Differential expression analysis using limma and edgeR

Mark van de Wiel
mark.vdwiel@vumc.nl

Department of Epidemiology & Biostatistics
Department of Mathematics
VUmc/VU, Amsterdam

Content

Overview

0.	Arrays vs Sequencing; Gaussian vs Counts
1.	Limma[1]
2.	RNAseq + Normalization
3.	Differential expression for RNAseq
4.	Overdispersion
5.	edgeR
6.	Limitations of edgeR

[1]Some slides by Wessel van Wieringen

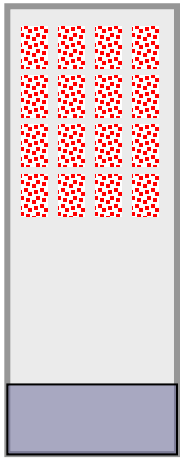
Arrays versus Sequencing

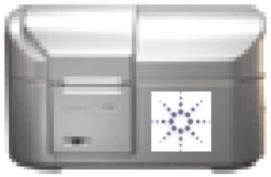
Arrays

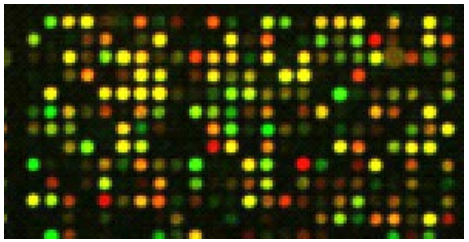
- Hybridization technique
- Relative measurement
- Probes

RNA sequencing

- Counting RNA sequences
- Absolute measurement
- Tags







Arrays versus Sequencing

Arrays

+ Cheap
+ Easy to analyse
+ Designed

- Bad sensitivity for lowly abundant genes
- Resolution fixed
- Only gene expression

RNA sequencing

- More expensive[1]
- Harder to analyse
- Not designed

+ Better sensitivity for lowly abundant genes
+ Arbitary resolution[2]
+ Also exon expression, mutations, inversions, etc.

[1]But becoming cheaper rapidly; [2]But this may come at a price…multiple testing

Limma

Limma

Limma

Limma is an R package that facilitates the analysis of microarray experiments in order to identify differentially expressed genes.

Limma uses standard regression models, but estimates the variance by “borrowing” information across all genes.

Hereto the distribution of the variance across all genes is estimated, and used to improve estimates of gene-specific variance (empirical Bayes).

Limma

Recall standard regression

Let:

- Y_i denote the response variable,
- $X_1,\ldots, X_p$ the independent variables,
- $X_{i1},\ldots, X_{ip}$ the values of the covariates at observation i, $\forall \beta_1,\ldots, \beta_p$ the regression coefficients (parameters), and
- n the number of observations.

The ***design space*** is the p-dimensional space spanned by the covariates.

The ***design matrix*** is the $(p\times n)$-matrix, denoted $\boldsymbol{X}$, with elements X_{ij} representing the value of the j-th covariate at the i-th observation.

Limma

Design matrix for single-color microarray experiments

array 1	*array 2*	*array 3*	*array 4*
A	A	B	B

$$\mathbf{X} = \begin{array}{c} \\ \textit{array 1} \\ \textit{array 2} \\ \textit{array 3} \\ \textit{array 4} \end{array} \begin{array}{c} \begin{array}{cc} \textit{factor A} & \textit{factor B} \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \end{array}$$

Limma

The standard regression model:

$$\mathbf{Y}_j = \mathbf{X}\,\boldsymbol{\alpha}_j + \boldsymbol{\varepsilon}_j \quad \text{with} \quad \mathrm{Var}(\varepsilon_{ij}) = \sigma_j^2$$

and contrast of interest: $\boldsymbol{\beta}_j = \mathbf{C}^T\,\boldsymbol{\alpha}_j$

Estimates:

$$\hat{\boldsymbol{\alpha}}_j = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{Y}_j$$

$$\mathrm{Var}(\hat{\boldsymbol{\alpha}}_j) = s_j^2\,(\mathbf{X}^T\mathbf{X})^{-1} = s_j^2\,\mathbf{V}$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_j) = s_j^2\,\mathbf{C}^T\,\mathbf{V}\,\mathbf{C}$$

Limma

Distributional assumptions (standard):

$$\hat{\beta}_{jk} \,|\, \beta_{jk}, \sigma_j^2 \;\sim\; N(\beta_{jk},\, \sigma_j^2\, (\mathbf{C}^T\, \mathbf{V}\, \mathbf{C})_{kk})$$

$$s_j^2 \,|\, \sigma_j^2 \;\sim\; \frac{\sigma_j^2}{s_j^2}\, \chi^2_{d_j}$$

where d_j the degrees of freedom.

The ordinary *t*-statistic:

$$t_{jk} \;=\; \hat{\beta}_{jk} \,/\, \sqrt{s_j^2\, (\mathbf{C}^T\, \mathbf{V}\, \mathbf{C})_{kk}} \;\sim\; t_{d_j}$$

Limma

Limma extends this to a hierarchical model:

$$\mathbf{Y}_j = \mathbf{X}\,\boldsymbol{\alpha}_j + \boldsymbol{\varepsilon}_j, \qquad \mathrm{Var}(\varepsilon_{ij}) = \sigma_j^2, \qquad \boldsymbol{\beta}_j = \mathbf{C}^T\,\boldsymbol{\alpha}_j$$

Prior for σ_j^2

Prior for β_{jk}

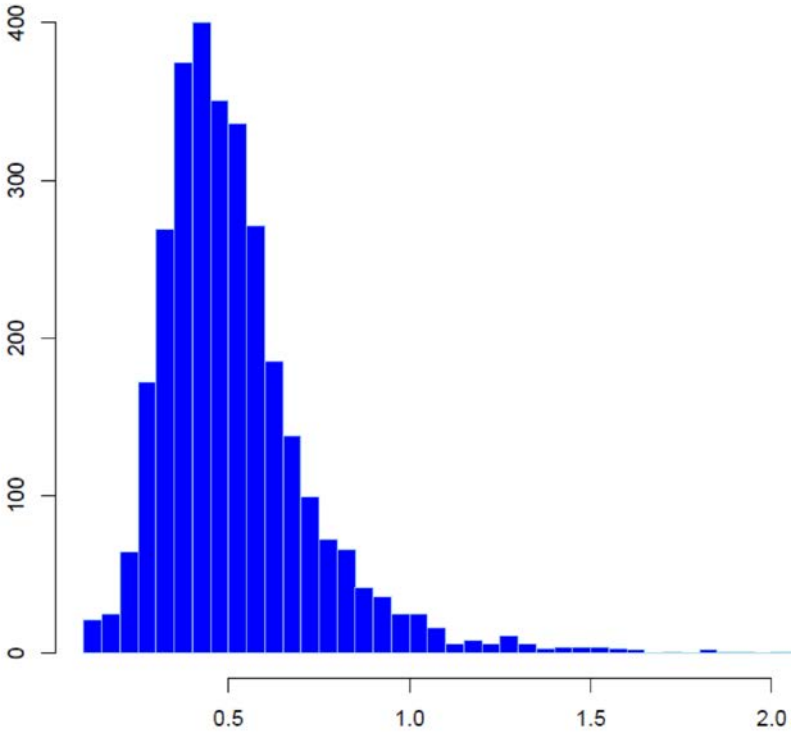
Priors:

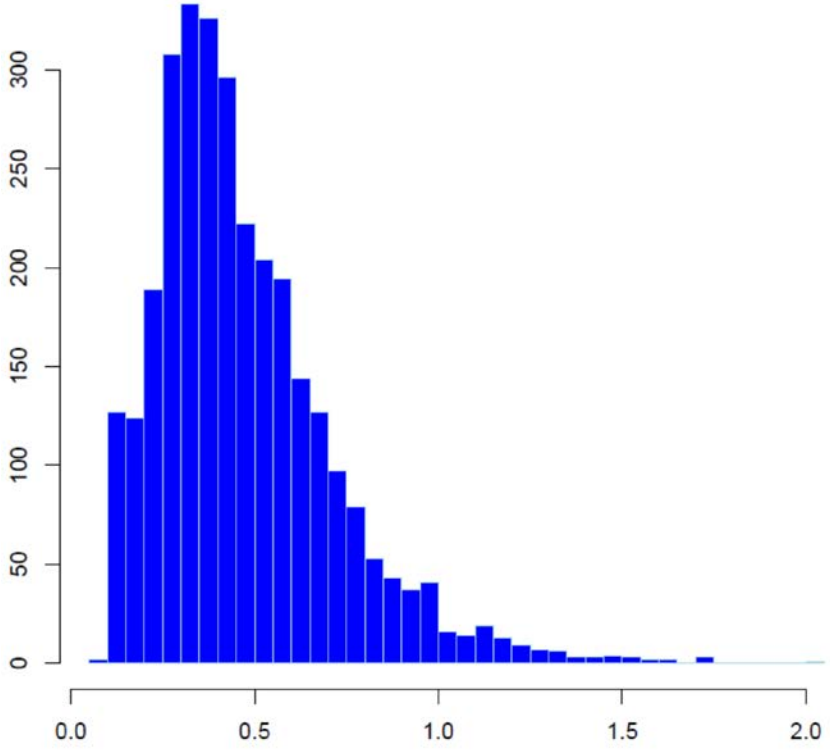
$$\frac{1}{\sigma_j^2} \sim \frac{1}{d_0\, s_0^2}\,\chi^2_{d_0}$$

$$\beta_{jk}\,|\,\sigma_j^2 \sim (1 - P(\beta_{jk} \neq 0)) + P(\beta_{jk} \neq 0)\, N(0, v_{0k}\,\sigma_j^2)$$

Limma

Are priors reasonable?
Check for sigmas in Golub data in both groups





Limma

The posterior mean of the variance[1]:

$$\tilde{s}_j^2 \;=\; E(\sigma_j^2 \mid s_j^2) \;=\; \frac{d_0\, s_0^2 + d_j\, s_j^2}{d_0 + d_j}$$

Observed variance is shrunken towards the prior mean.

