Alignment and Variant Calling

Introduction

Why use reference mapping as opposed to de novo assembly?

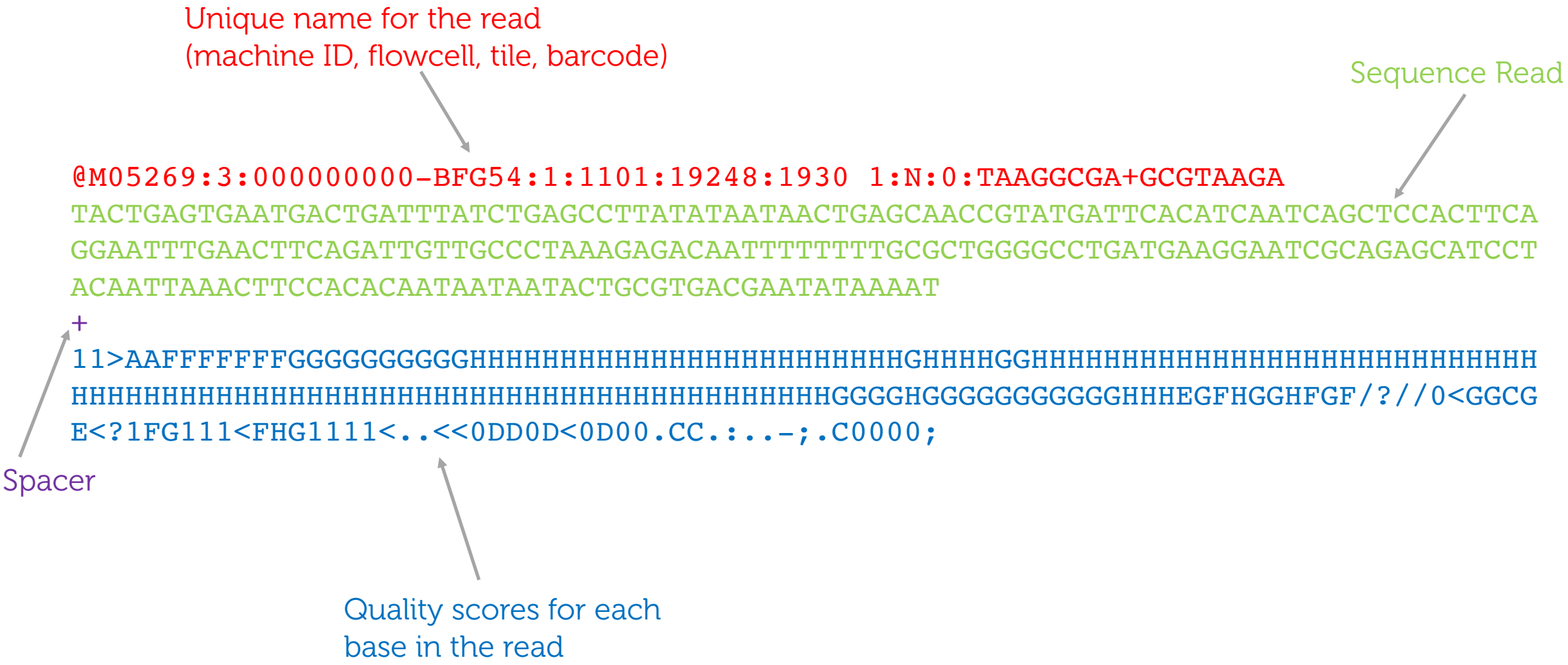
De novo assembly can computationally expensive and difficult to assess quality

Reference mapping is **rapid**, **accurate** and provides a mechanism for analysing sequence variation but **relies on a reference genome** for your organism.

Variants we can detect:

- single nucleotide polymorphisms (SNPs)
- small insertions/deletions (indels)
- structural variation (SVs)

Raw sequence data



ACCCCACAGTACTT

GTACTTATATACAT ATTCACCTGAACCTA

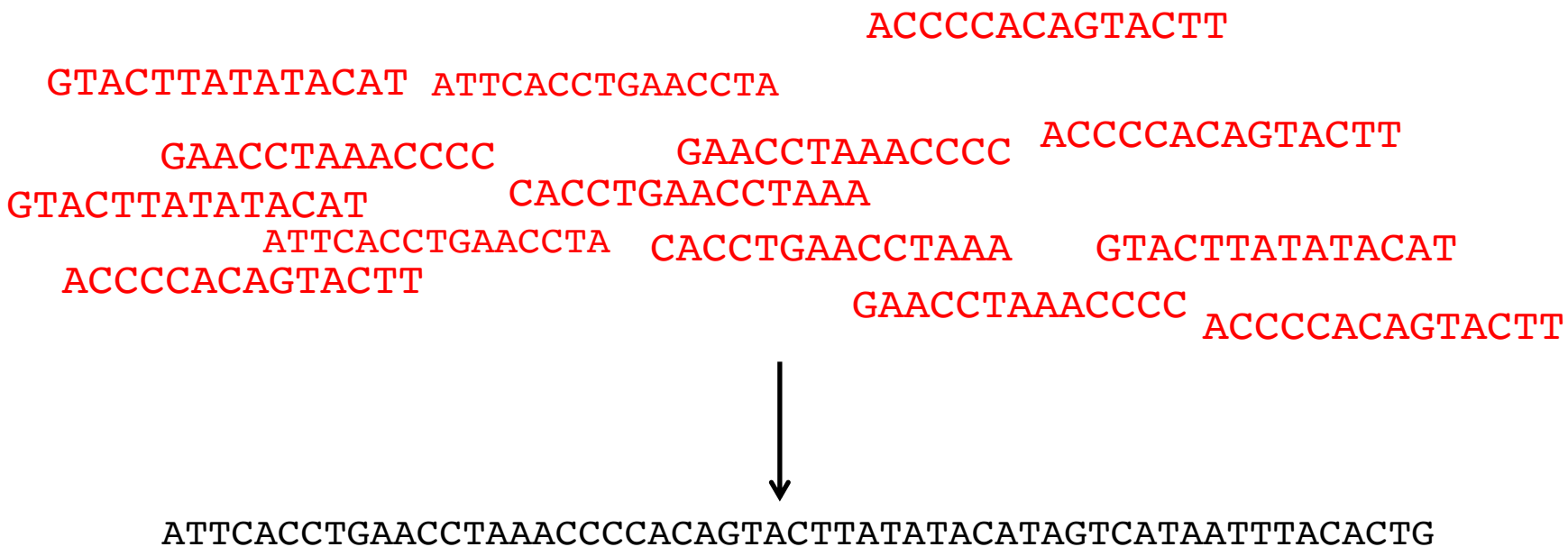
GAACCTAAACCCC GAACCTAAACCCC ACCCCACAGTACTT

