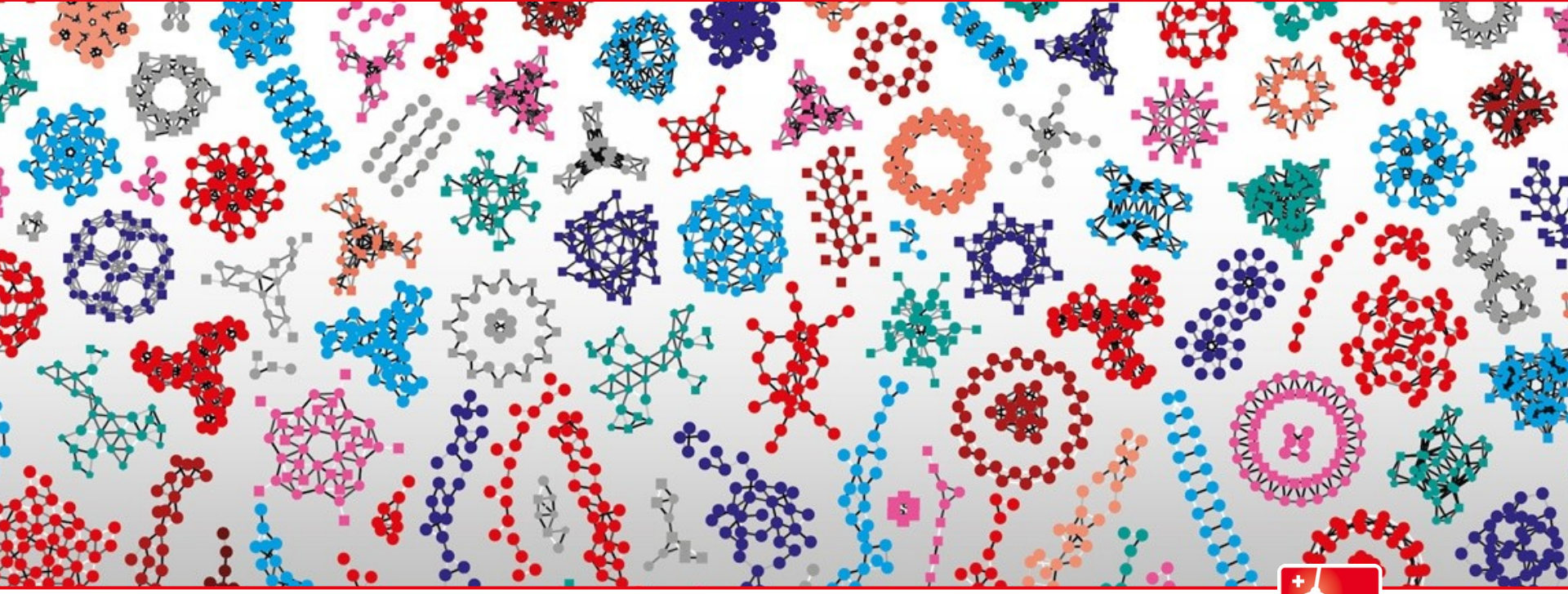


Swiss Institute of Bioinformatics

Introduction to RNA-Seq – Read Counting

Wandrille Duchemin

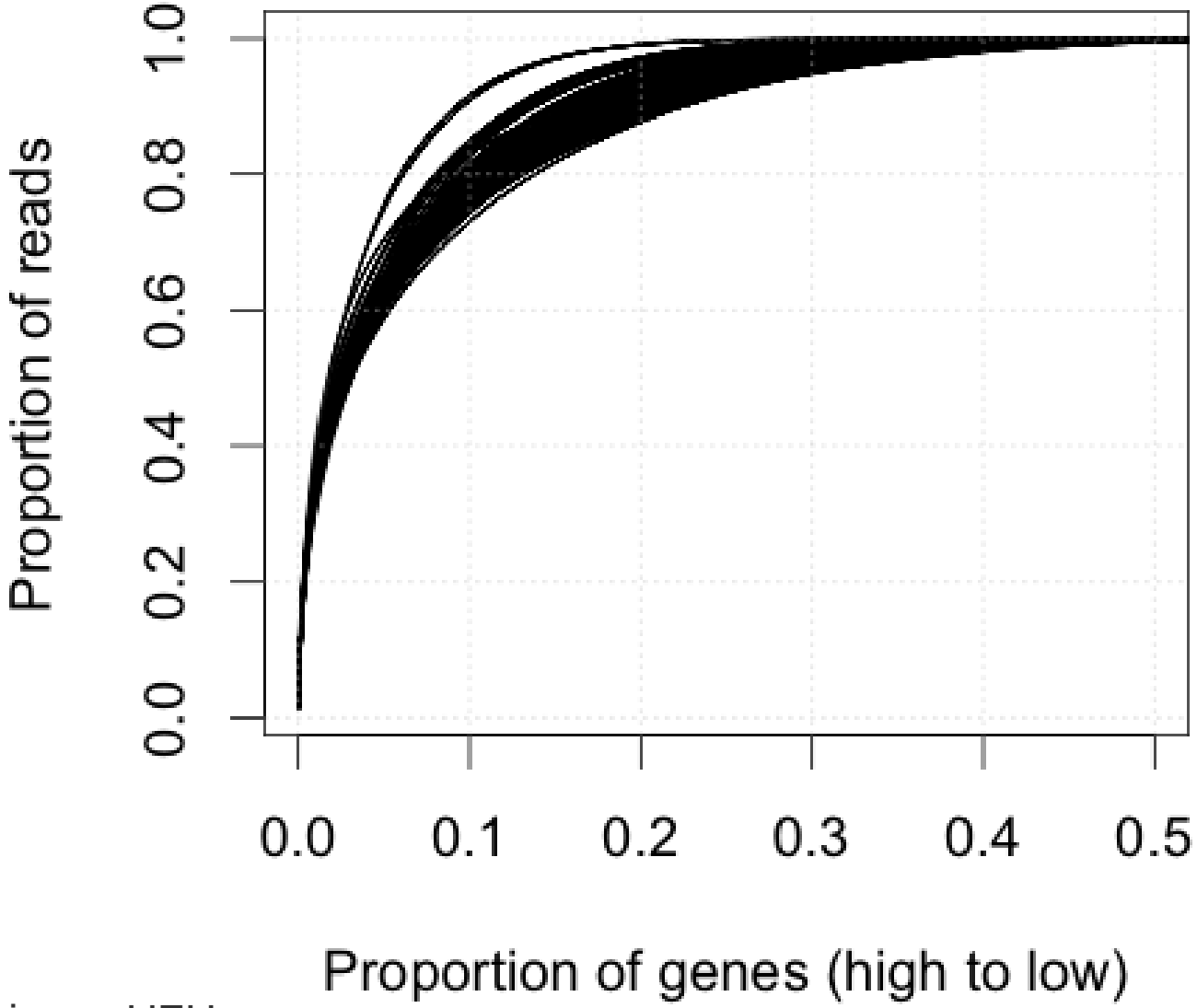
Is RNA-Seq expression inference reliable?

- **It's been known for many years that most Illumina-type RNA-Seq workflows are highly concordant with estimates from quantitative PCR methods**
- **Griffith *et al* (2010) Nature Methods**
  - Validation rates of ~85% for junction discovery and 88% for expression validation
- **Everaert *et al* (2017) Scientific Reports**
  - ~85% concordance between RNA-Seq and RT-qPCR
  - reproducibly inconsistent genes are typically small, with fewer exons, and lower overall expression

Read Counting – Initial Considerations

- **RNA-Seq comprises many technologies which are rapidly evolving**
- **The appropriate choice of methods highly depends on the question(s) you're asking**
  - Parameter space is important!
- **Proper gene/transcript model annotations are crucial**

How much sequencing goes to highly expressed genes?



