Weekly Report

# 21/03/2022

BOUDEMIA ALA EDDINE

TEAM CHEN

# Goal:

* Extracting mutational signatures acting in breast cancer genomes from TSS, TTS and Coding regions.

# Procedure:

## Extract TSS and TTS regions:

* Download GTF annotation from ENSEMBL.
* Define the set of chromosomes to work on (1:23 + X and Y).
* Read the GTF file line by line and search for the **Feature** and **Chromosome** columns.
* If the chromosome is defined and the feature is a transcript, then extract the following information:
  + The strand (+/-); The gene ID; The transcript ID; The Transcript starting and ending positions.
* If the strand is “+” then the TSS starting position will be the transcript starting position mince a predefined value (I used 300) and the ending point of this TSS will be the transcript starting position plus the predefined value. Conversely, the TTS will be defined using the transcript ending position.
* If the strand is “-” then the procedure will be inversed. The TSS will be at the ending position of the transcript while the TTS will be at the starting point.
* This algorithm yielded 244728 regions for each type.

## Map the mutations to the TTS and TSS regions:

* Read the MAF file as well as the obtained TSS and TTS files.
* From the TSS and TTS files, sort the ranges in a dictionary by strand, by chromosome and in an ascending order for each chromosome.
* It is possible to remove overlapping ranges to reduce the number (will be interesting if working with larger datasets).
* Then, for each mutation extract its chromosome, position, and the strand on which lies the transcript.
* Using the strand information choose which dictionary to work with and using the chromosome choose which list of ranges to perform on it a binary search.
* Allocate the index to the region if it maps correctly (there were approximately 6k mutations that mapped in both TSS and TTS regions).
* The mutations that do not map will be saved separately.

## Extract mutational signatures:

* Using SigProfilerMatrixGenerator, build the mutational catalogs for each region (TSS, TTS, AND WHAT REMAINS)
* Rung SigProfilerExtractor on these matrices to extract as minimal as 1 signature and as maximaum as 4 signatures as the best result on the total data was with 4 signatures.
* I reduced the number of iterations to 100 instead of 500 to accelerate the process and the results seems to be good.

