Week 5 notes:

- Journal club: signups via pull request
- Projects: some ideas
- Part 1 of the guts of limma (linear models for microarrays)

Journal club: signups via pull request

- Can be as early as next week
- Pull request to the README.md of the materials repo
- That way, the order of priority is determined
- Can have up to 2 per day, try to avoid 13th November, 11th December
- Fill in to the desired cell: title of paper w/ hyperlink to journal/preprint landing page + the initials of the speakers.

23.10.2017	Mark	limma 2		
30.10.2017	Hubert	RNA-seq quantification		
06.11.2017	Mark	edgeR+friends 1		
13.11.2017	Charlotte	hands-on session #1: RNA-seq	X	X
20.11.2017	Mark	edgeR+friends 2		
27.11.2017	Hubert	classification		
04.12.2017	Mark	single-cell		
11.12.2017	Gosia	hands-on session #2: mass cytometry	X	X
18.12.2017	Mark	epigenomics, DNA methylation, ChIP data, gene set analysis		

Project ideas

- As always, reproducing analyses from a paper or designing your own simulation to evaluate some methods is always a possibility
- I will put pressure on in a few weeks
- “Consulting” type possibilities:
  - Comparing fixed effects and mixed effects models for the paired comparison problem
  - Comparing properties of Nanopore and Illumina cDNA sequencing data (gene expression)
  - Comparing the design matrix and 2-group implementations in DRIM-Seq

From the feed: Terry's IMS Bulletin + "Over-optimism"

We will see a lot of methods in this course – **how do we evaluate what works well in practice ?**

http://bulletin.imstat.org/2012/11/terences-stuff-does-it-work-in-practice/

Gene expression Advance Access publication June 26, 2010

Over-optimism in bioinformatics: an illustration

Monika Jelizarow[1], Vincent Guillemot[1,2], Arthur Tenenhaus[2], Korbinian Strimmer[3] and Anne-Laure Boulesteix[1,*]

[1]Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany, [2]SUPELEC Sciences des Systèmes (E3S)-Department of Signal Processing and Electronics Systems - 3, rue Joliot Curie, Plateau de Moulon, 91192 Gif-sur-Yvette Cedex, France and [3]Department of Medical Informatics, Statistics and Epidemiology, University of Leipzig, Härtelstr. 16-18, 04107 Leipzig, Germany

Associate Editor: John Quackenbush

"if the improvement of a quantitative criterion such as the error rate is the main contribution of a paper, the superiority of new algorithms should always be demonstrated on independent validation data."

In class exercise + discussion

- (5 minutes) Read the excerpt from “Terence’s Stuff” column
- (5-10 minutes; discuss with your neighbour) Answer the following 3 questions:
  1. How do we tell what works in practice?
  2. What problems arise using simulated (synthetic) data?
  3. What problems arise using real data?
  4. What are positive/negative controls?
- Discuss
- If simulation: what metrics could/would/should we use?

Differential expression, small sample inference

- Table of data (e.g., microarray gene expression data with replicates of each of condition A, condition B)
  - rows = features (e.g., genes), columns = experimental units (samples)
- Most common problem in statistical bioinformatics: want to infer whether there is a change in the response —> a statistical test for each row of the table.

