Introduction to Classification of Human Mutations

Contents

Simpati introductive vignette

Introduction of Simpati applied on somatic mutation data of cancer patients This workflow is recommended for who wants to understand why to use Simpati. It focuses on the output information that Simpati provides as pathway-based classifier.

Introduction because:

- It shows a wrapper that allows to run Simpati with one function on human somatic mutation data or RNAseq data
- It focuses on the workflow output and results

Requirements:

- Base R skills

What you will get:

- You will discover the reasons to use Simpati and how to apply it
- The descriptions of all the information and results that Simpati produces

Let's clean and prepare the enviroment for the workflow

We remove every variables, clean the RAM memory and load Simpati library We set the random seed in order to get always the same results out of this workflow We set the number of cores in order to run Simpati in parallel

```
#Clean workspace and memory ----
rm(list=ls())
gc()
#>           used (Mb) gc trigger (Mb) max used (Mb)
#> Ncells  547332 29.3    1252717   67   621331 33.2
#> Vcells 1033267  7.9    8388608   64  1600624 12.3

#Set working directory----
gps0=getwd()
gps0=paste(gps0,"/%s",sep="")
rootDir=gps0
setwd(gsub("%s","",rootDir))
```

```
#Load libraries ----
suppressWarnings(suppressMessages(
  library("Simpati", quietly = T)
  )
)

#Set variables ----
#Set seed for reproduce the results
seed=0
#Set TRUE if you are running a introduction vignette to understand how to work with Simpati
test_run=TRUE
#Set the number of cores to use in the workflow
n_cores=5
```

Simpati works with the patient's biological profiles (e.g. gene expression profiles), the classes of the patients (e.g. cases and controls), a list of pathways and a biological interaction network (e.g. gene-gene interaction network). Simpati is designed to handle multiple biological omics but requires that the type of biological feature (e.g. gene) describing the patients is the same one that composes the pathways and the network which models how the features interact or are associated (e.g. proteins require protein-protein network). In this study, we tested Simpati in the classification of early versus late cancer stage patients.

```
#Get omic-specific patient profiles and their clinical data
geno=tcga_data$LIHC$`LIHC_Mutation-20160128`$assay_df;see(geno)
#>      TCGA-BC-4073 TCGA-BC-A10Z TCGA-CC-5263 TCGA-CC-5264 TCGA-CC-A1HT
#> A1BG            0            0            0            0            0
#> NAT2            0            0            0            0            0
#> ADA             0            0            0            0            0
#> CDH2            0            0            0            0            0
#> AKT3            0            0            0            0            0
info=tcga_data$LIHC$`LIHC_Mutation-20160128`$clin_df;see(info)
#>                patientID patient.vital_status pathologic_stage
#> TCGA-BC-4073 TCGA-BC-4073               ALIVE             LATE
#> TCGA-BC-A10Z TCGA-BC-A10Z                DEAD            EARLY
#> TCGA-CC-5263 TCGA-CC-5263                DEAD             LATE
#> TCGA-CC-5264 TCGA-CC-5264               ALIVE             LATE
#> TCGA-CC-A1HT TCGA-CC-A1HT               ALIVE             LATE
#>              patient.histological_type
#> TCGA-BC-4073  HEPATOCELLULAR CARCINOMA
#> TCGA-BC-A10Z  HEPATOCELLULAR CARCINOMA
#> TCGA-CC-5263  HEPATOCELLULAR CARCINOMA
#> TCGA-CC-5264  HEPATOCELLULAR CARCINOMA
#> TCGA-CC-A1HT  HEPATOCELLULAR CARCINOMA

#Simpati wants the info matrix to be a two column matrix
#patient's names | patient's class (e.g. clinical information)
#Here we select the pathologic_stage of the patient's tumour
info=info[,c("patientID","pathologic_stage")];see(info)
#>                patientID pathologic_stage
#> TCGA-BC-4073 TCGA-BC-4073             LATE
#> TCGA-BC-A10Z TCGA-BC-A10Z            EARLY
#> TCGA-CC-5263 TCGA-CC-5263             LATE
#> TCGA-CC-5264 TCGA-CC-5264             LATE
#> TCGA-CC-A1HT TCGA-CC-A1HT             LATE
```

```
#Set name of the dataset
dataset_name="LIHC"
#Set the semantic type of the disease for the disgnet enrichment
disease_type=tcga_data$LIHC$semantic_type;cat(disease_type)
#> Neoplastic Process
#Set key words associated to the patient's disease
key_words=tcga_data$LIHC$key_words;cat(key_words)
#> Liver hepatocellular carcinoma

#Gene interaction network
net=huri_net_l$net_adj;see(net)
#>         TNMD BCL2L13 BNIP3L CD33 HHLA2
#> TNMD       0       1      1    1     1
#> BCL2L13    1       0      0    0     0
#> BNIP3L     1       0      0    0     0
#> CD33       1       0      0    0     0
#> HHLA2      1       0      0    0     0

#Pathway list
print(pathways_l[1:2])
#> $`PROTEIN CITRULLINATION source-HUMANCYC source-PWY-4921`
#> [1] "PADI6" "PADI3" "PADI2" "PADI4" "PADI1"
#>
#> $`METHYLGLYOXAL DEGRADATION III source-HUMANCYC source-PWY-5453`
#> [1] "AKR1B10" "AKR1B1"  "CYP2E1"
```

