UNIVERSITEIT GENT

VIB science meets life

fwo Research Foundation Flanders Opening new horizons

Scalable differential transcript usage analysis for single-cell applications

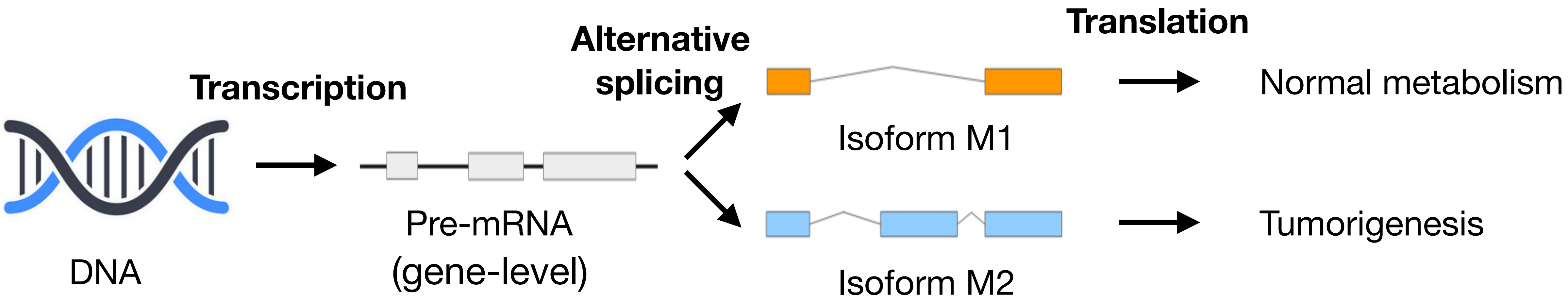
JEROEN GILIS

EuroBioc2019 presentation

Promotor: Prof. Lieven Clement

Supervisor: Dr. Koen Van den Berge

Differential Transcript Usage (DTU)



DNA

Transcription

Pre-mRNA
(gene-level)

Alternative splicing

Isoform M1

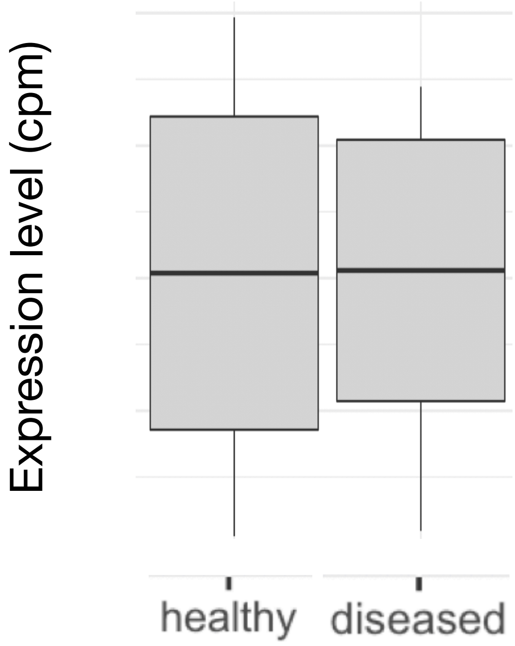
Isoform M2

Translation

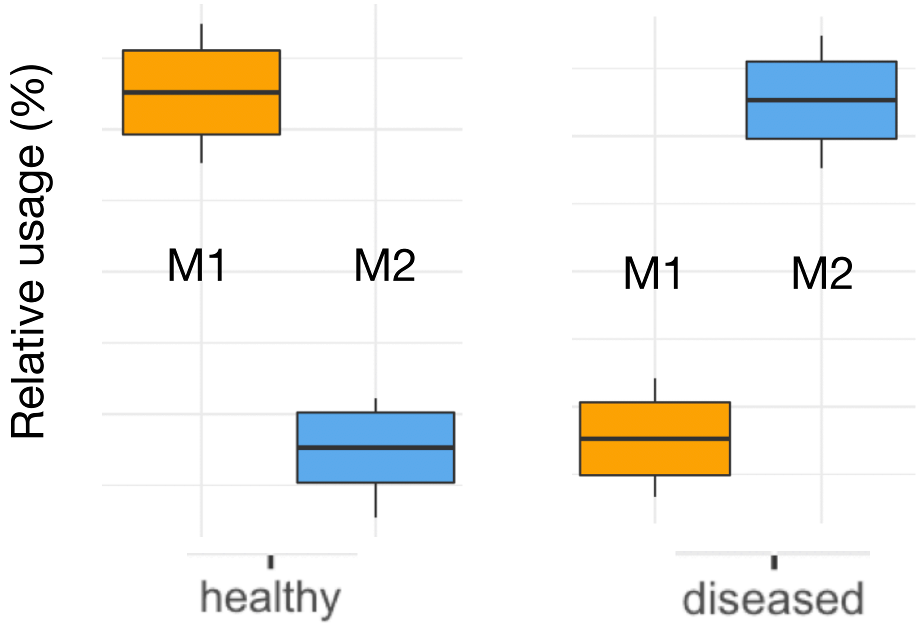
Normal metabolism

Tumorigenesis

Gene-level analysis



Transcript-level analysis



Method development

- Our workflow unlocks edgeR for DTU analysis

$$\textbf{DGE} \begin{cases} \mathrm{Y}_{\mathrm{gi}} \sim NB\,(\mu_{\mathrm{gi}}, \varphi_{\mathrm{g}}) \\ \log\,(\mu_{\mathrm{gi}}) = \eta_{\mathrm{gi}} \\ \eta_{\mathrm{gi}} = \beta_{0} + \beta^{\mathrm{C}}_{\mathrm{gc}} + \log\,(\mathrm{S}_{\mathrm{i}}) \end{cases}$$

Method development

- Our workflow unlocks edgeR for DTU analysis

$$\textbf{DTE} \quad \begin{cases} Y_{ti} \sim NB\,(\mu_{ti}, \varphi_{t}) \\ \log\,(\mu_{ti}) = \eta_{ti} \\ \eta_{ti} = \beta_{0} + \beta_{tc}^{C} + \log\,(S_{i}) \end{cases}$$

Method development

- Our workflow unlocks edgeR for DTU analysis

$$\underline{\textbf{DTU}} \quad \begin{cases} Y_{ti} \sim NB\,(\mu_{ti}, \varphi_{t}) \\ \log\,(\mu_{ti}) = \eta_{ti} \\ \eta_{ti} = \beta_{0} + \beta_{tc}^{C} + \log\,(T_{ti}) \end{cases}$$

- Our workflow takes the **gene-level counts (total counts, T_{ti}) as offsets** to the GLM framework → edgeR-total

Method development

- Our workflow unlocks edgeR for DTU analysis

$$\underline{\textbf{DTU}} \quad \begin{cases} Y_{ti} \sim NB(\mu_{ti}, \varphi_t) \\ \log(\mu_{ti}) = \eta_{ti} \\ \eta_{ti} = \beta_0 + \beta_{tc}^{C} + \log(T_{ti}) \end{cases}$$

- Our workflow takes the **gene-level counts (total counts, T_{ti}) as offsets** to the GLM framework → edgeR-total

