

Readme for examples

Shixiang Sun
08/31/2021

1	Input files	2
1.1.	Somatic mutations	2
1.1.1.	VCF file	2
1.1.2.	TSV file.....	2
1.1.3.	Count matrix file	3
1.2.	Group file	3
1.3.	Signature file	3
1.4.	Activate file.....	4
1.5.	Seeds file	4
2	General output files	4
3	Extract function.....	5
3.1.	MutationalPatterns tool	5
3.1.1.	NMF method	5
3.1.2.	Bayesian NMF method	5
3.2.	SomaticSignatures tool	5
3.2.1.	NMF method	5
3.2.2.	PCA method.....	6
3.3.	hdp tool	6
3.3.1.	<i>de novo</i> analysis	6
3.3.2.	analysis with prior signatures.....	6
3.4.	SignatureToolsLib tool	6
3.4.1.	Regular analysis	6
3.4.2.	Fixed signatures during analysis	7
3.5.	SigProfiler tool.....	7
4	Similarity function	7
5	Fit function.....	7
5.1.	MutationalPatterns tool	7
5.1.1.	NMF method	7
5.1.2.	strictNMF method	7
5.1.3.	bootstrappedNMF method	8
5.2.	mmsig tool	8
5.2.1.	bootstrapped analysis	8
5.2.2.	regular analysis	8
5.3.	SignatureToolsLib.....	8
5.4.	SigProfiler	8
5.4.1.	regular analysis	8
5.4.2.	fit somatic mutations and <i>de novo</i> signatures with user defined signatures	9

1 Input files

1.1. Somatic mutations

The input file for somatic mutations could be VCF files compressed in “.tar.gz” or “.zip” format, aggregated somatic mutations in “.tsv” format, and mutation count matrix in “.txt” format.

1.1.1. VCF file

Typical VCF file contains meta-information lines, a header line, and data lines each containing information about a position in the genome. In SomaMutDB, the meta-information lines and header line are not required for upload. Here is the example from part of “BF2_S3.vcf” in “test_small.hg19.tar.gz”.

```
##fileformat=VCFv4.3
##source=SCcallerV2.0.0
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=BI,Number=1,Type=Float,Description="Amplification Bias">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="sequencing noise, amplification artifact, heterozygous SNV and homozygous SNV respectively">
##FORMAT=<ID=SO,Number=1,Type=String,Description="Whether it is a somatic mutation.">
##contig=<ID=1,length=249250621,assembly="hg19">
##contig=<ID=2,length=243199373,assembly="hg19">
##contig=<ID=3,length=198022430,assembly="hg19">
##contig=<ID=4,length=191154276,assembly="hg19">
##contig=<ID=5,length=180915260,assembly="hg19">
##contig=<ID=6,length=171115067,assembly="hg19">
##contig=<ID=7,length=159138663,assembly="hg19">
##contig=<ID=8,length=146364022,assembly="hg19">
##contig=<ID=9,length=141213431,assembly="hg19">
##contig=<ID=10,length=135534747,assembly="hg19">
##contig=<ID=11,length=135006516,assembly="hg19">
##contig=<ID=12,length=133851895,assembly="hg19">
##contig=<ID=13,length=115169878,assembly="hg19">
##contig=<ID=14,length=107349540,assembly="hg19">
##contig=<ID=15,length=102531392,assembly="hg19">
##contig=<ID=16,length=90354753,assembly="hg19">
##contig=<ID=17,length=81195210,assembly="hg19">
##contig=<ID=18,length=78077248,assembly="hg19">
##contig=<ID=19,length=59128983,assembly="hg19">
##contig=<ID=20,length=63025520,assembly="hg19">
##contig=<ID=21,length=48129895,assembly="hg19">
##contig=<ID=22,length=51304566,assembly="hg19">
##contig=<ID=X,length=155270560,assembly="hg19">
##contig=<ID=Y,length=59373566,assembly="hg19">
##contig=<ID=MT,length=16569,assembly="hg19">
##reference=file:///gs/gsf0/users/shsun/my-script/SingleCell_Index/reference/b37/human_g1k_v37_decoy.fasta
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT IL11
1 5349909 . A G 18.0 . NS=1 GT:SO:AD:BI:GQ:PL 0/1:True:16,9:0.68:18:132,90,73,439
1 5498791 . C T 36.0 . NS=1 GT:SO:AD:BI:GQ:PL 0/1:True:27,16:0.596:36:246,160,124,956
1 5545619 . T C 16.0 . NS=1 GT:SO:AD:BI:GQ:PL 0/1:True:36,16:0.568:16:235,165,149,1090
```

Each VCF file should only include one sample and all VCF files need be compressed using “tar” or “zip” command as follows:

```
tar -zcf test_small.hg19.tar.gz *.vcf
zip -q test_small.hg19.zip *.vcf
```

1.1.2. TSV file

TSV (Tab-separated values) file records aggregated somatic mutations from all the samples. In TSV file, we only need five columns, including chromosome, position, reference allele, alt allele and sample ID. Here is the example from part of “test_small.hg19.tsv”.

