Quality check and trimming of sequencing reads

Bioinformatics Applications (PLPTH813)

Sanzhen Liu

2/23/2021

Outline

- Sequence data format
- Sequence quality
- Quality checking (fastQC)
- Sequence data processing (seqtk)
- Sequence trimming (trimmomatic)

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
```

Quality coding

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 = $-10 \times \log_{10}(p)$

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

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

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

Question

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

Can you think of a way to code quality data in different way to match sequencing data?

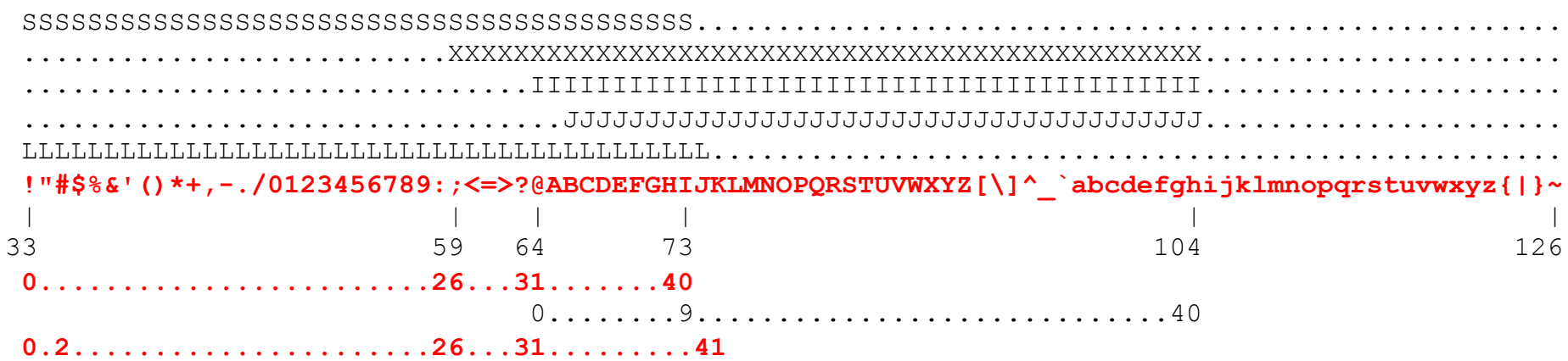
