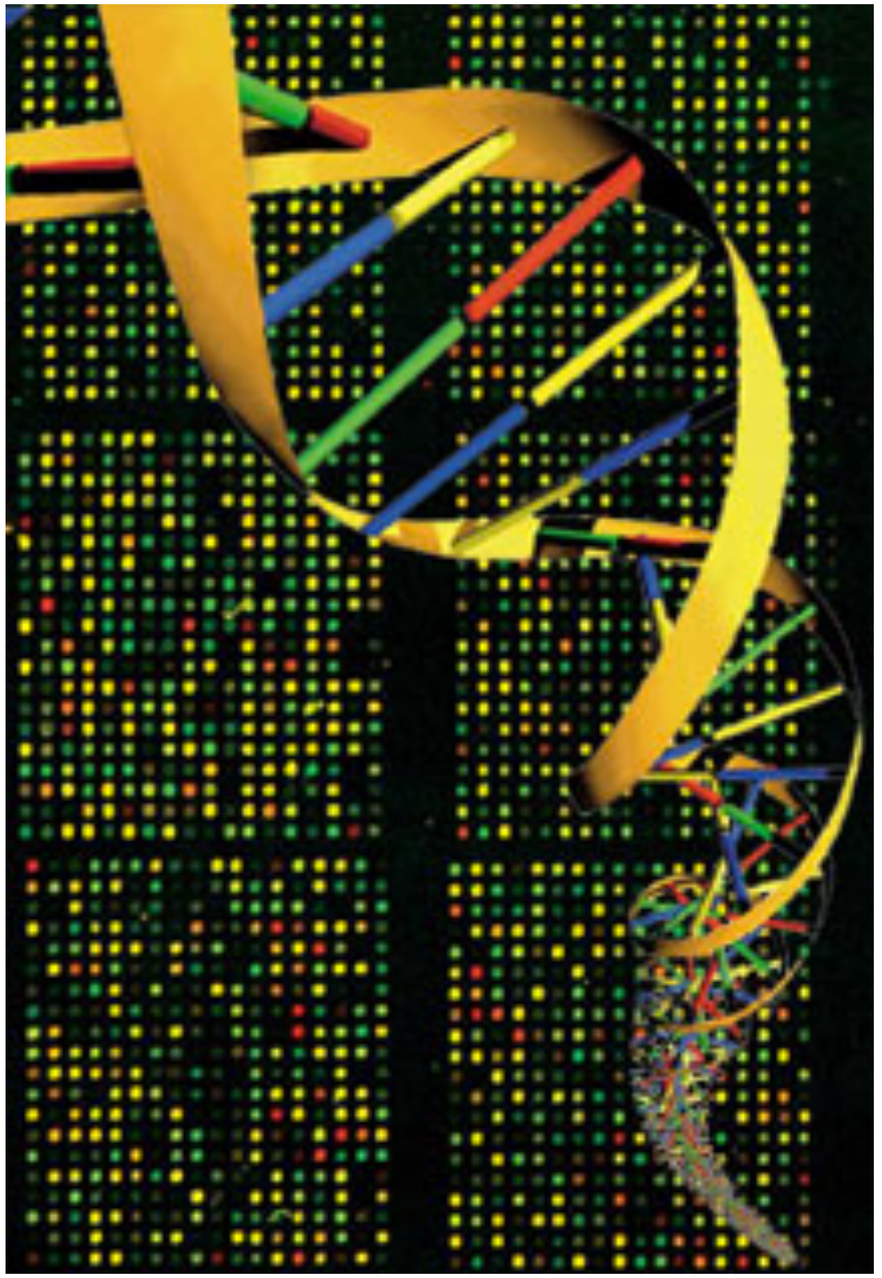
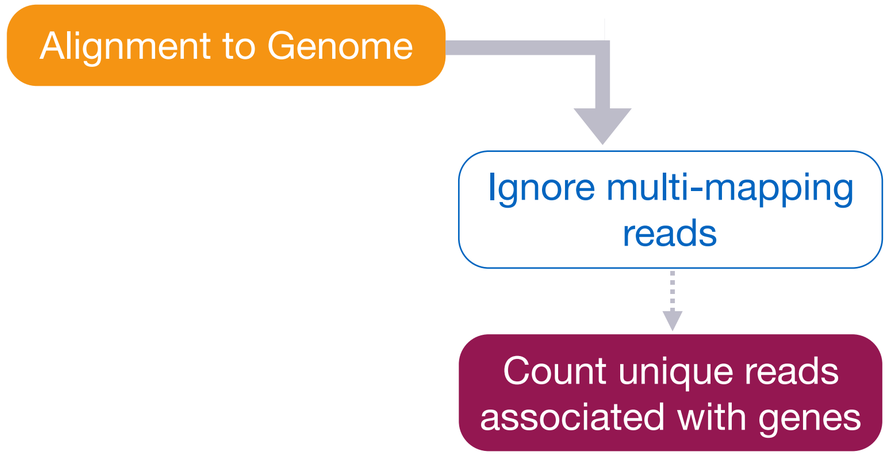
RNA-seq statistical analysis and gene-level differential expression

Transcriptomics

- to discover functional patterns of biological response to conditions of interest (treatments, environmental influences, mutations etc.)




Alignment to Genome
Ignore multi-mapping reads
Count unique reads associated with genes

High-throughput sequencing data

	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

High-throughput sequencing data

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

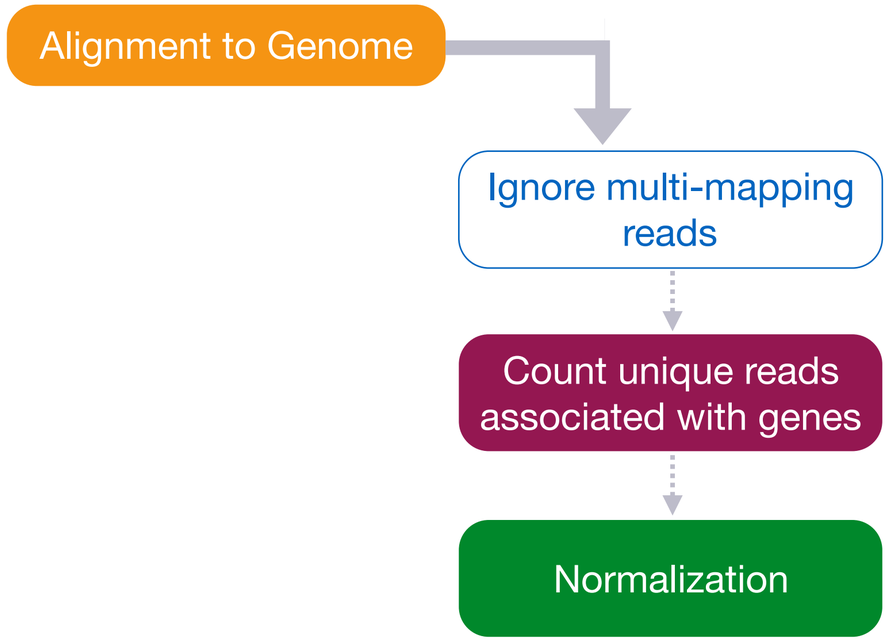
High-throughput sequencing data

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

features (e.g. genes)

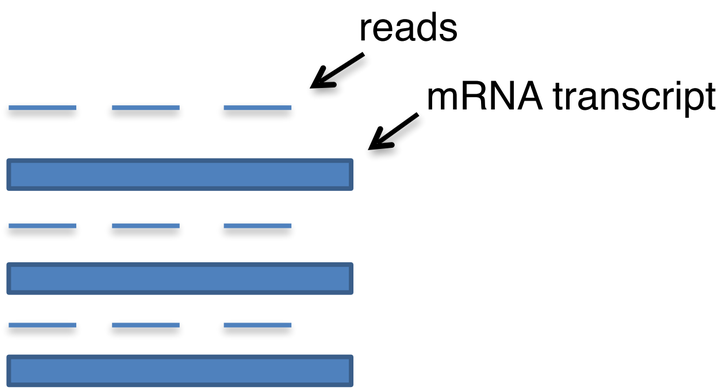
	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516
ENSG00000000003	679	448	873	408	1138
ENSG00000000005	0	0	0	0	0
ENSG00000000419	467	515	621	365	587
ENSG00000000457	260	211	263	164	245
ENSG00000000460	60	55	40	35	78

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



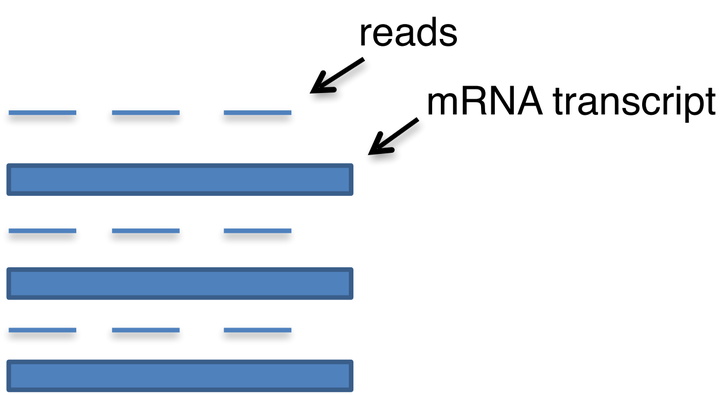
DE workflow :: normalization

mRNAs to reads



Slide adapted from MI Love: RNA-seq statistical analysis

mRNAs to reads

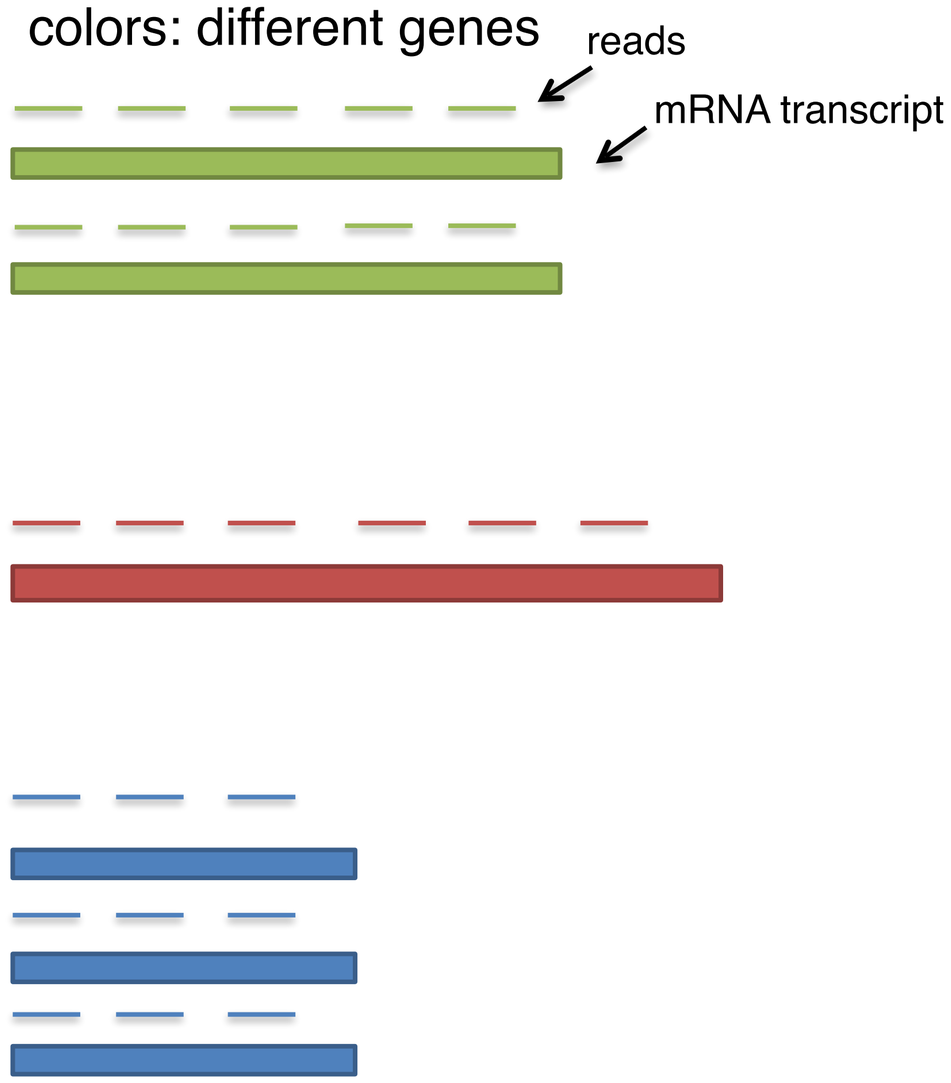


count of mapped reads proportional to:

- expression of RNA
- length of gene
- sequencing depth
- library prep. factors (PCR)
- etc...