Simpati considers the patient's biological profiles (e.g. genes per patients) divided into classes based on a clinical information (e.g. cases versus controls). It prepares the profiles singularly applying guilty-by-association approach to determine how much each biological feature is associated and involved with the other ones and so to the overall patient's profile. Higher is the guilty score and more the biological feature is involved in the patient's biology. Simpati proceeds by building a pathway-specific patient similarity network (psPSN). It determines how much each pair of patients is similarly involved in the pathway. If the members of one class are more similar (i.e. stronger intra-similarities) than the opposite patients and the two classes are not similar (i.e. weak inter-similarities), then Simpati recognizes the psPSN as signature. If the classes are likely to contain outlier patients (i.e. patients not showing the same pathway activity as the rest of the class), then Simpati performs a filtering to keep only the biggest and most representative subgroups and re-test the psPSN for being signature. Unknown patients are classified in the best pathways based on their similarities with known patients and on how much they fit in the representative subgroups of the classes (more you are friend with the leader of one group and more you are associated to that). As results, Simpati provides the classes of the unknown patients, the tested statistically significant signature pathways divided into up and down involved (new pathway activity paradigm based on similarity of propagation scores), the biological features which contributed the most to the similarities of interest, the guilty scores associated to the biological features and all the data produced during the workflow in a vectorial format easy to share or analyse.

```
#Simpati classification
Simp_res=wrapper_human_mutations(geno,info,net,pathways_l,dataset_name,disease_type,key_words,
                                 n_cores=n_cores,test_run=test_run,seed=seed)
#> *Data preparation
#> >>There are 8784 genes matching between network and profiles
#> >>There are 6147 genes matching between pathways and profiles
#> >>Patient's genetic profiles and clinical data match
#> *Input data are well formatted
```

```
#> >>Class names: EARLY LATE
#> >>Class sizes: 7 7
#> *Variables ready
#> *Performing propagation
#> >>The following number of profile's genes are missing as nodes: 14145
#> >>Row normalization of the network
#> >>Parallel for windows
#> >>Finished in: 0.2385 mins
#> *Cleaning and mapping pathway specific gene sets
#> >>Parallel for windows
#> >>Pathway list ready to use with: 500 sets
#> *Making pathway specific predictions of testing profiles
#> Parallel for windows
#> auroc: 0.75  aupr: 0.8333333
#> Ended classification in 2.144532
#> Starting pathway enrichment with classification results
#> *Data preparation
#> >>There are 8784 genes matching between network and profiles
#> >>There are 3922 genes matching between pathways and profiles
#> >>Patient's genetic profiles and clinical data match
#> *Input data are well formatted
#> >>Class names: EARLY LATE
#> >>Class sizes: 6 8
#> *Variables ready
#> *Performing propagation
#> >>The following number of profile's genes are missing as nodes: 14145
#> >>Row normalization of the network
#> >>Parallel for windows
#> >>Finished in: 0.2405 mins
#> *Cleaning and mapping pathway specific gene sets
#> >>Parallel for windows
#> >>Pathway list ready to use with: 71 sets
#> Parallel for windows
#> Preparing the final results
#> Parallel for windows
#> Saving results and workflow data
```