```
>SEQ_ID
ATCAACTGATGCATC
```

ASCII Table (partial)

ASCII stands for American Standard Code for Information Interchange.

32	space	52	4	72	H	92	\	112	p
33	!	53	5	73	I	93	]	113	q
34	“	54	6	74	J	94	^	114	r
35	#	55	7	75	K	95	_	115	s
36	$	56	8	76	L	96	`	116	t
37	%	57	9	77	M	97	a	117	u
38	&	58	:	78	N	98	b	118	v
39	‘	59	;	79	O	99	c	119	w
40	(	60	<	80	P	100	d	120	x
41	)	61	=	81	Q	101	e	121	y
42	*	62	>	82	R	102	f	122	z
43	+	63	?	83	S	103	g	123	{
44	,	64	@	84	T	104	h	124	|
45	–	65	A	85	U	105	i	125	}
46	.	66	B	86	V	106	j	126	~
47	/	67	C	87	W	107	k		
48	0	68	D	88	X	108	l		
49	1	69	E	89	Y	109	m		
50	2	70	F	90	Z	110	n		
51	3	71	G	91	[	111	o		

Quality codes in FASTQ



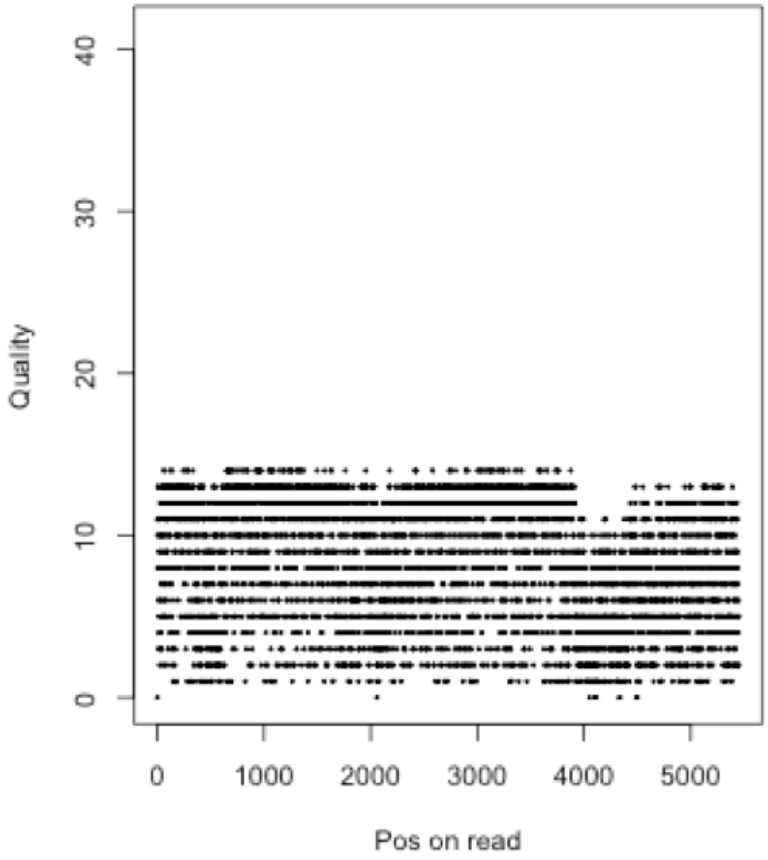
Sanger **Phred+33, raw reads typically (0, 40)**

