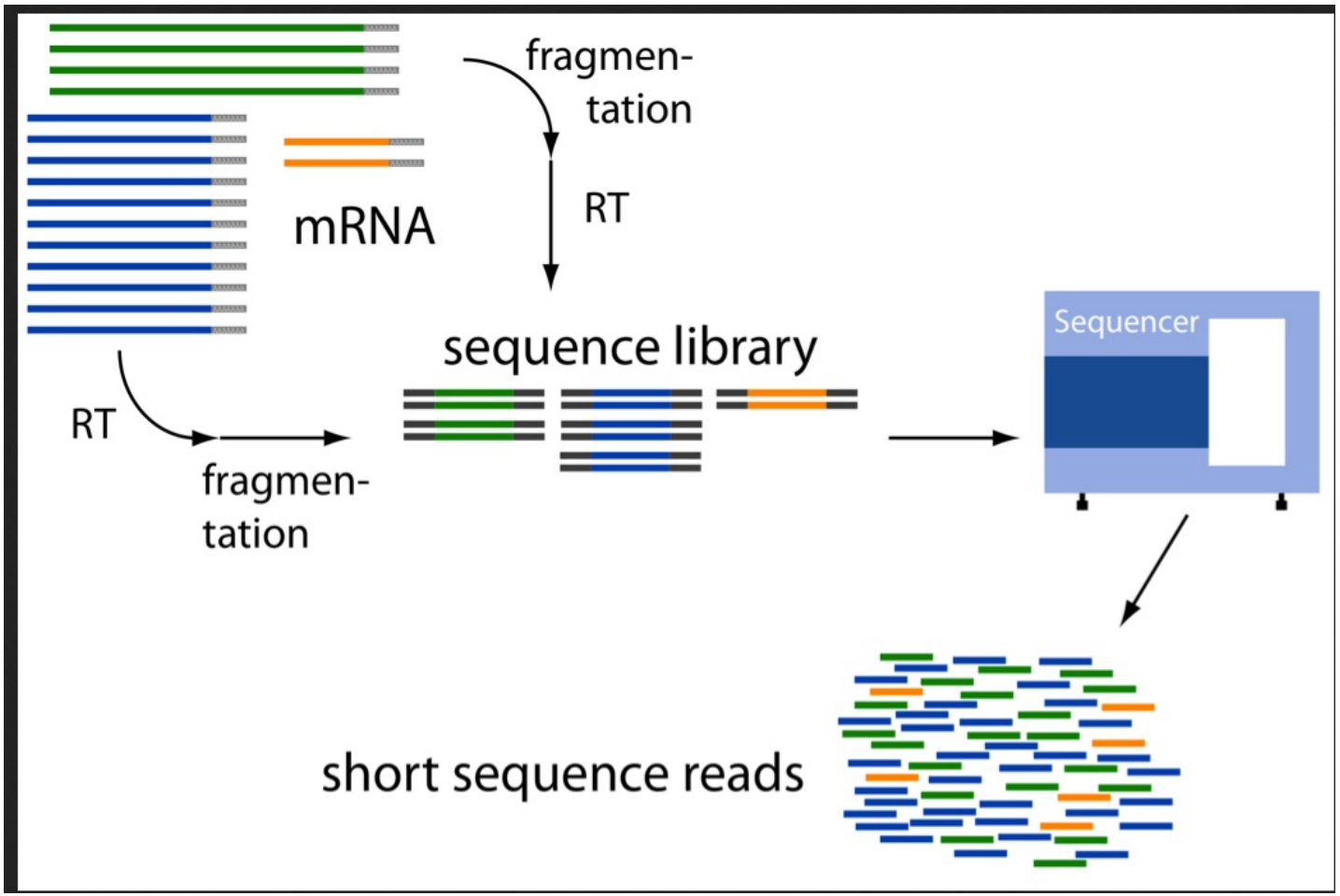
RNA-seq introduction

RNA-seq data analysis

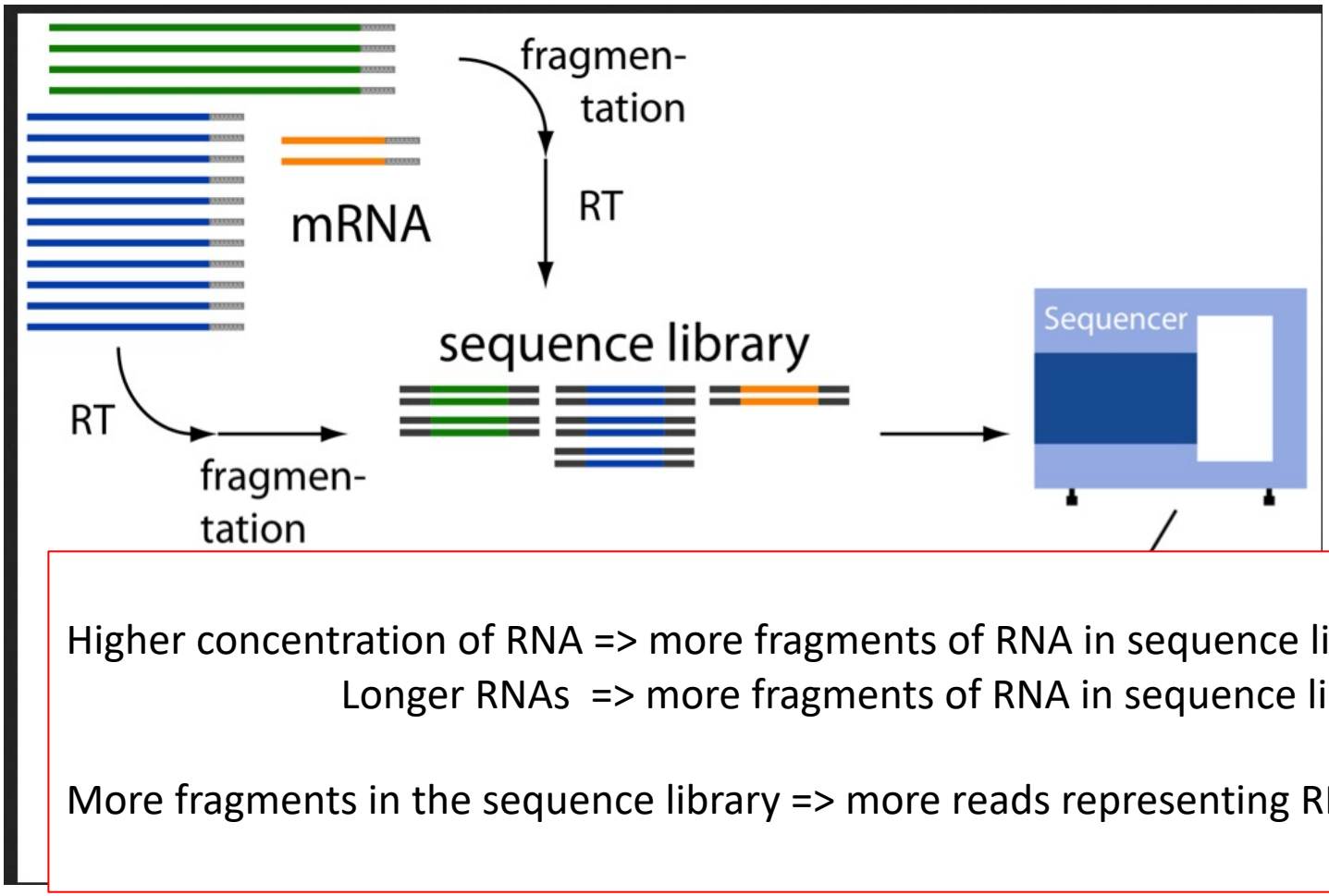
Johan Reimegård | 15-November-2021

How are RNA-seq data generated?



Sampling process

How are RNA-seq data generated?



Higher concentration of RNA => more fragments of RNA in sequence library

Longer RNAs => more fragments of RNA in sequence library

More fragments in the sequence library => more reads representing RNA

Sampling process

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	~10 *10 =100	10
Gene B	10	100	~10 *1 = 10	1
Gene C	100	100	100*10 = 100	10

