RNA-seq gene-level analysis and differential expression

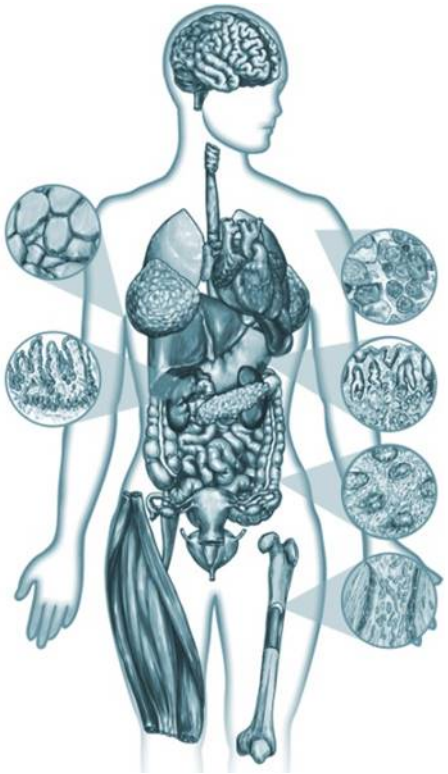
Michael Love
Biostatistics Department
Genetics Department
UNC Chapel Hill

Technology and quantification

RNA-SEQ PART I

Gene expression

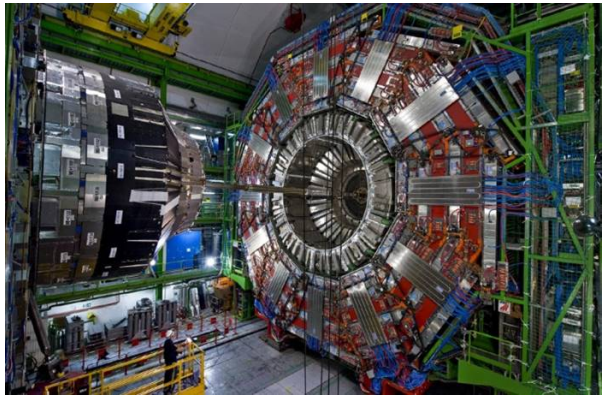
- Dynamic across time, tissue, individuals
- Measurement is harder than for the genome



[Roadmap Epigenomics Project]

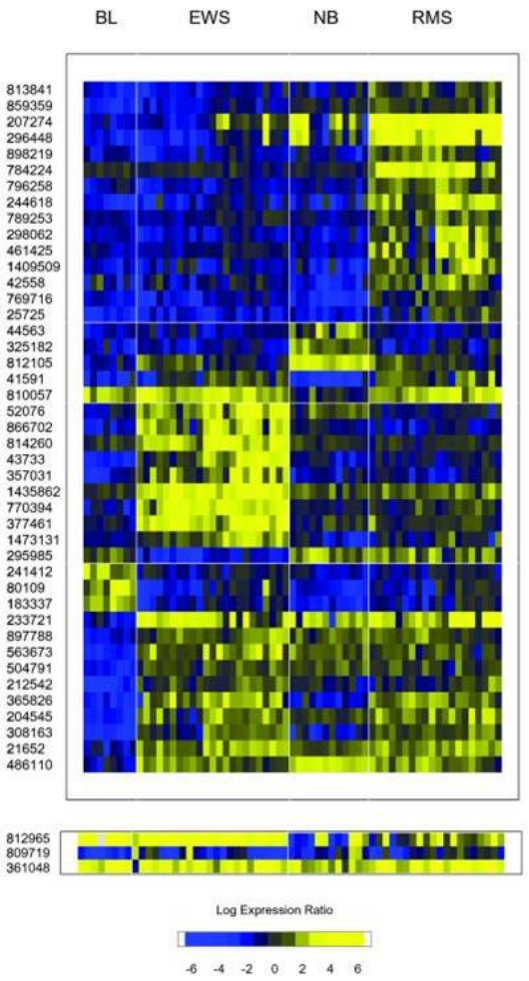
Why gene expression?

- Basic biology: transcription, translation, RNA enzyme
- As a phenotype: easier to measure than proteins
- Research: find interesting genetic loci
- Diagnostic: classify cancer subtypes ← FDA (2007)
- New and better measuring devices drive discovery



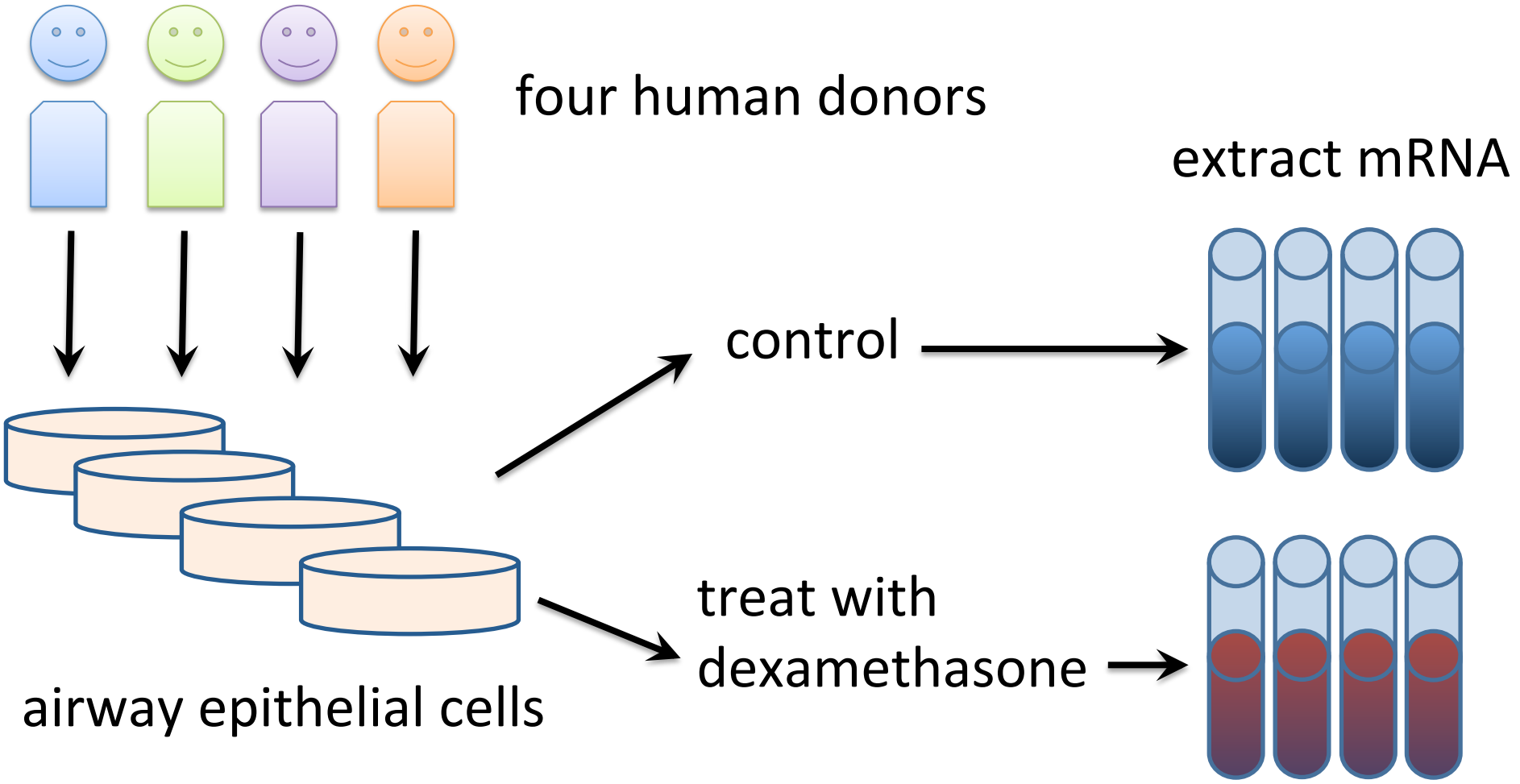
Statistical analysis of gene expression

- Previously: Northern blot, qPCR
- Era of microarrays: 1990s-now
- Clustering, differential expression
- Seems simple, but statistical methods offer huge benefit
- Key insight about expression data:
  - Costly, often few replicates (3-5)
  - Many genes over which to learn about parameters

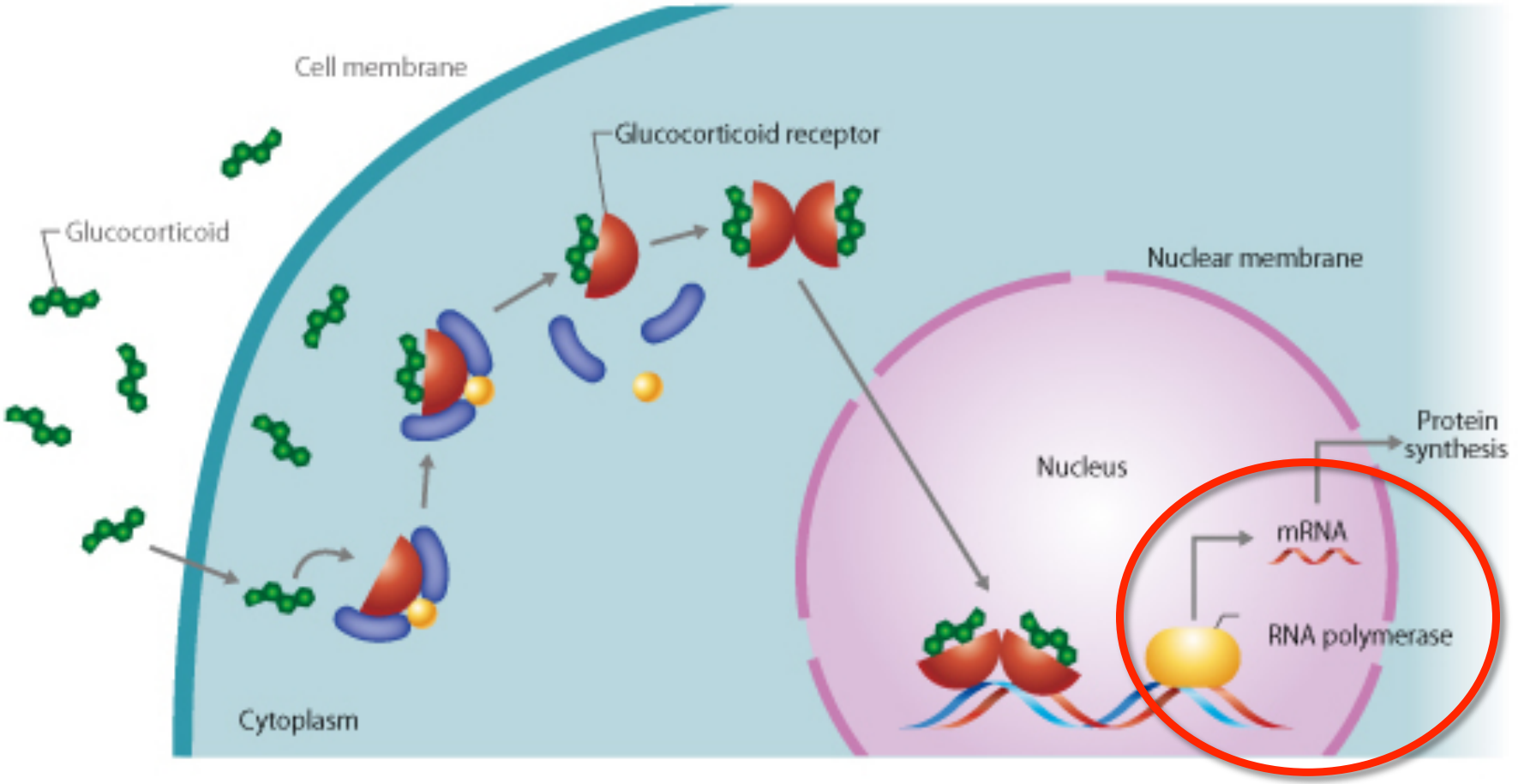


Tibshirani et al (2002)

