Multiple group comparisons for RNA-Seq and stable effect size estimates

Michael Love

Irizarry Lab
Department of Biostatistics
Dana Farber Cancer Institute &
Harvard School of Public Health

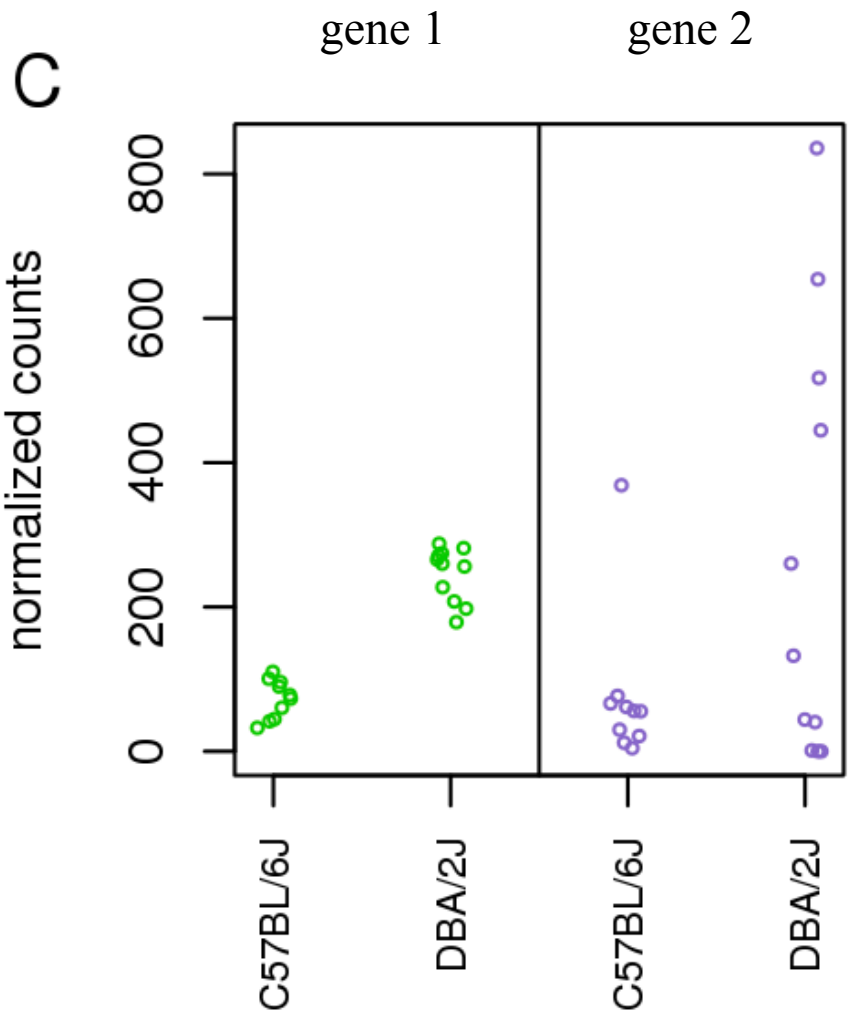
DANA-FARBER
CANCER INSTITUTE

HARVARD
SCHOOL OF
PUBLIC HEALTH

DESeq2
4/2013 Bioc package
2/2014 preprint bioRxiv

Simon Anders, EMBL
Wolfgang Huber, EMBL

Estimating effect sizes for counts



effect of:

- treatment
- species
- tissue, etc.

Hammer et al, mRNA-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain. Genome Research 2010.

Modeling differences in counts

For read count *K* for gene *i*, sample *j*...

mean dispersion

$$K_{ij} \sim NB(\mu_{ij}, \alpha_i)$$

size factor

$$\mu_{ij} = s_{ij} q_{ij}$$

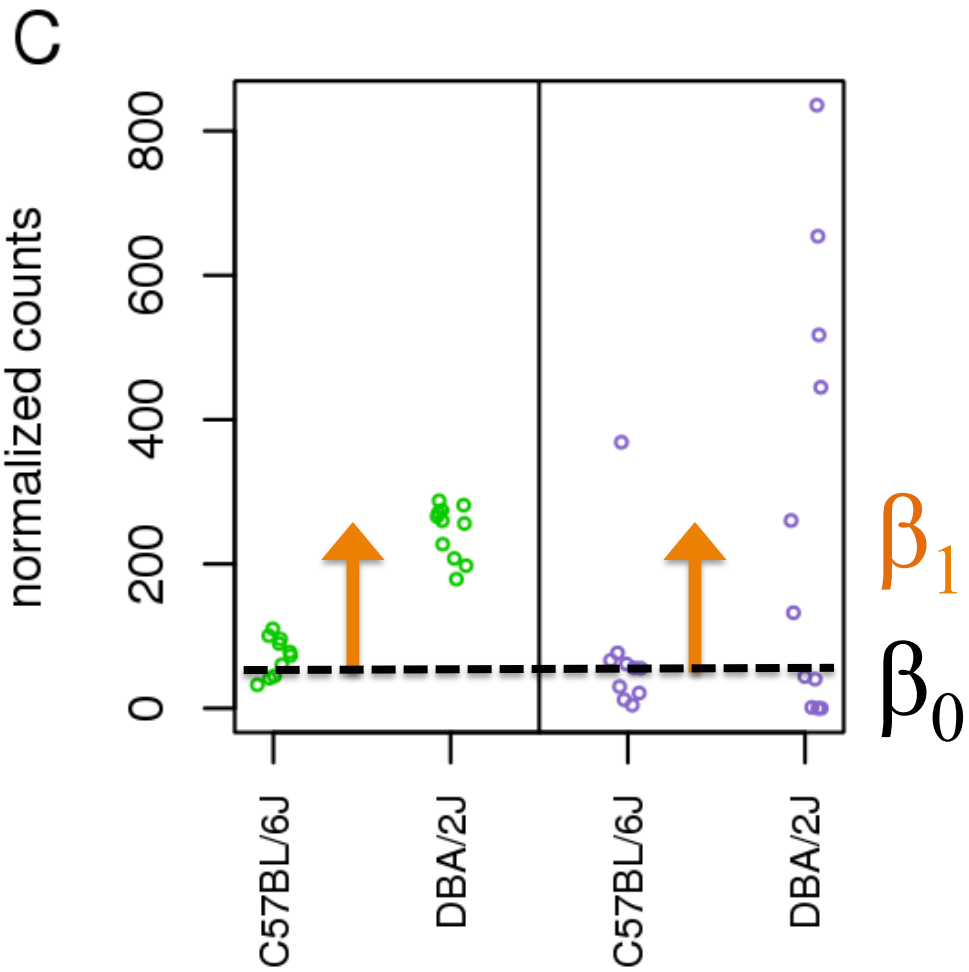
$$\log q_{ij} = \sum_r x_{jr} \beta_{ir}$$

