NBIS SciLifeLab

Quantification

RNA-seq data analysis

Johan Reimegård | 13-May-2019

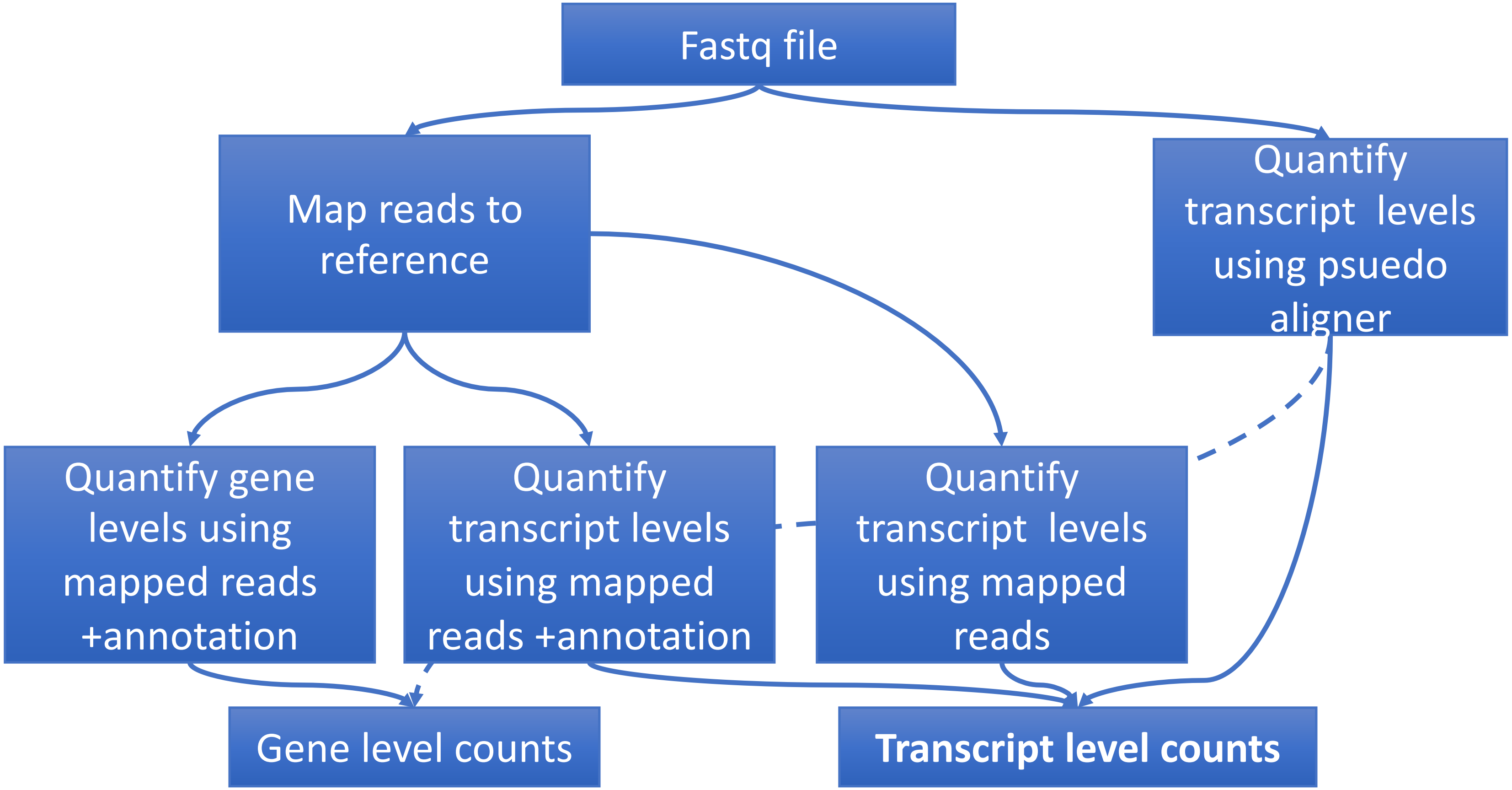
Initial steps in RNA-seq data processing

(for species with a reference genome)

