Genômica Computacional

Analise de variantes

Professor: Ricardo A. Vialle

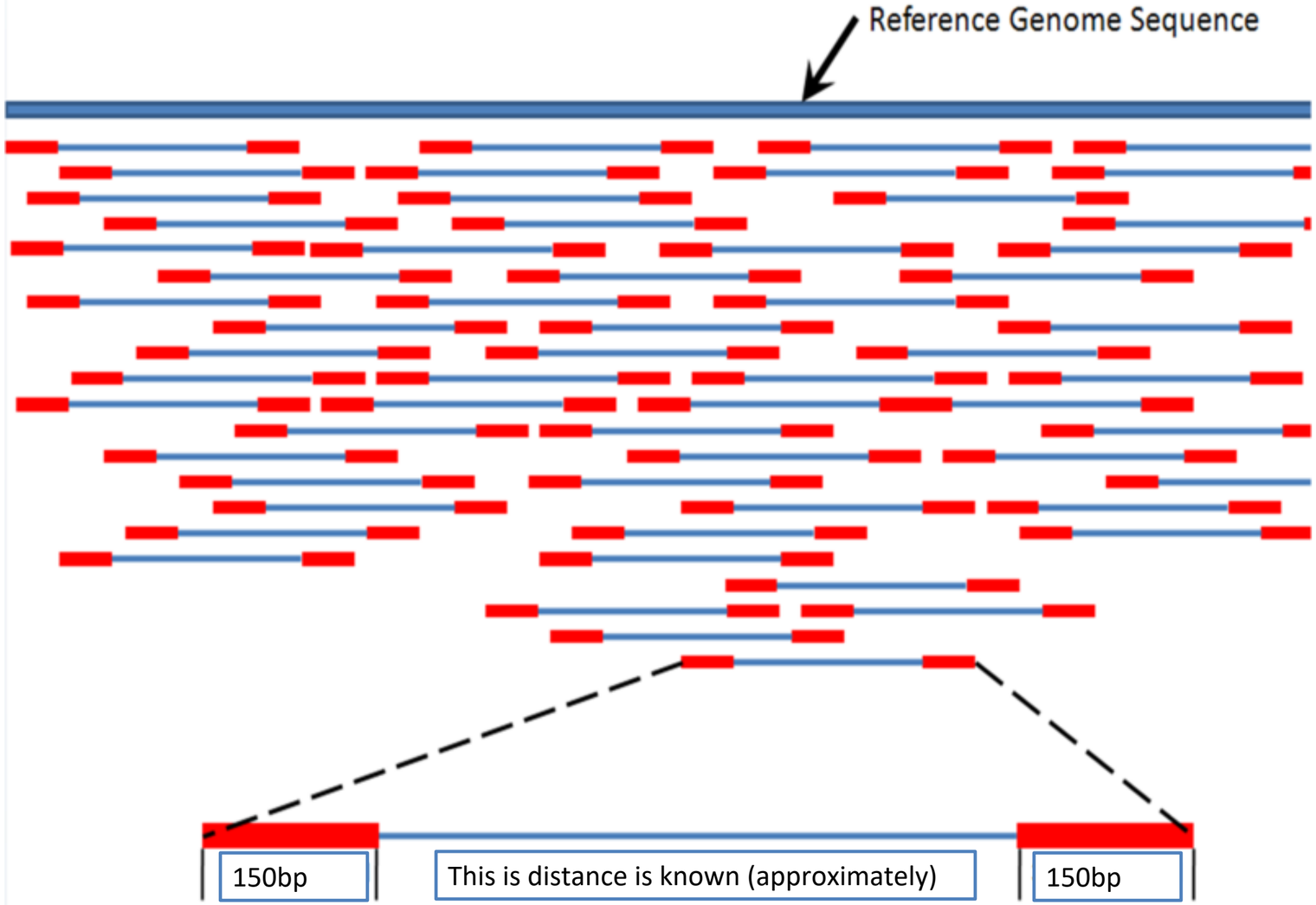
CS31 - Genômica Computacional

22 de Novembro de 2023

Cronograma

Data	Tema
~~11-Oct~~	~~Introdução a Genômica, Sequenciamento (teórica)~~
~~18-Oct~~	~~Bioinformática - Linux - Processamento de dados de sequenciamento (teórico-prática)~~
~~1-Nov~~	~~Montagem de genomas (teórico-prática)~~
~~8-Nov~~	~~Anotação de genomas (teórico-prática)~~
22-Nov	Analise de variabilidade genética (teórico-prática)

