Table of Contents

1. Introduction

Oncodrive-fm is an approach to uncover driver genes or gene modules. It computes a metric of functional impact using three well-known methods (SIFT, PolyPhen2 and MutationAssessor) and assesses how the functional impact of variants found in a gene across several tumor samples deviates from a null distribution. It is thus based on the assumption that any bias towards the accumulation of variants with high functional impact is an indication of positive selection and can thus be used to detect candidate driver genes or gene modules.

Oncodrive-fm starts by computing three metrics of the functional impact of each non-synonymous SNVs (nsSNVs) found in genes across a list of tumor samples. Any measure of the impact of nsSNVs on protein function (or FI score) could in principle be used here. We have chosen three well-known methods whose scores may be obtained in a high-throughput manner to evaluate hundreds of nsSNVs in a few minutes. Stop-gain SNVs (stSNVs) and frameshift-causing indels (fsindels) are incorporated to the bias analysis by assigning them scores that are comparable to the highest-ranking tier of nsSNVs. Finally, synonymous SNVs (sSNVs) are taken into account with scores equal to those of bottom ranking nsSNVs (see Supplementary Method for details).

The second step starts by averaging the FI scores of variants per gene and comparing them to the distribution of scores of variants in functionally similar genes. If somatic SNVs were obtained using a whole-genome or whole-exome sequencing approach, the null distribution contains all SNVs and fsindels detected across tumor samples. We call this the internal null distribution. On the other hand, if only a limited number of genes have been sequenced, the null distribution of each gene is composed of nsSNVs that occur naturally in human populations, or external mull distribution. Three FM bias are thus obtained, one for the average of the scores provided by each original method. These three FM biase are then combined using Fisher's method and a combined p-value is subsequently obtained. We regard this FM bias as a measure of the bias toward the accumulation of functional impact in a gene across all tumor samples. The higher the FI of SNVs and fsindels found in a gene, the higher its FM bias will be.

We have applied the Oncodrive-fm approach to three datasets of genes with SNVs and fsindels in samples of different tumor types: glioblastoma multiforme (gbm, ref.), and serous ovarian carcinoma (soc) produced within The Cancer Gene Atlas (TCGA) project and chronic lymphocytic leukemia (cll, ref.), produced within the International Cancer Genomes Consortium (ICGC) initiative. We were able to detect most genes also pinpointed by MutSig (a method that searches recurrently mutated genes) as significantly biased in gbm and soc. Moreover, we were able to detect recurrent genes with low functional impact which may not constitute true drivers and we uncovered other top-ranking functionally affected genes, some of which could be lowly recurrent drivers.

The OncodriveFM approach is described in detail in the paper **Functional impact bias reveals cancer drivers** by Gonzalez-Perez et al., *NAR*., doi:10.1093/nar/gks743.

2. System requirements

Oncodrive-fm is programmed in PERL, so you will need a **PERL interpreter** in your computer to run it. The scripts in our implementation assume that the PERL interpreter is installed in /usr/bin. If in your local configuration the PERL interpreter is installed in another directory, please edit all the scripts accordingly. It will require also that you install the PERL **package Statistics::Descriptive**, which may

be readily obtained from the cpan repository. Also, Oncodrive-fm uses the statistical analysis framework **R**, in two of the scripts. The scripts assume that R may be invoked directly. If this is not the case, please modify the scripts accordingly or create a symbolic link in your system that allows you to do so.

3. Installation

Just unpack the oncodrive-fm tar.gz package within the directory of your choice. This will create a subdirectory structure, where you will find all scripts and data files needed to directly execute oncodrive-fm. Also, two examples of mutations files (GBM and CLL) are included in the package to help you test the method.

