



3000788 Intro to Comp Molec Biol

Lecture 11: RNA sequencing data processing

September 21, 2023



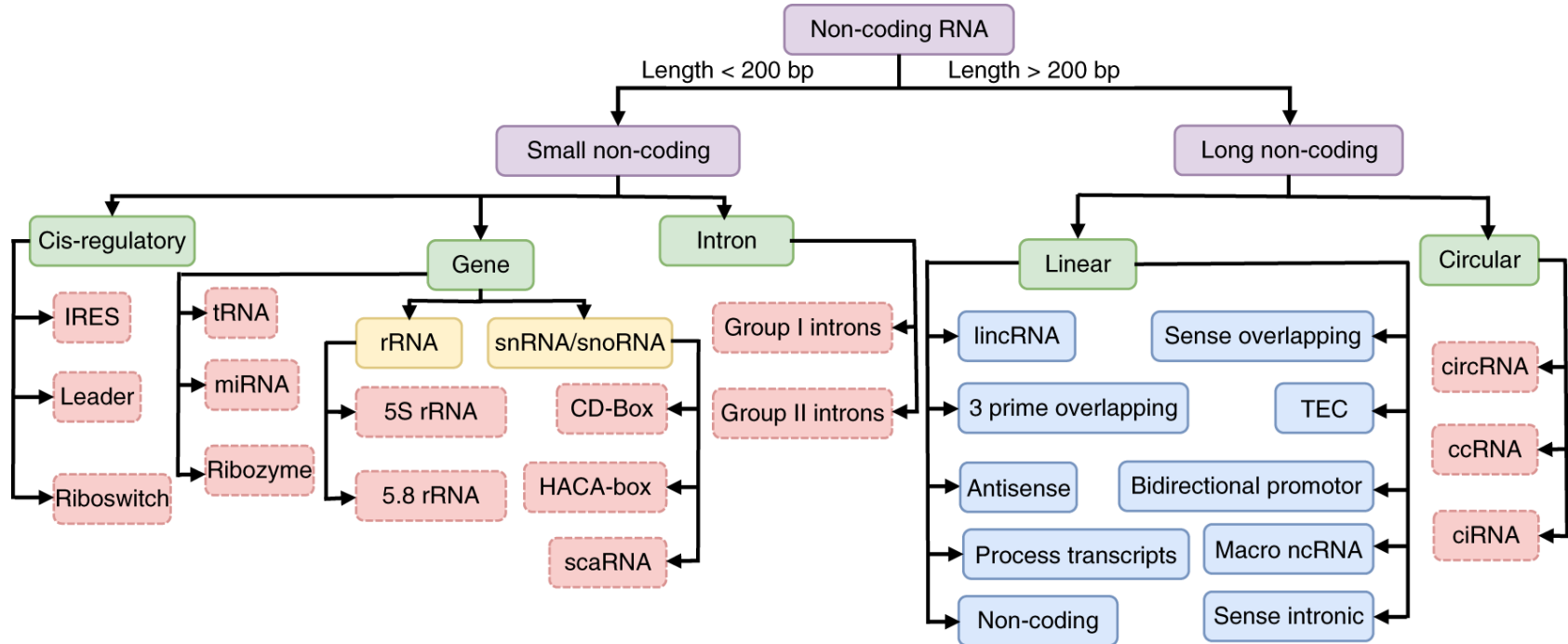
Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

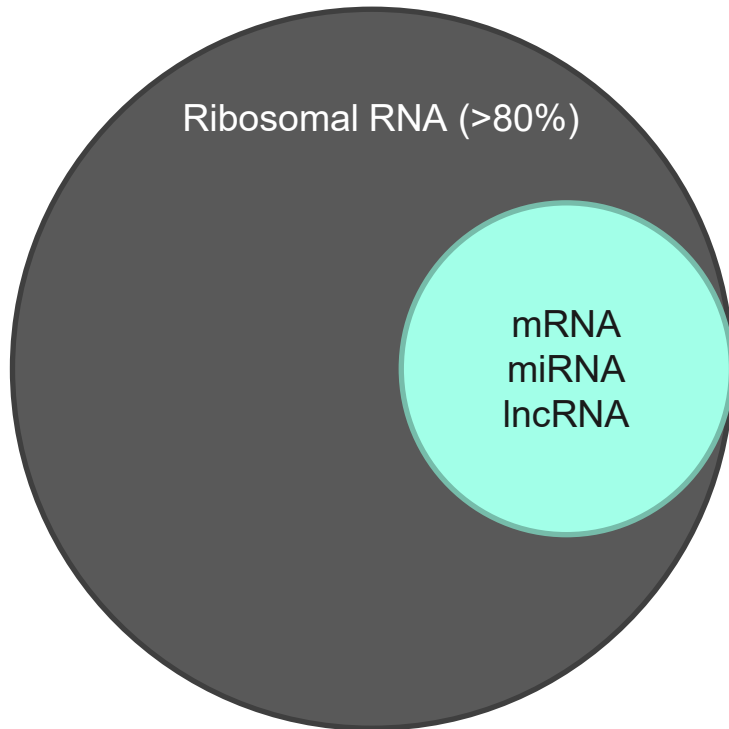


RNA-seq techniques

Non-coding RNAs



Total RNA sequencing

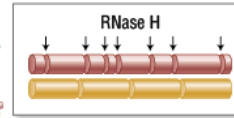
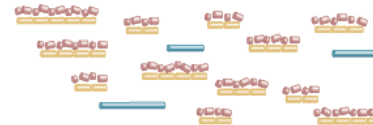


Binding of ssDNA Probes



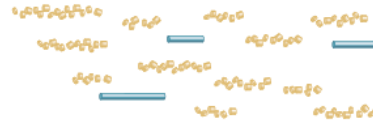
Single-stranded DNA probes hybridize specifically to rRNA molecules.

rRNA Degradation by Ribonuclease H (RNase H) Enzyme



RNase H degrades the hybridized RNA (rRNA).

Probe Degradation by DNase I Enzyme & Clean Up



DNase I degrades the DNA probes.

rRNA-depleted RNA

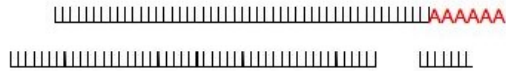


Non-rRNA species (blue) are enriched.

Source: New England BioLabs

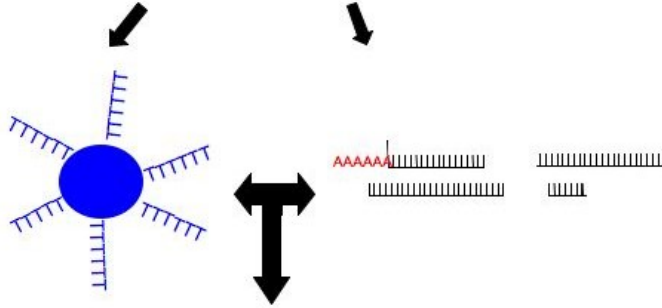
mRNA and miRNA sequencing

Isolate Total RNA

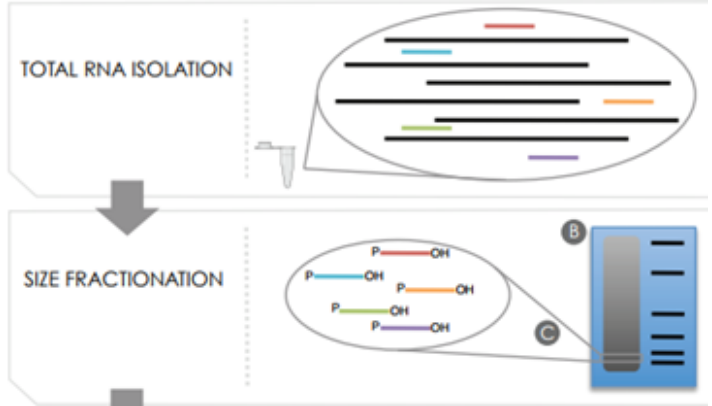
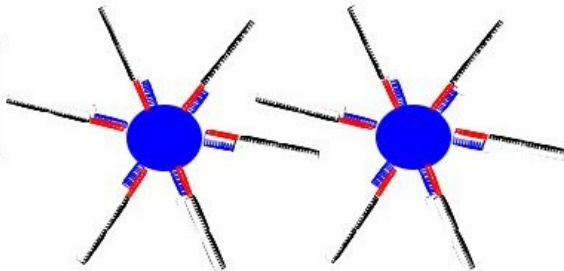