predictors

e.g.

$$\begin{bmatrix} \log q_1 \\ \log q_2 \\ \log q_3 \\ \log q_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Two genes with equal effect size



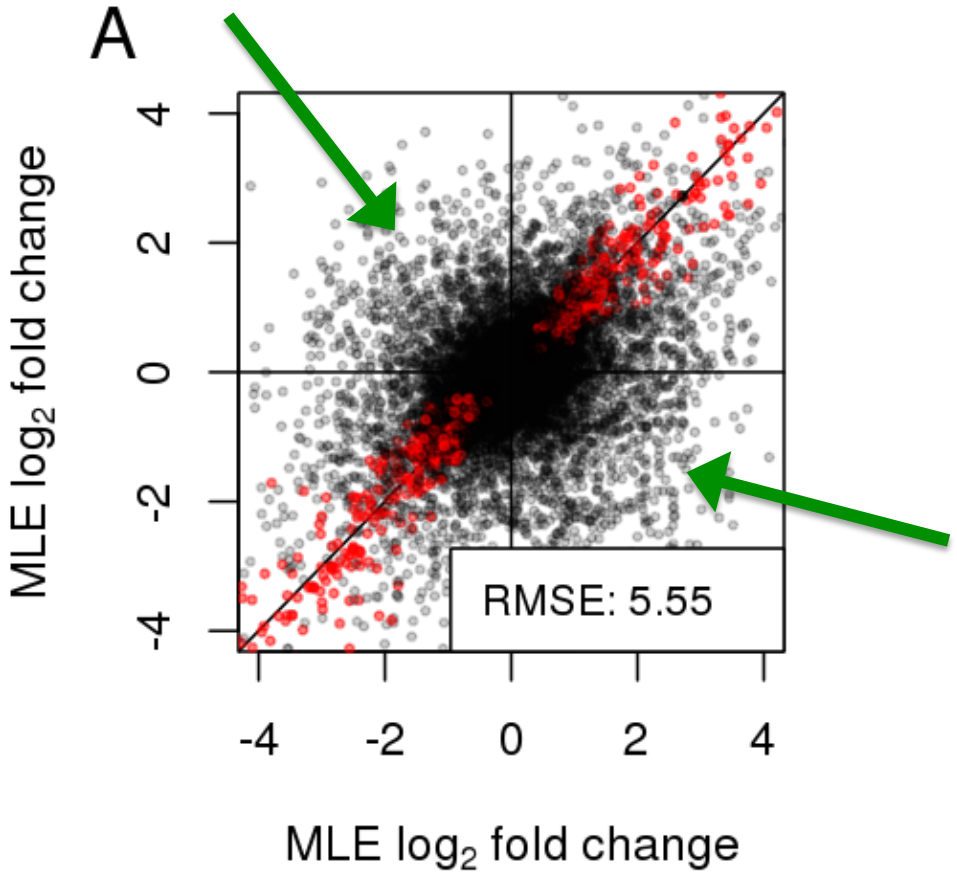
Hammer et al, mRNA-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain. Genome Research 2010.

Maximum likelihood estimates can have high variance

Split 10 vs 11 samples:

- 5 vs 5
- 5 vs 6

For every gene, we get a β from both comparisons.

