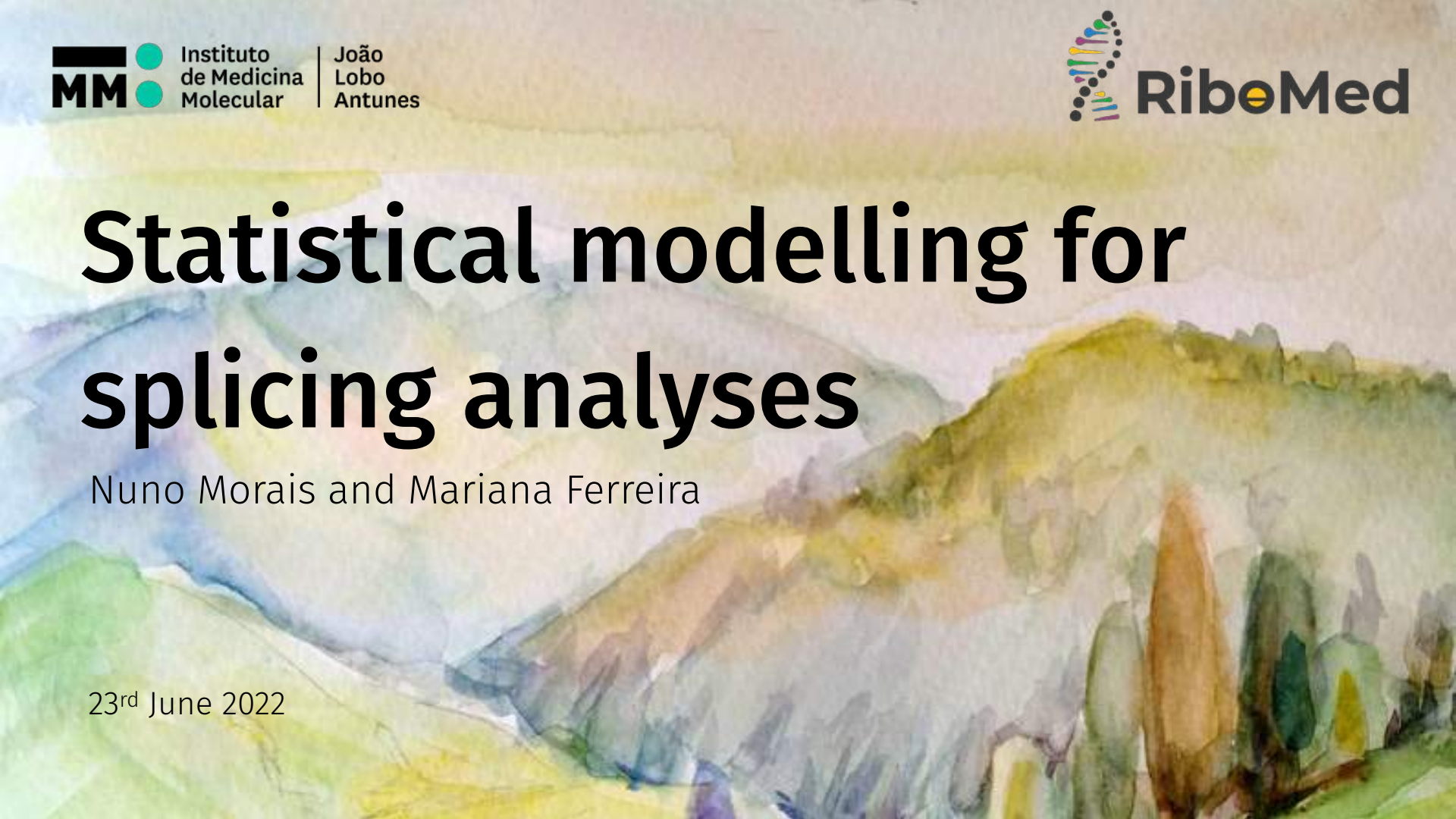


Instituto de Medicina Molecular | João Lobo Antunes

RiboMed

Statistical modelling for splicing analyses

Nuno Morais and Mariana Ferreira

23rd June 2022

Agenda

- **Session 1:** Quantification of alternative splicing inclusion levels from RNA sequencing data
  - Introduction of the speakers and their research interests
  - Our previous work on alternative splicing evolution
  - Estimating alternative sequence inclusion levels from RNA-sequencing data:
    - Sources of uncertainty in individual samples
    - Similarities and differences from gene expression quantification
    - Different approaches and their interpretations

    - The (mathematical) nature of the percent spliced-in (PSI) ratio
      - Constrained range, intrinsically normalised, logit transformation
      - Relationship with the beta distribution
    - Differential splicing with many samples (*psichomics*)

- **Session 2:** Differential splicing analyses
  - Modelling the estimation uncertainty in individual samples while accounting for variability among replicates using beta distributions
  - Other differential splicing approaches (e.g., non-parametric tests)
  - Sum-up and assessment of learning outcomes

Introduction

Nuno Morais

(group leader)

Mariana Ferreira

(PhD student in Nuno's lab)

Overall learning goals

Identify and understand the sources of **uncertainty** when estimating the inclusion levels of alternative sequences from RNA sequencing data and how those impact differential splicing analyses.

Estimate the **effect size** of alternative splicing changes between conditions **and its statistical significance,** using beta distributions for the analysis and visualisation of inclusion levels of alternative sequences from an event-centered perspective.

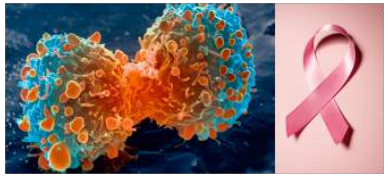
What we do: disease transcriptomics

our GOAL

Understand how ageing-associated molecular (RNA) changes in human tissues increase proneness to disease

(disease) models

Neurodegeneration



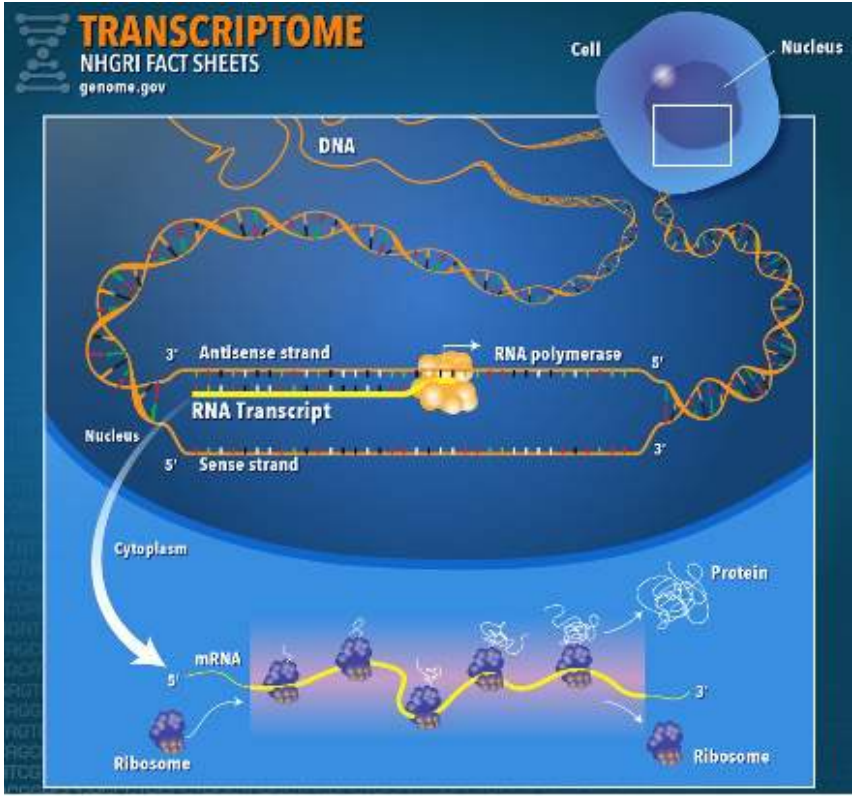
Cancer

Ageing

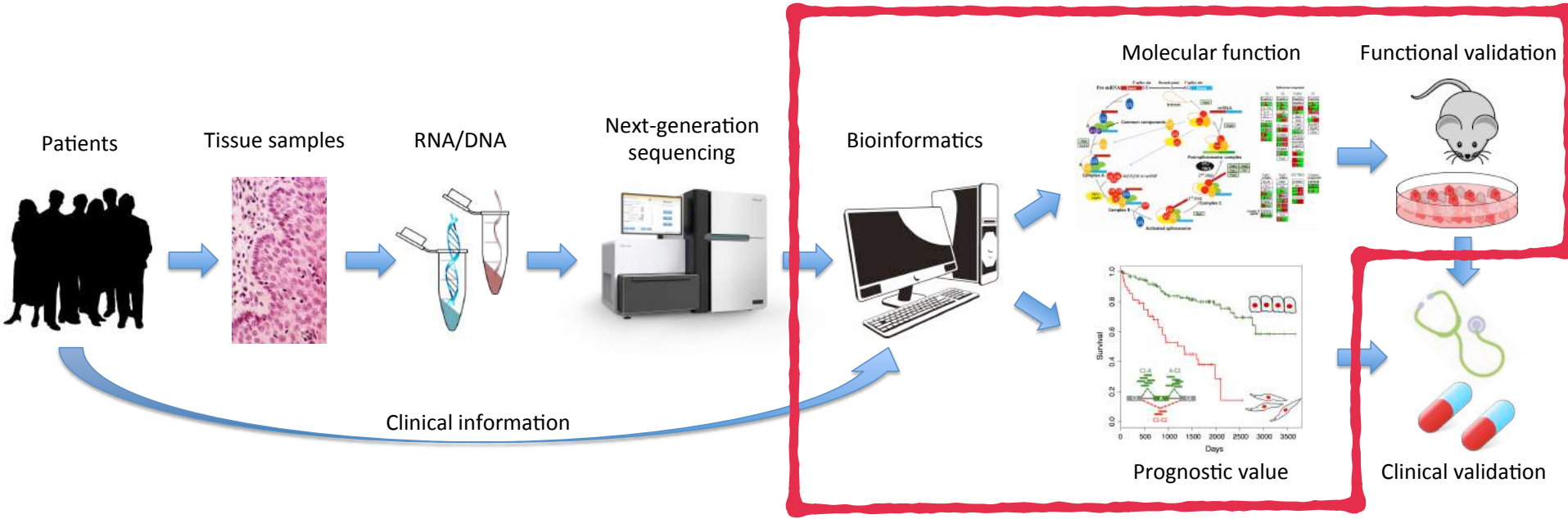
Why the transcriptome?

the set of all RNA molecules in one cell or a population of cells

- Early and accurately profileable measure of cells' response to stimuli
- Our expertise: **computational biology** + transcriptional regulation
- Timely: availability of (public) data

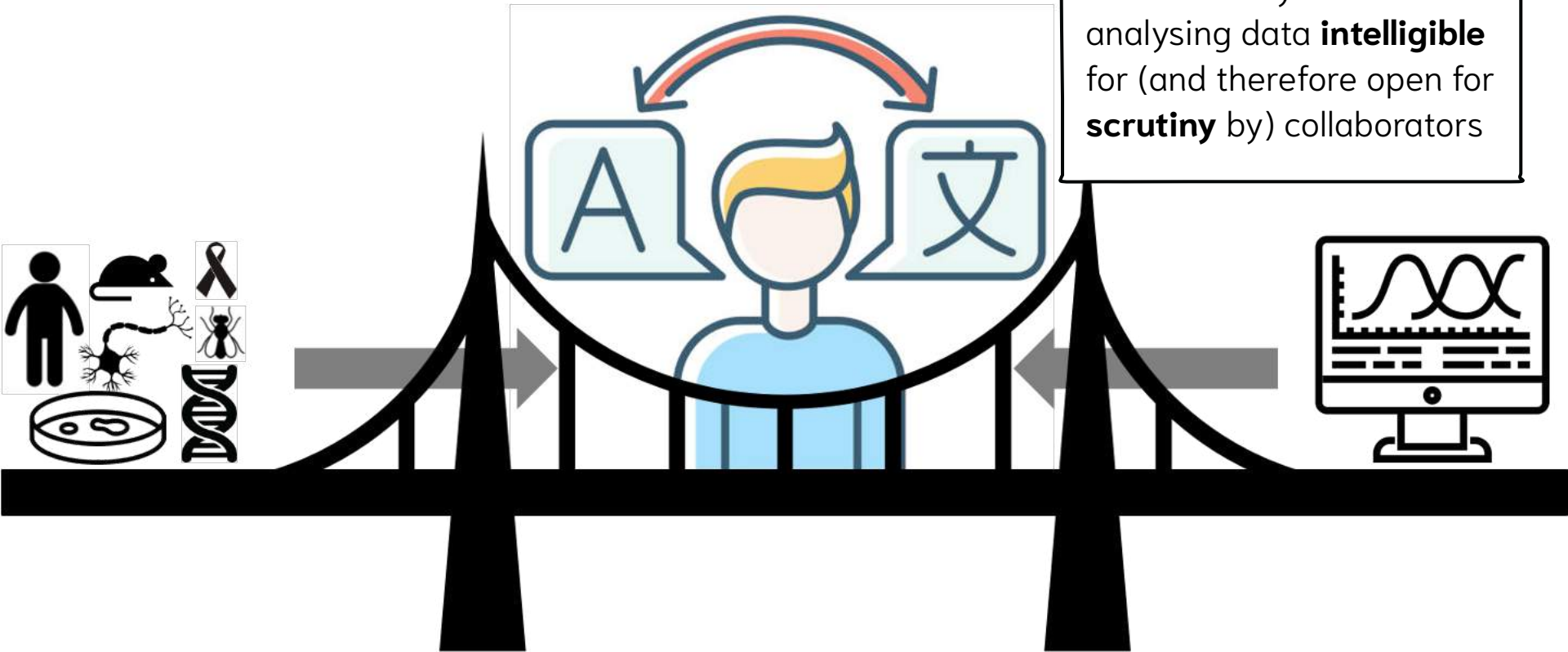


What we do: disease transcriptomics



What makes good computational biologists

Make our ways of analysing data **intelligible** for (and therefore open for **scrutiny** by) collaborators

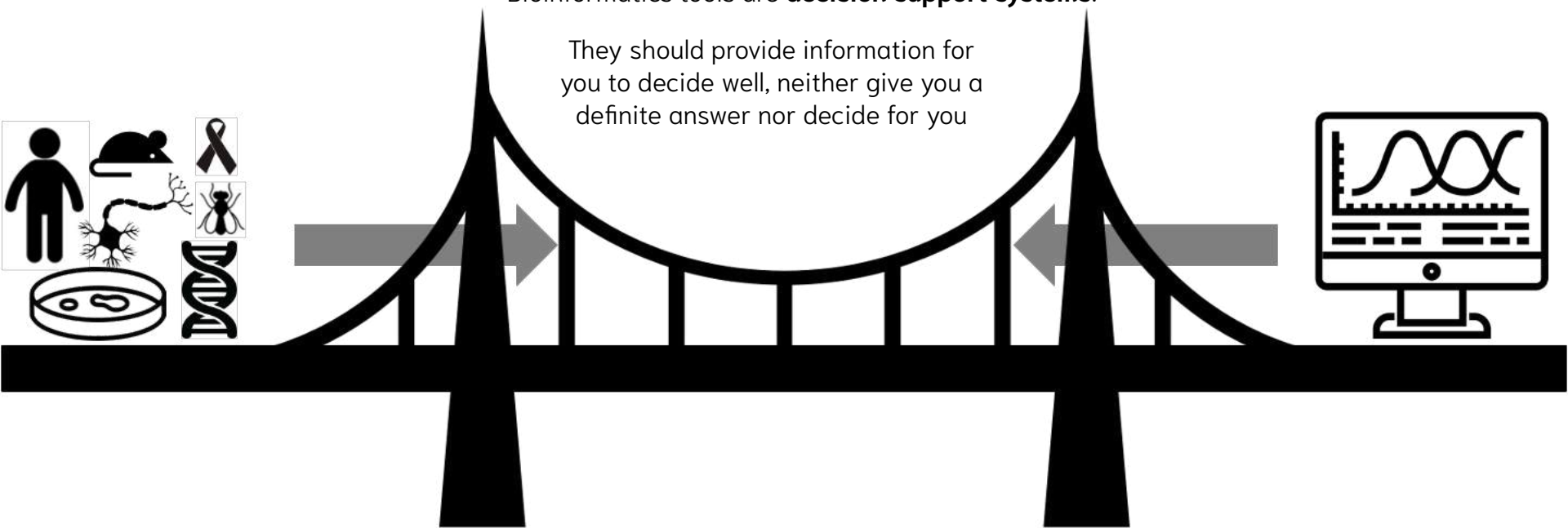


Our anti-black box efforts

Develop tools for empowering colleagues to perform similar analyses while **understanding decisions** needed to be made at their every stage

Bioinformatics tools are **decision support systems**!

They should provide information for you to decide well, neither give you a definite answer nor decide for you



Session 1

Quantification of alternative sequence inclusion levels from RNA sequencing data

Different complexity with comparable number of genes

6 000

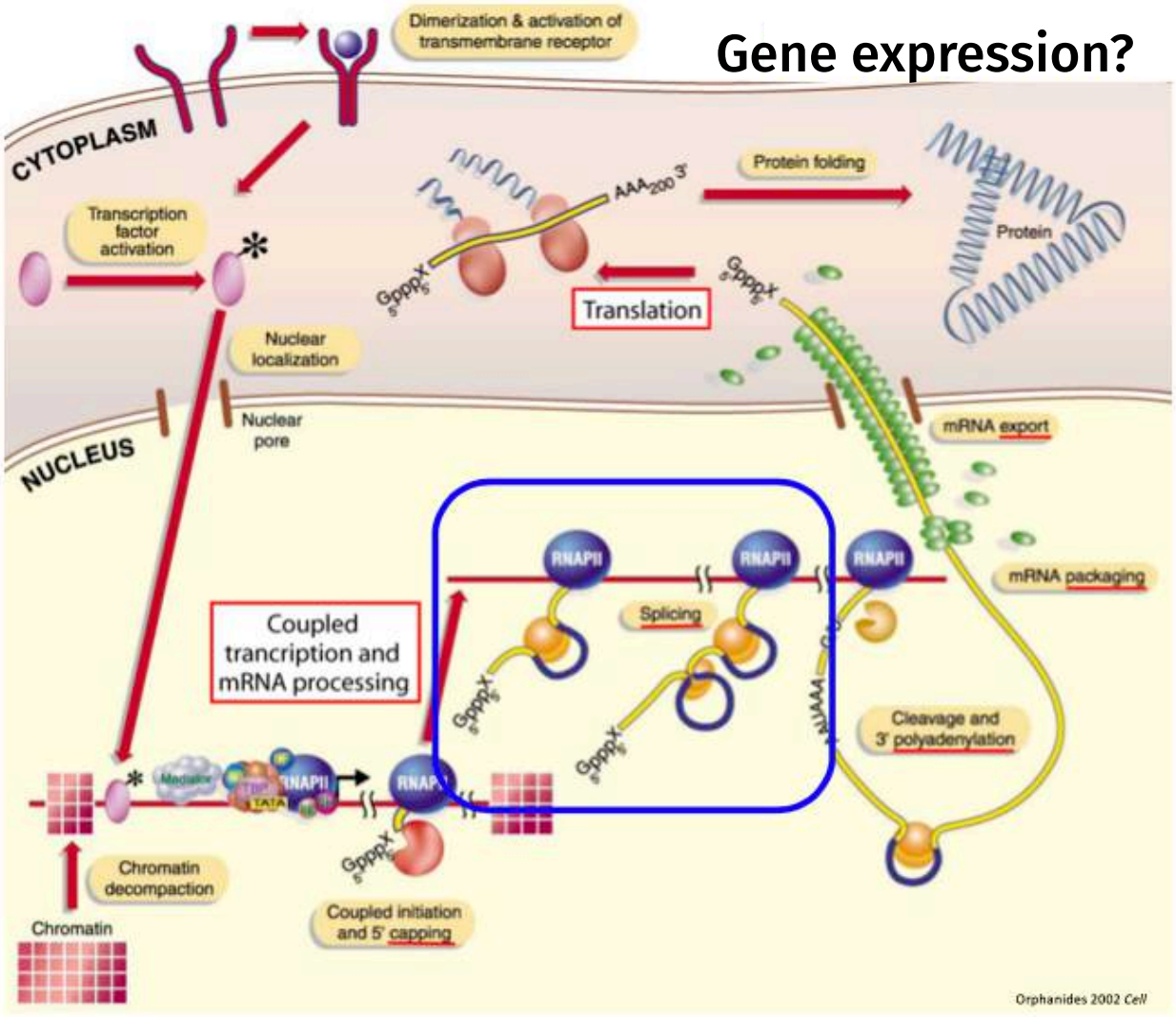
19 000

14 000

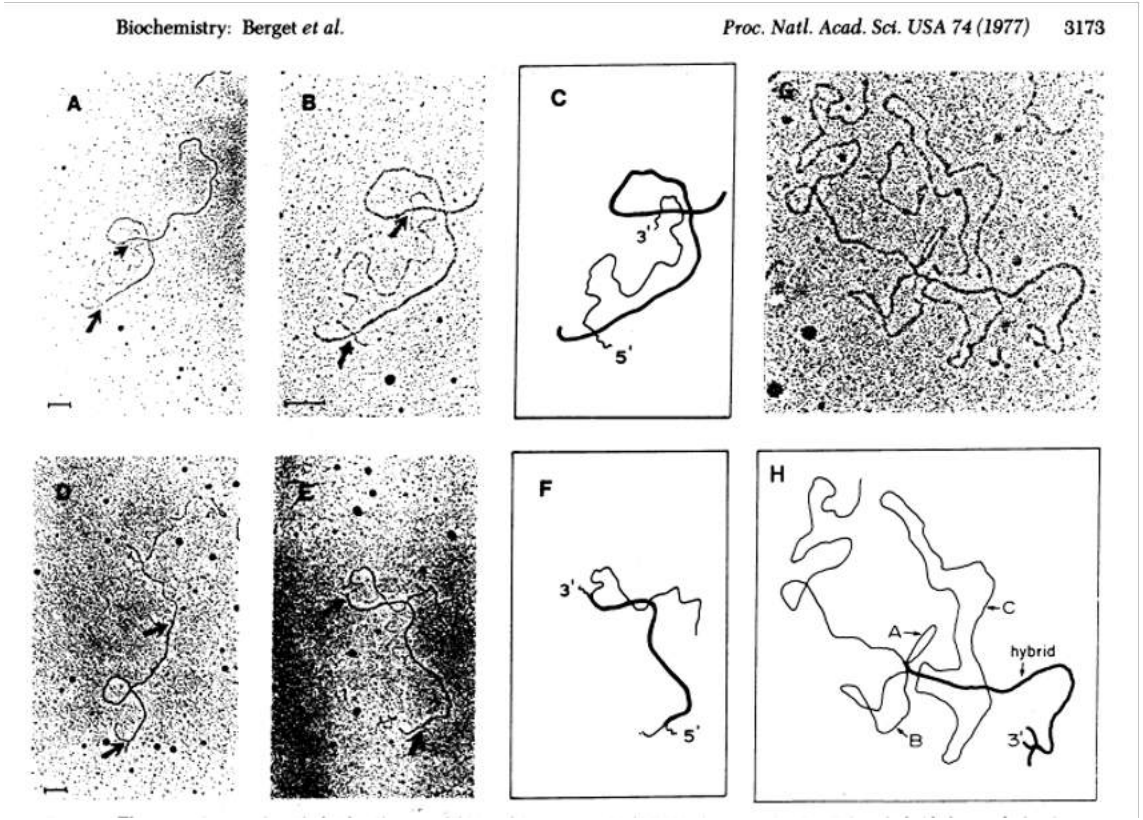
22 000

Number of **protein coding genes**

Gene expression?



mRNA splicing, the “discoveries of split genes”

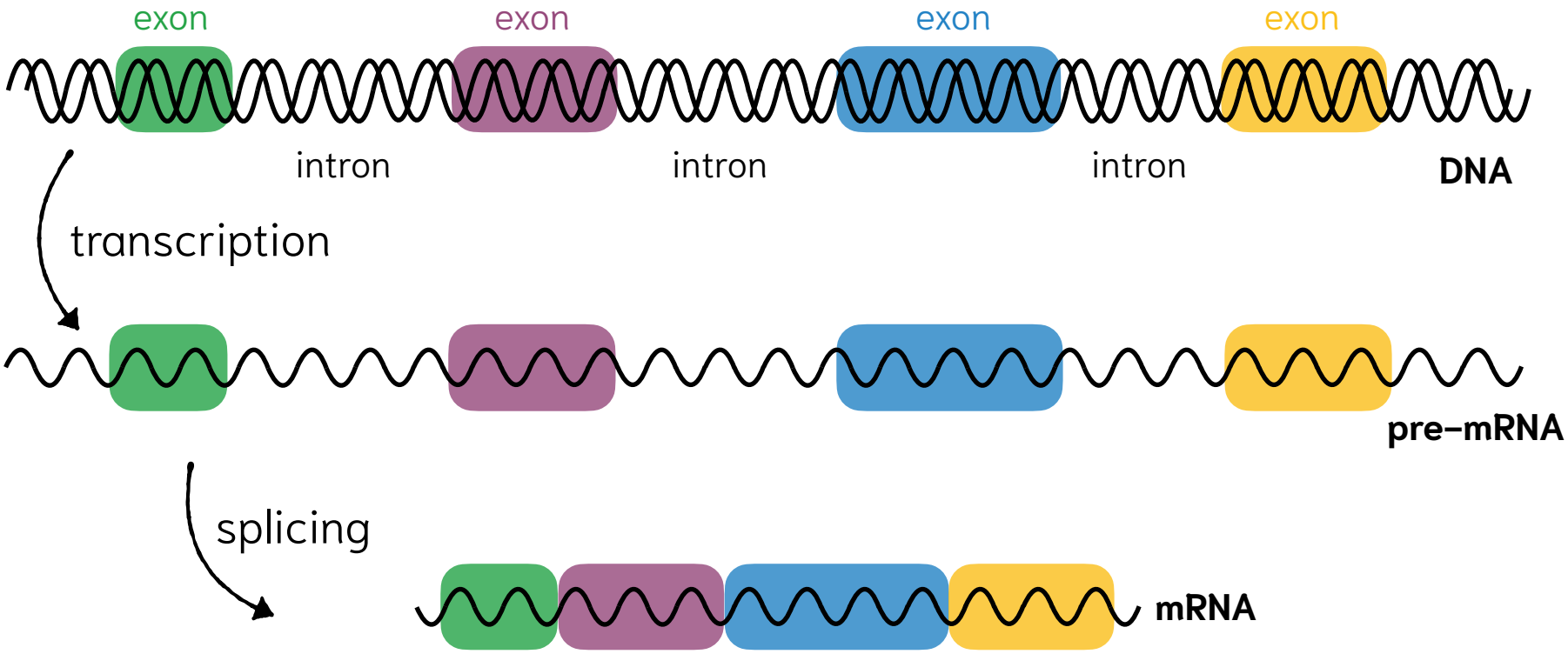


The Nobel Prize in Physiology or Medicine 1993

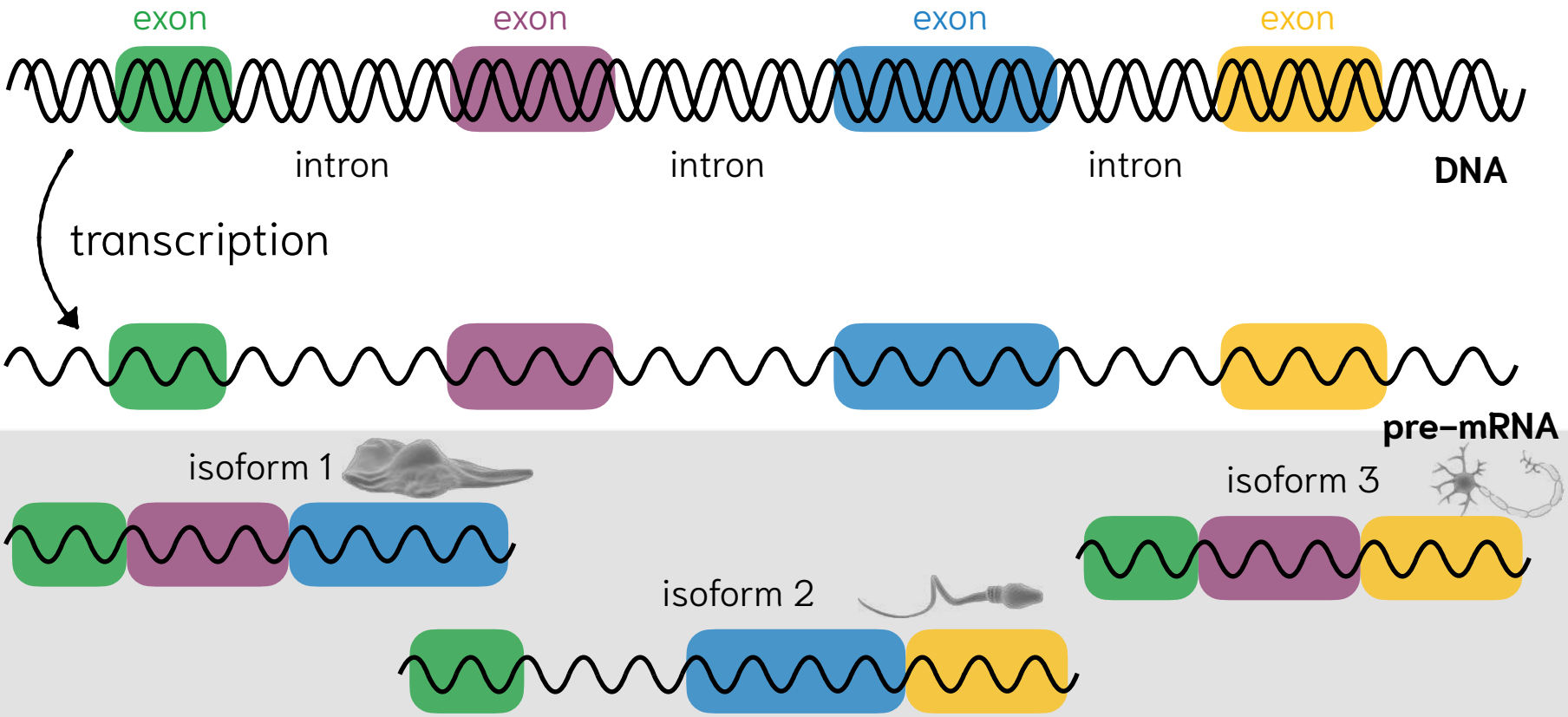
Richard J. Roberts

Phillip A. Sharp

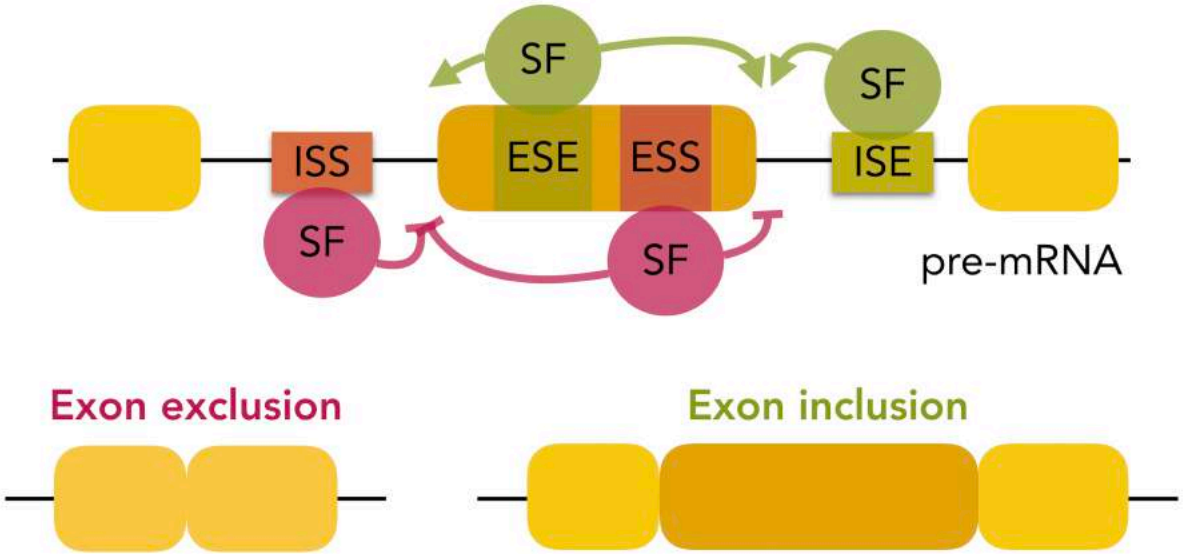
Alternative splicing functionally tunes transcriptomes



