Studying the transcriptome using RNA-seq

Cecilia Coimbra Klein

IBUB
Institut de Biomedicina
de la Universitat de Barcelona

UNIVERSITAT DE BARCELONA

CRG
Centre
for Genomic
Regulation

UVIC UNIVERSITAT DE VIC
UNIVERSITAT CENTRAL DE CATALUNYA

Master in Omics
Data Analysis

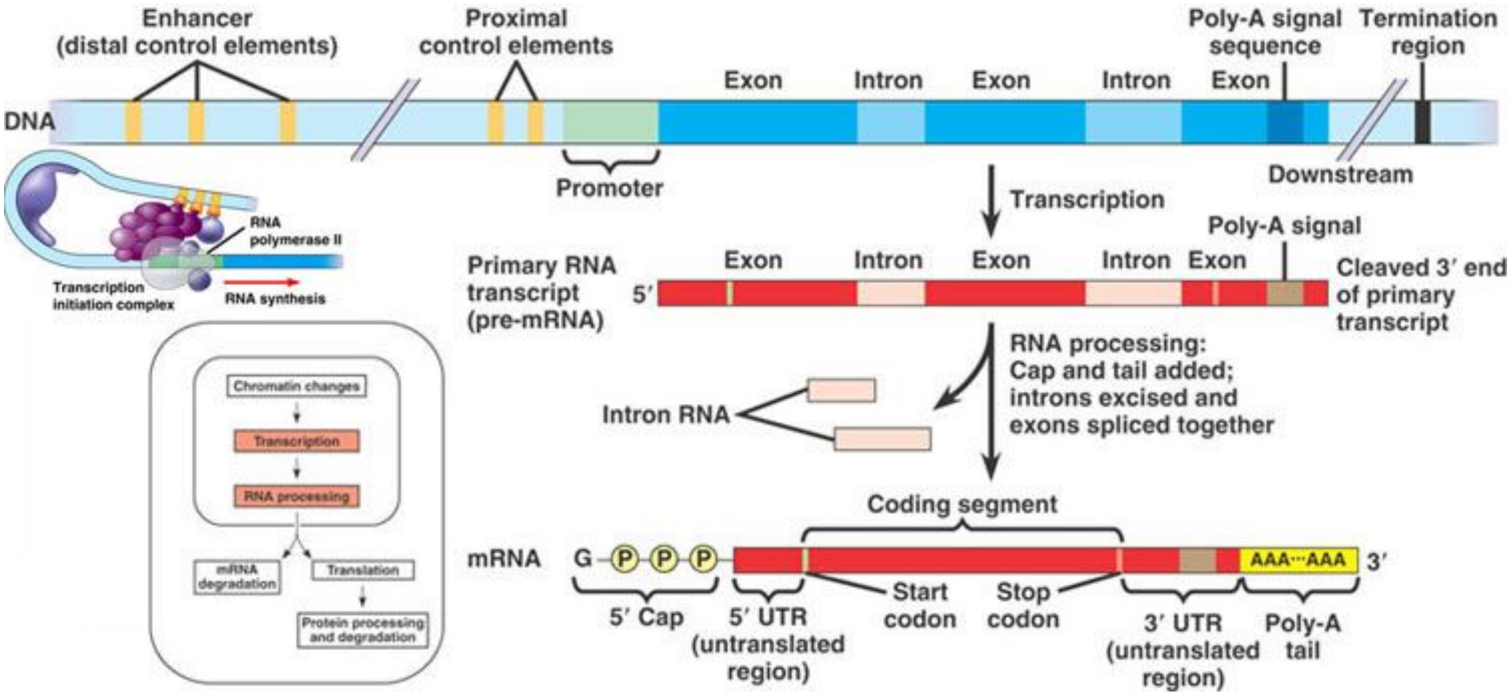
Outline

Outline

1. Introduction
2. Basic concepts
3. Short-read RNA-seq data processing
4. Gene level RNA-seq data analysis
5. Isoform level RNA-seq analyses
    - 5.1. AS events from genomic annotation
    - 5.2. PSI values
    - 5.3. Differential splicing analysis
    - 5.4. Functional analysis
6. Regulation of gene expression

