Differential Expression Analysis of Microarray Data

EMBO Practical Course

Analysis and Informatics of Microarray Data

October 2007

Denise Scholtens

Assistant Professor, Department of Preventive Medicine
Northwestern University Medical School
dscholtens@northwestern.edu

Simple question

- I have two sample types.
- Which genes represented on my microarray are *differentially expressed*?

- Assuming my experiments are done well…
  - `arrayQualityMetrics`

…and all uninteresting variation is accounted for…

- background correction, normalization (`rma, vsn, normexp`)

…what could possibly be so difficult?

Statistical Issues *NOT UNIQUE* to Microarray Data

- Scale of data
  - log transformation
- Test statistic
  - How do I find differences in expression?
- Statistical significance
  - How unusual are my observed data?

Statistical Issues *RELATIVELY UNIQUE* to Microarray Data (although often relevant in other settings)

- Multiplicity
  - Is the ability to test tens of thousands of genes simultaneously always helpful?
- Expense
  - Microarray experiments are fairly expensive, often resulting in small sample sizes.
- How do I interpret my results?
  - My ‘interesting gene list’ is really long…what do I do with it?

Synthesis

- How do we use the unique features of microarray data to address the more classic statistical problems?

Scale of data: logs

- Fold changes are often the preferred quantification of differential expression. Fold changes are essentially ratios.
- Notation for describing fold change is sometimes problematic: e.g. -2 mean 1/2, -3 means 1/3. Note that this would mean there are no values between -1 and 1.
- Ratios are not symmetric around 1 (the obvious ‘null’ value), making statistical operations difficult.

Scale of data: logs

- The intensity distribution of ratios has a fat right tail.
- Logs of ratios are symmetric around 0:
  - Average of 1/10 and 10 is about 5.
  - Average of log(1/10) and log(10) is 0.
  - Averaging ratios is in general a bad idea.

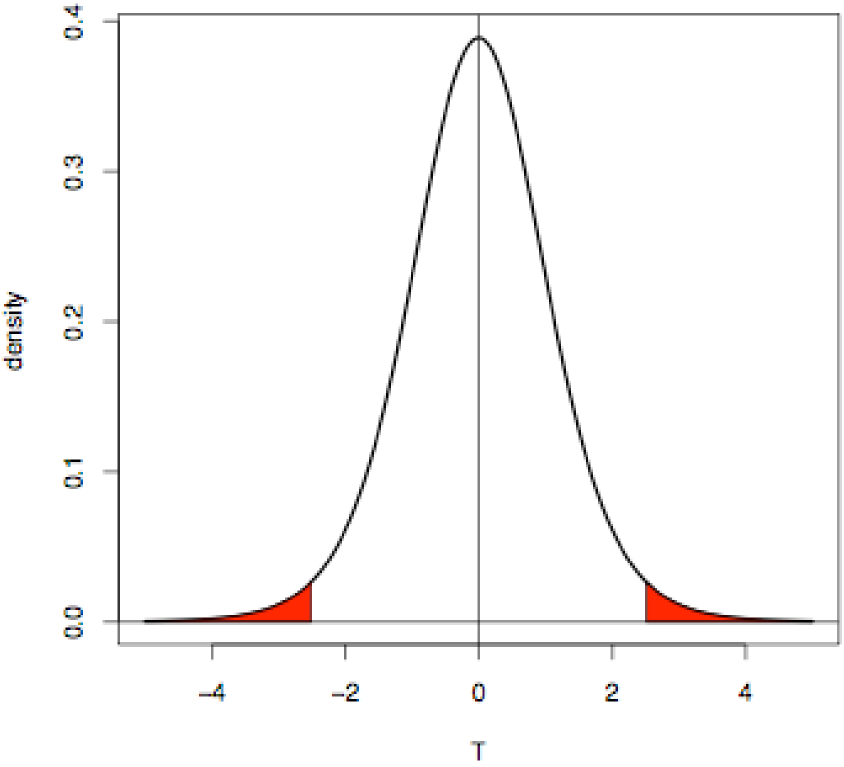
Statistical tests - example

- The two-sample *t*-statistic

$$T_g = \frac{\overline{X}_{g1} - \overline{X}_{g2}}{s_g \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

