RNA-seq quantification strategies

Harvard Chan Bioinformatics Core NGS Data Analysis Course

September 12, 2018

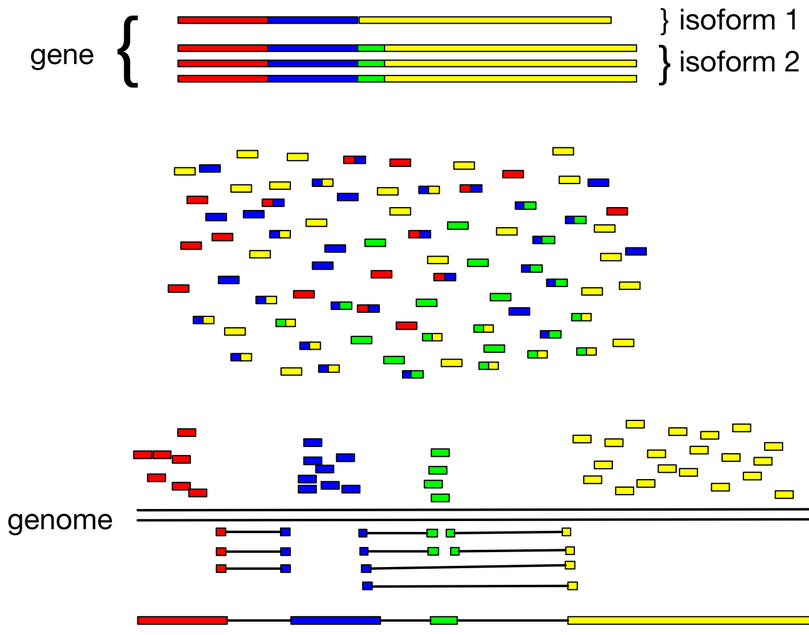
Rory Kirchner

kirchner@hsph.harvard.edu

rorykirchner

roryk

General RNA-seq quantification overview



transcript	sample1
isoform 1	1
isoform 2	3

Quantification units: RPKM, TPM, counts

- many tools report either counts, RPKM or TPM
- matrix of feature x sample read counts
- RPKM and TPM are normalized
- RPKM and TPM are both the number of reads assigned per kilobase of transcript per million reads assigned
- RPKM and TPM are calculated the same way, but the order of normalizations (kilobase of transcript and library depth) are reversed.
- Use TPM instead of RPKM.

RPKM (Reads Per Kilobase Million)

1. Divide by the read depth for each sample.
2. Divide by 1,000,000.
3. Divide by the length of the gene, in kilobases.

RPKM, starting counts

Gene Name	Rep1	Rep2	Rep3
A (2 kb)	10	12	30
B (4 kb)	20	25	60
C (1 kb)	5	8	15
D (10 kb)	0	0	1

example data from https://www.youtube.com/watch?time_continue=569&v=TTUrtCY2k-w

RPKM, calculate scaling factors

Gene Name	Rep1	Rep2	Rep3
A (2 kb)	10	12	30
B (4 kb)	20	25	60
C (1 kb)	5	8	15
D (10 kb)	0	0	1
total	35	45	106
scaling factor	3.5	4.5	10.6

RPKM, normalize by library depth

Gene Name	Rep1	Rep2	Rep3
A (2 kb)	10	12	30
B (4 kb)	20	25	60
C (1 kb)	5	8	15
D (10 kb)	0	0	1
total	35	45	106
scaling factor	3.5	4.5	10.6

Gene Name	Rep1	Rep2	Rep3
A (2 kb)	2.86	2.67	2.83
B (4 kb)	5.71	5.56	5.66
C (1 kb)	1.43	1.78	1.43
D (10 kb)	0	0	0.09

RPKM, normalize by transcript length

Gene Name	Rep1	Rep2	Rep3
A (2 kb)	2.86	2.67	2.83
B (4 kb)	5.71	5.56	5.66
C (1 kb)	1.43	1.78	1.43
D (10 kb)	0	0	0.09

Gene Name	Rep1	Rep2	Rep3
A (2 kb)	1.43	1.33	1.42
B (4 kb)	1.43	1.39	1.42
C (1 kb)	1.43	1.78	1.42
D (10 kb)	0	0	0.009

RPKM, before and after

Gene Name	Rep1	Rep2	Rep3
A (2 kb)	10	12	30
B (4 kb)	20	25	60
C (1 kb)	5	8	15
D (10 kb)	0	0	1
total	35	45	106

Gene Name	Rep1	Rep2	Rep3
A (2 kb)	1.43	1.33	1.42
B (4 kb)	1.43	1.39	1.42
C (1 kb)	1.43	1.78	1.42
D (10 kb)	0	0	0.009
total	4.29	4.5	4.25

TPM (transcripts per million)

1. Divide by the length of the gene in kilobases.
2. Divide by read depth.
3. Divide by 1,000,000.

TPM, normalize by transcript length

Gene Name	Rep1	Rep2	Rep3
A (2 kb)	10	12	30
B (4 kb)	20	25	60
C (1 kb)	5	8	15
D (10 kb)	0	0	1

Gene Name	Rep1	Rep2	Rep3
A (2 kb)	5	6	15
B (4 kb)	5	6.25	15
C (1 kb)	5	8	15
D (10 kb)	0	0	0.1

TPM, normalize by library depth

Gene Name	Rep1	Rep2	Rep3
A (2 kb)	5	6	15
B (4 kb)	5	6.25	15
C (1 kb)	5	8	15
D (10 kb)	0	0	0.1
total	15	20.25	45.1
scaling factor	1.5	2.025	4.51

Gene Name	Rep1	Rep2	Rep3
A (2 kb)	3.33	2.96	3.326
B (4 kb)	3.33	3.09	3.326
C (1 kb)	3.33	3.95	3.326
D (10 kb)	0	0	0.02

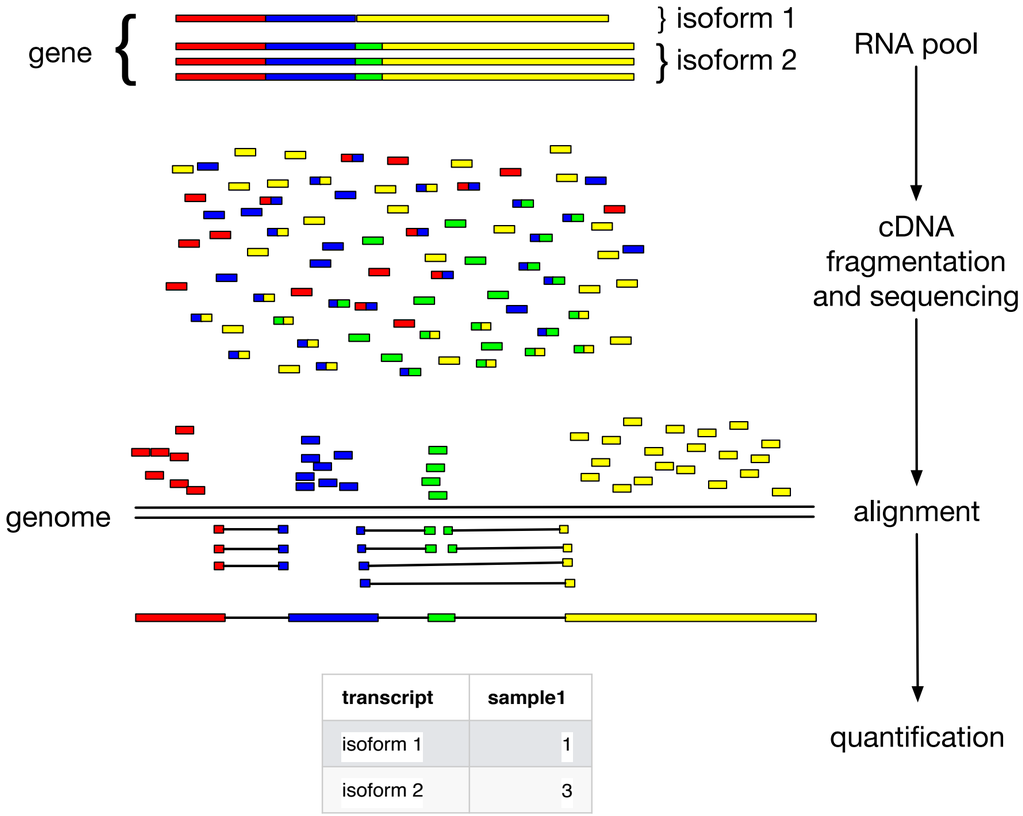
RPKM vs TPM

Gene Name	Rep1	Rep2	Rep3
A (2 kb)	1.43	1.33	1.42
B (4 kb)	1.43	1.39	1.42
C (1 kb)	1.43	1.78	1.42
D (10 kb)	0	0	0.009
total	4.29	4.5	4.25

Gene Name	Rep1	Rep2	Rep3
A (2 kb)	3.33	2.96	3.326
B (4 kb)	3.33	3.09	3.326
C (1 kb)	3.33	3.95	3.326
D (10 kb)	0	0	0.02
total	10	10	10