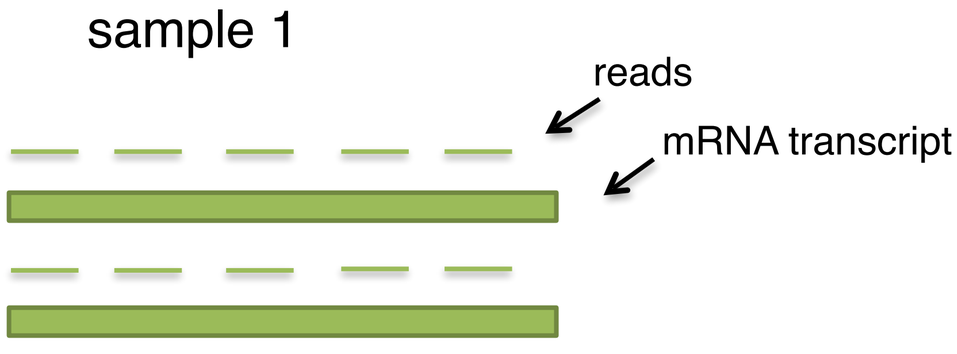
Slide adapted from MI Love: RNA-seq statistical analysis

Length of gene

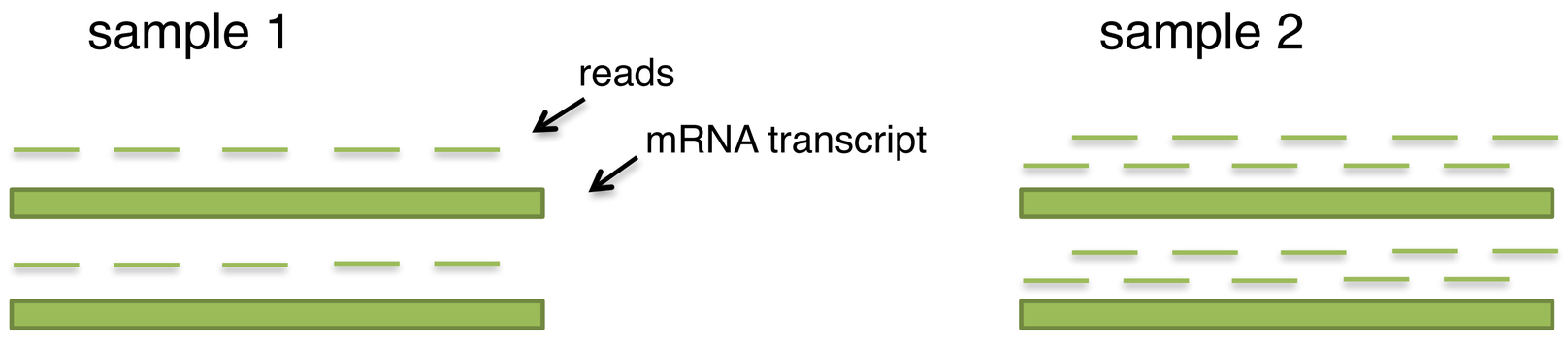


Slide adapted from MI Love: RNA-seq statistical analysis

Sequencing depth



Sequencing depth

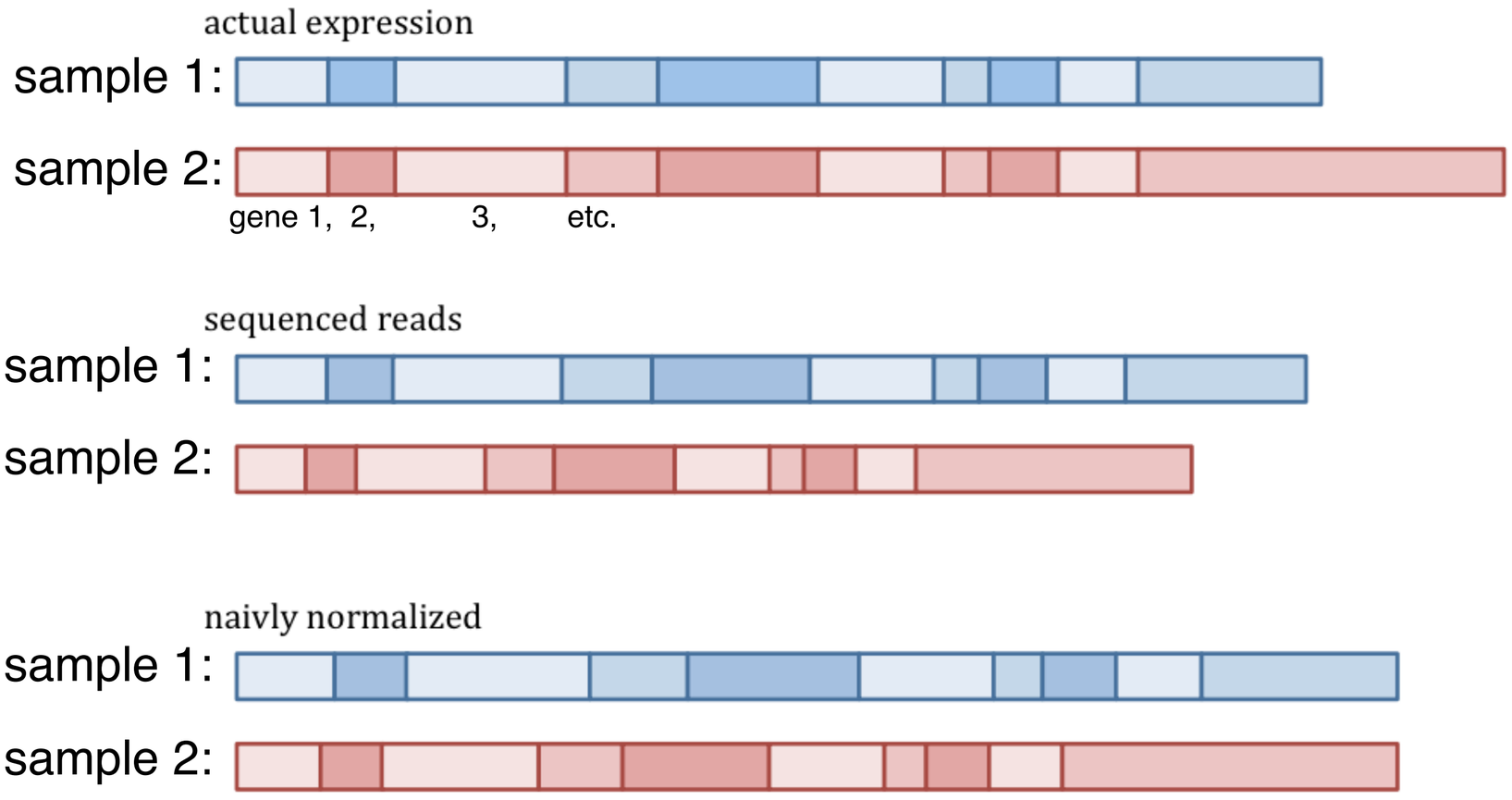


(Pre)-scaled measures of expression

- CPM (counts per million) counts scaled by the total number of reads
- TPM (transcripts per million) the proportion of transcripts in your RNA
- RPKM (reads aligned per kilobase of exon per million reads mapped) – Mortazavi et al 2008
- FPKM (fragments per kilobase of exon per million fragments mapped). Same idea for paired end sequencing

Actual counts contain information useful for statistical modeling, and should be used as input for differential expression tools

Need to have a robust estimator for sequencing depth

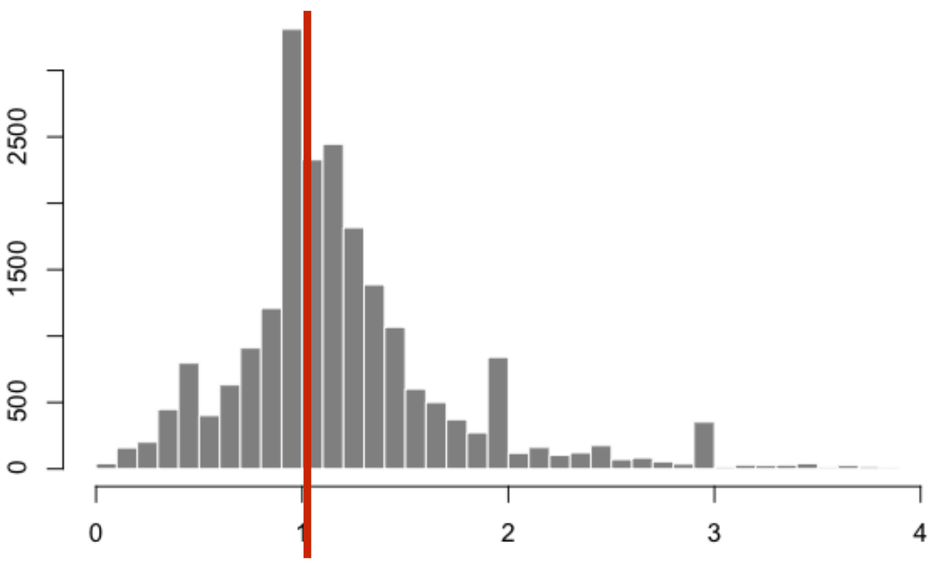


(slide from Simon Anders)

Median of ratios method

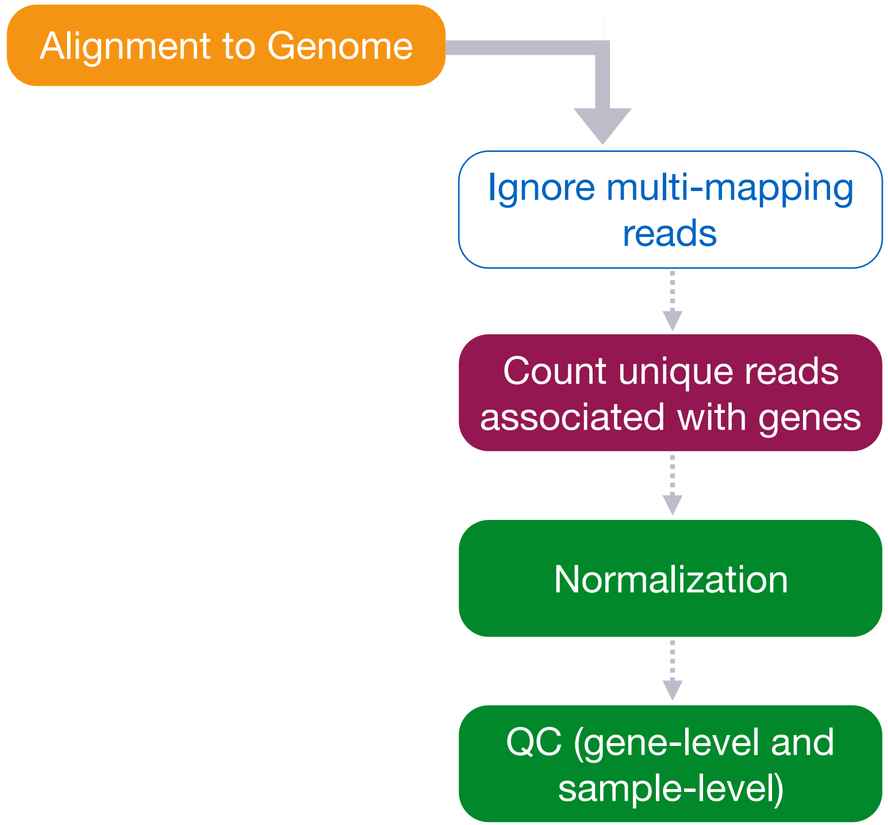
simple approach & works well
for each gene look at the count ratios:

sample 1 / pseudo-reference sample



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

Slide adapted from MI Love: RNA-seq statistical analysis



DE workflow :: quality control

QC: Gene-level filtering

	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

QC: Gene-level filtering

	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

Genes with zero counts

QC: Gene-level filtering

	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

Genes with extreme count outlier

Genes with zero counts

QC: Gene-level filtering

	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

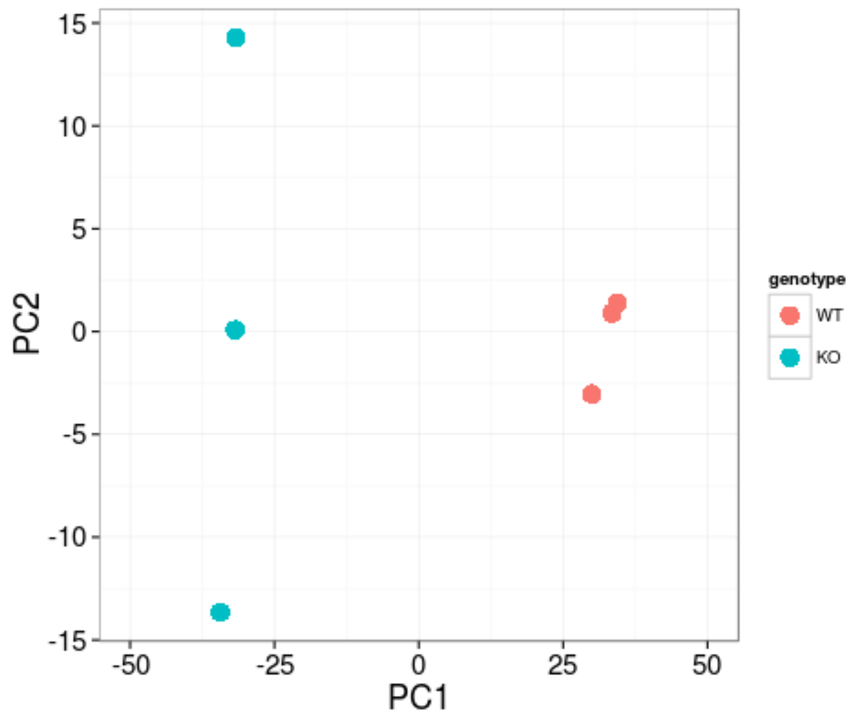
Genes with extreme count outlier

Genes with zero counts

Genes with low mean normalized counts (‘Independent filtering’)

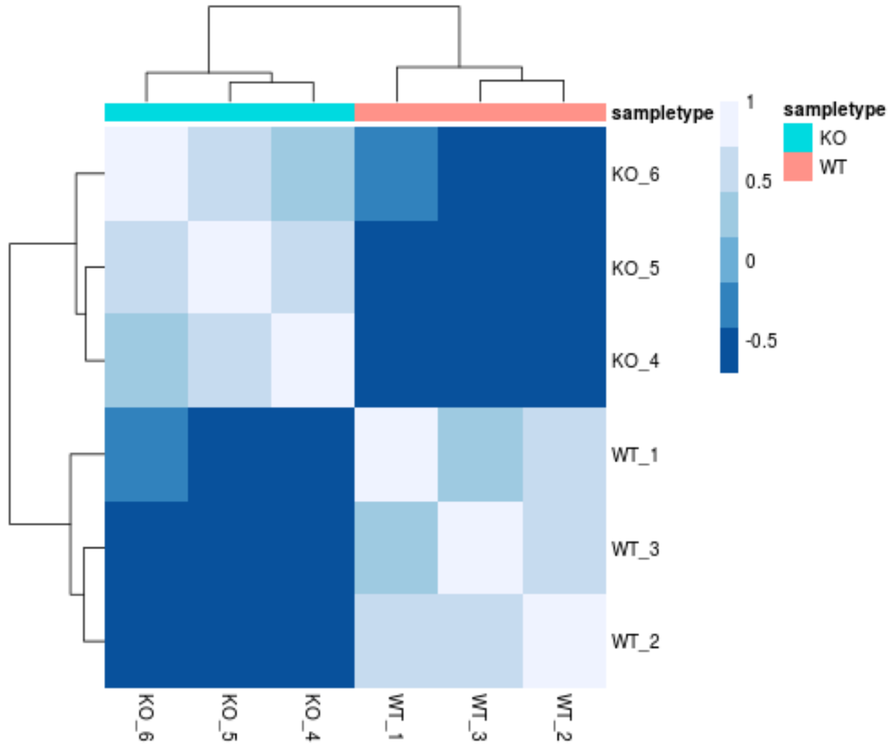
QC: Sample-level

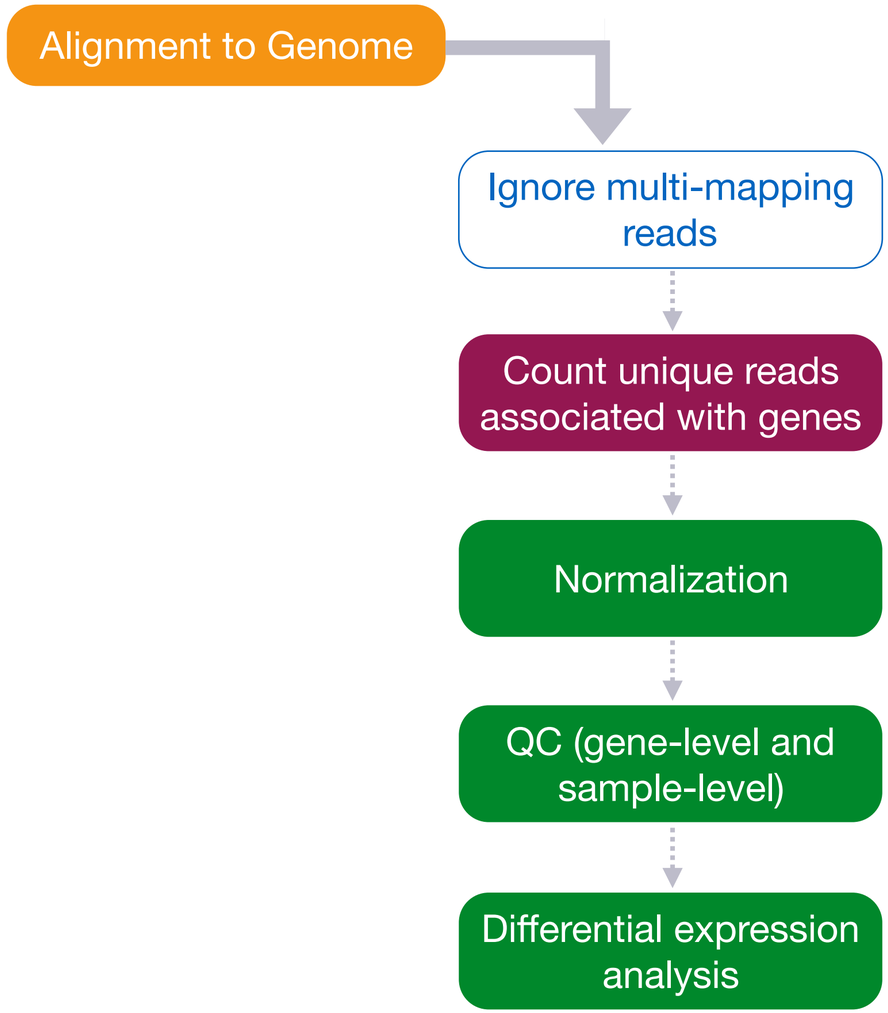
- **Principal Component Analysis (PCA):** A technique used to emphasize variation and bring out strong patterns in a dataset (dimensionality reduction)
- Project a line through the data points in n dimensional space (n = genes)
- Measure how much variance there is from that line (the distance from each point to the line).
- PC1 explains highest variance, PC2 next highest etc.



QC: Sample-level

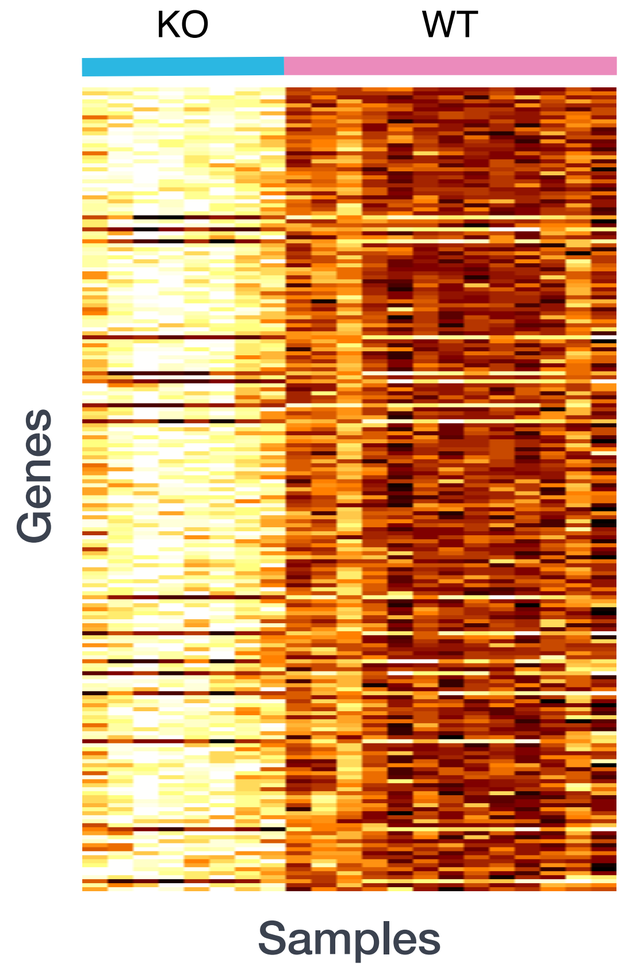
- Identify strong patterns in a dataset and potential outliers
- Correlation or distances for all pairwise combinations of samples.
- Generally high correlations with each other (values higher than 0.80)
- ‘Blocks’ indicate substructure in the data





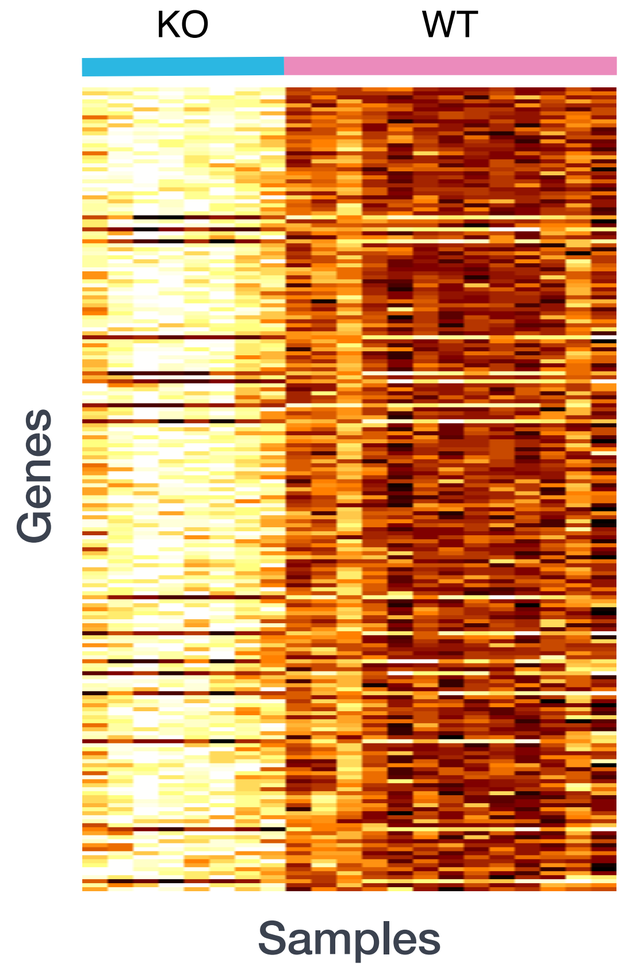
DE workflow :: differential expression

Identifying differences in gene expression



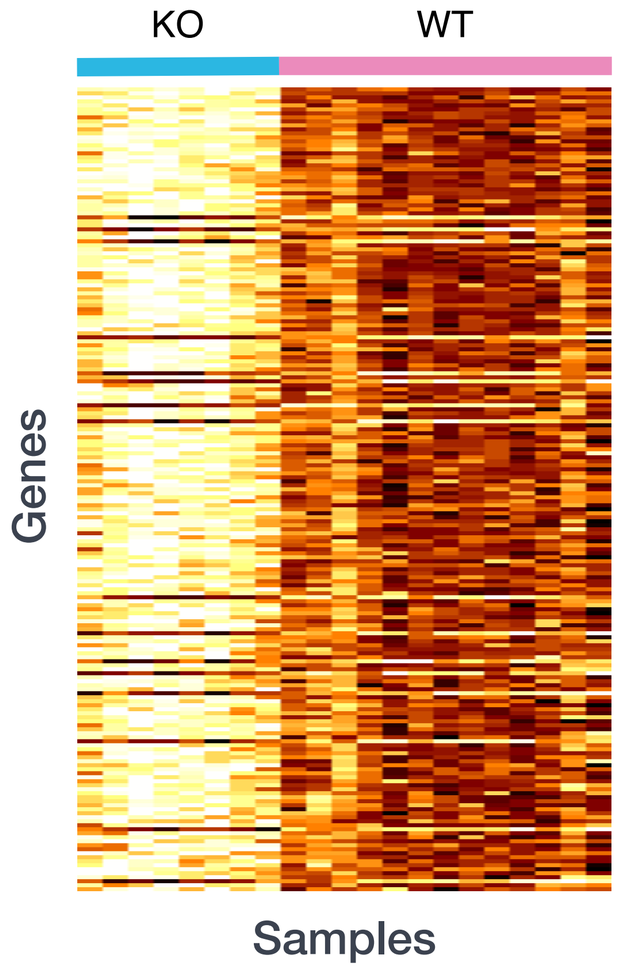
Identifying differences in gene expression

- Looking for genes that change in expression between two or more groups
    - case vs. control
    - correlation of expression with some variable or clinical outcome

