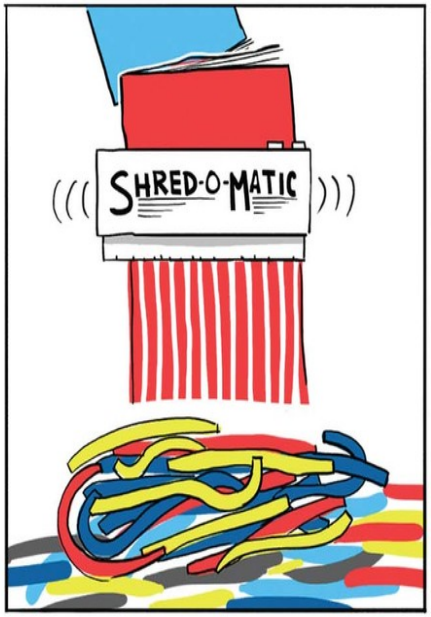
RNA-seq introduction

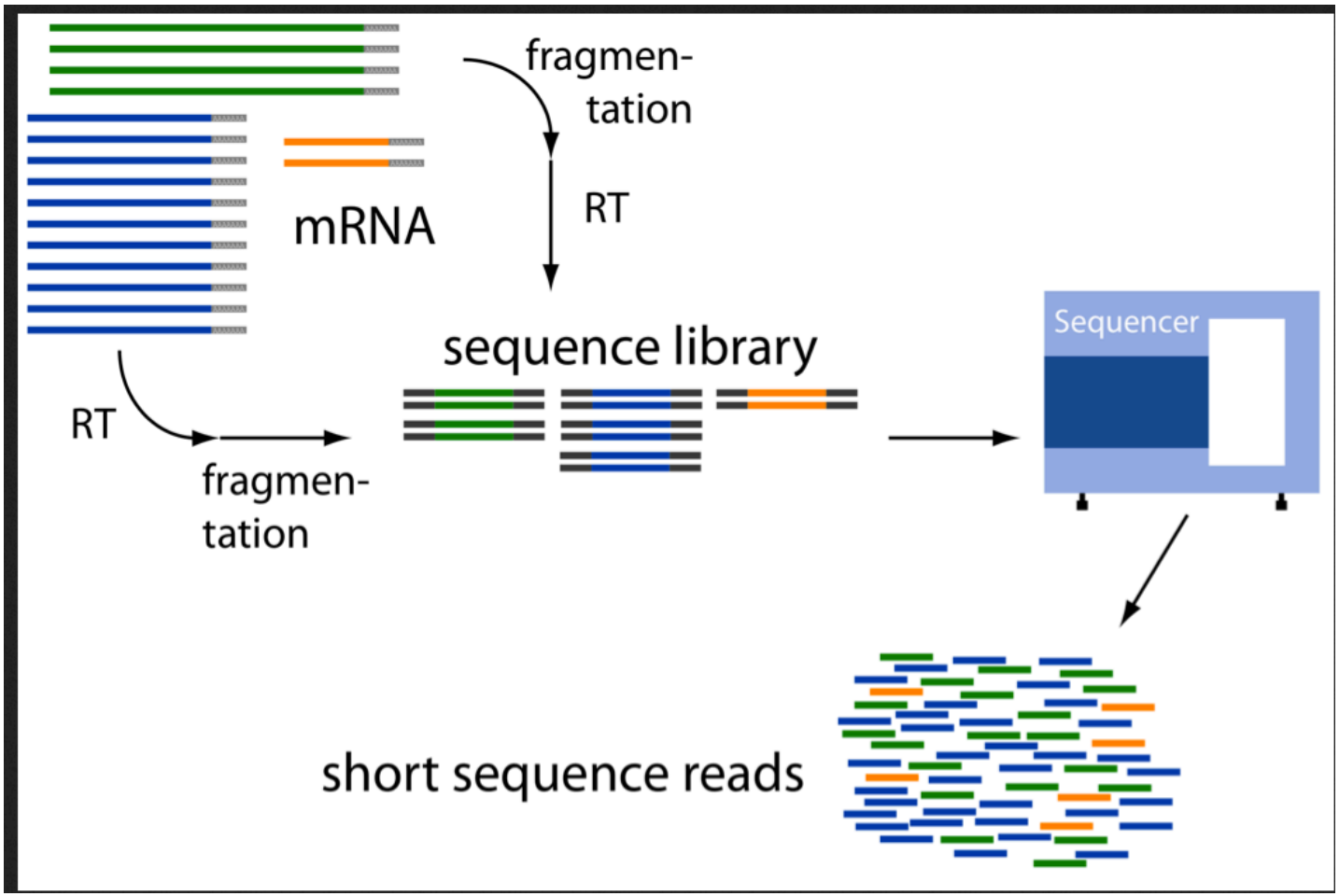
RNA-seq data analysis

Johan Reimegård | 13-May-2019

RNA-seq with short reads



How are RNA-seq data generated?



