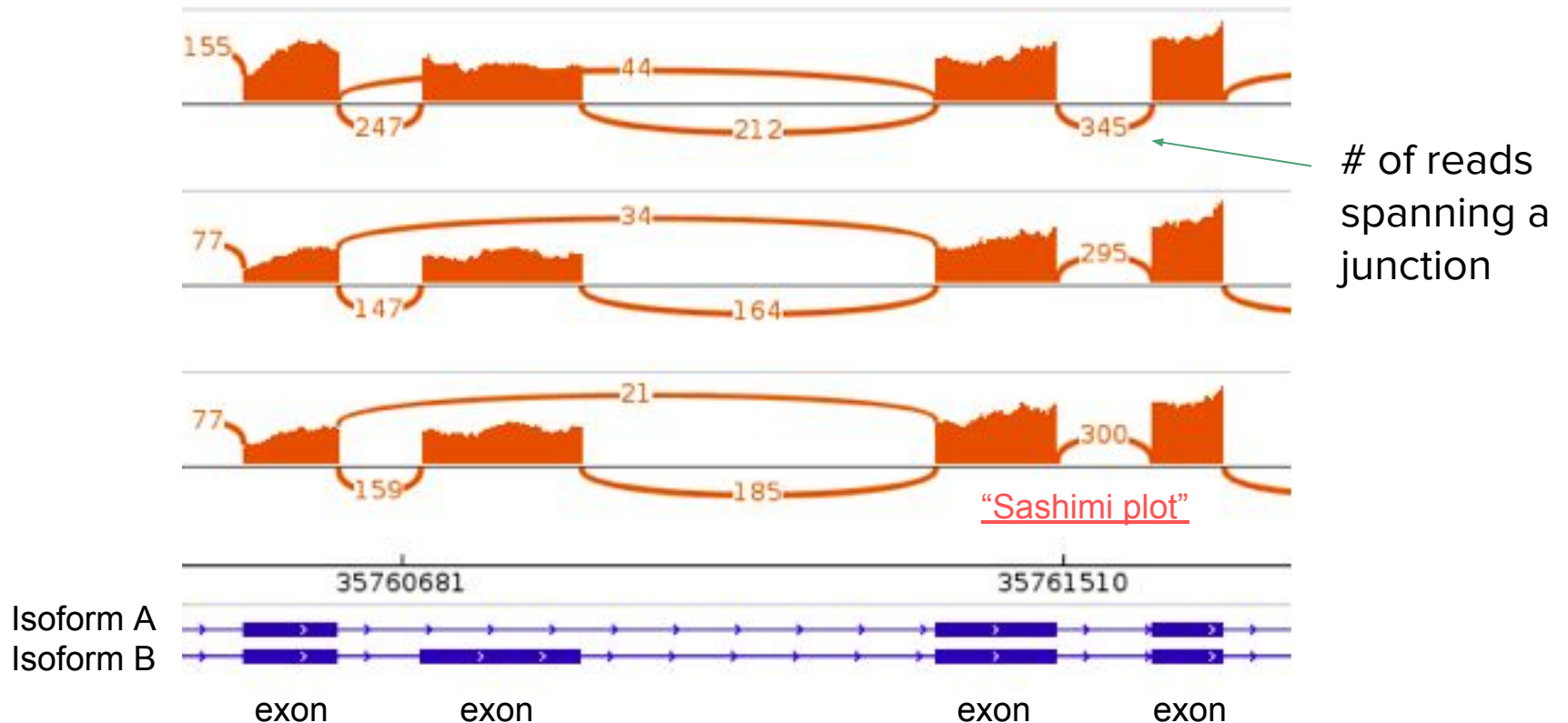


Isoform-level RNA-seq analysis

Michael Love
July 2018

RNA-seq gives us expression of isoforms



Expression of isoforms is tissue specific

582–592 *Nucleic Acids Research*, 2018, Vol. 46, No. 2
doi: [10.1093/nar/gkx1165](https://doi.org/10.1093/nar/gkx1165)

Published online 30 November 2017

Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues

Alejandro Reyes^{1,2,3,*} and Wolfgang Huber^{1,*}

(in these slides, red = hyperlink)

“...we investigated cell type-dependent differences in exon usage of over 18 000 protein-coding genes in 23 cell types from 798 samples of the Genotype-Tissue Expression Project (GTEx)”

