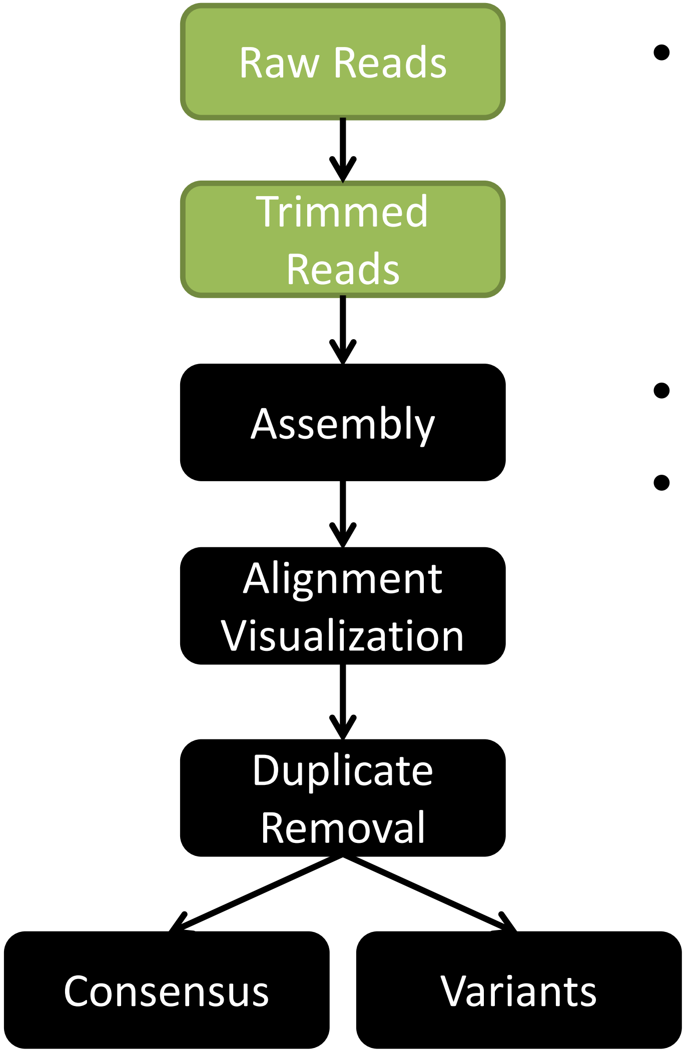
Reference Based Assembly

Why do we assemble reads?

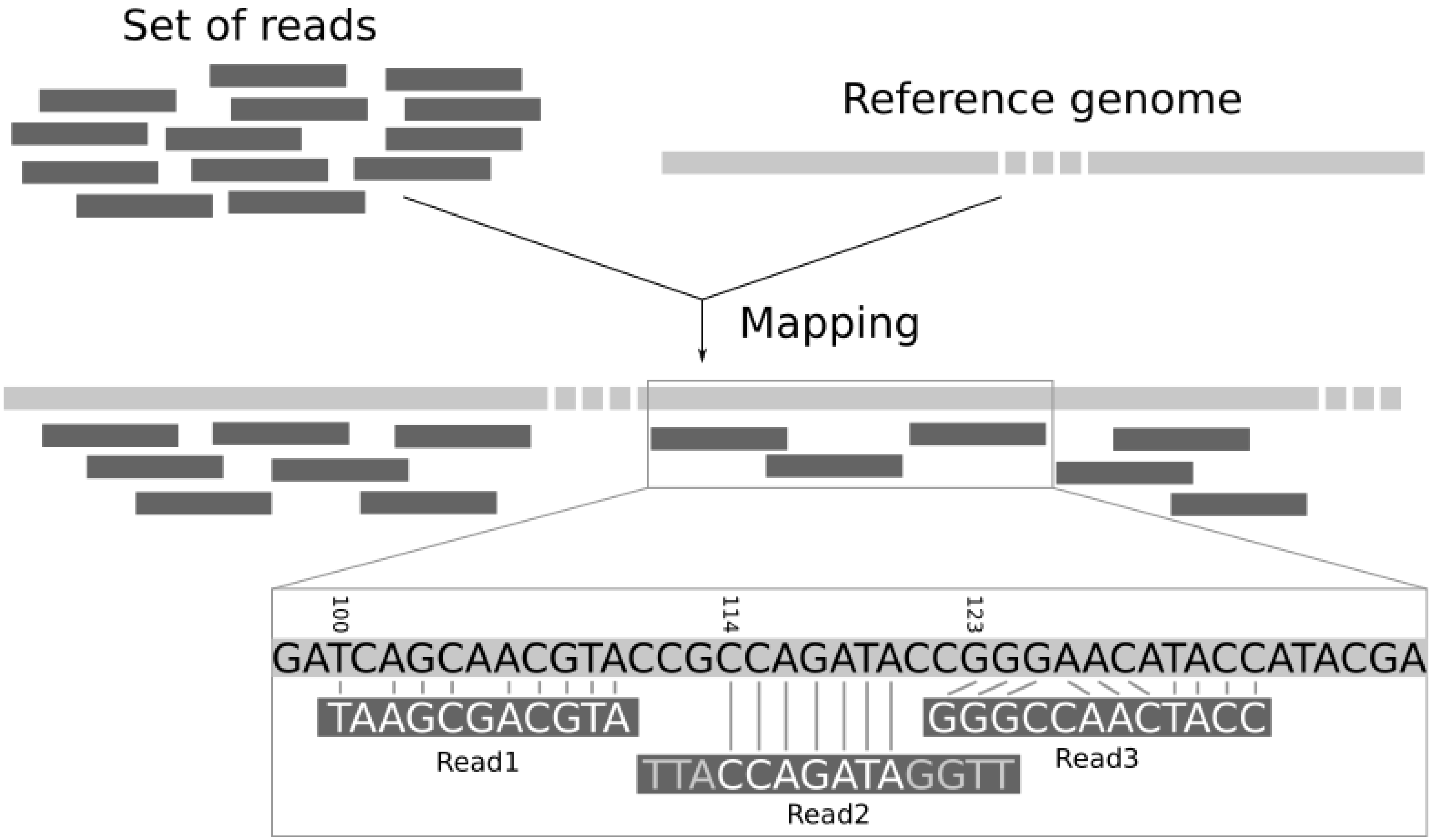
- Second Generation sequencing fragments genomic material which is then sequenced to give the reads
- We assemble these reads to identify the exact position of the genome they come from and their base-to-base correspondence.
- The two main methods of assembly are
  - ➢ Reference Bases Alignment
  - ➢ Denovo Assembly

Reference

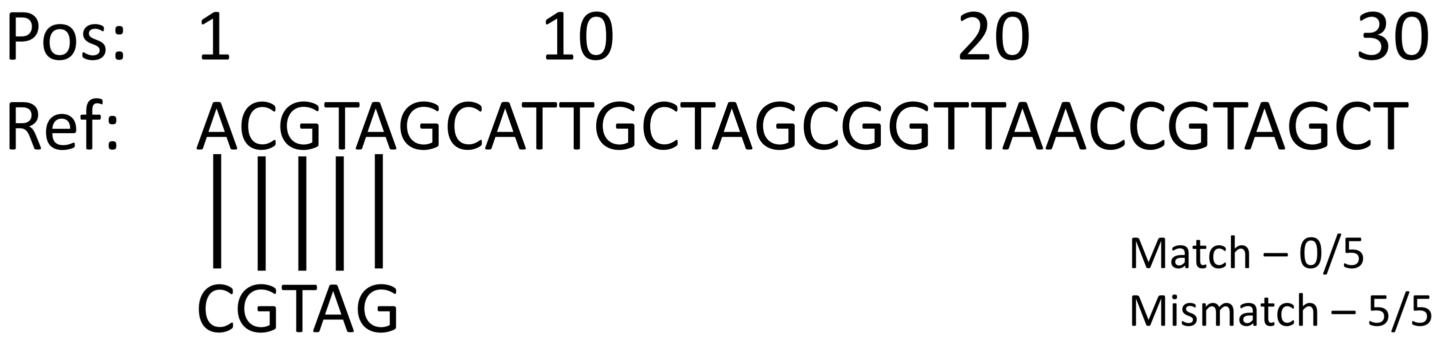


- All we need for reference alignment are the trimmed reads and a reference sequence.