Fragmentation
and/or Isolation

In this case, isolation via Poly(T)
coated magnetic beads

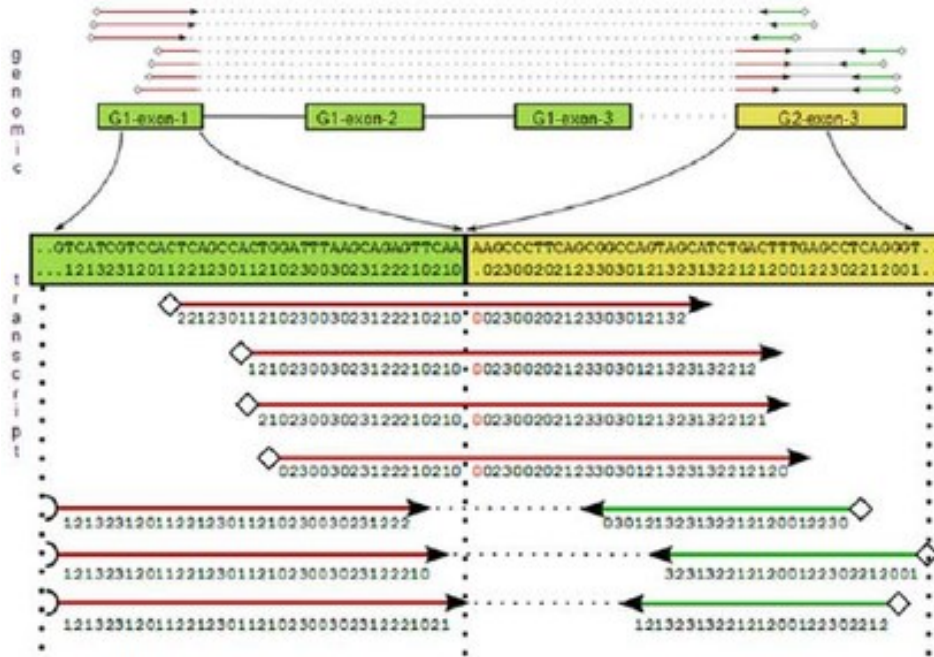


Poly(A) RNA molecules
bind to the Poly(T)
magnetic beads



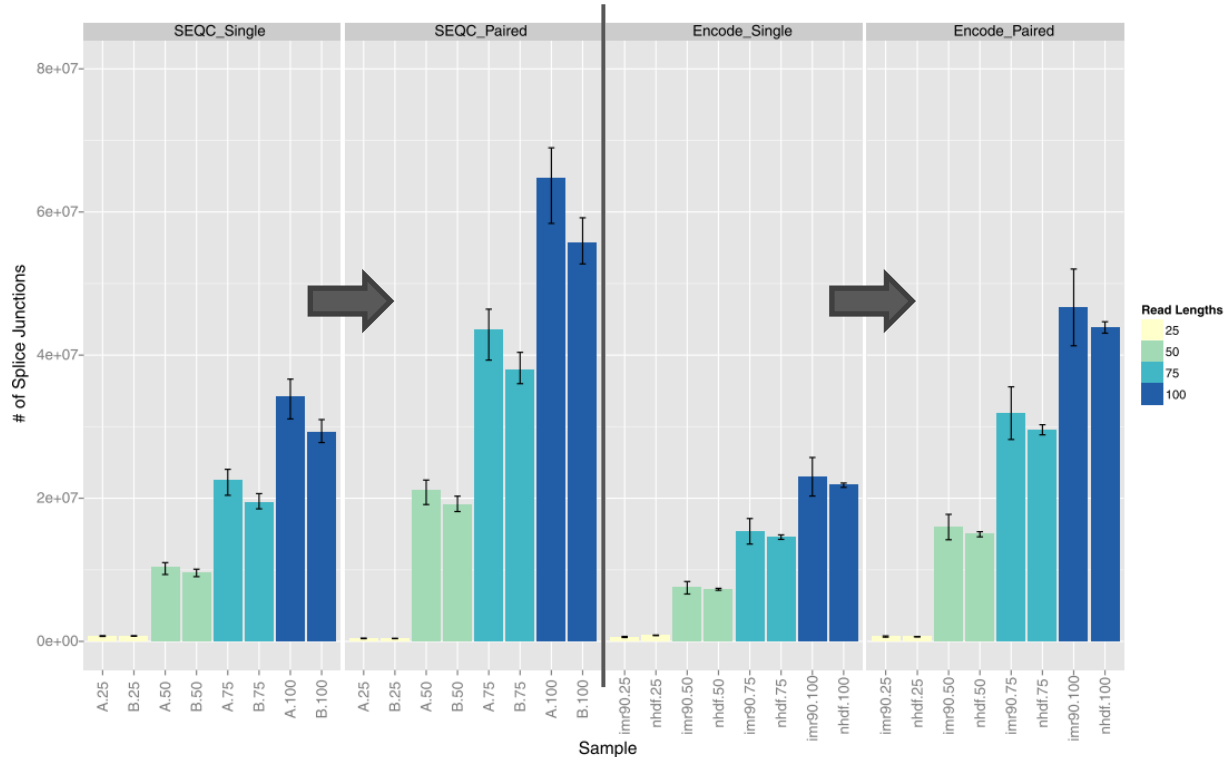
- Selection by polyT probe
- Size fractionation

Transcript isoform detection



- **Single-end sequencing:** The read must span the exon junction
- **Paired-end sequencing:** As long as the forward and reverse reads came from different exons

Impact of paired-end sequencing and read length



Full-length transcript sequencing



Choosing RNA-seq technique



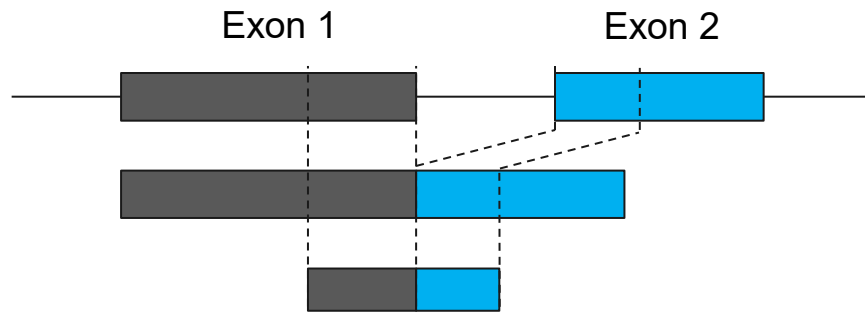
- miRNA or mRNA or total-RNA
- Single-end or paired-end
 - Single-end is ok for gene-level quantification
 - Paired-end is needed to distinguish isoforms
- Illumina or 3rd generation sequencer
 - Long-read data is helpful for genes with complex isoforms and for detecting novel isoforms



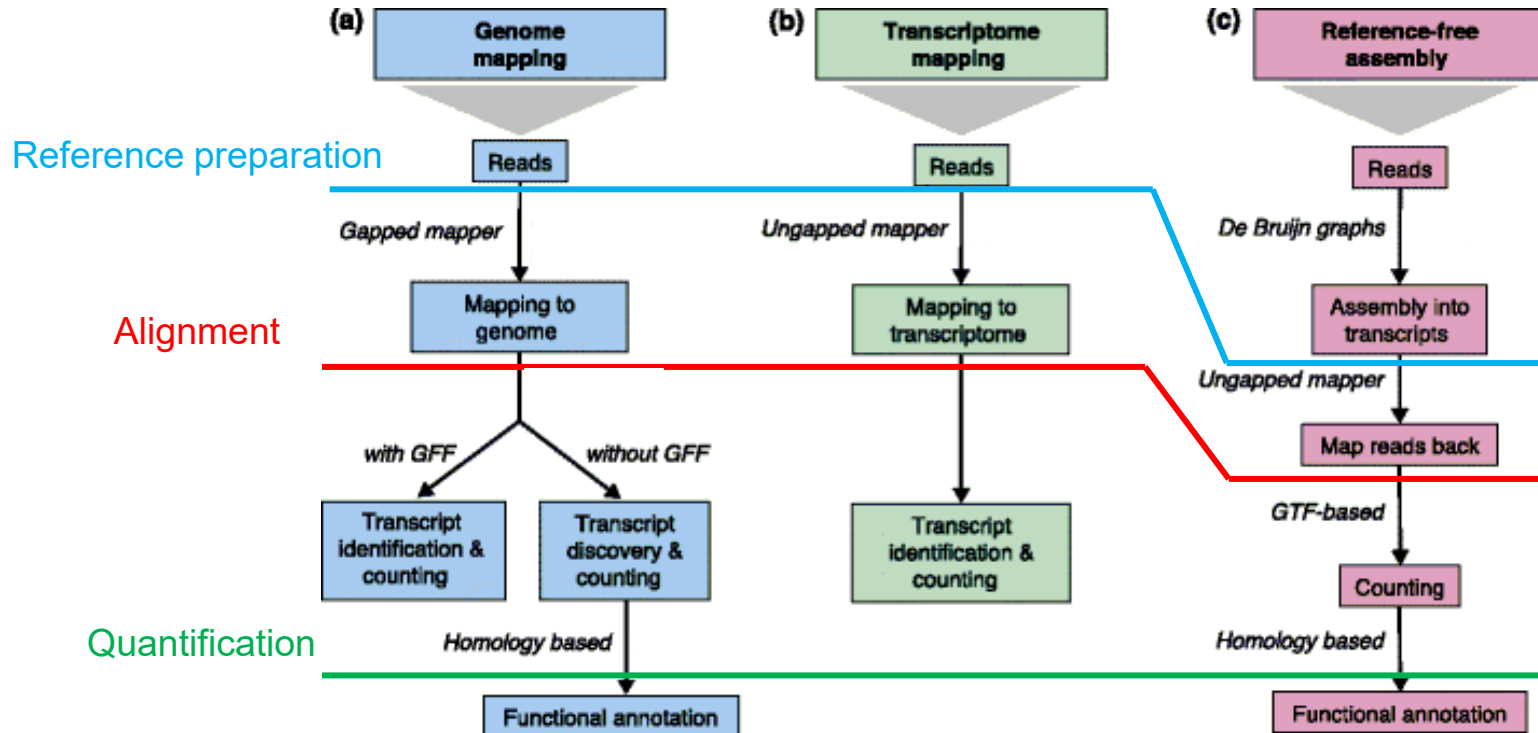
RNA-seq analysis

Three primary pipelines

- Reference-free
 - Novel species, rely on *de novo* assembly
- Reference transcriptome
 - **Fast**, cannot discover new isoform
 - Ungapped, *k*-mer-based alignment
- Reference genome
 - **Slow**, but can detect new isoforms
 - Gapped alignment

