Journal club

Papers to be selected by 18.00 on 16th October; please discuss it with Hubert and I before submitting pull request.

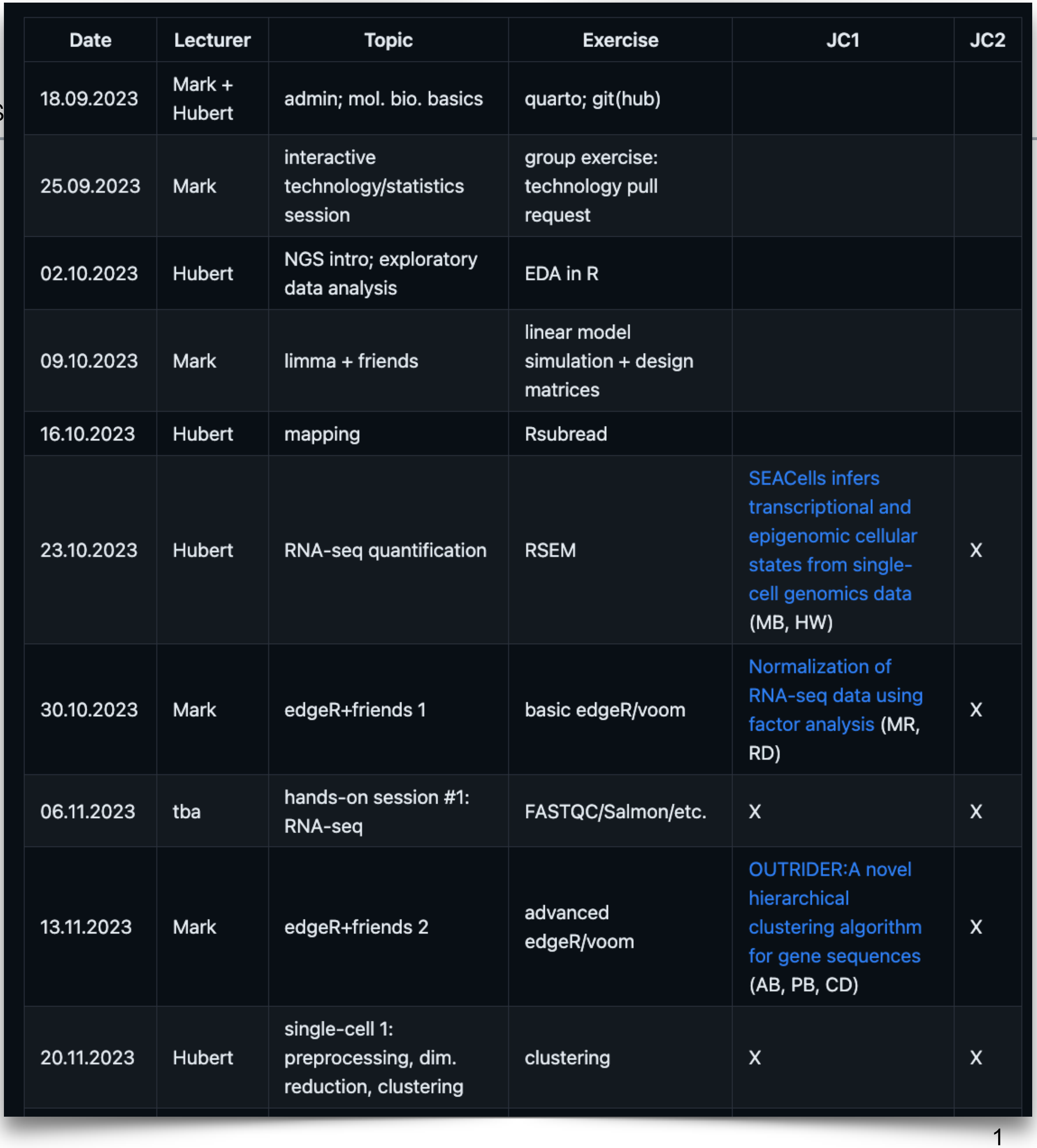
Start: Oct 23

Journal Club schedule to be finalized by 23rd October

Given the number of students, groups of 2-3.

Use the #journal_clubs channel (e.g., to find a group member). Hubert and I will put some suggestions there.

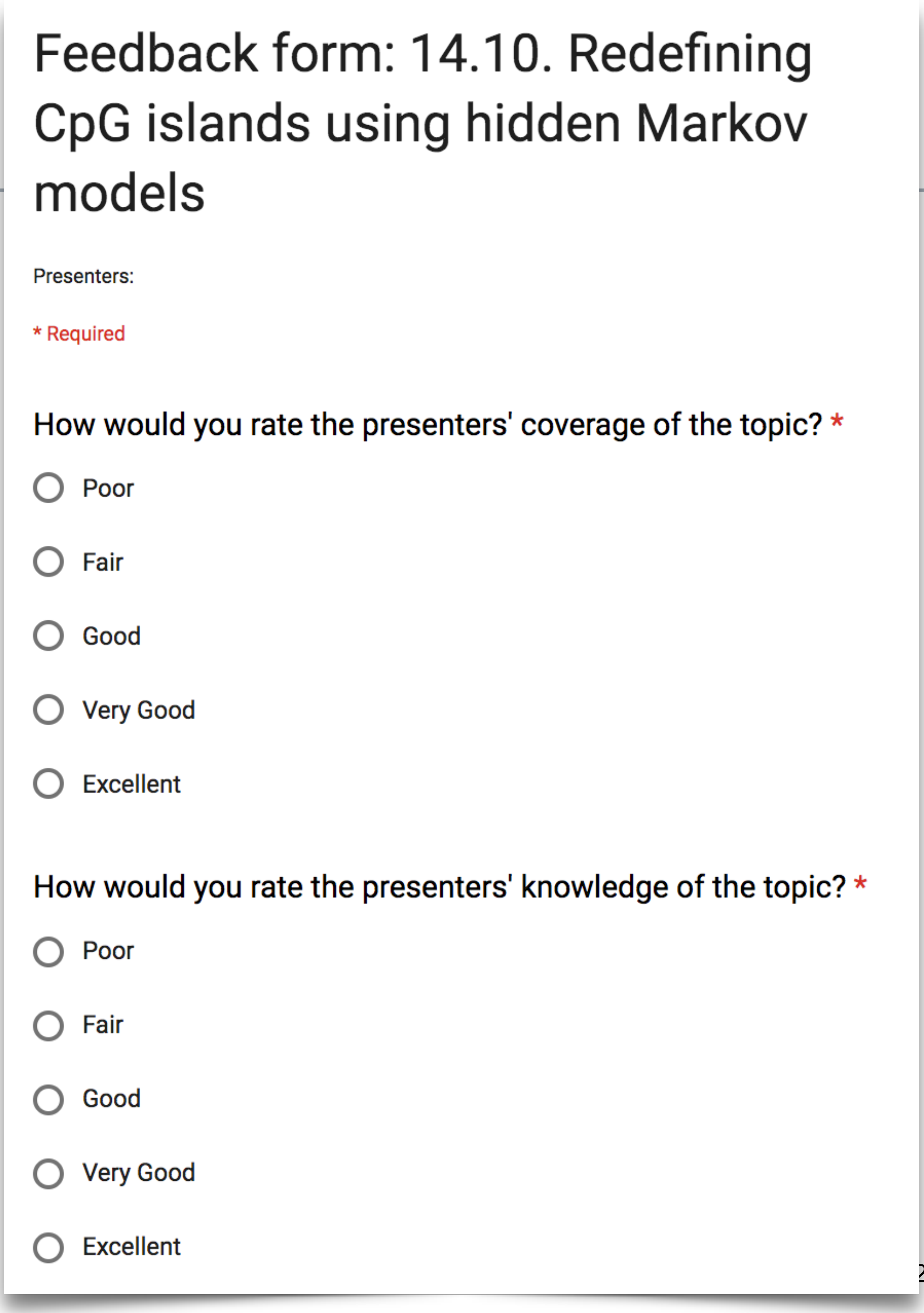
Sign up by pull request to the 'material' repo, README.md. "First come first served"