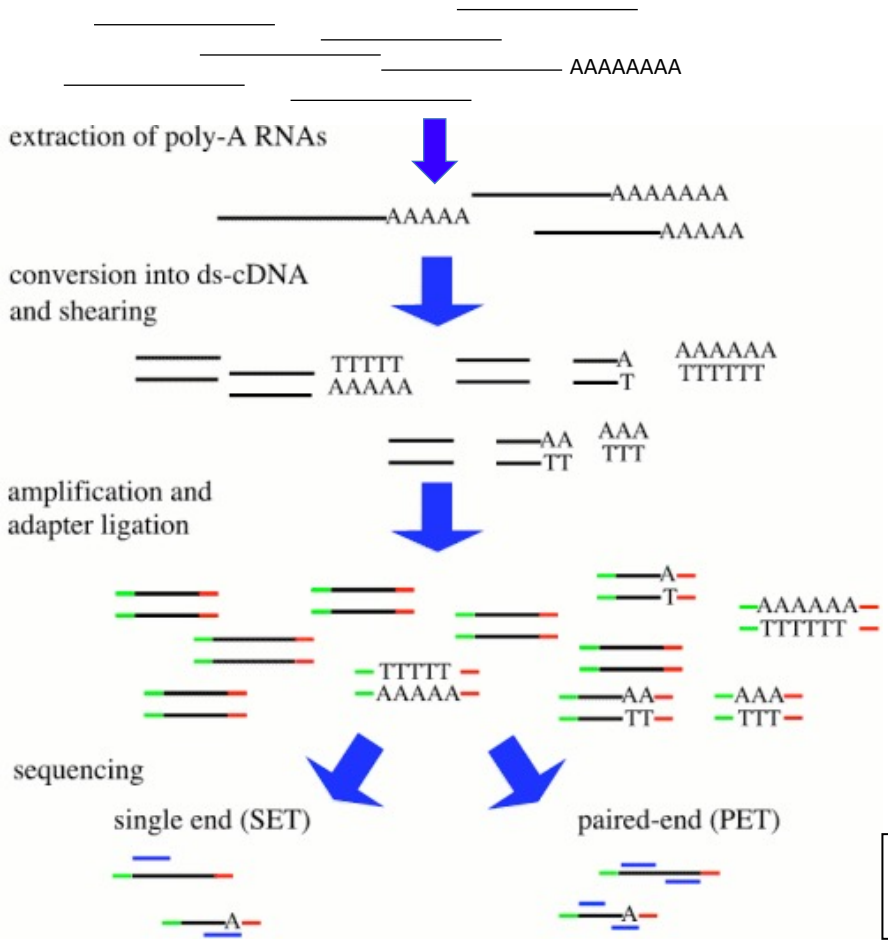
Depending on the different steps you can enrich for your genes of interest

RNA->

enrichments ->

library ->

reads ->



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

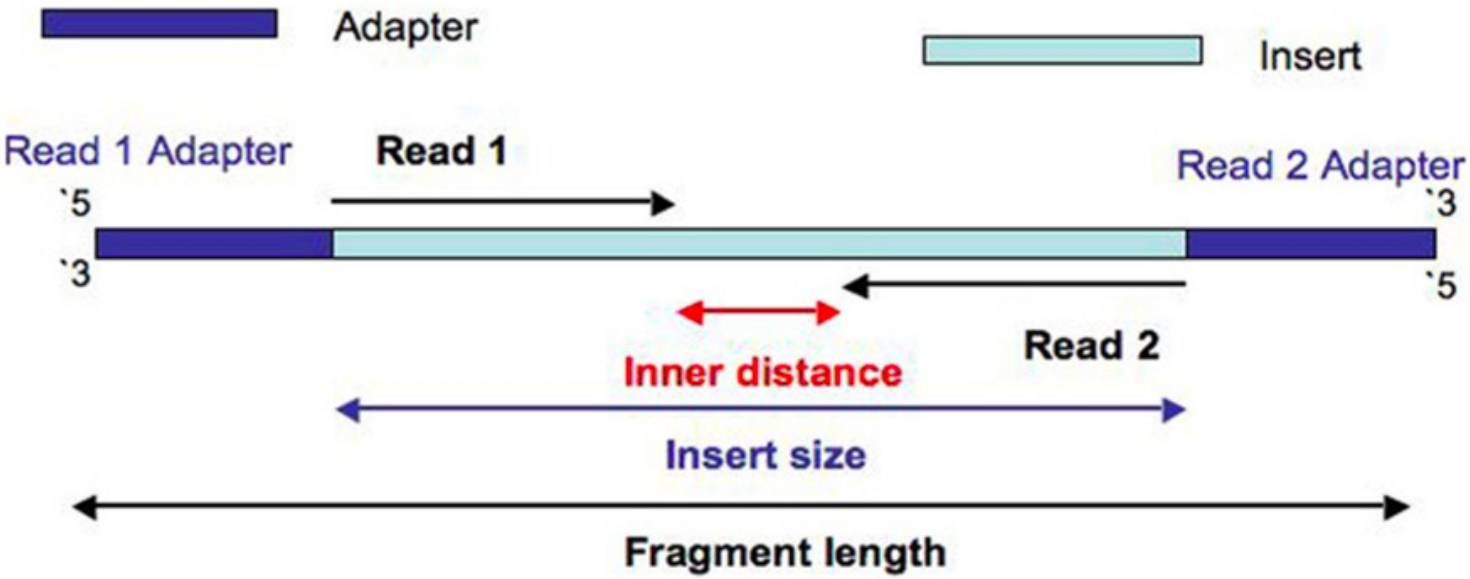
Size of fragment
Strand specific
5’ end specific
3’ end specific
…..