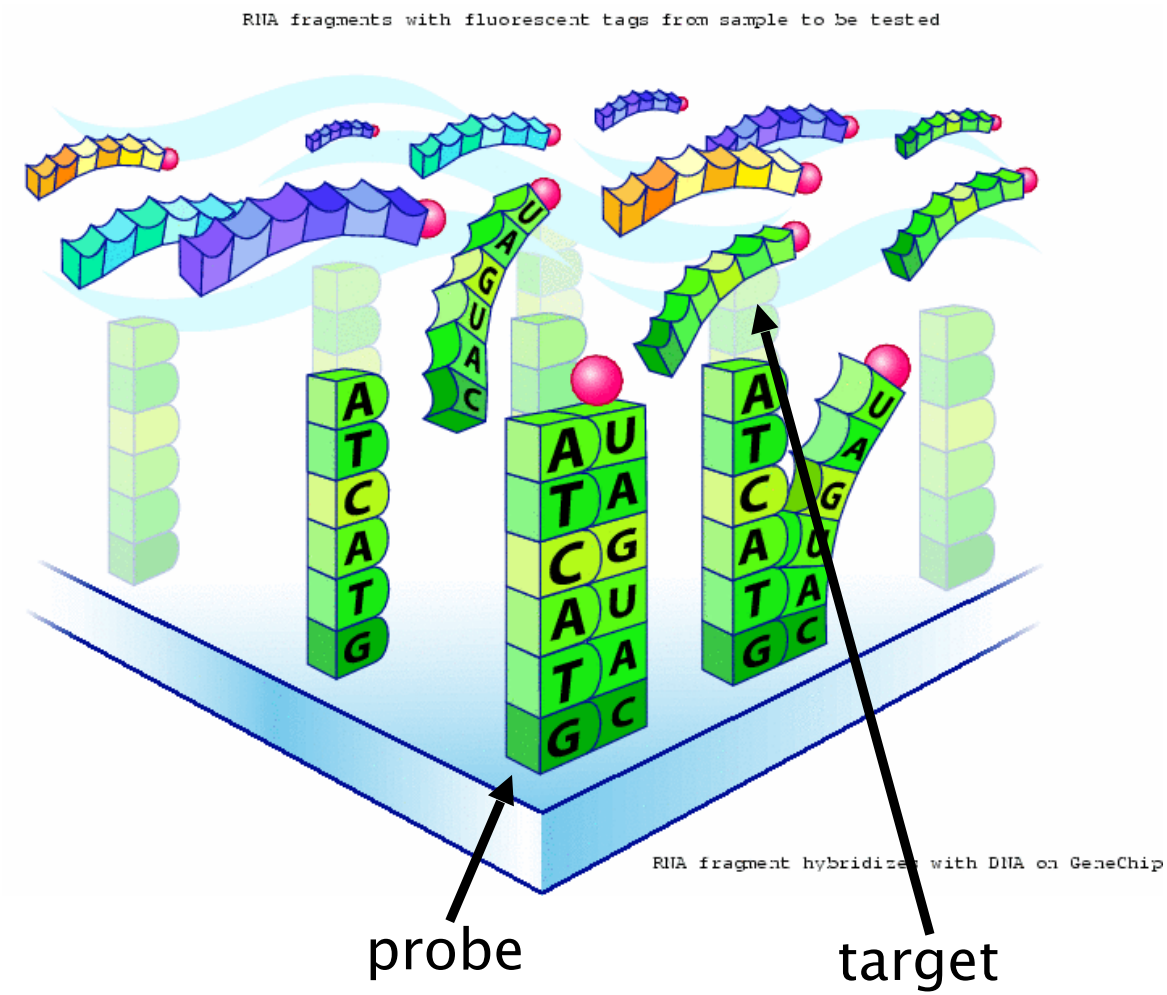
What test might you use? Why is this hard? What issues arise? How much statistical power is there [1] ?

```
> head(y)
        group0     group0      group0     group1     group1      group1
gene1 -0.1874854  0.2584037 -0.05550717 -0.4617966 -0.3563024 -0.03271432
gene2 -3.5418798 -2.4540999  0.11750996 -4.3270442 -5.3462622 -5.54049106
gene3 -0.1226303  0.9354707 -1.10537767 -0.1037990  0.5221678 -1.72360854
gene4 -2.3394536 -0.3495697 -3.47742610 -3.2287093  6.1376670 -2.23871974
gene5 -3.7978820  1.4545702 -7.14796503 -4.0500796  4.7235714 10.00033769
gene6  1.4627078 -0.3096070 -0.26230124 -0.7903434  0.8398769 -0.96822312
```

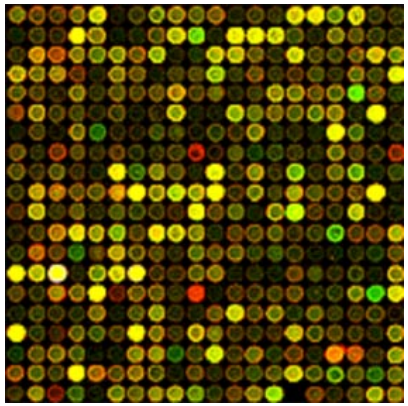
[1] http://www.stat.ubc.ca/~rollin/stats/ssize/n2.html

