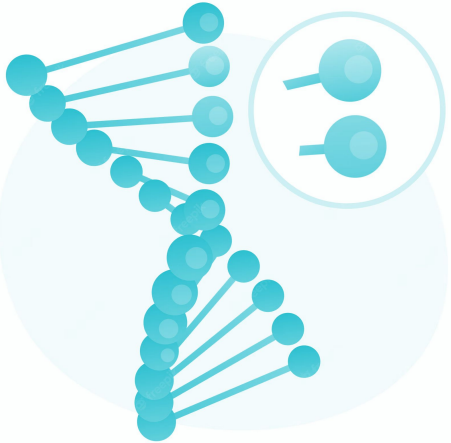
DNA-SEQ: UPSTREAM ANALYSIS

Presenter: Duy Dao

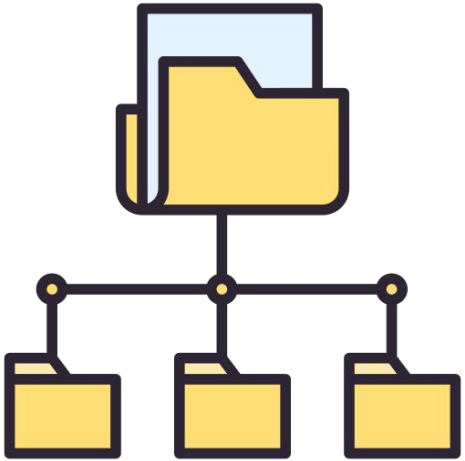
-05/07/2023-

TABLE OF CONTENTS

DNA SEQ: UPSTREAM ANALYSIS



PREPARATION



Setup working directory

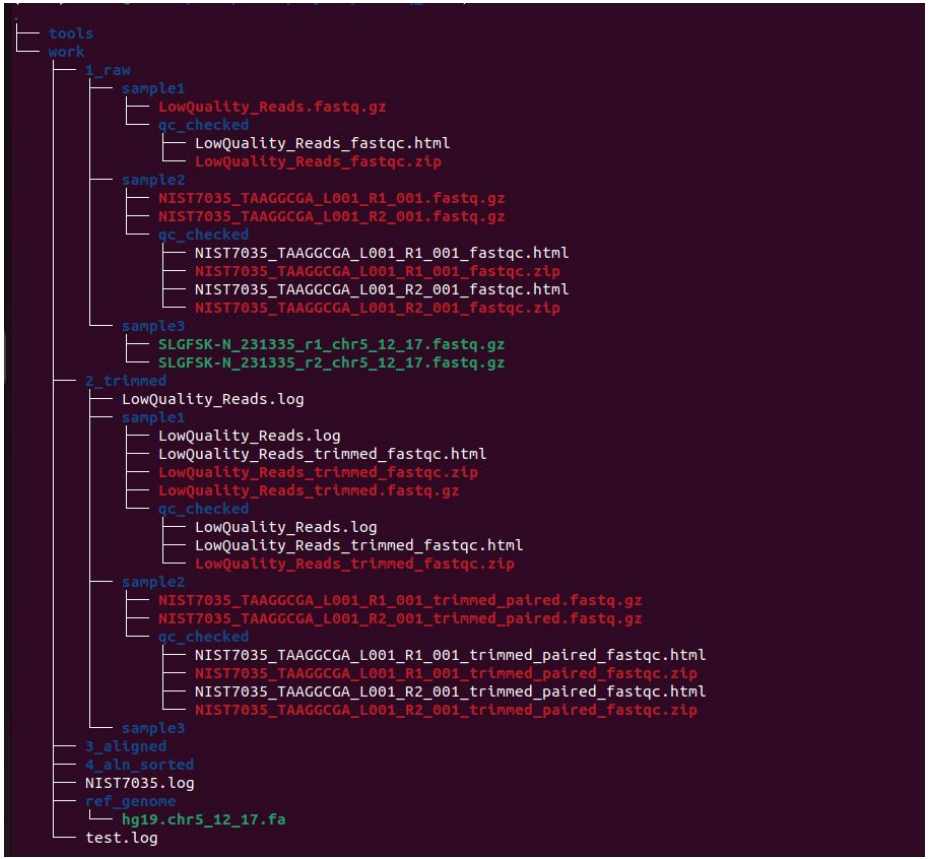
Download Data: Raw fastq &
Reference genome data

Install tools

PREPARATION

Setup working directory

```
dnaseq_work/
├── tools
│   ├── bwa
│   ├── FastQC
│   ├── gatk
│   ├── git-lfs-3.3.0
│   ├── htslib-1.17
│   ├── samtools-1.17
│   └── trimmomatic
└── work
    ├── 1_raw
    ├── 2_trim
    ├── 3_align
    └── ref_genome
```



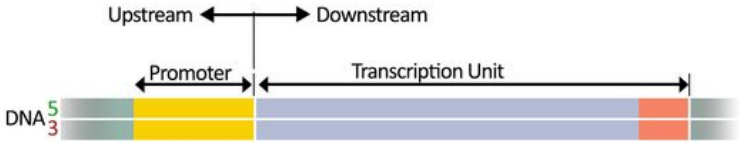
INTRODUCTION

INTRODUCTION

Concept

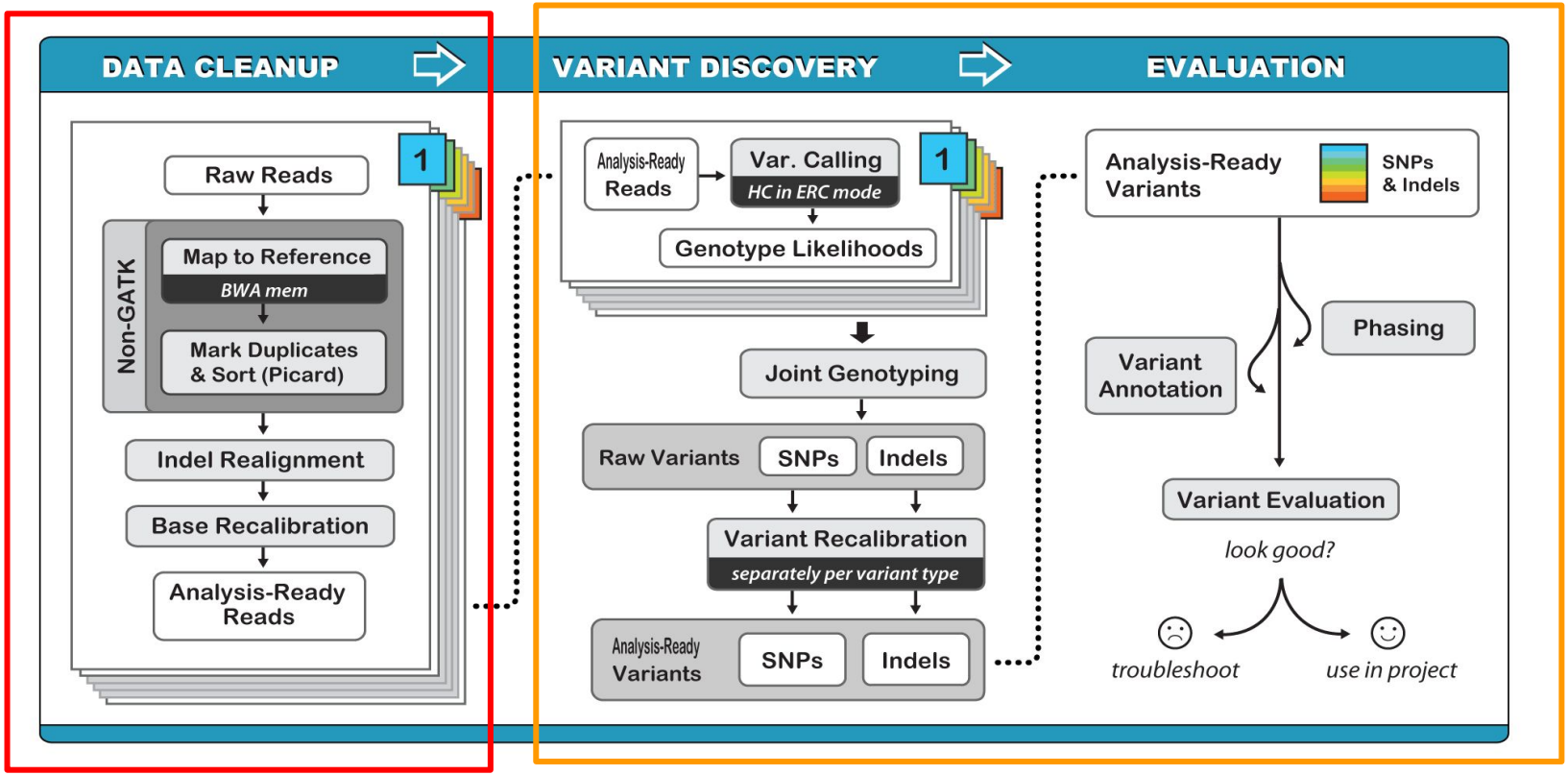
What is Upstream Analysis and why we have to do it?

What is Upstream Analysis workflow?



→ *Misconception*

DNA sequencing Workflow



Upstream Analysis

Downstream Analysis

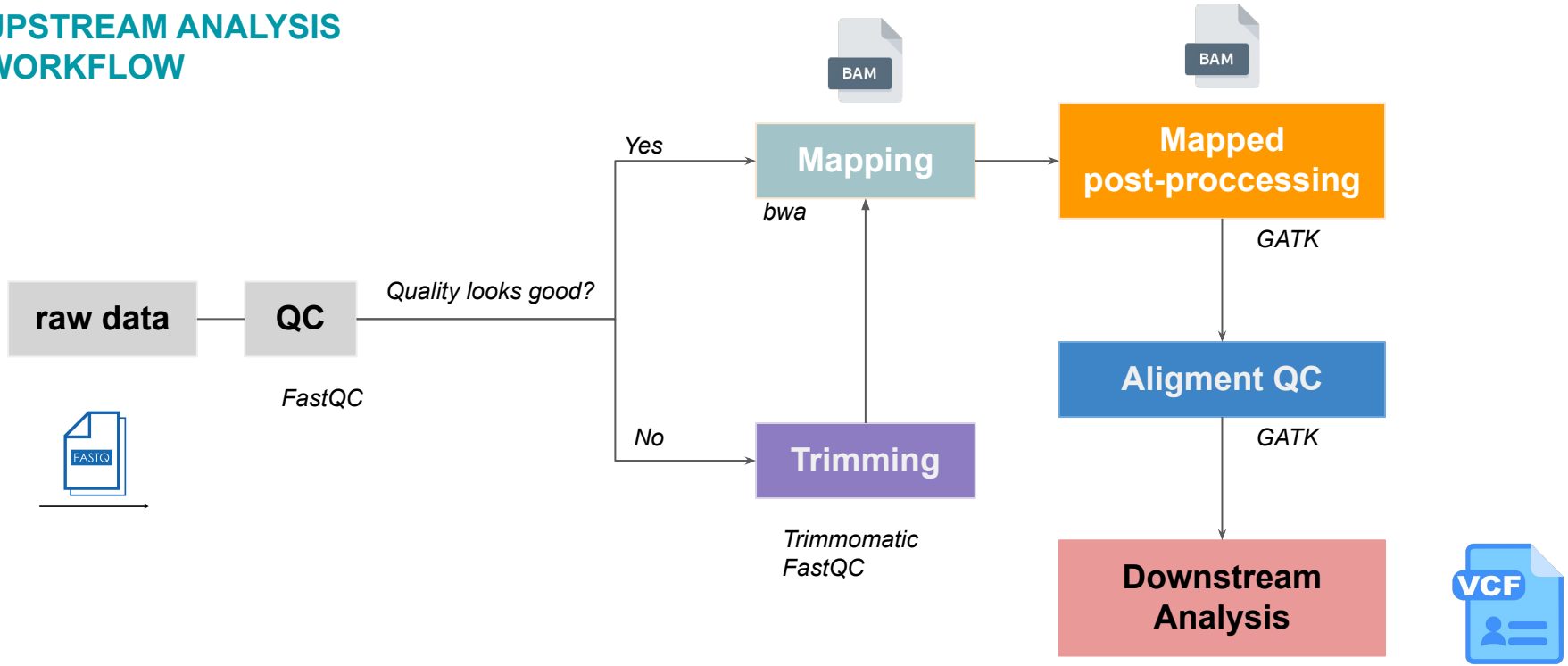
INTRODUCTION

CONCEPT

Why we need to pre-process the NGS data before further analysis?

INTRODUCTION

UPSTREAM ANALYSIS WORKFLOW



RAW DATA PRE-PROCESSING

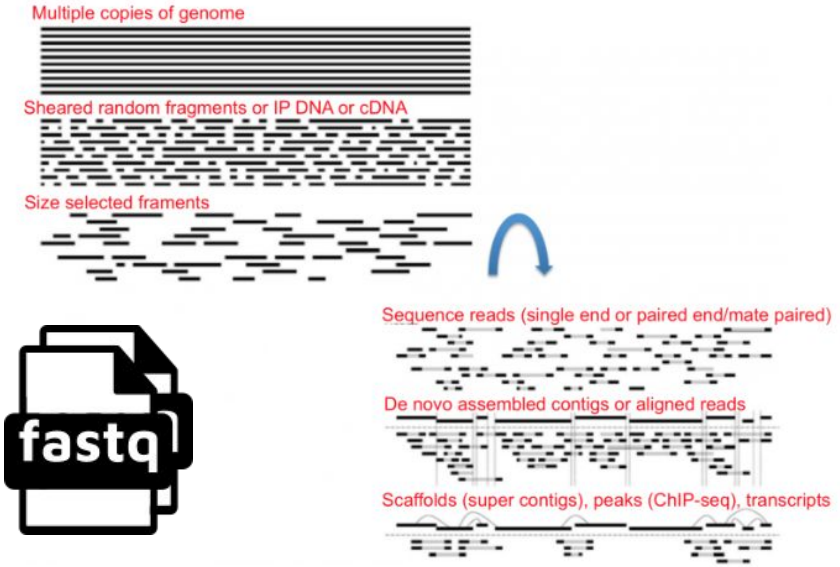
RAW DATA PROCESSING

- ★ *What is raw data?*
- ★ *What is a fastq file (Illumina)?*
- ★ *Does your data look good? How to check it?*
- ★ *How to keep the good and eliminate the bad quality?*

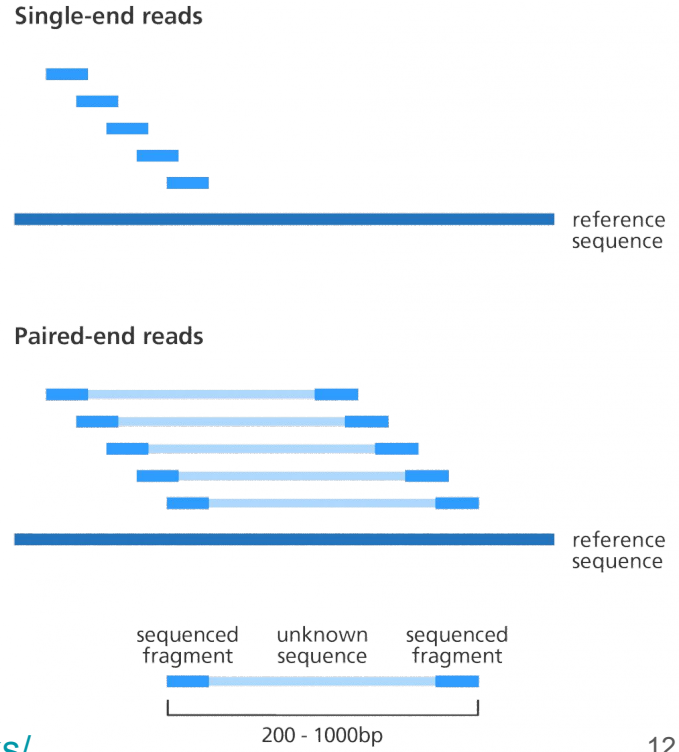
RAW DATA PROCESSING

★ *What is raw data?*

The reads (sequences)