1	962842	G	A	PD28690bx_2_a5
1	2092875	G	A	PD28690bx_2_a5
1	2857063	C	T	PD28690bx_2_a5
1	2883354	C	T	PD28690bx_2_a5
1	2901606	G	T	PD28690bx_2_a5

1.1.3. Count matrix file

Somatic mutations could be summarized into count matrix, of which first line records sample IDs and first column includes all mutation types. Each number represents the number of mutations found in the specific mutation type of certain sample. The first column of mutation type is required by SomaMutDB. Here is the example from part of “test_small.hg19.txt”.

Mutation Types	BF2_S3	IL11	PD28690bx_2_a5	PD28690bx_2_d5	PD34199a_c17	PD34199a_c27	PD34209ac	PD34209al
A[C>A]A	29	14	53	58	58	54	35	26
A[C>A]C	32	1	24	36	29	29	19	21
A[C>A]G	5	3	7	2	8	10	6	6
A[C>A]T	37	6	27	28	20	31	10	19
C[C>A]A	43	9	38	42	29	36	25	30
C[C>A]C	26	5	21	35	21	14	12	10

The top 96 lines with counts are SNV mutation counts, and the rest 83 lines are INDEL mutation counts. User could only provide SNV or INDEL counts for analysis.

1.2. Group file

There are only two columns required in group file: sample ID and group ID. Group ID could be tissue type, cell type and even serial numbers. Here is the example from “group_small.txt”.

PD28690bx_2_a5	Colon
PD28690bx_2_d5	Colon
PD34199a_c17	Colon
PD34199a_c27	Colon
PD36713b_lo013	Liver
PD36713b_lo023	Liver
PD36718b_lo001	Liver
PD36718b_lo002	Liver
PD34209ac	Lung
PD34209al	Lung
PD37451br	Lung
IL11	Fibroblast
BF2_S3	B

1.3. Signature file

Signature files are usually formatted as data matrix and saved as plain text. In signature files, first line records signature ID and first column is the mutation type, both of which are required by SomaMutDB. Here is the example from part of “prior_INDEL_hdp.txt”.

TYPE	ID1	ID2	ID3	ID4	ID5
C_deletion_1	0.000159889	4.82E-03	0.124727109	0.007249717	0.022202108
C_deletion_2	0.000773523	0.0000221	0.208876169	0.002734869	0.028547215
C_deletion_3	3.31E-18	3.11E-06	0.176324217	0.002041063	0.026596927
C_deletion_4	0.001907613	2.47E-03	0.064042764	0.001112283	0.014159122
C_deletion_5	0.00070599	0.003856976	0.043989808	0.001075684	0.002873422
C_deletion_6+	0.003370115	1.46E-02	0.022417486	0.002246178	0.002602449
T_deletion_1	1.31E-03	0.007905966	0.03759923	0.006635052	0.101481149
T_deletion_2	3.31E-18	3.95E-18	0.013309824	0.001622466	0.106099611
T_deletion_3	0.000936453	7.04E-03	0.016501851	0.002633405	0.114128464
T_deletion_4	0.002295855	0.020234743	0.020141477	0.000306038	0.088962208
T_deletion_5	0.004291422	4.71E-02	0.015814364	0.001023202	0.052259532
T_deletion_6+	3.31E-18	8.15E-01	0.008066314	0.000154442	0.030610523
C_insertion_0	3.31E-18	3.95E-18	0.002083092	6.62E-18	0.000542252
C_insertion_1	3.31E-18	0.000227026	0.001578887	0.000272399	0.001242898

The differences between SNV signatures and INDEL signature files are the mutation types. Specially, for the ‘Fit’ analysis of SigProfiler, the order of mutation type in SNV signature file is different. Please check the first

column in “SBS96_De-Novo_Signatures.txt” for details. Also, the name of mutation type in INDEL is different in ‘Fit’ analysis of SigProfiler, which is shown in the file “INDEL_rownames.txt”. The corresponding INDEL signature file could be prepared as file “COSMIC_ID83_Signatures.txt”.

