



4.4 Basic: Bioinformatics workshop on sequencing introducing data formats, analysis and visualization

Paco Hulpiau & Cedric Hermans

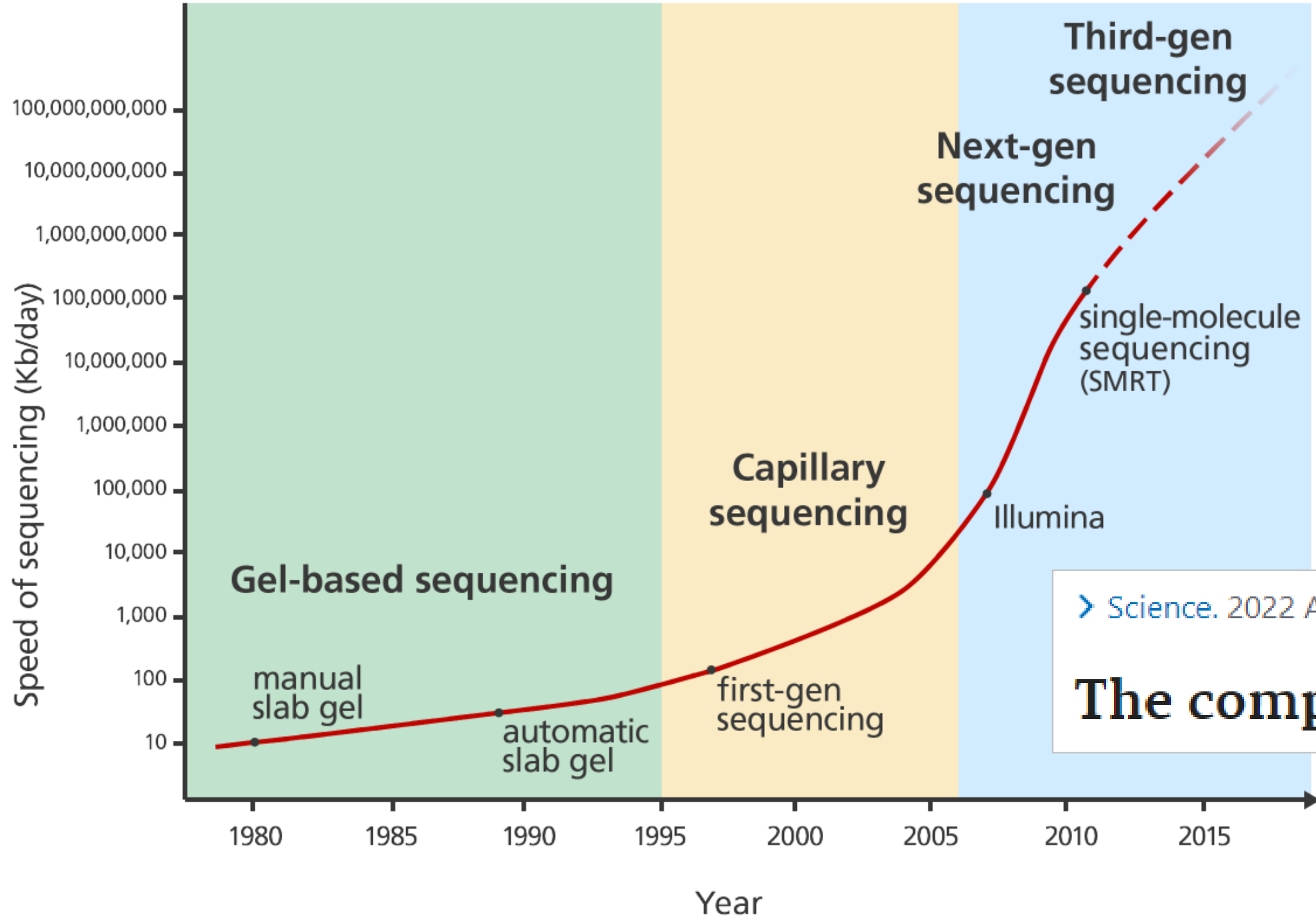
<https://www.bio-informatica.be/workshops/>

Sequencing technology

Human Genome Project

Started in 1990

~ 92% complete in 2003



> [Science](https://doi.org/10.1126/science.abj6987). 2022 Apr;376(6588):44-53. doi: 10.1126/science.abj6987. Epub 2022 Mar 31.

The complete sequence of a human genome

Sequencing technology

Bacteria ~ DNA code 5 000 000 letters

Short reads:

puzzle = 50 000 reads of 100 bases



Short-Read Sequencing

Long reads:

puzzle = 500 reads of 10000 bases



HiFi Sequencing

Source: PacBio HiFi sequencing

Sequencing technology

short read sequencing

next-generation of NGS

Illumina



long read sequencing

Oxford Nanopore Technologies



Accessing public sequencing data

- **Gene Expression Omnibus (GEO)** = international public repository for **high-throughput** microarray and next-generation sequencing **functional genomic data sets** submitted by researchers
- Supports archiving of **raw data**, **processed data** and **metadata** which are indexed, cross-linked and searchable

<https://www.ncbi.nlm.nih.gov/geo/>

Accessing public sequencing data

➤ Example: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE150727>

Series GSE150727		Query DataSets for GSE150727
Status	Public on Jul 15, 2020	
Title	Targeted sequencing of localized and metastatic cutaneous squamous cell carcinoma	
Organism	Homo sapiens	
Experiment type	Genome variation profiling by high throughput sequencing	
Summary	We report sequencing of <u>10 localized and 10 metastatic cutaneous squamous cell carcinomas from human subjects</u> . Sequencing was done on an oncology targeted gene mutation panel consisting of 76 genes.	
Overall design	Case-control study.	
Contributor(s)	Wysong A , Lobl M	
Citation(s)	Lobl MB, Hass B, Clarey D, Higgins S et al. Next-generation sequencing identifies novel single nucleotide polymorphisms in high-risk cutaneous squamous cell carcinoma: A pilot study. <i>Exp Dermatol</i> 2020 Jul;29(7):667-671. <u>PMID: 32479654</u>	
Submission date	May 17, 2020	

Accessing public sequencing data

➤ Example: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE150727>

Platforms (1) [GPL20301](#) Illumina HiSeq 4000 (Homo sapiens)

Samples (20)
[More...](#)
[GSM4557307](#) SCC_Localized [D685]
[GSM4557308](#) SCC_Localized [D720]
[GSM4557309](#) SCC_Localized [D686]

Relations

BioProject [PRJNA633390](#)
 SRA [SRP262054](#)

Download family	Format
SOFT formatted family file(s)	SOFT ?
MINiML formatted family file(s)	MINiML ?
Series Matrix File(s)	TXT ?