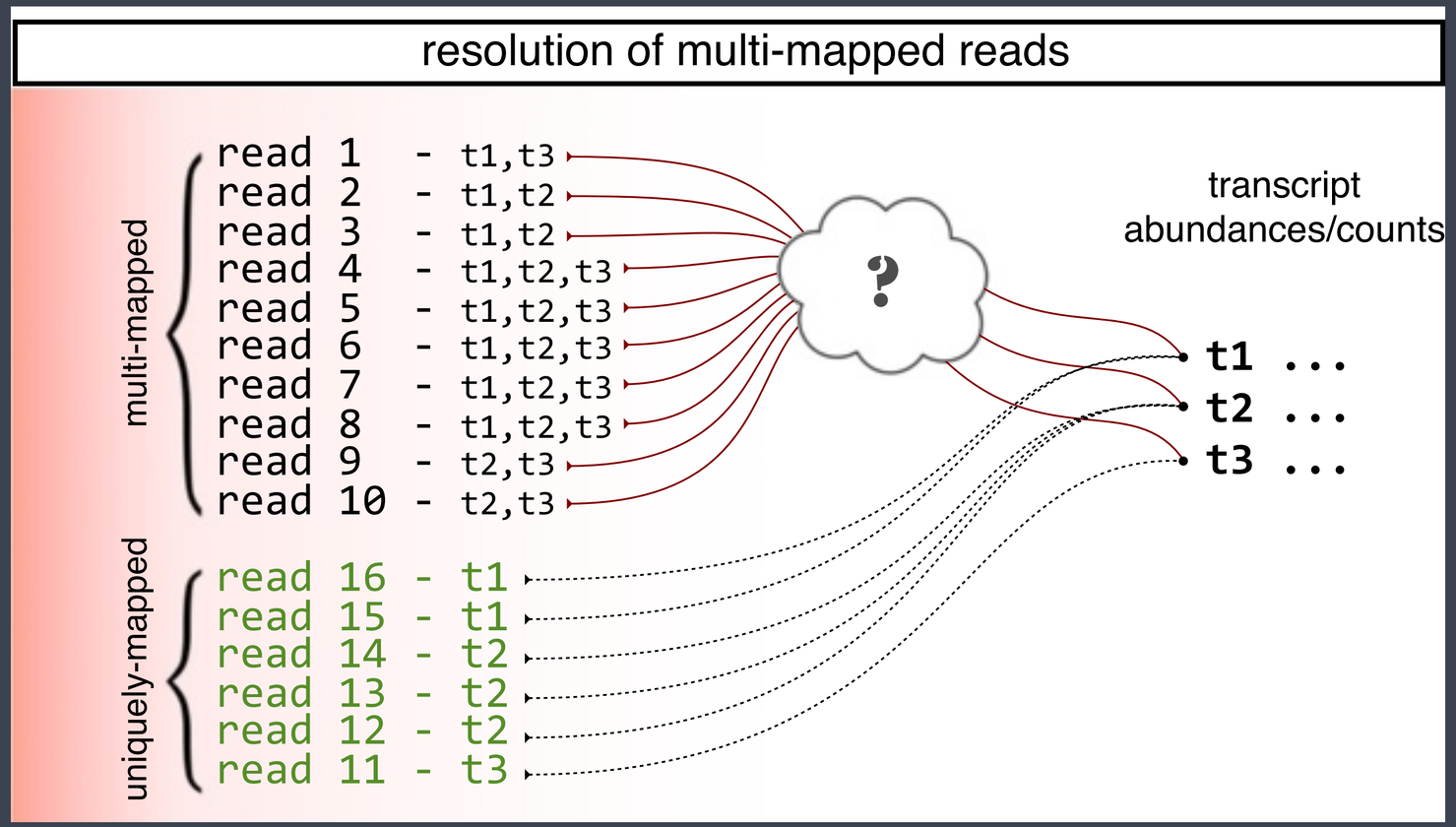
Three popular read assignment strategies

- align to genome and assign
- skip alignment, compare kmer content of read to kmer content of transcripts and assign
- quasialign to transcriptome and assign



transcript	sample1
isoform 1	1
isoform 2	3

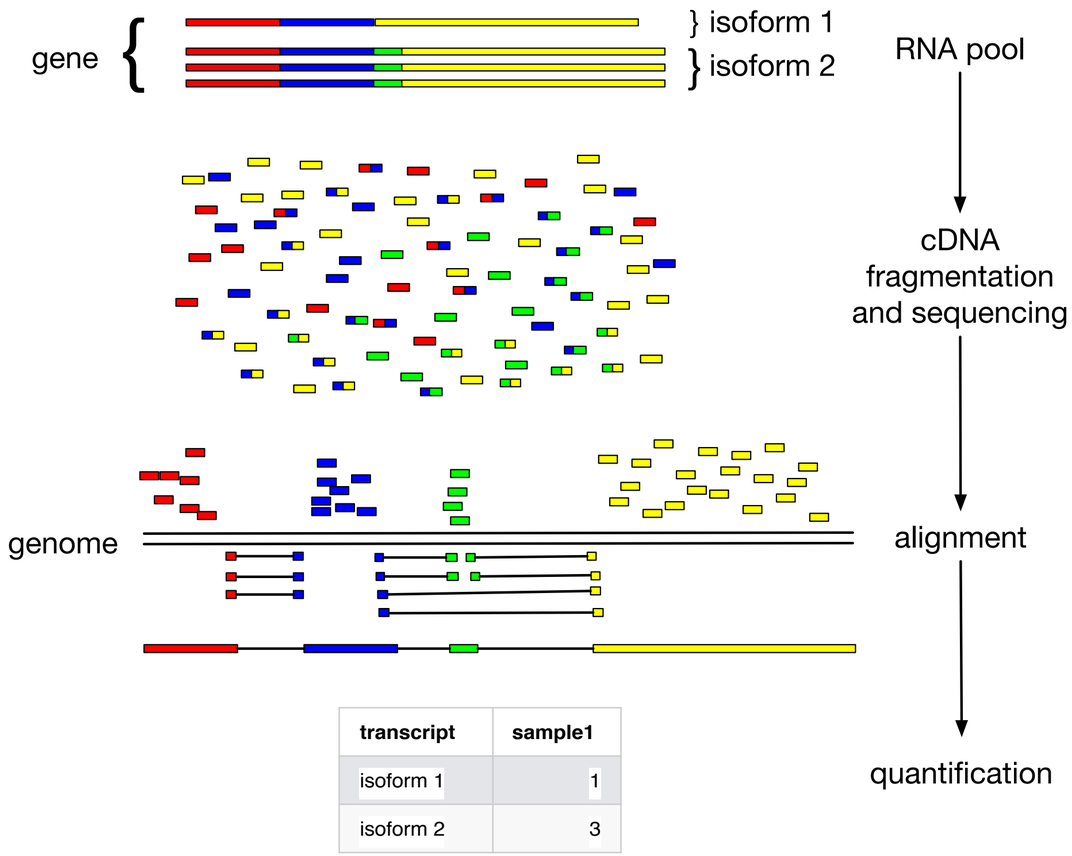
Ambiguous reads cause quantification noise



Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts, Genome Biology, 2016

align and count

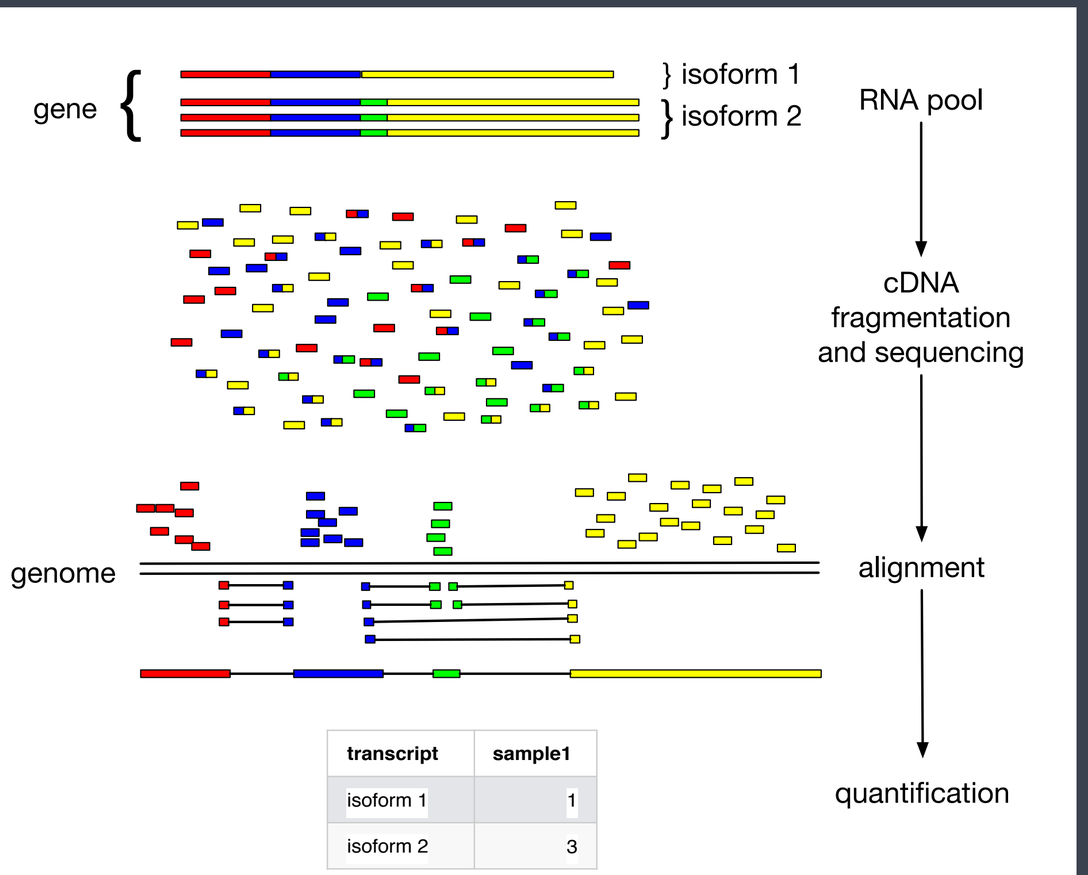
- align to genome with split-read aligner (STAR is the most common choice of aligner)
- overlay feature model
- count how many reads align to each feature
- what to do with reads that could come from more than one feature?



