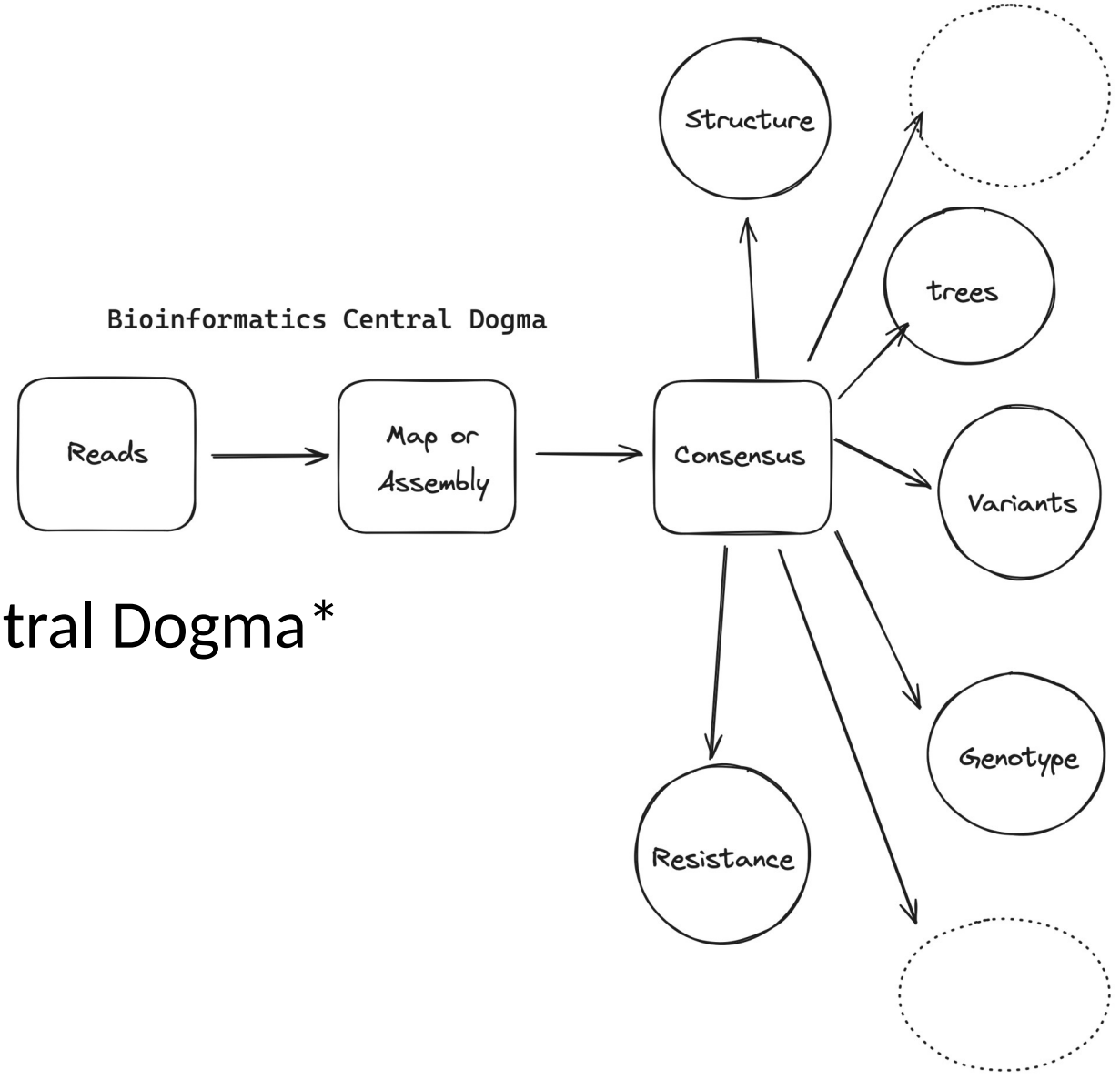
Genome Assembly

WHO Collaborating Centre
for Reference and
Research on Influenza
VIDRL

A joint venture between The University of Melbourne and The Royal Melbourne Hospital

Objective

- To learn the basics of genome assembly
- AKA Bioinformatics Central Dogma*

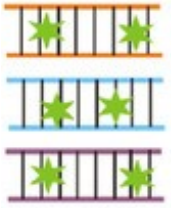


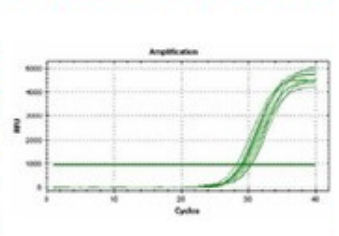
*There is no such thing I made it up!