Alternative splicing functionally tunes transcriptomes



Mechanism of splicing and its regulation



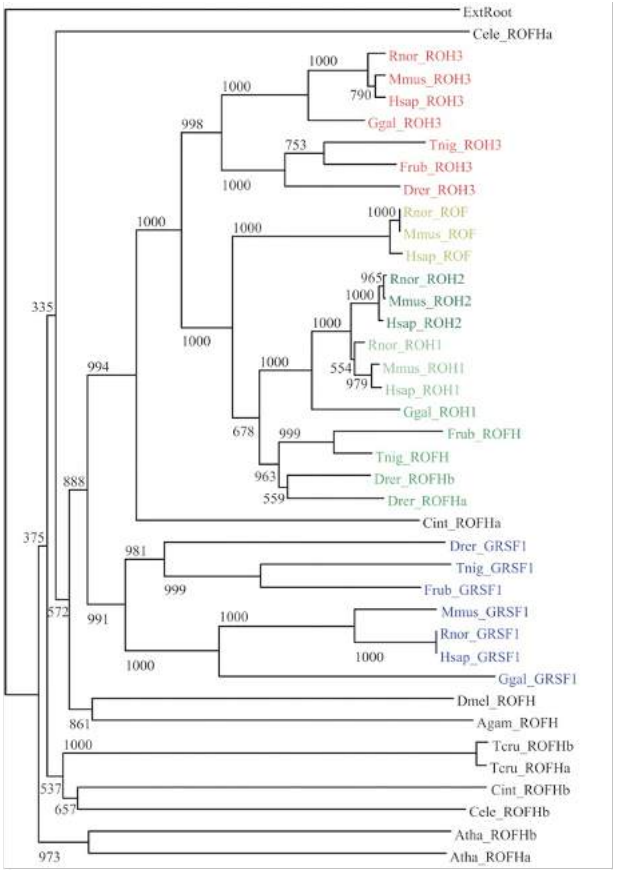
Gallego–Paez, L. M. et al. Alternative splicing: the pledge, the turn, and the prestige. Hum Genet 136, 1015–1042 (2017).

Alternative splicing parallels species complexity

- **Conservation** of the core spliceosomal proteins across eukaryotes

- Selective **expansions** of protein families known to **regulate alternative splicing** (e.g. SR proteins in metazoans, hnRNPs in vertebrates)

Controversy on how to best quantify AS abundance from EST data



Letter

Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion

Nuno L. Barbosa-Morais,[1,2] Maria Carmo-Fonseca,[2] and Samuel Aparício[1,3,4]

[1]University of Cambridge, Department of Oncology, Hutchison-MRC Research Centre, Cambridge CB2 2XZ, United Kingdom; [2]Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Av. Prof. Egas Moniz, 1649-028 Lisboa, Portugal

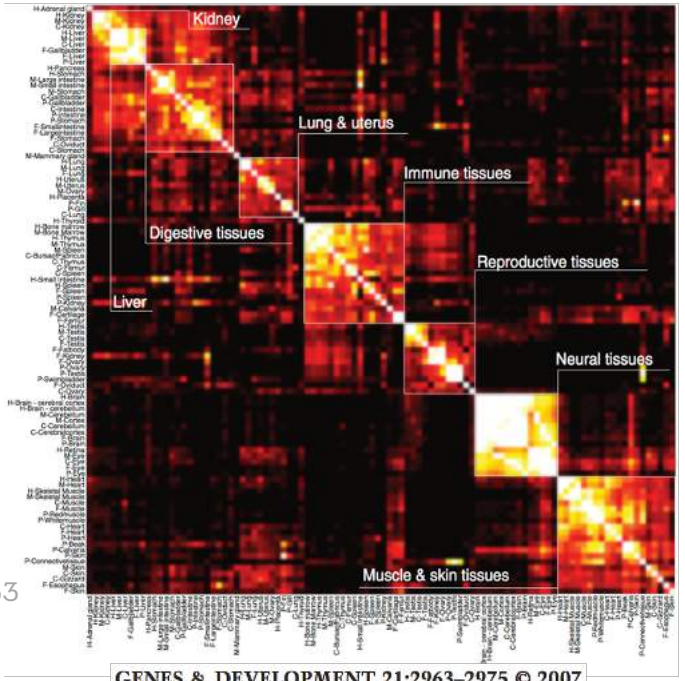
16:66–77 ©2006 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

Evolving vertebrate transcriptomes

Vertebrate species share similar repertoires of coding genes and their organ-specific expression is conserved

Ponting (2008) Nat Rev Genet 9:689

Changes in AS may represent a major source of species-specific differences



Chan et al (2009) J Biol 8:33

TRENDS in Genetics Vol.21 No.2 February 2005

Alternative splicing of conserved exons is frequently species-specific in human and mouse

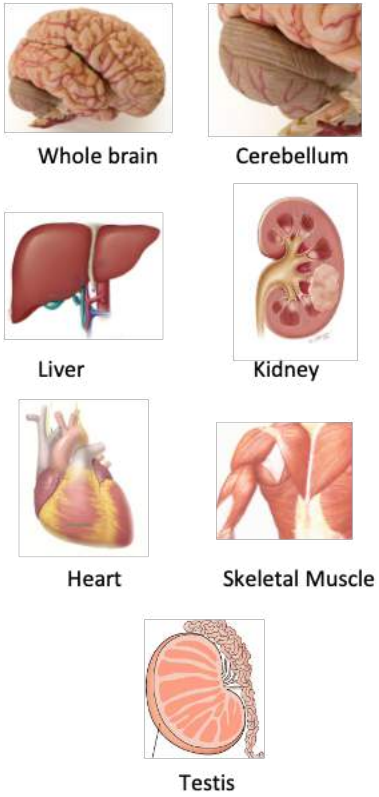
Qun Pan[1], Malina A. Bakowski[1], Quaid Morris[1,2], Wen Zhang[1], Brendan J. Frey[2], Timothy R. Hughes[1] and Benjamin J. Blencowe[1]

GENES & DEVELOPMENT 21:2963–2975 © 2007

Global analysis of alternative splicing differences between humans and chimpanzees

John A. Calarco,[1,2,8] Yi Xing,[3,4,8] Mario Cáceres,[5,6,8] Joseph P. Calarco,[1] Xinshu Xiao,[7] Qun Pan,[1] Christopher Lee,[5] Todd M. Preuss,[5,10] and Benjamin J. Blencowe[1,2,9]

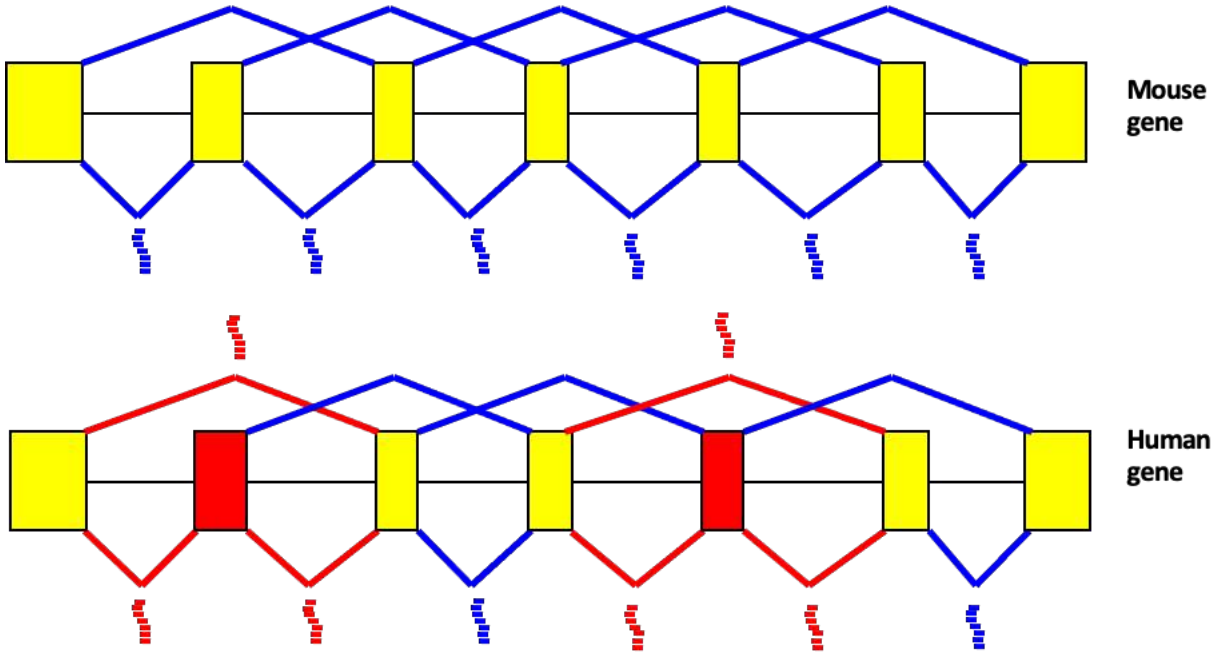
Data: RNA-Seq from physiologically-equivalent organs from vertebrate species spanning 350 Mys of evolution



Claudia Kutter, Stephen Watt (Odom Lab, Cambridge)

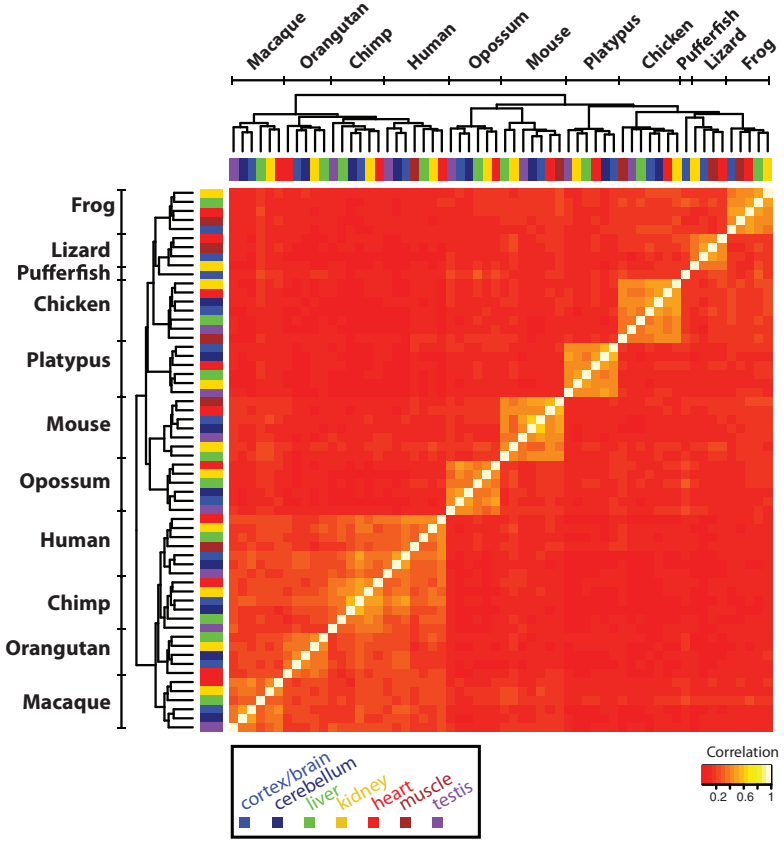
Brawand et al (2011) Nature 478:343

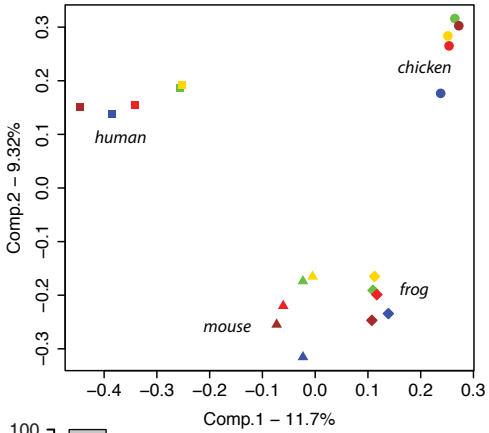
Comparing AS between species using RNA-Seq

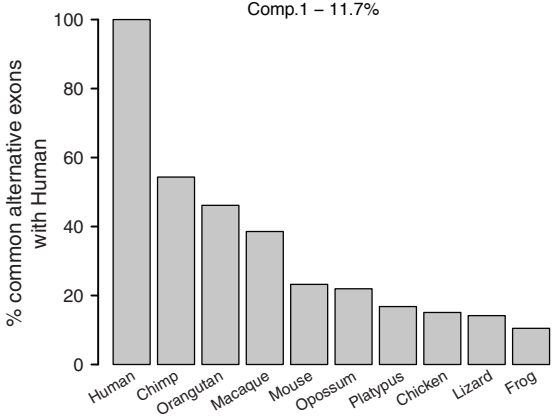


Exon orthology determined with Galaxy Lift-Overs:
https://main.g2.bx.psu.edu/tool_runner?tool_id=liftOver1

Rapid evolution of organ-specific alternative splicing

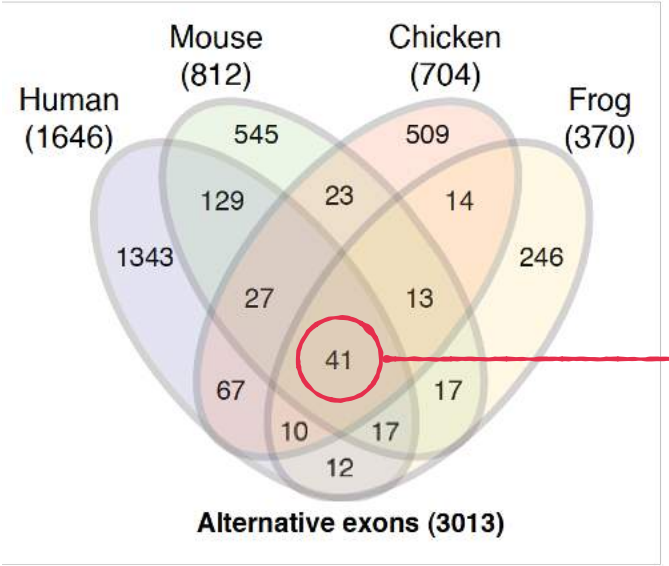


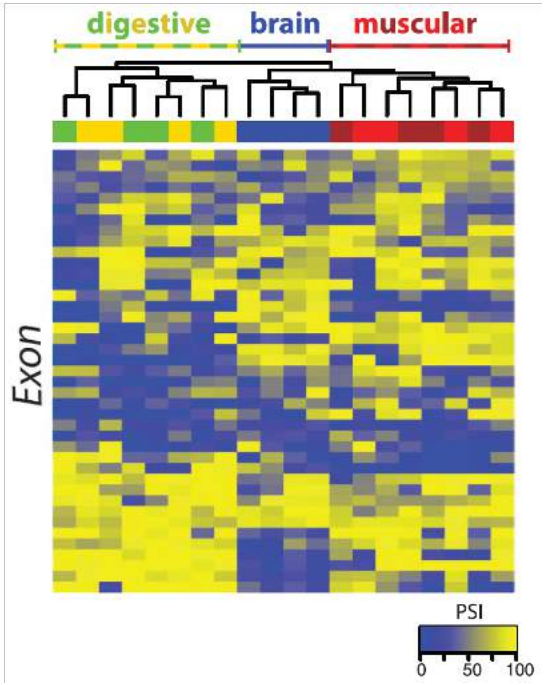




Barbosa-Morais et al (2012) Science 338:1587

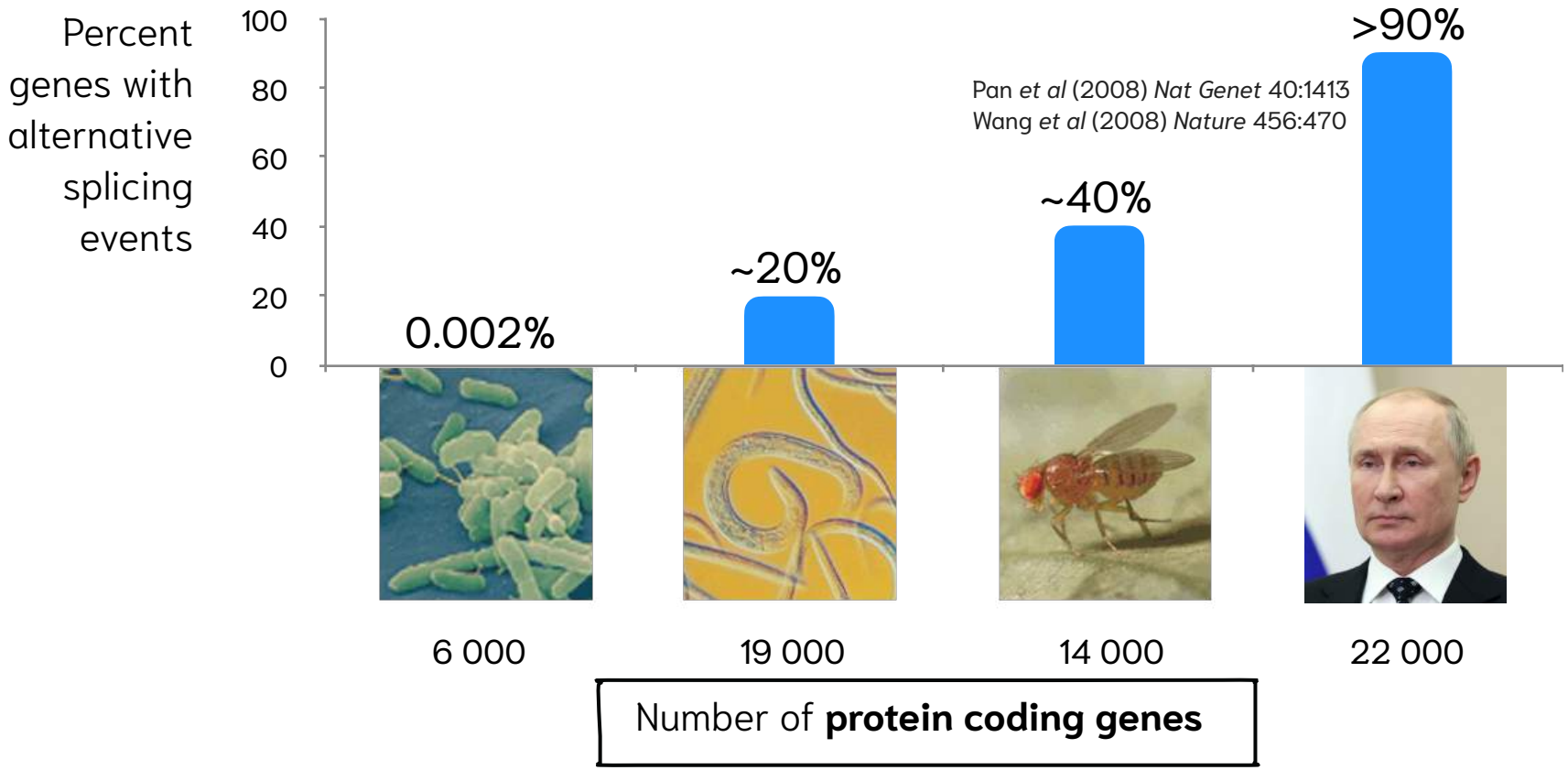
Conserved alternative splicing: organ-specific and highly regulated



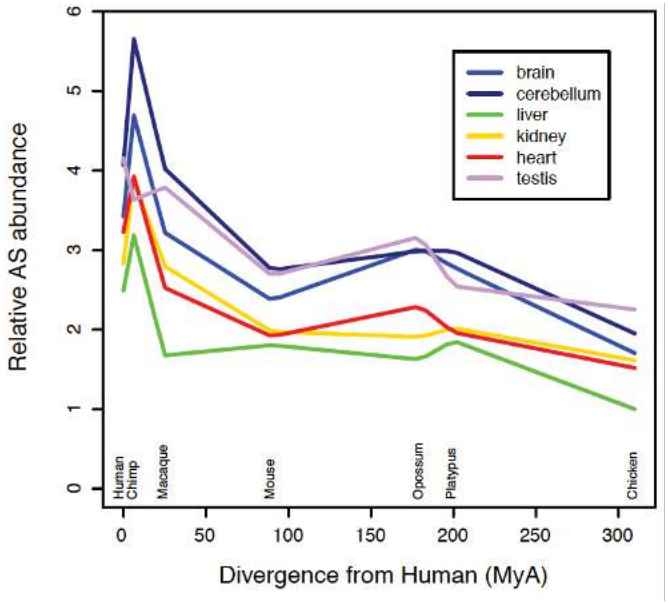


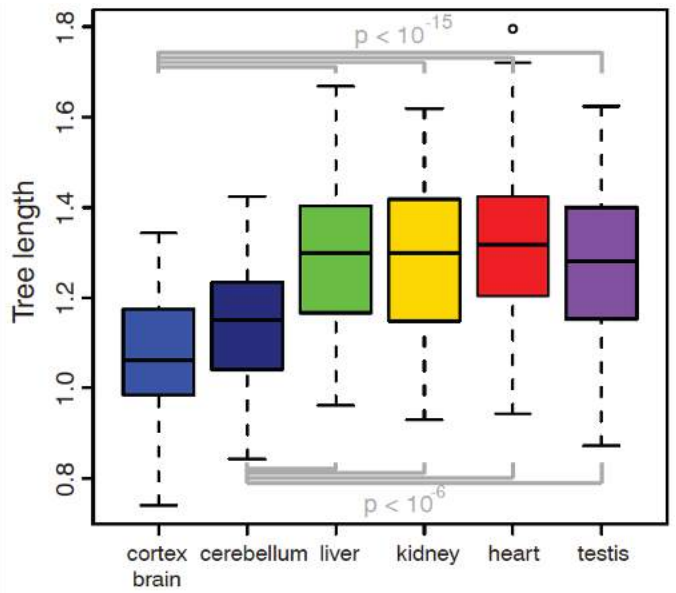
Barbosa-Morais et al (2012) Science 338:1587

Alternative splicing parallels species complexity



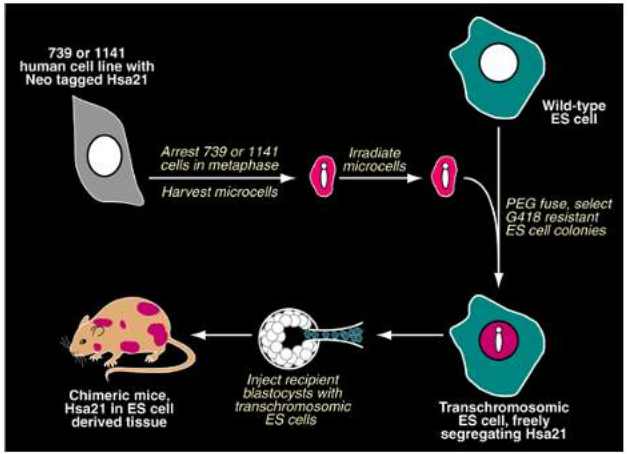
AS more abundant and more conserved in the brain





Barbosa-Morais et al (2012) Science 338:1587

The Tc1 Mouse

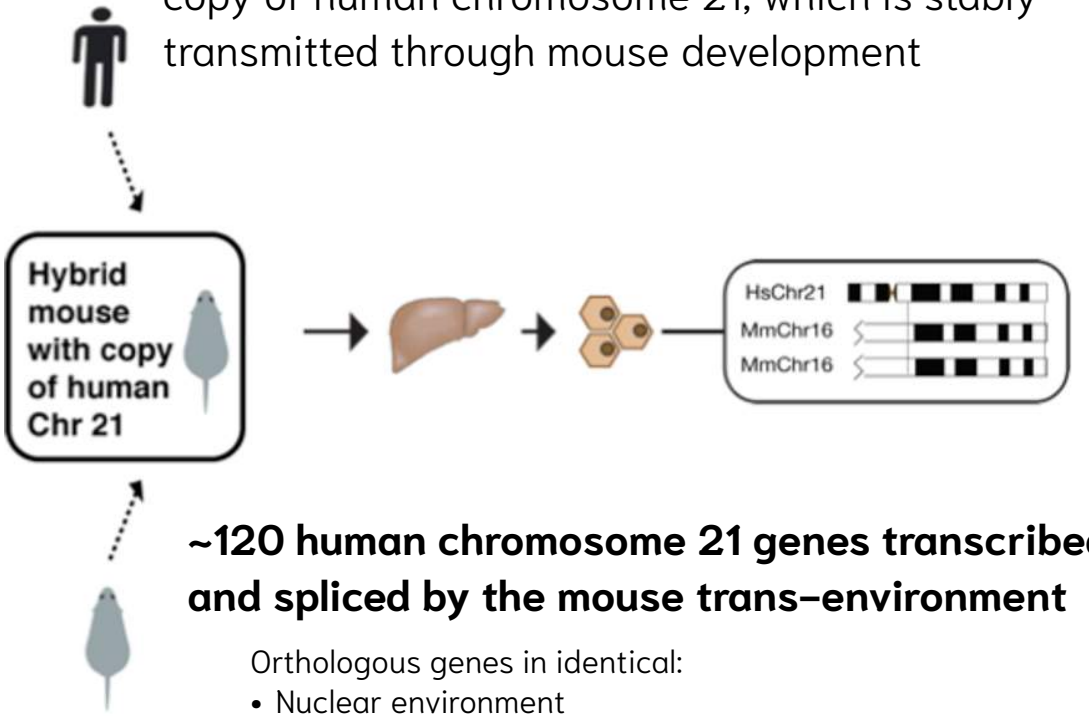


SCIENCE VOL 309 23 SEPTEMBER 2005 2033

An Aneuploid Mouse Strain Carrying Human Chromosome 21 with Down Syndrome Phenotypes