The moderated *t*-statistic:

$$\tilde{t}_{jk} \;=\; \hat{\beta}_{jk} \,/\, \sqrt{\tilde{s}_j^2\, (\mathbf{C}^T\, \mathbf{V}\, \mathbf{C})_{kk}}$$

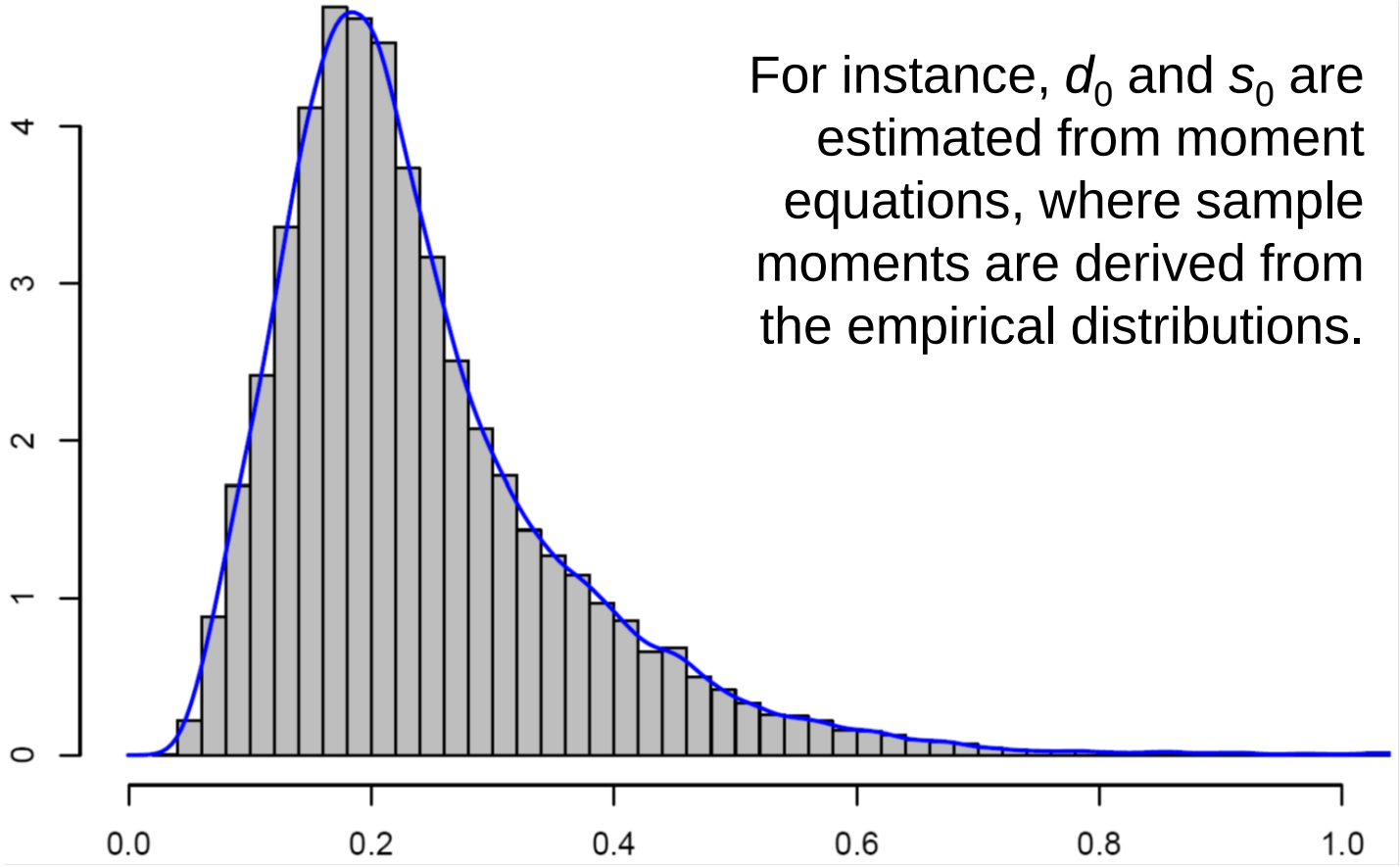
follows a *t*-distribution with $d_0 + d_j$ degrees of freedom

$$t_{jk} \;=\; \hat{\beta}_{jk} \,/\, \sqrt{s_j^2\, (\mathbf{C}^T\, \mathbf{V}\, \mathbf{C})_{kk}}$$

[1]Details: Smyth, G (2004) *SAGMB*

Limma

Hyperparameters are estimated from empirical distributions.

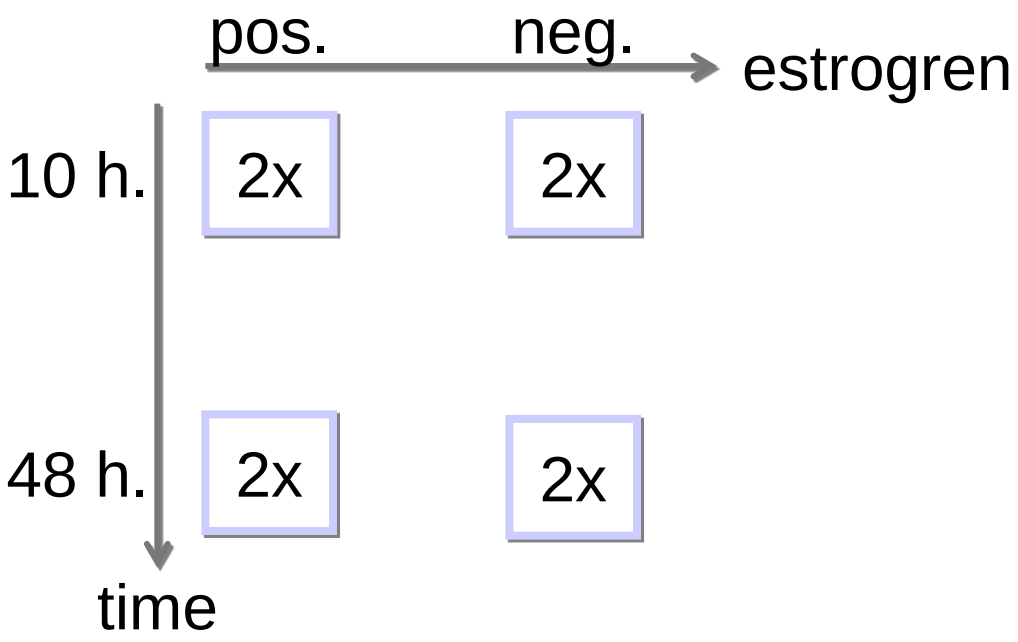


For instance, d_0 and s_0 are estimated from moment equations, where sample moments are derived from the empirical distributions.

Limma

Experiment

- A 2x2 factorial design
- Affymetrix hgu95av2
- Fully replicated



Model: expression = $\beta_0 + \beta_1$ x estrogen + β_2 x time

Limma

```
#install packages from Bioconductor
source("http://bioconductor.org//biocLite.R")
          biocLite(pkgs=c("estrogen", "limma", "hgu95av2cdf"))

# load necessary libraries
library(estrogen)
library(limma)
library(hgu95av2cdf)

# set working directory
setwd(system.file("extdata", package = "estrogen"))

# load experimental design
expDesign <- read.table("estrogen.txt", header = TRUE, row.names = 1)

# load data
data <- read.affybatch(c("low10-1.cel", "low10-2.cel", "high10-1.cel",
                    "high10-2.cel", "low48-1.cel", "low48-2.cel",
                    "high48-1.cel", "high48-2.cel"), phenoData =
                    new("AnnotatedDataFrame",
data=data.frame(expDesign)))
```

Limma

```
# normalize data
eset <- rma(data)

# create an appropriate design matrix
designMat <- model.matrix(~ factor(expDesign[,1])
                                                + factor(expDesign[,2]))
colnames(designMat)[2:3] <- c("estrogen", "time")

# fit linear model
fit <- lmFit(eset, designMat)

# assess significance and edit for presentation.
fit <- eBayes(fit)

# generate list of top diff.exp. genes with respect to time
topTable(fit, coef="time")
```

Limma

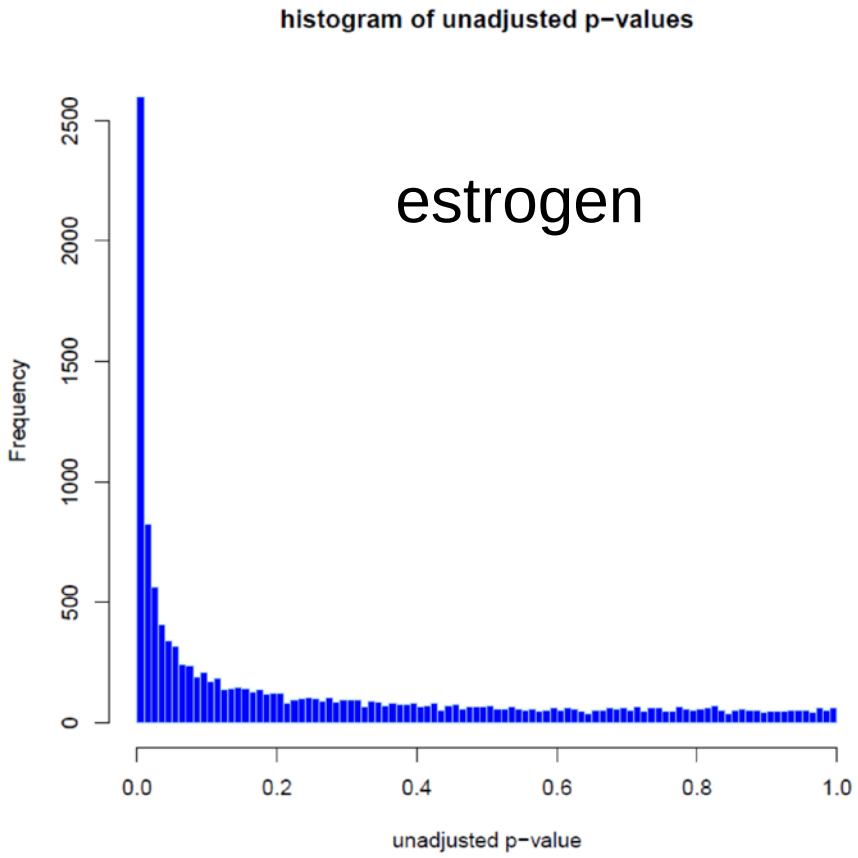
generate list of top diff.exp. genes with respect to time

```
topTable(fit, coef="estrogen")
```

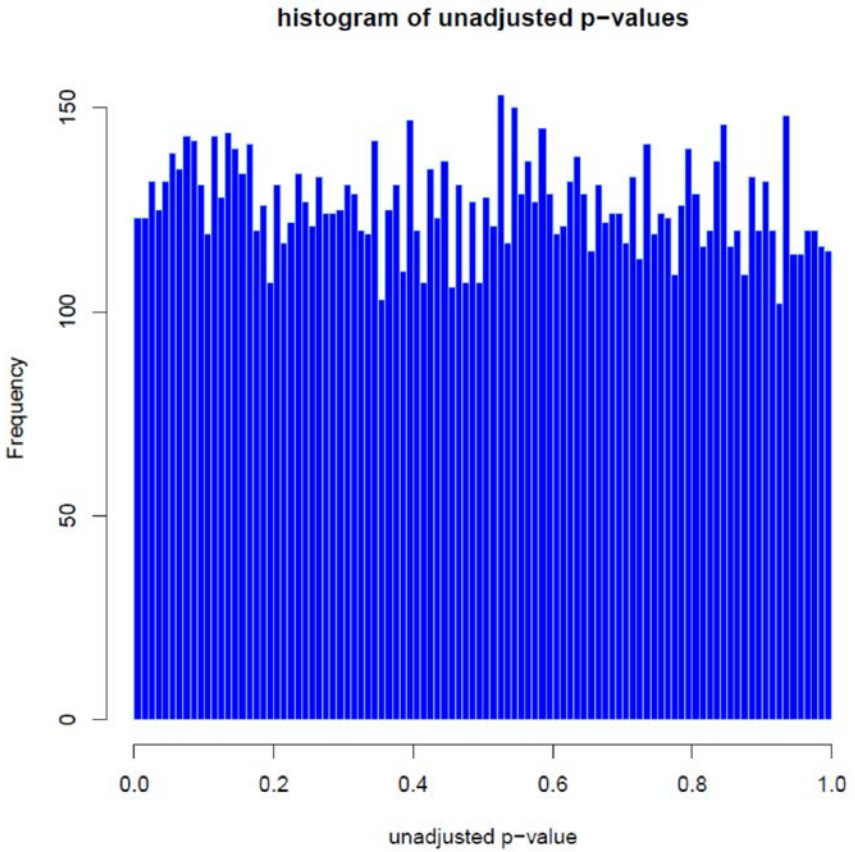
	ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
12573	AFFX-CreX-3_at	-6.770435	10.406343	-42.68670	3.508801e-12	3.698689e-08	14.430031
12575	AFFX-CreX-5_at	-7.167717	9.921026	-40.43357	5.859309e-12	3.698689e-08	14.258688
12563	AFFX-BioB-M_at	-3.601440	8.270105	-25.94820	3.828441e-10	1.611136e-06	12.336014
12571	AFFX-BioDn-5_at	-4.070924	8.348424	-24.03520	7.848047e-10	2.477040e-06	11.905782
12569	AFFX-BioDn-3_at	-2.655716	11.401080	-22.19835	1.650565e-09	4.167676e-06	11.429312
1115	2004_at	-2.062192	6.826916	-18.66861	8.272932e-09	1.740763e-05	10.292735
10167	40071_at	-1.818236	8.096773	-17.26175	1.708289e-08	3.024944e-05	9.738052
12567	AFFX-BioC-5_at	-2.243696	8.739444	-17.04775	1.916796e-08	3.024944e-05	9.647630
7082	37014_at	-1.388554	7.820017	-16.61719	2.426732e-08	3.404165e-05	9.460491
12565	AFFX-BioC-3_at	-3.534264	8.159920	-15.50963	4.575624e-08	5.776725e-05	8.944951