University of Zurich[UZH]
Institute of Molecular Life Sciences

DNA microarray: arrays of northern blots

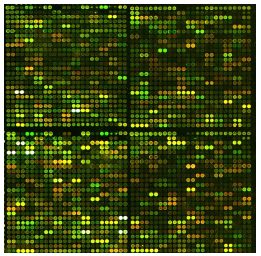


Abundance (of complementary DNA species) measured by flouresence intensity



Microarray expression measures

Two-colour



$$y_{ga} = \log_2(R/G)$$

array: g

probe or gene: a

Affymetrix

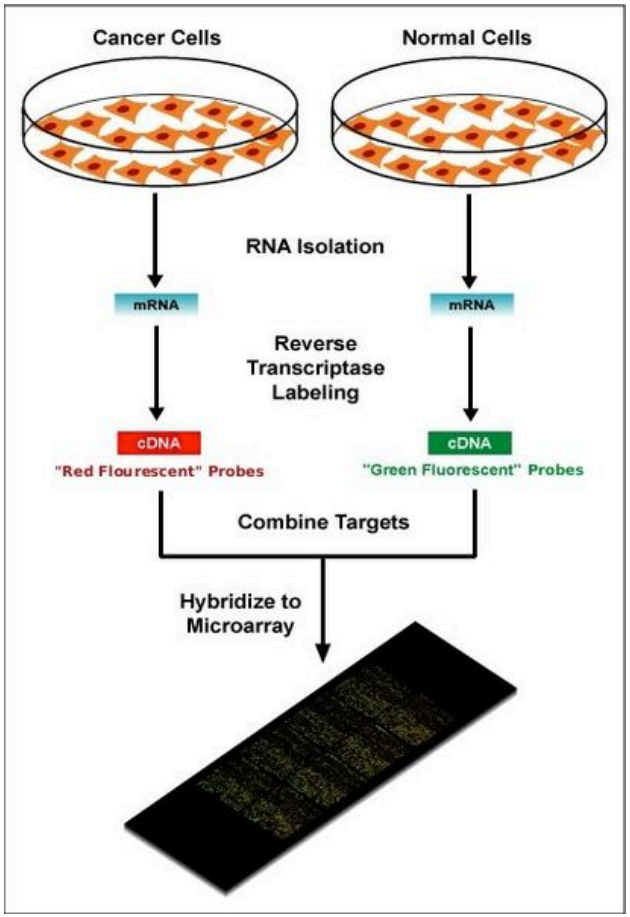
y_{ga} = log-intensity (summarized over probes)

Illumina

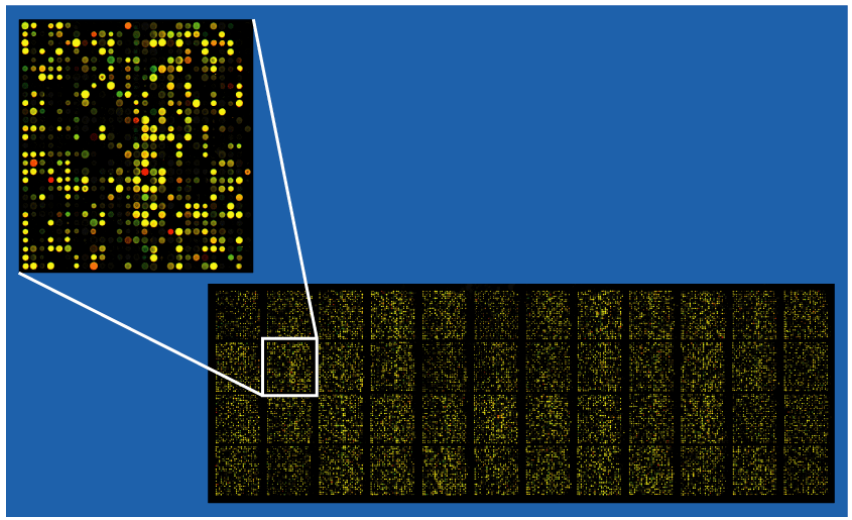
y_{ga} = log-intensity (summarized over beads)

