Subject Section

flexiMAP: A regression-based method for discovering differential alternative polyadenylation events in standard RNA-seq data

Krzysztof J. Szkop[1], David S. Moss[1] and Irene Nobeli[1,*]

1 Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck, Malet Street, London WC1E 7HX, UK

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: We present flexiMAP (flexible Modeling of Alternative PolyAdenylation), a new beta-regression-based method implemented in *R*, for discovering differential alternative polyadenylation events in standard RNA-seq data.

Results: We show, using both simulated and real data, that flexiMAP exhibits a good balance between specificity and sensitivity and compares favourably to existing methods, especially at low fold changes. In addition, the tests on simulated data reveal some hitherto unrecognised caveats of existing methods. Importantly, flexiMAP allows modeling of multiple known covariates that often confound the results of RNA-seq data analysis.

Availability: The flexiMAP *R* package is available at: https://github.com/kszkop/flexiMAP

Scripts and data to reproduce the analysis in this paper are available at:

https://doi.org/10.5281/zenodo.3689788

Contact: Irene Nobeli, i.nobeli@bbk.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Alternative polyadenylation (APA) is the selection of alternative cleavage and polyadenylation sites during transcription of eukaryotic genes, resulting in isoforms with distinct lengths. APA has been shown to be prevalent in mammalian transcripts and alternative isoforms are linked to different stages of development, cell types and disease status (Elkon *et al.*, 2013; Szkop *et al.*, 2017). APA events can be identified on a genome-wide scale using 3' end-focused sequencing (e.g. QuantSeq (Moll *et al.*, 2014)) or, more recently, long-read sequencing (Iso-seq (Anvar *et al.*, 2018) and nanopore-based sequencing (Garalde *et al.*, 2016)). However, as these methods are still not widely used and many legacy transcriptome surveys were carried out using standard RNA-seq sequencing instead, it would be useful to have computational methods that can identify differential APA in RNA-seq data. A few such methods exist already (Xia *et al.*, 2014; Grassi *et al.*, 2016; Ha *et al.*, 2018; Ye *et al.*, 2018; Arefeen *et al.*, 2018) but they have caveats (Szkop and Nobeli, 2017). For example, all methods must solve the problem of how to deal with biological replicates; some test the replicates individually, losing the advantage of having replicates in the first place; others, average values from replicates, effectively losing track of the within-group variability. In designing a method for differential APA analysis, we considered the following: a) the reconstruction and quantification of the

individual isoforms is both challenging and not strictly necessary for this task; b) the errors in modeling RNA-seq read counts are neither normal nor Poisson-distributed; c) multiple covariates can affect APA.

Inspired by the use of Generalized Linear Models (GLMs) in differential gene expression (Robinson *et al.*, 2010; Love *et al.*, 2014) we present here a regression-based method and associated pipeline (flexible modeling of APA or flexiMAP) that satisfactorily addresses the above requirements. We show, using simulated data, that the method is both sensitive and specific across a range of fold changes and numbers of samples and that its performance is superior to two alternatives (DaPars (Xia *et al.*, 2014), and APAtrap (Ye *et al.*, 2018)) in most tests we carried out. FlexiMAP is also outperforming both these methods and Roar (Grassi *et al.*, 2016), when additional covariates confound changes to the isoform ratios. Tested on real RNA-seq data, flexiMAP is slightly less specific than the other methods tested but outperforms all methods when the Matthews Correlation Coefficient is used as the measure of performance, indicating a better overall balance between specificity and sensitivity.
The method is available as an *R* package from: https://github.com/kszkop/flexiMAP

2 Methods

Our method can be applied to all pairs of polyadenylation sites in a gene, where one site is considered "distal" (i.e. located furthest away from the end of the coding region) and one is "proximal" (Supp. Fig. 1). Given a list of sites provided to the program, pairs of sites will be considered in turn, the most downstream site of the transcript being the distal site in all pairs. The proximal site separates the 3' UTR into two regions: the "short" region, starting at the end of the coding region and ending at the proximal site, and the "long" region starting at the proximal site and ending at the end of the transcript (Supp. Fig. 1). Assuming the separation of samples into groups based on the condition of interest, the question we want to answer is: given a total number of reads falling in the 3' UTR, is the proportion of reads falling in the long region dependent on the sample group membership?
We count RNA-seq reads falling in the "long" and "short" regions of the 3' UTR (N_{long}^{ij} and N_{short}^{ij} respectively), and define the ratio, R, for gene i in sample j as:

$$R_{ij} = \frac{N_{long}{}^{ij}}{N_{short}{}^{ij} + N_{long}{}^{ij}} \quad (1)$$

