3.3 - Setting Up Simulations

The Command Line

The Flux Simulator reads the parameter values from a file which is to be specified from the command line.

```
$ flux.sh -t simulator -x -l -s -p myParameters.par
```

The example carries out a complete simulation pipeline, including simulated expression ("-x" flag), library construction ("-l" flag), and sequencing ("-s" flag). With the "-p" flag a parameter file with the description of the corresponding pipeline is passed to the program.

A Minimal Example of a Simulation

The kind of parameters and values which are required to be set in the file myParameters.sh depend on the desired behavior of the simulation. Most of these parameters have default values, however, the Flux Simulator requires in any case a qualitative annotation of transcripts, i.e., their intron-exon structure described by genomic coordinates in a GTF file. Therefore, a minimal parameter file consists of exclusively the line

myParameters.par	
REF_FILE_NAME	myTranscriptome.gtf

Note: With exclusively qualitative data about the transcripts--i.e., their genomic location and their mutual overlap--no sequence-specific attributes are taken into account in the simulation; that includes experimental characteristics caused by sequence biases as well as the *in silico* production of read sequences, which potentially are affected by sequencing errors.

Requirement for Simulations on Transcript/Read Sequences

If the simulation should produce read sequences, respectively if intermediate steps of the simulation are to take into account sequence-dependent biases, the Flux Simulator program requires to provide the genomic sequences of the chromosomes respectively scaffolds on which the genes have been annotated.

myParameters.par	
REF_FILE_NAME	myTranscriptome.gtf
GEN_DIR	<path>/myGenome/

where <path> is the path in the file system pointing to the folder myGenome, which contains the (complete) set of chromosomes for transcripts in the annotation myTranscriptome.gtf; the names of the files in the myGenome folder **must** **coincide** with the names of the sequences provided in the first column of the myTranscriptome.gtf file:

```
$ head -n2 myTranscriptome.gtf

10  EnsEMBL  exon  123  456  .  +  .  gene_id="gene1"; transcript_id="transcript1";
10  EnsEMBL  exon  789  1012  .  +  .  gene_id="gene1"; transcript_id="transcript1";

$ ls myGenome

10.fa  12.fa  14.fa  16.fa  18.fa  1.fa   21.fa  2.fa  4.fa  6.fa  8.fa  X.fa
11.fa  13.fa  15.fa  17.fa  19.fa  20.fa  22.fa  3.fa  5.fa  7.fa  9.fa  Y.fa
```

My First Simulation

Let's consider the following parameter file

REF_FILE_NAME	**/Users/micha/annotations/hg19_RefSeq_2009-05-13.gtf**	
POLYA_SHAPE	5	
POLYA_SCALE	100	
FRAG_METHOD	**NB**	
FRAG_SUBSTRATE	**DNA**	
FILTERING	**ON**	

READ_LENGTH	**75**	
PAIRED_END	**TRUE**	

In this example, RNA molecules as annotated in the RefSeq annotation are simulated to be expressed with about normally distributed polyA-tail lengths of an average size of 100nt. Fragmentation by nebulization (FRAG_METHOD NB) is carried out after reverse transcription (FRAG_SUBSTRATE DNA). The default size selection is carried out (FILTERING ON). Finally, 75nt paired-end reads are obtained from the fragments.