Aideen O'Doherty,[1,3] Sandra Ruf,[1,3] Claire Mulligan,[4] Victoria Hildreth,[5] Mick L. Errington,[3] Sam Cooke,[3] Abdul Sesay,[3] Sonie Modino,[6] Lesley Vanes,[3] Diana Hernandez,[1,3] Jacqueline M. Linehan,[1,2] Paul T. Sharpe,[6] Sebastian Brandner,[1] Timothy V. P. Bliss,[3] Deborah J. Henderson,[5] Dean Nizetic,[4] Victor L. J. Tybulewicz,[3*] Elizabeth M. C. Fisher[1*]

Aneuploid strain that carries a nearly complete copy of human chromosome 21, which is stably transmitted through mouse development

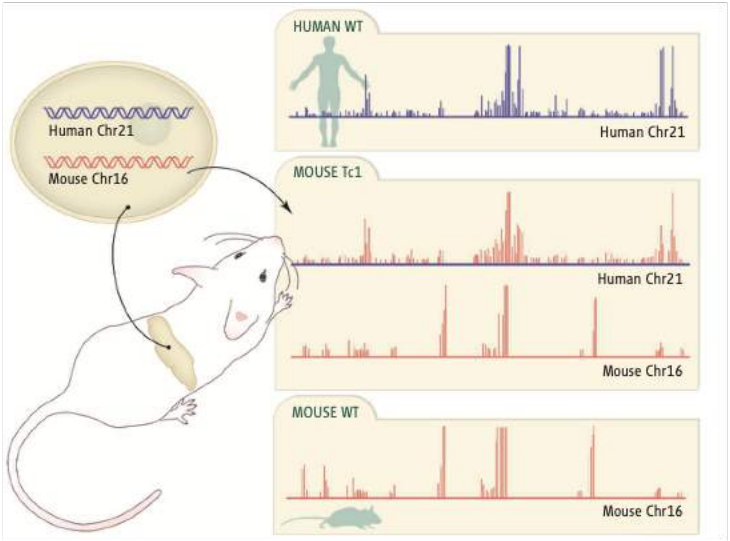


~120 human chromosome 21 genes transcribed and spliced by the mouse trans-environment

Orthologous genes in identical:

- Nuclear environment
- Developmental remodeling
- Concentration of transcription and splicing factors

Alternative splicing functionally tunes transcriptomes



17 OCTOBER 2008 VOL 322 **SCIENCE** www.sciencemag.org

Species-Specific Transcription in Mice Carrying Human Chromosome 21

Michael D. Wilson,[1]* Nuno L. Barbosa-Morais,[1,2]* Dominic Schmidt,[1,2] Caitlin M. Conboy,[3] Lesley Vanes,[4] Victor L. J. Tybulewicz,[4] Elizabeth M. C. Fisher,[5] Simon Tavaré,[1,2,6] Duncan T. Odom[1,2]†

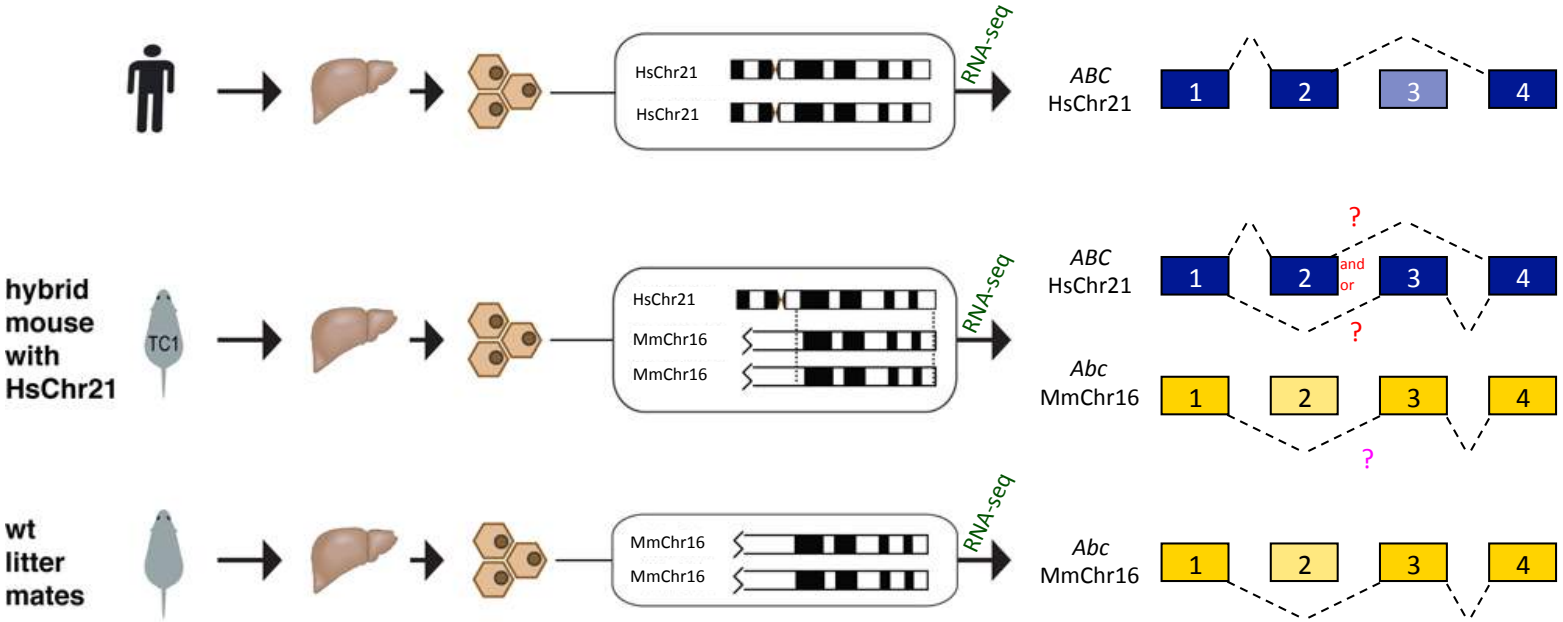
PERSPECTIVES

GENETICS

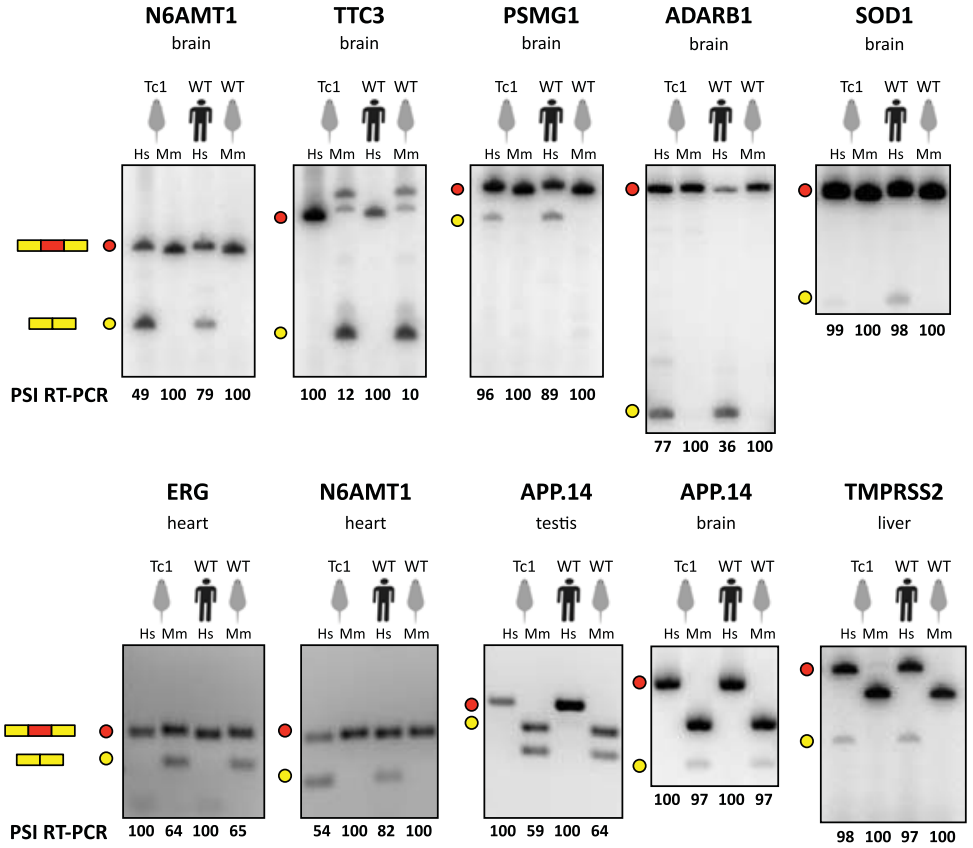
It's the Sequence, Stupid!

Hilary A. Coller[1] and Leonid Kruglyak[2]

Interspecies differences in alternative splicing: *cis* or *trans*?



Changes in *cis*-regulatory elements sufficient to direct the majority of the species-specific AS patterns



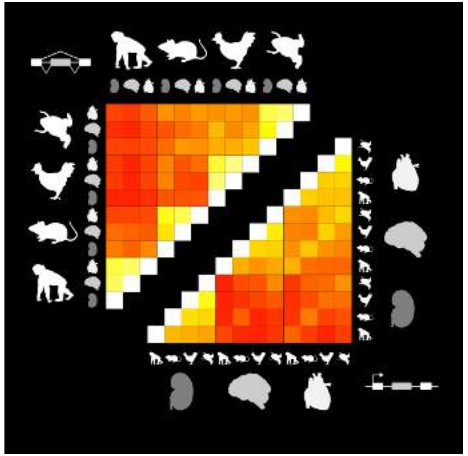
Barbosa-Morais et al (2012) Science 338:1587

Alternative splicing functionally tunes transcriptomes

www.sciencemag.org **SCIENCE** VOL 338 21 DECEMBER 2012

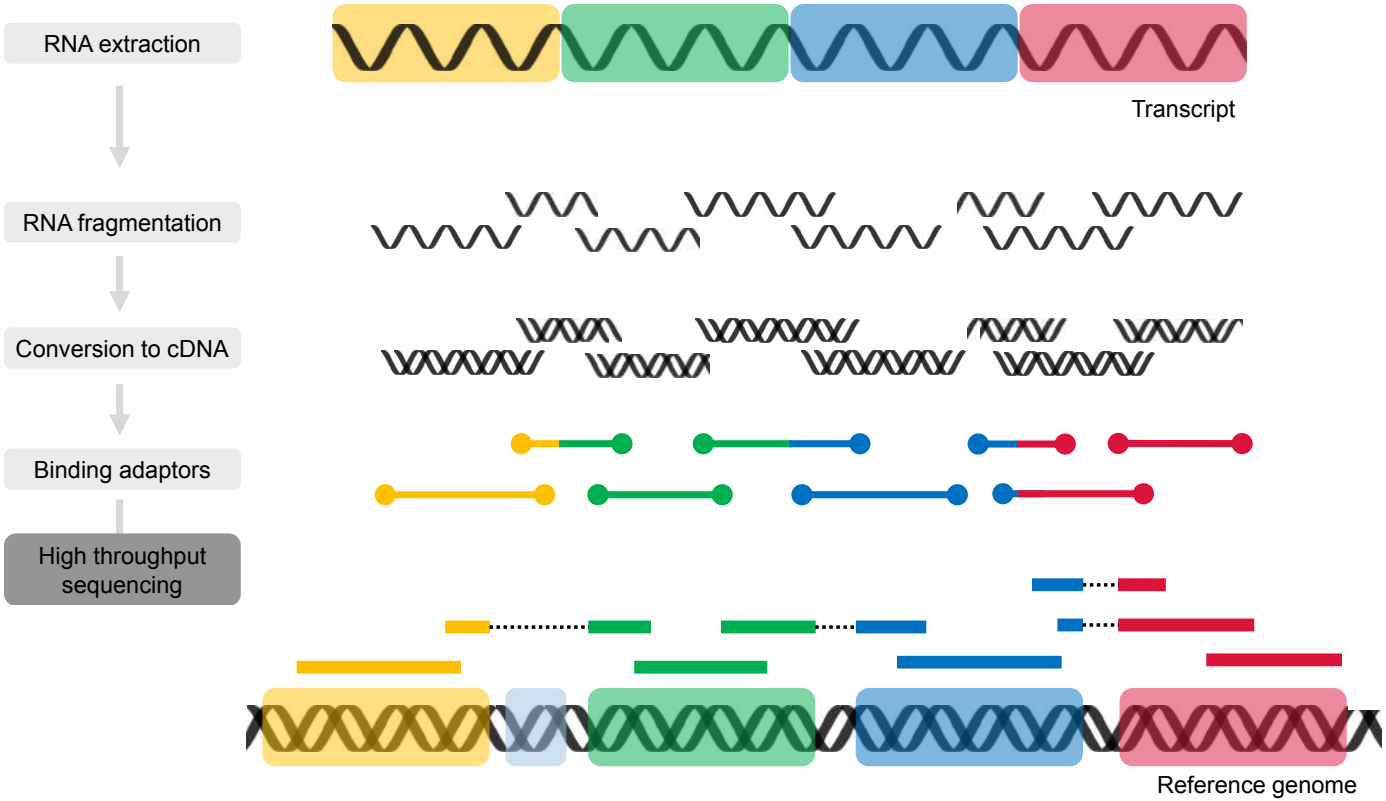
The Evolutionary Landscape of Alternative Splicing in Vertebrate Species

Nuno L. Barbosa-Morais,[1,2] Manuel Irimia,[1*] Qun Pan,[1*] Hui Y. Xiong,[3*] Serge Gueroussov,[1,4*] Leo J. Lee,[3] Valentina Slobodeniuc,[1] Claudia Kutter,[5] Stephen Watt,[5] Recep Çolak,[1,6] TaeHyung Kim,[1,7] Christine M. Misquitta-Ali,[1] Michael D. Wilson,[4,5,7] Philip M. Kim,[1,4,6] Duncan T. Odom,[5,8] Brendan J. Frey,[1,3] Benjamin J. Blencowe[1,4]†



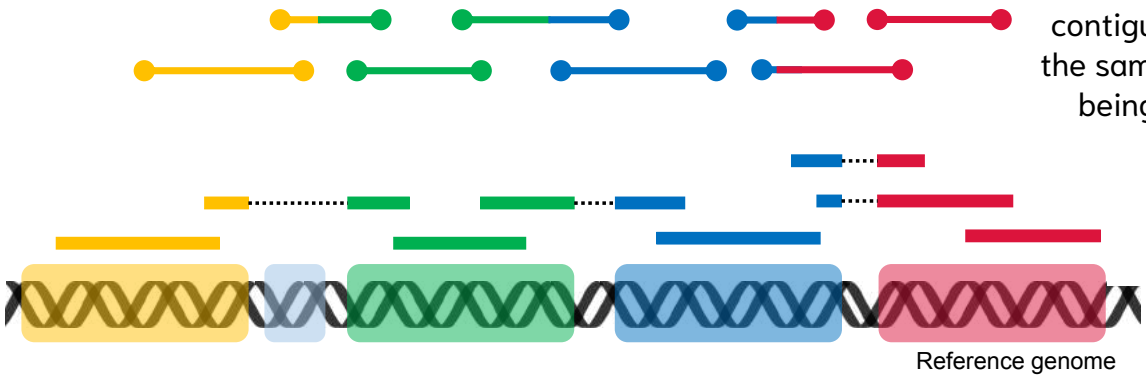
Estimating alternative sequence inclusion levels from RNA-sequencing data

Estimating alternative sequence inclusion levels from RNA-sequencing data



Estimating alternative sequence inclusion levels from RNA-sequencing data

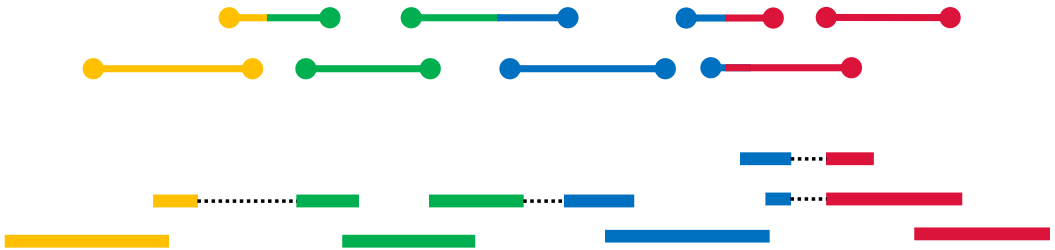
Mertes, C. et al. Detection of aberrant splicing events in RNA–seq data using FRASER. Nat Commun 12, 529 (2021).



RNA–seq split reads: ends align to two non–contiguous regions of the same chromosome, being evidence of splicing

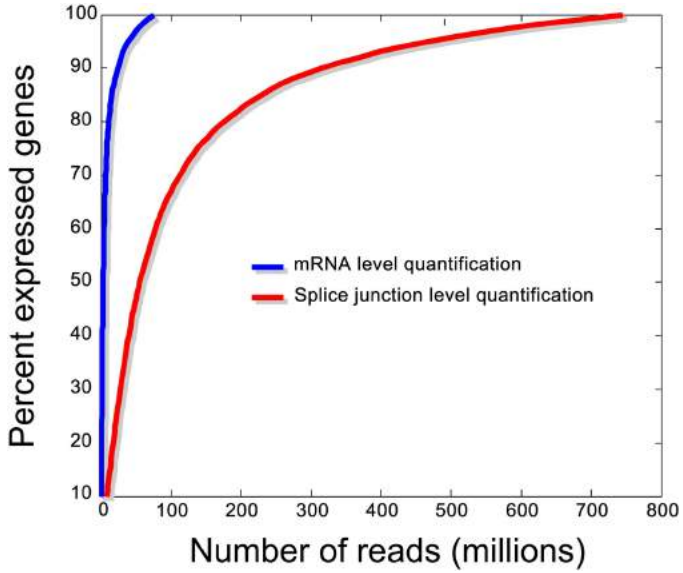
Estimating alternative sequence inclusion levels from RNA-sequencing data

- **small fraction of short** reads provide evidence for splicing
- **short reads fail** to disambiguate full-length alternative transcript isoforms

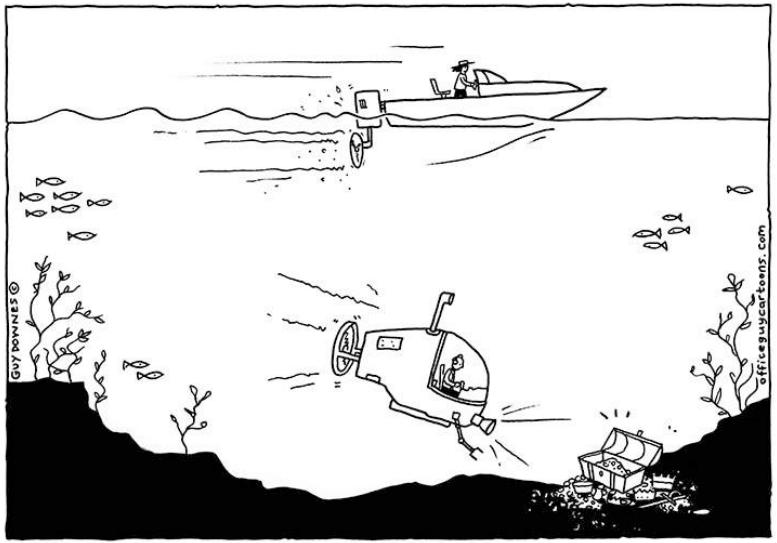


Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5, 621–628 (2008).

You need to sequence deeper to quantify alternative splicing

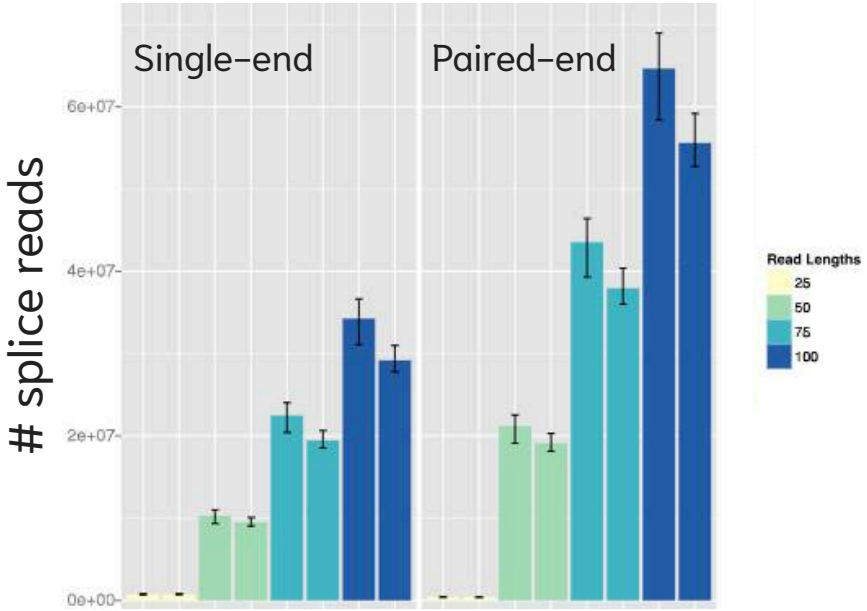


"Treasure is rarely found by skimming across the surface."



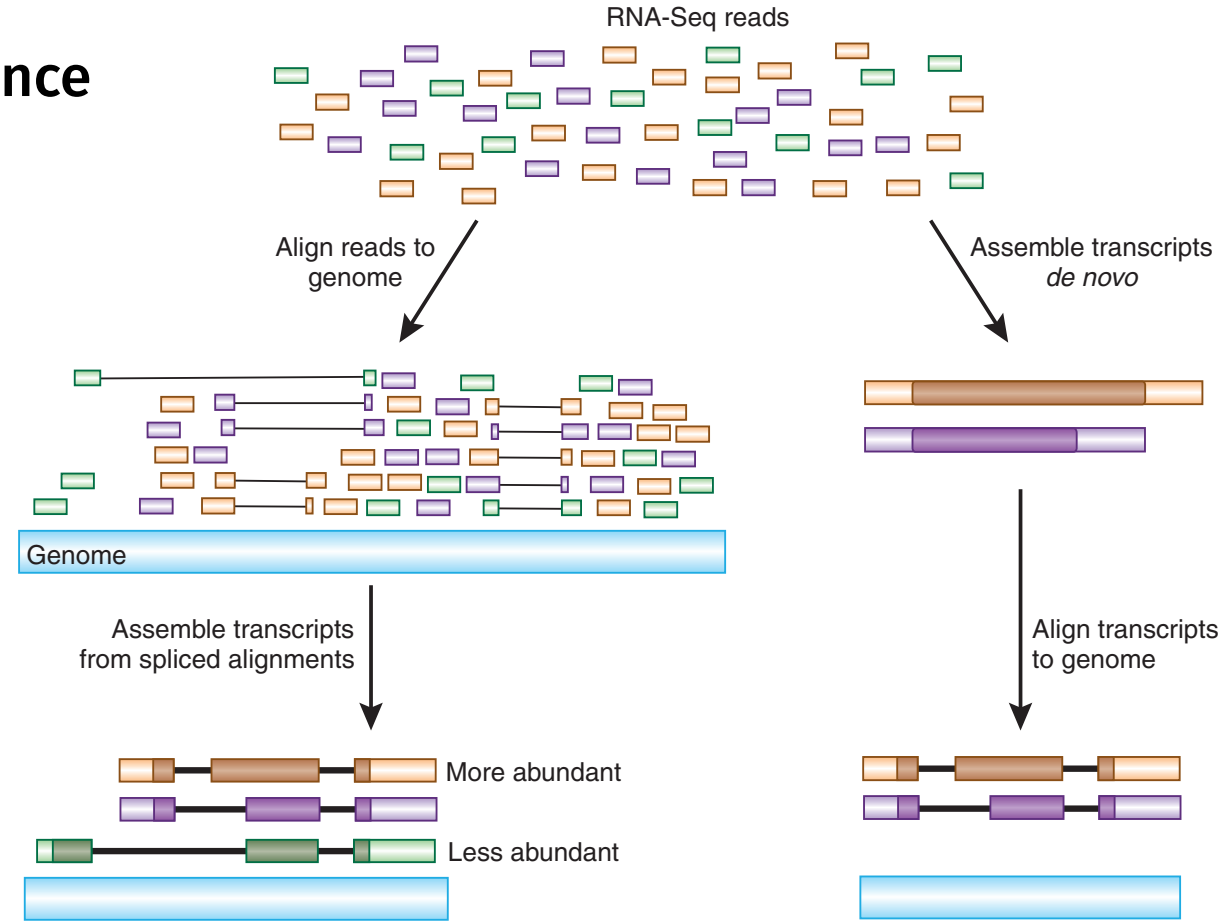
Office Guy Cartoons

Longer and paired-end reads favor detection of splice junctions



Chhangawala, S., Rudy, G., Mason, C. E. & Rosenfeld, J. A. The impact of read length on quantification of differentially expressed genes and splice junction detection. Genome Biol 16, 131 (2015).

From sequence to sense

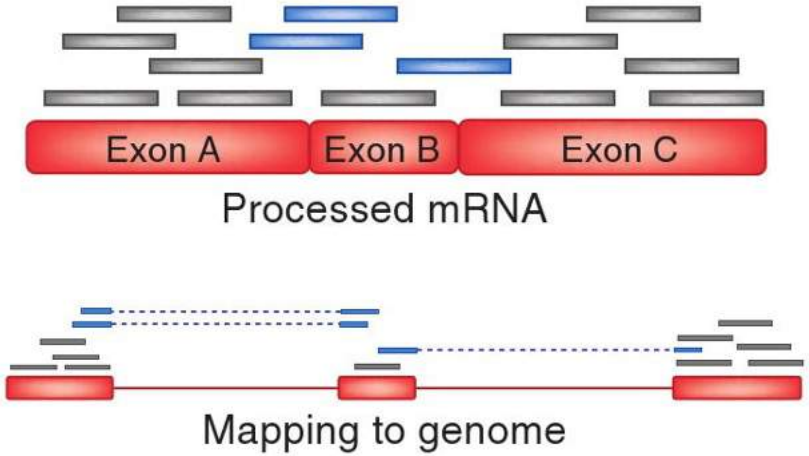


very noisy with short reads

Hass & Zody (2010) Nature Biotechnology

Reference-based: genome or transcriptome?

Trapnell & Salzberg (2009)
Nature Biotechnology



Genome mapping:

- Can identify novel features
- Trickier to align splice reads
- More difficult to reconstruct isoform and gene structures

Transcriptome mapping:

- Reference not repetitive
- No complex structures
- Harder to find novel features
- Highly dependent on annotation

Do we really need to align?

much **faster**

locate **genome coordinates** from which reads originate

identify reads' potencial transcripts of **origin**

rather than performing a full alignment, records information about **the set of transcripts a read is compatible with**

2009

TopHat/TopHat2
STAR
HISAT/HISAT2

2011

eXpress
RSEM

2016

kallisto
Salmon

Yi, L., Liu, L., Melsted, P. & Pachter, L. A direct comparison of genome alignment and transcriptome pseudoalignment. Biorxiv 444620 (2018) doi:10.1101/444620.