Limma

Histogram of unadjusted *p*-values reveals a lot of signal



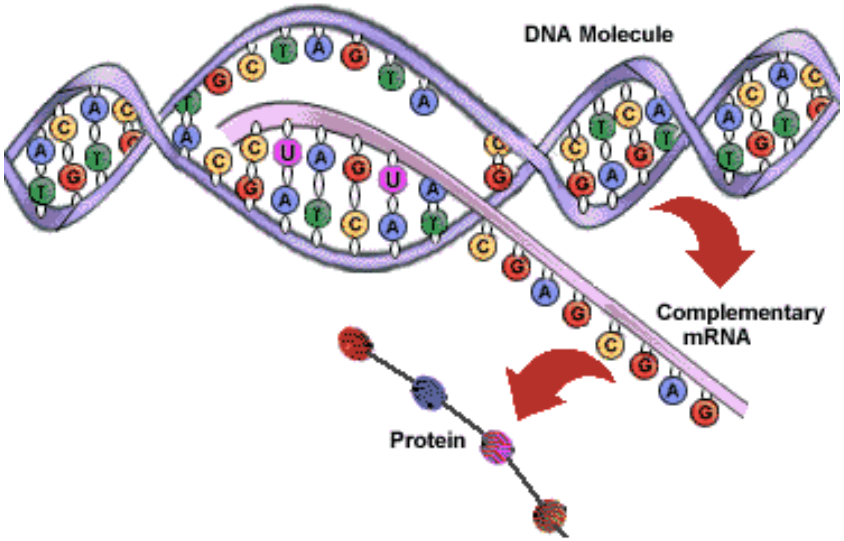
estrogen



random covariate

Intro
RNA Sequencing

RNA Sequencing



- mRNA is transcribed from DNA
- Only those pieces that the cell needs are transcribed (predominantly genes)
- Different pieces of DNA are transcribed in different quantities

- RNA-sequencing (RNA-seq): counts the number of copies of each piece for a given unit (e.g. tag, gene) for a pool of cells

Sequencing history

- 1977 - technologies published
- 1980 - Nobel prize
  - Wally Gilbert & Fred Sanger
- 1982 - GenBank started
- 1987 - 1st automated sequencer
- 1990 - 2003 – Humane Genome Project completed
- 2006 – next generation sequencing

Next generation sequencing

- Massively parallel sequencing
- Millions of sequence reads
- Short reads (25 – 500 nucleotides)
- Up to complete genome in one run
- Used for
  - Individual samples to find genomic events that may relate to disease
  - Denovo assembly of new genomes for many different organisms

