Contextualization of Boolean models using omics data

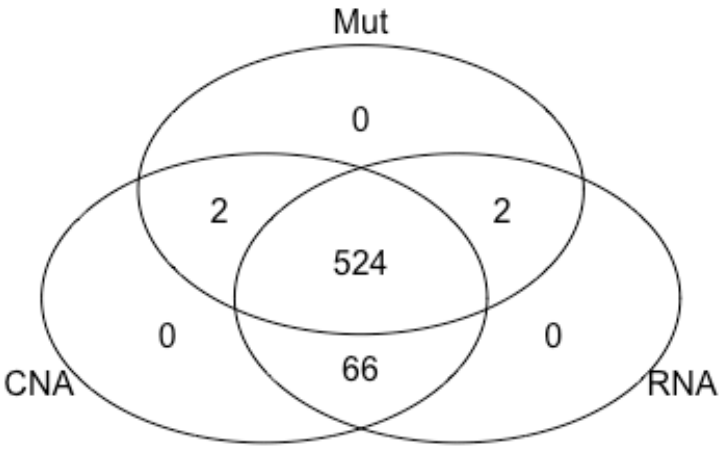
With a set of TCGA colorectal cancer data and a Boolean model of the early steps of metastasis, we will show the different steps to follow to provide patient-specific models on which personalized medicine can be applied.

1. The inputs

Select the data

The data can be downloaded from the cBioPortal (http://www.cbioportal.org/datasets) under the label *Colorectal Adenocarcinoma (TCGA, PanCancer Atlas)*. We selected only clinical data (data_clinical_patient.txt), the mutation data (data_mutations_extended.txt), CAN data (data_CNA.txt) and RNAseq data (data_RNA_Seq_v2_expression_median.txt). Additional survival information is taken from Table S1 in Liu et al. (DOI : https://doi.org/10.1016/j.cell.2018.02.052, file mmc1.xlsx)

The folder contains 594 samples, among them 524 have the three omics layers (mutations, CAN and RNA levels).



The data is gathered in the folder: **/Data/TCGA_colon/**

Another folder containing reference datasets such as HUGO names (HUGO_Entrez.txt), a correspondence of HUGO and Entrez gene names and the expected effect of the possible variants (allAnnotatedVariants.txt), and 20/20+ pan cancer list of driver genes (2020_pancancer.csv) can be found at the following address: **/Data/Common/**.

Select the model

The model we use for this tutorial is a model of the early steps of metastasis (Cohen et al. 2015, PLoS Comp. Biol.).
It is saved in MaBoSS format (two files: *.BND for the model description and *.CFG for the parameter configuration) at the address: **/Models/Cohen/**
A tab-delimited txt file also needs to be provided where a correspondence between the names used in the model and the HUGO name are specified: Cohen_namesToHugo_curated.txt

