3000788 Intro to Comp Molec Biol

Lecture 11: Differential expression analysis

Fall 2025

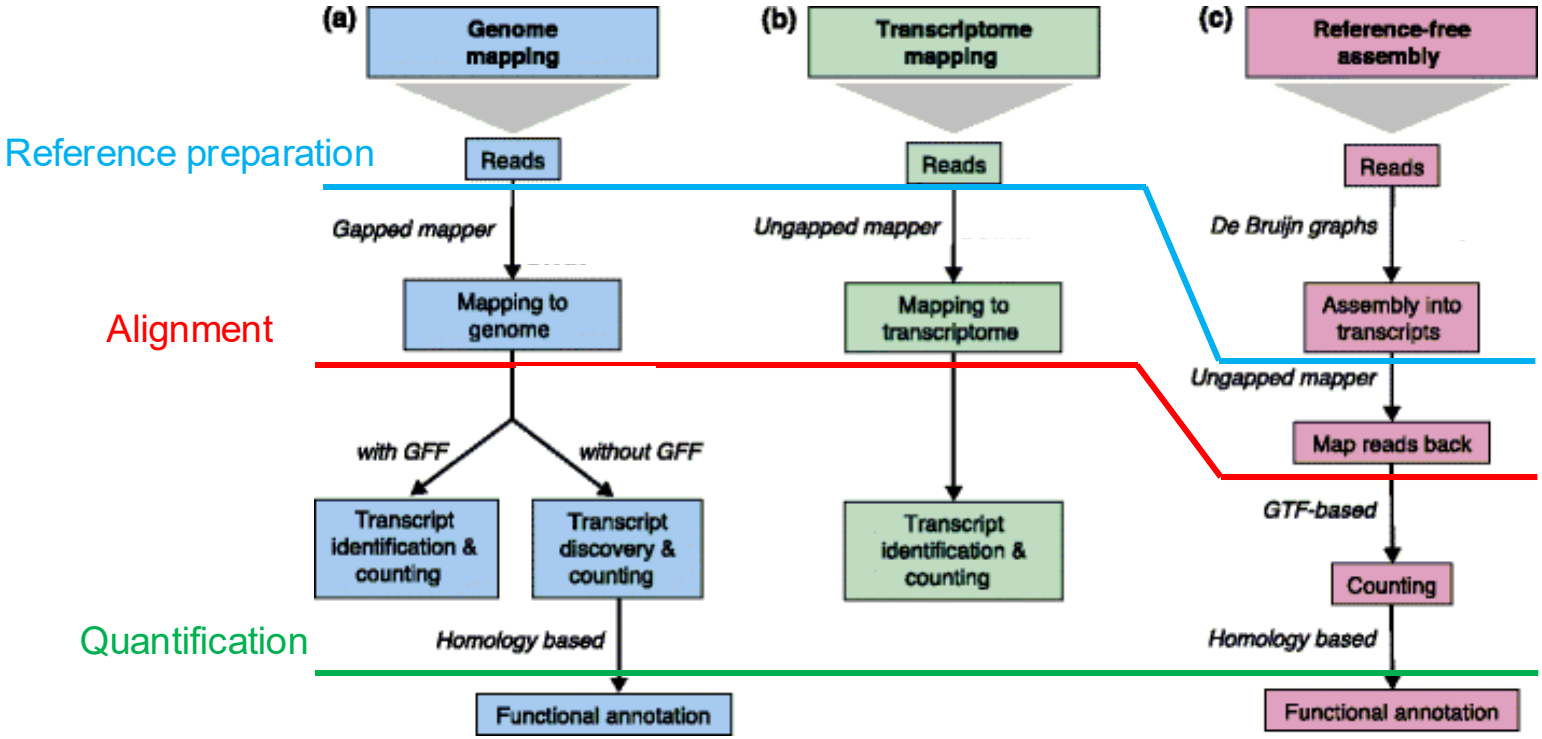
Sira Sriswasdi, PhD

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

Today's agenda

- Rapid RNA-seq alignment with *k*-mer
- Units of gene expression
- Negative binomial model for gene expression data
- Differential expression analysis

Recap: RNA-seq analysis pipelines

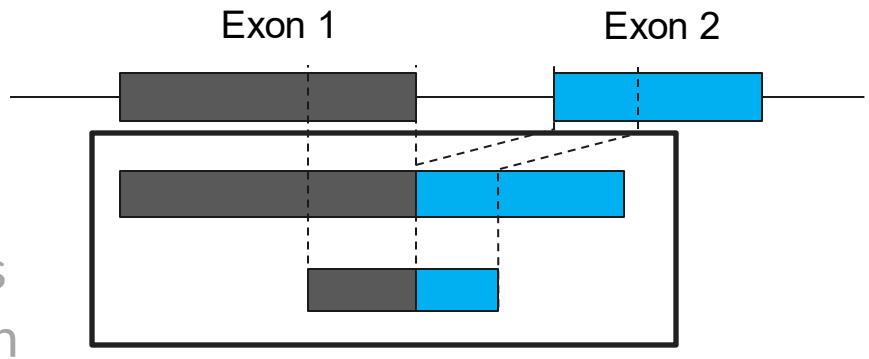


Conesa *et al.* Genome Biology 17:13 (2016)

