Motivating exact matching via read mapping

CMSC701

UNIVERSITY OF MARYLAND

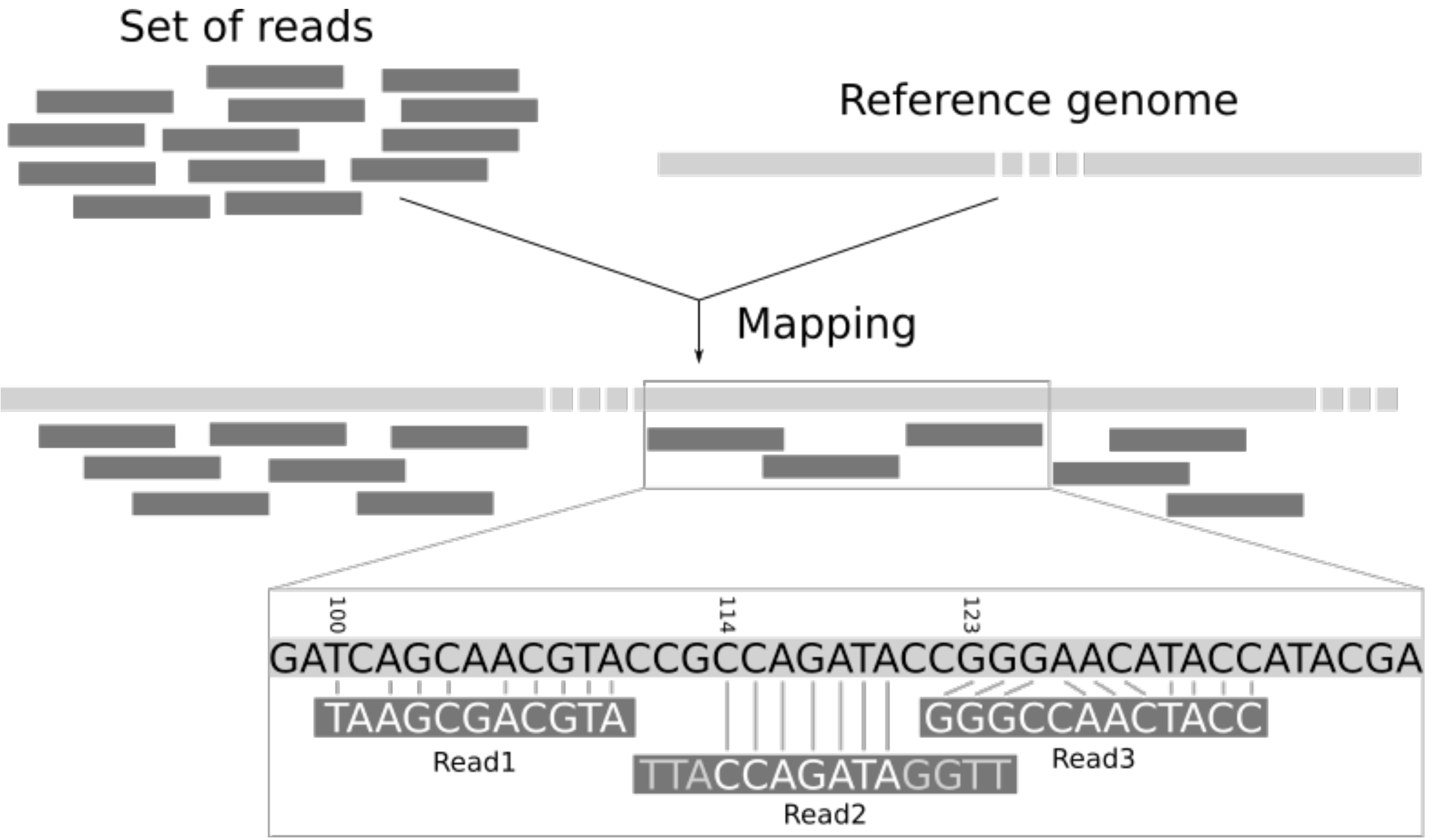
(Short) Read mapping/alignment

Read mapping / alignment is one of the most fundamental computational tasks in genomics.

Performing read mapping is often the first step in *many* different analyses.

