Introduction to read alignment pipelines and gene expression estimates

The Counts They Are a-Changin'

Johan Reimegård

Enabler for Life Sciences

Karolinska Institutet

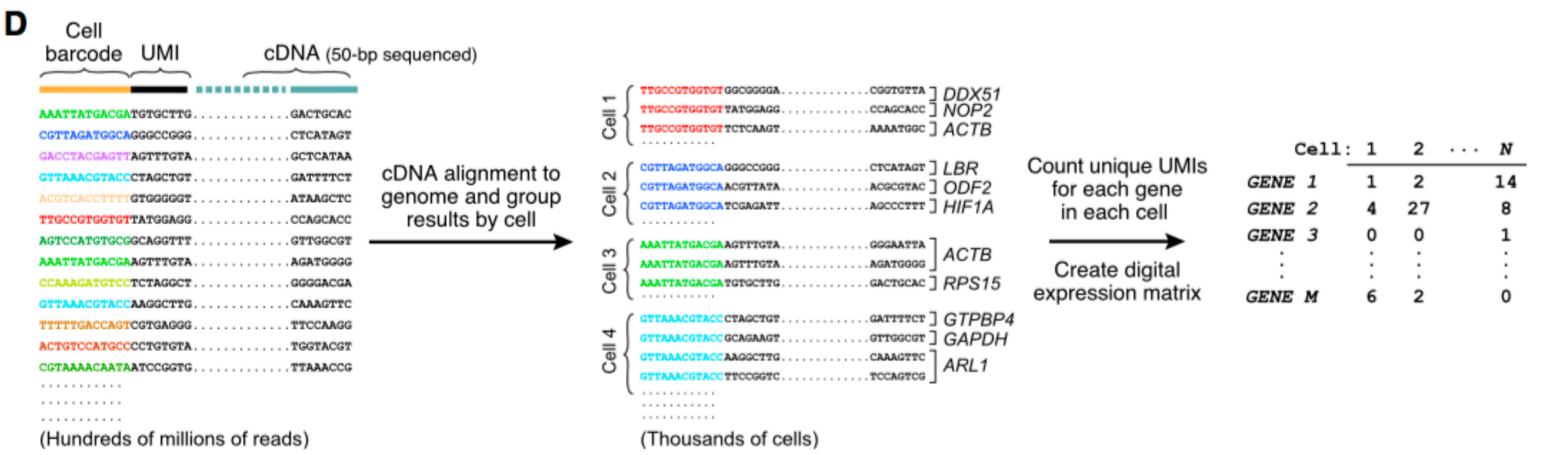
KTH Royal Institute of Technology

Stockholm University

Uppsala Universitet

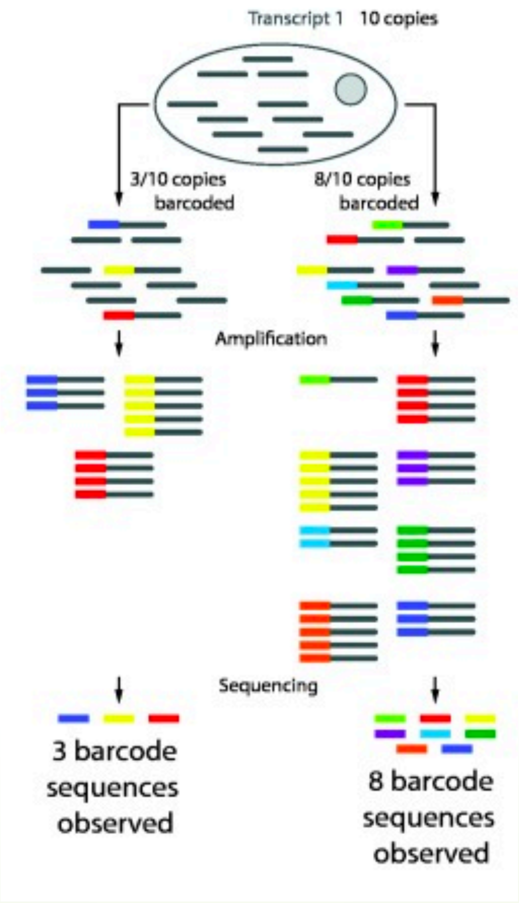
SciLifeLab

How to get from hundred of million of reads to a count table


D
Cell barcode
UMI
cDNA (50-bp sequenced)
AAATTATGACGATGTGCTTG..............GACTGCAC
CGTTAGATGGCAGGGCCGGG..............CTCATAGT
GACCTACGAGTTAGTTTGTA..............GCTCATAA
GTTAAACGTACCCTAGCTGT..............GATTTTCT
ACGTCACCTTTTGTGGGGGT..............ATAAGCTC
TTGCCGTGGTGTTATGGAGG..............CCAGCACC
AGTCCATGTGCGGCAGGTTT..............GTTGGCGT
AAATTATGACGAAGTTTGTA..............AGATGGGG
CCAAAGATGTCCTCTAGGCT..............GGGGACGA
GTTAAACGTACCAAGGCTTG..............CAAAGTTC
TTTTTGACCAGTCGTGAGGG..............TTCCAAGG
ACTGTCCATGCCCCTGTGTA..............TGGTACGT
CGTAAAACAATAATCCGGTG..............TTAAACCG
....
....
....
(Hundreds of millions of reads)
cDNA alignment to genome and group results by cell
Cell 1
TTGCCGTGGTGTGGCGGGGA..............CGGTGTTA
DDX51
TTGCCGTGGTGTTATGGAGG..............CCAGCACC
NOP2
TTGCCGTGGTGTTCTCAAGT..............AAAATGGC
ACTB
....
Cell 2
CGTTAGATGGCAGGGCCGGG..............CTCATAGT
LBR
CGTTAGATGGCAACGTTATA..............ACGCGTAC
ODF2
CGTTAGATGGCATCGAGATT..............AGCCCTTT
HIF1A
....
Cell 3
AAATTATGACGAAGTTTGTA..............GGGAATTA
AAATTATGACGAAGTTTGTA..............AGATGGGG
ACTB
AAATTATGACGATGTGCTTG..............GACTGCAC
RPS15
....
Cell 4
GTTAAACGTACCCTAGCTGT..............GATTTTCT
GTPBP4
GTTAAACGTACCGCAGAAGT..............GTTGGCGT
GAPDH
GTTAAACGTACCAAGGCTTG..............CAAAGTTC
GTTAAACGTACCTTCCGGTC..............TCCAGTCG
ARL1
....
....
....
(Thousands of cells)
Count unique UMIs for each gene in each cell
Create digital expression matrix

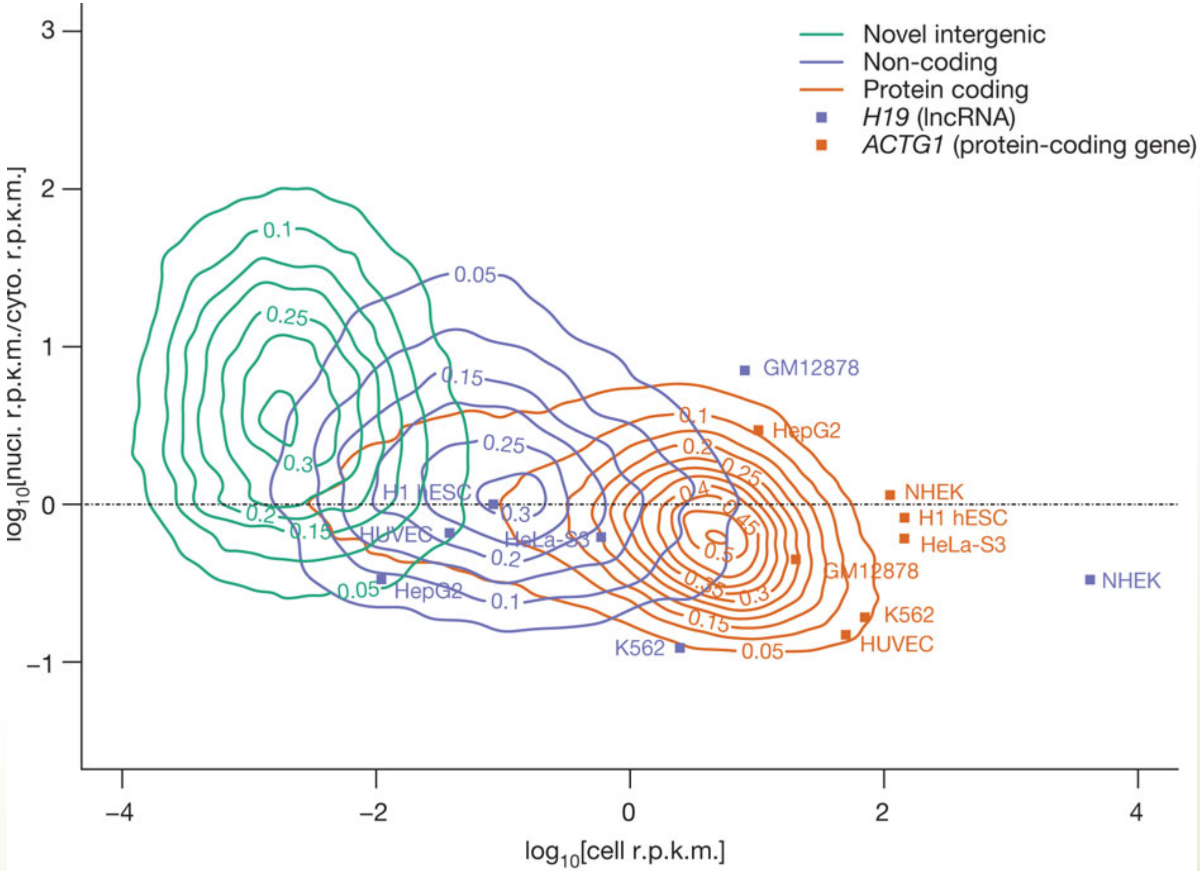
Cell:	1	2	···	N
GENE 1	1	2		14
GENE 2	4	27		8
GENE 3	0	0		1
⋮	⋮	⋮		⋮
GENE M	6	2		0

UMI (Unique molecular identifiers) will make sure that one fragment is counted as one read