- DEXSeq

Counts

	Sample 1	…	Sample m
Tx 1	112	…	15
Tx t	…	…	…
Tx n	62	…	348

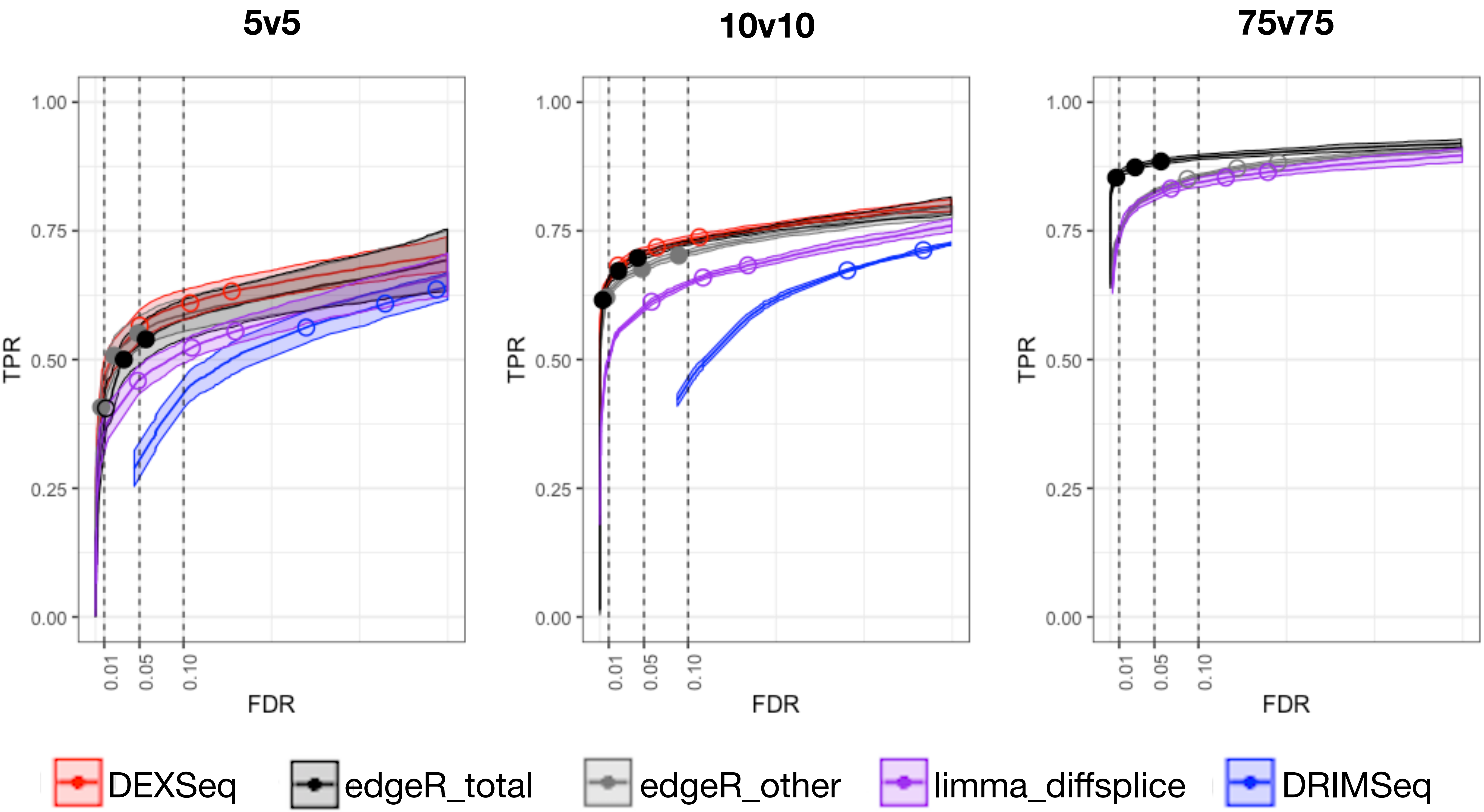
	Sample 1	…	Sample m
Tx 1	25	…	3
Tx t	…	…	…
Tx n	88	…	212

‘other’ counts

- Our second workflow takes the **other counts as offsets** → edgeR-other

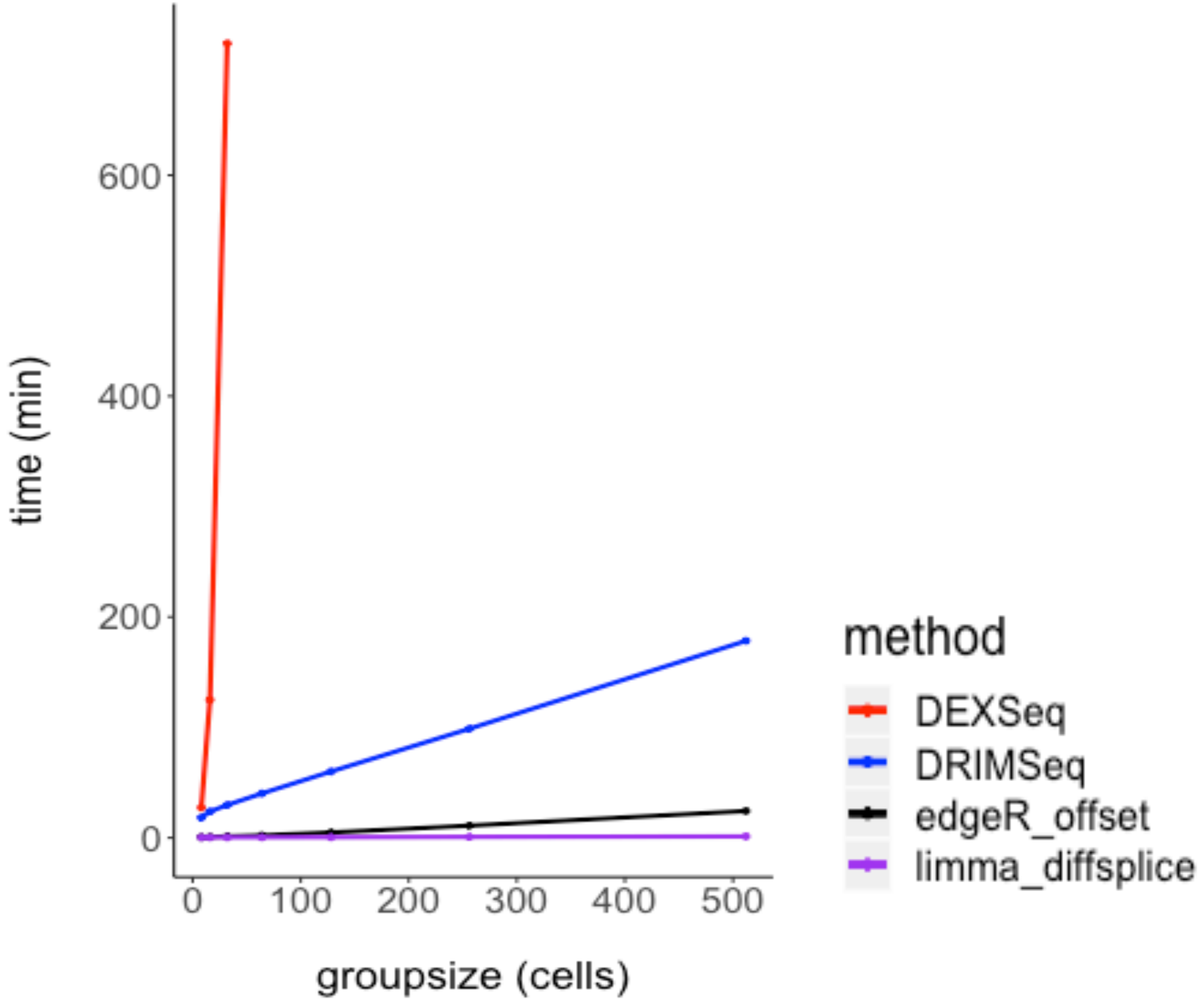
Performance evaluation on real bulk data

Gtex dataset, Nature Genetics 45, 580-585 (2013)



