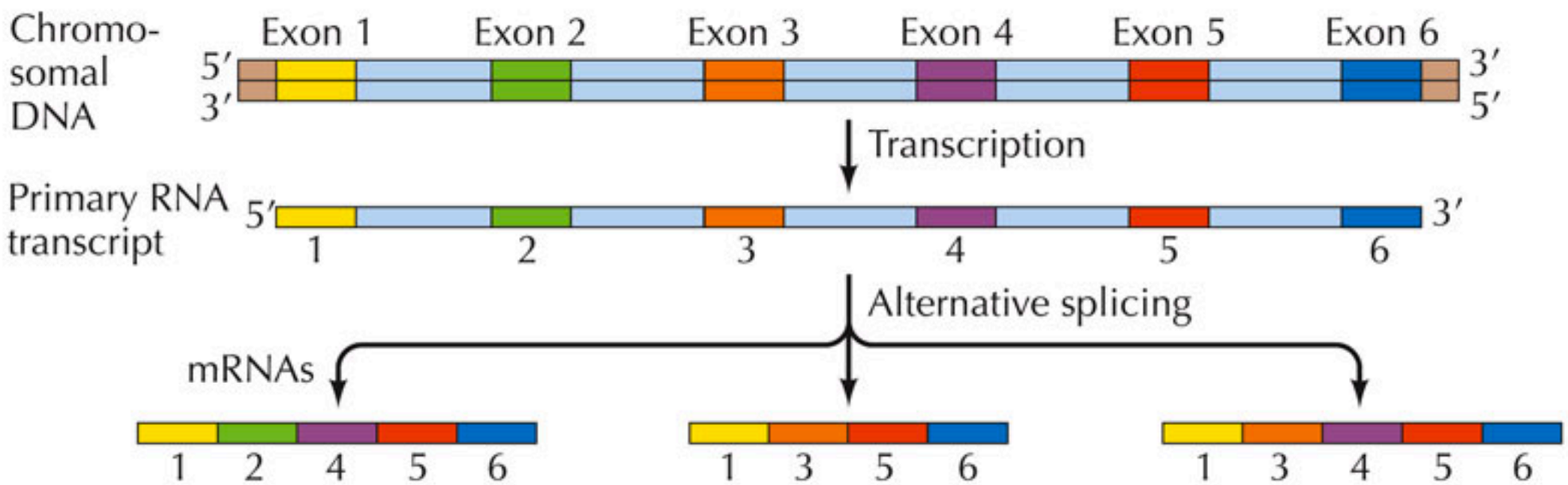
Differential expression

Alejandro Reyes
T: @areyesq89

Workshop on Transcriptomics
September 13th, 2017

Overview of exons, genes and transcripts



What is your biological question?

Given a gene, test for:

- Whether transcripts levels change between conditions? (differential exon usage, DGE)
- Whether transcript isoform proportions change between conditions? (differential transcript usage, DTU)
- Whether individuals exons are differentially used? (differential exon usage, DEU)

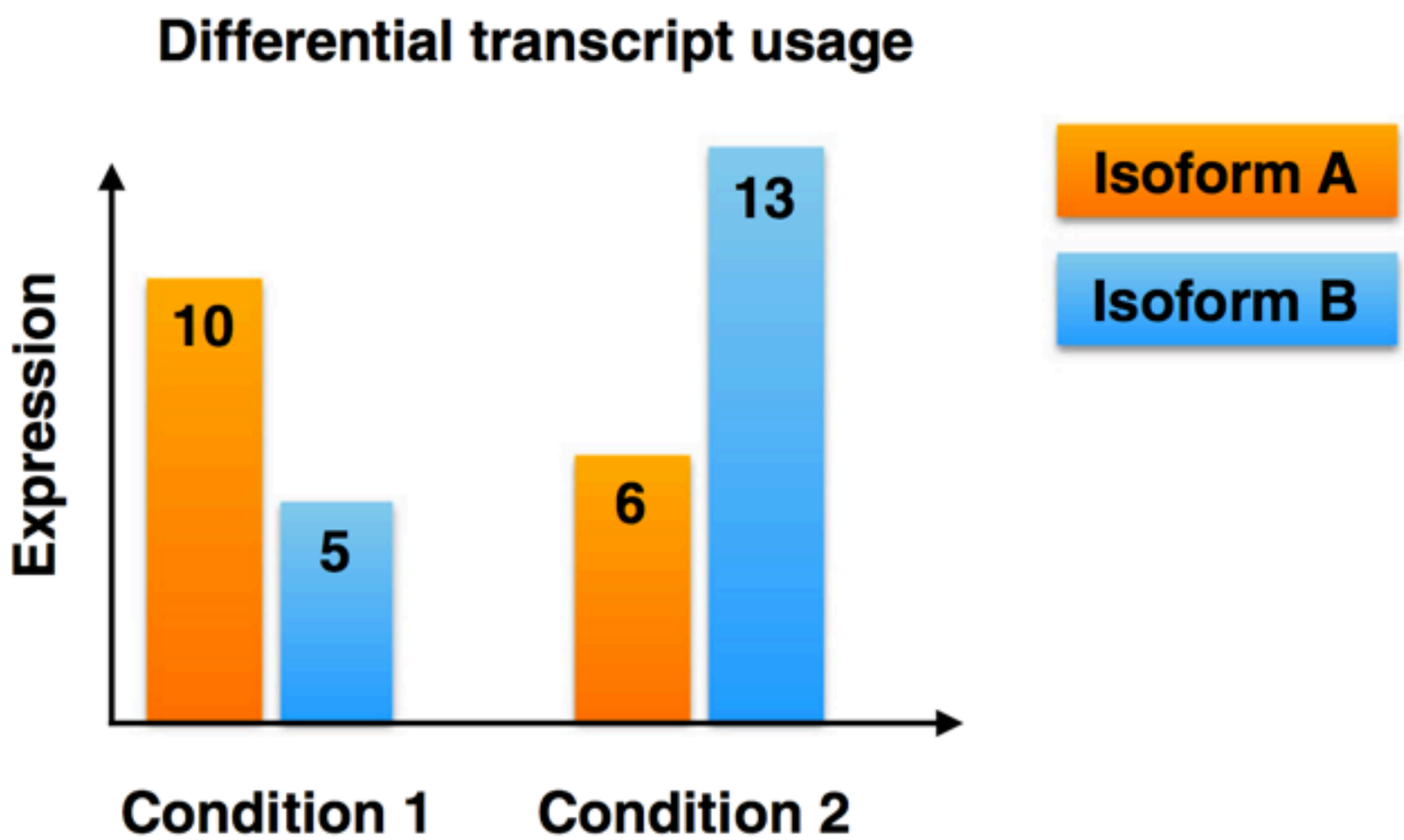
Differential problems: DGE

Differential transcript expression (DTE)

Differential gene expression (DGE)

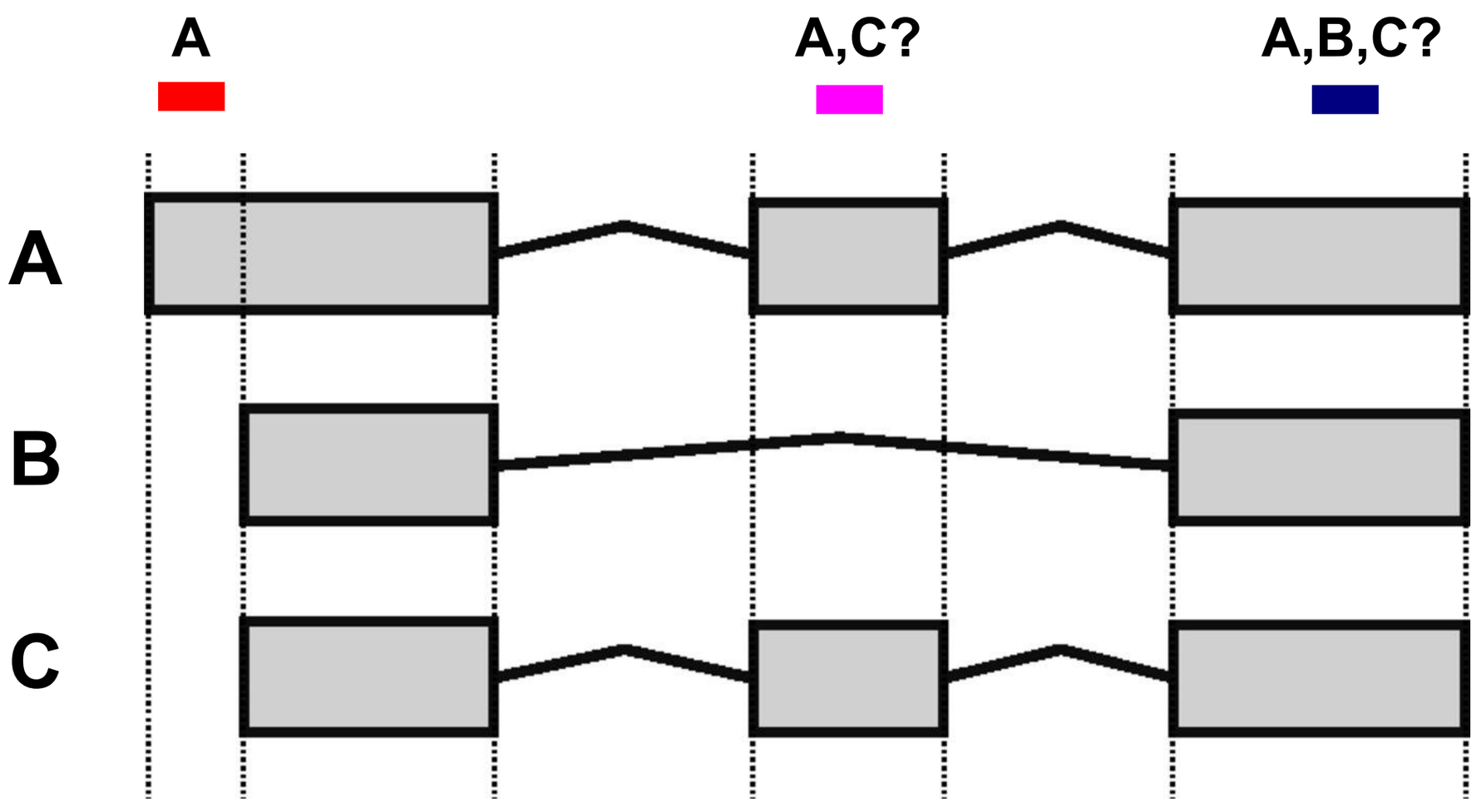
Modified slide from Mark Robinson

Differential problems: DTU

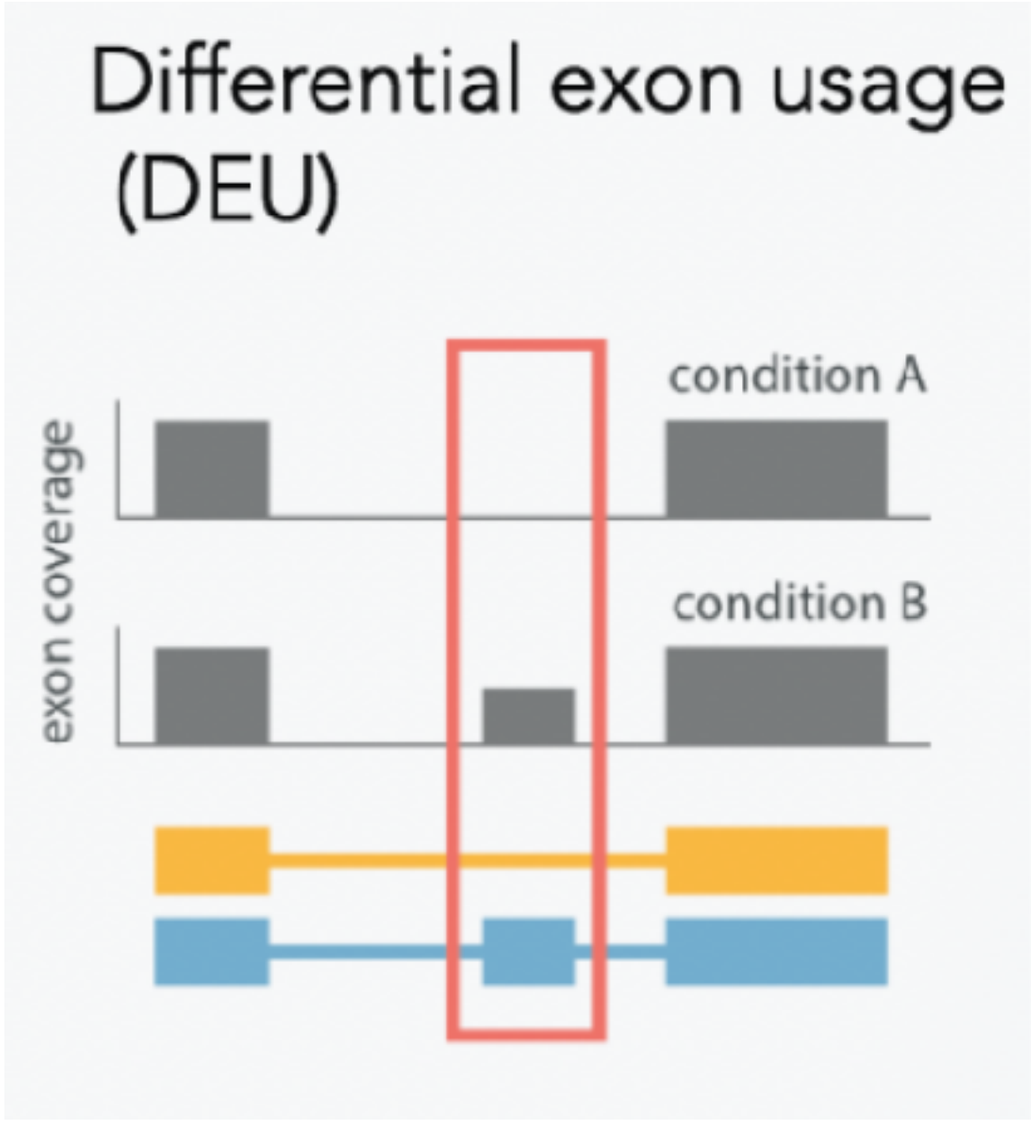


Soneson et al, 2014

Unambiguous assignment of reads to single transcripts



Differential problems: DEU



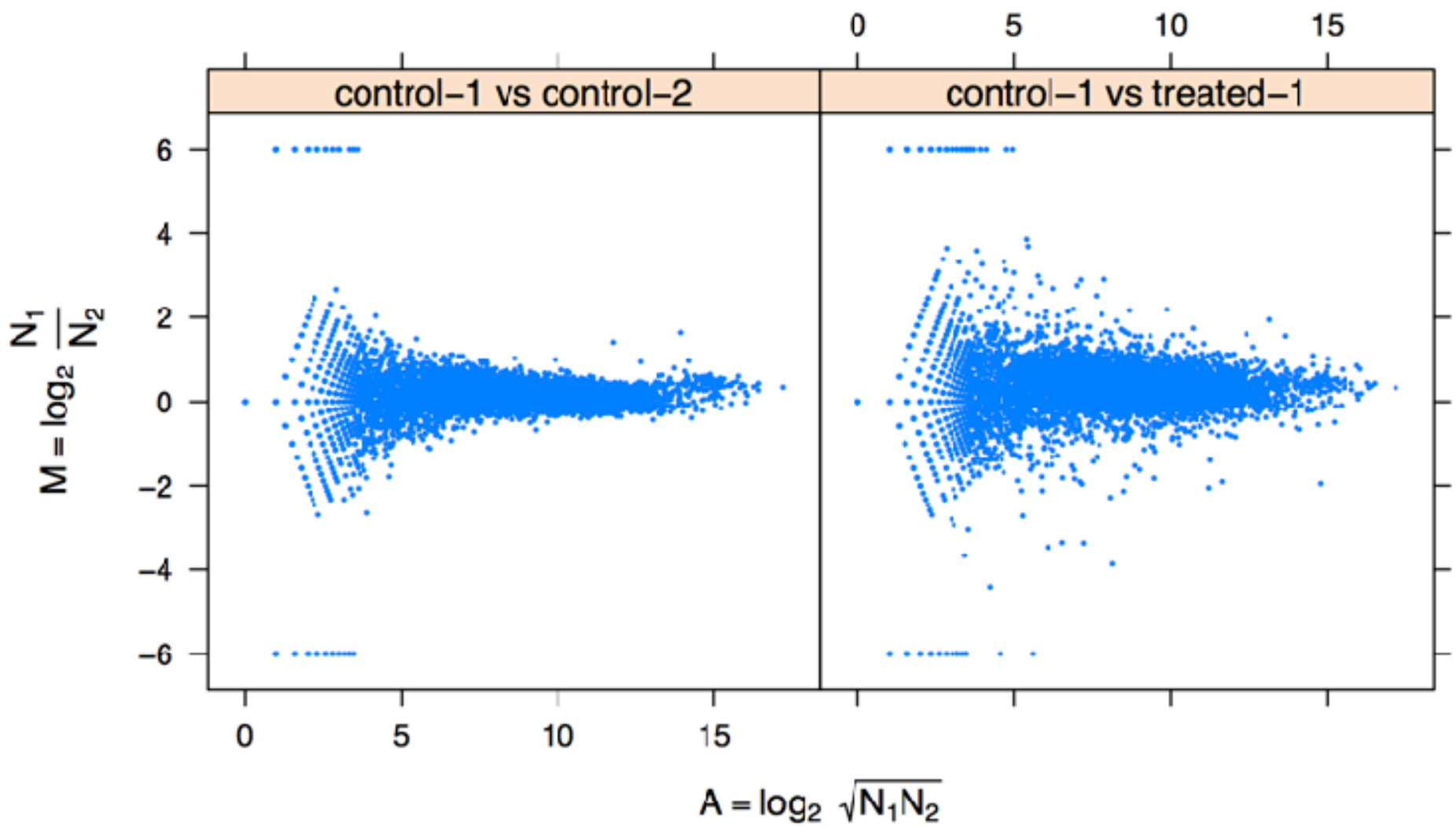
Modified slide from Mark Robinson

What is your biological question?

Given a gene, test for:

- **Whether transcripts levels change between conditions? (differential gene expression, DGE)**
- Whether transcript isoform proportions change between conditions? (differential transcript usage, DTU)
- Whether individuals exons are differentially used? (differential exon usage, DEU)

