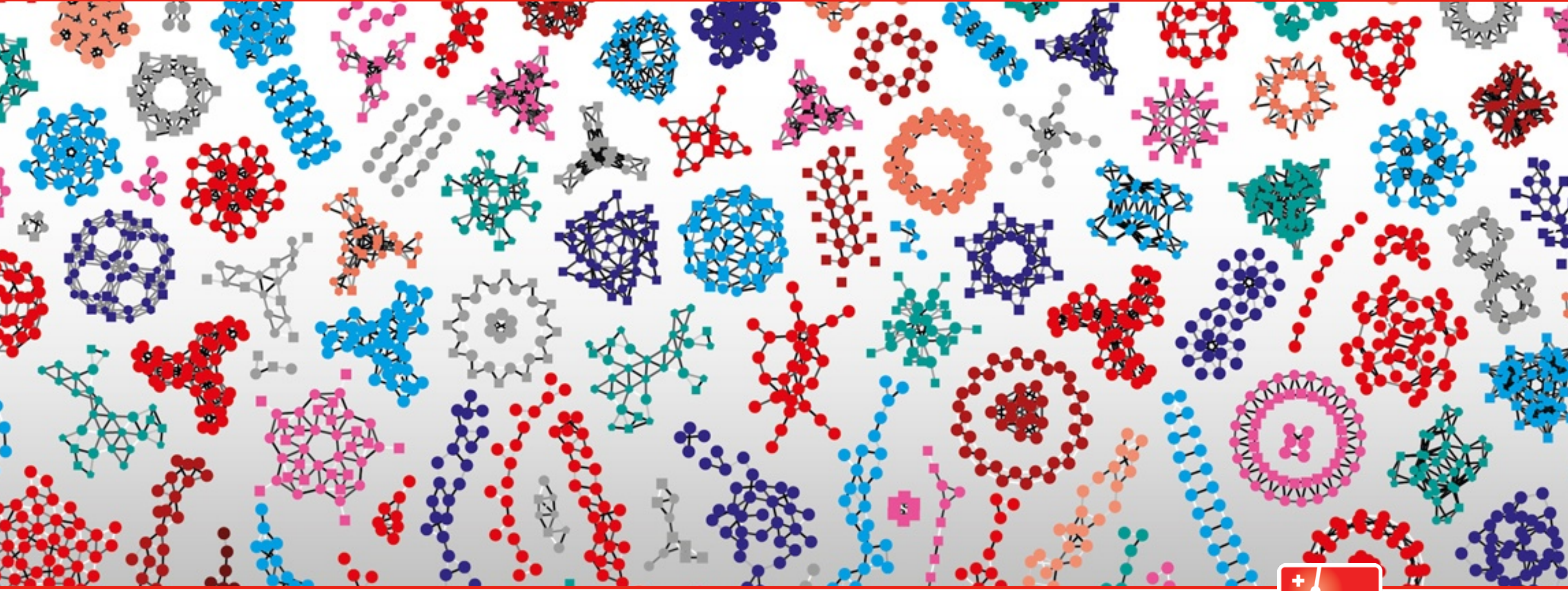


Swiss Institute of Bioinformatics

Bioinformatics pipeline overview for genomics and metagenomics

ESCMID Postgraduate Technical Workshop
Clinical bioinformatics for microbial genomics and metagenomics

Dr Aitana Lebrand | Lausanne, 9-12 September 2019

www.sib.swiss

(Meta)genomics for infectious diseases

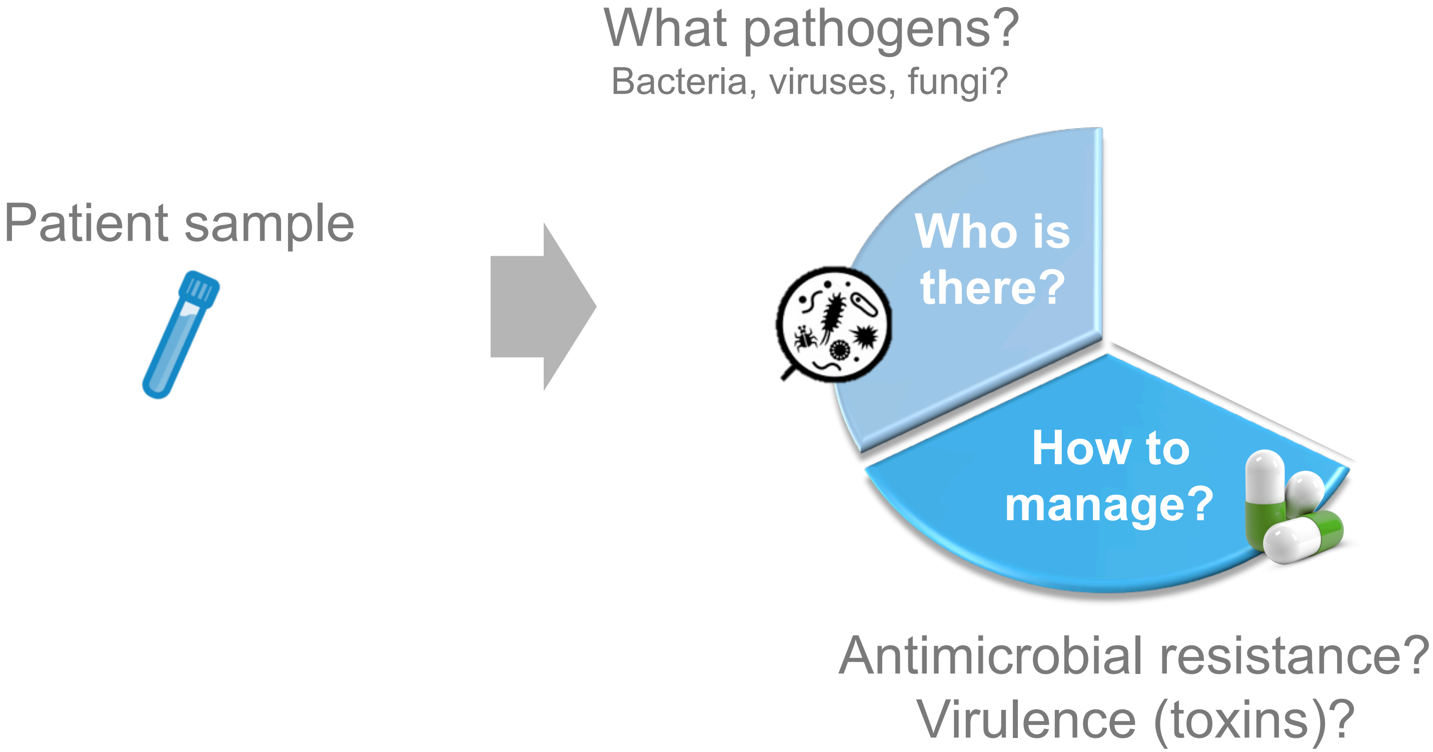
Patient sample

What pathogens?

Bacteria, viruses, fungi?

Who is there?

(Meta)genomics for infectious diseases



Patient sample

What pathogens?

Bacteria, viruses, fungi?

Antimicrobial resistance?

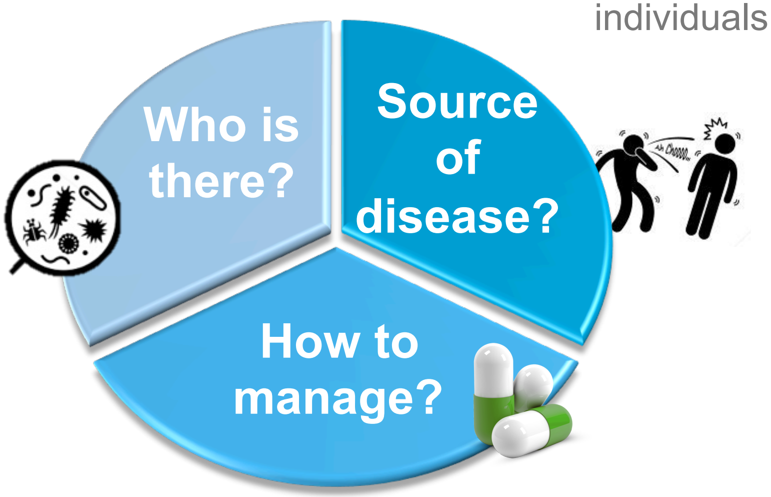
Virulence (toxins)?

(Meta)genomics for infectious diseases

Patient sample

What pathogens?
Bacteria, viruses, fungi?

Transmission routes?
Compare pathogens between individuals



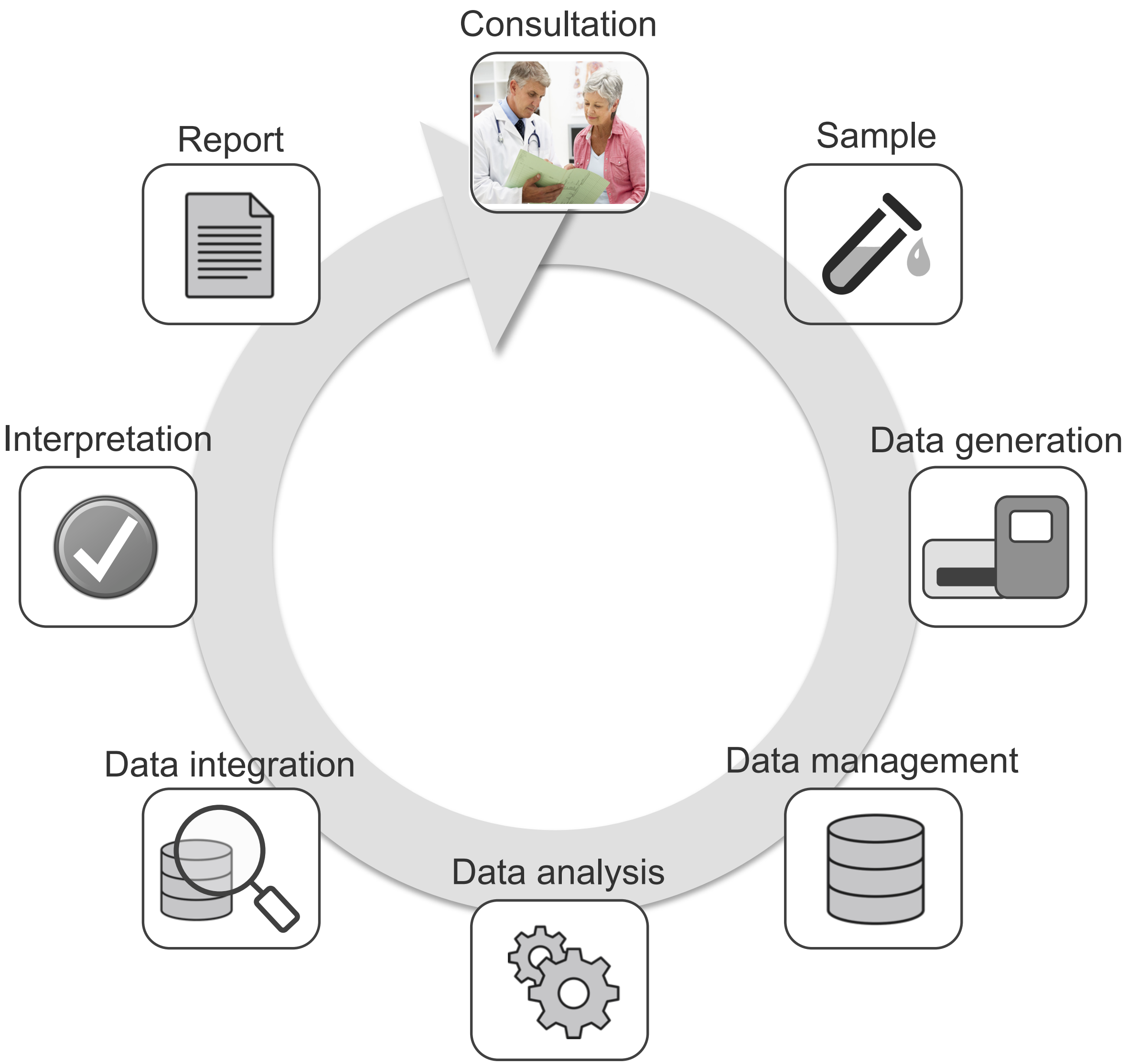
Who is there?

Source of disease?

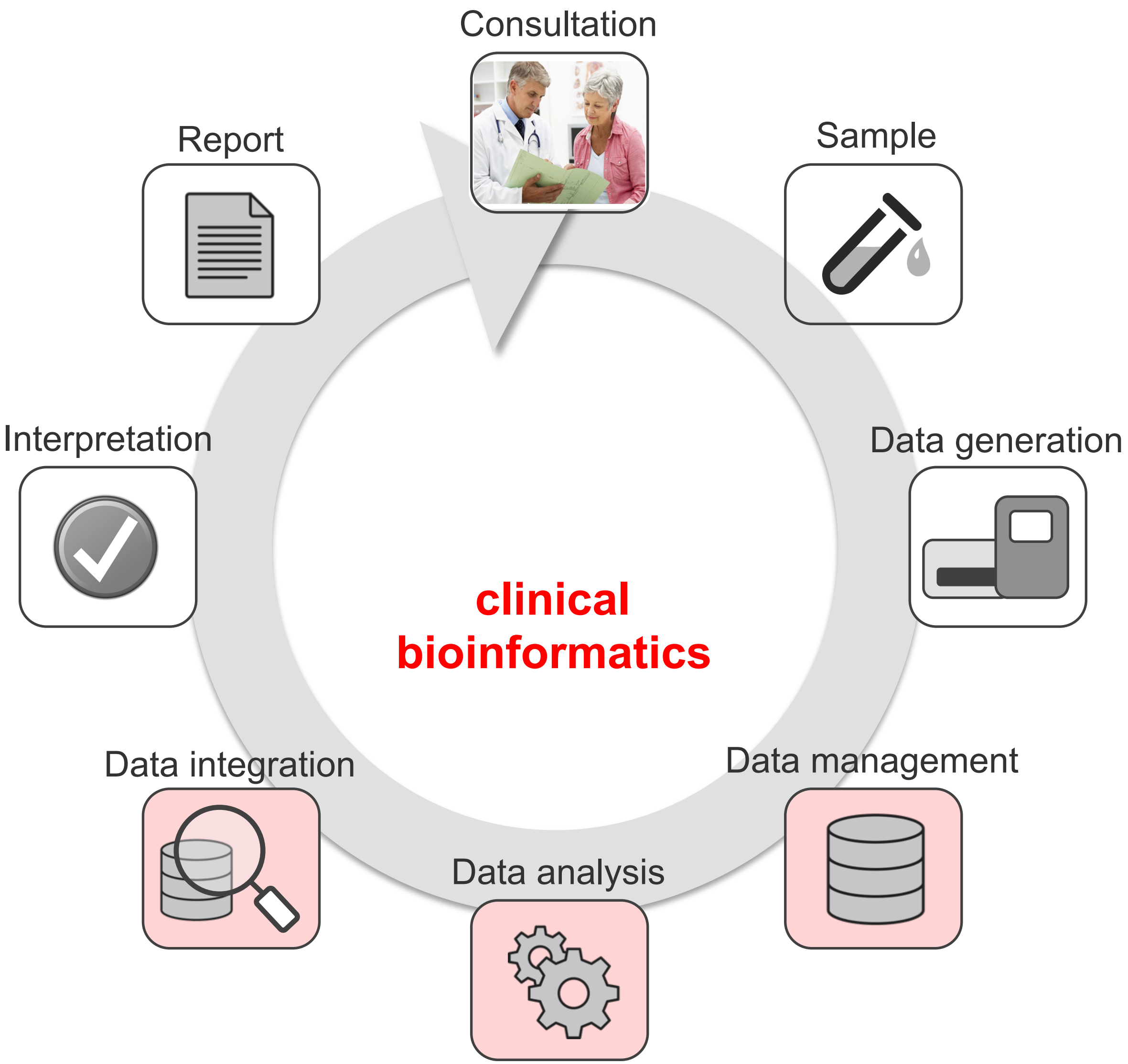
How to manage?

Antimicrobial resistance?
Virulence (toxins)?

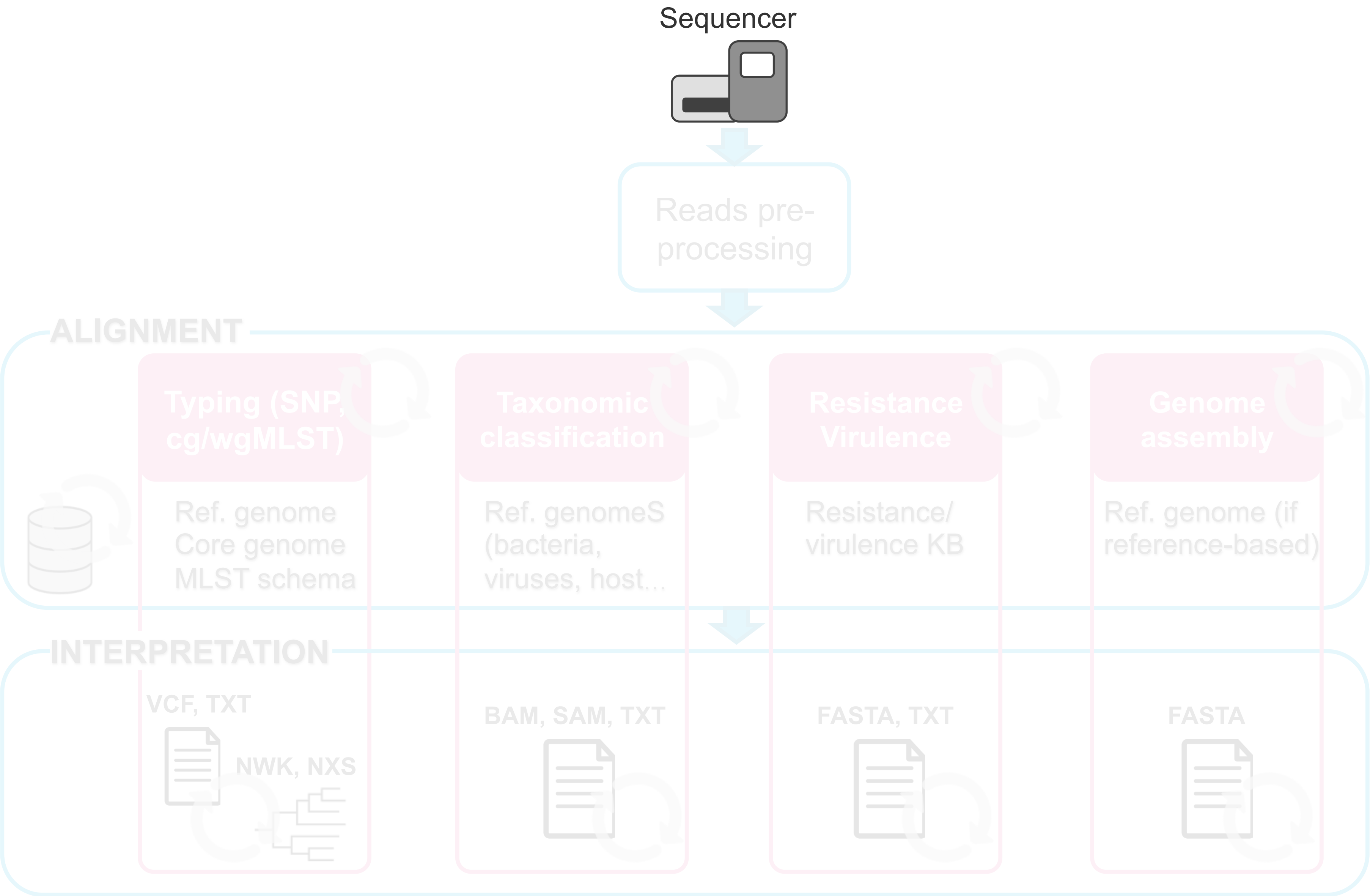
Clinical NGS pipeline



Clinical NGS pipeline



Overview of NGS bioinformatics pipelines



Sequencer

Reads pre-processing

ALIGNMENT

Typing (SNP, cg/wgMLST)

Ref. genome
Core genome
MLST schema

Taxonomic classification

Ref. genomeS (bacteria, viruses, host…

Resistance Virulence

Resistance/ virulence KB

Genome assembly

Ref. genome (if reference-based)

INTERPRETATION

VCF, TXT

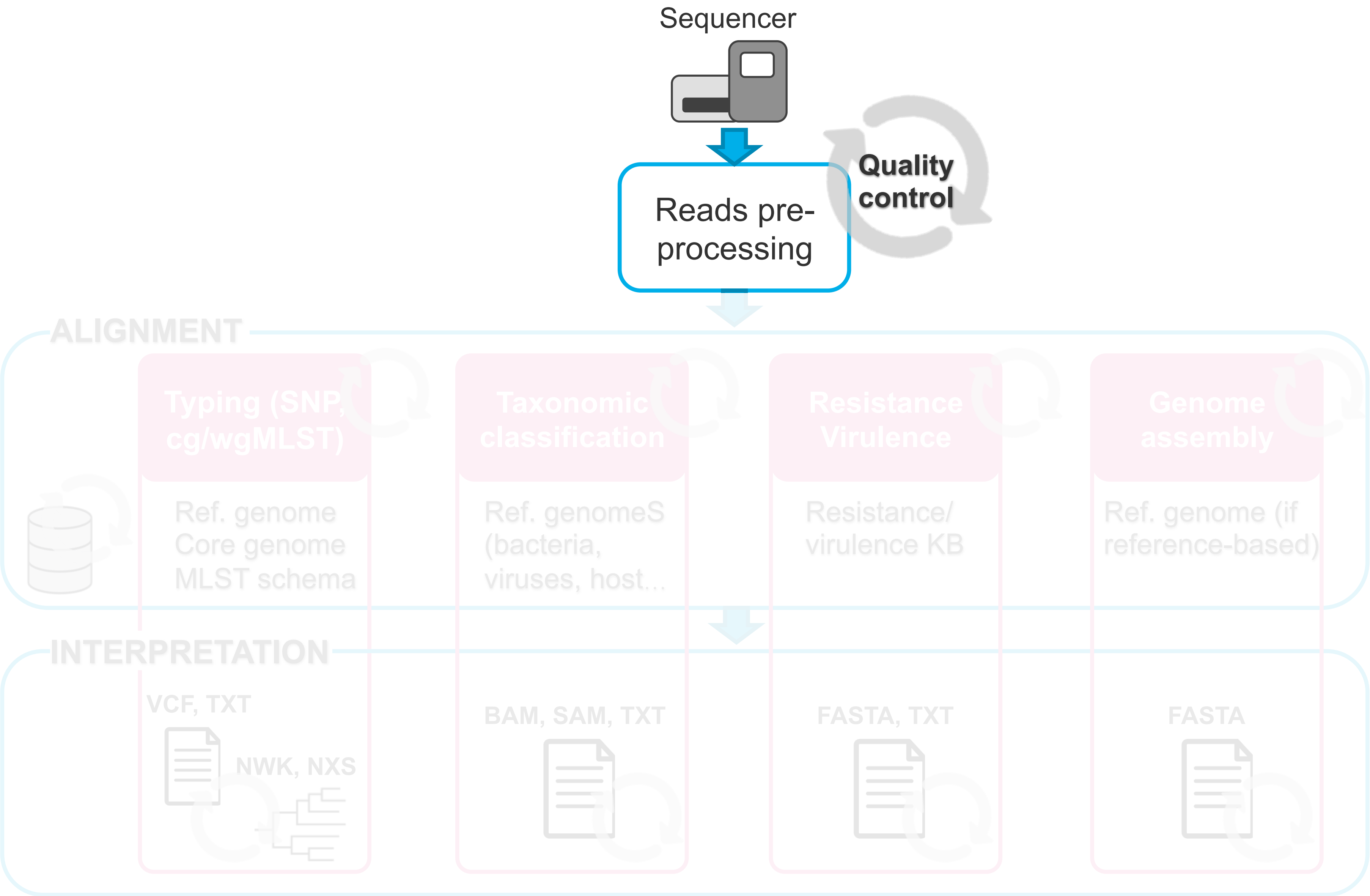
NWK, NXS

BAM, SAM, TXT

FASTA, TXT

FASTA

Overview of NGS bioinformatics pipelines



Sequencer

Reads pre-processing

Quality control

ALIGNMENT

Typing (SNP, cg/wgMLST)

