* 2004, 32(9):e74.

29. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M *et al*: Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* 2000, 24(3):227-235.
30. Nacht M, Dracheva T, Gao Y, Fujii T, Chen Y, Player A, Akmaev V, Cook B, Dufault M, Zhang M *et al*: Molecular characteristics of non-small cell lung cancer. *Proc Natl Acad Sci U S A* 2001, 98(26):15203-15208.
31. Porter DA, Krop IE, Nasser S, Sgroi D, Kaelin CM, Marks JR, Riggins G, Polyak K: A SAGE (serial analysis of gene expression) view of breast tumor progression. *Cancer Res* 2001, 61(15):5697-5702.
32. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000, 25(1):25-29.
33. Spinella DG, Bernardino AK, Redding AC, Koutz P, Wei Y, Pratt EK, Myers KK, Chappell G, Gerken S, McConnell SJ: Tandem arrayed ligation of expressed sequence tags (TALEST): a new method for generating global gene expression profiles. *Nucleic Acids Res* 1999, 27(18):e22.
34. Unneberg P, Wennborg A, Larsson M: Transcript identification by analysis of short sequence tags--influence of tag length, restriction site and transcript database. *Nucleic Acids Res* 2003, 31(8):2217-2226.
35. Quackenbush J: Computational analysis of microarray data. *Nat Rev Genet* 2001, 2(6):418-427.
36. Cai L, Huang H, Blackshaw S, Liu JS, Cepko C, Wong WH: Clustering analysis of SAGE data using a Poisson approach. *Genome Biol* 2004, 5(7):R51.
37. Butte A: The use and analysis of microarray data. *Nat Rev Drug Discov* 2002, 1(12):951-960.
38. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P: Coexpression Analysis of Human Genes Across Many Microarray Data Sets. *Genome Res* 2004, 14(6):1085-1094.
39. Yeung K, Medvedovic M, Bumgarner R: From co-expression to co-regulation: how many microarray experiments do we need? *Genome Biology* 2004, 5(7):R48.
40. Pruitt KD, Katz KS, Sicotte H, Maglott DR: Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends in Genetics* 2000, 16(1):44-47.
41. Mammalian Gene Collection Program Team*, Strausberg RL, Feingold EA, Grouse LH, Derge JG, Klausner RD, Collins FS, Wagner L, Shenmen CM, Schuler GD *et al*: Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *PNAS* 2002, 99(26):16899-16903.
42. Davidson GS, Wylie BN, Boyack KW: Cluster Stability and the Use of Noise in Interpretation of Clustering. In*: 2001*: IEEE Computer Society; 2001: 23.

43. De Hoon MJ, Imoto S, Nolan J, Miyano S: Open source clustering software. *Bioinformatics* 2004.

44. Ihaka R, Gentleman R: R: A language for data analysis and graphics. In: *Journal of Computational & Graphical Statistics.* vol. 5: American Statistical Association; 1996: 299.

Figure Legends

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.

Figure 2. 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 3. 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.

Figure 4. 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.

Figure 5. 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.

Figure 6. GO Correlation Range Analysis.

Comparison to data based on randomized correlations shows that a smaller number of gene pairs with very low correlations (<0.2) share GO terms. At higher correlations (in particular, $r > 0.8$) gene pairs are more likely to have similar function, 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. Random lines represent mean values from 1000 randomizations. Error bars indicate one standard deviation.

Supplementary Figure Legends

Suppl. Figure 1. Internal consistency analysis based on random division of experiments.

Analysis is identical to Figure 2, 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 2. 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 3. Effect of correlation cutoff on r_c.

Platform comparisons (Figure 3) 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 [26]. Asterisks indicate increased r_c values which were also found to be significant ($p<0.05$).

Suppl. Figure 4. Expanded Go Analysis including hierarchical relationships.

Analysis performed as for Figure 5, 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.

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.

Platform	**Division**	**MCE cutoff**	**Gene pairs**	**r_c value**
Affymetrix	Random	100	4,149,092	0.925
	By GSE series	95	3,427,174	0.257
		100	3,260,557	0.253
cDNA Microarray	Random	100	10,429,219	0.889
	By author	28	11,178,346	0.253
		100	9,747,169	0.273
SAGE	Random	100	2,635	0.776
	By tissue	23	577,820	0.253
		100	1,518	0.660
	Cancer/Normal	10	1,631,419	0.204
		23	448,691	0.324
		80	1,253	0.707

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

SUPPLEMENTARY MATERIALS

All necessary data will be provided on a supplementary materials webpage hosted by the GSC and also through a gene co-expression resource webpage soon to be developed (http://www.bcgsc.ca/gc/bomge/coexpression/). Gene pairs found to be co-expressed above a reasonable threshold of confidence are being added to a co-expression database (containing data from several species in addition to *H. sapiens*) as part of the Sockeye project. Sockeye users will ultimately be able to pull up co-expressed genes for any gene they are interested in and use these to help identify upstream regulatory elements.

