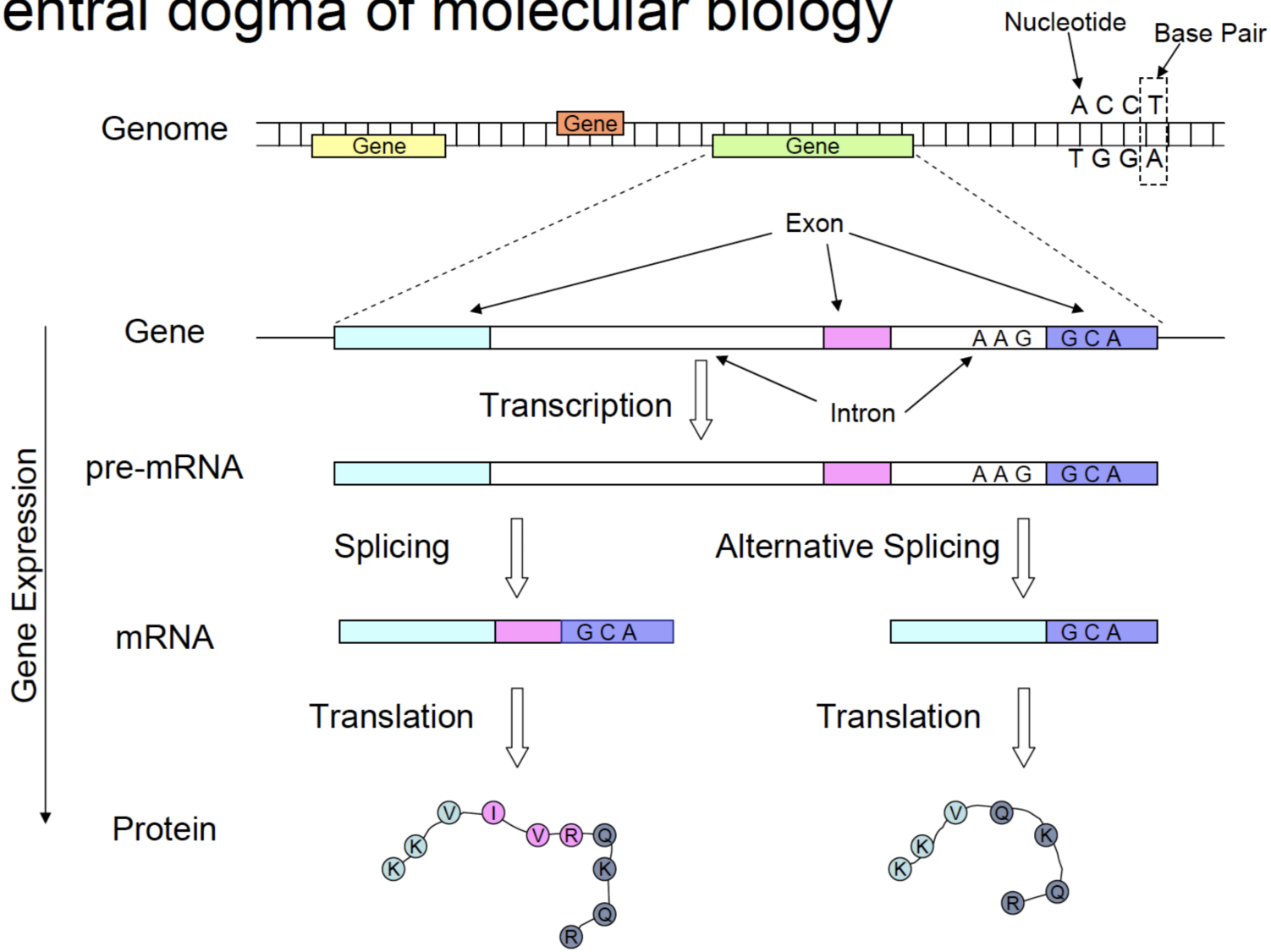
Whole transcriptome sequencing data analysis workshop

