University of Zurich[UZH]

Swiss Institute of Bioinformatics

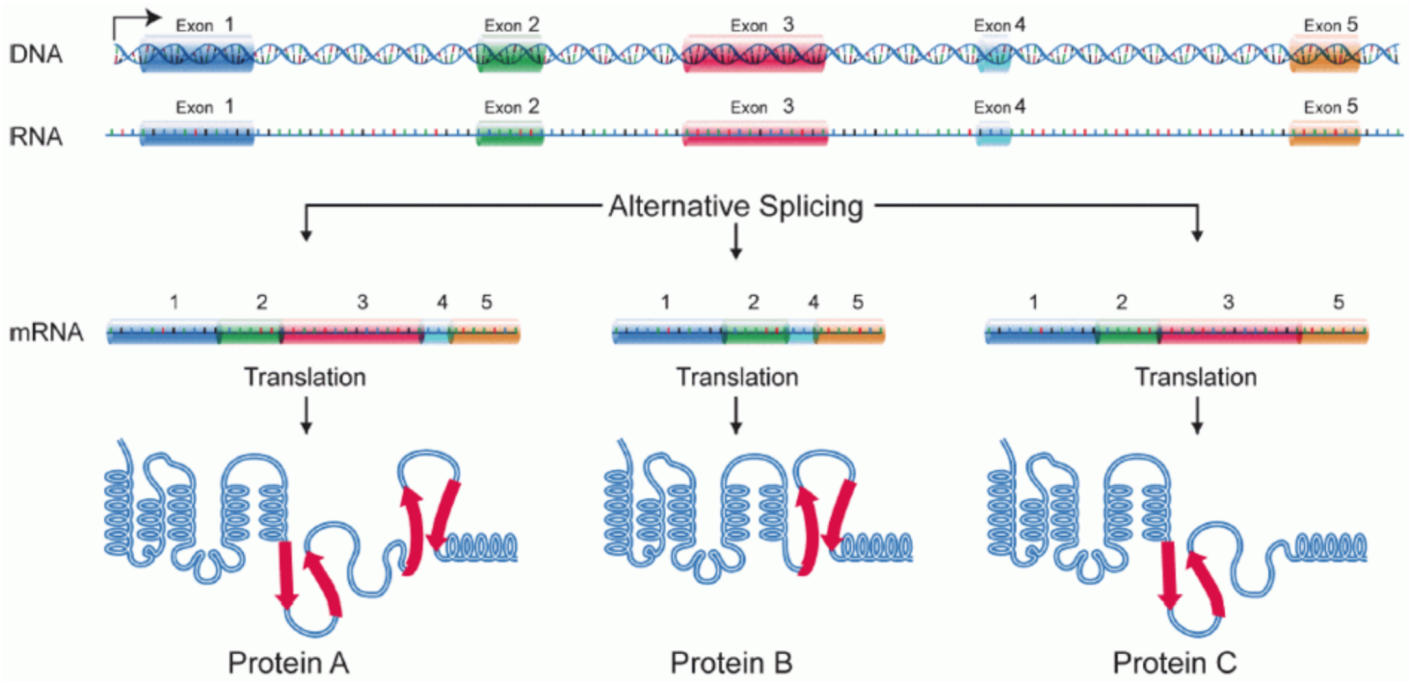
BANDITS: Bayesian differential splicing accounting for sample-to-sample variability and mapping uncertainty

Simone Tiberi and Mark D Robinson

Institute of Molecular Life Sciences, University of Zurich
SIB Swiss Instute of Bioinformatics, University of Zurich

December 10, 2019

Differential splicing



- Differential splicing (DS) studies how alternative splicing patterns change between conditions.
- We present a hierarchical Bayesian method for DS, based on RNA-seq data.
- The tool is distributed as a Bioconductor R package: BANDITS.

Mapping uncertainty

Slide adapted from Trapnell et al. (2013), Nat Biotech

- A big mathematical challenge in differential splicing analyses is that transcript level counts are not observed because most reads map to multiple transcripts.
- Most DS methods use transcript level estimated counts, obtained via EM algorithms (e.g., Salmon and kallisto); however the uncertainty in their estimate is typically neglected.
- Other methods instead (including BANDITS), avoid the quantification step and input the equivalence classes of reads (i.e., what transcripts each read is compatible with): BANDITS samples the transcript (and gene) allocation of reads.

