Phred scale Given a probability $0 < p \leq 1$, the phred scale of p equals $-10 \log_{10} p$, rounded to the closest integer.

1.2.1 Character set restrictions

Reference sequence names, CIGAR strings, and several other field types are used as values or parts of values of other fields in SAM and related formats such as VCF. To ensure that these other fields' representations are unambiguous, these field types disallow particular delimiter characters.

Query or read names may contain any printable ASCII characters in the range `[!-~]` apart from '`@`', so that SAM alignment lines can be easily distinguished from header lines. (They are also limited in length.)

Reference sequence names may contain any printable ASCII characters in the range `[!-~]` apart from backslashes, commas, quotation marks, and brackets—i.e., apart from '`\,"‘’()[]{}<>`'—and may not start with '`*`' or '`=`'.[4]

Thus they match the following regular expression:

```
[0-9A-Za-z!#$%&+./:;?@^_|~-][0-9A-Za-z!#$%&*+./:;=?@^_|~-]*
```

For clarity, elsewhere in this specification we write this set of allowed characters as a character class `[[:rname:]]` and extend the POSIX regular expression notation to use `^*=` to indicate the omission of '`*`' and '`=`' from the character class. Thus this regular expression can be written more clearly as `[[:rname:^*=]][[:rname:]]*`.

1.3 The header section

Each header line begins with the character '`@`' followed by one of the two-letter header record type codes defined in this section. In the header, each line is TAB-delimited and, apart from `@CO` lines, each data field follows a format '`TAG:VALUE`' where `TAG` is a two-character string that defines the format and content of `VALUE`. Thus header lines match `/^@(HD|SQ|RG|PG)(\t[A-Za-z][A-Za-z0-9]:[-[:print:]]+)+$/` or `/^@CO\t.*/`.[5] Within each (non-`@CO`) header line, no field tag may appear more than once and the order in which the fields appear is not significant.

The following table describes the header record types that may be used and their predefined tags. Tags listed with '*' are required; e.g., every `@SQ` header line must have `SN` and `LN` fields. As with alignment optional fields (see Section 1.5), you can freely add new tags for further data fields. Tags containing lowercase letters are reserved for local use and will not be formally defined in any future version of this specification.[6]

Tag	Description
`@HD`	File-level metadata. Optional. If present, there must be only one `@HD` line and it must be the first line of the file.
`VN*`	Format version. *Accepted format*: `/^[0-9]+\.[0-9]+$/`.

[3] Chimeric alignments are primarily caused by structural variations, gene fusions, misassemblies, RNA-seq or experimental protocols. They are more frequent given longer reads. For a chimeric alignment, the linear alignments constituting the alignment are largely non-overlapping; each linear alignment may have high mapping quality and is informative in SNP/INDEL calling. In contrast, multiple mappings are caused primarily by repeats. They are less frequent given longer reads. If a read has multiple mappings, all these mappings are almost entirely overlapping with each other; except the single-best optimal mapping, all the other mappings get mapping quality <Q3 and are ignored by most SNP/INDEL callers.

[4] Characters that are *not* disallowed include '`|`', which historically appeared in reference names derived from NCBI FASTA files, and '`:`', which appears in HLA allele names. Appendix A describes approaches for parsing *name*`[:`*begin*`-`*end*`]` region notation unambiguously even though *name* may itself contain colons.

[5] `[[:print:]]` indicates that header field values contain printable characters, i.e., non-control characters. For fields limited to ASCII, which is the majority, this is equivalent to `[ -~]`.

[6] Best practice is to use lowercase tags while designing and experimenting with new data field tags or for fields of local interest only. For new tags that are of general interest, raise an `hts-specs` issue or email `samtools-devel@lists.sourceforge.net` to have an uppercase equivalent added to the specification. This way collisions of the same uppercase tag being used with different meanings can be avoided.