Questions of Interest

- What genes have changed in expression? (e.g. between disease/normal, affected by treatment) Gene discovery, differential expression
- Is a specified group of genes all up-regulated in a particular condition? Gene **set** differential expression
- Can the expression profile predict outcome? Class prediction, classification
- Are there tumour sub-types not previously identified? Do my genes group into previously undiscovered pathways? Class discovery, clustering



Two colour microarrays



http://en.wikipedia.org/wiki/DNA_microarray

Preprocessing: additive + multiplicative error model

Observe intensity for one probe on one array

Intensity = background + signal

$$I = B + S$$

additive errors

multiplicative errors

This idea underlies variance stabilizing transformations vsn (two colour data) and vst (for Illumina data)

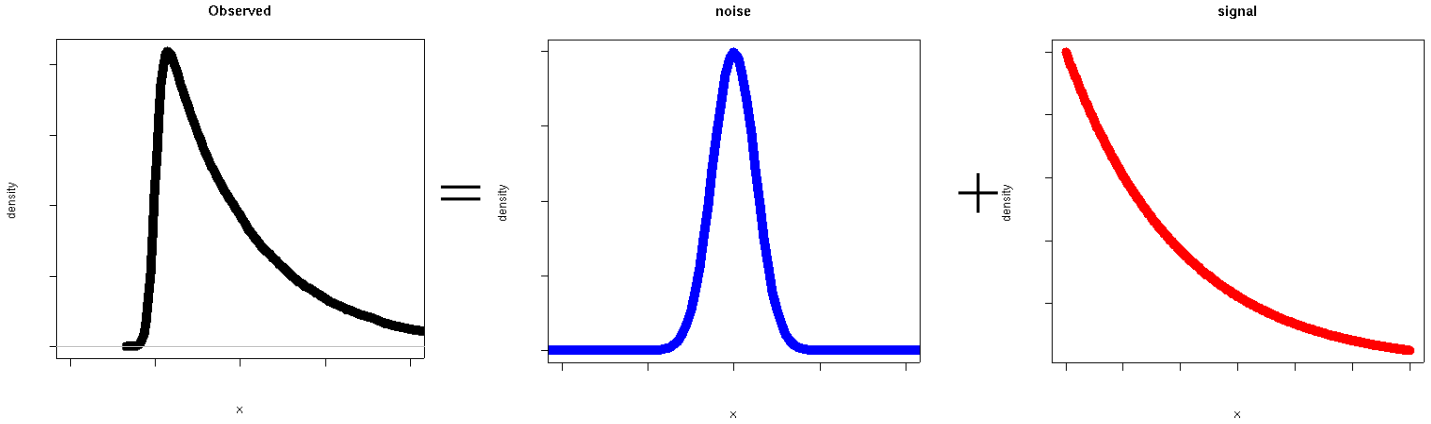
normexp convolution model

Microarray background correction: maximum likelihood estimation for the normal–exponential convolution

JEREMY D. SILVER
Bioinformatics Division, Walter and Eliza Hall Institute, Parkville 3050, Victoria, Australia and Department of Biostatistics, University of Copenhagen, Øster Farimagsgade 5, Entrance B, PO Box 2099, DK-1014 Copenhagen K, Denmark
j.silver@biostat.ku.dk
MATTHEW E. RITCHIE
Department of Oncology, University of Cambridge, Cambridge CB2 0RE, UK
GORDON K. SMYTH*
Bioinformatics Division, Walter and Eliza Hall Institute, Parkville 3050, Victoria, Australia
smyth@wehi.edu.au