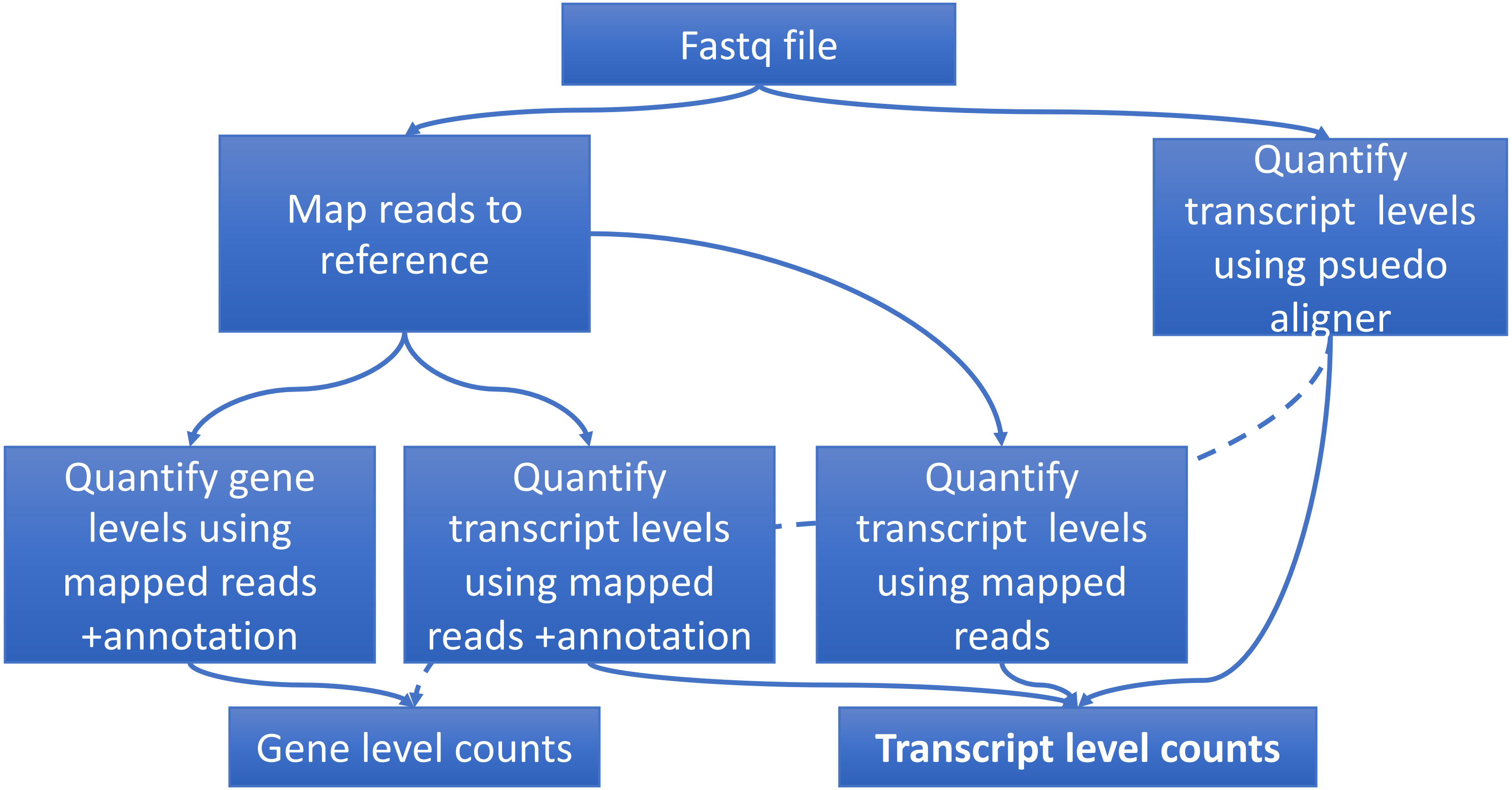
1. Quality checks on reads
2. Trim 3' adapters (optional)
3. Index reference genome
4. Map reads to genome (output in SAM or BAM format)
5. Convert results to a sorted, indexed BAM file
6. Quality checks on mapped reads
7. Visualize read mappings on the genome

Followed by further analyses…

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 Luypaert[4], Jesper L. V. Maag[5], Quek Xiu Cheng[5], Marcel E. Dinger[5], Jan Hellemans[4] & Pieter Mestdagh[1,2,3]