Reads falling in the long region can only originate from transcripts using the distal site, whereas reads falling in the short region may come from transcripts using either the distal or the proximal site. The ratio R_{ij} is the proportion of reads falling in the long region and is thus strictly contained in the interval [0,1]. We note that the extreme value of zero is only encountered in the complete absence of a long isoform, whereas values greater than 0.5 would be observed only in cases where the long region is longer than the short region, or where strong 3' biases in the read coverage are observed.
Our initial tests modeling APA events using logistic regression with quasi-binomial error distribution (within the Generalised Linear Model framework) showed that this approach was not sensitive enough for small numbers of samples or small fold changes. To allow more flexibility in modeling errors, we adopted instead a model where the response variable is assumed to be beta-distributed. This beta-regression model was implemented using the *betareg* package in *R* (Cribari-Neto and Zeileis, 2010). In addition, the quasi-binomial GLM is implemented in our software and used for transcripts where the number of reads falling in the long region is zero, as the ratio in these cases falls outside the permitted values for modelling with beta regression.
Finally, our method incorporates two filtering steps to improve accuracy, employing TIN (=Transcript Integrity Number) values (Wang *et al.*, 2012, 2016) to filter on transcript integrity and removing transcripts with too few reads mapping to the short region (see Supplemental Methods for details).

3 Results

We compared flexiMAP to three existing methods for APA analysis (DaPars (Xia *et al.*, 2014), Roar (Grassi *et al.*, 2016) and APAtrap (Ye *et al.*, 2018)) using simulated data we produced with the *polyester* R package (Frazee *et al.*, 2015) (see Supplementary Methods for details). In these tests, our method is specific (none of the transcripts with no APA events are predicted as having such events) and outperforms in sensitivity DaPars and APAtrap up to a fold change of 4 (Fig. 1A, Supp. Fig. 2). For larger fold changes, all methods appear to perform equally well. Surprisingly, the application of post-detection filters recommended by the developers of both DaPars and APAtrap appear to remove the majority of significant events across all fold changes, which renders questionable the usefulness of these filters (Supp. Fig. 2). In these simulations, Roar is more sensitive than flexiMAP at small fold changes but it is also the least specific, having the largest number of false positives of all methods compared. We note that the performance of Roar is dependent on the parameter value that controls the filtering of significant events (nUnderCutoff; set here to 50%) and that the specificity of the method can be improved by increasing this parameter, albeit at a great cost in sensitivity at low fold changes (Supp. Fig. 2).

All methods, including flexiMAP, were sensitive to the expression level of the transcript tested for differential polyadenylation (Supp. Fig. 3). APA events that were missed originated in transcripts of lower overall expression but the beta-regression approach displayed improved sensitivity over all of the other methods, except Roar. Unlike methods that average across samples from the same condition, the performance of flexiMAP depends on the number of samples available in each group, as expected for a method that needs to model the variance within each group (see Supp. Fig. 4). However, flexiMAP is much more sensitive than the GLM-quasi-binomial method at small sample sizes (<6), often encountered in RNA-seq datasets. Finally, flexiMAP's sensitivity does not seem to be affected by the length of the 3' UTR (Supp. Fig. 5).

Although simulated datasets are important for benchmarking tests, eventually methods are only useful if they can be applied to real data. The dataset we used here is the same used by both DaPars and APAtrap in their respective publications and contains RNA-seq data from the Human Brain Reference and the Universal Human Reference MAQC samples (Bullard *et al.*, 2010). 3' sequencing data (PolyA-seq) for the same samples was downloaded from the UCSC genome browser (processed with an independent method, DEXSeq (Anders *et al.*, 2012), to call the "true" differential polyadenylation events, as described in Supplementary Methods). The results of applying all methods to this