Differential problems: DGE

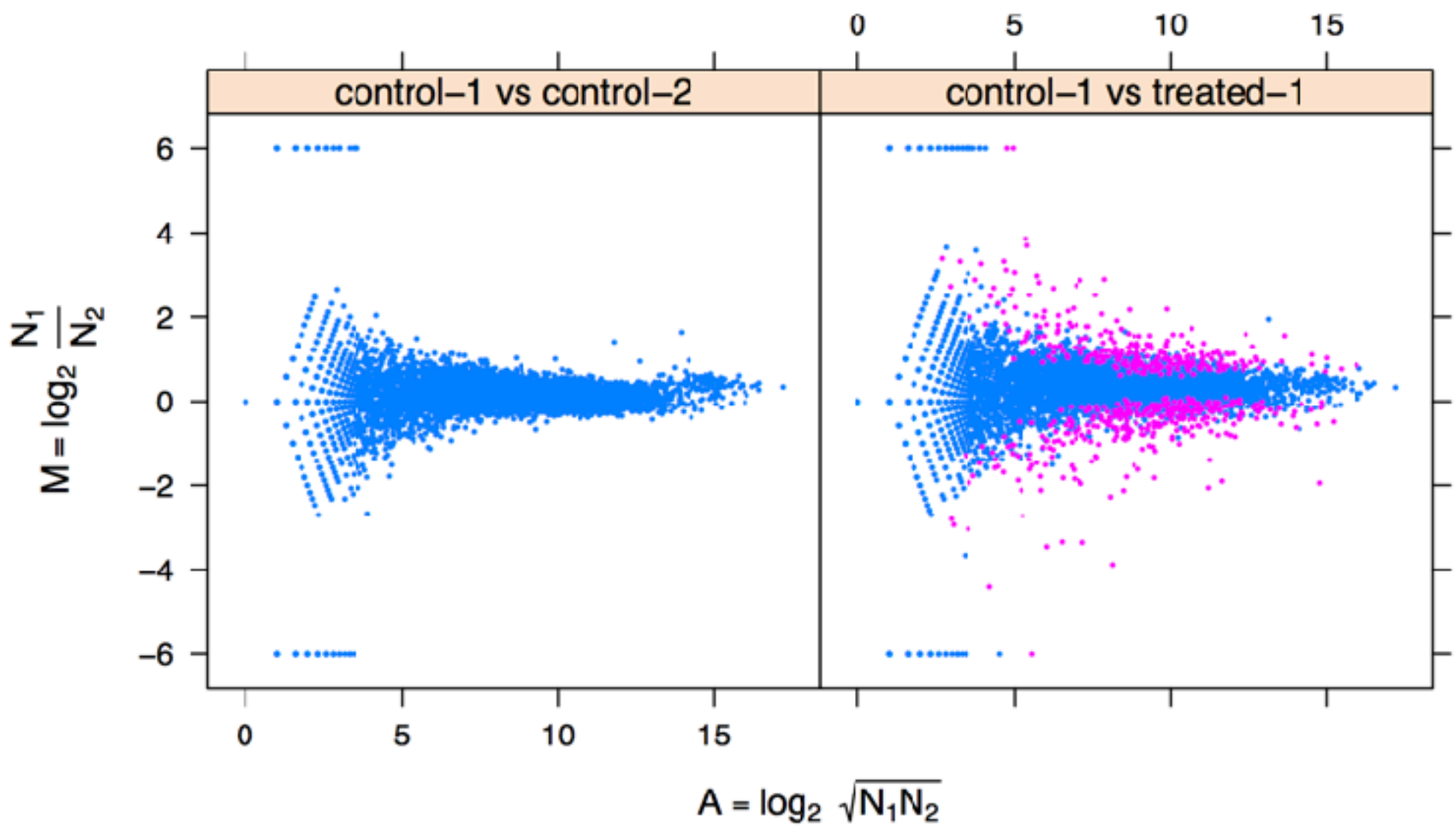


two biological replicates

treatment vs control

Modified from Wolfgang Huber

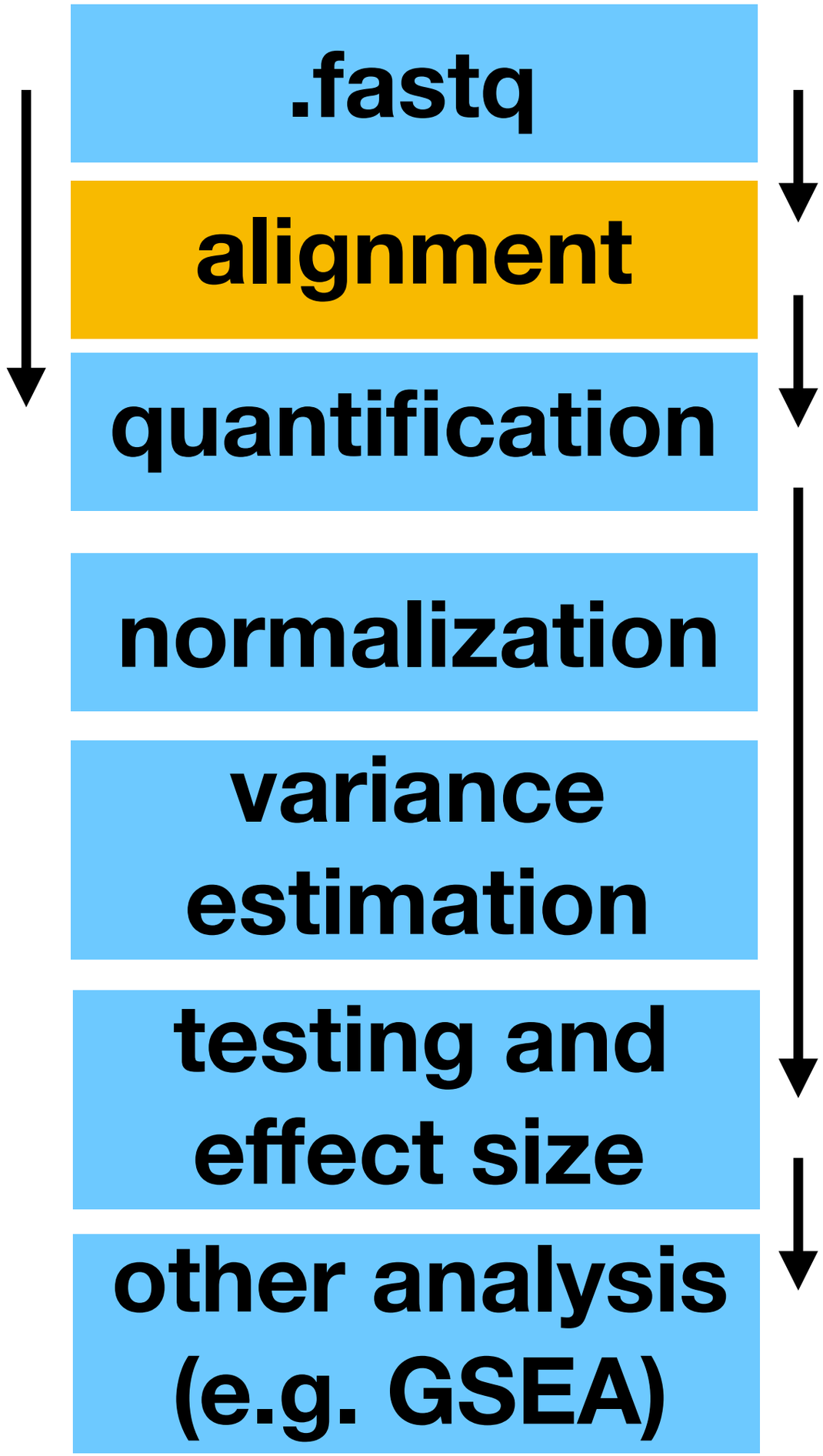
Differential problems: DGE

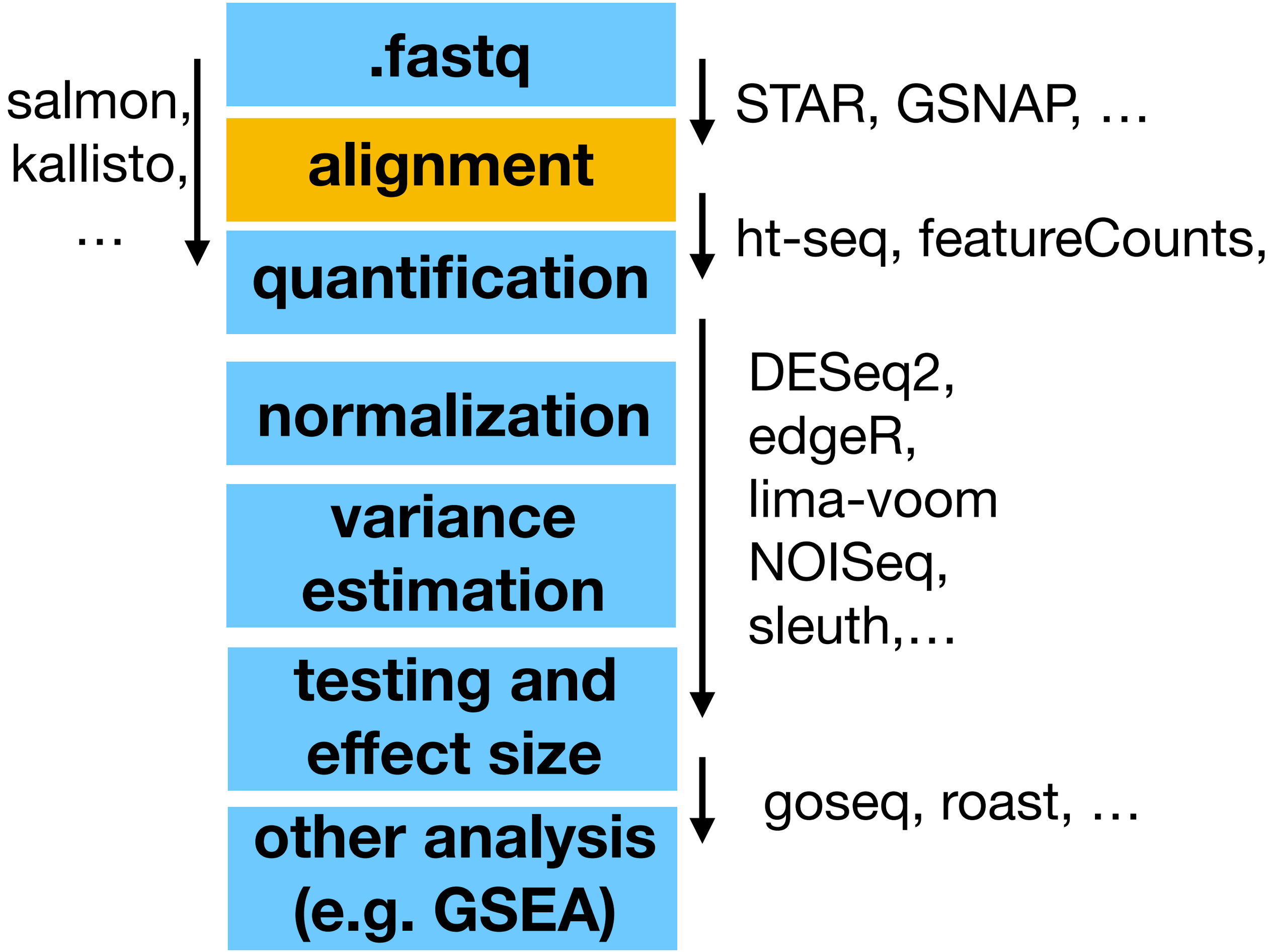


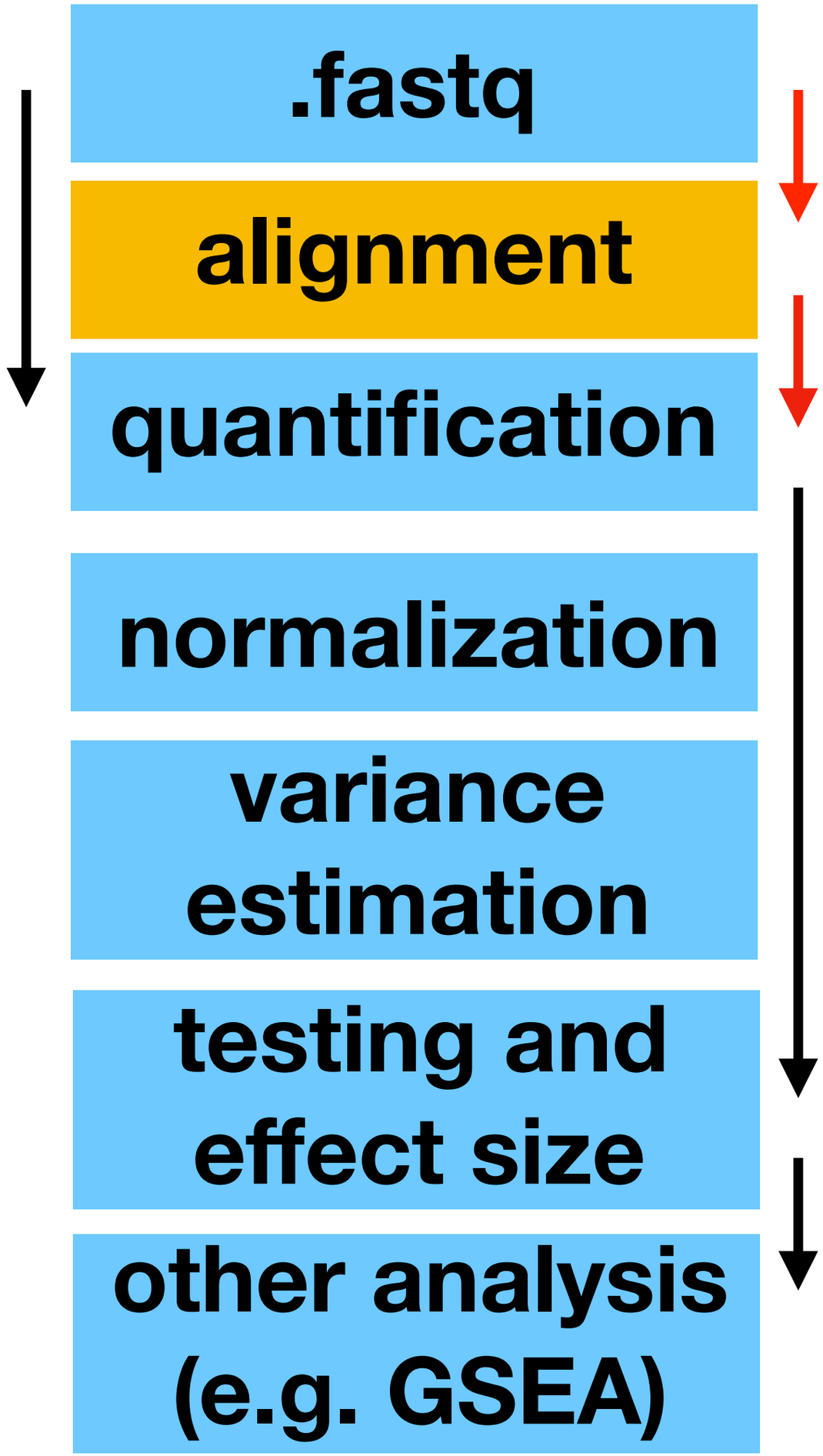
two biological replicates

treatment vs control

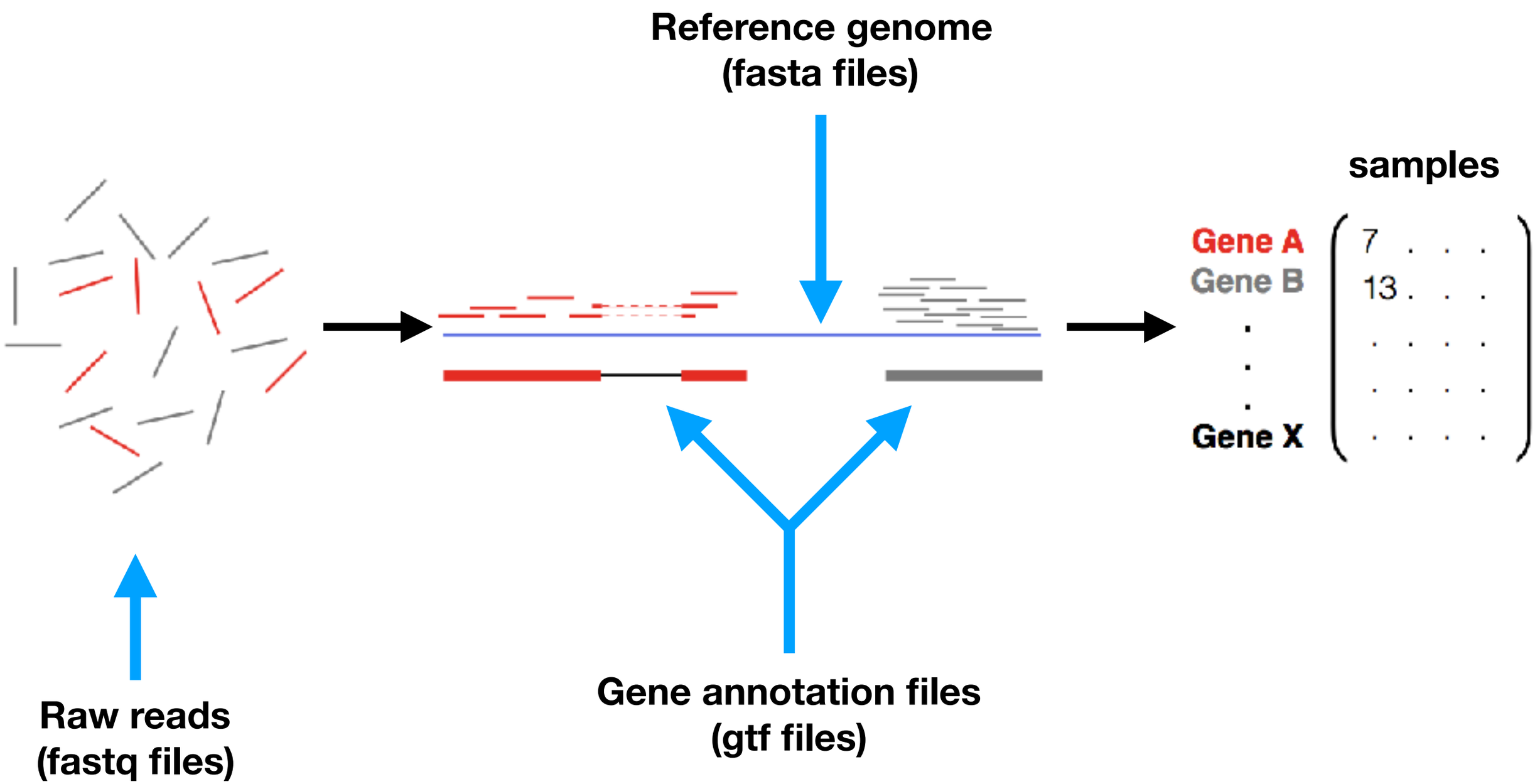
Modified from Wolfgang Huber


.fastq
alignment
quantification
normalization
variance estimation
testing and effect size
other analysis (e.g. GSEA)

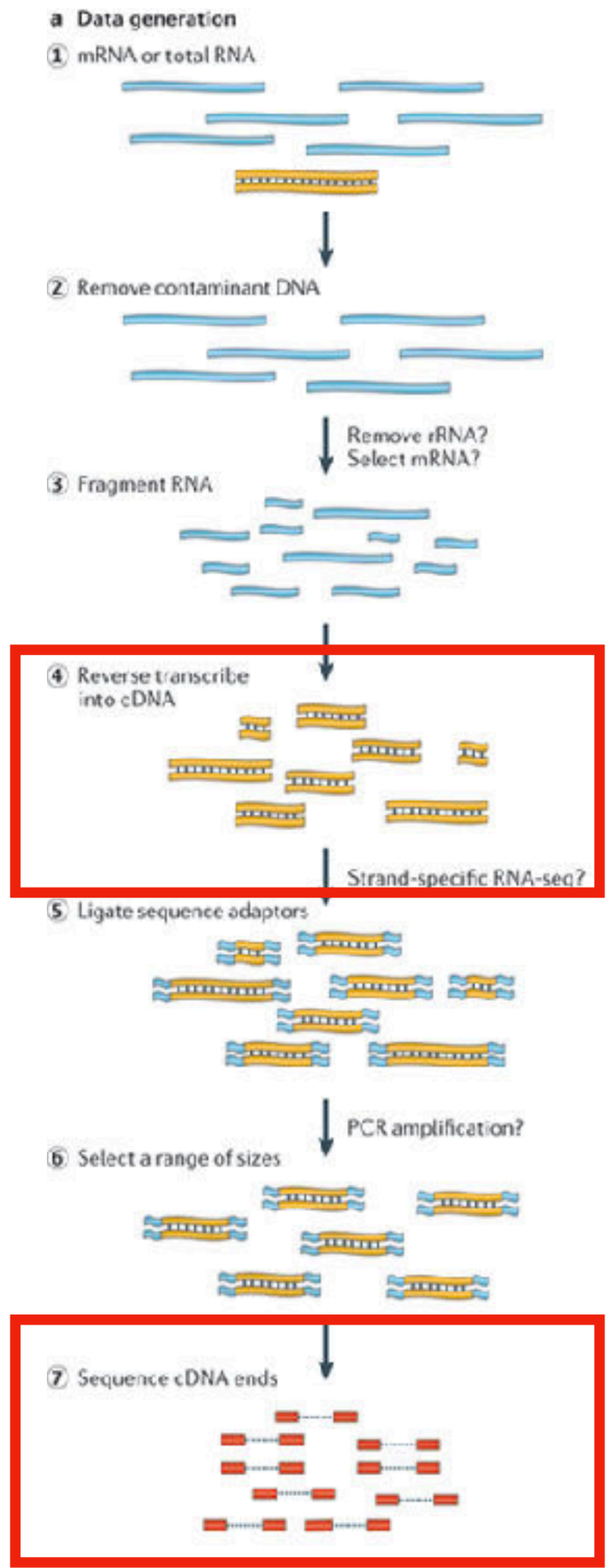

.fastq
alignment
quantification
normalization
variance estimation
testing and effect size
other analysis (e.g. GSEA)
salmon, kallisto, …
STAR, GSNAP, …
ht-seq, featureCounts,
DESeq2, edgeR, lima-voom NOISeq, sleuth,…
goseq, roast, …


.fastq
alignment
quantification
normalization
variance estimation
testing and effect size
other analysis (e.g. GSEA)

Alignment-based abundance estimation workflow

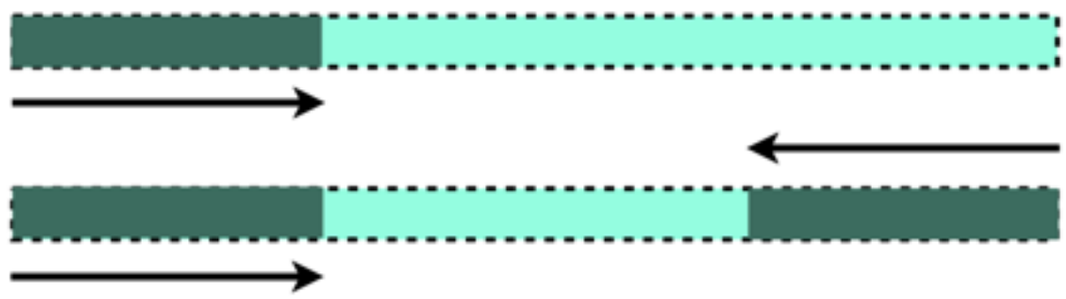


Modified from Charlotte Soneson



Martin et al, 2011

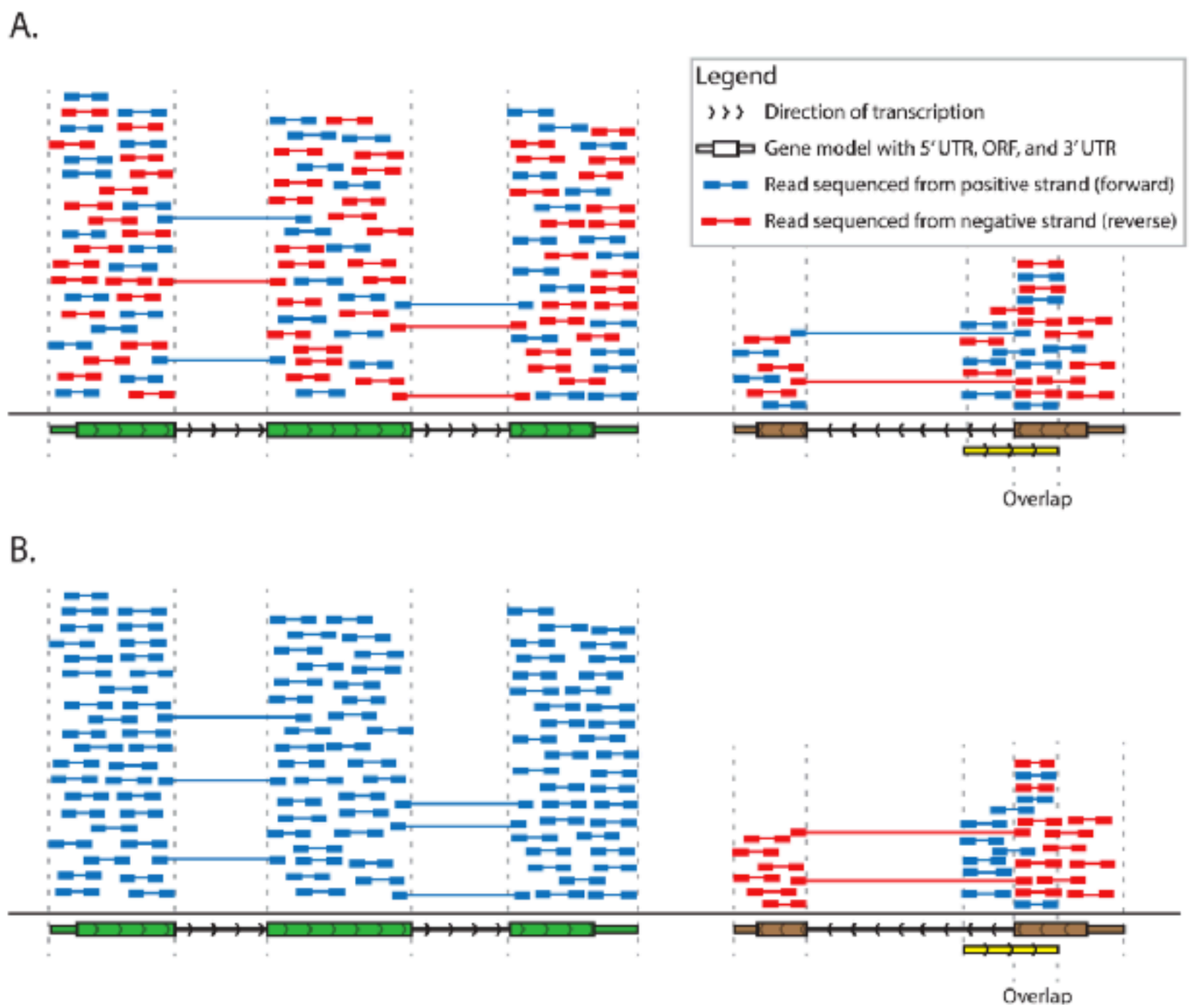
Are my data single-end or paired-end?



Tip: Look at the number of files per sample

Modified from Charlotte Soneson

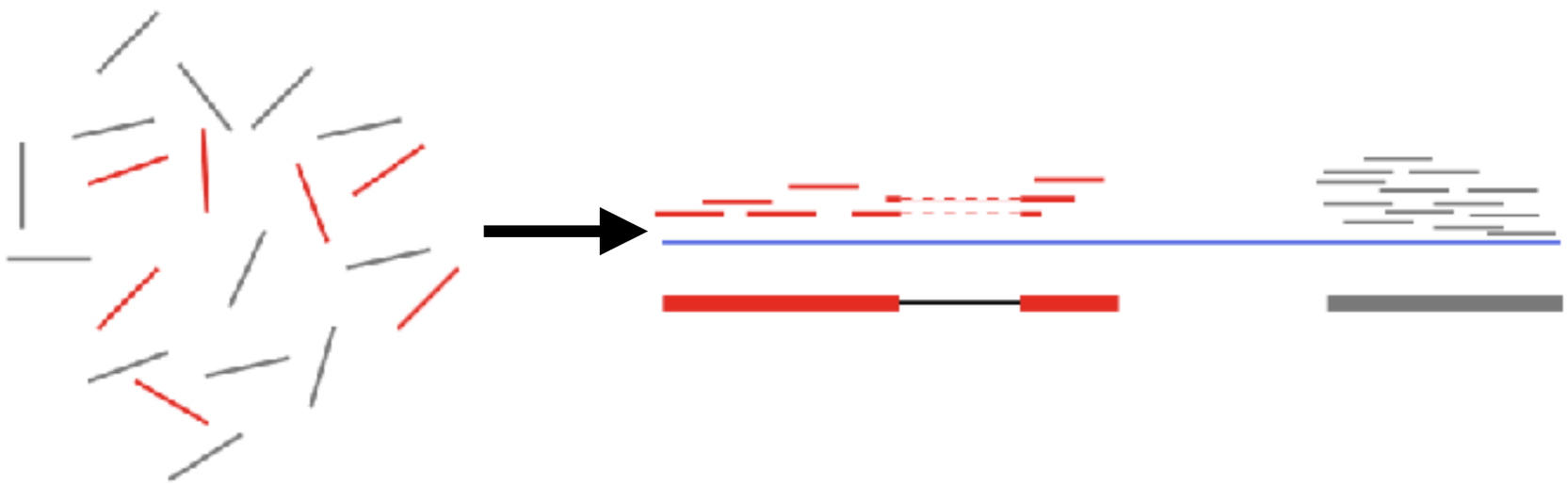
Are my data strand specific?