Estimating alternative sequence inclusion levels from RNA-sequencing data

Transcript isoform **1**

Transcript isoform **2**

Isoform–centred approach

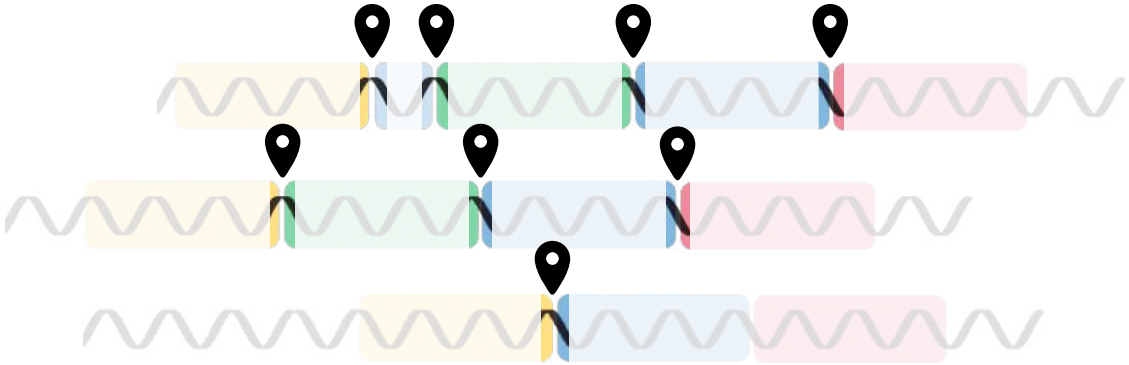
How much of the gene's expression is transcript isoform 1?

estimate full-length transcripts' relative abundances based on sequencing reads

- `Cufflinks/cuffdiff2`
- `MISO`
- `salmon`
- `kallisto`
- `STAR`
- `(…)`

Reference genome

Estimating alternative sequence inclusion levels from RNA-sequencing data



align spliced reads to a **database** of known splice junctions

- rMATS
- MISO
- SUPPA
- vast-tools
- MAJIQ
- (…)

Event-centred approach

What is the proportion of the gene's transcripts that **include** the **alternative exon**

Reference genome

Different approaches and their interpretations

AltAnalyze.org
Alternative Splicing and Functional Prediction

SpliceSeq

SUPPA2

Whippet

AJIQ

VAST-TOOLS

Multivariate Analysis of Transcript Splicing (MATS)

But... Non-computational biologists **not using** or **misusing** them:

- user-unfriendly
- hard-to-interpret event annotation and statistics
- resource-demanding for basic analysis of large datasets (e.g. TCGA)

Differential splicing with many samples

psichomics

Established tools for AS analysis of RNA-seq 😩

- Resource-demanding for basic analysis of large datasets
- No support for imputing **pre-processed data** (e.g. splice junction read counts, available for GTEx, TCGA and SRA)

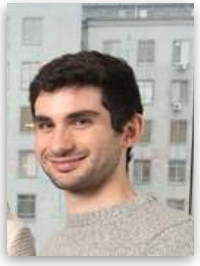
- **Limited** (e.g., restricted to pairwise comparisons) or **hard-to-interpret** statistics
- Unintelligible AS event annotation
- No incorporation of **molecular/clinical/design information** (e.g. survival)
- No **sub-setting** of events of interest for **interactive exploration**
- No user-friendly **interactive graphical interface**
- No support for **customisable statistical plots**

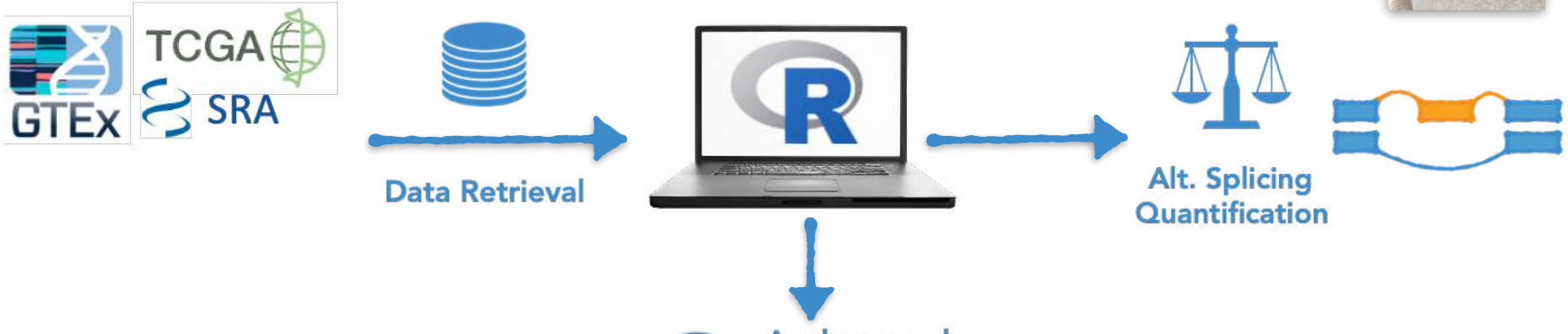
The (mathematical) nature of the **percent spliced–in (PSI)** ratio

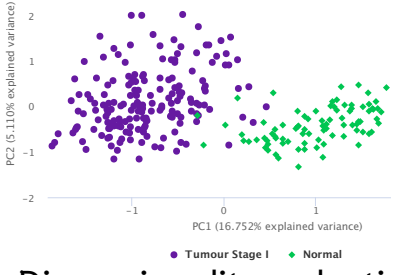
Nucleic Acids Res, 2019, Vol. 47, No. 2 e7–e7 doi: doi: 10.1093/nar/gky888

psichomics: graphical application for alternative splicing quantification and analysis

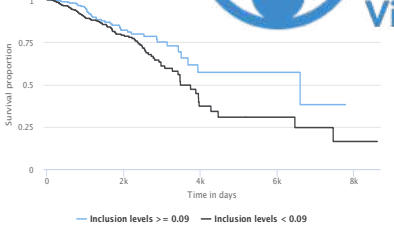
Nuno Saraiva-Agostinho and Nuno L. Barbosa-Morais



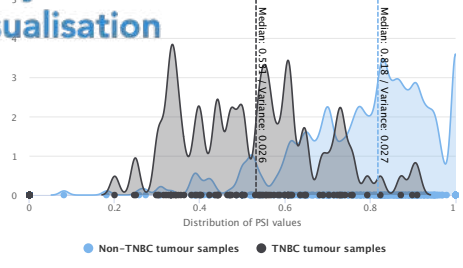




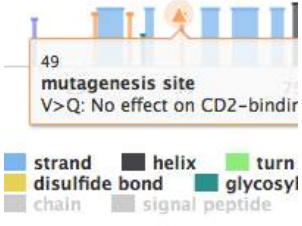
Dimensionality reduction techniques



Survival analysis



Differential splicing and gene expression analysis

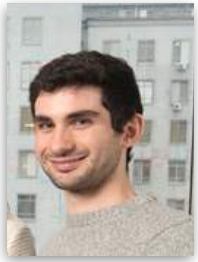


Gene, RNA and protein information

Data Retrieval

Alt. Splicing Quantification

Analyses and Visualisation



psichomics

Clinical data

Junction read counts

Gene expression

TCGA

GTEx Portal

recount2

User-provided data

33 human cancer types

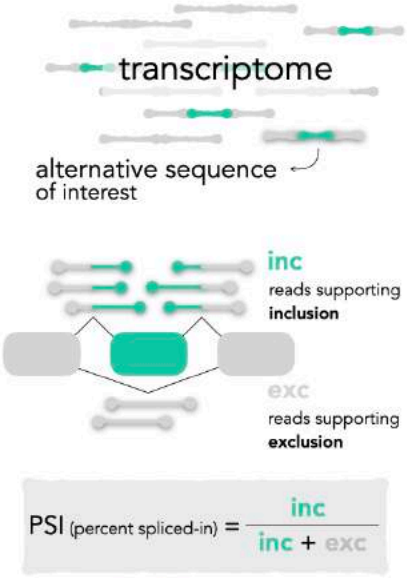
54 human normal tissues

>2000 human studies from SRA

Alternative splicing annotation

- Human annotation (hg19 + hg38)
- Annotation for 13 new species based on VAST-TOOLS
- User-provided custom annotation files

MOUSE
CHICKEN
ZEBRAFISH
FROG
FRUIT FLY
ARABIDOPSIS
...

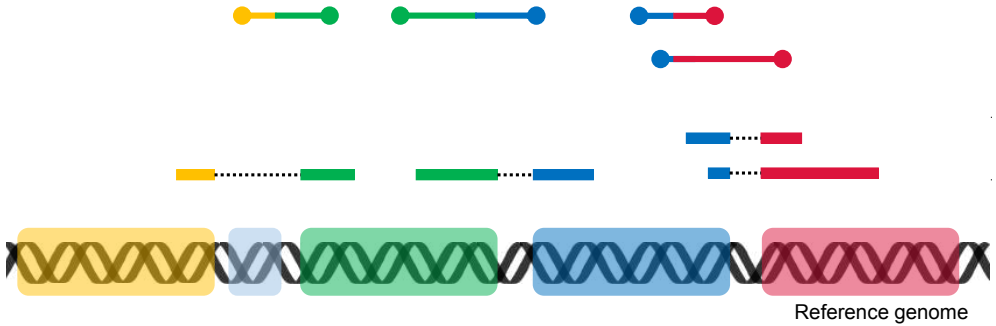


Event-centred approach

What is the percentage of transcripts that **include** the **alternative exon**

RNA-seq split reads or **junction reads**

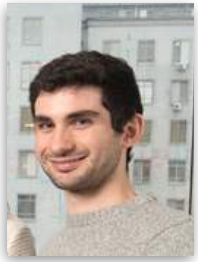
support inclusion or exclusion of the alternative sequence defined by those junctions



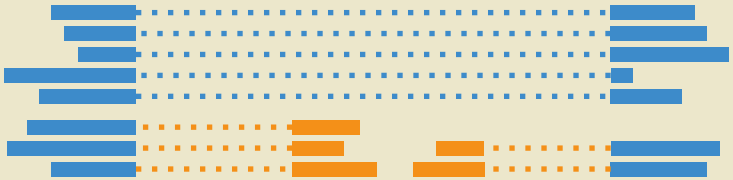
Data Retrieval

Alt. Splicing Quantification

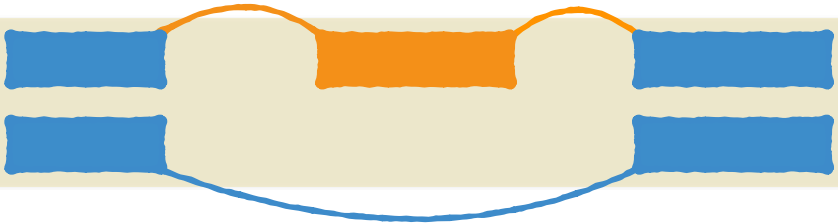
Analyses and Visualisation



psichomics



Junction read counts



Alternative splicing annotation

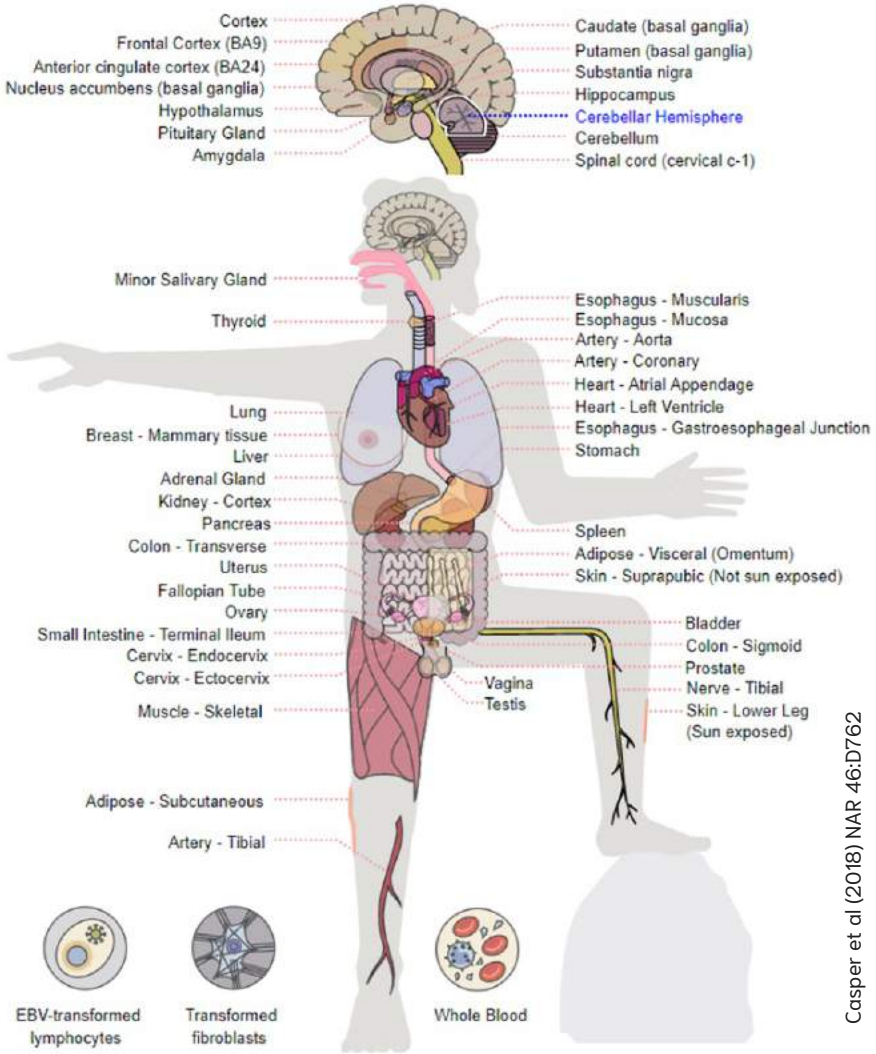
$$\text{Percent Spliced-In (PSI)} = \frac{\text{inclusion reads}}{\text{inclusion + exclusion reads}}$$

Analyse publicly available alternative splicing data using psichomics

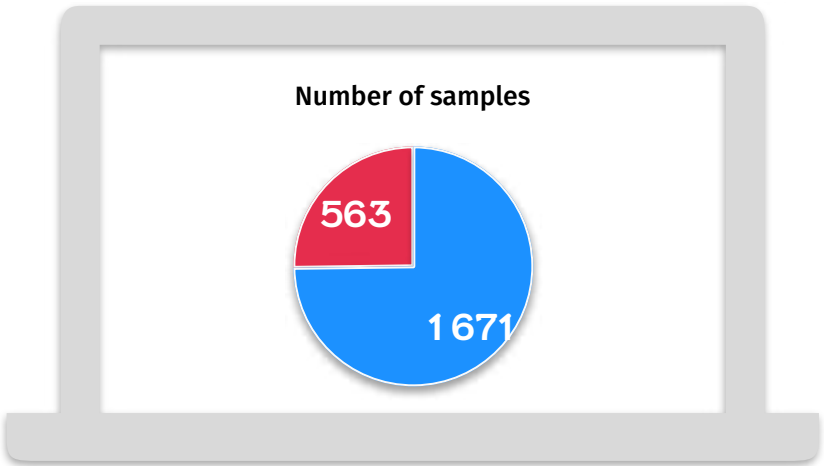
Genotype–Tissue Expression

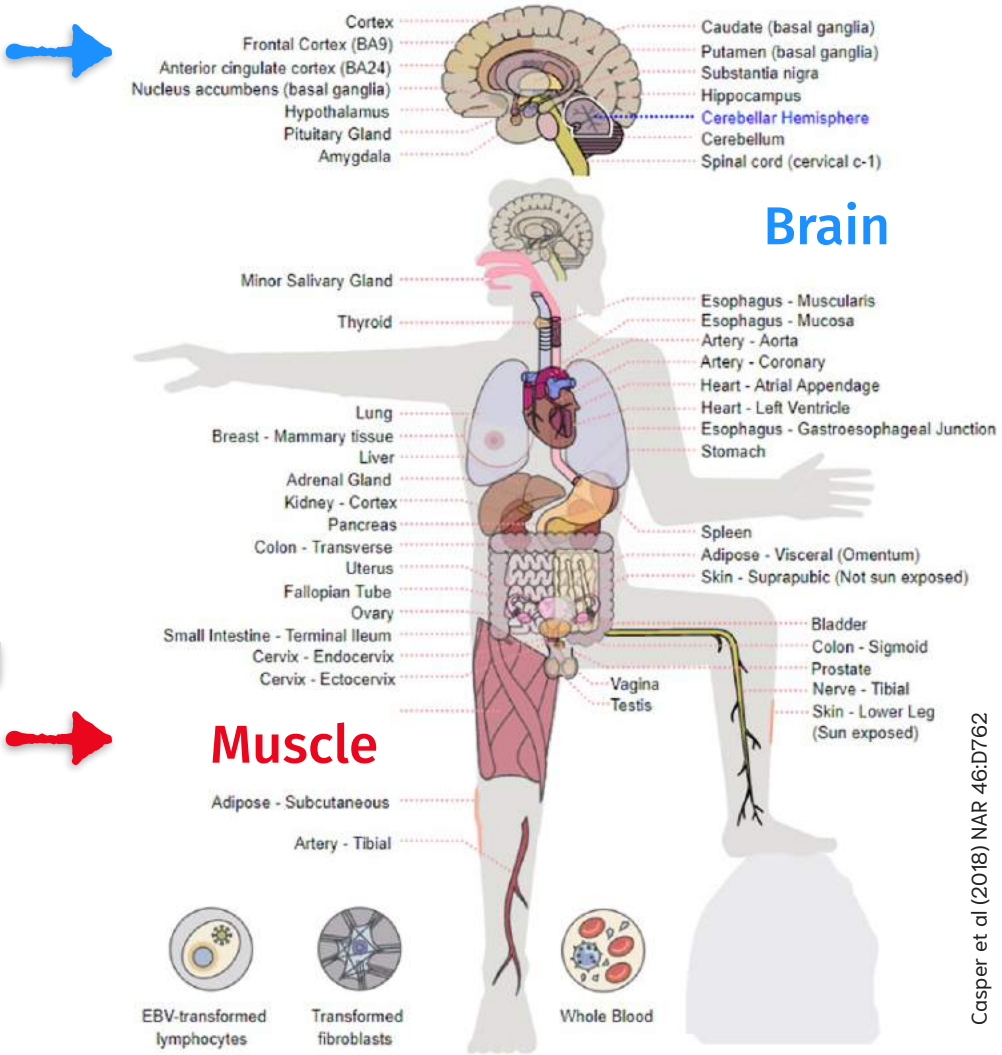
RNA–seq and genotype from 53 human tissues

- > 10 000 samples (100s per tissue)
- clinically annotated (age, gender, ethnicity, cause of death, diseases)



Casper et al (2018) NAR 46:D762





Casper et al (2018) NAR 46:D762

Structure of PSI tables other psichomics objects

```
loadGtexData()
```

```
data = c("sampleInfo", "subjectInfo", "junctionQuant"),
tissue = c("Brain", "Muscle")
```

for a particular GTEx `release`
with files stored locally in a given `folder`

Sample metadata

Clinical data

Junction read counts

Structure of PSI tables other psichomics objects

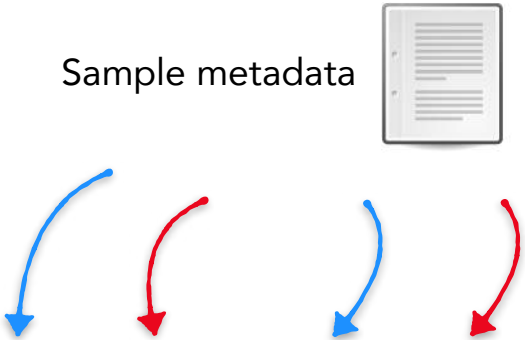
`loadAnnotation()`

for a particular `species, assembly, date`

Alternative splicing annotation

Structure of PSI tables other psichomics objects

Sample metadata



for a particular `eventType` and `minReads`

`quantifySplicing()`

Junction read counts

Alternative splicing annotation

events

GTEx samples

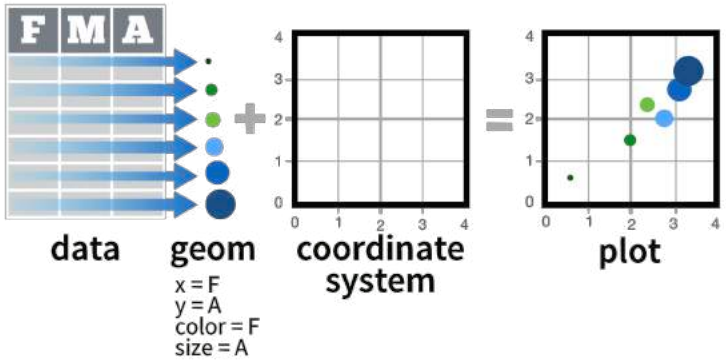
	GTEX-1117F-3266-SM-5N9CT	GTEX-111CU-2026-SM-5GZZC	GTEX-111FC-0326-SM-5GZZ1	GTEX-111FC-3126-SM-5GZZ2
SE_1_+_764484_783034_(...)_LOC643837	0.2727273	0.757575758	0.5402299	0.53333333
SE_1_-_1198726_1192690_(...)_UBE2J2	1.0000000	1.000000000	0.9850746	0.96938776
SE_1_-_1198726_1197770_(...)_UBE2J2	0.0000000	0.005263158	0.0000000	0.01176471
SE_1_-_1198726_1197770_(...)_UBE2J2	0.0000000	0.005263158	0.0000000	0.00000000

Note regarding ggplot()

events

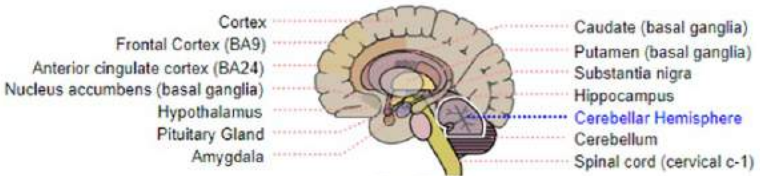
GTEx samples

	GTEX-1117F-3266-SM-5N9CT	GTEX-111CU-2026-SM-5GZZC	GTEX-111FC-0326-SM-5GZZ1	GTEX-111FC-3126-SM-5GZZ2
SE_1_+_764484_783034_(...)_LOC643837	0.2727273	0.757575758	0.5402299	0.53333333
SE_1_-_1198726_1192690_(...)_UBE2J2	1.0000000	1.000000000	0.9850746	0.96938776
SE_1_-_1198726_1197770_(...)_UBE2J2	0.0000000	0.005263158	0.0000000	0.01176471
SE_1_-_1198726_1197770_(...)_UBE2J2	0.0000000	0.005263158	0.0000000	0.00000000

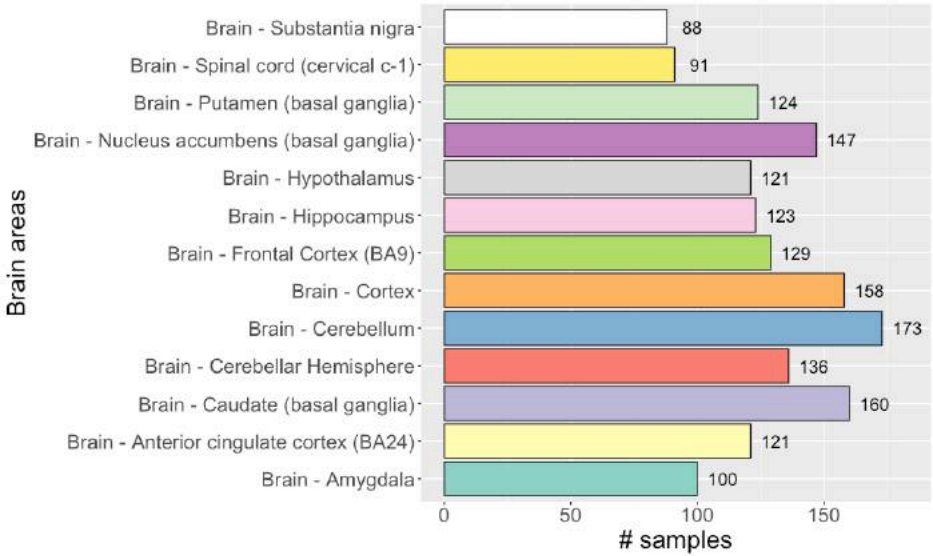


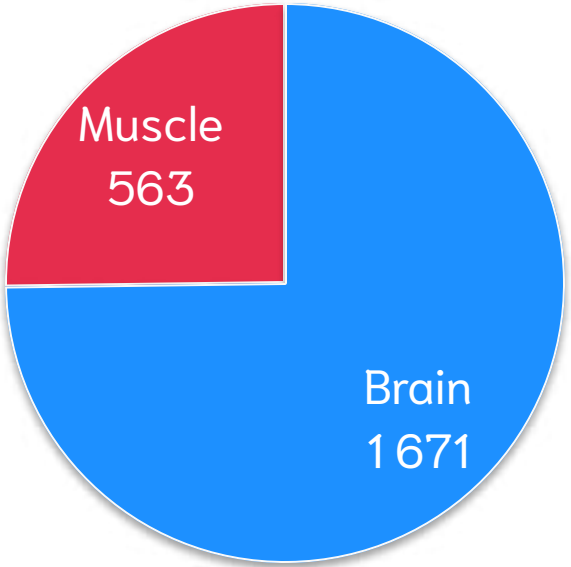
Ex 1:

Explore the distribution of brain samples by brain areas.



Load inclusion levels from GTEx (using `psichomics`)

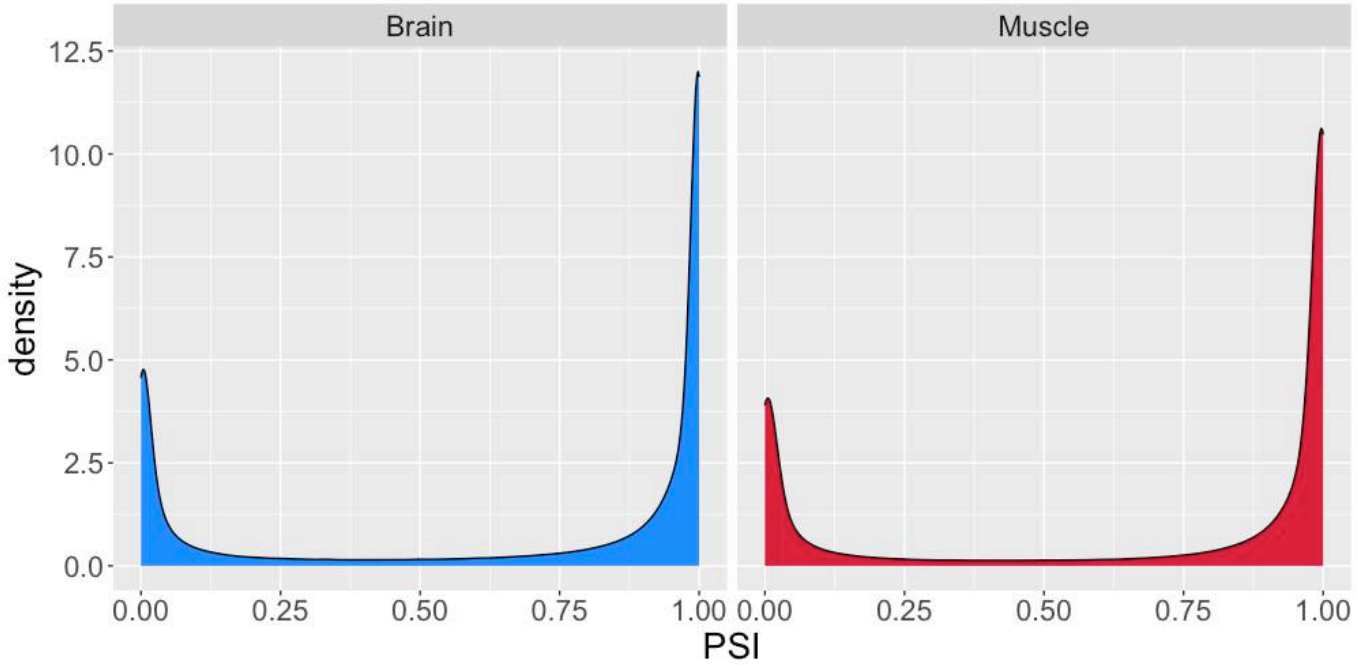




Ex 2:

How does the PSI distribution look like?
Is it similar for both tissues?

