H&TRC
Health and Technology Research Center

ESCOLA SUPERIOR DE TECNOLOGIA DA SAÚDE DE LISBOA
INSTITUTO POLITÉCNICO DE LISBOA

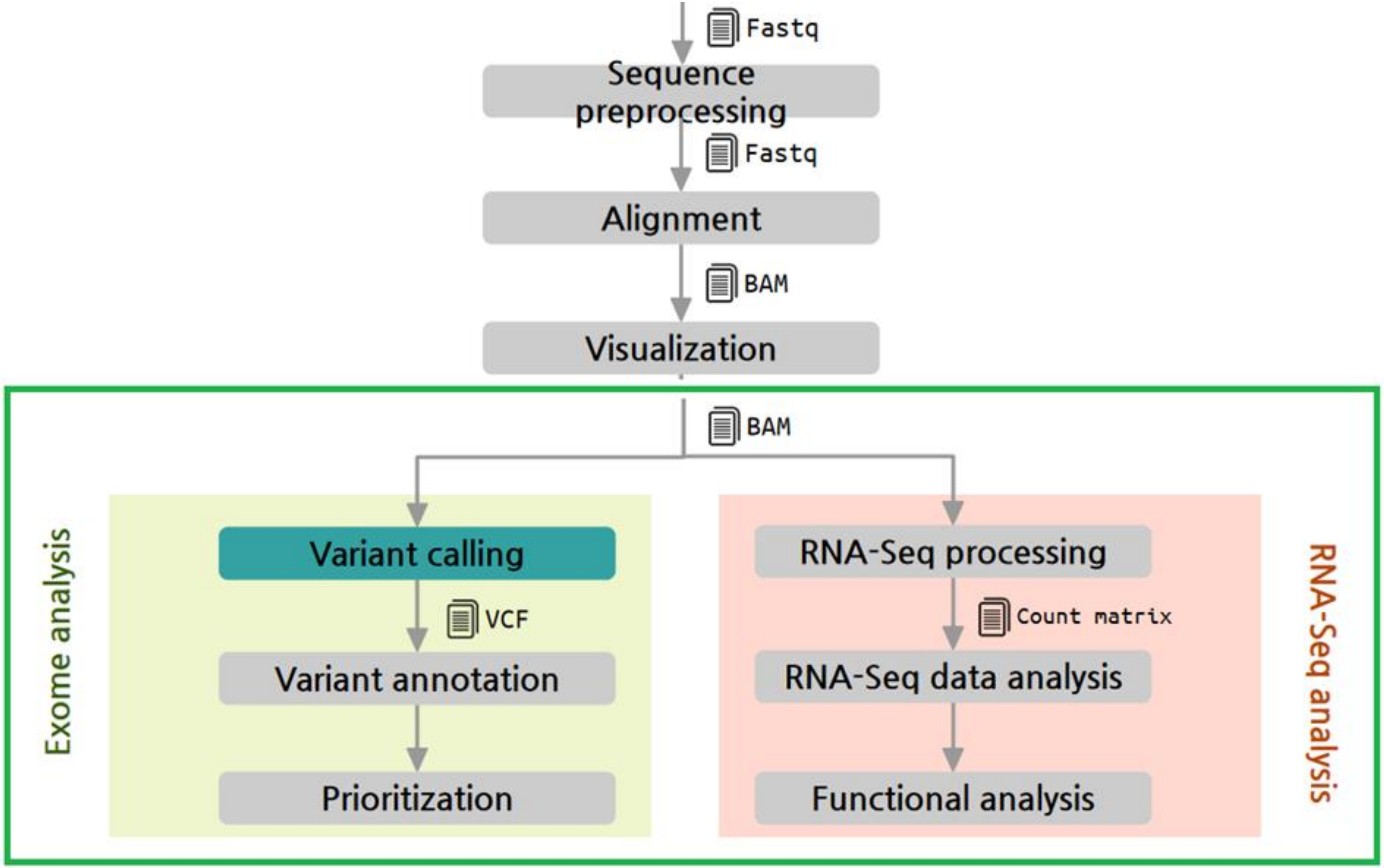
NEXT GENERATION SEQUENCING (NGS)

Introduction to NGS data analysis
Parte II

Carina Silva

2 – 4 ABRIL 2018, ESTeSL-IPL

AGENDA



Fastq

Sequence preprocessing

Fastq

Alignment

BAM

Visualization

BAM

Exome analysis

Variant calling

VCF

Variant annotation

Prioritization

RNA-Seq analysis

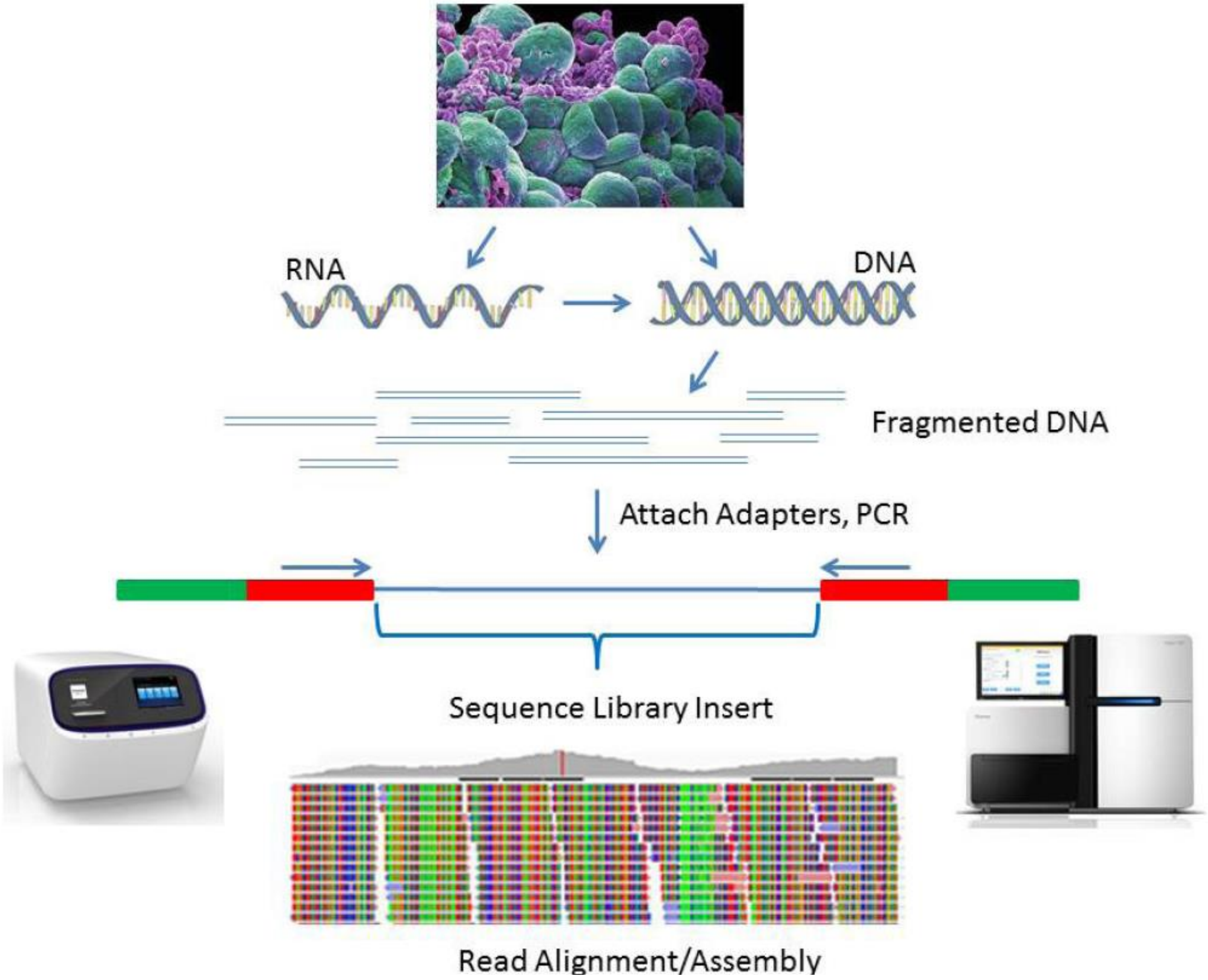
RNA-Seq processing

Count matrix

RNA-Seq data analysis

Functional analysis

A (very simple) NGS Workflow for library preparation



RNA

DNA

Fragmented DNA

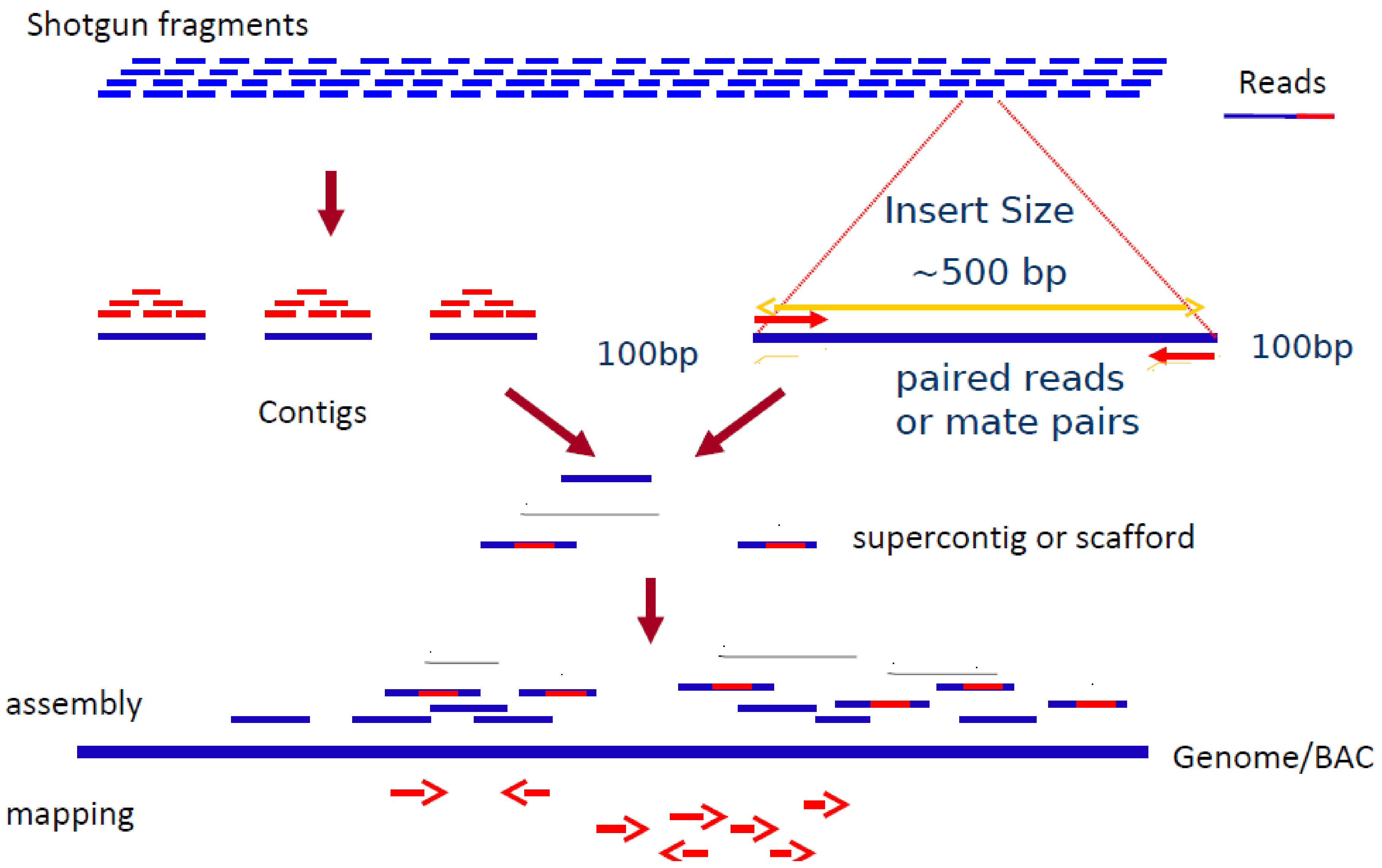
Attach Adapters, PCR

Sequence Library Insert

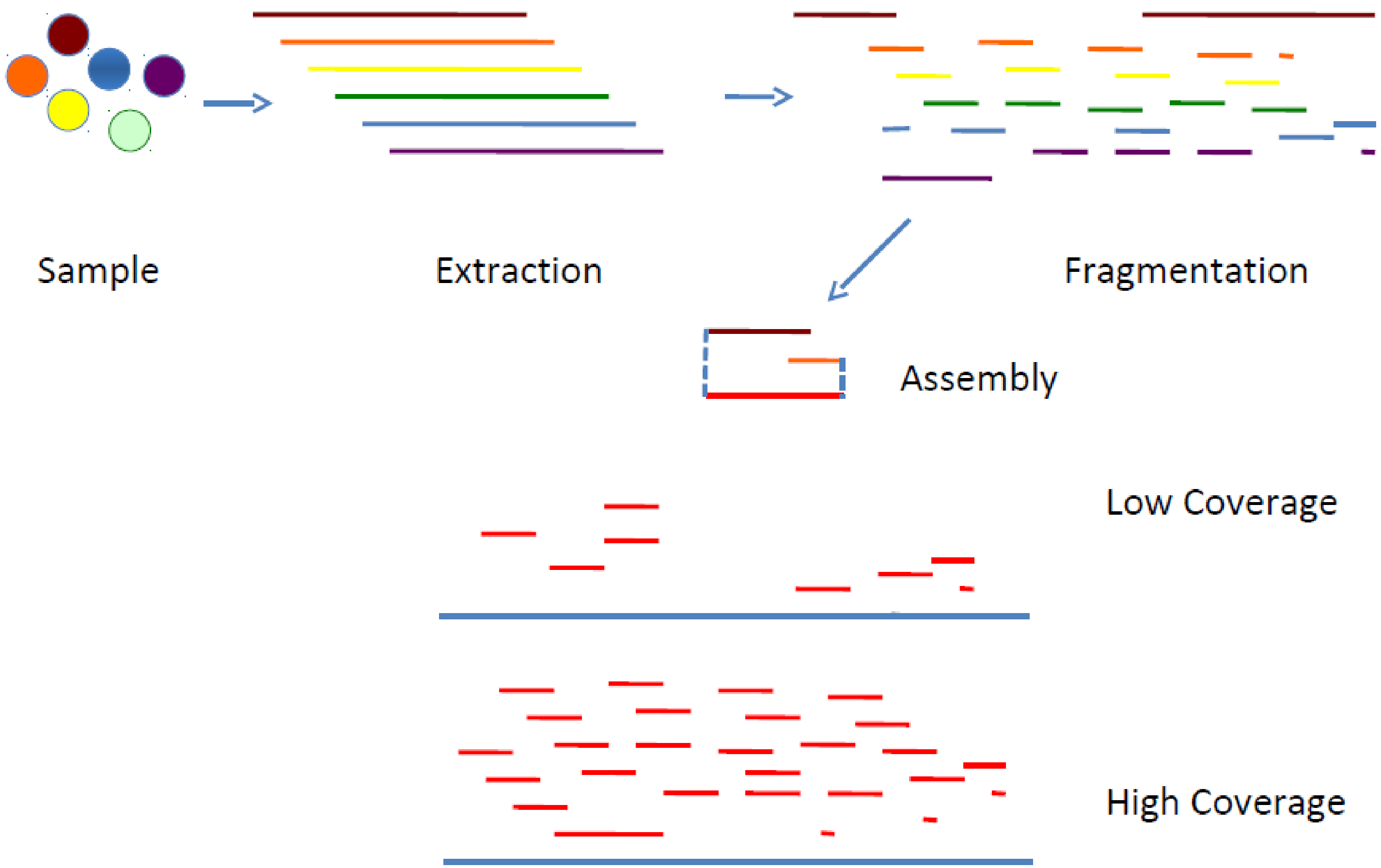
Read Alignment/Assembly

Head *et al.* (2014). Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*, 56: 61-77.

