Gene Expression Platforms for Global Co-Expression Analyses

A Comparison of spotted cDNA microarrays, Affymetrix microarrays, and SAGE

Obi Griffith, Erin Pleasance, Debra Fulton, Misha Bilenky, Sheldon McKay, Mehrdad Oveisi
Peter Ruzanov, Kim Wong, Scott Zuyderduyn, Asim Siddiqui, and Steven Jones

British Columbia Cancer Agency
Vancouver, British Columbia, Canada

Canada's Michael Smith Genome Sciences Centre
www.bcgsc.ca

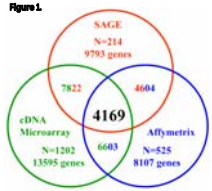
1. Abstract

We have conducted a comprehensive comparison of three major expression platforms: cDNA microarray, oligonucleotide microarray and serial analysis of gene expression (SAGE) using large sets of available data for Homo sapiens. Several studies have compared two of the three platforms to evaluate the consistency of expression profiles for a single tissue or sample set but none have determined if these translate into reliable co-expression patterns for global analyses across many conditions. To this end we analysed a recently published data set of 1202 cDNA microarray experiments (Stuart et al. 2003), 214 SAGE libraries from CGAP and internal sources, and 525 Affymetrix (HG-U133A) oligonucleotide microarray experiments from GEO. All expression data were assigned to LocusLink Ids resulting in an overlap set of 4169 unambiguously mapped genes represented in all three platforms. Using standard co-expression analysis methods, we have assessed each platform for internal consistency and performed all pairwise platform comparisons. Internal consistency was determined by randomly dividing the datasets in half and comparing the Pearson distances for each subset: Affymetrix gave an $r = 0.96$, SAGE an $r = 0.92$, and microarray an $r = 0.58$ ($p < 0.001$). Despite these levels of internal consistency, all pairwise comparisons found poor correlation between platforms ($r < 0.1$, $p < 0.001$). Comparison against the Gene Ontology (GO) showed that all three platforms identify more co-expressed gene pairs with common GO biological process annotations than random data. However, SAGE and Affymetrix performed equally best with microarray performing only slightly better than random.

2. Gene Expression Data

Human gene expression data for three major expression platforms (see sidebar) were collected. We used a recently published data set of 1202 cDNA microarray experiments (Stuart et al., 2003), 214 SAGE libraries from the Cancer Genome Anatomy Project (CGAP) and internal sources, and 525 Affymetrix HG-U133A oligonucleotide microarray experiments from the Gene Expression Omnibus (GEO). SAGE tags were mapped to genes by the lowest sense tag predicted from Refseq or MGC sequences. Gene Ids from all three platforms were then mapped to LocusLink and the intersection determined.

Figure 1.



3. Methods

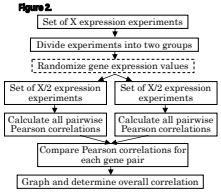
Gene Expression Analysis [sections 4-6]

Pearson correlations between genes were calculated using a modified version of the C clustering library (De Hoon *et al.*, 2004). Correlations of correlations were calculated using the R statistical package (v. 1.8.1) and plotted with the R hexbin function.

Internal Consistency Analysis (section 4)

To evaluate the consistency of co-expression observed with each platform, we divide the experiments available and determine co-expression for each subset independently. The results are then compared by calculating a correlation of the gene correlations. If the 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.

Figure 2.



Platform Comparison Analysis (section 5)

As with the internal consistency analysis, a correlation of gene correlations was determined except for each of the three pairwise platform comparisons instead of between subsets of one platform. If the two platforms being compared report the same distance between each gene pair, the overall correlation between platforms should be near 1.

Ranked Best Match Analysis (section 6)

Instead of considering the actual Pearson correlation between each gene pair and comparing between platforms, the correlation rank was considered. For example, 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.

Gene Ontology Analysis (section 7)

The Gene Ontology (GO) MySQL database dump (release 200402 of assocdb) was downloaded and a GO MySQL database was constructed. The most specific GO annotations for all genes found in common with our dataset were extracted (1908 genes excluding those inferred from electronic annotations (IEAs) and 2627 genes when IEAs were included) and written out to files. PERL scripts were developed to evaluate the number of gene pairs annotated to a common GO term node across a gene's expression similarity neighborhood for each platform. Similar analyses were implemented to evaluate ranked best match genes between platforms found at common GO terms and to evaluate the cardinalities of these gene pair neighborhoods within each platform.