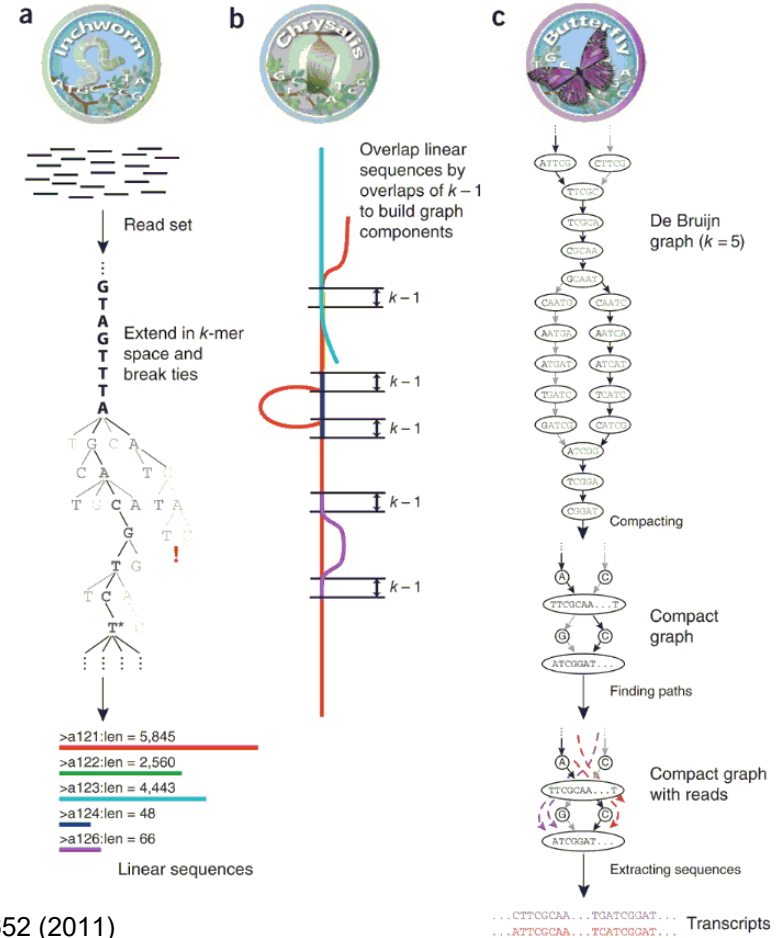


Pipeline overview

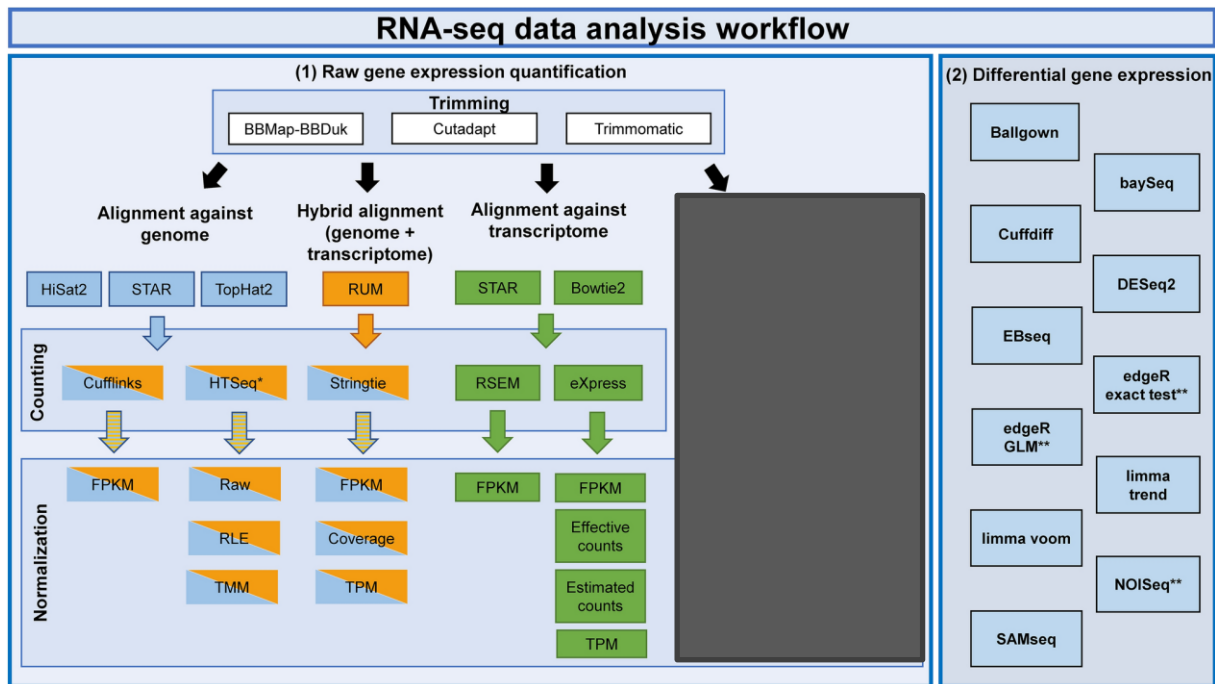
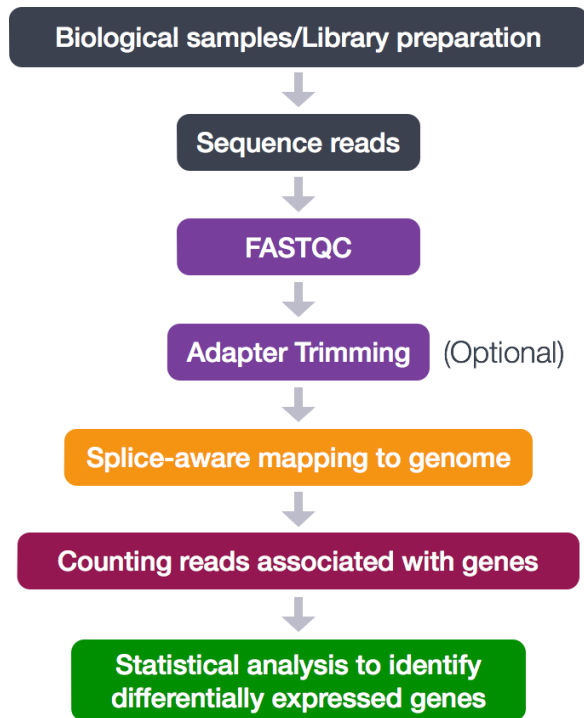


De novo transcript assembly

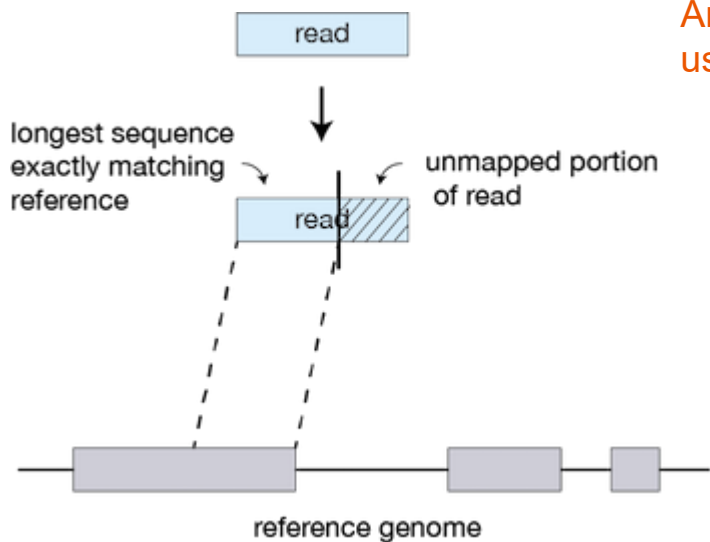
- Similar to genome assembly
 - Connect overlapping reads into **contigs**
- Handle the existence of isoforms
 - Cluster **contigs with shared sequences**
 - Generate all possible isoforms
- Used as database for downstream analysis



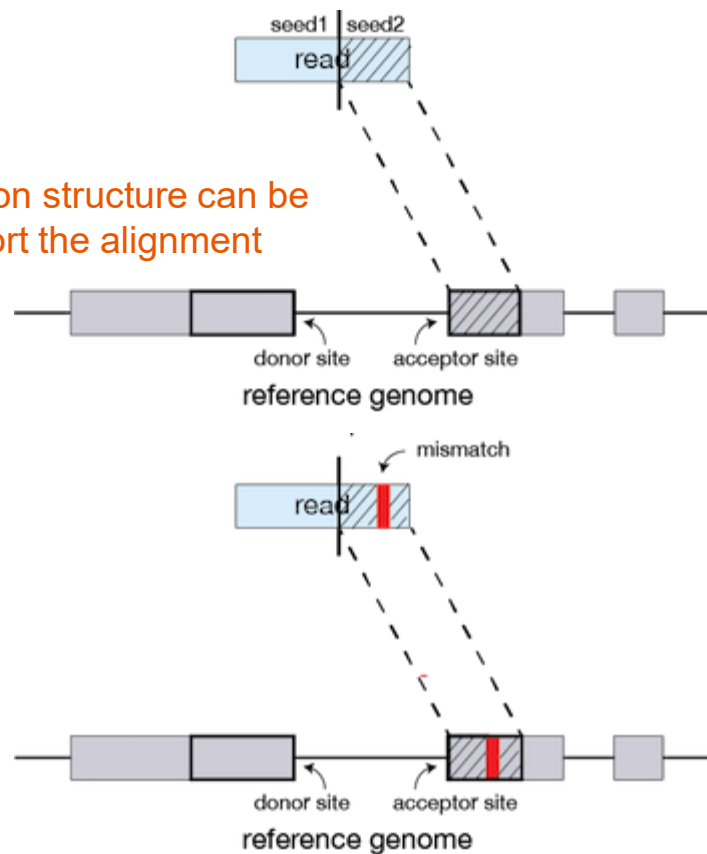
Alignment-based pipelines



Gapped alignment



Annotated exon structure can be used to support the alignment



GTF/GFF genome annotation format



Sample GTF output from Ensembl data dump:

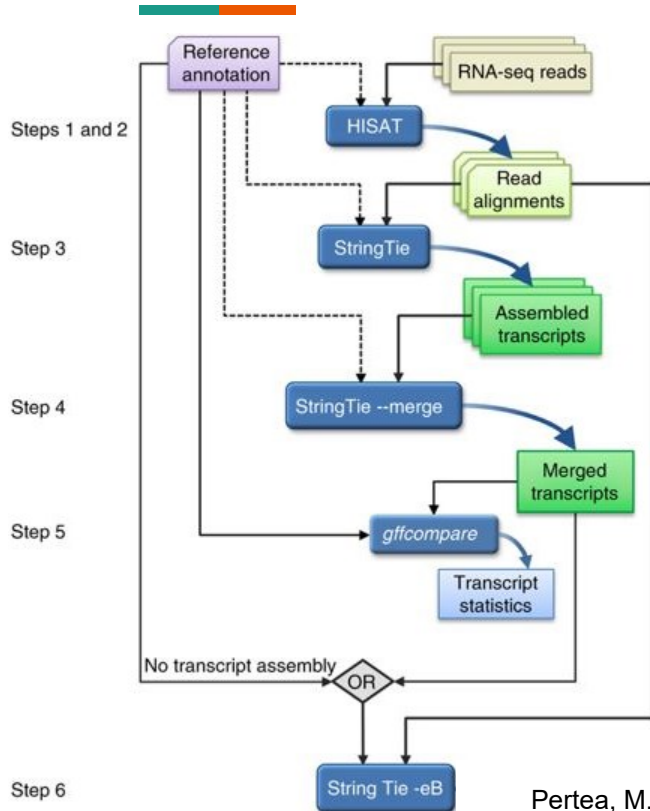
```
1 transcribed_unprocessed_pseudogene   gene      11869 14409 . + . gene_id "ENSG00000223972"; gene_name "DDX11L1"; gene_source "havana";
1 processed_transcript                  transcript 11869 14409 . + . gene_id "ENSG00000223972"; transcript_id "ENST00000456328"; gene_name
```

Sample GFF output from Ensembl export:

X	Ensembl Repeat	2419108	2419128	42	.	.	hid=trf; hstart=1; hend=21
X	Ensembl Repeat	2419108	2419410	2502	-	.	hid=AluSx; hstart=1; hend=303
X	Ensembl Repeat	2419108	2419128	0	.	.	hid=dust; hstart=2419108; hend=2419128
X	Ensembl Pred.trans.	2416676	2418760	450.19	-	2	genscan=GENSCAN00000019335
X	Ensembl Variation	2413425	2413425	.	+	.	
X	Ensembl Variation	2413805	2413805	.	+	.	