- What is a suitable reference?
- Can you map any reads to any reference?



https://galaxyproject.github.io/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html



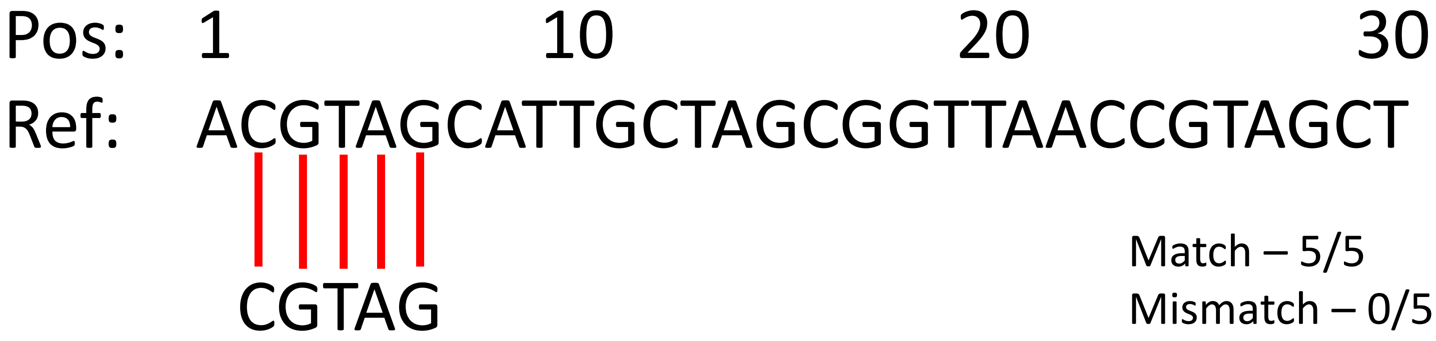
Pos: 1 10 20 30

Ref: ACGTAGCATTGCTAGCGGTTAACCGTAGCT

CGTAG

Match – 0/5

Mismatch – 5/5



Pos: 1 10 20 30

Ref: ACGTAGCATTGCTAGCGGTTAACCGTAGCT

CGTAG

Match – 5/5

Mismatch – 0/5

Perfect Alignment

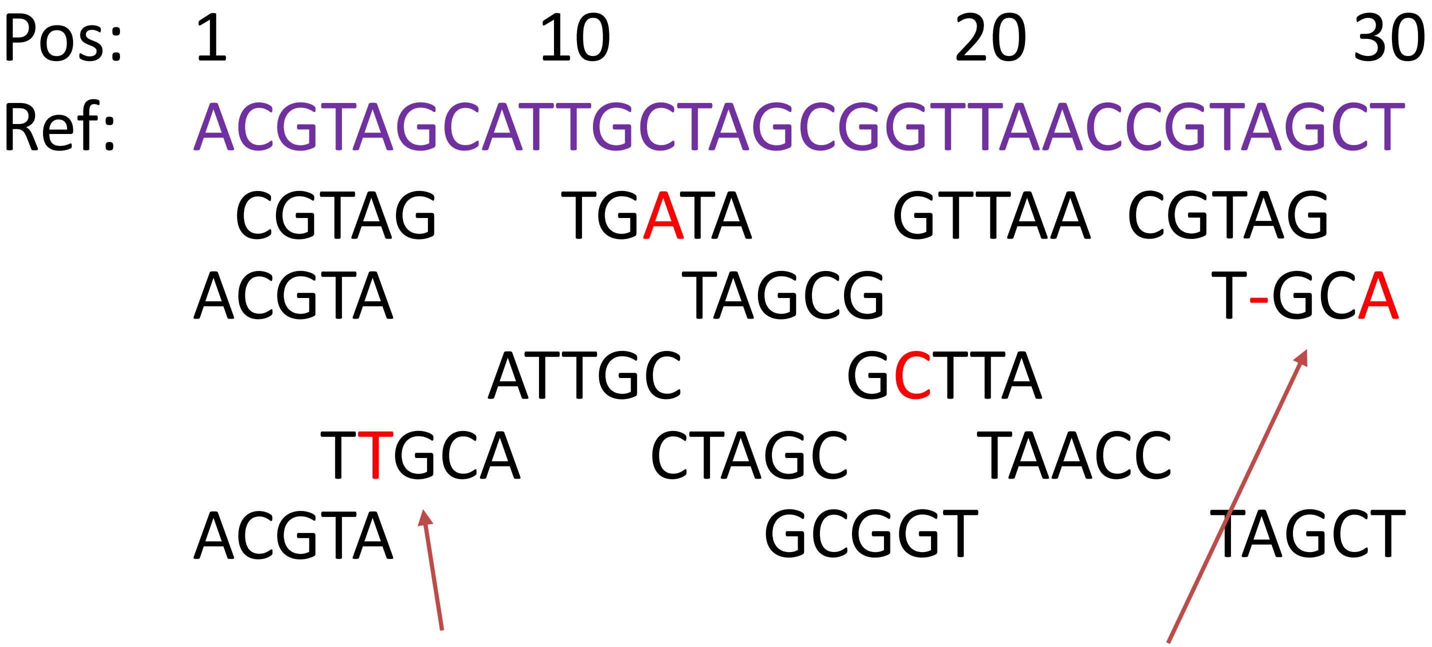
```
Pos:  1       10        20        30
Ref:  ACGTAGCATTGCTAGCGGTTAACCGTAGCT
       CGTAG   TGCTA   GTTAA CGTAG
      ACGTA       TAGCG        TAGCT
             ATTGC    GGTTA
         TAGCA   CTAGC   TAACC
      ACGTA         GCGGT      TAGCT
```

Viral Sequencing

- Viruses exist as a quasispecies distribution which introduces variants
- Variants can also be introduced by PCR errors and poor-quality sequencing
- Variants can be identified from read alignments

```
Reference CCGTTAGAGTTACAATTCGA
Read 2       TTAGAGTAACAA
Read 3    CCGTTAGAGTTA
Read 4             TTACAATTCGA
Read 5          GAGTAACAA
Read 6       TTAGAGTAACAAT
```

