

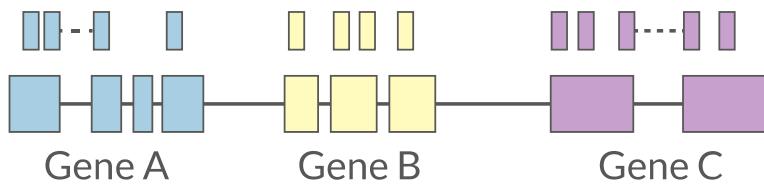
Quantification

RNA-seq data analysis

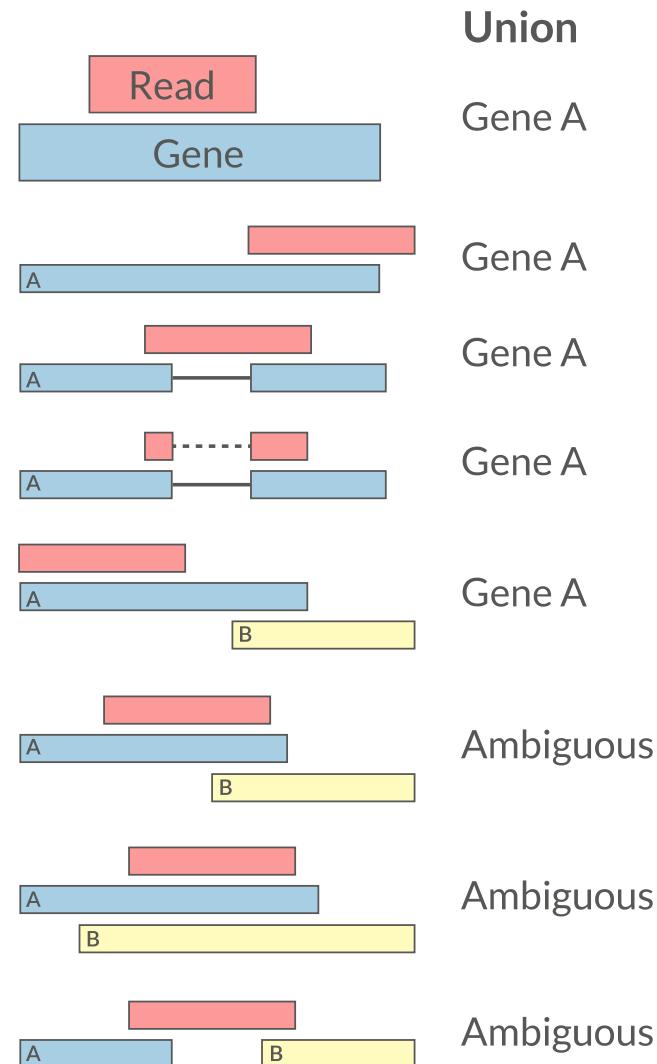
Johan Reimegård

Count the reads

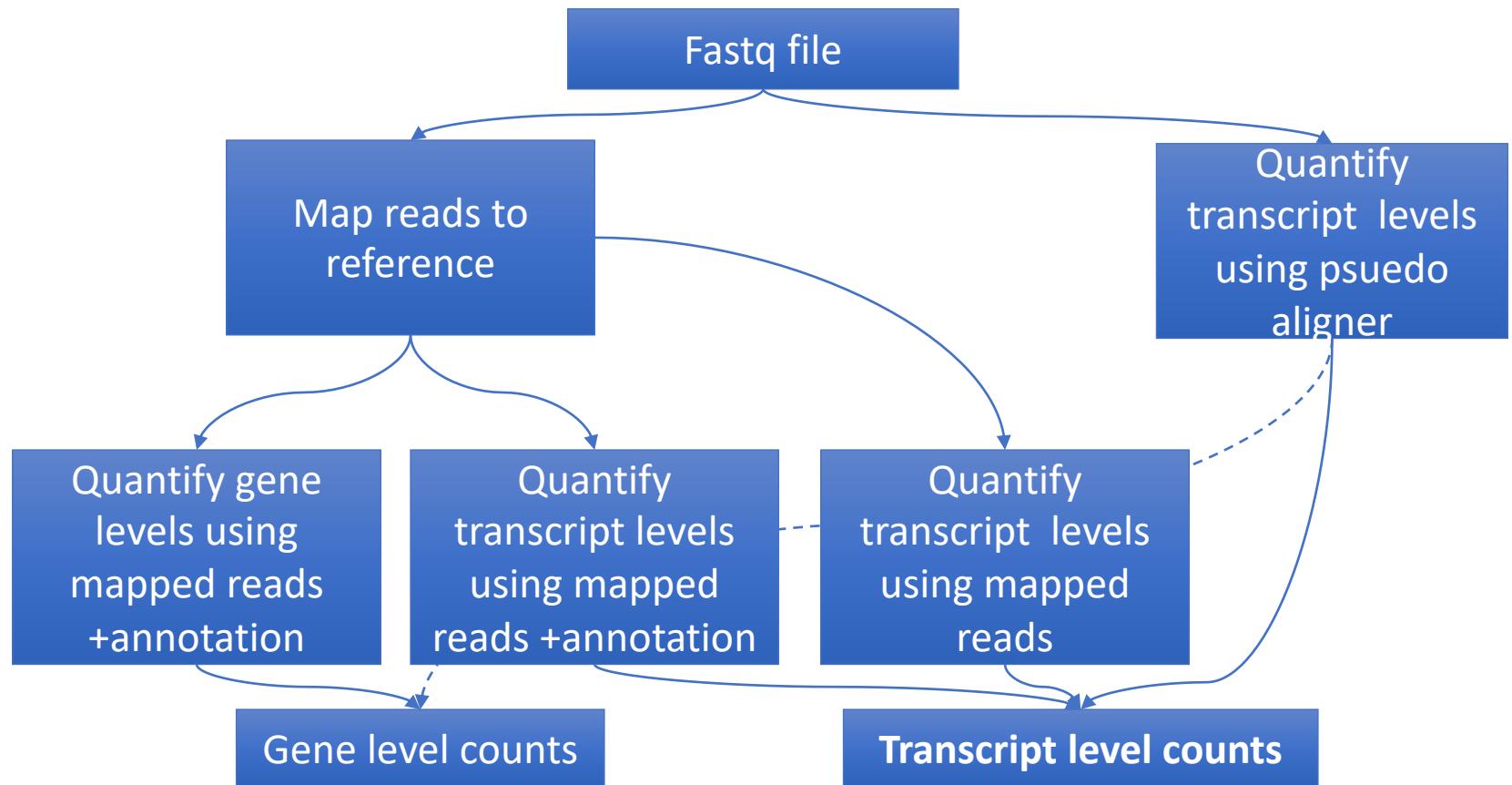
- Read counts = gene expression
- Reads can be quantified on any feature (gene, transcript, exon etc)
- Intersection on gene models
- Gene/Transcript level