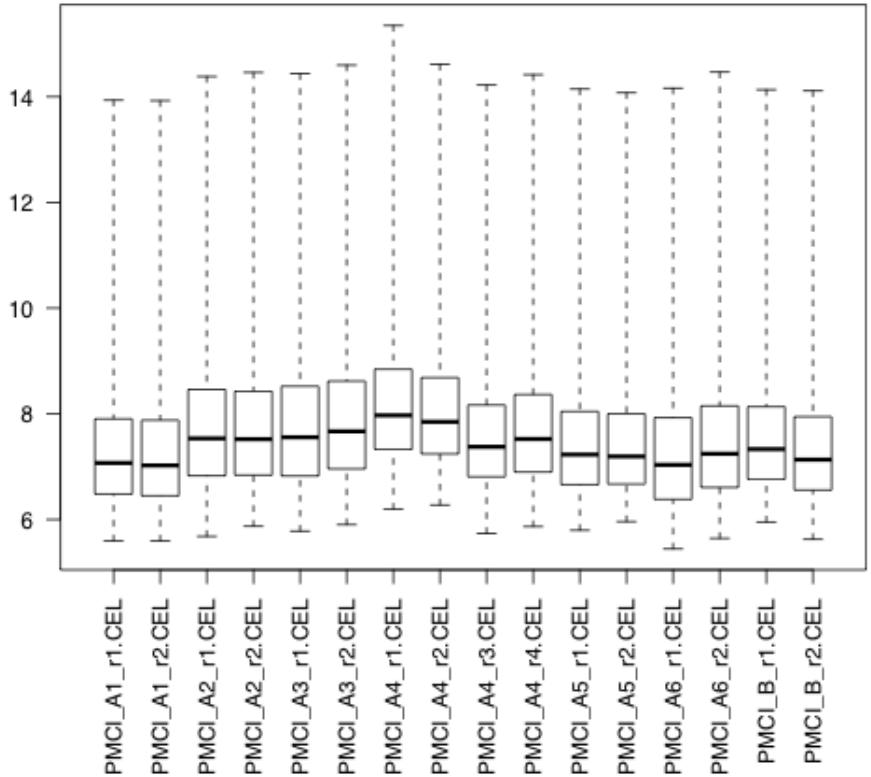
Intensity = Background + Signal

$N(\mu,\sigma^2)$ Exponential(α)

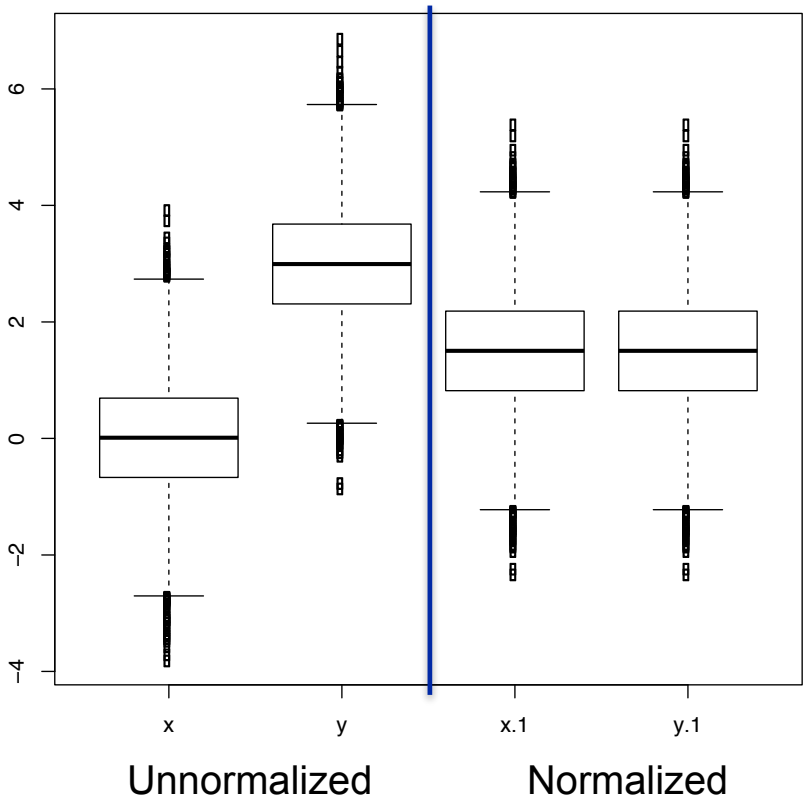


Normalization: one-colour

Similarly for single channel data, adjustments need to be made for all samples to be comparable.