Expression of isoforms is tissue specific

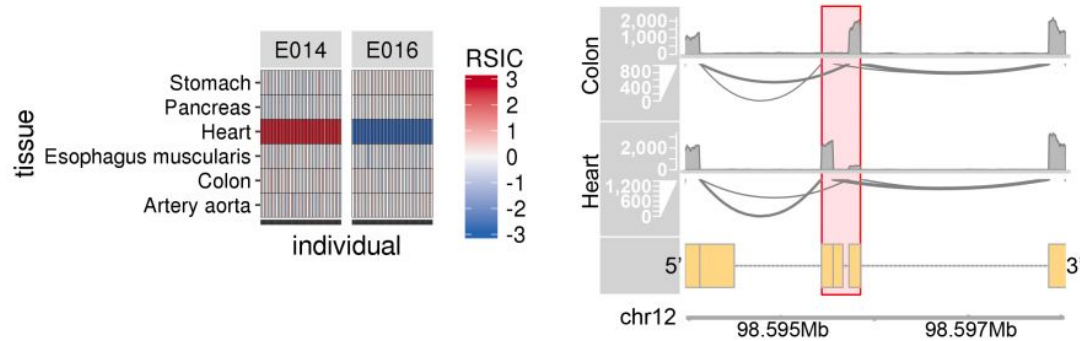
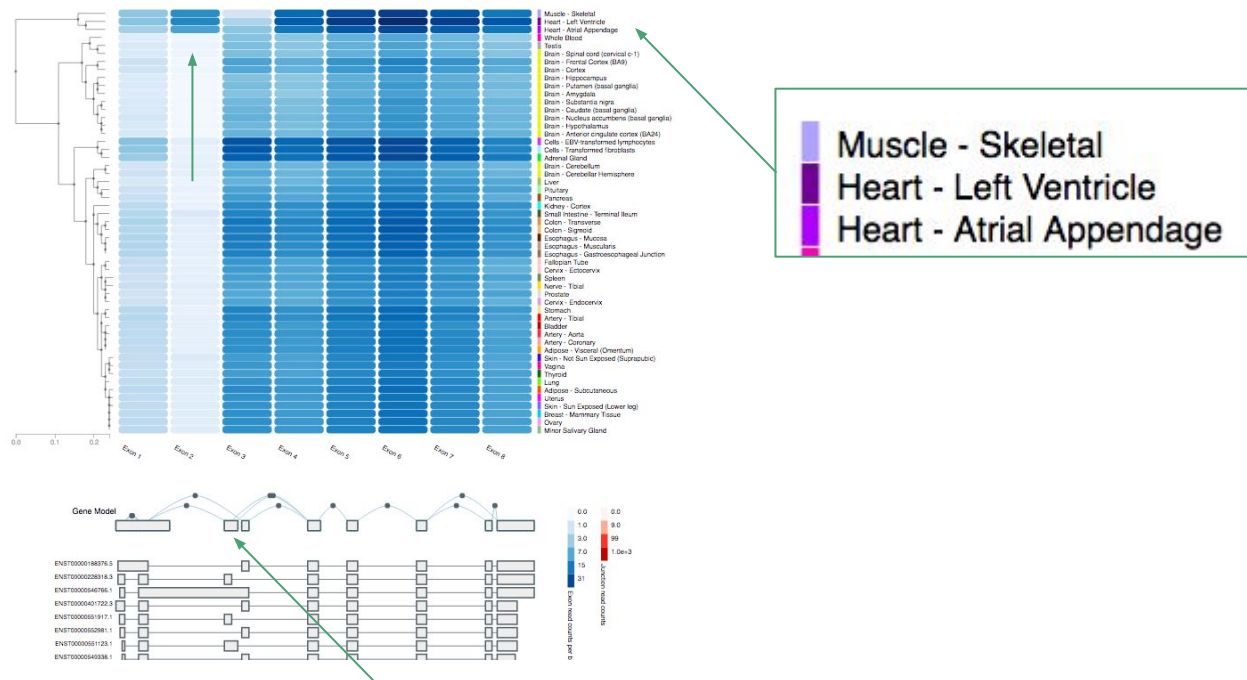


Figure S2: Panel A shows a heatmap representation of relative spliced-in coefficients (RSICs) for two exonic regions (E014 [q-value=0, tissue score=2.17] and E016 [q-value=0, tissue score=2.32]) of the gene *SLC25A3* on *subset C* of the *GTEx* data. Panel B shows sashimi plots of the RNA-seq data from the colon and heart samples from individual 11178. The highlighted area corresponds to the genomic coordinates of the exons shown in Panel A. These mutually exclusive exons were initially described in an independent study^{SR1} that also used RNA-seq data but different bioinformatic methods. As shown in these two panels as well as in figure 1A from the previous study^{SR1}, this splicing event is regulated differently in heart as compared to other tissues.

SLC25A3 also described in Wang et al (2008) [PMC2593745](#)

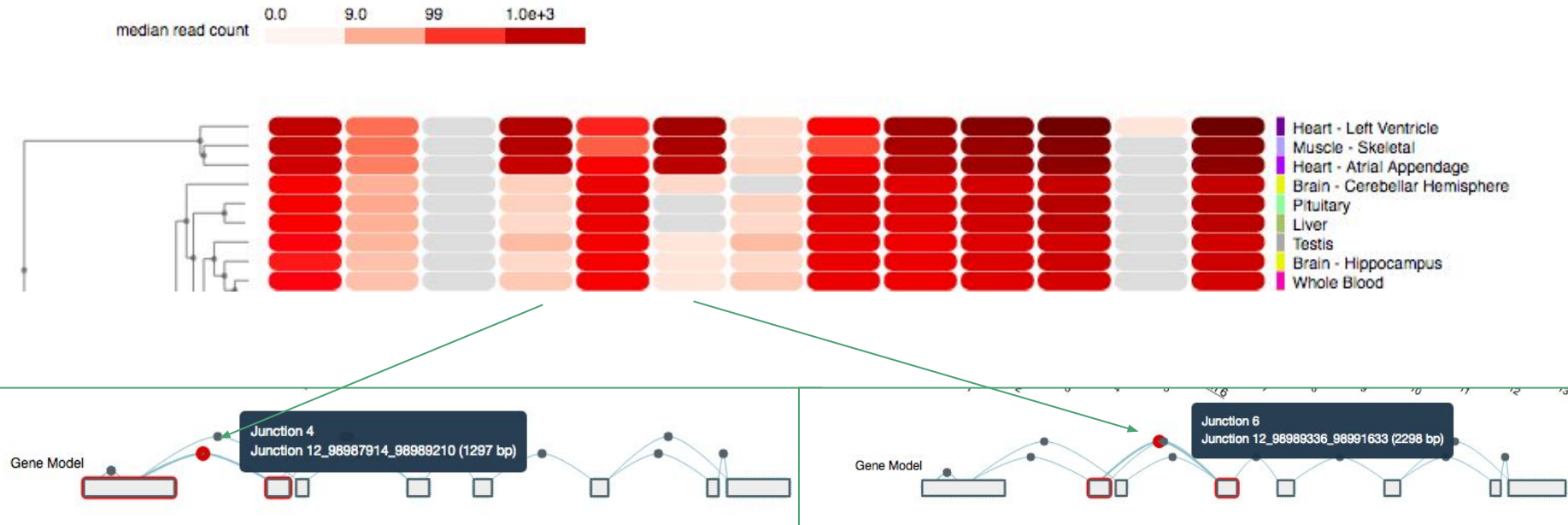
Expression of isoforms is tissue specific (exon)

