RNA-seq data analysis and differential expression

Michael Love
Biostatistics Department
UNC Chapel Hill

Outline

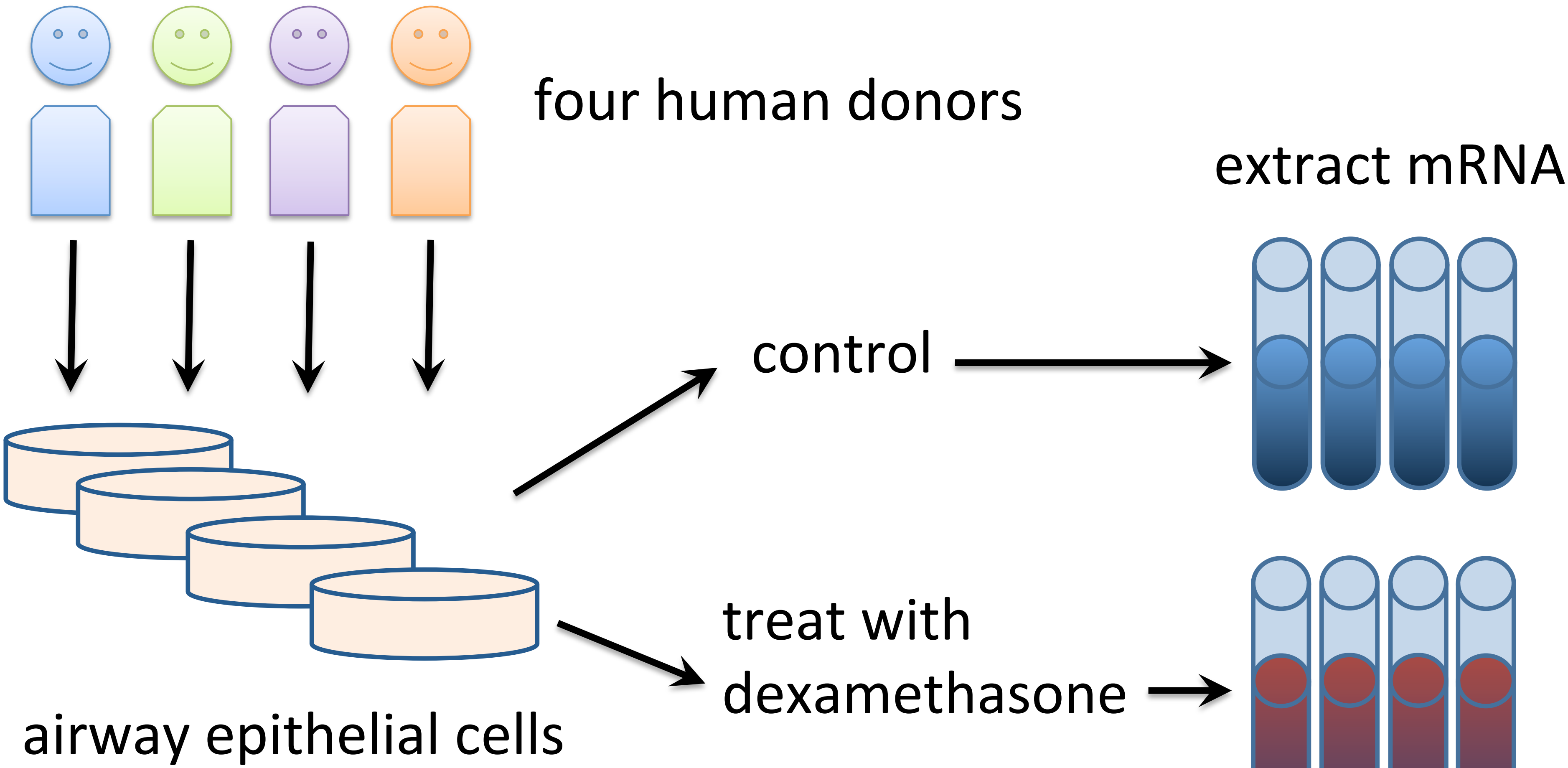
1. Example RNA-seq experiment

2. Statistical analysis of RNA-seq counts

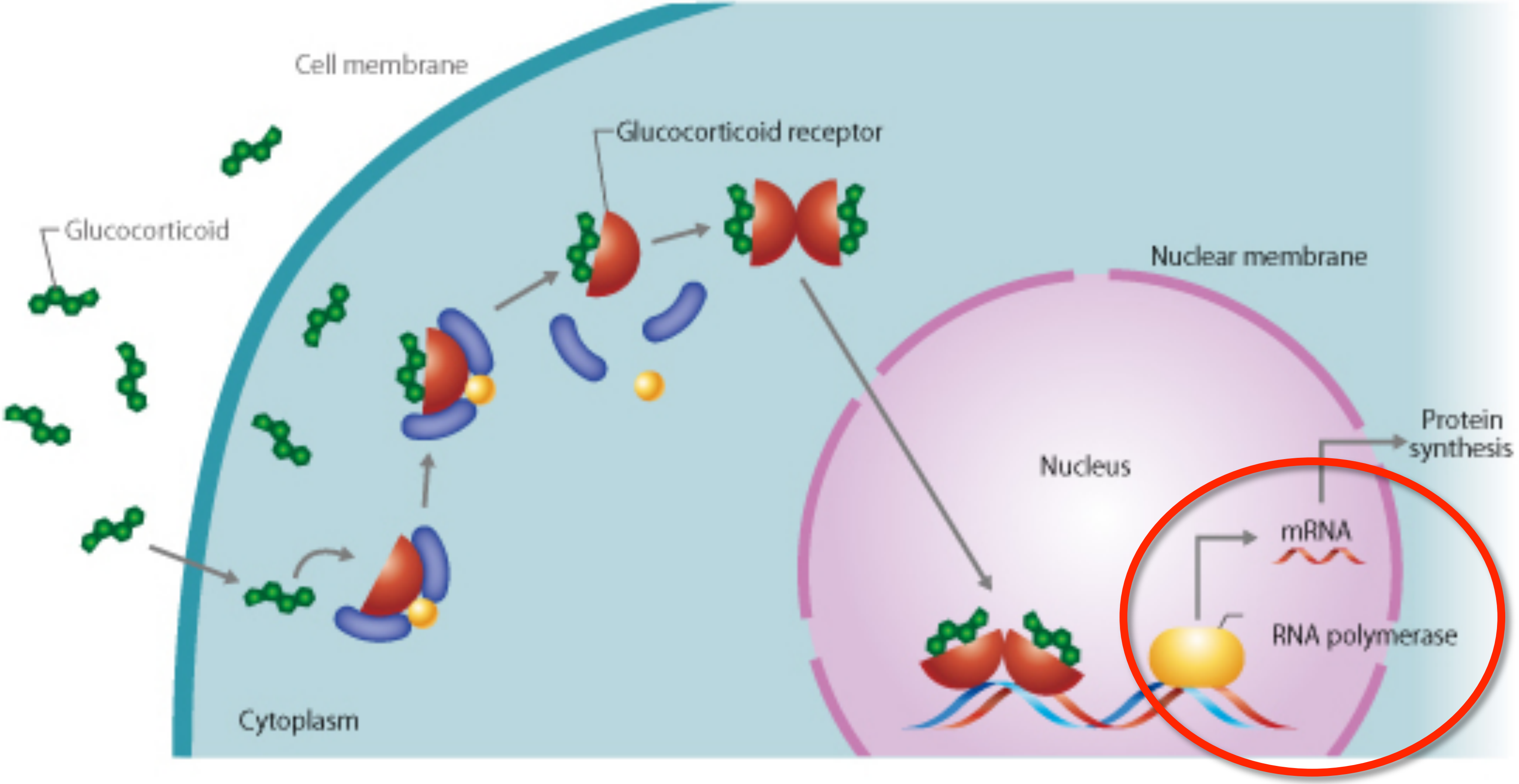
3. Theory of shrinkage estimation

4. Testing steps & statistical power

Our goal: what is airway transcriptome response to glucocorticoid hormone?



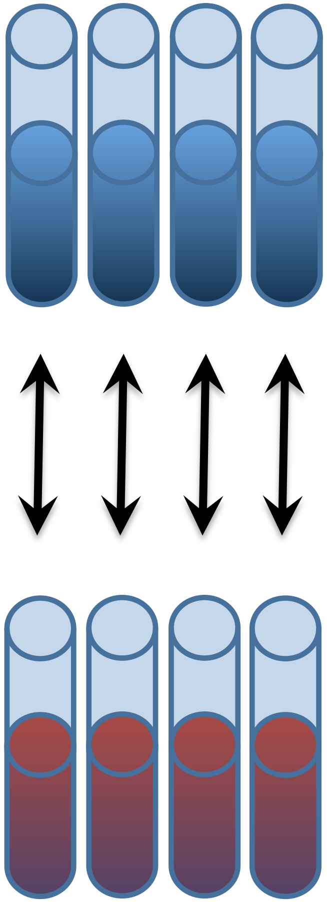
Glucocorticoid mechanism of action



(C) CSLS / University of Tokyo http://csls-text3.c.u-tokyo.ac.jp/

Compare gene expression across treatment, within cell line

cDNA libraries



control

treated with dexamethasone

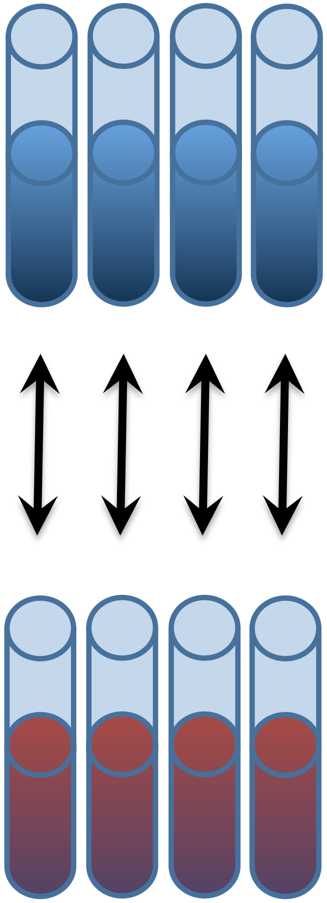
✓ Visualize differences between samples

✓ Test for differences in gene expression, one gene at a time

✓ Visualize differences across all genes

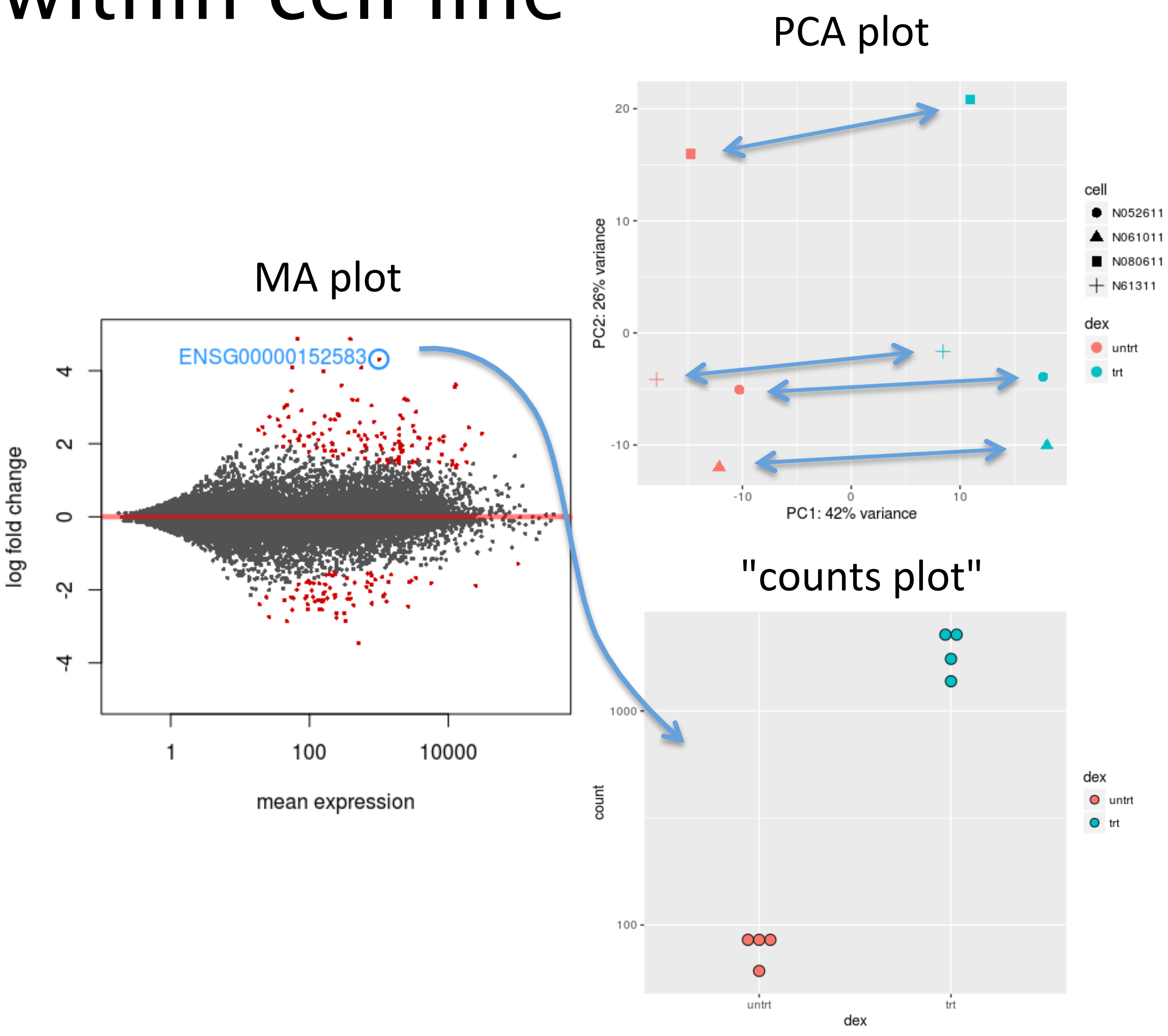
Compare gene expression across treatment, within cell line

cDNA libraries



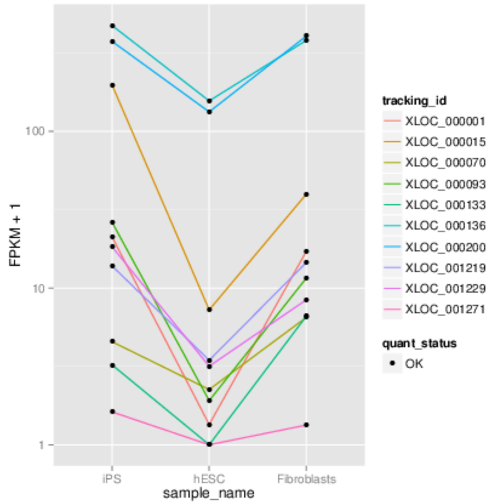
control

treated with dex.



