Differential expression

Introductory Bioconductor Workshop

Fred Hutchinson Cancer Research Center

27 April 2009

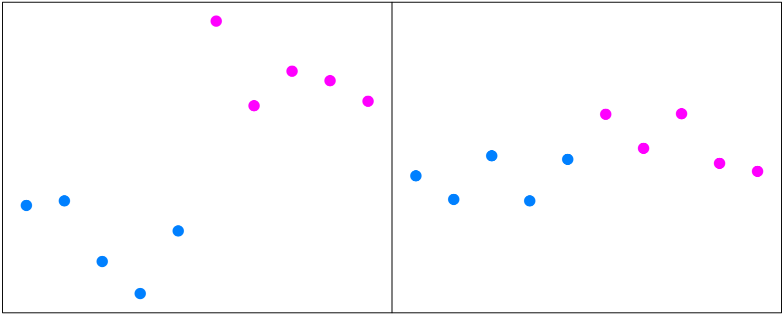
- Goal: find statistically significant associations of biological conditions or phenotypes with gene expression.
- Consider the two class problem.
- Data: n points in a p-dimensional space.
- $n \approx 10 - 100, p \approx 5000 - 30000$

A	A	A	A	A	B	B	B	B	B
$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	$x_{1,5}$	$x_{1,6}$	$x_{1,7}$	$x_{1,8}$	$x_{1,9}$	$x_{1,10}$
$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	$x_{2,5}$	$x_{2,6}$	$x_{2,7}$	$x_{2,8}$	$x_{2,9}$	$x_{2,10}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$x_{p,1}$	$x_{p,2}$	$x_{p,3}$	$x_{p,4}$	$x_{p,5}$	$x_{p,6}$	$x_{p,7}$	$x_{p,8}$	$x_{p,9}$	$x_{p,10}$

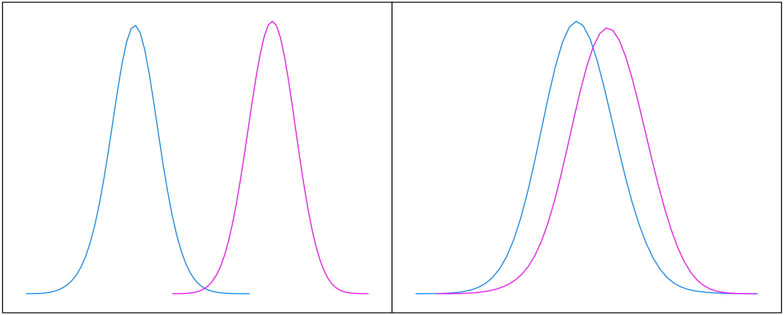
$p >> n$

- Problem: There are infinitely many ways to separate the space into two regions by a hyperplane such that the two groups are perfectly separated.
- This is a simple geometrical fact and holds as long as $n < p$!
- Answer: regularization. Rather than searching in the huge space of all hyperplanes in p-dimensional space, restrict ourselves to a smaller and biologically meaningful space.
- Two major approaches:
  - only hyperplanes perpendicular to the p coordinate axes (gene-by-gene discrimination, geneby-gene hypothesis testing)
  - any other reasonable, not too complex set of hypersurfaces (machine learning)

- Goal: find statistically significant associations of biological conditions or phenotypes with gene expression.
- The gene-by-gene approach:



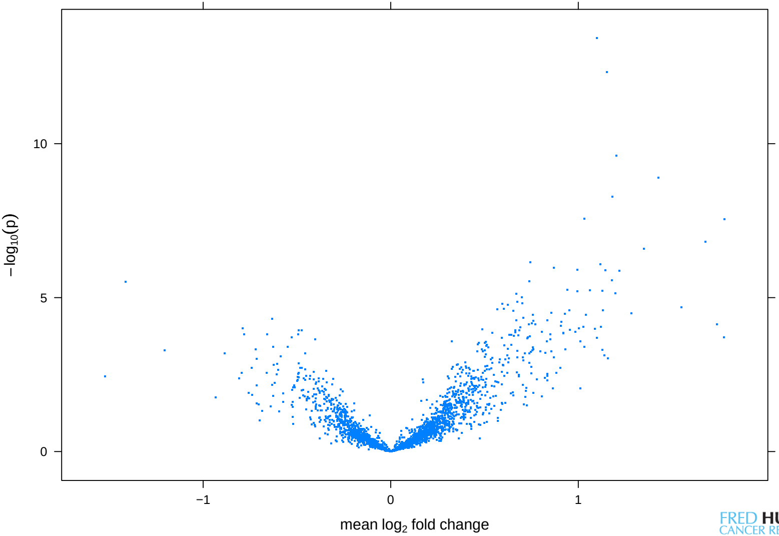
- Goal: find statistically significant associations of biological conditions or phenotypes with gene expression.
- The gene-by-gene approach:



Fold change vs *p*-value

- Two basic selection strategies are widely used
- Fold change (effect size):
  - Genes are deemed to be interesting if the effect size is large
  - For two sample comparisons we often call this the fold-change
  - Often values like 1.5 or 2.0 are used
- *p*-value:
  - Genes are deemed to be interesting if the *p*-value is small

