Weekly Report

# 07/03/2022

BOUDEMIA ALA EDDINE

TEAM CHEN

# **Goal:**

* Extraction of mutational signatures in breast/bladder cancers using SigProfiler.

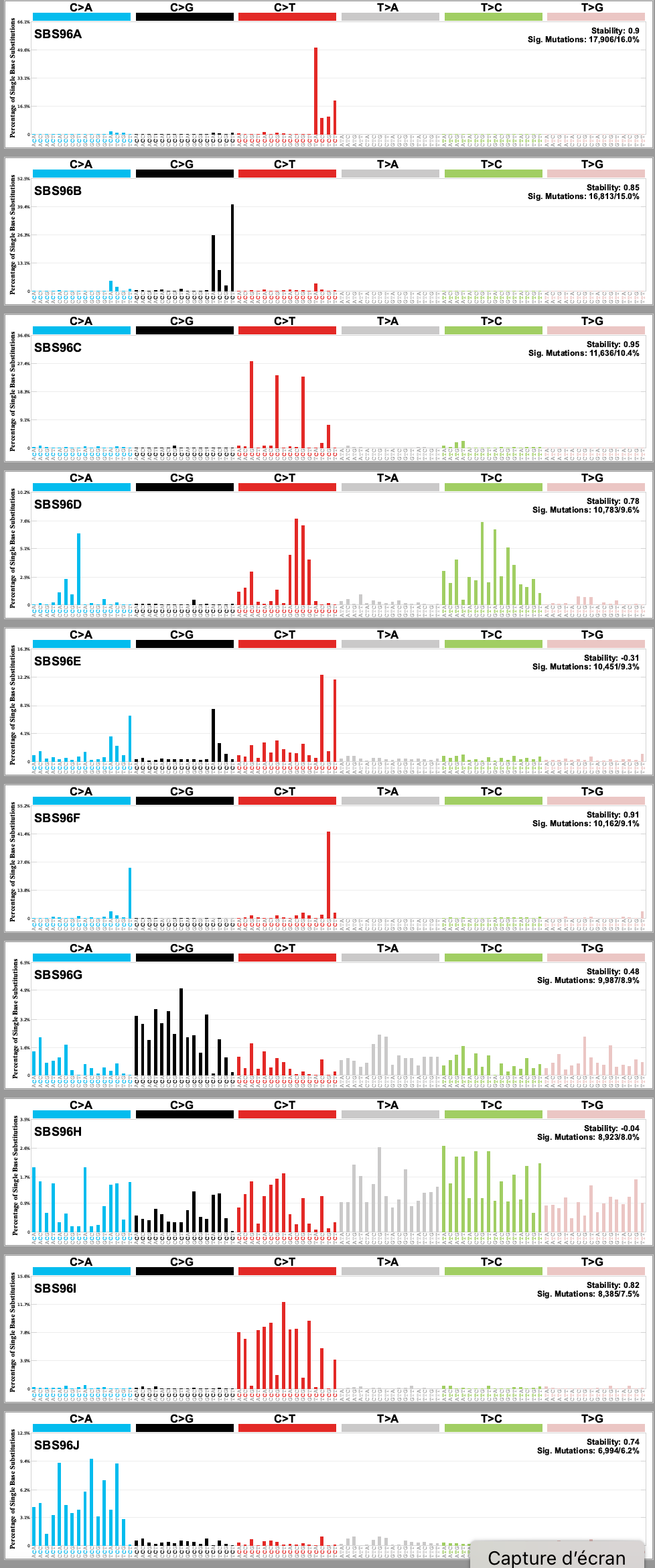
# **Procedure:**

* Download the data from TCGA (BRCA & BLCA) as MAF files. These datasets come from Whole Exome Sequencing.
* Generate mutational spectra using SigProfilerMatrixGenerator (BLCA not yet).
* Extract mutational signatures using SigProfilerExtractor (BRCA running for 1:13 signatures / BLCA will be run for 7 signatures).
* Analysis of the MAF files using maftools[[1]](#footnote-1)
* Map mutational signatures obtained to their genomic regions (find out how to do it first).

# **Results:**

## Mutational Signatures (10 signatures):

* Average stability is 0.61, we can observe SBS1, SBS2, SBS10(a&b), SBS13 and they present a good stability.