Supplementary file	Size	Download	File type/resource
GSE150727_RAW.tar	1.0 Mb	(http)(custom)	TAR (of XLSX)

[SRA Run Selector](#) [?](#)

Raw data are available in SRA

Processed data provided as supplementary file

Accessing public sequencing data

- GPLxxxxx for Platform records
- **GSMxxxxx for Sample records**
- **GSExxxxx for Series records**
- GDSxxxxx for DataSets
- **PRJNAxxxxx → BioProject**
- SRPxxxxx → raw data of Project in Sequence Read Archive (SRA)

Accessing public sequencing data

- **Sequence Read Archive (SRA)** = raw sequencing data and alignment information from high-throughput sequencing platforms is stored
→ available to enhance reproducibility and allow new discoveries by comparing data

- Platforms include: Roche 454, Illumina, Applied Biosystems SOLiD System, Complete Genomics, Oxford Nanopore and Pacific Biosciences

Accessing public sequencing data

➤ Example: <https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA633390>

Found 20 Items

<input checked="" type="checkbox"/> <input type="checkbox"/>	Run ¹	BioSample ²	AvgSpotLen ³	Bases ⁴	Bytes ⁵	Experiment ⁶	GEO_Accession ⁷	Sample Name ⁸	Tumor_location ⁹
<input type="checkbox"/> 1	SRR11804698	SAMN14943414	109	18.67 M	14.43 Mb	SRX8356142	GSM4557307	GSM4557307	Localized
<input type="checkbox"/> 2	SRR11804699	SAMN14943413	115	23.54 M	18.42 Mb	SRX8356143	GSM4557308	GSM4557308	Localized
<input type="checkbox"/> 3	SRR11804700	SAMN14943412	89	15.98 M	12.90 Mb	SRX8356144	GSM4557309	GSM4557309	Localized
<input type="checkbox"/> 4	SRR11804701	SAMN14943411	112	19.54 M	15.38 Mb	SRX8356145	GSM4557310	GSM4557310	Localized
<input type="checkbox"/> 5	SRR11804702	SAMN14943410	110	17.49 M	13.56 Mb	SRX8356146	GSM4557311	GSM4557311	Localized
<input type="checkbox"/> 6	SRR11804703	SAMN14943409	115	23.58 M	18.45 Mb	SRX8356147	GSM4557312	GSM4557312	Localized
<input type="checkbox"/> 7	SRR11804704	SAMN14943408	115	25.10 M	19.56 Mb	SRX8356148	GSM4557313	GSM4557313	Localized
<input type="checkbox"/> 8	SRR11804705	SAMN14943407	115	22.68 M	17.66 Mb	SRX8356149	GSM4557314	GSM4557314	Localized
<input type="checkbox"/> 9	SRR11804706	SAMN14943406	114	29.44 M	22.90 Mb	SRX8356150	GSM4557315	GSM4557315	Localized
<input type="checkbox"/> 10	SRR11804707	SAMN14943405	115	24.33 M	18.98 Mb	SRX8356151	GSM4557316	GSM4557316	Localized
<input type="checkbox"/> 11	SRR11804708	SAMN14943404	113	56.00 M	43.57 Mb	SRX8356152	GSM4557317	GSM4557317	Metastatic
<input type="checkbox"/> 12	SRR11804709	SAMN14943403	114	24.04 M	18.85 Mb	SRX8356153	GSM4557318	GSM4557318	Metastatic

Accessing public sequencing data

➤ **iGenomes** = collection of reference sequences and annotation files for commonly analyzed organisms

https://support.illumina.com/sequencing/sequencing_software/igenome.html

Species	Source	Build(s)
<i>Homo sapiens</i>	Ensembl	GRCh37
	NCBI	<ul style="list-style-type: none"> GRCh38 GRCh38Decoy
		<ul style="list-style-type: none"> build37.2 build37.1 build36.3
UCSC	hg38	<ul style="list-style-type: none"> hg19–Does not have annotation files. hg19–Has the latest annotation files. Use with LRM DNA Amplicon Analysis modules v1.1 and v2.0 hg19–Use with LRM DNA Amplicon Analysis module v1.0

Accessing public sequencing data

- **NCBI Genome** → **NCBI Datasets** = resources include information on large-scale genomics projects, **genome** sequences and **assemblies**, and mapped **annotations**
- Example: <https://www.ncbi.nlm.nih.gov/datasets/taxonomy/28901/>

NCBI Datasets Taxonomy Genome Gene Command-line tools Documentation

Genome

Browse all 517,426 genomes

Subspecies

Salmonella enterica subsp. *arizonae*

Genomes

504

Salmonella enterica subsp. *diarizonae*

754

Salmonella enterica subsp. *enterica*

214,412

Reference genome

ASM694v2

Washington University Genome Sequencing Center (2016). Strain: LT2.

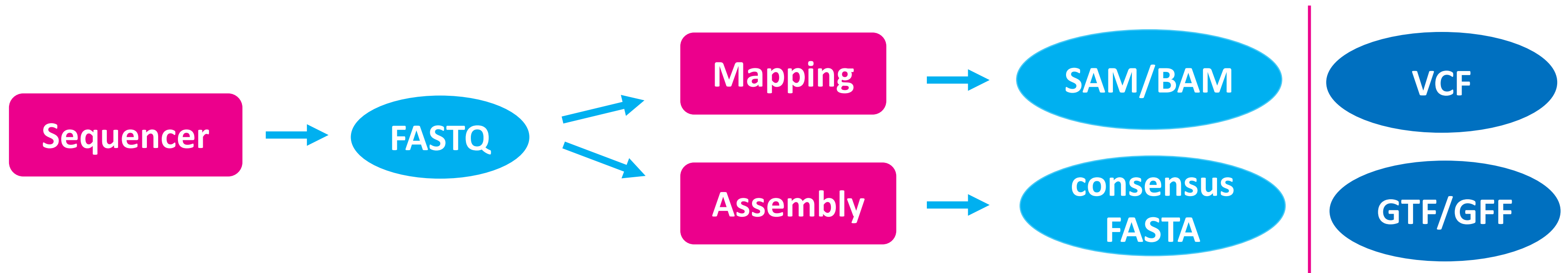
RefSeq GCF_000006945.2

Download

Accessing public sequencing data

➤ **FASTQ** = common file format for sequence read data

combining both sequence and associated per base quality score (~ FASTA + quality)



Data formats

FASTA format

- **FASTA format** = simple text-based format for representing sequences (nt and aa)
 - first line (header): summary description of sequence
 - starts with ">" followed by accession number or other unique identifier


 SARS-CoV-2_Belgium_2021.fasta - Notepad

File Edit Format View Help

```
>OL672836.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/BEL/reg-20174/2021, complete genome
AGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACCTCGGCTGCATGCTTAGTGCACTCACG
CAGTATAATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCT
TACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTTGTCCGGGTGTGACCGAAAGGTAA
GATGGAGAGCCTTGTCCTGGTTTCAACGAGAAAACACACGTCCAACCTCAGTTTGCCTGTTTTACAGGTT
CGCGACGTGCTCGTACGTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTA
AAGATGGCACTTGTGGCTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTT|
CATCAAACGTTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAA
GGCATTACAGTACGGTCGTAGTGGTGAGACACTTGGTGTCCCTGTCCCTCATGTGGGCGAAATACCAGTGG
CTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCT
```

FASTA format

- **FASTA format** = simple text-based format for representing sequences (nt and aa)
 - actual sequence on the line(s) following the first header line
 - filename extensions: .fasta , .fa, .fna , .faa

 SARS-CoV-2_Belgium_2021.fasta - Notepad

File Edit Format View Help

```
>OL672836.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/BEL/reg-20174/2021, complete genome
AGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACCTCGGCTGCATGCTTAGTGCACTCACG
CAGTATAATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCT
TACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTTGTCCGGGTGTGACCGAAAGGTAA
GATGGAGAGCCTTGTCCTGGTTTCAACGAGAAAACACACGTCCAACCTCAGTTTGCCTGTTTTACAGGTT
CGCGACGTGCTCGTACGTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTA
AAGATGGCACTTGTGGCTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTT|
CATCAAACGTTCCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAA
GGCATTACAGTACGGTCGTAGTGGTGAGACACTTGGTGTCCCTGTCCCTCATGTGGGCGAAATACCAGTGG
CTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCT
```