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

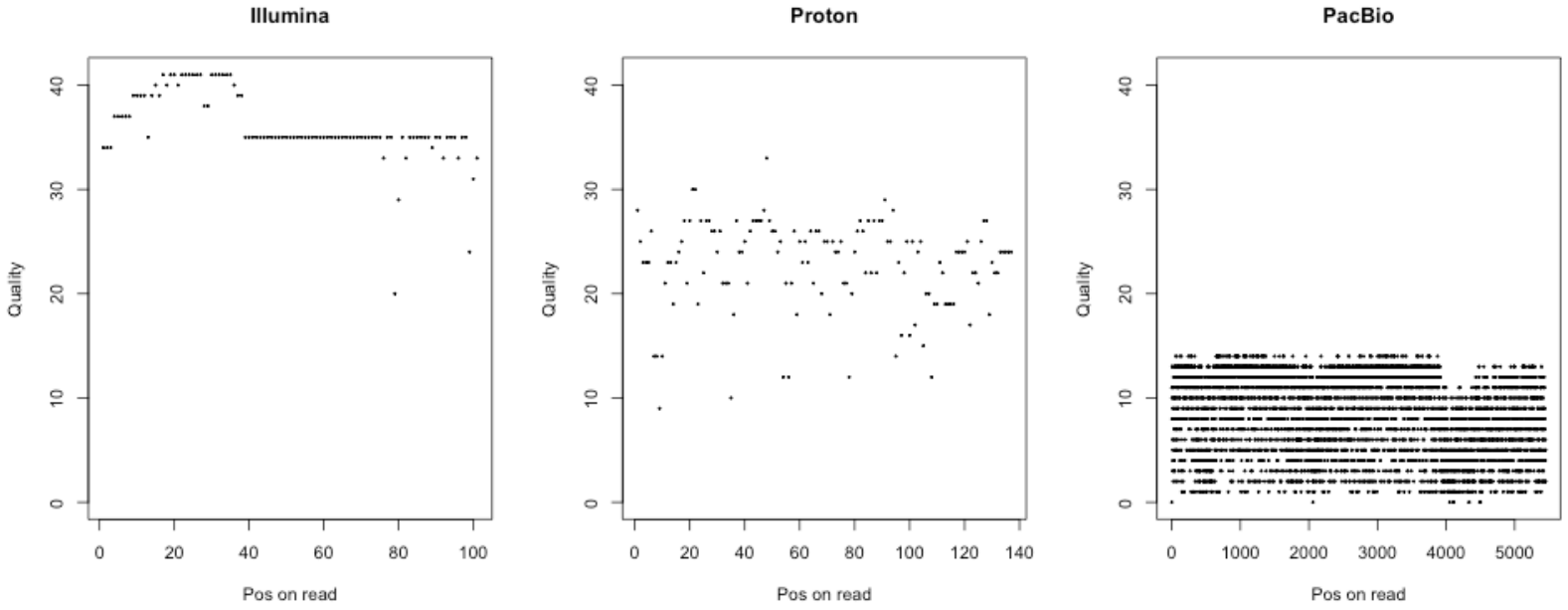
What platform was this read generated from?



Estimate average Phred quality?

Average probability of the error rate?

Typical reads in different platforms



Read length

Read quality

Data - FASTQ

Standard data format - FASTQ

@HWI-EAS225:3:1:2:854#0/1
GGGGGGAAGTCGGCAAAATAGATCCGTAACTTCGGG
+HWI-EAS225:3:1:2:854#0/1
a`abbbbabaabbbababb^`[aaa`_N]b^ab^``a

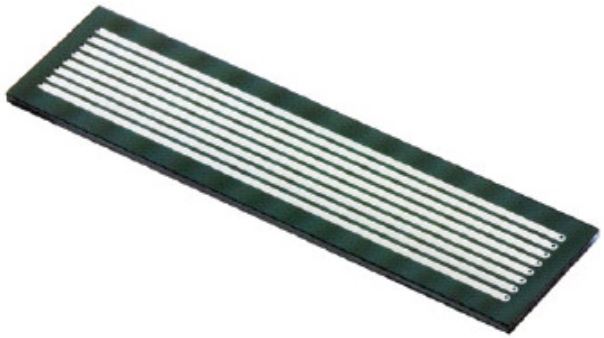
http://en.wikipedia.org/wiki/FASTQ_format

```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

```
@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 fails filter (read is bad), N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence



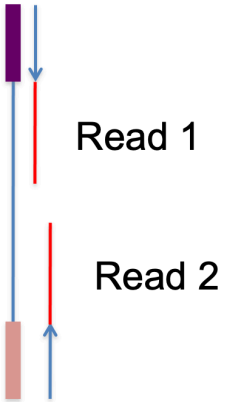