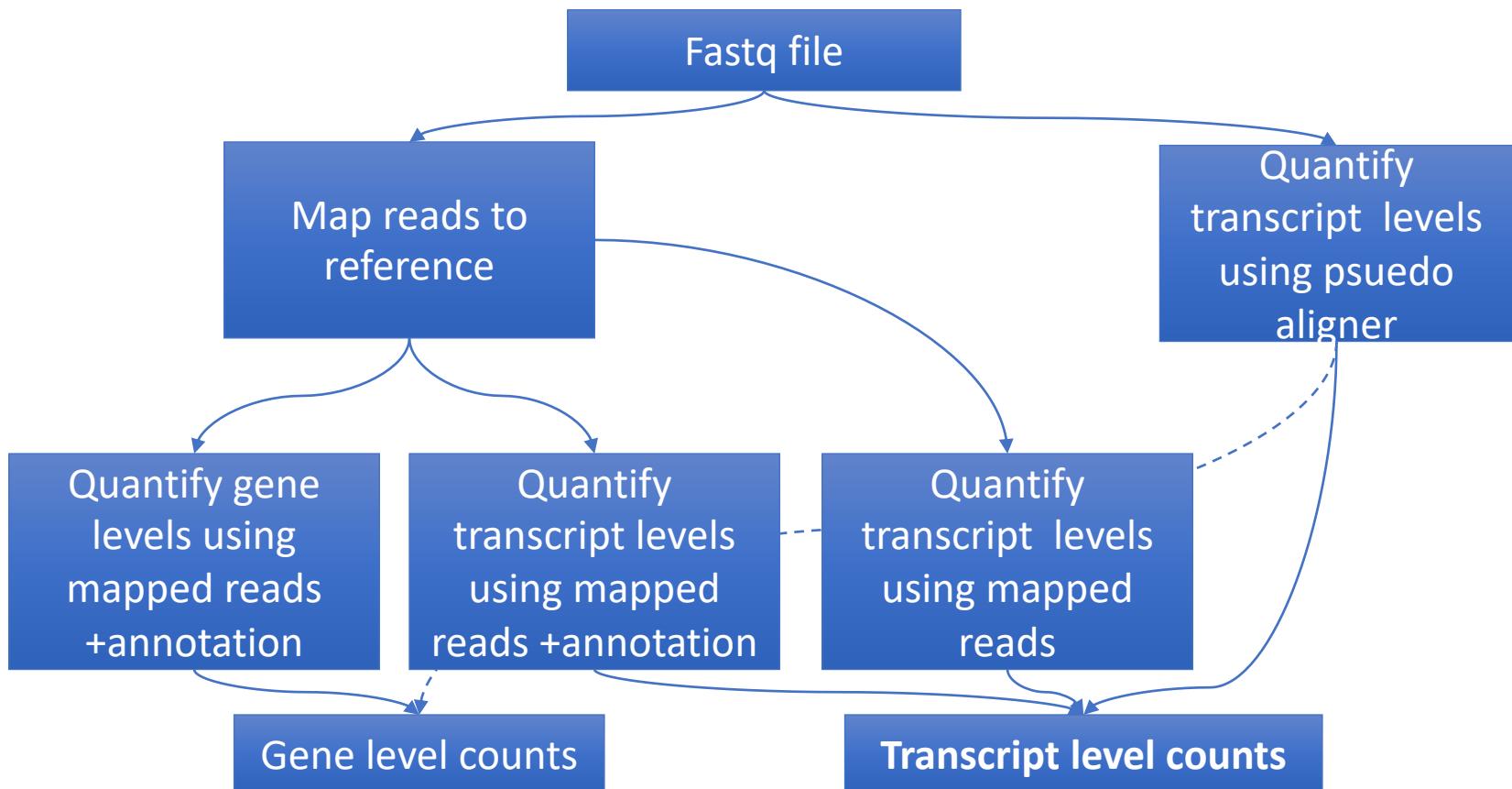
[featureCounts](#), [HTSeq](#)



Different paths to get a count table



Good news is that they are all working very well!!