Viral sequencing review

1. Clinical sample

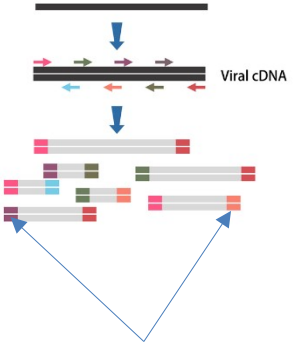
2. Diagnostic PCR





3. Amplify viral RNA

Tiled amplicon PCR



PCR primers

4. Sequencing

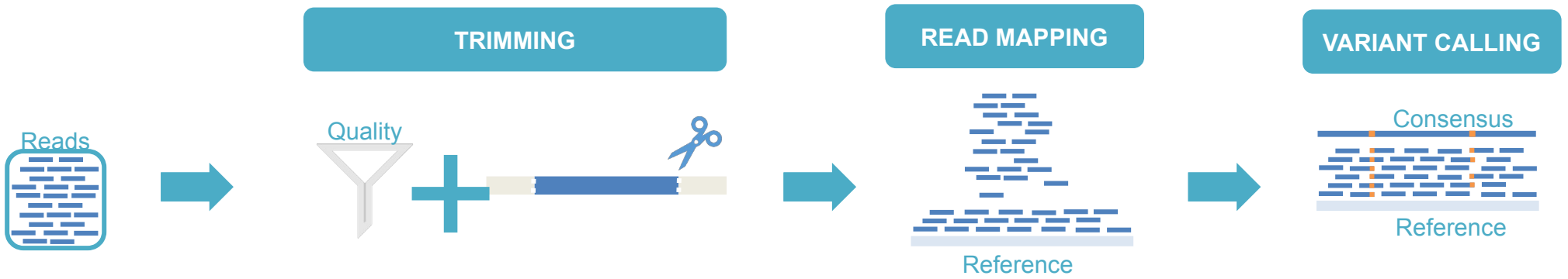
Short-read Illumina sequencing

OR

Long-read Nanopore sequencing

5. ?????