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



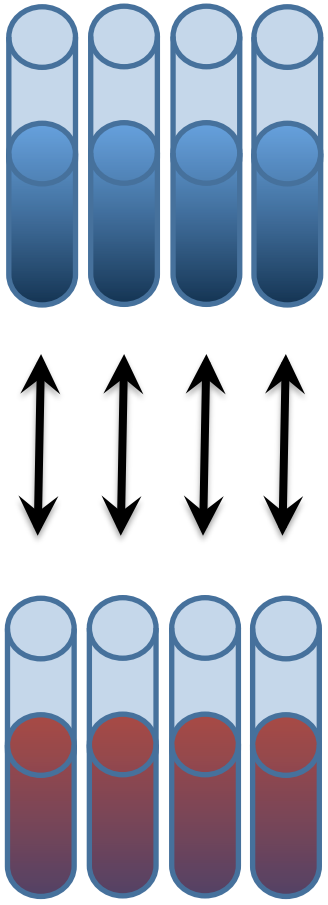
Glucocorticoid mechanism of action



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

Compare gene expression across treatment, within cell line

cDNA libraries



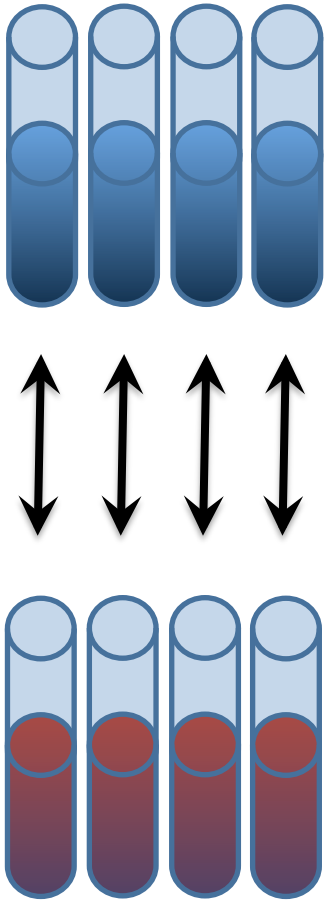
control

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

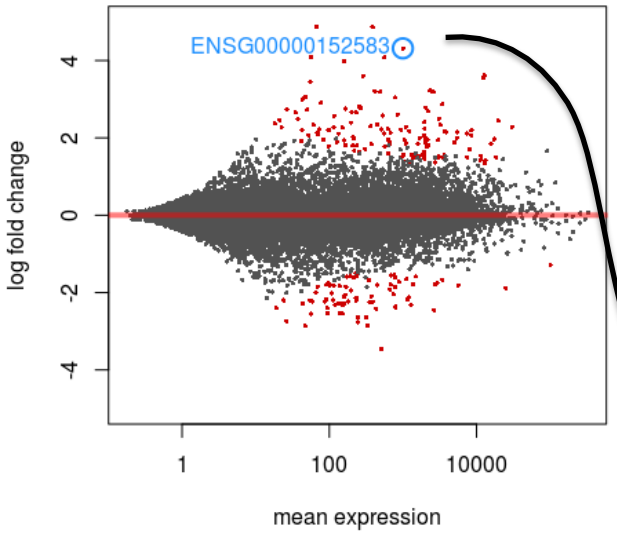
cDNA libraries



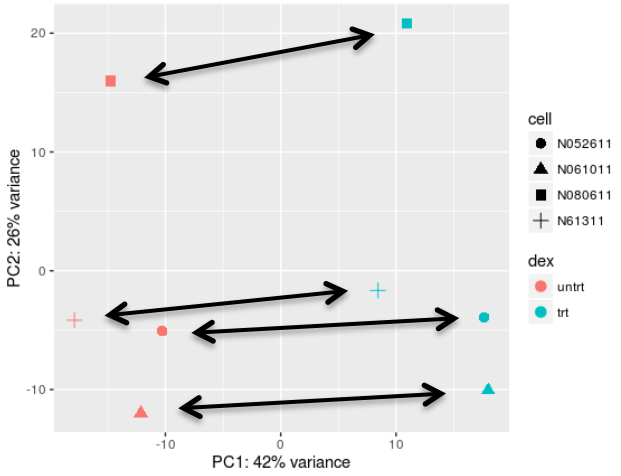
control

treated with dex.

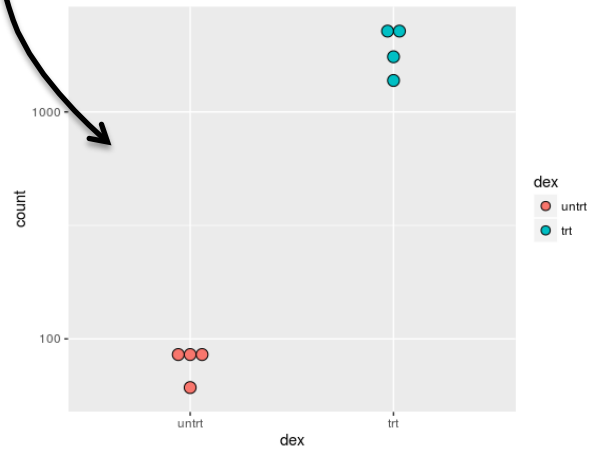
MA plot



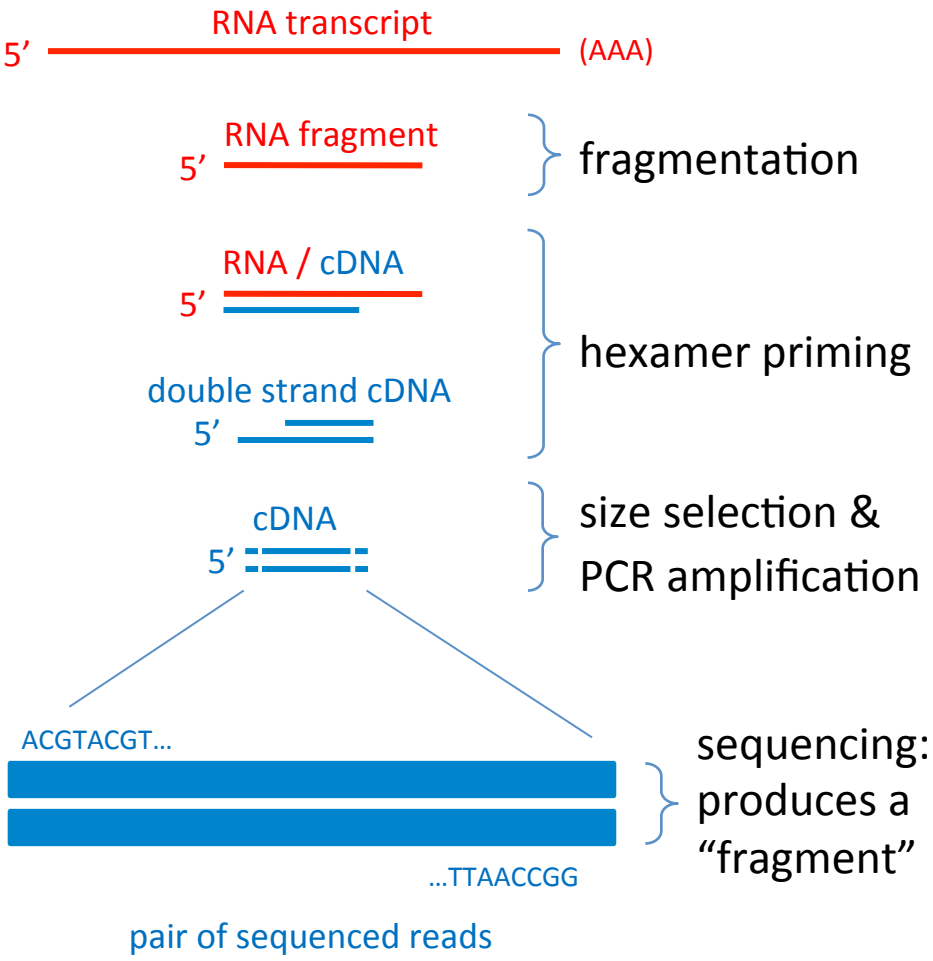
PCA plot

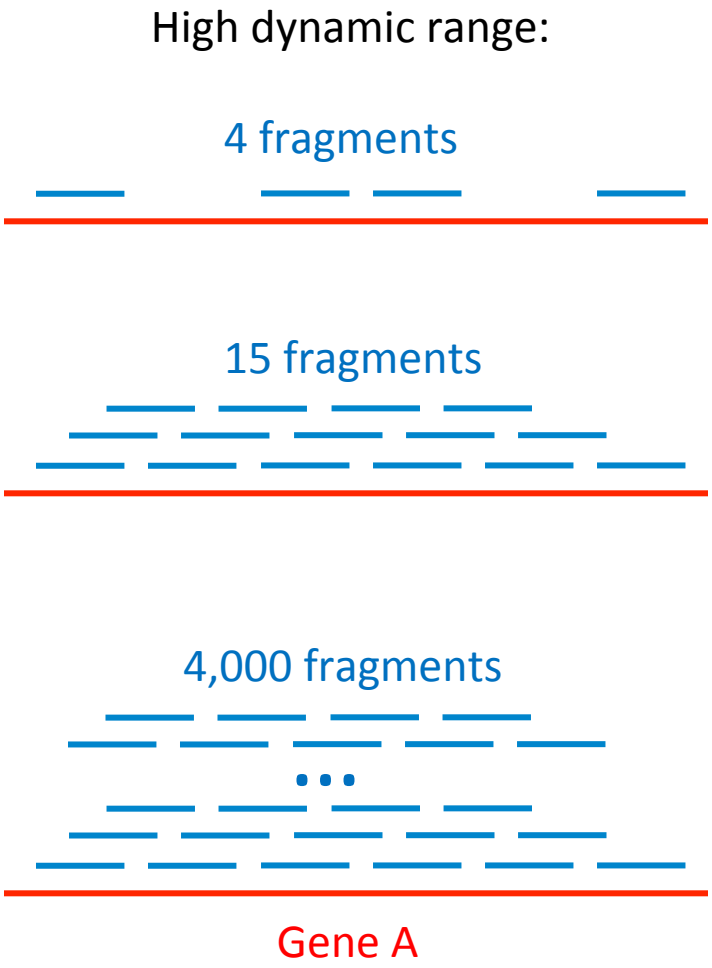


"counts plot"

