Molecular In My Pocket™...

Bioinformatics: Basic File Formats Part II

BAM, BED, and VCF (Secondary and Tertiary Analysis)

Binary Alignment Map (BAM): A compressed binary version of the text-based Sequence Alignment/Map (SAM) format used to store sequence alignment data in a tab delimited text file. BAM files contain a header section and an alignment section. Note: certain platforms generate raw data in an unmapped BAM format.

Header Section: It begins with "@" and contains information about the entire file such as chromosomes, sample name, and alignment method. The optional tags are described in http://samtools.sourceforge.net/.

Alignment Section: It contains read name, read sequence, read quality, and alignment information. The rest of the fields are described in http://samtools.sourceforge.net/.

Common Field in Alignment Section	Field Description
QNAME	Read name.
FLAG	Alignment information about the read. The value is explained in https://broadinstitute.github.io/picard/explain-flags.html.
RNAME	Chromosome the read is mapped to.
POS	Position the read is mapped to. The first base in a chromosome is numbered 0.
MAPQ mapping	Mapping quality.
CIGAR	Summary of the alignment. The description can be found in http://samtools.sourceforge.net/.
SEQ	The sequence of the read.
QUAL	The quality scores of the read.

Example 1: The header.

```
@HD     VN:1.4
@SQ     SN:chr1     LN:249250621
...
@SQ     SN:chrY     LN:59373566
@RG     ID:sample1
@PG     ID:bwa
```

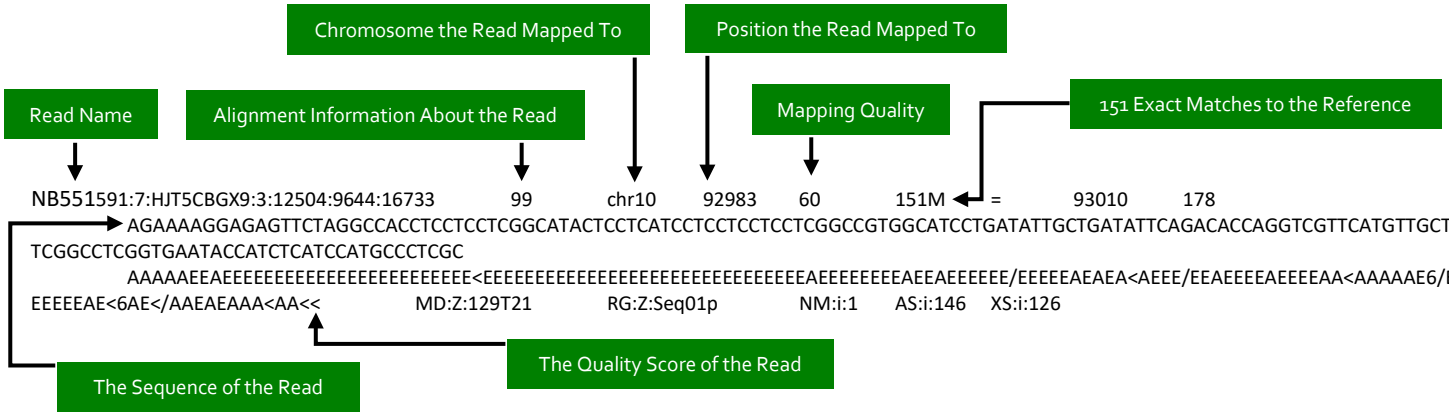
Format Version

Chromosome Name & Chromosome Length

Sample Name

Program Used for the Alignment

Example 2: The alignment of a mapped bam file.



