5.3 Poly-dT Priming and DNAse Digestion (S.cerevisiae)

The simulation of RNA-Seq in *Saccharomyces cerevisiae* joins a reverse transcription model by poly-dT primers with subsequent fragmenation by DNAseI. Sequence biases that have been reported for the DNAseI fragmentation process (Hansen et al. 2010) are captured in the simulation by a position weight matrix (DNAseI.pwm).

Input

Download

Reference Annotation

Parameter File

Reference Genome

Parameter

Expression		
NB_MOLECULES	5,000,000	Number of RNA molecules initially in the experiment
TSS_MEAN	25	Average deviation from the annotated transcription start site (TSS)
POLYA_SCALE	80	Scale of the Weibull distribution, shifts the average length of poly-A tail sizes
POLYA_SHAPE	2	Shape of the Weibull distribution describing poly-A tail sizes
Reverse Transcription		
RTRANSCRIPTION	YES	Switch on the reverse transcription
RT_PRIMER	PDT	Use poly-dT primers used for first strand synthesis
RT_LOSSLESS	YES	Flag to force every molecule to be reversely transcribed
RT_MIN	500	Minimum length observed after reverse transcription of full-length transcripts
RT_MAX	2,500	Maximum length observed after reverse transcription of full-length transcripts
Fragmentation		
FRAG_SUBSTRATE	DNA	Specifies DNA as the substrate of fragmentation
FRAG_METHOD	EZ	Enzymatic digestion as fragmentation method
FRAG_EZ_MOTIF	DNAseI.pwm	Fragmentation by enzymatic digestion
Amplification and Size Segregation		
PCR_DISTRIBUTION	default	Default PCR distribution with 15 rounds and 20 bins
GC_MEAN	0.5	Mean value of a gaussian distribution that reflects GC bias amplification probability
GC_SD	0.1	Standard deviation of a gaussian distribution that reflects GC bias amplification probability
FILTERING	YES	Enables size filtering of fragments
SIZE_SAMPLING	MH	The Metropolis-Hastings algorithm is used for filtering
Sequencing		
READ_NUMBER	1,000,000	Produce 1 million reads
READ_LENGTH	36	Each read sequence is 36nt long
PAIRED_END	NO	Single reads are simulated, one per fragment

Output

```
[INFO] I am collecting information on the run.
    initializing profiler  **********
[INFO] Checking GTF file
*[WARN] Unsorted in line 5 - cannot perform gene clustering: chrI + YAL069W @ 335 after YAL012W @ 130799
```

```
********* OK (00:00:02)
[GTF FILE] The GTF reference file given is not sorted, but we found a sorted version.
[GTF FILE] The Simulator will use /Users/micha/Desktop/sacCer3_SGDGenes_fromUCSC120515_sorted.gtf
[GTF FILE] You might want to update your parameters file
[PROFILING] I am assigning the expression profile
********** OK (00:00:02)
    Reading reference annotation *[WARN] merging exon (31229,35248) with exon (29935,31227) in transcript
YBL100W-B because intervening intron has 4 or less nt.
[WARN] merging exon (222636,226598) with exon (221330,222634) in transcript YBL005W-B because intervening
intron has 4 or less nt.
*********[WARN] merging exon (-854953,-856257) with exon (-850989,-854951) in transcript YPR158C-D because
intervening intron has 4 or less nt.
 OK (00:00:01)
    found 6664 transcripts
[PROFILING] Parameters
    NB_MOLECULES    5000000
    EXPRESSION_K    -0.6
    EXPRESSION_X0    5.0E7
    EXPRESSION_X1    9500.0
    PRO_FILE_NAME    /Users/micha/Desktop/sacCer3_enzyme.pro
    profiling ********** OK (00:00:00)
    Updating .pro file  ********** OK (00:00:00)
    molecules    4999971
[LIBRARY] creating the cDNA libary
    Initializing Fragmentation File ********** OK (00:00:04)
    4999971 mol initialized
[LIBRARY] Reverse Transcription
[LIBRARY] Configuration
        Mode: PDT
        PWM: No
        RT MIN: 500
        RT MAX: 2500
    Processing Fragments ********** OK (00:00:15)
        4999971 mol: in 4999971, new 0, out 4999971
        avg Len 969.7831, maxLen 2500
    preparing transcript sequences *[WARN] merging exon (31229,35248) with exon (29935,31227) in transcript
YBL100W-B because intervening intron has 4 or less nt.
*********[WARN] merging exon (-854953,-856257) with exon (-850989,-854951) in transcript YPR158C-D because
intervening intron has 4 or less nt.
 OK (00:00:02)
[LIBRARY] Enzymatic Digestion
[LIBRARY] Configuration
Left Flank : 100
Right Flank : 300
Motif: DNAseI.pwm
    Processing Fragments ********** OK (00:02:38)
        60604099 mol: in 4999971, new 55604128, out 60604099
        avg Len 80.00923, maxLen 2500
        initializing Selected Size distribution
[LIBRARY] Segregating cDNA (MCMC Filter)
    Processing Fragments ********** OK (00:01:47)
        60604099 mol: in 60604099, new 0, out 25719279
        avg Len 47.310493, maxLen 276
        start amplification
[INFO] Loading default PCR distribution
[LIBRARY] Amplification
[LIBRARY] Configuration
        Rounds: 15
        Mean: 0.5
        Standard Deviation: 0.1
    Processing Fragments ********** OK (00:01:05)
    Amplification done.
    In: 25719279 Out: 693695450
        25719279 mol: in 25719279, new 0, out 693695450
        avg Len 47.319595, maxLen 266
    Copied results to /Users/micha/Desktop/sacCer3_enzyme.lib
    Updating .pro file  ********** OK (00:00:00)
[SEQUENCING] getting the reads
    Initializing Fragment Index
    Indexing ********** OK (00:00:14)
    13804020 lines indexed (693695450 fragments, 6534 entries)
```

```
    sequencing *[WARN] merging exon (31229,35248) with exon (29935,31227) in transcript YBL100W-B because
intervening intron has 4 or less nt.
*********[WARN] merging exon (-854953,-856257) with exon (-850989,-854951) in transcript YPR158C-D because
intervening intron has 4 or less nt.
 OK (00:14:03)
    693695450 fragments found (13804020 without PCR duplicates)
    998612 reads sequenced
    226528 reads fall in poly-A tail
    407504 truncated reads
    Moving temporary BED file
    Updating .pro file  ********** OK (00:00:00)
    Updating .pro file  ********** OK (00:00:00)
    Updating .pro file  ********** OK (00:00:00)
    Updating .pro file  ********** OK (00:00:00)
[END] I finished, took me 1305 sec.
```