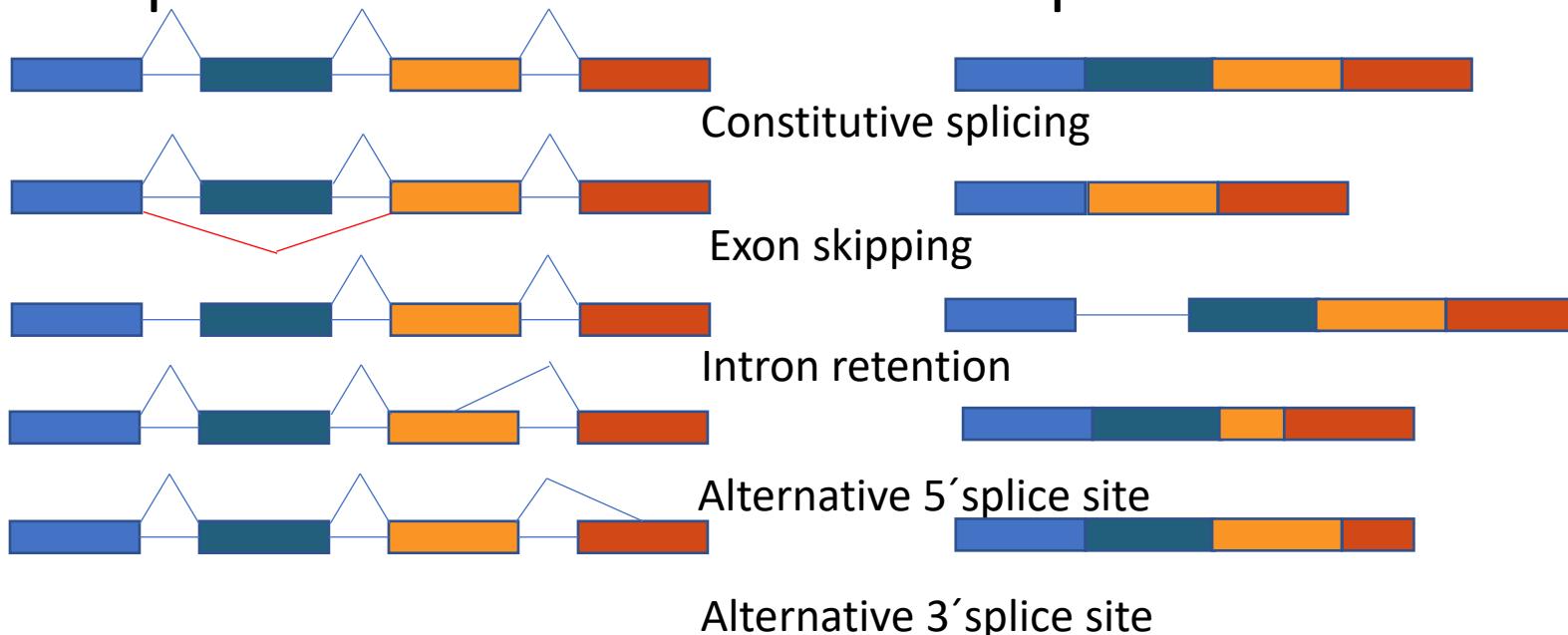
Gene expression estimates



- Expression estimates on gene level



- Expression estimates on transcript level



Gene level analysis

SCIENTIFIC REPORTS



OPEN

Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data

Received: 18 July 2016

Accepted: 3 April 2017

Published online: 08 May 2017

Celine Everaert^{1,2,3}, Manuel Luypaert⁴, Jesper L. V. Maag  ⁵, Quek Xiu Cheng⁵, Marcel E. Dinger  ⁵, Jan Hellemans⁴ & Pieter Mestdagh^{1,2,3}

In this paper they had done qPCR expression levels on all genes in two states. And at the same time created RNA seq libraries for the two states.