Q: For each read, where might I have sequenced it from on the genome, and how does the read differ from the reference at the mapped position?

(Short) Read mapping/alignment



Short reads:

- 10^6-10^9 reads
- 100-300 nt per end
- Often “paired-end”

Genome:

- 10^6-10^{10} nt long
- May contain gaps / Ns

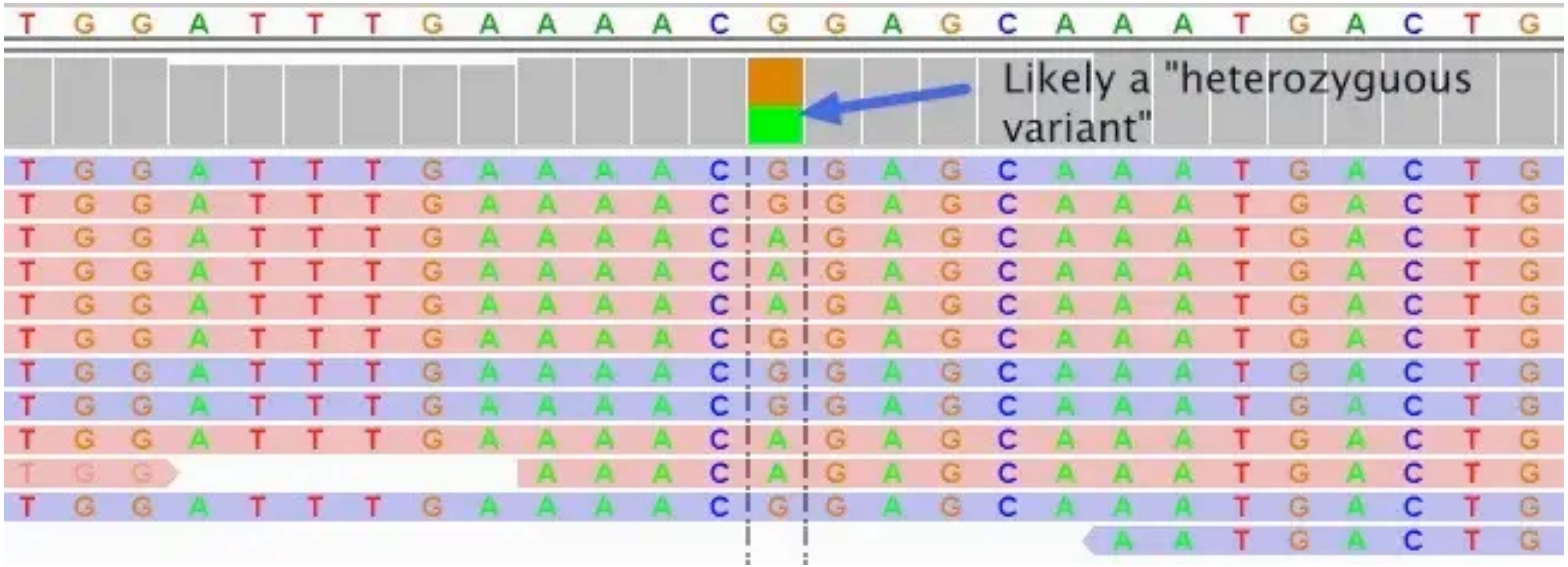
Q: For each read, where might I have sequenced it from on the genome?

