# Materials and Methods:

## Working Environment:

Most of this analysis was conducted on an iMac running MacOS Monterey version 12.2.1 Equipped with a 9th generation 6 cores Intel Core i5 processor unit, with a 1 Tb of storage and 16 Gb of DDR4 RAM. The construction of mutational matrices and the extraction of mutational signatures were performed using Institut Curie’s High-Performance Computing (HPC) Cluster Centos.

All scripts used in this analysis are available on the GitHub repository named TRC-related-mutational-signatures (<https://github.com/Ala-Eddine-BOUDEMIA/TRC-related-mutational-signatures>).

The required packages for the python scripts to run are listed in the requirements file available in the GitHub repository. All the python scripts are developed to run on a UNIX-based environment, to that end, the file path semantics are not appropriate for WINDOWS systems.

The required packages for the R scripts to run are “recount2” and “maftools” and they are available through Bioconductor (Huber et al., 2015).

## Data acquisition:

### Mutation Annotation Format (MAF) Files:

This study was focused on breast cancer (BRCA) and bladder cancer (BLCA) for which MAF files were obtained from TCGA through the GDC portal. The BRCA MAF file had the identifier “6f5cde97-d259-414f-8122-6d0d66f49b74” and comprised data for 986 samples. On the other hand, the BLCA MAF file had the identifier “995c0111-d90b-4140-bee7-3845436c3b42” and contained data for 412 samples.

Both datasets resulted from mutation calling on Whole Exome Sequencing (WES) data using the “MUTECT” method (Cibulskis et al., 2013). WES was performed using either Illumina HiSeq2000, Illumina 2000, or Illumina Genome Analyzer II and the resulting reads were aligned on the Genome Reference Consortium Human Reference 38 (GRCh38). However, the sequencing library preparation protocols were not mentioned.

To date, the two files are no longer accessible on the GDC portal with the previously mentioned identifiers. However, similar files are available at the GDC Legacy Archive with credentials “1285b1e8-5fe7-4481-b9f7-4b60fce9d45e" for BRCA and “f416b22d-fc7e-400c-b852-ccb65a6610f9" for BLCA.

### GRCh38 Generic Feature Format Version 3 (GFF3) Annotation:

In preparation to investigate the mutational processes operational at TSS and TTS regions of the protein coding genes, the human genome’s gene annotation for the GRCh38.p13 version was obtained as a GFF3 file from the GENCODE website <https://www.gencodegenes.org/human/>.

### HeLa cells data:

As a means to validate the TSS and TTS regions that were extracted, HeLa cells data were used to plot the average profiles of R-loops, phosphorylated RNA Protein A (pRPA), and Replication Fork Directionality (RFD). HeLa cells were the first human cell lines to become immortalized in tissue culture (Gey et al., 1952). For the sake of the validation process, the DNA-RNA hybrid Immunoprecipitation Sequencing (DRIP-Seq), Chromatin Immunoprecipitation Sequencing (CHIP-Seq) for pRPA datasets (Promonent et al., 2020), and Okazaki Sequencing (OK-Seq) (Petryk et al., 2016) were used. Preprocessed BigWig files were directly obtained from the team. The average enrichment profiles were computed and plotted utilizing deeptools (Ramírez et al., 2016).

### Transcriptomic data:

Next, in an effort to explore differences in mutational signatures between TSS and TTS of active and inactive genes. Transcriptomic data for breast and bladder cancers were downloaded from Recount2 project (Collado-Torres et al., 2017) as Rdata files. Recount2 project does not provide the read counts directly, instead, it provides the coverage per base matrices. The advantage of Recount2 is that all the samples have been reprocessed uniformly eliminating any computational biases. The actual read counts were obtained by dividing the sum of coverage per base by the read length of each read using a function provided by recount2 package (Collado-Torres et al., 2017). The read counts matrices were saved as tabula separated value (tsv) files. Metadata describing each sample were also obtained from the recount2 database and it has been used to filter samples that do not come from primary tumors. The BRCA dataset contained 1093 unique patients from which 983 patients match the MAF dataset patients. As for the BLCA dataset, 412 unique patients matched 408 patients in the MAF dataset. Both datasets provide expression levels for 58037 genes.