Expression levels are similar between RT-qPCR and RNA-seq data

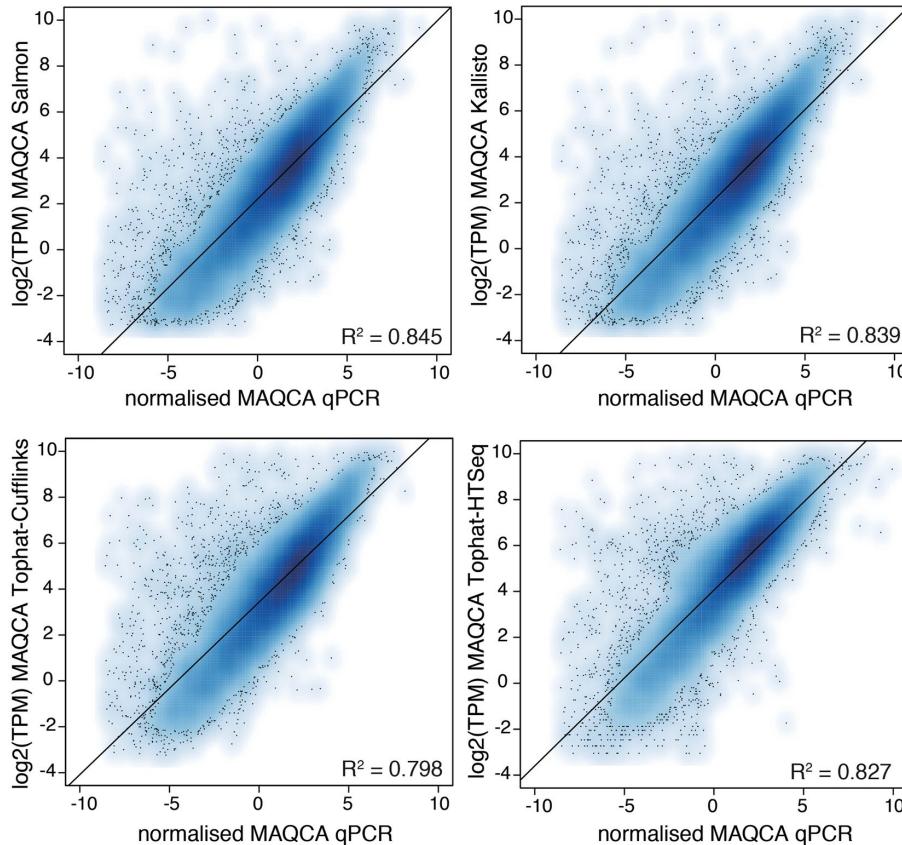