## Mutational Summary:

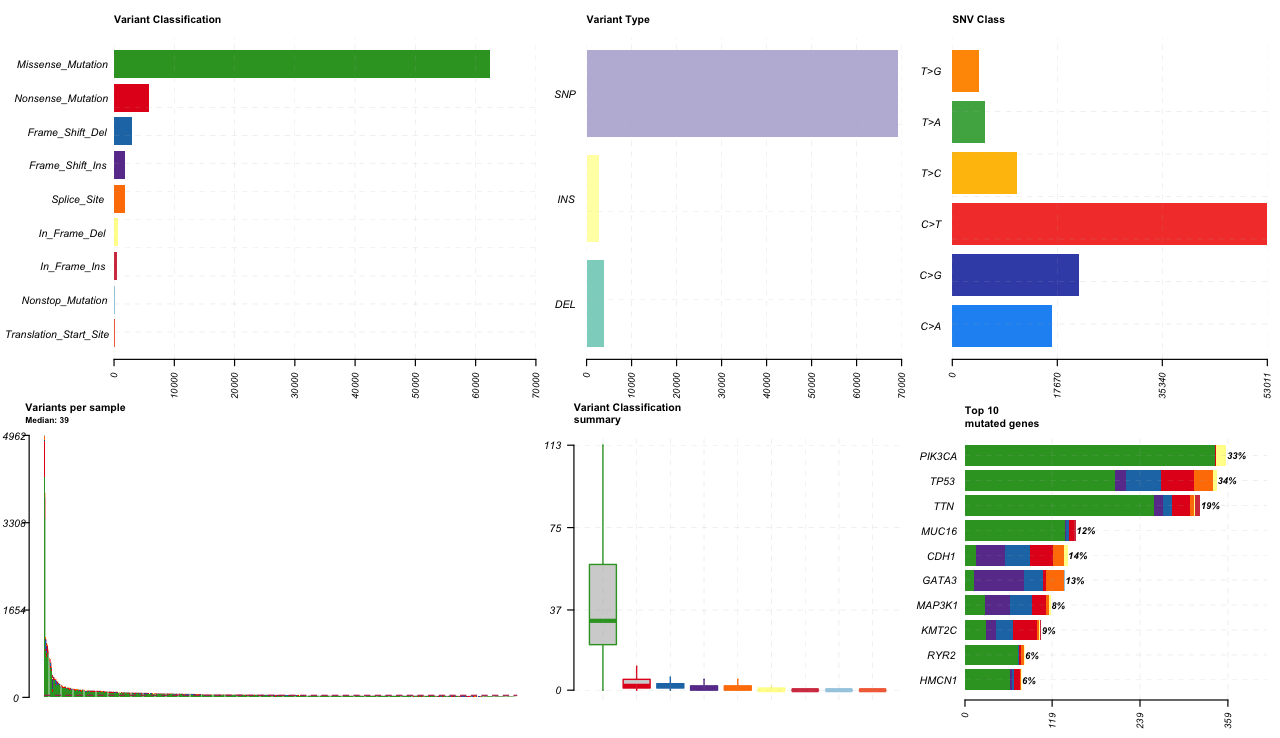


Figure 1 MAF summary of BRCA project

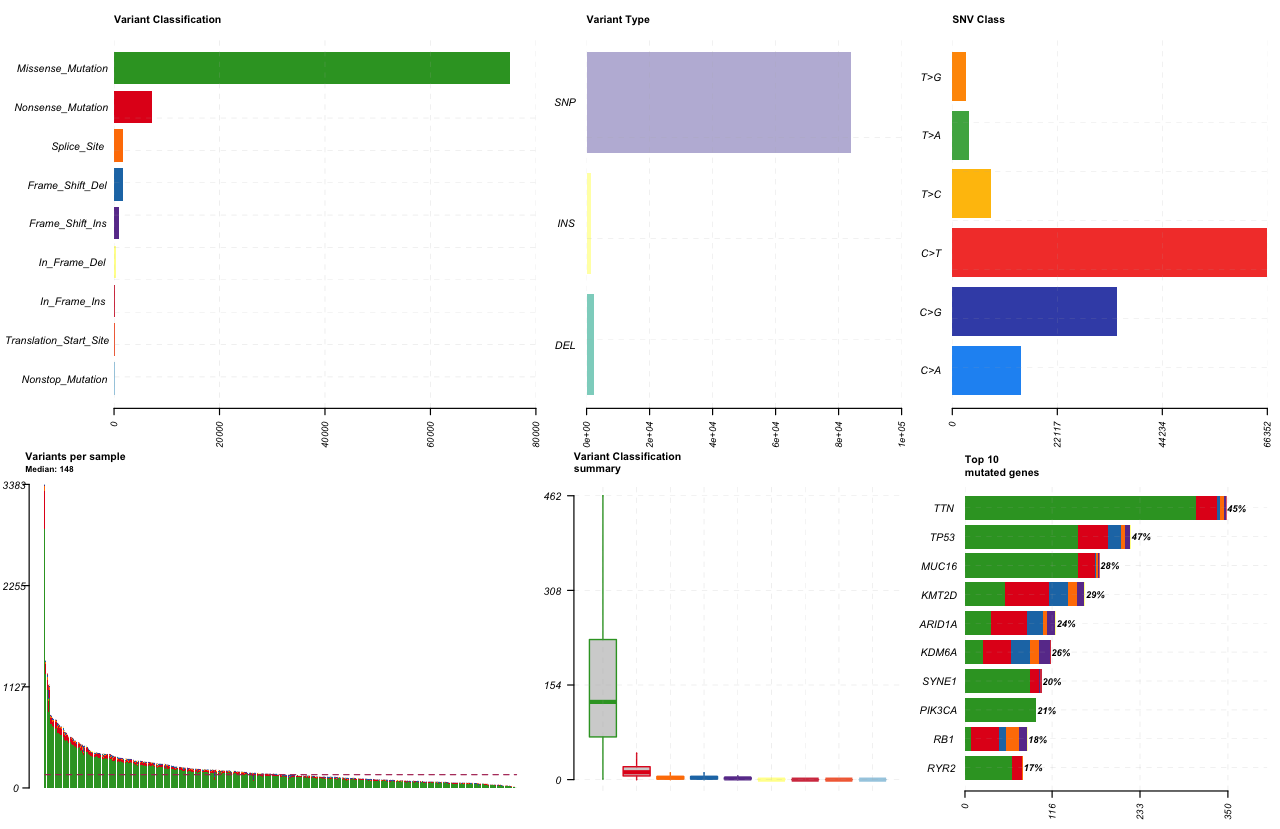


Figure 2 MAF summary of BLCA project

Table 1 Comparison Between BRCA and BLCA mutation types

|  |  |  |
| --- | --- | --- |
| Mutation type | BRCA | BLCA |
| Frame Shift Deletion | 2947 | 1641 |
| Frame Shift Insertion | 1849 | 1002 |
| In Frame Deletion | 574 | 341 |
| In Frame Insertion | 420 | 214 |
| Missense Mutation | 62316 | 75116 |
| Nonsense Mutation | 5734 | 7291 |
| Nonstop Mutation | 86 | 127 |
| Splice Site | 1804 | 1762 |
| Translation Start Site | 81 | 130 |
| Total | 75811 | 87624 |
| Silent Variants | 45177 | 46889 |

Table 2 Statistics from BRCA project

|  |  |
| --- | --- |
|  | BRCA |
| Count | 985 |
| Mean | 112,25 |
| Std | 334,97 |
| Min | 7 |
| 25% | 36 |
| 50% | 57 |
| 75% | 100 |
| Max | 7109 |

