Differential Expression Detection in Microarray Data

Jianhua Hu
Department of Bioinformatics and Computational Biology

2.11.08

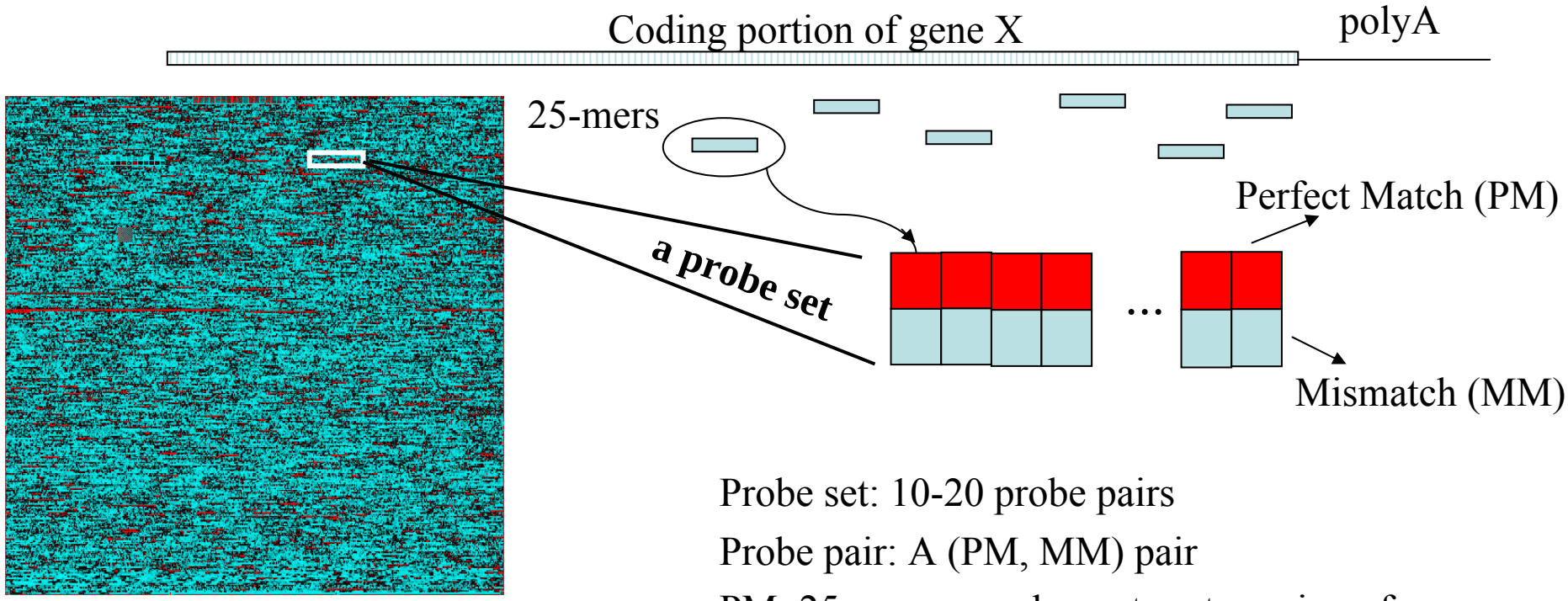
Outline

- Introduction
- Penalized t-statistics by Tusher et al. (2001) with implementation in the software ``SAM".
- Moderated t-statistics by Smyth et al. (2004) with implmentation in R package ``limmod".
- Likelihood-based identifying of differentially expressed genes by Hu and Wright (2007).

Introduction

- DNA microarrays play an important role in many areas of biomedical research.
- Two popular types: Spotted cDNA microarrays, multiprobe **oligonucleotide arrays (Affymetrix Genechip)**.
- Multiprobe oligonucleotide array: probe redundancy, one “color”.

An example Affymetrix genechip array:



Probe set: 10-20 probe pairs

Probe pair: A (PM, MM) pair

PM: 25-mers complementary to region of gene

MM: Middle base is different to that of PM

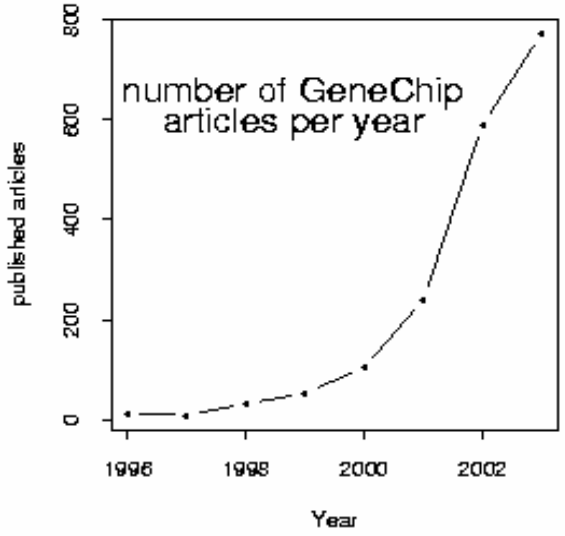
Current methods

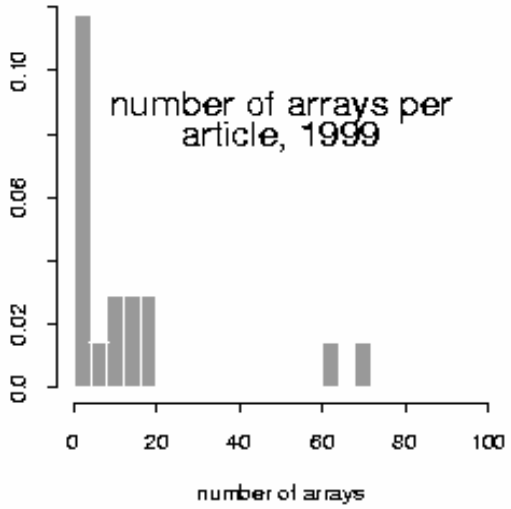
- Simple rule without accounting for expression variation:
  Chen et al. (1997) - “fold changes” rule.

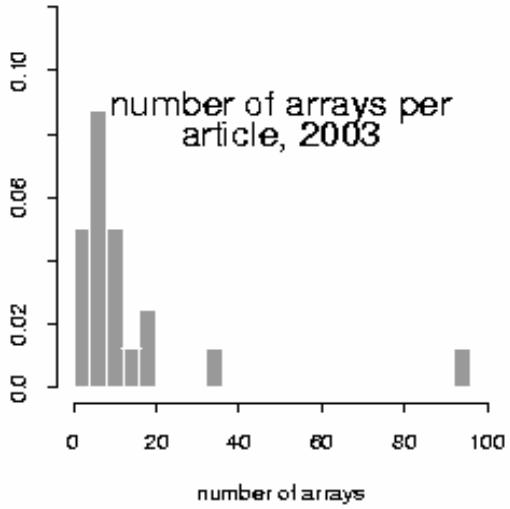
- Ordinary two-sample t-statistic:
  Dudoit et al. (2002), **Thomas et al. (2001)**

- Modified two-sample t-statistic:
  Tusher et al. (2001), **Efron et al. (2001), Smyth (2004).**

- "Borrowing" information from across the genes:
  - **Eaves et al. (2002)** - weighted average of the sample variance and a local variance estimate for groups of genes.
  - Bayesian approaches:
    Newton et al. (2001), **Baldi and Long (2001),**
    Ibrahim et al. (2002).
- Uncertainty in the variance is an acute problem when the sample size is small.







- The median number of arrays was 8 both in 1999 *and* 2003.
- Illustrates importance of dealing with a wide range of sample sizes.

Significance analysis of Microarrays (SAM)

Tusher, Tibshirani, Chu (2001)

The Problem:

- Identifying differentially expressed genes
- Determine which changes are significant
- Enormous number of genes

Reminder: t-Test

- t-Test for a single gene:
- We want to know if the expression level changed from condition A to condition B.
- Null assumption: no change
- Sample the expression level of the genes in two conditions, A and B.
- Calculate $\overline{x}_A, \overline{x}_B$
- H_0: The groups are not different, $E(\overline{x}_A - \overline{x}_B) = 0$

t-Test Cont'd

- Under H_0, and under the assumption that the data is normally distributed,

t-Statistic

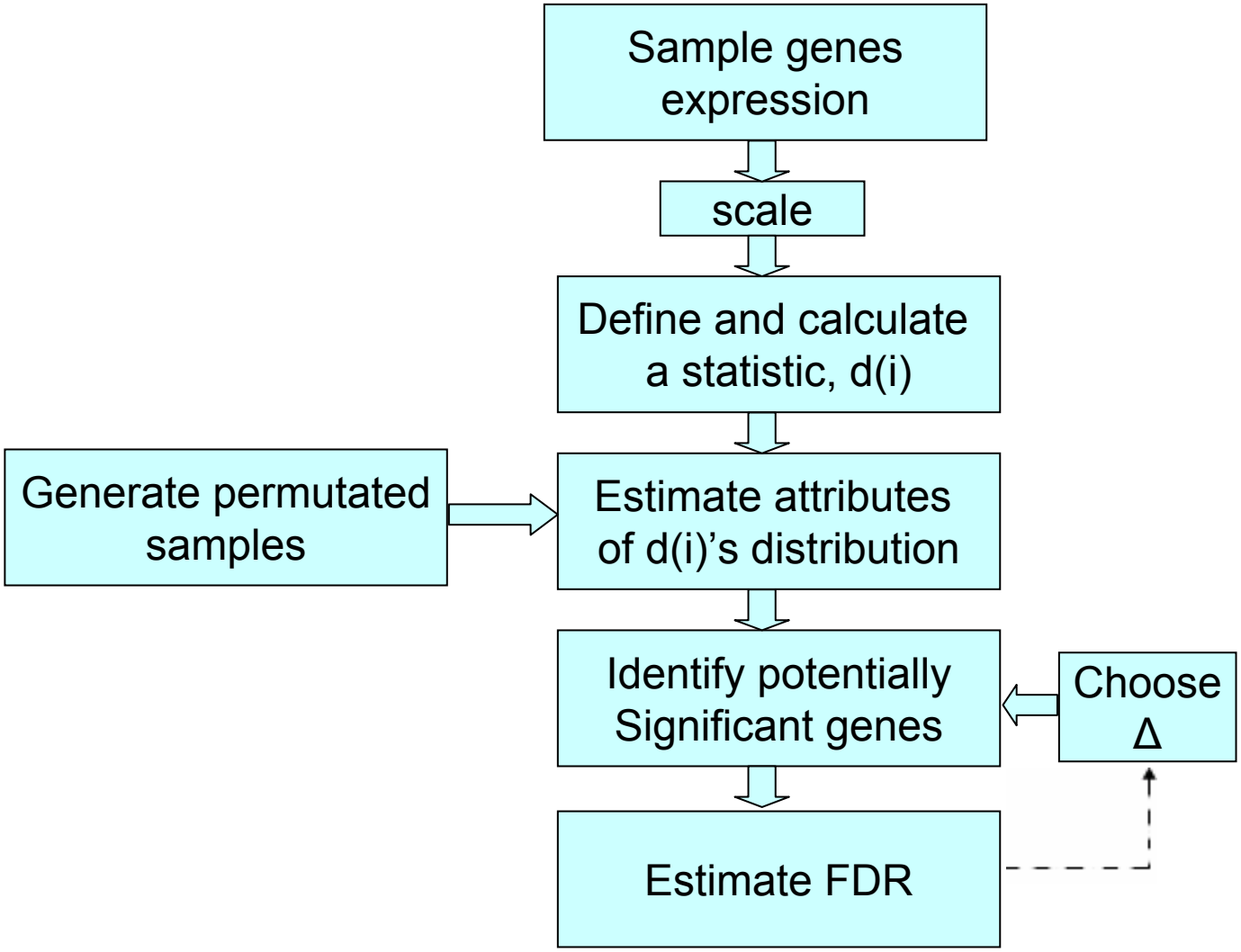
$$\frac{(\bar{x}_A - \bar{x}_B) - 0}{\hat{\sigma}(\bar{x}_A - \bar{x}_B)} \sim t$$

- Use the distribution table to determine the significance of your results.

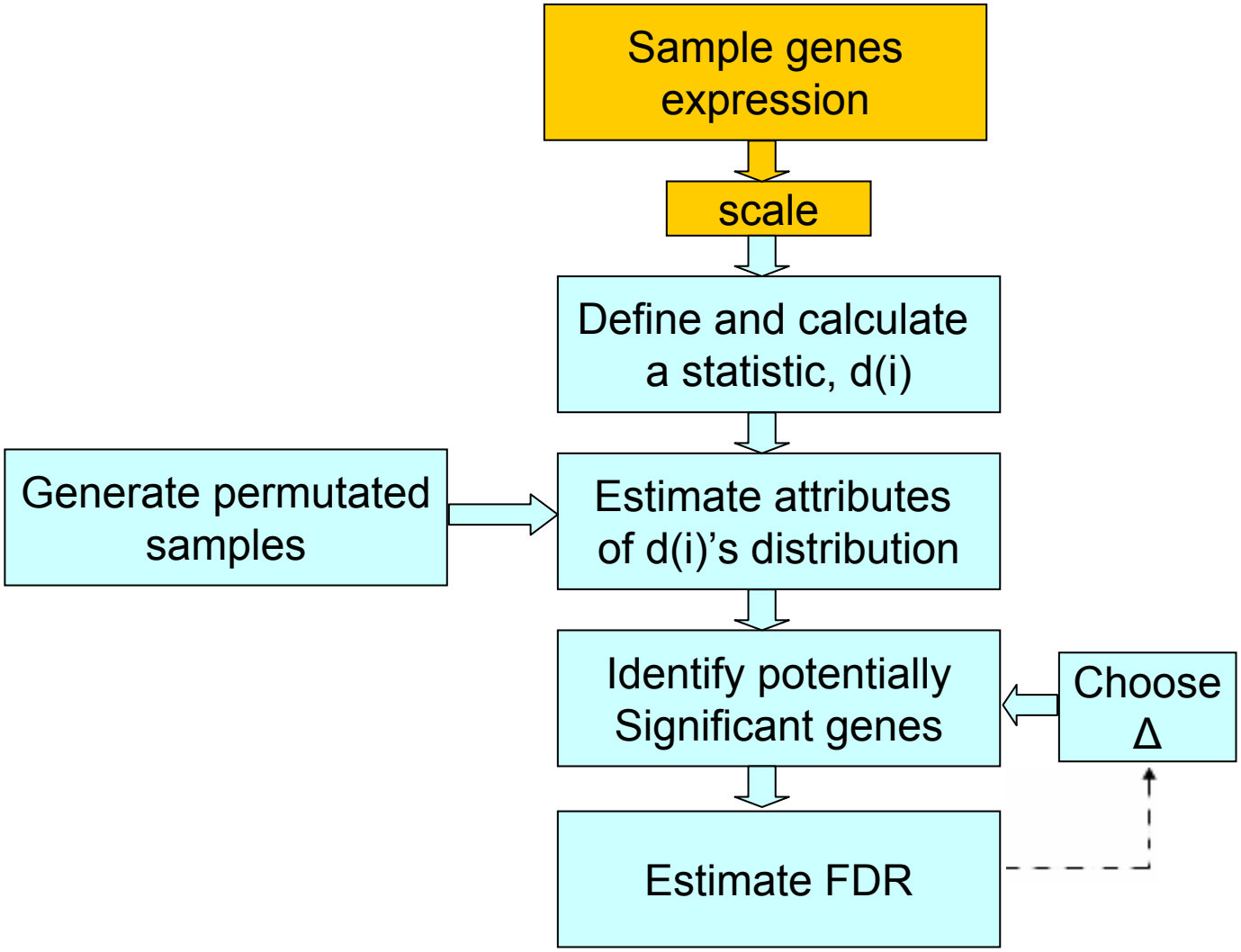
Multiple Hypothesis Testing

- Naïve solution: do t-test for each gene.
- Multiplicity Problem: The probability of error increases.
- We've seen ways to deal with it, that try to control the FWER or the FDR.
- Today: SAM (estimates FDR)