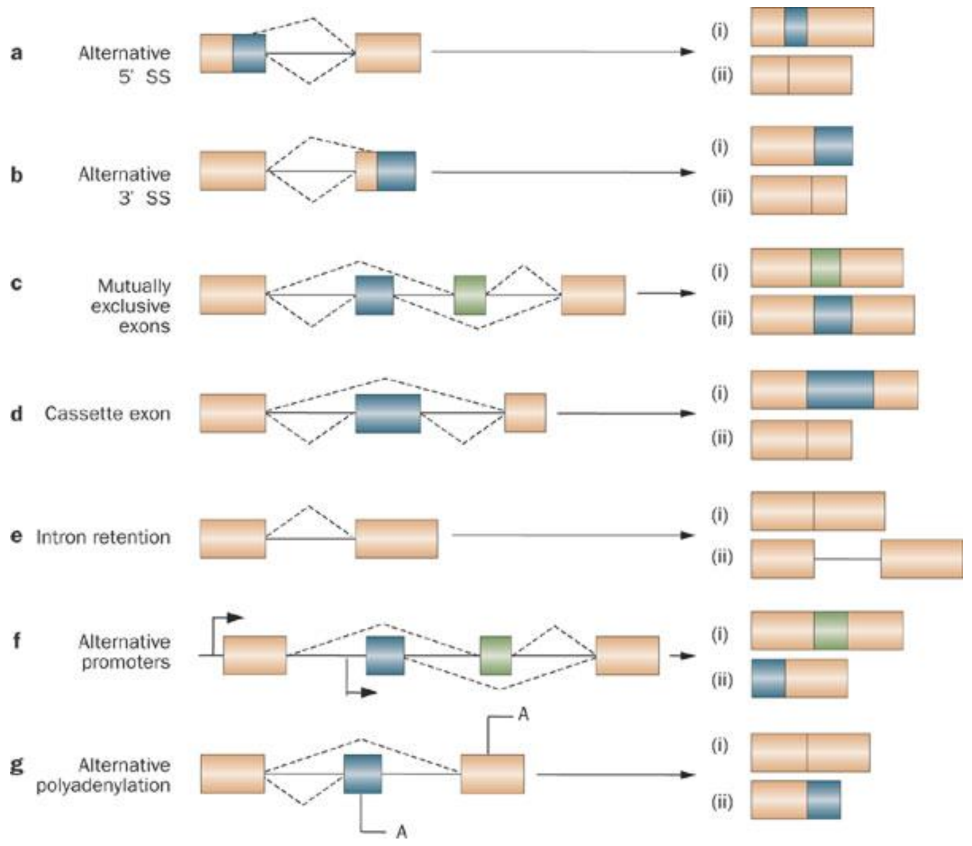
- Will remove errors that occur during the amplification step.
- Will not handle sampling bias

Different kind of RNAs have different expression values

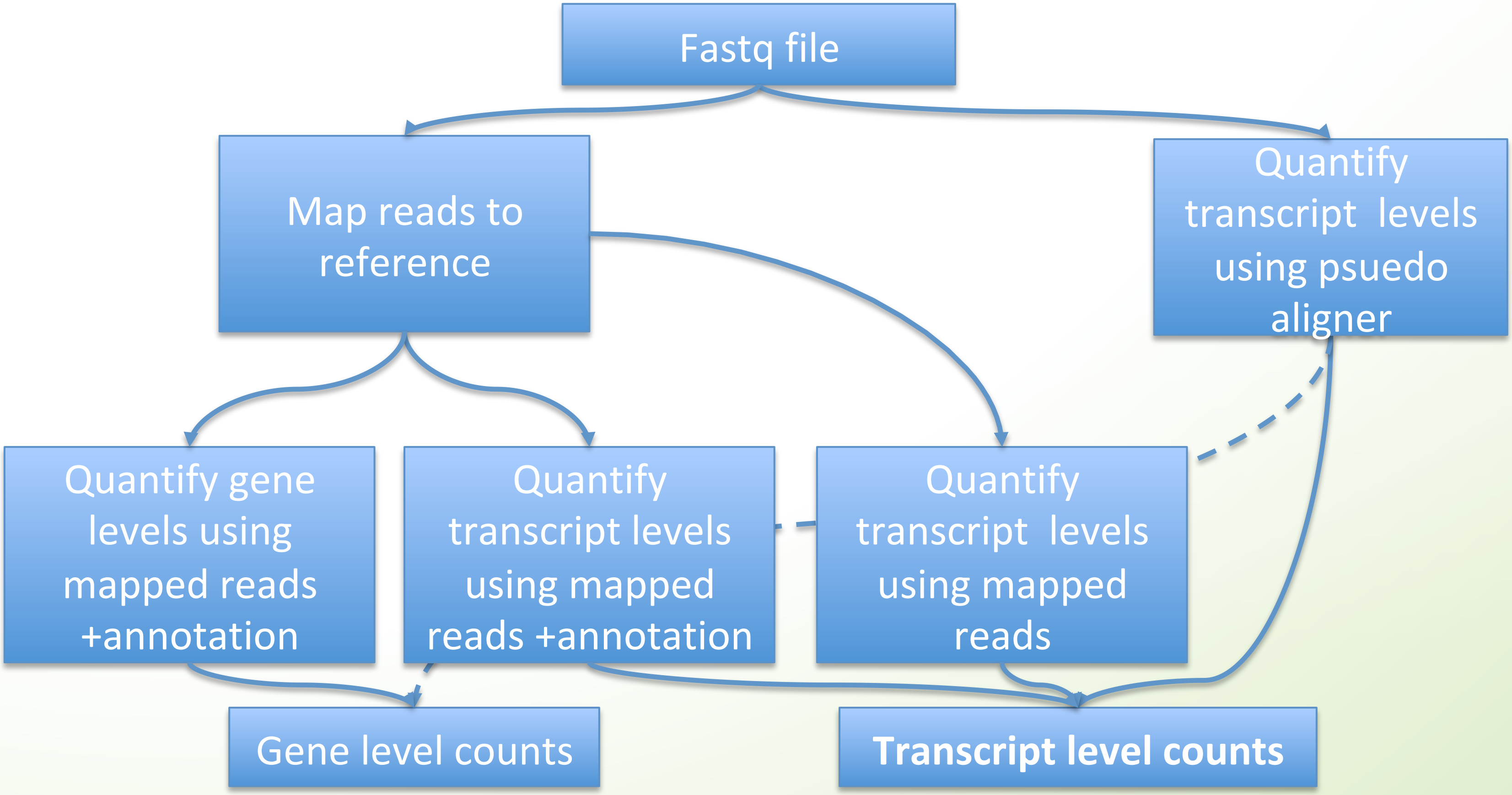


Landscape of transcription in human cells, S Djebali *et al.* ***Nature 2012***

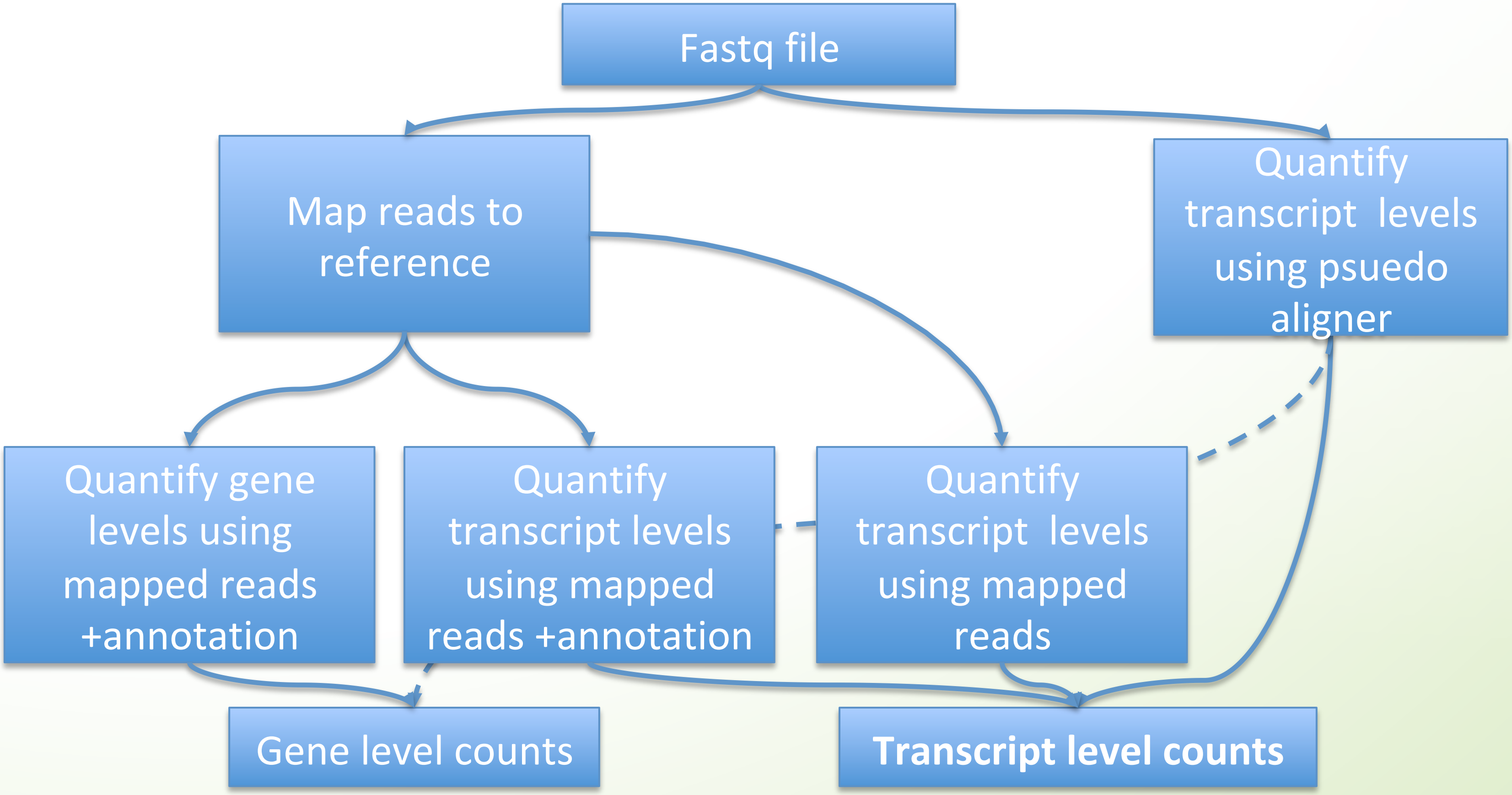
One gene many transcripts



Different paths to get a count table

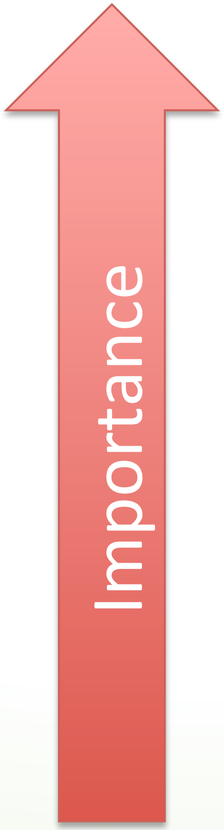

Fastq file
Map reads to reference
Quantify transcript levels using psuedo aligner
Quantify gene levels using mapped reads +annotation
Quantify transcript levels using mapped reads +annotation
Quantify transcript levels using mapped reads
Gene level counts
Transcript level counts

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


Fastq file
Map reads to reference
Quantify transcript levels using psuedo aligner
Quantify gene levels using mapped reads +annotation
Quantify transcript levels using mapped reads +annotation
Quantify transcript levels using mapped reads
Gene level counts
Transcript level counts

How important is mapping accuracy?