Read mapping

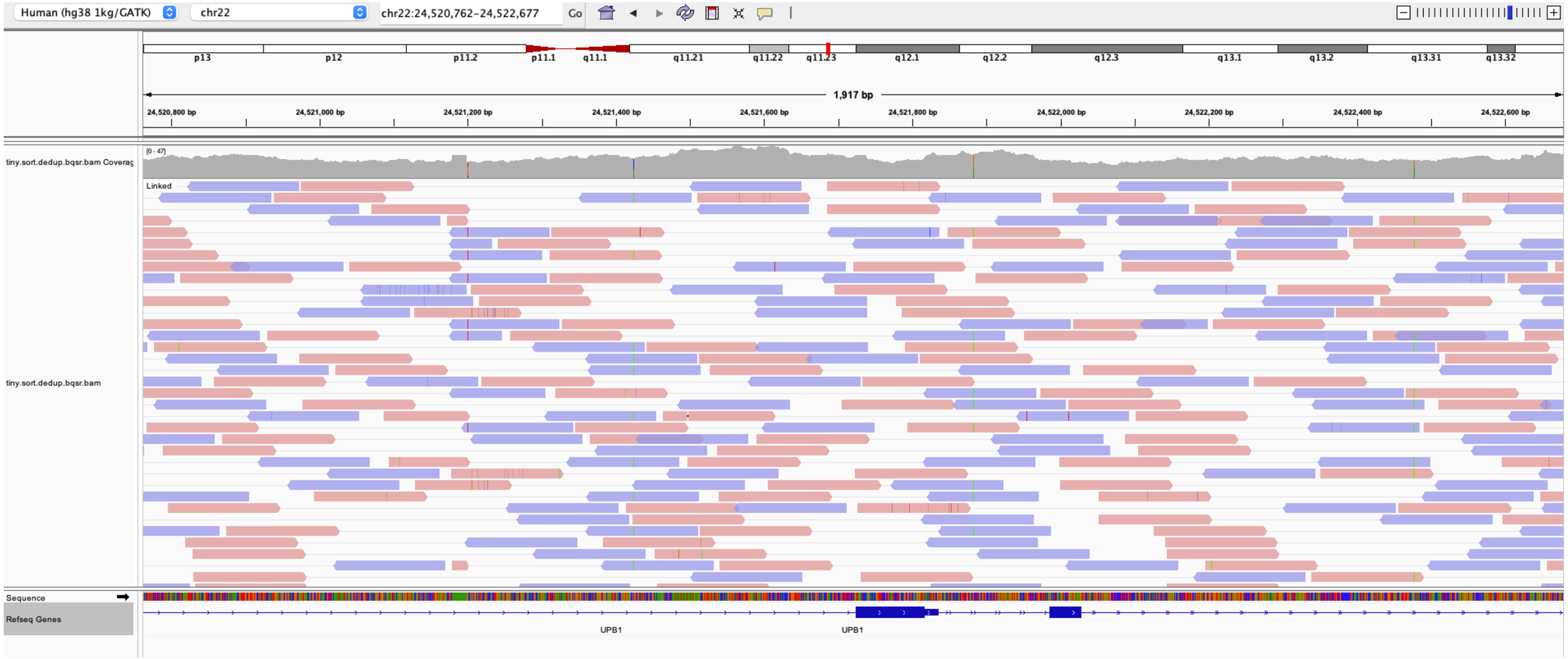


https://biocorecrg.github.io/RNAseq_course_2019/alignment.html

Read mapping

https://dmnfarrell.github.io/img/bam-pileup.png

IGV browser


Human (hg38 1kg/GATK)
chr22
chr22:24,520,762-24,522,677
Go
p13
p12
p11.2
p11.1
q11.1
q11.21
q11.22
q11.23
q12.1
q12.2
q12.3
q13.1
q13.2
q13.31
q13.32
1,917 bp
24,520,800 bp
24,521,000 bp
24,521,200 bp
24,521,400 bp
24,521,600 bp
24,521,800 bp
24,522,000 bp
24,522,200 bp
24,522,400 bp
24,522,600 bp
tiny.sort.dedup.bqsr.bam Coverag
[0 - 47]
Linked
tiny.sort.dedup.bqsr.bam
Sequence
Refseq Genes
UPB1
UPB1