Short reads



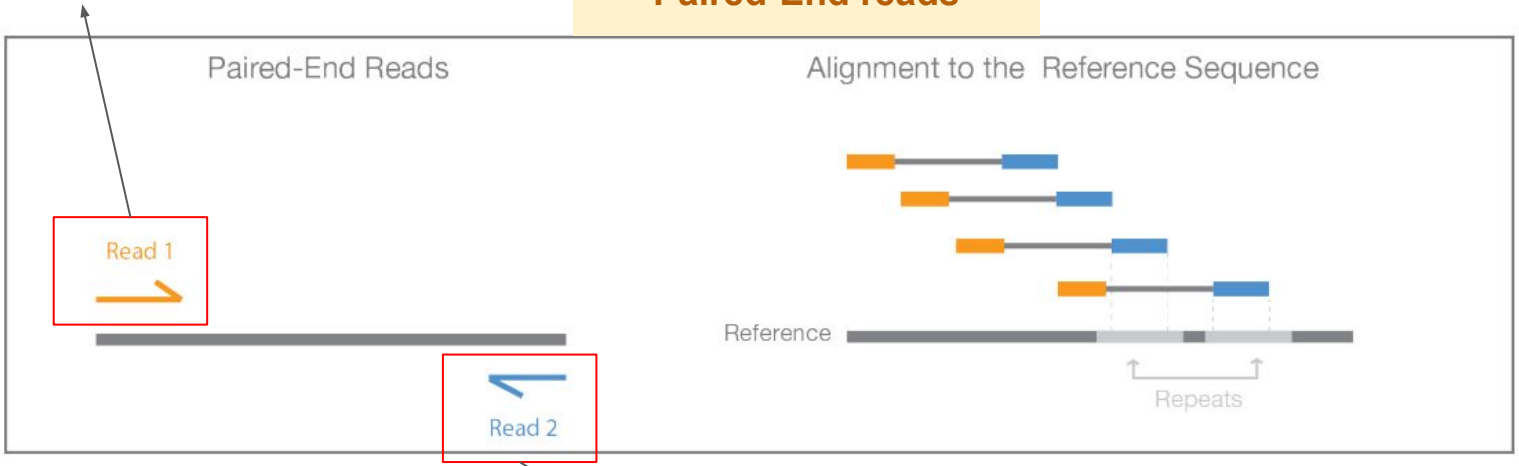
https://learn.gencore.bio.nyu.edu/ngs-file-formats/how-sequencing-works/

RAW DATA PROCESSING

★ *What is raw data?*

R1.fastq (forward read)

Paired-End reads



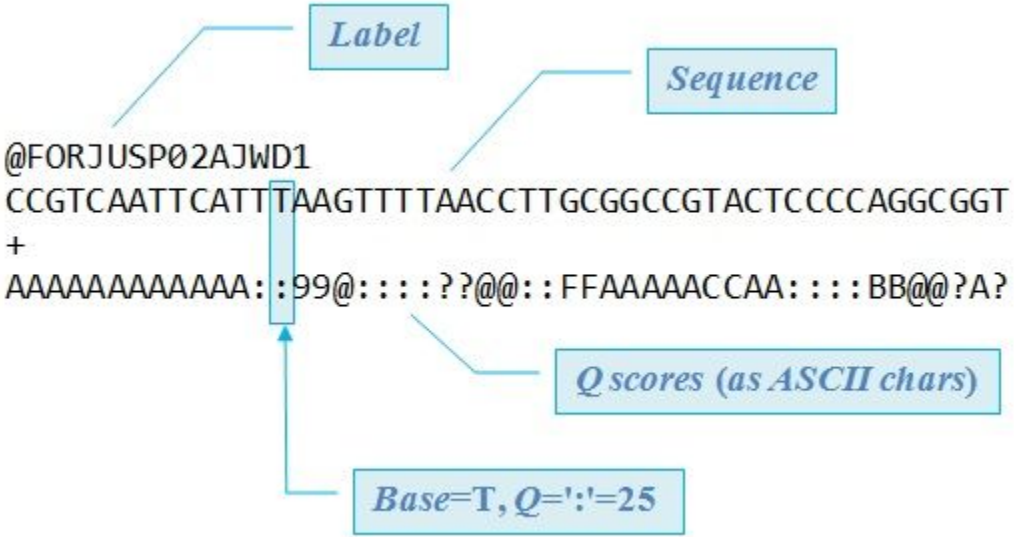
R2.fastq (reverse read)

```
sample2
    NIST7035_TAAGGCGA_L001_R1_001.fastq.gz
    NIST7035_TAAGGCGA_L001_R2_001.fastq.gz
```

https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

RAW DATA PROCESSING

INTRODUCE TO FASTQ FILE



File Format

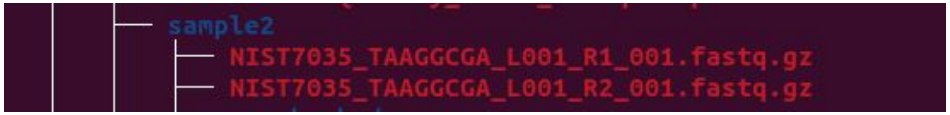
★ *What is a fastq file (Illumina)?*

```
@<title and optional description>
<sequence line>
+<optional repeat of title line>
<quality line>
```

RAW DATA PROCESSING

INTRODUCE TO FASTQ FILE

Illumina FASTQ file naming scheme



NIST7035_TAAGGCGA_L001_R1_001.fastq.gz

- sample_name: NIST7035
- barcode_sequence: TAAGGCGA
- lane: L001
- read_number: R1
- set_number: 001

> What is the meaning of fastq file's name?

RAW DATA PROCESSING

INTRODUCE TO FASTQ FILE

Illumina FASTQ read naming scheme (header of each read)

```
@@@D?BD?A>CBDCED;EFGF;@B3???8))0)8?B>B@FGCFEEBC################
@HWI-D00119:50:H7AP8ADXX:1:1101:1242:2178 1:N:0:TAAGGCGA
ACATAGTGGTCTGTCTTCTGTTTATTACAGTACCTGTAATAATTCTTGATGCTGTCCAGACTT
+
@@@DDBAAB=AACEEEFEEAHFEEECHAH>CH4A<+:CF<FFE<<DF@BFB9?4?<*?9D;D
@HWI-D00119:50:H7AP8ADXX:1:1101:1326:2059 1:N:0:TAAGGCGA
ATCTCTAGATCTATCTATCGTCTACCATTATCCATCGACCTATCTGTCTCTATTATATACACC
+
?@?DDDDDDHHHDIIJIJJJJJJJJJJIIIEIIJJJIHHIGHGHJIIJJIIJJIIJJJJIIII
@HWI-D00119:50:H7AP8ADXX:1:1101:1267:2070 1:N:0:TAAGGCGA
GTATTAGTTTGGAGCAACGGGAGGGAGGGTCTGGAGGAGACTCCCTCGGGCGCGCCGCGGGTA
+
@?@DFDFFHDFDFHGEHIIHGHIGIGEGHIHGIEGEGCGHIIIGIIIFEHDDDBDBDDDDB5?
@HWI-D00119:50:H7AP8ADXX:1:1101:1322:2086 1:N:0:TAAGGCGA
CTCTCTGATGATGGCATCCTGCCCAGCCCTTTGTCCCAGGTCCACGGGTAGGGATACTCACGA
+
CCCFFFFFHHHHHIJJJJJJIJJJJJJJJJJJHIJJJIIJJJJJJJJJJHIJJJIJJJJJJJJH
@HWI-D00119:50:H7AP8ADXX:1:1101:1294:2087 1:N:0:TAAGGCGA
TTGTTATTTCATTTTCCAAACTCTGGAAGTATTTTATTTTATTTAAAATATTTTCAGCATCTT
+
@@?BDDAADDHHFJJ>BGACHIJJEEFHGA<CFCIIIHGAHHHIIIFGHHIJIAHGG:BBGHE
@HWI-D00119:50:H7AP8ADXX:1:1101:1457:2135 1:N:0:TAAGGCGA
```

The header Begin witn an '@'

fastq header format (version > 1.8)

Sequence Header +Sequence ID

a b c d e f g h i j k

@HWI-ST486:166:C06K9ACXX:7:1101:1443:1995 1:N:0:ACAGTG

a. unique instrument name
b. run id
c. flowcell id
d. flowcell lane
e. tile number within the flowcell lane
f. x-coordinate of the cluster within the tile
g. y-coordinate of the cluster within the tile

h. the member of a pair, 1 or 2 (paired-end or mate-pair reads only)
i. Y if the read fails filter (read is bad), N otherwise
j. 0 when no control bits are on
k. index sequence

> What is the meaning of the reads' name?

RAW DATA PROCESSING

INTRODUCE TO FASTQ FILE

Base Quality Score (Q-score)

Table 5.2. Base Quality and ASCII Encoding.

ASCII character	Decimal value	Phred score
!	33	0
"	34	1
#	35	2
$	36	3
⋮	⋮	⋮
A	65	22
B	66	23
⋮	⋮	⋮
x	120	87
y	121	88
z	122	89
{	123	90
|	124	91
}	125	92
~	126	93

Phred 33

;B<97><A89;;<9?>?>9?<9=66;<<6@A@B?7<@<99@7<8:6?=66;@:6<;666778

"Measure of the confidence level in the accuracy of each nucleotide base call."

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%

Q30 is considered a benchmark for quality in next-generation sequencing.

RAW DATA PROCESSING

Classwork 1

Interpret these headers

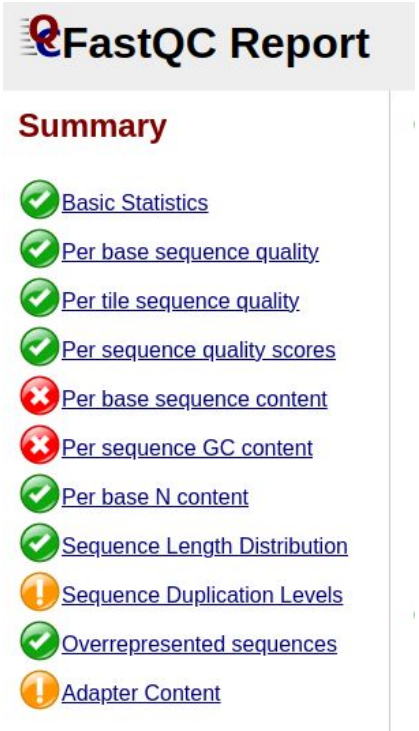
1
```
@HWI-D00107:50:H6BP8ACWV:5:2204:10131:51624 2:N:0:AGGCAGAA
```

2
```
@Machine42:1:FC7:7:19:4229:1044 1:N:0:TTAGGC
```

- ❑ What is the ID of the instrument?
- ❑ What is the ID of the flowcell?
- ❑ Which lane is this read in?
- ❑ What tile is this read in?
- ❑ Forward or Reverse read?

SEQUENCE QUALITY CONTROL (FASTQC)

> FASTQC Summary



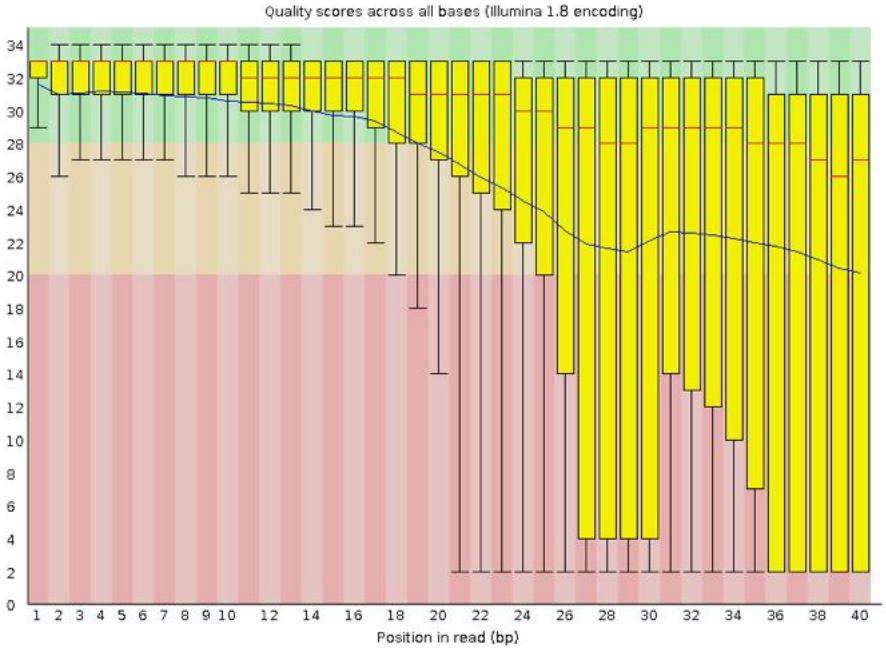
"FASTQC is a useful tool to check sequences quality."

Basic Statistics

Measure	Value
Filename	NIST7035_TAAGGCGA_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	20203002
Total Bases	2 Gbp
Sequences flagged as poor quality	0
Sequence length	101
%GC	49

SEQUENCE QUALITY CONTROL (FASTQC)

> Per base sequence quality



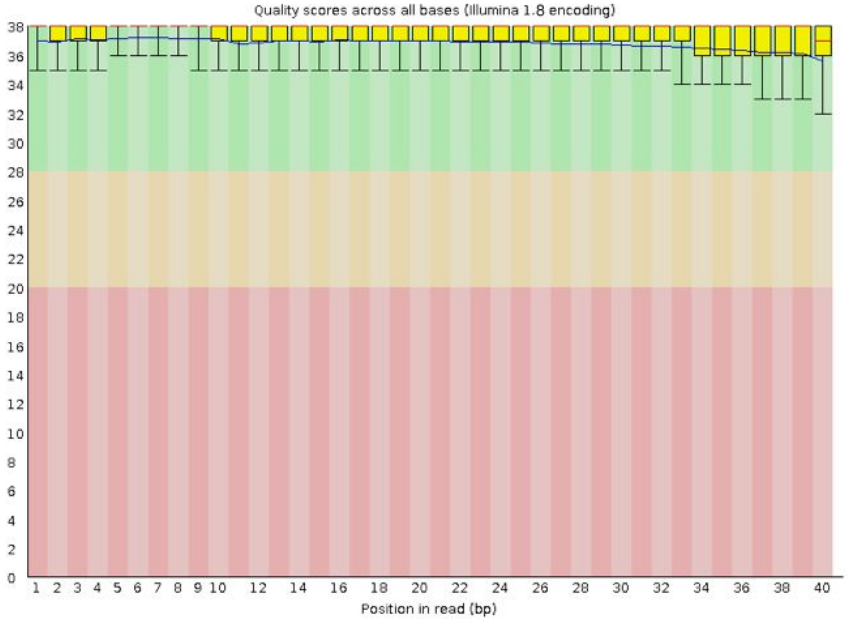
This module evaluates the quality at each base for all reads.

