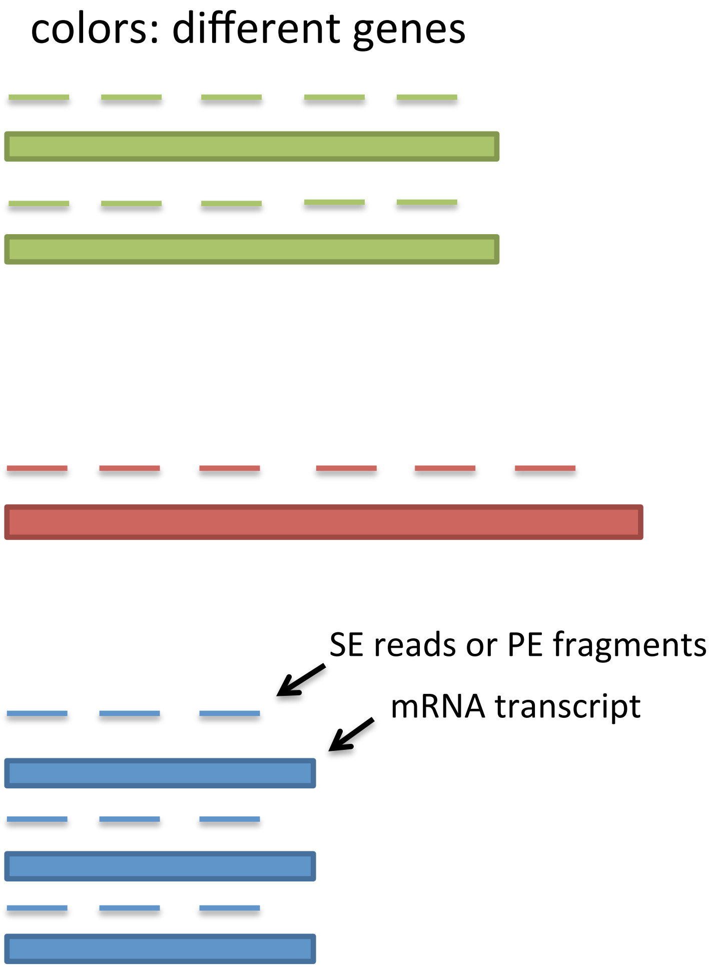
RNA-seq data analysis and differential expression part II

Michael Love
Irizarry Lab
Dept. of Biostatistics
Dana-Farber Cancer Institute &
Harvard T.H. Chan SPH

Outline

1. counts and sampling
2. shrinkage estimators
   - dispersion
   - fold changes
   - regularized logarithm
3. statistical power
   - independent filtering
   - threshold tests

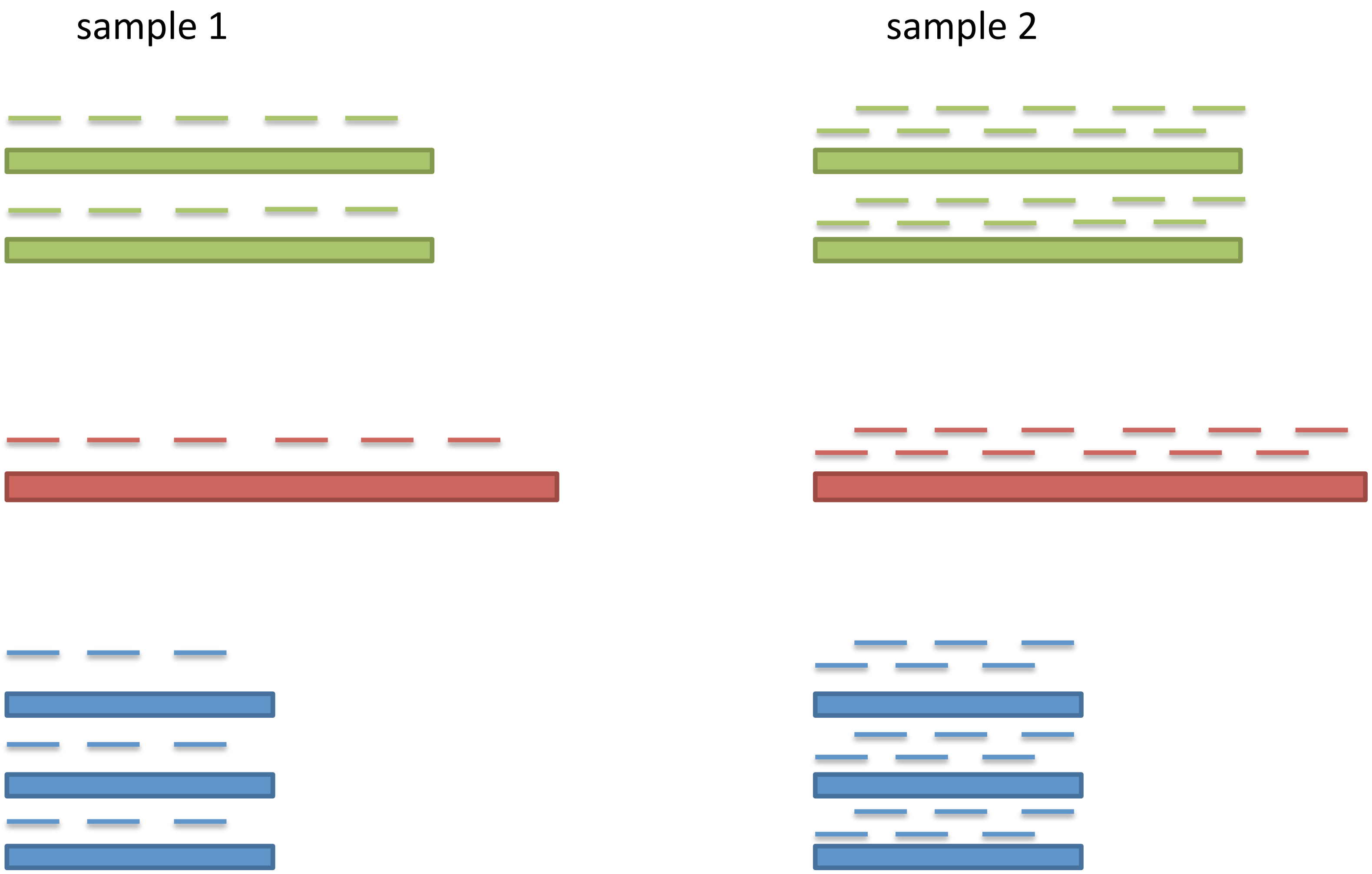
mRNAs to fragments



number of mapped fragments proportional to:

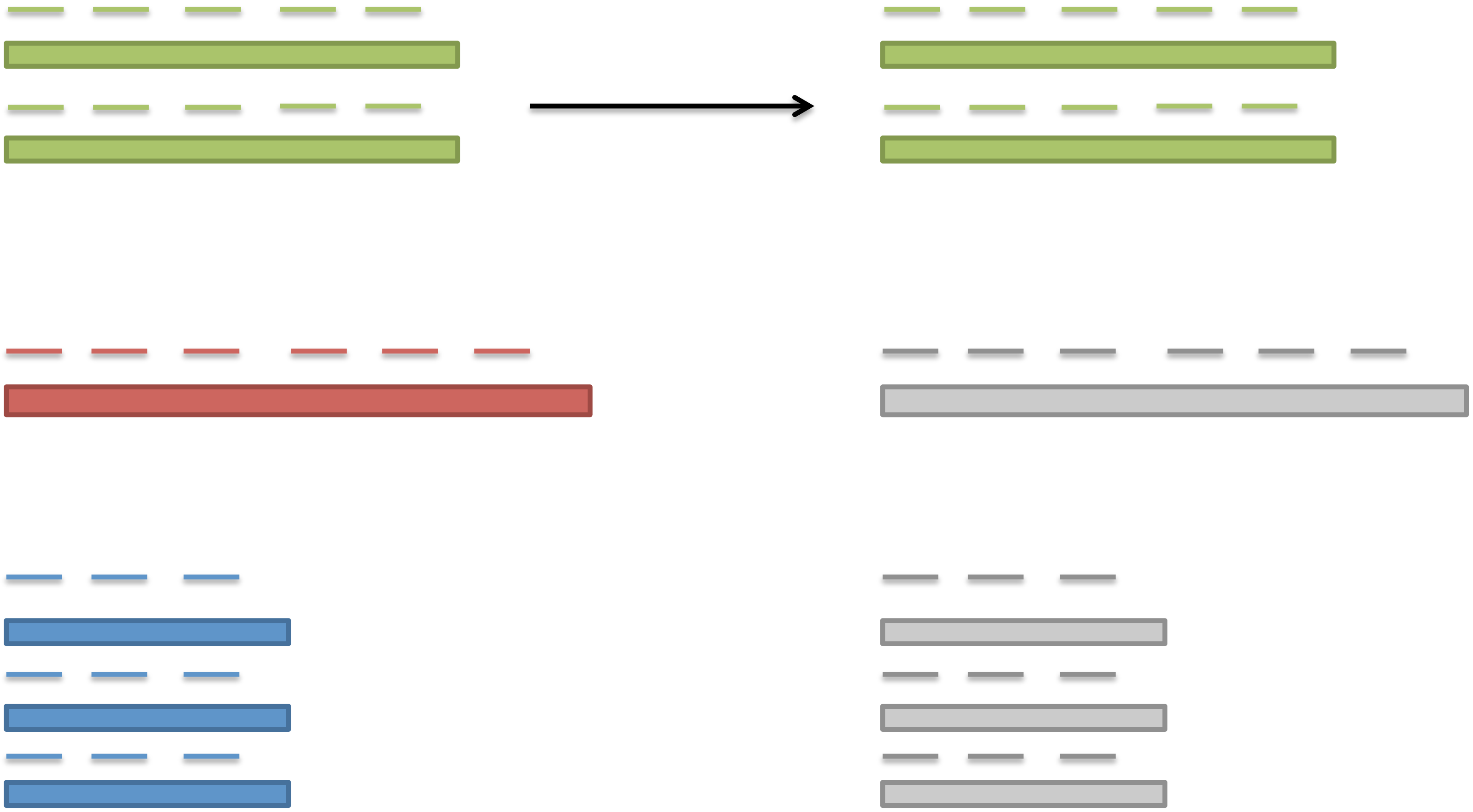
- expression of RNA
- length of gene
- sequencing depth
- lib. prep. factors (PCR)
- in silico factors (alignment)
- ...