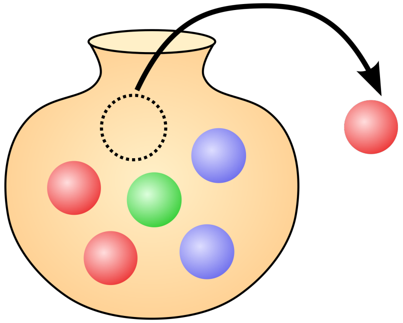
2. FPKM/TPM vs counts

- FPKM: fragments per kilobase per million mapped reads
- TPM: transcripts per million
- FPKM/TPM ∝ gene expression comparable across genes

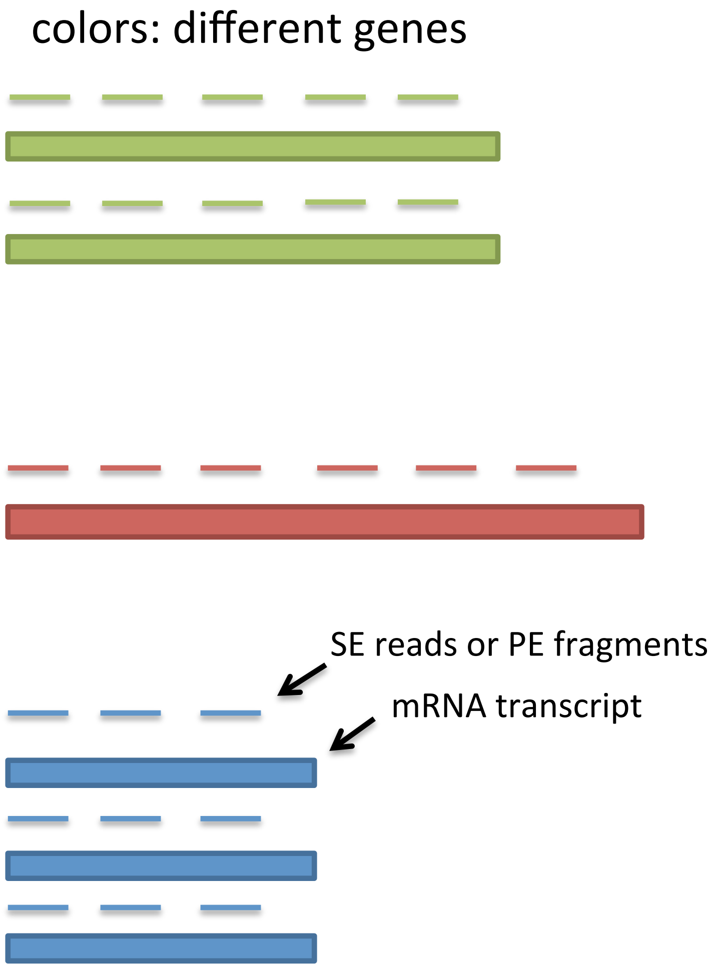


cummeRbund

- Counts have extra information: useful for statistical modeling



mRNAs to RNA-seq fragments

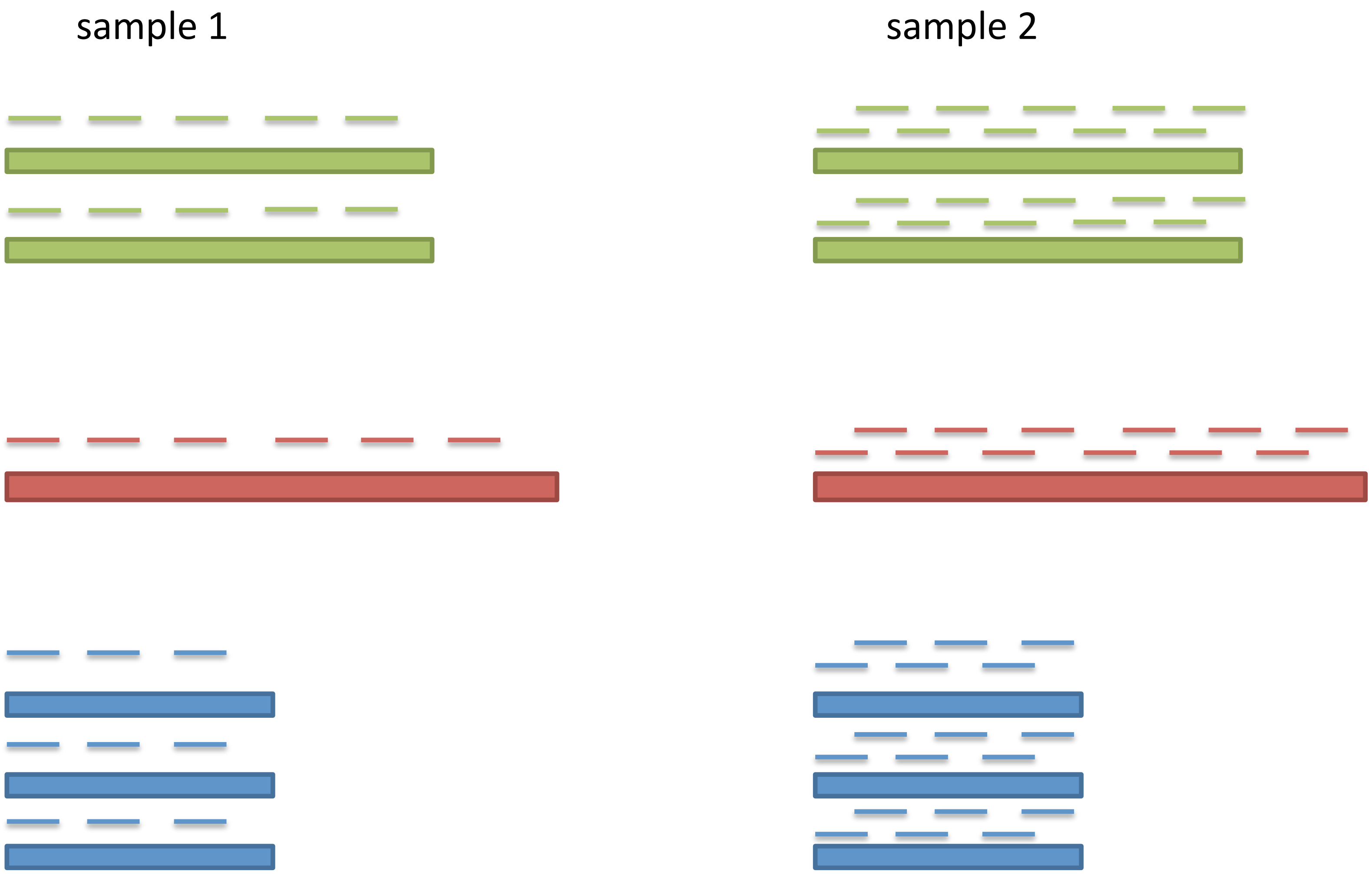


K_{ij} = count of fragments aligned to gene i, sample j

is proportional to:

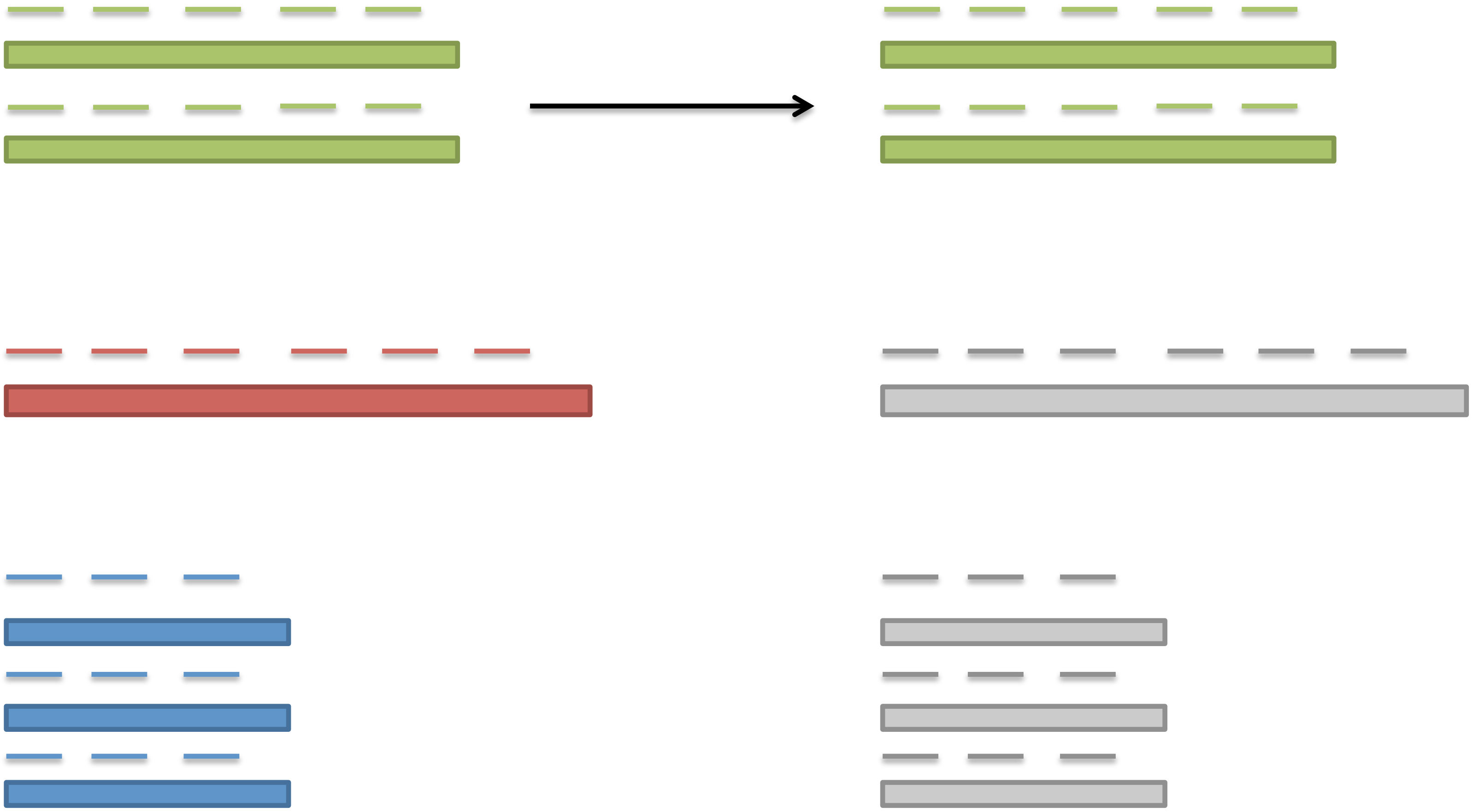
- expression of RNA
- length of gene
- sequencing depth
- lib. prep. factors (PCR)
- in silico factors (alignment)
- ...