## Transition and Transversion Ratios:

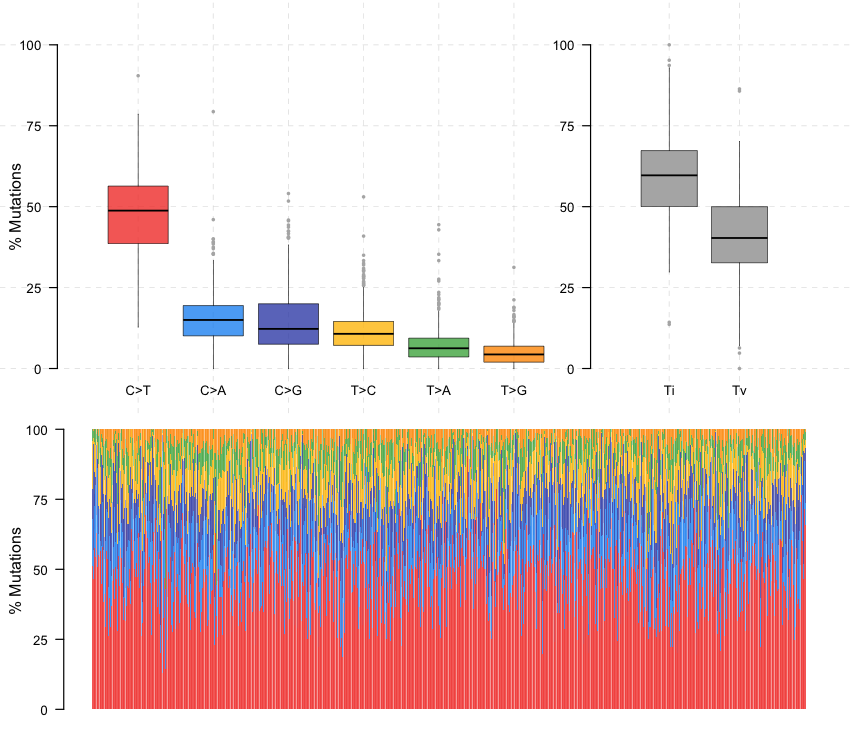


Figure 3 BRCA transition-transversion ratios

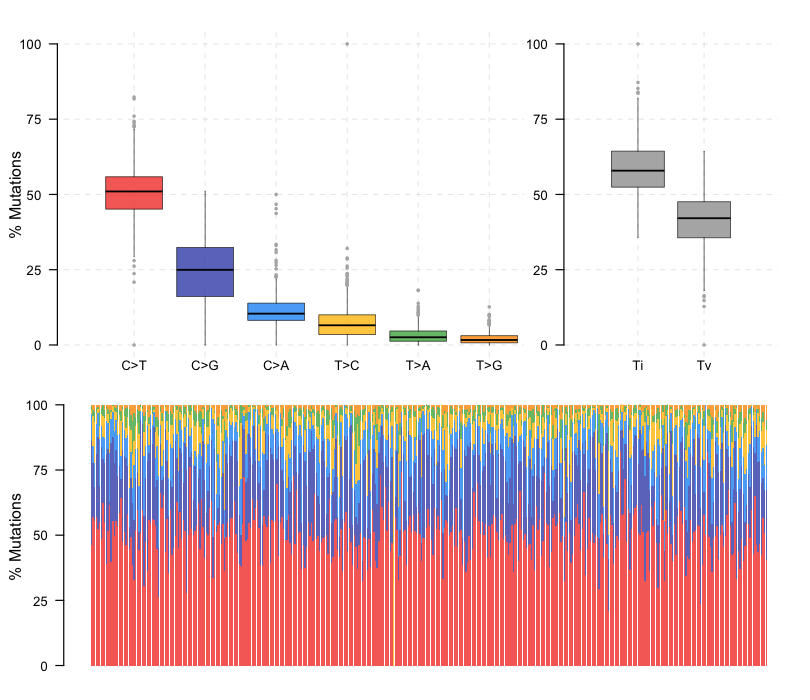


Figure 4 BLCA transition-transversion ratios

## Mutational Load in TCGA datasets:

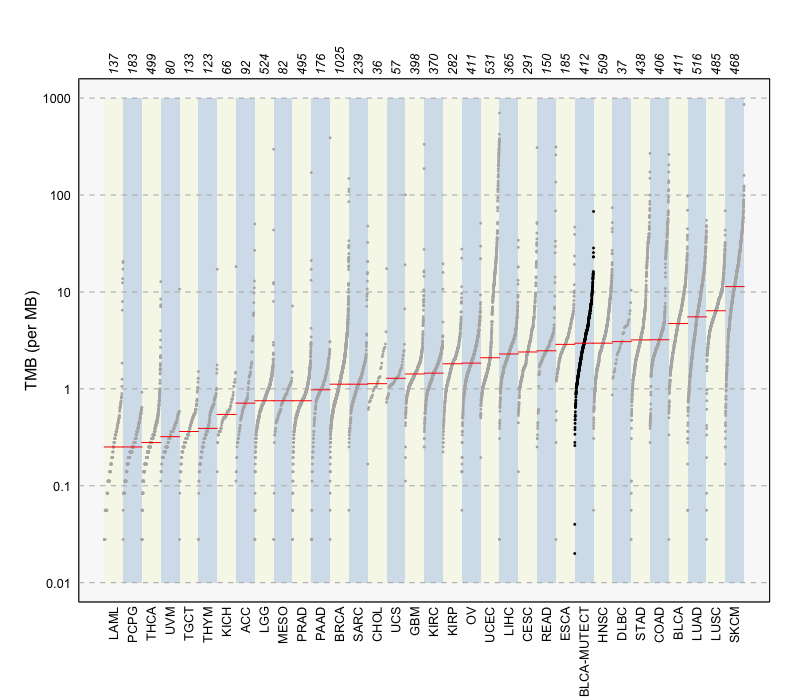


Figure 5 Mutational load

## APOBEC Enrichment:

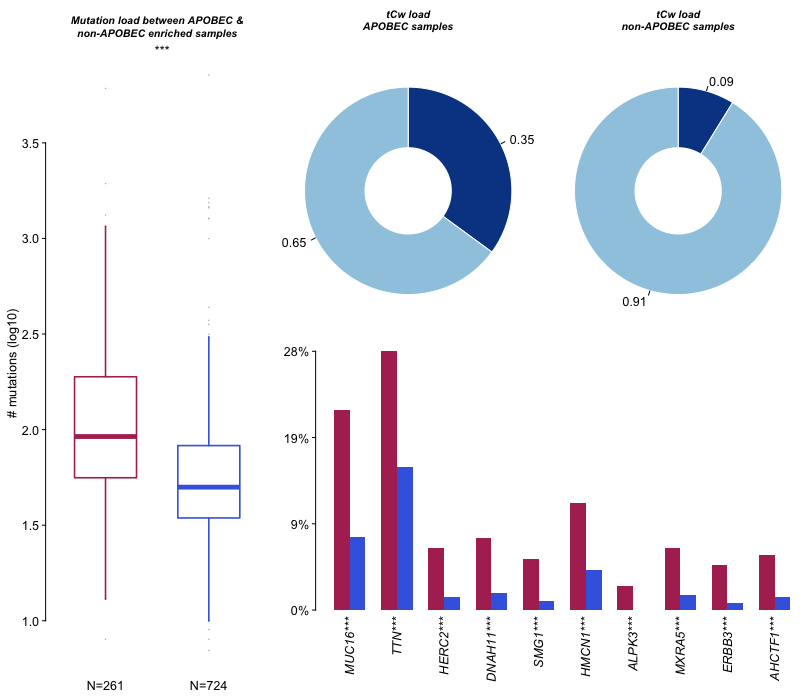


Figure 6 APOBEC Enrichment Analysis in BRCA

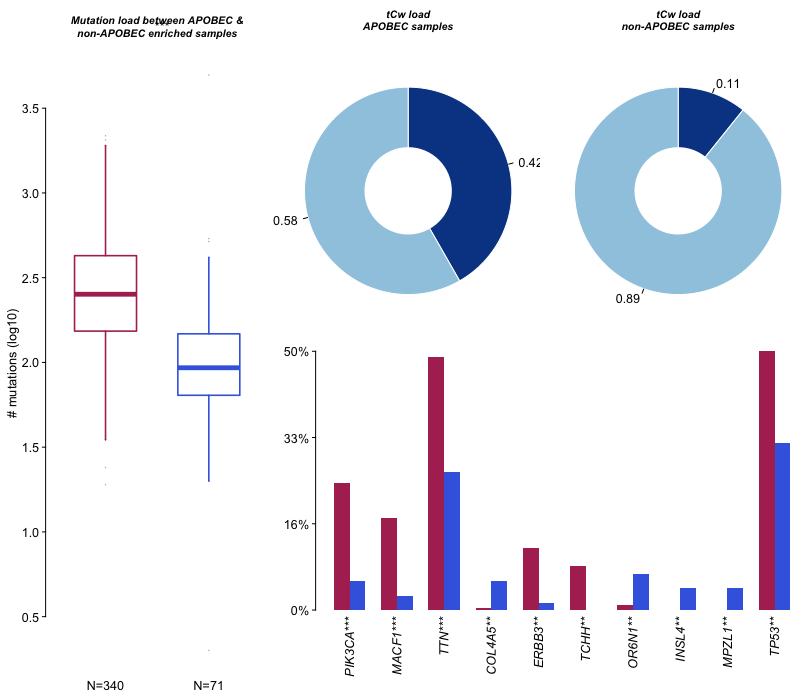


Figure 7 APOBEC Enrichment Analysis in BLCA

The de-novo process of extracting the mutational signatures from BRCA is currently running the 11th process and it should finish it today 11/03/2022. The stability of the process is getting worse at each process, the selection plot won’t be available until it is done. The fact that the stability is worsening could be explained by the fact that the data are extracted from WES therefore some mutational signatures associated with other regions of the genome can’t be extracted but also the average number of mutations in a cancer genome is low. The median is 39 when the outliers are removed and is 57 when they are not removed while the mean is around 100. This value doesn’t allow deciphering a significant number of mutational signatures with a good accuracy as demonstrated by (Alexandrov et al; 2013) and shown in **figure 08 & 09.**

It is easier to extract mutations from few samples with high mutational burden than from a lot of samples with low mutational burden.

