RNASEQ - DIFFERENTIAL EXPRESSION STATISTICS

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

From counts to differential expression statistics

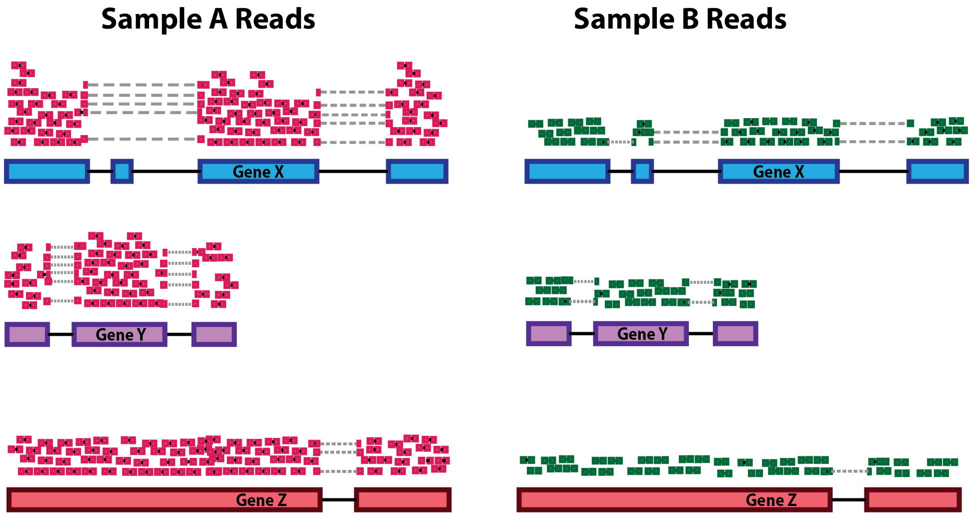
Differential expression statistics

- Its just a count matrix! Simple! Right?

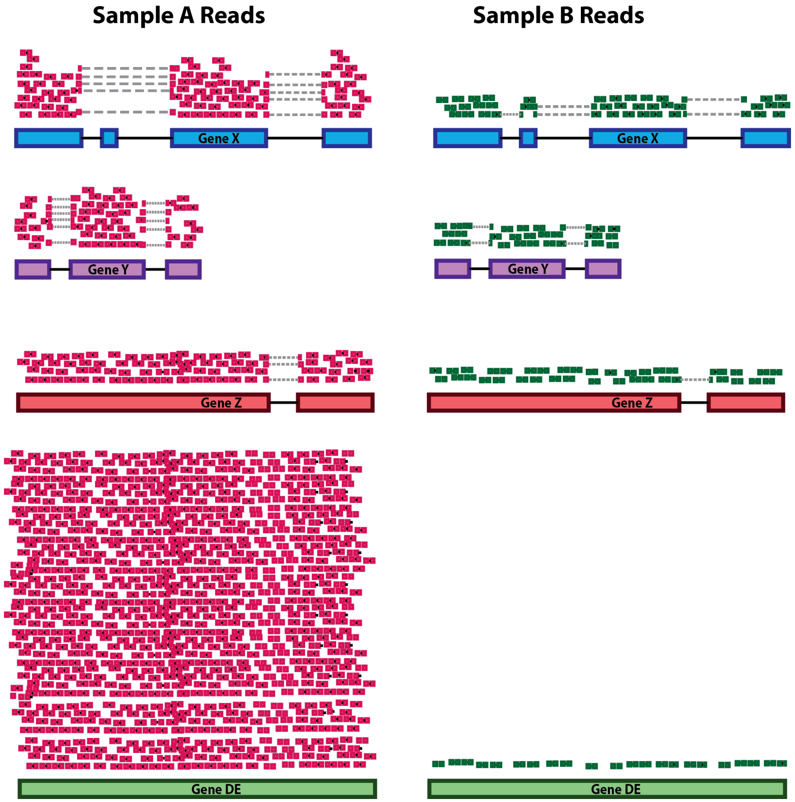
	24_GA-CL	24_GA-CP	24_GA-CR	GA-CL	GA-COL
Gene1	8	3	9	7	7
Gene2	4	0	1	2	7
Gene3	19	13	29	27	35
Gene4	147	56	102	60	73
Gene5	778	212	380	149	266

