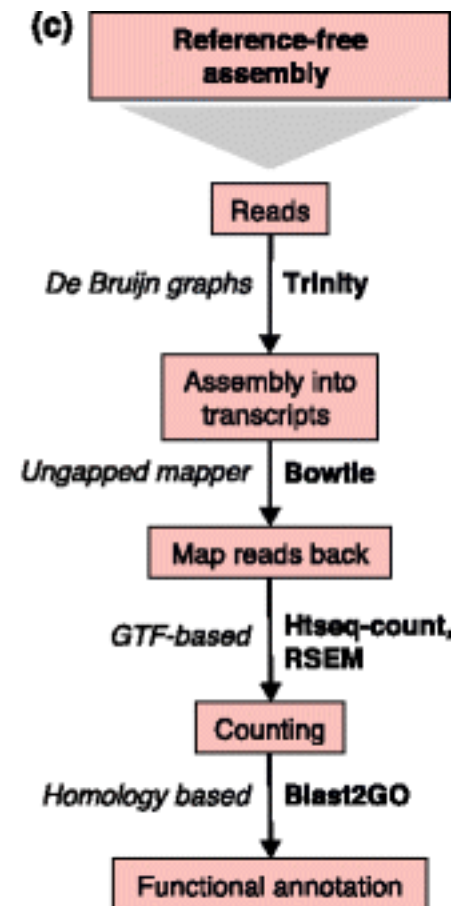
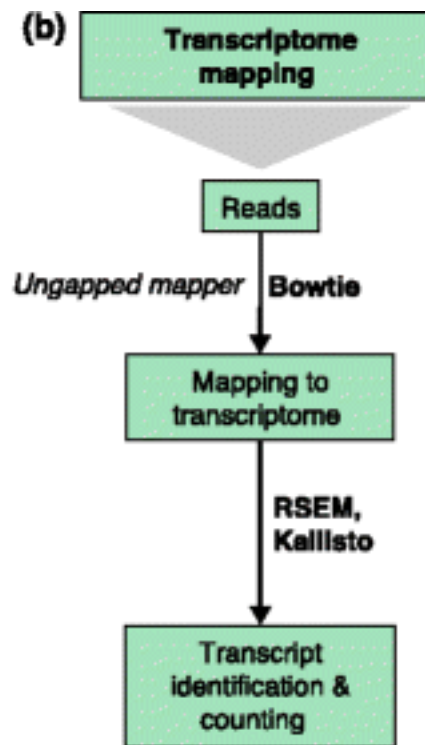
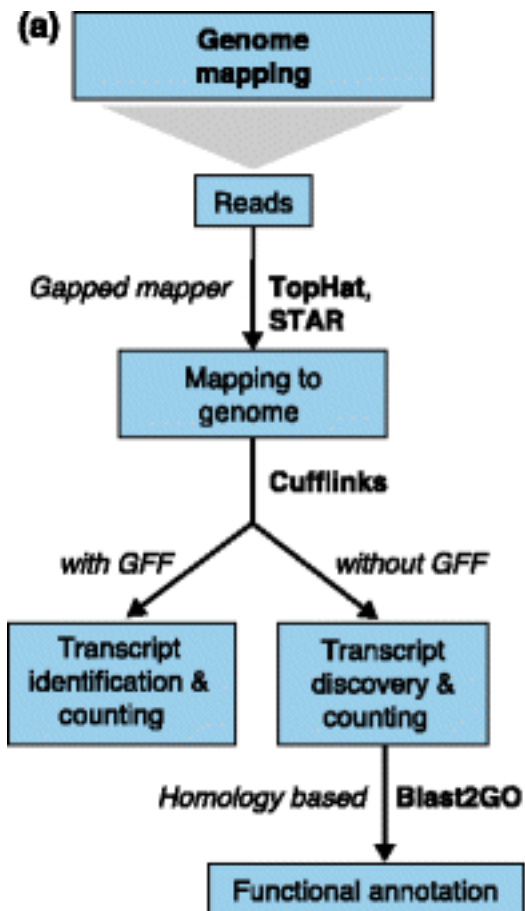