NGS Terminology


Shotgun fragments
Reads
Insert Size
~500 bp
100bp
100bp
paired reads
or mate pairs
Contigs
supercontig or scafford
assembly
Genome/BAC
mapping

NGS Terminology – Shotgun strategy



Sample

Extraction

Fragmentation

Assembly

Low Coverage

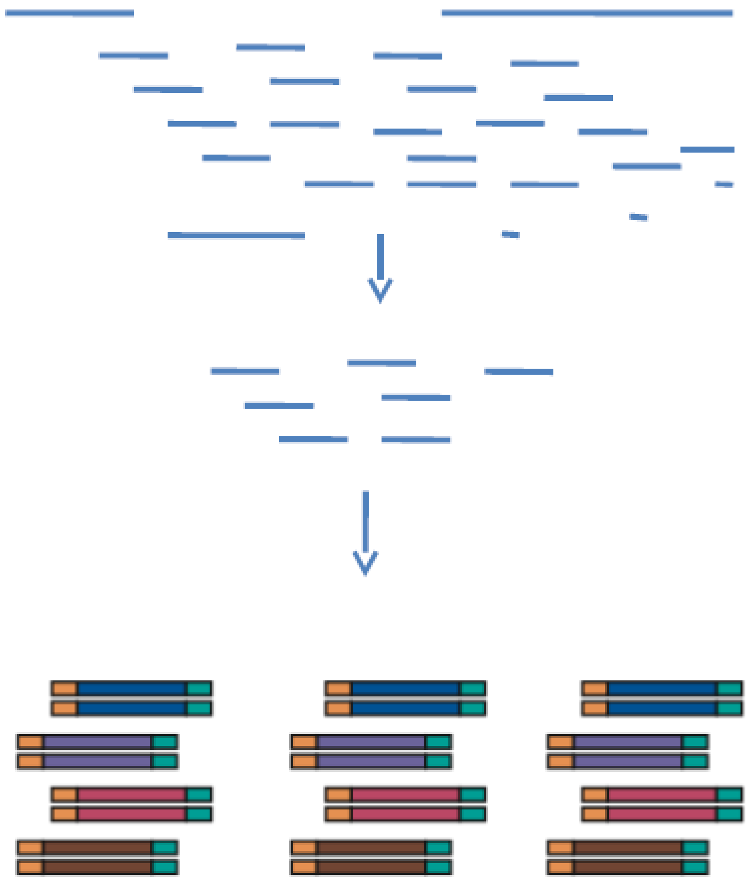
High Coverage

DNA is broken up randomly into numerous small segments, which are sequenced using the chain termination method to obtain ***reads***.

Multiple overlapping reads for the target DNA are obtained by performing several rounds of this fragmentation and sequencing – ***contig****.*

Coverage – (read depth or depth) is the average number of reads representing a given nucleotide in the reconstructed sequence. It can be calculated from the length of the original genome (G), the number of reads (N) and the average read length (L) as N*(L/G).

NGS Terminology – Library



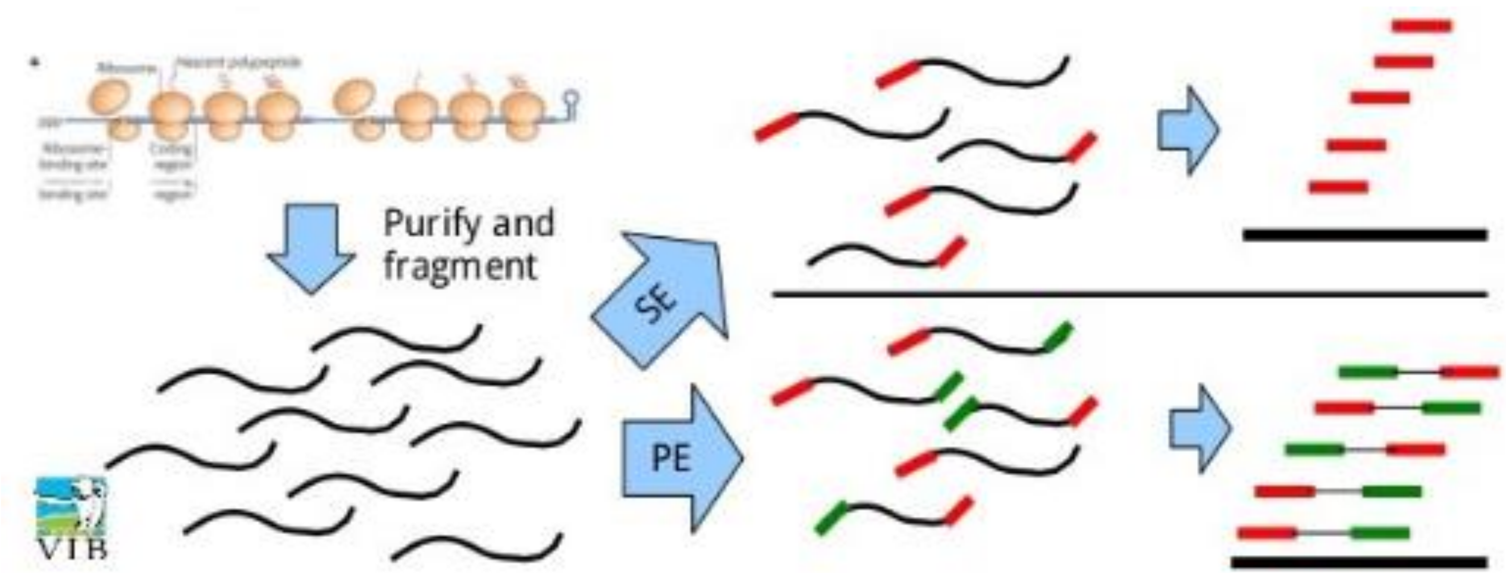
Size Selection

Library construction

During the library preparation stage, the sample DNA is fragmented, and the fragments of a specific size (typically 200–500 bp, but can be larger) are ligated or "inserted" in between two oligo adapters. The original sample DNA fragments are also referred to as "inserts".

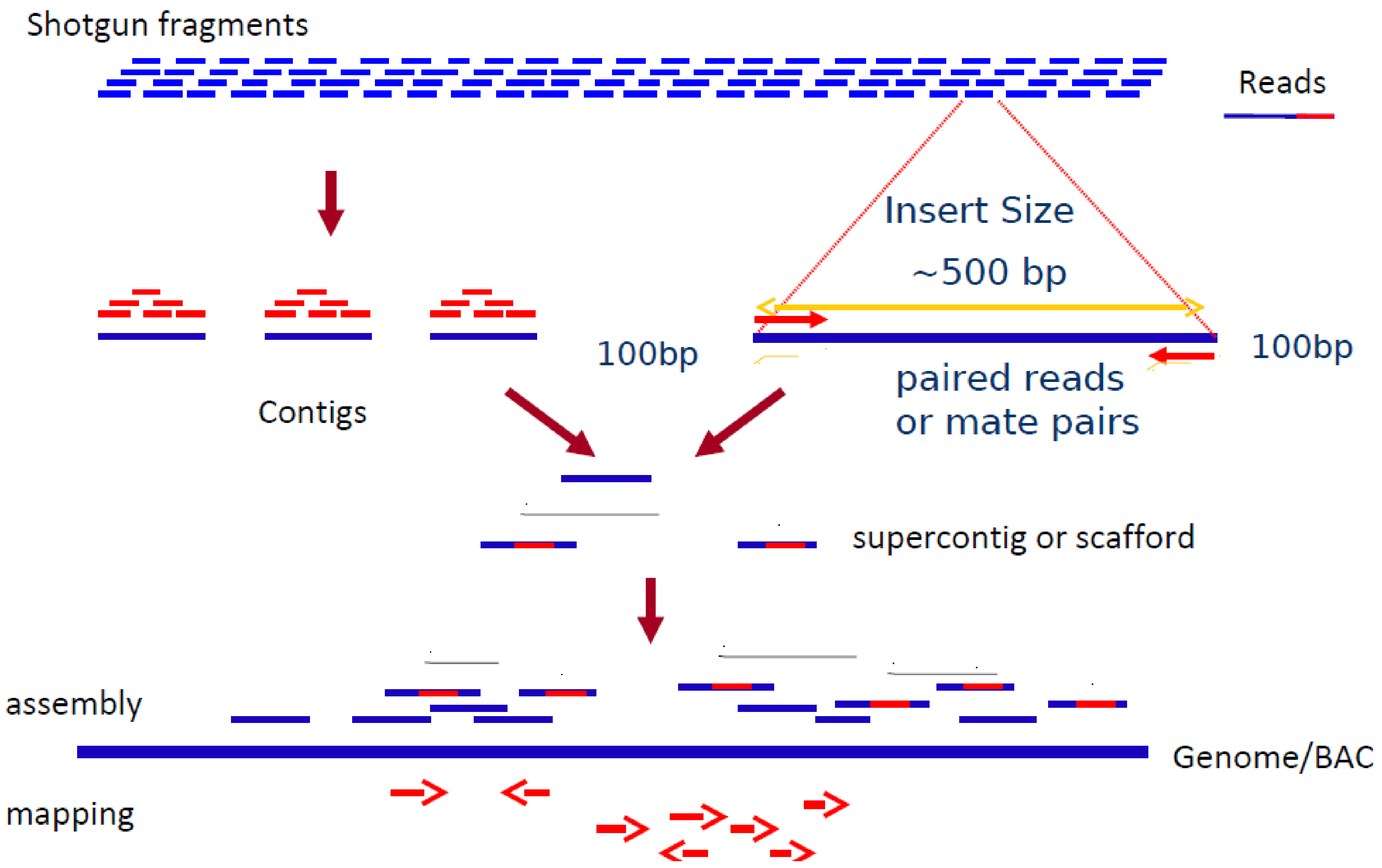
NGS Terminology – Single end or Paired end

Single end (SE): from each cDNA fragment only one is read.
Paired end (PE): the cDNA fragment is read from both ends



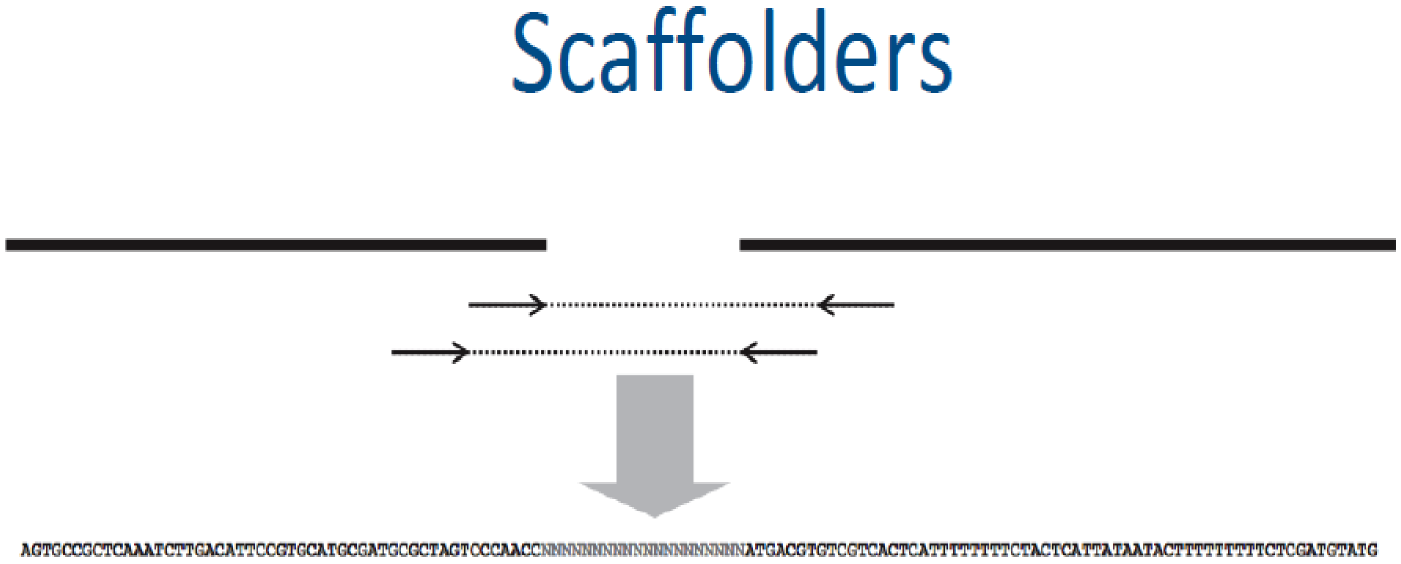
Paired end sequencing:
- Improves read alignment and therefore variant calling;
- Helps to detect structural variation;
- Can detect gene fusions and splice junctions;
- Useful for *de novo* assembly.

NGS Terminology


Shotgun fragments
Reads
Insert Size
~500 bp
100bp
100bp
paired reads
or mate pairs
Contigs
supercontig or scafford
assembly
Genome/BAC
mapping

NGS Terminology – Supercontigs or scaffolders

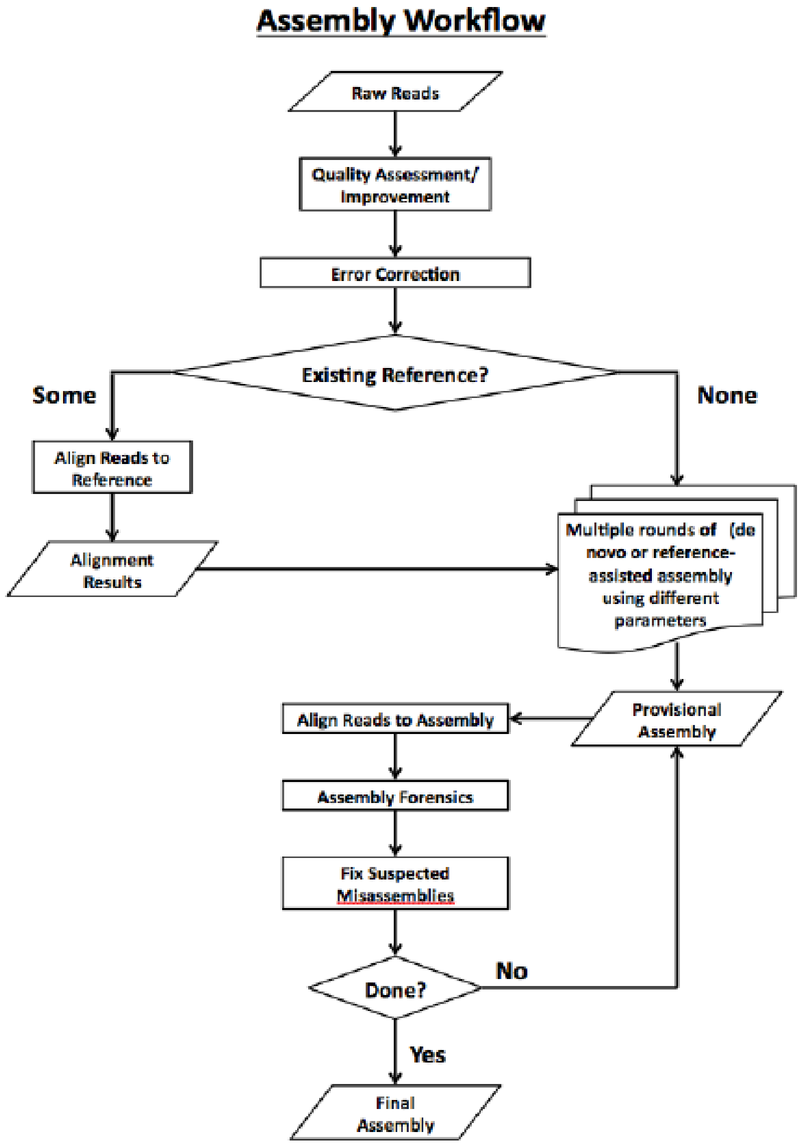
A scaffold is composed by contigs and gaps


Scaffolders

AGTGCCGCTCAAATCTTGACATTCCGTGCATGCGATGCGCTAGTCCCAACCNNNNNNNNNNNNNNNNNNNNNATGACGTGTCGTCACTCATTTTTTTTCTACTCATTATAATACTTTTTTTTTCTCGATGTATG

NGS Terminology – Assembly

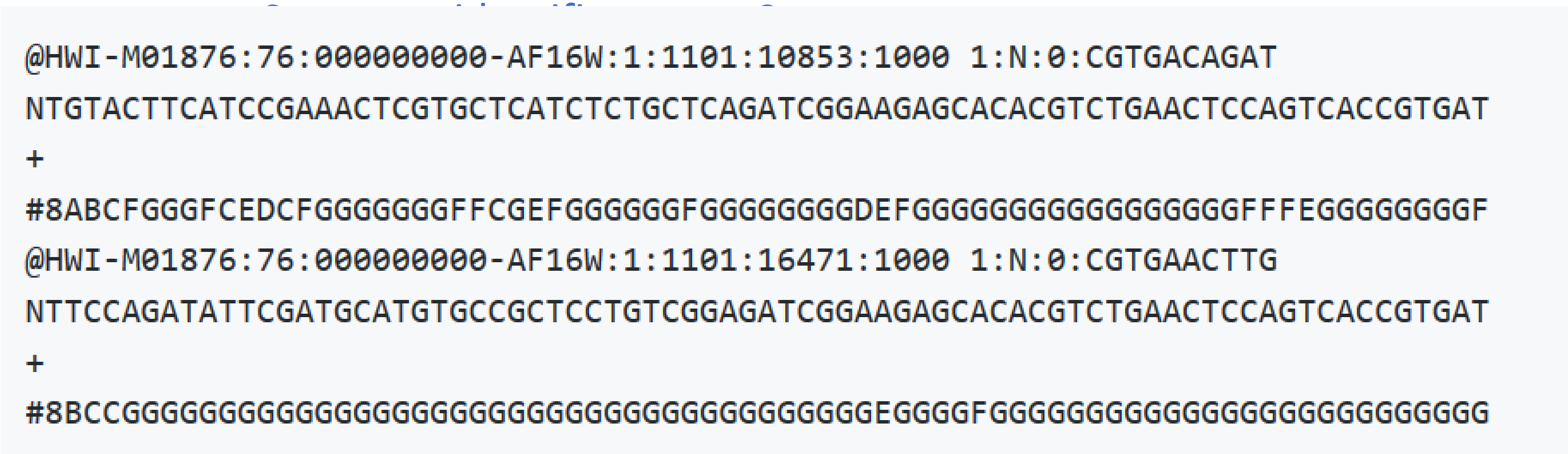
When we sequence the genome of a species that has not previously been characterized, *de novo* ("from new") assembly is required.