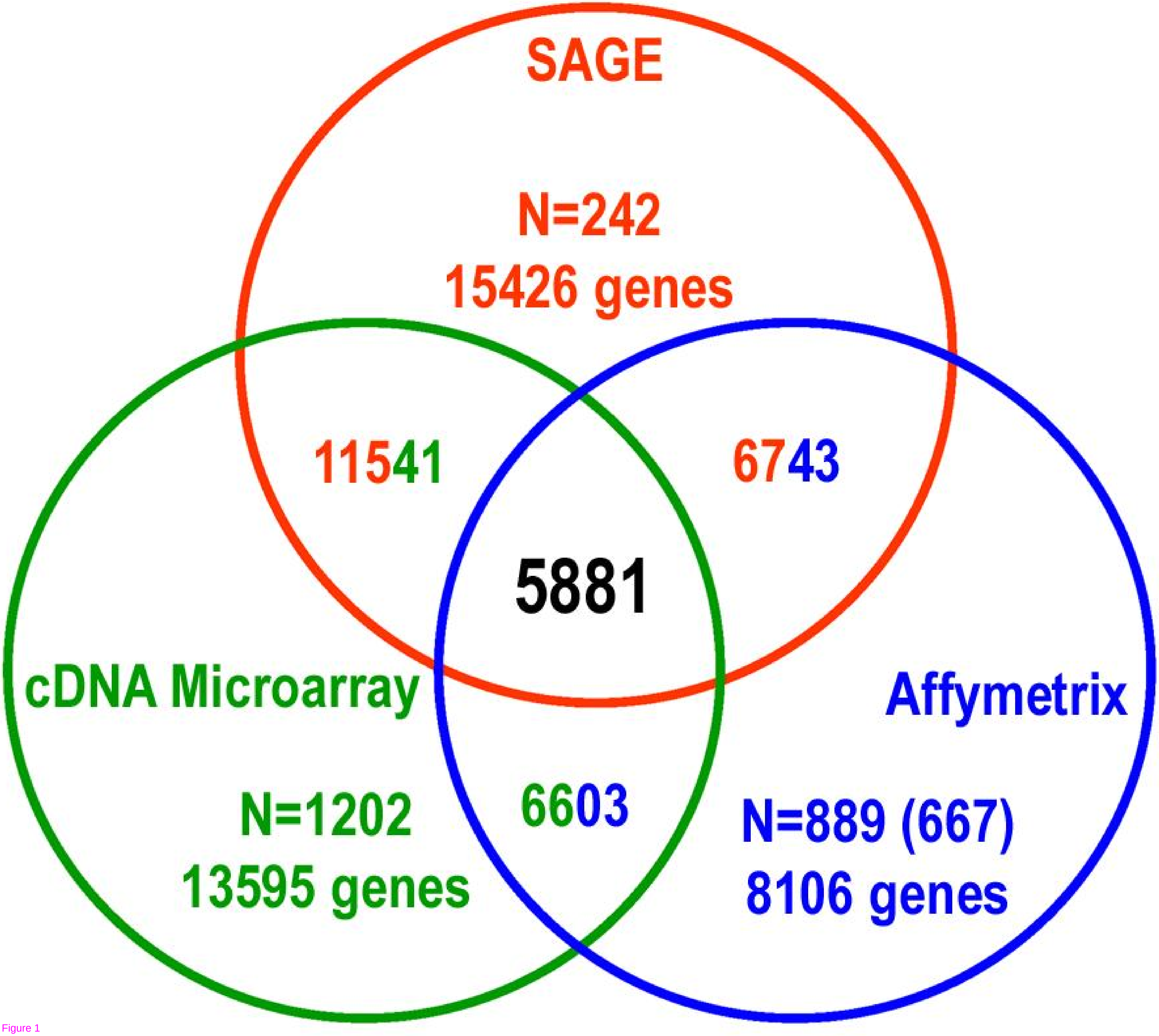
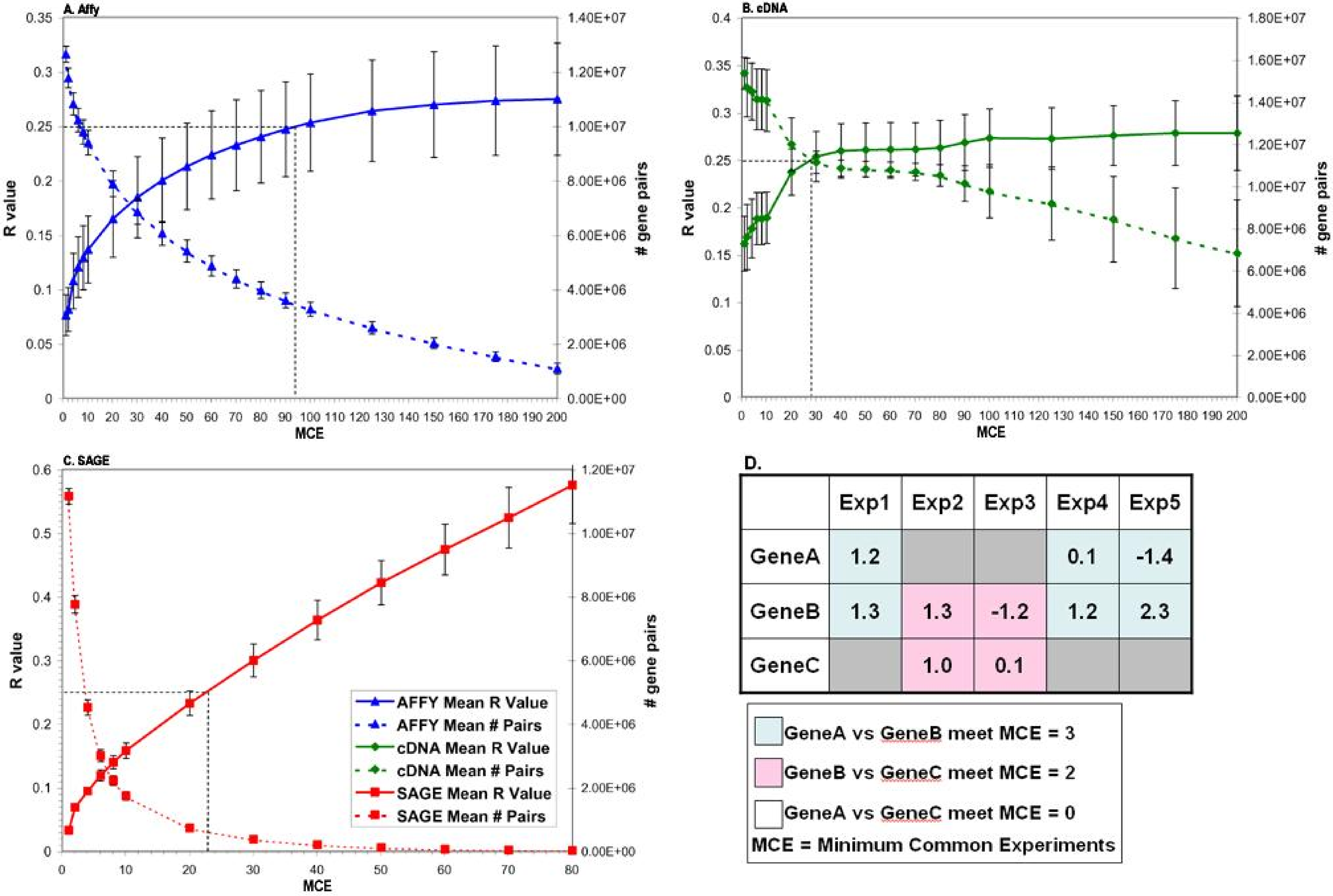