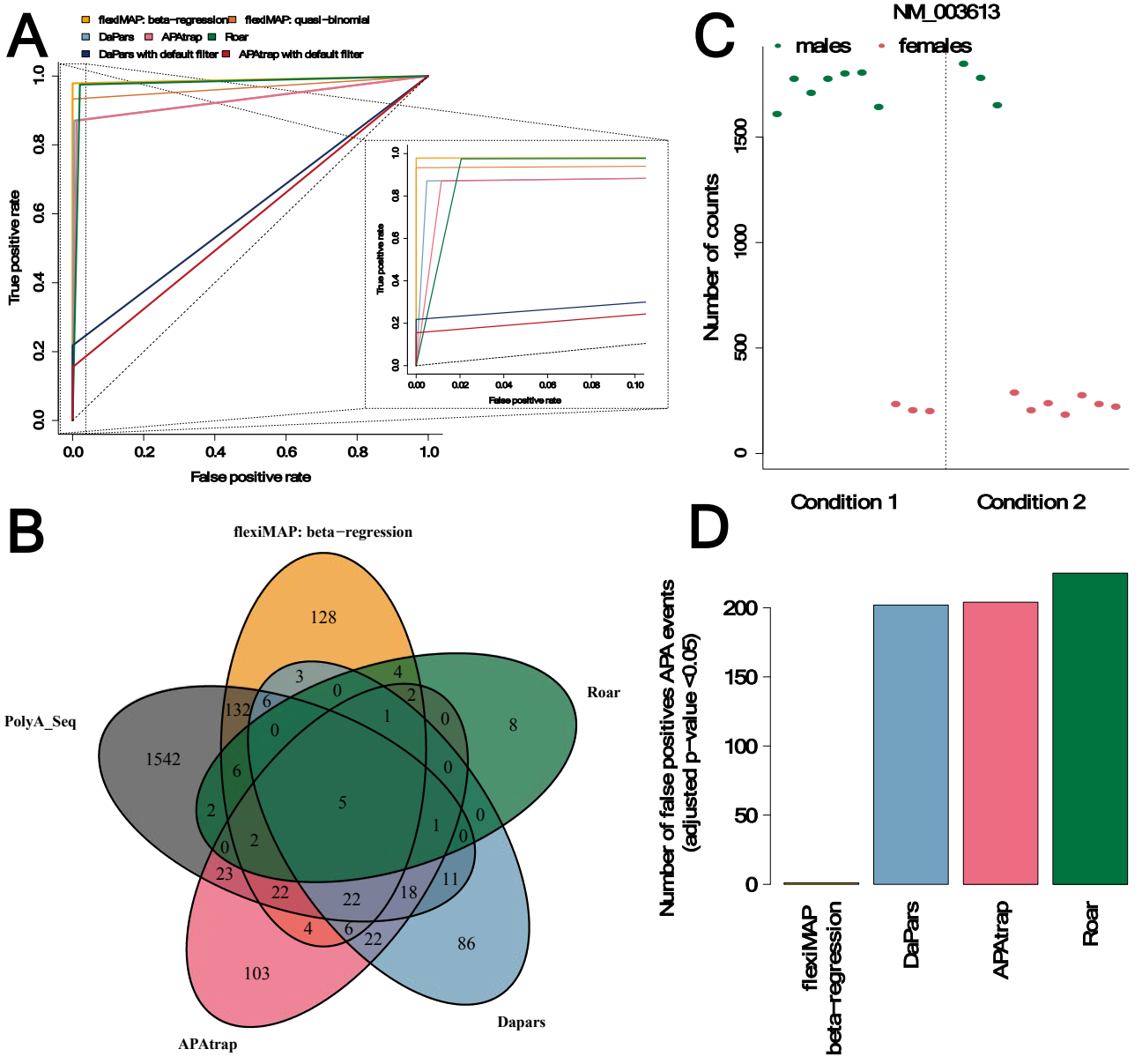
dataset (Fig. 1B) demonstrate that all four miss a large number of events called by DEXseq but flexiMAP is the most sensitive method as well as the one with the highest Matthews Correlation Coefficient (MCC; 0.27 for flexiMAP as compared with 0.23 (Roar), 0.15 (APAtrap) and 0.1 (DaPars)). FlexiMAP's specificity is lower in this dataset compared with other methods but remains over 0.9. Given these results, we believe that although filters or more conservative cut-offs for significance could reduce the number of false positive events called by flexiMAP, they may only be useful in practice when very high specificity is required.

anced datasets, where APA originates from the sex attribute of the samples rather than the condition of interest (Fig. 1C-D). Similar results are obtained with a more complex simulated dataset with two covariates (see description in Supplementary Methods and results in Supp. Fig. 6). Clearly, this is still an artificially simple scenario and one would expect more false positives in real data where at least some of the batch effects might be unknown and hence not included in the modelling. In addition, many real RNA-seq datasets still do not have enough samples to allow successful modelling of multiple covariates so flexiMAP's accuracy as measured in these simulations is likely to be lower with real data. However, it is clear that methods that are not designed to take into account multiple covariates will naturally misinterpret the origin of the variation, resulting in increased false positive rates.

The development of flexiMAP was primarily driven by the need to model multiple known covariates in APA analysis. Indeed, flexiMAP successfully discriminates between the effect of the condition of interest and that of an additional covariate in a simple simulated scenario of imbal

4 Conclusion

We presented here flexiMAP, a beta-regression-based method for detecting alternative polyadenylation events in RNA-seq data, given a list of putative polyadenylation sites. Our method is both sensitive and specific, even when small numbers of samples are used, and has the distinct advantage of being able to model contributions from known covariates that would otherwise confound the results of APA analysis.



FlexiMAP compares favourably with existing alternatives in tests involving simulated datasets. Importantly, these tests have highlighted some hitherto overlooked caveats of existing methods. Real datasets remain a challenge for all methods, not least because it is difficult to define objectively the ground truth, but flexiMAP is still outperforming other methods, when both specificity and sensitivity are taken into account using the Matthews Correlation Coefficient.