Assembly Workflow
Raw Reads
Quality Assessment/ Improvement
Error Correction
Existing Reference?
Some
None
Align Reads to Reference
Alignment Results
Multiple rounds of (de novo or reference-assisted assembly using different parameters
Provisional Assembly
Align Reads to Assembly
Assembly Forensics
Fix Suspected Misassemblies
Done?
No
Yes
Final Assembly

Interpret and Manipulate raw sequencing data
FastQ file

FastQ file

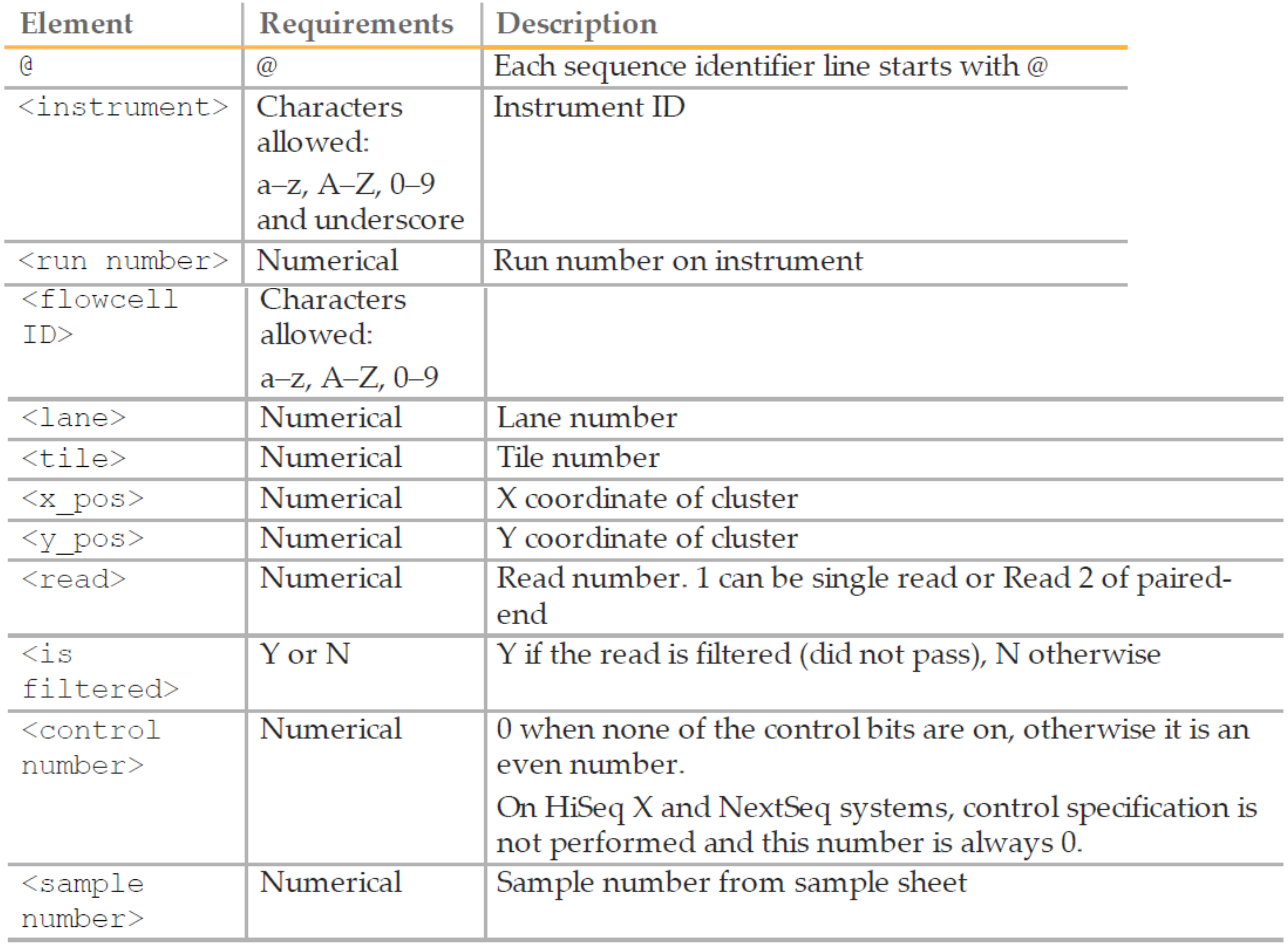
- Standard Format for NGS data.
- Extension of the Fasta format.
- Paired-end sequencing produces two fastq files per lane or multiplex.
- Unaligned read sequences with base qualities.
- Text-based formats (easy to use!)
- If not compressed, it can be huge.
- FASTQ files are saved compressed in zip format indicated by the .gz file extension.
- Each entry in a FASTQ file consists of four lines:



```
@HWI-M01876:76:000000000-AF16W:1:1101:10853:1000 1:N:0:CGTGACAGAT
NTGTACTTCATCCGAAACTCGTGCTCATCTCTGCTCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCGTGAT
+
#8ABCFGGGFCEDCFGGGGGGGFFCGEFGGGGGGFGGGGGGGGDEFGGGGGGGGGGGGGGGGGFFFEGGGGGGGGF
@HWI-M01876:76:000000000-AF16W:1:1101:16471:1000 1:N:0:CGTGAACTTG
NTTCCAGATATTCGATGCATGTGCCGCTCCTGTCGGAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCGTGAT
+
#8BCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGG
```

FastQ file - »Sequence identifier

```
@<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos> <read>:<is filtered>:<control number>:<sample number>
```



Element	Requirements	Description
@	@	Each sequence identifier line starts with @
<instrument>	Characters allowed: a–z, A–Z, 0–9 and underscore	Instrument ID
<run number>	Numerical	Run number on instrument
<flowcell ID>	Characters allowed: a–z, A–Z, 0–9	
<lane>	Numerical	Lane number
<tile>	Numerical	Tile number
<x_pos>	Numerical	X coordinate of cluster
<y_pos>	Numerical	Y coordinate of cluster
<read>	Numerical	Read number. 1 can be single read or Read 2 of paired-end
<is filtered>	Y or N	Y if the read is filtered (did not pass), N otherwise
<control number>	Numerical	0 when none of the control bits are on, otherwise it is an even number. On HiSeq X and NextSeq systems, control specification is not performed and this number is always 0.
<sample number>	Numerical	Sample number from sample sheet

FastQ file – Quality scores

Each base has a quality character associated with it, representing how confidently the machine identified (called) the base. The probability of error per base is given as a **Phred score** (https://en.wikipedia.org/wiki/Phred_quality_score), calculated from an integer value (Q) derived from the quality character associated to the base. The probability of error is given by the Phred score using P(Q)=10^(-Q/10). Useful reference values of Q include:

* Q=10 - 90% accuracy (0.1 error)

* Q=20 - 99% accuracy (0.01 error)

* Q=30 - 99.9% accuracy (0.001 error)

* Q=40 - 99.99% accuracy (0.0001 error)

Although there's theoretically no limit, Q usually goes up to around 40 in recent illumina machines.

FastQ file – Quality scores

To obtain this Q value from the character associated to the quality of the base, we have to know that each character (such as '#') has an ASCII (https://en.wikipedia.org/wiki/ASCII) decimal value associated (for example, '#' has a value of 35). The Q value of a character is the decimal value corresponding to the entry of that character in the ASCII table, subtracted by 33. For example Q('#') = 35 – 33.

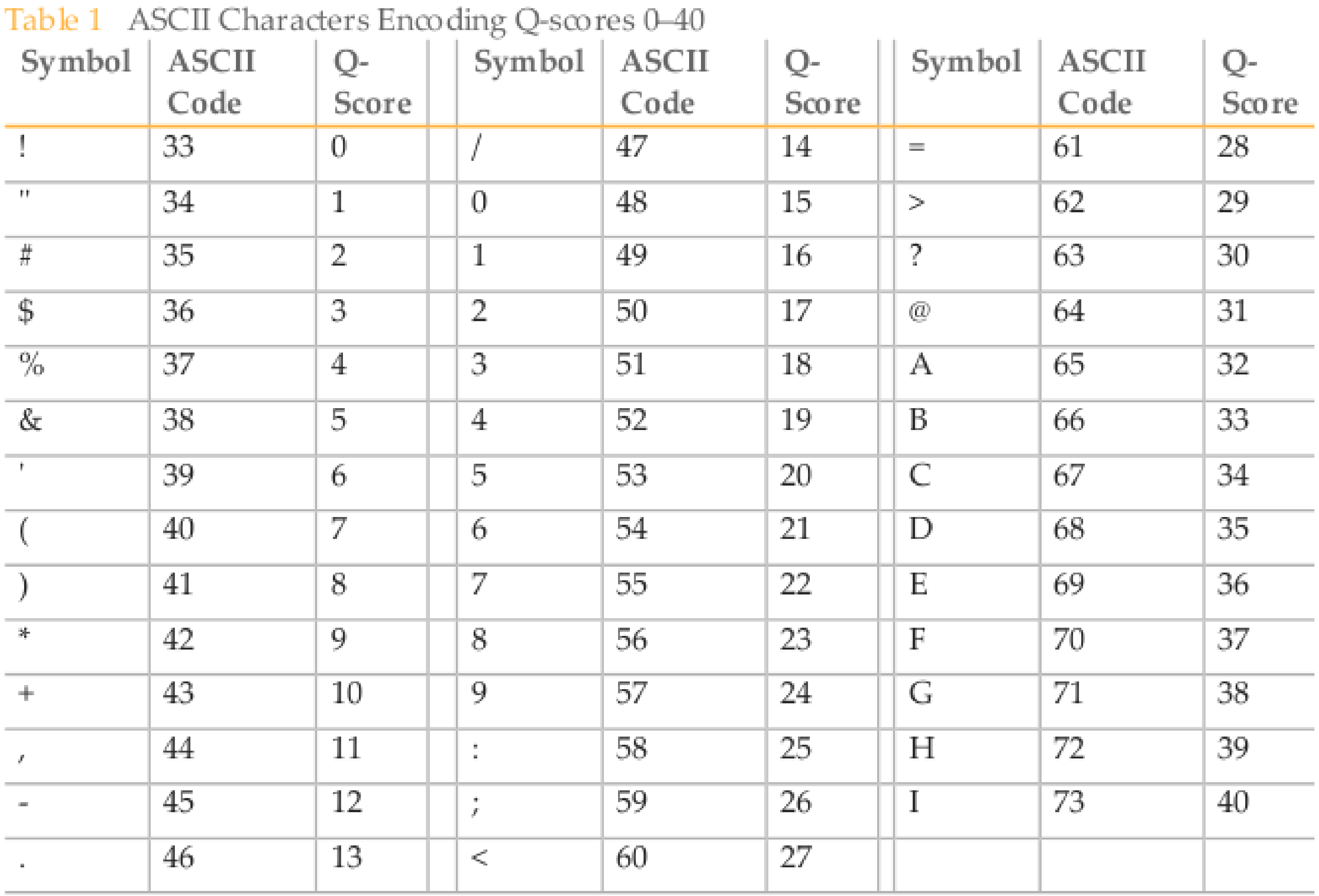


Table 1 ASCII Characters Encoding Q-scores 0–40

Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score
!	33	0	/	47	14	=	61	28
"	34	1	0	48	15	>	62	29
#	35	2	1	49	16	?	63	30
$	36	3	2	50	17	@	64	31
%	37	4	3	51	18	A	65	32
&	38	5	4	52	19	B	66	33
'	39	6	5	53	20	C	67	34
(	40	7	6	54	21	D	68	35
)	41	8	7	55	22	E	69	36
*	42	9	8	56	23	F	70	37
+	43	10	9	57	24	G	71	38
,	44	11	:	58	25	H	72	39
-	45	12	;	59	26	I	73	40
.	46	13	<	60	27			

Exercises: FastQ file – Quality scores

Go to

https://github.com/CarinaSilva/Curso-Int-NGS/tree/master/NGS

Exercises 1-5:

Quality Control
FastQC software

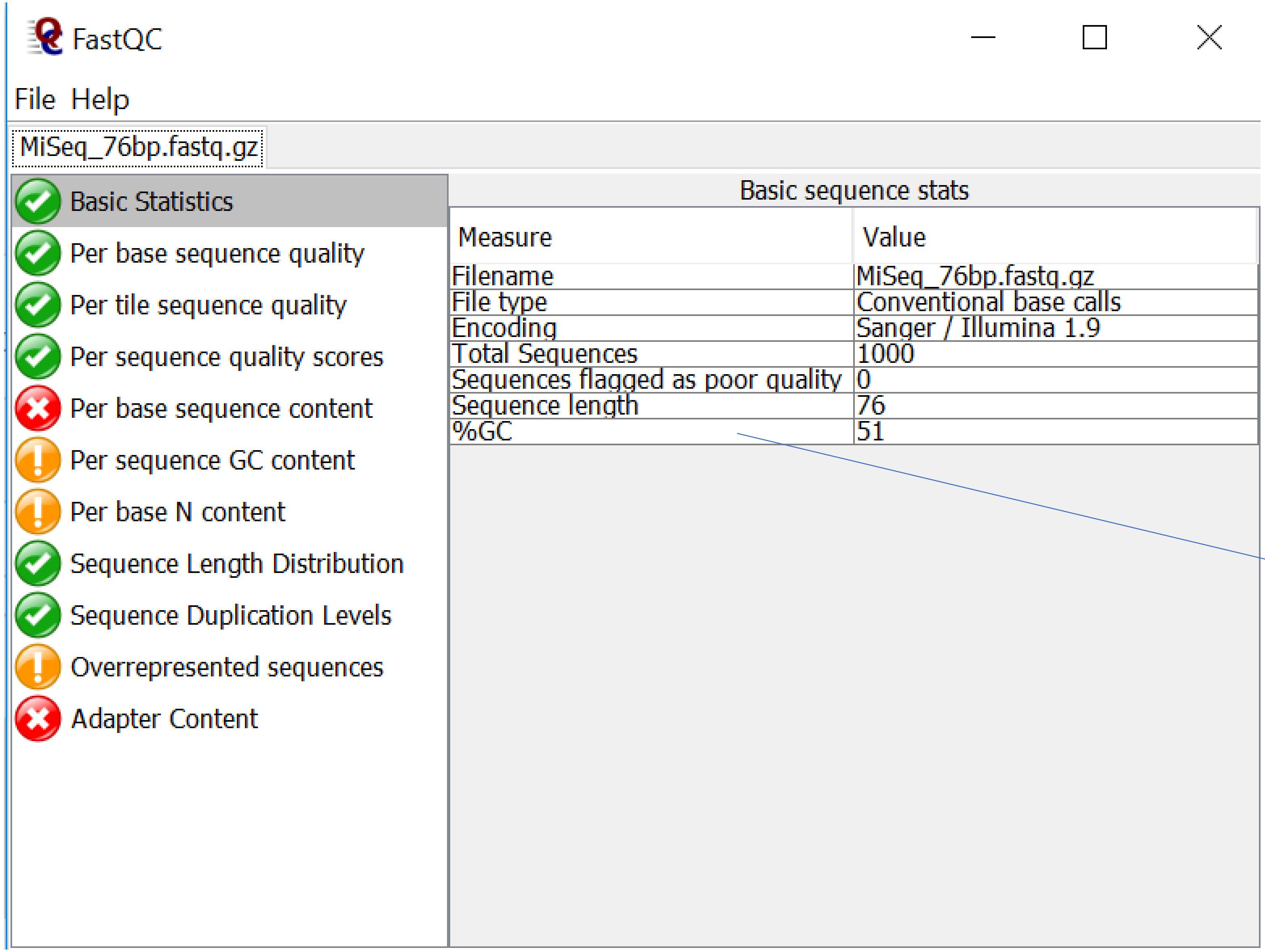
Exercises: FastQC software

Go to

https://github.com/CarinaSilva/Curso-Int-NGS/tree/master/NGS

Exercise 6

FastQC – Basic Statistics

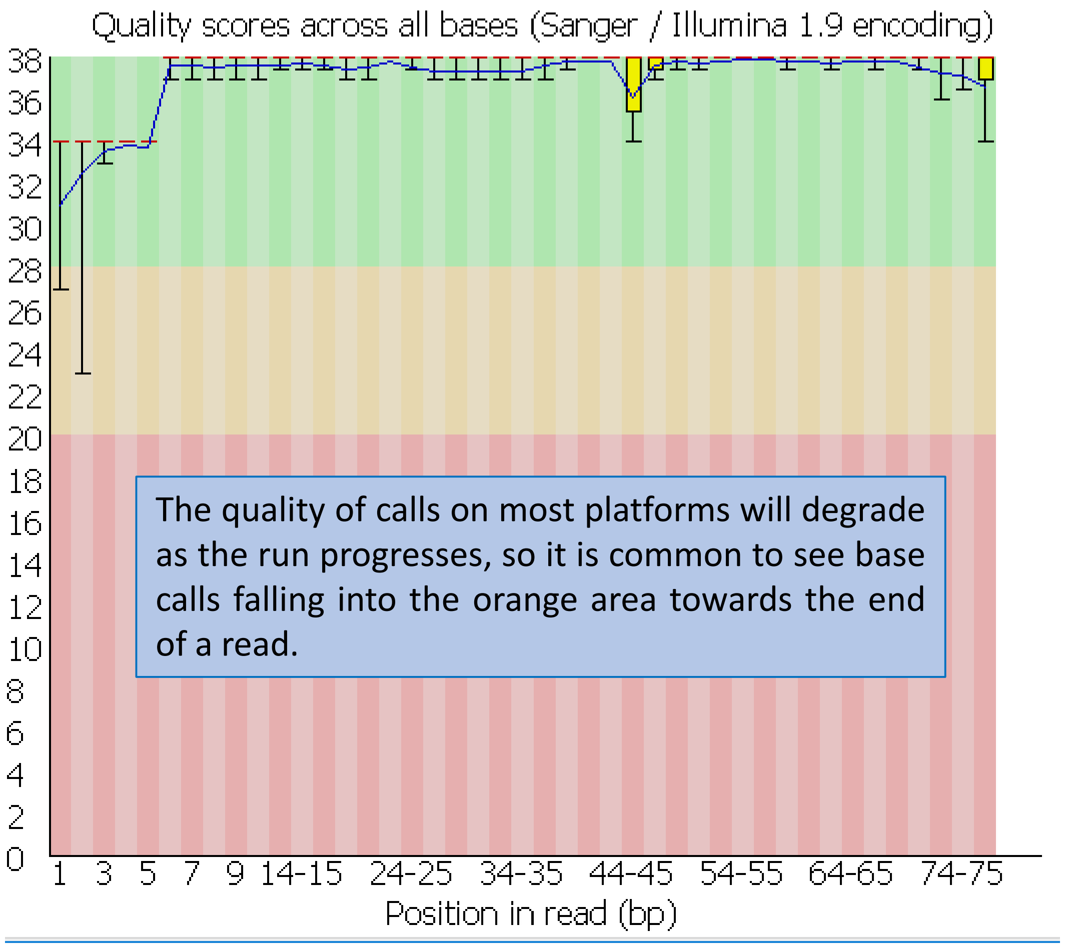

FastQC
File Help
MiSeq_76bp.fastq.gz
✓ Basic Statistics
✓ Per base sequence quality
✓ Per tile sequence quality
✓ Per sequence quality scores
✗ Per base sequence content
! Per sequence GC content
! Per base N content
✓ Sequence Length Distribution
✓ Sequence Duplication Levels
! Overrepresented sequences
✗ Adapter Content

Basic sequence stats

Measure	Value
Filename	MiSeq_76bp.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1000
Sequences flagged as poor quality	0
Sequence length	76
%GC	51

%GC: The overall %GC of all bases in all sequences

FastQC – Per base sequence quality



The quality of calls on most platforms will degrade as the run progresses, so it is common to see base calls falling into the orange area towards the end of a read.

-For each position a BoxWhisker type plot is drawn .

- The upper and lower whiskers represent the 10% and 90% percentiles.

-The blue line is the mean.

- The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red).

- The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red).

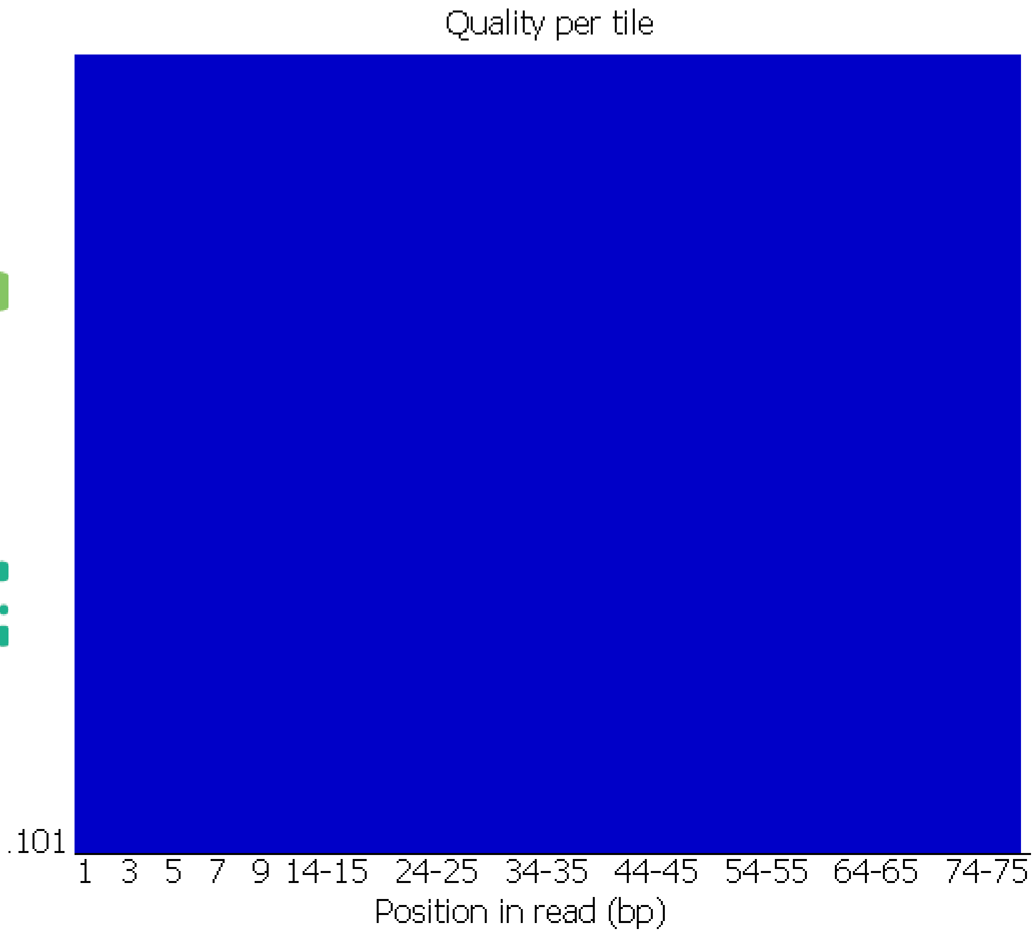
Warning

A warning will be issued if the lower quartile for any base is less than 10, or if the median for any base is less than 25.

Failure

This module will raise a failure if the lower quartile for any base is less than 5 or if the median for any base is less than 20.

FastQC – Per tile sequence quality



This graph will only appear in your analysis results if you're using an Illumina library which retains its original sequence identifiers.

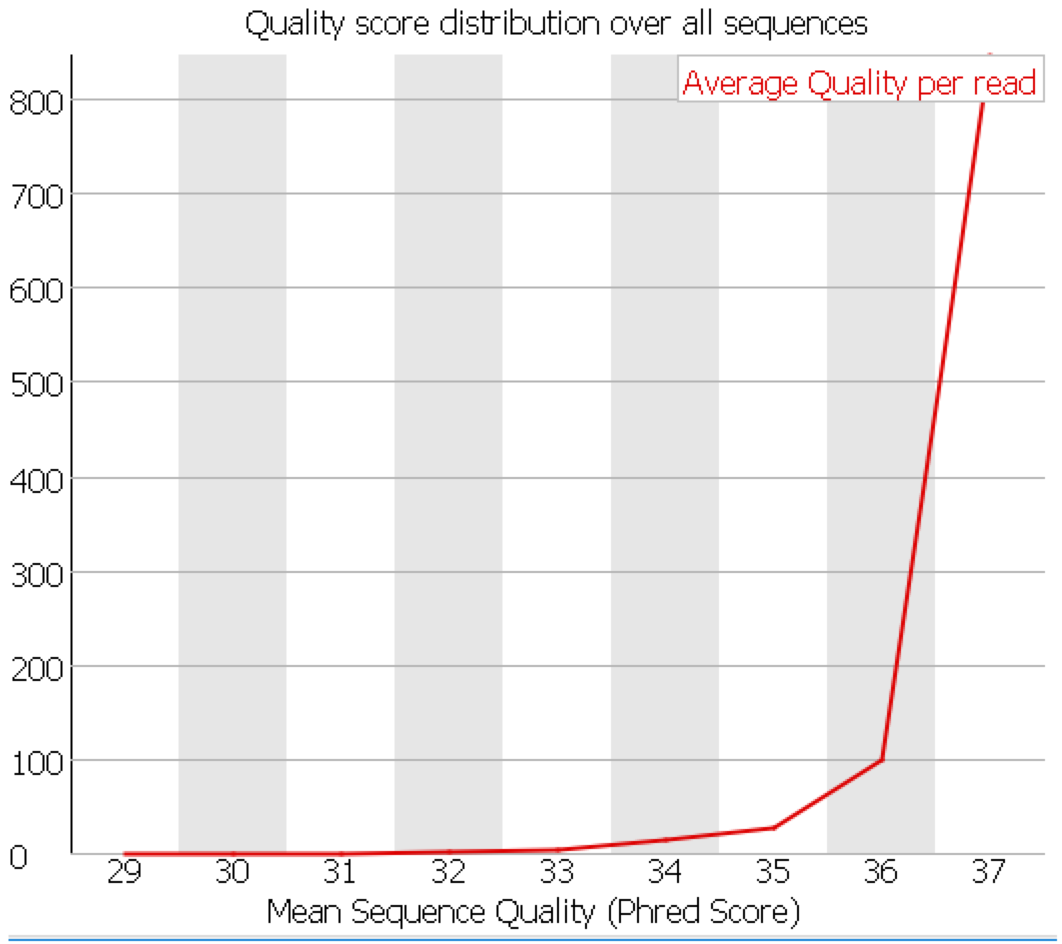
The graph allows you to look at the quality scores from each tile across all of your bases to see if there was a loss in quality associated with only one part of the flowcell.

The colors are on a cold to hot scale, with cold colors being positions where the quality was at or above the average for that base in the run.
A good plot should be blue all over.

Warning
This module will issue a warning if any tile shows a mean Phred score more than 2 less than the mean for that base across all tiles.

Failure
This module will issue a warning if any tile shows a mean Phred score more than 5 less than the mean for that base across all tiles.

FastQC – Per sequence quality scores



This plot allows you to see if a subset of your sequences have universally low quality values. It is often the case that a subset of sequences will have universally poor quality, often because they are poorly imaged (on the edge of the field of view etc), however these should represent only a small percentage of the total sequences.

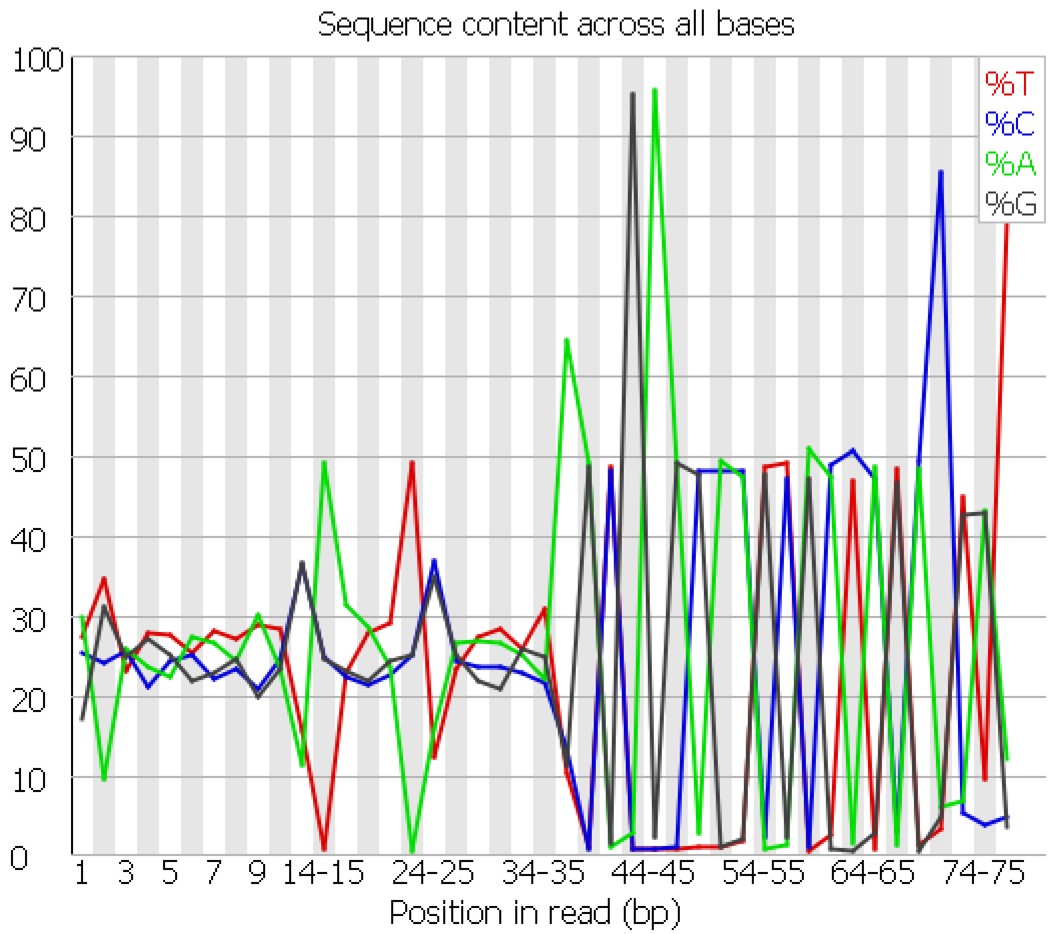
Warning

A warning is raised if the most frequently observed mean quality is below 27 - this equates to a 0.2% error rate.

Failure

An error is raised if the most frequently observed mean quality is below 20 - this equates to a 1% error rate.

FastQC – Per base sequence content



Per Base Sequence Content plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called.