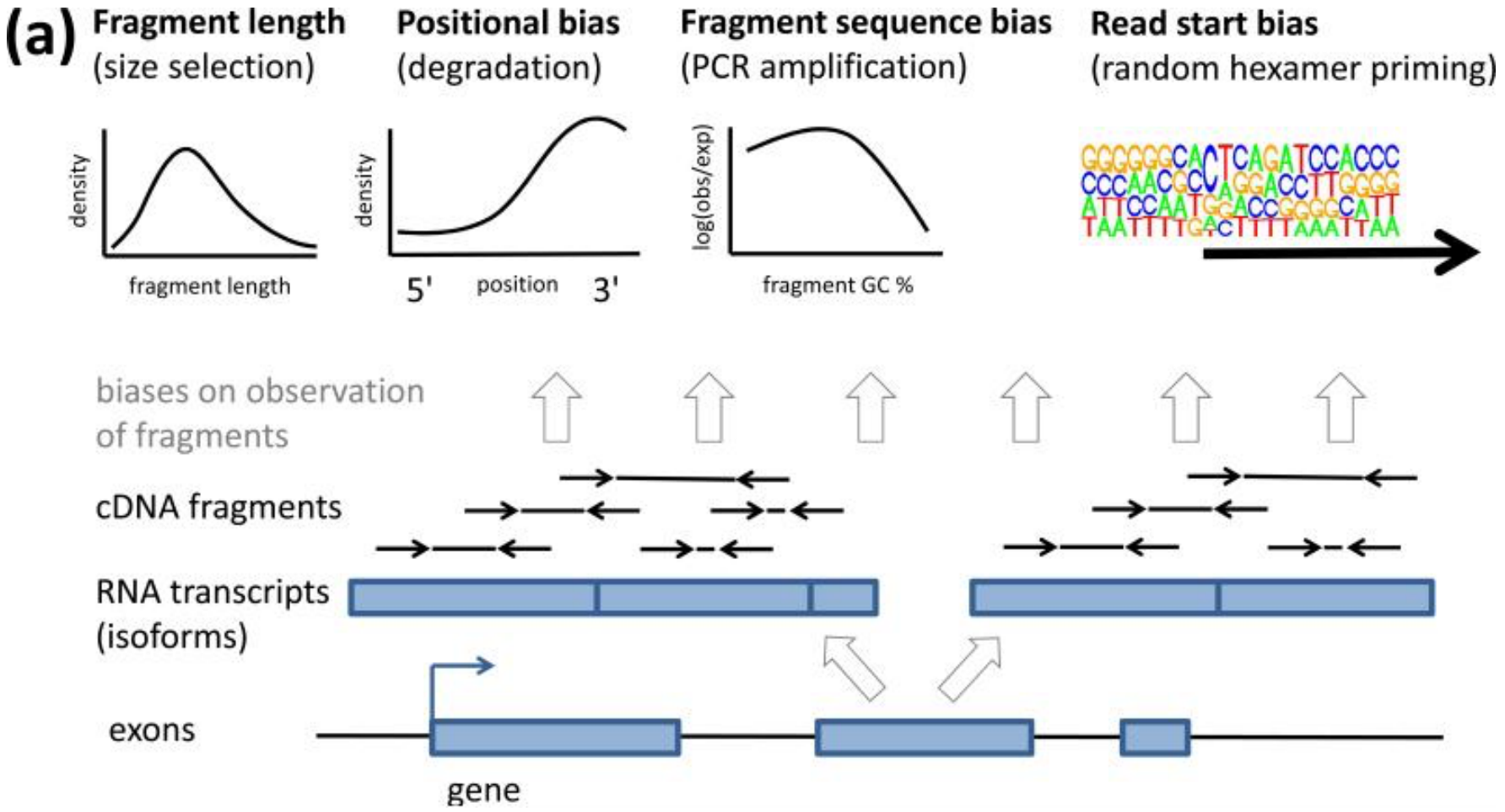


RNA sequencing protocol





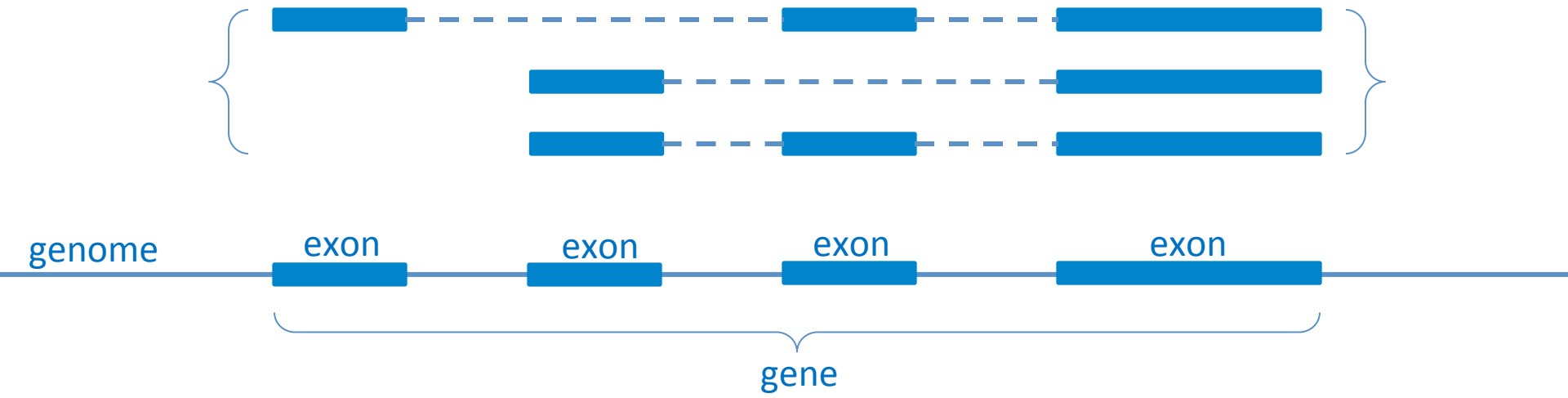
Biases of RNA-seq experiment



Love, "Modeling of RNA-seq fragment sequence bias..." (2016)

More complex

- Gene: region of genome
- Three “isoforms”, also called “transcripts”



→ Isoform usage differs across tissue, relevant in disease, including cancer

High-throughput sequencing data

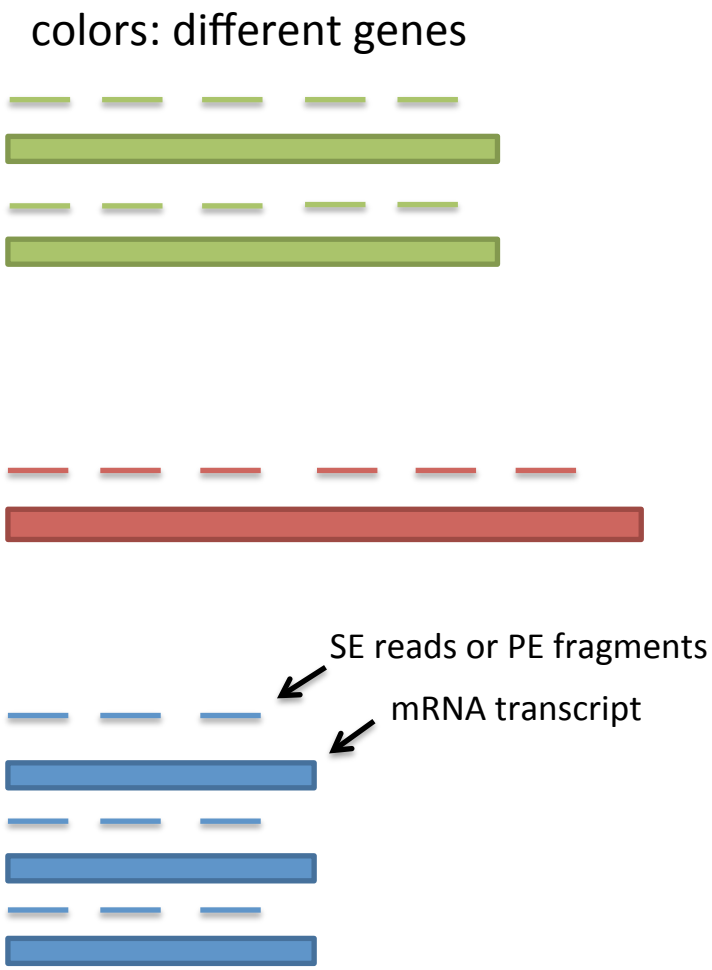
- often observed data consists of *counts* of reads/fragments across features (rows) and samples (columns)
- counts need an appropriate statistical model (normalization and variance modeling)

samples: want to see if differences across condition are significant
(w.r.t. biological and technical variation)

features (e.g. genes)

	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516
ENSG00000000003	679	448	873	408	1138
ENSG00000000005	0	0	0	0	0
ENSG00000000419	467	515	621	365	587
ENSG00000000457	260	211	263	164	245
ENSG00000000460	60	55	40	35	78

mRNAs to RNA-seq fragments



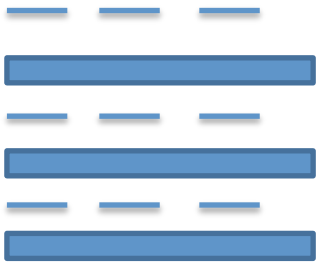
K_{ij} = count of fragments aligned to gene i, sample j

is proportional to:

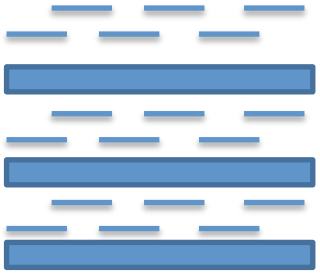
- expression of RNA
- length of gene
- sequencing depth
- lib. prep. factors (PCR)
- in silico factors (alignment)
- etc.

Sequencing depth

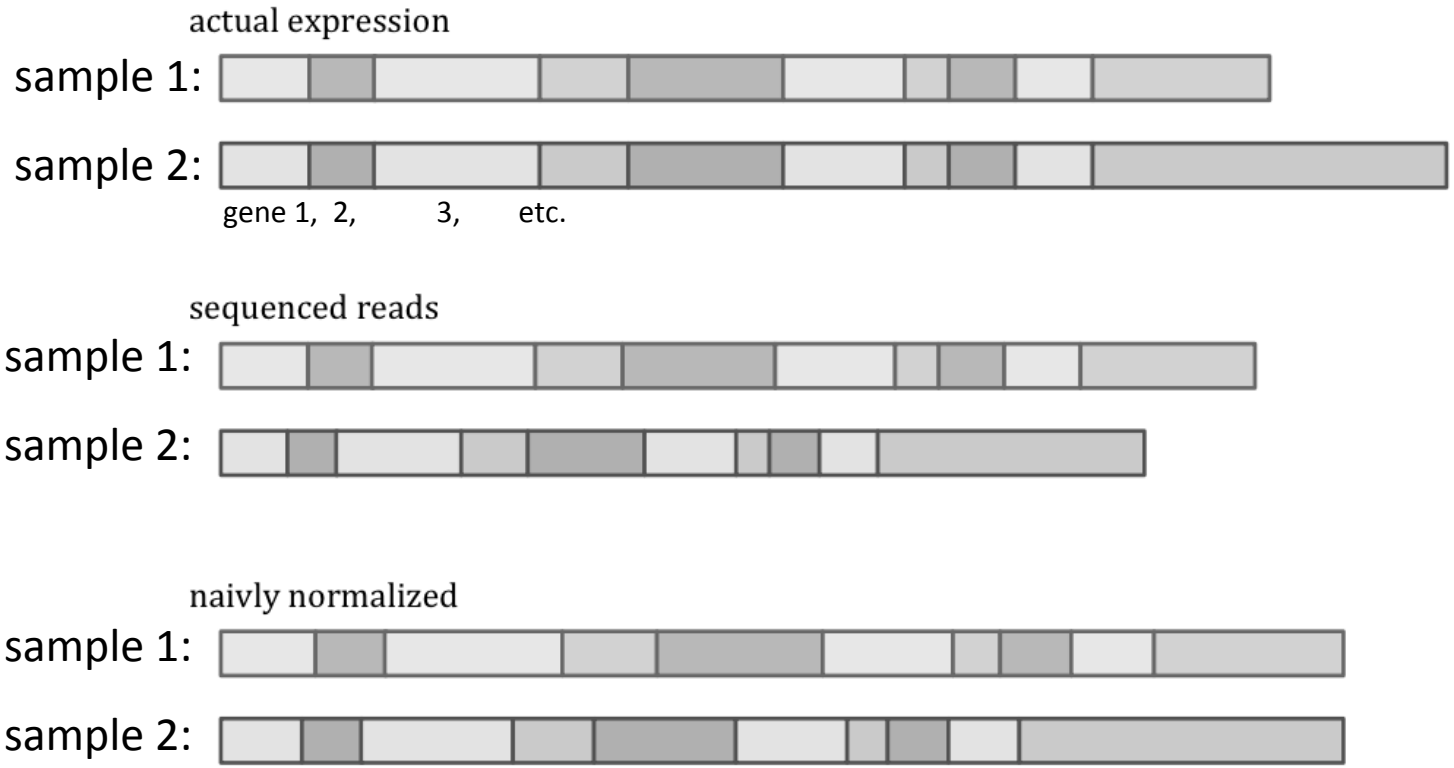
sample 1



sample 2



Need to have a robust estimator for sequencing depth

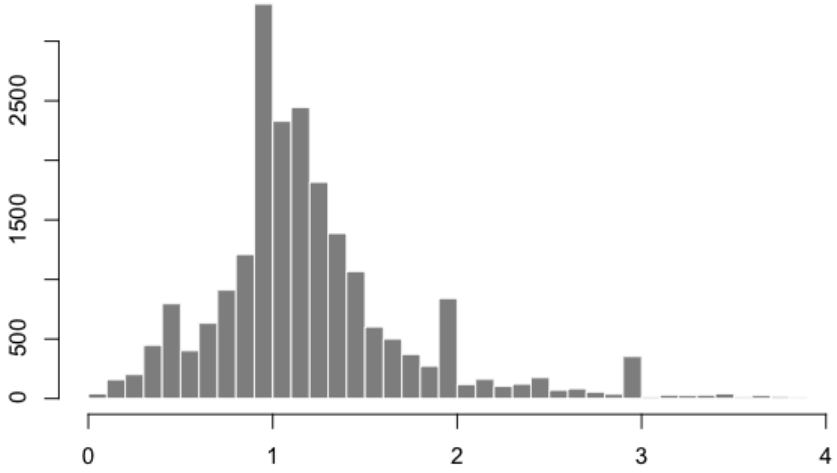


(slide from Simon Anders)

