Hands on – SigProfiler suite of tools

1. Introduction to SigProfiler

SigProfiler provides a comprehensive and integrated suite of bioinformatic tools for performing mutational signature analysis. The software covers the analytical lifecycle starting with the generation of the mutational matrices (**SigProfilerMatrixGenerator**) and finishing with signature extraction (**SigProfilerExtractor**) and assignment (**SigProfilerAssignment**), as well as supporting functionality for plotting and simulation.

As part of this hands-on section of the course, we will cover the whole cycle of a mutational signature analysis, both using the default test data available within SigProfiler, as well as using publicly available data from the Memorial Sloan Kettering Cancer Center cBioPortal platform.

SigProfiler packages have been developed using Python. However, an **R wrapper** is available for most of the tools (note the final **R** added to the name of the packages). Since we will use **RStudio** to perform our analyses, we will be using mostly these R wrappers. Although the packages are already installed in the virtual machines, it's important to consider for future applications that both the R wrapper and the original python packages need to be installed. Just installing the R wrapper is not enough to be able to run the tools.

2. Introduction to SigProfilerMatrixGeneratorR

The first step for performing mutational signature analysis is the generation of the input mutational matrices. These mutational matrices correspond to the categorization of the mutations present in our samples into a set of mutational contexts.

SigProfilerMatrixGeneratorR is the SigProfiler tool designed for this specific task, and it generates mutational matrices, as well as mutational profile plots for all the input samples provided. For example, for single base substitutions (SBS), SigProfilerMatrixGeneratorR automatically gather the nucleotides before and after a particular mutation from the reference genome. In this way, the mutation can be classified in one of the 96 subtypes defined in the SBS classification.

2.1 Reference genomes

The first step to run SigProfilerMatrixGeneratorR is installing a reference genome, that should match the one used for the alignment of the next generation sequencing data. We have already preinstalled human reference genomes GRCh37 and GRCh38 in the virtual machines, but in case you need to install these genomes (or different ones) on a different computer you can follow the R code in the **`SigProfilerMatrixGenerator_tutorial.R`** script.

3. Generating mutational matrices and mutational profile plots with SigProfilerMatrixGeneratorR

In order to generate mutational matrices and mutational profile plots, VCF or MAF files can be used

As a first example, we will use the VCF files found in the `./datasets/21BRCA_vcf/` folder, which correspond to **21 breast cancer samples** from Nik-Zainal *et al.* 2012 Cell.

Subsequently, we will use a set of mutation data that has been prepared from the **TCGA female breast cancer dataset**, downloaded from cBioPortal. The datasets have been created based on the molecular subtypes, namely:

- Basal (171 samples)
- Her2 (78 samples)
- LumA (499 samples)
- LumB (197 samples)

These datasets are found in the `./datasets/tcga_brca` folder, and consist of the mutation data in CSV format, found in the folder `./datasets/tcga_brca/original_data/`. For this example, we will only generate the mutational matrices for the **her2 subtype**. You can find these data in the following path: `./datasets/tcga_brca/original_data/tcga_brca_her2.csv`

You can run SigProfilerMatrixGeneratorR for both datasets using the code in the **`SigProfilerMatrixGenerator_tutorial.R`** script.

4. Outputs of SigProfilerMatrixGeneratorR

The provided code will generate all types of mutational matrices. For the **21 breast cancer dataset**, the matrices will be in the generated `./21BRCA_vcf/output/` folder. We will particularly use the following three matrices for our downstream mutational signature analysis:

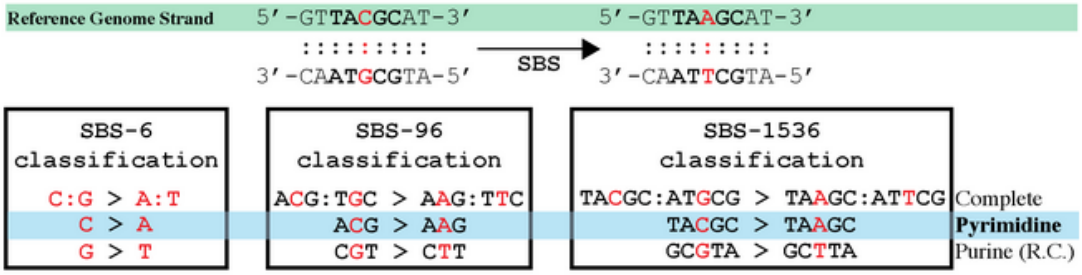
- SBS96
- DBS78
- ID83

Note that for this particular set of samples we will only obtain SBS and DBS mutational matrices, since the input VCF files only contain these variant types (indels were excluded from these files).

The output folder will contain the following subfolders:

- `./21BRCA_vcf/output/`**`SBS`**
- `./21BRCA_vcf/output/`**`plots`**
- `./21BRCA_vcf/output/`**`DBS`**

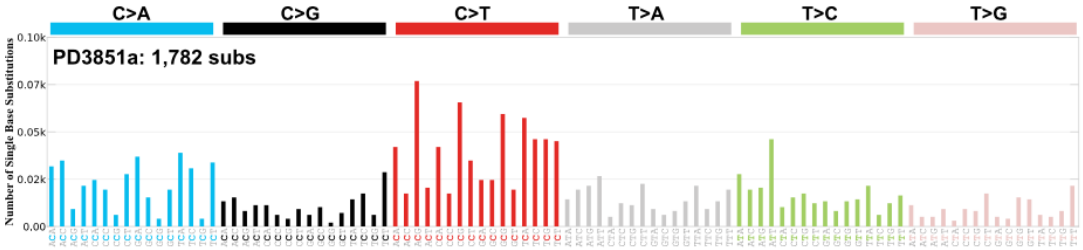
In the `./21BRCA_vcf/output/SBS` subfolder you will find all the mutational matrices for the SBS variant type, according to the different classifications that you can review here. For the purpose of this course, we are going to use the SBS96 matrix, which classifies the SBS mutations considering the mutation type, as well as the nucleotides immediately before and after the mutation (as reviewed during the lecture). You have an example of the classification of a specific mutation in the image below:



Similarly, in the `./21BRCA_vcf/output/DBS` subfolder you will find all the mutational matrices for the DBS variant type (which you can review here). In this case, we will focus on the DBS78 matrix.

Mutational profile plots are available for all samples used as input, as a visualization of the mutational matrices. They can be found in the `./21BRCA_vcf/output/plots` subfolder.

A visualization of a SBS96 mutational matrix for sample PD3851a is shown below:



Similarly, running SigProfilerMatrixGeneratorR with the TCGA breast cancer cohort data will generate all mutational matrices and profile plots, following a similar folder structure as in the case of the 21 breast cancer dataset for each of the variant types (SBS, DBS and ID).