Types of genetic variation

- **99% of DNA is shared** between two individuals
- Variation in the remainder explains all our **predisposition** differences
- **Remaining** phenotypic variation: environmental/stochastic differences

Name	**Example**	**Frequency in one genome**
Single nucleotide polymorphisms (**SNPs**)	GAGGAGAACG[C/G]AACTCCGCCG	1 per 1,000 bp
Insertions/deletions (**indels**)	CACTATTC[C/CTATGG]TGTCTAA	1 per 10,000 bp
Short tandem repeats (**STRs**)	ACGGCAGTCGTCGTCGTCACCGTAT	1 per 10,000 bp
Structural variants (**SVs**) / Copy Number Variants (**CNVs**)	Large (median 5,000 bp) deletions, duplications, inversions	1 per 1,000,000 bp

VCF and BCF file format

VCF

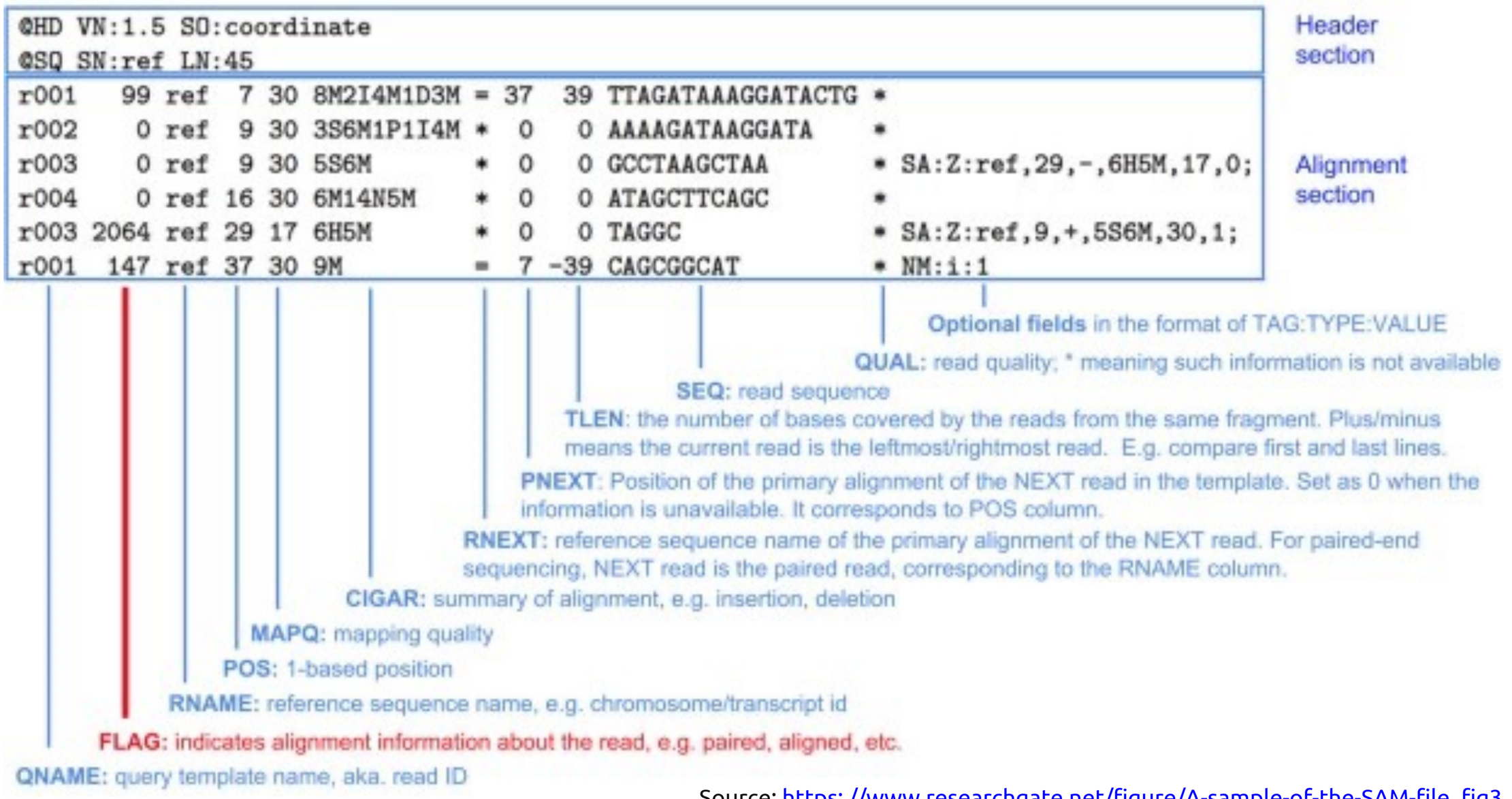
```
##fileformat=VCFv4.2
##contig=<ID=2,length=51304566>
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
#CHROM  POS ID  REF ALT QUAL FILTER  INFO FORMAT     SAMPLE1     SAMPLE2     SAMPLE3     SAMPLE4     SAMPLE5     SAMPLE6     SAMPLE7
2  81170  .  C  T   .  .   AC=9;AN=7424   GT:DP:GQ   0/0:4:12    0/0:3:9     0/1:1:3     0/1:9:24    1/0:4:12    0/0:5:15    0/0:4:12
2  81171  .  G  A   .  .   AC=6;AN=7446   GT:DP:GQ   0/1:4:12    0/0:3:9     0/0:1:3     0/0:9:24    0/1:4:12    0/1:5:15    0/0:4:12
2  81182  .  A  G   .  .   AC=5;AN=7506   GT:DP:GQ   0/0:5:15    0/0:4:12    0/0:5:15    0/0:9:24    0/0:4:12    0/0:4:12    0/0:4:12
2  81204  .  T  G   .  .   AC=2;AN=7542   GT:DP:GQ   1/0:5:15    0/0:9:27    0/0:10:30   0/0:15:39   0/0:9:27    1/0:13:39   0/1:14:42
```

