Quality check and data trimming

Bioinformatics Applications (PLPTH813)

Sanzhen Liu

2/7/2016

NGS Platforms and their features

Platform	Read length	reads per run	Error type	Error rate
Illumina Miseq	up to 2x300	Up to 10 Gb	substitutions	~0.1-1%
Illumina Hiseq	up to 2x250	Hundreds of Gb	substitutions	~0.1-1%
Ion Torrent	300-500	10 Gb?	InDel	>1%
PacBio	Average >5,000	500 Mb – 1 Gb	InDel	~15%
Nanopore	>1,000	? Mb per MinION	Del?	~10-20%?

Outline

- Sequence data format
- Sequence quality
- Quality checking
- Quality trimming
- Adaptor trimming

FASTA

Sequence FASTA file

```
>SEQ_ID
ATCAACTGATGCATC
```

Quality FASTA file

```
>SEQ_ID
28 30 33 34 33 35 38 37 36 35 38 35 36 36 30
```

Phred quality score

Q = -10 x $\log_{10}(p)$

p = $10^{-Q/10}$

where Q is quality score and p is the probability of error

1) What does “Q = 30” indicate?
2) What is the quality score of a base call with p = 0.01?

FASTQ

Single-end FASTQ file

```
@SEQ_ID
ATCAACTGATGCATC
+SEQ_ID
!''*((((***+))%
```

Paired-end FASTQ files

File1: forward seq

```
@SEQ_ID/1
ATCAACTGATGCATC
+
!''*((((***+))%
```

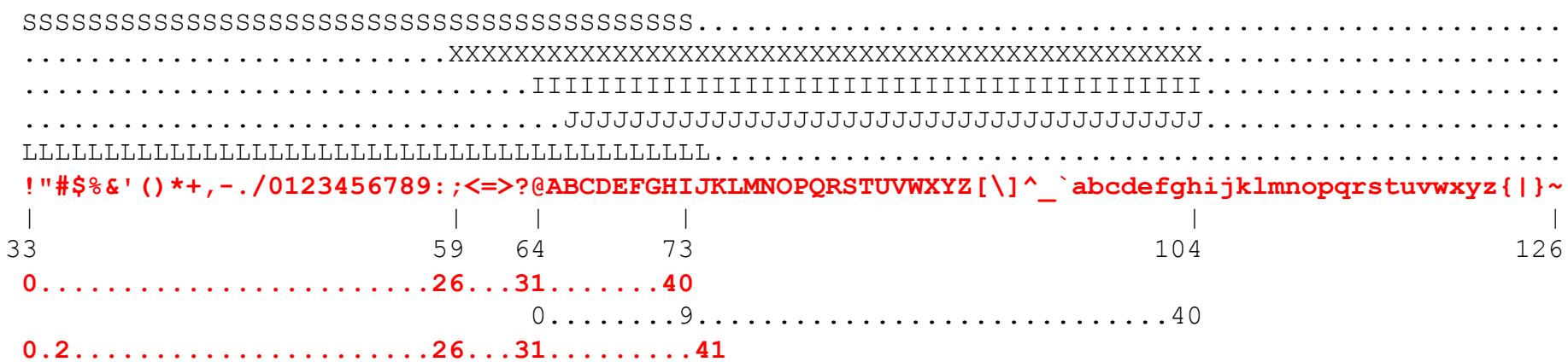
File2: reverse seq

```
@SEQ_ID/2
GATTTGGGGTTCCTG
+
)(%%%%).1***-+*
```

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

Quality codes in FASTQ



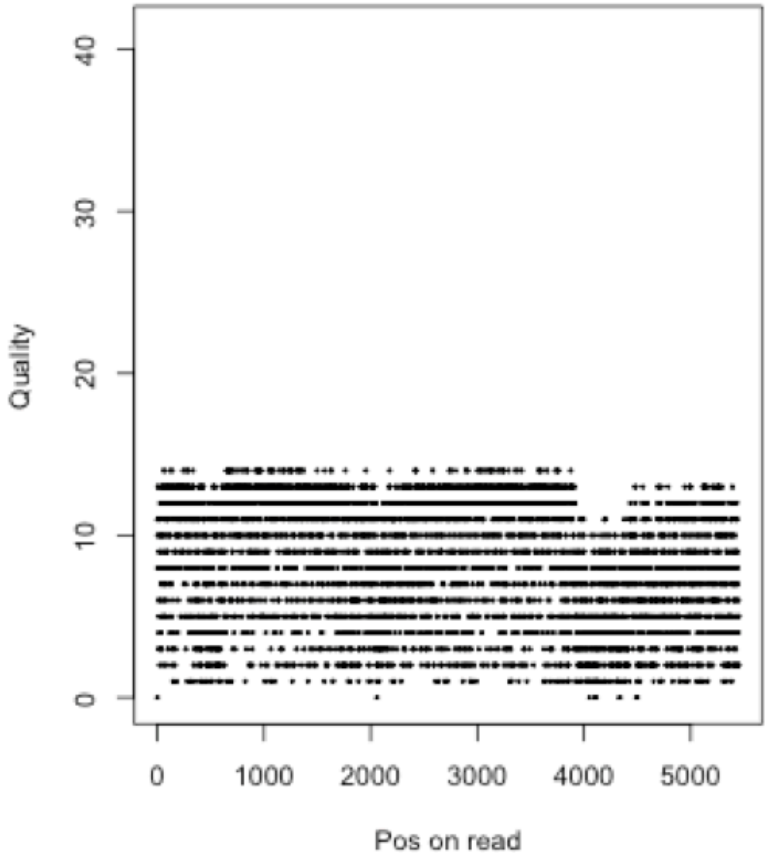
Sanger **Phred+33, raw reads typically (0, 40)**

Illumina 1.3+ Phred+64, raw reads typically (0, 40)

Illumina 1.8+ Phred+33, raw reads typically (0, 41)