- Box-plot: Yellow
- Median: Red line
- Mean: Blue line

< Example of a bad quality score reads (40 bp)

SEQUENCE QUALITY CONTROL (FASTQC)

> Per base sequence quality



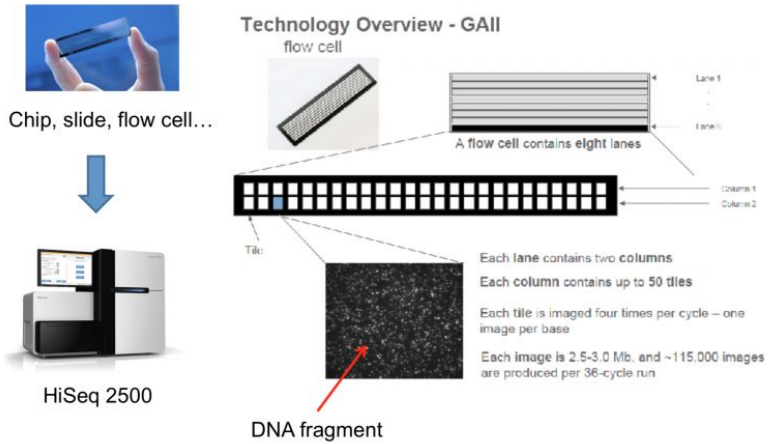
Good quality score (Q30)

Per base sequence quality

RAW DATA PROCESSING

SEQUENCE QUALITY CONTROL (FASTQC)

> Per tile sequence quality



What is a tile?

a b c d e f g h i j k

@HWI-ST486:166:C06K9ACXX:7:1101:1443:1995 1:N:0:ACAGTG

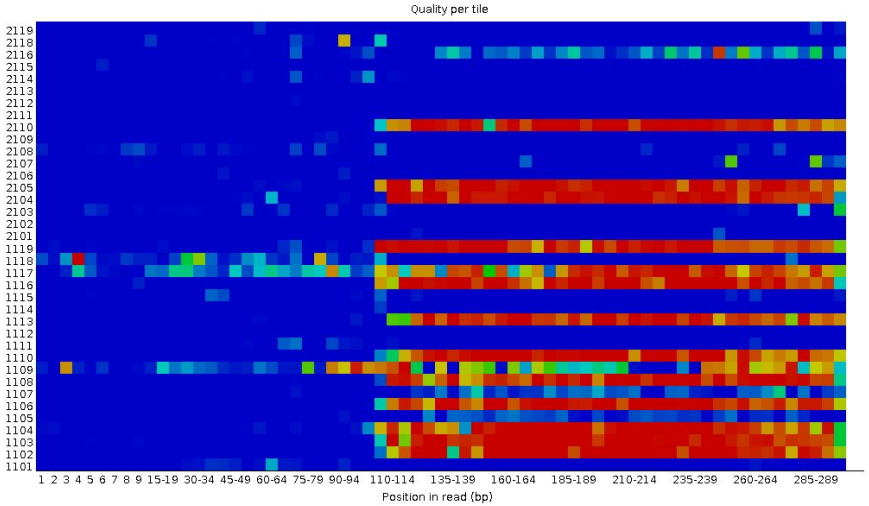
a. unique instrument name
b. run id
c. flowcell id
d. flowcell lane
e. tile number within the flowcell lane
f. x-coordinate of the cluster within the tile
g. y-coordinate of the cluster within the tile

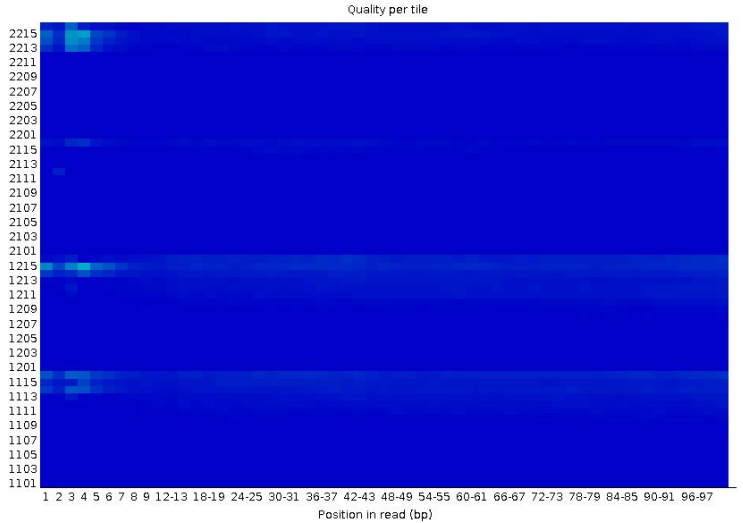
Information about tile stored in header of fastq

RAW DATA PROCESSING

SEQUENCE QUALITY CONTROL (FASTQC)

> Per tile sequence quality





Sequencing errors: bubbles, smudges, or dirt.

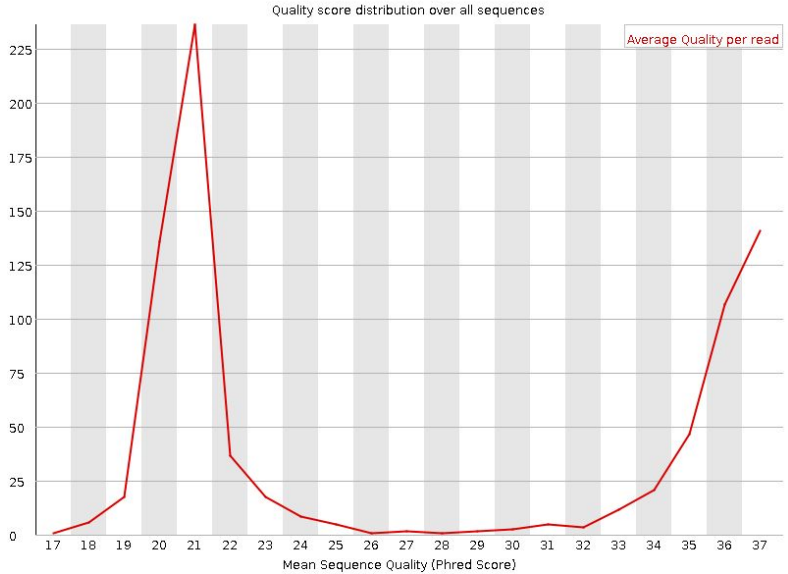
→ Cannot be fixed with bioinformatics

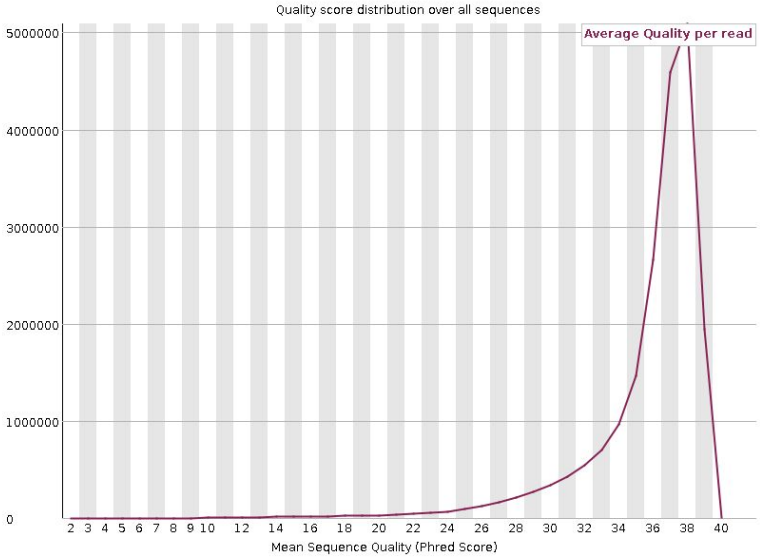
RAW DATA PROCESSING

SEQUENCE QUALITY CONTROL (FASTQC)

> Per sequence quality scores

"The average quality score over the full length of all reads."



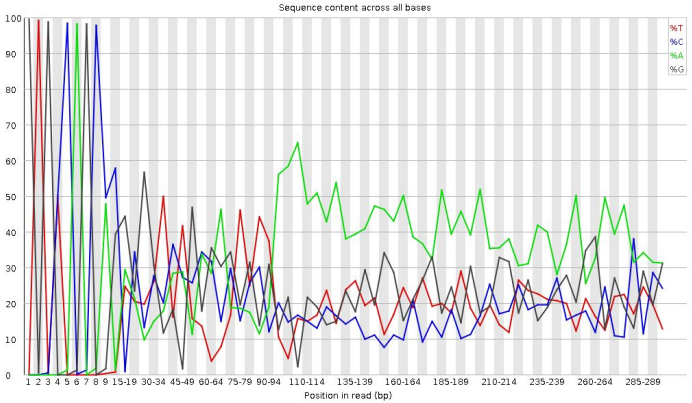


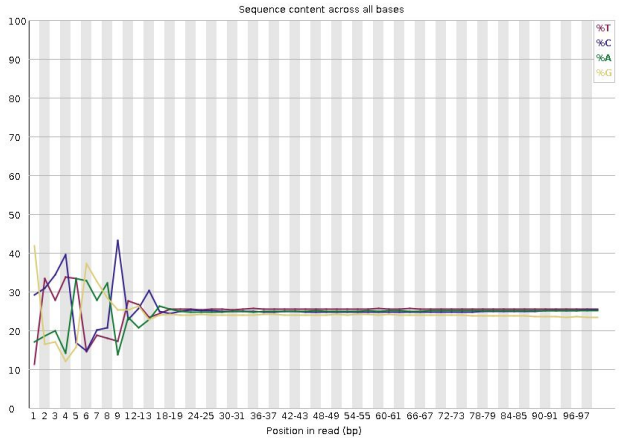
RAW DATA PROCESSING

SEQUENCE QUALITY CONTROL (FASTQC)

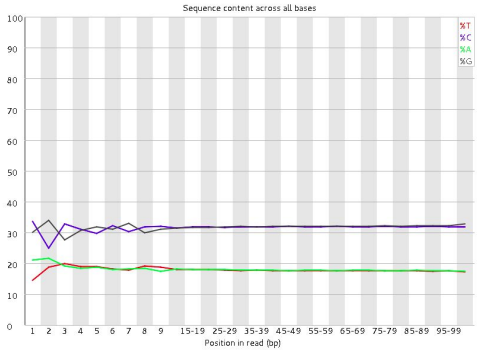
> Per Base Sequence Content

"Per Base Sequence Content" plots the percentage of each of the four nucleotides (T, C, A, G) at each position across all reads in the input sequence file.





Biased fragmentation (first 12 bp)



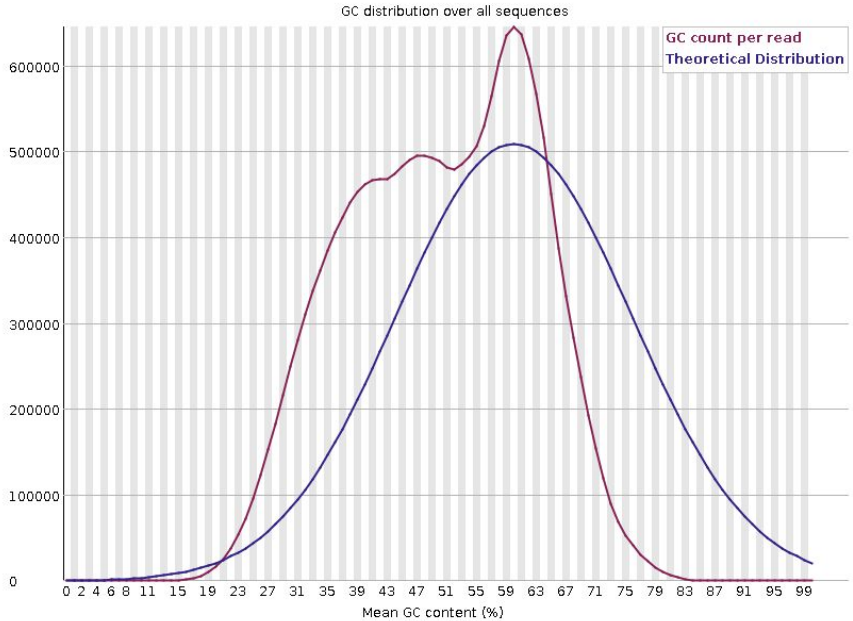
- Parallel
- %A=%T
- %G=%C

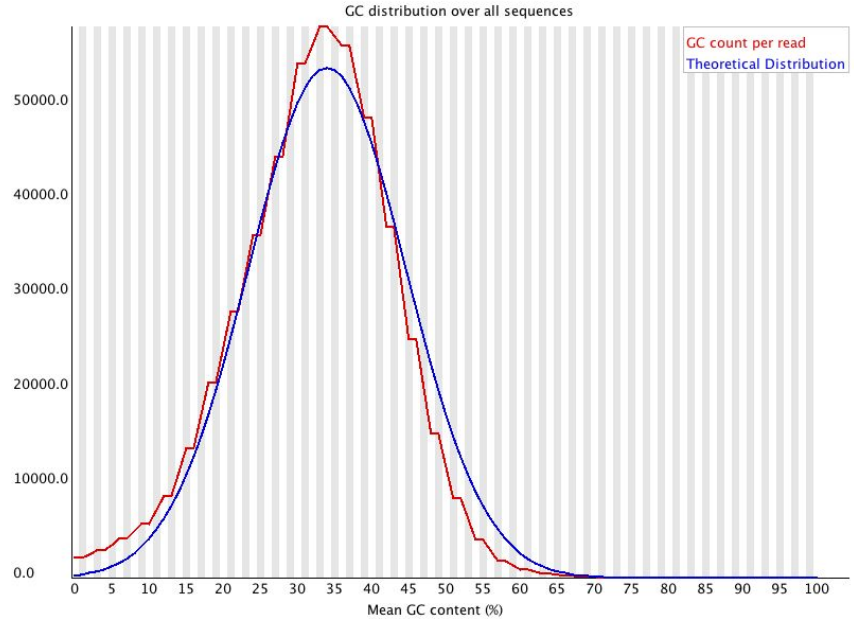
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

RAW DATA PROCESSING

SEQUENCE QUALITY CONTROL (FASTQC)

> Per sequence GC content



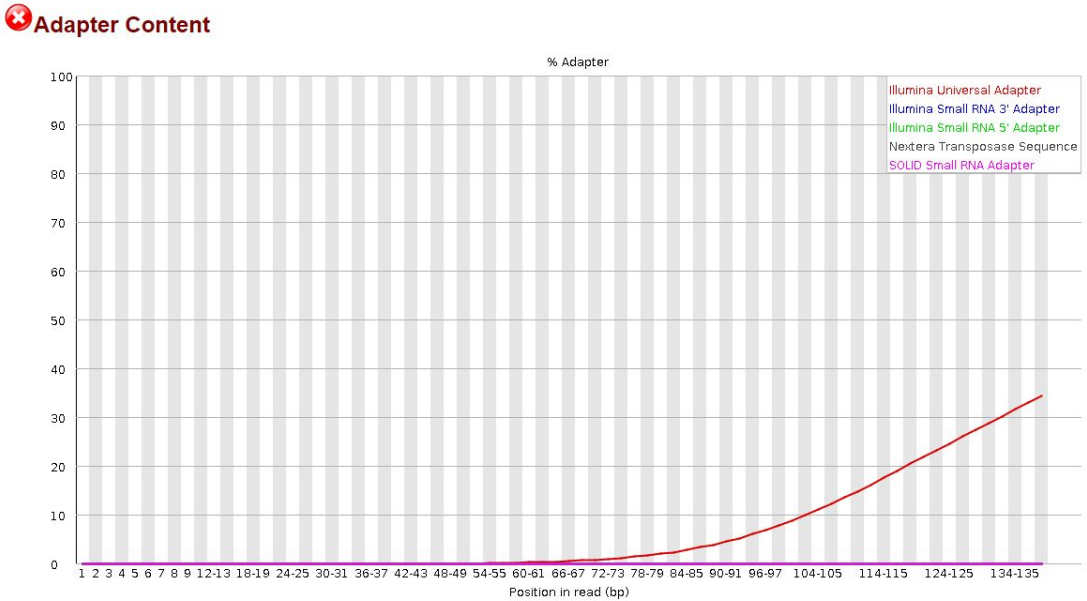


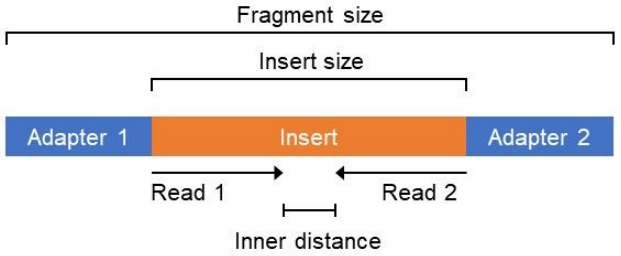
Deviations from this theoretical distribution often implies contamination of some kind (adapter/primer dimers, multiple species in the run)

RAW DATA PROCESSING

SEQUENCE QUALITY CONTROL (FASTQC)

> Adapter Content





The Adapters need to be removed

RAW DATA PROCESSING

SEQUENCE QUALITY CONTROL (FASTQC)