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

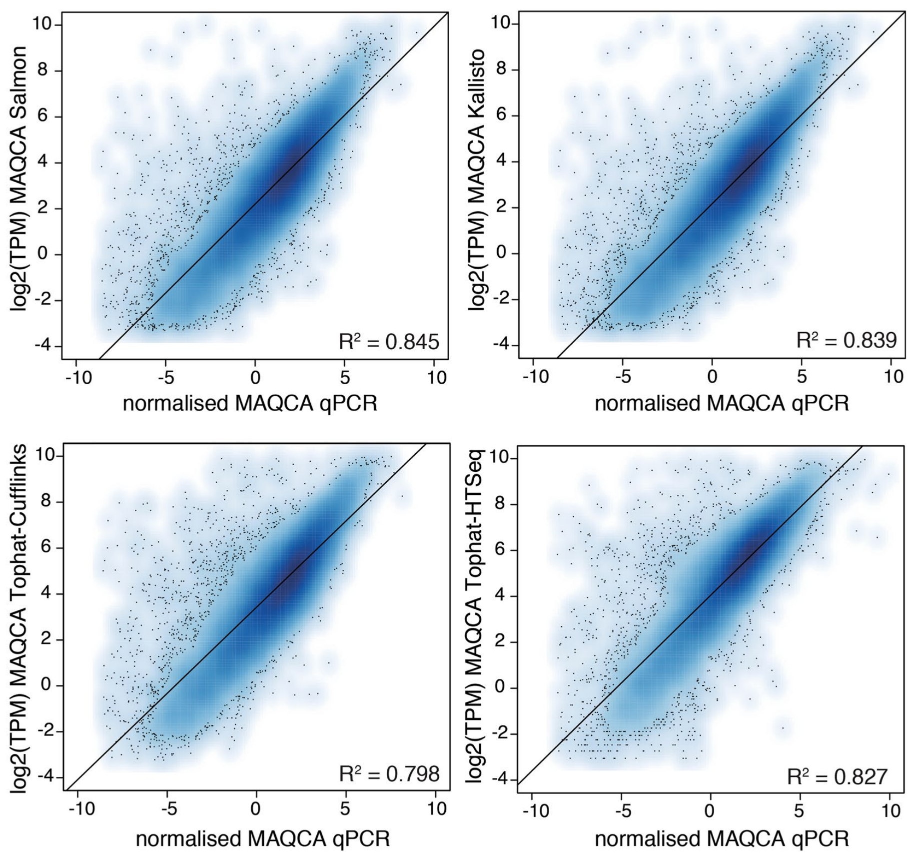