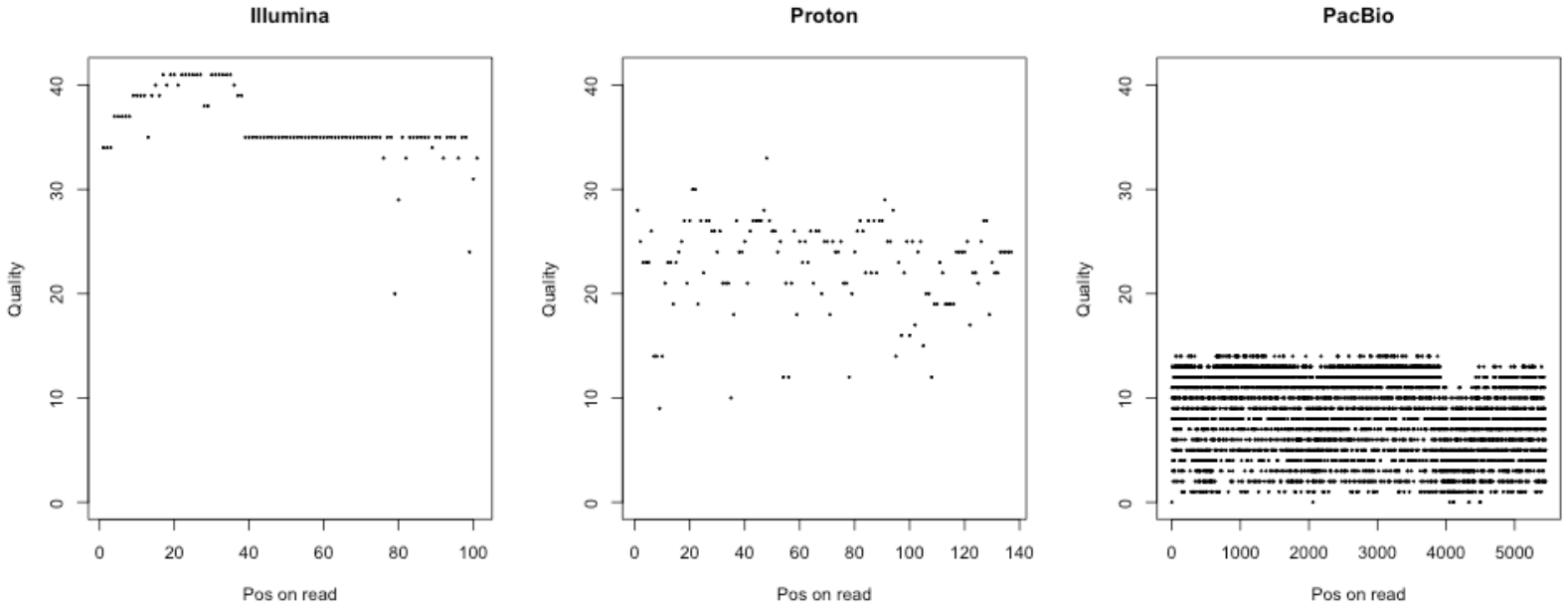
Source: en.wikipedia.org/wiki/FASTQ_format

What platform was this read generated from?



Average quality?
Average probability of error rate?

Typical reads in different platforms



Read length
Read quality

Overview sequencing data

Data QC - FASTQC

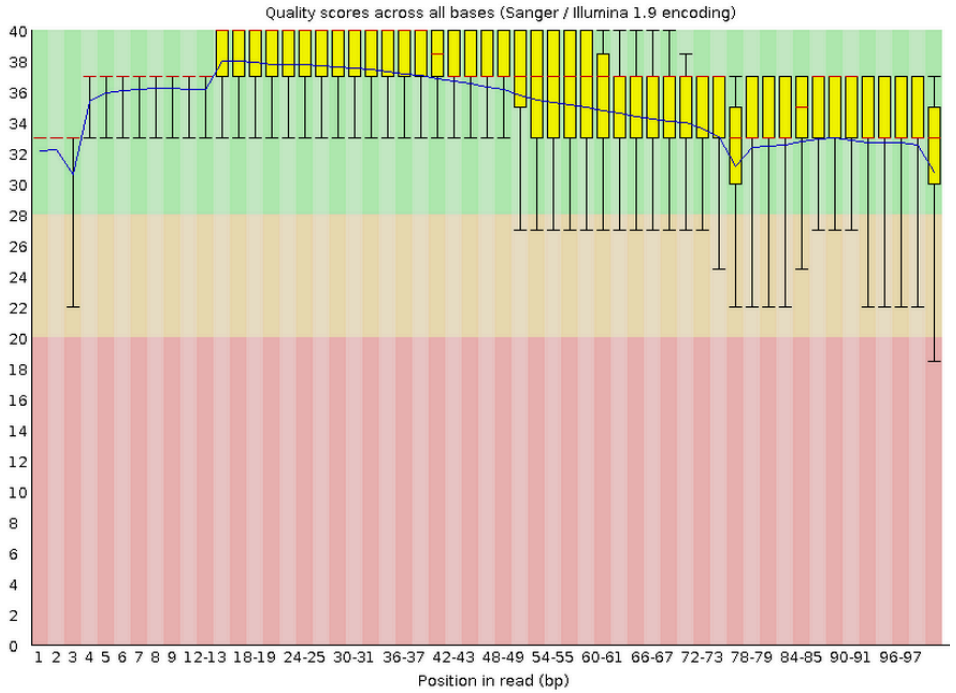
FASTQC is a popular tool to examine the quality of sequencing data

- Easy to run: `fastqc example.fastq`
- Rich output information
- Output present in the html format

Basic Statistics

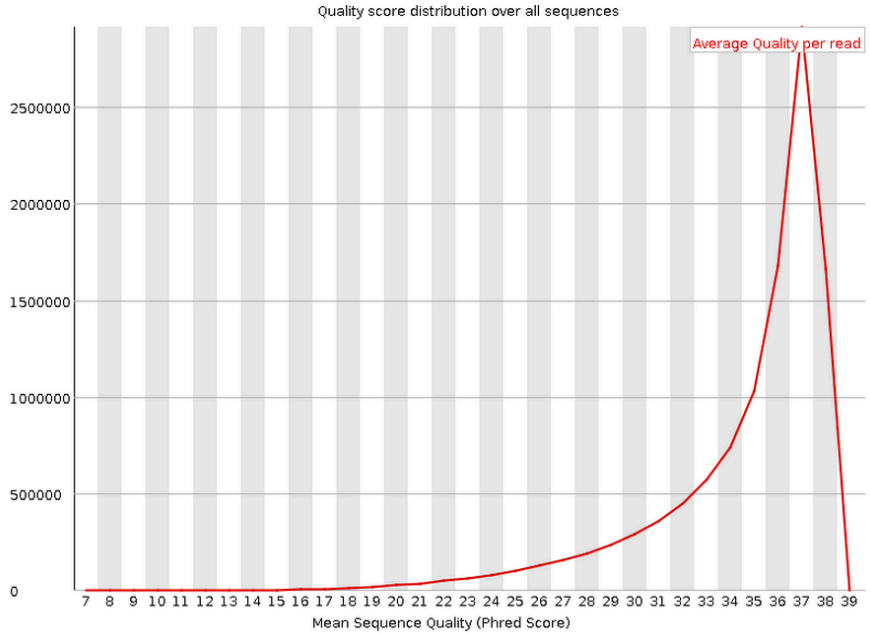
Measure	Value
Filename	example.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	10856448
Sequences flagged as poor quality	0
Sequence length	101
%GC	53