Viral Alignment



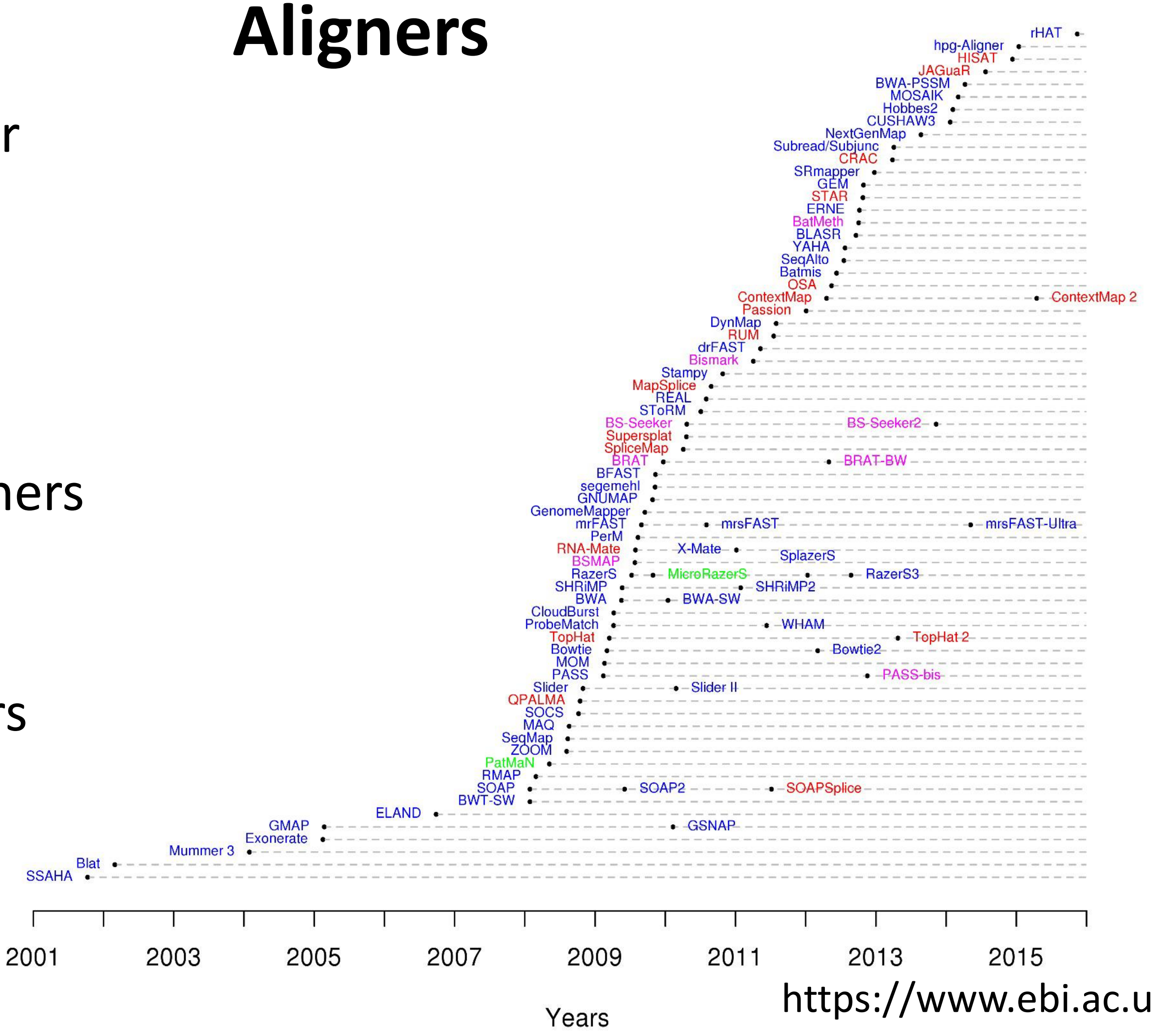
```
Pos:  1       10        20        30
Ref:  ACGTAGCATTGCTAGCGGTTAACCGTAGCT
       CGTAG   TGATA   GTTAA CGTAG
      ACGTA       TAGCG        T-GCA
             ATTGC    GCTTA
         TTGCA   CTAGC   TAACC
      ACGTA         GCGGT      TAGCT
```

CIGAR: Pos4 1M1X3M

CIGAR: Pos 26 1M1D2M1X

Aligners

- Short Read Aligner
  - BWA
  - Bowtie
  - Tanoti
  - NovoAlign
- Splice Aware Aligners
  - Tophat
  - BBMap
- Long Read Aligners
  - Minimap
  - LAST



https://www.ebi.ac.uk

Which Aligner?

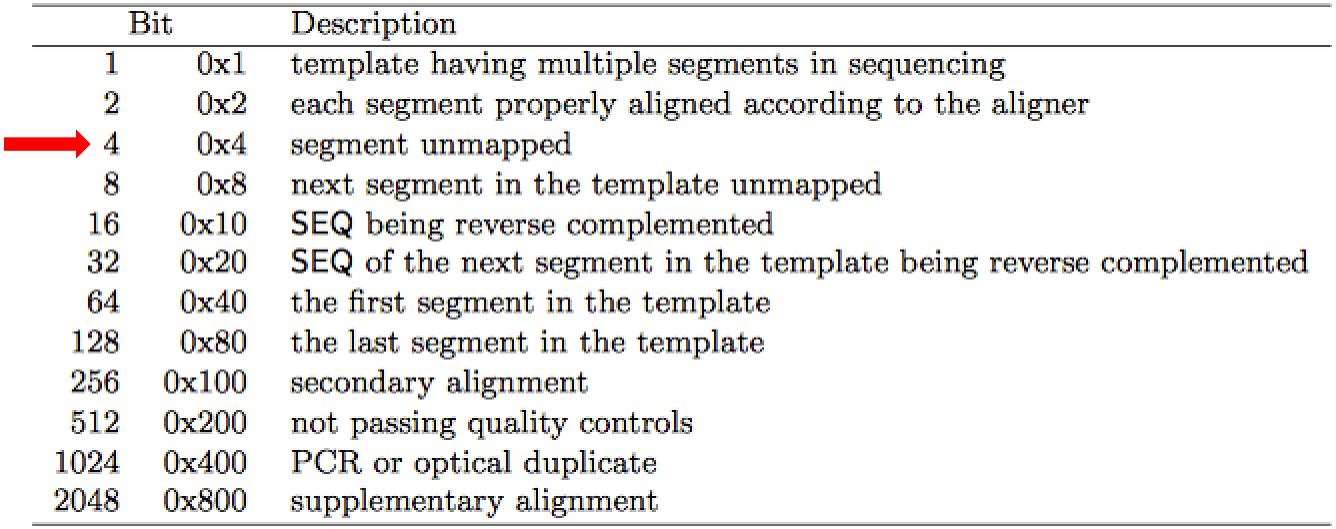
- Hash based Aligners
- Burrows Wheeler Aligners
- Global vs Local Aligners
- Effect on Consensus Calling
- Effect on Variant Calling

Always good to keep aligners consistent across any experiment

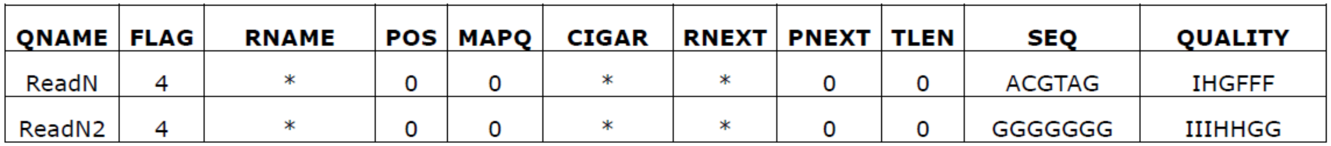
Which Reference?