SAGE
N=242
15426 genes
11541
6743
5881
cDNA Microarray
Affymetrix
N=1202
13595 genes
6603
N=889 (667)
8106 genes

Figure 1



	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

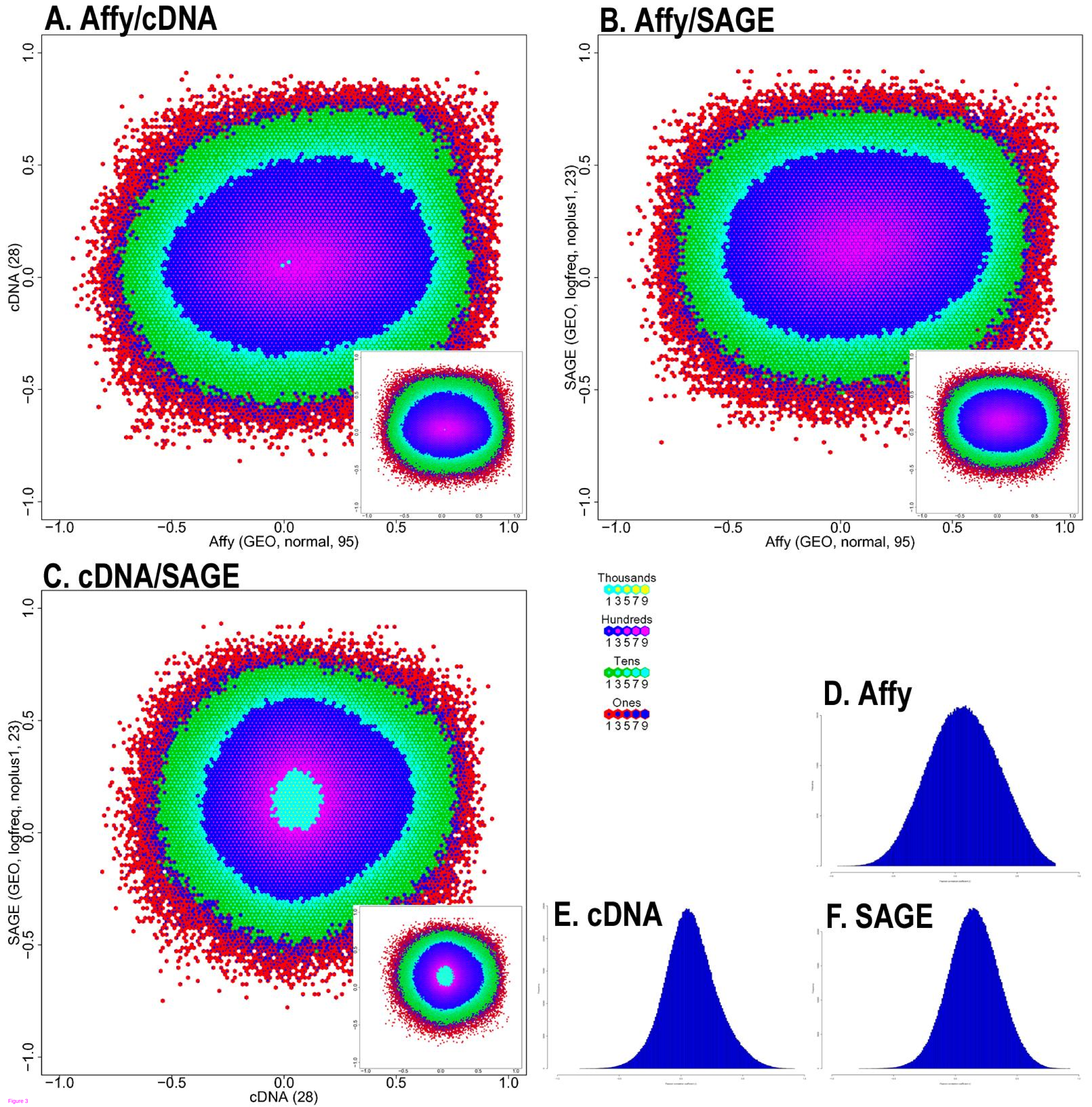

A. Affy/cDNA
cDNA (28)
Affy (GEO, normal, 95)
B. Affy/SAGE
SAGE (GEO, logfreq, noplus1, 23)
Affy (GEO, normal, 95)
C. cDNA/SAGE
SAGE (GEO, logfreq, noplus1, 23)
cDNA (28)
−1.0
−0.5
0.0
0.5
1.0
Thousands
1 3 5 7 9
Hundreds
1 3 5 7 9
Tens
1 3 5 7 9
Ones
1 3 5 7 9
D. Affy
E. cDNA
F. SAGE