Pr. Mark Robinson UZH

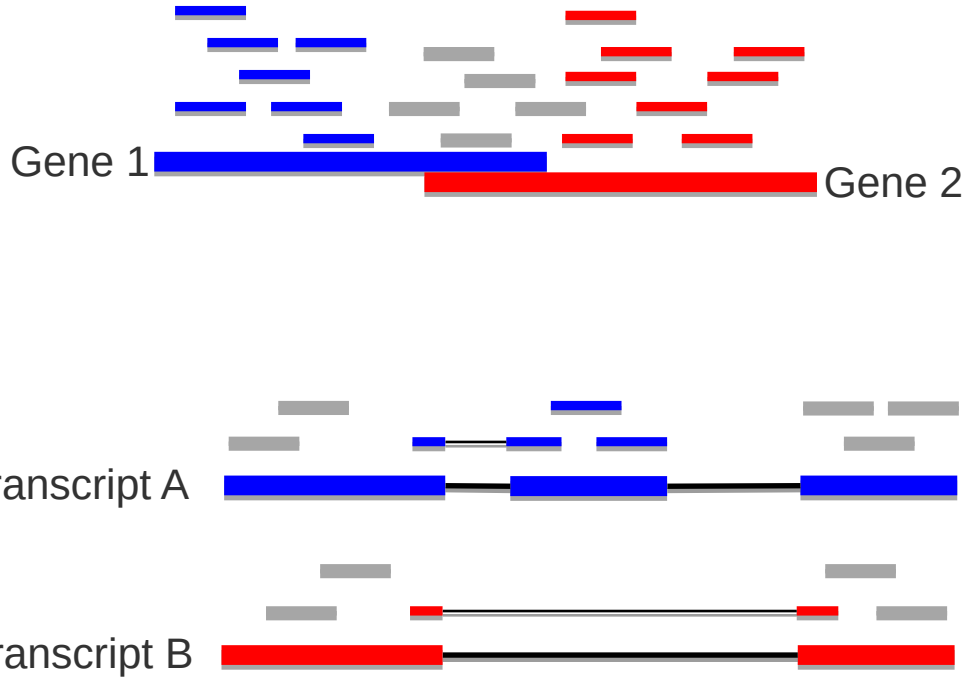
Basics of RNA-Seq Quantification

- **Remember - stochastic models underlie all methods for relative transcript abundance estimates**
- **First align reads against reference**
- **Count number of reads aligning to features**
  - "fragment assignment"
  - decide how to treat multi-mapping reads
- **Convert read counts to *relative abundance***
  - *"density deconvolution"*
- **Account for differences in:**
  - library size
  - feature lengths
  - sequence-based biases

Read Counting – Fundamental Problems

- **Aligners map reads to genomic coordinates and/or to all features associated to the mapped coordinates**
  - How to treat multi-mapping reads?
  - *eg* gene families, repetitive sequences, alternative splice forms