Alternative splicing

RNA transcription and processing



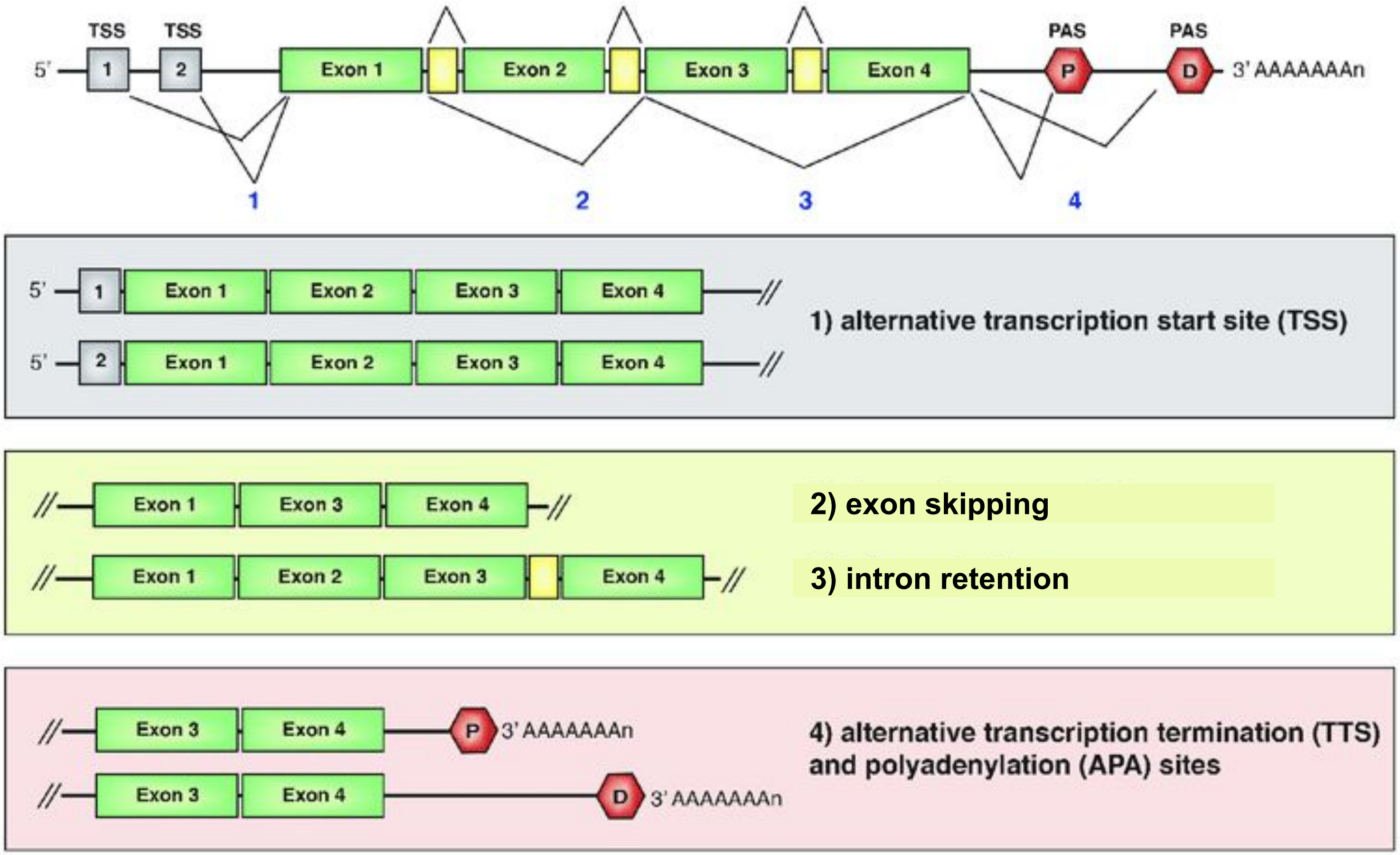
Primary RNA transcripts are extensively processed: capping, splicing, polyadenylation, editing

This process is highly regulated and results in a gene producing many distinct transcript isoforms: one gene, many transcripts

The transcriptome is distinct from and more complex than the genome

The transcriptome cannot be predicted from the genome sequence alone: it must be measured

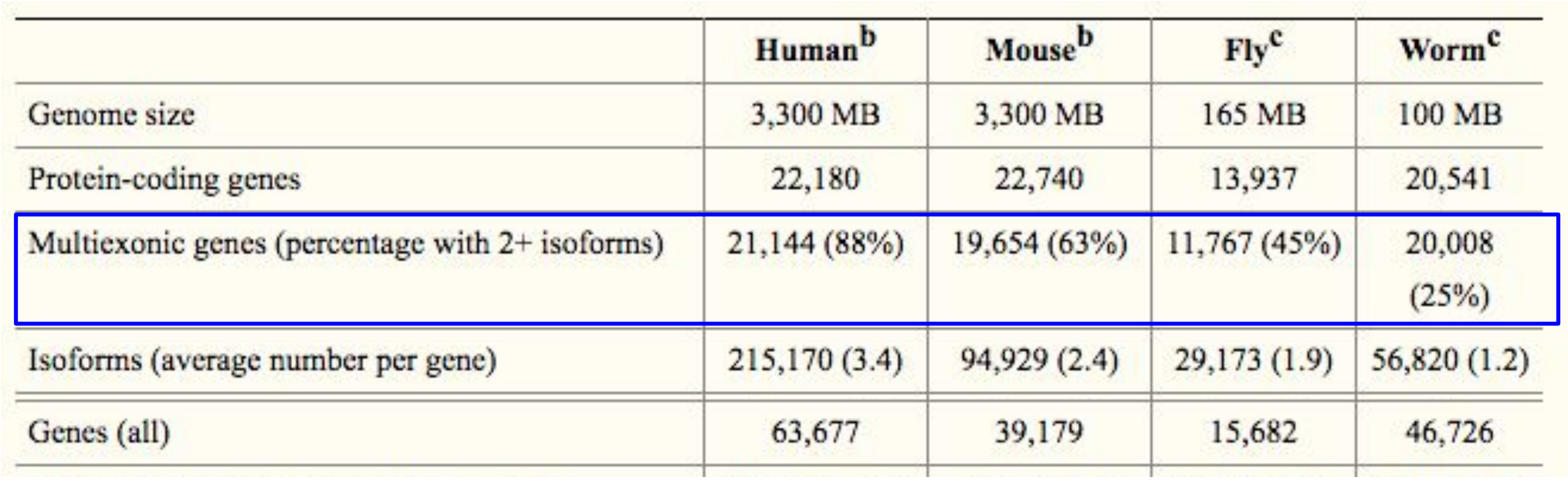
Complexity arising from differential processing



These processing events can result in different protein products, differentially (post-) transcriptionally regulated mRNAs or non-protein coding isoforms.