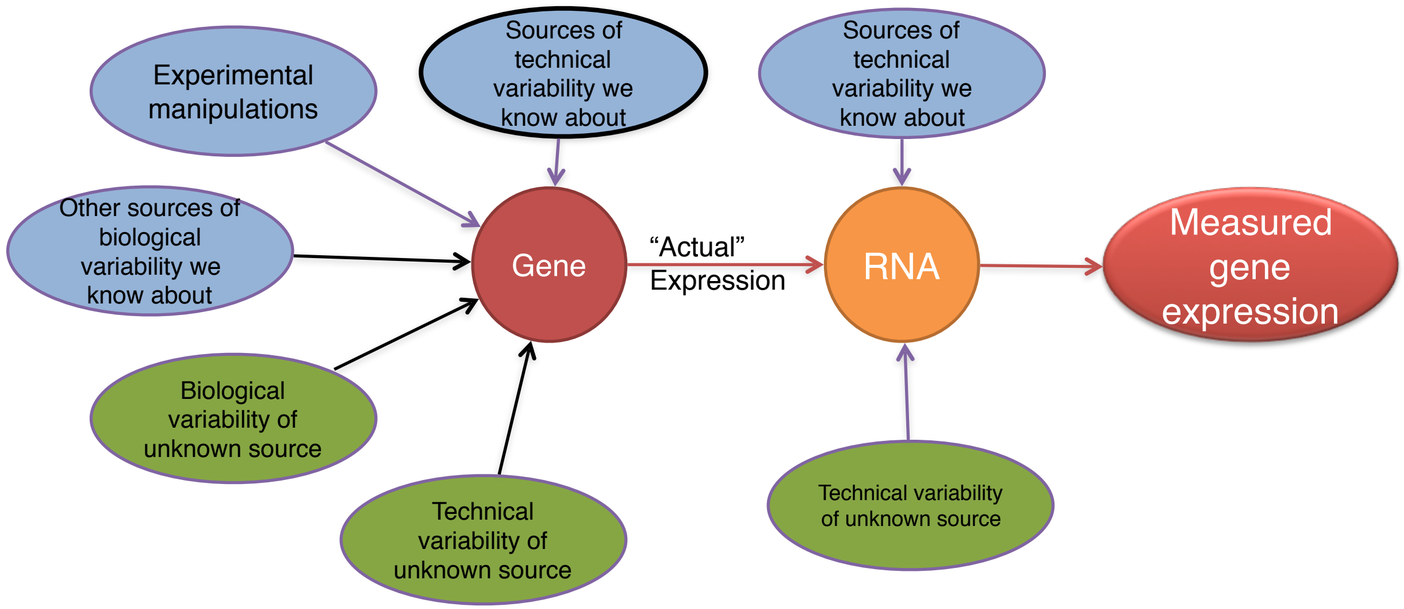


Identifying differences in gene expression

- Looking for genes that change in expression between two or more groups
    - case vs. control
    - correlation of expression with some variable or clinical outcome
- Rank the genes by how different they are between the two groups (based on fold change values. **Why does this not work?**



The measurement is the “sum” of many effects

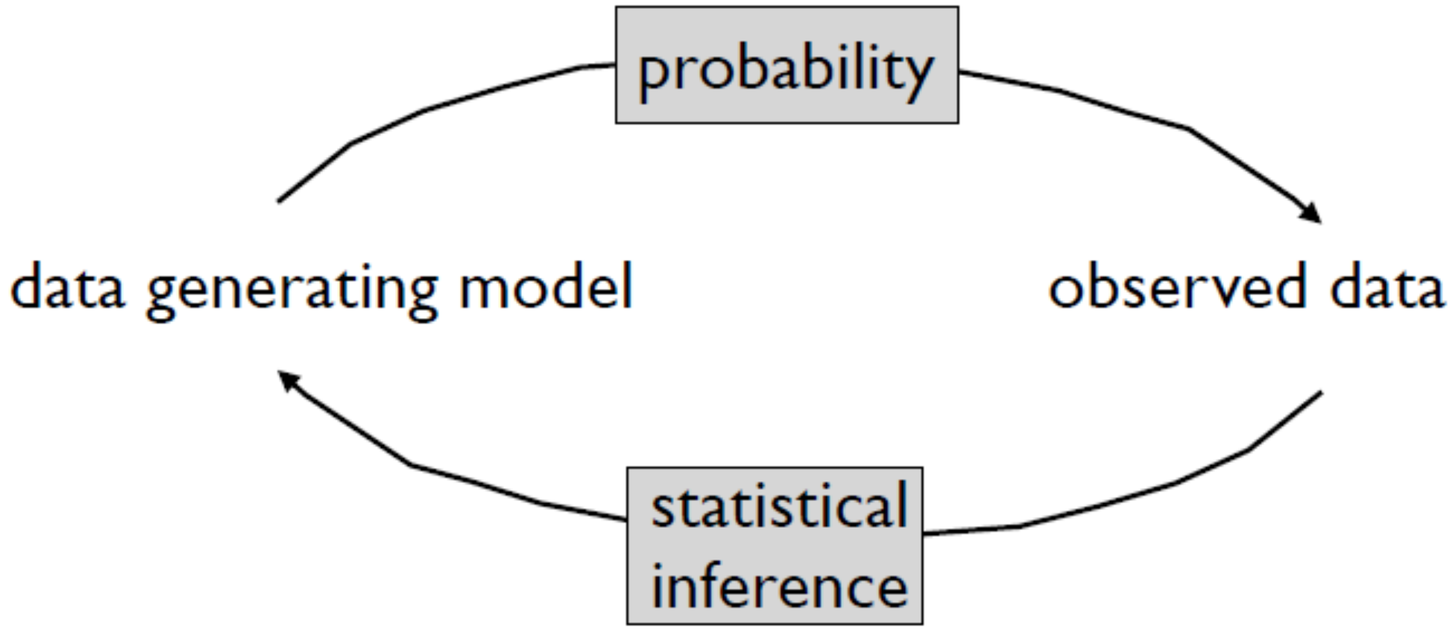


Modeling gene-level data

Courtesy of Paul Pavlidis, UBC

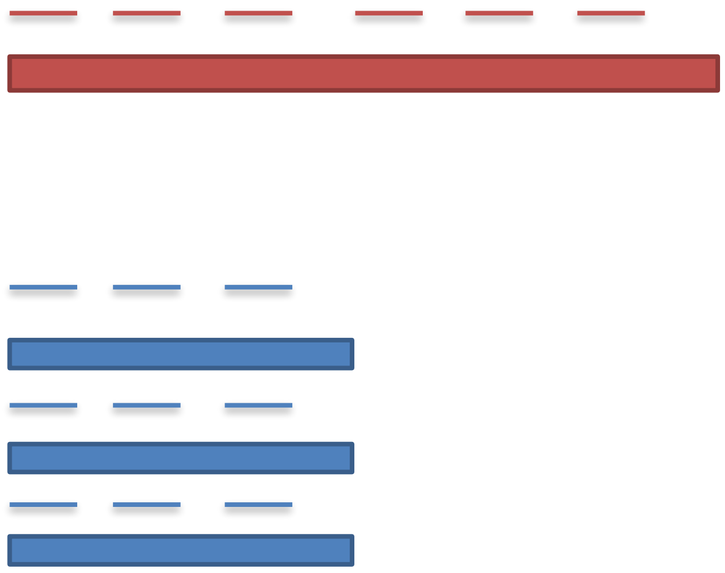
Making sense of data

Data is what we observe. We want to infer something about “where it came from”



Courtesy of Jenny Bryan, UBC

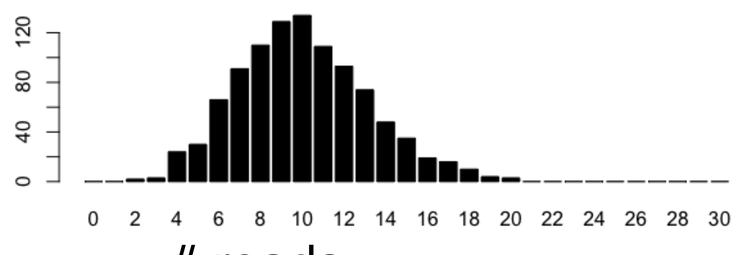
Variance of counts



Variance of counts

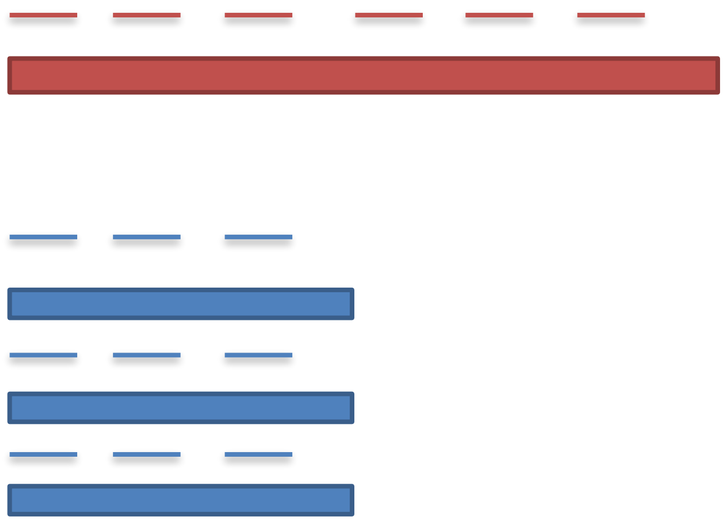
Consider one gene:

- Binomial sampling distribution

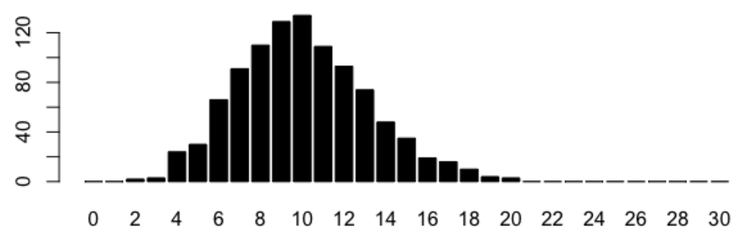


Variance of counts

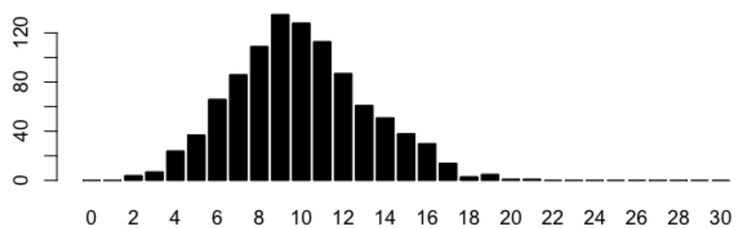
Consider one gene:



- Binomial sampling distribution



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



Technical vs biological replicates: modeling dispersion

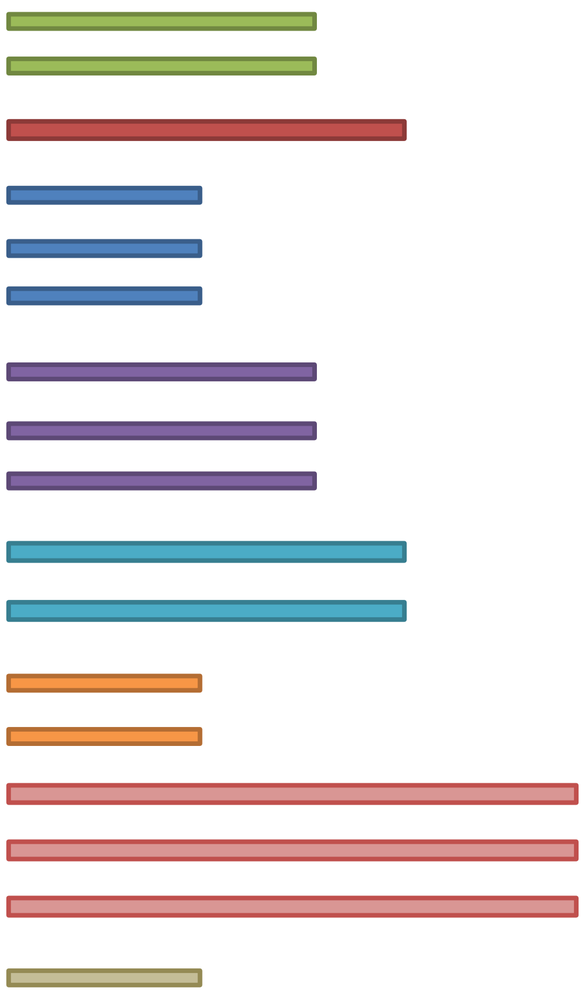
Biological vs. Technical Replicates

- **Biological replicates** contain multiple indviduals
  - **Technical replicates** contain one individual with some technical steps replicated

- Usually **biological variance > technical variance**, thus biological replicates are more useful. They also allow us to make inferences about treatment groups