Alignment to reference genome or transcriptome

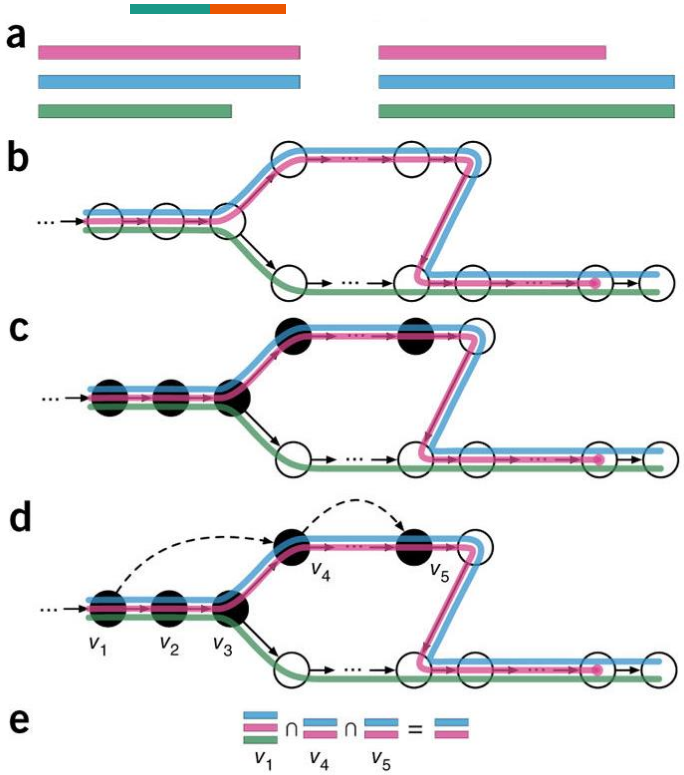
- Reference transcriptome
  - **Fast**, cannot discover new isoform
  - **Ungapped**, *k*-mer-based alignment
  - **salmon / kallisto**

- Reference genome
  - **Slow**, but can detect new isoforms
  - **Gapped alignment**, allow for intron
  - Can be guided by exon annotations
  - **STAR, HISAT2**



k-mer pseudoalignment

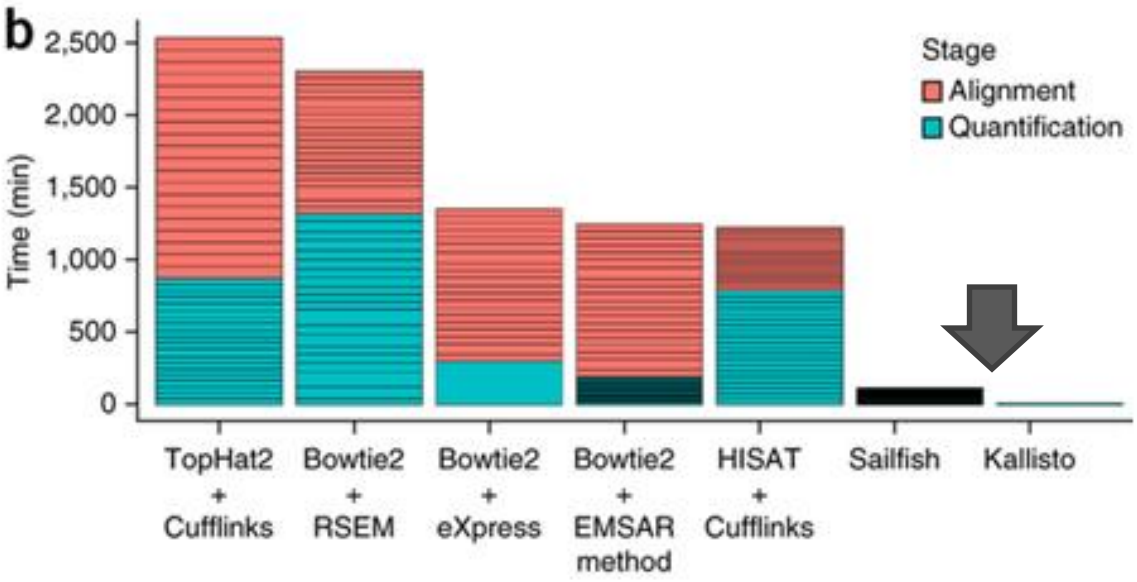
k-mer pseudoalignment algorithm sketch



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

- Create de Bruijn graph from *k*-mer of reference transcripts
  - Path = ordering of *k*-mer on a transcript
  - Drop uninformative *k*-mers (v_2 and v_3) to reduce search space
- **Step 1**: For each read, identify paths that are compatible with *k*-mers from the reads
- **Step 2**: Find the best path with matching ordering of *k*-mers to those on the reads

>100 fold speed up with pseudoalignment



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

- Similar gene expression quantification accuracy
- Pseudoalignment tools use the gained time to perform **bootstrapping**!
- Bootstrapping provides estimate of technical variance

Bootstrapping of RNA-seq data

- Multiple gene expression estimates across bootstrap replicates
- Estimate technical variance for each gene and sample

Units of gene expression

Units for transcript abundance

$$\text{CPM} = \frac{\text{Read Count}}{\sum \text{Read Count}} \times 1{,}000{,}000$$

Similar to percentage (but per million)

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

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

Long transcript generates more fragments and more read counts

Experiment with higher sequencing depth generates more read counts

- Read count (number of mapped reads)
- FPKM = Fragment per kilobase of exon per million reads mapped
- TPM = Transcript per million
- CPM = Count per million