Andreassi, C. *et al.* (2018). doi: 10.3389/fnmol.2018.00304

Complexity arising from differential processing

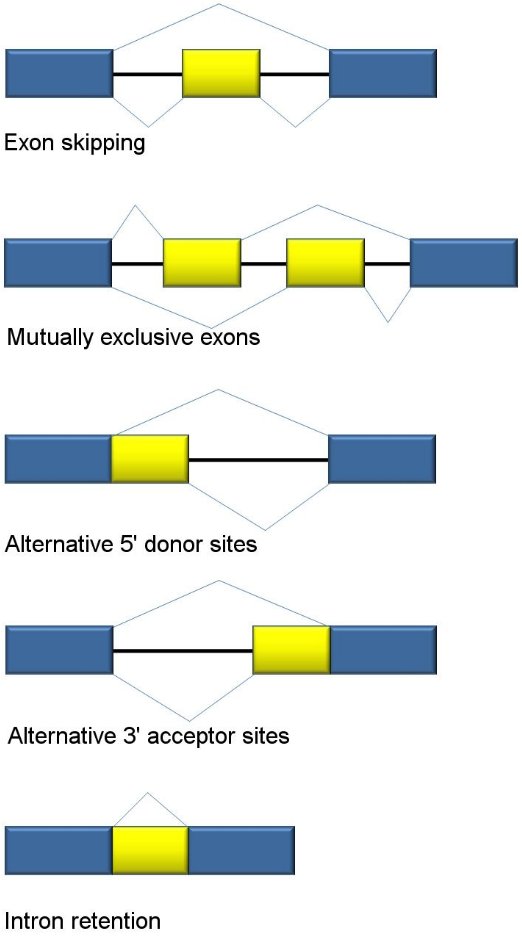


	Human[b]	Mouse[b]	Fly[c]	Worm[c]
Genome size	3,300 MB	3,300 MB	165 MB	100 MB
Protein-coding genes	22,180	22,740	13,937	20,541
Multiexonic genes (percentage with 2+ isoforms)	21,144 (88%)	19,654 (63%)	11,767 (45%)	20,008 (25%)
Isoforms (average number per gene)	215,170 (3.4)	94,929 (2.4)	29,173 (1.9)	56,820 (1.2)
Genes (all)	63,677	39,179	15,682	46,726

- pre-mRNA splicing scales with organismal complexity.
- Alternative pre-mRNA splicing occurs in ~88% of human genes, compared with ~63% of mouse genes.
- More recent deep RNA-seq data, 95% to 100% of human genes may encode two or more (2+) isoforms
- One function of alternative splicing is to significantly expand the form and function of the human proteome

Lee & Rio (2015). doi:10.1146/annurev-biochem-060614-034316

Modes of AS



Exon skipping

Mutually exclusive exons

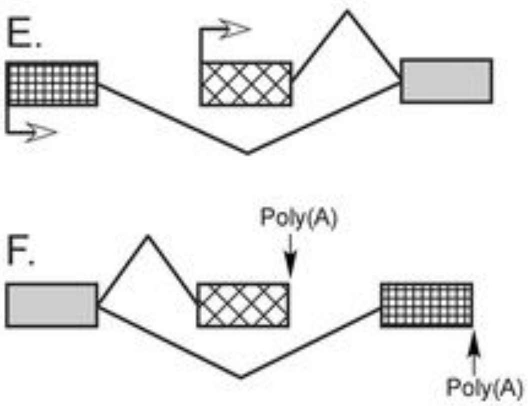
Alternative 5' donor sites

Alternative 3' acceptor sites

Intron retention

Exons are represented as blue and yellow blocks, introns as lines in between.

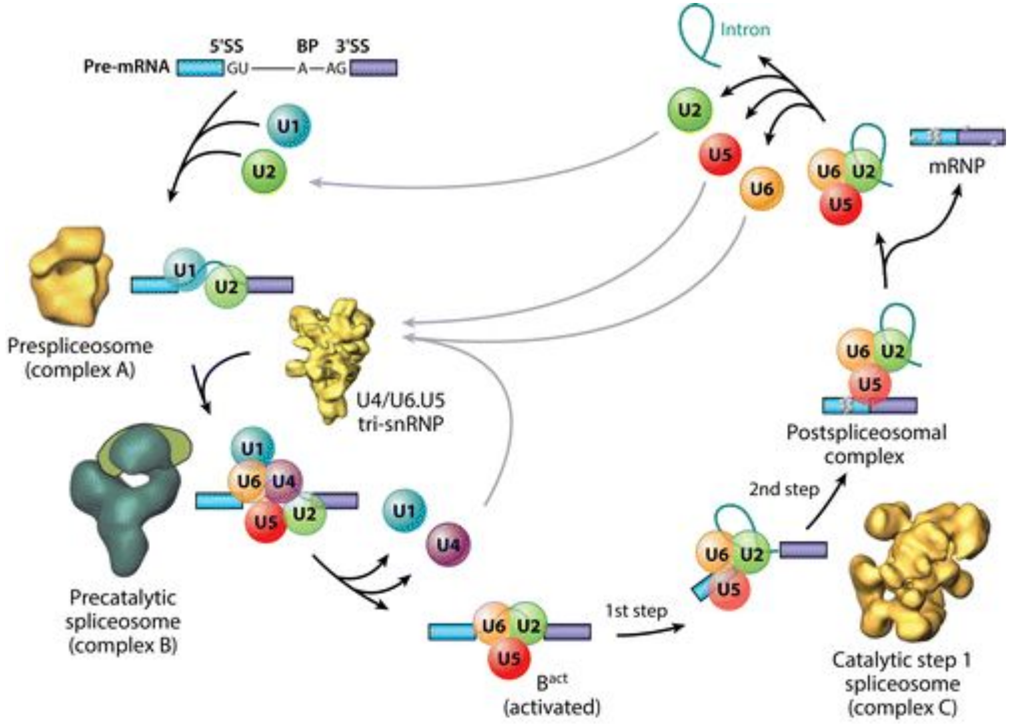
Alternative promoters and polyadenylation sites



Alternative promoters are primarily an issue of transcriptional control. Control of polyadenylation appears mechanistically similar to control of splicing.Both of these mechanisms are found in combination with alternative splicing and provide additional variety in mRNAs derived from a gene

Black (2003) doi: 10.1146/annurev.biochem.72.121801.161720
https://en.wikipedia.org/wiki/Alternative_splicing