Date	Lecturer	Topic	Exercise	JC1	JC2
18.09.2023	Mark + Hubert	admin; mol. bio. basics	quarto; git(hub)		
25.09.2023	Mark	interactive technology/statistics session	group exercise: technology pull request		
02.10.2023	Hubert	NGS intro; exploratory data analysis	EDA in R		
09.10.2023	Mark	limma + friends	linear model simulation + design matrices		
16.10.2023	Hubert	mapping	Rsubread		
23.10.2023	Hubert	RNA-seq quantification	RSEM	SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data (MB, HW)	X
30.10.2023	Mark	edgeR+friends 1	basic edgeR/voom	Normalization of RNA-seq data using factor analysis (MR, RD)	X
06.11.2023	tba	hands-on session #1: RNA-seq	FASTQC/Salmon/etc.	X	X
13.11.2023	Mark	edgeR+friends 2	advanced edgeR/voom	OUTRIDER:A novel hierarchical clustering algorithm for gene sequences (AB, PB, CD)	X
20.11.2023	Hubert	single-cell 1: preprocessing, dim. reduction, clustering	clustering	X	X

Journal Club procedure

- During/after journal clubs: give the presenters some constructive feedback
- Giving feedback (via Google form) is part of your JC grade! Feedback forms must be submitted within 1 week of presentation; comments will be sent to presenters (anonymously)
- JCs will be part of the recordings



Feedback form: 14.10. Redefining CpG islands using hidden Markov models

Presenters:

* Required

How would you rate the presenters' coverage of the topic? *

○ Poor

○ Fair

○ Good

○ Very Good

○ Excellent

How would you rate the presenters' knowledge of the topic? *

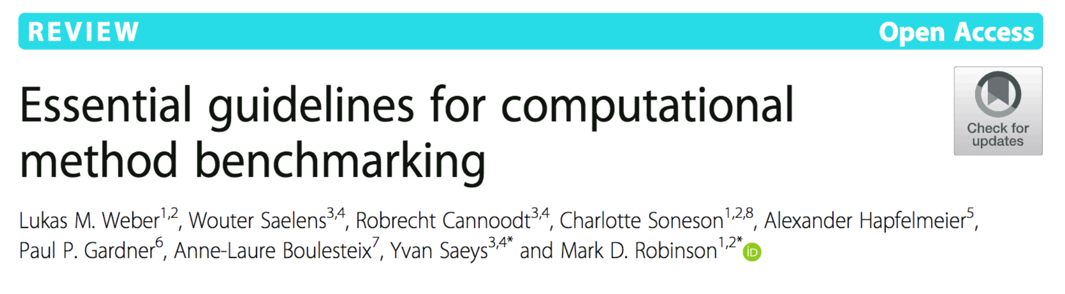
○ Poor

○ Fair

○ Good

○ Very Good

○ Excellent



REVIEW Open Access

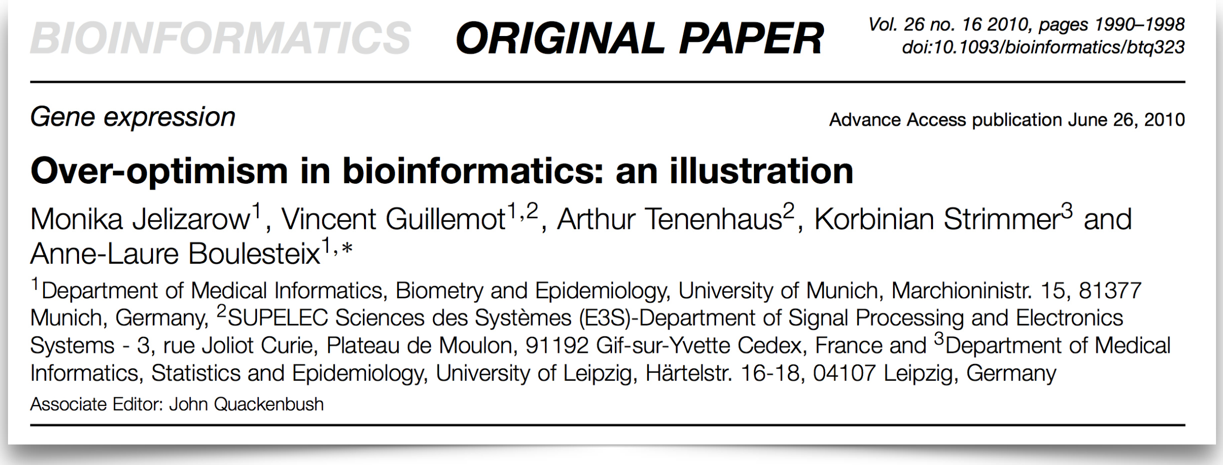
Essential guidelines for computational method benchmarking

Check for updates

Lukas M. Weber[1,2], Wouter Saelens[3,4], Robrecht Cannoodt[3,4], Charlotte Soneson[1,2,8], Alexander Hapfelmeier[5], Paul P. Gardner[6], Anne-Laure Boulesteix[7], Yvan Saeys[3,4*] and Mark D. Robinson[1,2*]

From the feed: “Over-optimism” + Terry’s IMS Bulletin

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



BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 16 2010, pages 1990–1998
doi:10.1093/bioinformatics/btq323

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.”

CORRESPONDENCE

The self-assessment trap: can we all be better than average?

"researchers wishing to publish their analytical methods are required by referees to compare the performance of their own algorithms against other methodologies, thus being forced to be judge, jury and executioner. The result is that the authors’ method tends to be the best .."

In class discussion

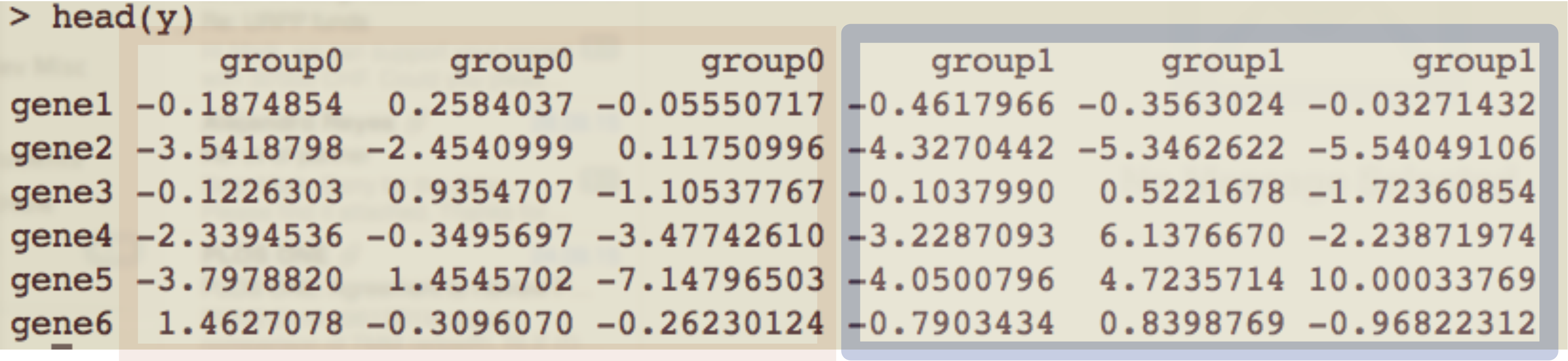
- **5 minutes**: read (the excerpt from “Terence’s Stuff”).
- **10 minutes**: Discuss with your neighbour/row and answer the following 4 questions:
  1. How do we tell what works in practice?
  2. What problems arise using simulated (synthetic) data?
  3. What problems arise using real data?
  4. What are positive/negative controls? Can you give an example positive control in omics data?
- **n.b. include this (method comparison) context in your Journal Club talks**

limma fundamentals

The simplistic view: 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] ?

