Available online at www.sciencedirect.com

SCIENCE DIRECT

Genomics 86 (2005) 476 – 488

GENOMICS

www.elsevier.com/locate/ygeno

Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses

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

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

Received 22 March 2005; accepted 16 June 2005
Available online 10 August 2005

Abstract

Large amounts of gene expression data from several different technologies are becoming available to the scientific community. A common practice is to use these 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 ($r_c < 0.11$). Assessment against 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.

Keywords: Gene expression; Gene expression profiling; Microarray analysis; cDNA microarray; Oligonucleotide microarray; Coexpression; Serial analysis of gene expression; Gene Ontology

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 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.

High degrees of consistency within a platform have been reported for cDNA microarrays and Affymetrix oligonucleo-

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.

* Corresponding author. Fax: +1 604 876 3561.
E-mail address: sjones@bcgsc.ca (S.J.M. Jones).

0888-7543/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2005.06.009

tide microarrays [9–11]. The reproducibility of SAGE has not been demonstrated given that the time and cost required to produce individual SAGE libraries are high. 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, personal communication). Cross-platform comparisons of gene expression values have found "reasonable" correlations for matched samples, especially for more highly expressed transcripts [11,13–19]. Other comparisons have reported "poor" correlations [15,18,20–24]. 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 *Caenorhabditis elegans* and *Saccharomyces cerevisiae* have been used with some success to identify gene function or genes that are coregulated [26–28]. 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 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 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 comprising 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.

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. We have compared large publicly available datasets for SAGE, cDNA microarray (cDNA), and Affymetrix oligonucleotide microarray (Affymetrix) platforms (Supplemental 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. 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 two-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 approach we identify a set of coexpressed gene pairs that represents the best of each.

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 (Figs. 1A–1C). 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 Materials and 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 the minimum required number of common or shared experiments for which any two genes actually have values available in their respective expression profiles (Fig. 1D).