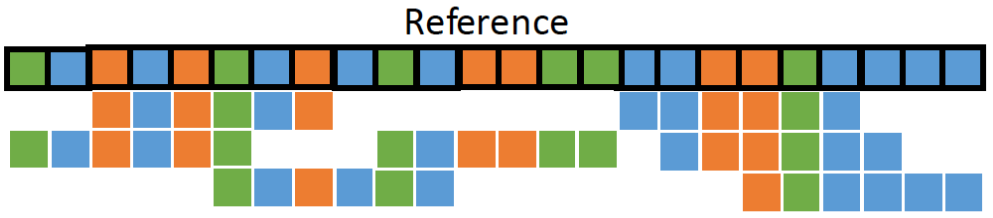
Processing amplicon sequencing data



Genome assembly from NGS data

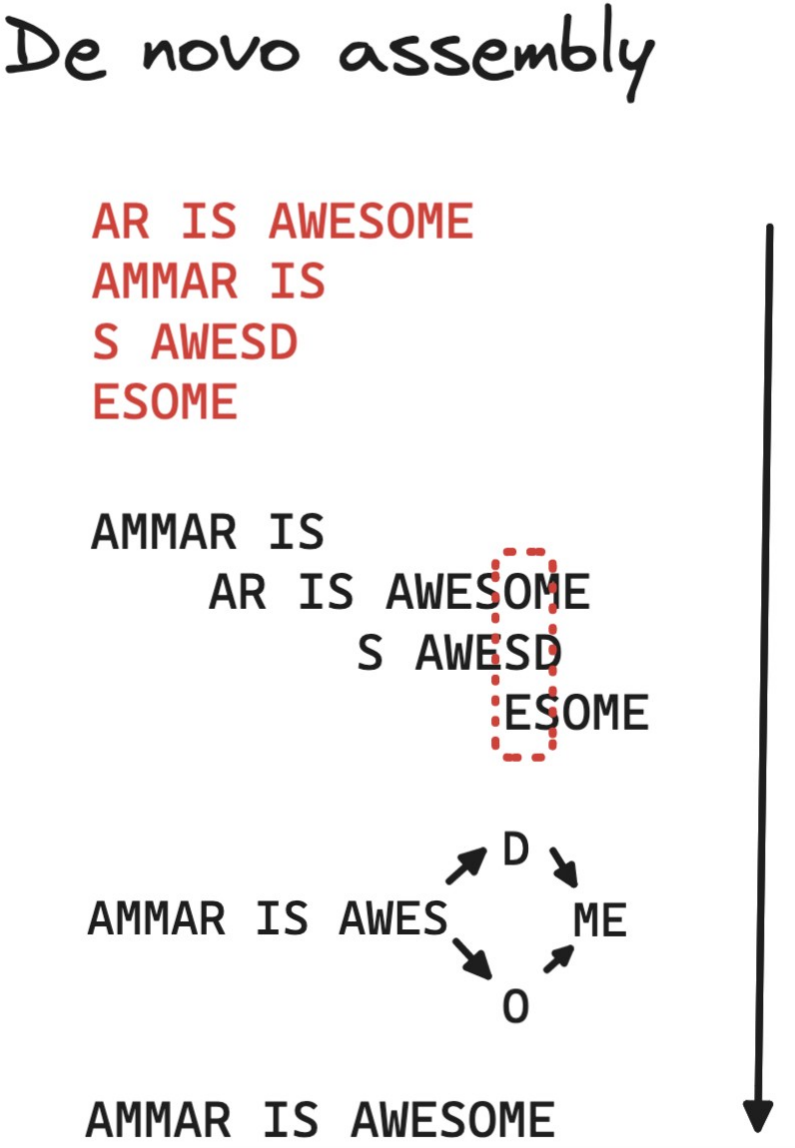
- ***De novo***
  - Latin for "from the new"
  - NO Reference sequence used
  - Results in **contigs** (contiguous sequences)
  - Requires a **scaffold** to put **contigs** into order across repetitive regions
  - Mostly used for metagenomics and assembling genomes that have no reference

- **Reference-based**
  - Individual reads are **mapped** directly to the position of the reference genome that they **align** the best
  - Disadvantage is that pathogens that evolve rapidly and are highly variable may not assemble if a close enough relative sequence is not used as the reference



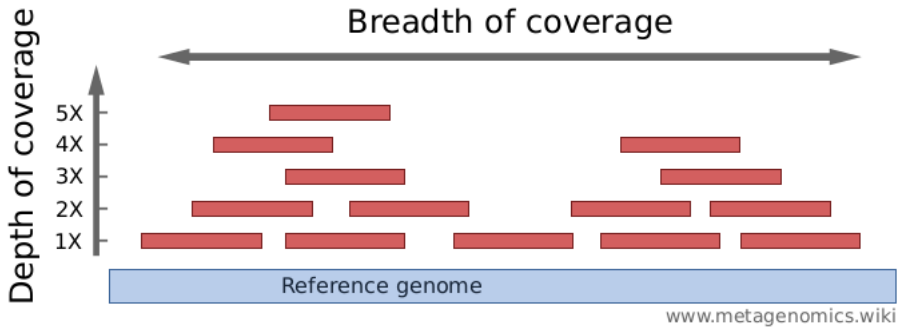
Genome assembly

- ***De novo***
  - Latin for “from the new”
  - NO Reference sequence used
  - Results in **contigs** (contiguous sequences)
  - Requires a **scaffold** to put **contigs** into order across repetitive regions
  - Mostly used for metagenomics and assembling genomes that have no reference