Figure 3

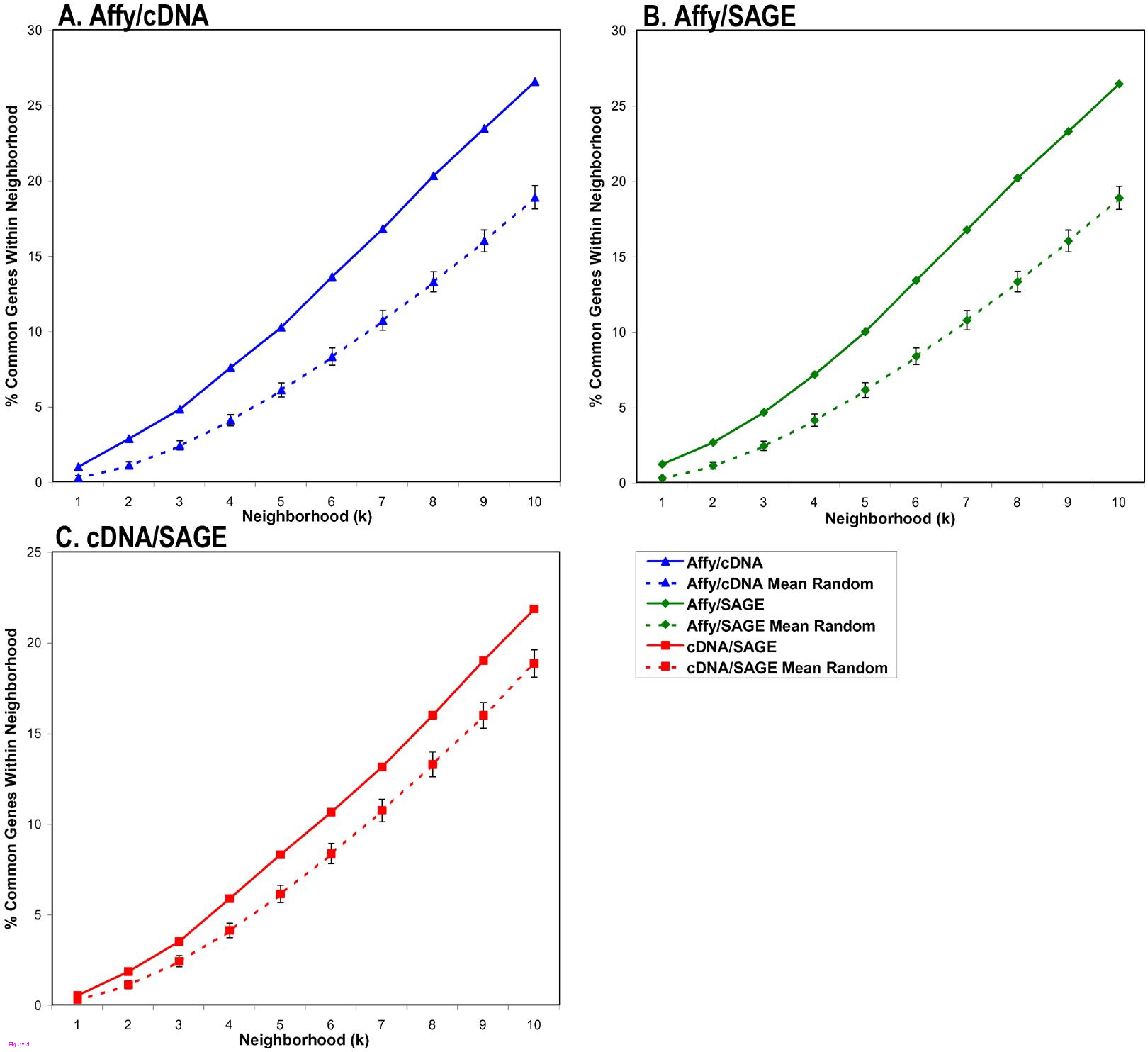

A. Affy/cDNA
B. Affy/SAGE
C. cDNA/SAGE
% Common Genes Within Neighborhood
Neighborhood (k)
0
5
10
15
20
25
30
1
2
3
4
5
6
7
8
9
10
Affy/cDNA
Affy/cDNA Mean Random
Affy/SAGE
Affy/SAGE Mean Random
cDNA/SAGE
cDNA/SAGE Mean Random