1.4. Activate file

Activate file records the contributions of mutations in each signature to the mutations of sample, which is only used in ‘Fit’ function of SigProfiler. The first line is signature IDs, and the first column is sample IDs. The numbers in the file could be the relative count and absolute count of the signatures belonging to each sample. Here is the example from “SBS96_De-Novo_Activities_refit.txt”.

Samples	SBS96A	SBS96B
BF2_S3	2954	479
IL11	736	0
PD28690bx_2_a5	0	3325
PD28690bx_2_d5	398	3125
PD34199a_c17	425	2258
PD34199a_c27	0	2936
PD34209ac	1443	322
PD34209al	2047	348
PD36713b_lo013	2742	0
PD36713b_lo023	3540	197
PD36718b_lo001	651	47
PD36718b_lo002	1339	92
PD37451br	1966	284

1.5. Seeds file

Seeds file provides seed numbers only for ‘Extract’ function of SigProfiler. Here is the example from “Seeds.txt”.

Replicates	Seeds
1	9745335
2	927617
3	839252
4	3409585
5	7924104
6	1597538
7	79364
8	8270294
9	6809254
10	6648386
11	8924081
12	2285059
13	1103912
14	5237750
15	9933304
16	5043209
17	7529075
18	5865754
19	1395095
20	275602

2 General output files

The common output files generated by the ‘Analysis’ function in SomaMutDB are described as follows.

- “SNV_mat.txt” and “INDEL_mat.txt” are the count matrices for analysis generated from uploaded files (the rows will be re-sorted for SigProfiler)

- “6types_SNV.pdf” and “16types_pdf” are the figures for the contributions to samples in simple mutation types of SNV and INDEL, respectively.
- “SNV_sig.txt” and “INDEL_sig.txt” are the count data for signature distributions, while “Sigs_SNV.pdf” and “Sigs_INDEL.pdf” are the corresponding figures.
- SNV_con.txt and INDEL_con.txt are the contributions/activates of signatures in each sample, while con_SNV.pdf and con_INDEL.pdf are the corresponding figures.
- File names beginning with “estimate” are the figures aiming to help users estimate the number of signatures used for signature extraction.
- Any files with “CellGroup” in name are the results for each group instead of samples according to the Group file.
- “fit.txt” and “fit.pdf” are the results for all tools in ‘Fit’ function.
- “config.json” records all the parameters used for signature analysis.

3 Extract function

The parameters for examples may be not suitable real analysis, please use default parameters or change the parameters only if you know its impacts in your analysis.

3.1. MutationalPatterns tool

3.1.1. NMF method

Only somatic mutations file and group file are required.

Input file:	test_small.hg19.tar.gz
Group file:	group_small.txt
Range of signature number to estimate:	2:3

The example results are compressed as “MutationalPatterns_NMF.tar.gz”.

3.1.2. Bayesian NMF method

Somatic mutations file and group file are required.

Input file:	test_small.hg19.tar.gz
Group file:	group_small.txt
Range of signature number to estimate:	2:3

The example results are compressed as “MutationalPatterns_BayesianNMF.tar.gz”.

3.2. SomaticSignatures tool

3.2.1. NMF method

Somatic mutations file and group file are required.

Input file:	test_small.hg19.tar.gz
-------------	------------------------

Group file: group_small.txt
Range of signature number to estimate: 2:3

The example results are compressed as “SomaticSignatures_NMF.tar.gz”.

3.2.2. PCA method

Somatic mutations file and group file are required.

Input file: test_small.hg19.tar.gz
Group file: group_small.txt
Range of signature number to estimate: 2:3

The example results are compressed as “SomaticSignatures_PCA.tar.gz”.

3.3. hdp tool

3.3.1. *de novo* analysis

Somatic mutations file and group file are required.

Input file: test_small.hg19.tar.gz
Group file: group_small.txt
burnin: 500

The example results are compressed as “hdp_denovo.tar.gz”. The files with “low” or “high” in names are the results at 95% confidential intervals (CI).

3.3.2. analysis with prior signatures

Somatic mutations file, group file and prior signature files are required.

Input file: test_small.hg19.tar.gz
Group file: group_small.txt
priorSNV (prior mode): prior_SNV_hdp.txt
priorINDEL (prior mode): prior_INDEL_hdp.txt
burnin: 500

The example results are compressed as “hdp_prior.tar.gz”. The files with “low” or “high” in names are the results at 95% confidential intervals (CI).

3.4. SignatureToolsLib tool

3.4.1. Regular analysis

Somatic mutations file and group file are required.