Concentration
normalisation

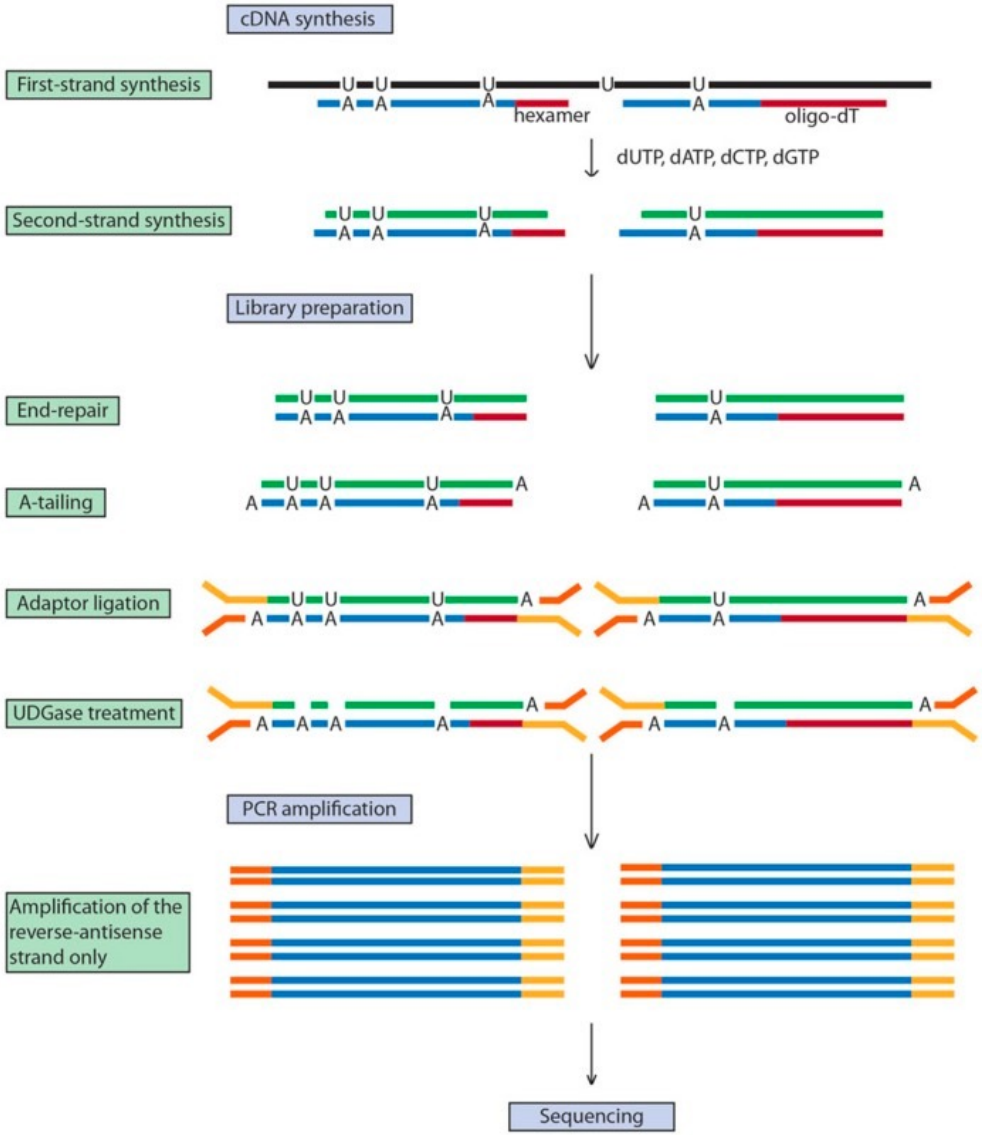
Single end (1 read per fragment)
Paired end (2 reads per fragment)