Ref. genome
Core genome
MLST schema

Taxonomic classification

Ref. genomeS (bacteria, viruses, host…

Resistance Virulence

Resistance/ virulence KB

Genome assembly

Ref. genome (if reference-based)

INTERPRETATION

VCF, TXT

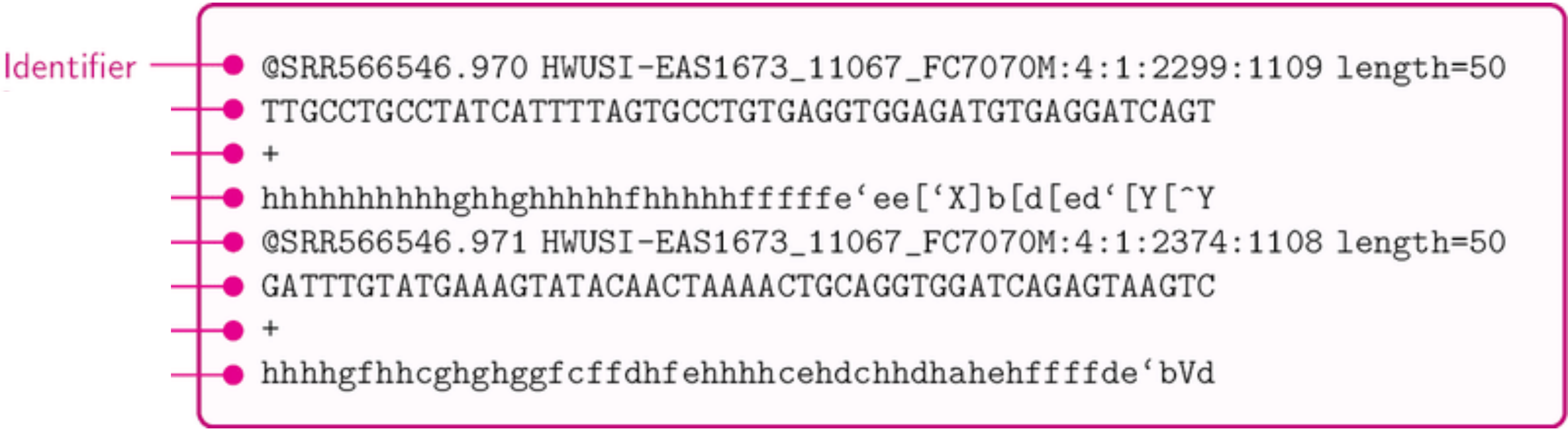
NWK, NXS

BAM, SAM, TXT

FASTA, TXT

FASTA

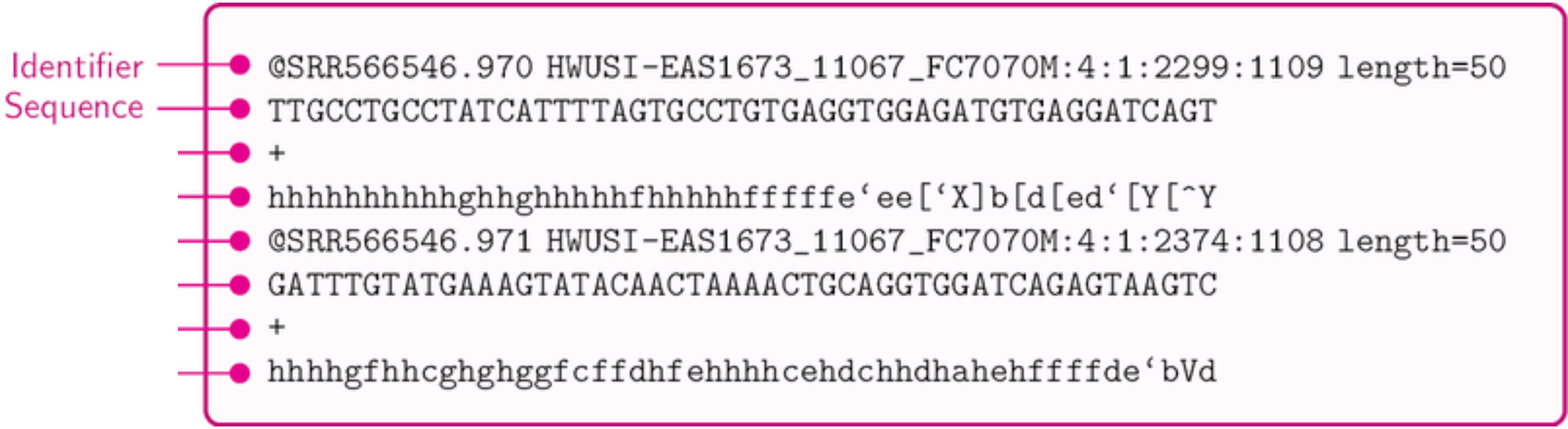
Out of the sequencer: FASTQ


Identifier

```
@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
+
hhhhhhhhhghhghhhhhfhhhhhfffffe‘ee[‘X]b[d[ed‘[Y[^Y
@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
+
hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde‘bVd
```

A Survey on Data Compression Methods for Biological Sequences. October 2016. Information (Switzerland) 7(4):56. DOI: 10.3390/info7040056

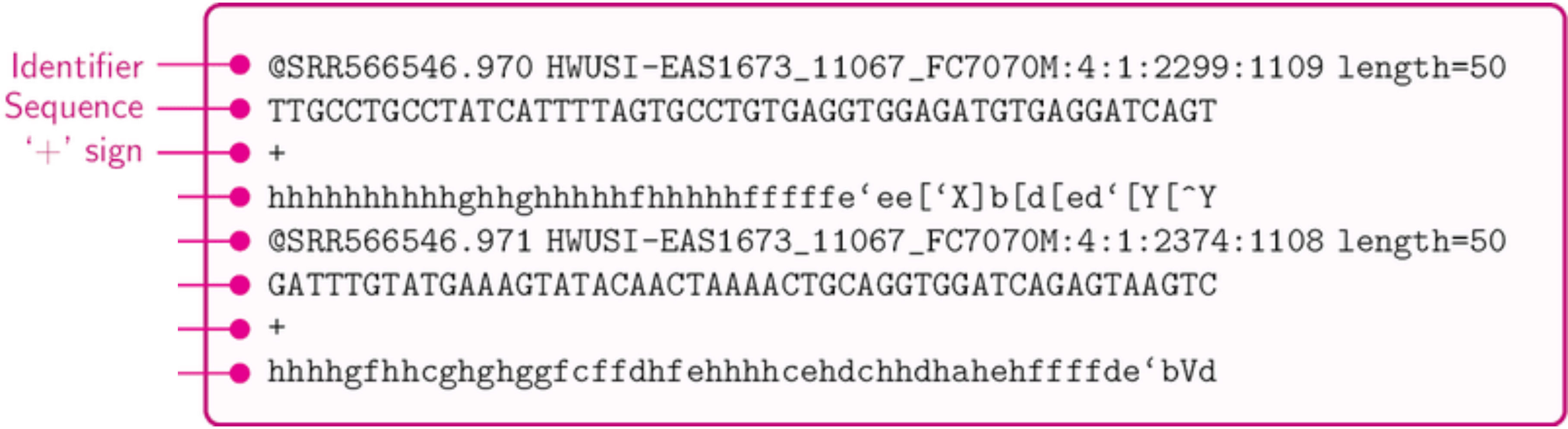
Out of the sequencer: FASTQ



```
@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
+
hhhhhhhhhghhghhhhhfhhhhhfffffe‘ee[‘X]b[d[ed‘[Y[^Y
@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
+
hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde‘bVd
```

A Survey on Data Compression Methods for Biological Sequences. October 2016. Information (Switzerland) 7(4):56. DOI: 10.3390/info7040056

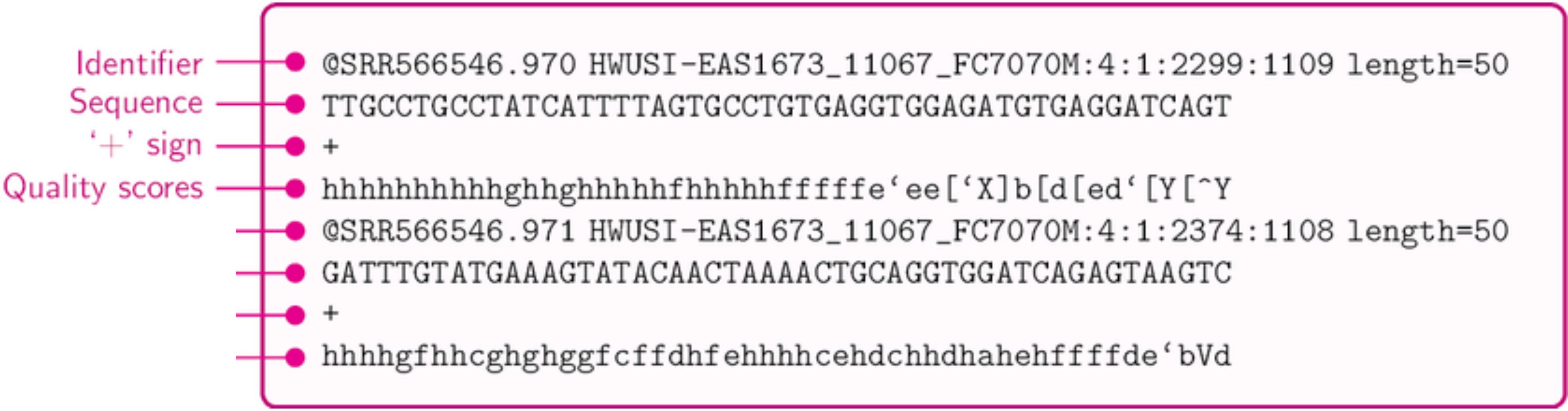
Out of the sequencer: FASTQ



```
@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
+
hhhhhhhhhhghhghhhhhfhhhhhfffffe`ee[`X]b[d[ed`[Y[^Y
@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
+
hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde`bVd
```

A Survey on Data Compression Methods for Biological Sequences. October 2016. Information (Switzerland) 7(4):56. DOI: 10.3390/info7040056

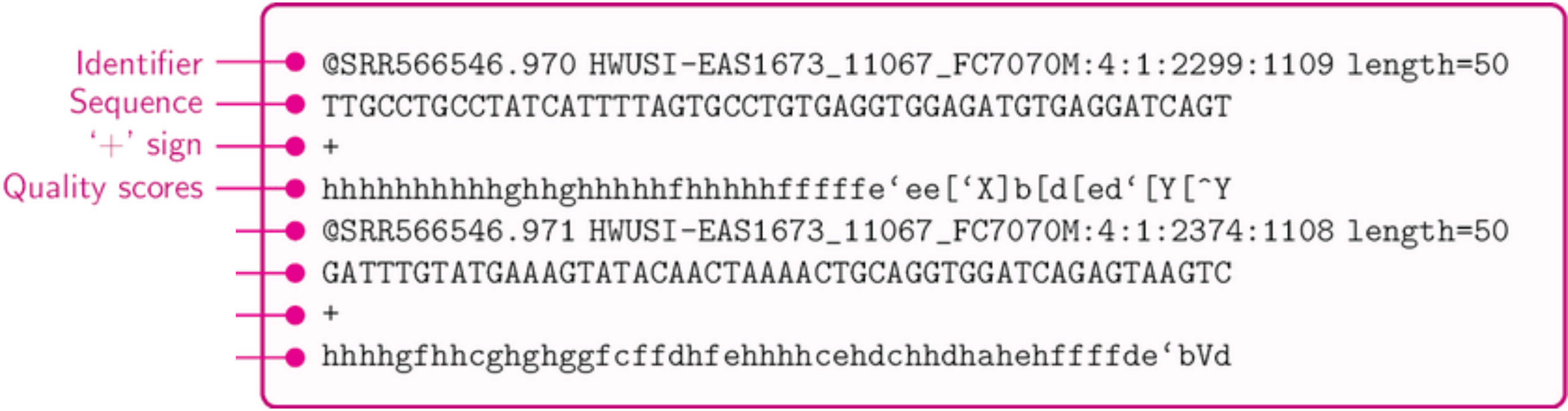
Out of the sequencer: FASTQ


Identifier
Sequence
'+' sign
Quality scores

```
@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
+
hhhhhhhhhhghhghhhhhfhhhhhfffffe‘ee[‘X]b[d[ed‘[Y[^Y
@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
+
hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde‘bVd
```

A Survey on Data Compression Methods for Biological Sequences. October 2016. Information (Switzerland) 7(4):56. DOI: 10.3390/info7040056

Out of the sequencer: FASTQ



```
@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
+
hhhhhhhhhghhghhhhhfhhhhhfffffe‘ee[‘X]b[d[ed‘[Y[^Y
@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
+
hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde‘bVd
```

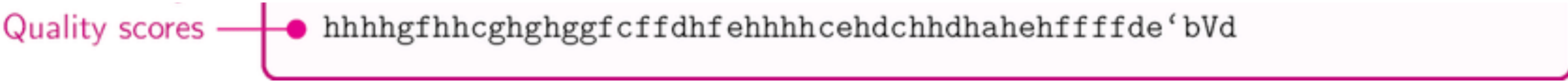
Each nucleotide has a **quality score** representing the probability that a base was miscalled by the sequencer

A Survey on Data Compression Methods for Biological Sequences. October 2016. Information (Switzerland) 7(4):56. DOI: 10.3390/info7040056

Quality scores: PHRED scores

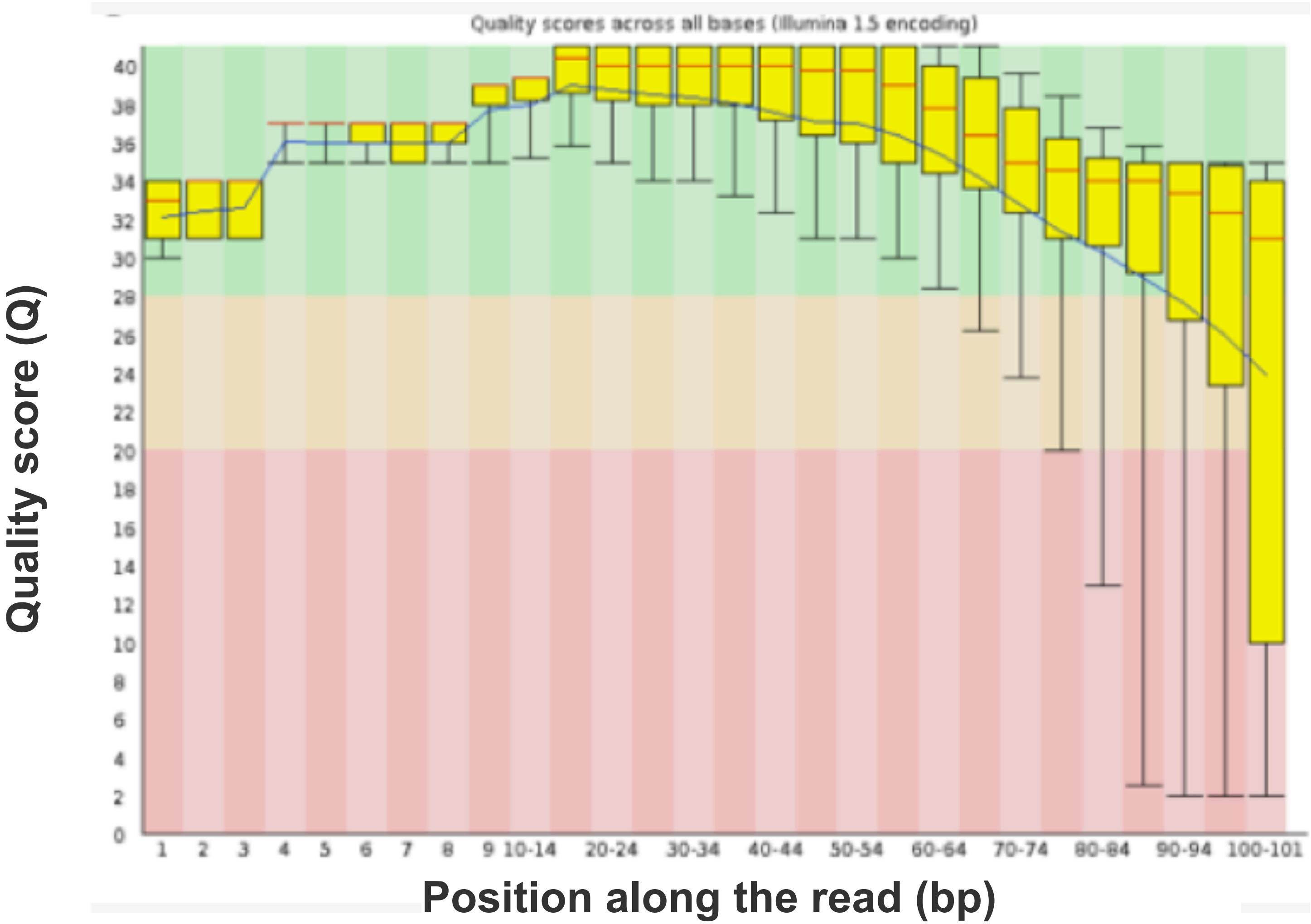
$$Q = -10\ \log_{10} P$$

Phred Quality Score	Prob. of incorrect base call	Base call accuracy	Code
10	1 in 10	90%	J
20	1 in 100	99%	T
30	1 in 1’000	99.9%	^
40	1 in 10’000	99.99%	h



```
hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde‘bVd
```

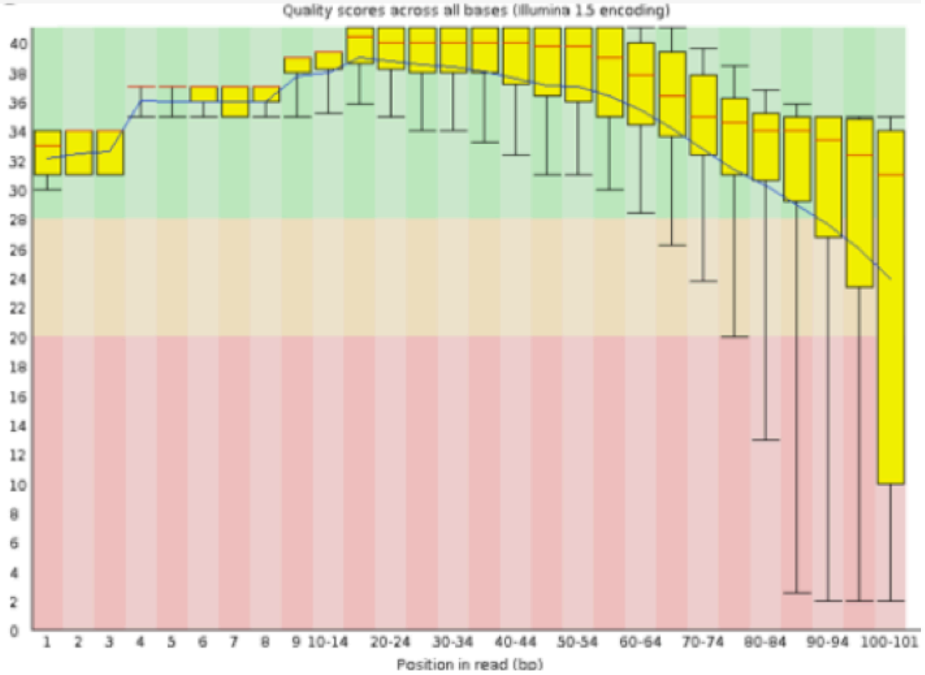
Quality-based reads trimming



CC BY-NC-SA SIB Clinical Bioinformatics 2019

Quality-based reads trimming

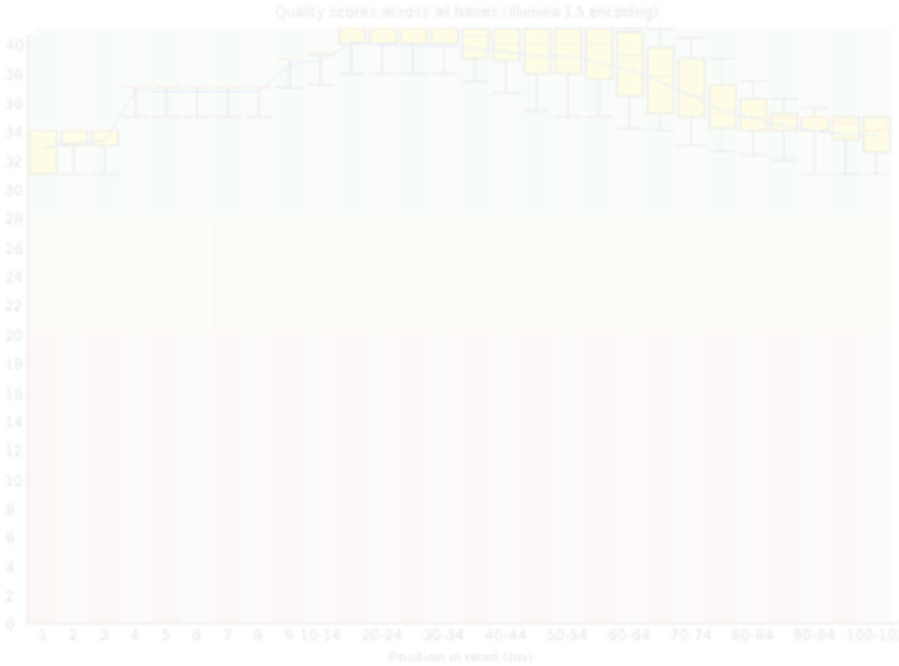
Before



Quality score (Q)

Position along the read (bp)