Sequencing depth



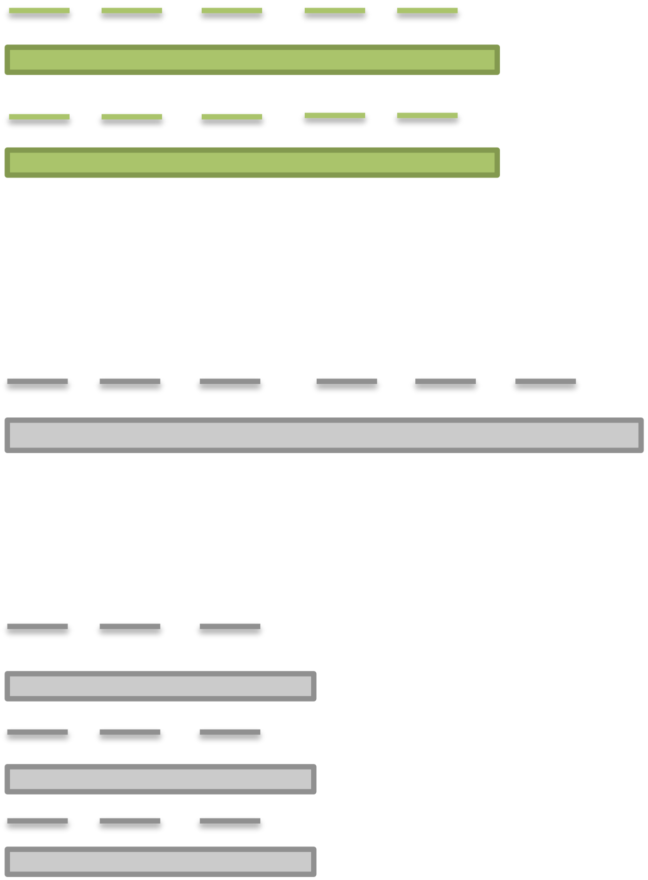
Variance of counts

Consider one gene:

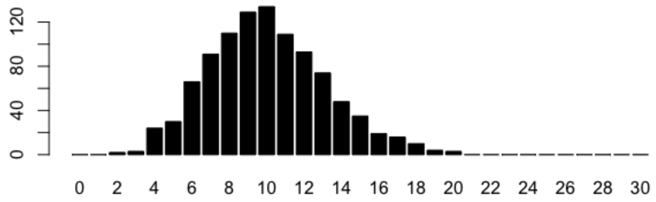


Variance of counts

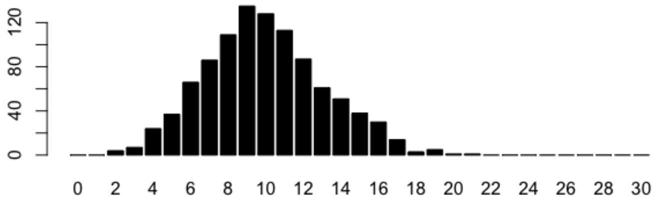
Consider one gene:



- Binomial sampling distribution

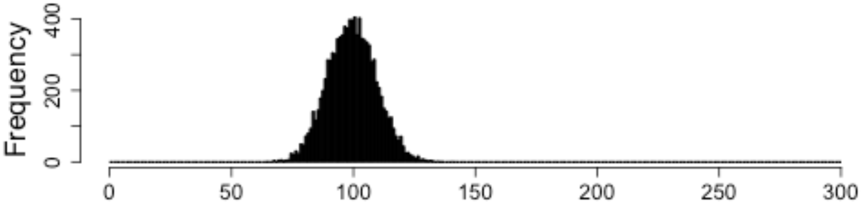


- With millions of reads & small proportion for each gene => Poisson sampling distribution

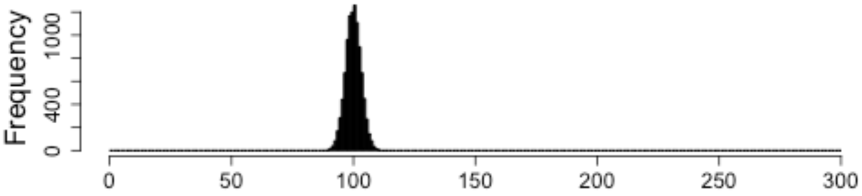


Raw counts vs. normalized counts

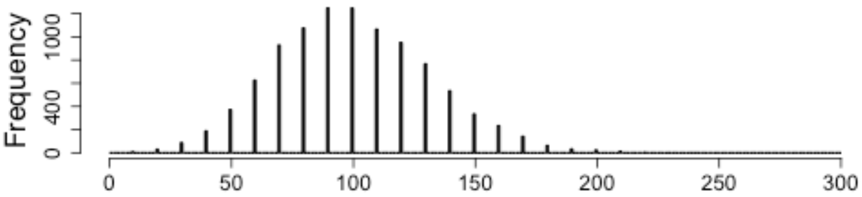
Raw count with mean of 100
Poisson sampling, so SD=10



Raw count mean = 1000
Scaled by 1/10
SD = ?



Raw count mean = 10
Scaled by 10
SD = ?



Raw counts vs normalized counts

raw count for gene i, sample j

normalization factor

quantity of interest

$$K_{ij} \sim \mathcal{L}(\mu_{ij} = s_{ij} q_{ij})$$

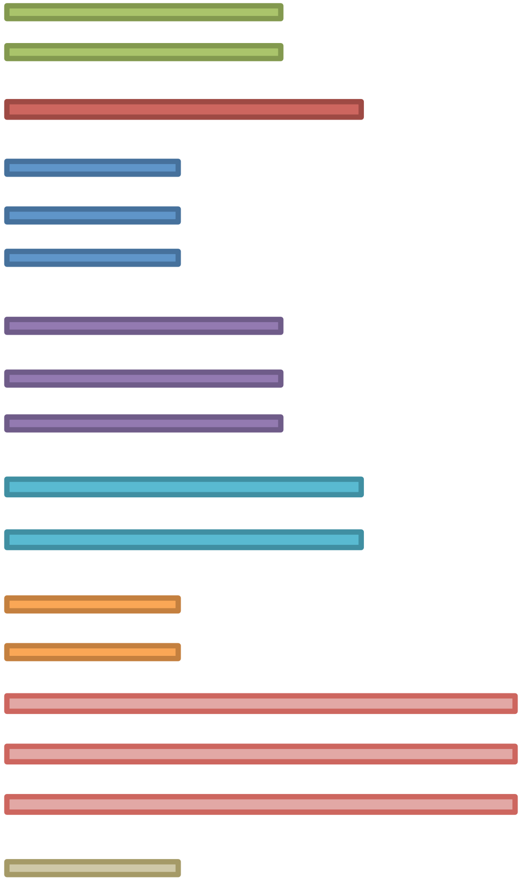
preferred

$$\frac{K_{ij}}{s_{ij}} \sim \mathcal{L}(\mu_{ij} = q_{ij})$$

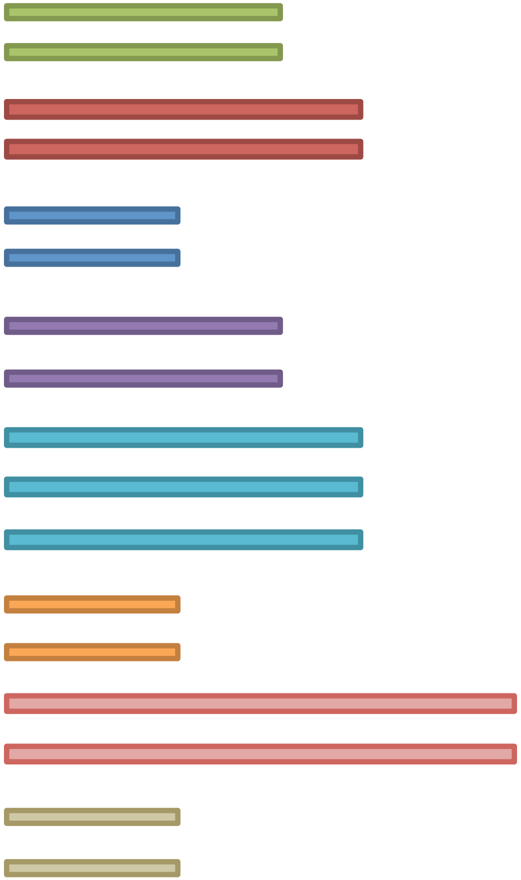
some distribution

Biological replicates

If the proportions of mRNA stays exactly constant ("technical replicate") we can expect Poisson dist.