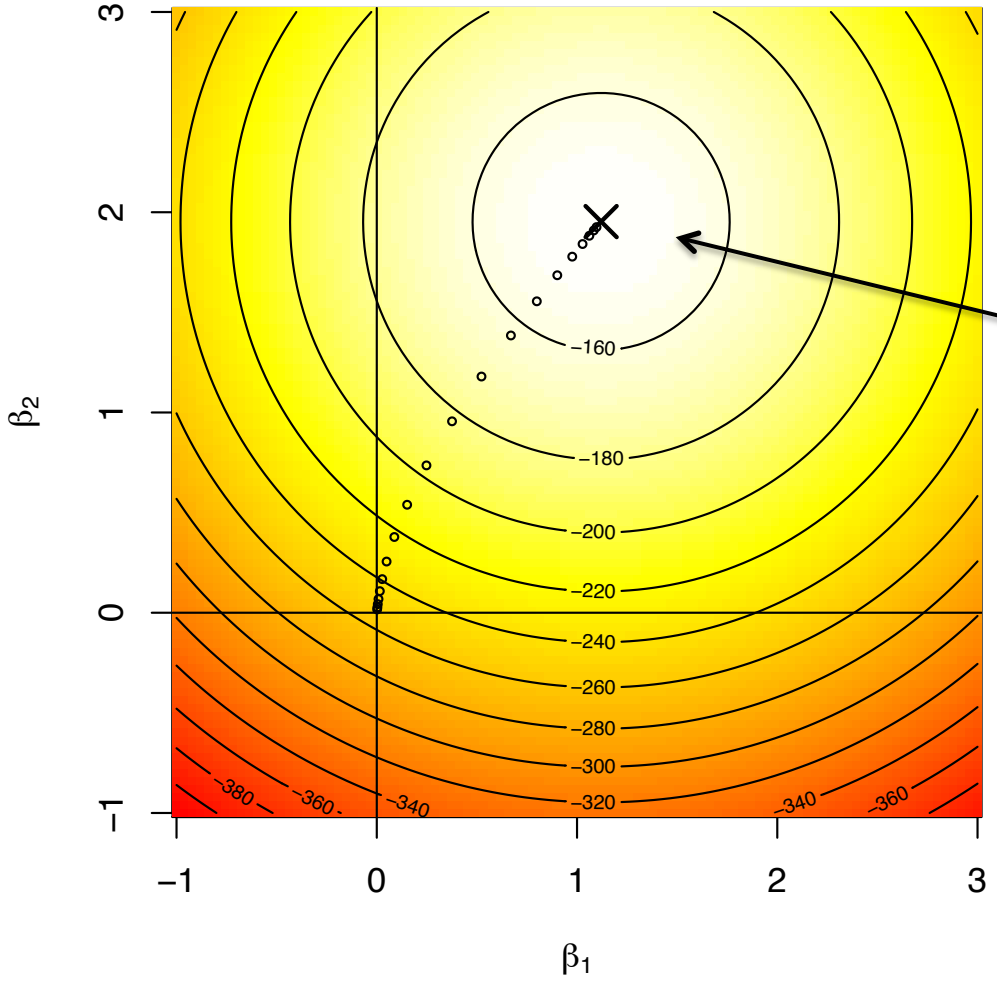


...especially for genes with low counts, but also for genes with high variance

Moderation stabilizes effect sizes

Moderation through a prior: $\beta_{ir} \sim N(0, \sigma_r^2)$ ...for r ≠ 0