Simpati provides the classification performances, collects the signature pathways used to predict, returns their corresponding PSNs in vectorial format and reports their related information to allow further analysis and considerations: the average of the intra and inter similarities to let understanding which is the most cohesive class, the psPSN power translated into a scale from 1 (poor separation between classes) to 10 (strong separation) to catch the pathways which most distinguish the classes in comparison, and a probability value (p.value). The latter is assessed testing the psPSN to retrieve the same original power or higher when patients are permutated between classes. This information allows to filter out pathways which have been detected as signature due to random.

classification_res: list which provides the classification performances

- testing_prof: Includes the testing patients that have been classified
- correct_classes: Includes the correct and original class of the testing patients
- predicted_classes: Includes the predicted class of the testing patients
- auroc: Includes the auroc performance measure
- aupr: Includes the aupr performance measure

```
Simp_res$classification_res
#> $testing_prof
#> [1] "E1" "L4" "L1" "E6"
#>
#> $correct_classes
#> [1] "E" "L" "L" "E"
#>
#> $predicted_classes
#> [1] "E" "L" "L" "L"
#>
#> $auroc
#> [1] 0.75
#>
#> $aupr
#> [1] 0.8333333
```

PSN_enr_df: matrix with the details of the enriched pathway-specific patient similarity networks found after the classification

- pathway_name: name of the pathway
- sign_class: name of the class which is up-involved signature (strongest one) in the PSN representing the pathway
- power: POWER of the pathway-specific patient similarity network
- direction: up-involved (the members of the signature class are similar in the features of the pathway due to
- they all have an high involvment of the features, while the non-signature class is heterogenous and not cohesive) or down-involved (the members of the signature class are similar in the features of the pathway due to they all have a low involvment of the features, while the non-signature class is heterogenous and not cohesive)
- Pvalue: probability value produced by testing the significativity of the pathway-specific PSN. Lower than 0.05 means that the PSN is signature not due to random
- records_db_SemType: number of records in the disgnet database which associate the pathway to the user-defined disease type of the patient's
- records_db_KeysDis: percentage of user-defined key words associated to the pathway by the disgnet database
- records_db_cancer: percentage of features (e.g. genes) of the pathway which are associated to the patient's cancer by the human protein atlas

```
head(Simp_res$PSN_enr_df)
#>                                                                                pathway_name
#> 1                          BIOCARTA_IL6_PATHWAY source-MSIGDB_C2 source-BIOCARTA_IL6_PATHWAY up-inv
#> 2                        BIOCARTA_MAPK_PATHWAY source-MSIGDB_C2 source-BIOCARTA_MAPK_PATHWAY down-inv
#> 3                        HALLMARK_ADIPOGENESIS source-MSIGDB_C2 source-HALLMARK_ADIPOGENESIS down-inv
#> 4          HALLMARK_ALLOGRAFT_REJECTION source-MSIGDB_C2 source-HALLMARK_ALLOGRAFT_REJECTION down-inv
#> 5 HALLMARK_CHOLESTEROL_HOMEOSTASIS source-MSIGDB_C2 source-HALLMARK_CHOLESTEROL_HOMEOSTASIS down-inv
#> 6      HALLMARK_FATTY_ACID_METABOLISM source-MSIGDB_C2 source-HALLMARK_FATTY_ACID_METABOLISM down-inv
#>   sign_class power direction Pvalue records_db_SemType records_db_KeysDis
#> 1          E     9    up-inv  0.035               1015          100.00000
#> 2          E    10  down-inv  0.005                624           66.66667
#> 3          L    10  down-inv  0.045                406          100.00000
#> 4          E    10  down-inv  0.005               2186          100.00000
#> 5          E    10  down-inv  0.005                122           66.66667
```

```
#> 6          E     9  down-inv  0.025                273          100.00000
#>   records_db_cancer records_db_normal
#> 1         100.00000                 0
#> 2         100.00000                 0
#> 3         100.00000                 0
#> 4         100.00000                 0
#> 5         100.00000                 0
#> 6          85.71429                 0
```