- Choosing the wrong species
- Choosing the wrong genotypes (divergent sequences)
- If you do not have information on correct reference may have to map to panel of references
- If species not known Denovo Assembly

Identifying incorrect Reference



Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

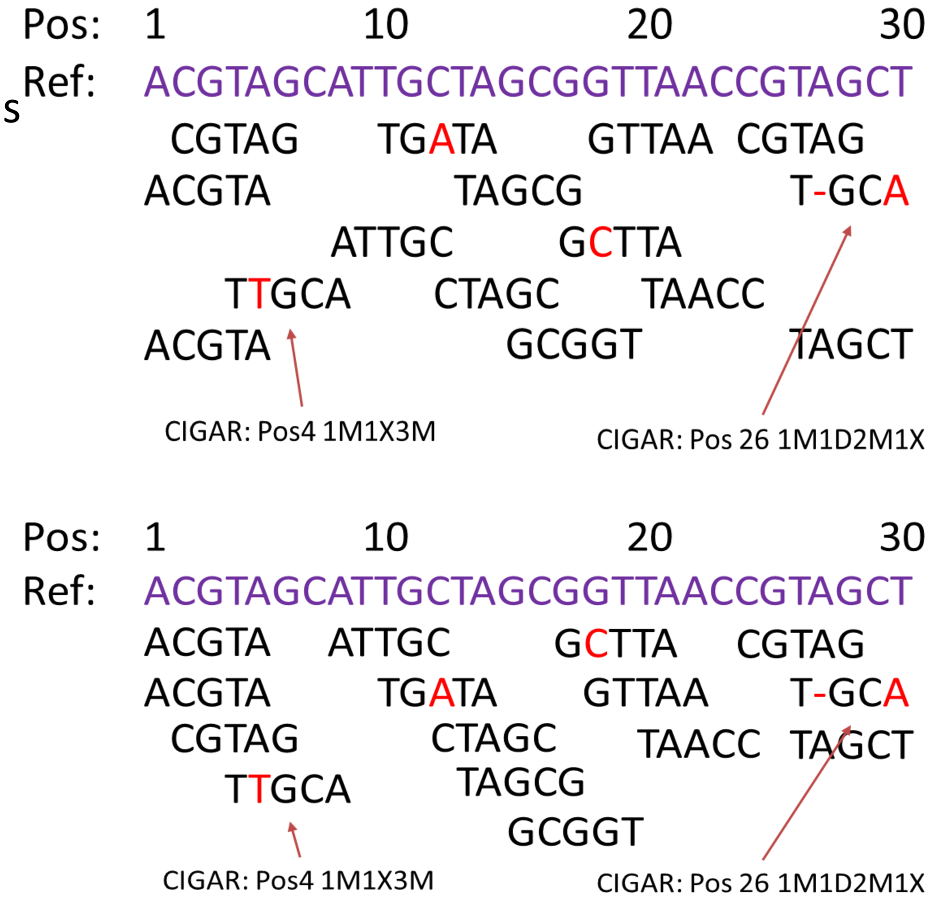


QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUALITY
ReadN	4	*	0	0	*	*	0	0	ACGTAG	IHGFFF
ReadN2	4	*	0	0	*	*	0	0	GGGGGGG	IIIHHGG

- Looking at assembly stats
- Visually inspecting the alignment

Sequence Alignment

- Usually output as SAM file
  - Sequence Alignment Map contains coordinates and quality of mapping.
- Converted to BAM file
  - Binary Alignment Map is compressed
- Files are sorted and indexed
  - Makes it faster to access and work with the data

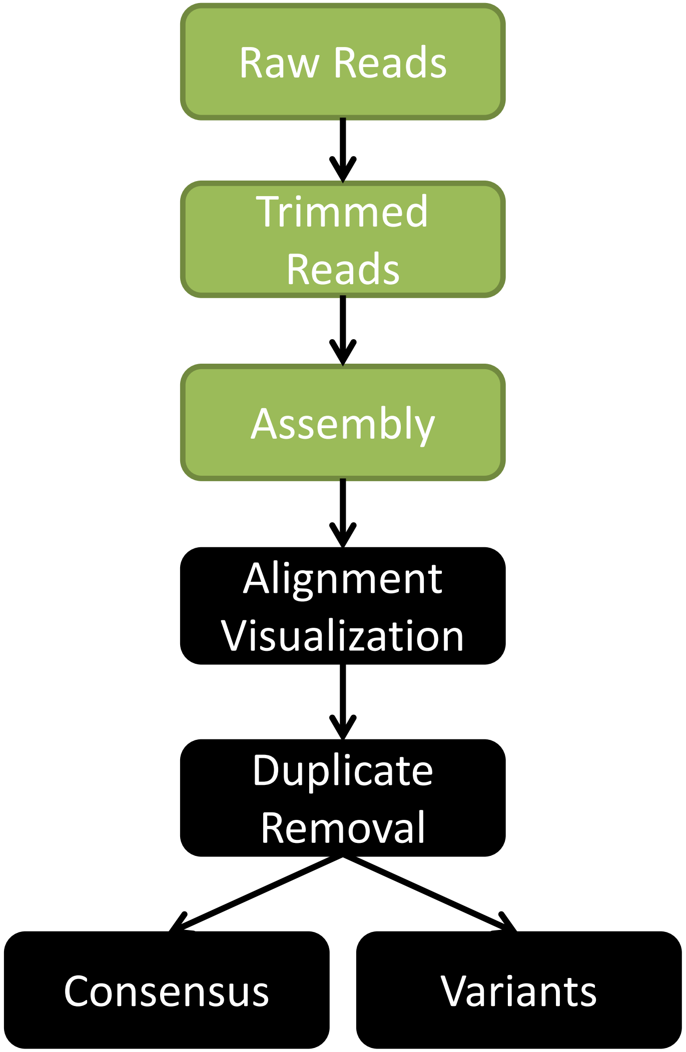


```
Pos:  1        10        20       30
Ref:  ACGTAGCATTGCTAGCGGTTAACCGTAGCT
       CGTAG   TGATA   GTTAA CGTAG
      ACGTA       TAGCG        T-GCA
             ATTGC    GCTTA
         TTGCA   CTAGC   TAACC
      ACGTA         GCGGT      TAGCT

        CIGAR: Pos4 1M1X3M      CIGAR: Pos 26 1M1D2M1X
```

```
Pos:  1        10        20       30
Ref:  ACGTAGCATTGCTAGCGGTTAACCGTAGCT
      ACGTA  ATTGC    GCTTA  CGTAG
      ACGTA    TGATA   GTTAA   T-GCA
       CGTAG     CTAGC   TAACC TAGCT
         TTGCA    TAGCG
                    GCGGT

        CIGAR: Pos4 1M1X3M      CIGAR: Pos 26 1M1D2M1X
```


Raw Reads
Trimmed Reads
Assembly
Alignment Visualization
Duplicate Removal
Consensus
Variants