### MCF-7 cells data:

MCF-7 cell lines are estrogen (E2) receptor positive breast cancer cells. They were first established from a 69-year-old woman with metastatic breast adenocarcinoma by Dr. Soule’s group at the Michigan Cancer Foundation (MCF) (Soule et al., 1973). With an aim to localize mutations that coincide with R-loops occurring around TSS and TTS regions, the browser extensible data (BED) files obtained from DRIP-Seq were used. The DRIP-Seq was performed on MCF7 control cells, MCF7 cells that undergone E2 treatment for 2 hours, and MCF7 cells that undergone E2 treatment for 24 hours (Stork et al., 2016).

## Methods:

### Tools:

#### SigProfiler:

SigProfiler is a tool developed and maintained by Alexandrov’s group which they first published in 2013 as a MATLAB code (Alexandrov et al., 2013), and recently, they have made it available as a python3 package. It has several functionalities that can be installed separately. SigProfilerMatrixGenerator (Bergstrom et al., 2019) and SigProfilerExtractor (Islam et al., 2020) are the main two packages used during this analysis.

SigProfilerMatrixGenerator was developed to construct and plot different mutational matrices considering different mutational contexts in a computationally efficient way. For instance, it provides six mutational matrices for SBSs. The simplest matrix takes in account only for the six mutation types represented by the pyrimidine bases following the conventions of the field [C>T, C>A, C>G, T>A, T>C, T>G]. The tool provides also mutational matrices with 96 and 1536 channels taking into account one or two adjacent bases at both 3’ and 5’ positions respectively. For each of these matrices it builds their alternatives considering transcriptional strand bias yielding 6 matrices in total with 6, 24, 96, 384, 1536, 6144 channels and it is capable of plotting all of them except the matrix with 6144 channels.

On the other hand, SigProfilerExtractor implements a non-negative matrix factorization algorithm (NMF) that has shown remarkable results in deciphering the mutational signatures from the mutational matrices provided by SigProfilerMatrixGenerator package or from a custom matrix of a supported format. This tool models the mutational signatures as a matrix with “k” rows and “n” columns. Each element of a row represents the probability of a given process to generate a specific mutation type, and “k” is the number of mutation types or channels. Whereas each column represents a mutational signature and “n” is the number of signatures. Furthermore, each process operates at a specific rate in a given genome. That is, each process causes a specific number of mutations. The rate at which a mutational signature is present in a genome is called exposure and it is represented as matrix of “n” rows and “g” columns, where each element in the row corresponds to the exposure of a given signature and each column represents one genome from the dataset. Hence, a mutational catalog of a cancer genome is the product of a probability mass function representing mutational signatures multiplied by the exposure of that signature in that genome plus the errors related to non-systematic analysis. Following this model, the NMF algorithm will approximate the mutational signatures (S) and exposure matrices (E) to reconstruct the original catalog (M). Mathematically, this could be represented as M = S \* E + Ɛ or in a simplified way M ≈ S \* E (Alexandrov et al., 2013).

#### Maftools:

Maftools is an R Bioconductor package that regroups several functionalities useful in downstream analysis of mutational data. Maftools functionalities can be grouped into three different modules: visualization, analysis, and variant annotation. Each module contains several functions as shown in Figure 1. In this project Maftools was used mainly for its visualization capabilities and the high-quality plots that it generates as a way to get a global overview of the data and to point out any peculiarities (Mayakonda et al., 2018).

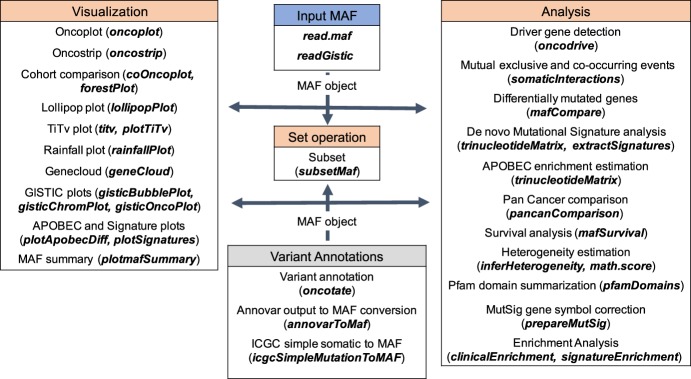


Figure 1 Overview of Maftools package.

Each table represented a module, and the name of the module is shown at the header of each table. Each table contains the name of the functions available in that module with a description of each function between parenthesis (from Mayakonda et al., 2018)

#### Deeptools:

Deeptools is similar to Maftools and combines several features commonly required for deep sequencing analysis. Deeptools is implemented in python3 and is available as a galaxy webserver. It has four modules where each module contains tools specific to quality control, data integration, data visualization, and downstream analysis. Deeptools was used to plot average RFD, R-loops, and CHIP-Seq p-RPA profiles around the TSS and TTS regions. This was performed using the *computeMatrix* function by scaling the regions of interest to the same length and computing the distribution of scores of these regions. The provided output is used by the *plotProfile* function to visualize the results (Ramírez et al., 2016).

### Pipeline:

The analysis workflow was designed to investigate the SBS signatures with 96 channels in a general-to-specific approach. It contains five major steps consisting in a global analysis of mutational signatures in BRCA and BLCA followed by analysis focused on specific regions of interest, namely, TSS and TTS of protein coding genes, TSS and TTS of active and inactive protein coding genes, regions between genes in a head-on (HO) or co-directional (CO) orientation and finally within and outside of R-loops (Figure 2).