PSI distributions are zero-one-inflated

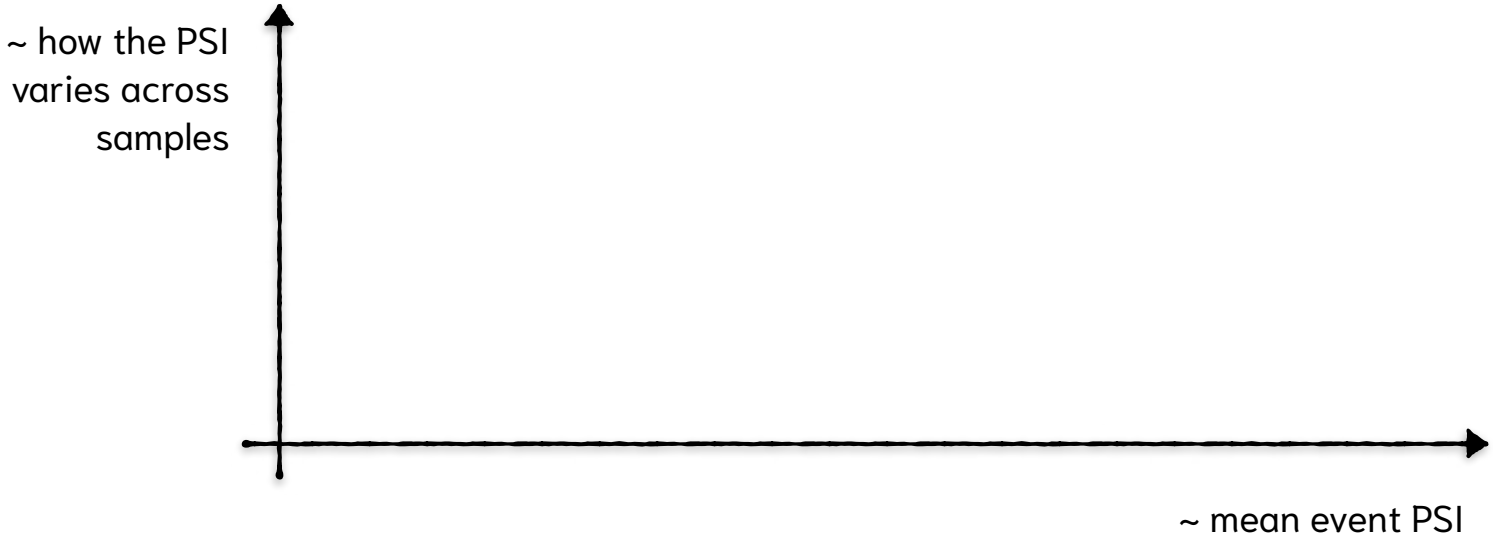


Ex 3:

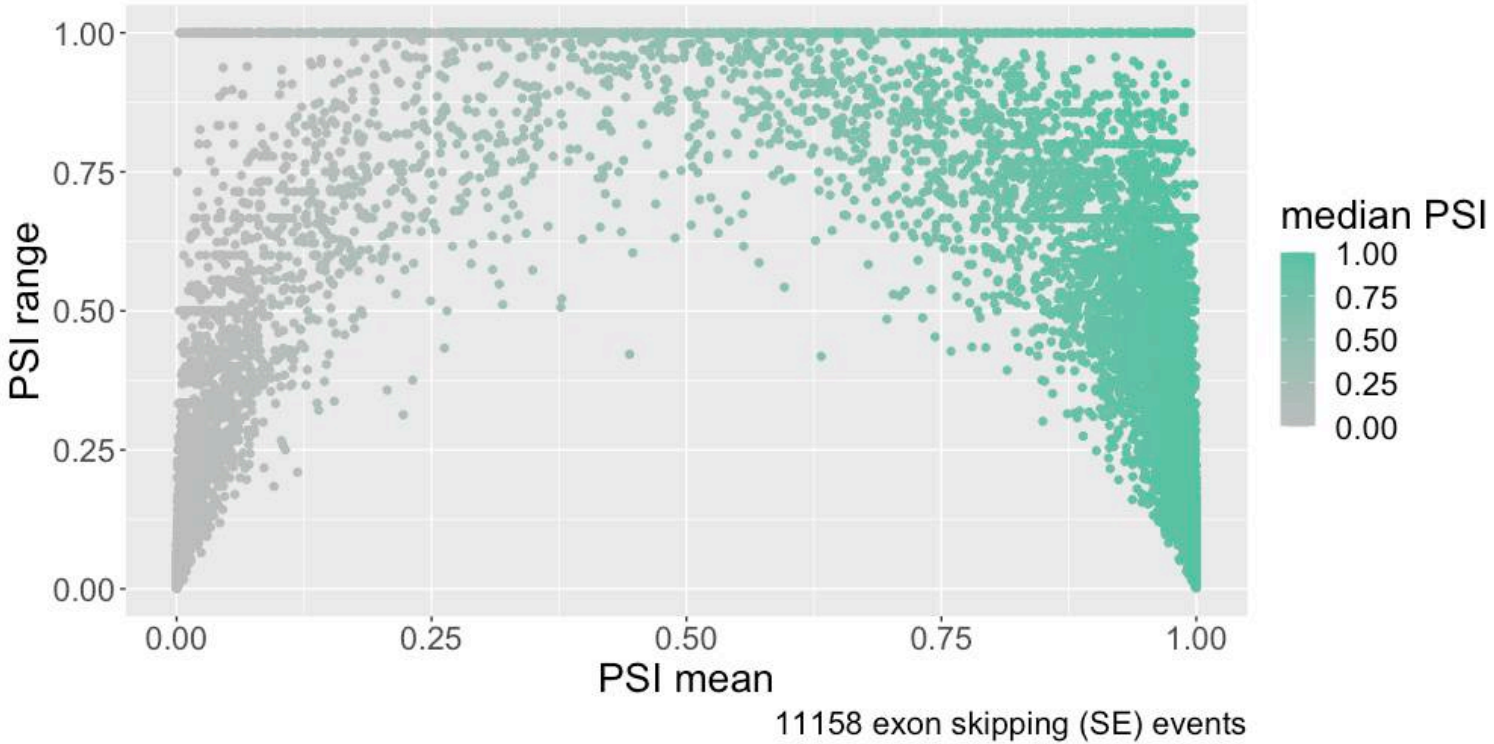
Are all events worth looking at?
If not, how do we select the most relevant?

Ex 3:

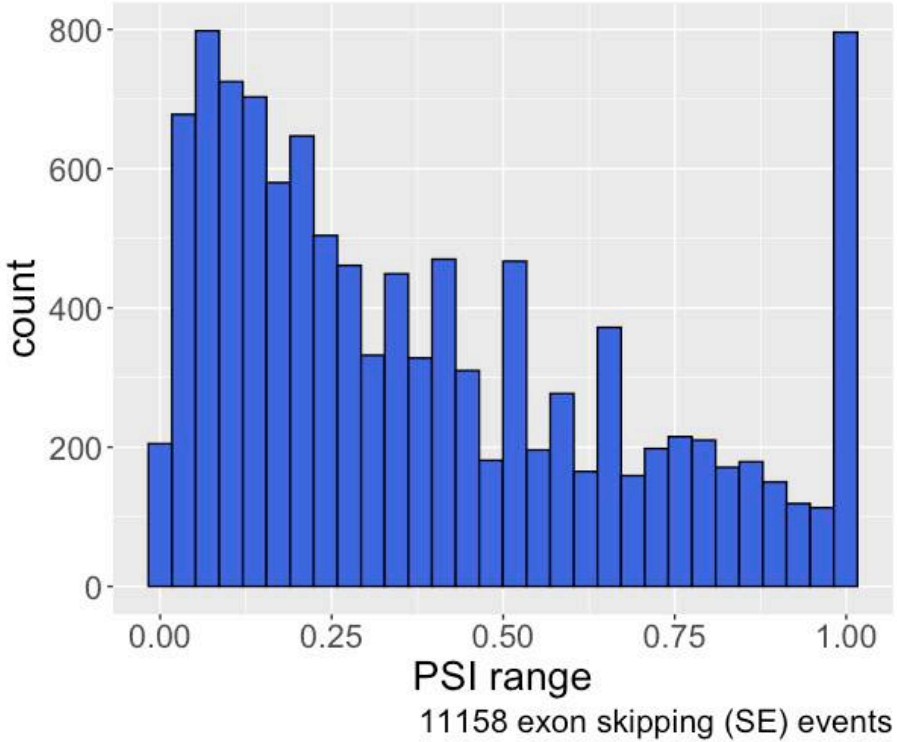
Are all events worth looking at?
If not, how do we select the most relevant?

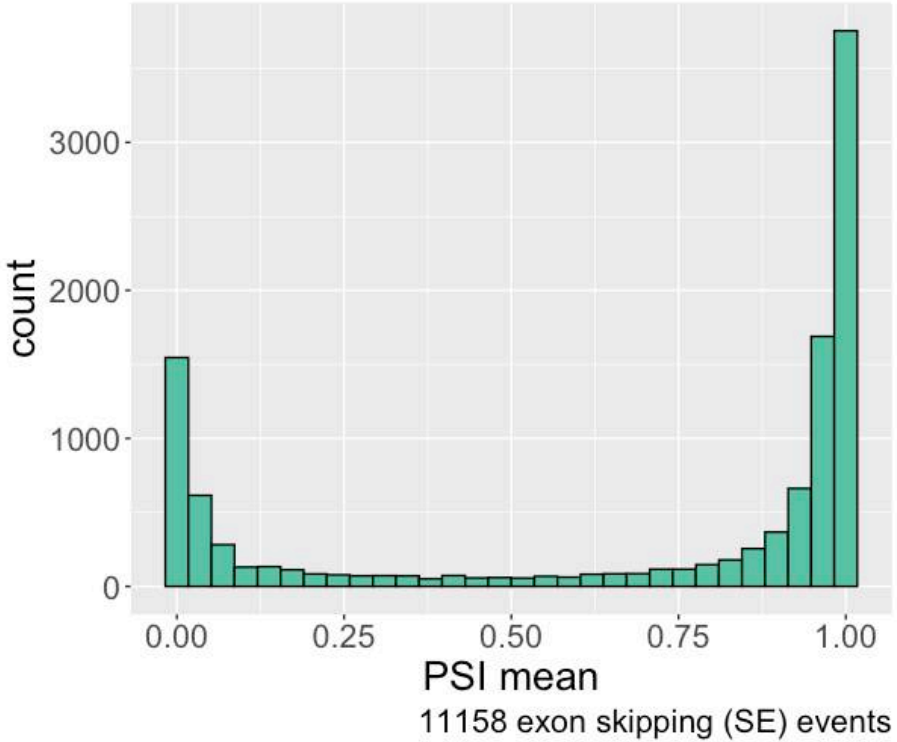


Selecting brain/muscle alternative splicing events



Selecting brain/muscle alternative splicing events

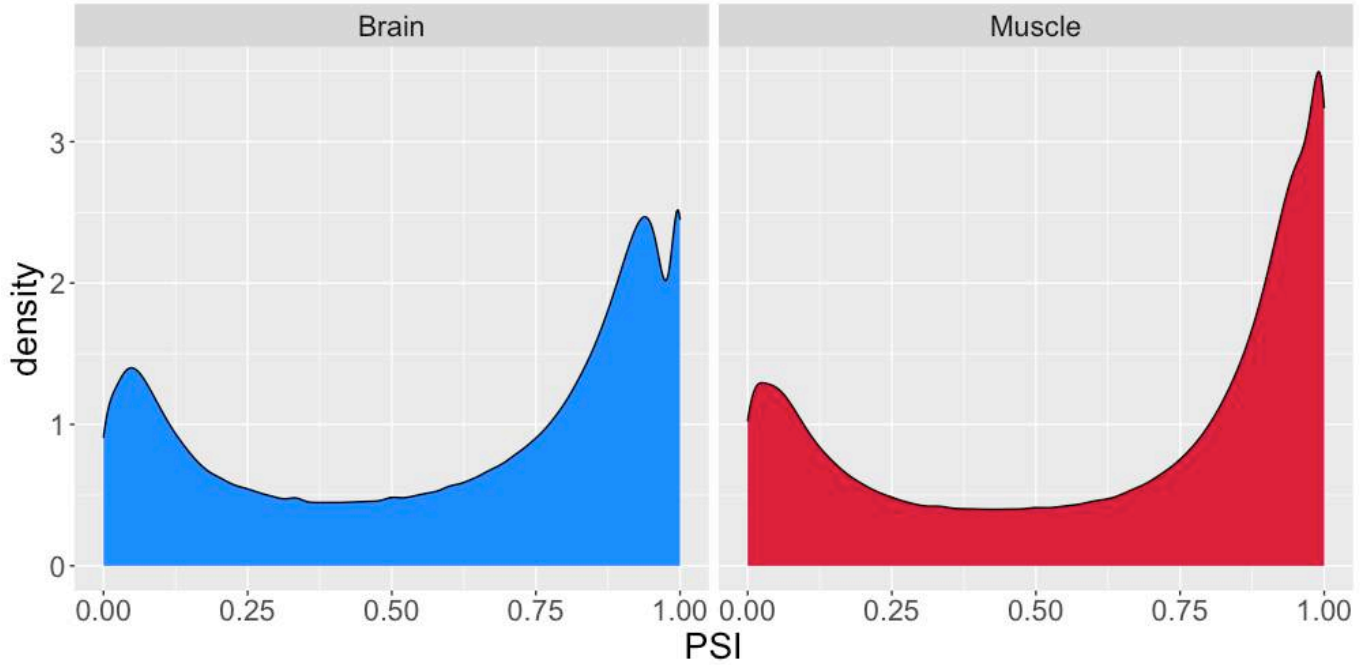




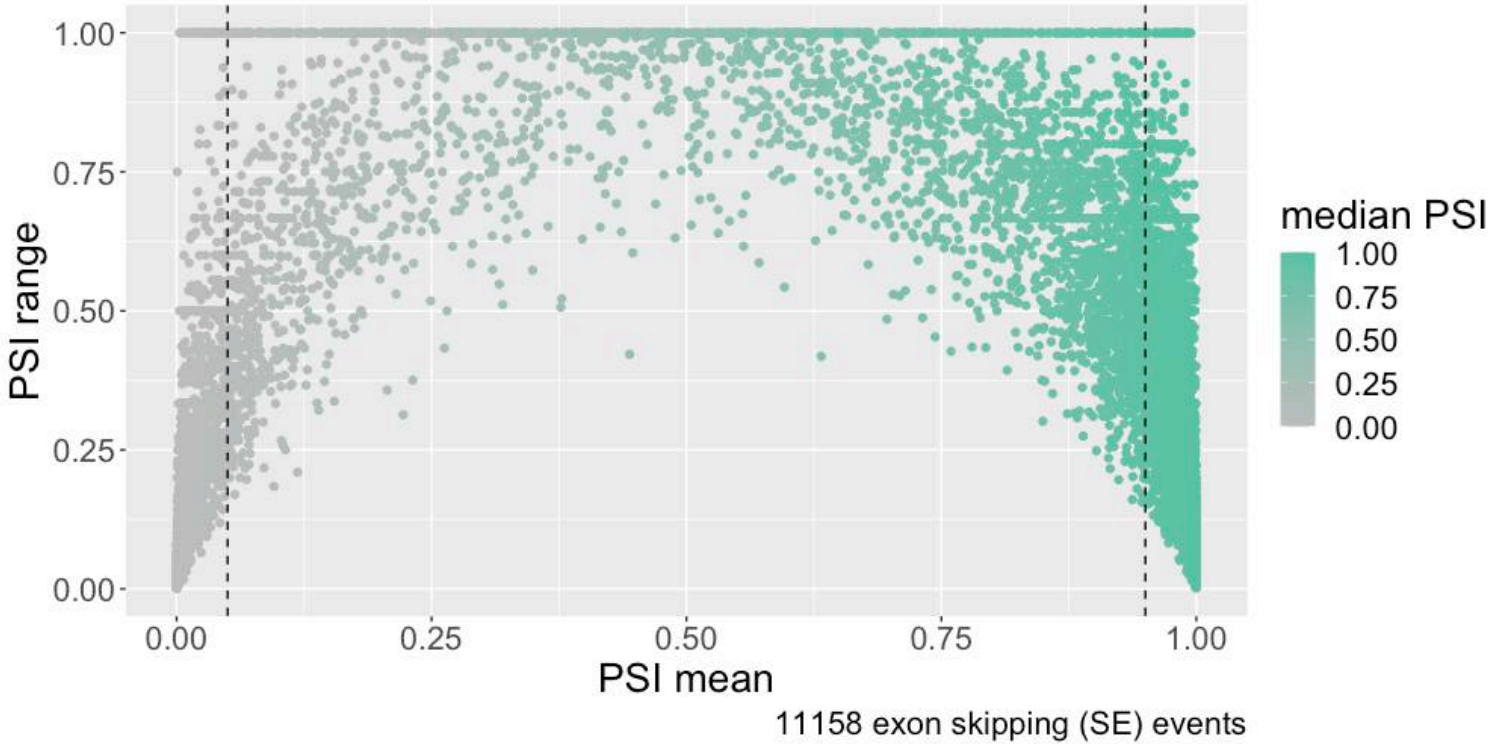
Ex 4:

Filter your PSI table to keep only
the most relevant (“alternative”) events.
Check how/if the distribution of PSIs has changed.

Selecting brain/muscle alternative splicing events



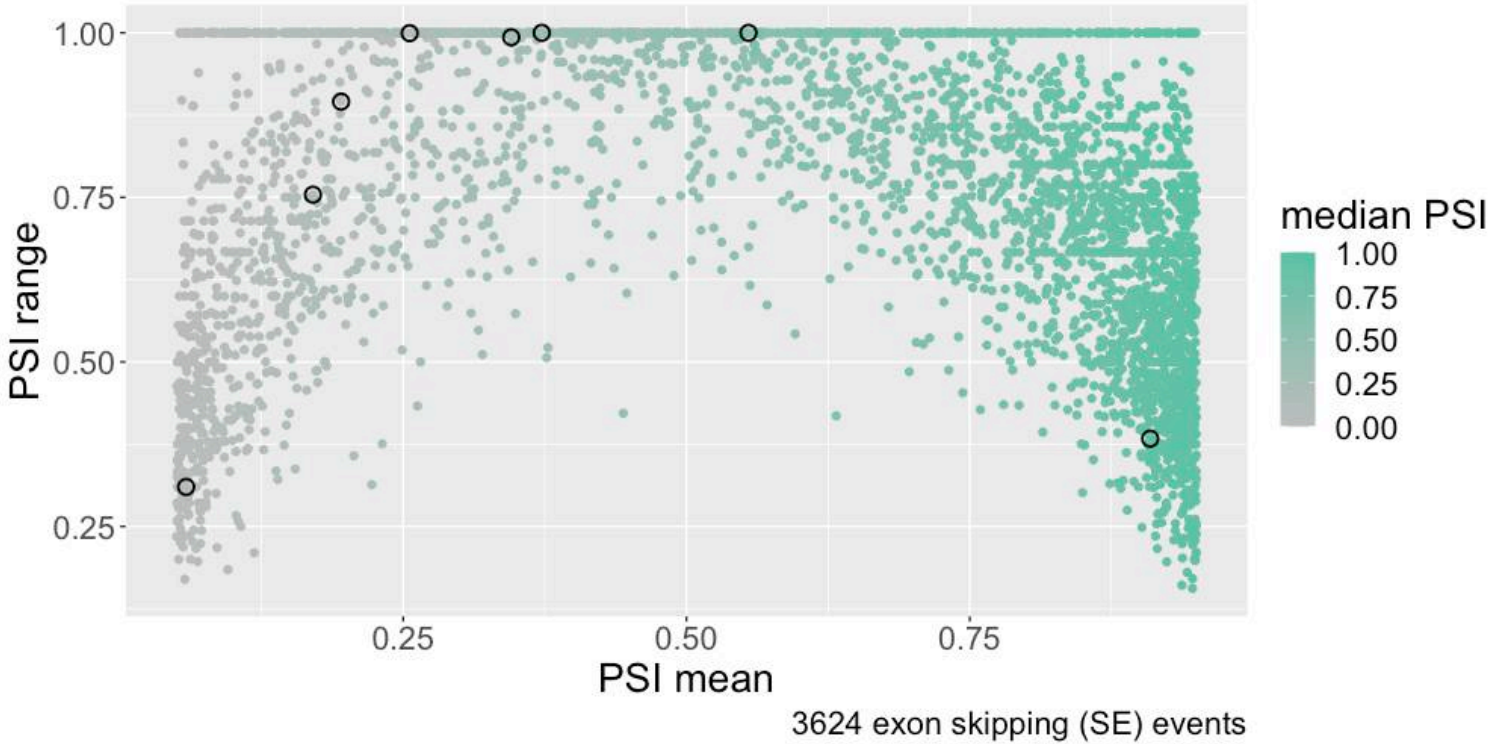
Selecting brain/muscle alternative splicing events



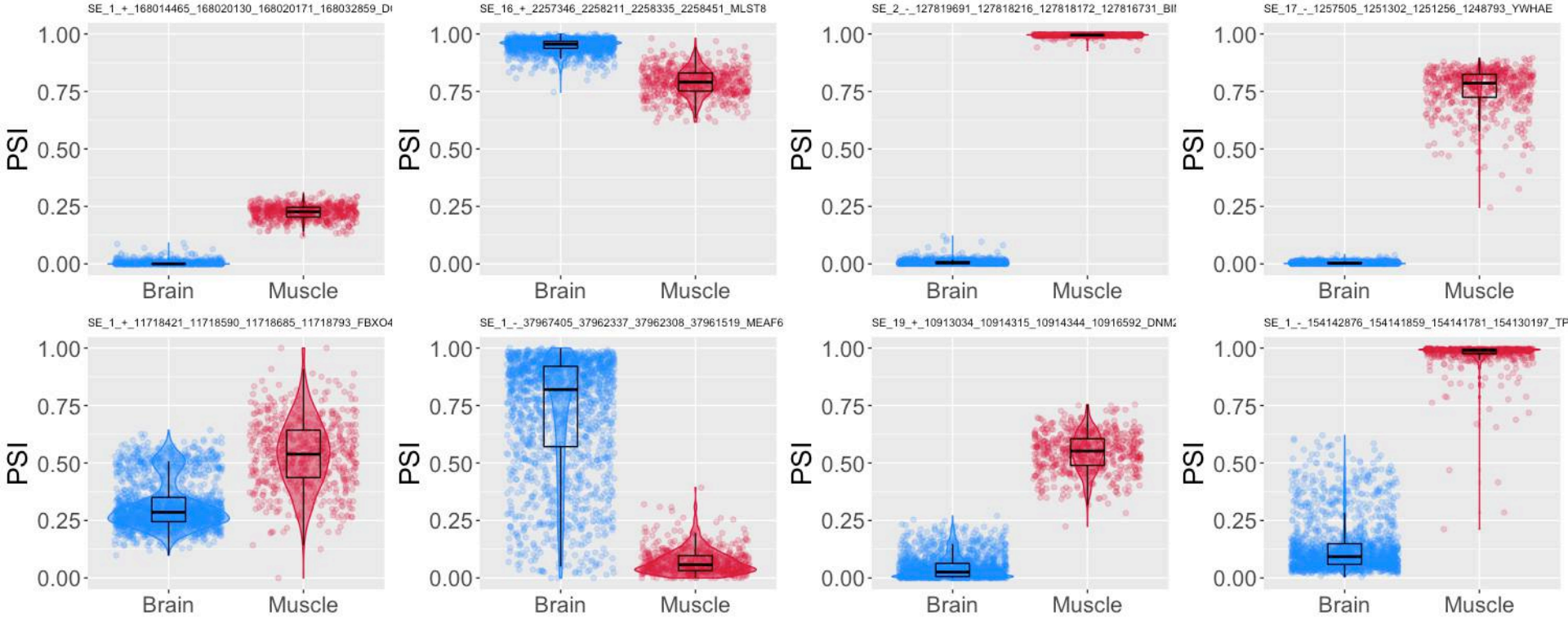
Ex 5:

Select some events in different areas of the plot. Can you predict their PSI distributions for samples from both tissues?

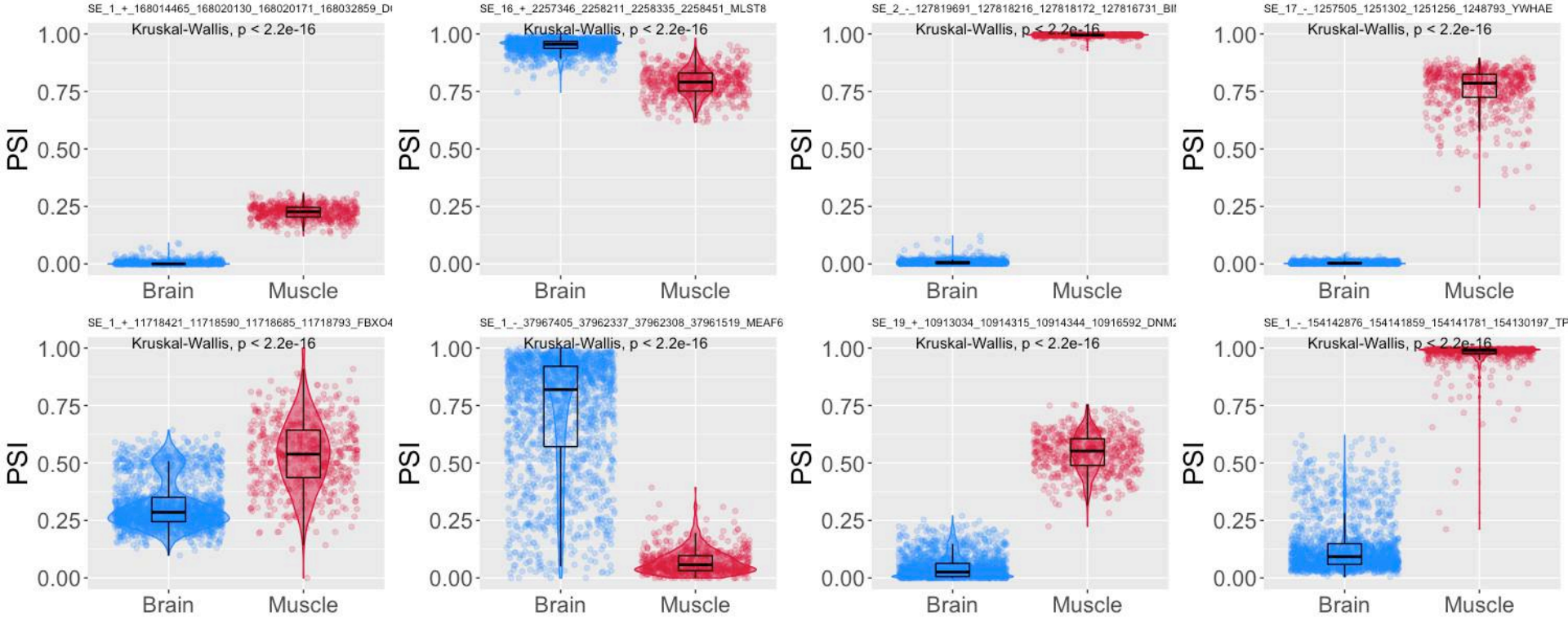
Selecting brain/muscle alternative splicing events



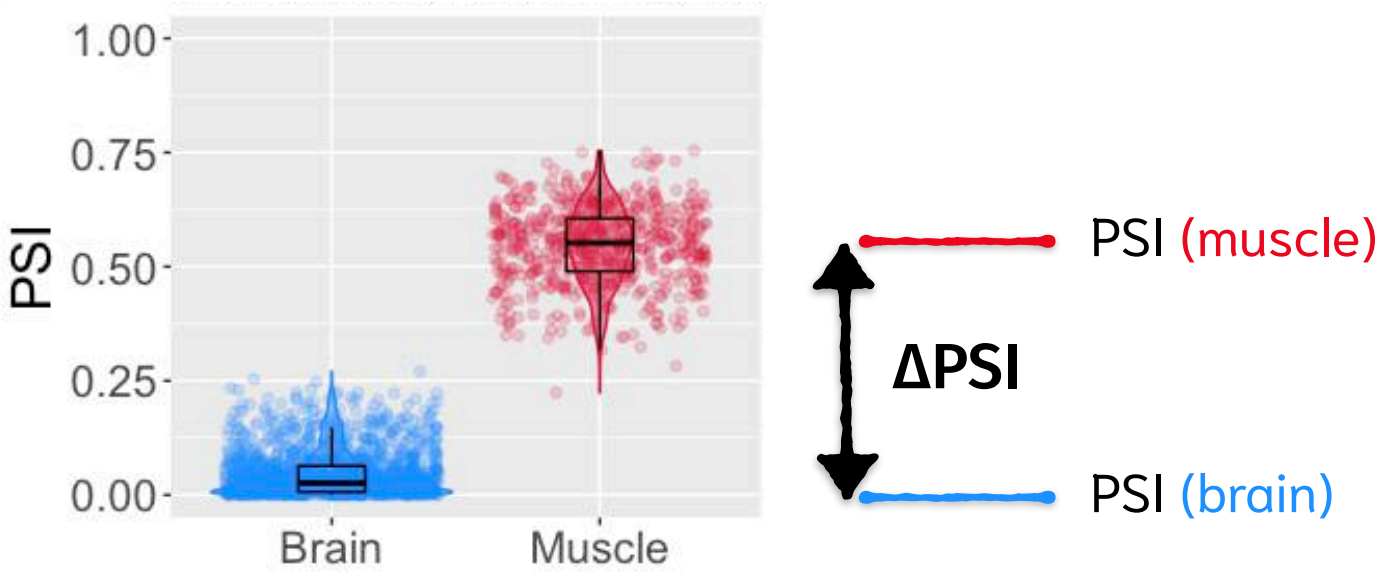
Measuring the effect size of splicing differences: muscle vs. brain