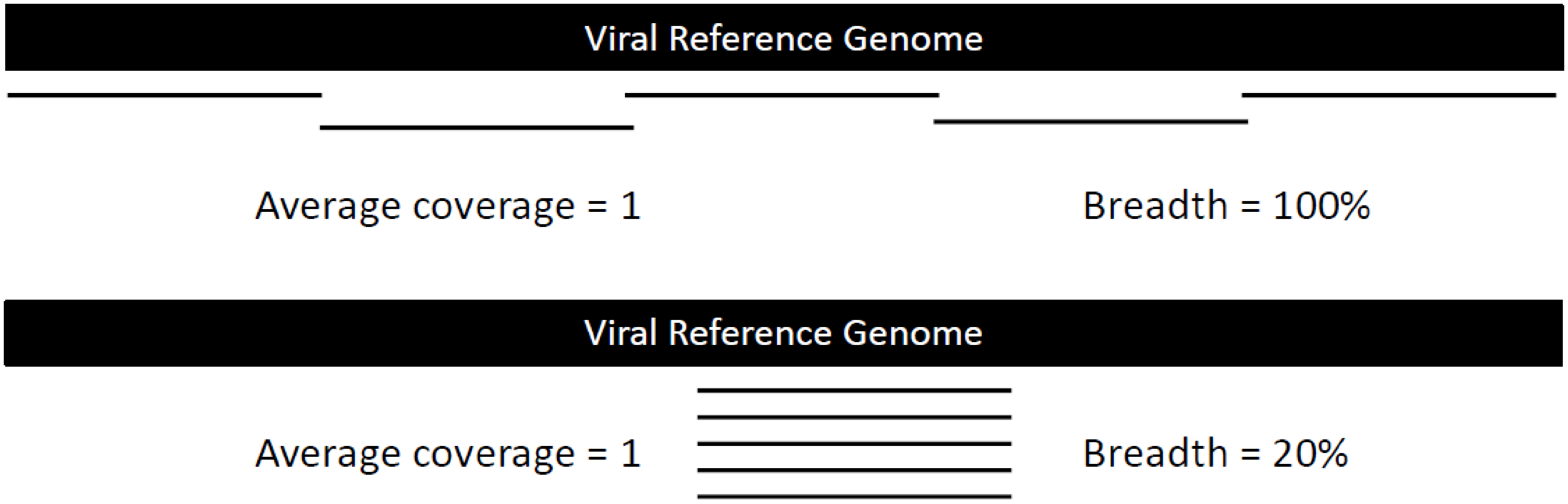
Alignment Stats/Visualization

- Summary stats give an idea of the mapping
  - Mapped vs Unmapped reads
  - Average depth of coverage
  - Breadth of Coverage
- Coverage Plots
- Visualization of complete alignment

Coverage

```
                  1         2         3
Pos:     1234567890123456789012345678901234
Ref:     ACGGTGACACGTAGCAGTACGCGGGTTACACAGA
         ACGGCGA       CAGTTCG      AC-CAGA
               AGACGTA       GCGGGTT
                   GTAGCAGT       TTACACAG
            GCGACAC        TCGCGGG
          CGGCGAC       AGTTCGC    TACACAT
                 ACG-AGC       GGGGTAC
Cov:     1223334333333233333343333344433331
```

