4.5.2 - Read Identifiers

The Read Identifier

The read identifier produced by the Flux Simulator encode some of the information about where the read originated from in the simulation. In the BED format, these read identifiers can be found in the 4th column, in FASTA/FASTQ files they correspond to the identifier line without the initial '>' respectively '@' character. Simulated read identifier are colon separated, where each token corresponds to a certain information.

Nr	Name	Type	Example	Description
1	Reference ID	String	chr1	Name of the reference sequence, usually the chromosome, from which a read has been sequenced
2	Locus ID	[0-9]+\-[0-9]+[WC]	4847775-4887990W	Genomic start and end position, and the strand of the locus from which the read has been obtained; W denotes to the Watson strand (i.e., transcribed RNA will have the same directionality as the genomic reference), C denotes the Crick strand (i.e., transcribed RNA sequences are reverse-complemented substrings of the genomic reference sequence)
3	Transcript ID	String	NM_001159750	Identifier of the transcript form that produced the read
4	Molecule Nr	Integer	1	Number, i.e., identifier, of the specific molecule that has been simulated from the transcript form
5	Annotated Length	Integer	2668	Length of the transcript form as annotated in the reference annotation, after removal of introns and without considering simulated variations of the transcription start respectively the poly-A tail
6	Fragment Start	Integer	917	Start position of the fragment from which the read has been derived. Coordinates are provided relative to the annotated transcription start, excluding introns (i.e., relative positions in the processed transcripts). Negative values can occur where in silico TSS variations move the transcription start site to further upstream locations, values greater than the annotated length are in the poly-A tail.
7	Fragment End	Integer	1137	End position of the fragment from which the read has been derived. Coordinates are provided relative to the annotated transcription start, excluding introns (i.e., relative positions in the processed transcripts). Negative values can occur where in silico TSS variations move the transcription start site to further upstream locations, values greater than the annotated length are in the poly-A tail.
8	Relative Orientation	[AS]	S	Orientation of the read relative to the transcription directionality. S stands for sense, A for anti-sense. Note that this is not the absolute directionality with respect to the reference chromosome seuqence, e.g., an anti-sense read of a form produced from a locus that is transcribed from the Crick strand reproduces a substring in the same orientation as the reference genomic sequence.

Example

```
chr1:4847775-4887990W:NM_001159750:1:2668:917:1137:S/2
```

Unique IDs for paired reads

In order to support ids for paired reads, the Simulator Parameters support a `UNIQUE_IDS` options. If this option si senabled and paired reads are simulated, the inforamation about the relative orientation is not added to the read ids. This allows the simulator to create read ids for a set of paired reads that are unique except for the `/1 /2` pair identifiers. Because it is a random process and does not effect the result, the simulator incorporates the information about the relative orientation into the /1, /2 pair identifier. If `UNIQUE_IDS` is enabled, all sense reads (S without `UNIQUE_IDS`) will be `/1`, and all anti-sense reads (A without `UNIQUE_IDS`) will be `/2`.

For example, here is a set of paired read ids **without** `UNIQUE_IDS`:

```
chr1:4847775-4887990W:NM_001159750:1:2668:917:1137:A/1
chr1:4847775-4887990W:NM_001159750:1:2668:917:1137:S/2
```

With `UNIQUE_IDS` **enabled**, the same ids become:

```
chr1:4847775-4887990W:NM_001159750:1:2668:917:1137/1
chr1:4847775-4887990W:NM_001159750:1:2668:917:1137/2
```


With UNIQUE_IDS enabled, in the current Simulator (version 1.2.1) the reads produced carry a strand information, as by mate 1 always is sense (and mate 2 correspondingly anti-sense). See also a corresponding discussion in the forum.