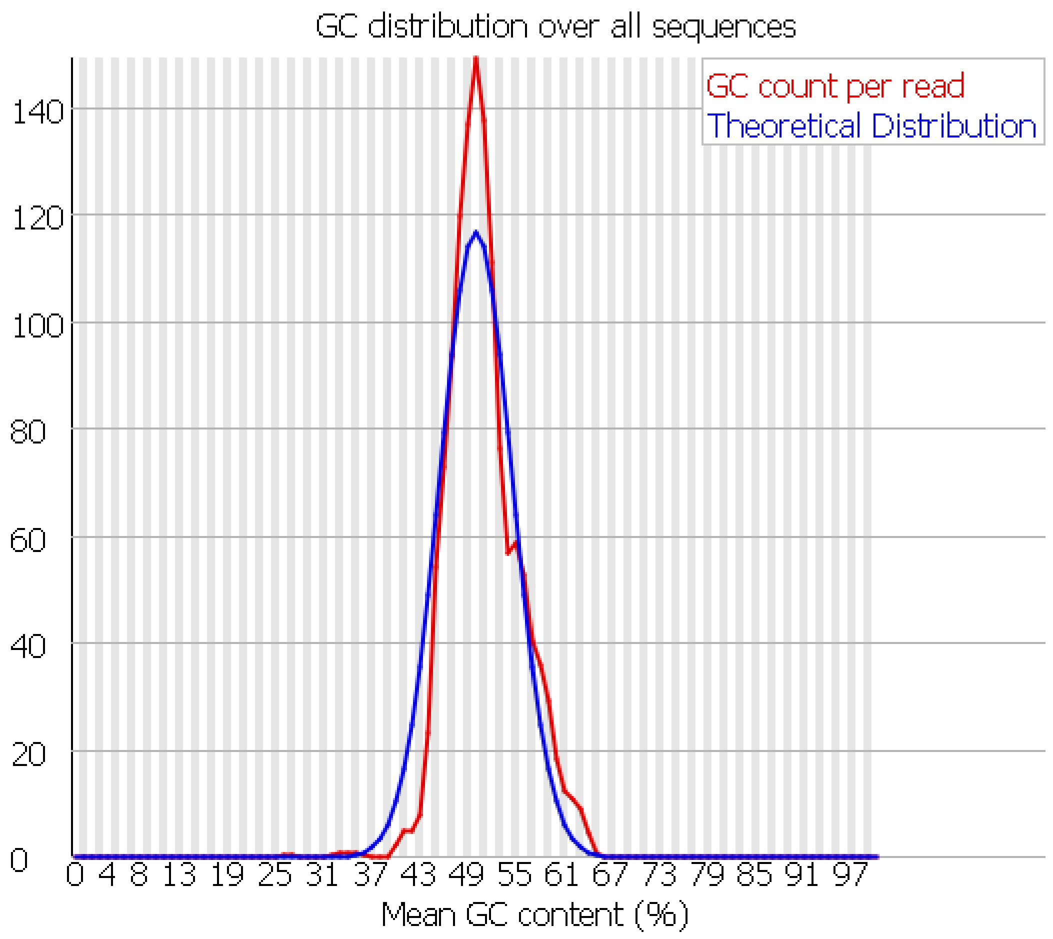
Warning

This module issues a warning if the difference between A and T, or G and C is greater than 10% in any position.

Failure

This module will fail if the difference between A and T, or G and C is greater than 20% in any position.

FastQC – Per sequence GC Content



This module measures the GC content across the whole length of each sequence in a file and compares it to a modelled normal distribution of GC content.

Warning

A warning is raised if the sum of the deviations from the normal distribution represents more than 15% of the reads.

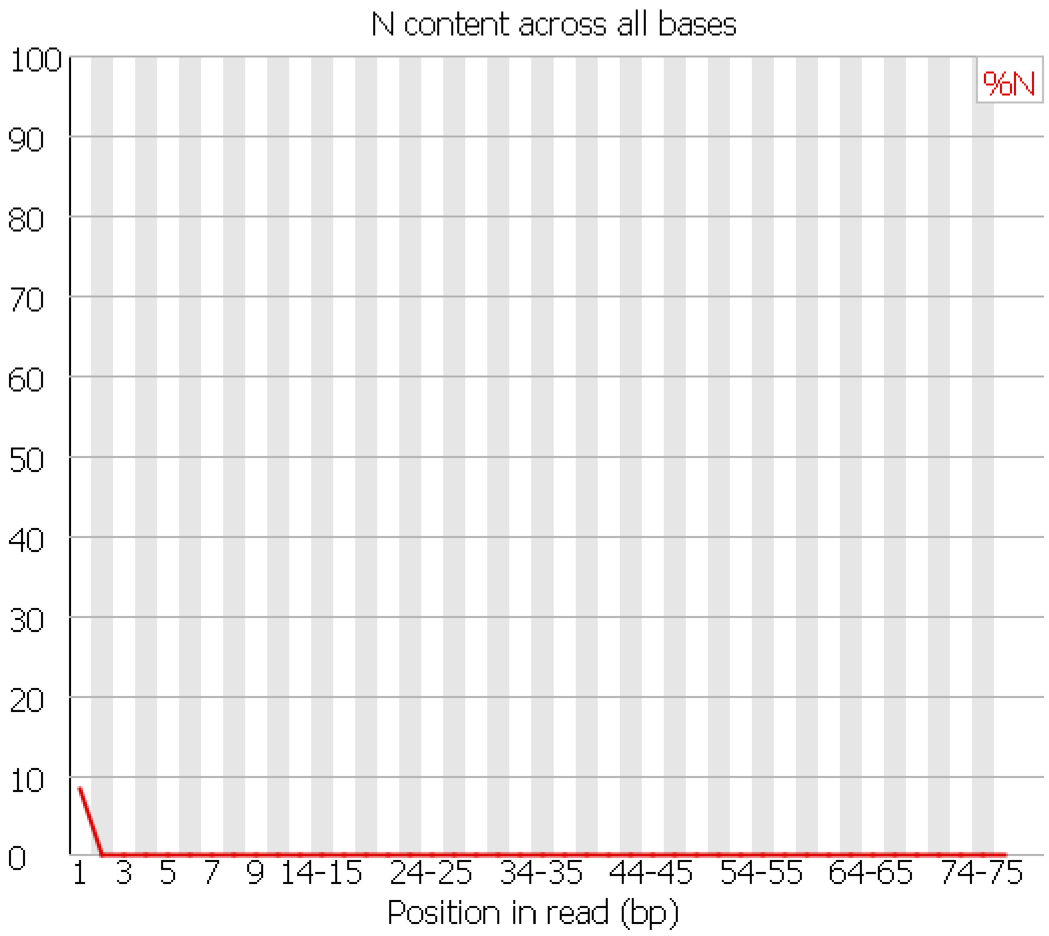
Failure

This module will indicate a failure if the sum of the deviations from the normal distribution represents more than 30% of the reads.

Common reasons for warnings

Warnings in this module usually indicate a problem with the library. Sharp peaks on an otherwise smooth distribution are normally the result of a specific contaminant (adapter dimers for example), which may well be picked up by the overrepresented sequences module. Broader peaks may represent contamination with a different species.

FastQC - Per base N content



If a sequencer is unable to make a base call with sufficient confidence then it will normally substitute an N rather than a conventional base call

This module plots out the percentage of base calls at each position for which an N was called.

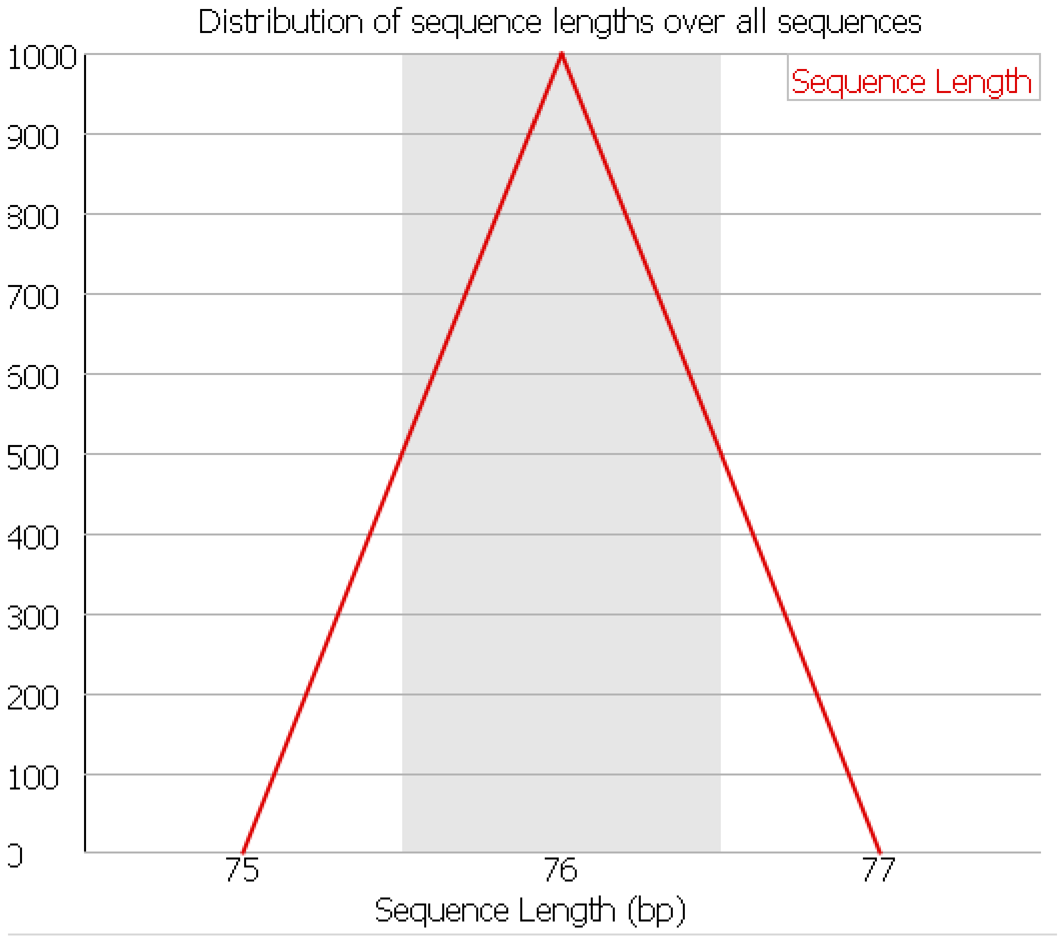
Warning

This module raises a warning if any position shows an N content of >5%.

Failure

This module will raise an error if any position shows an N content of >20%.

FastQC – Sequence length distribution



This module generates a graph showing the distribution of fragment sizes in the file which was analysed.

Warning

This module will raise a warning if all sequences are not the same length.

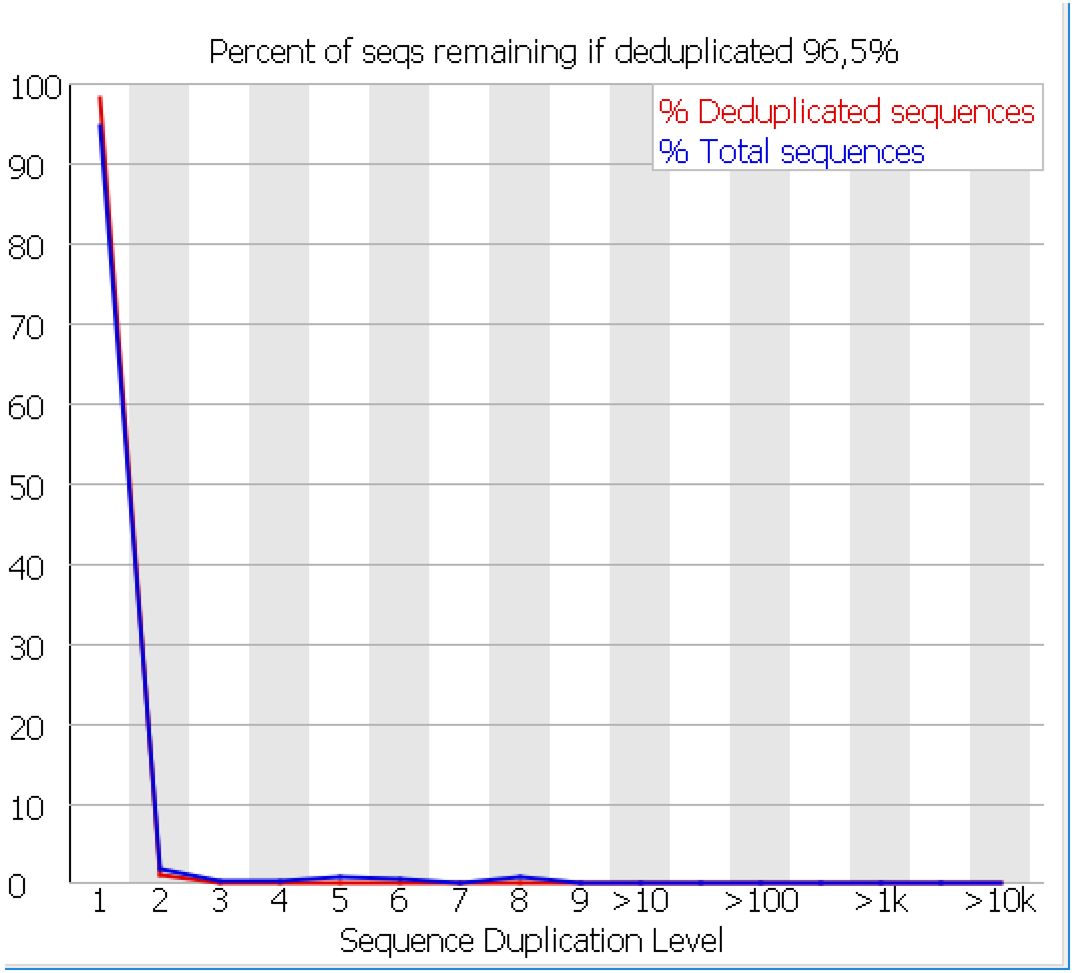
Failure

This module will raise an error if any of the sequences have zero length.

Common reasons for warnings

For some sequencing platforms it is entirely normal to have different read lengths so warnings here can be ignored.

FastQC – Sequence Duplication Levels



Note: A good site explanation: http://proteo.me.uk/2011/05/interpreting-the-duplicate-sequence-plot-in-fastqc/.

This module counts the degree of duplication for every sequence in a library and creates a plot showing the relative number of sequences with different degrees of duplication.

On a diverse library most sequences will occur only once in the final set. A low level of duplication may indicate a very high level of coverage of the target sequence, but a high level of duplication is more likely to indicate some kind of enrichment bias (eg PCR over amplification).

Warning

This module will issue a warning if non-unique sequences make up more than 20% of the total.

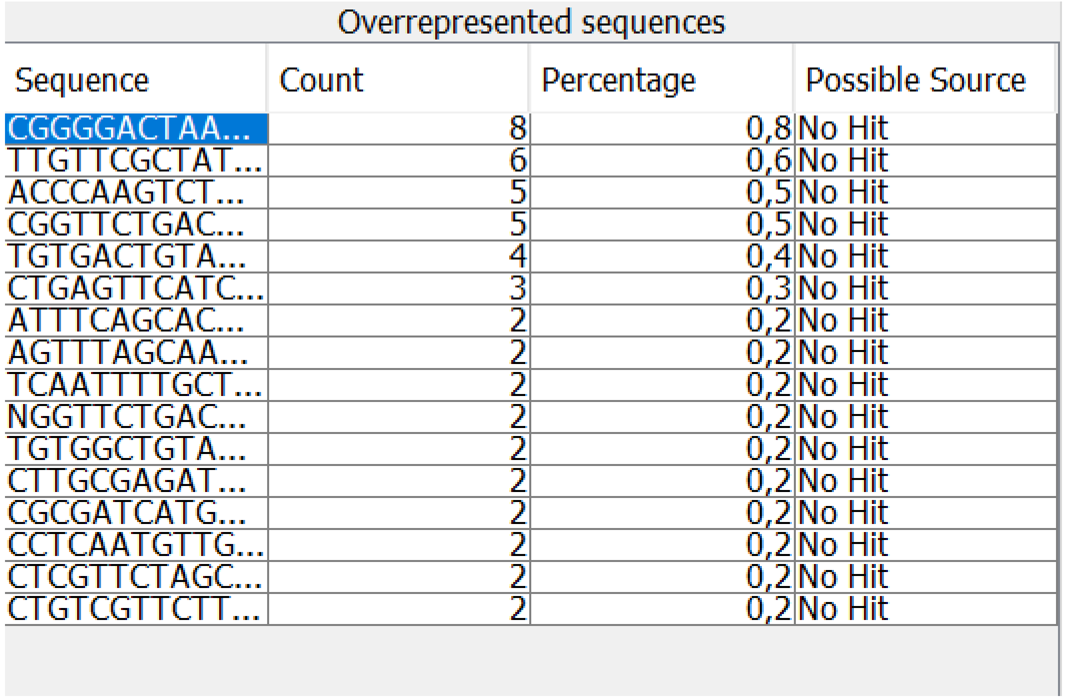
Failure

This module will issue a error if non-unique sequences make up more than 50% of the total.

Common reasons for warnings

The underlying assumption of this module is of a diverse unenriched library. Any deviation from this assumption will naturally generate duplicates and can lead to warnings or errors from this module.

