mTIM: margin-based transcript prediction from RNA-seq

Nico Görnitz,[1,2] Georg Zeller,[2,3] Jonas Behr,[2] Andre Kahles,[2]
Pramod Mudrakarta,[2] Sören Sonnenburg,[1] and Gunnar Rätsch [2]

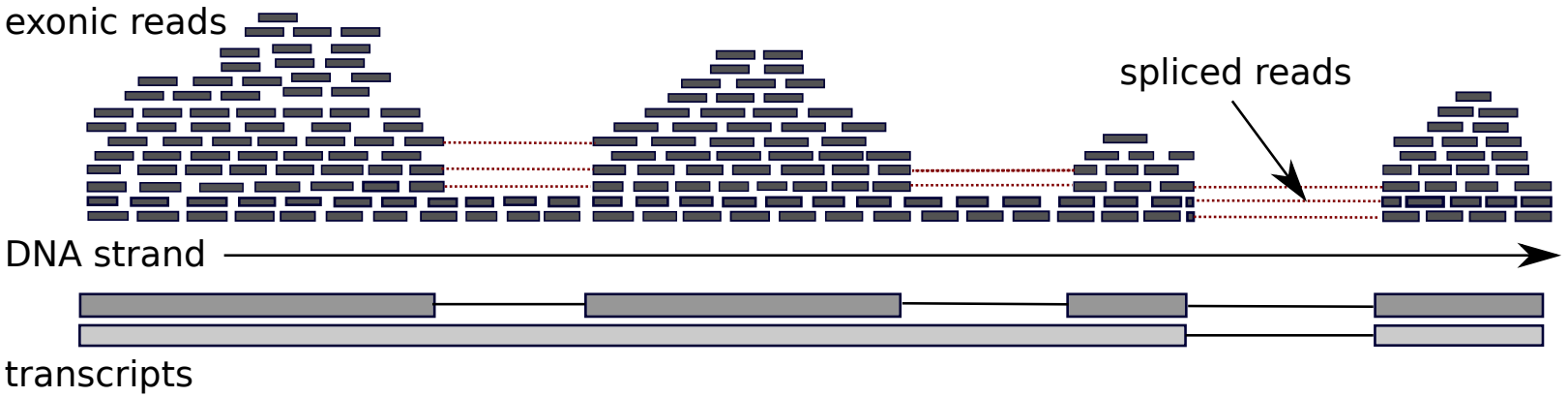






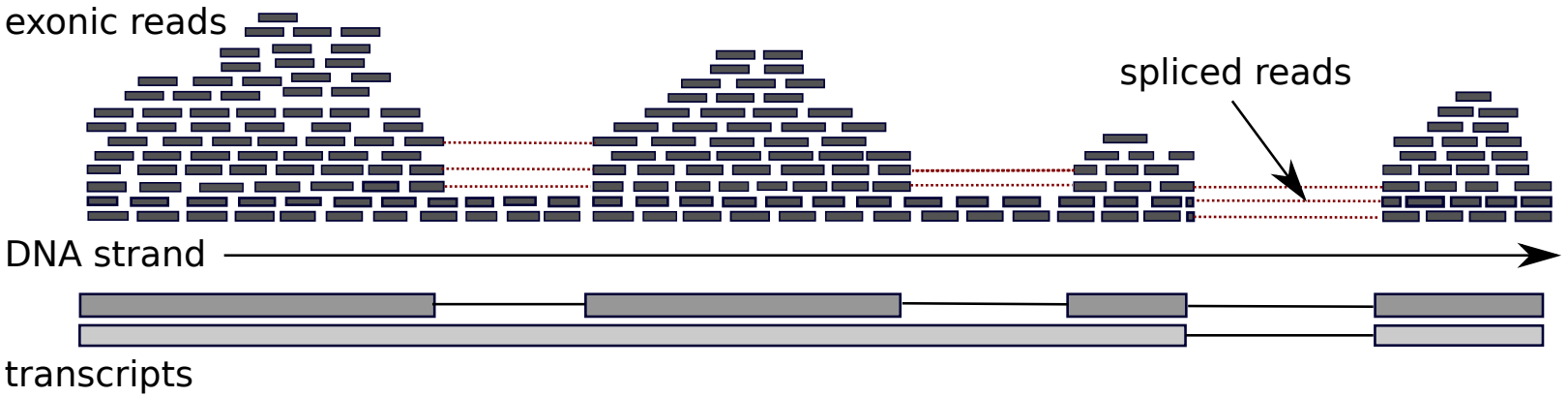
[1] Department of Software Engineering and Theoretical Computer Science, Technical University Berlin, Berlin, Germany
[2] Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany
[3] Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany

RNA-seq based transcript prediction



- reconstruct transcript using RNA-seq data

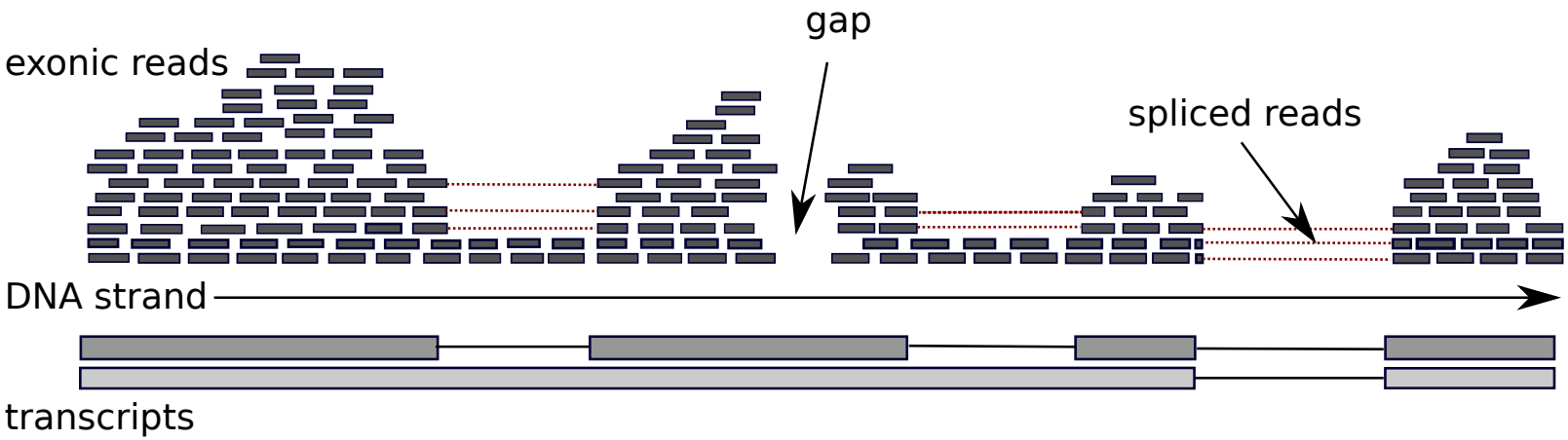
RNA-seq based transcript prediction



- reconstruct transcript using RNA-seq data
- existing methods: Cufflinks[1], Scripture[2], ..

[1] Trapnell et al. 2010 [2] Guttman et al., 2010

RNA-seq based transcript prediction



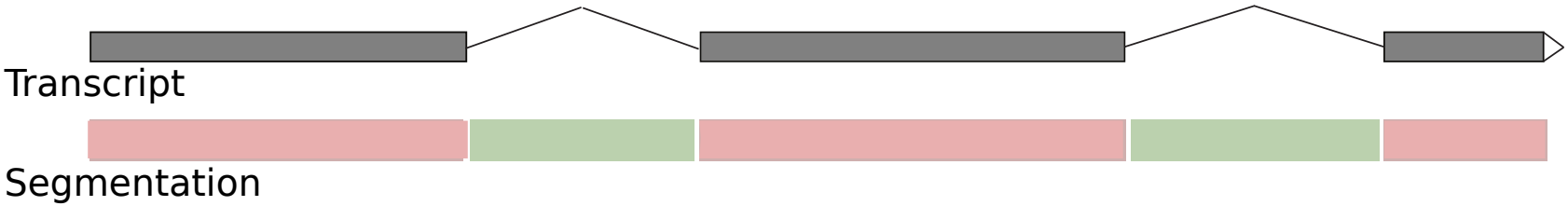
- reconstruct transcript using RNA-seq data
- existing methods: Cufflinks[1], Scripture[2], ..

Problems

- read coverage is too small or gaps exist
- depends massively on alignment quality

