Preprocessing steps

USF Math-Bio Group / Georgia Tech Storici Labs

May 7, 2022

Input

- SAM file format: https://github.com/samtools/hts-specs/blob/master/SAMv1.pdf
- SAM tags: https://github.com/samtools/hts-specs/blob/master/SAMtags.pdf
- Bowtie2 specific tags: http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#reporting

Common filters

- Input: SAM file from Bowtie2.

1. Discard if read does not align with position 1 on reference (`POS != 1`).
2. Discard if read did not align at all (`FLAG & 4 != 0`).

0 mut filters

- Input: SAM file from Bowtie2, `min_length`.

1. Apply common filters.
2. Discard if `XG` $>$ `0` (number of gap-extends, AKA in/dels, see Bowtie2 spec).
3. Discard if `len(seq)` `<` `min_length`.

Middle indel mut filters

- Input: SAM file from Bowtie2, `min_length`, `dsb_pos`.

1. Apply common filters.
2. Discard if `XG == 0` (number of gap-extends, AKA in/dels, see Bowtie2 spec).
3. Discard if `len(seq) < min_length`.
4. Set `indel_ranges` to contiguous ranges of indels (example later).
    a. Discard if more than one contiguous range (`len(indel_ranges) > 1`).
    b. Discard if DSB site does not touch the range (`dsb_pos not in range(indel_ranges[0][0], indel_ranges[0][1] + 1)`).

Example: getting in/del range

```
Pos   : 12  3456   7891111111
                        0123456
Ref   : CG--CGAT---CAGCTACTAG
Read  : CGATCGATTGC---CTACTAG
        ->
Range1 : [2, 2]
Range2 : [6, 9]
```