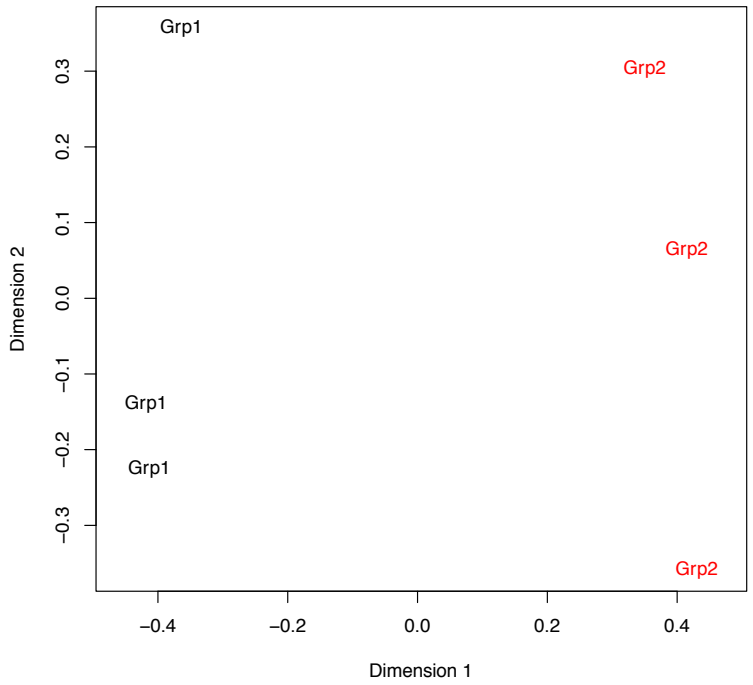
Quantile normalization




```
x <- rnorm(10000, mean=0, sd=1)
y <- rnorm(10000, mean=3)
z <- cbind(x,y)

# create "reference" distribution
s <- apply(z,2,sort)
sm <- rowMeans(s)

# impose ref. distribution by ranks
r <- apply(z,2,rank)
n <- apply(r,2,function(u) sm[u])

boxplot( data.frame(x=x,y=y,n) )

#> library(limma)
#> zn <- normalizeQuantiles(z)
#> all(zn==n)
#[1] TRUE
```

Quality assessments

Multidimensional scaling plot



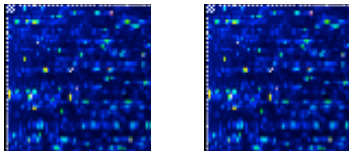

```
sd <- 0.3*sqrt(4/rchisq(1000,df=4))
x <- matrix(rnorm(1000*6,sd=sd),1000,6)
x[1:50,4:6] <- x[1:50,4:6] + 2

mds <- plotMDS(x)

> round(mds$distance.matrix,3)
      [,1]  [,2]  [,3]  [,4] [,5] [,6]
[1,] 0.000 0.000 0.000 0.000 0.00    0
[2,] 0.835 0.000 0.000 0.000 0.00    0
[3,] 0.850 0.793 0.000 0.000 0.00    0
[4,] 1.089 1.068 1.058 0.000 0.00    0
[5,] 1.050 1.058 1.072 0.863 0.00    0
[6,] 0.991 1.047 1.046 0.865 0.85    0
```

Limma concept: borrowing information across genes

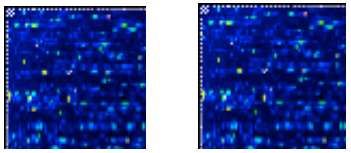
- Small data sets: few samples, generally under-powered for 1 gene
- Curse of dimensionality: many tests, need to adjust for multiple testing (= loss of power)
- Benefit of parallelism: same model is fit for every gene. Can borrow information from one gene to another
  - Hard: assume parameters are constant across genes
  - Soft: smooth genewise parameters towards a common value in a graduated way, e.g., Bayes, empirical Bayes, Stein shrinkage …

A very common experiment (1-colour)

Mutant x 2



WT x 2



Gene X

Which genes are differentially expressed?