PSNs_info[,1:6]: matrix which describes the pathway specific patient similarity networks learnt during the classification and used for the prediction of the testing patient's class

- pathway_name: name of the pathway
- class: name of the class which is up-involved signature (strongest one) in the PSN representing the pathway
- power: POWER of the pathway-specific patient similarity network
- minSIGN: low quantile of the signature class which is higher than the high quantile of the opposite class and of the interclass similarities
- maxWEAK1: high quantile of the non-signature class which is lower than the low quantile of the signature class
- maxWEAK2: high quantile of the interclass which is lower than the low quantile of the signature class

```
Simp_res$PSNs_info[1:5,1:6]
#>                                                                  pathway_name
#> 1                      PID_E2F_PATHWAY source-MSIGDB_C2 source-PID_E2F_PATHWAY
#> 2                      PID_E2F_PATHWAY source-MSIGDB_C2 source-PID_E2F_PATHWAY
#> 3                      PID_E2F_PATHWAY source-MSIGDB_C2 source-PID_E2F_PATHWAY
#> 4 PID_P53_DOWNSTREAM_PATHWAY source-MSIGDB_C2 source-PID_P53_DOWNSTREAM_PATHWAY
#> 5 PID_P53_DOWNSTREAM_PATHWAY source-MSIGDB_C2 source-PID_P53_DOWNSTREAM_PATHWAY
#>   class power   minSIGN  maxWEAK1  maxWEAK2
#> 1     L     3 0.4198617 0.3152399 0.4055884
#> 2     L     4 0.3285012 0.2969505 0.3240869
#> 3     L     4 0.4455857 0.3664509 0.4415346
#> 4     L     8 0.4389375 0.3182257 0.4363828
#> 5     L     9 0.4738674 0.4685179 0.4593195
```

PSNs_info[,c(1,2,7,10,16,17,21,22)]: matrix which describes how a pathway specific patient similarity network predicts a testing patient

- This section of the matrix tackles the black box effect of the algorithm.
- The first pathway PID_E2F_PATHWAY is a signature PSN for the L class.
- Stop: percentage of rapresentative patients of the Signature class in which the testing patient is similar to (lower and the better)
- Wtop: percentage of rapresentative patients of the non-Signature class in which the testing patient is similar to (lower and the better)
- A_strength: similarity of the testing patient with the members of the Signature class
- B_strength: similarity of the testing patient with the members of the non-Signature class
- testing_pat: name of the testing patientTh
- prediction: predicted class based on Stop, Wtop, A_strength and B_strength In this case, the testing patient E1 is predicted L because Stop is lower than Wtop and A_strength is greater than B_strength

```
Simp_res$PSNs_info[2,c(1,2,7,11,16,17,21,22)]
#>                                                pathway_name class Stop Wtop
#> 2 PID_E2F_PATHWAY source-MSIGDB_C2 source-PID_E2F_PATHWAY     L  0.8    1
#>   A_strength B_strength testing_pat prediction
#> 2     0.6714     0.7002          L4          L
```

vars_l: list that allows you to access to the data and variables used during the classification

