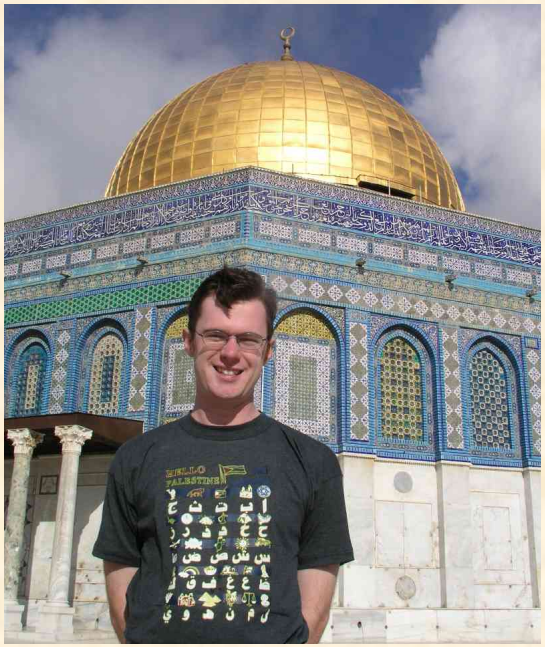
Sixth International Conference on
Intelligent Data Engineering and Automated Learning
IDEAL'05

Application of Mixture Models to Detect Differentially Expressed Genes

Liat Ben-Tovim Jones[1], Richard Bean[1], Geoff McLachlan[1,2,3] and Justin Zhu[1]

[1]ARC Centre in Bioinformatics, Institute for Molecular Bioscience, UQ
[2]Department of Mathematics, University of Queensland
[3]ARC Special Research Centre for Functional and Applied Genomics, UQ



Richard Bean

Justin Zhu

Outline of Talk

- The problem of detecting DE genes
- Multiple Hypothesis Testing and FDR
- Mixture Model Approach
- The Hedenfalk Breast Cancer Data
- Conclusions

The Challenge for Statistical Analysis of Microarray Data

Microarrays present new problems for statistics because the data is very high dimensional with very little replication.

The challenge is to extract useful information and discover knowledge from the data, such as gene functions, gene interactions, regulatory pathways, metabolic pathways etc.

Gene Expression Studies

- Pattern of genes expressed in a cell is characteristic of its current state

- Virtually all differences in cell state or type are correlated with changes in mRNA levels of many genes

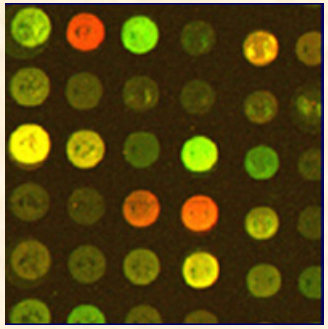
- DNA microarray technology originally conceived in order to detect expression of thousands of genes simultaneously

The Microarray Experiment

- DNA complementary to the genes of interest is generated and laid out in microscopic quantities on solid surfaces at defined positions