Differential splicing

Simone Tiberi, University of Zurich

1-8/02/2019

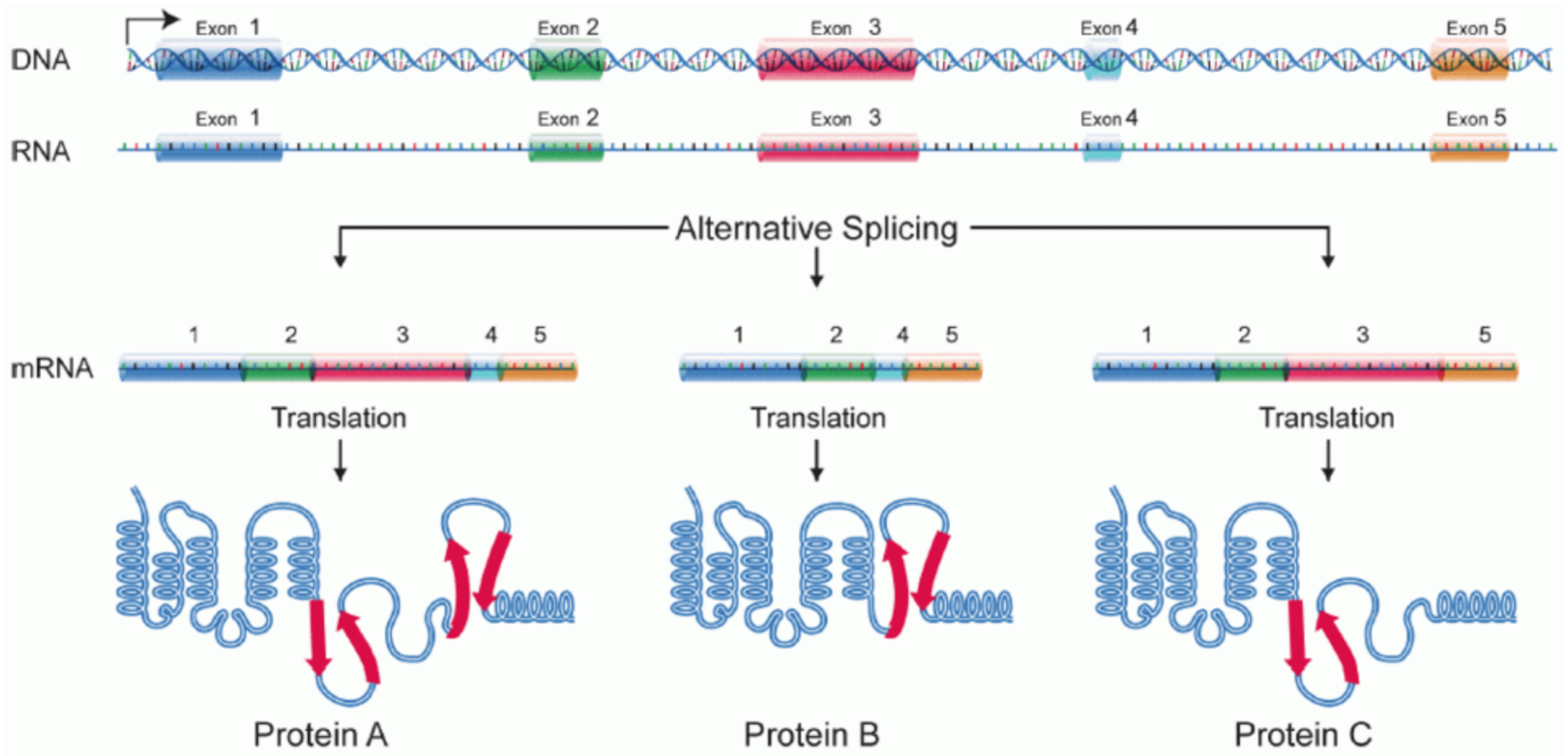
Central dogma of molecular biology



92–94% of human genes undergo alternative splicing,

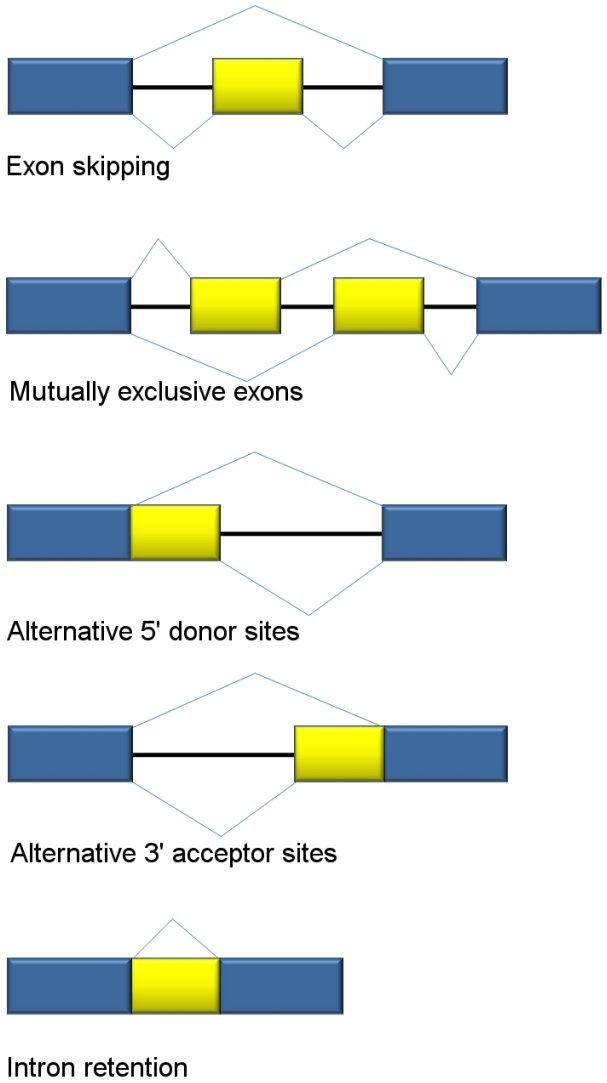
Hubert Rehrauer, ETH Zurich

Alternative splicing



Wikipedia

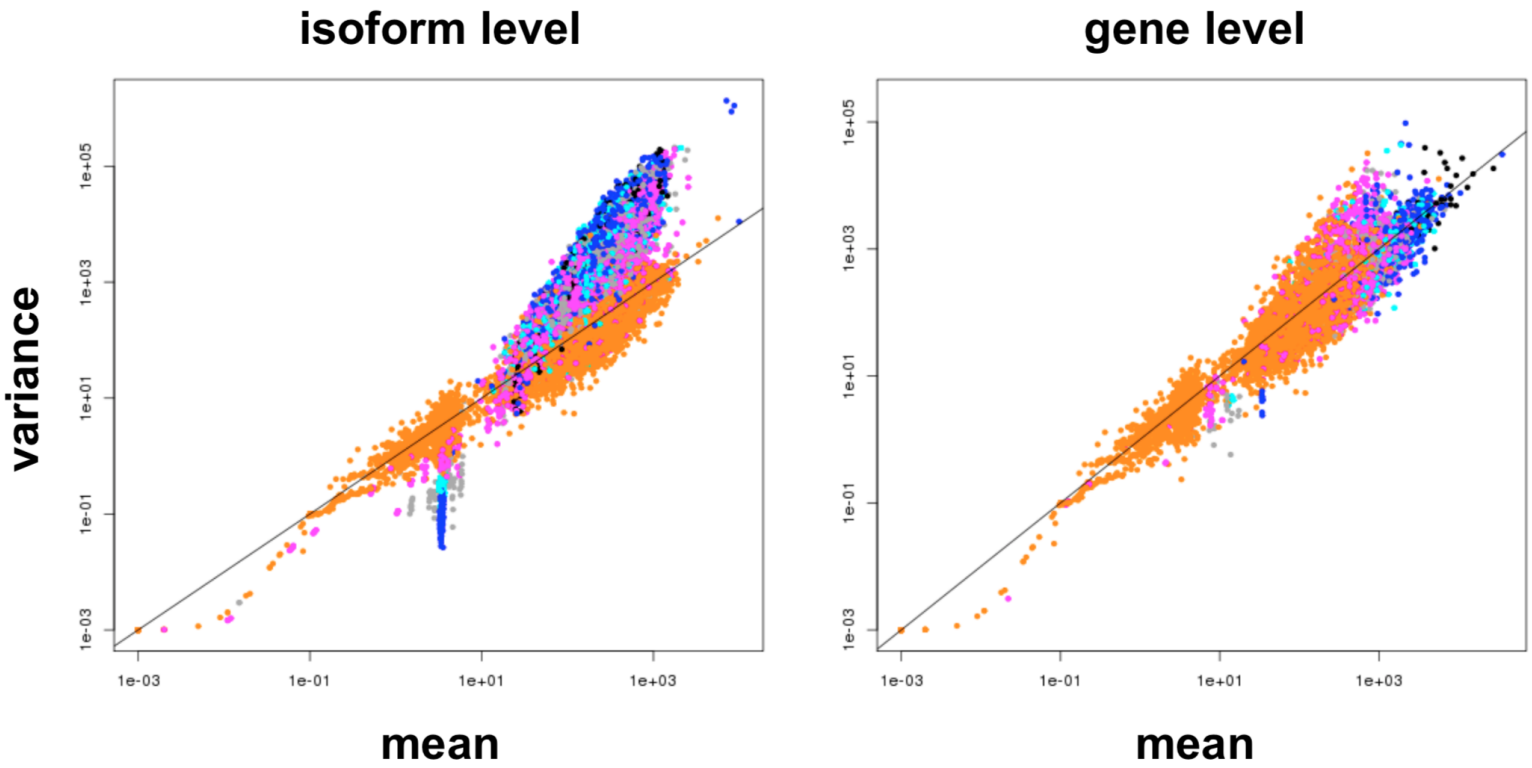