Per base sequence quality

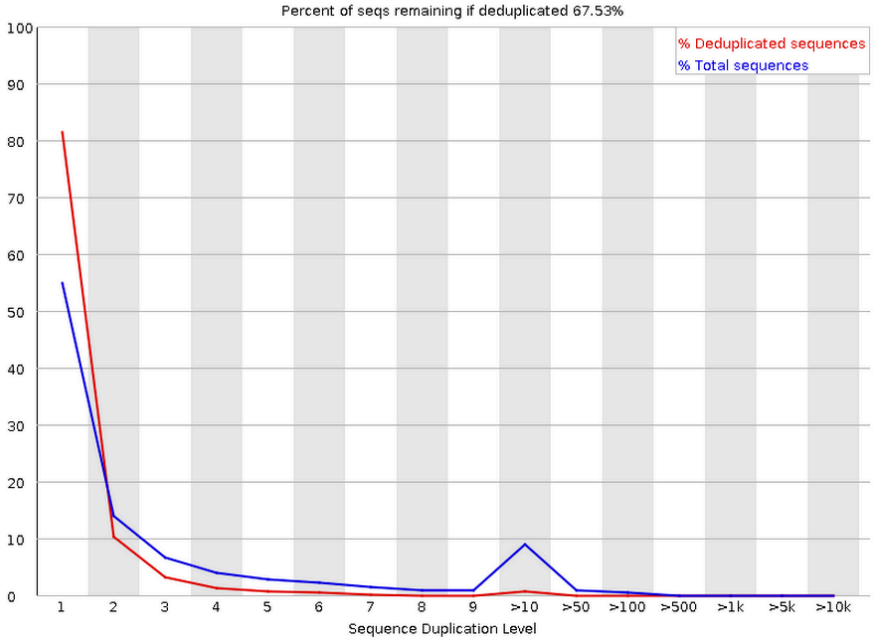


FASTQC – (2)

Per sequence quality scores

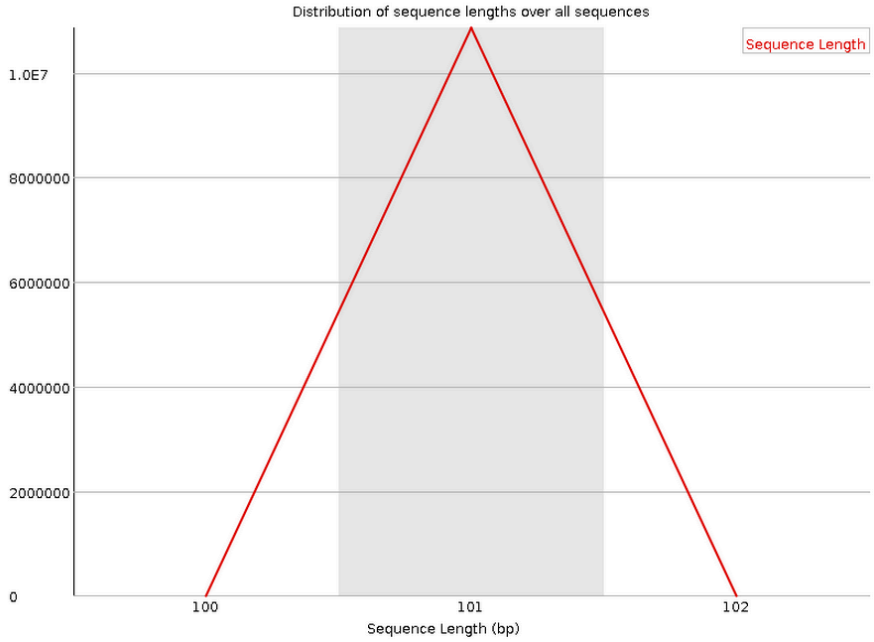


Sequence Duplication Levels

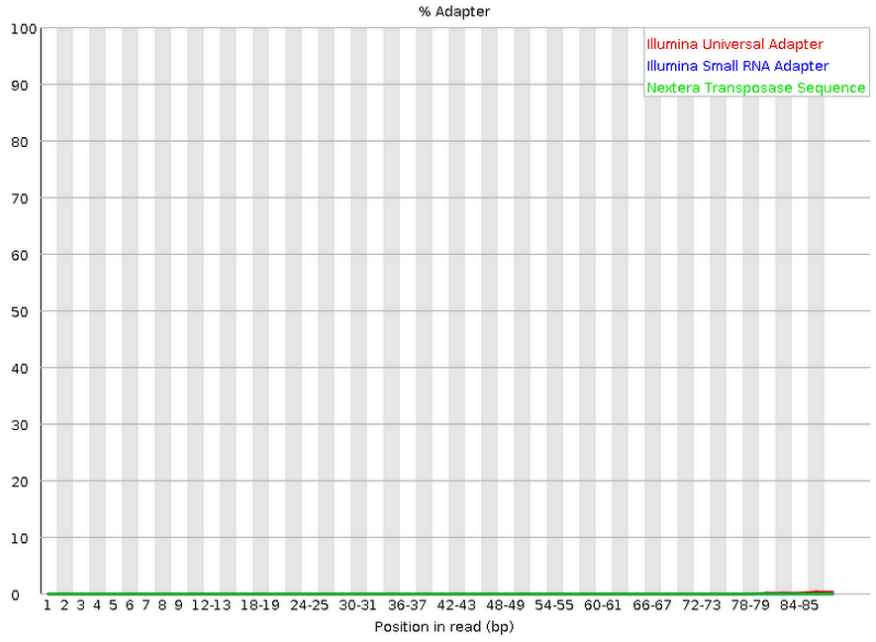


FASTQC (3)