4. Running on exemplary datasets

To execute the Oncodrive-fm analysis on the examples provided (gbm and cll), just execute the following command from the bin subdirectory generated after unpacking the tar.gz:

>./pipeline_laucher.pl ../config/[gbm|cll].config

Results will be written to subdirectories oncodrive/data/glioblastoma, oncodrive/data/CLL and oncodrive/data/glioblastoma/final_results, oncodrive/data/CLL/final_results. The pipeline is very verbose, and at the end it will indicate the names and locations of all files produced. The main files of the genes analysis are:

- oncodrive/data/[glioblastoma|CLL]/final_results/[glioblastoma|CLL].mutsummary

This file contains the list of all genes with at least two variants in the samples analyzed. The columns summarize the number of samples where the gene has been found with variants, the FM Zscore and the significance of this Zscore (using the internal or external null distributions).

The pvalue|qvalue column of this file may be used a one-column matrix to be visualized in Gitools.

- oncodrive/data/[glioblastoma|CLL]/[glioblastoma|CLL].fimp.pathways

This file contains the results of the pathways analysis. Its structure is similar to that of the previous file.

- oncodrive/data/[glioblastoma|CLL]/[glioblastoma|CLL].mutmatrix.[sift|pph2|ma]

This file contains the functional impact scores of all variants found across all tumor samples in every sequenced gene, assessed using three well-known methods.

5. Preparing your own data to run Oncodrive-fm

If you have sequenced a few hundred genes (or complete exomes) across tens (or hundreds) of tumor samples you may run Oncodrive-fm on it to detect highly FM biased genes.

While the original version of the oncodrivefm scripts was prepared to run on SIFT, PolyPhen and MutationAssessor scores (as covered in the previous section), the 1.1.0 version is prepared to run on any assortment of any number of functional impact scores (FIS). You'll just need to prepare an input file as follows:

Column 1: Ensembl id of the mutated gene

Column 2: FIS 1 of the variant

Column 3: FIS 2 of the variant

...

Column i: FIS i of the variant

...

Column n: ID of the sample where the variant was found

The exemplary files submitted with Oncodrive-fm contain, in addition another column that specifies the consequence type of the variant. This column is not mandatory: it's only used to produce a summary of the types of variants identified in the tumor samples.

Then, you'll need to produce the config file. The easiest way to go is to copy one of the exemplary *config* files prepared for the GBM and CLL datasets and edit the relevant input variables.

This is the general structure of the *config* file:

```
##########################################################

# Input data specific for the tumor under analysis

#tumor: This name will be used as prefix to name all intermediate and final pipeline files

tumor='Here introduce a name for your project'

#mutfile: File that contains the mutations data of the tumor you want to analyze. Each
```

row corresponds to the mutation of one gene in one sample. Its format should be:

#

####Ensembl_Gene_ID FIS1 FIS2 ... FISi... Sample_ID

mutfile='Here introduce the path to the file containing the list of mutations of your project'

####numFIS: number of functional scores included in the mutations file and used to compute the functional impact bias

numFIS='Here introduce the number of FIS you have used to score the mutations'

##

##

Common input data (change these only if you have downloaded different info files)

#genes2gos: File that contains the genes2gos mapping

genes2gos='../data/common/slimgos_distrib/genes2gos'

#gosdistribs: Directory with the files that contain the distributions of SIFT, PPH2 and MA scores for each slimGOA obtained from 1000genomes.

gosdistribs='../data/common/slimgos_distrib/'

#genes2symbols: File that contains the genes2symbols mapping obtained from BioMart. Its format should be:

#

####Ensembl_Gene_ID Gene_Symbol

genes2symbols='../data/common/genes2symbols.txt'

#outdir: Directory to write output files

outdir='Here introduce the path to the results directory; safest ../data/project_name'

#tmpdir: Directory to write intermediate files

tmpdir='../tmp'

#internal: whether the null distribution will be taken from variants observed in the tumor

internal='' Set this to 1 to use the internal null distribution or 0 to use the external null distribution.

##

Note: Use the external null distribution only if you have sequenced few genes or very few (less than 30) samples.