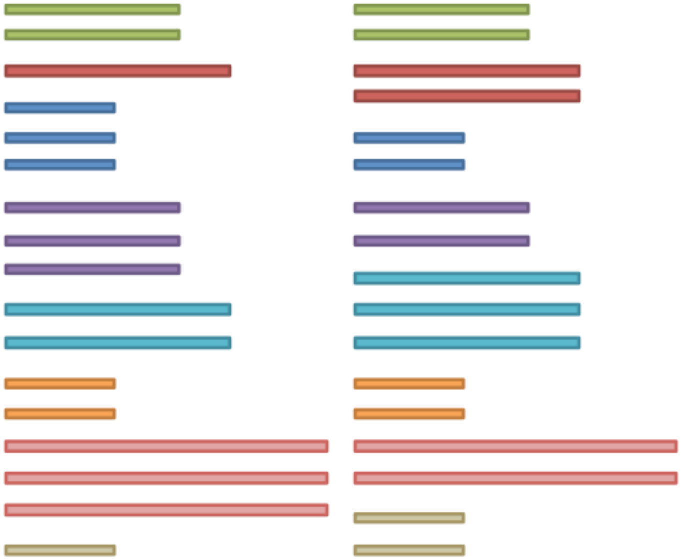


But realistically, biological variation across sample units is expected

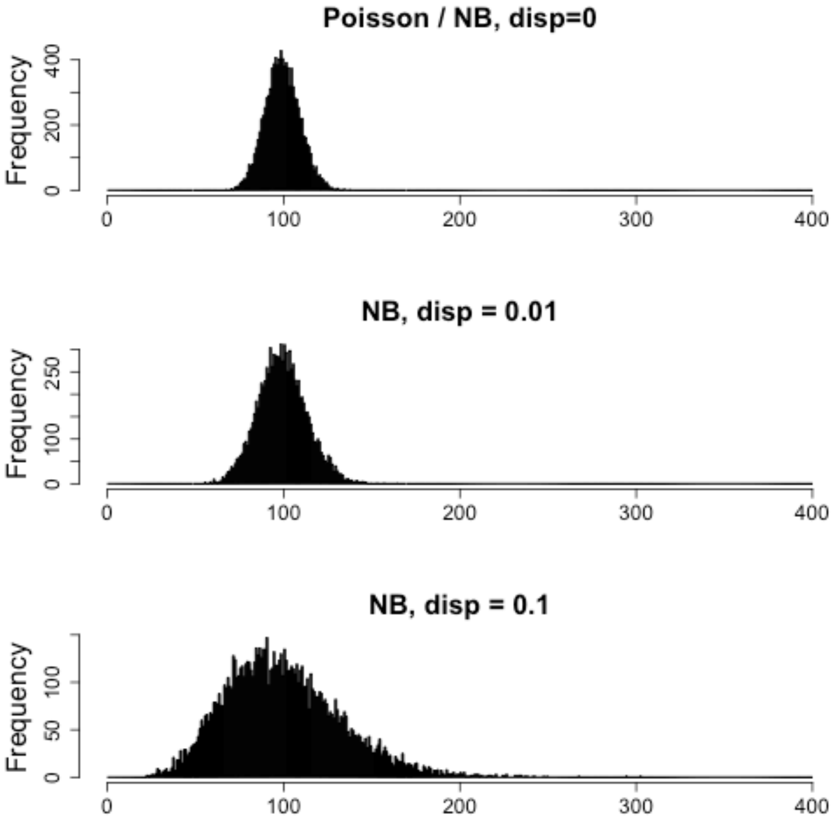


Biological replicates

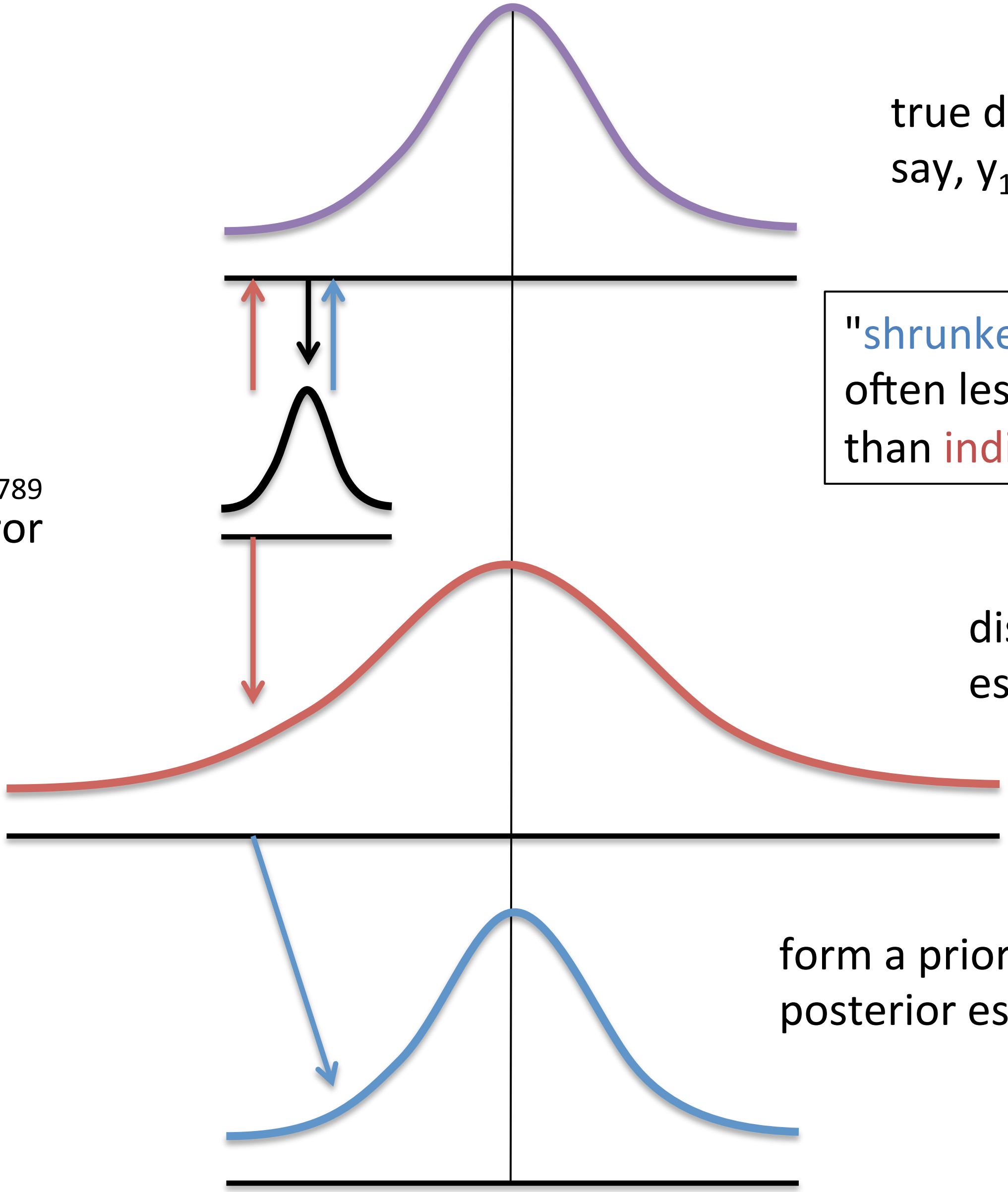
Biological variation for the abundance of a given gene produces "over-dispersion" relative to the Poisson dist.



Negative Binomial =
Poisson with a varying mean



2. Shrinkage estimators



true distribution of y's
say, y_1 - y_{1000}

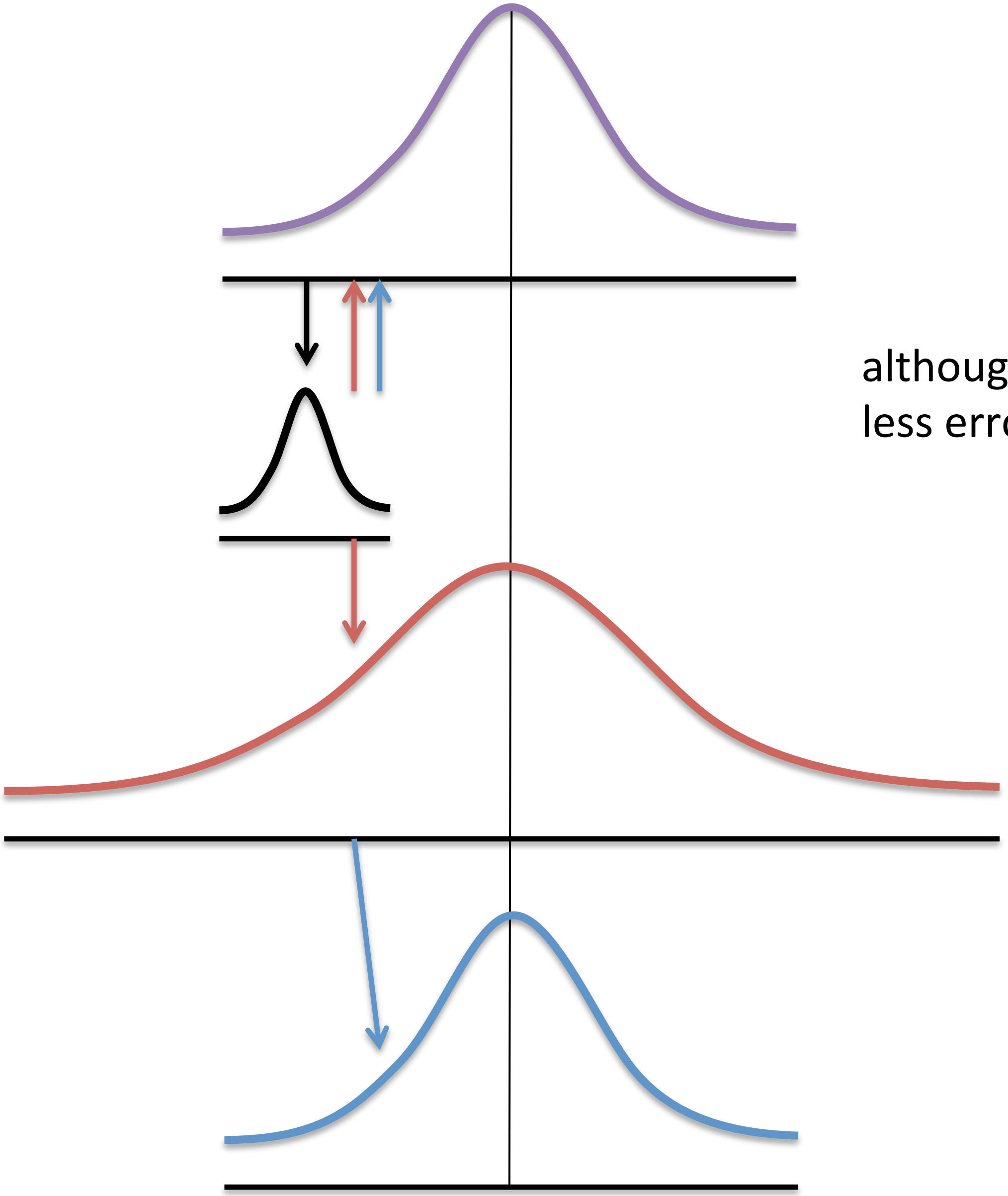
"shrunken" estimates
often less error overall
than individual estimates