Tip: Visualize mapped reads in a genome browser (e.g. IGV)

Griffith et al, 2015

Alignment



Need a splice aware aligner
(e.g. STAR, GSNAP, …)

NATURE METHODS | ANALYSIS

Simulation-based comprehensive benchmarking of RNA-seq aligners

Giacomo Baruzzo, Katharina E Hayer, Eun Ji Kim, Barbara Di Camillo, Garret A FitzGerald & Gregory R Grant

NATURE METHODS | ANALYSIS OPEN

Systematic evaluation of spliced alignment programs for RNA-seq data

Pär G Engström, Tamara Steijger, Botond Sipos, Gregory R Grant, André Kahles, The RGASP Consortium, Gunnar Rätsch, Nick Goldman, Tim J Hubbard, Jennifer Harrow, Roderic Guigó & Paul Bertone

Example using STAR: index generation

number of threads

output folder - name according to genome

```
$ STAR --runThreadN 24 \
       --runMode genomeGenerate \
       --genomeDir my_genome \
       --genomeFastaFiles my_genome.fa \
       --sjdbGTFfile my_genes.gtf \
       --sjdbOverhang 99
```

reference genome

gene annotation file

read length - 1

Modified from Charlotte Soneson

Example using STAR: alignment

created index

```
$ STAR --runThreadN 24 \
       --runMode alignReads \
       --genomeDir my_genome \
       --readFilesIn S1_read1.fq.gz \
         S1_read2.fq.gz \
```

read file(s)

[include sample ID]

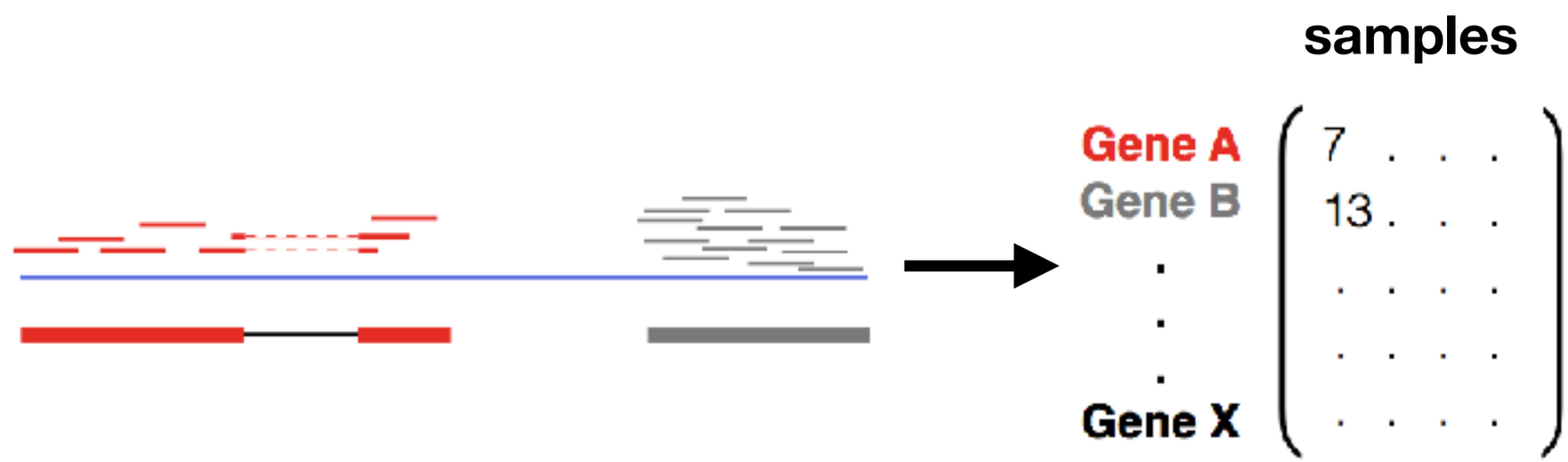
Modified from Charlotte Soneson

Aligner output: BAM files


```
9     1       11307   1       76M     =       11648   417     CTGCAGGGCCCTCTTGCTTACTGTATAGTGGT
JJJJJJJJJJJJIJJHIIJJIJJJJJJJJIJJJJDIFFFHFFDFDDEEEEEDDCDDDDDC    NH:i:4  HI:i:2  AS:i:150
5     1       11579   0       76M     =       11863   359     CAGAATTGTACTGTTCTGTATCCCACCAGCAA
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFHHIJJIJJJJJJIJ    NH:i:6  HI:i:3  AS:i:145
5     1       11606   0       69M7S   =       11652   460     AGCAATGTCTAGGAATACCTGTTTCTCCACAA
IJJJJHIJJJJJGIGHEG?FGGIGHJIIGGFIHIJJJBFHHIGIIGIJJJJCGFGGIJJI    NH:i:5  HI:i:4  AS:i:141
5     1       11625   0       47M338N29M      =       12083   534     TGTTTCTCCACAAAGTGTTTACTT
HHHHHJJHIHJJJJJJJJJIJJJJJJJJJJJJJJJJJJJDHIIJJJJJJJJJJJJJJJJJJJJHHH    NH:i:6  HI:i:5  AS:i:1
9     1       11648   1       76M     =       11307   -417    TTTGGATTTTTGCCAGTCTAACAGGTGAAGCC
JJJJIJJJJJIJJIJIJJJJJJJJJIJJJJJJJJJJJJJJJJJJJHHHHHFFFFFCCC    NH:i:4  HI:i:2  AS:i:150
```

read name **flag** **chr** **pos** **mapq** **CIGAR**

Alignment-based abundance estimation workflow



Modified from Charlotte Soneson

```
$ htseq-count  --format bam \
               --order=pos \
               --stranded=no \
               --type=exon \
               --idattr=gene_id \
               --mode=union \
               aligned.bam \
               my_genes.gtf
```

default=yes

check you
GTF file!

mapped reads

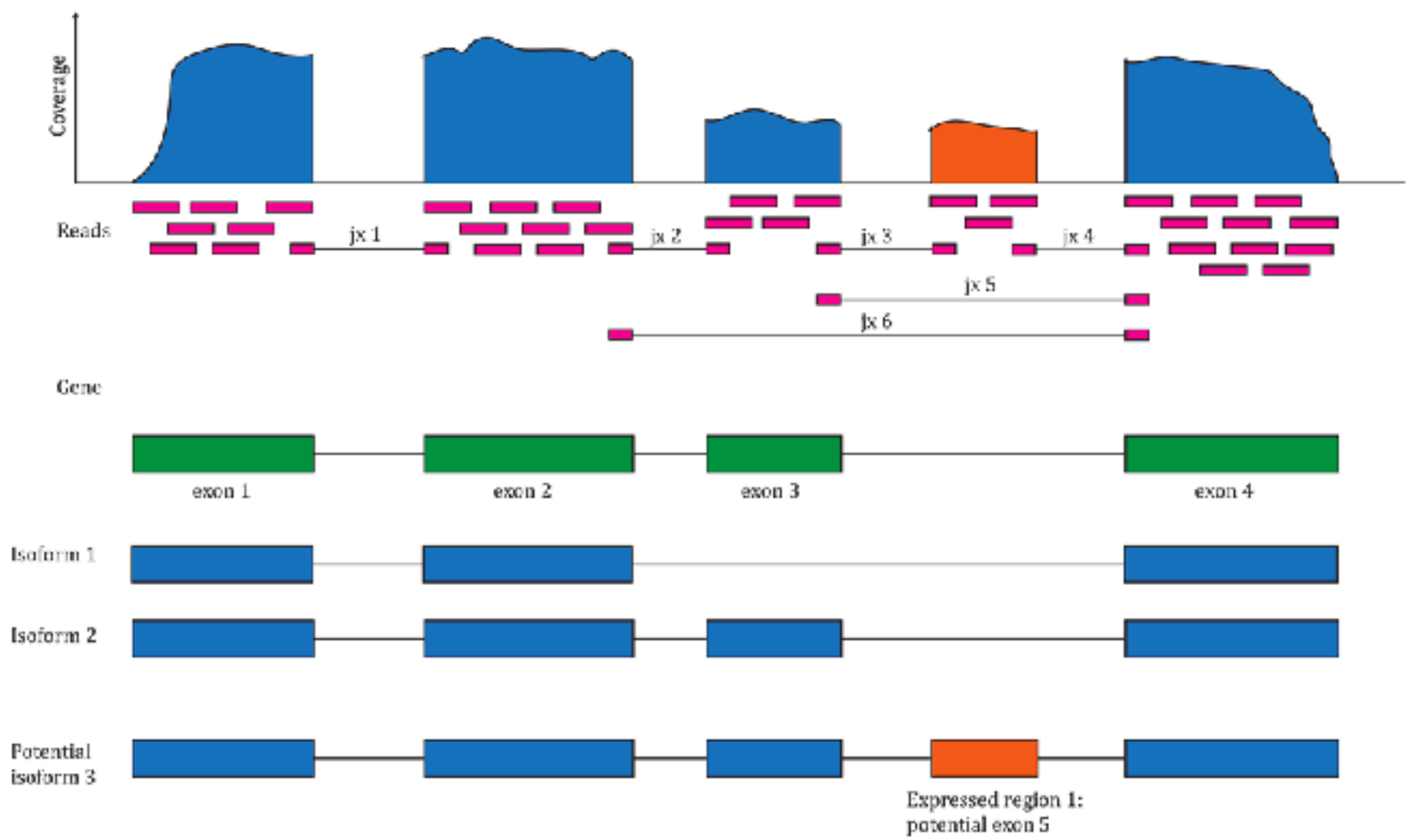
```
    protein_coding   exon  5139815    5141712    .    -    .    gene_id "FBgn0020621"; transcript_id
Btr0112897"; exon_number "10"; gene_name "Pkn"; gene_biotype "protein_coding"; transcript_name "Pkn-RG",
on_id "FBgn0020621:1";
```

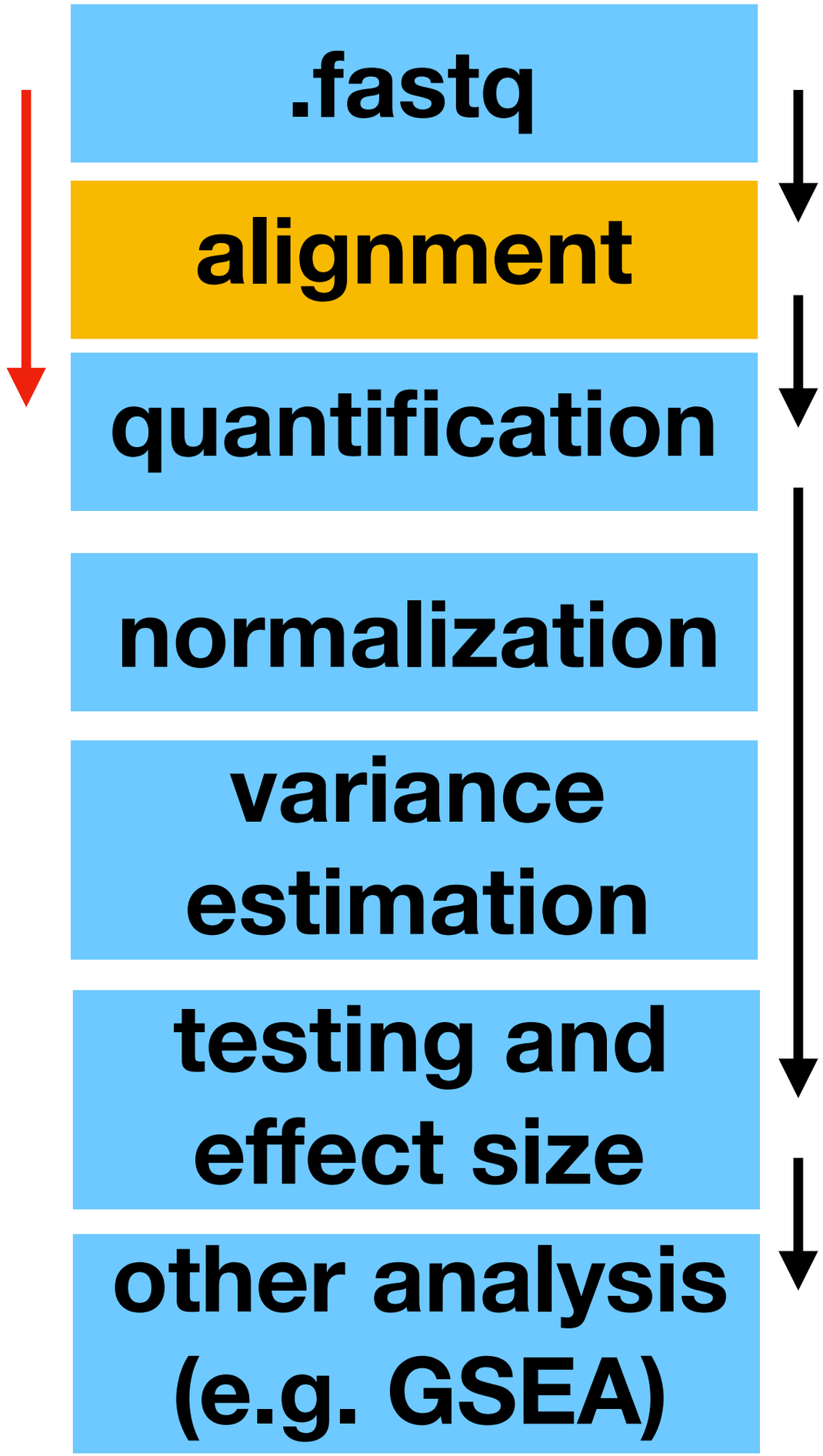
gene
annotation file (GTF)

Modified from Charlotte Soneson

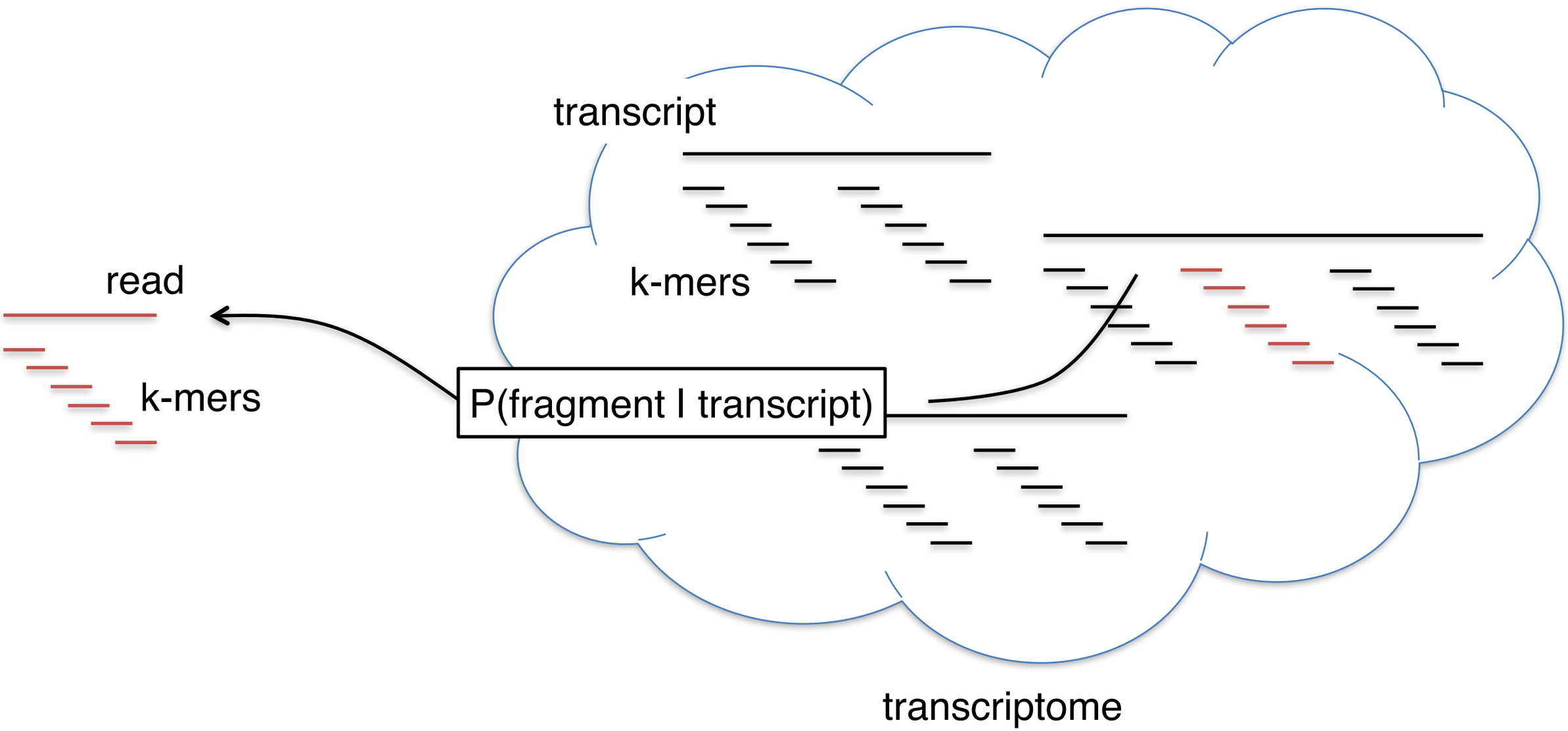
recount2

70,000 human RNA-seq samples already processed



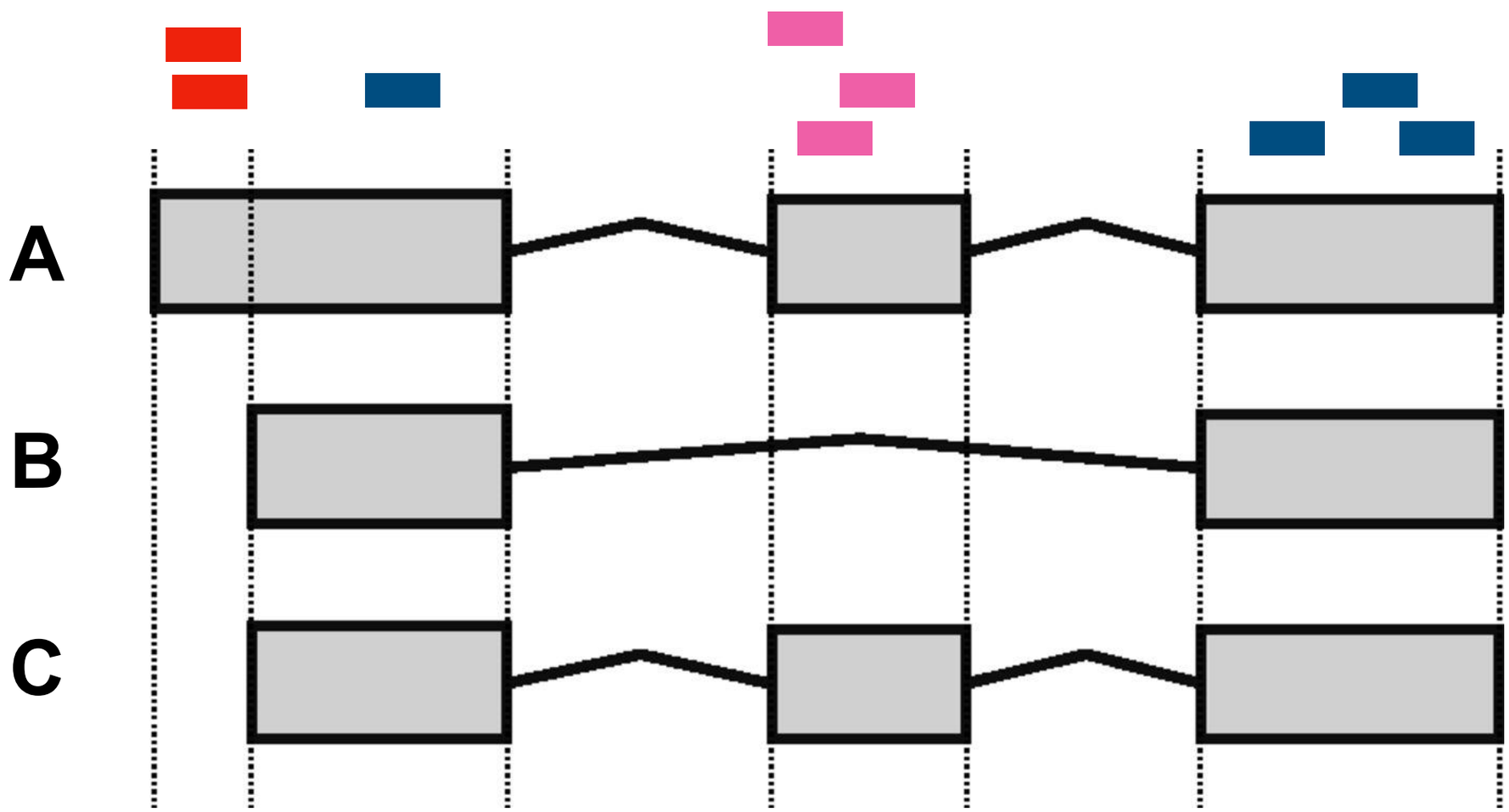

.fastq
alignment
quantification
normalization
variance estimation
testing and effect size
other analysis (e.g. GSEA)

Alignment-free abundance estimation workflows



Extremely fast compared to genome alignments!

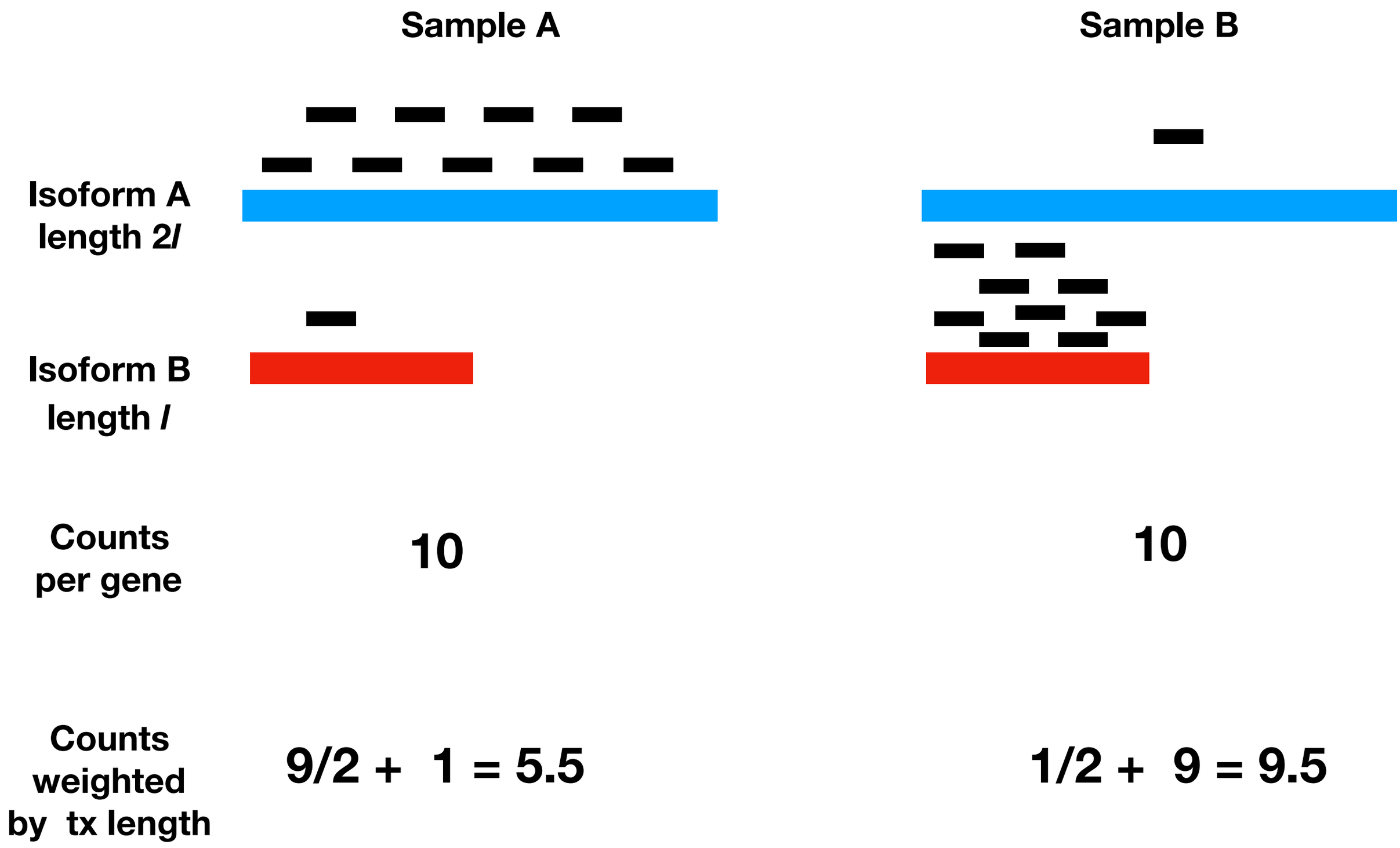
Alignment-free abundance estimation workflows



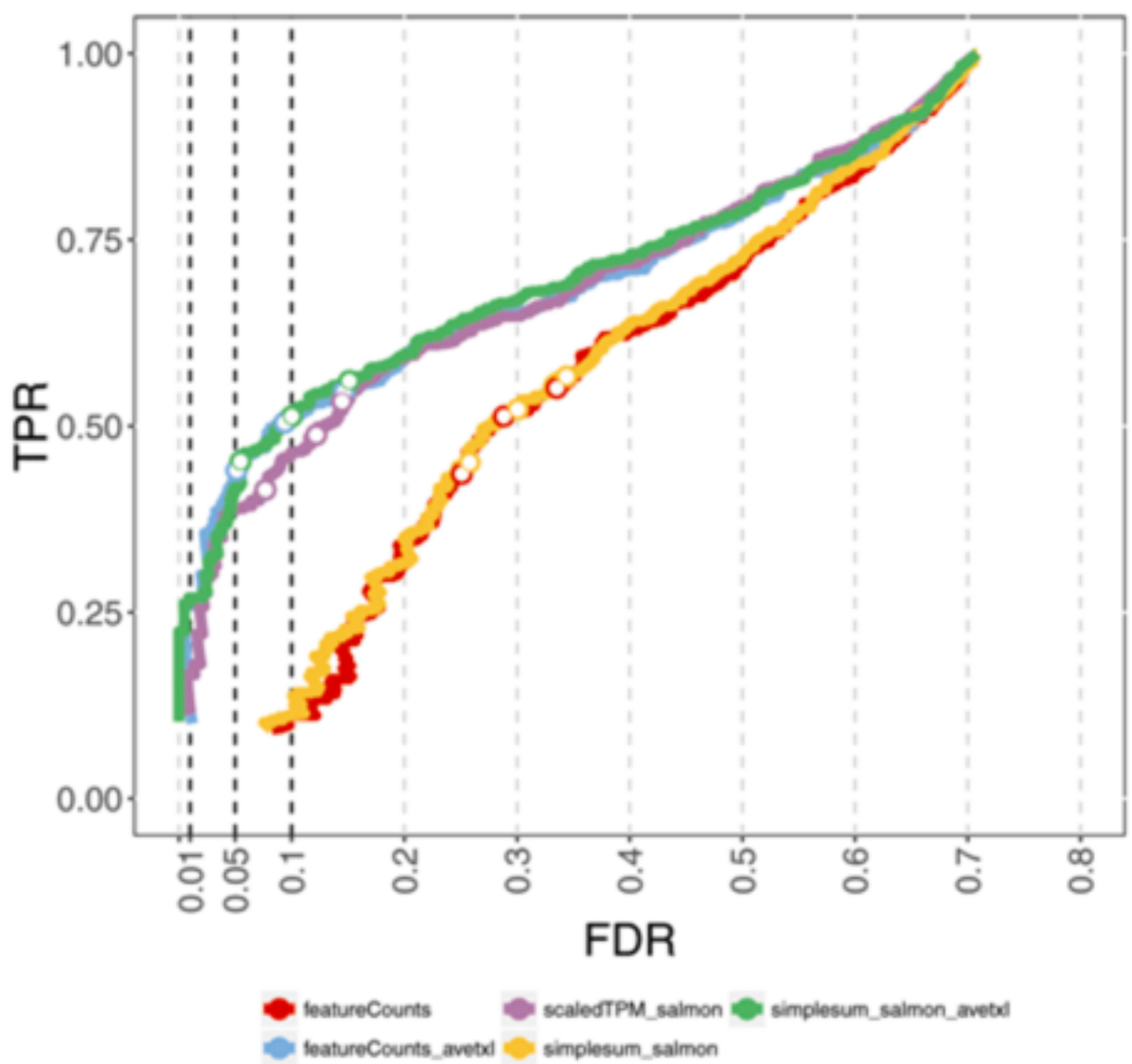
Equivalence classes: {A} = 2, {A,C}=3, {A, B, C}=4

Objective: estimate the transcript abundances that best explain the observed counts for the reference classes.