Scalability benchmark on real single-cell data

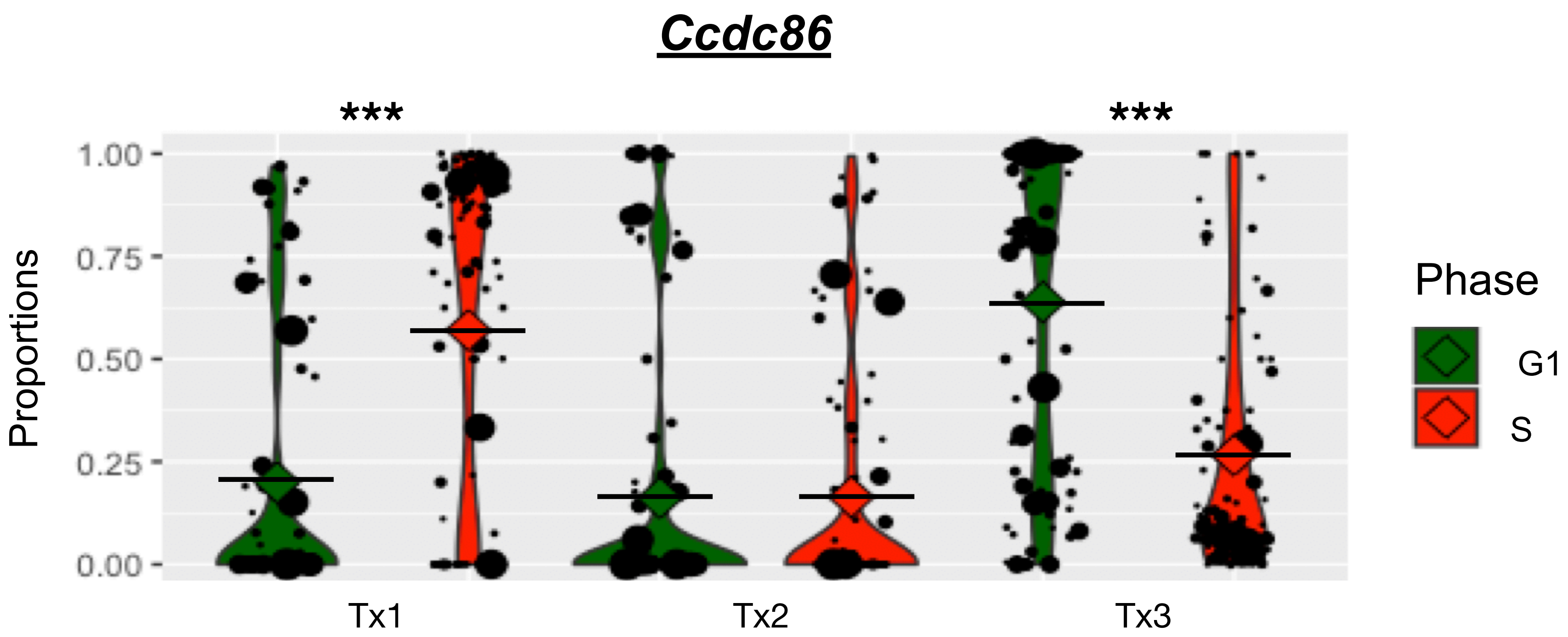
- Our workflow performs a DTU analysis between two groups of 512 cells in ~20 minutes
- DEXSeq scales quadratically



Single-cell transcriptomics case study

Dataset from Buettner et al., Nature Biotechnology 33; 155-160 (2015)

- Dataset; 288 mouse embryonic stem cells, different cell cycle stages (G1, S and G2M)
- Runtime; < 2 minutes
- Significant enrichment in cell cycle processes
- Several DTU genes are;
  - ✦ Biologically relevant
  - ✦ Not picked up in a gene-level analysis
  - ✦ Clearly differentially used when visualised



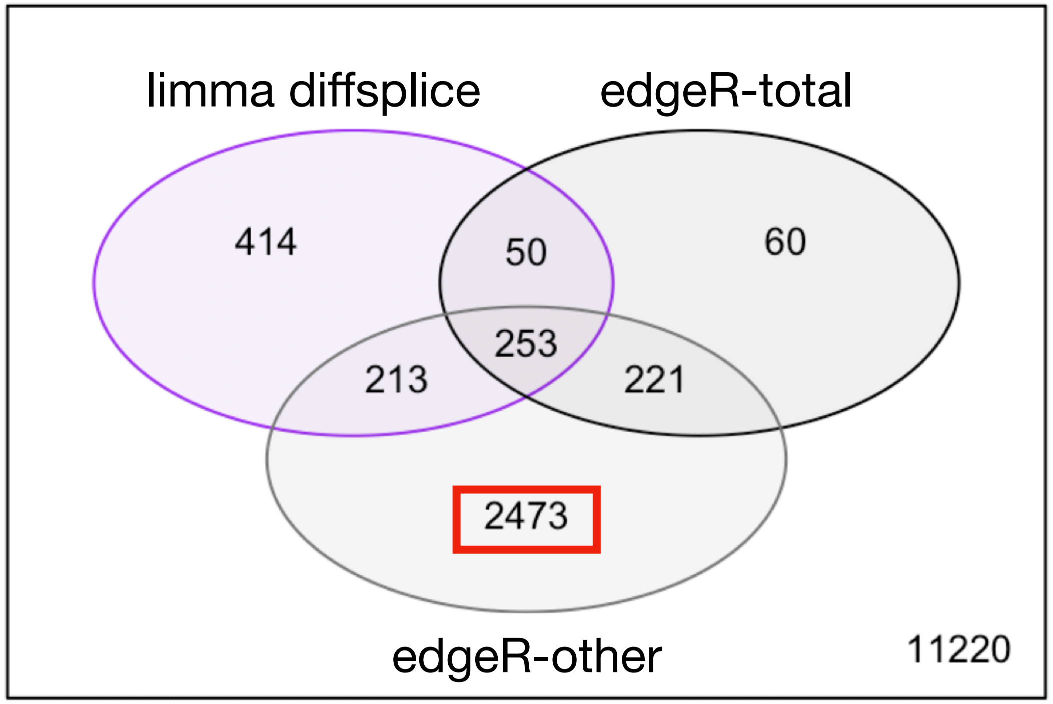
The size of the dots (which represent individual cells) are weighted according to the total expression of the gene in that cell.

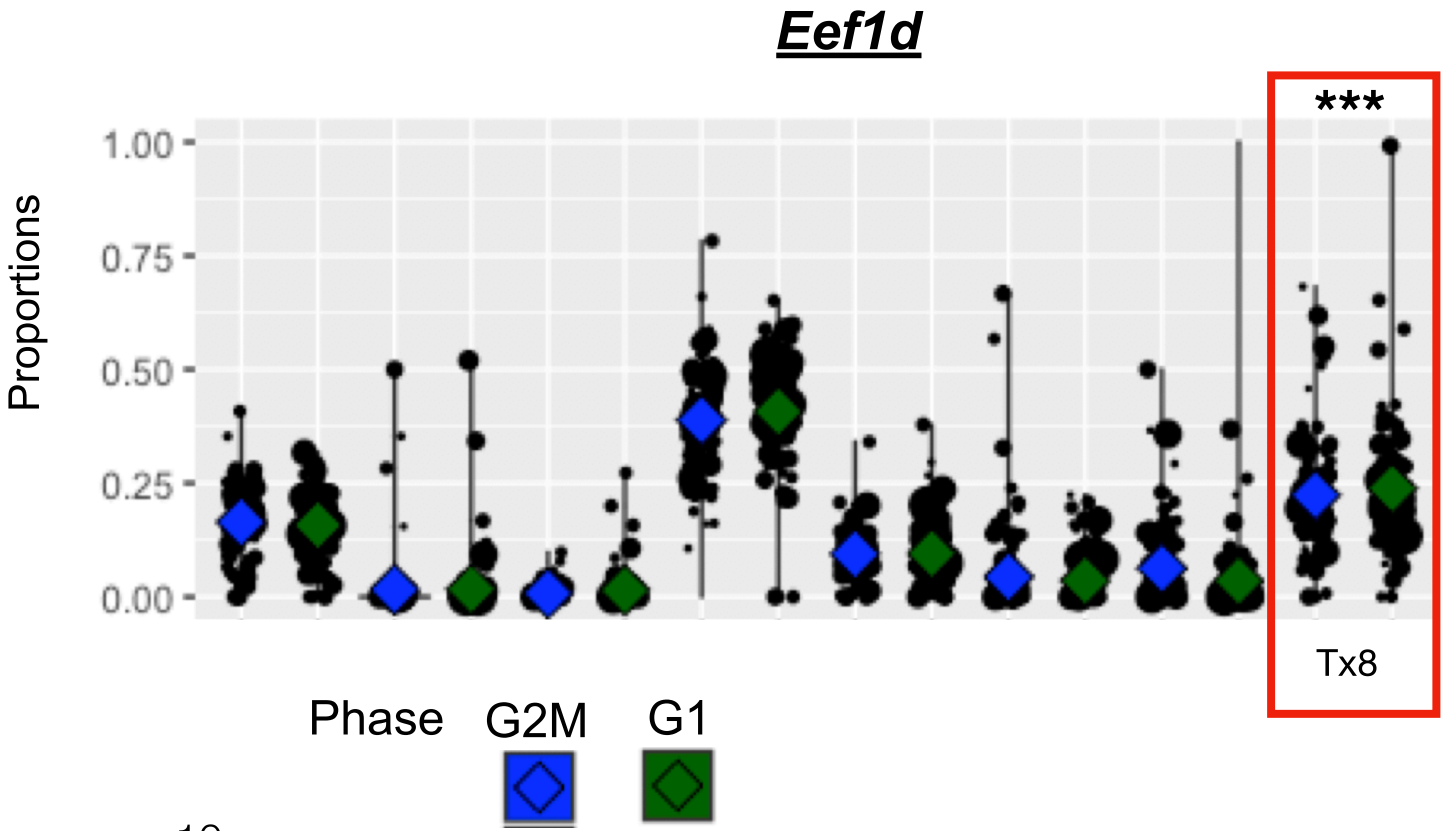
Single-cell transcriptomics case study

Buettner dataset, Nature Biotechnology 33; 155-160 (2015)