observations of y_{789}
with sampling error

distribution of
estimates of y

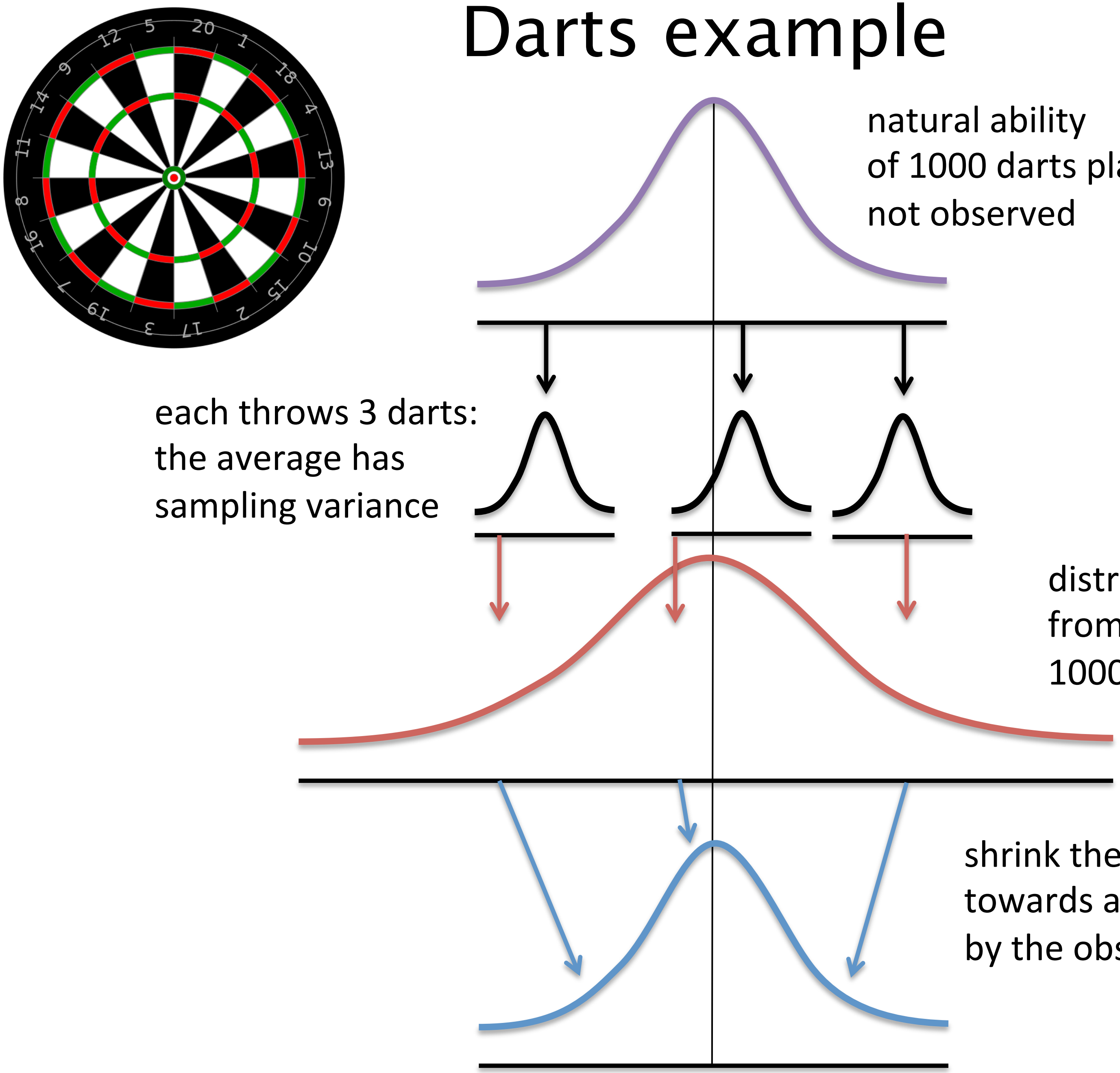
form a prior from the distribution
posterior estimates for each y

Shrinkage estimators



although not necessarily less error for *every* y

Darts example



natural ability
of 1000 darts players:
not observed

each throws 3 darts:
the average has
sampling variance

distribution of average
from 3 throws from all
1000 players

shrink the averages
towards a center defined
by the observed distribution

Shrinkage estimators in genomics

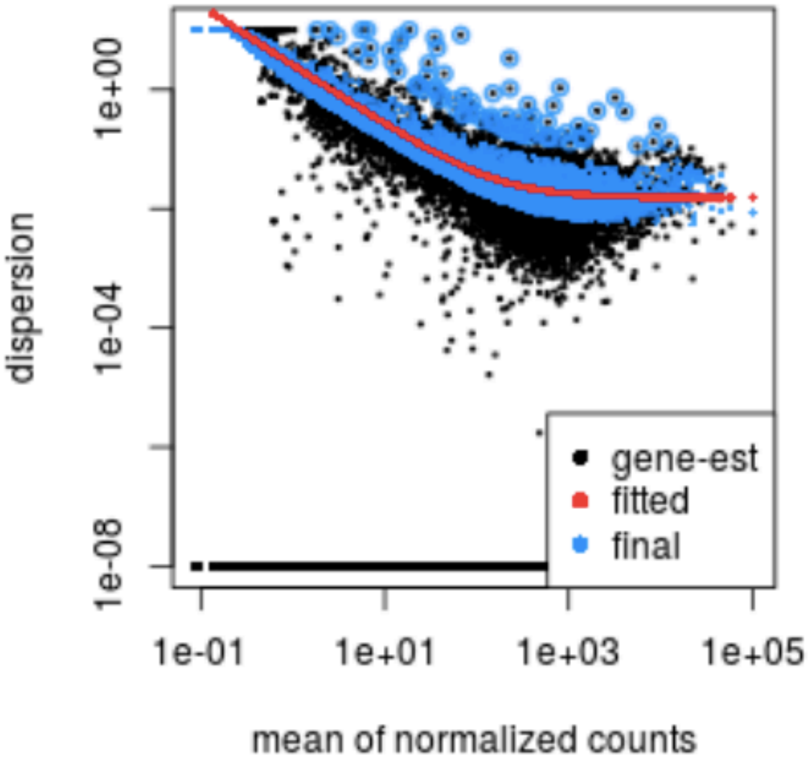
- Lönnstedt and Speed 2002: microarray
- Smyth 2004: limma for microarray
- Robinson and Smyth 2007: SAGE, digital gene exprs.
- Many adaptations: DSS and DESeq2 are a similar approach, data-driven strength of shrinkage

An introduction to shrinkage estimators:

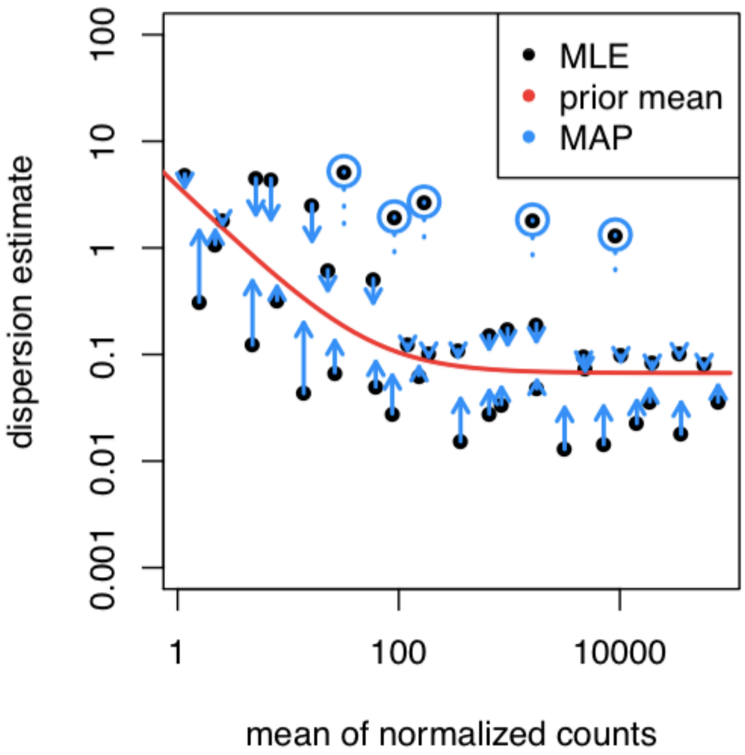
Baseball players as example
Efron and Morris 1977
"Stein's Paradox in Statistics"

Shrinkage of dispersion

all genes (Pasilla)