Now we are ready to perform mutational signature analysis. We can use the newly generated mutational matrices as input, or we can use again the same input VCF/MAF files. This last option is possible because SigProfilerExtractorR uses SigProfilerMatrixGeneratorR behind-the-scenes to automatically generate the mutational matrices, if not previously done by the user.

5. Introduction to SigProfilerExtractorR

SigProfilerExtractorR is the **R wrapper** for SigProfilerExtractor, developed in Python. It provides the function **sigprofilerextractor** which actually calls the python program **SigProfilerExtractor.** We will use it to extract *de novo* mutational signatures from a set of samples and decompose the *de novo* extracted signatures into the COSMIC signatures. To perform the decomposition into reference COSMIC signatures, SigProfilerAssignment is used behind-the-scenes.

5.1 Reference genomes used by SigProfilerExtractor

SigProfilerExtractor works with the same reference genomes used by SigProfilerMatrixGeneratorR. The reference genome is needed when we perform mutational signature extraction using VCF or MAF files as input. For more information check section 2.1 above.

5.2 The sigprofilerextractor function

sigprofilerextractor can extract *de novo* mutational signatures from **VCF** files or **tab delimited mutational tables representing** the mutation matrices.

The details of the SigProfilerExtractor function:

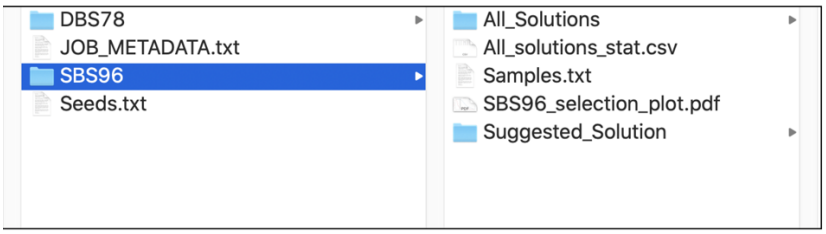
```
sigprofilerextractor(input_type,
                     out_put,
                     input_data,
                     reference_genome="GRCh37",
                     opportunity_genome = "GRCh37",
                     context_type = "default",
                     exome = False,
                     minimum_signatures=1,
                     maximum_signatures=25,
                     nmf_replicates=100,
                     resample = True,
                     batch_size=1,
                     cpu=-1,
                     gpu=False,
                     nmf_init="random",
                     precision= "single",
                     matrix_normalization= "gmm",
                     seeds= "none",
                     min_nmf_iterations= 10000,
                     max_nmf_iterations=1000000,
                     nmf_test_conv= 10000,
                     nmf_tolerance= 1e-15,
                     nnls_add_penalty=0.05,
                     nnls_remove_penalty=0.01,
                     initial_remove_penalty=0.05,
                     de_novo_fit_penalty=0.02,
                     get_all_signature_matrices= False)
```

Some key parameters

input_type	The type of input: "**vcf**": used for **vcf** format inputs or **maf** formats "**matrix**": used for tab-separated table format representing a mutation matrix
out_put	The name of the output folder. The output folder will be generated in the current working directory.
input_data	Name of the **input folder (in case of "vcf" type input)** or **the input file (in case of "table" type input)**. The project file or folder should be inside the current working directory. For the "vcf" type input, the project has to be a folder which will contain the vcf files in vcf format or text formats or the maf file.
reference_genome	The name of the reference genome. The default reference genome is "GRCh37". **This parameter is applicable only if the input_type is "vcf".**
context_type	A string of mutation context name/names separated by comma (","). The items in the list defines the mutational contexts to be considered to extract the signatures. The default value is "96,DINUC,ID", where "96" is the SBS96 context, "DINUC" is the DBS78 context and ID is ID83 context.
minimum_signatures	The minimum number of signatures to be extracted. The default value is 1.
maximum_signatures	The maximum number of signatures to be extracted. The default value is 25.
nmf_replicates	The **number of replicates** to be performed for each number of signatures. The default value is 100.
min_nmf_iterations	Value defines the **minimum number** of iterations to be completed before NMF converges. Default is 10000.
max_nmf_iterations	Value defines the **maximum number** of iterations to be completed before NMF converges. Default is 1000000.
cpu	The **number of processors (cores)** to be used to extract the signatures. The default value is -1 which will use all available processors.
gpu	Defines if the GPU resource will used if available. Default is False. **If True**, the GPU resources will be used in the computation.
batch_size	Will be effective **only if the GPU is used**. Defines the number of NMF replicates to be performed by each CPU during the parallel processing. Default is 1.

6. Outputs generated by SigProfilerExtractorR

The output folder of the sigprofilerextractor function will contain subdirectories for each **mutational context (SBS96, ID83, DBS78)** passed in the `context_type` parameter, a **JOB_METADATA.txt** file, and a **Seeds.txt** file. Below is a preliminary view of the files that will be generated in results.



6.1 JOB_METADATA.txt

This file contains all the metadata about the system and runtime of the job. The main sections include the following:

- System Info
- Python and Package Versions
- Execution Parameters
- Analysis Progress
- Job Status

6.2 Seeds.txt

A text file with a seed ID. This file can be passed through the `seed` parameter in order to reproduce a run.

6.3 Mutational context subdirectory

For this section, the subdirectory SBS96 and its contents will be used as an example. Different mutational contexts will share the same file structure. Each mutational context subdirectory (ex. SBS96, ID83, DBS78) contains the following files:

- All_Solutions subdirectory
- Suggested_Solution subdirectory
- All_solutions_stat.csv
- SBS96_selection_plot.pdf
- Samples.txt

6.3.1 All_Solutions subdirectory

The All_Solutions subdirectory contains the results from running extractions at each rank within the range of the input.

Each of the solution directories (SBS96_1_Signatures, …, SBS96_10_Signatures) contains the subdirectories **Activities**, **Signatures**, and **Solution_Stat**s. Each filename in the subdirectories is prepended with the mutational context and signature number (ex. SBS96_S1), with the exception of the signature plots. For example, for SBS96_S3, the files in its respective directories are listed below:

Activities
- SBS96_S3_NMF_Activities_SEM_Error.txt
- SBS96_S3_NMF_Activities.txt
- SBS96_S3_NMF_Activity_Plots.pdf
- SBS96_S3_TMB_NMF_plot.pdf

Signatures
- SBS96_S3_Signatures_SEM_Error.txt
- SBS96_S3_Signatures.txt
- Signature_plotSBS_96_plots_S3.pdf

Solution_Stats
- SBS96_S3_NMF_Convergence_Information.txt
- SBS96_S3_Samples_stats.txt
- SBS96_S3_Signatures_stats.txt

6.3.1.1 The Activities subdirectory

SBS96_S3_NMF_Activities_SEM_Error.txt – There are different activity matrices generated with each iteration, and from these the average is then calculated. The first column lists all of the samples, and the subsequent columns lists the error of the average (standard error) for each sample and the respective signature. Below is a screenshot of the first few rows of a sample file, SBS96_S3_NMF_Activities_SEM_Error.txt. There were three signatures identified, SBS96A, SBS96B, and SBS96C.