Normalizing - Major concerns for DE analysis

- Sequencing depth between samples



- mRNA composition



https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Normalizing – Other concerns

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

✓These concerns are can be ignored when our research question focused on gene differential expression analysis between treatment.

✓However, DEseq2 provided solution for above concerns by calculating gene specific normalization factor.

Normalizing

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads	• sequencing depth	gene count comparisons between replicates of the same samplegroup; NOT for within sample comparisons or DE analysis
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	• sequencing depth • gene length	gene count comparisons within a sample or between samples of the same sample group; NOT for DE analysis
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	• sequencing depth • gene length	gene count comparisons between genes within a sample; NOT for between sample comparisons or DE analysis
DESeq2's median of ratios	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	• sequencing depth • RNA composition	gene count comparisons between samples and for DE analysis; NOT for within sample comparisons
EdgeR's trimmed mean of M values (TMM)	uses a weighted trimmed mean of the log expression ratios between samples	• sequencing depth • RNA composition • gene length	gene count comparisons between and within samples and for DE analysis

DEseq2 Normalization – median of ratio

- Basically, calculate scaling factors that exclude or reduce the impact of highly expressed and highly DE genes

Gene	Sample A	Sample B	pseudo-Reference sample	Sample A / Ref Ratio	Sample B / Ref Ratio
EF2A	2800	700	1400	2800/1400 = **2**	700/1400 = **0.5**
ABCD1	60	15	30	60/30= **2**	15/30 = **0.5**
MEFV	1600	400	800	1600/800 = **2**	400/800 = **0.5**
BAG1	160	40	80	160/80 = **2**	40/80 = **0.5**
MOV10	10000	10000	10000	10000/10000 = **1**	10000/10000 = **1**
MA5	200	50	100	200/100=**2**	50/100=**0.5**
QB6	600	150	300	600/300=**2**	150/300=**0.5**
Scaling factor	…	…	…	**A-Median ratio=2**	**B-Median ratio=0.5**

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Statistical Models - challenges & improvements

- Count data:
  - Non-normal distribution
  - Variation depend on mean

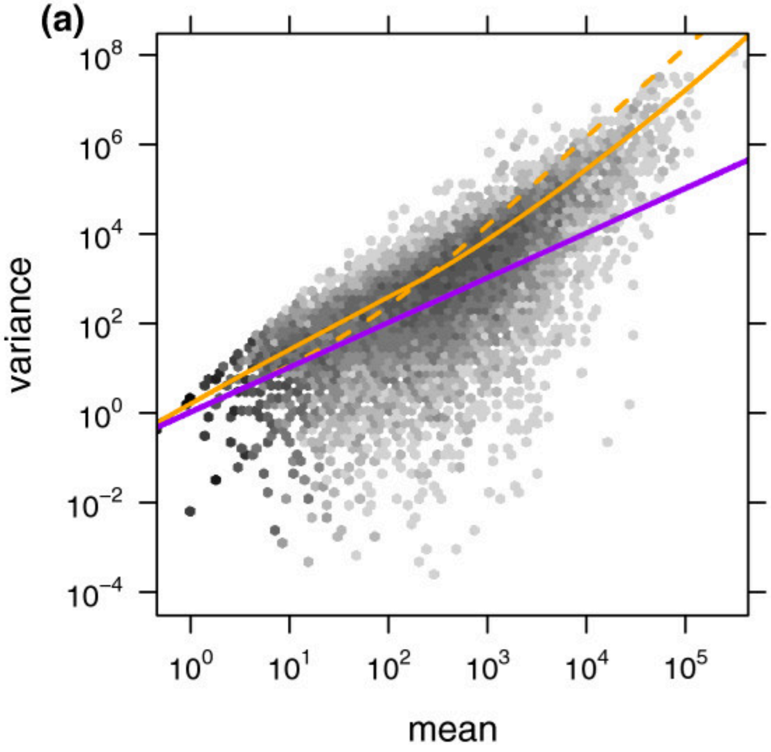
- Over-dispersion: sample variation exceeds sample mean