$n_1 = n_2 = 2$ Affymetrix arrays

~30,000 probe-sets

Ordinary t-tests (1-colour)

$$t_g = \frac{\overline{y}_{\mathrm{mu}} - \overline{y}_{\mathrm{wt}}}{s_g\, c}$$

give very high false discovery rates

$$c = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Residual df = 2

t-tests with common variance

$$t_{g,\text{pooled}} = \frac{\overline{y}_{\text{mu}} - \overline{y}_{\text{wt}}}{s_0\, c}$$

with residual standard deviation s_0 pooled across genes

More stable, but ignores gene-specific variability

$$c = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

A better compromise

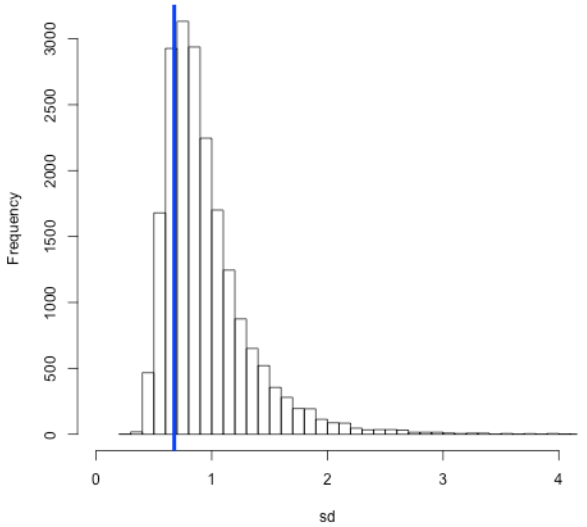
Shrink standard deviations towards common value

