Documentation of SAM format in CRAC

Nicolas PHILIPPE and Mikaël SALSON

April 4, 2014

Contents

1 Introduction

As a starter it seems useful to state that SAM format is not the most adequate for CRAC's philosophy. CRAC does not see a read as a whole, and that's what make it powerful on longer reads. Rather, it considers portions of the read. Hence defining

some properties for a read globally (whether the read is multiple or unique, at what location occurs the read) is not straightforward and makes little sense.

However SAM format is read-centered. We therefore followed as much as possible this view. But that explains why a read can both be considered both unique, duplicated and multiple. In fact such a thing would mean that **parts** of the read are unique, parts are duplicated and other parts are occurring many times.

In our view, we're working on parts of a read, and trying hard to determine where that part is coming from and more specifically we're more interested in what is happening in the read. Does the read contain a SNP? an exon-exon junction?

Since the last version of SAM format specification (28 feb 2014) the concept of multiple or chimeric alignments has been introduced, therefore we have decided to integrate these specifications in CRAC's SAM output.

2 Mandatory fields

2.1 FLAG (bit)

CRAC set every flags documented in SAM specifications from 0x1 to 0x800 excepted flag 0x400 (PCR duplicates) and flag 0x200 (not passing quality control) that will never be raised.

2.2 Cigar (string)

In CRAC the cigar string is consistent with SAM specifications, however we are using `M` operator to describe a sequence match and of course `X` for a sequence mismatch.

2.3 Tlen (int)

In CRAC, Tlen is given from a genomic point of view even if we are working on transcriptomic data. It means that we get the genomic distance from the leftmost mapped base to the rightmost mapped base. If both reads of the pair completely overlap each other (ie. same sequence), Tlen is set to `0` as a single-end read does.

3 Extra fields

SAM format allows to provide extra fields after the eleven mandatory fields. Some of those fields are defined in SAM format specification (see `http://samtools.github.io/hts-specs/SAMv1.pdf`) and some others are defined by CRAC. We will first review standard optional fields that CRAC provides, then we will talk about CRAC's owns.

3.1 Standard fields

3.1.1 Paired-end/mates

If you run CRAC with paired-end reads by providing two files to the `-r` option, these additional fields will be printed for the primary line of both reads.

1. R2 (string) Paired-end/mate read's sequence as displayed in the corresponding SAM record. That means if the paired-end read's sequence has been reverse complemented to match the forward strand, R2 will also be reverse complemented.
2. MC (string) Cigar of paired-end/mate read as displayed in the corresponding SAM record. If the mate has multiple alignments we provide the cigar of the mate's primary line (see SAM specifications).
3. MQ (int) Mapping quality of paired-end/mate read as displayed in the corresponding SAM record. If the mate has multiple alignments we provide the mapping quality of the primary alignment (see SAM specifications).

3.1.2 Multiple alignment

One read may results in several alignments, and therefore more that one SAM record for the read (query).

1. Multiple hits

   If option –treat-multiple is activated and CRAC has found multiple perfect alignment for the read that will result in multiple SAM record. Those records will have extra fields to give informations on the next hit.

   (a) CC (string) Reference name (chromosome) of the next hit (field 3 in SAM record).

   (b) CP (int) Leftmost coordinates of the next hit (field 4 in SAM record).

2. Chimeric alignments

   If CRAC has detected a chimeric alignment for one read, each part of this chimeric alignment will be printed in a different SAM line. A supplementary field is added to give information about those alignments.

   (a) SA (string) Other canonical alignments in a chimeric alignment, in the format of: (rname,pos,strand,CIGAR,mapQ,NM ;)+. Each element in the semi-colon delimited list represents a part of the chimeric alignment. Conventionally, at a supplementary line, the first element points to the primary line.

3.1.3 Others

Two more fields are added to the SAM record of each alignment.