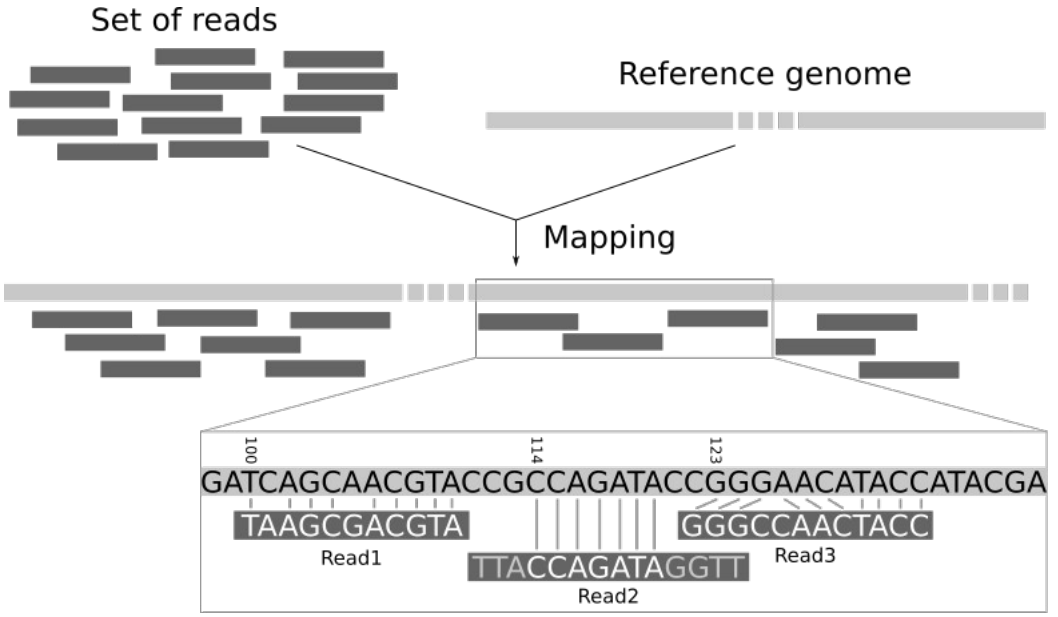
Coverage depth vs breadth

Sequencing depth, read depth, depth of coverage – number of times a specific position in a genome is sequenced
e.g. 50x depth coverage means that each base in the genome has been sequenced 50 times on average
Breadth of coverage – proportion of the genome that has been sequenced
e.g. 95% coverage means 95% of the length of the genome has been sequenced



Read mapping

Read mapping is the process of aligning reads to a reference genome



https://training.galaxyproject.org/training-material/topics/sequence-analysis/images/mapping/mapping.png

Read mapping Algorithms and Tools

Minimizers – collects unique kmers that are used as seeds to search for longer matches.
eg: minimap2

Position	1	2	3	4	5	6	7
Sequence	2	3	1	0	3	4	3
k-mers	2	3	1				
with		3	1	0			
minimizer			1	0	3		
in				**0**	**3**	**4**	
bold					3	4	3
	(a)						

Position	1	2	3	4	5	6	7	8	9	10	11	12
Sequence	4	2	6	4	7	2	8	1	4	7	5	1
k-mers	4	2	6	4	7	2	8					
with		**2**	**6**	**4**	**7**	**2**	**8**	**1**				
minimizer			6	4	7	2	8	1	4			
in				4	7	2	8	1	4	7		
bold					7	2	8	1	4	7	5	
	(b)					2	8	1	4	7	5	1

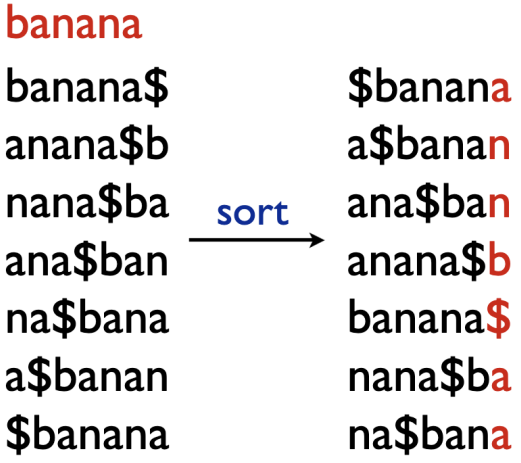
Read mapping Algorithms and Tools

Burrows-Wheeler Transformation (BWT)
- maps genomic sequences using a data compression technique that rearranges characters in a string to group identical ones together, facilitating efficient indexing and alignment

e.g. bwa or Bowtie2

Burrows–Wheeler Transform

Text transform that is useful for compression & search.