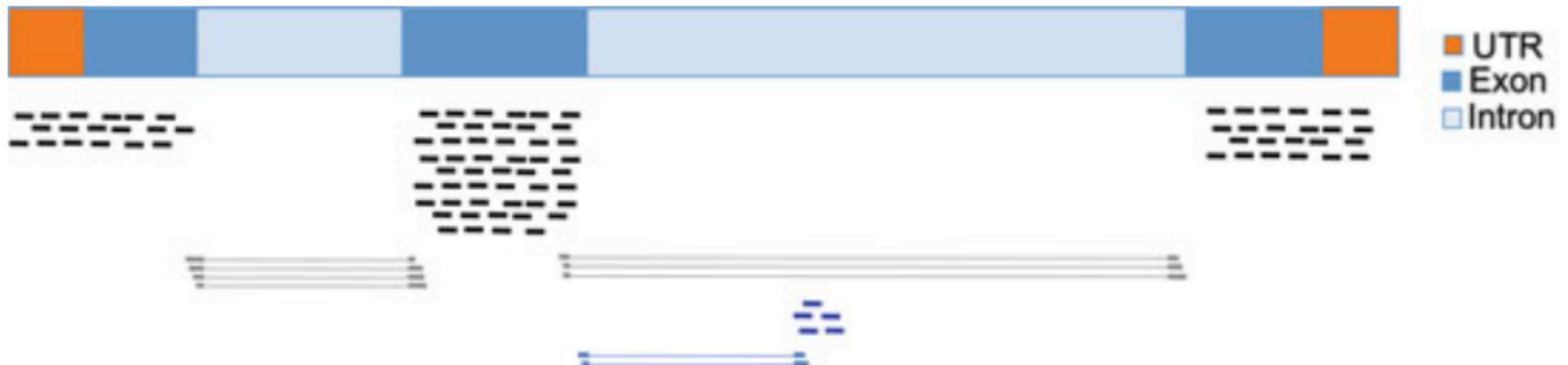
RNASEQ - DIFFERENTIAL EXPRESSION STATISTICS

Outline



Counting

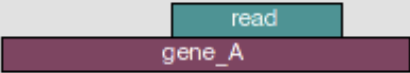
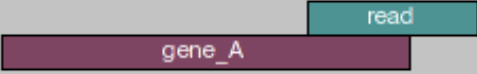


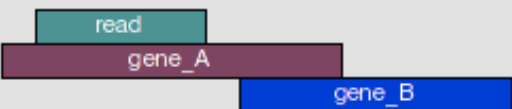

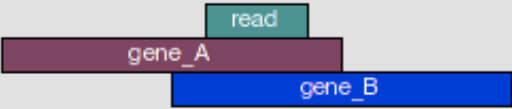
- From read alignments, you next need to summarize the reads into a “table of counts”
- Sounds easy. Its not!
- Gene-level counting vs.
- Transcript-level counting



HTSeq

- A Python framework to work with high-throughput sequencing data
- Comes with some very useful scripts, including one to count aligned reads
- Allows users to select from a number of counting strategies
- Can select any feature type from a gff3 file to be the “counting unit”
- Categorizes reads:
 - Maps to a feature
 - Does not map to a feature
 - Is ambiguous (could map to more than one feature)
 - Is too low quality
 - Is not aligned at all
 - Alignment is not unique

HTSeq

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Transcript level quantification

- (Isoform quantification)
- More complicated to compute, but gives information about alternative splicing
- Allocate multi-mapping reads among the possible transcripts. How?
- Requires statistical inference
 - Bayesian methods
 - Expectation maximization (EM)
- Many possible software options:
 - RSEM, kallisto, salmon
- Active area of research

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.