BCF

```
2  81170  .  C  T   .  .   AC=9;AN=7424   GT:0/0:0/0:0/1:0/1:1/0:0/0:0/0   DP:4:3:1:9:4:5:4   GQ:12: 9: 3:24:12:15:12
2  81171  .  G  A   .  .   AC=6;AN=7446   GT:0/1:0/0:0/0:0/0:0/1:0/1:0/0   DP:4:3:1:9:4:5:4   GQ:12: 9: 3:24:12:15:12
2  81182  .  A  G   .  .   AC=5;AN=7506   GT:0/0:0/0:0/0:0/0:0/0:0/0:0/0   DP:5:4:5:9:4:4:4   GQ:15:12:15:24:12:12:12
2  81204  .  T  G   .  .   AC=2;AN=7542   GT:1/0:0/0:0/0:0/0:0/0:1/0:0/1   DP:5:9:10:15:9:13:14   GQ:15:27:30:39:27:39:42
```

https://en.wikipedia.org/wiki/Variant_Call_Format#/media/File:Binary_BCF_versus_VCF_format.png

SAM and BAM file format



Source: https: //www.researchgate.net/figure/A-sample-of-the-SAM-file_fig3_309134977

SAM and BAM file format (FLAG)

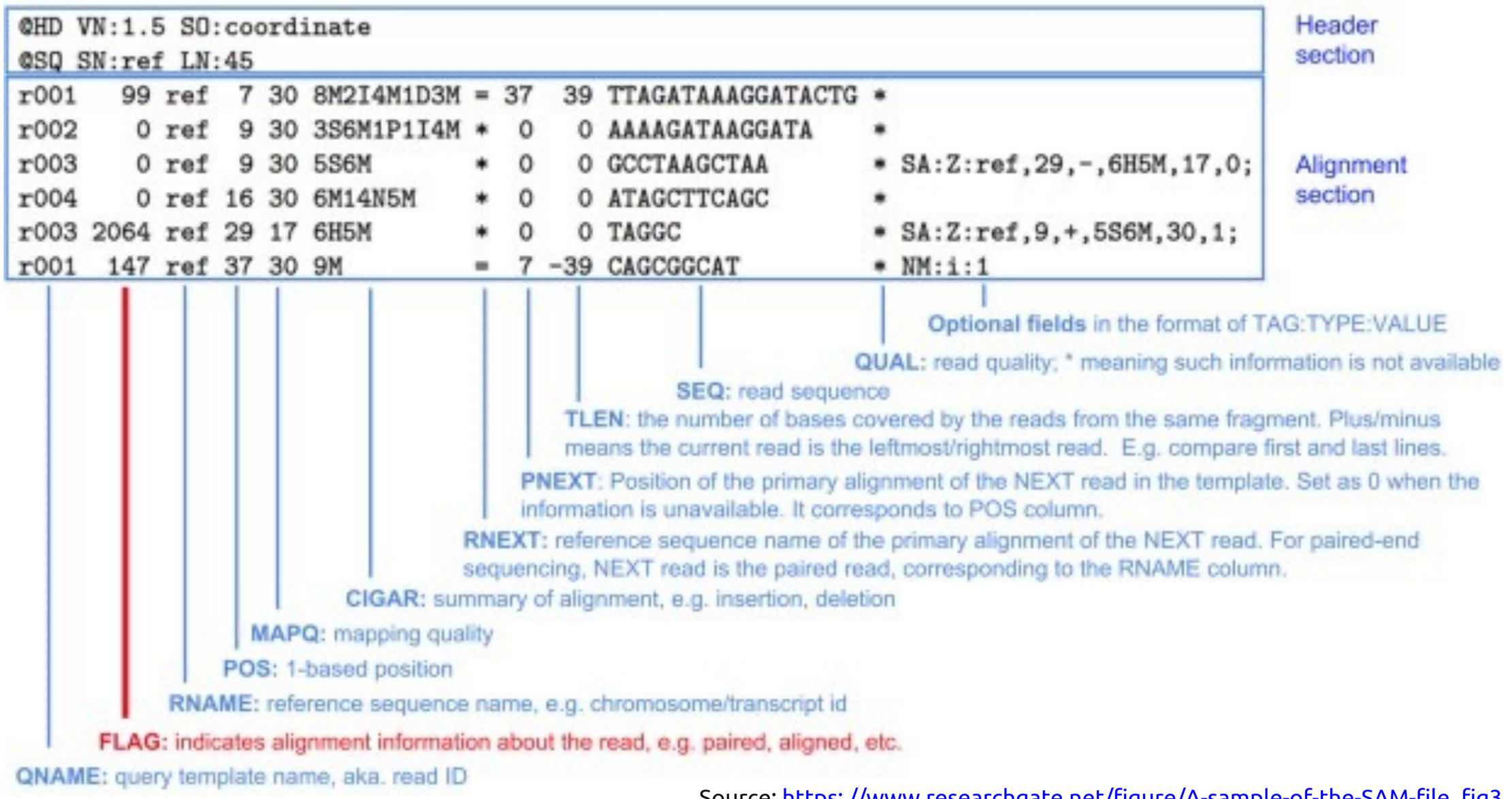
Bitwise Flags

Integer	Binary	Description (Paired Read Interpretation)
1	000000000001	template having multiple templates in sequencing (read is paired)
2	000000000010	each segment properly aligned according to the aligner (read mapped in proper pair)
4	000000000100	segment unmapped (read1 unmapped)
8	000000001000	next segment in the template unmapped (read2 unmapped)
16	000000010000	SEQ being reverse complemented (read1 reverse complemented)
32	000000100000	SEQ of the next segment in the template being reverse complemented (read2 reverse complemented)
64	000001000000	the first segment in the template (is read1)
128	000010000000	the last segment in the template (is read2)
256	000100000000	not primary alignment
512	001000000000	alignment fails quality checks
1024	010000000000	PCR or optical duplicate
2048	100000000000	supplementary alignment (e.g. aligner specific, could be a portion of a split read or a tied region)

https://en.wikipedia.org/wiki/SAM_(file_format)

Use ***samtools flagstat*** to summarize these metrics

SAM and BAM file format



Source: https: //www.researchgate.net/figure/A-sample-of-the-SAM-file_fig3_309134977