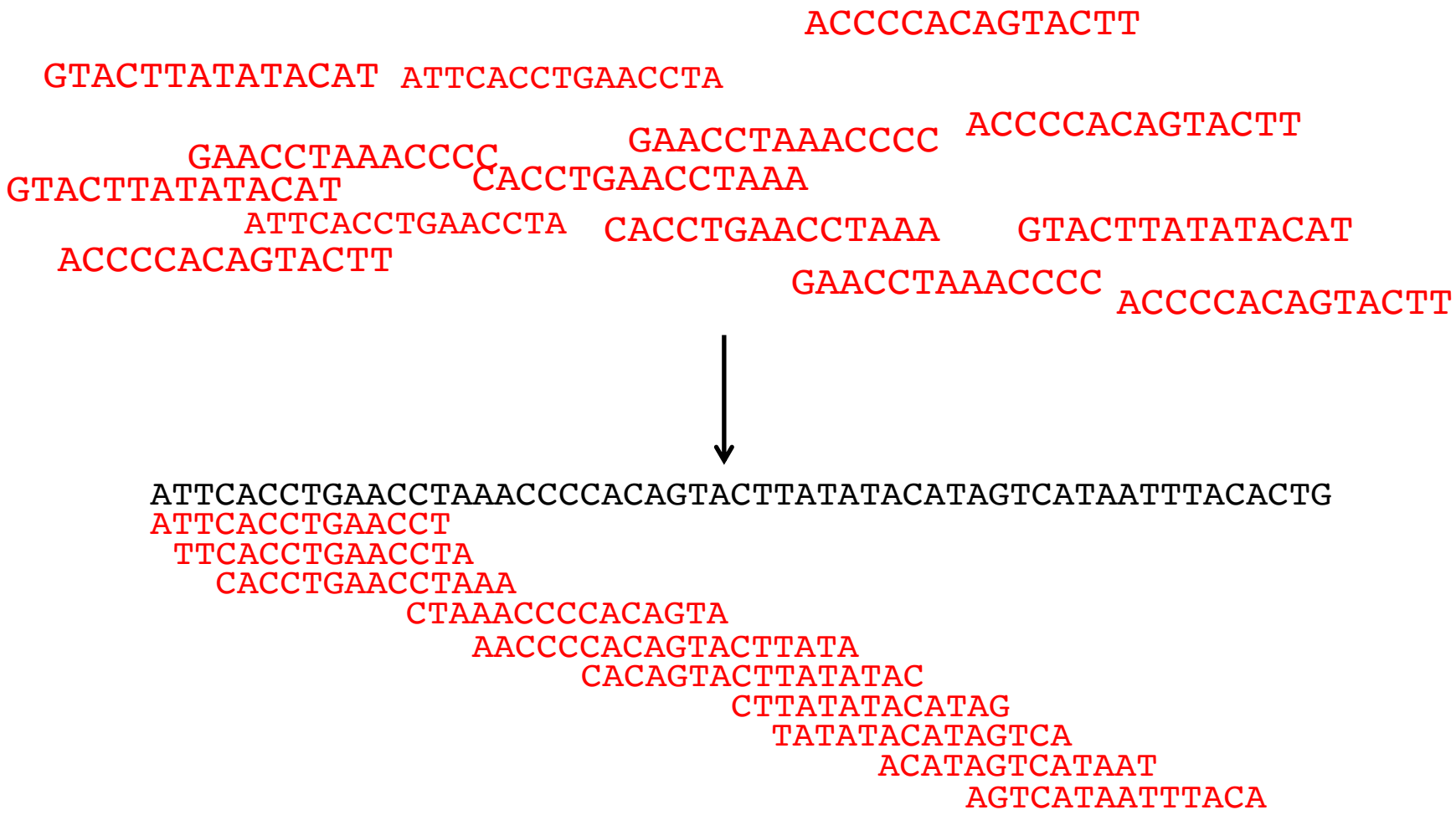
GTACTTATATACAT CACCTGAACCTAAA

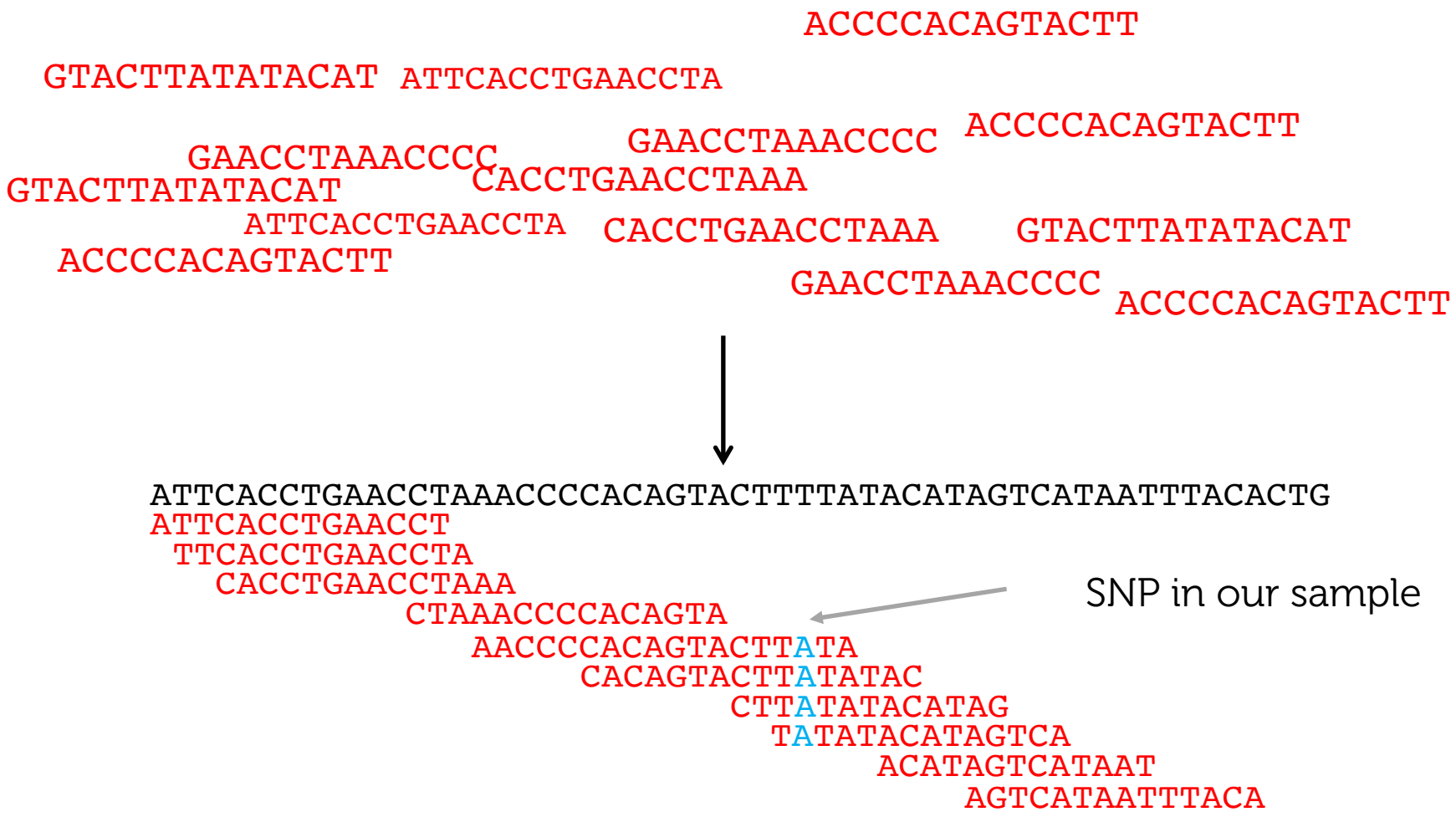
ATTCACCTGAACCTA CACCTGAACCTAAA GTACTTATATACAT

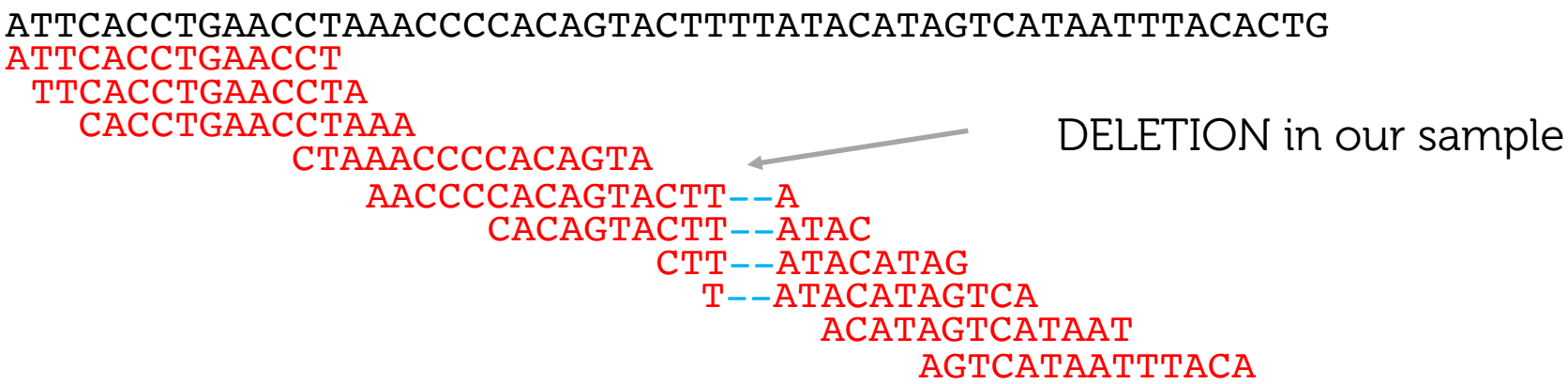
ACCCCACAGTACTT

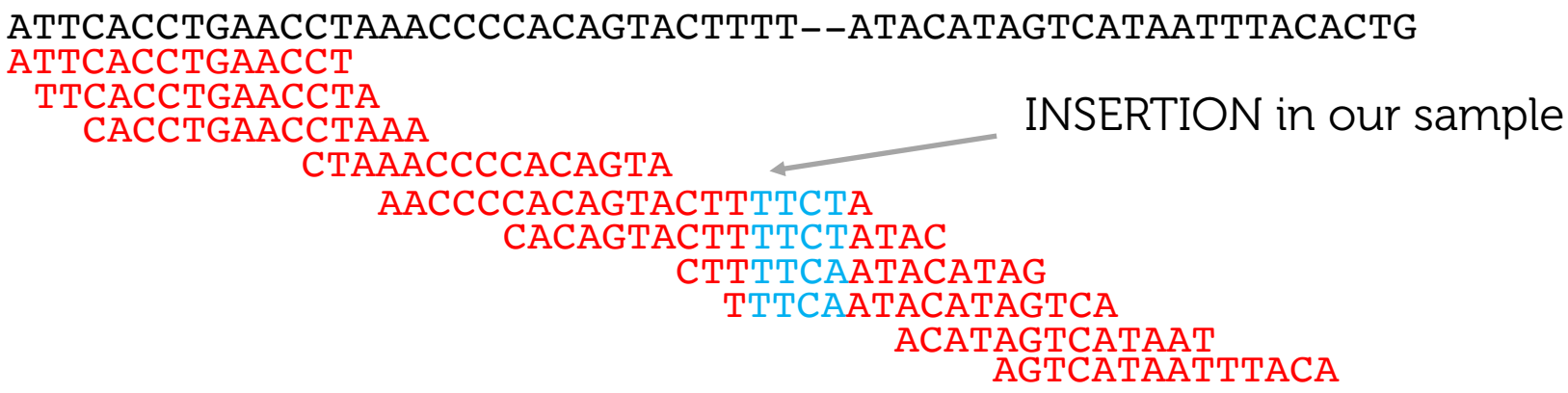
GAACCTAAACCCC ACCCCACAGTACTT


ACCCCACAGTACTT
GTACTTATATACAT
ATTCACCTGAACCTA
GAACCTAAACCCC
GAACCTAAACCCC
ACCCCACAGTACTT
GTACTTATATACAT
CACCTGAACCTAAA
ATTCACCTGAACCTA
CACCTGAACCTAAA
GTACTTATATACAT
ACCCCACAGTACTT
GAACCTAAACCCC
ACCCCACAGTACTT
ATTCACCTGAACCTAAACCCCACAGTACTTATATACATAGTCATAATTTACACTG