Figure 4

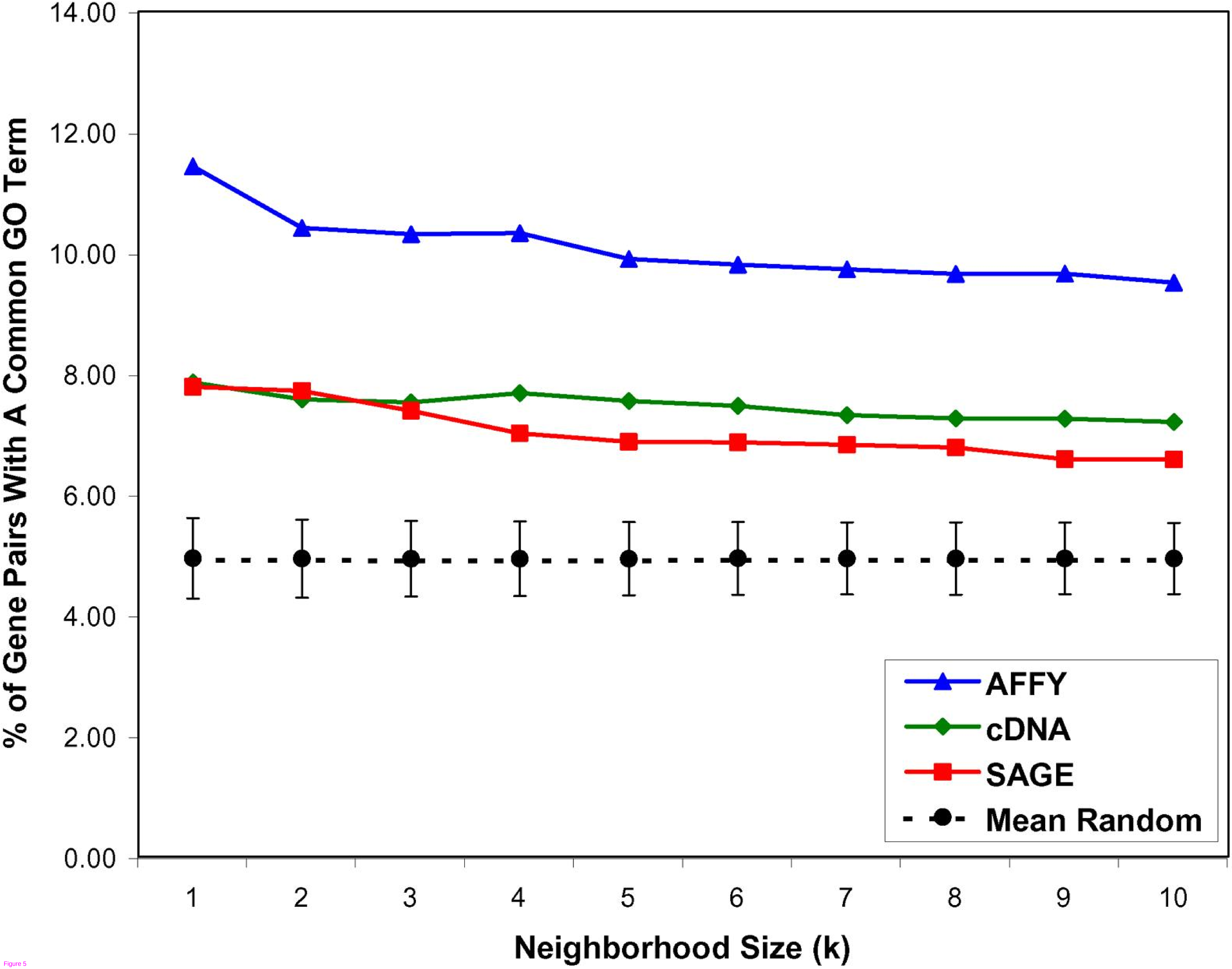

14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
% of Gene Pairs With A Common GO Term
1
2
3
4
5
6
7
8
9
10
Neighborhood Size (k)
AFFY
cDNA
SAGE
Mean Random

Figure 5

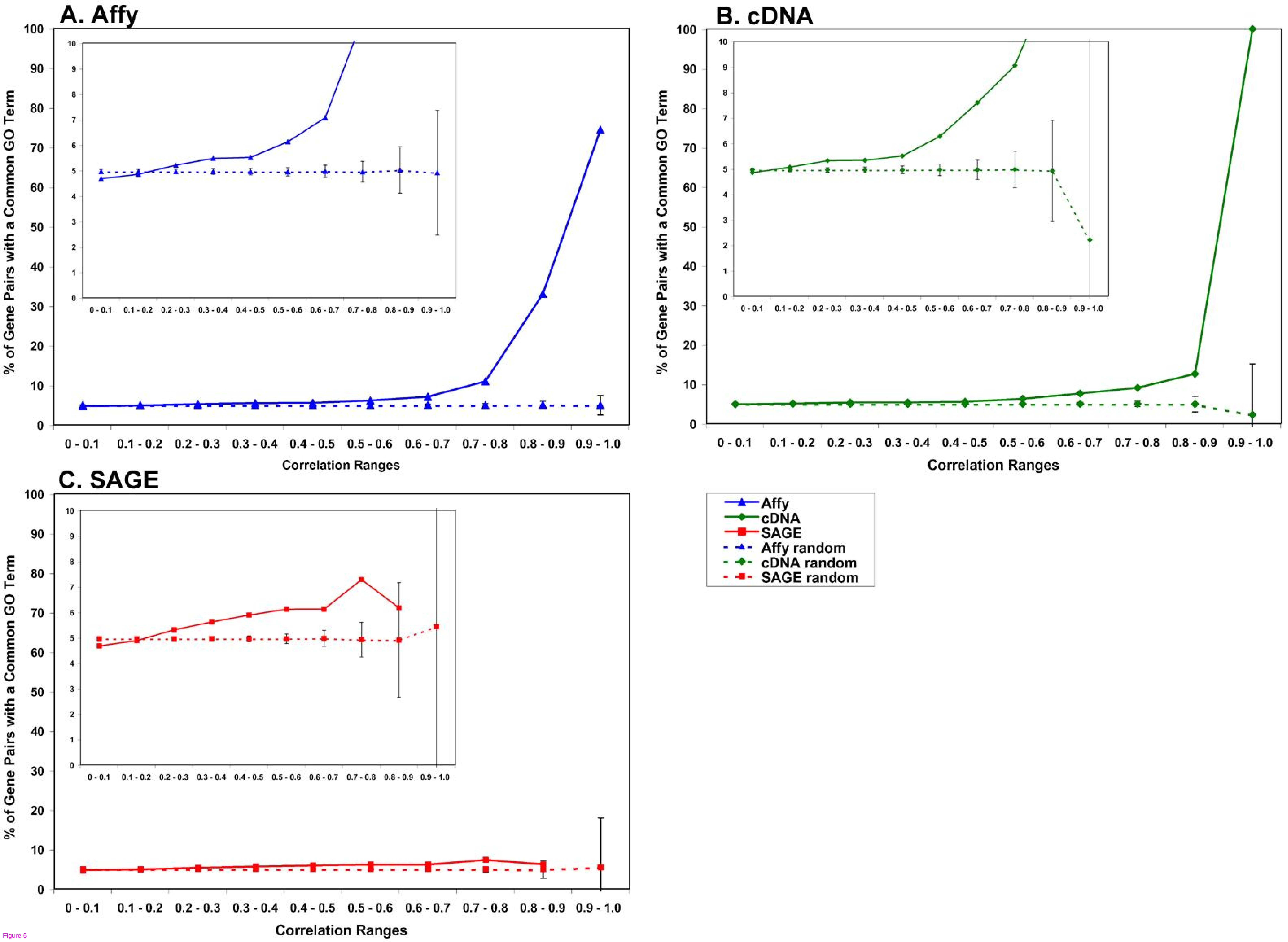

A. Affy
B. cDNA
C. SAGE
% of Gene Pairs with a Common GO Term
100
90
80
70
60
50
40
30
20
10
0
10
9
8
7
6
5
4
3
2
1
0
0 - 0.1
0.1 - 0.2
0.2 - 0.3
0.3 - 0.4
0.4 - 0.5
0.5 - 0.6
0.6 - 0.7
0.7 - 0.8
0.8 - 0.9
0.9 - 1.0
Correlation Ranges
Affy
cDNA
SAGE
Affy random
cDNA random
SAGE random