From counts to differential expression statistics

Software

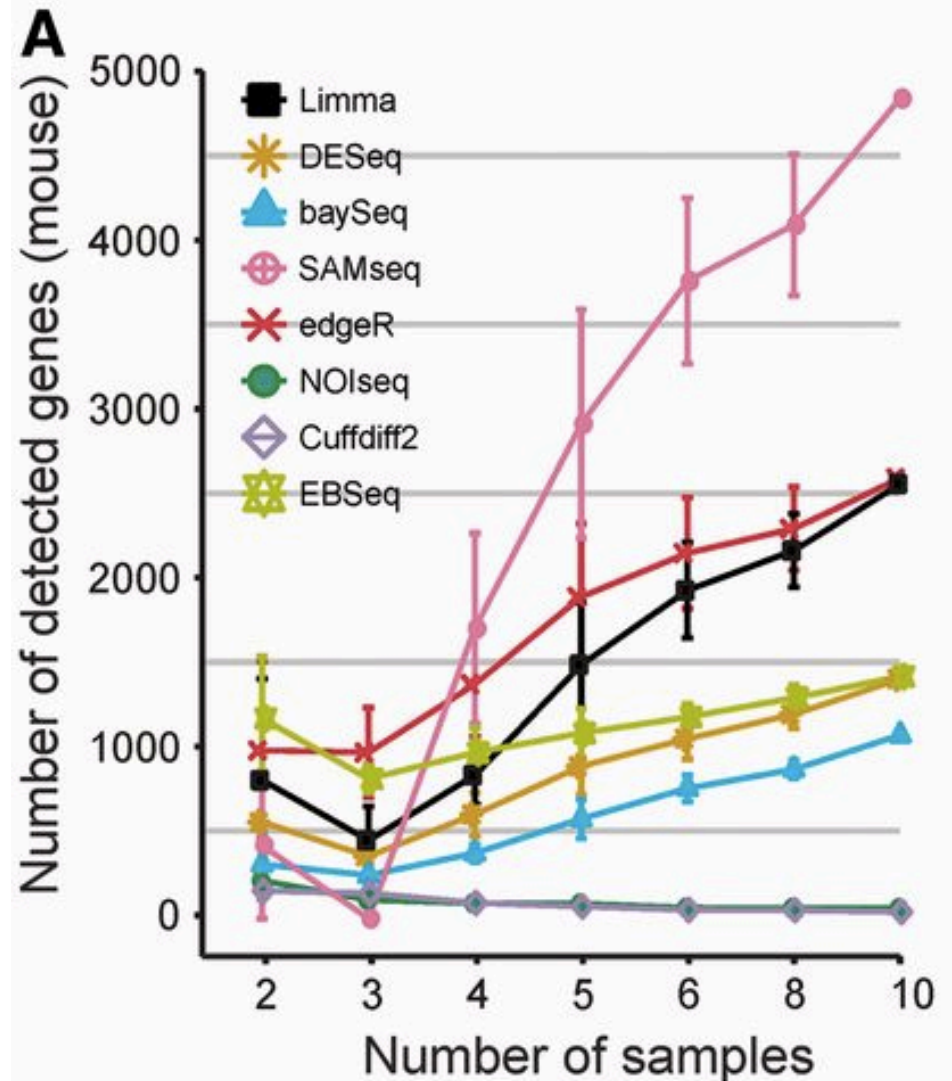
- Cuffdiff2
- Sleuth <- allows for isoform level abundance analysis
requires use of kallisto for read quantification