ACCCCACAGTACTT
GTACTTATATACAT
ATTCACCTGAACCTA
GAACCTAAACCCC
ACCCCACAGTACTT
GAACCTAAACCCC
CACCTGAACCTAAA
GTACTTATATACAT
ATTCACCTGAACCTA
CACCTGAACCTAAA
GTACTTATATACAT
ACCCCACAGTACTT
GAACCTAAACCCC
ACCCCACAGTACTT
ATTCACCTGAACCTAAACCCCACAGTACTTATATACATAGTCATAATTTACACTG
ATTCACCTGAACCT
TTCACCTGAACCTA
CACCTGAACCTAAA
CTAAACCCCACAGTA
AACCCCACAGTACTTATA
CACAGTACTTATATAC
CTTATATACATAG
TATATACATAGTCA
ACATAGTCATAAT
AGTCATAATTTACA


ACCCCACAGTACTT
GTACTTATATACAT
ATTCACCTGAACCTA
GAACCTAAACCCC
ACCCCACAGTACTT
GAACCTAAACCCC
CACCTGAACCTAAA
GTACTTATATACAT
ATTCACCTGAACCTA
CACCTGAACCTAAA
GTACTTATATACAT
ACCCCACAGTACTT
GAACCTAAACCCC
ACCCCACAGTACTT
ATTCACCTGAACCTAAACCCCACAGTACTTTTATACATAGTCATAATTTACACTG
ATTCACCTGAACCT
TTCACCTGAACCTA
CACCTGAACCTAAA
CTAAACCCCACAGTA
AACCCCACAGTACTTATA
CACAGTACTTATATAC
CTTATATACATAG
TATATACATAGTCA
ACATAGTCATAAT
AGTCATAATTTACA
SNP in our sample


ATTCACCTGAACCTAAACCCCACAGTACTTTTATACATAGTCATAATTTACACTG
ATTCACCTGAACCT
TTCACCTGAACCTA
CACCTGAACCTAAA
CTAAACCCCACAGTA
AACCCCACAGTACTT--A
CACAGTACTT--ATAC
CTT--ATACATAG
T--ATACATAGTCA
ACATAGTCATAAT
AGTCATAATTTACA
DELETION in our sample


ATTCACCTGAACCTAAACCCCACAGTACTTTT--ATACATAGTCATAATTTACACTG
ATTCACCTGAACCT
TTCACCTGAACCTA
CACCTGAACCTAAA
CTAAACCCCACAGTA
AACCCCACAGTACTTTTCTA
CACAGTACTTTTCTATAC
CTTTTCAATACATAG
TTTCAATACATAGTCA
ACATAGTCATAAT
AGTCATAATTTACA
INSERTION in our sample

Alignment Algorithms

- Finding a match for your read in the genome can be slow
- Alignment algorithms make a special genome index for fast mapping
- Very fast for exact matches
- Slower for inexact matches with multiple possible locations
- Main alignment software you will use is BWA-MEM

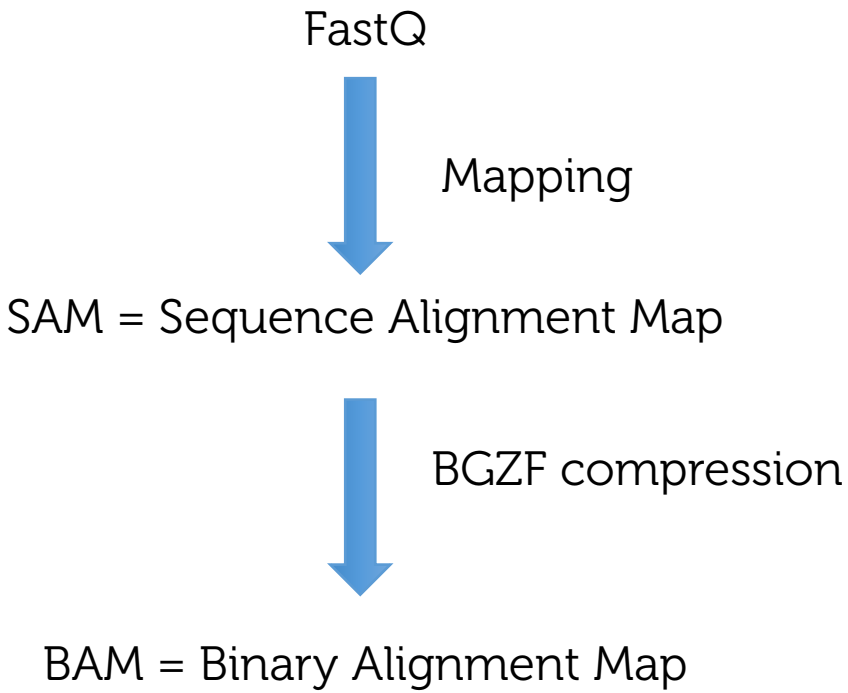
Mapping algorithms

- BWA-MEM & Bowtie2

- Perform alignment of single or paired end reads
- Generally recommended for high quality queries as they are faster and more accurate than predecessors.

Analysing data: alignment

WHAM, BAM?, SAM?


FastQ
Mapping
SAM = Sequence Alignment Map
BGZF compression
BAM = Binary Alignment Map

SAM Format

All short read aligners generate text based SAM output.

SAM files contain detailed information regarding the alignment results per read including mapping quality/Phred scores, mismatches/substitutions/indels (cigar line).