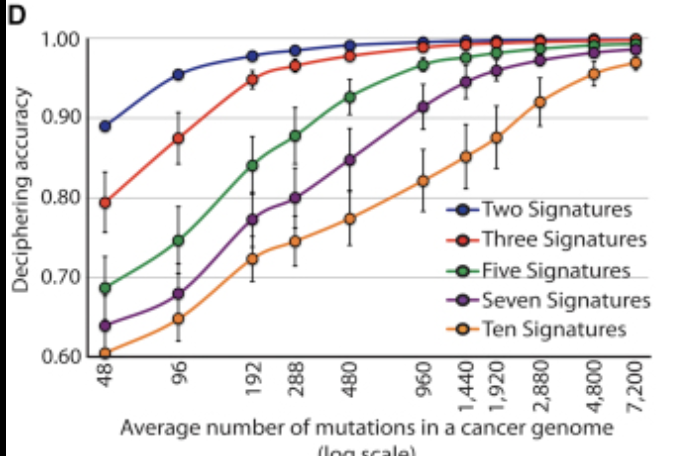


Figure 8 Deciphering accuracy in function of the average number of mutations in a cancer genome

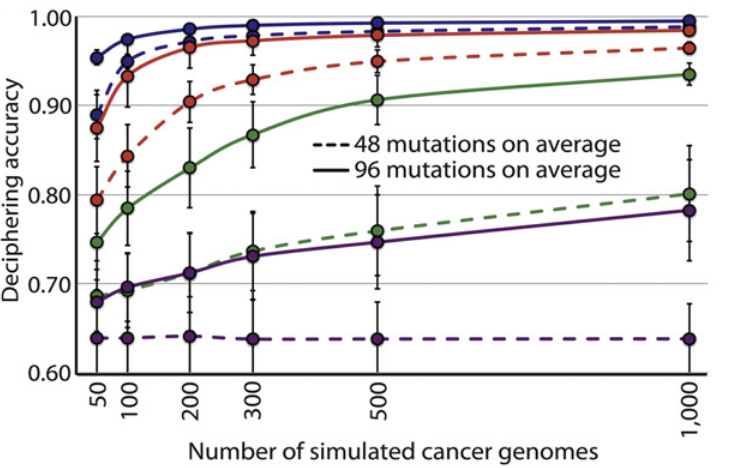


Figure 9 Deciphering accuracy in function of the number of cancer genomes with fixed mutational load

## Mapping signatures to mutations:

* The de-novo extraction of mutational signatures is an NP-hard problem that falls in the category of blind separation problems. Sigprofiler approximates the solution using NMF, however during the process it regroups all the mutations by their type to generate the 96 channel profiles. The latter is necessary for the algorithm to function, but it leads to a loss of information because grouping mutation types will discard their positional information. Similar to de-novo methods, the fitting methods also discard the position of each mutation making it difficult to recognize which signature generated which mutation.
* One way to go is to use the statistical profiles of the signatures as well as their contributions to map the mutations to their signatures, but this method is limited by the fact that several signatures could contribute to the same mutation type, so unless there is a specific context known for the signatures to operate upon it will be impossible to tell if a given mutation in a specific position is related to a given signature or to the other one.
* The other way is to limit the data to specific regions of interest and perform the analysis as many times as the number of regions. The limitation of this method is that the total number of mutations will drop drastically. For example, limiting the BRCA data to only 3’UTR resulted in only 8000 mutations out of 120000. This could be somewhat resolved by using data from different cancer types and especially WGS data but there will always be a problem of the average number of mutations per sample.

## Next week planning:

* SigProfiler should finish running on the BRCA project, therefore I will analyze the results but instead of running SigProfiler again on BLCA project I will use maftools to determine the best number of signatures then run SigProfiler to that exact number of signatures. This should save me a lot of time.
* Because it is hard to map the signatures to the exact mutations or at least to the regions, I will be selecting mutations based on their genomic regions or genes of interest or other specific spots (to be discussed, I need to learn how to map mutations to the genome first).
* I will follow the training for the cluster, it is available on the gitlab

Note: It is possible to check the altered pathways using maftools; If there is any pathway that worth checking ?

1. Mayakonda A, Lin D, Assenov Y, Plass C, Koeffler PH (2018). “Maftools: efficient and comprehensive analysis of somatic variants in cancer.” *Genome Research*. doi: [10.1101/gr.239244.118](https://doi.org/10.1101/gr.239244.118). [↑](#footnote-ref-1)