> Overrepresented sequences

Overrepresented sequences

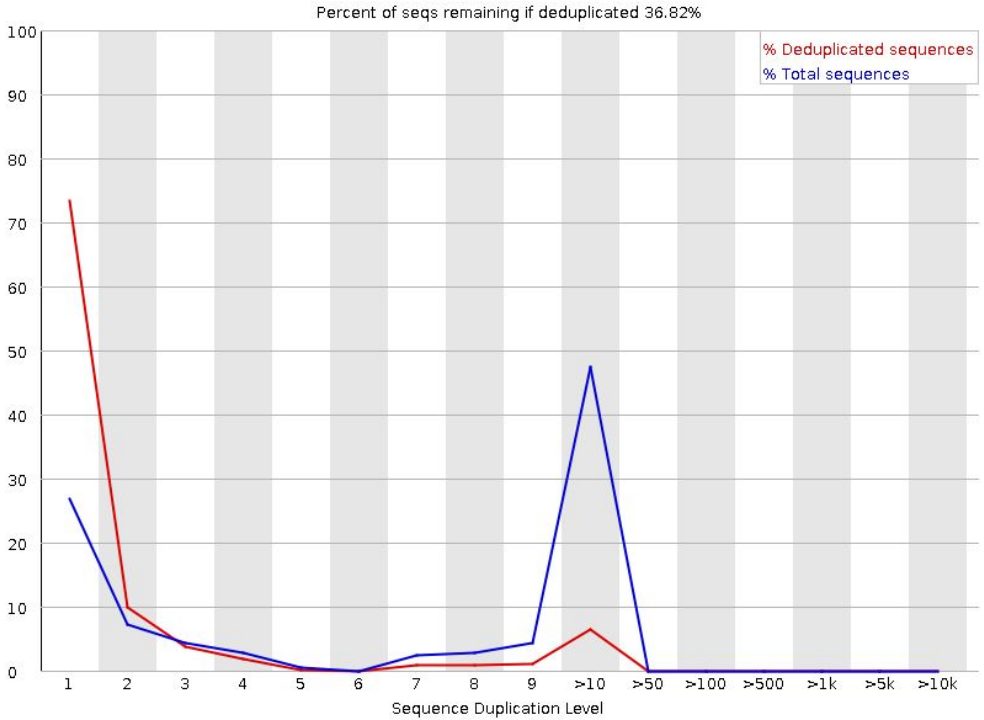
Sequence	Count	Percentage	Possible Source
GTCGGTAAAACTCGTGCCAGCCACCGCGGTCATACGATTAACCCAAGCTA	14869	13.721346572662508	No Hit
GTCGGTAAAACTCGTGCCAGCAAACTGGGATTAGATACCCCACTATGCTG	10094	9.314901627846886	No Hit
GTCGGTAAAACTCGTGCCAGCAGCCGCGGTAATACAGAGGTCTCAAGCGT	9386	8.66154811561035	No Hit
GTCGGTAAAACTCGTGCCAGCAGCCGCGGTAATACAGAGGTCCCAAGCGT	8815	8.134620353622974	No Hit
GTCGGTAAAACTCGTGCCAGCGGCGTTTCTTCGAATCCGGTGCCCAATAG	4442	4.0991473182975895	No Hit
GTCGGTAAAACTCGTGCCAGCCACCGCGGTCACACGATTAACCCAAGTCA	3578	3.3018345576021555	No Hit
GTCGGTAAAACTCGTGCCAGCCCACGAGACCAAACTGGGATTAGATACCC	1941	1.7911852644789783	No Hit
GTCGGTAAAACTCGTGCCAGCACGATTTCGGAGGGCGTTGCAATATTGGC	1040	0.9597283230593187	No Hit
GTCGGTAAAACTCGTGCCAGCAGCCGCGGTAAGACAGAGGTCCCGAGCGT	692	0.6385884611125466	No Hit
GTCGGTAAAACTCGTGCCAGCAGCCGCGGTAATACGGAGGATCCGAGCGT	669	0.6173637001218116	No Hit
GTCGGTAAAACTCGTGCCAGCCTAAATTATCATACGAACAAACTGGGATT	623	0.5749141781403417	No Hit
GTCGGTAACACTCGTGCCAGCCACCGCGGTCATACGATTAACCCAAGCTA	575	0.5306190247683733	No Hit
GTCGGTAAAACTCGTGCCAGCAGAGACAGCAAACTGGGATTAGATACCCC	570	0.5260049462921266	No Hit

FastQC lists all of the sequence which make up more than 0.1% of the total.

- Contaminated?
- Adapter?
- RNA transcripts?

SEQUENCE QUALITY CONTROL (FASTQC)

> Sequence duplication levels



2 types of Duplication Error:
- PCR duplication
- Optical duplication

→ Map and dectect which type.

RAW DATA PROCESSING

★ *How to deal with these problems?*

Things that we (bioinformaticians) can do:

- Filter and remove bad sequences
- Remove duplicated sequences
- Remove adapter
- Crop bad quality bases at the head/tail of sequences.

Things we can't:

- Remove contaminants
- Control GC contents.
  (Control previous steps: library prep, sequencing)

RAW DATA PROCESSING

Class work 2:

1. Do quality check for Sample1 (using FASTQC). This is a fastq file that stores singled-end reads. Interpret the results.

2. Do quality check for Sample2 (using fastqc), contains paired-end fastq files. Interpret the results.

RAW DATA PROCESSING

READ TRIMMING & FILTERING

usadellab/
Trimmomatic

2 Contributors | 25 Issues | 131 Stars | 56 Forks

This program does adaptive quality trimming, head and tail crop, and adaptor removal.

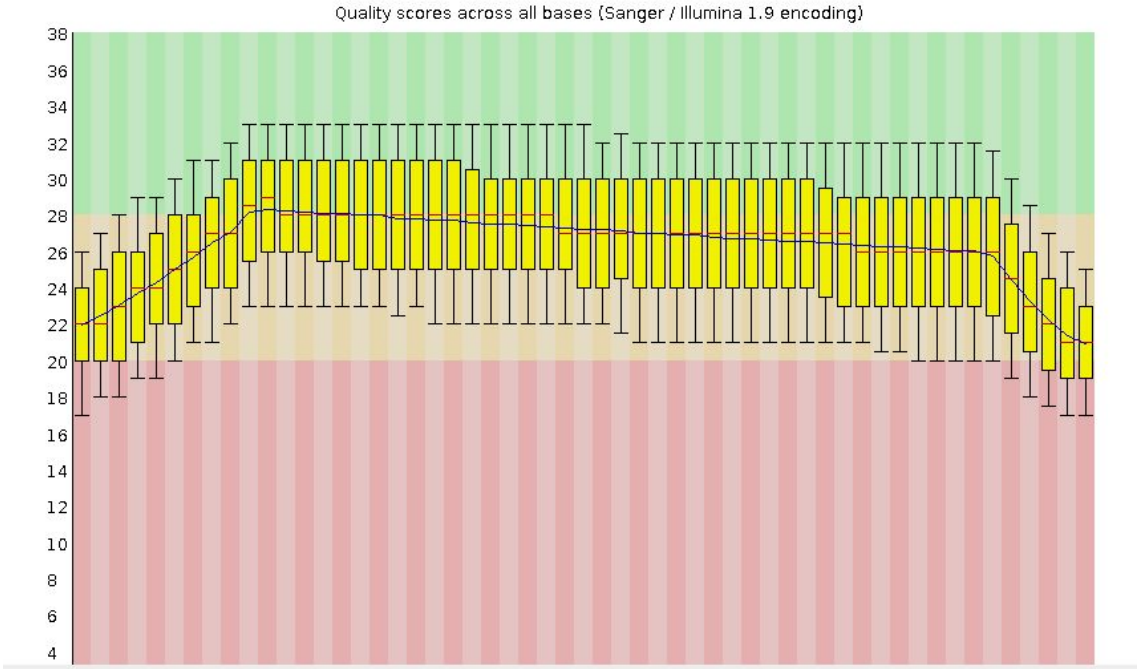
Check QC → Trim → Check QC again.

Trimming:
- Quality trimming
- Adapter trimming.

https://home.cc.umanitoba.ca/~psgendb/doc/TrimmomaticManual.pdf

Classwork 3: Trimming with Trimmomatic

1 Base Quality Trimming



Problem:
Base Quality of the reads in Sample1 are not good.

Requirement:
Increase the base quality score by filtering them with Trimmomatic and make sure that:

1) All reads have Q > 24
2) All reads have Q30

RAW DATA PROCESSING

Classwork 3: Trimming with Trimmomatic - Base Quality Trimming

Syntax:

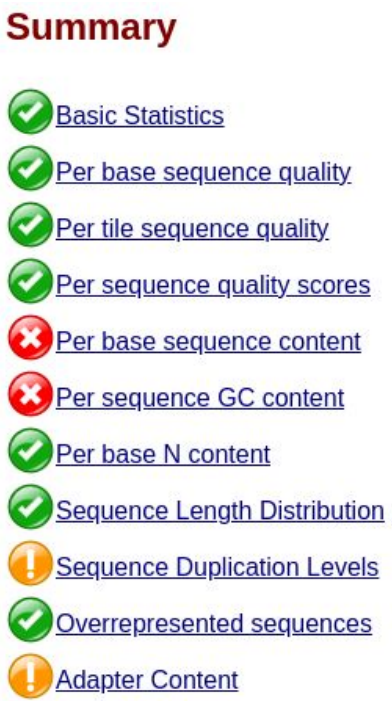
```
trimmomatic SE \
-phred33 \
-threads 4 \
-trimlog LowQuality_Reads.log \
$p_raw/sample1/LowQuality_Reads.fastq.gz \
$p_trim/sample1/LowQuality_Reads_trimmed.fastq.gz \
SLIDINGWINDOW:4:15 \
MINLEN:20
```

#Hint: Modify the SLIDINGWINDOW option.

SLIDINGWINDOW:4:15

Meaning: “Scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15”

Classwork 3: Trimming with Trimmomatic



2 Quality Trimming & Adapter removal

Problem:
Some parts of Sample2 are not good when checking with FASTQC.

Requirement:
Do trimming to increase Per base sequence content & remove adapter content. Check FASTQC after trim.

RAW DATA PROCESSING

Quality Trimming & Adapter removal

Syntax:

```
trimmomatic PE \
-phred33 \
-threads 4 \
-trimlog $p_trim/sample2/NIST7035.log \
$p_raw/sample2/NIST7035_TAAGGCGA_L001_R1_001.fastq.gz \
$p_raw/sample2/NIST7035_TAAGGCGA_L001_R2_001.fastq.gz \
$p_trim/sample2/NIST7035_TAAGGCGA_L001_R1_001_trimmed_paired.fastq.gz \
$p_trim/sample2/NIST7035_TAAGGCGA_L001_R1_001_trimmed_unpaired.fastq.gz \
$p_trim/sample2/NIST7035_TAAGGCGA_L001_R2_001_trimmed_paired.fastq.gz \
$p_trim/sample2/NIST7035_TAAGGCGA_L001_R2_001_trimmed_unpaired.fastq.gz \
ILLUMINACLIP:/home/duydao/dnaseq_work/tools/trimmomatic/share/trimmomatic-0.39-2/adapters/NexteraPE-PE.fa:2:30:10:8:3:true \
HEADCROP:10 \
LEADING:3 \
TRAILING:10 \
SLIDINGWINDOW:4:30 \
MINLEN:36
```

```
NexteraPE-PE.fa
TruSeq2-PE.fa
TruSeq2-SE.fa
TruSeq3-PE-2.fa
TruSeq3-PE.fa
TruSeq3-SE.fa
```

RAW DATA PROCESSING

TRIMMOMATIC WORKFLOW

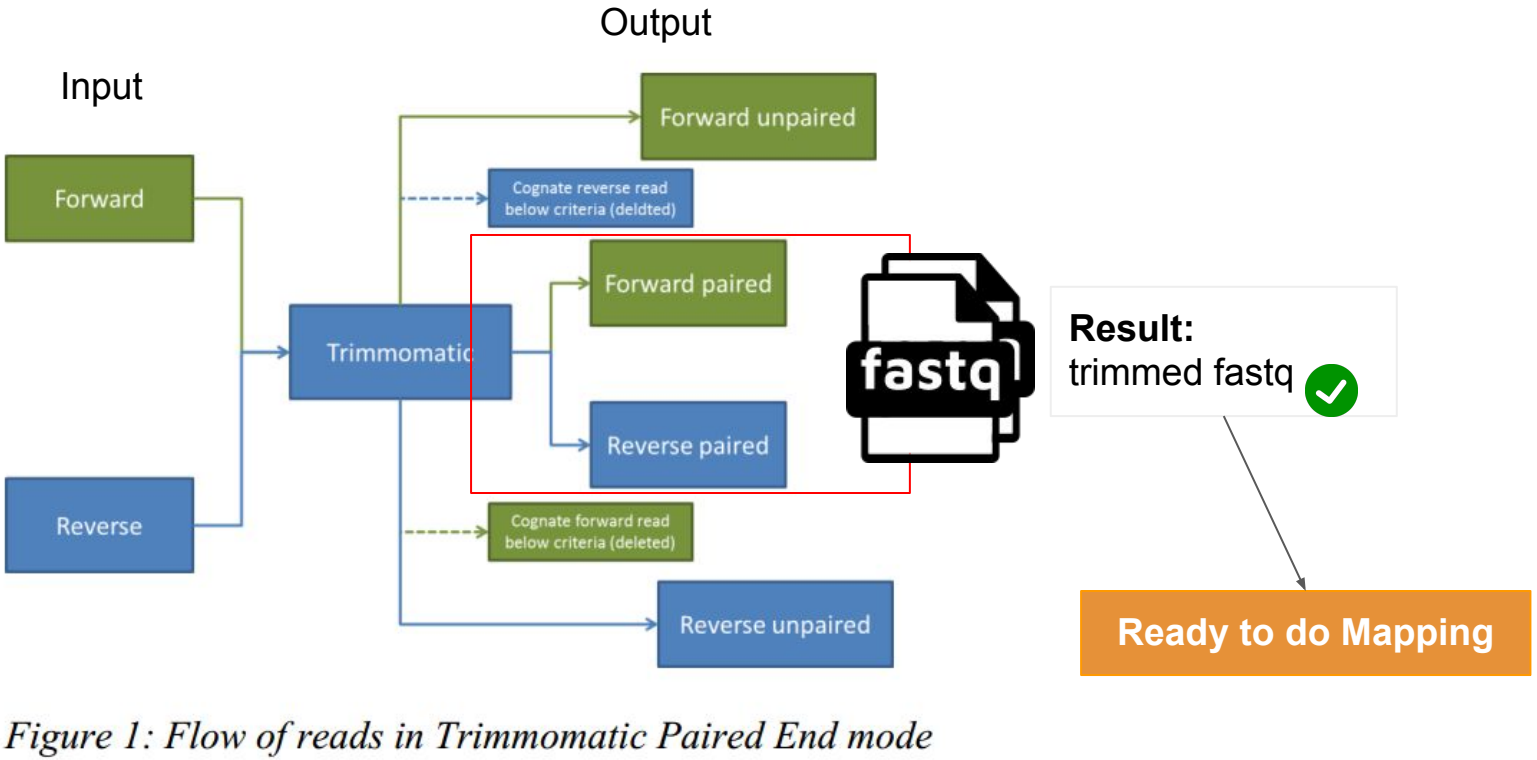


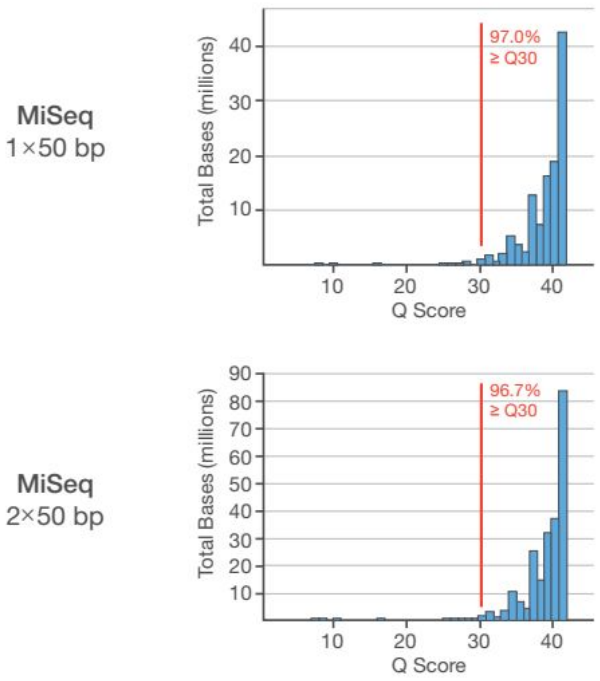
Figure 1: Flow of reads in Trimmomatic Paired End mode

RAW DATA PROCESSING

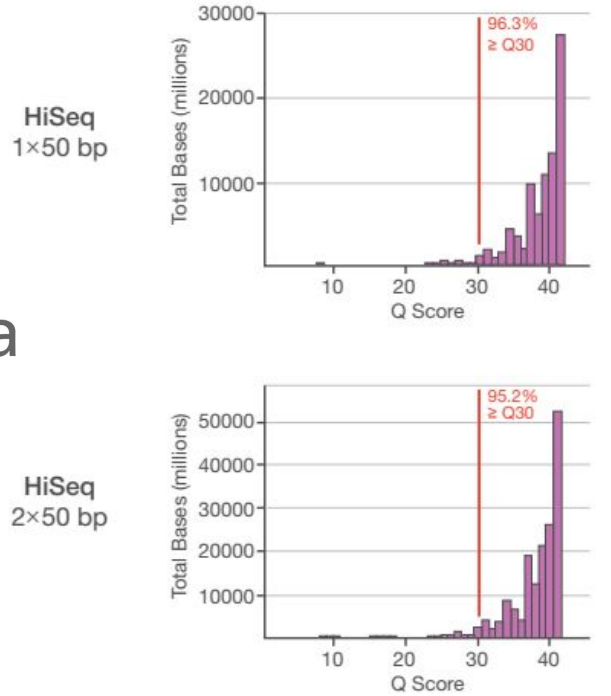
★ *Is trimming really necessary nowaday?*

RAW DATA PROCESSING

Illumina sequencers (2011)



illumina



At both 1 x 50 bp and 2 x 50 bp read lengths, virtually all bases are above Q30 across both the HiSeq and MiSeq systems.

