User manual for the hrDetect command line interface

signature.tools.lib version: 2.4.0
latest edit: 06/04/2023

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

1. Introduction

HRDetect is a classifier for homologous recombination deficiency in tumour samples that uses mutational signatures as input features.

This document describes how to use the `hrDetect` command line script, which is a wrapper for the `HRDetect_pipeline` function in the signature.tools.lib R package.

The `HRDetect_pipeline` function is a flexible interface for the HRDetect classifier. HRDetect scores can be computed directly from input features, or the input features can be computed by the function providing the somatic mutations of the tumours, including single nucleotide variants, indels, copy number variants and structural variants.

2. Installation

The script hrDetect 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 hrDetect script, which is located in the scripts folder in the github repository. For easy access, add a copy of the hrDetect script to a location in your command line PATH.

3. hrDetect options

The list of available options can be accessed by typing:

```
hrDetect --help
```

This is the current output:

```
This script runs the HRDetect pipeline of the signature.tools.lib R package.

Run this script as follows:

hrDetect [OPTIONS]

Available options:
  -i, --input=INPUTTABLE    Tab separate input table with the list of files for
                              each sample. Columns of INPUTTABLE should be:
                              sample, SNV_vcf_files, SNV_tab_files,
                              Indels_vcf_files, Indels_tab_files, CNV_tab_files,
                              SV_bedpe_files. Note that only one column of
                              SNV_vcf_files and SNV_tab_files is necessary
  -o, --outdir=OUTDIR       Name of the output directory. If omitted a name will
                              be given automatically.
  -O, --organ=ORGAN         When using RefSigv1 or RefSigv2 as SNVSV or SVSV,
                              organ-specific signatures will be used.
                              If SNVSV is COSMICv2 or COSMICv3.2, then a
                              selection of signatures found in the given organ
```

```
                              will be used. Available organs depend on the selected
                              SNVSV and SVSV. 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.
-s, --snvsigversion=SNVSV Either COSMICv2, COSMICv3.2, RefSigv1 or RefSigv2.
                              When SNVSV=RefSigv2 and an organ is specified,
                              signature fit for SNVs will be performed with FitMS
-S, --svsigversion=SVSV   Currently only RefSigv1 is available for SV signatures
-l, --snvsignames=SNVSN   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.
-L, --svsignames=SVSN     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.
-q, --snvcstier=SNVCSTIER
                          SNVCSTIER is either T1, T2 or T3. For each organ, T1 indicates
                              to use the common organ-specific substitution signatures, while T2
                              indicates to use the corresponding reference signatures.
                              In general, T1 should be more appropriate for organs
                              where there are no mixed organ-specific signatures, e.g.
                              SBS1+18 or SBS2+13, while T2 might be more suitable
                              for when such mixed signatures are present, so that
                              each signature can be fitted, e.g. fitting the two
                              signatures SBS1 and SBS18, instead of a single SBS1+18.
                              T3 is a mix of T1 and T2, where only the mixed organ signatures
                              are replaced with the correspondiing reference signatures.
                              If not specified SNVCSTIER=T1.
-Q, --svcstier=SVCSTIER
                          SVCSTIER is either T1, T2 or T3. For each organ, T1 indicates
                              to use the common organ-specific rearrangement signatures, while T2
                              indicates to use the corresponding reference signatures.
                              T3 is a mix of T1 and T2, where only the mixed organ signatures
                              are replaced with the correspondiing reference signatures.
                              If not specified SVCSTIER=T1.
-b, --bootstrap           Request HRDetect with bootstrap
-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. Set THRPERC
                              to -1 to deactivate.
-P, --thresholdnmuts=THRNMUTS
                          THRPERC is a threshold in number of mutations
                              in a sample, only exposures larger than THRNMUTS are
                              considered. If not specified THRNMUTS=10. Set THRNMUTS
                              to -1 to deactivate.
-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. The
                              threshold is computed as (1-Gini(signature))*GINISCALING,
                              and will be used as a percentage of mutations in a sample
                              that the exposure of "signature" need to be larger than.