Fold change vs *p*-value: Volcano plot



Modeling Considerations

- Parametric assumptions hard to justify with few arrays
- Nonparametric assumption:
  - Permutation tests or similar non-parametric tools are tempting
  - Such assumptions reduce power and hence ability to discriminate
  - With not much data (samples), a model is needed to help make inference
- A useful strategy is to aggregate information across genes

Gene by gene tests

- Examples:
    - t-test
    - Wilcoxon
    - F-test / more complex linear models
    - Cox regression
- Treating each gene independently of each other wastes information
- Many properties may be shared among genes; e.g., their within-group variability

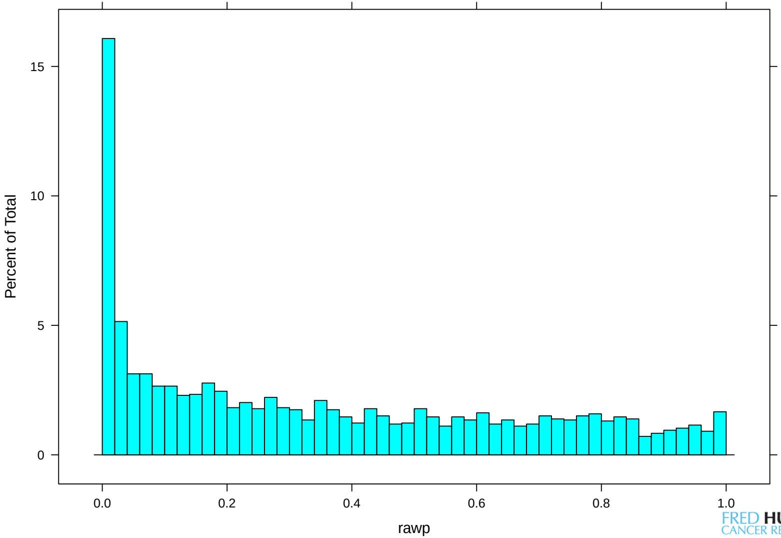
t-test

- Test for differences in means between two groups given the variability within each group

$$\frac{\bar{X}_1 - \bar{X}_2}{SE(\bar{X}_1 - \bar{X}_2)}$$

difference between group means / variability of groups

Distribution of *p*-values



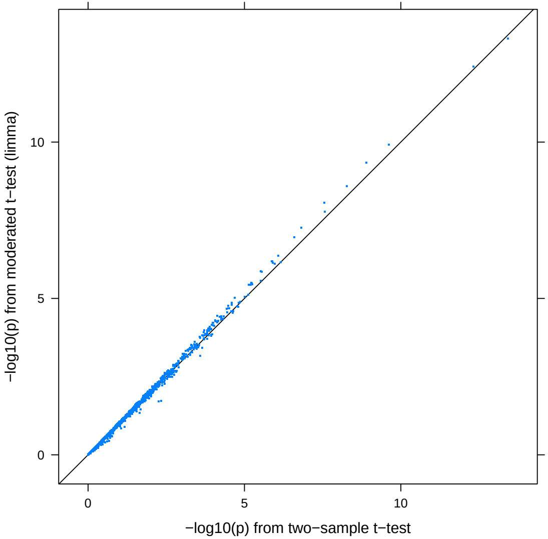
Moderated / Bayesian *t*-tests

- Rather than estimating within-group variability (denominator of t-test) over and over again for each gene, pool the information from many similar genes
  - Baldi, Long 2001 Tusher et al. (SAM) 2001
  - Lönnstedt and Speed 2002
  - Kendziorski et al. (Ebarrays) 2003
  - Smyth (limma) 2004
- Advantages:
  - eliminate occurrence of accidentally large *t*-statistics due to accidentally small within-group variance
  - effectively introduce a “fold-change” criterion

Moderated / Bayesian t-tests

- Typical approach
  - An overall estimate of the variance, s_0^2, is computed
  - then for each gene, an estimate of the per gene variance, s_g^2, is computed
  - the variance used is a weighted average of s_0^2 and s_g^2
  - the actual method of estimating the overall variance and the method of averaging is slightly different in different contexts

Moderated / Bayesian *t*-tests

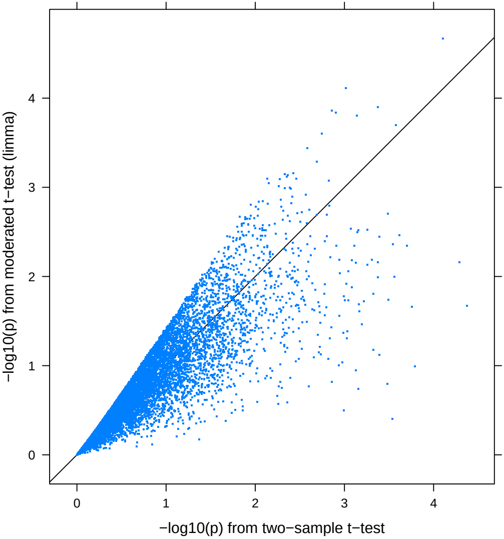


FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE

Moderated / Bayesian *t*-tests

- In this example with 79 samples, there is no big difference between ordinary and the moderated t-statistic.
- But for smaller data sets the differences will be larger.