Dirichlet-Multinomial hierarchical model

- Consider a gene with K transcripts and N samples in a given group.
- The transcript level counts for an individual sample are assumed to follow a Multinomial distribution:

$$X^{(i)}|\pi^{(i)} \sim Multinom\left(n^{(i)}, \pi^{(i)}\right), i = 1, ..., N, \quad (1)$$

  where $\pi^{(i)} = \left(\pi_1^{(i)}, ..., \pi_K^{(i)}\right)$ indicates the relative expression of transcripts $1, ..., K$ within the gene and $n^{(i)} = \sum_{k=1}^{K} X_k^{(i)}$.
- $\pi^{(i)}$ is assumed to vary between samples due to biological variation; *a priori* we assume:

$$\pi^{(i)} \sim Dirichlet(\delta), i = 1, ..., N, \quad (2)$$

  where $\delta = (\delta_1, \ldots, \delta_K)$.
- We test if the mean relative abundance of transcripts, $\bar{\pi} = \frac{\delta}{\sum_{k=1}^{K} \delta_k}$, varies between conditions.

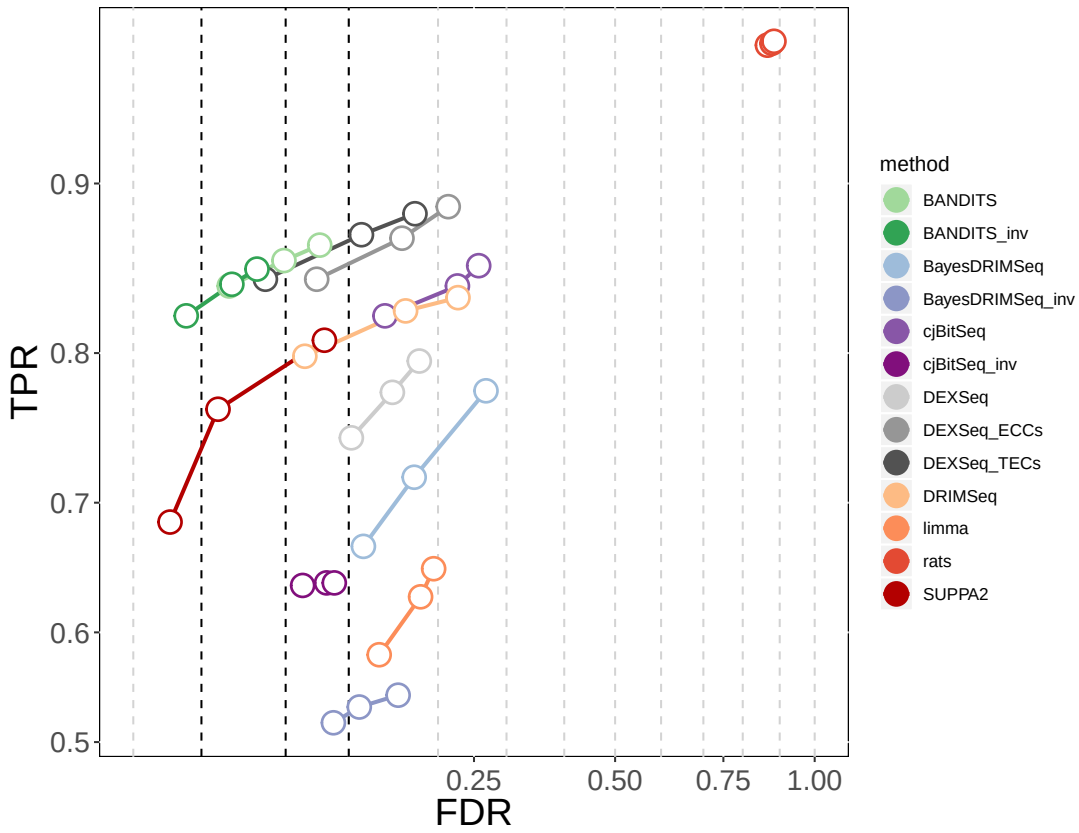
In a nutshell

- BANDITS:
  - inputs equivalence classes of reads and samples their transcript (and gene) allocations;
  - allows reads to be aligned to the transcriptome (with Salmon or kallisto) or to the genome (with STAR);
  - uses a hierarchical structure, to model the variability between biological replicates;
  - tests for differential splicing, both, at the gene and transcript level;
  - corrects for the different lengths of transcripts;
  - is computationally efficient: a 6 vs 6 group comparison (human genome) runs in a laptop in < 2 h;
  - also provides a conservative score (BANDITS_inv) which accounts for the inversion of the dominant transcript.

