

sambar_tutorial

December 1, 2019

1 SAMBAR usage example

Author: Genís Calderer - Kuijjer Lab (NCMM) - genis.calderer@gmail.com

SAMBAR is a method to identify subtypes of samples based on annotated mutation data. It uses gene mutation counts and membership of those genes in a set of pathways to summarize the mutation rate in a “pathway mutation score” thus desparasifying the mutation data. For each sample and each pathway a score is computed and it can be used to compare the samples to identify subtypes. The SAMBAR package uses binomial distance and hierarchical clustering to find subgroups of samples with similar pathway mutation patterns. The method was developed and published in the following paper: Cancer subtype identification using somatic mutation data, Kuijjer ML, Paulson JN, Salzman P, Ding W, Quackenbush J, British Journal of Cancer (May 16, 2018), doi: 10.1038/s41416-018-0109-7, <https://www.nature.com/articles/s41416-018-0109-7>, BioRxiv, doi: <https://doi.org/10.1101/228031>

This guide will use the toy data included in the SAMBAR package to showcase the usage of this package.

1.1 1. Importing SAMBAR from netZooPy

In order to use the SAMBAR functions it has to be imported from the netZooPy as follows:

```
In [11]: from netZooPy import sambar
```

To see the parameters of the main function one can use the following line:

```
In [ ]: help(sambar.sambar)
```

1.2 2. Selecting input files

The program requires a gene mutation dataset, a list of gene sizes, a list of cancer-associated genes (optional) and a list of pathways with its genes.

```
In [ ]: # These are the names of the files of the toy dataset.
        # The program by default runs with the toy data.
        mut_file = "mut.ucec.csv"
        cangenes = "genes.txt"
        sign_file = "h.all.v6.1.symbols.gmt"
        esize_file = "esizef.csv"
```

1.3 3. Run SAMBAR

The main SAMBAR function takes as input the filepaths of the datasets and returns a pathway score dataframe and a sample clustering dataframe for different cuts in the linkage tree. It also outputs a csv file for the adjusted mutation scores, pathway scores and clustering. The slow step in this method is the computation of the distance matrix, this matrix is also exported in case it's needed and rerunning the whole process is not wanted. The function runs first the desparcification and then the clustering.

```
In [ ]: pathway_scores, cluster_groups = sambar.sambar(mut_file, esize_file, cangenes, sign_file)
        #This will not run in this notebook as the specified files are in the package files and
```

```
In [15]: pathway_scores, cluster_groups = sambar.sambar() #Runs with the default files.
```

Sambar runtime: 2.9713971614837646

Clustering runtime: 4.223911762237549

1.4 4. Results

The pathway mutation scores and the sample groups are the output of the method.

```
In [16]: pathway_scores.head(10)
```

```
Out[16]:
```

	TCGA-A5-AOG3	TCGA-A5-AOG5	TCGA-A5-AOG9	\
HALLMARK_ADIPOGENESIS	0.000000	0.000000	0.001195	
HALLMARK_ALLOGRAFT_REJECTION	0.002480	0.000000	0.000266	
HALLMARK_ANDROGEN_RESPONSE	0.000000	0.000000	0.001152	
HALLMARK_ANGIOGENESIS	0.000000	0.000000	0.001392	
HALLMARK_APICAL_JUNCTION	0.000000	0.000000	0.001352	
HALLMARK_APICAL_SURFACE	0.000000	0.000000	0.000000	
HALLMARK_APOPTOSIS	0.000875	0.000000	0.000311	
HALLMARK_BILE_ACID_METABOLISM	0.000000	0.000987	0.000000	
HALLMARK_CHOLESTEROL_HOMEOSTASIS	0.000000	0.000000	0.000000	
HALLMARK_COAGULATION	0.001983	0.000000	0.000000	

	TCGA-A5-AOGA	TCGA-A5-AOGB	TCGA-A5-AOGD	\
HALLMARK_ADIPOGENESIS	0.000369	0.000409	0.0	
HALLMARK_ALLOGRAFT_REJECTION	0.000000	0.000124	0.0	
HALLMARK_ANDROGEN_RESPONSE	0.000000	0.000000	0.0	
HALLMARK_ANGIOGENESIS	0.000000	0.000000	0.0	
HALLMARK_APICAL_JUNCTION	0.000364	0.000152	0.0	
HALLMARK_APICAL_SURFACE	0.000000	0.000000	0.0	
HALLMARK_APOPTOSIS	0.000238	0.000085	0.0	
HALLMARK_BILE_ACID_METABOLISM	0.000000	0.000000	0.0	
HALLMARK_CHOLESTEROL_HOMEOSTASIS	0.000814	0.000105	0.0	
HALLMARK_COAGULATION	0.000633	0.000102	0.0	

	TCGA-A5-AOGE	TCGA-A5-AOGH	TCGA-A5-AOGJ	\
--	--------------	--------------	--------------	---

HALLMARK_ADIPOGENESIS	0.000000	0.000000	0.000000
HALLMARK_ALLOGRAFT_REJECTION	0.000000	0.000000	0.000000
HALLMARK_ANDROGEN_RESPONSE	0.000000	0.000000	0.000768
HALLMARK_ANGIOGENESIS	0.000000	0.000000	0.000000
HALLMARK_APICAL_JUNCTION	0.005127	0.000167	0.000000
HALLMARK_APICAL_SURFACE	0.000000	0.000000	0.000000
HALLMARK_APOPTOSIS	0.000000	0.000000	0.000000
HALLMARK_BILE_ACID_METABOLISM	0.000000	0.000000	0.000000
HALLMARK_CHOLESTEROL_HOMEOSTASIS	0.000000	0.000000	0.000000
HALLMARK_COAGULATION	0.000000	0.000000	0.000000