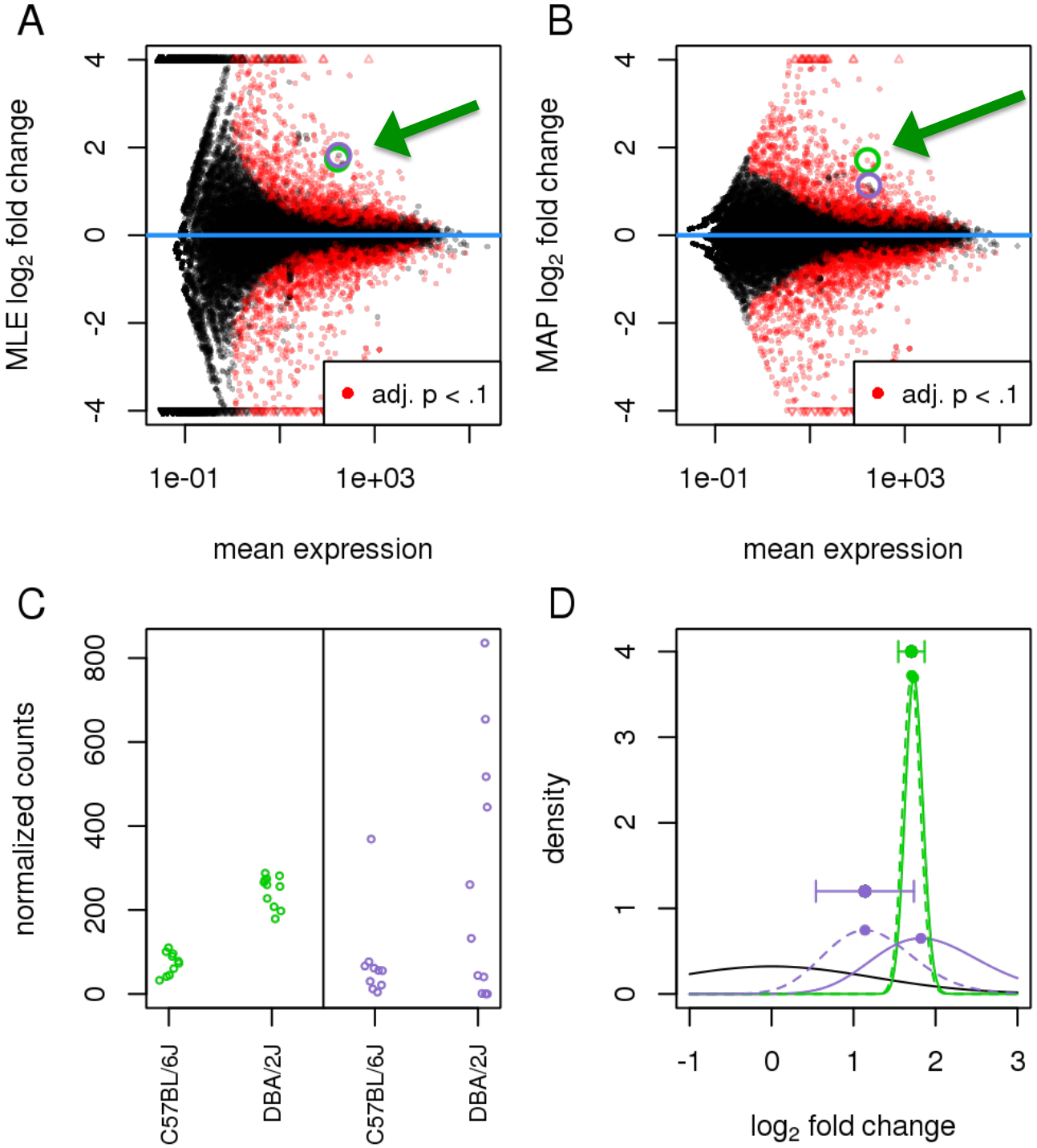
estimated from MLE effect sizes

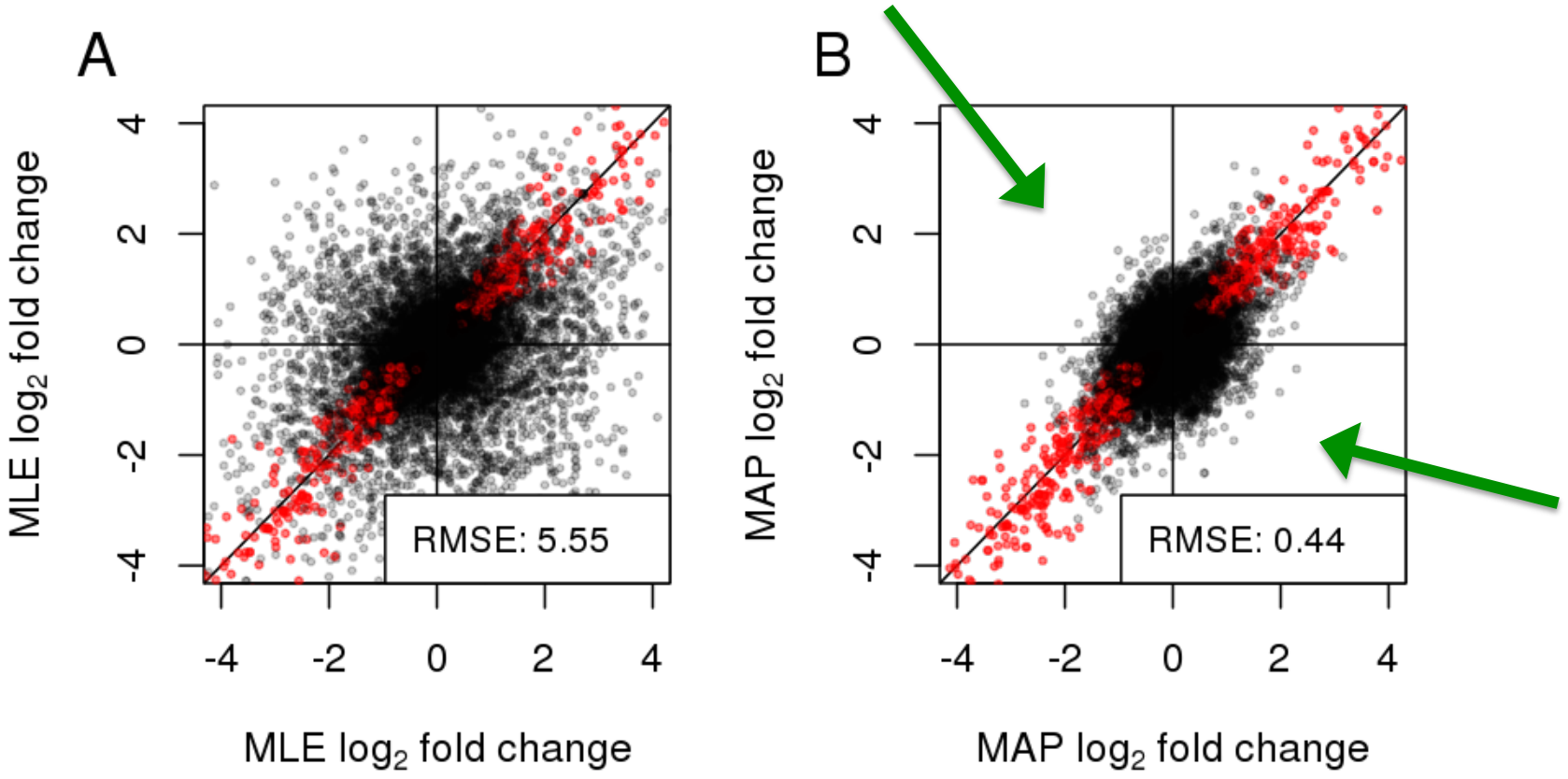


MLE

the max posterior is pulled in towards 0

(here with $\sigma_1=\sigma_2$)