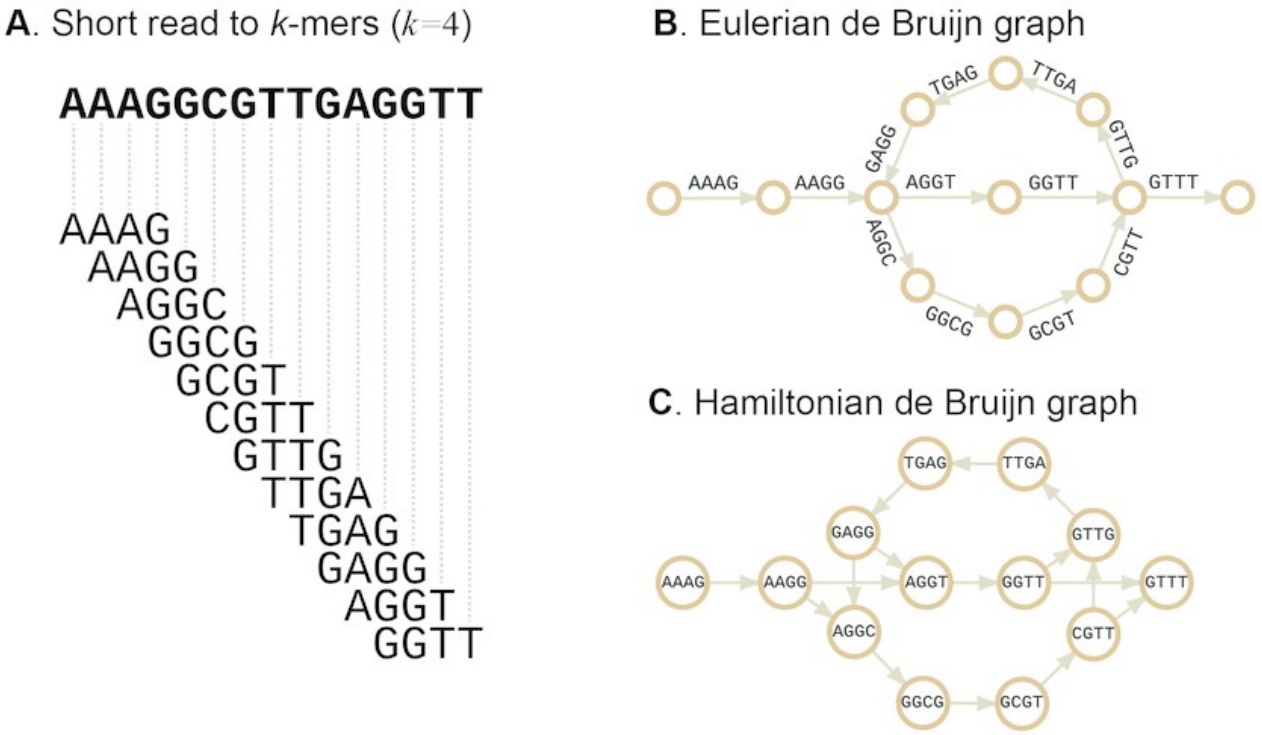
BWT(banana) =
annb$aa

Tends to put runs of the same character together.

Makes compression work well.

“bzip” is based on this.