Next generation sequencing platforms

- Roche 454
- Illumina / Solexa
- ABI SOLiD
- Helicos
- Pacific Biosystems
- …

Costs of sequencing has decreased tremendously over the last decade


Solexa
1G GENETIC ANALYZER

Raw data file

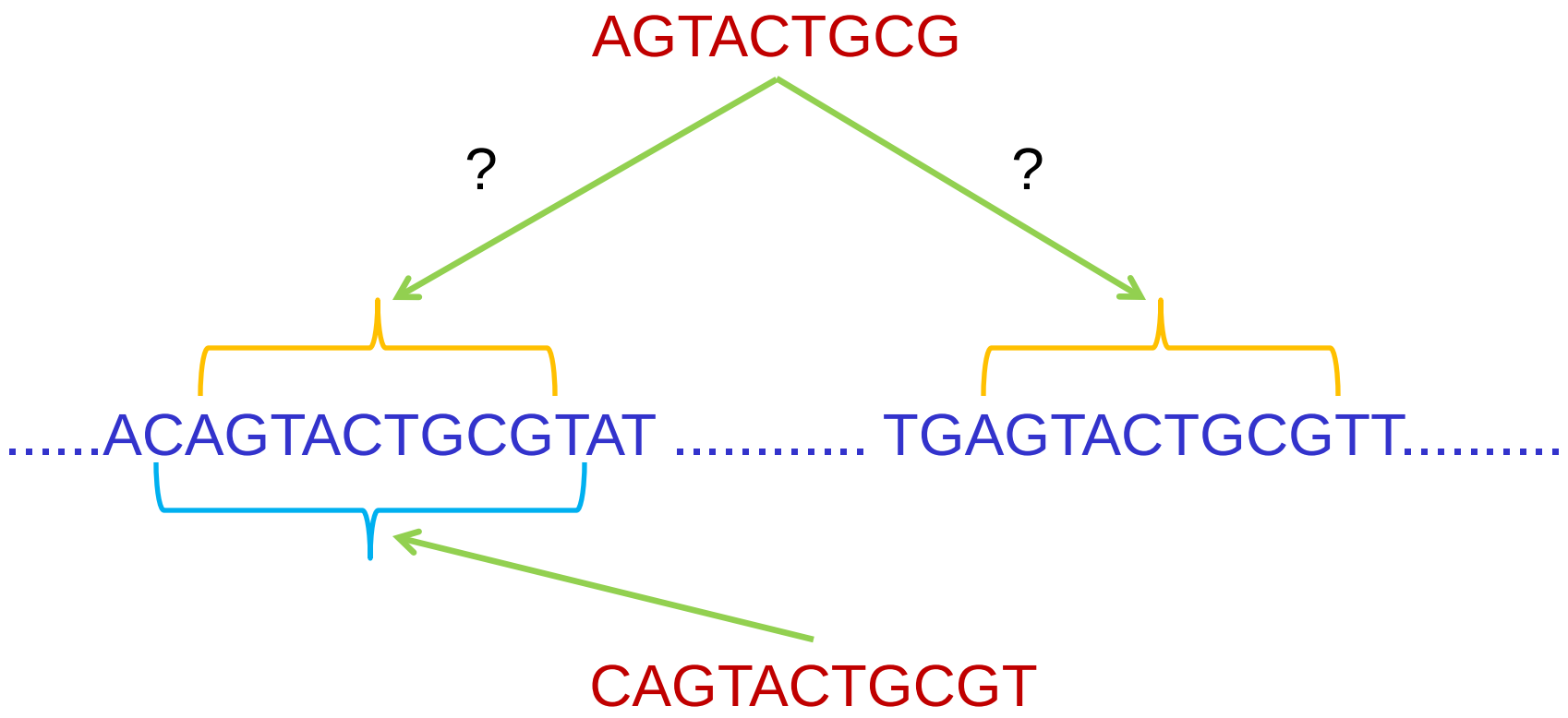
```
@SRR042267.1 HWI-EA332_308KJAAXX:2:1:1638:1923
GGGAACACACTCCAGAGTCGTATGCCGTCTTCTGCT
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@SRR042267.2 HWI-EA332_308KJAAXX:2:1:772:1307
GTGTTTTAAGGCTAAATTCGTATGCCGTCTTCTGCT
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@SRR042267.3 HWI-EA332_308KJAAXX:2:1:1470:339
GTGCCTCACAACCATCCTCGTATGCCGTCTTCTGCT
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@SRR042267.4 HWI-EA332_308KJAAXX:2:1:771:1911
GTAGTATTGGCTGGAATTCGTATGCCGTCTTCTGCT
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIII,II
@SRR042267.5 HWI-EA332_308KJAAXX:2:1:899:190
GATTGTCAAAAAAAAAATCGTATGCCGTCTTCTGCT
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@SRR042267.6 HWI-EA332_308KJAAXX:2:1:1778:1916
GAAAAATAACGTGTGTTTCGTATGCCGTCTTCTGCT
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@SRR042267.7 HWI-EA332_308KJAAXX:2:1:846:527
GTGTTTTAAAGCTAAATCGTATGCCGTCTTCTGCTT
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@SRR042267.8 HWI-EA332_308KJAAXX:2:1:279:1552
GCATGCCCTCTGGATCAGTCGTATGCCGTCTTCTGC
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

header
sequence
custom
quality score

Alignment

RNA strands: variable lengths, depending on technology, typically 10's to 100's

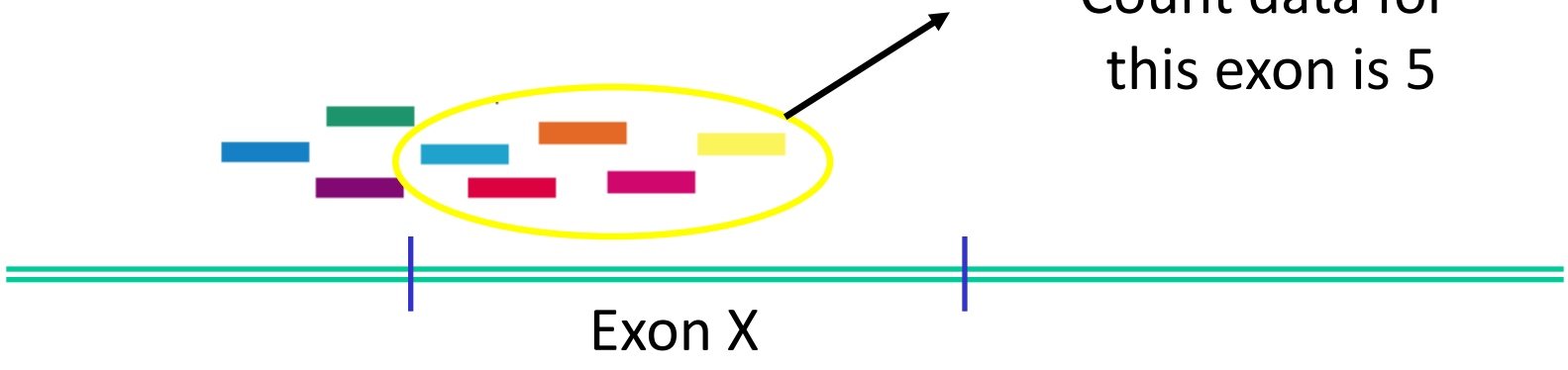
Humane Genome: length ~ $3*10^9$



Coverage

Source: www.illumina.com

Counting sequences in exons



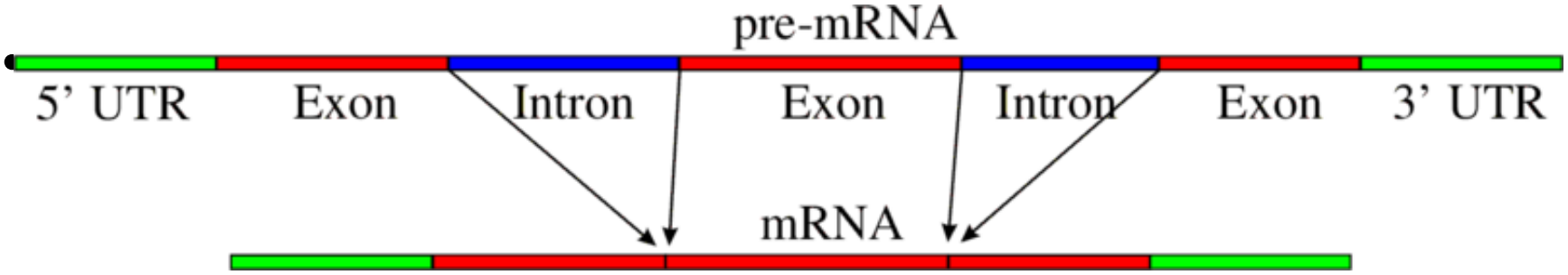
Gene id	WT_SHAM	WT_SHAM	WT_SHAM	WT_SHAM	WT_SHAM	WT_SHAM	WT_CSD1	WT_CSD2	WT_CSD3	WT_CSD4	WT_CSD4	WT_CSD5	WT_CSD6	RQ_SHAM	RQ_SHAM	RQ_SHAM	RQ_SHAM	RQ_SHAM	RQ_SHAM	RQ_CSD1	RQ_CSD2	RQ_CSD3	RQ_CSD4	RQ_CSD5_
ENSMUSG	553	1167	660	1736	365	760	971	961	974	864	926	1458	2843	556	871	1263	1142	1043	1919	969	794	771	682	394
ENSMUSG	2	11	3	8	2	8	1	10	3	30	9	5	19	1	5	34	13	1	3	8	13	5	1	1
ENSMUSG	26	50	50	42	100	140	29	31	41	28	41	147	99	20	43	74	135	63	92	55	44	23	68	41
ENSMUSG	381	664	442	762	476	653	618	376	353	79	727	1271	1938	381	306	859	687	249	1596	773	635	240	562	171
ENSMUSG	399	956	531	1725	1217	1111	668	928	350	151	398	828	2199	480	702	1611	1448	331	678	1199	1014	652	913	258
ENSMUSG	776	2144	1500	3594	2453	3111	1570	1955	948	379	1363	2021	5833	1087	1401	3664	3103	1414	1999	2526	2110	1241	1805	655
ENSMUSG	58	131	74	211	170	150	68	95	83	81	74	203	282	78	65	113	115	60	143	132	101	90	119	33
ENSMUSG	22	30	20	26	35	44	30	19	18	0	17	38	78	14	19	97	86	10	23	41	34	12	14	26
ENSMUSG	31	100	39	138	87	68	66	82	72	19	72	166	277	53	78	209	64	49	173	118	109	85	67	31
ENSMUSG	25	67	45	103	307	485	34	44	129	0	82	162	187	46	68	324	298	190	164	59	68	32	229	40
ENSMUSG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSMUSG	14	39	21	41	33	47	30	20	91	33	49	110	58	17	18	44	76	48	100	48	42	28	43	34
ENSMUSG	0	1	1	0	0	5	1	1	0	0	0	3	6	1	0	1	0	0	3	0	0	0	0	0
ENSMUSG	84	131	64	144	100	67	112	113	171	0	88	111	268	81	90	195	190	103	188	142	115	90	77	50
ENSMUSG	5	10	8	32	8	23	13	6	0	0	13	35	21	4	4	0	0	0	23	12	15	5	13	2
ENSMUSG	34	81	62	89	49	73	50	48	7	0	185	285	143	47	29	27	80	0	285	87	69	46	45	17
ENSMUSG	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ENSMUSG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSMUSG	5	21	5	27	33	6	13	14	11	0	11	1	50	11	7	15	48	58	23	22	21	14	19	10
ENSMUSG	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSMUSG	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	9	0	0	2	0	0	0	1
ENSMUSG	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0
ENSMUSG	354	732	494	1263	467	621	574	615	292	135	429	909	1615	380	558	674	642	292	1256	906	816	484	616	237
ENSMUSG	0	1	0	1	0	0	0	0	0	0	1	10	0	0	0	0	0	0	0	6	0	1	0	0
ENSMUSG	4	28	13	16	21	23	16	14	27	0	11	34	52	6	11	90	0	58	66	12	13	6	29	11

RNAseq

- Sequence specific regions at high coverage OR whole-genome at lower coverage

- Coverage: average number of reads representing a given nucleotide. Calculated as N*L/G with G: length of the original genome, N: the number of reads, L: the average read length

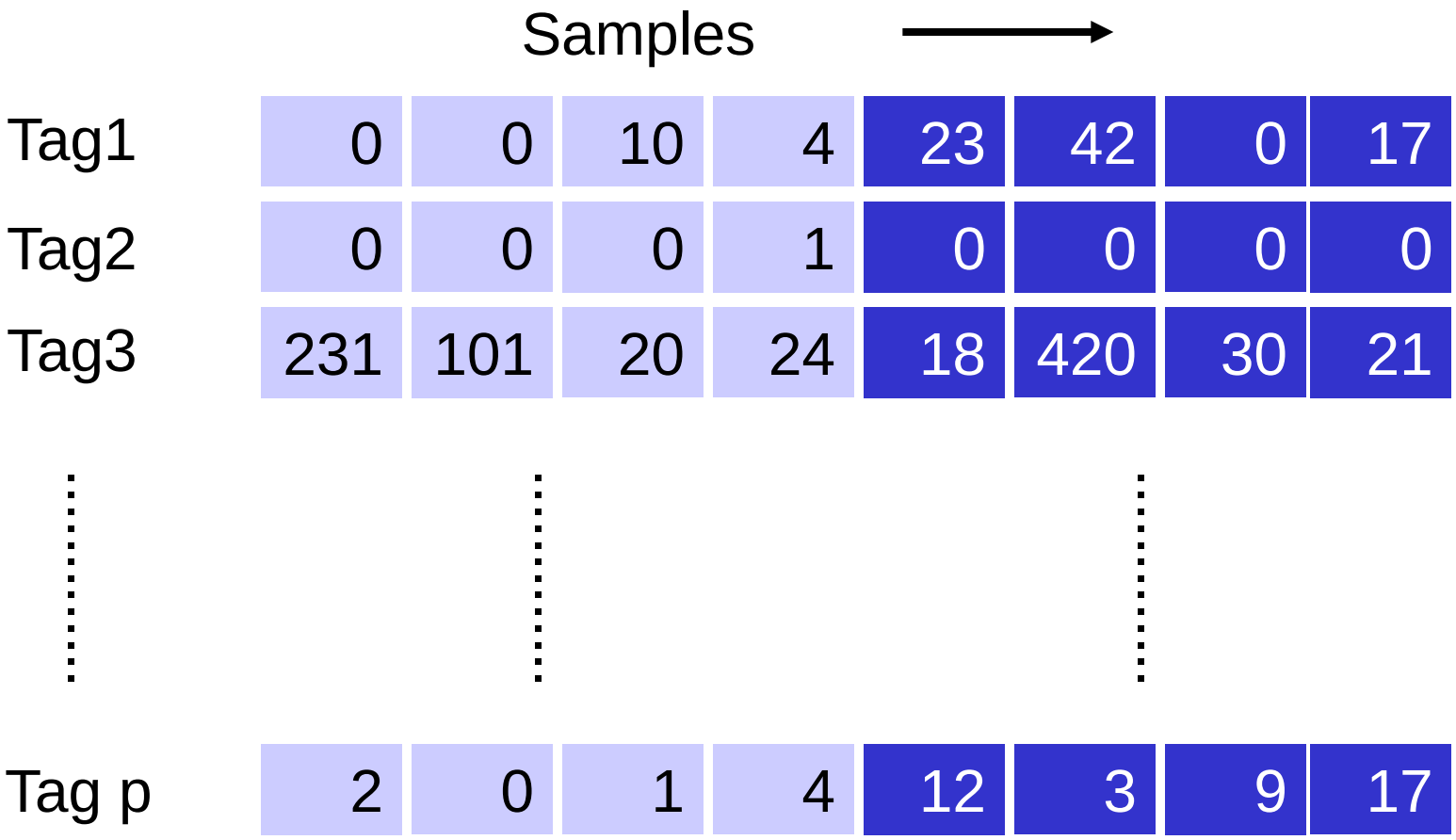
- RNAseq should be better than microarrays, in particular for lowly abundant mRNAs + provides more detailed information, e.g. on pieces of genes (exons)



Normalization for RNAseq data

RNAseq

Data: preprocessed; Tag: piece of RNA. May be a gene/exon.



Counts!

RNAseq:normalization

Library sizes (unit 10^6)

L1= 2.8	L2= 3.2	L3= 0.9	L4= 4.5	L5= 0.7	L6= 3.7	L7= 2.1	L8= 1.3

Mean = $2.4*10^6$

Correction multiplicative factor: Ci = Mean/Li for i = 1, ..., 8

Tag3	231	101	20	24	18	420	30	21
Ci	0.9	0.8	2.7	0.5	3.4	0.6	1.1	1.8

X

Tag3	198	76	53	13	62	272	34	39

RNAseq:normalization

Samples →

Tag1	0	0	27	2	79	27	0	31
Tag2	0	0	0	1	0	0	0	0
Tag3	198	76	53	13	62	272	34	39
⋮	⋮		⋮			⋮		
Tag p	2	0	3	2	41	2	10	31
∑ (*10^6)	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.4

Normalized library sizes are equal.

TMM and upper quartile

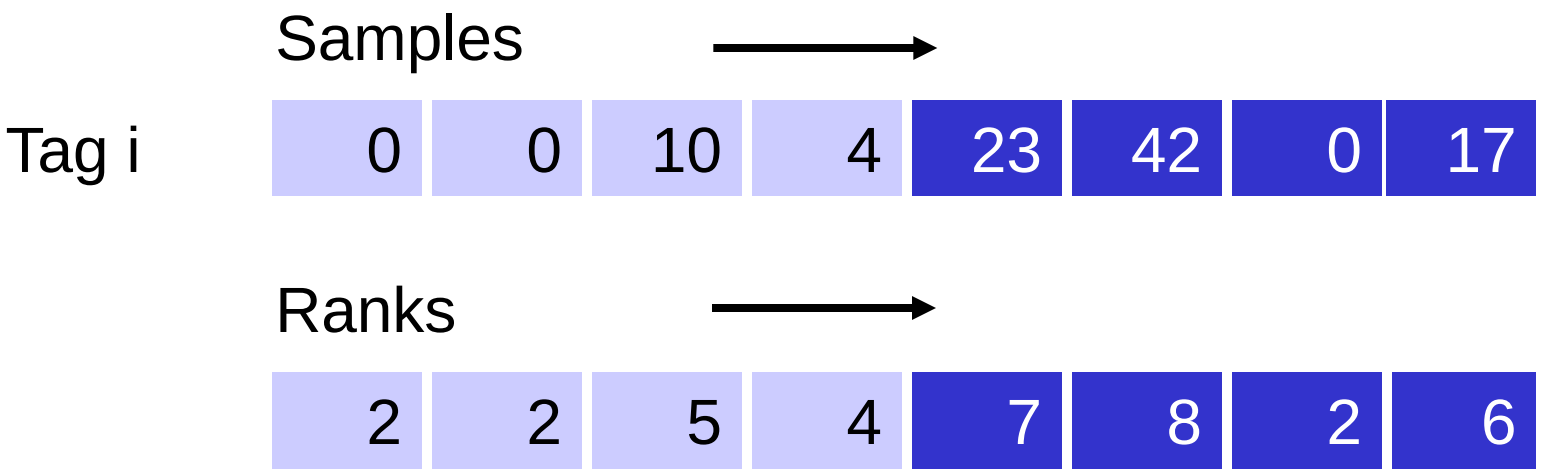
- Upper quartile:
  - Let $L_j(0.75)$ be the 75% quantile of library j.
  - med = median($L_1(0.75)$, ..., $L_n(0.75)$)
  - Then multiply vector $\mathbf{y}_{.j}$ by med/$L_j(0.75)$
- TMM
  - Similar, but uses weigthed trimmed mean of log-ratios (M-values) with respect to reference [usually sample j that corresponds to med (above)]

Differential expression for RNAseq

Differential expression for RNAseq

These are usually very non-normal data!

Permutation tests can be applied, but lack of power for small sample sizes.



Differential expression for RNAseq

R example

