User manual for the signatureFit command line interface

signature.tools.lib version: 2.1.1
latest edit: 28/02/2022

Andrea Degasperi, University of Cambridge, UK
ad923@cam.ac.uk

1. Introduction

Mutational signature fit analysis attempts to identify the presence of a given set of mutational signatures in the somatic mutations of a cancer sample.

This document describes how to use the signatureFit command line script, which is a wrapper for the signatureFit_pipeline function in the signature.tools.lib R packages, which in turn is an interface for the Fit and FitMS mutational signature fit analysis functions.

The signatureFit_pipeline function is a flexible interface for mutational signature fit analysis. Users can provide mutation calls as input or a pre-built mutational catalogue, and then they can perform either an automated analysis, with very few options required such as the organ of origin of the sample, or use the options to perform a more tailored analysis.

2. Installation

The script signatureFit is included in the signature.tools.lib R package. Thus, in order to use it, one is required to install signature.tools.lib, which is available on GitHub:

https://github.com/Nik-Zainal-Group/signature.tools.lib

After the installation of signature.tools.lib, one can run the signatureFit script, which is located in the scripts folder in the github repository. For easy access, add a copy of the signatureFit script to a location in your command line PATH.

3. signatureFit options

The list of available options can be accessed by typing:

```
signatureFit --help
```