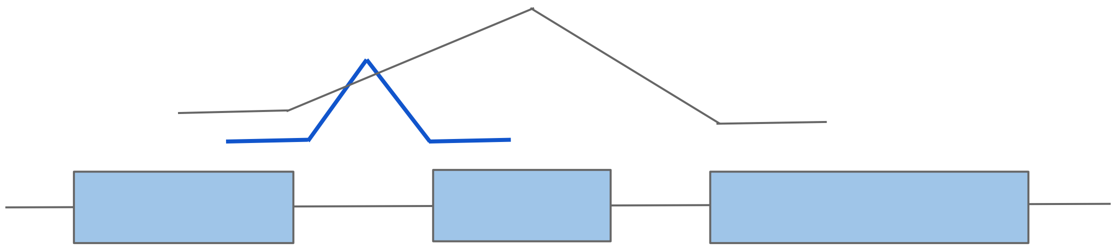
General splicing mechanism



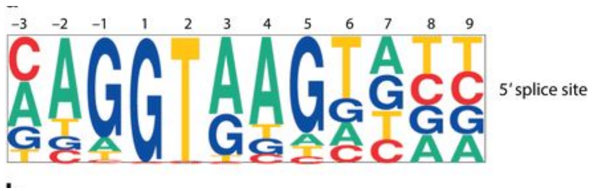
Lee Y, Rio DC. 2015.
Annu. Rev. Biochem. 84:291–323

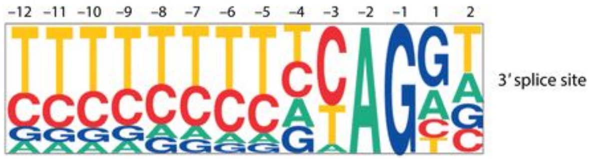
Lee & Rio (2015). doi:10.1146/annurev-biochem-060614-034316

Junctions



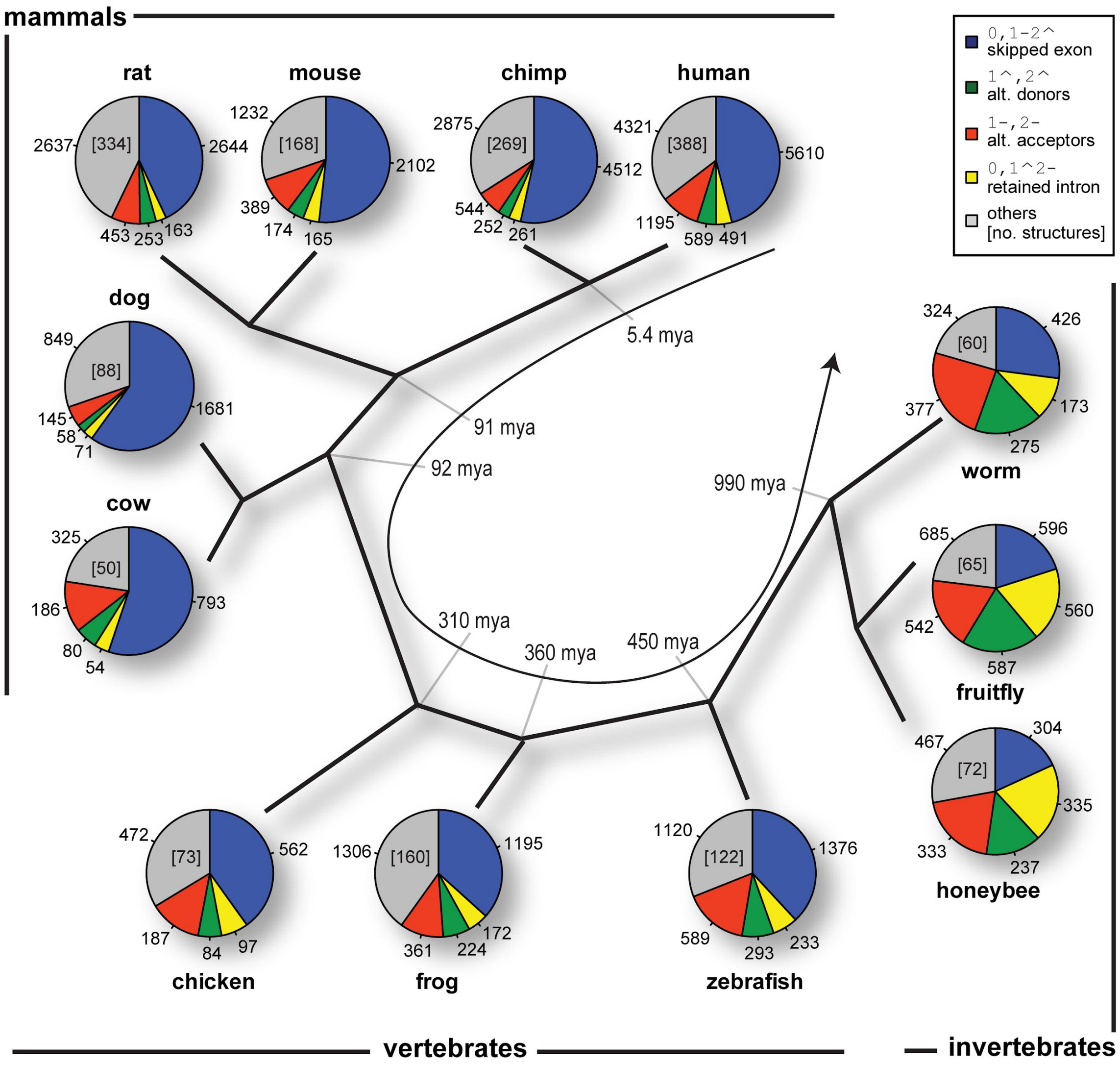
Splice sites in the human genome:





Lee Y, Rio DC. 2015.
Annu. Rev. Biochem. 84:291–323

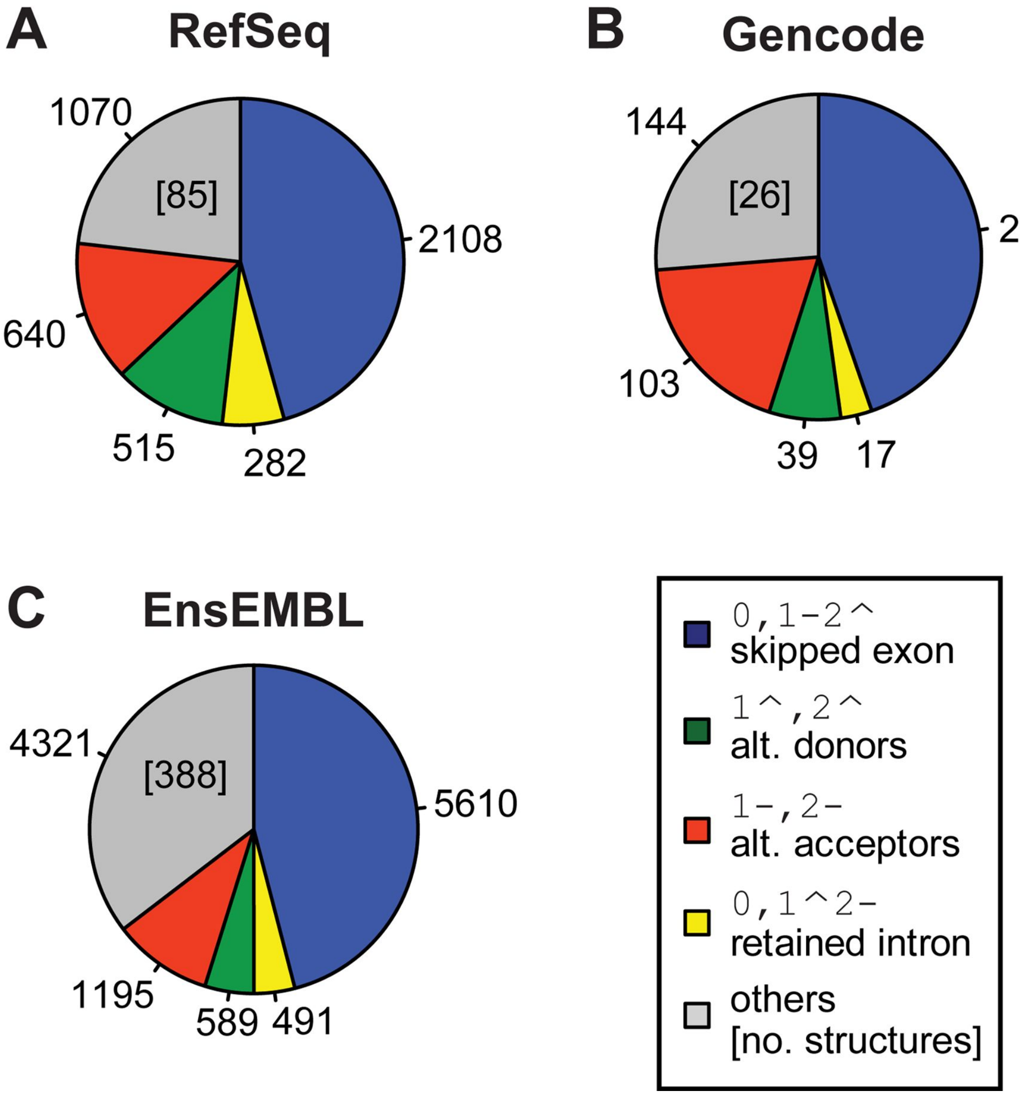
Comparative genomics of the AS landscape in 12 metazoa



Sammeth, Foissac , Guigó (2008) PLoS Comput Biol 4(8): e1000147

Cecilia Coimbra Klein

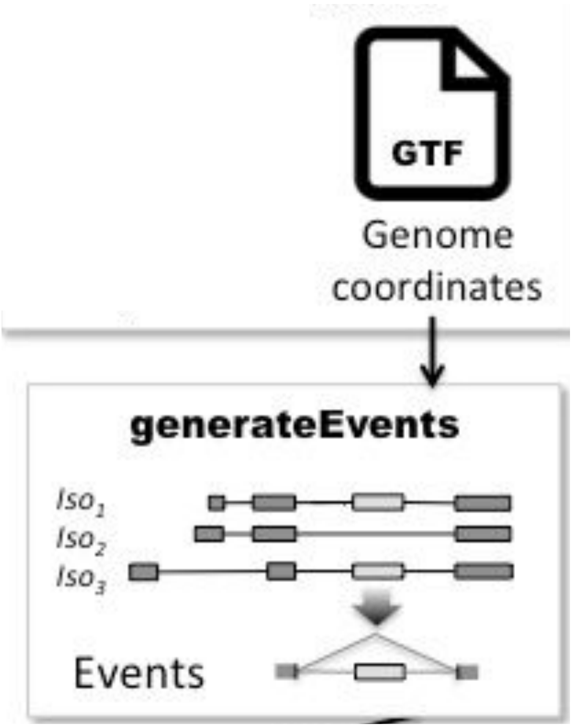
AS landscape in human reference annotations