- DNA (mRNA) from samples is eluted over the surface – complementary DNA binds

- Presence of bound DNA is detected by fluorescence following laser excitation

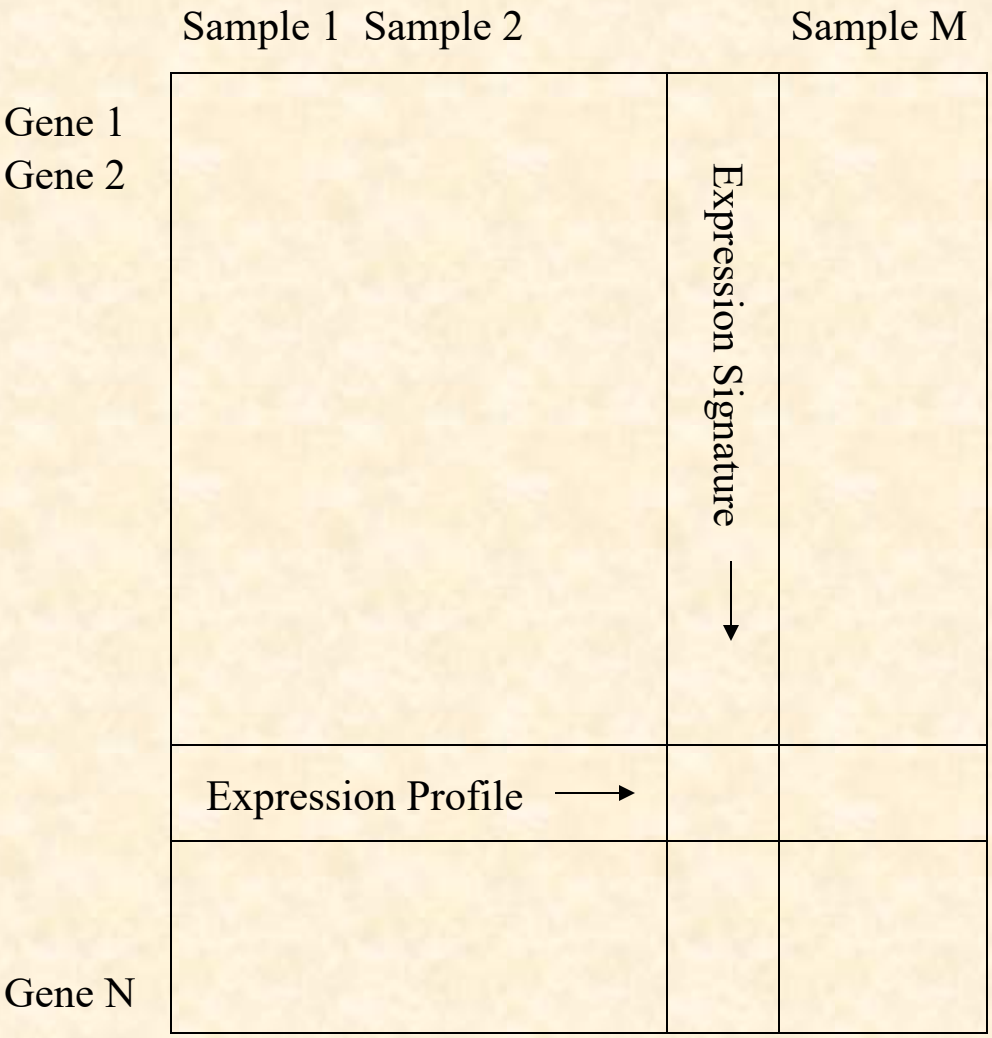


DE genes

- A major goal is to find differentially expressed (DE) genes in a given number of tissue classes, e.g. genes over- or under-expressed in tumour vs. normal, or in different subtypes of cancer.

- These genes may be useful for making new biological discoveries, or form part of a diagnostic kit in medicine (marker genes).

Microarray Data represented as *N* x *M* Matrix



M columns (samples) ~ 10^2

N rows (genes) ~ 10^4

Finding DE genes in known tissue classes

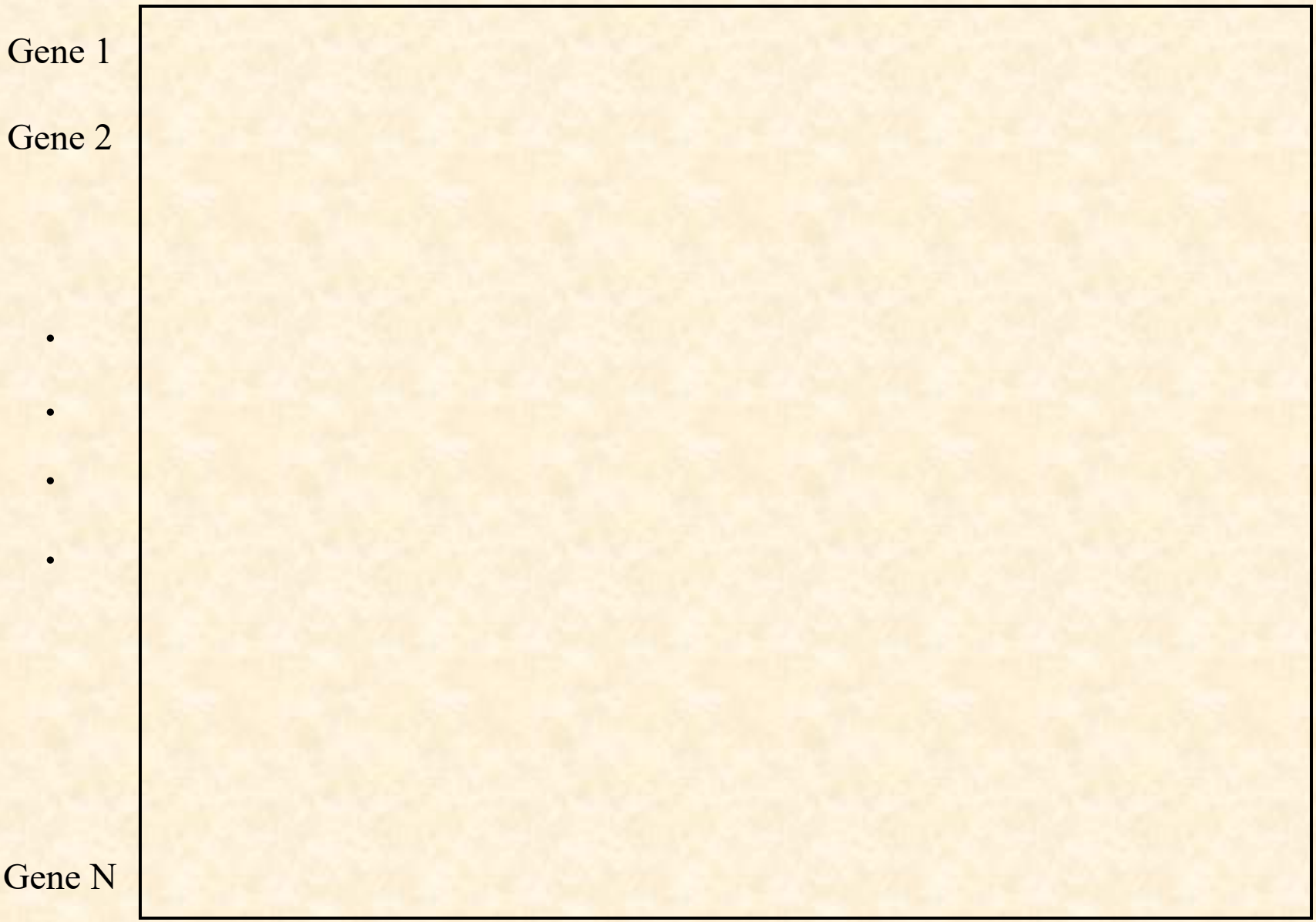
- Classify M samples with respect to g classes on the basis of the N gene expressions