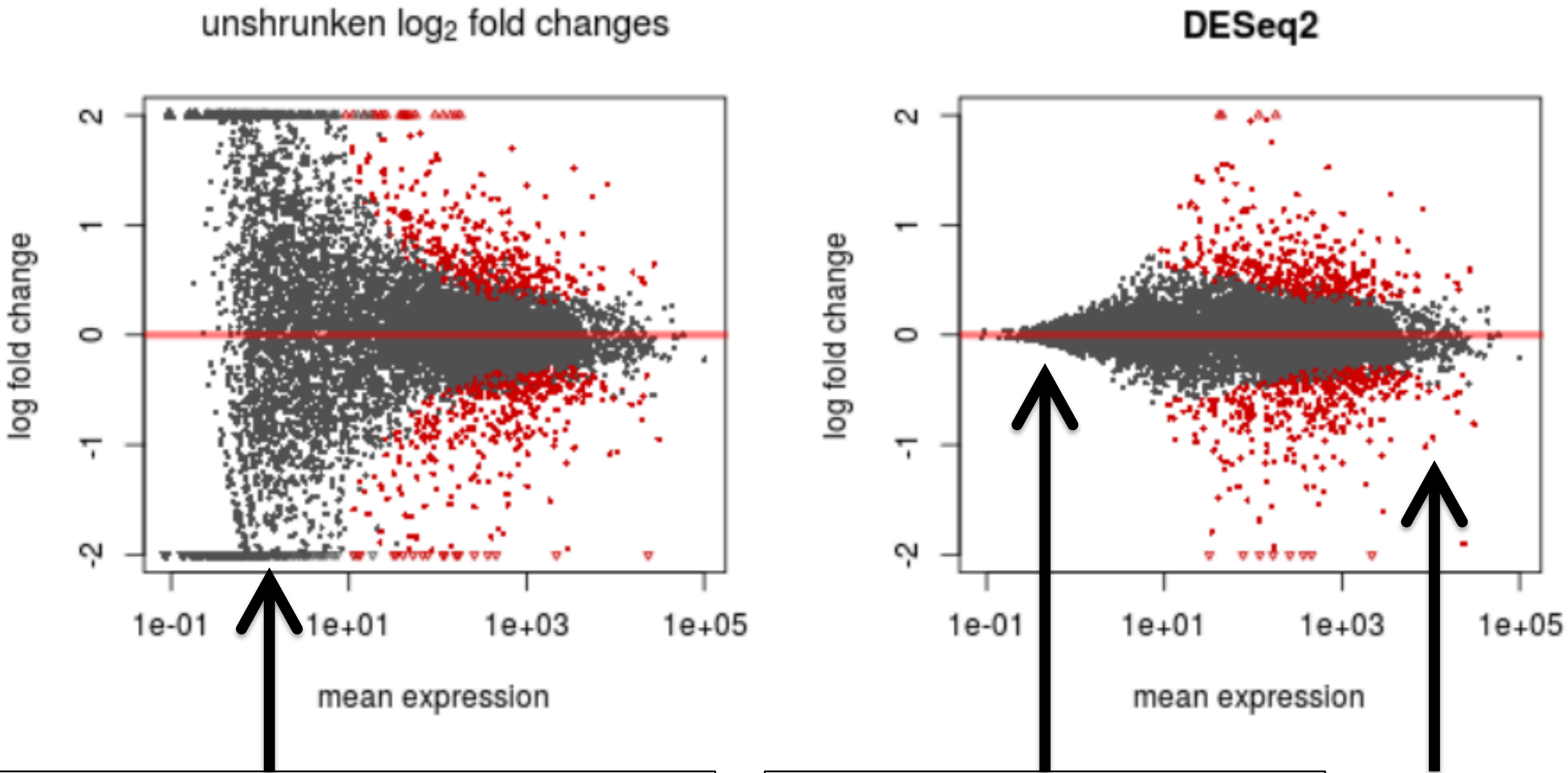


a subset of genes (Pickrell)



1. Gene estimate = maximum likelihood estimate (MLE)
2. Fitted dispersion trend = the mean of the prior
3. Final estimate = maximum a posteriori (MAP)

Shrinkage of fold changes

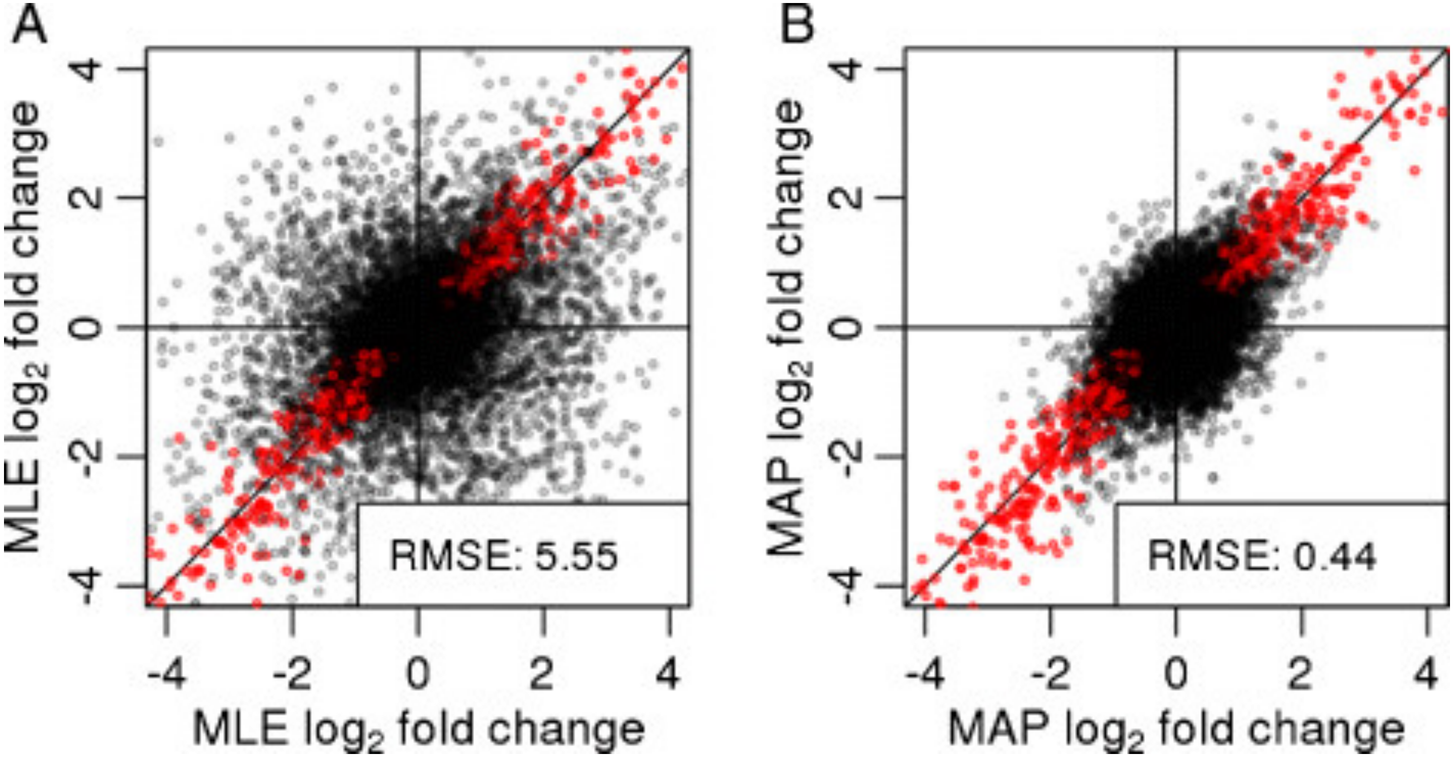


noisy estimates due to low counts
large FDR from the statistical model,
but we shouldn't trust the estimate itself

shrinkage is not equal.
strong moderation for low
information genes: low counts

little to no
shrinkage

Why shrink fold changes?



Split a dataset into two equal parts, compare LFC

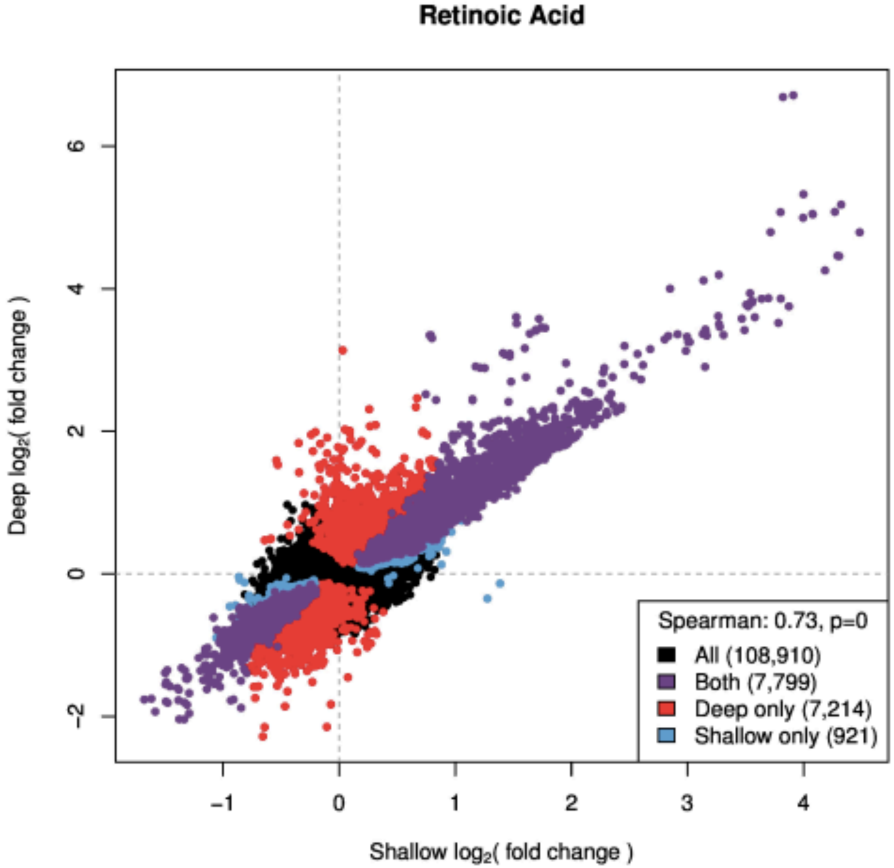
Why shrink fold changes?

Comparison of log fold changes across two experiments.

