Differential analysis of RNA-Seq data: gene vs transcript level

Ecole de Bioinformatique AVIESAN-IFB – Roscoff – November 2018

Hugo Varet – `hugo.varet@pasteur.fr`

Transcriptome & Epigenome Platform – Biomics Pole – Citech
Bioinformatics & Biostatistics Hub – C3BI & USR 3756 CNRS

aviesan
alliance nationale
pour les sciences de la vie et de la santé

ifb

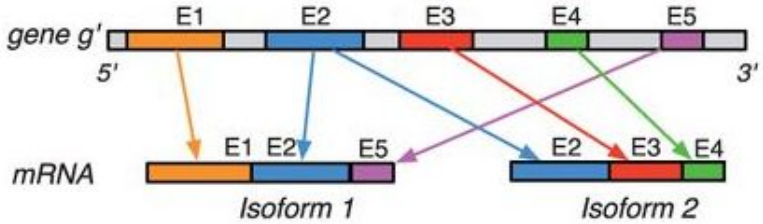
CNRS UPMC
Station Biologique
Roscoff

Outline

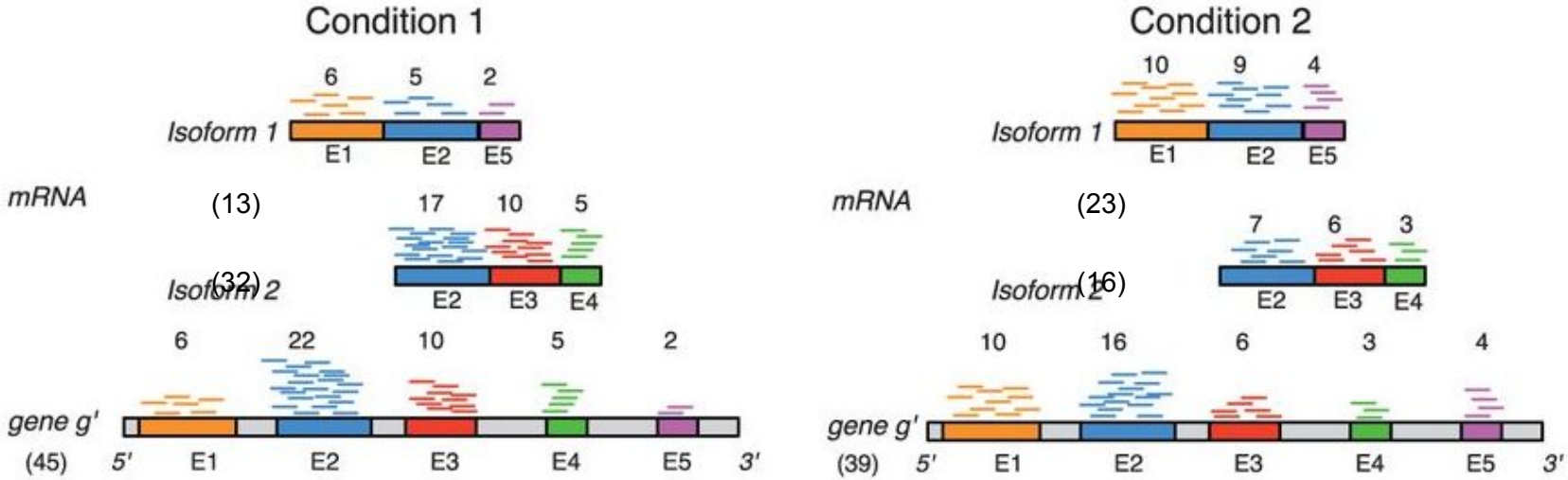
1. Gene vs isoform: several analysis strategies
2. Quantification and testing steps
3. A small example

Context

Gene vs transcript/isoform:

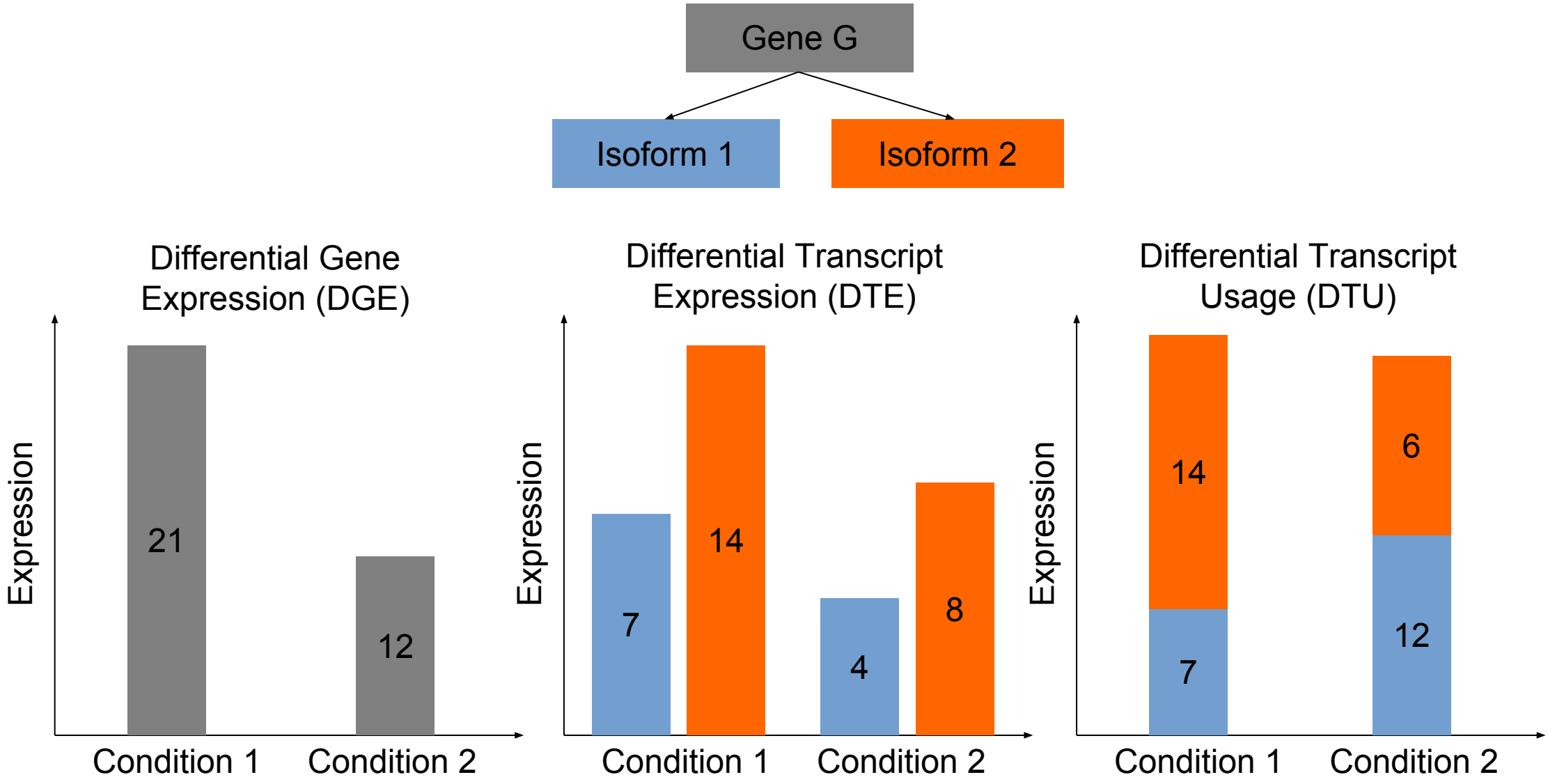


Which quantification level?



