uncomplemented 5' end of the SEQ field. This number series is comparable to the numbers in an MD tag, albeit counting specific base types only and potentially reverse-complemented.

For example 'C+m,5,12,0;' tells us there are three potential 5-Methylcytosine bases on the top strand of SEQ. The first 5 'C' bases are unmodified and the 6th, 19th and 20th have modification status indicated by the corresponding probabilities in the ML tag. The 12 cytosines between the 6th and 19th cytosine are unmodified. Modification probabilities for the 17 skipped cytosines are not provided.

When the '?' flag is present the tag 'C+m?,5,12,0;' tells us the modification status of the first five cytosine bases is unknown, the sixth cytosine is called (as either modified or unmodified), followed by 12 more unknown cytosines, and the 19th and 20th are called.

Similarly 'G-m,14;' indicates the 15th 'G' there might be a 5-Methylcytosine on the opposite strand (still counting using the top strand base calls from the 5' end). When the alignment record is reverse complemented (SAM flag 0x10) these two examples do not change since the tag always refers to the as-sequenced orientation. See the test/SAMtags/MM-orient.sam file for examples.

This permits modifications to be listed on either strand with the rare potential for both strands to have a modification at the same site. If SAM FLAG 0x10 is set, indicating that SEQ has been reverse complemented from the sequence observed by the sequencing machine, note that these base modification field values will be in the opposite orientation to SEQ and other derived SAM fields.

Note it is permitted for the coordinate list to be empty (for example 'MM:Z:C+m;'), which may be used as an explicit indicator that this base modification is not present. It is not permitted for coordinates to be beyond the length of the sequence.

When multiple modifications are listed, for example 'C+mh,5,12,0;', it indicates the modification may be any of the stated bases. The associated confidence values in the ML tag may be used to determine the relative likelihoods between the options. The example above is equivalent to 'C+m,5,12,0;C+h,5,12,0;', although this will have a different ordering of confidence values in ML. Note ChEBI codes cannot be used in the multi-modification form (such as the 'C+mh' example above).

If the modification is not one of the standard common types (listed below) it can be specified as a numeric ChEBI code. For example 'C+76792,57;' is the same as 'C+h,57;'.

An unmodified base of 'N' means count any base in SEQ, not only those of 'N'. Thus 'N+n,100;' means the 101st base is Xanthosine (n), irrespective of the sequence composition. A fundamental base of 'N' may also be used with a base-specific modification code to force the counting to be applied per base rather than per base-type.

The standard code types and their associated ChEBI values are listed below, taken from Viner *et al.*[5] Additionally ambiguity codes 'A', 'C', 'G', 'T' and 'U' exist to represent unspecified modifications bases of their respective canonical base types, plus code 'N' to represent an unspecified modification of any base type.

[5] Coby Viner *et al.*, *Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet*, https://www.biorxiv.org/content/10.1101/043794v1.

Unmodified base	Code	Abbreviation	Name	ChEBI
C	m	5mC	5-Methylcytosine	27551
C	h	5hmC	5-Hydroxymethylcytosine	76792
C	f	5fC	5-Formylcytosine	76794
C	c	5caC	5-Carboxylcytosine	76793
C	C		Ambiguity code; any C mod	
T	g	5hmU	5-Hydroxymethyluracil	16964
T	e	5fU	5-Formyluracil	80961
T	b	5caU	5-Carboxyluracil	17477
T	T		Ambiguity code; any T mod	
U	U		Ambiguity code; any U mod	
A	a	6mA	6-Methyladenine	28871
A	A		Ambiguity code; any A mod	
G	o	8oxoG	8-Oxoguanine	44605
G	G		Ambiguity code; any G mod	
N	n	Xao	Xanthosine	18107
N	N		Ambiguity code; any mod	
N	any		Mod applied to any base	

ML:B:C,`scaled-probabilities`

The optional `ML` tag lists the probability of each modification listed in the `MM` tag being correct, in the order that they occur. The continuous probability range 0.0 to 1.0 is remapped in equal sized portions to the discrete integers 0 to 255 inclusively. Thus the probability range corresponding to integer value N is $N/256$ to $(N+1)/256$.

The SAM encoding therefore uses a byte array of type '`C`' with the number of elements matching the summation of the number of modifications listed as being present in the `MM` tag accounting for multi-modifications each having their own probability.

For example '`MM:Z:C+m,5,12;C+h,5,12;`' may have an associated tag of '`ML:B:C,204,89,26,130`'.

If the above is rewritten in the multiple-modification form, the probabilities are interleaved in the order presented, giving '`MM:Z:C+mh,5,12; ML:B:C,204,26,89,130`'. Note where several possible modifications are presented at the same site, the `ML` values represent the absolute probabilities of the modification call being correct and not the relative likelihood between the alternatives. These probabilities should not sum to above 1.0 (≈ 256 in integer encoding, allowing for some minor rounding errors), but may sum to a lower total with the remainder representing the probability that none of the listed modification types are present. In the example used above, the 6th `C` has 80% chance of being `5mC`, 10% chance of being `5hmC` and 10% chance of being an unmodified `C`.

`ML` values for ambiguity codes give the probability that the modification is one of the possible codes compatible with that ambiguity code. For example `MM:Z:C+C,10; ML:B:C,229` indicates a C call with a probability of 90% of having some form of unspecified modification.

MN:i:`length`

The length of the `SEQ` field at the time the `MM` value was last written.

Some processing of aligned data, such as the use of hard-clipping tools, may alter `SEQ` sequence data. If the sequence is shortened in this manner then the base offsets in `MM` and `ML` become invalid unless they are also updated accordingly.

Some hard-clipping tools will update `MM`/`ML` but others do not, so the `MN` tag offers a simple sanity check. Software that wishes to validate `MM` should compare the length of the `SEQ` field with the contents of the `MN` tag—if they differ, the `MM` and `ML` values should be considered out-of-date. The tag is optional, but recommended, and if it is absent then there is an implicit assumption that the `MM` data is valid unless evidence implies otherwise (e.g., by having coordinates beyond the end of the sequence).