

satuRn: Scalable analysis of differential transcript usage for bulk and single-cell RNA-sequencing applications

This manuscript ([permalink](#)) was automatically generated from [jgilis/satuRn_md@c9d4d76](#) on August 30, 2022.

Authors

- **Jeroen Gilis**

 [0000-0001-8415-0943](#) ·  [jgilis](#) ·  [GilisJeroen](#)

Department of Applied Mathematics, Computer science and Statistics, Ghent University · Funded by FWO

- **Kristoffer Vitting-Seerup**

 [XXXX-XXXX-XXXX-XXXX](#) ·  [kvingtingseerup](#)

Department of Something, University of Whatever; Department of Whatever, University of Something

- **Koen Van den Berge**

 [XXXX-XXXX-XXXX-XXXX](#) ·  [koenvandenberge](#)

Department of Something, University of Whatever; Department of Whatever, University of Something

- **Lieven Clement**

 [XXXX-XXXX-XXXX-XXXX](#) ·  [statOmics](#)

Department of Something, University of Whatever; Department of Whatever, University of Something

✉ — Correspondence possible via [GitHub Issues](#)

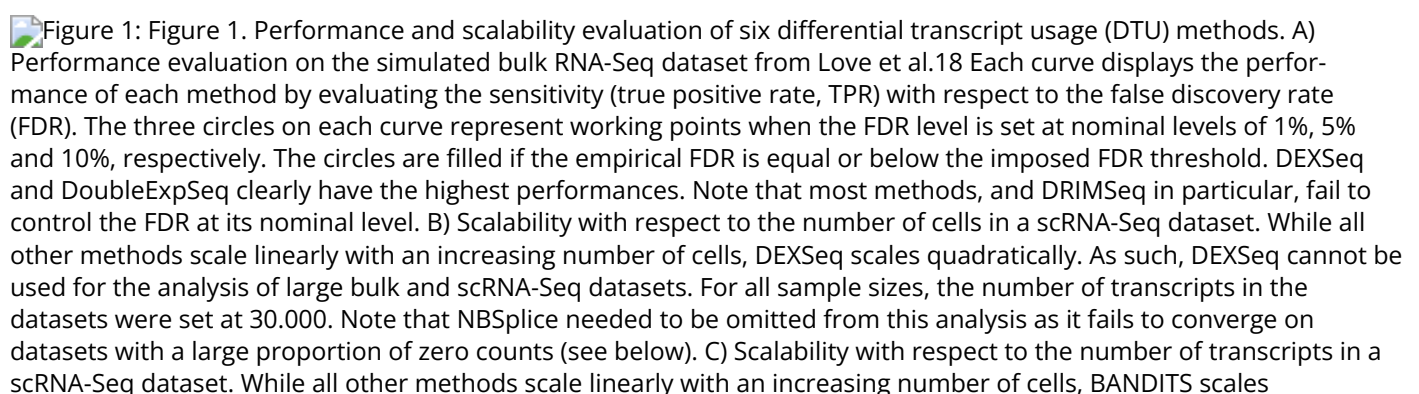
Abstract

Alternative splicing produces multiple functional transcripts from a single gene. Dysregulation of splicing is known to be associated with disease and as a hallmark of cancer. Existing tools for differential transcript usage (DTU) analysis either lack in performance, cannot account for complex experimental designs or do not scale to massive single-cell transcriptome sequencing (scRNA-seq) datasets. We introduce satuRn, a fast and flexible quasi-binomial generalized linear modelling framework that is on par with the best performing DTU methods from the bulk RNA-seq realm, while providing good false discovery rate control, addressing complex experimental designs, and scaling to scRNA-seq applications.

Introduction

Studying differential expression (DE) is one of the key tasks in the downstream analysis of RNA-seq data. Typically, DE analyses identify expression changes on the gene level. However, the widespread adoption of expression quantification through pseudo-alignment^[1,2], which enables fast and accurate quantification of expression at the transcript level, has effectively paved the way for transcript-level analyses. Here, we specifically address differential transcript usage (DTU) analysis, one type of transcript-level analysis that studies the change in relative usage of transcripts/isoforms within the same gene. DTU analysis holds great potential: previous research has shown that most multi-exon human genes are subject to alternative splicing and can thus produce a variety of functionally different isoforms from the same genomic locus.^{3–5} The dysregulation of this splicing process has been reported extensively as a cause for disease,^{6–9} including several neurological diseases such as frontotemporal dementia, Parkinsonism and spinal muscular atrophy, and is a well-known hallmark of cancer.¹⁰

