SAM Format

SAM Header

Record Type	Field	Description	Ultima-Specific	Example	Notes
@HD	VN	Format version	No	VN:1.6	
	SO	Sorting order of alignments	No	SO:coordinate	
@RG	ID	Read group identifier	No	ID:UGAv3-1	
	DT	Date the run was produced	No	DT:2021-09-13T07:32:05-0400	
	PL	Platform used to produce the reads	No	PL:ULTIMA	
	SM	Sample	No	SM:sm1	
	PU	Platform unit	No	PU:0583_1.20210905.CACATCCTGCATGTGAT	
	mc	Maximum class (h'mer length)	Yes	mc:12	
	BC	Barcode sequence	No	BC:CACATCCTGCATGTGAT	
	FO	Flow order	No	FO:TGCATGCATGCA…	
	PM	Platform model	No	PM:V1.3	
	vn	Data pipeline version	Yes	vn:4.0.15	
	tp	Orientation of the tp auxiliary tag in each record. Can be *synthesis* or *reference*.	Yes	tp:reference	
@CO	ring	Ring index, 1-based. Ring 1 is outermost.	Yes	ring:1	
	radius	Radius of ring in millimeters	Yes	radius:32	
	StartTheta	Starting angular position of ring, in turns (radians / (2 * PI))	Yes	StartTheta:1.095293	
	ThetaStep	Angular distance between scan lines, in turns	Yes	ThetaStep:0.00433593	
	tileHeight	Height of tile, in pixels	Yes	tileHeight:2048	
	tileWidth	Width of tile, in pixels	Yes	tileWidth:8192	

Alignment Fields

Field	Type	Description
QNAME	String	Read name in the format `<runid>-<barcode name>-<bead index>`

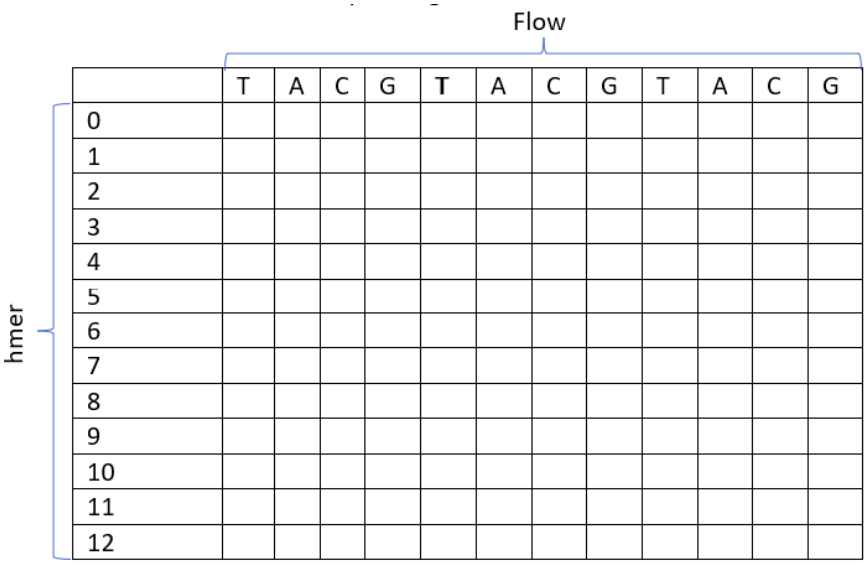
Auxiliary SAM Tags

Tag	Type	Format	Ultima-Specific	Description	Source	Example	Notes
RG	Z	RG:Z:readgroup	No	Read Group	Demultiplex	RG:Z:BC23	Required by spec
rq	f	rq:f:quality	Yes	Read Quality	Demultiplex	rq:f:1.73165	
pr	i	pr:i:index	Yes	Ring Index, 1-based	Demultiplex	pr:i:1	
pt	i	pt:i:index	Yes	Tile Index, 1-base	Demultiplex	pt:i:1	
px	i	px:i:xoffset	Yes	X position in tile	Demultiplex	px:i:3784	
py	i	py:i:offset	Yes	Y position in tile	Demultiplex	py:i:141	
tm	Z	tm:Z:[AQZ]+	Yes	Trimming Reasons: A=Adapter Q=Quality Z=Three Zeroes	probability	tm:Z:QZ	
si	i	si:i:intensity	Yes	Intensity of the first T flow	Demultiplex	si:i:10	
tp	B,c		Yes	Supplemental base quality information	probability	tp:B:c,2,-1,-2,-1,1,…	
t0	Z	t0:Z:string	Yes	Supplemental base quality information	probability		
X1	i	X1:i:count	No, but Ultima calculation	Count of alternate alignments with alignment score at least 90% of	UA	X1:i:16	