This is the current output:

```
This script runs the signature fit pipeline of the R package
signatures.tools.lib, using the Fit or FitMS functions.

Run this script as follows:

signatureFit [OPTIONS]

Available options:
  -o, --outdir=DIR          Name of the output directory. If omitted a name
                              will be given automatically.
  -b, --bootstrap           Request signature fit with bootstrap
  -c, --cataloguesfile=CFILE
```

```
                            CFILE is the name of a file containing mutational
                              catalogues. Each sample catalogue is in a column,
                              with sample names as column headers and channel
                              names as row names in the first column with no
                              header. Can be omitted if mutations are provided.
-x, --snvvcf=SNVVCF         SNVVCF is a tab separated file containing two
                              columns. The first column contains the sample
                              names, while the second column contains the
                              corresponding SNV vcf file names.
-X, --snvtab=SNVTAB         SNVTAB is a tab separated file containing two
                              columns. The first column contains the sample
                              names, while the second column contains the
                              corresponding SNV tab file names. Each SNV tab
                              file should have a header with the following
                              columns: chr, position, REF, ALT.
-y, --dnvvcf=DNVVCF         DNVVCF is a tab separated file containing two
                              columns. The first column contains the sample
                              names, while the second column contains the
                              corresponding DNV vcf file names. VCF files can
                              contain both DNVs and SNVs and if two SNVs are
                              adjacent they will be merged into additional DNVs.
-Y, --dnvtab=DNVTAB         DNVTAB is a tab separated file containing two
                              columns. The first column contains the sample
                              names, while the second column contains the
                              corresponding DNV tab file names. Each DNV tab
                              file should have a header with the following
                              columns: chr, position, REF, ALT. Tab files can
                              contain both DNVs and SNVs and if two SNVs are
                              adjacent they will be merged into additional DNVs.
-z, --svbedpe=SVBEDPE       SVBEDPE is a tab separated file containing two
                              columns. The first column contains the sample
                              names, while the second column contains the
                              corresponding BEDPE file names. Each BEDPE file
                              is a tab separated file with header: chrom1,
                              start1, end1, chrom2, start2, end2, and sample. In
                              addition, another column with header svclass should
                              indicate the type of SV: translocation, inversion,
                              deletion, or tandem-duplication.
-w, --signaturesfile=SFILE
                            SFILE is the name of a file containing mutational
                              signatures. Each signature is in a column, with
                              signature names as column hearders and channel
                              names as row names in the first column with no
                              header. Each column must sum to 1. Use only to
                              provide your own signatures. When fitmethod=FitMS,
                              these signatures are considered common signatures.
-W, --raresignaturesfile=RSFILE
                            RSFILE is the name of a file containing mutational
                              signatures. Each signature is in a column, with
                              signature names as column hearders and channel
                              names as row names in the first column with no
                              header. Each column must sum to 1. Use only to
                              provide your own signatures. When fitmethod=FitMS,
                              these signatures are considered rare signatures.
-s, --sigversion=SIGVERSION
                            Either COSMICv2, COSMICv3.2, RefSigv1 or RefSigv2.
                              If not specified SIGVERSION=RefSigv2.
-O, --organ=ORGAN           When using RefSigv1 or RefSigv2 as SIGVERSION,
                              organ-specific signatures will be used.
                              If SIGVERSION is COSMICv2 or COSMICv3.2, then a
                              selection of signatures found in the given organ
                              will be used. Organ names depend on the selected
                              SIGVERSION. For RefSigv1 or RefSigv2: Biliary,
                              Bladder, Bone_SoftTissue, Breast, Cervix (v1 only),
                              CNS, Colorectal, Esophagus, Head_neck, Kidney,
                              Liver, Lung, Lymphoid, NET (v2 only),
```

```
                              Oral_Oropharyngeal (v2 only), Ovary, Pancreas,
                              Prostate, Skin, Stomach, Uterus.
-l, --signames=SIGNAMES   If no ORGAN is specified, SIGNAMES can be used to
                              provide a comma separated list of signature names
                              to select from the COSMIC or reference signatures,
                              depending on the SIGVERSION requested. For example,
                              for COSMICv3.2 use: SBS1,SBS2,SBS3.
-e, --genomev=GENOMEV     Genome version to be used: hg19, hg38 or mm10.
                              If not specified GENOMEV=hg19.
-m, --fitmethod=FITMETHOD Either Fit or FitMS. If not specified FITMETHOD=FitMS
-M, --optmethod=OPTMETHOD Optimisation objective function, either KLD or NNLS.
                              If not specified OPTMETHOD=KLD.
-t, --filtertype=FTYPE    FTYPE is either fixedThreshold or giniScaledThreshold.
                              When using fixedThreshold, exposures will be removed
                              based on a fixed percentage with respect to the
                              total number of mutations (THRPERC will be
                              used). When using giniScaledThreshold each signature
                              will used a different threshold calculated as
                              (1-Gini(signature))*GINISCALING. If not specified
                              then FTYPE=fixedThreshold
-p, --thresholdperc=THRPERC
                            THRPERC is a threshold in percentage of total mutations
                              in a sample, only exposures larger than THRPERC are
                              considered. If not specified THRPERC=5.
-d, --giniscaling=GINISCALING
                            GINISCALING is a scaling factor for the threshold type
                              giniScaledThreshold, which is based on the Gini score
                              of a signature. If not specified GINISCALING=10.
-u, --thresholdpval=THRPVAL
                            THRPVAL is a p-value to determine whether an exposure
                              is above the THRPERC or the threshold calculated with
                              Gini scaling, when using bootstrap. In other words,
                              this is the empirical probability that the exposure
                              is lower than the threshold. If not specified then
                              THRPVAL=0.05.
-a, --fitmsmode=FMSMODE   FMSMODE is either constrainedFit, partialNMF,
                              errorReduction, or cossimIncrease. If not specified
                              FMSMODE=errorReduction.
-T, --raresigtier=RSTIER  RSTIER is either T1 or T2. For each organ we provide
                              two lists of rare signatures that can be used.
                              Tier 1 (T1) are rare signatures that were observed
                              in the requested organ. The problem with T1 is that
                              it may be that a signature is not observed simply
                              because there were not enough samples for a certain
                              organ in the particular dataset that was used to
                              extract the signatures. So in general we advise to
                              use Tier 2 (T2) signatures, which extend the rare
                              signature to a wider number of rare signatures.
                              If not specified RSTIER=T2.
-i, --residualnegprop=RNP RNP is the maximum proportion of mutations (w.r.t.
                              total mutations in a sample) that can be in the
                              negative part of a residual when using the
                              constrained least squares fit when fitMS mode is
                              FMSMODE=constrainedFit. If not specified then
                              RNP=0.003.
-R, --minresidualmuts=MINRM
                            MINRM is the minimum number of mutations in a residual
                              when FMSMODE=constrainedFit or FMSMODE=partialNMF.
                              Deactivated by default (MINRM=NULL).
-C, --mincossimraresigs=MINCSRS
                            MINCSRS is the minimum cosine similarity between a
                              residual and a rare signature for considering the
                              rare signature as a candidate in a sample when
                              FMSMODE=constrainedFit or FMSMODE=partialNMF.
                              If not specified, MINCSRS=0.8.
-E, --minerrorredperc=MINERPERC
```

```
                            MINERPERC is the minimum percentage of error reduction
                              for a rare signature to be considered as candidate
                              in a sample when FMSMODE=errorReduction. The error
                              is computed as mean absolute deviation. If not
                              specified MINERPERC=15.
  -I, --mincossimincr=MINCSINCR
                            MINCSINCR is the minimum cosine similarity increase
                              for a rare signature to be considered as candidate
                              in a sample when FMSMODE=cossimIncrease. If not
                              specified MINCSINCR=0.02.
  -k, --maxraresigs=MAXRS   MAXRS is the maximum number of rare signatures that.
                              are allowed to be present in each sample. If not
                              specified MAXRS=1.
  -n, --nparallel=NPARALLEL Number of parallel CPUs to be used.
  -f, --nboot=NBOOT         Number of bootstrap to be used when bootstrap is
                              requested (-b), if not specified, NBOOT=200.
  -r, --randomSeed=SEED     Specify a random seed to obtain always the same
                              identical results.
  -v, --verbose             Verbose option for additional output.
  -h, --help                Show this explanation.
```

4. Using the signatureFit command line interface

4.1 Mutational catalogues

A mutational catalogue is a vector that contains counts of specific classes (channels) of mutations, which depend on the type of mutations considered. For example, SNV mutations will use the 96-channel catalogue type, which contain substitution counts in their trinucleotide context (e.g., A[C>T]A).