Sequencing depth



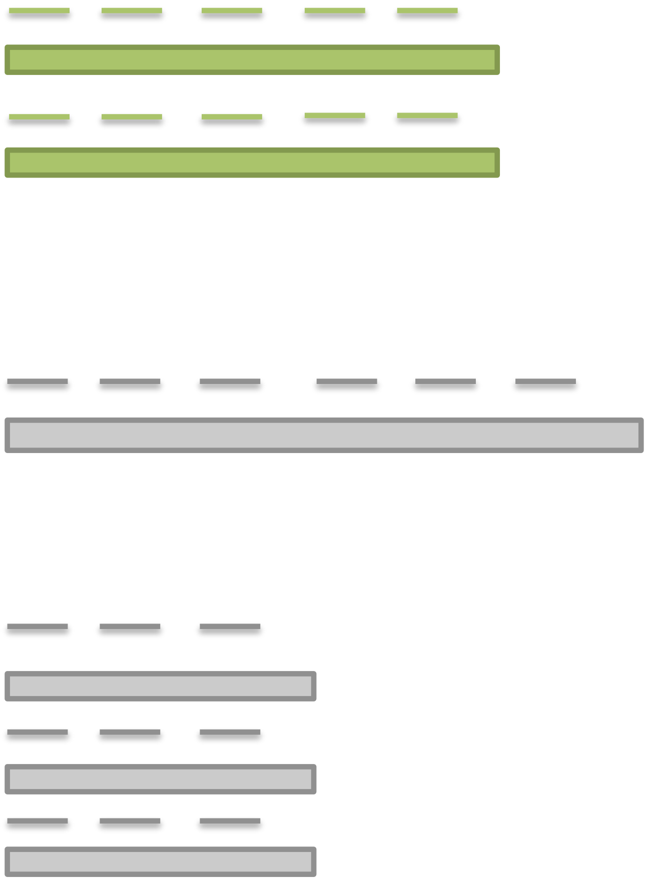
Variance of counts

Consider one gene:

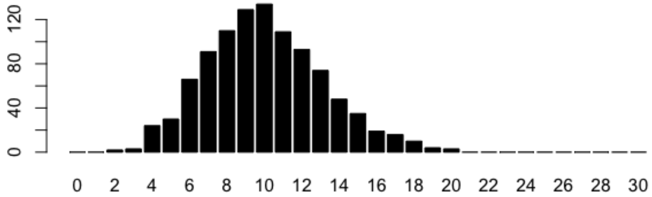


Variance of counts

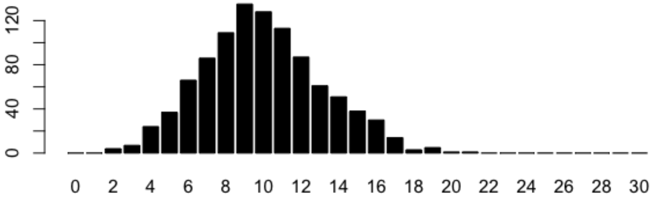
Consider one gene:



- Binomial sampling distribution

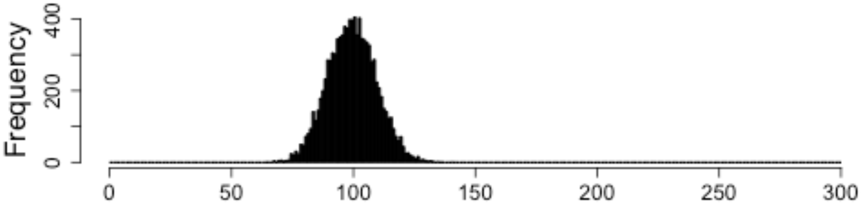


- With millions of reads & small proportion for each gene –> Poisson sampling distribution

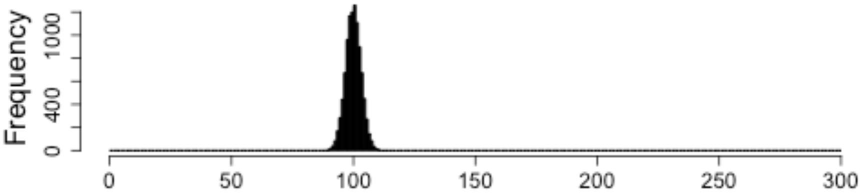


Raw counts vs. normalized counts

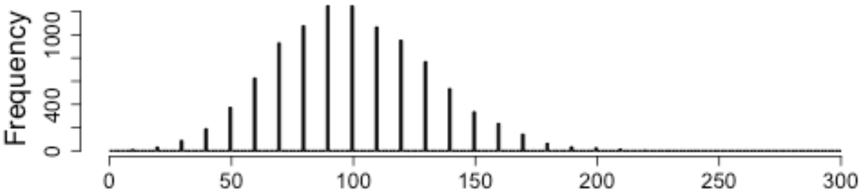
Raw count with mean of 100
Poisson sampling, so SD=10



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

∝ gene expression

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

statistical inference "for free"
edgeR, DESeq2

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

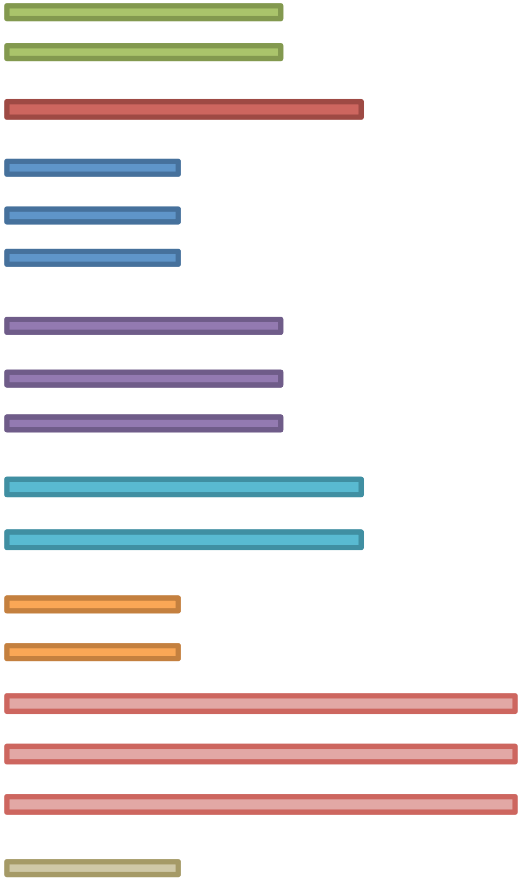
can be made to work with extra modeling
e.g. limma-voom