                              Set GINISCALING to -1 to deactivate.
-D, --giniscalingnmuts=GINISCALINGNMUTS
                          GINISCALINGNMUTS is a scaling factor for the threshold type
                              giniScaledThreshold, which is based on the Gini score
                              of a signature. If not specified GINISCALINGNMUTS=50. The
                              threshold is computed as (1-Gini(signature))*GINISCALINGNMUTS,
                              and will be used as number of mutations in a sample
                              that the exposure of "signature" need to be larger than.
                              Set GINISCALINGNMUTS to -1 to deactivate.
-x, --snvfitfile=SNVFF    SNVFF is the file name of an rData file containing a Fit
```

```
                            or FitMS result object. This parameter should be used
                            when the user wants to customise the subs fit outside
                            the HRDetect pipeline, e.g. using the signatureFit
                            command line script. If custom signatures were used,
                            values CSNV3 and CSNV8 can be used to specify which
                            custom signatures correspond to the HRDetect parameters
                            SNV3 and SNV8.
  -y, --snv3altname=CSNV3   Custom signature name that will be considered as SNV3
                            input for HRDetect. Useful for when snvfitfile is
                            provided and custom signatures are used.
  -z, --snv8altname=CSNV8   Custom signature name that will be considered as SNV8
                            input for HRDetect. Useful for when snvfitfile is
                            provided and custom signatures are used.
  -X, --svfitfile=SVFF      SVFF is the file name of an rData file containing a Fit
                            or FitMS result object. This parameter should be used
                            when the user wants to customise the rearr fit outside
                            the HRDetect pipeline, e.g. using the signatureFit
                            command line script. If custom signatures were used,
                            values CSV3 and CSV5 can be used to specify which
                            custom signatures correspond to the HRDetect parameters
                            SV3 and SV5.
  -Y, --sv3altname=CSV3     Custom signature name that will be considered as SV3
                            input for HRDetect. Useful for when svfitfile is
                            provided and custom signatures are used.
  -Z, --sv5altname=CSV5     Custom signature name that will be considered as SV5
                            input for HRDetect. Useful for when svfitfile is
                            provided and custom signatures are used.
  -w, --snvrstier=SNVRSTIER
                           SNVRSTIER is either T0, T1, T2, T3 or T4. For each organ,
                            T0 are rare substitution signatures that were observed in the
                            requested organ, including low quality signatures
                            (QC amber and red signatures).
                            T1 are high quality (QC green) rare signatures that
                            were observed in the requested organ. T2-T4 signatures
                            extend the rare signatures set to what has been observed
                            also in other organs. T2 includes all QC green signatures
                            that were classified as rare at least twice (SBS only)
                            in Degasperi et al. 2022 Science. T3 includes all QC green
                            signatures (if not SBS, T3=T2). T4 includes all signatures
                            including QC amber and red.
                            In general, we advise to use T2 signatures
                            If not specified SNVRSTIER=T2.
  -a, --snvmaxrs=SNVMAXRS   Maximum number of rare signatures allowed in a sample when
                            using FitMS to fit SNV signatures (which is the default if.
                            ORGAN is given). If not specified SNVMAXRS=1. Sometimes it
                            is useful to increase this to check whether additional rare
                            signatures might be present, e.g. SNVMAXRS=2.
  -e, --genomev=GENOMEV     Genome version to be used: hg19, hg38 or mm10. If not
                            specified GENOMEV=hg19.
  -n, --nparallel=NPARALLEL Number of parallel CPUs to be used
  -f, --nbootFit=NBOOTFIT   Number of bootstrap to be used in signature fit. If
                            not specified NBOOTFIT=100.
  -g, --genomeplot          Request to plot genomeplots for each sample, which
                            show all the mutations on a circle plot.
  -r, --randomSeed=SEED     Specify a random seed to obtain always the same
                            identical results.
  -h, --help                Show this explanation.
```

4. Examples

4.1 Using organ-specific mutational signatures and bootstrap HRDetect

In this example, we compute the HRDetect score for two breast cancer samples using mutation files, and request HRDetect with boostrap.

```
hrDetect -O Breast -b -o outfolder -i inputTable.tsv
```