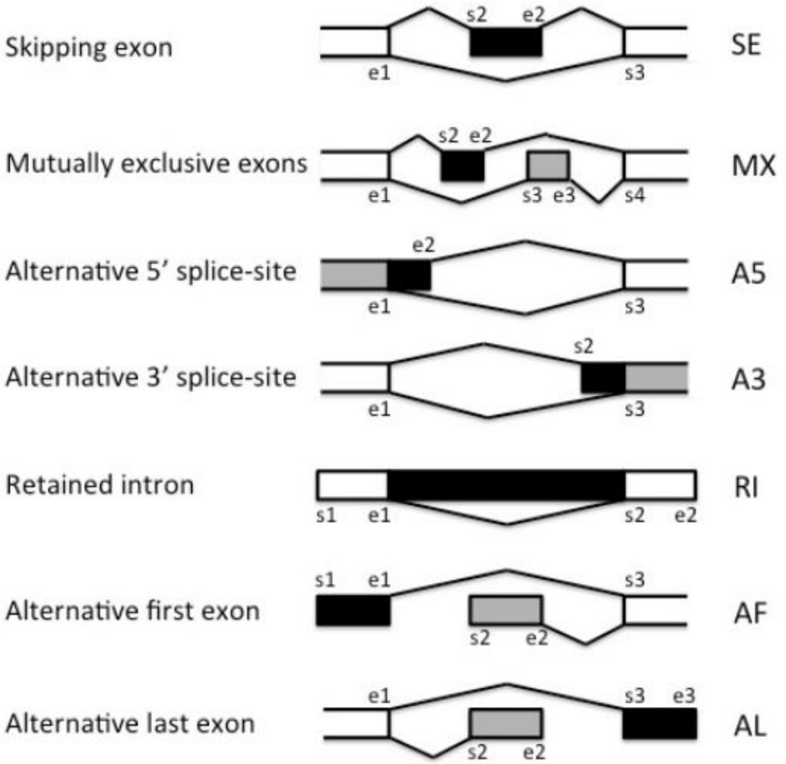


Sammeth, Foissac , Guigó (2008) PLoS Comput Biol 4(8): e1000147

Cecilia Coimbra Klein

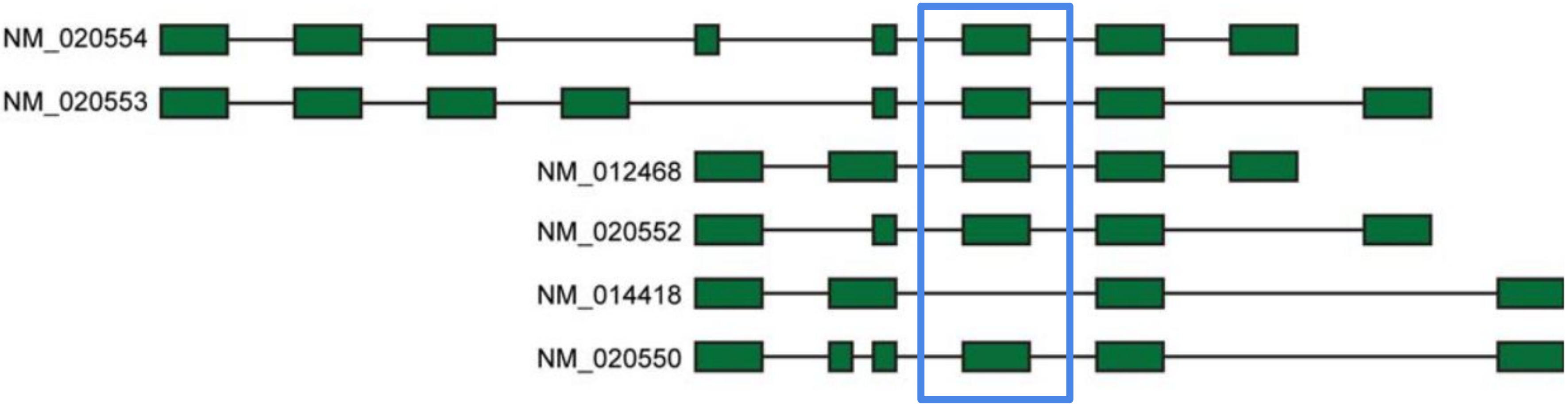
SUPPA: generate events based on gene annotation



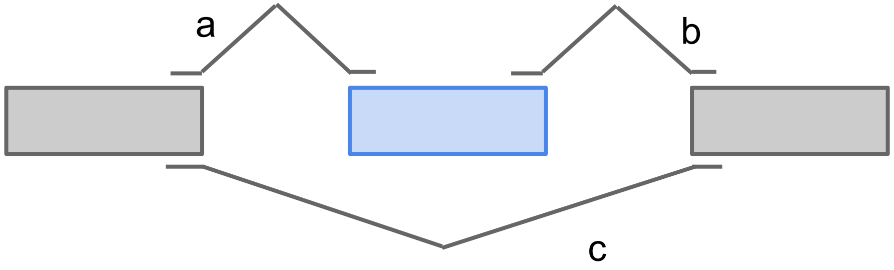


https://bitbucket.org/regulatorygenomicsupf/suppa

Alternative Splicing (AS)



PSI = percent-spliced-in = the number of transcripts in which the given exon is included as a fraction of the number of transcripts in which it is included or excluded

