RNASEQ - DIFFERENTIAL EXPRESSION STATISTICS

I. Differential expression statistics
II. DESeq2 – more details
III. Transcript level quantification

From counts to differential expression statistics

Differential expression statistics

- Its just a count matrix! Simple! Right?

	24_GA_CL	24_GA_CP	24_GA_CR	GA-CL	GA-COL
Gene1	8	3	9	7	7
Gene2	4	0	1	2	7
Gene3	19	13	29	27	35
Gene4	147	56	102	60	73
Gene5	778	212	380	149	266

Normalizing

Biggest concern:

- Different numbers of reads per sequence run

Other concerns

- Different lengths of transcripts
- Differing ability to map reads
- GC sequencing bias

Can safely ignore these other concerns by comparing changes in gene expression *within the same gene* across samples

Normalization – the Old Guard

RPKM = Reads Per Kilobase of exon per million Mapped reads

$$\frac{\text{count} \; * \; 10^9}{\text{transcript length} * \text{total reads sequenced}}$$

To account for paired end reads:

FPKM = Fragments Per Kilobase of exon per Million Mapped reads (paired end reads)

Lior Patcher, Models for transcript quantification from RNA-Seq, ArXiv http://arxiv.org/abs/1104.3889

Normalization – Evolving Thoughts

- Original strategies suffer from some significant problems
- Read densities even after compensating for total read depth, may not be directly comparable
- A small number of genes can consume a significant fraction of sequencing resources, and if their expression changes, this skews the count distribution for remaining genes
- Put another way - the proportional representation of each gene is dependent on the expression levels of all other genes

New Normalization

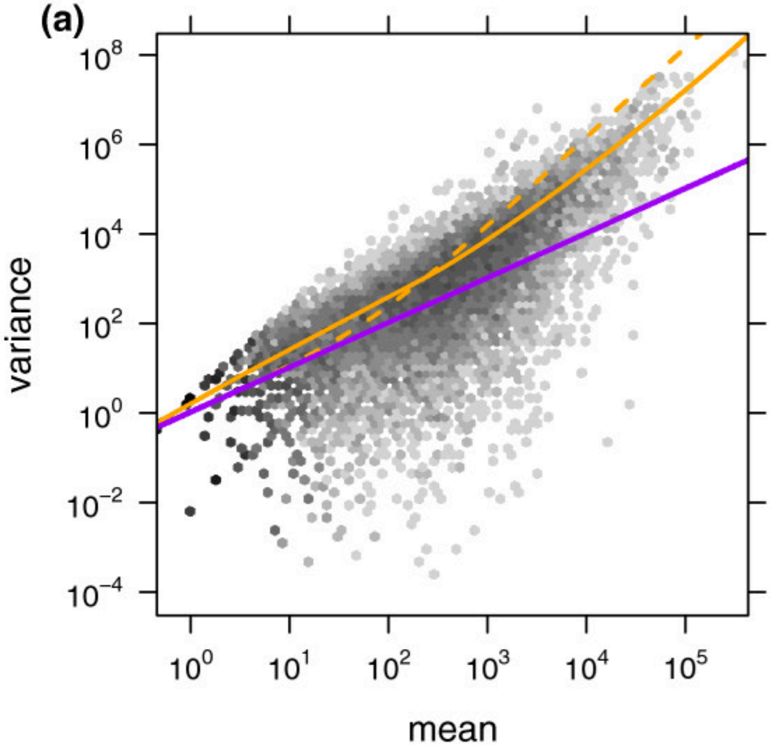
- Slightly more complicated to compute, usually implemented in a software package to make life easier
- Basically, need scaling factors that exclude or reduce the impact of highly expressed and highly DE genes
- DESeq
  - Start by computing the median of the ratio, for each gene, of its read count over its geometric mean across all samples. With the assumption that most genes are not DE, then uses this median of ratios to obtain the scaling factor associated with each sample
- EdgeR
  - TMM – trimmed means of M values
  - Exclude genes that have overall very high expression or very different expression. Then compute a scaling factor

Statistical Models

- Originally (pre-2010) Poisson distribution was often used to model counts
- Data was found to be over-dispersed (ie. it predicts less variation than what is seen in the data)
- Leads to higher false positives

Negative Binomial

- a good substitute for an over-dispersed poisson (sample variance exceeds sample mean)
- Allows mean and variance to be different



Purple = predicted variance implied by Poisson

Orange = variance used by edgeR (ie using negative binomial)