Depends what you want to do:



Identify novel genetic variants or RNA editing

Allele-specific expression

Genome annotation

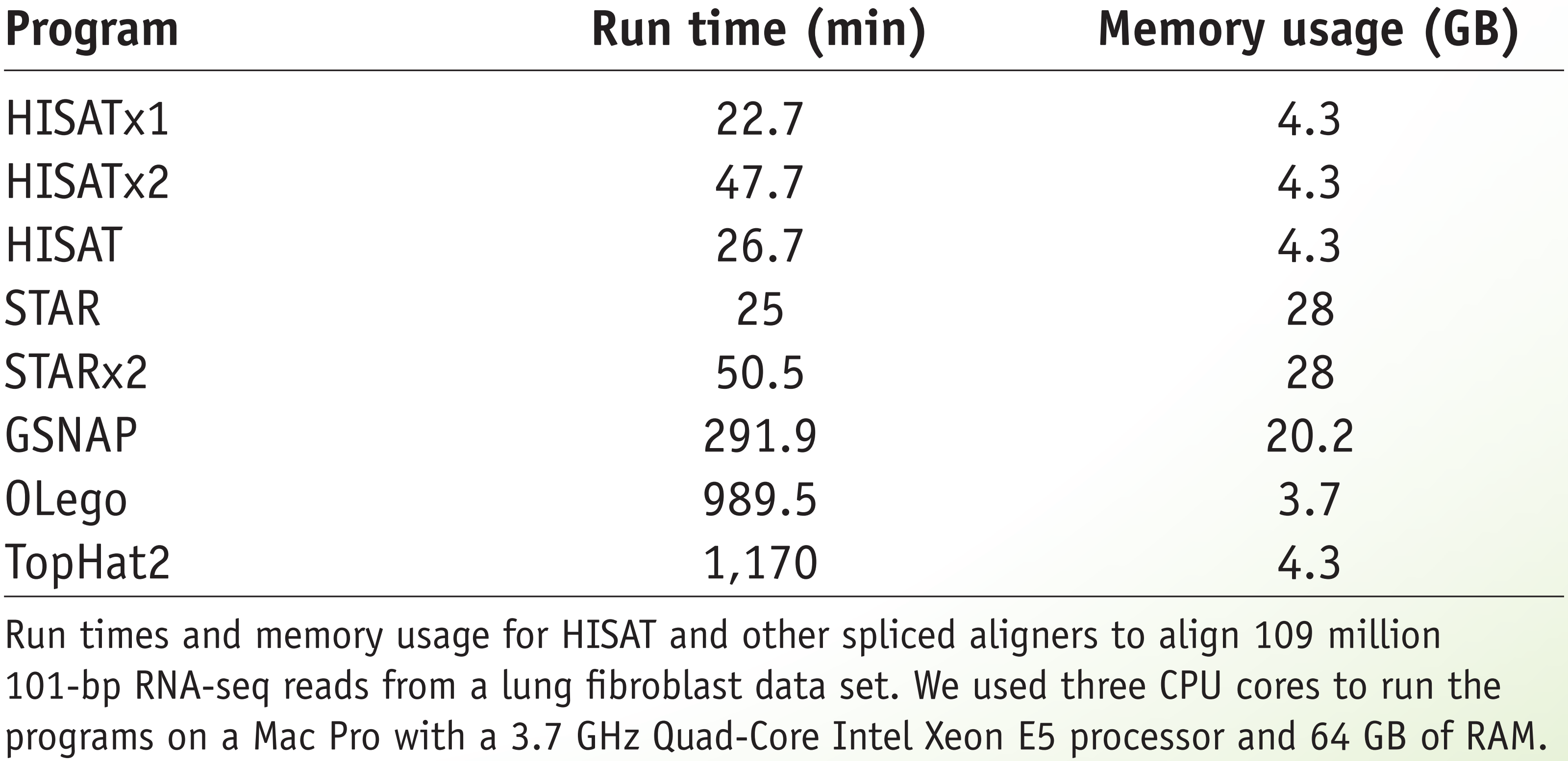
Gene and transcript discovery

Differential expression

Current RNA-seq aligners

TopHat2	Kim et al. *Genome Biology* 2013
HISAT2	Kim et al. *Nature Methods* 2015
STAR	Dobin et al. *Bioinformatics* 2013
GSNAP	Wu and Nacu *Bioinformatics* 2010
OLego	Wu et al. *Nucleic Acids Research* 2013
HPG aligner	Medina et al. *DNA Research* 2016
MapSplice2	http://www.netlab.uky.edu/p/bioinfo/MapSplice2

Compute requirements



Program	Run time (min)	Memory usage (GB)
HISATx1	22.7	4.3
HISATx2	47.7	4.3
HISAT	26.7	4.3
STAR	25	28
STARx2	50.5	28
GSNAP	291.9	20.2
OLego	989.5	3.7
TopHat2	1,170	4.3

Run times and memory usage for HISAT and other spliced aligners to align 109 million 101-bp RNA-seq reads from a lung fibroblast data set. We used three CPU cores to run the programs on a Mac Pro with a 3.7 GHz Quad-Core Intel Xeon E5 processor and 64 GB of RAM.

Kim et al. *Nature Methods* 2015

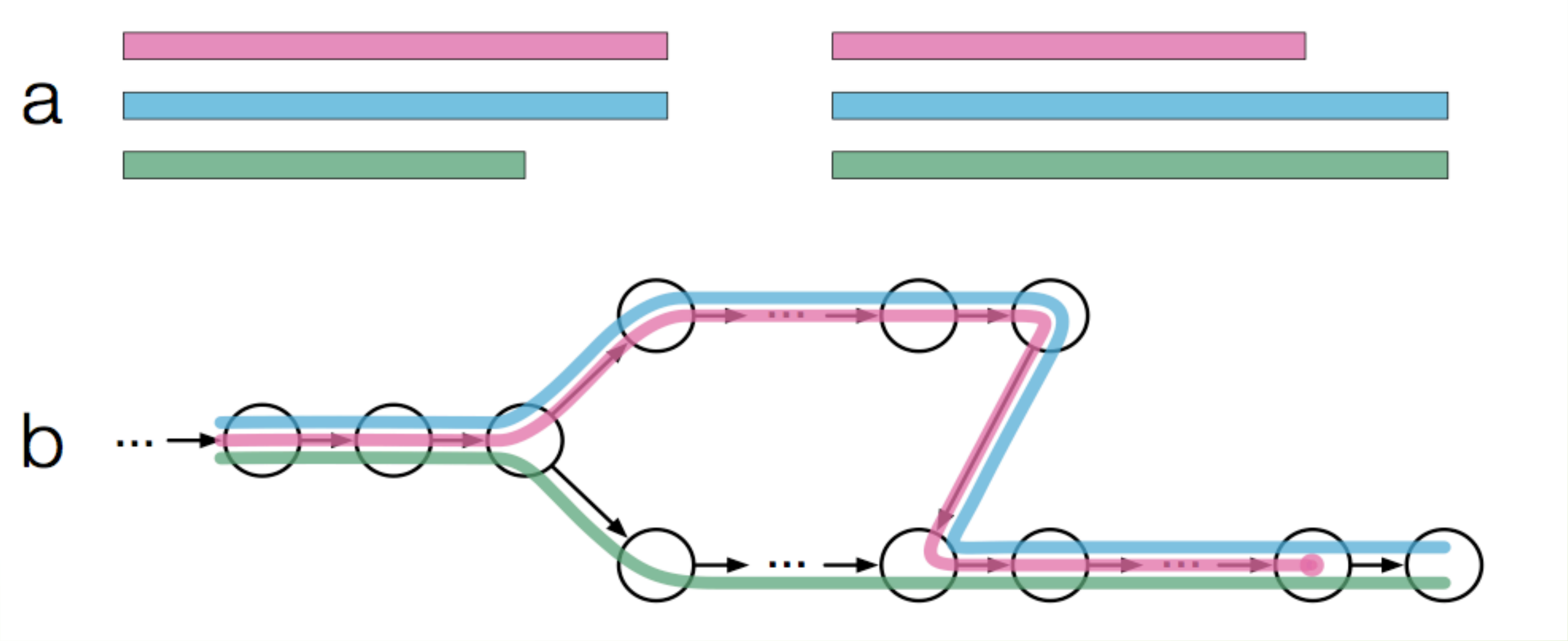
Innovations in RNA-seq alignment software