After

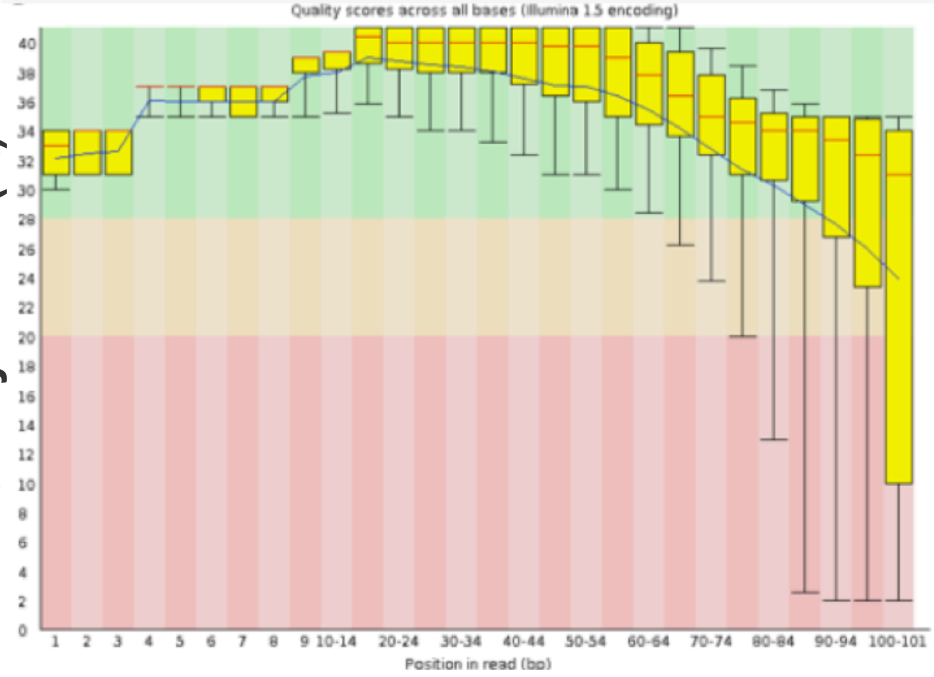


Position along the read (bp)

Quality-based reads trimming

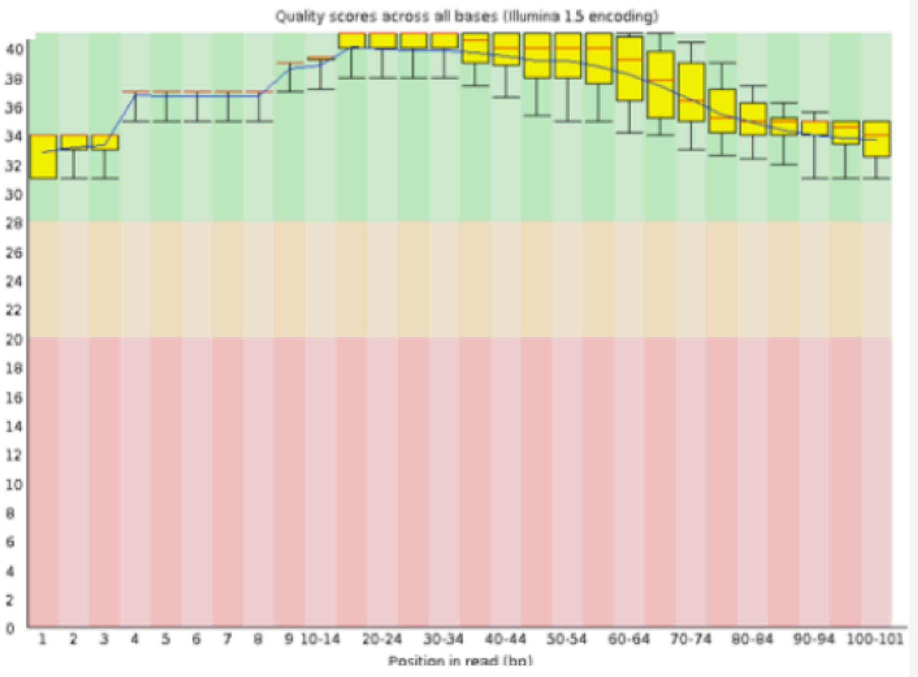
Before

Quality score (Q)



Position along the read (bp)

After

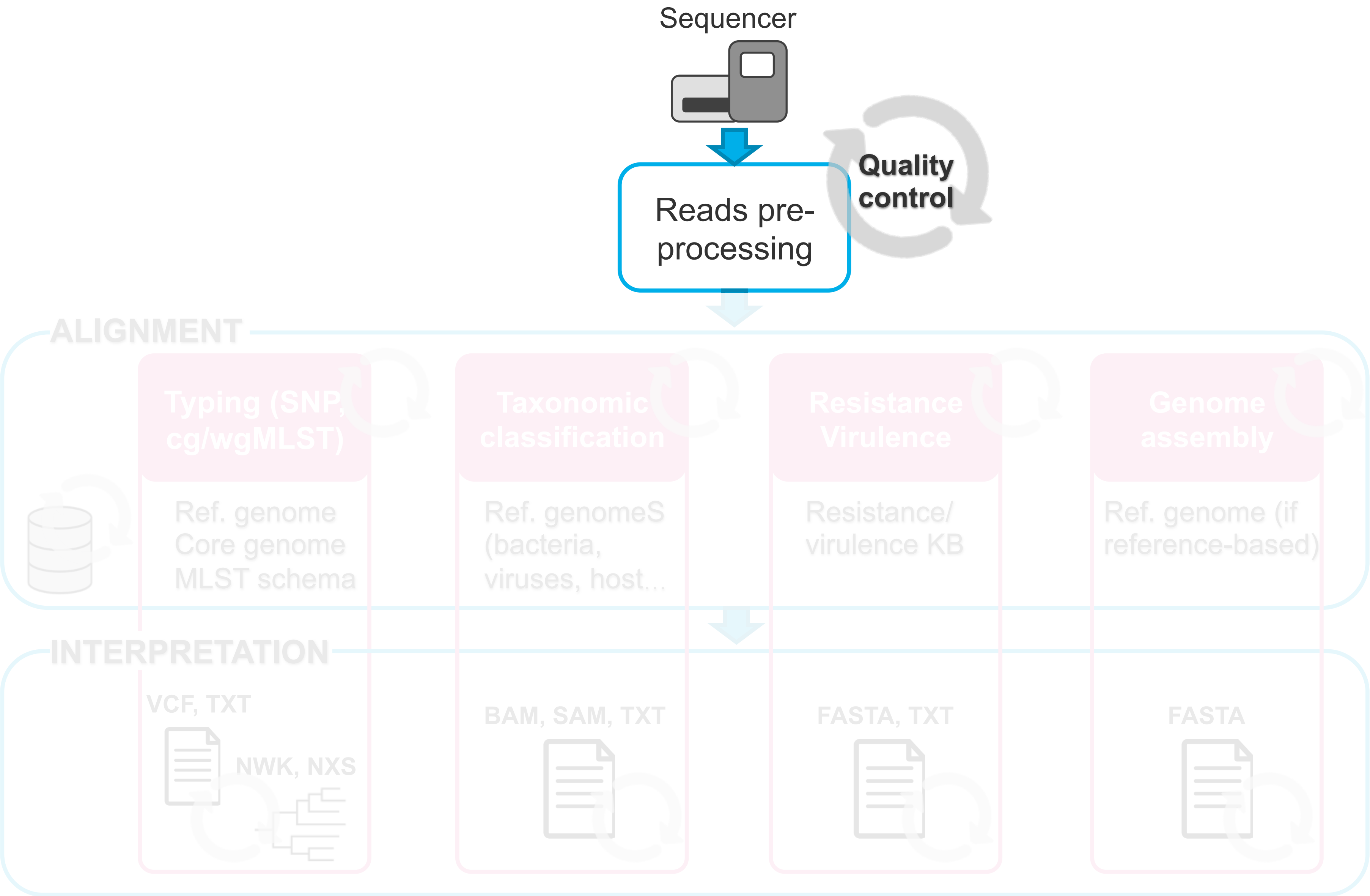


Position along the read (bp)

Remove adapters

- Adapter sequences should be removed from reads because they interfere with downstream analyses.
- The adapters contain the sequencing primer binding sites, the index sequences, and the sites that allow library fragments to attach to the flow cell lawn.

Overview of NGS bioinformatics pipelines



Sequencer

Reads pre-processing

Quality control

ALIGNMENT

Typing (SNP, cg/wgMLST)

Ref. genome
Core genome
MLST schema

Taxonomic classification

Ref. genomeS
(bacteria,
viruses, host…

Resistance Virulence

Resistance/
virulence KB

Genome assembly

Ref. genome (if
reference-based)

INTERPRETATION

VCF, TXT

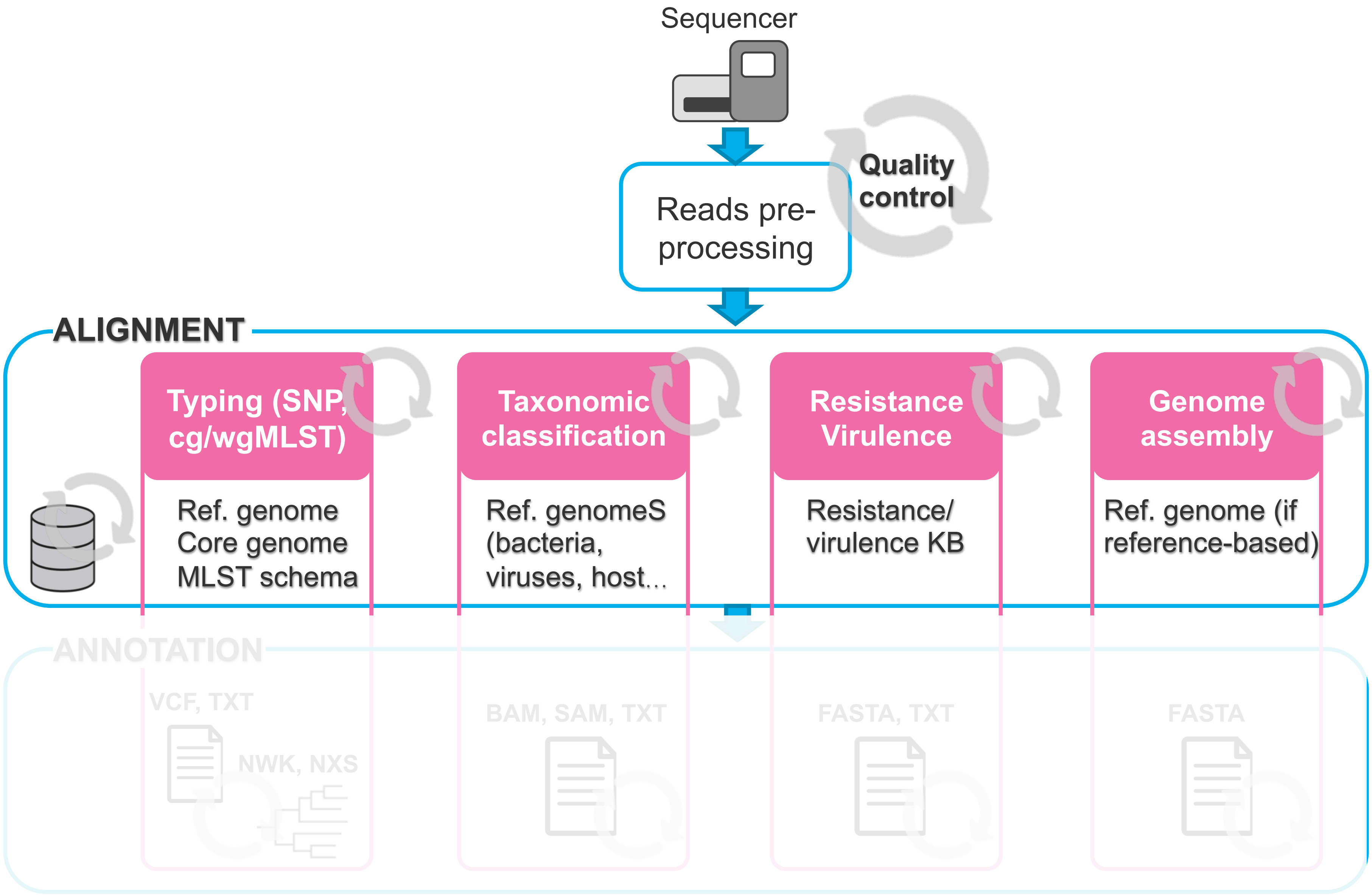
NWK, NXS

BAM, SAM, TXT

FASTA, TXT

FASTA

Overview of NGS bioinformatics pipelines



Sequencer

Reads pre-processing

Quality control

ALIGNMENT

Typing (SNP, cg/wgMLST)

Ref. genome
Core genome
MLST schema

Taxonomic classification

Ref. genomeS (bacteria, viruses, host…

Resistance Virulence

Resistance/ virulence KB

Genome assembly

Ref. genome (if reference-based)

ANNOTATION

VCF, TXT

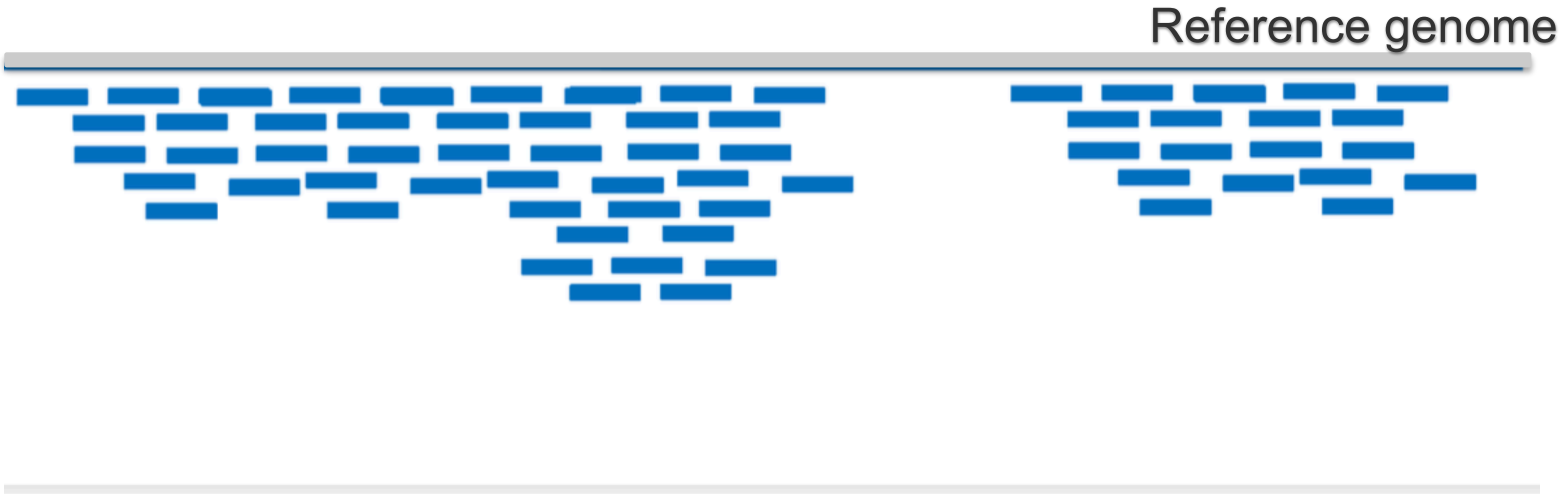
NWK, NXS

BAM, SAM, TXT

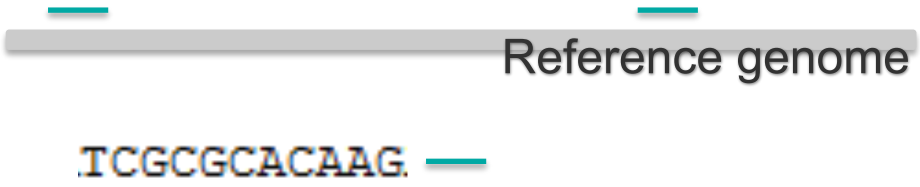
FASTA, TXT

FASTA

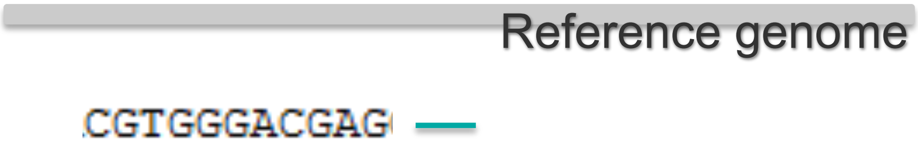
Now that we have clean reads, let’s align them!



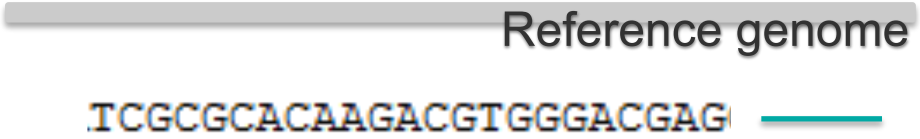
Alignment: a complex “simple” problem



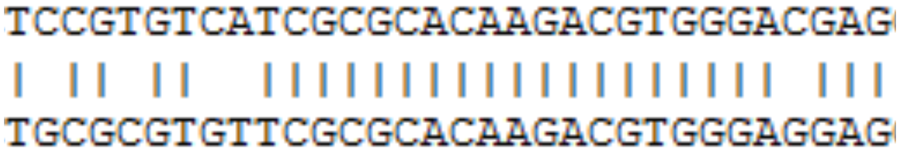
! Short reads are likely to map at several positions along the reference genome



! Mismatches and **gaps** allowed
→ algorithms have scoring functions



! Longer reads are less ambiguous
→ but computationally more expensive

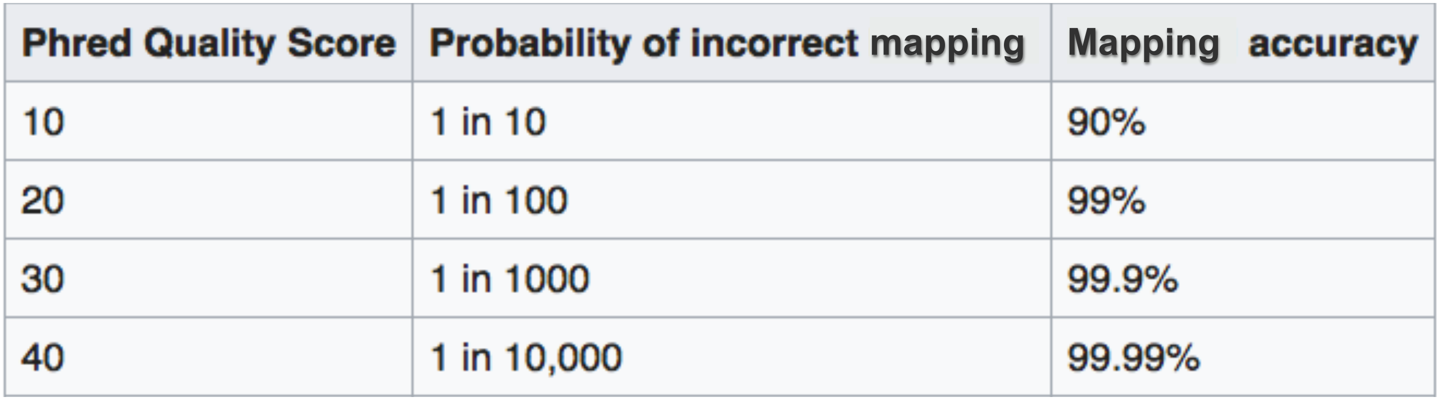


```
TCCGTGTCATCGCGCACAAGACGTGGACGAG
| || ||  |||||||||||||||||| |||
TGCGCGTGTTCGCGCACAAGACGTGGAGGAG
```

Alignment score and mapping quality score

- **Alignment score (AS)**
  - Generated by the aligner.
  - Reflects how many mismatches and gaps you need to align the read at a particular position.

- **Mapping quality score (MAPQ)**
  - Reflects the probability that the read was wrongly mapped, i.e. not aligned where it should.
  - Usually reported on a PHRED scale.



Phred Quality Score	Probability of incorrect mapping	Mapping accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%

Alignment quizz

We have a read with the following scores… what does it mean?

ALIGN SCORE	MAP SCORE	Conclusion
👍	👍	
👍	👎	
👎	👍	
👎	👎	

Alignment quizz

We have a read with the following scores… what does it mean?

ALIGN SCORE	MAP SCORE	Conclusion
👍	👍	Read maps unambiguously and is very similar to reference sequence at that position
👍	👎	
👎	👍	
👎	👎	

Alignment quizz

We have a read with the following scores… what does it mean?

ALIGN SCORE	MAP SCORE	Conclusion
👍	👍	Read maps unambiguously and is very similar to reference sequence at that position
👍	👎	Read is very similar to reference sequence at that position, but maps at several positions
👎	👍	
👎	👎	

Alignment quizz

We have a read with the following scores… what does it mean?

ALIGN SCORE	MAP SCORE	Conclusion
👍	👍	Read maps unambiguously and is very similar to reference sequence at that position
👍	👎	Read is very similar to reference sequence at that position, but maps at several positions
👎	👍	Read maps unambiguously, but aligns with several mismatches to reference sequence at that position
👎	👎	

Alignment quizz

We have a read with the following scores… what does it mean?

ALIGN SCORE	MAP SCORE	Conclusion
👍	👍	Read maps unambiguously and is very similar to reference sequence at that position
👍	👎	Read is very similar to reference sequence at that position, but maps at several positions
👎	👍	Read maps unambiguously, but aligns with several mismatches to reference sequence at that position
👎	👎	Reads aligns with several mismatches at this and several other positions on the ref. seq.

Alignment quizz

How relevant are these cases for clinical use?