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

Read mapping → variant calling

Reference genome

Mapped reads from sample



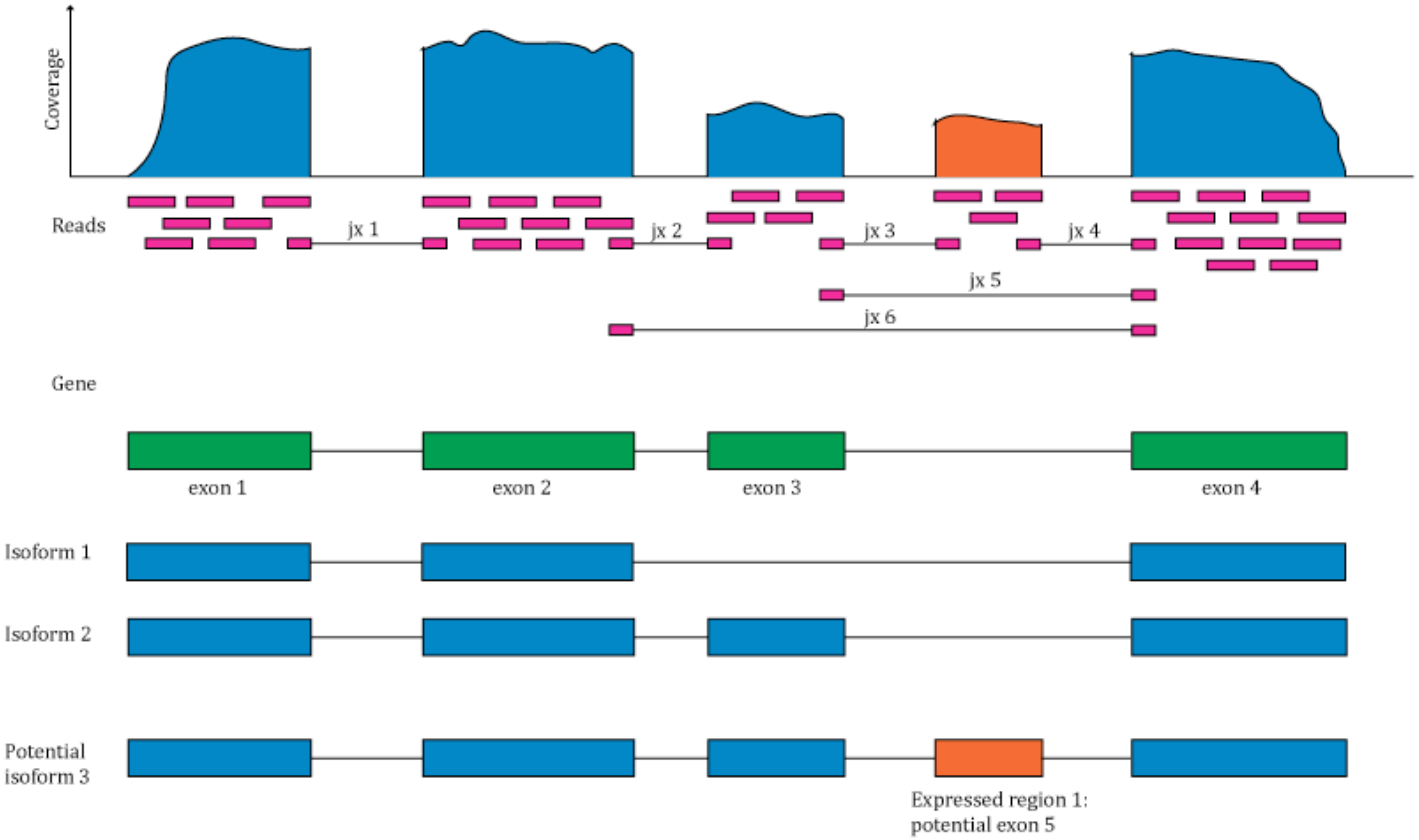
```
T G G A T T T G A A A A C G G A G C A A A T G A C T G

T G G A T T T G A A A A C G G A G C A A A T G A C T G
T G G A T T T G A A A A C G G A G C A A A T G A C T G
T G G A T T T G A A A A C A G A G C A A A T G A C T G
T G G A T T T G A A A A C A G A G C A A A T G A C T G
T G G A T T T G A A A A C A G A G C A A A T G A C T G
T G G A T T T G A A A A C G G A G C A A A T G A C T G
T G G A T T T G A A A A C G G A G C A A A T G A C T G
T G G A T T T G A A A A C G G A G C A A A T G A C T G
T G G A T T T G A A A A C A G A G C A A A T G A C T G
T G G             A A A C A G A G C A A A T G A C T G
T G G A T T T G A A A A C G G A G C A A A T G A C T G
                                        A A T G A C T G
```

Given the alignment of many reads to the “reference”, how and where does my sample differ from the reference?

https://towardsdatascience.com/simple-convolution-neural-network-for-genomic-variant-calling-with-tensorflow-c085dbc2026f

Read mapping → count/census



Given a sequencing sample, which genes (isoforms) are generating reads, and how many reads are coming from each (quantitative measure of expression level).

https://rna.recount.bio/docs/

The utility of *exact* matching here

As *loose* motivation, consider the problem of mapping a read r to the genome G.

