Documentation of homemade output formats in CRAC

Nicolas PHILIPPE and Mikaël SALSON

April 4, 2014

Contents

1 Introduction

SAM format is read-centered and, in CRAC's philosophy, we do not see a read as a whole. Rather, it considers portions of the read. Hence, we propose homemade formats for CRAC to classify all breaks for each read. In other words, a same read can be classified[1] several times in different files.

2 Description of each field

In each homemade format, there are some header lines. In each header line, there are some field and it is not easy to understand these fields without any explanation. This is a description of the fields.

First of all, in every field, when we talk about position `pos` or location `loc`, the position or the location always start at 0.

- `read_id` is the read number in input file. In case of paired reads, numbers for read_{id} are interlaced, ie. 0,2,4,...,N-2 for reads of the first paired file and 1,3,5,...,N-1 for the second paired file.

- `single_loc_on_genome` is the coordinate chr|strand,$\text{relative}_{\text{pos}}$ of a representative k-mer on the reference index. In fact, this coordinates identify the location of the read on the reference index used for the mapping process.

[1] We cannot properly talk about classification, since a read may contain at the same time a SNP, a splice junction and a sequencing error. Therefore it can be "classified" in three different places. However we use the term classification as it is more convenient.

- `occurrence_loc_on_genome` is the same as above except that no representative k-mer with a single loc has been found, so one of the multiple locations was given.

- `pos_single_loc_on_read` is the position of the k-mer used for the `single_loc_on_genome` in the read.

- `read` corresponds to the nucleotide sequence of the read.

- `p_support` represents a profile of all k-mers support along the read.

- `p_loc` represents a profile of all k-mers location (number of locations on the reference for each k-mer) along the read.

- `pos_start_repeat_on_read` is the position that corresponds to the beginning of a repeated factor in the read. A repeated factor is a factor which is located inside a repeated region on the reference index.

- `pos_end_repeat_on_read` is the position that corresponds to the end of a repeated factor in the read.

- `tag_snv`, `tag_indel`, `tag_splice` or `tag_chimera` is a tag to indicate if the biological event is ambiguous or not. Tag with a value to "single" means that the event is unique and a tag with a value to "duplicate" means that the event is ambiguous.

- `score` is a score given by CRAC to give a relevance for a sequence error or a biological event. The threshold is 0, a negative score means that a biological event is found while a positive score means that a sequence error is found.

- `snv` is a chain composed by two nucleotides and the symbol "->". The first nucleotide corresponds to the reference index and the second corresponds to the read.

- `loc_snv_on_genome` is the location of the k-mer on the genome immediately before the snv.

- `pos_snv_on_read` is the position of the snv on the read.

- `splice_length` corresponds to the length of the splice, ie. the distance between the end of the first exon and the start of the second exon.

- `loc_end_first_exon_on_genome` is the location of the last k-mer located just before the junction, ie. the last k-mer of the first exon.

- `loc_start_second_exon_on_genome` is the location of the first k-mer located just after the junction, ie. the first k-mer of the second exon.

- `pos_junction_on_read` is the position of the junction in the read.

- `bioUndetermined_cause_features` or `undetermined_cause_features` is a message to indicate why the break could not be classified somewhere above.

- chimera_flag is a flag to explain why we have classified a biological event as a chimera. This is a bitwise flag described by the following table.

Bit	Description
0x01 (1)	The exons are located on different chromosomes.
0x02 (2)	The exons are colinear but (likely) belong to different genes; this must be checked with annotation.
0x04 (4)	The exons are on the same chromosome and same strand, but not in the order in which they are found on DNA, and they do not overlap each other.
0x08 (8)	The exons are on the same chromosome but on different strands.