FastQC – Overrepresented Sequences



Overrepresented sequences

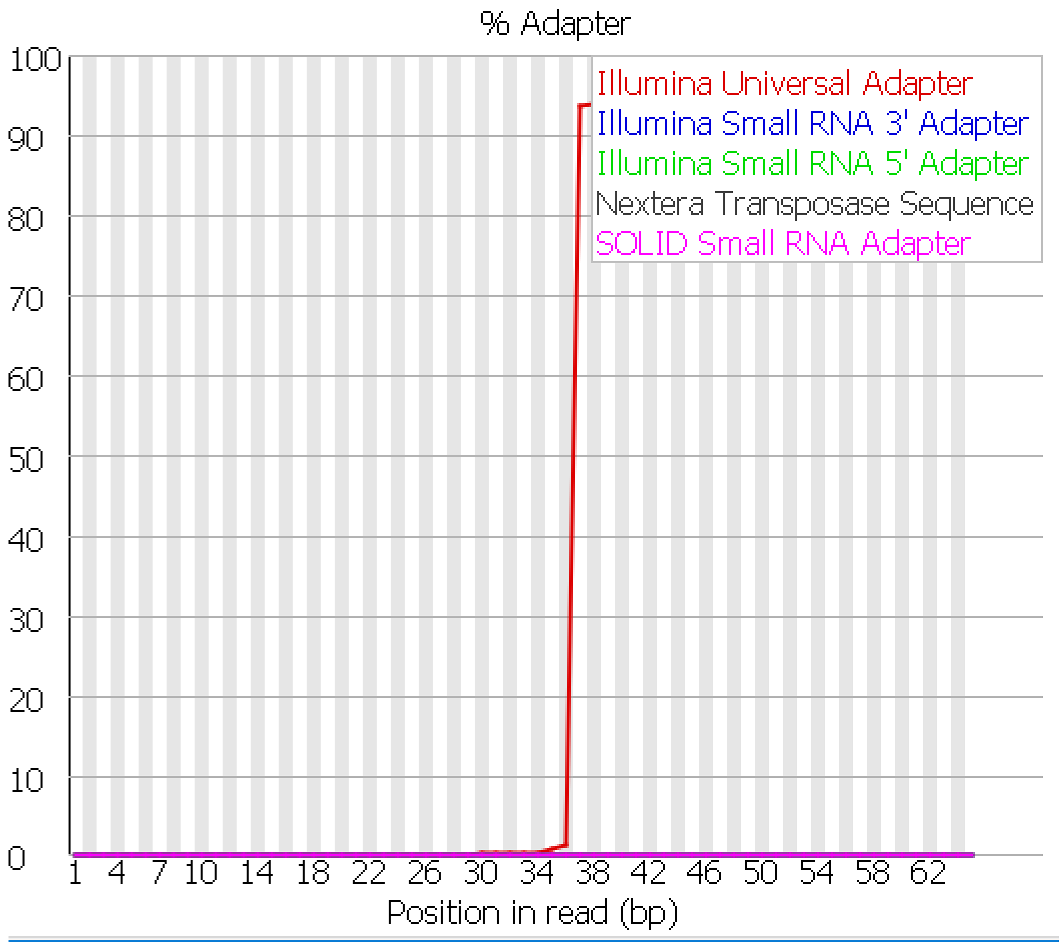
Sequence	Count	Percentage	Possible Source
CGGGGACTAA...	8	0,8	No Hit
TTGTTCGCTAT...	6	0,6	No Hit
ACCCAAGTCT...	5	0,5	No Hit
CGGTTCTGAC...	5	0,5	No Hit
TGTGACTGTA...	4	0,4	No Hit
CTGAGTTCATC...	3	0,3	No Hit
ATTTCAGCAC...	2	0,2	No Hit
AGTTTAGCAA...	2	0,2	No Hit
TCAATTTTGCT...	2	0,2	No Hit
NGGTTCTGAC...	2	0,2	No Hit
TGTGGCTGTA...	2	0,2	No Hit
CTTGCGAGAT...	2	0,2	No Hit
CGCGATCATG...	2	0,2	No Hit
CCTCAATGTTG...	2	0,2	No Hit
CTCGTTCTAGC...	2	0,2	No Hit
CTGTCGTTCTT...	2	0,2	No Hit

A normal high-throughput library will contain a diverse set of sequences, with no individual sequence making up a tiny fraction of the whole. Finding that a single sequence is very overrepresented in the set either means that it is highly biologically significant, or indicates that the library is contaminated, or not as diverse as you expected.

Because the duplication detection requires an exact sequence match over the whole length of the sequence any reads over 75bp in length are truncated to 50bp for the purposes of this analysis. Even so, longer reads are more likely to contain sequencing errors which will artificially increase the observed diversity and will tend to underrepresent highly duplicated sequences.

You can copy this sequences and save them to a text file. You can try BLAST

FastQC – Adapter Content



The plot itself shows a cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position. Once a sequence has been seen in a read it is counted as being present right through to the end of the read so the percentages you see will only increase as the read length goes on.

Warning
This module will issue a warning if any sequence is present in more than 5% of all reads.

Failure
This module will issue a warning if any sequence is present in more than 10% of all reads.

Common reasons for warnings
Any library where a reasonable proportion of the insert sizes are shorter than the read length will trigger this module. This doesn't indicate a problem as such - just that the sequences will need to be adapter trimmed before proceeding with any downstream analysis.

Exercises: FastQC software

Go to

https://github.com/CarinaSilva/Curso-Int-NGS/tree/master/NGS

Exercises 7-8

Improve quality of data in fastq files

Filtering and Trimming

As you may have noticed before, reads tend to lose quality towards their end, where there is a higher probability of erroneous bases being called.

To avoid problems in subsequent analysis, you should remove bases with higher probability of error, usually by trimming poor quality bases from the end.

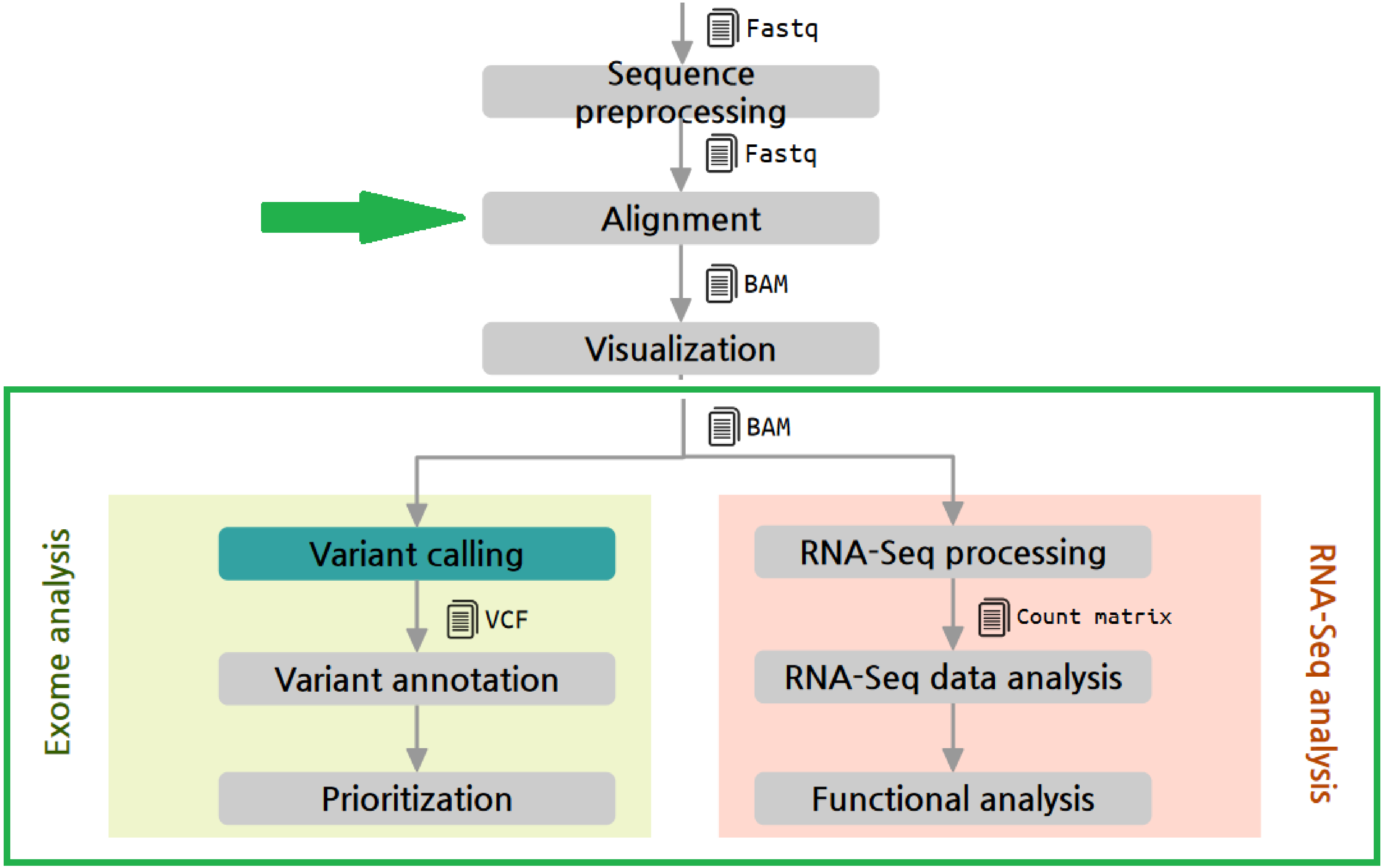
Exercises

Go to

https://github.com/CarinaSilva/Curso-Int-NGS/tree/master/NGS

Exercises 9-11

Where we are?



Fastq
Sequence preprocessing
Fastq
Alignment
BAM
Visualization
BAM
Exome analysis
Variant calling
VCF
Variant annotation
Prioritization
RNA-Seq processing
Count matrix
RNA-Seq data analysis
Functional analysis
RNA-Seq analysis

Alignment

Sequence alignment in NGS is
● *Process of determining the most likely source within the reference genome sequence that the observed DNA sequencing read is derived from.*

Principles and approaches to sequence alignment have not changed much since 80’s.

Basic Local Alignment Search Tool (BLAST)

NGS: Nucleotide based alignment

Alignment

Our goal is to align sequences to a reference genome.

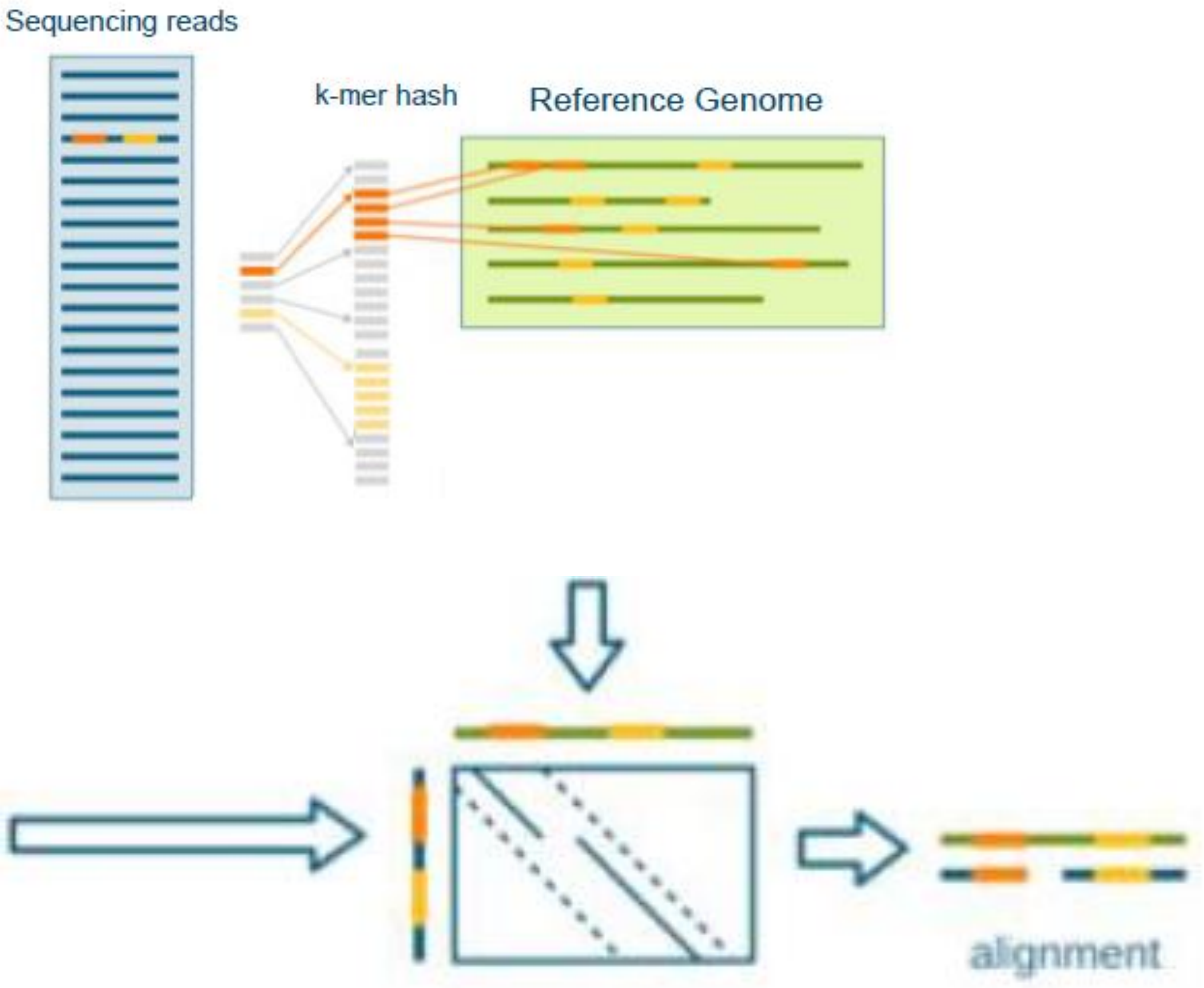
It is possible to obtain this references from sources such as Ensembl, NCBI and UCSC.

There are many popular aligners, including BWA, Bowtie2, SOAP, MAQ and Novoalign, which vary in speed and accuracy.

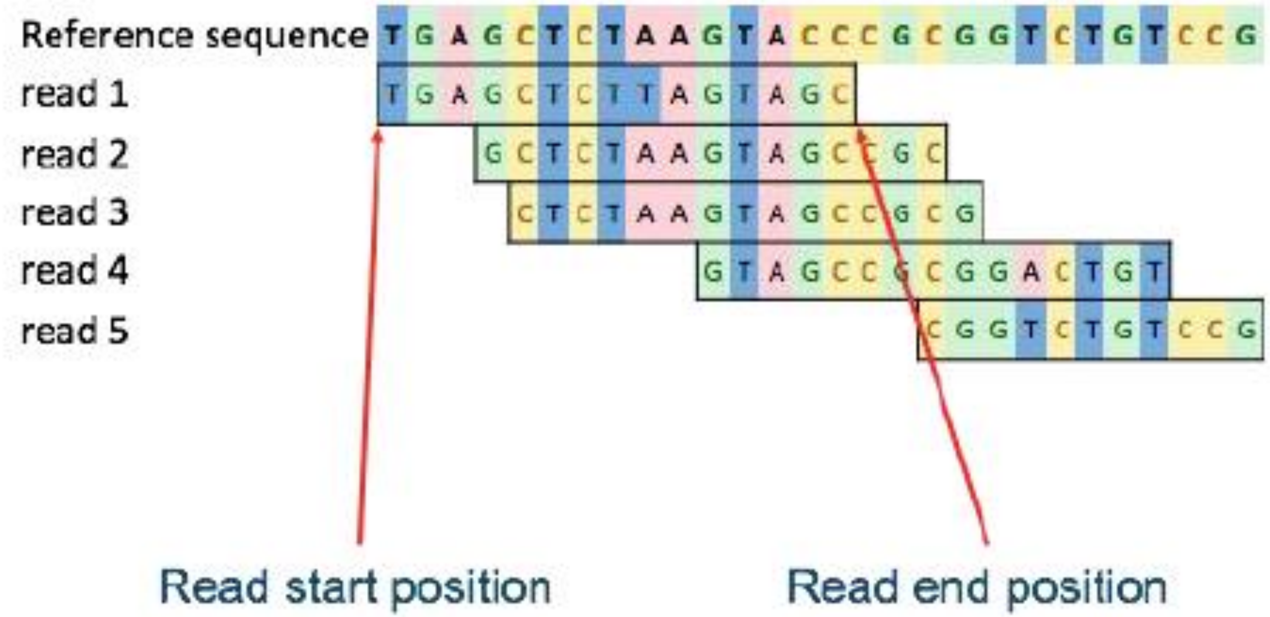
Firstly we have to index the reference genome. For a small set of sequences this may take a few seconds, for a entire genome the indexing may require hours.

The aligners will produce a SAM (sequence alignment/map) file.

