**Weekly report**

**28/02/2022-04/03/2022**

**BOUDEMIA Ala Eddine**

**Team CHEN**

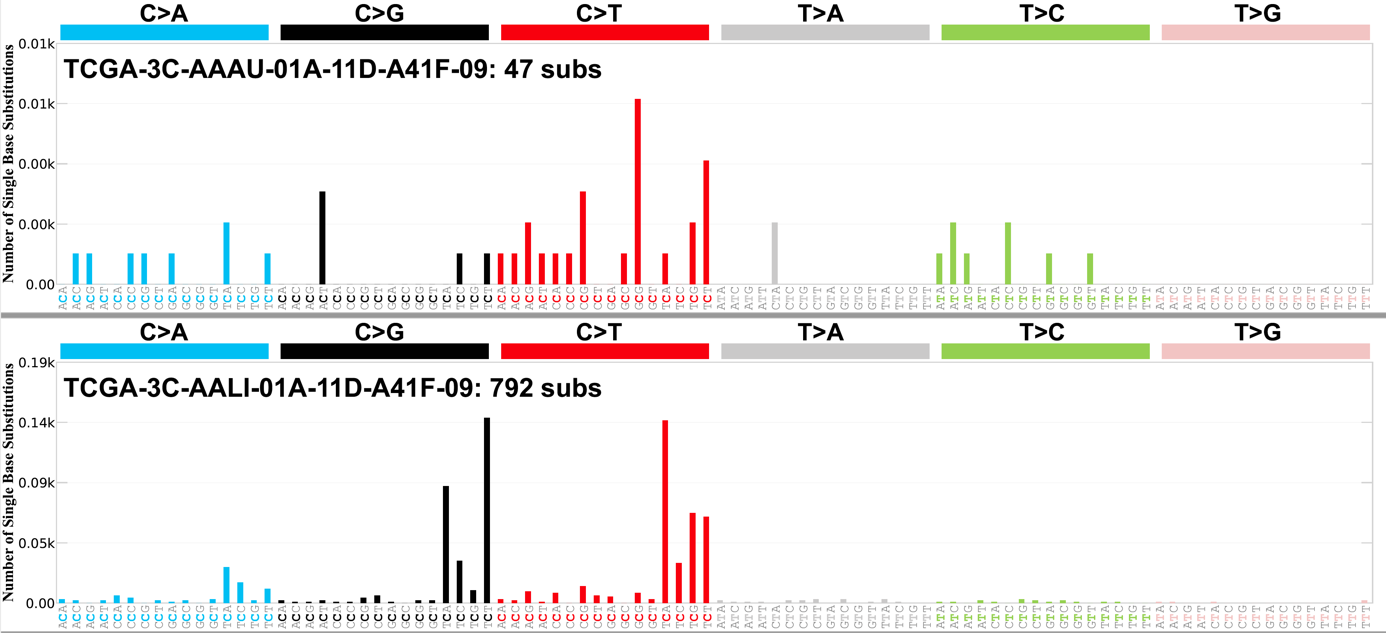
**Objective/Goal** (sereval sentence or one paragraph should be fine)

**Description of work progress:**

The data from TCGA BRCA and TCGA BLCA projects were obtained as a MAF files. Mutation calling was done by MUTECT. <https://portal.gdc.cancer.gov/files/995c0111-d90b-4140-bee7-3845436c3b42>

