Genomics 110 (2018) 67–73

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

How many differentially expressed genes: A perspective from the comparison of genotypic and phenotypic distances

Bi Zhao, Aqeela Erwin, Bin Xue*

Department of Cell Biology, Microbiology and Molecular Biology, School of Natural Sciences and Mathematics, College of Arts and Sciences, University of South Florida, 4202 East Fowler Ave. ISA2015, Tampa, Florida, 33620, USA

ARTICLE INFO

Keywords:
Microarray data analysis
Differentially expressed genes
Genetic distance
Genotypically and phenotypically significant DEGs
gpsDEGs

ABSTRACT

Identifying differentially expressed genes is critical in microarray data analysis. Many methods have been developed by combining *p*-value, fold-change, and various statistical models to determine these genes. When using these methods, it is necessary to set up various pre-determined cutoff values. However, many of these cutoff values are somewhat arbitrary and may not have clear connections to biology. In this study, a genetic distance method based on gene expression level was developed to analyze eight sets of microarray data extracted from the GEO database. Since the genes used in distance calculation have been ranked by fold-change, the genetic distance becomes more stable when adding more genes in the calculation, indicating there is an optimal set of genes which are sufficient to characterize the stable difference between samples. This set of genes is differentially expressed genes representing both the genotypic and phenotypic differences between samples.

1. Introduction

Microarray is a powerful technique for studying gene expression profiles at genome level [1]. This technique has been broadly used in different species, tissues, cell lines, and under different conditions [2] since it was invented in the early 80s of last century [3]. To ensure the reliability of experimental results, different but related samples are often used in the same microarray experiment. The samples used as a reference are called “control” samples, while the other samples that are under different phenotypic status or subject to various treatments are called “treated” or “target” samples. There may be multiple groups of treated samples in one microarray experiment. In many cases, each group contains several samples, which are also known as replicates. Among these groups, the expression levels of genes may be different. When the expression levels are significantly different, the corresponding genes are called Differentially Expressed Genes (DEGs). These DEGs are assumed to be the molecular driving force and/or the molecular biomarkers of different phenotypes.

In microarray data analysis, the difference of expression level of a gene between different samples is often shown by a ratio of the expression level of the gene in the target sample to the ratio of the gene in the control sample. This ratio is further scaled using base 2 logarithm to make another quantity called log2 ratio, the absolute value of log2 ratio is known as fold-change (FC) [4]. FC is a very important quantity to show the change of expression levels of genes. Fold-change analysis is actually a very intuitive method to identify DEGs [5]. In fold-change analysis, 2 or 1.5 is often used as the cutoff to choose DEGs of which the FC values are larger than the cutoff. Student *t*-test is also used to measure the statistical significance of each gene under the null hypothesis that the gene is not differentially expressed in different samples [6]. After using *t*-test, each gene is assigned a *p*-value. DEGs can also be selected based on *p*-value cutoffs. Since student *t*-test may be over simplified, SAM (significance analysis of microarrays) [7] was later designed to refine the results of *t*-test and has become very popular since then. In addition, many new methods were also recently developed based on new models, statistics, and algorithms, such as Rank product [8], Outlier Robust [9], Outlier Sums [10,11], Ranking Analysis [12–14], Mao and Sugihara's method [15], and SPRING [16]. While being designed to solve specific problems, these methods may not be very effective for other issues, which eventually lead to high false-positive rates and low reproducibility that are very common in microarray data analysis [17–19].

In a recent comprehensive comparison using datasets generated by the Microarray Quality Control (MAQC) project [17], the combination of non-stringent *p*-value from *t*-test and FC ranking was shown to improve the identification of DEGs significantly [20]. This study provides a generic strategy for the identification of DEGs, which has become rather prevailing in microarray data analysis. Nonetheless, it is still a challenge to determine the cutoff values and the actual numbers of DEGs solely based on that strategy. In this direction, several new

* Corresponding author.
E-mail address: binxue@usf.edu (B. Xue).

http://dx.doi.org/10.1016/j.ygeno.2017.08.007
Received 15 May 2017; Received in revised form 17 August 2017; Accepted 21 August 2017
Available online 24 August 2017
0888-7543/

strategies have been tested. Probability-based decision-tree was applied on breast cancer microarray data to identify DEGs and the results were consistent with the results from the method combining non-stringent *p*-value and 2-fold FC cutoff [21]. In another study, the distance between the distribution of a gene's expression profiles in different groups of samples was calculated and the statistical significance associate with the distance was used to select DEG [22]. While having provided novel implements, these "data analysis" based methods may not adequately address the biology behind the data of microarray experiments. By notation, DEGs are genes that differentiate the samples at both genotypic and phenotypic levels, and therefore the DEGs should be linked to a quantitative measure that counts for both the genotypic and phenotypic differences between samples. Based on this assumption, we developed a distance-based method to identify those genotypically and phenotypically and significant DEGs (gpsDEGs). These gpsDEGs are associated with the stable genetic distances between different groups of samples.

2. Method

2.1. Microarray data

Eight sets of microarray data were extracted from the Gene Expression Omnibus (GEO) database [23]. The microarray experiment for each of these datasets was designed to evaluate the responses of different treatments on a specific type of carcinomatous cell line. All the experiments used Affymetrix platform. The eight types of cancer are: acute lymphoblastic leukemia, breast cancer, colorectal cancer, kidney cancer, liver cancer, lung cancer, pancreatic cancer, and prostate cancer. The corresponding GEO entries are: GSE19315 [24], GSE53394 [25], GSE7259 [26], GSE65168 [27], GSE41804 [28], GSE6400 [29], GSE57728 [30], and GSE22606 [31]. A summary of these GEO entries was presented in Table 1.

2.2. Bioconductor and DEG analysis

The CEL format files of afore-mentioned microarray experimental data were downloaded from GEO, and then processed using RMA [32] for background correction and normalization. Afterwards, Limma [33,34] was used to calculate *p*-value, adjusted *p*-value, and FC for all the genes in the array. Then, an adjusted *p*-value cutoff 0.1 was used to select a preliminary list of genes, and this list of genes were ranked using FC as suggested by a recent benchmark study [20]. Clearly, in the final list of ranked genes, different sets of top genes can be selected by using different FC cutoffs. These different sets of genes are various sets of DEGs associated with different FC cutoffs [20].

2.3. Intra-group and inter-group distances

The microarray experiments selected in this study each contains one control group and normally three treated groups, with each group having often three replicates. Once having selected a set of ranked genes, the scaled expression levels of these ranked genes in any two samples can be used to calculate the distance between these two samples using the following equation:

$$d_{j,j2} = \frac{1}{N}\sqrt{\sum_{k=1}^{N}(E_{j,k} - E_{j2,k})^2}$$

where, $d_{j,j2}$ is the distance between sample j and j2. N is the total number of ranked genes in the calculation. $E_{j,k}$ is the scaled expression level of the k-th gene in the j-th sample. The scaled expression level is the ratio of the expression level of a gene in a treatment group to the averaged expression level of the same gene in the control group.

In addition, all the samples in the i-th group can be characterized by an intra-group distance d_i, which is calculated by:

$$d_i = \frac{1}{N^{*}\mathrm{C}(\mathrm{J},2)}\sqrt{\sum_{j=1}^{J-1}\sum_{j2=j+1}^{J}\sum_{k=1}^{N}(E_{i,j,k} - E_{i,j2,k})^2}$$

where, $E_{i,j,k}$ is the scaled expression level of the k-th gene in the j-th sample that belongs to the i-th group. J is the total number of samples in the i-th group. "j" and "j2" stand for two different samples in the i-th group. C(J,2) is the number of pairwise combinations of all the J samples in the i-th group.

Similarly, the inter-group distance $D_{i,i2}$ can be calculated using:

$$D_{i,i2} = \frac{1}{N^{*}J^{*}J2}\sqrt{\sum_{j=1}^{J}\sum_{j2=1}^{J2}\sum_{k=1}^{N}(E_{i,j,k} - E_{i2,j2,k})^2}.$$

Here, i and i2 represent the i-th and i2-th groups, respectively. $E_{i,j,k}$ and N have the same meaning as explained before. J and J2 are the numbers of samples in the i-th and i2-th group, accordingly.

Clearly, the distance is a measure of the genetic distance between any two samples, or among samples in one group (intra-group distance), or between different groups (inter-group distance). The inter-group distance by its notation can also be used to measure the phenotypic difference between groups. For all the above-mentioned types of distance, the distance is determined by the number of ranked genes and the expression levels of genes. Different sets of ranked genes may yield different values of distance. Since the genes are ranked by FC from

Table 1
Summary of the eight sets of microarray data.

GSE No.	Type of cancer	Tissue/cell line	GPL	Year and reference	Description
GSE19315 [ref. 24]	Acute lymphoblastic leukemia	THP-1	GPL570	2010	Cells were either untreated or treated by Shiga toxin type 1 or LPS. Three sample groups with three replicates in each group.
GSE53394 [ref. 25]	Breast cancer	MCF-7	GPL96	2013	There are four groups of samples and two replicates in each group.
GSE7259 [ref. 26]	Colorectal cancer	Caco-2	GPL571	2007	Cells were either untreated or treated with quercetin for five or ten days. Four groups of samples and two replicates in each group.
GSE65168 [ref. 27]	Kidney cancer	RCC (786-O)	GPL6244	2015	Cells were untreated or transfected with VHL under nomoxia- and hypoxia-conditions, resulting four groups of samples with two replicates in each group.
GSE41804 [ref. 28]	Liver cancer	(See the right-hand-side)	GPL570	2012	Heptocellular carcinoma and non-cancerous tissues with IL28B and TG/GG genotypes. There are four groups and ten replicates in each group.
GSE6400 [ref. 29]	Lung cancer	A549	GPL570	2006	Cells were treated with Manitol, actinomycin D, or two doses of sapphyrin PCI-2050. There are four groups of samples, with three replicates in each group.
GSE57728 [ref. 30]	Pancreatic cancer	AsPC1	GPL570	2014 [?]	Cells were treated with ICG-001, siRNA-mediated knockdown of CTNNB1, or DMCO (vehicle control) for 6, 24, or 48 h, resulting six groups of samples with two or three replicates in each group.
GSE22606 [ref. 31]	Prostate cancer	LNCaP, SRF silenced LNCaP	GPL570	2010 [?]	Non-organ-confined prostate cancer. LNCaP cells were treated with ethanol vehicle, or R1881, or siRNA-mediated SRF silencing plus R1181. There are four groups of samples with three replicates in each group.