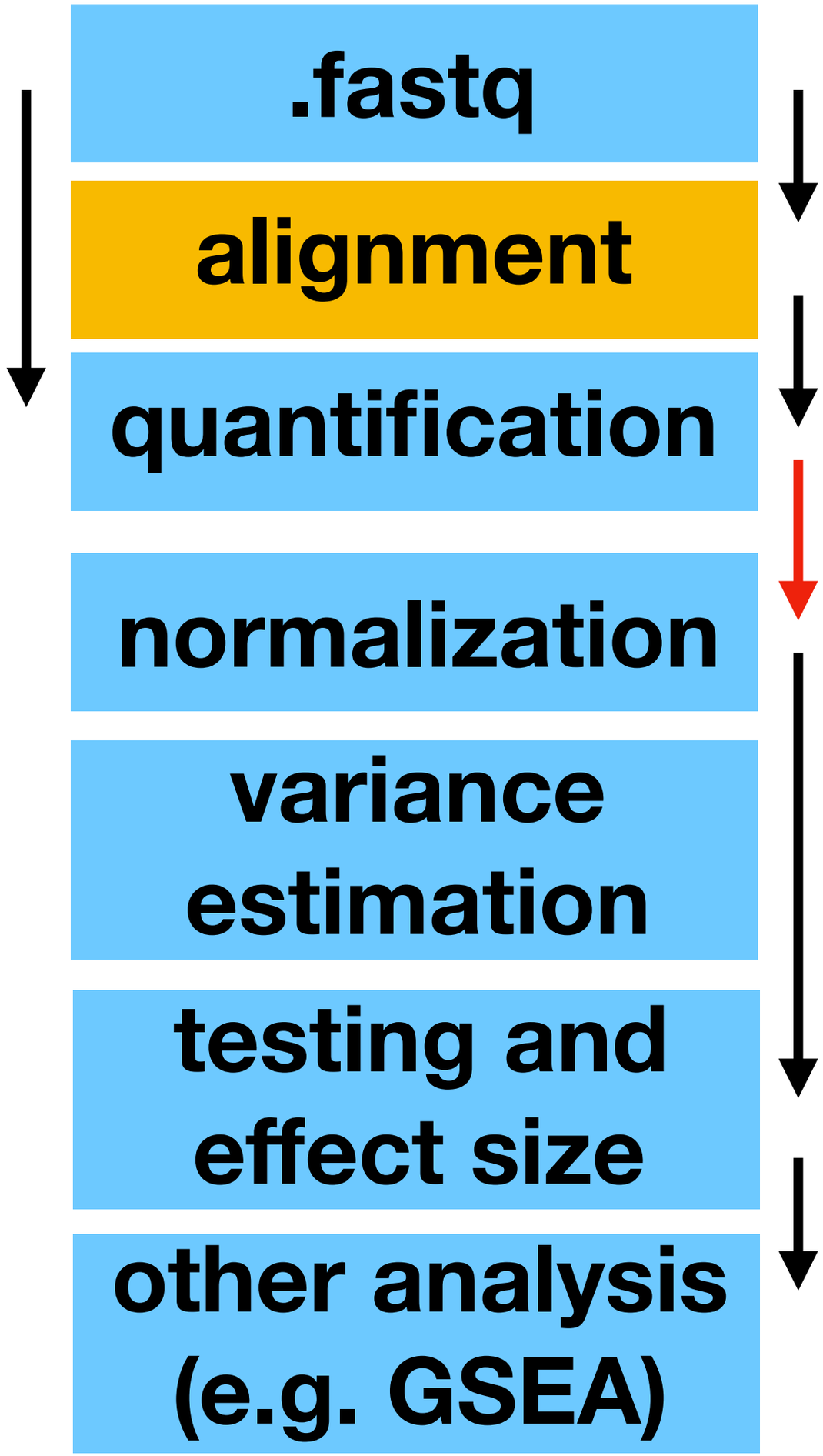
Why is it important to consider transcript length information? (hypothetical example)



Considering average transcript lengths improves estimates of DGE

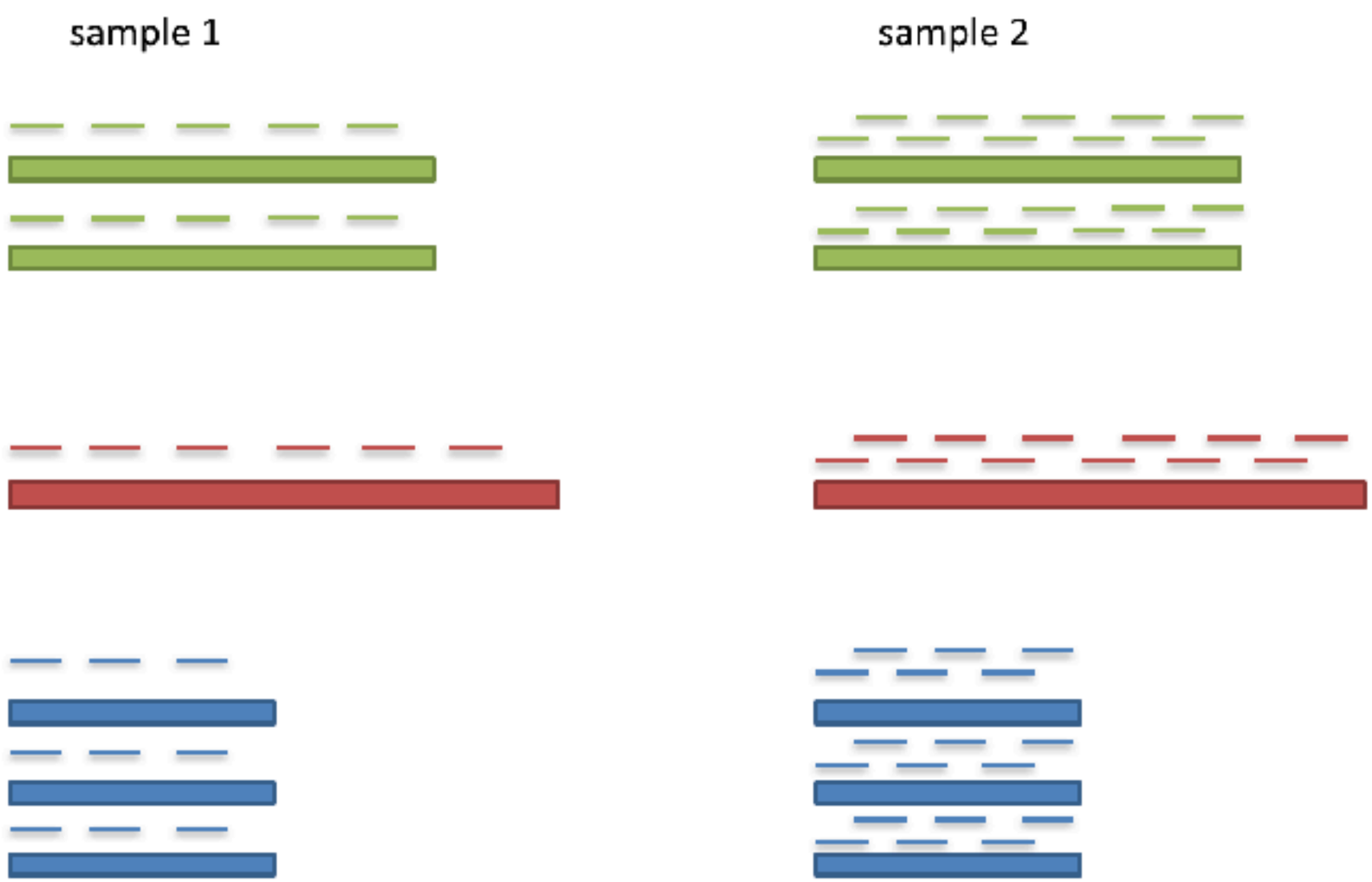


Soneson et al, 2015


.fastq
alignment
quantification
normalization
variance estimation
testing and effect size
other analysis (e.g. GSEA)

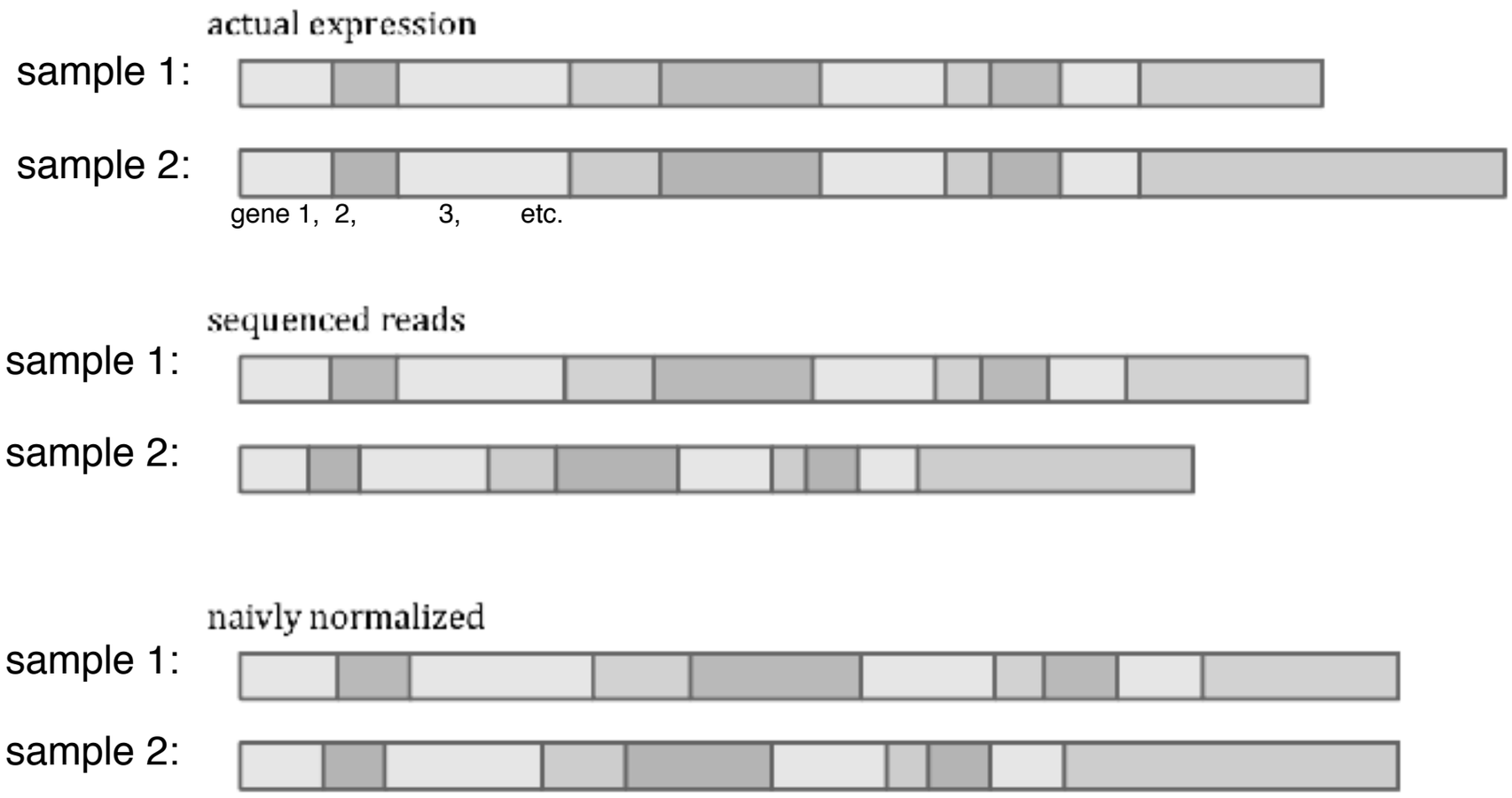
$$
\begin{array}{c}
\textbf{Gene A} \\
\textbf{Gene B} \\
\vdots \\
\textbf{Gene X}
\end{array}
\begin{pmatrix}
7 & \cdot & \cdot & \cdot \\
13 & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot
\end{pmatrix}
$$

Sequencing depth



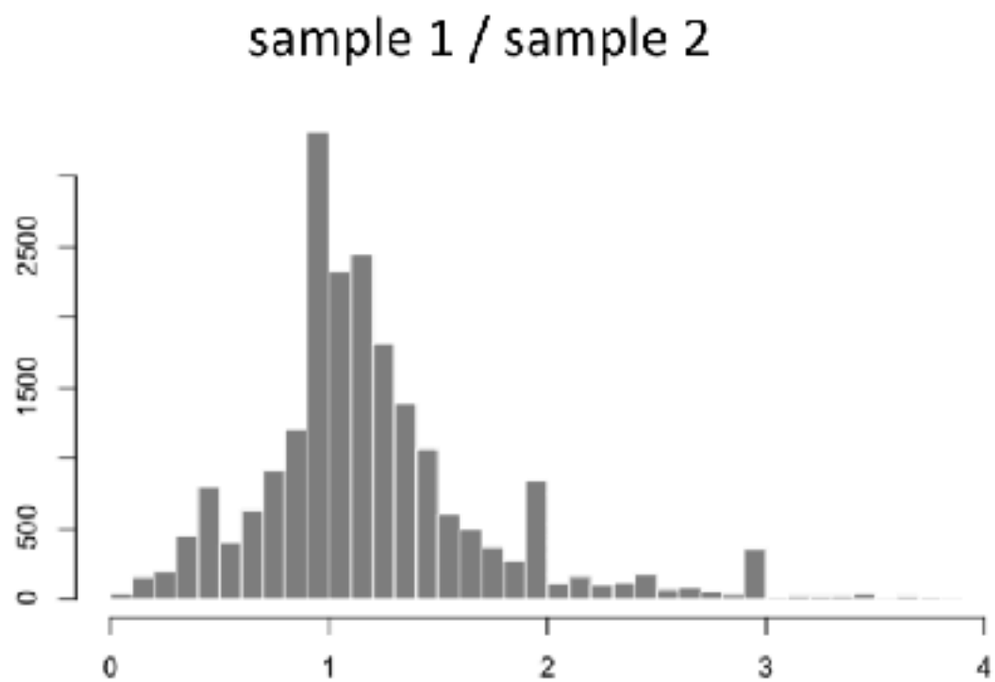
Modified from Mike Love

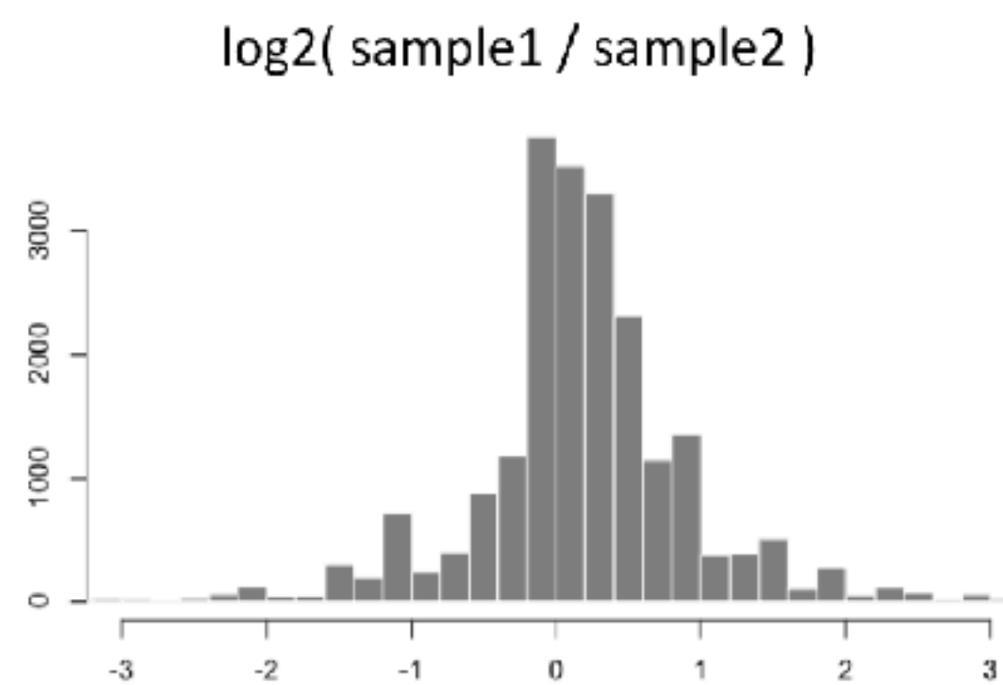
Normalization for sequencing depth



Modified from Simon Anders

DESeq2 uses a median of ratios method





- Create a reference sample by calculating the geometric mean across samples for each gene.
- For each sample, take the ratio with respect to the reference sample.
- Take the median across genes for each sample.

Modified from Mike Love

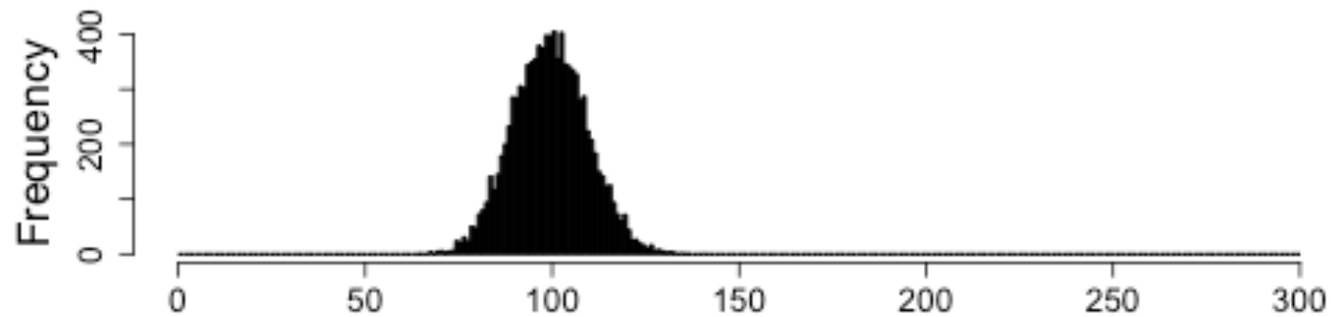
DESeq2 scaling factors or normalization factors?

s_j	s_{ij}
"size factor"	"normalization factor"
per sample *j*	per sample *j* & per gene *i*
sequencing depth	sequencing depth and other factors differ across samples (technical bias: cqn or EDASeq) (gene length: tximport)
robustly estimated with median ratio	median ratio method for sequencing depth can be estimated on top

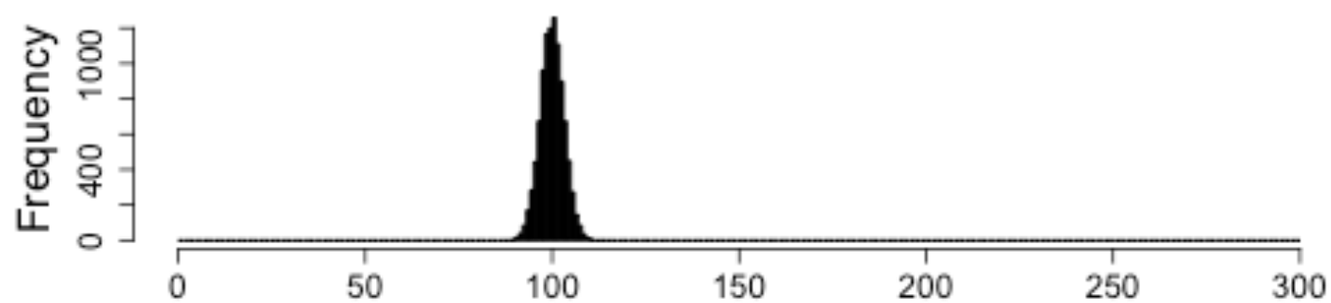
Modified from Mike Love

But, why counts and not other transformed data (e.g. FPKMs)?

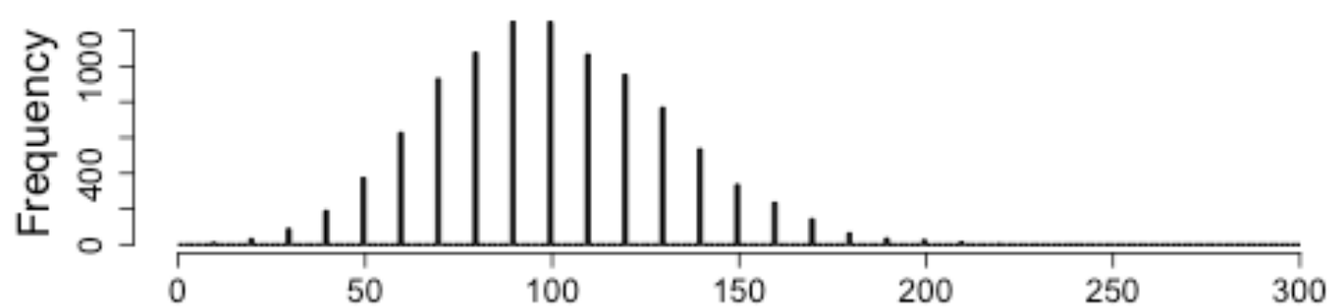
Raw count with mean of 100
Poisson sampling, so SD=10

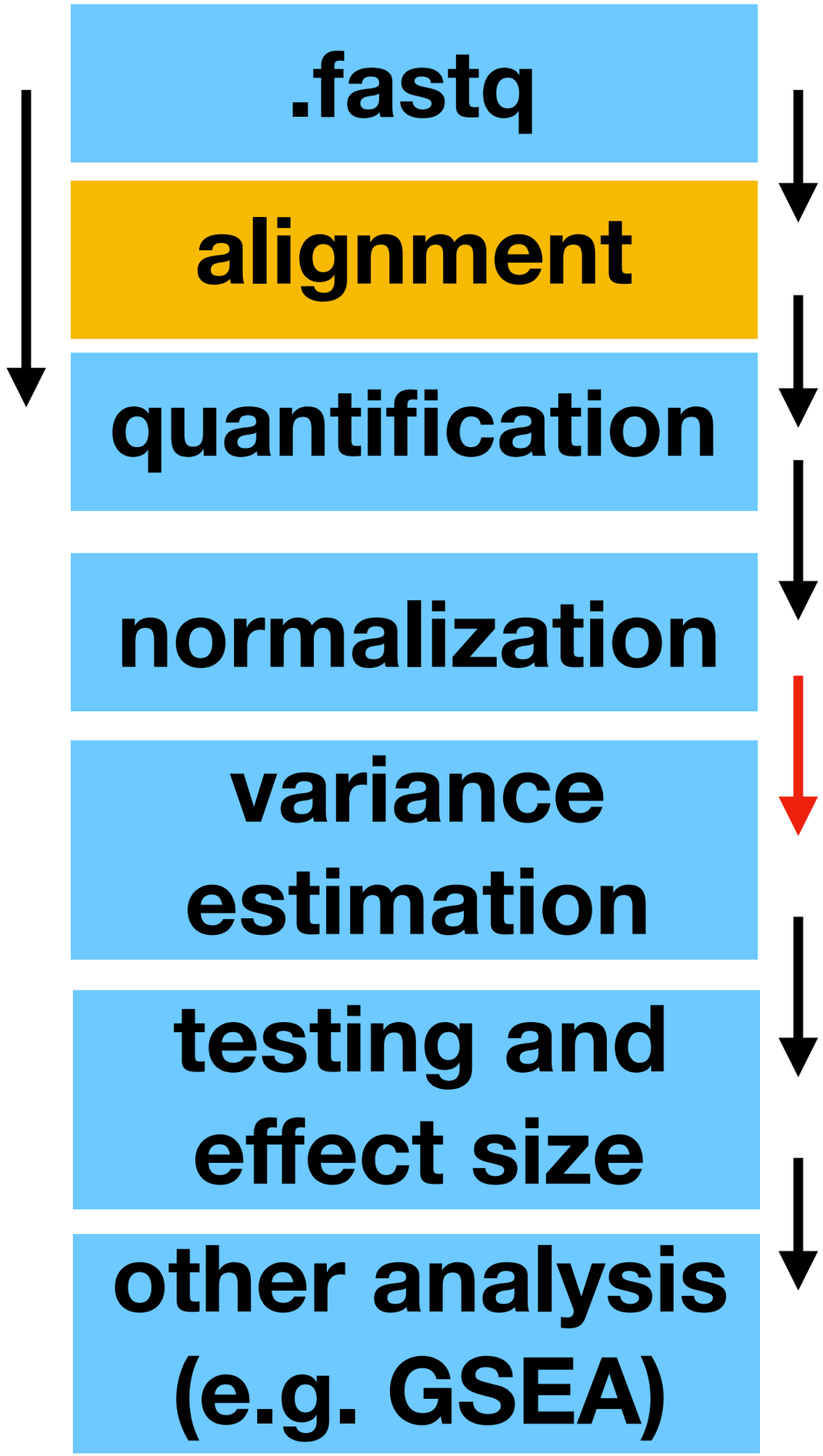


Raw count with mean of 100
scale by 1/10
SD = ?