BCL format

➤ **BCL format = base call files**

➤ Sequencing run software

→ generates BCL file containing base calls and associated quality scores (Q-scores)

➤ Most data analysis applications require **FASTQ** files as input

→ illumina **bcl2fastq** conversion software

https://emea.support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html

FAST5 format

- **FAST5 format** = standard sequencing output in which raw signals are stored by Oxford Nanopore Technologies (ONT) sequencers e.g. MinION
 - based on the hierarchical data format HDF5 format which enables storage of large and complex data
 - basecalling via guppy (or bonito) ONT basecallers
- **POD5 format** = more recent prototype file format for raw signal data
 - POD5 is replacing FAST5 as native file format

<https://github.com/nanoporetech/pod5-file-format>

FASTQ format

- Basecalling software reads signals from sequencer
 - calls bases and assigns a quality value to each base called
 - introduced **PHRED quality score** of a base call
 - PHRED scores are now standard for representing sequencing read base qualities

$$Q_{\text{PHRED}} = -10 \times \log_{10}(P_e)$$

P_e is the estimated probability of error (in this case, the estimated probability of the base the call being wrong)

- PHRED scores used in FASTQ, also used in SAM format

FASTQ format

- Storing **PHRED scores** as single characters (or bytes)

→ space efficient encoding

- File to be human readable and easily edited

→ restricted choices to ASCII printable

characters (! to ~) = 33–126 (decimal)

→

ASCII code	Char
64	@

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

FASTQ format

- Phred assigns Q score of 30 (Q30) to a base → equivalent to probability of incorrect base call 1 in 1000 times
- **Q30 considered benchmark for quality in NGS**
- Q20 → 99% → incorrect base call 1 of 100

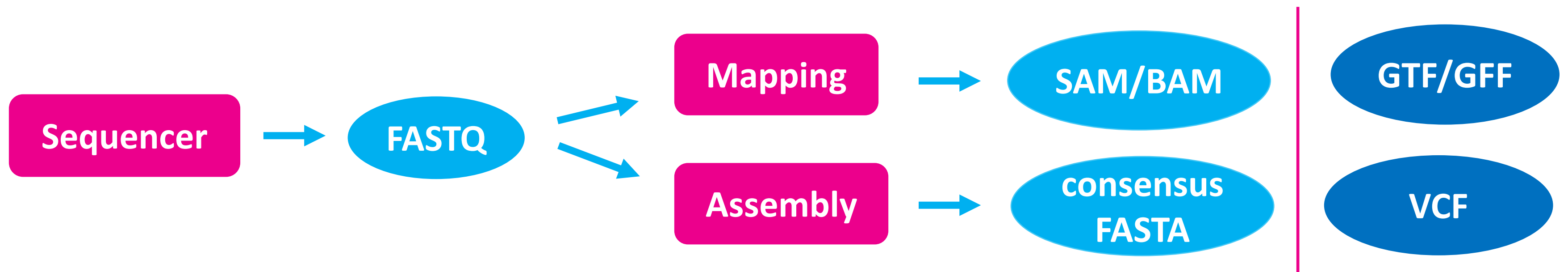
Table 1: Quality Scores and Base Calling Accuracy

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
➔ 20	1 in 100	99%
➔ 30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

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

➤ **FASTQ** = common file format for sequence read data

combining both sequence and associated per base quality score (~ FASTA + quality)



GFF and GTF format

- **GFF** and **GTF** are file formats with **genomic annotation information** for next-generation sequencing data analysis
- GFF = General Feature Format, current version GFF3 (.gff)
- GTF = General Transfer format, identical to GFF version 2 (.gtf)

GFF and GTF format

- Each feature is represented on one line of text and consists of nine columns (fields) of data values
- Data field is tab-separated and must contain value (empty → .)
- Optional: track definition lines

```
#!genome-build GRCh38
#!genome-date 2013-12
#!genome-build-accession NCBI:GCA_000001405.15
#!genebuild-last-updated 2014-08
1 Havana    gene  11869  14409  .  +  .  gene_id "ENSG00000223972"
1 Havana    exon  11869  14409  .  +  .  gene_id "ENSG00000223972"
1 Havana    exon  11869  12227  .  +  .  gene_id "ENSG00000223972"
1 Havana    exon  12613  12721  .  +  .  gene_id "ENSG00000223972"
```

1	2	3	4	5	6	7	8	9
chr1	unknown	stop_codon	1197845	1197847	.	+	.	gene_id "TTLL10"; gene_name "TTLL10"; p_id "P14573"; transcript_id "NM_001130045";
chr1	unknown	exon	1203508	1203960	.	-	.	gene_id "TNFRSF18"; gene_name "TNFRSF18"; p_id "P10164"; transcript_id "NM_148901"; tss_id
chr1	unknown	exon	1203508	1203968	.	-	.	gene_id "TNFRSF18"; gene_name "TNFRSF18"; p_id "P37213"; transcript_id "NM_148902"; tss_id
chr1	unknown	exon	1203508	1203968	.	-	.	gene_id "TNFRSF18"; gene_name "TNFRSF18"; p_id "P26347"; transcript_id "NM_004195"; tss_id
chr1	unknown	stop_codon	1203591	1203593	.	-	.	gene_id "TNFRSF18"; gene_name "TNFRSF18"; p_id "P10164"; transcript_id "NM_148901"
chr1	unknown	CDS	1203594	1203960	.	-	1	gene_id "TNFRSF18"; gene_name "TNFRSF18"; p_id "P10164"; transcript_id "NM_148901"; tss_id

➤ Structure of the **feature line**:

1. Name of sequence where feature is located
2. Keyword identifying source of feature
3. Feature type name (e.g. gene, exon, CDS)
4. Start (1-base offset)
5. End (1-base offset)
6. Score
7. Strand ("+" positive or "-" negative or "." undetermined)
8. Frame (GTF) or phase (GFF3) of CDS features
9. Attributes

Example: [genes1000.gtf](#)

SAM/BAM/BAI format

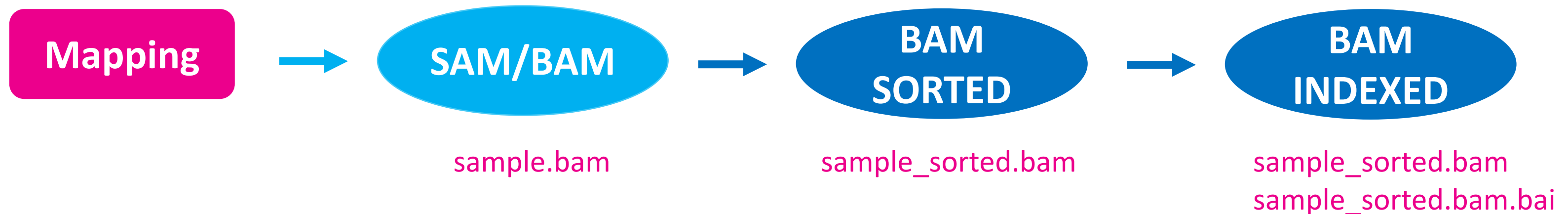
- **Raw sequence reads** from sequencers have no genomic position information and **must be mapped or aligned to a known reference genome**
- **Sequence Alignment/Map** or **SAM** format → generic alignment format for storing read alignments against reference sequences
- Supports short and long reads (up to 128Mbp) produced by different sequencing platforms

SAM/BAM/BAI format

- **SAM file** is a tab-delimited text file
- **BAM format** is the **compressed binary version** of the SAM format
- BAM files can be indexed for fast retrieval of alignments overlapping a specified region without going through whole alignment
 - **companion index file** of a bam file is in the **BAI format** (.bai)