	TCGA-A5-A0GM	...	TCGA-D1-A160	\
HALLMARK_ADIPOGENESIS	0.000000	...	0.000525	
HALLMARK_ALLOGRAFT_REJECTION	0.000000	...	0.000381	
HALLMARK_ANDROGEN_RESPONSE	0.000000	...	0.000395	
HALLMARK_ANGIOGENESIS	0.000000	...	0.000000	
HALLMARK_APICAL_JUNCTION	0.000703	...	0.000313	
HALLMARK_APICAL_SURFACE	0.000000	...	0.000000	
HALLMARK_APOPTOSIS	0.001057	...	0.000084	
HALLMARK_BILE_ACID_METABOLISM	0.000000	...	0.000000	
HALLMARK_CHOLESTEROL_HOMEOSTASIS	0.003094	...	0.000000	
HALLMARK_COAGULATION	0.000000	...	0.000253	

	TCGA-D1-A161	TCGA-D1-A167	TCGA-D1-A16F	\
HALLMARK_ADIPOGENESIS	0.000000	0.000507	0.000000	
HALLMARK_ALLOGRAFT_REJECTION	0.000645	0.000161	0.000111	
HALLMARK_ANDROGEN_RESPONSE	0.000000	0.000231	0.000000	
HALLMARK_ANGIOGENESIS	0.000000	0.000000	0.000000	
HALLMARK_APICAL_JUNCTION	0.000431	0.000429	0.000278	
HALLMARK_APICAL_SURFACE	0.000000	0.000000	0.000000	
HALLMARK_APOPTOSIS	0.000853	0.000184	0.000220	
HALLMARK_BILE_ACID_METABOLISM	0.000000	0.000217	0.000000	
HALLMARK_CHOLESTEROL_HOMEOSTASIS	0.000000	0.000000	0.000000	
HALLMARK_COAGULATION	0.000000	0.000757	0.002150	

	TCGA-D1-A16X	TCGA-D1-A16Y	TCGA-D1-A17Q	\
HALLMARK_ADIPOGENESIS	0.000102	0.000034	0.000266	
HALLMARK_ALLOGRAFT_REJECTION	0.000210	0.000146	0.000257	
HALLMARK_ANDROGEN_RESPONSE	0.000040	0.000000	0.000182	
HALLMARK_ANGIOGENESIS	0.000000	0.000000	0.000424	
HALLMARK_APICAL_JUNCTION	0.000358	0.000325	0.000245	
HALLMARK_APICAL_SURFACE	0.000000	0.000000	0.000088	
HALLMARK_APOPTOSIS	0.000248	0.000400	0.000258	
HALLMARK_BILE_ACID_METABOLISM	0.000306	0.000000	0.000299	
HALLMARK_CHOLESTEROL_HOMEOSTASIS	0.000000	0.000000	0.000028	
HALLMARK_COAGULATION	0.000235	0.001148	0.000455	

TCGA-D1-A1NX TCGA-EY-A1GS TCGA-EY-A212

HALLMARK_ADIPOGENESIS	0.000000	0.000210	0.0
HALLMARK_ALLOGRAFT_REJECTION	0.000000	0.000000	0.0
HALLMARK_ANDROGEN_RESPONSE	0.000000	0.000000	0.0
HALLMARK_ANGIOGENESIS	0.000000	0.000000	0.0
HALLMARK_APICAL_JUNCTION	0.000519	0.000085	0.0
HALLMARK_APICAL_SURFACE	0.000000	0.000000	0.0
HALLMARK_APOPTOSIS	0.000000	0.000211	0.0
HALLMARK_BILE_ACID_METABOLISM	0.000000	0.000000	0.0
HALLMARK_CHOLESTEROL_HOMEOSTASIS	0.000000	0.000000	0.0
HALLMARK_COAGULATION	0.001008	0.001918	0.0

[10 rows x 247 columns]

In [17]: cluster_groups.head()

```
Out[17]:
```

	TCGA-A5-A0G3	TCGA-A5-A0G5	TCGA-A5-A0G9	TCGA-A5-A0GA	TCGA-A5-A0GB	\
X2	0	0	0	1	1	
X3	0	0	1	2	2	
X4	0	0	1	2	3	

	TCGA-A5-A0GD	TCGA-A5-A0GE	TCGA-A5-A0GH	TCGA-A5-A0GJ	TCGA-A5-A0GM	...	\
X2	0	0	0	0	0	...	
X3	0	0	1	0	0	...	
X4	0	0	1	0	0	...	

	TCGA-D1-A160	TCGA-D1-A161	TCGA-D1-A167	TCGA-D1-A16F	TCGA-D1-A16X	\
X2	1	0	1	0	1	
X3	2	1	2	1	2	
X4	2	1	3	1	3	

	TCGA-D1-A16Y	TCGA-D1-A17Q	TCGA-D1-A1NX	TCGA-EY-A1GS	TCGA-EY-A212
X2	1	1	0	1	0
X3	2	2	0	2	0
X4	3	3	0	2	0

[3 rows x 247 columns]

In []: