4.5.3 - Output Read Sequences

Parameter Name	Default Value	Description
FASTA	false	Creates .fasta/.fastq output. Requires the genome sequences in a folder specified by GEN_DIR. If a quality model is provided by parameter ERR_FILE, a .fastq file is produced. Otherwise read sequences are given as .fasta.
GEN_DIR	null	Path to the directory with the genomic sequences, i.e., one fasta file per chromosome/scaffold/contig with a file name corresponding to the identifiers of the first column in the GTF annotation.
ERR_FILE	null	Path to the file with the error model. With the values '35' or '76', default error models are provided for the corresponding read lengths, otherwise the path to a custom error model file is expected.

Given a directory with genomic sequences split by chromosome GEN_DIR, Flux Simulator provides the possibility to additionally output the read sequences in FASTA or FASTQ format. If no error model ERR_FILE is provided, read sequences are an exact copy of the genomic sequence. Sequences of reads that are sequenced in antisense to the cDNA molecule are reverse complemented. Parts of the read that fall into the poly-A tail are correspondingly filled with a, respectively t characters whenever the read is produced in antisense direction. As in the BED file, the read identifiers are unique tags, composed of locus, transcript and fragment information from which they have been derived.

Example

A BED line

```
Chr1    28795    28871    Chr1:23259-31337W:AT1G01046.1:1:207:65:258:A    0    -    .    .    0,0,0    1
76    0
```

translates in a FASTA file to the line tuple

```
>Chr1:23259-31337W:AT1G01046.1:1:207:65:258:A
AACAAAGAAGCGTTAATTTATCGGTTATATCATTAAATTGTTAAAGTGAAAAGAATTTCTTATAACCTGACTGTTC
```

and can produce the FASTQ lines

```
@Chr1:23259-31337W:AT1G01046.1:1:207:65:258:A
AACAAAGAAGCGTTAATTTATCGGTTATATCATTAAATTGTTAAAGTGAAAAGAATTTCTTATAACCTGACTGTTC
+
IIIIHIIIIIIIIIIIIIIHG<2BBIIIIIIFIIIIE<BEHIIIIBGDDIHFG<ACCCCCDD:66CFEGHIFFBHI
```

Examples:
Here an example for a BED line that represents a spliced read

```
chr1 2082 2503 chr1:1116-4272W:uc009vip.1:105:2772:695:1003:968:1003|P2 0 - 0 0 0,0,0 2 8,28 0,393
```

The complete region of the read spans from 2083 (note the 0-base in BED format) to position 2503 (which is the first excluded position in BED format and therefore directly translates to the last included position in a 1-based coordinate system) on the reference sequence chr1. The the read alignment is split in two parts, one from 2083 to 2083+8-1=2090, and the other one from 2083+393=2476 to 2476+28-1=2502. The name field denotes that the read has been the downstream mate P2 of a read pair, derived from the 105*th* transcript copy of the annotated uc009vip.1 structure (which has spliced length 2772) in splicing locus chr1:1116-4272W. The fragment of this transcript that has been sequenced starts at position 695 and ends at position 1003 in the spliced sequence, relative to the annotated transcription start. From this fragment, the subarea 968-1003 relative to the annotated transcription start has generated the read.

This bed line is translated to FASTA format as:

```
>chr1:1116-4272W:uc009vip.1:105:2772:695:1003:968:1003|P2|chr1:2082:2503:-:8,28:0,393

GAAGGGCATGCCTGGCATCACCACACACTGGCCTAG
```

where the tag corresponds to the name field (number 4) in the BED format and the information about its genomic location as by BED coordinates concatenated from fields 1,2,3,6,11 and 12.