Leng et al. *EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments*, Bioinformatics, 2013

Several strategies



Institut Pasteur

Quantification step

Gene-level quantification:

- Mapping on the **genome** using `STAR` or `tophat`
- Counting using `featureCounts` or `htseq-count`

Transcript-level quantification:

- Transcripts Per Million (TPM) estimation using a fast "mapping" on the **transcriptome** with `salmon`
- Possibility to aggregate at the gene level

Testing step

1. **Gene-level quantification** + DGE: one test per gene

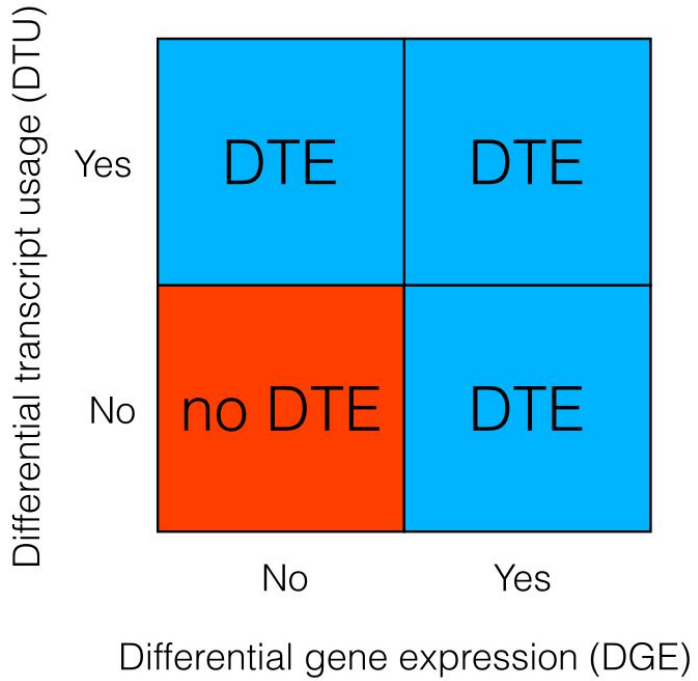
Transcript-level quantification and...
2. Gene aggregation + DGE: one test per gene
3. DTE: one test per transcript
4. DTE + post-analysis gene aggr.: one test per gene
5. DTU: one test per transcript or gene

Quantification level ≠ testing level

DTE, DGE & DTU

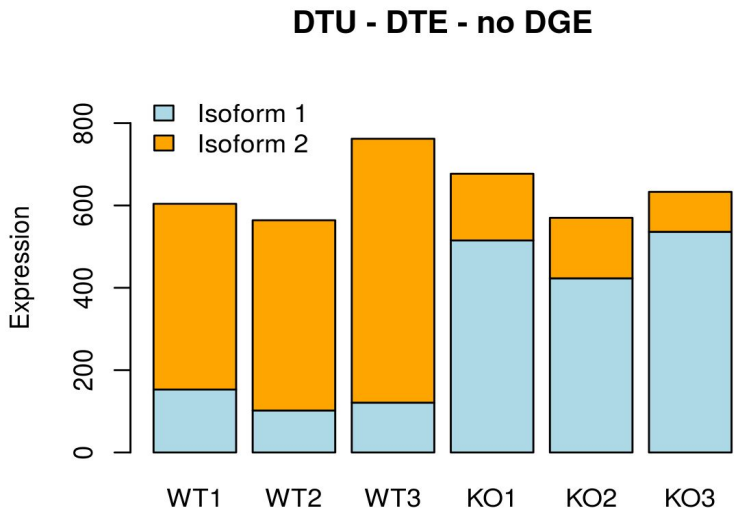
If a transcript is DE (DTE): what about the other transcripts of the gene?

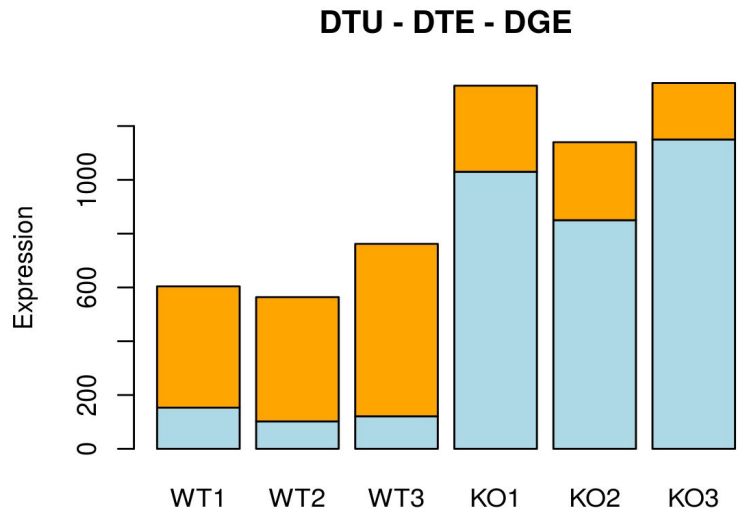
- Because of DGE?
- Because of DTU?
- Because of both DGE and DTU?