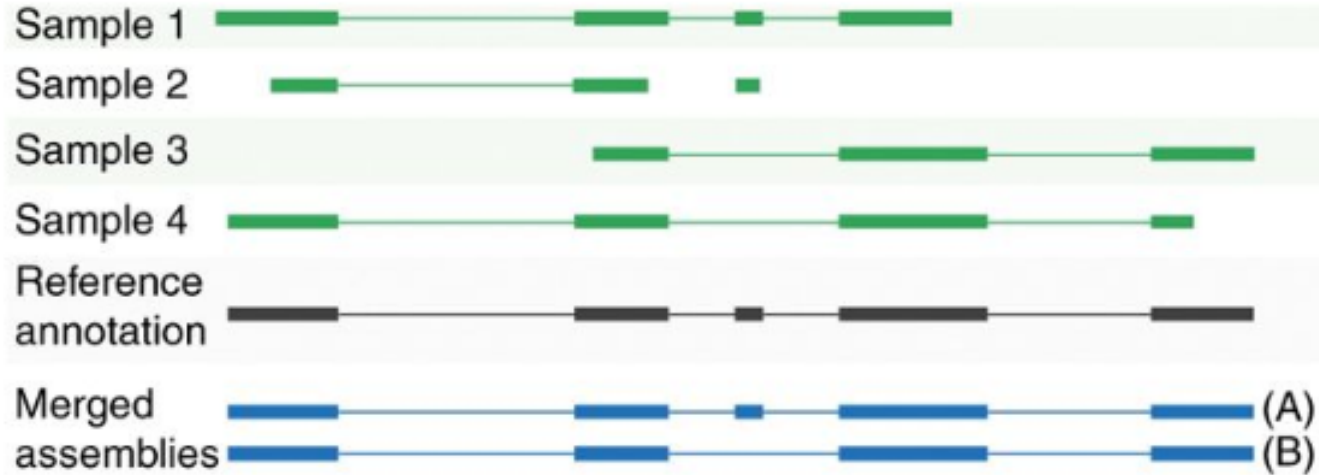
- Tab-separated text file
- Chromosome ID, object name, base pair positions, strand, and other annotation details

Multi-step post-alignment processing



- Initial alignment
- Assemble potential novel isoforms
- Merge isoforms across samples
- Re-quantify isoform abundances using the merged database of isoforms

Importance of merging isoforms



Pertea, M. et al. Nature Protocols 11:1650-1667 (2016)

- Rare isoform may be missing in some samples
- Reads can get misinterpreted if the correct isoform is not in the reference

GTF with abundance annotation



chr1	StringTie	transcript	36534	36849	1000	.	.	gene_id "STRG.1"; transcript_id "STRG.1.1"; cov "19.614035"; FPKM "6.688056"; TPM "10.944590";
chr1	StringTie	transcript	35245	36073	1000	-	.	gene_id "STRG.2"; transcript_id "STRG.2.1"; reference_id "ENST00000461467.1"; ref_gene_id "ENSG00000237613.2";
								ref_gene_name "FAM138A"; cov "0.327684"; FPKM "0.111735"; TPM "0.182847";
chr1	StringTie	transcript	52473	53312	1000	+	.	gene_id "STRG.3"; transcript_id "STRG.3.1"; reference_id "ENST00000606857.1"; ref_gene_id "ENSG00000268020.3";
								ref_gene_name "OR4G4P"; cov "0.119048"; FPKM "0.040593"; TPM "0.066429";
chr1	StringTie	transcript	137682	137965	1000	-	.	gene_id "STRG.4"; transcript_id "STRG.4.1"; reference_id "ENST00000595919.1"; ref_gene_id "ENSG00000269981.1";
								ref_gene_name "RP11-34P13.16"; cov "0.000000"; FPKM "0.000000"; TPM "0.000000";
chr1	StringTie	transcript	139283	139642	1000	.	.	gene_id "STRG.5"; transcript_id "STRG.5.1"; cov "3.111111"; FPKM "1.060837"; TPM "1.735993";

- Different tool outputs transcript abundance in different format
- GTF can accommodate abundance annotation in the last columns
 - Coverage (cov) = fraction of transcript length with mapped read
 - FPKM = **F**ragment **p**er **k**ilobase of exon per **m**illion reads mapped
 - TPM = **T**ranscript **p**er **m**illion

Units for transcript abundance

$$\text{FPKM} = \frac{\text{Read Count}}{\frac{\text{Transcript Length}}{1,000} \times \frac{\text{Total Read Count}}{1,000,000}}$$

Long transcript generates more fragments and more read counts

Experiment with higher sequencing depth generates more read counts

Similar to percentage (but per million)

$$\text{TPM} = \frac{\text{FPKM}}{\sum \text{FPKM}} \times 1,000,000$$

- Read count (number of mapped reads)
- FPKM = **F**ragment **p**er **k**ilobase of exon per **m**illion reads mapped
- TPM = **T**ranscript **p**er **m**illion

Alignment-based pipeline summary

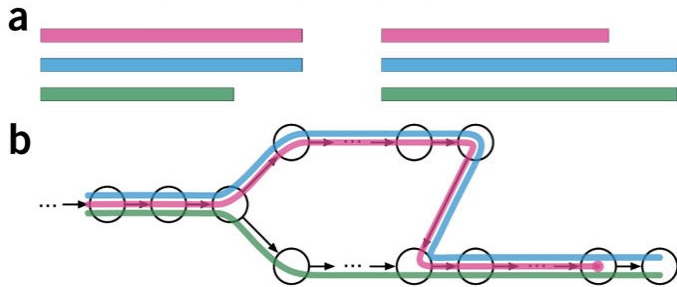


- Initial alignment to reference genome (with annotated gene structure)
 - STAR / HISAT2
- [Optional]
 - Identify novel isoforms
 - Merge isoforms across samples
- Quantify transcript abundances
 - Read count / FPKM / TPM
 - StringTie2 / htseq-count



k-mer pseudoalignment

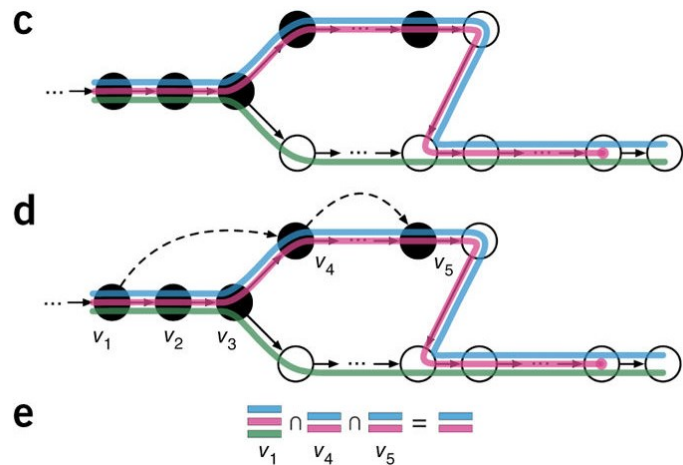
k -mer database for transcriptome



Bray *et al.* Nat Biotech 34:525-527 (2016)

- Create de Bruijn graph with k -mer as nodes
- Map node to transcripts with that k -mer
- Contig = a path on de Bruijn graph that mapped to the same transcript

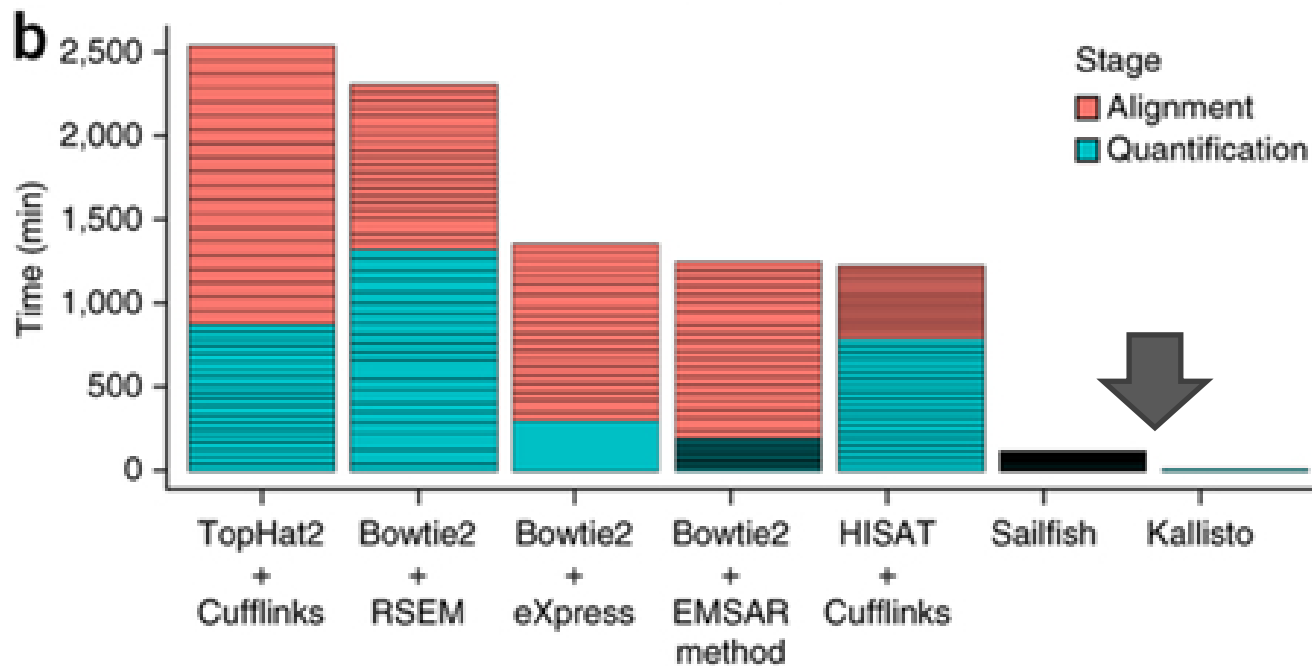
k -mer pseudoalignment



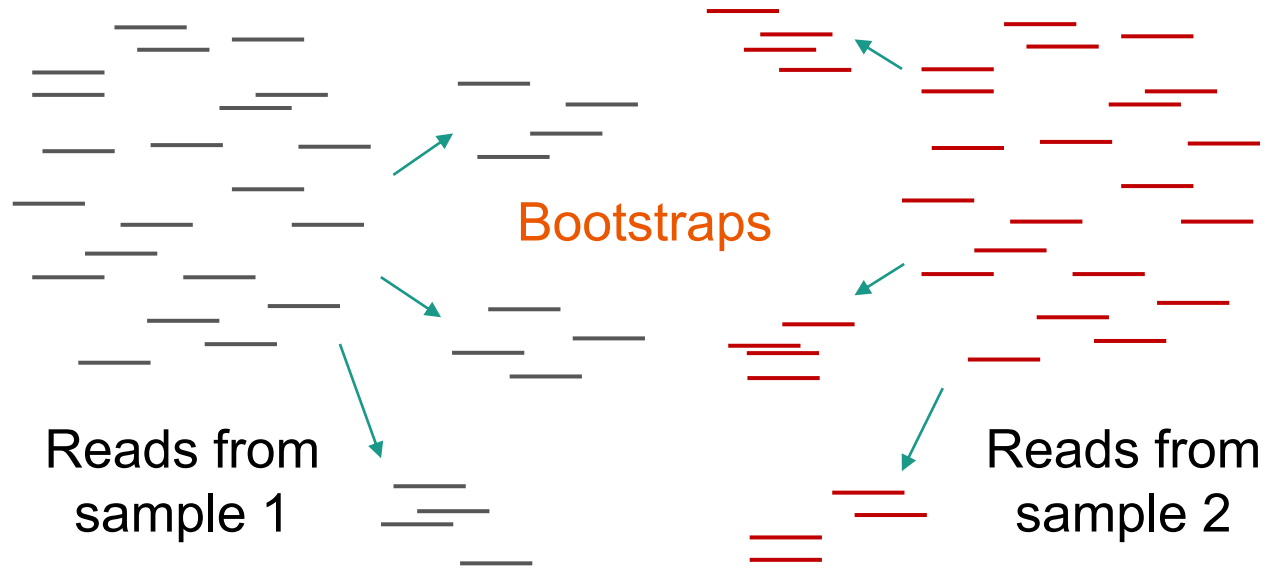
Bray *et al.* Nat Biotech 34:525-527 (2016)