Alignment


Sequencing reads
k-mer hash
Reference Genome
alignment

Alignment


Reference sequence T G A G C T C T A A G T A C C C G C G G T C T G T C C G
read 1 T G A G C T C T T A G T A G C
read 2 G C T C T A A G T A G C C G C
read 3 C T C T A A G T A G C C G C G
read 4 G T A G C C G C G G A C T G T
read 5 C G G T C T G T C C G
Read start position
Read end position

SAM/BAM format

SAM (Sequence Alignment/Map) format

- Single unified format for storing read alignments to a reference genome
- Developed by the 1000 Genomes group in 2009.

BAM (Binary Alignment/Map) format

- Binary equivalent of SAM.
- Developed for fast processing/indexing.

Key features

- Can store alignments from most aligners
- Supports multiple sequencing technologies
- Supports indexing for quick retrieval/viewing
- Compact size (e.g. 112Gbp Illumina = 116Gbytes disk space)
- Reads can be grouped into logical groups e.g. lanes, libraries, samples
- Widely supported by variant calling software packages

SAM/BAM tools

Several tools and programming APIs for interacting with SAM/BAM files

Samtools - Sanger (http://samtools.sourceforge.net)

- Convert SAM <-> BAM <-> CRAM
- Sort, index, BAM files
- Flagstat - summary of the mapping flags
- Merge multiple BAM files
- Rmdup - remove PCR duplicates from the library preparation

Picard tools - Broad Institute (https://www.broadinstitute.org/gatk/)

- MarkDuplicates, CollectAlignmentSummaryMetrics, CreateSequenceDictionary, SamToFastq, MeanQualityByCycle, FixMateInformation etc.

Others

- Bio-SamTool - Perl (http://search.cpan.org/~lds/Bio-SamTools/)
- Pysam - Python (https://github.com/pysam-developers/pysam)
- R - Bioconductor/Rsamtools

BAM Visualisation

- BamView, LookSeq, Gap5, Tablet, Ensembl, UCSC, Bambino, Biodalliance...
- **IGV: http://www.broadinstitute.org/igv/**

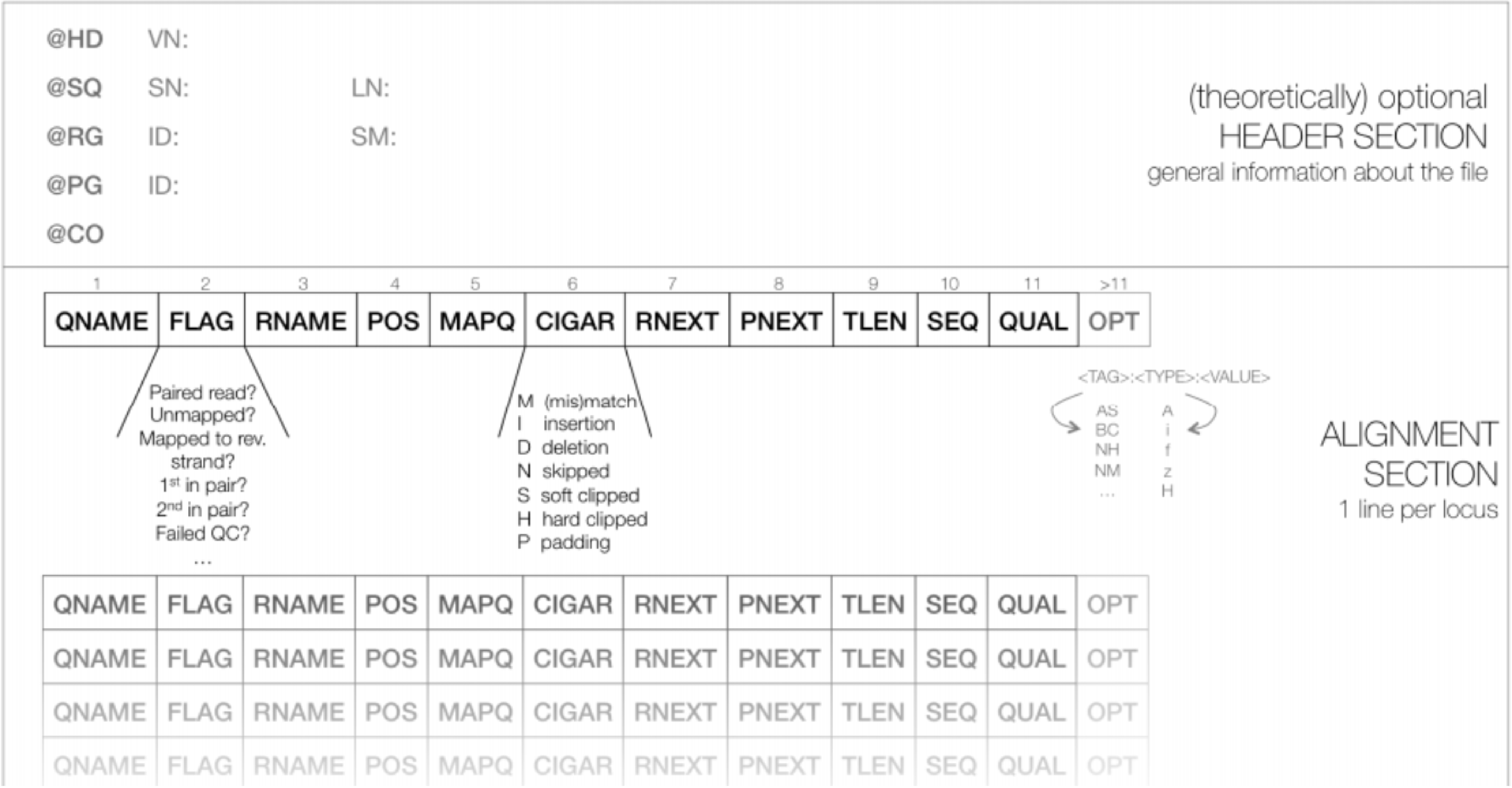
Exercises

Go to

https://github.com/CarinaSilva/Curso-Int-NGS/tree/master/NGS

Exercises 12-15

SAM file



@HD VN:
@SQ SN: LN:
@RG ID: SM:
@PG ID:
@CO

(theoretically) optional
HEADER SECTION
general information about the file

1	2	3	4	5	6	7	8	9	10	11	>11
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT

Paired read?
Unmapped?
Mapped to rev. strand?
1st in pair?
2nd in pair?
Failed QC?
...

M (mis)match
I insertion
D deletion
N skipped
S soft clipped
H hard clipped
P padding

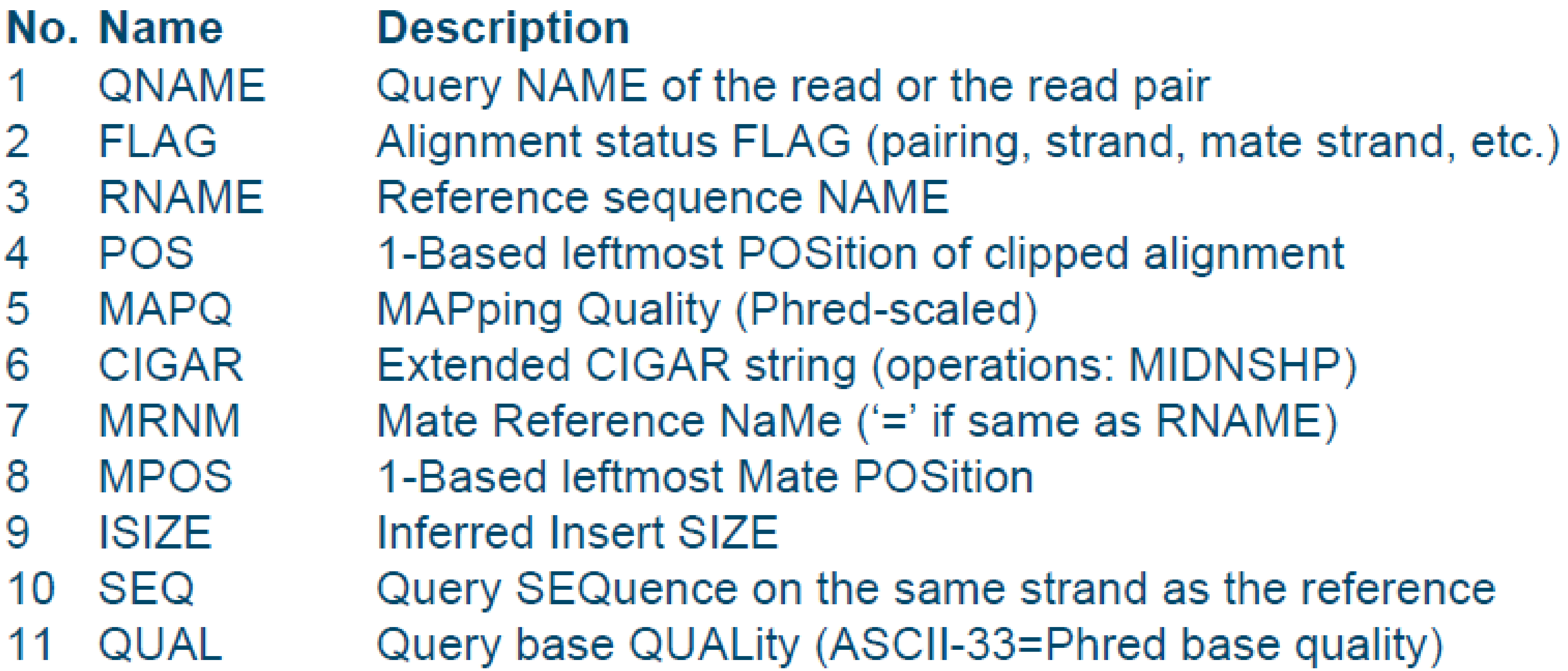
<TAG>:<TYPE>:<VALUE>

AS	A
BC	i
NH	f
NM	z
...	H

ALIGNMENT
SECTION
1 line per locus

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	OPT

SAM file



No.	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Alignment status FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe (‘=’ if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query base QUALity (ASCII-33=Phred base quality)

CIGAR format

Cigar has been traditionally used as a compact way to represent a sequence alignment

Operations include

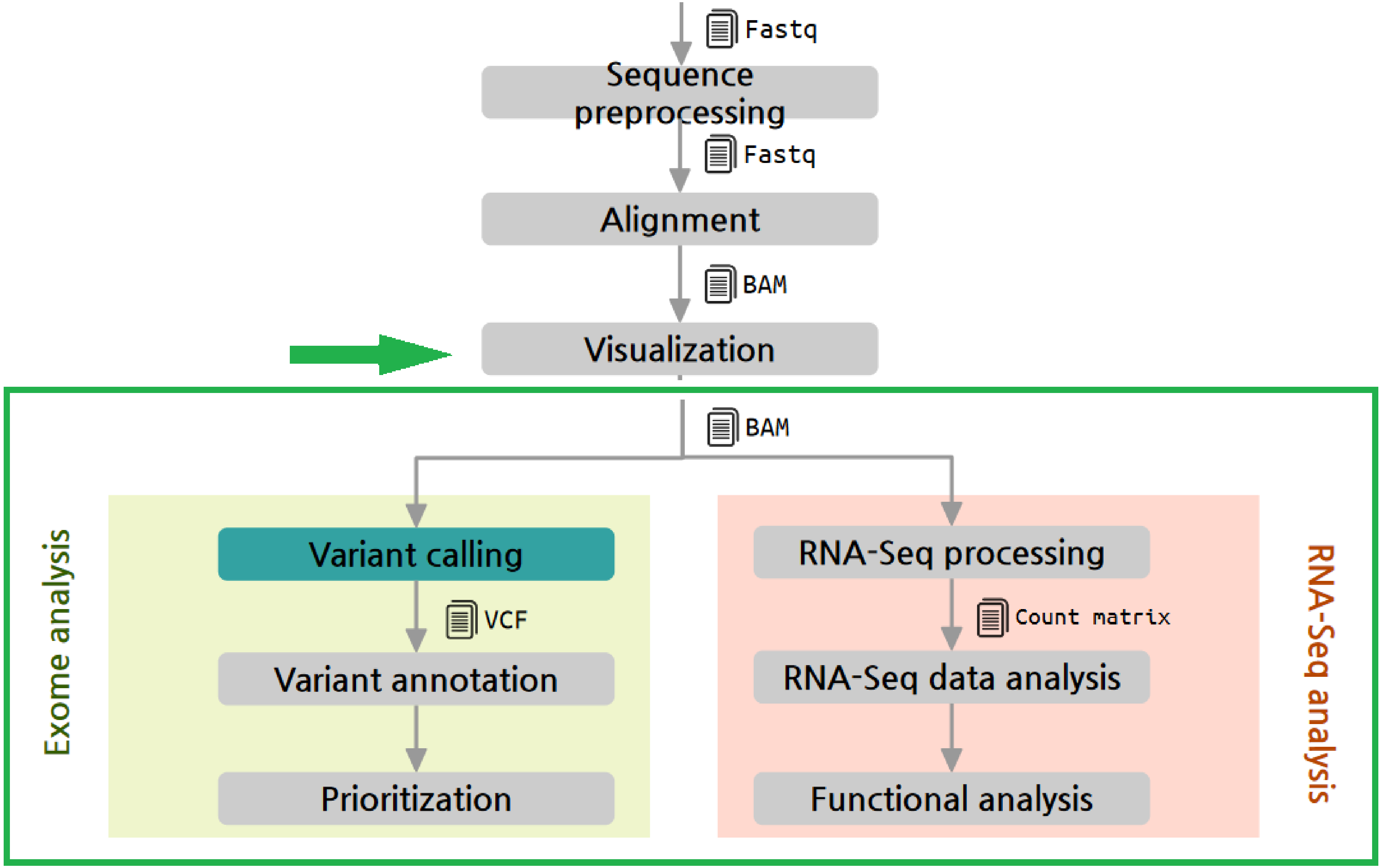
- M - match or mismatch
- I - insertion
- D - deletion

SAM extends these to include

- S - soft clip (ignore these bases)
- H - hard clip (ignore and remove these bases)

E.g.Read: `ACGCA-TGCAGTtagacgt`
Ref: `ACTCAGTG--GT`
Cigar: 5M1D2M2I2M7S

Where we are?


Fastq
Sequence preprocessing
Fastq
Alignment
BAM
Visualization
BAM
Exome analysis
Variant calling
VCF
Variant annotation
Prioritization
RNA-Seq processing
Count matrix
RNA-Seq data analysis
Functional analysis
RNA-Seq analysis

Qualimap

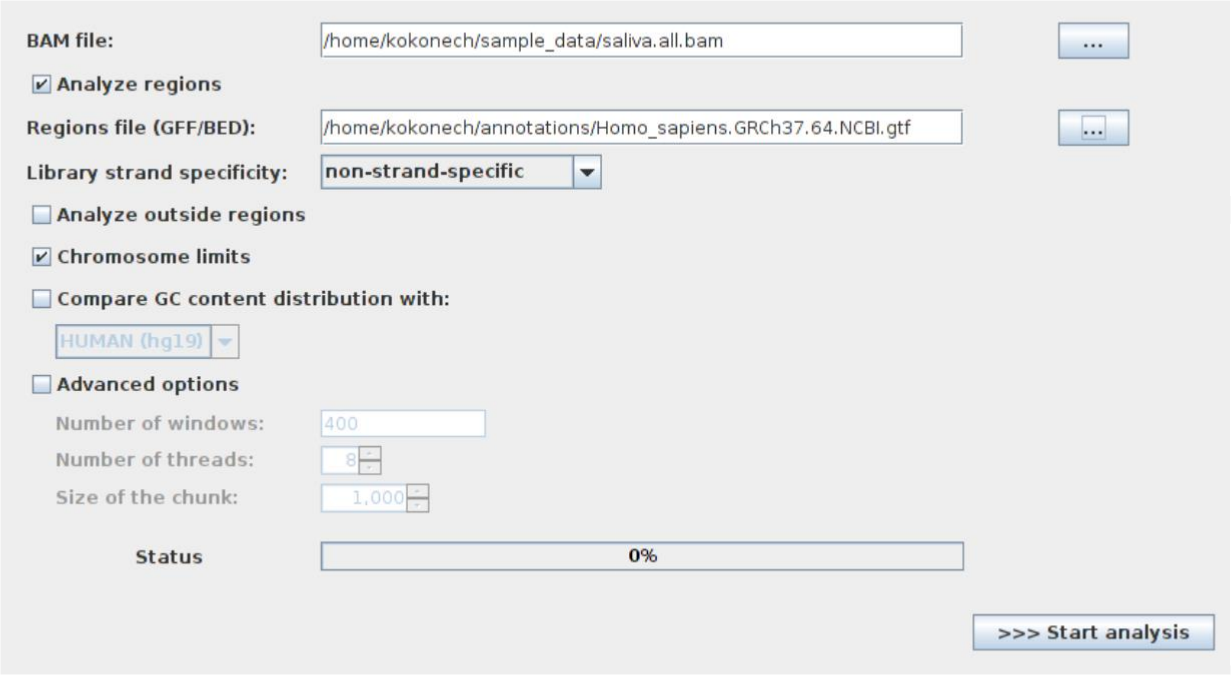
Exercises

Go to

https://github.com/CarinaSilva/Curso-Int-NGS/tree/master/NGS

Exercises 16

Qualimap



BAM file: /home/kokonech/sample_data/saliva.all.bam ...