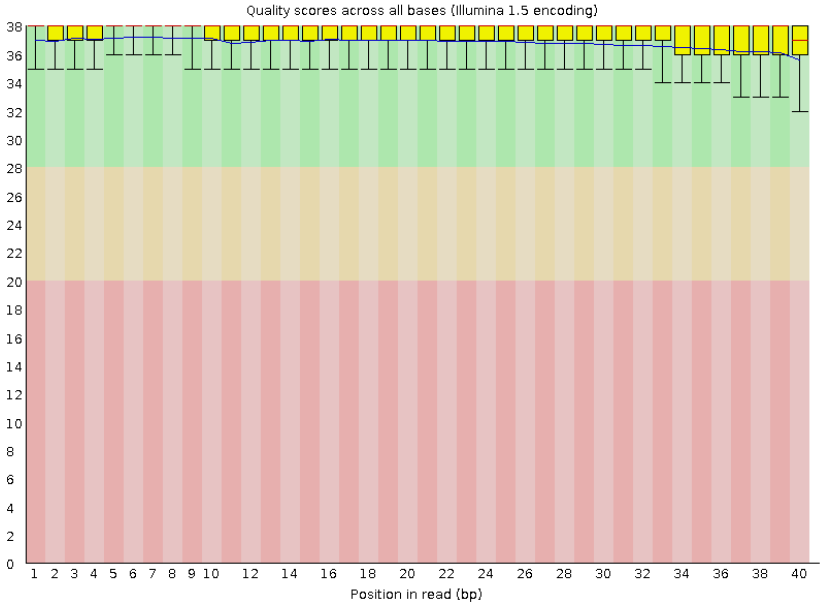
Sequence Length Distribution

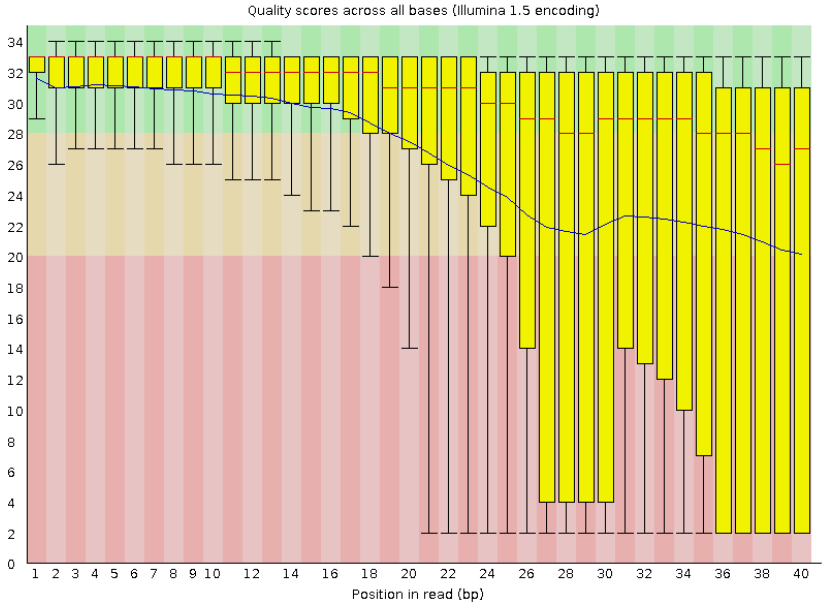


Adapter Content



Good and Bad data





More information, please read:
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Tools for FAST[AQ] - seqtk

seqtk is a fast tool for processing sequences in the FASTA/Q format.

seq	common transformation of FASTA/Q
comp	get the nucleotide composition of FASTA/Q
sample	subsample sequences
subseq	extract subsequences from FASTA/Q
fqchk	fastq QC (base/quality summary)
mergepe	interleave two PE FASTA/Q files
trimfq	trim FASTQ using the Phred algorithm
mutfa	point mutate FASTA at specified positions
mergefa	merge two FASTA/Q files
dropse	drop unpaired from interleaved PE FASTA/Q
cutN	cut sequence at long N
randbase	choose a random base from hets
hety	regional heterozygosity
listhet	extract the position of each het

seqtk examples (1)

- Conversion of a FASTQ to a FASTA

```
seqtk seq -A in.fq > out.fa
seqtk seq -A in.fq.gz > out.fa
```

- Reverse complement FASTA/Q:

```
seqtk seq -r in.fq > out.fq
```

- Extract sequences with names in file name.lst, one sequence name per line:

```
seqtk subseq in.fq name.lst > out.fq
```

github.com/lh3/seqtk

seqtk examples (2)

- Subsample 10,000 read pairs from two large paired FASTQ files

#(remember to use the same random seed to keep pairing):

```
seqtk sample -s100 read1.fq 10000 > sub1.fq
seqtk sample -s100 read2.fq 10000 > sub2.fq
```

- Trim 5bp from the left end of each read and 10bp from the right end:

```
seqtk trimfq -b 5 -e 10 in.fa > out.fa
```

- Trim low-quality bases from both ends using the Phred algorithm:

```
seqtk trimfq in.fq > out.fq
```

Quality trimming

github.com/lh3/seqtk

Sequence trimming

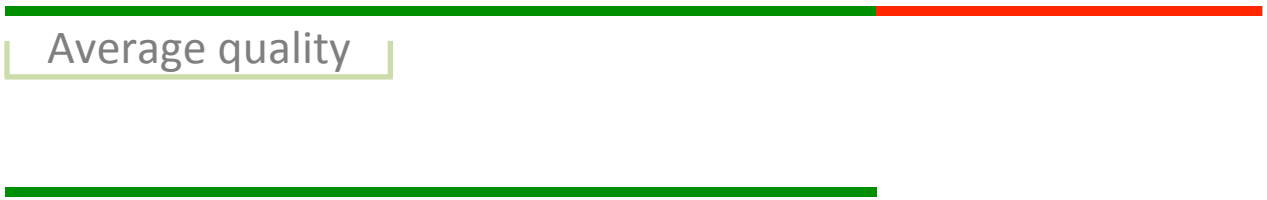
- **Quality trimming**: to remove low quality sequences
- **Adaptor trimming**: to remove adaptor contamination

Quality trimming

- **Window scan method**

the major steps in the quality trimming process involve calculating average quality within certain windows along the sequence