1. NH (int) Number of reported alignment for the read.
2. IH (int) Number of stored alignment in SAM for the read.

3.2 CRAC fields

3.2.1 XU (integer)

- 0 : read is not considered as unique.
- 1 : read is considered as unique (meaning that sufficiently k-mers are mapped unambiguously on the genome)

3.2.2 XD (integer)

- 0 : read is not considered as duplicated
- 1 : read is considered as duplicated (meaning that sufficiently k-mers are duplicated on the genome)

3.2.3 XM (integer)

- 0 : read is not considered as multiple
- 1 : read is considered as multiple (meaning that sufficiently k-mers are mapped at many positions on the genome)

3.2.4 XN (integer)

- 0 : read is not considered as normal
- 1 : read is considered as normal (meaning that nothing specific occurs in the read and the support is stable)
- 2 : read is considered as almost normal (the support is not stable, but almost stable)

3.2.5 XO (string)

Location on the genome of one *k*-mer in the read. The location is displayed in the following form: C|S,P where C is the chromosome, S is the strand (1 or -1) and P the position on the chromosome.

3.2.6 XQ (int)

Position in the read of the k-mer whose position is given in XQ.

3.2.7 XC (int)

Number of causes found in the read (ie. SNP, junctions, ...)

3.2.8 XX (string)

Details repetition portion(s) of the read is there is some. The syntax of the associated string is `start,end` where `start` if the first k-mer of the repetition and `end` the last one. `start` and `end` can be set to "unknown" if we do not now the start or the end of a repetition.

3.2.9 XP (string)

Aggregation of piece of informations about paired-end reads mapping. This field is only displayed for the first read of the pair (flagged with bit 0x40). Each piece of information is constituted by a couple of key, value : `(key:value;)+`. The following explanations give some details about the format of "value" depending on the "key" :

key = loc

The "key" `loc` gives us some details about the mapping of the paired-end read, the syntax of the "value" is `unique:duplicate:multiple` where `unique`, `duplicate` and `multiple` are respectively based on fields XU, XD, XM of the paired-end read associated with the query of the current record.

key = chimera

The "key" `chimera` is displayed when location of both read implicates a chimeric alignment. The syntax is `loc1,loc2`, where `loc1` correspond to the value given by field `XO` of the current read, and `loc2` the value given by the field `XO` of the paired-end read.

3.2.10 XE (string)

Details each cause found in the read. The string has the following form `(c:b:cause:...;)+` where `c` is the cause number (starting at 0), `b` is the break number (starting at 0, we have generally one cause per break) and `cause` is the name of the cause (SNP, Ins, Del, Junction, Error, chimera, Undetermined or BioUndetermined). The remaining of the string depends on the cause and is explained below

1. SNP `:score:position:loc:expected:actual`

    - `score` score of the SNP, the more negative the score is, the more confident the prediction, the closest to 0, the less confident it is.
    - `position` position in the read of the SNP (starting at 0)
    - `loc` location on the genome of SNP (the format used for describing the location is given in the description of XQ flag)
    - `expected` expected nucleotide (the nucleotide on the genome)

- `actual` actual nucleotide (the nucleotide in the read)

2. Ins/Del `:score:position:loc:nb`

   The three first fields are the same as in SNP. The last `nb` field is the number of nucleotide inserted or deleted.

3. Junction `:type:pos:loc:gap`

   - `type` is the type of junction, either `normal` or `coverless`. The latter corresponds to junctions that are not covered by many reads (in general only one).
   - `pos` is the position in the read of the junction
   - `loc` is the location on the genome of the end of the 5' exon
   - `gap` is the number of nucleotides that have been spliced

4. Error `:type:pos:score[:other1:other2]`

   - `type` is the type of sequence error either `Sub` (for one substitution), `Ins` (for an insertion), `Del` (for a deletion), `Unknown` (for an unknown error, when it is located on a read end). Depending on the type of error, there won't have the same number of fields or the same content. The two following fields are always the same.
   - `pos` position of the error in the read (starts at 0)
   - `score` score of the error, the higher the score, the more confident the prediction; the closest to 0, the less confident it is.
   - `other1` (does not exist for `Unknown` errors) either contains the original nucleotide on the genome (for `Sub` errors) or the number of inserted or deleted nucleotides (for `Del` and `Ins`)
   - `other2` (does not exist for `Unknown` errors) either contains the substituted nucleotide (`Sub` error) or the inserted or deleted sequence, if it is short enough, or <snip> if it is too long.