	SO	Sorting order of alignments. *Valid values*: `unknown` (default), `unsorted`, `queryname` and `coordinate`. For coordinate sort, the major sort key is the RNAME field, with order defined by the order of `@SQ` lines in the header. The minor sort key is the POS field. For alignments with equal RNAME and POS, order is arbitrary. All alignments with '`*`' in RNAME field follow alignments with some other value but otherwise are in arbitrary order. For queryname sort, no explicit requirement is made regarding the ordering other than that it be applied consistently throughout the entire file.[7]
	GO	Grouping of alignments, indicating that similar alignment records are grouped together but the file is not necessarily sorted overall. *Valid values*: `none` (default), `query` (alignments are grouped by QNAME), and `reference` (alignments are grouped by RNAME/POS).
	SS	Sub-sorting order of alignments. Valid values are of the form *sort-order*`:`*sub-sort*, where *sort-order* is the same value stored in the `SO` tag and *sub-sort* is an implementation-dependent colon-separated string further describing the sort order, but with some predefined terms defined in Section 1.3.1. For example, if an algorithm relies on a `coordinate` sort that, at each coordinate, is further sorted by query name then the header could contain `@HD SO:coordinate SS:coordinate:queryname`.[8] If the primary sort is not one of the predefined primary sort orders, then `unsorted` should be used and the sub-sort is effectively the major sort. For example, if sorted by an auxiliary tag `MI` then by coordinate then the header could contain `@HD SO:unsorted SS:unsorted:MI:coordinate`. *Regular expression*: `(coordinate|queryname|unsorted)(:[A-Za-z0-9_-]+)+`
`@SQ`		Reference sequence dictionary. The order of `@SQ` lines defines the alignment sorting order.
	SN*	Reference sequence name. The `SN` tags and all individual `AN` names in all `@SQ` lines must be distinct. The value of this field is used in the alignment records in RNAME and RNEXT fields. Regular expression: `[[:rname:]^*=]][[:rname:]]*`
	LN*	Reference sequence length. *Range*: $[1, 2^{31} - 1]$
	AH	Indicates that this sequence is an alternate locus.[9] The value is the locus in the primary assembly for which this sequence is an alternative, in the format '*chr*`:`*start*`-`*end*', '*chr*' (if known), or '`*`' (if unknown), where '*chr*' is a sequence in the primary assembly. Must not be present on sequences in the primary assembly.
	AN	Alternative reference sequence names. A comma-separated list of alternative names that tools may use when referring to this reference sequence.[10] These alternative names are not used elsewhere within the SAM file; in particular, they must not appear in alignment records' RNAME or RNEXT fields. *Regular expression*: *name*`(,`*name*`)*` where *name* is `[[:rname:]^*=]][[:rname:]]*`
	AS	Genome assembly identifier.
	DS	Description. UTF-8 encoding may be used.
	M5	MD5 checksum of the sequence. See Section 1.3.2
	SP	Species.
	TP	Molecule topology. *Valid values*: `linear` (default) and `circular`.[11]
	UR	URI of the sequence. This value may start with one of the standard protocols, e.g., '`http:`' or '`ftp:`'. If it does not start with one of these protocols, it is assumed to be a file-system path.
`@RG`		Read group. Unordered multiple `@RG` lines are allowed.
	ID*	Read group identifier. Each `@RG` line must have a unique `ID`. The value of `ID` is used in the RG tags of alignment records. Must be unique among all read groups in header section. Read group IDs may be modified when merging SAM files in order to handle collisions.
	BC	Barcode sequence identifying the sample or library. This value is the expected barcode bases as read by the sequencing machine in the absence of errors. If there are several barcodes for the sample/library (e.g., one on each end of the template), the recommended implementation concatenates all the barcodes separating them with hyphens ('`-`').
	CN	Name of sequencing center producing the read.

[7] It is known that widely used software libraries have differing definitions of the queryname sort order, meaning care should be taken when operating on multiple files of varying provenance. Tools may wish to use the sub-sort field to explicitly distinguish between natural and lexicographical ordering. See Section 1.3.1.

[8] The repetition of *sort-order* enables a limited form of validation. For example, `@HD SO:queryname SS:coordinate:TLEN` would indicate that the data has been re-sorted (by query name) by a non-`SS`-aware tool and the `SS` field should be ignored.

[9] See `https://www.ncbi.nlm.nih.gov/grc/help/definitions` for descriptions of *alternate locus* and *primary assembly*.

[10] For example, given '`@SQ SN:MT AN:chrMT,M,chrM LN:16569 TP:circular`', tools can ensure that a user's request for any of 'MT', 'chrMT', 'M', or 'chrM' succeeds and refers to the same sequence.

[11] The previous footnote's example identifies MT as a circular chromosome. The `TP` field is often omitted, which implies linear.