Alternative splicing



Wikipedia

More variable counts

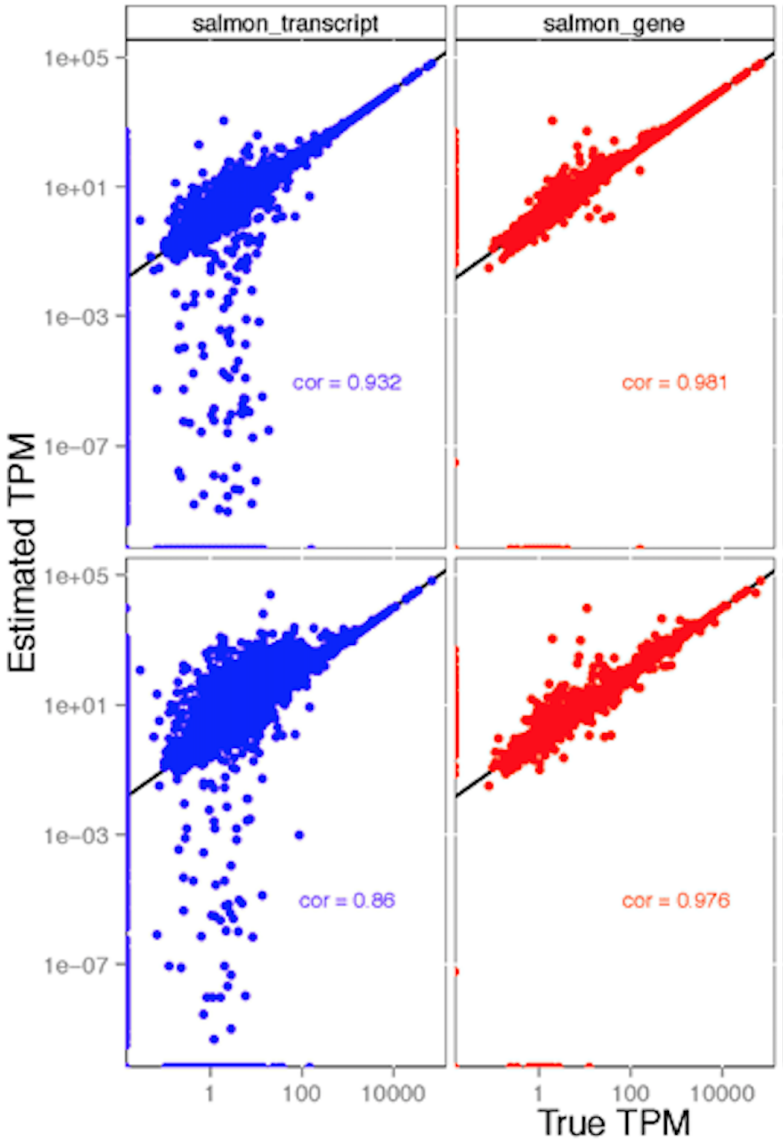
- Transcript level counts have higher variability than gene level counts: more biological variability.



Hubert Rehrauer, ETH Zurich

Less accurate estimates

- Transcript level estimates are less accurate than gene level estimates: higher measurement error.