- Assume that there are n_i tissue samples from each class C_i (i = 1, …, g), where

$$M = n_1 + \ldots + n_g.$$

- Take a simple case where g = 2

- The aim is to detect whether some of the thousands of genes have different expression levels in class C_1 than in class C_2.


Sample 1
Sample 2
Sample M
Gene 1
Gene 2
Gene N


Sample 1
Sample 2
Sample M
Gene 1
Gene 2
Gene N
Class 1
Class 2

Fold change is the simplest method

Calculate the log ratio of the average expression between the two classes and consider all genes that differ by more than an arbitrary cutoff value to be differentially expressed. A two-fold difference is often chosen.

Fold change is not a statistical test.

Test of a Single Hypothesis

For gene *j*, let $H_j = 0$ denote that the null hypothesis of no association between its expression level and its class membership holds, where ($j = 1, \ldots, N$).

$H_j = 0$ Null hypothesis for the j^{th} gene holds.
$H_j = 1$ Null hypothesis for the j^{th} gene does not hold.

	Retain Null	Reject Null
$H_j = 0$	√	type I error
$H_j = 1$	type II error	√

Gene Statistics: Two-Sample *t*-Statistic

Student's *t*-statistic

$$T_j = \frac{\bar{\mathbf{y}}_{1j} - \bar{\mathbf{y}}_{2j}}{\sqrt{s_{1j}^2/n_1 + s_{2j}^2/n_2}}$$

Pooled form of the Student's *t*-statistic, assumed common variance in the two classes

$$T_j = \frac{\bar{\mathbf{y}}_{1j} - \bar{\mathbf{y}}_{2j}}{s_j\sqrt{1/n_1 + 1/n_2}}$$

Modified *t*-statistic of Tusher et al. (2001)

$$T_j = \frac{\bar{\mathbf{y}}_{1j} - \bar{\mathbf{y}}_{2j}}{s_j\sqrt{1/n_1 + 1/n_2} + a_0}$$

Multiplicity Problem

Perform a test for each gene to determine the statistical significance of differential expression for that gene.

Problem: When many hypotheses are tested, the probability of a type I error (false positive) increases sharply with the number of hypotheses.

Further: Genes are co-regulated, subsequently there is correlation between the test statistics.

Methods for dealing with the Multiplicity Problem

- The Bonferroni Method

controls the family wise error rate (FWER) i.e.
the probability that at least one false positive error will be made

  - Too strict for gene expression data, tries to make it unlikely that even one false rejection of the null is made, may lead to missed findings

- The False Discovery Rate (FDR)

emphasizes the proportion of false positives among the identified differentially expressed genes.

  - Good for gene expression data – says something about the chosen genes

False Discovery Rate

Benjamini and Hochberg (1995)

$$\text{FDR} \approx \frac{\#(\text{false positives})}{\#(\text{rejected hypotheses})}$$

The FDR is essentially the expectation of the proportion of false positives among the identified differentially expressed genes.