SAM- procedure overview

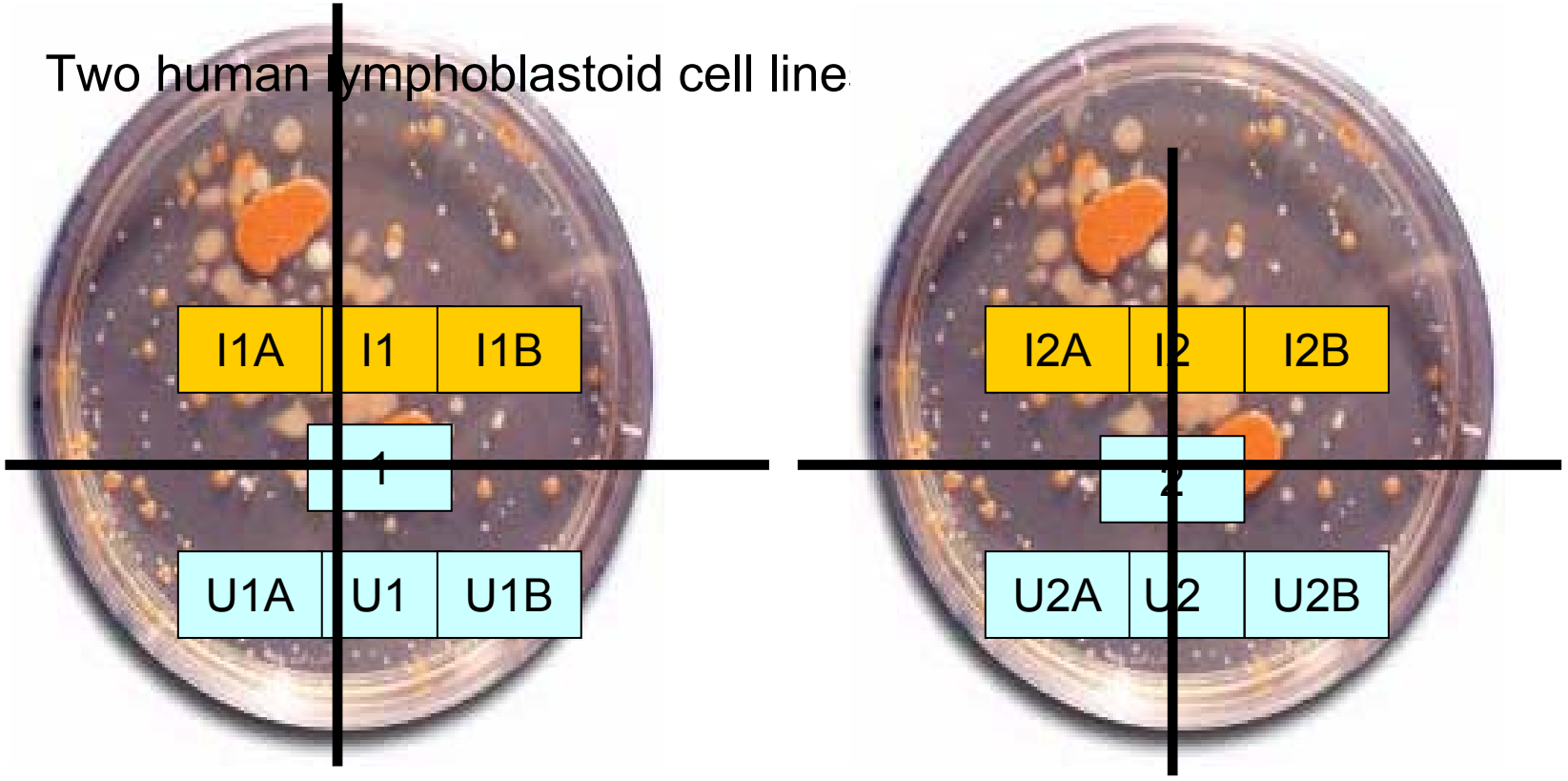


SAM- procedure overview



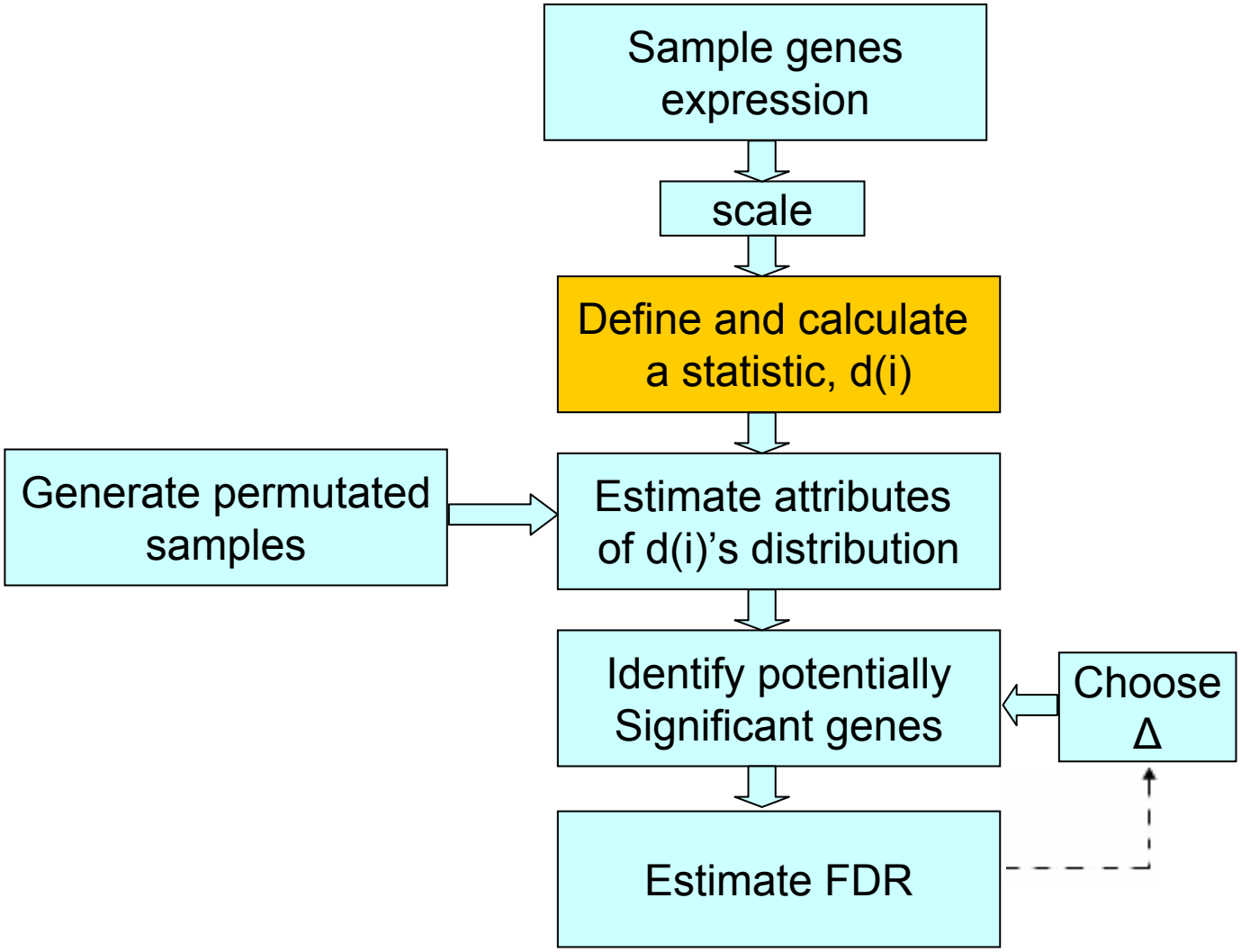
The Experiment

Two human lymphoblastoid cell line



Eight hybridizations were performed.

SAM- procedure overview



SAM's statistic- Relative Difference

- **Define a statistic, based on the ratio of change in gene expression to standard deviation in the data for this gene.**

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}$$

Difference between the means of the two conditions

Estimate of the standard deviation of the numerator

Fudge Factor

$$s(i) = \sqrt{\left(\frac{\frac{1}{n_1} + \frac{1}{n_2}}{n_1 + n_2 - 2}\right)\left\{\sum_m [x_m(i) - \bar{x}_I(i)]^2 + \sum_m [x_m(i) - \bar{x}_I(i)]^2\right\}}$$

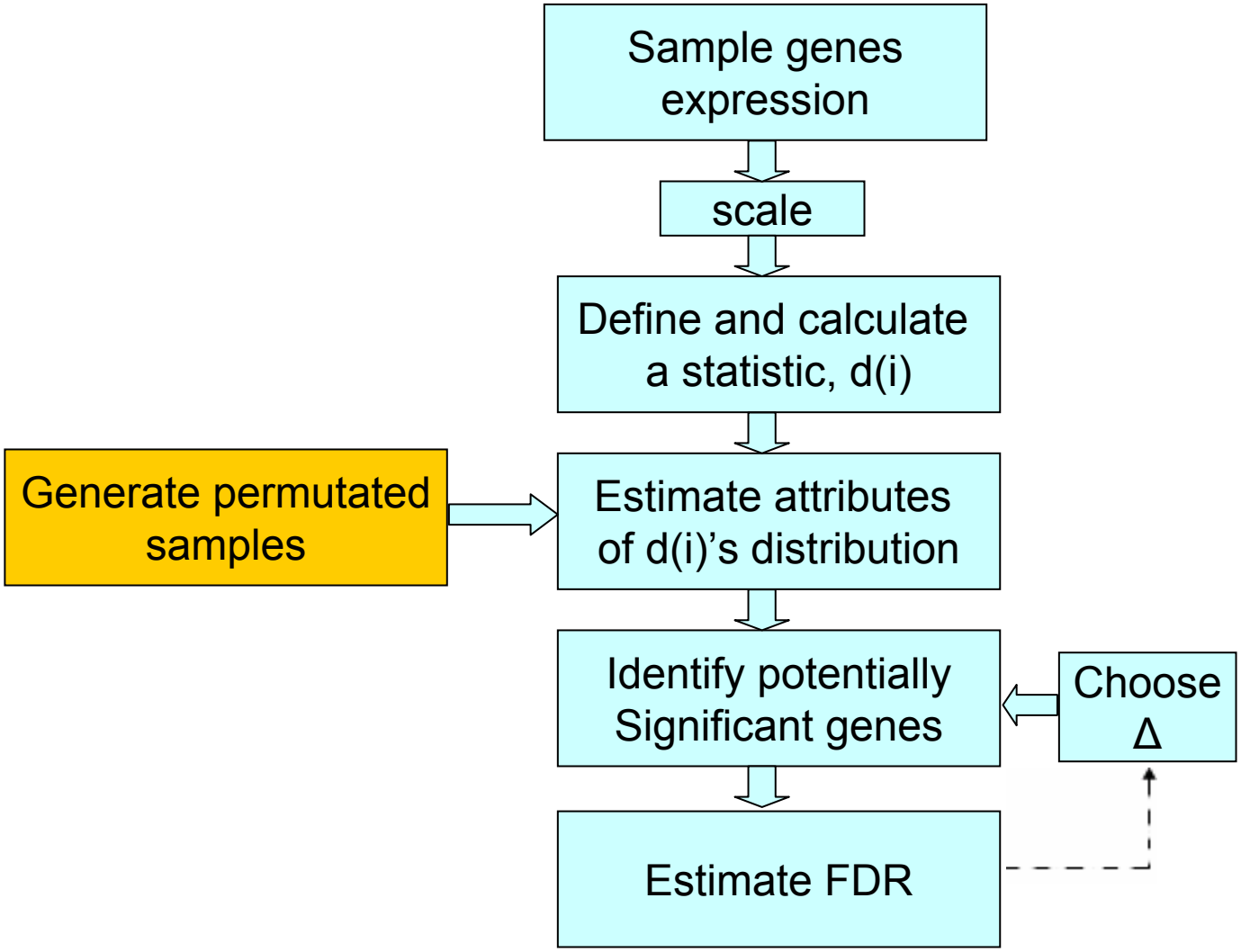
Why s_0?

- At low expression levels, variance in d(i) can be high, due to small values of s(i).
- To compare d(i) across all genes, the distribution of d(i) should be independent of the level of gene expression and of s(i).
- Choose s_0 to make the coefficient of variation of d(i) approximately constant as a function of s(i).

Now what?

- We gave each gene a score.
- At what threshold should we call a gene significant?
- How many false positives can we expect?

