Motif enrichment analysis based on post-hoc inference of large-scale multiple testing

Benjamin Sadacca

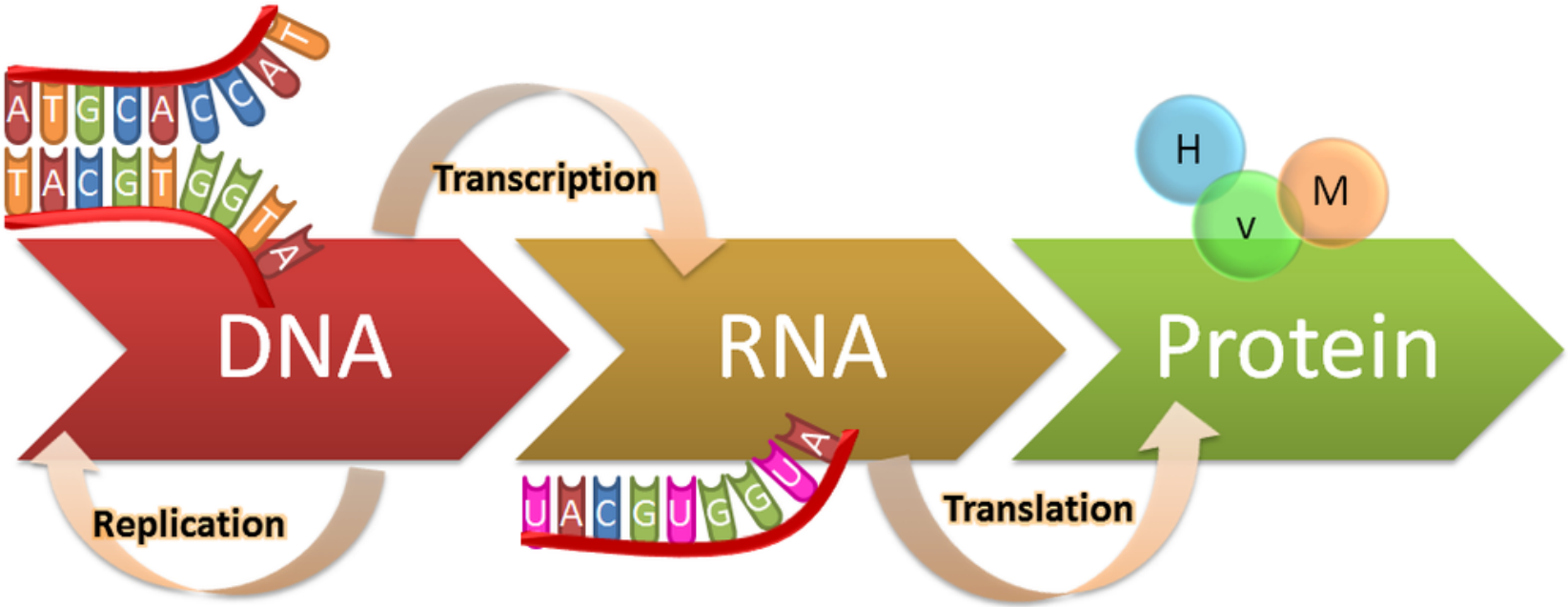
Joint work with Pierre Neuvial

Institut Curie
Translational Research Dpt & Centre d'immunotherapie - U932
Team of J. Waterfall & S. Amigorena

Institut de Mathématiques de Toulouse

May, 23rd, 2018

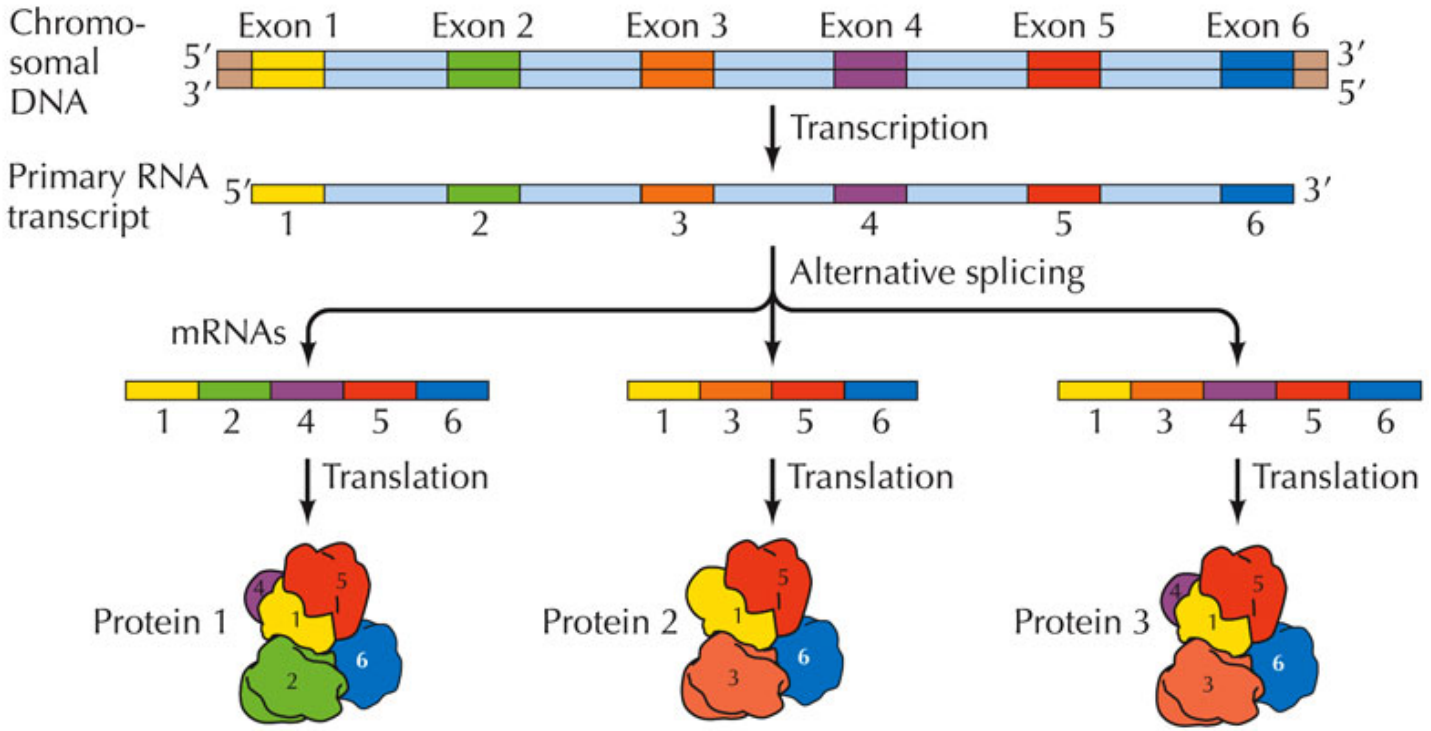
Protein synthesis - The central dogma of molecular biology



How can only 25,000 genes encodes for more than 100,000 proteins ?