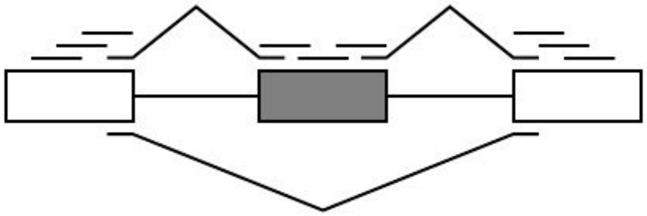


$$PSI = \frac{a+b}{a+b+2c}$$

More than one way to define PSI

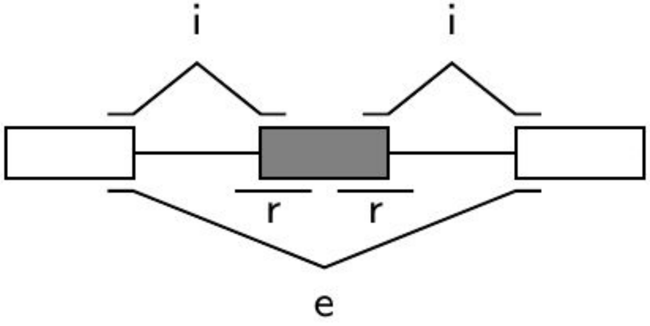
PSI = Percent-Spliced-In

Transcript-centric



$$\Psi = \frac{t_i}{t_i + t_e}$$

Exon-centric



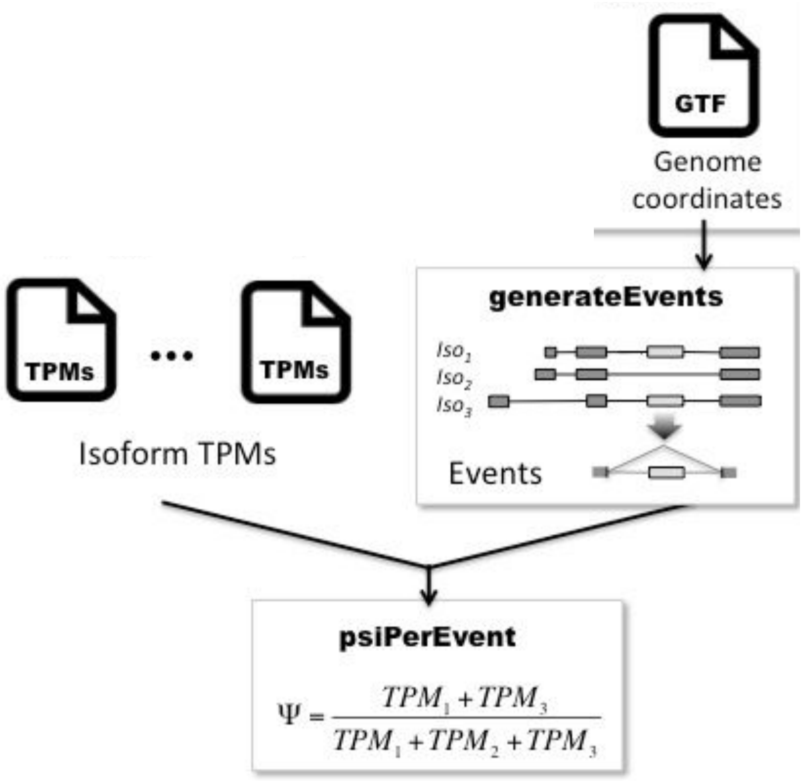
$$\Psi = \frac{i}{i + e}$$

i = inclusion

e = exclusion

r = retention

SUPPA: Quantify event inclusion levels (PSIs)

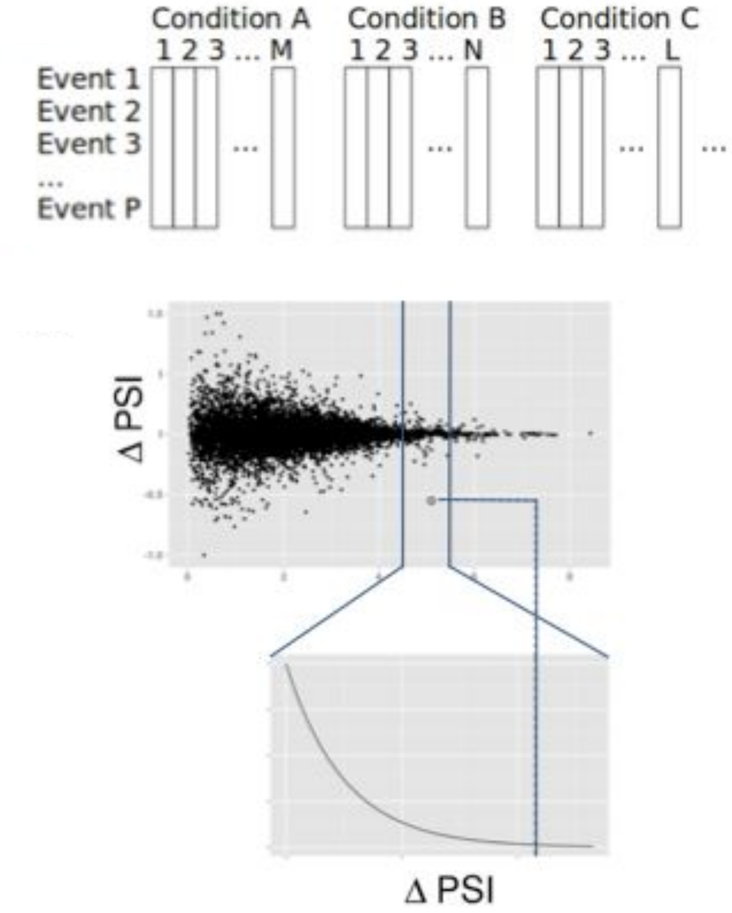


$$\Psi = \frac{TPM_1 + TPM_3}{TPM_1 + TPM_2 + TPM_3}$$

https://bitbucket.org/regulatorygenomicsupf/suppa

Cecilia Coimbra Klein

SUPPA: compare conditions



- SUPPA calculates the magnitude of splicing change (ΔPSI) and their significance across multiple biological conditions, using two or more replicates per condition.

- Statistical significance is calculated by comparing the observed ΔPSI between conditions with the distribution of the ΔPSI between replicates as a function of the gene expression (measured as the expression of the transcripts defining the events).

