Appendix D - Java Utilities and output files description

Description of the Java tools and details of the output are detailed. Java utilities were authored by Sol Shenker.

Trimming tool : TrimUnmapped.jar trim3p

This utility trims the adapter sequence and polyA sequence from the reads that are unmapped in the first mapping step.

Execute as:

```
$ java -jar TrimUnmapped.jar trim3p -bam /path/to/file.bam\
-out output_file.fastq
```

Options:

-bam : the bam file from the first round of mapping, containing the unmapped reads.

-out : the name of the fastq file of trimmed unmapped reads to output

-adapter : polyA and adaptor sequences to be trimmed using regular expression as:
Default: A{5,}G(A(T(C(.*))?)?)?

If using a different adapter on the RT oligo than the one provided in the protocol, change the regular expression to identify the new adapter by using the -adapter option. Otherwise the default shown above will be used to trim the adapter reported in the protocol.

Internal priming mask tool: ThreeSeqPipeline.jar IdentifyInternalPriming

This utility identifies all nucleotide positions that are upstream of genomic stretches that contain a certain percentage of Adenosines (As). This can be used as a mask to filter out derived cleavage sites that could likely have been produced by binding of the RT primers to A-rich regions of transcribed RNA, leading to a determination of a false positive 3' end. The mask will be used in the clustering and site derivation step.

Execute as:

```
$ java -jar TrimUnmapped.jar IdentifyInternalPriming \
-fa genome.fasta -maxAdenosine 9 —primingMask \
primingMask.bed -window 16
```

Options:

-fa : fasta reference sequence for genome that reads are mapped to

-maxAdenosine : maximum number of adenines in a window before it is called
internal priming
Default: 9
-primingMask : file to which the BED file of potential internal priming regions are
written to
-window : window in which A's will be counted downstream of the cleavage site
Default: 16

The maxAdenosine and window parameter can be adjusted to filter ends that are upstream of sequence of different A-richness. This parameter can be derived empirically by assessing the impact on the identification of already annotated 3' end events.

Clustering, 3' end calling and quantification: TrimUnmapped.jar DefineClusters

This utility takes the trimmed reads and uses these to call a final cleavage site of events clustered in a certain window. This step pools together events that arise from slight differences in cleavage site (in this case events within a 25 bp window). The major site is derived by analyzing all libraries provided. The event in the window with the most reads will be the final cleavage site called by the utility. The untrimmed reads which 3' end falls in the 50 nt upstream of the identified cluster of trimmed reads will be counted towards the called 3' end to quantify the final event for each library. A minimum number of required reads to call a cleavage site can be input by the user, as well as parameters to filter the reads by mapping quality. Cleavage sites that coincides with positions in the internal priming mask will be filtered, unless they meet criteria outlined below in the description of the gtf_files found in the out folder (see end of this document).

Execute as:

```
java -Xmx4g -Djava.io.tmpdir=temp_dir \
-jar ThreeSeqPipeline.jar DefineClusters \
-minDistinctReads 3 -inDir mapped -trimmed trimmed \
-untrimmed untrimmed -primingMask /path/to/primingMask.bed\
-outDir gtf_files -baseNames sample_names.txt
```

In the example above, 3 unique reads are required to call an event.

-baseNames : a file with the names per line, without extension, of all samples used to derive the final atlas of 3' end annotations.
E.g., if sample is 'H.bam' it's base-name should be H.
Each sample should be reported on a separate line.

-clusterWidth : cleavage sites closer than this distance will be merged into a single cluster
Default: 25

-inDir : the directory in which the mapped BAM files live

-maxHits : the maximum number of multi-mapper locations before a read is discarded
Default: 5

-minDistinctReads : the minimum number of reads with distinct 5p positions before a site can be considered for defining cleavage position
Default: 3

-minMappingQuality : positions with all mapped reads with MAPQ below or equal to this value will not be used to define the position of cleavage sites
Default: 5

-outDir : the directory to which filtered BAM files will be output to

-primingMask : BED file of potential internal priming regions derived in previous steps

-trimmed : The subfolder of the [inDir] in which the trimmed reads live
Default: trimmed

-untrimmed : The subfolder of the [inDir] in which the trimmed reads live
Default: untrimmed

Output:
atlas.gtf : The utility will output an atlas of the 3' end positions quantified for each library in gtf format. Metadata with the quantification of the events (total, only trimmed reads, only untrimmed reads) for each library as well as other information is also reported.

Description of atlas.gtf columns:
chromosome
pipeline author
exon - the 3' end position is reported as an exon feature
start
end
score - default:0.0
strand

metadata:

the quantification for the event (sample_name.3p=counts for event if using only the trimmed reads for quantification, sample_name.un= counts for event if using only the untrimmed reads for quantification, sample_name= counts for event using both trimmed and untrimmed reads for quantification)

n = total number of events in all libraries using both trimmed and untrimmed reads for quantification

site_id = unique site ID

cluster = with the format *chr:start:end:strand* the window within events where used to quantify the final major site reported in the atlas in this gtf line.

The *gtf files* in the out file contain a breakdown of how each read was considered when building the atlas.gtf for record. Each folder contains a subset of reads, with the genomic position of the 3' end of the read in the gtf file followed by the number of reads with 3' ends at that position as metadata. These positions can be used to filter the bam files generated in the previous steps (for example to filter out the internal primed events in the bam file) and to follow how each read is processed. Outlined below are rules for rescuing reads that were initially flagged as internally primed events.

counted - the counts for the number of reads with trimmed terminal As. The region in the GTF is the position of the 3' end of the read.
counted_templated - the counts of reads without trimmed terminal As. The position in the GTF is also the 3' end of the read.
retained - counts of reads with trimmed terminal As that pass the internal priming filter. Position in the GTF is same as above
removed - counts of reads with trimmed terminal As that fail the internal priming filter. Position in the GTF is same as above
removed/rescued/ - if a read that fails the internal priming filter is less than 50nt upstream of a valid cleavage site, that read is rescued. This file contains counts of reads in this category. The position is the 3' end of the read.
removed/removed/ - if a read that fails the internal priming filter is more than 50nt upstream of a valid cleavage site, that read is put in this file.
consolidated - counts reads without trimmed As are assigned to the closest cleavage site downstream within 50nt. The counts represent the sum of reads with trimmed terminal As, untrimmed A's less than 50nt upstream, and rescued reads less than 50nt upstream. The position in the GTF is the position of the closest cleavage site as defined by reads with trimmed As.
unconsolidated - counts reads without trimmed As for which there is no cleavage site within 50nt downstream. The position in the GTF is 3' end of the untrimmed read.