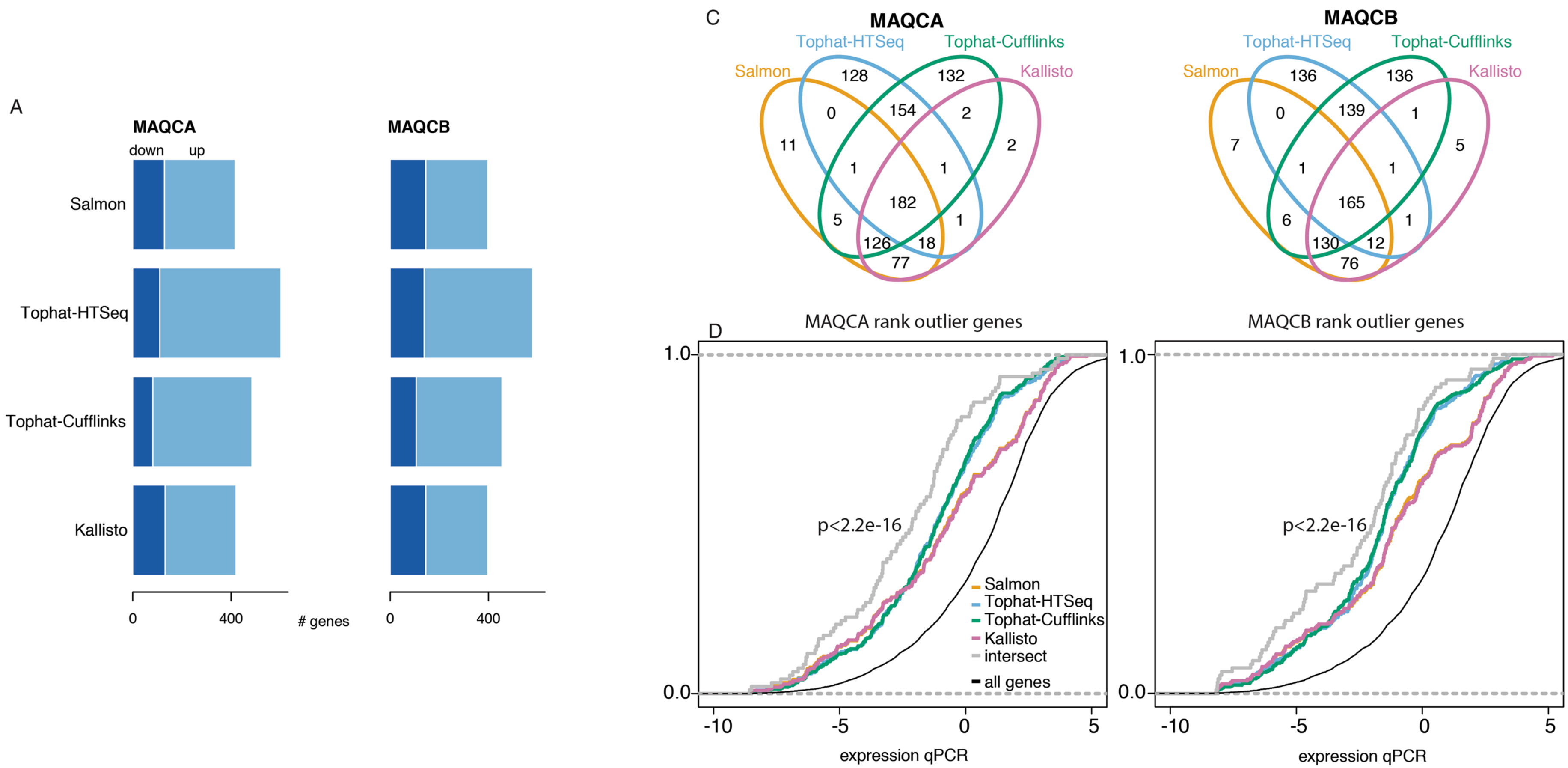
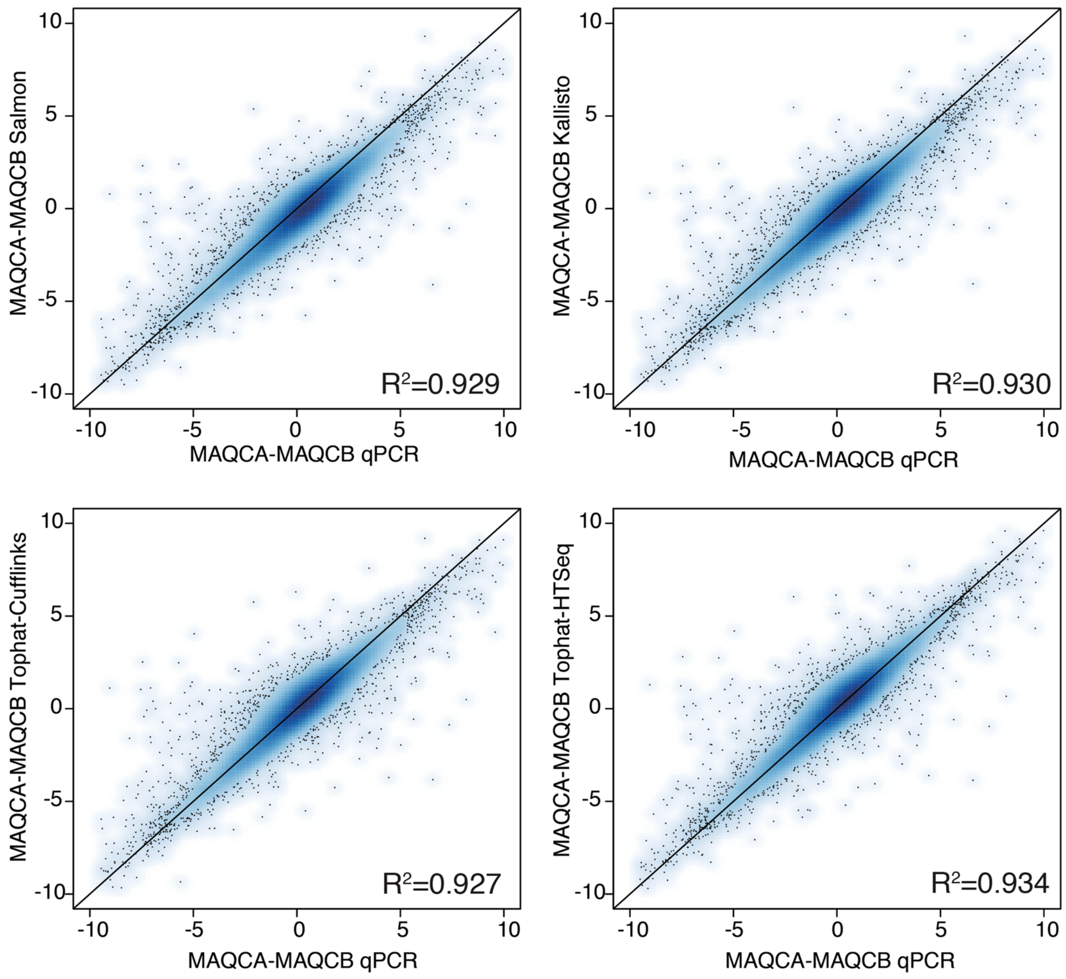