4. Internal Consistency Analysis

Figure 3. SAGE

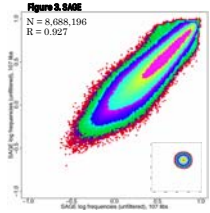
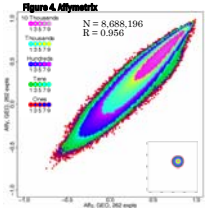


Figure 4. Affymetrix



Figures 3-5: Internal consistency of expression datasets. Affymetrix shows the highest internal correlation of 0.96, then SAGE with correlation 0.92, and cDNA microarray with correlation 0.56.

Inset: same plot for randomized data.

Figure 6: The low internal consistency observed in Fig. 5 is due to a large number of missing values in the cDNA microarray data. Different arrays were used in different experiments, and not all genes are present on all the arrays. Thus, for genes that are rarely present on the same array, Pearson correlations are calculated based on very few overlapping data points. Increasing the required number of overlapping data points increases the internal consistency.

Figure 7: When gene pairs sharing less than 100 data points are excluded, the internal consistency of the cDNA microarray dataset rises to 0.88, comparable to that seen for SAGE and Affymetrix.

Figure 5. cDNA Microarray

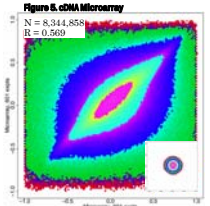


Figure 6.

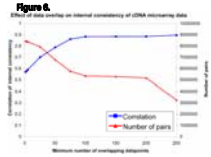
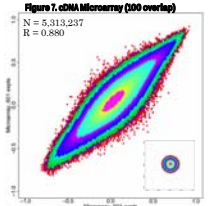


Figure 7. cDNA Microarray (100 overlap)



5. Platform Comparison Analysis

Figures 8-10: Cross platform comparisons. Despite the high levels of internal consistency observed above, surprisingly poor levels of consistency were observed between platforms. The distribution for each platform appeared nearly random and showed correlations of $r < 0.1$. Affymetrix versus SAGE showed the best correlation of 0.094, then Affymetrix versus cDNA microarray with 0.093, and finally cDNA microarray versus SAGE with 0.014. There are several possible explanations for this observation: One possibility is that one platform is correct and the others incorrect. A more likely explanation is that each platform identifies different co-expression patterns because the available data for each platform represents different tissue sources and experimental conditions. Yet another possibility is that few genes are actually consistently co-expressed in biological systems.

Figure 8. Affymetrix vs. SAGE

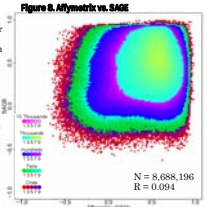


Figure 9. cDNA Microarray vs. SAGE

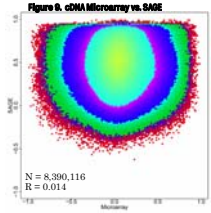
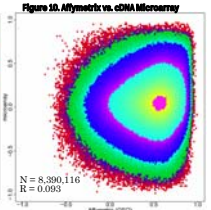


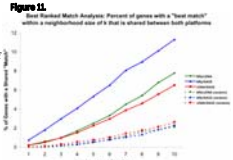
Figure 10. Affymetrix vs. cDNA Microarray



6. Ranked Best Match Analysis

Figure 11: The ranked best match analysis shows that different expression platforms do sometimes identify the same co-expressed genes. The Affymetrix versus SAGE platforms showed the best agreement with 11.3% of genes having a co-expressed gene of Pearson rank 10 or better confirmed by both platforms. Affymetrix versus cDNA microarray agreed for 7.7% of genes, and cDNA microarray versus SAGE for 6.5%. To clarify, a shared best match would be something like 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 counts as one shared 'match' for a neighborhood of k = 3 for the Affymetrix versus SAGE comparison.

Figure 11.



7. Gene Ontology (GO) Analysis

Figure 12.