In this context, full-length single-cell RNA-Seq (scRNA-seq) technologies such as Smart-Seq²¹¹ and Smart-Seq³¹² hold the promise to further increase the resolution of DTU analysis from bulk RNA-seq data towards the single-cell level, where differences in transcript usage are expected to occur naturally between cell types. However, only a few bespoke DTU methods have been developed for scRNA-seq data and they lack biological interpretation. Indeed, methods specifically developed for scRNA-seq data are either restricted to exon/event level^{13,14} analysis (e.g. pinpointing exons involved in splicing events), or they can only pinpoint DTU genes without unveiling the actual transcripts that are involved.¹⁵ Interestingly, many DTU methods for bulk RNA-seq do provide inference at the transcript level and their performance has already been extensively profiled in benchmark studies.^{16–18} Based on a subset of the simulated RNA-seq dataset from Love et al.¹⁸ (see Methods), we show the performance of six DTU tools; DEXSeq,¹⁹ DoubleExpSeq,²⁰ DRIMSeq,²¹ edgeR diffSplice,²² limma diffSplice²³ and NBSplice²⁴ (Figure 1). DEXSeq and DoubleExpSeq have a higher performance than the other methods. In addition, we observe that most methods, and DRIMSeq in particular, fail to control the false discovery rate (FDR) at its nominal level, which is in line with previous reports.^{16–18}

Figure 1: Figure 1. Performance and scalability evaluation of six differential transcript usage (DTU) methods. A) Performance evaluation on the simulated bulk RNA-Seq dataset from Love et al.¹⁸ Each curve displays the performance of each method by evaluating the sensitivity (true positive rate, TPR) with respect to the false discovery rate (FDR). The three circles on each curve represent working points when the FDR level is set at nominal levels of 1%, 5% and 10%, respectively. The circles are filled if the empirical FDR is equal or below the imposed FDR threshold. DEXSeq and DoubleExpSeq clearly have the highest performances. Note that most methods, and DRIMSeq in particular, fail to control the FDR at its nominal level. B) Scalability with respect to the number of cells in a scRNA-Seq dataset. While all other methods scale linearly with an increasing number of cells, DEXSeq scales quadratically. As such, DEXSeq cannot be used for the analysis of large bulk and scRNA-Seq datasets. For all sample sizes, the number of transcripts in the datasets were set at 30,000. Note that NBSplice needed to be omitted from this analysis as it fails to converge on datasets with a large proportion of zero counts (see below). C) Scalability with respect to the number of transcripts in a scRNA-Seq dataset. While all other methods scale linearly with an increasing number of cells, BANDITS scales

quadratically. Moreover, BANDITS failed to run on our system for datasets with 7,500 transcripts or more. As such, it had to be omitted from panels A and B. A performance and scalability evaluation of BANDITS on datasets with an (artificial) lower number of transcripts is provided as Extended data figures S1 and S3.25

Figure 1: Figure 1. Performance and scalability evaluation of six differential transcript usage (DTU) methods. A) Performance evaluation on the simulated bulk RNA-Seq dataset from Love et al.¹⁸ Each curve displays the performance of each method by evaluating the sensitivity (true positive rate, TPR) with respect to the false discovery rate (FDR). The three circles on each curve represent working points when the FDR level is set at nominal levels of 1%, 5% and 10%, respectively. The circles are filled if the empirical FDR is equal or below the imposed FDR threshold. DEXSeq and DoubleExpSeq clearly have the highest performances. Note that most methods, and DRIMSeq in particular, fail to control the FDR at its nominal level. B) Scalability with respect to the number of cells in a scRNA-Seq dataset. While all other methods scale linearly with an increasing number of cells, DEXSeq scales quadratically. As such, DEXSeq cannot be used for the analysis of large bulk and scRNA-Seq datasets. For all sample sizes, the number of transcripts in the datasets were set at 30,000. Note that NBSplice needed to be omitted from this analysis as it fails to converge on datasets with a large proportion of zero counts (see below). C) Scalability with respect to the number of transcripts in a scRNA-Seq dataset. While all other methods scale linearly with an increasing number of cells, BANDITS scales quadratically. Moreover, BANDITS failed to run on our system for datasets with 7,500 transcripts or more. As such, it had to be omitted from panels A and B. A performance and scalability evaluation of BANDITS on datasets with an (artificial) lower number of transcripts is provided as Extended data figures S1 and S3.25

References

1. **Near-optimal probabilistic RNA-seq quantification**
Nicolas L Bray, Harold Pimentel, Páll Melsted, Lior Pachter
Nature Biotechnology (2016-04-04) <https://doi.org/f8nvsp>
DOI: [10.1038/nbt.3519](https://doi.org/10.1038/nbt.3519) · PMID: [27043002](https://pubmed.ncbi.nlm.nih.gov/27043002/)
2. **Salmon provides fast and bias-aware quantification of transcript expression**
Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, Carl Kingsford
Nature Methods (2017-03-06) <https://doi.org/gcw9f5>
DOI: [10.1038/nmeth.4197](https://doi.org/10.1038/nmeth.4197) · PMID: [28263959](https://pubmed.ncbi.nlm.nih.gov/28263959/) · PMCID: [PMC5600148](https://pubmed.ncbi.nlm.nih.gov/PMC5600148/)