MSMuTect, Version 0.5 manual

Installation methods:

PIP

```
pip3 install msmutect (not on pypi yet; this will not work yet)
```

Github

# cython must be preinstalled; if it is not, run: pip3 install Cython

```
git clone https://github.com/MaruvkaLab/MSMuTect_0.5
cd MSMuTect_0.5
sh rename.sh
pip3 install .
```

Defintions:

Histogram - a histogram is a structure containing indel information about a microsatellite locus. For instance, if the reference motif repeat length is 5, but some cells have a motif-length deletion, a histogram could be: 5_10, 4_6. This means there were 10 reads with a motif repeat length of 5, and 6 reads with a motif repeat length of 4

Alleles - Alleles are the alleles called by MSMuTect's rigorous statistical noise model for each microsatellite locus. The output file contains the called alleles, their projected frequencies, and the extent to which the call precisely models the data (log likelihood).

Mutations - Mutations are the mutations called by MSMuTect, by analyzing a normal and tumor sample from the same patient. It uses a rigorous noise model to weed out noisy loci, only analyzing loci that are good candidates for a mutation. The final step in the mutation calling process is a Fisher Exact Test for the histograms of the normal and tumor file for a locus. The p value of this test is also written to the output file (note that the majority of microsatellite loci will not have a Fisher Exact test performed since they do not pass the noise model).

Flags:

Run msmutect --help to see help message.

All flags

```
  -T TUMOR_FILE, --tumor_file TUMOR_FILE
                        Tumor BAM file
  -N NORMAL_FILE, --normal_file NORMAL_FILE
                        Non-tumor BAM file
  -S SINGLE_FILE, --single_file SINGLE_FILE
                        Analyze a single file for histogram and/or alleles
  -l LOCI_FILE, --loci_file LOCI_FILE
                        Loci to be processed and included in the output
  -O OUTPUT_PREFIX, --output_prefix OUTPUT_PREFIX
                        prefix for all output files
  -c CORES, --cores CORES
                        Number of cores to run MSMuTect on
  -b BATCH_START, --batch_start BATCH_START
                        1-indexed number locus to begin analyzing at (Inclusive)
  -e BATCH_END, --batch_end BATCH_END
                        1-indexed number locus to stop analyzing at (Inclusive)
  -H, --histogram       Output Histograms
  -A, --allele          Output alleles
  -m, --mutation        Output mutations
  -F FLANKING, --flanking FLANKING
                        Length of flanking on both sides of an accepted read
  -f, --force           overwrite pre-existing result files, if they exist
```

MSMuTect can run for pairs of files (ie. a regular and a tumor), or call histograms/alleles for a single file. The -S flag denotes a single file, and -T and -N flags denote Tumor and Normal files, respectively. The -H flag will generate a histogram, the -A flag will generate allele calls, and the -m flag will generate mutation calls (for pairs of files only). MSMuTect will always show preceding steps. In other words, the -m flag will output a file with alleles and histograms in addition to the mutations, and the alleles flag will output a file with histograms as well. In addition, by default, MSMuTect will do more work, not less. So, if you give MSMuTect a single file, it will call alleles and histograms for it, and only by passing the -H flag (and not the -A flag) will it only call histograms.

Some examples:

```
$ msmutect -I generic_loci.phobos -S myBam.bam -O myout [-A]
```

Will generate a file containing the alleles and histograms for myBam.bam called myout.all.tsv.
The -A argument is unnecessary, calling alleles is the default behavior.

```
$ msmutect -I generic_loci.phobos -S myBam.bam -H -O myout
```

Will generate a file holding only histograms called myout.hist.tsv. It’s first columns will be identical to myout.all.tsv

For pairs of files, by default MSMuTect will only analyze loci that are considered candidates for a mutation after analysis of the Normal file. So, for instance,

```
$ msmutect -I generic_loci.phobos -N normalBam.bam -T tumorBam.bam -O myout [-m]
```

Will output a file with mutations and some non mutations called myout.partial.mut.tsv

If you would like to see full data for a pair, for all loci, simply feed the -A or -H flags, along with the -m flag

```
$ msmutect -I generic_loci.phobos -N normalBam.bam -T tumorBam.bam -O myout -m -A
```

Will output a file myout.full.mut.tsv with all Alleles, Histograms, and Mutations for the Allele and tumor files

This will incur a significant performance penalty since all loci must be evaluated for both files

All files will record the mutation call for convenience

Please note, the -S and -T/-N flags cannot be used in tandem. -S is for single files. -T and -N is for running pairs

Tips:

Always pass the -c flag if you have more than one core. The runtime for a single core is very high

Count the mutations quickly and efficiently with cut -f49 [prefix].partial.mut.tsv | sort | uniq -c

Notes for Developers:

The Noise Table is written inline to ease distribution with pip. In any case, the hope is to replace this noise table with something more exact in the near future. The code is not completely pruned, however it is largely clear and concise. For now, the files are kept as .py to ease in debugging, which is the most important step in pre-1.0 versions. Many are changed by rename.sh before compilation. When the 1.0 version is created and uploaded to PYPI this will be changed

Support:

For support, please email k.avraham@technion.ac.il or yosi.maruvka@bfe.technion.ac.il

The code is hosted at Github at https://github.com/MaruvkaLab/MSMuTect_0.5 . If there is some feature you would like added to make MSMuTect work for you, please do not hesitate to email us and we will consider adding it. The code is licensed under an MIT license