SAM/BAM/BAI format

- Before indexing, BAM must be sorted by reference ID and leftmost coordinate



- To convert SAM <-> BAM and sort/index the files → **samtools** (<http://www.htslib.org/>)

SAM/BAM/BAI format

@HD VN:1.5 SO:coordinate											Header section
@SQ SN:ref LN:45											
r001	99	ref	7	30	8M2I4M1D3M	=	37	39	TTAGATAAAGGATACTG	*	Alignment section
r002	0	ref	9	30	3S6M1P1I4M	*	0	0	AAAAGATAAGGATA	*	
r003	0	ref	9	30	5S6M	*	0	0	GCCTAAGCTAA	* SA:Z:ref,29,-,6H5M,17,0;	
r004	0	ref	16	30	6M14N5M	*	0	0	ATAGCTTCAGC	*	
r003	2064	ref	29	17	6H5M	*	0	0	TAGGC	* SA:Z:ref,9,+,5S6M,30,1;	
r001	147	ref	37	30	9M	=	7	-39	CAGCGGCAT	* NM:i:1	

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

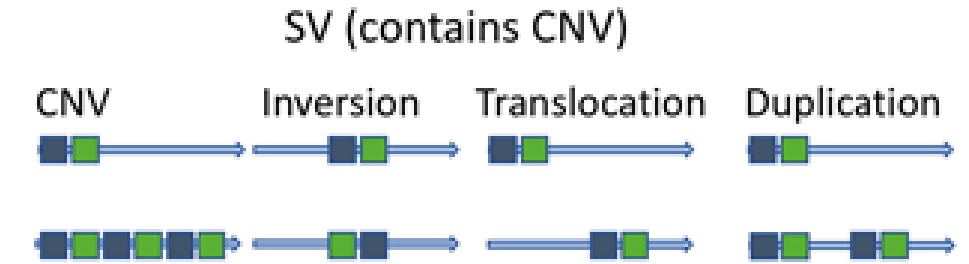
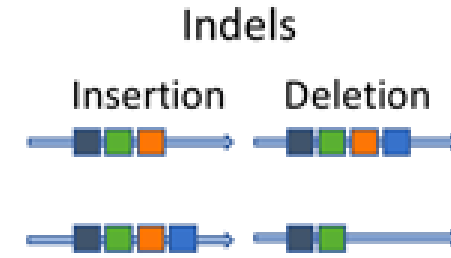
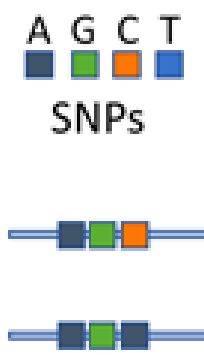
SAM/BAM/BAI format

- SAM/BAM files contain a **header section** (optional) and an **alignment section**
- Each **header line** begins with @ character followed by one of two-letter header record type codes: @HD (header definition start), @SQ (reference sequence dictionary), @RG (read group information), @PG (program information), @CO (one-line text comment)

SAM/BAM/BAI format

- The **alignment section** contains sequences with genomic position and other descriptive information → each single sequence (short read from FASTQ) and its associated information are presented as one line text
→ consists of 11 mandatory, tab-delimited text fields

VCF



- **Variant Call Format (VCF)** is a text file format for **variation data** such as:
 - ✓ single nucleotide variant (SNP)
 - ✓ insertion/deletion (indel)
 - ✓ copy number variant (CNV)
 - ✓ structural variant (SV)

Ensembl Variation - Variant classification

Sequence variants

Type	Description	Example (Reference / Alternative)	
SNP	Single Nucleotide Polymorphism	Ref: ...TTG A CGTA...	Alt: ...TTG G CGTA...
Insertion	Insertion of one or several nucleotides	Ref: ...TTGACGTA...	Alt: ...TTGAT G CGTA...
Deletion	Deletion of one or several nucleotides	Ref: ...TTG AC GTA...	Alt: ...TTGGTA...
Indel	An insertion and a deletion, affecting 2 or more nucleotides	Ref: ...TTG AC GTA...	Alt: ...TTG GCT CGTA...
Substitution	A sequence alteration where the length of the change in the variant is the same as that of the reference.	Ref: ...TTG AC GTA...	Alt: ...TTG TAG TA...

Structural variants

Type	Description	Example (Reference / Alternative)	
CNV	Copy Number Variation: increases or decreases the copy number of a given region	Reference: 	"Gain" of one copy: "Loss" of one copy:
Inversion	A continuous nucleotide sequence is inverted in the same position	Reference: 	Alternative:
Translocation	A region of nucleotide sequence that has translocated to a new position	Reference: 	Alternative:

VCF

Example: [129P2_1000lines.vcf](#)

- Meta-information lines (##)
- A tab-delimited header line (#)
- Data lines each containing information about a position in the genome (also tab-delimited)

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT 129P2_01aHsd
1 3000019 . G GA 20.8655 MinDP;MinAB INDEL;DP4=1,0,3,0;DP=4;CSQ=A|||intergenic_variant|
1 3001236 . A ATTTT,ATTT 179 Het INDEL;DP4=10,3,5,17;DP=35;CSQ=TTTT|||intergenic_va
```

VCF

Example: [129P2_1000lines.vcf](#)

➤ The **header line** starting with #CHROM contains 8 fixed, mandatory columns

1. #CHROM = chromosome identifier from reference genome
2. POS = reference position with 1st base having position 1, sorted numerically
3. ID = semi-colon separated list of identifiers
4. REF = reference base(s) in uppercase (A,C,G,T,N)
5. ALT = comma separated list of alternate non-reference alleles
6. QUAL = phred-scaled quality score for assertion in ALT (high → high confidence)
7. FILTER = PASS if this position has passed all filters (if not: semi-colon list of codes)
8. INFO = additional information as semicolon separated series of <key>=<value>

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT 129P2_01aHsd
1 3000019 . G GA 20.8655 MinDP;MinAB INDEL;DP4=1,0,3,0;DP=4;CSQ=A|||intergenic_variant|
1 3001236 . A ATTTT,ATTT 179 Het INDEL;DP4=10,3,5,17;DP=35;CSQ=TTTT|||intergenic_va
```

VCF

Example: [129P2_1000lines.vcf](#)

- If **genotype data** is present in the file → **FORMAT column**
- First format is given specifying data types and order (e.g. **GT:GQ:DP:HQ**)
- Followed by one field (column) per sample (e.g. **A|A:::23:23,34**)
- **GT** = genotype, **DP** = read depth, **FT** = sample genotype filter,
GL = three likelihoods for AA, AB, BB genotypes (A=ref, B=alt),
GQ = genotype quality, **HQ** = haplotype quality

Data analysis and visualization

Quality control

➤ Quality control (QC) of sequencing data (FASTQ files) generated by high throughput technologies is important for meaningful downstream analysis

➤ One of most popular QC tools is **FastQC**

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

→ runs a **series of tests** on .fastq(.gz) file and generates comprehensive QC report

Quality control

- Each test is flagged as a **pass**, **warning** or **fail** in comparison with what you expect from a normal large dataset
- Warnings (and even failures) do not necessarily mean there is a problem with the data but tells you it is unusual
- Look at the typical results in the output html file: [SRR11804708_fastqc.html](#)



Quality control

- Fastqc documentation explains each flag in the report:

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/>

- **Per base sequence quality**

→ shows overview of range of quality values across all bases at each position in file

→ each position box-whisker-plot:

red line for median quality, **blue line** is mean quality, **yellow box** for IQR