A
MLE log2 fold change
4
2
0
-2
-4
adj. p < .1
1e-01
1e+03
mean expression
B
MAP log2 fold change
4
2
0
-2
-4
adj. p < .1
1e-01
1e+03
mean expression
C
normalized counts
0
200
400
600
800
C57BL/6J
DBA/2J
C57BL/6J
DBA/2J
D
density
0
1
2
3
4
-1
0
1
2
3
log2 fold change



...introduces bias but reduces mean squared error.
for an estimator: MSE = bias^2 + variance

...but moderation with ≥ 3 groups is not symmetric

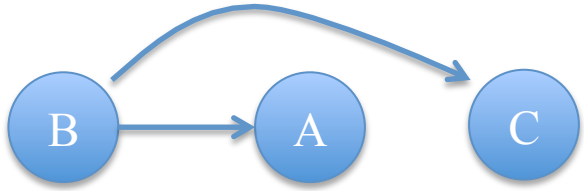
		$\hat{\beta}_{B-A}$	$\hat{\beta}_{C-A}$
A	1	0	0
	1	0	0
B	1	1	0
	1	1	0
C	1	0	1
	1	0	1

β_0
β_{B-A}
β_{C-A}

$\neq$

		$\hat{\beta}_{A-B}$	$\hat{\beta}_{C-B}$
A	1	1	0
	1	1	0
B	1	0	0
	1	0	0
C	1	0	1
	1	0	1

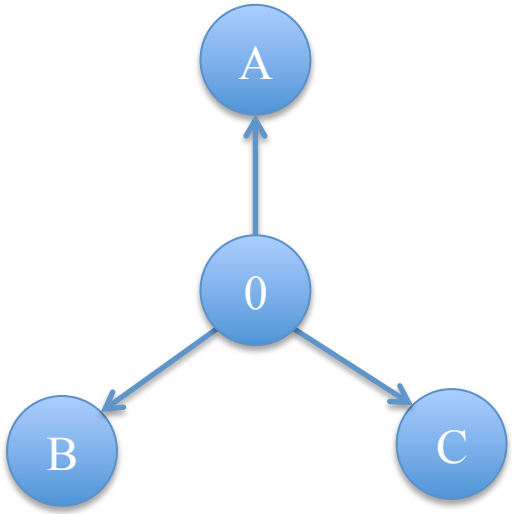
β_0
β_{A-B}
β_{C-B}



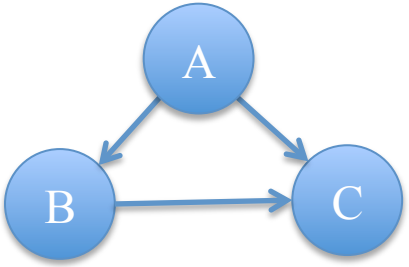
A solution

Expand by adding a column,
shrinking all groups towards an intercept.

	△	△	△		
A	1	1	0	0	
A	1	1	0	0	β_0
B	1	0	1	0	β_A
B	1	0	1	0	β_B
C	1	0	0	1	β_C
C	1	0	0	1	



Estimate prior using all MLE contrasts:



Statistical inference

Test a Wald statistic for each coefficient: $$W = \frac{\hat{\beta}_{ir}}{\widehat{\mathrm{SE}}(\hat{\beta}_{ir})}$$

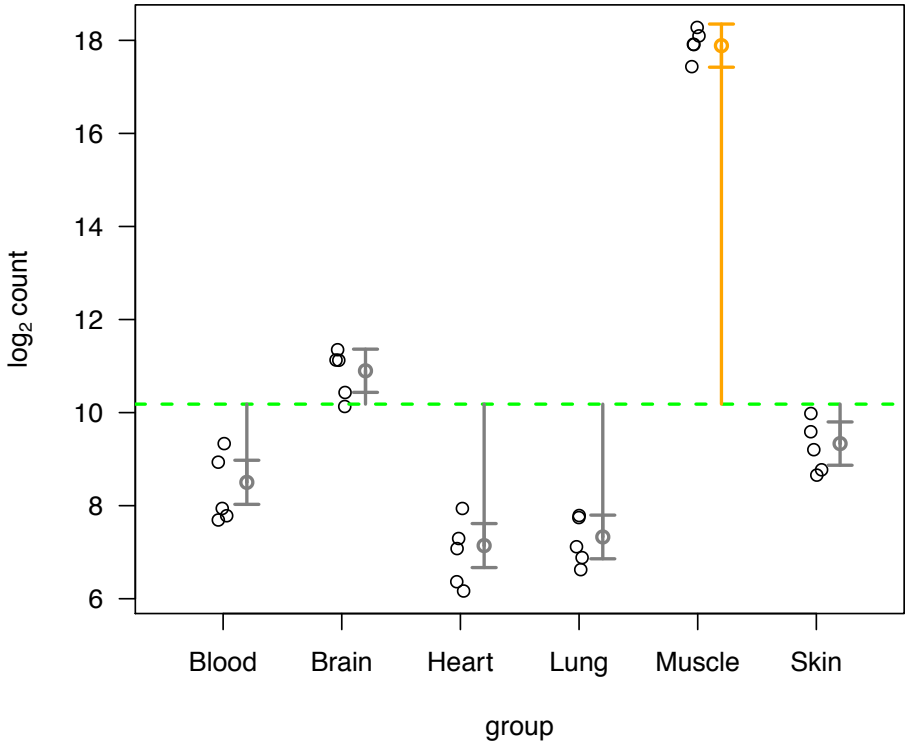
Contrasts can be used to compare multiple levels of a factor, e.g., $c^t = [0,0,-1,1]$