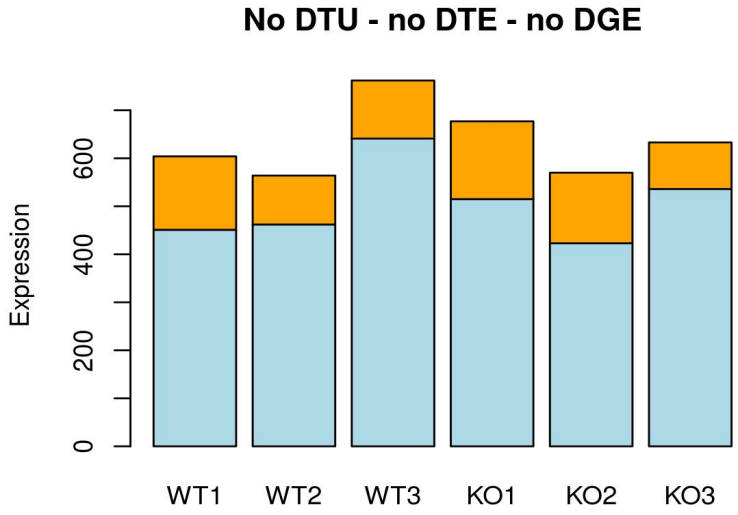


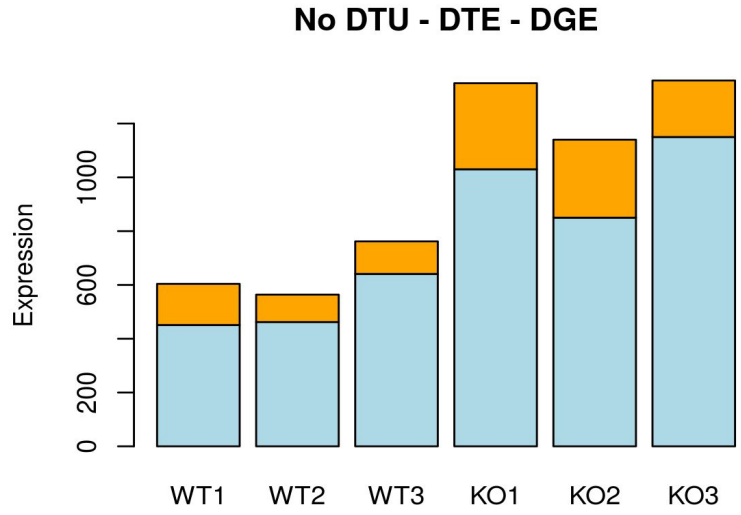
Soneson et al. *Differential analyses for RNA-Seq: transcript-level estimates improve gene-level inferences*, F1000Research, 2016

DTE, DGE & DTU









Institut Pasteur

salmon's output

geneID	txID	WT1	WT2	WT3	KO1	KO2	KO3
geneA	A1	51	21	45	20	23	16
geneB	B1	153	102	121	515	323	536
geneB	B2	451	462	641	162	147	97
geneC	C1	1015	1256	998	1475	1678	1459
geneC	C2	2301	3261	4841	2157	5114	5101
geneC	C3	124	148	187	178	196	209
...	...	...	...	...	...	...	...