FASTQ

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

Single-end FASTQ file

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

Paired-end FASTQ files



File1: forward seq

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

File2: reverse seq

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

QC sequencing data

Data QC – FASTQC (I)

FASTQC is a tool to examine the quality of sequencing data

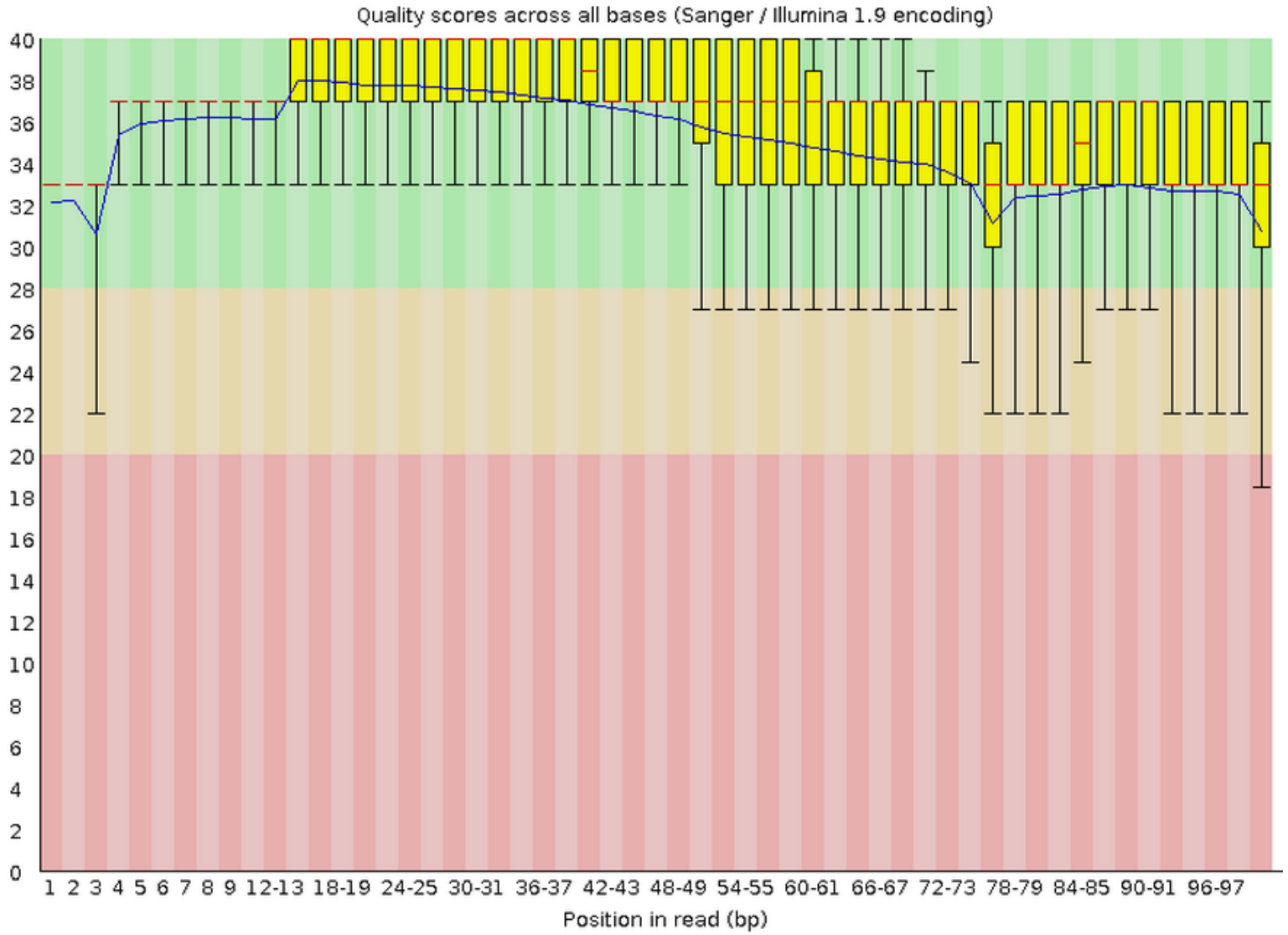
- Easy to run: `fastqc example.fastq`
- Rich output information
- Output presented in the html format

Basic Statistics

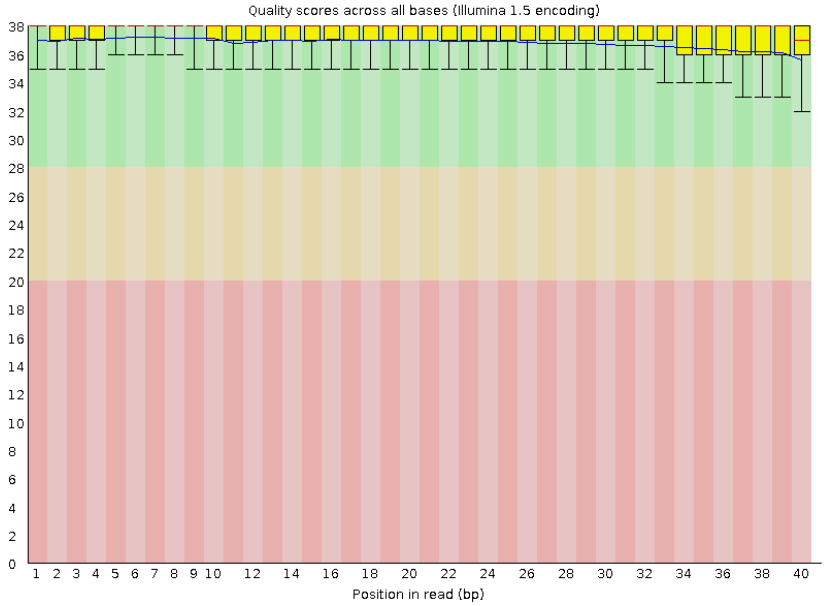
Measure	Value
Filename	example.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	10856448
Sequences flagged as poor quality	0