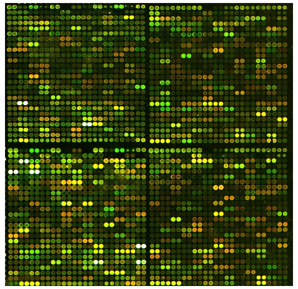


```
> head(y)
          group0     group0      group0     group1     group1      group1
gene1 -0.1874854  0.2584037 -0.05550717 -0.4617966 -0.3563024 -0.03271432
gene2 -3.5418798 -2.4540999  0.11750996 -4.3270442 -5.3462622 -5.54049106
gene3 -0.1226303  0.9354707 -1.10537767 -0.1037990  0.5221678 -1.72360854
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

Microarray expression measures

Two-colour



array

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

probe or gene

Affymetrix

$$y_{ga} =$$

log-intensity (summarized over probes)

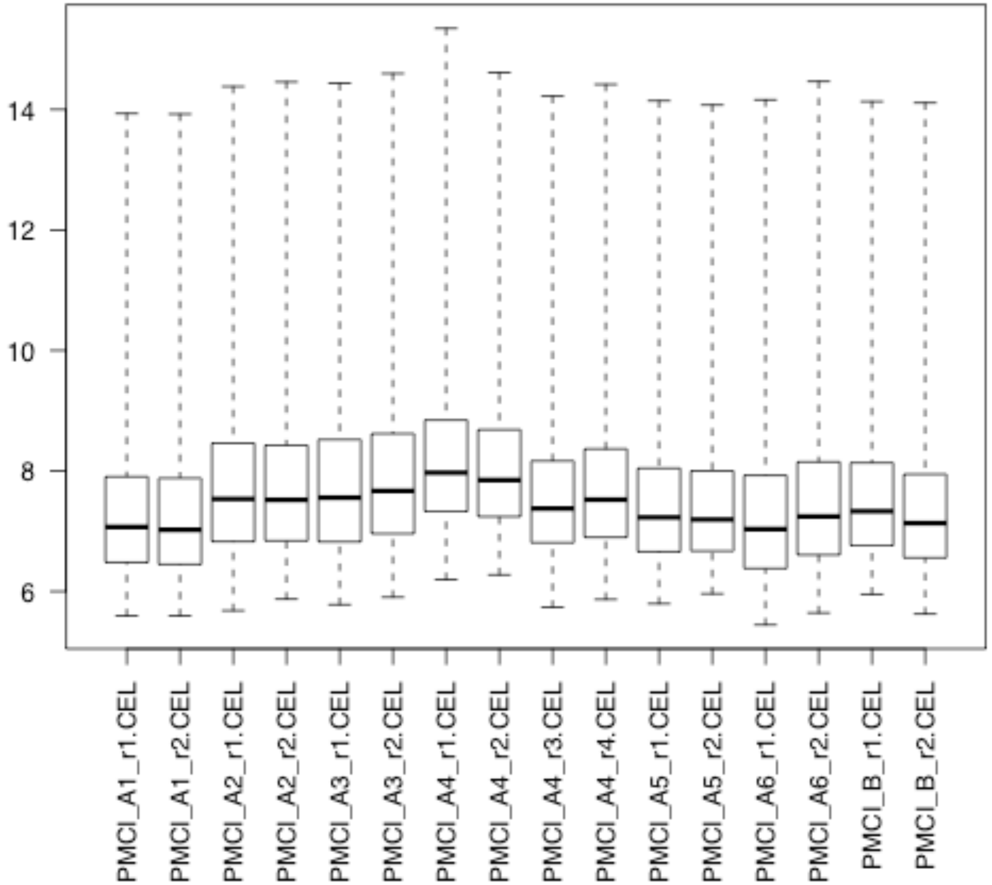
Illumina

$$y_{ga} =$$

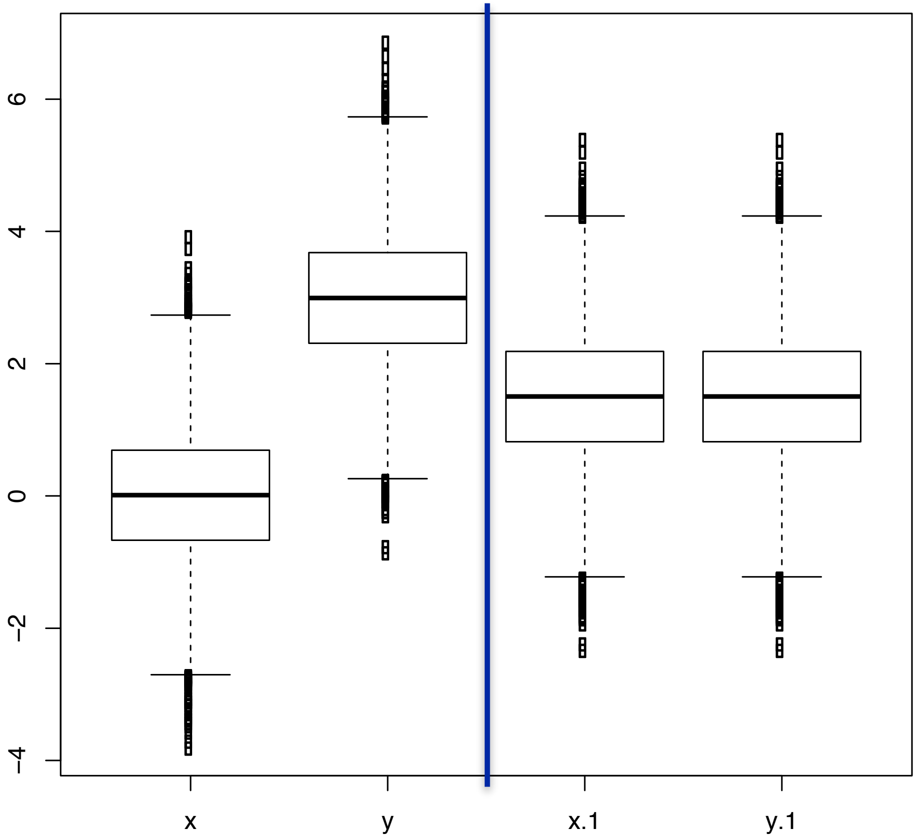
log-intensity (summarized over beads)

Normalization: one-colour

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



Quantile normalization



Unnormalized Normalized