Read Counting – Fundamental Problems

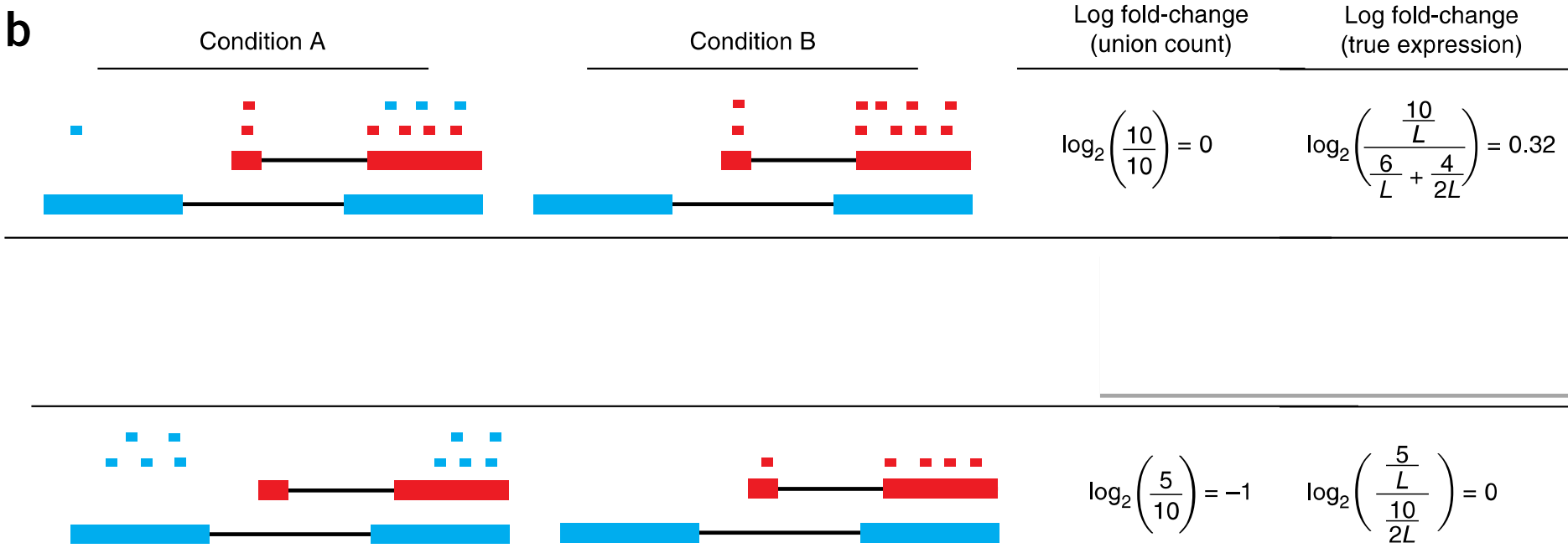


Solutions to multi-mapping reads

- **Discard all multi-reads, estimate abundance based on uniquely mapping reads only**
  - Loss of information
  - Potentially biased abundance estimates
  - Appropriate for edgeR/DESeq2, expected that samples being compared have same distribution of multi-reads
- **“Rescue” multireads by fractional allocation**
  - Estimate abundances based on uniquely mapping reads
  - Divide multireads between features based on abundance estimates from uniquely mapped reads
  - Recompute abundances based on updated counts
  - Used by tools like Cufflinks

Counting/Quantification

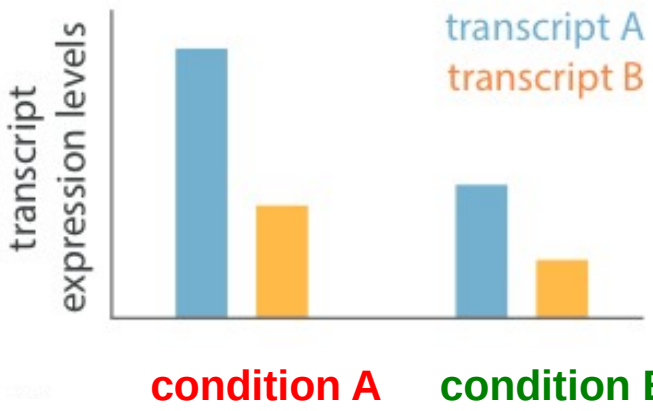
union counters -> simple sum of all reads
transcript counters -> sum of length-normalized reads
(often unknown which reads map to which transcript)



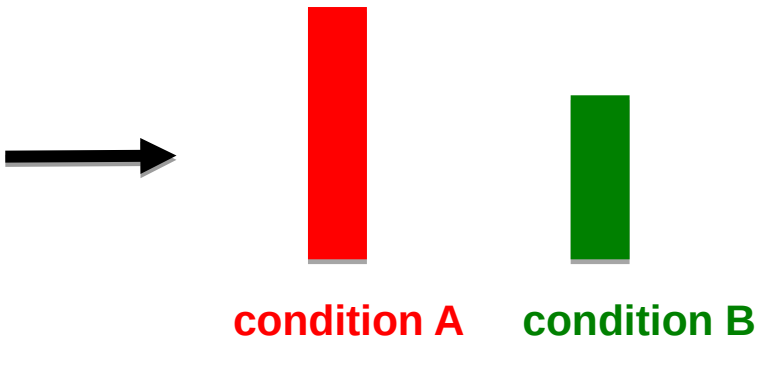
adapted from Trapnell *et al* 2013 Nat Biotech

Define the differential problem

Differential transcript expression (DTE)



Differential gene expression (DGE)