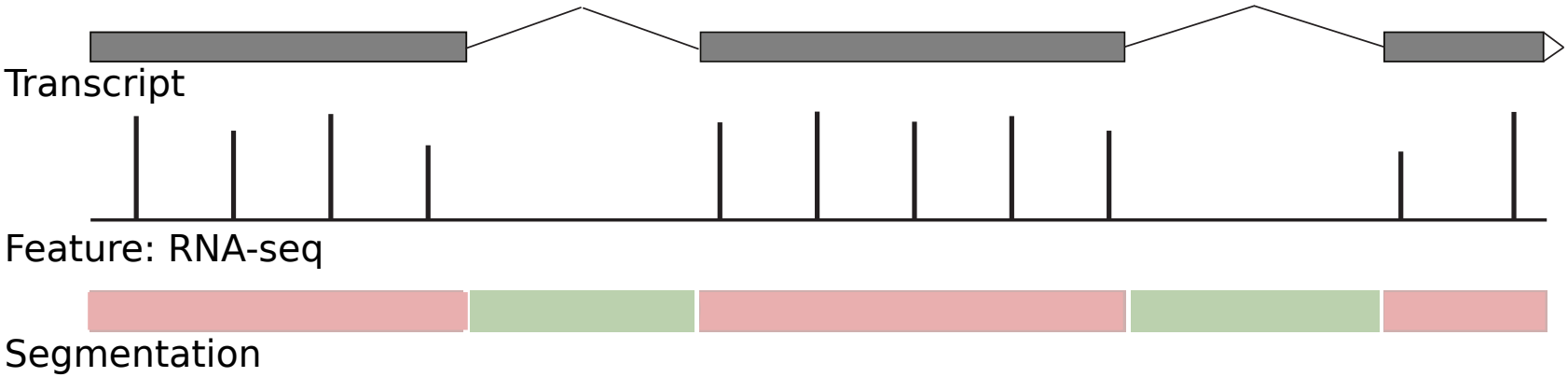
[1] Trapnell et al. 2010 [2] Guttman et al., 2010

mTIM: transcript reconstruction as segmentation task



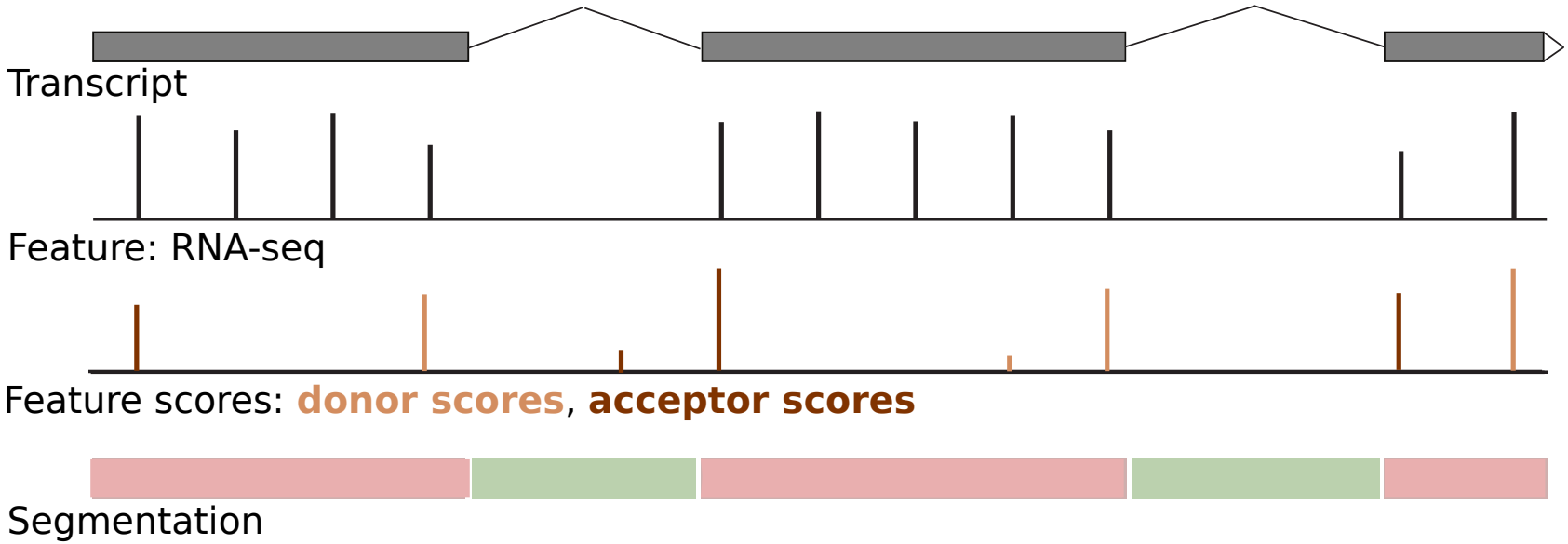
- transcript reconstruction as segmentation task

mTIM: transcript reconstruction as segmentation task



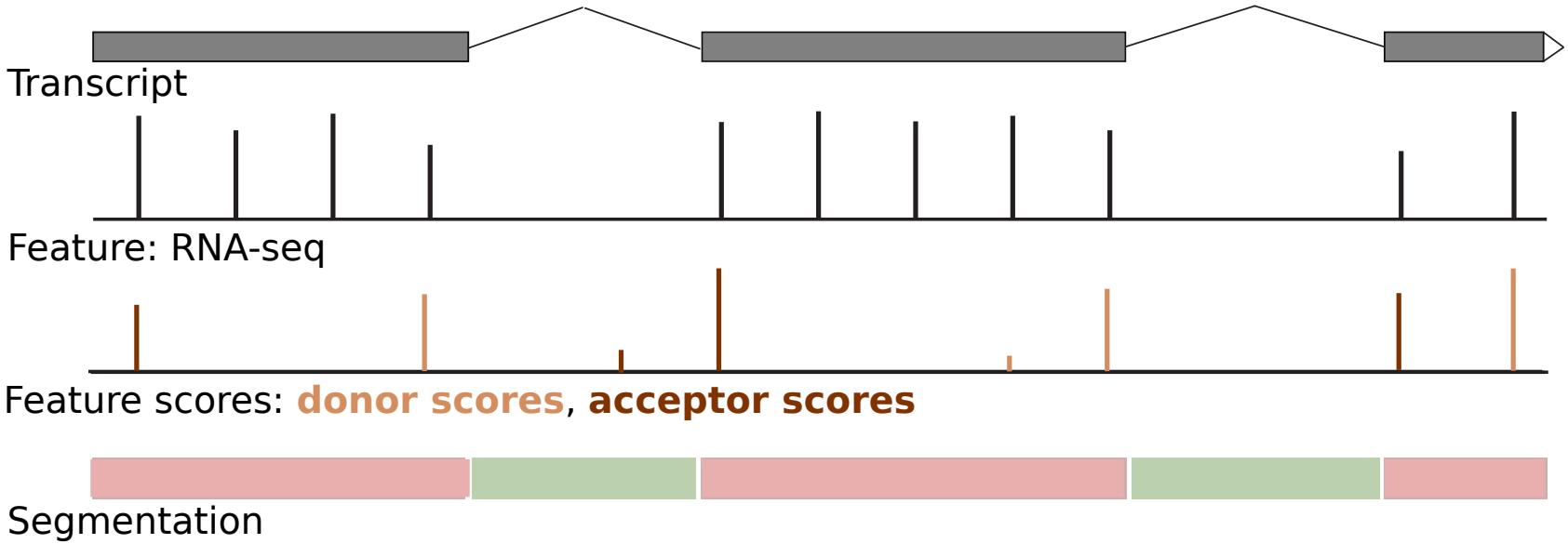
- transcript reconstruction as segmentation task
- use RNA-seq as feature

mTIM: transcript reconstruction as segmentation task



- transcript reconstruction as segmentation task
- use RNA-seq as feature
- use splice-signals as features

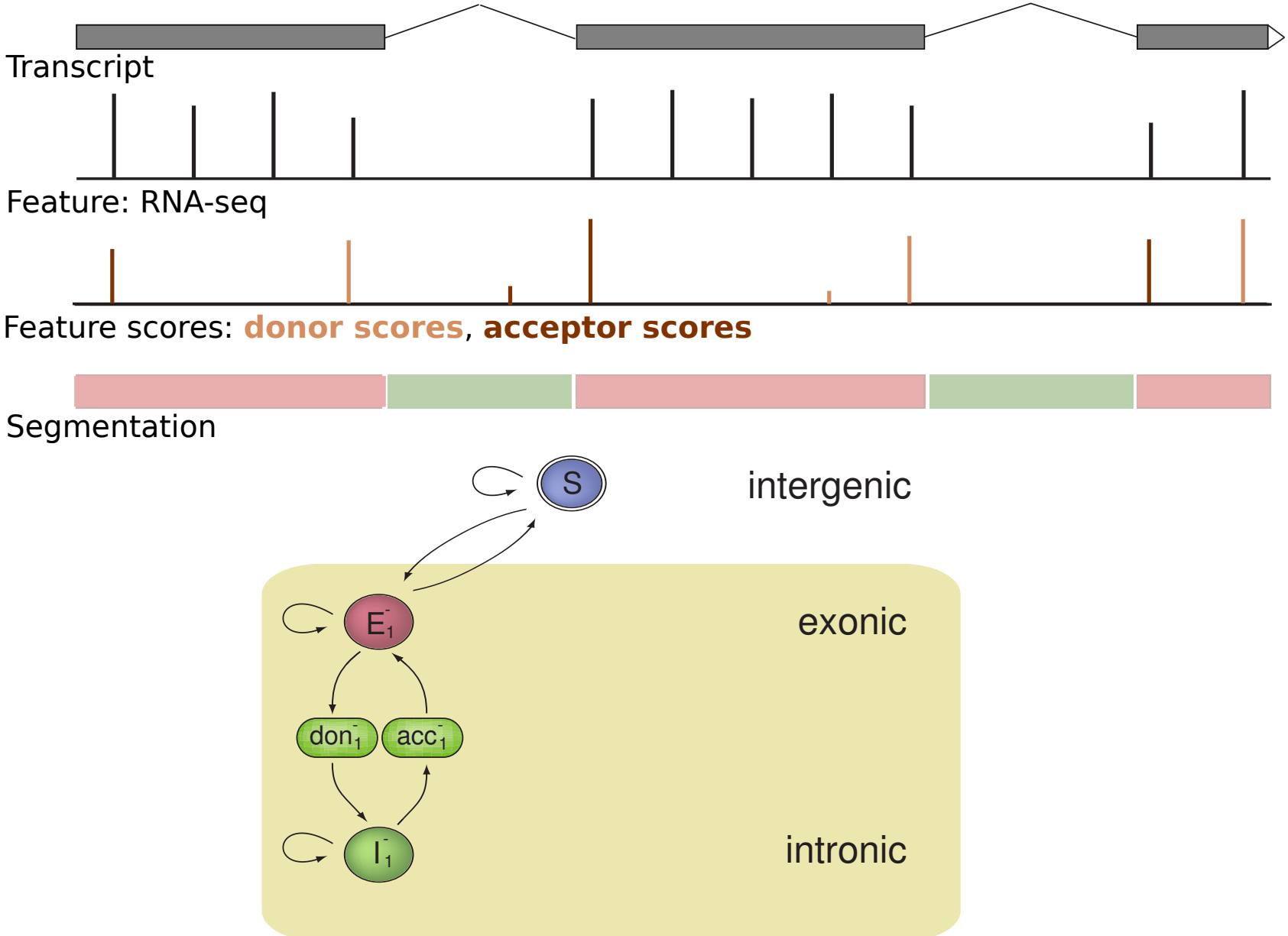
mTIM: transcript reconstruction as segmentation task