Sampling process

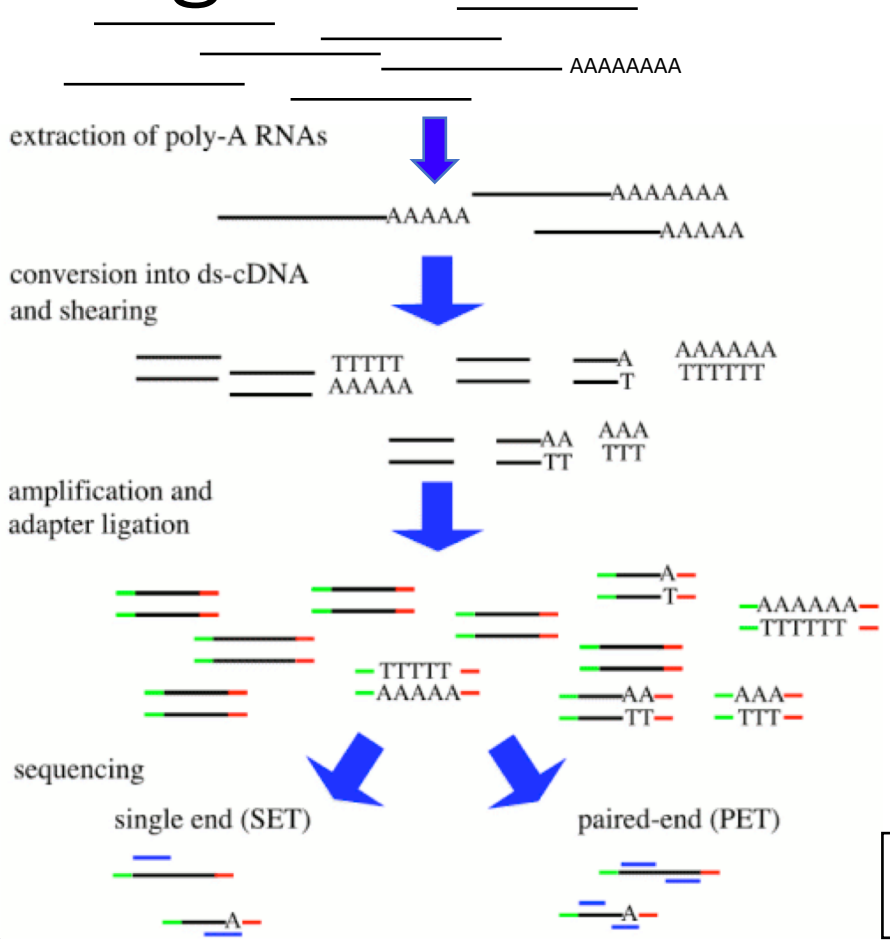
Depending on the different steps you will get different results