Measuring the effect size of splicing differences: muscle vs. brain

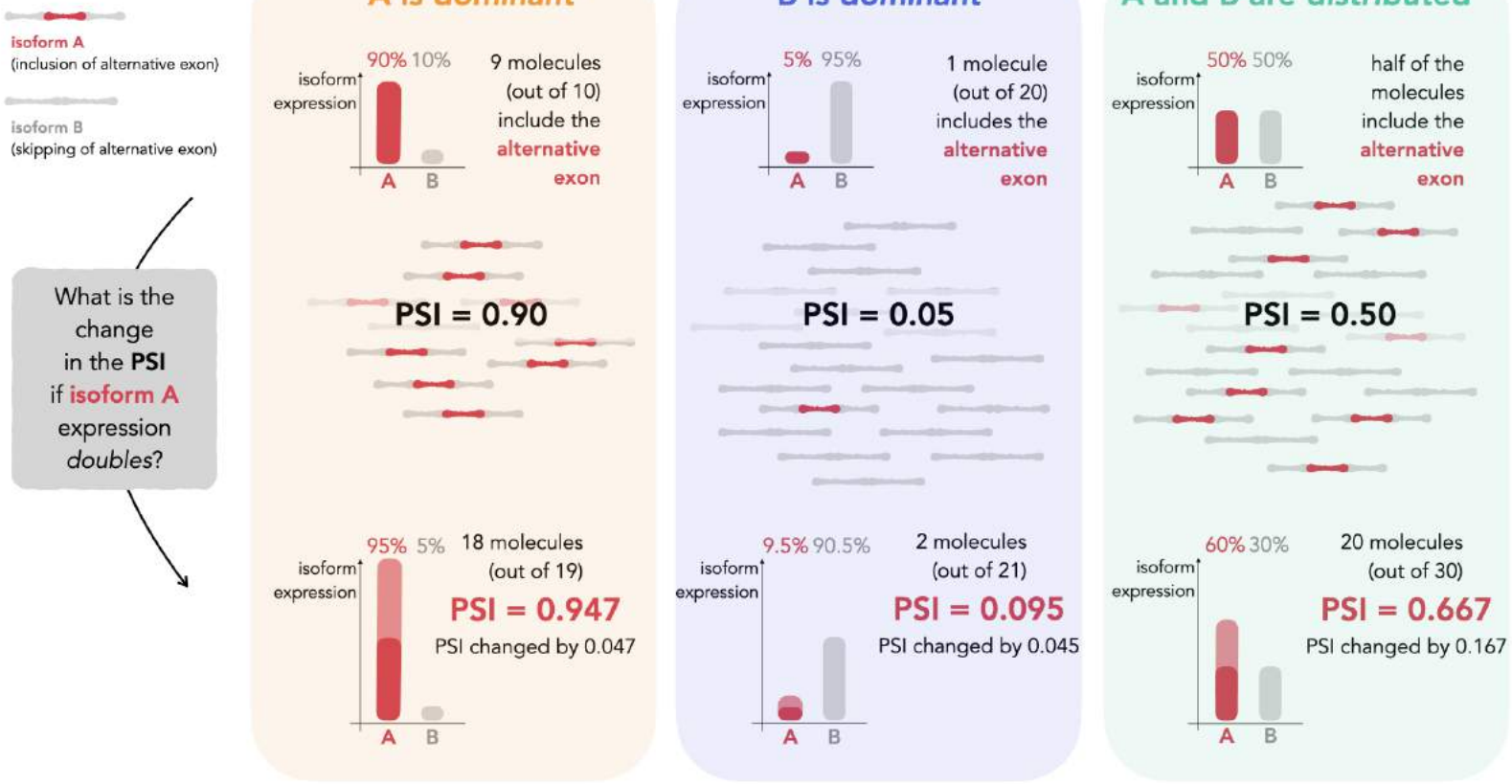


Measuring the effect size of splicing differences

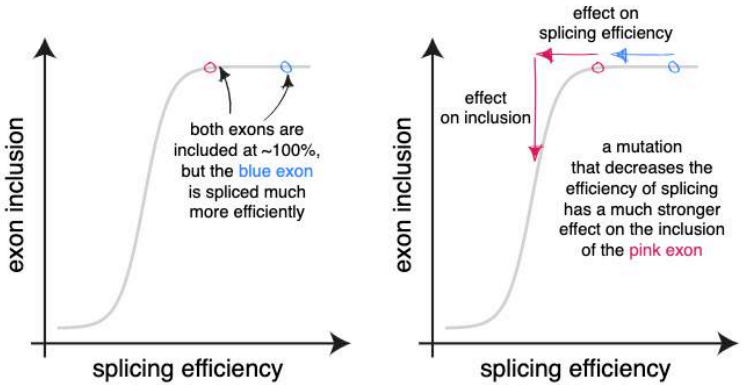


What does a given ΔPSI mean biologically?

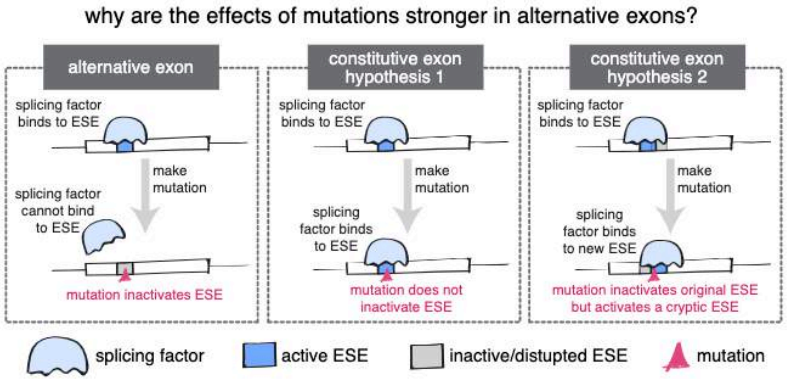
How isoform changes in expression impact PSI values



Scaling of (mutation) effects in alternative exons



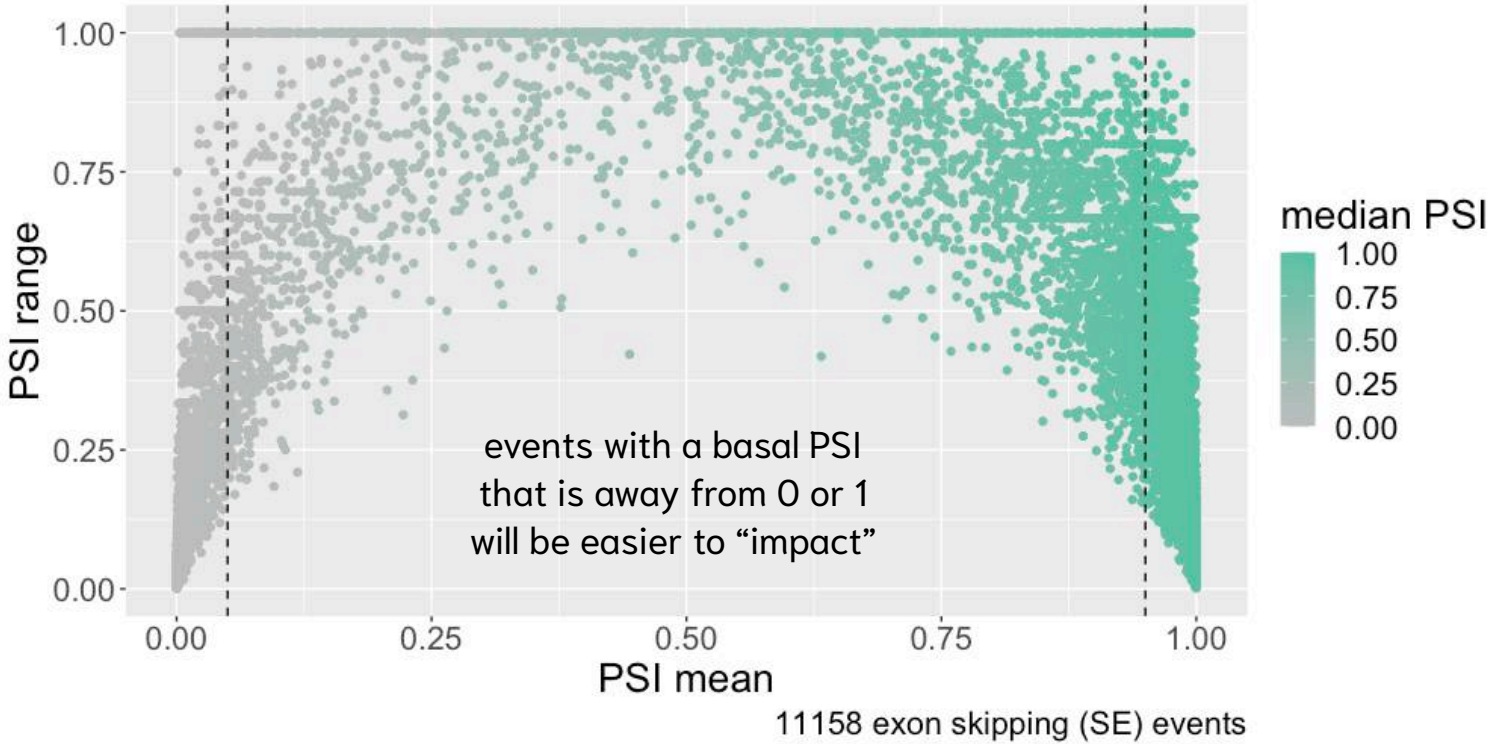
(figure 9)



(figure 7)

Baeza-Centurion, P., Miñana, B., Valcárcel, J. & Lehner, B. Mutations primarily alter the inclusion of alternatively spliced exons. Elife 9, e59959 (2020).

Scaling of (mutation) effects in alternative exons



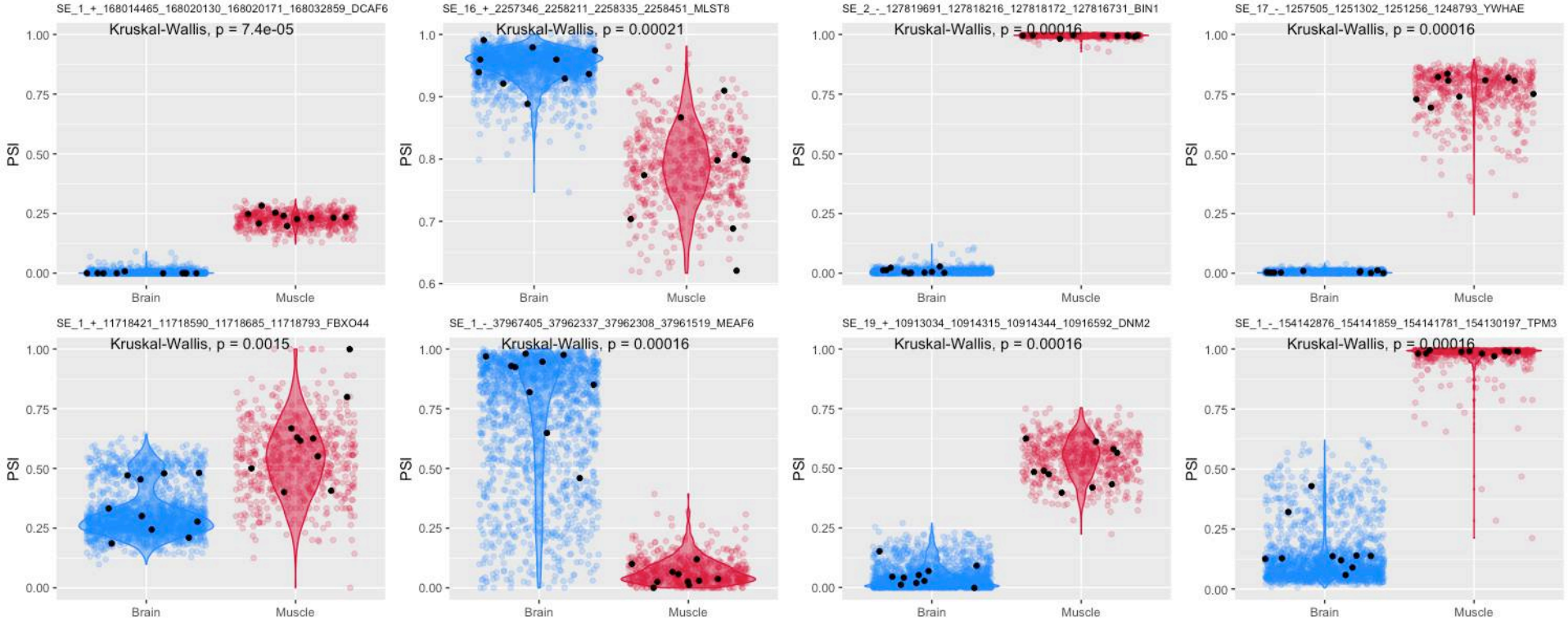
What happens in “real life”*?

*when you don’t have hundreds of replicates

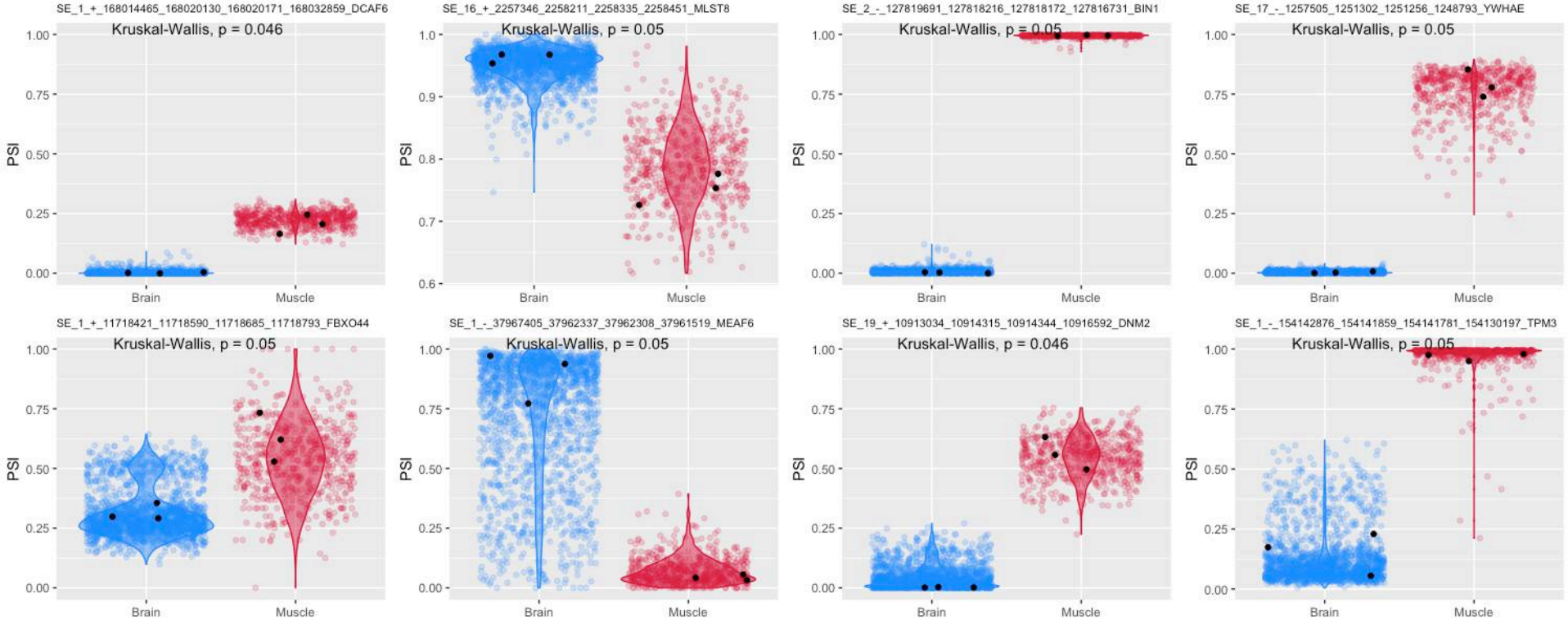
Ex 6:

What if we have only a subset of these samples?
N = 10? N = 3?

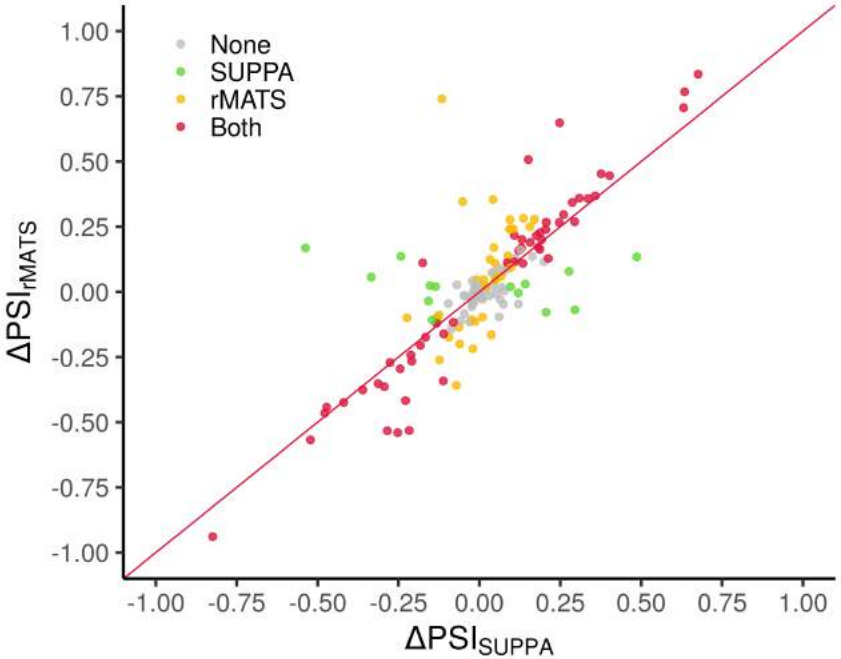
PSI differences (ΔPSI) between conditions: muscle vs. brain

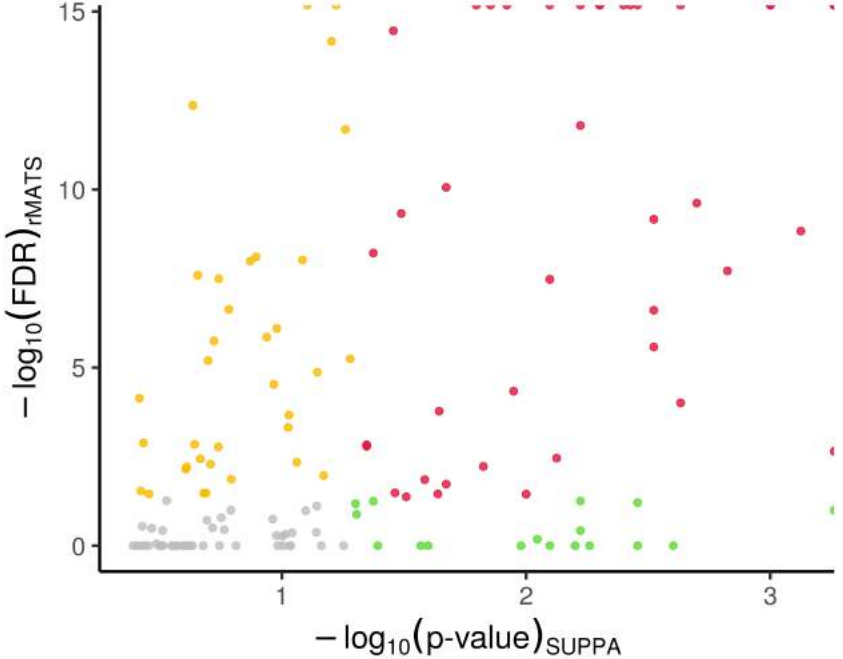


PSI differences (ΔPSI) between conditions: muscle vs. brain



Differential splicing tools provide discrepant significance results





Trincado, J. L. et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol 19, 40 (2018).
Shen, S. et al. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc National Acad Sci 111, E5593–E5601 (2014).

Differential splicing analysis with small sample size

Most tools overlook this!

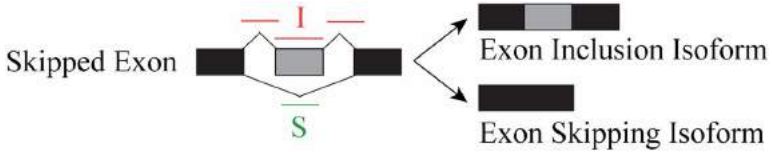
Compromise between modelling the **estimation uncertainty in individual samples** and accounting for **variability among replicates**.

Differential splicing analysis with small sample size

Most tools overlook this! But not all!

rMATS

Shen et al. (2014), *PNAS* 11:E5593



$$I_{ijk} | \psi_{ijk} \sim \text{Binomial}\left(n = I_{ijk} + S_{ijk}, p = f_i(\psi_{ijk})\right)$$

$$\text{logit}(\psi_{ijk}) \sim \text{Normal}\left(\mu = \text{logit}(\psi_{ij}), \sigma^2 = \sigma_{ij}^2\right)$$

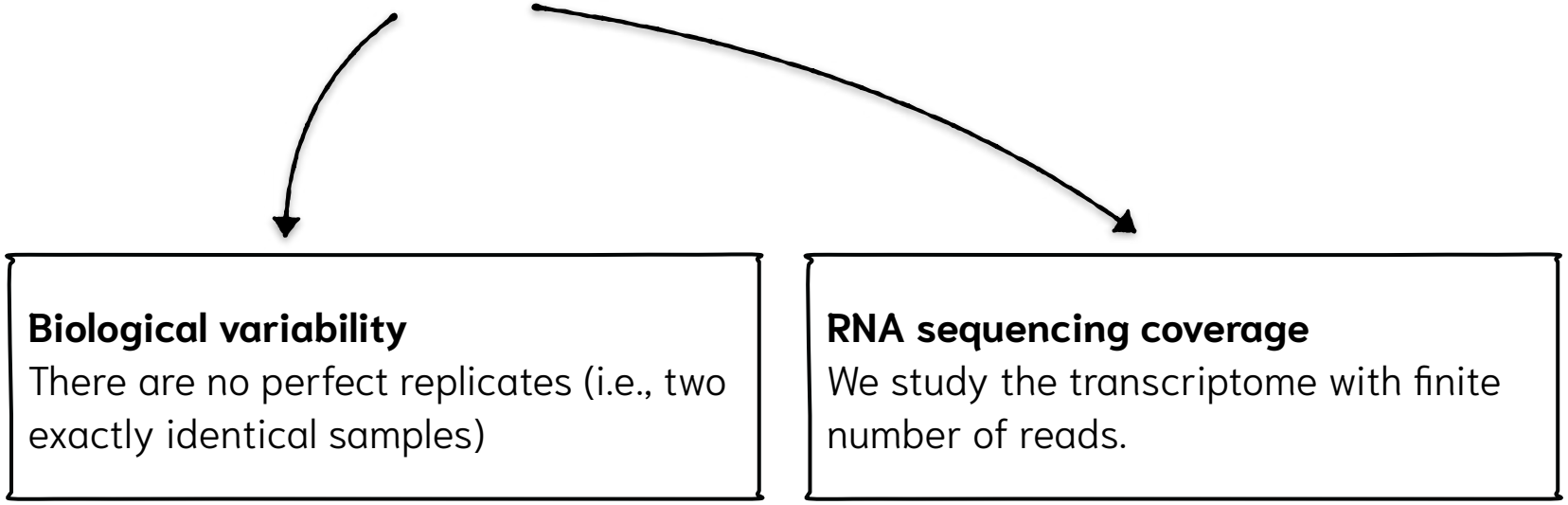
Likelihood Ratio Test: $-2(logL_{|\Psi_{i1}-\Psi_{i2}|\leq c} - logL)$

exon *i*, sample group *j*=1,2, replicate *k*
I_{ijk} - inclusion read counts
S_{ijk} - skipping read counts
ψ_{ijk} - exon inclusion levels
$f_i(\psi_{ijk})$ - length normalization
ψ_{ij} - mean inclusion level of group *j*
σ_{ij} - variance of group *j*
c - user-defined threshold

But:

- Maths not intelligible to non-experts
- Each replicate's contribution to the final stats hidden in the "black box"

Sources of uncertainty in quantification of alternative splicing



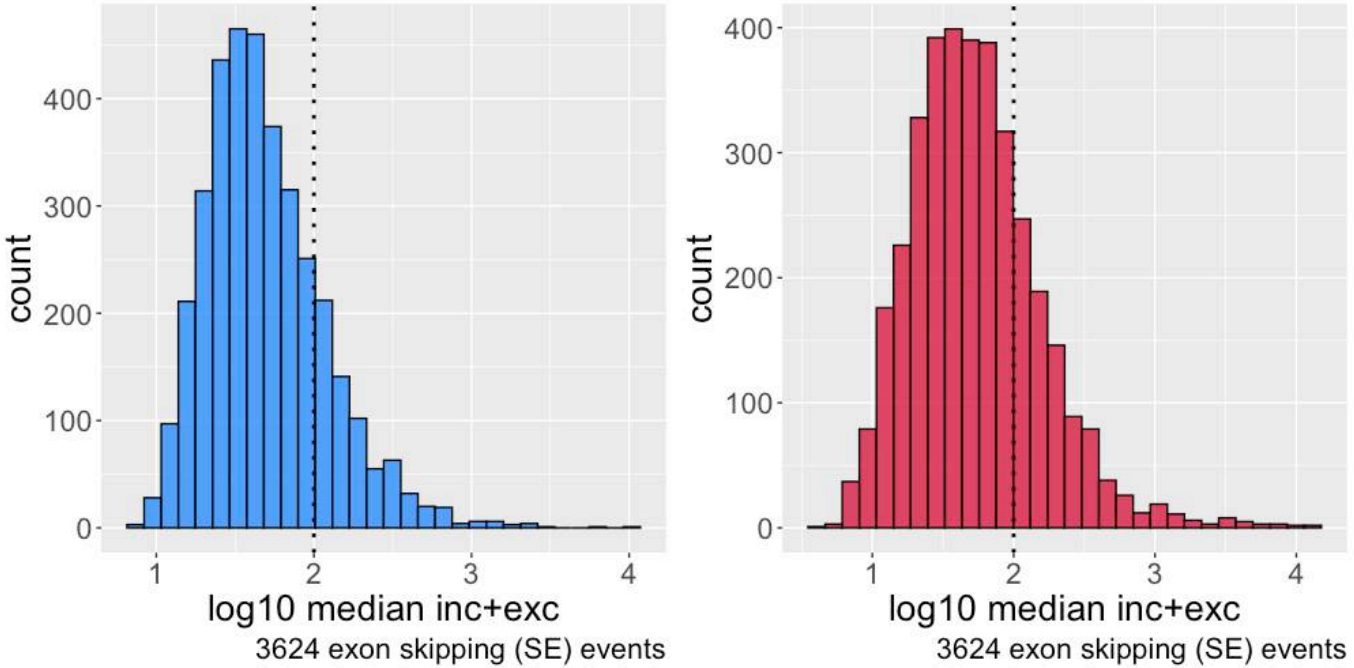
Ideally, all uncertainty would be **biological**

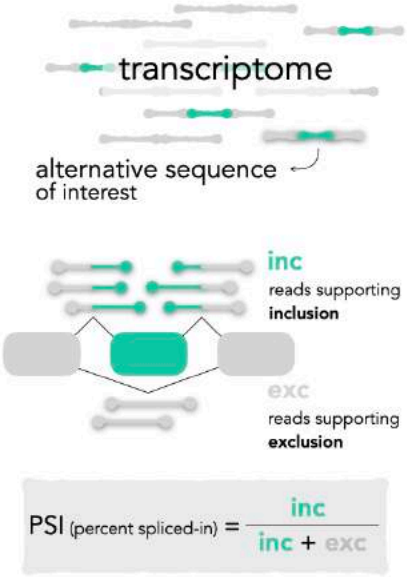
Profiling alternative splicing should robustly **account for this**

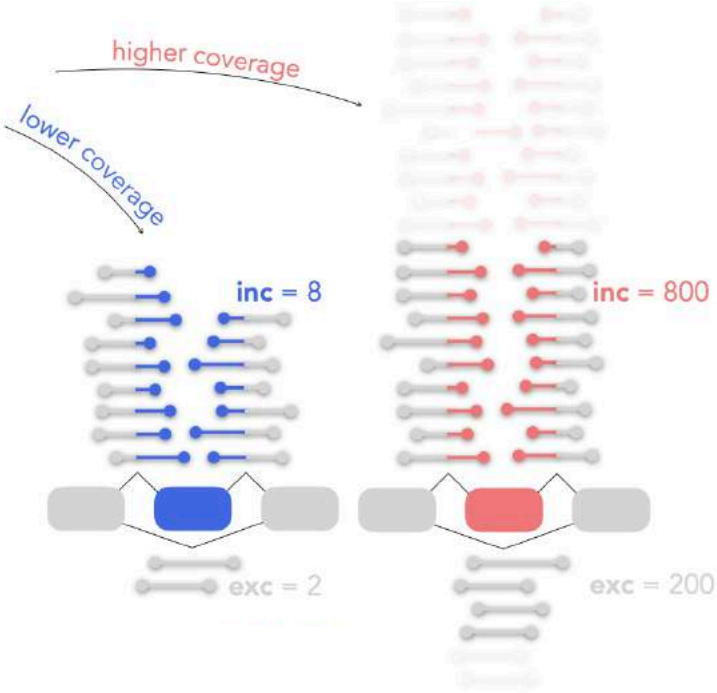
Ex 7:

How does the coverage (inc + exc) distribution look like?
Is it the similar for both tissues?