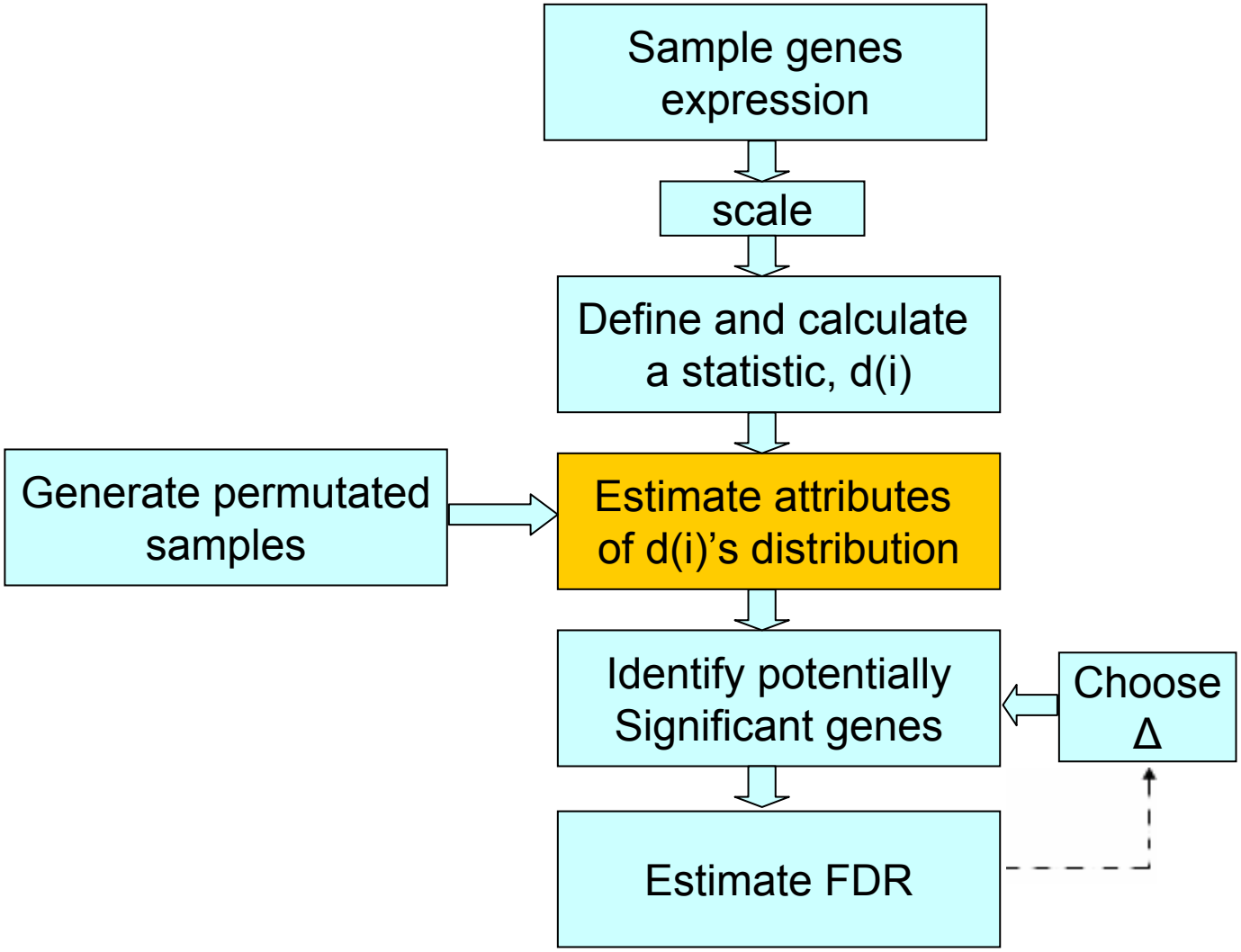
SAM- procedure overview



More data required

- Experiments are expensive.
- Instead, generate permutations of the data (mix the labels)
- Can we use all possible permutations?

SAM- procedure overview



Estimating d(i)'s Order Statistics

- For each permutation p, calculate $d_p(i)$.

$$d_p(i) = \frac{\bar{x}_{G1}(i) - \bar{x}_{G2}(i)}{s(i) + s_0}$$

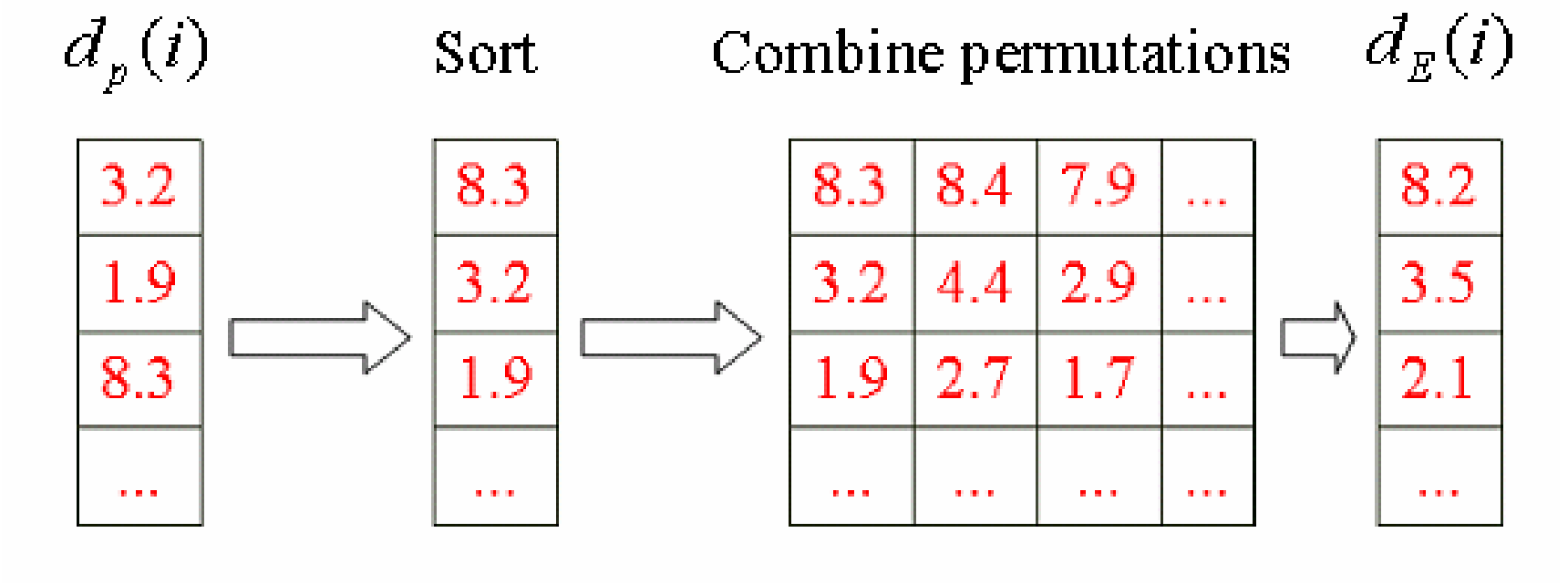
- Rank genes by magnitude:

$$d_p(1) \geq d_p(2) \geq d_p(3) \geq \ldots$$

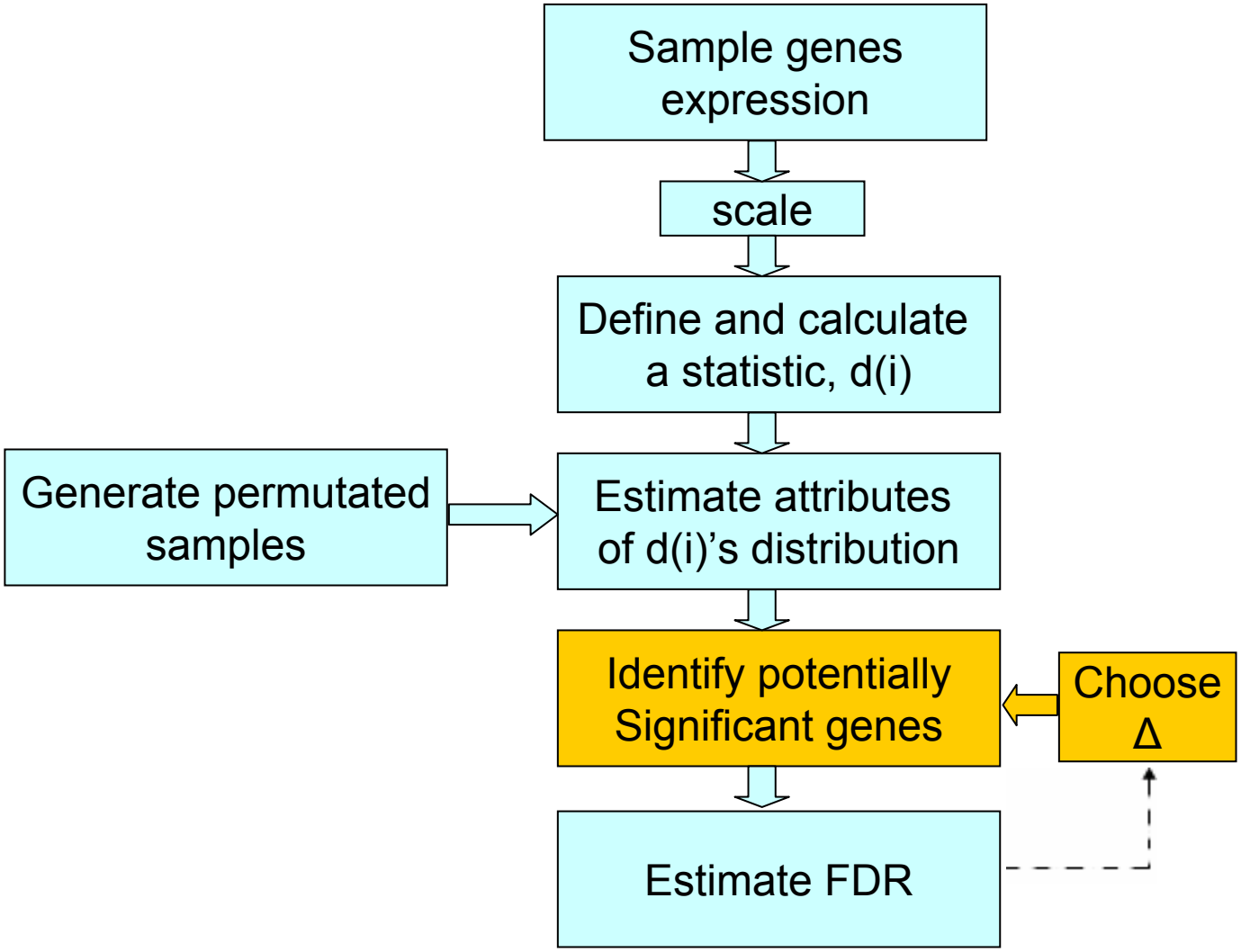
- Define:

$$d_E(i) = \sum_p \frac{d_p(i)}{36}$$

Example


$d_p(i)$
Sort
Combine permutations
$d_E(i)$
3.2
1.9
8.3
...
8.3
3.2
1.9
...
8.3 8.4 7.9 ...
3.2 4.4 2.9 ...
1.9 2.7 1.7 ...
... ...
8.2
3.5
2.1
...

SAM- procedure overview



Identifying Significant Genes

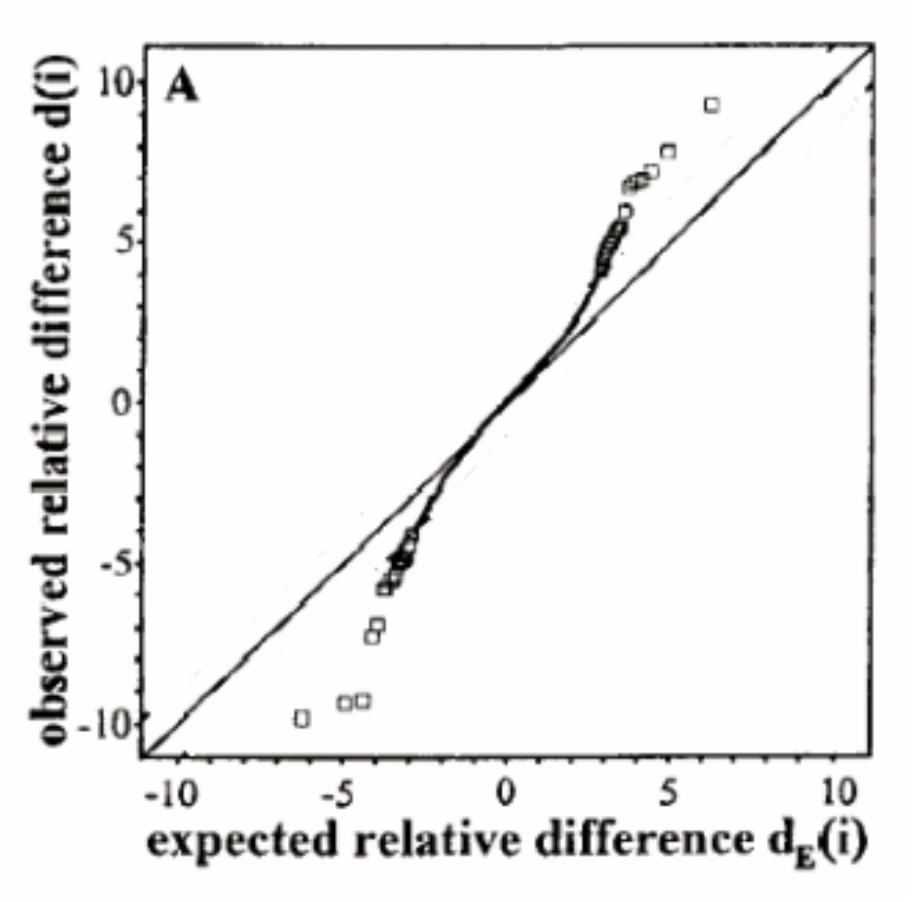
- Now Rank the original d(i)'s:

$$d(1) \geq d(2) \geq d(3) \geq \ldots$$

- Plot d(i) vs. $d_E(i)$:

- For most of the genes,

$$d(i) \cong d_E(i)$$



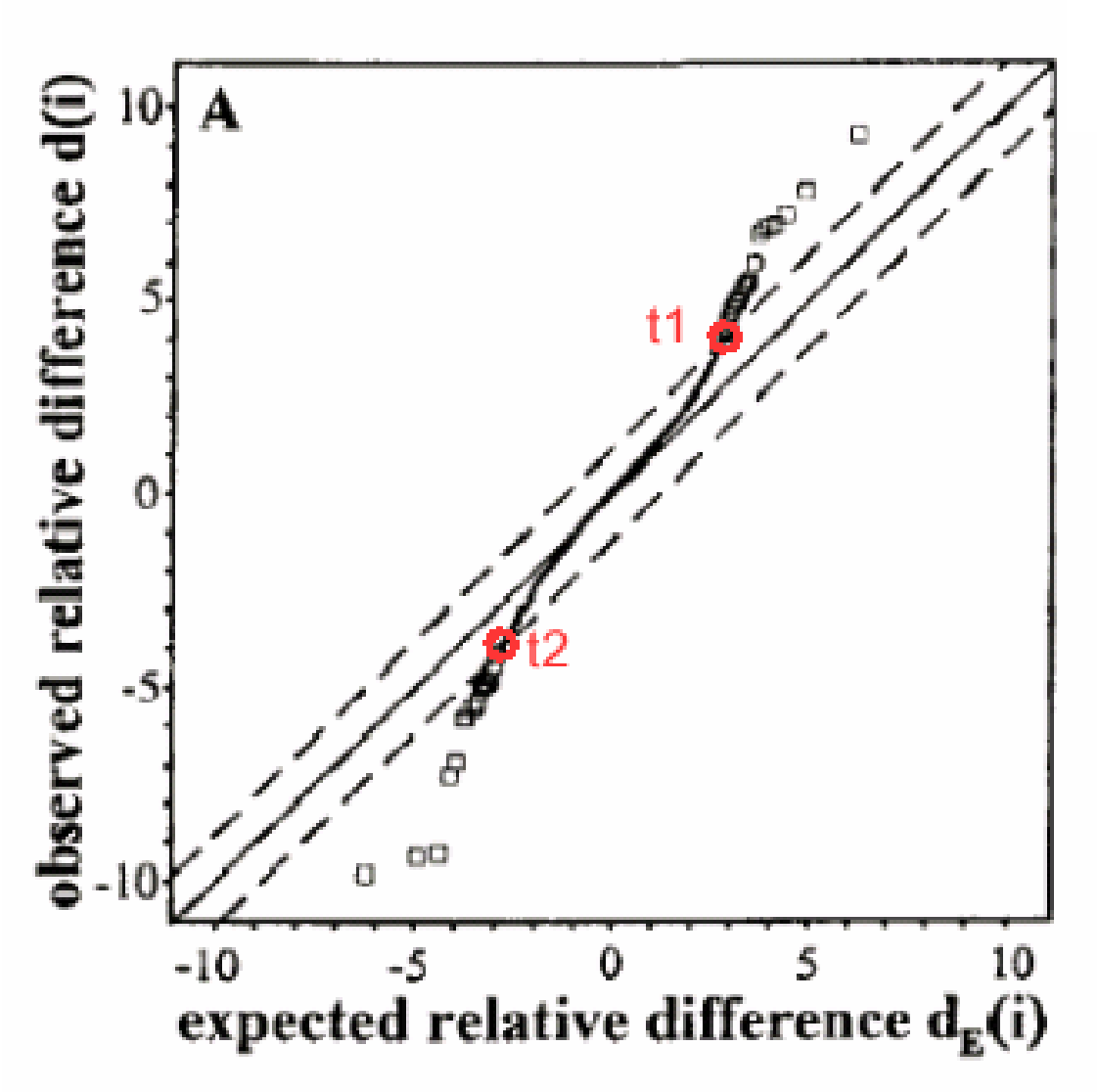
Identifying Significant Genes

- Define a threshold, Δ.
- Find the smallest positive d(i) such that

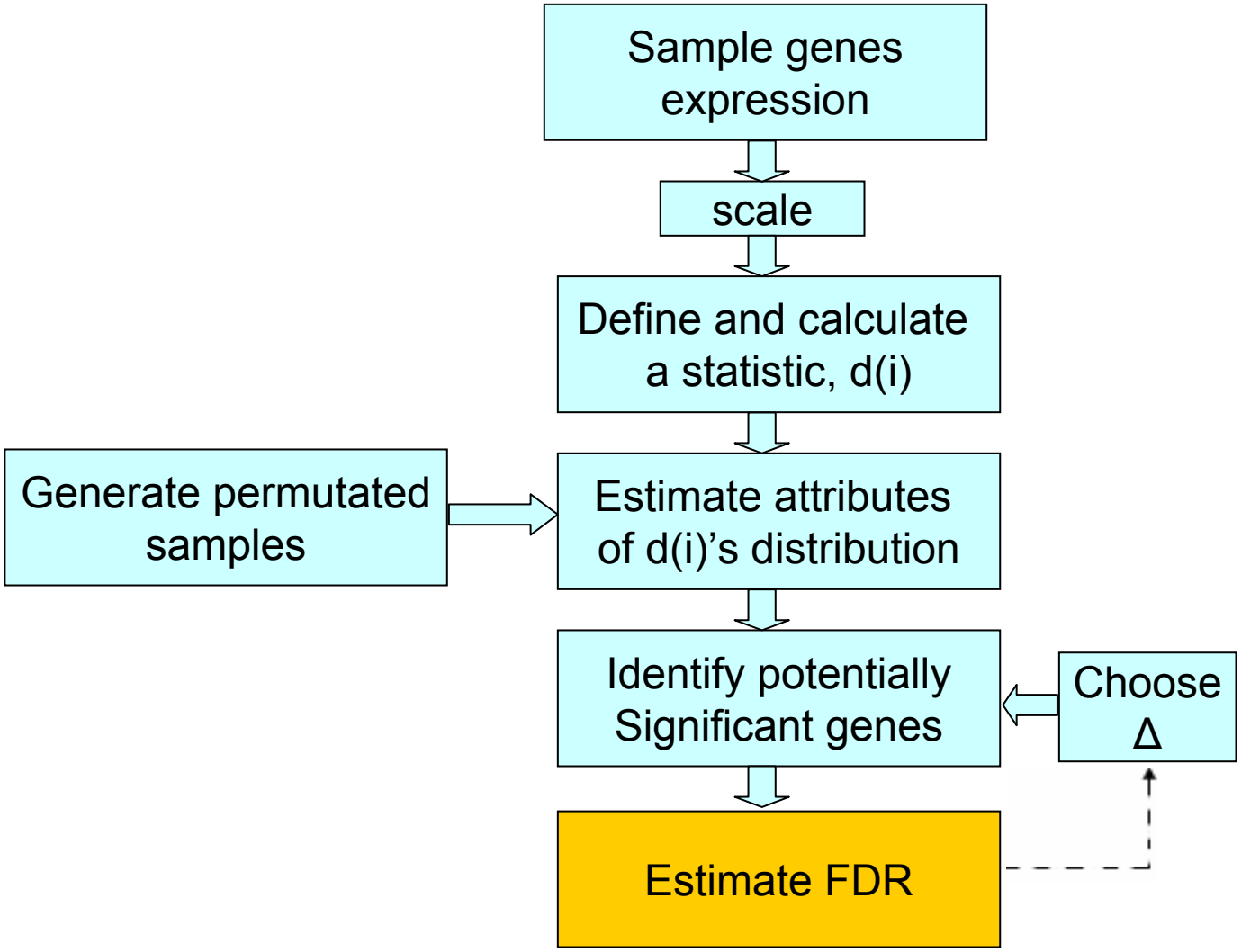
$$\left|d(i) - d_E(i)\right| \geq \Delta$$

  call it t_1.
- In a similar manner, find the largest negative d(i). Call it t_2.
- For each gene i, if, $d(i) \geq t_1 \vee d(i) \leq t_2$

  call it potentially significant.


A
t1
t2
observed relative difference d(i)
expected relative difference $d_E(i)$
10
5
0
-5
-10
-10
-5
0
5
10

SAM- procedure overview



Estimate FDR

- t_1 and t_2 will be used as cutoffs.
- Calculate the average number of genes that exceed these values in the permutations.
- Estimate the number of falsely significant genes, under H_0:

$$\frac{1}{36}\sum_{p=1}^{36}\#\{i \mid d_p(i) \geq t_1 \vee d_p(i) \leq t_2\}$$

- Divide by the number of genes called significant

FDR cont'd

$$FDR \approx \frac{\frac{1}{36}\sum_{p=1}^{36} \#\{i \mid d_p(i) \geq t_1 \vee d_p(i) \leq t_2\}}{\#\{i \mid d(i) \geq t_1 \vee d(i) \leq t_2\}}$$

- Note: Cutoffs are asymmetric

Example

	$d(i)$	$d_p(i)$			
	8.3	8.3	8.4	7.9	8.1
t_1	4.2	3.2	4.4	2.5	1.6
	2.9	1.9	2.7	1.7	0.1
t_2	−0.5	0.3	−0.6	1.0	−2.1

$$FDR \approx \frac{\frac{7}{4}}{3} = 0.5833$$

How to choose Δ?

Parameter	Number falsely significant	Number called significant	FDR
SAM			
$\Delta = 0.4$	134.9	288	47%
$\Delta = 0.5$	78.1	192	41%
$\Delta = 0.6$	56.1	162	35%
$\Delta = 0.9$	19.1	80	24%
$\Delta = 1.2$	8.4	46	18%

Omitting s_0 caused higher FDR.

Test SAM's validity

- 10 out of 34 genes found have been reported in the literature as part of the response to IR
- 19 appear to be involved in the cell cycle
- 4 play role in DNA repair
- Perform Northern Blot- strong correlation found
- Artificial data sets- some genes induced, background noise

Moderated t-statistic

Smyth G. K. (2004)

Linear model estimates

Obtain a linear model for each gene g

$$E(\mathbf{y}_g) = \mathbf{X}\boldsymbol{\beta}_g, \quad var(\mathbf{y}_g) = W_g^{-1}\sigma_g^2.$$

Estimate model by *robust regression*, *least squares*, or *generalized least squares* to obtain

coefficients, $\hat{\beta}_{gj}$

estimators of σ_g^2, s_g^2

standard errors, $\mathrm{se}(\hat{\beta}_{\mathrm{gj}})^2 = \mathrm{c}_{\mathrm{gj}}\mathrm{s}_{\mathrm{g}}^2$.

Parallel inference for genes

1. 10,000-40,000 linear models.
2. High dimensionality:
   Need to adjust for multiple testing, e.g., control family-wise error rate (FWE) or false discover rate (FDR).
3. The key: borrow information across genes.

Hierarchical model

$$\hat{\beta}_{gj} \sim N(\beta_{gj}, c_{gj}\sigma_g^2)$$

$$P(\beta_{gj} \neq 0) = p$$

$$\beta_{gj} \mid \beta_{gj} \neq 0 \sim N(0, c_{0j}\sigma_g^2)$$

$$s_g^2 \sim \frac{\sigma_g^2}{d_g}\chi_{d_g}^2$$

$$\sigma_g^2 \sim s_0^2(\chi_{d_0}^2/d_0)^{-1}$$

Posterior statistics

1. Posterior variance estimators

$$\tilde{s}_g^2 = E(\sigma_g^2 \mid s_g^2) = \frac{s_g^2 d_g + s_0^2 d_0}{d_g + d_0}$$

2. Moderated t-statistic

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{c_{gj}}}$$

3. The goal: eliminates large t-statistics merely from very small s.

Marginal distributions

The marginal distributions of the sample variances are moderated t-statistics are mutually independent

$$s_g^2 \sim s_0^2 F_{d,d_0}$$

$$\tilde{t}_g \sim \begin{cases} t_{d_0+d} & \text{with prob } 1-p \\ \dfrac{t_{d_0+d}}{\sqrt{1+c_0/c}} & \text{with prob } p \end{cases}$$

Degrees of freedom add!

Estimating prior parameters

Marginal moments of $\log(s^2)$ lead to estimators of s_0 and d_0:
Estimate d_0 by solving

$$\psi'(d_0/2) = \text{mean}\{\text{ns}_\text{e}^2 - \psi'(\text{d}_\text{g}/2)\}$$

where

$$e_g = \log(s_g^2) - \psi(d_g/2) + \log(d_g/2),$$

and

$$s_e^2 = (e_g - \bar{e})^2/(n-1).$$

Finally

$$s_0^2 = \exp\{\bar{\text{e}} + \psi(\text{d}_0/2) - \log(\text{d}_0/2)\}.$$

Where $\psi()$ and $\psi'()$ are the digamma and trigamma functions respectively.

Estimating prior parameters

$s_1, s_2, \cdots, s_g \longrightarrow \tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_g \longrightarrow s_0$
$t_1, t_2, \cdots, t-g \longrightarrow \tilde{t}_1, \tilde{t}_2, \cdots, \tilde{t}_g \longrightarrow t_{g,\text{pooled}}$
The data decides whether $\tilde{t}_g$ should be closer to $t_{g,\text{pooled}}$ or to t_g.

Quantile estimation of c_0

let r be rank of $\mid \tilde{t}_g \mid$ in descending order, and let $F()$ be the distribution function of the t-distribution. c_0 can be estimated by equating empirical to theoretical quantiles:

$$2[pF(-\sqrt{\frac{c_g}{c_g + c_0}} \mid \tilde{t}_g \mid; d_0 + d_g) + (1-p)F(- \mid \tilde{t}_g \mid; d_0 + d_g)] = \frac{r - 0.5}{n}$$

Get overall estimator of c_0 by averaging the individual estimators from the top $p/2$ proportion of the $\mid \tilde{t}_g \mid$.

Posterior odds

Posterior probability of differential expression for any gene is

$$\frac{p(\beta \neq 0 \mid \hat{\beta}, s^2)}{p(\beta = 0 \mid \hat{\beta}, s^2)} = \frac{p}{1-p} \left(\frac{c}{c+c_0} \right)^{1/2} \left\{ \frac{\tilde{t}^2 + d + d_0}{\tilde{t}^2 \frac{c}{c+c_0} + d + d_0} \right\}^{\frac{1+d+d_0}{2}}$$

It is a monotonic function of $\tilde{t}^2$ for constant d.

Limma

- Limma is an R package to find differentially expressed genes
- it uses linear models
  - fitted to normalized intensities (one-color)
  - or log-ratios (two-color)
- assumption: normal distribution
- output: p-values (adjusted for multiple testing)

Documentation

- limma User's Guide, Gordon Smyth, Natalie Thorne, James Wettenhall
- help documents for each function
- Smyth, GK (2004). SAGMB 3 (1) article 3
- de Menezes RX, Boer JM, van Houwelingen JC (2004). Applied Bioinformatics 3: 229-235
- background on linear models: tech note by Renee de Menezes

limma

- linear models
  - can be used to compare two or more groups
  - can be used for multifactorial designs
    - e.g. genotype and treatment
- uses empirical Bayes analysis to improve power in small sample sizes
  - borrowing information across genes

Pre-analysis steps

- read data into limma/affy
- basic quality control features
- background correction
- within-array normalization
- between-array normalization
- if duplicate spotting: sort data so that duplicates are together

Linear model

- make design matrix
- fit a linear model to estimate all the fold changes
- [make contrasts matrix]
- apply Bayesian smoothing to the standard errors (very important!)
- output: moderated t-statistics

Two color - start

- working directory containing
  - *.gpr files
  - targets.txt file
  - *.gal file (optional)

Reading in data

- basically the same as Anja Schiel has shown for Quality Control packages
  - read in a targets file including
    - file names for *.gpr files
    - cy3 and cy5 samples
  - read in *.gpr files using read.maimages()
  - option to use GenePix flag information
  - print layout (from *.gpr or *.gal file)
  - option to define spot types (controls)

Other BioC packages

- Limma package can work with microarray objects derived by these packages:
- marray: marrayRaw and marrayNorm
- affy: single channel (exprSet)

Exploring data

- automate the production of plots for all arrays in an experiment
  - imageplot3by2
    - array image of R, Rb, G, Gb, M (R/G) (un)norm
  - plotMA3by2
    - MA plots before/after normalization
  - plotDensities
    - histogram of all intensities before/after normalization

Background correction

- default = subtract
  - disadvantage: negative values -> NAs
- “normexp”, offset = 50
  - adjusts fg to bg to yield strictly positive intensities
  - use of an offset damps the variation of the log-ratios for very low intensities towards 0, i.e. stabilizes the variability of the M-values as a function of intensity
  - this is important for the empirical Bayes methods

Normalization 1

- normalizeWithinArray
  - normalizes M-values of each array separately
  - default = print-tip loess
  - not appropriate for e.g. Agilent arrays, which do not have print groups: method = “loess”
  - assumes bulk of probes not changed
  - symmetrical change is not required
  - spot quality weights (in RG) are used by default; weight = 0 will not influence normalization of other spots, but will be kept and normalized

Normalization 2

- normalizeBetweenArray
  - intensities of single-channel microarrays
  - log-ratios of two-color microarrays as a second step after within array normalization of the M-values
  - because: loess normalization doesn’t affect the A-values
  - quantile normalization results in equal distributions across channels and arrays

Normalization 3

- normalizeBetweenArrays directly on two-color data
  - quantile normalization directly to individual red and green intensities
  - vsn normalization should always be used directly on raw intensities
    - background subtraction is allowed,
    - but no correction (e.g. normexp) or loess!!!