# Results:

## TSS signatures:

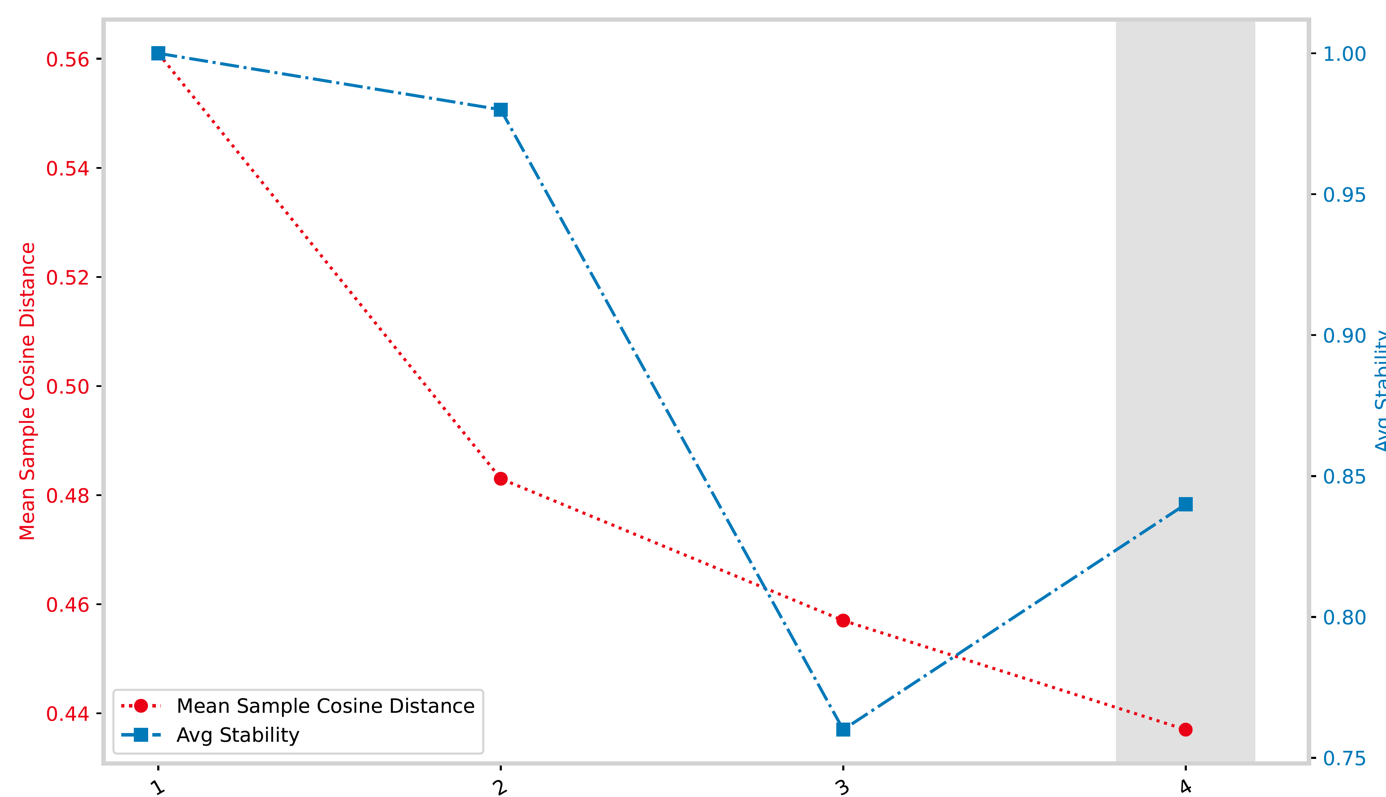


Figure 1 Selection plot for TSS regions

Une image contenant texte, capture d’écran, intérieur

Description générée automatiquement

Figure 2 Decomposition plot for SBSA of the TSS region

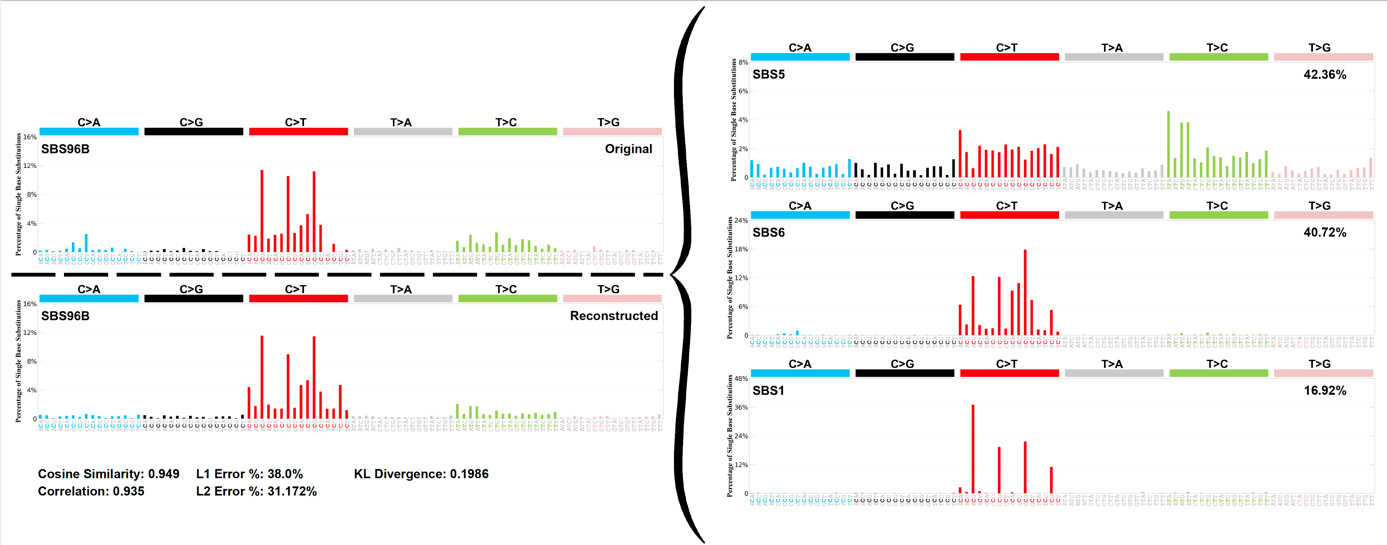


Figure 3 Decomposition plot for SBSB of the TSS region