Moderated / Bayesian *t*-tests



p-value corrections

- Problem: we perform a large number of tests and the resulting *p*-values are difficult to interpret
- Band-aid: statisticians have turned *p*-value corrections into an industry, but they are really more of a band-aid than a solution
- Solution: test fewer, more directed hypotheses. We still need to correct, but the amount of correction needed will be much smaller

p-value corrections

- Methodology: there are now more methods than we could ever consider
- Basic idea: reduce the critical value used to reject
  - since truly false hypotheses tend to have smaller *p*-values, this adjustment enriches those rejected for those that are truly false
  - but among the casualties are those hypotheses that are truly false, but which did not obtain an extraordinarily small *p*-value
- Trade-off between sensitivity and specificity

p-value corrections

- The `multtest` package (by K. Pollard, Y. Ge and S. Dudoit) provides a wide variety of *p*-value correction methods
  - provides a variety of *t*- and *F*-tests, including robust versions of each test
  - Single-step and step-down minP and maxT methods can be used to control the chosen type I error rate
  - criteria for error rate control include FWER, gFWER, FDR
- Check the vignette and other package documentation for more deatils

FWER

Family wise error rate: Probability of at least one false positive.

```
> sum(resT$rawp < 0.05)

[1] 577

> sum(resT$adjp < 0.05)

[1] 34
```

This is a large loss of power!

FDR

False Discovery Rate:

$$E\left(\frac{FP}{FP+TP}\right)$$

```
> res <- mt.rawp2adjp(rawp, proc = "BH")
> sum(res$adjp[, "BH"] < 0.05)

[1] 209
```

Data Reduction

- Typically, most genes do not show differences in expression across arrays
- Should consider a reduction in the set of gene/probes that are under consideration:
  - not all genes are expressed in all tissues
  - one of the basic assumptions of normalization is that most of the genes have not changed expression levels across conditions
  - these observations argue in favor of reducing the set of genes
- We recommend using some form of non-specific filtering

Filtering on variability

- The expression estimate itself does not reflect mRNA abundance
- Only within-gene, between-array comparisons are valid
- Filtering on absolute expression values (e.g., removing those below 100) is falling into that same trap: absolute numbers do not tell us about the true mRNA abundance
- We recommend filtering genes by some measure of the variability (MAD, IQR, etc) across arrays
- genes that show no variation across the conditions measured are not interesting

Discrimination scores - ROC curve analysis

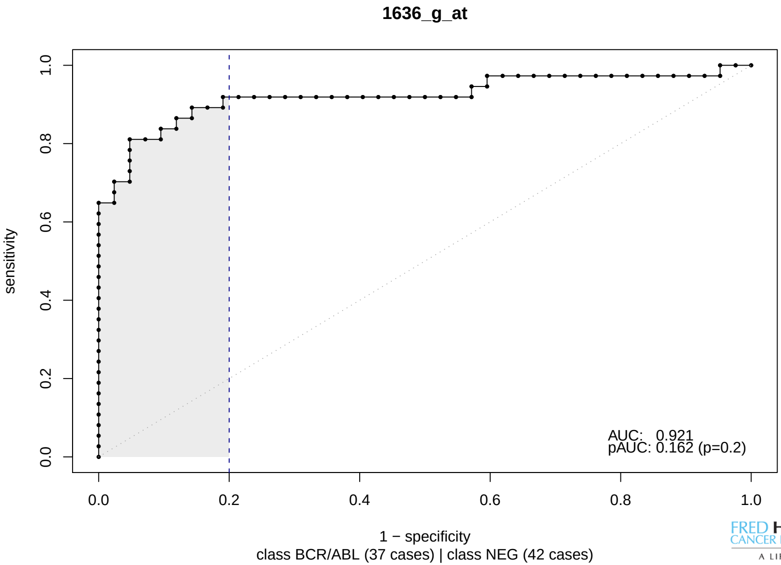
- Classification based approach (Pepe et al, 2003)
- Find potential marker genes

  Gene expression should discriminate between groups

ROC curve

- Gene g, two groups (A and B)
- For any cutoff θ
  - classify sample i to group B if $x_{g,i} \geq \theta$
  - Specificity: proportion of true positives
  - Sensitivity: proportion of true negatives
- ROC curve: plot of Sensitivity *vs* 1 - Specificity

ROC curve



1636_g_at
sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1 − specificity
class BCR/ABL (37 cases) | class NEG (42 cases)
AUC: 0.921
pAUC: 0.162 (p=0.2)

Labs from Bioconductor Case Studies

- Chapter 1: The ALL Data Set
- Chapter 6: Easy Differential Expression
- Chapter 7: Differential Expression