some distribution

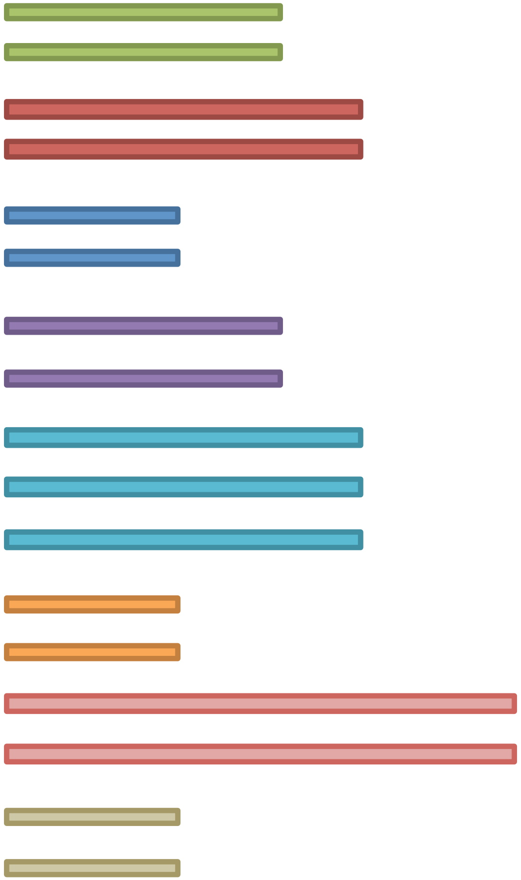
mean parameter

Biological replicates

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

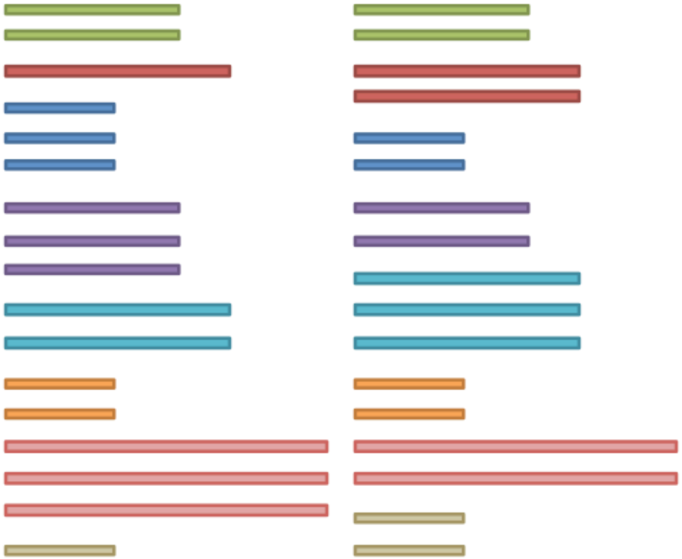


But realistically, biological variation across sample units is expected

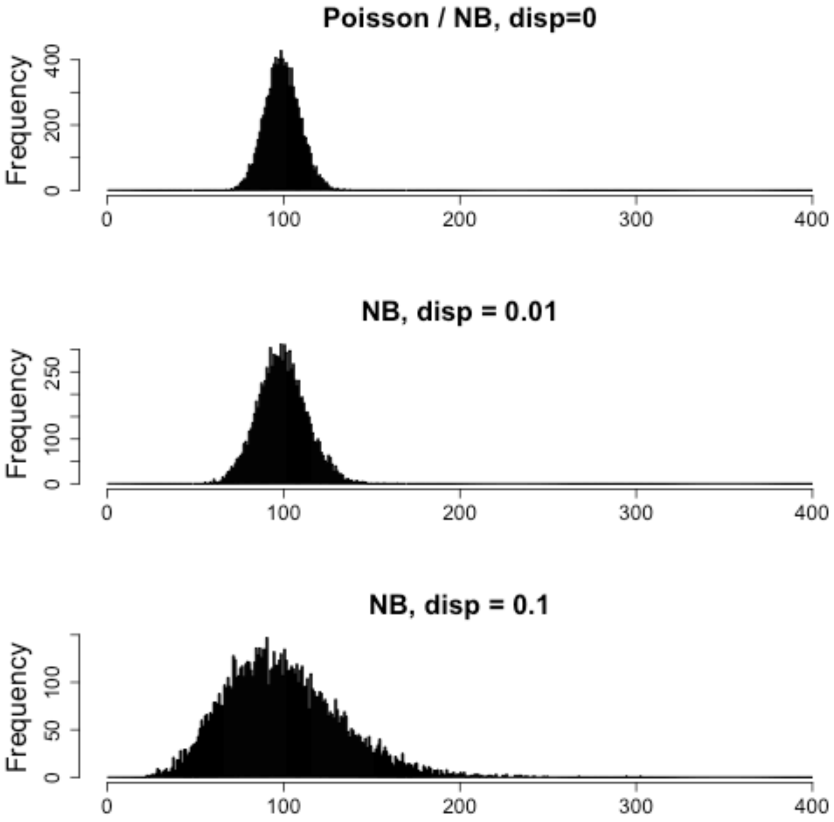


Biological replicates

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



Negative Binomial =
Poisson with a varying mean


Poisson / NB, disp=0

NB, disp = 0.01

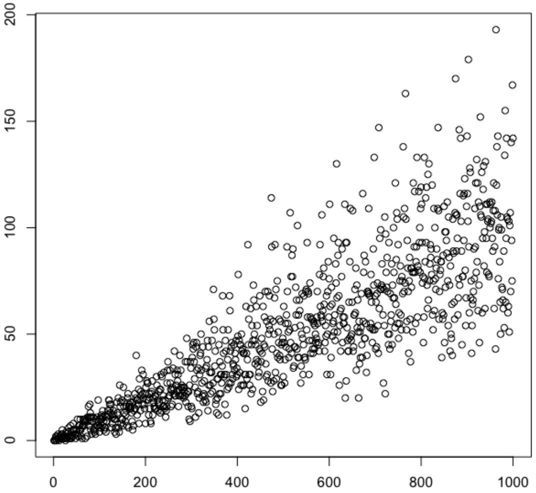
NB, disp = 0.1

Dispersion parameter

$$\mathrm{Var}(K_{ij}) = \mu_{ij} + \alpha_i \mu_{ij}^2$$

Poisson part:
sampling fragments

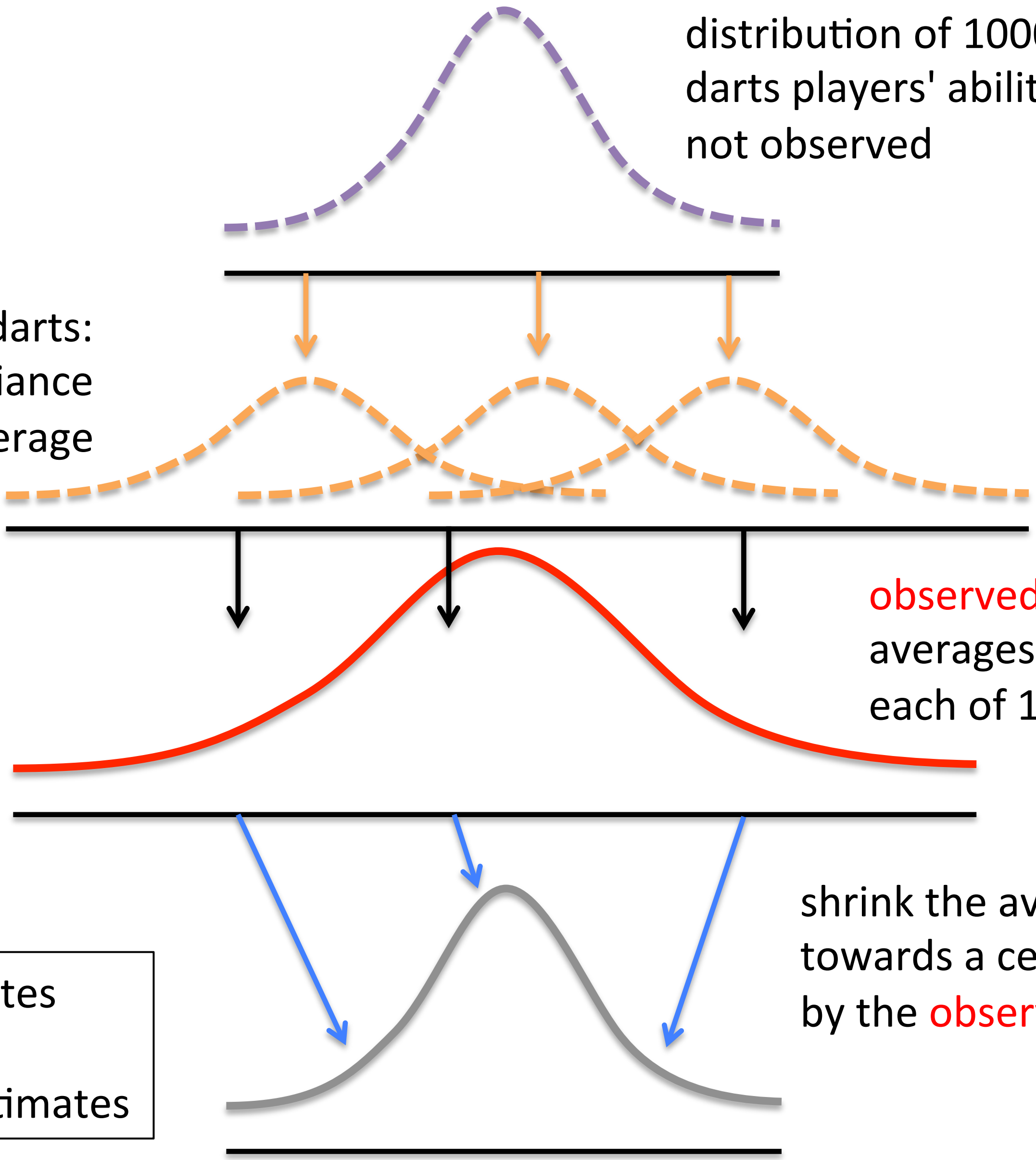
Extra variation
due to biological variance



for large counts: $\sqrt{\alpha_i} \approx \dfrac{\sigma}{\mu} \equiv CV$ (coefficient of variation)

disp = 0.01 -> CV 10%
disp = 0.25 -> CV 50%

3. Shrinkage estimation



distribution of 1000
darts players' ability:
not observed

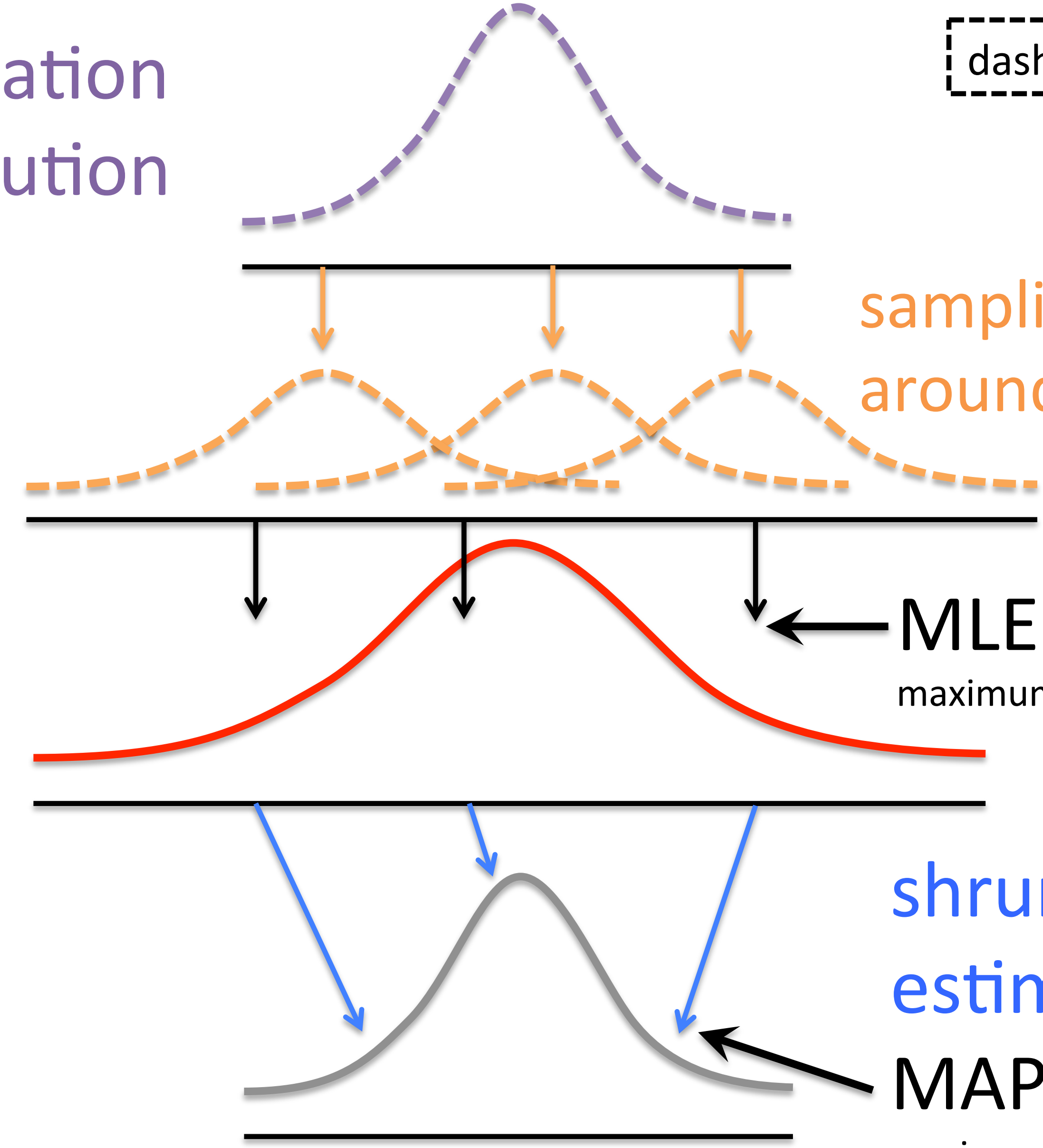
each throws 3 darts:
sample variance
of the average