Benchmarking

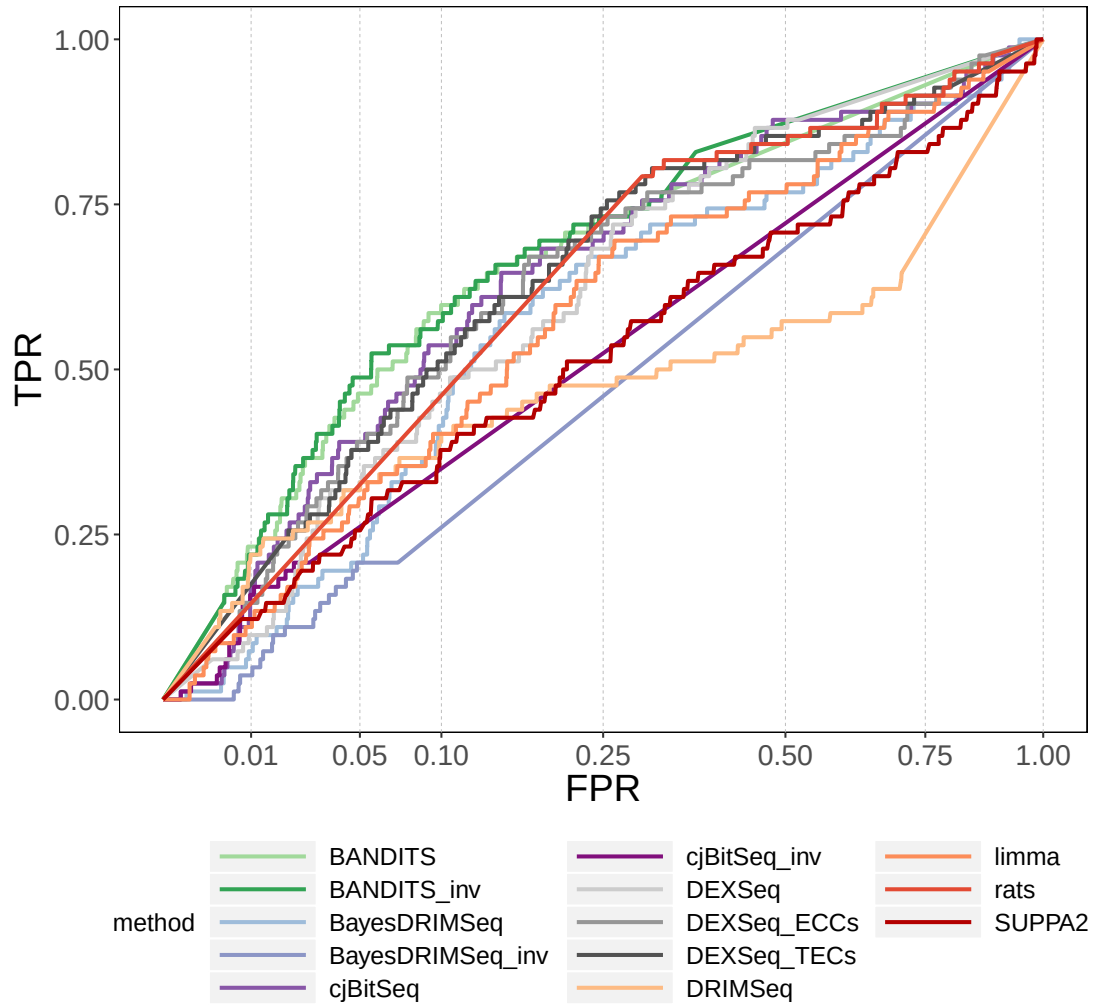
- We benchmarked our method against several competitors in three simulated and two experimental datasets (all human genome).
- Simulation studies:
  - a 3 vs 3 two-group comparison (not shown);
  - a 6 vs 6 two-group comparison (not shown);
  - a 6 vs 6 two-group comparison (with transcript pre-filtering);
- Experimental data:
  - "Best et al. data" (Best et al., 2014), with a 3 vs 3 two-group comparison, with 82 validated genes (via PCR);
  - a "null" experimental dataset (Kim et al., 2013), with a 3 vs 3 two-group comparison, where all samples belong to the same group of healthy patients.

Simulation study



TPR vs FDR for the 6 vs 6 simulation study with transcript pre-filtering.