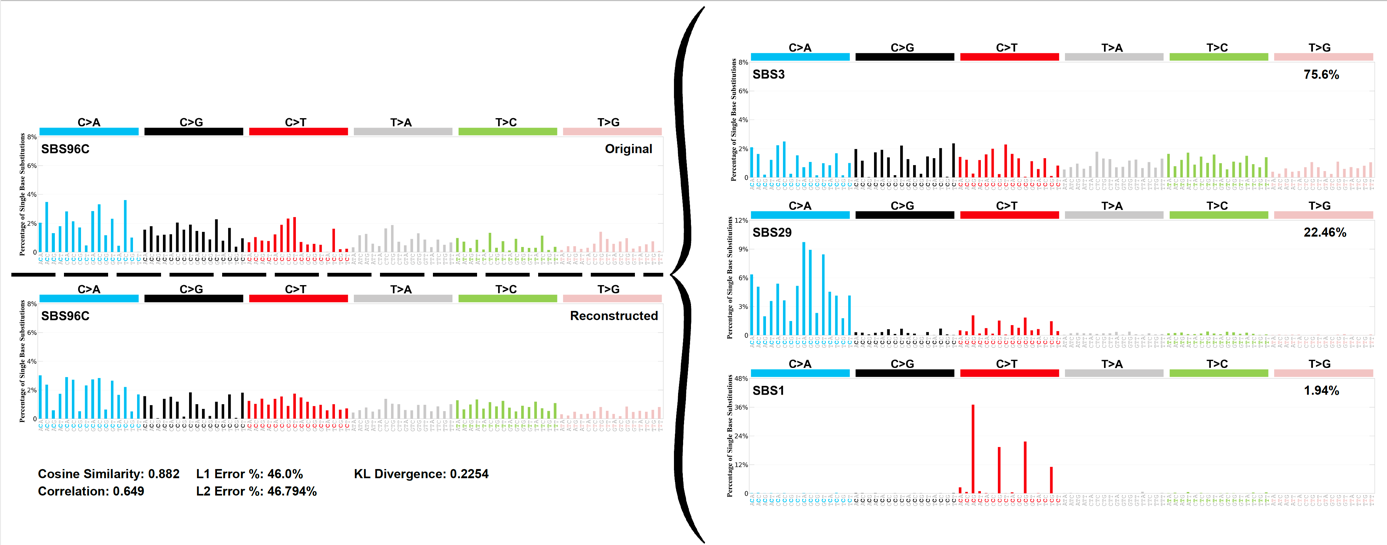


Figure 4 Decomposition plot for SBSC of the TSS region

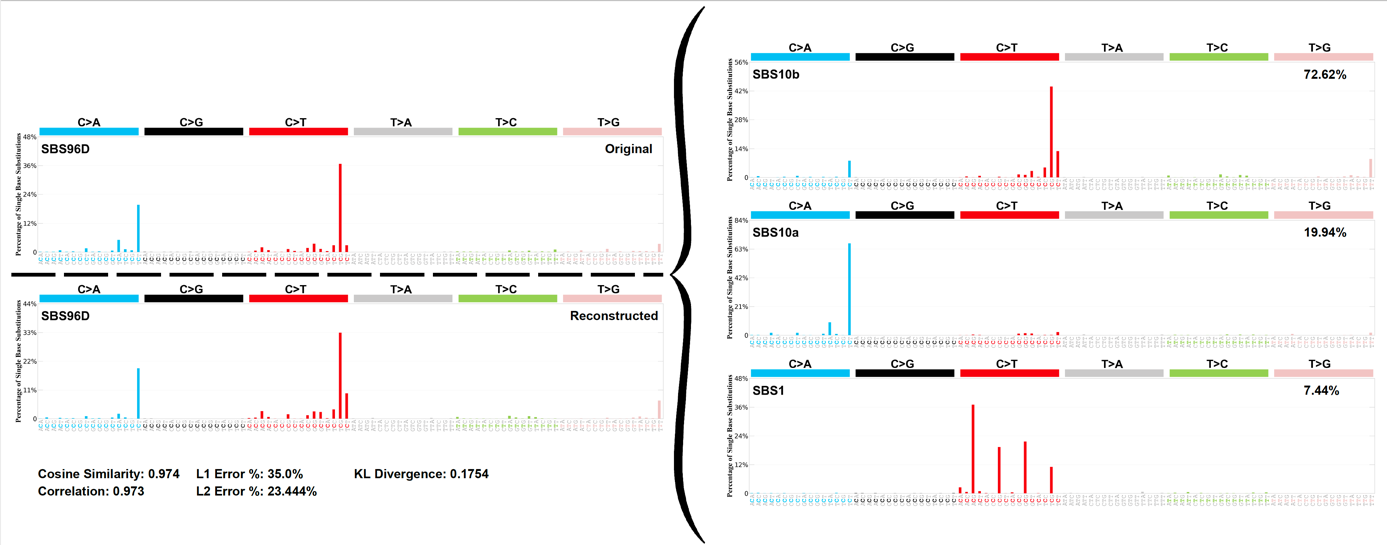


Figure 5 Decomposition plot for SBSD of the TSS region

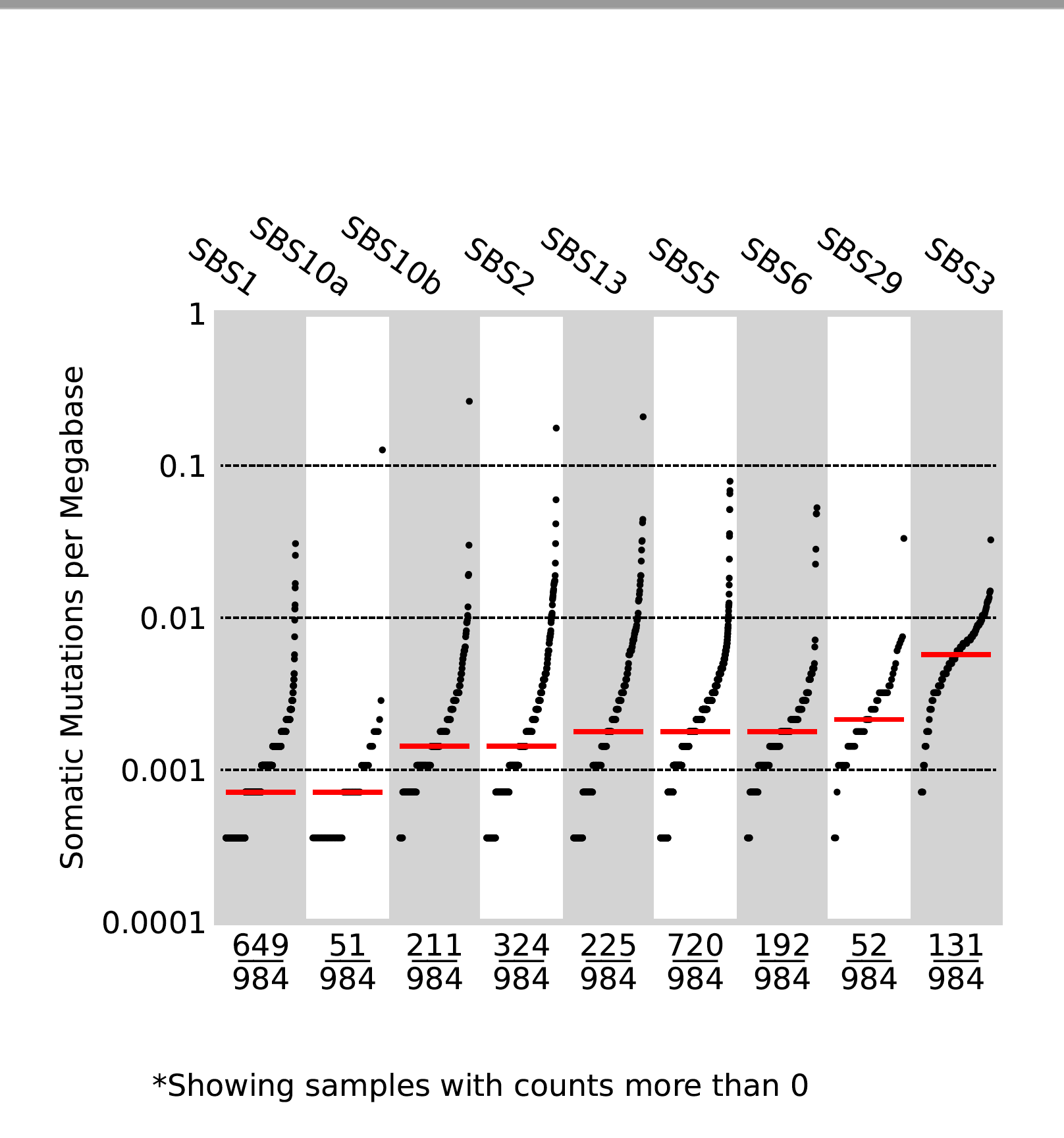


Figure 6 TMB plot of the decomposed COSMIC signatures for the TSS region

## TTS signatures:

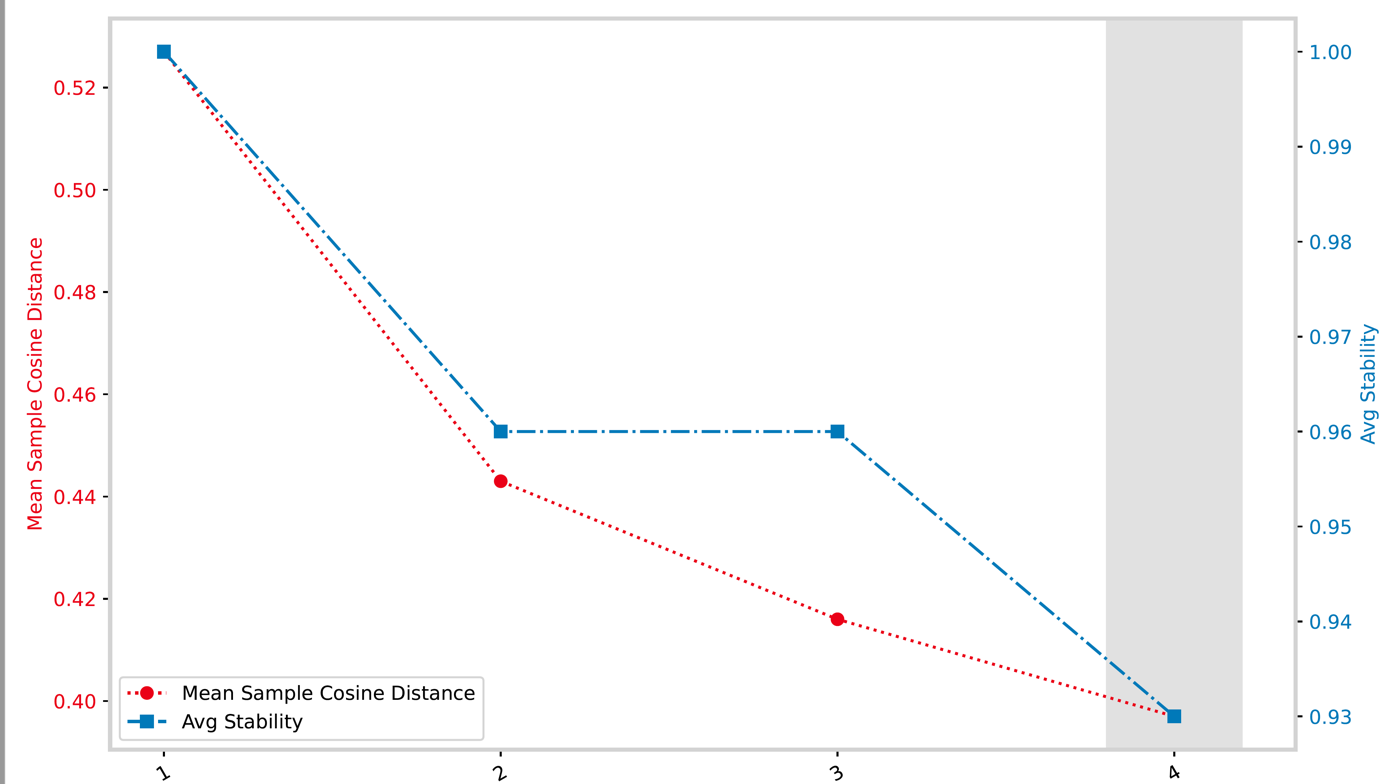