Quick notes about gene expression units

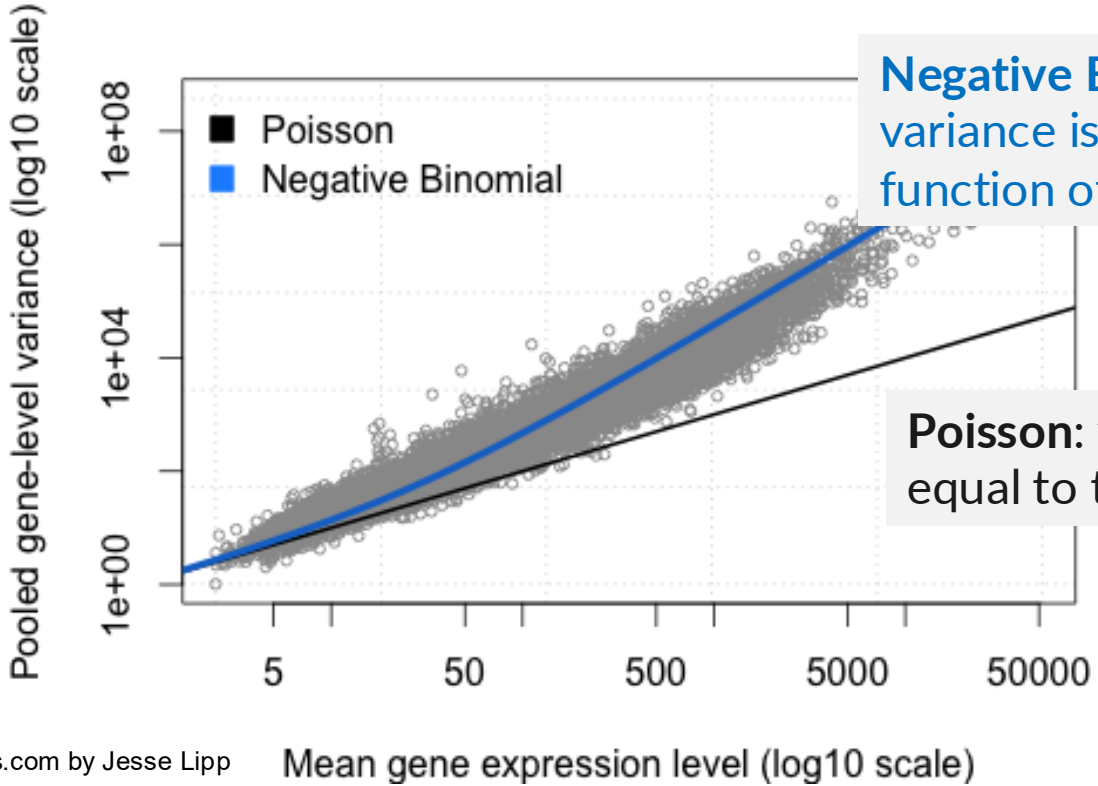
- **Read count** is not normalized, but can be modeled with statistics
  - This is needed for all **differential expression analysis** tools

- **CPM** and **FPKM** are rarely used nowadays

- **TPM** is the most normalized, and is used to **visualize** and **cluster** data
  - Show composition of transcripts
  - Similar concept as microbiome composition

Negative binomial model for read count

The distribution of RNA-seq read count



Negative Binomial: variance is a parabolic function of the mean

Poisson: variance is equal to the mean

Image from bioramble.wordpress.com by Jesse Lipp

Negative binomial model

- Repeat a series of Bernoulli trials, each with probability of success p, until we obtain r successes. How many failures, k, would we observe?
  - X O O X X X O O X O O = 5 failures observed until 6 successes were obtained

- $P_{NB}(k; r, p) = \binom{k+r-1}{k}(1-p)^k p^r$
  - $k + r - 1$ locations to place k failures (the last location must be success)

- Mean = $\frac{pr}{(1-p)}$ failures (proportional to the number of successes)
- Variance = $\frac{pr}{(1-p)^2} = \frac{pr}{(1-p)} + \left(\frac{pr}{1-p}\right)^2 \frac{1}{r}$ (a parabolic function of the mean)

Another interpretation of negative binomial model

- $\mathrm{P}_{\mathrm{NB}}(k; r, p) = \int_0^\infty \mathrm{P}_{\mathrm{Poisson}(\lambda)}(k) \cdot \mathrm{P}_{\mathrm{Gamma}\left(r, \frac{1-p}{p}\right)}(\lambda)\, d\lambda$
- **Negative Binomial** is a continuous mixture of **Poisson**, with **Gamma**-distributed weights

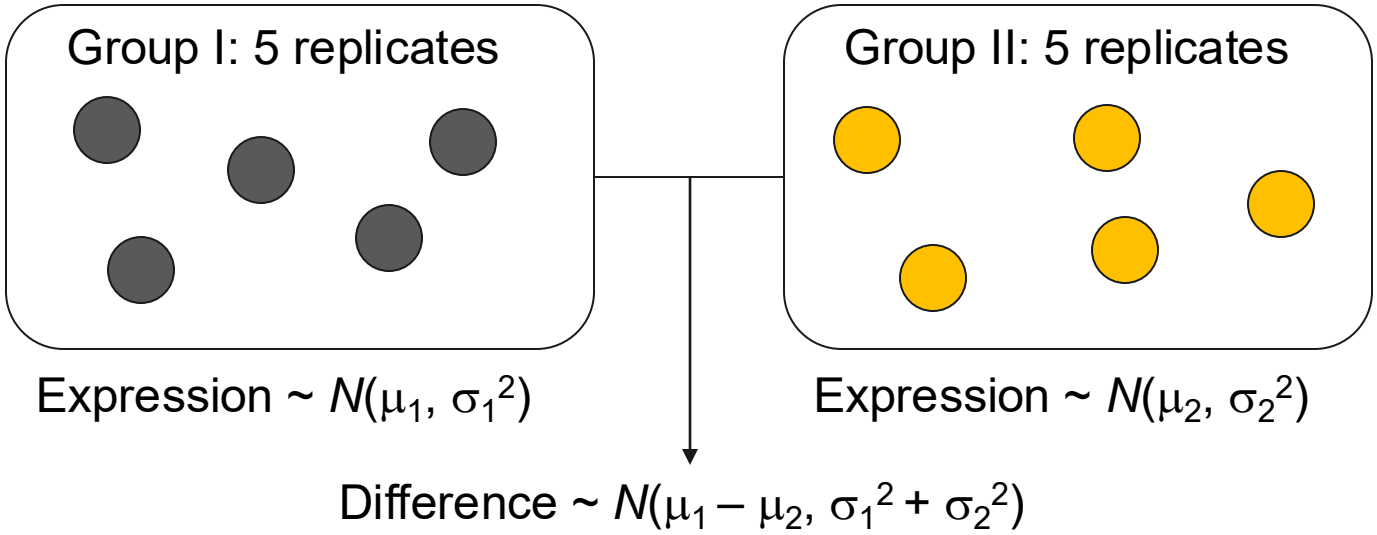
- Bulk gene expression is an average over many cells

- Imagine read counts from each cell following **Poisson** distribution, and that each cell type has a population following **Gamma** distribution
  - **Negative Binomial** is a weighted sum of single-cell gene expression!

