satuRn:
Scalable Analysis of differential Transcript Usage for bulk and single-Cell RNA-sequencing applications

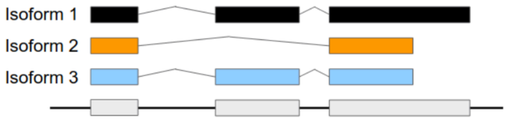
Jeroen Gilis[1,2,5], Kristoffer Vitting-Seerup[3], Koen Van den Berge[1,4,5] and Lieven Clement[1,5]

1. Department of Applied Mathematics, Computer Science & Statistics, Ghent University, Belgium.
2. Data Mining and Modeling for Biomedicine, VIB Center for Inflammation Research, Ghent, Belgium.
3. Department of Biology, The Bioinformatics Centre, University of Copenhagen
4. Department of Statistics, University of California, Berkeley, CA, USA
5. Bioinformatics Institute Ghent, Ghent University, Ghent, Belgium.

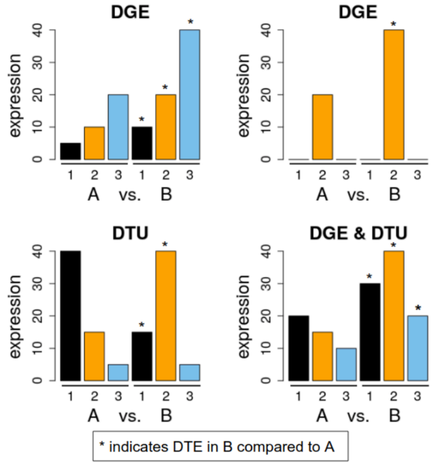
Presenter: Jeroen Gilis
Email: jeroen.gilis@ugent.be
GitHub: https://github.com/jgilis
Twitter: https://twitter.com/GilisJeroen

Background and objectives

Differential transcript usage (DTU) analyses study the change in usage of the different isoforms within a gene between conditions of interest.



Reference: Van den Berge et al. (2019), Annual Reviews of Biomedical Data Science.



Objectives: to develop a method for DTU analysis that

1. Is highly performant
2. Scales to large datasets
3. Provides a strict control of the false discovery rate (FDR)
4. Allows for modelling multi-factor experimental designs
5. Handles real-life proportions of zero counts in single-cell data

Methods

- Denote the expression of transcript *t* of gene *g* in sample *i* as Y_{gti}
- The total expression of gene *g* in sample *i* can then be expressed as:

$$Y_{g.i} = \sum_{t \in \tau_g} Y_{gti} \quad (1)$$

- Denote the usage of transcript *t* of gene *g* in sample *i* as:

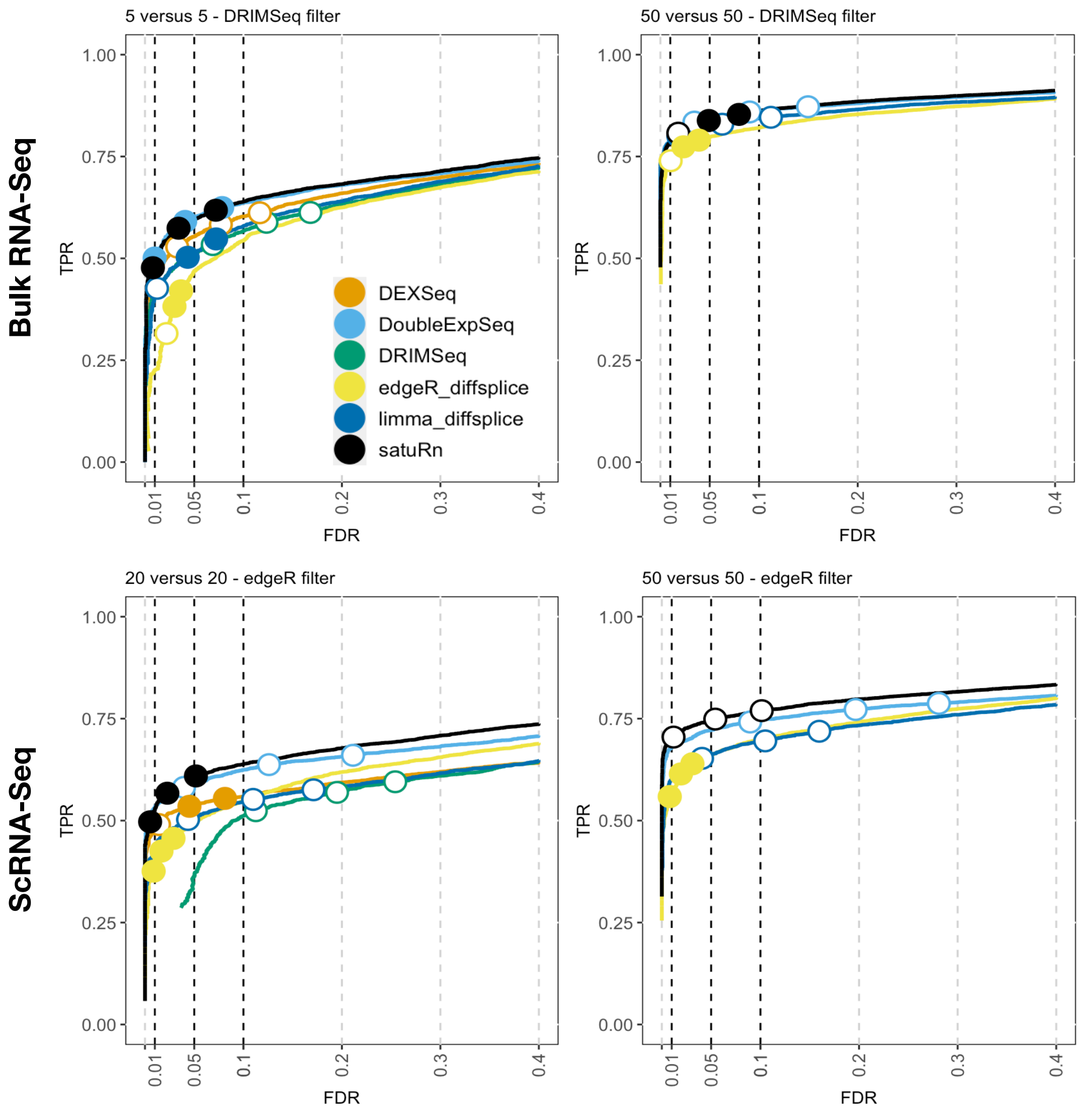
$$U_{gti} = \frac{Y_{gti}}{Y_{g.i}} \quad (2)$$

- Describe the quasi-binomial generalised linear model as:

$$\begin{cases} E\left[U_{gti} \middle| \boldsymbol{X_i}, Y_{g.i}\right] = \pi_{gti} \\ \log\left(\frac{\pi_{gti}}{1-\pi_{gti}}\right) = \eta_{gti} \\ \eta_{gti} = \boldsymbol{X_i^T \beta_{gt}} \end{cases}$$