Alternative splicing

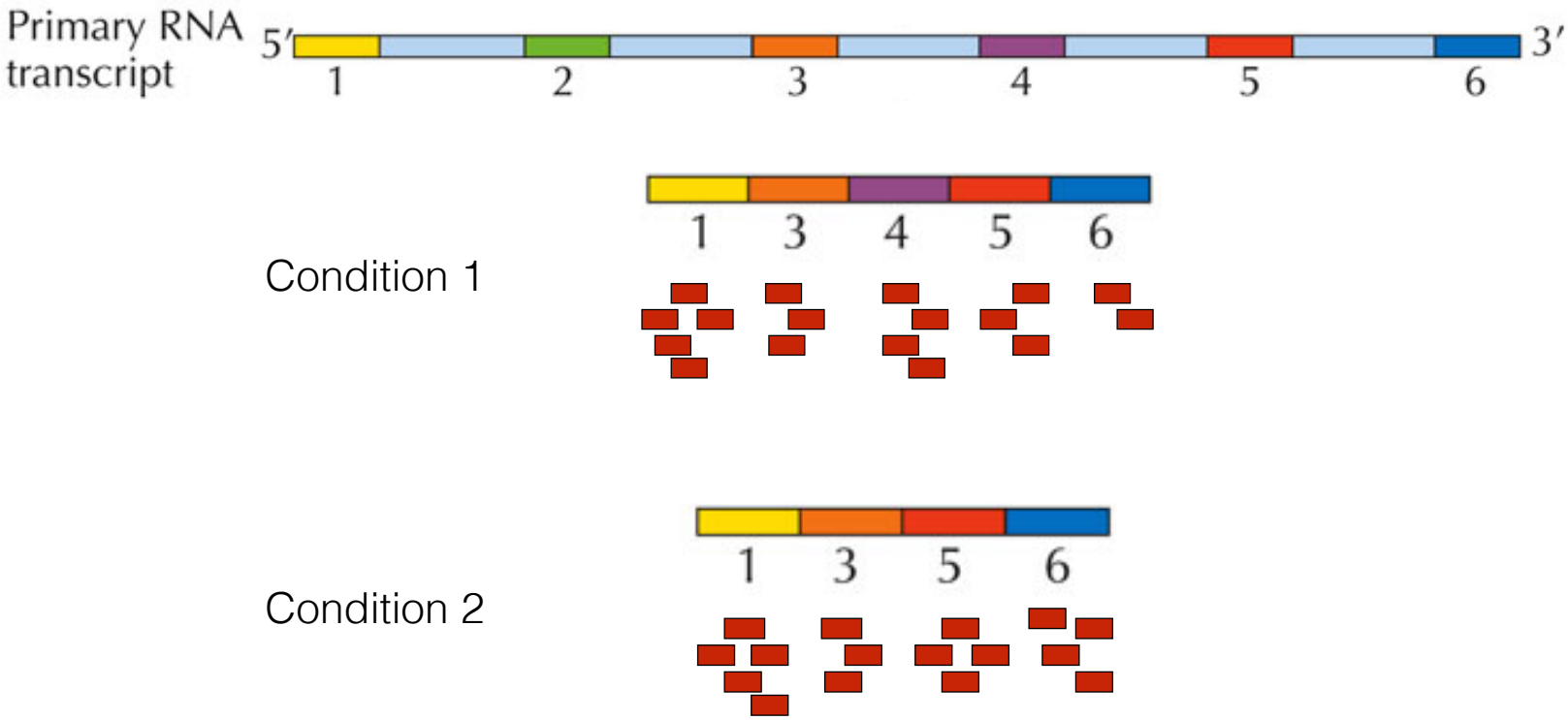
- Each gene may encode several proteins by a process call alternative splicing
- One gene makes different mRNA products (isoformes) and hence different proteins



THE CELL, 4th Edition, Figure 5.5

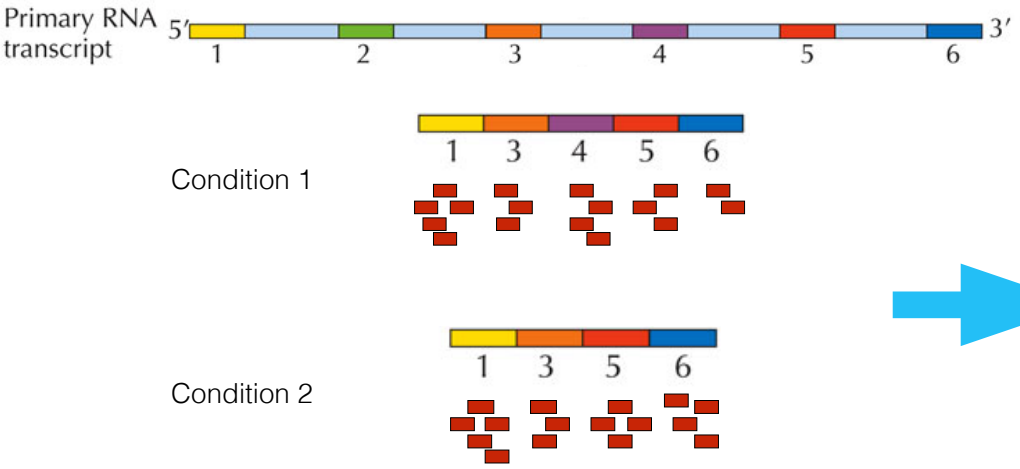
Differential alternative splicing analysis

- Find differences in exon splicing patterns among different biological conditions
- Detect the differences by analyzing distribution of reads (expression level)



Differential alternative splicing analysis

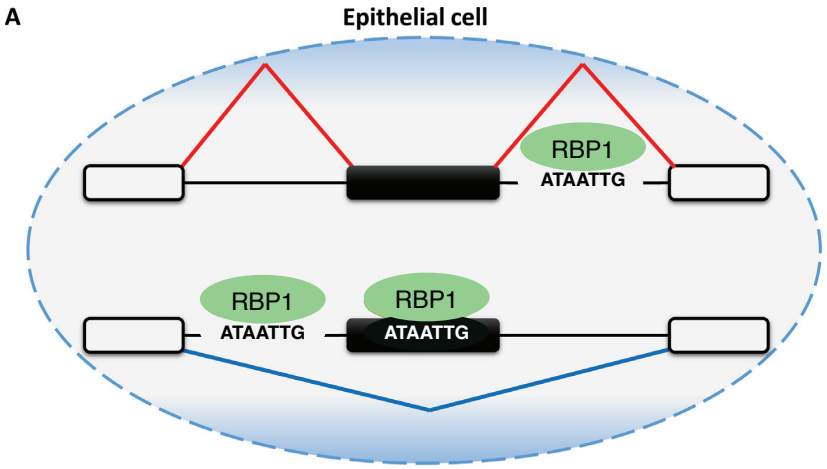
Result : list of differentially spliced exons with their inclusion level :
1 is included
-1 is excluded



Gene	Exon	Exon inclusion levels	p-value
Gene1	4	1	5.5e-10
Gene2	3	-1	6.7e-7
Gene3	8	-1	7.1e-7
Gene4	2	1	2.7e-6
…	…	…	…

Alternative splicing is regulated by RNA Binding Proteins

- RBP are splicing factors
- Proteins that regulate post-transcription of mRNAs
- Bind to specific DNA sequences (motifs)
- Binding position has a key role in splicing regulation



Park et al, NAR 2016

RBPs can produce specific isoforms associated with cancer development or chemo-resistance

→ Need to reliably identify the RBPs involved in the regulation of differentially spliced exons

RBPs action at a specific position may change between physiological and cancerous states

→ Need to reliably identify their binding position

How to identify the RBPs that regulates a set of exons ?

- RBPs bind specific sequences (motifs) on the genome
- Look at whether their motifs are overrepresented in the intronic sequences of the exons regulated

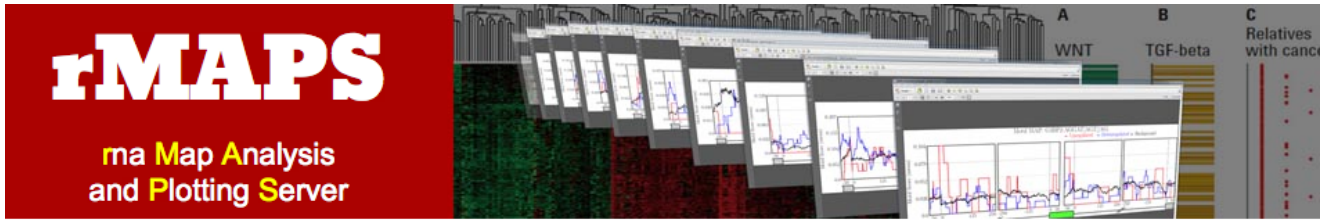
Exons differentially spliced between 2 conditions

TAATCACAATAATTGTGGGGAAGAAGCTAGGAAGTTTTATCCGCGACGAC EXON 1
ACTACTACTAAATCGACCTATAATTGAGAACAAGATAATTGGTGTATATTG EXON 2
GGCACCTCGAGCGTGTTTGATATCCAGGTGGCGCCCCGAACGCTTGTCTT EXON 3
TCCCGATTCCTATAATTGAATAATTGAAGACGAATCGTGGCATAATTGTG EXON 4
⋮
CATAATTGAACGTCCGAAACTGGGCATACGATCTAAAATAAGCGAAAAGC EXON M