```
x <- c(0,0,10,4)
y <- c(23,42,0,17)
wilcox.test(x, y)

        Wilcoxon rank sum test with continuity correction

data:  x and y
W = 3, p-value = 0.1832
alternative hypothesis: true location shift is not equal to 0
```

More powerful alternatives: edgeR and ShrinkBayes[1]

[1]lectures 11,12

Counting process

Poisson distribution

$$Y_{ik} \sim \text{Poisson}(\lambda_i) \text{ with } \lambda_i = p_i r$$

i: gene, exon, ….

k: technical repeat

λ_i: true expression

Y_{ik}: observed expression

p_i: probability

r: total number of RNA molecules

Poisson distribution: good model when only technical repeats present

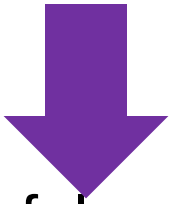
Overdispersion

Poisson: variance = mean = λ_i

Biological variation between subjects increases variance

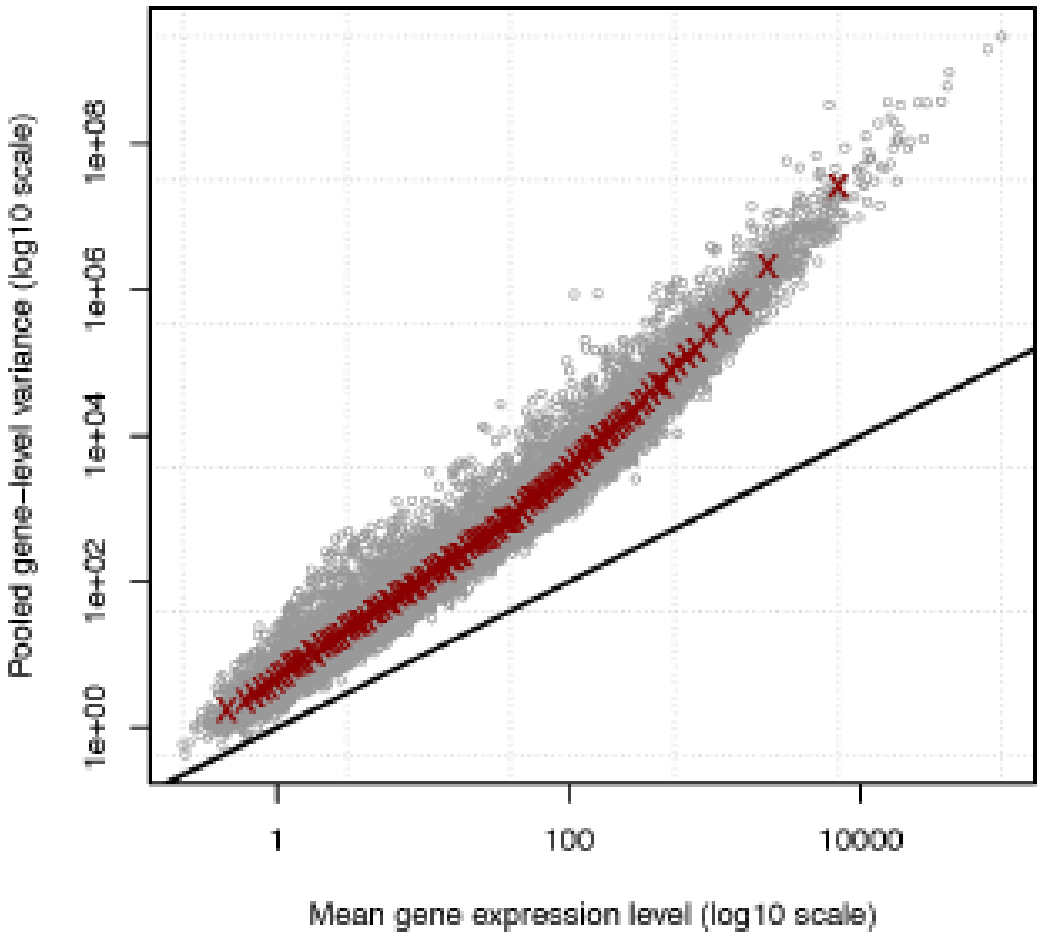
Count Y_{ij} for j: biological replication

Use of Poisson: underestimation of the variance



Many false positives

Overdispersion in our data



Counting process

Negative Binomial distribution

If Y ~ Poisson(λ) and λ ~ Gamma(f(μ, φ),g(μ,φ))

Then,Y ~ NB(μ, φ), meaning $P(Y=y_j)$ =

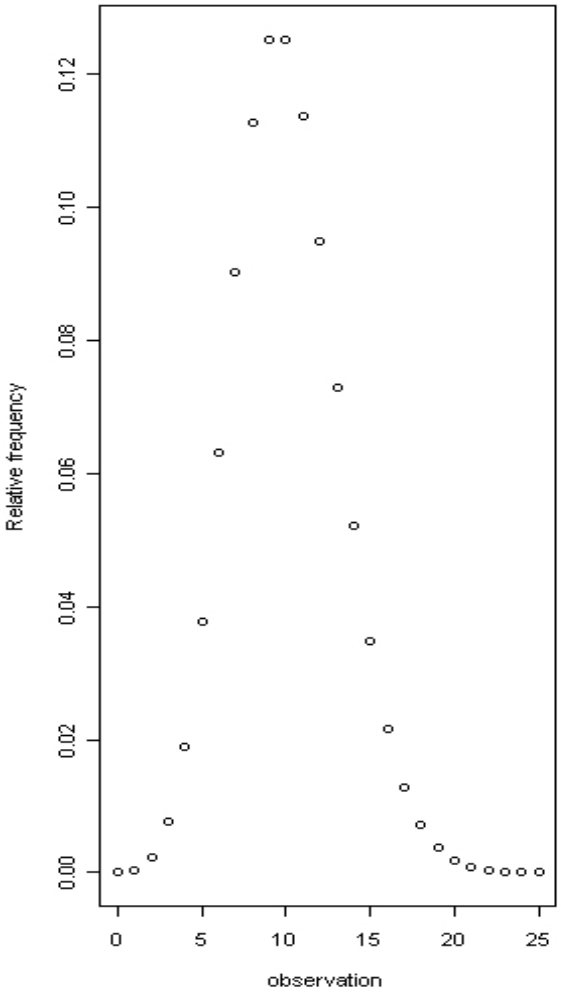
$$\frac{\Gamma(y_j+\phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y_j+1)}\left(\frac{1}{1+\mu\phi}\right)^{\phi^{-1}}\left(\frac{\mu}{\phi^{-1}+\mu}\right)^{y_j}$$

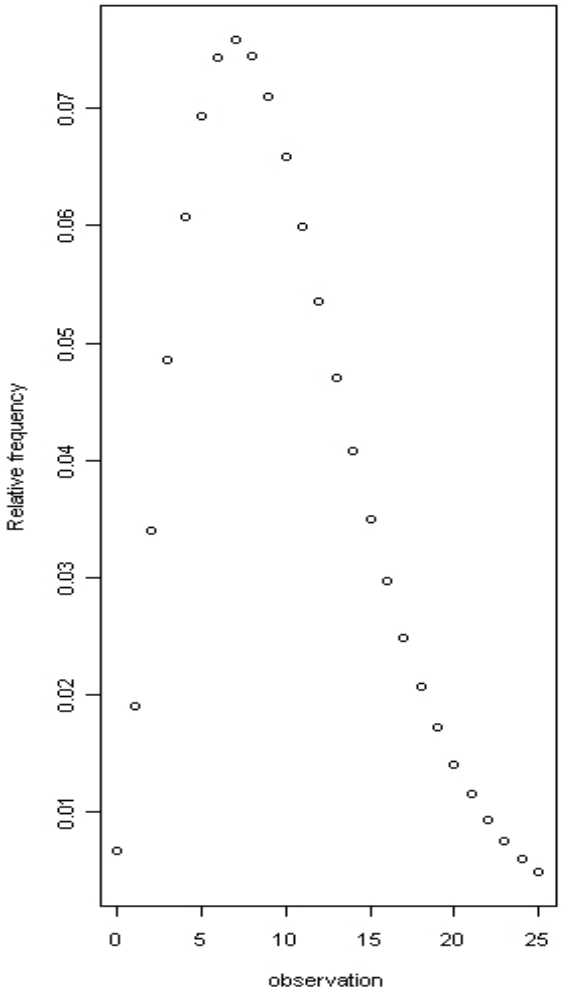
E(Y) = μ

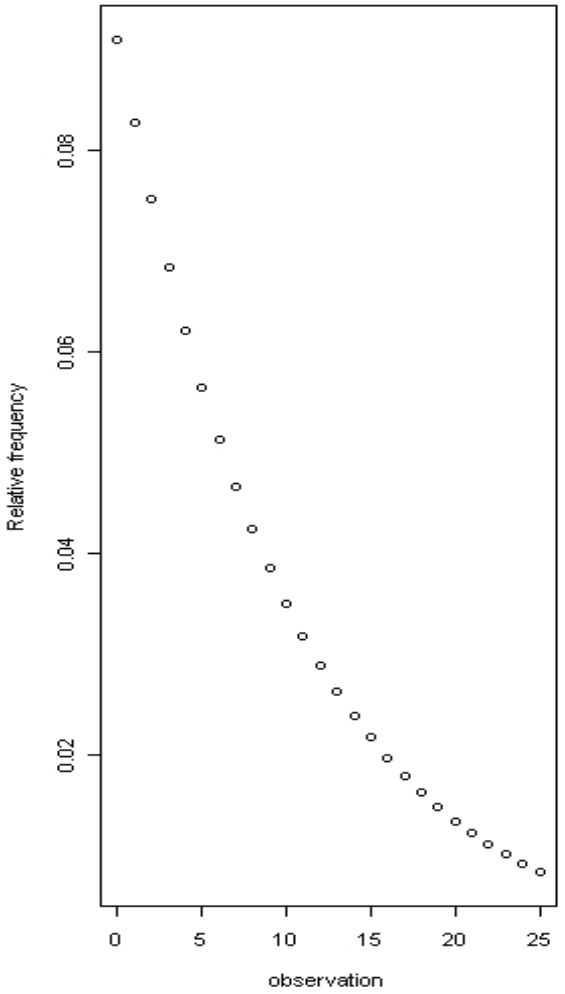
V(Y) = μ(1+ μ φ)

What are f() and g()? [Exer]

Count data





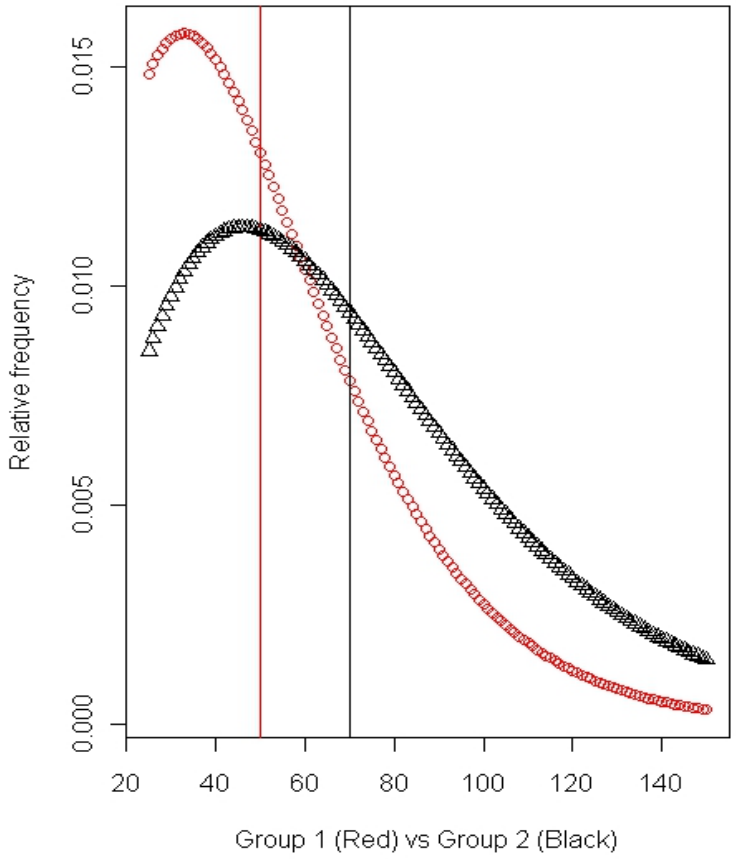


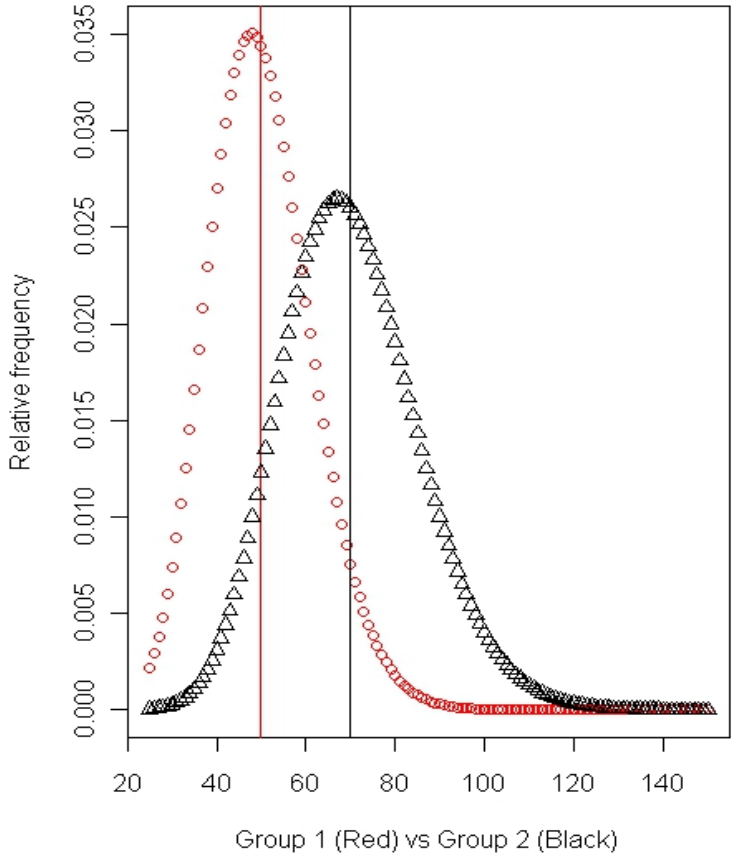
Poisson (left) + 2 negative binomials (mid,right), all mean 10

Count data

Negative binomial distribution has a *dispersion* parameter that tunes to the observed variability (between samples)

Dispersion parameter essential for the final p-value





edgeR

- R-package by Robinson et al.

- Package available from Bioconductor

MODEL

$$Y_{ij} \sim NB(\mu_{ij}, \varphi_i)$$

$$\mu_{ij} = g^{-1}(\mathbf{X_j\beta})$$

Here, g denotes the link function.

For φ_i known $\rightarrow$ NB in exponential family $\rightarrow$ GLM setting