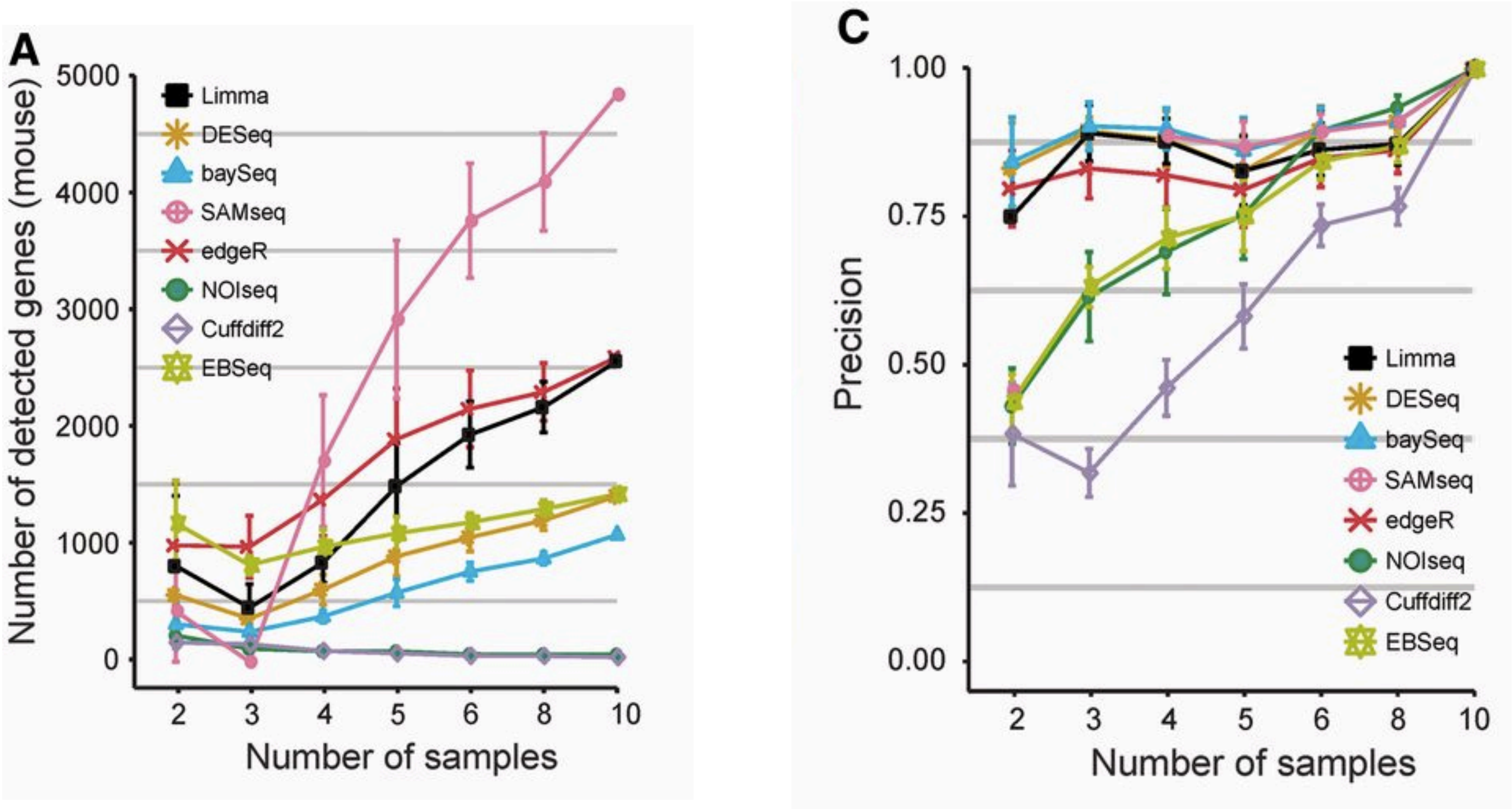
Anders and Huber 2010

Many approaches

- Negative Binomial - edgeR, DESeq, DESeq2
- Beta negative binomial – cufflinks/cuffdiff2
- Poisson - DEGseq, Myrna, PoissonSeq
- Bayesian - baySeq
- Non-parametric - SAMseq, NOIseq

Each has further variation in normalization procedures (RPKM, upper quartile, median, TMM, Quantile) and testing strategy (fishers exact, likelihood ratio, parallelized permutation test, score test, Wald test, posterior probability, Wilcoxon test)

Different approaches give different results



Seyednasrollah et al., 2013
Comparison of software packages for detecting differential expression in RNA-seq studies

How to choose?

- DESeq, DESeq2, EdgeR generally dominate the market right now. (See many references at end of presentation)

- Right now decisions are largely driven by limited biological replicates. With low numbers of replicates there is not enough power to accurately estimate mean and variance of expression for each gene.

- One day when we have lots of affordable biological replicates? Nonparametric may take over.