observed distribution:
averages of 3 throws from
each of 1000 players

shrink the averages
towards a center defined
by the observed distribution

"shrunken" estimates
less error *overall*
than **individual** estimates

Shrinkage estimation



dashed = unobserved

population distribution

sampling variance around true ability

empirical distribution

the center defines the prior mean

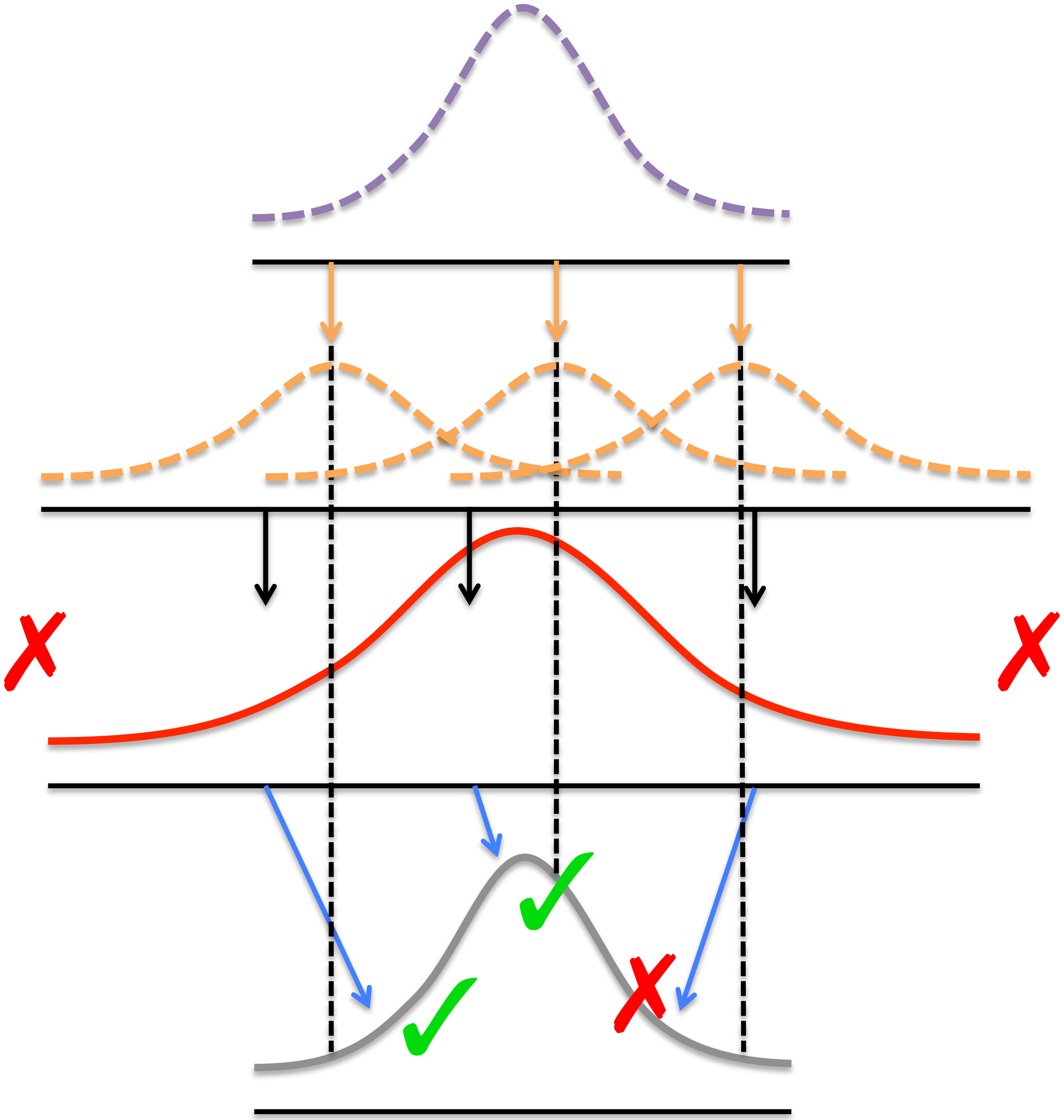
MLE

maximum likelihood estimates

shrunken estimates or MAP

maximum a posteriori

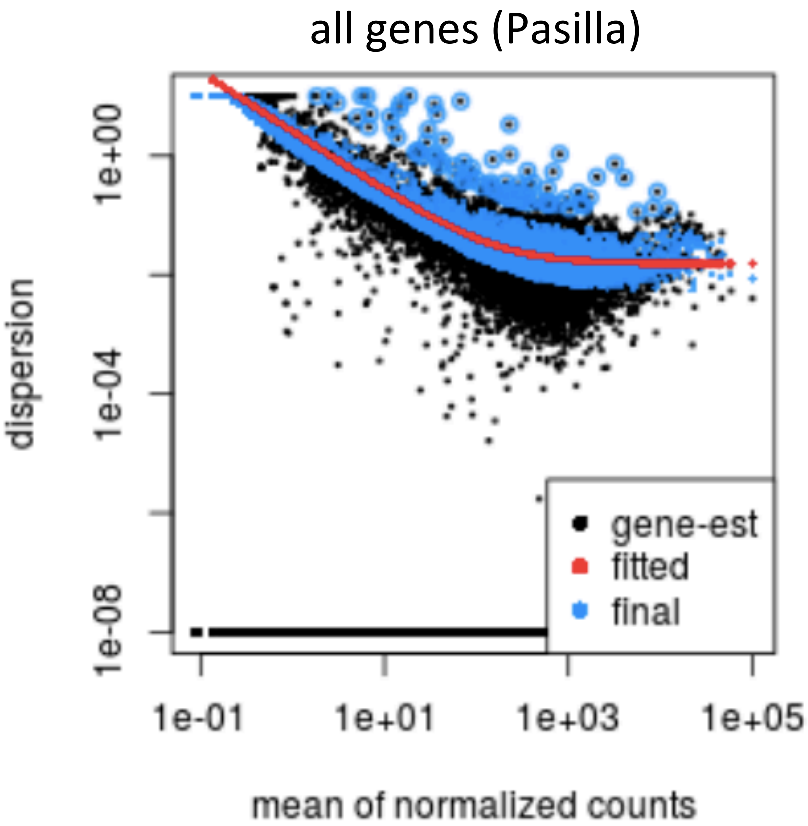
Shrinkage estimation

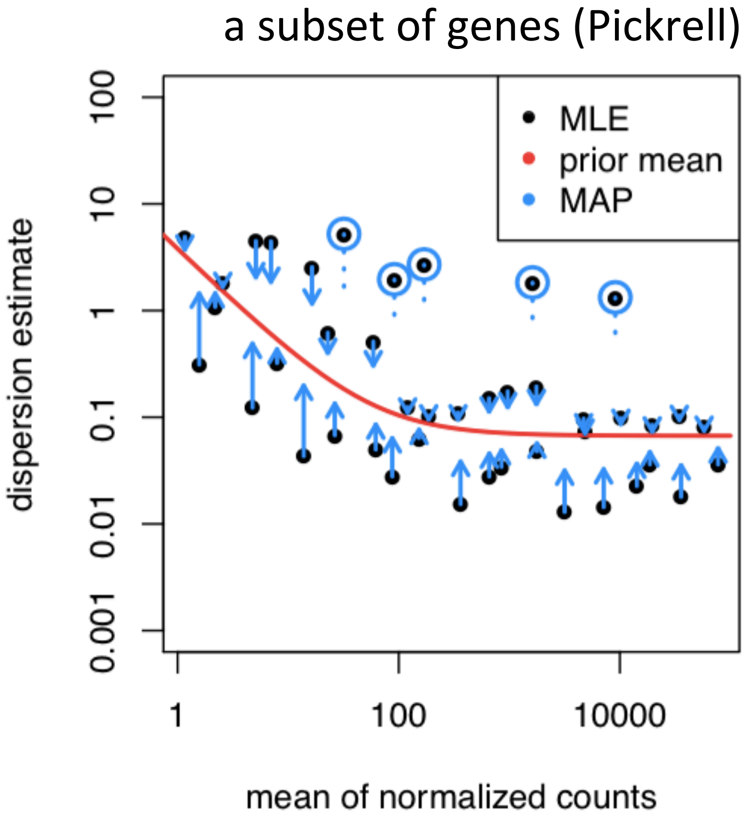


Shrinkage estimators in genomics

- Lönnstedt and Speed 2002: microarray
- Smyth 2004: limma for microarray
- Robinson and Smyth 2007:
  edgeR for SAGE and then applied to RNA-seq
- Many adaptations: DSS and DESeq2 are a similar approach, data-driven strength of shrinkage

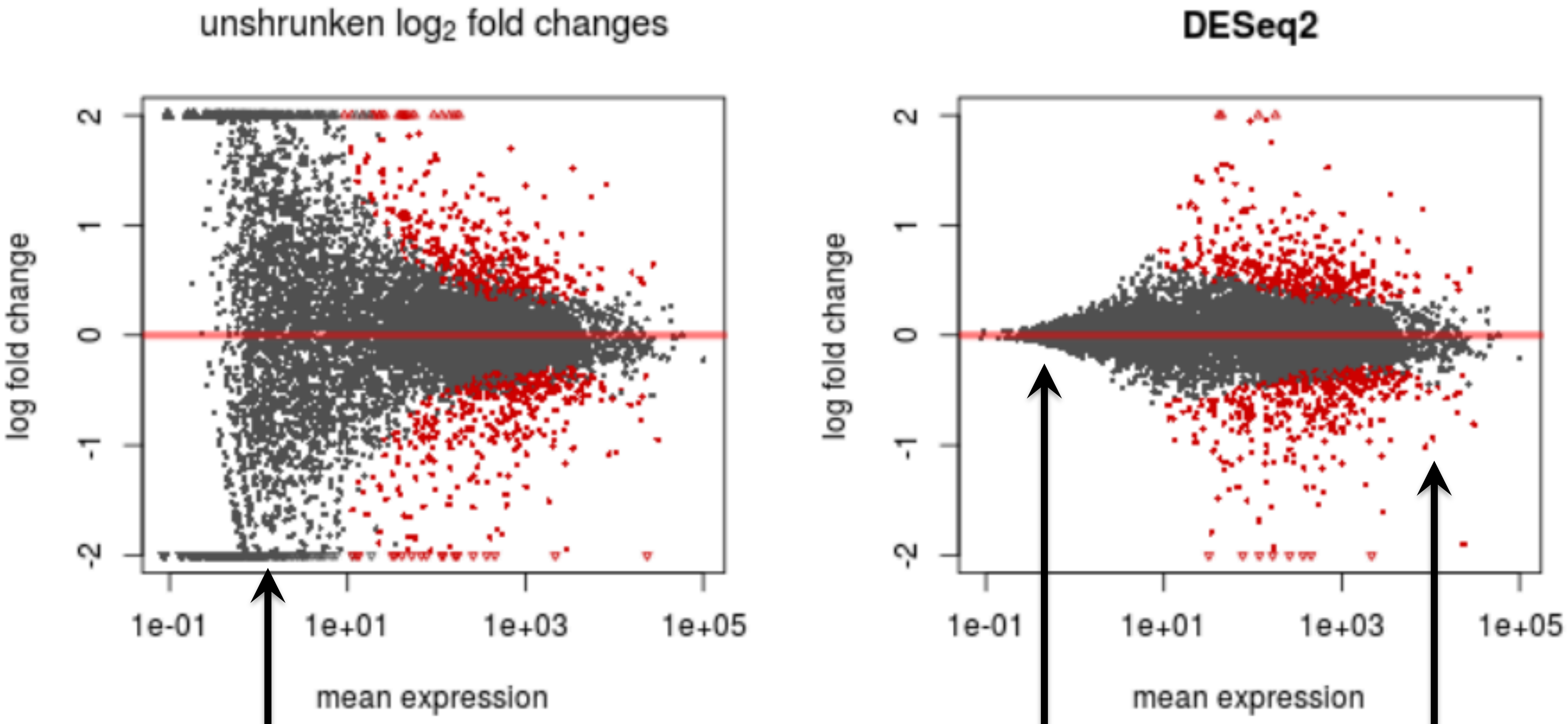
Shrinkage of dispersion for RNA-seq





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 for RNA-seq

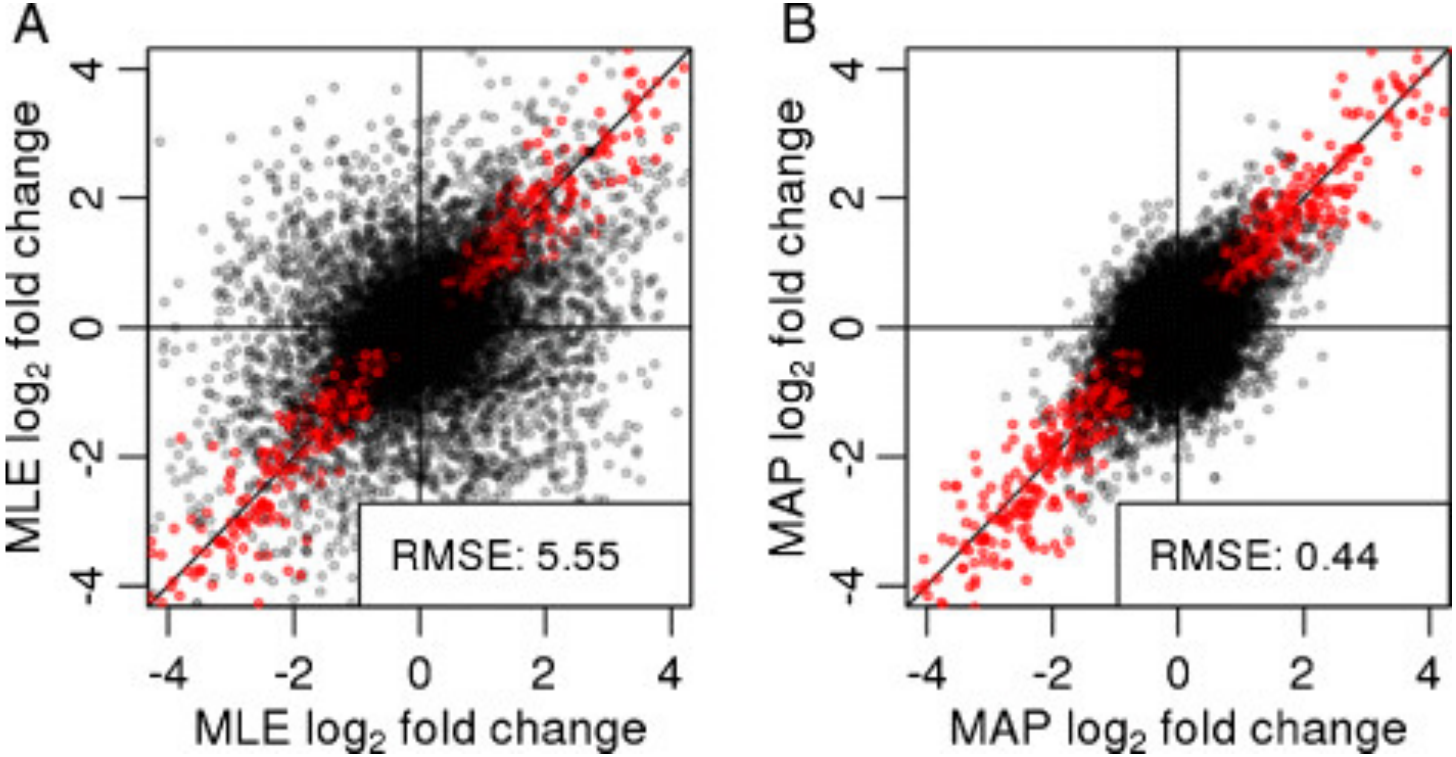


noisy estimates due to low counts
large FDR from the statistical model,
but we shouldn't trust the estimate itself

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

almost no
shrinkage

Why shrink fold changes?



Split a dataset into two equal parts, compare LFC

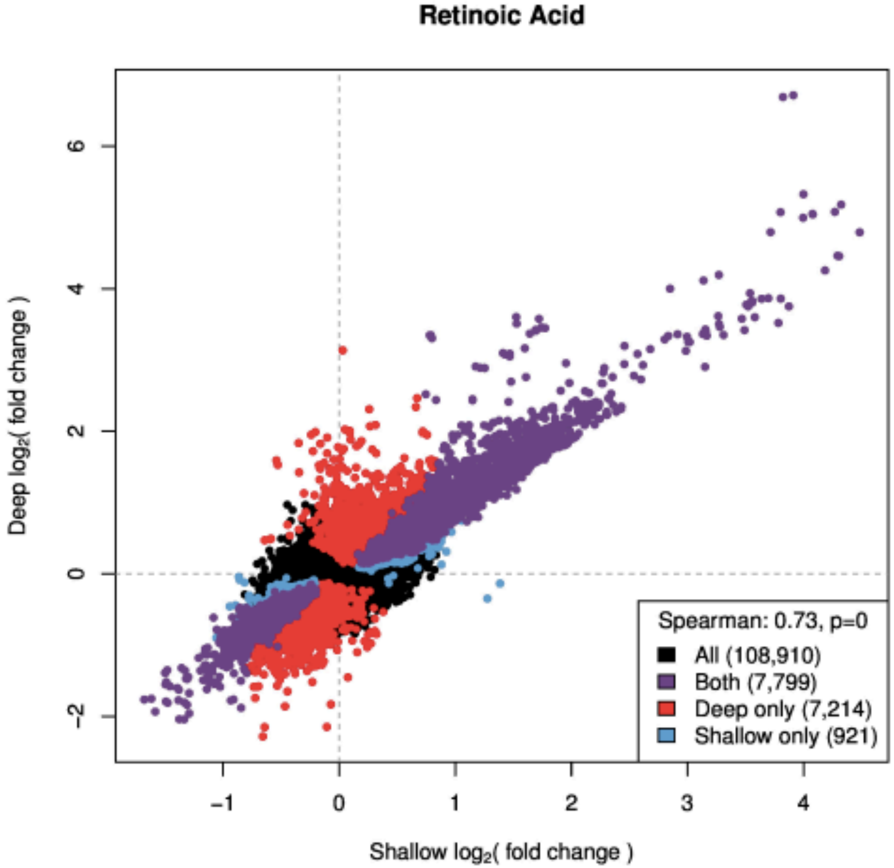
Why shrink fold changes?