is used to test equality of the group means μ_1 and $\mu_{2.}$

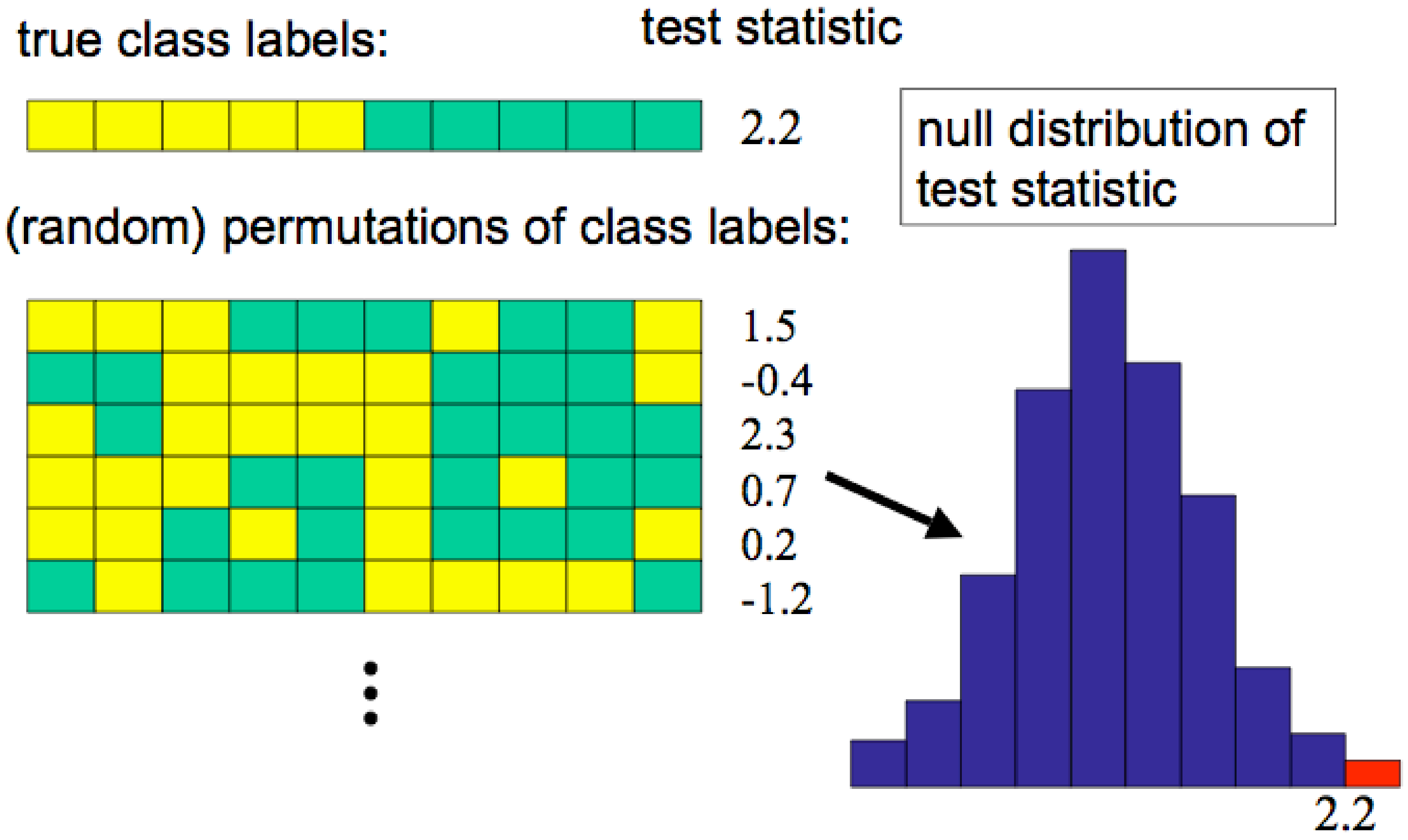
- The *p*-value p_g is the probability under the null hypothesis (here: $\mu_1 = \mu_2$) that the test statistic is at least as extreme as the observed value T_g.



Statistical tests - Variations on the theme

- Standard *t*-tests: assumes Normally distributed data in each class (almost always questionable), equal variances within classes
- Welch *t*-test: as above, but allows for unequal variances
- Wilcoxon test: non-parametric, rank-based
- Permutation test: estimate the distribution of the test statistic (e.g. the *t*-statistic) under the null hypothesis by permutation of the sample labels. The *p*-value p_g is given as the fraction of permutations yielding a test statistic that is at least as extreme as the observed one.
- Moderated *t*-statistic: the one that is often used for microarray data sets with small sample size (to be discussed in more detail)

Permutation tests



true class labels:

test statistic

null distribution of test statistic

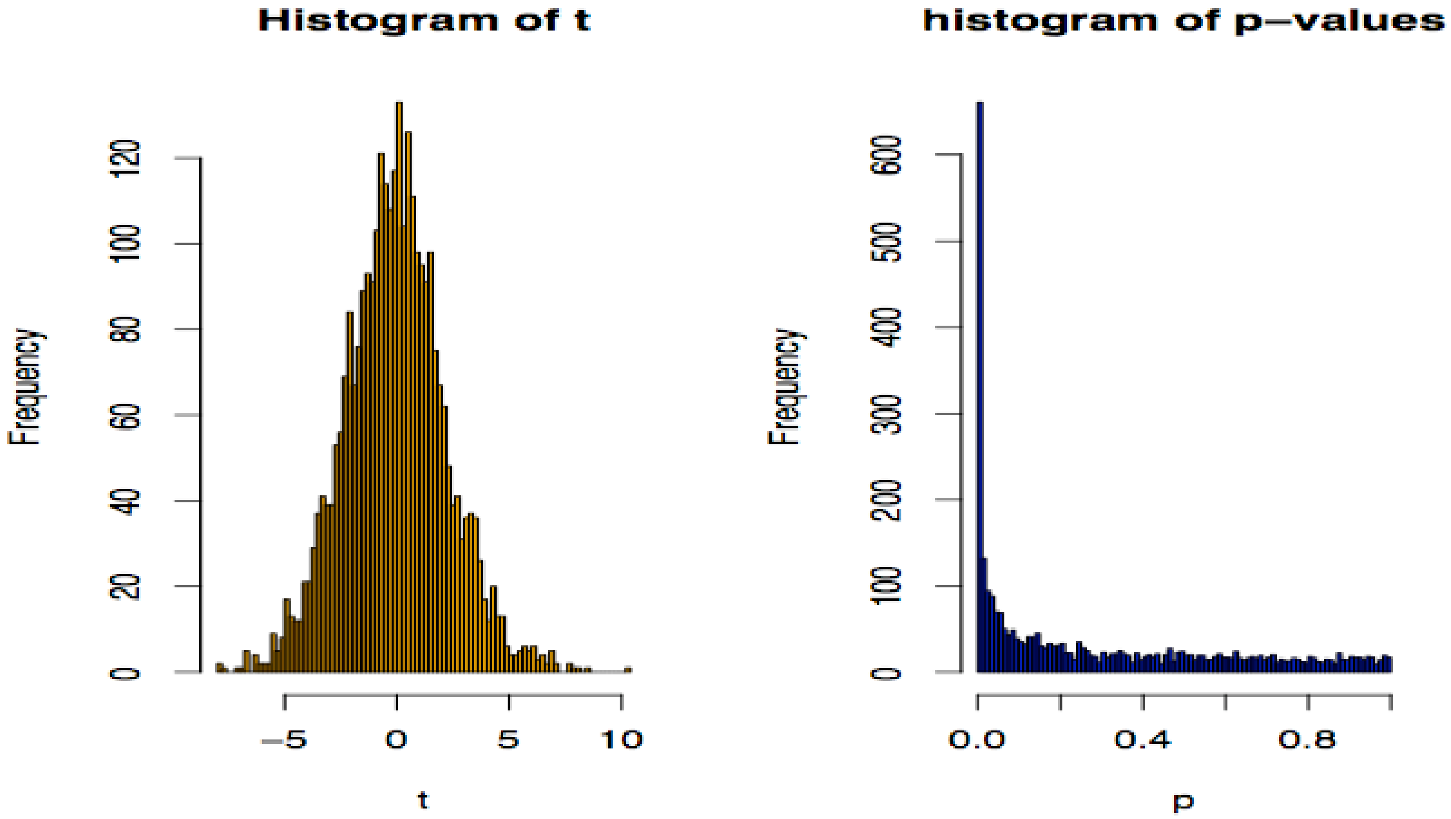
(random) permutations of class labels:

Statistical tests - Different settings

- Comparison of two classes (e.g. tumor vs. normal, treated vs. untreated cell line)
- Paired observations from two classes: e.g. the *t*-test for paired samples is based on the within-pair differences
- More than two classes and/or more than one categorical or continuous factor: linear models
  - Linear model framework encompasses two class problems described above

Example

Golub data, 27 ALL vs. 11 AML samples, 3,051 genes.



t-test: 1045 genes with $p < 0.05$.

Multiple testing: the problem

- Thousands of hypotheses are tested simultaneously.
- Increased chance of false positives.
- E.g. suppose you have 10,000 genes on a chip and not a single one is differentially expressed. You would expect 10000*0.01=100 of them to have a *p*-value < 0.01.
- Multiple testing methods help to account for this extra amount of ‘chance’ findings.

Multiple hypothesis testing

	# non–rejected hypotheses	# rejected hypotheses	
# true null hypotheses (non-diff. genes)	U	V Type I error	m_0
# false null hypotheses (diff. genes)	T Type II error	S	m_1
	$m - R$	R	m

From Benjamini & Hochberg (1995).

Controlling Type I Error Rates

- Family-wise error rate (FWER)
  - *FWER* is defined as the probability of at least one Type I error (false positive) among the genes selected as significant.

$$FWER = \Pr(V > 0)$$

- False discovery rate (FDR)
  - *FDR* is defined as the expected proportion of Type I errors (false positives) among the rejected hypotheses.

$$FDR = E(Q) \text{ with } Q = \begin{cases} V/R, & \text{if } R > 0, \\ 0, & \text{if } R = 0. \end{cases}$$

FWER: The Bonferroni Correction

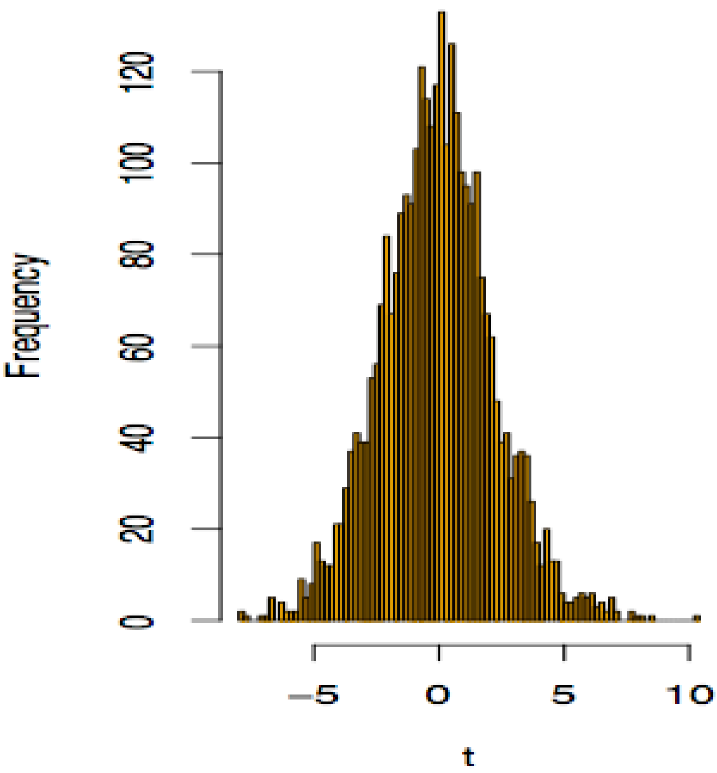
- Suppose we conduct a hypothesis test for each gene $g=1,\ldots,m$, producing an observed test statistic T_g and an unadjusted *p*-value p_g.
- Bonferroni adjusted *p*-values:

$$\breve{p}_g = \min(mp_g, 1).$$

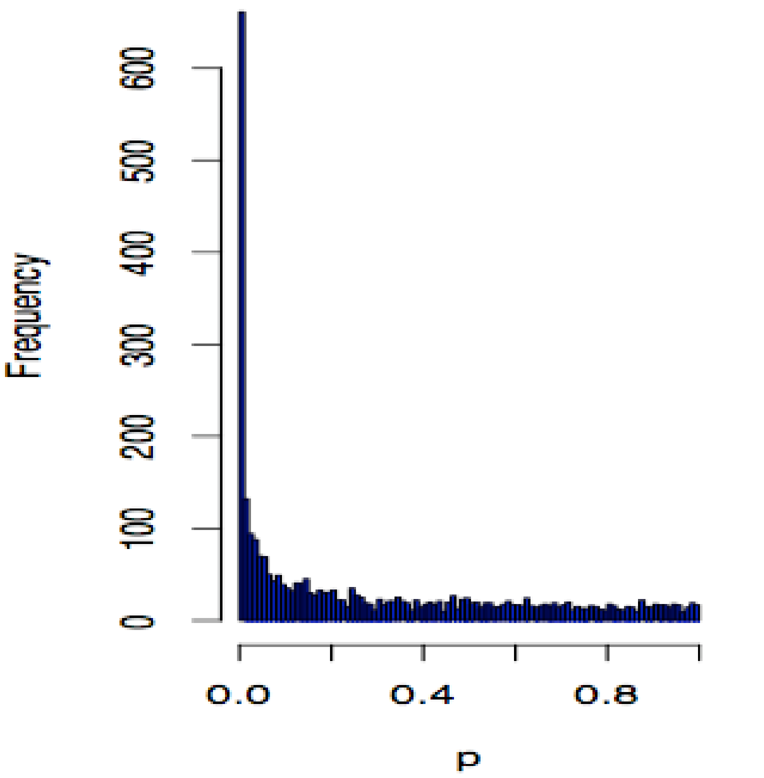
- Selecting all genes with $\breve{p}_g \le \alpha$ controls the FWER at level α, i.e. $Pr(V>0) \le \alpha$.

Example: Bonferroni correction

Golub data, 27 ALL vs. 11 AML samples, 3,051 genes.



Histogram of t



histogram of p−values

98 genes with Bonferroni-adjusted $\tilde{p}_g < 0.05 \Leftrightarrow p_g < 0.000016$

FWER: Alternatives to Bonferroni

- There are alternative methods for FWER *p*-value adjustment which can be more powerful.
- The permutation-based Westfall-Young method takes the correlation between genes into account and is typically more powerful for microarray data.
- The Bioconductor package `multtest` facilitates many approaches to multiple testing correction.