Simple differential expression analysis

For log-normally distributed data

Differential expression for normally distributed data



- Simple *t*-test works

Preparing Microarray / Nanostring data for *t*-test

Control Treatment

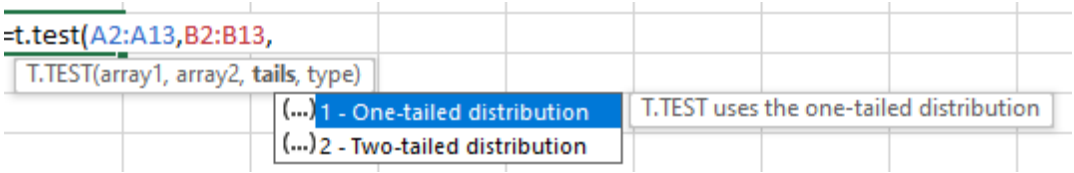
J15

	A	B	C	D	E	F	G	H
1	Acc ID	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	
2	NM_007818	67540.89	70924.09	80243.76	3501.2	5697.47	2426.72	
3	NM_001105160	811.93	801.36	740.71	128.67	104.42	101.33	
4	NM_028089	190.41	211.06	236.19	9.05	23.33	8.44	
5	NM_016696	66.77	57.56	101.09	750.9	659.84	491.89	
6	NM_013459	3.3	11.29	1.89	735.82	816.46	118.22	
7	NM_007809	45.34	36.12	51.02	245.27	372.13	335.67	
8	NM_009999	103.04	370.21	200.29	17.09	13.33	8.44	
9	NM_133960	7708.78	6976.38	6569.04	1731	1641.81	1853.55	
10	NM_027881	31.32	10.16	24.56	268.39	186.62	135.11	
11	NM_054053	31.32	24.83	19.84	323.68	428.78	116.11	
12	NM_007377	47.81	89.17	70.86	370.93	378.79	279.72	
13	NM_028064	703.95	689.62	662.29	214.11	168.85	144.61	
14	NM_008182	222.56	339.73	226.75	30.16	63.32	26.39	
15	NM_013661	12.36	11.29	8.5	97.51	77.76	71.78	
16	NM_007815	20613.09	25218.13	31540.46	5209.07	7680.3	6312.2	

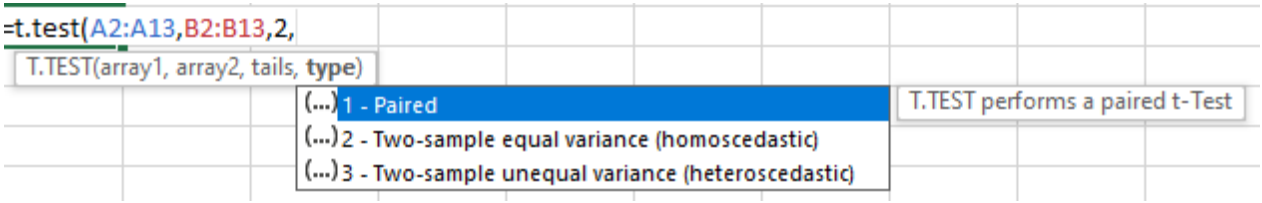
- Remember to check if the data have been log-transformed
- *t*-test can be performed on each gene
- Correct the p-values for multiple testing

http://homer.ucsd.edu/homer/basicTutorial/clustering.html

Flavors of *t*-test



- One or two-tailed depends on your hypothesis, two-tailed in general



- Paired for **before & after treatment data of the same samples**
- Equal or unequal variance across groups depends on assumption

p-value correction methods

Correction with Bonferroni method

- Divide the p-value cutoff by the number of test
- **Example**: Adjusted p-value cutoff = 0.05 / 1000 = 0.00005

- Easy to perform but lose power (fail to reject Null Hypothesis when the Alternative is true)

- P-value is calculated based on the assumption that Null Hypothesis is true
  - Doesn't tell us directly about **False Positives**
  - Out of 1,000 genes that passed the cutoff, how many are false?

False discovery rate (FDR)

- **False Discovery Rate (FDR)** = Probability of getting a false positive
  - Probability that a gene that passed the cutoff is not truly differentially expressed

- But FDR involves alternative hypothesis is difficult to calculate

- There are ways to **control FDR through p-value**!

Benjamini-Hochberg procedure

- Valid under broad assumptions (independence, positively correlated, etc.)
- Statistical tests with raw p-values, $p_1, p_2, \ldots, p_n$

- To control **FDR ≤ 0.05** (not p-value ≤ 0.05)
  - **Step 1**: Sort p-values from low to high: $p'_1, p'_2, \ldots, p'_n$
  - **Step 2**: Go through each p-value and check whether $\boldsymbol{p'_i \leq 0.05 \times i / n}$
  - **Step 3**: Stop when the condition fail (found $\boldsymbol{p'_k > 0.05 \times k / n}$)
  - **Step 4**: Reject null hypothesis for tests corresponding to $p'_1, p'_2, \ldots, p'_{k-1}$

- **Observations**:
  - For the smallest p-value, whether $\boldsymbol{p'_1 \leq 0.05 \times 1 / n}$ is equivalent to **Bonferroni**
  - For other p-values, whether $\boldsymbol{p'_i \leq 0.05 \times i / n}$ is more relaxed

Benjamini-Yekutieli procedure

- Valid under broader assumption (some dependency between tests is allowed)

- Same procedure as **Benjamini-Hochberg**
- But the cutoff for each p-value is more stringent (smaller cutoff)

- For the smallest p-value, the test is again equivalent to **Bonferroni**

- If your smallest p-value fails Bonferroni correction, don't have to try another test