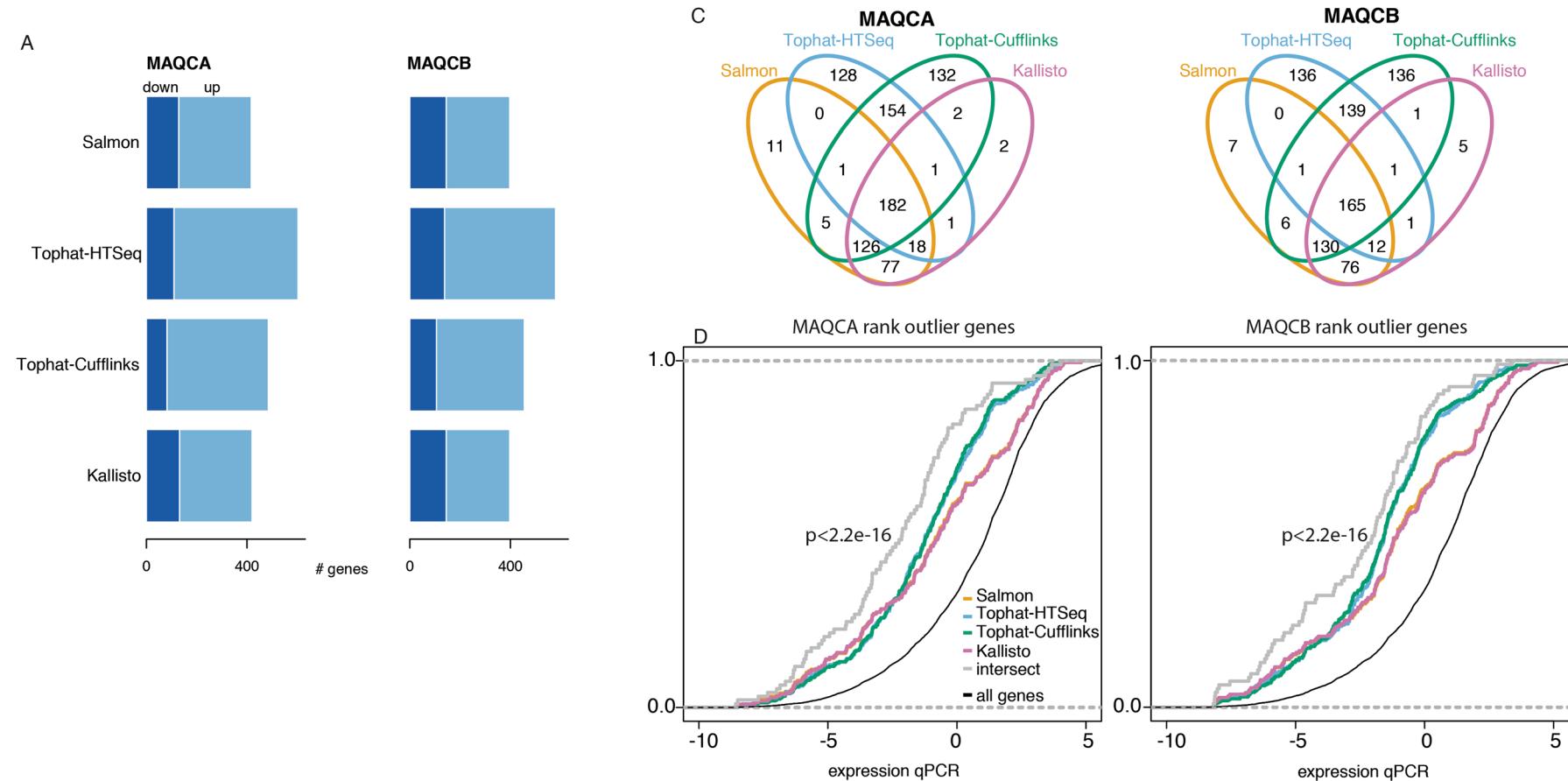
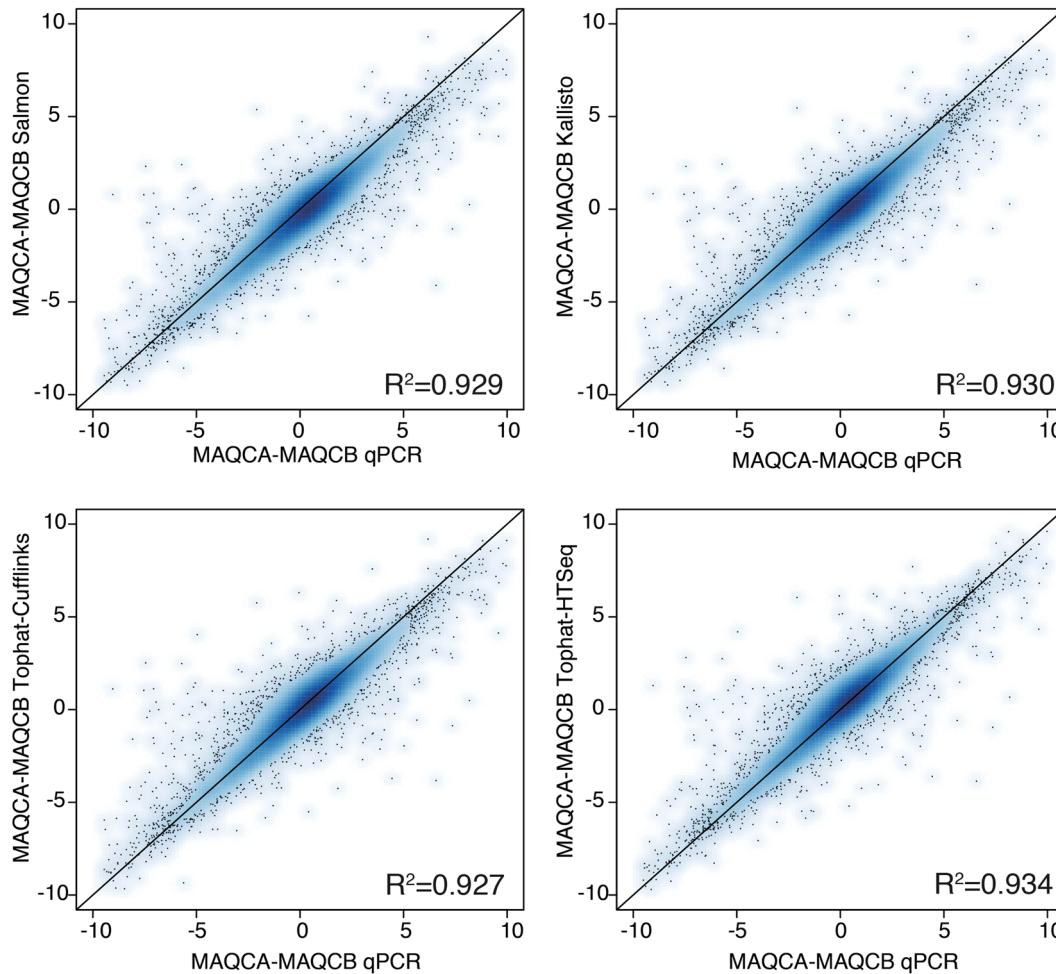