- DESeq
- DESeq2
- EdgeR
- SAMSeq



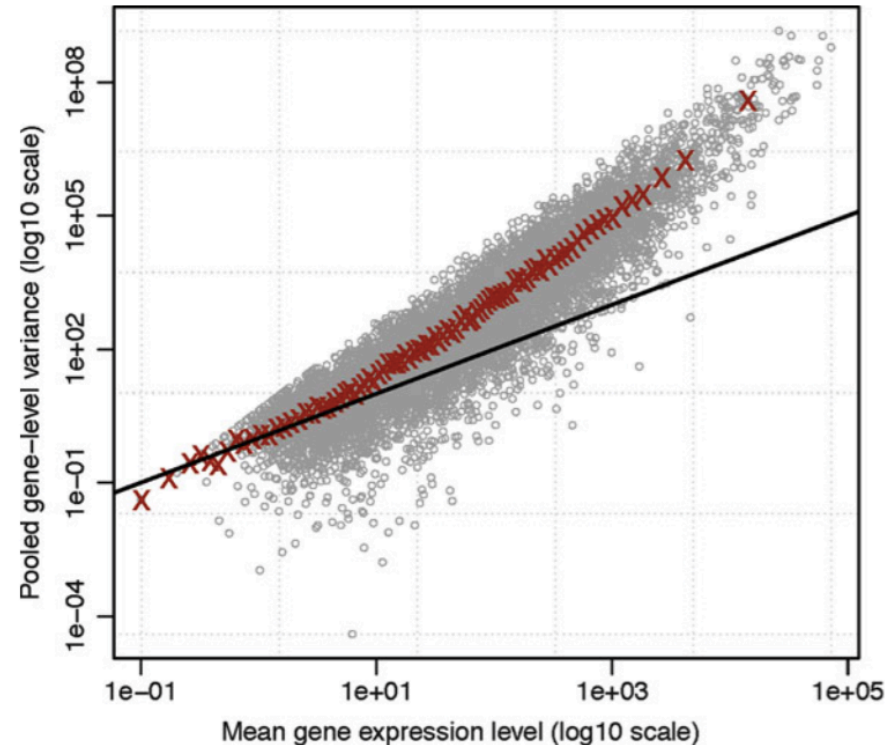
Different approaches give different results

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



Statistical Models

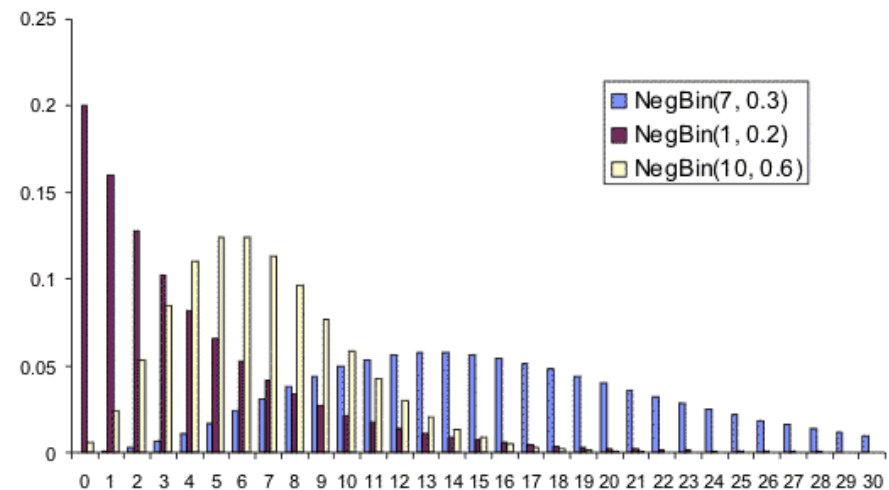
- Originally (pre-2010) Poisson distribution was used to model counts
- Data was found to be over-dispersed
- Negative binomial now preferred



Mean–variance plot for the Parikh et al. *Dictyostelium* dataset (Parikh et al. 2010). The variability in this biologically replicated RNA-seq dataset exhibits prominent extra-Poisson variability.

Negative Binomial

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



Why do we need a distribution?