- For a new read, all of its k -mers are mapped against contigs
 - Ignore the ordering of k -mers on the read
- Report only contigs that are compatible with all k -mers
- Speed up by skipping uninformative k -mer
 - V_2 and V_3 regions
 - Only 2-4 k -mer lookups are enough

>100 fold speed up with pseudoalignment

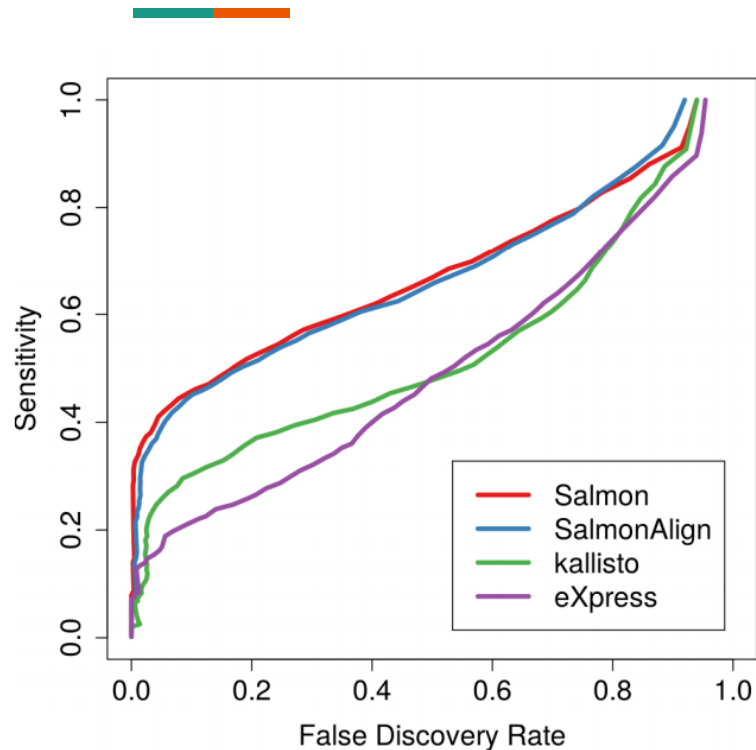


Bootstrapping enabled by pseudoalignment



- Bootstrapping estimates technical variances

Salmon: improved k -mer alignment

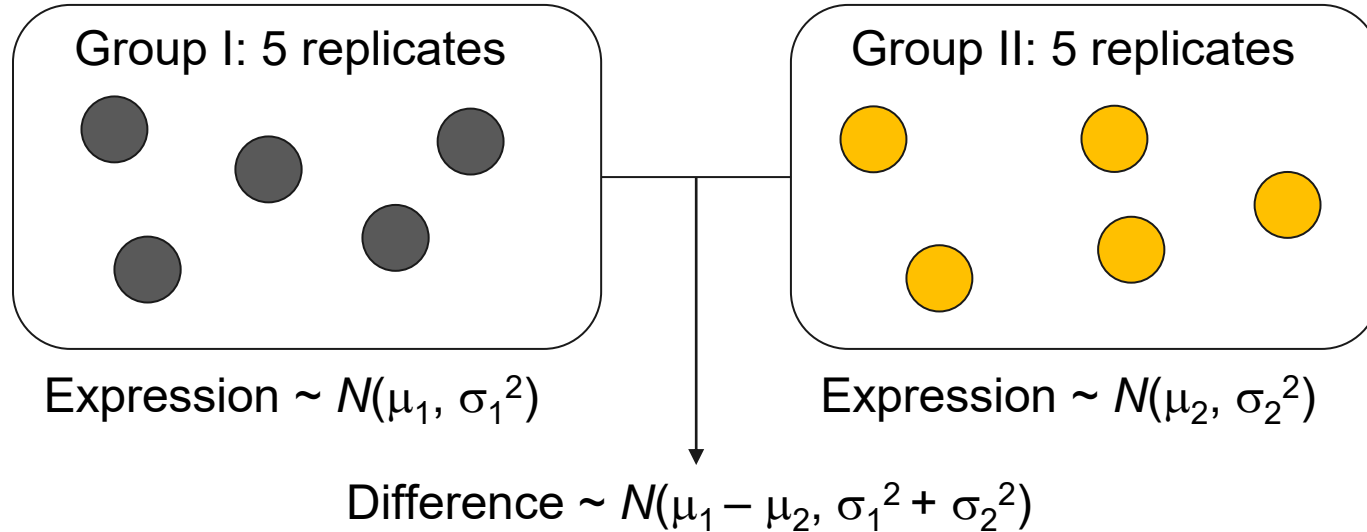


- Also track the location of k -mer on the input read → semi-alignment
- Correct quantification based on GC content and 3' / 5' amplification biases



Differential expression analysis

DE for microarray & nanostring



- Simple t -test for normally distributed abundance data

DE as nested model testing / likelihood ratio test



Hypothesis 1



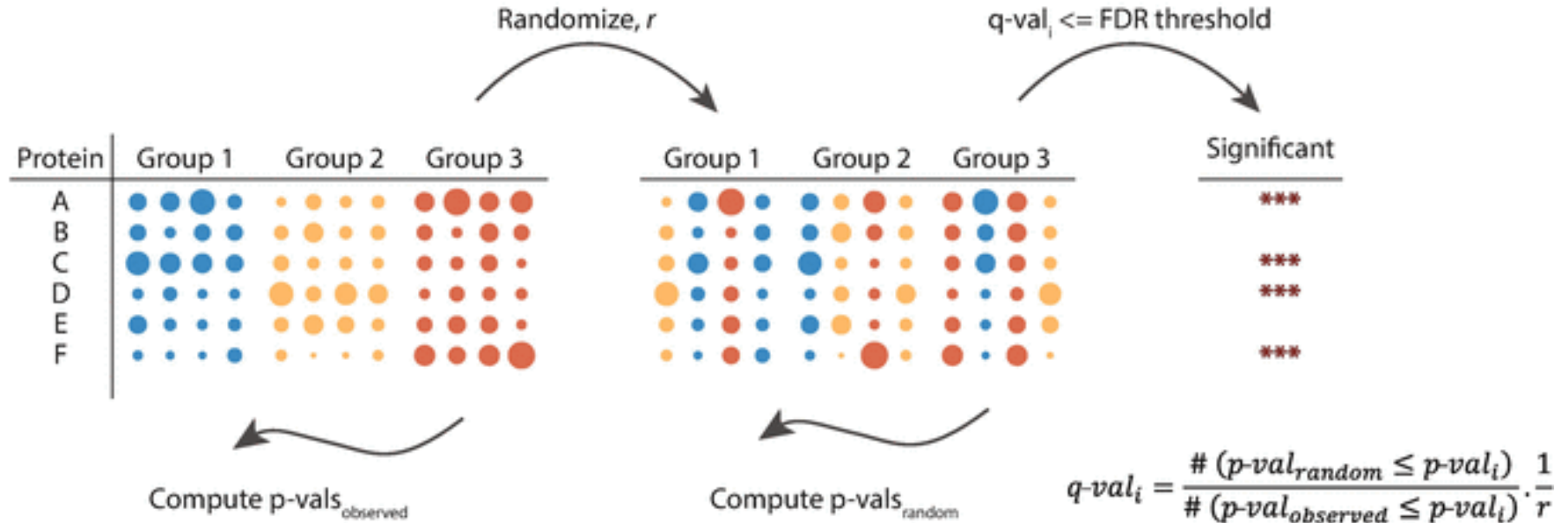
Hypothesis 2



Hypothesis 3



DE as permutation test



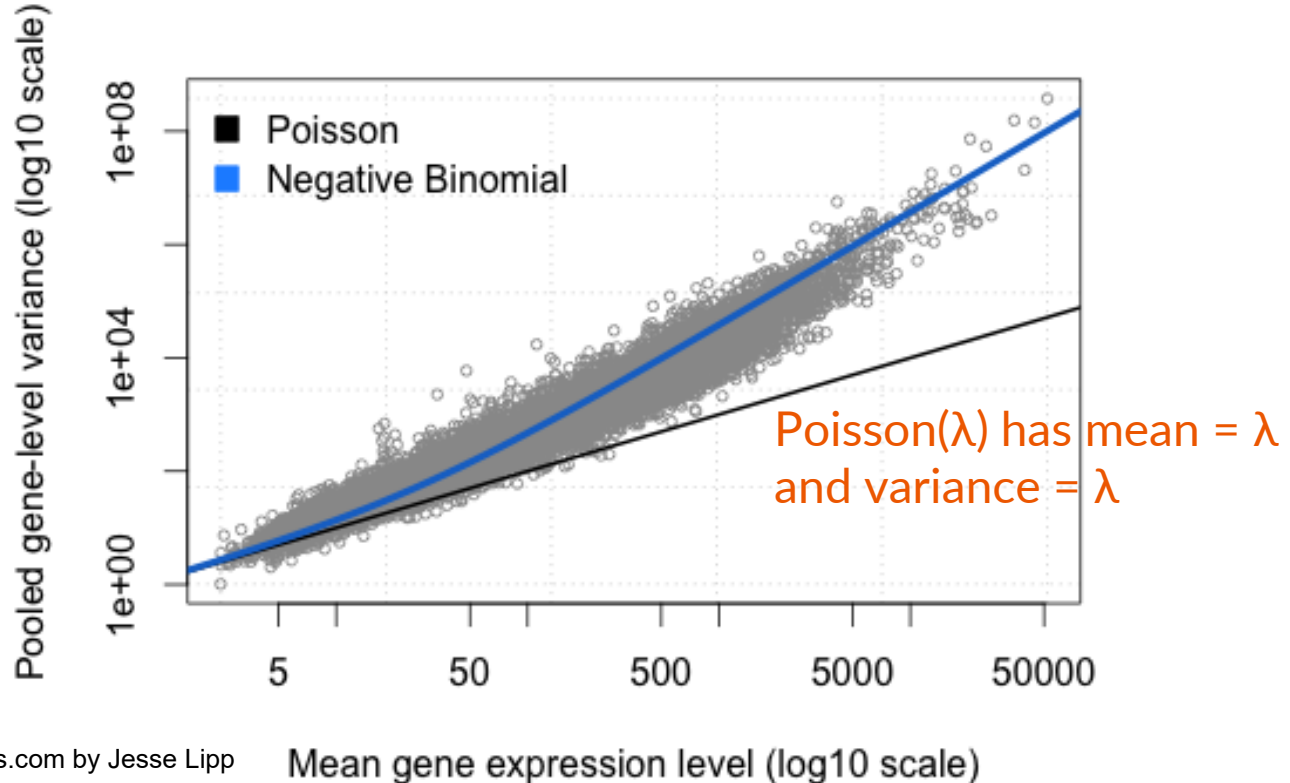
Tyanova and Cox. Cancer Systems Biology pp 133-148. (2018)