- Small numbers of replicates: makes the within-group variation hard to estimate.

Statistical Models

- Originally (pre-2010) Poisson distribution was often used to model counts
- Data was found to be over-dispersed (i.e. 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)

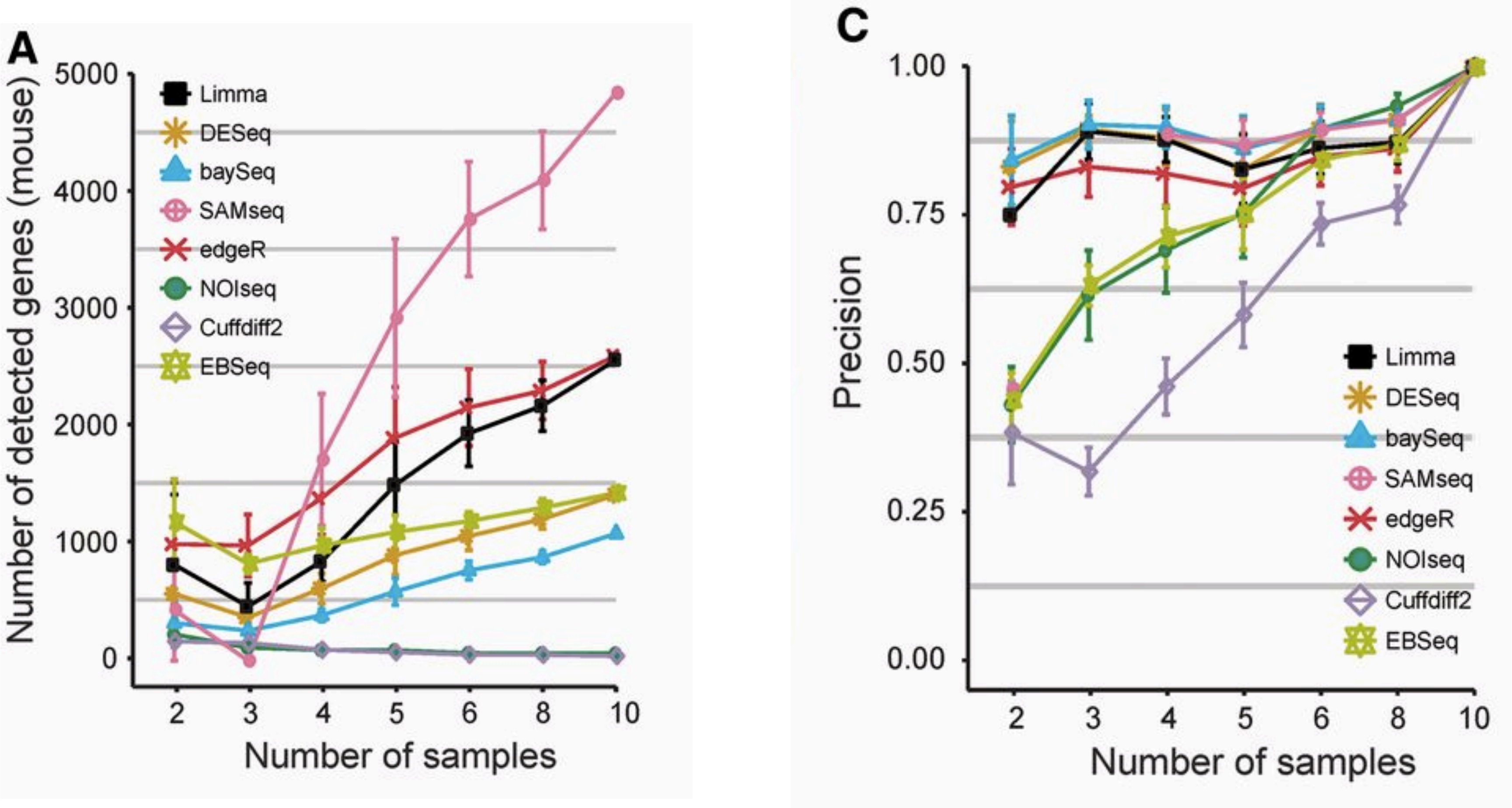
Anders and Huber 2010

Many approaches - Statistical Models

- 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 - principles

DESeq2 - Test for differential expression

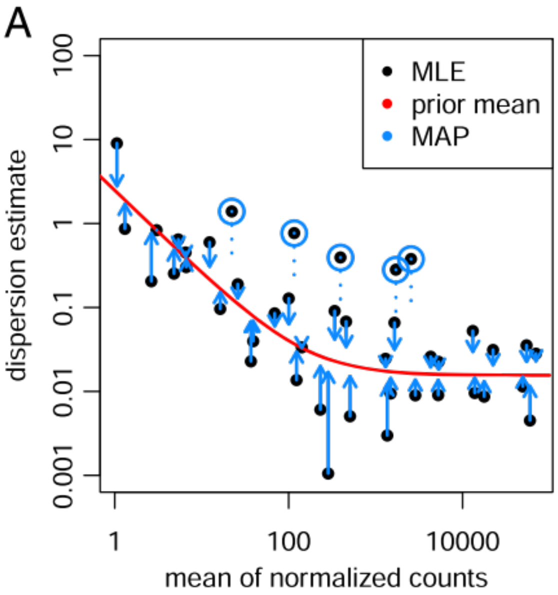
- Accepts only raw counts as input: median of ratio for normalization
- Negative binomial distribution: solve over-dispersion problem
- DESeq2 approach:
- Generalized linear model is fit for each gene
    - Flexible - allows for complex designs