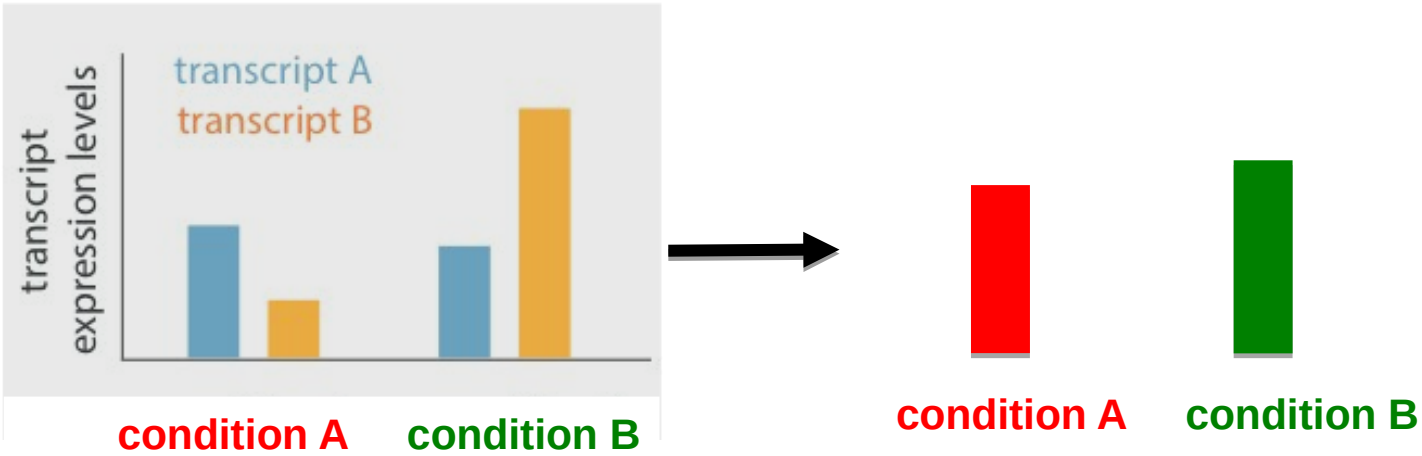
Pr. Mark Robinson UZH

Slide adapted from Mar Gonzàlez-Porta's talk at ECCB 2014
http://radiant-project.eu/ECCB/gonzalez-porta-140907065638-phpapp01.pdf

Define the differential problem

Differential transcript usage (DTU)

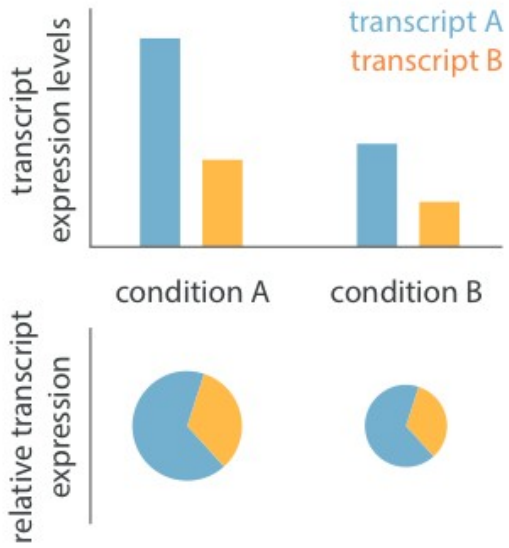
Differential gene expression (DGE)



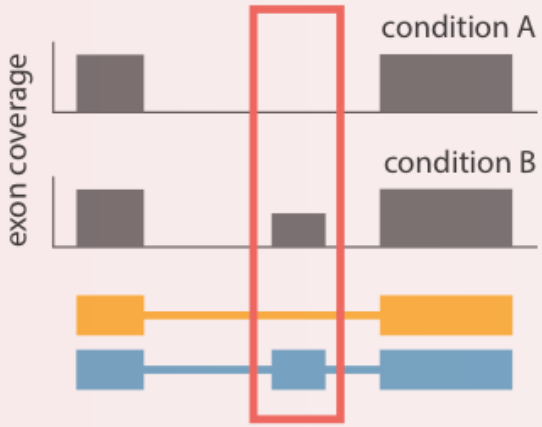
Slide adapted from Mar Gonzàlez-Porta's talk at ECCB 2014
http://radiant-project.eu/ECCB/gonzalez-porta-140907065638-phpapp01.p

Define the differential problem

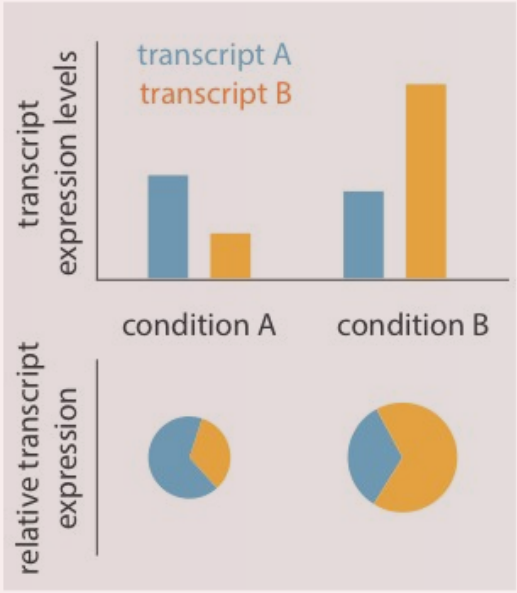
Differential transcript expression (DTE)