☑ Analyze regions

Regions file (GFF/BED): /home/kokonech/annotations/Homo_sapiens.GRCh37.64.NCBI.gtf ...

Library strand specificity: non-strand-specific

☐ Analyze outside regions

☑ Chromosome limits

☐ Compare GC content distribution with:

HUMAN (hg19)

☐ Advanced options

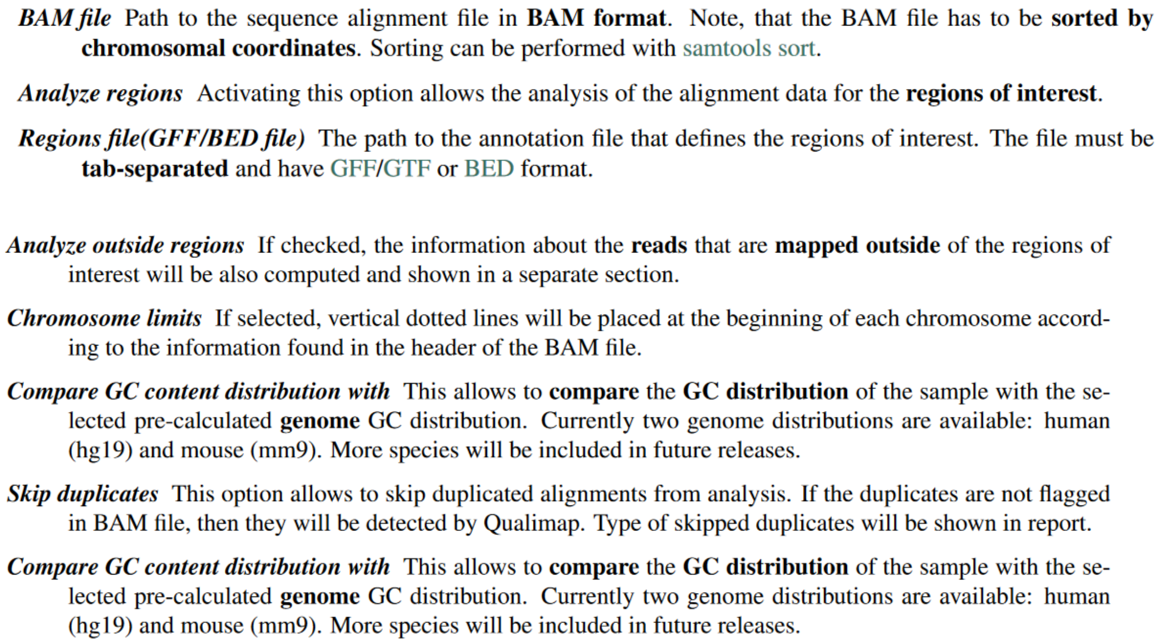
Number of windows: 400

Number of threads: 8

Size of the chunk: 1,000

Status 0%

>>> Start analysis



BAM file Path to the sequence alignment file in **BAM format**. Note, that the BAM file has to be **sorted by chromosomal coordinates**. Sorting can be performed with samtools sort.

Analyze regions Activating this option allows the analysis of the alignment data for the **regions of interest**.

Regions file(GFF/BED file) The path to the annotation file that defines the regions of interest. The file must be **tab-separated** and have GFF/GTF or BED format.

Analyze outside regions If checked, the information about the **reads** that are **mapped outside** of the regions of interest will be also computed and shown in a separate section.

Chromosome limits If selected, vertical dotted lines will be placed at the beginning of each chromosome according to the information found in the header of the BAM file.

Compare GC content distribution with This allows to **compare** the **GC distribution** of the sample with the selected pre-calculated **genome** GC distribution. Currently two genome distributions are available: human (hg19) and mouse (mm9). More species will be included in future releases.

Skip duplicates This option allows to skip duplicated alignments from analysis. If the duplicates are not flagged in BAM file, then they will be detected by Qualimap. Type of skipped duplicates will be shown in report.

Compare GC content distribution with This allows to **compare** the **GC distribution** of the sample with the selected pre-calculated **genome** GC distribution. Currently two genome distributions are available: human (hg19) and mouse (mm9). More species will be included in future releases.

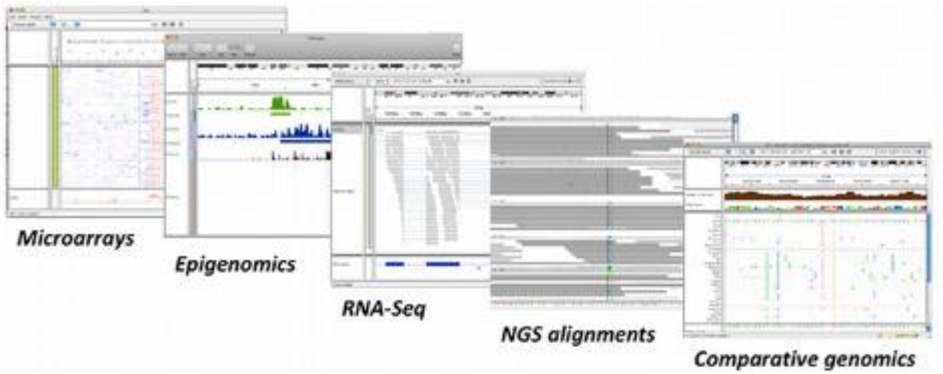
Manual: https://hpc.nih.gov/docs/QualimapManual.pdf

IGV

IGV

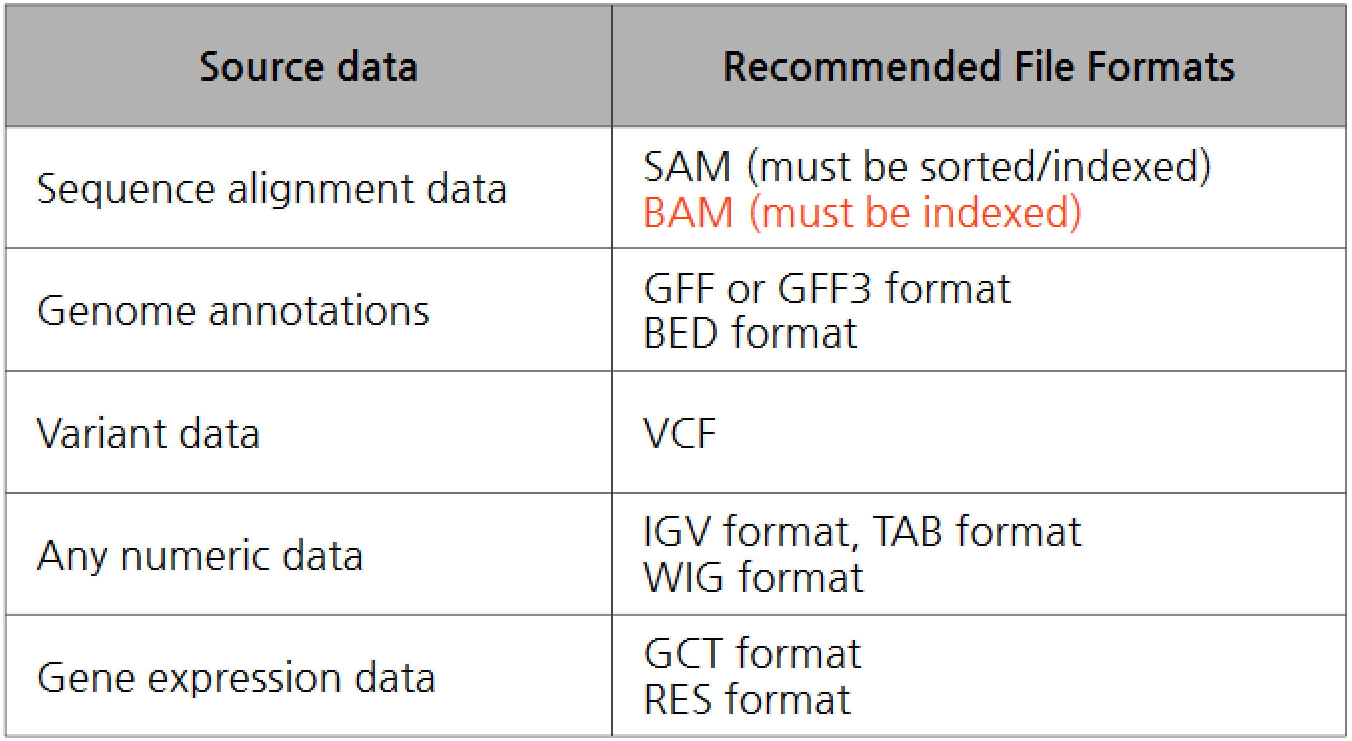
- Integrative Genomics Viewer (**IGV**)
  - **Integrate** different data types simultaneously
  - View **large datasets** easily
  - Faster navigation or browsing
  - Runs **locally** on your desktop
  - Used by large-scale projects
  - Open source and **freely available**

Integrative Genomics Viewer


Microarrays
Epigenomics
RNA-Seq
NGS alignments
Comparative genomics

Manual: http://software.broadinstitute.org/software/igv/userguide

Recommended files



Source data	Recommended File Formats
Sequence alignment data	SAM (must be sorted/indexed) BAM (must be indexed)
Genome annotations	GFF or GFF3 format BED format
Variant data	VCF
Any numeric data	IGV format, TAB format WIG format
Gene expression data	GCT format RES format

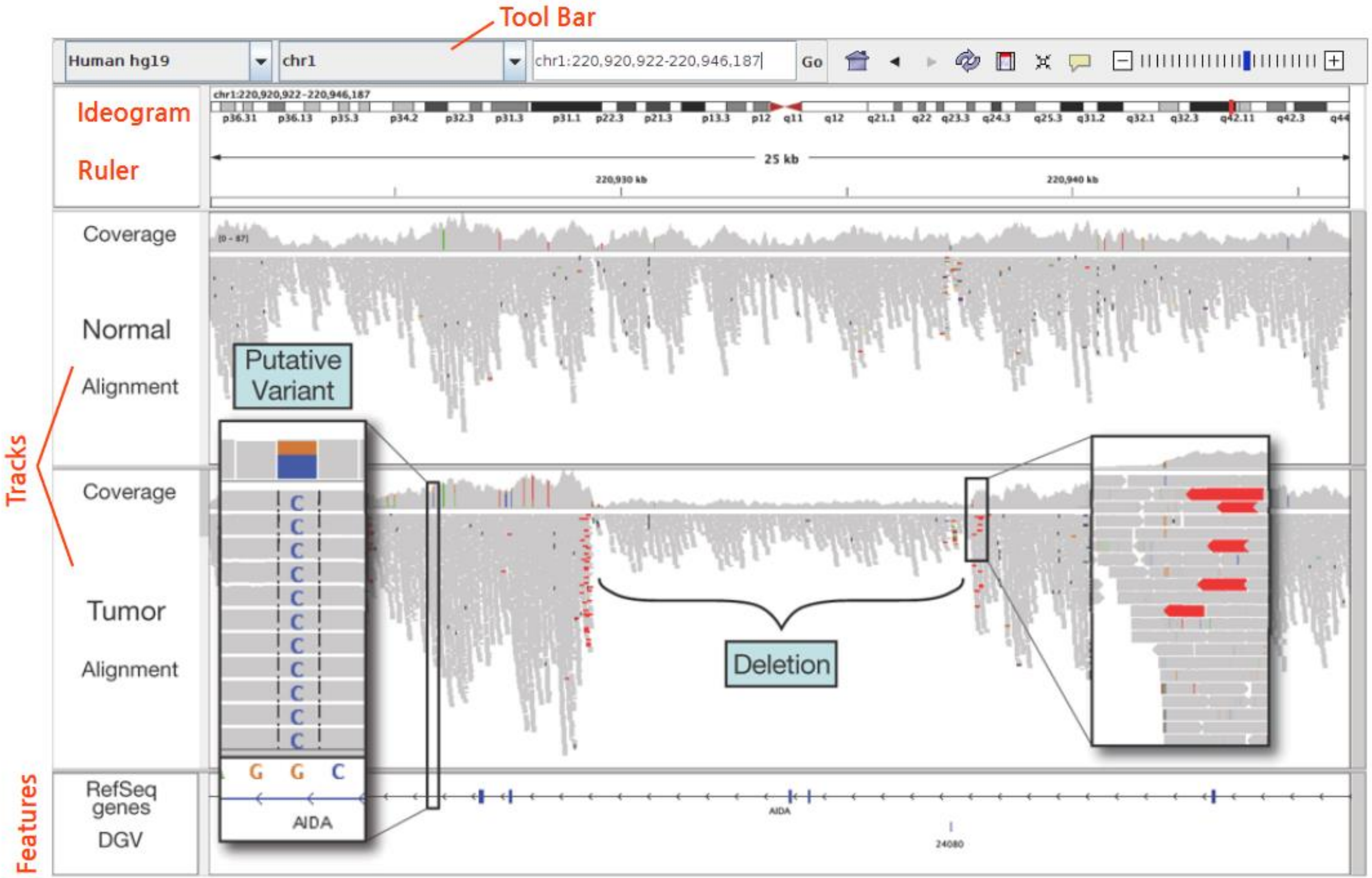
Exercises

Go to

https://github.com/CarinaSilva/Curso-Int-NGS/tree/master/NGS

Exercise 17

IGC interface



Human hg19

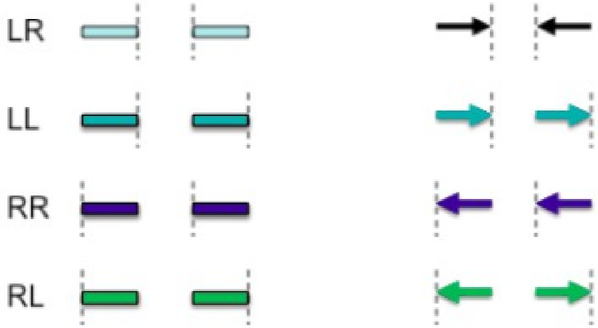
chr1

chr1:220,920,922-220,946,187

Tool Bar
Go
Ideogram
Ruler
chr1:220,920,922-220,946,187
25 kb
220,930 kb
220,940 kb
Coverage
Normal
Alignment
Putative Variant
Tracks
Coverage
Tumor
Alignment
Deletion
C
G
G
C
AIDA
RefSeq genes
DGV
Features
24080

Color interpretation

Read pair orientation



LR	Normal reads. The reads are left and right (respectively) of the unsequenced part of the sequenced DNA fragment when aligned back to the reference genome.
LL,RR	Implies inversion in sequenced DNA with respect to reference.
RL	Implies duplication or translocation with respect to reference.

Insert size

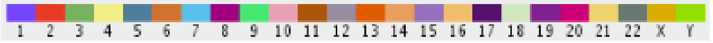
Larger than expected (Deletion)

🟥

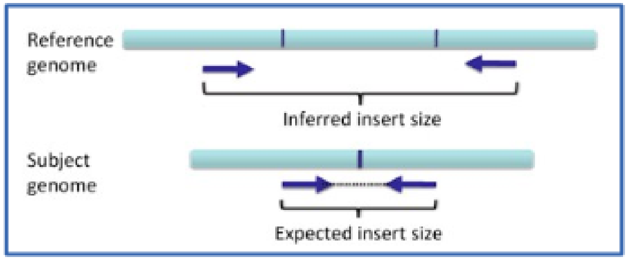
Smaller than expected (Insertion)

🟦

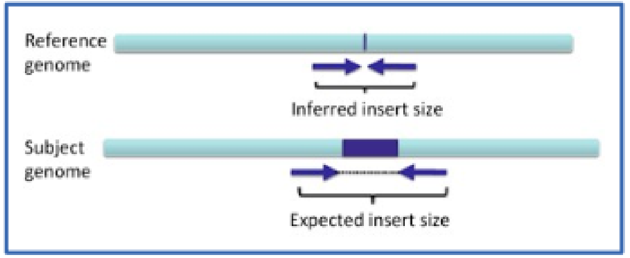
Mate of paired end reads that map to other chromosomes


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y

Deletion


Reference genome
Inferred insert size
Subject genome
Expected insert size

Insertion


Reference genome
Inferred insert size
Subject genome
Expected insert size

To be continued…

Oh, my God!
What should I do now?
FASTA
DELIVERY SERVICE