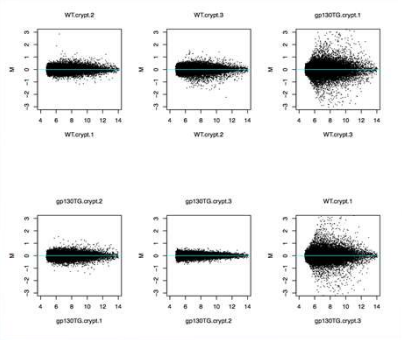
Pipeline for Integrated Microarray Expression Normalization Toolkit (PIMENTo)

Thomas Nash[1], Matthew Huff[2], Sean M Courtney[1,3], E. Starr Hazard[1,4], Gary Hardiman[1,5] `[musc-bioinformatics/]$`

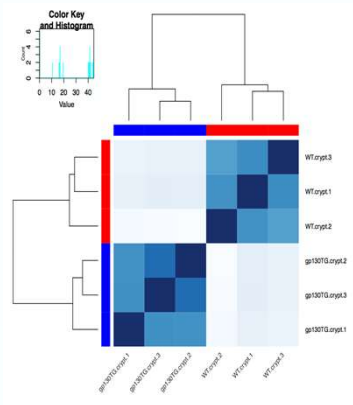
[1]MUSC Bioinformatics, Center for Genomics Medicine, [2]MS in Biomedical Sciences Program, [3]Department of Pathology & Laboratory Medicine, [4]Library Science and Informatics, [5]Departments of Medicine & Public Health, Medical University of South Carolina (MUSC), Charleston SC

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



Example MA plots of quantile normalized Illumina BeadArray data



Example similarity matrix (Euclidean distances between the samples) demonstrating within group and between group biological replicate comparisons. Small intestinal crypts were isolated from wild-type and villin-gp130Act small intestines subjected to Illumina BeadArray analysis as described in Taniguchi et al., [4]. Samples are clustered by similarity. The blocks in the comparison matrix are scaled by color and the most similar are dark blue and least similar are white. The samples from wild-type and villin-gp130Act cluster together indicating global transcriptome differences between the two sample groups.

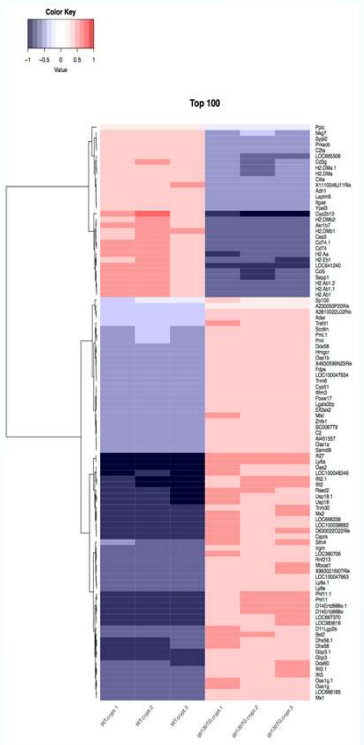
```
# Preprocess data: normalization, MA plots, clustering
example.preprocessed <- preprocessData(inputFile = "example_data_sheet.xlsx",
                                       fileSheet = 1, ntext = 7, dataCol = 8:19,
                                       symbolIndex = 3, idIndex = 4)

# Background subtraction: remove noise and arrays below defined threshold
backgroundCutoff(example.preprocessed, method = "quantile")
backgroundCutoff(example.preprocessed, method = "quantile", xlim.lo = 6.5, xlim.hi = 7.5)
example.subtracted <- backgroundSubtraction(example.preprocessed, method = "quantile",
                                            cutoff = 7.1)

# SAM and sample similarity: significance analysis of genes and compare array groups
example.significant.genes.AvsB <- runSAM(backgroundSub.obj = example.subtracted,
                                         classCompareCols = 8:13,
                                         classCompareName = "AvsB",
                                         fdr.cutoff = 0.1,
                                         response = c(rep(1,3),rep(2,3)))

# Heatmap creation: create heatmaps of expression levels across chosen genes
pathwayHeatmap(runSAM.obj = example.significant.genes.AvsB, pathwaysDir = "heatmap-AvsB",
               fileFormat = "symbol")
```

Example pipeline code to perform preprocessing, background subtraction, normalization, SAM, and heatmap generation of microarray data through the R command-line. In this example, twelve arrays are in the input file but the first six are used with SAM. Three belong to class A and three to class B, significant genes are found amongst this comparison which are then selected by the user and used for downstream analyses including heatmap creation.



Example heatmap of top 100 genes from the 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 compared groups.

Data preprocessing through quantile and mLOESS normalization,

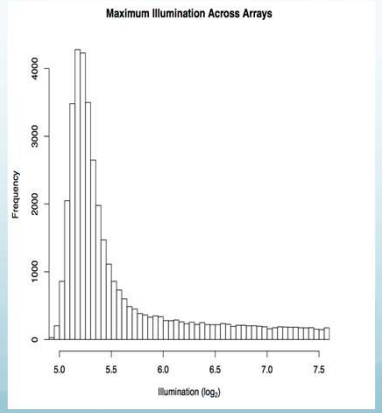
Background subtraction is achieved from user-defined cutoff points based on inspection of distribution plots of (log) expression levels

SAM (Significance Analysis of Microarrays) to identify significant genes, and **sample similarity** comparisons using **PCA** (principle component analysis)

Heatmaps of significant genes based on user-defined lists, identified pathways and ranked gene lists

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 Bead Studio 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 formatting of the file.
- 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 multi-class comparisons.
- Currently the pipeline is accessible through the R command-line or using R studio.



Example histogram of (log) expression levels which is used to identify the threshold value and determine the background cutoff point

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" [5]. 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] were analyzed with PIMENTo.
- Currently the pipeline is available for use as an R package through GitHub at https://www.github.com/TomNash/PIMENTo
- A future release will utilize R Shiny to expand the platform to an easy to use user-interface (UI).

Bibliography

1. Rouse RJ, Field K, Lapira J et al. Development and application of a microarray meter tool to optimize microarray experiments. BMC research notes 1 45 (2008).
2. Hardiman G. Microarray platforms--comparisons and contrasts. Pharmacogenomics 5(5), 487-502 (2004).
3. Trachtenberg AJ, Robert JH, Abdalla AE et al. A primer on the current state of microarray technologies. Methods in molecular biology (Clifton, N.J.) 802 3-17 (2012).
4. Taniguchi K, Wu LW, Grivennikov SI et al. A gp130-Src-YAP module links inflammation to epithelial regeneration. Nature 519(7541), 57-62 (2015).
5. Tibshirani R, Chu G, Narasimhan B, Li J. samr: Significance analysisof microarrays. R package version 2 (2011).