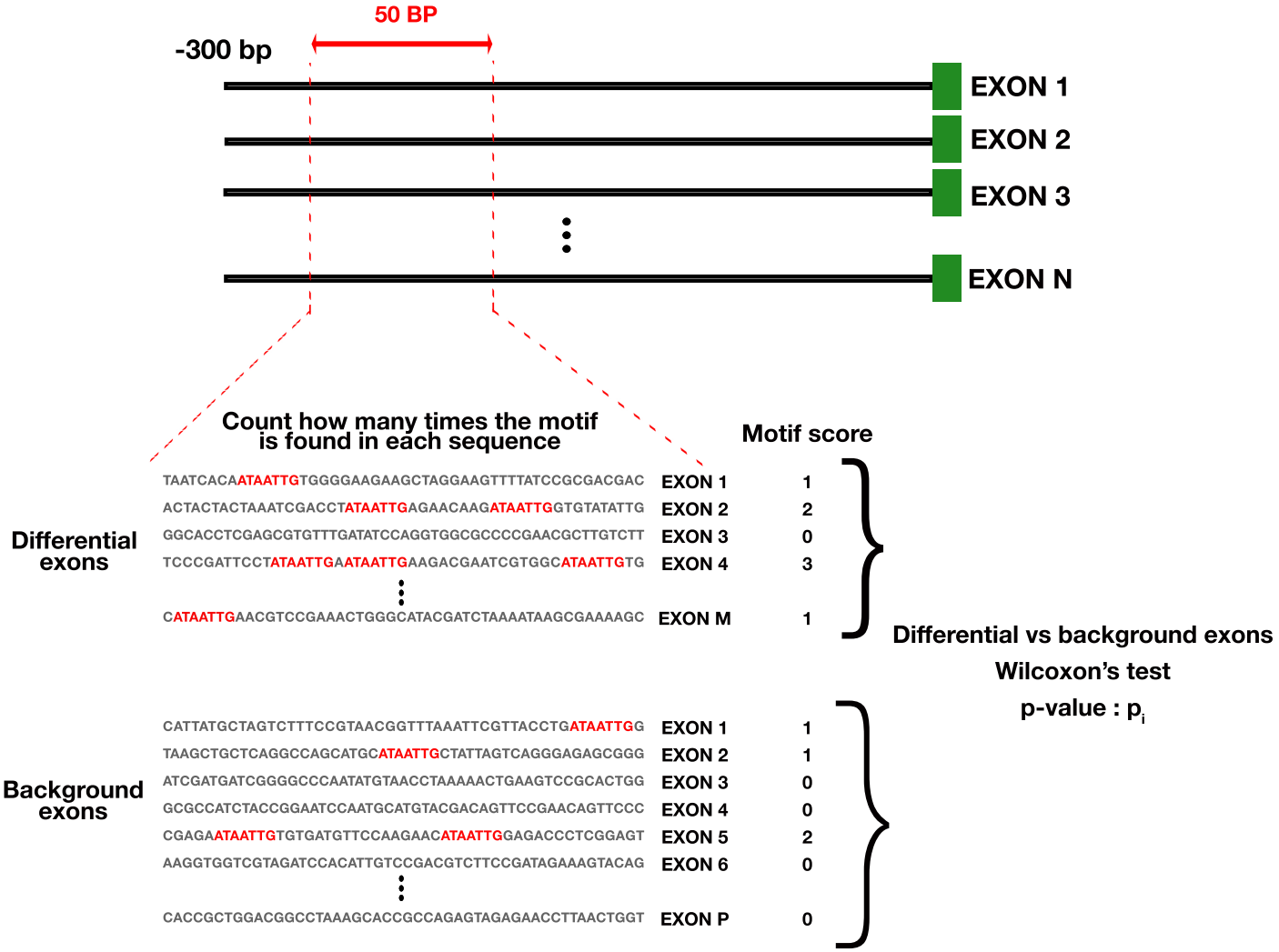
Exons not differentially spliced (controls)

CATTATGCTAGTCTTTCCGTAACGGTTTAAATTCGTTACCTGATAATTGG EXON 1
TAAGCTGCTCAGGCCAGCATGCATAATTGCTATTAGTCAGGGAGAGCGGG EXON 2
ATCGATGATCGGGGCCCAATATGTAACCTAAAAACTGAAGTCCGCACTGG EXON 3
GCGCCATCTACCGGAATCCAATGCATGTACGACAGTTCCGAACAGTTCCC EXON 4
CGAGAATAATTGTGTGATGTTCCAAGAACATAATTGGAGACCCTCGGAGT EXON 5
AAGGTGGTCGTAGATCCACATTGTCCGACGTCTTCCGATAGAAAGTACAG EXON 6
⋮
CACCGCTGGACGGCCTAAAGCACCGCCAGAGTAGAGAACCTTAACTGGT EXON P

Is the motif of a given RBP more often identified in the sequences of the regulated exons than in the sequences of the control exons ?

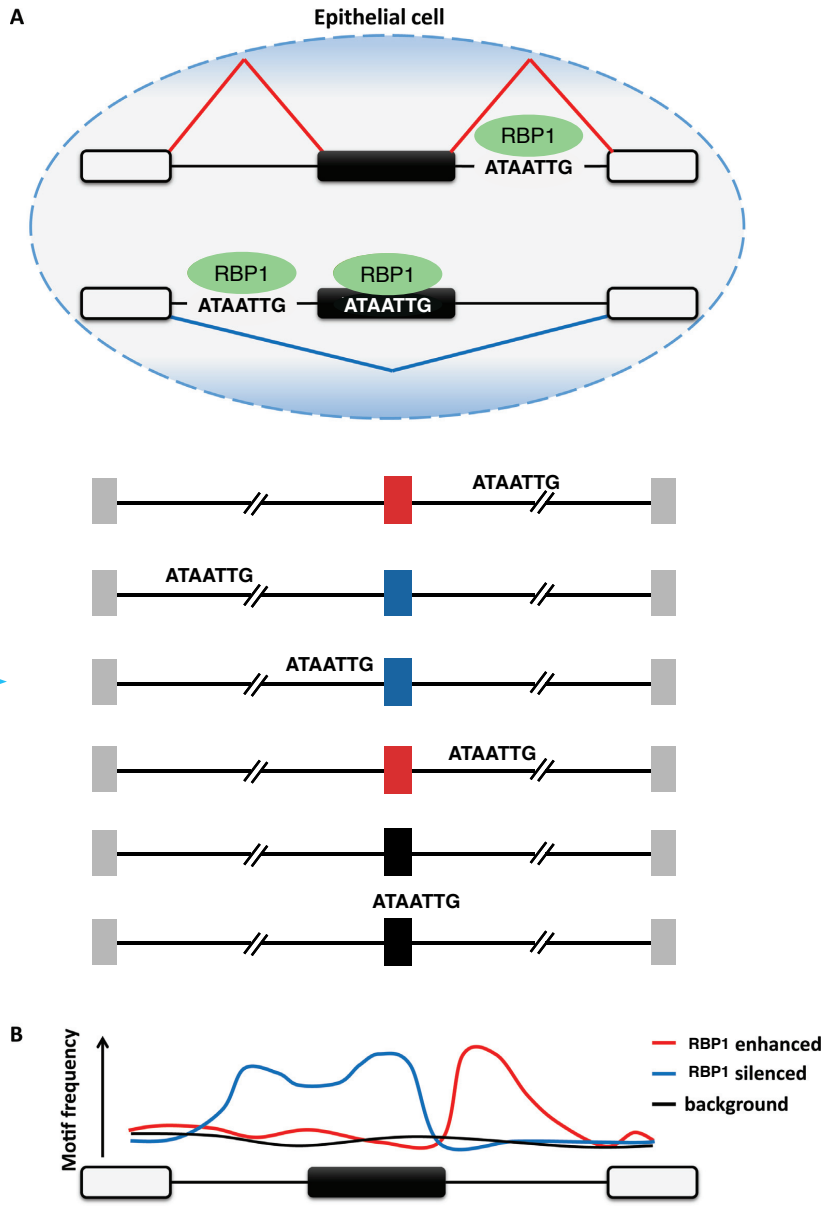


Goal : Identify the motifs that are overrepresented in a given set of sequences and ensure that this result is not observed by chance.