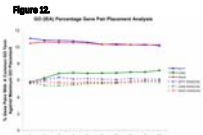


Figure 12: GO biological process domain knowledge was used to evaluate gene co-expression predictions for each platform. The proportion of gene pairs annotated at a specific common GO term for a given gene were enumerated and compared against the maximum number of gene pairs that share GO terms for a given gene across each neighborhood distance. In general, the three platforms perform better than random. Affymetrix and Sage platforms placed 10-11% (4-5% above random) of their co-expressed gene pairs at common GO terms, while the cDNA microarray platform placed 5-7% (1% above random) of it's maximum GO placements.

Figure 13.

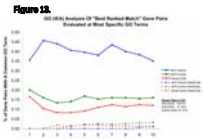
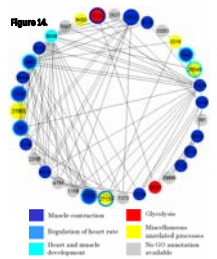


Figure 13: Gene pairs identified by more than one platform in the ranked best match analysis (Section 6) were assessed for sharing common GO Biological process descriptions. Although all platform comparisons performed better than random, the proportions of gene pairs found with common GO terms were relatively low. The Affymetrix versus SAGE comparison was able to identify 0.35-0.46% gene pairs in common GO terms (0.15-0.25% more than the other platform comparisons). The Affymetrix versus cDNA comparison identified 0.14-0.20% GO placements while the cDNA versus SAGE comparison only placed 0.08-0.16% of it's gene pairs.

Coexpression network example

Figure 14: Identifying groups or networks of coexpressed genes is the starting point for analyses such as functional annotation of novel genes or identification of transcription factor binding sites. As the co-expressed gene pairs found by both Affymetrix and SAGE platforms appear to be the most reliable, we used these pairs to identify an example of a co-expression network. The network was constructed starting from the alpha actin gene (LocusLink ID 70) and iteratively expanded to include all genes co-expressed with any other genes in the network. Those which have at least two connections to other genes within the network were visualized using Cytoscape (Shannon *et al*, 2003), and coloured based on their GO biological process annotation. Interestingly, nearly all genes in this network are involved in processes related to muscle function and development, demonstrating that co-expression can be closely related to gene function.

Figure 14.



8. Conclusions

- Co-expressed genes can be identified based on large-scale gene expression data
- Internal consistencies are fairly high for co-expression patterns identified by Affymetrix (R=0.96), SAGE (R=0.92) and cDNA microarray (R=0.56)
- cDNA microarray data are more consistent when sufficient data is available (R=0.88)
- Direct comparison of correlation values between platforms yields poor correlations (R<0.1)
- Comparison of gene rank shows significant overlap in coexpressed pairs identified by different platforms, particularly between Affymetrix and SAGE
- Gene pairs identified as coexpressed are more likely to share the same function
- Coexpressed gene pairs identified by more than one platform which also share functional annotations are most likely to 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

Acknowledgments

funding | Natural Sciences and Engineering Council of Canada (for OG and EP); Michael Smith Foundation for Health Research (for OG, SJ and EP); CIHR/MSFHR Bioinformatics Training Program (for DF); Killam Trusts (for EP); Genome BC

references | 1. Stuart *et al.* 2003. *Science.* 302(5643):249-255; 2. De Hoon *et al.* 2004. *Bioinformatics.* Feb 10 [epub ahead of print]; 3. Shannon *et al.* (2003). *Genome Res* 13:2498-2504.

SAGE

Serial analysis of gene expression (SAGE) is a method of large-scale gene expression analysis that involves sequencing small segments of expressed transcripts ("SAGE tags") in such a way that the number of times a SAGE tag sequence is observed is directly proportional to the abundance of the transcript from which it is derived.

A description of the protocol and other references can be found at www.sagenet.org.

cDNA Microarrays

cDNA Microarrays simultaneously measure expression of large numbers of genes based on hybridization to cDNAs attached to a solid surface. Measures of expression are relative between two conditions.

For more information, see www.microarrays.org.

Affy Oligo Arrays

Affymetrix oligonucleotide arrays make use of tens of thousands of carefully designed oligos to measure the expression level of thousands of genes at once. A single labeled sample is hybridized at a time and an intensity value reported. Values are the based on numerous different probes for each gene or transcript to control for non-specific binding and chip inconsistencies.

For more information, see www.affymetrix.com.