Linear models

- design matrix
  - indicates which RNA samples have been applied to each array
  - rows: arrays; columns: coefficients
- contrast matrix
  - specifies which comparisons you would like to make between the RNA samples
  - for very simple experiments, you may not need a contrast matrix

Look at the result

- topTable(fit, adjust="fdr")
  - gives the top10 of differentially expressed genes (for each contrast)
- plotMA(fit)
- decideTests
  - makes a matrix with 0 (not selected) and -1/1 (selected for a specific p-value)
  - visualize by Venn diagram

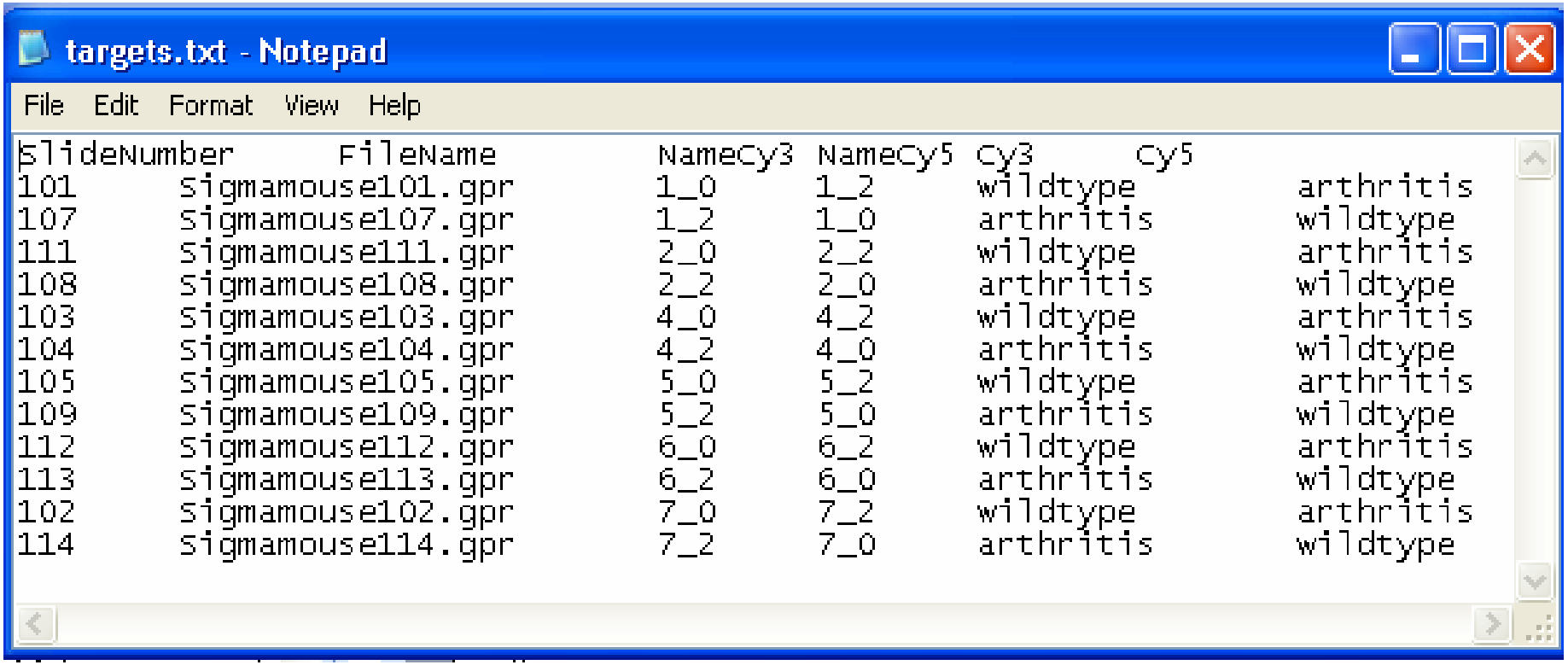
Limma objects

- RGList (Red-Green, raw data)
  - generated by read.maimages
- MAList (M- and A-values, normalized data)
  - generated by MA.RG or normalizeWithinArrays
- MArrayLM (result of fitting linear model)
  - generated by lmFit
- TestResults (results of testing a set of contrasts equal to 0 for each probe)
  - generated by decideTests

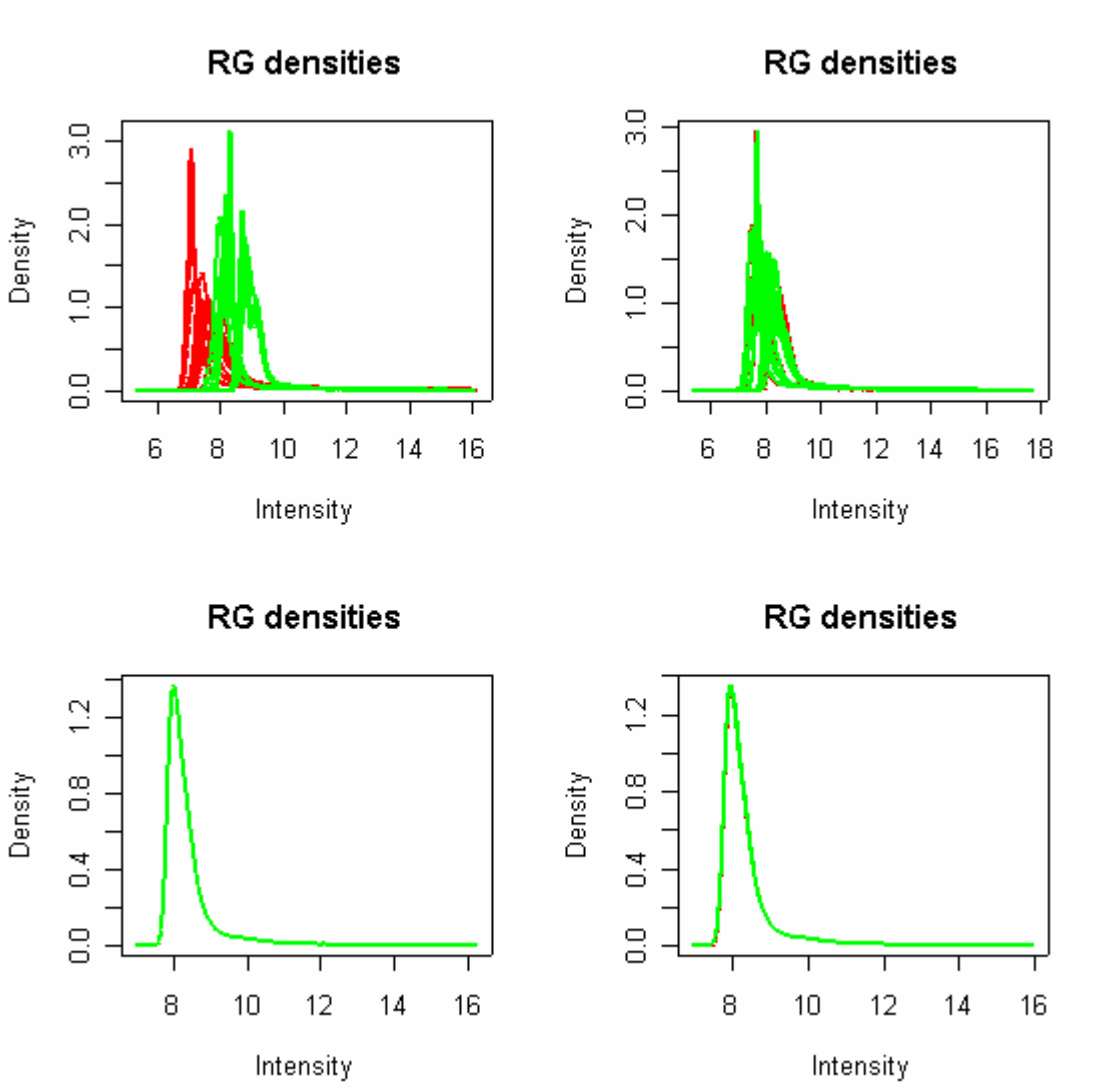
Example 1: paired design

- direct two-color design including dye-swap
- dataset "arthritis", Maaike van den Hoven
- platform: Sigmamouse, 23232 single spots
- 12 arrays, 2 groups:
  - untreated (6 biological replicates)
  - arthritis (6 biological replicates)
- question: find differentially expressed genes after induction of arthritis

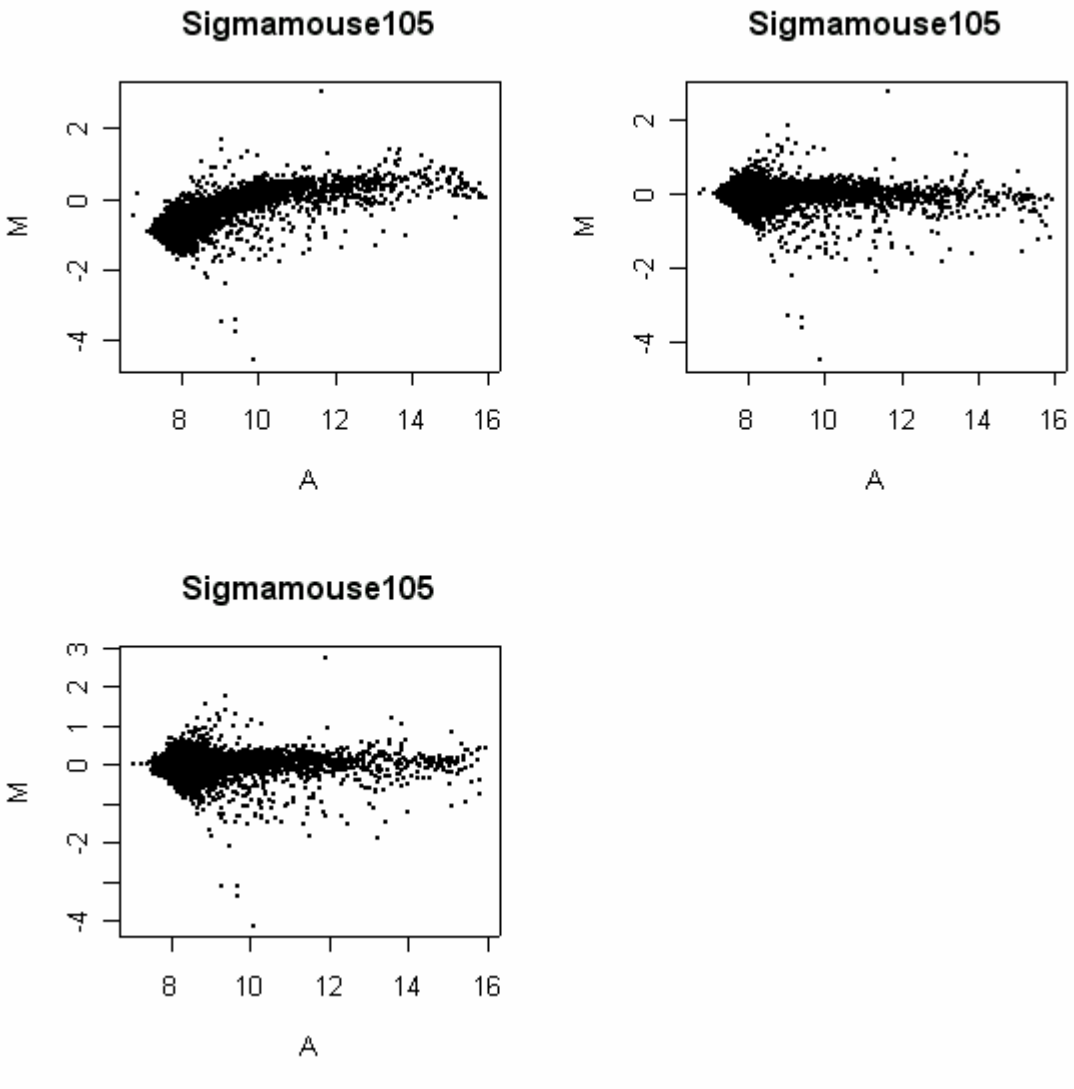
targets.txt


targets.txt - Notepad
File Edit Format View Help
SlideNumber FileName NameCy3 NameCy5 Cy3 Cy5
101 Sigmamouse101.gpr 1_0 1_2 wildtype arthritis
107 Sigmamouse107.gpr 1_2 1_0 arthritis wildtype
111 Sigmamouse111.gpr 2_0 2_2 wildtype arthritis
108 Sigmamouse108.gpr 2_2 2_0 arthritis wildtype
103 Sigmamouse103.gpr 4_0 4_2 wildtype arthritis
104 Sigmamouse104.gpr 4_2 4_0 arthritis wildtype
105 Sigmamouse105.gpr 5_0 5_2 wildtype arthritis
109 Sigmamouse109.gpr 5_2 5_0 arthritis wildtype
112 Sigmamouse112.gpr 6_0 6_2 wildtype arthritis
113 Sigmamouse113.gpr 6_2 6_0 arthritis wildtype
102 Sigmamouse102.gpr 7_0 7_2 wildtype arthritis
114 Sigmamouse114.gpr 7_2 7_0 arthritis wildtype

plotDensities(RGb, MA, MA.q, MAq)



plotMA(RGb, MA, MAq)



topTable(MA.q, adjust="fdr")

Block	Row	Column	ID	Name	M	A	t	P.Value	B
1838	4	18	12	NM_026004	NA	1.996	9.31	9.58	0.00577	6.88
9277	20	4	15	NM_018762	NA	0.392	10.88	8.63	0.00577	5.91
5551	12	11	7	NM_017372	NA	1.741	9.37	8.55	0.00577	5.74
14031	29	22	17	AmbionSpike5	NA	1.053	14.24	8.43	0.00577	5.66
18056	38	7	16	NM_020611	NA	0.187	8.21	8.52	0.00577	5.52
15529	33	2	19	U52197	NA	0.340	8.83	8.28	0.00598	5.47
13274	28	10	8	X83919	NA	0.407	8.56	8.11	0.00598	5.18
22079	46	14	13	NM_026542	NA	2.017	9.97	8.08	0.00598	5.18
1155	3	9	11	X14097	NA	0.251	8.07	8.46	0.00577	5.16
13559	29	1	7	AmbionSpike5	NA	1.034	13.93	7.93	0.00598	5.03

AmbionSpike5 was spiked in at 2-fold change arthritis/untreated: log-ratio 1

The likelihood based approach

Hu and Wright (2007)

Notation

- A simple family of *t*-like statistics for gene *i*:

$$t_i^a = \frac{\bar{x}_{1i} - \bar{x}_{2i}}{s_i + a}, \text{ with } s_i = \sqrt{s_{1i}^2 / n_1 + s_{2i}^2 / n_2},$$

  t^0 is the "ordinary" Welch statistic, $f = n_1 + n_2 - 2.$

- $\delta_i = \frac{\mu_{1i} - \mu_{2i}}{\sqrt{\frac{\sigma_{1i}^2}{n_1} + \frac{\sigma_{2i}^2}{n_2}}}$, governs the power of the statistics.