CIGAR

For example:

```
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:

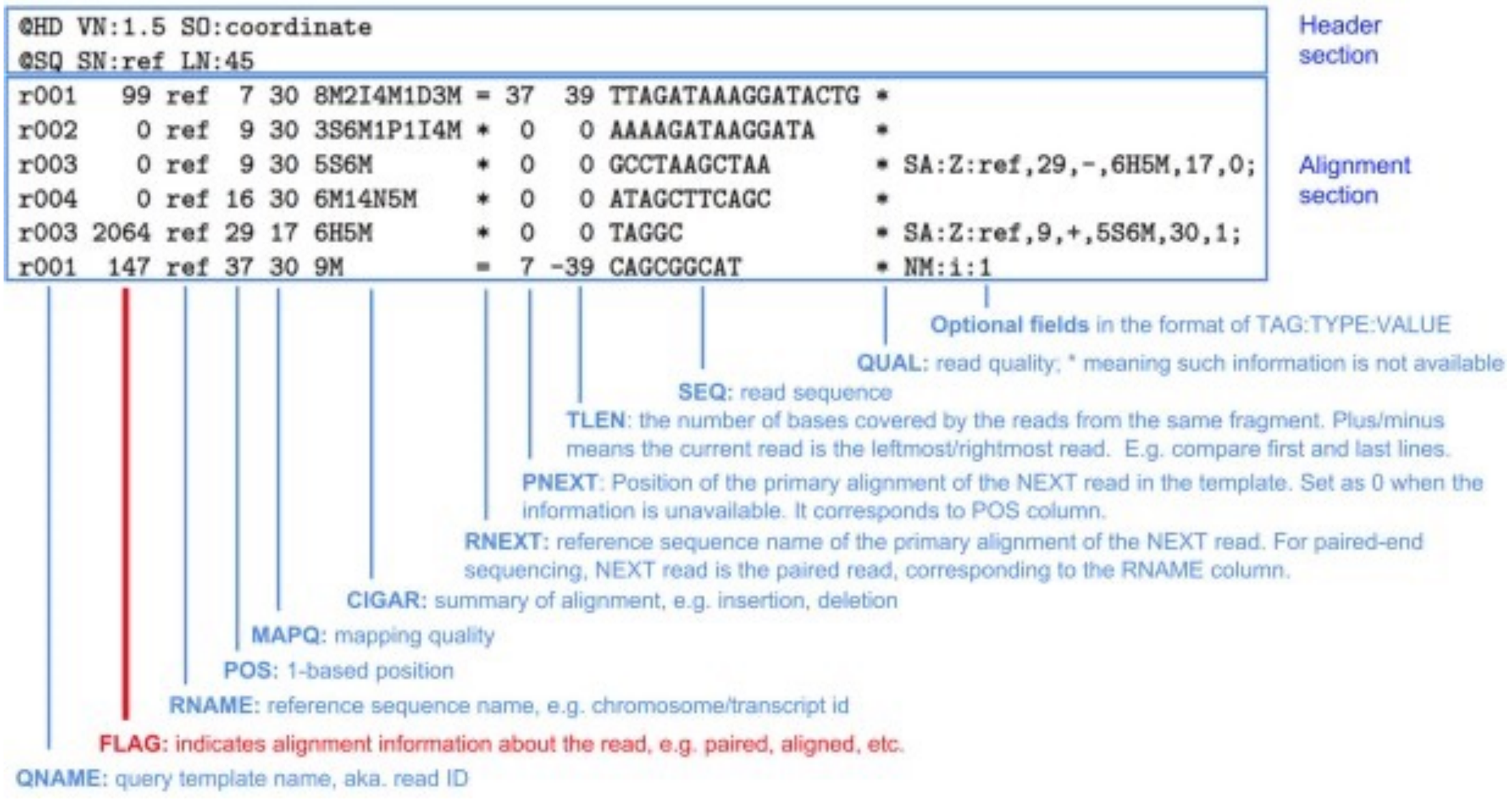
```
RefPos:     1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T  A  A  C
Read:                  A  C  T  A  G  A  A     T  G  G  C  T
```

With the alignment above, you get:

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

https://genome.sph.umich.edu/wiki/SAM

SAM and BAM file format



SAM and BAM file format