Quality control

Summary

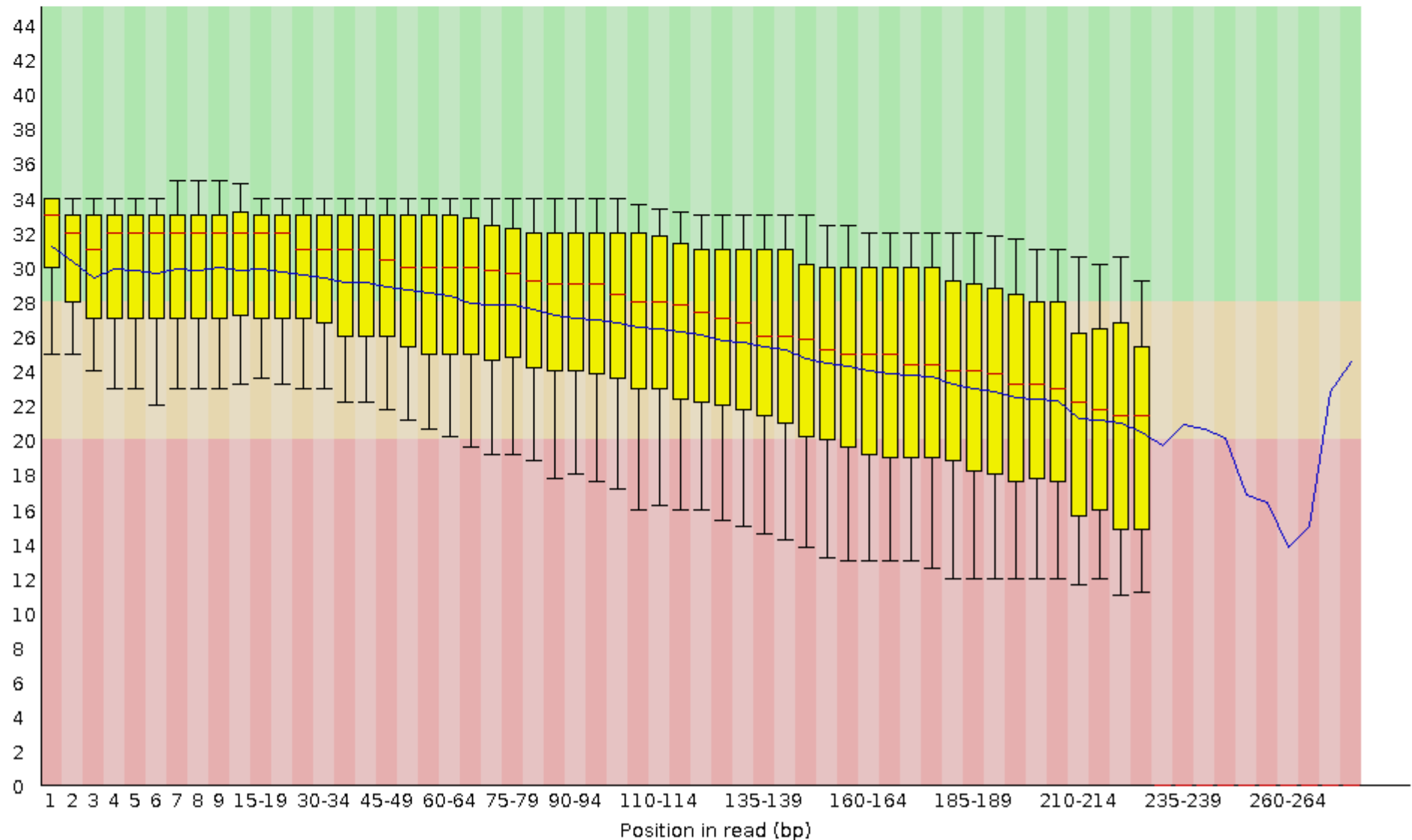
- ✔ [Basic Statistics](#)
- ! [Per base sequence quality](#)
- ✔ [Per sequence quality scores](#)
- ✘ [Per base sequence content](#)
- ! [Per sequence GC content](#)
- ✔ [Per base N content](#)
- ! [Sequence Length Distribution](#)
- ✘ [Sequence Duplication Levels](#)
- ! [Overrepresented sequences](#)
- ✔ [Adapter Content](#)

✔ Basic Statistics

Measure	Value
Filename	SRR11804708.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	492006
Sequences flagged as poor quality	0
Sequence length	8-276
%GC	47

! Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

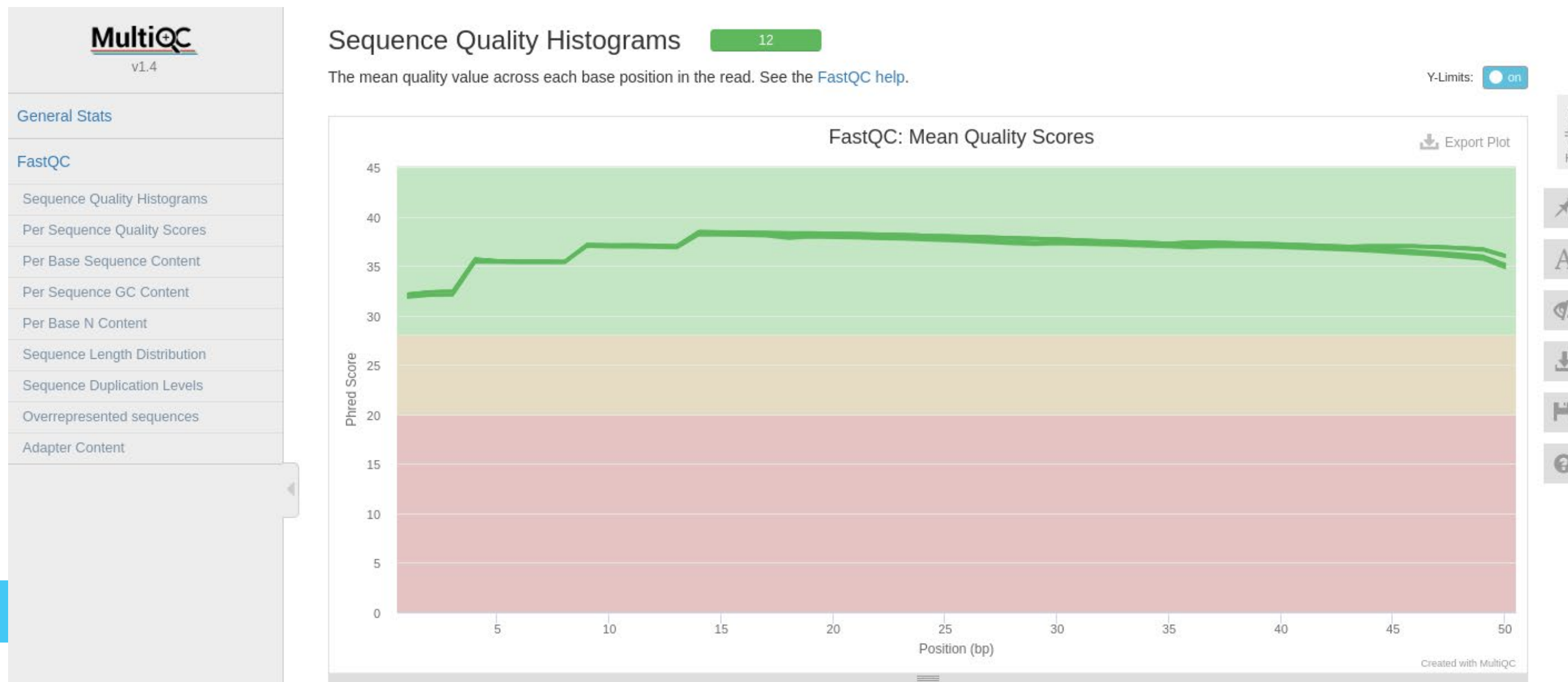


Example: [SRR11804708_fastqc.html](#)