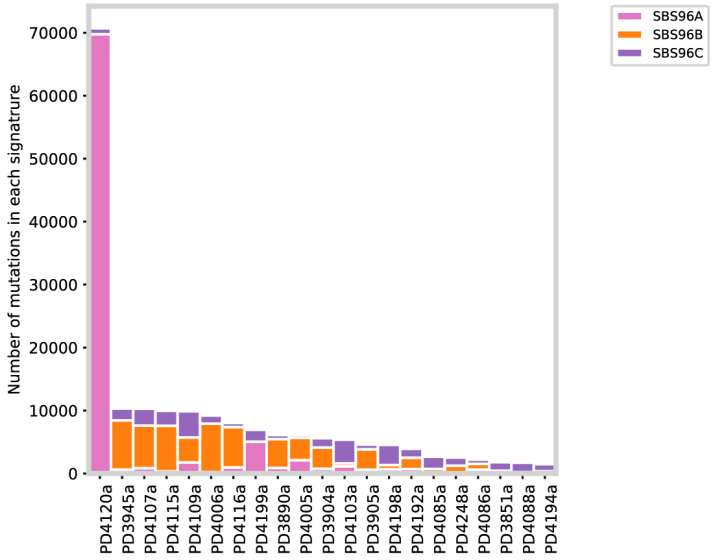
Samples	SBS96A	SBS96B	SBS96C
PD3851a	2.57E+00	1.25E+01	1.14E+01
PD3890a	8.77E+00	1.54E+01	1.95E+01
PD3904a	6.86E+00	1.88E+01	2.18E+01
PD3905a	6.72E+00	1.44E+01	1.71E+01
PD3945a	1.43E+01	3.23E+01	4.1E+01
PD4005a	6.68E+00	1.41E+01	1.57E+01
PD4006a	1.4E+01	2.69E+01	3.63E+01
PD4085a	3.79E+00	2.01E+01	1.83E+01

SBS96_S3_NMF_Activities.txt – This file contains the activity matrix for the signature. The first column lists all of the samples and the second column lists the calculated activity value for the respective

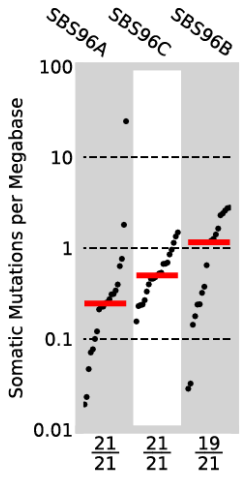
signature. The number of columns is the number of signatures identified. Below is a screenshot of the first few rows of a sample file, SBS96_S3_Activities.txt. There were three signatures, SBS96A, SBS96B, SBS96C, that were identified.

Samples	SBS96A	SBS96B	SBS96C
PD3851a	53	400	1329
PD3890a	874	4592	658
PD3904a	764	3375	1469
PD3905a	590	3251	746
PD3945a	634	7802	1872
PD4005a	2130	3539	435
PD4006a	215	7694	1285
PD4085a	64	673	1936

SBS96_S3_NMF_Activity_Plots.pdf – This plot shows the number of mutations in each signature on the y-axis and the sample name on the x-axis. The colors indicate which signature had the mutations and which signatures were found in each sample.



SBS96_S3_TMB_NMF_plot.pdf – This file contains a tumor mutational burden plot. The y-axis is the somatic mutations per megabase and the x-axis is the number of samples plotted over the number of samples included. The column names are the mutational signatures and the plot is ordered by the median somatic mutations per megabase.



*Showing samples with counts more than 0

6.3.1.2 The Signatures subdirectory

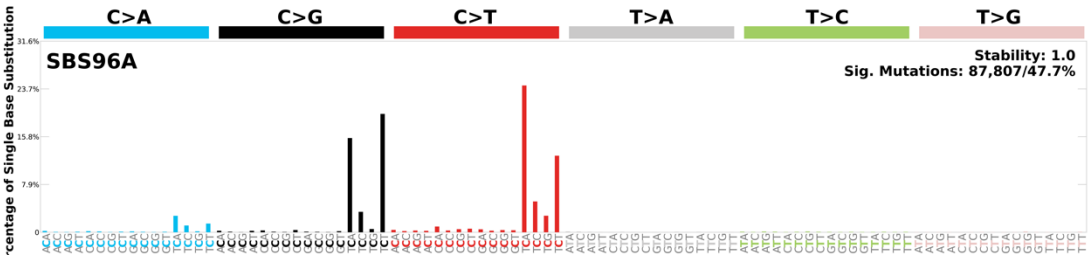
SBS96_S3_Signatures_SEM_Error.txt – Information about the different signature matrices generated during the run is stored in in this file. There are different signature matrices from each iteration from which the average is then calculated. The first column lists all of the samples and the subsequent columns lists the error of the average (standard error) for each signature and the respective signature.

Samples	SBS96A	SBS96B	SBS96C
PD3851a	2.57E+00	1.25E+01	1.14E+01
PD3890a	8.77E+00	1.54E+01	1.95E+01
PD3904a	6.86E+00	1.88E+01	2.18E+01
PD3905a	6.72E+00	1.44E+01	1.71E+01
PD3945a	1.43E+01	3.23E+01	4.1E+01
PD4005a	6.68E+00	1.41E+01	1.57E+01
PD4006a	1.4E+01	2.69E+01	3.63E+01
PD4085a	3.79E+00	2.01E+01	1.83E+01

SBS96_S3_Signatures.txt – This file contains the contribution of each mutation to the observed signature. The first column lists each mutation possible in the mutational context. There are 96 possible mutation types that are considered for SBS96. The following columns are the signatures. In the example, the signatures are SBS96A, SBS96B, and SBS96C. Only the first few rows are shown in the image below; however, the sum of each column is 1, and each value in a column indicates the contribution of a mutational context to the signature.

MutationType	SBS96A	SBS96B	SBS96C
A[C>A]A	0.002233656161697580	0.021869640555232800	0.020044031580910100
A[C>A]C	0.000661568925570464 0	0.018994258753955400	0.015116348611190900
A[C>A]G	0.000381261857837671 00	0.002663589318981390 0	0.002905753884697330
A[C>A]T	0.000936798184702639 0	0.019618410300463400	0.011061954498291000
A[C>G]A	0.002500529708340760	0.01701367654837670	0.005774734576698390
A[C>G]C	0.000993777128751390 0	0.009572514193132 52	0.006354089551605280
A[C>G]G	0.000194899615707982 00	0.004668717703316360	0.001644179549475670
A[C>G]T	0.002439858963480220 0	0.016890175649896300	0.006604584231972700
A[C>T]A	0.003914246312342580	0.017154338406398900	0.022299492601305200

Signature_plotSBS_96_plots_S3.pdf – It has a plot for each signature identified that depicts the proportion of the mutations for that signature and X is that number. For more details on the plots, plotting tools, and interpretation of the plots please refer to the SigProfilerPlotting tool.
In the example below, the plot generated for the first signature (SBS96A) identified in the sample input is shown. The top right corner lists the stability, total number of mutations, and the percentage of total mutations assigned to this mutational signature.