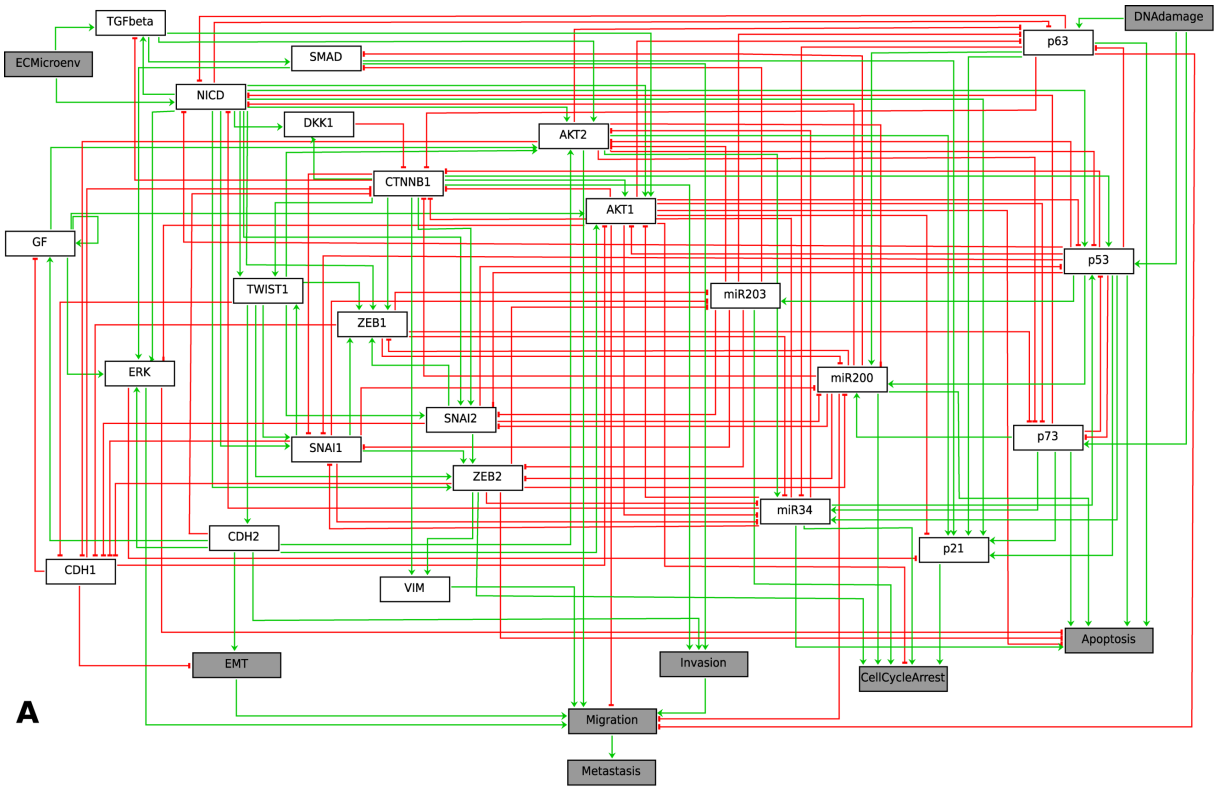


Figure from Cohen DPA, Martignetti L, Robine S, Barillot E, Zinovyev A, Calzone L (2015) Mathematical Modelling of Molecular Pathways Enabling Tumour Cell Invasion and Migration. PLoS Comput Biol 11(11): e1004571.

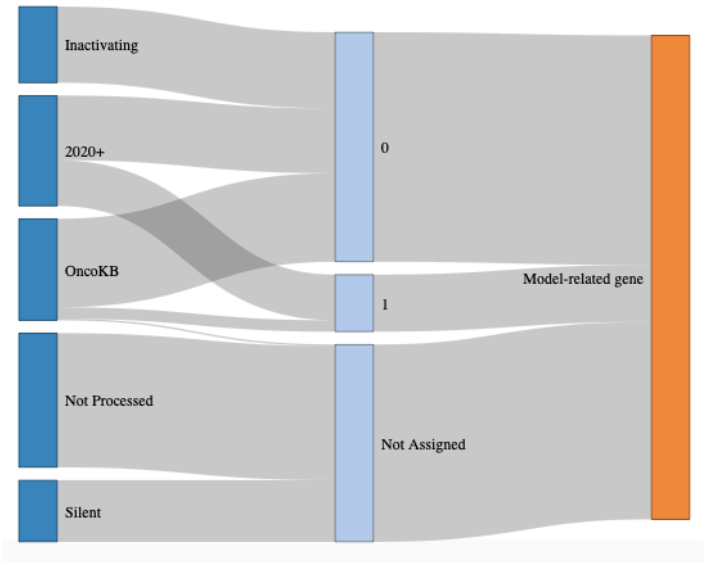
2. Creation of the patient profiles

In Scripts/Profiles open RStudio and run the Rmd file (Cohen_TCGA_colon_profiles.Rmd) It may take some minutes to create the profiles. Please note that analysis can be restrained to model-related genes only to reduce computation time.

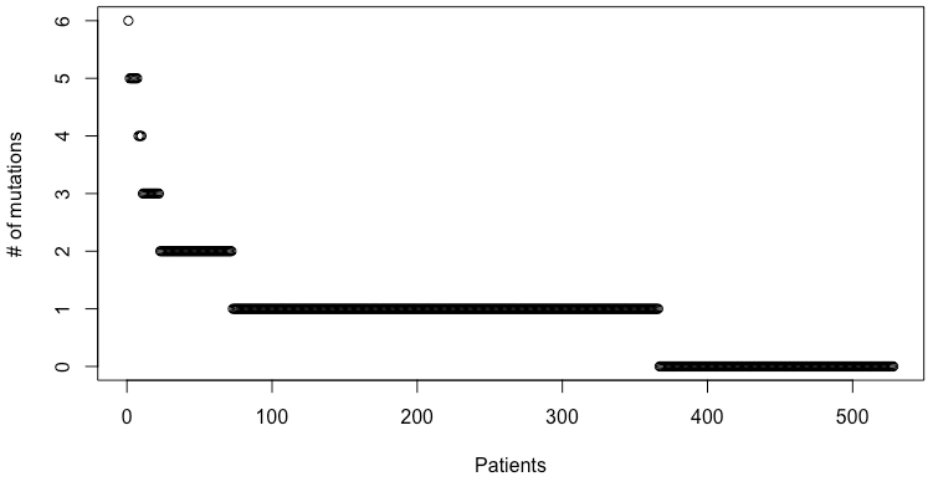
From the data, data tables will be created.