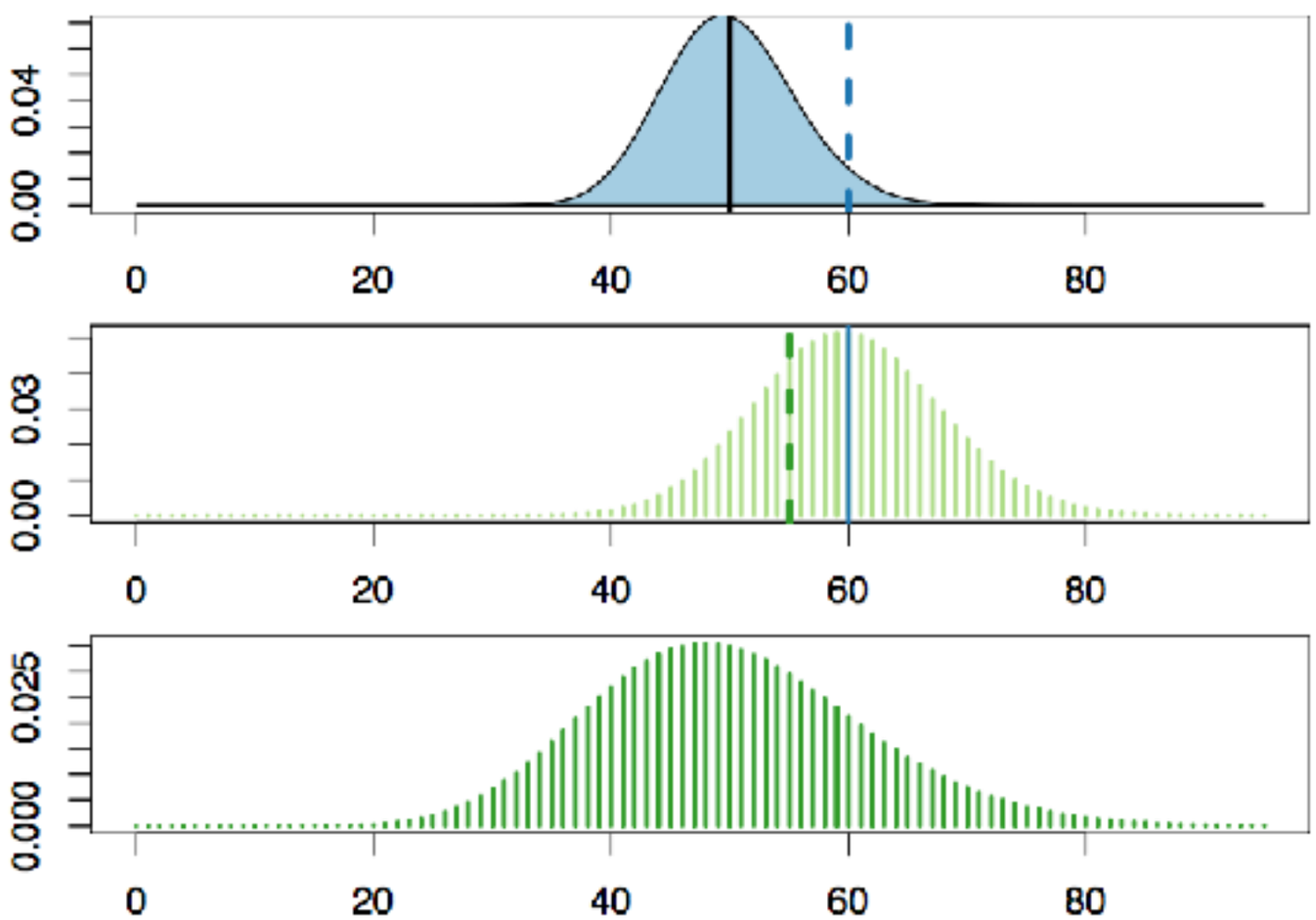


Raw count with mean of 100
scale by 10
SD = ?




.fastq
alignment
quantification
normalization
variance estimation
testing and effect size
other analysis (e.g. GSEA)

Variance of a gene: technical noise + biological noise



Biological sample with variance Γ

Poisson sampling variance = Λ

Gamma-poisson process

$$NB(\mu, \sigma^2 + \mu) = \Lambda(\Gamma(\mu, \sigma^2))$$

Poisson process in which the mean is gamma distributed

gene *j*

gene *i*

$$\begin{pmatrix} 7 & \cdot & \cdot & \cdot \\ 13 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

normalization factor

quantity of interest

dispersion parameter

$$K_{ij} \sim \mathrm{NB}(s_{ij} q_{ij}, \alpha_i)$$

$$\mathrm{Var}(K_{ij}) = \underbrace{\mu_{ij}}_{\text{technical noise}} + \alpha_i \underbrace{\mu_{ij}^2}_{\text{extra biological variability}}$$

Challenge: small number of replicates!

Dispersion estimation

$$\hat{\alpha}_{\mathrm{MLE}} = \underset{\alpha}{\operatorname{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu}))$$

$$\mathrm{CR}(\alpha) = -\frac{1}{2}\log(\det(X^t W X))$$

$$\hat{\alpha}_{\mathrm{CR}} = \underset{\alpha}{\operatorname{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu}) + \mathrm{CR}(\alpha))$$

$$\mathrm{prior}(\alpha) = f_{\mathcal{N}}(\log(\alpha); \log(\alpha_{\mathrm{fit}}), \sigma^2_{\alpha\text{-prior}})$$

$$\hat{\alpha}_{\text{CR-MAP}} = \underset{\alpha}{\operatorname{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu}) + \mathrm{CR}(\alpha) + \log(\mathrm{prior}(\alpha)))$$

Dispersion estimation

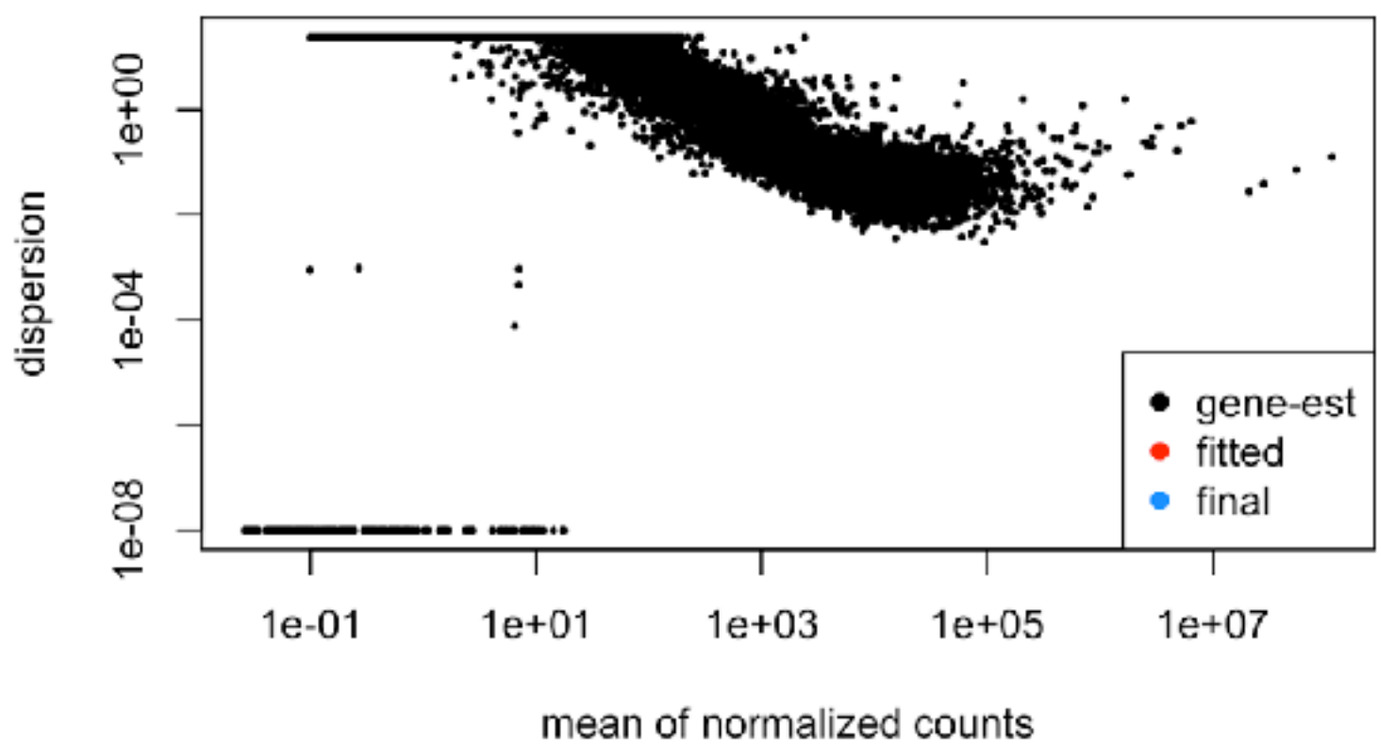
$$\hat{\alpha}_{\text{MLE}} = \underset{\alpha}{\text{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu}))$$ ← **maximum-likelihood estimates**

$$\text{CR}(\alpha) = -\frac{1}{2}\log(\det(X^t W X))$$ ← **Cox-Reid bias term**

$$\hat{\alpha}_{\text{CR}} = \underset{\alpha}{\text{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu}) + \text{CR}(\alpha))$$ ← **Cox-Reid ML estimate**

$$\text{prior}(\alpha) = f_{\mathcal{N}}(\log(\alpha); \log(\alpha_{\text{fit}}), \sigma^2_{\alpha\text{-prior}})$$

$$\hat{\alpha}_{\text{CR-MAP}} = \underset{\alpha}{\text{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu}) + \text{CR}(\alpha) + \log(\text{prior}(\alpha)))$$



Dispersion estimation

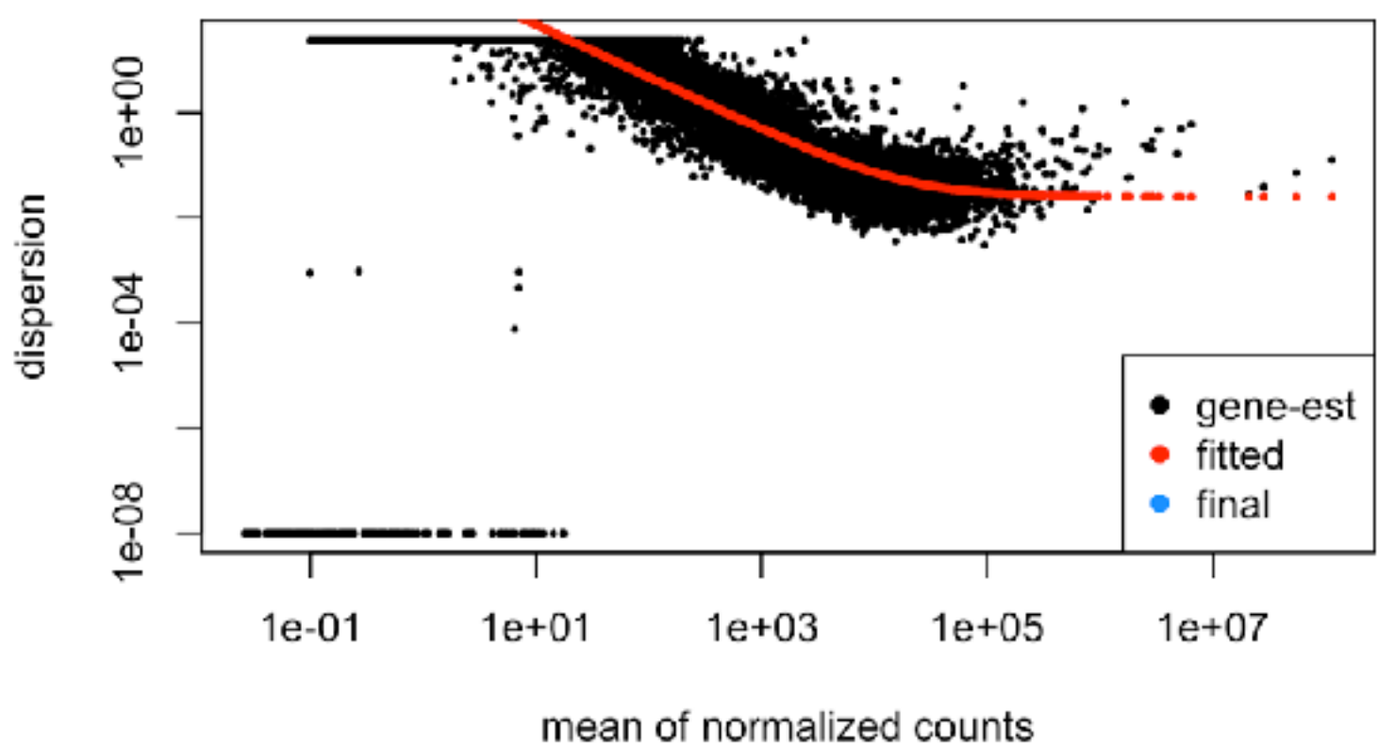
$$\hat{\alpha}_{\text{MLE}} = \underset{\alpha}{\text{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu}))$$ ← **maximum-likelihood estimates**

$$\text{CR}(\alpha) = -\frac{1}{2}\log(\det(X^t W X))$$ ← **Cox-Reid bias term**

$$\hat{\alpha}_{\text{CR}} = \underset{\alpha}{\text{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu}) + \text{CR}(\alpha))$$ ← **Cox-Reid ML estimate**

$$\text{prior}(\alpha) = f_{\mathcal{N}}(\log(\alpha); \log(\alpha_{\text{fit}}), \sigma^2_{\alpha\text{-prior}})$$ ← **alpha prior by information sharing across genes**

$$\hat{\alpha}_{\text{CR-MAP}} = \underset{\alpha}{\text{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu}) + \text{CR}(\alpha) + \log(\text{prior}(\alpha)))$$



Dispersion estimation

$$\hat{\alpha}_{\text{MLE}} = \underset{\alpha}{\text{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu}))$$ ← **maximum-likelihood estimates**

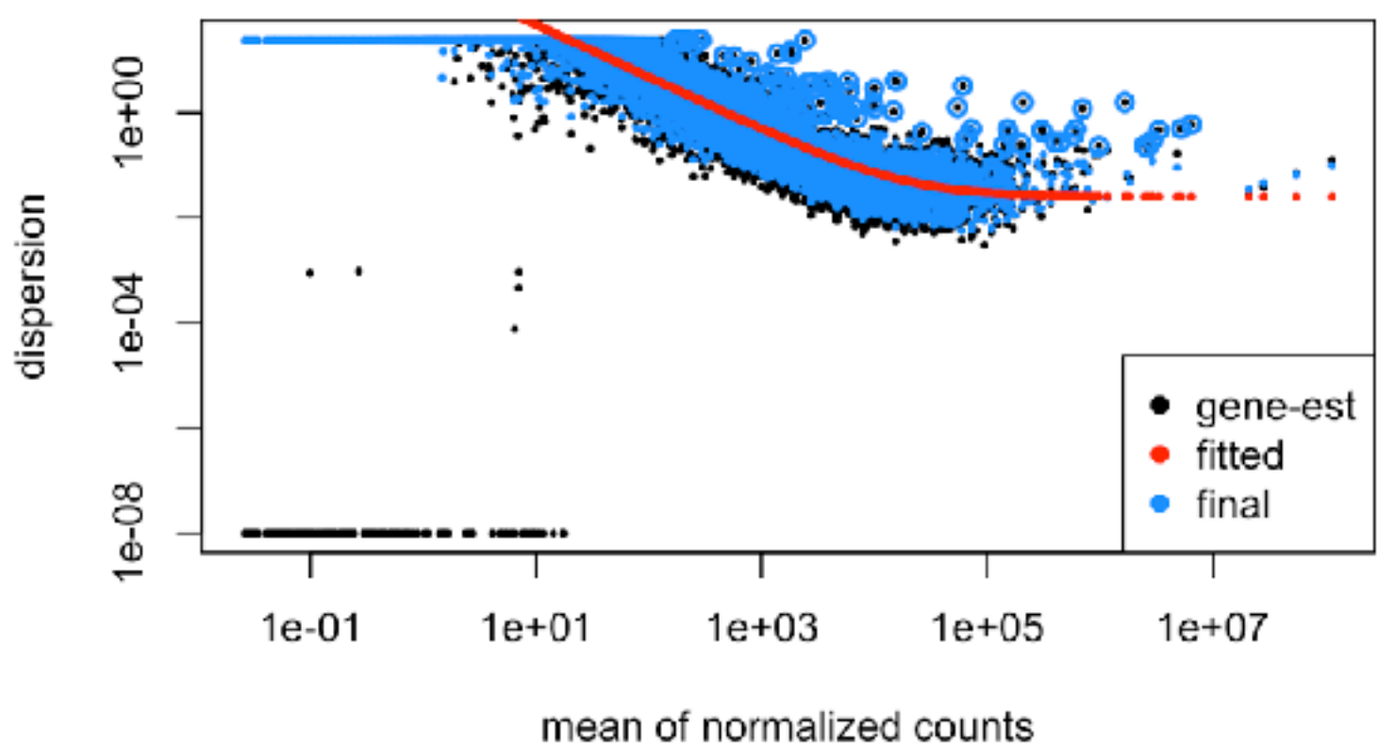
$$\text{CR}(\alpha) = -\frac{1}{2}\log(\det(X^t W X))$$ ← **Cox-Reid bias term**

$$\hat{\alpha}_{\text{CR}} = \underset{\alpha}{\text{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu}) + \text{CR}(\alpha))$$ ← **Cox-Reid ML estimate**

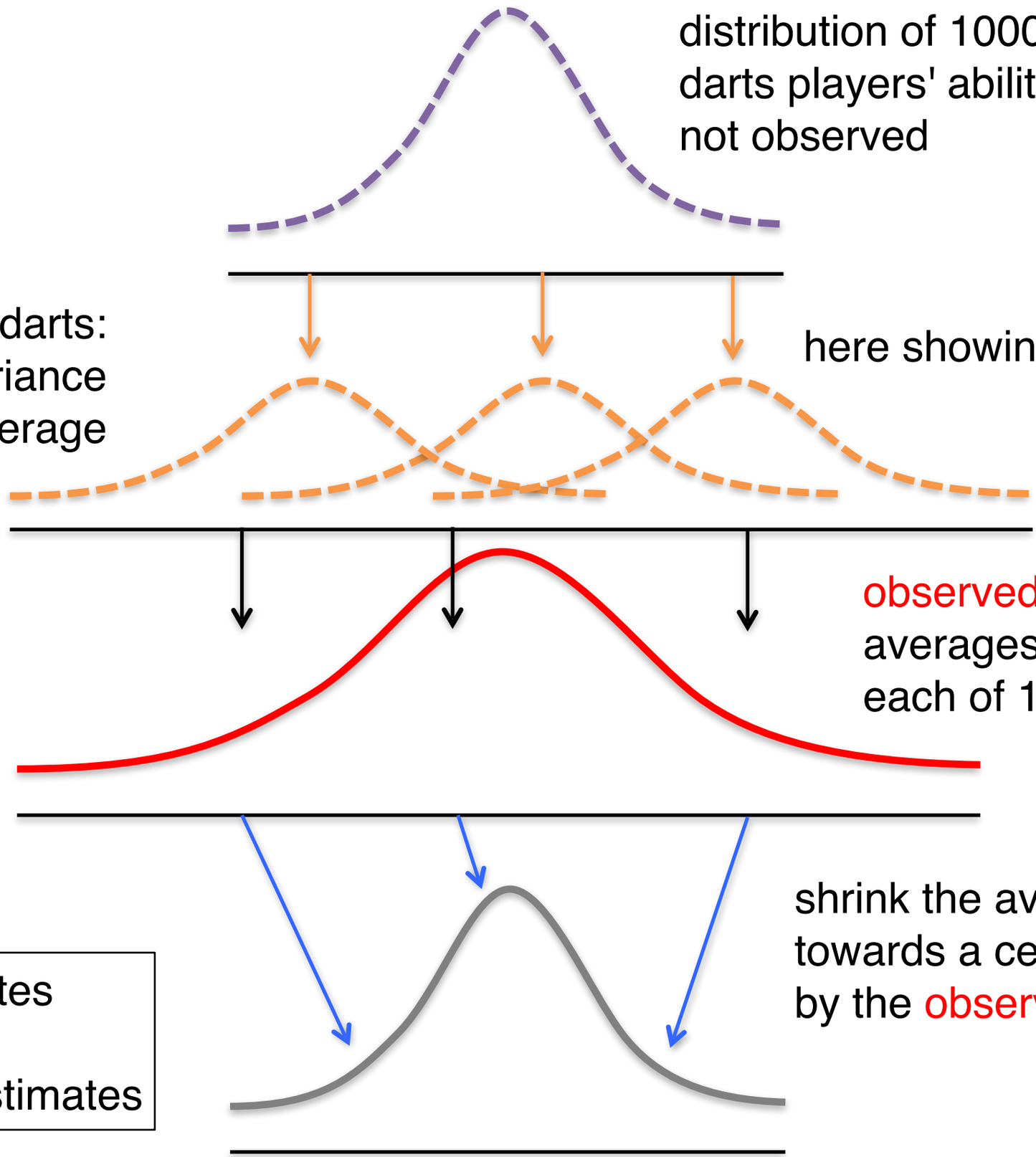
$$\text{prior}(\alpha) = f_{\mathcal{N}}(\log(\alpha); \log(\alpha_{\text{fit}}), \sigma^2_{\alpha\text{-prior}})$$ ← **alpha prior by information sharing across genes**

$$\hat{\alpha}_{\text{CR-MAP}} = \underset{\alpha}{\text{argmax}}(\ell(\alpha|\vec{k}, \hat{\mu}) + \text{CR}(\alpha) + \log(\text{prior}(\alpha)))$$