In reality, we would not use exact matching for this; why?

However, exact matching is useful here:

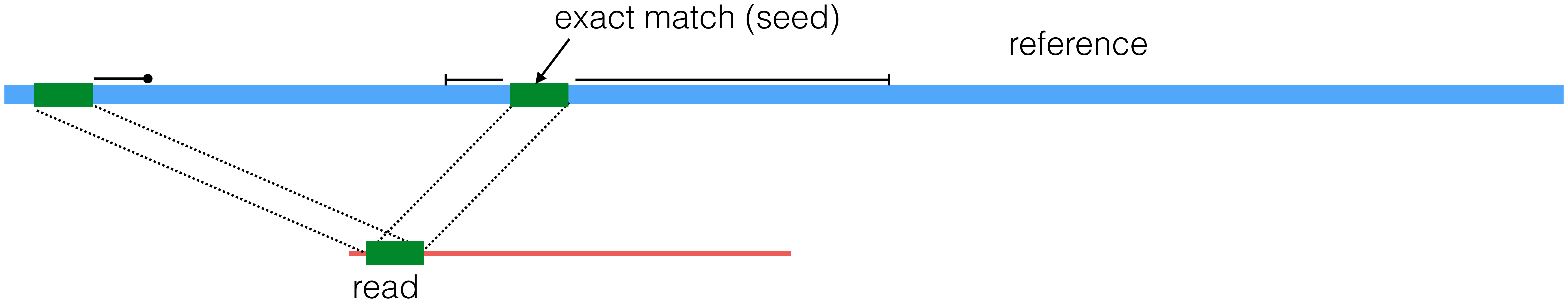
- Find all places where a substring of the query matches the reference exactly (seeds)
- Filter out regions with insufficient exact matches to warrant further investigation
- Perform a “constrained” alignment that includes these exact matching “seeds”

Requires efficient exact search

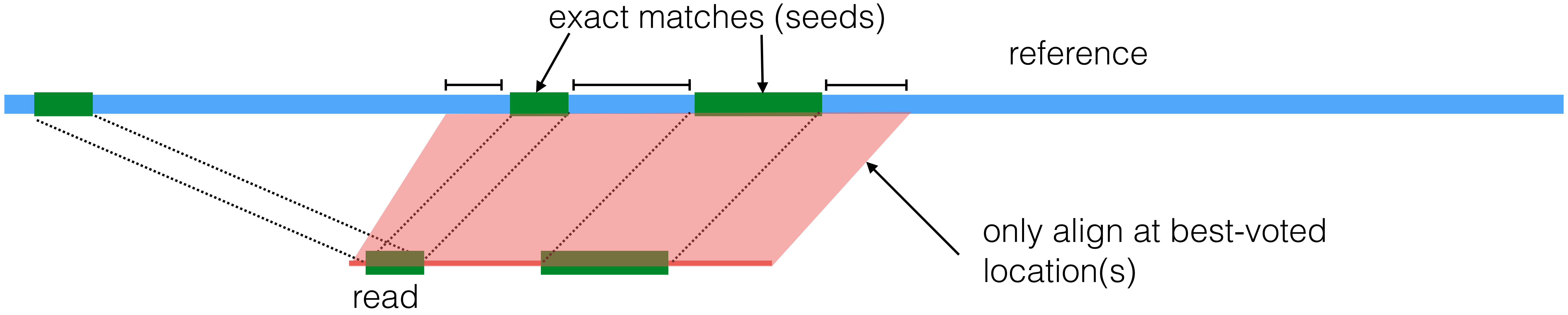
Here is where we use efficient algorithms for inexact matching (alignment)

Typical Strategies

Seed & Extend:

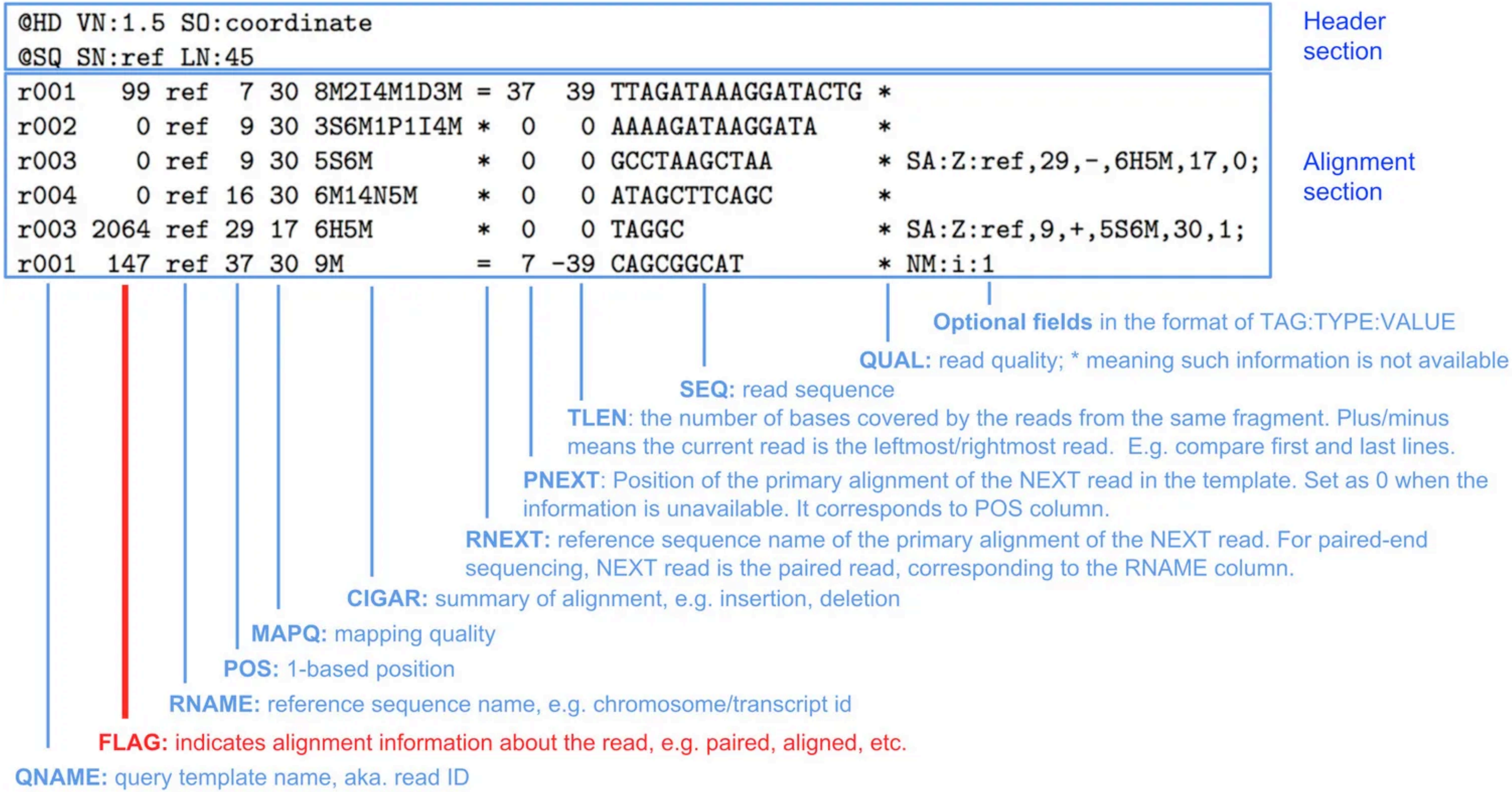


Seed & Vote:



only align at best-voted location(s)

Representing alignments



```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M * 0    0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       * 0    0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    * 0    0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       * 0    0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         = 7  -39 CAGCGGCAT         * NM:i:1
```

