Package: terminalMapper (v4.1)
Automated pipeline for the mapping, processing, filtering and analysis of CC-seq NGS data (Bash/Perl/R)

Dependencies: Perl (5.25.9+), Bash (v4.1+), R (v3.4+), Bowtie2 (v2.2.4+)
Input: Paired-End _R1 and _R2 FASTQs, Indexed Reference Genome, User Configuration
Output: 1bp Histograms, Molecule Sizes/Frequencies, Alignment Logs

terminalMapper constitutes a novel, low memory software package for the automated batch processing of CC-seq data (sae2Δ Spo11 DSBs, Topo II DSB/SSBs) or any library containing informative Read-1/2 5' ends. terminalMapper, run on the command line, requires three arguments:

```
Usage: termMapper.sh -i [INPUT FOLDER] -c [CONFIGURATION FILE] -o [OUTPUT FOLDER]
```

-i INPUT: Input data folder containing paired-end FASTQ files.
-c CONFIG: Configuration file specifying user-parameters (termMapper.config).
-o OUTPUT: Output data folder.

Configuration
terminalMapper is initially configured via an external .config file, specifying key variables:

(i) LIBRARY_TYPE (SINGLE/DOUBLE) — specification of library type, which in turn differentiates the type of coordinates called and the analysed performed (single cuts/double cuts)
(ii) SPACE_SAVER (Y/N) — when enabled, terminalMapper will reduce the disk footprint of the pipeline, progressively deleting non-essential files including .SAM and .FASTQ files
(iii) CORE — no. of CPU cores available/to utilise
(iv) READ1/2_EXT — FASTQ file extension for automated detection of paired end samples (e.g. _R1 and _R2)
(v) GLOBAL_OPTIONS — specification of Bowtie2 parameters for end-to-end alignment. Default settings: -X 1000 --no-discordant --very-sensitive --mp 5,1
(vi) GENOME_DIR, GENOME-NAME — directory and filename of a Bowtie2 indexed FASTA reference genome
(vii) FASTA_INDEX — directory and filename of the reference genome FASTA index
(viii) MAPQ_FILTER — terminalMapper will discard all reads below the specified MAPQ filter (e.g. 5)
(ix) REPEAT_LIST (Optional) — directory and filename of a file containing blacklisted regions to discard from analysis (see examples).

.SAM File Processing and Filtering
Tab-delimited .SAM files, the primary output of Bowtie2 alignment, specify (i) 1-based leftmost coordinates for mapped reads (ii) numerical "flags" denoting the aligned identify of each read pair—terminalMapper only handles paired, fully aligned reads (i.e. 99 - Read-1, Watson | 147 - Read-2, Crick // 83 - Read-1 Crick | 163 - Read-2, Watson) (iii) Alpha-numeric CIGAR codes describing base-by-base alignment and detailing the presence of INDELs—terminalMapper utilises such information for accurate coordinate calling. For example, 5M2I30M1D25M denotes:

- 5bp reference match (5M) ("M" may contain unspecified mismatches)
- 2bp insertion in the read (relative to the reference) (2I)
- 30bp reference match (30M)
- 1bp deletion in the read (relative to the reference) (1D)
- 25bp reference match (25M)

(iv) Alpha-numeric MD:Z tags denoting the position and base composition of SNPs/deletions present in the read relative to the reference. Insertions are not specified. terminalMapper utilises MD:Z tags to detect and discard reads with ambiguous ends. In LIBRARY_TYPE = SINGLE mode, >=2bp of mismatch at the Read-1 5'-end is defined as ambiguous and disqualifies the read pair from the main dataset. In LIBRARY_TYPE = DOUBLE mode, >=2bp of mismatch at either the Read-1 5' or Read-2 5' end disqualifies the read pair. Non-informative 3'-ends are not considered. For example, an MD:Z-tag of 0T0C25A5^T10 denotes:

- An initial 2bp mismatch (reference specifies TC, read contains alternative bases) (0T, 0C)

- 25bp of precise reference:read match (25) followed by a 1bp mismatch (25A)
- 5bp of precise reference:read match (5) followed by a 1bp deletion in the read (reference contains a T) (5^T)
- 10bp of precise reference:read match (10)

ATGAGCGTACCTGTAAATAAGAAGATCGATCGA_GGTACATACT — READ (0T0C25A5^T10)
TCGAGCGTACCTGTAAATAAGAAGATCAATCGATGGTACATACT — REF

For unambiguous 99-147 and 83-163 read pairs, coordinate positions of the informative ends are calculated. As SAM files specify 1-based leftmost coordinates—the 5' end is readily called by Bowtie2 for Watson (+) reads (99 or 163). In contrast, for Crick (-) reads (83 and 163), the leftmost base is the 3' end of the read. To call 5' Crick (-) coordinates, CIGAR codes are parsed and scored to determine the mapped read length—according to the following rules: (M = 1, D = 1, I = 0)—and the sum is added to the 3' coordinate. Insertions (I) (in the read) are ignored in order to call coordinates accurate to the utilised reference. As an example, a Crick (-) read with a CIGAR code of 75M and a leftmost coordinate is 10200 is called as 102074 (10200+75-1). A 1bp adjustment is made as the leftmost base is included as part of 75M. A more complex Crick (-) read with a CIGAR code of 35M2D10M3I30M and a leftmost coordinate of 10200 is called as 10276 (10200 + 35 + 2 + 10 + 30 -1).

Output Files

(i) Sparsely-formatted, 1bp histograms detailing the total number of Watson (+) and Crick (-) hits for all mapped base-pairs, on a per-chromosome basis. Unambiguous and ambiguous read-pairs are separated.
(ii) In LIBRARY_TYPE = DOUBLE mode, terminalMapper calculates molecule sizes as the absolute distance between the Read-1 5' and Read-2 5' ends, for each unambiguous read pair.
(iii) In LIBRARY_TYPE = DOUBLE mode, terminalMapper records the frequency with which any given pair of 5' coordinates is observed.
(iv) Log files detailing alignment quality and coordinate calling statistics.