**2.1 Hypothesis Testing**

xxx

**2.2 The Multiple Testing Problem**

Typically, the goal of differential expression analysis is to identify genes whose expression levels show a statistically significant difference (in expression) between **two groups.**  This involves comparing the means of each group, and can be accomplished using a statistical test. In the case of microarrays, *t*-tests are frequently used. For each gene, the null hypothesis that “there is no difference in expression between the two sample groups” is tested.

Because a typical microarray study generates a gene expression matrix with tens of thousands of rows, thousands of hypothesis tests will need to be performed. It can be expected that a number of the tests will show significance just by chance. These represent false positives, or Type I errors. For example, suppose an expression matrix contains 1000 genes, and *t*-tests are performed on each gene using a significance level of 0.05. The significance level is the probability of a Type I error and indicates the likelihood that the experiment will result in a mistake (false positive). So, it can be expected that 50 genes (1000 x 0.05) will falsely show as differentially expressed when they actually occur by chance. It is possible to have more false positives than actual differentially expressed genes. This “multiple testing problem” is addressed by adjusting the *p*-values computed for each gene. Procedures for controlling the family-wise error rate (FWER) or the false discovery rate (FDR) [Benjamini and Hochberg, 1995] [Dudoit, *et al*, 2003] are commonly used for recomputing *p*-values.

**2.3 Gene Filtering**

Gene filtering can be used to reduce the number of hypothesis tests performed, thus reducing the impact of multiple testing correction and also eliminated uninformative genes. Filtering can be based on biological knowledge, as well as statistics. Statistical filtering considers gene expression levels and aims to remove genes whose measurements are uninformative, generally due to low intensity or low variability. Genes are removed that do not meet some minimum user-specified threshold.

Many genes are only expressed at certain times, under certain conditions, or in certain tissues. Furthermore, many will not show significant variation is expression across sample groups. Therefore, these can be removed before multiple testing to reduce its impact. Considerations for filtering genes commonly include intensity level and variability across sample groups.

2.3.1 Threshold/Ceiling.

* Before filtering begins, microarray analysis may apply preprocessing to remove platform noise by setting floor and ceiling values. Each expression value lower than the floor is raised to the floor value. Each expression value higher the ceiling is lowered to the ceiling value.

2.3.2 Intensity based filters remove genes with a low overall intensity.

* Filtering by ***average expression level***

Genes with an average express level below the threshold are removed.

* Filtering by ***maximal expression level***

Genes with all expression values below the threshold are eliminated.

* Filtering by ***coefficient of variation***

Coefficient of variationis defined as the standard deviation divided by the absolute value of the mean: *cv* = σ/|μ|.

* Filtering by ***minimum sample count***

Genes are retained only if there are sufficient samples with intensities above the threshold. For example, *genefilter*, a component of Bioconductor [Biocondcutor 2016], implements this with filter functions *kOverA* and *pOverA*.

2.3.3 Variability based filters remove genes with a low overall variability across samples, as they show little discriminatory power.

* Filtering by ***delta***

Genes with delta (*max*-*min*) less than the threshold are removed.

* Filtering by ***fold change***

Genes with fold change (*max*/*min*) less than the threshold are removed.

* Filtering by ***inter-quartile range (IRQ)***

Genes with IQR of expression values that fall below the threshold are removed. For example, in [Torgo 2010] genes with variability that was smaller than 1/5 of the global IQR were removed.

**References**

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*. Series B (Methodological) 57(1): 289–300.

**Bioconductor** [http://www.bioconductor.org]

Draghici S (2012). *Statistics and Data Analysis for Microarrays Using R and Bioconductor*. New York, NY: Chapman & Halll/CRC.

Dudoit S, Shaffer JP, Boldrick JC (2003). *Multiple hypothesis testing in microarray*

*experiments*. Stat Sci 71–103.

**Gene Expression Omnibus (GEO)** [http://www.ncbi.nlm.nih.gov/geo]

**R** [http://www.r-project.org]

Torgo L (2010). *Data Mining with R: Learning with Case Studies*. New York, NY: Chapman & Halll/CRC.