PhiX quality scores for the MiSeq® and HiSeq® systems show that nearly all bases have scores > Q30 for single and paired-end reads (2011)

https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf

RAW DATA PROCESSING

MGI sequencers (2022)

WGS

Case 2: Human WGS

Sample: 1025 DNA samples of Han Chinese in the Central Plains
Library: MGIEasy PCR-Free DNA Library Preparation Set
Sequencing Strategy: DNBSEQ-G400 PE150
Results:

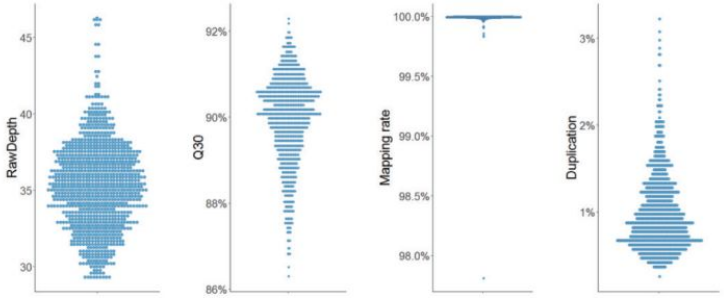


Figure 2. Excellent overall sequencing quality

Table 2-1 Sequencing data quality

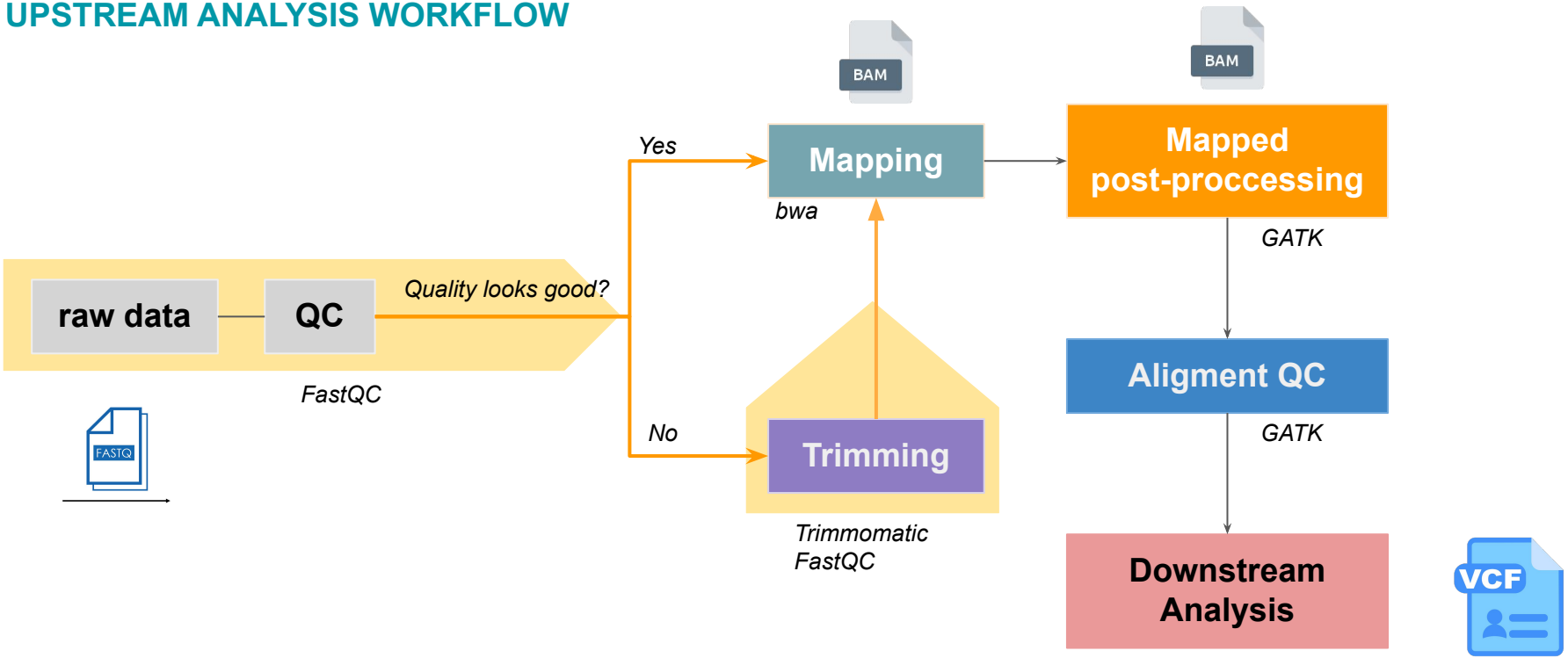
		Min	Median	Mean	Max	High quality	PASS
Total Reads		601727956	726056164	726494436	952285662	/	/
Mean Reads Length		150	150	150	150	/	/
Reads*	R1	100%	100%	100%	100%	=100%	=100%
	R2	100%	100%	100%	100%	=100%	=100%
Q30	R1	87.21%	90.43%	90.34%	92.91%	>=85%	>=80%
	R2	84.22%	89.79%	89.56%	92.00 %	>=85%	>=80%

*passed filter

https://en.mgi-tech.com/Download/download_file/id/22

RAW DATA PROCESSING

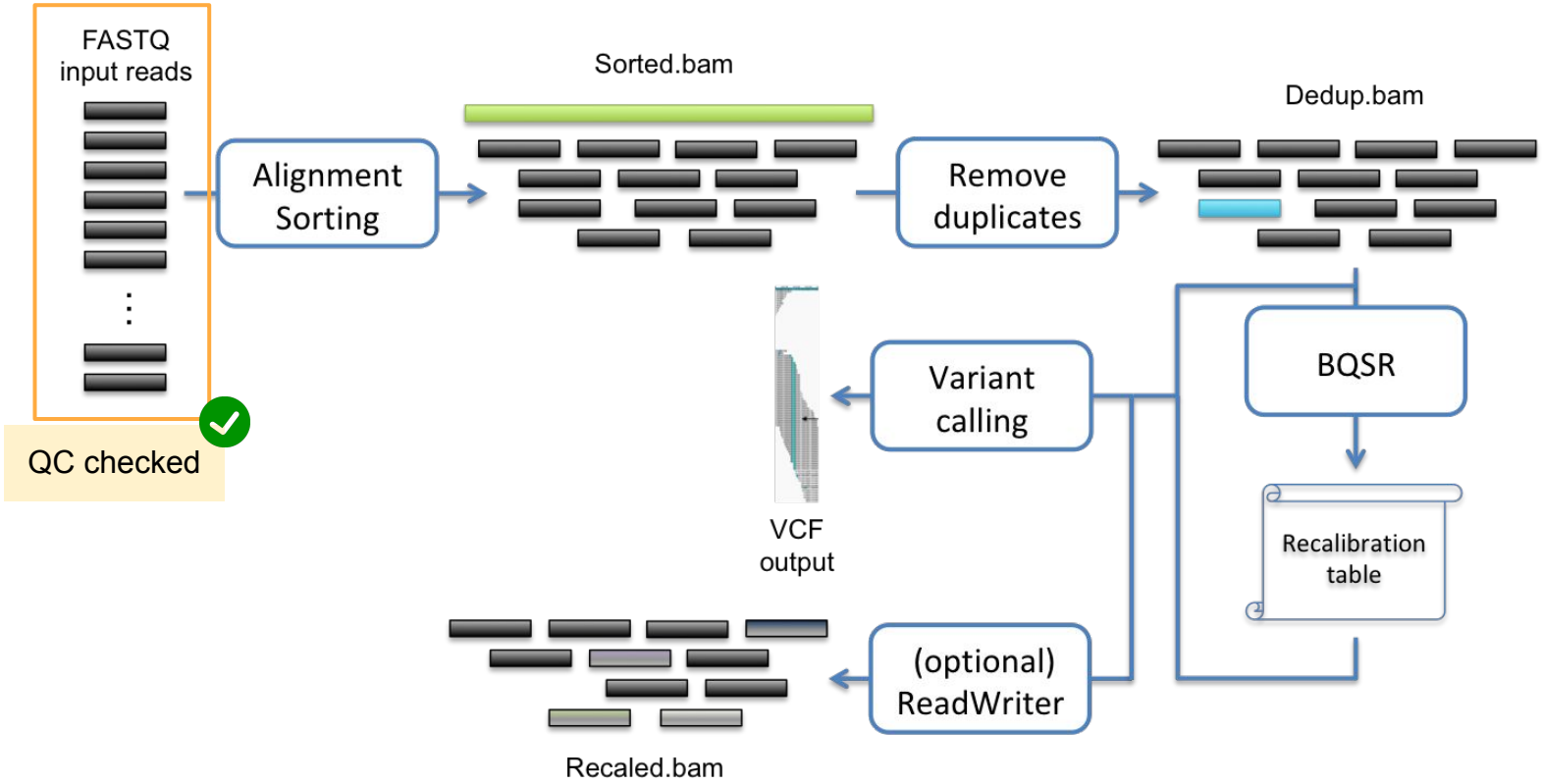
UPSTREAM ANALYSIS WORKFLOW



ALIGNMENT

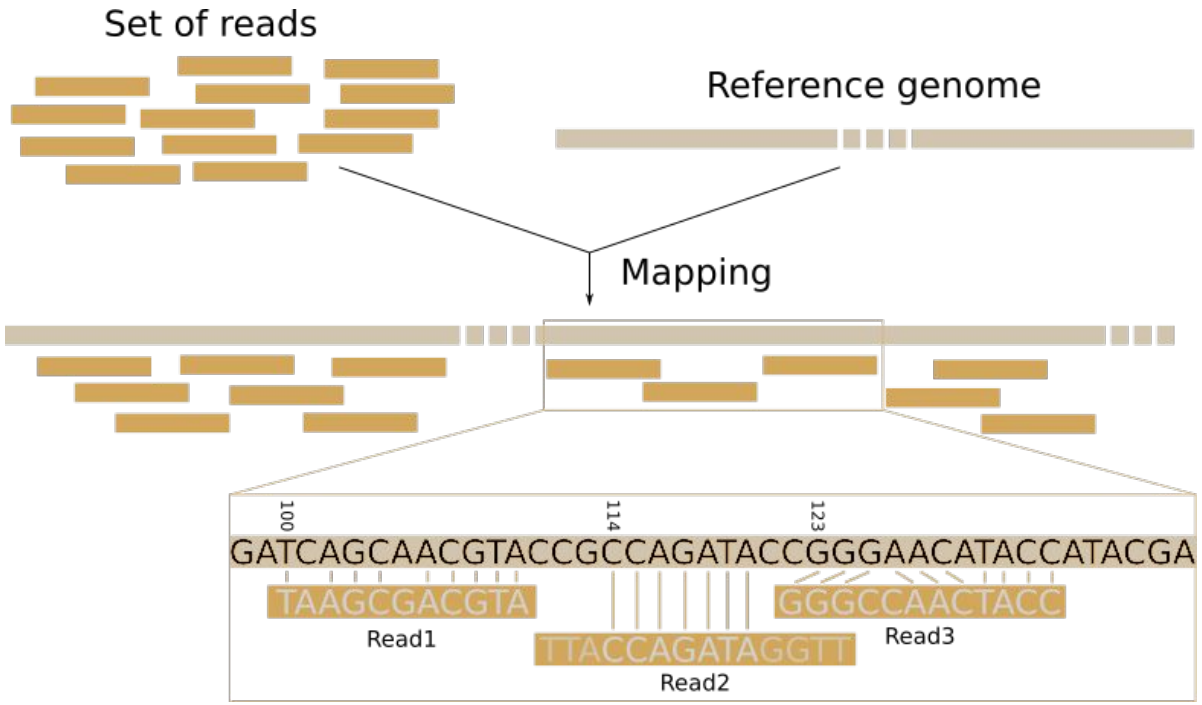
ALIGNMENT & MAPPED POST-PROCESSING

ALIGNMENT & POST-PROCESSING WORKFLOW



ALIGNMENT / MAPPING

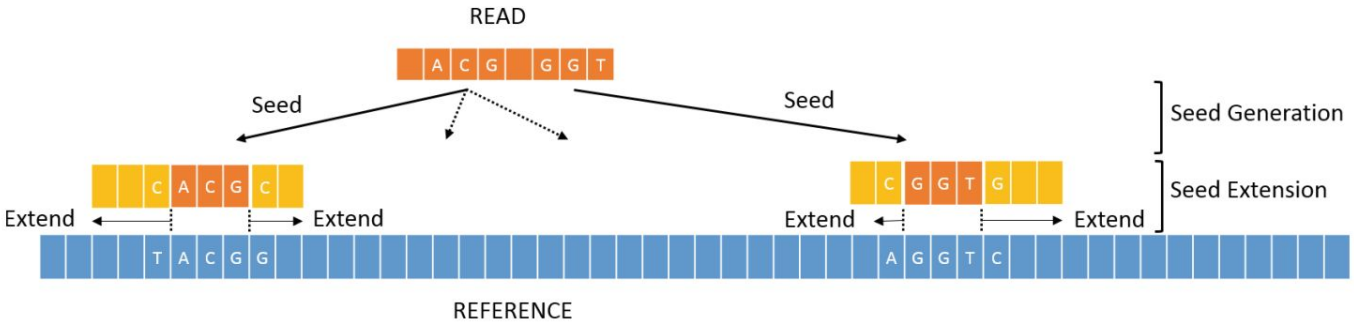
ALIGNMENT



ALIGNMENT

BWA: Burrows-Wheeler Alignment tool

- The most widely used aligners for Illumina data
- Based on BWT Algorithms
- **BWA-MEM** was developed for reads of ≥70 bases, for shorter reads it is advisable to use the standard BWA algorithm. Both are provided by the same tool.



BWA mem algorithm

ALIGNMENT

Hands-on: Using BWA to Align NIST7035 data (sample2). (This may take a while)

First: Index the reference genome.

Syntax:

```
bwa index -a bwtsw hs38DH.fa
```

When finished, we will have the following files like this:

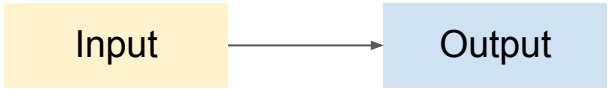
ALIGNMENT / MAPPING

ALIGNMENT

Hands-on: Using BWA to align NIST7035 data (sample2).