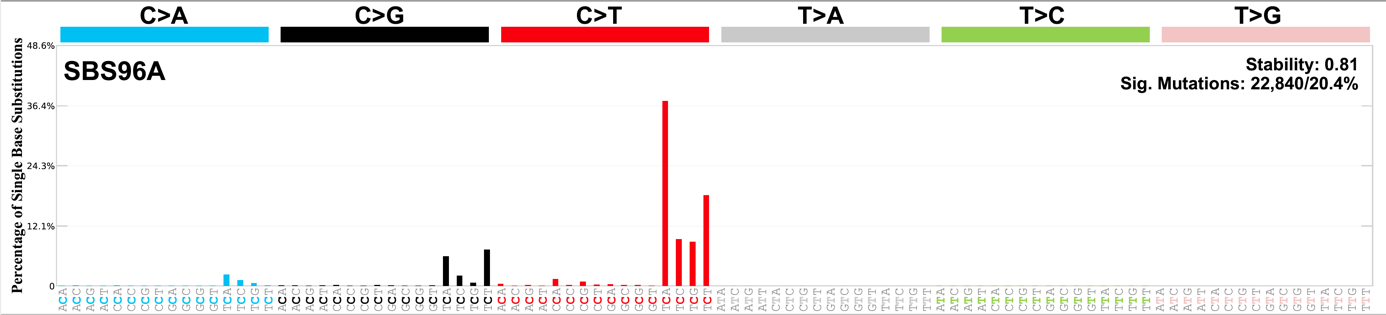
The re-fitting approach requires VCF files, therefore I had to operate with the De-Novo extraction approach. Firstly, I ran SigProfilerMatrixGenerator to create the mutational spectra of each sample giving a total number of 986 mutational catalog for breast cancer. This number is enough to perform the de-novo method as Alexandrov’s paper of 2013 shows that a set of 100 catalogs is enough to decipher 20 signatures.

The SigProfilerMatrixGenerator resulted in the INDEL, DBS, SBS catalogs (all the possible channels) from which I am showing two catalogs from the SBS-96.

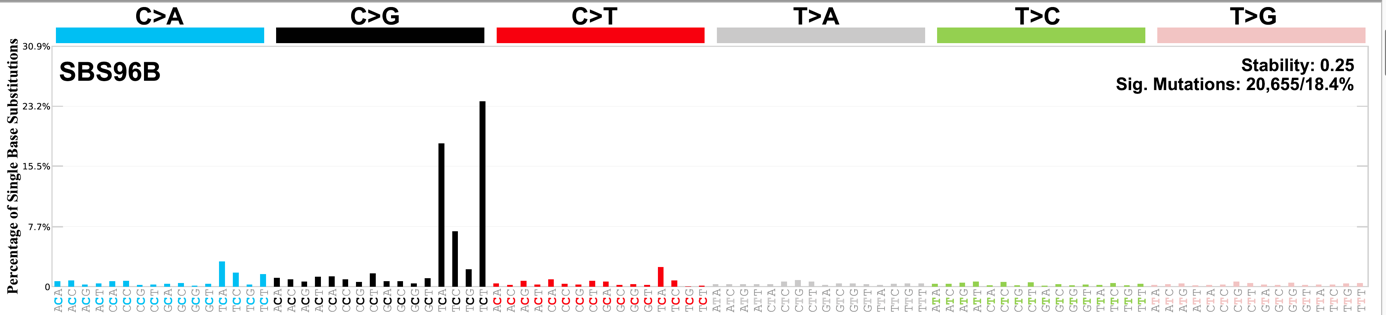


Then, I ran SigProfilerExtractor on the SBS-96 catalogs to extract as minimal as 1 mutational signature and as maximum as 13 signatures. This is justified by the fact that 13 signatures were shown to be associated with breast cancer (11 signatures if we consider Signal instead of COSMIC).

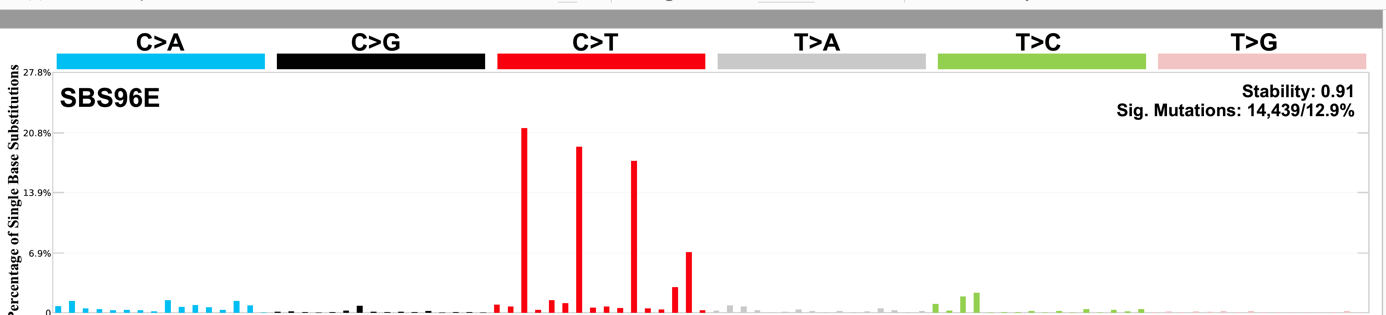
The 8th process is on-going therefore I am showing the results of 7 mutational signatures:



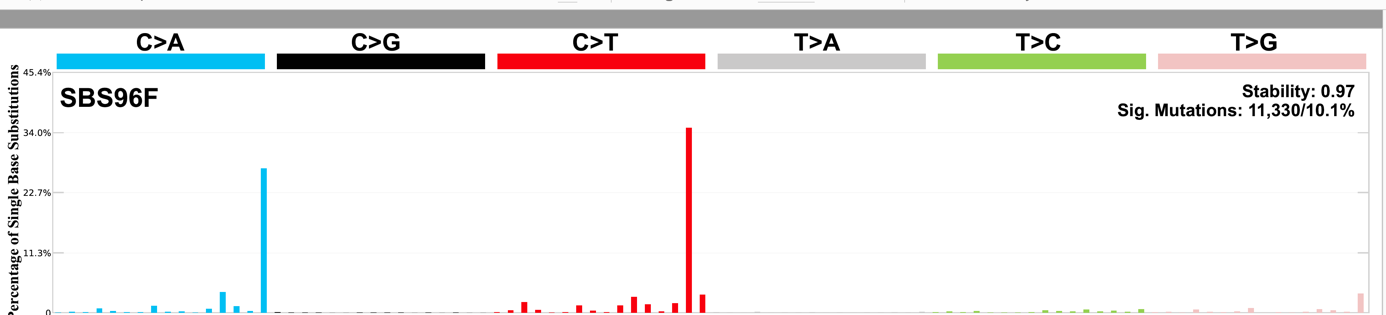
The C>T pattern is like the SBS2 signature.



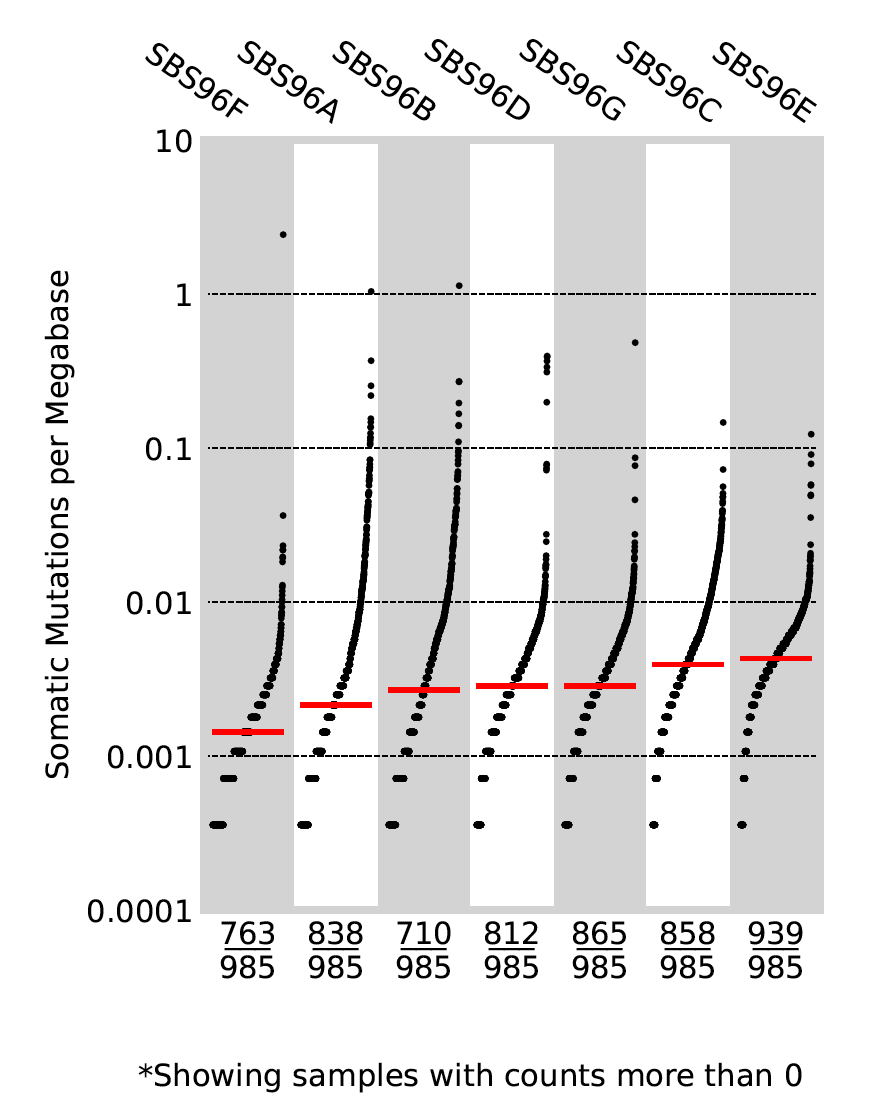
This one is like SBS13 however it lacks reproducibility as the stability is 0.25