For *mutation profiles*, a Boolean effect will be assigned to each mutation: either 0 (inactivating) or 1 (activating). A mutation can stay unassigned in absence of any evidence.

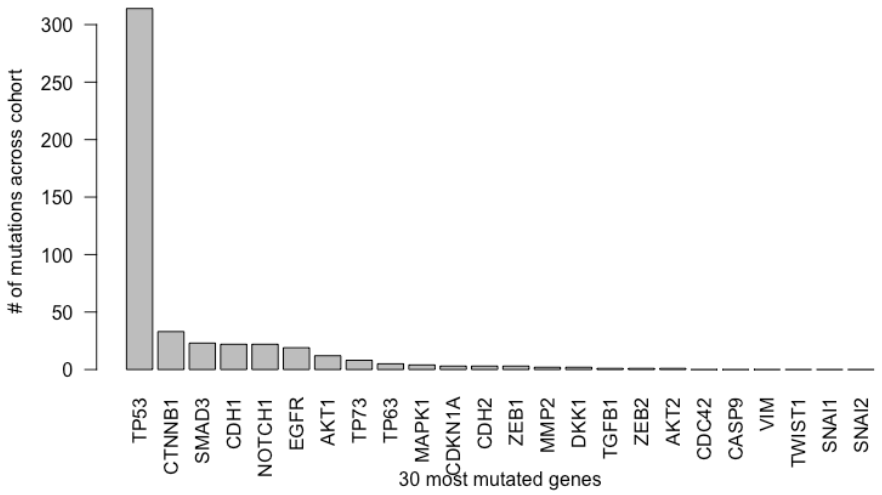
The script shows how many genes were assigned 0s and 1s and how many remain unassigned with a Sankey plot. This is a fast way to verify that the method is not assigning only 0s or 1s. It is also possible to check the distribution of assigned mutations per patient in the cohort.