Coverage Depth vs Breadth



Average coverage = 1

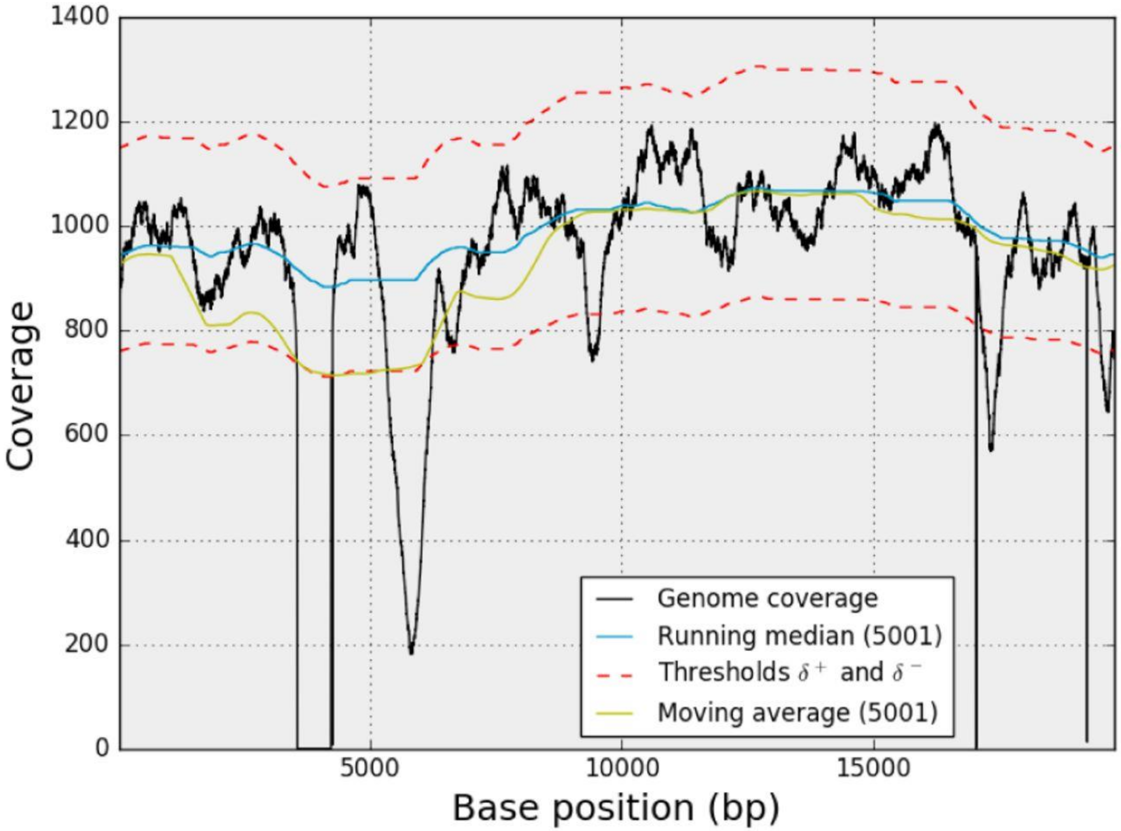
Breadth = 100%

Average coverage = 1

Breadth = 20%

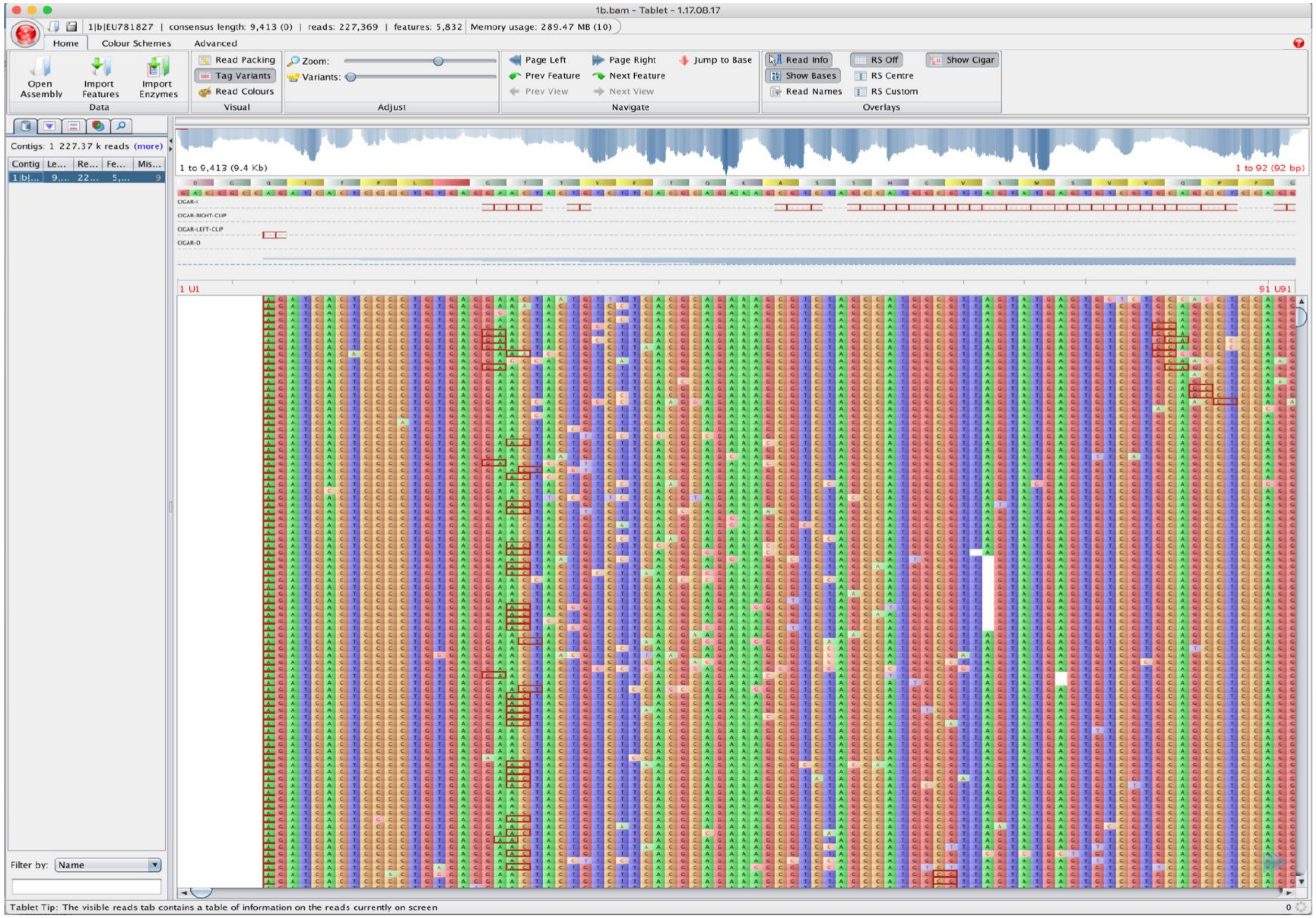
Mode, Median, Quartiles would be different

Coverage Plot



https://www.biorxiv.org/content/10.1101/092478v1.full

Visualization - Tablet



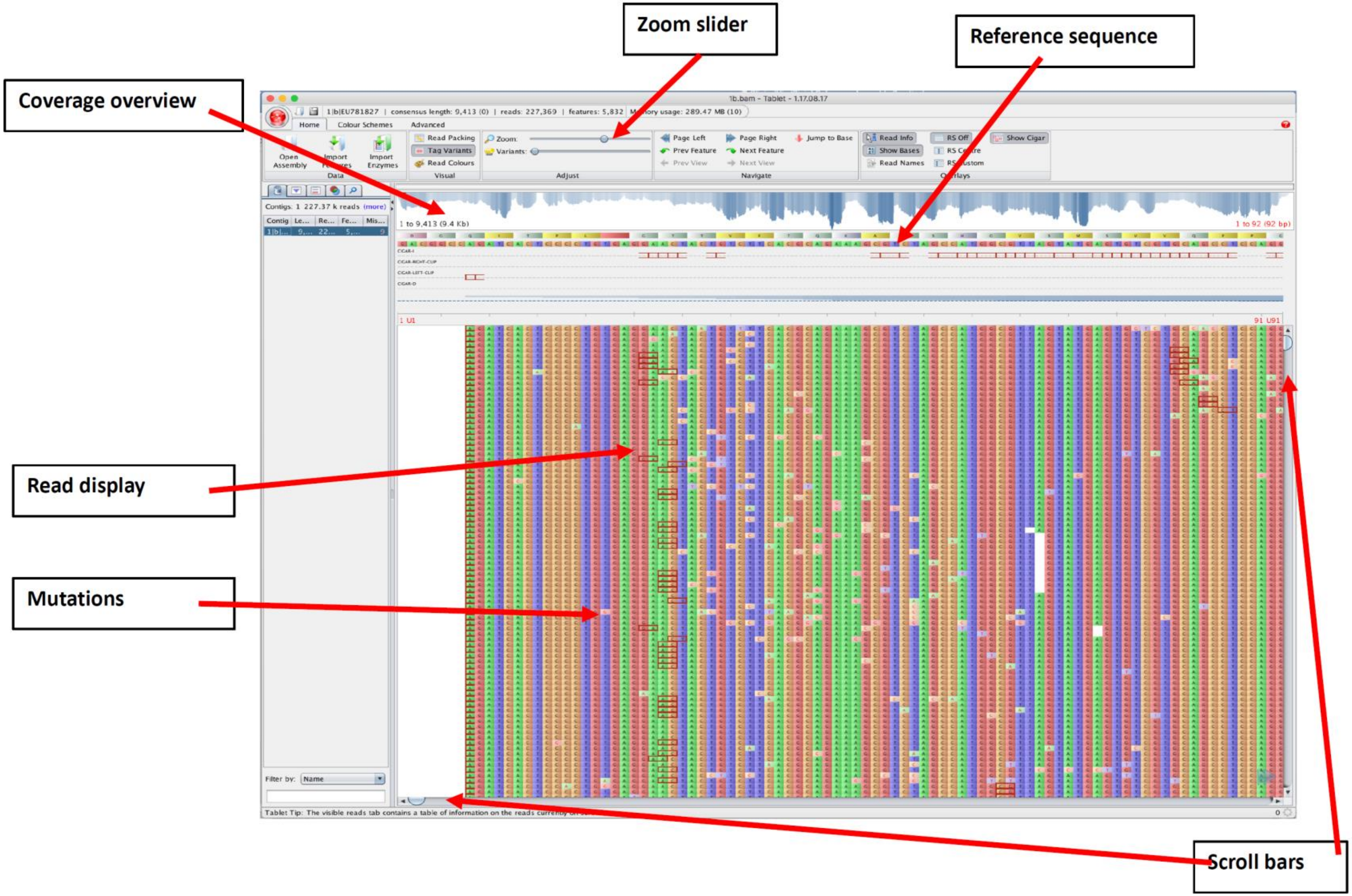
1|b|EU781827 | consensus length: 9,413 (0) | reads: 227,369 | features: 5,832 | Memory usage: 289.47 MB (10)

Tablet Tip: The visible reads tab contains a table of information on the reads currently on screen

BAM File + Reference File

https://ics.hutton.ac.uk/tablet/

Workshop On The Analysis Of Next-Generation HIV Sequence Data



1|b|EU781827 | consensus length: 9,413 (0) | reads: 227,369 | features: 5,832 | Memory usage: 289.47 MB (10)

Contigs: 1 227.37 k reads (more)

1 to 9,413 (9.4 Kb)

1 to 92 (92 bp)

Tablet Tip: The visible reads tab contains a table of information on the reads currently on screen