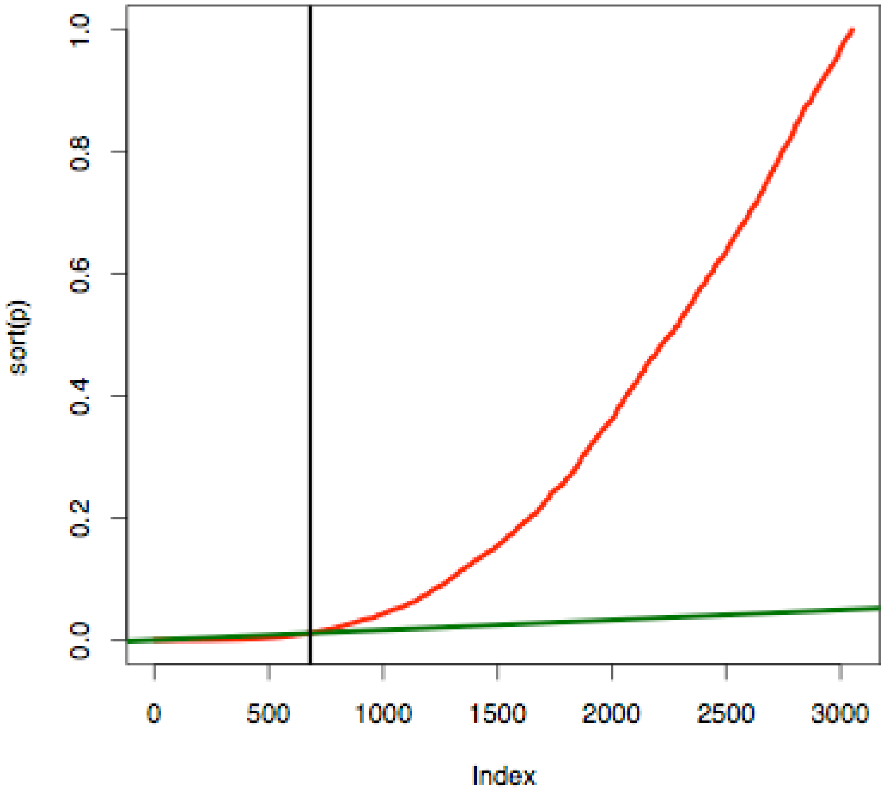
FDR: Benjamini-Hochberg

- FDR: the expected proportion of false positives among the significant genes.
- Ordered unadjusted *p*-values: $p_{r1} \leq p_{r2} \leq \ldots \leq p_{rm}$.
- To control *FDR = E(V/R)* at level α, let

$$j^* = \max\{j : p_{rj} \leq (j/m)\alpha\}.$$

  Reject the hypotheses H_{rj} for $j=1,\ldots j^*$.
- Is valid for independent test statistics and for some types of dependence. Tends to be conservative if many genes are differentially expressed. Implemented in `multtest`.

FDR: Benjamini-Hochberg



Golub data: 681 genes with BH–adjusted $p < 0.05$.

FWER or FDR?

- Choose control of the FWER if high confidence in all selected genes is desired. Loss of power due to large number of tests: many differentially expressed genes may not appear significant.
- If a certain proportion of false positives is tolerable, then procedures based on FDR are more flexible. The researcher can decide how many genes to select based on practical considerations.

Focusing analyses

- More is not always better!
- Suppose you use a focused array with 500 genes you are particularly interested in.
- If a gene on this array has an unadjusted *p*-value of 0.0001, the Bonferroni-adjusted *p*-value is still 0.05.
- If instead you use a genome-wide array with 50,000 genes, this gene would be much harder to detect. Roughly 5 genes can be expected to have such a low *p*-value simply by chance.
- Therefore, it may be worthwhile to focus on genes of particular biological interest from the beginning.

Pre-filtering

- What about pre-filtering genes according to criteria not specific to the experiment to reduce the proportion of false positives?
- This can be useful since genes with low intensities in most of the samples or low variance across the samples are less likely to be interesting.
- In order to maintain control of the Type I error, the criteria must be independent of the distribution of the test statistic under the null hypothesis.