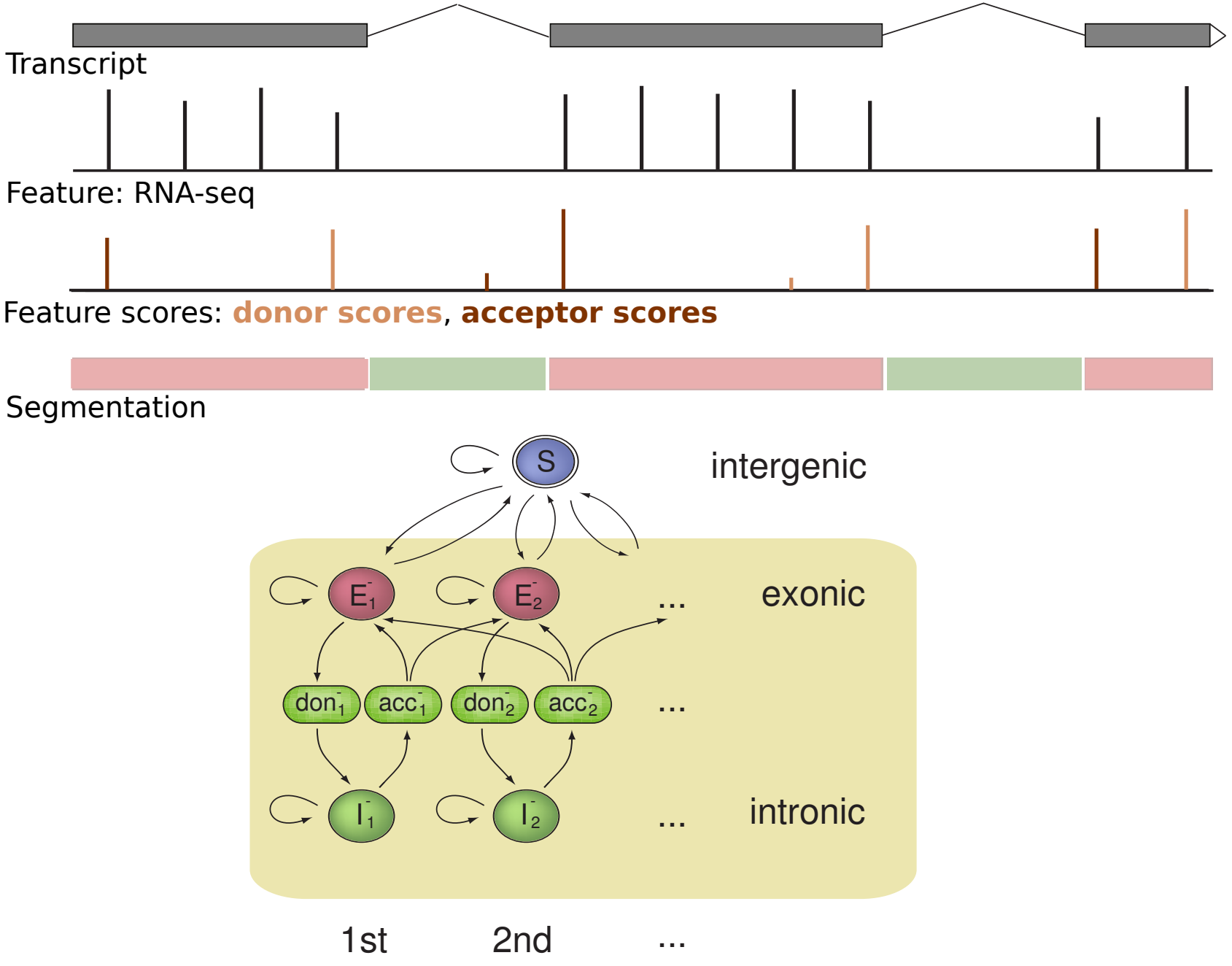
- transcript reconstruction as segmentation task
- use RNA-seq as feature
- use splice-signals as features