https://bitbucket.org/regulatorygenomicsupf/suppa

Hands-on

Setup environment 1

RNA-seq data analysis 4

https://public_docs.crg.es/rguigo/Data/cklein/courses/UVIC/handsOn/

Hands-on

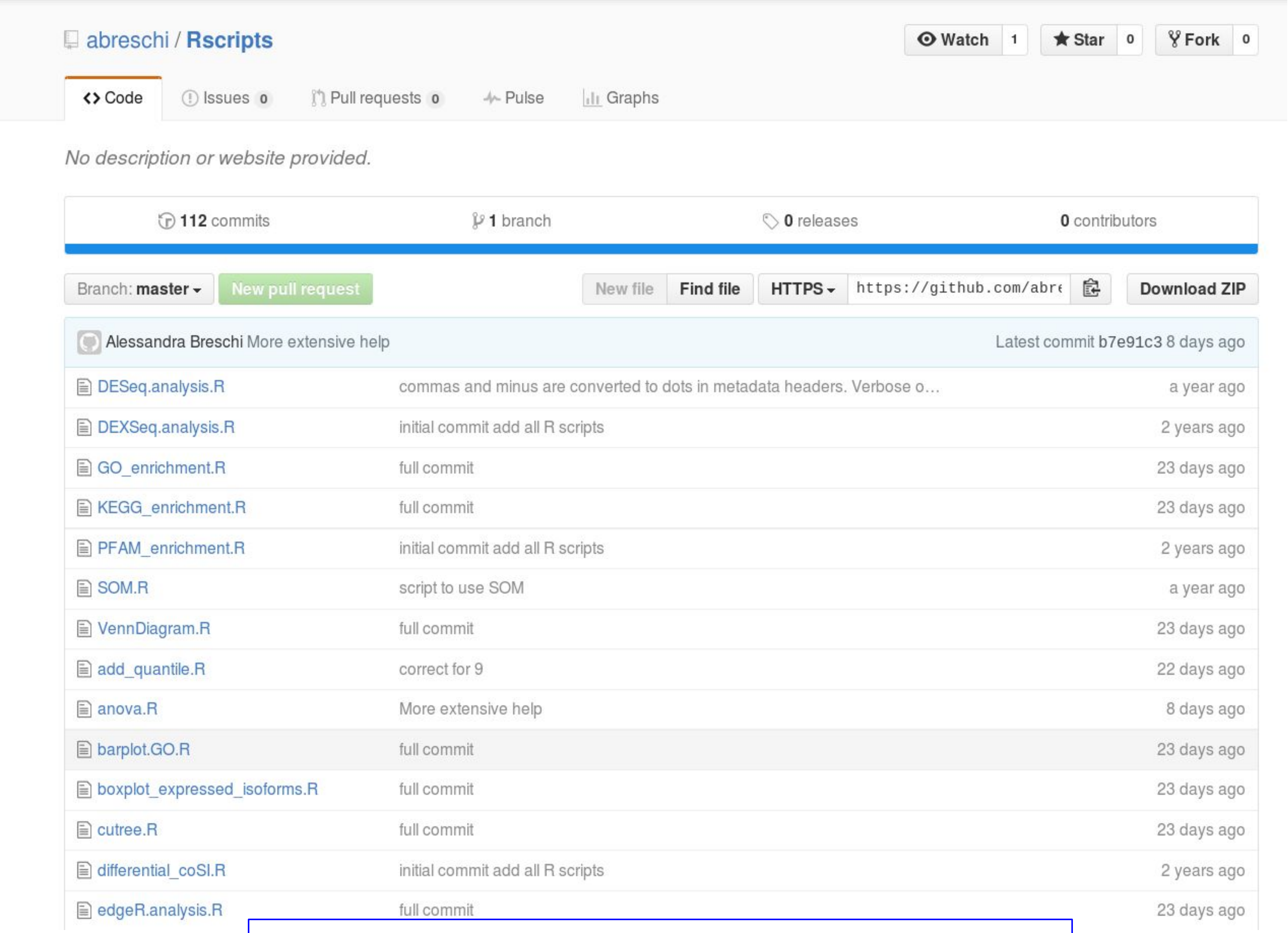
- Forebrain, heart and liver of 12.5 days mouse embryos
    - 2 bio replicates
    - RNA-seq, ChIP-seq and ATAC-seq

- References:
    - mouse genome – mm10 assembly
    - gene annotation – gencode vM4

- Processing:
    - References: a small sample of the genome and annotation (21 chromosomes, 1Mb long)
    - Data: one sample only (100,000 alignment-based pre-filtered reads)

- Analysis:
    - all samples

https://public_docs.crg.es/rguigo/Data/cklein/courses/UVIC/handsOn/



abreschi / **Rscripts**

No description or website provided.

112 commits · **1** branch · **0** releases · **0** contributors

Alessandra Breschi More extensive help — Latest commit b7e91c3 8 days ago

DESeq.analysis.R	commas and minus are converted to dots in metadata headers. Verbose o…	a year ago
DEXSeq.analysis.R	initial commit add all R scripts	2 years ago
GO_enrichment.R	full commit	23 days ago
KEGG_enrichment.R	full commit	23 days ago
PFAM_enrichment.R	initial commit add all R scripts	2 years ago
SOM.R	script to use SOM	a year ago
VennDiagram.R	full commit	23 days ago
add_quantile.R	correct for 9	22 days ago
anova.R	More extensive help	8 days ago
barplot.GO.R	full commit	23 days ago
boxplot_expressed_isoforms.R	full commit	23 days ago
cutree.R	full commit	23 days ago
differential_coSI.R	initial commit add all R scripts	2 years ago
edgeR.analysis.R	full commit	23 days ago

https://github.com/abreschi/Rscripts

--help
will provide input/output parameters

```
Rscript rpkm_fraction.R --help

Usage: rpkm_fraction.R [options] file

Options:
    -i INPUT_MATRIX, --input_matrix=INPUT_MATRIX
        the matrix you want to analyze [default=stdin]
    -m METADATA, --metadata=METADATA
        tsv file with metadata on matrix experiment
    -o OUTPUT, --output=OUTPUT
        additional tags for otuput
    -c COLOR_BY, --color_by=COLOR_BY
        choose the color you want to color by. Leave empty for no color
    -y LINETYPE_BY, --linetype_by=LINETYPE_BY
        choose the factor you want the linetype by. Leave empty for no linetype
    -f FILE_SEL, --file_sel=FILE_SEL
        list of elements of which computing the proportion at each point
    --out_file=OUT_FILE
        store the coordinates in a file [default=NULL]
    -P PALETTE, --palette=PALETTE
        file with the colors
    -t TAGS, --tags=TAGS
        choose the factor by which grouping the lines [default=labExpId]
    -h, --help
        Show this help message and exit
```