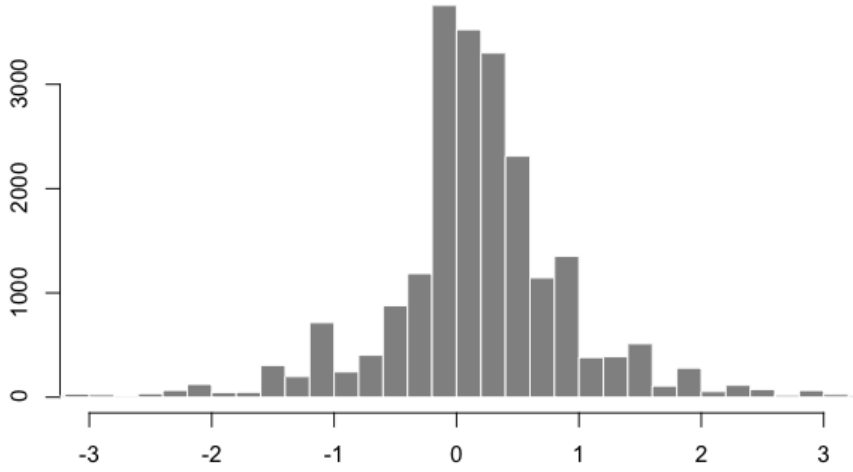
Median of ratios method

simple approach & works well
for each gene look at count ratios:

sample 1 / sample 2



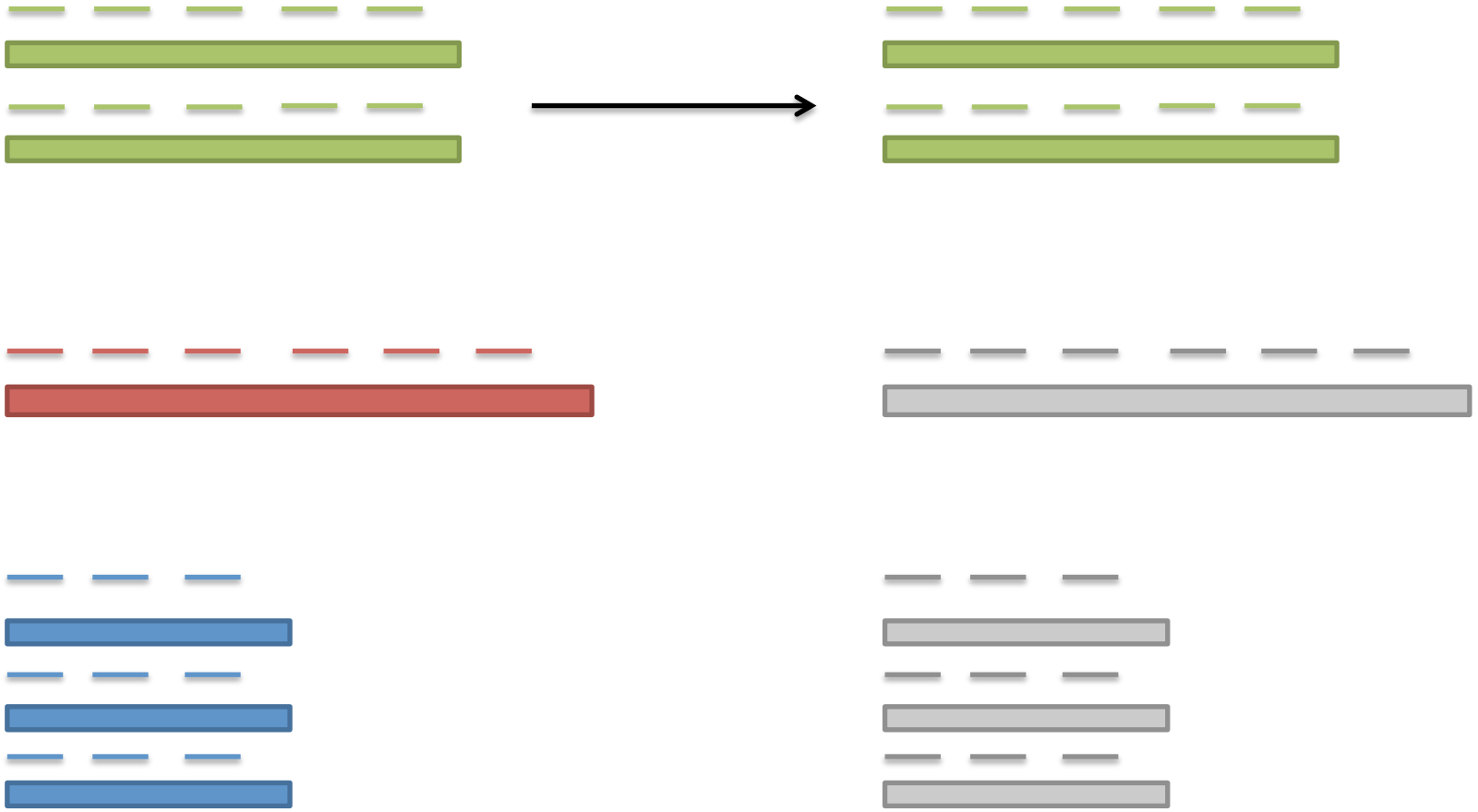
log2(sample1 / sample2)



- in general: create a pseudo-reference-sample (row-wise geometric mean)
- calculate ratio of each sample to the reference
- assumes that not *ALL* genes are DE (differentially expressed)
- robust to imbalance in up-/down- regulation and large numbers of DE genes

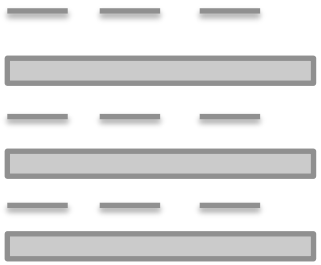
Variance of counts

Consider one gene:

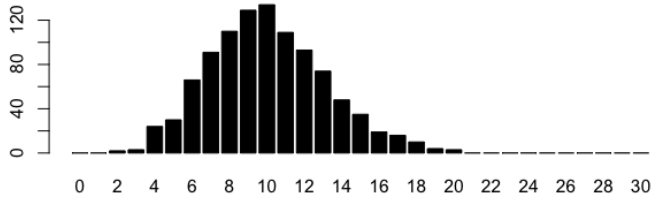


Variance of counts

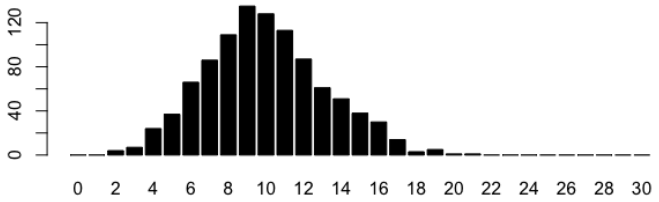
Consider one gene:



- Binomial sampling distribution

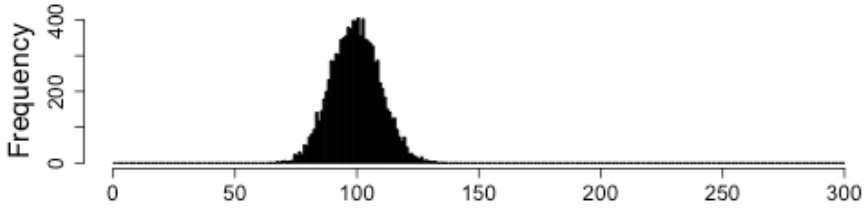


- With millions of reads & small proportion for each gene → Poisson sampling distribution

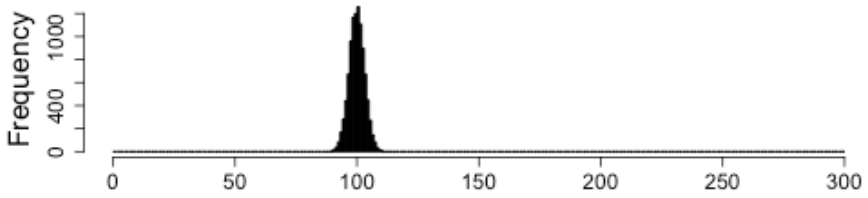


Raw counts vs. normalized counts

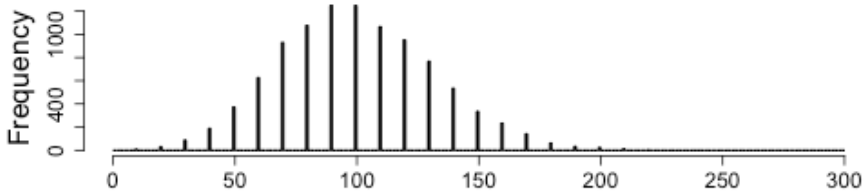
Raw count with mean of 100
Poisson sampling, so SD=10



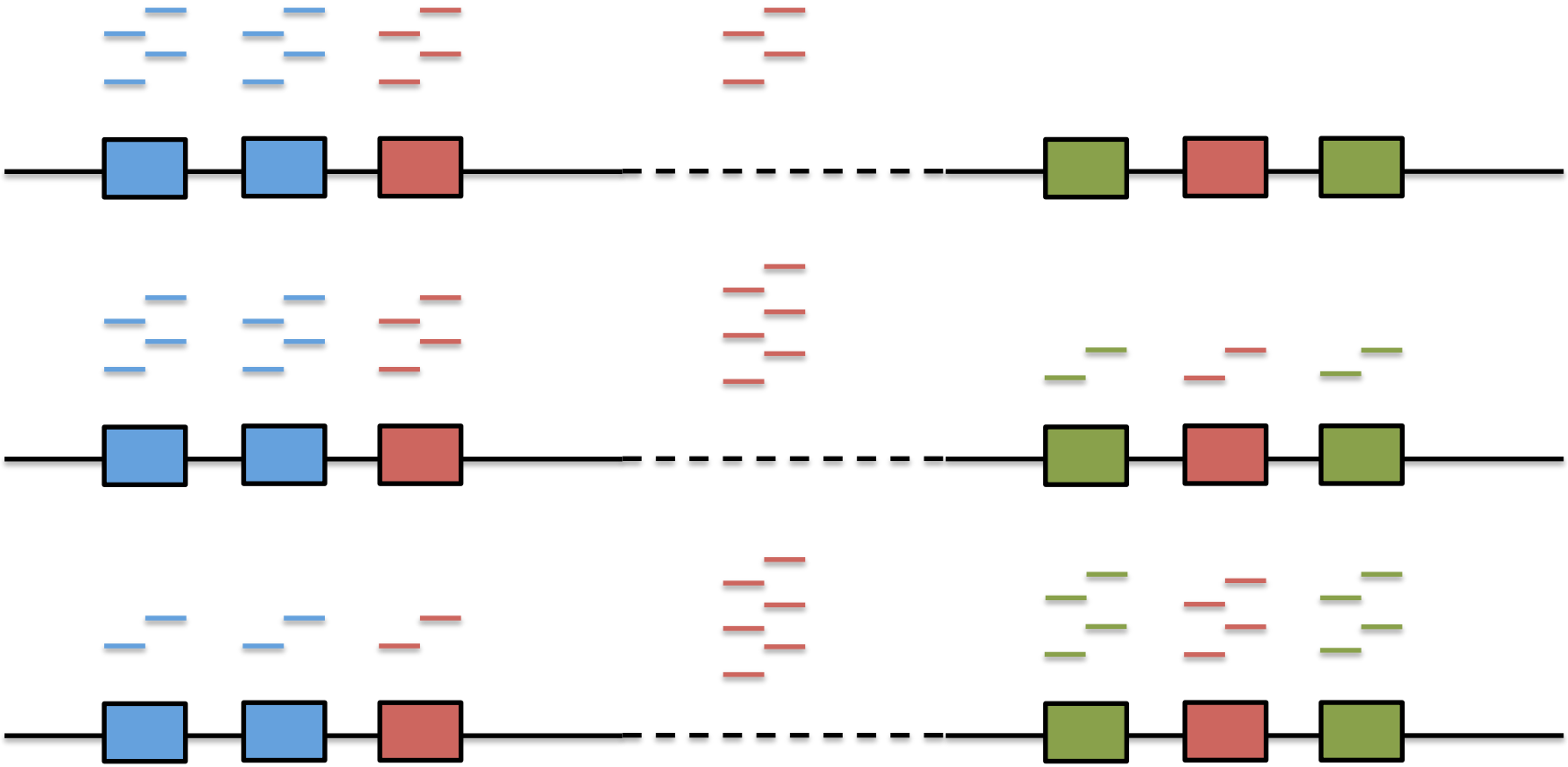
Raw count mean = 1000
Scaled by 1/10
SD = ?