maximum a posteriori, penalized likelihood

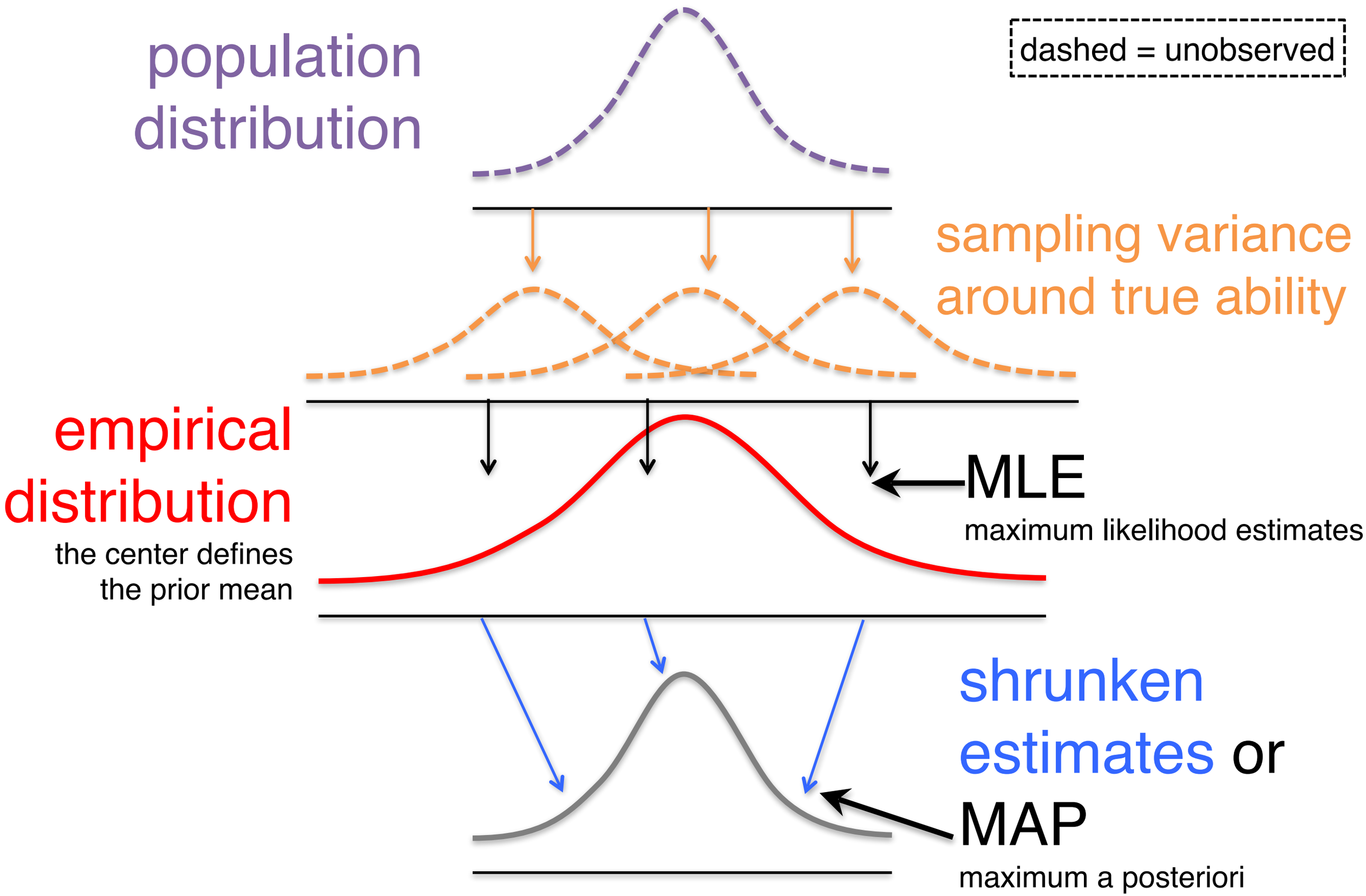


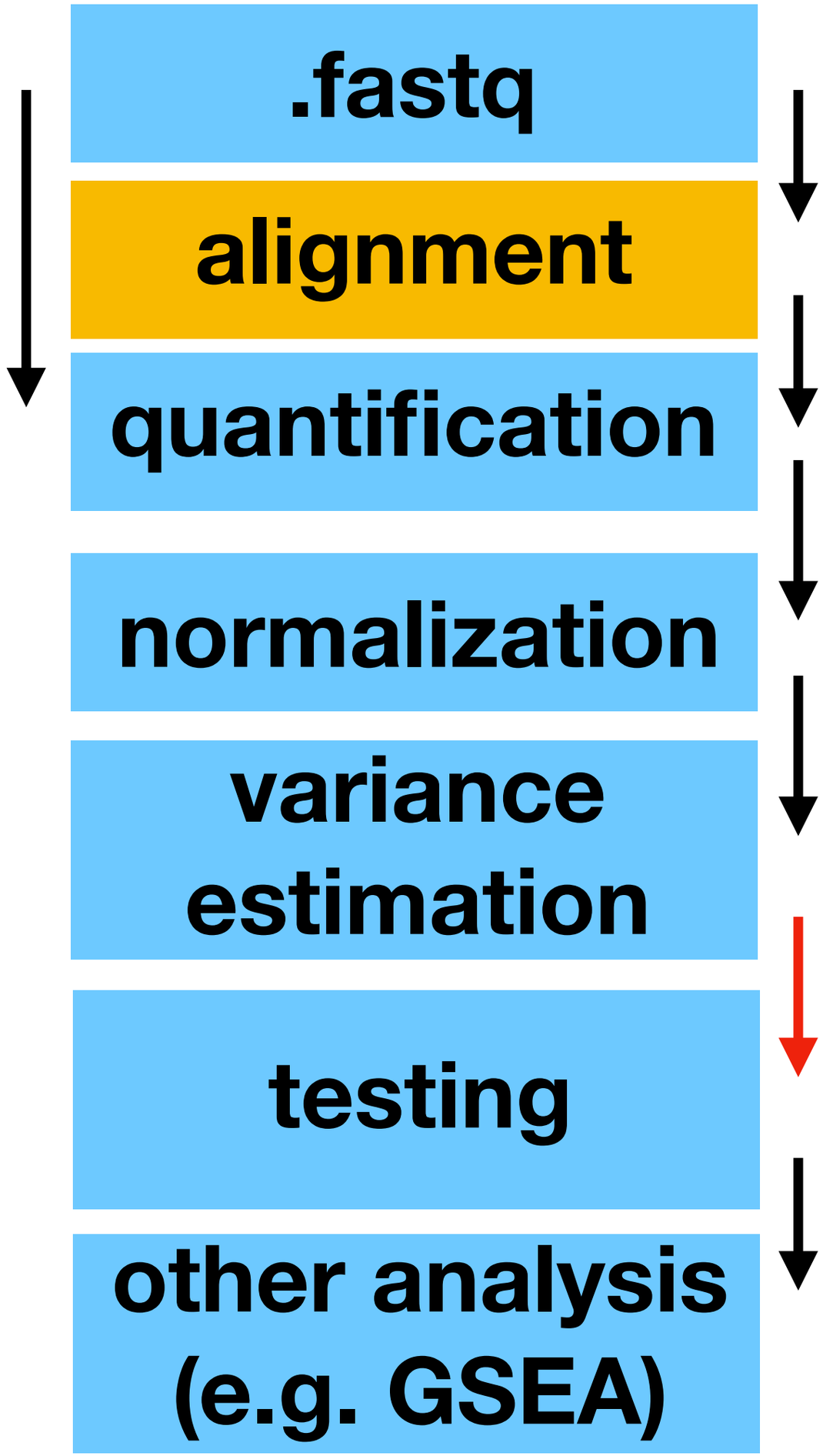
Why shrinkage?



"shrunken" estimates
less error *overall*
than **individual** estimates

Shrinkage estimation




.fastq
alignment
quantification
normalization
variance estimation
testing
other analysis (e.g. GSEA)

Generalized linear models

$$\log \mu_{ij} = \sum_{k} \beta_{ik} x_{kj}$$

response
(expected counts)

predictors
(or differential expression effects)

design matrix

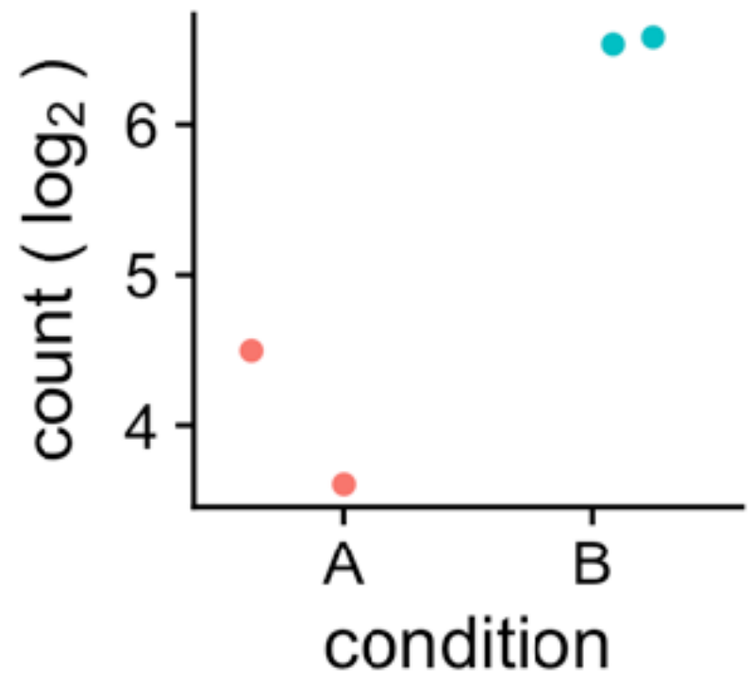
Simplest case: two-group comparison

DESeq2 design = ~ condition

condition

A
A
B
B

$$\begin{matrix} \log_2 u_1 \\ \log_2 u_2 \\ \log_2 u_3 \\ \log_2 u_4 \end{matrix} = \begin{array}{c} x_1 \\ \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \end{array} \begin{pmatrix} \beta_{\text{Intercept}} \\ \beta_{\text{condition\_A\_vs\_B}} \end{pmatrix}$$



estimated log expression

condition A = $\beta_{\text{Intercept}}$

condition B = $\beta_{\text{Intercept}}$ **+** $\beta_{\text{condition\_A\_vs\_B}}$

Multi-level comparisons

DESeq2 design = ~ condition

$$
\begin{array}{cc}
\textbf{condition} & \\
\textbf{A} & \log_2 u_1 \\
\textbf{A} & \log_2 u_2 \\
\textbf{B} & \log_2 u_3 \\
\textbf{B} & \log_2 u_4 \\
\textbf{C} & \log_2 u_5 \\
\textbf{C} & \log_2 u_6
\end{array}
=
\begin{array}{c}
\begin{array}{ccc} & x_1 & x_2 \end{array} \\
\begin{pmatrix}
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 1 & 0 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & 0 & 1
\end{pmatrix}
\end{array}
\begin{pmatrix}
\beta_{\text{Intercept}} \\
\beta_{\text{condition\_A\_vs\_B}} \\
\beta_{\text{condition\_A\_vs\_C}}
\end{pmatrix}
$$

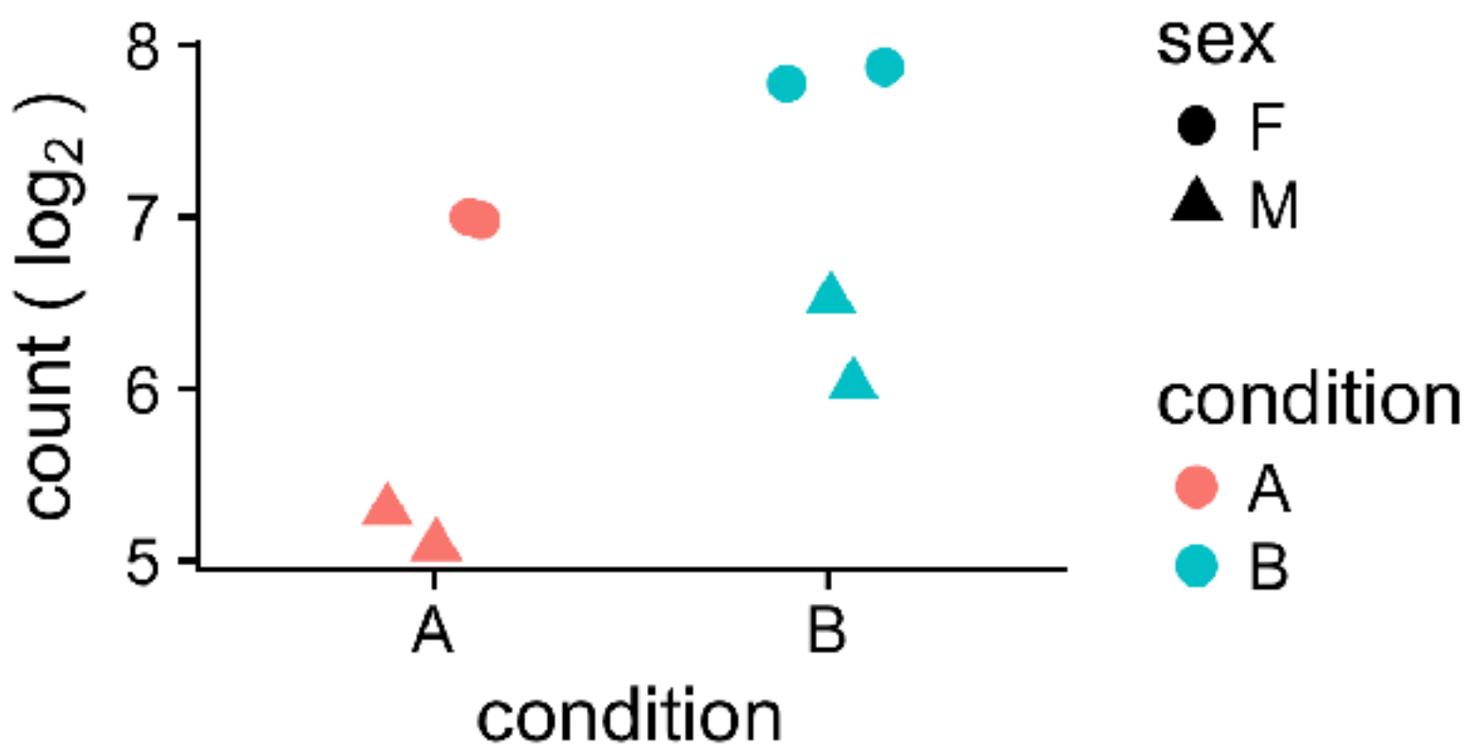
estimated log expression

condition A = $\beta_{\text{Intercept}}$

condition B = $\beta_{\text{Intercept}}$ **+** $\beta_{\text{condition\_A\_vs\_B}}$

condition C = $\beta_{\text{Intercept}}$ **+** $\beta_{\text{condition\_A\_vs\_C}}$

Comparisons with blocking factors



Comparisons with blocking factors

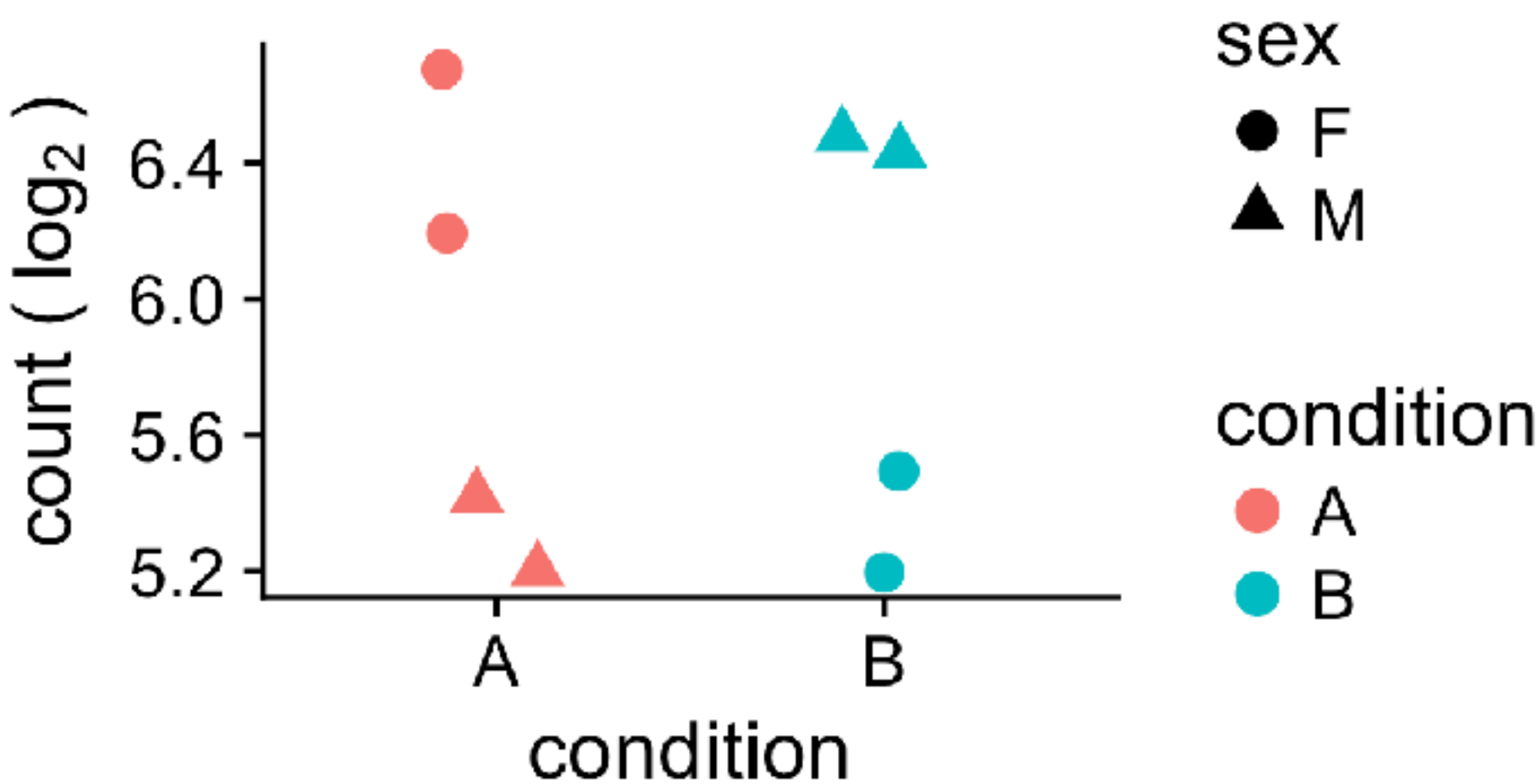
DESeq2 design = ~ sex + condition

sex	**condition**	
M	**A**	$\log_2 u_1$
M	**A**	$\log_2 u_2$
M	**B**	$\log_2 u_3$
M	**B**	$\log_2 u_4$
F	**A**	$\log_2 u_5$
F	**A**	$\log_2 u_6$
F	**B**	$\log_2 u_7$
F	**B**	$\log_2 u_8$

$$= \begin{array}{c} \begin{array}{ccc} & x_1 & x_2 \end{array} \\ \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \end{array} \begin{pmatrix} \beta_{\text{Intercept}} \\ \beta_{\text{sex\_female\_vs\_male}} \\ \beta_{\text{condition\_A\_vs\_B}} \end{pmatrix}$$

What would be the predicted log expression for females in condition B?

Interactions



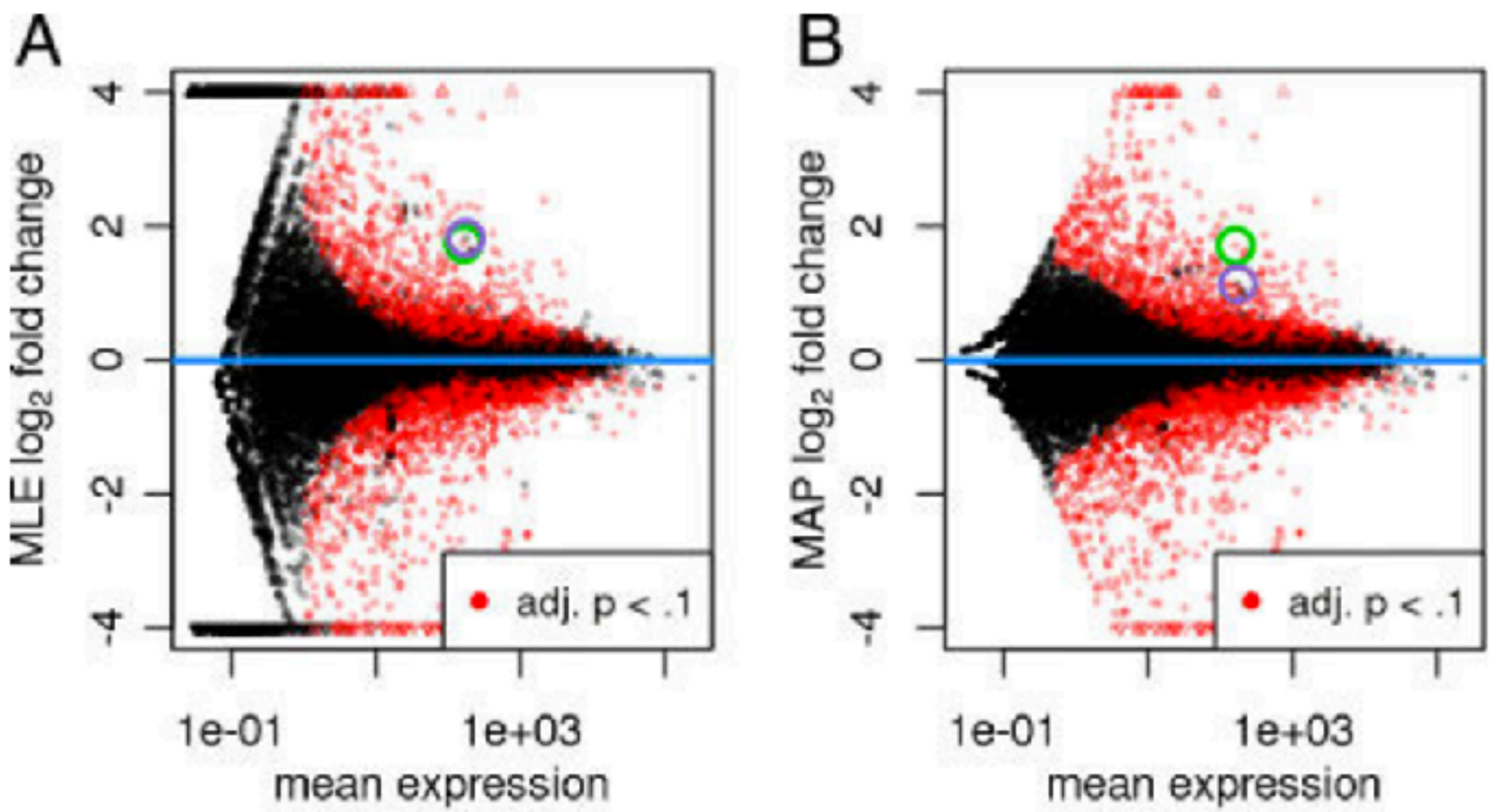
Comparisons with blocking factors

DESeq2 design = ~ sex + condition + sex:condition

sex	condition	
M	**A**	$\log_2 u_1$
M	**A**	$\log_2 u_2$
M	**B**	$\log_2 u_3$
M	**B**	$\log_2 u_4$
F	**A**	$\log_2 u_5$
F	**A**	$\log_2 u_6$
F	**B**	$\log_2 u_7$
F	**B**	$\log_2 u_8$

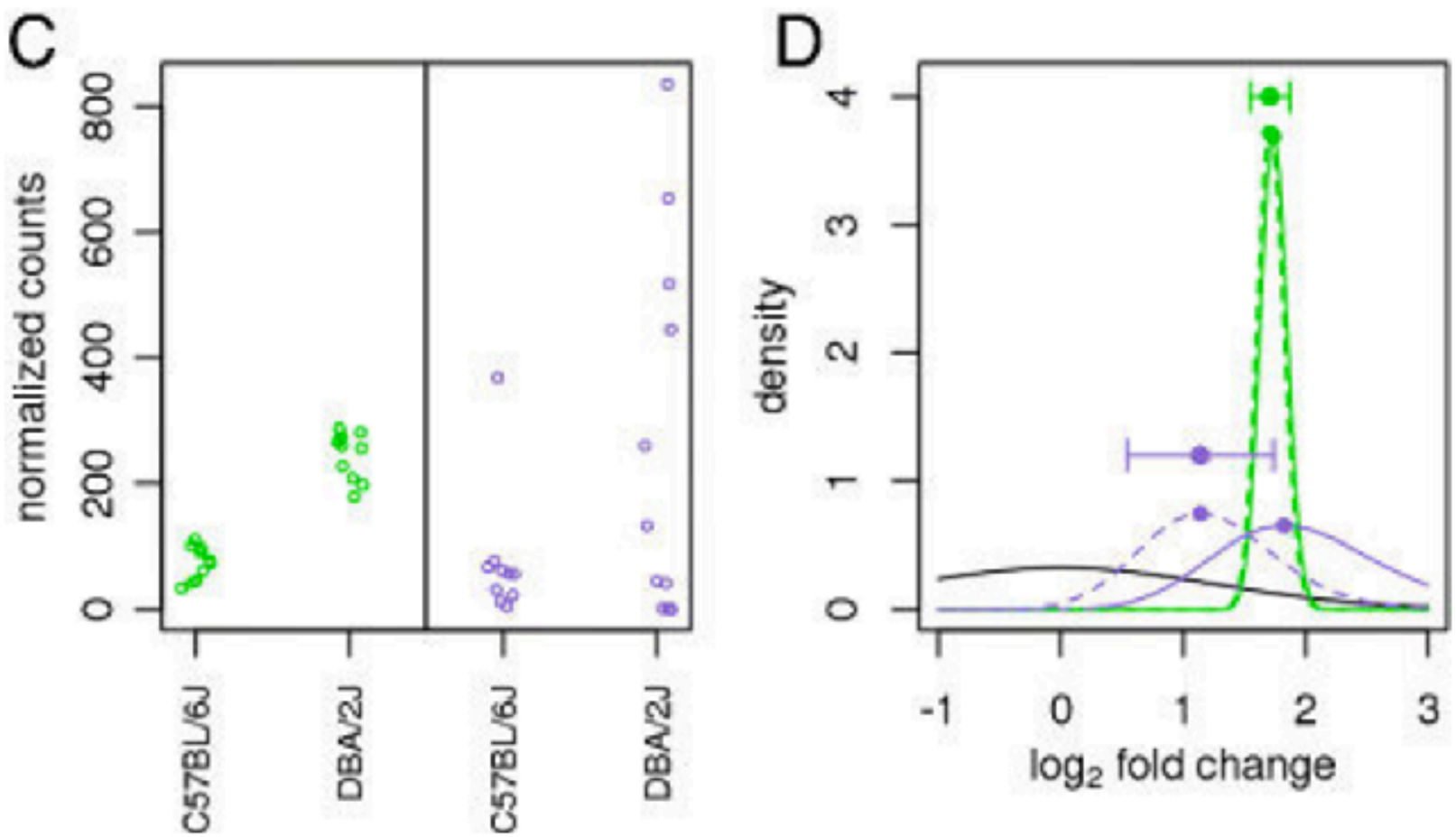
$$= \begin{array}{c} \begin{array}{cccc} & x_1 & x_2 & x_3 \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix} \end{array} \begin{pmatrix} \beta_{\text{Intercept}} \\ \beta_{\text{sex\_female\_vs\_male}} \\ \beta_{\text{condition\_A\_vs\_B}} \\ \beta_{\text{interaction}} \end{pmatrix}$$