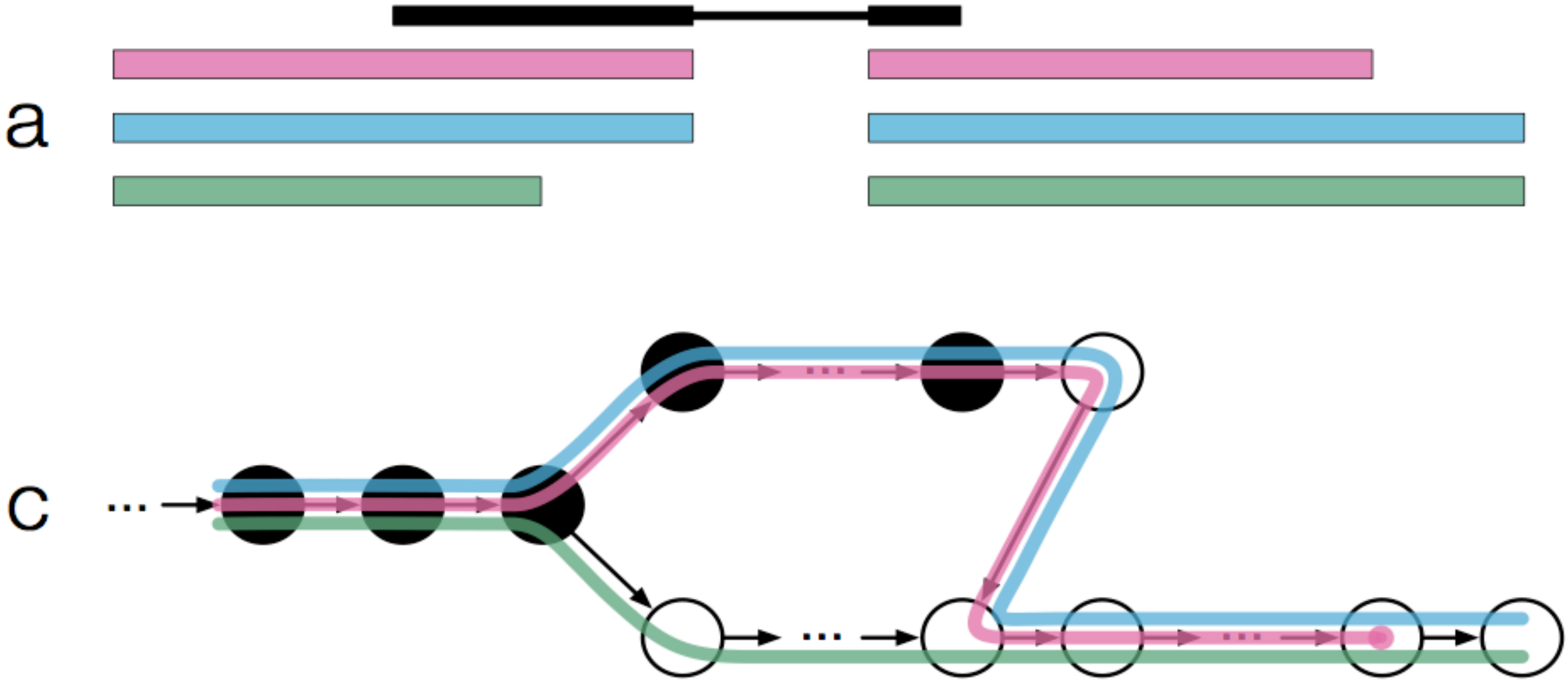
- Read pair alignment
- Consider base call quality scores
- Sophisticated indexing to decrease CPU and memory usage
- Map to genetic variants
- Consider junction annotation
- Two-step approach (junction discovery & final alignment)

Recommendations when using mapping programs

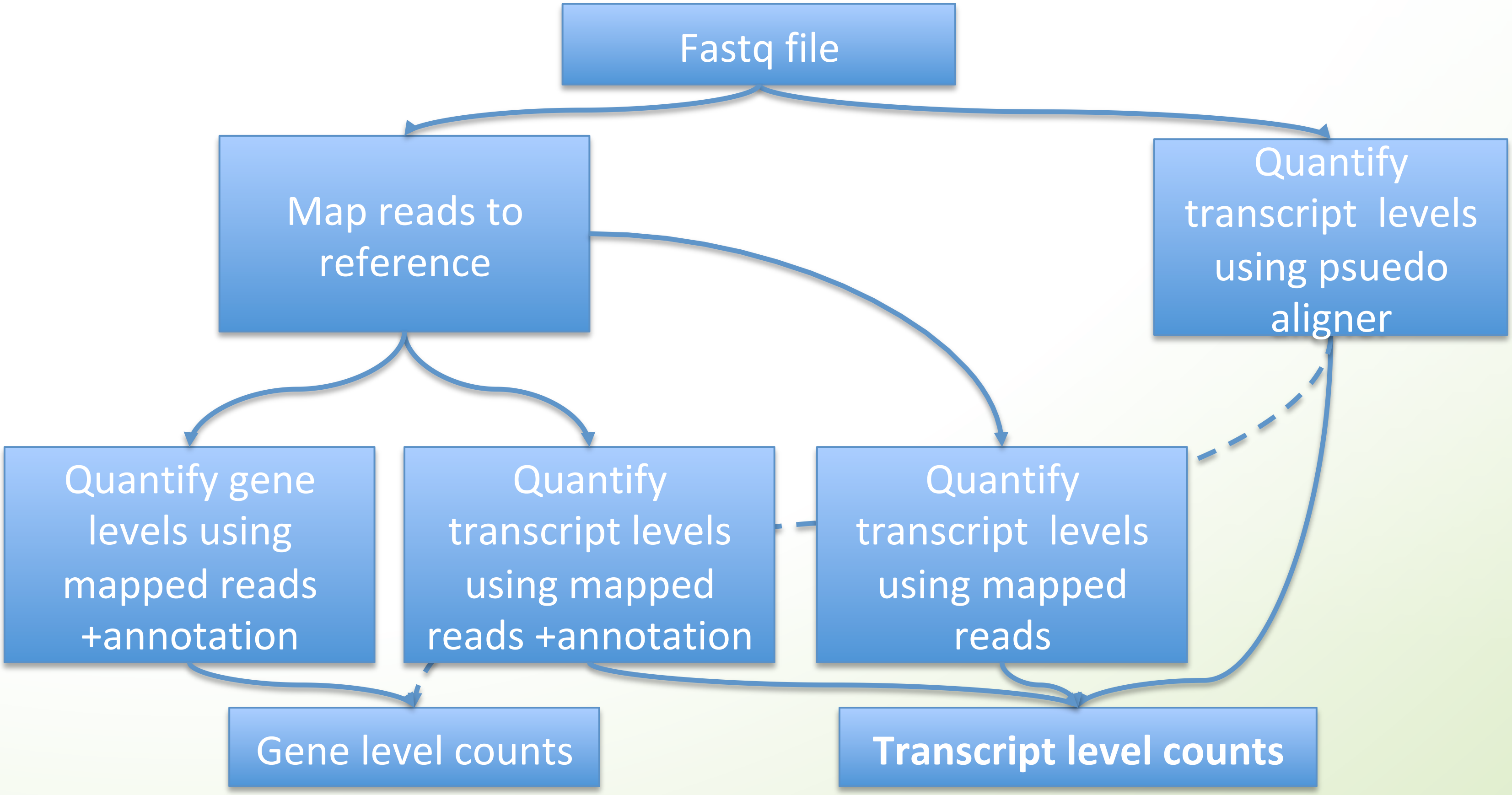
- Use STAR, HISAT2
- STAR and HISAT2 are the fastest
- HISAT2 uses the least memory
- Always check the results!

“Pseudoalignments” in calisto




a
c
...

Different paths to get a count table


Fastq file
Map reads to reference
Quantify transcript levels using psuedo aligner
Quantify gene levels using mapped reads +annotation
Quantify transcript levels using mapped reads +annotation
Quantify transcript levels using mapped reads
Gene level counts
Transcript level counts