Distribution of assigned mutations per patient in TCGA cohort

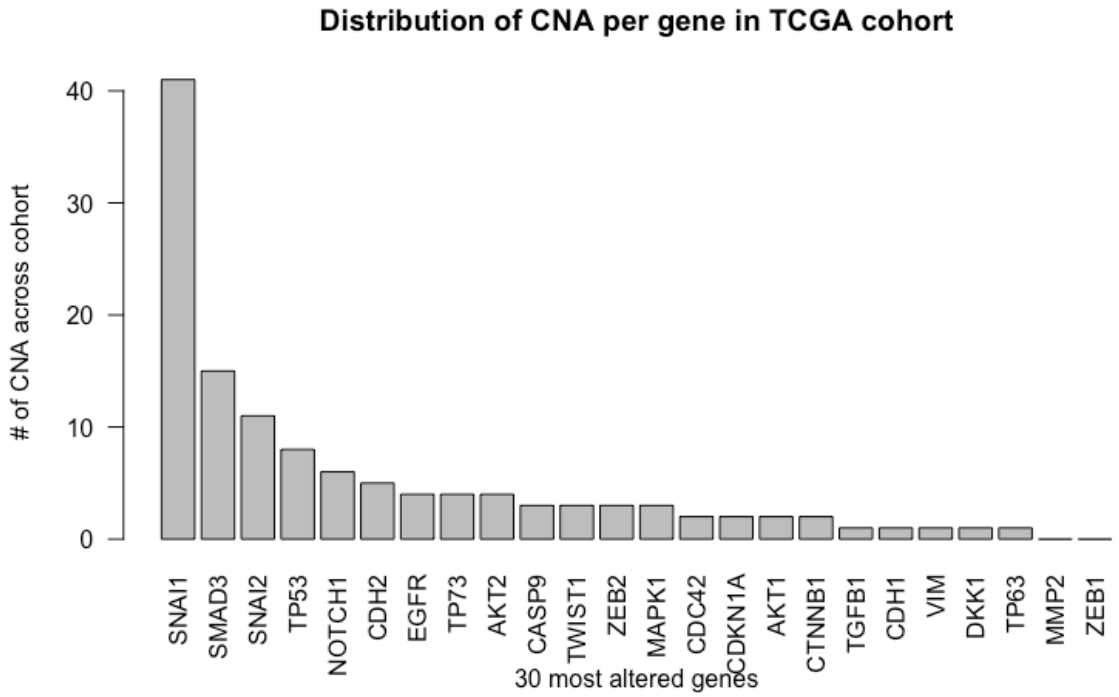


Distribution of mutations per gene in TCGA cohort



Here, we see that a small proportion of patients have 2 or more mutations, and the most mutated gene is TP53.

For the *CNA profiles*, only the amplifications (+2) and the homozygous deletions (-2) are considered here (GISTIC results). The same types of plots are generated showing that SNAI is the most amplified gene and



For the *RNA profiles*, the pipeline is launched to determine the genes that are: discarded, zero-inflated, unimodal or bimodal. The data is either binarized or normalized according to the situation.

The profiles are then exported as csv files and will be the inputs for the simulations of the Boolean models. They are saved in **/Results/Profiles/**
- Cohen_TCGA_colon_mutations.csv (only mutations used per patient)
- Cohen_TCGA_colon_CNA.csv (only CNA used per patient)
- Cohen_TCGA_colon_mutCNA.csv (both mutations and CAN are used per patient)
- Cohen_TCGA_colon_RNA_norm.csv (normalized RNAseq data are used per patient)
- Cohen_TCGA_colon_RNA.csv (binarized RNAseq data are used per patient)

Note that, in the nomenclature of the files, unless explicitly stated with the mention _norm, the data is binarized.

The Rmd, TCGA_colon_simulation_analysis_tuto.Rmd, saves a html file of the outputs in the folder **/Scripts/Profiles/**

3. Simulations of the Boolean models

For each patient, a profile exists, either for the mutations, the CNA, and both, or for RNA seq data. For each of these profiles / patients, a simulation will be performed using the studied Boolean model.

In a terminal, launch the shell command: TCGA_Cohen_shell.sh

In this shell, the model is chosen:
Model=Cohen

The name of the output files is chosen:
sim_case=TCGA_CNA_asMutants_RNA_asTransition

The list of outputs is determined:
list1="Apoptosis,Migration,Invasion,EMT,CellCycleArrest"

Note that many lists can be defined. It is better not to put too many nodes in the same list.

In the following command, the user must choose which methodology to select:
Mutations or/and CNA as mutants (nodes are forced to 0 or 1)
RNAseq normalized data as transitions (transition rates amplified with a factor of 100 (-rf 100)

```
$  python3 Scripts/Simulations/MaBoSS_specific.py $model -sy Mac -p 1
"resultsN_"$sim_case"_"$i".txt" -o $listnodes -s "list"$i"_"$sim_case -m
"Results/Profiles/Cohen_TCGA_colon_CNA.csv" -rb
"Results/Profiles/Cohen_TCGA_colon_RNA_norm.csv" -rf 100
```

The number of simulations correspond to the number of samples with both CNA and RNAseq data. For this case, there are 524 samples at the intersection of both data types.