Figure 1. Gene expression correlation between RT-qPCR and RNA-seq data. The Pearson correlation coefficients and linear regression line are indicated. Results are based on RNA-seq data from dataset 1.

Lowly expressed genes are more problematic to identify using RNA seq



Most problems are consistent so they disappear when you do diff-exp analysis



Transcript level analysis

Zhang *et al.* BMC Genomics (2017) 18:583
DOI 10.1186/s12864-017-4002-1

BMC Genomics

RESEARCH ARTICLE

Open Access

Evaluation and comparison of computational tools for RNA-seq isoform quantification



CrossMark

Chi Zhang¹, Baohong Zhang¹, Lih-Ling Lin² and Shanrong Zhao^{1*}

Methods used in paper

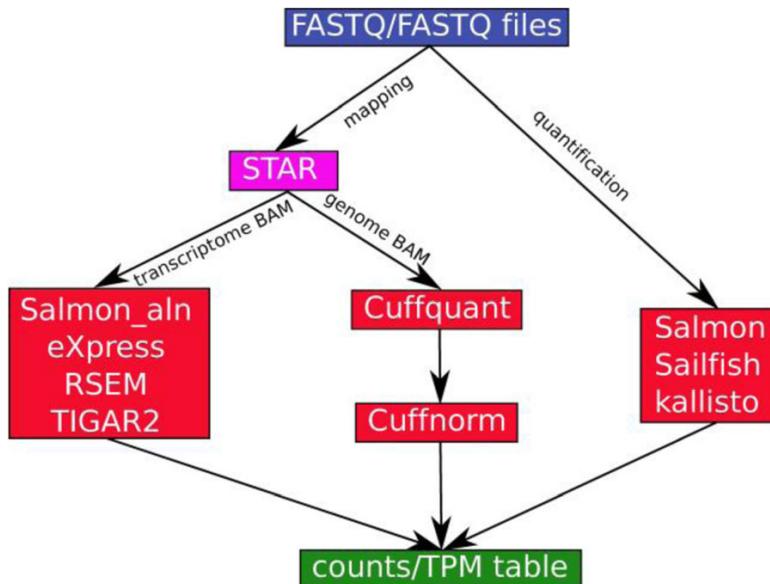


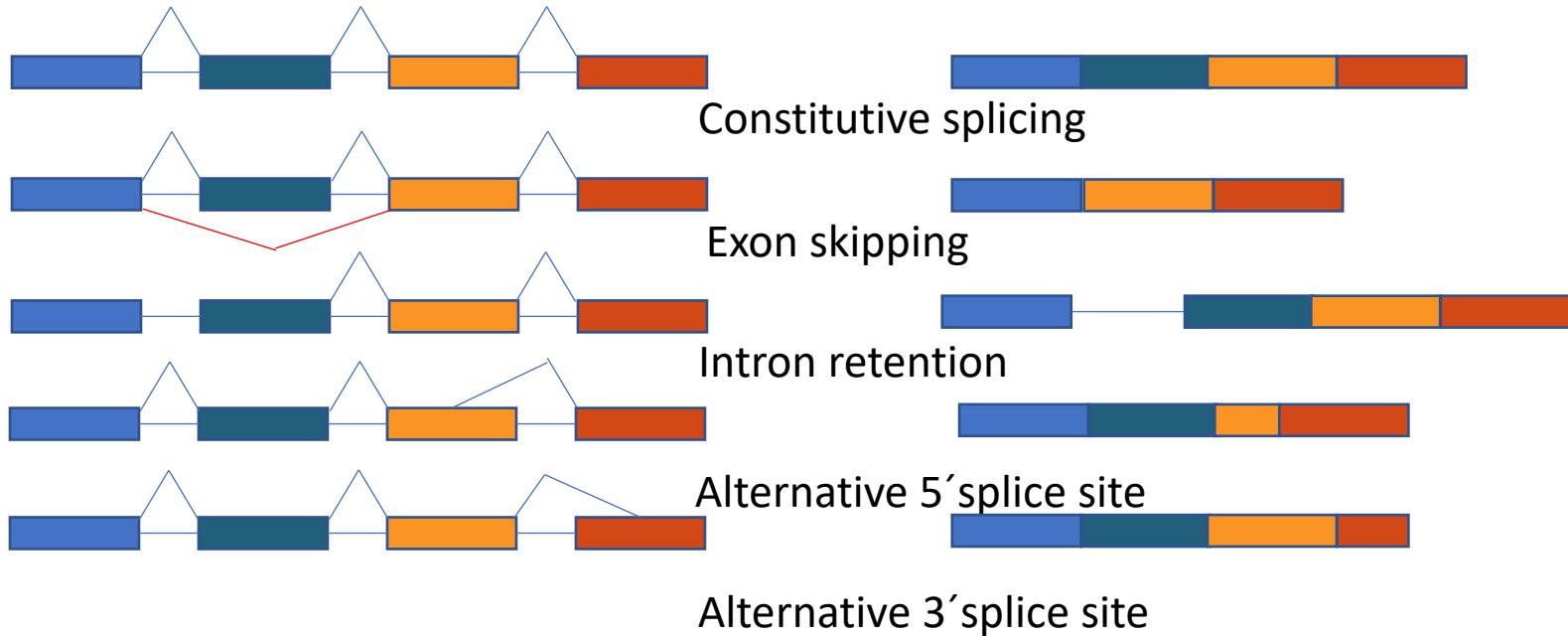
Table 1 Run time metrics of each method on 50 million paired-end reads of length 76 bp in an high performance computing cluster

| | Memory (Gb) | Run time (min) | Algorithm | Multi-thread |
|------------|-------------|----------------|-----------|--------------|
| Cufflinks | 3.5 | 117 | ML | Yes |
| RSEM | 5.6 | 154 | ML | Yes |
| eXpress | <u>0.55</u> | 30 | ML | No |
| TIGAR2 | 28.3 | 1045 | VB | Yes |
| kallisto | 3.8 | 7 | ML | Yes |
| Salmon | 6.6 | 6 | VB/ML | Yes |
| Salmon_aln | 3 | 7 | VB/ML | Yes |
| Sailfish | 6.3 | <u>5</u> | VB/ML | Yes |