Comparing correction methods

P-value	Bonferroni	B-H	B-Y
Smallest	0.0005	0.0005	0.0005
2nd smallest	0.0005	0.001	0.000667
3rd smallest	0.0005	0.0015	0.000818
4th smallest	0.0005	0.002	0.00096
5th smallest	0.0005	0.0025	0.001095

- There are n = 100 tests
- Target FDR = 0.05
- **B-H** and **B-Y** gradually increase p-value cutoff while **Bonferroni** does not

Power of different correction methods

Gene	p-value (sorted)	Bonferroni Result	Benjamini-Hochberg Cutoff	B-H Result	Benjamini-Yekutieli Cutoff	B-Y Result
Gene M	0.00001	Pass	0.0005	Pass	0.00050	Pass
Gene S	0.00035	Pass	0.0010	Pass	0.00067	Pass
Gene A	0.00062	Fail	0.0015	Pass	0.00082	Pass
Gene C	0.00110	Fail	0.0020	Pass	0.00096	Fail
Gene P	0.06014	Fail	0.0025	Fail	0.00110	Fail

DESeq2 differential expression model

Negative Binomial model for gene expression

- Read count $K_{i,j} \sim NB(\mu_{i,j}, \sigma_{i,j}^2)$, for gene *i* in sample *j*
  - $\mu_{i,j}$ = sample effects x sequencing effects
  - $\sigma_{i,j}^2 = \mu_{i,j}$ + gene-specific effects x $\mu_{i,j}^2$ (parabolic function of the mean)

- Sequencing effects
  - Sequencing depth of the sample
  - GC content of the reads
  - Length of the gene

Negative Binomial model for gene expression

- Read count $K_{i,j}$ ~ NB($\mu_{i,j}$, $\sigma_{i,j}^2$), for gene *i* in sample *j*
  - $\mu_{i,j}$ = sample effects x sequencing effects
  - $\sigma_{i,j}^2$ = $\mu_{i,j}$ + gene-specific effects x $\mu_{i,j}^2$ (parabolic function of the mean)

- Sample effects
  - **Linear effect model**
    - **Log Fold-Change** = $\sum_r x_{j,r}\beta_{i,r}$ where $x_{j,r}$ are design parameters for sample *j* and $\beta_{i,r}$ are the effect sizes for gene *i*
  - **Design parameters**
    - **Experiment conditions**: control, treatment, etc.
    - **Confounding factors**: age, time after treatment, dosage, etc.

Modeling effects through design parameters

Model

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

E(y) =	1.03				=	1.03	*(for control)*
E(y) =	1.03 +	1.09			=	2.12	*(for treatment I)*
E(y) =	1.03 +		1.97		=	3.00	*(for treatment II)*
E(y) =	1.03 +	1.09 +	1.97 +	0.82	=	4.90	*(for treatments I & II)*

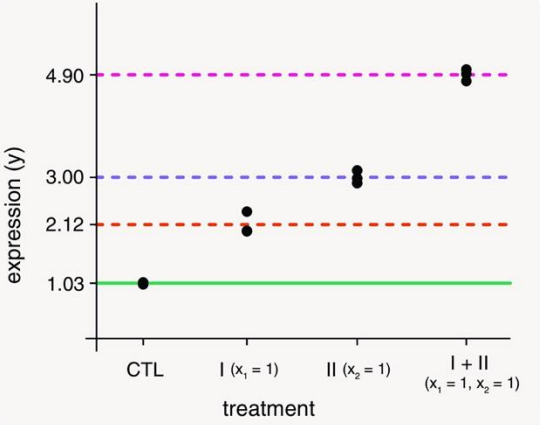
Matrix

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

	(Intercept)	treat1YES	treat2YES	treat1YES:treat2YES
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	1	1	0	0
5	1	1	0	0
6	1	1	0	0
7	1	0	1	0
8	1	0	1	0
9	1	0	1	0
10	1	1	1	1
11	1	1	1	1
12	1	1	1	1

Plot

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



Differential expression as a test of effect size

- **Linear effect model**
  - **Log Fold-Change** = $\sum_r x_{j,r}\beta_{i,r}$ where $x_{j,r}$ are design parameters for sample *j* and $\beta_{i,r}$ are the effect sizes for gene *i*

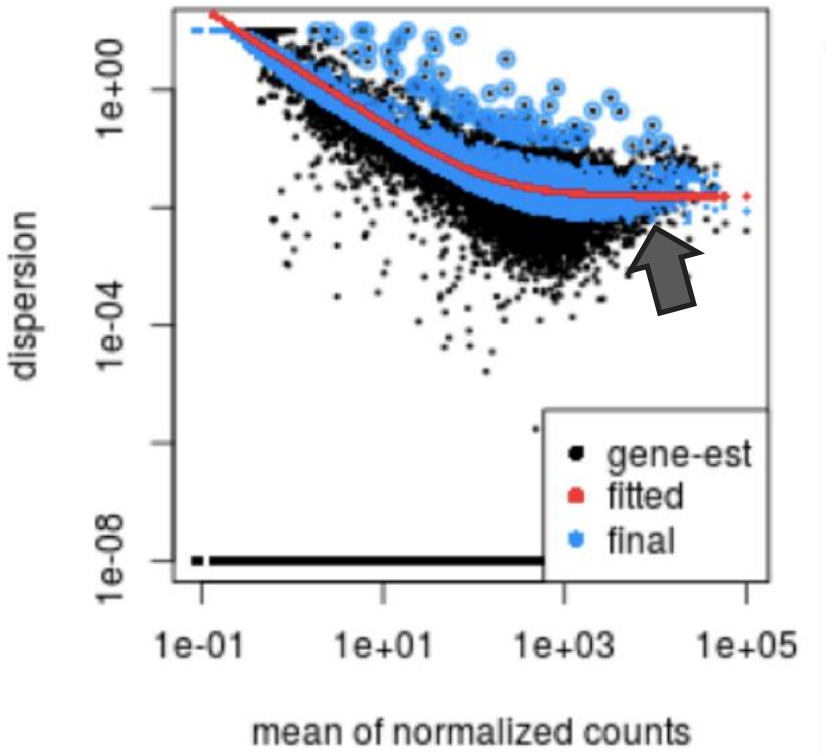
- **Wald test** whether each $\beta_{i,r}$ is significantly different from zero
  - **Assumption**: $\frac{\beta_{i,r}}{\mathrm{SE}(\beta_{i,r})} \sim$ Standard Normal
  - Reject Null Hypothesis = design parameter *r* affected the gene expression