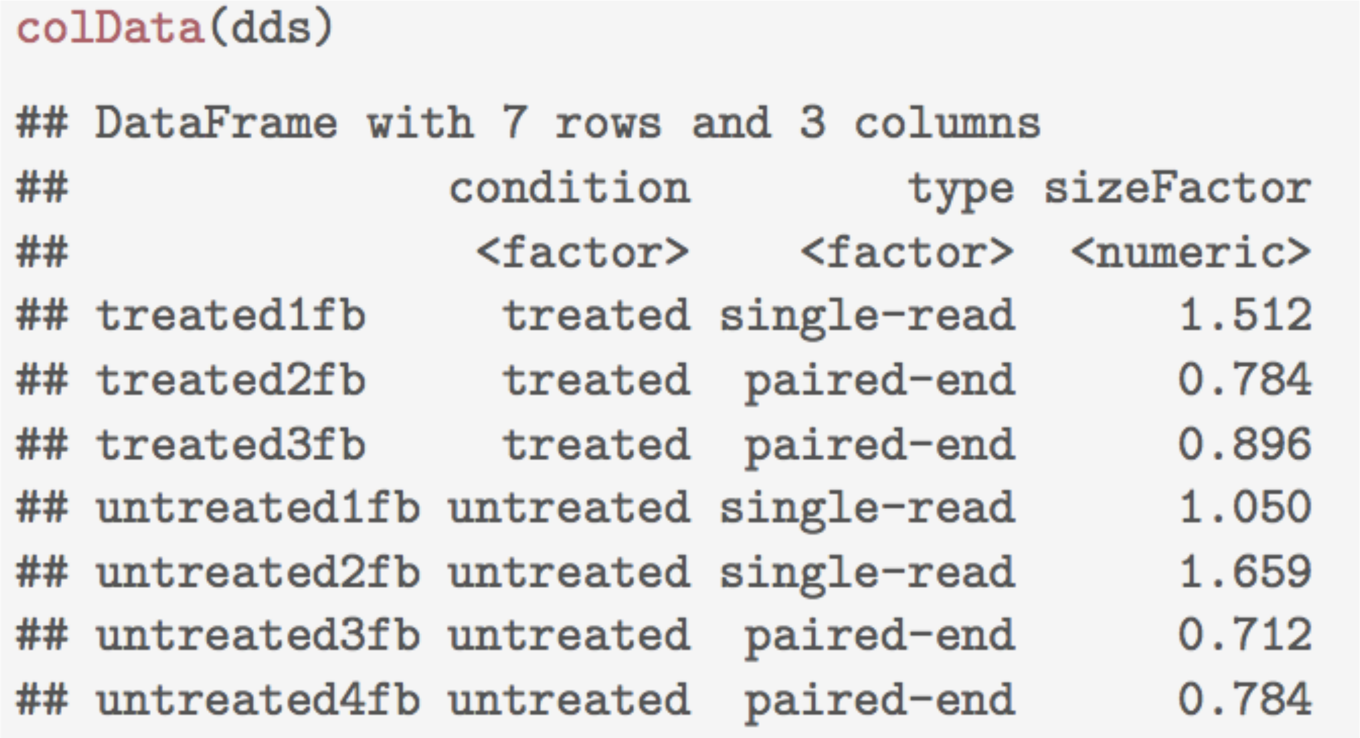
DESeq2

DESeq2 - Test for differential expression

- Accepts only raw counts as input
- DESeq2 approach:
- Generalized linear model is fit for each gene
  - Flexible - allows for complex designs
- Wald test is the default test
  - An adjusted log fold change is used, resulting in a z-statistic
  - Test for each coefficient of GLM or contrasts of coefficients
  - No need for a reduced model
- Likelihood Ratio Test also available
  - Do need a reduced model

- Need to adjust for multiple testing (of many genes)
  - Benjamini and Hochberg

Multi-factor Design



```
colData(dds)

## DataFrame with 7 rows and 3 columns
##              condition        type sizeFactor
##               <factor>    <factor>  <numeric>
## treated1fb     treated single-read      1.512
## treated2fb     treated  paired-end      0.784
## treated3fb     treated  paired-end      0.896
## untreated1fb untreated single-read      1.050
## untreated2fb untreated single-read      1.659
## untreated3fb untreated  paired-end      0.712
## untreated4fb untreated  paired-end      0.784
```

```
design(ddsMF) <- formula(~ type + condition)
```

The variable of interest goes at the end of the formula. Thus the results of this design will by default pull the condition results

Interaction term

- Interaction terms can be added to the design formula, in order to test if the log2 fold change attributable to a given condition is different based on a second variable
- for example if the treatment effect differs based on another grouping variable like species
- Colon used to specify interaction

```
design(ddsMF) <- formula(~ type + condition
+ type:condition)
```