DTU analysis

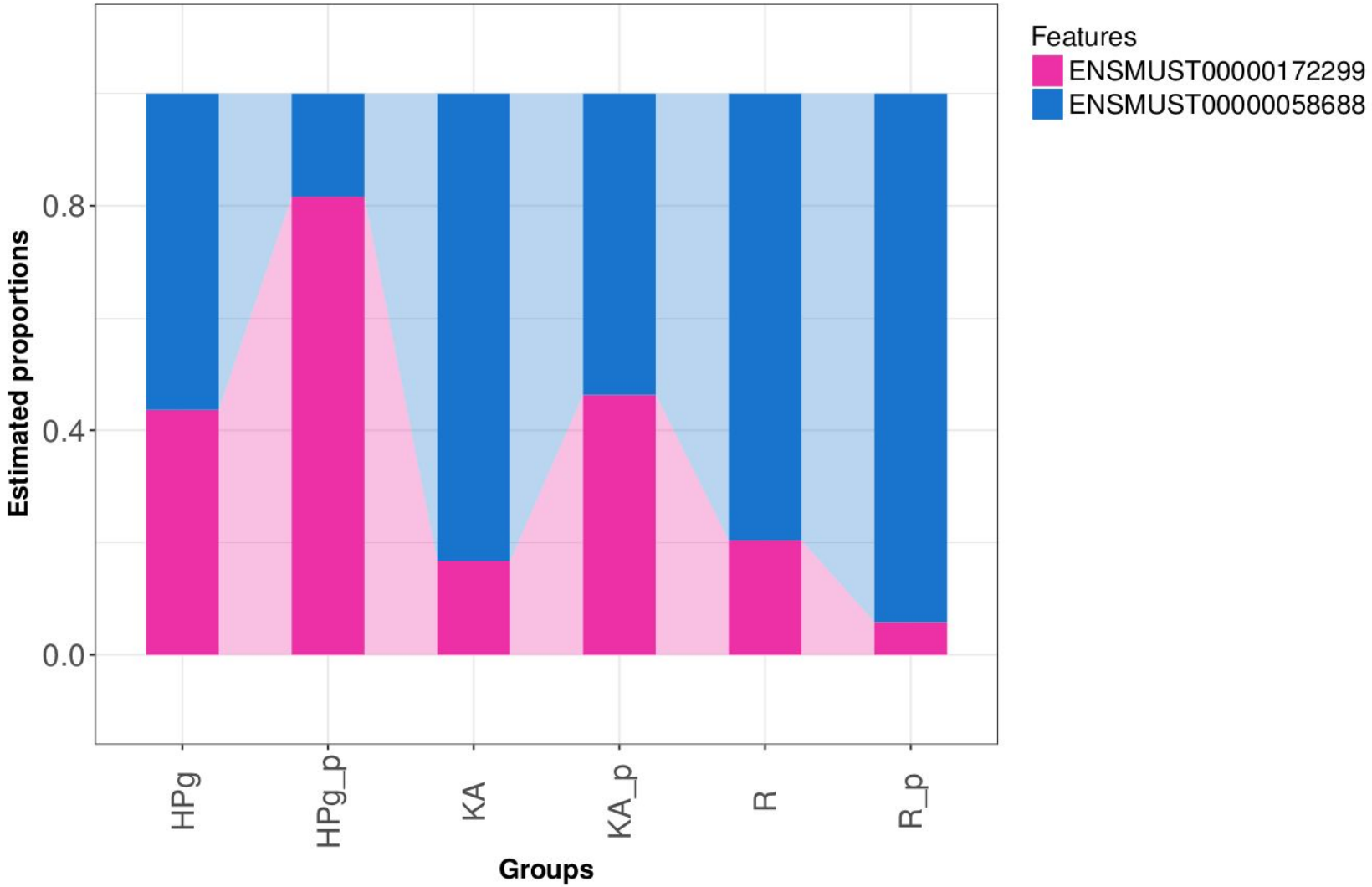
F1000Research F1000Research 2016, 5:1356 Last updated: 13 JUN 2016

METHOD ARTICLE

DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics [version 1; referees: awaiting peer review]

Malgorzata Nowicka[1,2], Mark D. Robinson[1,2]

DTU for *Mus musculus*



Take home message

- Analyses at the isoform level require longer reads: e.g. PE100 instead of the classic SR50

- Need a well defined transcriptome: all the possible isoforms must be present

  → isoform discovery is another question!