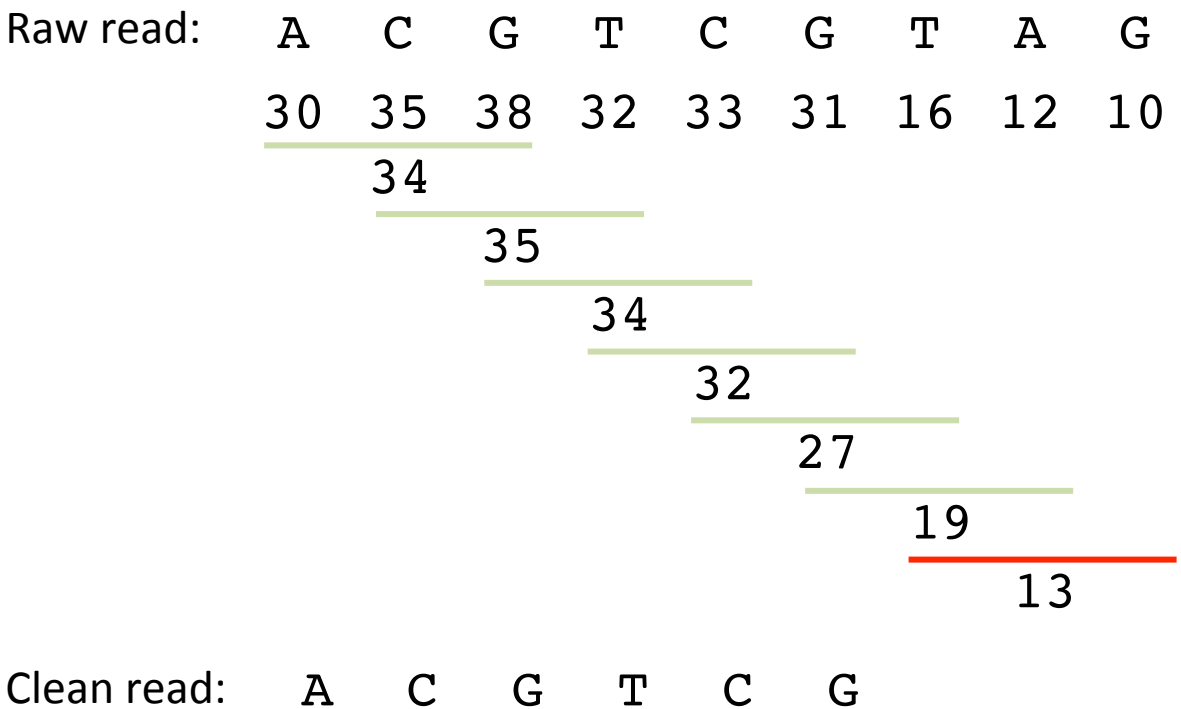
1. Sliding windows (window size and step size)
2. Maximum average errors (minimum average quality)



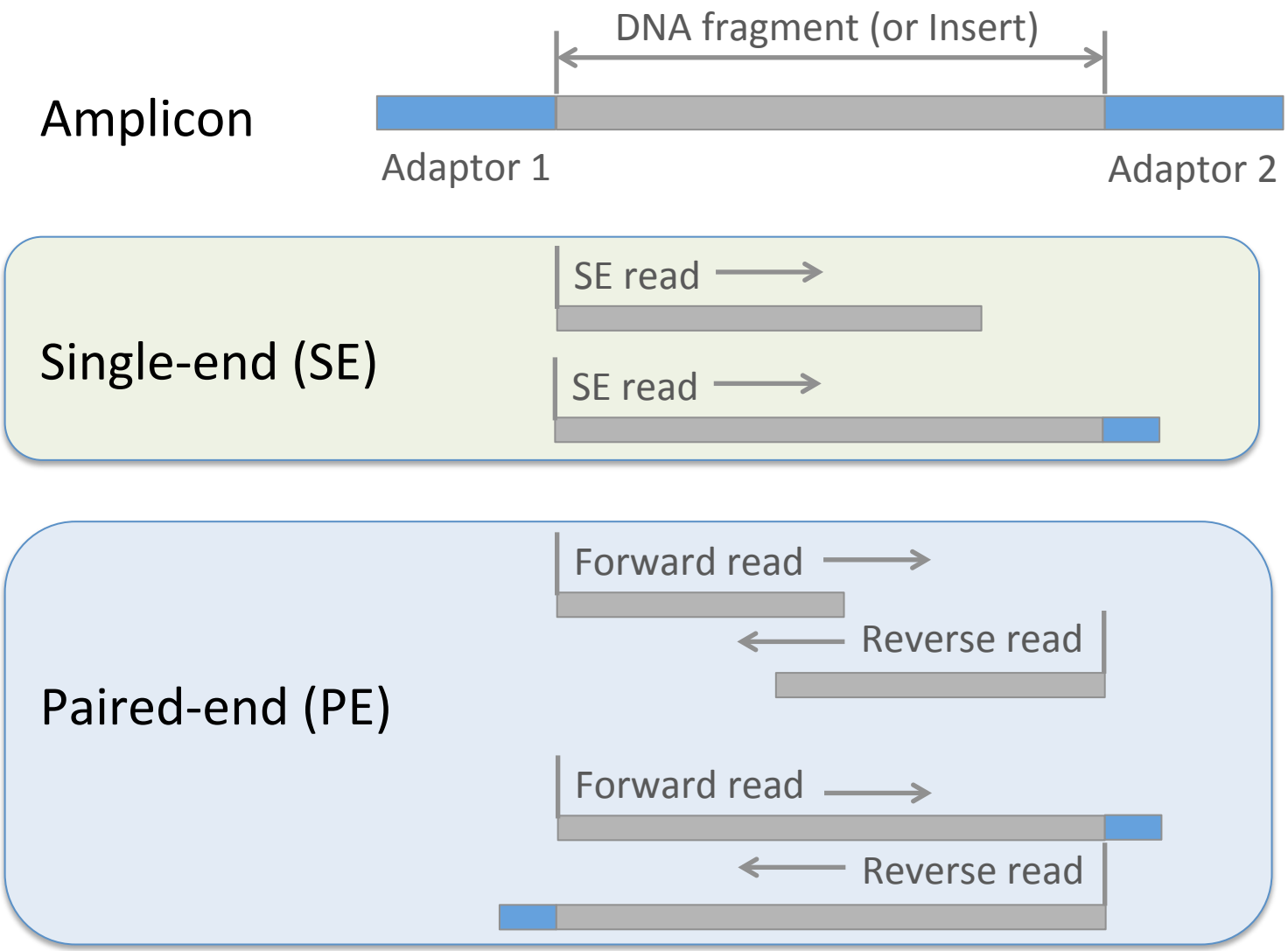
Bioinformatics, 2001, 17:1093-1104

Quality trimming example

1. Window = 3 bp
2. Step = 1 bp
3. Minimum average quality score = 15



Adaptor contamination and trimming



Trimmomatic – an innovative trimming tool

BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 15 2014, pages 2114–2120
doi:10.1093/bioinformatics/btu170