6.3.1.3 The Solution_Stats subdirectory

SBS96_S3_NMF_Convergence_Information.txt – This file contains the L1 norm (calculated as the sum of the absolute values of the vector), L2 norm (calculated as the square root of the sum of the squared vector values), KL divergence, and correlation between the original and reconstructed mutational matrix for each NMF replicate, as well as the number of convergence iterations.

NMF_Replicate	L1	L1 %	L2	L2 %	KL Divergence	Correlation	Convergence Iterations
1	926.734	18.527	146.812	19.731	0.032	0.944	20000.0
2	902.929	18.157	143.034	19.329	0.033	0.947	20000.0
3	938.456	18.691	146.938	19.512	0.034	0.948	20000.0
4	938.055	18.467	149.179	19.452	0.033	0.948	30000.0
5	907.963	18.122	142.063	18.919	0.031	0.95	30000.0
6	933.192	18.649	147.668	19.772	0.032	0.945	20000.0
7	934.838	18.314	148.178	19.207	0.031	0.949	20000.0
8	937.074	18.733	147.329	19.463	0.033	0.948	30000.0
9	907.963	17.966	145.276	18.992	0.03	0.951	20000.0
10	942.938	18.847	152.369	20.315	0.034	0.942	20000.0
11	936.241	18.618	147.177	19.566	0.033	0.946	20000.0
12	915.206	18.123	145.281	19.182	0.032	0.949	30000.0

SBS96_S3_Samples_stats.txt – This file contains the statistics for each sample including the total number of mutations, cosine similarity, L1 norm (calculated as the sum of the absolute values of the vector), L1 norm percentage, L2 norm (calculated as the square root of the sum of the squared vector values), and L2 norm percentage, along with the KL divergence and correlation.

Sample Names	Total Mutations	Cosine Similarity	L1 Norm	L1_Norm_%	L2 Norm	L2_Norm_%	KL Divergence	Correlation
PD3851a	1782	0.978	344.001	19.304%	48.148	21.003%	0.03111	0.941
PD3890a	6124	0.981	1021.516	16.681%	151.353	19.195%	0.02143	0.949
PD3904a	5608	0.984	919.18	16.391%	129.846	17.748%	0.02025	0.96
PD3905a	4587	0.991	544.507	11.871%	74.893	13.063%	0.0113	0.974
PD3945a	10308	0.991	1147.176	11.129%	169.428	13.69%	0.01244	0.968
PD4005a	6104	0.985	1005.018	16.465%	187.649	17.179%	0.02041	0.979

SBS96_S3_Signatures_stats.txt – This file contains the statistics for each of the signatures identified and includes their stability value (calculated average silhouette coefficient) and total number of mutations assigned.

Signatures	Stability	Total Mutations
SBS96A	1.0	87807
SBS96B	0.975	60781
SBS96C	0.97	35328

6.3.2 All_solutions_stat.csv

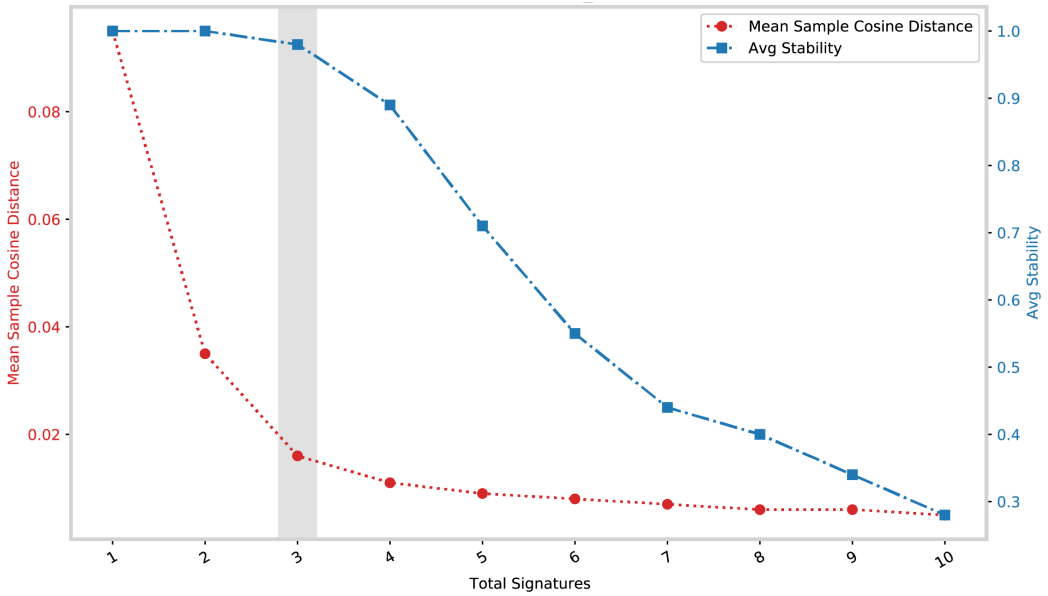
This file contains the record of the relative reconstruction error (the squared distance between the original data and its "estimate") and process stability. This file contains columns with the following values for each signature identified from the input samples:

- Stability (calculated average silhouette coefficient)
- Minimum Stability
- Considerable Solution
- P-value
- Matrix Frobenius %
- Mean sample L1% (calculated as the sum of the absolute values of the vector) %
- Maximum sample L1%
- Mean sample L2% (calculated as the square root of the sum of the squared vector values) %
- Maximum sample L2%
- Mean sample KL (Kullback-Leibler divergence)
- Maximum sample KL
- Mean Cosine Distance
- Max Cosine Distance
- Mean Correlation
- Minimum Correlation