Differential exon usage (DEU)



Differential transcript usage (DTU)

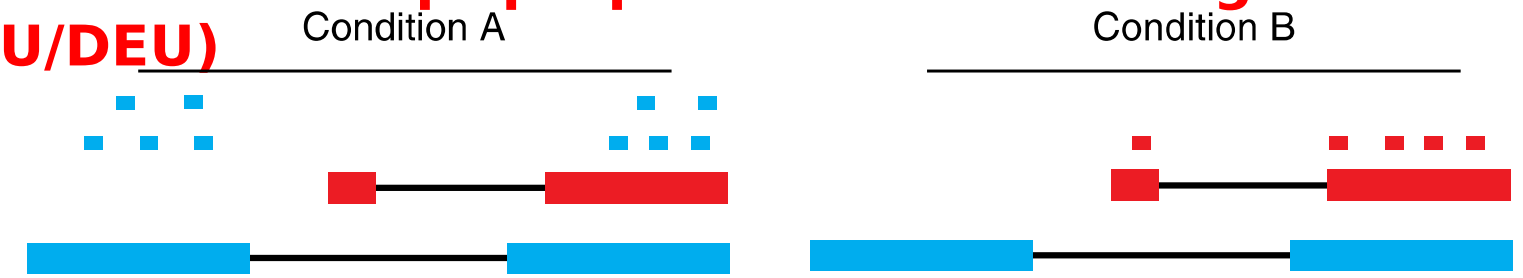


See also Soneson, Matthes et al., 2016, Genome Biology (comparison of DTU methods)

Slide adapted from Mar Gonzàlez-Porta's talk at ECCB 2014
http://radiant-project.eu/ECCB/gonzalez-porta-140907065638-phpapp01.pdf

What do you want to know?

- **whether individual transcripts have changed? (DTE)**
- **whether *any* transcripts in gene have changed? (DTE->G)**
- **whether the overall output has changed? (DGE)**
- **whether transcript proportions have changed? (DTU/DEU)**



Blue/red transcript changed? Yes, Yes

Any transcripts changed? Yes

Overall expression change? No

Transcript proportions changed? Yes

Transcript-Level Counting

- **More informative to understand regulation of alternative transcript usage**
- **Enables novel transcript discovery**
- **Primary drawbacks:**
  - requires complex statistical modeling, often difficult to interpret. see Pachter's 2013 keynote address describing how Cufflinks was (not) reviewed
  - highly dependent on the quality of feature annotation
  - Many more transcripts than genes, thus higher multiple testing penalty and potentially lower sensitivity
  - Generally introduces extra noise
- **Long-read sequencing is a solution here**

Transcript-Level Counting & Alternative splicing

- **splice junction counting as a proxy for differential isoform expression**
  - JunctionSeq , Hartley & Mullikin (2016) Nucleic Acids Research
  - WHIPPET , Blencowe et al (2018) Molecular Cell

Gene-Level Counting

- **Collapsing reads from all alternative spliced transcripts to one gene feature simplifies counting**
- **Recent insights indicate gene-level counting is preferred due to performance and interpretability**
- **However, differential isoform usage can lead to inflated false discovery rates when gene-level counting**
  - this effect is **relatively minor in most real datasets**
  - can be addressed by incorporating offsets from transcript-level abundance estimates
  - → see the *tximport* Bioconductor package Soneson *et al* (2016) F1000Research 4:152

Approaches to RNA-Seq Abundance Estimation

- **RPKM/FPKM/TPM**
  - Normalization for feature length and library size
  - Cufflinks combines FPKM counts with complex models for density deconvolution
- **“Raw counts” used for subsequent abundance estimates by fitting to negative binomial distribution**
  - Technical and biological noise is estimated from data
  - Employed by edgeR, DESeq2

RPKM/FPKM and TPM

- **Reads Per Kilobase per Million mapped reads**
- **Fragments Per Kilobase per Million mapped reads**
  - Same as RPKM but accounts for paired-end reads