$$\beta_i^c = \vec{c}^{\,t}\vec{\beta}_i$$

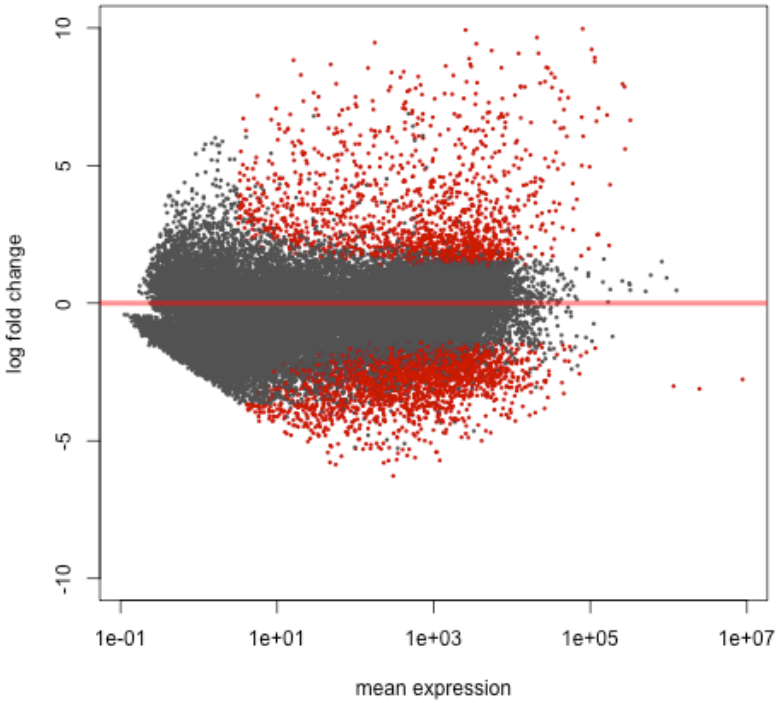
$$\mathrm{SE}(\beta_i^c) = \sqrt{\vec{c}^{\,t}\Sigma_i\vec{c}}$$