Sequence length	101
%GC	53

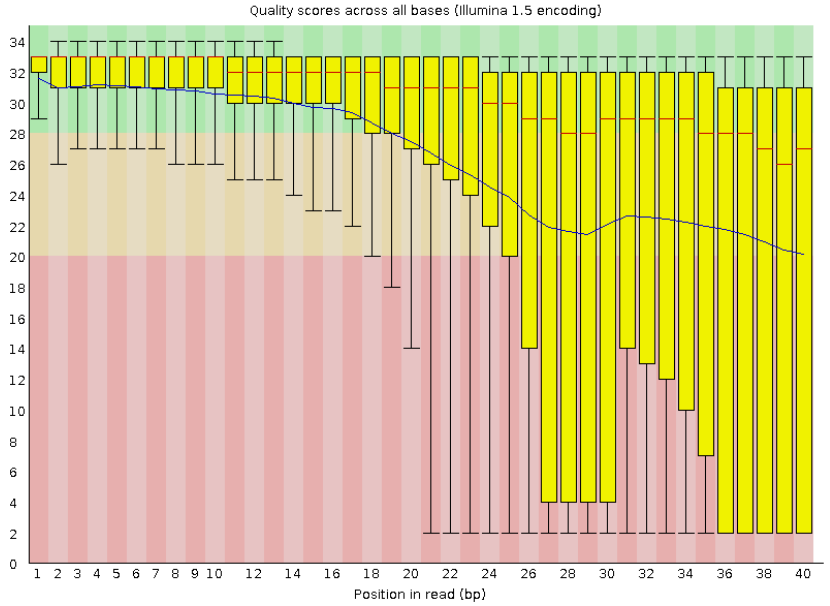
FASTQC (II)

Per base sequence quality



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 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
hety	regional heterozygosity
gc	identify high- or low-GC regions
mutfa	point mutate FASTA at specified positions
mergefa	merge two FASTA/Q files
famask	apply a X-coded FASTA to a source FASTA
dropse	drop unpaired from interleaved PE FASTA/Q
rename	rename sequence names
randbase	choose a random base from hets
cutN	cut sequence at long N
listhet	extract the position of each het

seqtk examples (I)

- 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 (II)

- 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

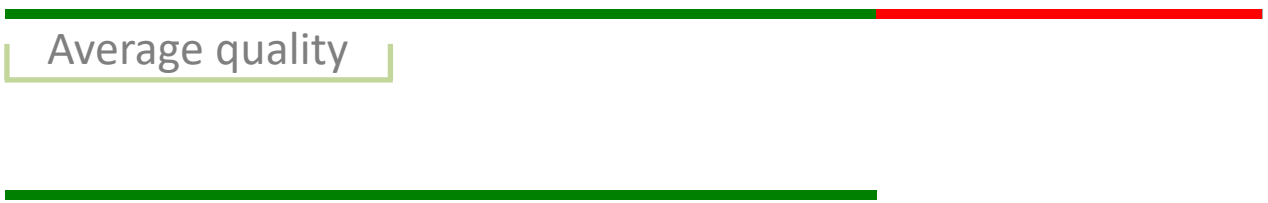