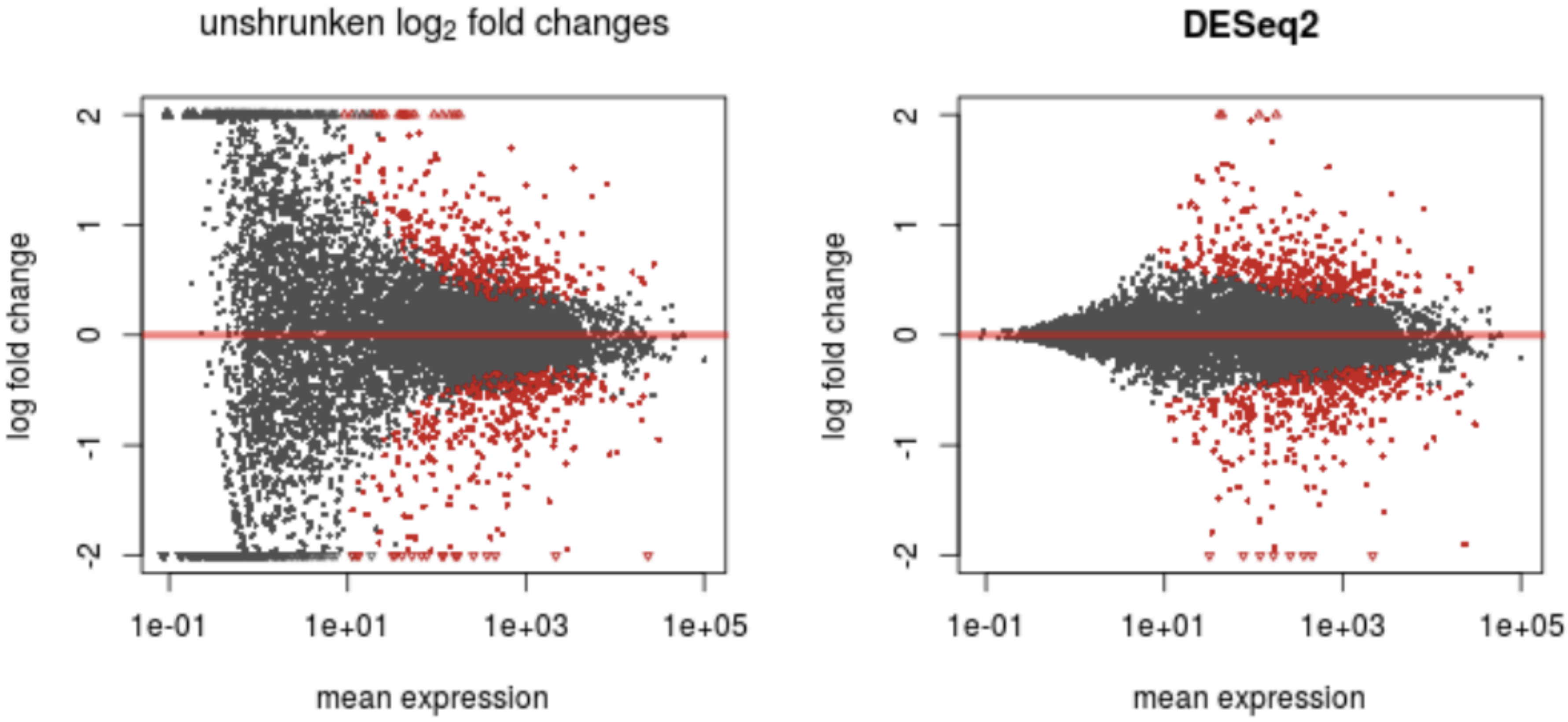
What else can DESeq2 do?

- Vignette and manual available from Bioconductor site
- http://bioconductor.org/packages/release/bioc/html/DESeq2.html

  - Likelihood Ratio Test
  - Contrasts
  - MA plot
  - Count data transformations
  - Heatmap
  - Sample clustering
  - Principal Components Plot

MA Plot

- Mean (normalized) expression vs log fold change (LFC)
- DESeq2 also performs shrinkage of LFC
- Genes with high dispersion + low counts get a relatively lower LFC



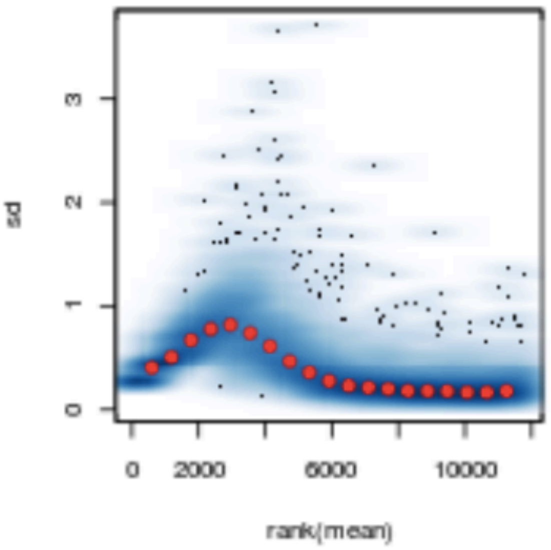
unshrunken $\log_2$ fold changes

DESeq2

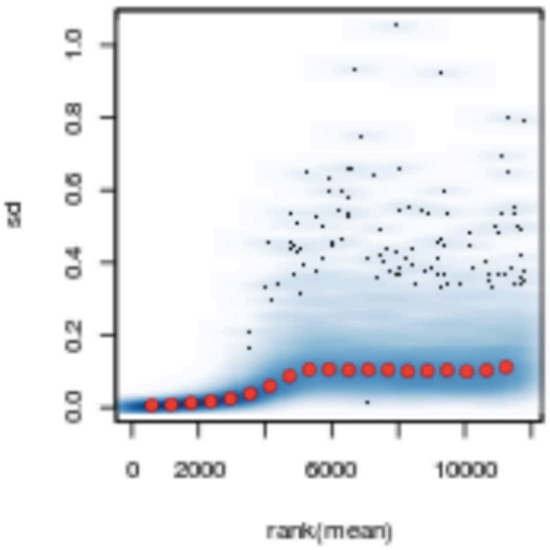
Count Data Transformation

- DESeq2 operates on raw read counts
- This is not optimal for downstream analysis such as visualization or clustering
- DESeq2 offers two transformation options