Syntax:

```
bwa mem -t 4 \
-R '@RG\tID:rg1\tSM:NA12878\tPL:illumina\tLB:lib1\tPU:H7AP8ADXX:1:TAAGGCGA' \
$p_ref/hg38/hs38DH.fa \
$p_trim/sample2/NIST7035_TAAGGCGA_L001_R1_001_trimmed_paired.fastq.gz \
$p_trim/sample2/NIST7035_TAAGGCGA_L001_R2_001_trimmed_paired.fastq.gz >
$p_align/sample2/NIST7035_aln.sam
```



- *R1.fastq*
- *R2.fastq*
- *reference.fa*

- *SAM/BAM*

ALIGNMENT / MAPPING

ALIGNMENT

Introduce to SAM/BAM file > SAM format

Sequence Alignment/Map (SAM) format is a tab-delimited text format that aims to be a universal format for storing alignments of NGS reads to a reference genome.

Header (begin with '@') + Alignment section.

```
@SQ     SN:chrUn_KI270744v1     LN:168472
@SQ     SN:chrUn_KI270745v1     LN:41891
@SQ     SN:chrUn_KI270746v1     LN:66486
@SQ     SN:chrUn_KI270747v1     LN:198735
@SQ     SN:chrUn_KI270748v1     LN:93321
@SQ     SN:chrUn_KI270749v1     LN:158759
@SQ     SN:chrUn_KI270750v1     LN:148850
@SQ     SN:chrUn_KI270751v1     LN:150742
@SQ     SN:chrUn_KI270752v1     LN:27745
@SQ     SN:chrUn_KI270753v1     LN:62944
@SQ     SN:chrUn_KI270754v1     LN:40191
@SQ     SN:chrUn_KI270755v1     LN:36723
@SQ     SN:chrUn_KI270756v1     LN:79590
@SQ     SN:chrUn_KI270757v1     LN:71251
@SQ     SN:chrX LN:156040895
@SQ     SN:chrX_KI270880v1_alt  LN:284869
@SQ     SN:chrX_KI270881v1_alt  LN:144206
@SQ     SN:chrX_KI270913v1_alt  LN:274009
@SQ     SN:chrY LN:57227415
@SQ     SN:chrY_KI270740v1_random       LN:37240
@RG     ID:rg1  SM:NA12878      PL:illumina     LB:lib1 PU:H7AP8ADXX:1:TAAGGCGA
@PG     ID:bwa  PN:bwa  VN:0.7.17-r1188 CL:bwa mem -t 8 -R @RG\tID:rg1\tSM:NA12878\tPL:illumina\tLB:lib1\tPU:H7AP8ADXX:1:TAAGGCGA /DA
@PG     ID:samtools     PN:samtools     PP:bwa  VN:1.13 CL:samtools view -Sb NIST7035_aln.sam
@PG     ID:samtools.1   PN:samtools     PP:samtools     VN:1.13 CL:samtools view -h NIST7035_aln.bam
HWI-D00119:50:H7AP8ADXX:1:1101:1322:2086        113     chr17   74243430        60      86M     =       74243430        0       GCCTC
HWI-D00119:50:H7AP8ADXX:1:1101:1322:2086        177     chr17   74243430        60      86M     =       74243430        0       GCCTC
```

Header of a SAM file

https://en.wikipedia.org/wiki/SAM_(file_format)

ALIGNMENT / MAPPING

ALIGNMENT

SAM/BAM file > SAM format

Alignment sections have 11 mandatory fields, as well as a variable number of optional fields.

Alignment section

Table 9.1. Mandatory fields of the SAM Format.

Col	Field	Description	Example
1	QNAME	Query template NAME	read_1
2	FLAG	Bitwise FLAG	0
3	RNAME	Reference sequence NAME	chrE
4	POS	Left-most mapping POSition (1-based)	11
5	MAPQ	MAPping Quality	37
6	CIGAR	CIGAR string	10M
7	RNEXT	Ref. name of the mate or NEXT read	*
8	PNEXT	Position of the mate or NEXT read	0
9	TLEN	Observed Template LENgth	0
10	SEQ	Segment SEQuence	ACGCATACTG
11	QUAL	Base QUALity string	DIGAFHHBCA

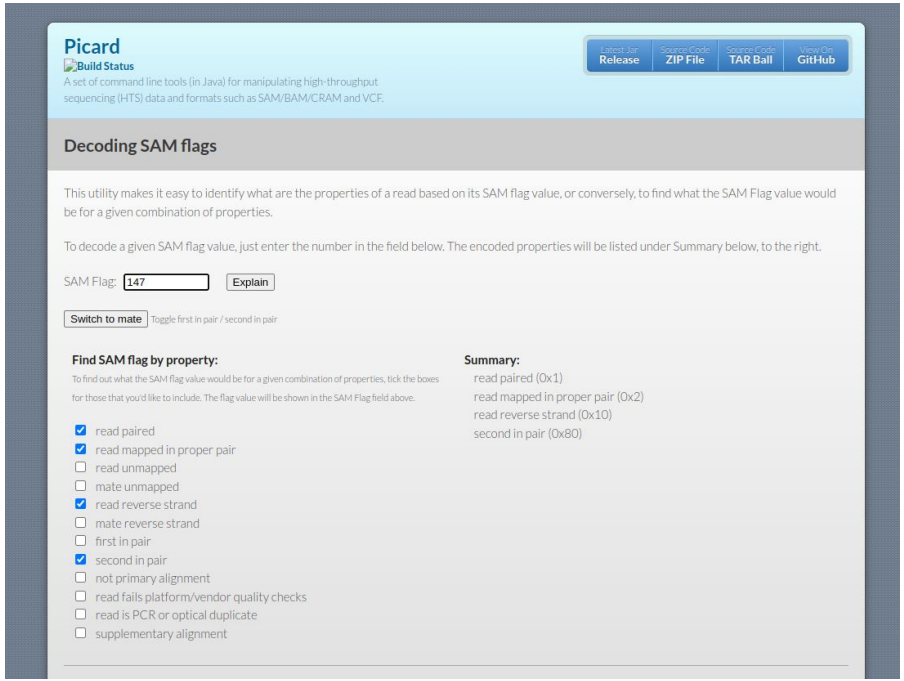
Note: Each line in the alignment section of a SAM file comprises 11 mandatory fields.

SAM format: 11 mandatory fields

ALIGNMENT / MAPPING

ALIGNMENT

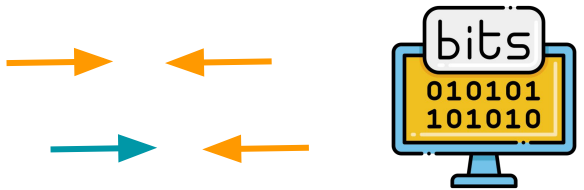
SAM/BAM file > Bit wise Flag



Website interpret bitflag

FLAG
113
177
65
129
65
129
65
129
81
161
113
177
113
177
113
177
65
129
113
177
65
129
65
129
65
129
65
129
113
177

"The bitwise flag is a 16-bit integer that encodes various properties of the read and its alignment to the reference genome."



https://broadinstitute.github.io/picard/explain-flags.html

ALIGNMENT

SAM/BAM file > Bit wise Flag

Flagstats

```
samtools flagstat NIST7035_aln.bam

#
28137263 + 0 in total (QC-passed reads + QC-failed reads)
28133082 + 0 primary
0 + 0 secondary
4181 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
28132795 + 0 mapped (99.98% : N/A)
28128614 + 0 primary mapped (99.98% : N/A)
28133082 + 0 paired in sequencing
14066541 + 0 read1
14066541 + 0 read2
0 + 0 properly paired (0.00% : N/A)
28128614 + 0 with itself and mate mapped
0 + 0 singletons (0.00% : N/A)
1382652 + 0 with mate mapped to a different chr
16 + 0 with mate mapped to a different chr (mapQ>=5)
```

The flagstat function of SAMtools provides a summary of the number of records corresponding to each of the bit flags.

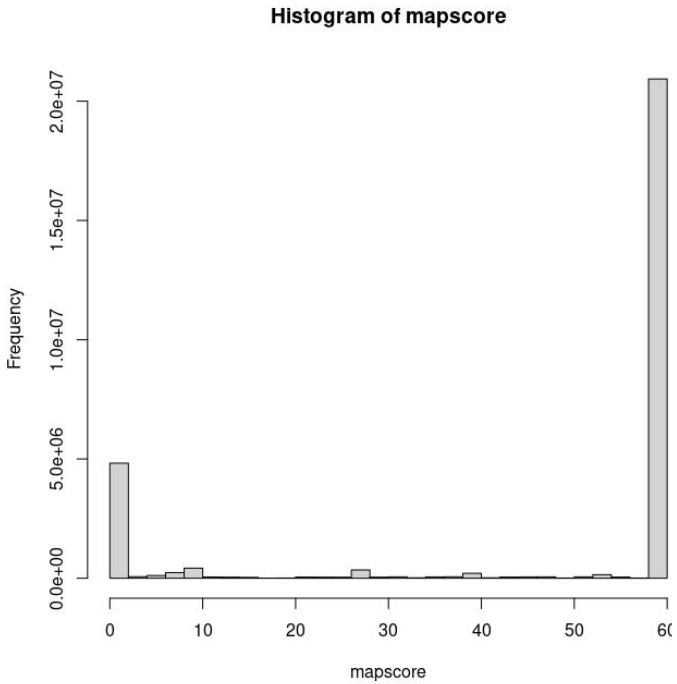
ALIGNMENT / MAPPING

ALIGNMENT

SAM/BAM file > MAPQ

MAPQ

```
60
60
0
0
60
60
27
27
46
46
0
0
60
60
60
60
0
0
60
60
60
60
```



“Mapping Quality Scores (MAPQ) quantify the probability that a read is misplaced.”

In BWA-MEM, MAPQ score:

- Based on Phred score
- Range: 0 to 60
- Higher scores indicating greater confidence in the mapping.
  - 0: read could map multiple locations.
  - 60: unique, highly confident mapping.
- Different mapping tools may produce a different MAPQ.

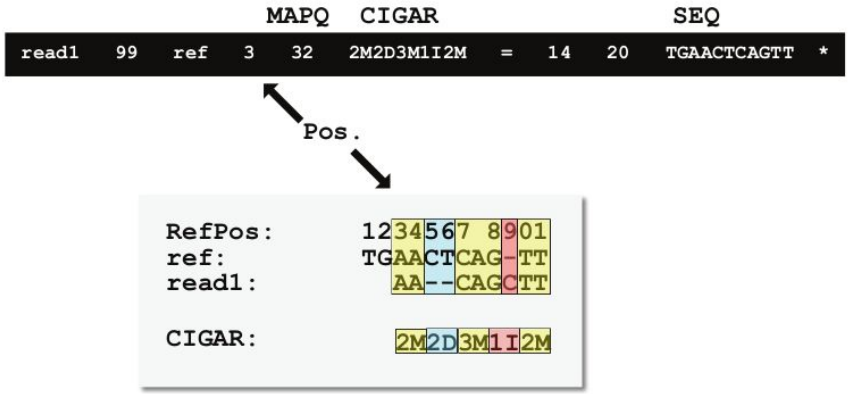
ALIGNMENT / MAPPING

ALIGNMENT

SAM/BAM file > CIGAR

CIGAR

86M
86M
86M
86M
86M
86M
86M
86M
86M
86M
86M
86M
86M
86M
86M
86M
86M
86M
86M
86M
86M
86M
86M



CIGAR: 2M2D3M1I2M

- 2 matches
- 2 deletes
- 3 matches
- 1 insert
- 2 matches

CIGAR string describes how each read aligns to the reference genome.

Syntax:

`<length><operation>`

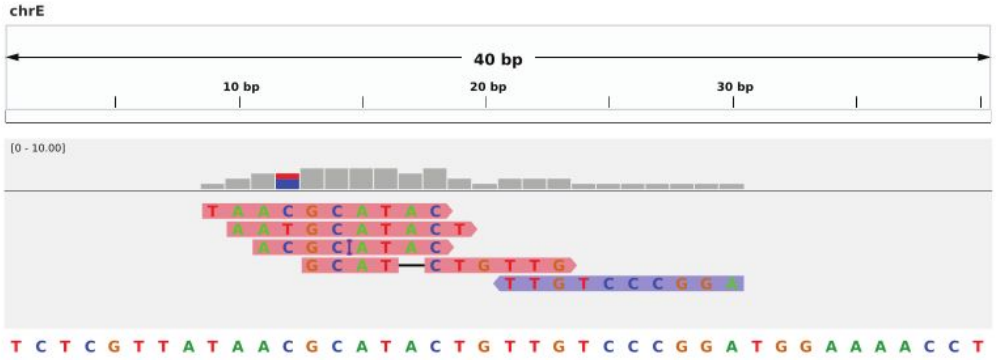
Table 9.2. CIGAR operations

Op	Description
M	alignment match (sequence match or mismatch)
I	insertion (additional non-reference base)
D	deletion (reference base missing in the read)
N	skipped region from the reference
S	soft clipping (clipped sequences still present in SEQ)
H	hard clipping (clipped sequences not present in SEQ)
P	padding (silent deletion from padded reference)
=	sequence match
X	sequence mismatch

Classwork 4: Interpret SAM file

```
@SQ SN:chrE LN:40
read_1   0 chrE  9 37    10M * 0 0 TAACGCATAC JJJJIGIHIJ
read_2   0 chrE 10 37    10M * 0 0 AATGCATACT JIJIIIIHGD
read_3   0 chrE 11 37 4M2I4M * 0 0 ACGCAAATAC IIIIIIHHFH
read_4   0 chrE 13 37 4M1D6M * 0 0 GCATCTGTTG JHHJIGIHIJ
read_5  16 chrE 21 37    10M * 0 0 TTGTCCCGGA HFHHIIIIII
```

(a) SAM file



(b) Integrative Genomics Viewer (IGV)

1

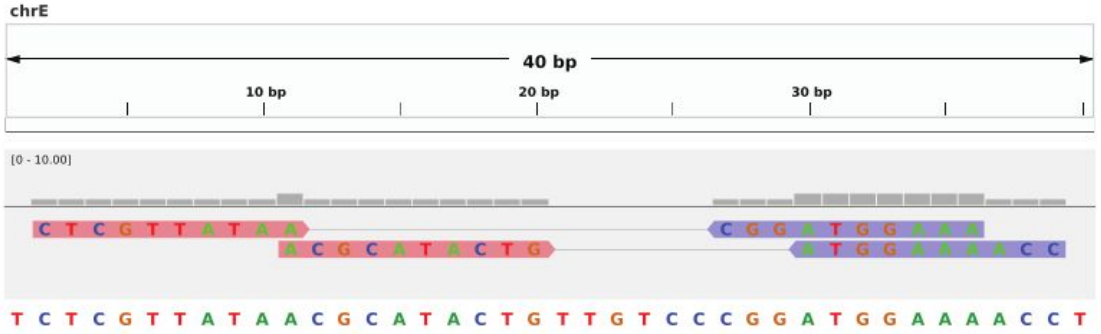
Interpret BAM:

- Header
- Alignment
  - QNAME
  - FLAG (0 & 16)
  - RNAME (chrE)
  - POS (9)
  - MAPQ (37)

Classwork 4: Interpret SAM file

```
@SQ SN:chrE LN:40
read_1  99 chrE  2 42 10M = 27  35 CTCGTTATAA JJJJIGIHIJ
read_2  99 chrE 11 42 10M = 30  29 ACGCATACTG JHHJIGIHIJ
read_1 147 chrE 27 42 10M =  2 -35 CGGATGGAAA DGHIIIIJIJ
read_2 147 chrE 30 42 10M = 11 -29 ATGGAAAACC HFHHIIIIII
```

(a) SAM file



(b) Integrative Genomics Viewer (IGV)

