MANTIS

MANTIS (Microsatellite Analysis for Normal-Tumor InStability) is a program developed for detecting microsatellite instability in paired tumor-normal samples. To perform analysis, the program needs a tumor BAM and a matched normal BAM file (produced using the same pipeline) to determine the instability score between the two samples within the pair. Longer reads (ideally, 100 bp or longer) are recommended, as shorter reads are unlikely to entirely cover the microsatellite loci, and will be discarded after failing the quality control filters.

Requirements

MANTIS is written in Python. Later versions of Python 2 (e.g. 2.7.1, 2.7.8) are compatible, but use of Python 3 is encouraged. The program utilizes the NumPy (http://www.numpy.org/) and Pysam (https://github.com/pysam-developers/pysam) libraries, which must be pre-installed to work in the environment. Additionally, a copy of the reference genome (e.g. hg19) in FASTA format must be available.

Download

The program is freely available under the GPLv3 license from GitHub at:
https://github.com/OSU-SRLab/MANTIS

Usage

The tool can be run with default parameters by executing:

```
python mantis.py --bedfile /path/to/loci.bed --genome /path/to/genome.fasta -n
/path/to/normal.bam -t /path/to/tumor.bam -o /path/to/output/file.txt
```

More detailed information about the parameters can be found in the sections below. Please note that the BED file has certain expectations, which are listed below.

Microsatellite Loci BED File Format

The tool requires microsatellite loci of interest to be input in a 6-column BED format. The fourth (name) column of the BED file must contain the targeted repeating k-mer (e.g. AC) and the reference repeat count for it, e.g. “(AC)12”. A sample entry in your BED file might look like:

```
chr15 33256217    33256249    (AC)16          0     +
```

With the format being:

Column	1	2	3	4	5	6
Description	Chromosome	Locus Start	Locus End	K-Mer Feature	Unused	Unused

The program will perform minor internal realignment to account for differences in BED file formats, e.g. whether a ‘chr’ prefix is used with chromosome, and whether the genomic start and end positions are 0- or 1-based.

Configuration File

To facilitate running many samples, or incorporation of MANTIS in a pipeline, a configuration file may be used. By default, the program will search for a configuration file with the filename “mantis_config.cfg” in the root folder of the program. Alternatively, a path to the configuration file can be provided using the `-cfg/--config` command line parameter.

The file should contain settings that match the named parameters of the program according to the following format:

```
genome = /path/to/reference/genome.fasta
bedfile = /path/to/my/loci.bed
```

Configuration Parameter Priority

Parameters specified on the command line have highest priority, followed by settings in the configuration file, and then by default values. For example, the minimum read quality setting has a value of 20.0 by default. If a different value (e.g. `-mrq = 30.0`) is specified in the configuration file, it will override the default. However, if the command line parameter (e.g. `-mrq 25.0`), is supplied when running the program, the value of 25.0 will be used, as command line parameters override both the configuration file and the default.

Multithreading Support

MANTIS provides support for using multiple threads/cores to perform the analysis. By default, the program will only use a single thread. By using the --threads parameter, you can specify the use of more threads. As much of the computation speed is bound by the rate at which reads can be retrieved from the BAM files, there will be diminishing returns beyond a certain number of threads as the device (hard disk) will have limited reading speed. The exact number of threads recommended will depend on your system configuration.

Parameters

The software will use default parameters if the user chooses not to customize the settings. These default settings have been selected during testing on various datasets. However, to customize the usage of the tool to better fit the user’s data, one could provide various command line parameters to the program, either directly on the command line and/or with a configuration file (see above). The available parameters are listed below:

Flag(s)	Name	Description
-cfg/--config	cfg	Path to the default configuration file being used. Optional. Note: If you have a mantis_config.cfg file in the MANTIS folder, it will be used by default without needing to be explicitly specified.
-n/--normal	normal	Path to the BAM file for the normal sample.
-t/--tumor	tumor	Path to the BAM file for the tumor sample.
--threads	threads	How many threads to use for multiprocessing. Optional. Default: 1
-b/--bedfile	bedfile	Path to the BED file containing the targeted MSI loci. Requires the format specified in the BED file section above.
--genome	genome	Path to the reference genome in FASTA format.
-o/--output	output	Path to the output file.
-mrq/--min-read-quality	mrq	Minimum average per-base read quality for a read to pass the quality control filters. Default: 25.0
-mlq/--min-locus-quality	mlq	Minimum average per-base quality for the bases contained within the microsatellite locus. Reads that pass the read quality filter (above) will still fail quality control if the locus quality scores are too low. Default: 30.0
-mrl/--min-read-length	mrl	Minimum read length for a read to pass quality control. Only bases that are not clipped will be considered; in other words, soft-clipped or hard-clipped parts of the read do not count towards the length. Default: 35
-mlc/--min-locus-coverage	mlc	Minimum coverage (after QC filters) required for each of the normal and tumor samples for a locus to be considered in the calculations. Default: 30
-mrr/--min-repeat-reads	mrr	Minimum reads supporting a specific repeat count. Repeat counts that have less than this value will be discarded as part of outlier filtering. Default: 3
-sd/--standard-deviations	sd	Standard deviations from the mean before a repeat count is considered an outlier and discarded. Default: 3.0

Whole-exome usage

Note that the above default quality thresholds are intended for use in situations such as targeted resequencing, in which locus coverage is less of an issue than with whole-exome data. Therefore, we recommend a less stringent set of thresholds for whole-exome data, as follows:

```
-mrq 20.0
-mlq 25.0
-mlc 20
-mrr 1
```

RepeatFinder

Included with MANTIS is a tool, RepeatFinder, for finding microsatellites within a reference genome. RepeatFinder is written in C++, and should compile with GCC using the included Makefile on almost any Linux system. RepeatFinder can be run with default parameters by executing:

```
./RepeatFinder –i /path/to/genome.fasta -o /path/to/loci.bed
```

The resulting BED file is suitable for use with MANTIS immediately (with the `-b/--bedfile` option), or it may be filtered with bedtools to intersect for regions of interest.

Several parameters are available to customize microsatellite finding with RepeatFinder, as follows:

Flag(s)	**Description**
-m	Minimum number of bases that a repeat region must span to call a microsatellite. Default: 10
-r	Minimum number of k-mer repeats to call a microsatellite. Default: 3
-l	Minimum k-mer length (bp). Default: 1
-L	Maximum k-mer length (bp). NOTE: Considering 6-mers is not recommended, as this will include telomere repeats. Default: 5
-i	Path to the reference genome in FASTA format.
-o	Output path to the desired BED file containing the targeted MSI loci.