Fig. 1. flexiMAP detects differential polyadenylation events with a good balance of specificity and sensitivity. a) Receiver operating characteristic (ROC) curves representing the accuracy of detecting differential alternative polyadenylation events using flexiMAP, DaPars, APAtrap and Roar. DaPars and APAtrap make their own prediction of polyadenylation sites, not always agreeing with the annotated sites used in this study. To avoid inflating the error rate of these programs by including sites that do not map the annotation (and hence, differential events called at these sites would be automatically considered as false positives), only transcripts where the polyadenylation site was correctly predicted by DaPars and APAtrap are included in this plot. FlexiMAP clearly outperformed DaPars, APAtrap and Roar by perfect specificity and improved sensitivity in this simulated experiment. Although application of the DaPars' PDUI (= Percentage of Distal polyA site Usage Index) post-hoc filter (dark blue) and APAtrap's PD (= Percentage Difference) filter (dark red) corrected the false positives problem of these methods, they did so at a heavy cost on sensitivity. b) Venn diagram showing the overlap of "true" differential polyadenylation events in the MAQC samples PolyA-seq data (as called by DEXSeq; grey) with predictions from all four methods tested here: flexiMAP (orange), DaPars (light blue), APAtrap (pink) and Roar (green). c) Example from the imbalanced simulated dataset of a situation where a covariate of no interest (in this case, sex) affects the ratio of reads assigned to short and long isoforms. Male samples display much higher expression of the short region of transcript NM_003613 compared with female ones, regardless of the condition group samples belong to. In addition, the dataset is imbalanced, with more males present in condition 1 than condition 2. The mean expression for condition 1 is thus higher than the mean for condition 2, but the effect is due to the covariate sex, not the condition to which the samples belong to. d) DaPars, APAtrap and Roar report a large number of false positives for an imbalanced simulated dataset. In contrast, flexiMAP reports only one false positive in this case, highlighting its main advantage over alternative approaches.

Acknowledgements

The authors gratefully acknowledge help from Dr Elena Grassi and Dr Congting Ye in carrying out the comparison of flexiMAP to Roar and APAtrap respectively.

Funding

This work was supported by grants from the Birkbeck / Wellcome Trust Institutional Strategic Support Fund awarded to KJS and IN.

Conflict of Interest: none declared.

References

Anders,S. *et al.* (2012) Detecting differential usage of exons from RNA-seq data. *Genome Res.*, **22**, 2008–2017.

Anvar,S.Y. *et al.* (2018) Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biol.*, **19**, 46.

Arefeen,A. *et al.* (2018) TAPAS: tool for alternative polyadenylation site analysis. *Bioinformatics (Oxford, England)*, **34**, 2521–2529.

Bullard,J.H. *et al.* (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.

Cribari-Neto,F. and Zeileis,A. (2010) Beta Regression in R. *Journal of Statistical Software*, **34**, 1–24.

Elkon,R. *et al.* (2013) Alternative cleavage and polyadenylation: extent, regulation and function. *Nature reviews. Genetics*, **14**, 496–506.

Frazee,A.C. *et al.* (2015) Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, **31**, 2778–2784.

Garalde,D.R. *et al.* (2016) Highly parallel direct RNA sequencing on an array of nanopores. *bioRxiv*, 068809.

Grassi,E. *et al.* (2016) Roar: detecting alternative polyadenylation with standard mRNA sequencing libraries. *BMC bioinformatics*, **17**, 423.

Ha,K.C.H. *et al.* (2018) QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome biology*, **19**, 45.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**, 550.

Moll,P. *et al.* (2014) QuantSeq 3′ mRNA sequencing for RNA quantification. *Nature Methods*, **11**, i–iii.

Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, **26**, 139–140.

Szkop,K.J. *et al.* (2017) Dysregulation of Alternative Poly-adenylation as a Potential Player in Autism Spectrum Disorder. *Front. Mol. Neurosci.*, **10**.

Szkop,K.J. and Nobeli,I. (2017) Untranslated Parts of Genes Interpreted: Making Heads or Tails of High-Throughput Transcriptomic Data via Computational Methods: Computational methods to discover and quantify isoforms with alternative untranslated regions. *BioEssays : news and reviews in molecular, cellular and developmental biology*, **39**, 1700090.

Wang,L. *et al.* (2012) RSeQC: quality control of RNA-seq experiments. *Bioinformatics (Oxford, England)*, **28**, 2184–2185.

Wang,Liguo *et al.* (2016) Measure transcript integrity using RNA-seq data. *BMC bioinformatics*, **17**, 58.

Xia,Z. *et al.* (2014) Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nature communications*, **5**, 5274.

Ye,C. *et al.* (2018) APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. *Bioinformatics (Oxford, England)*, **34**, 1841–1849.