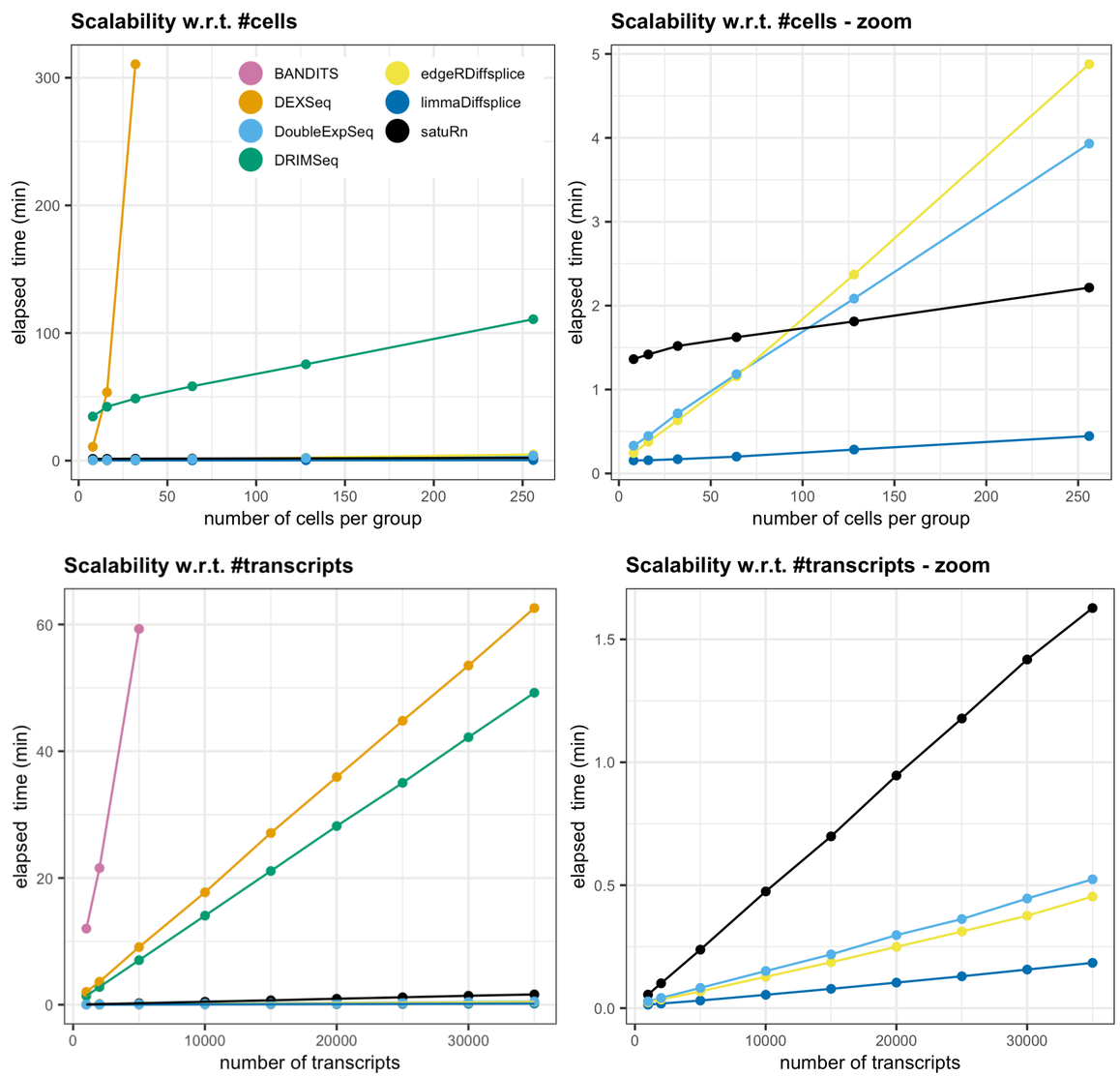
- With variance: $Var\left[U_{gti} \middle| \boldsymbol{X_i}, Y_{g.i}\right] = \frac{\pi_{gti} * (1-\pi_{gti})}{Y_{g.i}} * \phi_{gt}$

Performance



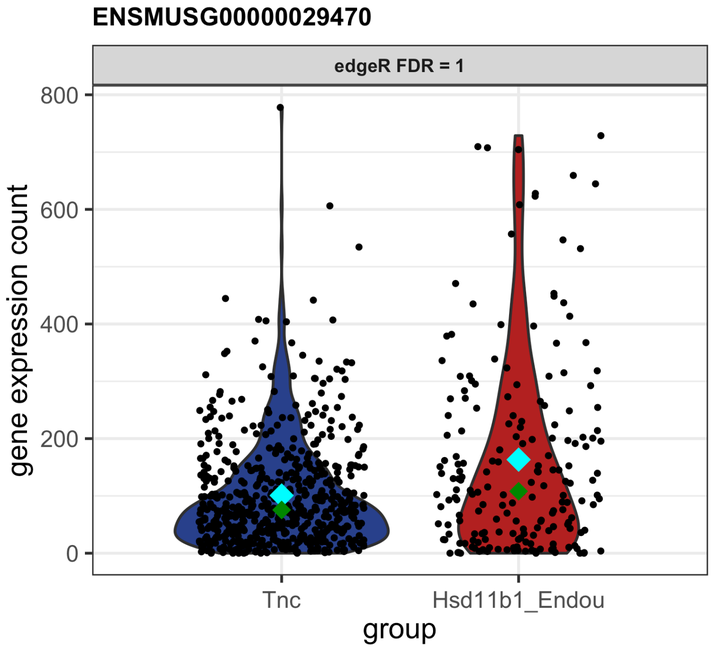
***satuRn* displays an excellent performance,** both in bulk and scRNA-Seq datasets. The high performance is achieved over a large range of sample sizes and in two distinct filtering criteria. In addition, *satuRn* provides an accurate control of the FDR, even in large sample sizes.