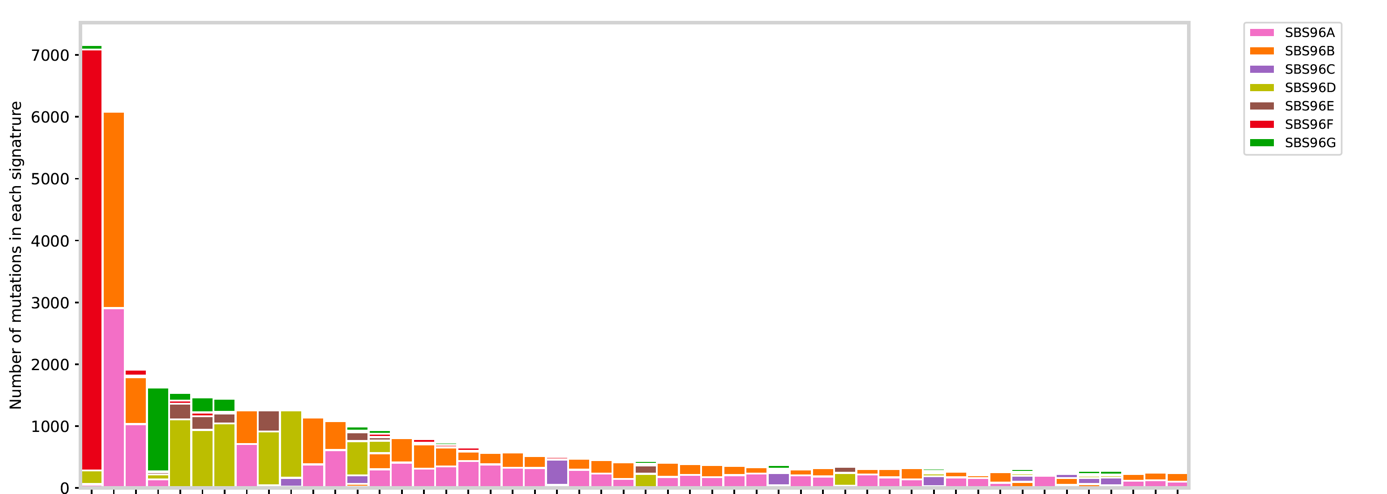
This SBS1 like signature



Finally, this is a combination of SBS10a and SBS10b signatures, I expect that they will deconvolute later.



In here we can observe the activities of each signature, all the extracted signatures are present in most cases. The median of all the 7 signatures is in the same range but this is not reflecting the fact that some samples have very high activities compared to the median, this might be related to the subtype of breast cancer. This is clearer on the next figure that shows the activities of each signature on a sample level



It shows that the first sample has a high mutational burden mainly contributed by SBS96F which is the equivalent of SBS10, but it is mainly present in only one sample. SBS96B however is present in a good number of samples, and it is the equivalent of SBS13 that is attributed to the activity of APOBEC (Note this is one figure out of 20).

**Working planning:**

Next week should be about running SigProfiler on the bladder cancer data and exploring maftools as it seems interesting and provides a lot of methods to analyze both the MAF files and the mutational signatures and it also has an APOBEC enrichment analysis and some other interesting methods. Also, it could be interesting to check Helmsman package for its ability to run in parallel and to handle large datasets because SigProfiler is slow and requires a lot of time to run on the 986 samples of breast cancer. Besides that, I will try to look for a way to assign mutational signatures to genomic locations.