Signatures	Stability (Avg Silhouette)	Minimum Stability	Considerable Solution	P-value	Matrix Frobenius%	Mean Sample L1%	Maximum Sample L1%	Mean Sample L2%	Maximum Sample L2%	Mean Sample KL	Maximum Sample KL	Mean Cosine Distance	Max Cosine Distance	Mean Correlation	Minimum Correlation
1	1.0	1.0	NO	N/A	71.26%	30.48%	112.11%	42.78%	71.67	0.1225	0.7597	0.095	0.192	0.835	0.584
2	1.0	1.0	YES	1.61E-04	5.87%	19.15%	40.47%	22.66%	47.62	0.0435	0.138	0.035	0.116	0.914	0.778
3*	0.98	0.97	YES	5.75E-02	5.78%	16.1%	26.16%	17.75%	26.08	0.0225	0.0538	0.016	0.034	0.961	0.911
4	0.89	0.79	YES	Most Stab Sigs	2.24%	11.98%	25.96%	13.05%	25.75	0.0185	0.0532	0.011	0.033	0.974	0.933
5	0.71	0.26	NO	N/A	1.85%	10.98%	25.92%	11.14%	25.69	0.0165	0.0527	0.009	0.033	0.977	0.934
6	0.55	-0.08	NO	N/A	1.68%	10.04%	22.7%	10.95%	21.37	0.014	0.0446	0.008	0.022	0.98	0.954
7	0.44	-0.4	NO	N/A	1.64%	9.65%	22.84%	10.93%	22.87	0.0133	0.0402	0.007	0.026	0.981	0.955
8	0.4	-0.12	NO	N/A	1.55%	9.04%	21.08%	10.06%	20.93	0.0114	0.0344	0.006	0.021	0.984	0.96
9	0.34	-0.16	NO	N/A	1.44%	8.71%	19.23%	9.67%	19.02	0.0107	0.0302	0.006	0.018	0.985	0.963
10	0.28	-0.23	NO	N/A	1.39%	8.68%	15.92%	8.57%	16.57	0.01	0.0219	0.005	0.014	0.987	0.962

6.3.3 SBS96_selection_plot.pdf

This file contains a plot between the mean sample cosine distance and the average stability. **The vertical gray bar indicates the optimal number of signatures selected by SigProfilerExtractor.**



6.3.4 Samples.txt

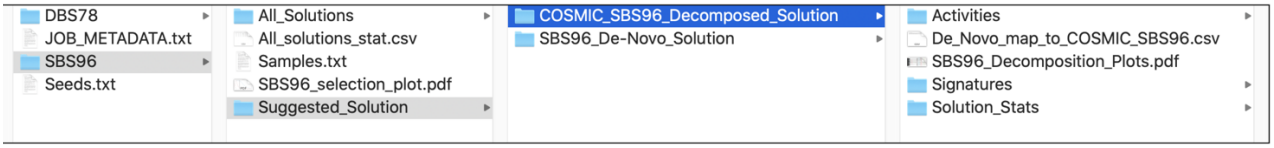
This file contains the original input mutational matrix (containing the number of mutations found in each of the samples corresponding to each mutational context). For example, in the file below, each row corresponds to a SBS96 mutational context A[C>A]A and every column corresponds to a sample (PD10010a). Thus, the numbers represent the number of mutations found in that context in each sample.

Mutation Types	PD3851a	PD3890a	PD3904a	PD3905a	PD3945a	PD4005a	PD4006a	PD4085a	PD4086a	PD4088a	PD4103a	PD4107a	PD4109a	PD4115a	PD4116a	PD4120a	PD4192a	PD4194a	PD4198a	PD4199a	PD4248a
A[C>A]A	31	110	122	94	243	74	198	61	31	34	112	210	122	228	128	165	48	28	58	58	64
A[C>A]C	34	91	112	69	163	66	173	51	19	18	50	176	133	169	138	55	42	24	72	36	43
A[C>A]G	9	9	13	11	24	12	33	7	8	3	7	33	12	21	15	20	16	6	12	13	7
A[C>A]T	21	87	107	65	155	64	192	49	23	26	44	176	96	158	142	65	40	13	45	37	35
A[C>G]A	13	100	52	66	130	56	164	18	18	15	30	126	95	128	102	191	60	8	37	23	25
A[C>G]C	15	46	42	41	78	34	96	18	14	11	28	85	72	74	71	72	26	5	27	19	18
A[C>G]G	8	18	18	15	19	16	56	7	0	2	9	61	25	21	22	18	28	1	8	2	4
A[C>G]T	11	101	63	71	116	79	156	25	22	16	33	118	104	130	103	170	43	8	32	26	31
A[C>T]A	41	99	127	65	168	61	170	62	28	46	99	191	162	179	127	345	51	24	75	38	56
A[C>T]C	17	62	54	34	80	49	85	28	18	15	52	68	87	89	63	133	39	19	29	30	21
A[C>T]G	75	95	94	84	150	108	152	128	57	125	242	157	328	194	79	213	120	81	204	121	75
A[C>T]T	20	79	96	53	171	78	170	28	23	29	60	176	113	149	94	162	36	11	46	40	22
A[T>A]A	14	53	40	39	131	25	58	22	17	12	29	65	68	79	95	112	34	11	64	12	24
A[T>A]C	19	40	31	30	92	36	63	30	19	10	32	66	61	74	62	66	34	14	36	29	21
A[T>A]G	21	36	49	40	150	32	100	29	14	10	29	91	65	99	60	57	37	9	36	22	19
A[T>A]T	26	51	63	58	182	34	121	38	21	14	53	121	186	141	97	53	42	14	86	27	51
A[T>C]A	27	89	100	54	201	72	126	56	37	20	72	169	179	173	135	89	51	27	94	28	60
A[T>C]C	19	44	37	31	63	34	99	22	18	10	27	95	73	80	67	40	24	10	25	22	21
A[T>C]G	20	72	52	53	127	61	107	34	40	18	49	115	118	120	96	84	40	15	52	37	34
A[T>C]T	45	104	73	79	205	89	156	64	50	27	69	180	157	203	113	87	59	19	60	32	39
A[T>G]A	11	37	23	25	55	16	52	13	13	4	22	53	47	51	33	30	28	3	50	6	9
A[T>G]C	5	20	10	9	27	7	21	9	6	1	7	23	19	18	23	23	24	1	23	7	6
A[T>G]G	5	39	28	42	72	38	87	7	11	5	10	71	44	67	46	31	29	2	14	3	10
A[T>G]T	9	71	22	34	60	32	43	8	11	5	27	45	43	48	41	31	29	7	24	2	22
C[C>A]A	24	108	99	86	199	66	175	36	30	38	64	178	136	184	163	135	45	14	62	37	42
C[C>A]C	19	66	72	55	173	77	147	33	17	12	32	167	81	139	159	86	28	10	40	27	21
C[C>A]G	6	10	7	17	19	5	32	3	2	4	12	29	15	16	28	25	15	4	8	8	7
C[C>A]T	27	67	84	64	180	46	153	23	22	25	66	188	112	149	124	82	20	7	40	27	29
C[C>G]A	11	61	56	42	105	56	106	8	9	8	28	93	72	100	79	219	36	5	21	20	10
C[C>G]C	6	47	41	30	89	49	74	16	6	7	21	82	44	93	70	61	26	3	17	14	15

6.3.5 Suggested_Solution subdirectory

The Suggested_Solution subdirectory contains the optimal solution. It contains 2 folders (in the case of SBS96 signatures):