- Permuting sample labels = remove condition-specificity



DESeq2 model for read count

The distribution of RNA-seq read count



Negative binomial model

- $NB(r, p)$ = the number of failures that we will see in a series of Bernoulli trials with probability of success p until we obtain r successes
 - $X O O X X X O O X O O$ = 5 failures until 6 successes
- $P_{NB}(k; r, p) = \binom{k+r-1}{k} (1-p)^k p^r$
 - $k + r - 1$ locations to place k failures (the last location must be success)
- Mean = $\frac{pr}{(1-p)}$
- Variance = $\frac{pr}{(1-p)^2} = \frac{pr}{(1-p)} \frac{1}{(1-p)} = \frac{pr}{(1-p)} \left(1 + \frac{p}{1-p}\right)$
$$= \frac{pr}{(1-p)} + \frac{p^2 r}{(1-p)^2} = \frac{pr}{(1-p)} + \left(\frac{pr}{(1-p)}\right)^2 \frac{1}{r}$$

Another view of negative binomial model



- $P_{\text{NB}}(k; r, p) = \int_0^\infty P_{\text{Poisson}(\lambda)}(k) \cdot P_{\text{Gamma}\left(r, \frac{1-p}{p}\right)}(\lambda) d\lambda$
- Negative binomial distribution is a continuous mixture of Poisson distribution, with mixing weights Gamma-distributed
 - Same as Gamma site-specific mutation rates
- Bulk gene expression is an average over many cells
- Mixture of read counts from multiple cells, each following $\text{Poisson}(\lambda)$

DESeq2 model of gene expression



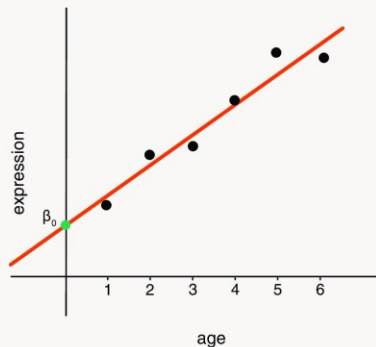
- Read count $K_{i,j} \sim \text{NB}(\mu_{i,j}, \sigma_{i,j}^2)$, for gene i from sample j
 - $\mu_{i,j}$ = sample effects x sequencing effects
 - $\sigma_{i,j}^2 = \mu_{i,j} + \text{gene-specific effects} \times \mu_{i,j}^2$
- Sample effects
 - Control / Treatment
 - Confounding factors: age, time after treatment, etc.
 - Log FC = $\sum_r x_{j,r} \beta_{i,r}$ where $x_{j,r}$ are design parameters for sample j and $\beta_{i,r}$ are the effect sizes
 - Linear effect model

Linear effect models

Covariates: quantitative measurements (e.g. age)

Regression model

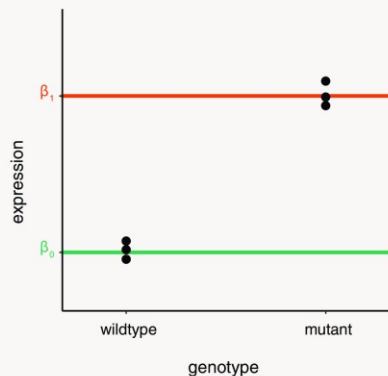
$$\text{expression} = \beta_0 + \beta_1 \text{age}$$



Factors: categorical variables (e.g. genotype)

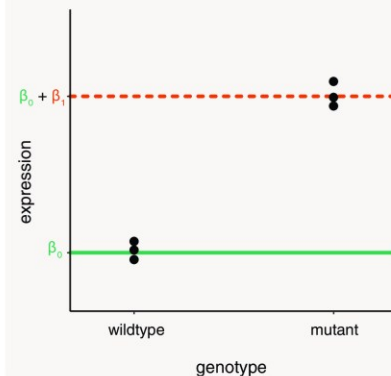
Means model

$$\text{expression} = \beta_1 \text{wildtype} + \beta_2 \text{mutant}$$



Mean-reference model

$$\text{expression} = \beta_1 + \beta_2 \text{mutant}$$



Legend

- Original data points

— Expected gene expression
(based on model)

- - - Expected gene expression
(of non-reference levels in mean-reference model)

Law, C.E. et al. F100Res 9:1444 (2020)

Linear model for multiple effects

Model

$$E(y) = 1.03 + 1.09x_1 + 1.97x_2 + 0.82x_1x_2$$

$$E(y) = 1.03 = 1.03 \quad (\text{for control})$$

$$E(y) = 1.03 + 1.09 = 2.12 \quad (\text{for treatment I})$$

$$E(y) = 1.03 + 1.97 = 3.00 \quad (\text{for treatment II})$$

$$E(y) = 1.03 + 1.09 + 1.97 + 0.82 = 4.90 \quad (\text{for treatments I \& II})$$

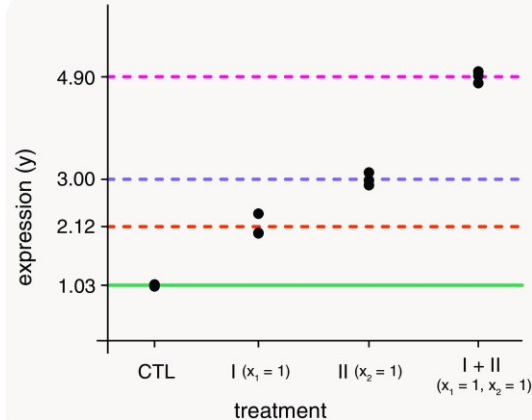
Matrix

```
> model.matrix(~treat1 * treat2)
```

$$\begin{array}{c}
 \text{(Intercept)} \\
 \text{treat1YES} \\
 \text{treat2YES} \\
 \text{treat1YES:} \\
 \text{treat2YES}
 \end{array}
 \begin{array}{c}
 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12
 \end{array}
 \begin{pmatrix}
 1 & 0 & 0 & 0 \\
 1 & 0 & 0 & 0 \\
 1 & 0 & 0 & 0 \\
 1 & 1 & 0 & 0 \\
 1 & 1 & 0 & 0 \\
 1 & 1 & 0 & 0 \\
 1 & 0 & 1 & 0 \\
 1 & 0 & 1 & 0 \\
 1 & 0 & 1 & 0 \\
 1 & 1 & 1 & 1 \\
 1 & 1 & 1 & 1 \\
 1 & 1 & 1 & 1
 \end{pmatrix}$$

Plot

Law, C.E. et al. F100Res 9:1444 (2020)



DESeq2 model of gene expression



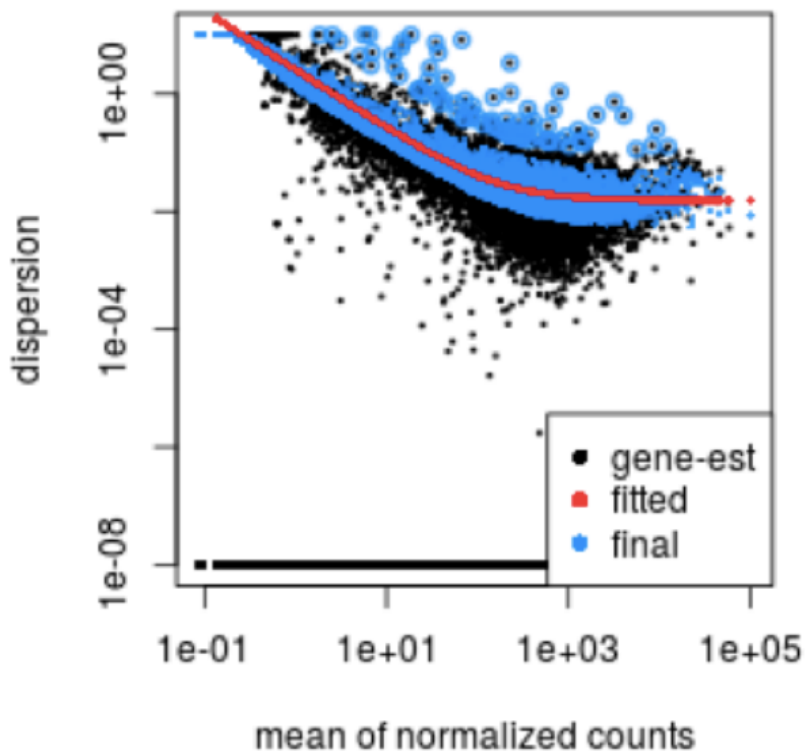
- Read count $K_{i,j} \sim \text{NB}(\mu_{i,j}, \sigma_{i,j}^2)$, for gene i from sample j
 - $\mu_{i,j}$ = sample effects x sequencing effects
 - $\sigma_{i,j}^2 = \mu_{i,j} + \text{gene-specific effects} \times \mu_{i,j}^2$
- Sequencing effects
 - Sequencing depth (sample-specific)
 - GC content (gene-specific)
 - Gene length (gene-specific)