- Dataset; 288 mouse embryonic stem cells, different cell cycle stages (G1, S and G2M)
- Runtime; < 2 minutes for offset-based methods
- Significant enrichment in cell cycle processes
- Some DTU genes display clear DTU in visualisation and are biologically relevant

- edgeR_other method large number of (false) positive results; sensitive to outliers (?)
- Discrepancy between edgeR-total and limma diffsplice; asses formally in single-cell benchmark





Take-home messages

We are developing a workflow for studying DTU that;

1. Has a performance similar to that of DEXSeq
2. Correctly controls the false discovery rate
3. Scales towards large transcriptomics datasets

Scalable differential transcript usage analysis for single-cell applications

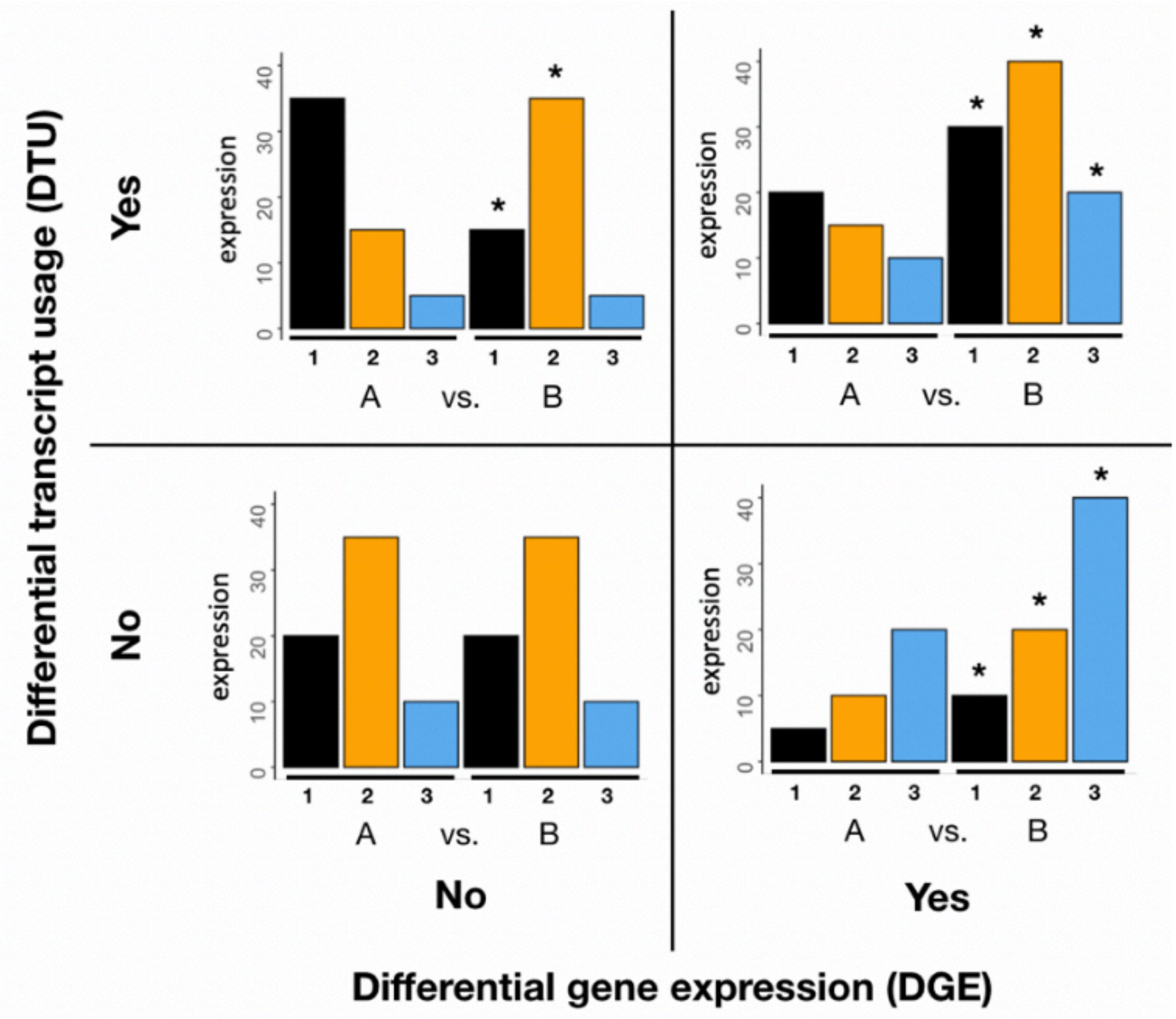
JEROEN GILIS

EuroBioc2019 presentation

Promotor: Prof. Lieven Clement

Supervisor: Dr. Koen Van den Berge

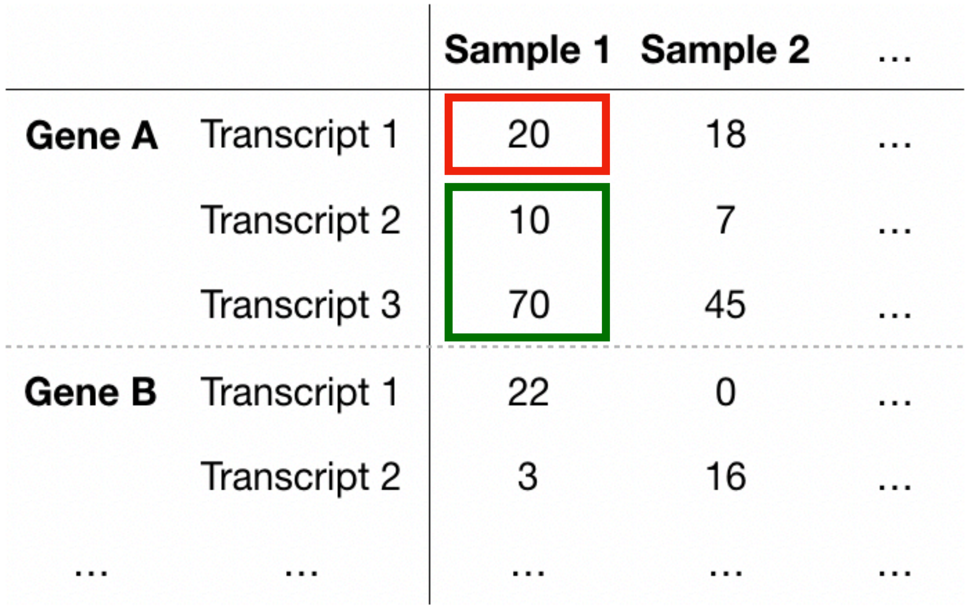
Background - DTU



Background - DEXSeq

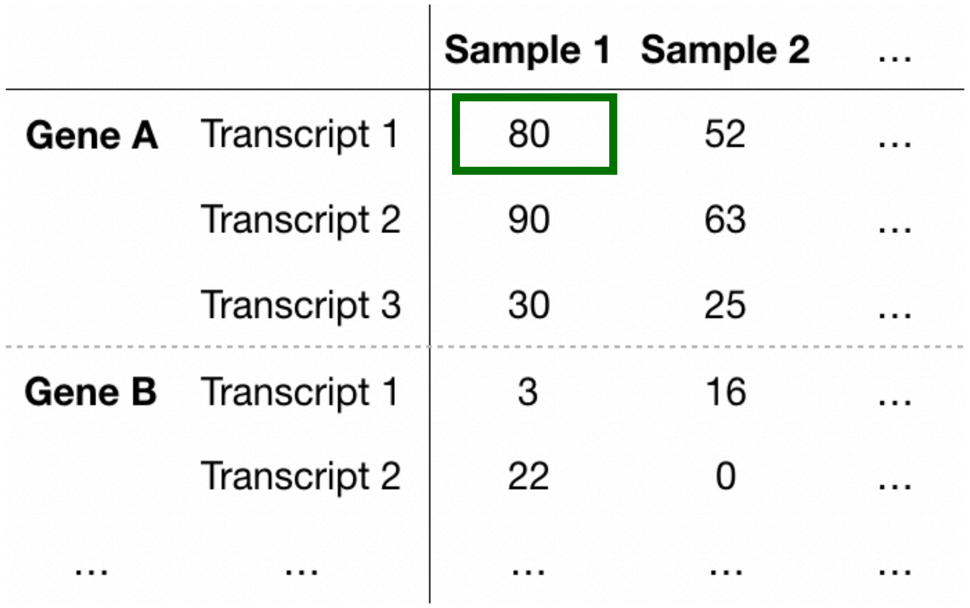
- **Input**: matrix of transcript-level counts (e.g. Salmon or kallisto)

Transcript-level counts



		Sample 1	Sample 2	...
Gene A	Transcript 1	20	18	...
	Transcript 2	10	7	...
	Transcript 3	70	45	...
Gene B	Transcript 1	22	0	...
	Transcript 2	3	16	...
...	...	...	...	...