"A new two-step high-throughput approach:

1. gene expression screening of a large number of conditions

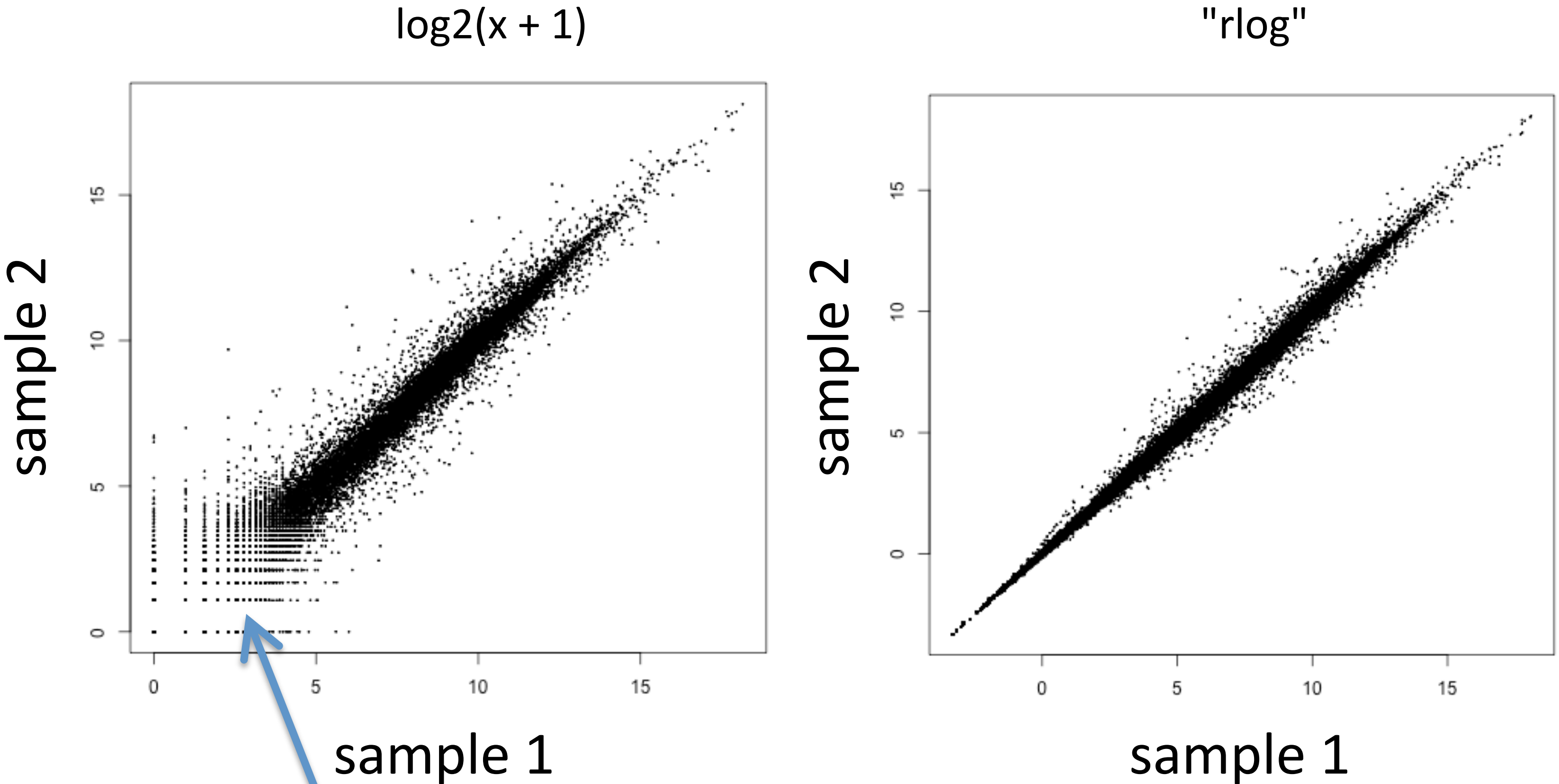
2. deep sequencing of the most relevant conditions"



G. A. Moyerbrailean et al. "A high-throughput RNA-seq approach to profile transcriptional responses" http://dx.doi.org/10.1101/018416

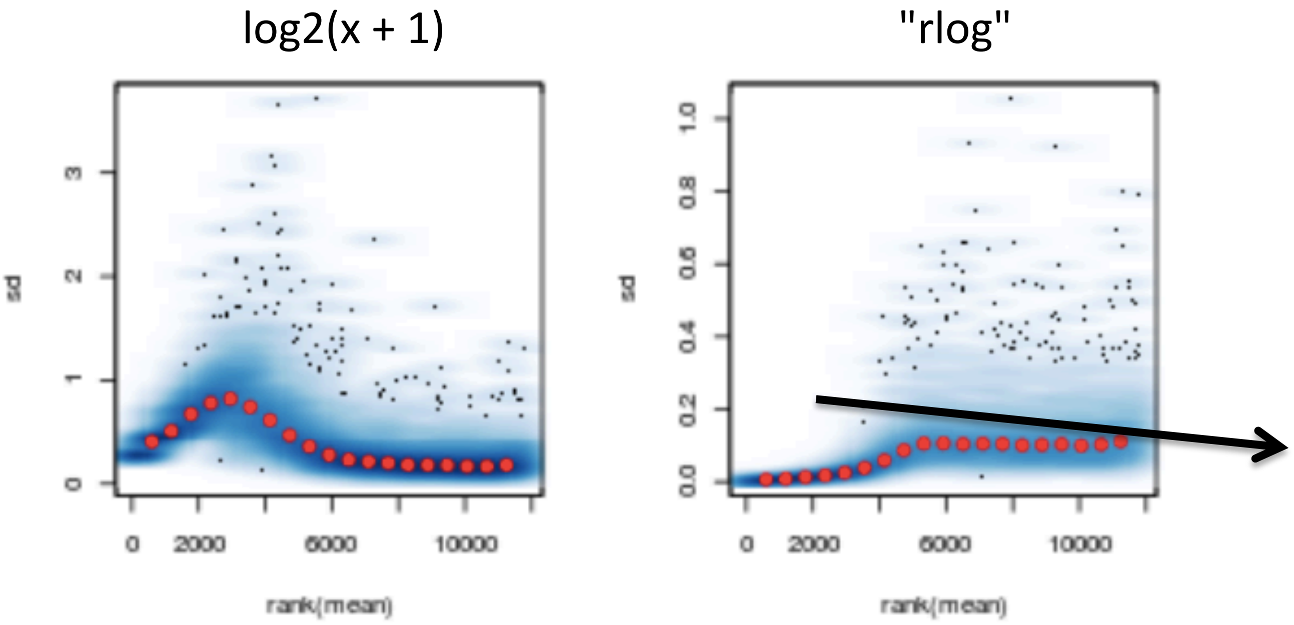
Regularized logarithm, "rlog"

similar idea, but now shrink sample/sample fold changes



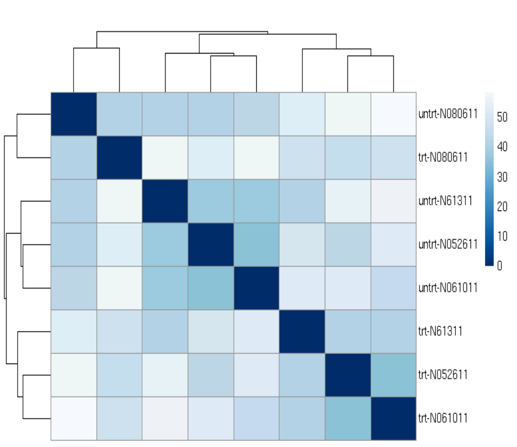
Poisson noise from low counts, when squared
a big contribution to Euclidean distance between samples

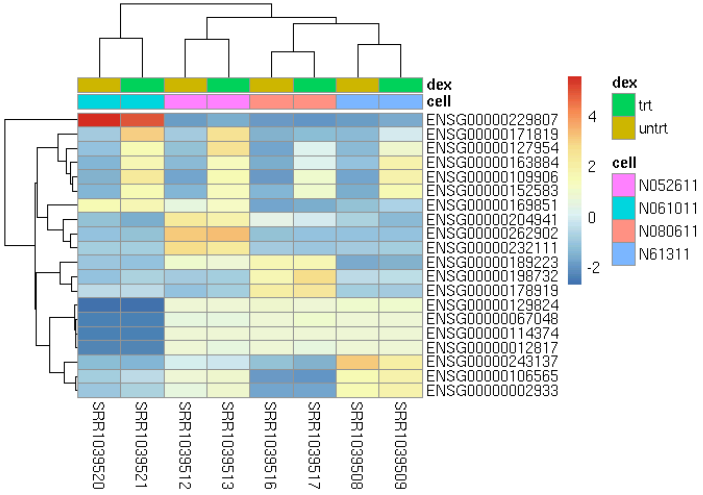
rlog stabilizes variances along the mean



corrects *systematic* dependencies, doesn't force all variances equal.