- COSMIC_SBS96_Decomposed_Solution
- SBS96_De-Novo_Solution



6.3.5.1 COSMIC_SBS96_Decomposed_Solution

There are two files in COSMIC_SBS96_Decomposed_Solution that are not found in All Solutions, namely:

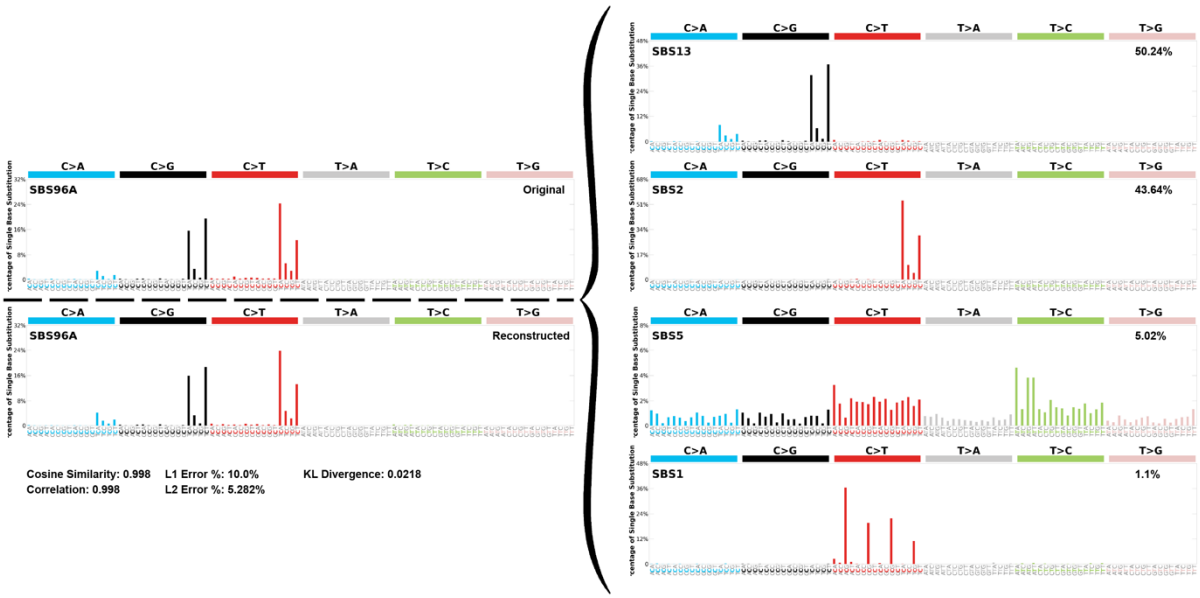
- De_Novo_map_to_COSMIC_SBS96.csv
- SBS96_Decomposition_Plots.pdf

De_Novo_map_to_COSMIC_SBS96.csv – This file contains data on how the *de novo* extracted signatures are decomposed using the COSMIC reference signatures. Additionally, it also contains information on the L1 error %, L2 error %, KL divergence, cosine similarity, and correlation of this decomposition. Below is an example of what the file contains.

De novo extracted	Global NMF Signatures	L1 Error %	L2 Error %	KL Divergence	Cosine Similarity	Correlation
Signature 96-A	Signature SBS1 (1.10%) & Signature SBS2 (43.64%) & Signature SBS5 (5.02%) & Signature SBS13 (50.24%)	10.03	6.00	0.022	1.00	1.00
Signature 96-B	Signature SBS1 (0.58%) & Signature SBS3 (66.72%) & Signature SBS5 (14.14%) & Signature SBS8 (18.56%)	20.20	23.58	0.042	0.97	0.87
Signature 96-C	Signature SBS1 (17.30%) & Signature SBS5 (21.84%) & Signature SBS40 (60.86%)	25.40	26.98	0.066	0.96	0.92

SBS96_Decomposition_Plots.pdf – This file contains a visualization of the results from De_Novo_map_to_COSMIC_SBS96.csv. There are two plots to the left of the curly brace. One is the original *de novo* signature plot and the other is the reconstruction of the *de novo* signature. On the right

side of the curly brace are the COSMIC signatures that the *de novo* signature is decomposed. Additionally, below the reconstructed plot are the data for cosine similarity, correlation, L1 error %, L2 error %, and KL divergence of the decomposition.



6.3.5.2 SBS96_De_Novo_Solution

There are two files in the SBS96_De_Novo_Solution directory not found in All_Signatures.

- De_Novo_Mutation_Probabilities_refit.txt
- SBS96_De-Novo_refit_Signature_Assignment_log.txt

De_Novo_Mutation_Probabilities_refit.txt - This file contains the mutation probability for each sample, mutational context and identified signature. Each of the signature mutation probabilities for a given sample and mutational context add up to 1.

Sample Names	MutationTypes	SBS96A	SBS96B	SBS96C
PD3851a	A[C>A]A	0.0	0.2577593416046730	0.7422406583953270
PD3851a	A[C>A]C	0.0	0.28568071906941200	0.7143192809305880
PD3851a	A[C>A]G	0.0	0.22586069589810300	0.7741393041018970
PD3851a	A[C>A]T	0.0	0.360808119175553	0.6391918808244470
PD3851a	A[C>G]A	0.0	0.48393305849328500	0.5160669415067150
PD3851a	A[C>G]C	0.0	0.32409455733785500	0.675905442662145
PD3851a	A[C>G]G	0.0	0.47472850683138200	0.5252714931686180
PD3851a	A[C>G]T	0.0	0.44871920209607100	0.5512807979039290
PD3851a	A[C>T]A	0.0	0.19668745019707400	0.8033125498029260
PD3851a	A[C>T]C	0.0	0.20133136402012900	0.7986686359798710
PD3851a	A[C>T]G	0.0	0.02766370375804570	0.9723362962419540
PD3851a	A[C>T]T	0.0	0.3261240712937690	0.6738759287062310

SBS96_De-Novo_refit_Signature_Assignment_log.txt - This file logs the events that occur when *de novo* extracted signatures are assigned to samples.

```
************************ Stepwise Description of Signature Assignment to Samples ************************



                    ################ Sample 1 #################
############################### Initial Composition ####################################
   SBS96A  SBS96B  SBS96C
0    53.0   400.0  1329.0
############################### Composition After  Remove ##############################
       SBS96B       SBS96C
0  430.366769  1351.633231
L2 Error %: 0.22
Cosine Similarity: 0.97



                    ################ Sample 2 #################
############################### Initial Composition ####################################
   SBS96A  SBS96B  SBS96C
0   874.0  4592.0   658.0
############################### Composition After  Remove ##############################
       SBS96A       SBS96B
0  917.262193  5206.737807
L2 Error %: 0.21
Cosine Similarity: 0.98



                    ################ Sample 3 #################
############################### Initial Composition ####################################
   SBS96A  SBS96B  SBS96C
0   764.0  3375.0  1469.0
############################### Composition After  Remove ##############################
   SBS96A  SBS96B  SBS96C
0   764.0  3375.0  1469.0
L2 Error %: 0.18
Cosine Similarity: 0.98
```

7. Extracting *de novo* mutational signatures using the test data available in SigProfilerExtractorR

7.1 Extracting signatures using the test data available in SigProfilerExtractorR

For our first example we will continue using the dataset corresponding to **21 breast cancer samples** from Nik-Zainal *et al.* 2012 Cell, which is available in SigProfilerExtractorR for testing purposes.