RNA->

enrichments ->

library ->

reads ->



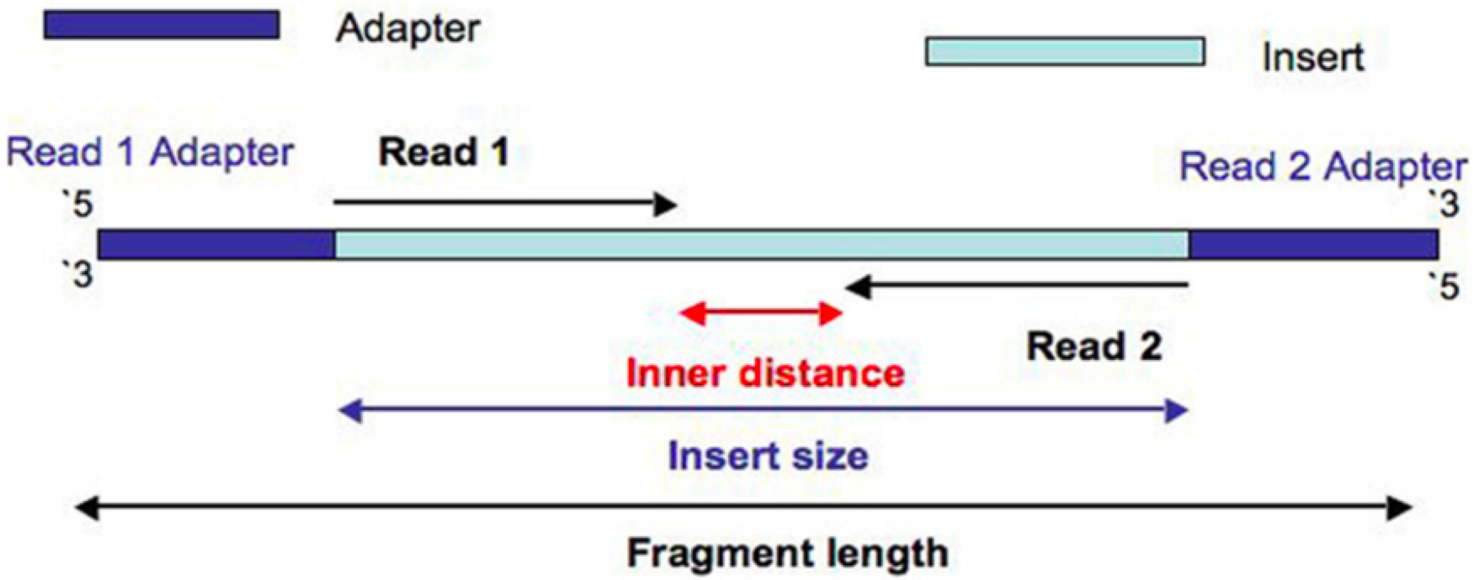
PolyA	(mRNA)
RiboMinus	(- rRNA)
Size <50 nt	(miRNA)
…..	

Size of fragment
Strand specific
5’ end specific
3’ end specific
…..

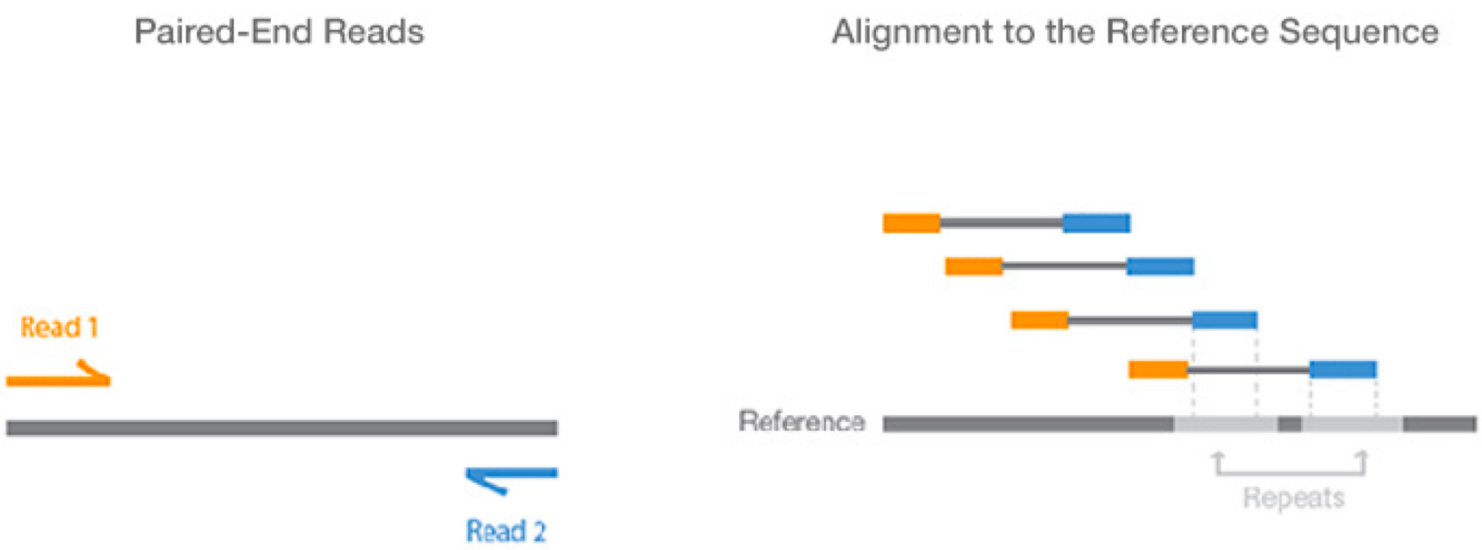
Single end (1 read per fragment)
Paired end (2 reads per fragment)