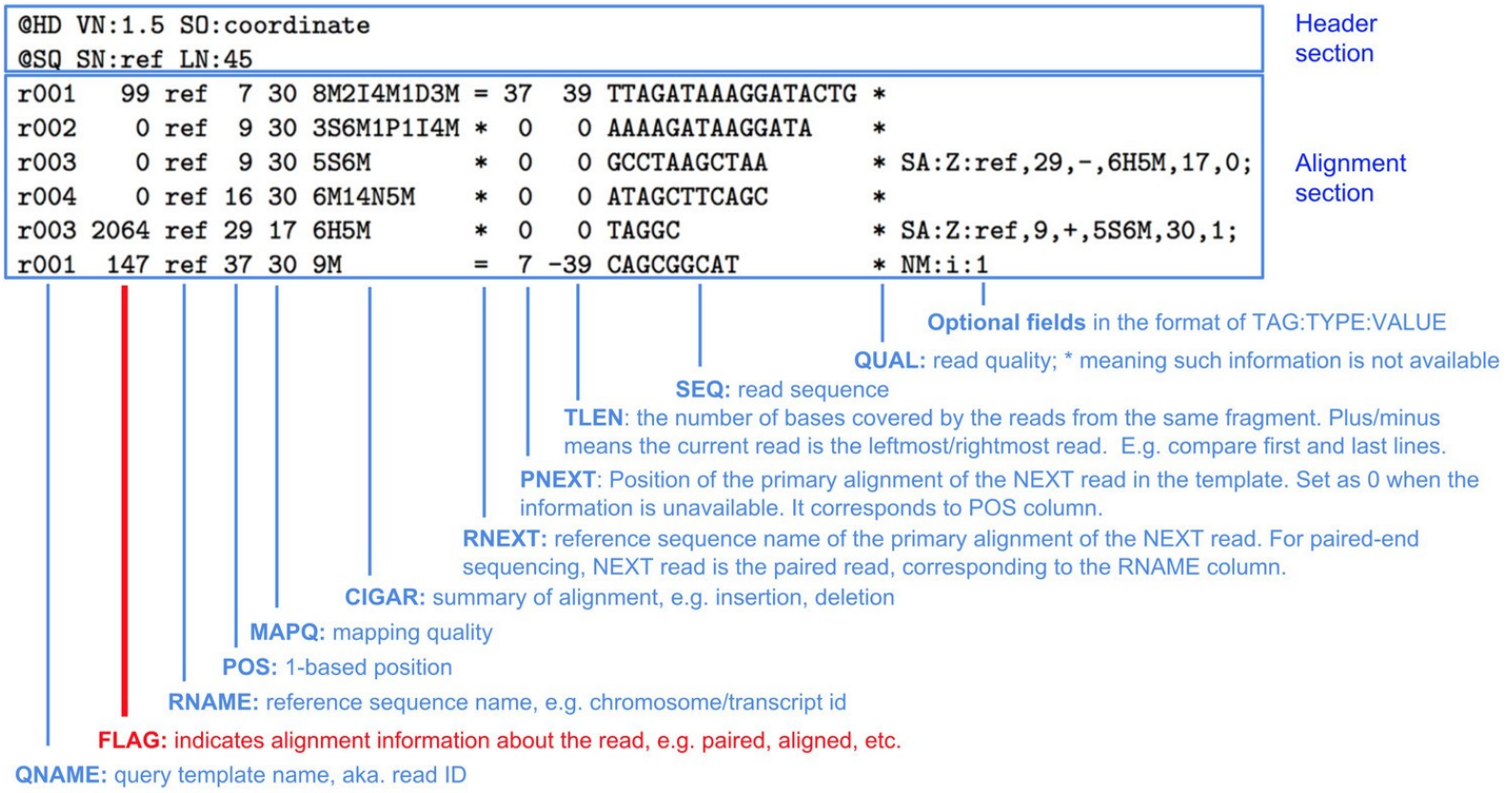
Read mapping tools



Sequence Alignment/Map (SAM)

SAM is a file format used to store information about the alignment of reads to a sequencing genome
Human-readable text files consist of tab-separated columns, with each line representing a single read alignment and each column providing specific information the alignment

SAM file format



https://www.samformat.info/sam-format-flag

Binary Alignment/Map (BAM)

- same as SAM but encoded in binary (e.g. 10110)
- more compact and efficient in terms of storage and processing