TIPseqHunter

TIPseqHunter was developed by Java (version 7) and R (version 3.2) languages and tested under Linux operating system.

***Prerequisites:***

(1) At least 10GB memory is needed if the number of sequencing read-pairs is larger than 20M.

(2) Bowtie 2 alignment software (version 2.2.3 used for testing)

Website: <http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>

(3) Samtools software (latest version)

Website: <http://samtools.sourceforge.net/>

(4) Trimmomatic software (version 0.32 used for testing)

Website: <http://www.usadellab.org/cms/?page=trimmomatic>

(5) Java packages:

sam-1.112.jar, commons-math3-3.4.1.jar, jfreechart-1.0.14.jar, jcommon-1.0.17.jar, itextpdf-5.2.1.jar, biojava3-core-3.0.1.jar

(6) R packages:

pROC, ggplot2, caret, e1071

***Important note:***

BAM file has to be generated by bowtie2 alignment with "XM" tag

***Testing data and masked and bowtie-built reference genome are available from following website:***

<http://openslice.fenyolab.org/data/tipseqhunter/test_data>

***Usage of TIPseqHunter***

***(1) quality control, alignment, feature selection, modeling, prediction***

Usage: *./TIPseqHunterPipelineJar.sh fastq\_path output\_path fastq\_r1 key\_r1 key\_r2 num\_rp*

(Note: Detail information is also provided in TIPseqHunterPipelineJar.sh file. There are some parameters needed to be pre-set.)

Parameters:

fastq\_path: path of the fastq files (Note: this is the only path and file name is not included)

output\_folder: path of the output files (Note: this is the only path and file name is not included)

fastq\_r1: read 1 file name of paired fastq files

key\_r1: key word to recognize read-1 fastq file (such as "\_1" is the key word for CAGATC\_1.fastq file) (Note\*\*\*\*\*: key has to be unique in the file name)

key\_r2: key word to recognize read-2 fastq file and replaceable with the read-1 key word to match to read-1 file (such as "\_2" is the key word for CAGATC\_2.fastq fastq file) (Note\*\*\*\*\*: key has to be unique in the file name)

num\_rp: the total number of the read pairs in the paired fastq files (Note: it is the total number of read-pairs, i.e. either the total number of read1 or read2 but not together.) (This number is for normalization purpose)

***(2) somatic insertions***

Usage: *TIPseqHunterPipelineJarSomatic.sh repred\_path control\_path repred\_file control\_file*

(Note: Detail information is also provided in TIPseqHunterPipelineJarSomatic.sh file. There are some parameters needed to be pre-set.)

repred\_path: path of “model” folder under output folder

control\_path: path "TRLocator" folder under output folder

repred\_file: file with suffix ".repred" and generated from P11 in repred\_path (Note: file name should be ending with ".repred".) (such as 302\_T\_GTCCGC.wsize100.regwsize1.minreads1.clip1.clipflk5.mindis150.FP.uniqgs.bed.csinfo.lm.l1hs.pred.txt.repred)

control\_file: file with suffix “.bed” in control\_path (Note: file name should be ending with ".bed".) (such as 302\_N\_GTGAAA.fastq.cleaned.fastq.pcsort.bam.w100.minreg1.mintag1.bed)