Raw count mean = 10
Scaled by 10
SD = ?

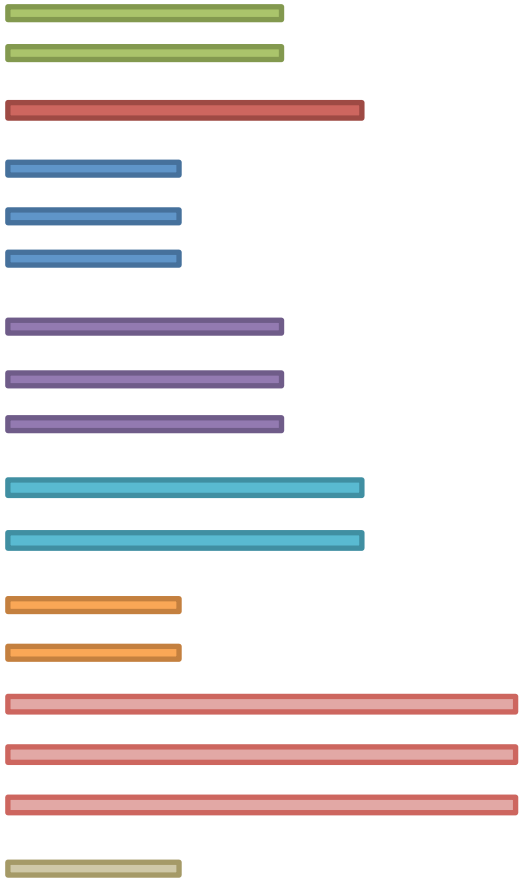


Raw gene counts (htseq, featureCounts, etc.) and estimated counts (Salmon, etc.)

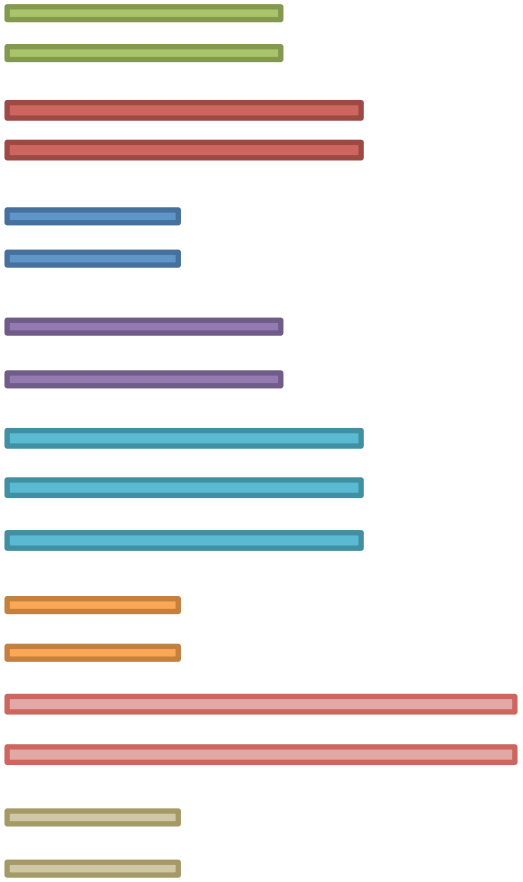


Biological replicates

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



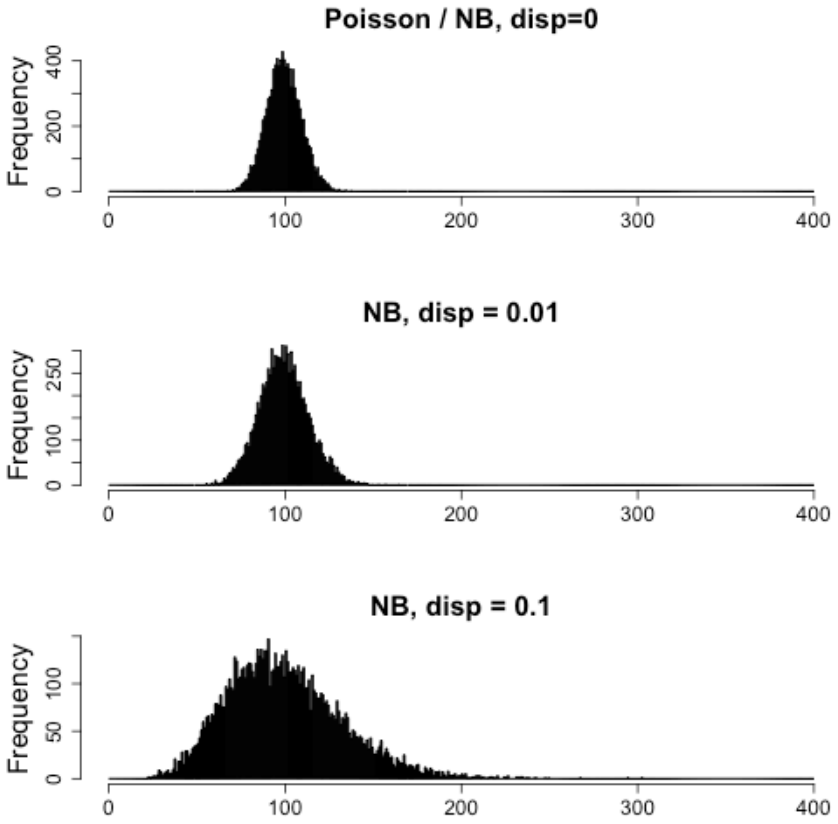
But realistically, biological variation across sample units is expected



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



Dispersion parameter

raw count for gene i, sample j

normalization factor

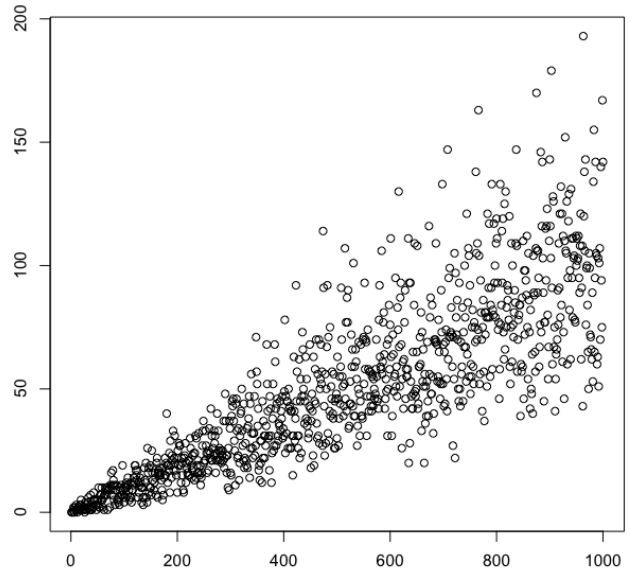
quantity of interest

one dispersion per gene

$$K_{ij} \sim \mathrm{NB}(s_{ij} q_{ij}, \alpha_i)$$

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

variance depends on mean value though

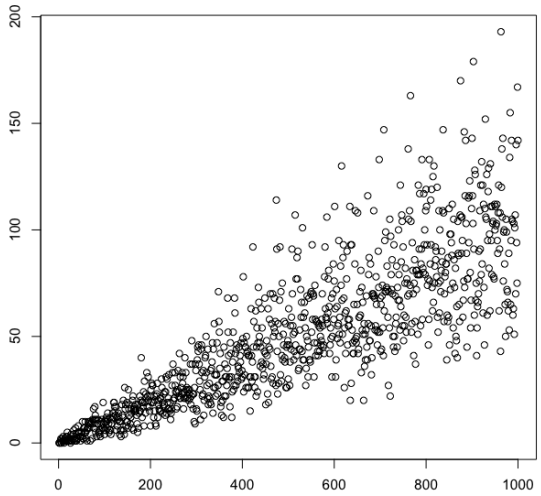


Dispersion parameter

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

Poisson part:
sampling fragments

Extra variation
due to biological variance



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

disp = 0.01 → CV 10%
disp = 0.25 → CV 50%

Transformations and power

RNA-SEQ PART II

Two paths in RNA-seq analysis

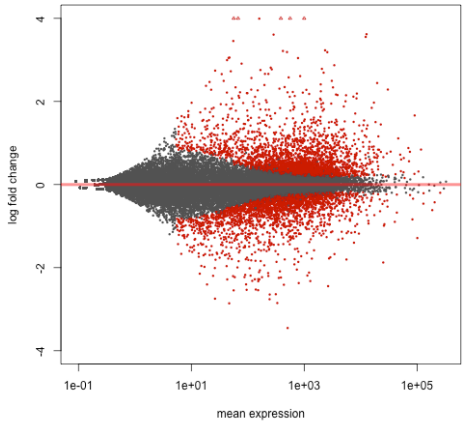
Count matrix

Differential expression

testing, p-values, FDR

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

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

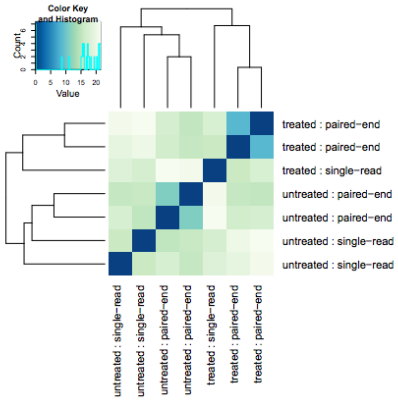


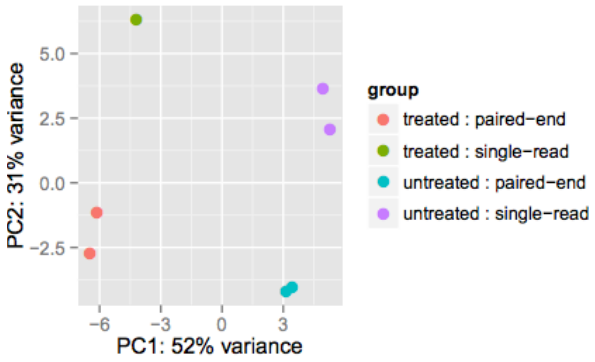
Transformations and Exploratory Data Analysis (EDA)