Scalability



***satuRn* scales to large scRNA-Seq datasets.** The top panels demonstrate the lineair scalability profile of satuRn with respect to the number of cells or samples in the dataset. The bottom panels show the scalability profile with respect to the number of transcripts in the dataset.

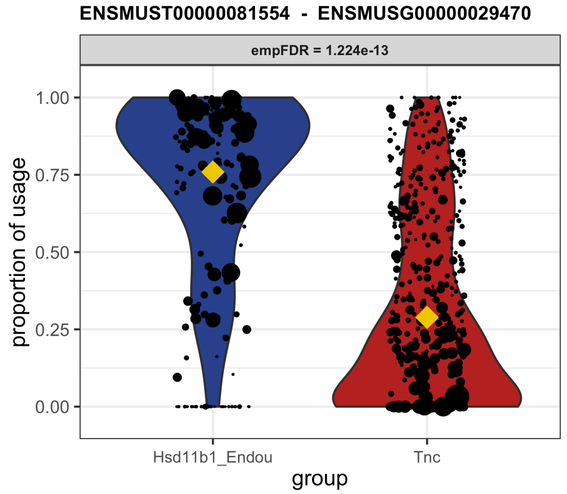
Case study

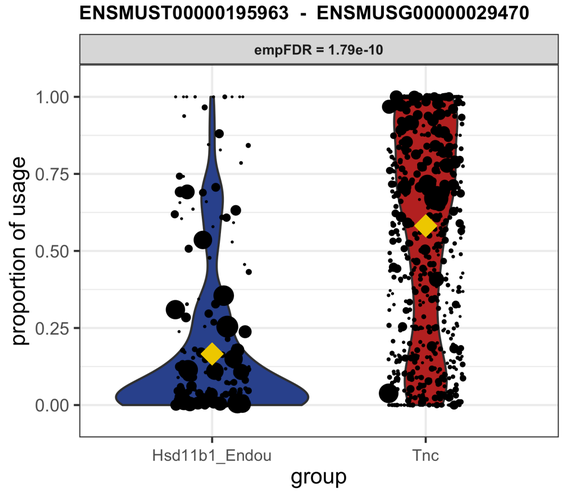


In a canonical edgeR analysis, **no evidence for differential gene expression** of P2rx4 was found between both mouse neocortex cell types

Dataset obtained from Tasic et al. (2018), Nature 563, 72–78

***satuRn* identifies biologically relevant differences in transcript usage between cell types**





- In Hsd11b1-Endou cells, the isoform in the top panel is the dominant isoform of the P2rx4 gene (estimated usage of 76%)
- However, the isoform at the bottom is dominant in Tnc cells (estimated usage = 58%)
- Crucially, the isoform at the top is protein coding, while the isoform at the bottom is not

satuRn:
Scalable Analysis of differential Transcript Usage for bulk and single-Cell RNA-sequencing applications

Jeroen Gilis[1,2,5], Kristoffer Vitting-Seerup[3], Koen Van den Berge[1,4,5] and Lieven Clement[1,5]

1. Department of Applied Mathematics, Computer Science & Statistics, Ghent University, Belgium.
2. Data Mining and Modeling for Biomedicine, VIB Center for Inflammation Research, Ghent, Belgium.
3. Department of Biology, The Bioinformatics Centre, University of Copenhagen
4. Department of Statistics, University of California, Berkeley, CA, USA
5. Bioinformatics Institute Ghent, Ghent University, Ghent, Belgium.

Presenter: Jeroen Gilis
Email: jeroen.gilis@ugent.be
GitHub: https://github.com/jgilis
Twitter: https://twitter.com/GilisJeroen

Take-home messages

- ***satuRn* is a novel tool for DTU analysis that:**
  1. Has a similar performance as the state-of-the-art DTU tools
  2. Scales to large datasets
  3. Provides a strict control of the FDR
  4. Allows for modelling multi-factor experimental designs
  5. Handles real-life proportions of zero counts in single-cell data
- ***satuRn* adopts quasi-binomial GLMs to assess DTU between conditions of interest**
- ***satuRn* detects differences in transcript usage between cell types that show evidence of biological relevance.**
- ***satuRn* will soon be published on bioRxiv and available as an R package from GitHub.**