- orig_geno: the original user-provided matrix of patient's profiles
- geno: the matrix of patient's profiles normalized and processed to use in Simpati workflow
- net: the original feature association network
- pathways_l: the original pathways (aka feature sets)
- info: the original info matrix with the patient'ids and classes mapped to the new labels
- tab_status: the classes compared in Simpati workflow

```
Simp_res$vars_l$info
#>                 patientID pathologic_stage shortID
#> TCGA-BC-A10Z TCGA-BC-A10Z            EARLY      E1
#> TCGA-DD-A114 TCGA-DD-A114            EARLY      E2
#> TCGA-DD-A118 TCGA-DD-A118            EARLY      E3
#> TCGA-DD-A11B TCGA-DD-A11B            EARLY      E4
#> TCGA-DD-A11D TCGA-DD-A11D            EARLY      E5
#> TCGA-DD-A1EB TCGA-DD-A1EB            EARLY      E6
#> TCGA-DD-A1EC TCGA-DD-A1EC            EARLY      E7
#> TCGA-BC-4073 TCGA-BC-4073             LATE      L1
#> TCGA-CC-5263 TCGA-CC-5263             LATE      L2
#> TCGA-CC-5264 TCGA-CC-5264             LATE      L3
#> TCGA-CC-A1HT TCGA-CC-A1HT             LATE      L4
#> TCGA-DD-A116 TCGA-DD-A116             LATE      L5
#> TCGA-DD-A119 TCGA-DD-A119             LATE      L6
#> TCGA-DD-A1EH TCGA-DD-A1EH             LATE      L7
```

PSN_data_l$outlier_df: matrix which indicates the likelihood of each patient to be outlier for its class

```
Simp_res$PSN_data_l$outlier_df
#>    Var1      Freq
#> 1    E1 48.484848
#> 7    L1 39.393939
#> 13   L7 39.393939
#> 8    L2 30.303030
#> 9    L3 27.272727
#> 2    E2 21.212121
#> 6    E6 21.212121
#> 11   L5 21.212121
#> 10   L4 18.181818
#> 3    E3 15.151515
#> 12   L6 15.151515
#> 5    E5 12.121212
#> 14   L8 12.121212
#> 4    E4  6.060606
```

PSN__data__l$PSN__comp__l: list in which each element includes the vectorized form of a pathway-specific PSN These data are handy to plot the PSN of interest or to analyse manually the pathway-specific PSN

```
head(Simp_res$PSN_data_l$PSN_comp_l$`PID_RB_1PATHWAY source-MSIGDB_C2 source-PID_RB_1PATHWAY down-inv`$
#> [1] 1.0000000 0.9960091 0.9979586 0.9972884 0.9970337 0.9960098
head(Simp_res$PSN_data_l$PSN_comp_l$`PID_RB_1PATHWAY source-MSIGDB_C2 source-PID_RB_1PATHWAY down-inv`$
#> [1] "E1" "E2" "E3" "E4" "E5" "E6"
#You can convert the vectorized form of a PSN to get its adjacency matrix and plot it or elaborate it
#Let's take the name of the most powerful signature PSN
pathway_name=Simp_res$PSN_enr_df$pathway_name[order(Simp_res$PSN_enr_df$power,decreasing = T)][1]
#Take its vector
pathway_data=Simp_res$PSN_data_l$PSN_comp_l[pathway_name]
pathway_PSN_v=pathway_data[[1]][["m_sim_l"]]
#Convert it to adjancency matrix
pathway_PSN_m=vec2m(pathway_PSN_v)
see(pathway_PSN_m)
#>           E1        E2        E3        E4        E5
#> E1 1.0000000 0.9796465 0.9887328 0.9819582 0.9892160
#> E2 0.9796465 1.0000000 0.9916920 0.9853584 0.9928535
#> E3 0.9887328 0.9916920 1.0000000 0.9916173 0.9984636
#> E4 0.9819582 0.9853584 0.9916173 1.0000000 0.9927618
#> E5 0.9892160 0.9928535 0.9984636 0.9927618 1.0000000
#Plot it
plot_network(pathway_PSN_m,image_name=pathway_name)
#>
#>  I work on the first class of patients
#>
#>  I work on the second class of patients
#> Plotting
#> Registered S3 methods overwritten by 'huge':
#>   method    from
#>   plot.sim  BDgraph
#>   print.sim BDgraph
#> Output stored in C:/Users/norm/Documents/R/PHOENIX/package/Simpati/BIOCARTA_MAPK_PATHWAY source-MSIG
```

‘‘‘