Quality control

Example: [multiqc_report.html](https://multiqc.info)

- **MultiQC** (<https://multiqc.info>) is a reporting tool that parses summary statistics from results and log files generated by other bioinformatics tools such as fastqc



Preprocessing data: read trimming

- Raw reads from Next Generation Sequencing are processed prior to analysis
- One of most used preprocessing procedure is **read trimming**
 - **remove low quality bases** while preserving longest high quality part of read
 - trimming is beneficial in RNA-seq, SNP identification and genome assembling
 - some tools can also **remove** (Illumina) **adapters**
- Trimming tools can be classified in two classes based on algorithm:
e.g. **Trimmomatic** (window based) and Cutadapt (running sum)

After trimming...

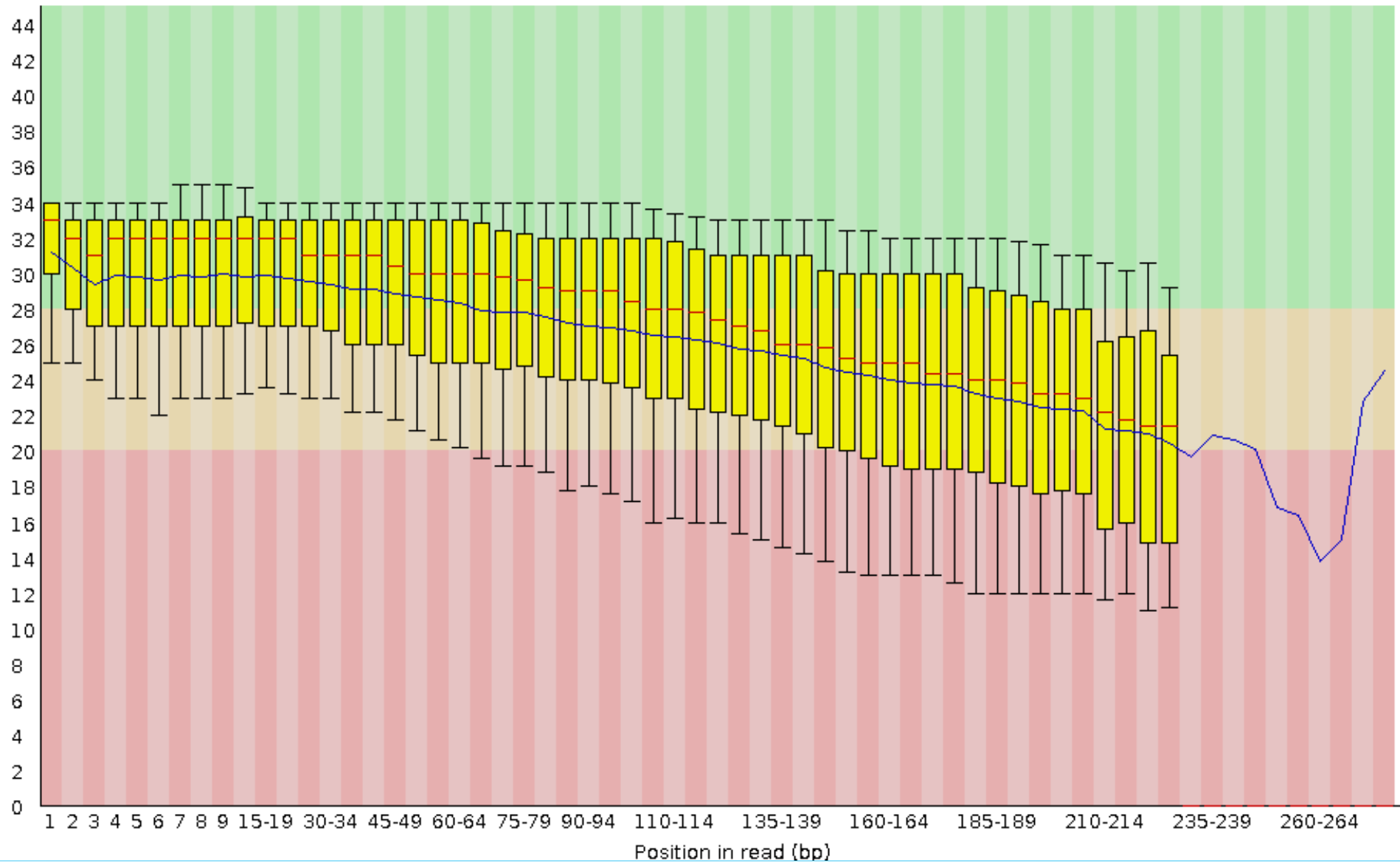
Input Reads: 492006 Surviving: 450080 (91.48%) Dropped: 41926 (8.52%)