Figure 7 Selection plot fot the TTS region

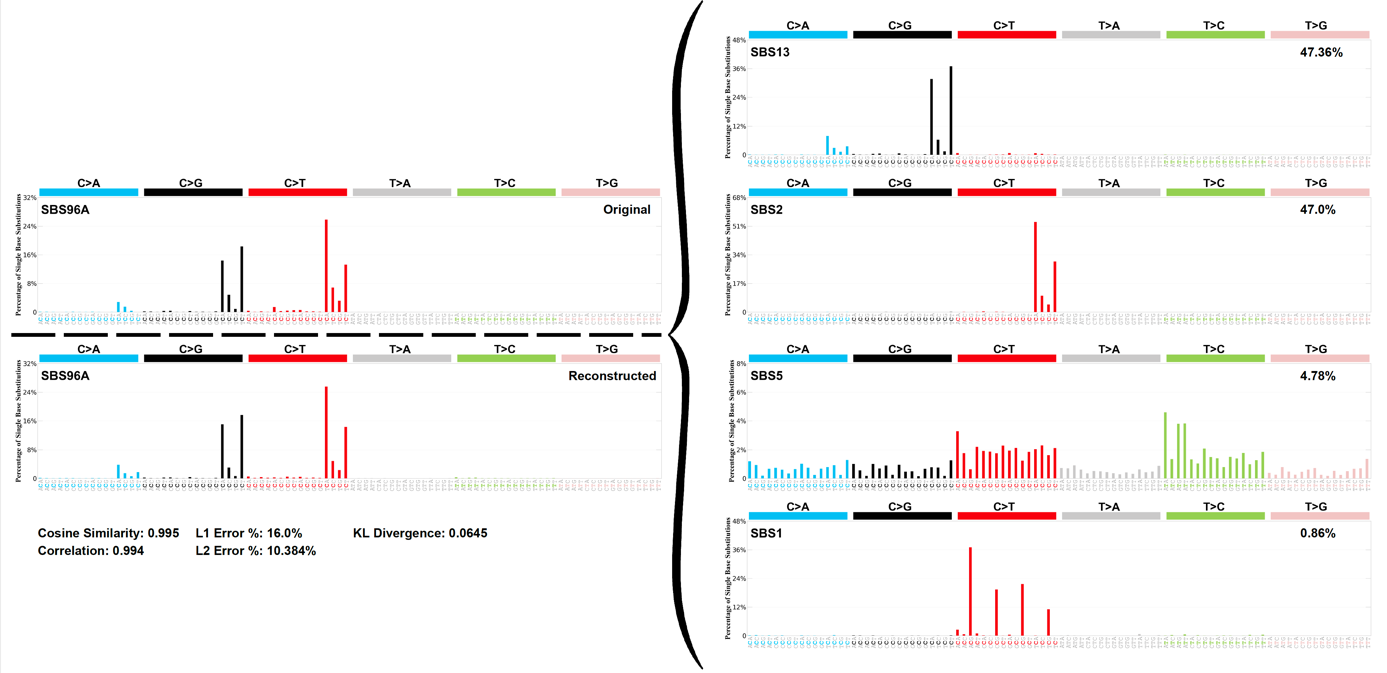


Figure 8 Decomposition plot for SBSA of the TTS region

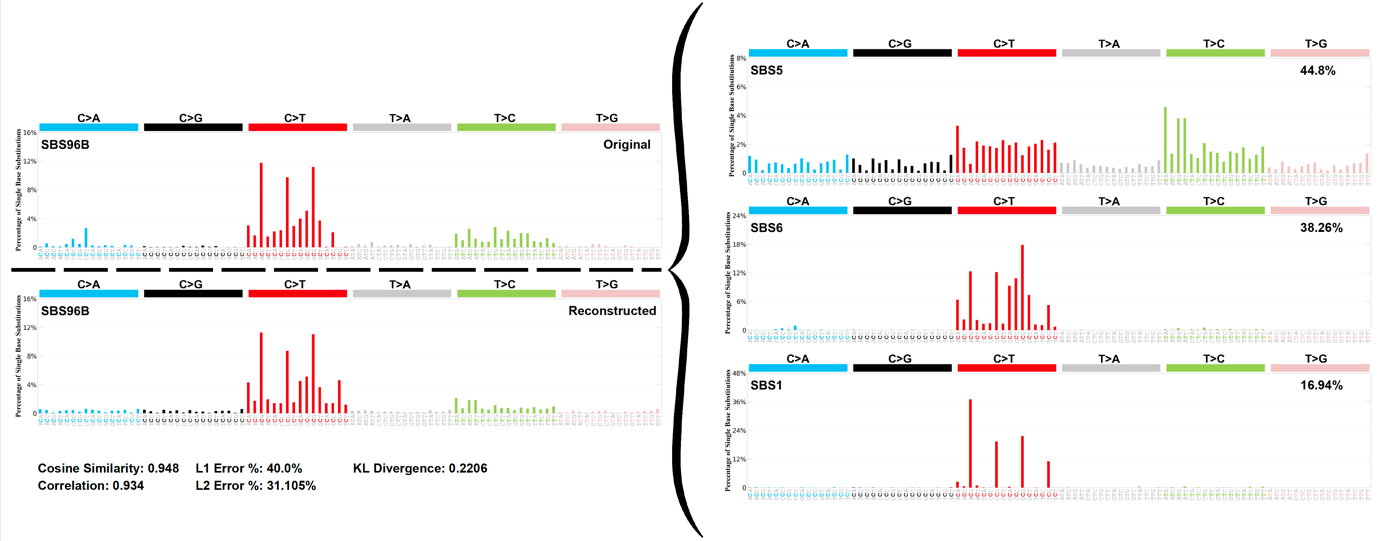


Figure 9 Decomposition plot for SBSB of the TTS region

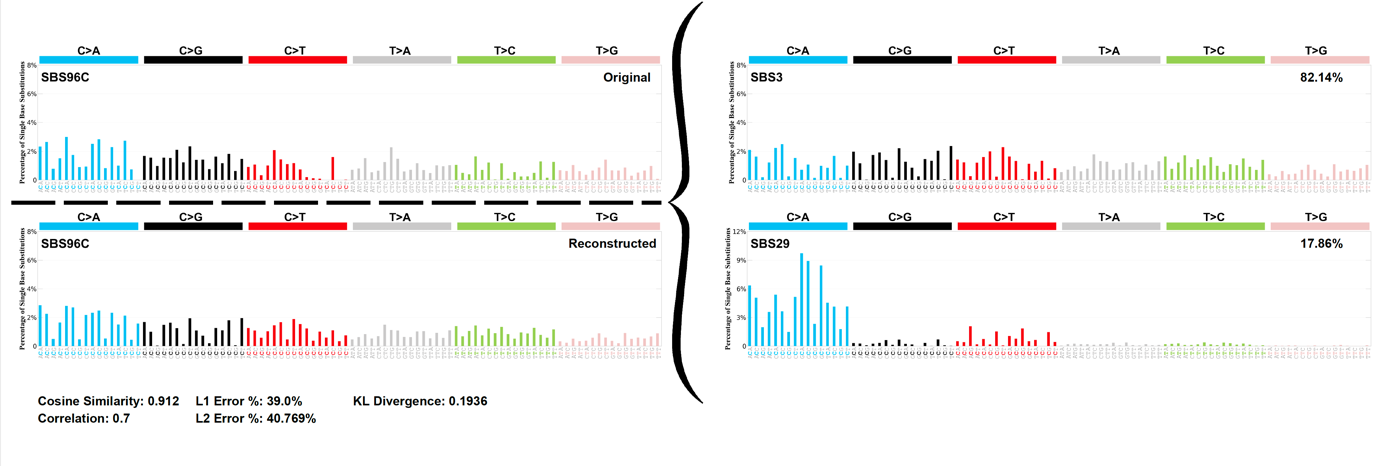


Figure 10 Decomposition plot for SBSC of the TTS region

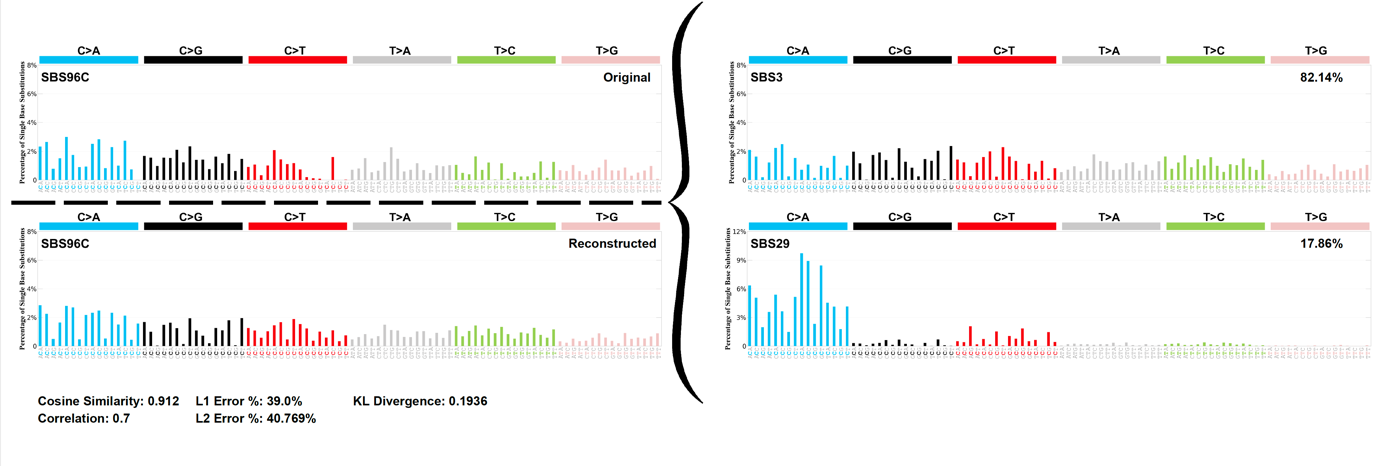