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

Two paths in RNA-seq analysis

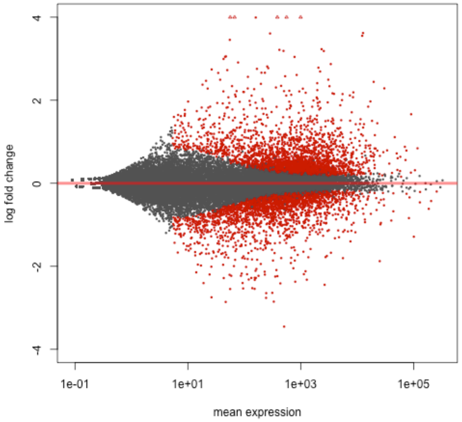
Count matrix

Differential expression

testing, p-values, FDR

`DESeq()`
`results()` } DESeq2

`glmLRT()`
`topTags()` } edgeR

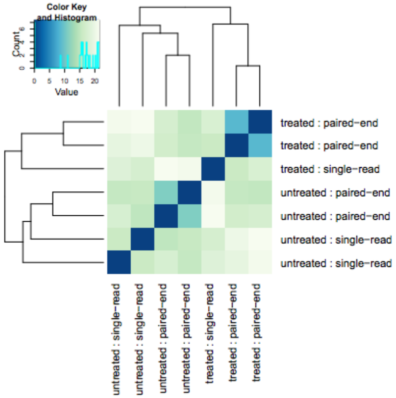


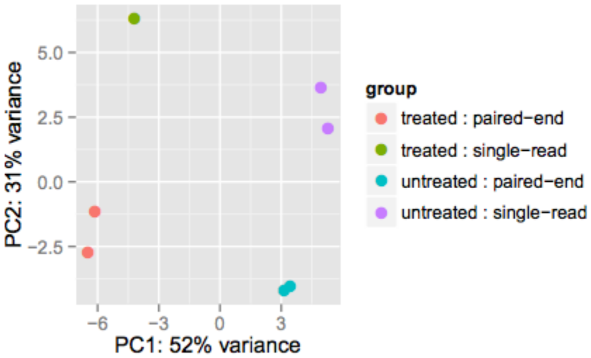
Transformations and Exploratory Data Analysis (EDA)

clustering, heatmaps,
sample-sample distances

DESeq2 { `vst(), rlog(), plotPCA()`

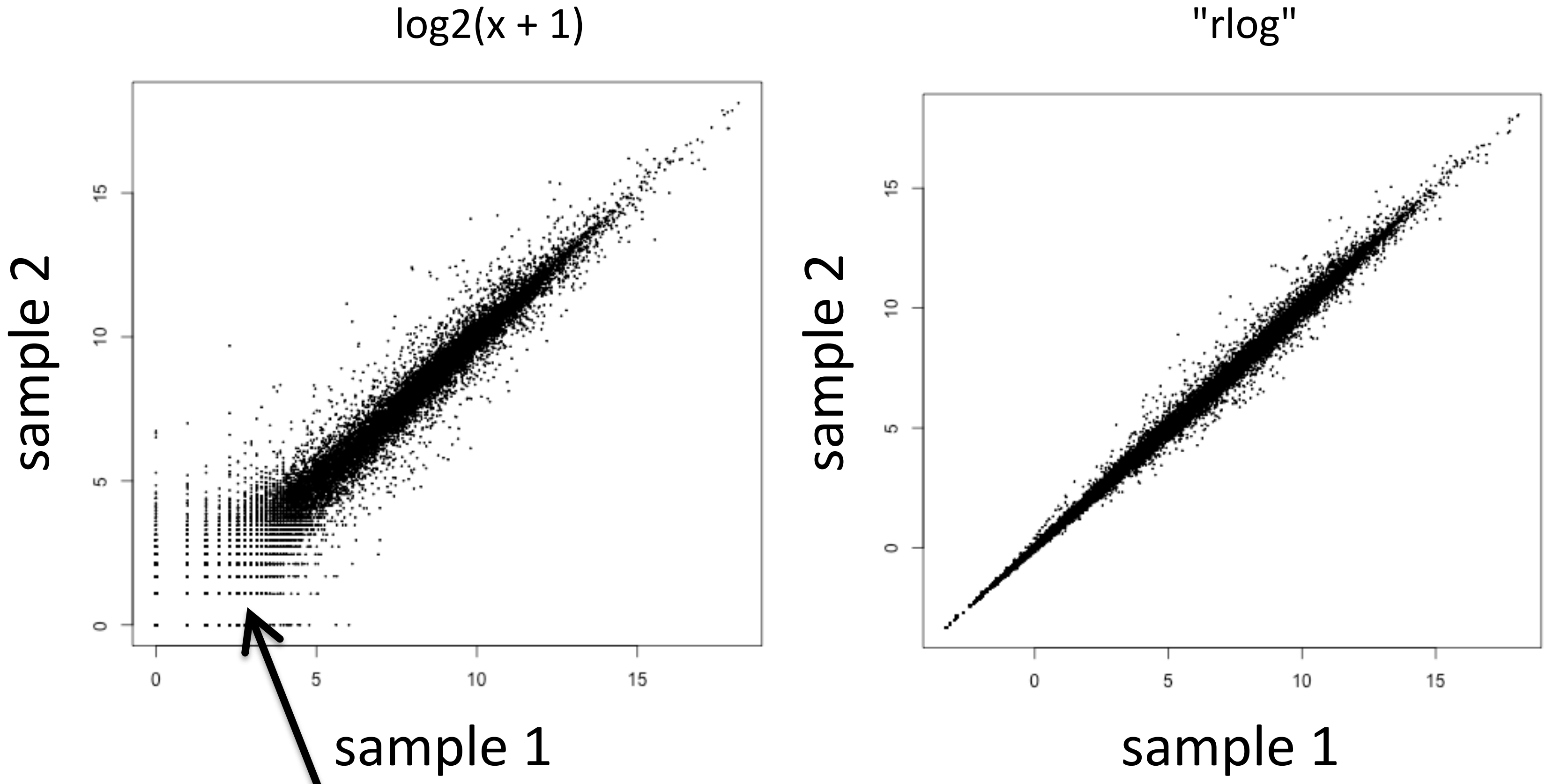
edgeR { `cpm(), plotMDS()`





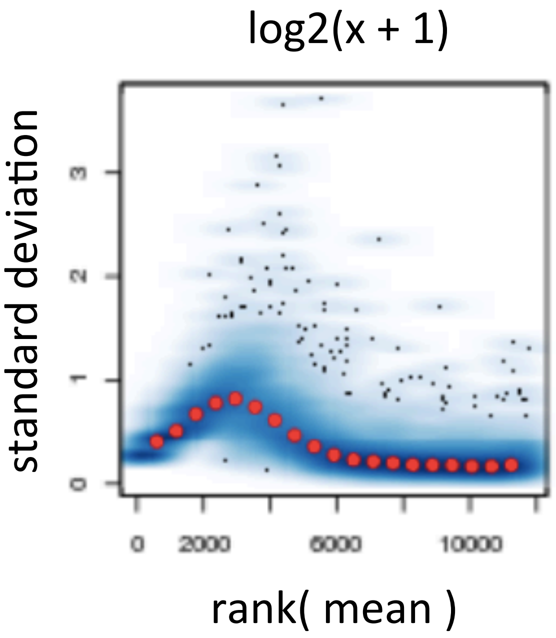
Regularized logarithm, "rlog"

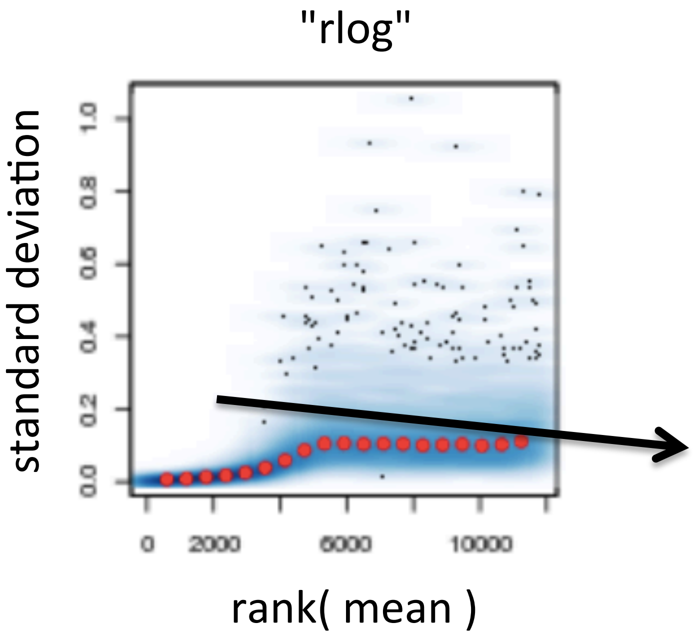
similar idea as fold change shrinkage,
now sample-to-sample fold changes



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

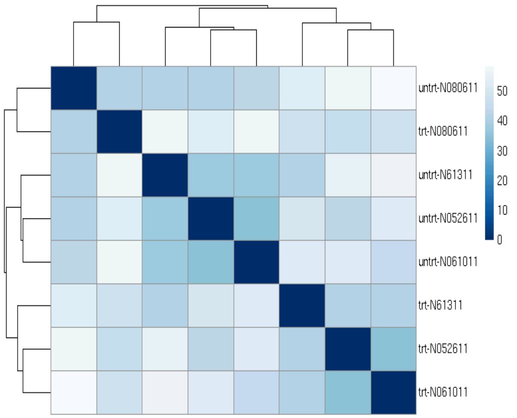
rlog stabilizes variance along the mean

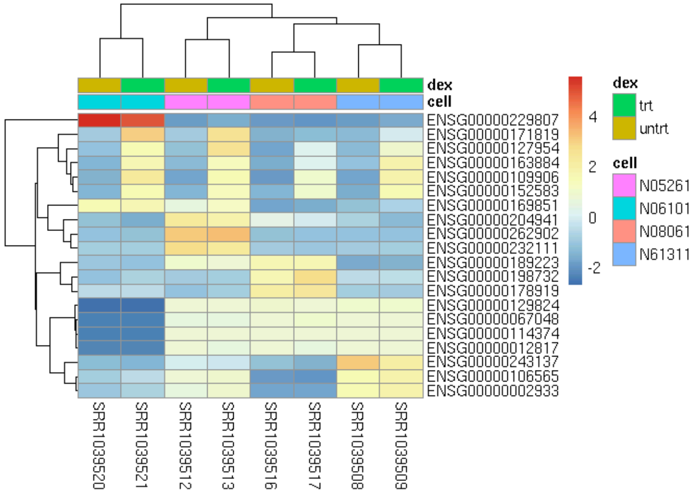




corrects *systematic* dependencies, doesn't force all variances equal.

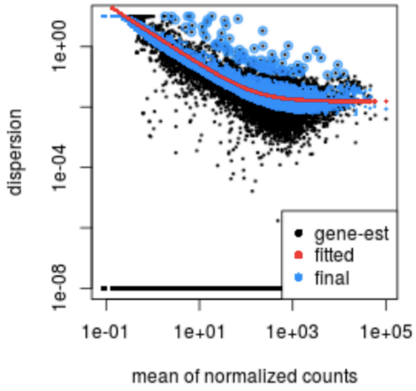
improving distances, clustering, visualizations





Also in DESeq2: VST

- *Variance stabilizing transformation*: calculate the dependence of variance on the mean (using the dispersion trend)



- Closed-form expression f(x) for stabilizing

- `vst()` is a *faster* implementation

4. Testing steps

count matrix (from featureCounts,
summarizeOverlaps,
htseq, tximport, etc.)

1. size factors (sequencing depth)
2. dispersion (additional variance)
3. *Wald* test or *likelihood ratio* test
4. build a results table

Statistical power

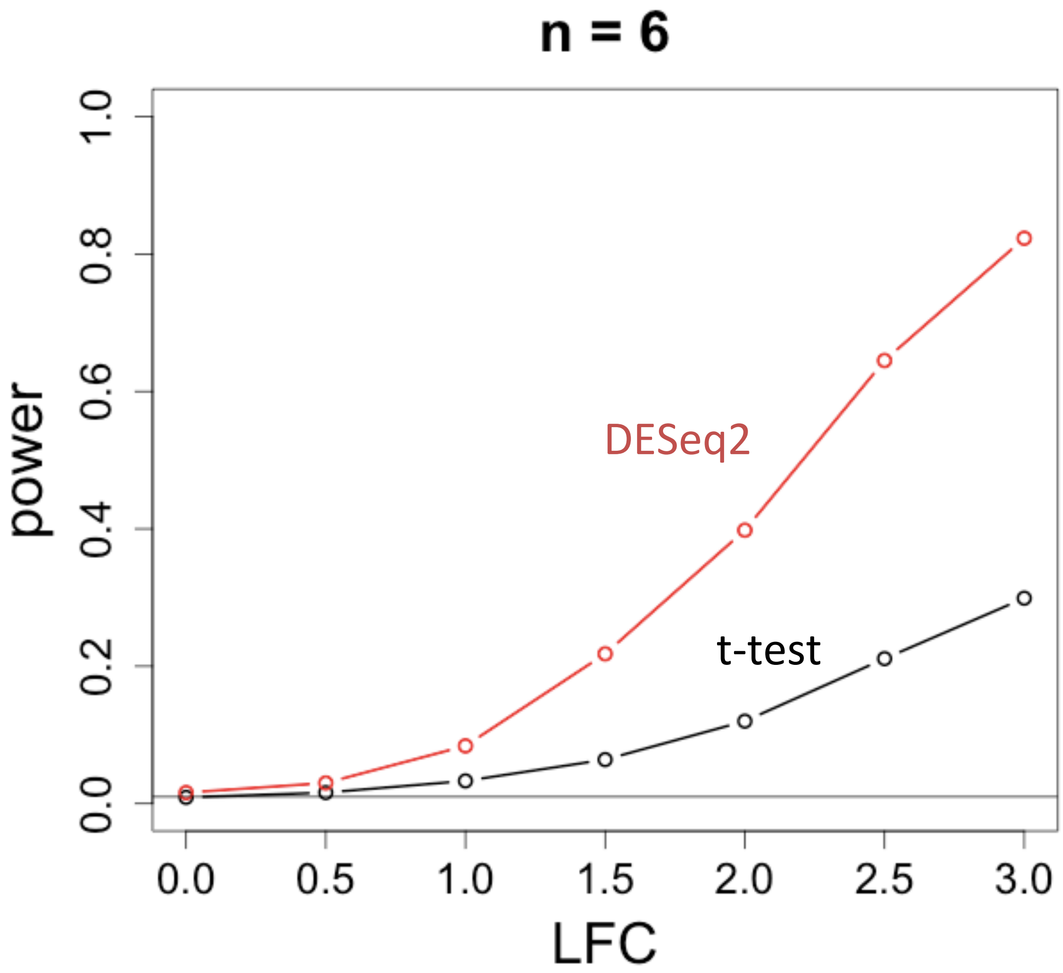
- False positive rate (1 - specificity):
  under the null (no differences),
  how many called positives?