Single end vs paired end reads

Single end only contains one read per fragment (Read 1)
Paired end reads contains two reads per fragment (Read 1 and Read2)



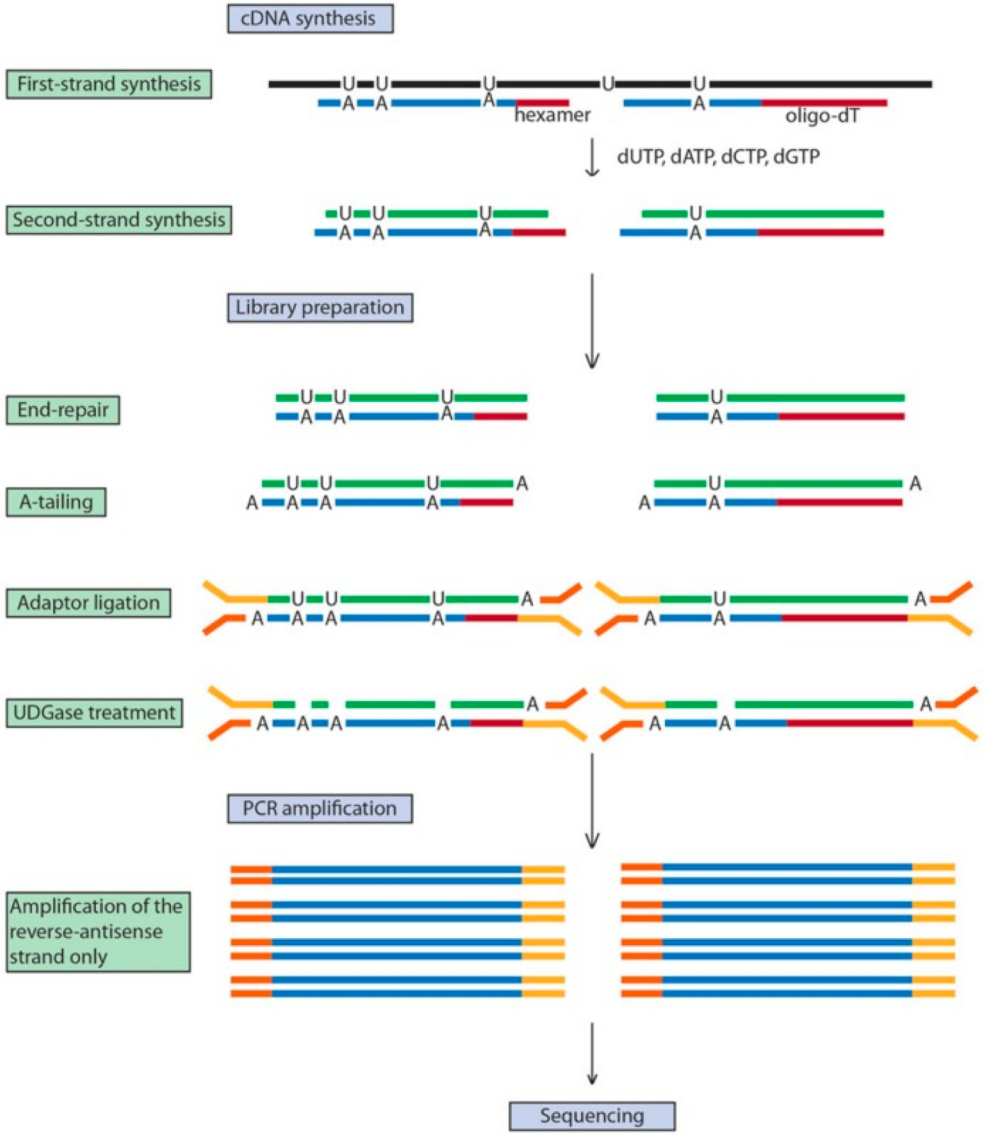
Advantage with paired end reads



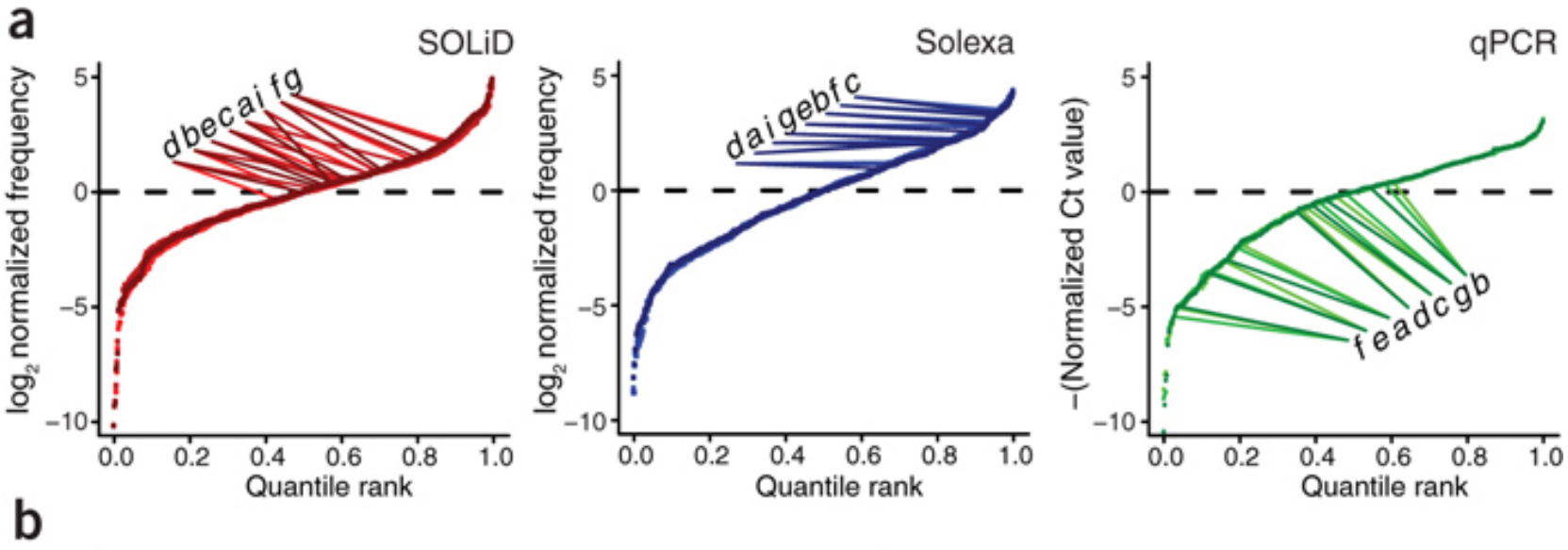
Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

Illumina

Strand specific sequencing



Different sequencing teqhniques have different preferences



Sequencing frequency of 472 artificial mircoRNAs in equal abundance

(Figure from Linsen *et al.,* Nature Methods. 2009)

