.diff Specifications

The .diff ("Diff ") file has at most four fields per row. Three of these are found in the original SAM/BAM file, and one field (the "mod" field) is generated by pTools, as described in the table below. The original SAM fields can be seen here, https://samtools.github.io/hts-specs/SAMv1.pdf, section 1.4- The Alignment Section: Mandatory Fields.

Col	Field	Type	Regexp/Range	Brief Description
1	CIGAR	String	*|([0-9+[MIDNSHPX=])+	CIGAR string
2	MOD	String	([ATGCatgc]+:[SHIDX]-*)+	Modification string (generated by pTools)
3	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33
4	MD	String	(\[MD:Z:] [0-9]+[ATGCatgc]+)*	MD:Z string (optional field)
5	AS	String	(\[AS:i:] [0-9])*	AS:i string (optional field)
6	NM	String	(\[NM:i] [0-9])*	NM:i string (optional field)

1. CIGAR: The CIGAR string, described in the table below (from https://samtools.github.io/hts-specs/SAMv1.pdf)

OP	Description	Consumes Query	Consumes Reference
M	alignment match (can be sequence match or mismatch)	yes	yes
I	Insertion to the reference	yes	no
D	deletion from the reference	no	yes
N	skipped region from the reference	no	yes
S	soft clipping (clipped sequences present in SEQ)	yes	no
H	hard clipping (clipped sequences NOT present in SEQ)	no	no
P	padding (silent deletion form padded reference)	no	no
=	sequence match	yes	yes
X	sequence mismatch	yes	yes

2. MOD: The MODification string. Generated by pTools. All MOD strings follow the format nucleotide(s):modification type-*. Examples can be seen in the table below*:

Reference Sequence	CIGAR	SAM/BAM read sequence	MOD
AAAGCAAGGTGAGGA	3S4M2D3M3H	GCAATGA	AAA:S-GG:D-GGA:H
TGCCATA	2M1I5M	TGACCATA	A:I
GGGGGGGGGGA	10M1S	GGGGGGGGGG	A:S
CCGTG	2M3I1D2M	CCAAATG	AAA:I-G:D

3. QUAL: ASCII of base QUALity plus 33 (same as the quality string in the Sanger FASTQ format). A base quality is the phred-scaled base error probability which equals −10 log10 Pr{base is wrong}. This field can be a '*' when quality is not stored. If not a '*', SEQ must not be a '*' and the length of the quality string ought to equal the length of SEQ (from SAMtools specification).
4. MD: The MD:Z tag in the original SAM/BAM file (optional). Optional fields are usually displayed as TAG:TYPE:VALUE. From https://samtools.github.io/hts-specs/SAMtags.pdf:

   "MD:Z: [0-9]+(([A-Z]|\^[A-Z]+)[0-9]+)* String for mismatching positions. The MD field aims to achieve SNP/indel calling without looking at the reference. For example, a string '10A5^AC6' means from the leftmost reference base in the alignment, there are 10 matches followed by an A on the reference which is different from the aligned read base; the next 5 reference bases are matches followed by a 2bp deletion from the reference; the deleted sequence is AC; the last 6 bases are matches. The MD field ought to match the CIGAR string."

*Removed (deleted, soft clipped, or hard clipped) or inserted base pairs have been colored gray and red, respectively, solely to aid in the readability of this document; .diff files do not have colored base pairs.