- **Quality trimming**: to remove low quality sequences
- **Adaptor trimming**: to remove adaptor contamination

Quality trimming

- **Window scan method**

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

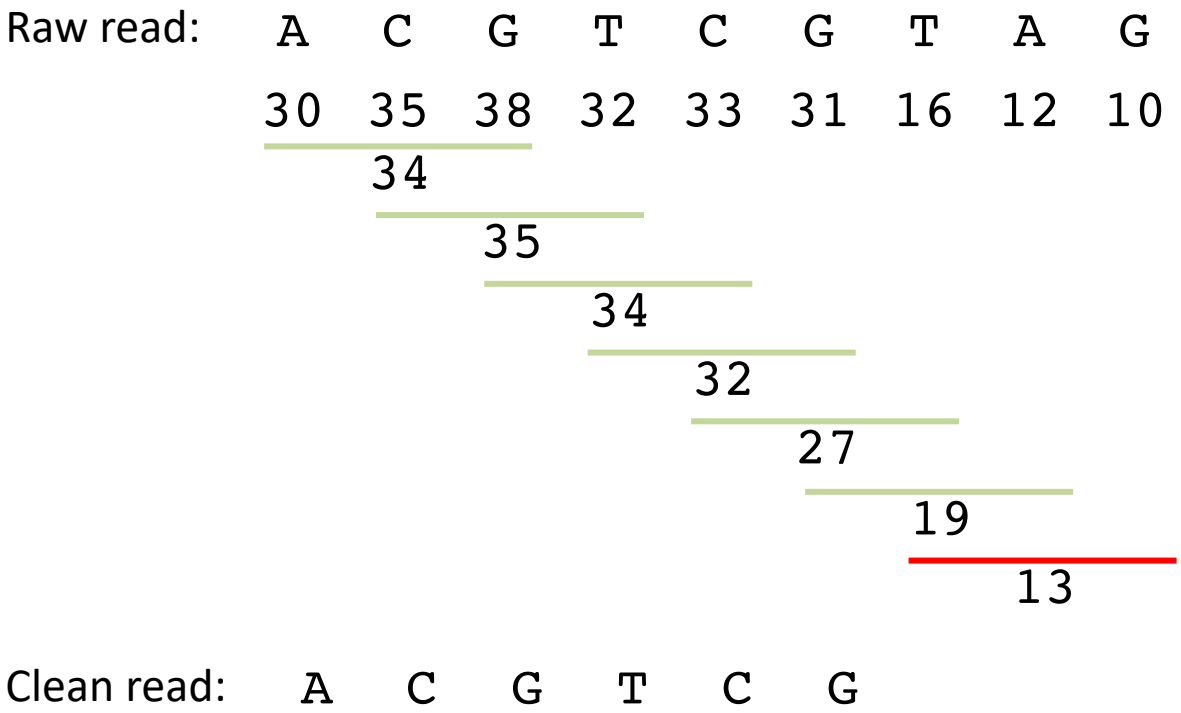
1. Sliding windows (window size and step size)
2. Maximum average errors (minimum average quality)



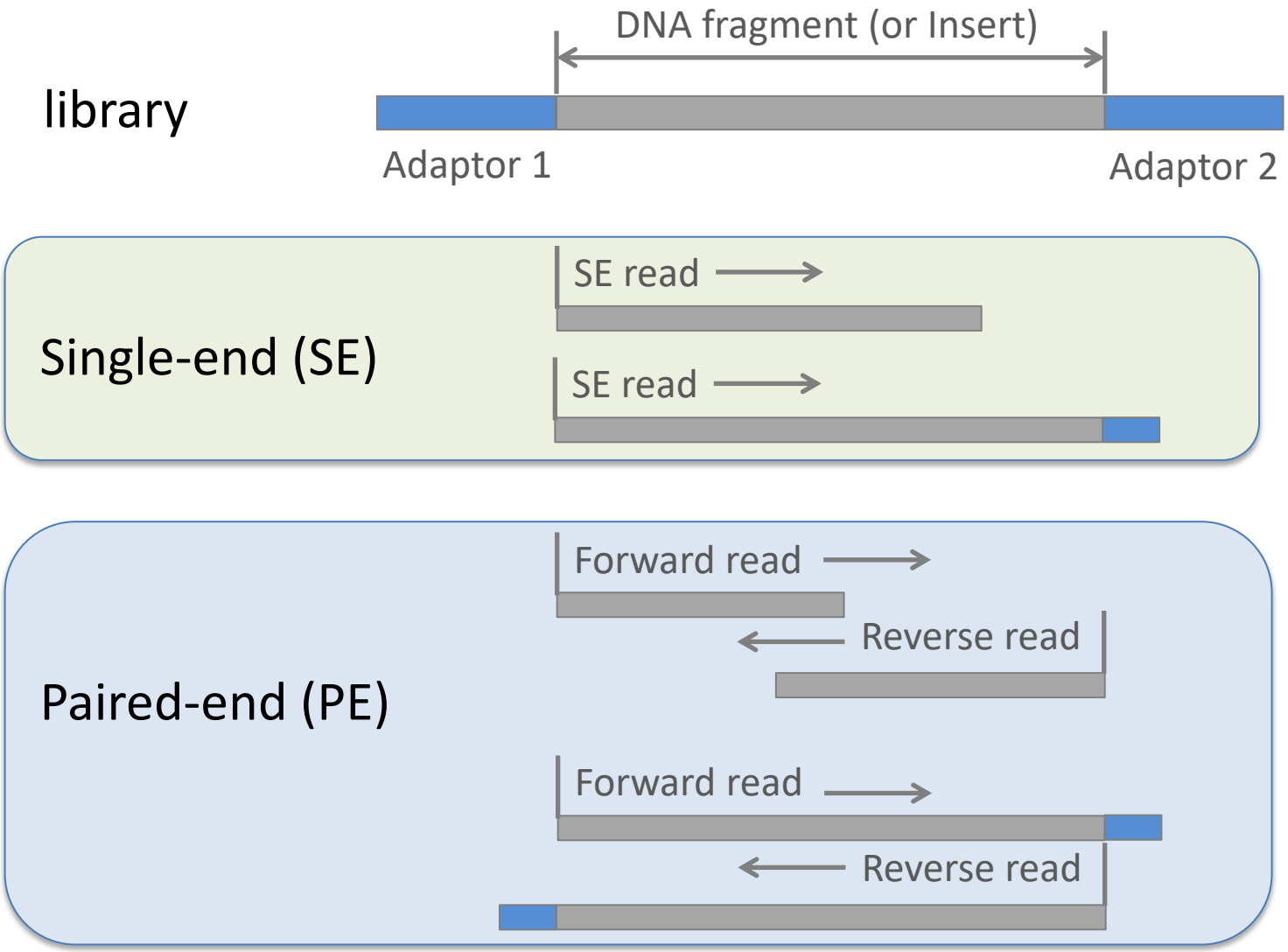
Bioinformatics, 2001, 17:1093-1104

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

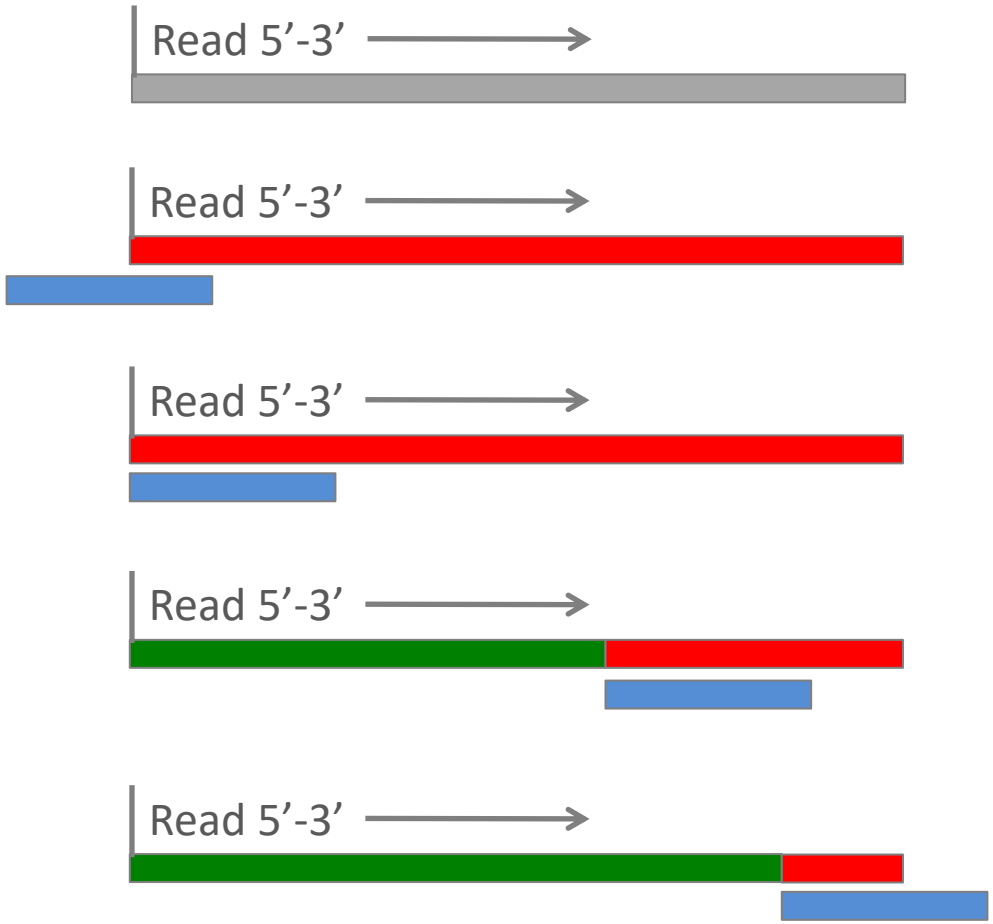
Trimmomatic: a flexible trimmer for Illumina sequence data

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

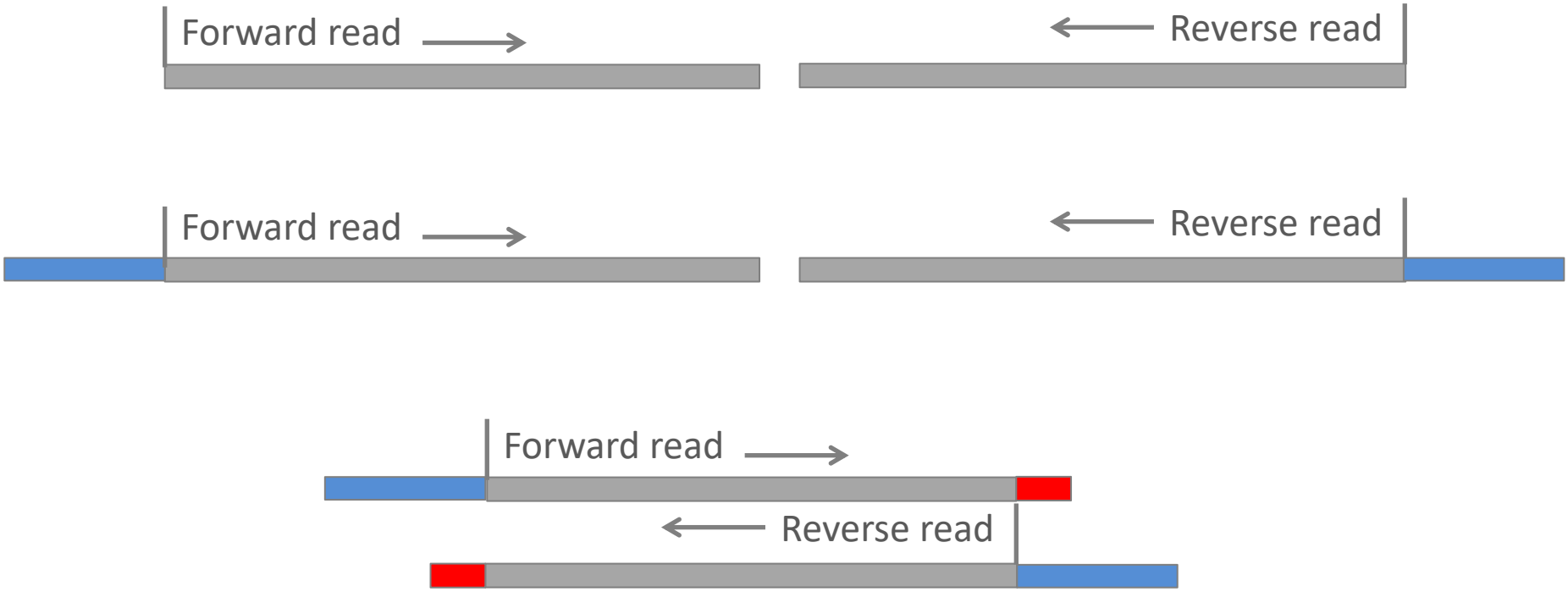
Simple mode: pro and cons

- 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***.