Single end vs paired end reads

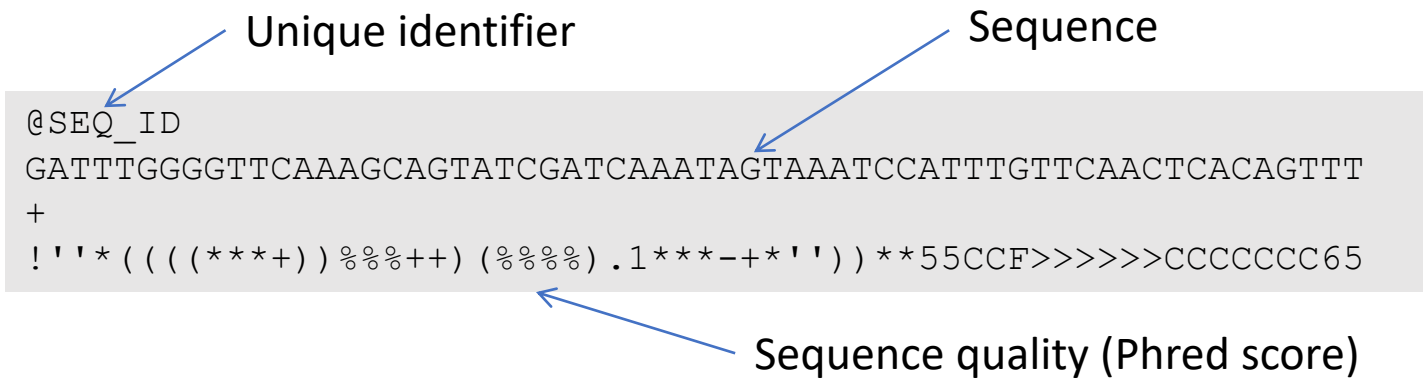
Single end only contains one read per fragment (Read 1)
Paired end reads contains two reads per fragment (Read 1 and Read2)



Strand specific sequencing



Fastq – read file format



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)

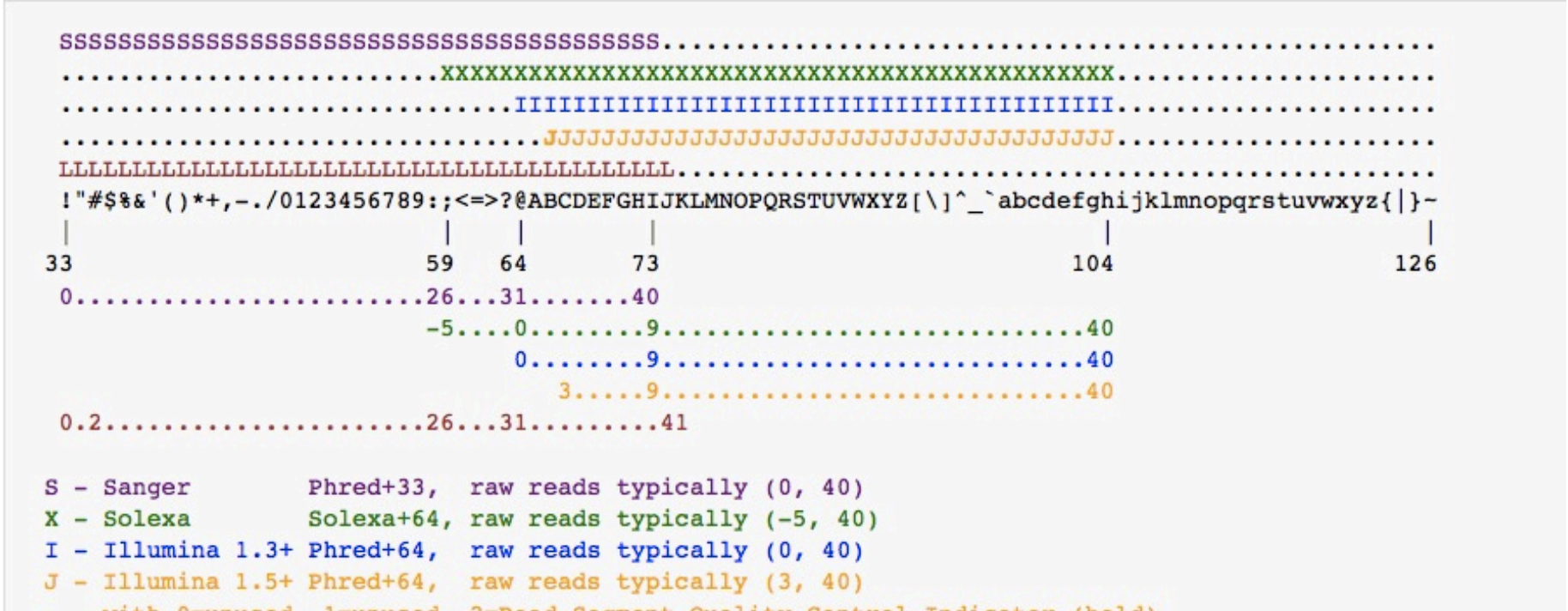
Fastq – read file format

Unique identifier

Sequence

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Sequence quality (Phred score)



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).

Thank you

Johan Reimegård