```
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
```

Preprocessing: additive + multiplicative error model

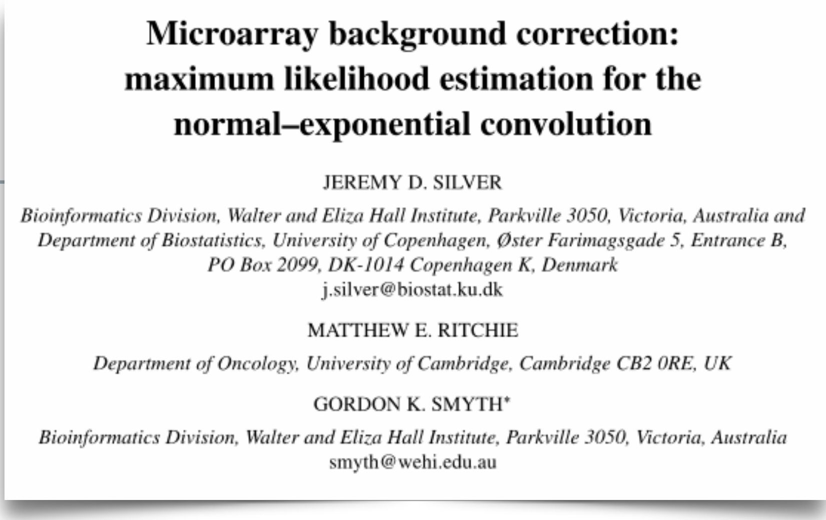
Observe intensity for one probe on one array

Intensity = background + signal

$$I = \underset{\substack{\text{additive}\\ \text{errors}}}{\underset{\nearrow}{B}} + \underset{\text{multiplicative errors}}{\underset{\nwarrow}{S}}$$

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

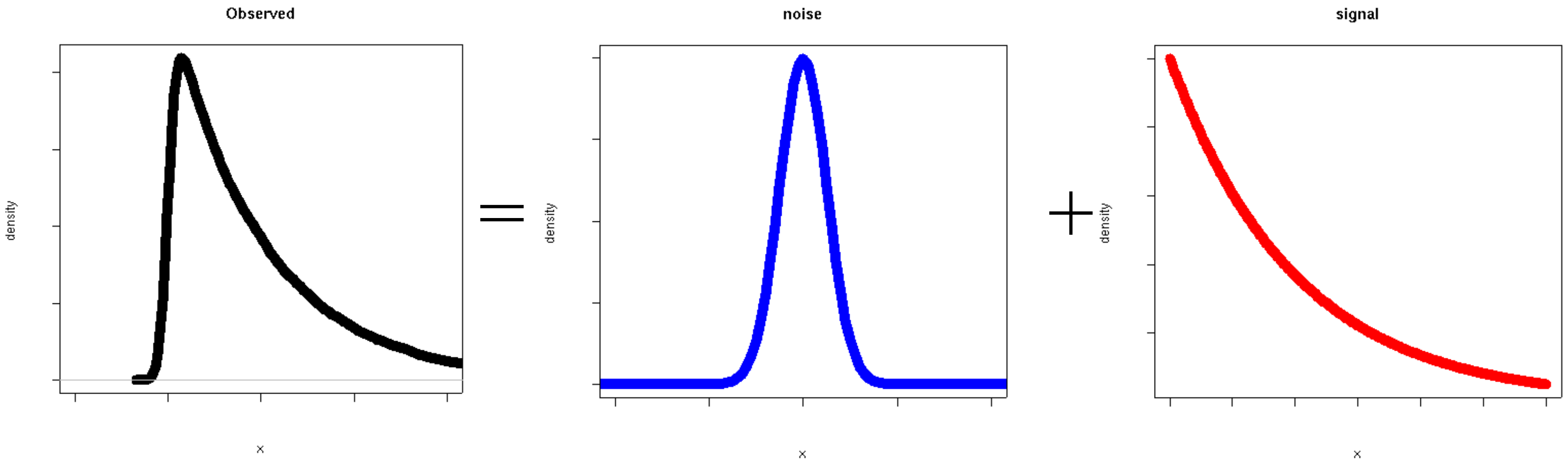
Department of Oncology, University of Cambridge, Cambridge CB2 0RE, UK

GORDON K. SMYTH*

Bioinformatics Division, Walter and Eliza Hall Institute, Parkville 3050, Victoria, Australia
smyth@wehi.edu.au

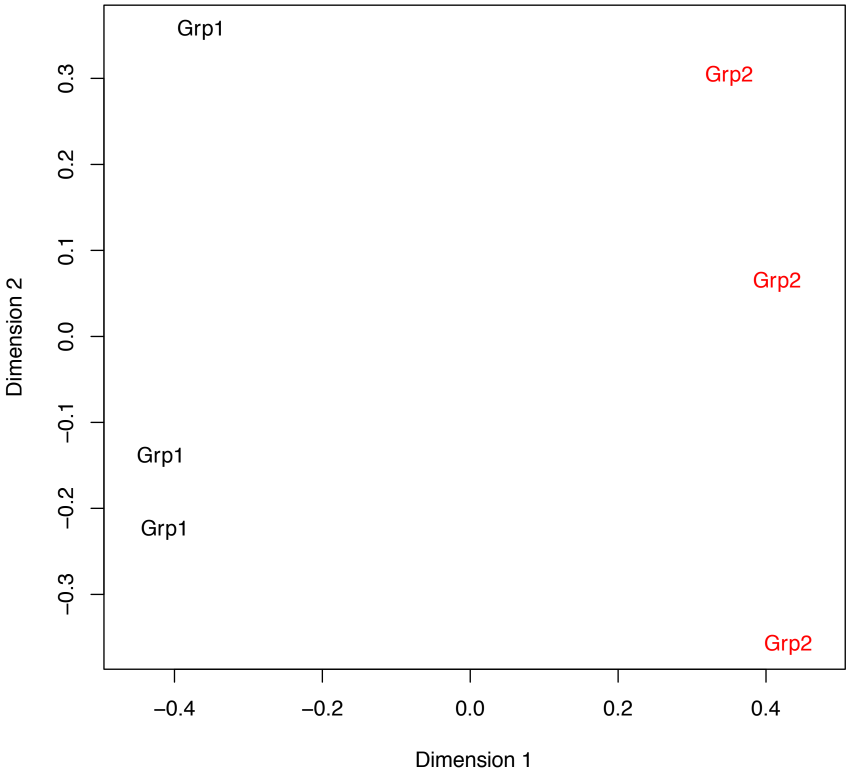
Intensity = Background + Signal

$N(\mu,\sigma^2)$ Exponential(α)



Quality assessments / spot checks

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
```

“To consult the statistician after an experiment is finished is often merely to ask him[her] to conduct a post mortem examination. He[She] can perhaps say what the experiment died of.” R. A. Fisher

Motivation for exploratory data analysis: Case Study

(from Stefano, a former M.Sc. student in my Institute)

He is studying gene expression in fruitfly and is interested in transcriptional responses following “heat shock”.

Basic schematic of experiment:

CTL	**t0**		**t12**		
TRT		**t4**	**t12**	**t24**	**t72**

↑

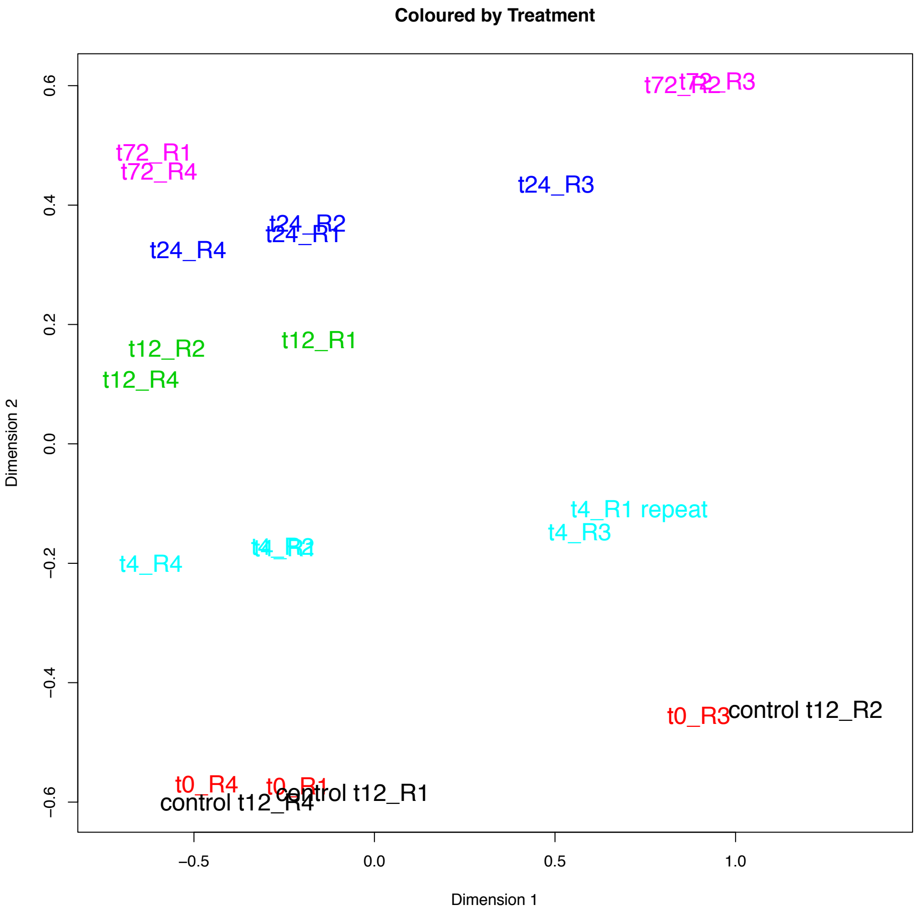
Change to lower temperature.