				selected alignment			
AS	i	AS:i:score	No	Alignment Score	UA	AS:i:91	
NM	i	NM:i:count	No	Edit distance to the reference	UA	NM:i:7	deleted during CRAM compression; can be recomputed
SA	Z	SA:Z:(rname ,pos ,strand ,CIGAR ,mapQ ,NM ;)+	No	Other canonical alignments in a chimeric alignment	UA	SA:Z:chr17,41490821,-,24H19M1D15M1I12M50H,0,2;	
XA	Z	XA:Z: (chr,pos,CIGAR,NM;)*	No	Alternative hits	UA		
XM	Z	XM:Z:string	No	Methylation call string	UA		
XG	Z	XG:Z:[CT,GA]	No	Which base conversion was performed on the reference: CT or GA	UA		
XR	Z	XG:Z:[CT,GA]	No	Which base conversion was performed on the read: CT or GA	UA		

Internal representation of a UG read – “flow matrix”

Each read is encoded by n_hmers x n_flows matrix. Position `(h, f)` in the matrix describes the probability that the signal at flow `f` was generated from the true flow `h`. Note that the convention we use are that the flow probabilities are normalized to 1. We call this representation “flow matrix”



Example for a flow matrix with 12 flows (3 cycles)

Calculation of QUAL and `tp` tag

1. Qual string and the `tp` tag encode the columns of the flow matrix for non-zero flows.
   a. A non-zero flow is a flow (column) where the maximal probability is not at H=0.
2. Specifically for hmer=H we encode up to min(4, floor((H+1)/2)) error probabilities.
3. QUAL encodes values of the probabilities while `tp` encodes the value of the error relatively to the called hmer (I.e. error h=3 if the called hmer is 4 will be encoded as --1)
4. Probabilities in QUAL are in phred-encoding
   a. See FASTQ format, Quality Score Encoding
   b. Notice usually QUAL encodes the probability that there is some error in the base, and here QUAL encodes the probability that a specific type of error will occur.
5. For convenience, the errors are encoded symmetrically relative to the middle of the hmer, with the nucleotide on each side of the hmer capturing half of the error probability
   a. The QUAL of the most outward bases of an hmer, will encode the most frequent error, and the quals of the most inward bases of an hmer encode the least frequent error (which is still represented)

Examples:

Example1

A T flow called as a 4-mer (TTTT) with the following probabilities:

	T
0	0
1	0
2	0
3	0.025
4	0.875
5	0.1
6	0
7	0
8	0
9	0
10	0
11	0
12	0

Bases	QUAL	tp
TTTT	.44.	+1, -1, -1, +1

Explanation:

The most likely error is H=5 with probability of 0.1

The Q-score for this error is: round(-10 * $\log_{10}$(0.1)) = 10.

The Q-score for **half** of this error probability is: round(-10 * $\log_{10}$(0.05)) = 13.

The encoding for this error is thus '.' , which will be encoded in association with the two outermost bases of the 4-mer.

The associated `tp` for this error is +1, since 5 = 4 +1

The next likely error is H=3 with probability of 0.025.

The Q-score for **half** of this error probability is: round(-10 * $\log_{10}$(0.025 / 2)) = 19.

The Q-score encoding for 19 is 4, which will be encoded in in association with the two innermost bases of the 4-mer.

The associated `tp` for this error is -1, since 3 = 4 -1

Example2

A T flow called as a 3-mer (TTT) with the following probabilities:

	T
0	0
1	0
2	0.025
3	0.875
4	0.1
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0

Bases	QUAL	tp
TTT	.1.	+1, -1, +1

Explanation:

The most likely error is H=4 with probability of 0.1.

The Q-score for **half** of this error probability is: round(-10 * $\log_{10}$(0.05)) = 13.

The encoding for this error is thus '.' , which will be encoded in association with the two outermost bases of the h-mer

The associated tp for this error is +1, since 4= 3 +1

The next likely error is H=2 with probability of 0.025.