All this information is compressed into a BAM file and then indexed significantly reducing file size and allowing for rapid processing at the variant detection stage.

SAM Format

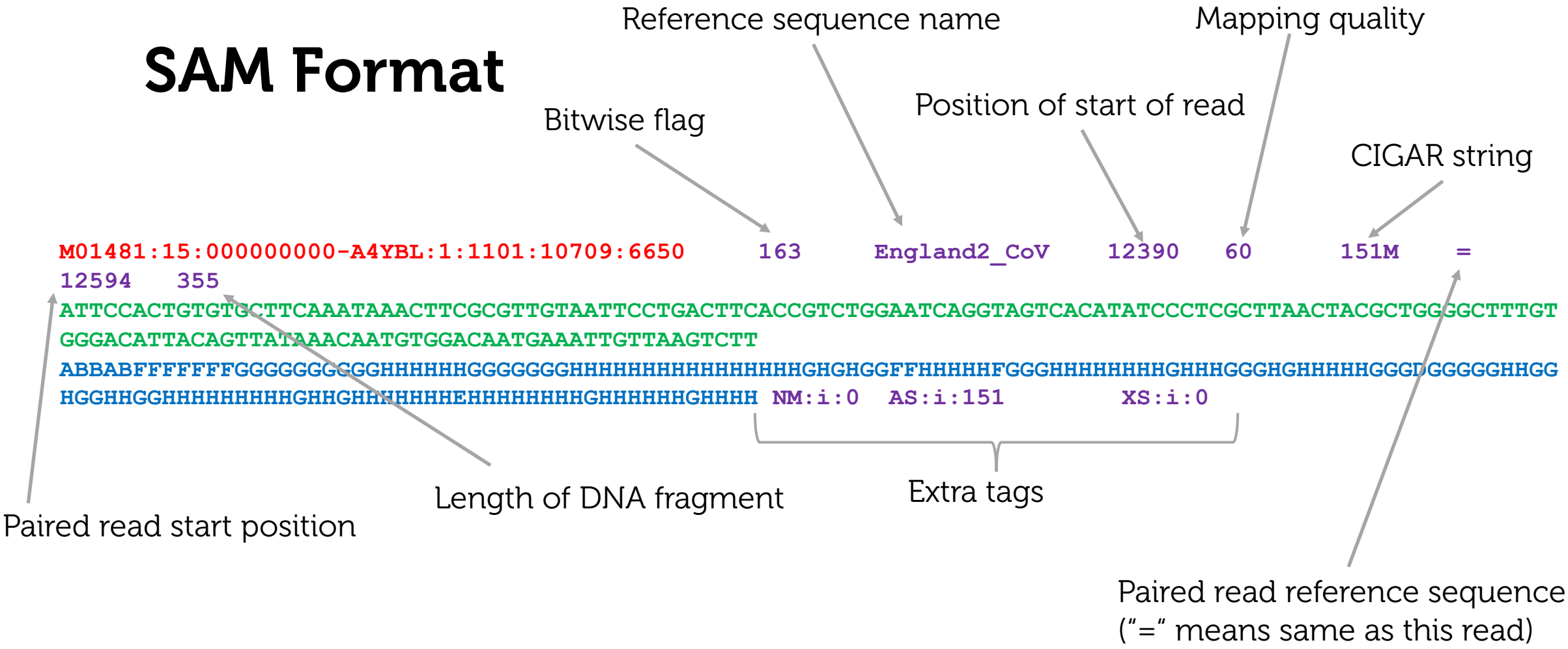
Unique name for the read
(machine ID, flowcell, tile, barcode)

Sequence Read

```
M01481:15:000000000-A4YBL:1:1101:10709:6650     163     England2_CoV    12390   60      151M    =
12594   355
ATTCCACTGTGTGCTTCAAATAAACTTCGCGTTGTAATTCCTGACTTCACCGTCTGGAATCAGGTAGTCACATATCCCTCGCTTAACTACGCTGGGGCTTTGT
GGGACATTACAGTTATAAACAATGTGGACAATGAAATTGTTAAGTCTT
ABBABFFFFFFFGGGGGGGGGGHHHHHHHGGGGGGGHHHHHHHHHHHHHHHHHGHGHGGFFHHHHHFGGGHHHHHHHHHGHHHGGGHGHHHHHGGGDGGGGGHHGG
HGGHHGGHHHHHHHHHHGHHGHHHHHHHEHHHHHHHHGHHHHHHGHHHH NM:i:0  AS:i:151        XS:i:0
```

Quality scores for each
base in the read

SAM Format



Google 'SAM Format specification' or go to https://github.com/samtools/hts-specs

Bitwise flag

numeric	binary	description
1	00000001	template has multiple fragments in sequencing
2	00000010	each fragment properly mapped according to aligner
4	00000100	fragment is unmapped
8	00001000	mate is unmapped
16	00010000	sequence is reverse complemented
32	00100000	sequence of mate is reversed
64	01000000	is first fragment in template
128	10000000	is second fragment in template

Decoding SAM flags

This utility makes it easy to identify what are the properties of a read based on its SAM flag value, or conversely, to find what the SAM Flag value would be for a given combination of properties.

To decode a given SAM flag value, just enter the number in the field below. The encoded properties will be listed under Summary below, to the right.

SAM Flag: 163 Explain

Switch to mate Toggle first in pair / second in pair

Find SAM flag by property:

To find out what the SAM flag value would be for a given combination of properties, tick the boxes for those that you'd like to include. The flag value will be shown in the SAM Flag field above.

