Pipeline for Integrated Microarray Expression Normalization Toolkit (PIMENTo)

Thomas Nash1, Matthew Huff2, Sean M Courtney1,3, E. Starr Hazard1,4, Gary Hardiman1,5

1MUSC Bioinformatics, Center for Genomics Medicine, 2MS in Biomedical Sciences Program, 3Department of Pathology & Laboratory Medicine, 4Library Science and Informatics, 5Departments of Medicine & Public Health, Medical University of South Carolina (MUSC), Charleston SC

Background

• DNA microarray technology has been used for genome-wide gene expression studies that incorporate molecular genetics and computer science analyses on massive levels [1-3]. The availability of microarrays permit the simultaneous analysis of tens of thousands of genes for the purposes of gene discovery, disease diagnosis, improved drug development, and therapeutics tailored to specific disease processes.
• We have developed a Pipeline for Integrated Microarray Expression & Normalization Toolkit (PIMENTo).
• The objective was to integrate existing open source software and processes and in house scripts into a simple, easy-to-use interface and tool kit.
• The longer term goal is to create a pipeline which researchers with varying levels of programming experience can fully implement with ease.
• A prototype has been built, tested, and exploited for series of analyses. PIMENTo integrates disparate open-source components into an integrated package that rapidly automates background subtraction, normalization and data QC, and produces both text and graphic experimental summaries.

Methodology

• Probes whose expression level exceeds a threshold value in at least one sample are called detected. The threshold value is found by inspection from the distribution plots of (log) expression levels. Expression level data from the Illumina BeadStudio software were normalized using quantile or mLOESS algorithms.
• The user provides array intensity data in CSV, Excel, or tab-separated format along with information with regard to the samples.
• Each step of the pipeline allows the user to set limits or thresholds on parameters, such as false discovery rate (FDR), background subtraction, and normalization method. Furthermore, the user can perform sub-setting of arrays for downstream analysis including heatmap creation.

Usage

• The pipeline incorporates many open-source packages freely available through the Bioconductor project to perform the majority of the operations, including “limma” and “affy” for quantile and LOESS normalization, respectively. Significance testing is carried out using the code for Significance Analysis of Microarrays from the R package “samr”. All outputs are saved in both Postscript and PDF formats, heatmaps are further saved as TIFF and FIG.
• Small intestinal crypt expression profiles from wild-type and villin-gp130Act mice as described in Taniguchi et al. [4] study. Genes were ranked by absolute fold change. Weakly expressed genes are removed before normalization and square-root scaling across all arrays. Ward’s method is used for clustering and Euclidean distance is used as the distance function. The colors qualitatively correspond to fold changes with respect to a reference which is calculated as the mid-point between grouped samples.

Bibliography