**SMuRF – Significantly Mutated Region Finder**

SMuRF identifies significantly mutated genomic regions in a set of samples.

Developed by Paul Guilhamon

**Requirements**

SMuRF is implemented in Bash and R; it runs on any platform with Bash (≥4.1.2), R (≥3.3.0) and BEDTools (≥2.23.0). It requires the following R packages: GenomicRanges, gtools, gplots, ggplot2, data.table, psych, and dplyr.

**Input**

* Variants file(s): These should contain variants of interest. SMuRF has options for applying SNP filtering but will not perform arbitrary quality filters. They should be in either VCF format as multiple files (1 per sample) under the same directory OR in BED format, with all mutations for all samples within the same file, and exactly 4 columns: chr/start/stop/name.

NOTE: if using the BED format, keep one row per mutation per sample, ie: if a mutation is recurrent and found in multiple samples these should be placed on separate rows. Also note that the start and stop positions are NOT the same; the start column corresponds to the POS column of a VCF-formatted file minus 1; (BED and VCF follow different numbering conventions; see <https://www.biostars.org/p/84686/> for more information).

See format example below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Example input variants in BED format** | | | |
| chr10 | 101776718 | 101776719 | Sample1 |
| chr10 | 101877881 | 101877882 | Sample2 |
| ... | ... | ... | ... |

* Genomic Regions: a BED file of genomic regions of interest. The format is chr/start/stop/name ; Note: the 'name' column is required but can be any string; it is recommended to use it for the names of biological samples from which each genomic region was identified for easy trace-back in downstream analyses, although this is not required for SMuRF. These genomic regions and the variants specified above do not need to come from matched samples. Unlike the set of input variants above, it is important that the genomic regions be mutually exclusive. If in your cohort regions from two different samples overlap, they should be merged as one and the sample name column can have all the contributing sample names in a comma-separated list.

See format example below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Example input genomic regions in BED format** | | | |
| chr1 | 10027 | 10612 | Sample1 |
| chr1 | 237701 | 237901 | Sample3, Sample5, Sample22 |
| ... | ... | ... | ... |

* Annotation files: Gencode promoter annotation, and dbSNP sets for both hg19 and hg38 are provided within the SMuRF directory.
  + Promoter annotation: The Gencode set of TSS coordinates for each genome build (hg19: v19, hg38: v24) was downloaded from the UCSC table browser. Promoters were annotated from -2.5kb to +0.5kb around the TSS.
  + SNPs: dbSNP sets v147 were downloaded from the UCSC table browser for each genome build. The “Common” version of the database is made available with the tool; this can be used to remove only those that have a minor allele frequency of >= 1%, of use if one wants to avoid removing potentially interesting somatic mutations that also appear as very rare polymorphisms in the population. The larger file containing all SNPs (including those found at frequency < 1% ) can be downloaded from the dbSNP website.

**Output**

SMuRF outputs a number of files, some for display and evaluation purposes, others as end results, and a final set for input into downstream tools, such as C3D or GO annotation.