clustering, heatmaps,
sample-sample distances

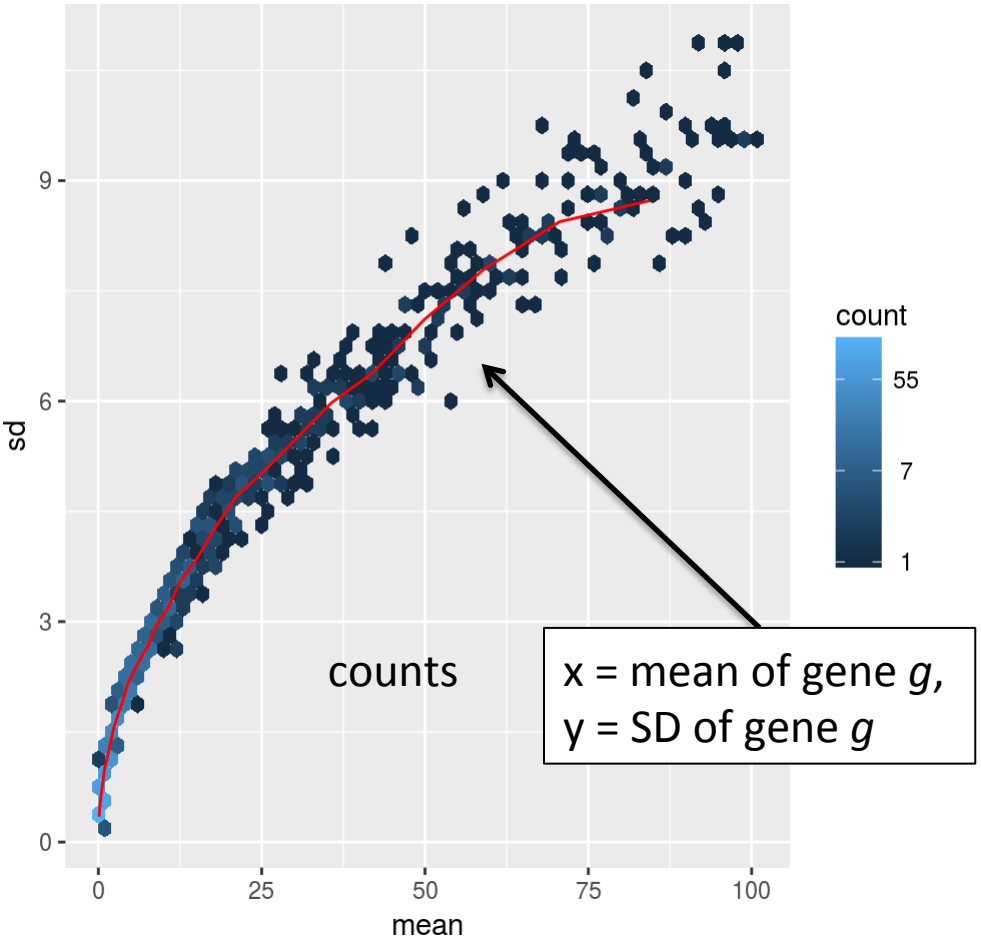
DESeq2 { `vst(), rlog(), plotPCA()`

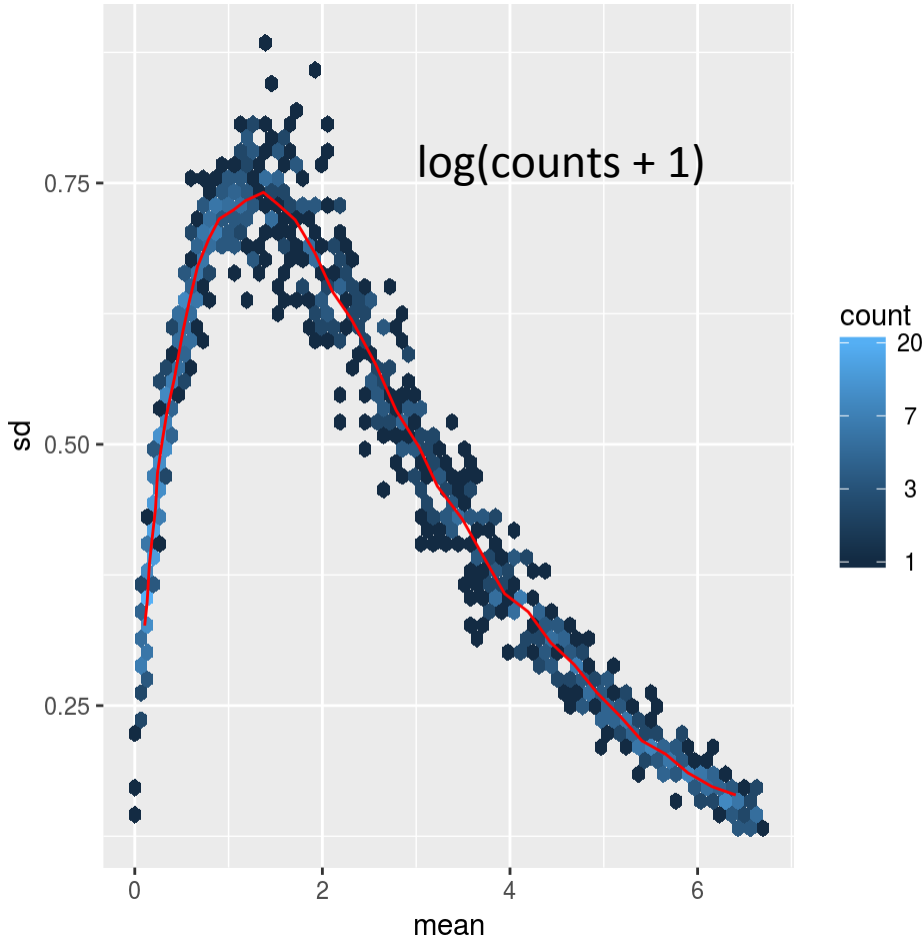
edgeR { `cpm(), plotMDS()`



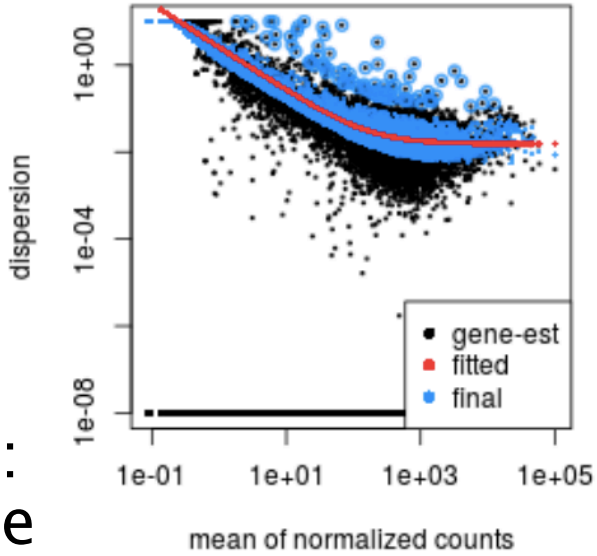


How to transform data for distances?





Variance stabilizing transformation

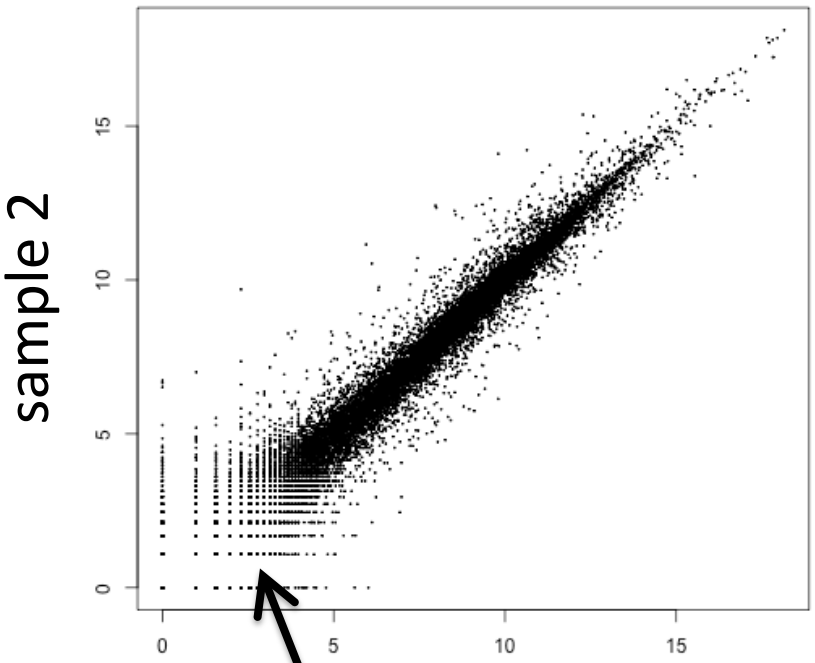


- *Variance stabilizing transformation*: calculate the dependence of variance on the mean (using the dispersion trend)
- Closed-form expression f(x) for stabilizing Var
- `vst()` is a *faster* implementation than `rlog()`

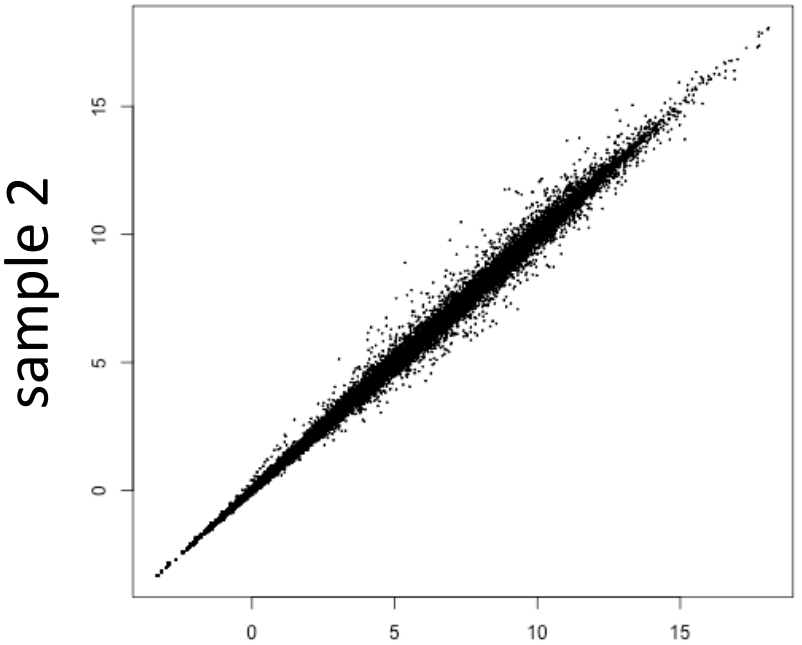
Regularized logarithm, "rlog"

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

log2(x + 1)

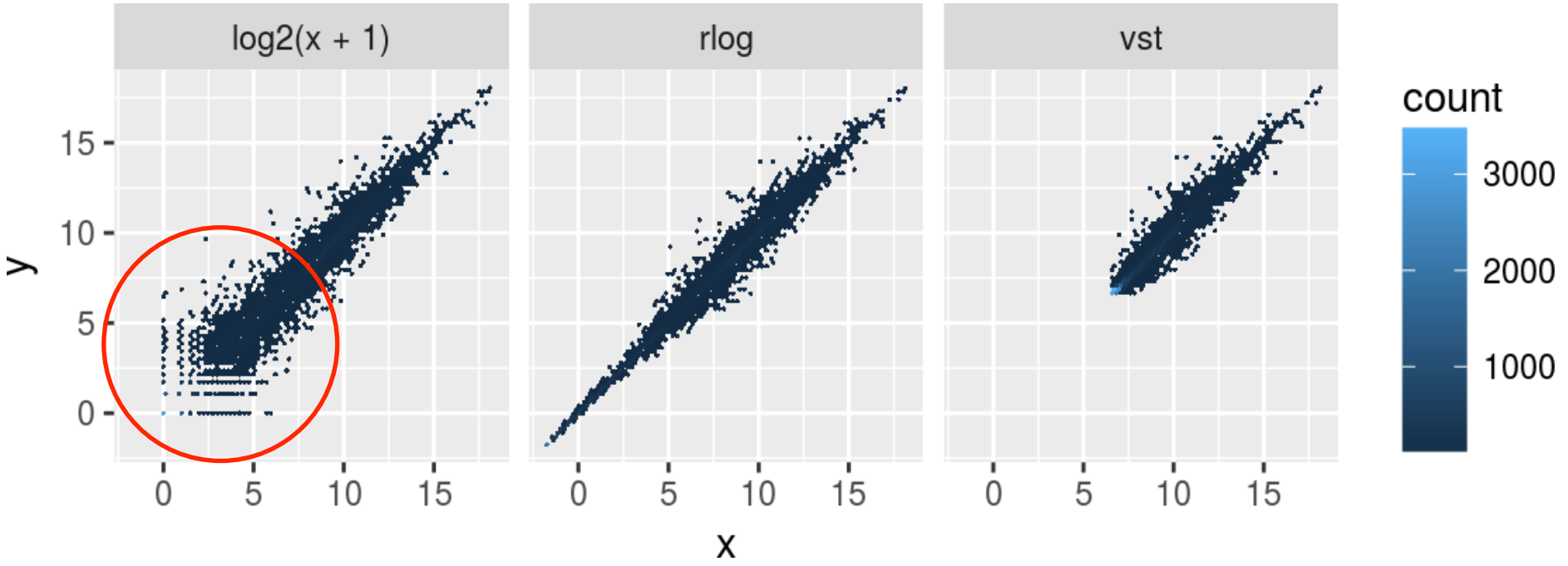


"rlog"



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

VST and rlog vs log(x+1)



Essentially provides a similar outcome as filtering at T and/or adding a pseudocount of X, but parameter estimation is data-driven.