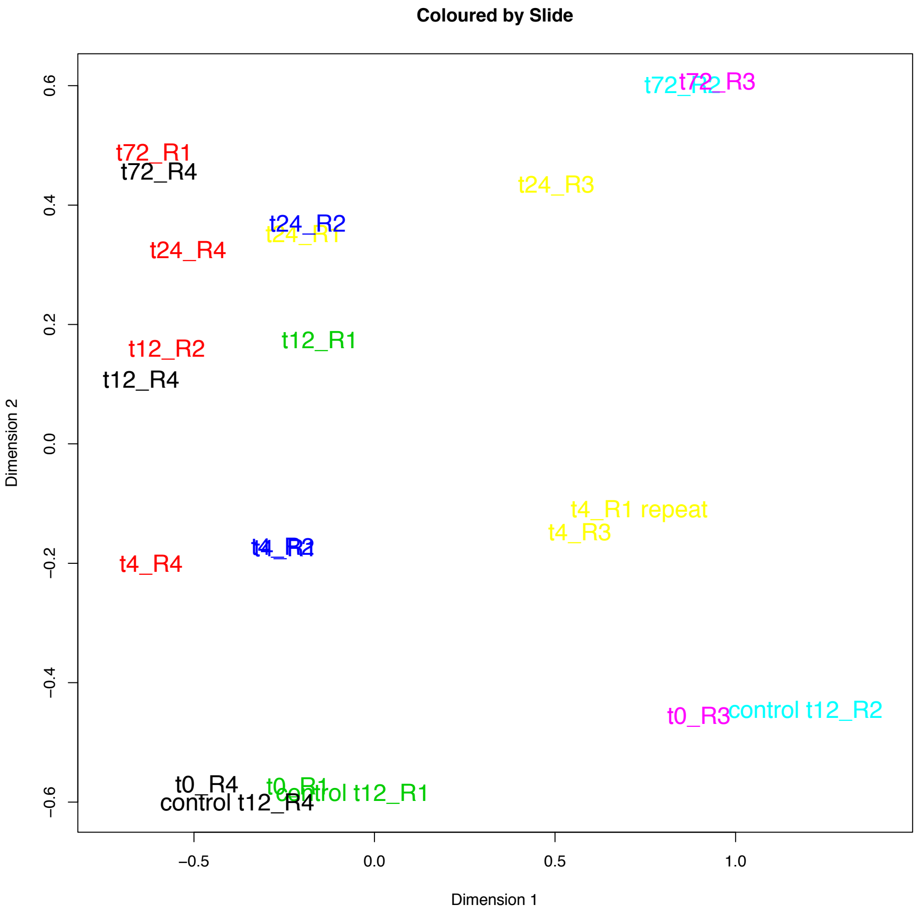
~4 replicates for each condition

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

```
library(limma)
plotMDS(d)  # ‘d’ is a matrix
```

Take a close look at where the replicates are to each other relative to the X- and Y-axes





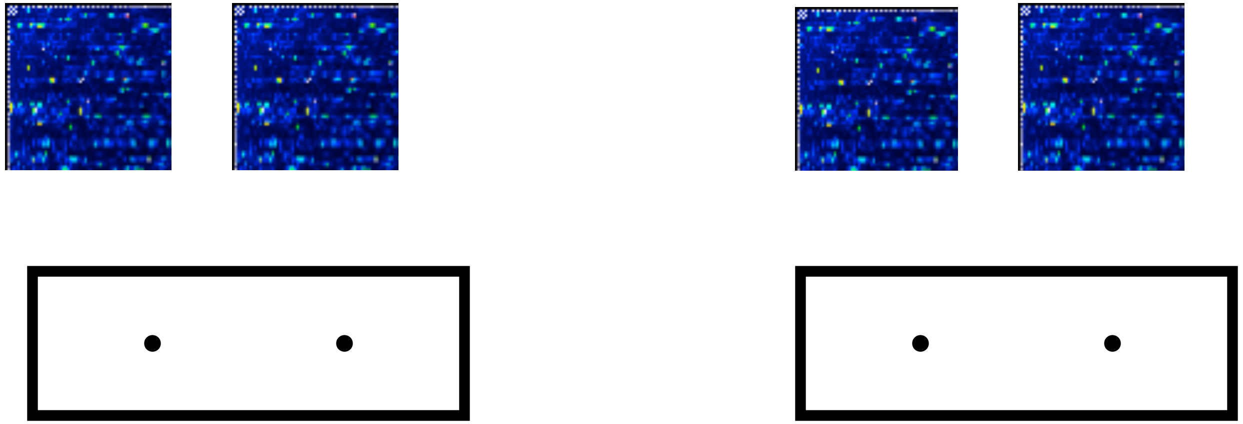
22 samples x ~20,000 genes

reduced to 22 samples x 2 dimensions

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)

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

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

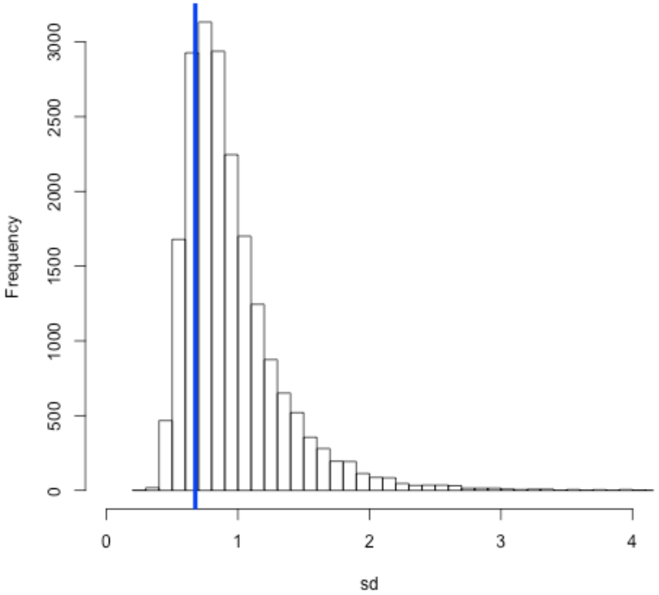
Shrink standard deviations towards common value