Pre-filtering

- Common filters:
  - Low intensity across all (or most) samples
  - Low variance/IQR across samples

- The Bioconductor package `genefilter` can be used for pre-filtering.

Few replicates: moderated *t*-statistics

- With the *t*-test, we estimate the variance of each gene individually. When there are only a few replicates (say 2-5 per group), the variance estimates are unstable.
- The Bioconductor packages `limma` and `siggenes` offer moderated *t*-statistics as an aid for this problem.

`limma`:
Linear Models for Microarray Analysis

- Highly used Bioconductor package for microarray data analysis
- Handles data import, some QA, background correction, normalization, linear modeling, multiple testing correction, sorting and display of results
- In particular, applies linear models to microarray data.
  - Linear models encompass the two-sample problem we have discussed to this point.

Why `limma`? Statistical reasons

- While `limma` provides convenient handling and linear modeling capabilities for microarray data, linear model parameters can be estimated using all standard statistical software.
- The statistical novelty and power for `limma` are harnessed in the `eBayes()` function.
- In particular, `eBayes()` provides moderated *t*-statistics and resultant corrected *p*-values.

Linear models

- $y_j = \mu_j + \beta_{1j}x_1 + \beta_{2j}x_2 + \ldots + \beta_{kj}x_k$
- x's are covariates
- β_j's are measures of the effect of the covariate for gene *j*
- Often covariates represent treatments applied to cell lines or samples from individuals with different disease types
- Must specify a *design matrix* and a *contrast matrix*
  - *Design matrix* indicates which samples have been applied to each array
  - *Contrast matrix* specifies which comparisons you would like to make between the samples

Ordinary *t*-statistics

- Assume a simple model with only one covariate of interest $y_j = \mu_j + \beta_j x$
- Then the ordinary *t*-statistic to evaluate differential expression for gene *j* is

$$t_j = \bar{\beta}_j / (u_j s_j)$$

where $\bar{\beta}_j$ is the estimated coefficient in the linear model for the *j*th gene, u_j is the unscaled standard deviation and s_j^2 is the sample residual variance.

- The *p*-value is then calculated according to a Student's *t* distribution with f_j degrees of freedom.

`eBayes()`: Empirical Bayes variance adjustment

- General Bayesian paradigm:
  - Bayesian statistical analyses begin with ‘prior’ distributions describing beliefs about the values of parameters in statistical models prior to analysis of the data at hand
  - Bayesian analyses require specification of these parameters
  - So called ‘Empirical Bayes’ methods use the data at hand to guide prior parameter specification
  - Then given the data, these prior distributions are updated to give posterior results

`eBayes():` Empirical Bayes variance adjustment

- Instead of usual *t*-statistics comparing two sample types, `limma` returns moderated *t*-statistics
- The interpretation of the usual and moderated statistics is the same, except the standard errors for the moderated statistics are shrunk toward a common value
- Moderated *t*-statistics lead to *p*-values, but the degrees of freedom increase reflecting the strength in borrowing information across genes

`eBayes()`: Empirical Bayes variance adjustment

- Assume an inverse Chi-square prior for the true gene-specific residual variances with mean s_0^2 and degrees of freedom f_0.
- Then the posterior residual variances are given by