edgeR

- Most difficult part: estimation of φ_i .
- Ordinary likelihood may render biased results
- edgeR uses concept of conditional marginal likelihood
- Idea: condition on sufficient statistic for μ_i, total count Z_i:

$$Z_i = \sum_{j=1}^{n} Y_{ij}$$

- In case covariates are absent we have $Y_{ij} \sim NB(\mu_i, \varphi_i)$, hence $Z_i \sim NB(n\mu_i, \varphi_i/n)$

edgeR, conditional likelihood

Drop index i.

$$f(y;\mu,\phi) = P(Y_j = y_j) = \frac{\Gamma(y_j+\phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y_j+1)}\left(\frac{1}{1+\mu\phi}\right)^{\phi^{-1}}\left(\frac{\mu}{\phi^{-1}+\mu}\right)^{y_j}$$

$$Z = \sum_{j=1}^{n} Y_j =^d \mathrm{NB}(n\mu, \phi/n)$$

$$l(\boldsymbol{y};\mu,\phi|z) = l(\boldsymbol{y};\mu,\phi|Z=z) = \sum_{j=1}^{n}\log[P(Y_j=y_j|Z=z)] = l(\boldsymbol{y},z;\mu,\phi) - l(z;\mu,\phi) = l(\boldsymbol{y};\mu,\phi) - l(z;\mu,\phi)$$

$$\begin{aligned} l(\boldsymbol{y};\mu,\phi) = \sum_{j=1}^{n}\Big(& \log(\Gamma(y_j+\phi^{-1})) - \log(\Gamma(\phi^{-1})) - \log(\Gamma(y_j+1)) - \phi^{-1}\log(1+\mu\phi) \\ & + y_j(\log(\mu) - \log(\phi^{-1}+\mu))\Big) \end{aligned}$$

$$\begin{aligned} l(z;\mu,\phi) = & \log(\Gamma(z+n\phi^{-1})) - \log(\Gamma(n\phi^{-1})) - \log(\Gamma(z+1)) - n\phi^{-1}\log(1+\mu\phi) \\ & + z(\log(\mu) - \log(\phi^{-1}+\mu)) \end{aligned}$$

$$l(\boldsymbol{y};\mu,\phi|z) = l(\boldsymbol{y};\phi|z) = \sum_{j=1}^{n}\log(\Gamma(y_j+\phi^{-1})) - \log(\Gamma(z+n\phi^{-1})) - n\log(\Gamma(\phi^{-1})) + \log(\Gamma(n\phi^{-1})) + C,$$

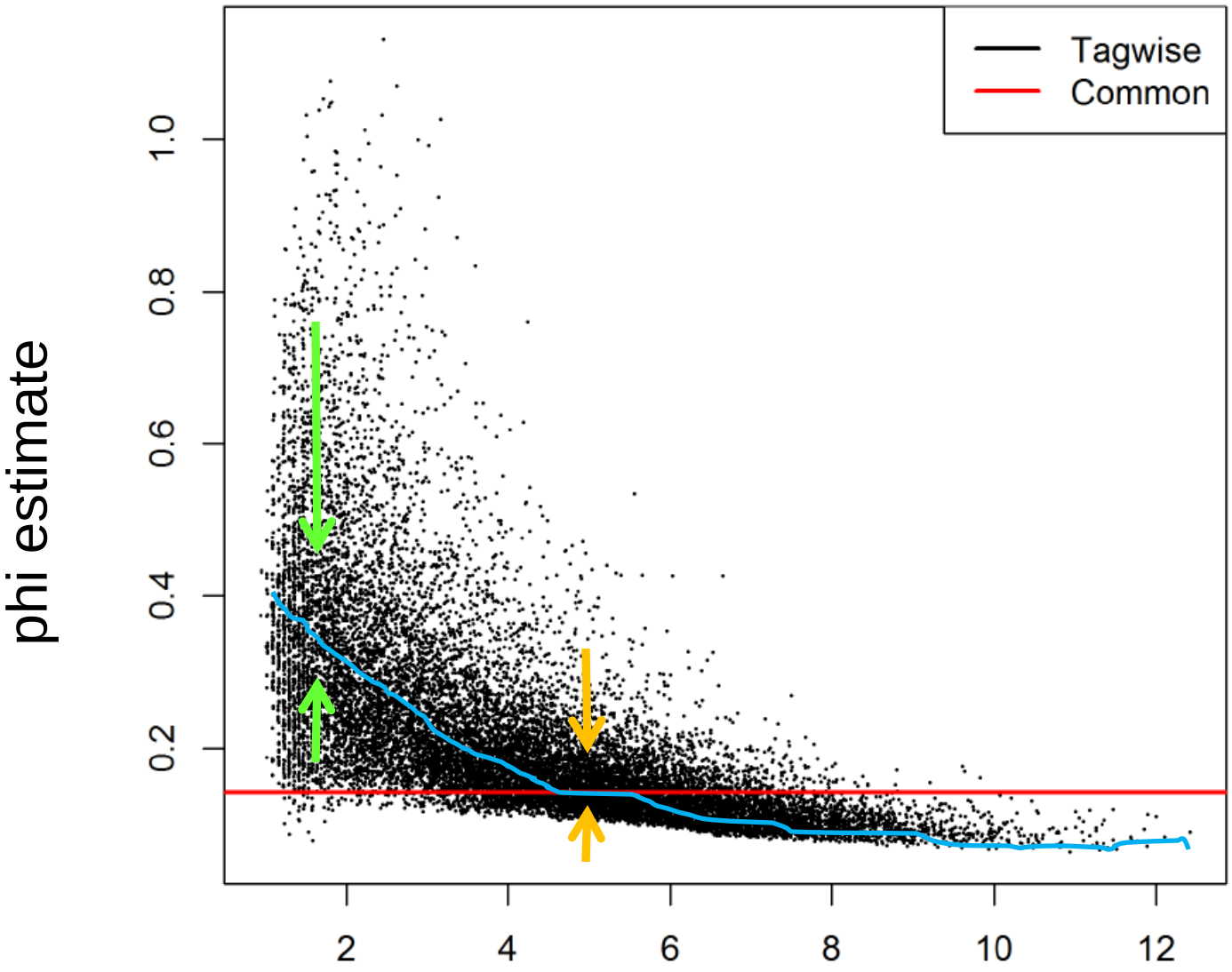
so all terms involving μ cancel, using the fact that $z = \sum_{j=1}^{n} y_j$.

edgeR

- Estimation of **common** dispersion φ: maximize the common conditional likelihood (not depending on μ):
  - $l_C(\mathbf{Y};\varphi \mid \mathbf{Z}) = \Sigma_{i=1\ldots p}\, l_i\,(\mathbf{y}_i;\varphi \mid Z_i)$
- Shrink individual likelihoods to the common likelihood
  - $WL(\varphi_i) = l_i\,(\mathbf{y}_i;\varphi_i \mid Z_i) + \alpha l_C(\mathbf{Y};\varphi_i \mid \mathbf{Z})$
- α is estimated using Empirical Bayes principles, similar to those of limma and ShrinkBayes[1]
