How do I run MutSig?

To run MutSigCV locally, you can download it from this site, and follow the directions below.

MutSigCV is also available to run on our public GenePattern server. After a free registration, you can access it here:

http://genepattern.broadinstitute.org/gp/pages/index.jsf?lsid=MutSigCV (http://genepattern.broadinstitute.org/gp/pages/index.jsf?lsid=MutSigCV)

Input files

The following three data files are required:

- mutations.maf mutation table
- coverage.txt coverage table
- covariates.txt covariates table

MUTATION TABLE

This is a list of all the mutations to be analyzed, with all patients concatenated together. It should be a tab-delimited file with a header row. The positions of the columns are not important, but the header names are. The program will look up the information it needs based on the names of the headers.

The file can be a MAF file, and some of the columns come directly from the standard columns of the MAF format. However, there are a few additional columns required.

The four columns required by the program are:

- "gene" = name of the gene that the mutation was in. (can also be called "Hugo_Symbol")
- "patient" = name of the patient that the mutation was in. (can also be called "Tumor_Sample_Barcode")
- "effect" = what broad class of effect does the mutation exert on the gene? This should be one of "nonsilent" (it changes the protein sequence or splice-sites), "silent" (it is a synonymous change), or "noncoding" (it is intronic or otherwise in a flanking noncoding region.). (Note, in version1.0.0, this information was split into two columns, "is_coding" and "is_silent", and this method is still supported.)
- "categ" = mutation category. MutSigCV splits mutations into different categories depending on their DNA context (e.g. was the mutation in a CpG dinucleotide? in another C:G basepair? in an A:T basepair?). For each category, there is a different set of "bases at risk" for that category of mutation. There is also a special category outside this system called "null/indel", into which all truncating mutations and indels are placed. By convention, the "bases at risk" for a null/indel mutation is taken to be the entire territory of the gene.

The standard set of categories we have used for many sequencing projects is as follows:

- 1. CpG transitions
- 2. CpG transversions
- 3. C:G transitions
- 4. C:G transversions
- 5. A:T transitions
- 6. A:T transversions
- 7. null+indel mutations

To compute the "categ" column, information from the reference genome is required: specifically, the identity of the nucleotides directly on each side of the mutation site. Together with that information and the Variant_Classification, Reference_Allele, and Tumor_Seq_Allele1+2 .columns, "categ" can be computed

Starting with MutSigCV version 1.3, an integrated preprocessing module assists with the calculation of these categ numbers, and also .enables the automated determination of the optimal set of categories to be used for a given dataset

COVERAGE TABLE

The coverage table tells how many nucleotides were sequenced to adequate depth for mutation calling. Coverage is tabulated for each patient, and for each gene. It is also broken down by "categ" and "effect" (as listed above). Again, "effect" can be either "noncoding" (this refers to the flanking territory outside exons), "nonsilent" (this refers to bases which, when mutated, yield a change in the protein sequence--including splice-sites), or "silent" (this refers to bases which give a synonymous change when mutated). Note, some coding positions can contribute fractionally to the "silent" and "nonsilent" zones, in a ratio of 1/3 to 2/3 (or vice versa), depending on the consequences of mutating to each of the three possible alternate bases.

The columns required in the coverage table are:

- "gene": the gene name, corresponding to the "gene" column in the mutation_table.

- "effect": whether this row tabulates the "silent", "nonsilent", or "noncoding" territory for this gene. (Note, in version 1.0.0 column was called "zone", and this name is still allowed.)
- "categ": which category this row tabulates-- should be same as in mutation_table

- <patient_name_1>: number of sequenced bases for patient#1 in this gene and effect/categ bin
- <patient_name_2>: number of sequenced bases for patient#2 in this gene and effect/categ bin
- <patient_name_3>
- (etc.)

We recognize that detailed coverage information is not always available. In such cases, a reasonable approach is to carry out the computation assuming full coverage. We have prepared a file that can be used for this purpose: it is a "full coverage" file, or more accurately a "territory" file: the only information it contributes is a tabulation of how the reference sequence of the human exome breaks down by gene, categ, and effect. To download this file, see the section below about MutSigCV v1.3.

COVARIATES TABLE

This table lists genomic parameters for each gene being analyzed. They are called covariates because they co-vary with mutation rate. They will be used to calculate distances between pairs of genes in a "covariate space" in order to determine the nearest neighbors of each gene, in order to pool information among them about the local background mutation rate (BMR).

The columns of this file are:

- "gene": the gene name, should match those used in the first two tables.
- <covariate_name_1>: the value of the first covariate for each gene
- <covariate_name_2>: the value of the second covariate for each gene
- <covariate_name_3>: the value of the third covariate for each gene
- etc.