Figure 11 Decomposition plot for SBSD of the TTS region

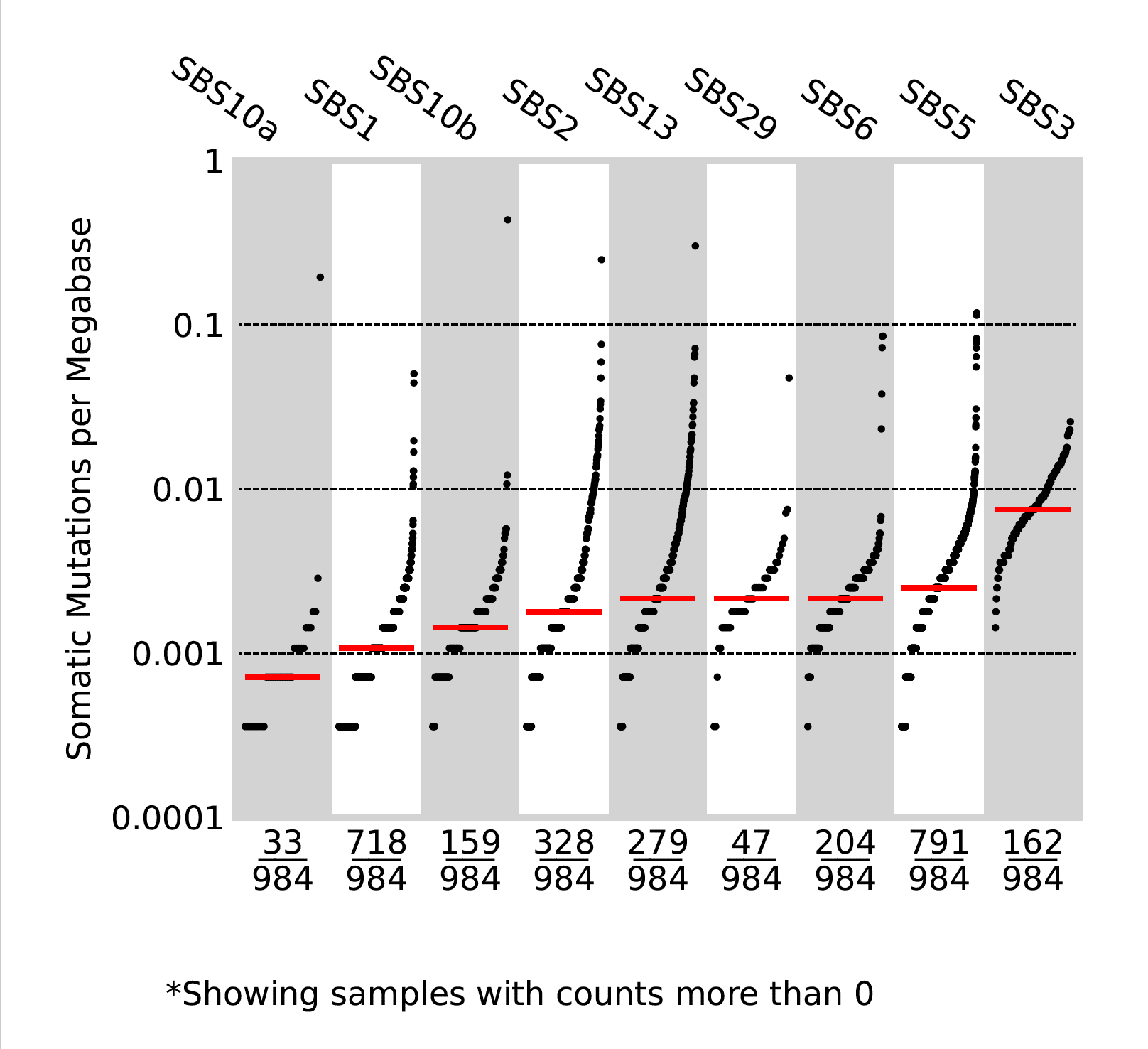


Figure 12 TMB plot for the decomposed COSMIC signatures acting in TTS

## Signatures of the coding regions (not TSS not TTS):

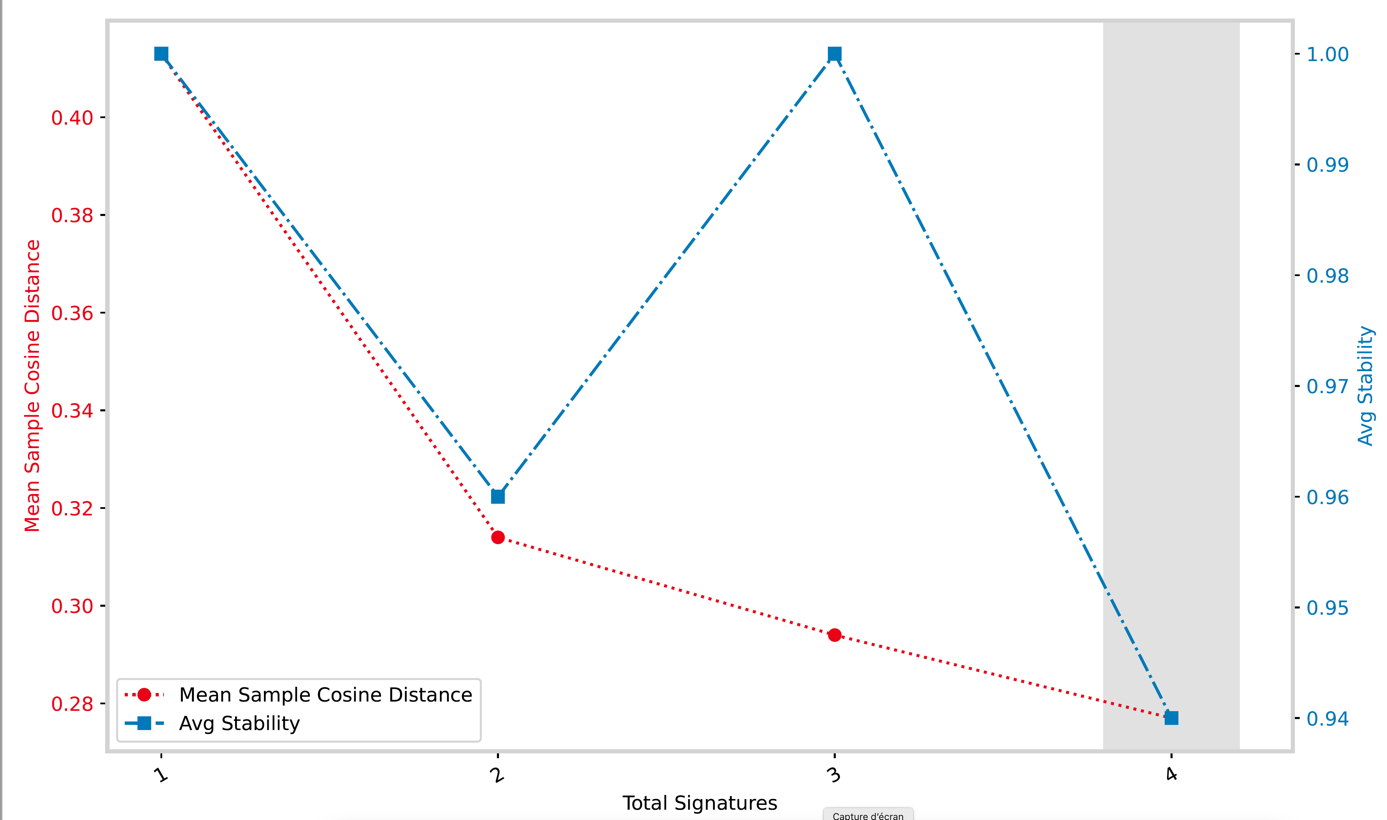


Figure 13 Selection plot

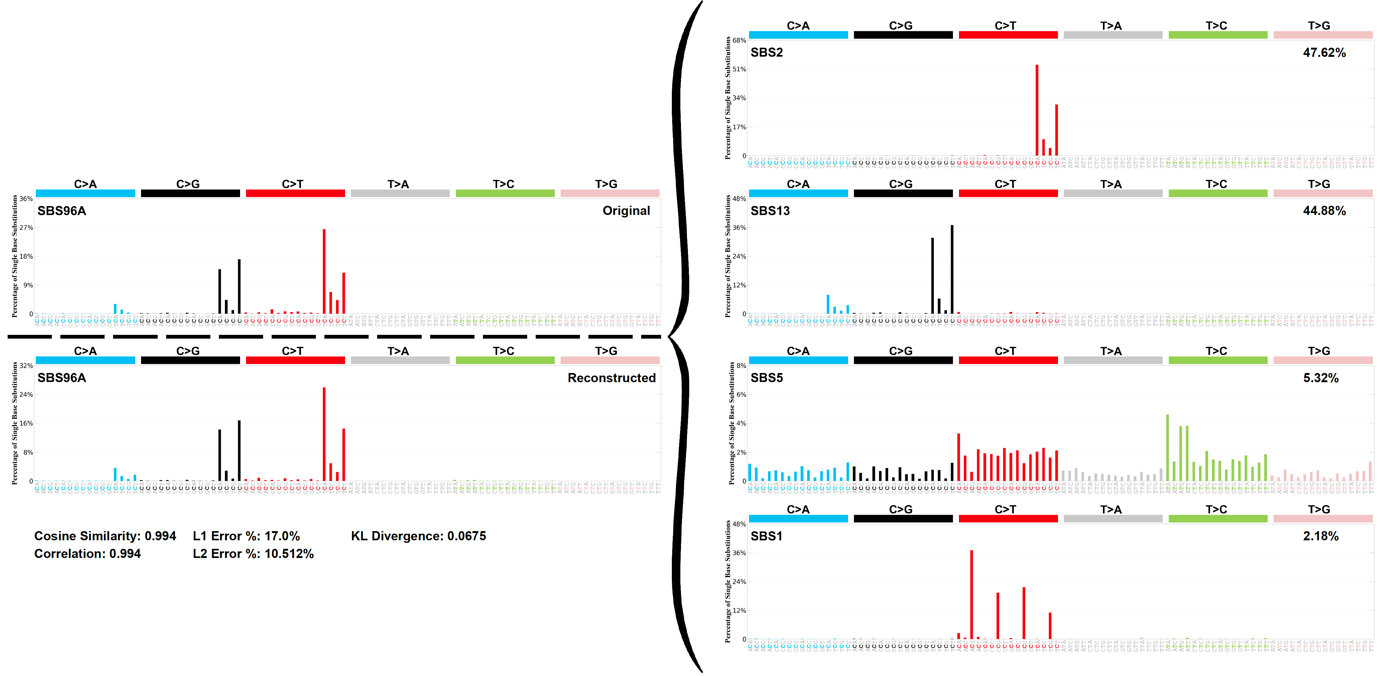


Figure 14 Decomposition plot for SBSA