Negative Binomial model for gene expression

- Read count $K_{i,j}$ ~ NB($\mu_{i,j}$, $\sigma_{i,j}^2$), for gene *i* in sample *j*
  - $\mu_{i,j}$ = sample effects x sequencing effects
  - $\sigma_{i,j}^2$ = $\mu_{i,j}$ + gene-specific effects x $\mu_{i,j}^2$ (parabolic function of the mean)

- Gene-specific effects on variance
  - **Assumption**: Genes with similar expression should have similar variances
  - Provide more robust variance estimates across genes
  - Regression of gene-specific effects versus $\mu_{i,j}$

Regression of variance as a function of mean expression

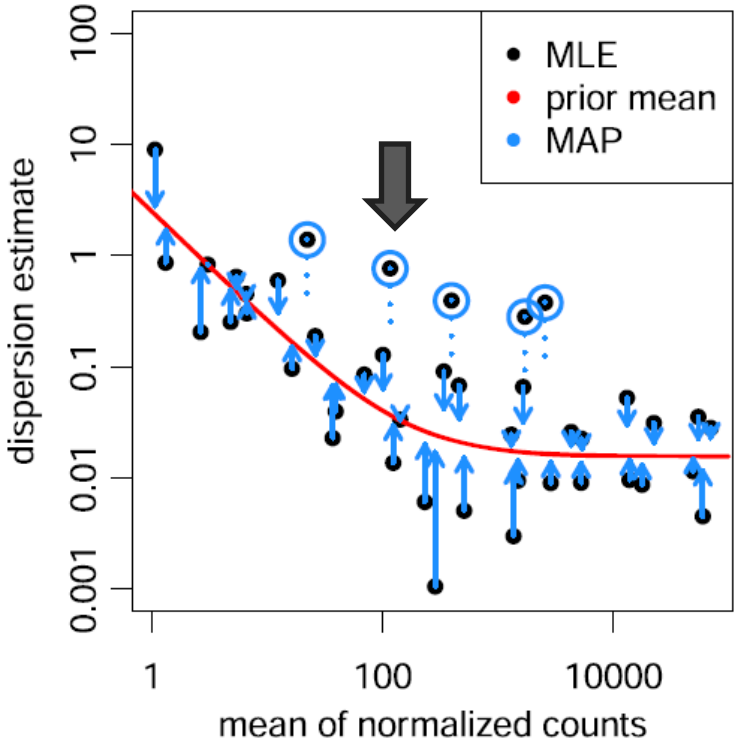


- Dispersion $= \frac{\sigma_{i,j}^2 - \mu_{i,j}}{\mu_{i,j}^2} = \left(\frac{\sigma_{i,j}}{\mu_{i,j}}\right)^2 - \frac{1}{\mu_{i,j}}$
- For highly expressed genes,

  $\mathrm{Log}(\mathrm{Dispersion}) \approx 2 \cdot \mathrm{Log}\left(\frac{\sigma_{i,j}}{\mu_{i,j}}\right)$
- Fit trend using local regression
  - Similar to moving average

Love, Huber, and Anders. Genome Biol. 15:550 (2014)

Two-step Bayesian approach for dispersion fitting



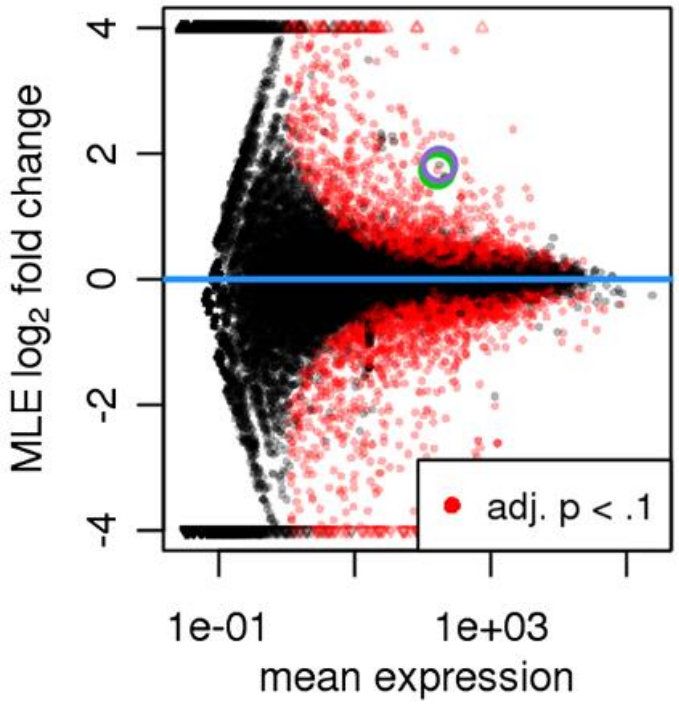
- Observed variance can be noisy if there are few replicate samples