Complementary counts



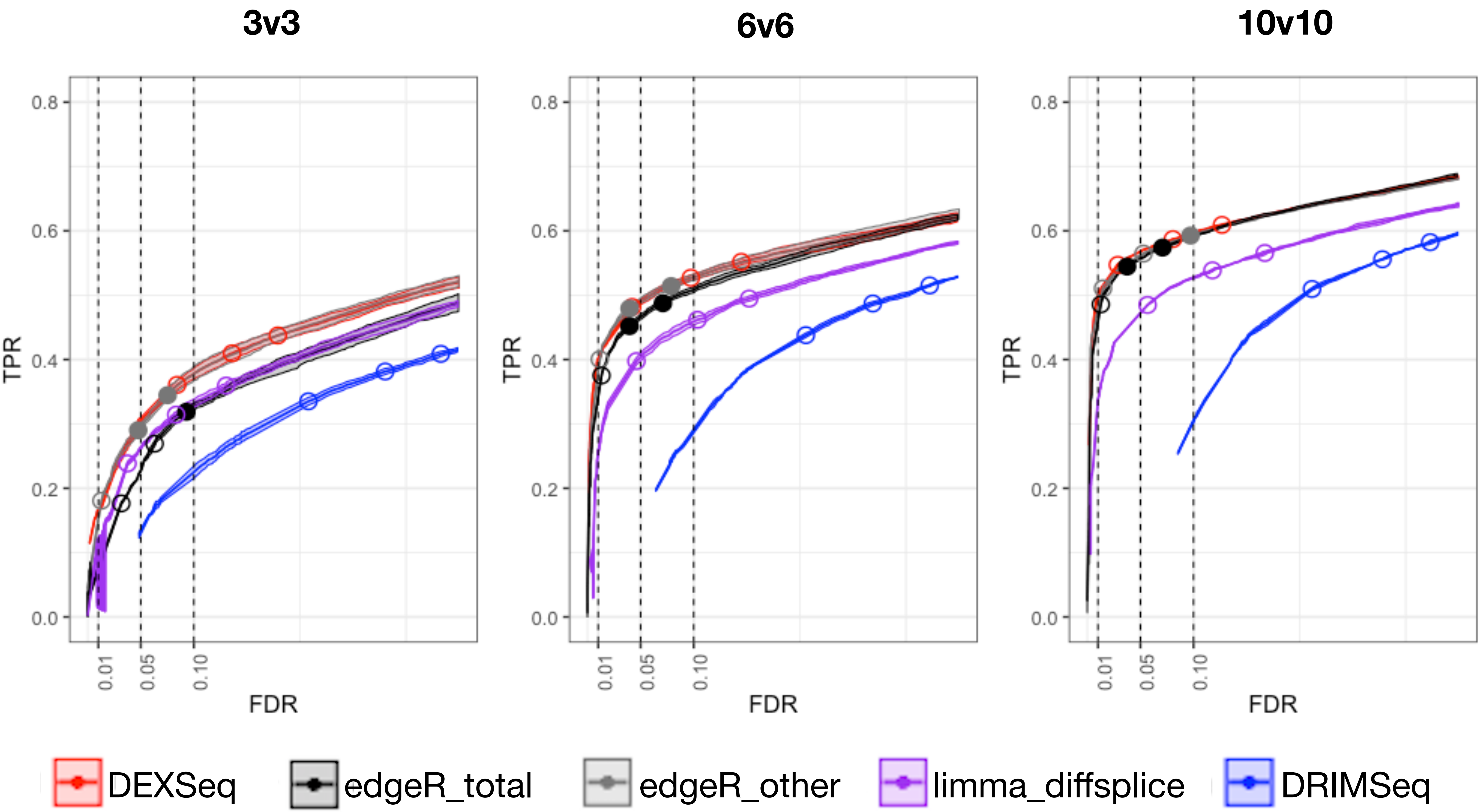
		Sample 1	Sample 2	...
Gene A	Transcript 1	80	52	...
	Transcript 2	90	63	...
	Transcript 3	30	25	...
Gene B	Transcript 1	3	16	...
	Transcript 2	22	0	...
...	...	...	...	...

- **Statistical model:**

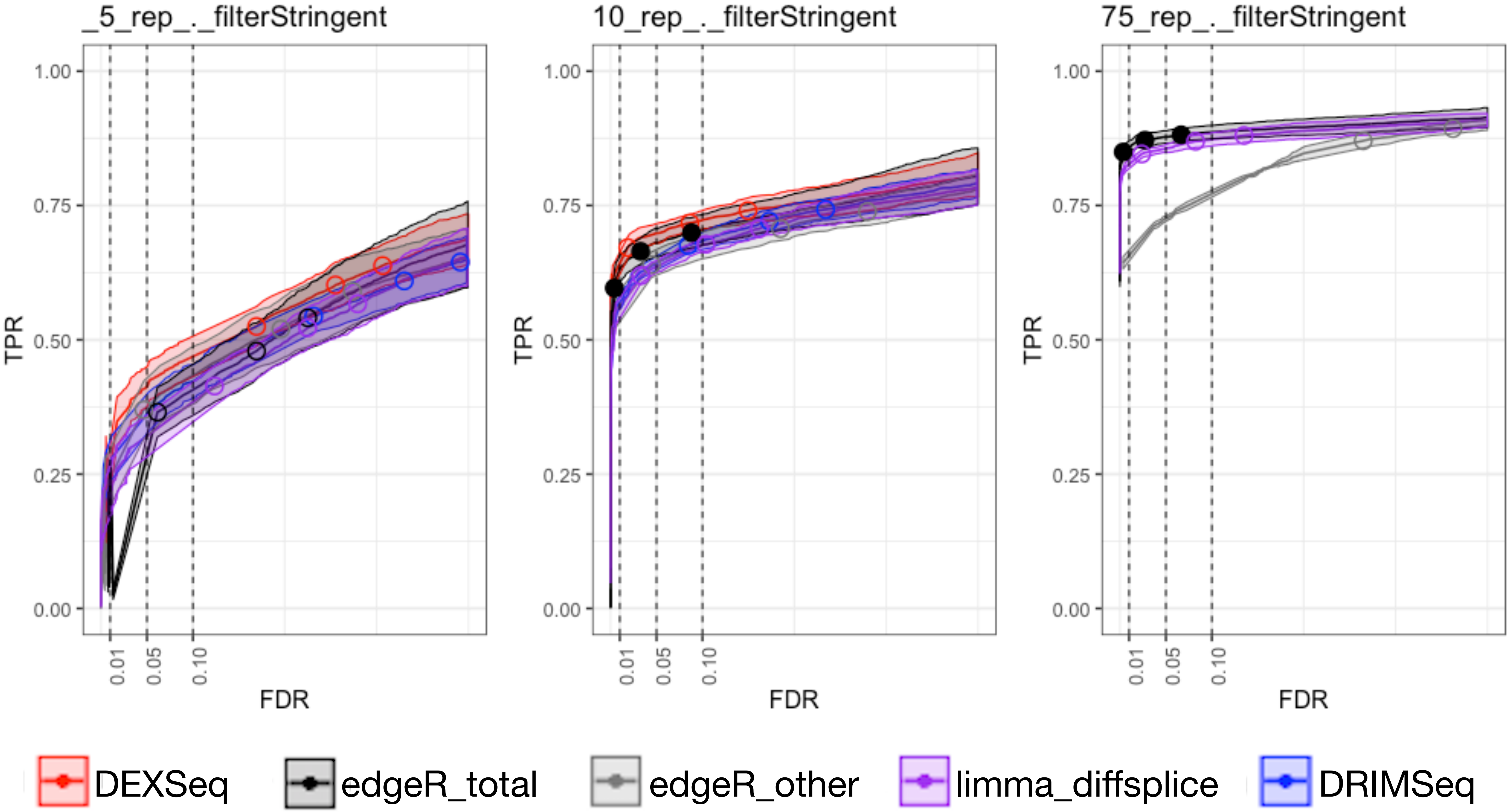
$$\begin{cases} Y_{ti} \sim NB(\mu_{ti}, \varphi_t) \\ \log(\mu_{ti}) = \eta_{ti} \\ \eta_{ti} = \boldsymbol{\beta}^{\mathbf{S}}_{\mathbf{ti}} + \boldsymbol{\beta}^{\mathbf{T}}_{\mathbf{t}} + \boxed{\boldsymbol{\beta}^{\mathbf{TC}}_{\mathbf{tc_i}}} \end{cases}$$