- [x] read paired
- [x] read mapped in proper pair
- [ ] read unmapped
- [ ] mate unmapped
- [ ] read reverse strand
- [x] mate reverse strand
- [ ] first in pair
- [x] second in pair
- [ ] not primary alignment
- [ ] read fails platform/vendor quality checks
- [ ] read is PCR or optical duplicate
- [ ] supplementary alignment

Summary:

read paired (0x1)
read mapped in proper pair (0x2)
mate reverse strand (0x20)
second in pair (0x80)

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

The CIGAR string

- Compress the read alignment into a short easy-to-parse format
- M – match
- I – insertion (base in the read, not the reference)
- D – deletion (base in the reference, not in the read)

```
Reference: CCCTACGTCCCAGTC-AC
             CTACGTCCCAG        11M
              TAC--CAC          3M2D3M
                    CCAGTCAAC   6M1I2M
```

The CIGAR line

```
M01481:15:000000000-A4YBL:1:1101:10709:6650     163     England2_CoV    12390   60  127M1D5I17M  =
12594   355
ATTCCACTGTGTGCTTCAAATAAACTTCGCGTTGTAATTCCTGACTTCACCGTCTGGAATCAGGTAGTCACATATCCCTCGCTTAACTACGCTGGGGCTTTG
TGGGACATTACAGTTATAAACAATGTGGACAATGAAATTGTTAAGTCTT
ABBABFFFFFFFGGGGGGGGGGHHHHHHHGGGGGGGHHHHHHHHHHHHHHHHHGHGHGGFFHHHHHFGGGHHHHHHHHGHHHGGGHGHHHHHGGGDGGGGGHHG
GHGGHHGGHHHHHHHHHGHHGHHHHHHHEHHHHHHHHGHHHHHHGHHHH NM:i:0  AS:i:151        XS:i:0
```

127 Matches,1 Deletion, 5 Insertions, 17 Matches when compared to the reference sequence.

Error probabilities

M01481:15:000000000-A4YBL:1:1101:10709:6650 163 England2_CoV 12390 60 151M = 12594 355
ATTCCACTGTGTGCTTCAAATAAACTTCGCGTTGTAATTCCTGACTTCACCGTCTGGAATCAGGTAGTCACATATCCCTCGCTTAACTACGCTGGGGCTT
TGTGGGACATTACAGTTATAAACAATGTGGACAATGAAATTGTTAAGTCTT
ABBABFFFFFFFGGGGGGGGGGHHHHHHHGGGGGGGHHHHHHHHHHHHHHHHHGHGHGGFFHHHHHFGGGHHHHHHHHGHHHGGGHGHHHH
HGGGDGGGGGHHGGHGGHHGGHHHHHHHHHGHHGHHHHHHHEHHHHHHHHGHHHHHHGHHHH NM:i:0 AS:i:151 XS:i:0

Phred quality scores are logarithmically linked to error probabilities

Phred quality score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%
60	1 in 1000000	99.9999%

Issues with alignment

- Repeat regions
  - Read may align equally well to multiple regions
  - Paired end reads have distance information which is also weighted
  - Alignment between the read and true source in the genome may have more differences than alignment with a repeat (read will be misplaced)

- Choice of reference genome
  - There may be many nucleotide differences between the reads and the most closely related reference genome (a significant problem in HIV and HCV viral NGS mapping)
  - Reads are easier to align with fewer variants
  - Leads to bias in alignment/variant calling

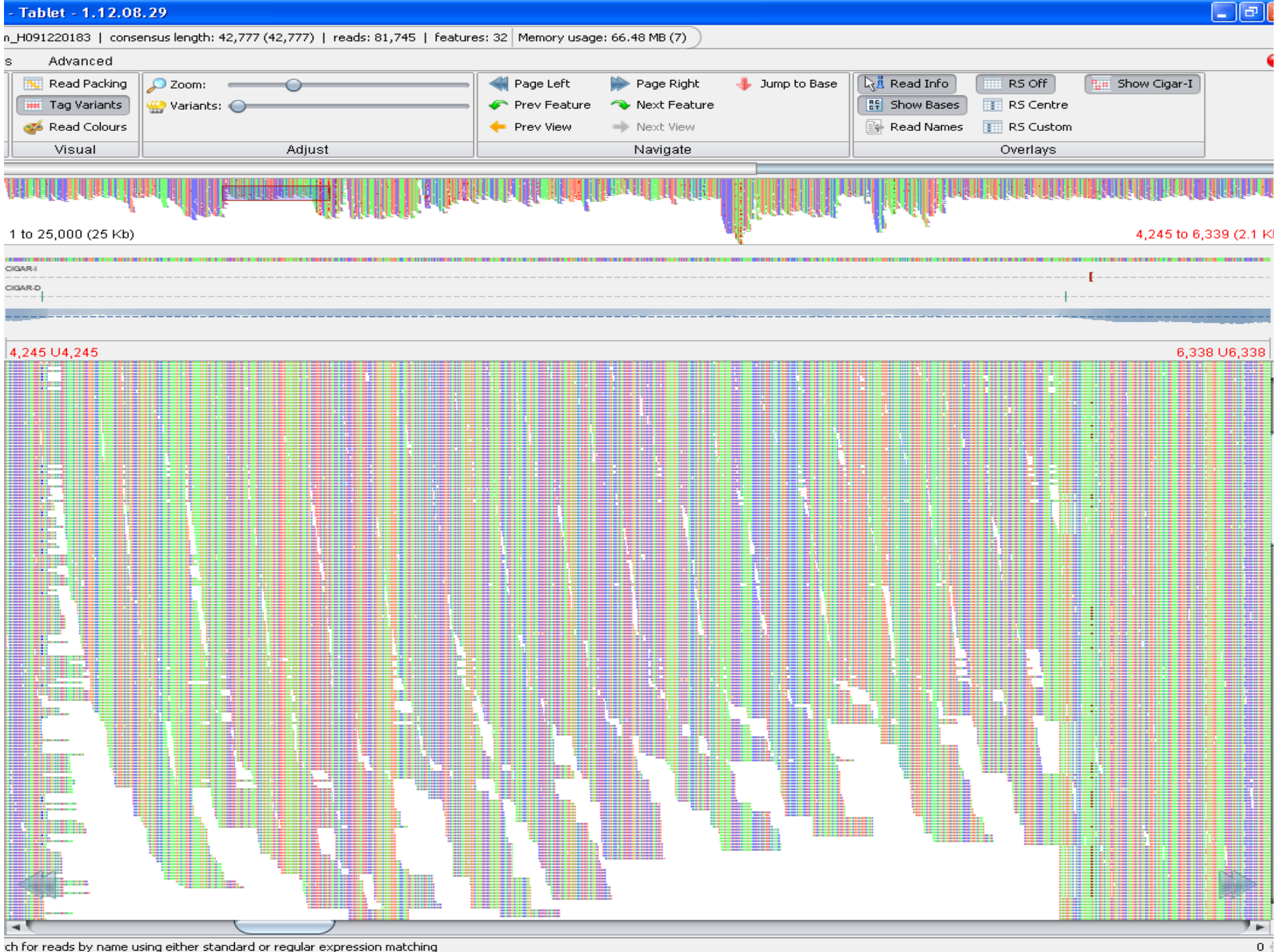
- Alignment of long reads with high error rate is difficult