* Results Files
  + filtered vcf files: the vcf files will be filtered for SNPs as specified by the user
  + FiltVarsInPeaks.txt: a concatenated list of unique variants, overlapped with the input genomic regions, and including the names of samples they were identified in
  + summary.txt: summary of the number of input regions, number mutated, number mutated significantly, and their distribution in promoter and enhancer regions
  + annotated\_regions.txt: all input genomic regions annotated to either a promoter region (using Gencode as described above) OR to "DistalRE" (=distal regulatory element) if they fall outside promoter regions
  + named\_counts.txt: a count of unique variants in each sample
  + Mutated\_Regions.txt and other files with that prefix: all information on the mutated regions and stats on significance of mutation. Other files include suffixes indicating the subsequent filtering steps
    - Freq3: only includes regions with >=3 mutations from >= 3 different samples
    - qval0.05: only includes regions significantly mutated
* Results Plots
  + QQplot\_Freq3.pdf: self-explanatory;
  + barplot\_mut\_inRegions\_Counts\_and\_Perc.pdf: summary plots of the number of variants per sample and how many overlap with the provided genomic intervals
  + Mutation\_Rate\_Plot\_Freq3\_qval0.05.pdf: scatter plot of significantly mutated regions:-log10(q-value) vs region mutation rate
  + Mutation\_Rate\_Plot\_Freq3\_qval0.05\_Promoters\_VS\_DistalREs.pdf: same as above with added promoter/DistalRE annotation;
  + Sample\_Frequency\_Plot\_Freq3\_qval0.05.pdf and Sample\_Frequency\_Plot\_Freq3\_qval0.05\_Promoters\_VS\_DistalREs.pdf: Same as the two above, but with the number of unique samples mutated in the region as the x-axis as opposed tot he mutation rate
  + Sample\_Frequency\_Plot\_Freq3\_qval0.05\_DistalREs.pdf and Sample\_Frequency\_Plot\_Freq3\_qval0.05\_Promoters.pdf: same as Sample\_Frequency\_Plot\_Freq3\_qval0.05\_Promoters\_VS\_DistalREs.pdf but separating out those regions annotated to promoters and DistalREs
* Files for use in dowstream analyses
  + Mutated\_Regions\_sigmut0.05\_DistalRE\_forc3d.bed: significantly mutated regions annotated to enhancers, to use as input for C3D to identify downstream target promoters
  + Mutated\_Regions\_sigmut0.05\_DistalReAndProm.bed: all significantly mutated regions; for use in GO applications for example, or TF motif enrichment
  + Mutated\_Regions\_allmut\_DistalREAndProm.bed: all mutated regions

**Arguments**

All arguments are required and no defaults are set:

* **o**: full path to output directory
* **f**: format chosen for input variants; this must be one of *vcf* or *bed*.
  + *vcf*: individual VCF formatted files for each sample must be placed under a single directory. The full path to that directory is what should be supplied to the **-v** argument described below. **NOTE: when using this option the sample names used for output files will be automatically generated from the names of the VCF files themselves: the ".vcf" at the end of the filename will be clipped off and the rest will be used as sample name.**
  + *bed*: a single BED file containing all variants to be analysed. The format must be the following: chr/start/stop/samplename/. The full path to that bed file is what should be supplied to the **-v** argument described below.
* **v**: the input variants themselves; depending on the format chosen above, either the path to a directory containing individual VCF files OR the path to a single BED file
* **s**: SNP filter to apply to the input variants. Must be either *n* (no SNP filter), or the full path to a SNP file of your choice that will be filtered out of your input variants file. Files for common and all snps from dbSNP147 for the hg19 and hg38 human genome builds are available in the SMuRF directory for convenience.
* **p**: promoter annotation file; those derived from Gencodev19 (human genome build hg19) and Gencodev24 (human genome build hg38) are provided in the SMuRF directory. These can be replaced with any annotation of your choice, but the output figures by SMuRF will automatically colour-code regions by either the gene name linked to a promoter region or the mention “DistalRE” for distal regulatory element, as this tool was originally designed to look for mutation-enriched non-coding regions.
* **r**: genomic regions of interest that will be checked for significant enrichment of mutations; must be supplied in the format chr/start/stop/name; the 4th column is required but can be any string. Having the name there could make downstream analyses easier but isn't required by SMuRF
* **m**: method used to calculate the background mutation rate in the binomial test; Must be either *allsamples* or *regionsamples*
  + *allsamples*: bmr is total number of mutations (Across all samples) in regions divided by total bp in regions
  + *regionsamples*: bmr is calculated based on which samples have mutations in that region; the individual sample bmr is calculated as for global, and the average of the BMRs for the samples represented in the region is taken as the region BMR

**Usage**

sh smurf.sh \

-o ... \

-f ... \

-v ... \

-s ... \

-p … \

-r ... \

-m ...