ALIGN SCORE	MAP SCORE	Conclusion
👍	👍	Read maps unambiguously and is very similar to reference sequence at that position
~~👍~~	~~👎~~	~~Read is very similar to reference sequence at that position, but maps at several positions~~
👎	👍	Read maps unambiguously, but aligns with several mismatches to reference sequence at that position
~~👎~~	~~👎~~	~~Reads aligns with several mismatches at this and several other positions on the ref. seq.~~

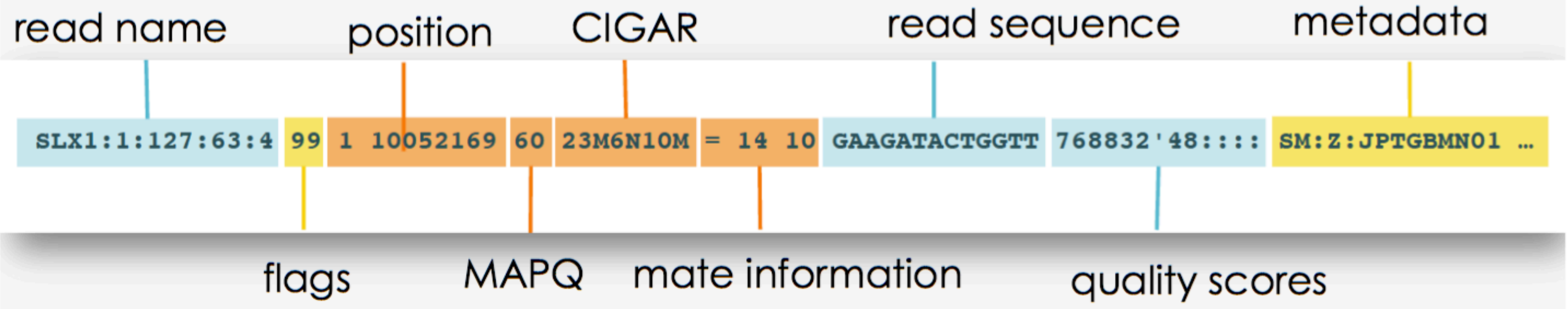
Out of the mapper: SAM - BAM

Header

@SQ Reference Sequence: SN name, LN length

@RG Read Group: e.g. grouping samples

Records



read name

position

CIGAR

read sequence

metadata

```
SLX1:1:127:63:4 99 1 10052169 60 23M6N10M = 14 10 GAAGATACTGGTT 768832'48:::: SM:Z:JPTGBMN01 …
```

flags

MAPQ

mate information

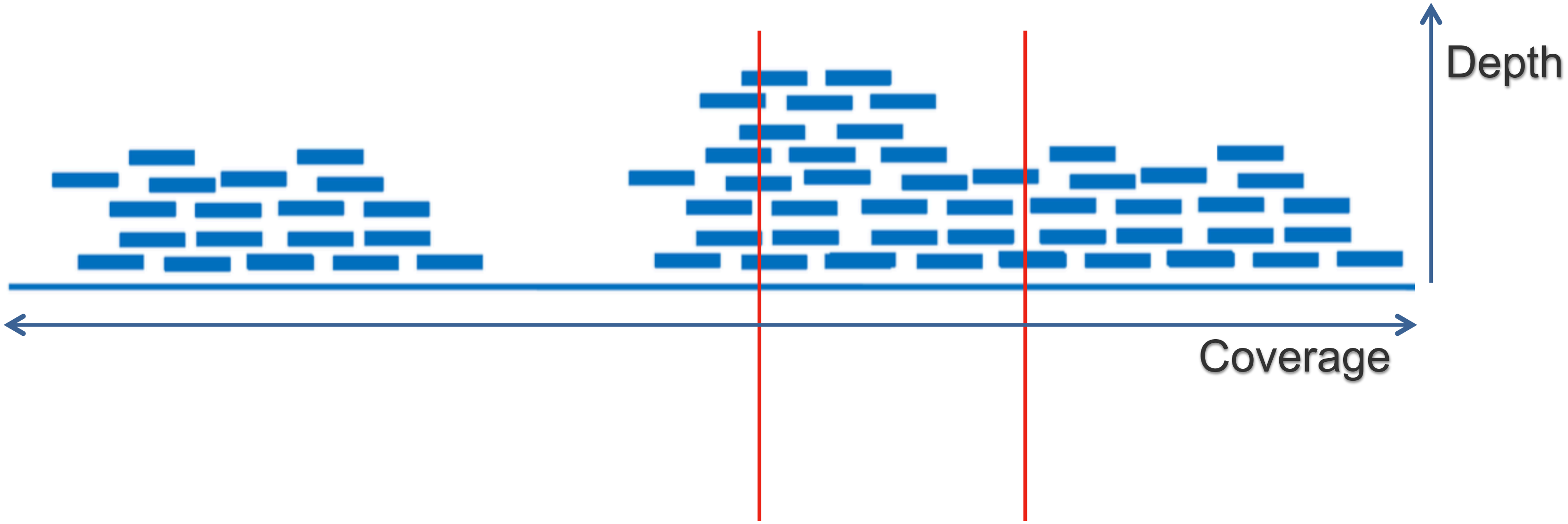
quality scores

BAM is the binary version of the SAM file (i.e. compressed, human non-readable).

Depth and coverage

Depth = number of reads that include a given nucleotide, e.g. 1000X at a given position.

Coverage = percentage or number of bases of a reference genome that are covered with a certain depth, e.g. 90% at 5X

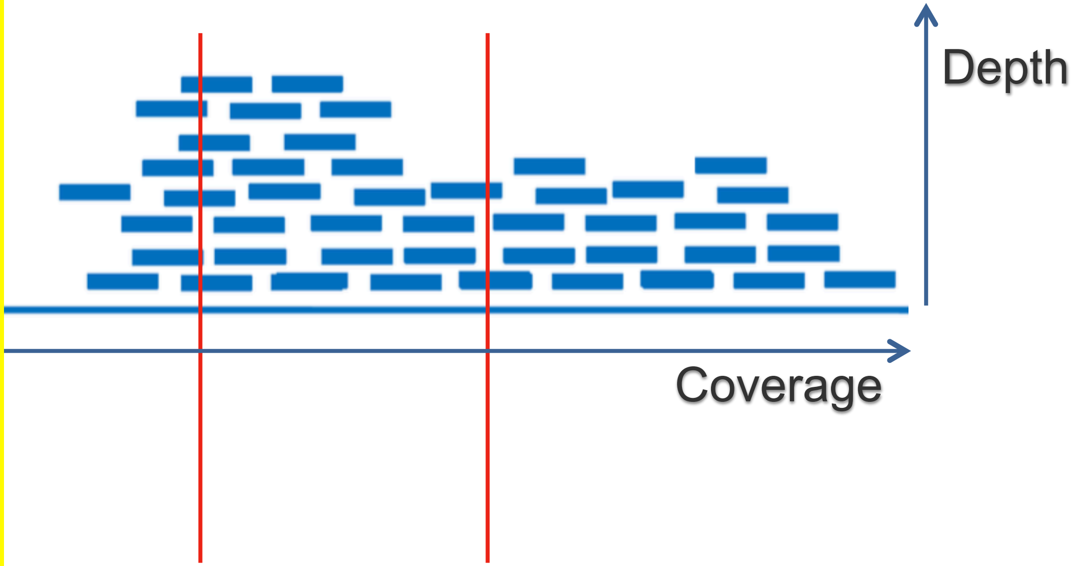


Depth and coverage

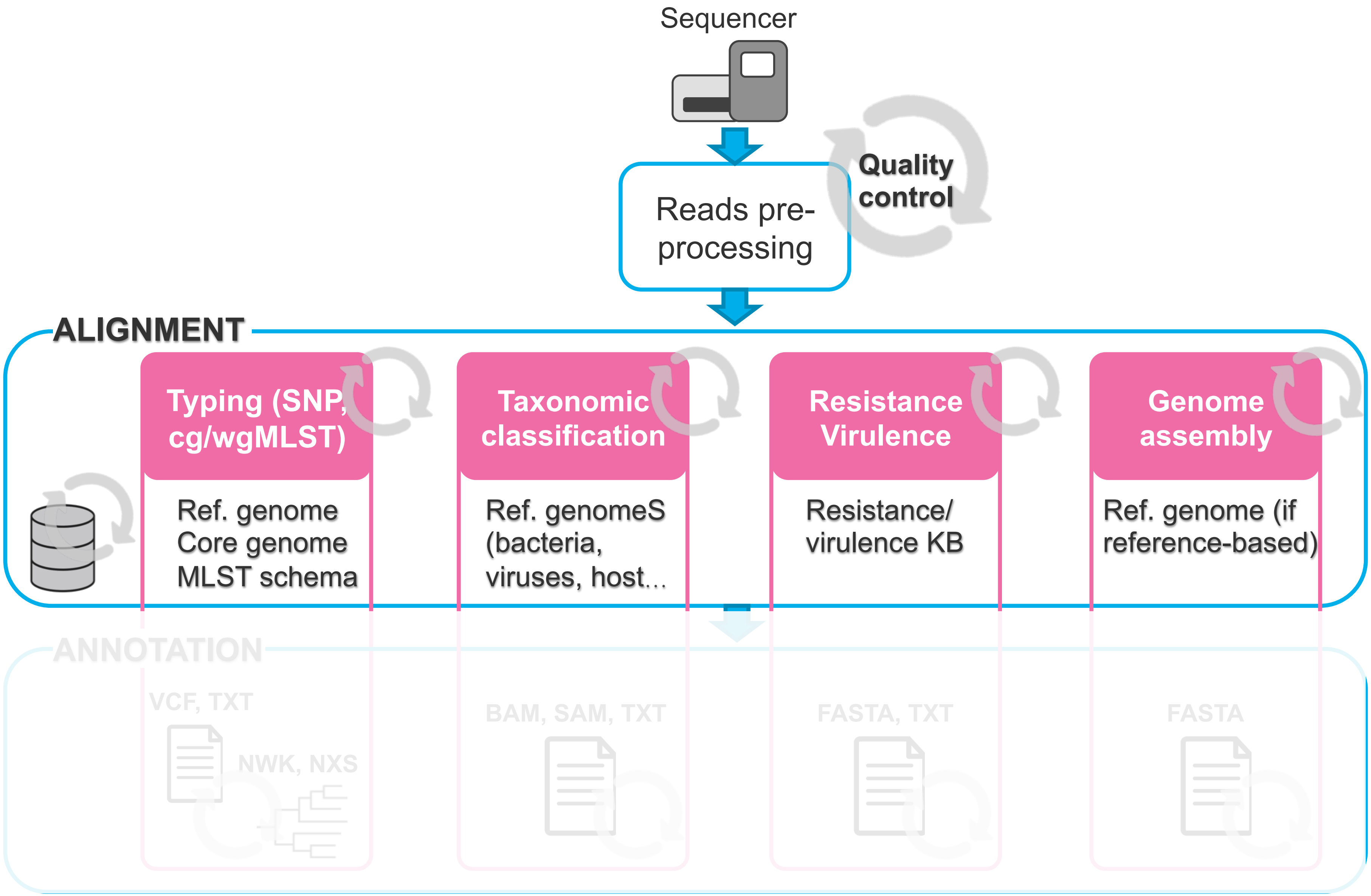
Depth = number of reads that include a given nucleotide, e.g. 1000X at a given position.

Coverage = percentage or number of bases of a reference genome that are covered with a certain depth, e.g. 90% at 5X

Many people use "coverage" for "depth". Watch out if % or X



Overview of NGS bioinformatics pipelines



Sequencer

Reads pre-processing

Quality control

ALIGNMENT

Typing (SNP, cg/wgMLST)

Ref. genome
Core genome
MLST schema

Taxonomic classification

Ref. genomeS (bacteria, viruses, host…

Resistance Virulence

Resistance/ virulence KB

Genome assembly

Ref. genome (if reference-based)

ANNOTATION

VCF, TXT

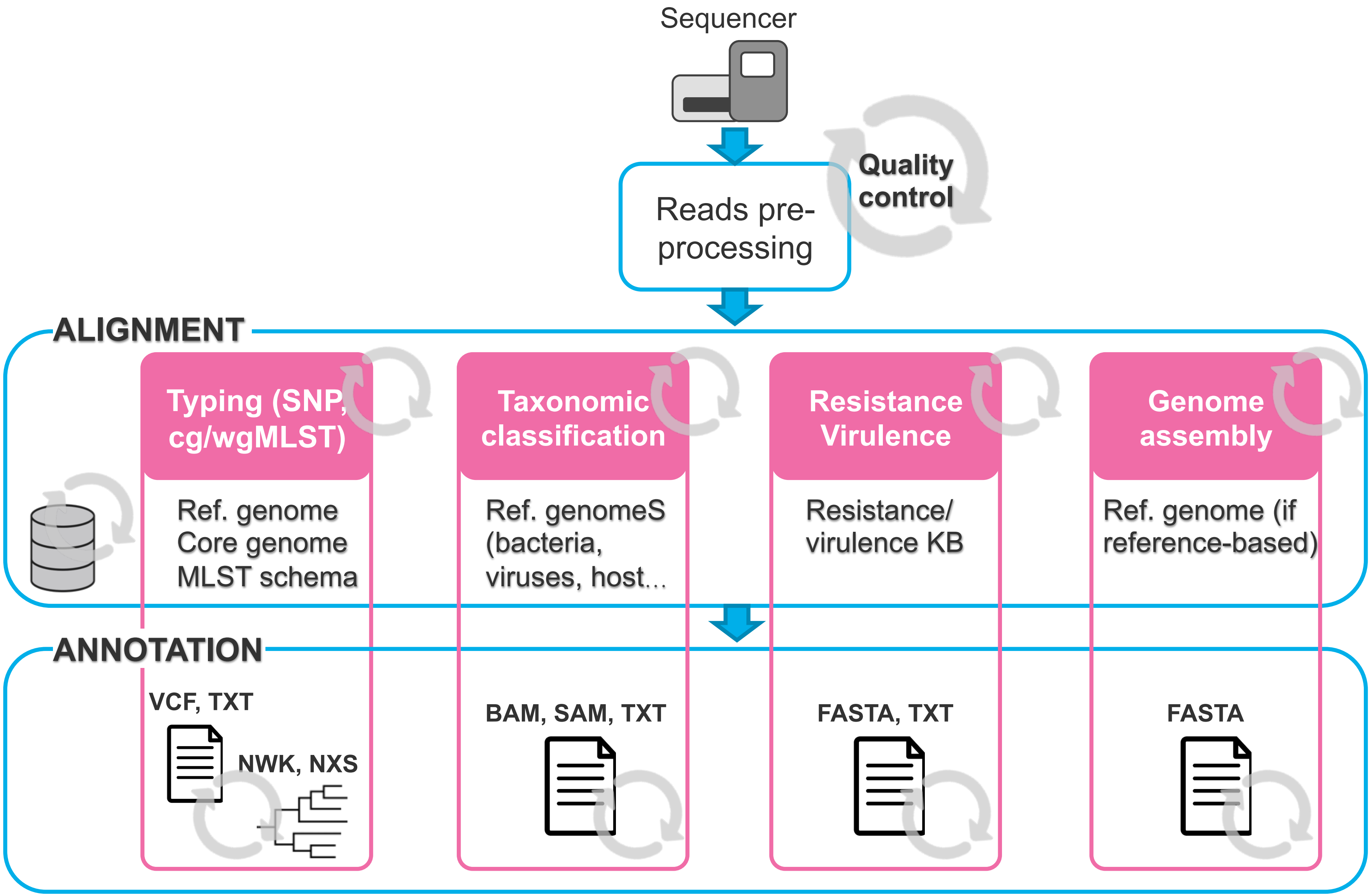
NWK, NXS

BAM, SAM, TXT

FASTA, TXT

FASTA

Overview of NGS bioinformatics pipelines



Sequencer

Reads pre-processing

Quality control

ALIGNMENT

Typing (SNP, cg/wgMLST)

Ref. genome
Core genome
MLST schema

Taxonomic classification

Ref. genomeS (bacteria, viruses, host…

Resistance Virulence

Resistance/ virulence KB

Genome assembly

Ref. genome (if reference-based)

ANNOTATION

VCF, TXT

NWK, NXS

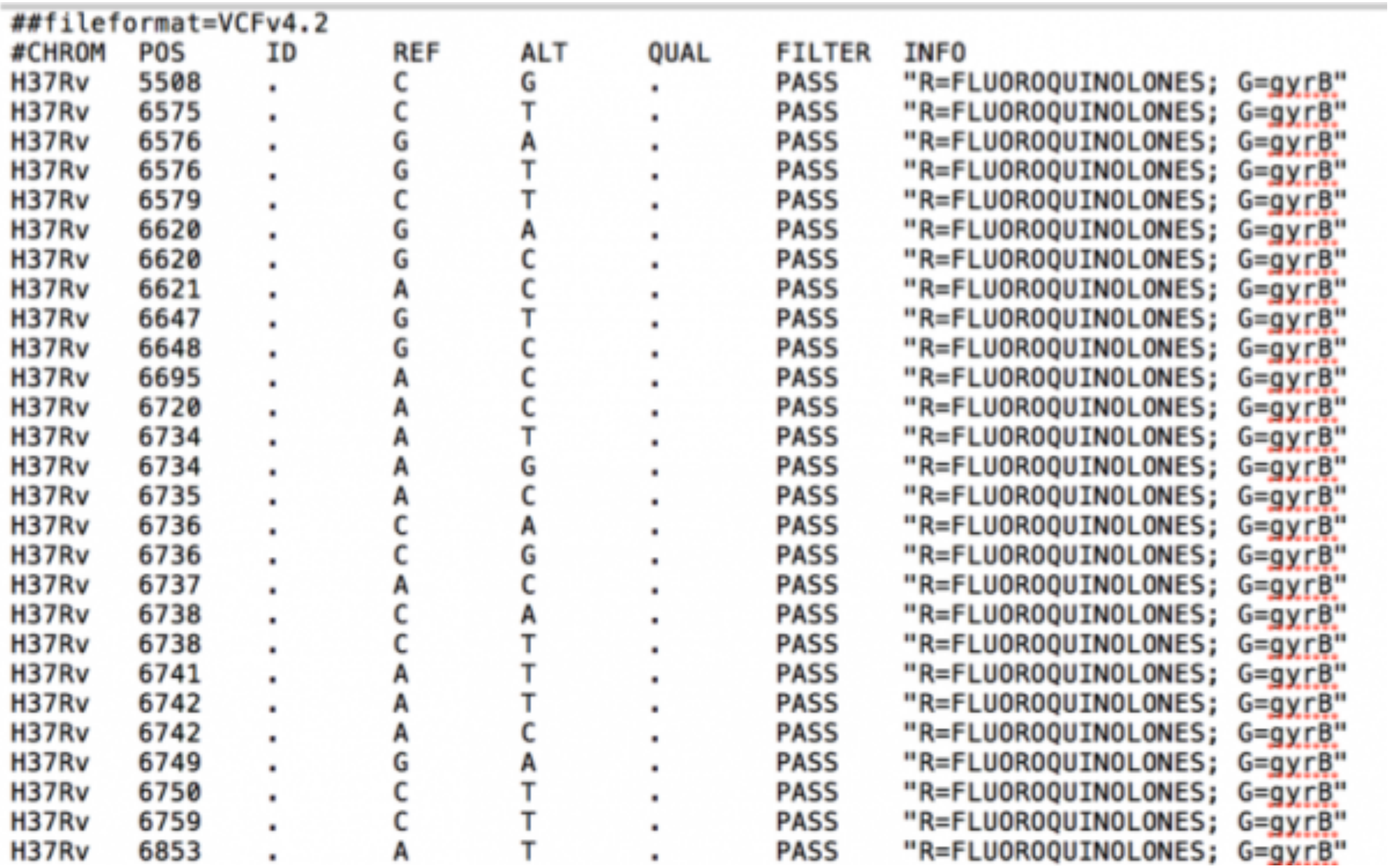
BAM, SAM, TXT

FASTA, TXT

FASTA

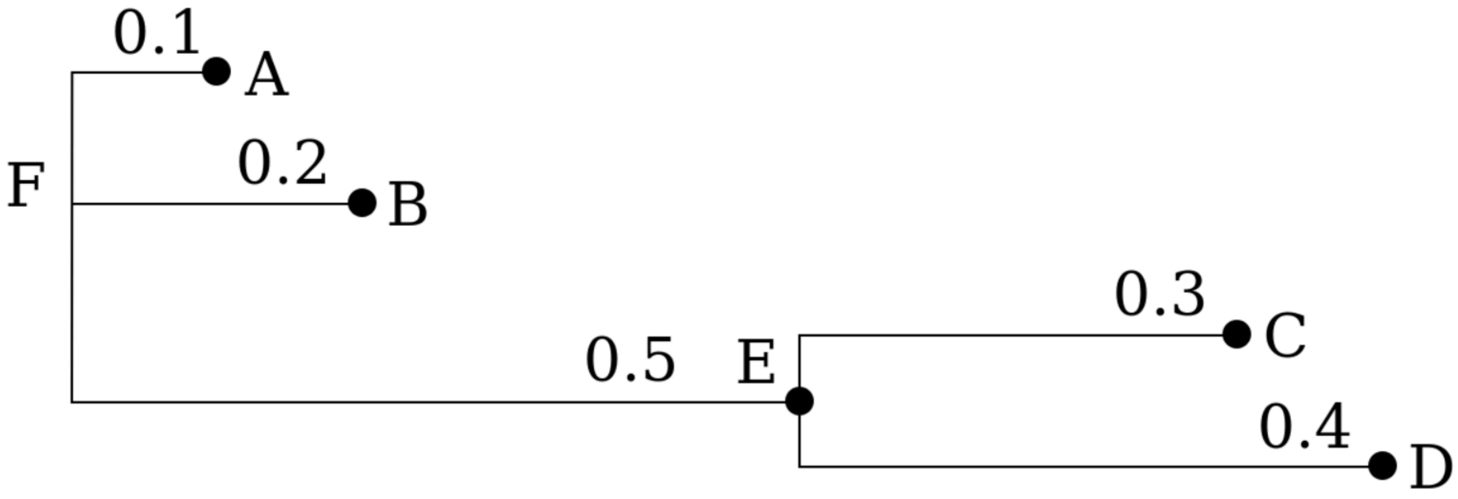
Standard formats are important in bioinformatics for automating parsing and analyses

VCF file format for variants (e.g. SNPs)

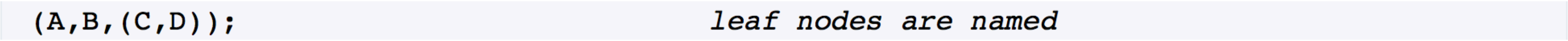


```
##fileformat=VCFv4.2
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO
H37Rv   5508    .       C       G       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6575    .       C       T       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6576    .       G       A       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6576    .       G       T       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6579    .       C       T       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6620    .       G       A       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6620    .       G       C       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6621    .       A       C       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6647    .       G       T       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6648    .       G       C       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6695    .       A       C       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6720    .       A       C       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6734    .       A       T       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6734    .       A       G       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6735    .       A       C       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6736    .       C       A       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6736    .       C       G       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6737    .       A       C       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6738    .       C       A       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6738    .       C       T       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6741    .       A       T       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6742    .       A       T       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6742    .       A       C       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6749    .       G       A       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6750    .       C       T       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6759    .       C       T       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
H37Rv   6853    .       A       T       .       PASS    "R=FLUOROQUINOLONES; G=gyrB"
```