Figure 6

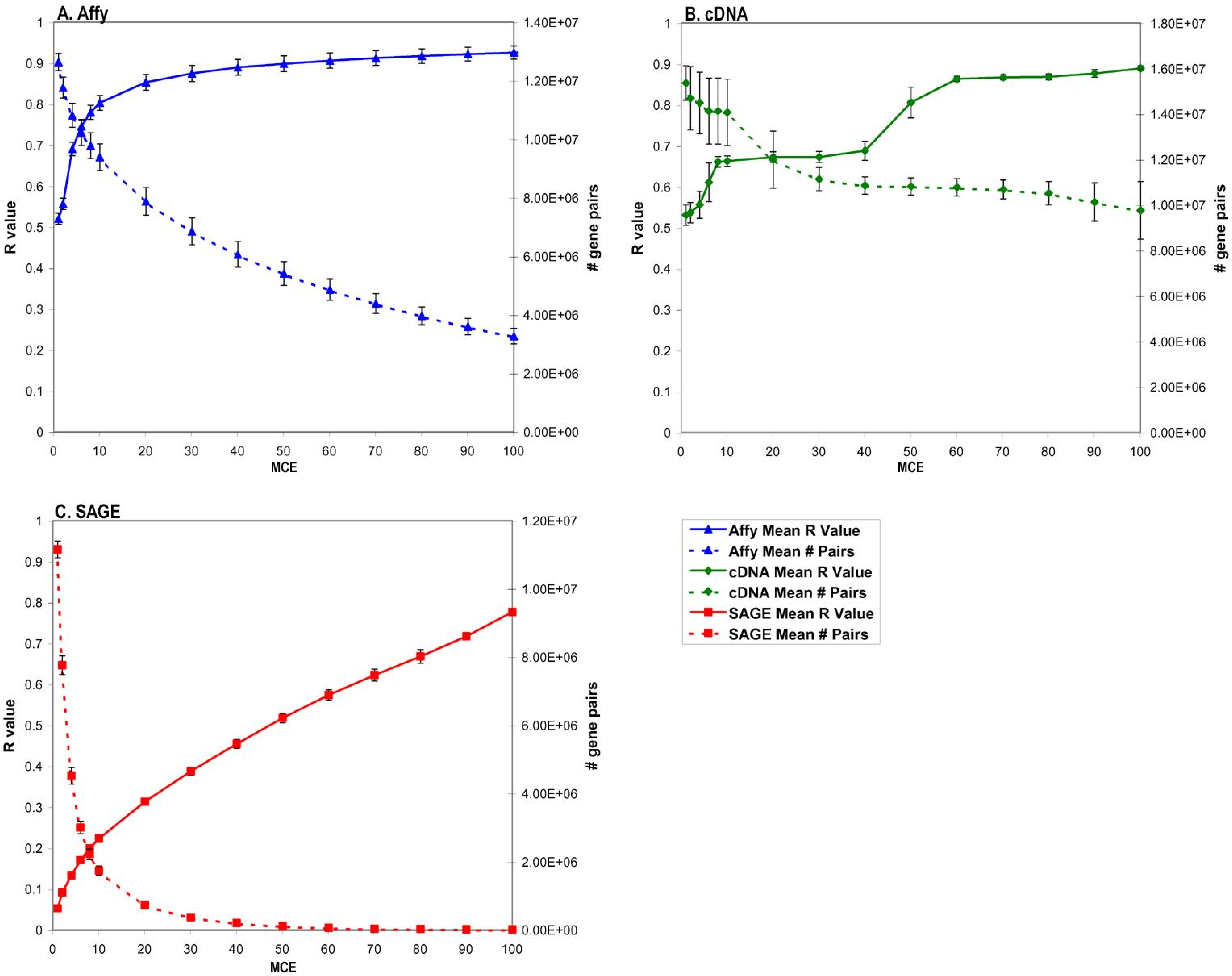

A. Affy
R value
# gene pairs
MCE
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.00E+00
2.00E+06
4.00E+06
6.00E+06
8.00E+06
1.00E+07
1.20E+07
1.40E+07
0
10
20
30
40
50
60
70
80
90
100
B. cDNA
R value
# gene pairs
MCE
0.00E+00
2.00E+06
4.00E+06
6.00E+06
8.00E+06
1.00E+07
1.20E+07
1.40E+07
1.60E+07
1.80E+07
C. SAGE
R value
# gene pairs
MCE
0.00E+00
2.00E+06
4.00E+06
6.00E+06
8.00E+06
1.00E+07
1.20E+07
Affy Mean R Value
Affy Mean # Pairs
cDNA Mean R Value
cDNA Mean # Pairs
SAGE Mean R Value
SAGE Mean # Pairs