But evens out over longer RNAs

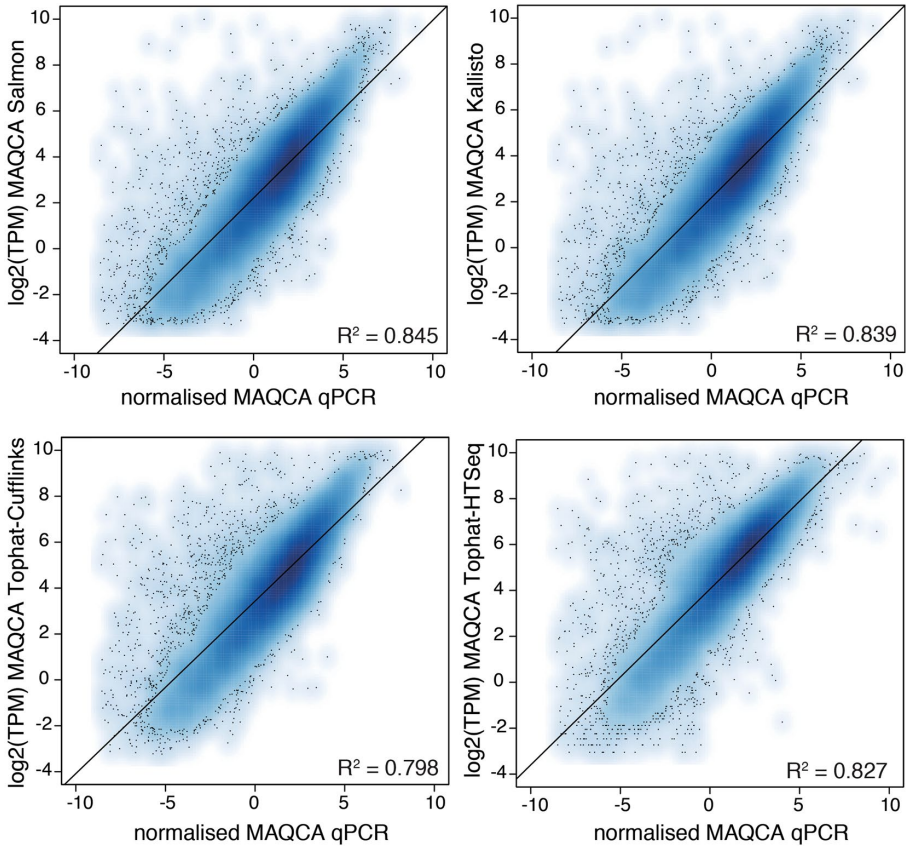
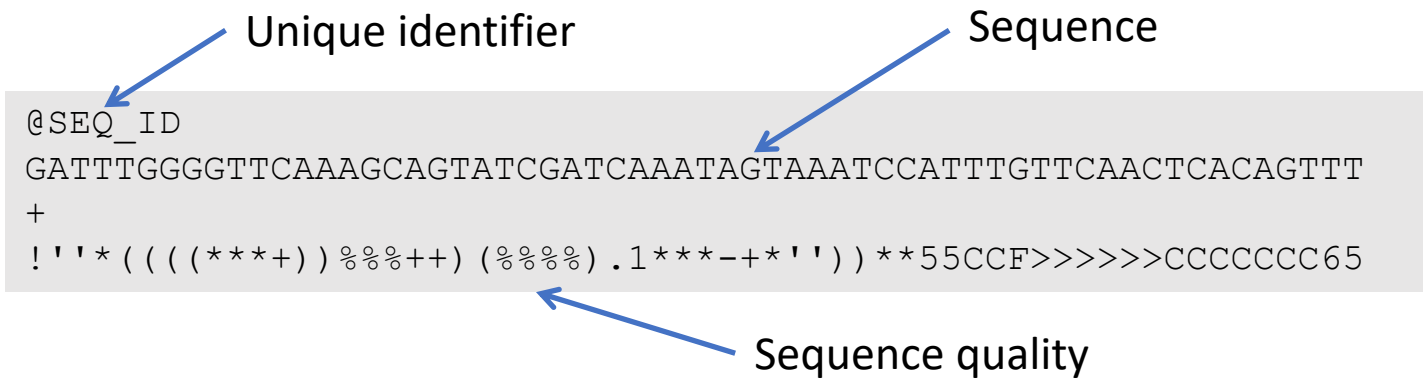


Figure 1. Gene expression correlation between RT-qPCR and RNA-seq data. The Pearson correlation coefficients and linear regression line are indicated. Results are based on RNA-seq data from dataset 1.

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

Fastq – read file format



Paired end data usually in format sampleX_1.fastq and sampleX_2.fastq with same SEQ_ID for both mate pairs, followed by /1 and /2 (or _f and _r)

Sequence quality (phred-score)

Definition [edit]

Phred quality scores Q are defined as a property which is logarithmically related to the base-calling error probabilities P.[2]

$$Q = -10 \log_{10} P$$

or

$$P = 10^{\frac{-Q}{10}}$$

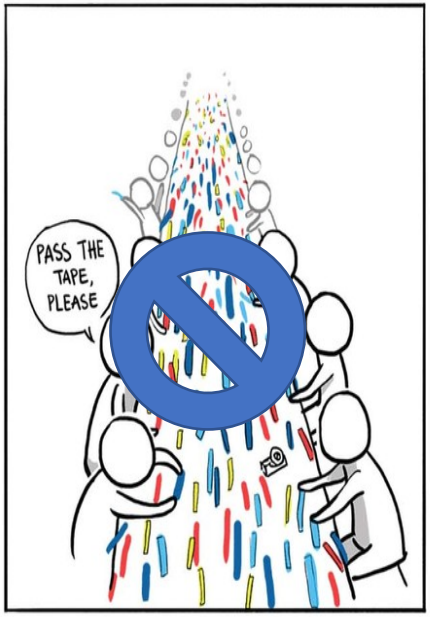
For example, if Phred assigns a quality score of 30 to a base, the chances that this base is called incorrectly are 1 in 1000.

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

The phred quality score is the negative ratio of the error probability to the reference level of $P = 1$ expressed in Decibel (dB).