Parametric bulk simulation study

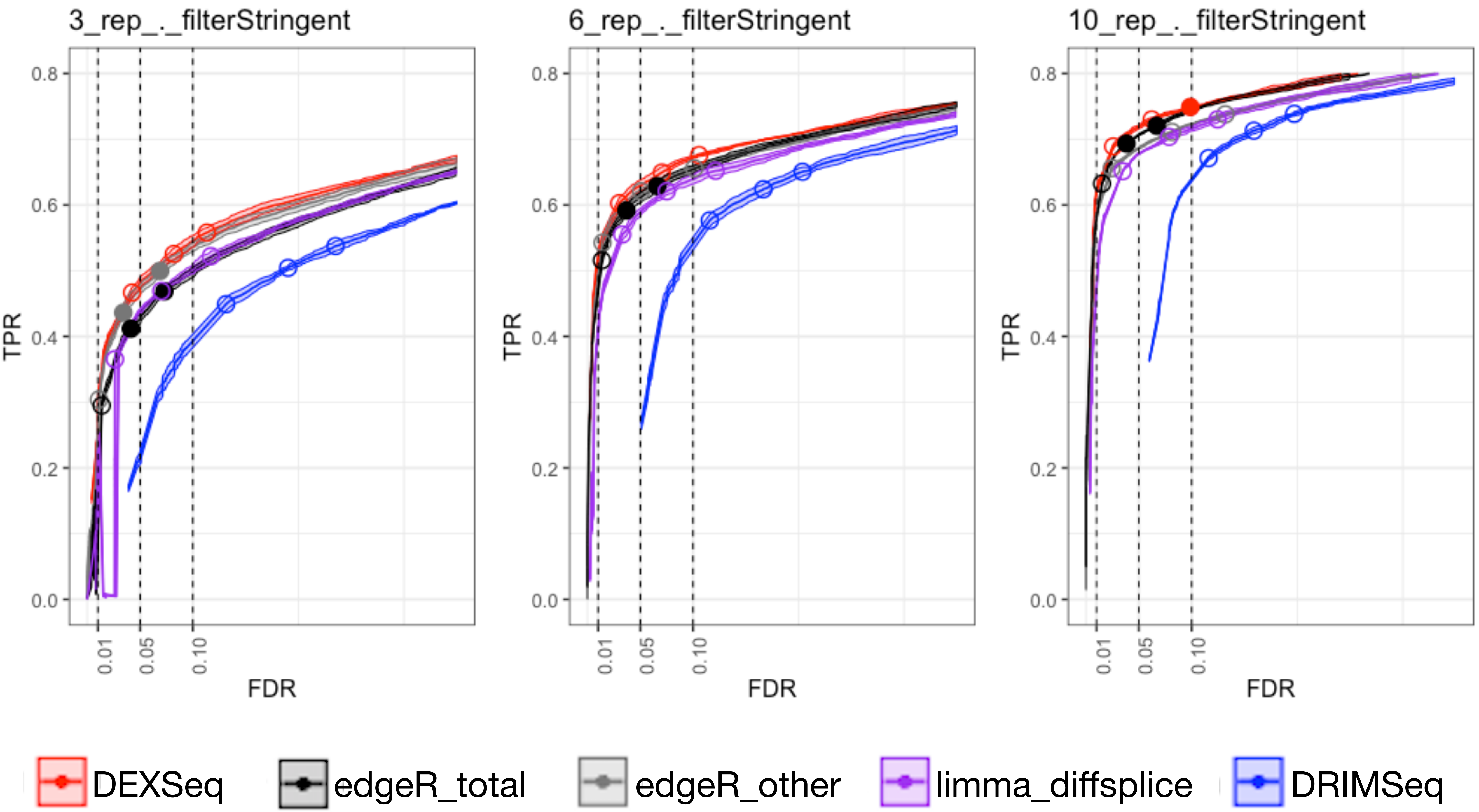
Dataset from Love et al., F1000Research, 7:952 (2018)