$\Rightarrow$ **mTIM finds plausible introns**

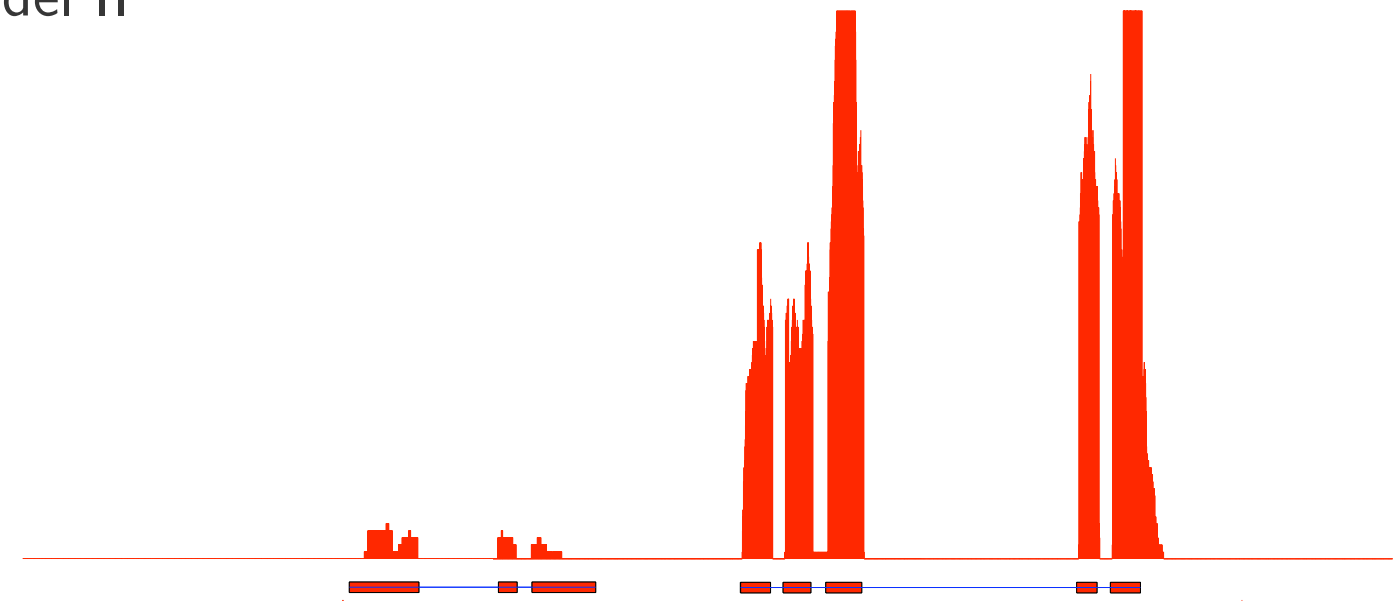
State Model I

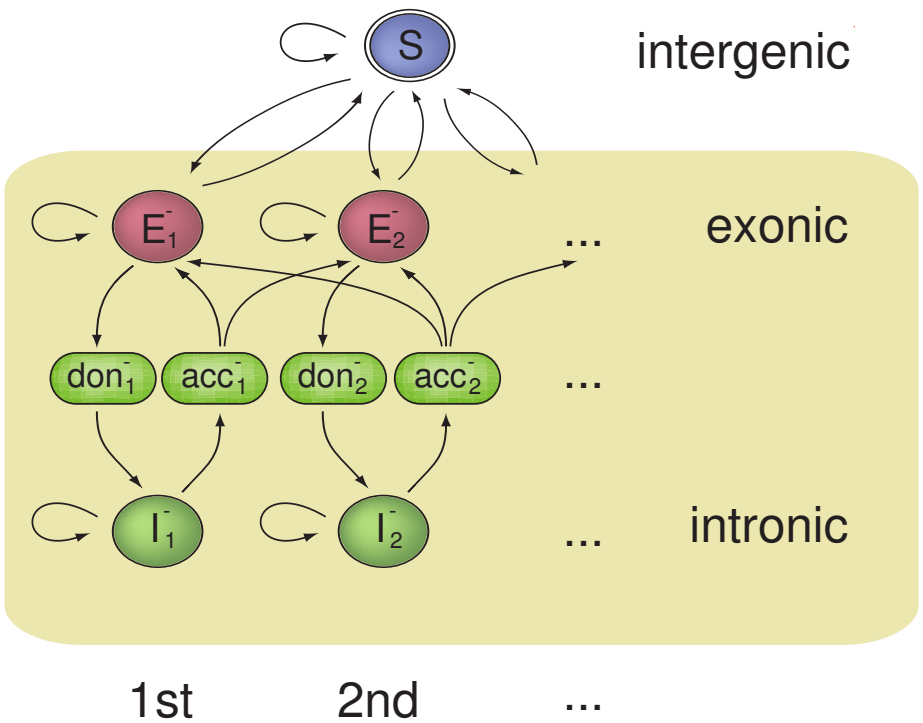


State Model II

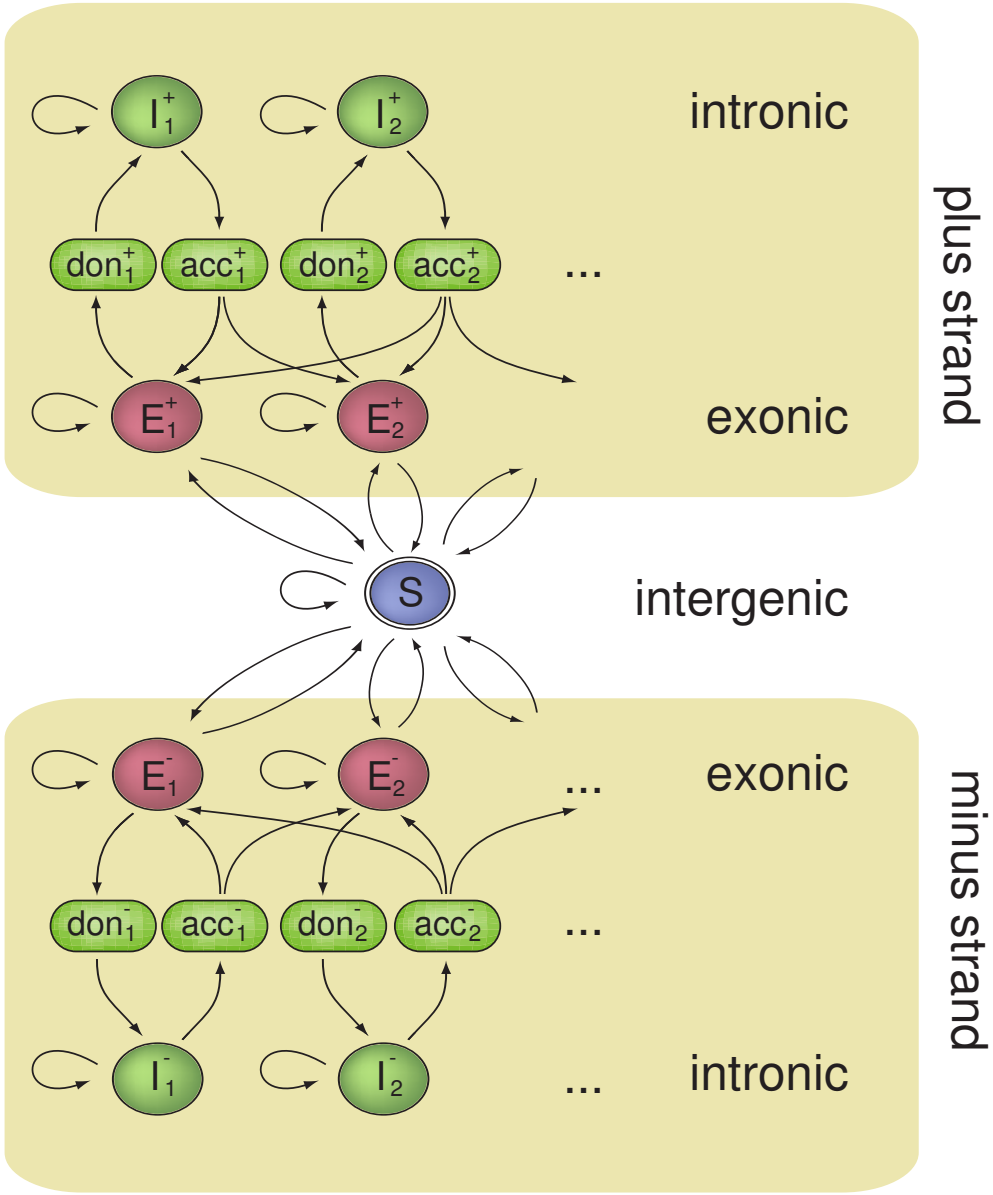


State Model II

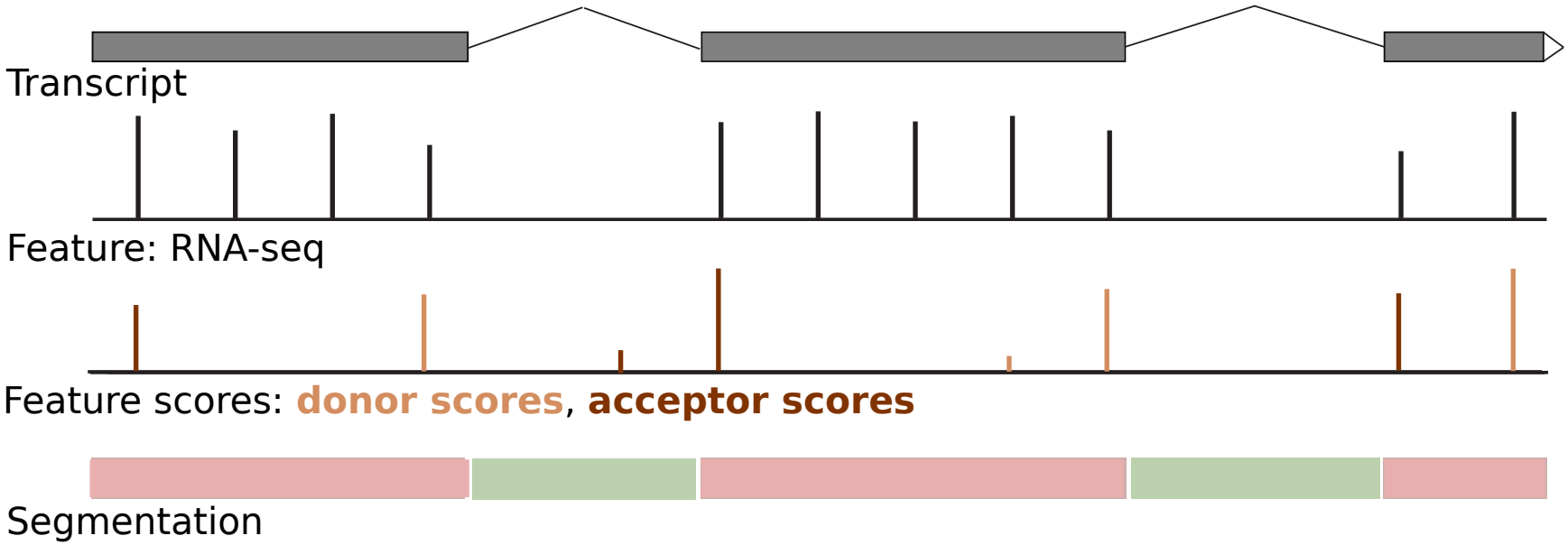




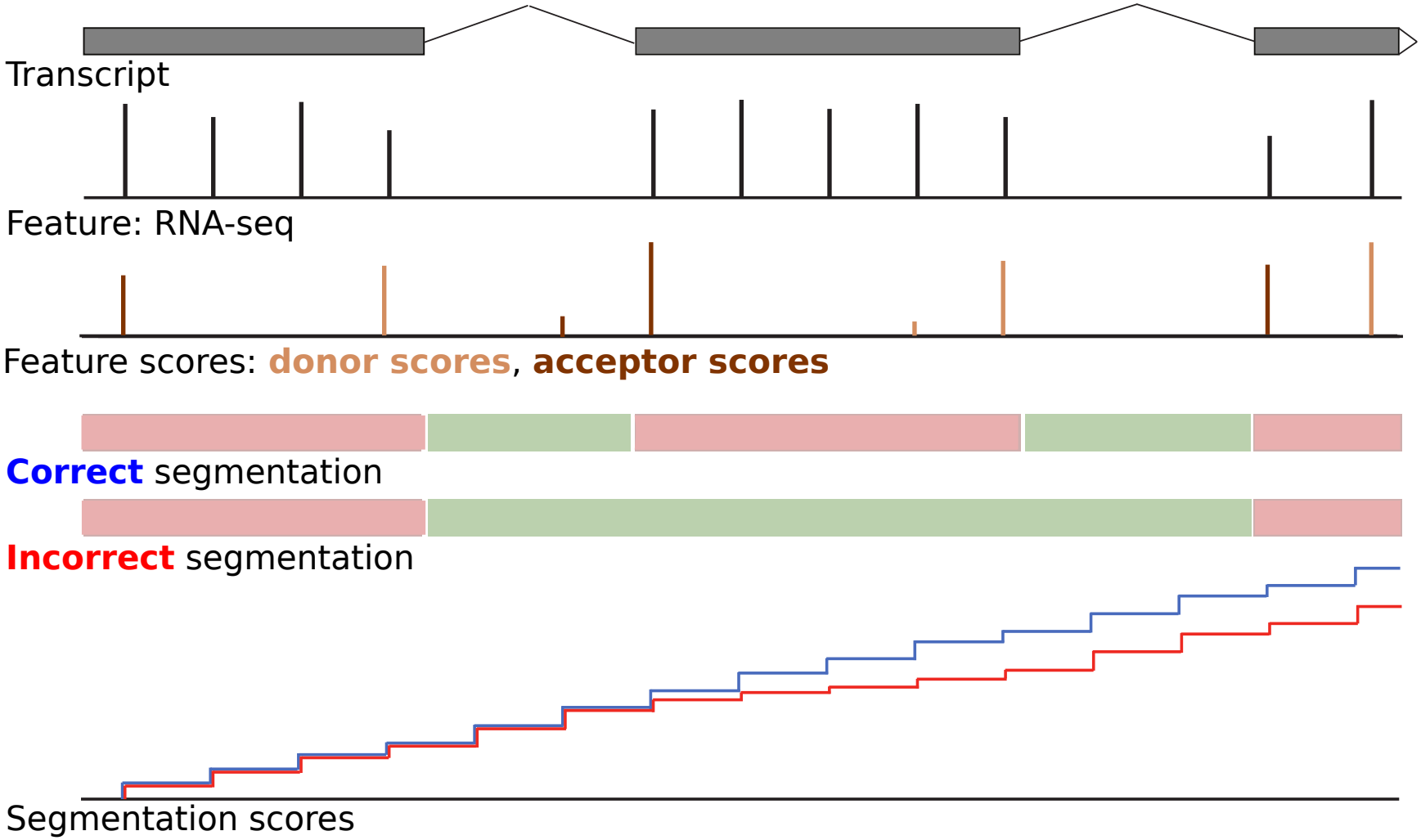
State Model III



Hidden-Markov SVM: Application



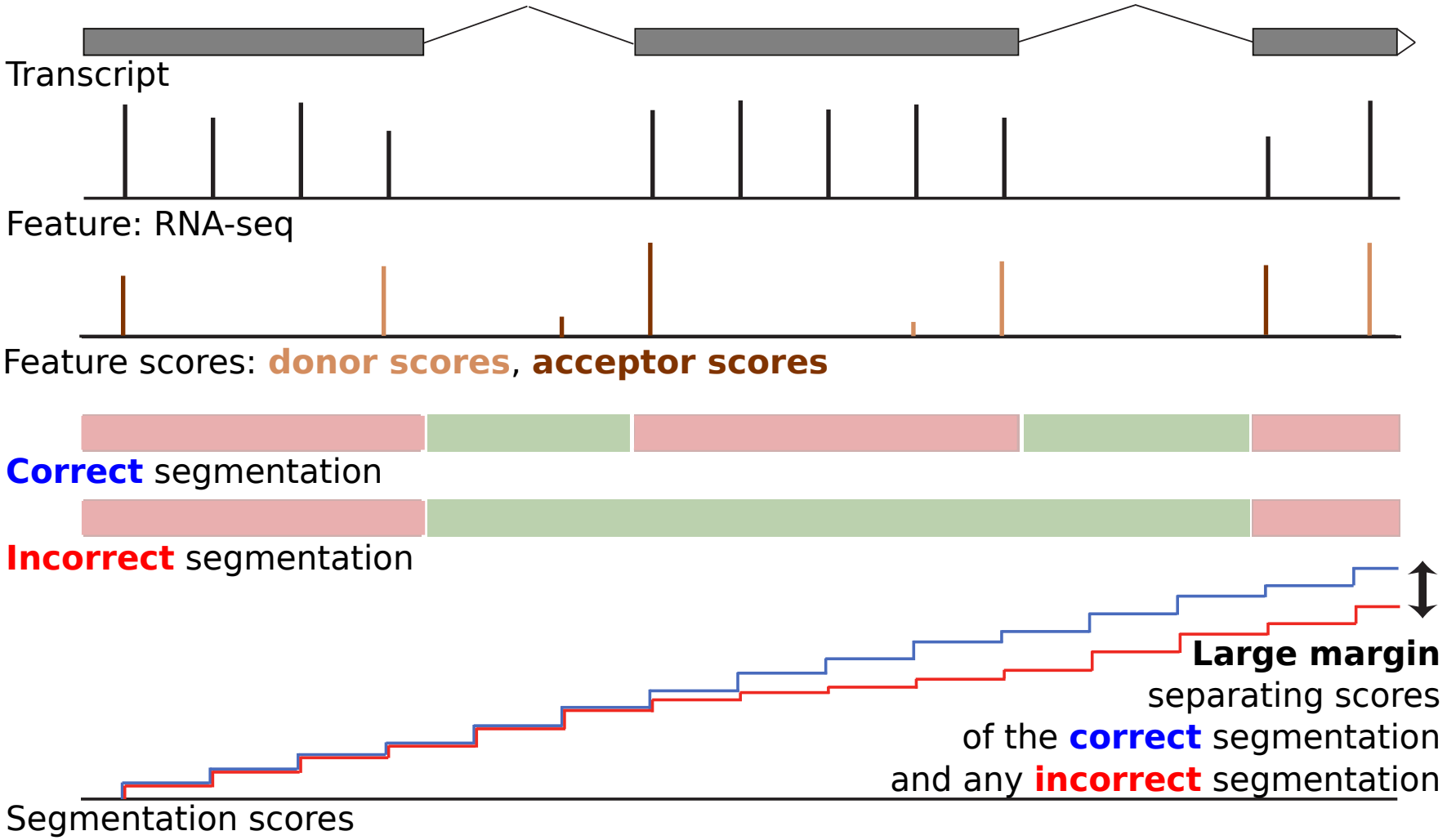