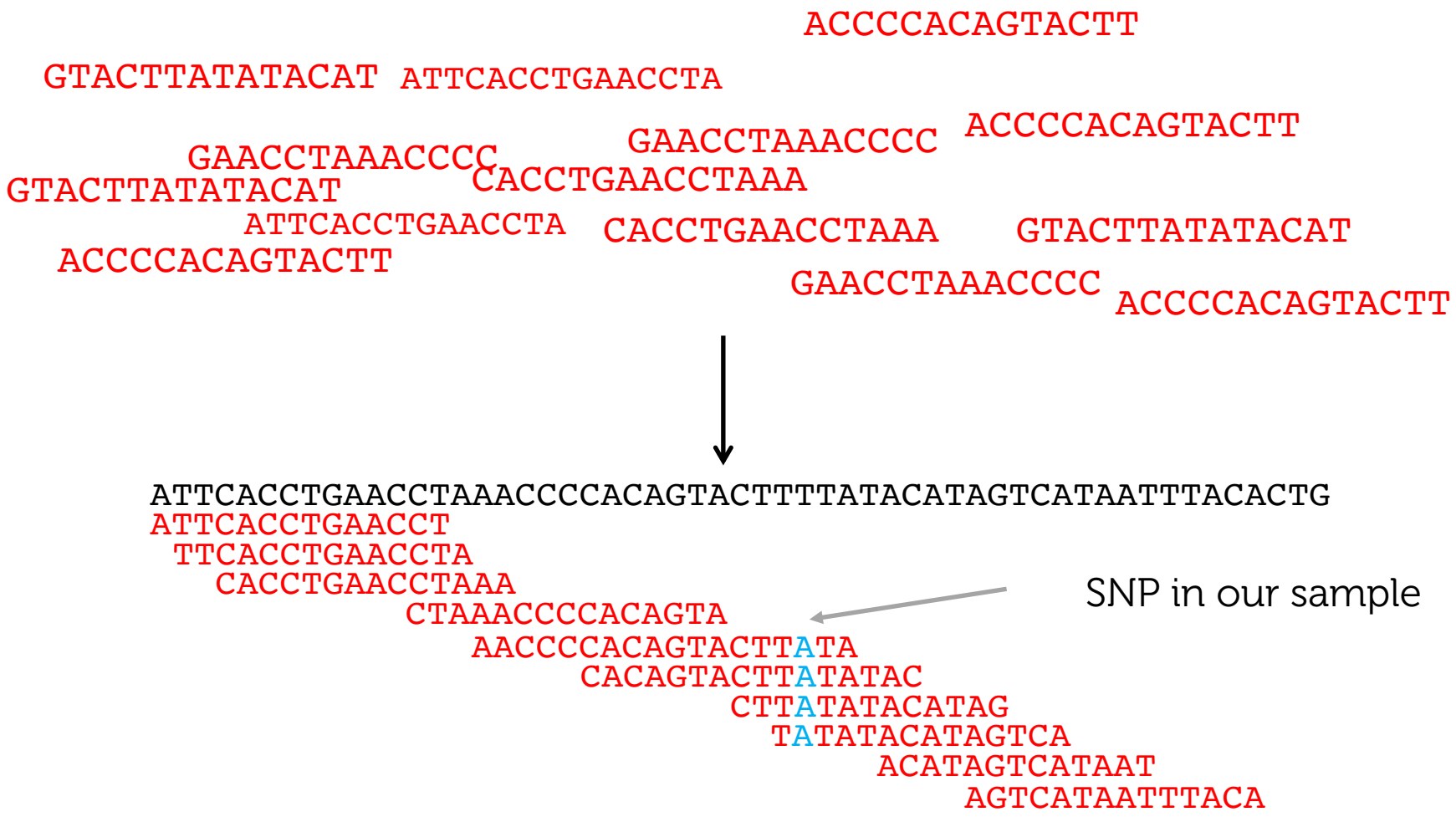
Visualisation

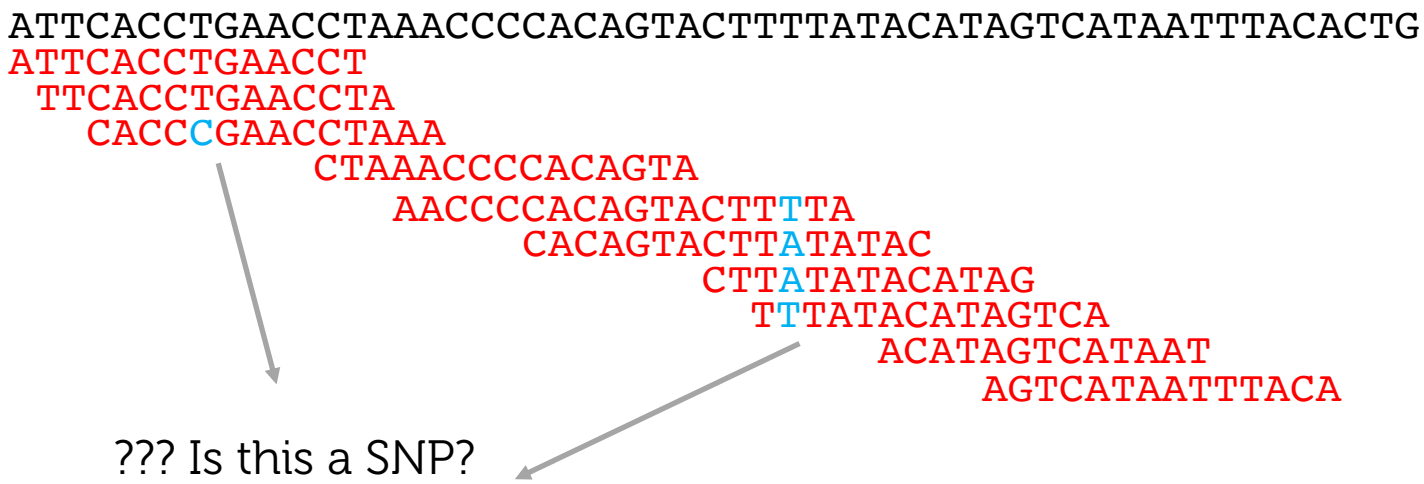
- Many editors exist for visualising NGS data enabling you to view the read pileup.