- t^0 can be viewed as an **estimate** of δ.

  Its performance can be examined by the positive FDR in **Storey (2001)**,

$$FDR = Pr(H_0 \mid T \geq c)$$

FDR property of t^0

- FDR reaches a limit as $c \to \infty$,
- δs, independent random variable, with

$$\delta_i = \begin{cases} 0 \text{ w.p. } \pi_0 \\ \delta' \text{ w.p. } \pi_1 \end{cases}$$

- **Theorem 1** $\lim_{c\to\infty} FDR = \pi_0 / (\pi_0 + \pi_1 Q)$, where

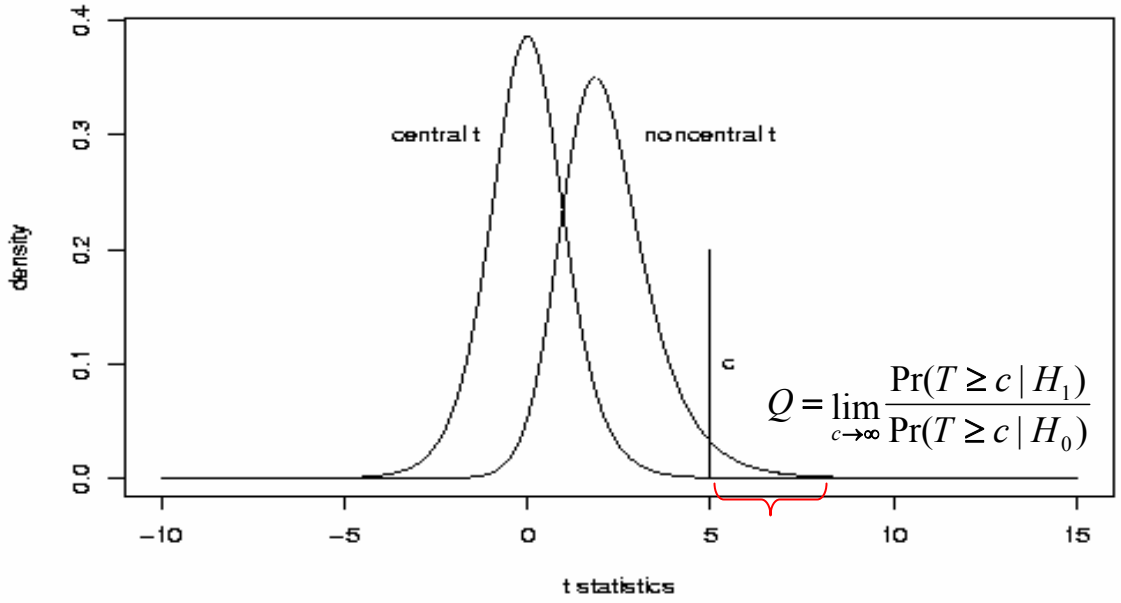
$$Q = \lim_{c\to\infty} \frac{\Pr(T \geq c \mid H_1)}{\Pr(T \geq c \mid H_0)} = \frac{2\Phi(\delta')E(Y^f)}{E(Z^f)},$$

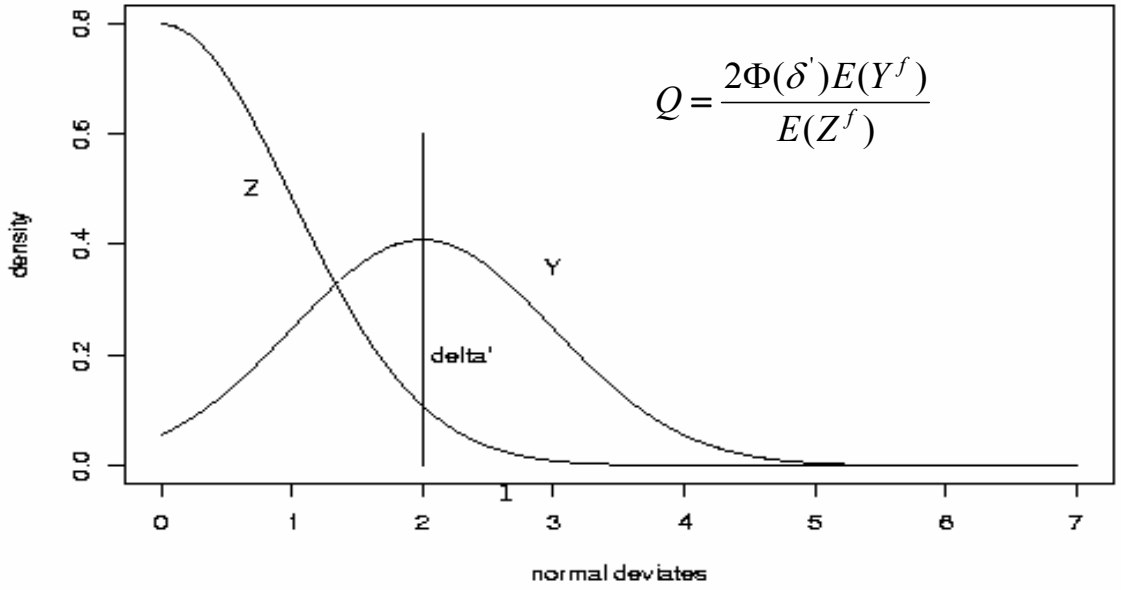
Φ, the standard normal CDF;
Y and Z, random variables with truncated normal densities,

$$p(y) = \exp(-(y-\delta')^2/2)/(\sqrt{2\pi}\Phi(\delta')), y \geq 0$$

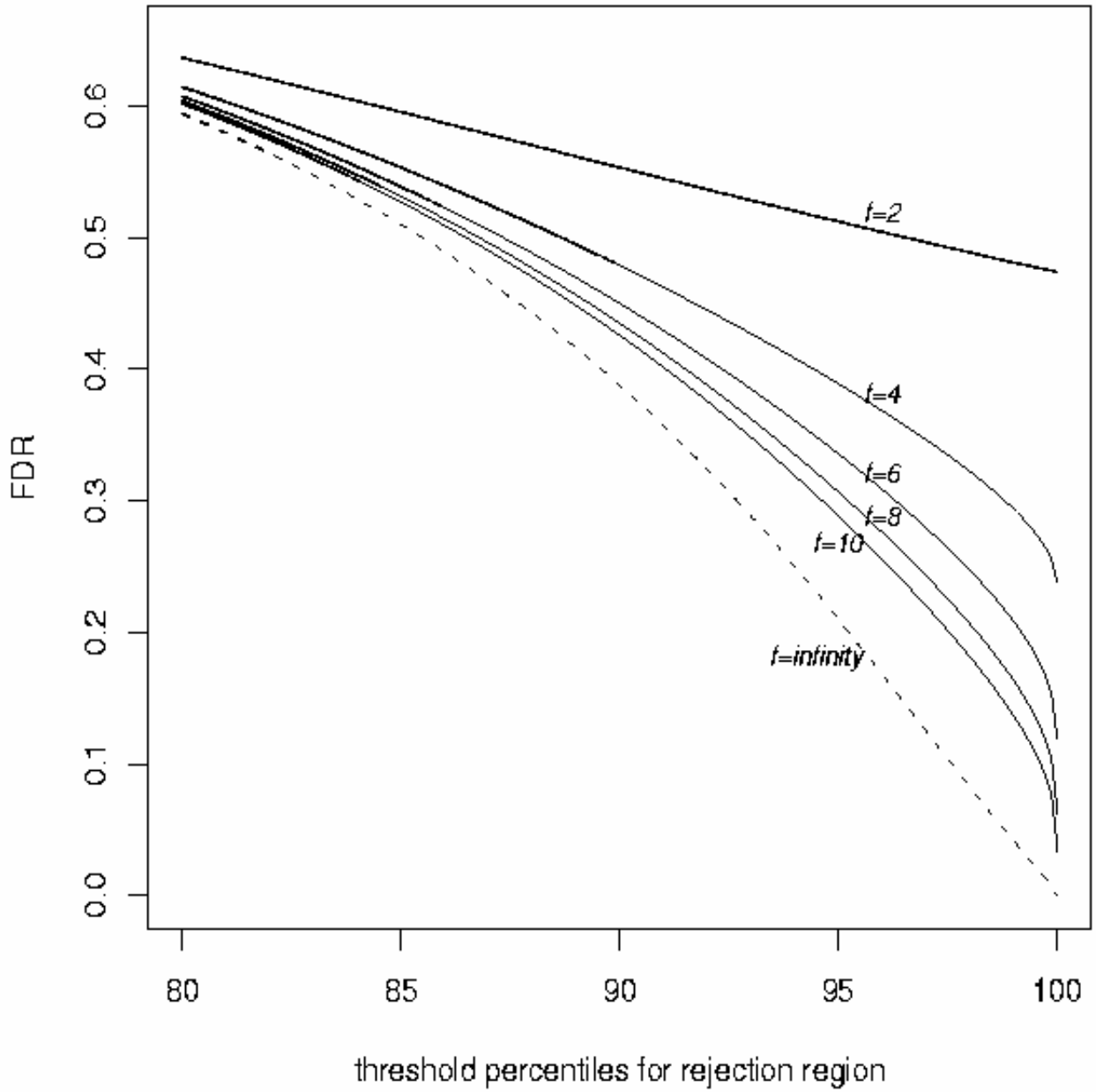
$$p(z) = 2\exp(-z^2/2)/\sqrt{2\pi}, z \geq 0$$