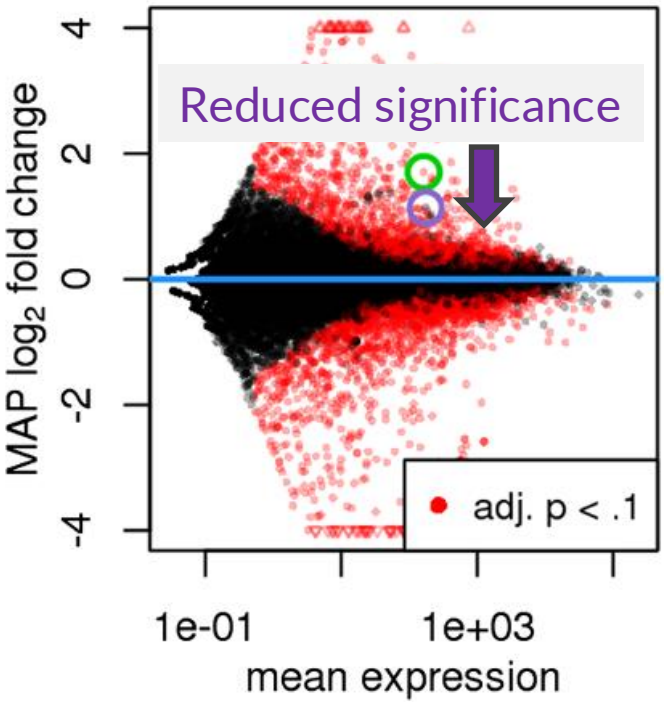
- MLE = variance of each gene
- Prior mean = Fitted regression trend
- MAP = **Bayesian** update

- Genes with very high dispersions may reflect true biological variations, otherwise, trust the regression trend

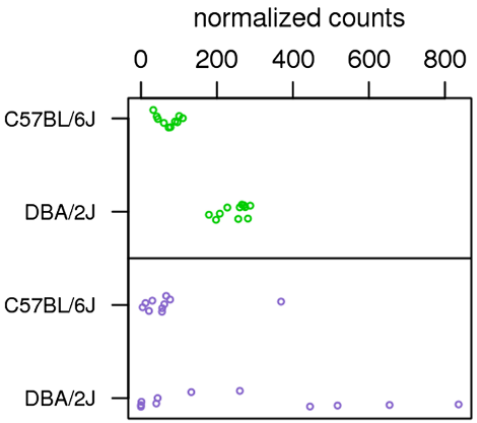
Love, Huber, and Anders. Genome Biol. 15:550 (2014)

Dispersion fitting highlights true differences





Difference come from real shift



Difference come from dispersion

Love, Huber, and Anders. Genome Biol. 15:550 (2014)

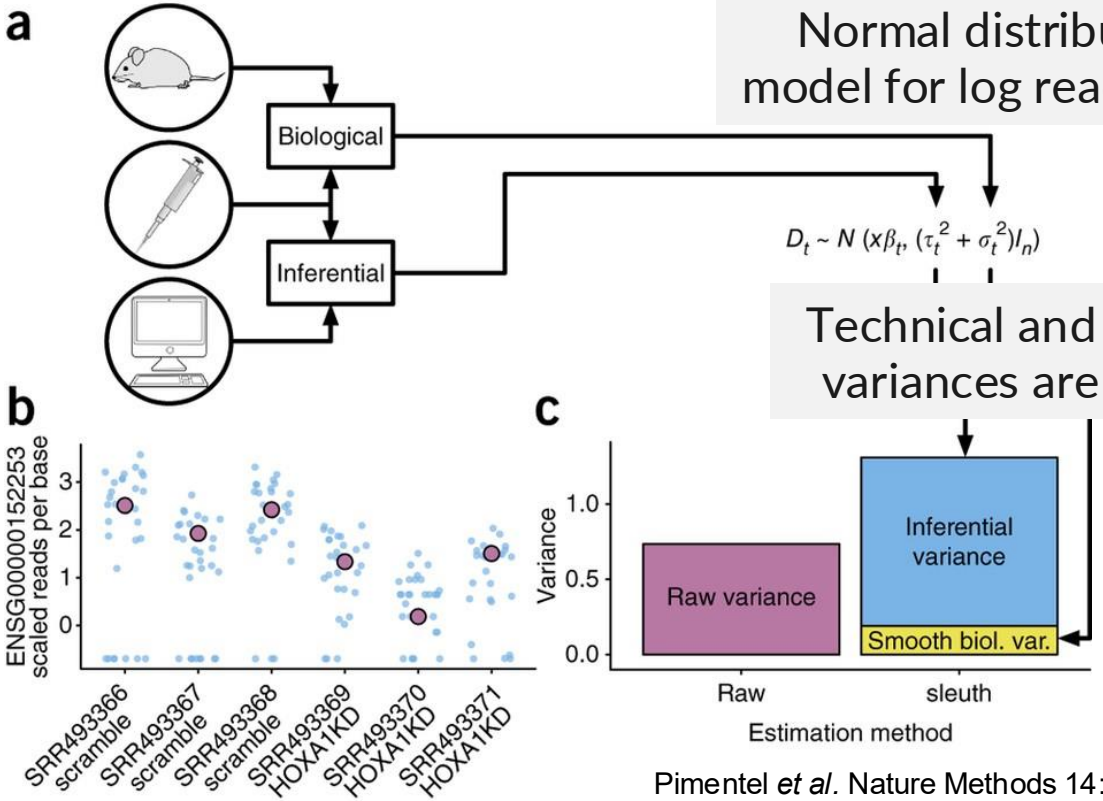
sleuth differential expression model

Segregation of biological and technical variances

Normal distribution model for log read count

Technical and biological variances are separate

Technical variance estimates from bootstrapping



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

Normal distribution model for log read count

- **True expression:** $y_{t,i} = x_i^T \beta_t + \varepsilon_{t,i}$ for sample *i* and transcript *t*
- **Observed expression:** $D_{t,i} = y_{t,i} + \zeta_{t,i}$

- **Transcript-specific noises:** $\varepsilon_{t,i} \sim N(0, \sigma_t^2)$ and $\zeta_{t,i} \sim N(0, \tau_t^2)$

- **Full model:** $D_t \sim N(x^T \beta_t, (\sigma_t^2 + \tau_t^2) I_n)$

Likelihood ratio test for differential expression

- **Full model:** $D_t \sim N(x^T \beta_t, (\sigma_t^2 + \tau_t^2) I_n)$
  - Know τ_t^2 from bootstrapping (unique to pseudoalignment)
  - Estimate σ_t^2 from data (same as DESeq2)

