pSAM Specifications

The original SAM fields can be seen here, https://samtools.github.io/hts-specs/SAMv1.pdf, section 1.4- The Alignment Section: Mandatory Fields

pSAM files have identical header and mandatory field entries as their corresponding SAM files, with the exception of the CIGAR field (column 6), the SEQ field (column 10), and the QUAL field (column 11). In a pSAM file, the CIGAR field is modified to reflect a match of the length of SEQ in the original SAM file (the sum of the lengths of the original CIGAR M/I/S/=/X operations), and the SEQ and QUAL fields are converted to '*'. Differences between pSAM and SAM are highlighted in red.

Please note that while the '*' ASCII character in a Phred +33 score of a SAM file corresponds to a P(error) of the match of 0.12589, in a pSAM file, there is no such meaning. The '*' is strictly a placeholder and is kept as such so that the file can later be converted back to the original SAM format if needed. The mandatory fields are described below.

Col	Field	Type	Regexp/Range	Brief description
~~1~~	~~QNAME~~	~~String~~	~~[!-?A-~]{1,254}~~	~~Query template~~ NAME
2	FLAG	Int	[0,216-1]	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	[0,231-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,28-1]	MAPping Quality
6	CIGAR	String	*|([0-9]+[M])	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	[0,231-1]	Position of the mate/next read
9	TLEN	Int	[-231+1, ,231-1]	Observed Template LENgth
10	SEQ	String	Reference Sequence	Segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

Furthermore, SAM files have optional fields which could potentially lead to additional private information leakage. pSAM files, therefore, have two optional fields that are modified when present: the AS:i field, NM:i field and MD:Z field.

From https://samtools.github.io/hts-specs/SAMv1.pdf, section 1.5:

All optional fields follow the TAG:TYPE:VALUE format where TAG is a two-character string that matches /[A-Za-z][A-Za-z0-9]/. Each TAG can only appear once in one alignment line. A TAG containing lowercase letters are reserved for end users. In an optional field, TYPE is a single case-sensitive letter which defines the format of VALUE:

Standard TAGs (from https://samtools.github.io/hts-specs/SAMtags.pdf, section 1), with their modification in red.

TAG	TYPE	DESCRIPTION
AS	i	Alignment score generated by aligner - Modified when present to reflect perfect alignment with reference genome.
MD	Z	String for mismatching positions - Modified when present to match read length
NM	i	Edit distance to the reference - Modified to 0 when present

Note that there may be other optional fields that could potentially leak sensitive information. We can easily develop additional modules to the current pTools to modify any optional tag in a SAM file.