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

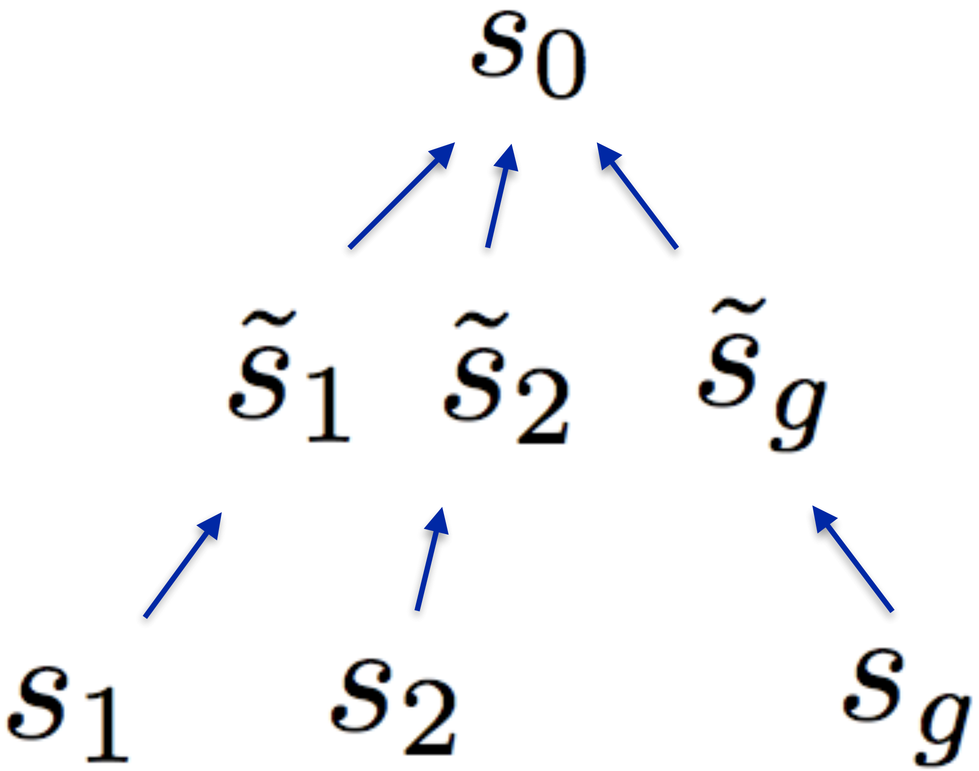
d = degrees of freedom

Moderated t-statistics

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



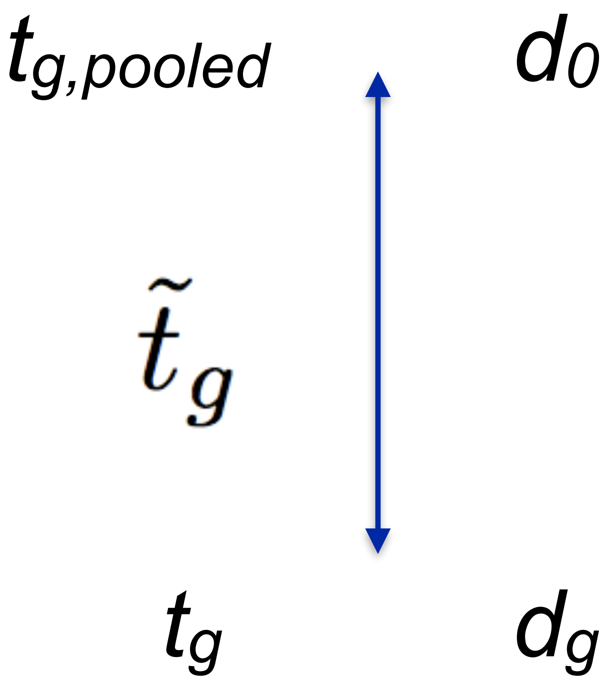
Shrinkage of standard deviations



s_0

$\tilde{s}_1$ $\tilde{s}_2$ $\tilde{s}_g$

s_1 s_2 s_g



$t_{g,pooled}$ d_0

$\tilde{t}_g$

t_g d_g

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”/„moderate“ 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 positive 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.

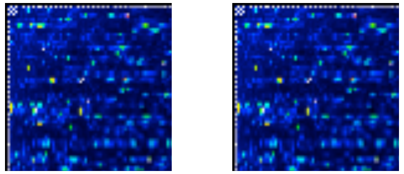
Linear Models

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

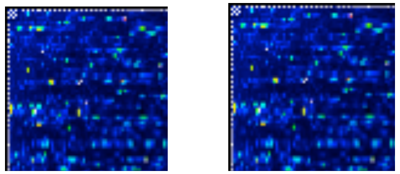
$$\underset{\substack{\nearrow\\ \text{vector of}\\ \text{observed}\\ \text{data}}}{y} = \underset{\substack{\uparrow\\ \text{design}\\ \text{matrix}}}{X}\underset{\substack{\nwarrow\\ \text{Vector of}\\ \text{parameters to}\\ \text{estimate}}}{\beta} + \epsilon$$

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

$E[y_1]=E[y_2]=\beta_1$

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

Layers to add ..

- Where does the moderated variance come from?
- Why the degrees of freedom add: $d_0 + d$
- empirical Bayes: how to estimate the hyperparameters (d_0 and s_0)
- Design matrices + contrast matrices in practice

Unexpected mathematics: Why do degrees of freedom add?

The construction of the classical t-statistic:

$$Z = \left(\overline{X}_n - \mu\right)\frac{\sqrt{n}}{\sigma}$$

$$V = (n-1)\frac{S_n^2}{\sigma^2}$$

$$T \equiv \frac{Z}{\sqrt{V/\nu}} = \left(\overline{X}_n - \mu\right)\frac{\sqrt{n}}{S_n},$$

Optional Exercise: what are a, b above?

If T is distributed as $(a/b)^{1/2}Z/U$ where $Z \sim N(0,1)$ and $U \sim \chi_\nu$, then T has density function

$$p(t) = \frac{a^{\nu/2}b^{1/2}}{B(1/2,\nu/2)(a+bt^2)^{1/2+\nu/2}}$$

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

Optional exercise: Derive the posterior

Data

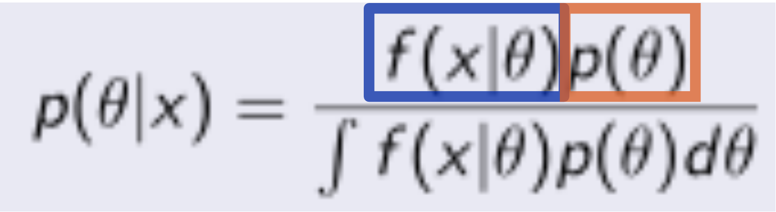
$$\boxed{s_g^2 \sim \sigma_g^2 \frac{\chi^2_{d_g}}{d_g}}$$

Prior

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

Posterior

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



$$p(\theta|x) = \frac{\boxed{f(x|\theta)}\boxed{p(\theta)}}{\int f(x|\theta)p(\theta)d\theta}$$

Optional Exercise:

Sketch: i) Let $x=s^2$, $\theta=\sigma^{-2}$; ii) Using the functional form of chi-squared distribution, calculate only the numerator (since denominator does not contain θ); iii) collect terms and see if you can identify the distribution and the parameters of it; iv) What is the mean of this distribution?

Unexpected mathematics: Why do degrees of freedom add?

$$p(\hat{\beta}, s^2 \mid \beta = 0) = \int \boxed{p(\hat{\beta} \mid \sigma^{-2}, \beta = 0)}\,\boxed{p(s^2 \mid \sigma^{-2})}\,\boxed{p(\sigma^{-2})}\,d(\sigma^{-2})$$

The integrand is

$$\begin{aligned}
& \boxed{\frac{1}{(2\pi v \sigma^2)^{1/2}} \exp\left(-\frac{\hat{\beta}^2}{2v\sigma^2}\right)} \\
\times & \boxed{\left(\frac{d}{2\sigma^2}\right)^{d/2} \frac{s^{2(d/2-1)}}{\Gamma(d/2)} \exp\left(-\frac{ds^2}{2\sigma^2}\right)} \\
\times & \boxed{\left(\frac{d_0 s_0^2}{2}\right)^{d_0/2} \frac{\sigma^{-2(d_0/2-1)}}{\Gamma(d_0/2)} \exp\left(-\sigma^{-2}\frac{d_0 s_0^2}{2}\right)} \\
= & \frac{(d_0 s_0^2/2)^{d_0/2}(d/2)^{d/2} s^{2(d/2-1)}}{(2\pi v)^{1/2}\Gamma(d_0/2)\Gamma(d/2)} \\
& \sigma^{-2(1/2+d_0/2+d/2-1)} \exp\left\{-\sigma^{-2}\left(\frac{\hat{\beta}^2}{2v} + \frac{ds^2}{2} + \frac{d_0 s_0^2}{2}\right)\right\}
\end{aligned}$$

Unexpected mathematics: Why do degrees of freedom add?

$$p(\hat{\beta}, s^2 \,|\, \beta = 0) = \int p(\hat{\beta} \,|\, \sigma^{-2}, \beta = 0) p(s^2 \,|\, \sigma^{-2}) p(\sigma^{-2}) d(\sigma^{-2})$$

$$= \frac{(d_0 s_0^2/2)^{d_0/2} (d/2)^{d/2} s^{2(d/2-1)}}{(2\pi v)^{1/2} \Gamma(d_0/2) \Gamma(d/2)} \boxed{\sigma^{-2(1/2+d_0/2+d/2-1)} \exp\left\{-\sigma^{-2}\left(\frac{\hat{\beta}^2}{2v} + \frac{ds^2}{2} + \frac{d_0 s_0^2}{2}\right)\right\}}$$