Using the signatureFit interface, it is possible to provide a mutational catalogue directly, or, alternatively, provide a list of mutation files that will be automatically converted into catalogues. Only one list of mutation files of one single type of mutations can be provided, e.g., a list of vcf files containing SNVs can be provided using the --snvvcf option. If a catalogue is provided directly (option --cataloguesfile), the mutation type will be inferred by the row names, which are the catalogue channels.

It is also possible to provide both a catalogues file and a list of mutation files. Provided that the type of mutations is the same, the new catalogues built from the mutation files will be added to the catalogues provided.

4.2 Mutational signatures

The signatureFit script can use different strategies to estimate how many mutations in a catalogue are associated with a certain mutational signature. To do so, it requires an *a priori* set of mutational signatures to use.

With signatureFit, it is possible to provide files containing the mutational signatures to use, or use available options to select a set of signatures automatically. If mutational signatures are provided manually, this will have the priority, and the options used to select signatures automatically, such as --organ, --sigversion and --signames, will be ignored.

Depending on the fit method, one or two sets of signatures is required. If --fitmethod=Fit, then just one set of signatures will be used, which can be specified manually with --signaturesfile, while

if --fitmethod=FitMS, then two sets of signatures are required, one set of common signatures (--signaturesfile) and one of rare signatures (--raresignaturesfile).

If no signatures have been provided manually, then signatureFit will check if the --organ option has been used, and select the appropriate signatures based on the --sigversion and --fitmethod options, as well as the inferred mutation type.

If no signatures have been provided manually and also no organ was specified, then signatureFit will select signatures based on the --sigversion and --signames parameters and the mutation type. For example, one could use --sigversion=COSMICv3.2 and --signames=SBS1,SBS2,SBS3 to fit only three COSMIC v3.2 signatures into a catalogue, provided it is an SNV catalogue.

When using --sigversion=RefSigv1 or --sigversion=RefSigv2, signatureFit will check the --organ option to select organ-specific mutational signatures. If the organ is not specified, then signatureFit will use the reference signatures instead, which are obtained as average of multiple similar organ-specific signatures.

5. Examples

5.1 FitMS example

In this example we run a signature fit analysis with FitMS, using vcf single nucleotide variant (SNV) files as input, assuming that the vcf are obtained from breast cancer samples.

```
signatureFit ––organ Breast –b –o outfolder –x snvvcf.tsv
```

Note that FitMS is the default fit method, so there is no need to specify --fitmethod=FitMS. FitMS will use the latest RefSigv2 signatures, which include the common and rare SBS signatures identified in the analysis of the Genomics England WGS cancer dataset. The flag -b requests a bootstrap analysis, -o indicates the output folder, and -x indicates the location of a tab separated file containing a list of sample names and corresponding vcf locations. The content of snvvcf.tsv could be as follows:

```
Sample1   sample1_snv.vcf
Sample2   sample2_snv.vcf
Sample3   sample3_snv.vcf
...
```

Finally, note that all the mutations in the input vcf files will be used, so they should already be filtered, e.g. containing only PASS variants.

5.2 Fitting COSMIC v3.2 signatures

In this example we show how to fit a specific set of SBS COSMIC signatures from the COSMIC v3.2 set.

```
signatureFit –b –o outfolder –x snvvcf.tsv –s COSMICv3.2 –m Fit –l
SBS1,SBS2,SBS3,SBS5,SBS8,SBS13,SBS17a,SBS17b,SBS18
```

In this case, we did not specify an organ, but rather used the -s option to request the use of COSMIC v3.2 signatures and the -l option to supply a list of signatures to use. We also specified to use the Fit algorithm with the option -m.

5.3 Fitting organ-specific rearrangement signatures

In this example, we fit organ-specific ovarian cancer rearrangement signatures, from the Degasperi et al 2020 Nature Cancer paper.

```
signatureFit ––organ Ovary –b –o outfolder –s RefSigv1 –m Fit –z svbedpe.tsv
```

In this case, we provided a list of bedpe files using the option -z. The content of svbedpe could be as follows:

```
Sample1   sample1_sv.bedpe
Sample2   sample2_sv.bedpe
Sample3   sample3_sv.bedpe
...
```

Note that we specified the option -s RefSigv1, because the latest organ-specific rearrangement signatures belong to this signature version, although, this could have been omitted, implying the default option -s RefSigv2, and signatureFit would have switched to -s RefSigv1 automatically with a warning message, so that the latest organ-specific rearrangement signatures available could be selected.