GTEx: tissue specific

one gene, highest Wald statistic



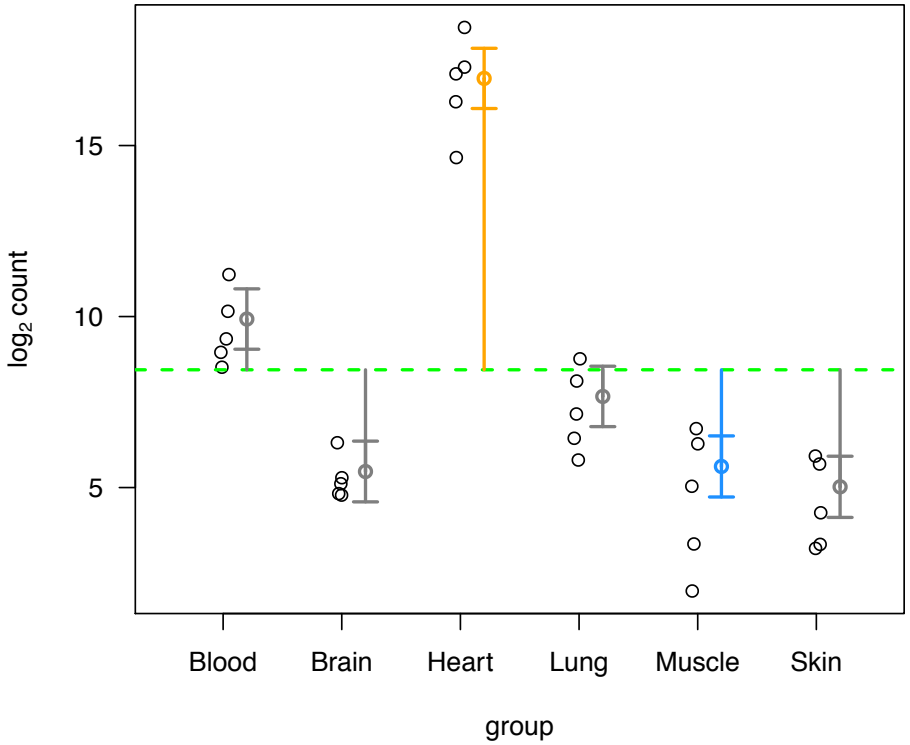
MA-plot all genes



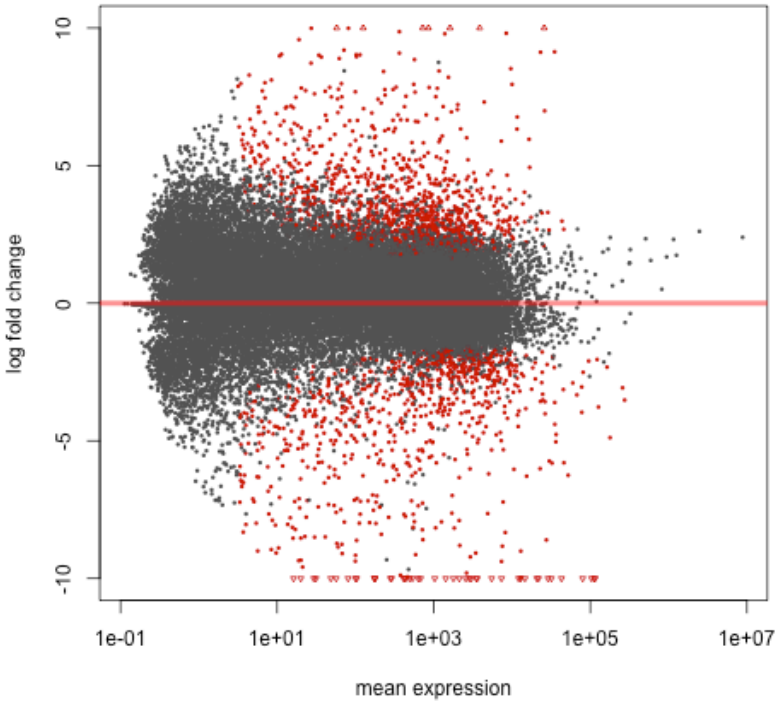
red points:
more than doubling
adjusted p-value < 0.1

GTEx: contrast two tissues

one gene, highest Wald statistic



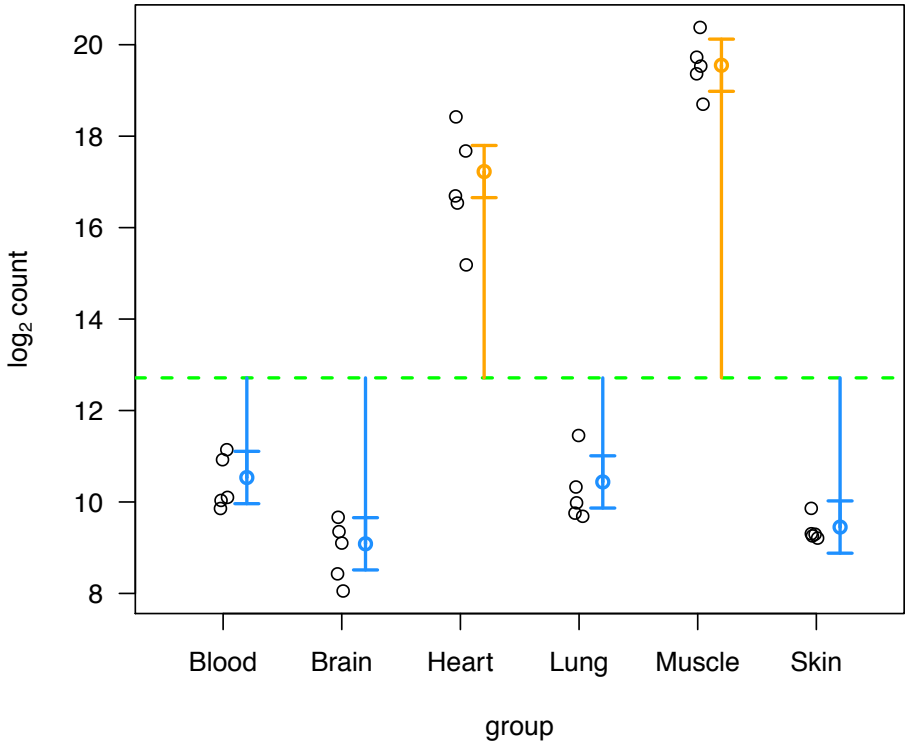
MA-plot all genes



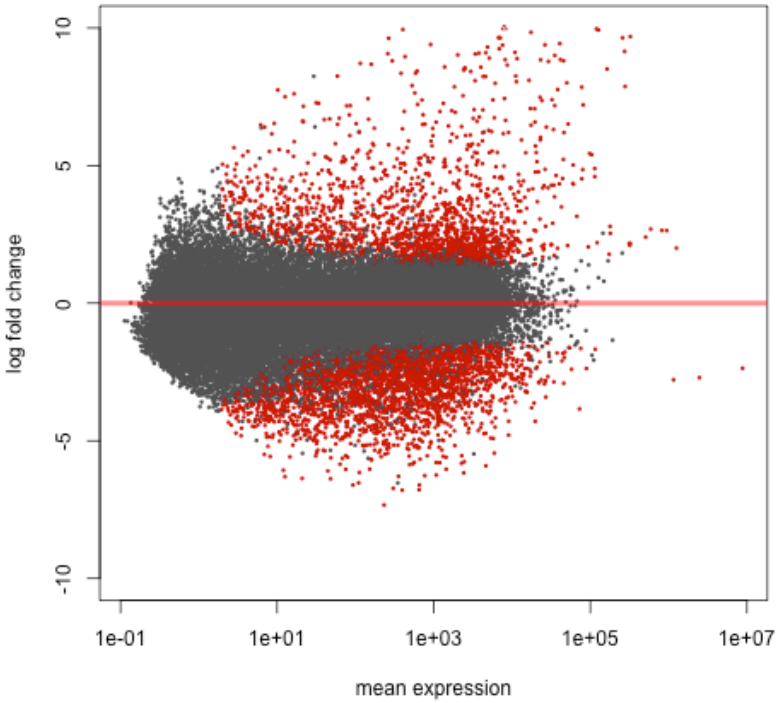
red points:
more than doubling
adjusted p-value < 0.1