Karolinska Institutet

ROYAL INSTITUTE OF TECHNOLOGY

Stockholm University

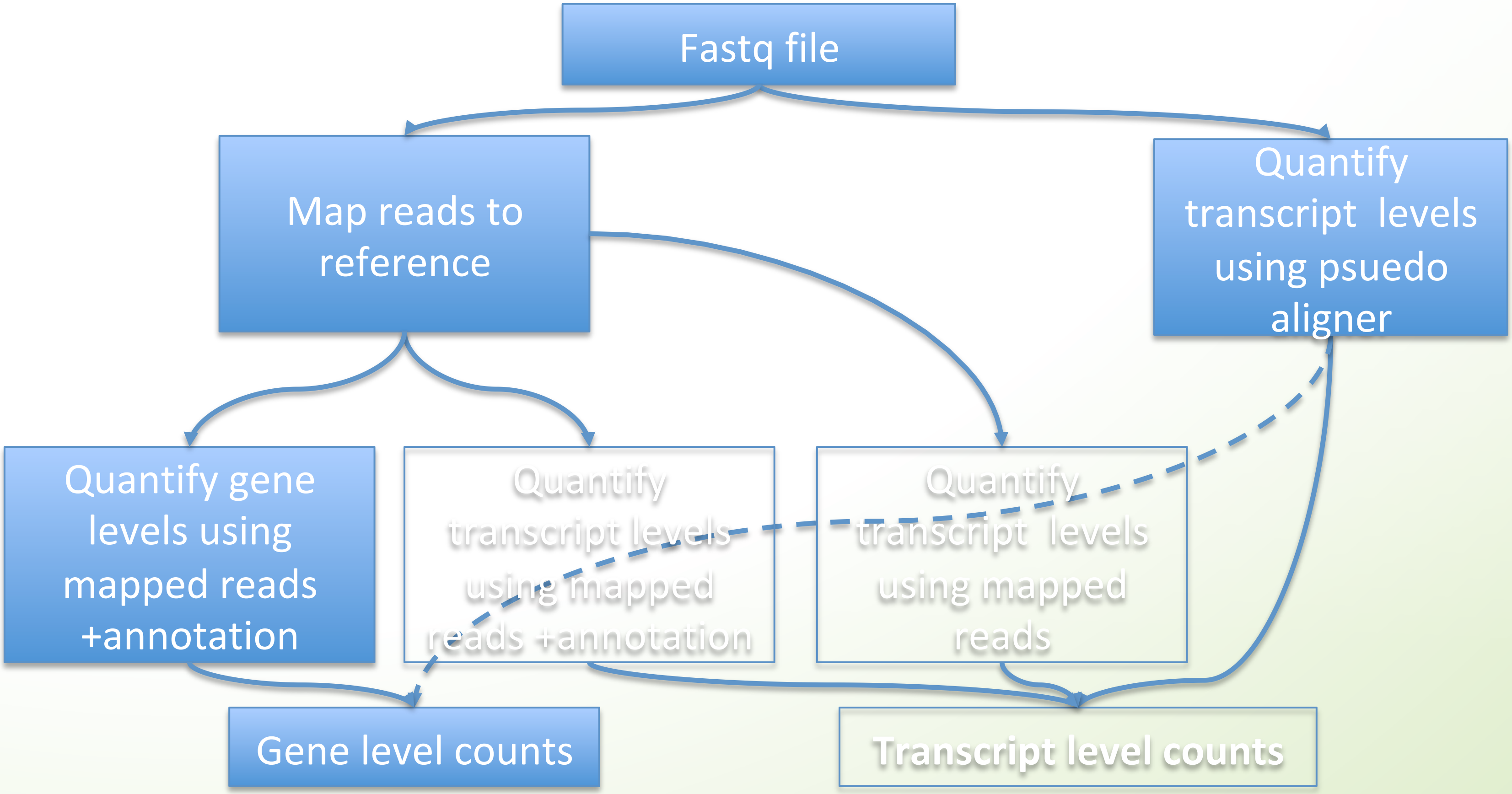
UPPSALA UNIVERSITET

SciLifeLab

Gene expression estimates

- Expression estimates on gene level
- Expression estimates on transcript level

Gene level analysis



Fastq file

Map reads to reference

Quantify transcript levels using psuedo aligner

Quantify gene levels using mapped reads +annotation

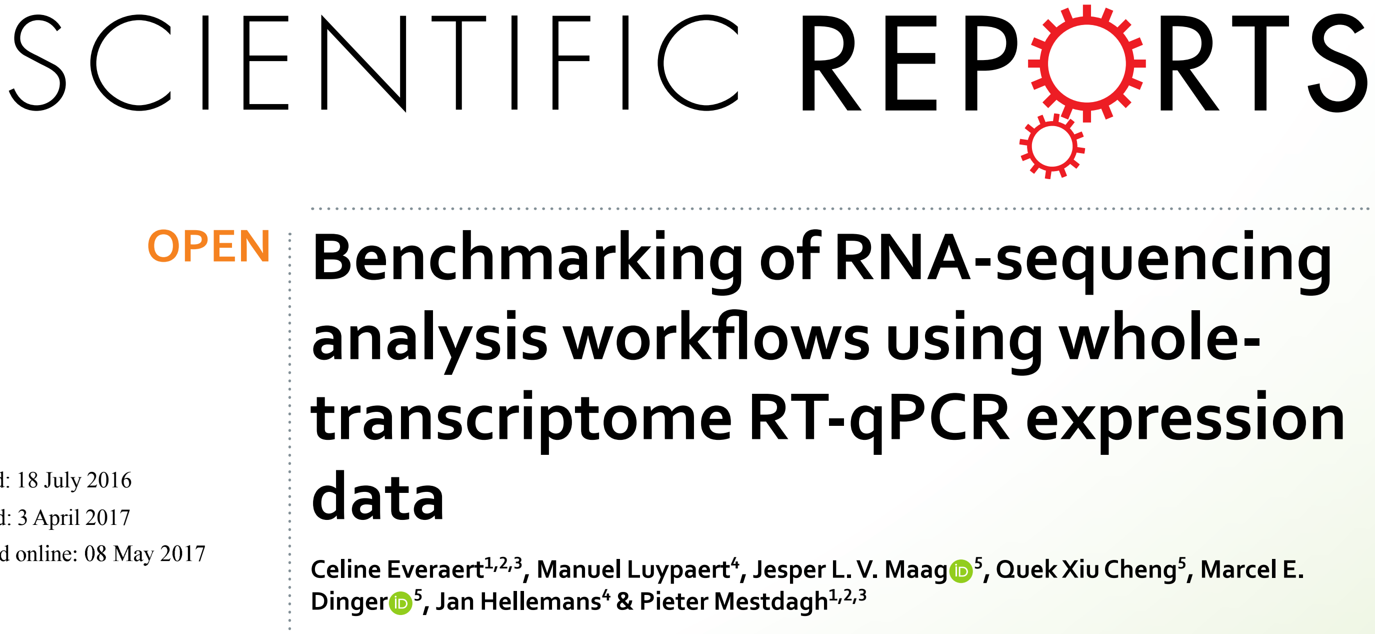
Quantify transcript levels using mapped reads +annotation

Quantify transcript levels using mapped reads

Gene level counts

Transcript level counts

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

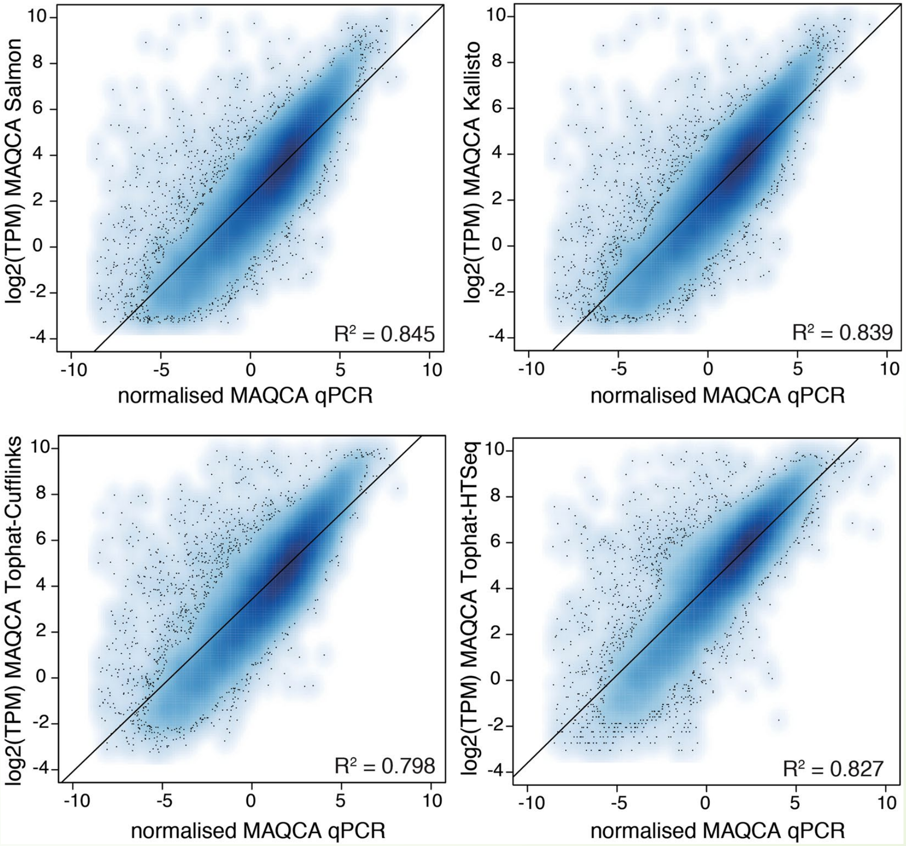
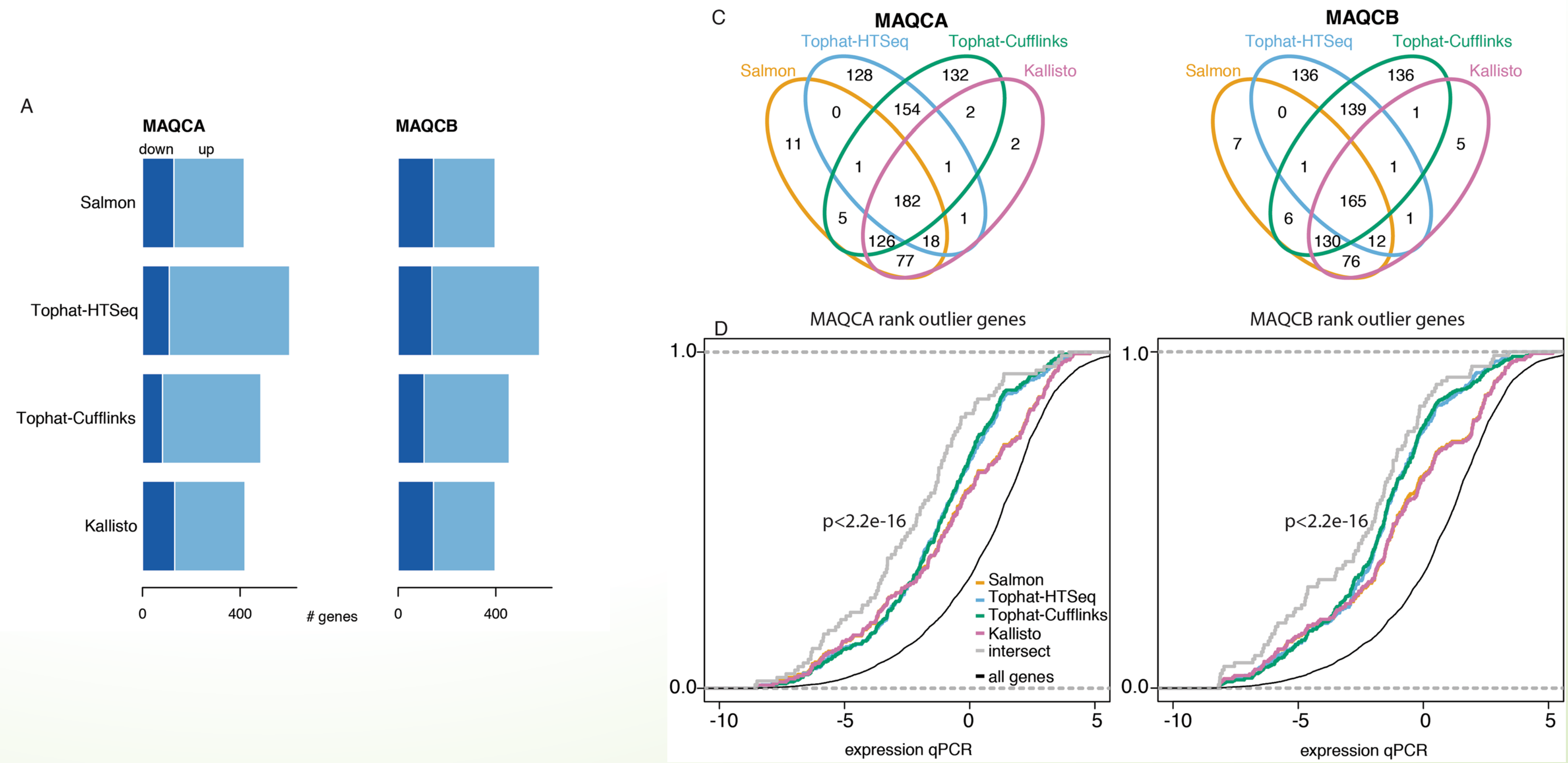
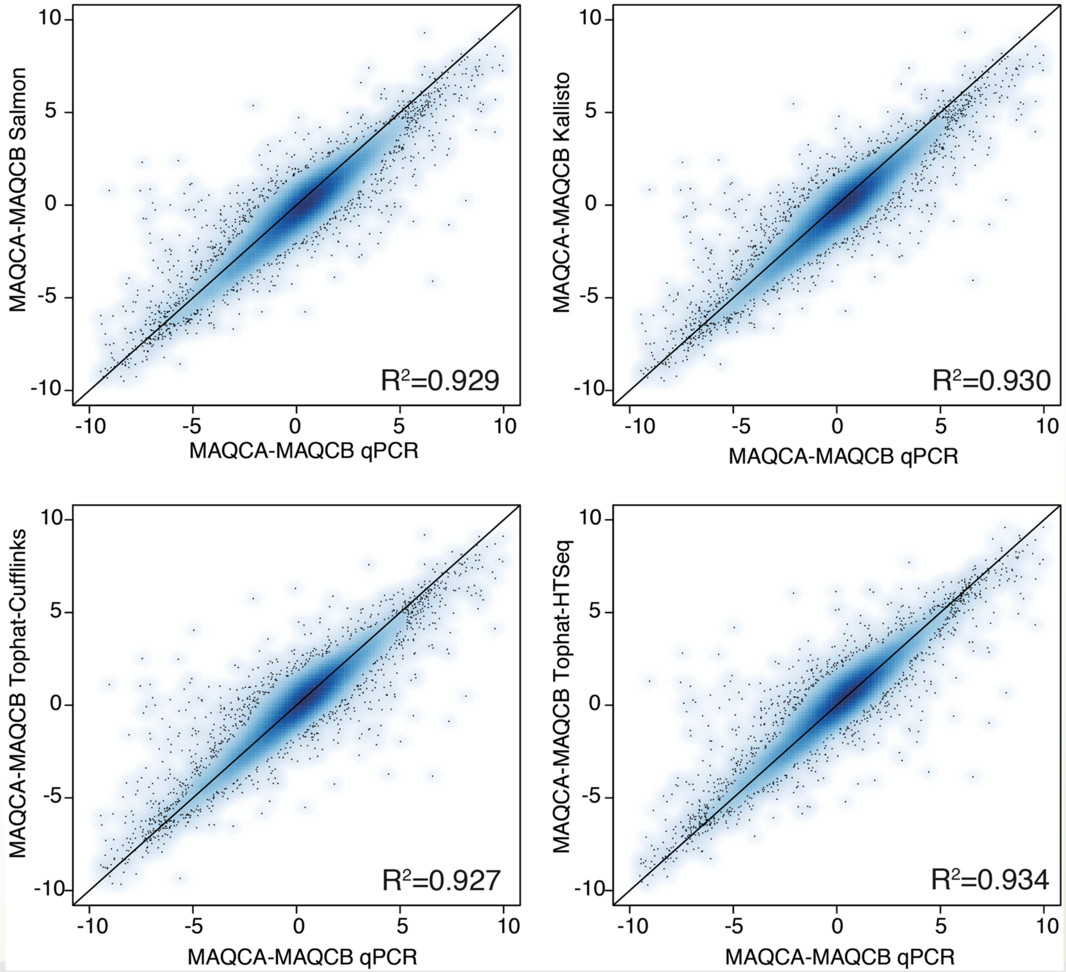