You can run SigProfilerExtractorR for this test dataset by following the **First Example** in the **`SigProfilerExtractor_tutorial.R`** script.

Note: We have restricted the **`nmf_replicates` to 5** so that it can run within a short time, and we can see some results. However, in real applications, you should leave this parameter to the default value of 100 or increase it to more iterations as we are solving an optimization problem. A larger number of replicates will converge to a more robust and stable solution.

The **output** generated will be as follows:

Hence it has extracted SBS96 and DBS78 signatures. If we explore down the specific folders of each signature sets and their suggested fitted COSMIC signatures, we will get the following outputs:

For SBS, open the file:

`SBS96/Suggested_Solution/COSMIC_SBS96_Decomposed_Solution/Signatures/SBS_96_plots_COSMIC_SBS96.pdf`

You will find COSMIC signatures SBS1, SBS2, SBS3, SBS5, SBS8, SBS13 & SBS40 fitted. Information can be obtained from the COSMIC website (https://cancer.sanger.ac.uk/signatures/):

Signature **COSMIC Proposed Etiology**

SBS1 An endogenous mutational process initiated by spontaneous or enzymatic deamination of 5-methylcytosine to thymine which generates G:T mismatches in double stranded DNA. Failure to detect and remove these mismatches prior to DNA replication results in fixation of the T substitution for C.

Comments: Signature SBS1 is clock-like in that the number of mutations in most cancers and normal cells correlates with the age of the individual. Rates of acquisition of Signature

SBS1 mutations over time differ markedly between different cancer types and different normal cell types. These differences correlate with estimated rates of stem cell division in different tissues and Signature SBS1 may therefore be a cell division/mitotic clock.

SBS2 Attributed to activity of the AID/APOBEC family of cytidine deaminases on the basis of similarities in the sequence context of cytosine mutations caused by APOBEC enzymes in experimental systems. APOBEC3A is probably responsible for most mutations in human cancer, although APOBEC3B may also contribute (these differ in the sequence context two bases 5' to the mutated cytosine, see 1,536 mutation classification signature extraction). SBS2 mutations may be generated directly by DNA replication across uracil or by error prone polymerases replicating across abasic sites generated by base excision repair removal of uracil.

Comments: SBS2 is usually found in the same samples as SBS13. It has been proposed that activation of AID/APOBEC cytidine deaminases in cancer may be due to previous viral infection, retrotransposon jumping, or tissue inflammation. Currently, there is limited evidence to support these hypotheses. Germline polymorphisms involving APOBEC3A and APOBEC3B are associated with predisposition to breast and bladder cancer as well as with mutation burdens of SBS2 and SBS13. Mutations of similar patterns to SBS2 and SBS13 are commonly found in the phenomenon of local hypermutation present in some cancers, known as kataegis, implicating AID/APOBEC enzymes in this process as well.

SBS3 Defective homologous recombination-based DNA damage repair which manifests predominantly as small indels and genome rearrangements due to abnormal double strand break repair but also in the form of this base substitution signature.

Comments: SBS3 is strongly associated with germline and somatic BRCA1 and BRCA2 mutations and BRCA1 promoter methylation in breast, pancreatic, and ovarian cancers. In pancreatic cancer, responders to platinum therapy usually exhibit SBS3 mutations. Together with associated indel and rearrangement signatures, SBS3 has been proposed as a predictor of defective homologous recombination-based repair and thus of response to therapies exploiting this repair defect.

SBS5 **Unknown**. SBS5 mutational burden is increased in bladder cancer samples with ERCC2 mutations and in many cancer types due to tobacco smoking.

Comments: SBS5 is clock-like in that the number of mutations in most cancers and normal cells correlates with the age of the individual. Rates of acquisition of SBS5 mutations over time differ between different cancer types and different normal cell types. These differences do not clearly correlate with estimated rates of stem cell division in different tissues nor with differences in SBS1 mutation rates. SBS5 may be contaminated by SBS16.

SBS8 **Unknown**

SBS13 Attributed to activity of the AID/APOBEC family of cytidine deaminases on the basis of similarities in the sequence context of cytosine mutations caused by APOBEC enzymes in experimental systems. APOBEC3A is probably responsible for most mutations in human cancer, although APOBEC3B may also contribute (these differ in the sequence context two

bases 5' to the mutated cytosine, see 1536 mutation classification signature extraction). SBS13 mutations are likely generated by error prone polymerases (such as REV1) replicating across abasic sites generated by base excision repair removal of uracil.

Comments: SBS13 is usually found in the same samples as SBS2. It has been proposed that activation of AID/APOBEC cytidine deaminases in cancer may be due to previous viral infection, retrotransposon jumping, or tissue inflammation. Currently, there is limited evidence to support these hypotheses. Germline polymorphisms involving APOBEC3A and APOBEC3B are associated with predisposition to breast and bladder cancer as well as with mutation burdens of **SBS2 and SBS13**. Mutations of similar patterns to SBS2 and SBS13 are commonly found in the phenomenon of local hypermutation present in some cancers, known as kataegis, implicating AID/APOBEC enzymes in this process as well.

SBS40 **Unknown**

For DBS, open the file:

`DBS78/Suggested_Solution/COSMIC_DBS78_Decomposed_Solution/Signatures/DBS_78_plots_COSMIC_DBS78.pdf`

You will find COSMIC signatures DBS2, DBS4, DBS6 & DBS11.

Signature COSMIC Proposed Etiology

DBS2 Exposure to tobacco smoking as well as other endogenous and/or exogenous mutagens (e.g., acetaldehyde).

Comments: DBS2 exhibits transcriptional strand bias with more GG>TT mutations than CC>AA on the untranscribed strands of genes indicative of damage on guanine and repair by transcription-coupled nucleotide excision repair. In addition to its presence in tobacco smoking induced cancers, DBS2 is also found in many cancer types unrelated to tobacco smoking. Its profile is similar to that of mutations in normal cells in mice. It may therefore also be an endogenously generated signature. Its mutation burden correlates with age of cancer diagnosis and this clock-like feature suggests that it is generated in normal human cells.

DBS4 **Unknown**.

Comments: Its profile is similar to that of a subset of mutations in normal mouse cells. It may therefore be an endogenously generated signature. Its mutation burden correlates with the age of cancer diagnosis and this clock-like feature suggests that it is generated in normal human cells.