<https://www.gtexportal.org/home/gene/SLC25A3>

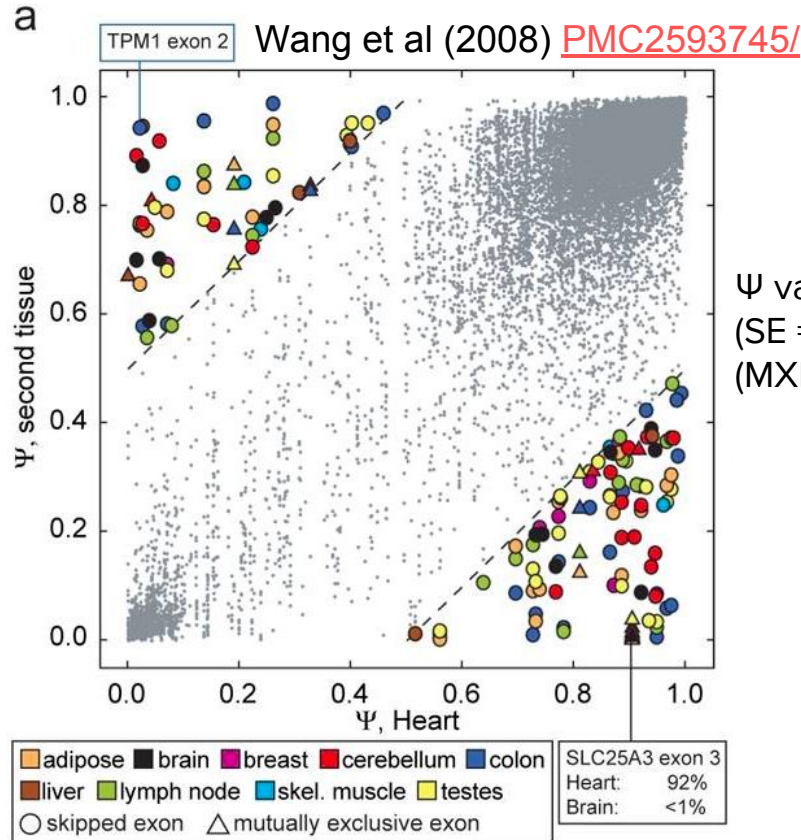


Expression of isoforms is tissue specific (junction)

<https://www.gtexportal.org/home/gene/SLC25A3>



Switch-like alternative splicing



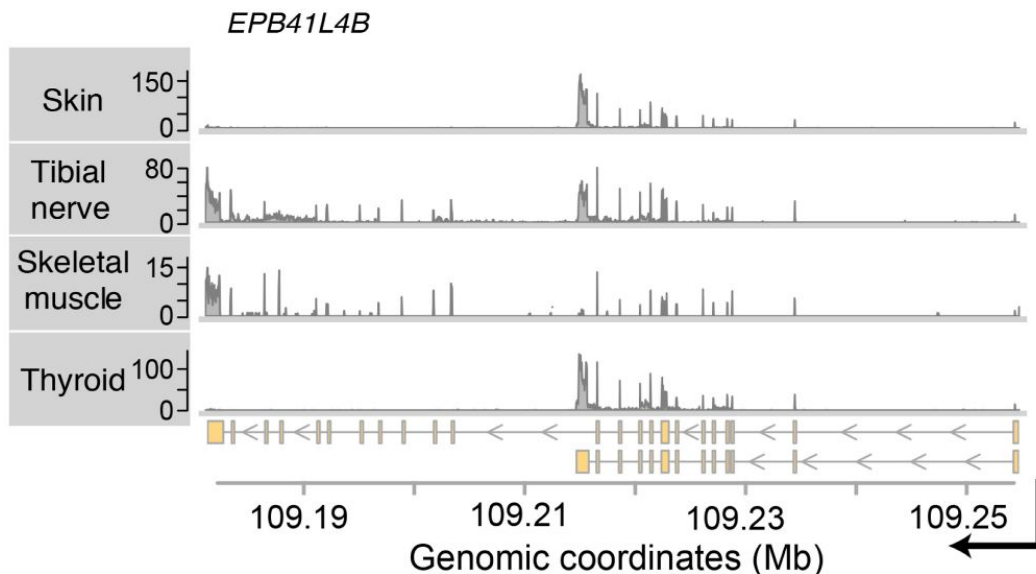
Ψ values of SEs and MXEs
(SE = skipped exons)
(MXE = mutually exclusive exons)

PSI = percent spliced in

Often the Greek symbol
Psi (Ψ) is shown.