DESeq2 shrinkage of log fold changes



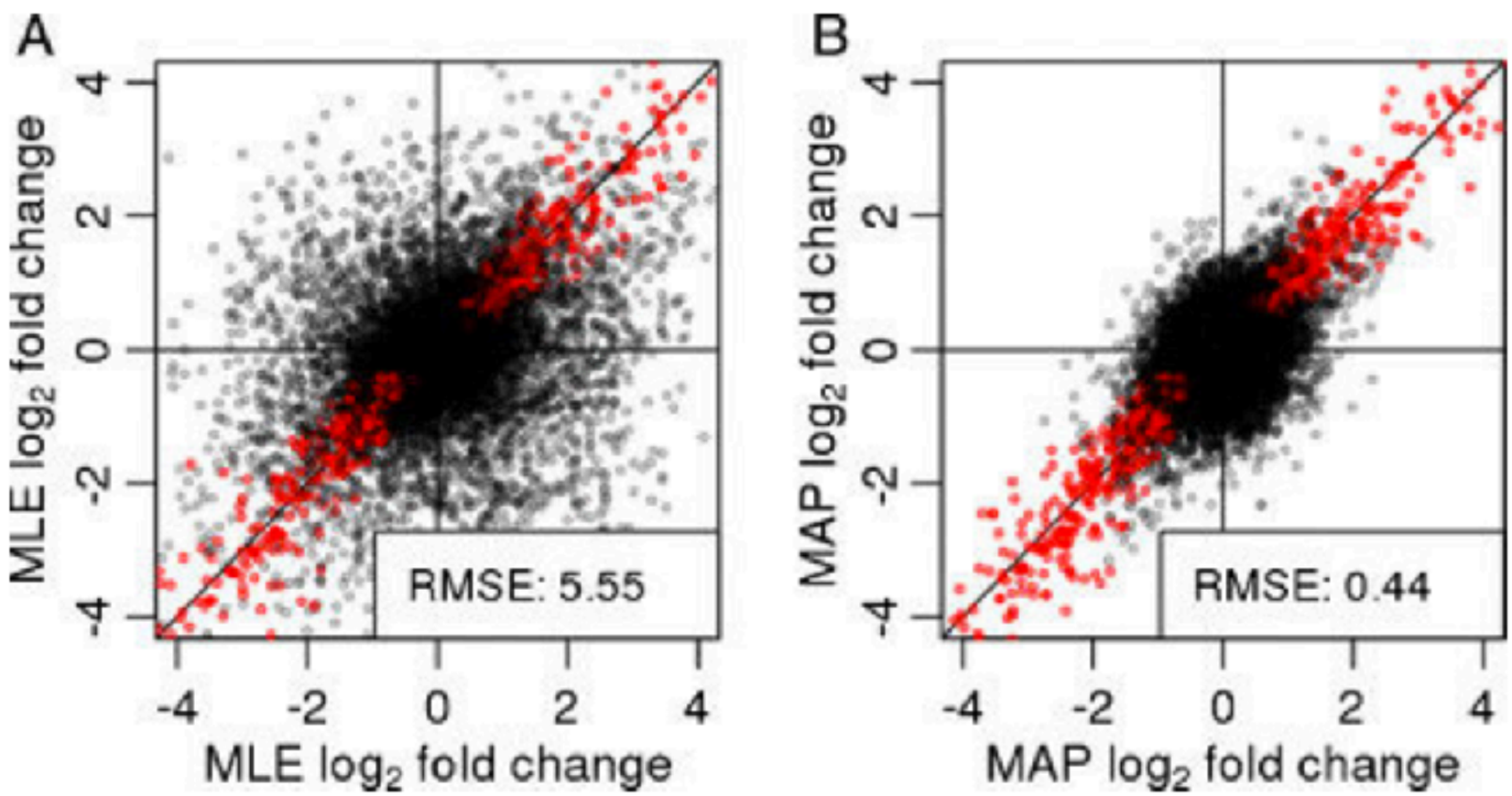
Love et al, 2014

DESeq2 shrinkage of log fold changes



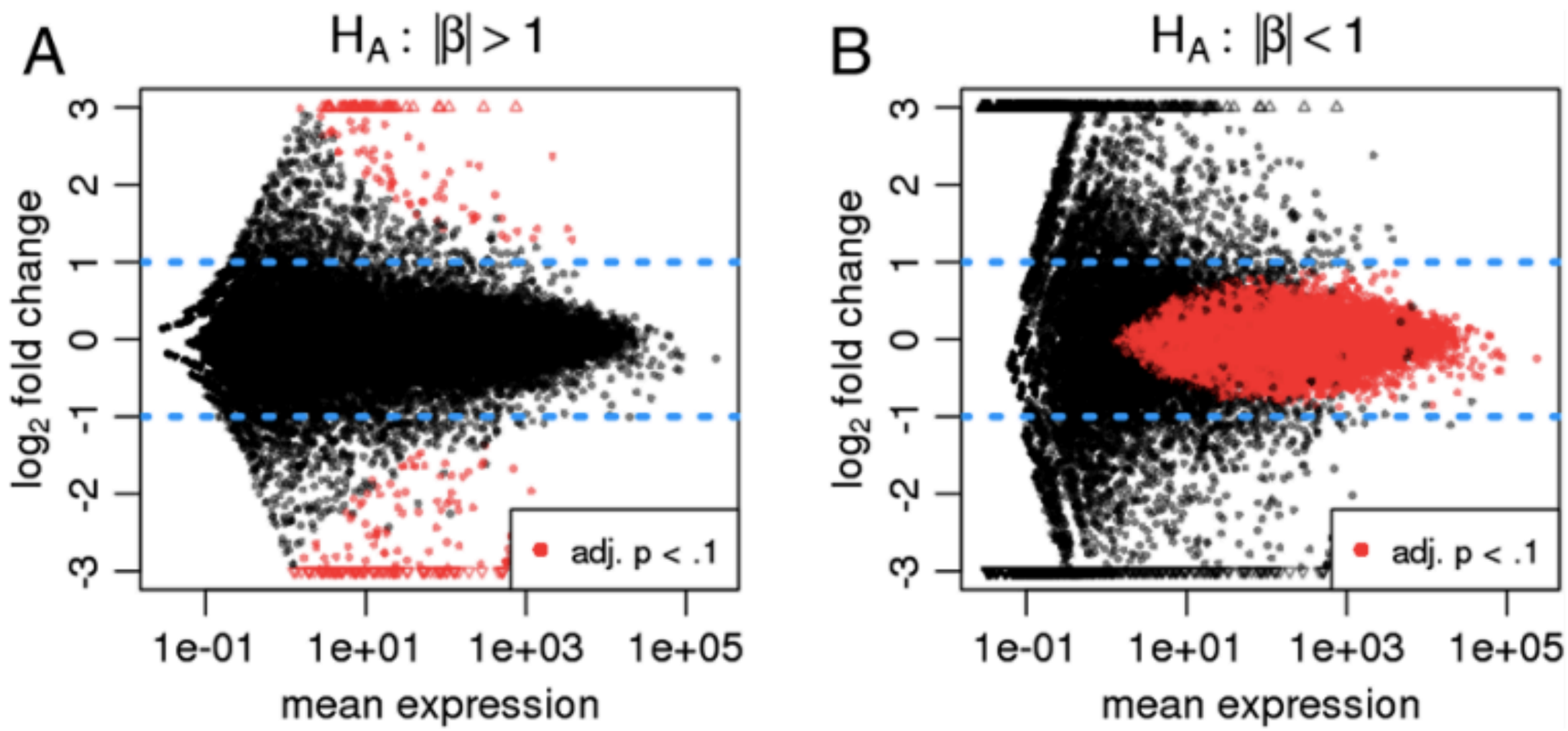
Love et al, 2014

DESeq2 shrinkage of log fold changes



Love et al, 2014

Other hypothesis testing options



Love et al, 2014

Confounders and experimental design

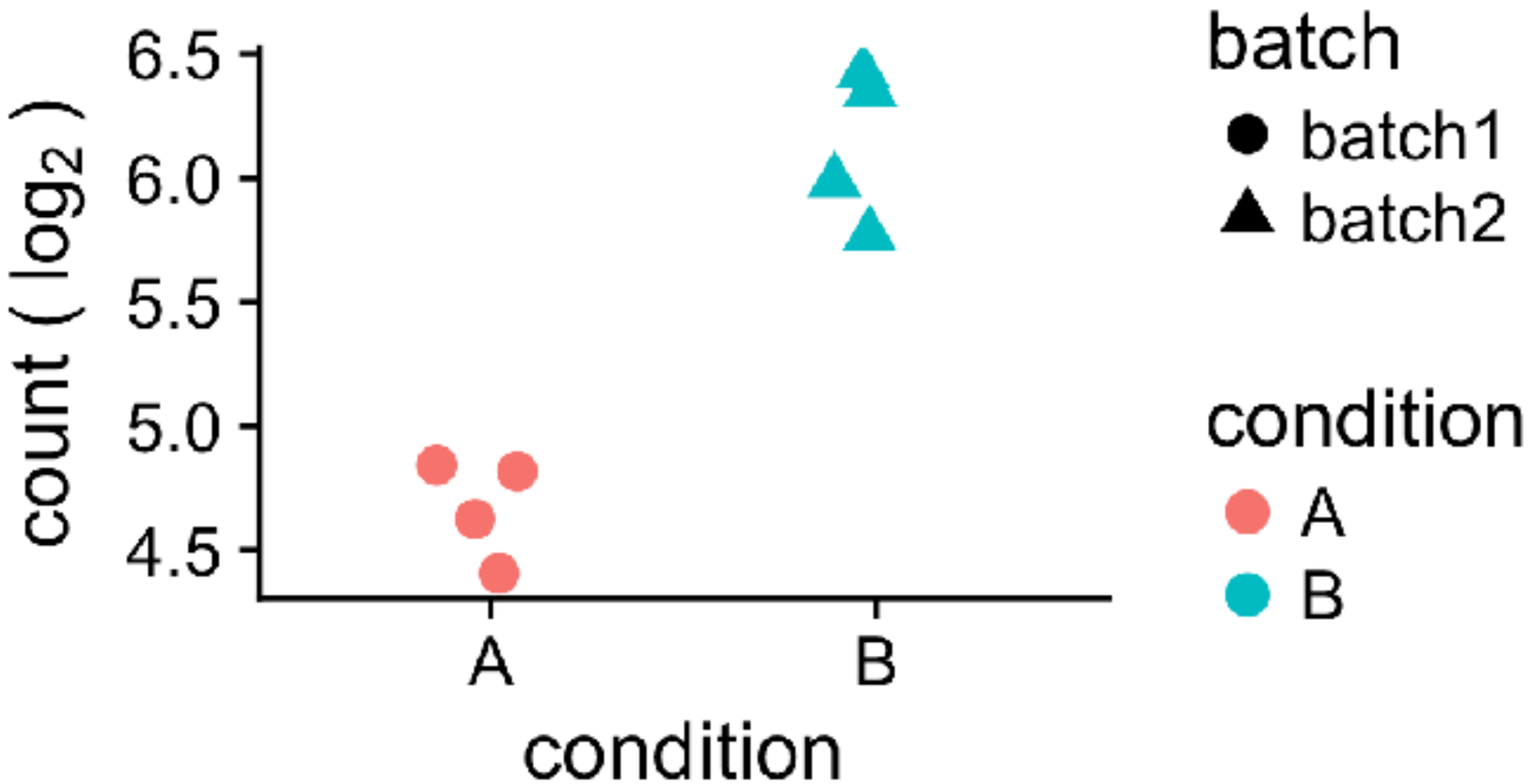
- Spielman et al., Nature Genetics 2007

  78% of genes are differentially expressed between human populations

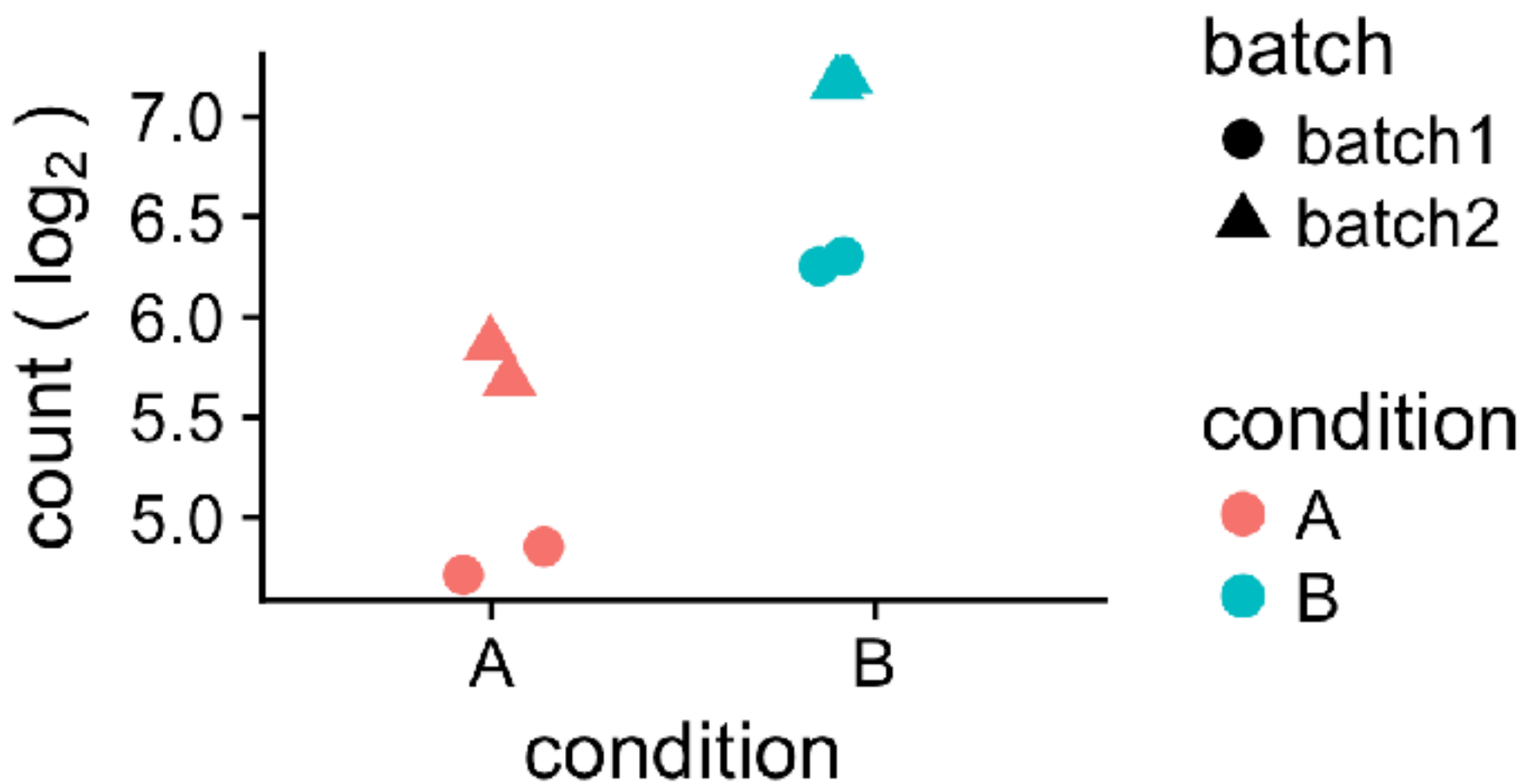
- Lin et al., PNAS, 2014

  Differences in gene expression across species are larger than between tissues of a same species.

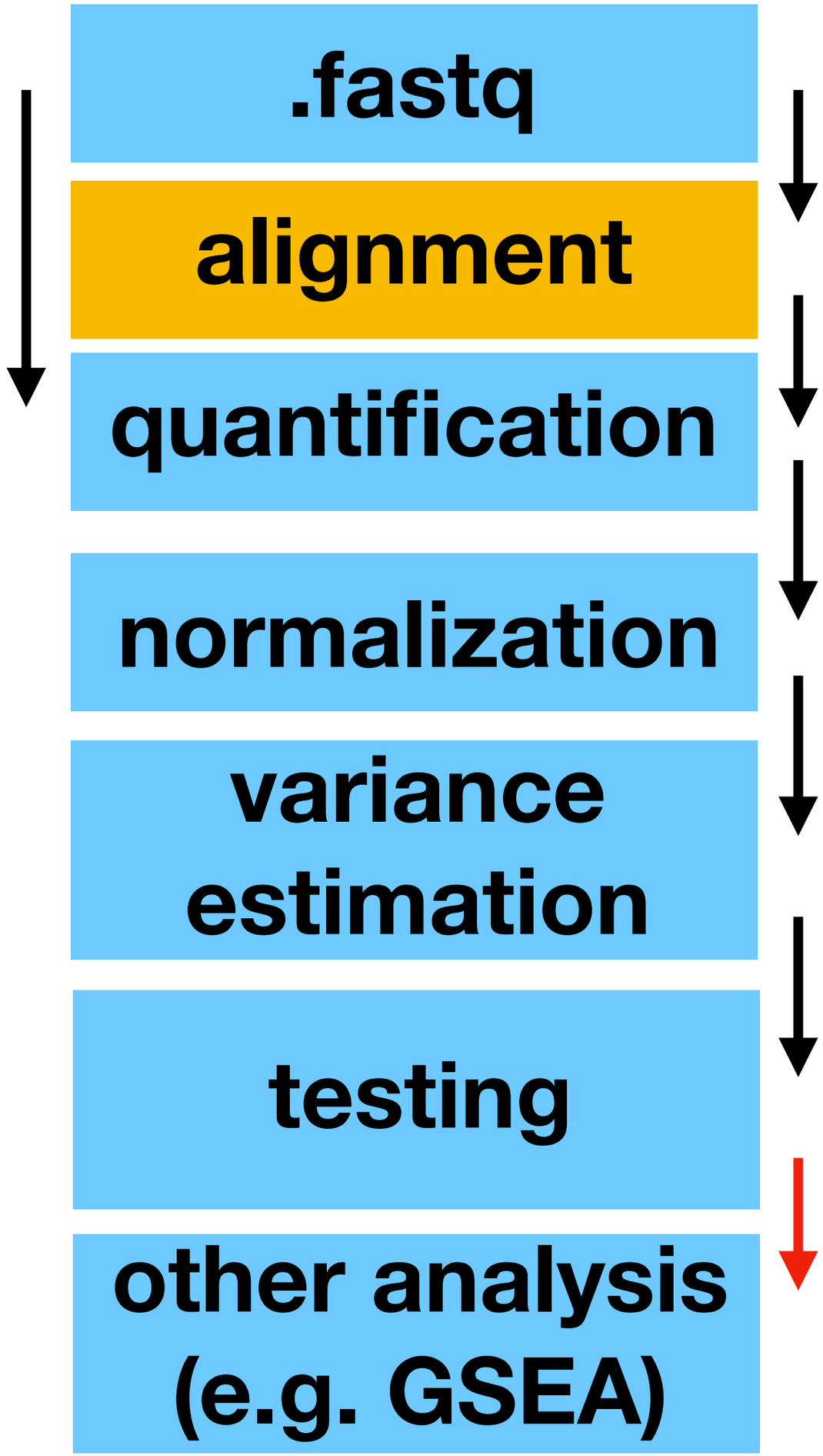
Confounders experiment



Confounders and experimental design



- Randomize conditions of interest in batches and include it as a blocking factor.


.fastq
alignment
quantification
normalization
variance estimation
testing
other analysis (e.g. GSEA)

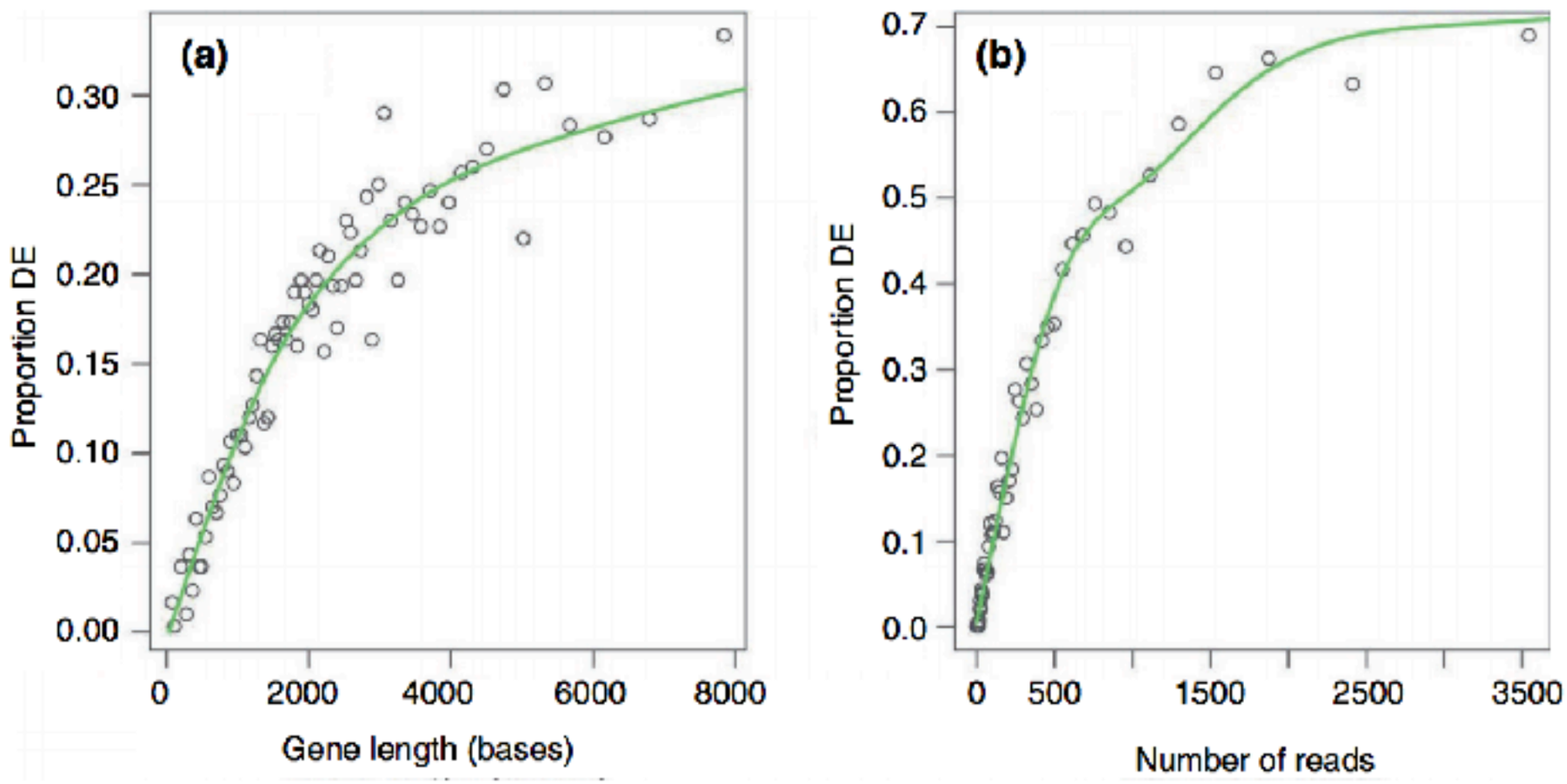
“Traditional” gene-set enrichment analysis