Figure 1. Gene expression correlation between RT-qPCR and RNA-seq data. The Pearson correlation coefficients and linear regression line are indicated. Results are based on RNA-seq data from dataset 1.

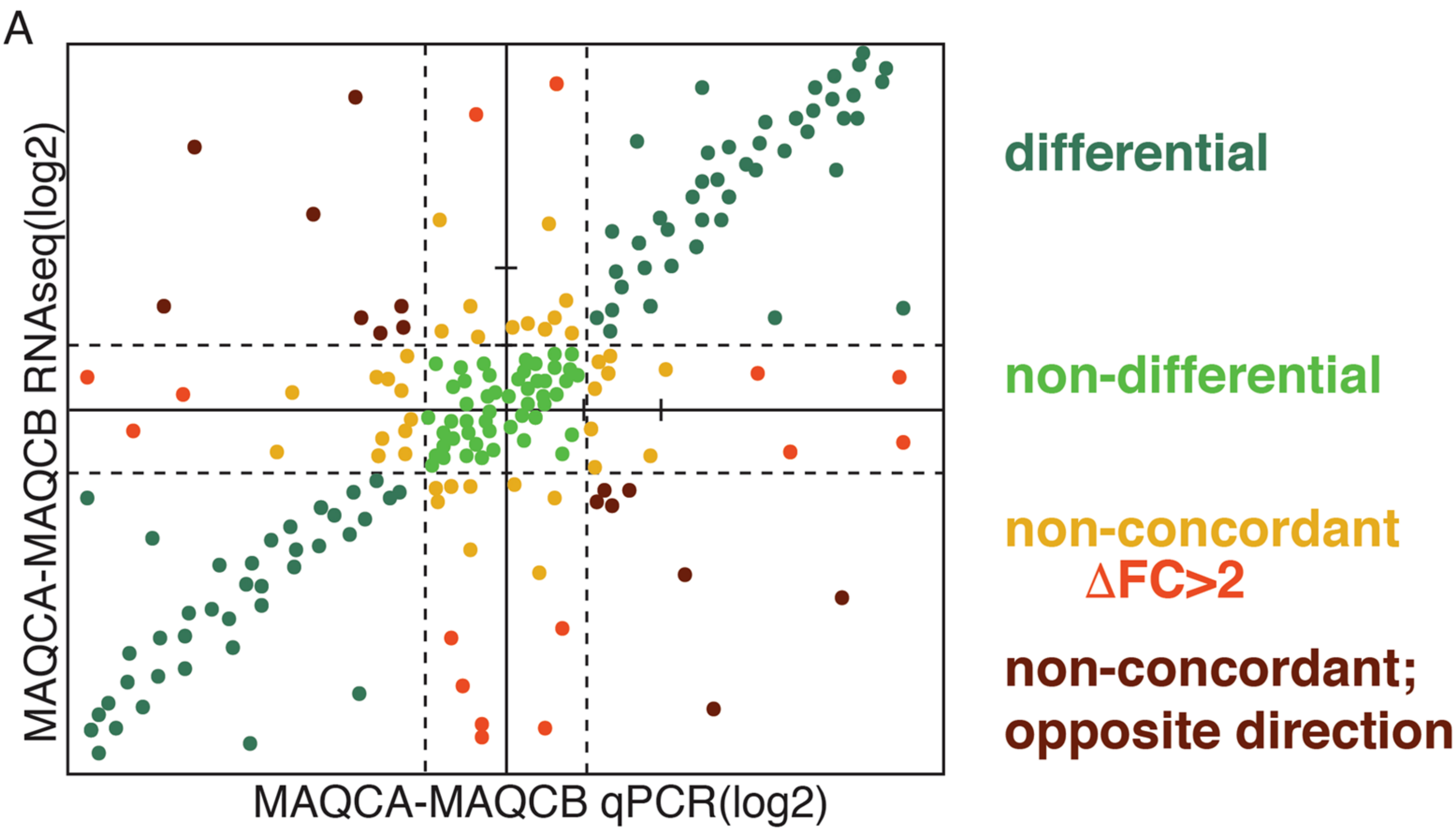
Lowly expressed genes are more problematic to identify using RNA seq



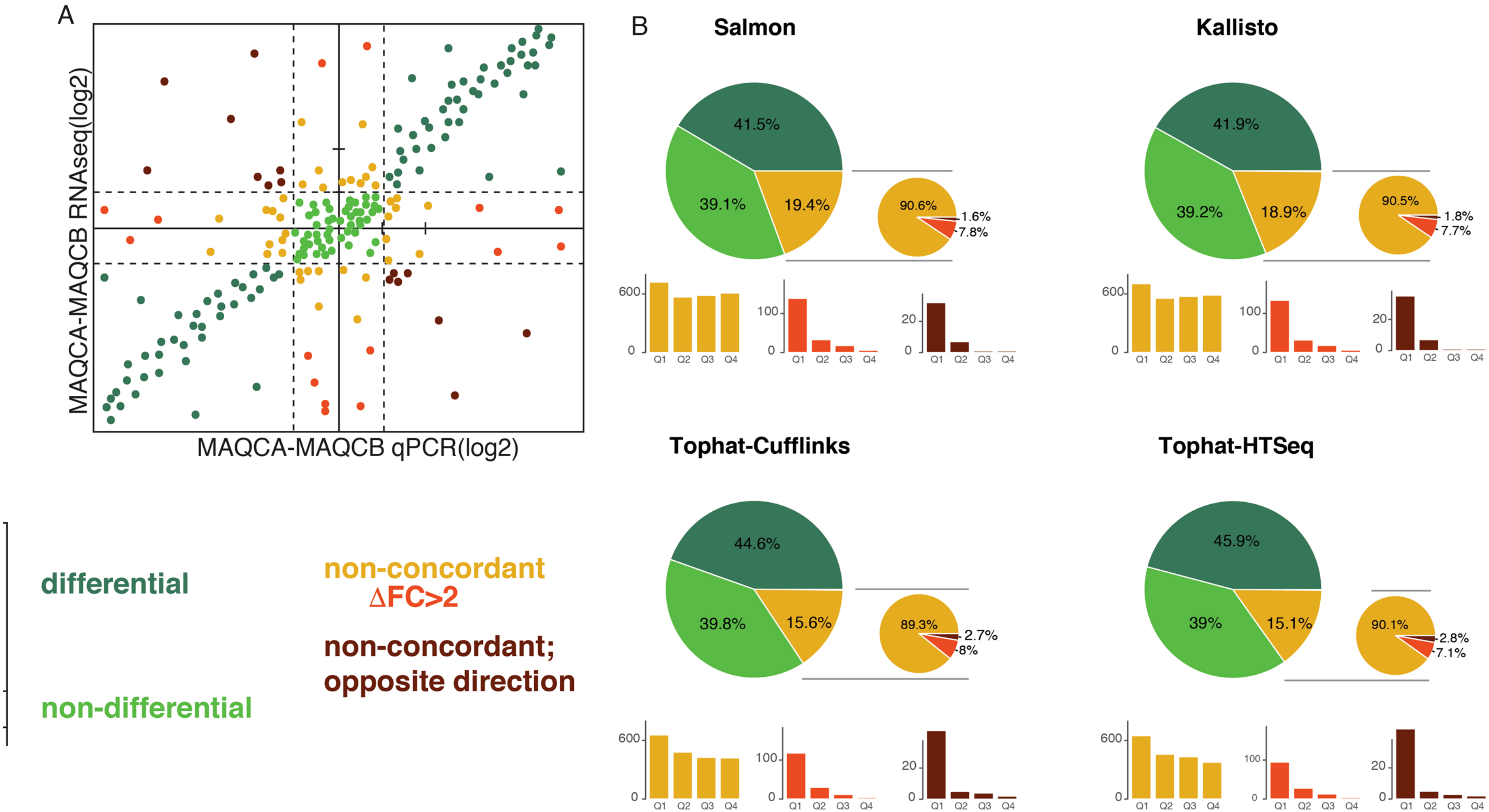
Most problems are consistent so they disappear when you do diff-exp analysis