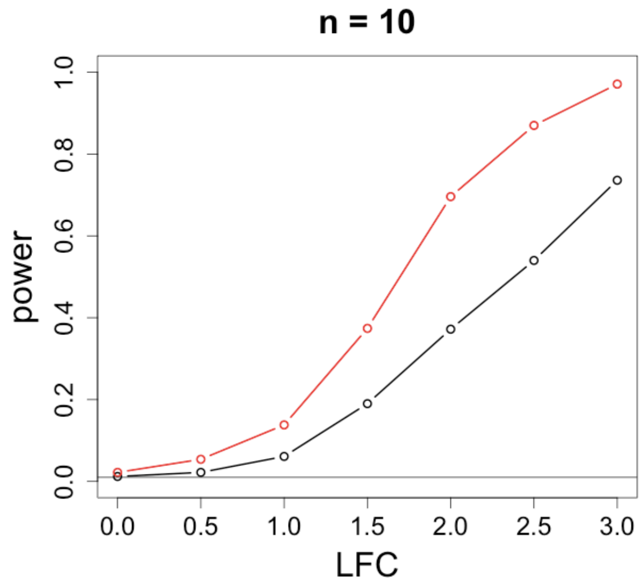
- Precision (1 - false discovery rate):
  of the positives (called DE),
  how many are true positives?

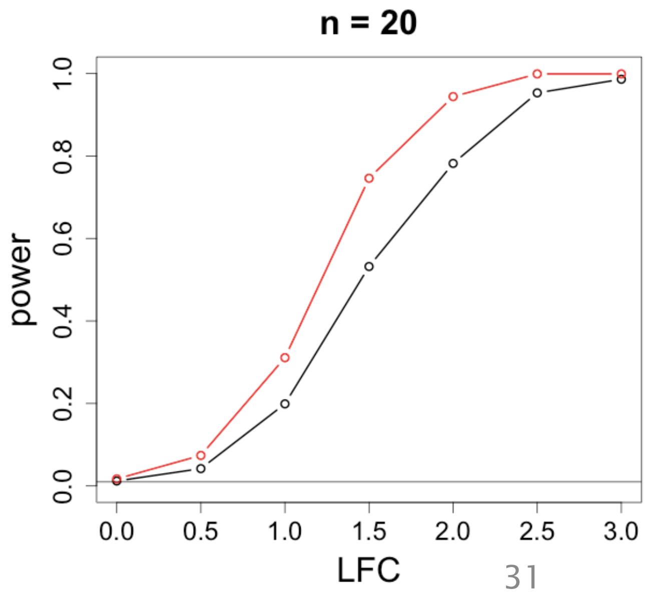
- Power (sensitivity):
  under the alternative to the null,
  how many called positive?

