Bioo Scientific®

MOLECULAR INDEXING™ ANALYSIS

The following describes the use of the script for Molecular Indexing™ Analysis based on stochastic labeling (STL). This software was provided by Weihong Xu from Stanford Genome Technology Center under GPL license to Bioo Scientific. The STL Method is included in Bioo Scientific's Nextflex™ qRNA-Seq™ Kit and Nextflex™ Rapid Directional qRNA-Seq™ Kit and is the essential prerequisite for accurate molecular quantification of mRNA transcription. Bioo Scientific is now providing the dqRNASeq script which allows the proper elimination of PCR duplicates from NGS data based on the STL method.

NGS libraries are routinely amplified by PCR in the last step. Some products are disproportionally amplified which causes the distortion of final results. These PCR duplicates are commonly removed based on the identification of their unique starts and stops (USS method). The USS method is built on the assumption that initial chemical fragmentation of RNA is random and therefore all RNA fragments generated in this way would differ in their start/stops sites. Consequently, only PCR duplicates will have identical start/stops and could be removed. However, detailed analyses revealed that DNA as well as RNA fragmentation is not random and that many original fragments have identical start and stop sites. Therefore, the application of the start/stop method alone is causing incorrect elimination of unique fragments from NGS data. To prevent this elimination, the STL method is used to provide additional identity to fragments. The principle of this protocol is the random ligation of 96 molecular labels at both ends of cDNA fragments prior to PCR amplification. In the case that two fragments will have the same start/stop but different labels, they will not be considered as PCR duplicates.

The percentage of fragments salvaged from incorrect elimination by the STL method depends on starting material, depth of sequencing and method of library preparation, but could be significant. Using read pairs aligned to transcripts and fastq files, this script will generate a table listing transcripts, their start/stops and STLs, as well as a table listing transcripts with the numbers of total read pairs and number of read pairs after STL/USS, USS, and STL correction.

1. System Requirement

any *nix system with BASH, AWK and SAMtools, including Unix, Linux, Mac OS and Cygwin. The script was tested on Mac Unix and Linux Ubuntu (version 14.04) systems.

dqRNASeq is Bash-based shell script, which takes advantage of the Unix C-program for computing.

2. Usage

dqRNASeq [options]
-b bamfile
-f R1_fastq
-r R2_fastq
-t STL_sequence_file [STL96.txt]
-s STL_size [8]
-q MAQ_Cutoff [30]
-m NREADS_Cutoff [1]
-h help

e.g. dqRNASeq -b BAMPREFIX.bam -f FASTQPREFIX_R1.fastq.gz -r FASTQPREFIX_R2.fastq.gz -t STL96.txt

e.g. dqRNASeq -b BAMPREFIX.bam -f FASTQPREFIX_R1.fastq -r FASTQPREFIX_R2.fastq

with BAMPREFIX and FASTQPREFIX to be the specific prefixes of input files.

3. Input

You need to provide at least three piece of information: a bam file with reads mapped to transcripts and a pair of fastq files. The bam file provides the link from read_id to transcript_id, and the fastq files link read_id to a pair of barcodes (from position 1 to STL_size, default 8). Additional parameters for filtering by Mapping Quality (MAQ_Cutoff) and Read Coverage (NREADS_Cutoff) are provided. The STL_sequence_file should contain one stochastic label per line, without anything else.

4. Output

There're 4 output files, with BAMPREFIX to be the prefix of the bam file.

BAMPREFIX.STLprefixPE8.MQ30_MINREAD1.txt --master table of all R1 and R2 8bp stochastic labels of sequences with mapping quality ≥30 (MQ30), requiring minimum 1 read pair mapped, and the 8bp STL sequences to be in the STL_sequence_file (the pre-defined stochastic labels), e.g.

transcript_id	start	stop	R1_STL	R2_STL	nReadPairs
NM_000051.3	8833	9139	CTTCGTTG	AACGCCAT	1
NM_000059.3	10931	11182	CCAAGGTT	AACGCCAT	1

BAMPREFIX.prefixPE8.MQ30_MINREAD1.txt --master table of all R1 and R2 8bp stochastic labels of sequences with mapping quality ≥30 (MQ30), requiring minimum 1 read pair mapped, and not requiring the 8bp STL sequences to be in the STL_sequence_file.

BAMPREFIX.uniqSTLprefixPE8.MQ30_MINREAD1.txt --expression matrix requiring 8 bp STL in the STL_sequence_file. nSTL is the number of unique stochastic labels, and nUSS is the number of unique start/stops, e.g.

transcript_id	nReadPairs	nSTL+nUSS	nUSS	nSTL
NM_000016.4	3	3	3	3
NM_000017.2	5	5	5	5

BAMPREFIX.uniqprefixPE8.MQ30_MINREAD1.txt --same as the previous file except not requiring the 8bp sequence to be in the STL_sequence_file.