Basic Statistics

Measure	Value
Filename	SRR11804708.fastq.gz

Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

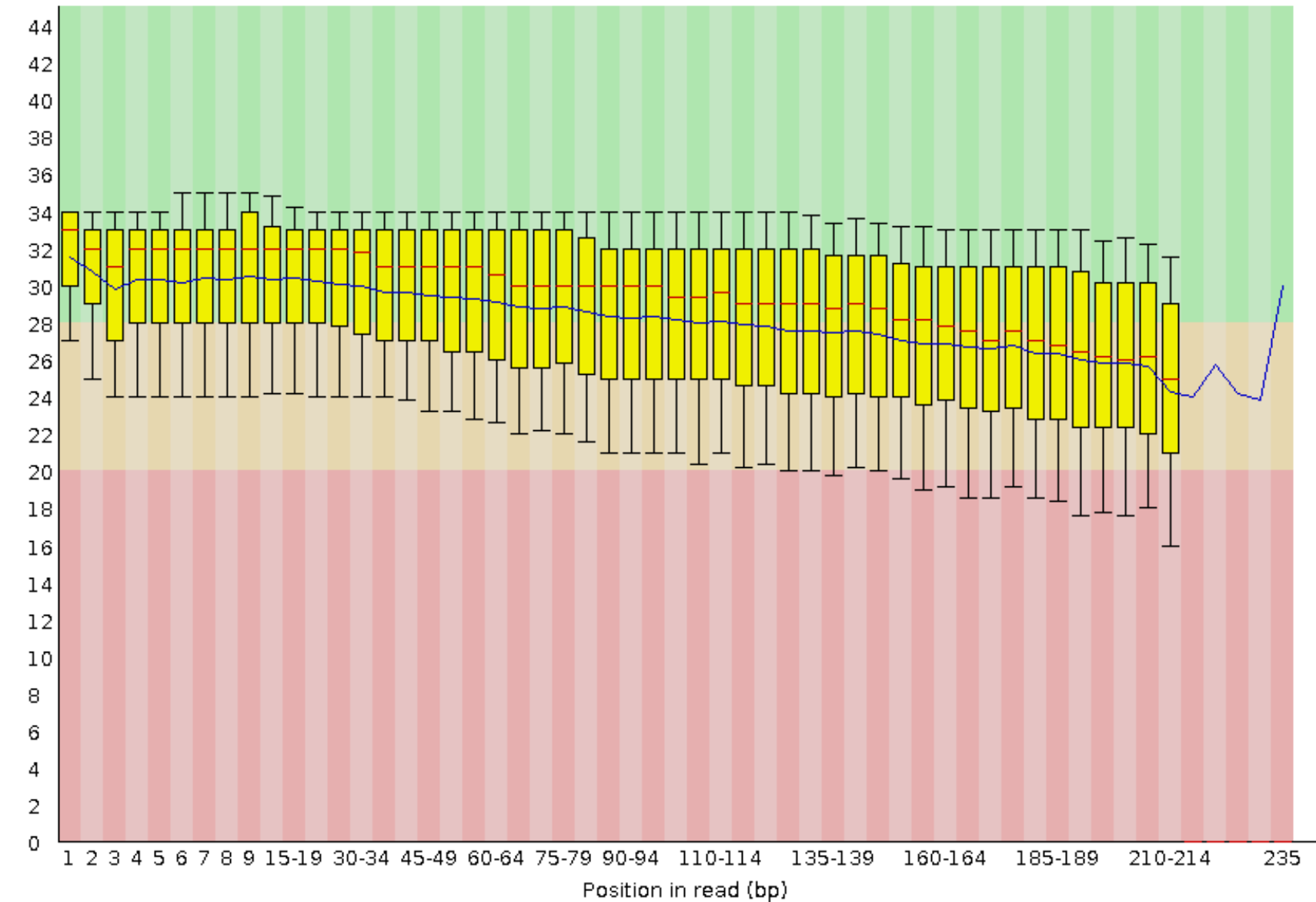


Basic Statistics

Measure	Value
Filename	SRR11804708_trimmed.fastq.gz

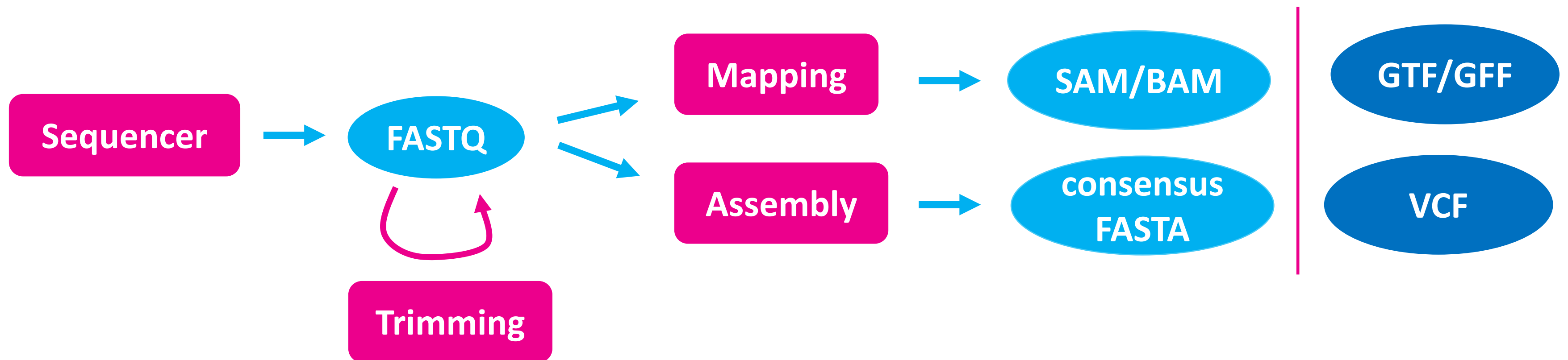
Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)



➤ **FASTQ** = common file format for sequence read data

combining both sequence and associated per base quality score (~ FASTA + quality)



Mapping data

- Fundamental step in high-throughput sequence analysis is alignment or **mapping of the reads to the reference sequence**
- Many software tools available
 - when considering choice → suitable for specific application?
 - type of data (DNA, RNA, miRNA, bisulfite), sequencing platform

Mappers

Color legend:

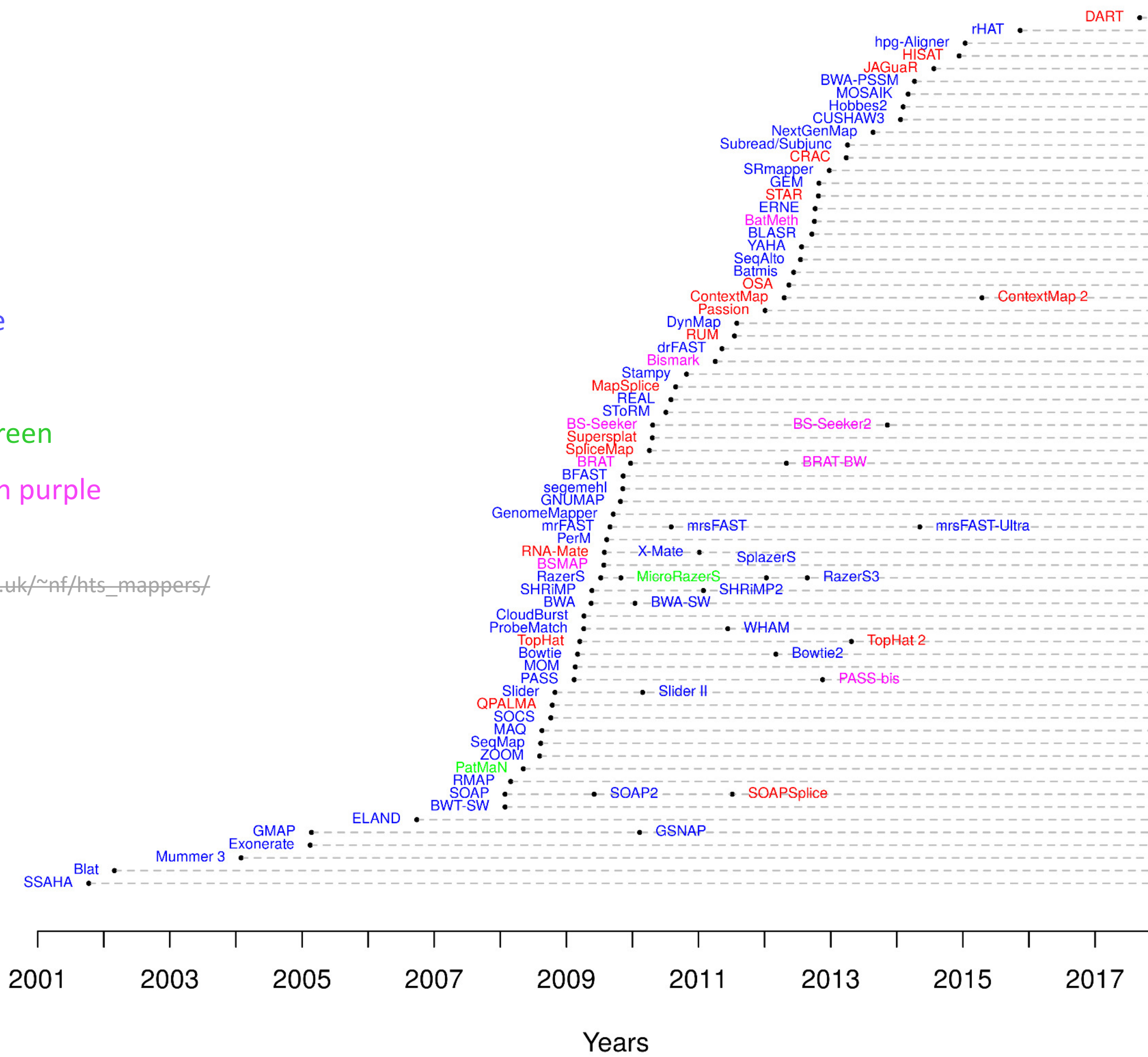
DNA mappers in blue

RNA mappers in red

miRNA mappers in green

Bisulphite mappers in purple

Source: https://www.ebi.ac.uk/~nf/hts_mappers/



Mappers

- **DNA** mappers: most used are **BWA** and **Bowtie2**
- **RNA** mappers: most used are Tophat2/**HISAT2** and **STAR**
- DNA/RNA mapper for long reads: **Minimap2**

IGV

➤ IGV = Integrative Genomics Viewer

→ open source (free) desktop **genome visualization tool** for Windows/Mac/Linux

→ also available as webapp: <https://igv.org/app/>

➤ Load the **.sorted.bam** and **.sorted.bai** files as demonstrated

to visualize the SNP [rs713598](#)

in region chr7:141,973,472-141,973,627

Geno	Mag	Summary
(C;C)	1.1	Can taste bitter.
(C;G)	1.1	Can taste bitter.
(G;G)	1.1	Possibly unable to taste bitter in some foods.