	In gene set	Not in gene set	
DGE	a	b	a + b
Background	c	d	c + d
	a + c	**b + d**	

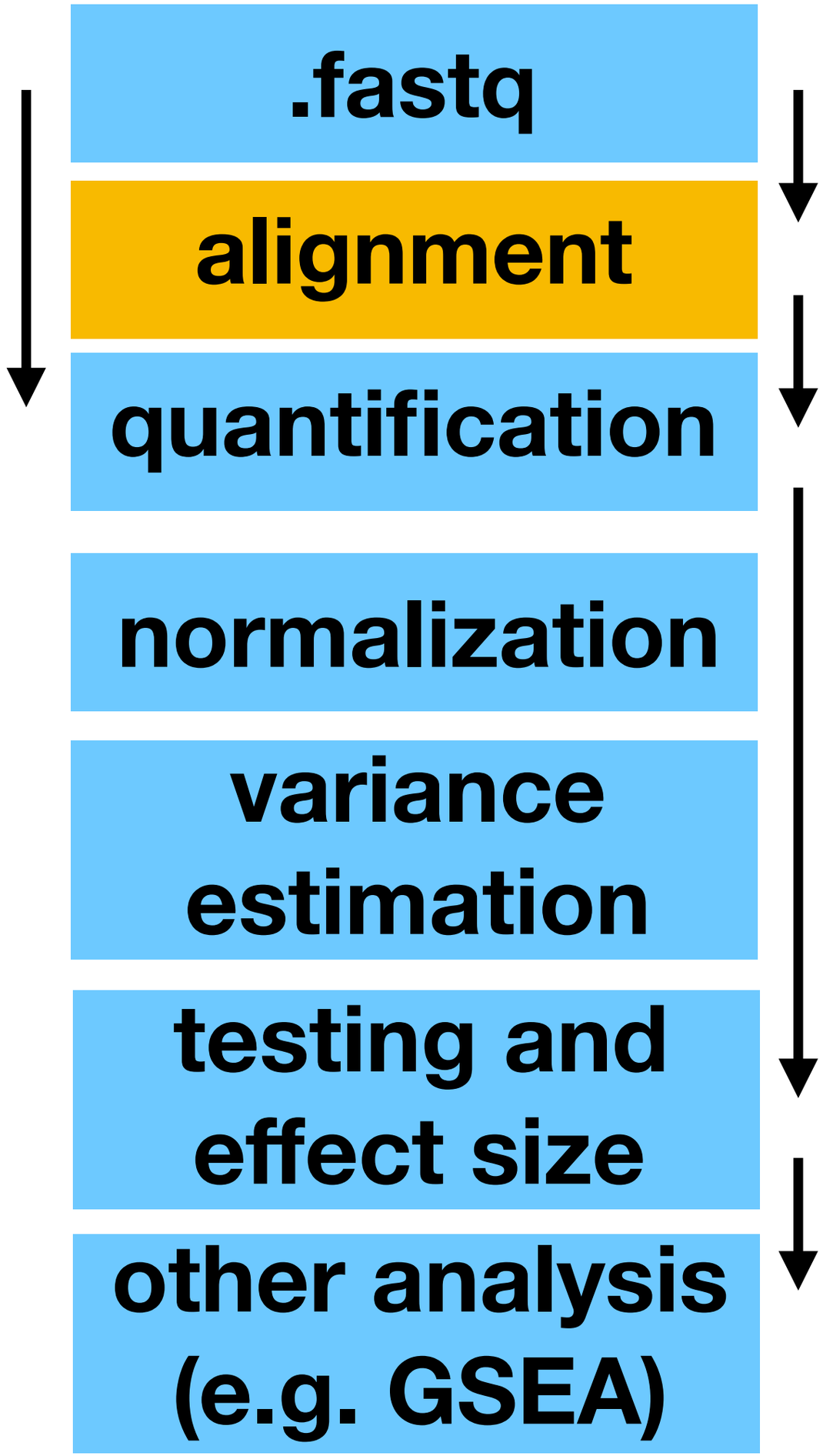
$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\;\;b!\;\;c!\;\;d!\;\;n!}$$

Fisher’s test

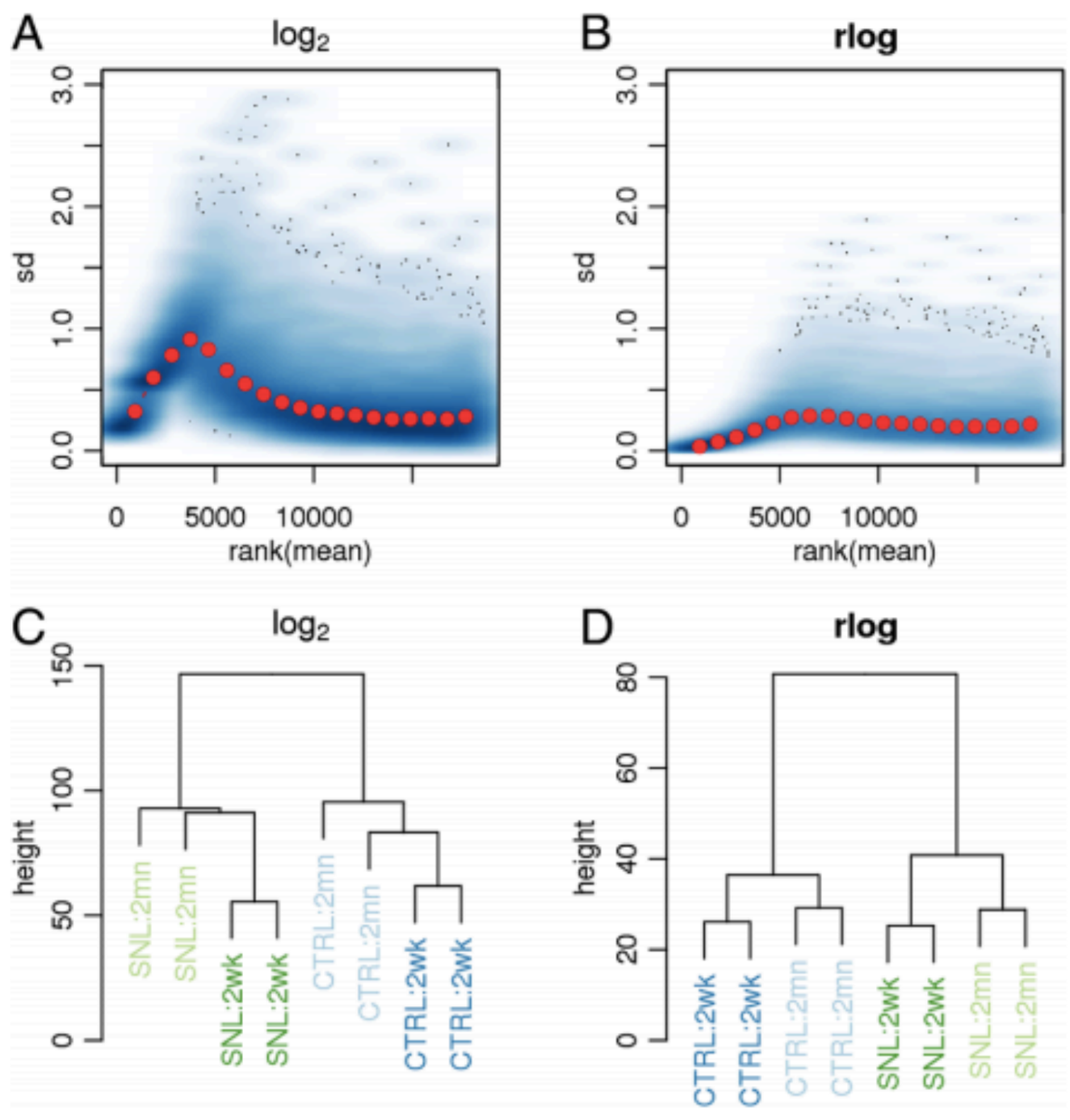
goseq estimates and corrects for RNA-seq biases in GSEA



Young et al, 2010

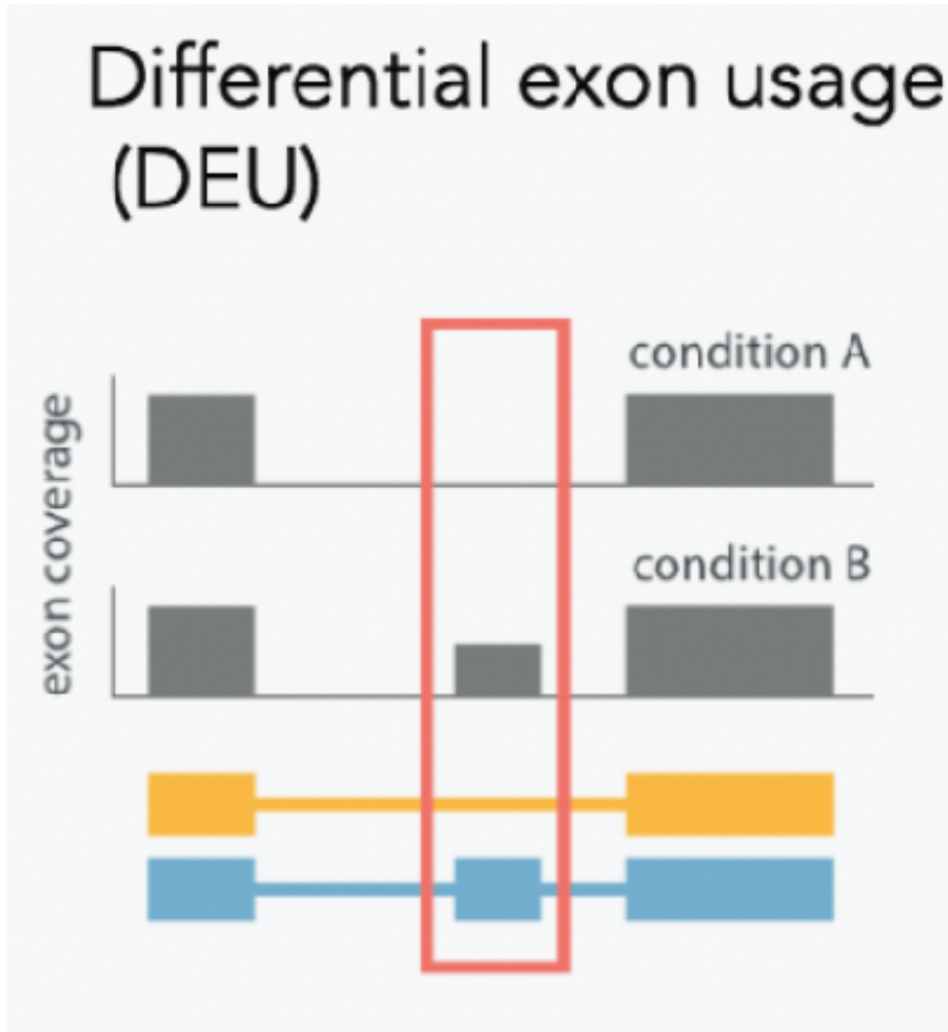

.fastq
alignment
quantification
normalization
variance estimation
testing and effect size
other analysis (e.g. GSEA)

For exploratory analysis, clustering, PCA: Use rlog or vst data!



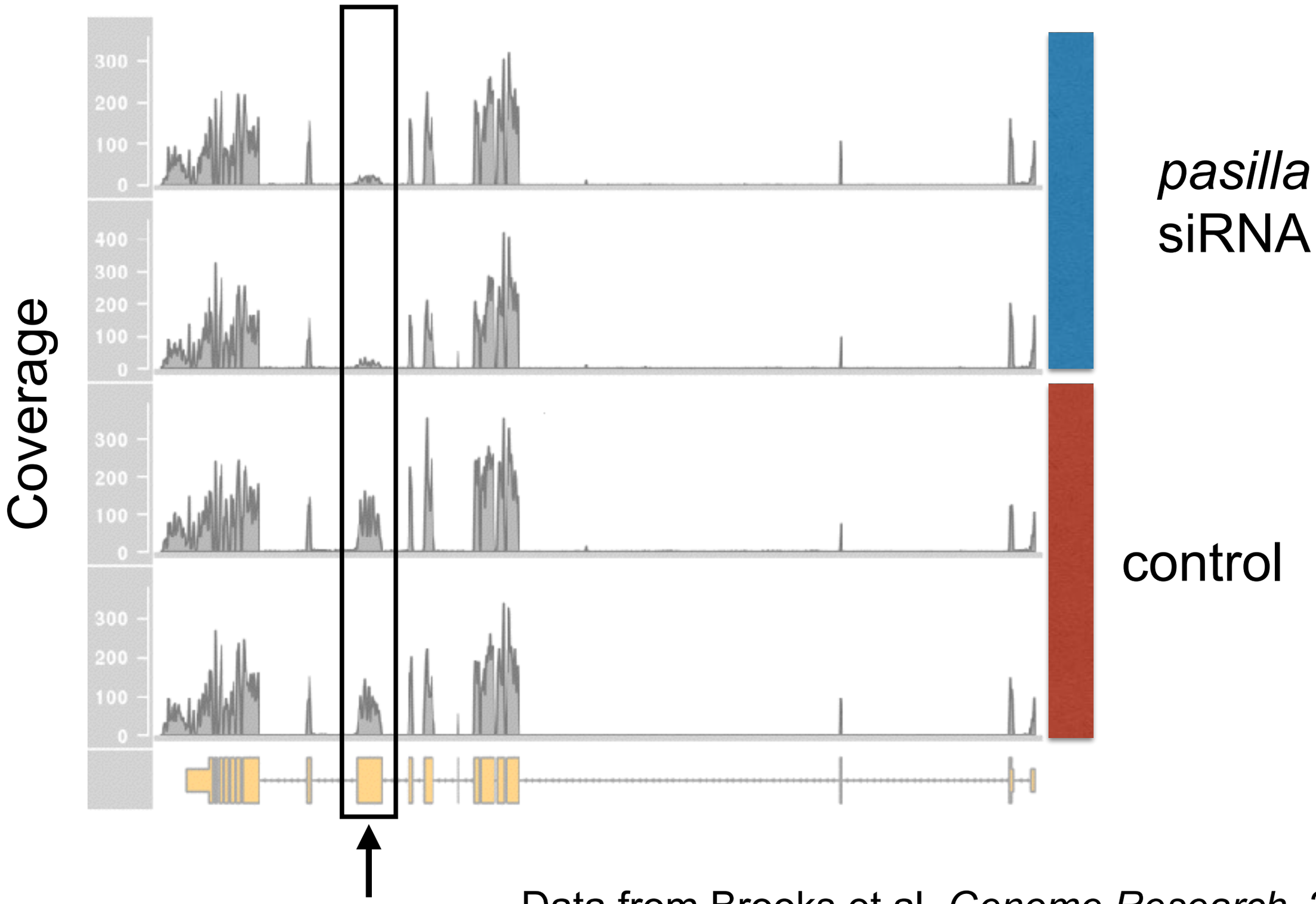
Love et al, 2014

Differential exon usage



Modified slide from Mark Robinson

High-throughput RNA sequencing enables an unbiased characterisation of isoform expression



Data from Brooks et al. *Genome Research*, 2011

DEXSeq: inference of differential exon usage

samples →

exons ↓

	treated1fb	treated2fb	treated3fb	untreated1fb	untreated2fb	untreated3fb	untreated4fb
E001	1997	494	562	1150	2514	570	547
E002	122	112	180	69	203	156	142
E003	276	293	305	190	398	312	259
E004	420	200	182	230	446	183	185
E005	416	217	279	146	170	237	231
E006	486	357	471	190	337	418	364
E007	574	465	536	469	805	480	496
E008	536	417	447	541	832	475	472
E009	191	237	216	217	427	286	222
E010	188	130	96	617	1177	520	508
E011	165	212	210	118	275	294	269
E012	536	437	414	441	792	619	504
E013	72	41	49	40	76	34	38
E014	3	0	33	5	0	2	42

$$\text{exon usage} = \frac{\text{\# of transcripts including an exon}}{\text{\# total transcripts}}$$

Anders et al, 2012

DEXSeq: inference of differential exon usage

exon *i*, sample *j*
I = 1 exon under consideration
I = 0 sum of gene count

$$\log \mu_{ijl} = \beta_{ij}^{S} + I\beta_{i}^{E} + \beta_{i\rho_j}^{EC}$$

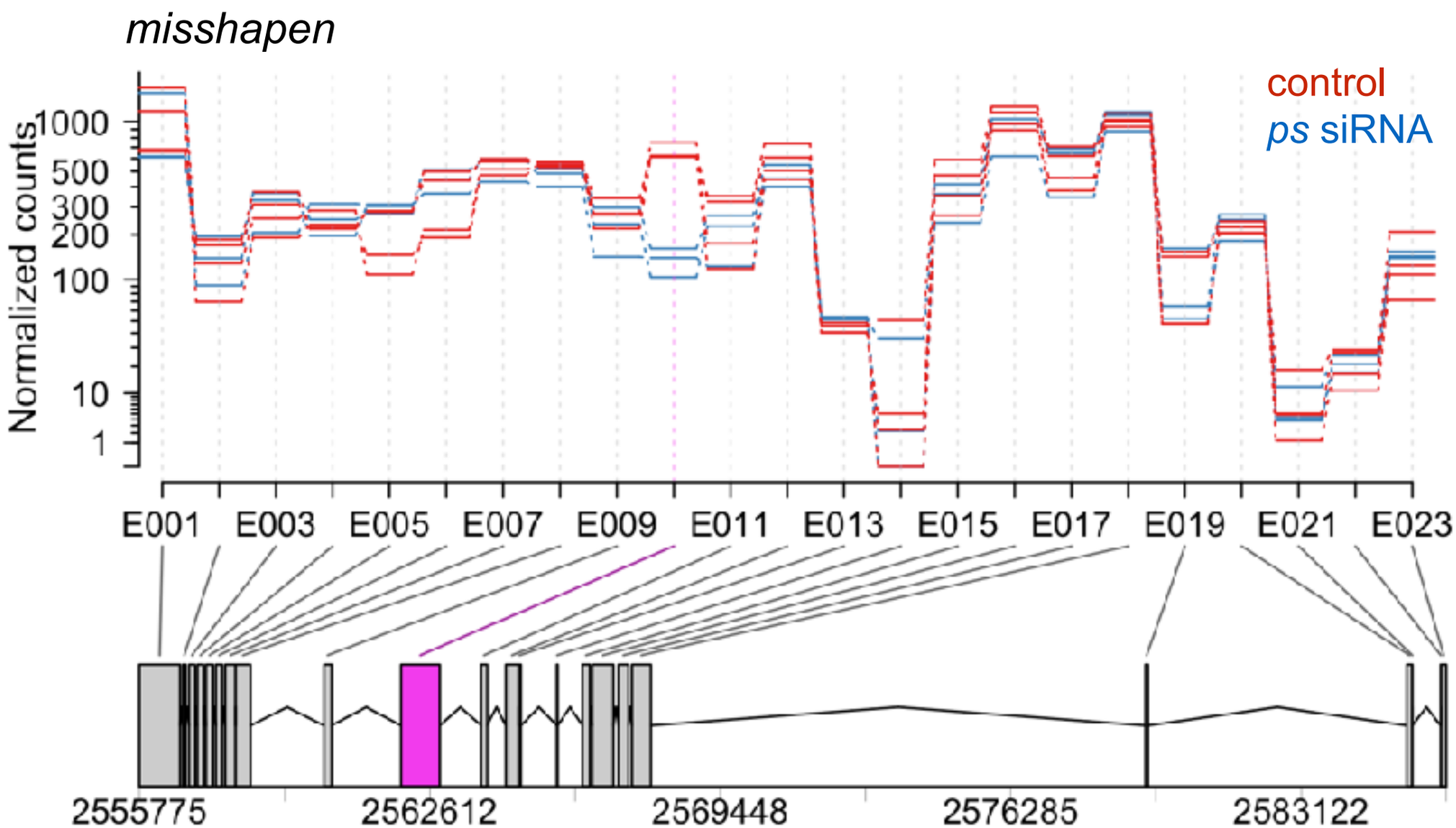
sample specific contributions (gene expression)

average exon usage

changes in exon usage due to the conditions

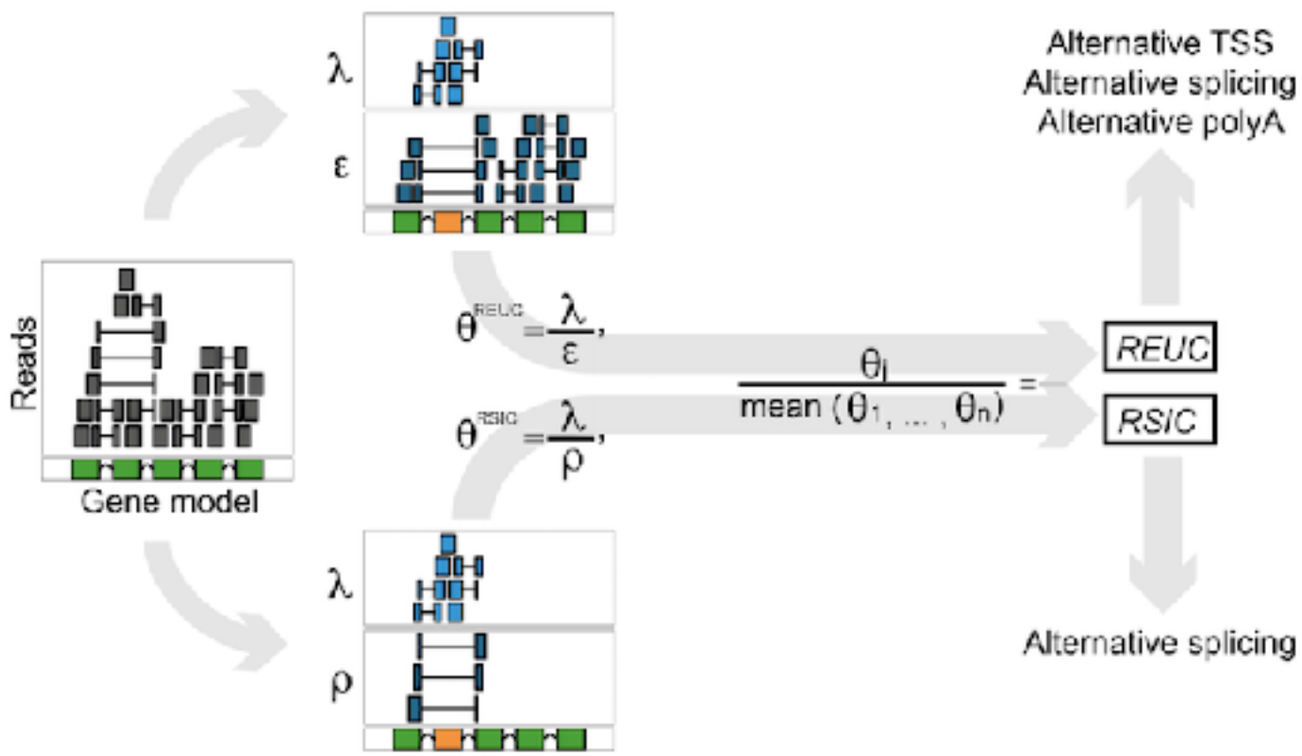
Anders et al, 2012

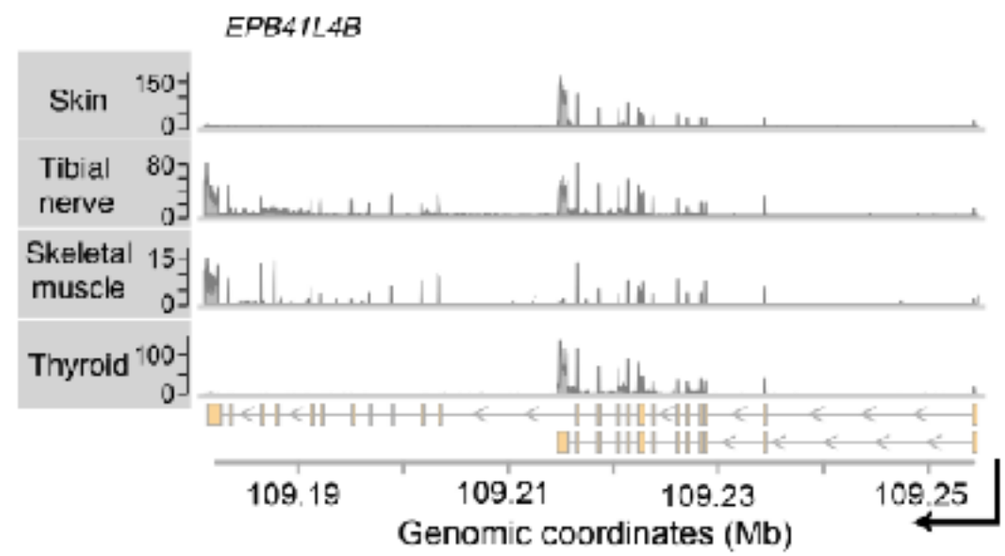
DEXSeq: inference of differential exon usage

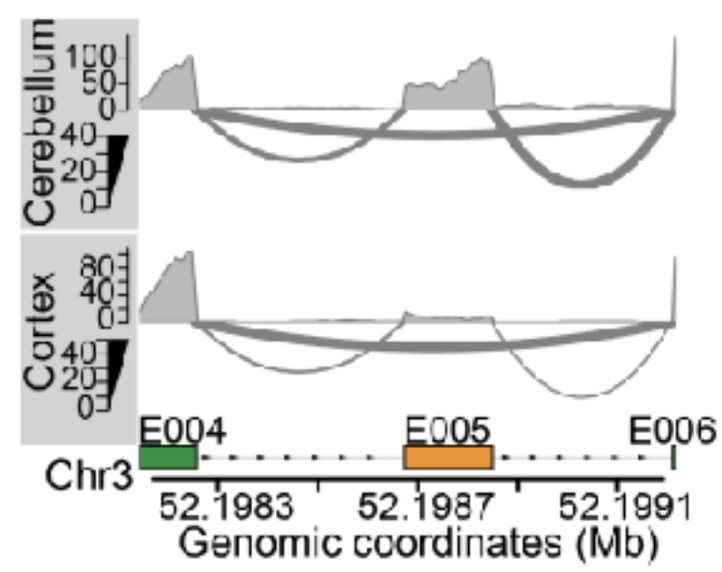


Anders et al, 2012

Detecting differential splicing vs differential usage of TSS and polyA







Reyes et al, 2017