Microarray Gene Expression: Bioinformatics, Class Discovery

Final deliverable includes the seven objectives herein and an executive summary detailing background, methods, results and conclusions in the format most familiar to the scientific community. Dynamic source documents include details of calculations and code used in the data analysis, as well as methodological descriptions that allow the sponsor to directly reproduce computational results for review by the scientific community.

Analysis reports are written using Sweave, a literate programming combination of Latex source and R code. Sweave is open source software that uses standardized report structures to provide the transparency required for reproducible research.

The use of best practices for reporting of research ensures that all analyses are fully open to inspection and verification essential to the peer review process.

**Objective 1 – Basic Quality Control.** Array quality assessment will be performed using the R/Bioconductor SimpleAffy package and arrayQualityMetrics package (Kauffmann et al, 2009).

**Objective 2 – Array Preprocessing.** Background correction and quantile normalization will be performed using Limma (Smyth et al., 2005, Bolstad et al., 2003) or Robust Multi-chip Average (RMA) (Irizarry et al., 2003).

**Objective 3 – Unsupervised Clustering.** Principle Components Analysis (PCA) and/or hierarchical clustering (HC) will be performed to determine the overall relationship between sample types. PCA will be performed using the R prcomp function. HC clustering will be performed using the R hclust function. Results will be depicted as a scatterplot (PCA) or as a clustergram.

**Objective 4 – Non-specific Filtering.** Based on the fact that most genes are not expected to show differences in expression across arrays we may filter probes to decrease noise in the dataset and reduce the impact that multiple testing adjustment has on statistical power (Bourgon et al., 2010). Non-specific filtering will be performed using the R/Bioconductor genefilter library (R package version 1.34.0).

**Objective 5 – Differential Expression.** Differentially Expressed Genes (DEGs) will be identified for each comparison using standard criteria (e.g., Log Ratio p-values < 0.01, and absolute fold change > 1.5). Comparisons will be made using the R/Bioconductor LIMMA library (Smyth et al., 2005). LIMMA fits gene-wise linear models to estimate log-ratios between two or more RNA samples simultaneously. Results of model fitting will be exported to a table, with headers as described in example Table 1.

**Example Table 1.** Example Column headers for LIMMA results.

<table>
<thead>
<tr>
<th>Column Heading</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeneName</td>
<td>The most common identifier for the gene, typically the gene symbol.</td>
</tr>
<tr>
<td>SystematicName</td>
<td>Accession ID for the Genbank, Refseq, or other publically available repositories of sequence data.</td>
</tr>
<tr>
<td>Description</td>
<td>A brief description of the gene corresponding to the probe sequence.</td>
</tr>
<tr>
<td>logFC</td>
<td>Estimate of the log2-fold change corresponding to the effect or contrast (for topTableF there may be several columns of log-fold changes)</td>
</tr>
<tr>
<td>AveExpr</td>
<td>average log2-expression for the probe over all arrays and channels</td>
</tr>
<tr>
<td>t</td>
<td>moderated t-statistic (omitted for topTableF)</td>
</tr>
<tr>
<td>P.Value</td>
<td>Raw p-value</td>
</tr>
<tr>
<td>adj.P.Val</td>
<td>Adjusted p-value or q-value</td>
</tr>
<tr>
<td>B</td>
<td>log-odds that the gene is differentially expressed</td>
</tr>
</tbody>
</table>

**Objective 6 – Gene Category Testing I.** To gain mechanistic insight gene set analysis (GSA) will be conducted using lists of DETs from Objective 5. GSA analysis compares the observed frequency of genes in a set that
belong to discrete category, such as a Gene Ontology (GO) category, or KEGG pathway, against the frequency of genes belonging to that category in a gene set of identical size chosen at random from the array. The difference between the observed and the expected frequency forms the basis of the statistical test, expressed as a p-value estimating the likelihood that the frequency of genes in the discrete category is random. GSA analysis will be performed using the R/Bioconductor GOstats library (Falcon et al., 2007).

**Objective 7 – Gene Category Testing II.** To gain additional biological understanding Gene Set Enrichment Analysis (GSEA) will be conducted using the R-GSEA R script available from the Broad Institute (Subramanian et al., 2005). The goal of GSEA analysis is to determine whether members of a gene set S tend to occur toward the top or bottom of a list of ranked genes (e.g., expression dataset L), in which case the gene set is correlated with the phenotypic class distinction. Gene sets for analysis will be selected from the C2 “functional” collection in the Molecular Signatures Database (MSigDB http://www.broadinstitute.org/gsea/msigdb/index.jsp).

**Reference and Further Readings**


**Contact for more information**

- **John Bauman, Ph.D.**
  Scientific Account Manager
  jbauman@ils-inc.com
  Office: 919-281-1110 x750
  Cell: 919-259-3172
- **Kenneth Phillips**
  Senior Bioinformatics Specialist
  kphillips@ilsgenomics.com
  Office: 919-825-1032
  Cell: 919-632-9865

ILS Genomics © 2012