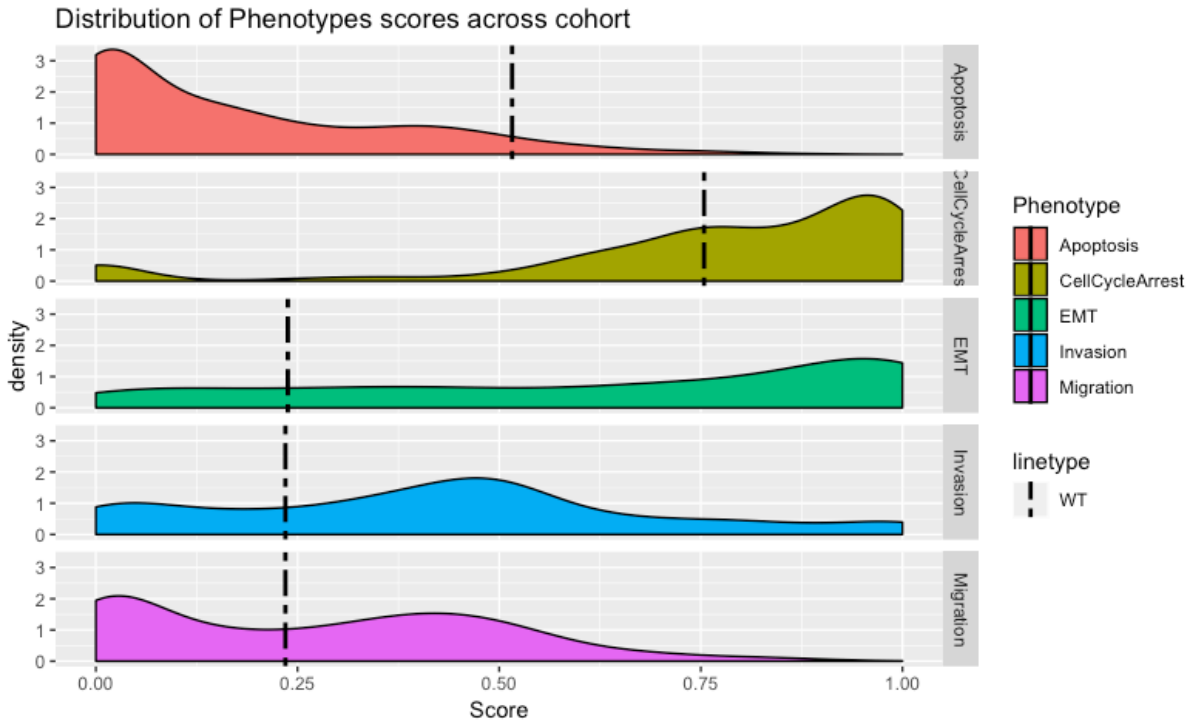
The results are saved in **/Results/Simulations/**

NB: It is very important to check that the correspondence file between nodes of the model and HUGO names are correct. They are case-sensitive!

4. Validation on clinical data

We use the results of the simulations to compare with the available clinical data. Some examples of such analyses can be found in the script:
TCGA_colon_simulation_analysis_tuto.Rmd

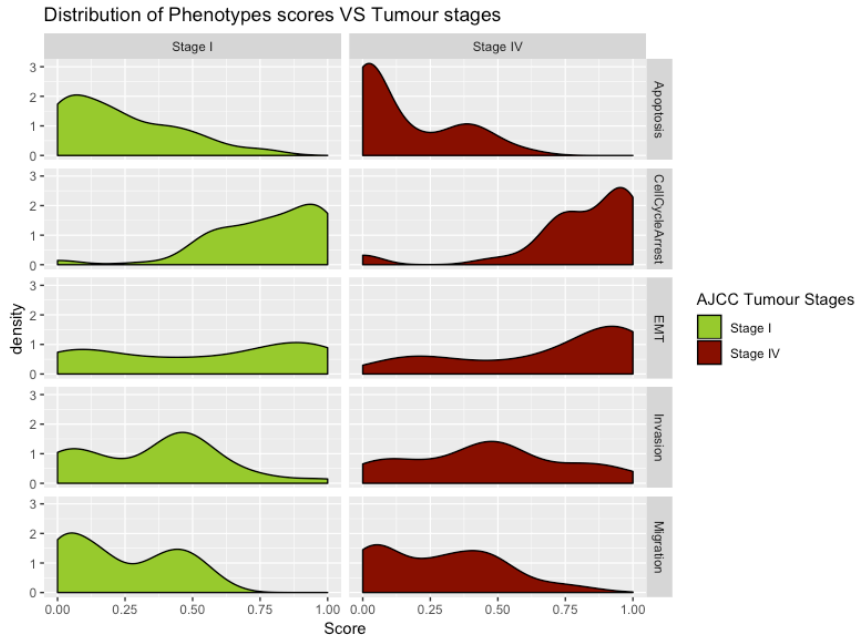
4.1. First, we check the scores of each phenotype (or node) across the cohort. For each of them, the WT score is informed.



The wild type values (non-personalized models) are:

Apoptosis	CellCycleArrest	EMT	Migration	Invasion
0.516	0.754	0.238	0.235	0.235

4.2. We can then perform several data analysis to investigates relations with clinical data, like tumor stages or survival data. Below, we can observe that high-stage tumors appear to be less apoptotic and slightly more prone to EMT. Other phenotypes do not capture significant differences.



…

4.3. When available, the results of the simulations can be compared to existing signatures. For more details, see the example of METABRIC or TCGA data with the model of Fumia et al. in the corresponding paper (https://www.frontiersin.org/articles/10.3389/fphys.2018.01965/full)