improving distances, clustering, visualizations





3. Statistical power

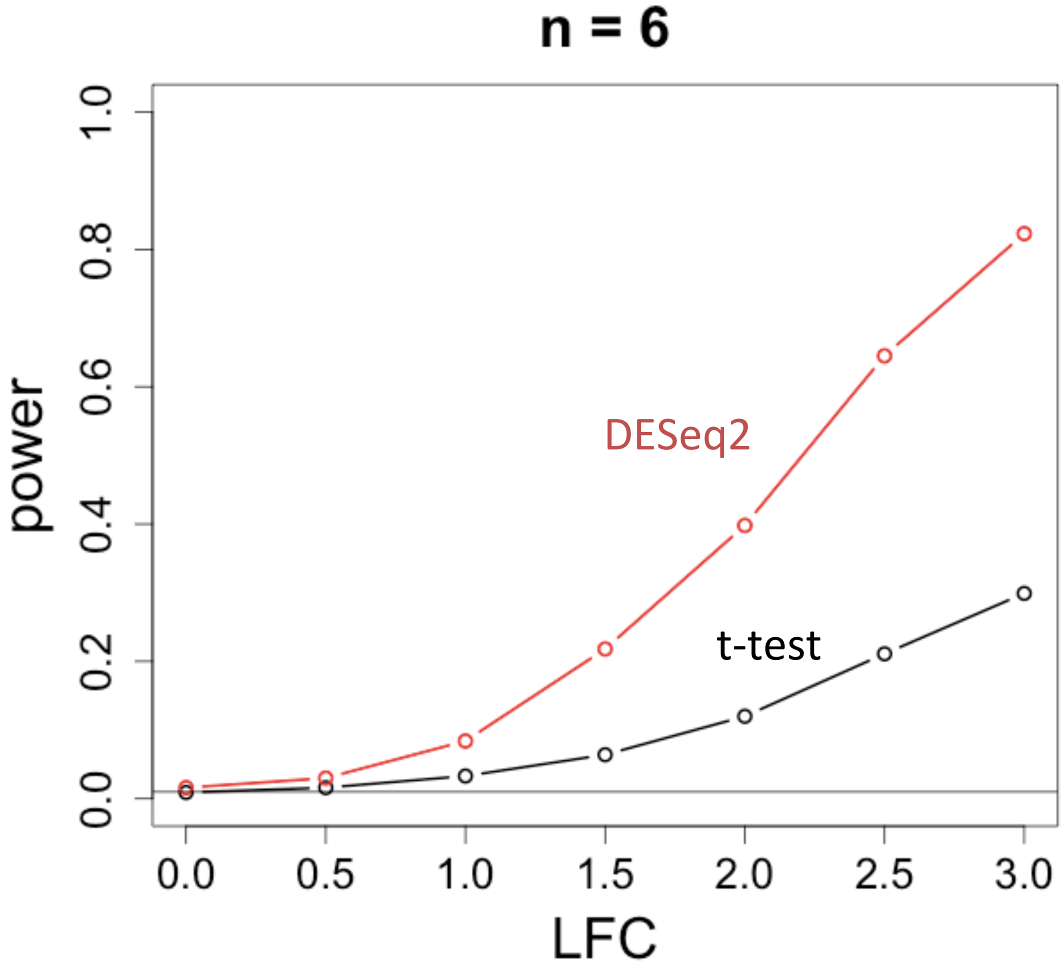
- False positive rate (1 - specificity): under the null (no differences), how many positives?

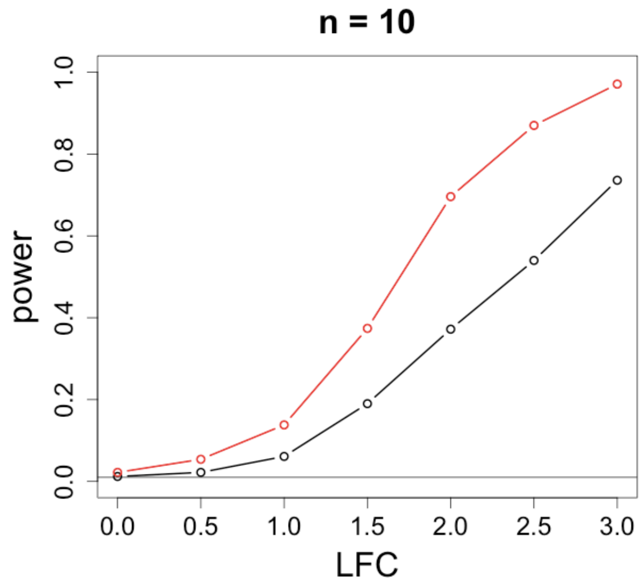
- Precision (1 - false discovery rate): of the positives (predicted to be DE), how many true?

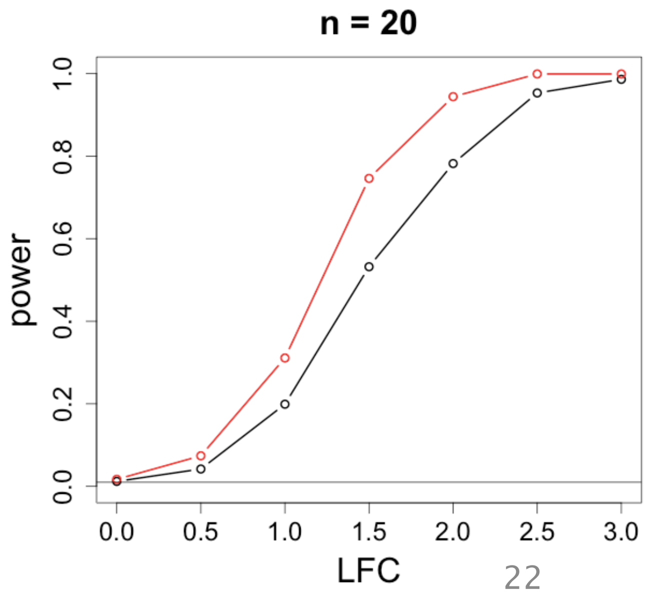
- Power (sensitivity): under the alternative to the null, how many positives (reject null)?

Statistical power

Why not just use a t-test on log normalized counts?



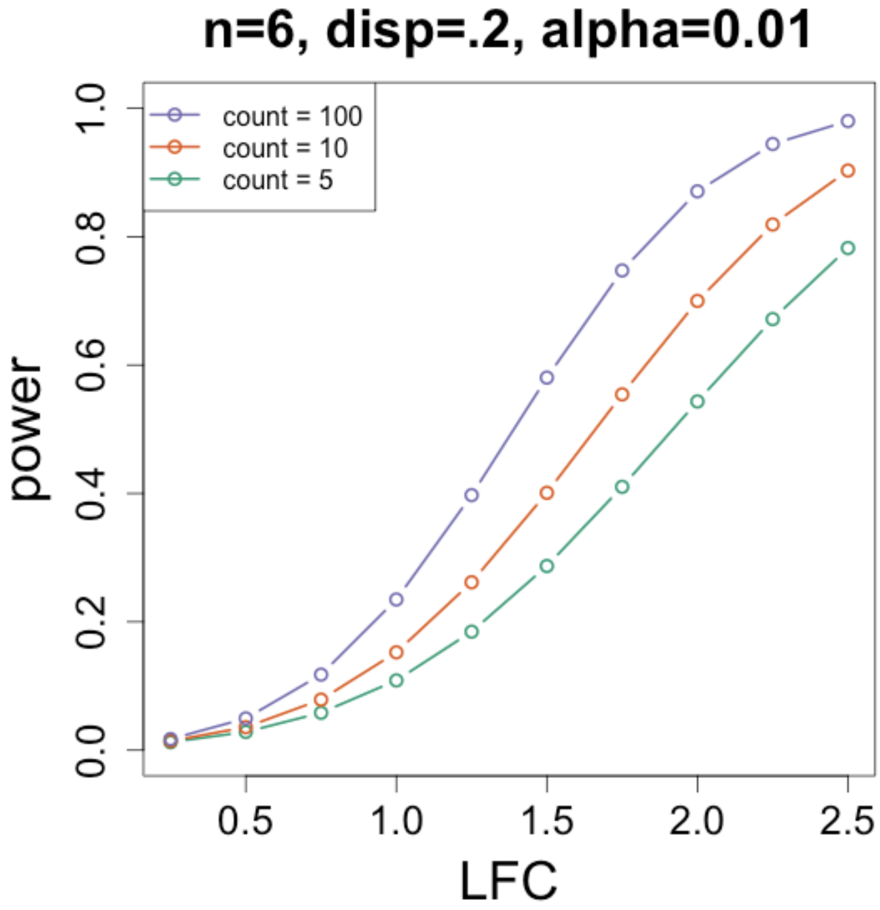




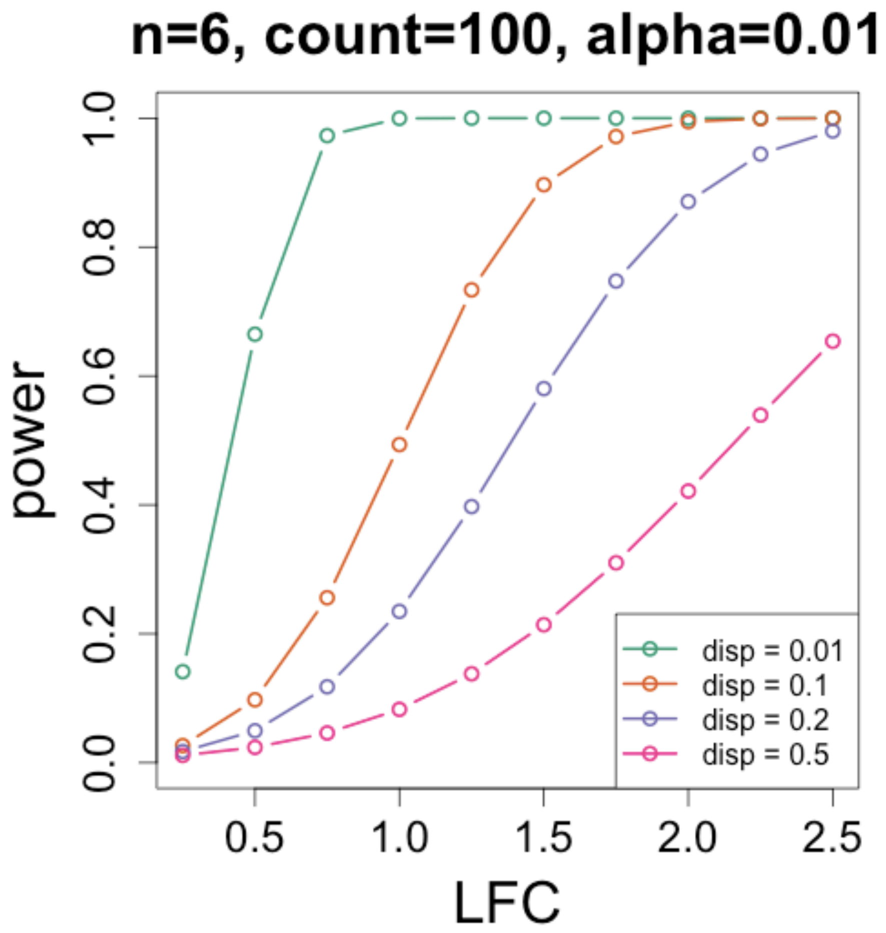
Factors influencing power

- Value of count
  - Sequencing depth
  - Expression
  - Gene length
- Sample size
- Dispersion
- True fold change