- Maximize $WL(\varphi_i)$ with respect to φ_i

[1]to be discussed

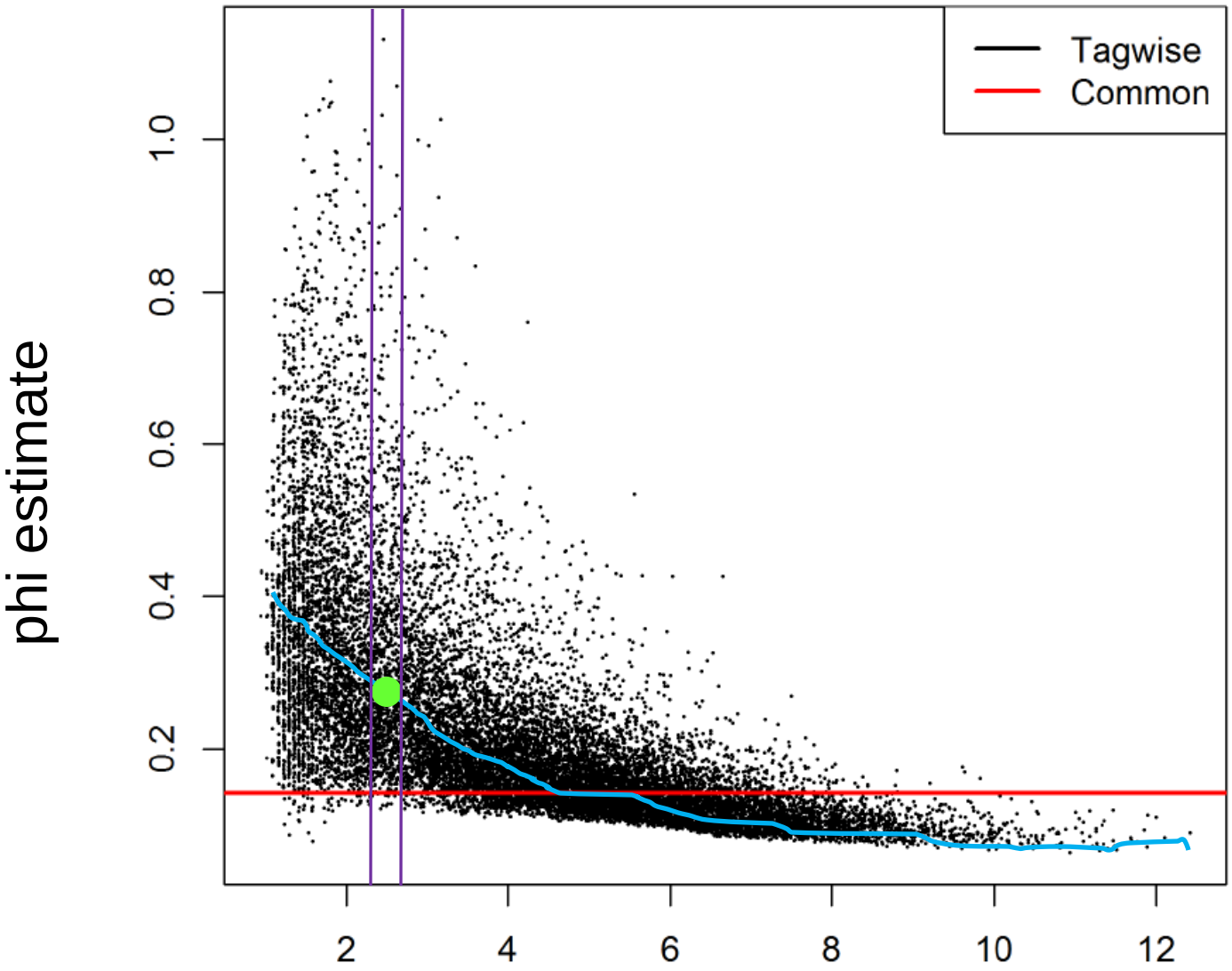
edgeR


Tagwise
Common
phi estimate
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
12
Abundance = total count

edgeR

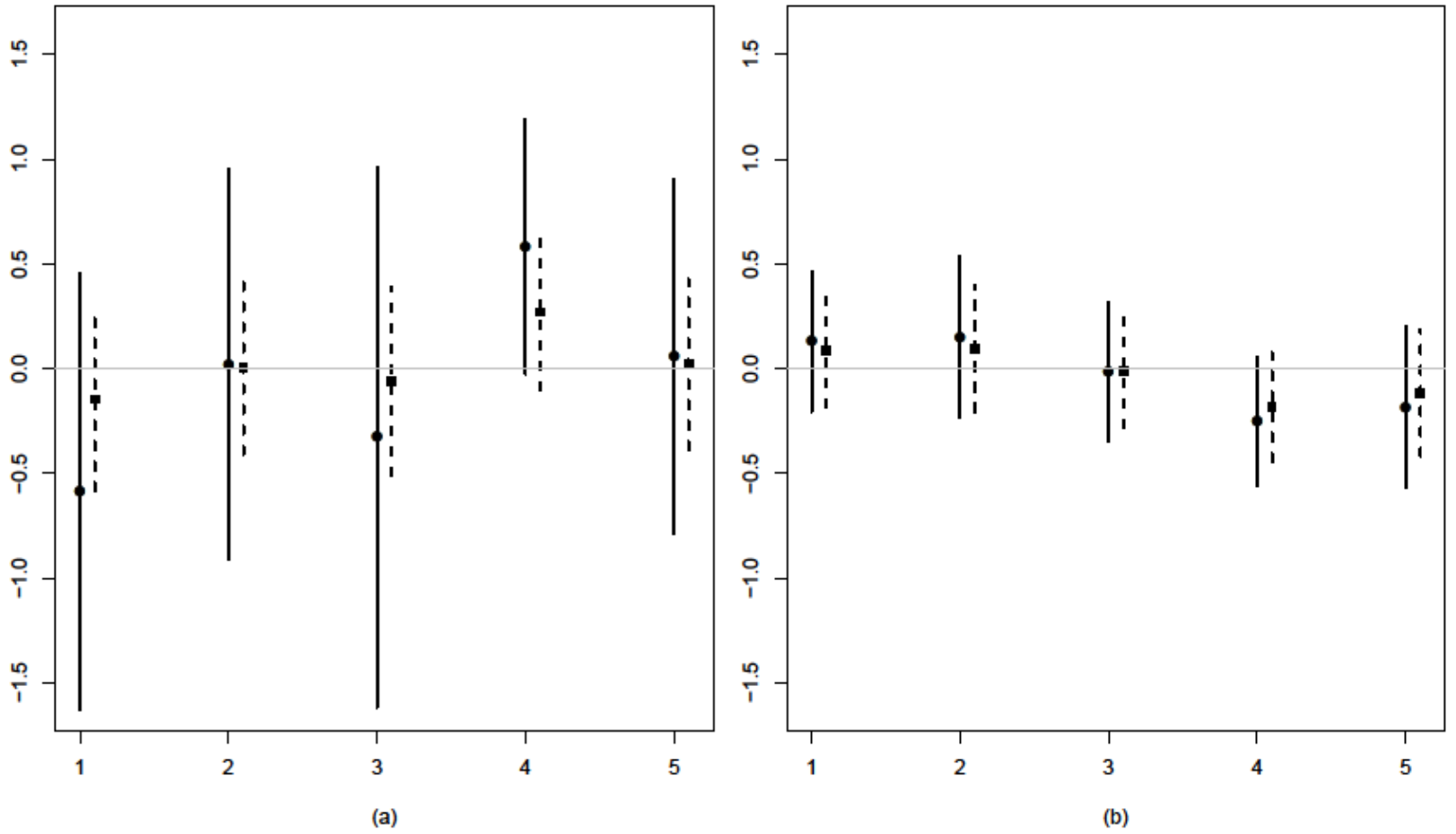
- Shrink individual likelihoods to common likelihood of those tags (i) that share a similar abundance Z_i.
- **Z**'(i) = $\{Z_j \mid |Z_j - Z_i| < \delta\}$
  - $WL(\varphi_i) = l_i(\mathbf{y}_i; \varphi_i \mid Z_i) + \alpha l_C(\mathbf{Y'}(i); \varphi_i \mid \mathbf{Z}'(i))$
- Again, α is estimated using Empirical Bayes principles
- Alternatively, use binning: **Z**'(i) consists of a fixed number of tags with abundances closest to Z_i

edgeR


Tagwise
Common
phi estimate
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
12
Abundance = total count

Shrinkage

Beneficial effect of shrinkage



5 repeated studies. Estimates of parameter of interest +/- sd.
Solid: no shrinkage; dashed: shrinkage. (a): n=5, (b): n=40.

edgeR, exact test

Two-group model[1], $X_j = 1$: sample j is in group 2

$$Y_{ij} \sim NB(\mu_{ij}, \varphi_i)$$
$$\mu_{ij} = g^{-1}(\beta_0 + \beta_1 X_j)$$

Define partial sums: $Z_1 = \sum_{j=1}^{n_1} Y_{ij}$ $Z_2 = \sum_{j=n_1+1}^{n_1+n_2} Y_{ij}$

Then under $H_0: \beta_1 = 0$,

$Z_1 \sim NB(n_1\mu, \varphi/n_1)$ $Z_2 \sim NB(n_2\mu, \varphi/n_2)$

[1]approach also valid for k-group setting

edgeR, exact test

P-value computation (similar to Fisher's exact test)

- Let $Z_1 = z_{1,obs}$, $Z_2 = z_{2,obs}$, $Z = Z_1+Z_2 = z$
- $p(z_{1,obs}, z_{2,obs} \mid z) = P(Z_1 = z_{1,obs}, Z_2 = z_{2,obs} \mid Z=z)$
- Let (z_1,z_2) be any integer pair for which $z_1+z_2=z$, p-value:

$$p = \sum_{(z_1,z_2):z_1+z_2=z} p(z_1,z_2 \mid z) I_{\{p(z_1,z_2|z) \le p(z_{1,obs},z_{2,obs}|z)\}}$$

$$= \frac{\sum_{(z_1,z_2):z_1+z_2=z} p(z_1)p(z_2) I_{\{p(z_1)p(z_2) \le p(z_{1,obs})p(z_{2,obs})\}}}{\sum_{(z_1,z_2):z_1+z_2=z} p(z_1)p(z_2)}$$

In words: sum of all partitions of z into z_1 and z_2 which are more extreme than $(z_{1,obs}, z_{2,obs})$

edgeR, exact test

