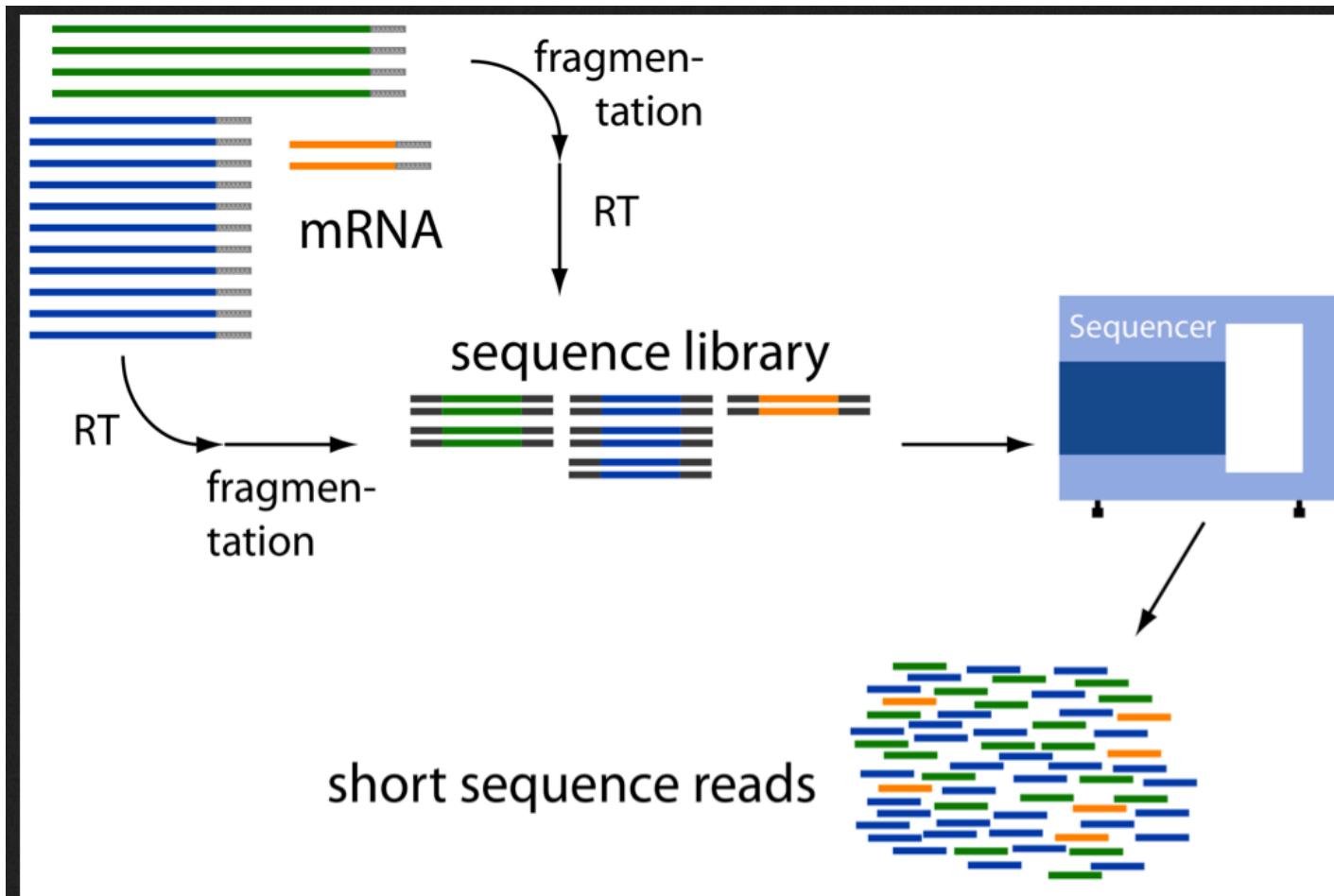


RNA-seq introduction

RNA-seq data analysis

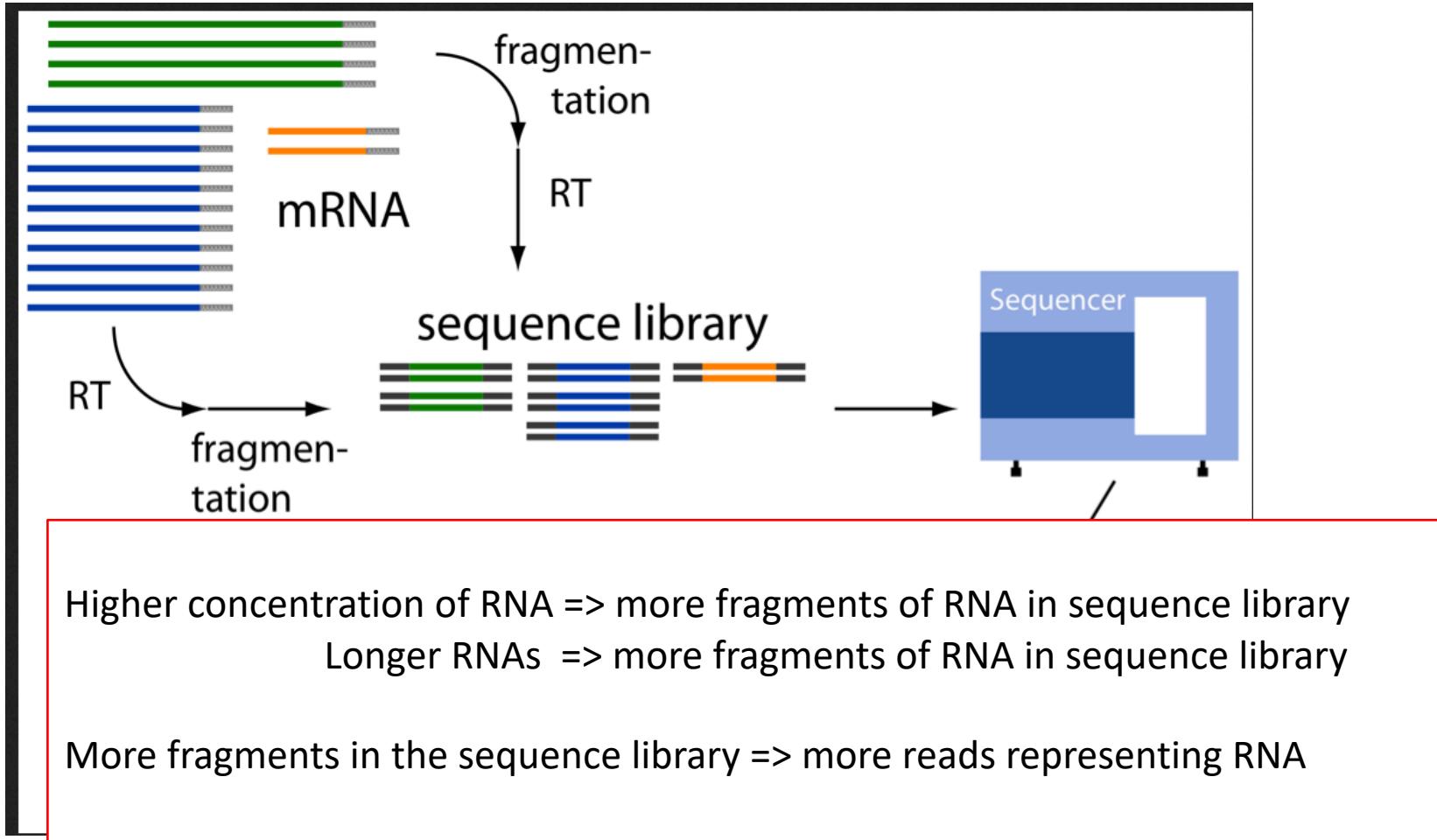
Johan Reimegård | 30-November-2020

How are RNA-seq data generated?



Sampling process

How are RNA-seq data generated?



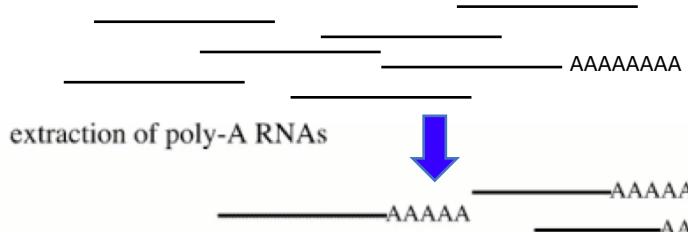
Example

Fragment size in Sequence library = 100
Percent of library being sequenced 10 %

RNA Name	RNA Concentration	Length	Sequence library	Reads
Gene A	10	1000	$\sim 10 * 10 = 100$	10
Gene B	10	100	$\sim 10 * 1 = 10$	1
Gene C	100	100	$100 * 10 = 100$	10