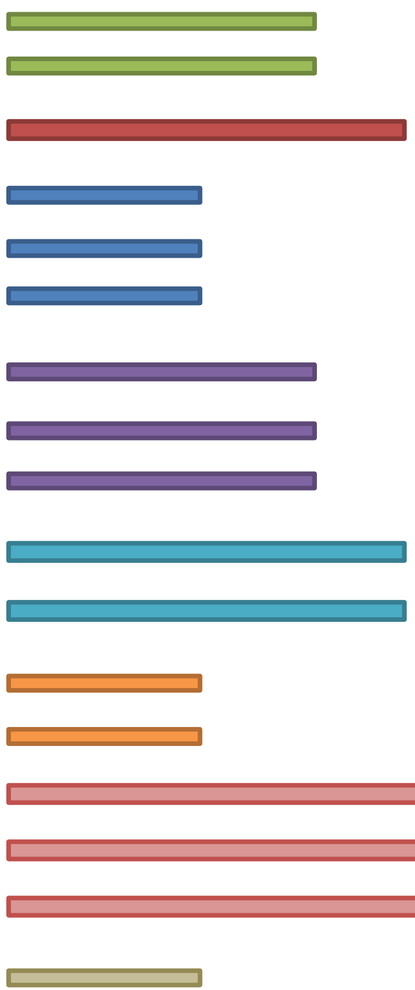
Biological replicates

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

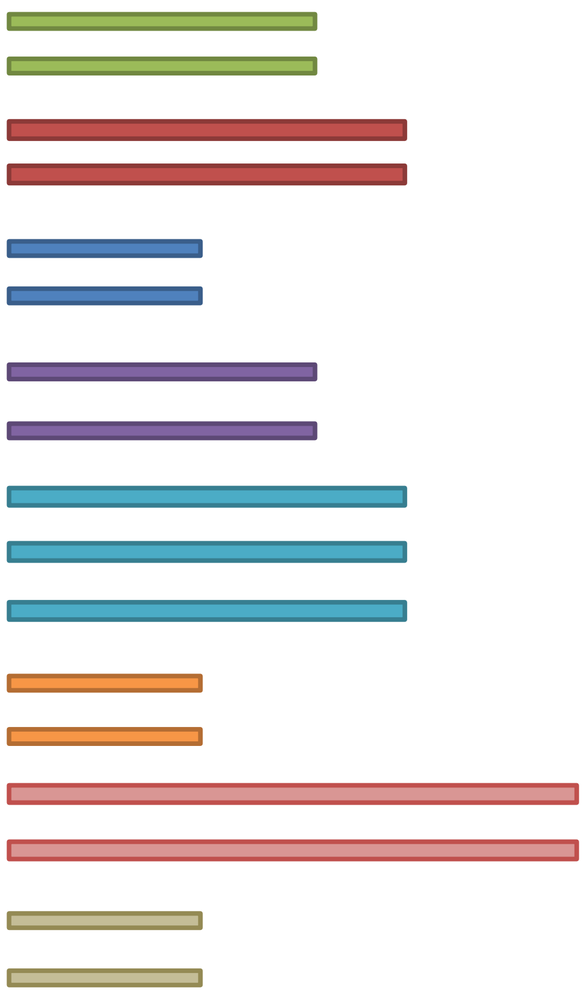


Biological replicates

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



But realistically, biological variation across sample units is expected



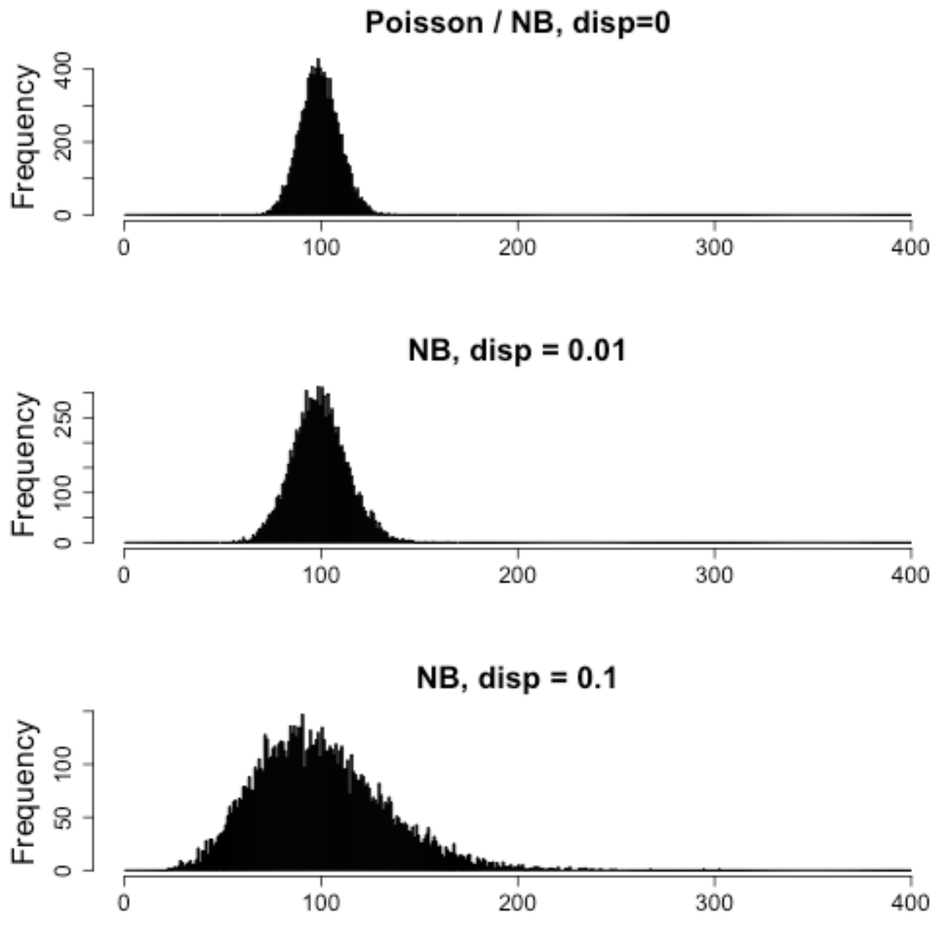
Biological replicates

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

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



Counts are modeled using the Negative Binomial (NB) distribution

raw count for gene i, sample j

The mean is taken as a quantity of interest scaled by a normalization factor

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

We model read counts K_{ij}, as following the negative binomial distribution with mean μ_i and **dispersion α_i**

Slide adapted from MI Love: RNA-seq statistical analysis

Counts are modeled using the Negative Binomial (NB) distribution

raw count for gene i, sample j

The mean is taken as a quantity of interest scaled by a normalization factor

one dispersion per gene

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

We model read counts K_{ij}, as following the negative binomial distribution with mean μ_i and **dispersion α_i**

Slide adapted from MI Love: RNA-seq statistical analysis

Within group variability is accounted for using the dispersion parameter

The **dispersion parameter** α_i, describes the variance of counts via:

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

Poisson part:
sampling fragments

Extra variation
due to biological variance

Dispersion is an approximation of the coefficient of variation

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

Dispersion is an approximation of the coefficient of variation

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

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

Dispersion is an approximation of the coefficient of variation

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

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

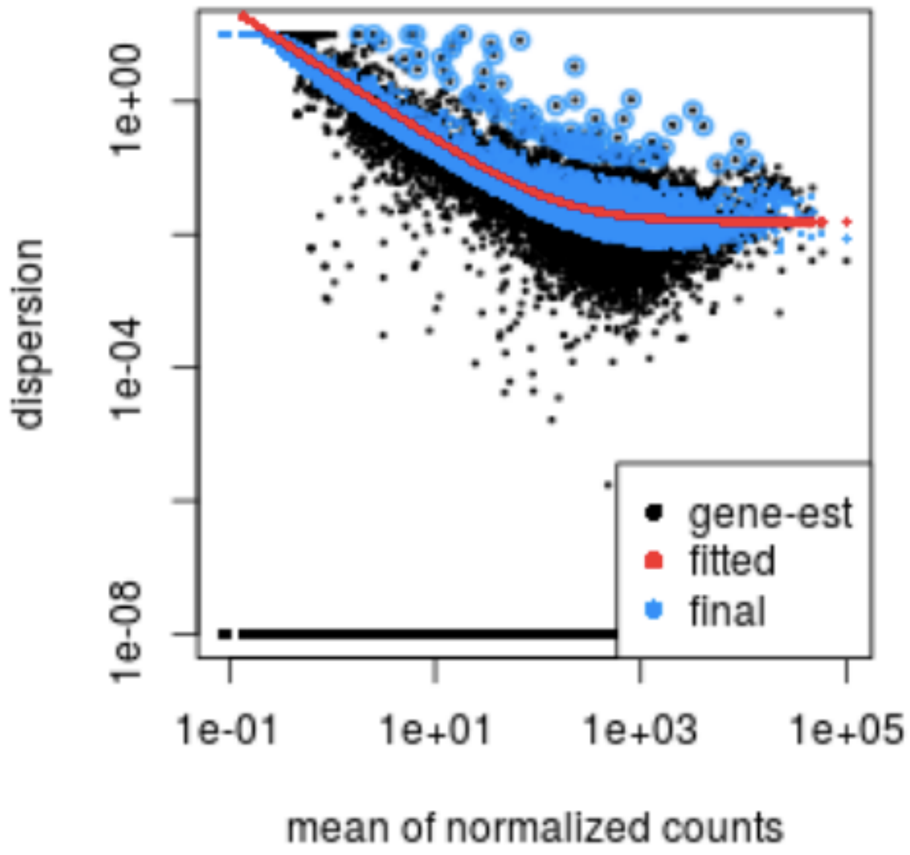
alpha = 0.01 => CV 10%
alpha = 0.25 => CV 50%

Shrinkage and dispersion

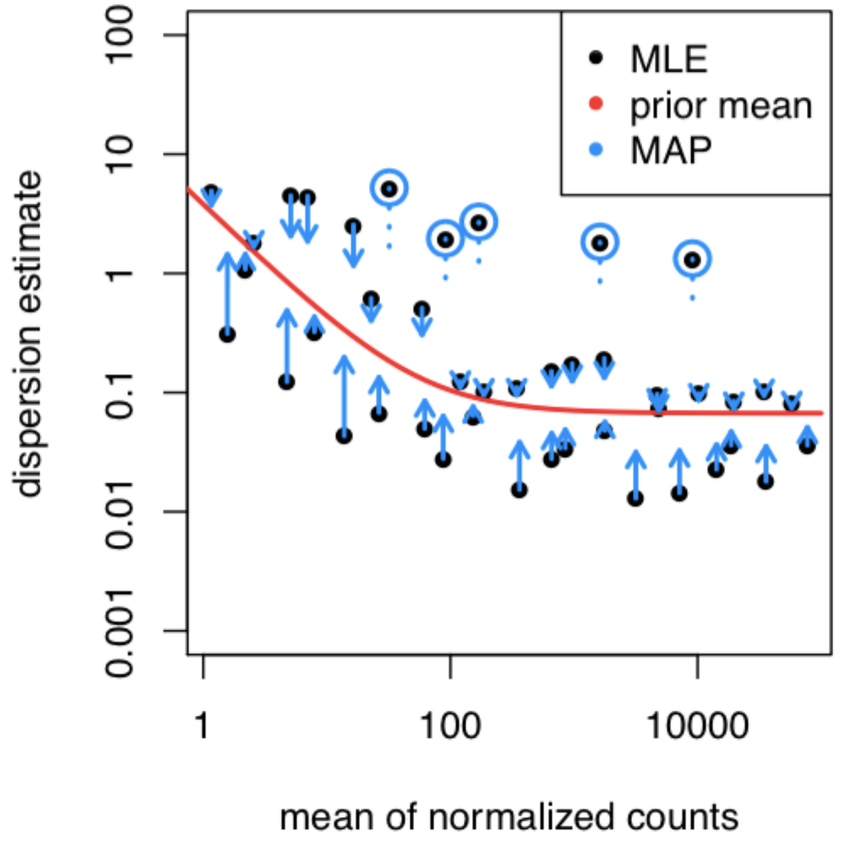
- Different genes naturally have different scale of biological variability
- Over all genes, there will be a distribution of reasonable estimates of dispersion
- With small sample size (n=3-5 replicates per group), we will make *very bad* estimates of gene-wise dispersion unless we share information across genes

Shrinkage of dispersion

all genes (Pasilla)



a subset of genes (Pickrell)



1. Gene-wise estimate = just look at one gene (MLE)
2. Fitted dispersion trend = the middle for the prior
3. Final estimate = posterior, uses shared information (MAP)

Differences across conditions

DESeq2

- Available through Bioconductor since 2013
- Publication: Genome Biology, Dec 2014.
  main text written with non-statisticians in mind
- Builds on good ideas for dispersion estimation and use of GLM from the DSS and edgeR methods
- See bioconductor.org/install for installation
- Note that the latest Bioconductor packages are only available with latest R version. Bioconductor and R versions are *linked*

DESeq2 steps

we tried to streamline the default pipeline.