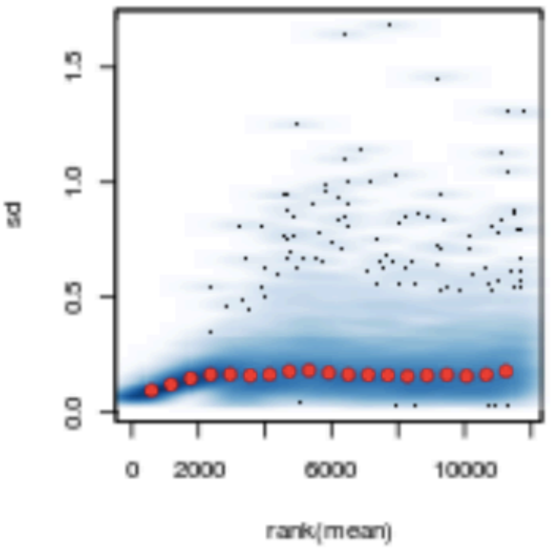
Shifted logarithm



Regularized log transformation



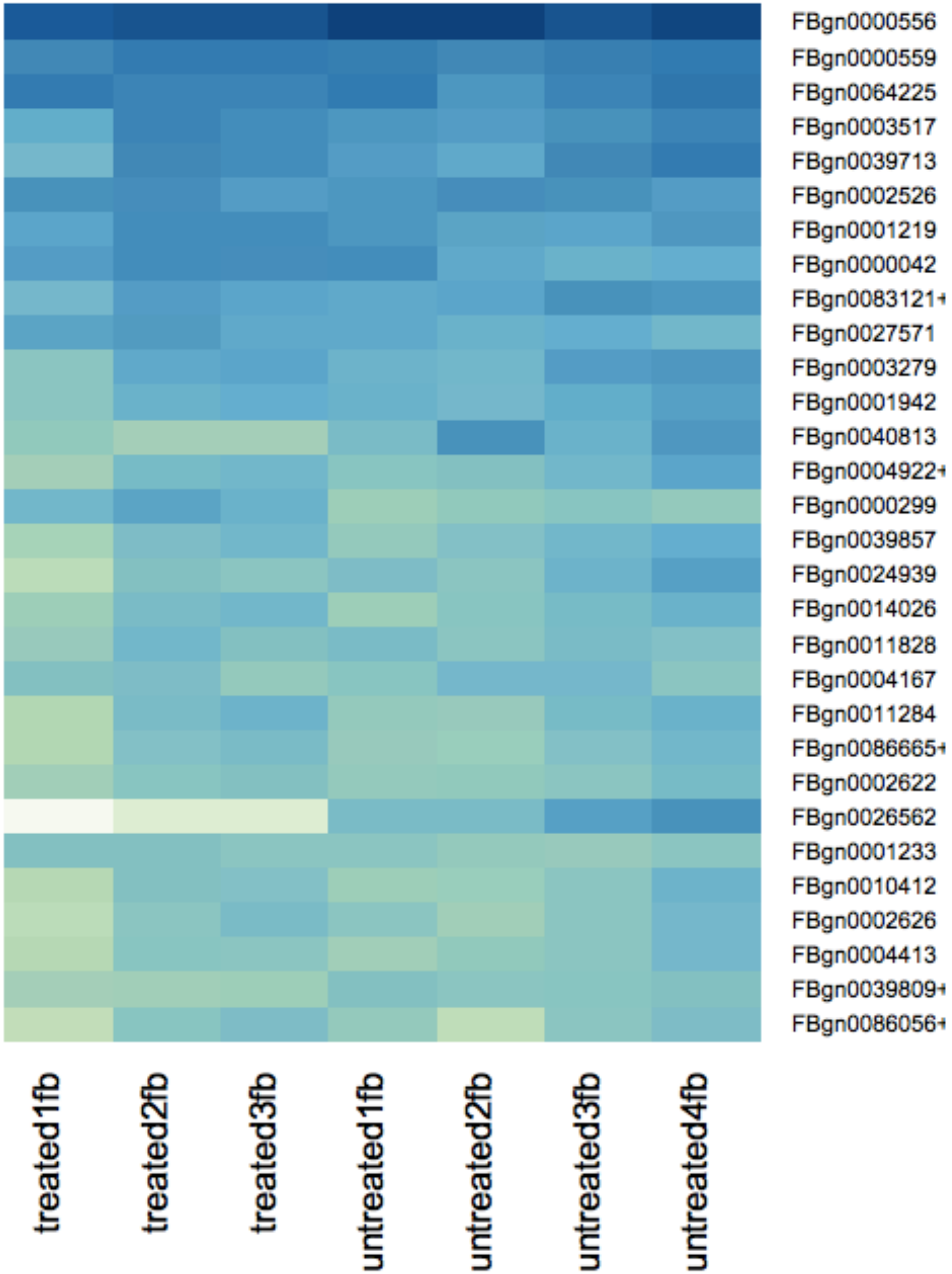
Variance stababilizing transformation



Heatmaps