GTEx: contrast 2 vs 4 tissues

one gene, highest Wald statistic



MA-plot all genes



red points:
more than doubling
adjusted p-value < 0.1

Acknowledgments

- Simon Anders, EMBL
- Wolfgang Huber, EMBL

- Rafael Irizarry, DFCI/HSPH
- Martin Vingron & Knut Reinert, IMPRS, MPIMG, FU

Related work:

- For RNA-Seq: GFOLD, BitSeq, ShrinkBayes, NPEBSeq
- Genereally: selection bias, "Winner's curse", adaptive shrinkage, Brad Efron, Noah Simon, Matthew Stephens et al.
- Background on bias-variance trade-off: "Elements of Stat. Learning"

More:

- Effect size shrinkage in `DESeq2`

- Preprint available on bioRχiv (search "`biorxiv DESeq2`")

- Code for GTEx example: `github.com/mikelove/multigroup`

GLM with ridge regularization

maximum *a posteriori*

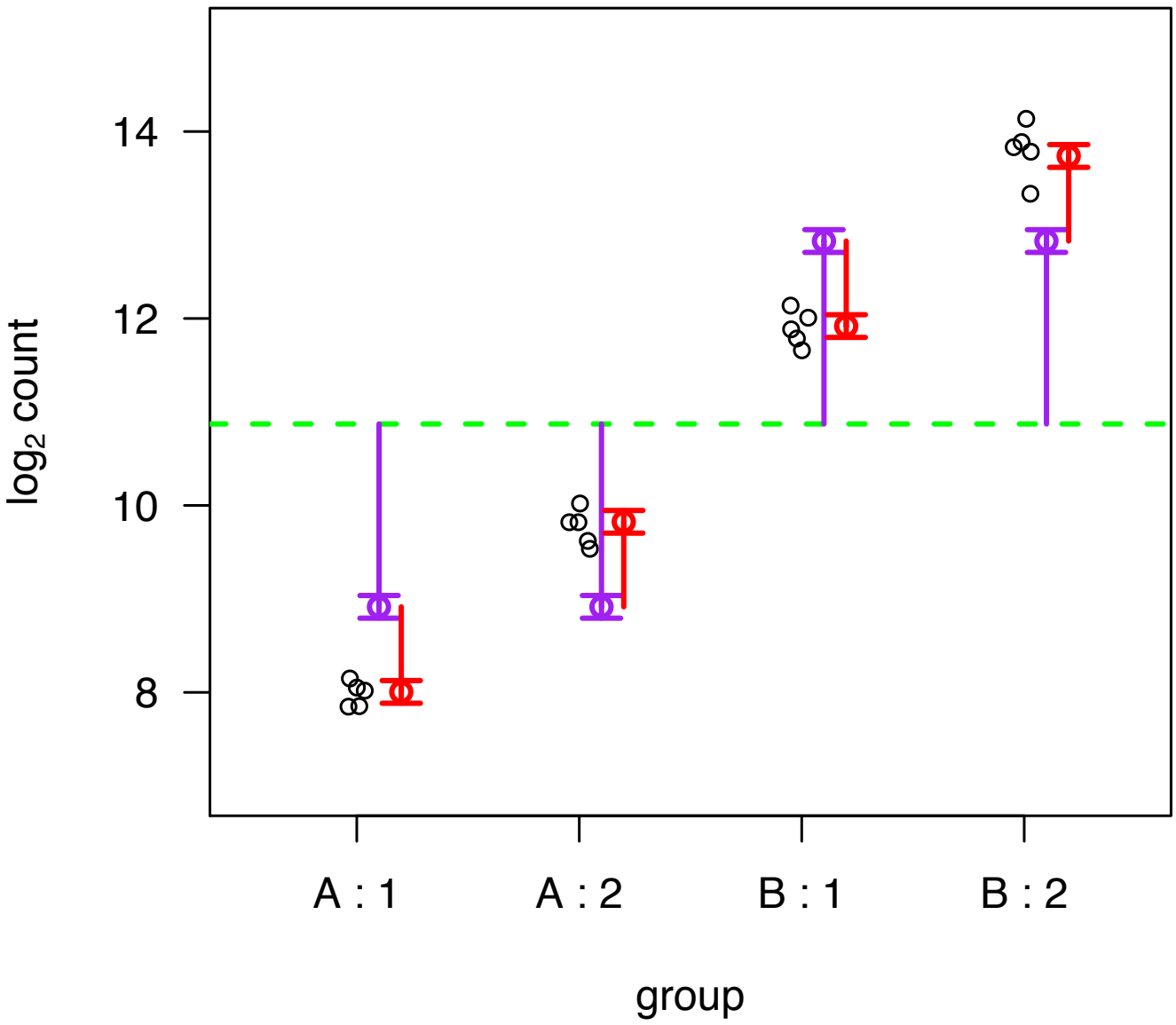
log likelihood

log prior

$$\vec{\beta}_i = \arg\max_{\vec{\beta}} \left(\sum_j \log f_{\mathrm{NB}} \left(K_{ij}; \mu_j(\vec{\beta}), \alpha_i \right) + \Lambda(\vec{\beta}) \right)$$

$$\beta_{ir} \sim N(0, \sigma_r^2) \longrightarrow \Lambda(\vec{\beta}) = \sum_r \frac{-\beta_r^2}{2\sigma_r^2}$$

the "penalty term" for large β


simulated multifactor
log2 count
14
12
10
8
A : 1
A : 2
B : 1
B : 2
group