Read Descriptors

The BARNA Descriptor

The BARNA (Barcelona Attributes for RNA-Seq) descriptor is a proposition for annotating unique read IDs with additional information yielded from special experiments, e.g., read mates derived from the same cDNA molecule, information about the original transcript sequence orientation, etc. It replaces the [[include FMRD descriptor proposition for the Flux Capacitor. BARNA expects all attributes as flags, i.e., a pre-defined set of characters, which are recognized in the read suffix after the last '/' character of the read ID. Currently defined flags are:

Flag	Meaning	Context
1	first mate of a read pair	paired-end sequencing
2	second mate of a read pair	paired-end sequencing
s	read in sense to transcription orientation	strand-specific RT
a	read in antisense to transcription orientation	strand-specific RT

Example:

```
chr1 10041 10109 BILLIEHOLIDAY:6:90:1240:1493/2a 1 + 0 0 0,0,0 1 68
chr1 10105 10181 TUPAC:7:103:90:14/1a 10 - 0 0 0,0,0 1 76 0
chr1 10138 10206 TUPAC:7:117:1290:277/2s 2 + 0 0 0,0,0 1 68 0
chr1 10223 10294 TUPAC:8:48:251:1564/1 10 - 0 0 0,0,0 1 71 0
```

These read identifiers have been retrieved from an experiment with strand-specific reverse transcription (RT) and paired-end sequencing.

1st line: the read is the (non-ordered) second mate of a read pair with the common and unique ID BILLIEHOLIDAY:6:90:1240:1493. Moreover, it is know that the read aligns in antisense to the transcription directionality and to the + strand of the genome (i.e., the so-called "Watson strand"). Thus, the transcript sequence it has been derived from stems from the - strand of the genomic sequence ("Crick strand"), and by mate orientations produced by current paired-end technologies the mating read BILLIEHOLIDAY:6:90:1240:1493/1s should map to the - strand of the reference genomic sequence.
2nd line: first mate of read pair TUPAC:7:103:90:14 which aligns to the negative strand. It's mate TUPAC:7:103:90:14/2s should align to the Watson strand.
3rd line: second mate on a transript on that has been transcribed from the + strand of the genomic sequence.
4th line: first mate of read pair TUPAC:8:48:251:1564, that aligns to the - strand of the genome for which the information about transcription directionalty has been lost (that happens). If the strandedness information has been conserved in its mate TUPAC:8:48:251:1564/2, it can be recovered.

Regular Expressions for Descriptors

Regular expressions allow for a flexible description of attribute retrieval from generic read descriptors. A summary of the regular expression system adopted in the Flux Capacitor can be foundthere.

Example 1

```
chr1 10033 10109 PAN:2:27:1091:1987/2
```

the example shows a standard Illumina/Solexa identifier for paired sequencing, where /1 indicates the first mate of a pair and /2 the second mate. A corresponding regular expression would target the suffix by

```
/([12])_strand([12])$
```

Example 2

```
chr1 10033 10109 PAN:2:27:1091:1987/2_strand1
```

where /2 indicates the 2nd mate, and '_strand1' additionally indicates read orientation in sense to the transcription directionality could be described by the regular expression

```
/([12])_strand([12]??)$
```

Note: exactly two symbols are expected for identifying paired mates as well as sense/anti-sense orientation. In the case the information is optional, add