transcript	sample1
isoform 1	1
isoform 2	3

Three popular read assignment strategies

- align to genome and assign
- skip alignment, compare kmer content of read to kmer content of transcripts and assign
- quasialign to transcriptome and assign



transcript	sample1
isoform 1	1
isoform 2	3

What is a kmer

ACGTACGTACGTAGCTAGCATC

ACGTACGTACGTAGCTAGCATC

ACGTACGTACGTAGCTAGCATC

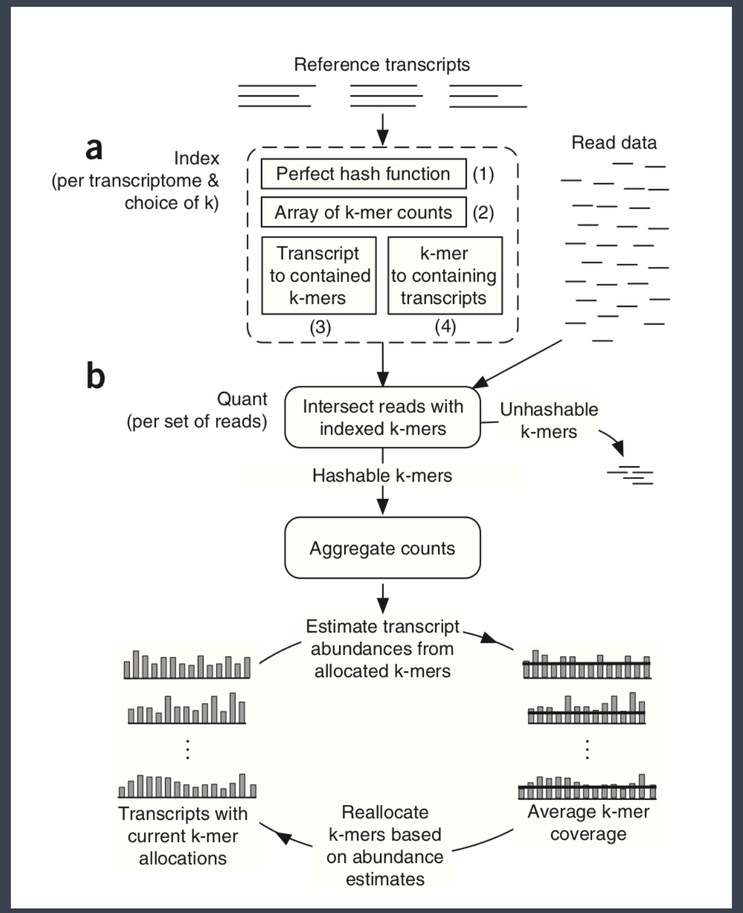
ACGTACGTACGTAGCTAGCATC

ACGTACGTACGTAGCTAGCATC

GTA	3	TAC	2
ACG	3	GCA	1
CGT	3	CAT	1
AGC	2	GCT	1
TAG	2	CTA	1

kmer signature

Sailfish: alignment via counting kmers



ACGTAC**GTACGTAG**CTAGCATC

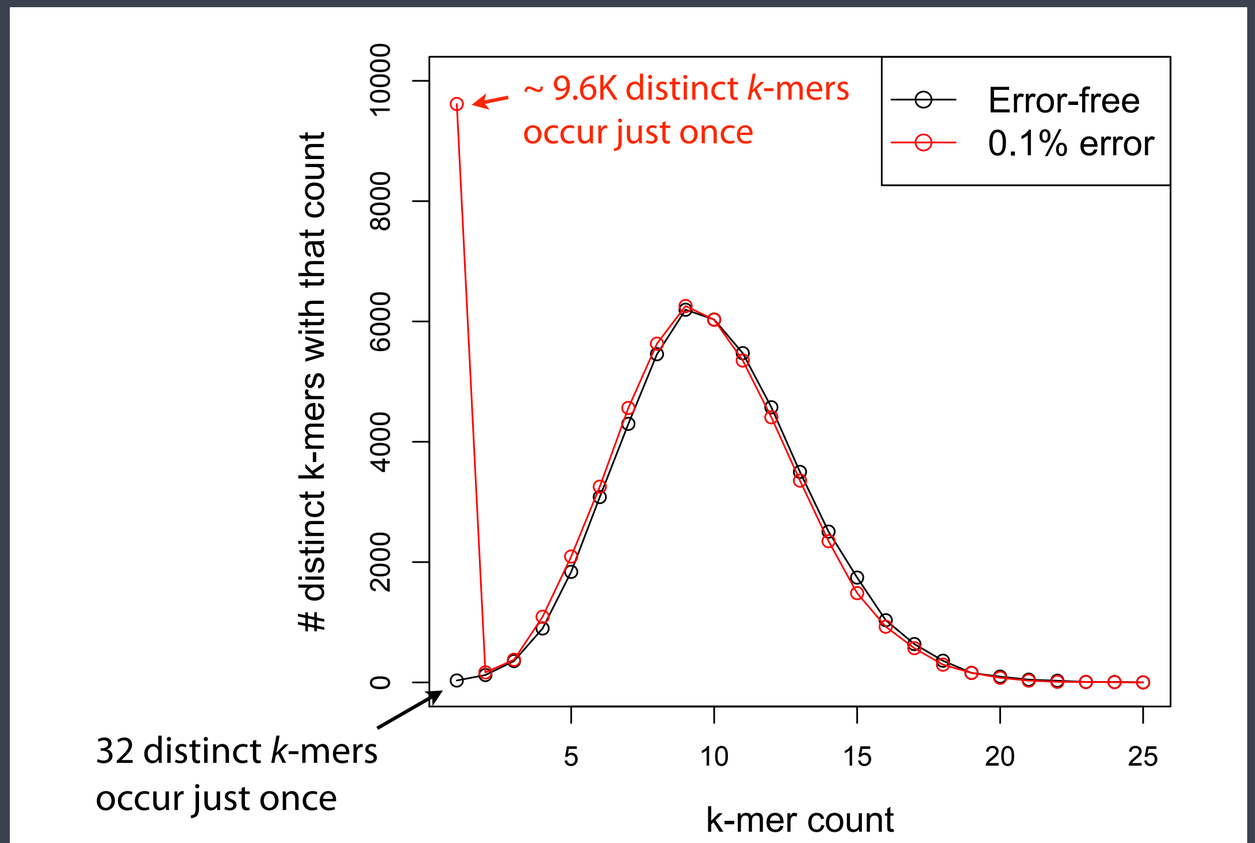
GTA	3	TAC	2
ACG	3	GCA	1
CGT	3	CAT	1
AGC	2	GCT	1
TAG	2	CTA	1

GTACGTTG

GTA	1	TAC	1
ACG	1	CGT	1
GTA	1	GTT	1

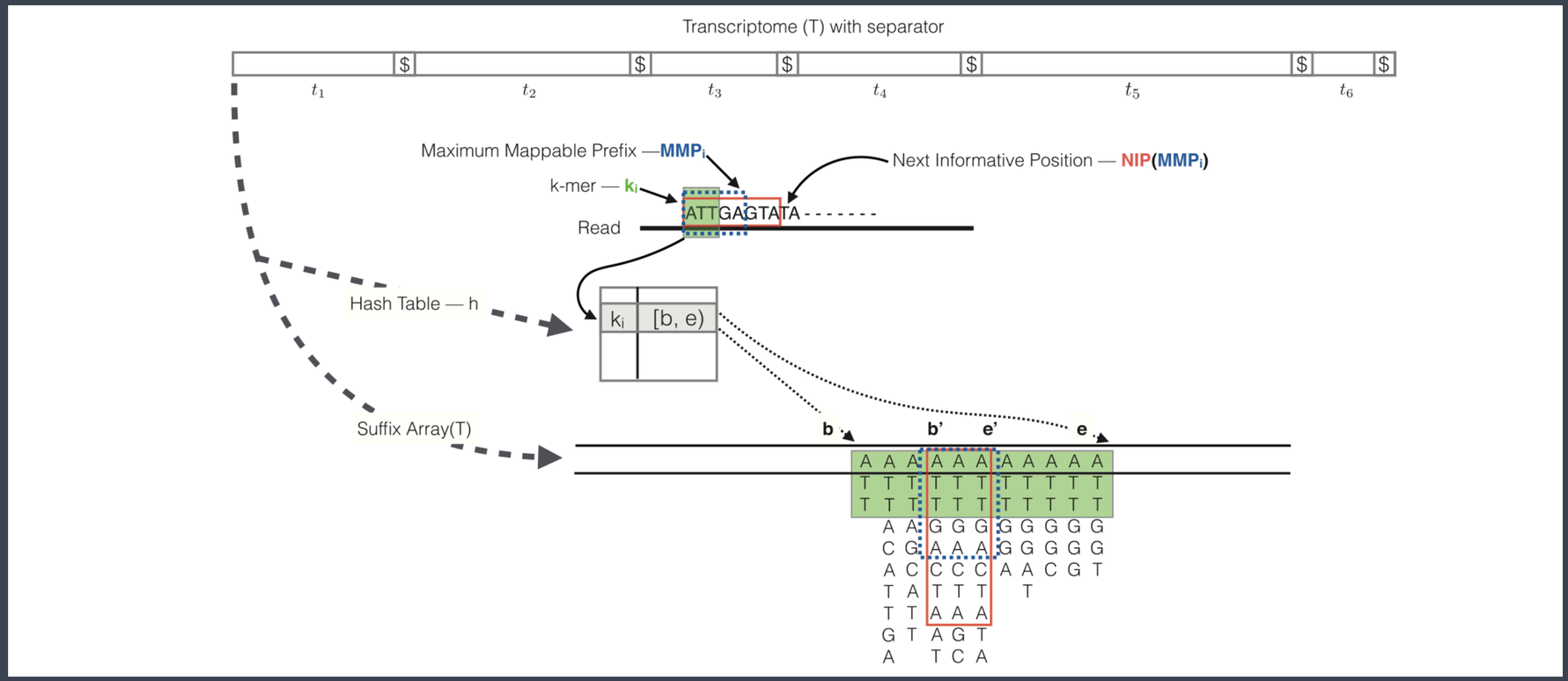
Patro, Rob, Stephen M. Mount, and Carl Kingsford. 2014. “Sailfish Enables Alignment-Free Isoform Quantification from RNA-Seq Reads Using Lightweight Algorithms.” Nature Biotechnology 32 (5): 462–64.