RNA-sequencing with long reads



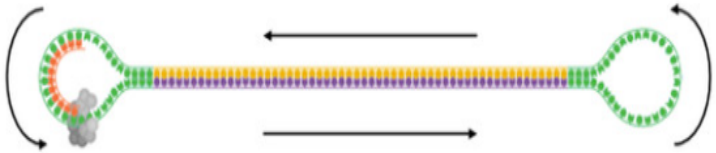
Long read sequencing

- Pacific Biosciences
  - Single molecule sequencing
  - Very long read lengths (up to 30 kb)
  - Rapid sequencing
  - Can detect base modifications (e.g. methylation)
  - Relatively low throughput

- Oxford Nanopore

Pacific Biosciences RSII

PacBio – Sequencing Template



SMRTbell™ Template

Polymerase Read

Definition:

- Sequence of nucleotides incorporated by polymerase while reading a template
- Includes adapters
- Often called "read"
- Includes adapters
- 1 molecule, 1 pol. read

Uses:

- QC of instrument run
- Benchmarking

Subread

Definition:

- Single pass of template
- Adapters removed
- 1 molecule, >=1 subread

Unique data:

- Kinetic measurements
- Rich QVs

Uses:

- Applications

Read (of Insert)

Definition:

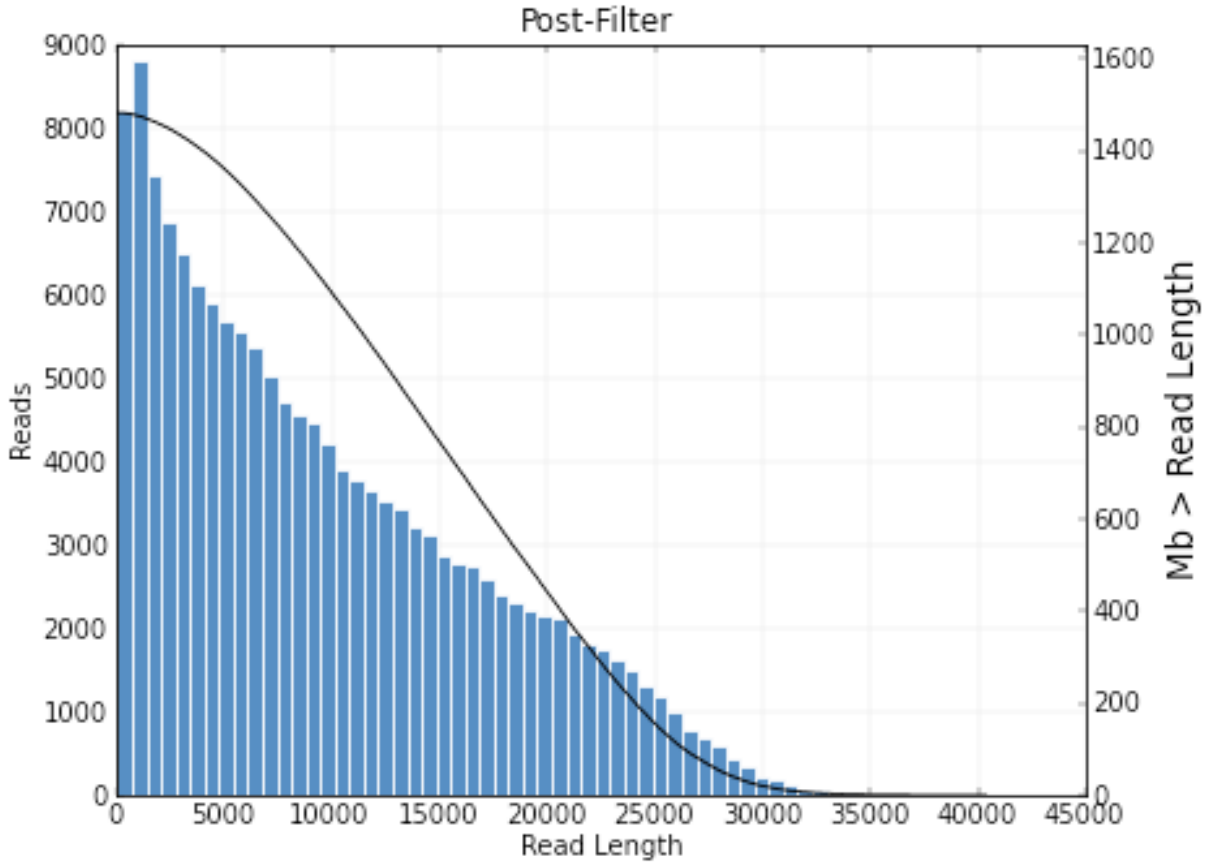
- Represents highest-quality single-sequence for an insert, regardless of number of passes
- Generalizes CCS for <2 passes & RQ <0.9
- 1 or more passes
- 1 molecule, 1 read