Genome analysis

Advance Access publication April 1, 2014

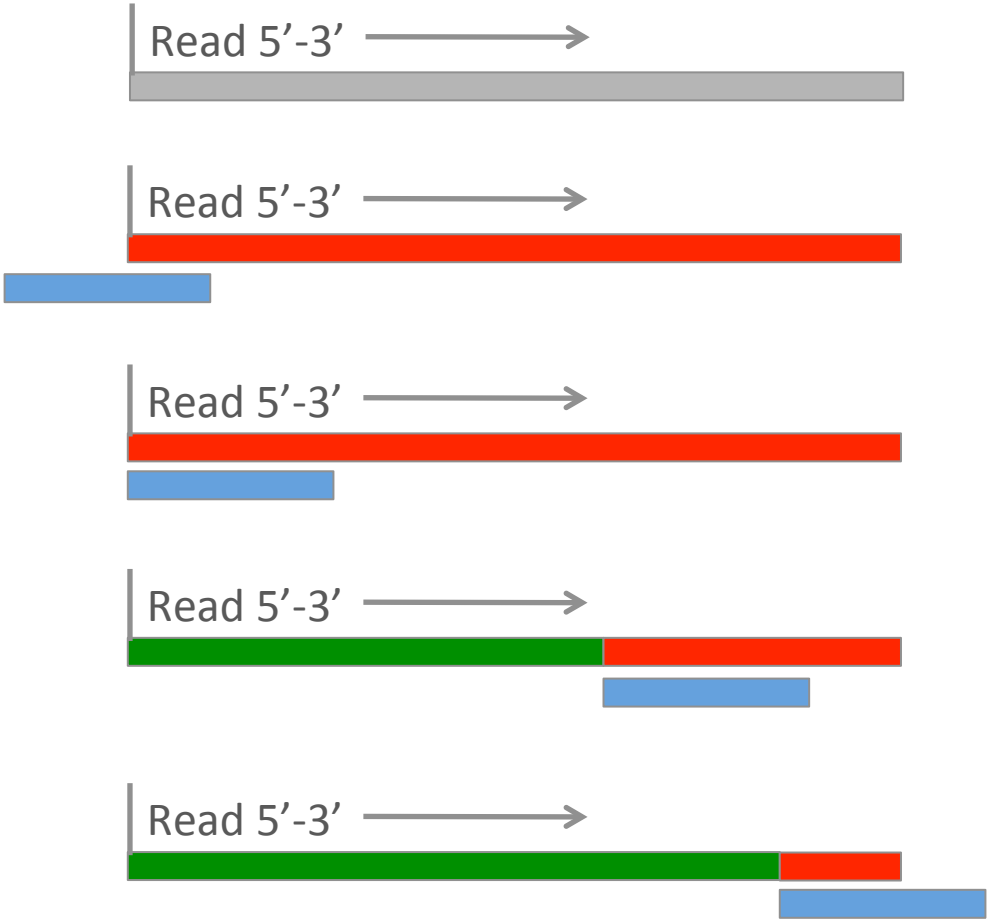
Trimmomatic: a flexible trimmer for Illumina sequence data

Anthony M. Bolger[1,2], Marc Lohse[1] and Bjoern Usadel[2,3,*]

[1]Department Metabolic Networks, Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm,[2]Institut für Biologie I, RWTH Aachen, Worringer Weg 3, 52074 Aachen and [3]Institute of Bio- and Geosciences: Plant Sciences, Forschungszentrum Jülich, Leo-Brandt-Straße, 52425 Jülich, Germany

Associate Editor: Inanc Birol

Trimmomatic – simple mode



If an alignment was identified, the alignment region plus the remainder after the alignment are removed.