kmers with errors tend to be unique



https://www.cs.jhu.edu/~langmea/resources/lecture_notes/error_correction.pdf

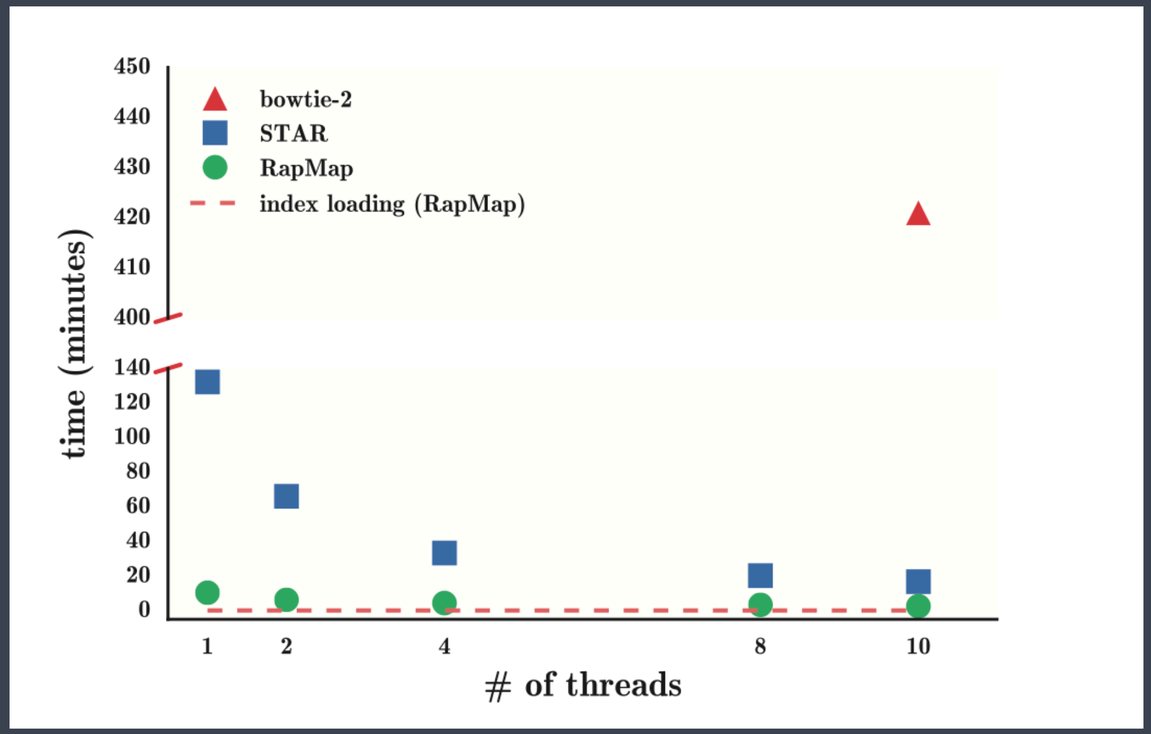
Salmon: quasialignment



Srivastava, Avi, Hirak Sarkar, Nitish Gupta, and Rob Patro. 2016. “RapMap: A Rapid, Sensitive and Accurate Tool for Mapping RNA-Seq Reads to Transcriptomes.” Bioinformatics 32 (12): i192–200.

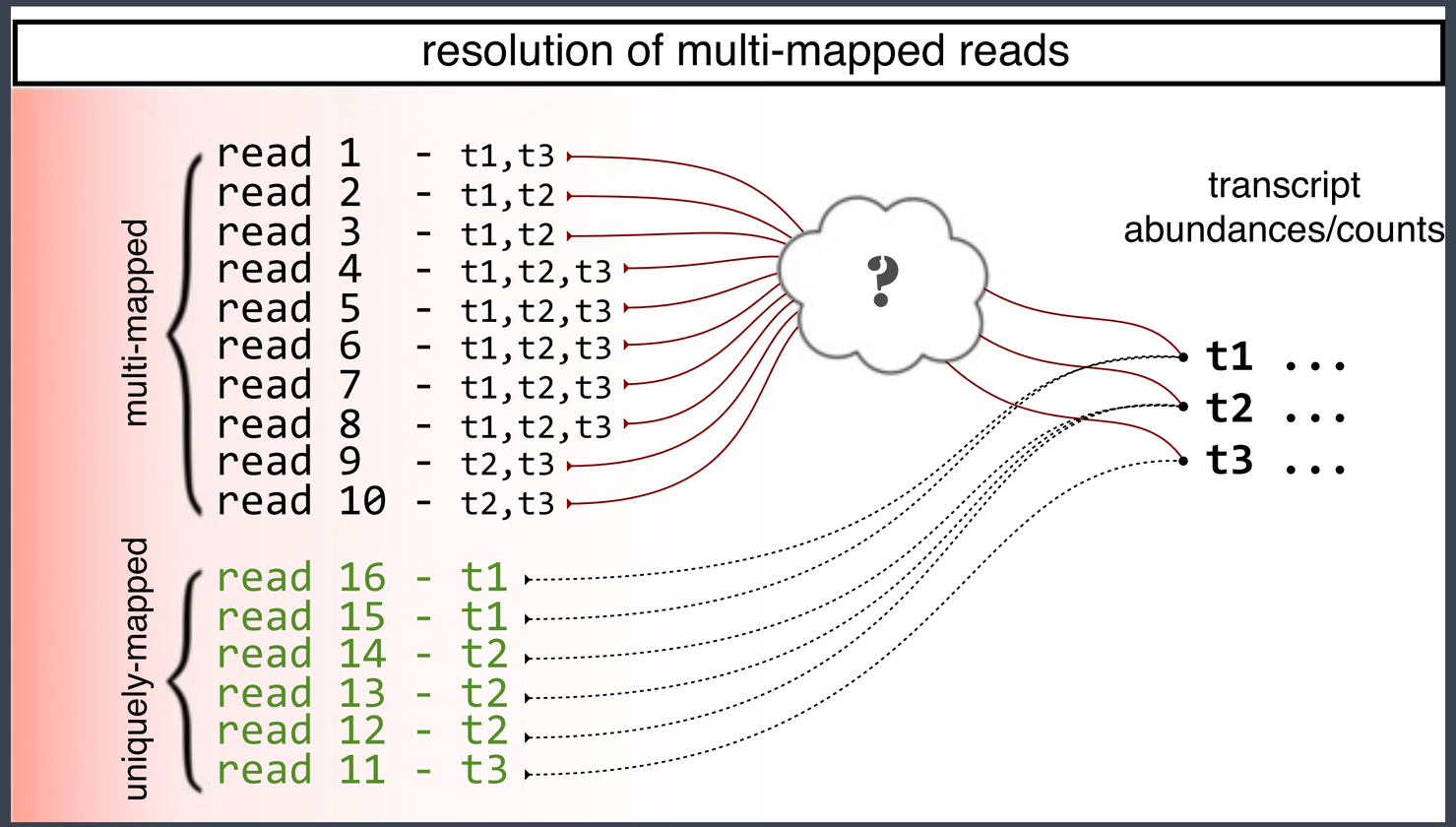
Why does being extremely fast matter?

- most experiments can be run on your laptop
- fast enough to allow for experimentation with little overhead
- bootstrapping
- reanalyzing huge datasets not as onerous



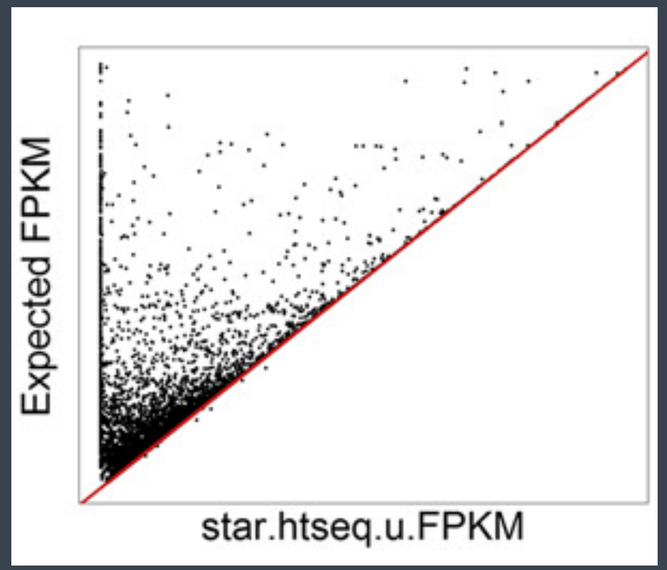
Srivastava, Avi, Hirak Sarkar, Nitish Gupta, and Rob Patro. 2016. “RapMap: A Rapid, Sensitive and Accurate Tool for Mapping RNA-Seq Reads to Transcriptomes.” Bioinformatics 32 (12): i192–200.