Hidden-Markov SVM: Application



$$\vec{y}_i = \text{argmax}_{\bar{\vec{y}}} F_{\vec{w}}(\bar{\vec{y}}, \vec{x}_i)$$

- parameter vector $\vec{w}$ needs to be adjusted using training data

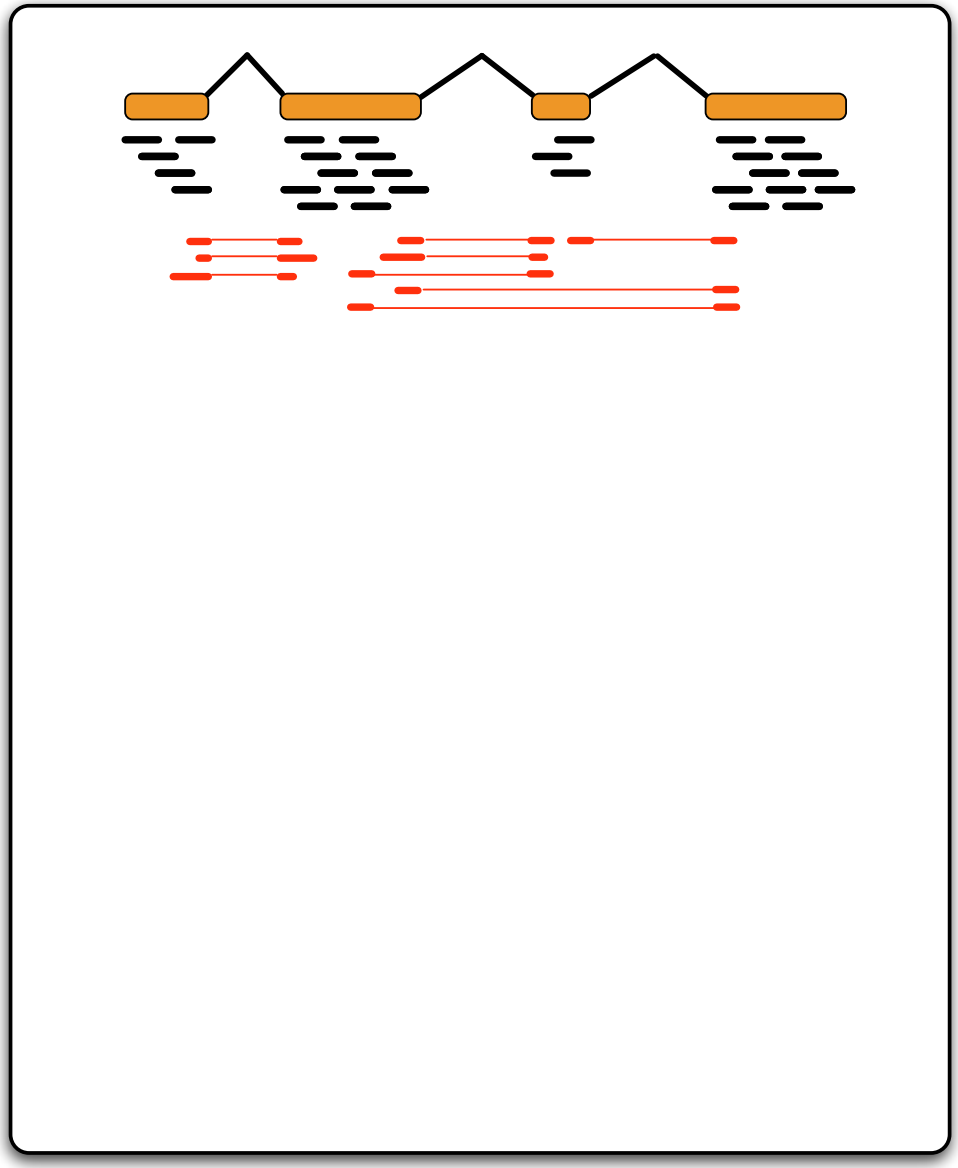
Hidden-Markov SVM: Training



$$\text{maximize}_{\vec{w}}(F_{\vec{w}}(\text{correct}, \vec{x}_i) - F_{\vec{w}}(\text{incorrect}, \vec{x}_i)) \quad \forall i$$