Uses:

- Library QC
- Applications

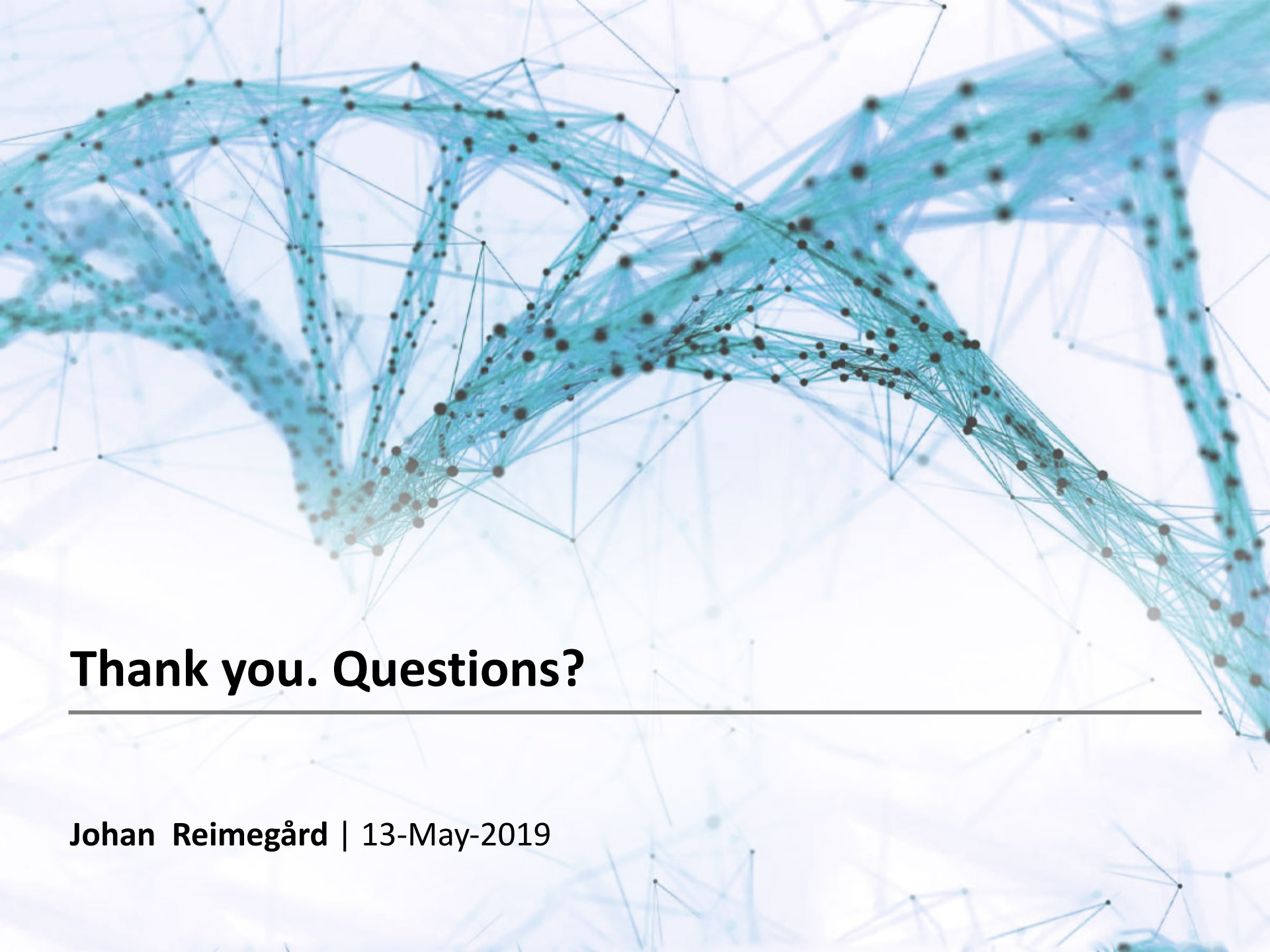
PacBio – Current read lengths

- >10kb average read lengths! (run from April 2014)



Iso-Seq: Full length RNA-seq on PacBio!

- Single molecule sequencing
  - One read – one transcript
- Transcript in full length
  - No assembly required
- No systematic bias
  - CG-rich, AT-rich, tandem repeats



Thank you. Questions?

Johan Reimegård | 13-May-2019