Since the h-mer is odd, we don't split the error probability, and get a Q-score of 16 (enocding=1), which will be encoded in association with the center base of the h-mer.

The associated `tp` for this error is -1, since 2 = 3 -1

Example 3

A T flow called as a 4-mer (TTTT) with the following probabilities:

	T
0	0
1	0
2	0.001
3	0.025
4	0.874
5	0.1
6	0
7	0
8	0
9	0
10	0
11	0
12	0

Bases	QUAL	tp
TTTT	.44.	+1, -1, -1, +1

Explanation:

The most likely error is H=5 with probability of 0.1

The Q-score for this error is: round(-10 * $\log_{10}$(0.1)) = 10.

The Q-score for **half** of this error probability is: round(-10 * $\log_{10}$(0.05)) = 13.

The encoding for this error is thus '.' , which will be encoded in association with the two outermost bases of the 4-mer.

The associated `tp` for this error is +1, since 5 = 4 +1

The next likely error is H=3 with probability of 0.025.

The Q-score for **half** of this error probability is: round(-10 * $\log_{10}$(0.025 / 2)) = 19.

The Q-score encoding for 19 is 4, which will be encoded in in association with the two innermost bases of the 4-mer.

The associated `tp` for this error is -1, since 3 = 4 -1

The next likely error is H=2 with probability of 0.001. However, since we can only represent up to two error probabilities (floor((4+1)/2)=2) it is not represented in the output.

Example 4

A T flow called as a 5-mer (TTTTT) with the following probabilities:

	T
0	0
1	0
2	0
3	0.001
4	0.025
5	0.874
6	0.1
7	0
8	0

9	0
10	0
11	0
12	0
13	0

Bases	QUAL	tp
TTTTT	.4?4.	+1, -1, -2,-1, +1

Explanation:

The most likely error is H=6 with probability of 0.1

The Q-score for this error is: round(-10 * $\log_{10}$(0.1)) = 10.

The Q-score for **half** of this error probability is: round(-10 * $\log_{10}$(0.05)) = 13.

The encoding for this error is thus '.' , which will be encoded in association with the two outermost bases of the 5-mer.

The associated `tp` for this error is +1, since 6 = 5 +1

The next likely error is H=4 with probability of 0.025.

The Q-score for **half** of this error probability is: round(-10 * $\log_{10}$(0.025 / 2)) = 19.

The Q-score encoding for 19 is 4, which will be encoded in in association with the next two inner bases of the 5-mer.

The associated `tp` for this error is -1, since 4 = 5 -1

The next likely error is H=3 with probability of 0.001.

The Q-score for this error probability is: round(-10 * $\log_{10}$(0.001)) = 30.

The Q-score encoding for 30 is ?, which will be encoded in in association with the innermost base of the 5-mer. Note that in this case the whole error probablity is placed on the base since the middle base of the hmer is symmetric to itself.

The associated `tp` for this error is -2, since 3 = 5 - 2

Calculation of `t0` tag

1. `t0` tag indicates the probability that one of the neighboring flows was called zero although in reality it should be non-zero. Since the flow length was called 0, there is no base to encode this error in the base string and thus we encode the probability of these errors on the neighboring bases.
2. Specifically, for every zero-called flows we calculate the total probability that the call originated from non-zero genome.
3. The value of the `t0` tag of a non-zero flow is the maximum of the total error probabilities of the 0-4 neighboring zero flows.
4. `t0` tag value is encoded in the Phred scale similarly to the base qualities and is placed on all bases of the homopolymer

Example

Consider the following flow matrix:

flow	T	G	C	A	T	G
nuc						
call	3	0	1	0	0	1

P(0)	0	0.95	0	0.9	1	0
P(1)	0	0.05	1	0.1	0	1
P(2)	0	0	0	0	0	0
P(3)	0.99	0	0	0	0	0
P(4)	0.01	0	0	0	0	0
P(5)	0	0	0	0	0	0

The sequence that will be reported is `TTTCG` and the value of the `t0` tag will be `...++`.

Explanation

For the first T-flow the highest error probability of neighboring 0-mer is 0.05, which corresponds to `.` in Phred encoding. We place this quality on all bases of the corresponding hmer. For the C-flow (the third flow) the maximal error probability is 0.1, corresponding to `+` and the same for the G-flow (6th flow).