Infer alternative isoforms

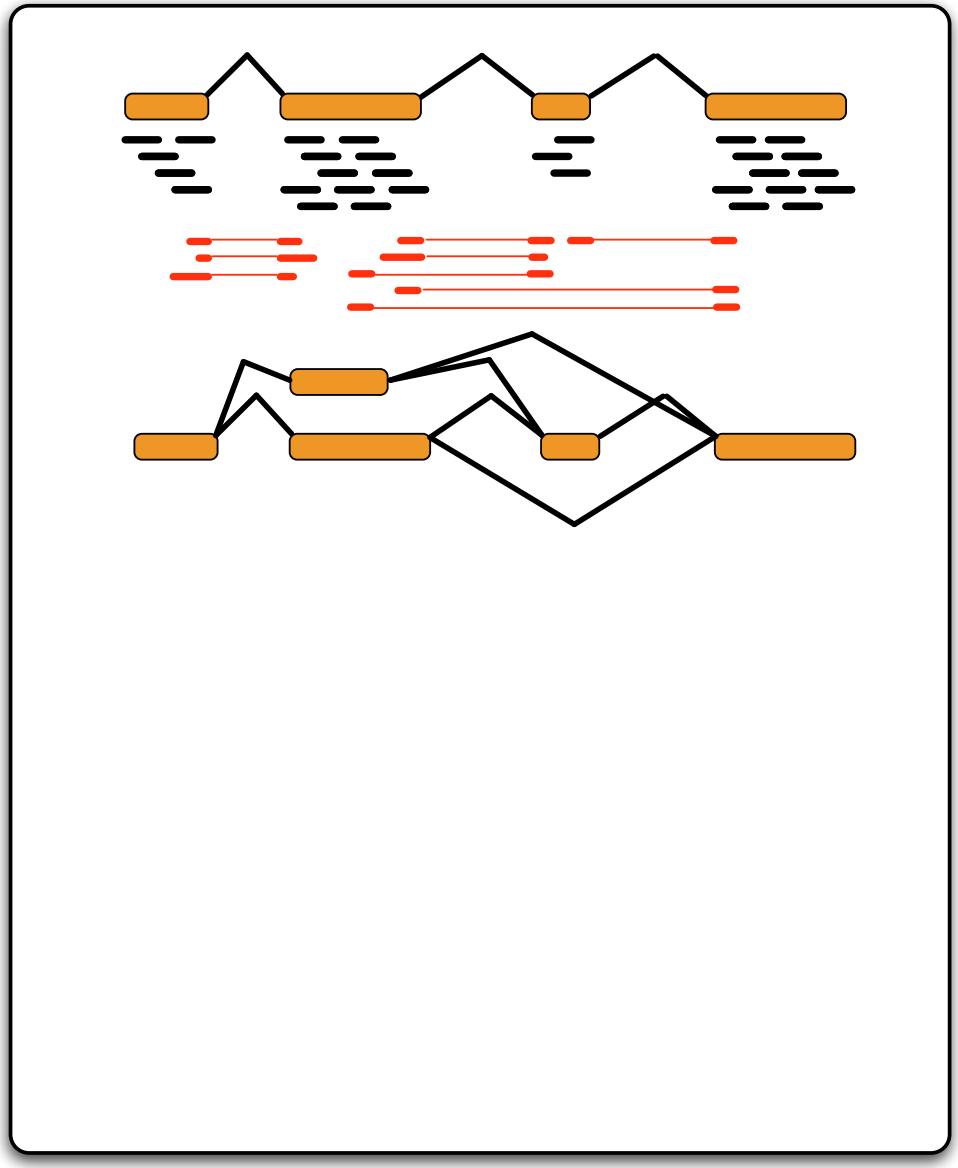
- mTIM prediction



Infer alternative isoforms

- mTIM prediction

- Build splicegraph using spliced reads

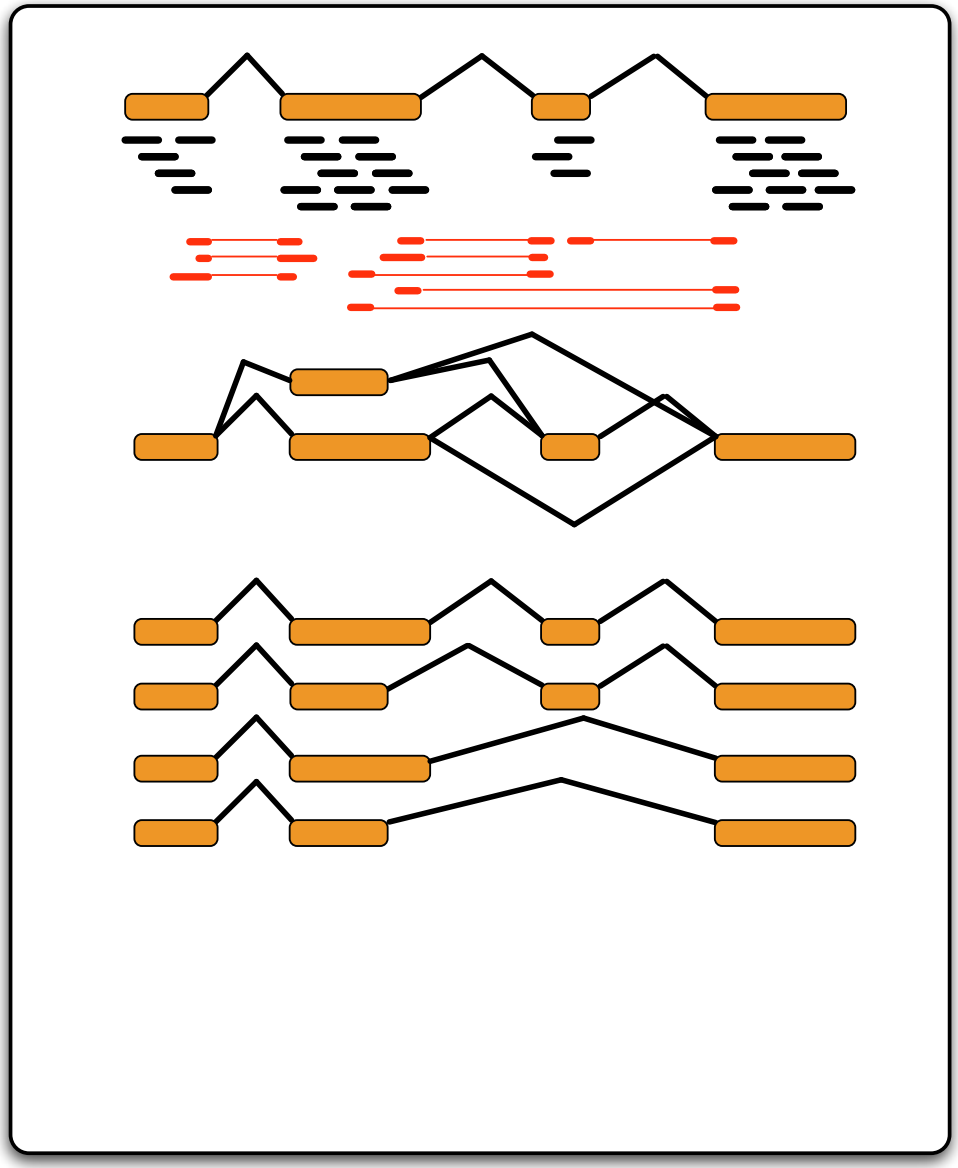


Infer alternative isoforms

- mTIM prediction

- Build splicegraph using spliced reads

- Generate transcripts from graph



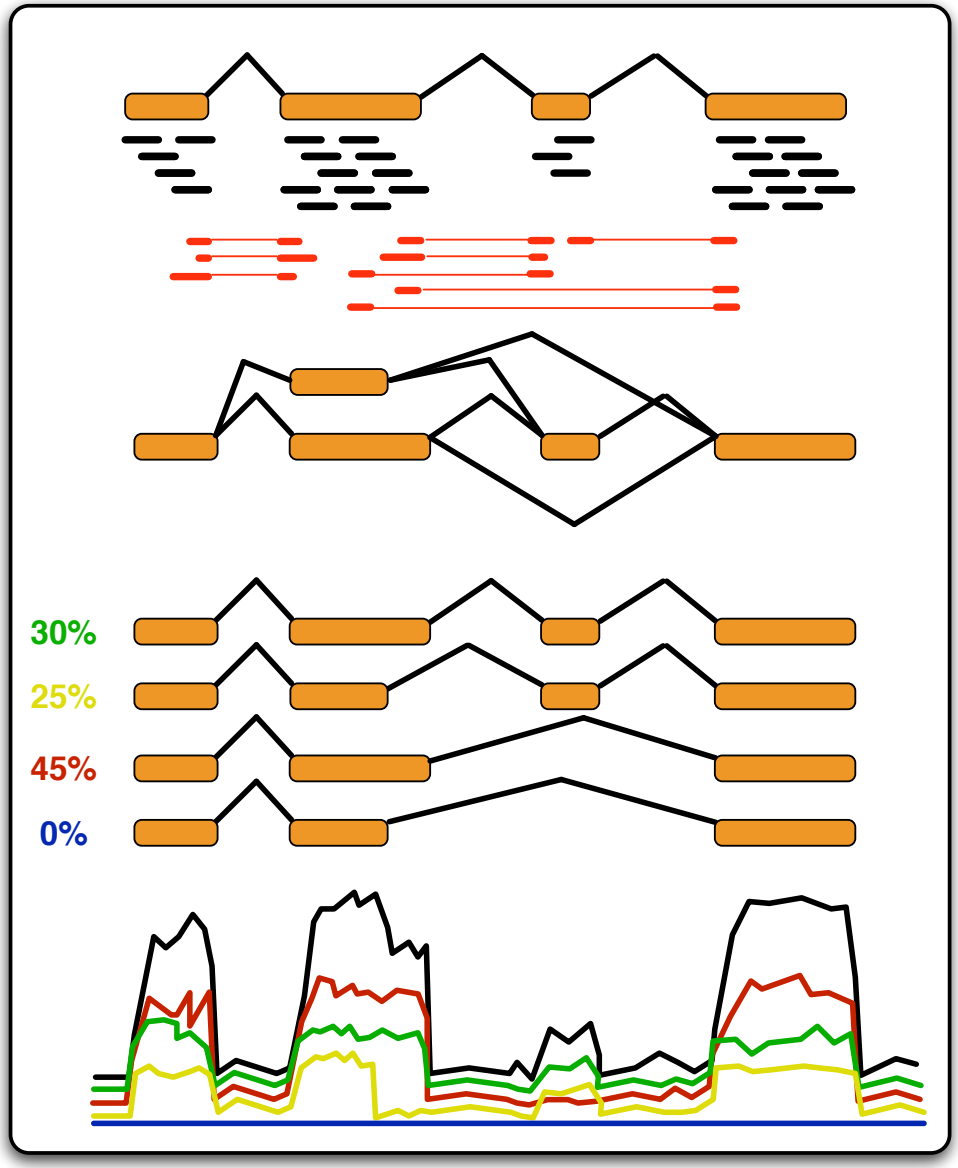
Infer alternative isoforms

- mTIM prediction

- Build splicegraph using spliced reads

- Generate transcripts from graph

- rQuant[1]: explain coverage by weighted sum of transcripts