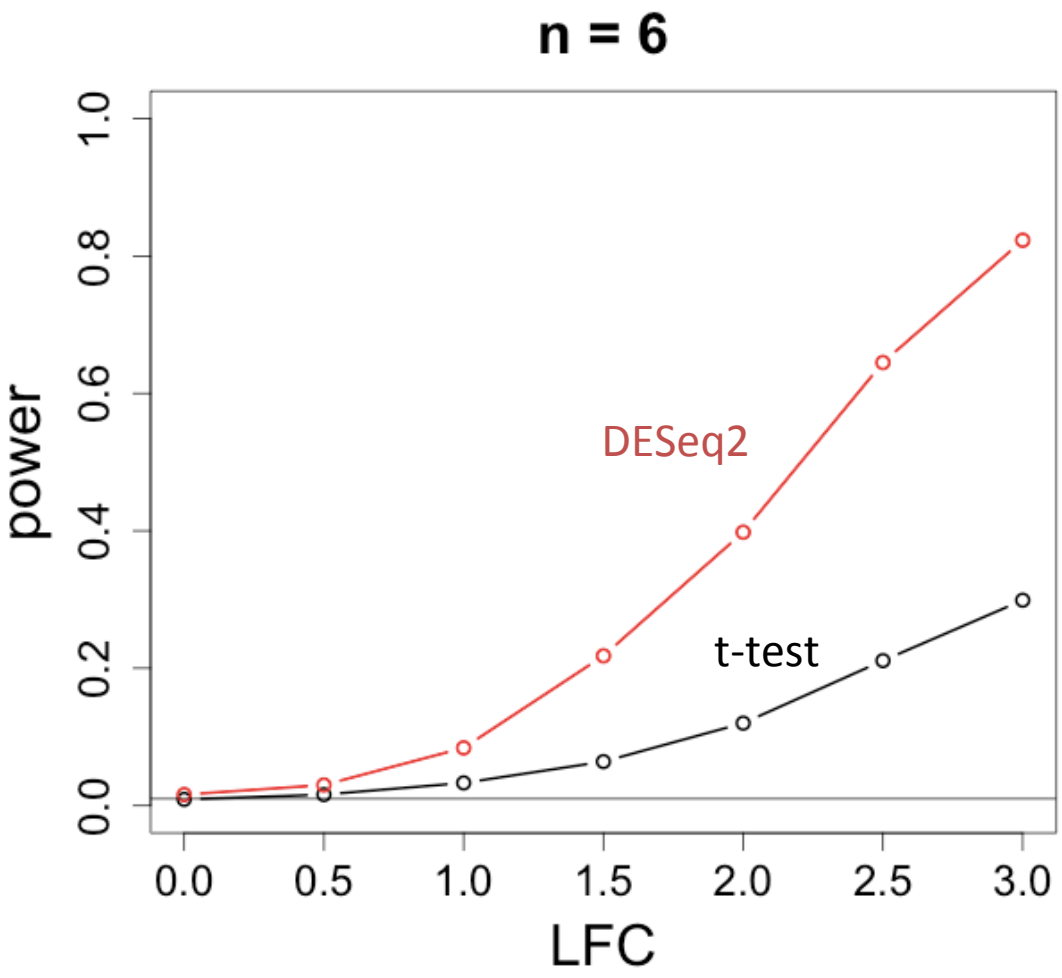
Statistical power

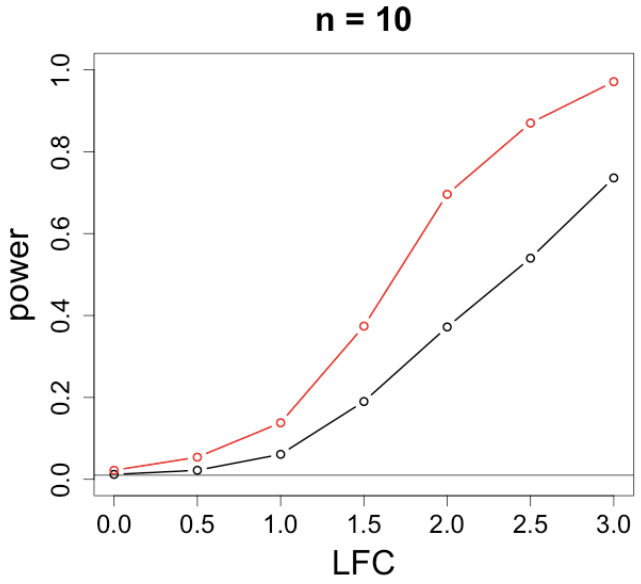
- False positive rate:
  of the null, how many positives?
- False discovery rate:
  of the positives, how many false positives?
- Power (sensitivity):
  of the non-null, how many positive?

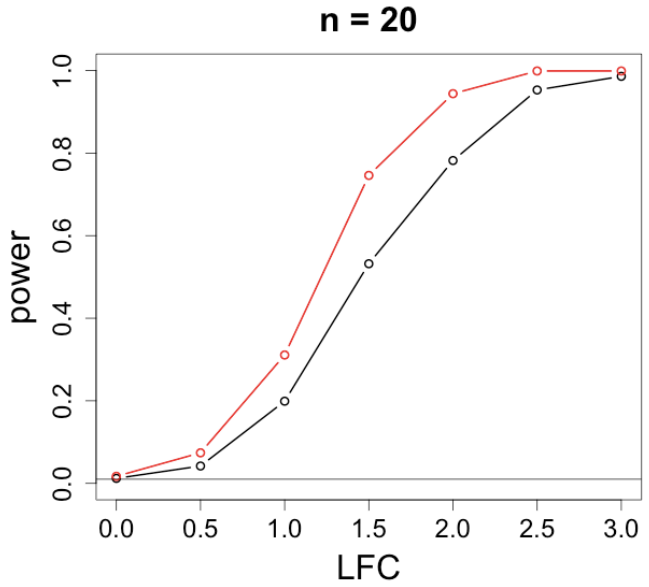
test: \ true:	null	non-null
negative	true negative	false negative
positive	false positive	true positive