Two-component mixture model

$$f(w_j) = \pi_0 f_0(w_j) + \pi_1 f_1(w_j)$$

π_0 is the proportion of genes that are not differentially expressed, and $\pi_1 = 1 - \pi_0$ is the proportion that are.

Efron et al. (2001)

Two-component mixture model

$$f(w_j) = \pi_0 f_0(w_j) + \pi_1 f_1(w_j)$$

π_0 is the proportion of genes that are not differentially expressed, and $\pi_1 = 1 - \pi_0$ is the proportion that are.

Then

$$\tau_0(w_j) = \frac{\pi_0 f_0(w_j)}{f(w_j)}$$

is the posterior probability that gene j is not differentially expressed.

Procedure

1. Form a statistic w_j for each gene. A large positive value of w_j corresponds to a gene that is DE.

2. Fit to $w_1,...,w_N$ a mixture of two normal densities with the 1st component as a standard normal - genes that are not DE. Assume that w_j have been transformed so that they are approx. normally distributed e.g for the ANOVA statistic F (Broet et al., 2004).

$$W_j = \frac{\left(1-\frac{2}{9(M-g)}\right)F_j^{\frac{1}{3}} - \left(1-\frac{2}{9(g-1)}\right)}{\sqrt{\frac{2}{9(M-g)}F^{\frac{2}{3}} + \frac{2}{9(g-1)}}}$$

3. Let $\hat{\tau}_0(w_j)$ denote the (estimated) posterior probability that gene j belongs to the first component of the mixture.

If we conclude that gene j is differentially expressed if:

$$\hat{\tau}_0(w_j) \leq c_0,$$

then this decision minimizes the (estimated) Bayes risk

$$\widehat{\text{Risk}} = (1 - c_0)\hat{\omega}\widehat{\text{FDR}} + c_0(1 - \hat{\omega})\widehat{\text{FNR}}$$

where

$$\hat{\omega} = N_r/N$$

Estimated FDR

$$\widehat{\mathrm{FDR}} = \sum_{j=1}^{N} \hat{\tau}_0(w_j)\, I_{[0,c_o]}(\hat{\tau}_0(w_j))/N_r$$

where

$$N_r = \sum_{j=1}^{N} I_{[0,c_o]}(\hat{\tau}_0(w_j))$$

Hedenfalk Breast Cancer Data

Hedenfalk *et al.* (2001) used cDNA arrays to obtain gene expression profiles of tumours from carriers of either the BRCA1 or BRCA2 mutation (hereditary breast cancers), as well as sporadic breast cancer.

We consider their data set of $M = 15$ patients, comprising two patient groups: BRCA1 (7) versus BRCA2 - mutation positive (8), with $N = 3{,}226$ genes.

The problem is to find genes which are differentially expressed between the BRCA1 and BRCA2 patients.

Hedenfalk et al. (2001) NEJM, **344**, 539-547

Two component model for the Breast Cancer Data

Fit $\pi_0 N(0,1) + \pi_1 N(\mu_1, \sigma_1^2)$

to the *N* values of w_j (based on pooled two-sample *t*-statistic)

j th gene is taken to be differentially expressed if:

$$\hat{\tau}_0(w_j) \leq c_0$$

Estimated FDR for various levels of c_0

c_0	N_r	$\hat{FDR}$
0.5	1702	0.29
0.4	1235	0.23
0.3	850	0.18
0.2	483	0.12
0.1	175	0.06

Significant Genes (Hedenfalk Breast Cancer Data)

- 175 genes selected as significant by us

- 137 of these over-expressed in BRCA-1 relative to BRCA-2, including MSH2 (DNA repair), PDCD5 (apoptosis)

Compare Storey and Tibshirani (160 genes) and Hedenfalk (176 genes), gives 23 genes unique to our set.

Storey and Tibshirani (2003) PNAS, **100**, 9440-9445

Uniquely Identified Genes

Gene Identifier	Functional Class
ITPK1, NALP1, GADD34	Cell death
MAPK6	Cell cycle
GATA3, TLE1, HDAC2, GTF2B	Transcription
ANXA1	Cell-to-cell signalling
COL5A1, ACTB1	Cell growth/adhesion/motility
EIF2S2	Protein synthesis
PRKACA, CSTB	Protein modification
OXCT1, POX1	Metabolism

SAM (v. 2) Method for finding DE genes

- 210 genes selected as significant with an FDR of 5%
- Compare to our 174 genes, 152 common genes
- Compare to 160 (Storey and Tibshirani), 132 common