```
@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

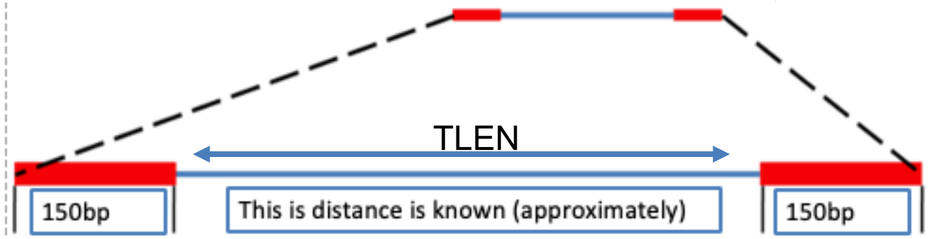
MAPQ: mapping quality

POS: 1-based position

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

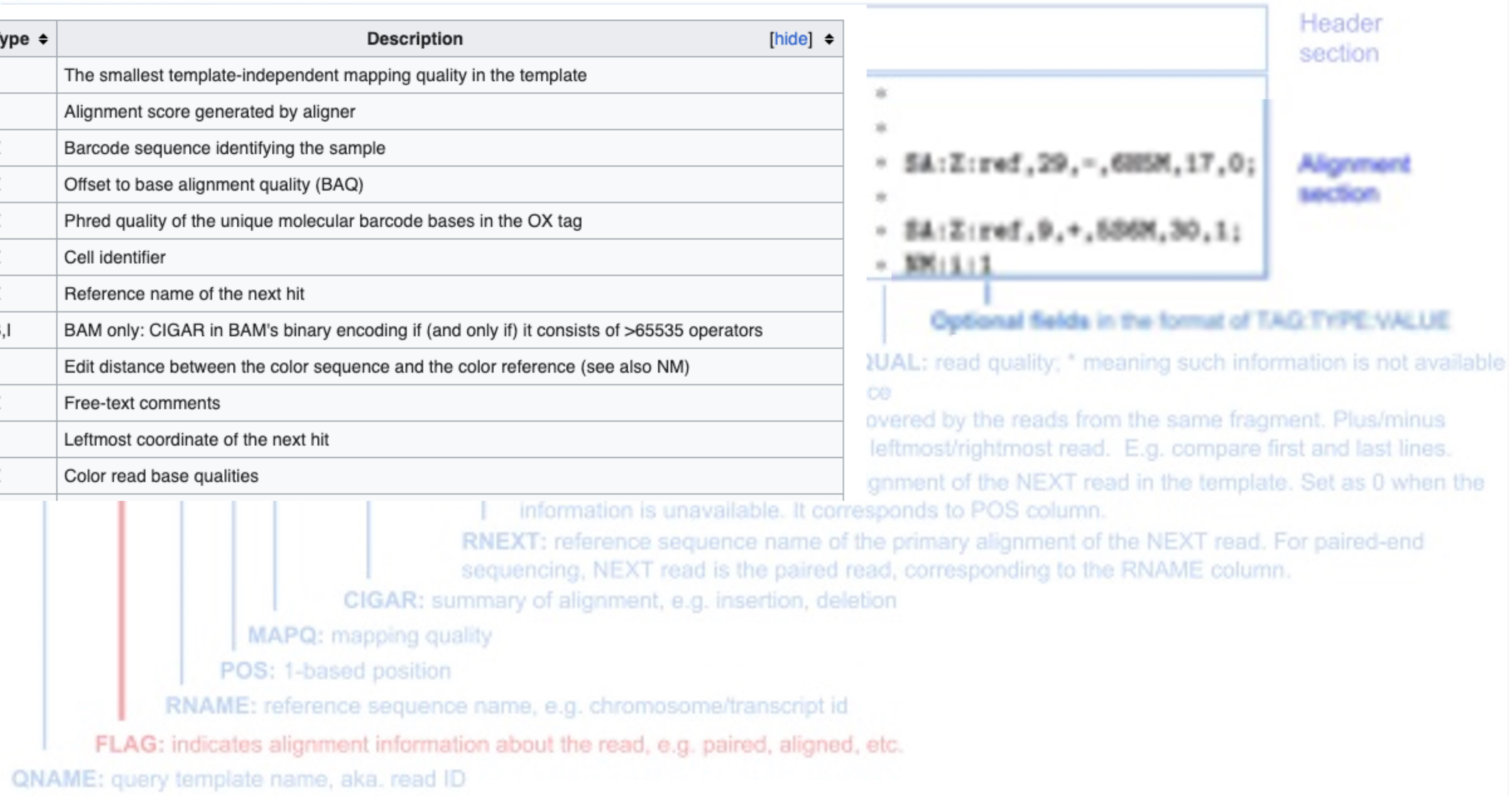
FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

