

Sequence Alignment/Map Format Specification

The SAM/BAM Format Specification Working Group

3 May 2023

The master version of this document can be found at <https://github.com/samtools/hts-specs>.
This printing is version 1f92c08 from that repository, last modified on the date shown above.

1 The SAM Format Specification

SAM stands for Sequence Alignment/Map format. It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. If present, the header must be prior to the alignments. Header lines start with '@', while alignment lines do not. Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information.

This specification is for version 1.6 of the SAM and BAM formats. Each SAM and BAM file may optionally specify the version being used via the @HD VN tag. For full version history see Appendix B.

~~Unless explicitly specified elsewhere, all fields are encoded SAM files are encoded in UTF-8, though most of their content is limited to ASCII. They must not begin with a byte order mark, and non-ASCII characters are permitted only in certain field values as individually specified.¹ Where it makes a difference, SAM file contents should be read and written using 7-bit US-ASCII² in using the POSIX /C locale. Regular expressions listed use the POSIX²~~

~~The regular expressions in this specification have been written using the POSIX /IEEE Std 1003.1 extended syntax.~~

1.1 An example

Suppose we have the following alignment with bases in lowercase clipped from the alignment. Read r001/1 and r001/2 constitute a read pair; r003 is a chimeric read; r004 represents a split alignment.

```
Coor      12345678901234 5678901234567890123456789012345
ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1      TTAGATAAAGGATA*CTG
+r002        aaaAGATAA*GGATA
+r003        gcctaAGCTAA
+r004                ATAGCT.....TCAGC
-r003                ttagctTAGGC
-r001/2                        CAGCGGCAT
```

¹~~Equivalently, SAM file content is primarily US-ASCII characters in the usual single-byte encoding; certain field values as specified may contain other Unicode characters and are encoded as UTF-8.~~

²~~Charset ANSI X3.4-1968 as defined in RFC1345.~~

²~~For example, floating-point values in SAM always use '.' (FULL STOP) for the decimal-point character.~~