Valid sequence

Trimmed sequence

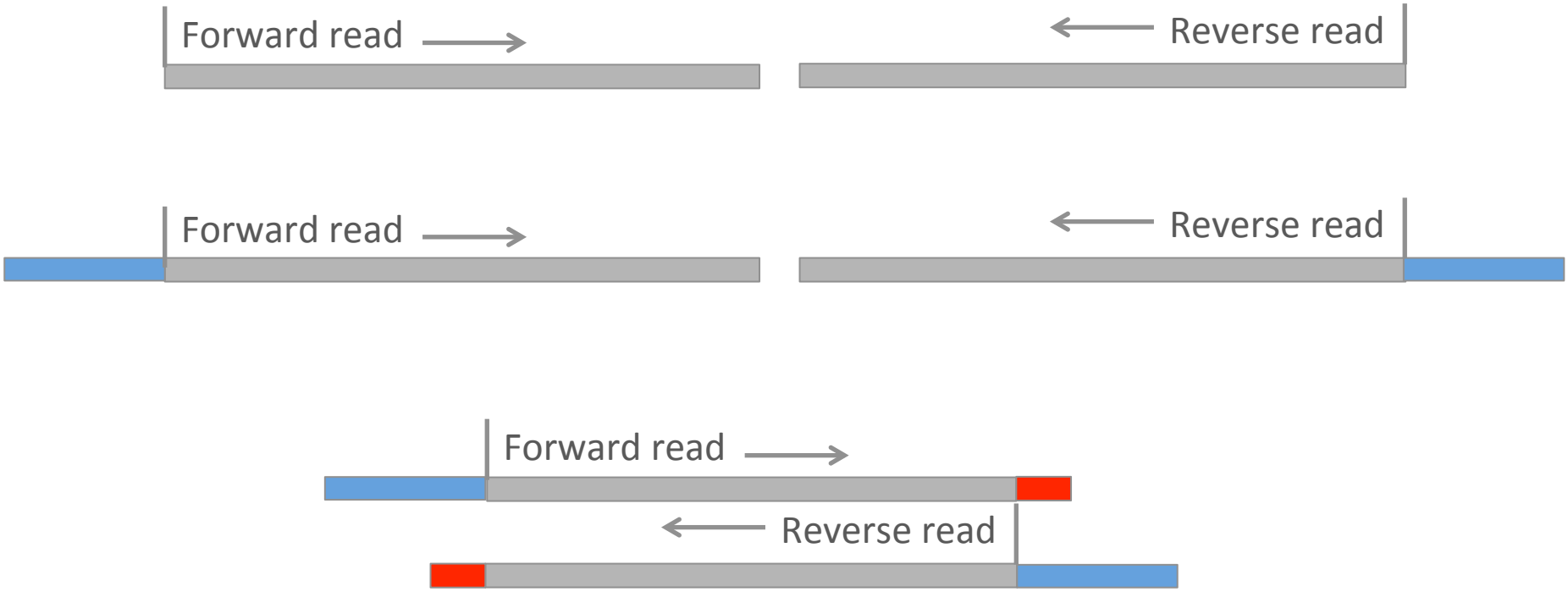
Simple mode: pro and cons

- Simple mode has the advantage that it can detect any technical sequence at any location in the read, provided that the alignment is sufficiently long and the read is sufficiently accurate.

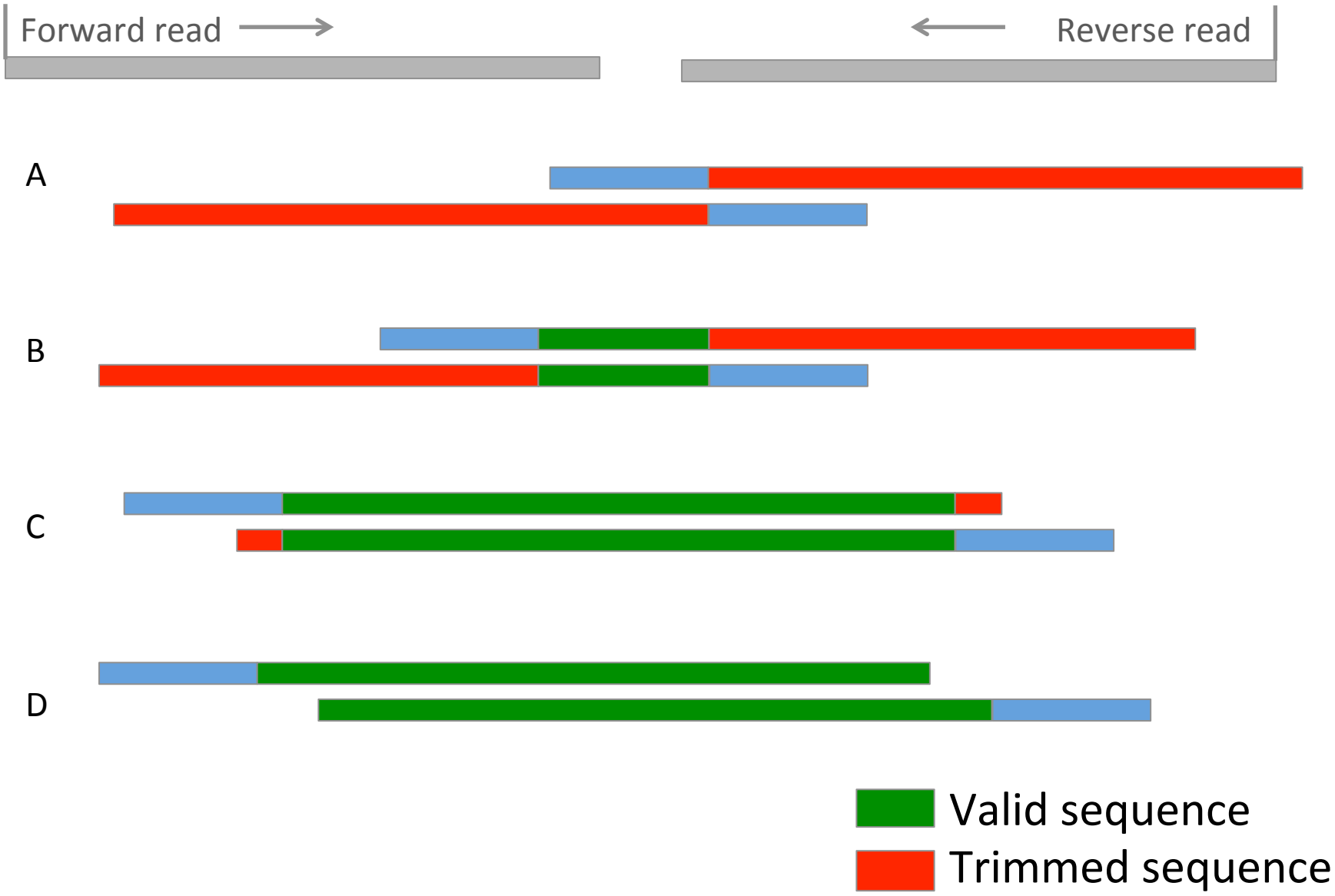
- Issue: if the adaptor sequence on the read is too short to make the alignment, the adaptor sequence can not be trimmed.

Trimmomatic – panlindrome mode

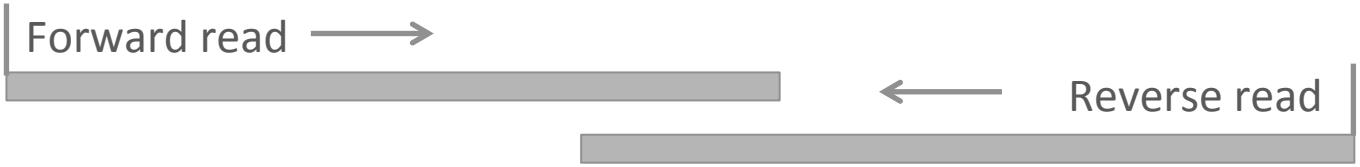
- the main algorithmic innovation is to identify adapter sequences through making use of paired information



Trimmomatic – panlindrome mode

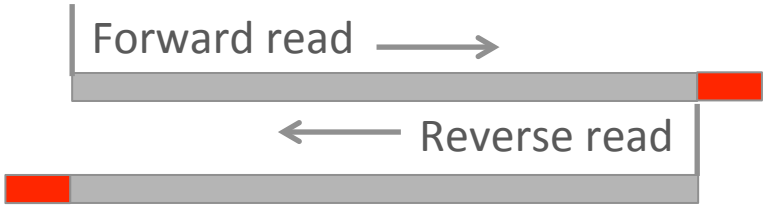


Trimmomatic – panlindrome mode



Three pre-required features:

1. both reads in a pair consist of an equal number of valid bases
2. the valid sequence of the two reads are reverse complements
3. the valid sequence of two read are followed by contaminating sequence from the “opposite” adapters



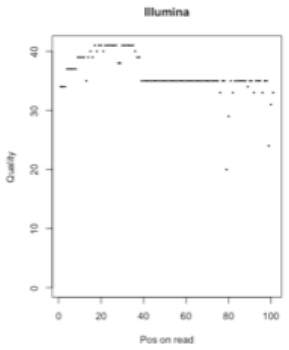
Quality trimming in Trimmomatic

- Sliding window quality trimming (SW)

Scan reads from the 5' end of the read, and remove the 3' end of the read when the average quality of a group of bases drops below a specified threshold.