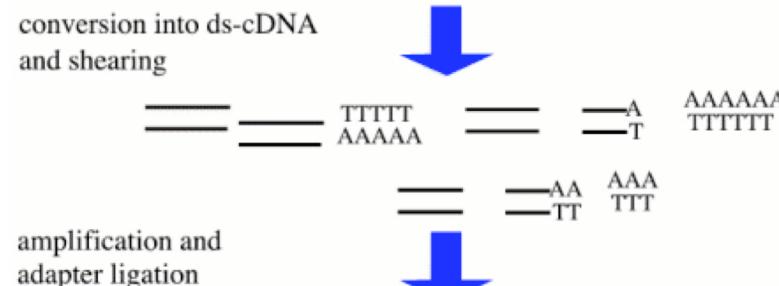
Depending on the different steps you can enrich for your genes of interest

RNA->



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

enrichments ->



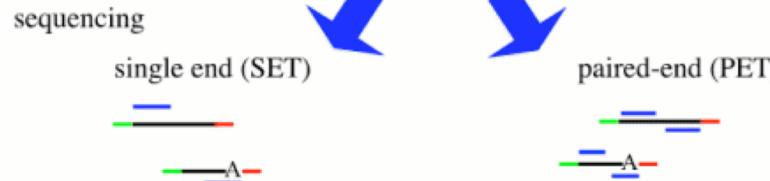
Size of fragment
Strand specific
5' end specific
3' end specific
.....

library ->



Concentration
normalisation

reads ->

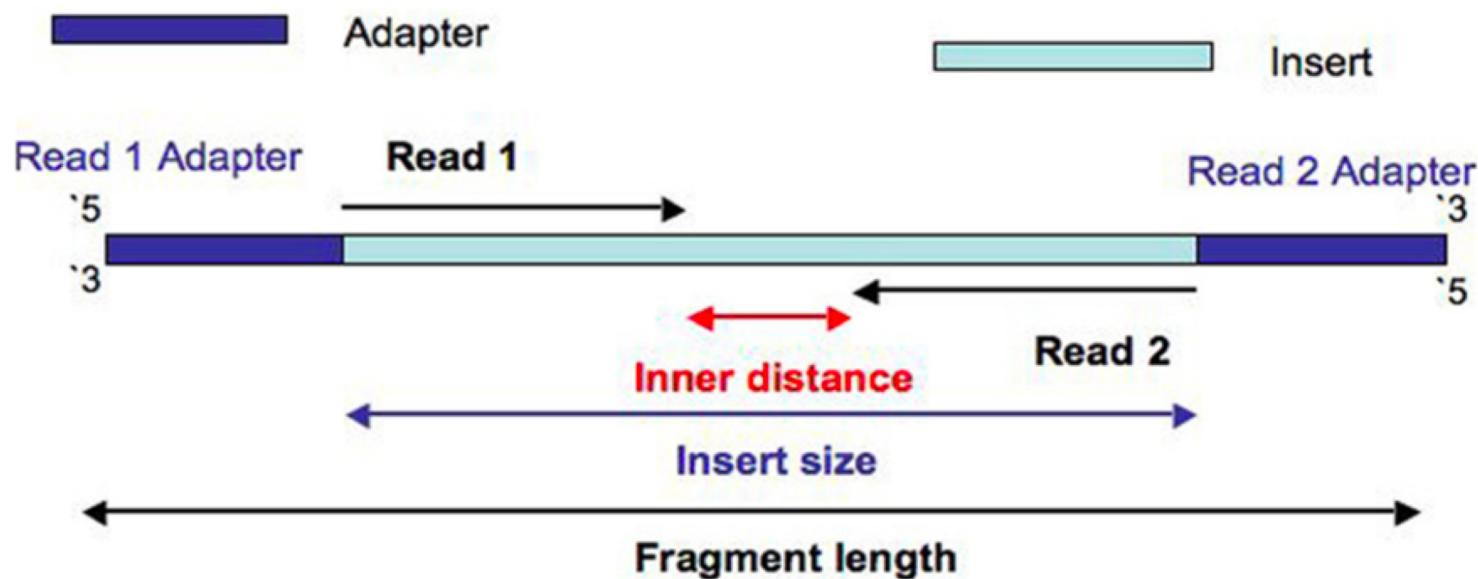


Single end (1 read per fragment)
Paired end (2 reads per fragment)