Figure 1. Gene expression correlation between RT-qPCR and RNA-seq data. The Pearson correlation coefficients and linear regression line are indicated. Results are based on RNA-seq data from dataset 1.

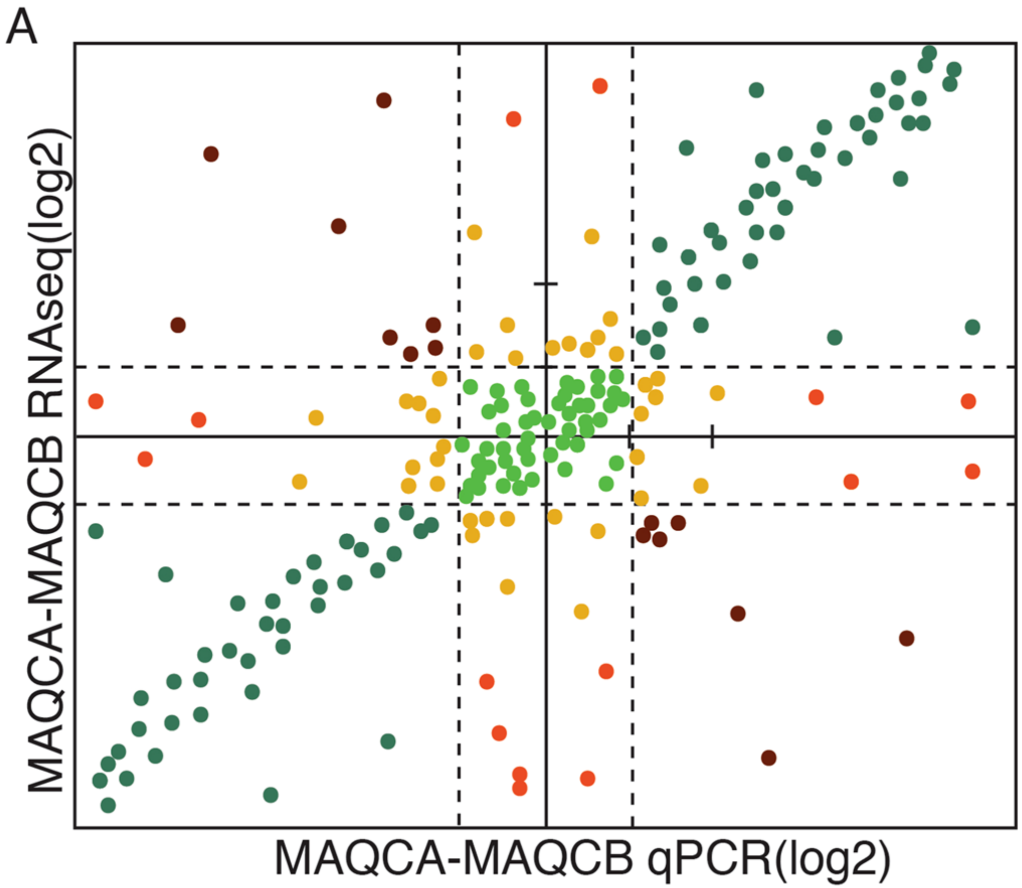
Lowly expressed genes are more problematic to identify using RNA seq