- Maximum Information (MI)

The trimming process becomes increasingly strict as it progresses through the read, rather than to apply a fixed quality threshold.



Output from Trimmomatic

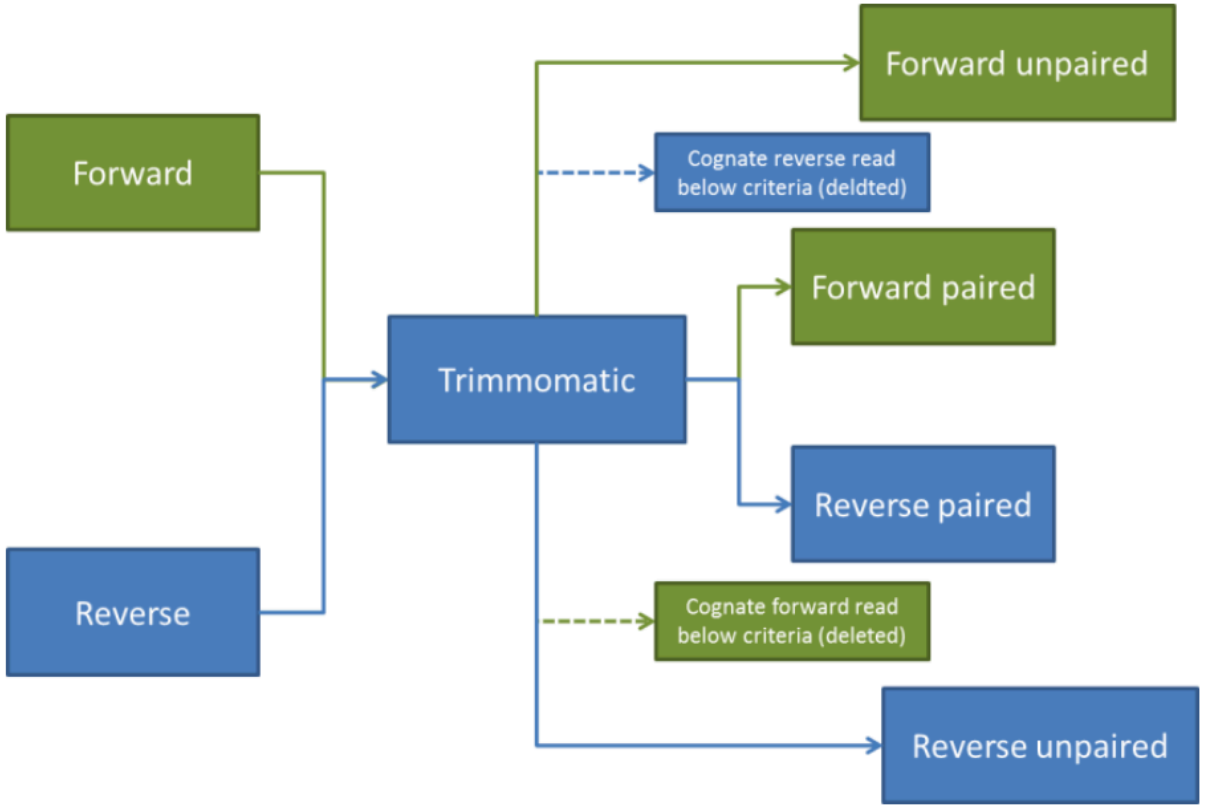


Figure 1: Flow of reads in Trimmomatic Paired End mode

Some convenience features of Trimmomatic

- Accept compressed input (gzip or bzip2)
- Automatically determine quality format
- Use multiple threads if multiple CPU cores are available
- Provide trimming log for each read

Comparison among trimming software packages

Table 1 Main features of various adapter trimmers

	Adapter trimming						Quality control		Other			
Method	**5′**	**3′**	**SE**	**PE**	**LMP**	**Multi**	**Ns**	**Q**	**Barcode**	**Merge**	**gzip Files**	**MT**
FastX	×	○	○	×	×	×	○	×	○	×	×	×
SeqTrim	×	○	○	×	×	○	○	○	×	×	○	○
TagCleaner	○	○	○	×	×	×	×	×	×	×	×	×
EA-Tools	×	○	○	○	×	×	○	○	○	×	○	×
Cutadapt	○	○	○	○	×	○	×	○	×	×	○	×
TrimGalore	×	○	○	○	×	×	×	○	×	×	○	×
SeqPrep	×	○	×	○	×	×	×	×	×	○	×	×
Btrim	○	○	○	○	×	×	×	○	○	×	×	×
Scythe	×	○	○	×	×	×	×	×	×	×	○	×
Flexbar	○	○	○	○	×	○	○	○	○	×	○	○
Trimmomatic	×	○	○	○	×	○	×	○	×	×	○	○
AdapterRemoval	○	○	○	○	×	×	○	○	×	○	×	×
AlienTrimmer	○	○	○	○	×	○	×	○	×	×	×	×
NextClip	×	×	×	×	○	×	×	×	×	×	×	×
Skewer	○	○	○	○	○	○	○	○	○	×	○	○

For each method, the table shows if it is able to: i) identify adapters in the 5′ end of reads, ii) identify adapters in the 3′ end of reads, iii) process single-end (SE) reads, iv) process paired-end (PE) reads, v) process Nextera long mate-pair (LMP) reads, vi) search for multiple different adapters (Multi), vii) trim subsequences of multiple degenerative characters (Ns), viii) trim low-quality nucleotides (Q), ix) separate multiplexed reads based on barcodes, x) merge overlapped pairs into longer single-end reads, xi) process gzip files directly, and xii) run with multiple threads simultaneously (MT). (○: Yes; ×: No).