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

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

Differences across two conditions

- Describe experiment with formula, e.g.: ~ condition
- Per gene the design matrix looks like:

$\log_2$ q1		1	0
$\log_2$ q2		1	0
$\log_2$ q3	=	1	1
$\log_2$ q4		1	1

Differences across two conditions

- Describe experiment with formula, e.g.: ~ condition
- Per gene the design matrix looks like:

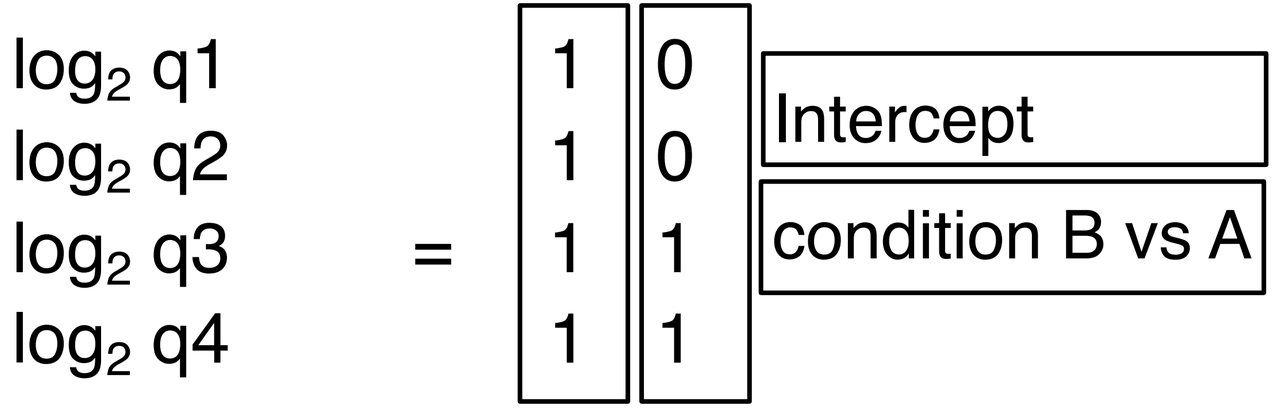
$$\begin{matrix} \log_2 q1 \\ \log_2 q2 \\ \log_2 q3 \\ \log_2 q4 \end{matrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Intercept

All samples get an Intercept term

Differences across two conditions

- Describe experiment with formula, e.g.: ~ condition
- Per gene the design matrix looks like:



All samples get an Intercept term

The B condition samples also get a term that accounts for the difference between B and A

Differences across multiple conditions

- The design matrix now uses a column for each condition

$\log_2$ q1		1	1	0	0	
$\log_2$ q2		1	1	0	0	Intercept
$\log_2$ q3	=	1	0	1	0	conditionA
$\log_2$ q4		1	0	1	0	conditionB
$\log_2$ q5		1	0	0	1	conditionC
$\log_2$ q6		1	0	0	1	

Slide adapted from MI Love: RNA-seq statistical analysis

Differences across multiple conditions

- The design matrix now uses a column for each condition

$$
\begin{matrix}
\log_2 q1 \\
\log_2 q2 \\
\log_2 q3 \\
\log_2 q4 \\
\log_2 q5 \\
\log_2 q6
\end{matrix}
=
\begin{bmatrix}
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1
\end{bmatrix}
\begin{matrix}
\\
\text{Intercept} \\
\text{conditionA} \\
\text{conditionB} \\
\text{conditionC} \\
\\
\end{matrix}
$$

Differences across multiple conditions

- The design matrix now uses a column for each condition

$$
\begin{array}{lcccc}
\log_2 q1 & & 1 & 1 & 0 & 0 \\
\log_2 q2 & & 1 & 1 & 0 & 0 \\
\log_2 q3 & = & 1 & 0 & 1 & 0 \\
\log_2 q4 & & 1 & 0 & 1 & 0 \\
\log_2 q5 & & 1 & 0 & 0 & 1 \\
\log_2 q6 & & 1 & 0 & 0 & 1
\end{array}
$$

Intercept
conditionA
conditionB
conditionC

Slide adapted from MI Love: RNA-seq statistical analysis

Differences across multiple conditions

- The design matrix now uses a column for each condition

$\log_2$ q1		1	1	0	0	
$\log_2$ q2		1	1	0	0	Intercept
$\log_2$ q3	=	1	0	1	0	conditionA
$\log_2$ q4		1	0	1	0	conditionB
$\log_2$ q5		1	0	0	1	conditionC
$\log_2$ q6		1	0	0	1	

Slide adapted from MI Love: RNA-seq statistical analysis

Differences across multiple conditions

- The design matrix now uses a column for each condition

$\log_2$ q1		1	1	0	0	
$\log_2$ q2		1	1	0	0	Intercept
$\log_2$ q3	=	1	0	1	0	conditionA
$\log_2$ q4		1	0	1	0	conditionB
$\log_2$ q5		1	0	0	1	conditionC
$\log_2$ q6		1	0	0	1	

Slide adapted from MI Love: RNA-seq statistical analysis

Differences across multiple conditions

- The design matrix now uses a column for each condition

$\log_2$ q1		1	1	0	0	
$\log_2$ q2		1	1	0	0	Intercept
$\log_2$ q3	=	1	0	1	0	conditionA
$\log_2$ q4		1	0	1	0	conditionB
$\log_2$ q5		1	0	0	1	conditionC
$\log_2$ q6		1	0	0	1	

Slide adapted from MI Love: RNA-seq statistical analysis

GLM: Generalized linear models

- Extension of linear models to non-normally distributed response data (in our case, negative binomial)
- Helps address the different mean-variance relationships
- GLM fit for a gene will return **coefficients indicating the overall expression** strength of the gene for each design matrix element

Contrasts

- A contrast is the comparison of coefficients we choose to evaluate

$$\begin{array}{lcccc} \log_2 q1 & & 1 & 1 & 0 & 0 \\ \log_2 q2 & & 1 & 1 & 0 & 0 \\ \log_2 q3 & = & 1 & 0 & 1 & 0 \\ \log_2 q4 & & 1 & 0 & 1 & 0 \\ \log_2 q5 & & 1 & 0 & 0 & 1 \\ \log_2 q6 & & 1 & 0 & 0 & 1 \end{array}$$

Intercept
conditionA
conditionB
conditionC

Contrasts

- A contrast is the comparison of coefficients we choose to evaluate

$\log_2$ q1		1	1	0	0	
$\log_2$ q2		1	1	0	0	Intercept
$\log_2$ q3	=	1	0	1	0	conditionA
$\log_2$ q4		1	0	1	0	conditionB
$\log_2$ q5		1	0	0	1	conditionC
$\log_2$ q6		1	0	0	1	

Contrasts

- A contrast is the comparison of coefficients we choose to evaluate

$\log_2$ q1		1	1	0	0	
$\log_2$ q2		1	1	0	0	Intercept
$\log_2$ q3	=	1	0	1	0	conditionA
$\log_2$ q4		1	0	1	0	conditionB
$\log_2$ q5		1	0	0	1	conditionC
$\log_2$ q6		1	0	0	1	

Contrasts

- A contrast is the comparison of coefficients we choose to evaluate

$\log_2$ q1		1	1	0	0	Intercept
$\log_2$ q2		1	1	0	0	
$\log_2$ q3	=	1	0	1	0	conditionA
$\log_2$ q4		1	0	1	0	conditionB
$\log_2$ q5		1	0	0	1	conditionC
$\log_2$ q6		1	0	0	1	

```
results(dds, contrast=c("condition","B","C"))
```

Hypothesis testing

Wald test

- use the shrunken estimate of the log2 fold-change divided by the SE
- this gives a Z-statistic which is compared to a standard normal distribution
- allows testing of individual coefficients, or contrasts of coefficients (i.e. two-level comparison)

Likelihood ratio test (LRT)

- examines two different models: full and reduced model (with some terms removed)
- determines if the increased likelihood of the data using the extra terms in the full model is more than expected if those extra terms are truly zero
- useful for identifying any gene that is changing in expression with respect to the biological factor of interest (useful for 3 or more levels)

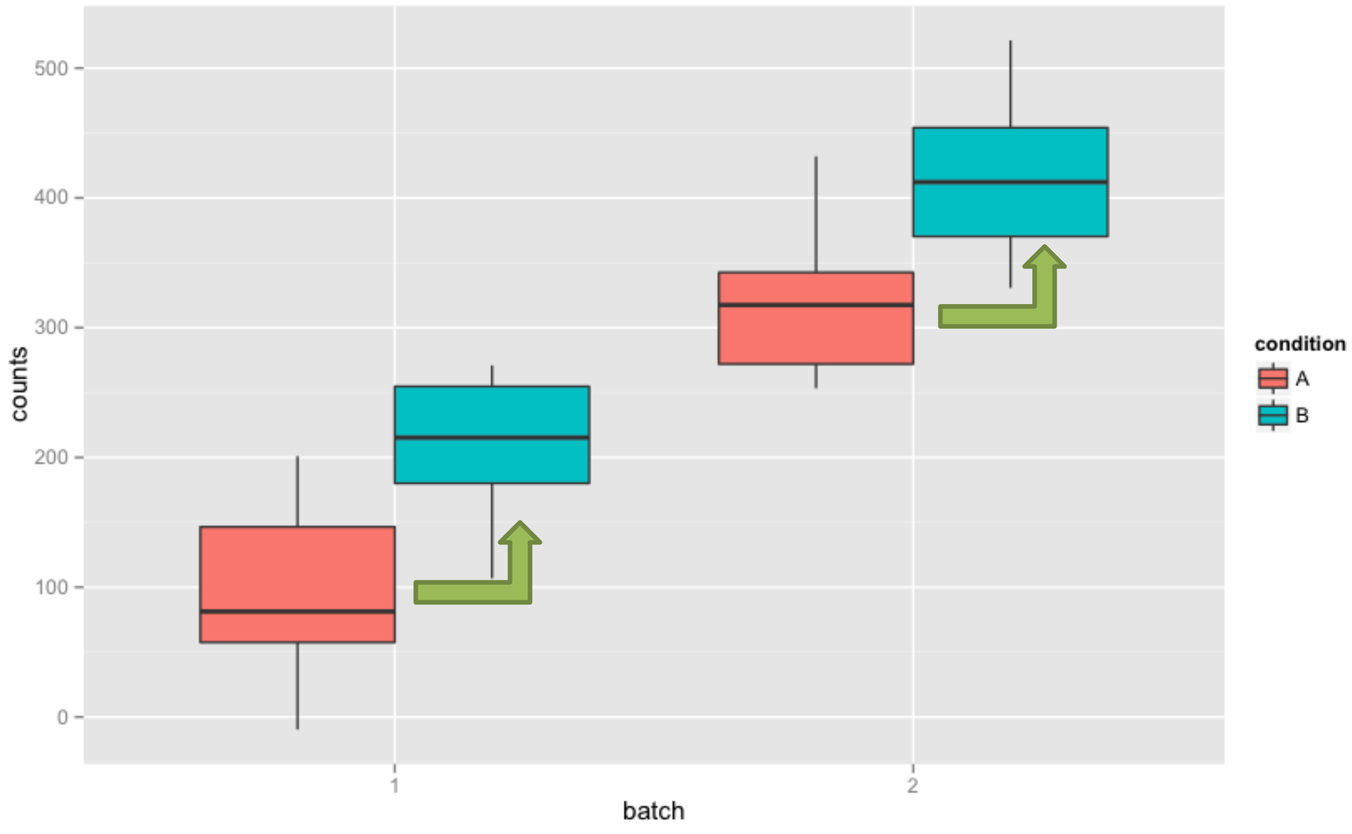
Controlling for different batches

Controlling for different batches

- Using a design formula: ~ batch + condition,
  adds terms that control for batch differences