Back to Reyes and Huber (2017) [PMC5778607](https://pubmed.ncbi.nlm.nih.gov/31511111/)

- “We found that there is tissue-specific regulation of alternative transcript isoform choice for a large fraction of the human genome, affecting about [half of multi-exonic genes](#).”
- “We estimated that alternative splicing explains tissue-dependent transcript differences for, at most, [35% of the genes](#).”
- “Nevertheless, for the majority of exonic regions with tissue-dependent usage (TDU), the TDU was consistent with [alternative transcription initiation and termination sites](#).”



Scotti & Swanson (2016) [PMC5993438](#)

RNA mis-splicing in disease

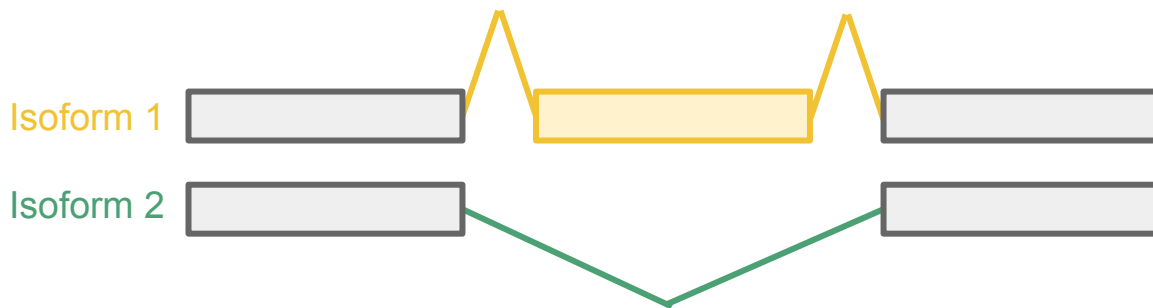
- “Here, we provide an overview of RNA splicing mechanisms followed by a discussion of disease-associated errors”
- “The most common type of mutations that alter splicing patterns are cis-acting and are located in either core consensus sequences (5'ss, 3'ss and branch point (BP)) or the regulatory elements that modulate spliceosome recruitment, including exonic splicing enhancer (ESE), exonic splicing silencer (ESS), intronic splicing enhancer (ISE) and intronic splicing silencer (ISS) elements.”
- “...mutations in core constituents of the spliceosome also underlie a discrete set of diseases, including retinal degenerative disorders and cancer”

Scotti & Swanson (2016) [PMC5993438](#)

- Core spliceosome mutations in [retinitis](#) pigmentosa
- Spliceosome dysregulation in [cancer](#)
- [Development and stress](#): key roles for the minor spliceosome
- Large introns and microexons in [neurological disorders](#)

Methods for quantifying/compared isoform expression

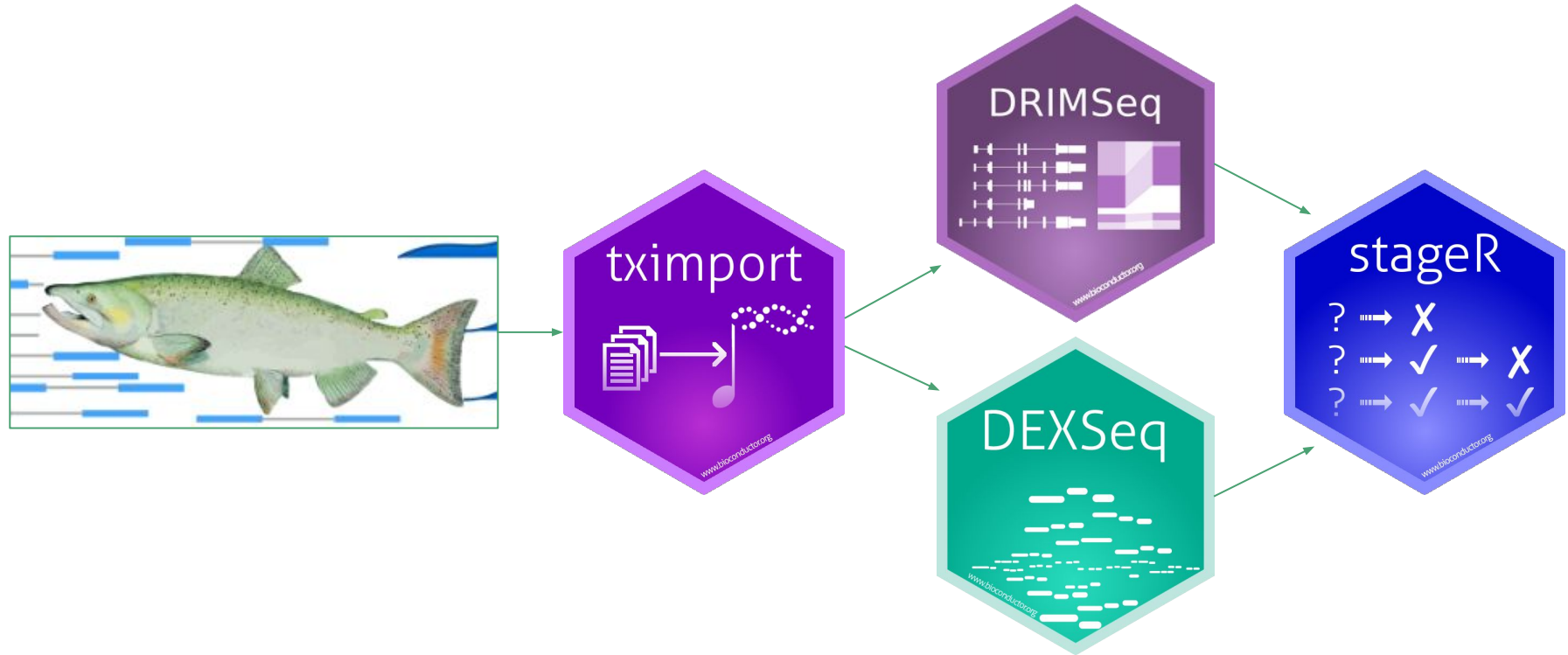
- Quantification of transcript/isoform expression:
 - Alignment-based: Cufflinks, MISO, BitSeq, RSEM, FlipFlop, and many more!
 - Alignment-free: Sailfish, kallisto, Salmon
- Counting reads in exons bins / reads falling on junctions
- Statistical analysis of differences across samples:
DEXSeq, Cuffdiff2, rMATS, MISO, BitSeq, rDiff, DiffSplice, diffSplice() in edgeR/limma, SUPPA2, LeafCutter, DRIMSeq, and many more!