SAM method of Tusher et al. (2001) PNAS, **98**, 5116-5121

Conclusions

- Mixture-model based approach to finding DE genes can yield new information

- Gives a measure of the posterior probability that a gene is not DE (i.e. a local FDR rather than global)

- Can be used in the spirit of the q-value, to bound the FDR at a desired level

Extra Slides

Possible Outcomes for *N* Hypothesis Tests

	Accept Null	Reject Null	Total
Null True	N_{00}	N_{01}	N_0
Non-True	N_{10}	N_{11}	N_1
Total	$N - N_r$	N_r	N

FWER is the probability of getting one or more false positives out of all the hypotheses tested:

$$\text{FWER} = \text{pr}\{N_{01} \geq 1\}$$

Bonferroni method for controlling the FWER

Consider N hypothesis tests:

$$H_{0j} \text{ versus } H_{1j}, \qquad j = 1, \ldots, N$$

and let $P_1, \ldots, P_N$ denote the N P-values for these tests.

The Bonferroni Method:

Given P-values $P_1, \ldots, P_N$ reject null hypothesis H_{0j} if

$$P_j < \alpha / N .$$

Possible Outcomes for *N* Hypothesis Tests

	Accept Null	Reject Null	Total
Null True	N_{00}	N_{01}	N_0
Non-True	N_{10}	N_{11}	N_1
Total	$N - N_r$	N_r	N

$$FDR \approx \frac{N_{01}}{N_r}$$

Positive FDR

$$\mathrm{pFDR} = \mathrm{E}\{N_{01} / N_r \mid N_r > 0\}$$

$$= \mathrm{FDR}/\mathrm{pr}\{N_r > 0\}$$

Benjamini-Hochberg (BH) Procedure

Controls the FDR at level α when the P-values following the null distribution are independent and uniformly distributed.

(1) Let $p_{(1)} \leq \cdots \leq p_{(N)}$ be the observed *P*-values.

(2) Calculate $\hat{k} = \arg\max_{1 \leq k \leq N} \{k : p(k) \leq \alpha k / N\}$.

(3) If $\hat{k}$ exists then reject null hypotheses corresponding to $p_{(1)} \leq \cdots \leq p_{(k)}$. Otherwise, reject nothing.

Bayes Decision Rule

$$\text{Risk} = (1-c)\pi_0 e_{01} + c\pi_1 e_{10}$$

Where e_{01} is the probability of a false positive and e_{10} is the probability of a false negative.

$$\widehat{\text{Risk}} = (1-c_0)\hat{\omega}\widehat{\text{FDR}} + c_0(1-\hat{\omega})\widehat{\text{FNR}}$$

where

$$\hat{\omega} = N_r/N$$

Suppose $\tau_0(w)$ is monotonic decreasing in w. Then

$$\hat{\tau}_{0(w_j)} \le c_0 \quad \text{for} \quad w_j \ge w_0.$$

$$\hat{FDR} = \hat{\pi}_0 \frac{1 - F_0(w_0)}{1 - \hat{F}(w_0)}$$

Suppose $\tau_0(w)$ is monotonic decreasing in w. Then

$$\hat{\tau}_{0(w_j)} \leq c_0 \quad \text{for} \quad w_j \geq w_0.$$

$$FDR = \hat{\pi}_0 \frac{1 - F_0(w_0)}{1 - \hat{F}(w_0)}$$

where $F_0(w_0) = \Phi(w_0)$

$$\hat{F}(w_0) = \hat{\pi}_0 \Phi(w_0) + \sum_{i=1}^{g} \hat{\pi}_i \Phi\left(\frac{w_0 - \hat{\mu}_i}{\hat{\sigma}_i} \right)$$

For a desired control level α, say $\alpha = 0.05$, define

$$w_0 = \arg\min_{w} \left\{ F\hat{D}R(w) \leq \alpha \right\} \qquad (1)$$

If $\pi_0 \dfrac{1 - F_0(w)}{1 - F(w)}$ is monotonic in w, then using (1)

to control the FDR [with $\hat{\pi}_0 = 1$ and $\hat{F}(w_0)$ taken to be the empirical distribution function] is equivalent to using the Benjamini-Hochberg procedure based on the P-values corresponding to the statistic w_j.

The SAM Method

Use the permutation method to calculate the null distribution of the modified *t*-statistic (Tusher et al., 2001).

$$\bar{t}_{0(j)} = (1/B) \sum_{b=1}^{B} t_{0(j)}^{(b)} \quad (j = 1, \ldots, N).$$

The order statistics $t_{(1)}, \ldots, t_{(N)}$ are plotted against their null expectations above.

A good test in situations where there are more genes being over-expressed than under-expressed, or vice-versa.