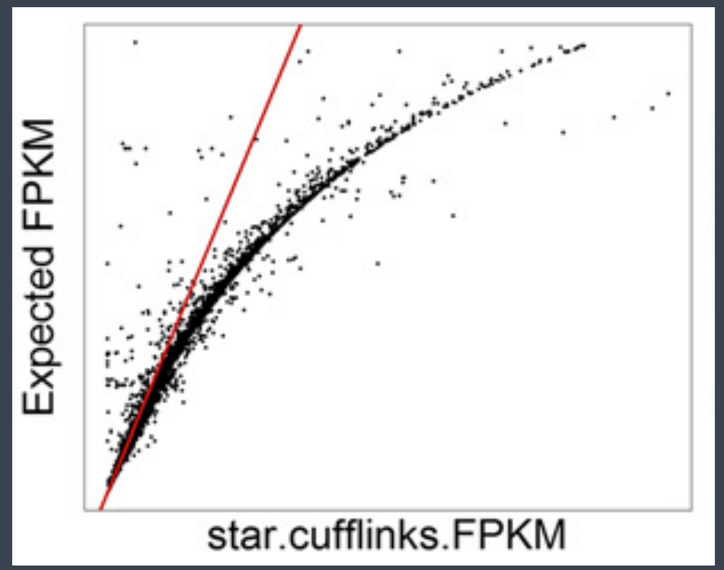
Ambiguous reads cause quantification noise

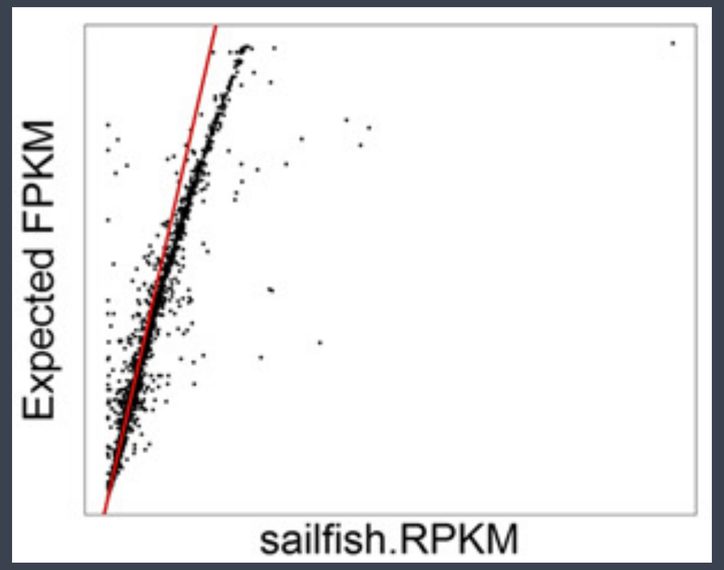


Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts, Genome Biology, 2016

Three multimapper resolution strategies

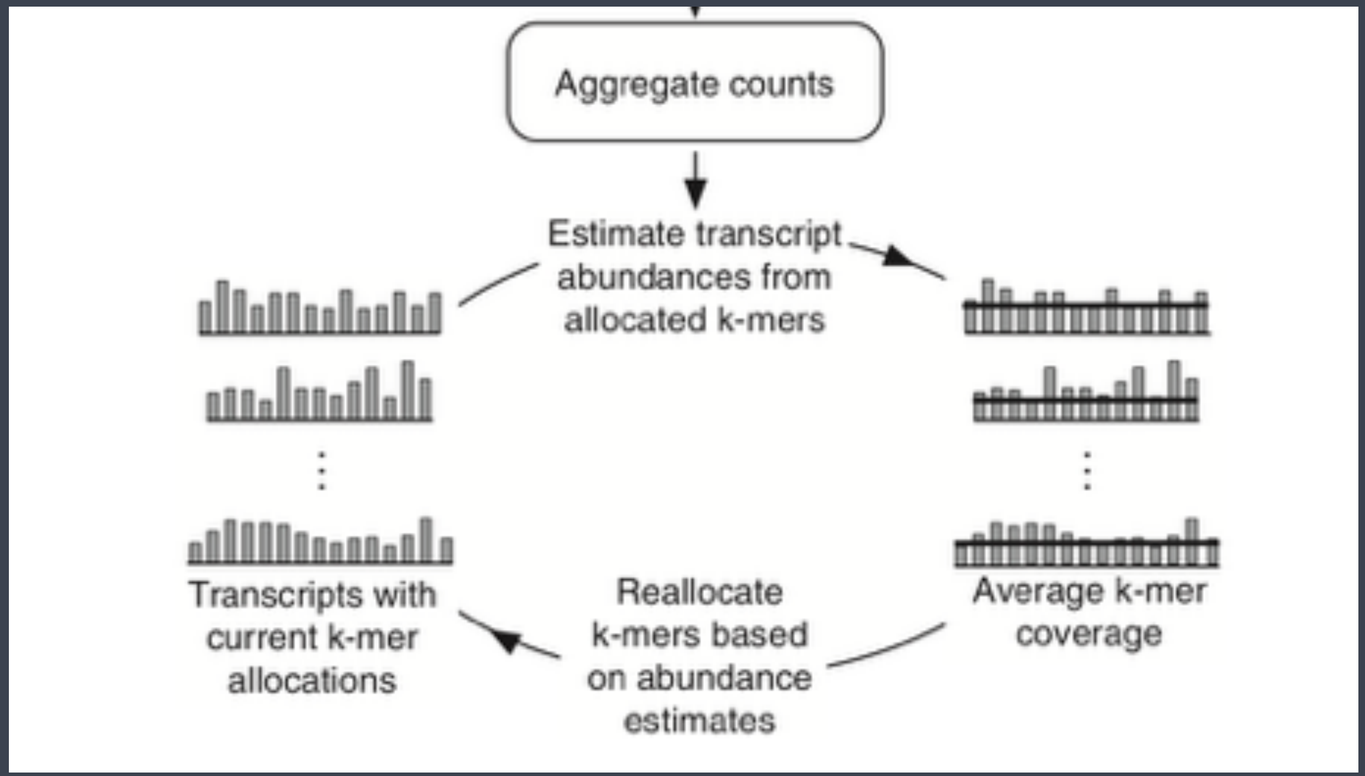




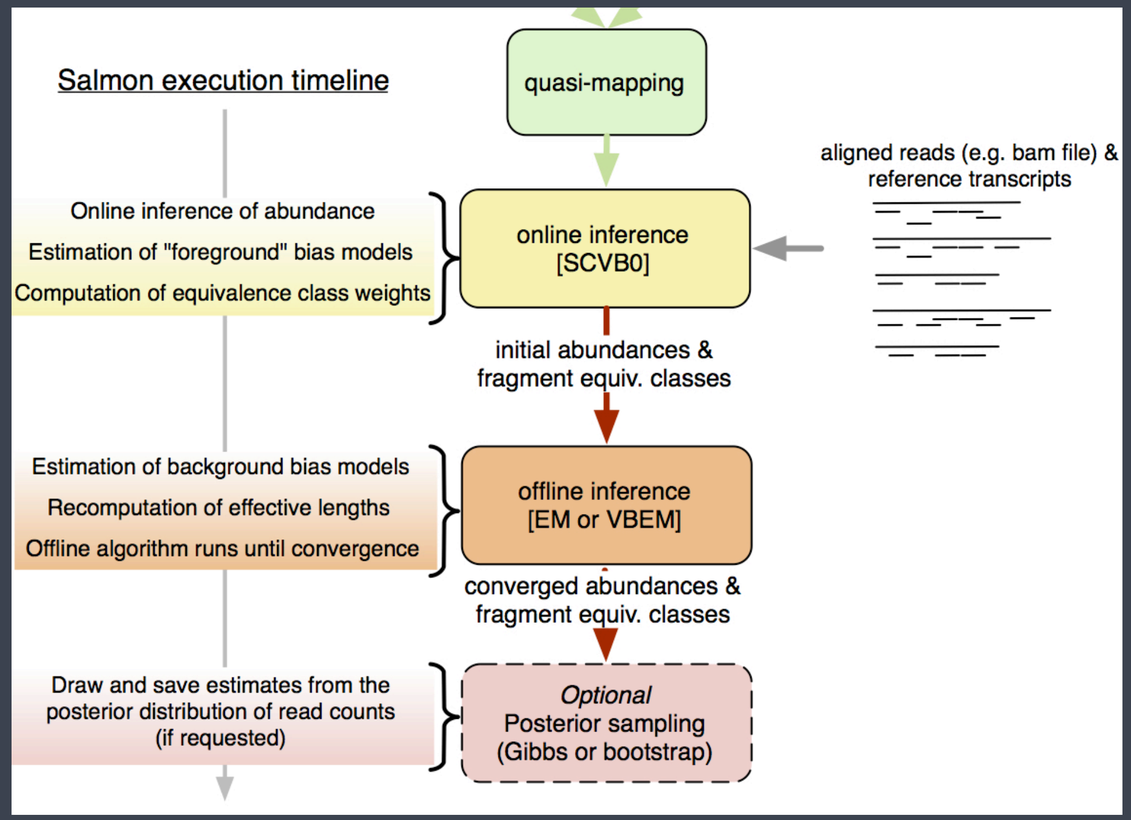


Robert, C., & Watson, M. (2015). Errors in RNA-Seq quantification affect genes of relevance to human disease. Genome Biology, 16(1), 177. Robert, C., & Watson, M. (2015). Errors in RNA-Seq quantification affect genes of relevance to human disease. Genome Biology, 16(1), 177. http://doi.org/10.1186/s13059-015-0734-x