5. chimera `:pos:loc1:loc2`

   chimera corresponds to a fusion gene or a fusion transcript or to a sequencing bias that may have created a chimera.

   - `pos` position of the fusion point in the read
   - `loc1` location on the genome of the 5' part of the read
   - `loc2` location on the genome of the 3' part of the read.

6. Undetermined `:message`

   - `message` is a string enclosed between [and]. The message should explain why an accurate prediction has not been made.

7. BioUndetermined `:pos:message`

   The difference between an undetermined error and a bioUndetermined error is that CRAC considers that the bioUndetermined error is not a sequencing error and has, therefore, some biological relevance.

   - `pos` position in the read of the error
   - `message` message explaining the error, enclosed between [and].

3.2.11 XB (string)

If CRAC was provided the `--detailed-sam` option, some additional information will be displayed for every break. This additional information is useful if one wants to post-process the results. We provide the value of several variables that are useful to CRAC for making its predictions.

The provided string has the following format `XB:Z:i:var1=value1;var2=value2;[...]` where `i` is the break number (starting at 0), and then the variables with their values are separated by semicolons. The displayed variables are the following ones:

- `is_duplicated` boolean (0 or 1) stating if the break is considered as having a duplication
- `genome_indels` number of indels on the genome between the 5' and 3' part of the break, can be either negative (insertion) or positive (deletion) or C INT_{MAX} value if the value does not make sense
- `score_intra` score for discriminating a sequencing error from a biological cause, and assuming that we're still on the same transcript (therefore the coverage should not vary tremendously)
- `score_inter` same as above but assuming that the two parts of the read can originate from two different transcripts, expressed at different levels
- `deviation` score computed, on the support profile inside the break, between the average of the 50% highest values and the average of the 50% lowest values. It is a way of identifying a great variation of the support profile inside a break
- `falling_left` a boolean stating if the support profile falls on the 5' border of the break
- `falling_right` same as above but for the 3' border
- `inside_first_quartile` 25% of the support profile inside the break is lower or equal to that value
- `inside_last_quartile` 25% of the support profile inside the break is greater or equal to that value
- `inside_score` the support profile average inside the break
- `outside_score` the support profile average outside the break

- `average_low_inside` the average of the 50% lowest values inside the break
- `average_high_inside` the average of the 50% highest values inside the break
- `has_no_start_break` boolean stating that the break starts at the read's 5' end
- `has_no_end_break` boolean stating that the break ends at the read's 3' end
- `is_deviated` boolean stating that the support profile inside the break is deviated, meaning that the support profile inside the break changes a lot
- `is_nice_break` boolean stating that this break is nice: corresponds to a break that could represent a substitution, an insertion or a deletion (including junctions) but not a chimera.
- `is_very_nice_break` boolean stating that this break is very nice: corresponds to a break that represents a substitution or a small indel.
- `pos_start_gap` 5' position of the break in the read
- `pos_end_gap` 3' position of the break in the read

3.2.12 XR (string)

If CRAC was provided the `--detailed-sam` option, some additional information will be displayed for the read. This additional information consists of the support and location profiles.

The provided string has the following format `XR:Z:i:p_support=[...];p_loc=[...]`

- `p_support` is a list of numbers. This list is the support profile. It consists of the number of occurrences of each k-mer from the read, in the read collection. In the list, numbers are separated by comma. A `p_support=3,2,2,3` means that first k-mer in the read appears 3 times among all the reads while, the following two appears twice and the last k-mer in the read appears three times in all the reads.
- `p_loc` is also a list of numbers, which has the same size as $p_{support}$. This list is the location profile. It consists of he number of occurrences of each k-mer from the read, on the genome. A `p_loc=1,1,0,0` means that the first two k-mers appear once in the genome while the two last k-mers do not appear in the genome.