- Fit β_t under various design matrices x (hypotheses)
  - Compare likelihoods across hypotheses (with different number of parameters)

Likelihood ratio test for differential expression

Hypothesis 1

No treatment effect

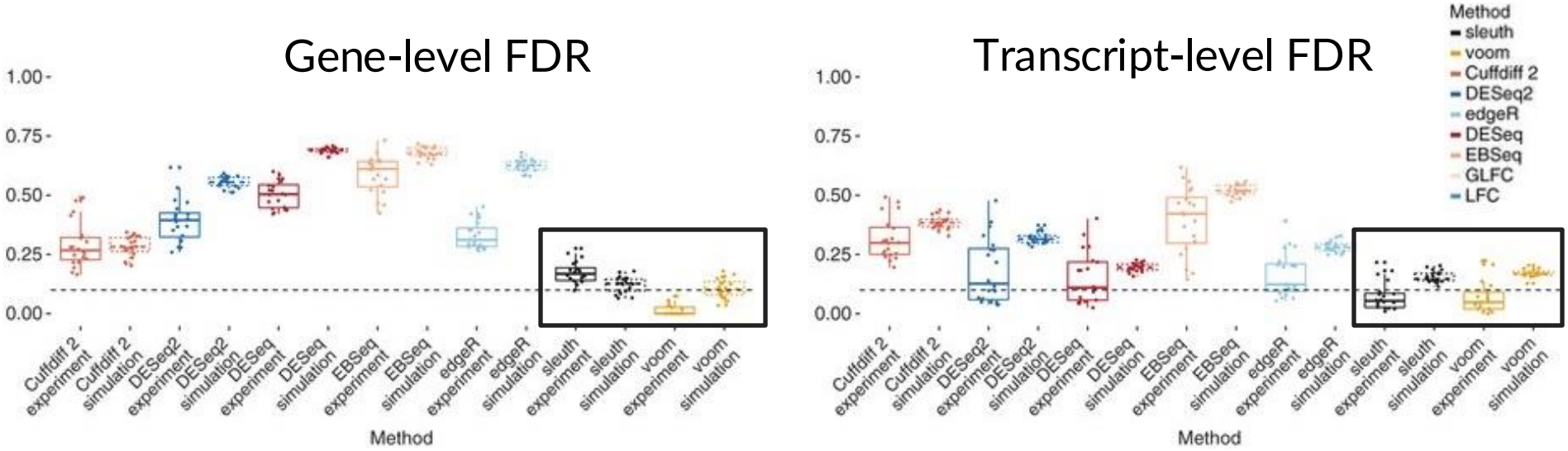
Hypothesis 2

Treatment effect

Hypothesis 3

Treatment and batch effects

Technical variance estimates improve accuracy



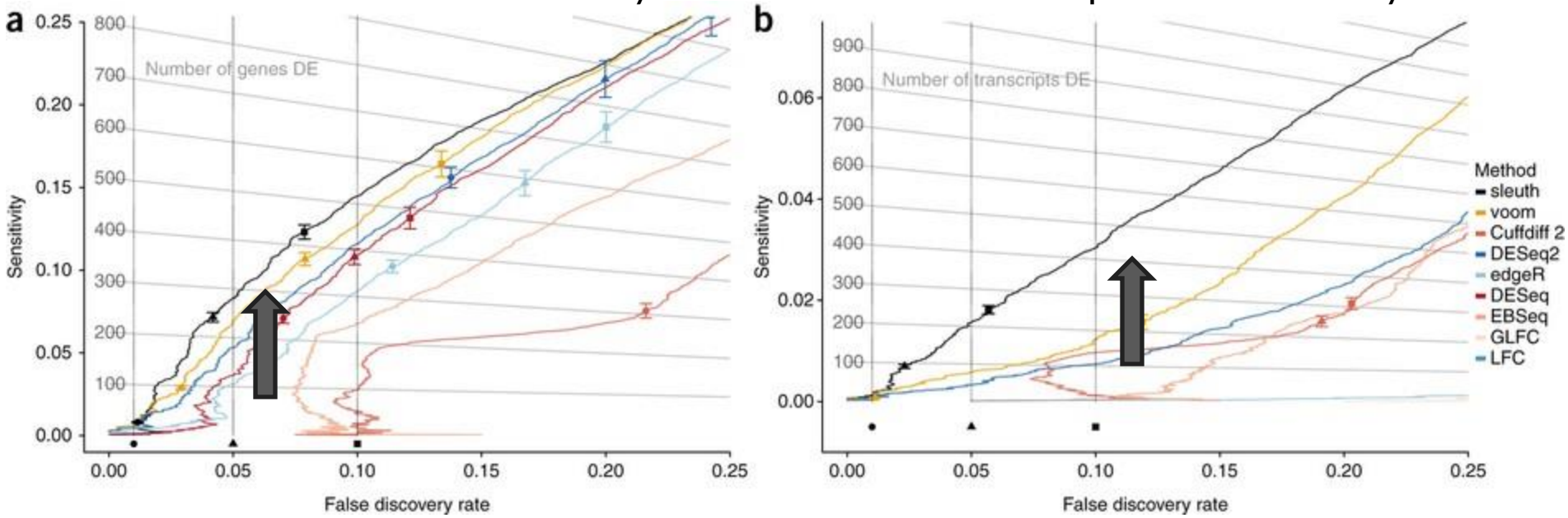
Pimentel *et al.* Nature Methods 14:687-690 (2017)

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

Technical variance estimates improve sensitivity

Gene-level sensitivity

Transcript-level sensitivity



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

Differential expression summary

- Microarray / Nanostring → ***t*-test** followed by p-value correction
- Nanostring data can also be analyzed with **DESeq2** (count data)
- RNA-seq aligned to genome (no bootstrapping) → **DESeq2**
- RNA-seq aligned to transcriptome (bootstrapping) → **DESeq2** or **sleuth**
  - **sleuth** exhibits superior sensitivity at transcript level

Any question?

- See you next time