Advantages/disadvantages of *transcript* analysis

- Estimated isoform counts benefit from sophisticated bias modeling
- We have very fast methods for computing transcript counts, and this may reduce the size of the problem
- *Transcripts* are the biological unit: what is actually being expressed
- “...with the emergence of longer reads (fragments), *transcript* quantifications will become more accurate” -DRIMSeq paper
- We need to know all the possible transcripts ahead of time, while exon-based or junction-based can discover novel isoforms
- With exon- or junction-based, maybe additional power from aggregating signal across transcripts if the exon or junction is the relevant feature

My recommended pipeline for differential transcript usage (DTU)



tximport (Soneson 2015) w/ countsFromAbundance

tximport has multiple methods for importing counts from *Salmon*, etc.

- counts + length offset
- scaledTPM
- lengthScaledTPM
- dtuScaledTPM

Counts have information about precision, offset corrects if the counts are biased by *effective* transcript or gene length

These scale up TPMs to the total number of mapped reads, so are count-scale data with some information about precision. Because generated from TPM, countsFromAbundance *obviate* the need for an offset

tximport with countsFromAbundance

tximport has multiple methods for importing counts from *Salmon*, etc.

- counts + length offset
- scaledTPM
- lengthScaledTPM
- dtuScaledTPM

Counts				Length				Abundance			
30	30	30	30	3	3	3	3	1	1	1	1
20	20	20	20	2	2	2	2	1	1	1	1
0	0	0	0	2	2	2	2	0	0	0	0
50	50	50	50	1	1	1	1	5	5	5	5

$$\Sigma = 100$$

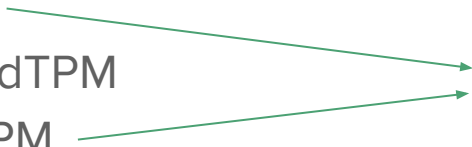
These latter two, similar to scaledTPM, but first multiply by the transcript length (ave over samples), or the median transcript length within gene. Therefore very close to counts, without need for offset

14.3	14.3	14.3	14.3
14.3	14.3	14.3	14.3
0.0	0.0	0.0	0.0
71.4	71.4	71.4	71.4

tximport with countsFromAbundance

tximport has multiple methods for importing counts from *Salmon*, etc.

- counts + length offset
- scaledTPM
- lengthScaledTPM
- dtuScaledTPM

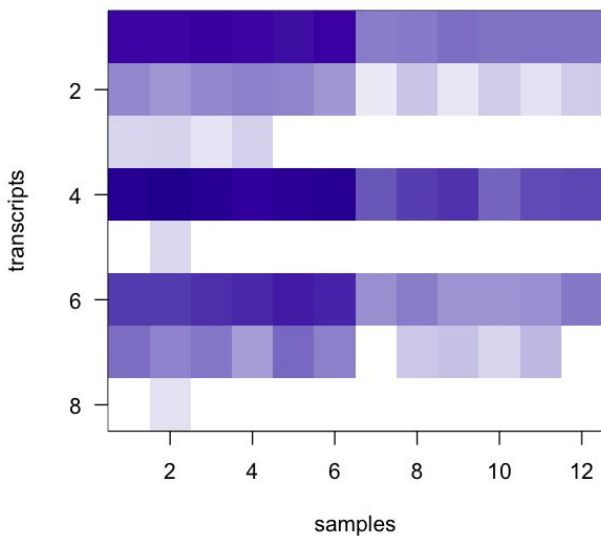


These two produce count-scale data such that counts within a gene are directly proportional to TPM ratio.

This is good for performing DTU analysis

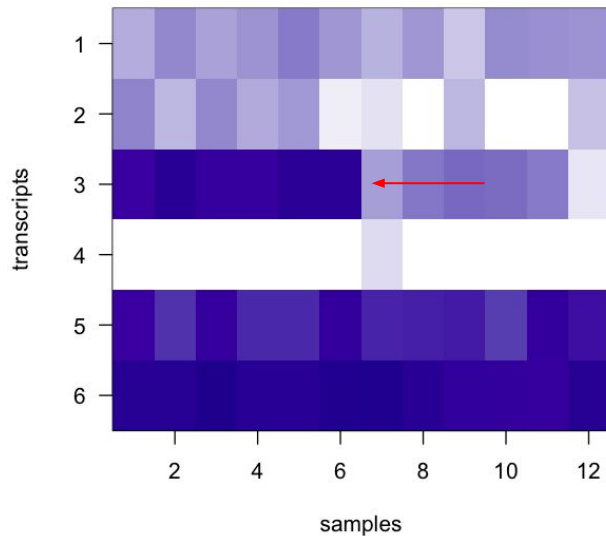
Explore different types of genes (simulated counts)