σ^{-2} is chi-squared (or gamma)

$$f(x;\, k) = \begin{cases} \frac{x^{(k/2)-1} e^{-x/2}}{2^{k/2} \Gamma\left(\frac{k}{2}\right)}, & x \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

http://en.wikipedia.org/wiki/Chi-squared_distribution

Unexpected mathematics: Why do degrees of freedom add?

$$p(\hat{\beta}, s^2 \,|\, \beta = 0) = \int p(\hat{\beta} \,|\, \sigma^{-2}, \beta = 0) p(s^2 \,|\, \sigma^{-2}) p(\sigma^{-2}) d(\sigma^{-2})$$

$$\begin{aligned} & p(\hat{\beta}, s^2 \,|\, \beta = 0) \\ & \quad = \frac{(1/2v)^{1/2} (d_0 s_0^2/2)^{d_0/2} (d/2)^{d/2} s^{2(d/2-1)}}{D(1/2, d_0/2, d/2)} \left(\frac{\hat{\beta}^2/v + d_0 s_0^2 + d s^2}{2} \right)^{-(1+d_0+d)/2} \end{aligned}$$

Unexpected mathematics: Why do degrees of freedom add?

$$\boxed{\begin{aligned} & p(\hat{\beta}, s^2 \,|\, \beta = 0) \\ & \quad = \frac{(1/2v)^{1/2}(d_0 s_0^2/2)^{d_0/2}(d/2)^{d/2} s^{2(d/2-1)}}{D(1/2, d_0/2, d/2)} \left(\frac{\hat{\beta}^2/v + d_0 s_0^2 + ds^2}{2}\right)^{-(1+d_0+d)/2} \end{aligned}}$$

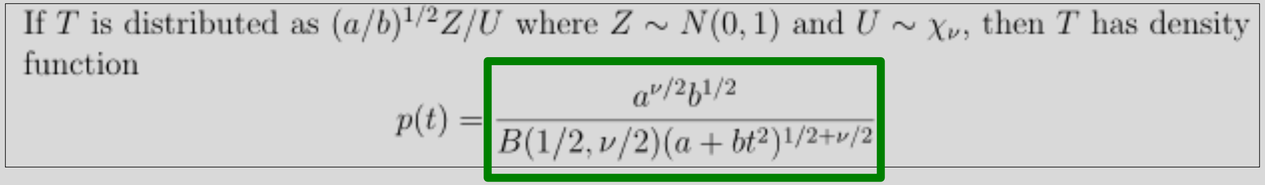
The null joint distribution of $\tilde{t}$ and s^2 is

$$p(\tilde{t}, s^2 \,|\, \beta = 0) = \tilde{s} v^{1/2} \boxed{p(\hat{\beta}, s^2 \,|\, \beta = 0)}$$

http://en.wikipedia.org/wiki/Random_variable#Distribution_functions_of_random_variables

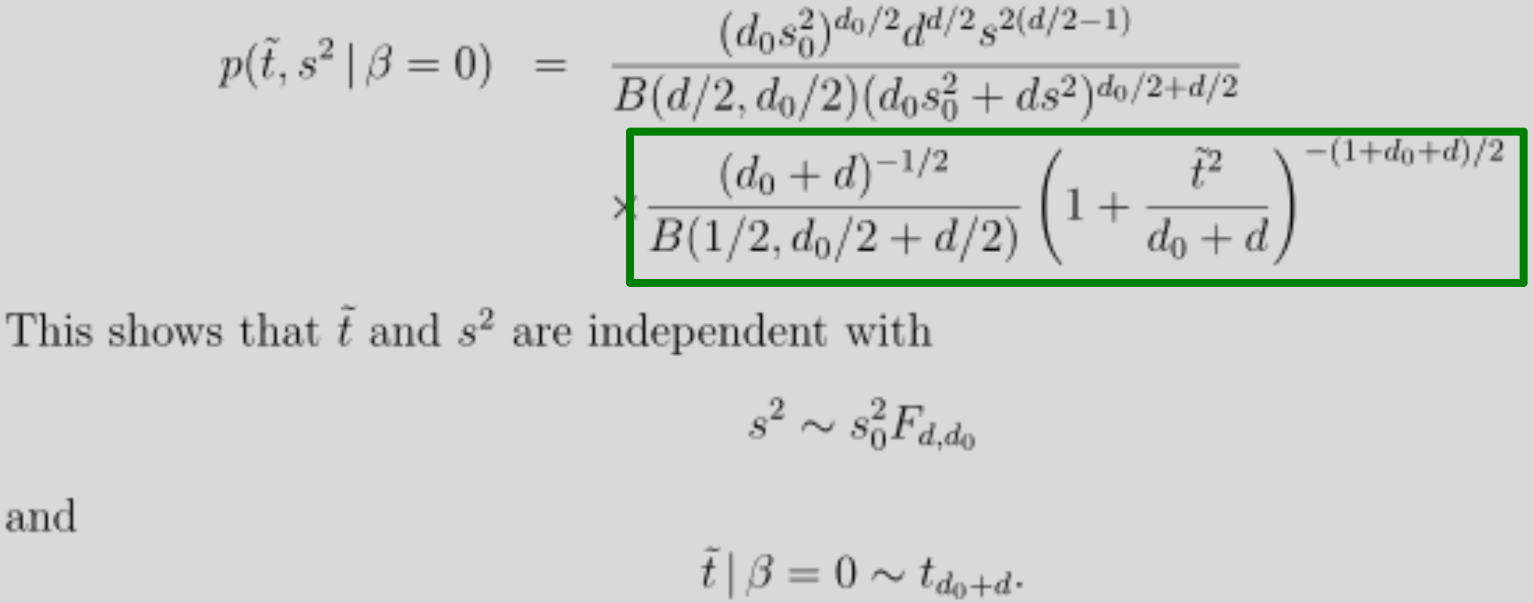
$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|$$

Unexpected mathematics: Why do degrees of freedom add?



If T is distributed as $(a/b)^{1/2}Z/U$ where $Z \sim N(0,1)$ and $U \sim \chi_\nu$, then T has density function

$$p(t) = \boxed{\frac{a^{\nu/2}b^{1/2}}{B(1/2,\nu/2)(a+bt^2)^{1/2+\nu/2}}}$$



$$p(\tilde{t}, s^2 \,|\, \beta = 0) = \frac{(d_0 s_0^2)^{d_0/2} d^{d/2} s^{2(d/2-1)}}{B(d/2, d_0/2)(d_0 s_0^2 + d s^2)^{d_0/2+d/2}} \times \boxed{\frac{(d_0+d)^{-1/2}}{B(1/2, d_0/2+d/2)} \left(1 + \frac{\tilde{t}^2}{d_0+d}\right)^{-(1+d_0+d)/2}}$$

This shows that $\tilde{t}$ and s^2 are independent with

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

and

$$\tilde{t} \,|\, \beta = 0 \sim t_{d_0+d}.$$

Linear Models

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

$$y = X\alpha + \epsilon$$

vector of observed data

design matrix

Vector of parameters to estimate

Obtain a linear model for each gene g

$$E(\underset{\sim}{y}_g) = X\underset{\sim}{\alpha}_g$$

$$\mathrm{var}(\underset{\sim}{y}_g) = W_g^{-1}\sigma_g^2$$

Contrasts -- `contrasts.fit()`

A *contrast* is any linear combination of the coefficients α_j which we want to test equal to zero.

Define contrasts

$$\underset{\sim}{\beta}_g = C^T \underset{\sim}{\alpha}_g$$

were C is the contrast matrix.