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



Toy example of differences between to methods that can arise



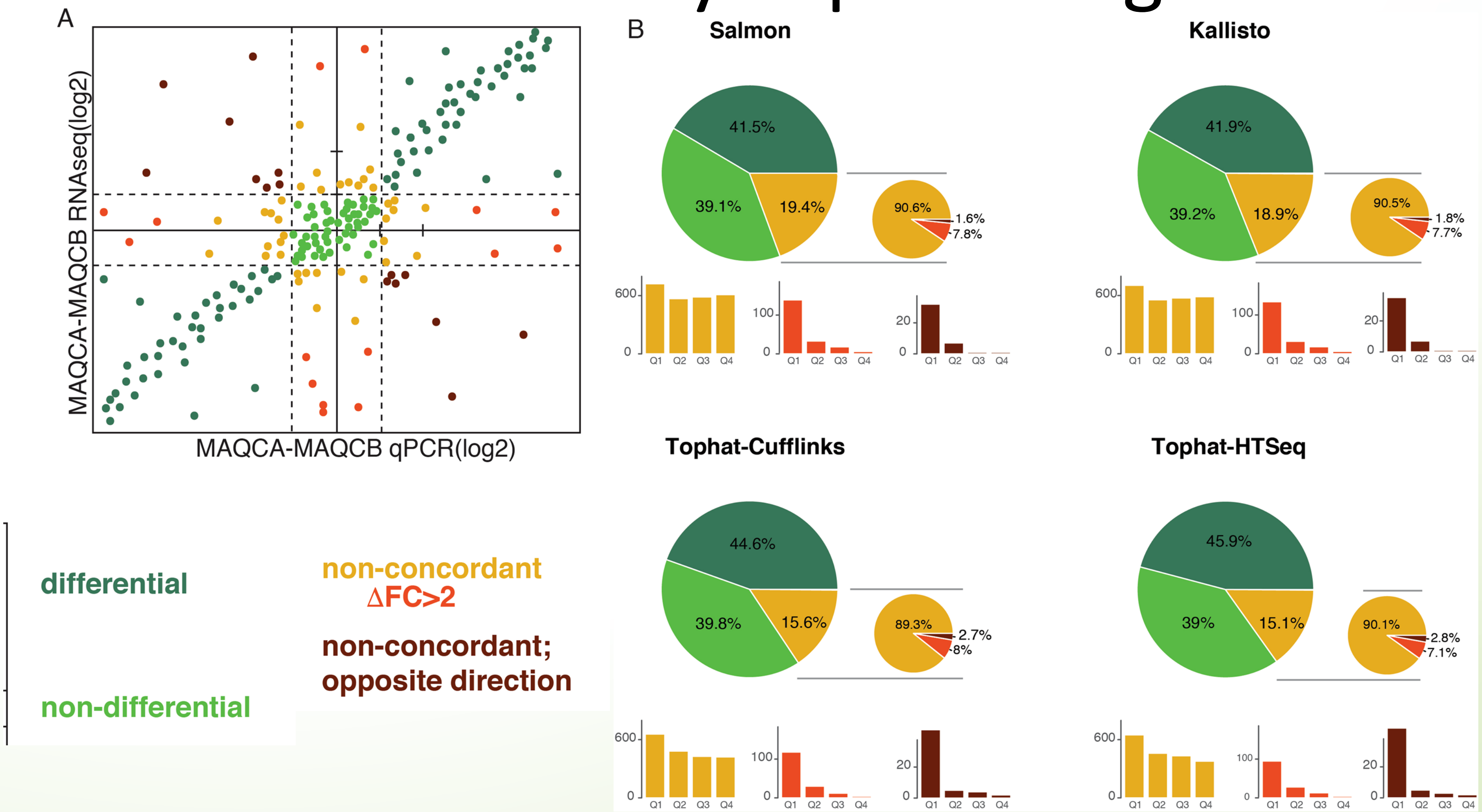
differential

non-differential

non-concordant
ΔFC>2

non-concordant;
opposite direction

Non-concordant results are often found in lowly expressed genes



Karolinska Institutet

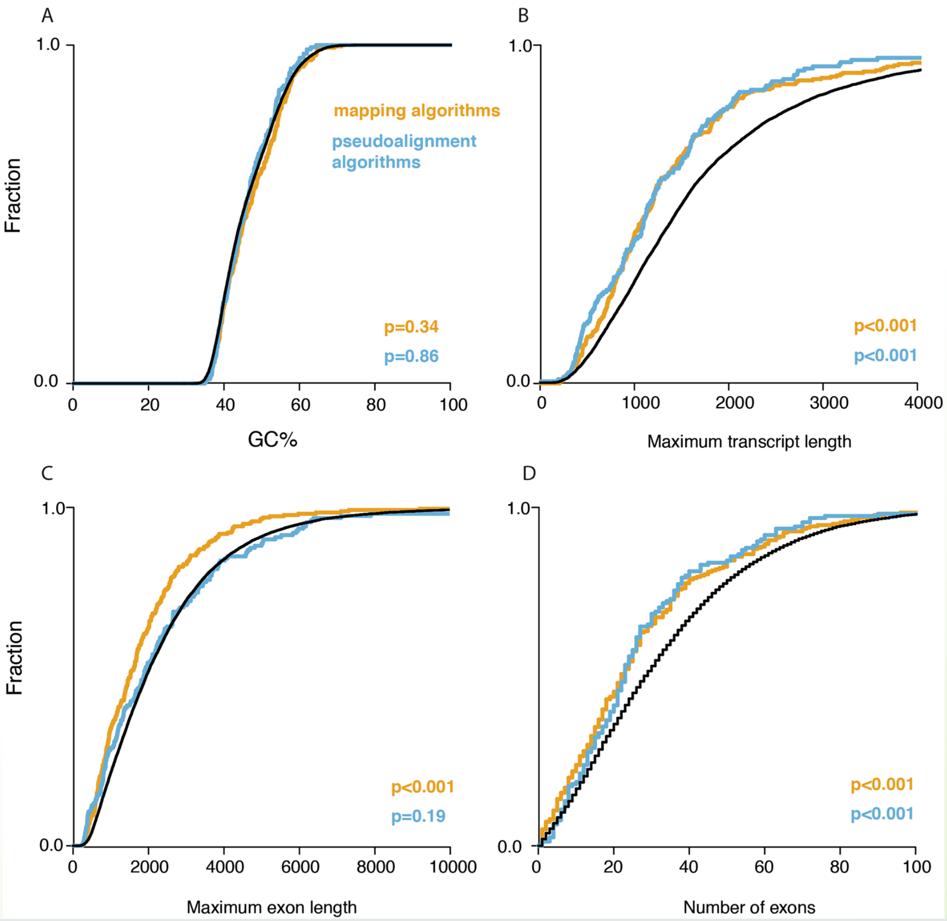
KTH VETENSKAP OCH KONST
ROYAL INSTITUTE OF TECHNOLOGY

Stockholm University

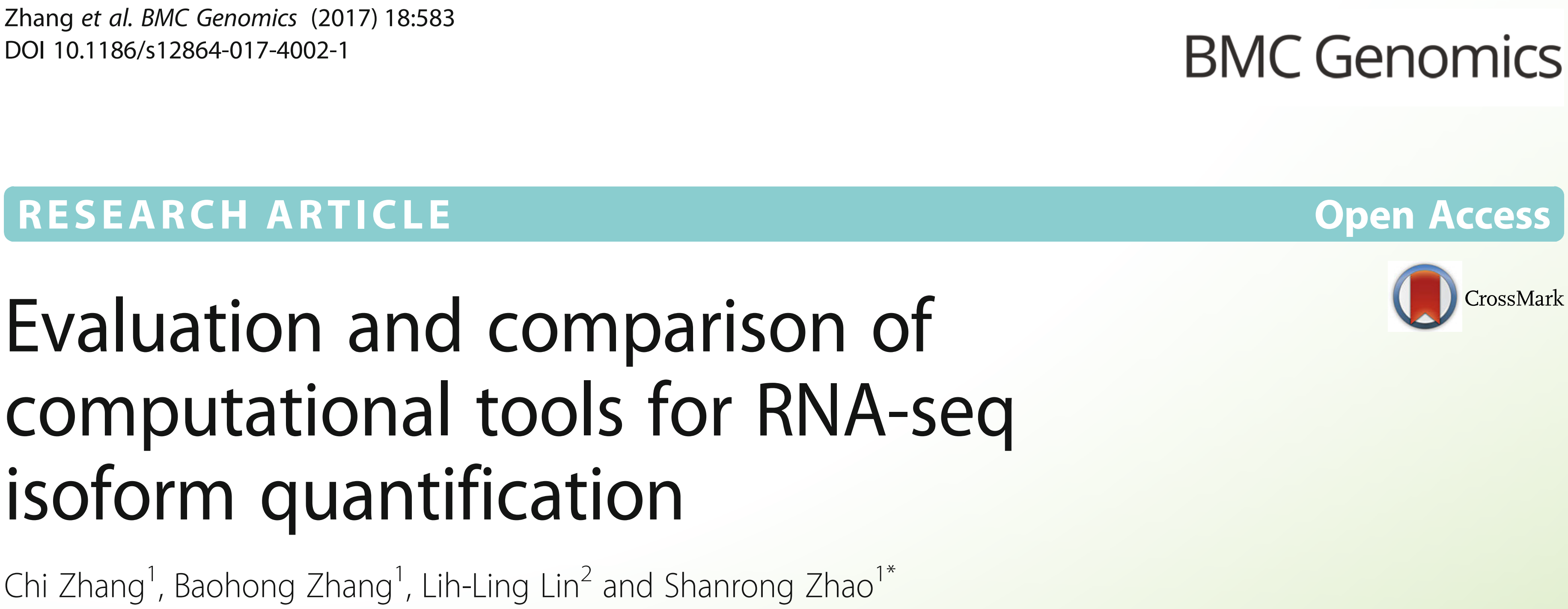
UPPSALA UNIVERSITET

SciLifeLab

Small transcripts are harder to to get correct values for



Transcript level analysis



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

BMC Genomics

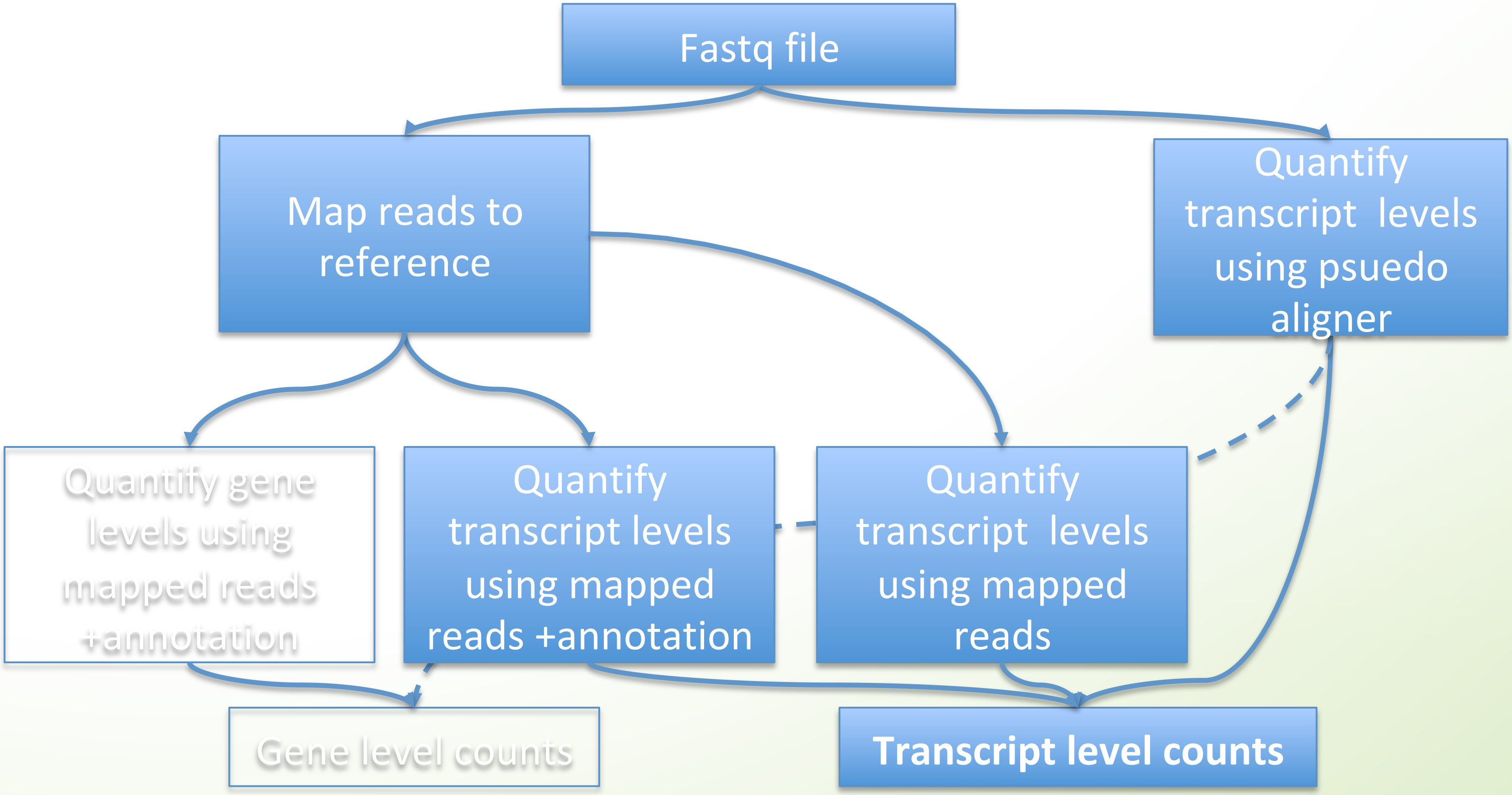
RESEARCH ARTICLE Open Access

CrossMark

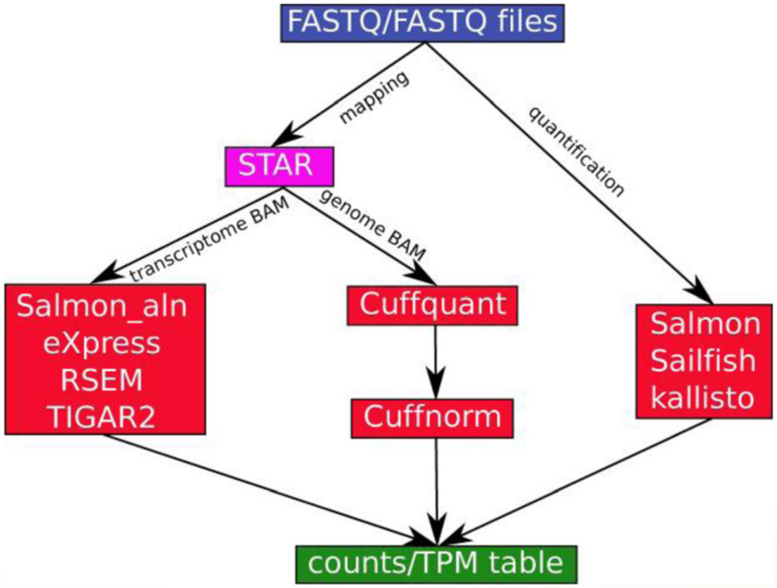
Evaluation and comparison of computational tools for RNA-seq isoform quantification

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

Transcript level analysis


Fastq file
Map reads to reference
Quantify transcript levels using psuedo aligner
Quantify gene levels using mapped reads +annotation
Quantify transcript levels using mapped reads +annotation
Quantify transcript levels using mapped reads
Gene level counts
Transcript level counts

Methods used in paper



FASTQ/FASTQ files

mapping

quantification

STAR

transcriptome BAM

genome BAM

Salmon_aln
eXpress
RSEM
TIGAR2

Cuffquant

Cuffnorm

Salmon
Sailfish
kallisto

counts/TPM table

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	0.55	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	5	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