Figure 7

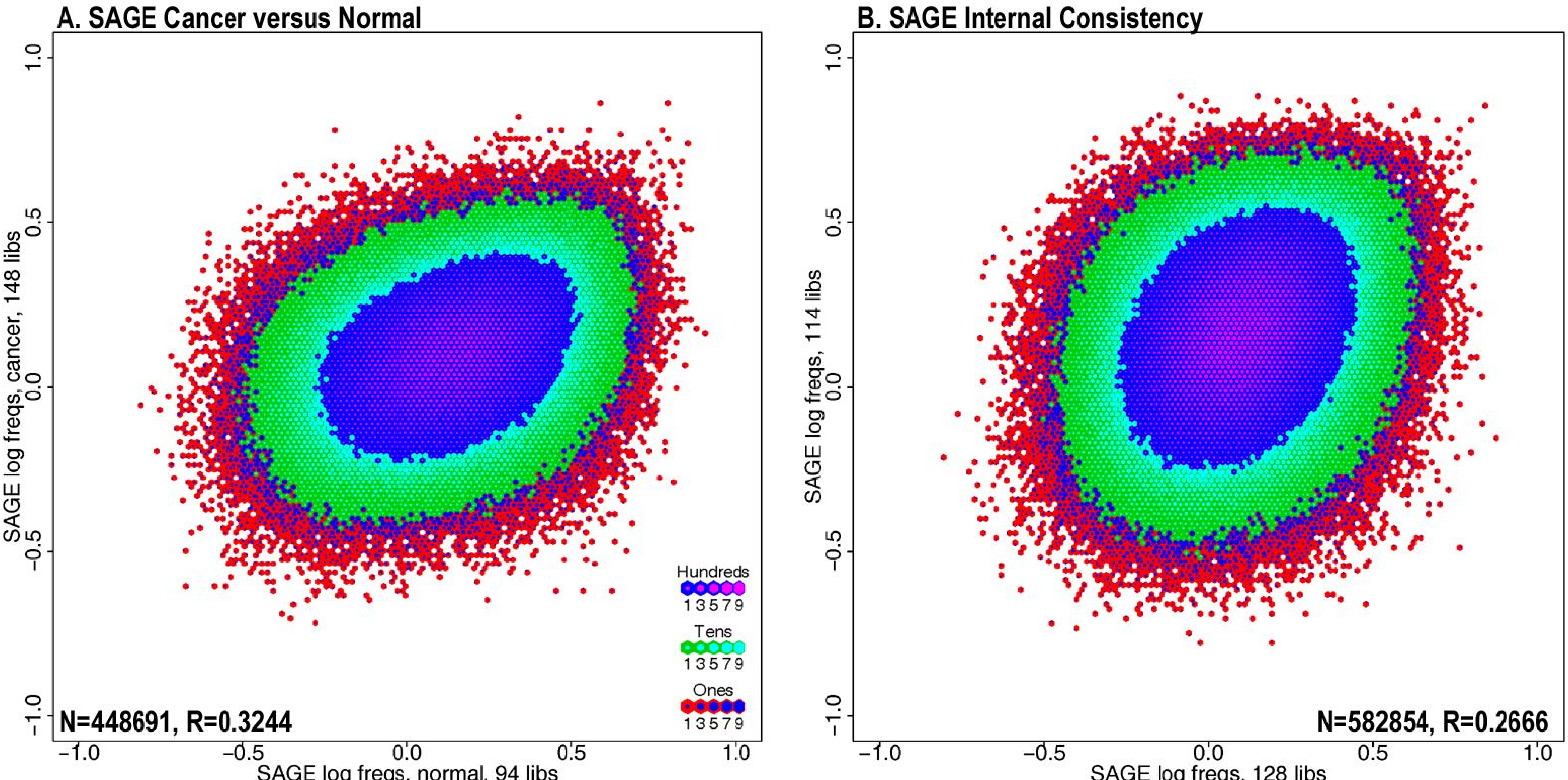

A. SAGE Cancer versus Normal
B. SAGE Internal Consistency
SAGE log freqs, cancer, 148 libs
SAGE log freqs, normal, 94 libs
SAGE log freqs, 114 libs
SAGE log freqs, 128 libs
1.0
0.5
0.0
−0.5
−1.0
Hundreds
Tens
Ones
1 3 5 7 9
N=448691, R=0.3244
N=582854, R=0.2666