→ **sum of all RPKM is not the same between samples**

- **Transcripts Per Million :**
  - idem but operation order differs

→ **proportionality constants are comparable between experiments**

  - Li & Dewey 2011, Wagner *et al* 2012, Dillies *et al* 2012

https://rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

Read Counting with STAR

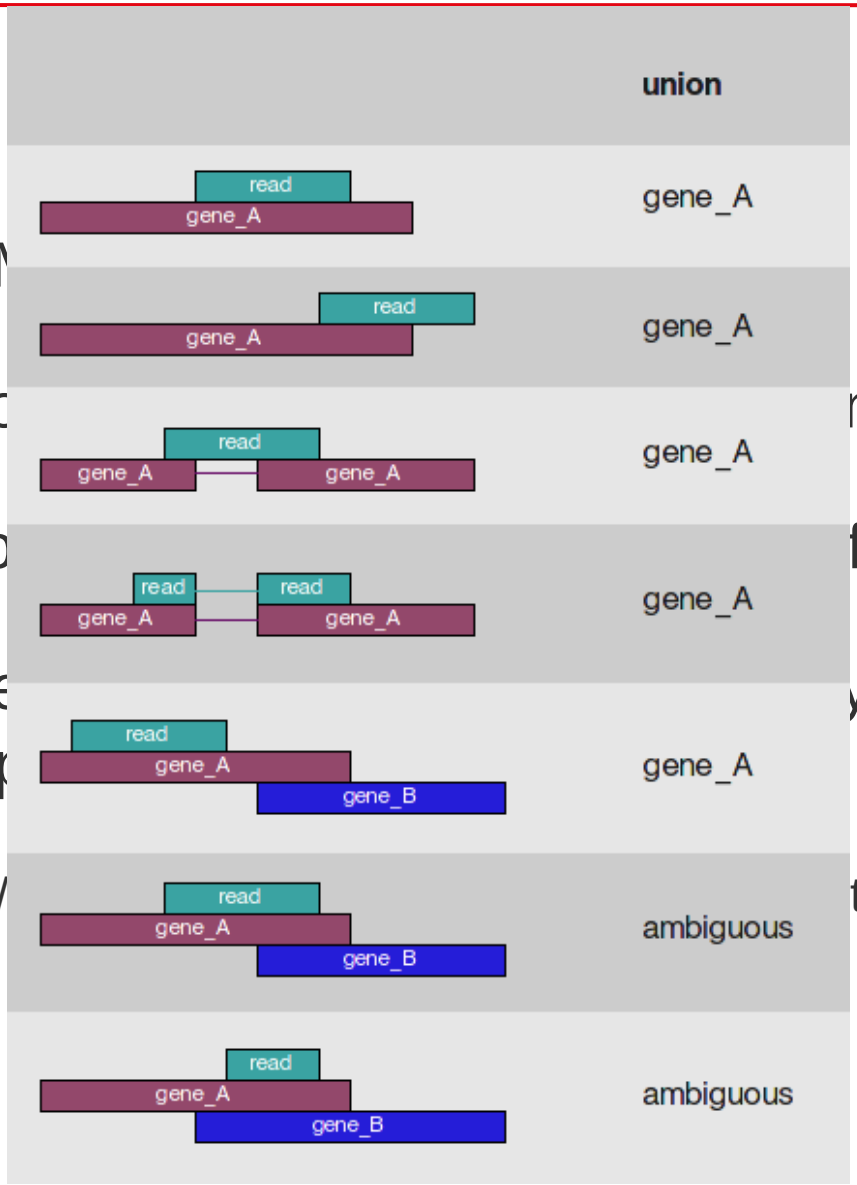
- **Use** --quantMode GeneCounts
- “A gene is counted if it overlaps one and only one gene”
- “Both ends of the paired-end are checked for overlaps”
- This coincides with the counts produced by htseq-count with default parameters :

https://htseq.readthedocs.io/en/master/count.html

Read Counting with STAR

- **Use** --quantM
- “A gene is co nly one gene”
- “Both ends o for overlaps”
- This coincide htseq-count with default p

https:// tml



Read Counting with FeatureCount

http://subread.sourceforge.net/featureCounts.html

- **FeatureCount is actually a part of the larger Subread package**
- **It summarizes the counts in one or several .bam/.sam files at a given level:**
  - **feature (*eg.* exon)**
  - **meta-feature (*eg.* gene)**

- **Requirements:**
  - an annotation file (gtf/gff)
  - Paired-end or single-end ?
  - Stranding information
  - a decision about how to treat multi-mapping/overlapping reads (generally discarded)

Read Counting with FeatureCount

http://subread.sourceforge.net/SubreadUsersGuide.pdf

- **Reads are counted if any overlap are found between read and feature.**

  change with –minOverlap

- **Multi-mapping reads : not counted**

  change with -M and –fraction

- **Multi-overlapping genes : not counted**

  change with -O and --fraction

Practical 5

- Go to the website and do the featureCount practical

REFERENCES

Griffith *et al* (2010) “Alternative expression analysis by RNA sequencing” Nature Methods 7:843-847.