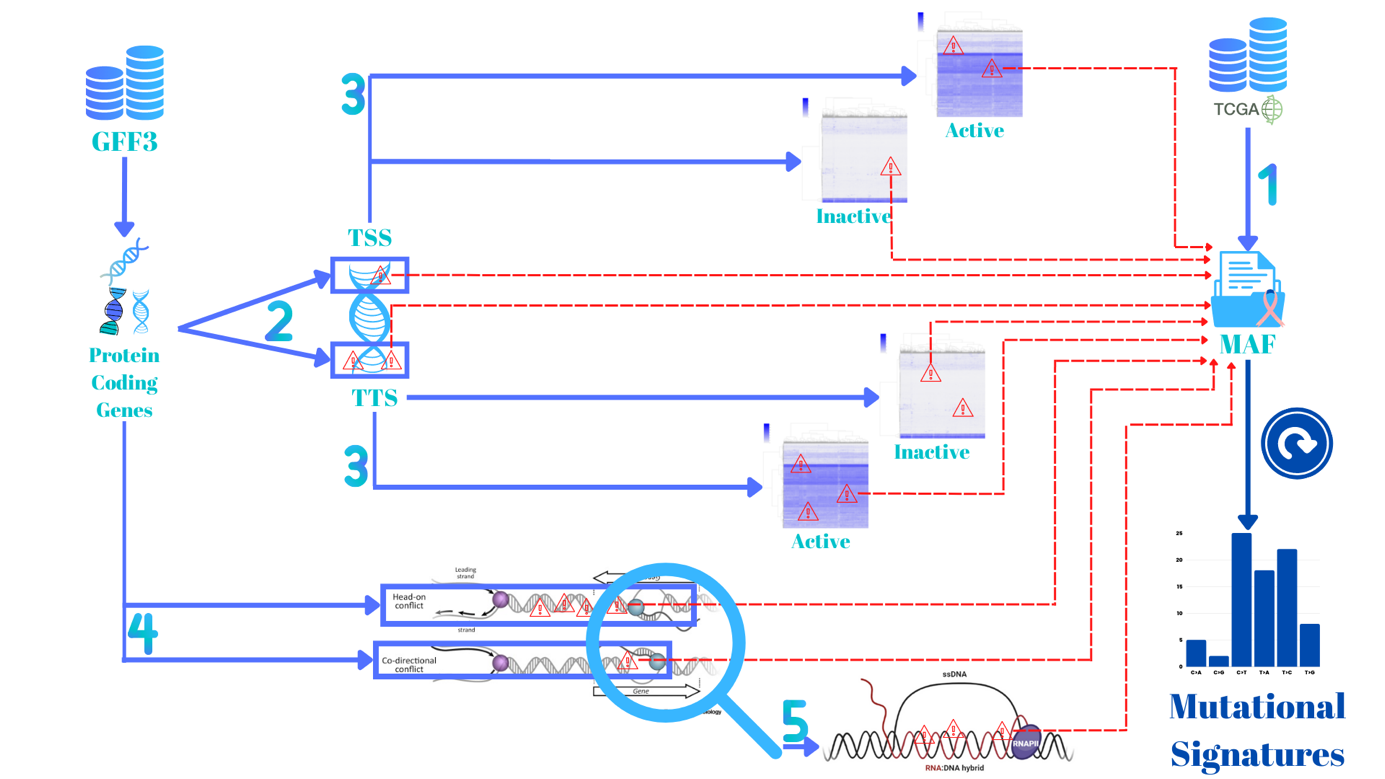


Figure 2 Analysis workflow of the project.

Step 1 consists in downloading MAF files from TCGA and extracting mutational signatures. Step 2 consists in extracting the coordinates of protein coding genes and their TSS and TTS regions, then to sort mutations by region and extract mutational signatures for each region. Step 3 is similar to step 2 but instead the TSS and TTS region are further separated based on active and inactive genes. Step 4 consists in extracting regions in CO or HO orientations and sort the mutations based on each region then decipher the signatures operating in each region. Step 5 consists in zooming-into the mutations that coincide with R-loops and the mutations outside of R-loops then extract the mutational signatures.

In order to investigate differences between TSS and TTS regions, the GFF3 file was used to extract the coordinates of protein coding genes and saved as a BED file. The genes in the BED file were then sorted by the DNA strand on which the transcript lies (+ or -), then sorted by chromosome and finally by starting position. Afterwards, the genes’ positions were compared by pairs searching for overlapping genes. In case of overlapping, the smallest gene gets discarded leading to a removal of isoforms or overlapping genes on the same strand. The final step consisted in defining the TSS and TTS regions for each gene. This was done with a window frame of 3000 base before and after a gene’s starting and ending position coordinates yielding a 6000 bases window frame for each region. Moreover, HeLa DRIP-Seq, OK-Seq and CHIP-Seq for pRPA were used to validate the selected regions as the average profiles should present an enrichment of R-loops at TSS and TTS regions, an enrichment of pRPA signal at TTS (Promonet et al., 2020) and a switch of replication fork directionality (Petryk et al., 2016).

The following phase of analysis was designed to see if mutational signatures vary between TSS, and TTS regions of active and inactive protein coding genes based on their expression levels. To that goal, genes with an average count per million (CPM) that is lower than 1 were considered to be inactive. This has yielded 14154 active genes and 3955 inactive genes for BRCA and 14140 active genes and 3969 inactive genes for BLCA. For each case the TSS and TTS region coordinates were defined in the same way as in the previous phase.