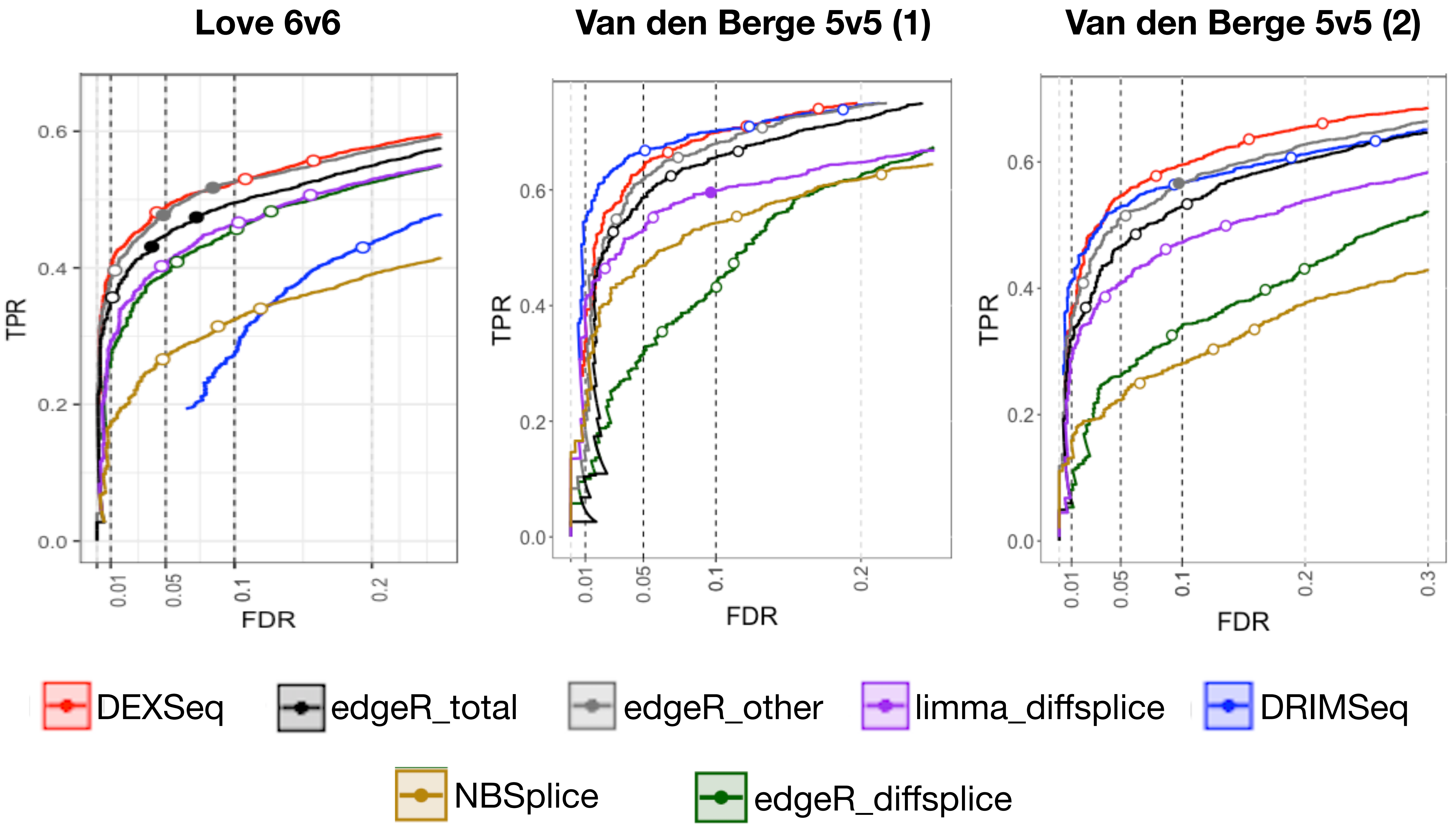
Gtex dataset stringent filtering



Love dataset stringent filtering



Other parametric bulk simulations and additional methods



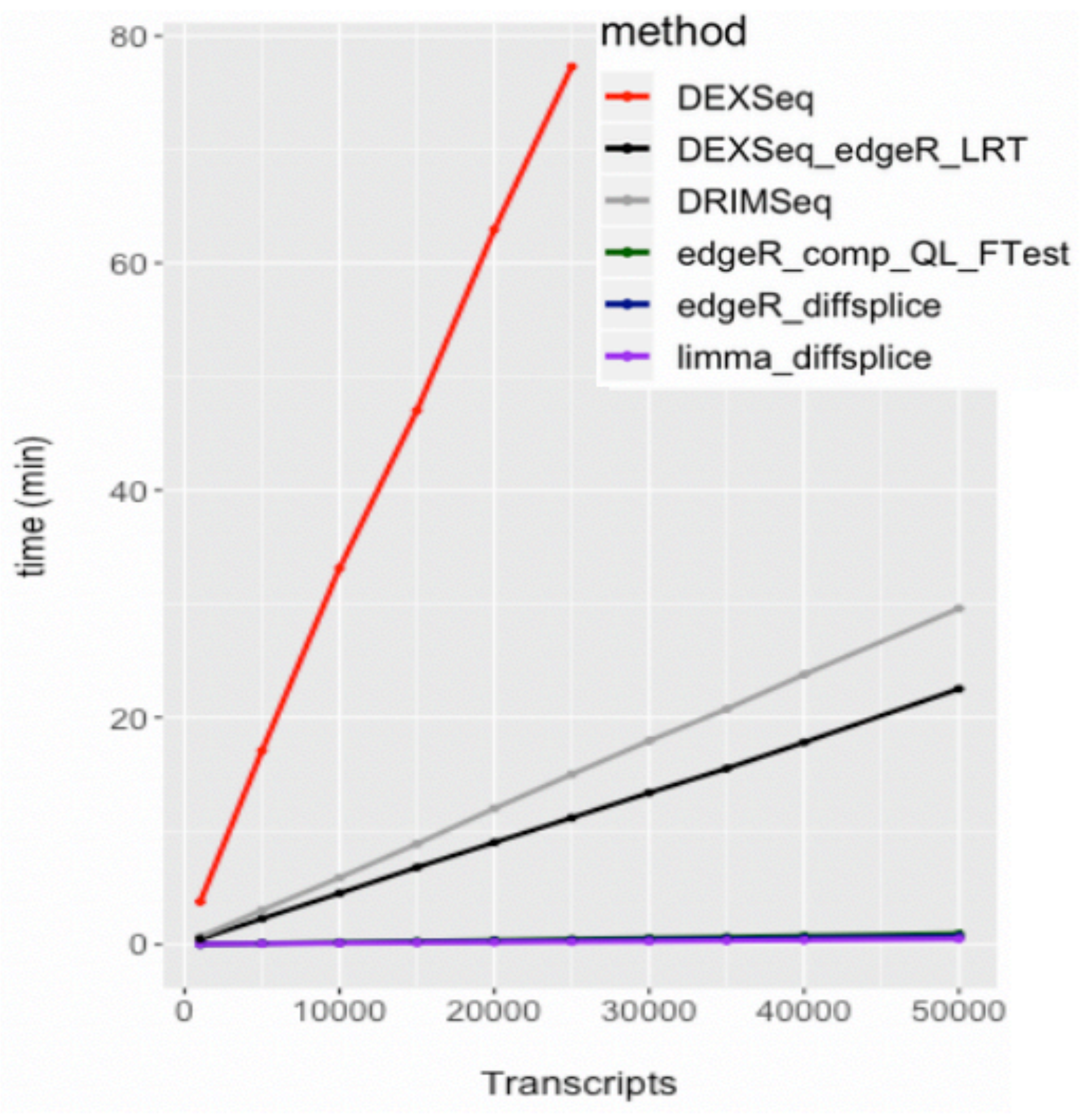
Love 6v6

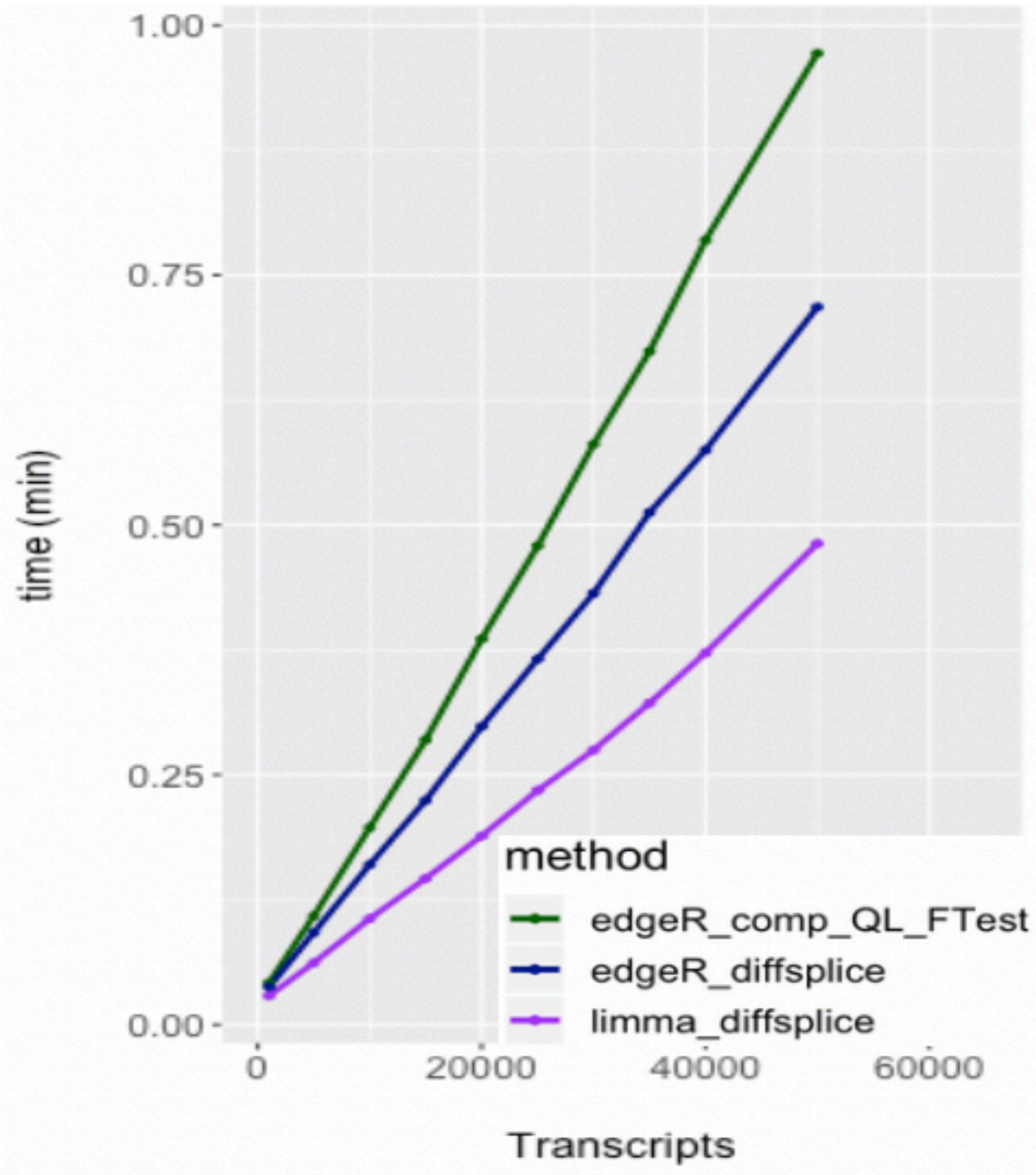
Van den Berge 5v5 (1)