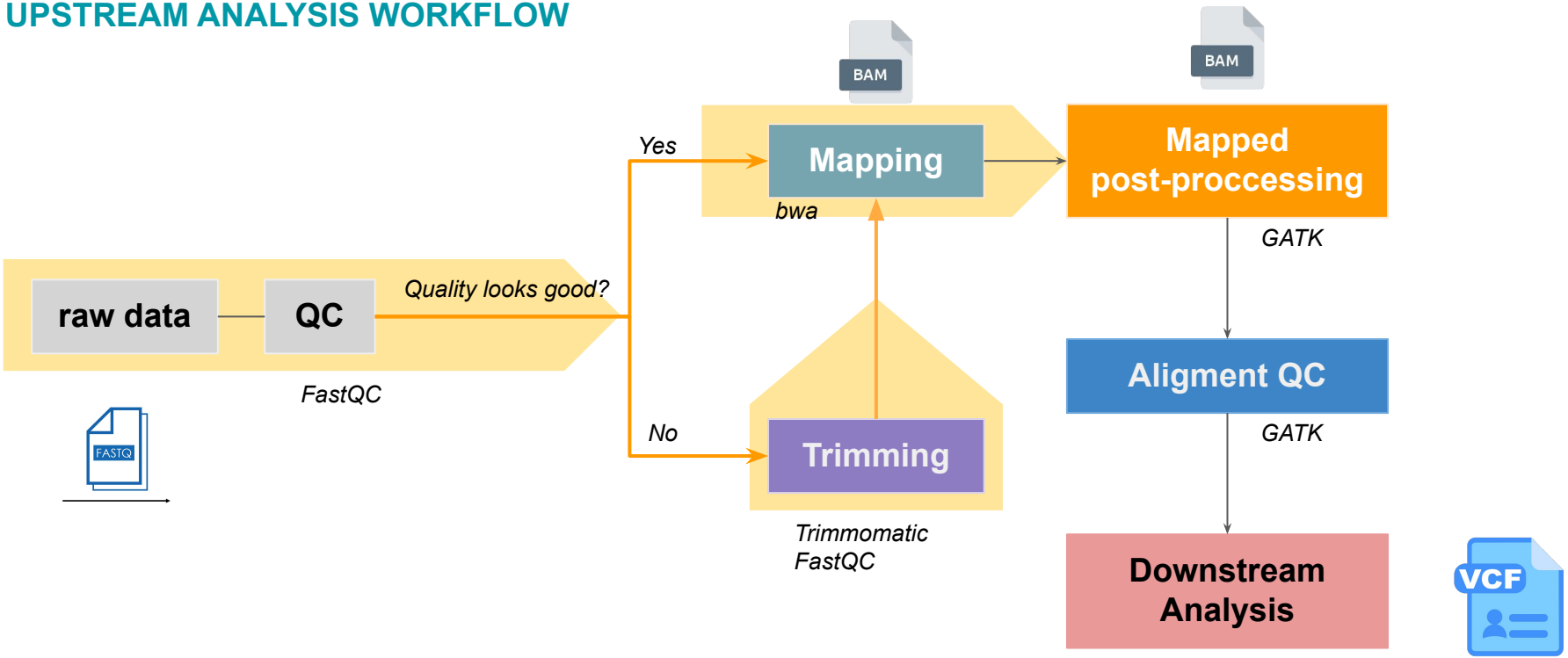
2

Interpret BAM:

- Header
- Alignment
  - QNAME
  - FLAG
  - RNAME
  - POS
  - MAPQ

ALIGNMENT / MAPPING

UPSTREAM ANALYSIS WORKFLOW



MAPPED READ POST-PROCESSING

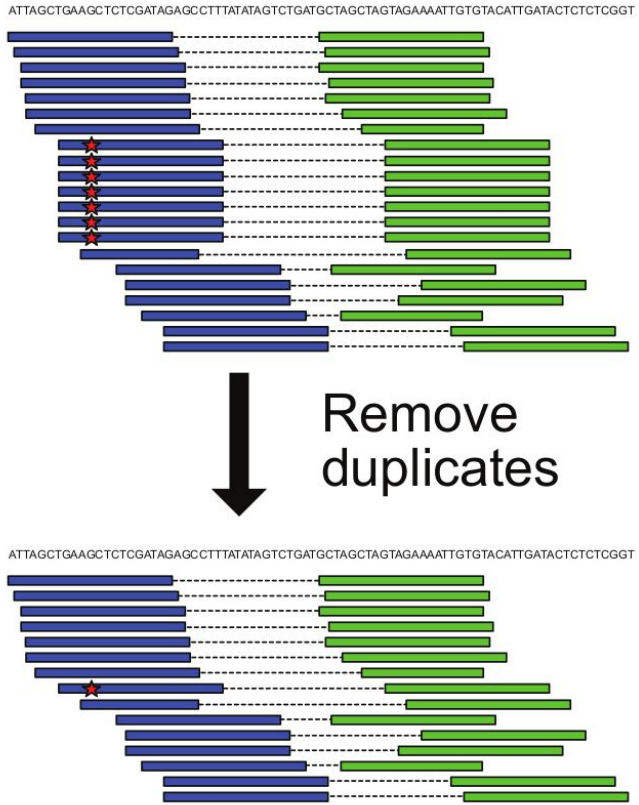
MAPPED READS POST-PROCESSING

Post-processing:

- Sorting, Indexing BAM file and Mark Duplicates
- Base Quality Score Recalibration

MAPPED READS POST-PROCESSING

Marking & removing duplicates

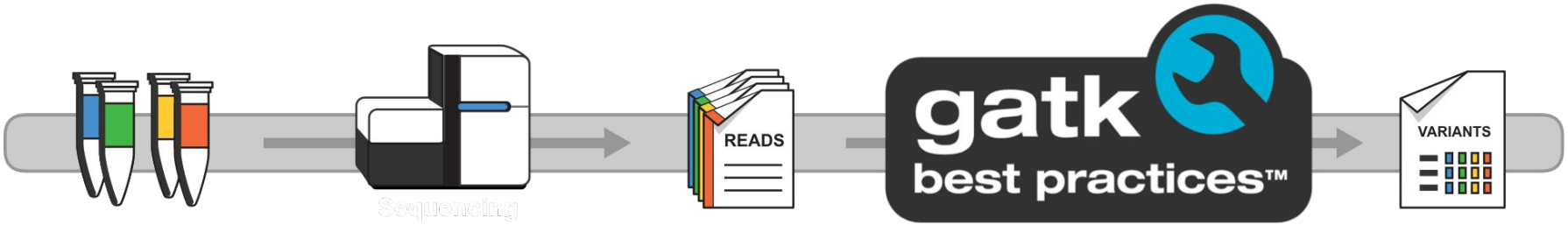


Duplicates read:

- PCR duplicates
- Optical duplicates

→ Duplicate reads can be problematic in downstream analyses, particularly in variant calling

MAPPED READS POST-PROCESSING

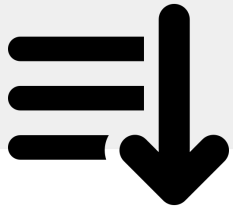


MAPPED READS POST-PROCESSING

Marking & removing duplicates

Sort the BAM file by coordinate

```
gatk SortSam \
--INPUT $p_align/sample2/NIST7035_aln.bam \
--OUTPUT $p_align/sample2/NIST7035_sorted.bam \
--SORT_ORDER coordinate
```



Mark Duplicates

```
gatk MarkDuplicates \
--INPUT $p_align/sample2/NIST7035_sorted.bam \
--OUTPUT $p_align/sample2/NIST7035_dedup.bam \
--METRICS_FILE $p_align/sample2/NIST7035.metrics
```

MAPPED READS POST-PROCESSING

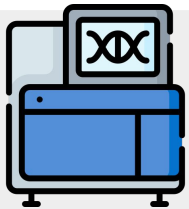
View the marked duplicate reads with flag

```
samtools view NIST7035_dedup.bam

samtools view -f 0x400 NIST7035_dedup.bam
```

Discriminate Optical and PCR duplication

Marked the read as an optical duplicated (DT:Z:SQ) & PCR duplicated (DT:Z:LB)

```
gatk MarkDuplicates \
--INPUT $p_align/sample2/NIST7035_sorted.bam \
--OUTPUT $p_align/sample2/NIST7035_dedup.bam \
--METRICS_FILE $p_align/sample2/NIST7035.metrics2 \
--TAGGING_POLICY All

samtools view -f 0x400 NIST7035_dedup.bam | grep DT:Z:SQ | less -S
```



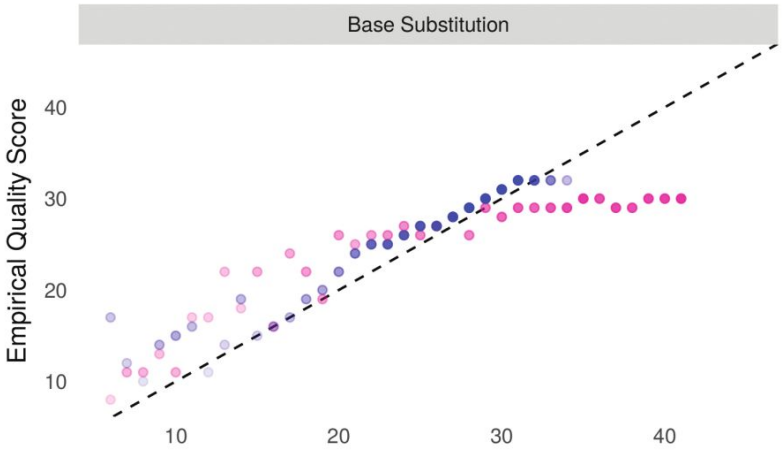
MAPPED READS POST-PROCESSING

Remove Duplicate

```
gatk MarkDuplicates \
--INPUT $p_align/sample2/NIST7035_sorted.bam \
--OUTPUT $p_align/sample2/NIST7035_remove_dup.bam \
--METRICS_FILE $p_align/sample2/NIST7035_remove_dup.metrics2 \
--REMOVE_DUPLICATES true
```

MAPPED READS POST-PROCESSING

BASE QUALITY SCORE RECALIBRATION



Base quality recalibration involves using a statistical model → adjust the base quality scores

Based on various features:

- Position of the base in the read,
- The sequence context,
- Quality scores of nearby bases.

→ Identify and correct for patterns of errors that are not random, but rather systematic and predictable.

MAPPED READS POST-PROCESSING

BASE QUALITY SCORE RECALIBRATION

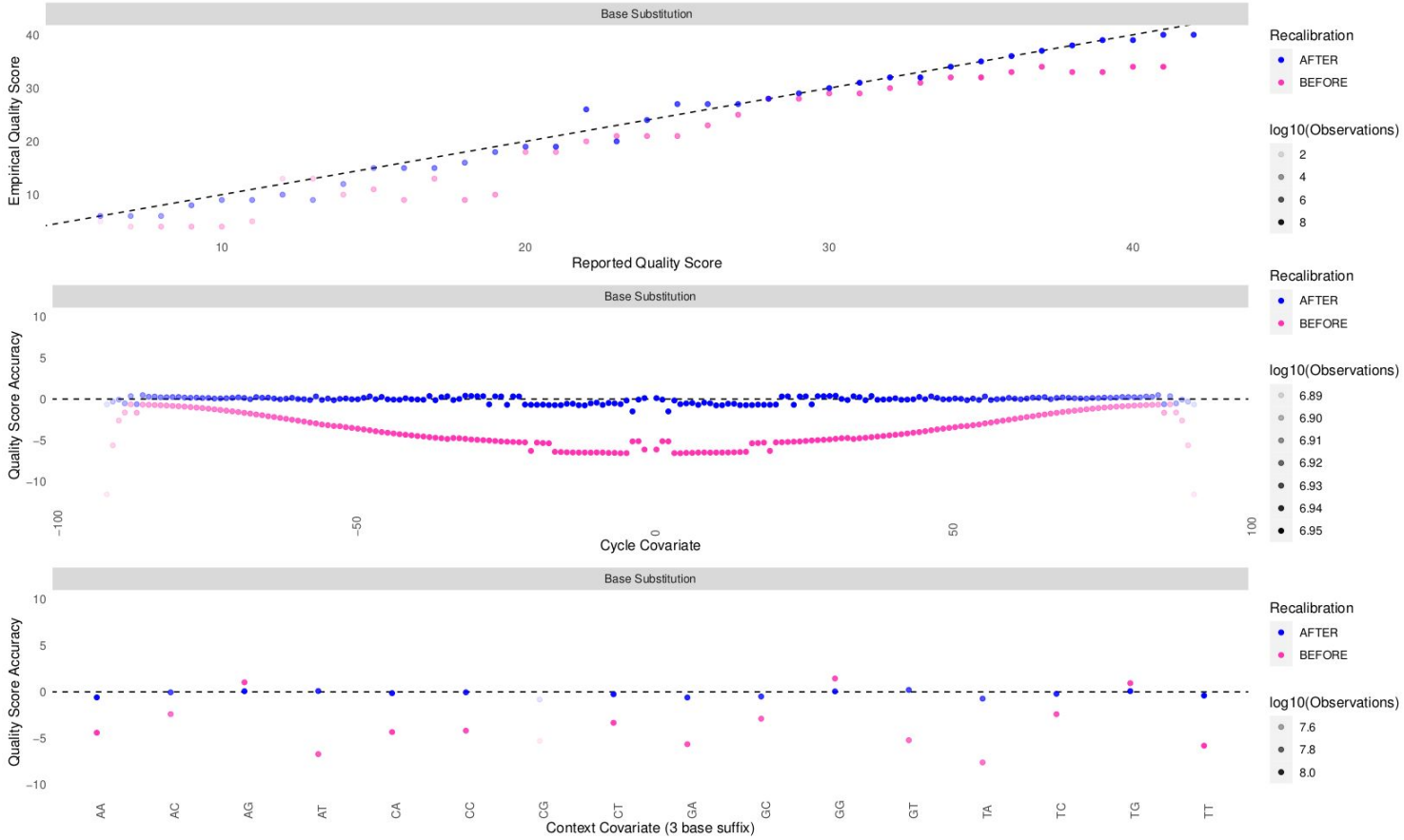
```
gatk BaseRecalibrator \
-R $p_ref/hg38/hs38DH.fa \
-I $p_align/sample2/NIST7035_dedup.bam \
-known-sites /mnt/portable_drive/Homo_sapiens_assembly38.dbsnp138.vcf \
-O $p_align/sample2/NIST7035-recal.table
#
gatk ApplyBQSR \
-R $p_ref/hg38/hs38DH.fa \
-I $p_align/sample2/NIST7035_dedup.bam \
-bqsr $p_align/sample2/NIST7035-recal.table \
-O $p_align/sample2/NIST7035_recal.bam
```

The base quality recalibration process typically involves the following steps:

- Generating a set of known variants
- Calculating empirical base quality scores
- Applying the recalibration model
- Re-running downstream analyses