Toy example of differences between to methods that can arise



Non-concordant results are often found in lowly expressed genes



Transcript level analysis

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

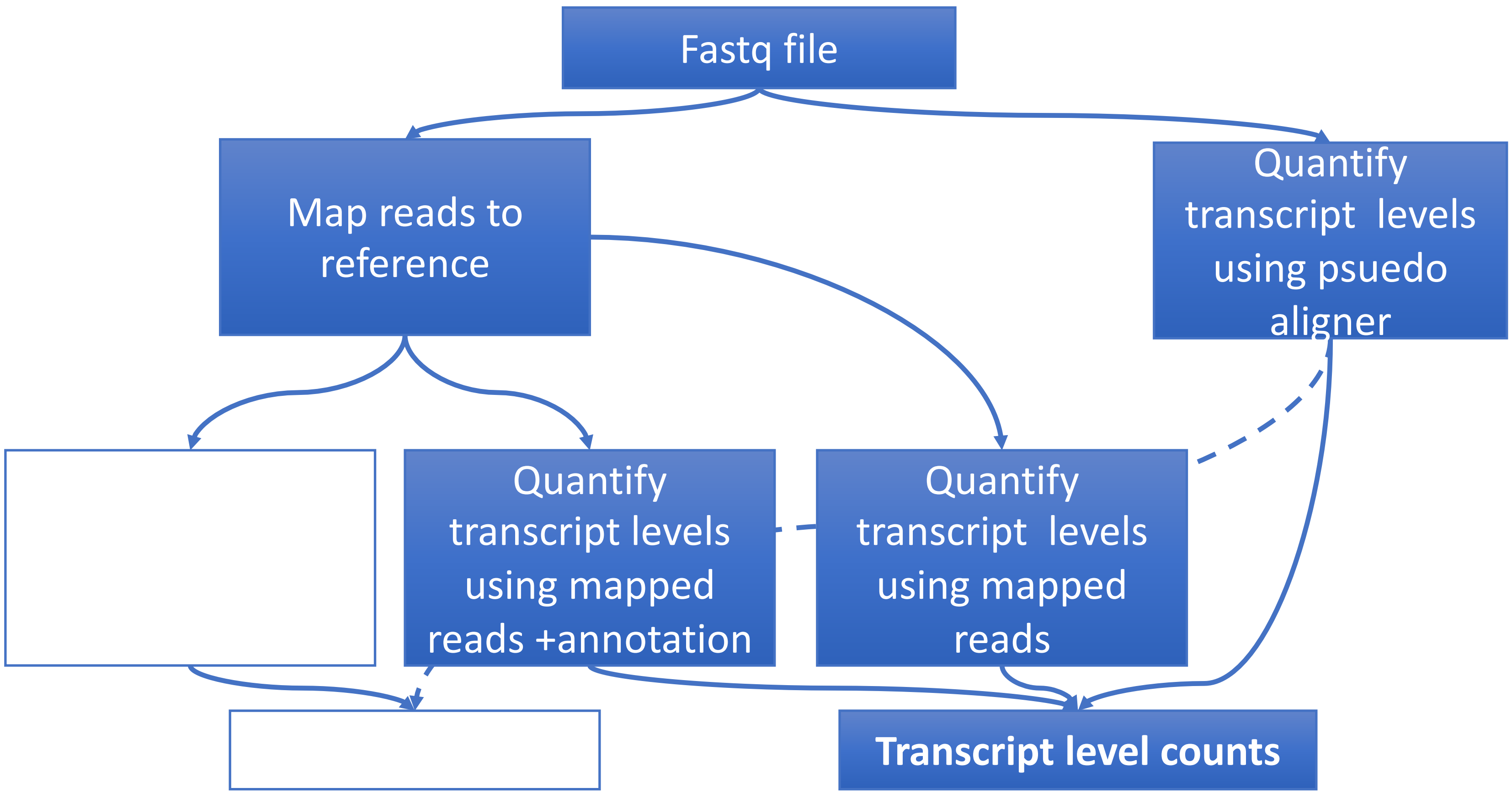
BMC Genomics

RESEARCH ARTICLE Open Access

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

Chi Zhang[1], Baohong Zhang[1], Lih-Ling Lin[2] and Shanrong Zhao[1*]

Transcript level analysis



Methods used in paper

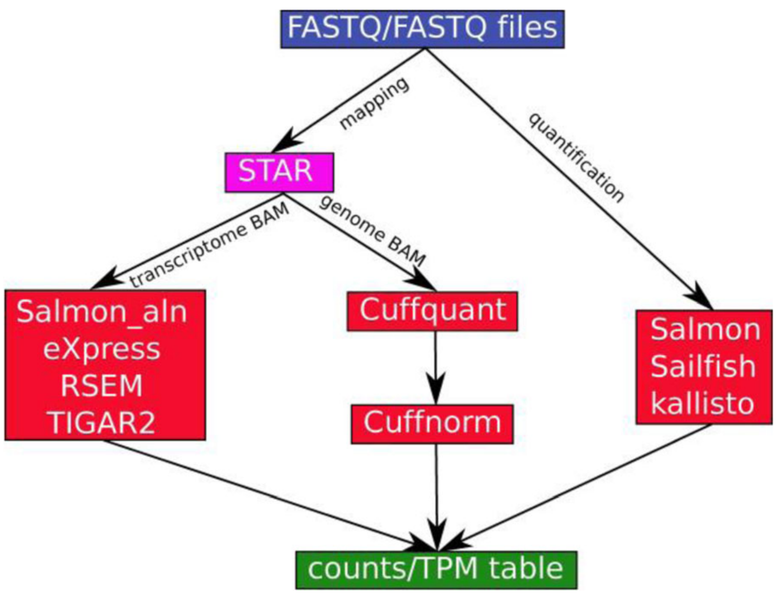


Table 1 Run time metrics of each method on 50 million paired-end reads of length 76 bp in an high performance computing cluster