NWK file format for trees



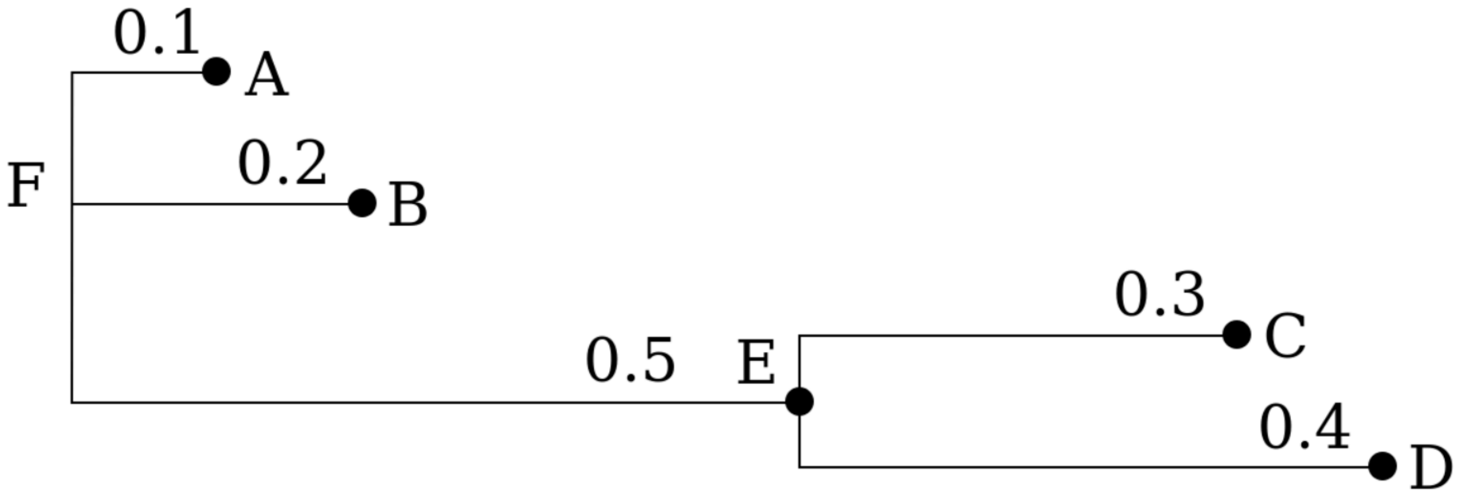
could be represented in Newick format in several ways



```
(A,B,(C,D));                                 leaf nodes are named
```

https://en.wikipedia.org/wiki/Newick_format

NWK file format for trees



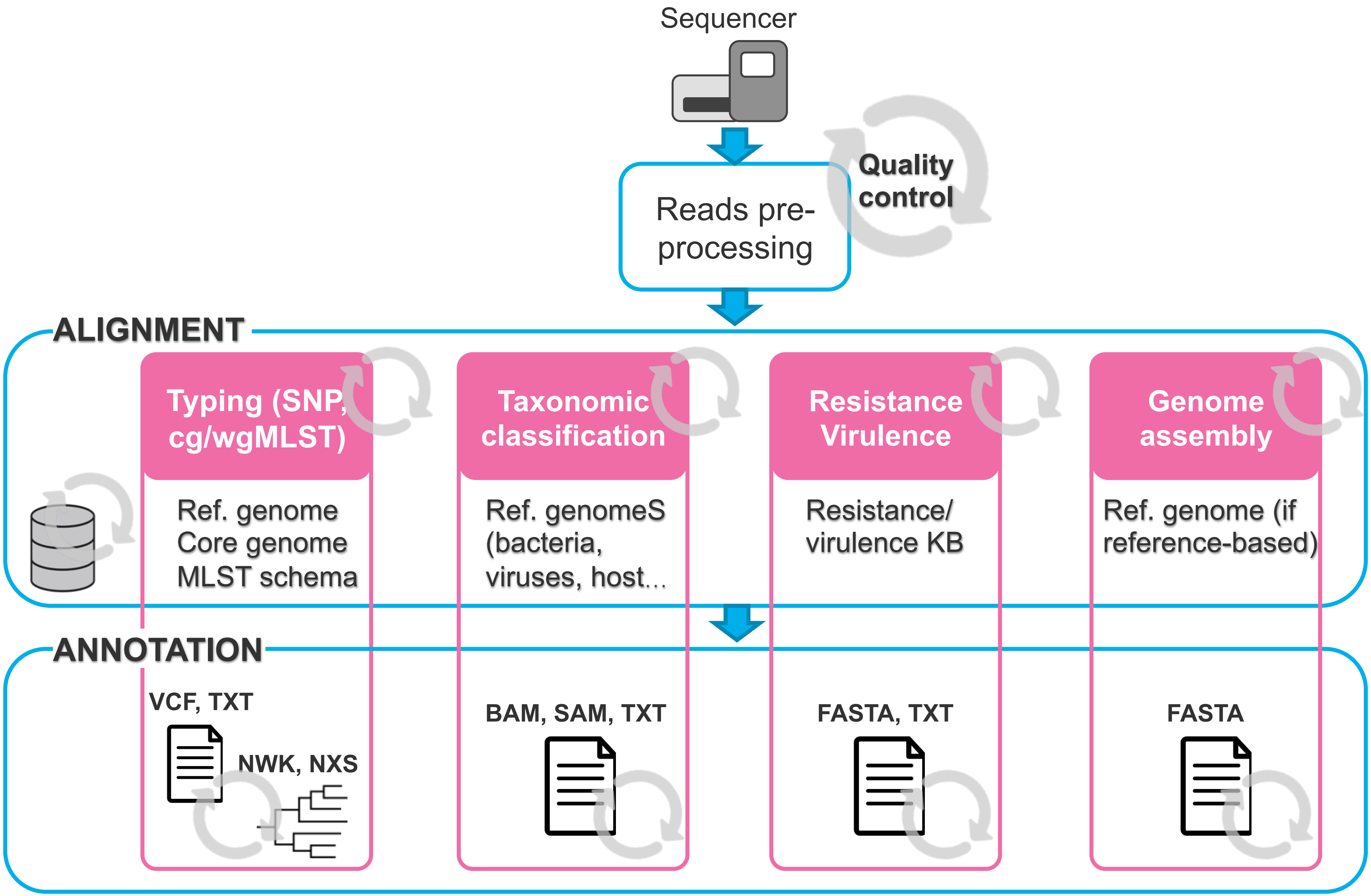
could be represented in Newick format in several ways

`(A,B,(C,D));` *leaf nodes are named*

`(A:0.1,B:0.2,(C:0.3,D:0.4):0.5);` *distances and leaf names* **(popular)**

https://en.wikipedia.org/wiki/Newick_format

Overview of NGS bioinformatics pipelines



Sequencer

Reads pre-processing

Quality control

ALIGNMENT

Typing (SNP, cg/wgMLST)

Ref. genome
Core genome
MLST schema

Taxonomic classification

Ref. genomeS
(bacteria,
viruses, host…

Resistance Virulence

Resistance/
virulence KB

Genome assembly

Ref. genome (if
reference-based)

ANNOTATION

VCF, TXT

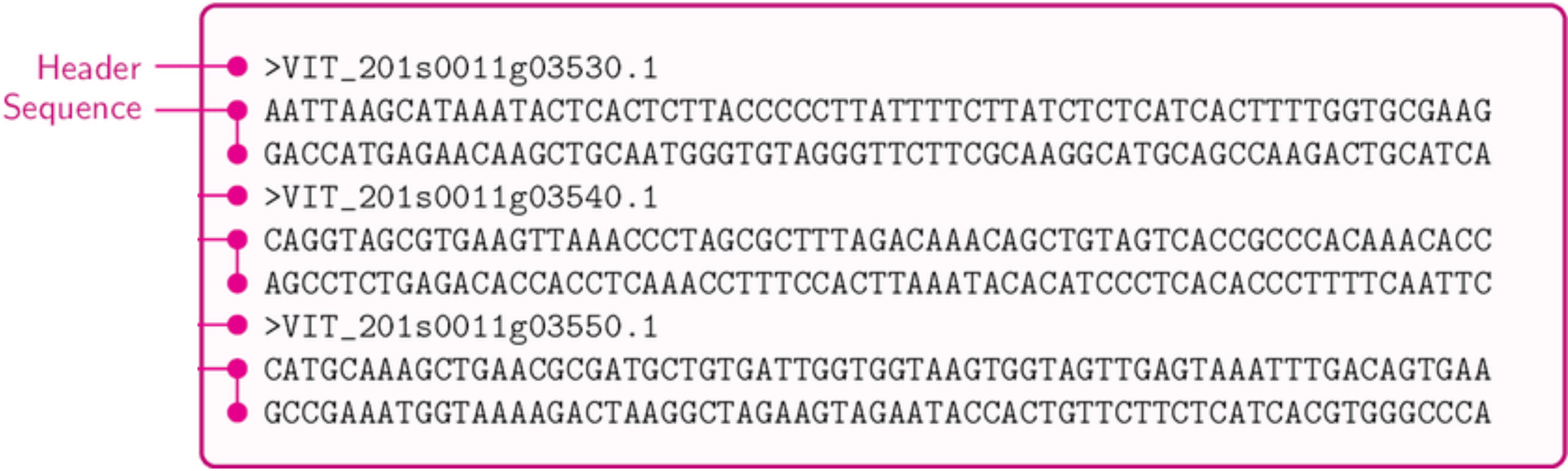
NWK, NXS

BAM, SAM, TXT

FASTA, TXT

FASTA

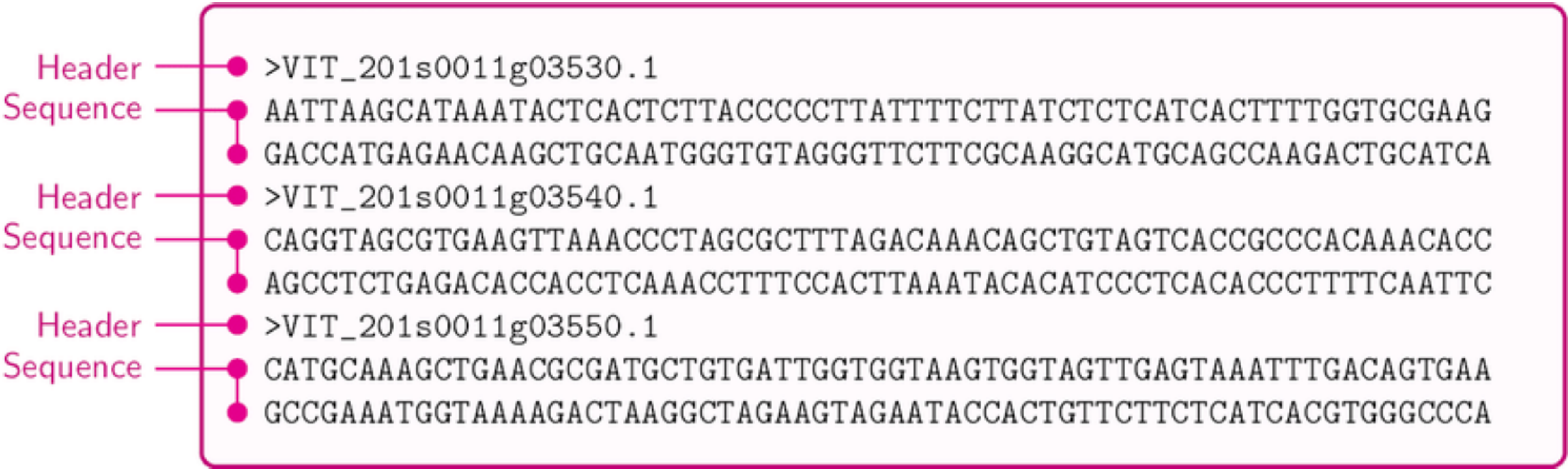
FASTA file format for sequences



```
>VIT_201s0011g03530.1
AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
>VIT_201s0011g03540.1
CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
AGCCTCTGAGACACCACCTCAAACCTTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
>VIT_201s0011g03550.1
CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA
```

A Survey on Data Compression Methods for Biological Sequences. October 2016. Information (Switzerland) 7(4):56. DOI: 10.3390/info7040056

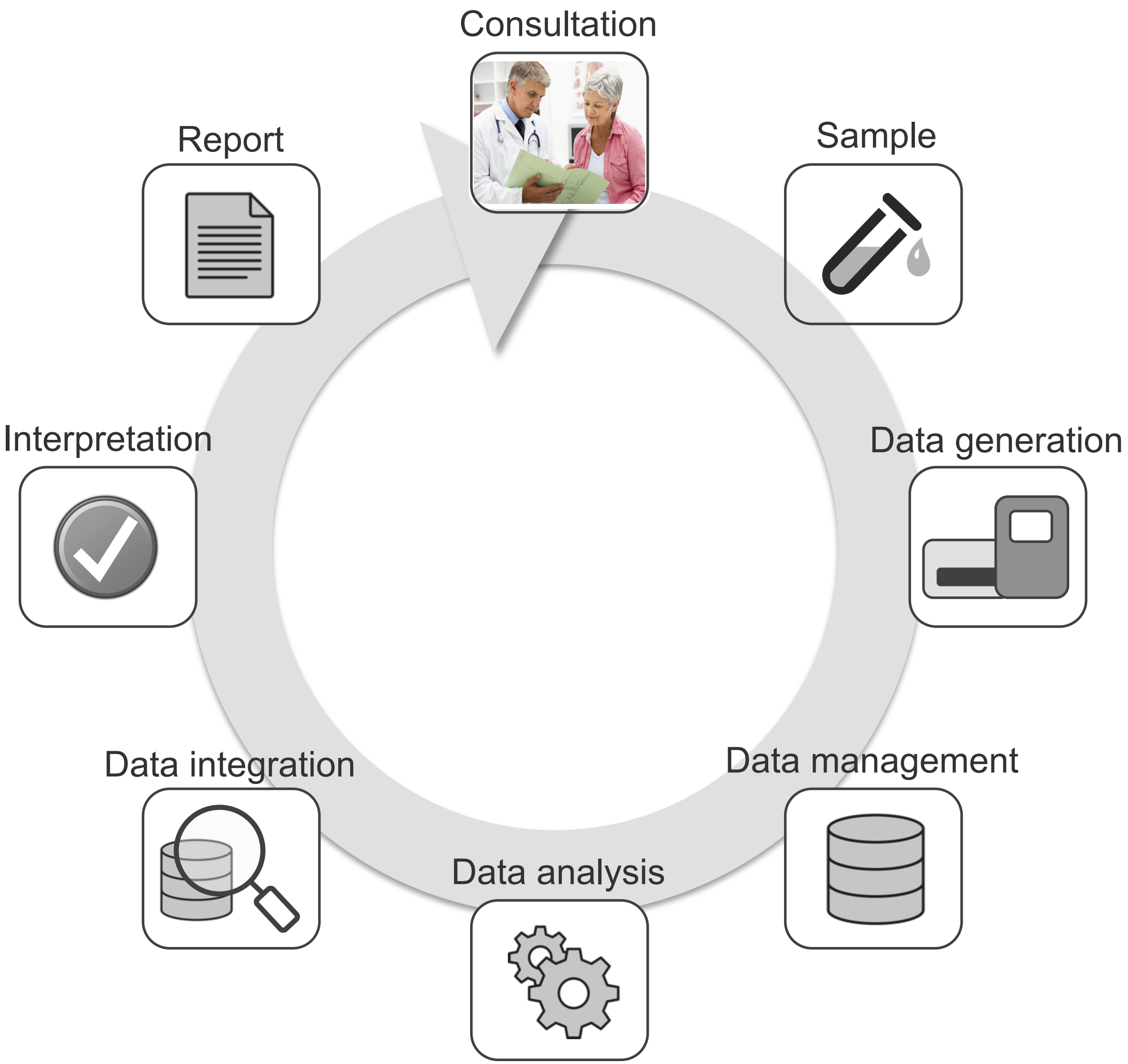
FASTA file format for sequences



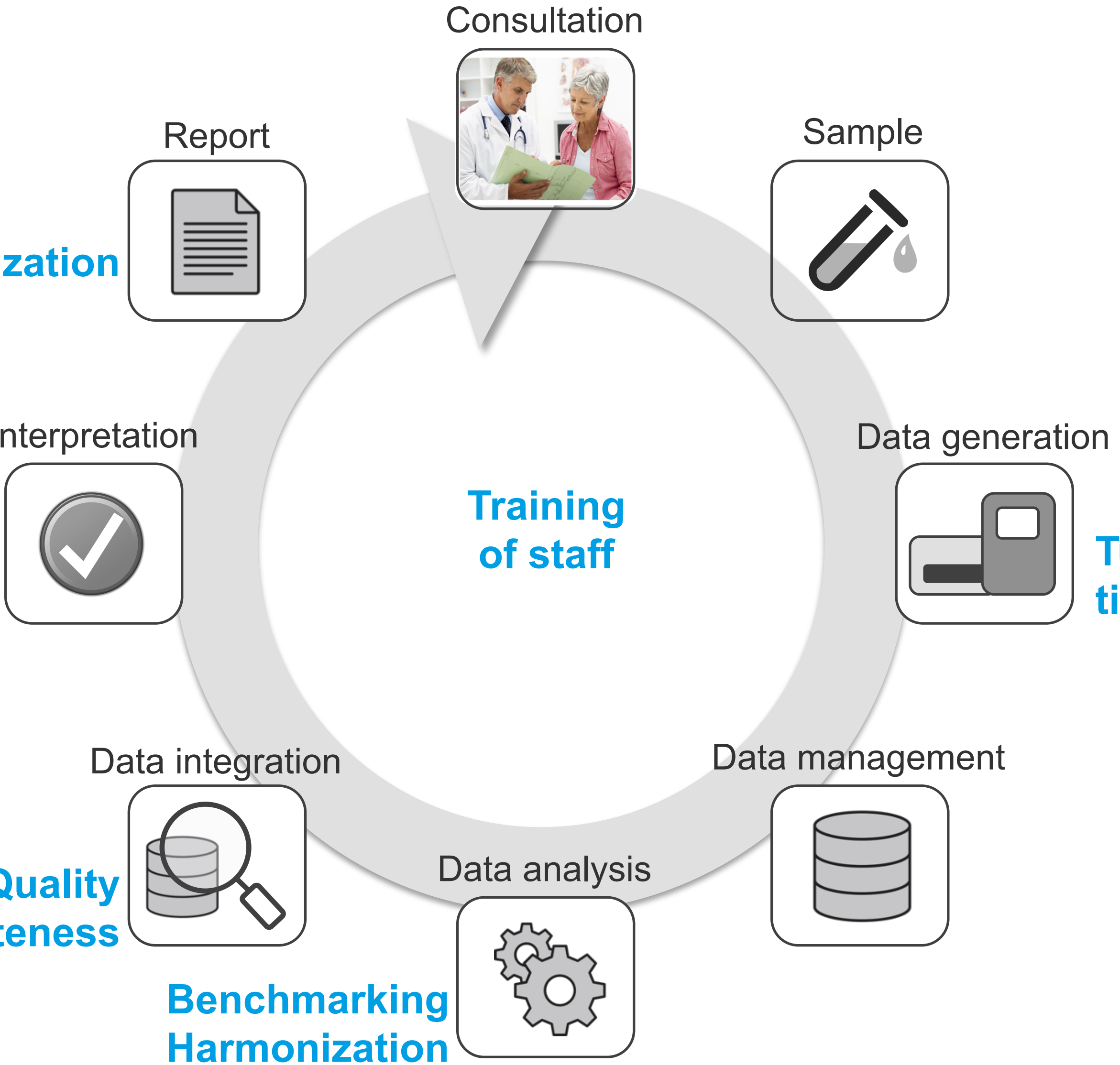
```
>VIT_201s0011g03530.1
AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
>VIT_201s0011g03540.1
CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
AGCCTCTGAGACACCACCTCAAACCTTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
>VIT_201s0011g03550.1
CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA
```

A Survey on Data Compression Methods for Biological Sequences. October 2016. Information (Switzerland) 7(4):56. DOI: 10.3390/info7040056

Clinical NGS pipeline



Clinical genomics pipeline: **main challenges**



Consultation

Sample

Data generation

Turnaround time and cost

Data management

Data analysis

Benchmarking Harmonization

Data integration

Quality Completeness

Interpretation

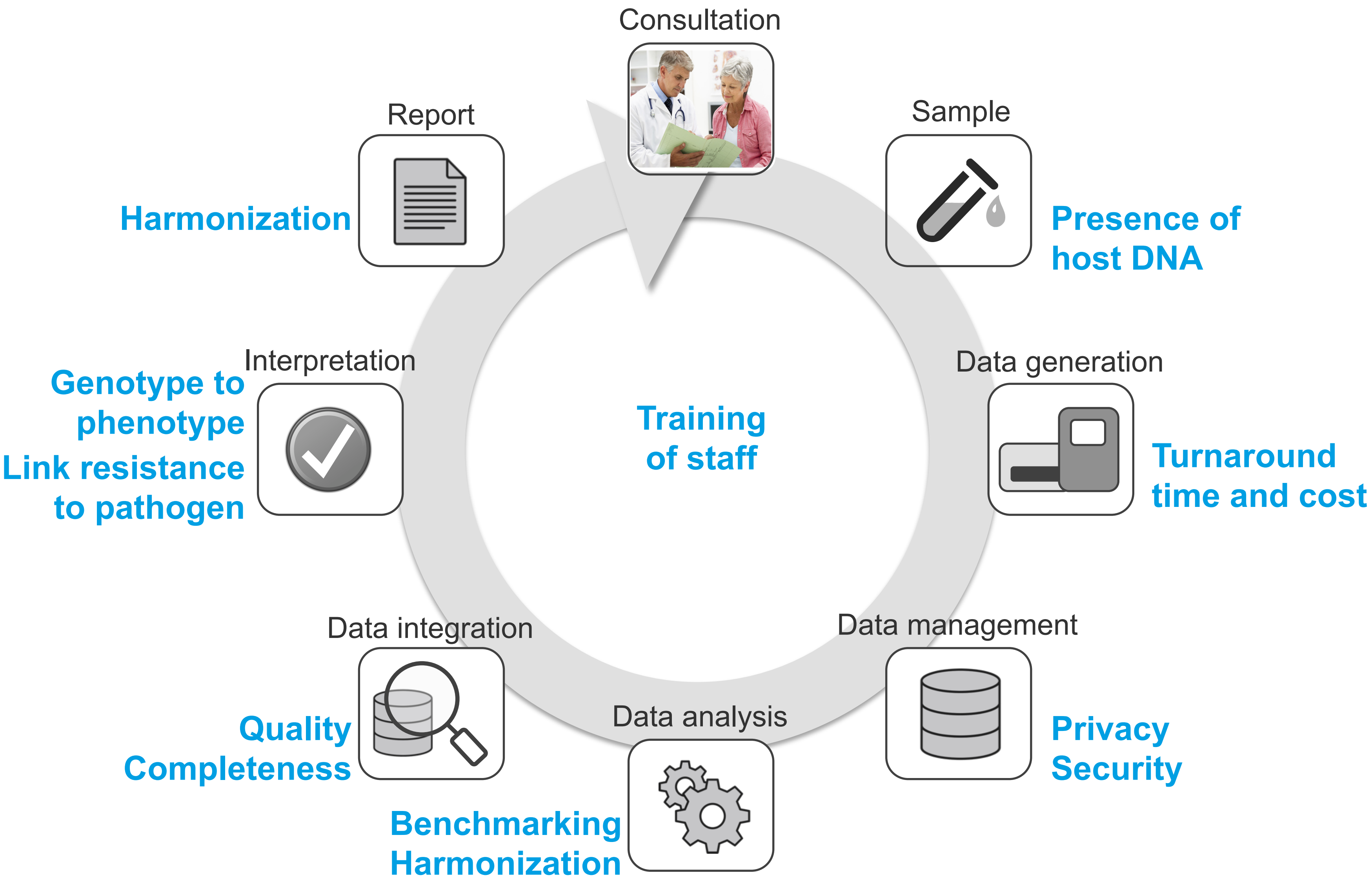
Genotype to phenotype

Report

Harmonization

Training of staff

Clinical **metagenomics** pipeline: main challenges



Swiss Institute of
Bioinformatics

Hands-on

Pre-processing of FASTQ datasets

Quality Control using FastQC

- FastQC aims to provide a QC report which can spot problems which originate either in the sequencer or in the starting library material
- It can either run as a stand alone interactive application for the immediate analysis of small numbers of FASTQ files
- Or run in a non-interactive mode where it would be suitable for integrating into a larger analysis pipeline
- INFO: https://rtsf.natsci.msu.edu/sites/_rtsf/assets/File/FastQC_TutorialAndFAQ_080717.pdf