- Wald test is the default test
    - An adjusted log fold change is used, resulting in a z-statistic
    - Test for each coefficient of GLM or contrasts of coefficients
- Need to adjust for multiple testing (of many genes)
    - Benjamini and Hochberg

Improvements in DEseq2 vs DEseq

- **Shrinkage for dispersion estimation**

- Small numbers of replicates: makes the within-group variation hard to estimate.

✓ Solution: pool variance information across genes. With the assumption that genes of similar average expression strength have similar dispersion.



Love, Michael I., Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome biology* 15.12 (2014): 550.

Improvements in DEseq2 vs DEseq

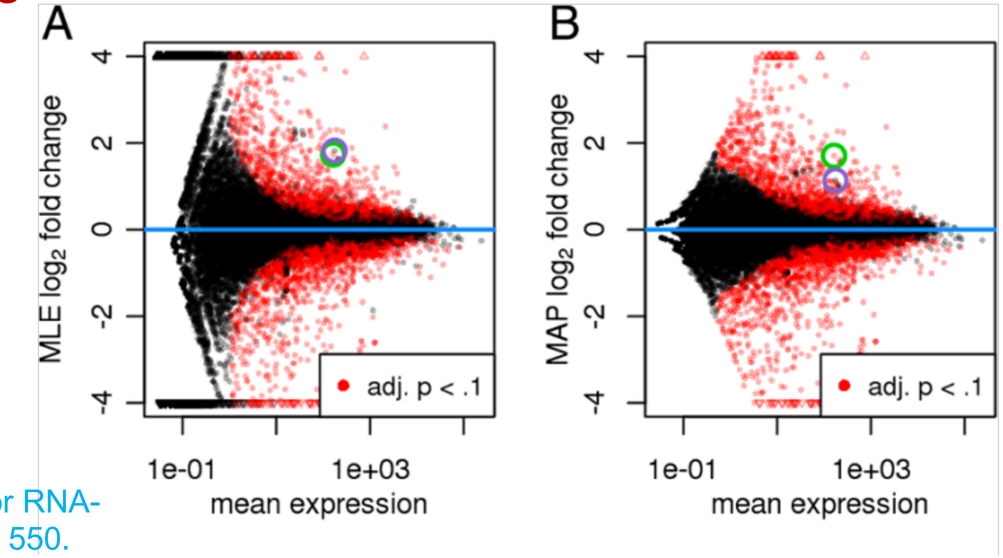
Shrinkage for log fold change (LFC)