Everaert *et al* (2017) “Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data” Scientific Reports 7:1559.

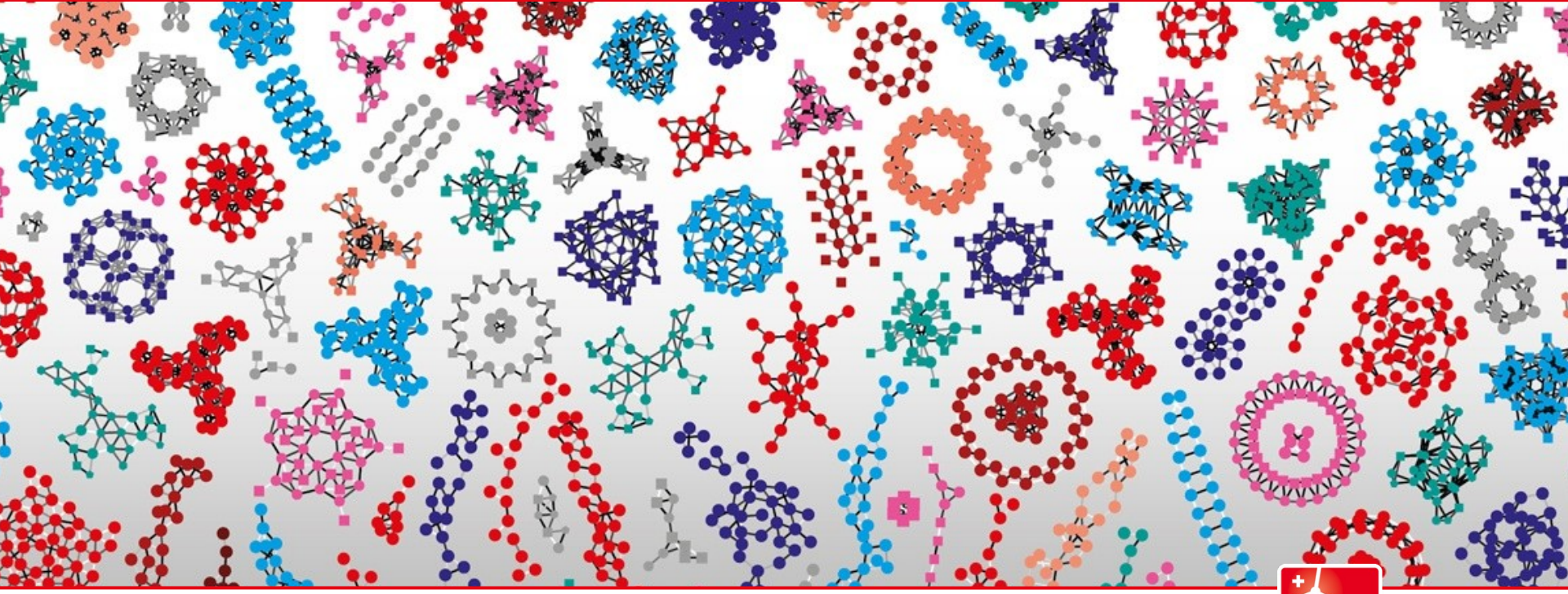
Soneson *et al* (2016) “Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences” F1000 Research 4:1521.

Li & Dewey (2011) BMC Bioinformatics 12:323.

Wagner *et al* (2012) Theory Biosciences 131(4):281-285.

Dillies et al (2012) “A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis” Briefings Bioinformatics 14(6):671-683.

Liao Y, Smyth GK and Shi W (2014). “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.” Bioinformatics, 30(7):923-30._



Swiss Institute of Bioinformatics

Contributors:

Wandrille Duchemin
Geoffrey Fucile
Walid Gharib
Pablo Escobar Lopez
Mark Robinson