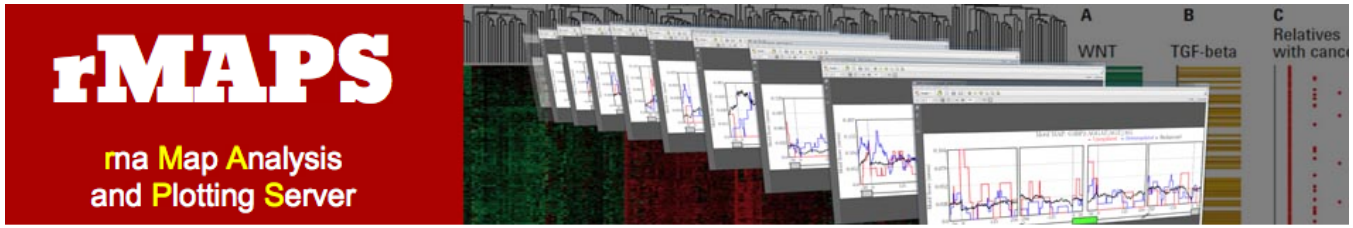


RNA Binding Map

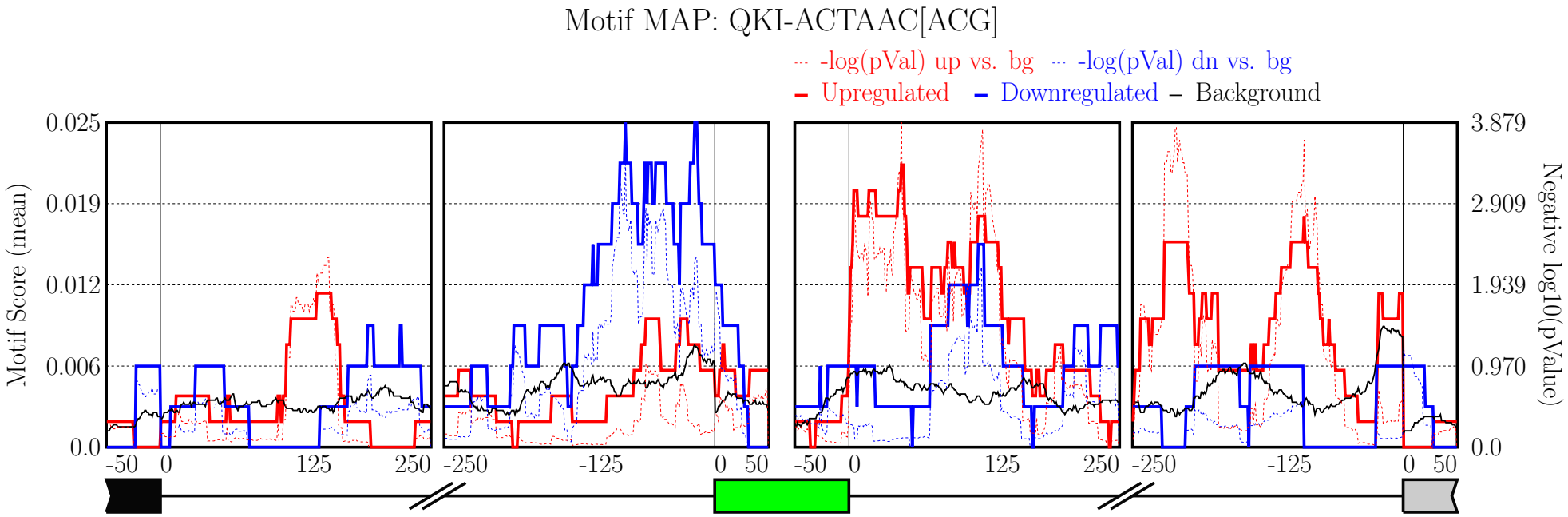
Gene	Exon	Exon inclusion levels	p-value
Gene1	4	1	5.5e-10
Gene2	3	-1	6.7e-7
Gene3	8	-1	7.1e-7
Gene4	2	1	2.7e-6
Gene4	5	1	0.2
Gene5	2	-1	0.07
…	…	…	…



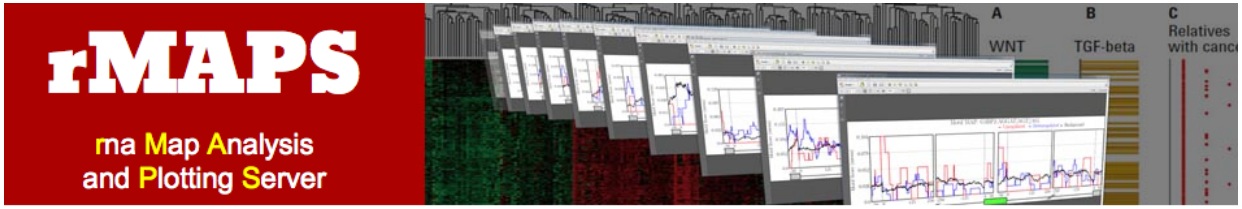
RBP motif enrichment analysis with rMAPS



- Web application
- rMATS suites (takes rMATS results as input)
- Generates RNA-maps for the analysis of RBPs binding sites
- Perform analysis of binding sites around differential alternative splicing events for over 100 of known RBPs



Limitations of rMAPS

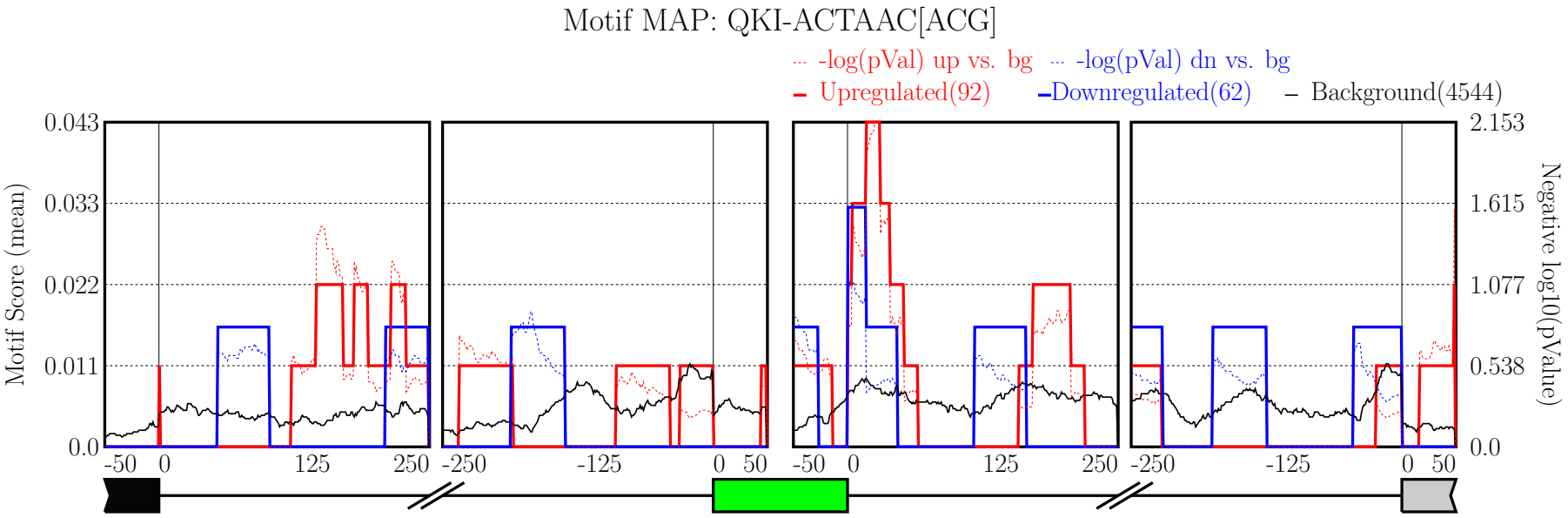


Practical limits

- Source code not available
- Not reproducible
- (Many server problems)

Methodological limits

- **Only the smallest p-value** among the upstream/downstream sequences is reported
- **No precise identification** of binding site
- The sliding window procedure produces 250 p-values that have to be corrected for **multiple testing**