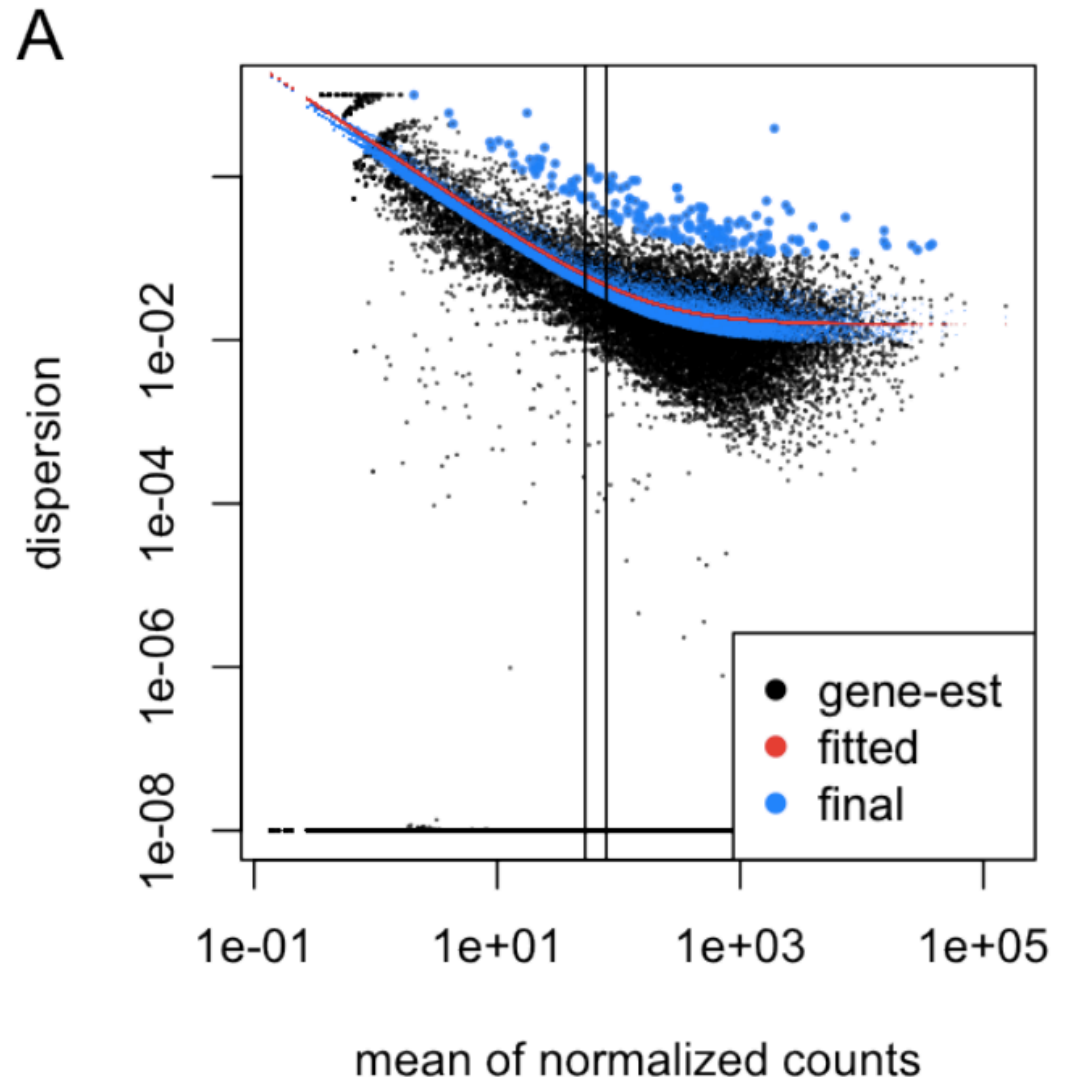
- Why not just use a non-parametric method?
 - More difficult to show significance with a non-parametric method with few replicates
- Rank order statistics will begin working well with ~ 10 biological replicates
 - SamSeq (<http://www.inside-r.org/packages/cran/samr/docs/SAMseq>)
 - NOIseq (<https://www.bioconductor.org/packages/release/bioc/html/NOISeq.html>)

Modeling dispersion

- Now we have a distribution that allows the dispersion to be different from the mean
- But we often still have very low sample numbers ($n = 2, 3, 4$), which is not good for modeling variance
- A variety of ways to handle this – usually share information across genes to measure variance
- Both DESeq2 and EdgeR assume that genes of similar average expression strength have similar dispersion
 - Use this information in slightly different ways to predict reasonable dispersions

DESeq2 Strategy for Dispersion Estimation – Empirical Bayes

1. Estimate dispersion for each gene
2. Plot and fit a curve
3. Adjust dispersion parameter toward curve (shrinking)



DESeq2 - Test for differential expression

- Accepts only raw counts
- Null hypothesis: the expression change in a gene is 0
- 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(~ set + condition  
+ set: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

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
- once differentially expressed contigs have been identified, interpretation is difficult, as many genes will be present in the list multiple times

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

- Corset
- RapClust