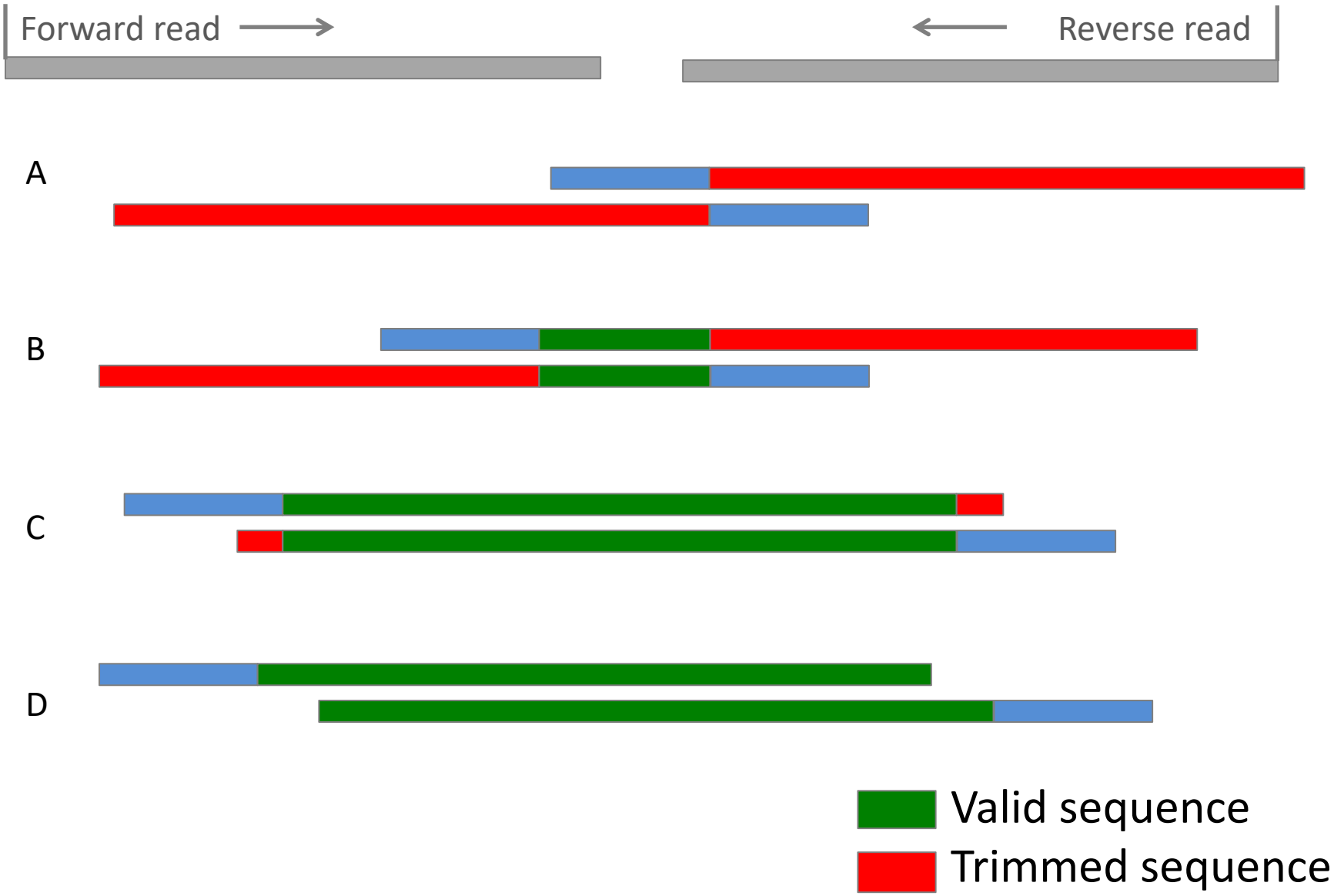
- Issue: if the adaptor sequence on the read is too short to make the alignment, the adaptor sequence can not be trimmed.