Figure 8

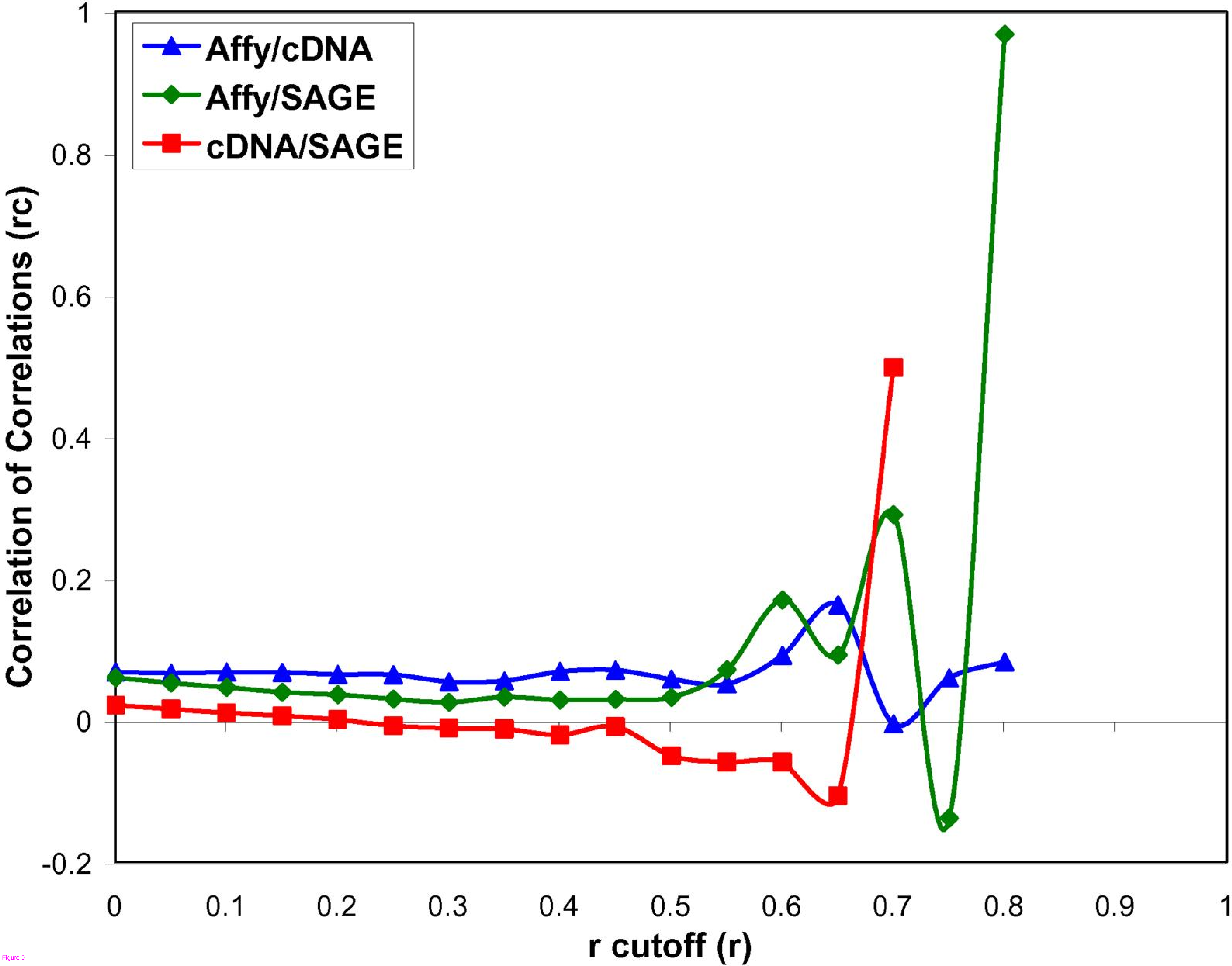
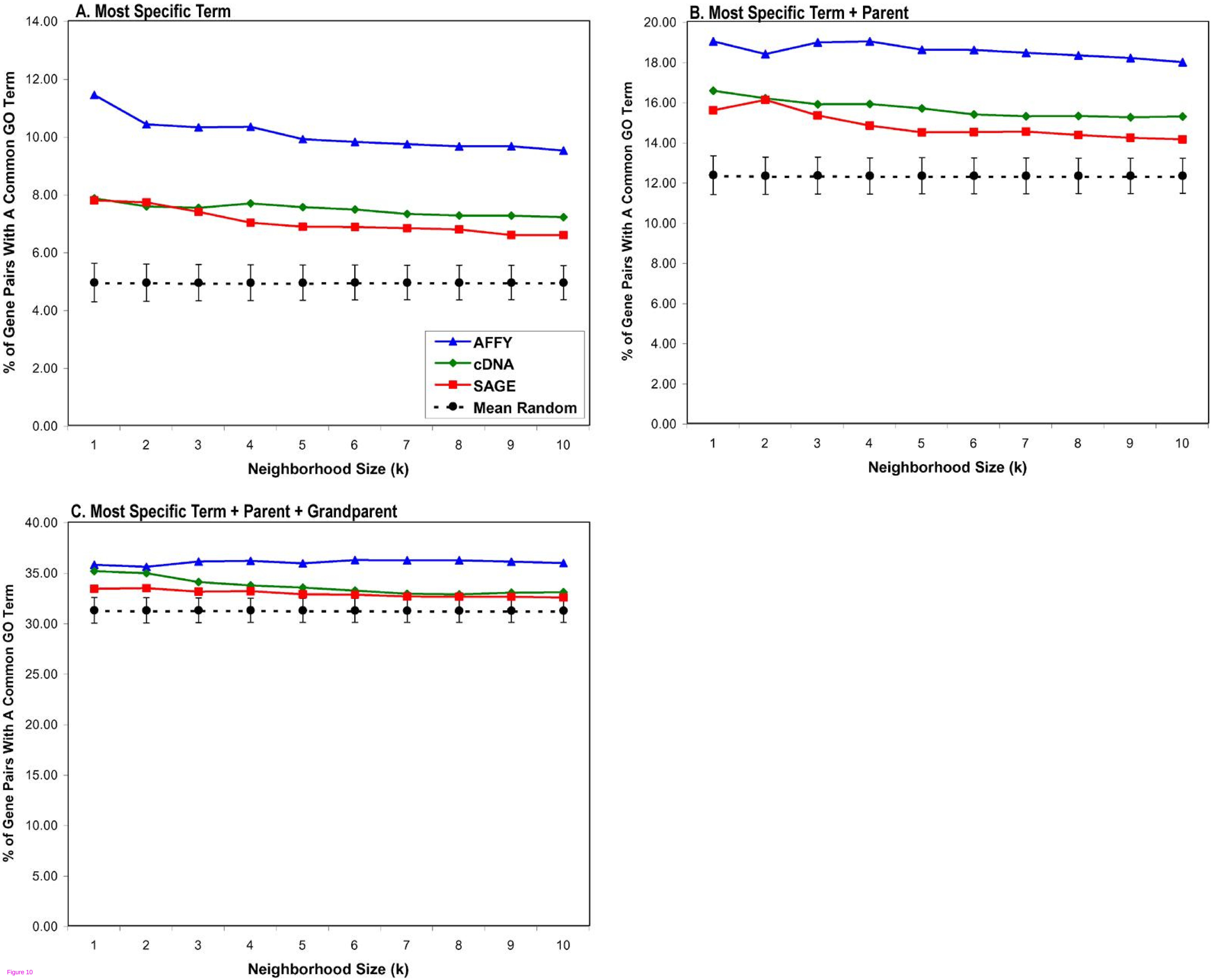

Affy/cDNA
Affy/SAGE
cDNA/SAGE
Correlation of Correlations (rc)
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
r cutoff (r)

Figure 9



A. Most Specific Term
% of Gene Pairs With A Common GO Term
0.00
2.00
4.00
6.00
8.00
10.00
12.00
14.00
1
2
3
4
5
6
7
8
9
10
Neighborhood Size (k)
AFFY
cDNA
SAGE
Mean Random
B. Most Specific Term + Parent
% of Gene Pairs With A Common GO Term
0.00
2.00
4.00
6.00
8.00
10.00
12.00
14.00
16.00
18.00
20.00
1
2
3
4
5
6
7
8
9
10
Neighborhood Size (k)
C. Most Specific Term + Parent + Grandparent
% of Gene Pairs With A Common GO Term
0.00
5.00
10.00
15.00
20.00
25.00
30.00
35.00
40.00
1
2
3
4
5
6
7
8
9
10
Neighborhood Size (k)

Figure 10