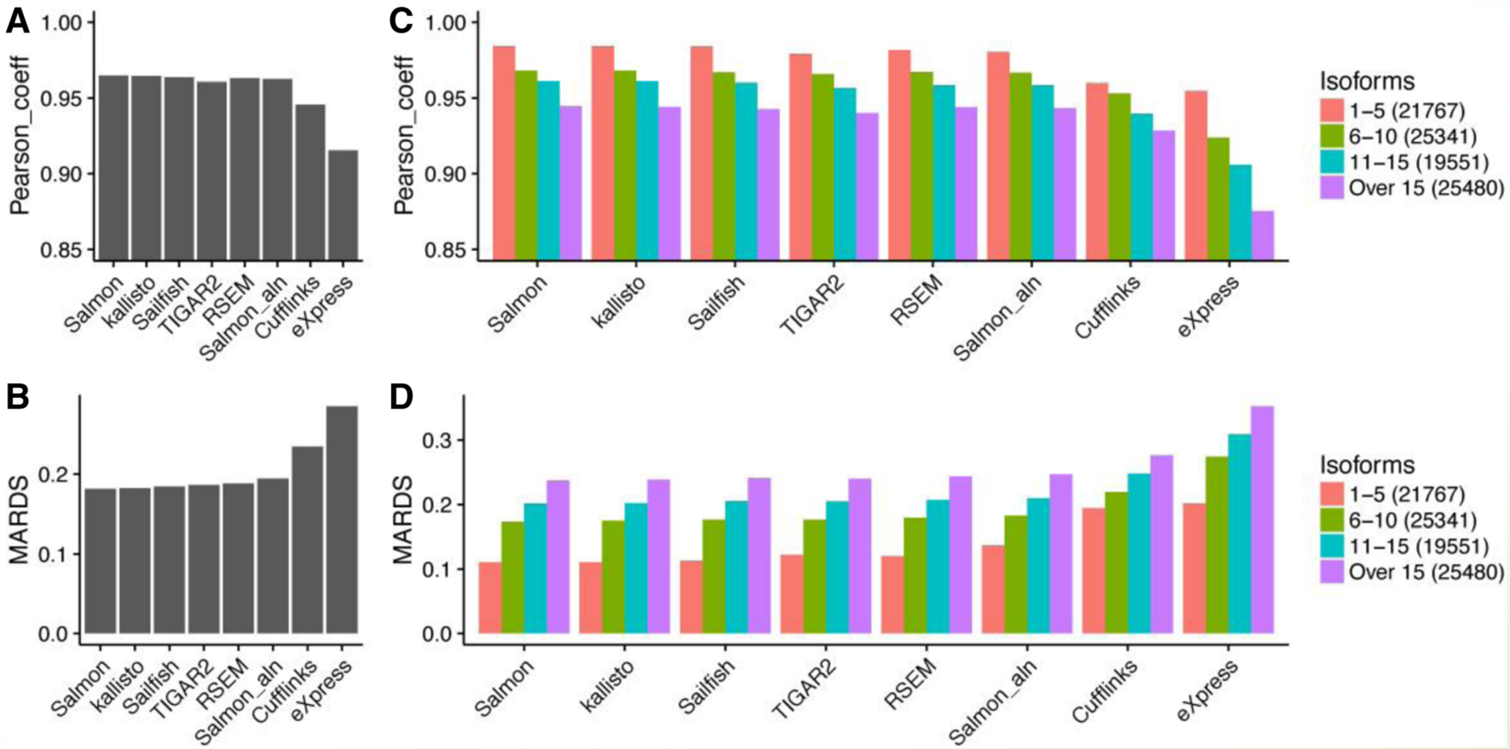


Fig. 2 Comparisons of the overall performance among different methods and the impact of the number of transcripts on the accuracy of isoform quantification. **a** Pearson correlation coefficient. **b** mean absolute relative differences and **c-d**) The above metrics were broken into separate groups 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

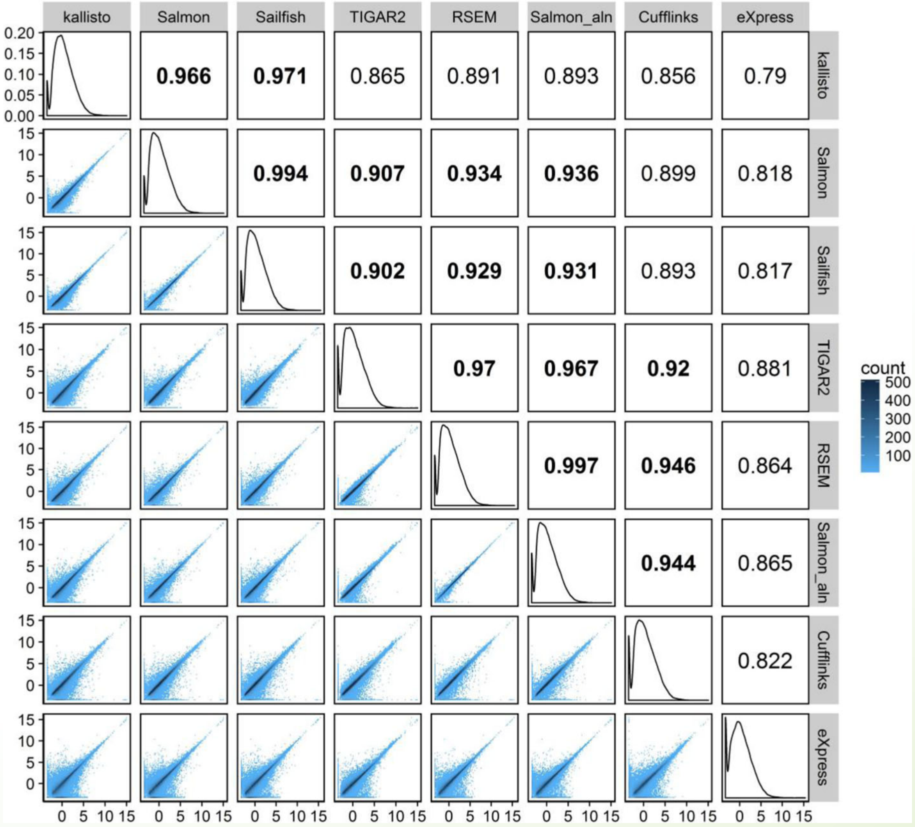
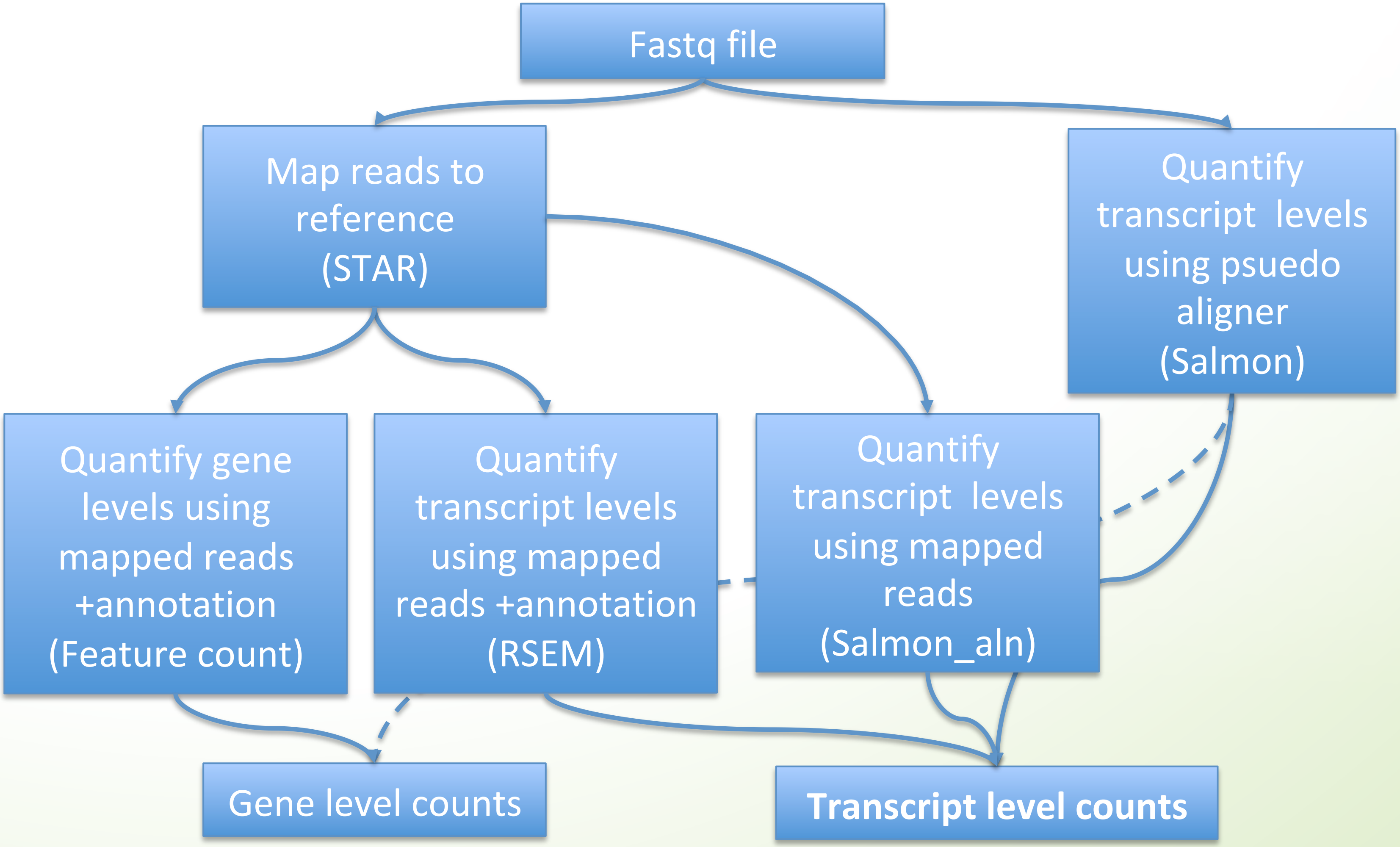


Fig. 5 Pairwise correlation of estimated TPM values for all transcripts between methods for the HBRR-C4 sample. The distribution of transcripts' TPMs from each method was plotted on the diagonal panels. Pairwise density plots and R^2 values are shown in the lower and upper triangular panels, respectively.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

What to choose?
My personal choices


Fastq file

Map reads to reference (STAR)

Quantify transcript levels using psuedo aligner (Salmon)

Quantify gene levels using mapped reads +annotation (Feature count)

Quantify transcript levels using mapped reads +annotation (RSEM)

Quantify transcript levels using mapped reads (Salmon_aln)

Gene level counts

Transcript level counts

Good luck!