Zoom slider

Reference sequence

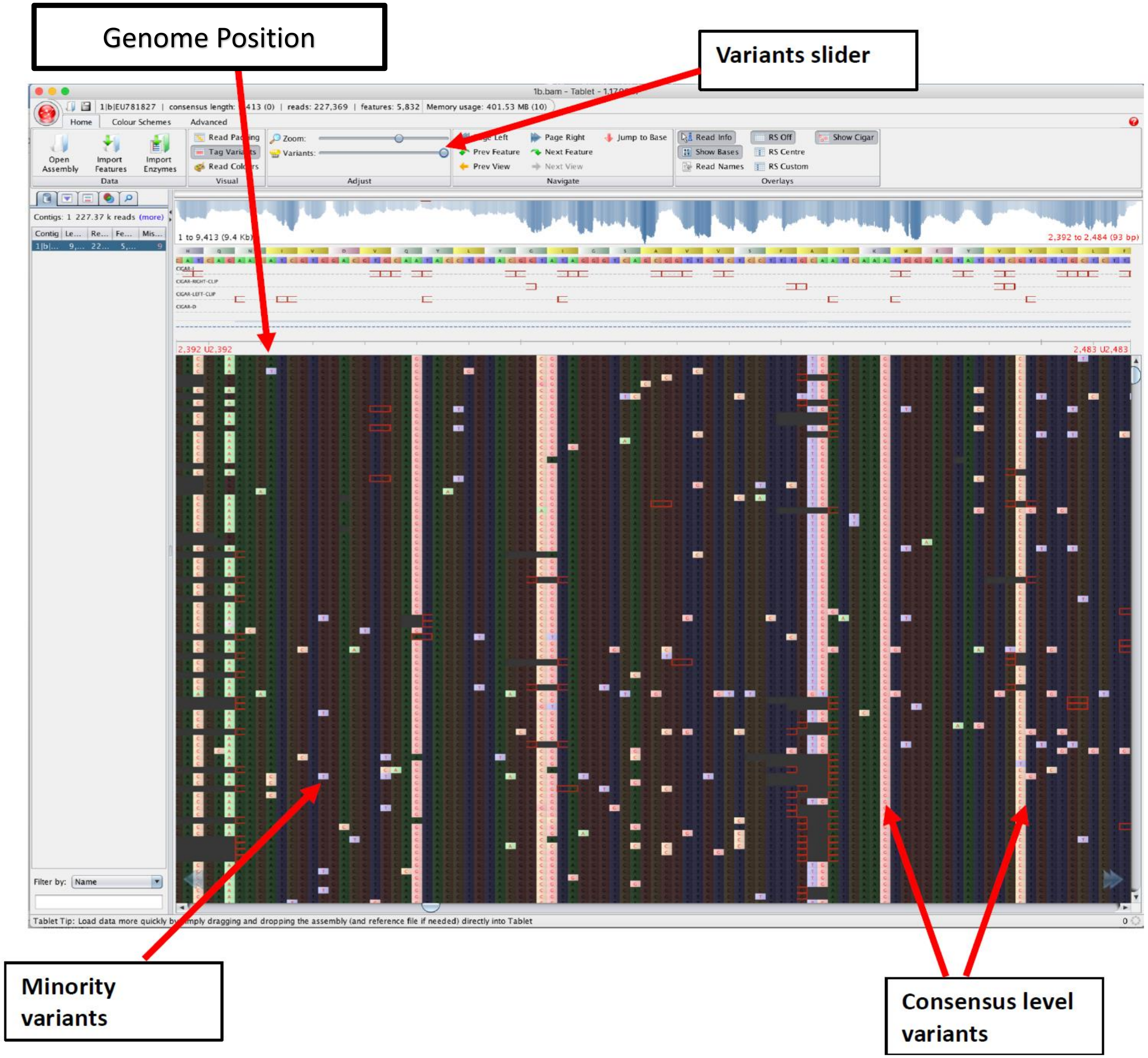
Coverage overview

Read display

Mutations

Scroll bars

WORKSHOP ON THE ANALYSIS OF NEXT-GENERATION HIV SEQUENCE DATA



Genome Position

Variants slider

Minority variants

Consensus level variants

1b.bam - Tablet - 1.17...

1|b|EU781827 | consensus length: ...413 (0) | reads: 227,369 | features: 5,832 | Memory usage: 401.53 MB (10)

Contigs: 1 227.37 k reads (more)

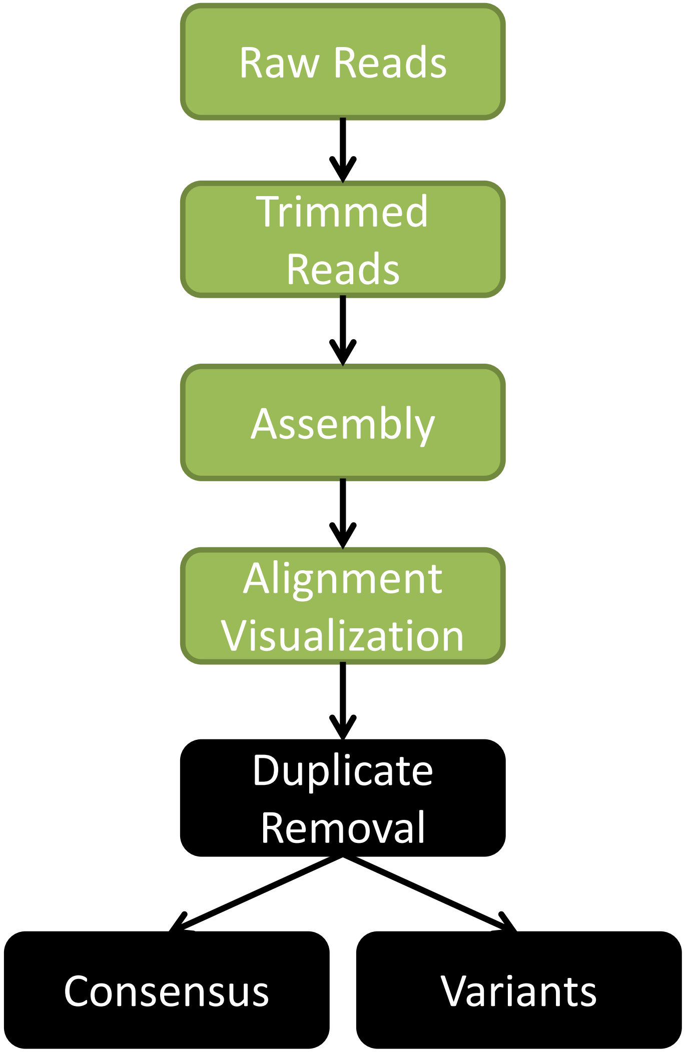
1 to 9,413 (9.4 Kb)

2,392 to 2,484 (93 bp)

2,392 U2,392

2,483 U2,483

Tablet Tip: Load data more quickly by simply dragging and dropping the assembly (and reference file if needed) directly into Tablet