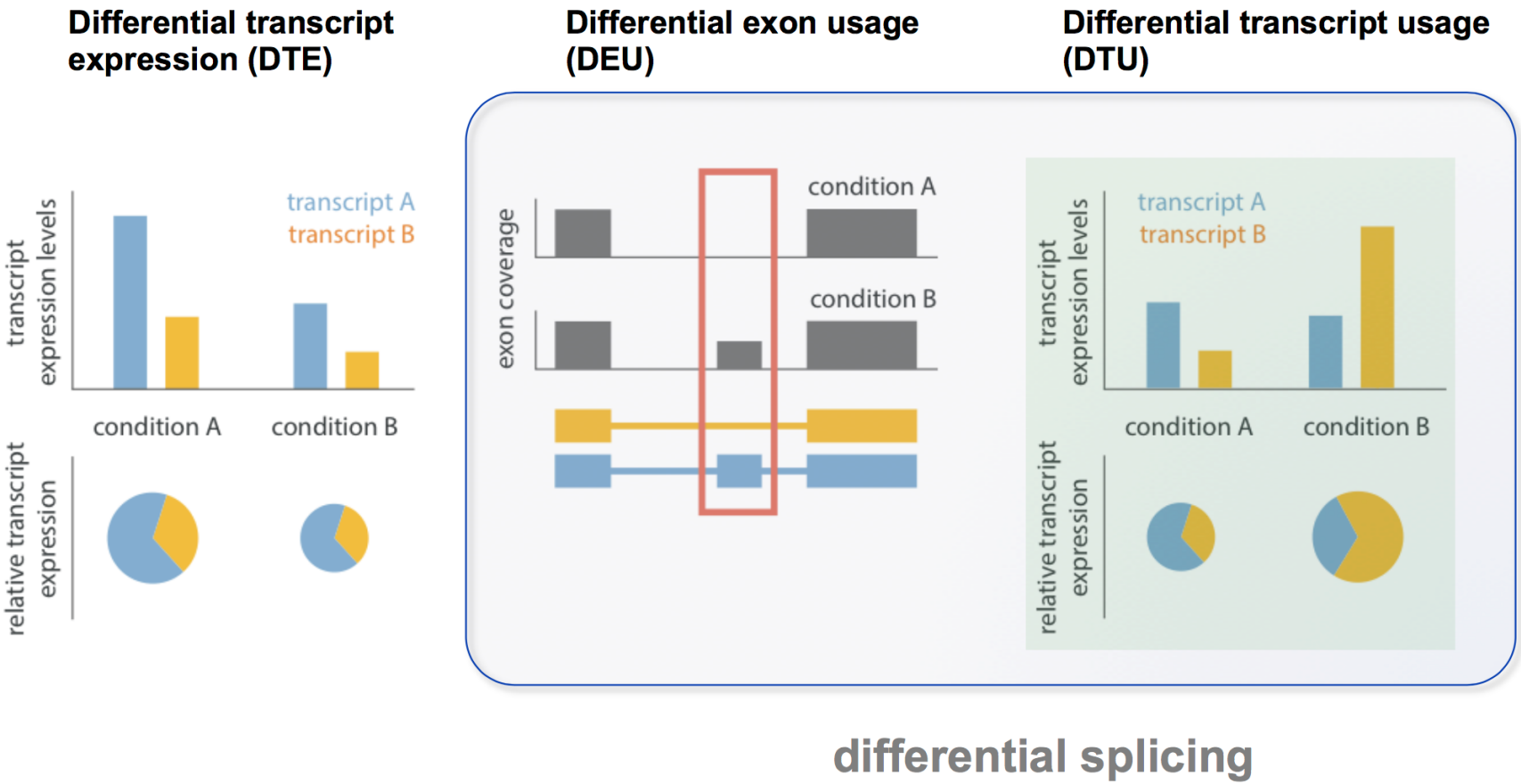


Soneson et al. F1000Research, 2015

DTE, DTU & DEU

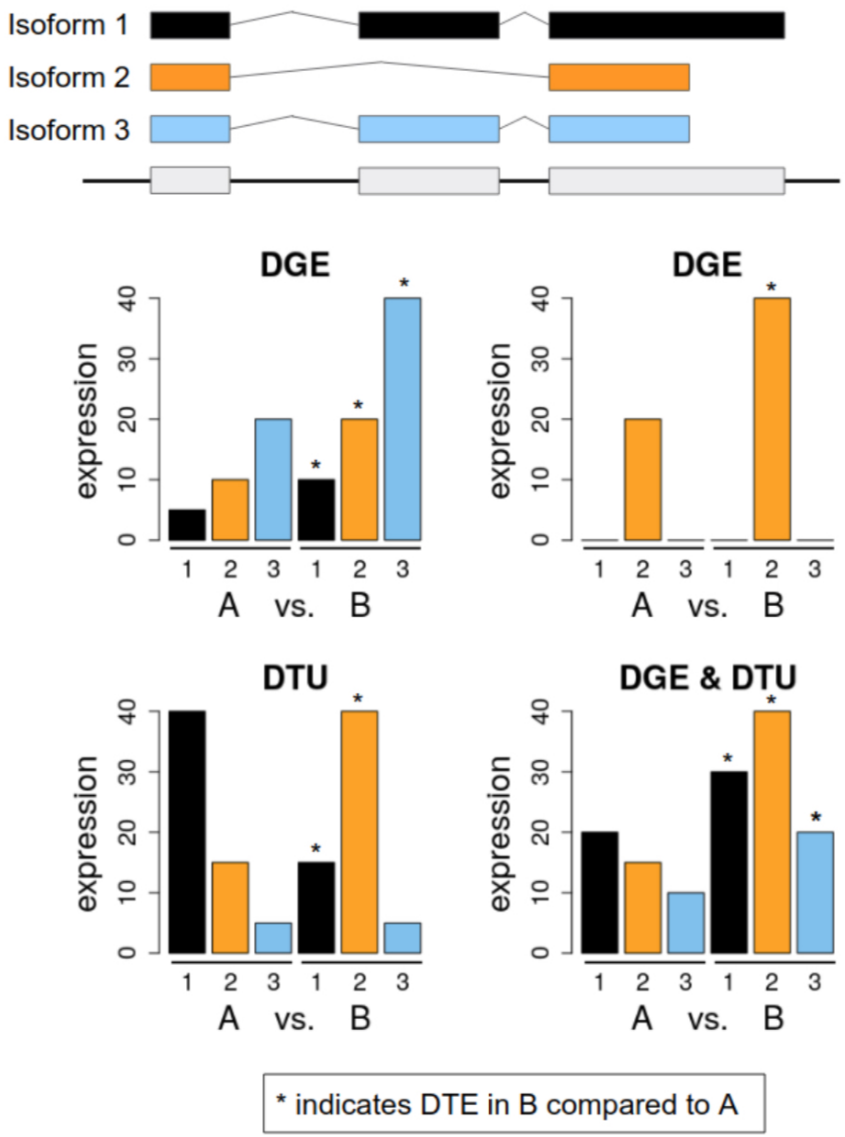
- Differential splicing (DS) is the field that studies if splicing patterns vary between conditions.
- The main branches for differential splicing are differential transcript expression (DTE), differential transcript usage (DTU) and differential exon usage (DEU).
- **DTE** happens when the overall expressions of transcripts change between conditions (similar to DGE on transcripts instead of genes).
- **DTU** happens when the relative abundances of transcripts within a gene change between conditions, i.e. when the proportions of transcripts change. DTU implied DTE, not viceversa.
- **DEU** happens when the relative abundance of exons within a gene change between conditions, i.e. when the proportions of exons change.
- DEU is similar to DTU but it focuses on exons instead of transcripts. DTU is more appropriate because it looks at transcripts, which we are interested in: DEU is a proxy for DTU which focuses on exons for convenience.

DTE, DTU & DEU



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

DTE, DTU & DEU



Van Den Berge et al., PeerJ Preprints, 2018

DTE vs DGE + DTU

- Conceptually, pure DTE points to all kinds of DE and while casting a wide net of potentially interesting genes might seem appealing, there are some considerations to be made.
- If a given transcript is DE, often the question becomes: what happens to the expression of the other transcripts for this gene? Are all transcripts changing in the same direction? If so, it may be better in terms of sensitivity and power to detect an aggregated output (i.e., DGE).
- Transcript level expression can be represented as a genewise multivariate outcome, and isoform switches considered collectively, i.e., by assessing DTU, which is not affected, in either direction, by DGE.
- DTU implies DTE while the opposite is not necessary true.
- Generally speaking, we favour the strategy of two clear, but orthogonal, analyses (DGE and DTU), over a catch-all DTE approach.

Transcript mapping uncertainty

Slide adapted from Trapnell et al. (2013), Nat Botech

- A big challenge in differential splicing analyses is that counts at the transcript level are not observed because many/most reads/fragments map to multiple transcripts.
- Three ways to overcome this issue are presented next.

Input data (1): estimated transcript level counts

Slide adapted from Trapnell et al. (2013), Nat Botech

- Many DTU & DTE methods use transcript level estimated counts, obtained via EM algorithms, e.g. Salmon and Kallisto, and use these counts as input (*plug-in* like approach).
- The drawback is that transcript level estimated counts are treated as true counts, hence neglecting the uncertainty in their estimate.
- Approach used by **BayesDRIMSeq**, **DRIMSeq**, **SUPPA2**, **limma** and **sleuth** (sleuth and rats uses bootstrap replicates to account for the uncertainty in the estimates).
- Input data:
  - $\hat{\theta}_{blue}$: 4.4 and $\hat{\theta}_{red}$: 5.6.

Input data (2): equivalence classes

Slide adapted from Trapnell et al. (2013), Nat Botech

- Other methods for DTU & DTE, such as **cjBitSeq** and **Bspliced** (soon out!), take as input the actual alignment of fragments in the genome and consider what transcripts each fragment maps to, e.g. via TopHat or STAR.
- The information about all fragments is typically summarized in equivalence classes: all fragments mapping to the same transcripts are grouped together in one equivalence class by counting the total number of fragments in the class.
- Input data:
  - Equivalence classes: $C_1 = \{blue\}$, $C_2 = \{red\}$ and $C_3 = \{blue, red\}$;
  - Counts: $f_1 = 1$, $f_2 = 2$ and $f_3 = 7$.