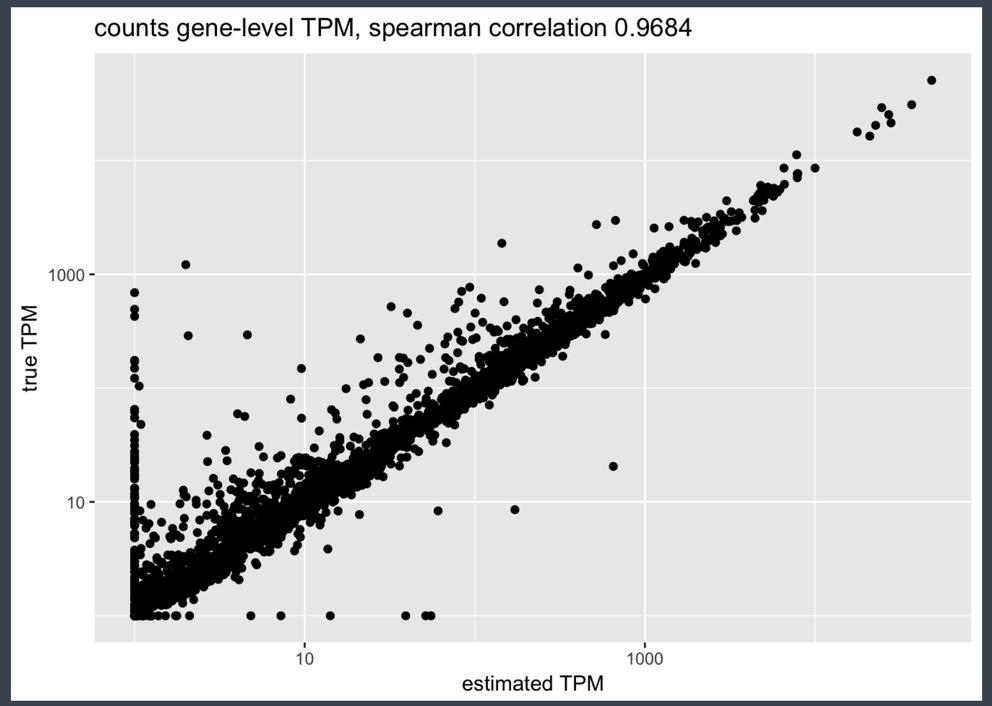
sailfish: quantification

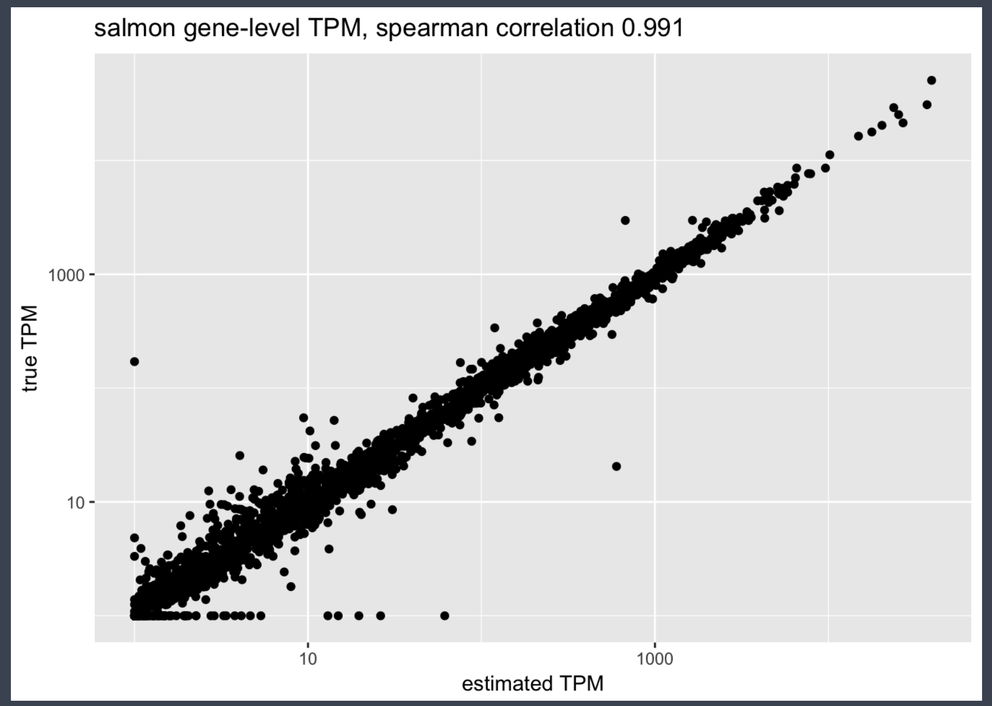


salmon: quantification



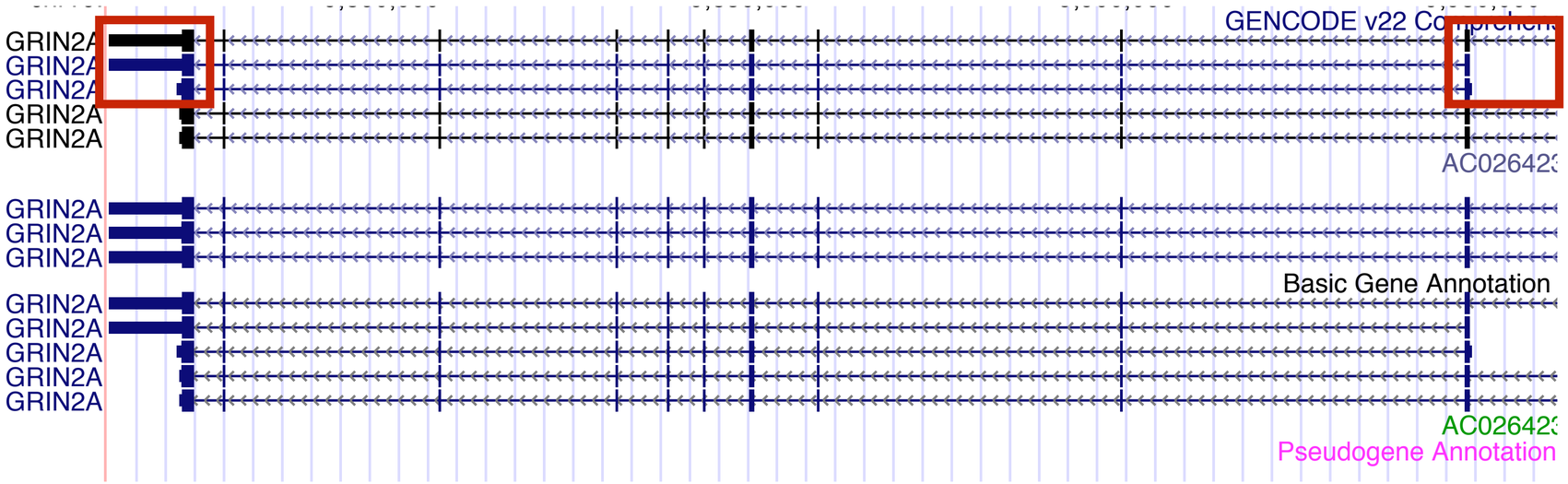
Salmon quantification vs align-and-count



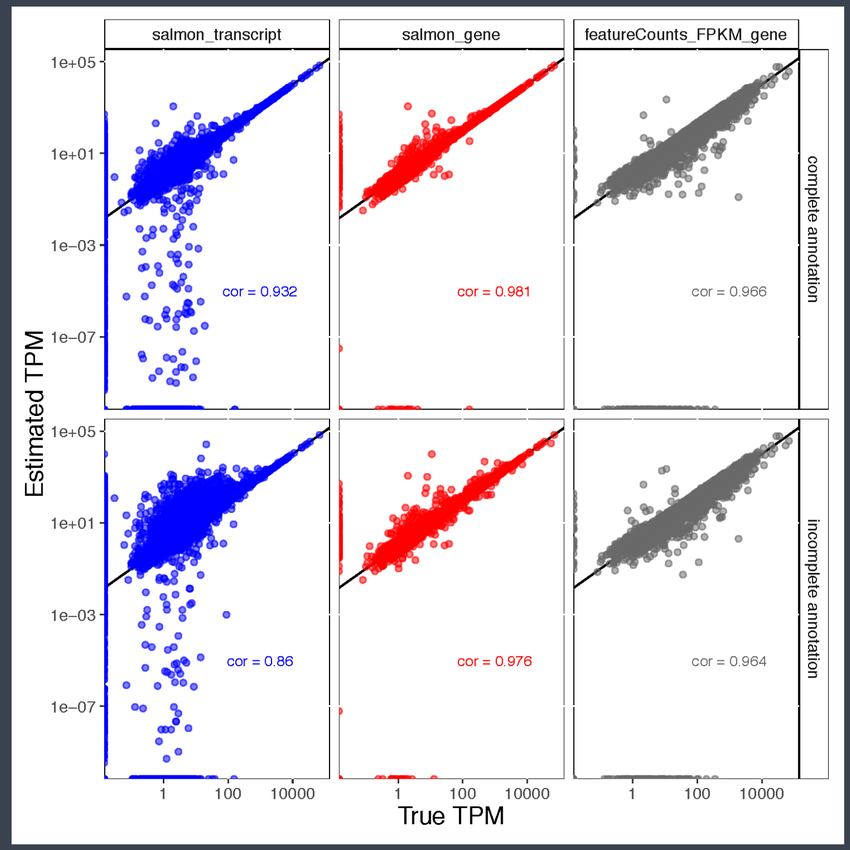


Dataset: E-MTAB-4119 - Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-4119/)