- Significant test are less reliable for gene with low counts as log fold change are more noisy.

✓ Solution: using shrinkage estimator for fold change - Empirical Bayes shrinkage

✓ As a result, log fold changes in below situations will be shrank toward zero

- Low count genes
- Dispersion is high
- Few replicates



Love, Michael I., Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome biology* 15.12 (2014): 550.

DEseq2 – usage and results

Formula - Multi-factor Design

	sampleID	Genotype	Treatment
1	sample_1	Sensitive	Drouht
2	sample_2	Sensitive	Drouht
3	sample_3	Sensitive	Drouht
4	sample_4	Sensitive	Norm
5	sample_5	Sensitive	Norm
6	sample_6	Sensitive	Norm
7	sample_7	Resistant	Drouht
8	sample_8	Resistant	Drouht
9	sample_9	Resistant	Drouht
10	sample_10	Resistant	Norm
11	sample_11	Resistant	Norm
12	sample_12	Resistant	Norm

```
design(ddsMF) <- formula(~ Genotype + Treatment)
```

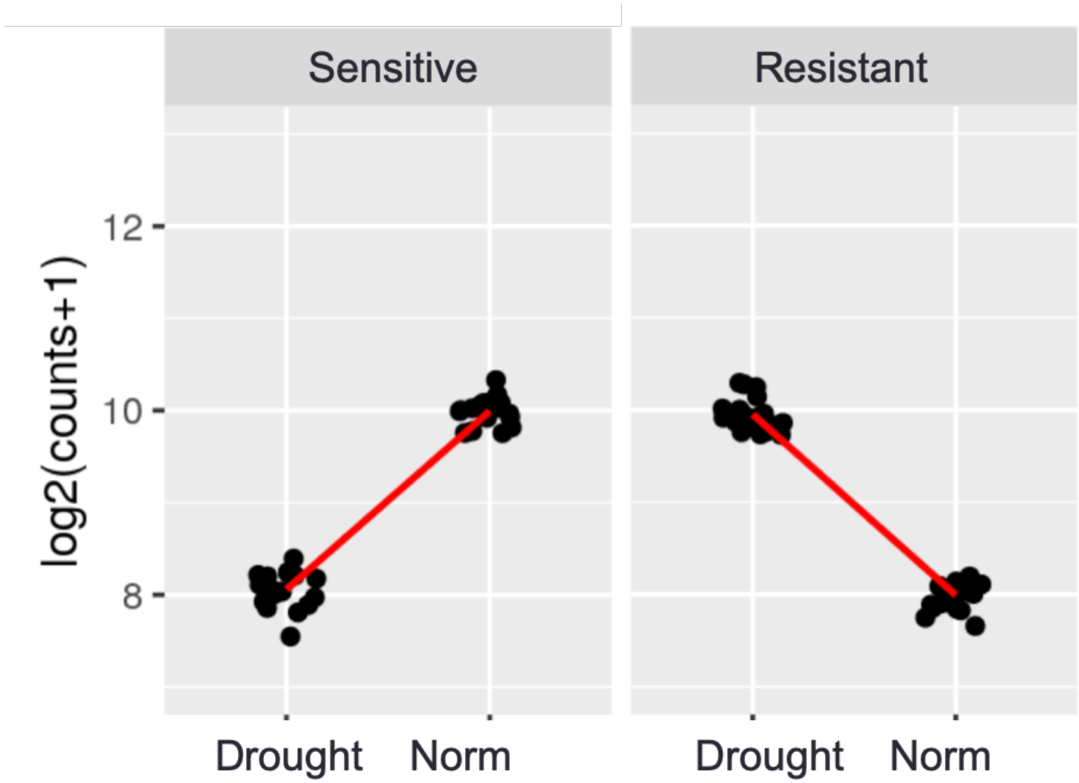
The variable of interest goes at the end of the formula. Thus the results of this design will by default pull the condition results