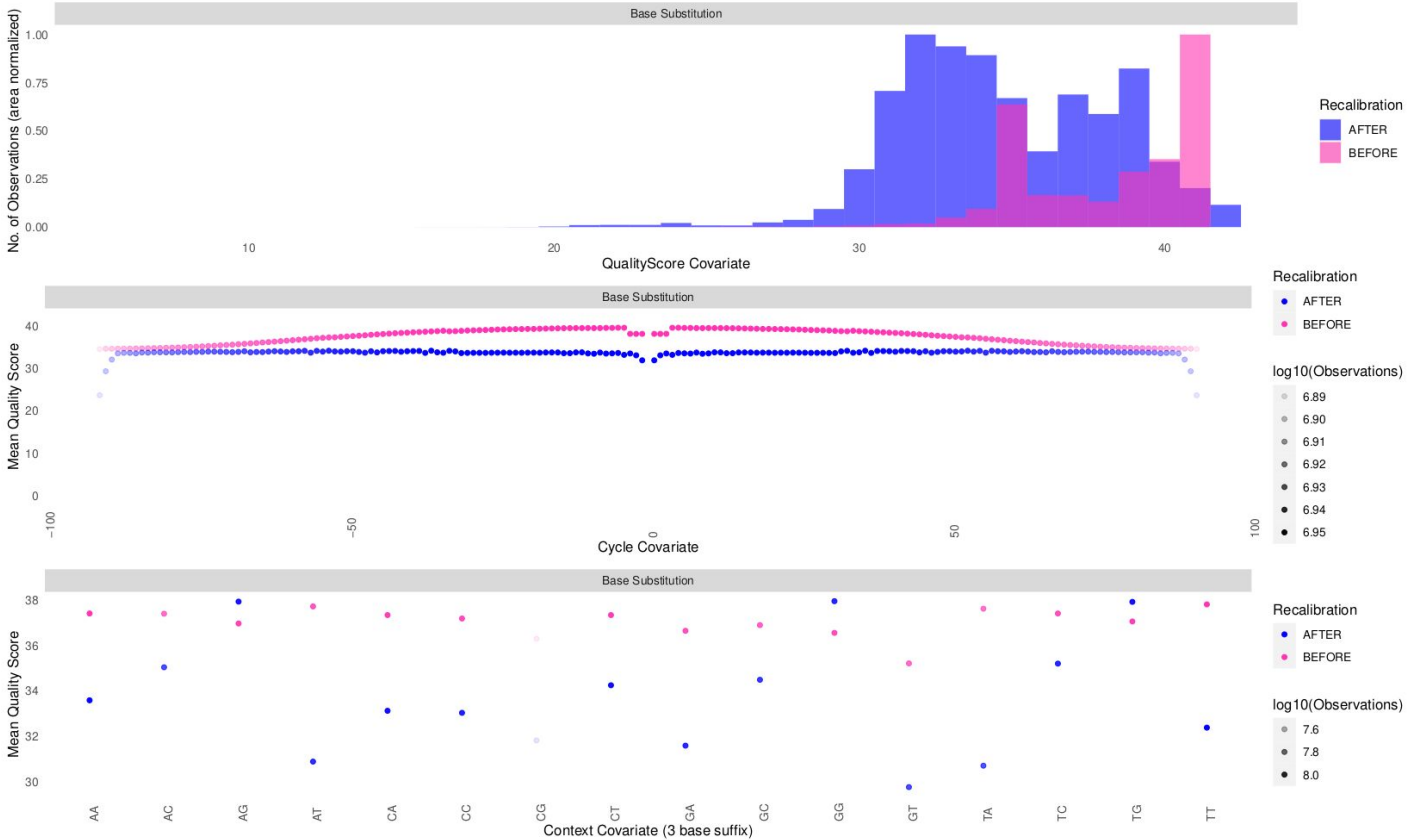
MAPPED READS POST-PROCESSING

BASE QUALITY SCORE RECALIBRATION



MAPPED READS POST-PROCESSING

BASE QUALITY SCORE RECALIBRATION

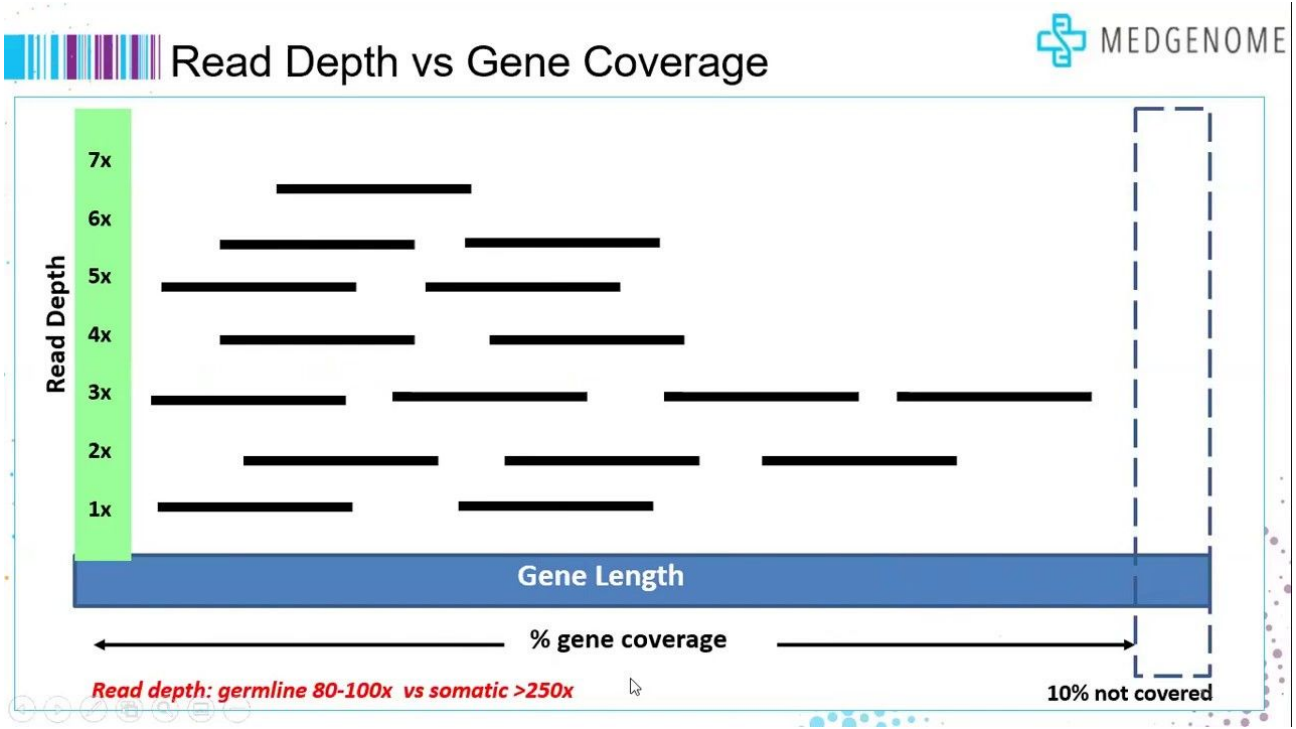


ALIGNMENT DATA: QUALITY CONTROL

2 key parameters for Alignment Quality Control:

- Read Depth
- Coverage

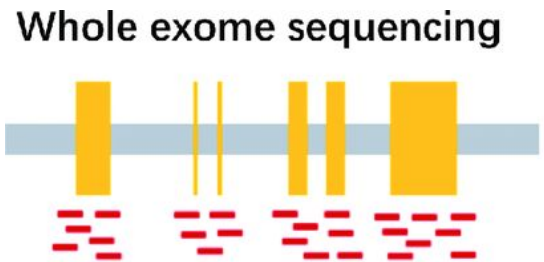
ALIGNMENT DATA: QUALITY CONTROL



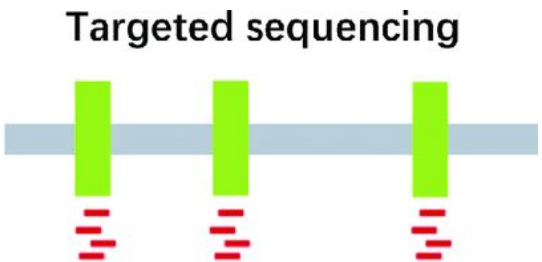
The more reads there are, the more certain we can be about the genotype at any given position.

ALIGNMENT DATA: QUALITY CONTROL

WES



- Region sequencing: whole exome
- Depth of coverage: 20X
- Identify all variants: SNV, INDELs, SV in coding Regions
- Cost effective



- Region sequencing: specific coding region
- Depth of coverage: 300X
- Identify alla variants: SNV, INDELs, SV in specific regions
- Most cost effective

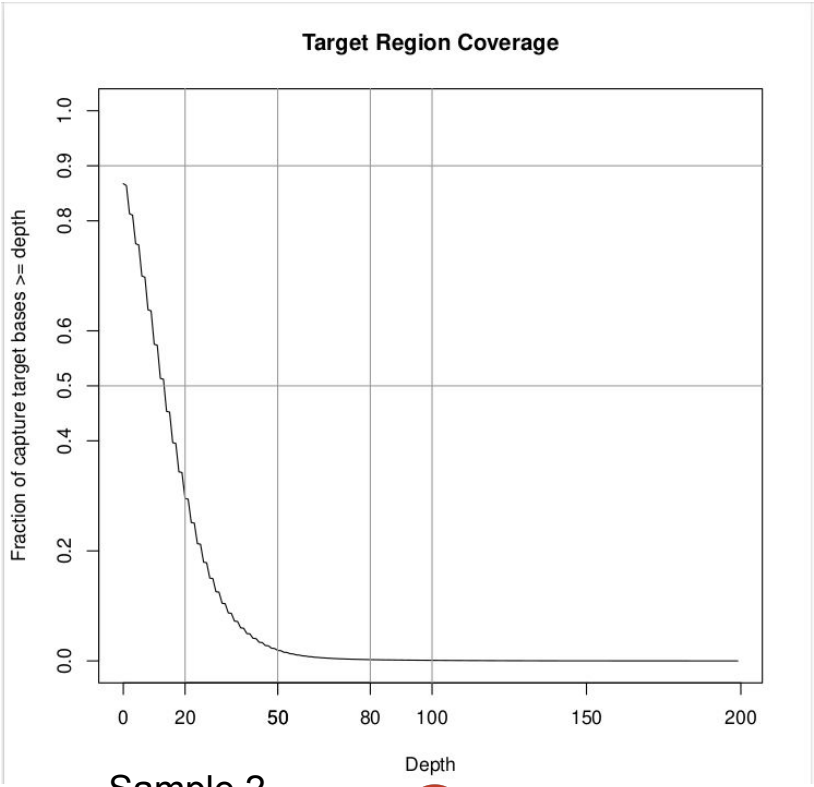

```
chr1    65564   65573
chr1    69036   70008
chr1    358066  358183
chr1    365564  365692
chr1    373143  373323
chr1    379768  379870
chr1    399040  399100
chr1    450739  451678
chr1    601397  601577
chr1    607954  608056
chr1    609082  609217
chr1    611111  611297
chr1    685715  686654
chr1    923455  923461
chr1    923615  924948
chr1    925921  926013
chr1    930154  930336
chr1    931038  931089
chr1    935771  935896
```

ALIGNMENT DATA: QUALITY CONTROL

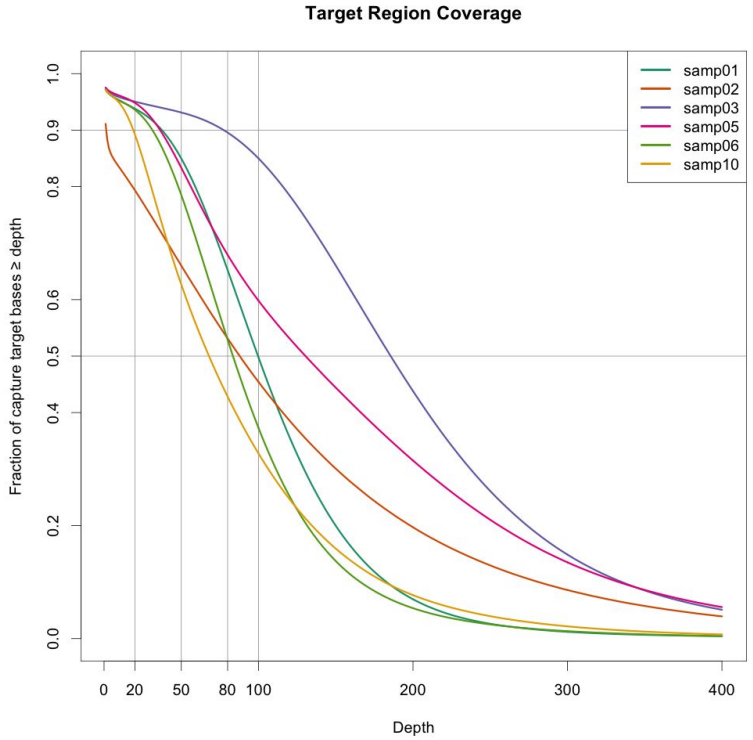
Calculate coverage

```
bedtools coverage \
-hist \
-a $p_ref/hg38_exome.bed \
-b $p_align/sample2/NIST7035_remove_dup.bam > NIST.bed.cov
#
grep ^all NIST.bed.cov > NIST.all.cov
```

ALIGNMENT DATA: QUALITY CONTROL



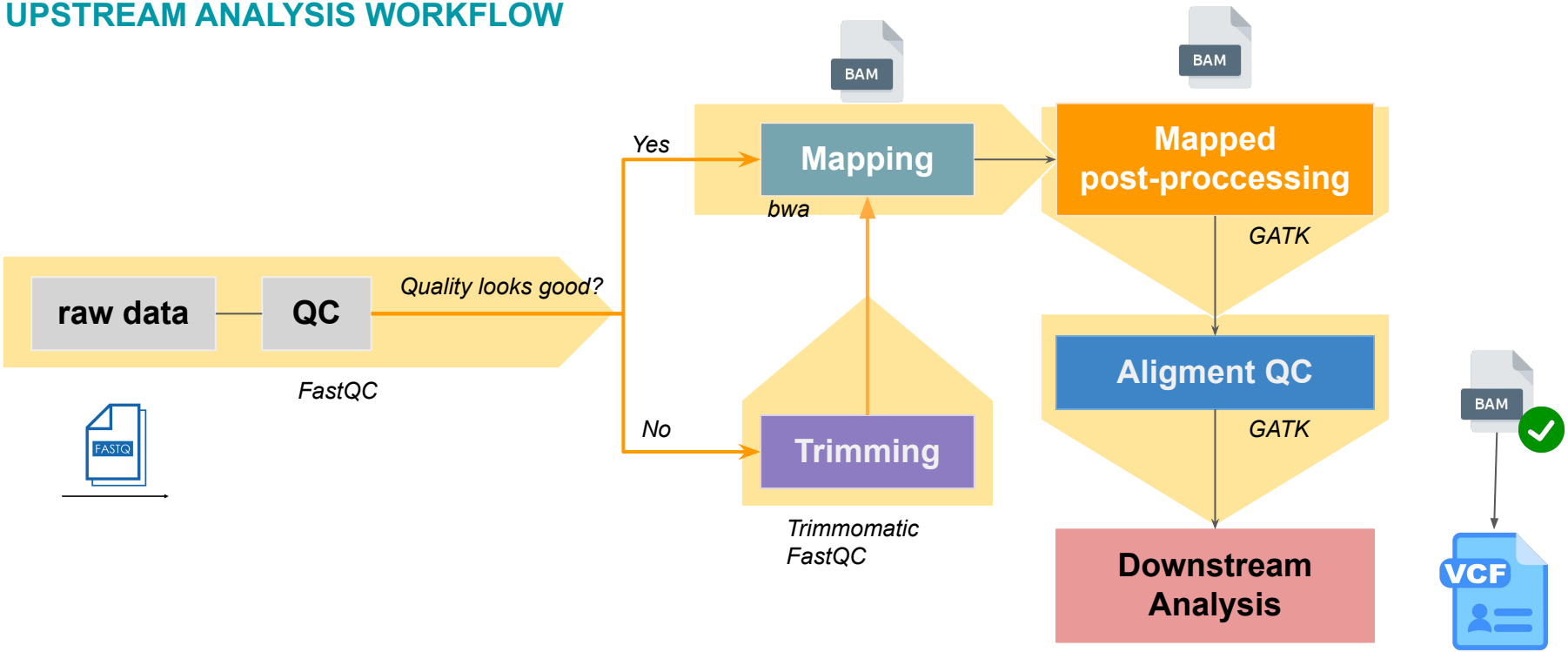
Sample 2



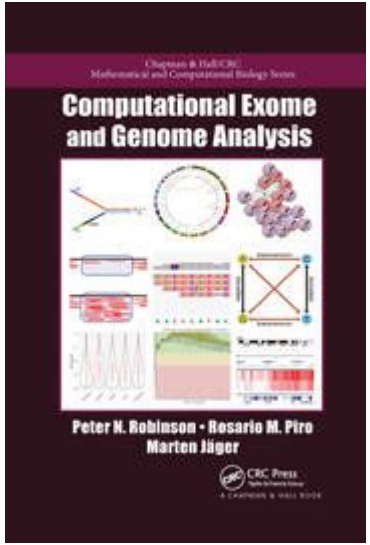
A depth of coverage of 30X or more is considered to be sufficient for most applications.

SUMMARY

UPSTREAM ANALYSIS WORKFLOW



PREFERENCE



Robinson, P.N., Piro, R.M., & Jager, M. (2017). Computational Exome and Genome Analysis (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315154770

THANK YOU