Best et al. data

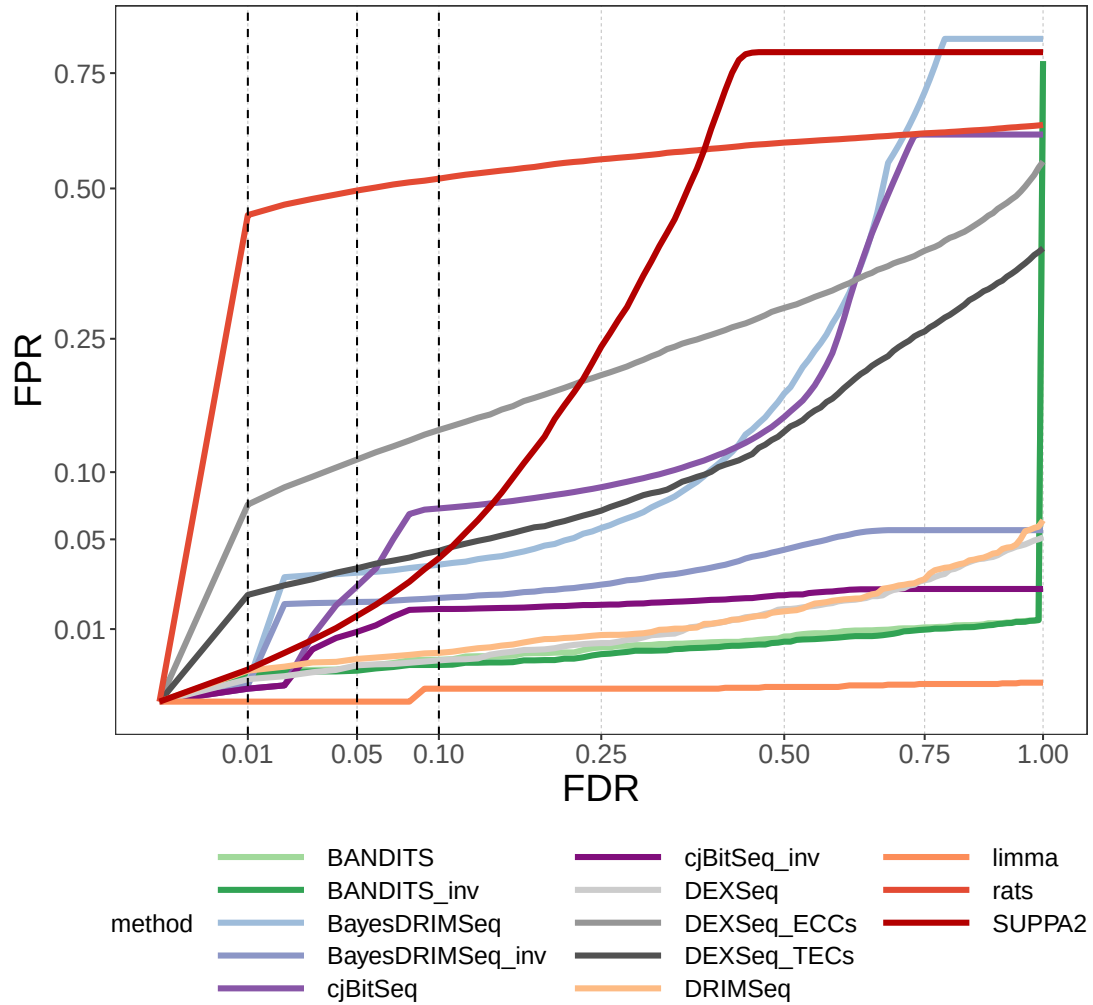


TPR vs FPR for the "Best et al." data analysis.

Best et al. data

	Median position	AUC	pAUC 0.1	pAUC 0.2
BANDITS_inv	596.00	0.81	0.04	0.11
BANDITS	672.75	0.80	0.04	0.11
cjBitSeq	900.00	0.79	0.04	0.10
rats	942.50	0.80	0.03	0.10
DEXSeq_TECs	968.00	0.79	0.03	0.09
DEXSeq_ECCs	1039.00	0.78	0.03	0.10
BayesDRIMSeq	1231.00	0.74	0.02	0.08
DEXSeq	1348.00	0.78	0.03	0.08
limma	1556.00	0.74	0.03	0.08
SUPPA2	2109.75	0.67	0.02	0.07
DRIMSeq	3248.00	0.59	0.03	0.07
cjBitSeq_inv	5146.50	0.59	0.02	0.05
BayesDRIMSeq_inv	5362.00	0.57	0.02	0.04

Null experimental data



FPR vs FPR for the "null." experimental data analysis.

Availability and Acknowledgements

- Bioconductor R package:
  - `https://bioconductor.org/packages/BANDITS`
  - `https://github.com/SimoneTiberi/BANDITS`
- Pre-print
  - Tiberi and Robinson, biorxiv (2019). BANDITS: Bayesian differential splicing accounting for sample-to-sample variability and mapping uncertainty.
    `https://www.biorxiv.org/content/10.1101/750018v1`
- Acknowledgements:
  - Mark D Robinson;
  - Charlotte Soneson and the Robinson lab;
  - Panagiotis Papastamoulis, Magnus Rattray and David Rossell.