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$$

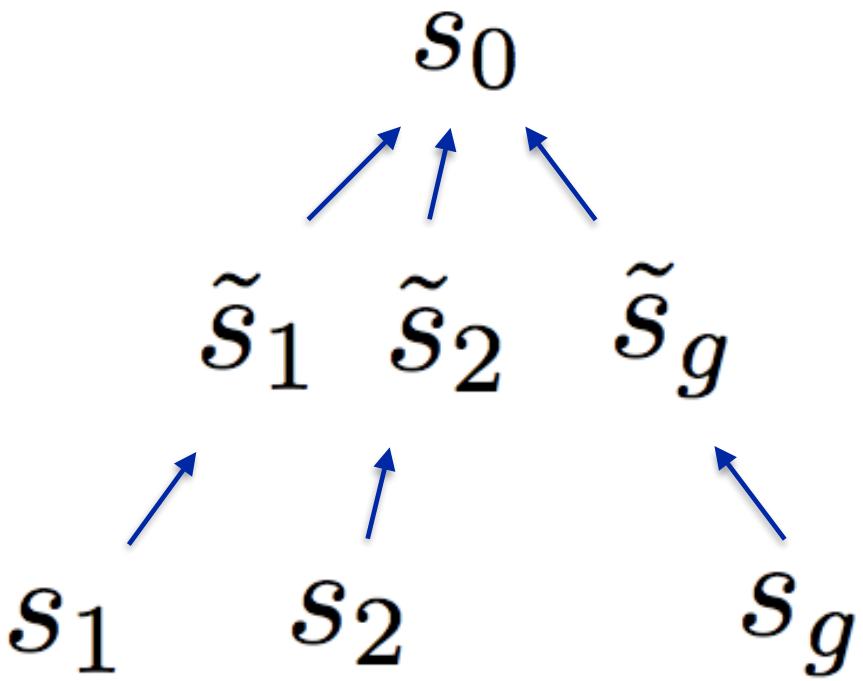
d = degrees of freedom

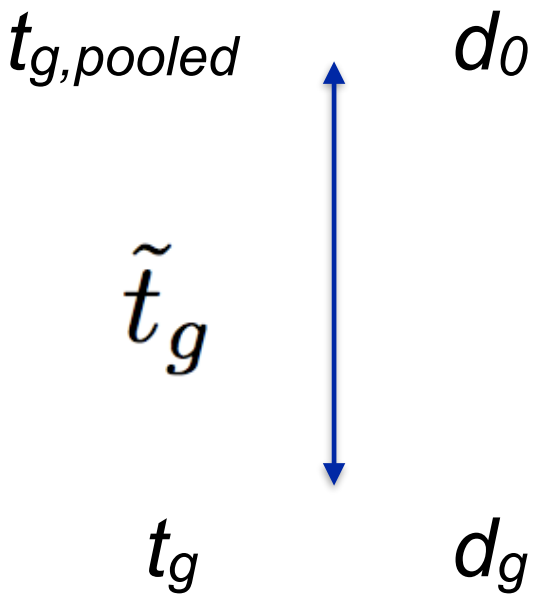
Moderated t-statistics

$$\tilde{t}_g = \frac{\overline{y}_{\text{mu}} - \overline{y}_{\text{wt}}}{\tilde{s}_g \, u}$$



Shrinkage of standard deviations





The **data decides** whether $\tilde{t}_g$ should be closer to $t_{g,pooled}$ or t_g

Why does it work?

- We learn what is the typical variability level by looking at all genes, but allow some flexibility from this for individual genes
- Adaptive – data (through hyperparameter estimates, d_0 and s_0) suggests how much to "squeeze" toward common value

Hierarchical model for variances

Data

$$s_g^2 \sim \sigma_g^2 \frac{\chi^2_{d_g}}{d_g}$$

Prior

$$\frac{1}{\sigma_g^2} \sim s_0^2 \frac{\chi^2_{d_0}}{d_0}$$

Posterior

$$E\left(\frac{1}{\sigma_g^2} \mid s_g^2\right) = \frac{d_0 + d_g}{s_0^2 d_0 + s_g^2 d_g}$$

Posterior Statistics

Posterior variance estimators

$$\tilde{s}_g^2 = \frac{s_0^2 d_0 + s_g^2 d_g}{d_0 + d_g}$$

Moderated t-statistics

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

Baldi & Long 2001, Wright & Simon 2003, Smyth 2004

Exact distribution for moderated t

An unexpected piece of mathematics shows that, under the null hypothesis,

$$\tilde{t}_g \sim t_{d_0+d_g}$$

The degrees of freedom add!

The Bayes prior in effect adds d_0 extra arrays for estimating the variance.

Wright and Simon 2003, Smyth 2004

Aside: Marginal Distributions to calculate

Under usual likelihood model, s_g is independent of the estimated coefficients.

Under the hierarchical model, s_g is independent of the moderated t-statistics instead

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

Multiple testing and adjusted p-values

- Each statistical test has an associated false error rate
- Traditional method in statistics is to control family wise error rate, e.g., by Bonferroni.
- Controlling the false discovery rate (FDR) is more **appropriate** in microarray studies
- Benjamini and Hochberg method controls expected FDR for independent or weakly dependent test statistics. Simulation studies support use for genomic data.
- All methods can be implemented in terms of adjusted p-values.

Linear Models

- In general, need to specify:
  - Dependent variable
  - Explanatory variables (experimental design, covariates, etc.)
- More generally:

$$y = X\beta + \epsilon$$

vector of observed data

design matrix

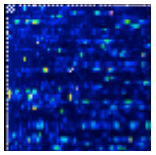
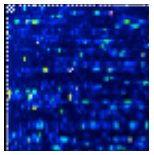
Vector of parameters to estimate

Linear Models for microarrays

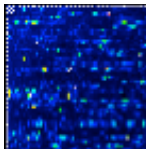
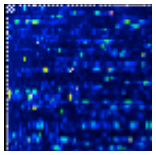
- Analyse all arrays together combining information in optimal way
- Combined estimation of precision
- Extensible to arbitrarily complicated experiments
- Design matrix: specifies RNA targets used on arrays
- Contrast matrix: specifies which comparisons are of interest

Design → Linear models

WT x 2

Mutant x 2

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

β_1 = wt log-expression

β_2 = mutant – wt

$E[y_1]=E[y_2]=\beta_1$ $E[y_3]=E[y_4]=\beta_1+\beta_2$