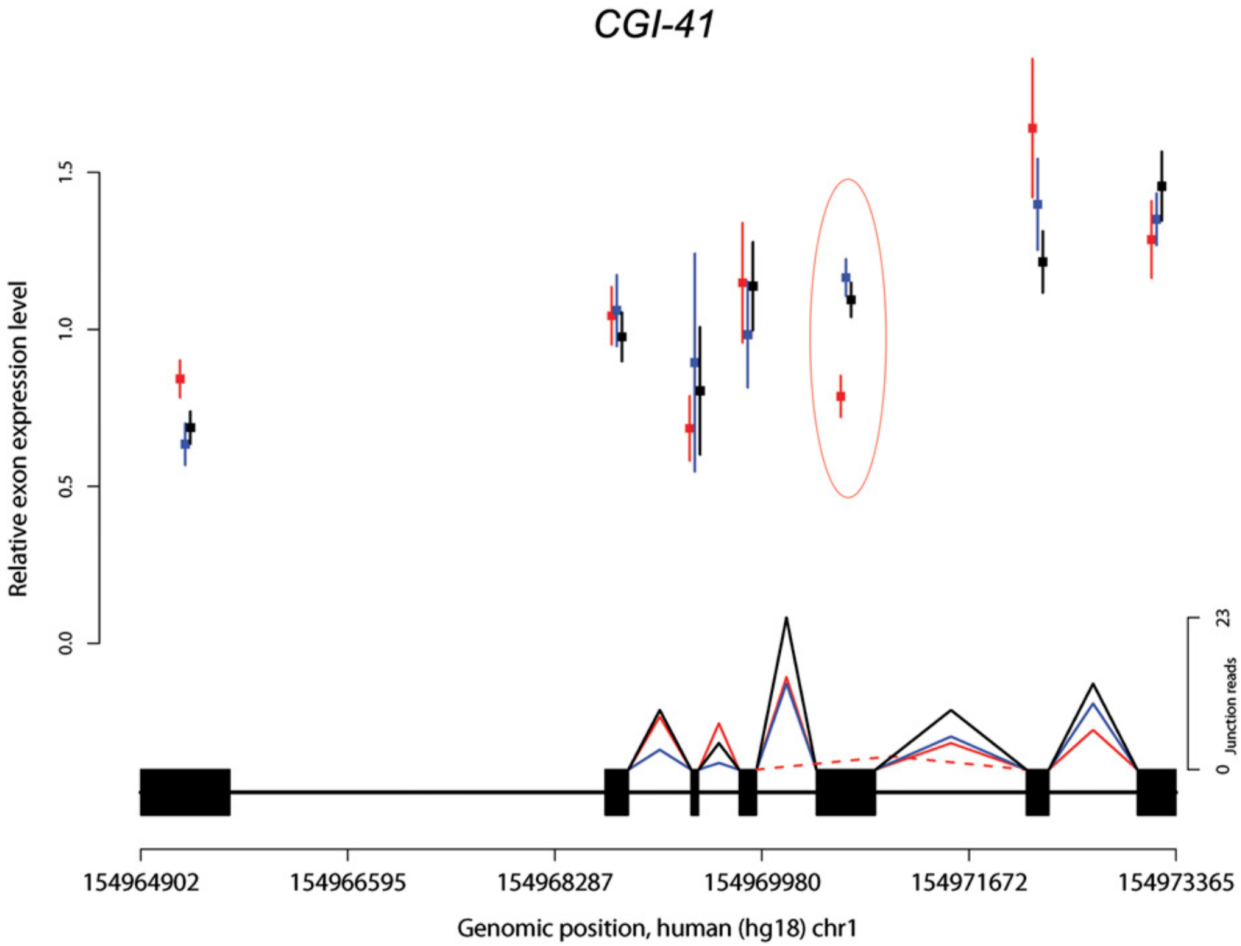
Input data (3): disjoint bin counts

Slide adapted from Trapnell et al. (2013), Nat Botech

- **DEXSeq**, instead of considering transcript level counts, focuses on exon-level counts (where there is no mapping uncertainty) and studies DEU.
- DEXSeq divides the genome into disjoint bins (4 bins in the previous example) and takes as input the counts in every bin, which (unlike transcript level counts) are observed.
- Input data:
  - Exon bin counts: $e_1 = 1$, $e_2 = 2$, $e_3 = 0$ and $e_4 = 7$.

An example of adaptation

- An example of human-specific exons skipping, between human (red), chimpanzee (blue) and rhesus macaque (black).



Blekhman et al. (2009). Sex-specific and lineage-specific alternative splicing in primates, Genome Res.

Overview

- DTE is very similar to DGE, the main differences are that transcript level counts are not observed directly (they are latent states) and that fewer counts are available (hence the variability is higher).
  - **sleuth** and **cjBitSeq** are popular methods for DTE. sleuth inputs transcript level estimated counts (input (1)), while cjBitSeq inputs equivalence classes (input (2)).
  - Alternatively, we can use the standard methods for DGE (**edgeR**, **DESeq2**, etc...) on transcript level estimated counts (input (1)) and test each transcript for DE separately.

DTU: Multinomial for one sample

- We consider one gene with K transcripts, where N samples are available.
- The transcript level counts for each sample of that condition are assumed, *a priori*, to have been generated from a Multinomial distribution:

$$X^{(i)}|\pi^{(i)} \sim Multinom(n^{(i)}, \pi^{(i)}), i = 1, ..., N, \quad (1)$$

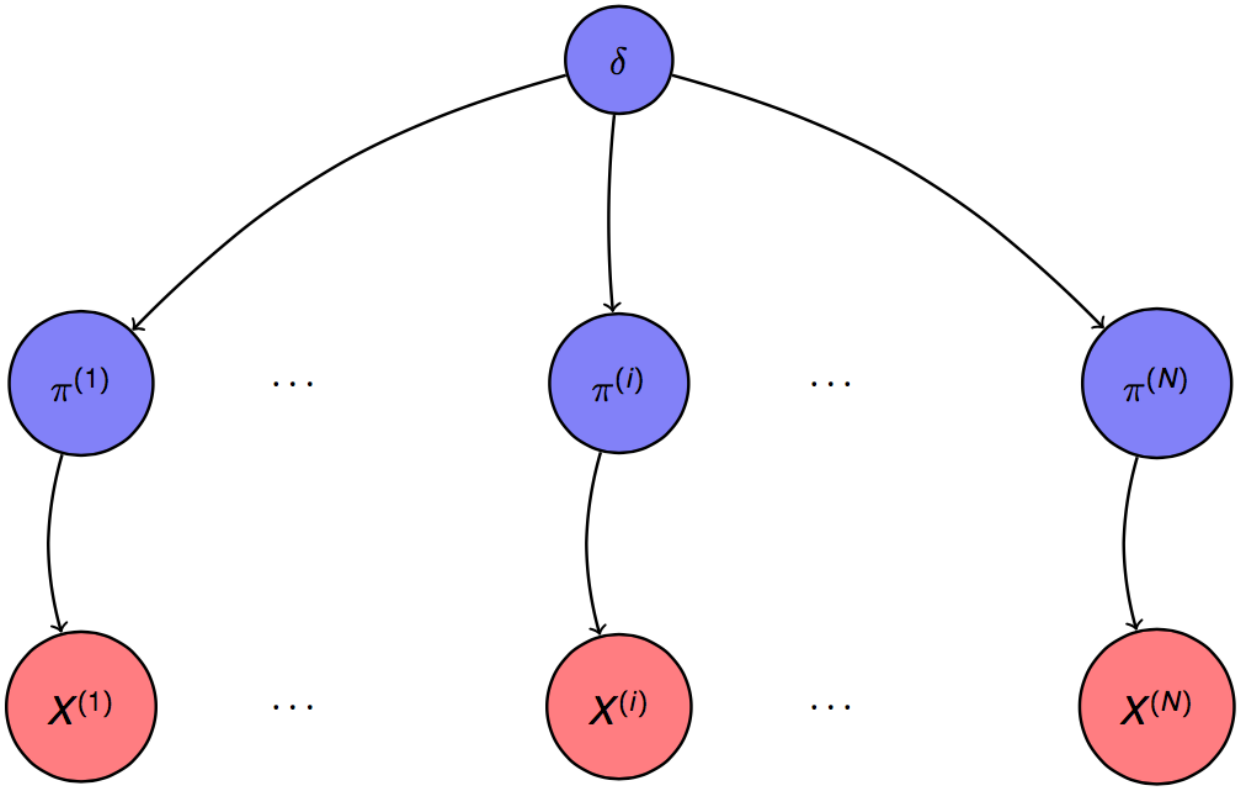
  where $\pi^{(i)} = (\pi_1^{(i)}, ..., \pi_K^{(i)})$ indicates the relative expression of transcripts $1, ..., K$ within the gene and $n^{(i)}$ represents the total number of counts aligning to the gene of interest in the i-th sample.
- When modelling gene expression, the negative binomial (NB) extends the Poisson distribution by allowing for over-dispersed data.
- Similarly, we extend the multinomial distribution allowing for extra variability, via random effects (frequentist statistics) or hierarchical modelling (Bayesian statistics).
- $\pi^{(i)}$ is assumed to vary between samples (under the same condition) due to biological variation.