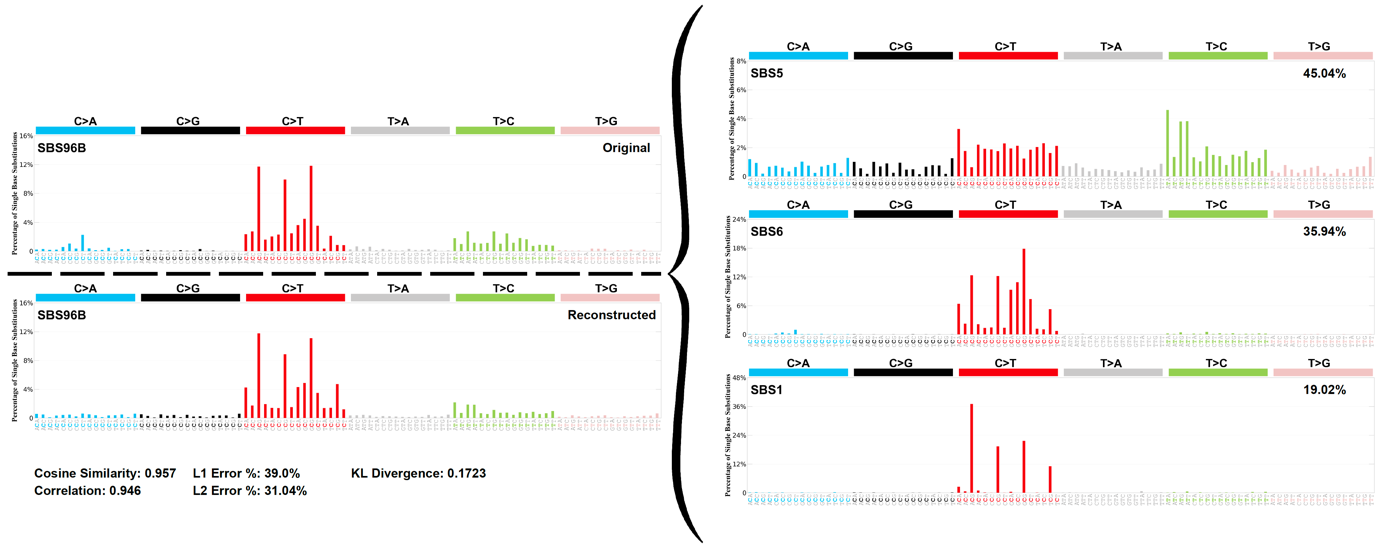


Figure 15 Decomposition plot for SBSB

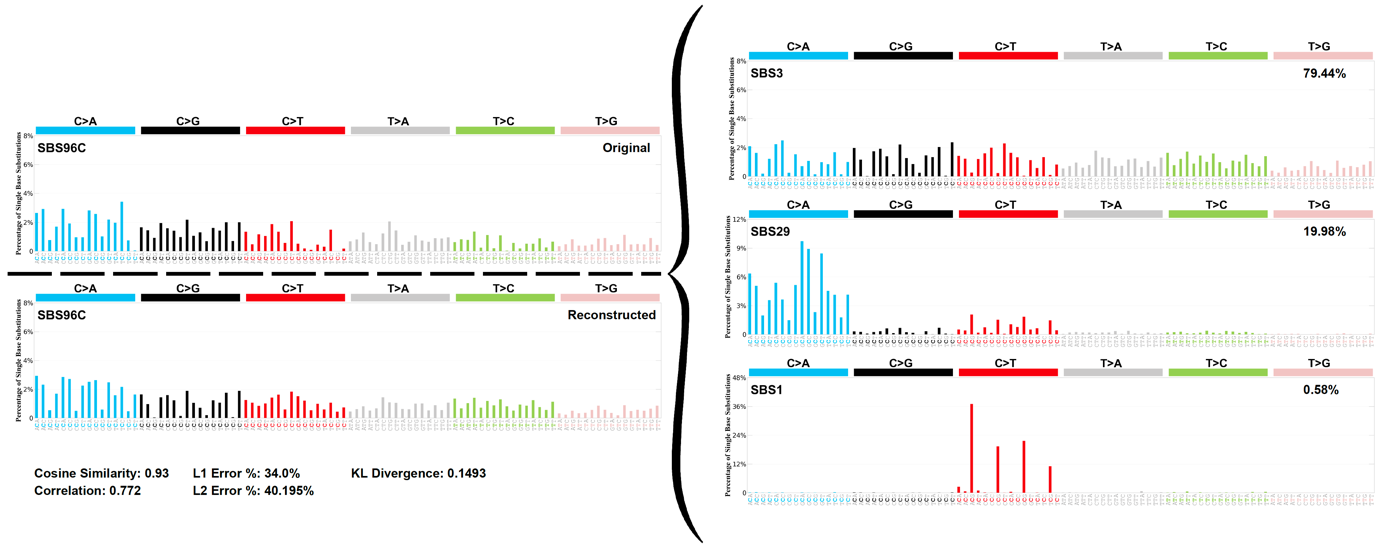


Figure 16 Decomposition plot for SBSC

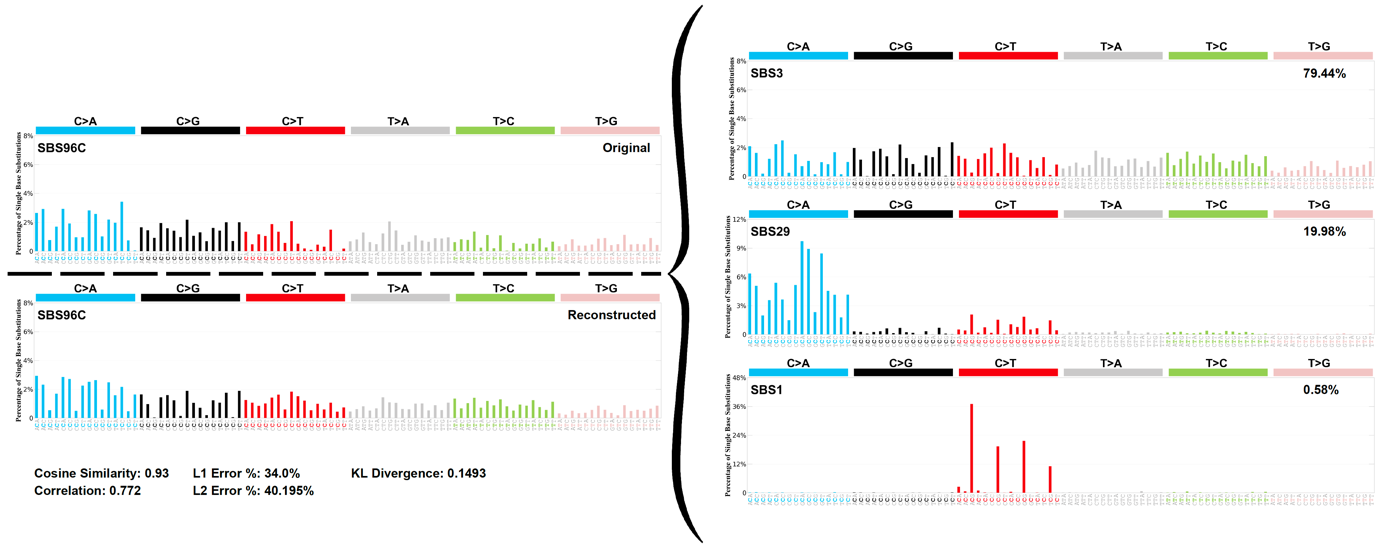


Figure 17 Decomposition plot for SBSD

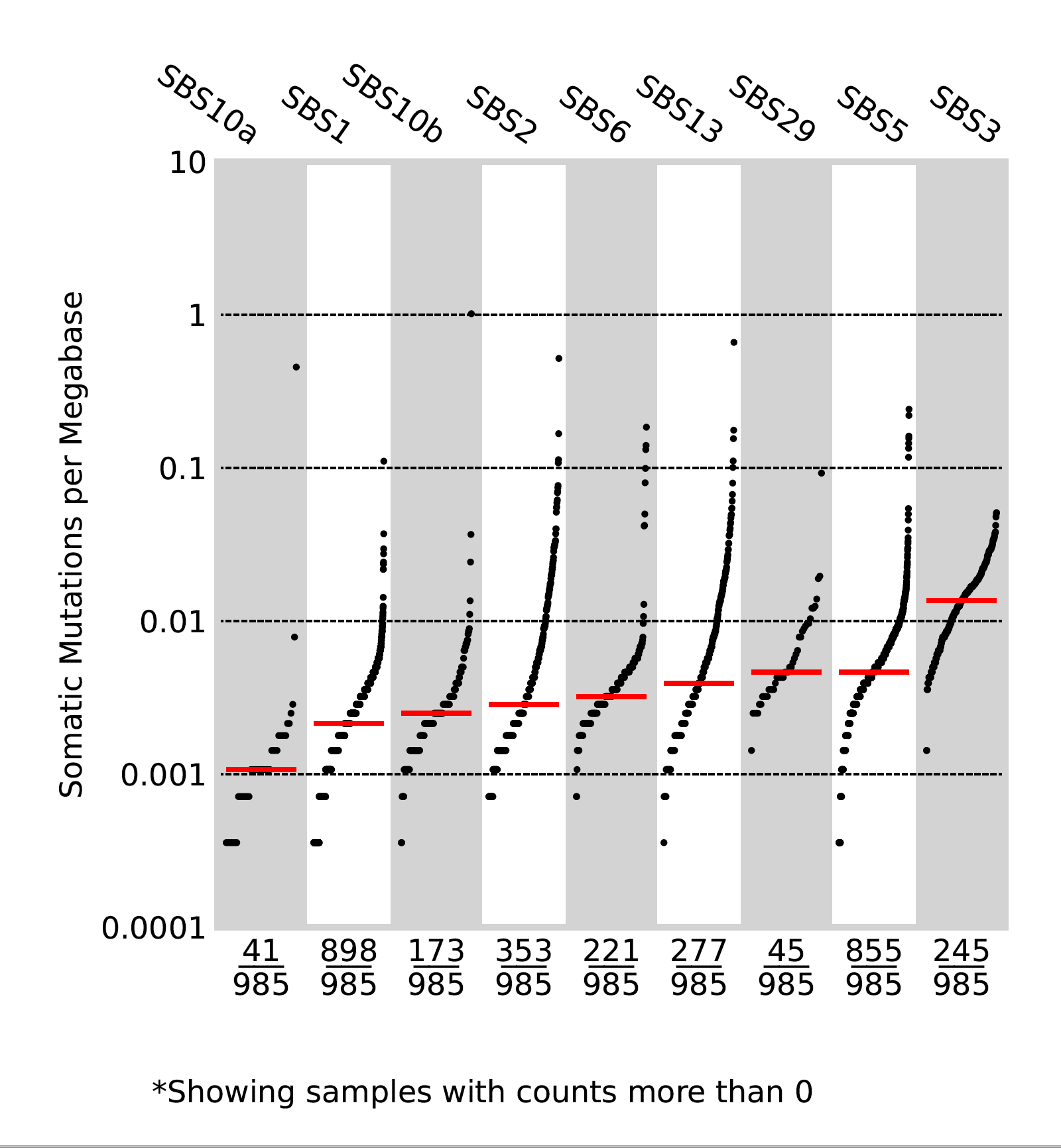


Figure 18 TMB plot for the decomposed COSMIC signatures

* The results shown here in terms of signatures acting in breast cancer are more or less similar between themselves and also with the results obtained previously from the whole dataset. There are some differences in the profiles, but this doesn’t impact the results that much. For instance, SBSB is identical between all the results, the decomposition of B yields in SBS5 with 45% contribution, SBS6 with 36% and SBS1% with 19% contributions.
* So, there is no difference between the signatures acting in these regions but there seems to be a difference in the activity of these signatures (number of samples and mutational burden) between the different regions as shown by the TMB plots.
* In order to be able to tell if there is a significant difference, statistical tests will be required.