DGE



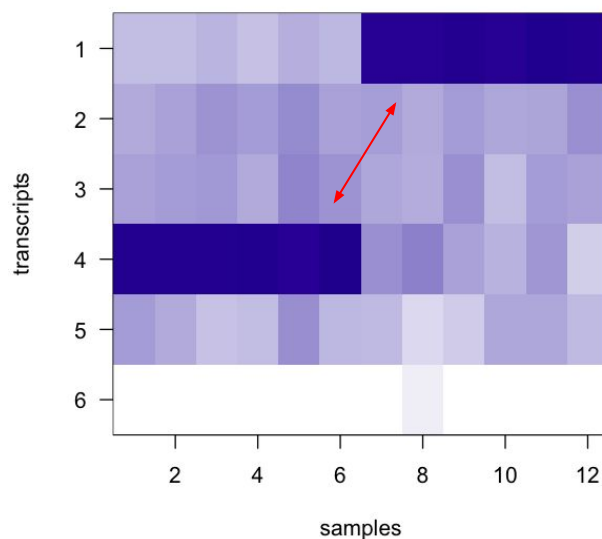
Differential gene expression:
Total expression changes

DTE



Differential transcript expression

DTU

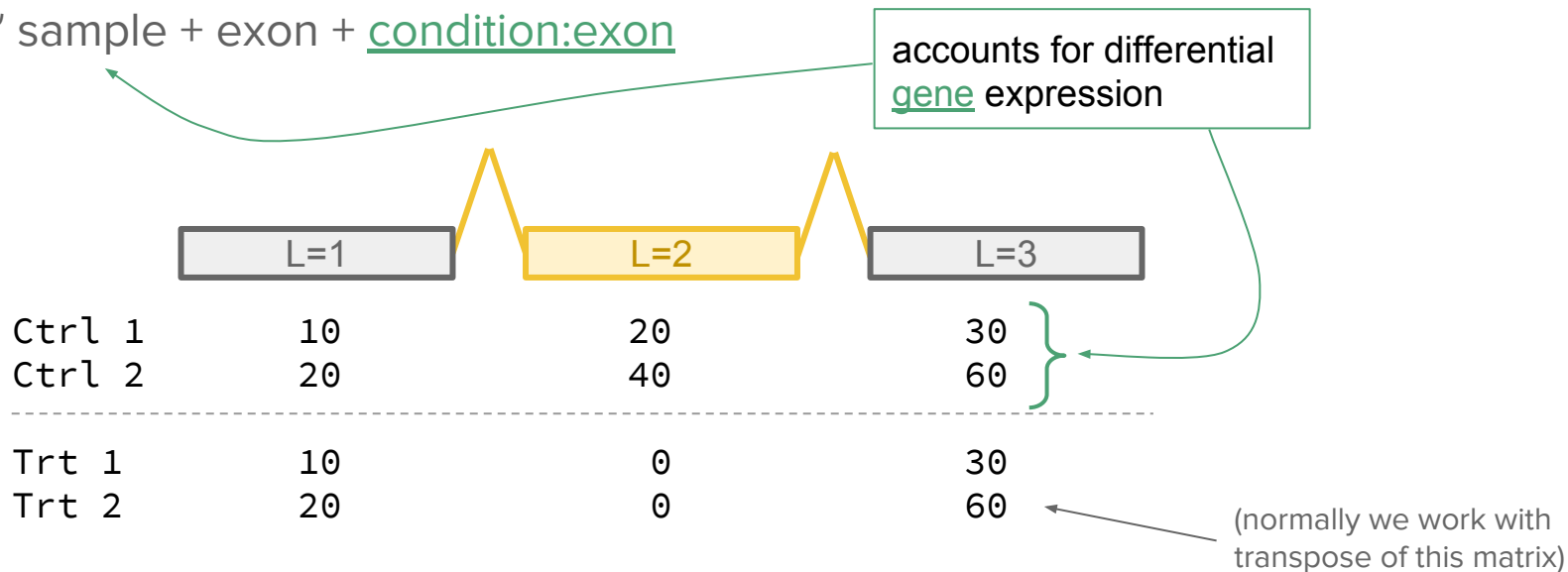


Differential transcript usage:
Switching isoforms, the total
expression may not change

DEXSeq (Anders 2012) with exon counts

Per exon, make a table with exon = “this” or “other”

design = ~ sample + exon + condition:exon

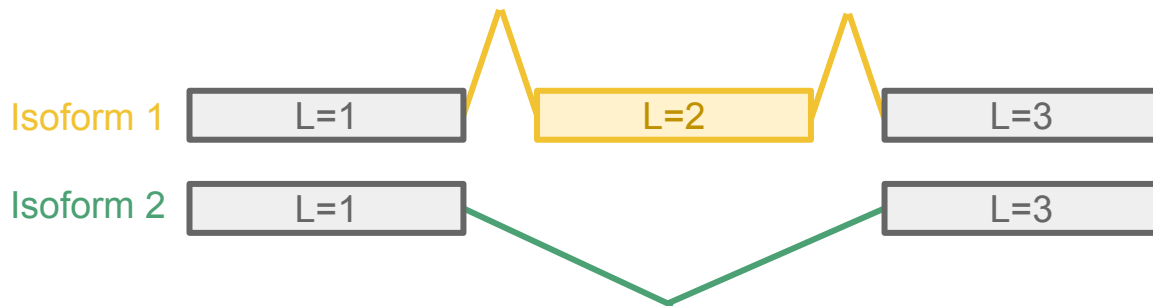


- DEXSeq will model the counts per feature as independent Negative Binomial counts (as in DESeq2)