Hierarchical structure/random effect model

- A hierarchical structure or random effect model is assumed where $\pi^{(i)}$ is assumed, *a priori*, to have been generated from a common distribution for all samples:

$$\pi^{(i)} \sim Dirichlet(\delta), i = 1, ..., N, \tag{2}$$

with $\delta = (\delta_1, ..., \delta_K)$.



DTU: Dirichlet-multinomial for multiple samples

- The marginal probability of the transcript level counts conditional on the hyper-parameter δ is

$$f(X^{(i)} = x|\delta) = \int f(x, \pi|\delta)\, d\pi = \int f_{Multin}(x|\pi)\, f_{Dir}(\pi|\delta)\, d\pi. \quad (3)$$

- This integral can be solved in closed form integrating out π, where (3) is the density of a Dirichlet-multinomial (DM) distribution.
- The distribution of $X^{(i)}$ given δ turns out to be a Dirichlet-multinomial (DM):

$$X^{(i)}|\delta \sim \mathcal{DM}(n^{(i)}, \delta), i = 1, ..., N. \quad (4)$$

- δ can be decomposed in:
  - $\delta_+ = \sum_{k=1}^{K} \delta_k$, the dispersion parameter indicating how proportions $\pi^{(i)}$ vary between samples;
  - $\bar{\pi} = \frac{\delta}{\delta_+}$, representing the mean relative abundance of transcripts (mean across samples).
- When comparing two conditions, interest lies primarily in testing whether $\bar{\pi}$ varies between conditions.

Methods for DTU

- DRIMSeq and BayesDRIMSeq use the Dirichlet-multinomial structure described, and both provide a global p-value at the gene level.
- Main tools for DTU: **DRIMSeq**, **BayesDRIMSeq**, **cjBitSeq**, **SUPPA2** and **Bspliced** (soon out!).
- Of these DRIMSeq, BayesDRIMSeq and SUPPA2 input estimated transcript-level counts (input (1)).
- cjBitSeq instead works with equivalence classes (input (1)), but it uses a simpler (non-hierarchical) multinomial framework.
- Bspliced, similarly to cjBitSeq, inputs equivalence classes counts, but also allows for to model the sample-to-sample variability via a Dirichlet-multinomial distribution.

DEU: DEXSeq

- **DEXSeq** has a similar mathematical structure to DGE methods: for every gene, it formulates a negative binomial (NB) model with a baseline coefficient (β^G), an exon coefficient (β^E), a condition effect (β^C) and an exon/condition interaction (β^{EC}).
- For every exon, DEXSeq tests if the interaction term, β^{EC}, is significant; i.e., if the condition significantly affects the expression of a specific exon.

$$K_{ijl} \sim NB\Big(\text{mean} = s_j \mu_{ijl}, \text{dispersion} = \alpha_{il}\Big), \quad (1)$$

where α_{il} is the dispersion parameter (a measure of the distribution's spread; see below) for counting bin (i, l), and the mean is predicted via a log-linear model as

$$\log \mu_{ijl} = \beta_i^{\mathrm{G}} + \beta_{il}^{\mathrm{E}} + \beta_{i\rho_j}^{\mathrm{C}} + \beta_{i\rho_j l}^{\mathrm{EC}}. \quad (2)$$

i – gene
j – sample … ρ_j is condition (categorical)
l – bin

β^G – baseline "expression strength"
β^E – "exon" (bin) effect
β^C – condition effect
β^{EC} – condition x "exon" interaction

DEU: DEXSeq

- A drawback of DEXSeq is that tests are performed on individual exons and not on the entire gene: the p-value for the gene is obtained as the minimum p-value across the exons (with a special correction for dependent p-values).
- DEXSeq has also been successfully used to perform DTU on transcript level estimated counts (input (1)).

Event-specific DS

- An alternative approach to perform differential splicing (DS) is to consider percent spliced in (PSI) values.
- PSIs can be computes for specific events (retained intron, cassette exon, etc.) or at the transcript level, and indicate the fraction of RNA-seq reads supporting the event.
- PSIs are obtained as the ratio between the number of reads including the event and the total number of reads including and excluding the event.
- The difference of the PSIs between conditions is then used to assess DS, which is performed separately for each event (or transcript).
- The main tools for DS based on PSIs are: **rMATS** and **SUPPA2**.

Multi-stage testing

- DS analyses can be approached at the gene, transcript, exon or even-specific level.
- While gene-level tests often have higher sensitivity (i.e., power), testing each individual transcript, exon or event provides increased resolution.
- However, this does not guarantee control of the false discovery rate on the full set (FDR).
- Stagewise testing procedures, instead, first screen for significant genes, and only consider significant transcripts, exon or events from those genes.
- This procedure gives gene-level FDR control also for transcript, exon or event level tests: allows them to leverage gene-level FDR control, while interpreting transcript, exon or event specific tests.
- The R package stageR implements multi-stage testing procedures in R.

Splicing quantitative trait loci (sQTL)

- sQTL in DS are the analogous of eQTL in DGE.
- eQTL: if we have information about phenotypes, via SNP data, we can test if a phenotype influences gene-expression; i.e., DGE between phenotypes instead of between conditions.
- Similarly, splicing quantitative trait loci (sQTL) tests if a phenotype alters alternative splicing profiles.
- In other words, we can perform DTE, DTU and DEU analyses (with the same methods described above) where the separation in groups is defined by the SNPs.
- As for eQTL, also in sQTL we perform many more tests than for DTE, DTU and DEU: for every gene we need to test many SNPs; we typically only test the SNPs in the neighbourhood of the gene we are considering.
- **DRIMSeq** has a separate function to perform for sQTL.