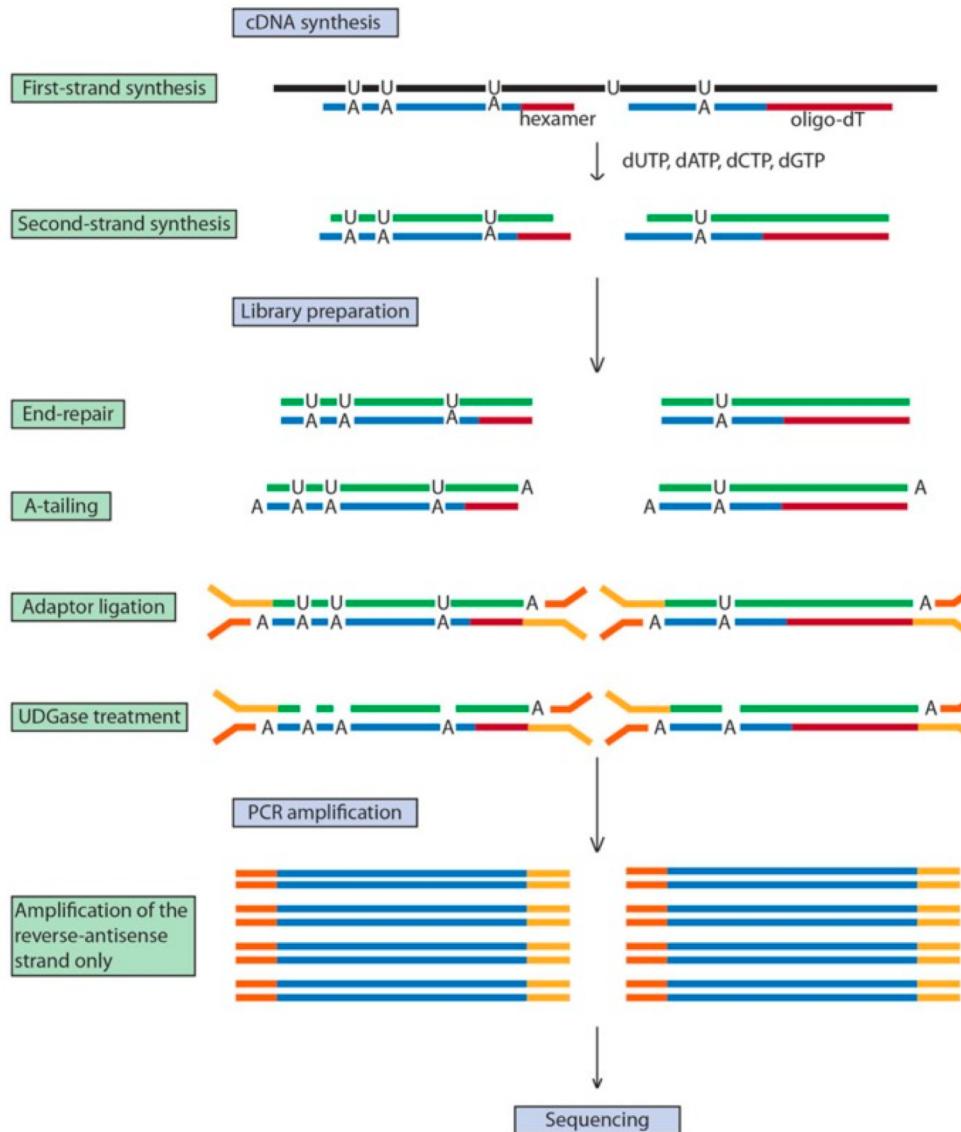
Single end vs paired end reads

Single end only contains one read per fragment (Read 1)

Paired end reads contains two reads per fragment (Read 1 and Read2)



Strand specific sequencing



Fastq – read file format

The diagram illustrates a single line of Fastq file content with three annotations:

- Unique identifier:** Points to the '@SEQ_ID' prefix.
- Sequence:** Points to the sequence of nucleotides: GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTTT.
- Sequence quality (Phred score):** Points to the quality scores represented by ASCII characters: + ! ' ' * (((***+) % % % ++) (% % % %) . 1 *** - + * ' ')) **55CCF>>>>CCCCCCC65.

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)

Fastq – read file format

Unique identifier Sequence
@SEQ_ID GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTTT
+ ! ' ' * (((***+)) % % % + +) (% % % %) . 1 * * * - + * ' ')) * * 55CCF>>>>CCCCCCCC65
Sequence quality (Phred score)

S - Sanger Phred+33, raw reads typically (0, 40)
 X - Solexa Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
 with compressed, interleaved 2-read Segment Quality Control Indicators (bold)

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).

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 complexity and connectivity, suggesting a social network or a complex system.

Thank you

Johan Reimegård