Statistical power

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



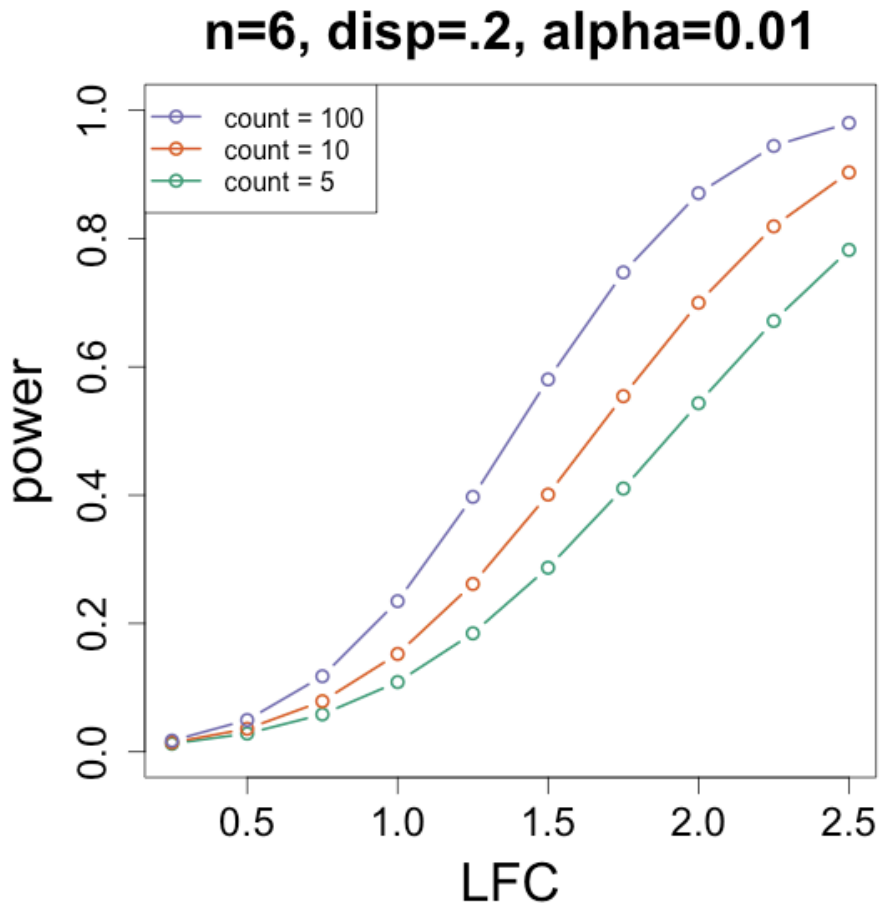


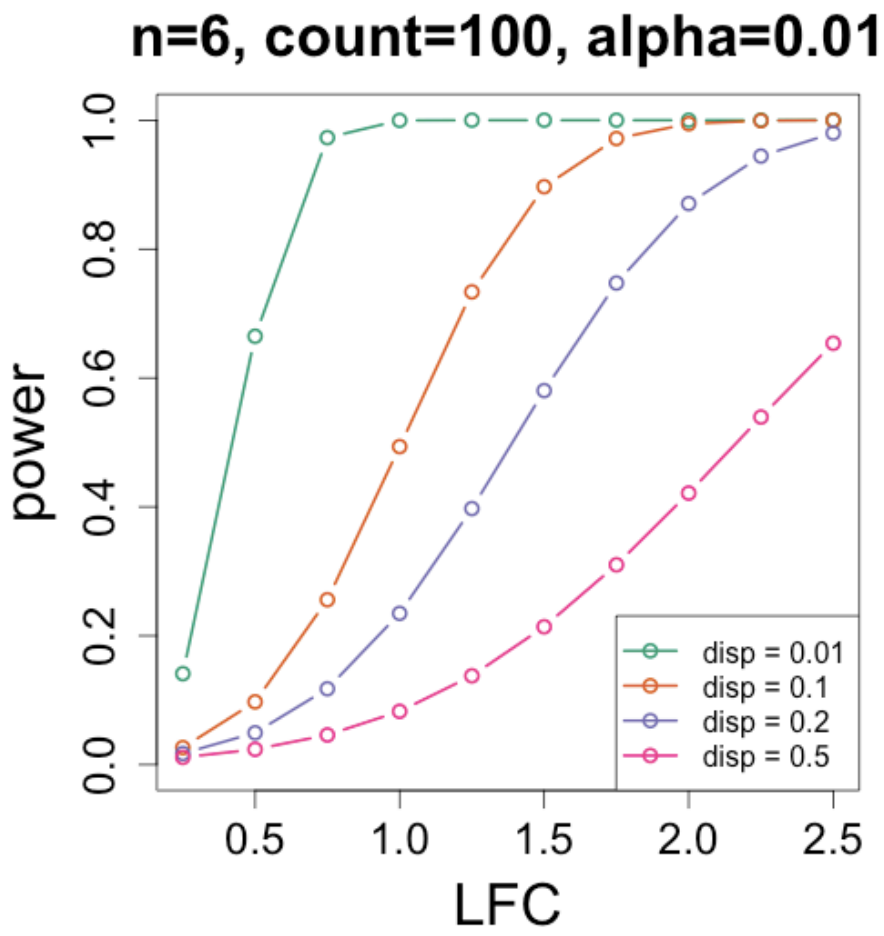


Factors influencing power

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

Bioc pkg: RNASeqPower





varying the count

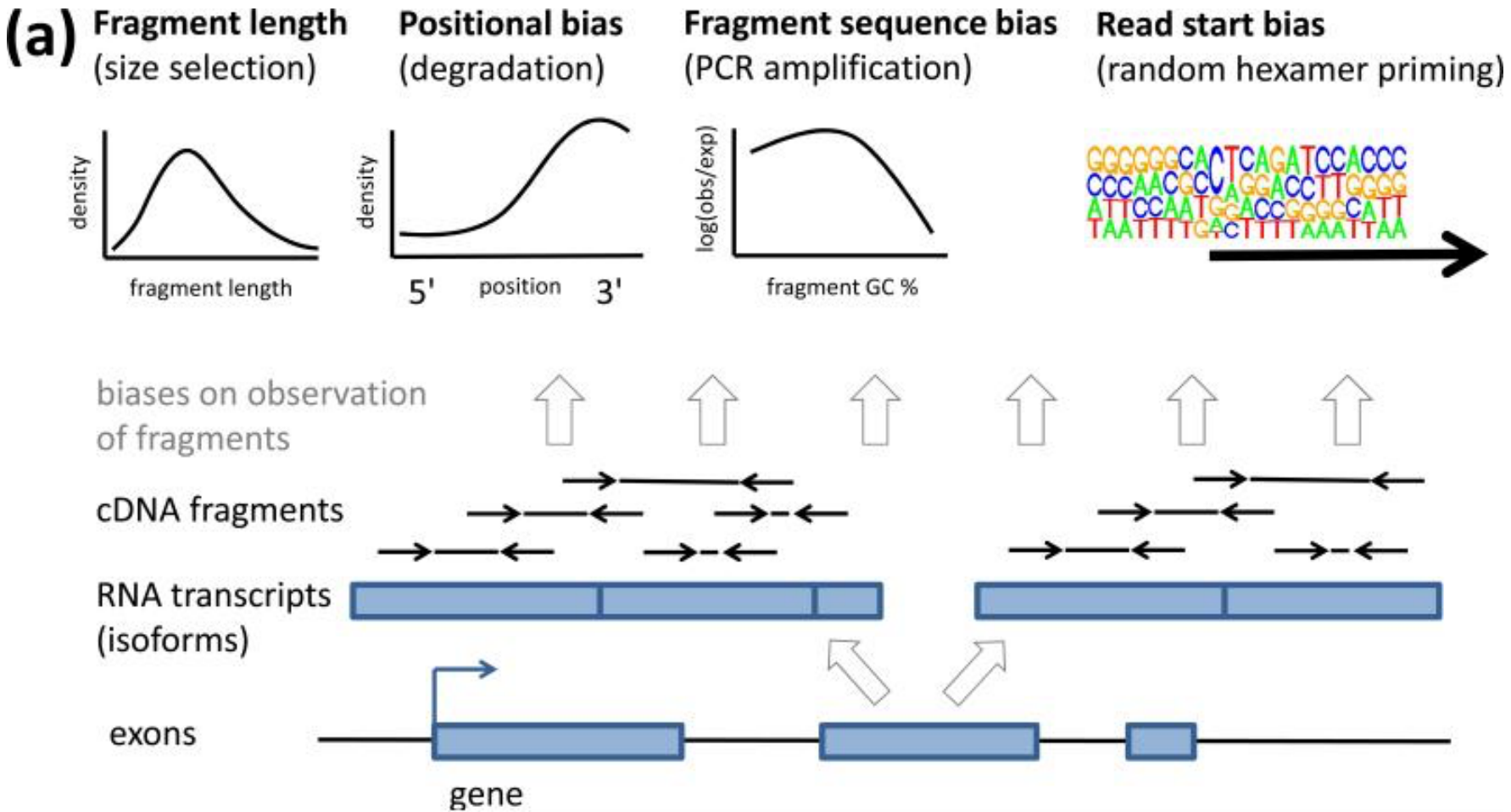
varying the dispersion

Salmon quantification

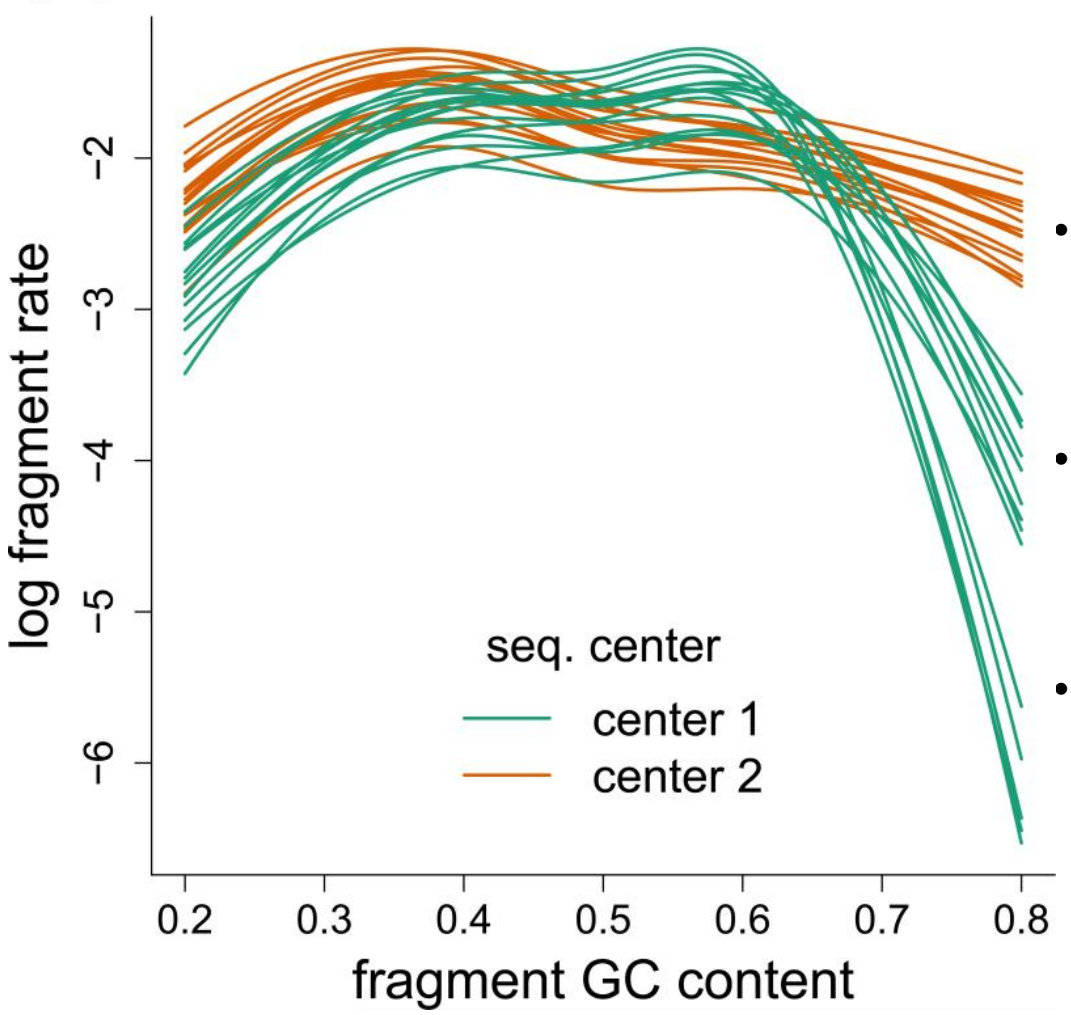
RNA-SEQ PART III

Biases estimated by Salmon

(default) **`--posBias --gcBias --seqBias`**



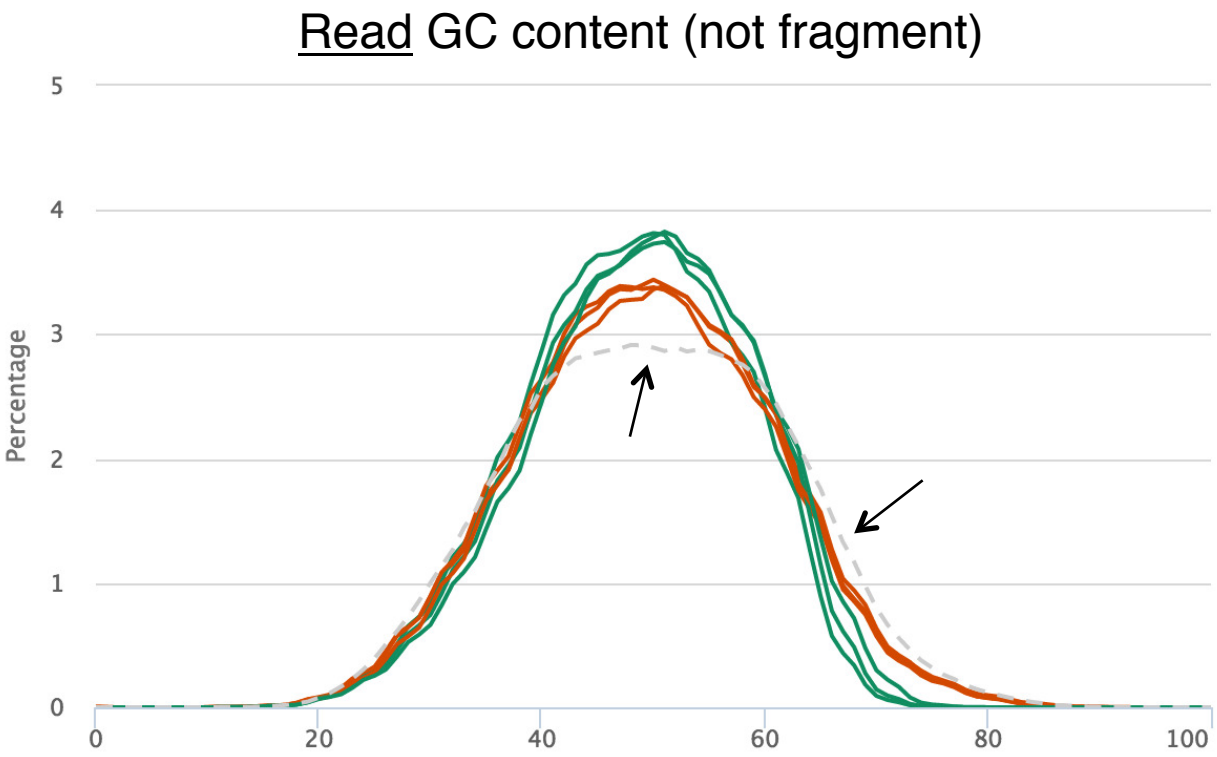
Fragment sequence bias



- Shown is *after* removing "hexamer bias" / "sequence bias"
- Can be attributed to differences in PCR amplification
- Sample- and batch-specific in comparison to "sequence bias" which doesn't vary much across samples

Love, Hogenesch, & Irizarry, "Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation" (2016) PMC5143225

We have a plugin for MultiQC



Theoretical GC Content

It is possible to plot a dashed line showing the theoretical GC content for a reference genome. MultiQC comes with genome and transcriptome guides for Human and Mouse. You can use these in your reports by adding the following MultiQC config keys (see Configuring MultiQC):

```
fastqc_config:
    fastqc_theoretical_gc: 'hg38_genome'
```

Only one theoretical distribution can be plotted. The following guides are available: `hg38_genome`, `hg38_txome`, `mm10_genome`, `mm10_txome` (txome = transcriptome).

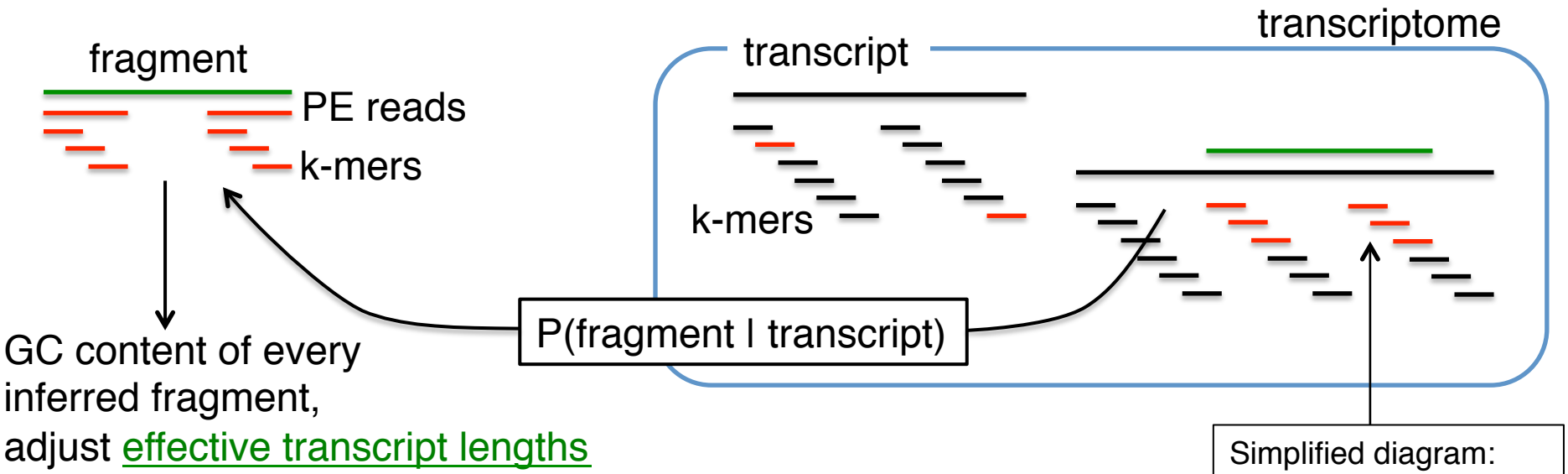
Lightweight quantifiers

Sailfish: Patro et al (2014), kallisto: Bray et al (2016), Salmon: Patro et al (2017)

Salmon maps reads to transcriptome with **RapMap:** Srivastava *et al* (2016)

Elements of mapping algorithm use ideas from **kallisto:** Bray *et al* (2016):
(1) skipping ahead to find Next Informative Position (NIP) and
(2) defining the consensus as ∩ of transcript sets from all hits

Salmon maps all PE reads, useful for estimating bias



~3 min for 45 mil PE reads (6 threads)
~5 min with --gcBias

Steps for running Salmon