isoform quantification is difficult

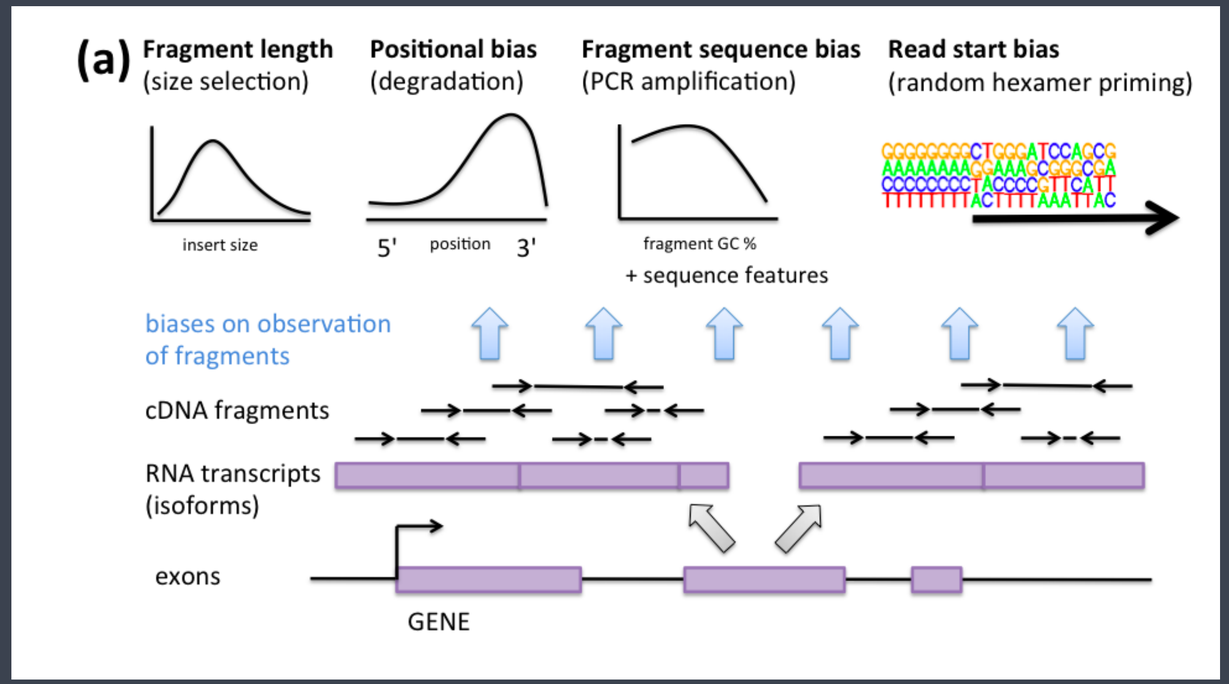


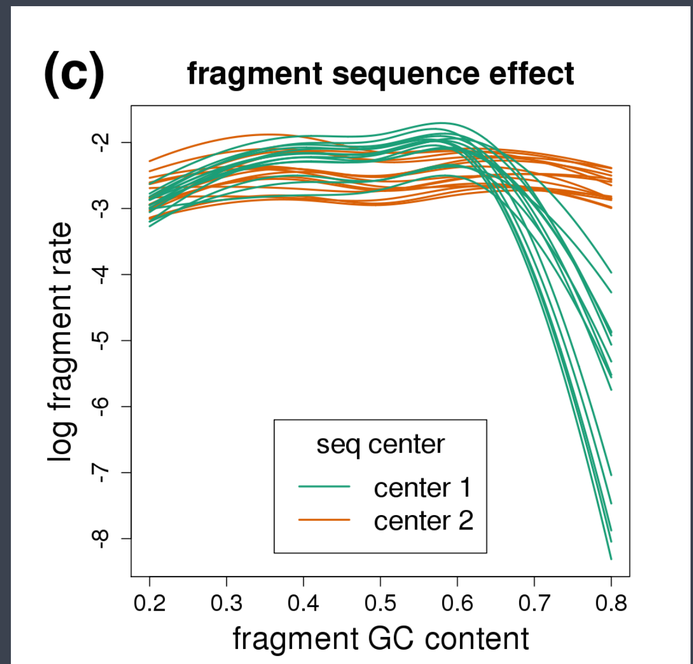
But, always quantitate at transcript level



Soneson, Charlotte, Michael I. Love, and Mark D. Robinson. 2015. “Differential Analyses for RNA-Seq: Transcript-Level Estimates Improve Gene-Level Inferences.” F1000Research 4 (December): 1521.

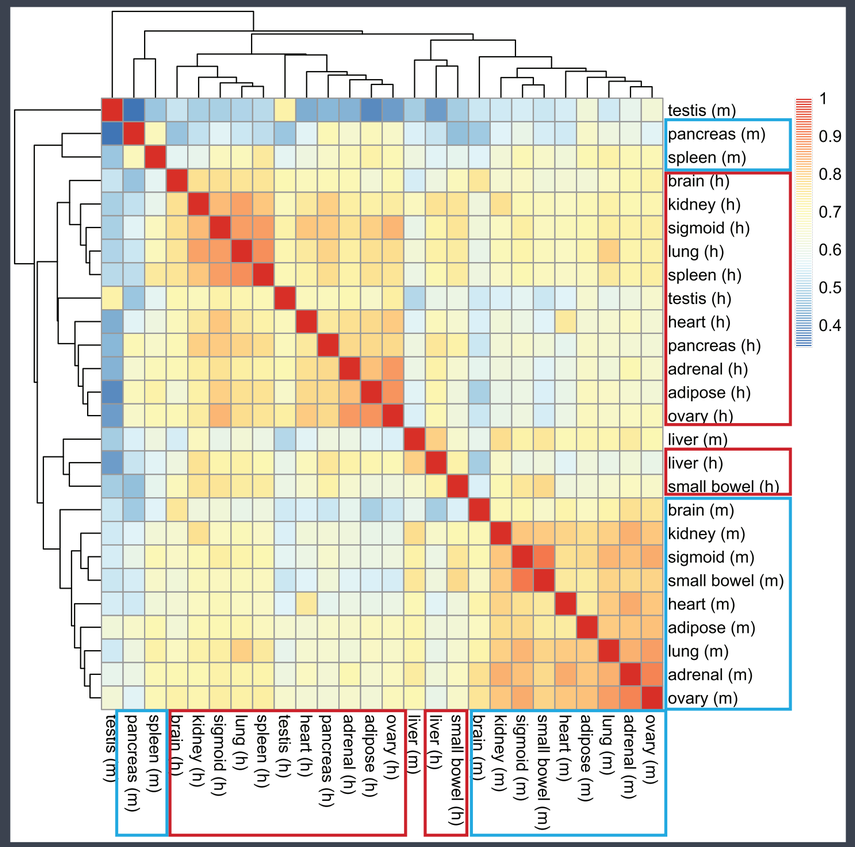
Biases affecting quantification

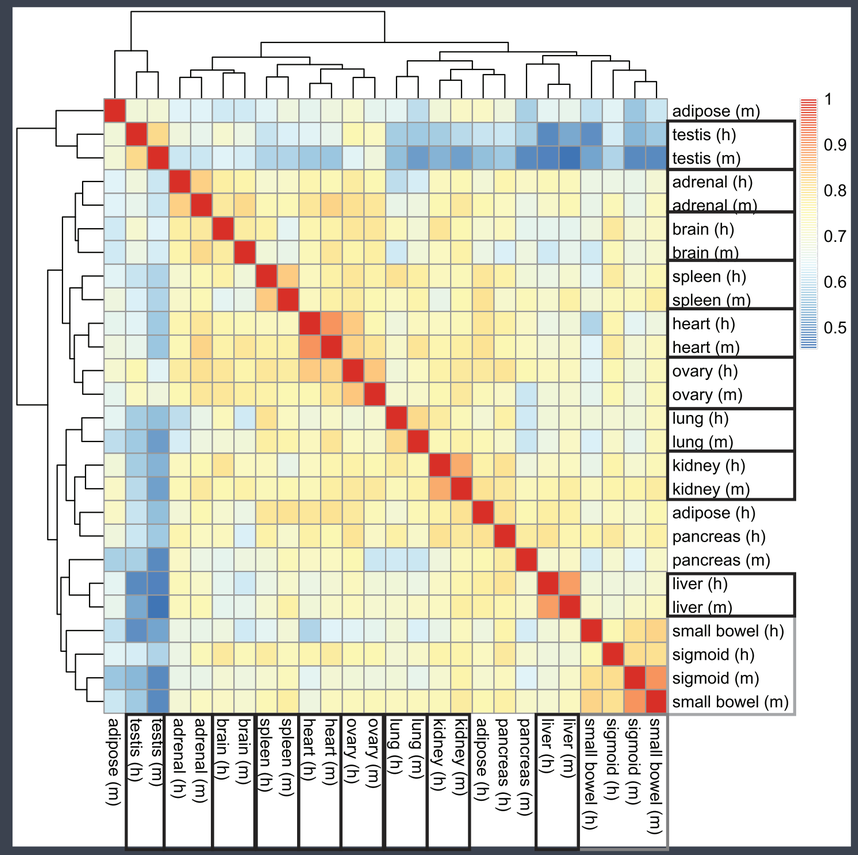




Love, Michael I., John B. Hogenesch, and Rafael A. Irizarry. 2016. “Modeling of RNA-Seq Fragment Sequence Bias Reduces Systematic Errors in Transcript Abundance Estimation.” Nature Biotechnology 34 (12): 1287–91.