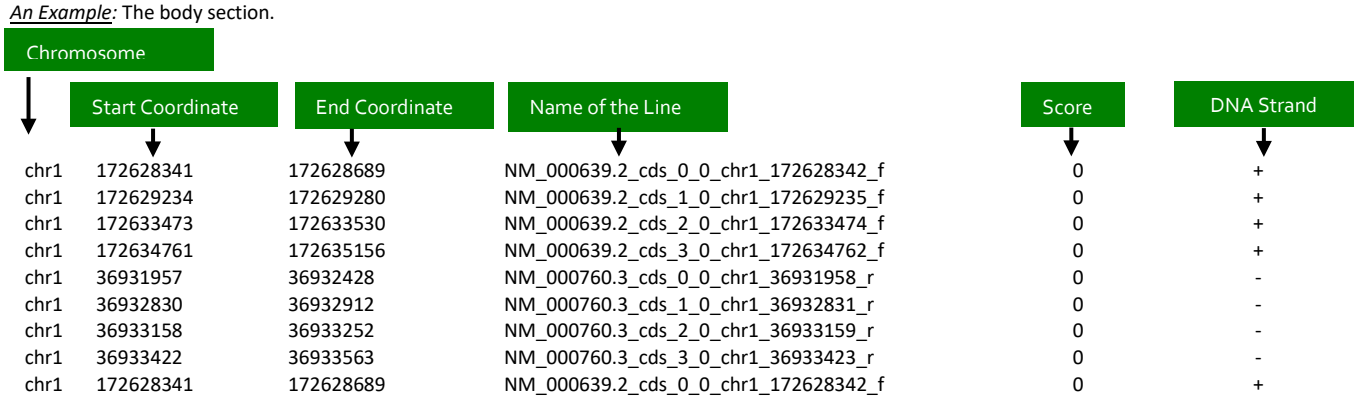
References

3. https://support.illumina.com/help/BS_App_RNASeq_Alignment_OLH_1000000006112/Content/Source/Informatics/BAM-Format.htm
4. http://samtools.sourceforge.net/
5. https://genome.ucsc.edu/FAQ/FAQformat.html#format1

Browser Extensible Data (BED): A text file format used to store genomic regions as coordinates and associated annotations. BED files contain a header section and a body section.
Header section: It is optional and begins with "browser or "track".
Body section: It follows the header and is whitespace or tab separated into 12 fields. The first 3 fields are required. The next 9 fields are optional and described in https://genome.ucsc.edu/FAQ/FAQformat.html#format1.

Common Field in Body Section	Field Description
chrom	Chromosome.
chromStart	Start coordinate on the chromosome. The first base in a chromosome is numbered 0.
chromEnd	End coordinate on the chromosome.
name (optional)	Name of the line in the BED file.
score (optional)	Score between 0 and 1000. The actual biological meaning depends on the performed experiment.
strand (optional)	DNA strand orientation.

An Example: The body section.



chr1	172628341	172628689	NM_000639.2_cds_0_0_chr1_172628342_f	0	+
chr1	172629234	172629280	NM_000639.2_cds_1_0_chr1_172629235_f	0	+
chr1	172633473	172633530	NM_000639.2_cds_2_0_chr1_172633474_f	0	+
chr1	172634761	172635156	NM_000639.2_cds_3_0_chr1_172634762_f	0	+
chr1	36931957	36932428	NM_000760.3_cds_0_0_chr1_36931958_r	0	-
chr1	36932830	36932912	NM_000760.3_cds_1_0_chr1_36932831_r	0	-
chr1	36933158	36933252	NM_000760.3_cds_2_0_chr1_36933159_r	0	-
chr1	36933422	36933563	NM_000760.3_cds_3_0_chr1_36933423_r	0	-
chr1	172628341	172628689	NM_000639.2_cds_0_0_chr1_172628342_f	0	+

Variant Call Format (VCF) format: The information about variants found at specific positions in a reference genome. VCF files contain a header section and a body section.
Header section: It begins with "##" and lists the annotations used in the file.
Body section: It follows the header and is tab separated into 10 fields. The column names begin with "#".

Field in Body section	Field Description
CHROM	The chromosome on which the variation is called.
POS	The position of the variation on the chromosome. The first base in a chromosome is numbered 1. The vcf standard has the indel shifted to the 5' most position on the reference.
ID	The identifier of the variation. Identifier can vary according to software (e.g., dbDNP id, COSMIC)
REF	The reference allele at the position on the chromosome.
ALT	The list of alternative alleles at the position.
QUAL	A quality score associated with the given alleles.
FILTER	The filters the variation has passed.
INFO	Fields describe the variation.
FORMAT (optional)	More fields describe the variation.
SAMPLE(s) (optional)	The values specified in the FORMAT.

An example: The body section.