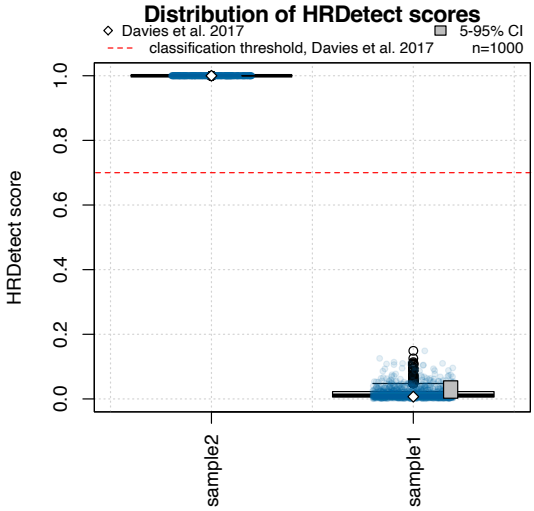
Using the -O option to specify an organ, the HRDetect pipeline will use organ-specific mutational signatures for estimating SNV and SV signature exposures.

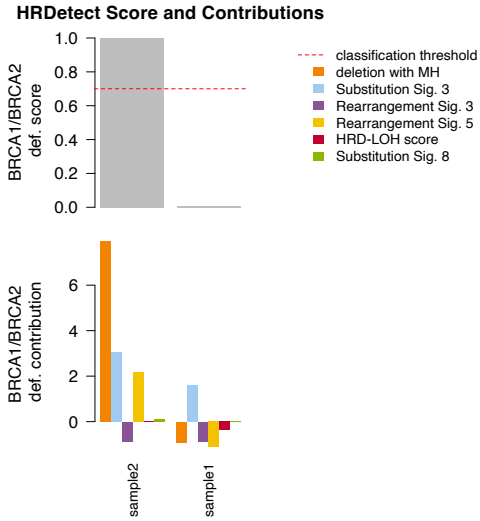
Note that FitMS is the default fit method for SNV organ-specific signatures, while the SV organ-specific signatures are fitted with the simpler Fit function. The flag -b requests a bootstrap HRDetect score, -o indicates the output folder, and -i indicates the location of a tab separated file containing a list of sample names and corresponding mutation files locations. The content of inputTable.tsv could be as follows:

```
sample    SNV_vcf_files     Indels_vcf_files  CNV_tab_files     SV_bedpe_files
Sample1   s1_snv.vcf        s1_id.vcf         s1_cnv.tsv        s1_sv.bedpe
Sample2   s2_snv.vcf        s2_id.vcf         s2_cnv.tsv        s2_sv.bedpe
Sample3   s3_snv.vcf        s3_id.vcf         s3_cnv.tsv        s3_sv.bedpe
...
```

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.

Below is an example of the pipeline output:





4.2 Using custom signature fit files

In this example, we assume that the user has performed a custom signature fit analysis using the command line script signatureFit, which automatically saves the fit results into a fitData.rData file, or alternatively using the Fit or FitMS functions and then saving the results using the saveFitToFile function. Let assume that the saved signature fit files are called fitSNV.rData and fitSV.rData.

The HRDetect pipeline will try to extract values for SNV3, SNV8, SV3, SV5, using the following signature names: SNV3 = "SBS3", "Signature3", "RefSig3"; SNV8 = "SBS8", "Signature8", "RefSig8"; SV3 = "RS3","RefSigR3"; SV5 = "RS5", "RefSigR5", "RefSigR9". If custom signature names have been used, then they can be provided using the flags --snv3altname, --snv8altname, --sv3altname, and --sv5altname.

The updated command line could then be as follows:

```
hrDetect -b -o outfolder -i inputTable.tsv -x fitSNV.rData -y "customSNV3name"
-z "customSNV8name" -X fitSV.rData -Y "customSV3name" -Z "customSV5name"
```

Given that the SNV and SV fit files are provided, then the input table should contain only the CNV and Indel files. The content of inputTable.tsv could be as follows:

```
sample    Indels_vcf_files  CNV_tab_files
Sample1   s1_id.vcf         s1_cnv.tsv
Sample2   s2_id.vcf         s2_cnv.tsv
Sample3   s3_id.vcf         s3_cnv.tsv
...
```