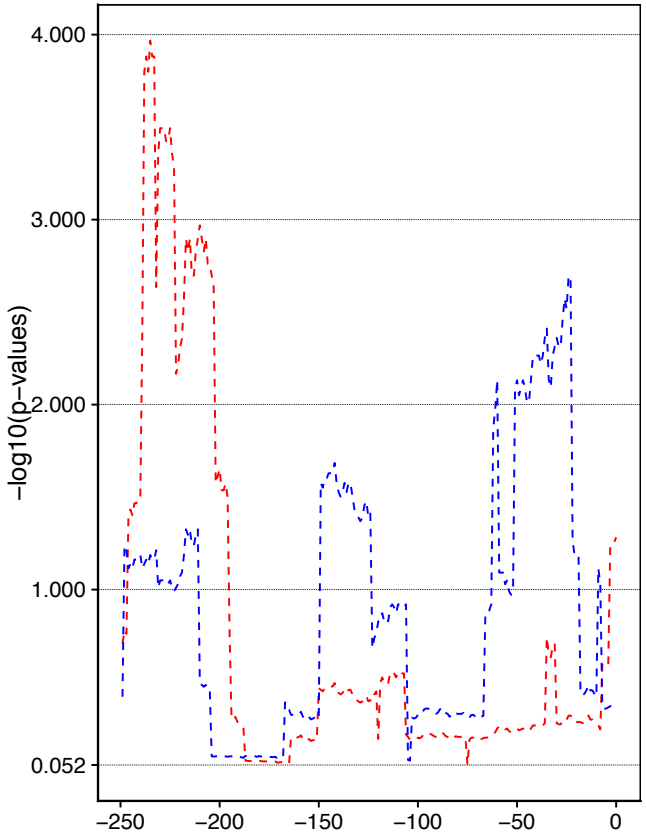
Extending rMAPS

We propose to extend rMAPS by :

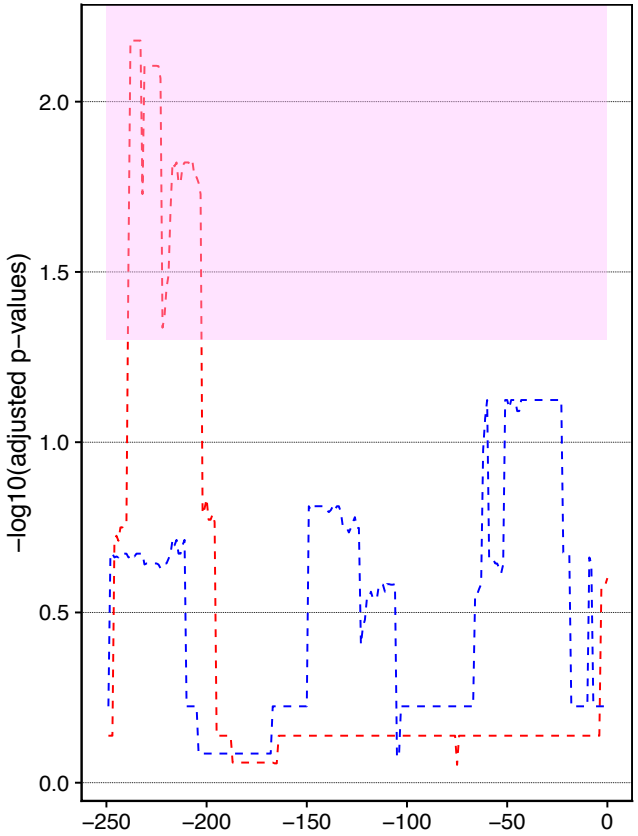
- Adding value on the biological side : finer localization of relevant binding regions
- Adding value on the statistical side : control of false positive
- Providing an open source solution

How to correct for the multiple testing ?

RNA Binding Map

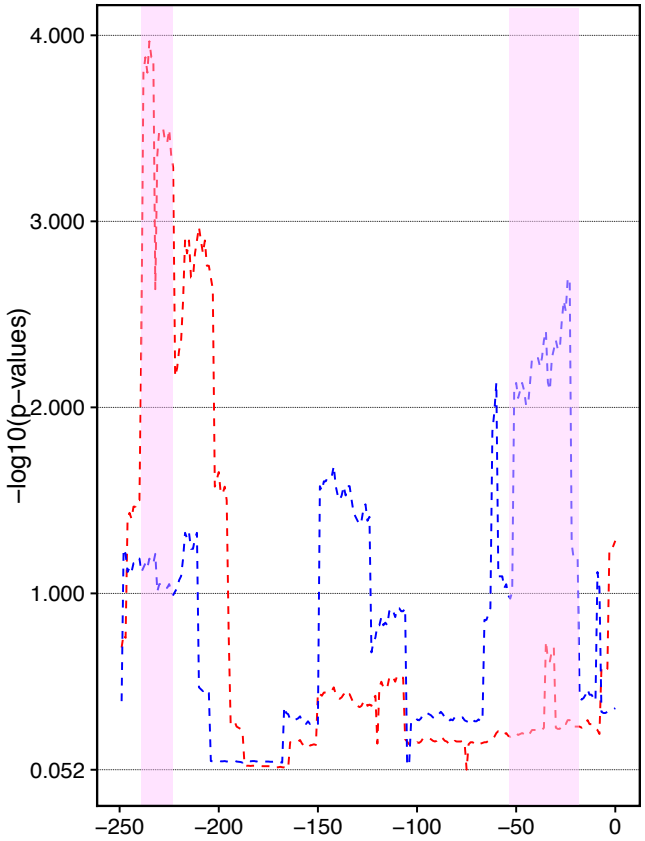


Classical Multiple Testing



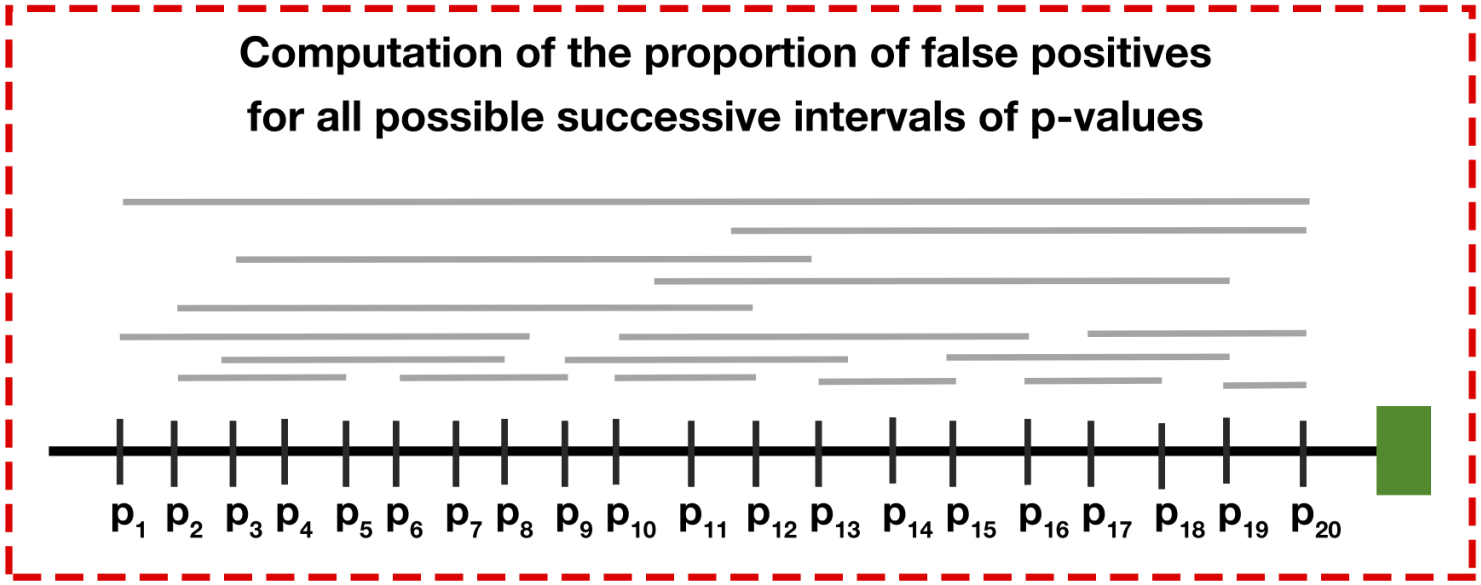
FDR ≤ 25%

Post Hoc Inference



Estimate the proportion of FP with probability ≥ 75% in these two regions

Find the largest significant intervals of successive p-values



Goal : with high probability, the proportion of false positives among any of the selected intervals does not exceed a user-defined threshold.