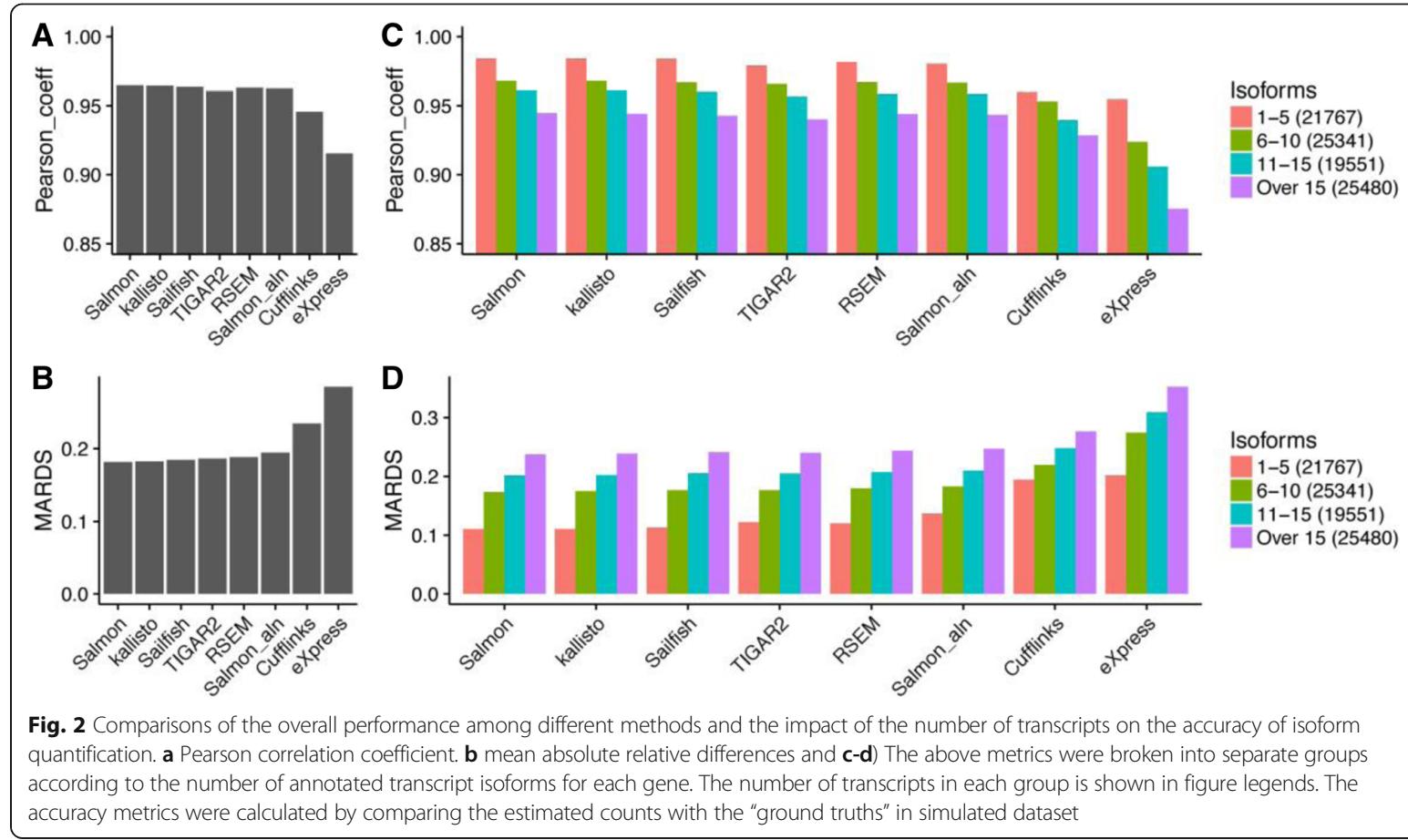
For methods that support multi-threading, eight threads were used. For alignment-free methods (Kallisto, Salmon and Sailfish), a mapping step was included. The best performer in each category is underlined and the worst performer is in bold
ML Maximum Likelihood, VB Variational Bayes