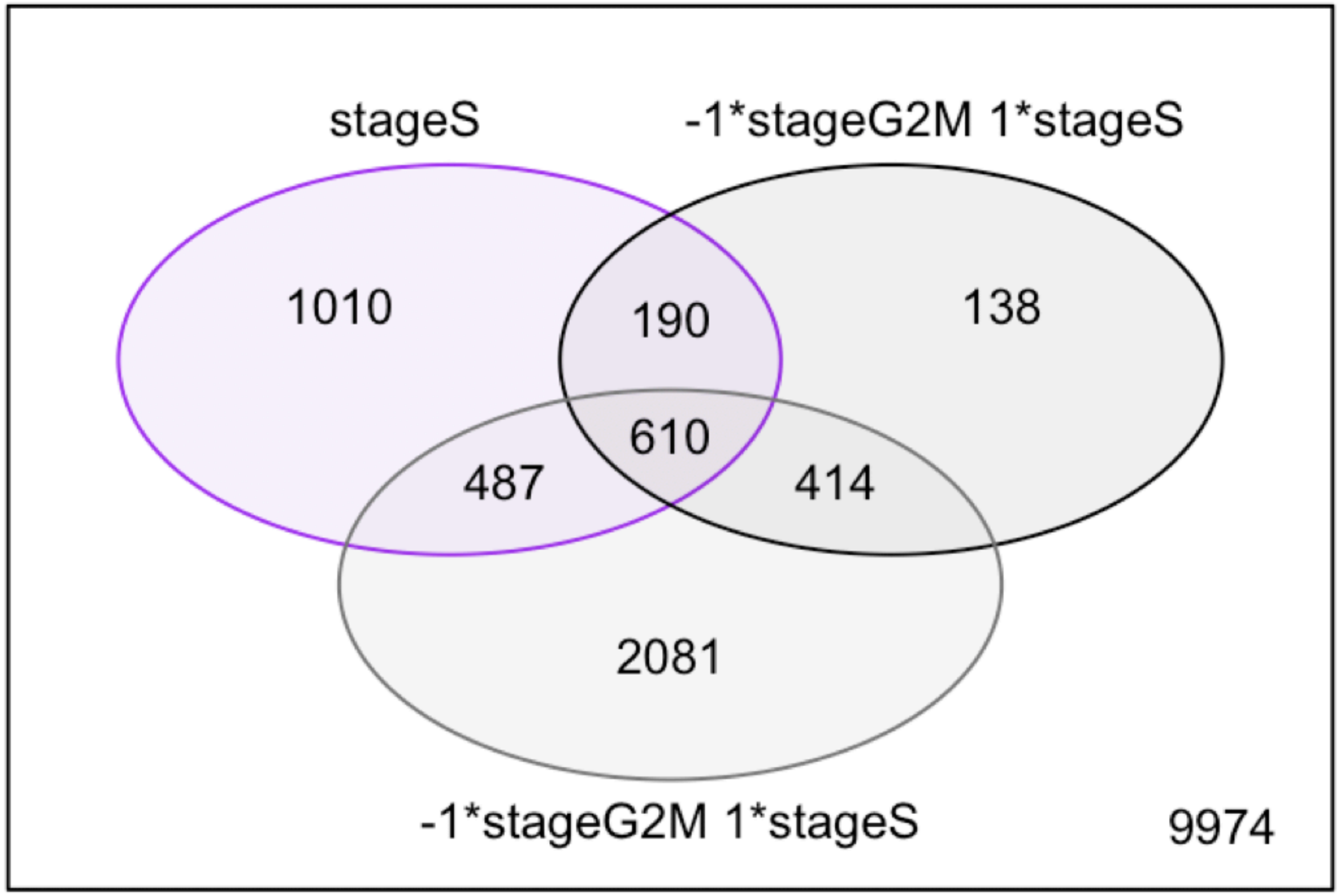
Van den Berge 5v5 (2)

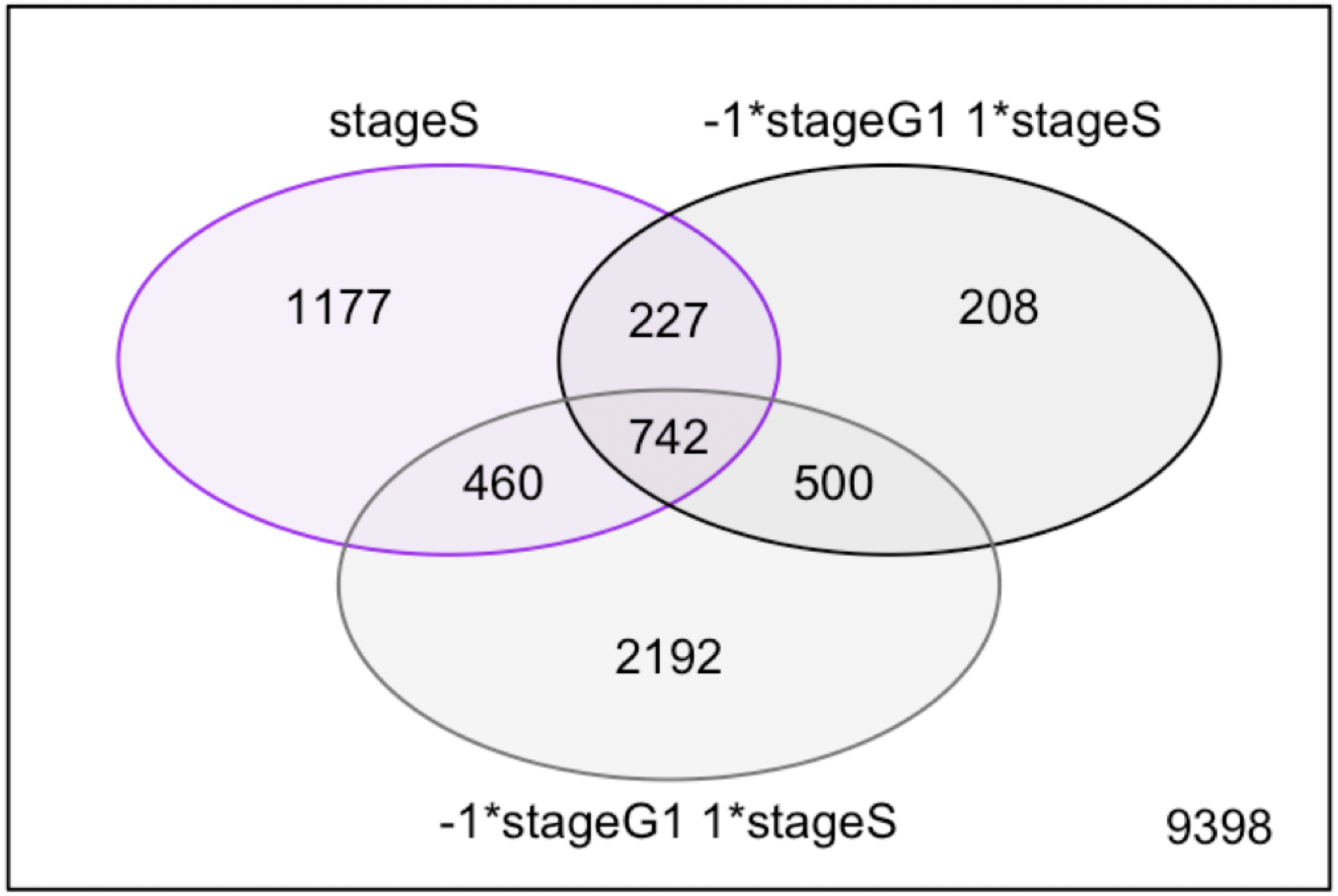
Results - Scalability

- Methods that require sample-level intercepts scale quadratically with the number of cells
- edgeR one order of magnitude faster than DESeq2
- All methods scale linearly with the number of transcripts






stageS
-1*stageG2M 1*stageS
1010
190
138
610
487
414
2081
-1*stageG2M 1*stageS
9974


stageS
-1*stageG1 1*stageS
1177
227
208
742
460
500
2192
-1*stageG1 1*stageS
9398