Visualizing the limit Q

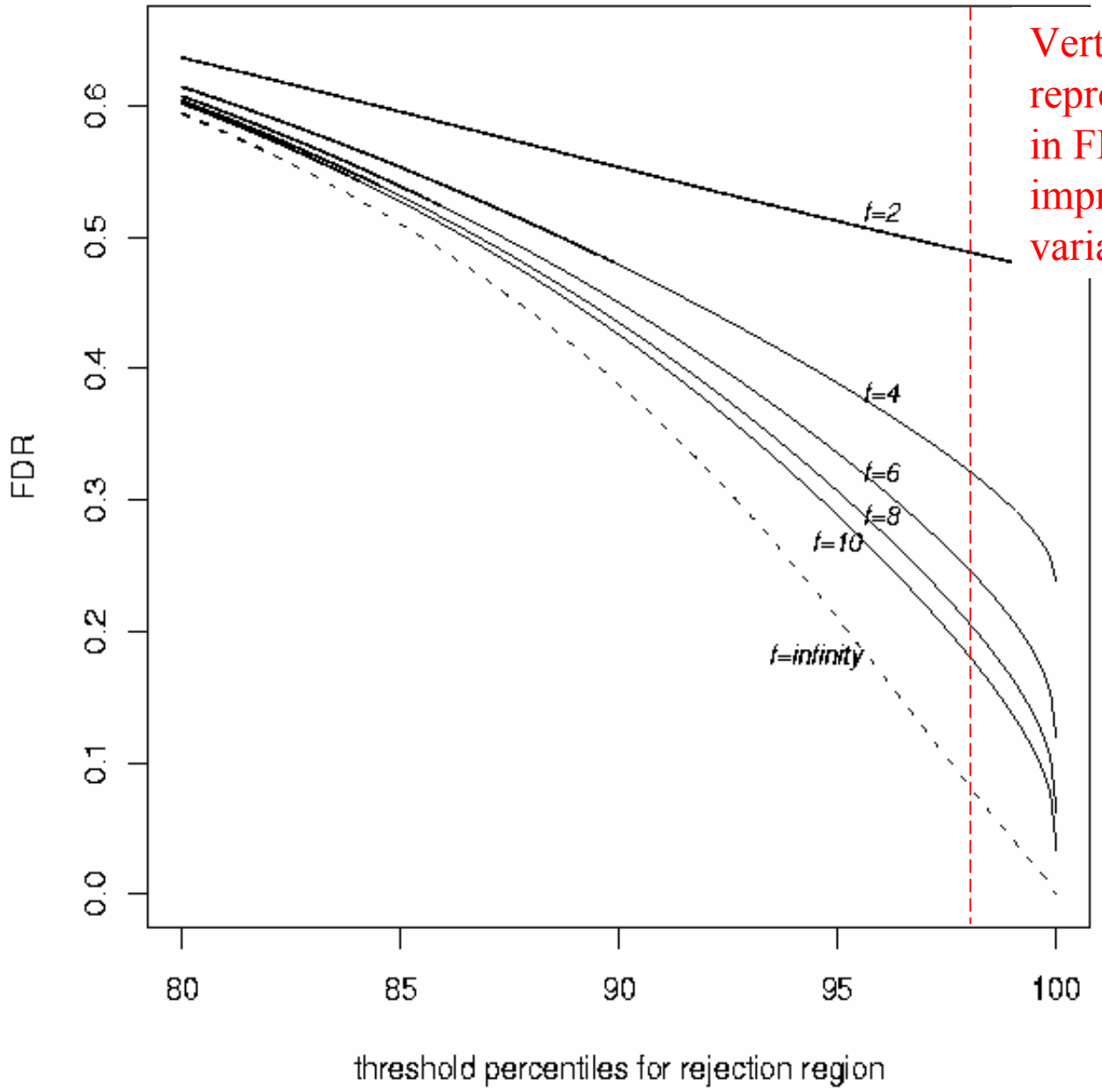




FDR curves for t^0 with $\pi_0 = 0.9$, $\delta' = 2$



FDR curves for t^0 *with* $\pi_0 = 0.9$, $\delta' = 2$

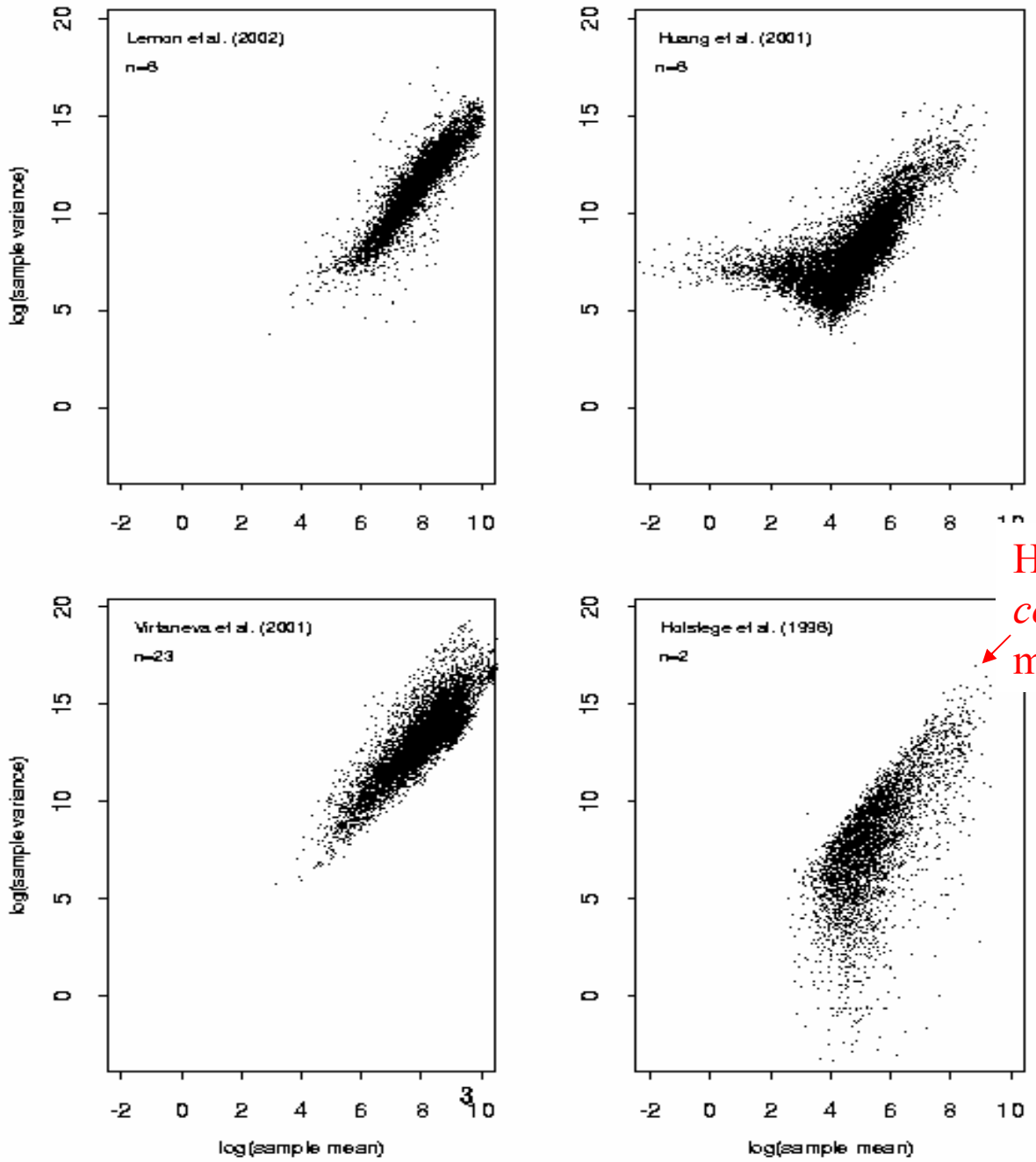


Vertical movement represents improvement in FDR purely due to improvement in variance estimates

Motivation

- The variance estimation plays an important role.
- A roughly linear relationship between $\log(s^2)$ and $\log(\overline{x})$ is observed.

$\log(s^2)$ vs. $\log(\bar{x})$



Heteroscedasticity is *consistent* with the model

Statistical model

- Under each experimental condition, the expression index estimate x_{ij} from $N(\mu_i, \sigma_i^2)$.
- Linear model is proposed,

$$\log(\sigma_i^2) = \alpha_0 + \alpha_1 \log(\mu_i) + \eta_i$$

  η_i, a gene-specific random effect term, from $N(0, \xi^2)$.
- Maximum likelihood estimation:
  We assume the hierarchical model $x_{ij} \mid \mu_i, \sigma_i^2 \sim N(\mu_i, \sigma_i^2)$,
  $\log \sigma_i^2 \mid \mu_i \sim N(\alpha_0 + \alpha_1 log(\mu_i), \xi^2)$,

$$L(\mu, \alpha_0, \alpha_1, \xi^2) = \prod_{i=1}^{m} \prod_{j=1}^{n} \int_{\eta_i} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp(-(x_{ij} - \mu_i)^2 / 2\sigma_i^2) \frac{1}{\sqrt{2\pi\xi^2}} \exp(-\eta_i^2 / 2\xi^2) d\eta_i$$

where σ_i^2 is substituted by $\exp(\alpha_0 + \alpha_1 \log(\mu_i) + \eta_i)$.

Parameter estimation

- Table 1: MLEs for the four data sets

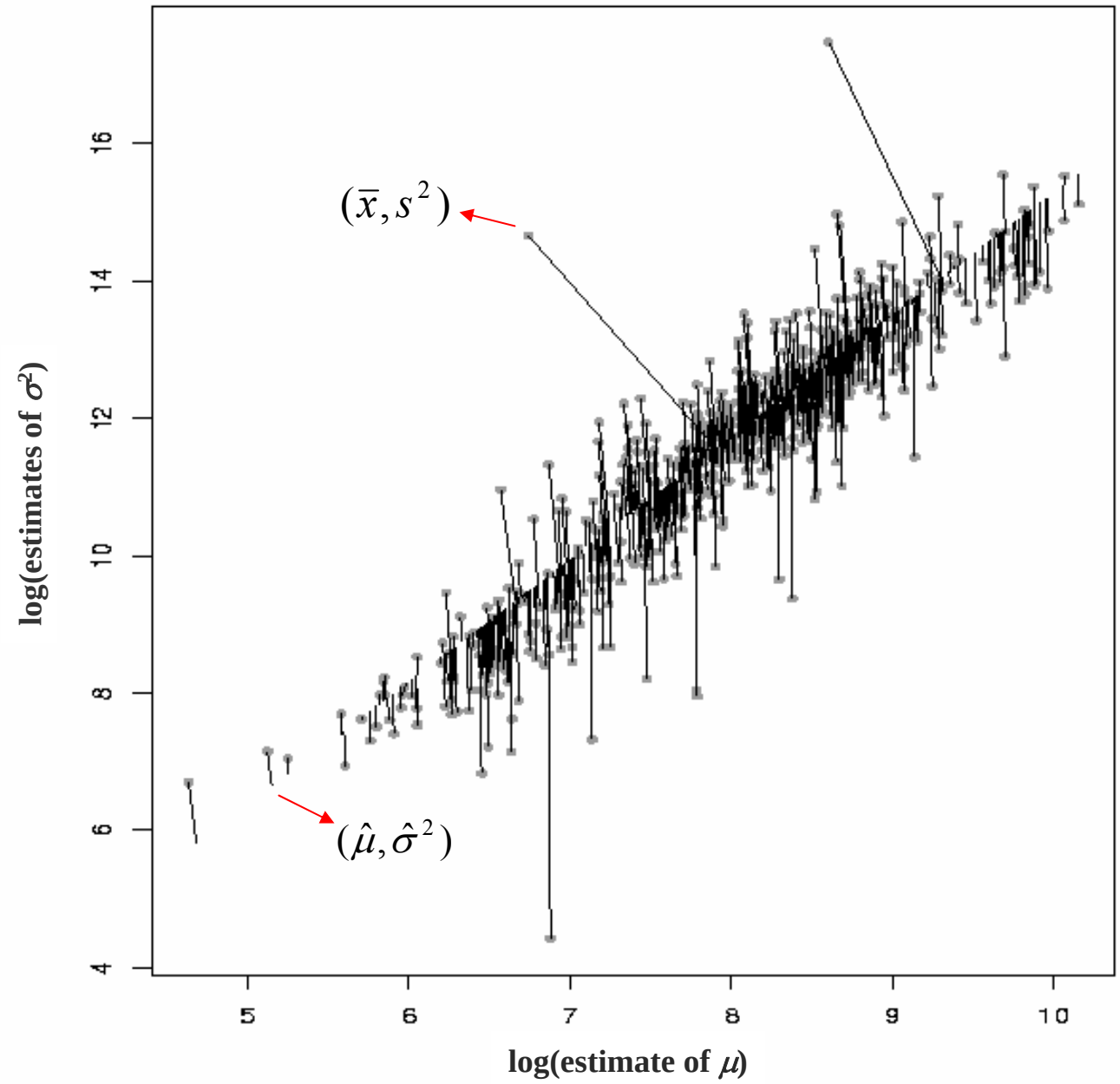
	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\xi}^2$
Lemon et al.	-5.853	1.697	0.049
Huang et al.	-5.08	2.27	0.80
Virtaneva et al.	-2.803	2.028	0.271
Bro et al.	-0.010	1.518	0.00

- The "best predictor" approach (McCulloch and Searle, 2001) is used to predict η_i:

$$BP(\eta_i) = E(\eta_i \mid x_i) = \int_{\eta_i} \eta_i p(\eta_i \mid x_i; \mu_i, \alpha_0, \alpha_1, \xi^2) d\eta_i$$

Then σ^2 can be predicted based on the fixed parameter estimates and predicted η_i.

The shrinkage effect in estimating μ, σ^2



- Eventually, MLE of δ_i can be obtained,

$$\hat{\delta}_i = \frac{\hat{\mu}_{1i} - \hat{\mu}_{2i}}{\sqrt{\frac{\hat{\sigma}_{1i}^2}{n_1} + \frac{\hat{\sigma}_{2i}^2}{n_2}}}$$

Small sample efficiency results

- The MSE of $\hat{\delta}_i$ nearly achieves the Cramér-Rao lower bound

$$CRLB(\hat{\delta}_i) = (1 + b'(\delta_i))^2 / \imath_{\cdot}(\delta_i),$$

$b(\delta_i) = \delta_i - E_{\delta_i}(\hat{\delta}_i)$, Fisher information $\imath_{\cdot}(\delta_i)$.

- Define

$$B = \frac{\imath_{\cdot}^{-1}(\delta_i)}{MSE(\hat{\delta}_i)} \leq \frac{CRLB(\hat{\delta}_i)}{MSE(\hat{\delta}_i)} \leq 1$$

B represents a conservative lower bound for the efficiency of $\hat{\delta}_i$.

- B nearly reaches 1, with no overdispersion (small sample size).

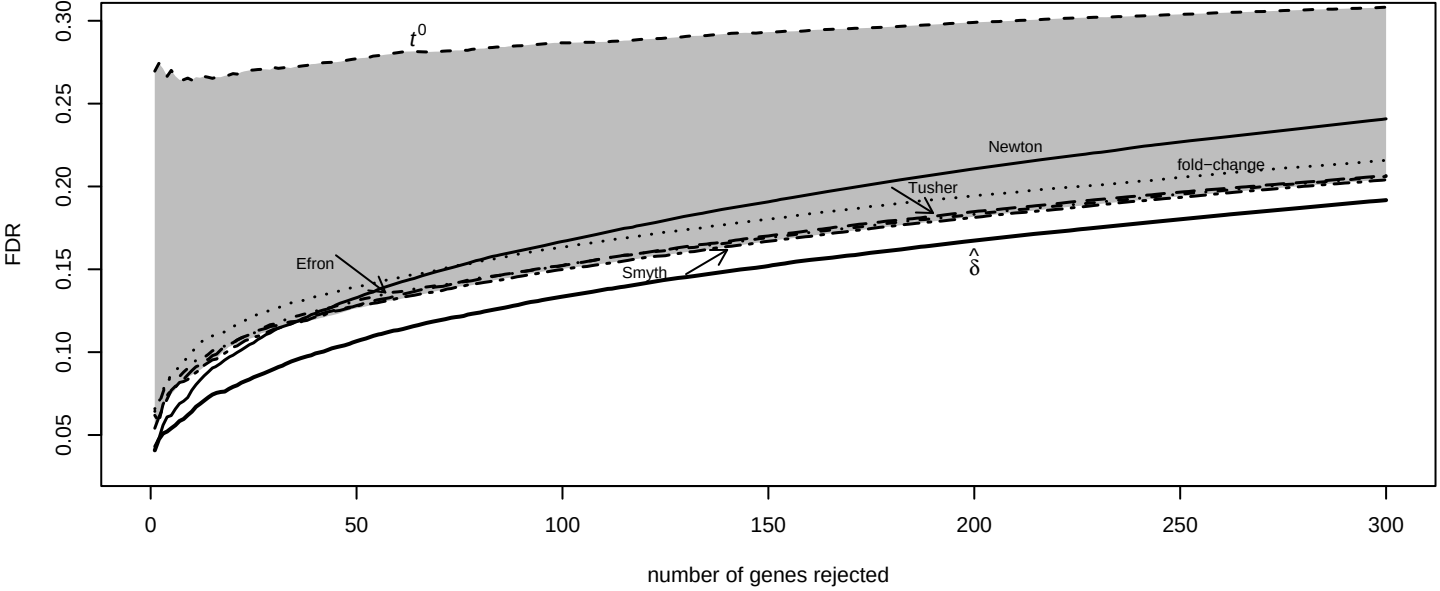
- The situation of overdispersion is more complex.