Input file: test_small.hg19.tar.gz
Group file: group_small.txt
Range of signature number to estimate: 2:3
nrepeats: 20

The example results are compressed as “SignatureToolsLib_normal.tar.gz”.

3.4.2. Fixed signatures during analysis

Somatic mutations file, group file and fixed signature files are required.

Input file:	test_small.hg19.tar.gz
Group file:	group_small.txt
Range of signature number to estimate:	2:3
matrix of fixed signatures SNV:	fix_SNV_SignatureToolsLib.txt
matrix of fixed signatures INDEL:	fix_INDEL_SignatureToolsLib.txt
nrepeats:	20

The example results are compressed as “SignatureToolsLib_fix.tar.gz”. Fixed signatures will be forced to use in signature extraction.

3.5. SigProfiler tool

Somatic mutations file and group file are required.

Input file:	test_small.hg19.tar.gz
Group file:	group_small.txt
Range of signature number to estimate:	2:3
Seed file:	Seeds.txt
nmf_replicates:	20

The example results are compressed as “SigProfiler.tar.gz”. Files with “sem” in name are the results of standard error of the mean, while files with “cosmic” in name are results for COSMIC signatures decomposition.

4 Similarity function

Only signatures files are required.

Signature file 1:	prior_INDEL_hdp.txt
Signature file 2:	fix_INDEL_SignatureToolsLib.txt

The example results are compressed as “CosmicSimilarity.tar.gz”.

5 Fit function

5.1. MutationalPatterns tool

5.1.1. NMF method

Somatic mutations file and signature file are required.

Input file:	test_small.hg19.txt
Prior file:	prior_INDEL_hdp.txt

The example results are compressed as “Fit_MutationalPatterns_NMF.tar.gz”.

5.1.2. strictNMF method

Somatic mutations file and signature file are required.

Input file:	test_small.hg19.txt
-------------	---------------------

Prior file: prior_INDEL_hdp.txt

The example results are compressed as “Fit_MutationalPatterns_strictNMF.tar.gz”.

5.1.3. bootstrappedNMF method

Somatic mutations file and signature file are required.

Input file: test_small.hg19.txt
Prior file: prior_INDEL_hdp.txt

The example results are compressed as “Fit_MutationalPatterns_bootstrappedNMF.tar.gz”.

5.2. mmsig tool

5.2.1. bootstrapped analysis

Somatic mutations file and signature file are required.

Input file: test_small.hg19.txt
Prior file: prior_SNV_hdp.txt
force_include: SBS1, SBS5
bootstrap: √
bootstrap reducing: √
iterations: 20

The example results are compressed as “Fit_mmsig_SNV_bootstrapped.tar.gz”.

5.2.2. regular analysis

Somatic mutations file and signature file are required.

Input file: test_small.hg19.txt
Prior file: prior_INDEL_hdp.txt
force_include: ID1, ID2
iterations: 100

The example results are compressed as “Fit_mmsig_INDEL.tar.gz”.

5.3. SignatureToolsLib

Somatic mutations file and signature file are required.

Input file: test_small.hg19.txt
Prior file: prior_INDEL_hdp.txt
nboot: 20

The example results are compressed as “Fit_SignatureToolsLib.tar.gz”.

5.4. SigProfiler

5.4.1. regular analysis

Somatic mutations file, signature file (prior file) and activity file (contributions of signatures to the somatic mutations) are required.

Input file: test_small.hg19.txt
Activities file: SBS96_De-Novo_Activities_fit.txt
Prior file: SBS96_De-Novo_Signatures.txt

The example results are compressed as “Fit_SigProfiler_SNV.tar.gz”. “SNV_con_fit.txt” and “SNV_con_fit.pdf” are the decomposition results based on COSMIC signatures. “SNV_con.txt” and “SNV_con.pdf” are the results for user defined signatures.

5.4.2. fit somatic mutations and *de novo* signatures with user defined signatures

Somatic mutations file, signature file, activity file (contributions of signatures to the somatic mutations) and user defined signature file (signature database file, e.g., COSMIC signatures) are required.

Input file: test_small.hg19.txt
Activities file: ID83_S2_NMF_Activities.txt
Prior file: INDEL_sig_denovo.txt
signature database file: COSMIC_ID83_Signatures.txt

The example results are compressed as “Fit_mmsig_INDEL.tar.gz”. “INDEL_con_fit.txt” and “INDEL_con_fit.pdf” are the decomposition results based on user defined signatures; “INDEL_con.txt” and “INDEL_con.pdf” are the results for *de novo* signatures by users.