DEXSeq with *transcript* counts

~ sample + exon + condition:exon

We still call them “exon”
when running DEXSeq



	Iso. 1	Iso. 2
Ctrl 1	60	0
Ctrl 2	120	0
<hr/>		
Trt 1	0	40
Trt 2	0	80

(normally we work with
transpose of this matrix)

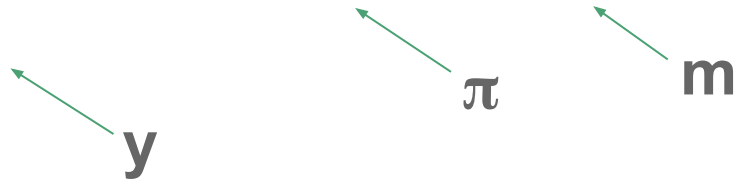
DRIMSeq (Nowicka 2016)

- DRIMSeq is a Dirichlet-Multinomial model for transcript counts
- What are these?
- Dirichlet - a probability distribution on vector of proportions,
[a, b, c] : $a + b + c = 1$
- Multinomial - a probability distribution on vector of counts with fixed total

E.g.:

```
> rmultinom(1, prob=c(.5,.3,.2), size=100)
```

52 28 20



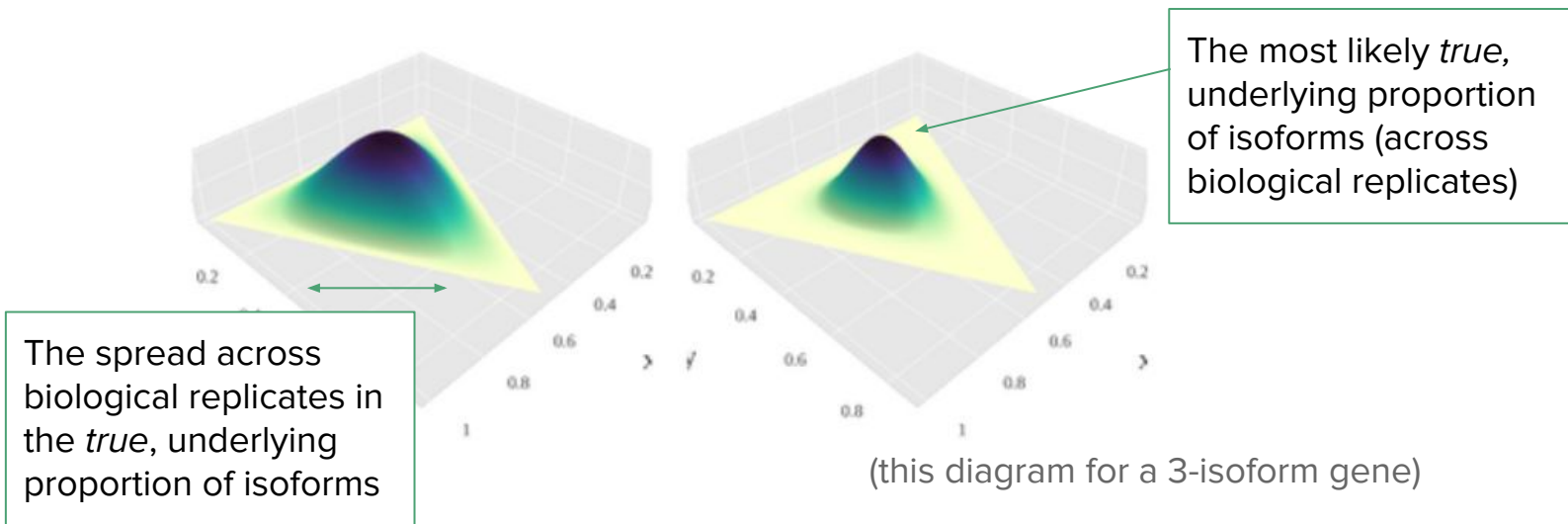
$$\mathbf{y} = (y_1, \dots, y_n)$$

$$m = \sum_{j=1}^q y_j$$

$$\boldsymbol{\pi} = (\pi_1, \dots, \pi_q),$$

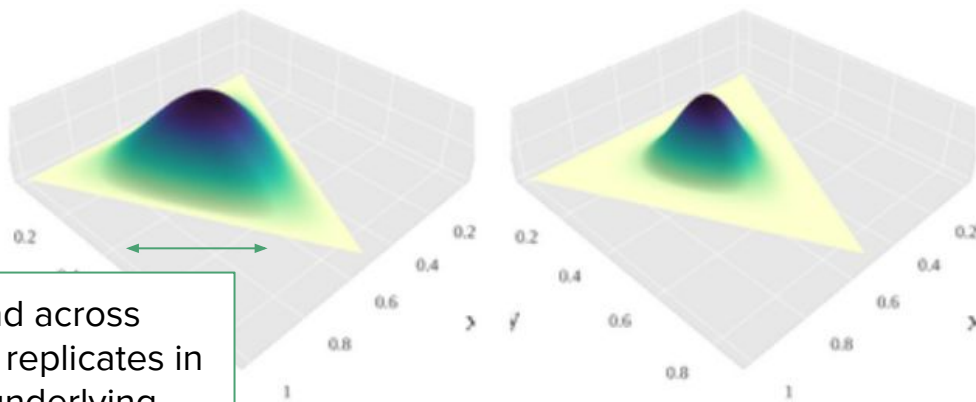
DRIMSeq: biological variation

- “To account for overdispersion due to true biological variation between experimental units as well as technical variation ..., we assume the feature proportions, π , follow the Dirichlet distribution”