These selected intervals will be called significant.

Limits of classical multiple testing procedures

- Do not control for the selection effect
- Do not control for the evaluation of multiple nested sets of p-values

$\Rightarrow$ **Post hoc inference**, as introduced by Goeman and Solari, gives a statistical guarantee for all possible sets of p-values simultaneously by providing simultaneous upper bounds on the number of false positives on the selected sets.

Joint Family-Wise Error Rate control for post hoc inference

The goal of post hoc inference is to build functional bounds V $(\cdot)$ defined on all subset of hypotheses, such that the following uniform guarantee holds :

$$\mathbb{P}(\forall S \in \{1, \ldots, m\}, |S \cap H_0| \leq V(S)) \geq 1 - \alpha$$

where S is a subset of hypotheses, m is the number of null hypotheses to be tested and $H_0 \subset 1, \ldots, m$ corresponds to the (unknown) set of true null hypotheses.

Joint Family-Wise Error Rate control for post hoc inference

Blanchard, Neuvial and Roquain have proposed to build $V(.)$ based on the control of a multiple testing criterion called "joint (family-wise) error rate" (JER).

Using the $k - FWER$ and a p-value thresholding approach based on Simes' inequality they show that

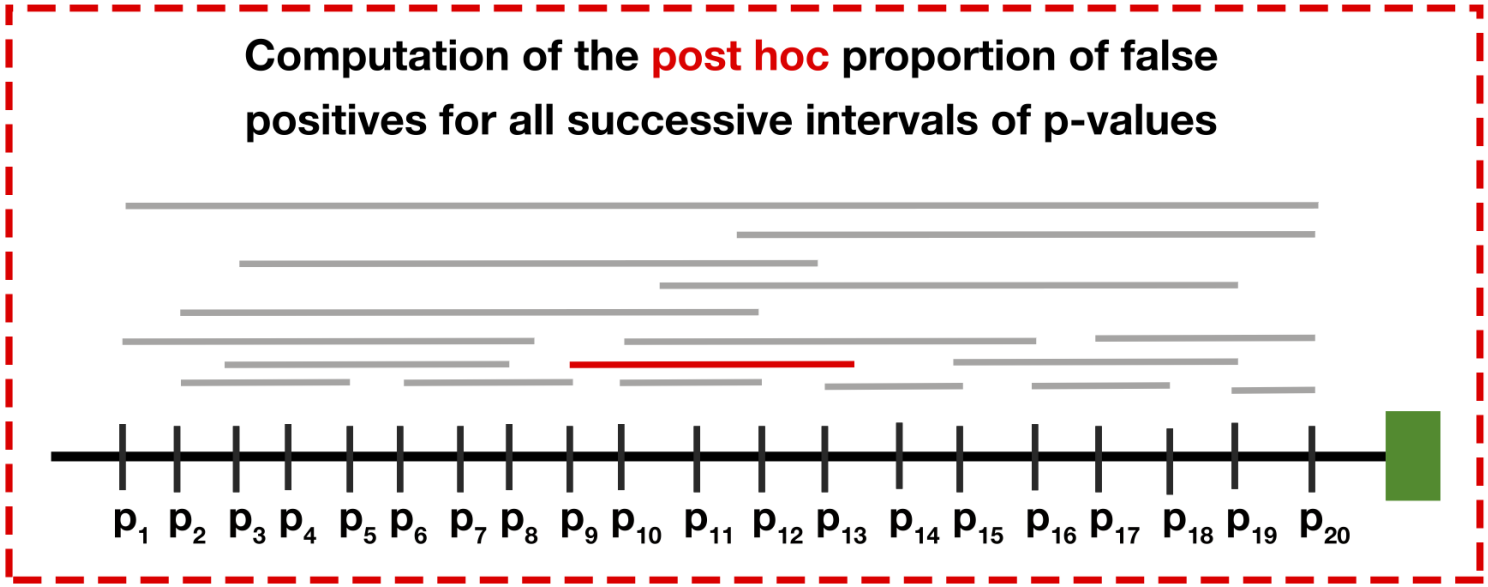
$$V_{\alpha}^{Simes}(S) = \min_{k \in 1 \leq \ldots \leq |S|} \{ \sum_{i \in S} \mathbb{1}_{\{p_i > \alpha k/m\}} + k - 1 \}$$

with $R \subset 1, \ldots, m$, $k \in 1, ..., K$ and $i \in S$.

This bound is the same than the one introduced by Goeman and Solari, but easier to interpret and to implement.

Joint Family-Wise Error Rate control for post hoc inference

We calculate the bound $V_{\alpha}^{Simes}(S)$ for all possible successive intervals of p-values, from which we derive a bound on the proportion of false positives ($VP_{\alpha}^{Simes}(S)$).



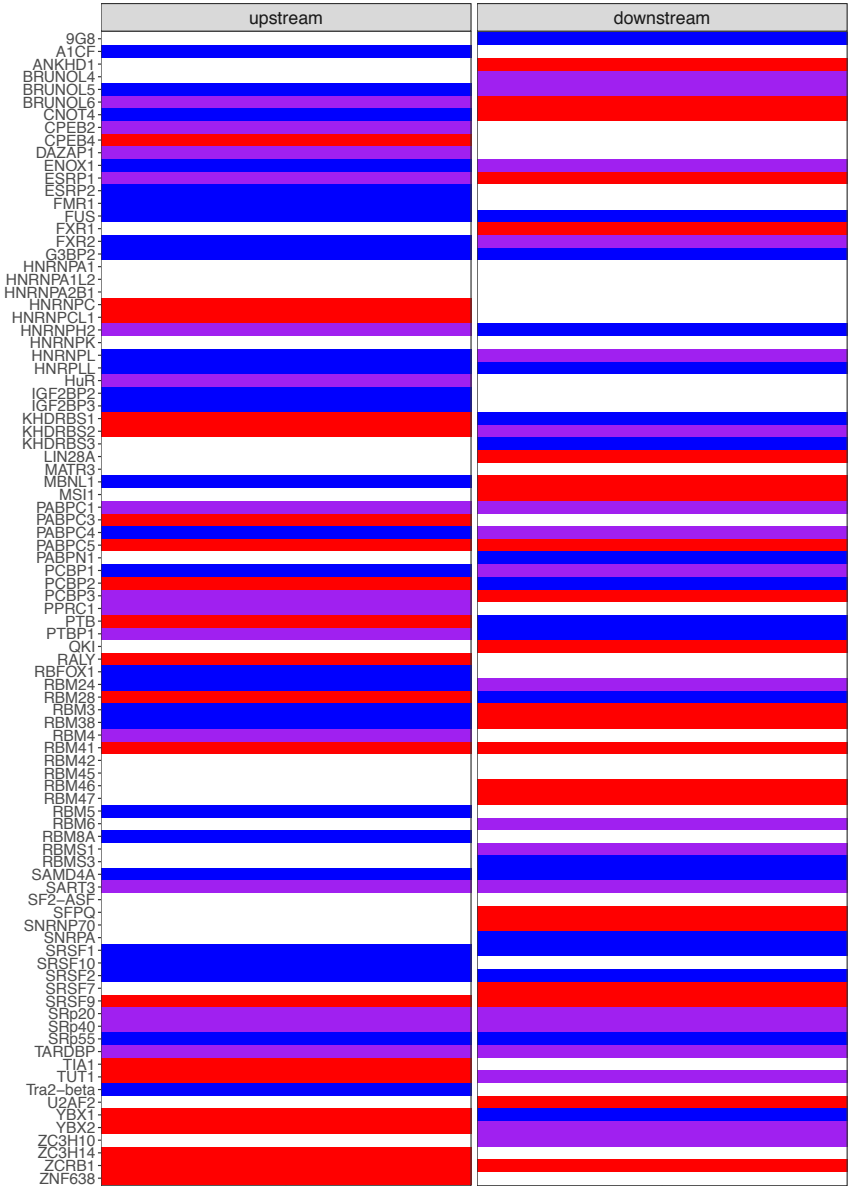
According to a threshold δ and if it exists, the largest interval as $VP_{\alpha}^{Simes}(S) \leq \delta$, informs us about the over-representation of the RBP's motif in the sequence and its binding position.