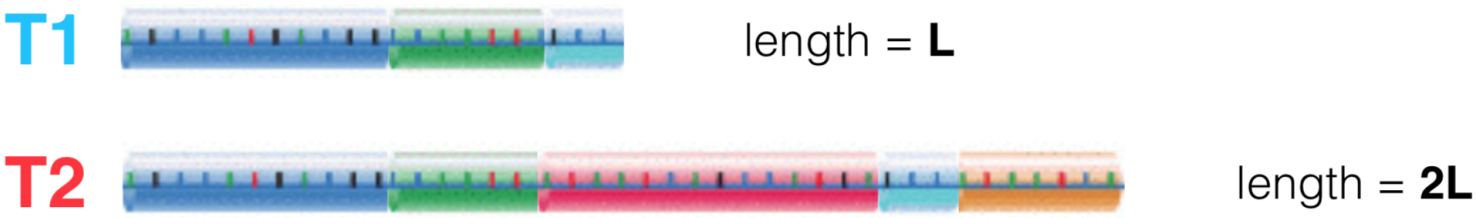
Transcript pre-filtering

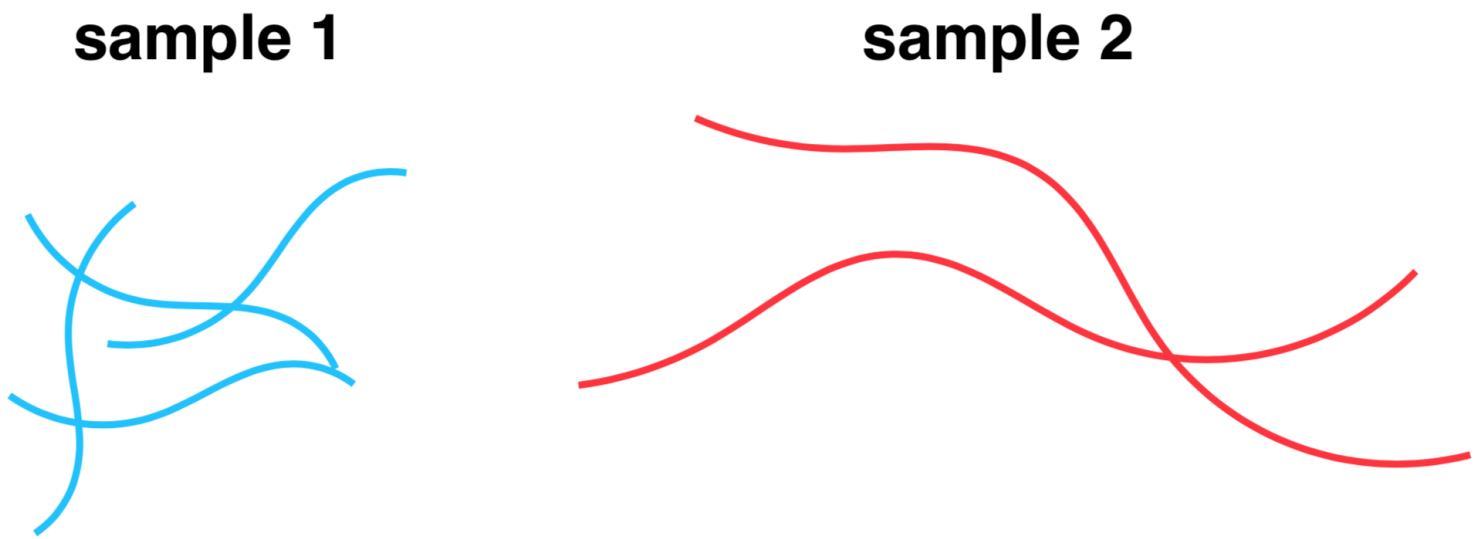
- Most methods for differential splicing, in particular all the ones mentioned here, rely on a transcriptome reference.
- Transcript pre-filtering: Soneson et al. (2016), Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage, Genome Biology.
- Filtering the reference transcriptome (i.e. the cdna fasta file) by removing lowly expressed transcripts improves the performance of differential splicing methods:
  - run transcriptome aligner (e.g. Salmon or kallisto) with standard reference transcriptome;
  - select the transcripts with estimated proportion $<$ threshold (5% seems an ideal choice) and filter the reference transcriptome by removing the lowly abundant transcripts and make a new reference transcriptome;
  - re-run the transcriptome aligner (e.g. Salmon or kallisto) with the new filtered reference.

Differential splicing affects DGE

- Since transcripts have different lengths, differential splicing can affect (in both directions) the detection of differentially expressed genes.
- Problem well described in Soneson et al. (2015), where the authors provide an R package (**tximport**), which allows to load easily transcript level estimates, counts and effective transcript lengths estimated from Salmon, kallisto, etc...
- Very simple idea, easy to implement into other DGE methods (edgeR, DESeq2, etc...) and extremely important.