Courtesy Dr Walid Gharib

Analysis Modules

1. Basic Statistics
2. Per Base Sequence Quality
3. Per Sequence Quality Scores
4. Per Base Sequence Content
5. Per Base GC Content
6. Per Sequence GC Content
7. Per Base N Content
8. Sequence Length Distribution
9. Duplicate Sequences
10. Overrepresented Sequences
11. Overrepresented Kmers

Courtesy Dr Walid Gharib

1. Basic Statistics

Filename

File type

Encoding

Total Sequences

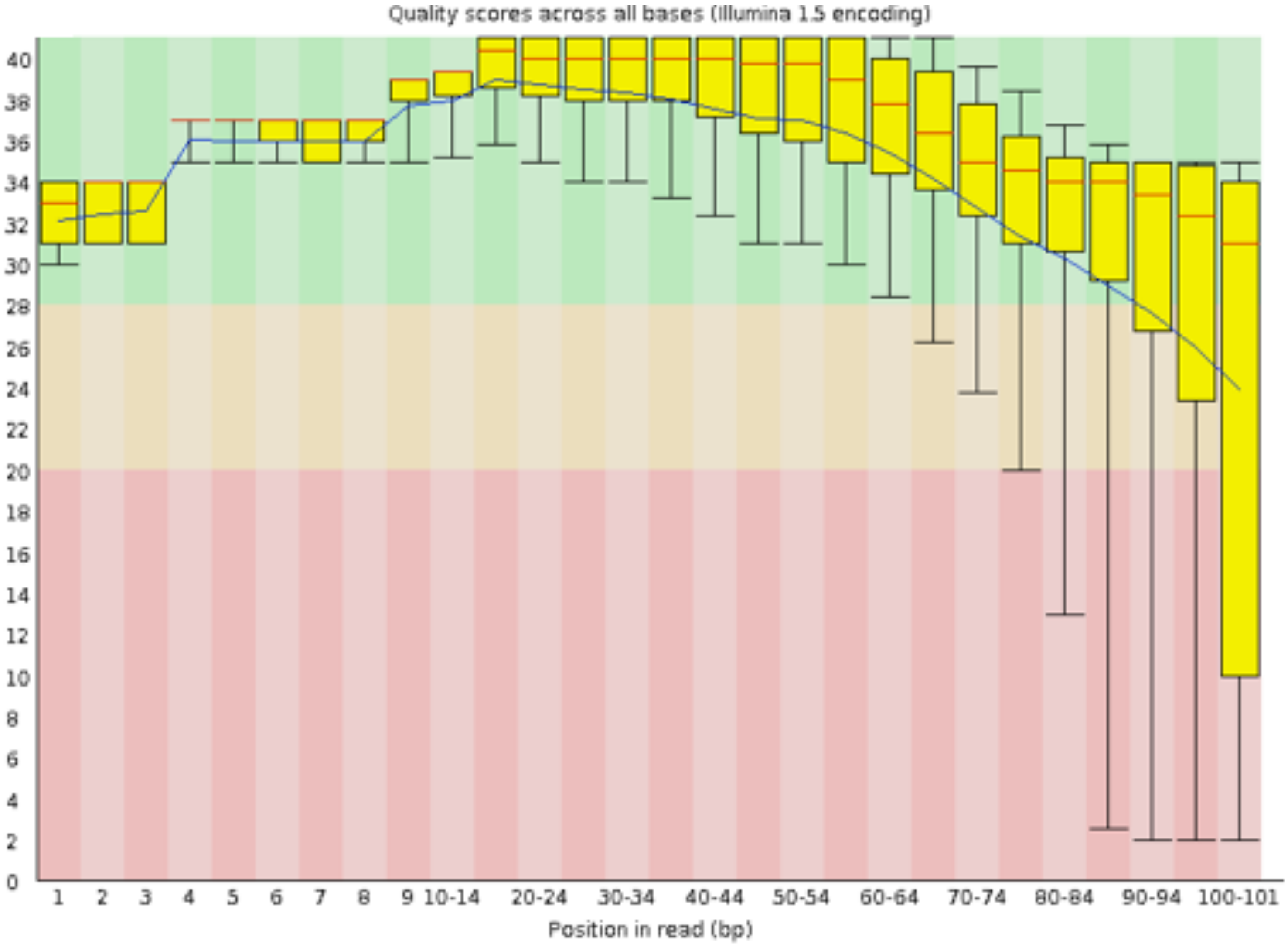
Filtered Sequences

Sequence Length

%GC

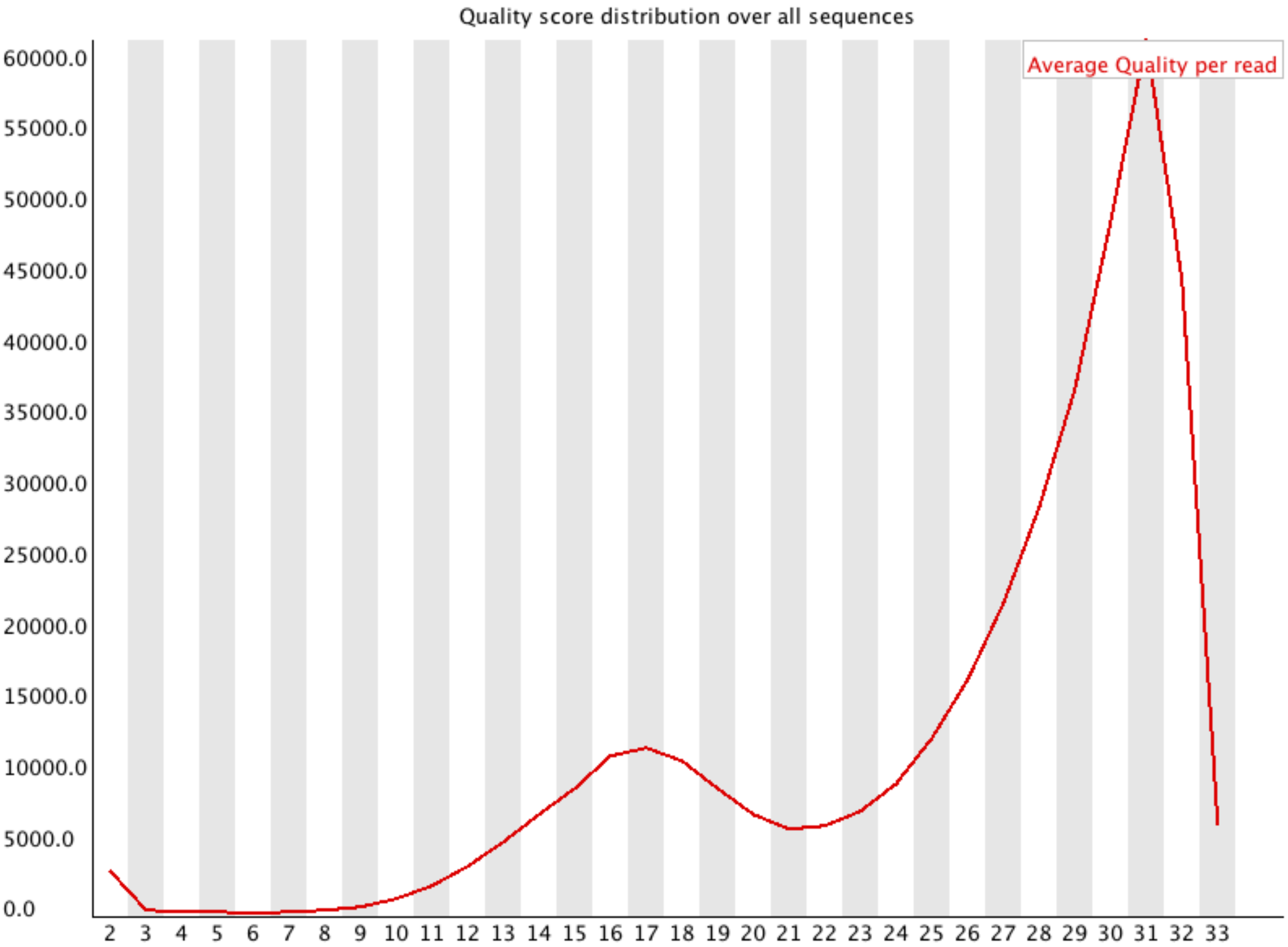
Courtesy Dr Walid Gharib

2. Per Base Sequence Quality



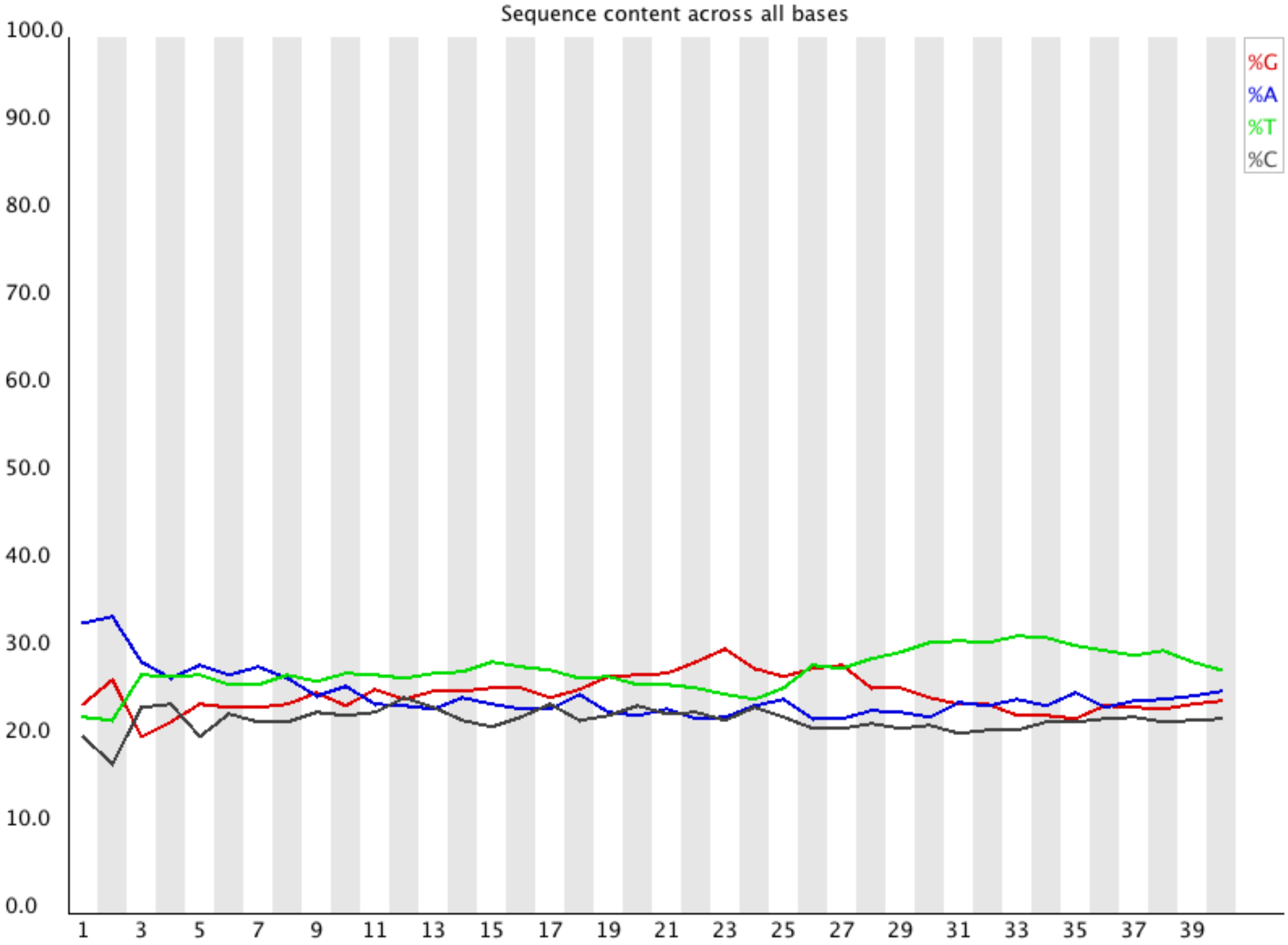
CC BY-NC-SA SIB Clinical Bioinformatics 2019