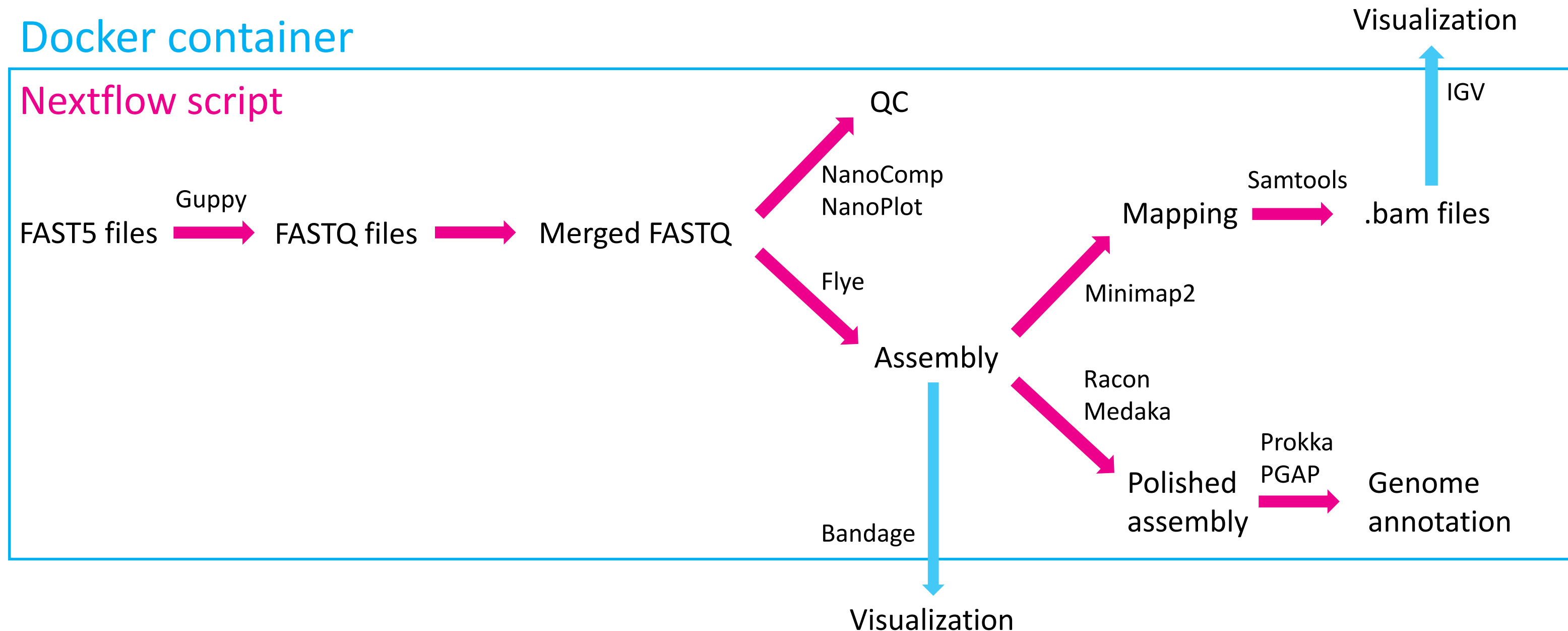
Reference GRCh38 38.1/141
Chromosome 7
Position 141973545
Gene TAS2R38

Example of a bioinformatics workflow / pipeline

➤ WGS of 50 bacterial strains → 2TB of raw data → 224GB fastq data → assembled in ~41 hours

Docker container

Nextflow script



Galaxy

- Open web-based platforms (125+)
 - facilitates centralized data analysis
- Analysis using available tools
 - integrate diverse set of readily available tools tailored for different analysis needs
- No IT-knowledge required
 - intuitive user interface and workflow, promoting accessibility

Galaxy e.g. usegalaxy.eu

Display and analysis screen

History

Tools

The screenshot shows the Galaxy Europe web interface. The top navigation bar includes the Galaxy logo, 'Europe', and a menu with 'Workflow', 'Visualize', 'Shared Data', 'Help', 'Log in or Register', and a 'Using 0%' indicator. On the left, a 'Tools' sidebar is visible, featuring a search bar, an 'Upload Data' button, and categories like 'GENERAL TEXT TOOLS' and 'GENOMIC FILE MANIPULATION'. The main content area is titled 'The European Galaxy server' and contains a descriptive paragraph, a 'Browse installed tools' button, a 'Request temporary increase of quota' button, a 'Request TlaaS' button, and a 'Check the status of the server' button. Below this are four informational cards for 'Projects', 'Communities', 'Citation', and 'Team'. On the right, a 'History' sidebar shows an 'Unnamed history' section with a message: 'This history is empty. You can load your own data or get data from an external source.'

WORKSHOP

Nanopore sequencing

8 & 10 November 2023

PROGRAMMA

- **Intake gesprek:**

Er zal eerst een online intakegesprek worden georganiseerd op basis van de beschikbare tijdstippen van alle deelnemers. Tijdens dit gesprek zullen de DNA stalen besproken worden die tijdens de workshop zullen worden gesequenced. **Het is dus mogelijk om uw eigen staal te sequencen!** Daarnaast zal ook de werking en techniek worden toegelicht.

- **8 november 2023:**

Tijdens deze dag gaan we aan de slag in het labo met al het beschikbare DNA materiaal. **We doen de nodige stappen voor het maken van een DNA library en we voeren de DNA sequencer uit op de beschikbare flowcellen.** We laten de runs lopen tot er geen actieve poriën meer beschikbaar zijn.

- **10 november 2023:**

We gaan aan de slag met de data van onze runs. Hiervoor gaan we aan de slag met een in-huis ontwikkelde tool, Epi2ME en stellen we Galaxy voor. **We geven jullie dus de kans om met verschillende tools te werken om zelf aan de slag te gaan met jullie data.**

PRAKTISCHE INFO

- **Hoe inschrijven?**

Bij interesse of meer info: marjolein.vandekerckhove@howest.be

Mailen is de boodschap om in te schrijven, de plaatsen zijn beperkt (maximum 8 deelnemers)

New workshops planned

➤ May 2024

➤ September 2024



bio-informatica.be

ADVANCED BACHELOR
OF BIOINFORMATICS

BIOINFORMATICS @HOME
VIA DISTANCE LEARNING

NANOPORE SEQUENCING
HANDS-ON WORKSHOP

AI FOR
HEALTHCARE