	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

Isoform quantification problematic for genes with many isoforms

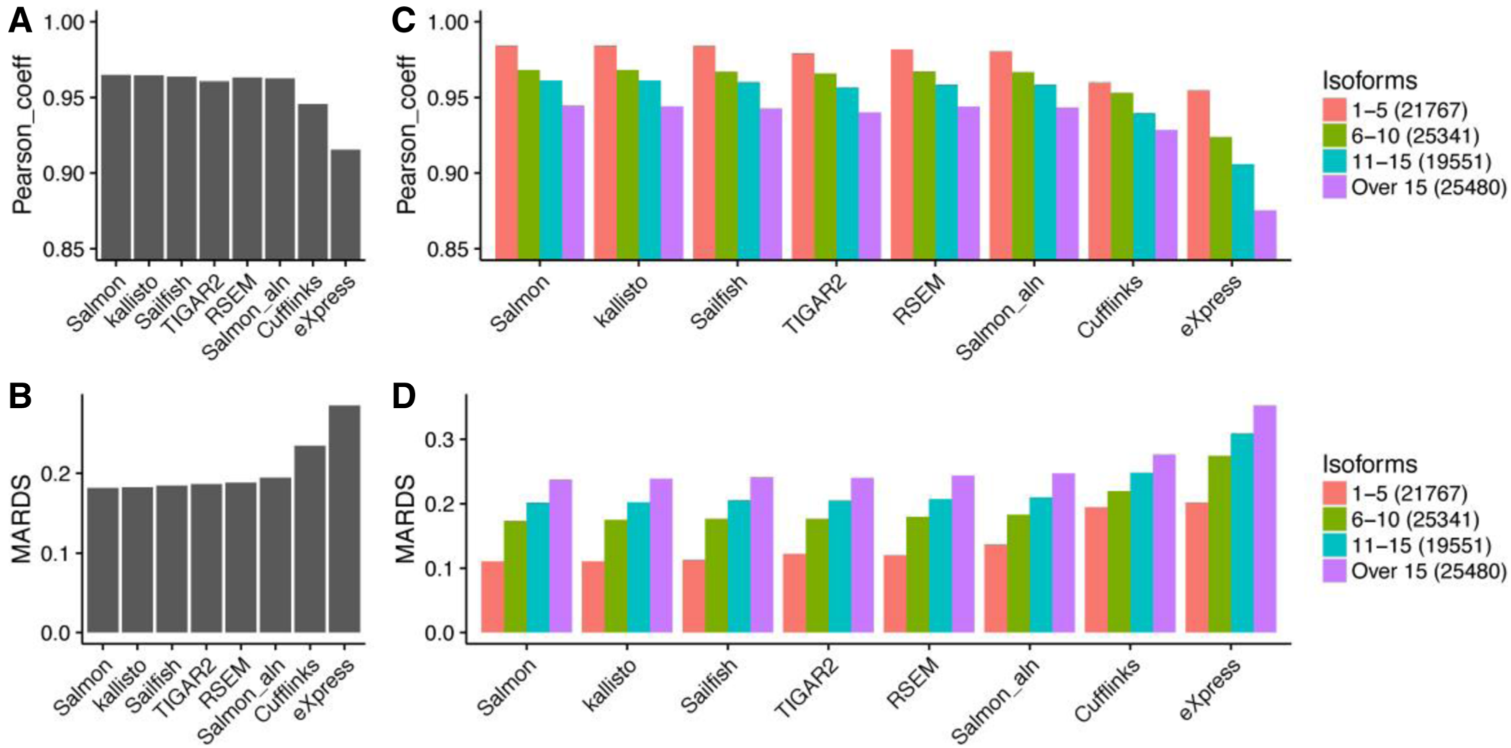


Fig. 2 Comparisons of the overall performance among different methods and the impact of the number of transcripts on the accuracy of isoform quantification. **a** Pearson correlation coefficient. **b** mean absolute relative differences and **c-d**) The above metrics were broken into separate groups according to the number of annotated transcript isoforms for each gene. The number of transcripts in each group is shown in figure legends. The accuracy metrics were calculated by comparing the estimated counts with the "ground truths" in simulated dataset

Results are very similar between methods

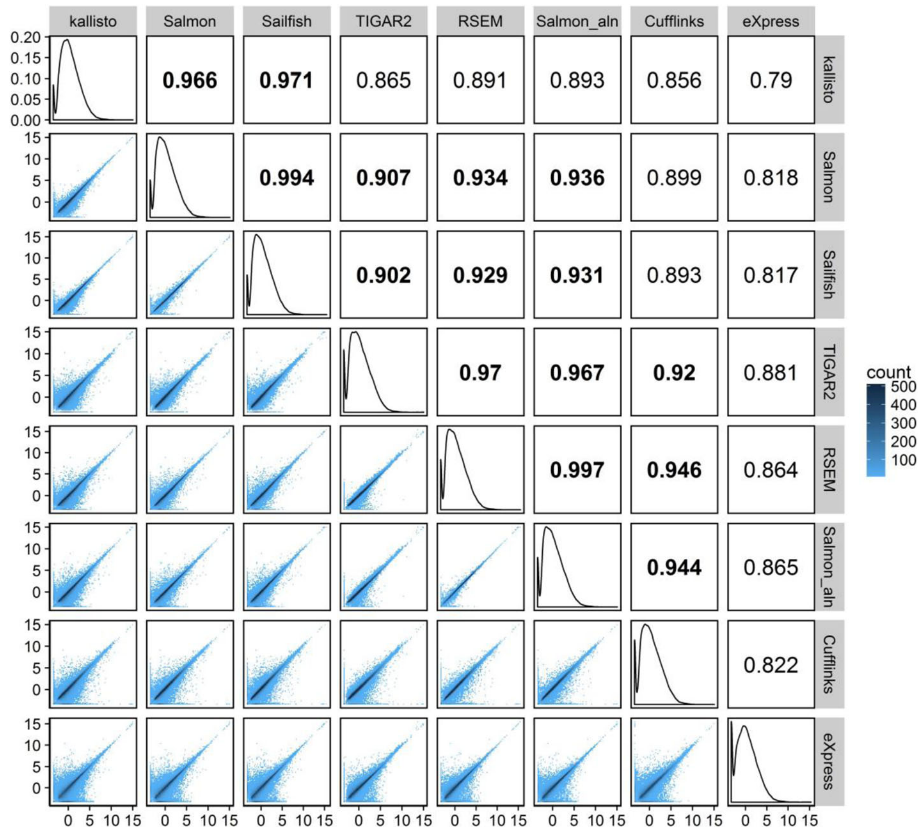


Fig. 5 Pairwise correlation of estimated TPM values for all transcripts between methods for the HBRR-C4 sample. The distribution of transcripts' TPMs from each method was plotted on the diagonal panels. Pairwise density plots and R^2 values are shown in the lower and upper triangular panels, respectively.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

Thank you. Questions?

Johan Reimegård | 13-May-2019