Want to test

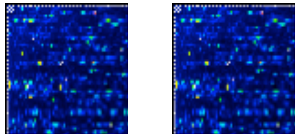
$$H_0 : \beta_{gj} = 0$$

vs

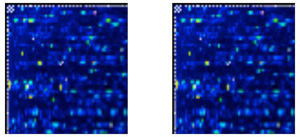
$$H_a : \beta_{gj} \neq 0$$

Analysis of Variance → Linear model

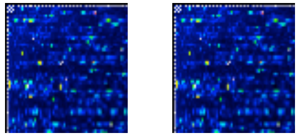
WT x 2



Cond A x 2



Cond B x 2



$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

α_1 = wt log-expression

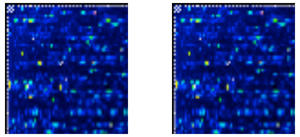
α_2 = Cond A - wt

α_3 = Cond B - wt

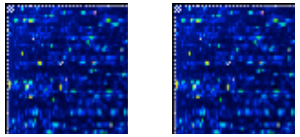
$E[y_1]=E[y_2]=\alpha_1$ $E[y_3]=E[y_4]=\alpha_1+\alpha_2$ $E[y_5]=E[y_6]=\alpha_1+\alpha_3$

Analysis of Variance → Linear model, alternative parameterization

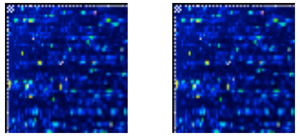
WT x 2



Cond A x 2



Cond B x 2



$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

α_1 = wt log-expression

α_2 = Cond A log-expression

α_3 = Cond B log-expression

$E[y_1]=E[y_2]=\alpha_1$ $E[y_3]=E[y_4]=\alpha_2$ $E[y_5]=E[y_6]= \alpha_3$

An example use of design and contrast matrices

design matrix

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

$E[y_1]=E[y_2]=\alpha_1$

$E[y_3]=E[y_4]=\alpha_2$

$E[y_5]=E[y_6]=\alpha_3$

contrast matrix

$$\beta = C\alpha = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} \alpha_2 - \alpha_1 \\ \alpha_3 - \alpha_2 \end{bmatrix}$$

Contrasts -- `contrasts.fit()`

A *contrast* is any linear combination of the coefficients α_j that we want to test equal to zero.

Define contrasts

$$\underset{\sim}{\beta}_g = C^T \underset{\sim}{\alpha}_g$$

were C is the contrast matrix.

Want to test

$$H_0 : \beta_{gj} = 0$$

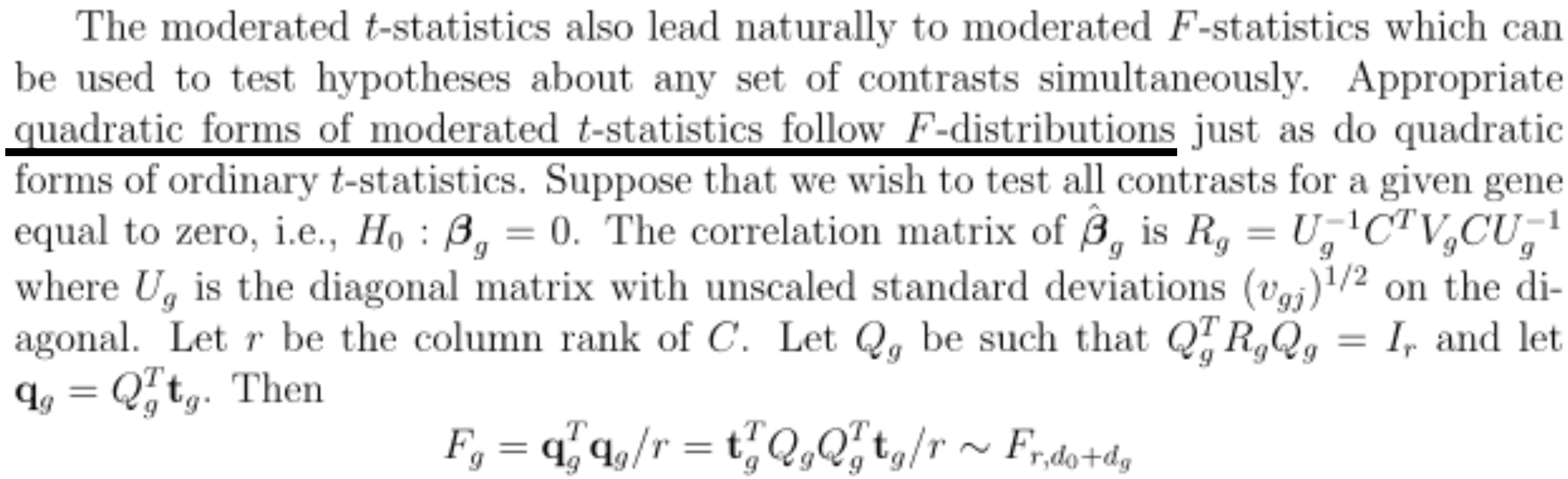
vs

$$H_a : \beta_{gj} \neq 0$$

Limma / Analysis of Variance

$$F = \frac{\text{variance between treatments}}{\text{variance within treatments}}$$

$$F = \frac{MS_{\text{Treatments}}}{MS_{\text{Error}}} = \frac{SS_{\text{Treatments}}/(I-1)}{SS_{\text{Error}}/(n_T - I)}$$



The moderated t-statistics also lead naturally to moderated F-statistics which can be used to test hypotheses about any set of contrasts simultaneously. Appropriate quadratic forms of moderated t-statistics follow F-distributions just as do quadratic forms of ordinary t-statistics. Suppose that we wish to test all contrasts for a given gene equal to zero, i.e., $H_0 : \boldsymbol{\beta}_g = 0$. The correlation matrix of $\hat{\boldsymbol{\beta}}_g$ is $R_g = U_g^{-1} C^T V_g C U_g^{-1}$ where U_g is the diagonal matrix with unscaled standard deviations $(v_{gj})^{1/2}$ on the diagonal. Let r be the column rank of C. Let Q_g be such that $Q_g^T R_g Q_g = I_r$ and let $\mathbf{q}_g = Q_g^T \mathbf{t}_g$. Then

$$F_g = \mathbf{q}_g^T \mathbf{q}_g / r = \mathbf{t}_g^T Q_g Q_g^T \mathbf{t}_g / r \sim F_{r, d_0 + d_g}$$

Aside: Marginal Distributions to calculate

Fun fact: 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

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

Thus, the set of s_g can be used to estimated d_0 and s_0

Section 6.2 limma paper: other tricks, such as Fisher's z distribution to estimate d_0 and s_0

Relate to limma objects

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

$E[y_1]=E[y_2]=\alpha_1$

$E[y_3]=E[y_4]=\alpha_2$

$E[y_5]=E[y_6]=\alpha_3$

$$\beta = C\alpha = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} \alpha_2 - \alpha_1 \\ \alpha_3 - \alpha_2 \end{bmatrix}$$

```
> design
  alpha1 alpha2 alpha3
1      1      0      0
2      1      0      0
3      0      1      0
4      0      1      0
5      0      0      1
6      0      0      1
> cont.matrix <- makeContrasts(beta1="alpha2-alpha1",
                 beta2="alpha3-alpha2",levels=design)
> cont.matrix
        Contrasts
Levels   beta1 beta2
  alpha1    -1     0
  alpha2     1    -1
  alpha3     0     1

fit <- lmFit(y,design)

fit.c  <- contrasts.fit(fit, cont.matrix)
fit.c  <- eBayes(fit.c)

> head(round(y,2),3)
       [,1]   [,2]  [,3]  [,4]   [,5]   [,6]
[1,]  -1.62   1.49  2.50  1.57  -0.71   0.38
[2,]  -4.50  -4.95 -3.66 -7.83  -1.59   6.94
[3,] -10.17 -21.90 14.03  3.66 -12.21 -15.26

> head(round(fit$coef,2),3)
     alpha1 alpha2 alpha3
[1,]  -0.07   2.03  -0.16
[2,]  -4.73  -5.75   2.67
[3,] -16.04   8.85 -13.74

> head(round(fit.c$coef,2),3)
      Contrasts
       beta1  beta2
  [1,]  2.10  -2.20
  [2,] -1.02   8.42
  [3,] 24.89 -22.59
```