Joint Family-Wise Error Rate control for post hoc inference

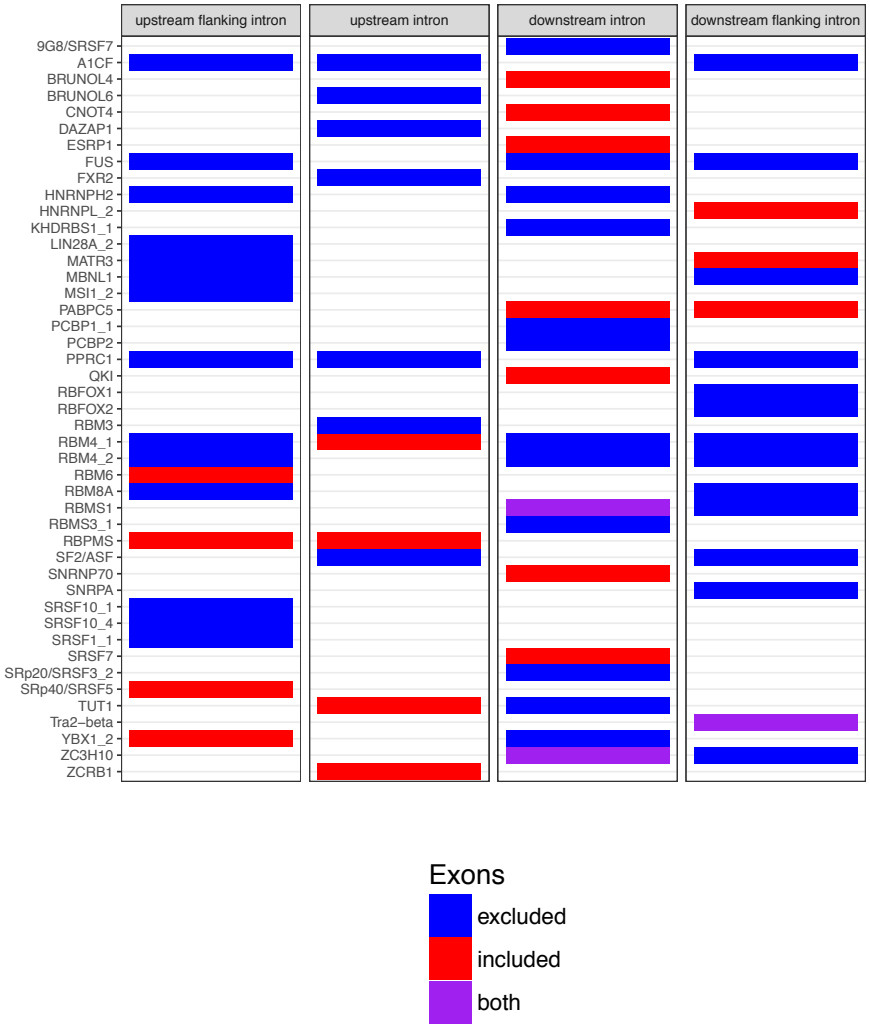
/Users/bsadacca/Thesis/Talk/Talk_RT2/Seminaire_Evry/fig/eval_interval.gif

Post hoc correction leads to fewer false positives

Classical rMAPS
raw p-values < 5%



rMAPS + post hoc
Bound controlled at 5%
Intervalles with maximum 25% of FP



Exons
excluded
included
both

More flexibility in user investigations

/Users/bsadacca/Thesis/Talk/Talk_RT2/Seminaire_Evry/fig/rMAPS_shiny_video_cut.mp4

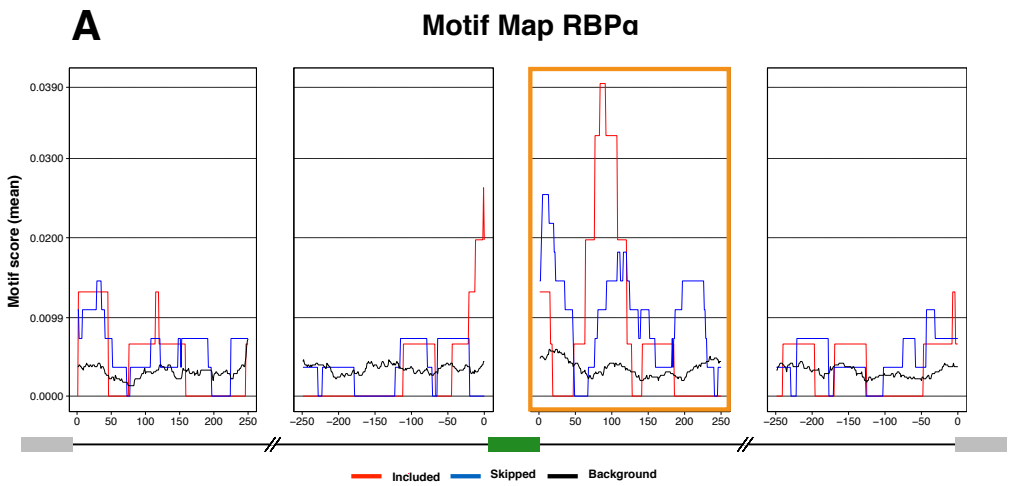
Application to Ewing's sarcoma cell lines

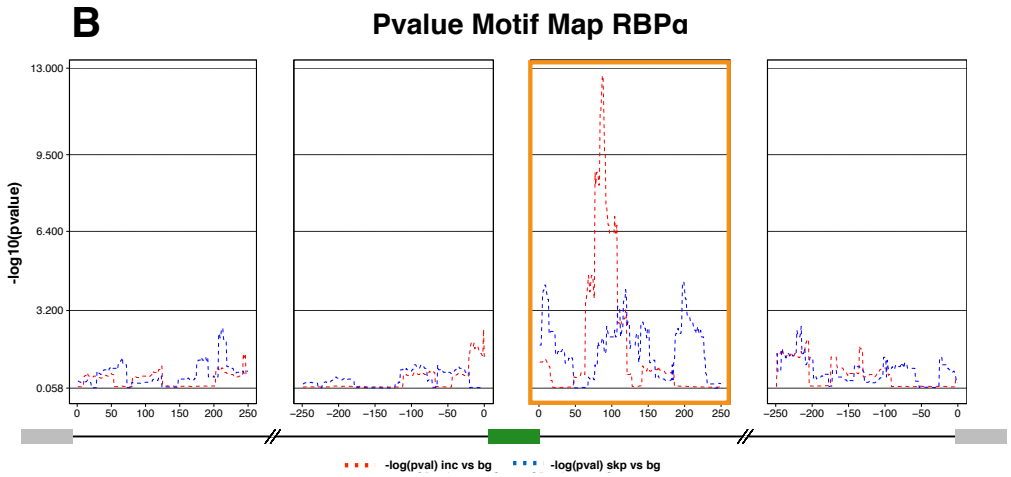
Collaboration with O. Saulnier from U830 (IC)

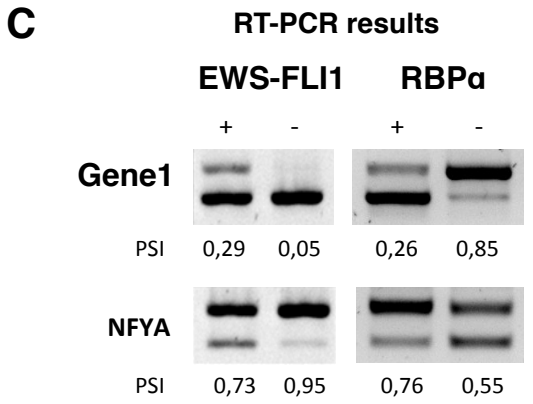
- Ewing's sarcoma is an aggressive cancer of bone and soft tissues affecting children and young adults.
- It is driven, in 85% of cases, by a a modification of the transcription factor EWS-FLI1 that also impacts splicing.