- Tablet is a popular lightweight editor requiring the sorted BAM, the index file (BAI) and the reference genome in FASTA format.

- Variation can be graphically viewed across the assembled reads.


- Tablet - 1.12.08.29
n_H091220183 | consensus length: 42,777 (42,777) | reads: 81,745 | features: 32 | Memory usage: 66.48 MB (7)
s Advanced
Read Packing
Tag Variants
Read Colours
Visual
Zoom:
Variants:
Adjust
Page Left
Page Right
Jump to Base
Prev Feature
Next Feature
Prev View
Next View
Navigate
Read Info
RS Off
Show Cigar-I
Show Bases
RS Centre
Read Names
RS Custom
Overlays
1 to 25,000 (25 Kb)
4,245 to 6,339 (2.1 K
CIGAR-I
CIGAR-D
4,245 U4,245
6,338 U6,338
ch for reads by name using either standard or regular expression matching
0

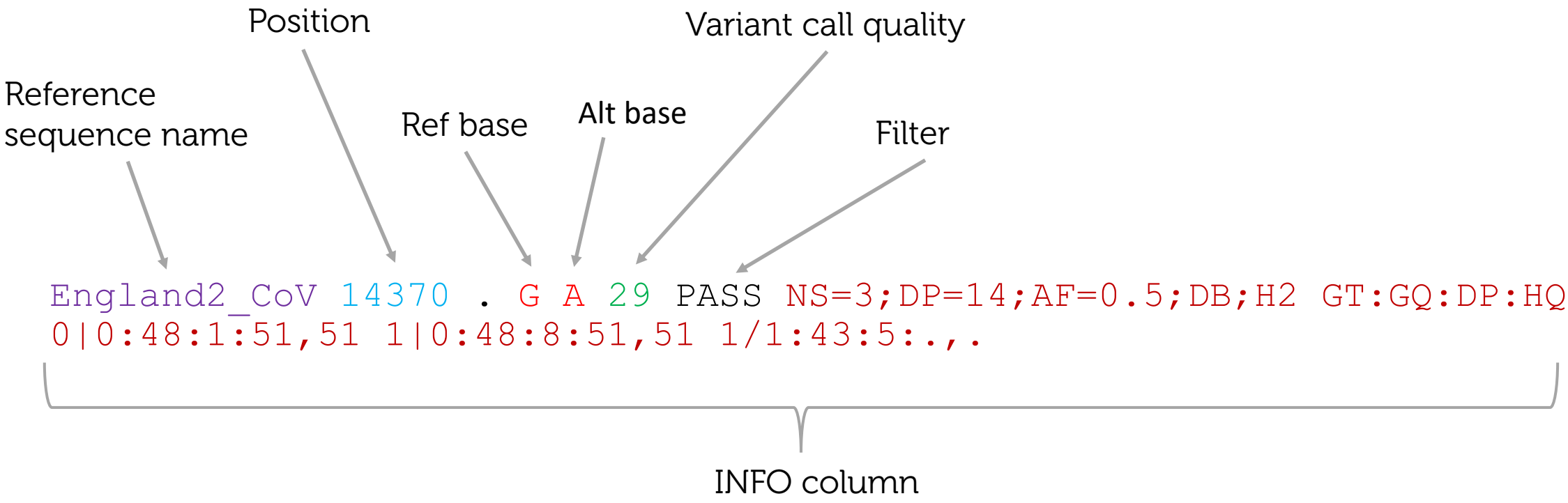

ACCCCACAGTACTT
GTACTTATATACAT
ATTCACCTGAACCTA
GAACCTAAACCCC
ACCCCACAGTACTT
GAACCTAAACCCC
CACCTGAACCTAAA
GTACTTATATACAT
ATTCACCTGAACCTA
CACCTGAACCTAAA
GTACTTATATACAT
ACCCCACAGTACTT
GAACCTAAACCCC
ACCCCACAGTACTT
ATTCACCTGAACCTAAACCCCACAGTACTTTTATACATAGTCATAATTTACACTG
ATTCACCTGAACCT
TTCACCTGAACCTA
CACCTGAACCTAAA
CTAAACCCCACAGTA
AACCCCACAGTACTTATA
CACAGTACTTATATAC
CTTATATACATAG
TATATACATAGTCA
ACATAGTCATAAT
AGTCATAATTTACA
SNP in our sample


ATTCACCTGAACCTAAACCCCACAGTACTTTTATACATAGTCATAATTTACACTG
ATTCACCTGAACCT
TTCACCTGAACCTA
CACCCGAACCTAAA
CTAAACCCCACAGTA
AACCCCACAGTACTTTTA
CACAGTACTTATATAC
CTTATATACATAG
TTTATACATAGTCA
ACATAGTCATAAT
AGTCATAATTTACA
??? Is this a SNP?

Filtering variant calls

- How can we tell if the variant is real?
- Require a minimum number of reads with the variant
- Require a minimum quality of the reads with the variant
- Require a minimum proportion of reads with the variant

Variant Call Format



Variant Call Format (VCF) INFO column

```
AB=0;ABP=0;AC=1;AF=1;AN=1;AO=125;DP=125;DPB=125;
LEN=1;MQM=60;MQMR=0;NS=1;TYPE=snp
```

- ‘NS’ is number of samples
- ‘DP’ is depth of reads
- ‘TYPE’ is type of variant
- All columns are defined at the top of the file

Variant Call Format (VCF) FORMAT column

```
GT:DP:AD:RO:QR:AO:QA:GL
1:125:0,125:0:0:125:4434:-399.193,0
```

- ‘GT’ is genotype. 0 is ref, 1 is alt
- ‘DP’ is depth (125)
- ‘GL’ is genotype likelihood for ref and alt, highest is best

Variant Annotation

- What does this variant actually do?
- Requires a reference genome file with gene features