[1] Bohnert, R et al., 2009

Experiments





Caenorhabditis elegans

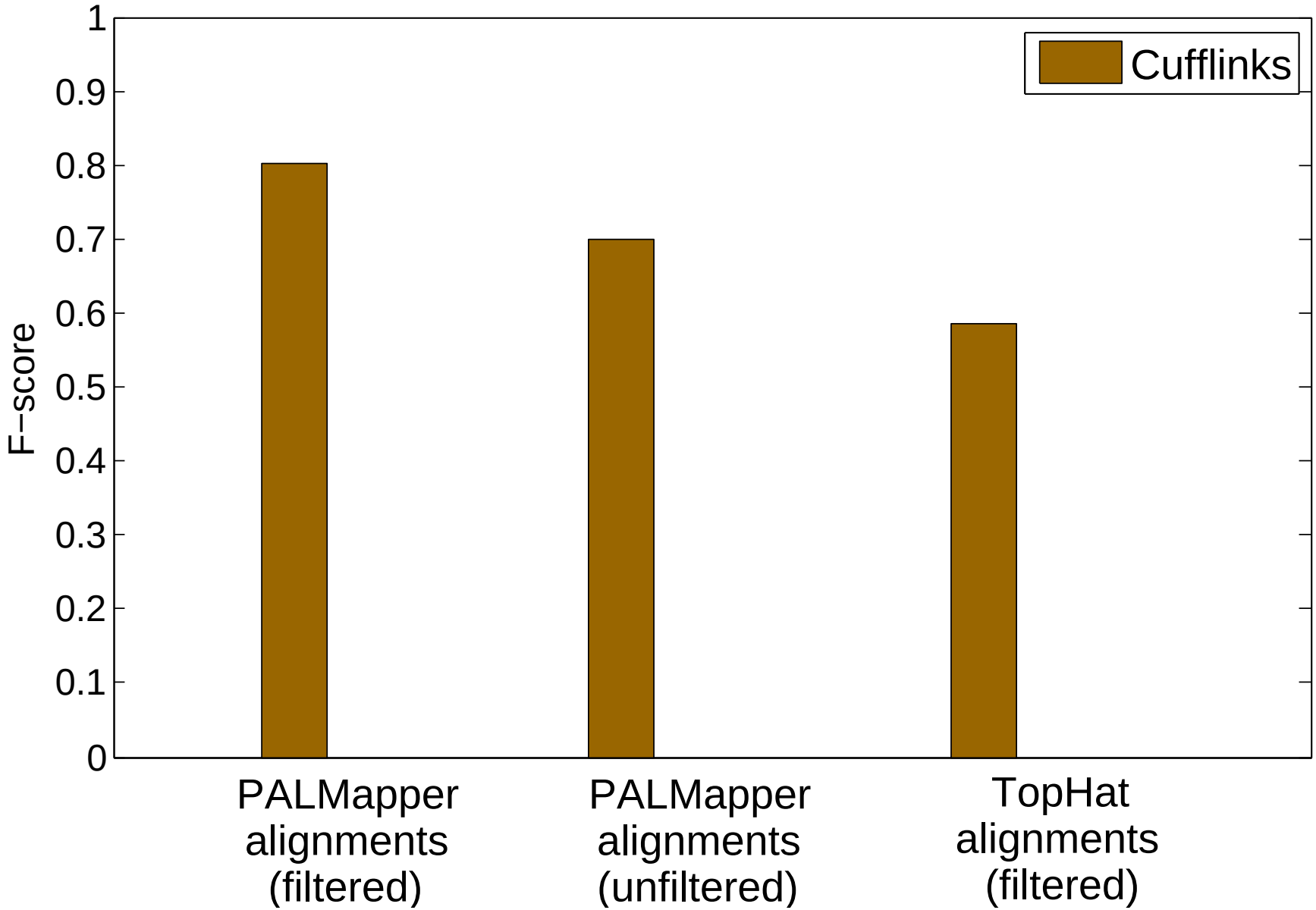
- strand-specific, paired-end Illumina RNA-seq data (2x76bp)
- alignment methods: TopHat[1], PALMapper[2]
- filtered and unfiltered alignments
- for filtering see also poster of Andre Kahles:
  "RGASP Evaluation of RNA-seq Read Alignment Algorithms"

Arabidopsis thaliana

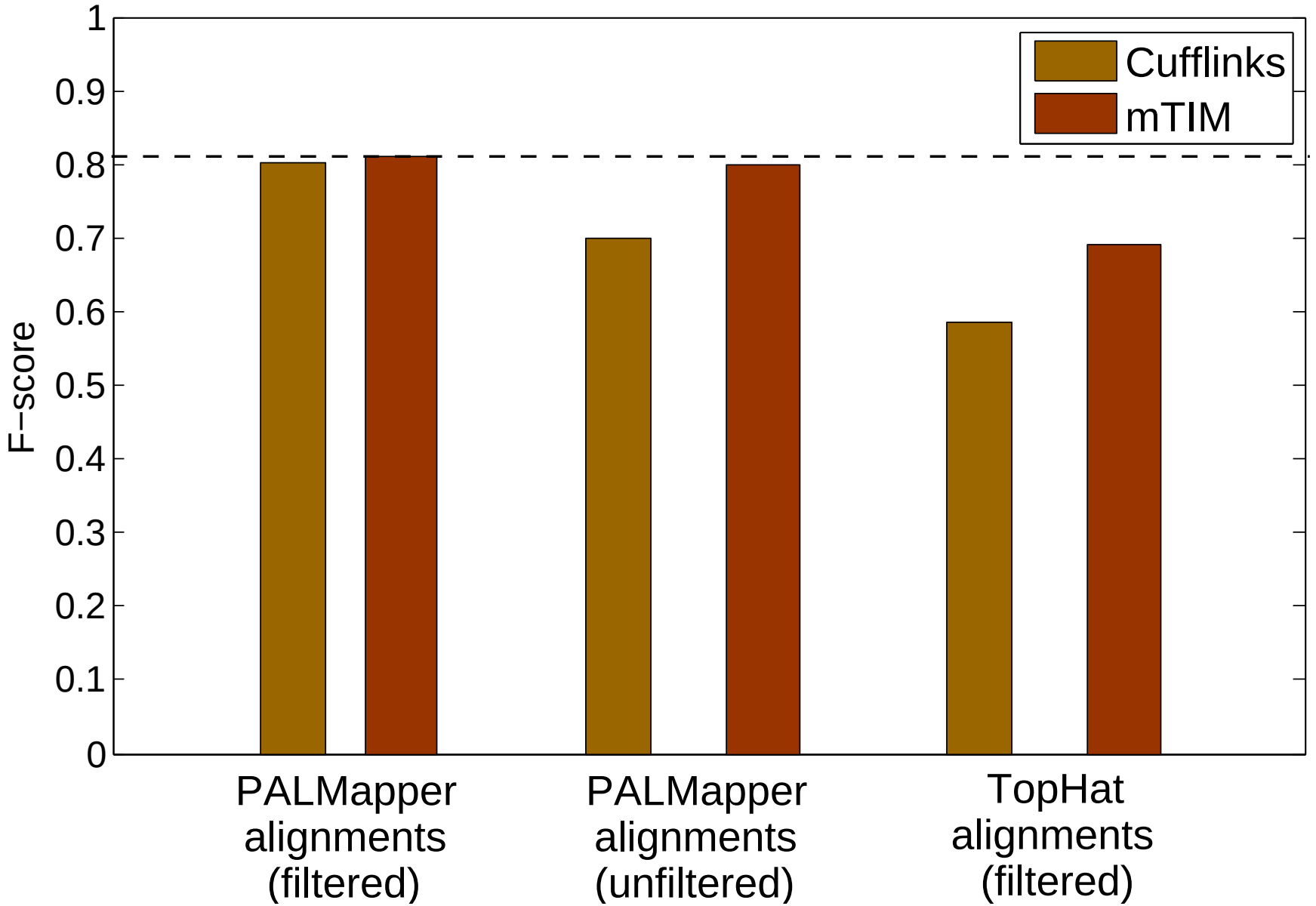
- non-strand-specific, single-end Illumina RNA-seq data (1x36bp)
- unfiltered PALMapper[2]