DRIMSeq estimation of dispersion

- To estimate the precision / dispersion parameter for each gene, [DRIMSeq](#) uses similar technique to [edgeR](#), but Dirichlet-Multinomial rather than the Negative Binomial: Cox-Reid adjusted profile likelihood + shrinkage



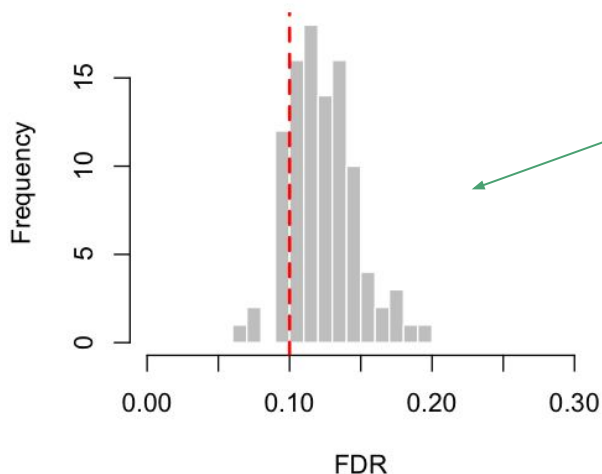
The spread across biological replicates in the *true*, underlying proportion of isoforms

DRIMSeq and DEXSeq differences in count model

- Unlike DEXSeq, DRIMSeq models correlation of the counts among the transcripts within a gene – if one goes up, other goes down
- In the current implementation, DRIMSeq models a single precision parameter per gene (related to the Multinomial dispersion), whereas DEXSeq models a different dispersion parameter for each feature (here, transcript)

Stage-wise testing framework (Van den Berge 2017)

- Adapted from Heller (2009) for gene set testing
- Idea: we can gain power by first screening at gene-level,
- Then later confirming which transcripts are significant
- What if we just screen and confirm without a framework?



Here we simulate 2000 genes with 10 isoforms each.

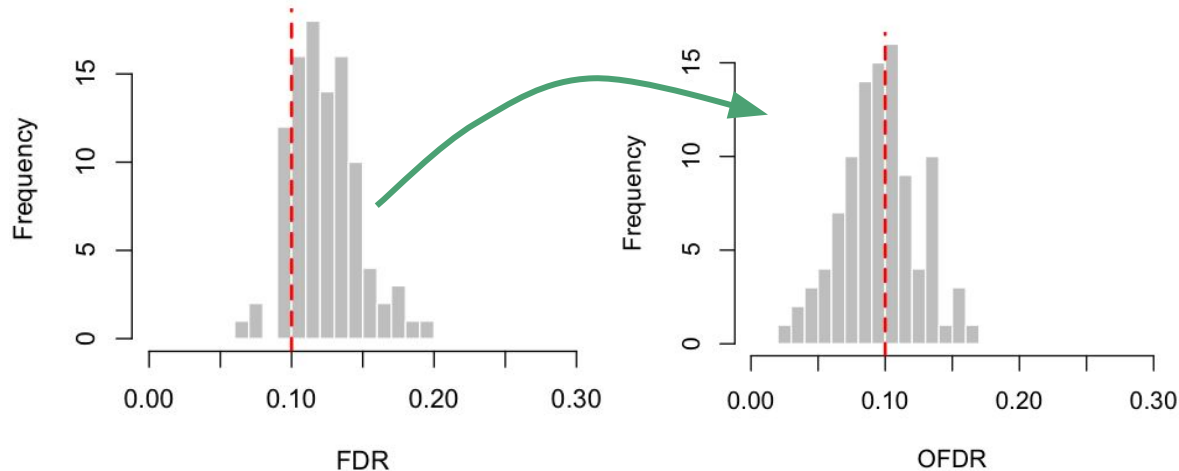
Insert 100 genes w/ two DTU isoforms each

Screen the genes, then compute adjusted p-values for transcripts. Assess the txp adjusted p-values.

We tend to exceed the target FDR in the confirmation set, because of txps from null genes coming along.

stageR for stage-wise testing

- Screening and confirmation stages, controls “overall false discovery rate”
- OFDR of 5% = expect no more than 5% of genes that pass screening will:
 - not contain any DTU, so be falsely screened genes, or
 - contain a transcript with a transcript adj. $p < 0.05$ which is falsely confirmed
- The OFDR is on the unit of genes



stageR for stage-wise testing

- How does it work?
- Take as input or apply BH adjustment to screening p-values (**G** genes)
- Screen at α_1 producing **R** rejections
- For genes passing screening, define $\alpha_2 = \mathbf{R} \alpha_1 / \mathbf{G}$
- Apply multiple testing* within each gene to provide FWER of α_2

Example: 1000 genes, screen at $\alpha_1=0.05$ producing 20 rejections.

Apply multiple testing within each gene to provide FWER of 0.001

* “According to the Shaffer method, the two most significant transcripts can be tested at a significance level of $\alpha_2 / (n_g - 2)$, and from the third most significant transcript onwards the procedure reduces to the Holm method.” -stageR paper

Acknowledgments

- Workflow coauthors: Charlotte Soneson and Rob Patro
- DRIMSeq: Malgorzata Nowicka and Mark D. Robinson
- DEXSeq: Simon Anders, Alejandro Reyes, and Wolfgang Huber
- stageR: Koen Van den Berge, Charlotte Soneson,
Mark D. Robinson, and Lieven Clement