Trimmomatic – panlindrome mode

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



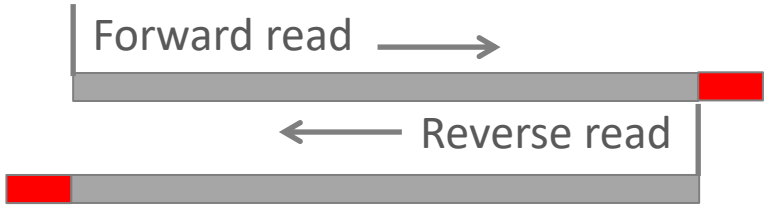
Trimmomatic – panlindrome mode



Trimmomatic – panlindrome mode

Pre-required features:

1. the valid sequence of the two reads are reverse complements
2. the valid sequence of two read are followed by contaminating sequence from the “opposite” adapters



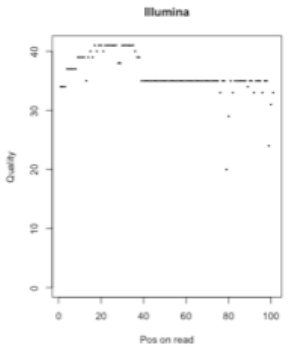
Quality trimming in Trimmomatic

- 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

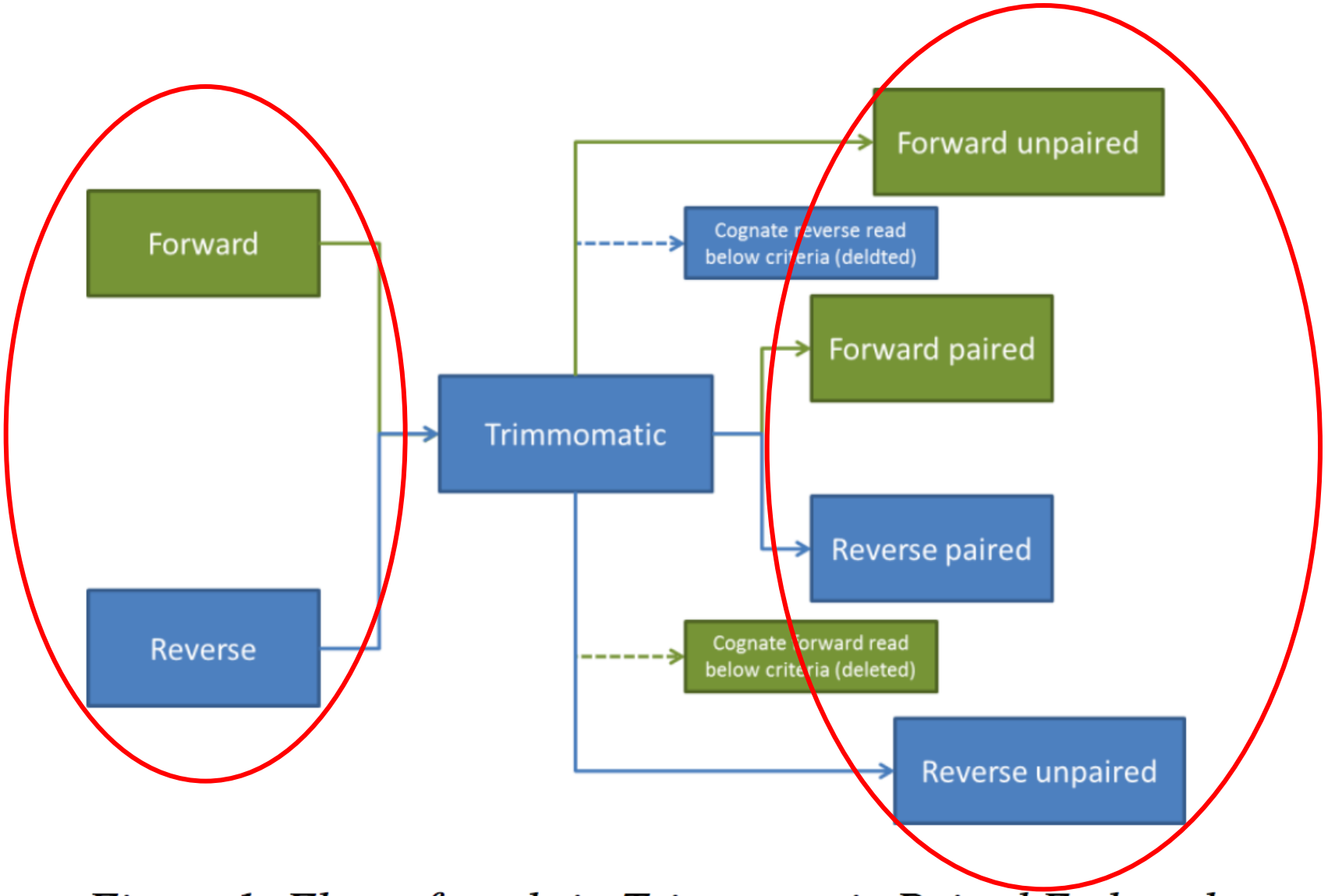


Figure 1: Flow of reads in Trimmomatic Paired End mode

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).

Summary

- FASTA and FASTQ
- Sequence quality (Phred)
- fastQC for quality checking
- Trimmomatic for quality and adaptor trimming