3. Per Sequence Quality Scores



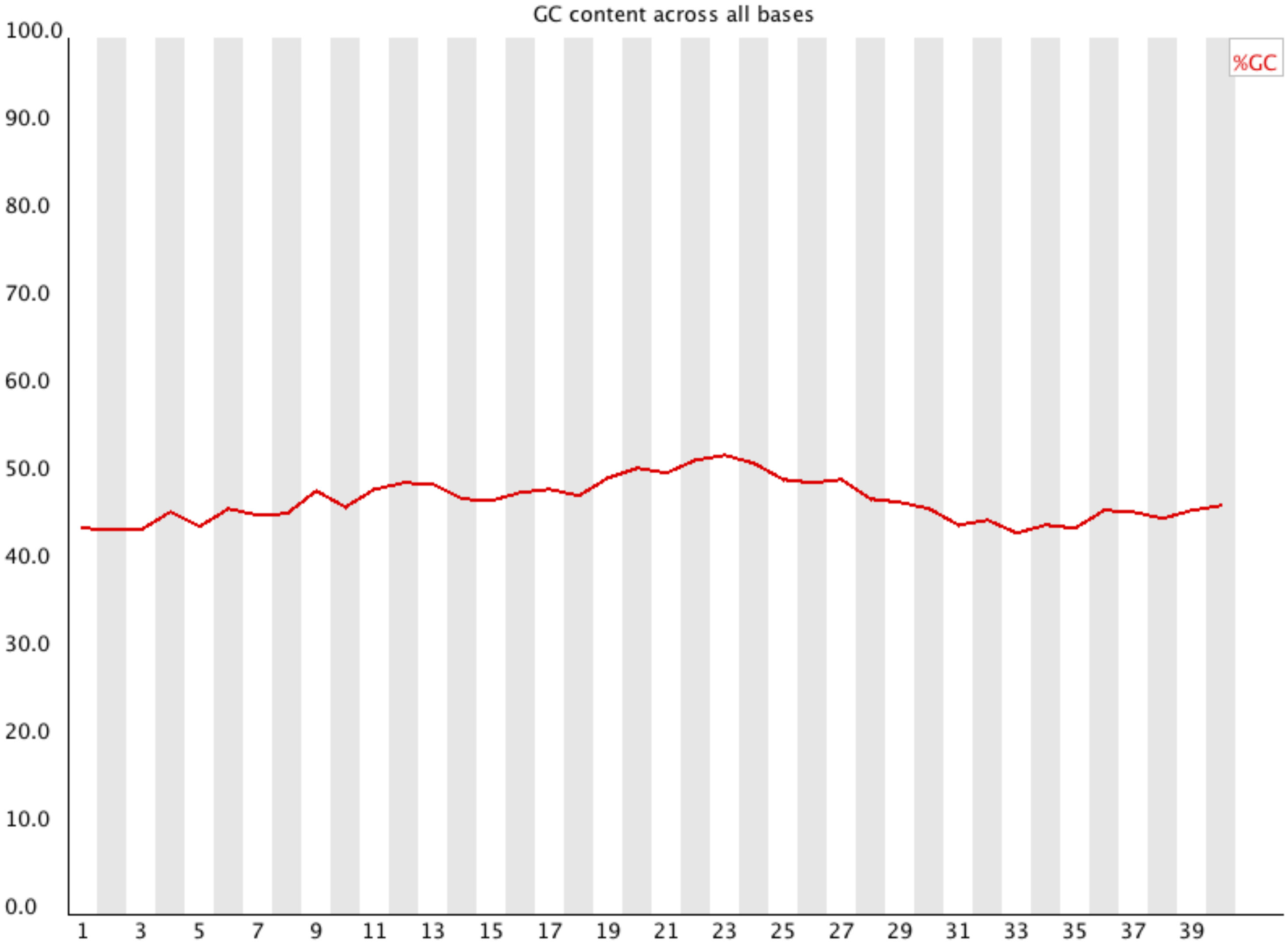
Courtesy Dr Walid Gharib

4. Per Base Sequence Content



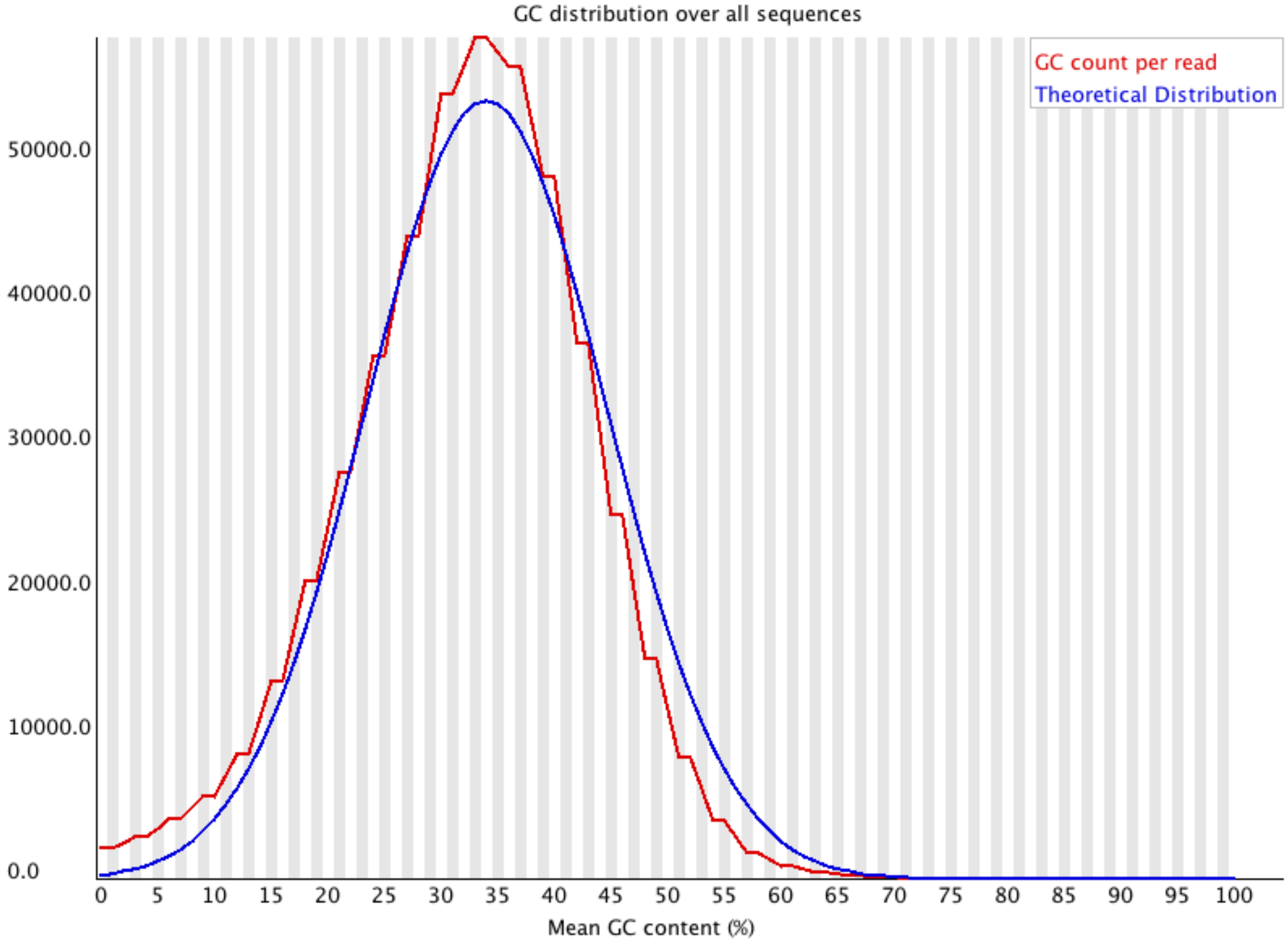
Courtesy Dr Walid Gharib

5. Per Base GC Content



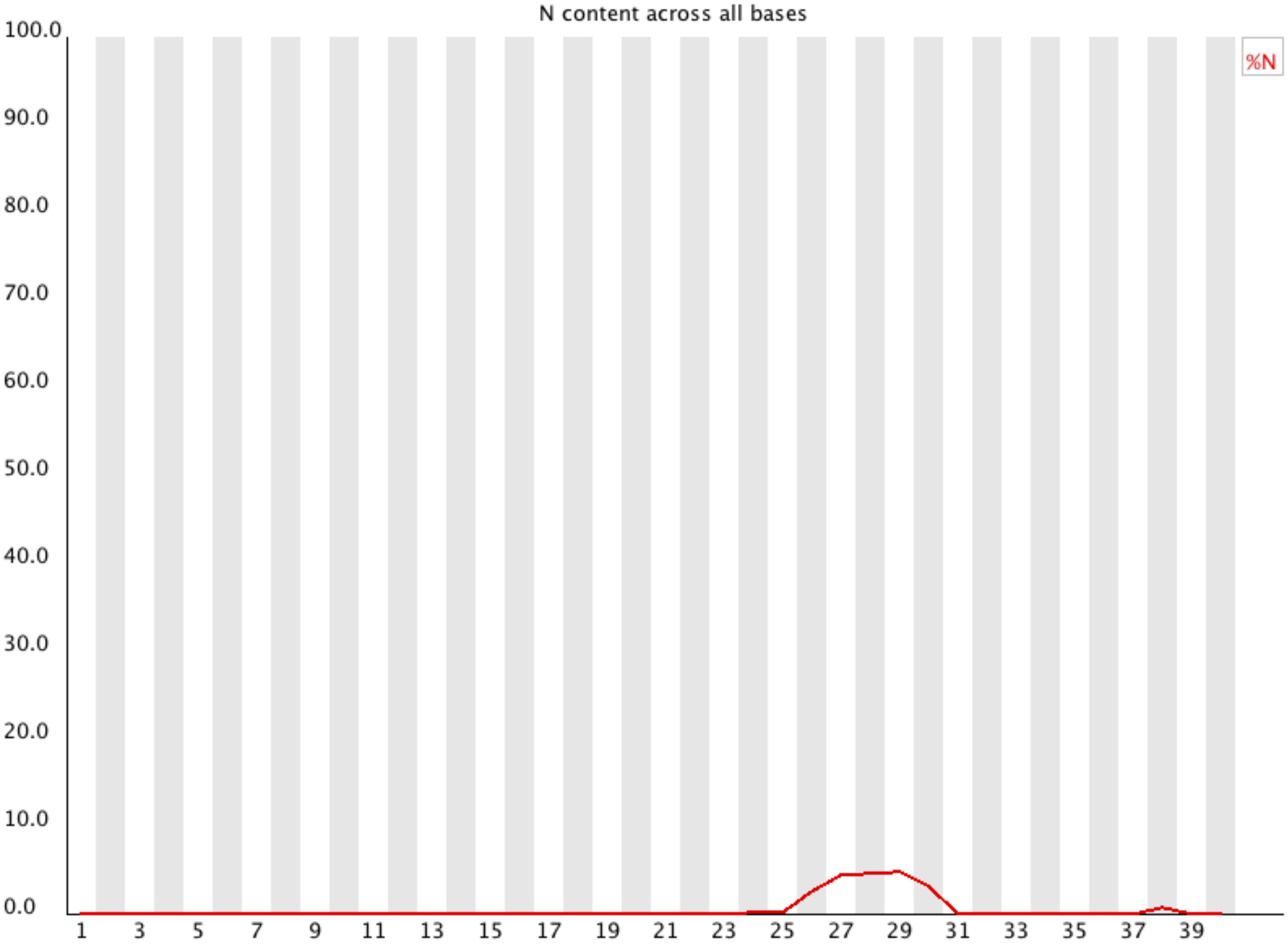
Courtesy Dr Walid Gharib

5. Per sequence GC Content (1)



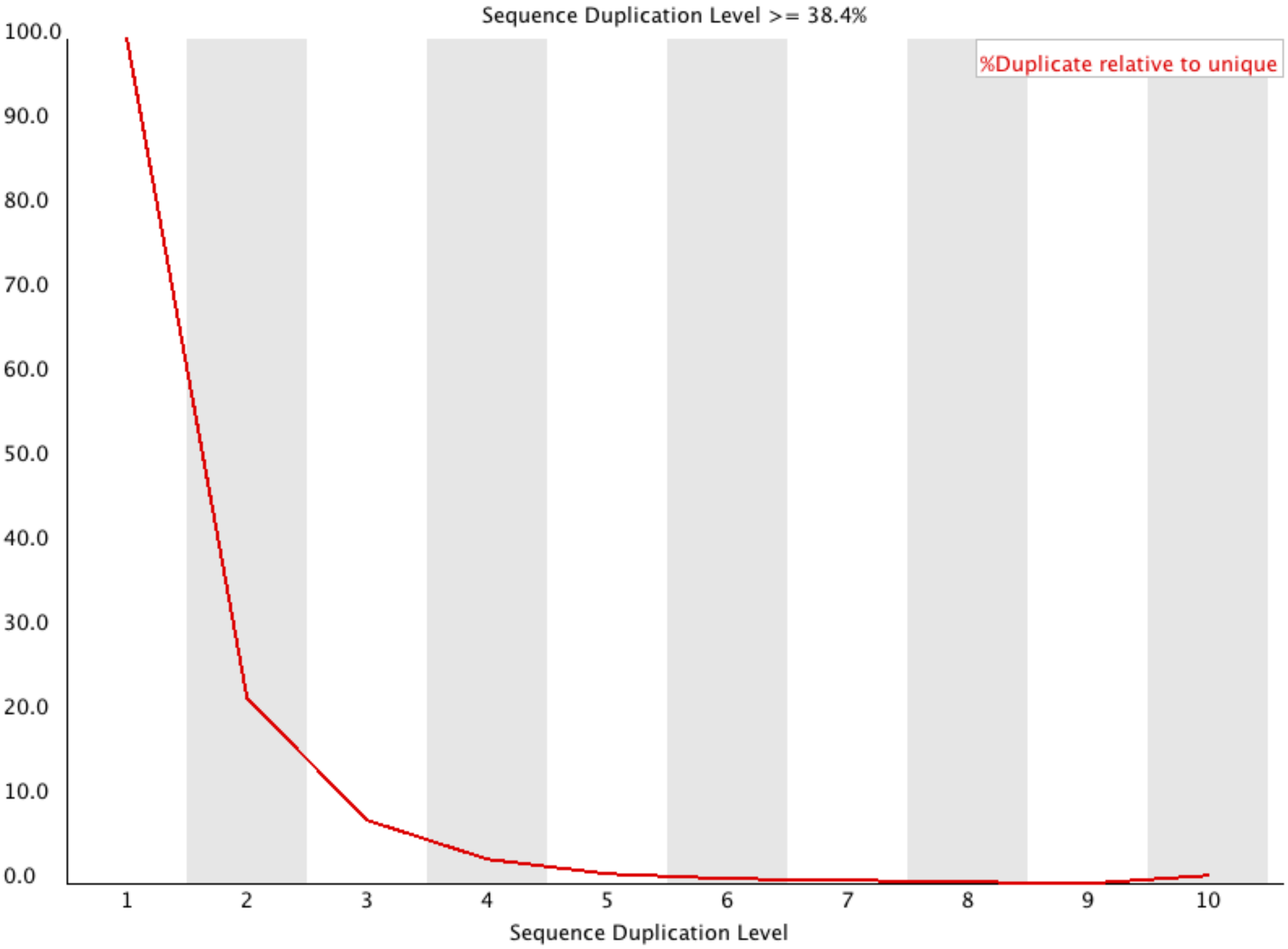
Courtesy Dr Walid Gharib

7. Per Base N Content



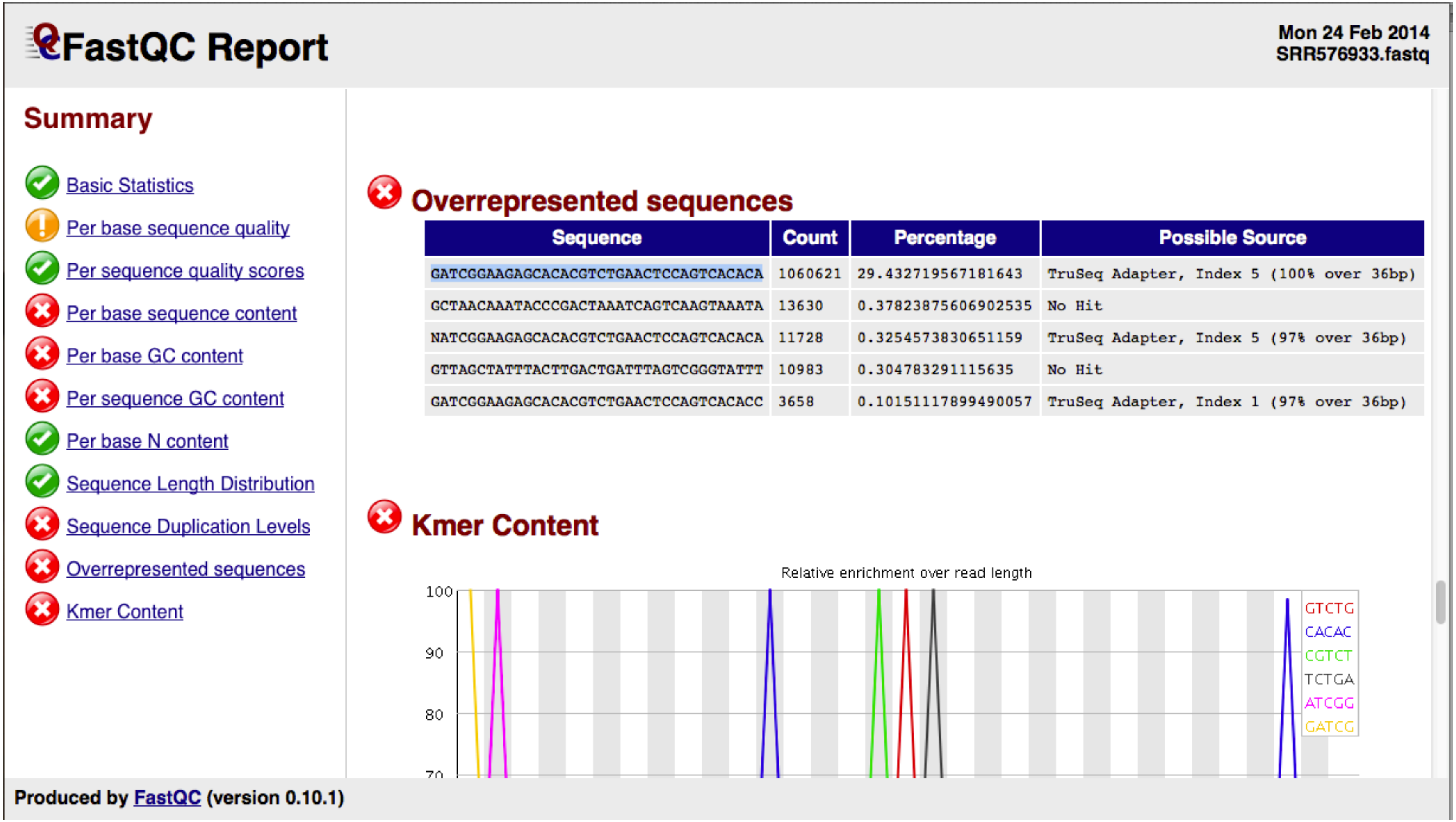
Courtesy Dr Walid Gharib

9. Duplicate Sequences



Courtesy Dr Walid Gharib

10. Overrepresented Sequences



FastQC Report

Mon 24 Feb 2014
SRR576933.fastq

Summary

- ✓ Basic Statistics
- ! Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per base GC content
- ✗ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✗ Overrepresented sequences
- ✗ Kmer Content

✗ Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACA	1060621	29.432719567181643	TruSeq Adapter, Index 5 (100% over 36bp)
GCTAACAAATACCCGACTAAATCAGTCAAGTAAATA	13630	0.37823875606902535	No Hit
NATCGGAAGAGCACACGTCTGAACTCCAGTCACACA	11728	0.3254573830651159	TruSeq Adapter, Index 5 (97% over 36bp)
GTTAGCTATTTACTTGACTGATTTAGTCGGGTATTT	10983	0.304783291115635	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACC	3658	0.10151117899490057	TruSeq Adapter, Index 1 (97% over 36bp)

✗ Kmer Content

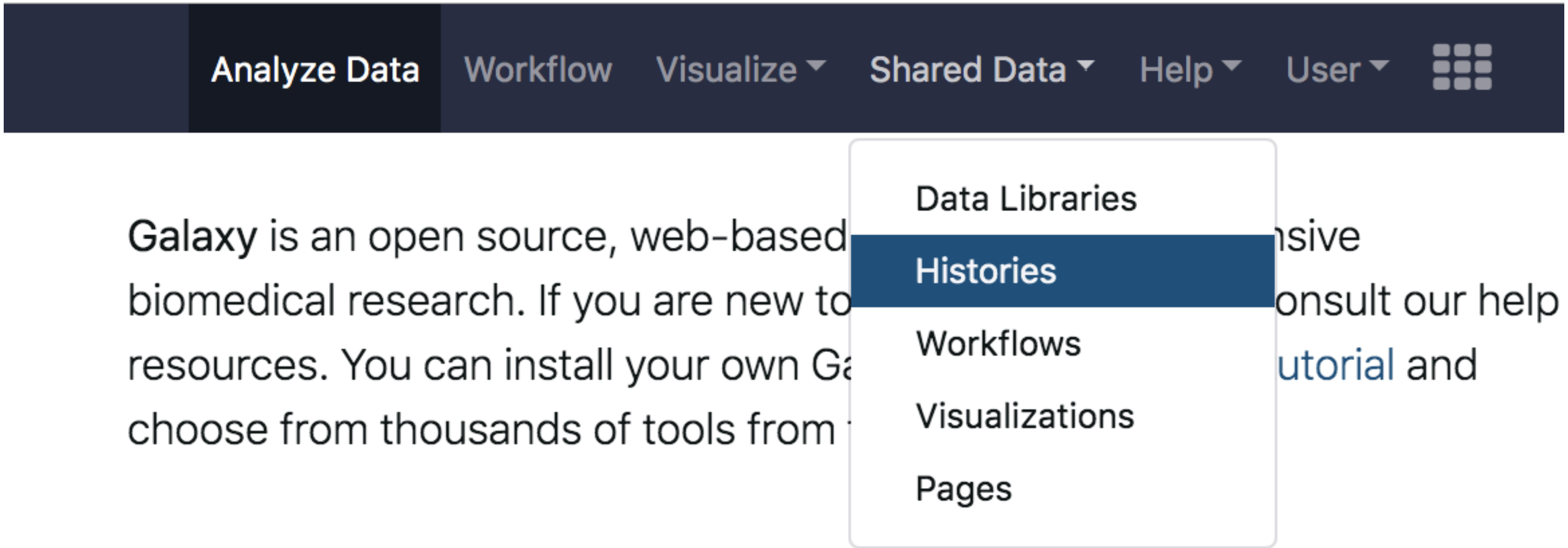
Produced by FastQC (version 0.10.1)

Courtesy Dr Walid Gharib

Run FastQC on our the course datasets

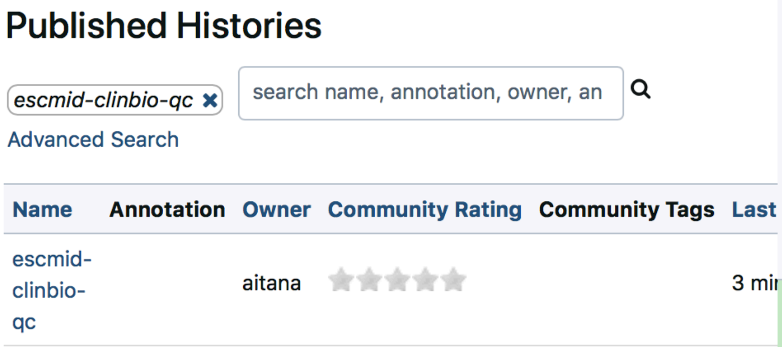
- Go to: https://usegalaxy.org
- Create an account (requires email validation)
- Go to:
  - Shared Data/Histories, search for *escmid-clinbio-qc* https://usegalaxy.org:/u/aitana/h/escmid-clinbio-qc
  - Import history
- You will find several datasets:
- WGS of S. aureus
- Metagenomics data (plasma spiked with viruses)

Go to Shared Data/Histories



Galaxy is an open source, web-based nsive biomedical research. If you are new to onsult our help resources. You can install your own Ga utorial and choose from thousands of tools from

Search for “escmid-clinbio-qc” and Import history (+)



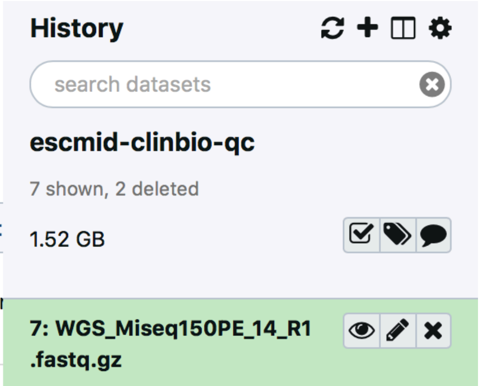
Published Histories

escmid-clinbio-qc ✖

search name, annotation, owner, an

Advanced Search

Name	Annotation	Owner	Community Rating	Community Tags	Last
escmid-clinbio-qc		aitana	☆☆☆☆☆		3 mi



History

search datasets

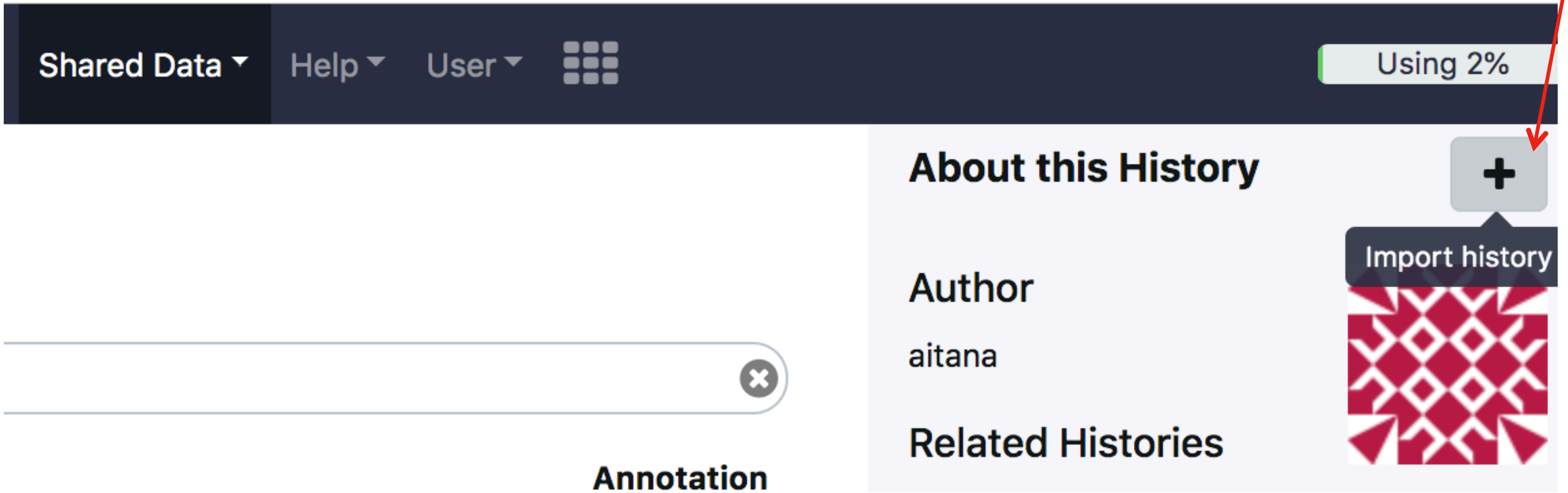
escmid-clinbio-qc

7 shown, 2 deleted

1.52 GB

7: WGS_Miseq150PE_14_R1.fastq.gz

THEN CLICK ON escmid-clinbio-qc and import History



About this History

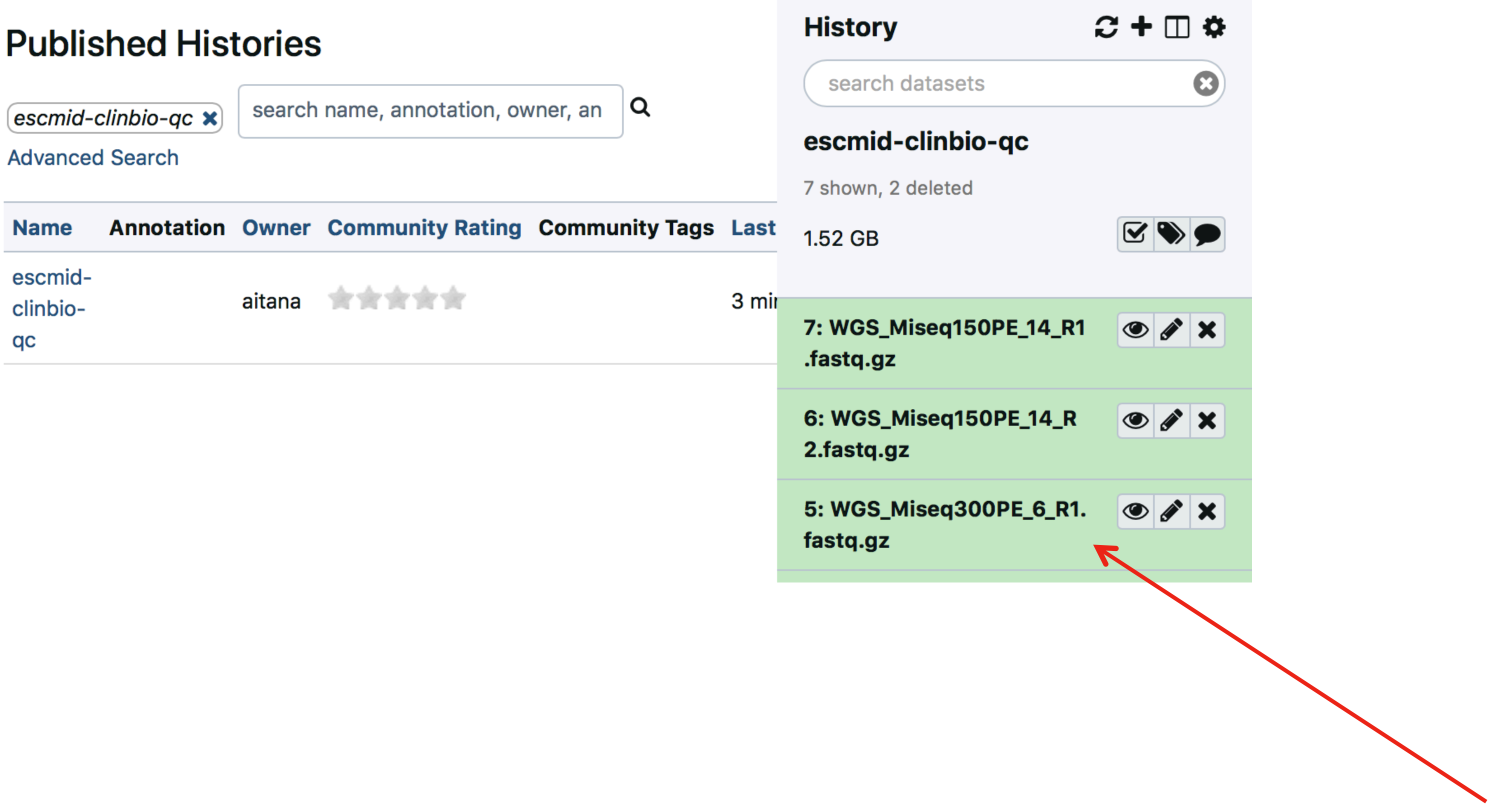
Author

aitana

Related Histories

Annotation

The datasets are now in your history



Published Histories

escmid-clinbio-qc ✖ search name, annotation, owner, an

Advanced Search

Name	Annotation	Owner	Community Rating	Community Tags	Last
escmid-clinbio-qc		aitana	☆☆☆☆☆		3 mi

History

search datasets

escmid-clinbio-qc

7 shown, 2 deleted

1.52 GB

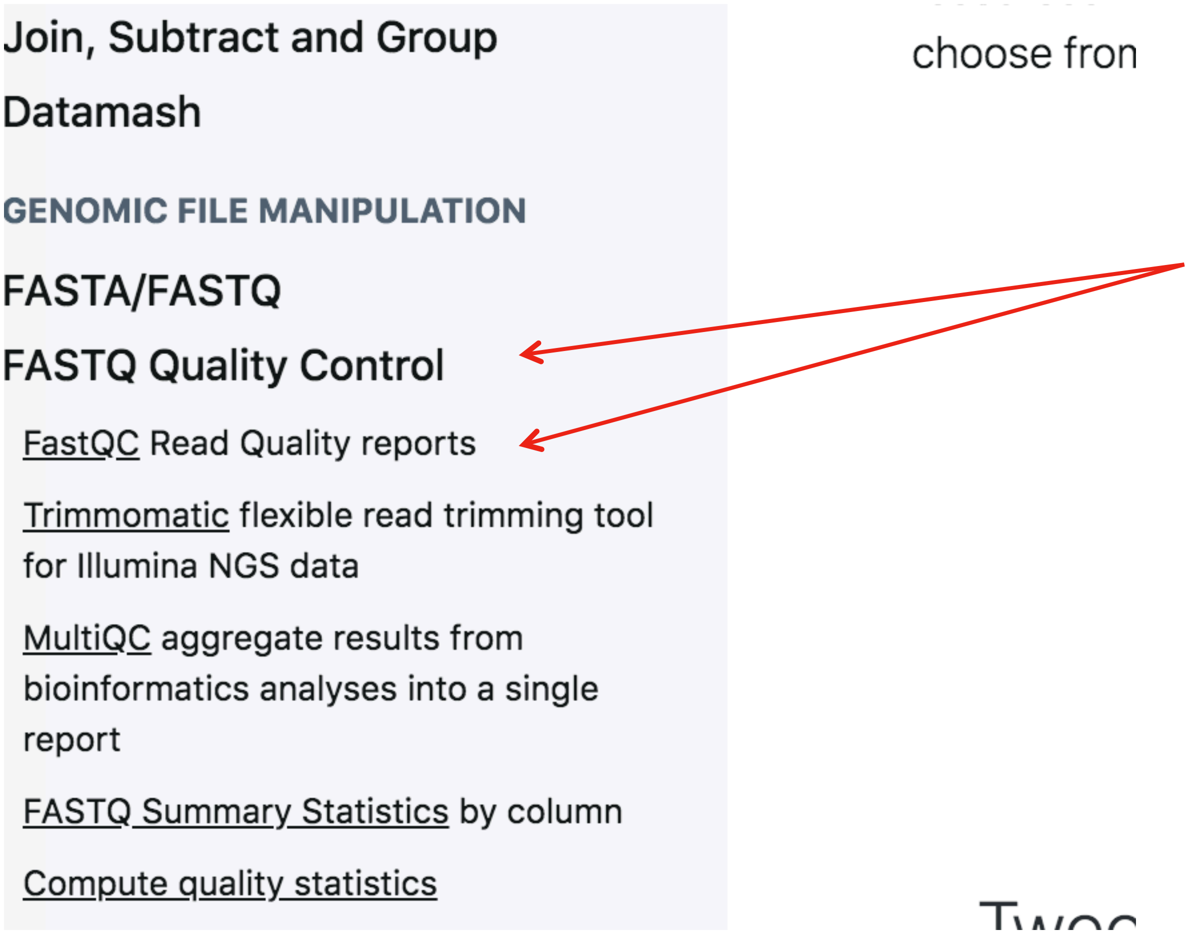
7: WGS_Miseq150PE_14_R1.fastq.gz

6: WGS_Miseq150PE_14_R2.fastq.gz

5: WGS_Miseq300PE_6_R1.fastq.gz

Run FastQC with default parameters

On the left menu, **select FASTQ Quality Control**, and then **FastQC Read Quality reports**