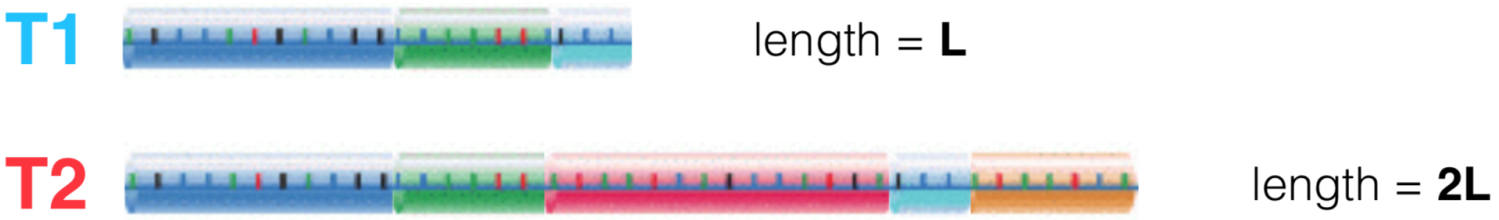
Alternative splicing affects DGE

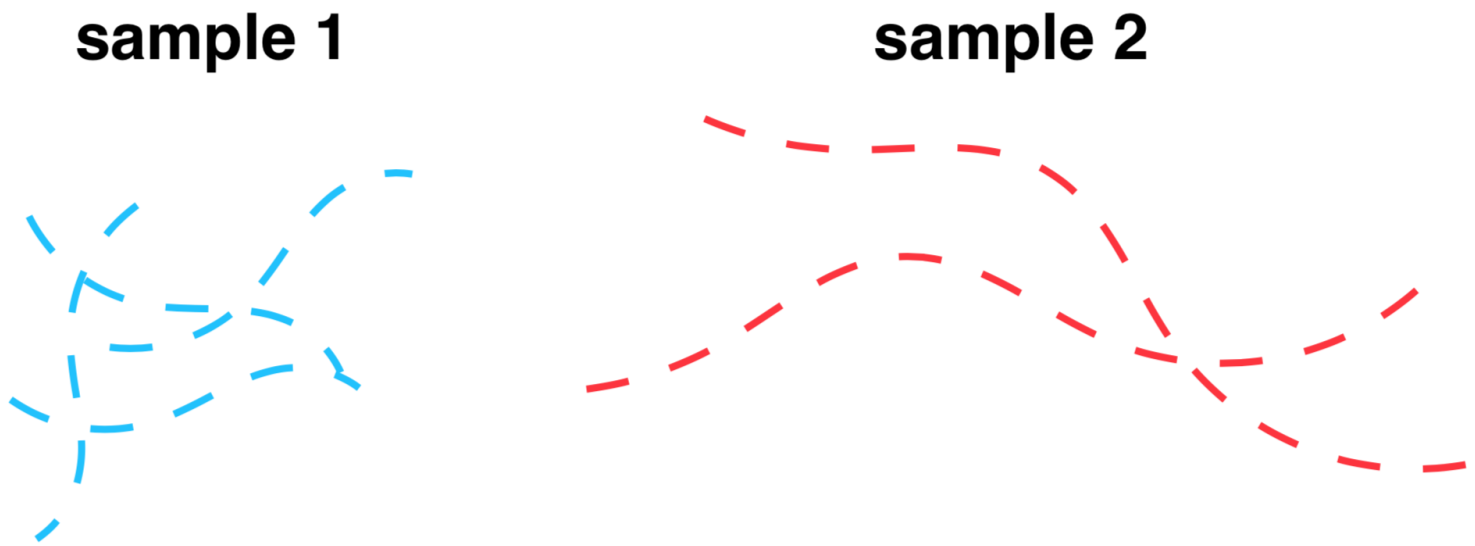




Charlotte Soneson, UZH

Alternative splicing affects DGE





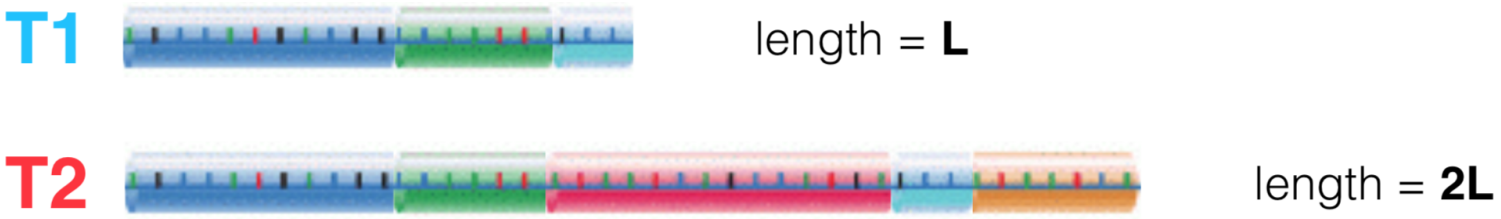
Charlotte Soneson, UZH

Alternative splicing affects DGE

Gene	S1	S2
Count	150	150

Charlotte Soneson, UZH

Average transcript length (ATL)

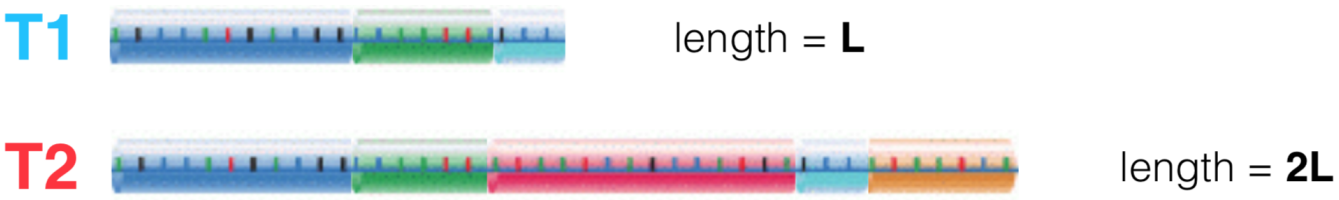


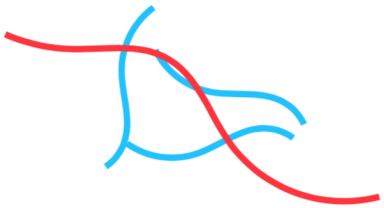
$$ATL_{g1} = 1 \cdot L + 0 \cdot 2L = L$$

$$ATL_{g2} = 0 \cdot L + 1 \cdot 2L = 2L$$

Charlotte Soneson, UZH

Average transcript length (ATL)





$$ATL_{g1} = 0.75 \cdot L + 0.25 \cdot 2L = 1.25L$$

$$ATL_{g2} = 0.5 \cdot L + 0.5 \cdot 2L = 1.5L$$

Charlotte Soneson, UZH

Include ATL in DGE analysis

raw count for gene *i* in sample *j*

relative abundance

scaling factor

$$C_{ij} \sim NB(\mu_{ij} = s_{ij} q_{ij}, \theta_i)$$

dispersion

- Extend scaling factor for given sample and gene to include the average length of the transcripts from the gene that are present in the sample

Charlotte Soneson, UZH

Include ATL in DGE analysis

- Similar to correction factors for library size, but sample-**and** gene-specific
- Transcript abundance levels (TPMs) can be obtained from (e.g.) Salmon or kallisto
- Average transcript length for gene *g* in sample *s*:

$$ATL_{gs} = \sum_{i \in g} \theta_{is} \bar{\ell}_{is}, \qquad \sum_{i \in g} \theta_{is} = 1$$

$\bar{\ell}_{is}$ = effective length of isoform i (in sample s)
θ_{is} = relative abundance of isoform i in sample s

Include ATL in DGE analysis

- The tximport vignette, in the "Use with downstream Bioconductor differential expression packages" section, explains how to incorporate the ATL in the DGE analyses of edgeR, DESeq2 and limma-voom.
- For more info, read:
  `https://bioconductor.org/packages/release/bioc/vignettes/tximport/inst/doc/tximport.html`

References

References I

- **DEU: DEXSeq**: Anders et al. (2012). Detecting differential usage of exons from RNA-seq data, Genome Research.
- **DTU: DRIMSeq**: Nowicka et al. (2016). DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics, F1000Research.
- **DTU: cjBitSeq & BayesDRIMSeq**: Papastamoulis et al. (2017). Bayesian estimation of differential transcript usage from RNA-seq data, Statistical Applications in Genetics and Molecular Biology.
- **DTE: cjBitSeq**: Papastamoulis et al. (2017). A Bayesian model selection approach for identifying differentially expressed transcripts from RNA sequencing data, J. Royal Statistical Society, Series C.
- **DTE: sleuth**: Pimentel et al. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty, Nature Methods.
- **DTU from PSIs: SUPPA2**: Trincado et al. (2018). SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions, Genome Biology.

References II

- **Transcript pre-filtering**: Soneson et al. (2016). Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage, Genome Biology.
- **tximport**: Soneson et al. (2016). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences, F1000Research.
- **iCOBRA**: Soneson et al. (2016). iCOBRA: open, reproducible, standardized and live method benchmarking, Nature Methods.
- **DTU pipeline**: Soneson et al. (2016). Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. F1000Research
- **stageR**: Van den Berge et al. (2017). stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage, Genome Biology.
- **RNA-seq overview**: Van den Berge et al. (2018). RNA sequencing data: hitchhiker's guide to expression analysis, PeerJ Preprints.

Questions?