$$\breve{s}_j^2 = \frac{f_0 s_0^2 + f_j s_j^2}{f_0 + f_j}$$

`eBayes()`: Empirical Bayes variance adjustment

- The moderated *t*-statistic is then

$$t_j = \overline{\beta}_j / (u_j \breve{s}_j)$$

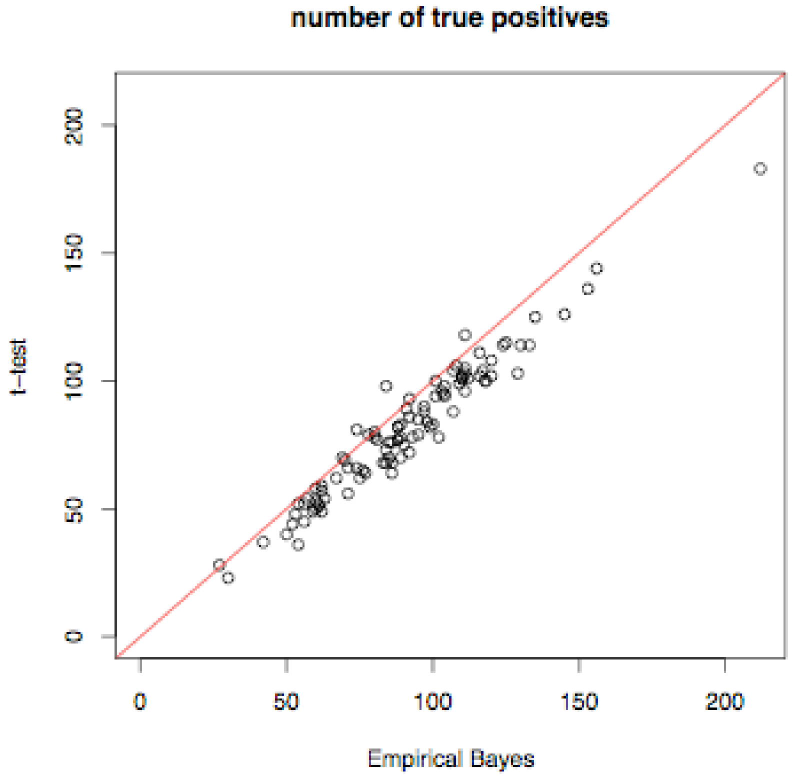
which follows a *t* distribution with f_0+f_j degrees of freedom under the null hypothesis.

`eBayes()`:
Empirical Bayes variance adjustment

- Summarize please?
- In a signal-to-noise ratio paradigm, we are all familiar with the idea of not wanting to attribute mistaken biology to signals that appear large only by random chance
- A misleadlingly small estimate of the variance will cause the same problem, and the empirical Bayes adjustment helps address this problem.
- Also, degrees of freedom (and therefore power for statistical inference) increase by harnessing information across all genes.
- All of these contribute to effective identification of differentially expressed genes, particularly when sample sizes are small.

Moderated *t*-test

Repeatedly draw 4 ALL and 4 AML samples out of the total 38 samples and apply the usual and moderated t–test (Bioconductor package limma) to them. Using a cut–off of $p < 0.05$, “true positives” are defined on the basis of the analysis of the whole data set (681 genes with FDR < 0.05).



Summary

- Classic statistical concerns such as suitable scale of data for analysis, appropriate test statistics, and statistical significance are all relevant.
- Additionally, the multiplicity of genes and the expense of microarray data often leading to small sample sizes must be accounted for.

Summary

- Log transforming data improves suitability of data for linear model analysis.
- Pre-filtering and multiple testing methods help address problems in simultaneously examining thousands of genes.
- Moderated *t*-statistics are helpful when sample sizes are small.

Next lecture and labs

- Practical steps to using `limma` and other Bioconductor packages
- A few options for what to do with the resultant gene lists

Slides largely adapted from

- Wolfgang Huber
- Anja von Heydebreck

- Sandrine Dudoit
- Axel Benner
- Rafael Irrizary

References

- Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JRSSB* 57:289-200.
- Dudoit, S. Shaffer, J.P., Boldrick, J.C. (2003) Multiple hypothesis testing in microarray experiments. *Stat Sci* 18:71-103.
- Smyth, G.K. (2004) Linear Models and Empirical Bayes Methods for Assessing Differential Expresion in Microarray Experiments, *SAGMB*, Vol.3, Article 3.
- Smyth, G.K. (2005) limma: Linear Models for Microarray Data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* Eds. Gentleman, R., et al. Springer.
- Storey, J.D. and Tibshirani, R. (2003) SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In: *The analysis of gene expression data: methods and software*. Eds. Parmigiani, G., et al. Springer.