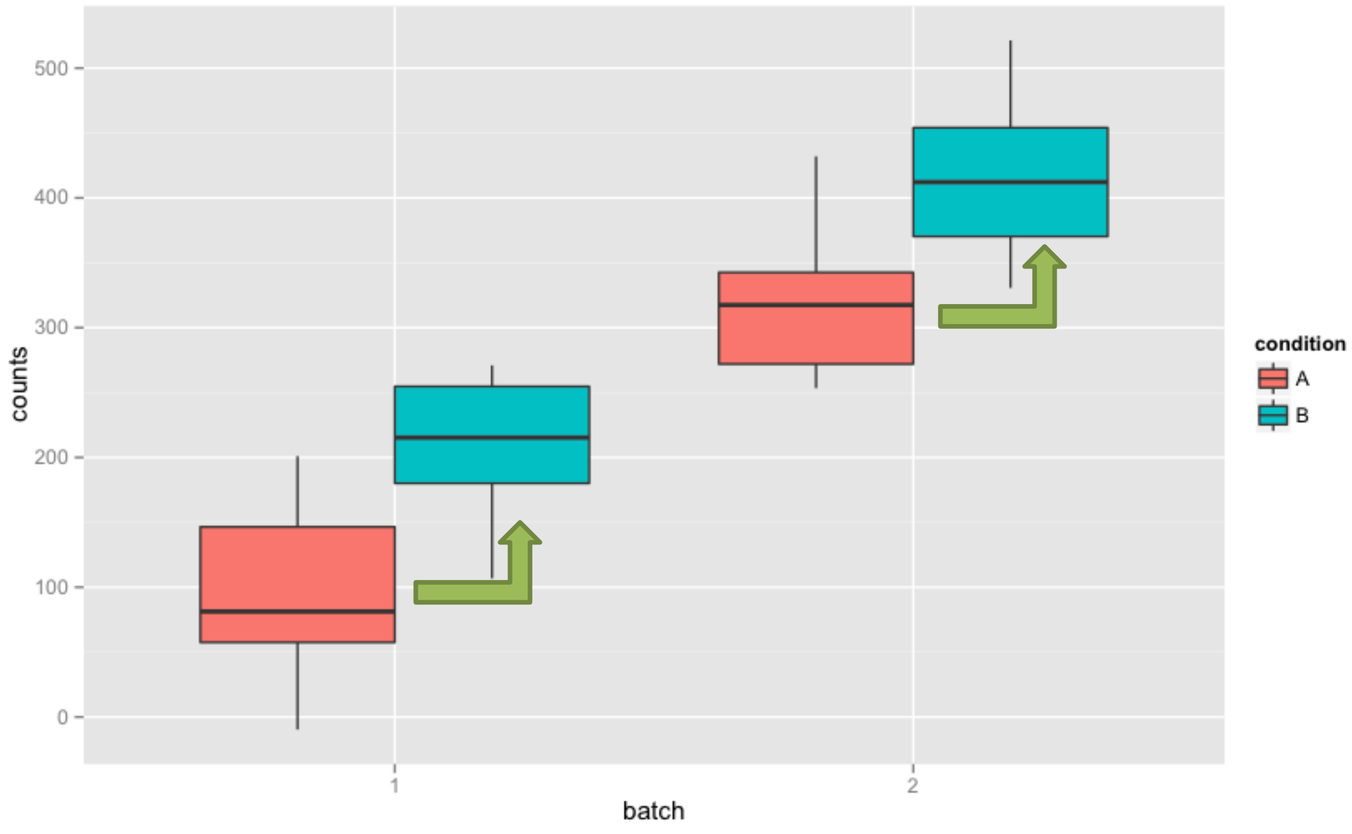
Controlling for different batches

- Using a design formula: ~ batch + condition,
  adds terms that control for batch differences



Controlling for different batches

- Using a design formula: ~ batch + condition,
  adds terms that control for batch differences
- If batches are unknown, possible to detect these with other methods: svaseq, RUVSeq



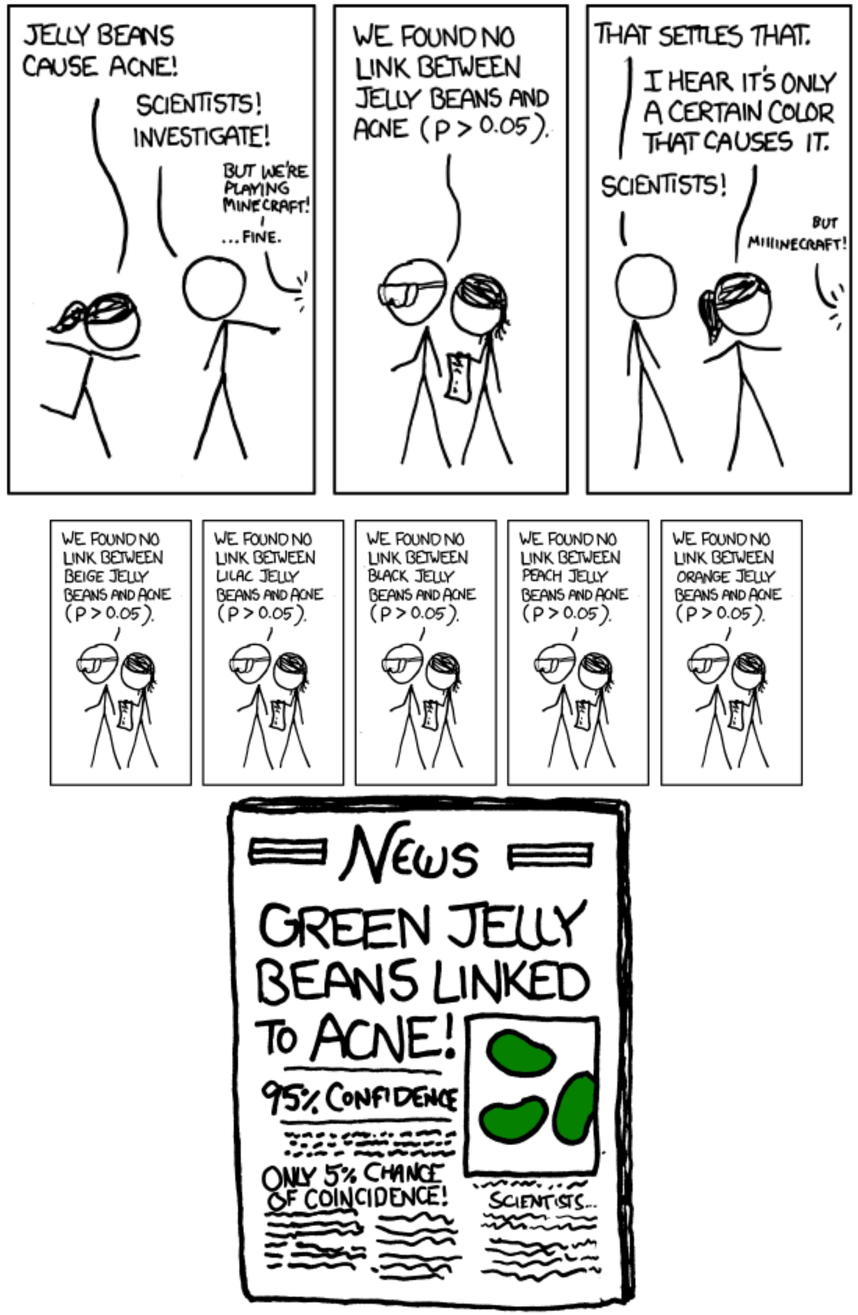
Complex designs

Want to test: treatment changes for enriched samples over baseline, controlling for individual effects

~individual + enrichment + treatment + enrichment:treatment

indiv	.	enrich.	treat.
1		input	control
1		IP	control
1		input	treat
1		IP	treat
2		input	control
2		IP	control
2		input	treat
2		IP	treat
...			

Multiple test correction



https://xkcd.com/882/

Multiple test correction

Multiple test correction

- In our example using condition A vs. condition B we found 2,898 genes to be differentially expressed at $p < 0.05$

Multiple test correction

- In our example using condition A vs. condition B we found 2,898 genes to be differentially expressed at $p < 0.05$
- This is 12% of the genes we tested

Multiple test correction

- In our example using condition A vs. condition B we found 2,898 genes to be differentially expressed at $p < 0.05$
- This is 12% of the genes we tested
- Expect 5% by chance (this is what $p < 0.05$)

Multiple test correction

- In our example using condition A vs. condition B we found 2,898 genes to be differentially expressed at $p < 0.05$
- This is 12% of the genes we tested
- Expect 5% by chance (this is what $p < 0.05$)
- Probably ~1/3 genes we found are false positives

Multiple test correction

- In our example using condition A vs. condition B we found 2,898 genes to be differentially expressed at $p < 0.05$
- This is 12% of the genes we tested
- Expect 5% by chance (this is what $p < 0.05$)
- Probably ~1/3 genes we found are false positives

This is the multiple testing problem. The more tests we perform, the more we inflate the number of false positives observed.

Multiple test correction

Controlling the FWER

- Control α, the probability of making an error (false positive)
- **Bonferroni:** Reject any hypothesis with p-value ≤ α/m
    - Conservative; high probability of false negatives

Controlling the FDR

- FDR: false discovery rate: the expected percent of false predictions in the set of predictions
- **Benjamini-Hochberg:** Rank *j / m* multiplied by the FDR level
    - designed to control the FDR
- **Q-value:** The minimum FDR that can be attained when calling that feature significant

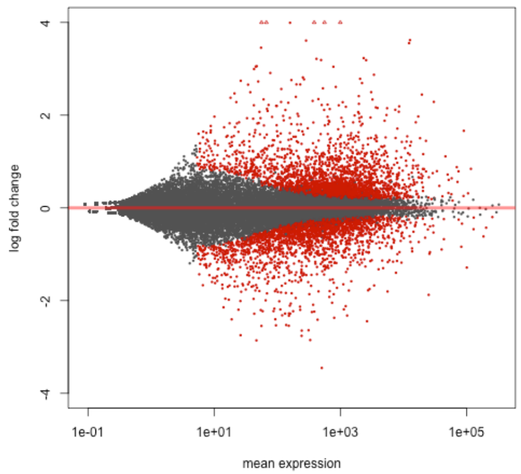
DE vs EDA

Two paths in DESeq2

Count matrix

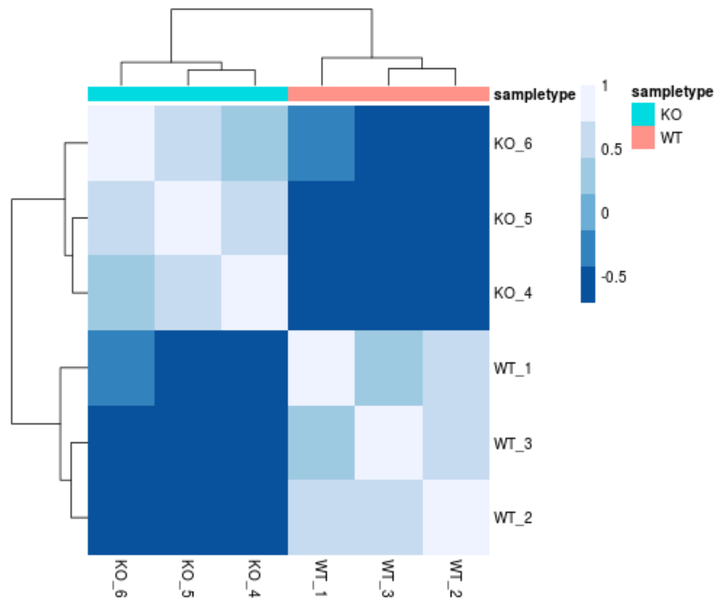
Differential expression

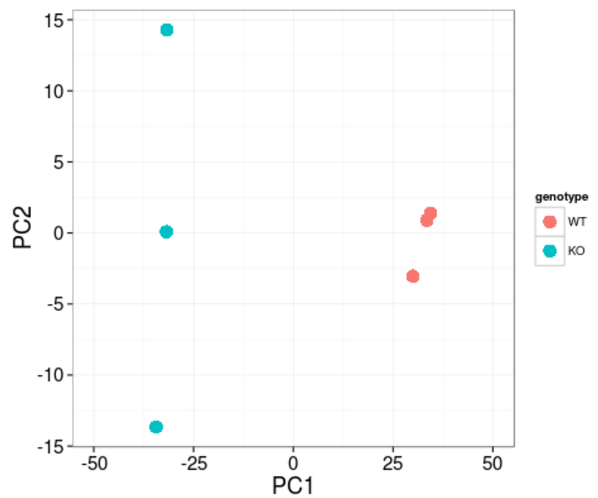
testing, p-values, FDR



Transformations and
Exploratory Data Analysis (EDA)

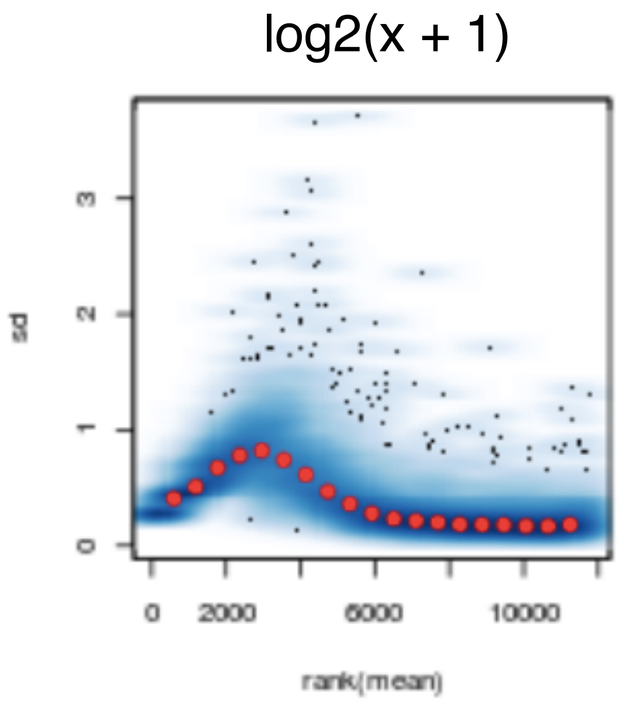
clustering, heatmaps,
sample-sample distances

