The Fist Draft

Mehrdad Samadishadlou

2022-09-19

Materilas and Methods

In order to come by sufficient classification power for MI samples and others, the sample size of the dataset should be relatively large. Therefore, NIH's Gene Expression Omnibus (GEO) repository from the National Institute of Health was searched. Finally, the MI and CAD samples were acquired from GSE59867 and control samples from GSE56609. Moreover, GSE62646 and GSE54475 were used as validation set in both DEGs and pathway analyse as well as machine learning section for. All four datasets are microarray experiments which have been conducted using Affymetrix Human Gene 1.0 ST Array (GPL6244) platform. More details about datasets are available in table ?.

Pre-processing

Raw data (CEL files) of the four datasets were downloaded from the GEO and preprocessed using the fRMA package M. N. McCall, Jaffee, and Irizarry (2012). fRMA allows to preprocess individual microarray samples and combine them consistently for analysis. For each dataset, background correction is performed and then it is quantile normalized based on the reference distribution. During summarization, batch effects are removed and variances of the gene expressions are estimated by taking into account these probe-specific effects. For those multiple probe sets matched to the identical gene, the mean log fold change was retained.

Statistical methods

Barcode algorithm

The barcode algorithm proposed by McCall et al. (Matthew N. McCall et al. 2011) transformes the actual expression values into binary barcode values. Huge sets of samples were collected and normalized using fRMA for several platforms. The distribution of the expressed and non-expressed observed intensities for each gene is estimated using these normalized sets. Genes are deemed expressed (and their value coded to 1) or unexpressed (and their value coded to 0) according to the following equation:

$$\hat{x}_{ij} = \begin{cases} 1 & \text{if } x_{ij} >= \mu^{ne} + C \times \sigma^{ne} \\ 0 & \text{otherwise} \end{cases}$$

where x_{ij} is the normalized intensity of gene i in sample j, C is a user-defined parameter, σ^{ne} is the standard deviation of the non-expressed distribution and μ^{ne} is the mean of the non-expressed distribution. The barcode representation of a sample is a vector of ones and zeros denoting which genes are estimated to be expressed (ones) and unexpressed (zeros). The barcode algorithm was implemented by the barcode function in the R fRMA package, and the default value of C was used.

Differentially expressed genes

To determine if the expressed ratios differed in the diseased group versus the control group, Fisher's exact test for individual genes was carried out upon the barcode values. Genes with a false discovery rate (FDR) of < 0.05, which was calculated through the Benjamini-Hochberg (BH) procedure to adjust for multiple testing issue, were considered as differentially expressed genes.

Pathway enrichment analysis

Using the R clusterProfilter package, the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis and Gene Ontology (GO) functional annotation were carried out on the differentially expressed genes. In these analyses, all default parameters were used.

Results

Refrences

McCall, M. N., B. M. Bolstad, and R. A. Irizarry. 2010. "Frozen Robust Multiarray Analysis (fRMA)." *Biostatistics* 11 (2): 242–53. https://doi.org/10.1093/biostatistics/kxp059.

McCall, M. N., H. A. Jaffee, and R. A. Irizarry. 2012. "fRMA ST: Frozen Robust Multiarray Analysis for Affymetrix Exon and Gene ST Arrays." *Bioinformatics* 28 (23): 3153–54. https://doi.org/10.1093/bioinformatics/bts588.

McCall, Matthew N., Karan Uppal, Harris A. Jaffee, Michael J. Zilliox, and Rafael A. Irizarry. 2011. "The Gene Expression Barcode: Leveraging Public Data Repositories to Begin Cataloging the Human and Murine Transcriptomes." *Nucleic Acids Research* 39 (suppl_1): D1011–15. https://doi.org/10.1093/nar/gkq1259.