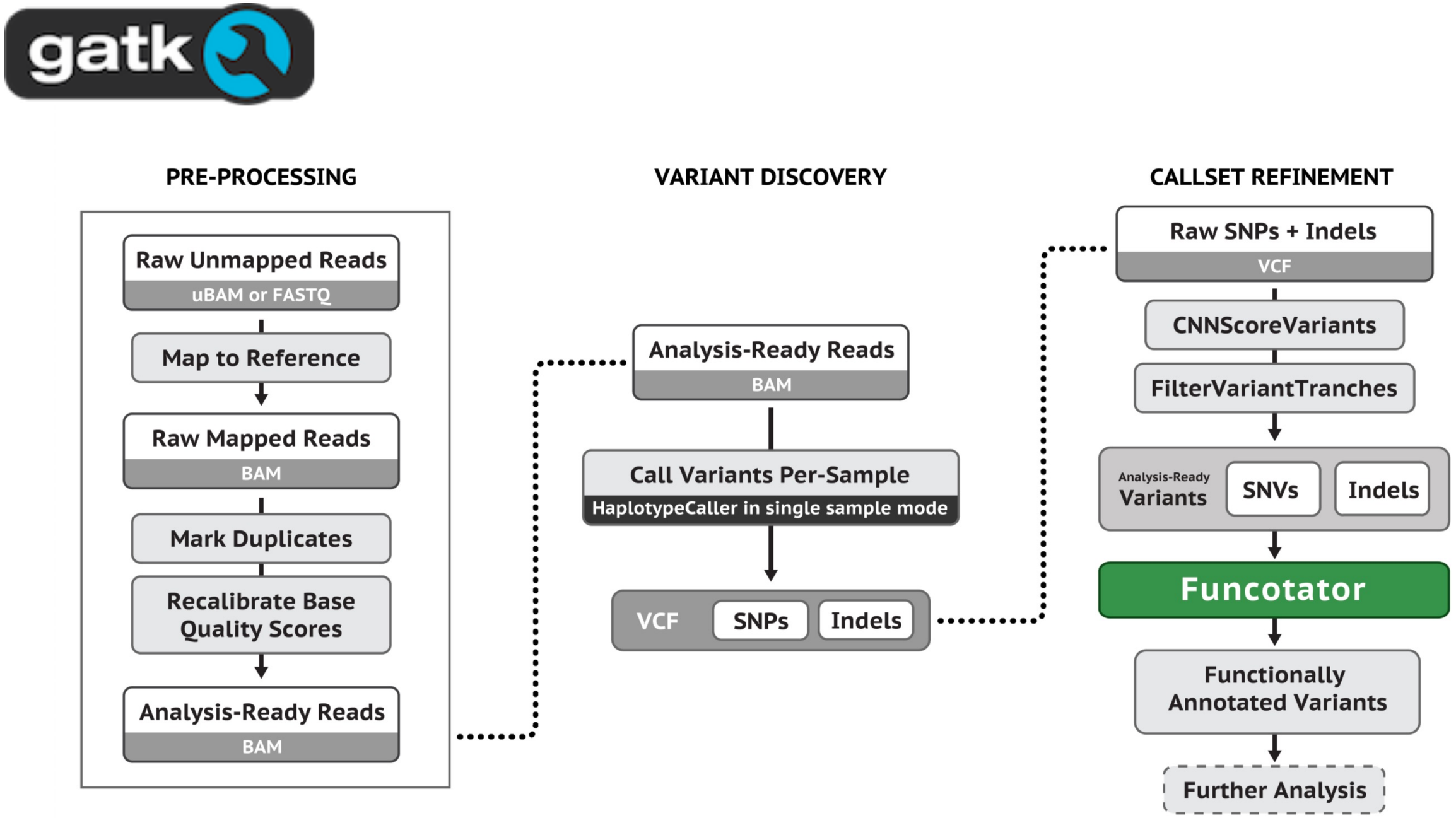
QNAME: query template name, aka. read ID



SAM and BAM file format

Tag	Type	Description [hide]
AM	i	The smallest template-independent mapping quality in the template
AS	i	Alignment score generated by aligner
BC	Z	Barcode sequence identifying the sample
BQ	Z	Offset to base alignment quality (BAQ)
BZ	Z	Phred quality of the unique molecular barcode bases in the OX tag
CB	Z	Cell identifier
CC	Z	Reference name of the next hit
CG	B,I	BAM only: CIGAR in BAM's binary encoding if (and only if) it consists of >65535 operators
CM	i	Edit distance between the color sequence and the color reference (see also NM)
CO	Z	Free-text comments
CP	i	Leftmost coordinate of the next hit
CQ	Z	Color read base qualities




PRE-PROCESSING
Raw Unmapped Reads
uBAM or FASTQ
Map to Reference
Raw Mapped Reads
BAM
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM
VARIANT DISCOVERY
Analysis-Ready Reads
BAM
Call Variants Per-Sample
HaplotypeCaller in single sample mode
VCF
SNPs
Indels
CALLSET REFINEMENT
Raw SNPs + Indels
VCF
CNNScoreVariants
FilterVariantTranches
Analysis-Ready Variants
SNVs
Indels
Funcotator
Functionally Annotated Variants
Further Analysis