Explore coverage across conditions: muscle vs. brain

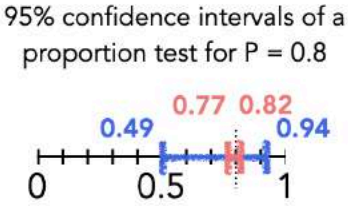


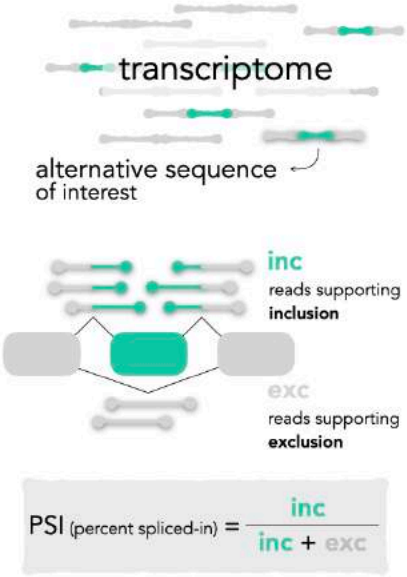


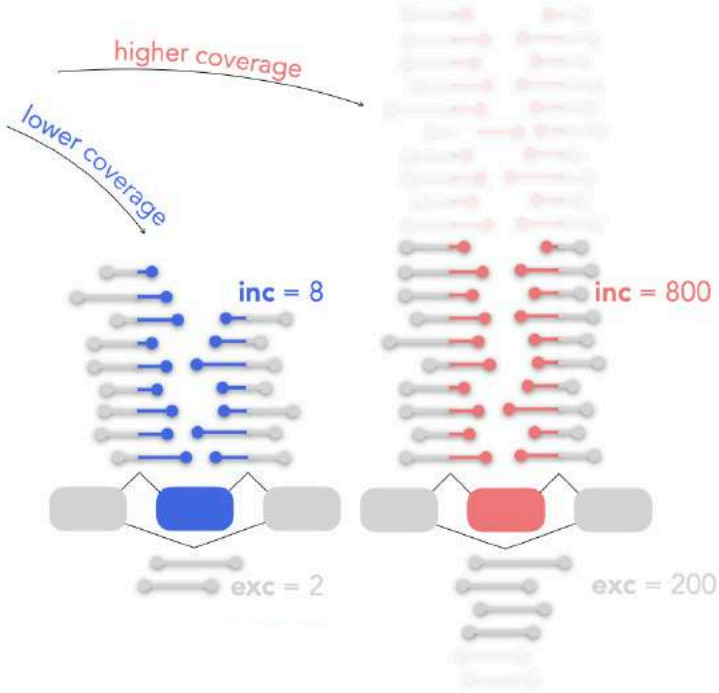


Ex 8:

Simulate PSI-associated confidence intervals from a proportion's test.

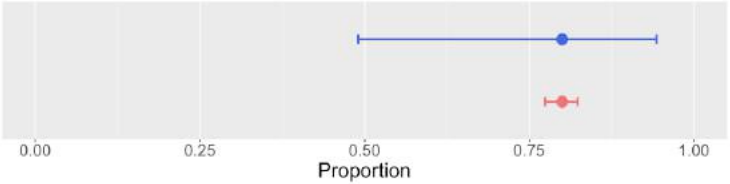


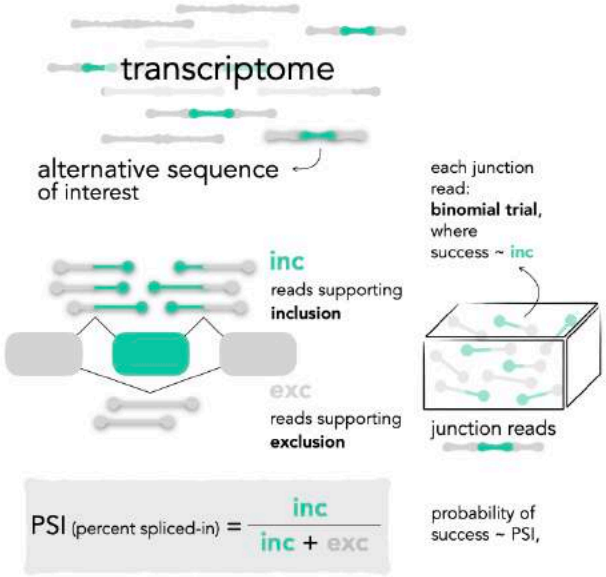




Ex 8:

Simulate PSI-associated confidence intervals from a proportion's test.




transcriptome
alternative sequence of interest
inc
reads supporting inclusion
exc
reads supporting exclusion
each junction read: binomial trial, where success ~ inc
junction reads
PSI (percent spliced-in) = inc / inc + exc
probability of success ~ PSI,

higher coverage
lower coverage
inc = 8
exc = 2
inc = 800
exc = 200
95% confidence intervals of a proportion test for P = 0.8
0.49
0.77
0.82
0.94
0
0.5
1

The Beta Distribution

package: `stats`

Density function	**`dbeta()`**
Distribution function	**`pbeta()`**
Quantile function	**`qbeta()`**
Random generation	**`rbeta()`**

for the Beta distribution with shape parameters **`shape1`** and **`shape2`**

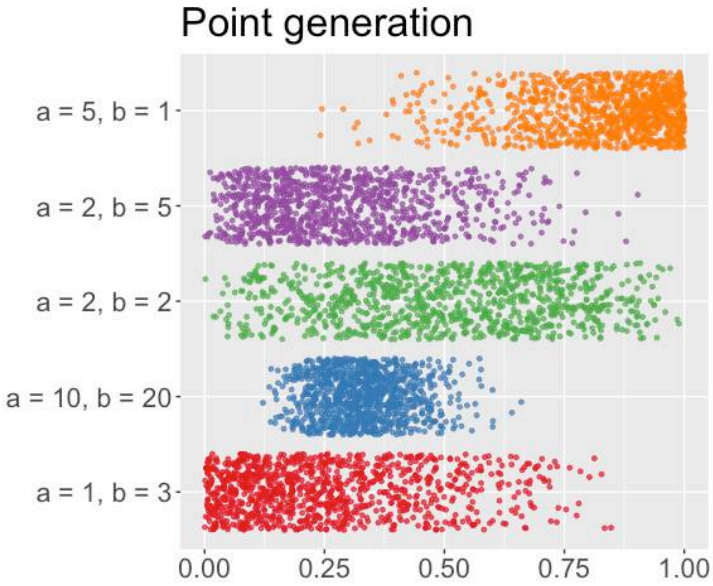
`Ex 9:`

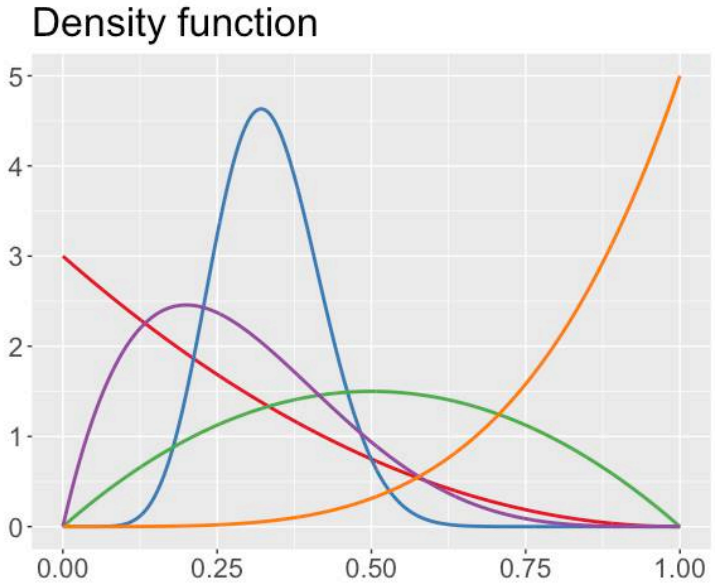
`Explore R stats beta distribution functions to generate beta distributions with different shapes.`

	shape1 = α	**shape2** = β
1	10	20
2	5	1
3	2	5
4	2	2
5	1	3

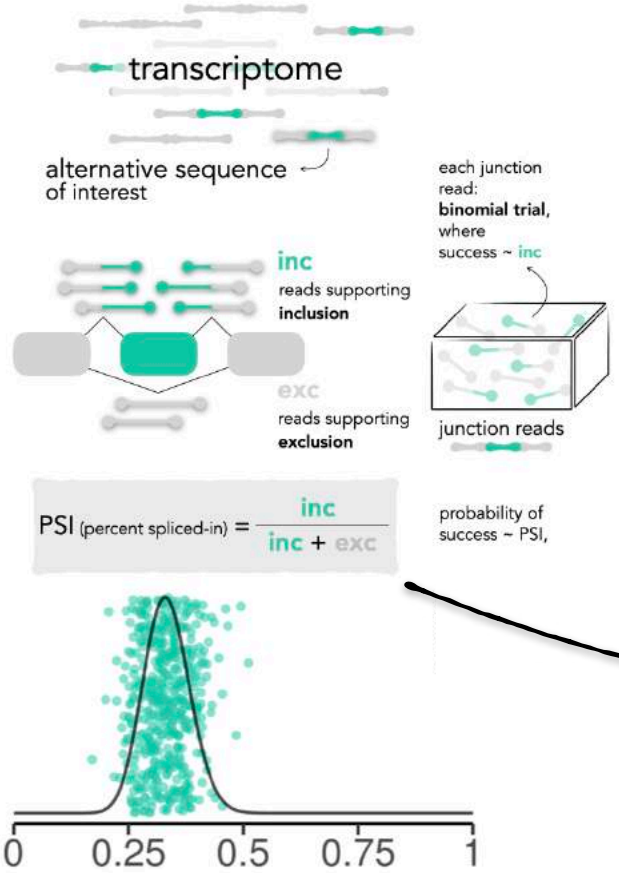
The Beta Distribution

package: `stats`





Wikipedia: “Beta distribution”

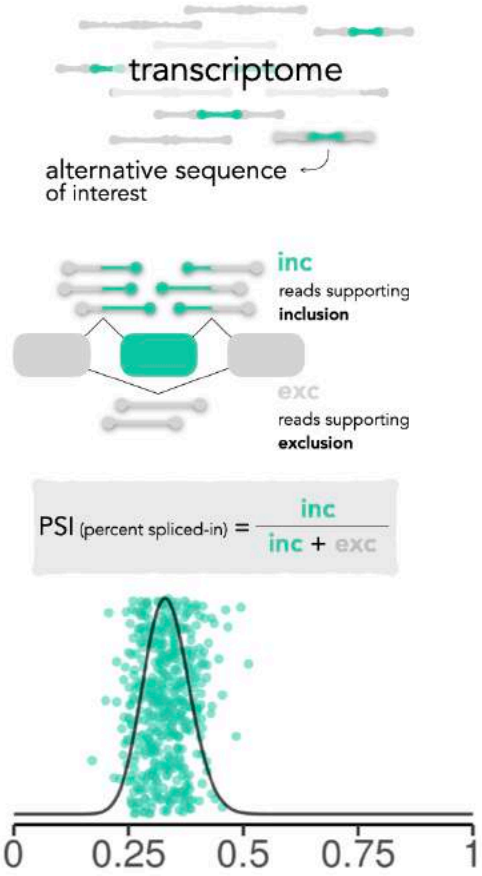


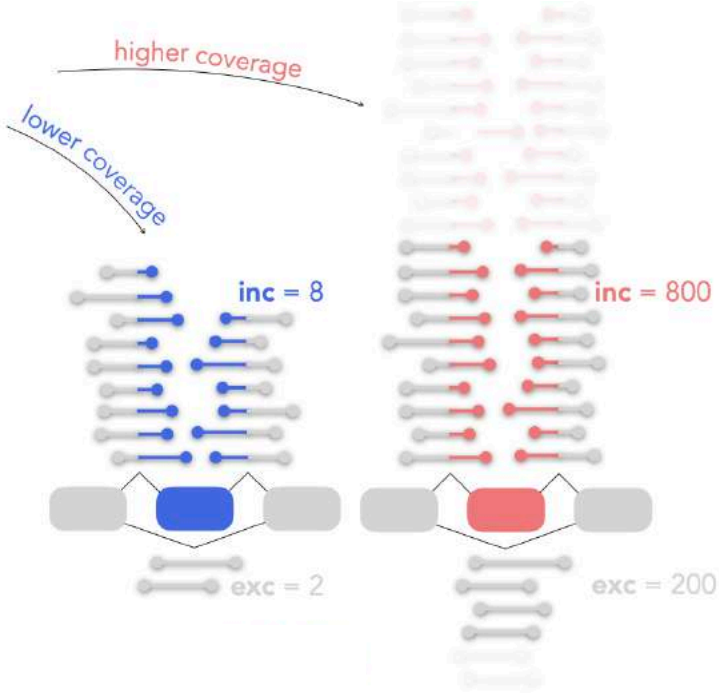
The Beta Distribution

- model **proportions**
- parameters **α** and **β** control the distribution shape
- ratio **α/(α+β)** defines the distribution mean
- get narrower as **α** and **β** increase

PSI ~ **α/(α+β)**

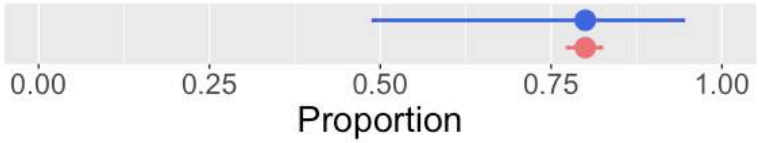
PSI can be modelled by a beta distribution with shape parameters such that the **mean ~ PSI**

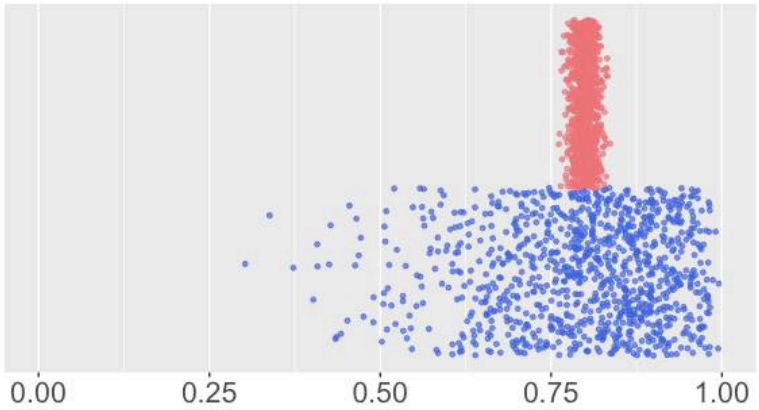


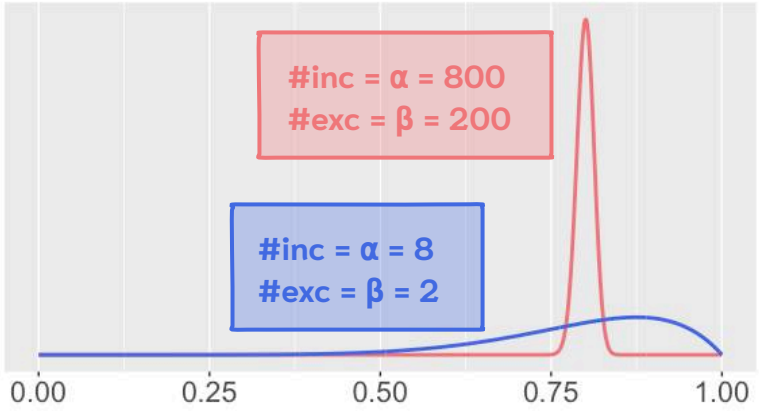


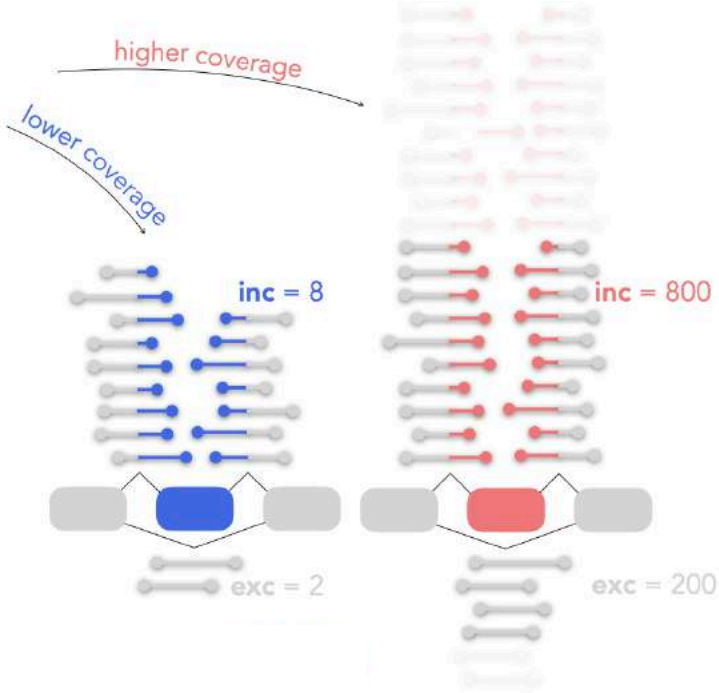
Ex 10:

Explore R stats beta distribution functions to generate beta distributions with different shapes.





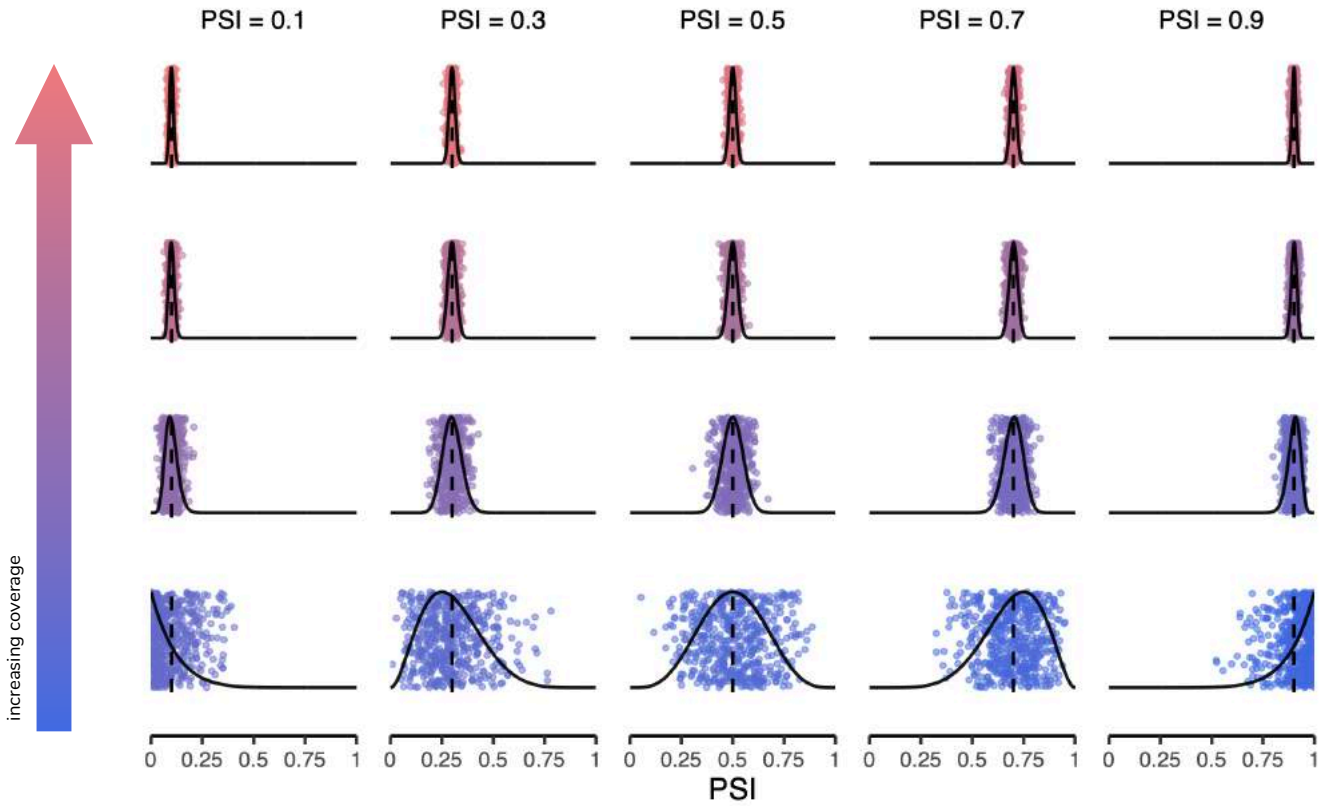


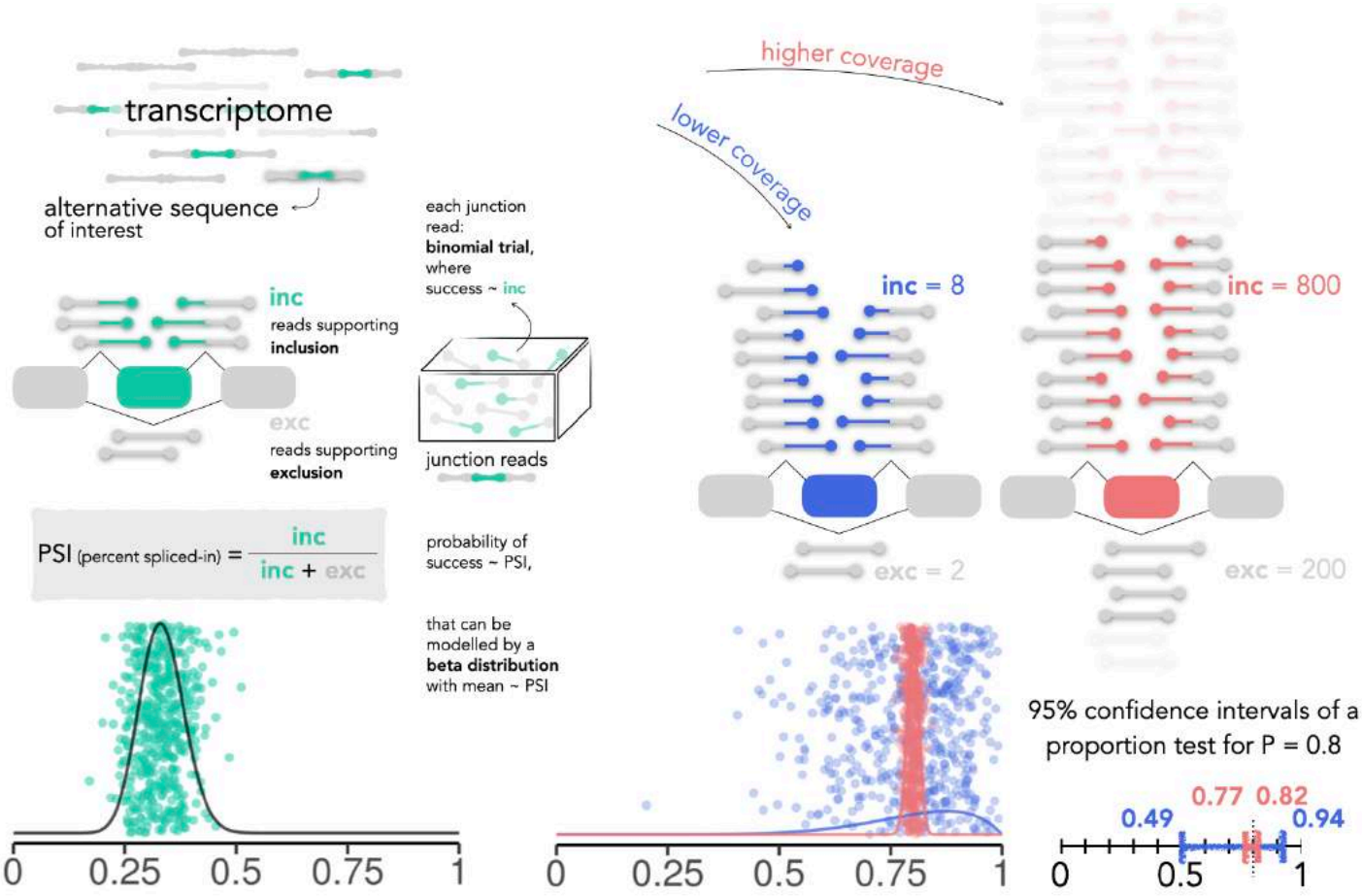


Ex 10:

Explore R stats beta distribution functions to generate beta distributions with different shapes.