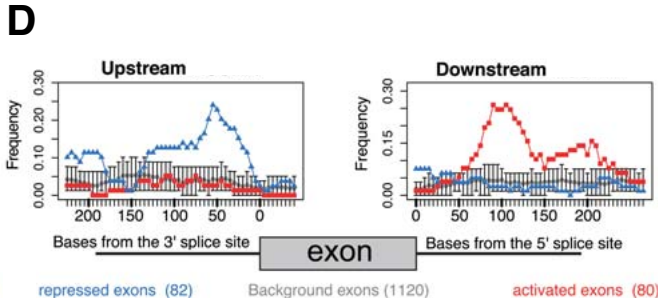
- Transcriptome-wide splicing events analysis on multiple Ewing's sarcoma cell lines following EWS-FLI1 knock-down.
- We used rMAPS-PH to identified RBPs that regulate alternative splicing in these cells.

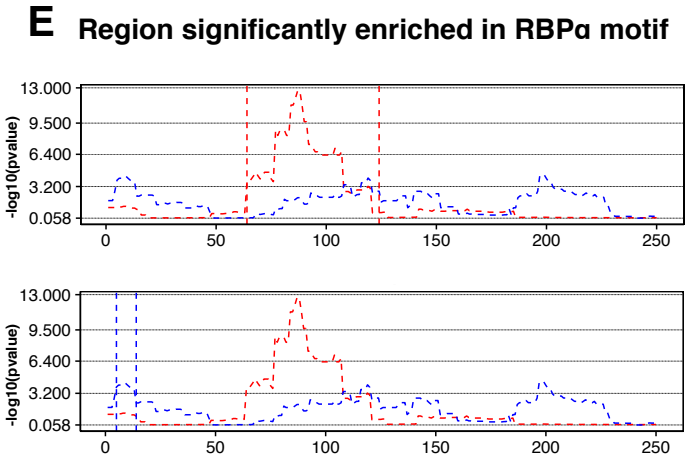
Application to Ewing's sarcoma cell lines











Conclusion

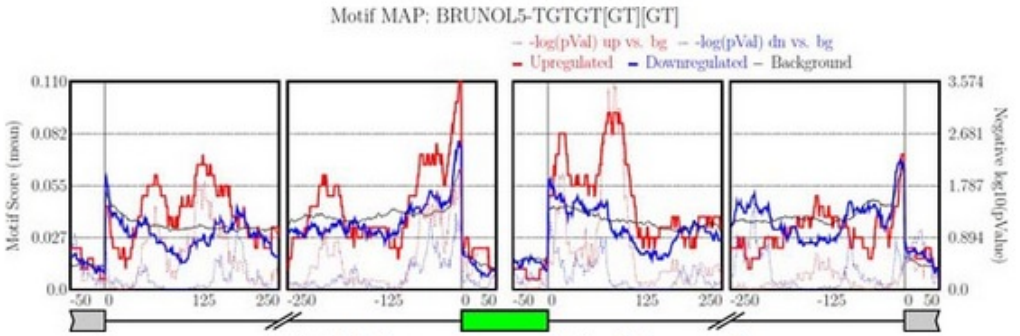
- Extended the statistical framework of rMAPS with post-hoc inference to correct for multiple testing
- Control of false positive in a complex multiple testing framework
- Precise identification of RBP binding sites
- Identify reliable RBPs that have been validated biologically
- Provide an open source solution as a R package

Perspectives

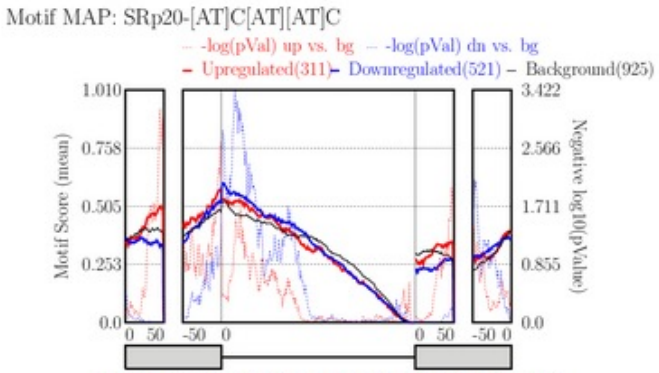
- Current approach uses the Simes' local test to compute the post-hoc bound of false positive $\Rightarrow$ PRDS assumption
- Use post hoc procedures that are adaptive to unknown dependence - JER with permutations

- Improve computing time for the identification of the binding site
- Extend the method to fit other common methods of alternative splicing analysis such as DEXSeq or Voom
- Extend the method to any kind of alternative splicing event

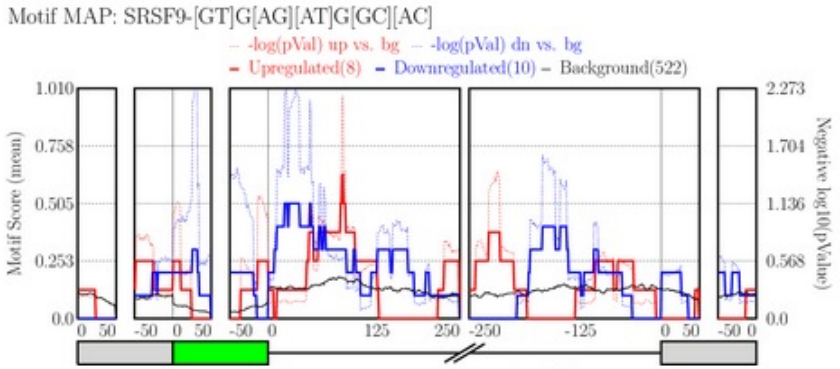
Perspectives



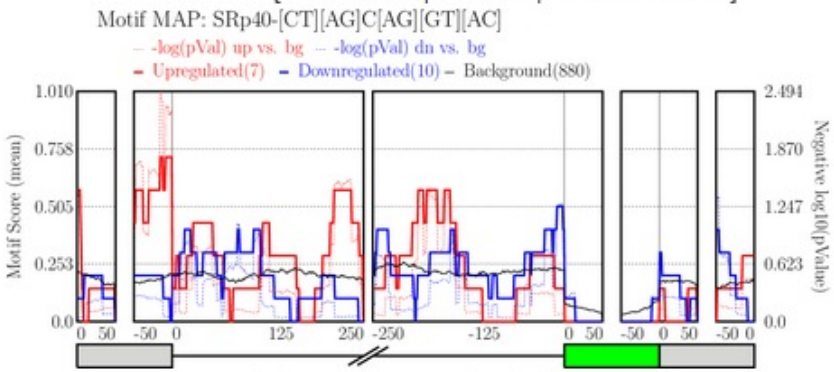
[rMAPS Output example for **SE** event]



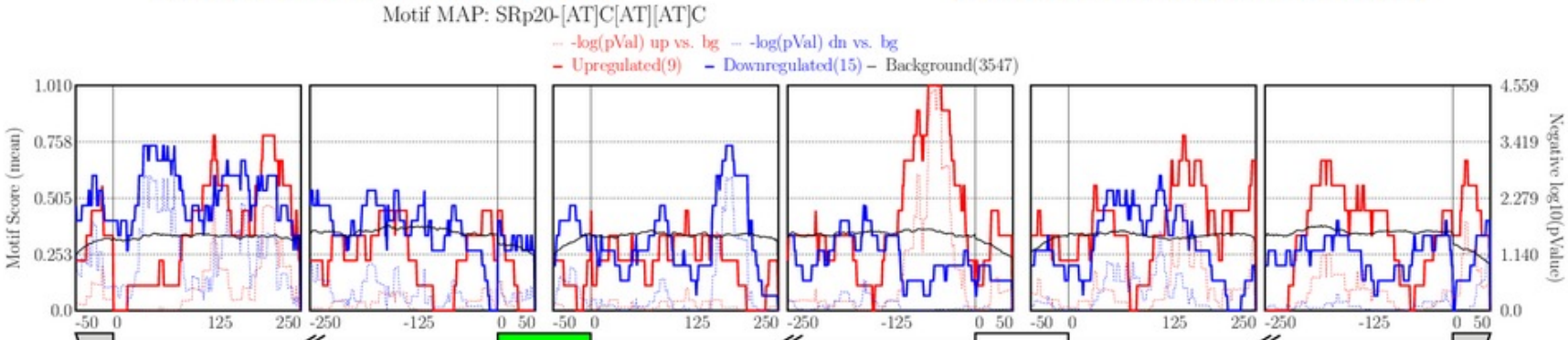
[rMAPS Output example for **RI** event]



[rMAPS Output example for **A5SS** event]



[rMAPS Output example for **A3SS** event]



[rMAPS Output example for **MXE** event]