Raw Reads

Trimmed Reads

Assembly

Alignment Visualization

Duplicate Removal

Consensus

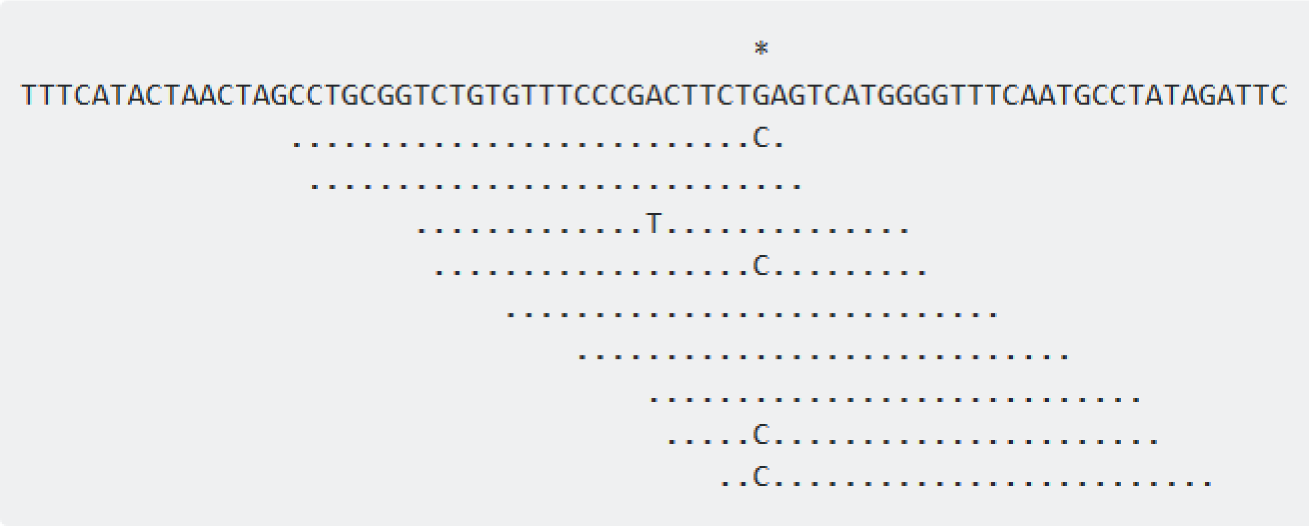
Variants

Duplicate Removal

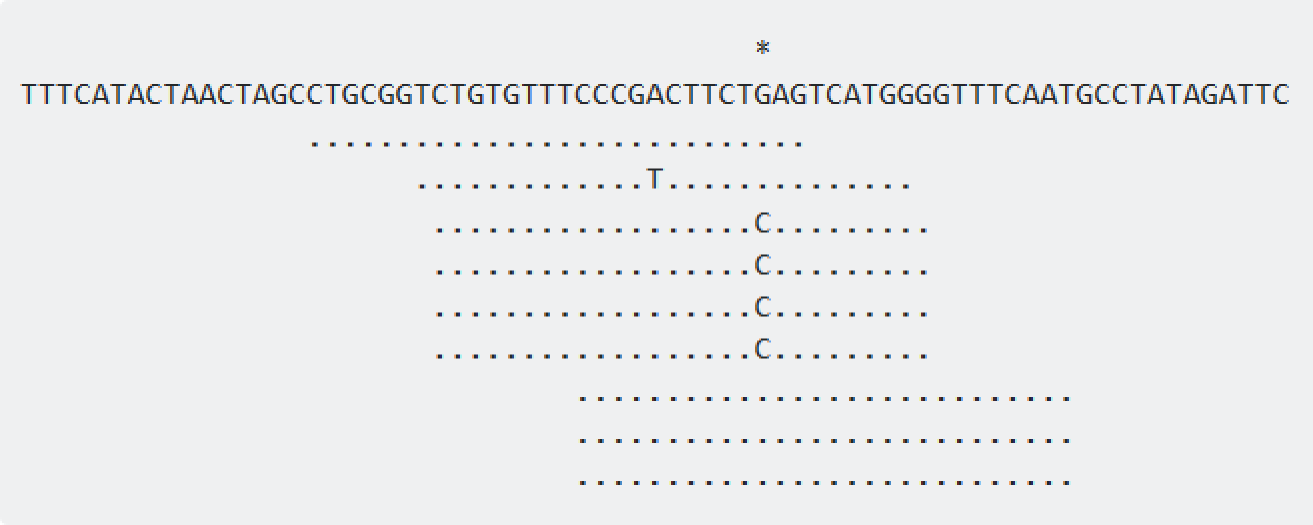
- Types of Duplicates
  - Optical Duplicates
  - PCR Duplicates
- Relationship of starting genomic material, sequencing depth and PCR Duplicates
- Why remove PCR Duplicates
  - Effect on Consensus Calling
  - Effect on Variant Calling

- **Note:** Only remove duplicates if library preparation has PCR steps

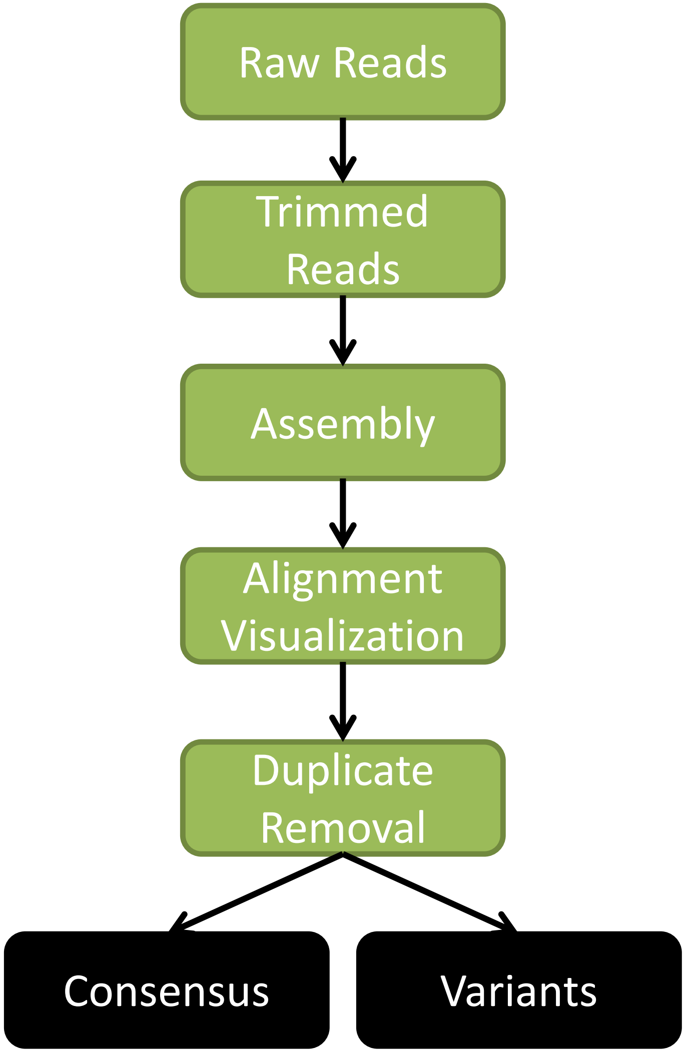
Duplicate Removal



```
                                         *
TTTCATACTAACTAGCCTGCGGTCTGTGTTTCCCGACTTCTGAGTCATGGGGTTTCAATGCCTATAGATTC
               ..........................C.
                ............................
                      .............T..............
                       ..................C.........
                           ............................
                               ............................
                                   ............................
                                    .....C......................
                                       ..C.........................
```



```
                                         *
TTTCATACTAACTAGCCTGCGGTCTGTGTTTCCCGACTTCTGAGTCATGGGGTTTCAATGCCTATAGATTC
                ............................
                      .............T..............
                       ..................C.........
                       ..................C.........
                       ..................C.........
                       ..................C.........
                               ............................
                               ............................
                               ............................
```


Raw Reads
Trimmed Reads
Assembly
Alignment Visualization
Duplicate Removal
Consensus
Variants