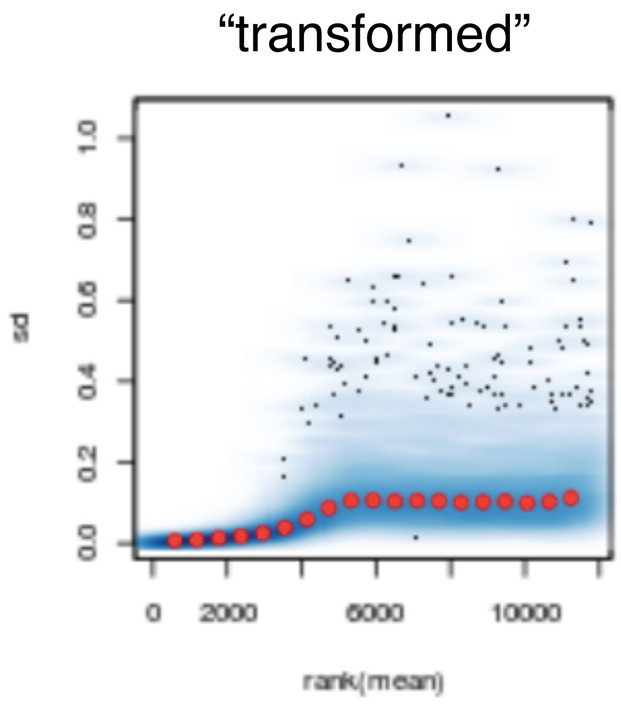




Transformations

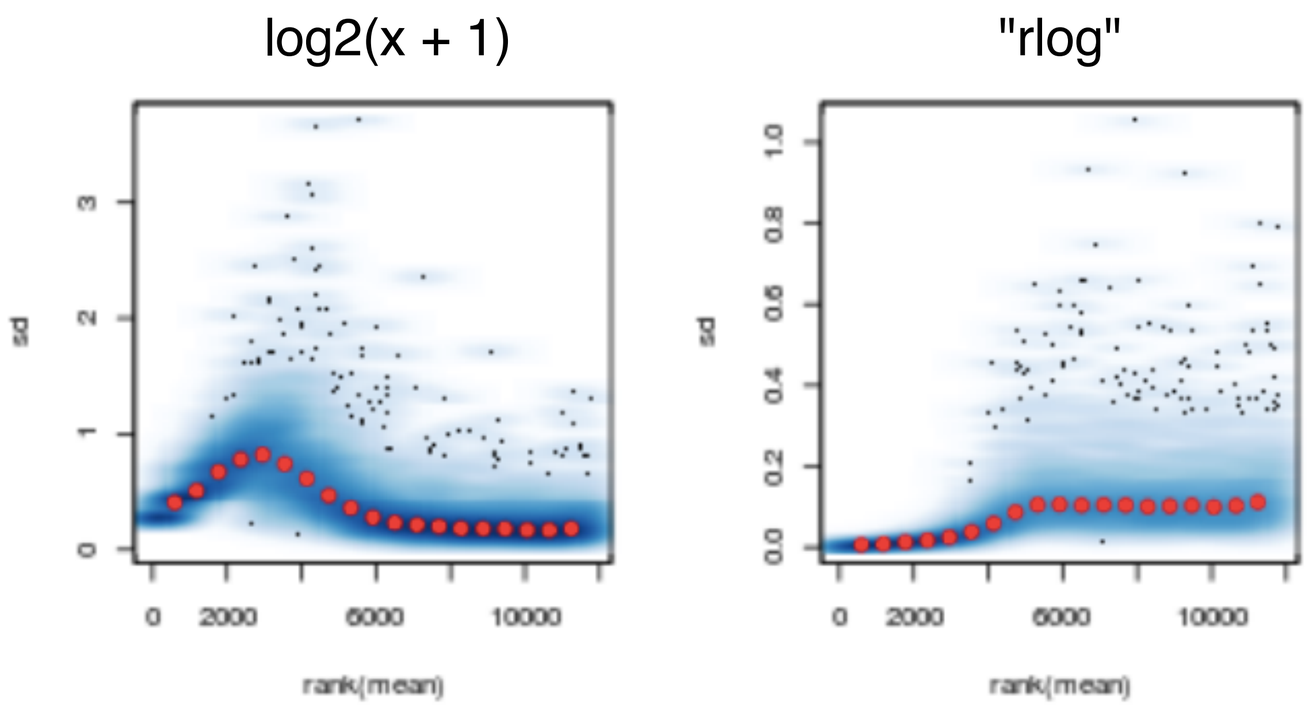
For comparison analyses when using unsupervised techniques, it can be useful to *transform* data. These techniques (VST and log) perform better when values have a similar dynamic range



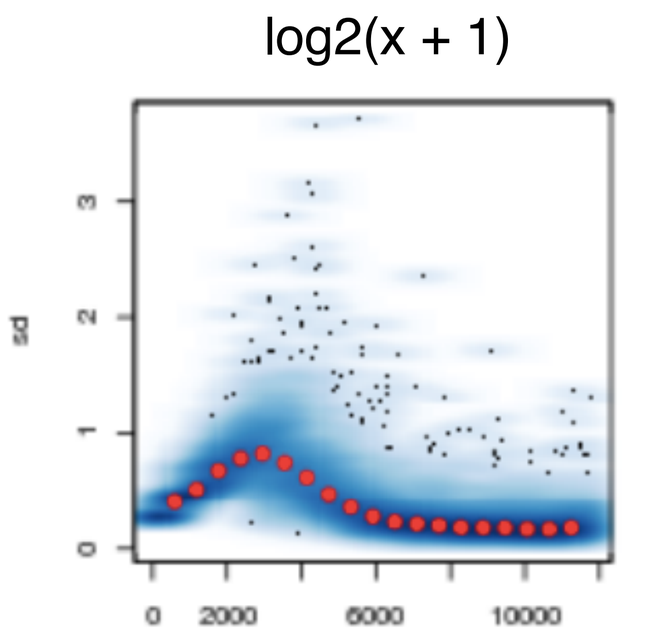


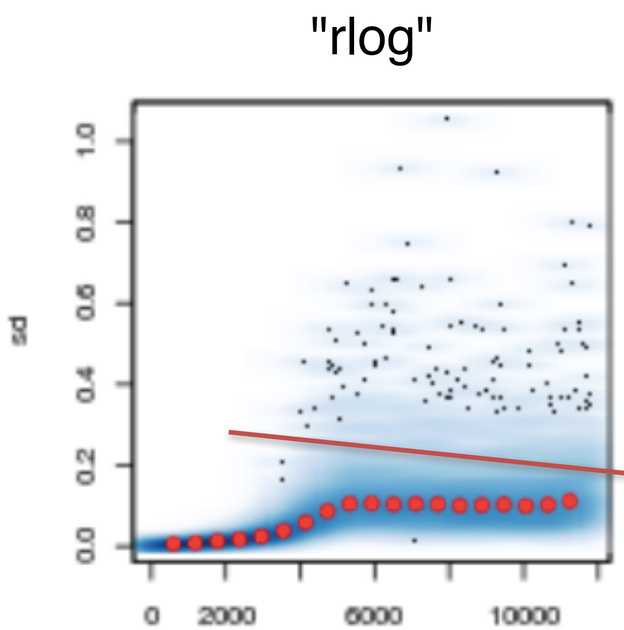
renders data *homoskedastic* (variance of the gene is stabilized across expression levels

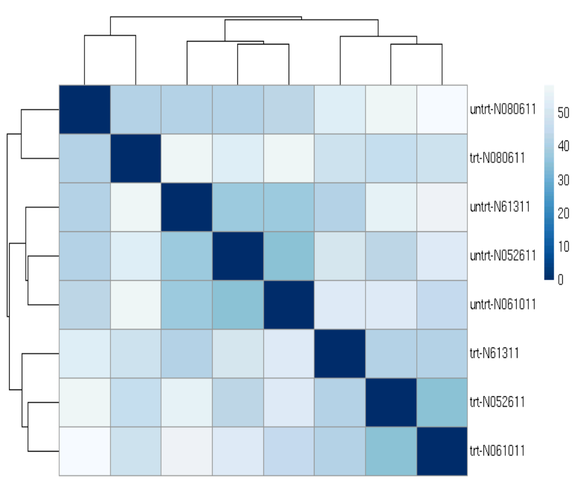
rlog stabilizes variances along the mean

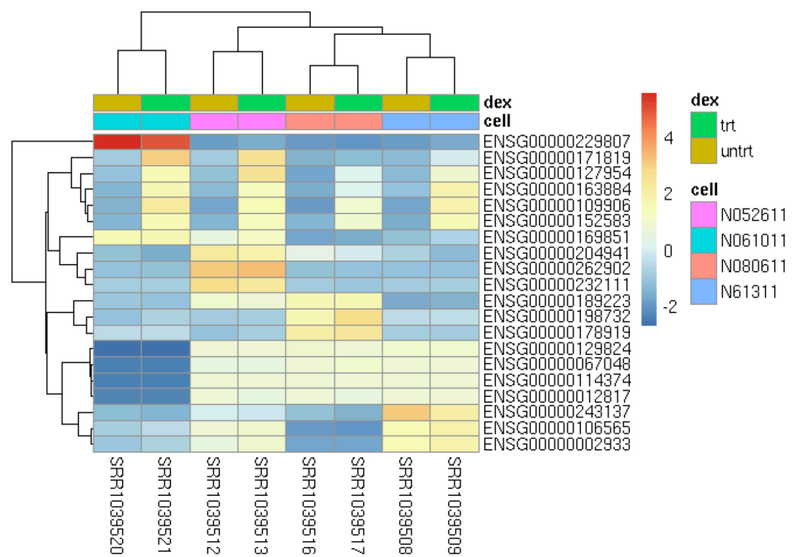


rlog stabilizes variances along the mean









Moderating the high variance / low count genes
Improves distances, clustering, visualizations

Variance stabilizing transform (VST)

- The variance stabilizing transformation is an earlier approach (from original DESeq) for transforming counts.

- Uses a function which is log-like but doesn't go to –Inf at x=0.

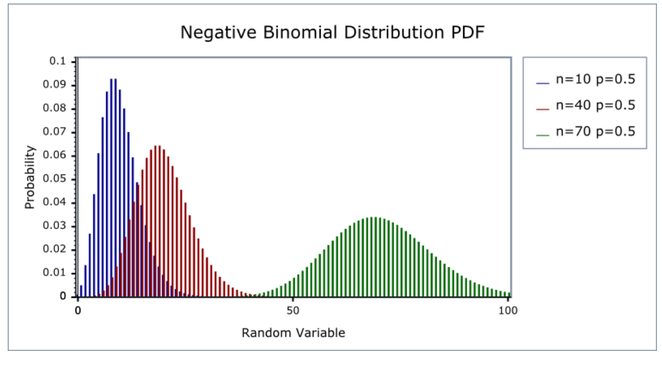
- VST doesn't use size factors so is better for data that has consistent sequence depth across samples

- VST is closed form, so can be better for large datasets (n > 50) due to speed.

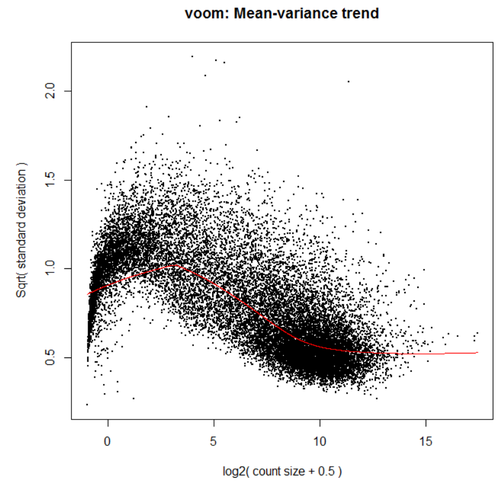
Comparison of methods

Count model vs linear model

- DESeq2 and edgeR similar approach, similar results
  - very sensitive, may sometimes underestimate FDR



- limma+voom uses a linear model, weights determined by variance over mean
  - strong control of FDR, may be less sensitive for small sample size
  - recommended when number of biological replicates per group grows large (e.g. > 20), because CLT kicks in



Credits

RNA-seq statistical analysis and gene-level differential expression

Mike Love @mikelove
Dept. of Biostatistics and Computational Biology
Dana-Farber Cancer Institute & Harvard TH Chan School of Public Health

Acknowledgements:

DESeq2/DEXSeq:

Wolfgang Huber
Simon Anders
Alejandro Reyes

Rafael Irizarry

Work supported by:

Dana-Farber Cancer Institute

Harvard TH Chan School of Public Health

NIH Training Grant