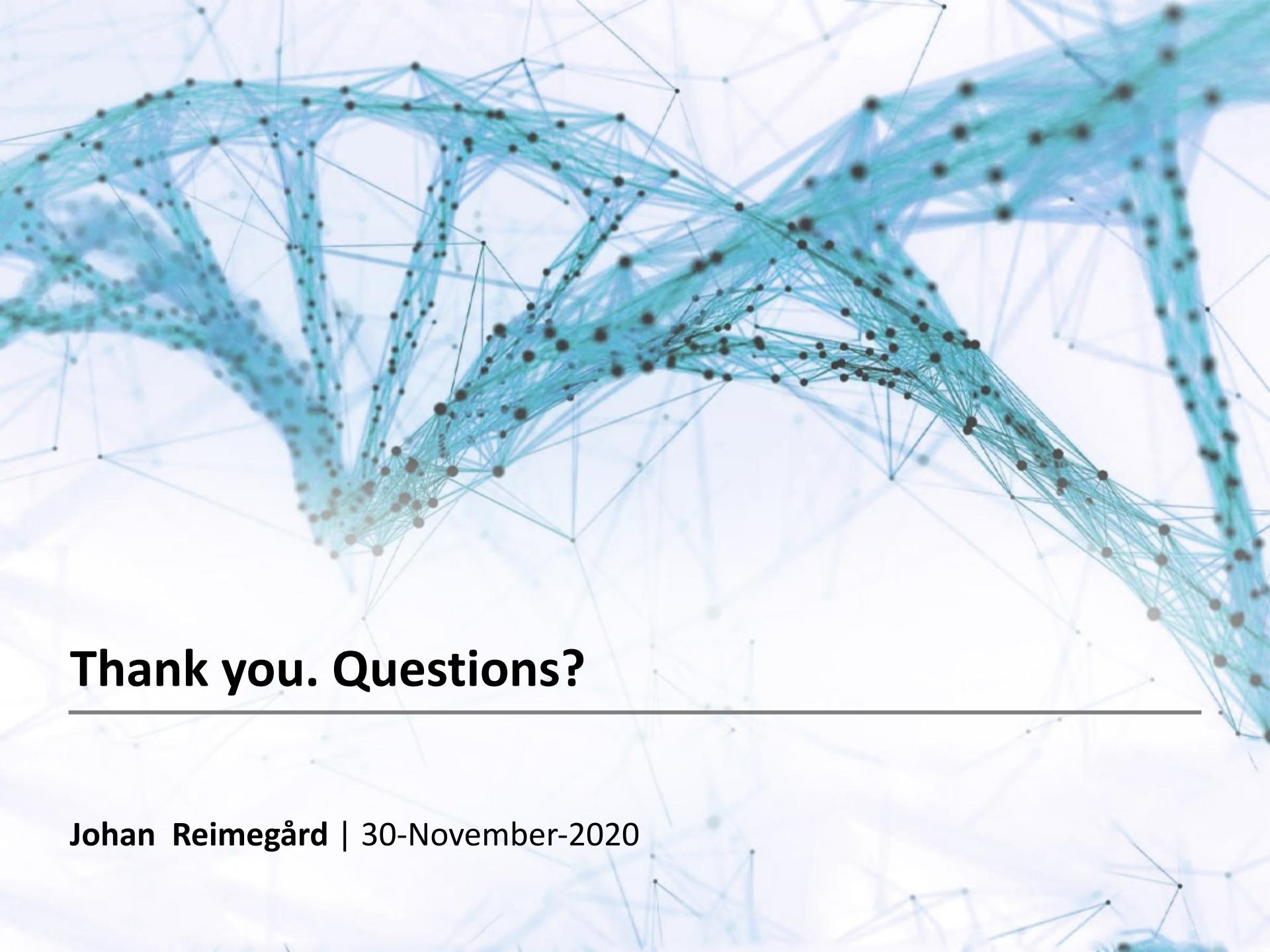
Reads are generated *in silico*

- Expression estimates on transcript level



Isoform quantification problematic for genes with many isoforms



The background of the slide features a complex, abstract network graph. It consists of numerous small, dark brown dots representing nodes, connected by a dense web of thin, translucent blue lines representing edges. The graph is highly interconnected, forming several large, irregular clusters and many smaller, isolated nodes. The overall effect is one of a complex system or a web of relationships.

Thank you. Questions?

Johan Reimegård | 30-November-2020

Results are very similar between methods

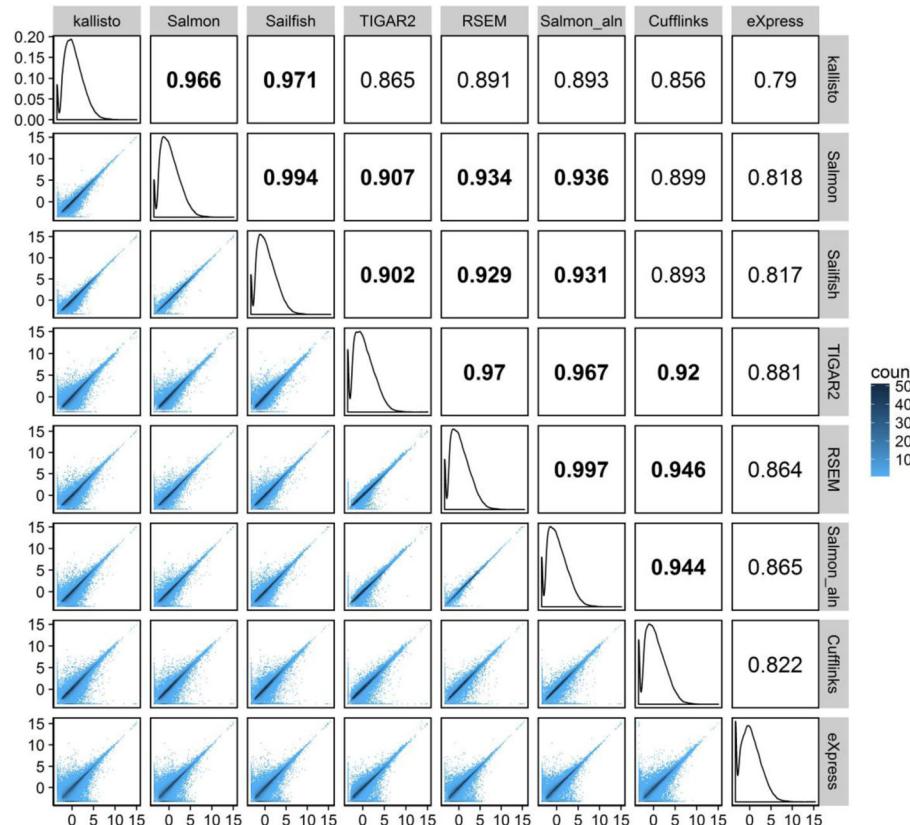


Fig. 5 Pairwise correlation of estimated TPM values for all transcripts between methods for the HBRR-C4 sample. The distribution of transcripts' TPMs from each method was plotted on the diagonal panels. Pairwise density plots and R^2 values are shown in the lower and upper triangular panels, respectively. R^2 values over 0.9 are in *bold*. Methods are grouped using hierarchical clustering