DBS6 **Unknown**

DBS11 **Unknown**

7.2 Extracting signatures from a TCGA breast cancer dataset – using mutation data

Now we will use SigProfilerExtractorR to extract *de novo* mutational signatures from the **TCGA Breast Cancer her2 subtype**, using the mutation data downloaded from cBioPortal.

You can run SigProfilerExtractorR for this dataset by following the **Second Example** in the **`SigProfilerExtractor_tutorial.R`** script.

7.3 Extracting signatures from a TCGA breast cancer dataset – using individual mutational matrices

If you only want to extract signatures from specific variant types, i.e., either SBS, DBS or ID, you can extract the different mutational matrices using **SigProfilerMatrixGeneratorR** and then run **SigProfilerExtractorR** with the corresponding matrix only.

You can run SigProfilerExtractorR for this dataset starting from the mutational matrix of your choice by following the **Third Example** in the **`SigProfilerExtractor_tutorial.R`** script.

8. Assigning reference mutational signatures to individual samples with SigProfilerAssignment

In case that the number of samples available is small, such as in clinical settings, it is not possible to accurately extract *de novo* mutational signatures. In those cases, we can assign a set of reference mutational signatures, which normally corresponds to the COSMIC signatures, by using **SigProfilerAssignment**. This assignment analysis is always performed in a **sample by sample basis**, and is also known as mutational signature refitting analysis.

Currently, there is no R wrapper available for SigProfilerAssignment. However, using the **reticulate** R package, is possible to run the python code directly using **RStudio**. Also, SigProfilerAssignment requires mutational matrices as input, so we will use the mutational matrices obtained as part of the previous section 3, where we run SigProfilerMatrixGeneratorR.

We will use **SigProfilerAssignment** to assign SBS96 COSMIC reference mutational signatures to individual samples from both the **21 breast cancers** cohort, and the **TCGA Breast Cancer her2 subtype** cohort.

You can run SigProfilerAssignment for both datasets by following the **`SigProfilerAssignment_tutorial.R`** script.

The **output** obtained from SigProfilerAssignment will include a new folder named `Assignment_Solution`. This folder follows the same structure as the output from SigProfilerExtractorR, including **Signatures**, **Activities** and **Solution_Stats** subfolders as presented in previous section 6.3.1. The main visualizations for the activity of the different mutational signatures in the evaluated samples are the **activity bar plot** and the **tumor mutational burden plot**, as shown in section 6.3.1.1.

9. Exercises

9.1 TCGA breast cancer her2 dataset

Using the results obtained from SigProfilerExtractorR in sections 7.2 and 7.3 (which should be analogous), as well as the results from SigProfilerAssignment in section 8, please answer the questions below.

Exercise 1

(a) Explore the SBS96 folder from the SigProfilerExtractorR output and list down the COSMIC signatures found in the her2 dataset

(b) What do these signatures correspond to? Endogenous or exogenous mutational processes?

(c) Order the COSMIC signatures based on the number of samples where they are present

(d) Order the COSMIC signatures based on the number of mutations assigned to them

Exercise 2

Repeat Exercise 1 for DBS78 signatures

Exercise 3

Repeat Exercise 1 for ID83 signatures

Exercise 4

Explore the differences between the assignment of SBS96 COSMIC mutational signatures done with SigProfilerExtractorR and SigProfilerAssignment

9.2 TCGA breast cancer basal dataset

Reproduce the analysis on sections 7 and 8, but in this case using the **basal subtype** samples from the TCGA breast cancer cohort. The original mutation data can be found in the following folder:

`./datasets/tcga_brca/original_data/tcga_brca_basal.csv`

Exercise 5

Transform the original mutational data to a MAF format (check the `SigProfilerMatrixGenerator_tutorial.R` script if needed) and save the output in a new folder:

`./datasets/SigProfilerMatrixGenerator_TCGA_BRCA_BASAL/`

Exercise 6

Run a *de novo* mutational signature extraction analysis with SigProfilerExtractorR and save the output in a new folder:

`./outputs/SigProfilerExtractor_output_TCGA_BRCA_BASAL_MAF/`

Exercise 7

Run a refitting mutational signature analysis with SigProfilerAssignment for the SBS96, ID83 and SBS78 mutational types and save the output in the following new folders:

`./outputs/SBS96_assignment_output_TCGA_BRCA_BASAL/`

`./outputs/ID83_assignment_output_TCGA_BRCA_BASAL/`

`./outputs/DBS78_assignment_output_TCGA_BRCA_BASAL/`

Exercise 8

List down the COSMIC SBS signatures found in the basal subtype using SigProfilerExtractorR and SigProfilerAssignment. Are they different from those in the her2 subtype?

Exercise 9

Repeat Exercise 6 for the DBS and ID signatures

10. Micro Project for Day 5

1) Perform a complete mutational signature analysis for the breast cancer subtypes **lumA** and **lumB** including extraction of *de novo* signatures and refitting of COSMIC reference signatures. The original mutation datasets are stored in the following folders:

```
./datasets/tcga_brca/lumA/
./datasets/tcga_brca/lumB/
```

2) Compare the signature profiles of the breast cancer subtypes **basal**, **lumA**, **lumB** and **her2**. For each of the subtypes, list down the mutational signatures resulting from endogenous and exogenous causes. What can you conclude about each of the breast cancer subtypes?

11. Conclusion

You have been taught how to extract and assign mutational signatures from example datasets and real-case datasets already prepared for you, using VCF/MAF and matrix formats as inputs.

If you wish to conduct your own research using novel data in VCF formats, you can just follow the steps listed above.

If you do not have your own data, you can still work with public domain data. cBioPortal is a very rich resource with cancer datasets already prepared for you. You can download the mutation data of your choice from there, stratify into the types of interest (subtype or race, if available, sex for cancers where you believe the cancer can affect males and females differently, etc.), and go ahead and perform the mutational signature analyses.

Note that SigProfilerExtractor methodology is very sensitive to the total number of samples and mutations. Therefore, mutational signature analyses work better with WGS rather than WES or gene panel data. So, please beware of the type of mutation data you are using.

Interesting links:

1) cBioPortal: https://www.cbioportal.org/

2) COSMIC Mutational Signatures: https://cancer.sanger.ac.uk/signatures/

3) SigProfilerExtractor: https://github.com/AlexandrovLab/SigProfilerExtractor

4) SigProfilerExtractorR – R wrapper: https://github.com/AlexandrovLab/SigProfilerExtractorR

5) SigProfilerMatrixGenerator: https://github.com/AlexandrovLab/SigProfilerMatrixGenerator

6) SigProfilerMatrixGeneratorR – R wrapper: https://github.com/AlexandrovLab/SigProfilerMatrixGeneratorR

7) SigProfilerAssignment: https://github.com/AlexandrovLab/SigProfilerAssignment