The covariates table provided in the Example Data has proven useful for analyzing many cancer types. The table contains one value per gene for: (1) global expression, derived from RNA-Seq data and summed across the 91 cell lines in the CCLE (Barretina et al.). (2) DNA replication time (from Chen et al.). (3) the HiC statistic, a measure of open vs. closed chromatin state (from Lieberman-Aiden et al.).

Running the algorithm

If you have a license for Matlab, you can run MutSigCV from its source code file: MutSigCV.m

Open Matlab and type the following command at the ">>" Matlab prompt.

```
MutSigCV('mutations.maf','coverage.txt','covariates.txt','output.txt')
```

If you do not have a license for Matlab, you can run the compiled version of MutSigCV using the free Matlab MCR:

```
run_MutSigCV.sh <path_to_MCR> mutations.maf coverage.txt covariates.txt output.txt
```

The algorithm will load the three input files, process them using the MutSigCV algorithm, and then finally write the output table to the file 'output.txt'.

Be sure to replace mutations.maf, etc with the actual paths to the input files.

Running MutSigCV when all you have is a MAF file

==========

Starting with v1.3 of the code, MutSigCV has a preprocessing module that takes care of organizing the "categ" and "effect" information. This makes it easy to run MutSigCV when all you have is a MAF file.

To run MutSigCV in this way, please first download the following four reference files:

- genome reference sequence: chr_files_hg18.zip (http://www.broadinstitute.org/cancer/cga/sites/default/files/data/tools/mutsig/reference_files/chr_files_hg18.zip) or chr_files_hg19.zip (http://www.broadinstitute.org/cancer/cga/sites/default/files/data/tools/mutsig/reference_files/chr_files_hg19.zip)
  - unzip this file to yield a directory (chr_files_hg18/ or chr_files_hg19/) of chr*.txt files
- mutation_type_dictionary_file.txt (http://www.broadinstitute.org/cancer/cga/sites/default/files/data/tools/mutsig/reference_files/mutation_type_dictionary_file.txt)
- exome_full192.coverage.txt.zip (http://www.broadinstitute.org/cancer/cga/sites/default/files/data/tools/mutsig/reference_files/exome_full192.coverage.zip)
  - unzip this file to yield exome_full192.coverage.txt
- gene.covariates.txt (http://www.broadinstitute.org/cancer/cga/sites/default/files/data/tools/mutsig/reference_files/gene.covariates.txt)

Then run the program with the following six arguments (instead of four):

- the name of your MAF file
- exome_full192.coverage.txt (after unzipping)
- gene.covariates.txt
- output filename stem <outputstem>, which will be suffixed for each output file
- mutation_type_dictionary_file.txt
- chr_files_hg18 or chr_files_hg19 (after unzipping)

From the Matlab prompt, that is:

MutSigCV ('my_mutations.maf','exome_full192.coverage.txt','gene.covariates.txt','my_results','mutation_type_dictionary_file.txt','chr_files_hg19')

or using the compiled version of MutSigCV and the free Matlab MCR:

run_MutSigCV.sh <path_to_MCR> my_mutations.maf exome_full192.coverage.txt gene.covariates.txt my_results mutation_type_dictionary_file.txt chr_files_hg19

OTHER BUILDS

Please note, currently the preprocessing module supports data from the human exome, in builds hg18 or hg19. Future work will enable use of other builds (e.g. mm9, canFam2, etc.)

Example data

The data used for the TCGA Lung Squamous paper (http://www.nature.com/nature/journal/v489/n7417/full/nature11404.html) is available here:

LUSC.MutSigCV.input.data.v1.0.zip (/cancer/cga/sites/default/files/data/tools/mutsig/LUSC.MutSigCV.input.data.v1.0.zip)

MutSigCV_example_data.1.0.1.zip (/cancer/cga/sites/default/files/data/tools/mutsig/MutSigCV_example_data.1.0.1.zip) (same data as v1.0, but renames files and includes the expected output file)

Unzip the archive to yield the following four files:

- LUSC.mutations.maf mutation table from the TCGA lung squamous publication
- LUSC.coverage.txt coverage table based on real exome-capture data, from the TCGA lung squamous publication
- gene.covariates.txt covariates table (expression, replication time, HiC compartment)
- LUSC.example.output.txt output generated by running the example data

The input data can be processed by invoking the following command from the Matlab ">>" prompt:

```
MutSigCV('LUSC.mutations.maf','LUSC.coverage.txt','gene.covariates.txt','LUSC.output.txt')
```

NOTE ABOUT VERSIONS

The list of significantly mutated genes found by MutSigCV1.0 is the same as that published in the LUSC manuscript, with the exception of one additional gene, FBXW7, which was not initially reported as significantly mutated, but now is. Also please note that the per-gene p-values differ due to the slightly different variants of the algorithm used. In particular, the dynamic range of MutSigCV1.0 now ends at ~10^{-16}.