Alternative splicing quantification using beta distributions



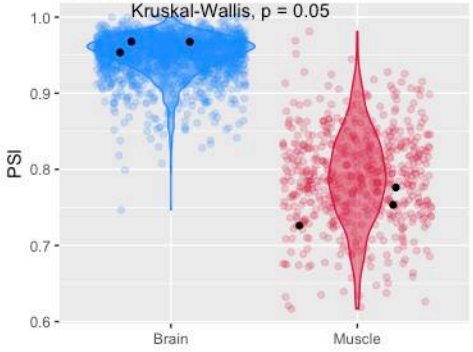

transcriptome
alternative sequence of interest
inc
reads supporting inclusion
exc
reads supporting exclusion
each junction read: binomial trial, where success ~ inc
junction reads
PSI (percent spliced-in) = inc / inc + exc
probability of success ~ PSI,
that can be modelled by a beta distribution with mean ~ PSI
0
0.25
0.5
0.75
1
higher coverage
lower coverage
inc = 8
exc = 2
inc = 800
exc = 200
95% confidence intervals of a proportion test for P = 0.8
0.49
0.77
0.82
0.94
0
0.5
1

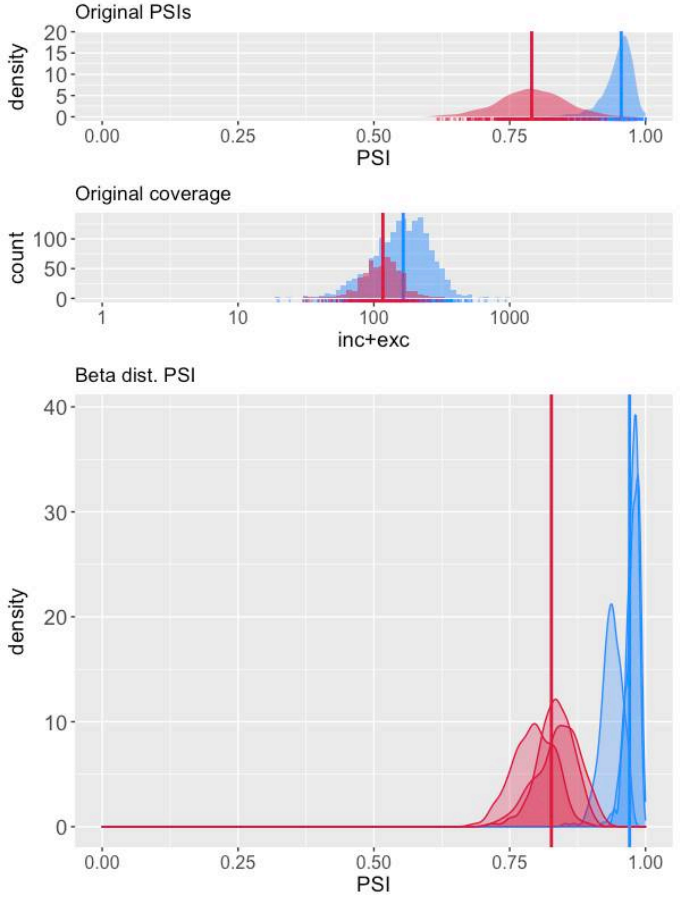
Session 2

Differential splicing analyses

Modelling the estimation uncertainty in individual samples while accounting for variability among replicates using beta distributions

PSI differences (ΔPSI) between conditions: muscle vs. brain





Differential splicing quantification using beta distributions

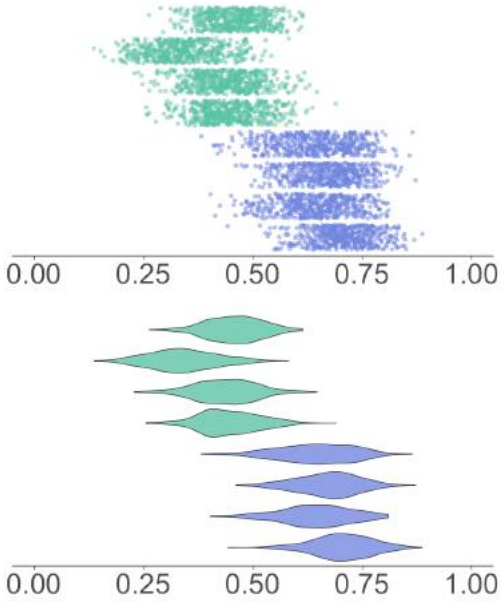
- for each considered alternative splicing event:

1 emit individual distributions per sample, with **inc** and **exc** as shape parameters

2 considering the experimental design, gather all emitted points per condition

condition A

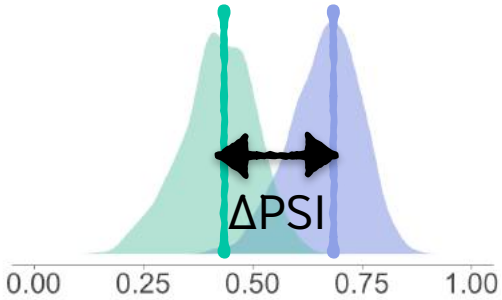
condition B



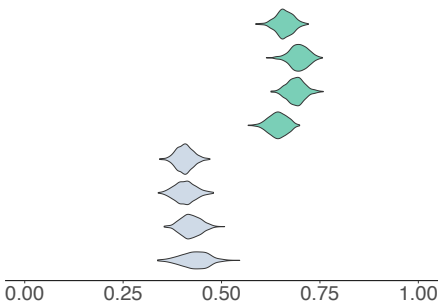
3 median of grouped beta distributions ~ group PSI

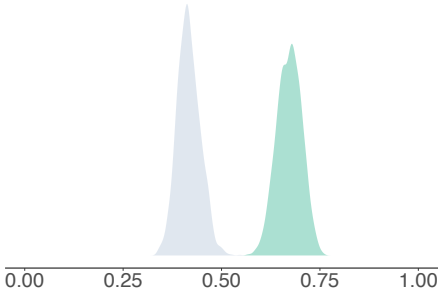
ΔPSI: estimated effect size

$\mathbf{P_{diff}}$: proportion of points A > points B
~ probability that random point A > random point B



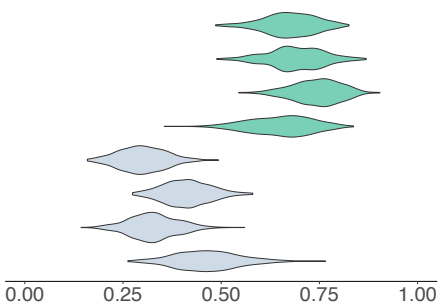
Differential splicing quantification using beta distributions

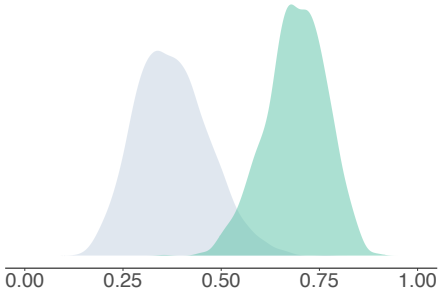




ΔPSI

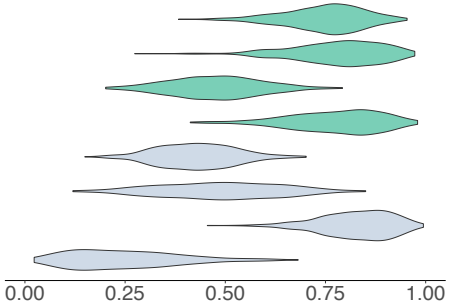
P(**B** > **A**) ~ 1

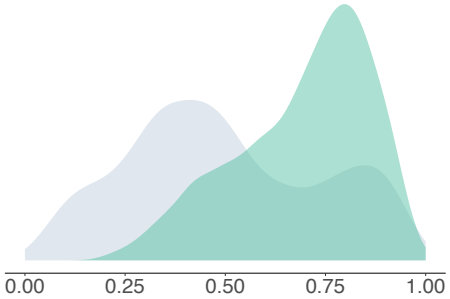




ΔPSI

P(**B** > **A**) ~ 0.95





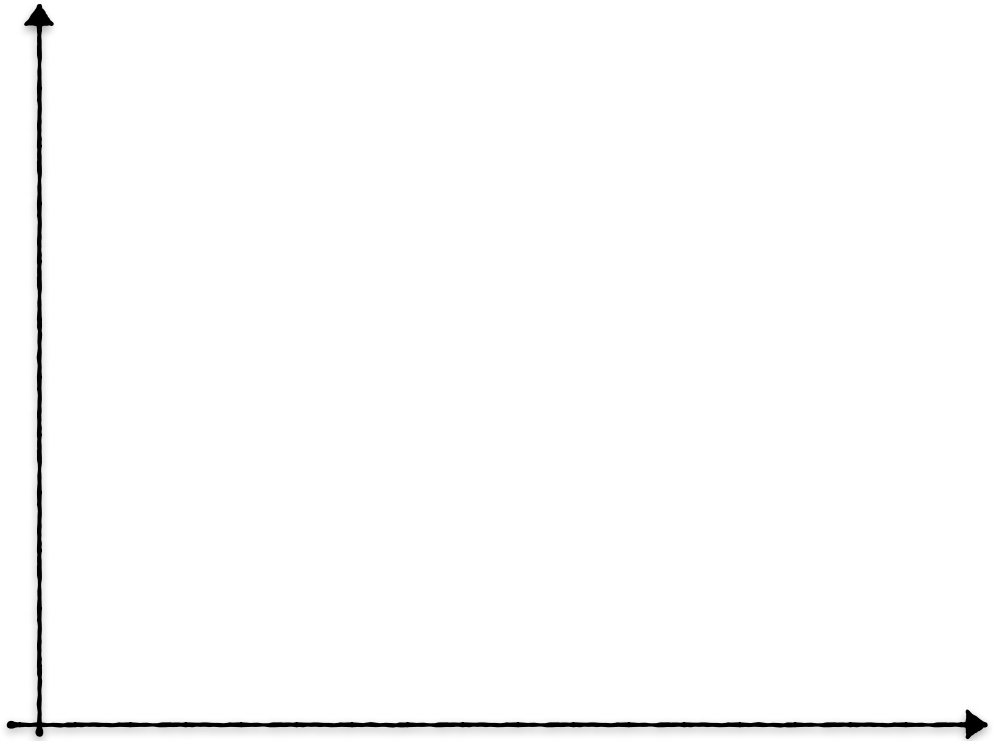
ΔPSI

P(**B** > **A**) ~ 0.75

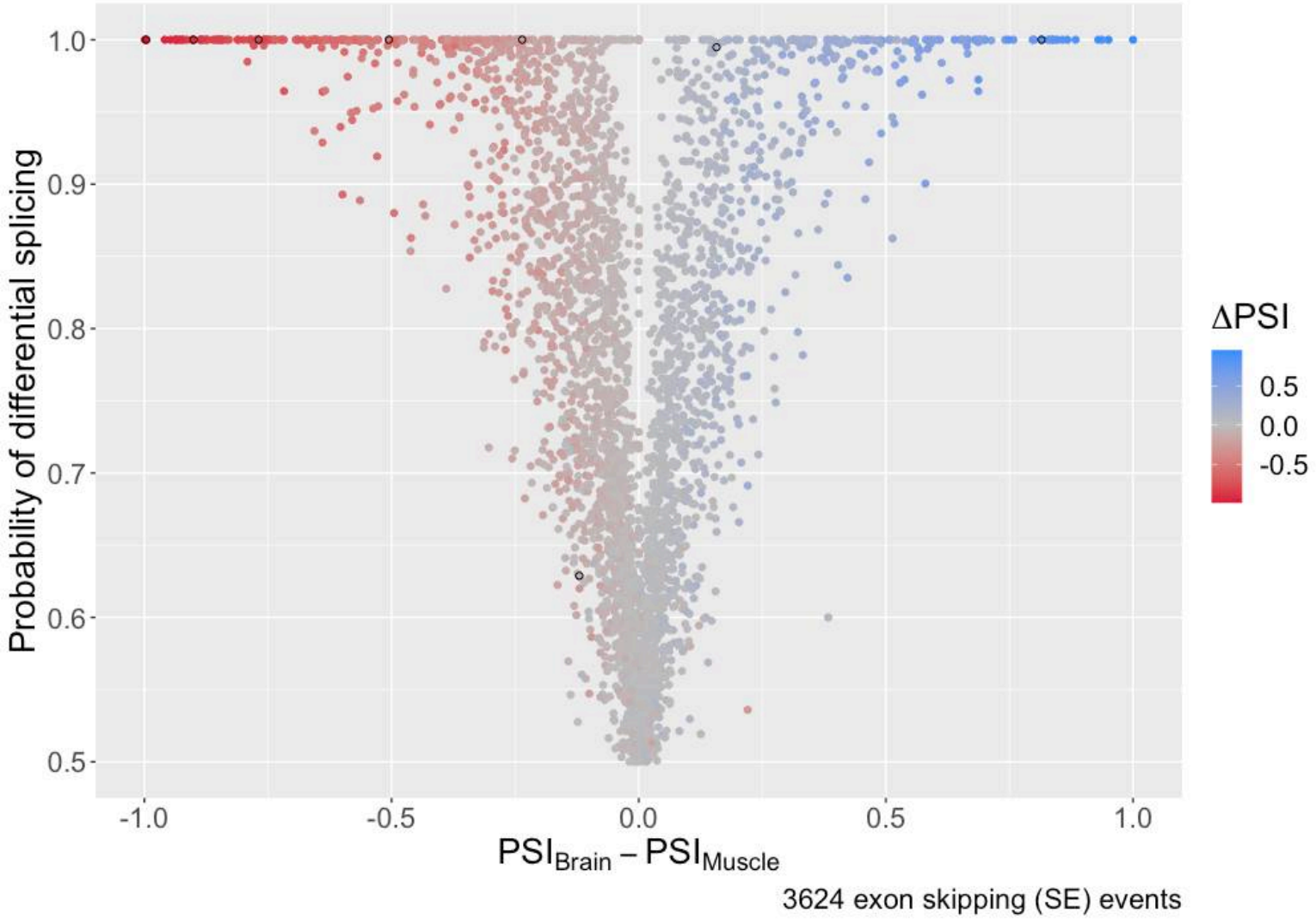
Differential splicing quantification using beta distributions

Probability of differential splicing

P(Brain > Muscle)

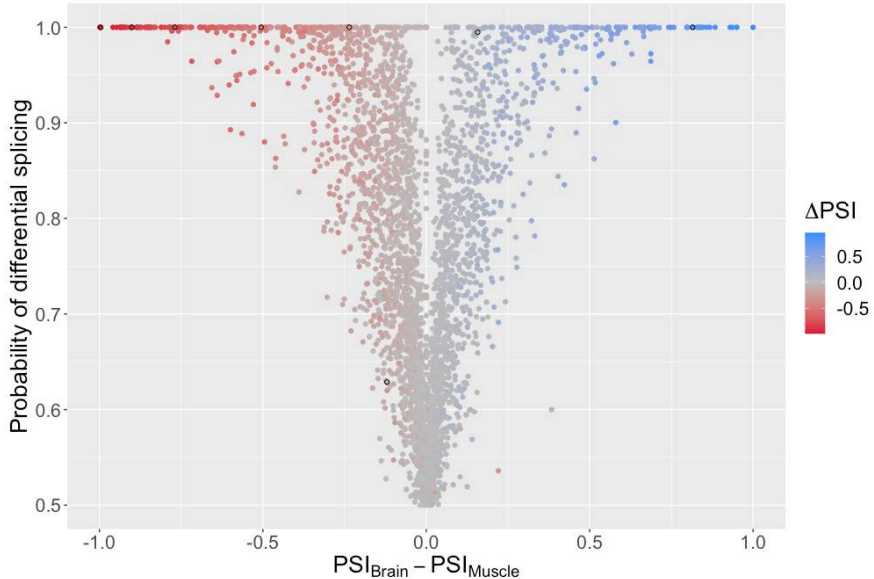


ΔPSI


Probability of differential splicing
1.0
0.9
0.8
0.7
0.6
0.5
-1.0
-0.5
0.0
0.5
1.0
PSI Brain − PSI Muscle
3624 exon skipping (SE) events
ΔPSI
0.5
0.0
-0.5

So… what are the most differentially spliced events?

$$F = \frac{\text{variation between sample means}}{\text{variation within the samples}}$$



Differential splicing quantification using beta distributions

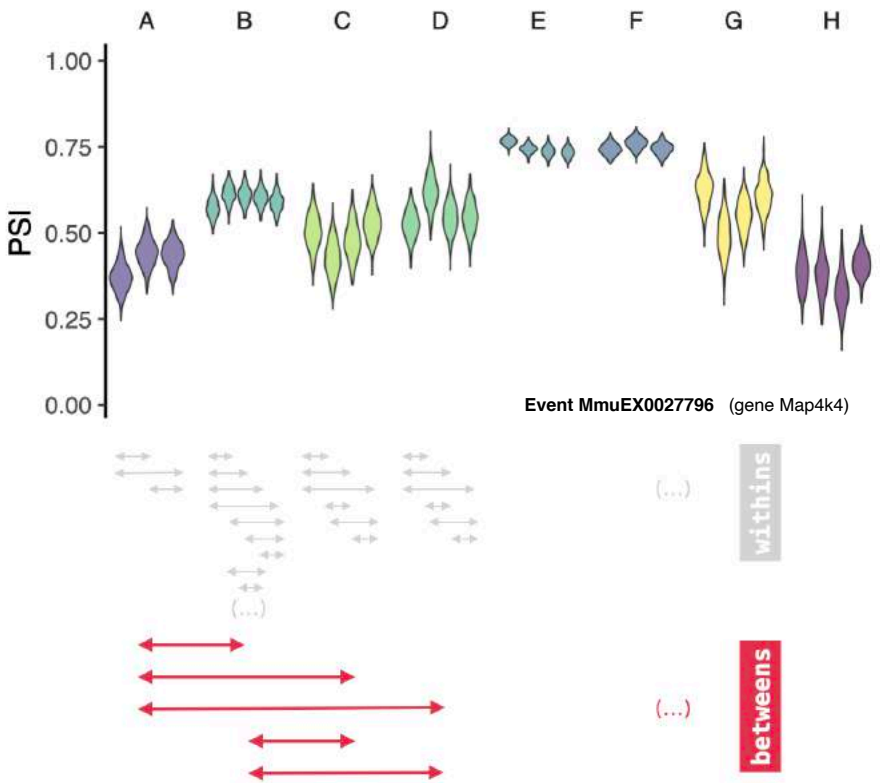
- inspired in the ANOVA approach
  for the analysis of variance:
  a measure of dispersion

- beta distribution points:
  a measure of PSI dispersion
  (that depends on coverage)

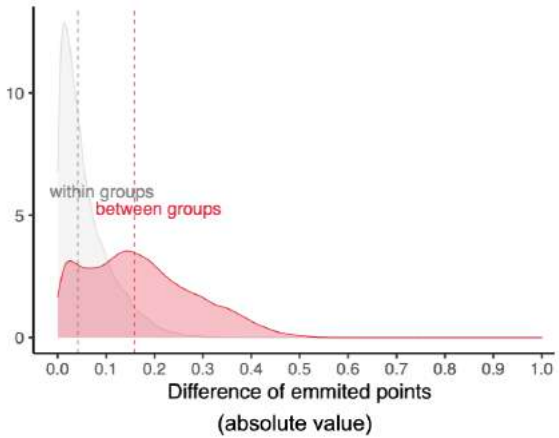
for each event of interest:

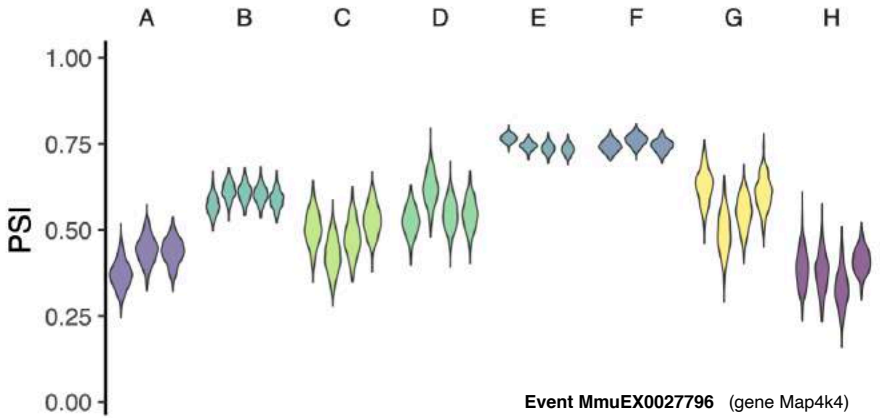
- per group:
`withins`
set of differences between samples

- overall:
`betweens`
set of differences between groups

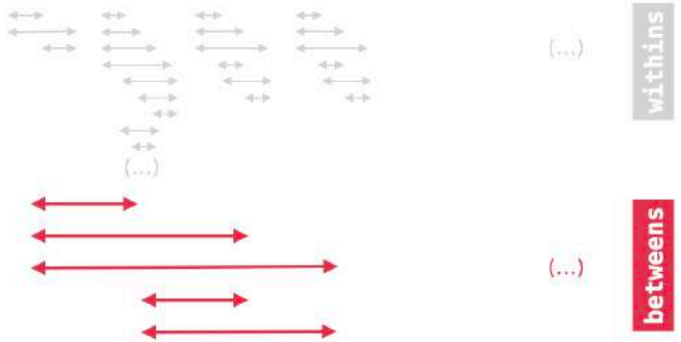


Differential splicing quantification using beta distributions



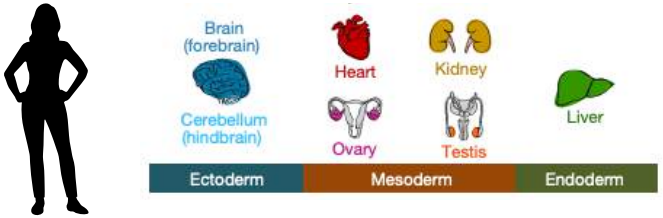


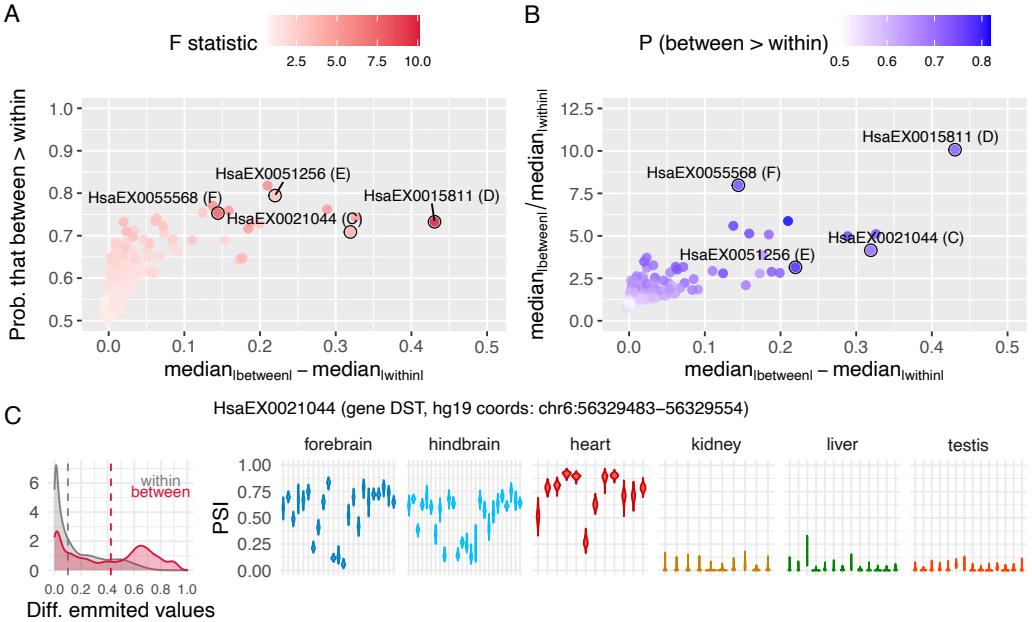
$$F = \frac{\text{variation between sample means}}{\text{variation within the samples}} = \frac{\text{median |between|}}{\text{median |within|}}$$



Differential splicing quantification using beta distributions

- applied into a dataset of human **multi tissue** transcriptomes





$$F = \frac{\text{variation between sample means}}{\text{variation within the samples}} = \frac{\text{median |between|}}{\text{median |within|}}$$

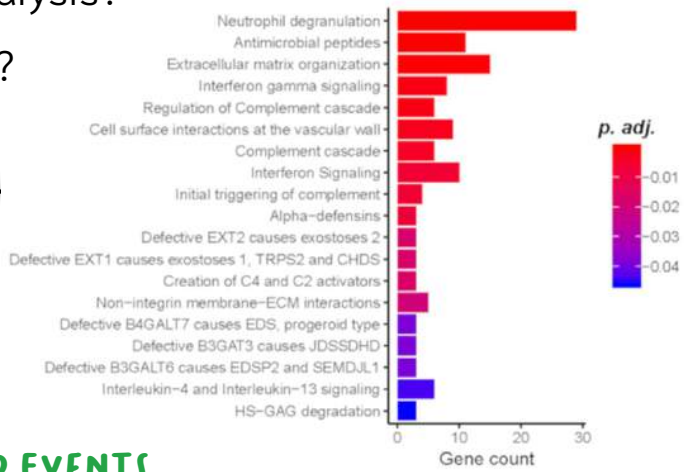
What else?

Exploring functional enrichment with differential splicing

Gene set enrichment analysis?

Gene ontology?

ranked gene list



SPLICED EVENTS

List of differentially ~~expressed genes~~:

- ranked by a statistic
- adjusted in some way for multiple comparison

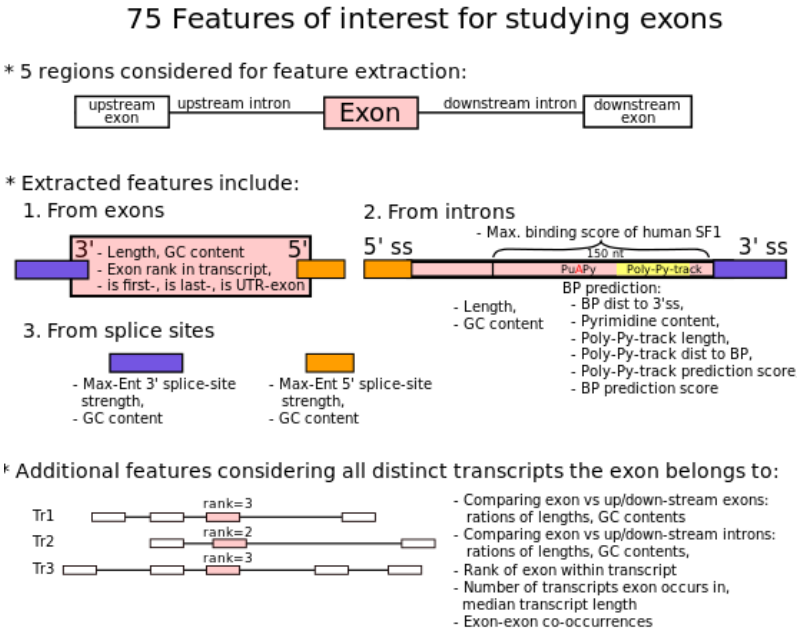
But... not that simple:

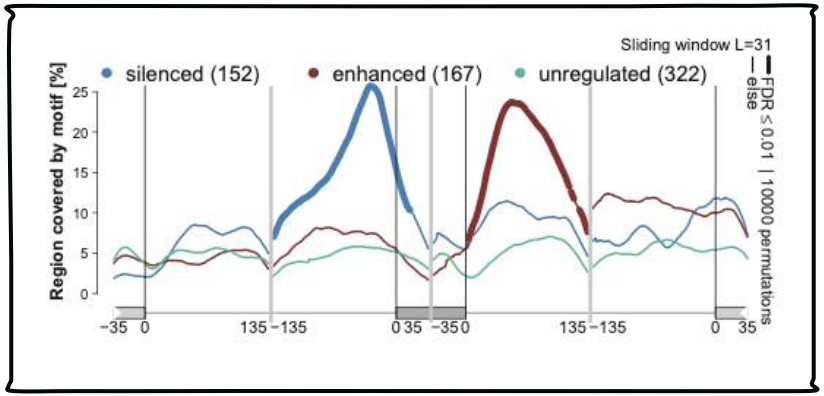
- Splicing differences are more precisely quantified in **highly expressed** genes (e.g. tissue-specific genes) – enriched in specific pathways?
- Some genes have **more splicing events** than others

Difficult to define an unbiased **background**

Exploring functional enrichment with differential splicing

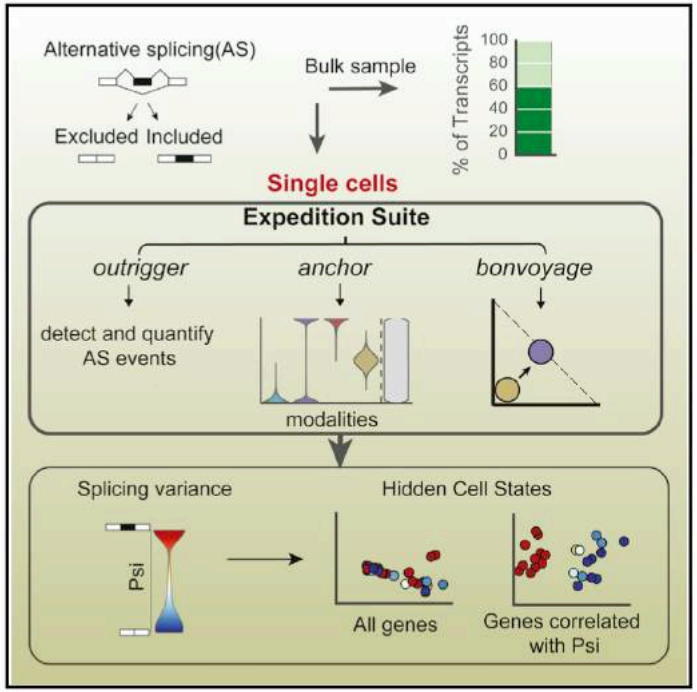
Matt: toolkit for feature analysis of alternative splicing events



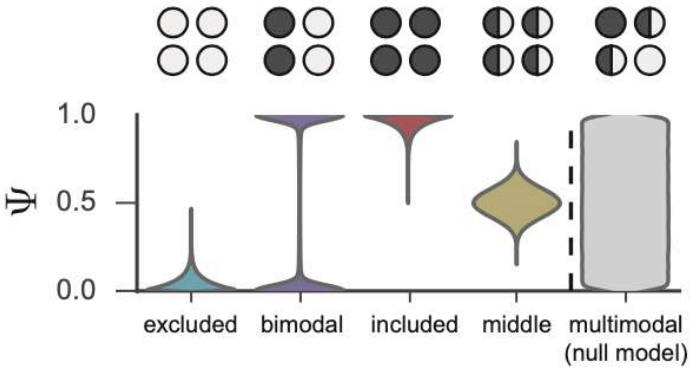


Gohr, A. & Irimia, M. Matt: Unix tools for alternative splicing analysis. Bioinformatics 35, 130–132 (2018).

Alternative splicing analysis with single-cell RNA-seq



Distribution of PSI for each AS event can be modelled as a Beta probability distribution parameterised by a and b:



Song, Y. et al. Single-Cell Alternative Splicing Analysis with Expedition Reveals Splicing Dynamics during Neuron Differentiation. Mol Cell 67, 148–161.e5 (2017).