Interaction term

- To test the **interaction effects**: If the treatment impacts differs between two genotypes

Interaction

```
design(ddsMF) <- formula (~Genotype + Treatment + Genotype : Treatment)
```



What else can DESeq2 do?

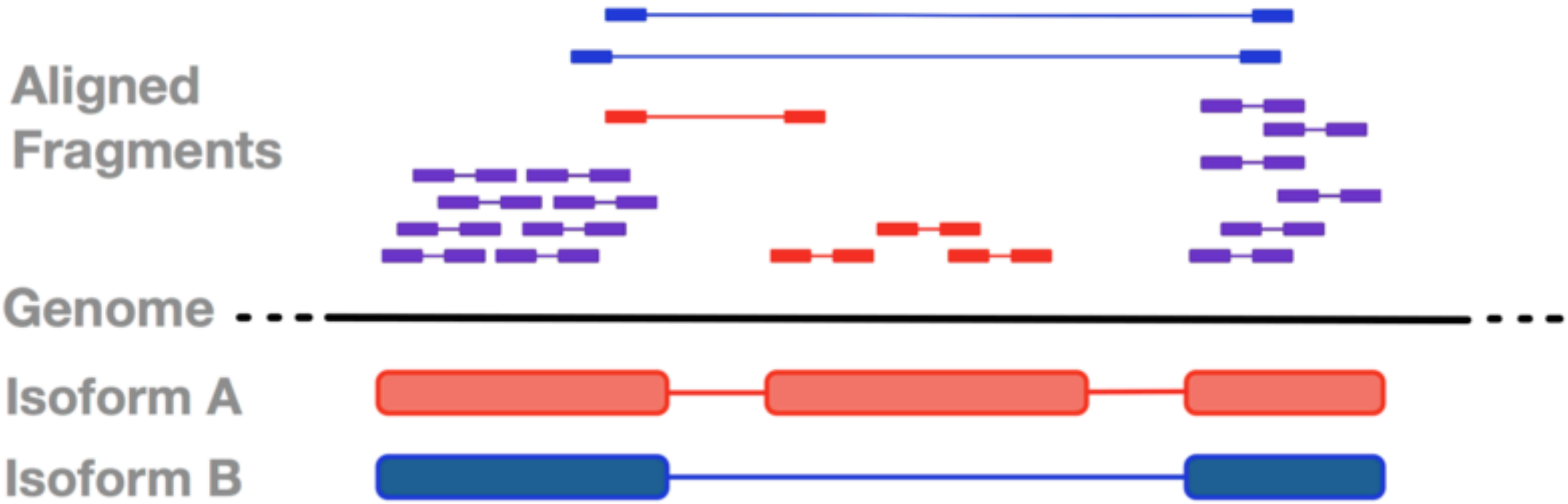
- Vignette and manual available from Bioconductor site
- http://bioconductor.org/packages/release/bioc/html/DESeq2.html

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

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:

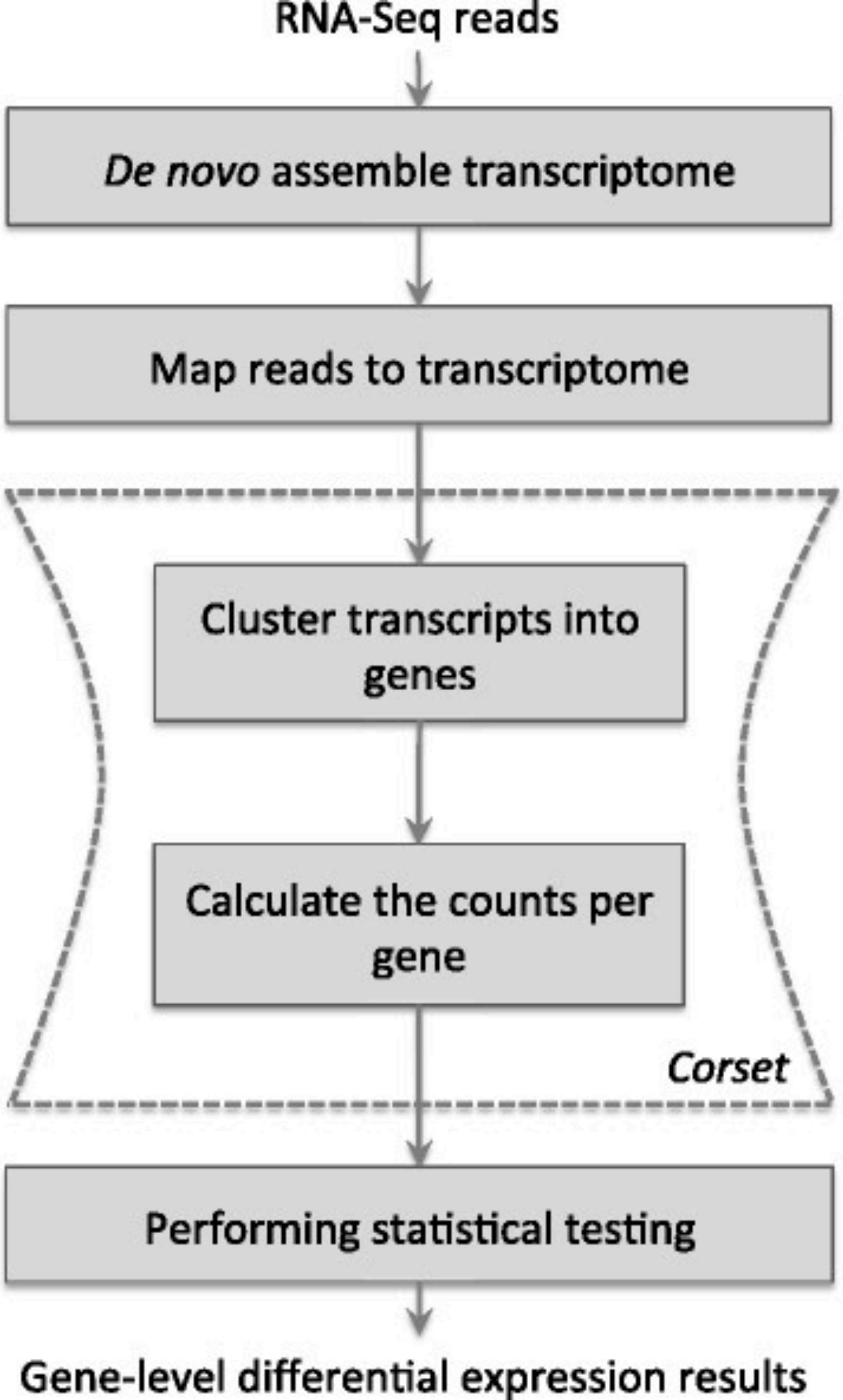
- Gene-level results are often more accurate, powerful and interpretable than transcript-level results
- Incorporating transcript-level estimates yields more accurate gene-level results.

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.