Biases affect everyone



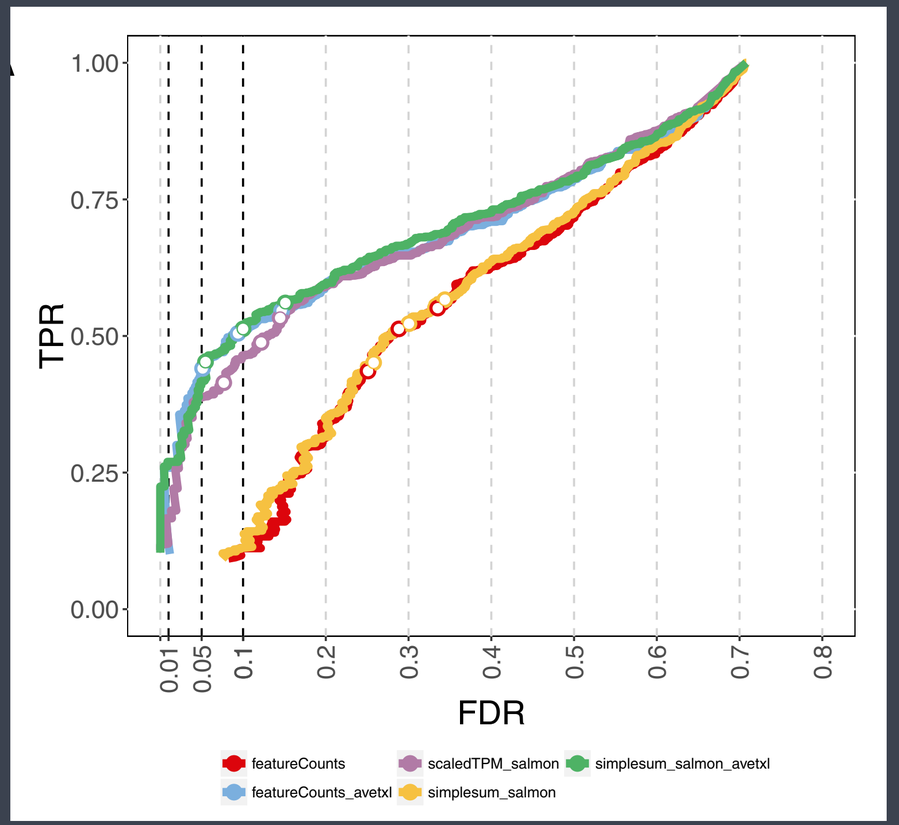


Gilad, Yoav, and Orna Mizrahi-Man. 2015. “A Reanalysis of Mouse ENCODE Comparative Gene Expression Data.” F1000Research 4 (May): 121.

Collapse to gene counts: tximport

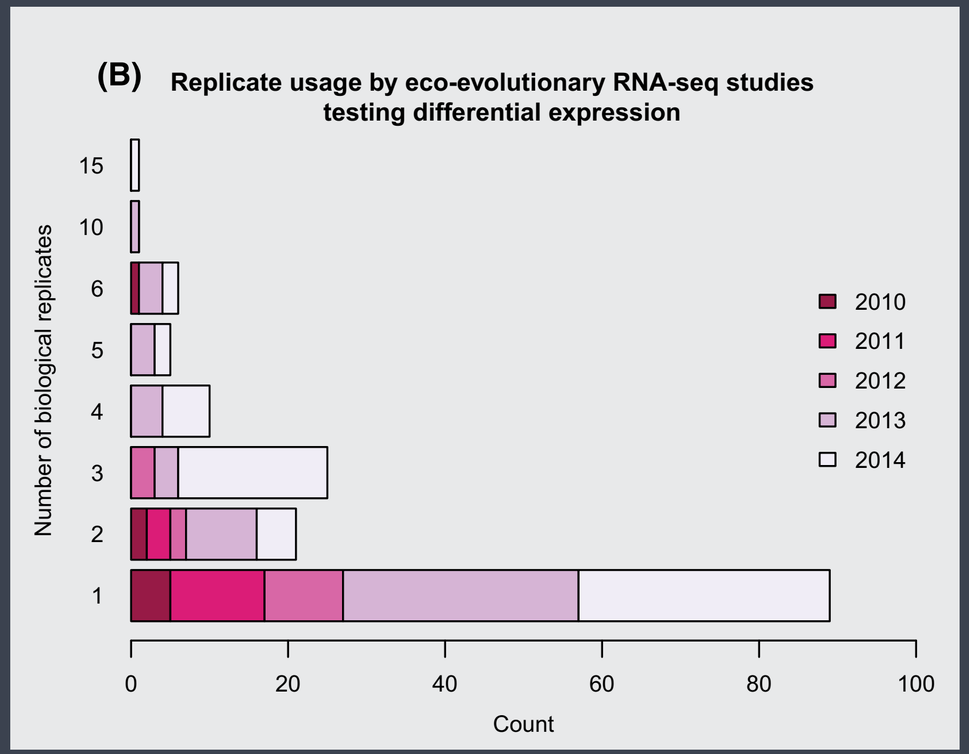
- Many downstream differential expression tools require counts, not TPM
- DESeq2 fits a negative binomial, which needs counts of observations, so we need a count matrix
- TPM and counts from Salmon have length biases in them, for example degradation in one sample will result in smaller effective transcript lengths and artificially smaller TPM
- Salmon reports “effective” transcript length, which is an estimate of the usuable surface of each transcript in each sample
- use tximport to generate counts from TPM with lengthScaledTPM parameter
- this generates gene or transcript level counts that have any effective length biases removed from them

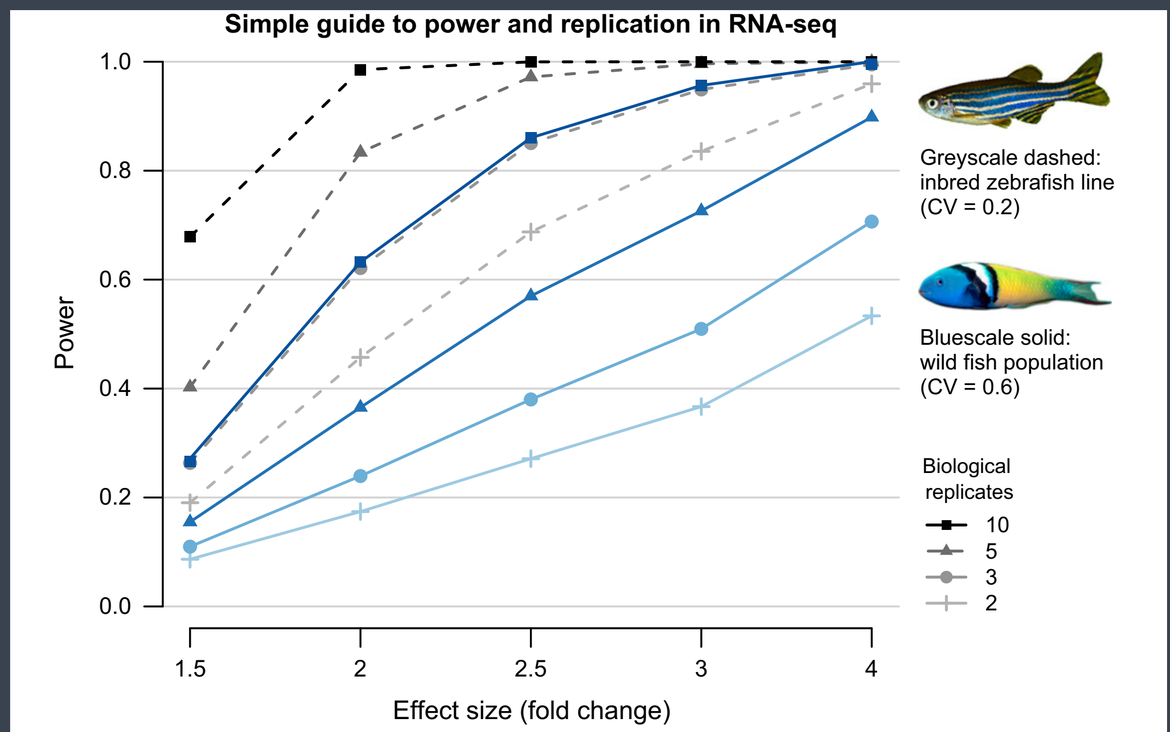
Removing length bias improves quantification



Soneson, Charlotte, Michael I. Love, and Mark D. Robinson. 2015. “Differential Analyses for RNA-Seq: Transcript-Level Estimates Improve Gene-Level Inferences.” F1000Research 4 (December): 1521.

You need more replicates than you think





Todd, Erica V., Michael A. Black, and Neil J. Gemmell. 2016. “The Power and Promise of RNA-Seq in Ecology and Evolution.” Molecular Ecology 25 (6): 1224–41.