Statistical power

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



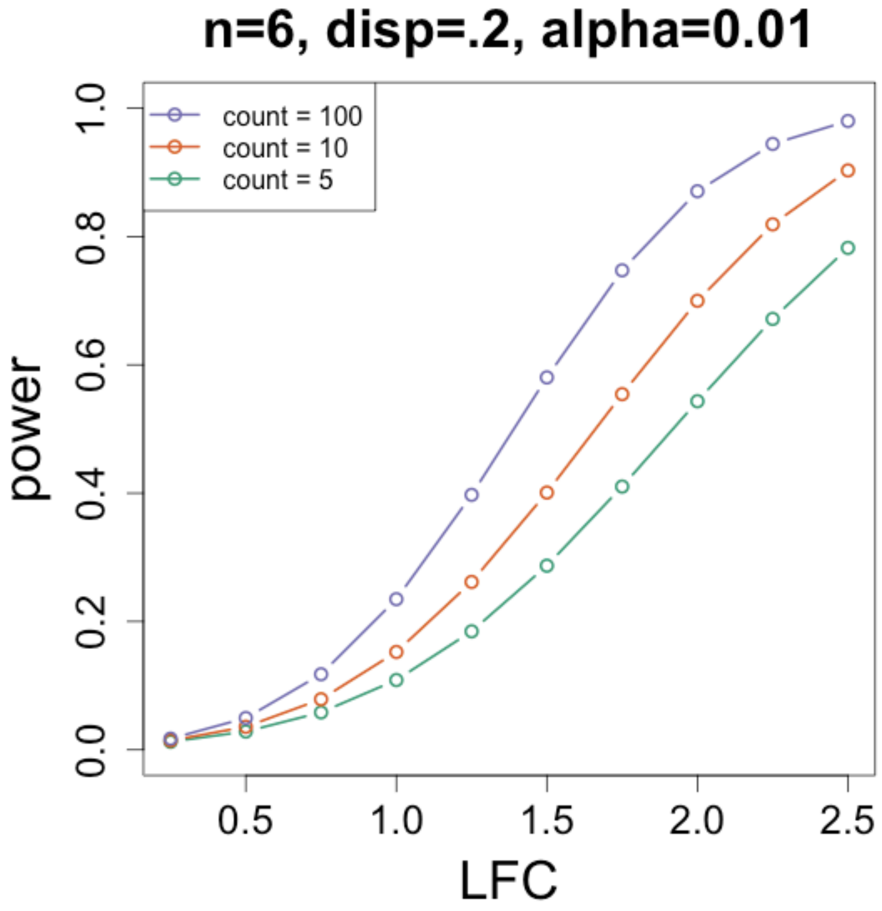




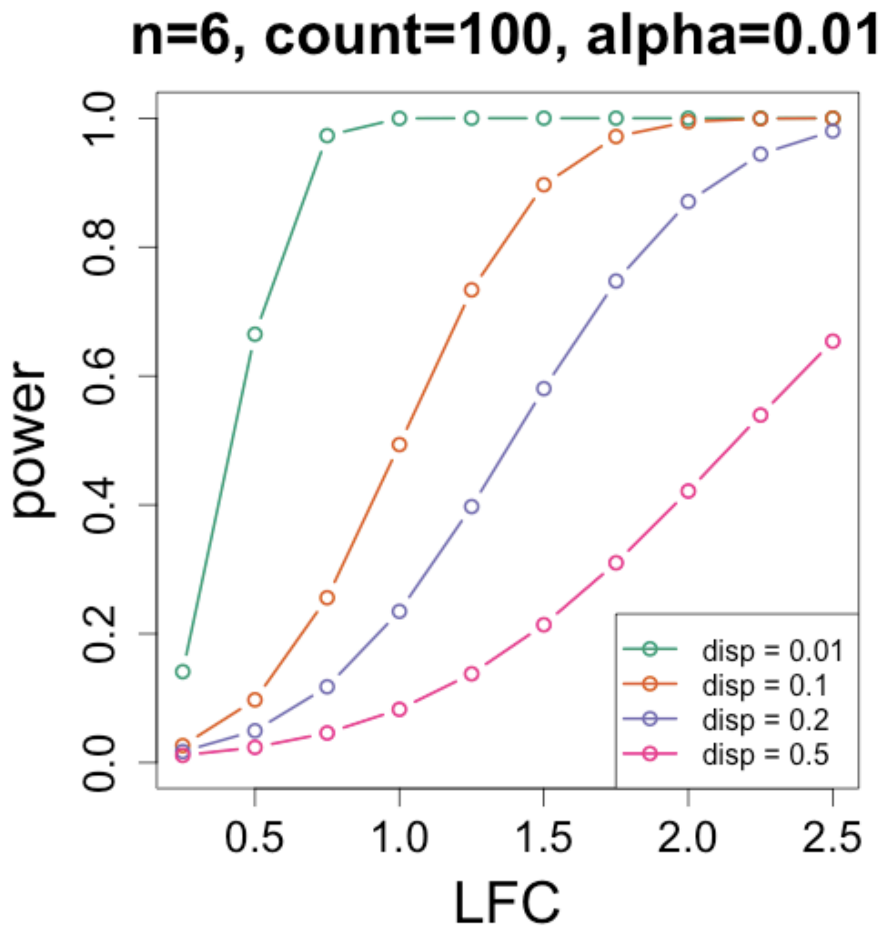
Factors influencing power

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

Bioc pkg: RNASeqPower

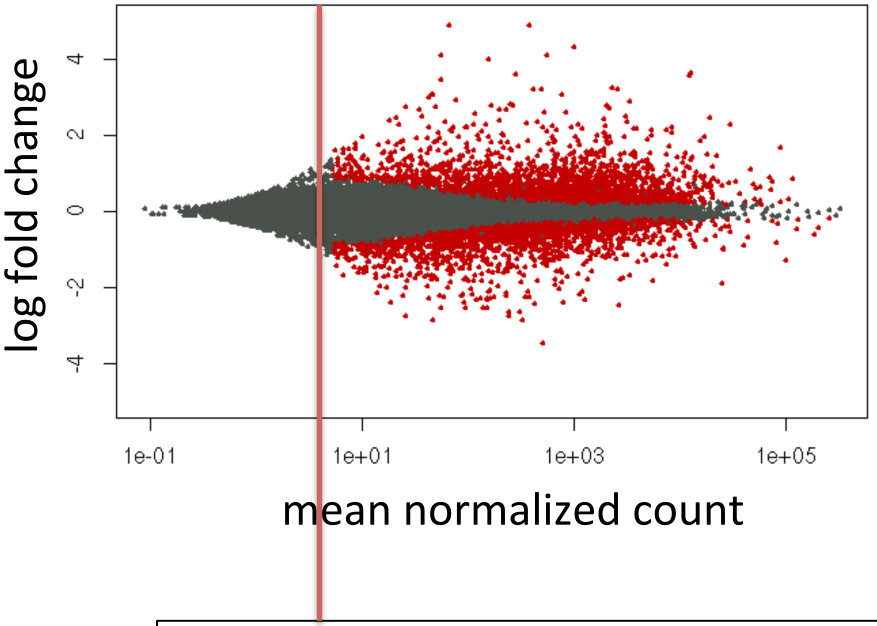


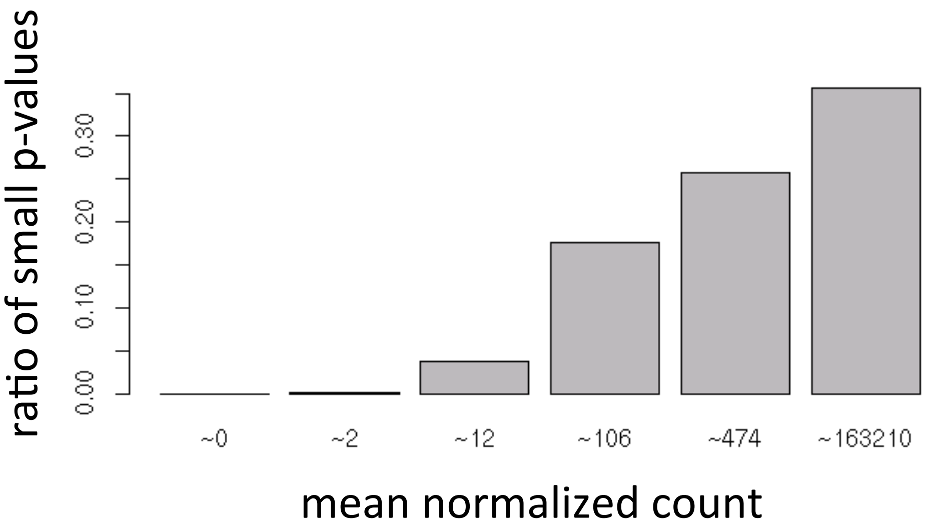
varying the count



varying the dispersion

Power depends on range of counts

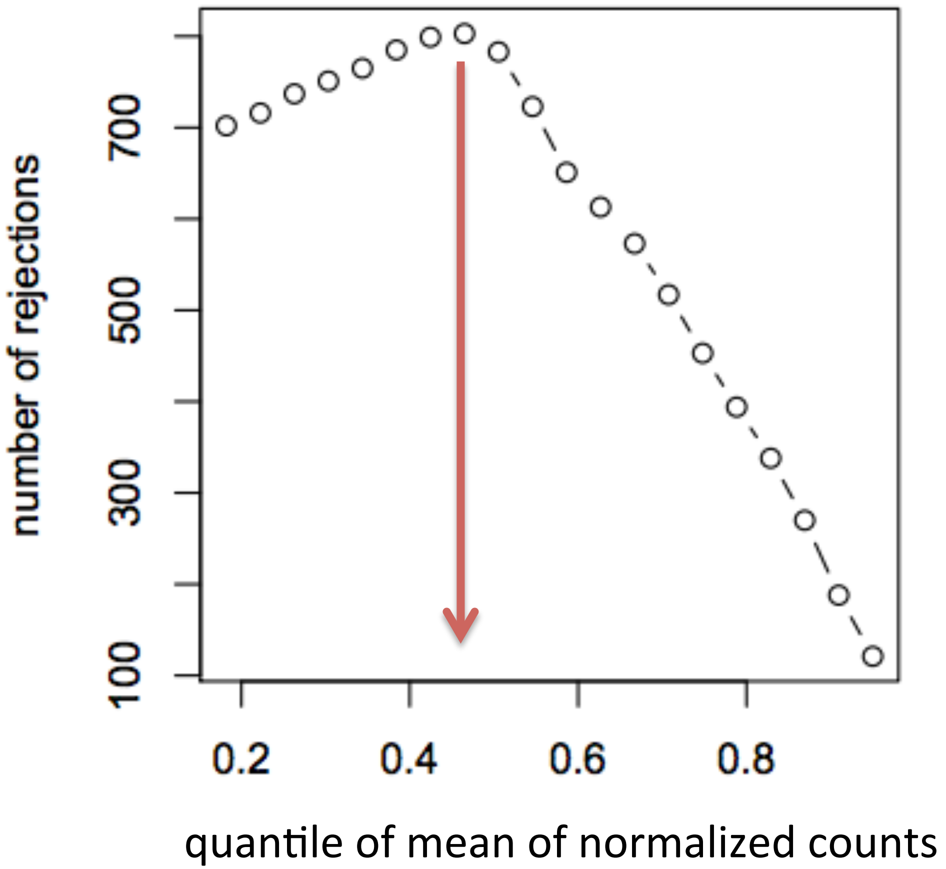




By excluding some tests, e.g. genes with mean normalized count < 5,

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

Power depends on range of counts



- 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.

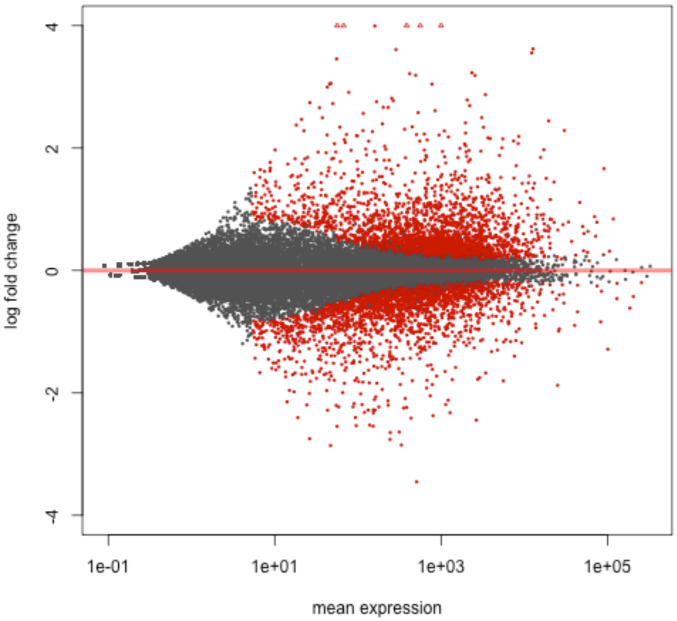
Independent Hypothesis Weighting

- Wolfgang will teach later this week...

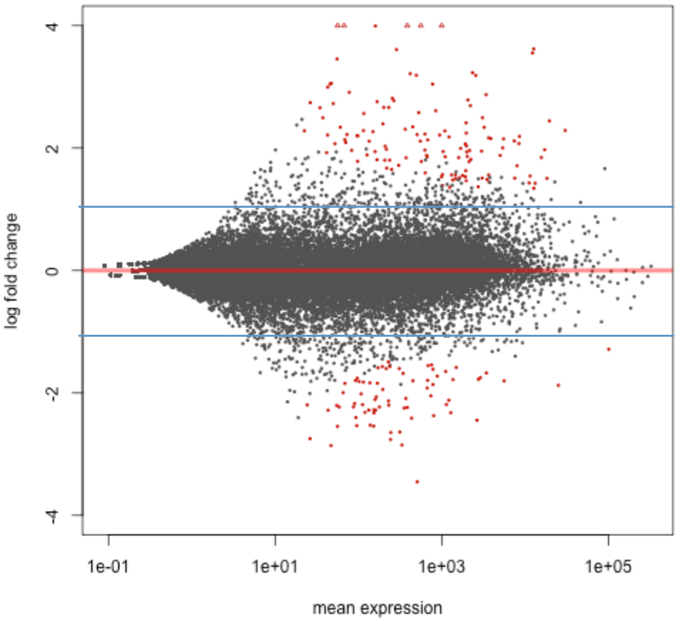
Testing against a threshold

"We get too many DEGs..."

using 'lfcThreshold' in results()



null hypothesis: fold change = 1



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**."

Bioconductor help

- Vignettes:

```
> browseVignettes("DESeq2")
> vignette("DESeq2")
```

- Type ? then the function name:

```
> ?results
```

Bioconductor help

```
results                 package:DESeq2                 R Documentation

Extract results from a DESeq analysis

Description:

     ‘results’ extracts a result table from a DESeq analysis giving
     base means across samples, log2 fold changes, standard errors,
     test statistics, p-values and adjusted p-values; ‘resultsNames’
     returns the names of the estimated effects (coefficents) of the
     model; ‘removeResults’ returns a ‘DESeqDataSet’ object with
     results columns removed.

Usage:

     results(object, contrast, name, lfcThreshold = 0,
       altHypothesis = c("greaterAbs", "lessAbs", "greater", "less"),
       listValues = c(1, -1), cooksCutoff, independentFiltering = TRUE,
       alpha = 0.1, filter, theta, pAdjustMethod = "BH",
       format = c("DataFrame", "GRanges", "GRangesList"), test, addMLE = FALSE,
       tidy = FALSE, parallel = FALSE, BPPARAM = bpparam())

...

Arguments:

  object: a DESeqDataSet, on which one of the following functions has
          already been called: ‘DESeq’, ‘nbinomWaldTest’, or
          ‘nbinomLRT’

contrast: this argument specifies what comparison to extract from the
          ‘object’ to build a results table. one of either:

            • a character vector with exactly three elements: the name
              of a factor in the design formula, the name of the
              numerator level for the fold change, and the name of the
              denominator level for the fold change (simplest case)
```

Bioconductor help

```
Value:

     For ‘results’: a ‘DESeqResults’ object, which is a simple subclass
     of DataFrame. This object contains the results columns:
     ‘baseMean’, ‘log2FoldChange’, ‘lfcSE’, ‘stat’, ‘pvalue’ and
     ‘padj’, and also includes metadata columns of variable
     information....

...

References:

     Richard Bourgon, Robert Gentleman, Wolfgang Huber: Independent
     filtering increases detection power for high-throughput
     experiments. PNAS (2010), <URL:
     http://dx.doi.org/10.1073/pnas.0914005107>

See Also:

     ‘DESeq’

Examples:

     ## Example 1: simple two-group comparison

     dds <- makeExampleDESeqDataSet(m=4)

...
```

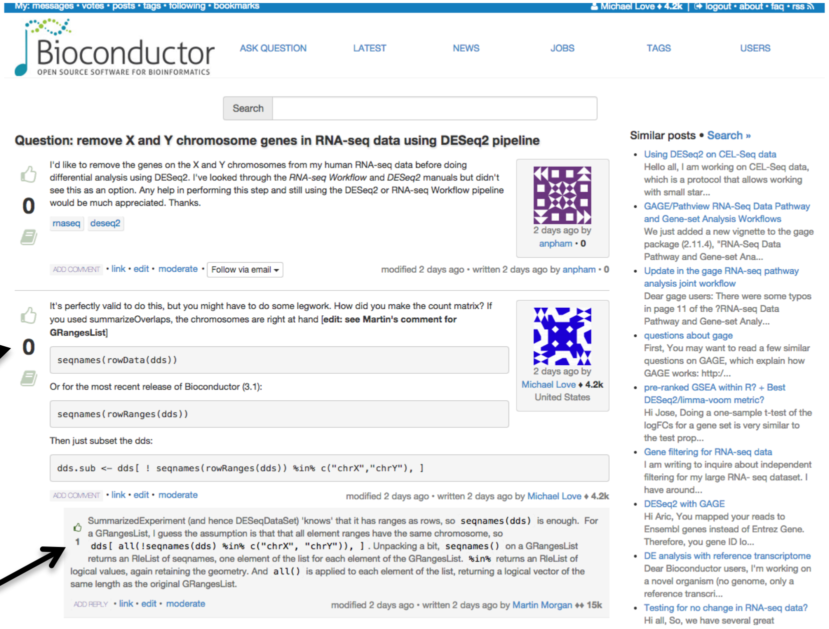
Looking up help for objects

```
> class(dds)
[1] "DESeqDataSet"
attr(,"package")
[1] "DESeq2"

> ?DESeqDataSet

> help(package="DESeq2", help_type="html")
```

Bioconductor support site

All questions about Bioconductor software post to:

support.bioconductor.org



My: messages • votes • posts • tags • following • bookmarks
Michael Love ♦ 4.2k | logout • about • faq • rss
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS
ASK QUESTION
LATEST
NEWS
JOBS
TAGS
USERS
Search

Question: remove X and Y chromosome genes in RNA-seq data using DESeq2 pipeline

0

I'd like to remove the genes on the X and Y chromosomes from my human RNA-seq data before doing differential analysis using DESeq2. I've looked through the *RNA-seq Workflow* and *DESeq2* manuals but didn't see this as an option. Any help in performing this step and still using the DESeq2 or RNA-seq Workflow pipeline would be much appreciated. Thanks.

rnaseq deseq2

2 days ago by anpham • 0

ADD COMMENT • link • edit • moderate • Follow via email

modified 2 days ago • written 2 days ago by anpham • 0

0

It's perfectly valid to do this, but you might have to do some legwork. How did you make the count matrix? If you used summarizeOverlaps, the chromosomes are right at hand [**edit: see Martin's comment for GRangesList**]

```
seqnames(rowData(dds))
```

Or for the most recent release of Bioconductor (3.1):

```
seqnames(rowRanges(dds))
```

Then just subset the dds:

```
dds.sub <- dds[ ! seqnames(rowRanges(dds)) %in% c("chrX","chrY"), ]
```

2 days ago by Michael Love ♦ 4.2k
United States

ADD COMMENT • link • edit • moderate

modified 2 days ago • written 2 days ago by Michael Love ♦ 4.2k

1

SummarizedExperiment (and hence DESeqDataSet) 'knows' that it has ranges as rows, so `seqnames(dds)` is enough. For a GRangesList, I guess the assumption is that that all element ranges have the same chromosome, so `dds[ all(!seqnames(dds) %in% c("chrX", "chrY")), ]`. Unpacking a bit, `seqnames()` on a GRangesList returns an RleList of seqnames, one element of the list for each element of the GRangesList. `%in%` returns an RleList of logical values, again retaining the geometry. And `all()` is applied to each element of the list, returning a logical vector of the same length as the original GRangesList.

ADD REPLY • link • edit • moderate

modified 2 days ago • written 2 days ago by Martin Morgan ♦♦ 15k

Similar posts • Search »

- Using DESeq2 on CEL-Seq data
  Hello all, I am working on CEL-Seq data, which is a protocol that allows working with small star...
- GAGE/Pathview RNA-Seq Data Pathway and Gene-set Analysis Workflows
  We just added a new vignette to the gage package (2.11.4), "RNA-Seq Data Pathway and Gene-set Ana...
- Update in the gage RNA-seq pathway analysis joint workflow
  Dear gage users: There were some typos in page 11 of the ?RNA-seq Data Pathway and Gene-set Analy...
- questions about gage
  First, You may want to read a few similar questions on GAGE, which explain how GAGE works: http:/...
- pre-ranked GSEA within R? + Best DESeq2/limma-voom metric?
  Hi Jose, Doing a one-sample t-test of the logFCs for a gene set is very similar to the test prop...
- Gene filtering for RNA-seq data
  I am writing to inquire about independent filtering for my large RNA- seq dataset. I have around...
- DESeq2 with GAGE
  Hi Aric, You mapped your reads to Ensembl genes instead of Entrez Gene. Therefore, you gene ID lo...
- DE analysis with reference transcriptome
  Dear Bioconductor users, I'm working on a novel organism (no genome, only a reference transcri...
- Testing for no change in RNA-seq data?
  Hi all, So, we have several great

edit posts

voting

comment / reply

always provide:

- biological question
- **all** code, any errors/warnings
- `sessionInfo()`