DESeq2 model of gene expression



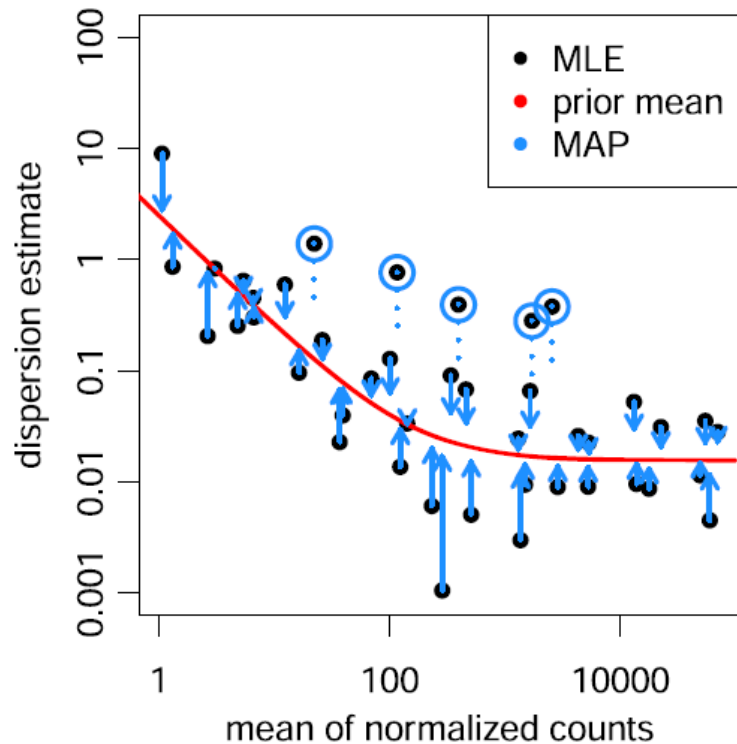
- Read count $K_{i,j} \sim \text{NB}(\mu_{i,j}, \sigma_{i,j}^2)$, for gene i from sample j
 - $\mu_{i,j}$ = sample effects x sequencing effects
 - $\sigma_{i,j}^2 = \mu_{i,j} + \text{gene-specific effects} \times \mu_{i,j}^2$
- Gene-specific effects on variance
 - **Assumption:** Genes with similar expression should have similar variances
 - Regression of **gene-specific effects** versus $\mu_{i,j}$
 - Also called **dispersion**

Two-step Bayesian approach for dispersion fitting



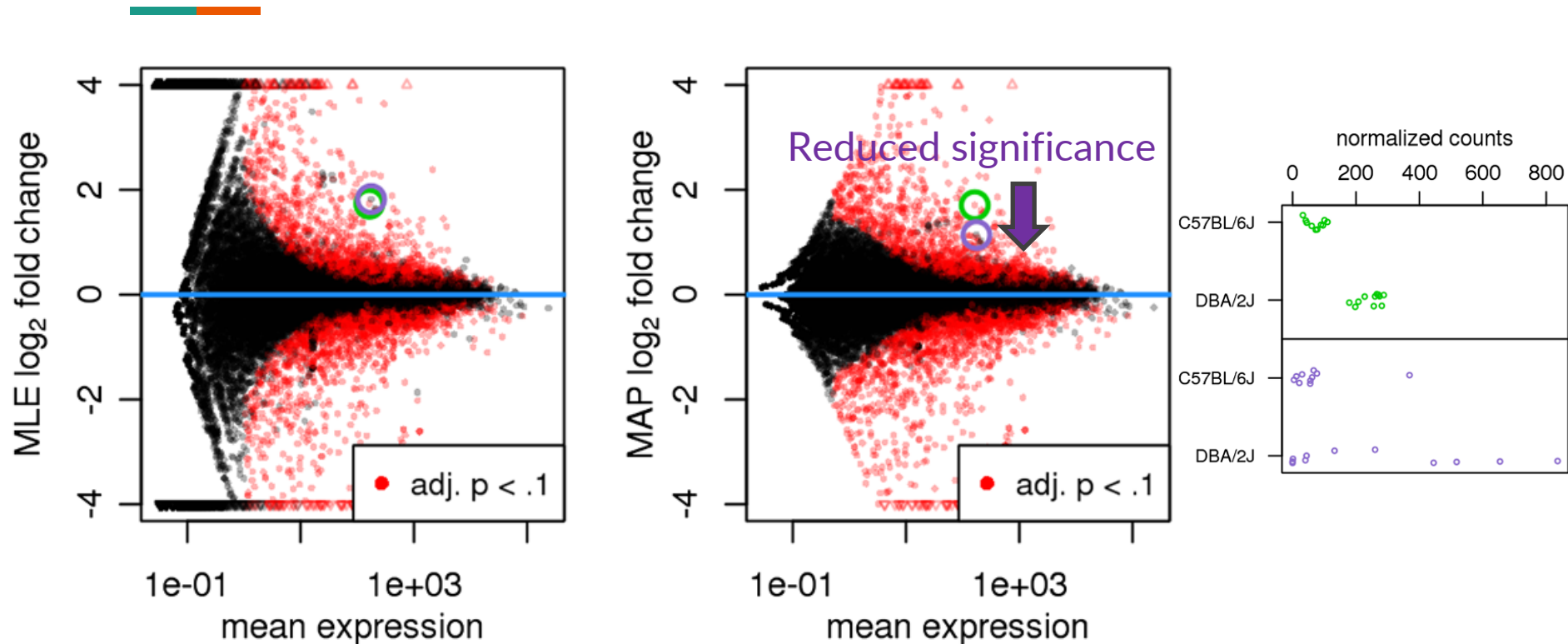
- **Dispersion** = $\frac{\sigma_{i,j}^2 - \mu_{i,j}}{\mu_{i,j}^2} = \left(\frac{\sigma_{i,j}}{\mu_{i,j}}\right)^2 - \frac{1}{\mu_{i,j}}$
- For genes with high expression level,
 $\text{Log}(\text{Dispersion}) \approx 2 \cdot \text{Log}\left(\frac{\sigma_{i,j}}{\mu_{i,j}}\right)$
- Fit trend using local regression
 - Similar to moving average

Two-step Bayesian approach for dispersion fitting



- Estimate of dispersion is noisy if there are few samples
- MLE = direct estimate
- MAP = Bayesian update using the fitted trend as prior
- Genes with very high dispersions may reflect true biological variations

Impact of two-step Bayesian update

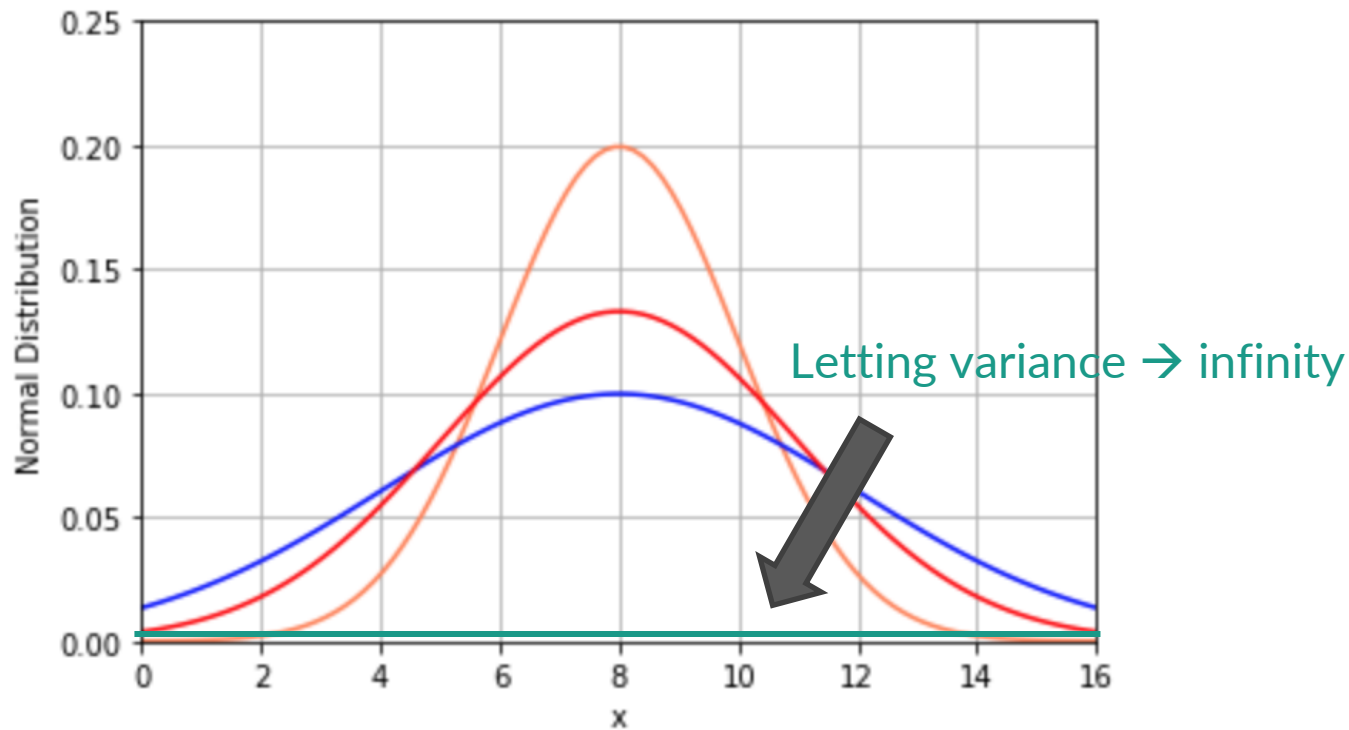


Caveats of two-step Bayesian update



- The used prior models are typically **conservative**
 - Dispersions follow the average fitted trend
 - Sample effect $\beta_{i,r}$ follows $\text{Normal}(0, \sigma_r^2)$
- Implicitly **favor the null hypothesis** of no differential expression
- What if the goal of the experiment is to show that two treatments **provide the same difference**?
 - Bayesian update in DESeq2 can be disabled
 - Uninformative prior: $\text{Normal}(0, \infty)$