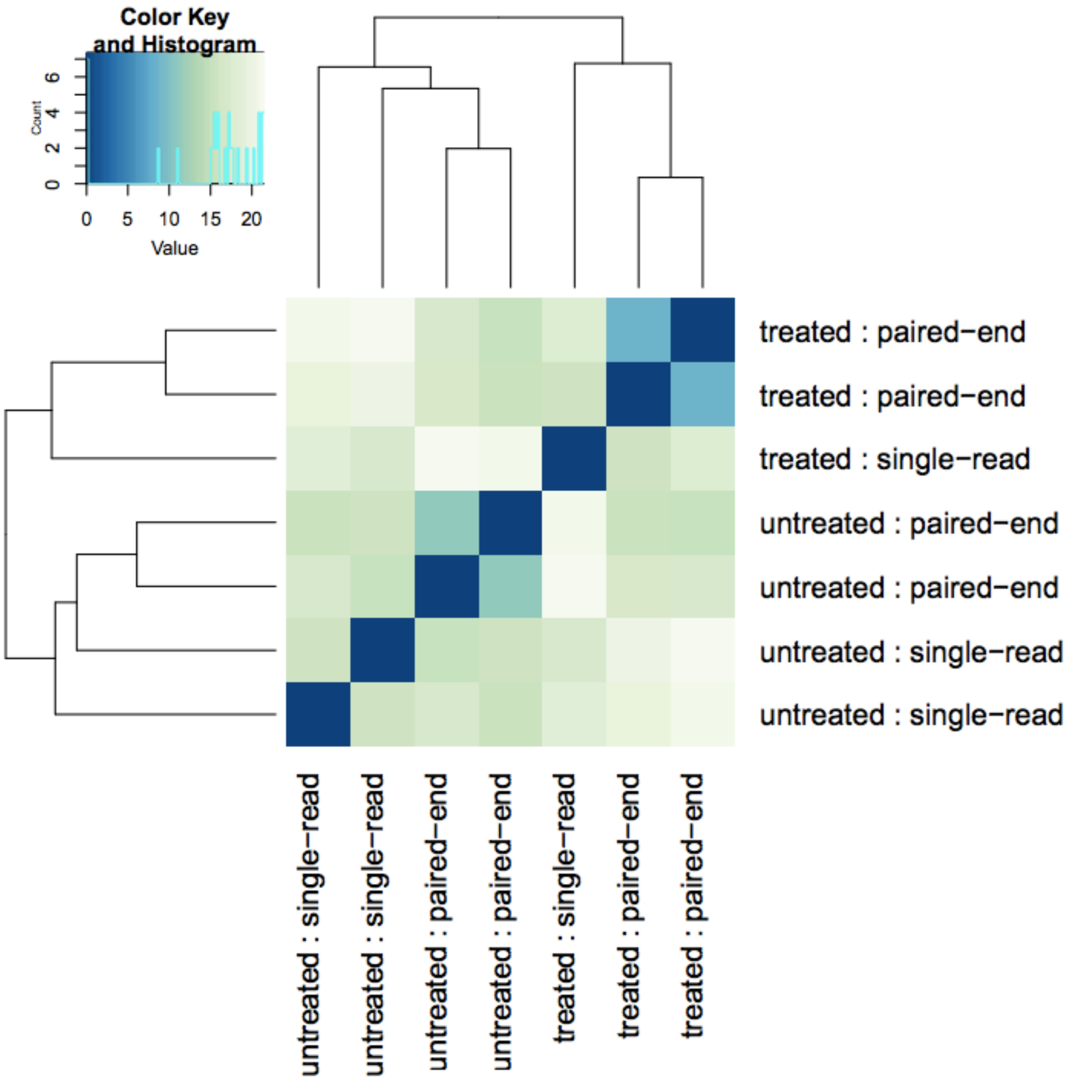
30 most highly expressed genes

VST



Sample clustering

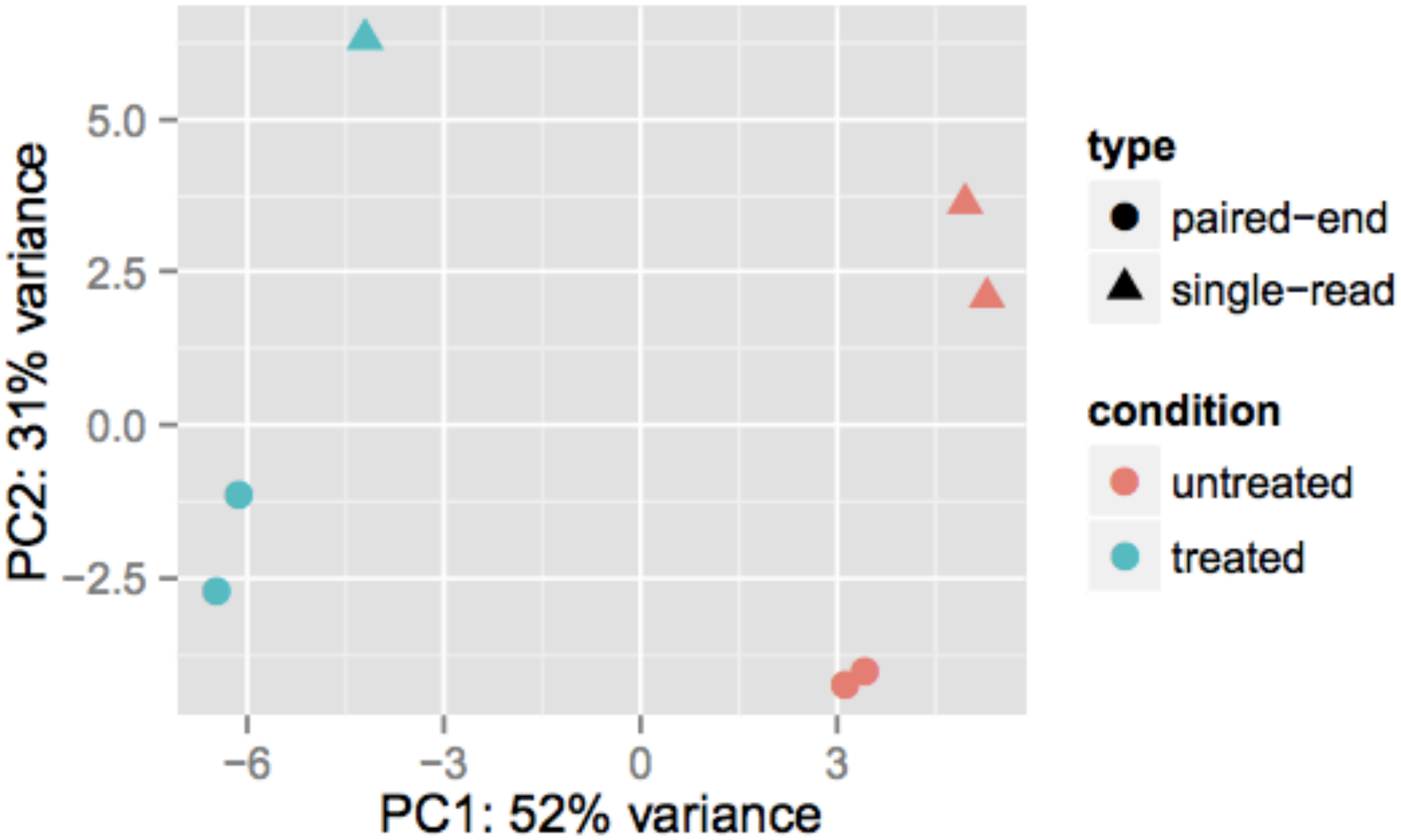
- Good for quality control
- Do any samples appear to be outliers?



