### Single Ended alignment of read length 50

Ref: TTGTGCTCTCGTCTCGTTATCAGCTAGACGATGCGAATATGGATAGCGGG Read:

CIGAR: 2S11M2D22M2I10M3S MD: 7A3^G^A14T12C4

NH tag: 3 SAM format Flag: 256

CIGAR:

### **Expand the CIGAR and the MD String**

 Both CIGAR and MD string is expanded to determine the nucleotides that are mismatches or are insertion/deletions

#### Overlay the reference and the read to determine indel and mismatched nucleotides

MD: Ref: CAGTGCTCTAGTCGATCGTTATCAGCTAGTCGATGCG TATGGCTAGCTAC TTGTGCTCTCGTC TCGTTATCAGCTAGACGATGCGAATATGGATAGCGGG Read: 

The reference sequence is adjusted to accommodate insertion nucleotides and the read sequence is adjusted to accommodate deletion nucleotides. Mismatches and insert nucleotides are represented by different characters.

Integrated CIGAR is constructed by including the soft-clips

The final Integrated CIGAR is constructed by replacing the

matching character with a letter to represent the SAM

Relevant information for each alignment is collected. NH

tag holds the number of time a read is aligned to the

#### **Construct the Integrated CIGAR**

Integrated CIGAR: TT7M)3M2D14M&7M!!5M&4MGGG

reference.

### **Construct the final Integrated CIGAR (No quality scores stored)**

SAM format Flag: 256

NH tag: 3

Integrated CIGAR: TT7M)3M2D14M&7M!!5M&4MGGG3 Integrated CIGAR: TT7B)3B2D14B&7B!!5B&4BGGG3

format flag. The value of the NH tag is appended.

# **Construct the final Integrated CIGAR (With quality scores)**

NH tag: 3 SAM format Flag: 256

Integrated CIGAR: TT7M)3M2D14M&7M!!5M&4MGGG3

Integrated CIGAR:  $TQ_1TQ_27B)Q_33B2D14B&Q_47B!Q_5!Q_65B&Q_74BGQ_8GQ_9GQ_{10}3$ 

If quality scores are requested to be stored, then the quality scores for only the nucleotides that are mismatched to the reference or inserts are stored.

# **Append MAPQ and AS tag Values**

AS tag: 50 MAP0: 255

Integrated CIGAR: TT7M)3M2D14M&7M!!5M&4MGGG3~50~255 Integrated CIGAR:  $TQ_1TQ_27B$ ) $Q_33B2D14B&Q_47B$ ! $Q_5$ ! $Q_65B&Q_74BGQ_8GQ_9GQ_{10}3\sim50\sim255$ 

value of the AS tag along with the MAPQ values are retained within the Integrated CIGAR.

# Add number of repetitions of the same read

Integrated CIGAR: TT7M)3M2D14M&7M!!5M&4MGGG3~50~255-4

Integrated CIGAR:  $TQ_1TQ_27B$ ) $Q_33B2D14B&Q_47B$ ! $Q_5$ ! $Q_65B&Q_74BGQ_8GQ_9GQ_{10}3~50~255-4$ 

Due to several rounds of PCR, multiple reads could be sequenced from the same nucleotide position of the reference. ABRIDGE will store a single representation and record the number of repetitions of the read.

If alignment scores are requested to be stored, then the

### Exact same mapping of adjoining sequence with different SAM format Flag

Integrated CIGAR1: TT7B)3B2D14B&7B!!5B&4BGGG3~50~255-4 Integrated CIGAR2: TT7C)3C2D14C&7C!!5C&4CGGG3~43~250-2 Final CIGAR2: C~43~250-2

If adjoining alignments have the same integrated CIGARR with only a different SAM Format Flag, then the representation is reduced to only the SAM Format Flag and the alignment scores with the repetition of reads