Header section

Alignment section

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

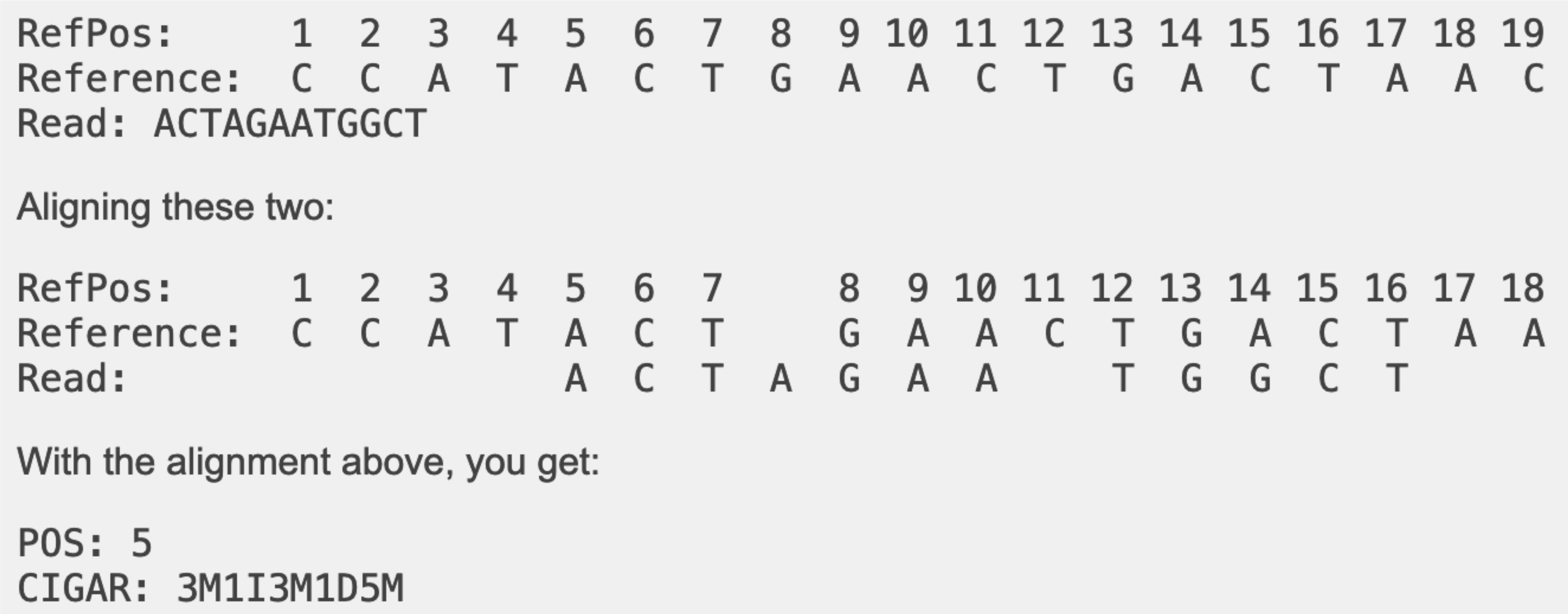
RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

Source: https://bioinformaticamente.com/2021/03/03/sam-bam/

What is a CIGAR string?



```
RefPos:     1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T  G  A  A  C  T  G  A  C  T  A  A  C
Read: ACTAGAATGGCT
```

Aligning these two:

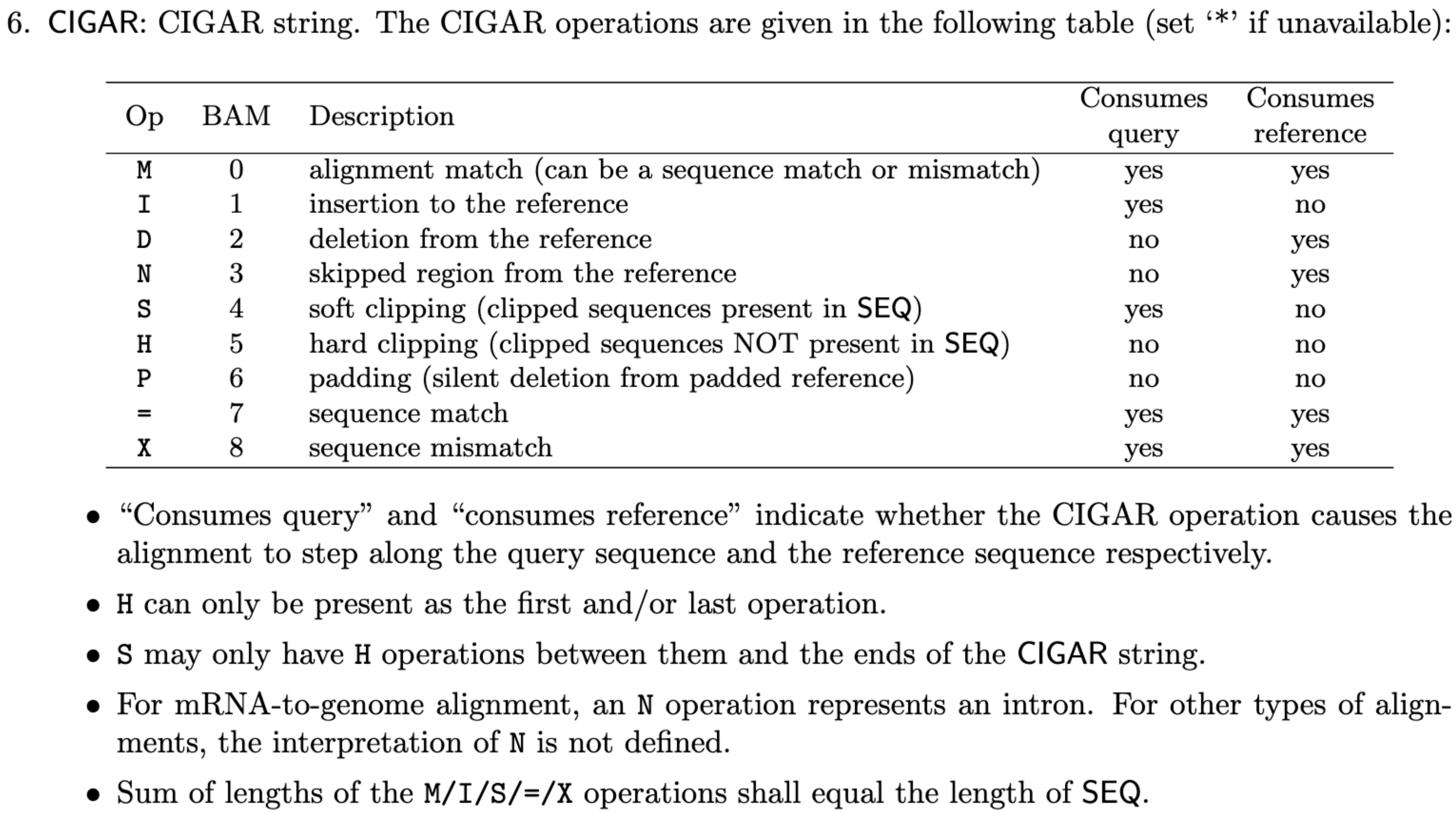
```
RefPos:     1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16 17 18
Reference:  C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T  A  A
Read:                   A  C  T  A  G  A  A     T  G  G  C  T
```

With the alignment above, you get:

```
POS: 5
CIGAR: 3M1I3M1D5M
```

https://sites.google.com/site/bioinformaticsremarks/bioinfo/sam-bam-format/what-is-a-cigar

What is a CIGAR string?



6. CIGAR: CIGAR string. The CIGAR operations are given in the following table (set '*' if unavailable):

Op	BAM	Description	Consumes query	Consumes reference
M	0	alignment match (can be a sequence match or mismatch)	yes	yes
I	1	insertion to the reference	yes	no
D	2	deletion from the reference	no	yes
N	3	skipped region from the reference	no	yes
S	4	soft clipping (clipped sequences present in SEQ)	yes	no
H	5	hard clipping (clipped sequences NOT present in SEQ)	no	no
P	6	padding (silent deletion from padded reference)	no	no
=	7	sequence match	yes	yes
X	8	sequence mismatch	yes	yes

- "Consumes query" and "consumes reference" indicate whether the CIGAR operation causes the alignment to step along the query sequence and the reference sequence respectively.
- H can only be present as the first and/or last operation.
- S may only have H operations between them and the ends of the CIGAR string.
- For mRNA-to-genome alignment, an N operation represents an intron. For other types of alignments, the interpretation of N is not defined.
- Sum of lengths of the M/I/S/=/X operations shall equal the length of SEQ.

https://samtools.github.io/hts-specs/SAMv1.pdf