Join, Subtract and Group

Datamash

GENOMIC FILE MANIPULATION

FASTA/FASTQ

FASTQ Quality Control

FastQC Read Quality reports

Trimmomatic flexible read trimming tool for Illumina NGS data

MultiQC aggregate results from bioinformatics analyses into a single report

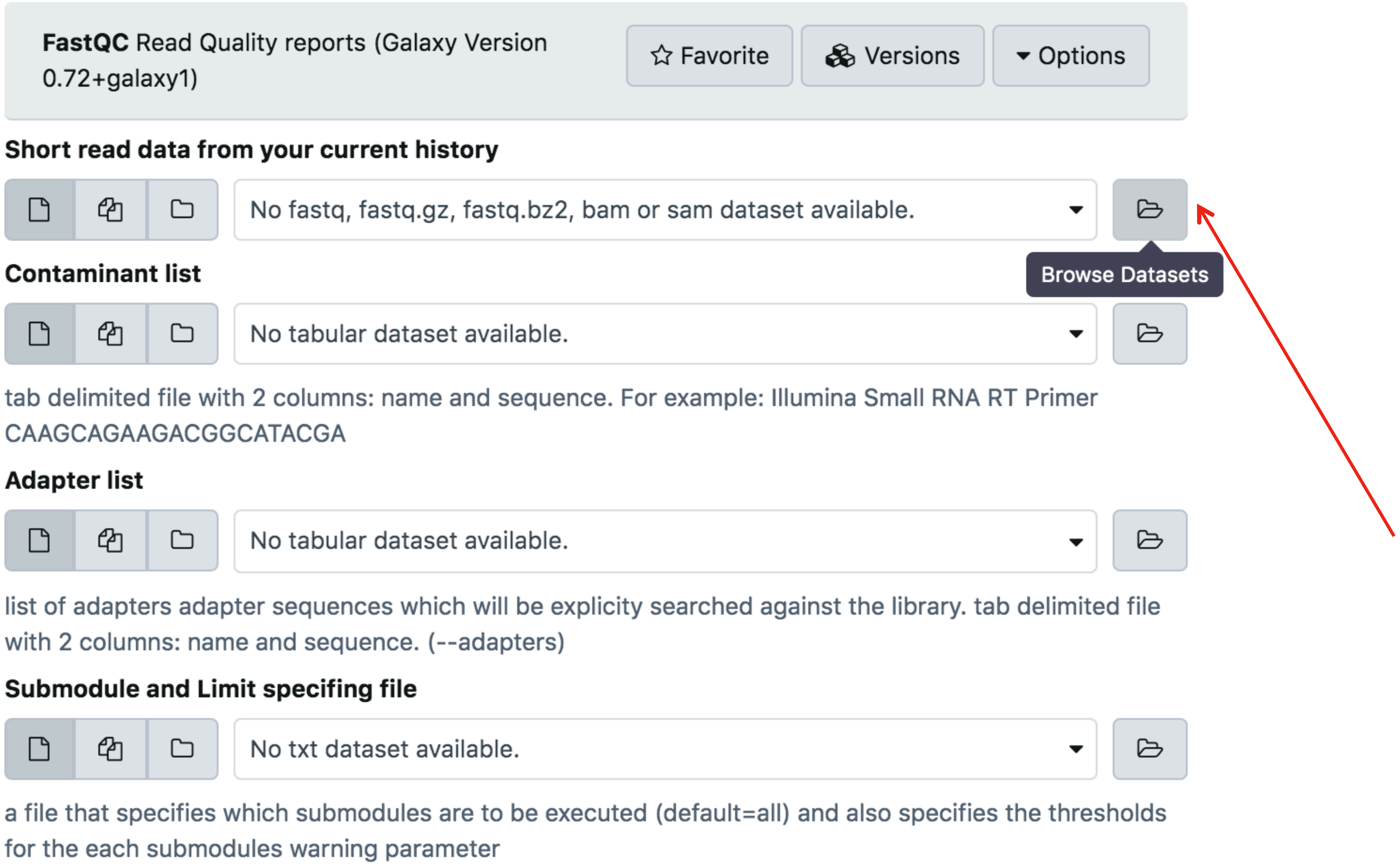
FASTQ Summary Statistics by column

Compute quality statistics

choose fron

Twoc

Browse datasets and select one FASTQ file



FastQC Read Quality reports (Galaxy Version 0.72+galaxy1)

Short read data from your current history

No fastq, fastq.gz, fastq.bz2, bam or sam dataset available.

Contaminant list

No tabular dataset available.

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Adapter list

No tabular dataset available.

list of adapters adapter sequences which will be explicity searched against the library. tab delimited file with 2 columns: name and sequence. (--adapters)

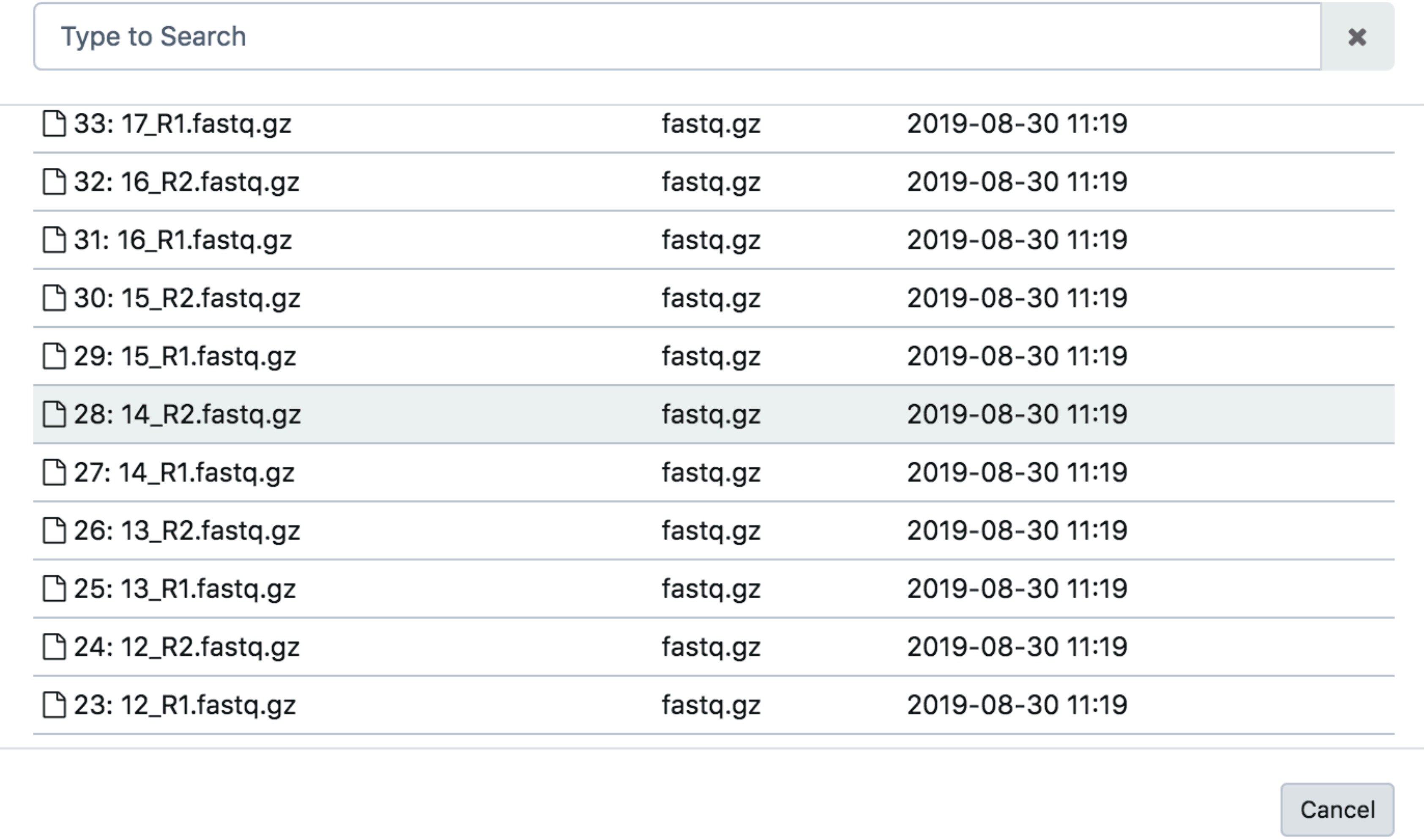
Submodule and Limit specifing file

No txt dataset available.

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Browse datasets and select FASTQ file

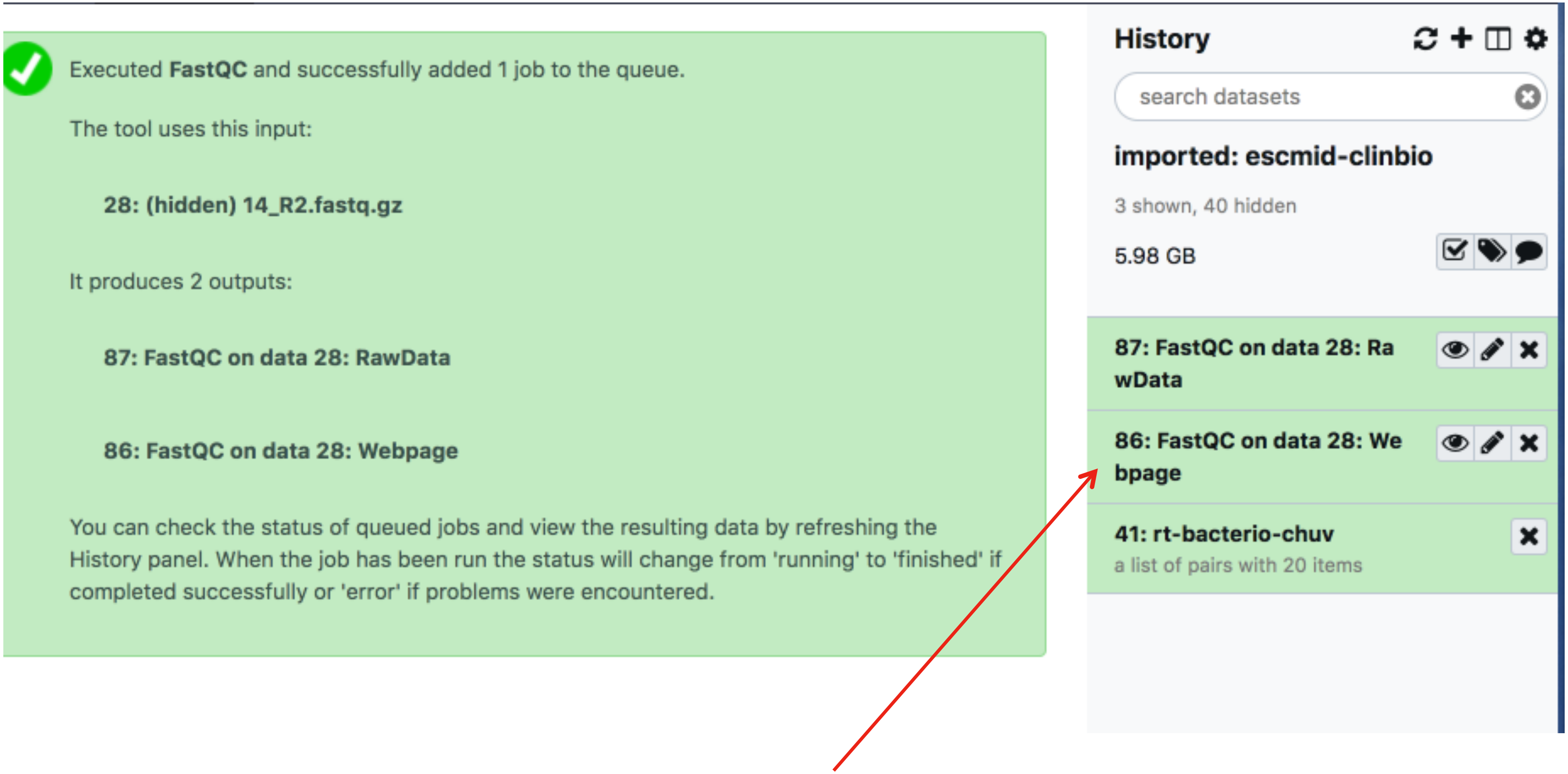
Click on “Execute” (blue button at the bottom)



33: 17_R1.fastq.gz	fastq.gz	2019-08-30 11:19
32: 16_R2.fastq.gz	fastq.gz	2019-08-30 11:19
31: 16_R1.fastq.gz	fastq.gz	2019-08-30 11:19
30: 15_R2.fastq.gz	fastq.gz	2019-08-30 11:19
29: 15_R1.fastq.gz	fastq.gz	2019-08-30 11:19
28: 14_R2.fastq.gz	fastq.gz	2019-08-30 11:19
27: 14_R1.fastq.gz	fastq.gz	2019-08-30 11:19
26: 13_R2.fastq.gz	fastq.gz	2019-08-30 11:19
25: 13_R1.fastq.gz	fastq.gz	2019-08-30 11:19
24: 12_R2.fastq.gz	fastq.gz	2019-08-30 11:19
23: 12_R1.fastq.gz	fastq.gz	2019-08-30 11:19

Wait for FastQC to finish running

Open FastQC on data xx: Webpage



✓ Executed **FastQC** and successfully added 1 job to the queue.

The tool uses this input:

28: (hidden) 14_R2.fastq.gz

It produces 2 outputs:

87: FastQC on data 28: RawData

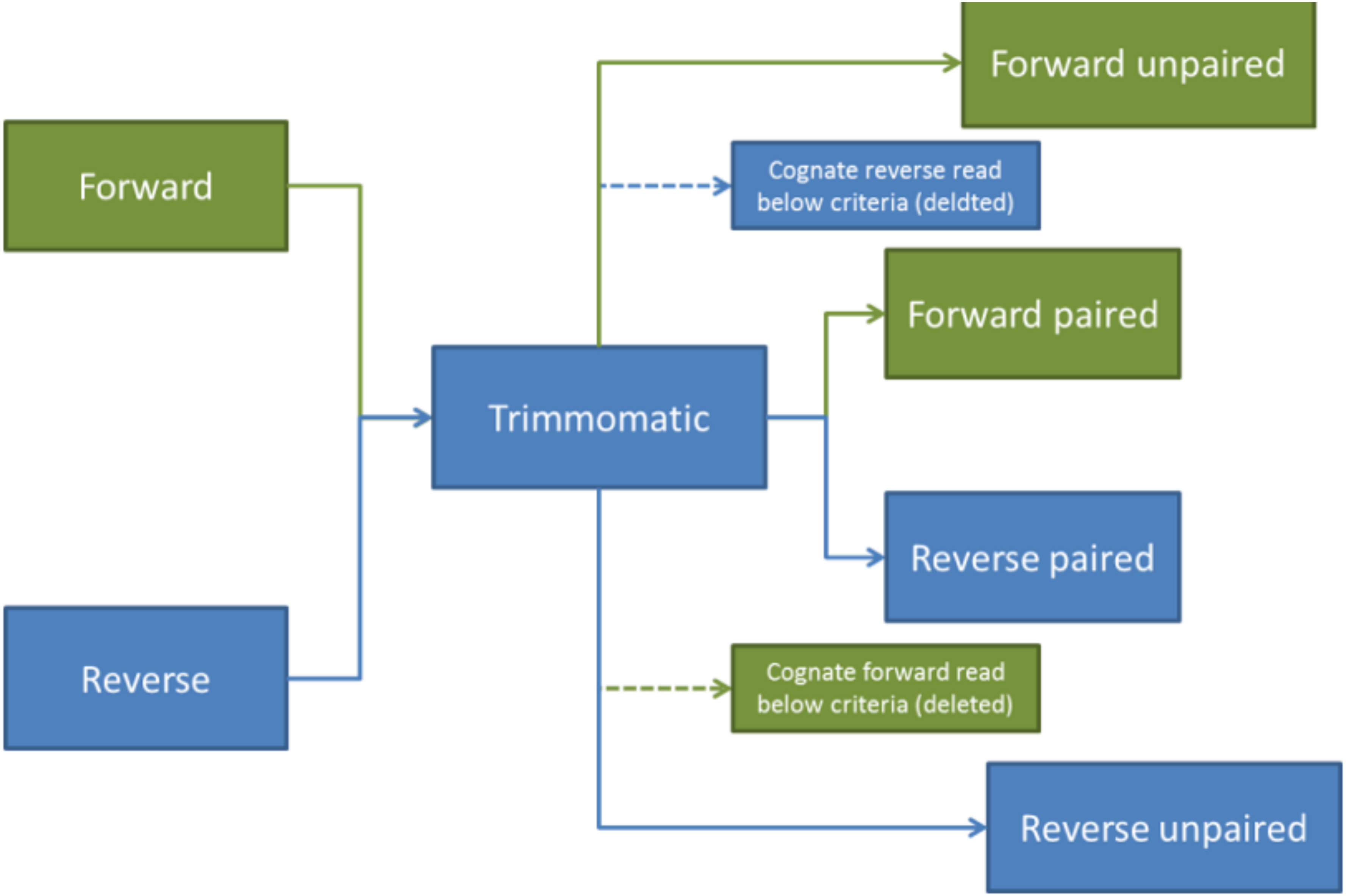
86: FastQC on data 28: Webpage

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

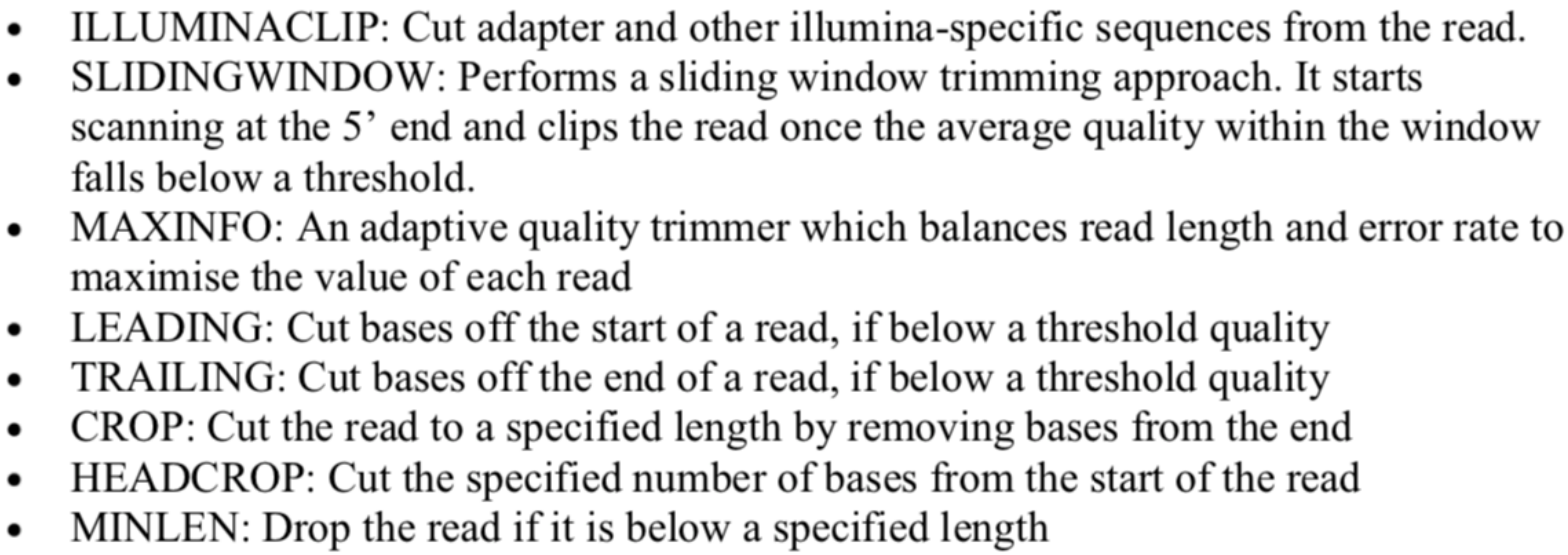
Repeat FastQC for other FASTQ files

- Discuss results

Using Trimmomatic


Forward unpaired
Forward
Cognate reverse read below criteria (deldted)
Forward paired
Trimmomatic
Reverse paired
Reverse
Cognate forward read below criteria (deleted)
Reverse unpaired

Using Trimmomatic



- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.
- SLIDINGWINDOW: Performs a sliding window trimming approach. It starts scanning at the 5’ end and clips the read once the average quality within the window falls below a threshold.
- MAXINFO: An adaptive quality trimmer which balances read length and error rate to maximise the value of each read
- LEADING: Cut bases off the start of a read, if below a threshold quality
- TRAILING: Cut bases off the end of a read, if below a threshold quality
- CROP: Cut the read to a specified length by removing bases from the end
- HEADCROP: Cut the specified number of bases from the start of the read
- MINLEN: Drop the read if it is below a specified length

Using Trimmomatic

Example code:

```
trimmomatic PE -phred33 \
input_forward.fq.gz input_reverse.fq.gz \
output_forward_paired.fq.gz output_forward_unpaired.fq.gz \
output_reverse_paired.fq.gz output_reverse_unpaired.fq.gz \
ILLUMINACLIP:TruSeq3-PE.fa:2:30:10
LEADING:3
TRAILING:3
SLIDINGWINDOW:4:15
MINLEN:36
```

On Galaxy, run Trimmomatic on one dataset

- Select a forward and reverse pair
- Run Trimmomatic with default options
- Re-run FastQC on the newly created (R1 paired, R2 paired)
- Compare the output from FastQC before/after Trimmomatic