User Manual

(January 7, 2019)

DriverML integrates the Rao’s score test and supervised machine learning to identify cancer driver genes. Rigorous and unbiased benchmark analysis and comparisons of DriverML with 20 other existing tools in 31 independent datasets from The Cancer Genome Atlas (TCGA) show that DriverML is robust and powerful among various datasets and outperforms the other tools with a better balance of precision and recall.

# Access

DriverML is a free software tool.

# Installation

tar –xzf DriverML\_v1.0.0.tgz

cd DriverML\_v1.0.0

chmod +x \*.pl \*.r \*.sh

# Running

Requirements: Perl and R are required on user’s Linux environment.

Usage: path/run.driverml.sh -w <AbsolutePath/DriverML\_v1.0.0> –i <AbsolutePath/mutation.maf> -f <AbsolutePath/reference file> -r <AbsolutePath/reference file> [options]

Required arguments:

|  |  |
| --- | --- |
| -w/--path | The absolute path to the DriverML\_v1.0.0. |
| -i/--input | The list of tumor mutations to be analyzed. |
| -f/--reference\_genome | The reference genome for input data. |
| -r/--reference\_training | The human reference genome file for training data. |

Options:

|  |  |
| --- | --- |
| -g/--mutation\_table | The predefined gene mutation table which could be found in the package of this application. Default: GRCh38 |
| -y/--tumor\_type | Training mutation data.Default: Pan-cancer |
| -m/--multicore | Set the number of parallel instances to be run concurrently. Default: 4 |
| -t/--simulation\_time | Set the number for Monte Carlo Simulation. Default: 10,000. |
| -o/--output | Set the prefix of the output file. Default: summary. |
| -c/--cluster\_number | Set the upper limit of cluster number for computing BMR. |
| -p/--prior | Set the prior information. Default: Non-TCGA genes from DriverDB and IntOGen databases. |
| -n/--interpolation\_number | Set the upper limit of interpolation number for making gene clusters. Default: 100. |
| -d/--indel\_ratio | The ratio of point mutation to indel in the background. Default: 0.05. |
| -h/--help | Help information. |
| -v/--version | Software version. |

# Notes

**The format of input mutation file**

The mutation file needs to be MAF format. Eight columns named Hugo\_Symbol, Chromosome, Start\_Position, Variant\_Classification, Variant\_Type, Reference\_Allele, Tumor\_Seq\_Allele2 and Tumor\_Sample\_Barcode are required. The first row in the mutation file must be the header including column names above (case sensitive).

Eight required columns are as below:

* Tumor sample barcode is the sample ID in which this mutation occurs.
* Hugo symbol requires HUGO symbols of genes.
* Chromosome that contains the gene.
* Start position requires the lowest numeric position of the reported variant on the genomic reference sequence (1-based coordinate system).
* Reference allele is the plus strand reference allele at this position. Include the sequence deleted for a deletion, or "-" for an insertion.
* Tumor allele2 is tumor sequencing (discovery) allele 2.
* Variant classification describes the translational effect of a variant allele. It is one of Silent, Missense\_Mutation, Nonsense\_Mutation, Splice\_Site, Frame\_Shift\_Del, Frame\_Shift\_Ins, In\_Frame\_Del, and In\_Frame\_Ins. Other mutation types will not be analyzed.
* Variant type is Type of mutation, such as SNP, DNP, TNP, INS, and DEL.

The mutation file needs to be Detailed information about the MAF format could be found at <https://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format+%28MAF%29+Specification>. An example file could be seen in /yourpath/DriverML\_v1.0.0/example\_data/UVM.txt.

# Output

The output of DriverML is a summary of putative driver genes, including the numbers of each mutation type, the value of the statistic, p-value, and FDR adjusted p-value.

# Example

nohup /AbsolutePath/DriverML\_v1.0.0/run.driverml.sh -w /AbsolutePath/DriverML\_v1.0.0 -i example/UVM.txt -f /AbsolutePath/GRCh38.fa -r /AbsolutePath/hg19.fa -c 1 -o UVM-summary.txt > UVM-nohup.out