Uninformative prior



DE as a test of effect size



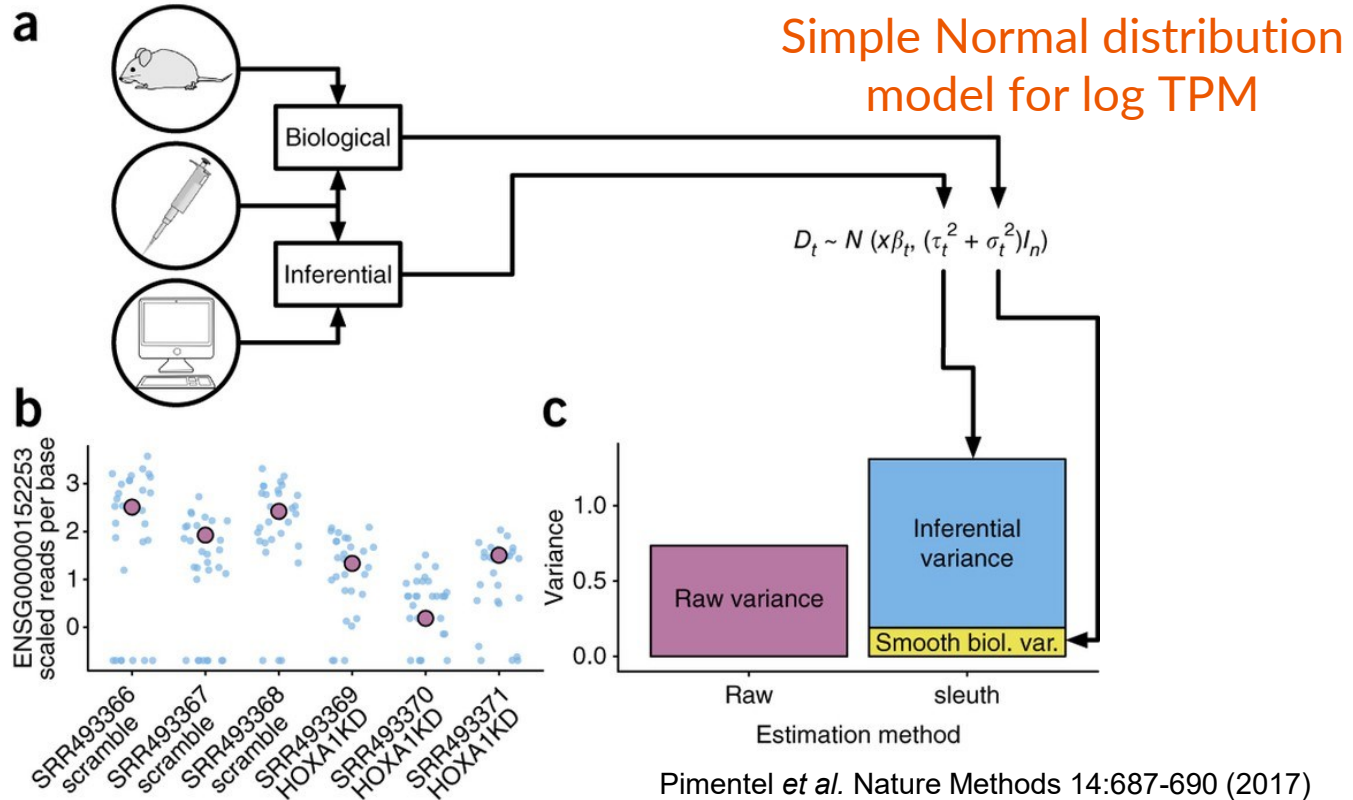
- Sample effects
 - $\text{Log FC} = \sum_r x_{j,r} \beta_{i,r}$ where $x_{j,r}$ are design parameters for sample j and $\beta_{i,r}$ are the effect sizes
- Wald test for each $\beta_{i,r}$: $\frac{\beta_{i,r}}{\text{SE}(\beta_{i,r})} \sim \text{Standard Normal}$



sleuth model for TPM

Making use of bootstrap to estimate variance

Technical
variance
estimates from
bootstrapping



Normal distribution model for log TPM



- True expression: $y_{t,i} = x_i^T \beta_t + \varepsilon_{t,i}$ for sample i and transcript t
- Observed expression: $D_{t,i} = y_{t,i} + \zeta_{t,i}$
- Noises are normally distributed: $\varepsilon_{t,i} \sim N(0, \sigma_t^2)$ and $\zeta_{t,i} \sim N(0, \tau_t^2)$
 - Transcript-specific
- Full model: $D_t \sim N(x^T \beta_t, (\sigma_t^2 + \tau_t^2) I_n)$

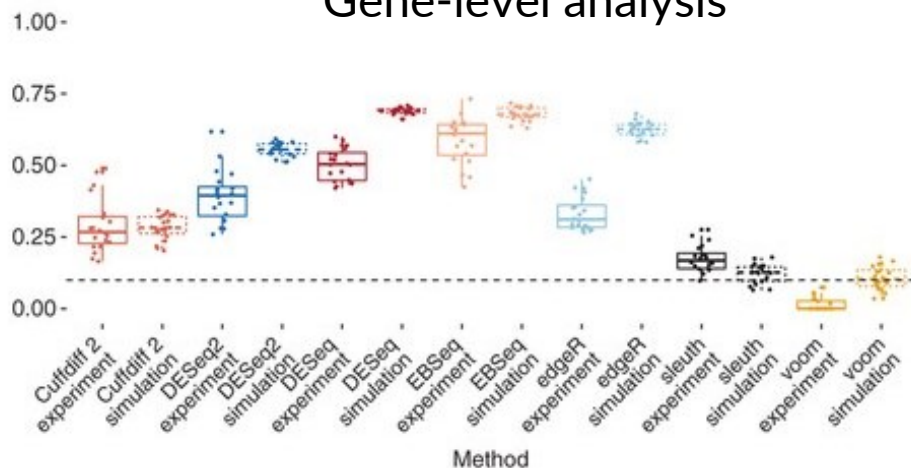
DE for sleuth



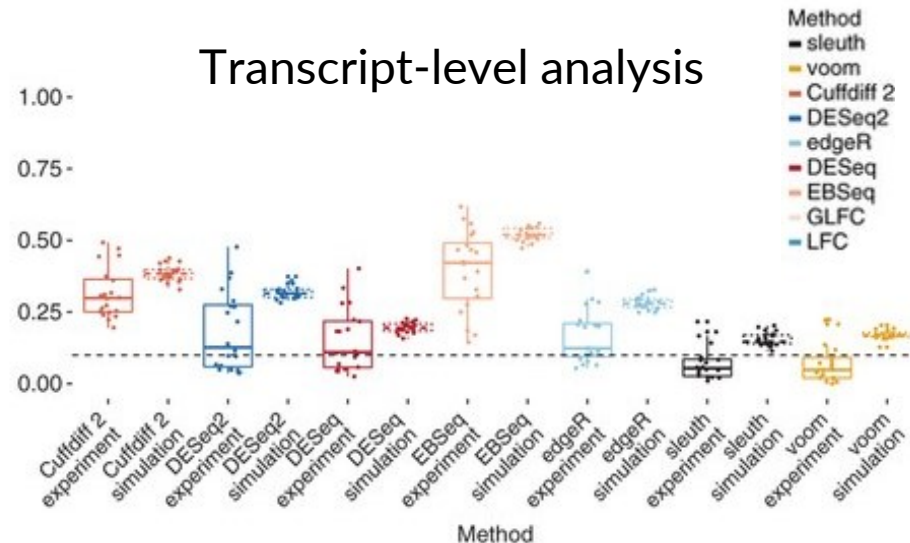
- Full model: $D_t \sim N(x^T \beta_t, (\sigma_t^2 + \tau_t^2) I_n)$
- Know from τ_t^2 bootstrapping
- Estimate from σ_t^2 data
- Fit β_t under various design matrices x (hypotheses)
- Compare likelihood ratios

Variance estimates improve accuracy of DE

Gene-level analysis



Transcript-level analysis



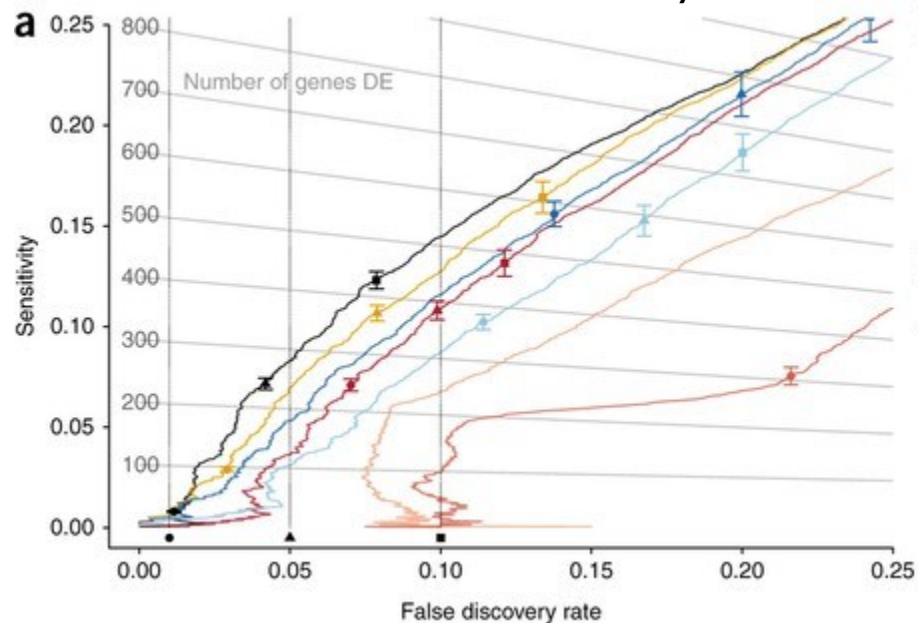
Pimentel *et al.* Nature Methods 14:687-690 (2017)

- All approaches were set to control False Discovery Rate at 10%
- Only **sleuth** and **voom** achieved the target

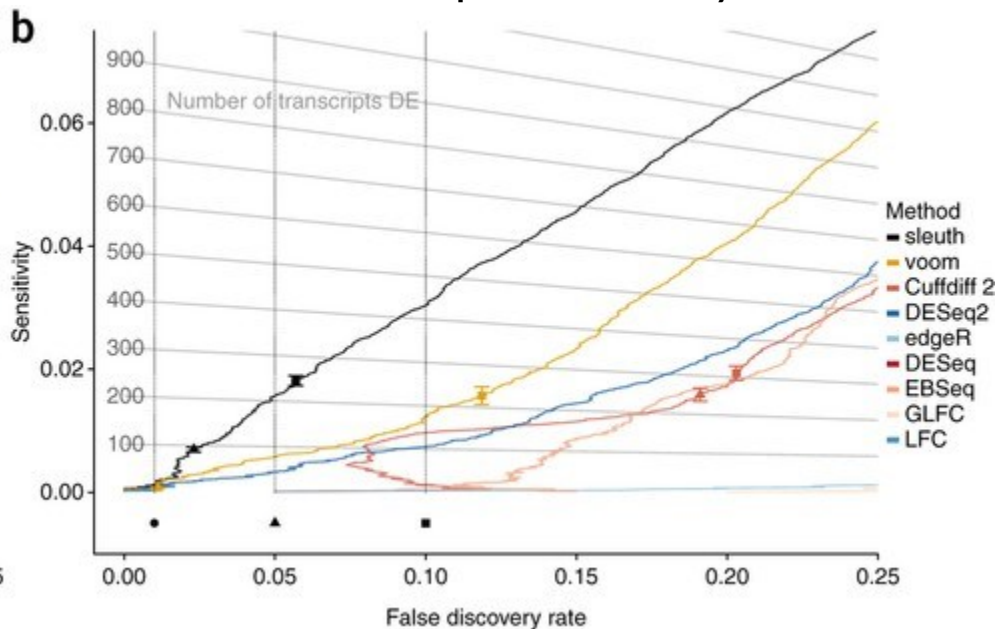
Variance estimates improve sensitivity of DE



Gene-level analysis



Transcript-level analysis



Differential expression summary



- DE can be formulated in multiple ways but depend heavily on the model of gene expression distribution
- **Read count** model using Negative Binomial distribution
 - Bayesian update to improve the estimate of variance
 - Tied to genome-based pipeline: **STAR**
- Log **TPM** model using Normal distribution
 - Estimate technical variance directly using bootstrapping
 - Tied to transcriptome-based pipeline with *k*-mer pseudoalignment
 - **kallisto** / **salmon**

Any question?



- See you on September 25