[1] Trapnell et al., 2009 [2] Jean et al., 2011

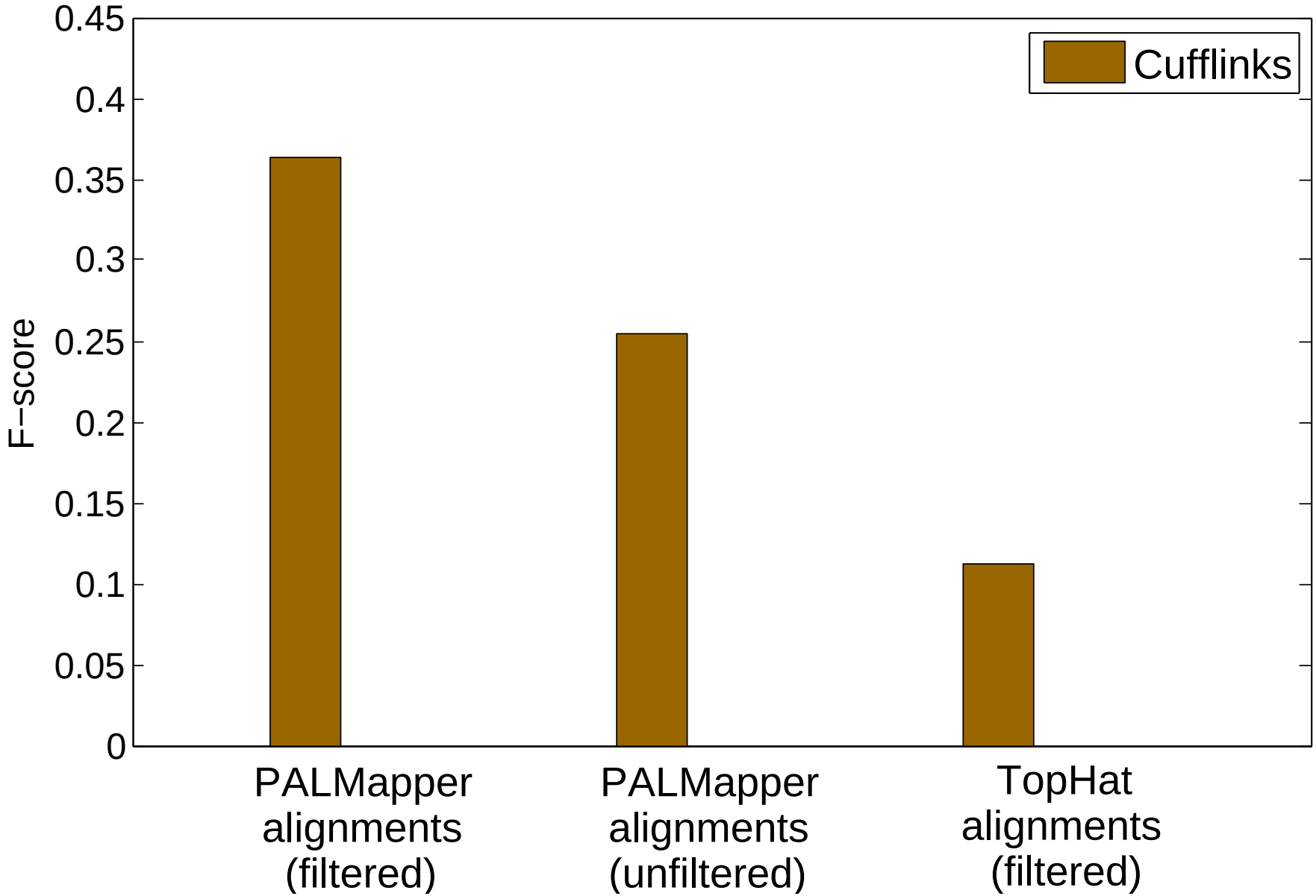
Results: C. elegans intron evaluation



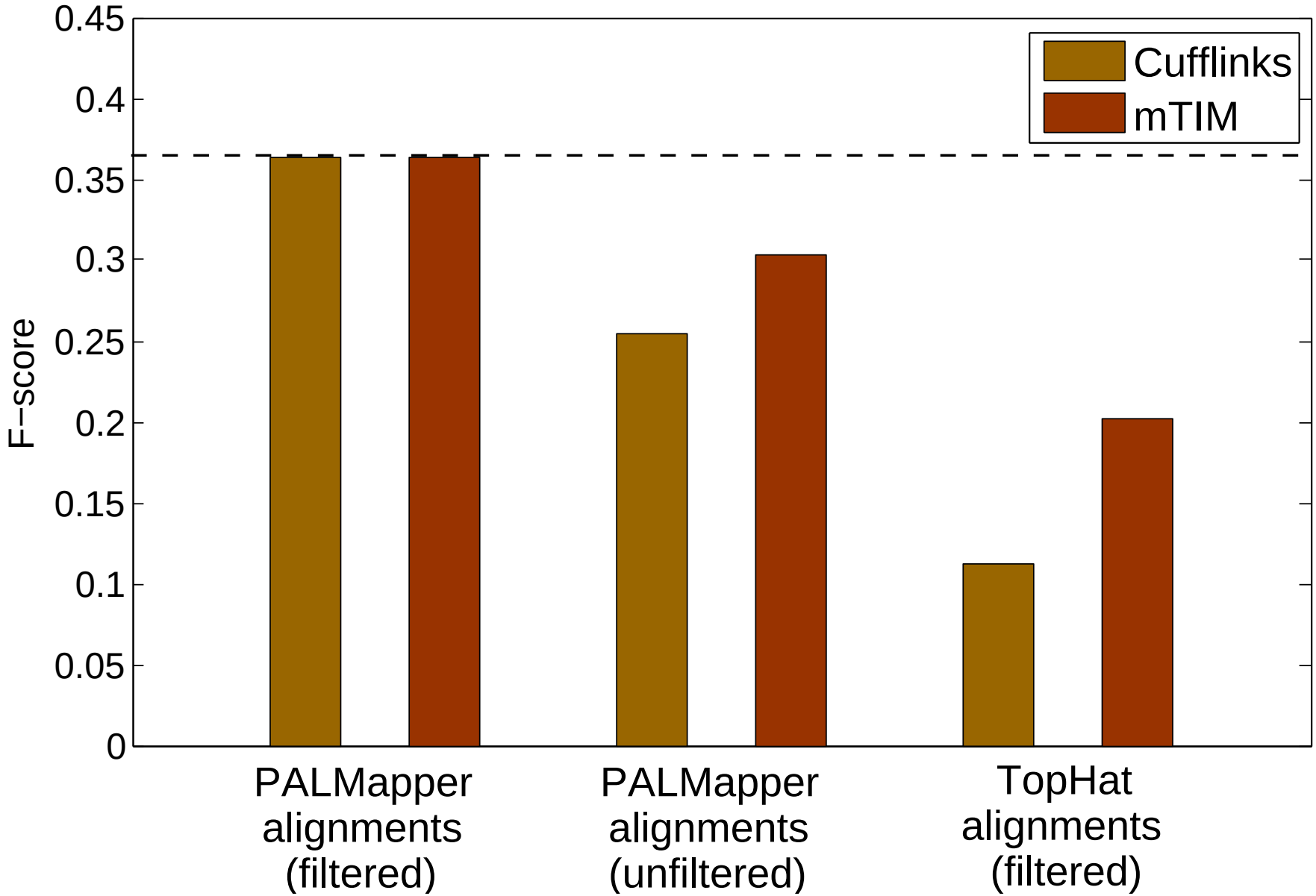
Results: C. elegans intron evaluation



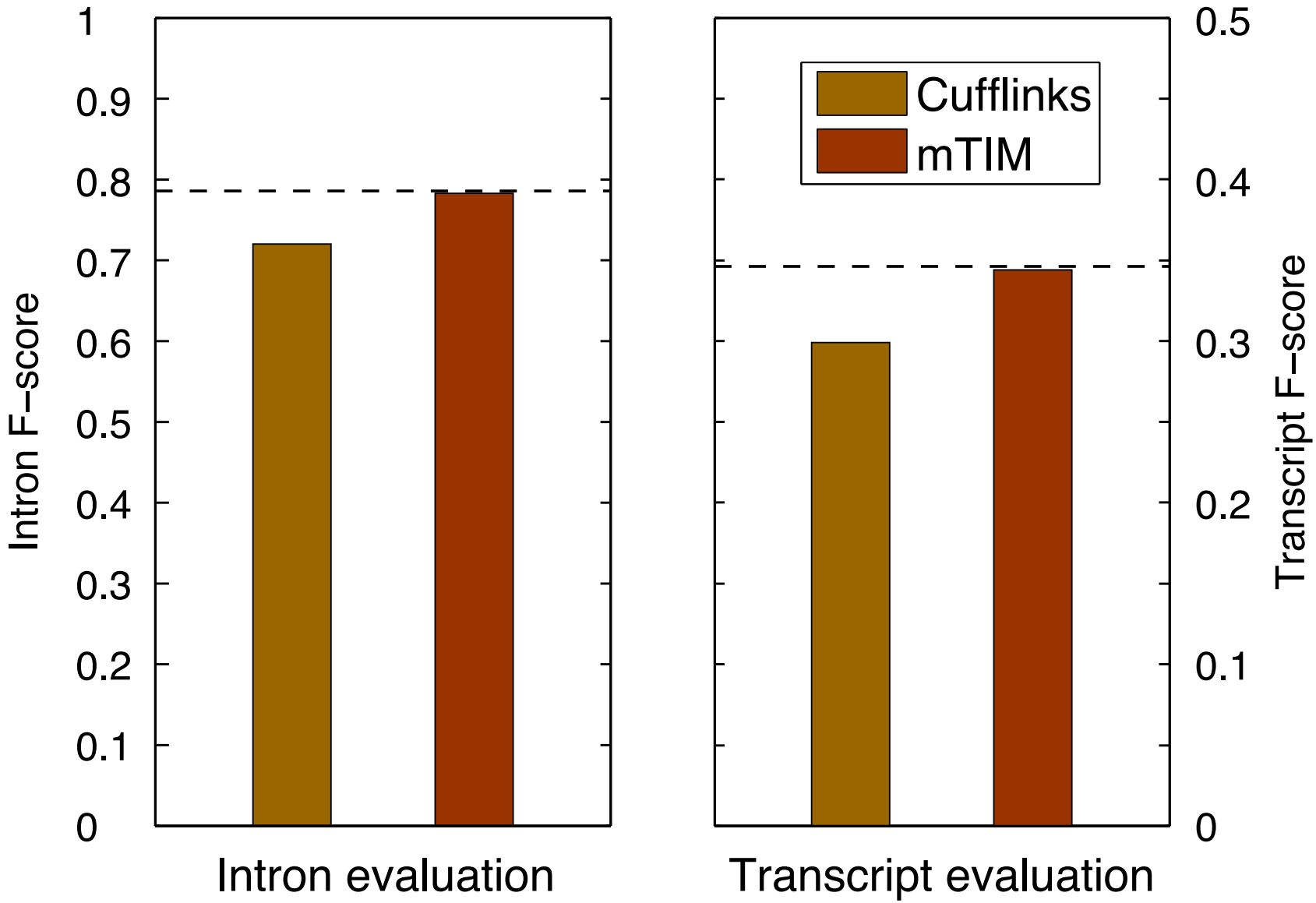
Results: C. elegans transcript evaluation



Results: C. elegans transcript evaluation



Results: A. thaliana evaluation



mTIM: accurate margin-based transcript identification method

Summary

- transcript prediction as segmentation task
- RNA-seq data & splice signals as features
- robust against noisy inputs (able to handle gaps)
- results better or comparable to Cufflinks

Availability

- **http://www.fml.mpg.de/raetsch/suppl/mtim**
  for abstract, slides, (upcoming) source & further information
- Galaxy package in a few weeks

Acknowledgements

Financial support by the German Research Foundation (DFG) under the grant MU 987/6-1 and RA 1894/1-1

Thanks to ..

Andre Kahles, Georg Zeller, Jonas Behr, Regina Bohnert, Geraldine Jean, Sören Sonnenburg, Gunnar Rätsch