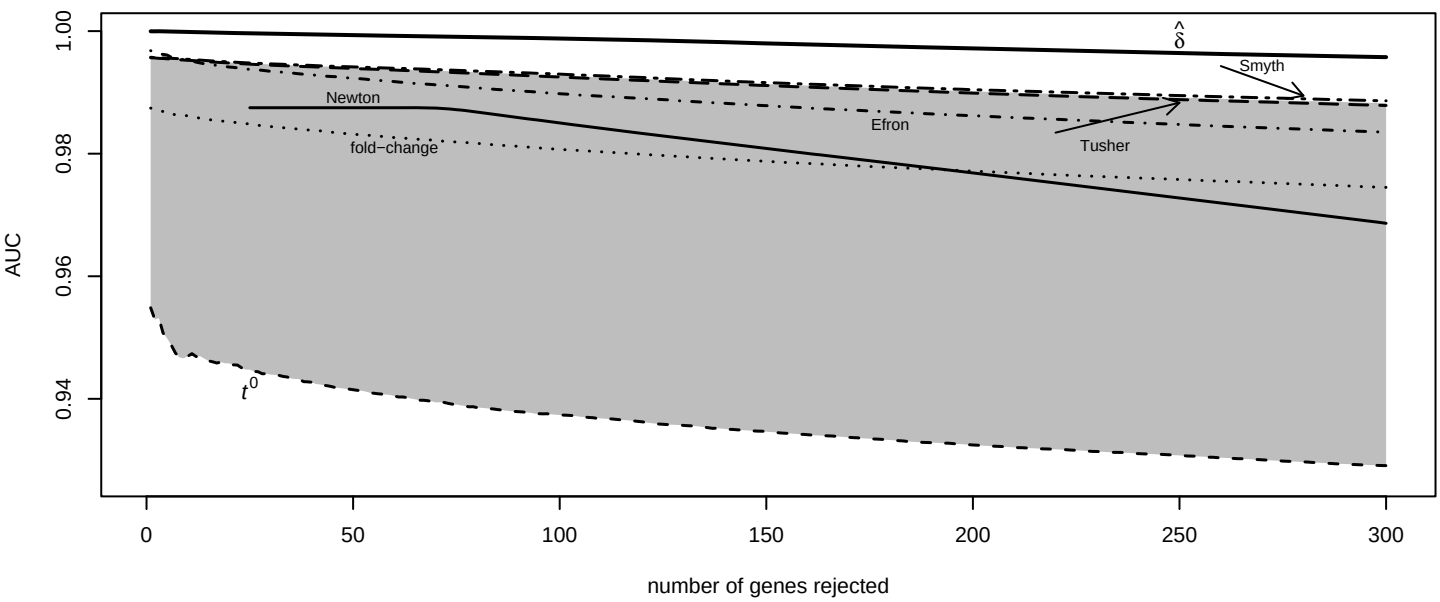
Criterion of comparing the statistics

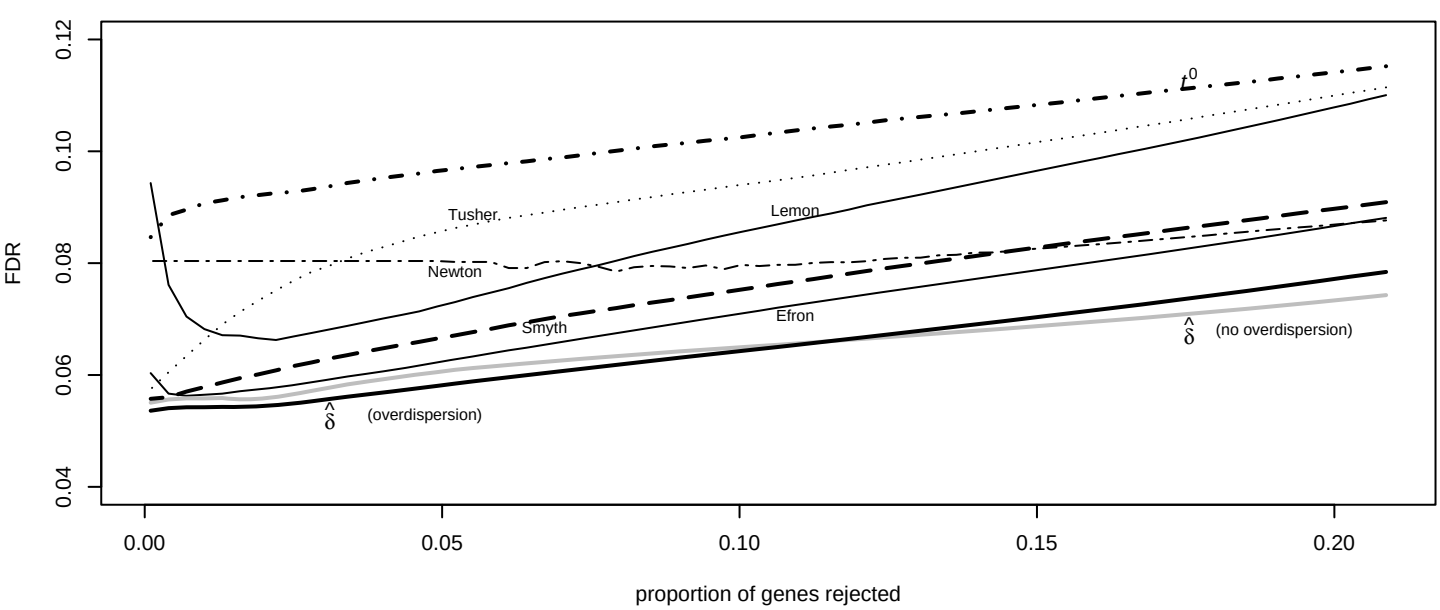
- General criterion: how well the statistics preserve the rank order of the δ_i.
- Discrete δ_i:
  FDR is used, fixing the number of rejected genes.
- Continuous δ_i:
  - Receiver-operator characteristic curve (ROC),
    comparing the conditional distribution of the δs for $T < c$ vs. $T \geq c$.
  - The area under the ROC curve (AUC), $Pr\,(\delta^R > \delta^A)$,
    the most commonly used summary.
- Relationship between FDR and AUC for discrete δ:
  Theorem 2 *AUC = 1/2 (1 – FDR + true accept/accept)*

Simulation study

- Common set-up:
  - $n_1 = n_2 = 3$, $m = 10000$.
  - In each case, 2000 simulations are performed.
  - Choices of a:
    - percentiles 25, 50, 75, 90 and 100 of s, and 2max(s)
    - $a = 0,\ \infty$
    - criterion proposed in **Tusher et al. (2001)**.
- Discrete δ:
  - "realized" FDR (**Genovese and Wasserman, 2002**)
  - a case of $\delta' = 1$ and $\pi_0 = 0.7$
- Continuous δ:
  Sampling from a double exponential distribution with location 0 and variance 1/2.


FDR
0.05
0.10
0.15
0.20
0.25
0.30
t^0
Newton
fold−change
Tusher
Efron
Smyth
$\hat{\delta}$
0
50
100
150
200
250
300
number of genes rejected


AUC
0.94
0.96
0.98
1.00
$\hat{\delta}$
Smyth
Newton
Efron
Tusher
fold−change
t^0
0
50
100
150
200
250
300
number of genes rejected

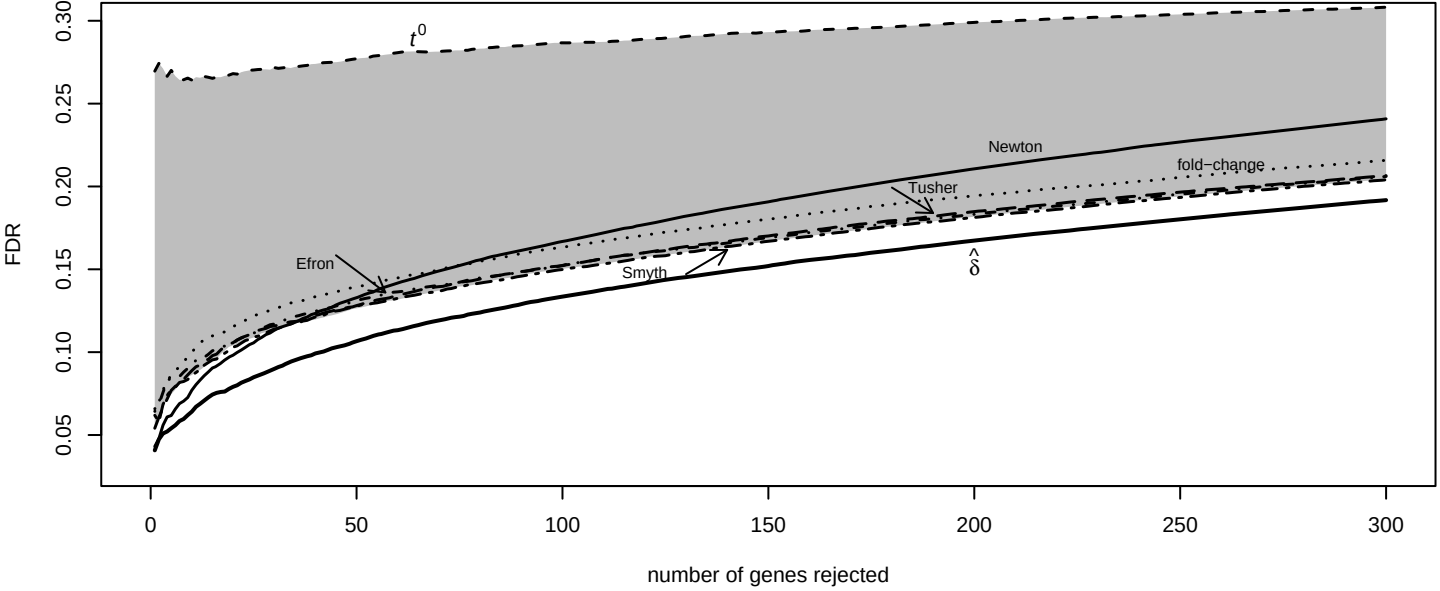

FDR
0.04
0.06
0.08
0.10
0.12
t^0
Tusher.
Lemon
Newton
Smyth
Efron
$\hat{\delta}$ (no overdispersion)
$\hat{\delta}$ (overdispersion)
0.00
0.05
0.10
0.15
0.20
proportion of genes rejected

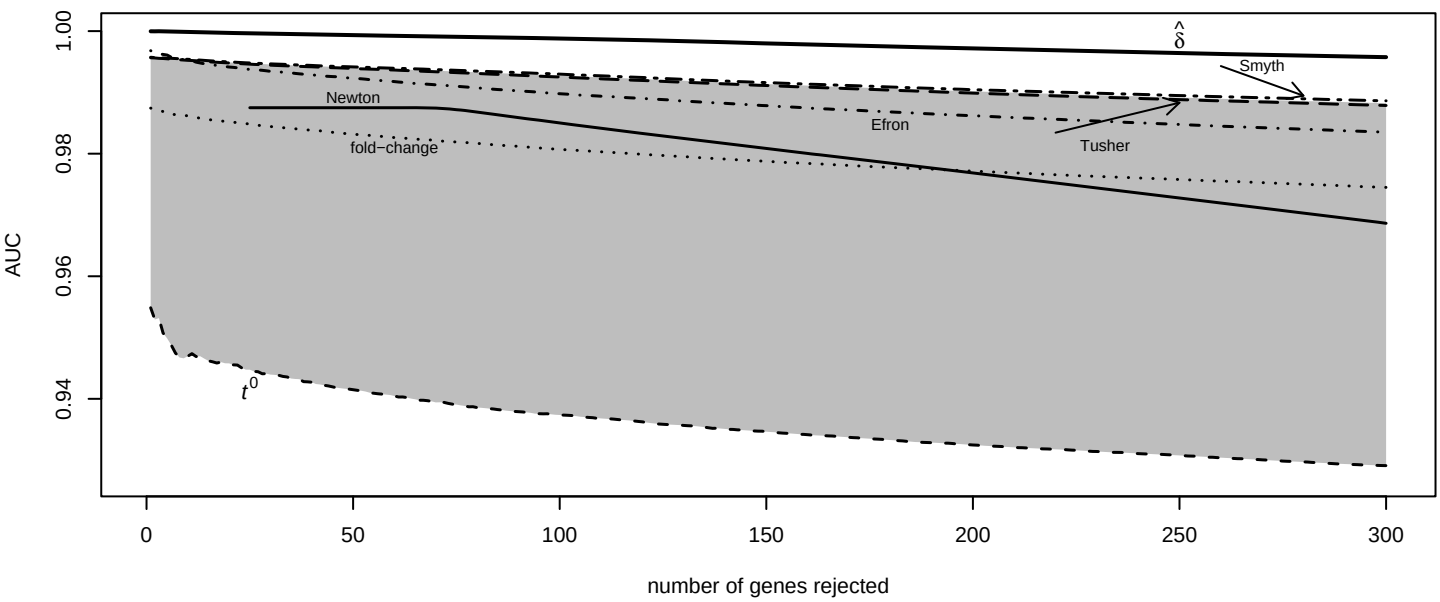
A case study

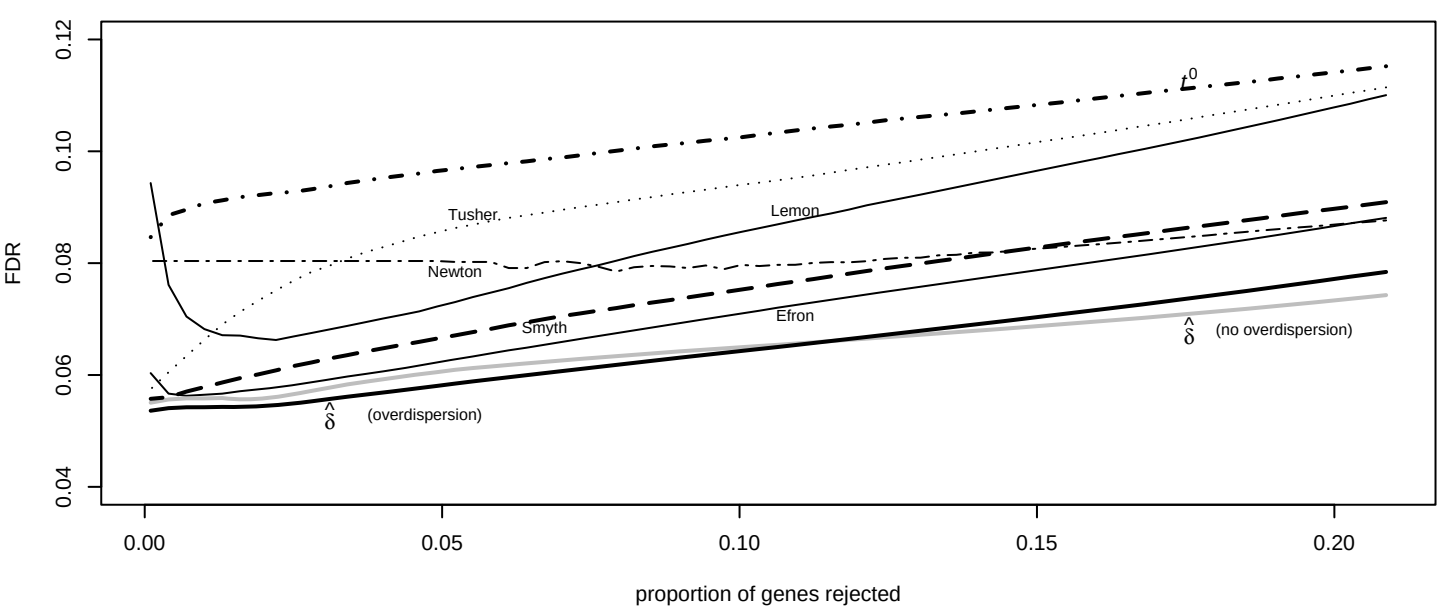
- Real data: A human fibroblast cell expression data set from **Lemon et al. (2002)** (6 vs. 6, under stimulated to 50:50 conditions)

- Small sample comparisons: $n_1 = n_2 = 2$

- Permutation procedures to estimate the FDR (**Storey and Tibshirani, 2001**)
  - "observed" distribution - $\binom{6}{2}\binom{6}{2}$ comparisons of two conditions.
  - an empirical "null" distribution - $\binom{12}{2}\binom{10}{2}$ total comparisons of 2 vs. 2 arrays.


FDR
0.05
0.10
0.15
0.20
0.25
0.30
t^0
Newton
fold−change
Tusher
Efron
Smyth
$\hat{\delta}$
0
50
100
150
200
250
300
number of genes rejected


AUC
0.94
0.96
0.98
1.00
$\hat{\delta}$
Smyth
Newton
Efron
Tusher
fold−change
t^0
0
50
100
150
200
250
300
number of genes rejected


FDR
0.04
0.06
0.08
0.10
0.12
t^0
Tusher.
Lemon
Newton
Smyth
Efron
$\hat{\delta}$ (no overdispersion)
$\hat{\delta}$ (overdispersion)
0.00
0.05
0.10
0.15
0.20
proportion of genes rejected