To extend the analysis further and based on the knowledge that conflicts between transcription and replication machineries are unavoidable the next step was guided towards intergenic regions between genes in CO or HO orientations. HO collisions between the two machineries tend to be more deleterious and it was shown that R-loops and pRPA are more enriched in TTS regions of converging genes indicating that these regions are hotspots for TRCs (Promonet et al., 2020). Taking that into consideration, genes within a distance of 10000 bases or less were considered as neighboring and depending on their TSS and TTS orientation, the regions between them were classified into three categories, namely co-directional, divergent, or convergent which are subclasses of the head-on orientation. However, to avoid any conflicts the overlapping genes on two different strands were also removed for this analysis phase, reducing the number of genes from 18340 to 16074. A CO region will be located between two genes that are transcribed in the same orientation. That is the transcribed template of the two genes are located on the same DNA strand whether it was the “+” or the “-” strand. Conversely, an HO region will be located between two genes where the transcribed template of one gene is located at the opposite DNA strand on which lies the transcript of the second.

Finally, using MCF-7 DRIP data, mutations were sorted on the basis of their occurrence, namely, if they co-occur with R-loops, if they co-occur with R-loops that are within a 3000 base-pair from the central position of the TSS or TTS regions, or if they occur outside of R-loops.

For each step described above mutations were sorted using binary search algorithm. Mutational profiles were constructed using SigProfilerMatrixGenerator and mutational signatures were extracted using SigProfilerExtractor with both 100 and 500 NMF iterations.

Bergstrom, E.N., Huang, M.N., Mahto, U. *et al.* SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* **20,** 685 (2019). <https://doi.org/10.1186/s12864-019-6041-2>

Cibulskis, K., Lawrence, M., Carter, S. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31,** 213–219 (2013). <https://doi.org/10.1038/nbt.2514>

Collado-Torres, L., Nellore, A., Kammers, K., Ellis, S. E., Taub, M. A., Hansen, K. D., Jaffe, A. E., Langmead, B., & Leek, J. T. (2017). Reproducible RNA-seq analysis using recount2. *Nature biotechnology*, *35*(4), 319–321. <https://doi.org/10.1038/nbt.3838>

Islam, S. M., Díaz-Gay, M., Wu, Y., Barnes, M., Vangara, R., Bergstrom, E. N., He, Y., Vella, M., Wang, J., Teague, J. W., Clapham, P., Moody, S., Senkin, S., Li, Y. R., Riva, L., Zhang, T., Gruber, A. J., Steele, C. D., Otlu, B., … Alexandrov, L. B. (2020). Uncovering novel mutational signatures by *de novo* extraction with Sigprofilerextractor. <https://doi.org/10.1101/2020.12.13.422570>

Gey, G. O., W. D. Coffman, and M. T. Kubicek. Tissue culture studies of the proliferative capacity of cervic

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., Pagès, H., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods*, *12*(2), 115–121. <https://doi.org/10.1038/nmeth.3252>

Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C., & Koeffler, H. P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome research*, *28*(11), 1747–1756. <https://doi.org/10.1101/gr.239244.118>

Petryk, N., Kahli, M., d'Aubenton-Carafa, Y., Jaszczyszyn, Y., Shen, Y., Silvain, M., Thermes, C., Chen, C. L., & Hyrien, O. (2016). Replication landscape of the human genome. *Nature communications*, *7*, 10208. <https://doi.org/10.1038/ncomms10208>

Ramírez, Fidel, Devon P. Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S. Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. [deepTools2: A next Generation Web Server for Deep-Sequencing Data Analysis](http://nar.oxfordjournals.org/content/early/2016/04/12/nar.gkw257.abstract). Nucleic Acids Research (2016). [doi:10.1093/nar/gkw257](http://doi.org/10.1093/nar/gkw257).

Soule HD, Vazguez J, Long A, Albert S and Brennan M: A human cell line from a pleural effusion derived from a breast carcinoma. J Natl Cancer Inst 51: 1409-1416, 1973.al carcinoma and normal epithelium. Cancer Res 1952. 12 :264 –265.

Stork, C. T., Bocek, M., Crossley, M. P., Sollier, J., Sanz, L. A., Chédin, F., Swigut, T., & Cimprich, K. A. (2016). Co-transcriptional R-loops are the main cause of estrogen-induced DNA damage. *eLife*, *5*, e17548. <https://doi.org/10.7554/eLife.17548>