PCA

- Use only for QC
- Returns two largest components
- Brush up on what a PCA is:

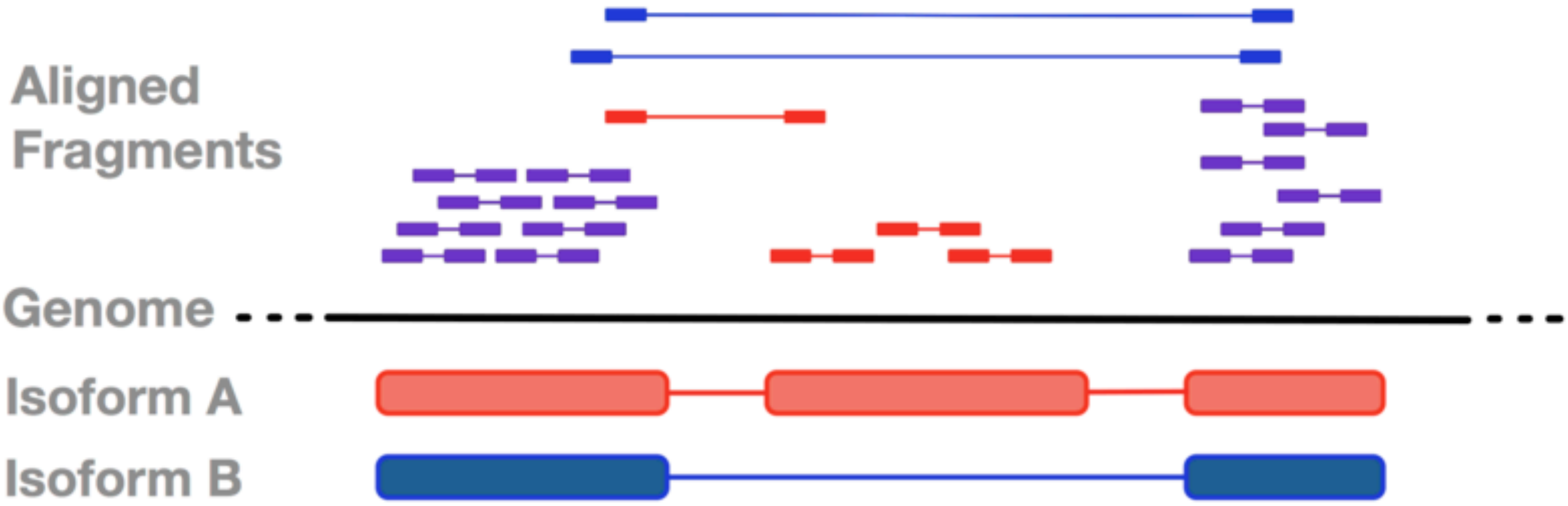
  https://en.wikipedia.org/wiki/Principal_component_analysis



Transcript level quantification

Transcript level quantification

- (Isoform quantification)



https://cgrlucb.wikispaces.com/Isoform+Deconvolution+and+Unannotated+Species

Transcript level quantification

- Active area of research, currently recommended by many in the field
- Allocate multi-mapping reads among the possible transcripts. How?
- Software to calculate transcript abundances:
  - Salmon (Patro et al. 2016)
  - Sailfish (Patro, Mount, and Kingsford 2014)
  - kallisto (Bray et al. 2016)
  - RSEM (Li and Dewey 2011)
- Can still use DESeq2
  - R package tximport

Soneson et al., 2016:

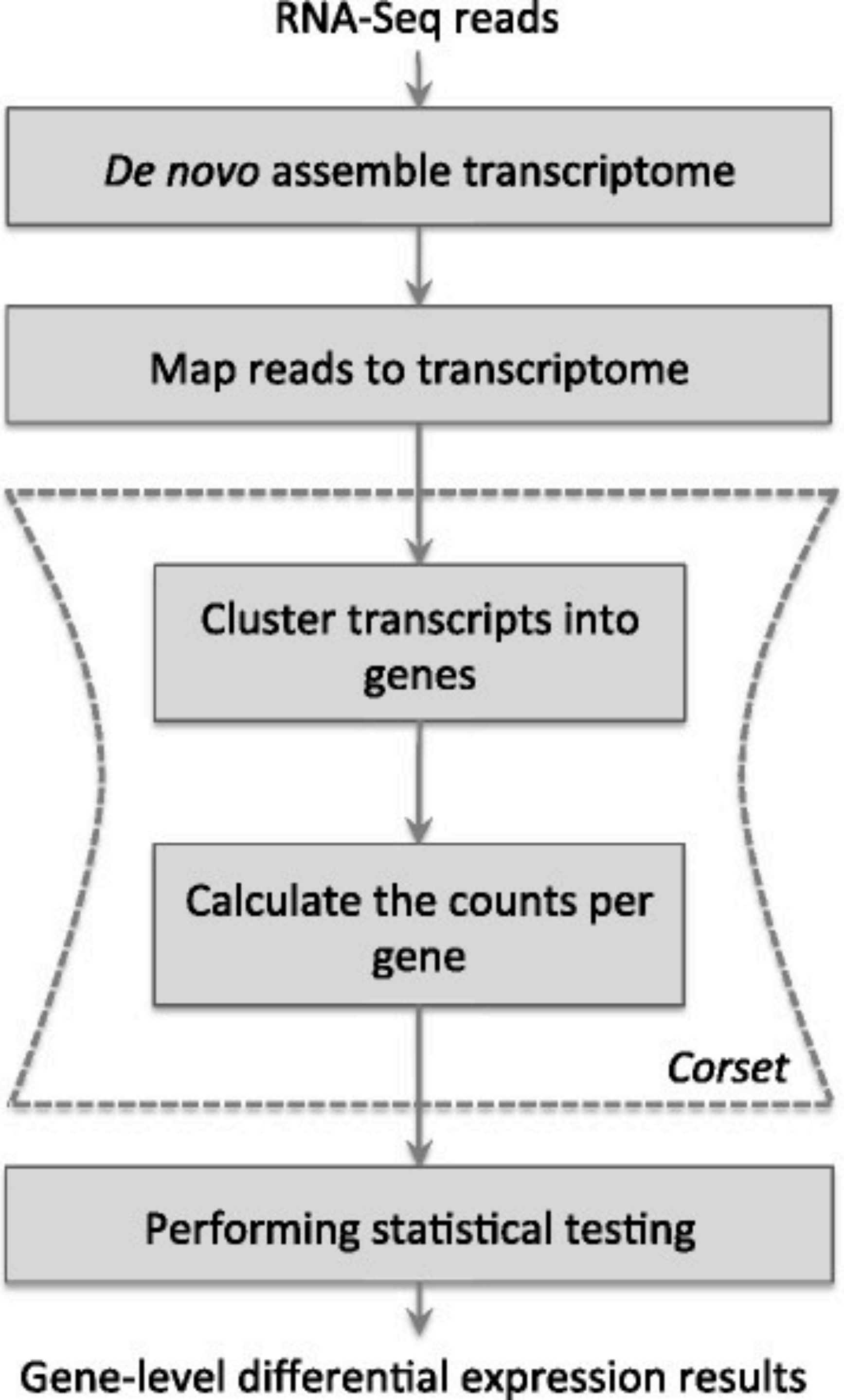
- Gene-level results are often more accurate, powerful and interpretable than transcript-level results
- Incorporating transcript-level estimates yields more accurate gene-level results.

What if you have a de novo assembled transcriptome?

- This presents some statistical problems – your “unigenes” or transcript contigs are often fragments of the same gene
- fewer reads can be aligned unambiguously (because of duplicated sequences)
- the statistical power of the test for differential expression is reduced as reads must be allocated amongst a greater number of contigs
- the adjustment for multiple testing is more severe

New packages are available to cluster contigs from de novo assemblies for more accurate quantification:

- Corset (Davidson et al., 2014)
- RapClust (Srivastava et al., 2016)


RNA-Seq reads
De novo assemble transcriptome
Map reads to transcriptome
Cluster transcripts into genes
Calculate the counts per gene
Corset
Performing statistical testing
Gene-level differential expression results

References

- Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. A survey of best practices for RNA-seq data analysis. Genome biology. 2016 Jan 26;17(1):13.
- Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome biology. 2013 Sep 10;14(9):3158.
- Huang HC, Niu Y, Qin LX. Differential expression analysis for RNA-Seq: an overview of statistical methods and computational software. Cancer informatics. 2015;14(Suppl 1):57.
- Finotello F, Di Camillo B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. Briefings in functional genomics. 2015 Mar 1;14(2):130-42.
- Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014 Dec 5;15(12):550.
- Love M, Anders S, Huber W. Differential analysis of count data–the DESeq2 package. Genome Biology. 2014 May 13;15:550.
- Anders S, Huber W. Differential expression analysis for sequence count data. Genome biology. 2010 Oct 27;11(10):R106.
- Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Briefings in bioinformatics. 2015 Jan 1;16(1):59-70.