```
#R example exact test; n1 = 4; n2 = 4;
#for n = 1: mu =6; size = phi^{-1} = 2.5; so for n=4: mu4 =
#4*6=24. phi4 = (1/(2.5))/4 = 1/10, so size4 = 10
#Observed sum: Z = z = 50; Z1 = z1obs = 10; Z2 = z2obs = 40.
#Due to equal sample sizes, all partitions of z for which
#min(z1,z2) <= 10 result in p(z1,z2) <= p(z1obs,z2obs)

mu4 = 24; size4=10

pval <- function(obs){
sumnbinomobs <- 2*sum(sapply(0:obs,function(i)
{dnbinom(i,mu=mu4,size=size4)*dnbinom(50-i,mu=mu4,size=size4)}))
sumnbinom <- 2*sum(sapply(0:25,function(i)
dnbinom(i,mu=mu4,size=size4)*dnbinom(50-i,mu=mu4,size=size4)))
return(sumnbinomobs/sumnbinom)
}

#pvalue for (z10bs,z2obs) given z=50:
pval(10)

#And now the entire null-distribution:
distr <- sapply(0:25,pval)
distr
```

edgeR, LRT

The edgeR model (NB-GLM) fits in the GLM setting: once φ_i known: standard likelihood-ratio testing applies.

1. Test H_0: $\beta_i = 0$
2. Perform maximum likelihood estimation under H_0
3. Perform unrestricted maximum likelihood estimation: **b**
4. LRT = log(ML_0(**Y; b**$_{(-i)}$) - log(ML(**Y; b**), where **b**$_{(-i)}$ denotes the estimate of β under restriction $\beta_i = 0$
5. Under H_0 : $LRT^1 \sim \chi^2(df=1)$

Also applies to k parameters: LRT ~ $\chi^2(df=k)$

edgeR

Demo

See ‘edgeRdemo.R’

edgeR

Plusses and Minuses

+ Easy to use

+ Fast

+ Solid, when sample sizes are medium to large

- A:Cannot deal well with multiple sources of variation (random effects). E.g. within person and between persons

- B: Cannot deal well with paired data

- C: Does not deal well with excess of zeros. Consider filtering out rows with, say, more than 70% zeros.

edgeR

Ad A: Multiple sources of variation (random effects)

E.g. Data on 6 individuals, multiple (3) locations of tumor

Ignore 2 sources of variation

↓

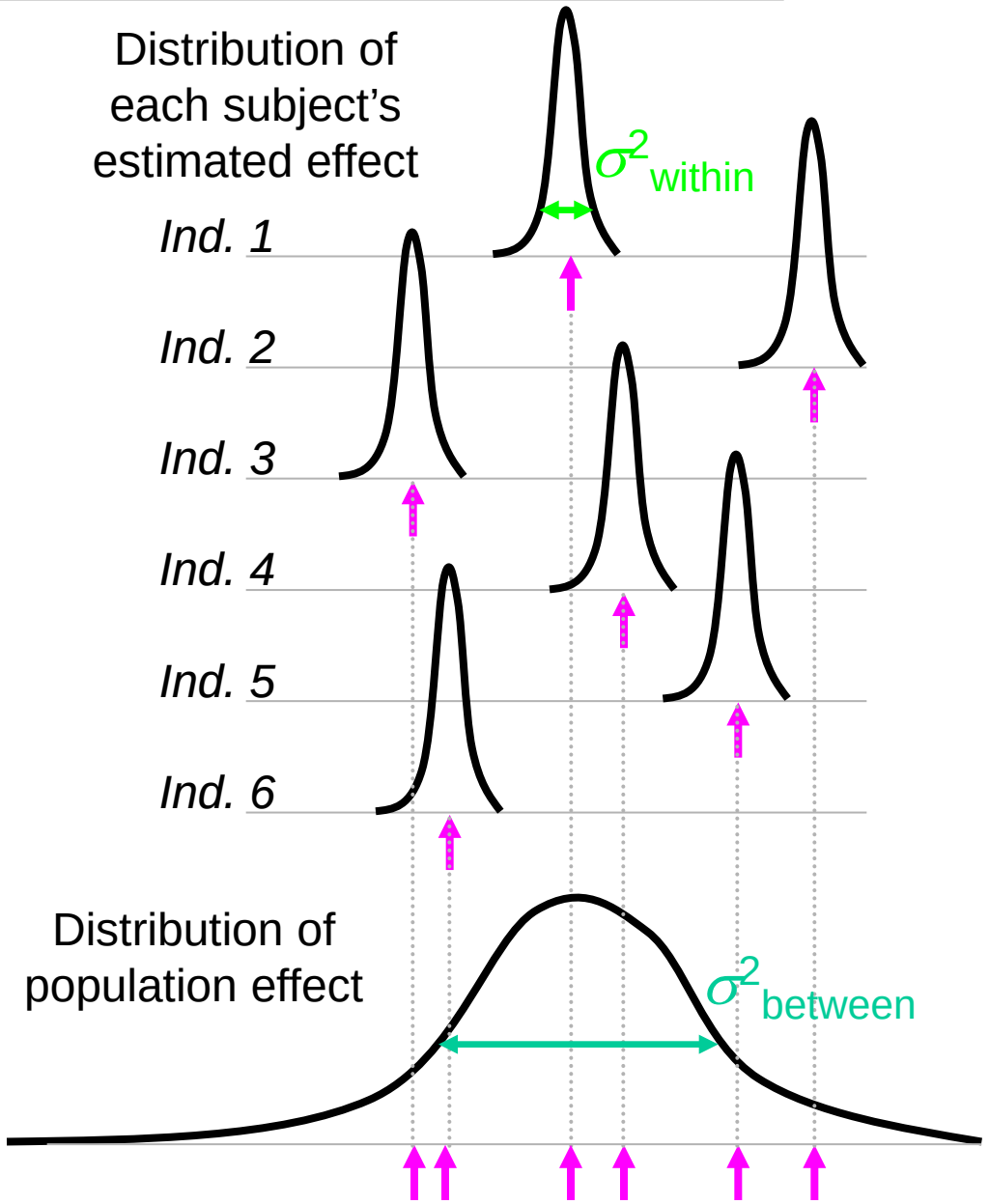
18 measurements treated as independent

↓

Too small sd σ

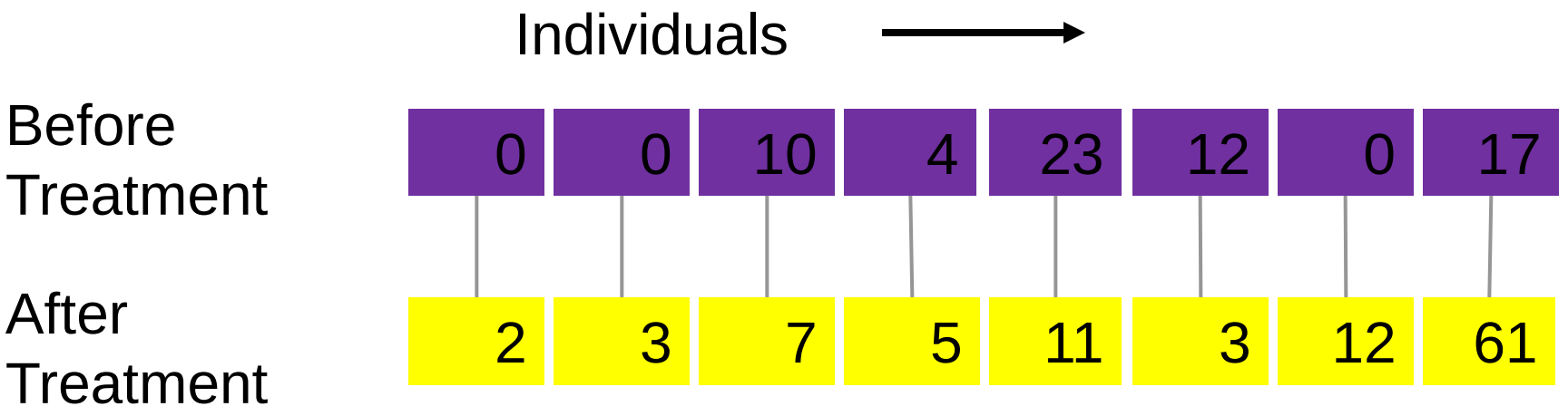
↓

Too optimistic p-values



edgeR

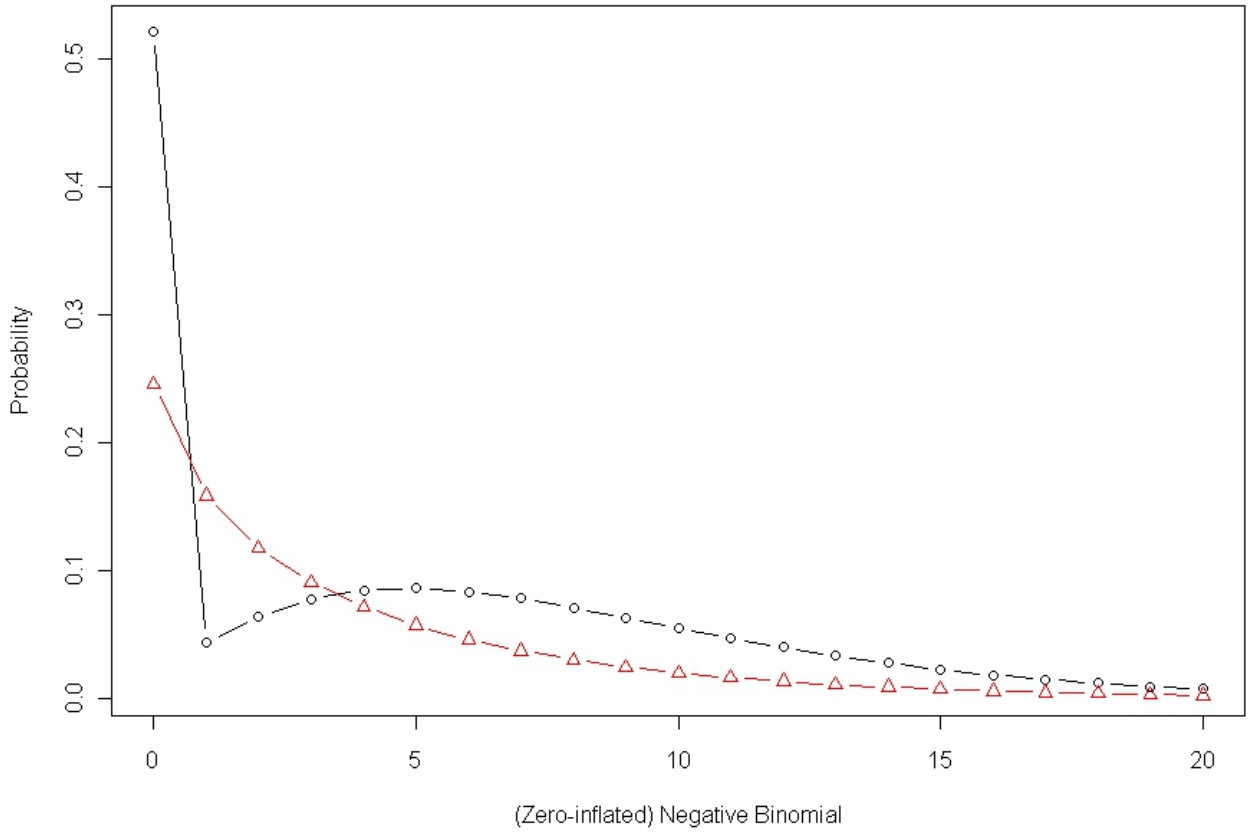
Ad B: paired data, tag i, i = 1, ..., p



edgeR

Ad C: Excess of zeros

0	12	0	0	0	3	5	2	10	7	8	0	0	9	0	3	0	1	2	0	0	14	5	6	4	0	11	0	0	7



Black: true (zero-inflated). Red: best fitting neg. binomial

ShrinkBayes

Alternative R-package (Next week)

+ Better for small sample sizes

+ Can deal with pairs, random effects, excess of zeros

+ More reproducible results

- More difficult to use

- More time-consuming

- Bayesian concept, less familiar to most biologists