Bioc pkg: RNASeqPower

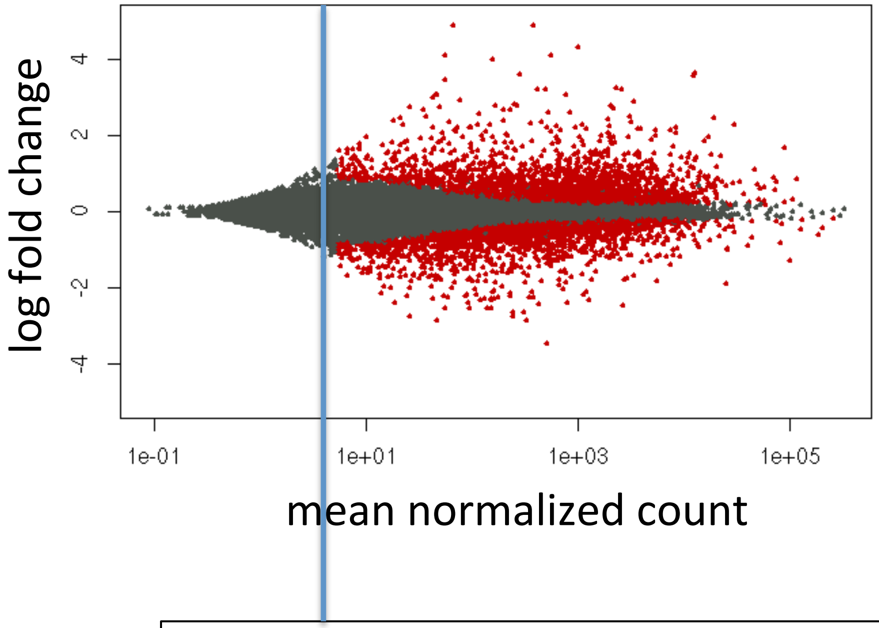


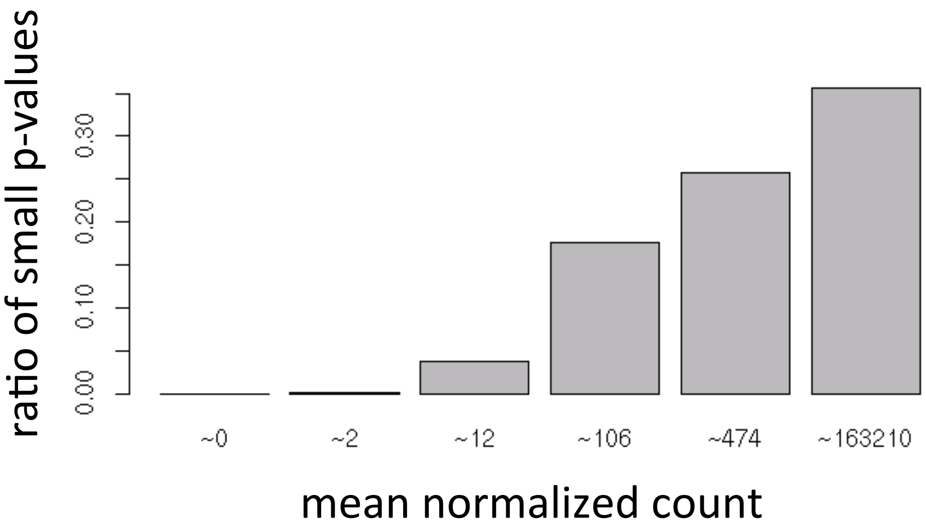
varying the count



varying the dispersion

Power depends on range of counts

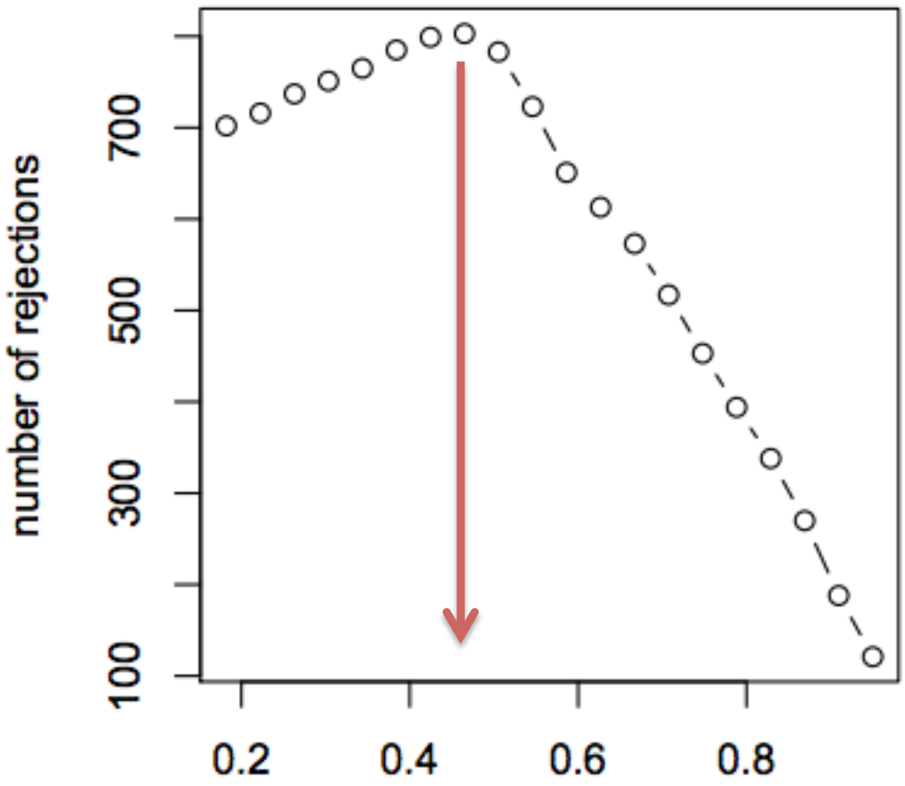




By excluding some tests, e.g. genes with mean normalized count < 5,

we reduce the penalty on adjusted p-values from multiple test correction.

Power depends on range of counts



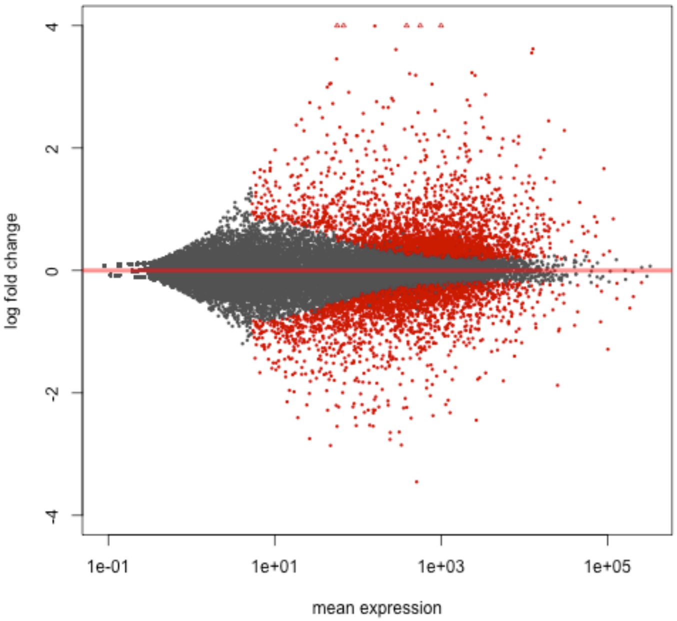
quantile of mean of normalized counts

- Filter on a statistic which is:
  - independent of the test statistic under the null
  - correlated under the alternate hypothesis

Bourgon, Gentleman and Huber, PNAS 2010.

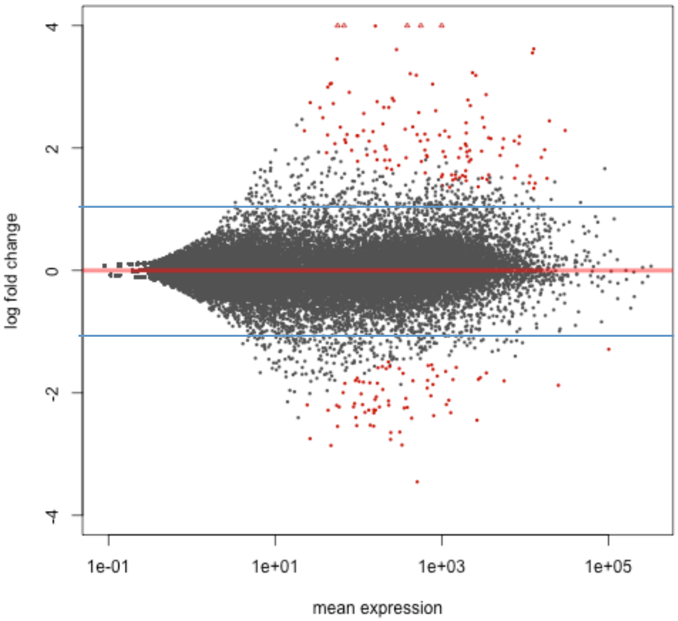
Testing against a threshold

"We get too many DEGs..."



null hypothesis: fold change = 1

using 'lfcThreshold' in results()



null hypothesis: fold change is < 2 or > 1/2

"For **well-powered experiments**, however, a statistical test against the conventional null hypothesis of zero LFC may report genes with statistically significant changes that are so weak in effect strength that they could be **considered irrelevant or distracting**."