WHAT ARE MUTATIONAL SIGNATURES?

ENVIRONMENT

Exogenous mutagenic processes

Endogenous mutagenic processes

DNA repair

Replication

MOLECULAR NETWORK

a Mutation counts for each mutation category

C > A, C > T, C > G, ...
Mutation category

b Inference of mutation signatures and their properties

Signatures

Frequencies of mutation categories

c Linking signatures to molecular causes

d Role in cancer and genome evolution and diversity

MAIN GOAL

An advanced toolkit offering end-to-end solutions for detecting, analyzing, and interpreting mutational signatures — created to assist bioinformaticians in gaining deeper insights from complex mutational datasets.
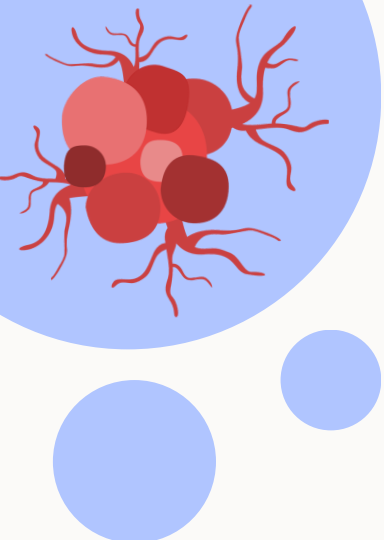
1
Downloading and preprocessing mutational data from TCGA database

2
Extraction of mutational signatures present in the data

3
Visualization and analysis of mutational signatures

DOWNLOADING AND PREPROCESSING MUTATIONAL DATA FROM TCGA DATABASE

First step involves creating custom mutational databases from large genomic datasets, which will later be used for mutational signature extraction. This is crucial for cancer genomics research, where identifying mutational patterns helps understand cancer etiology and development

Flowchart:
Request File → Database Filtering → VCF Generation → SigProfiler → Matrices (SBS/DBS/ID)

Input:
- mutations.parquet.gzip: Genomic mutations (Chromosome, Position, Alleles, etc.)
- request_file.txt: User-defined filters (e.g., TCGA-COAD, Colon)

Output structure:
```
mutational_matrices/
├────── DBS/        # Doublet base substitution matrices
├────── ID/         # Insertion/deletion matrices
├────── SBS/        # Single base substitution matrices
└────── vcf_files/  # Intermediate VCF files
```

EXTRACTION OF MUTATIONAL SIGNATURES PRESENT IN THE DATA

Second step matches mutation patterns in the data to known COSMIC signatures (e.g., SBS1 [aging], SBS6 [mismatch repair])
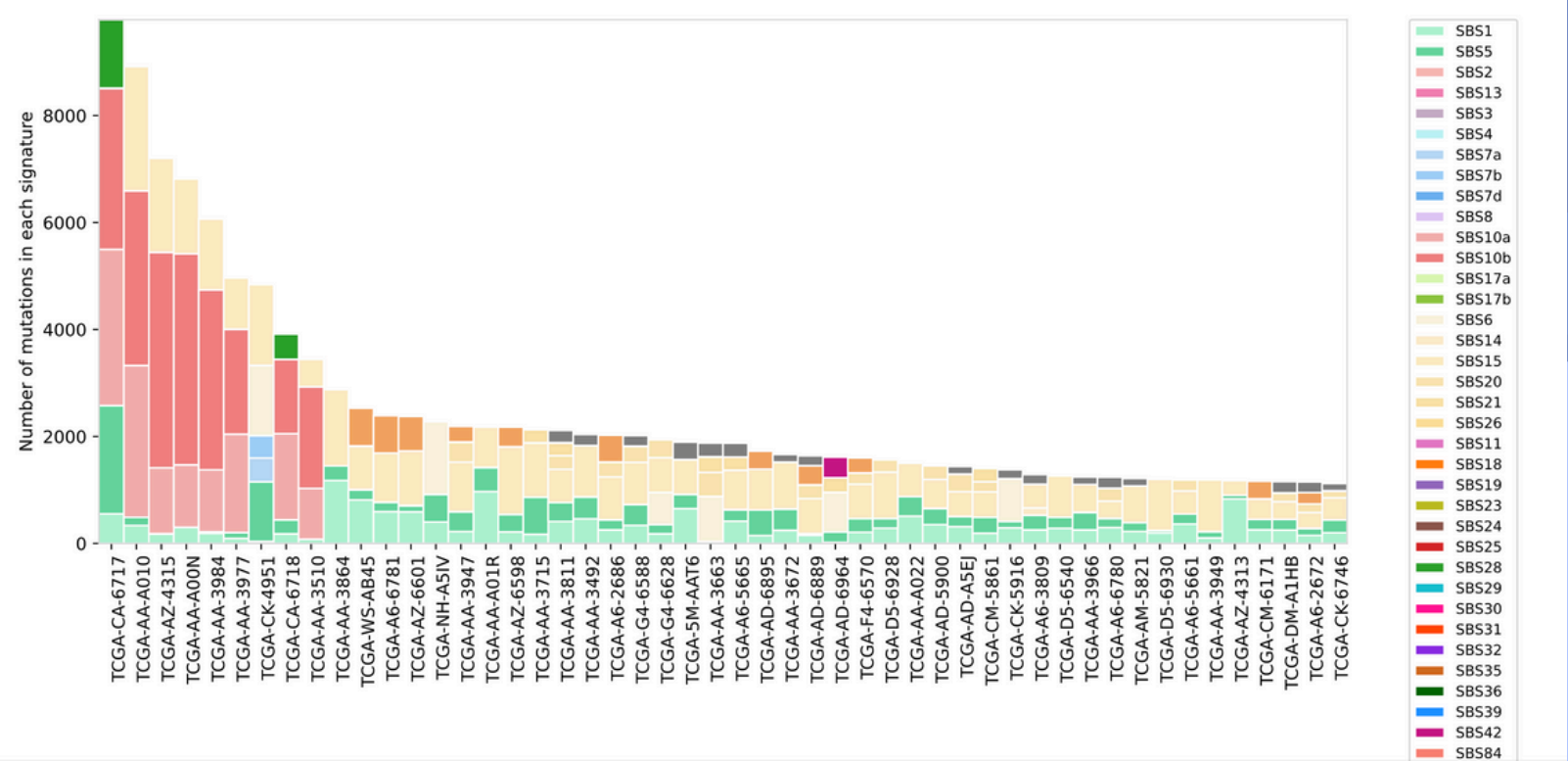
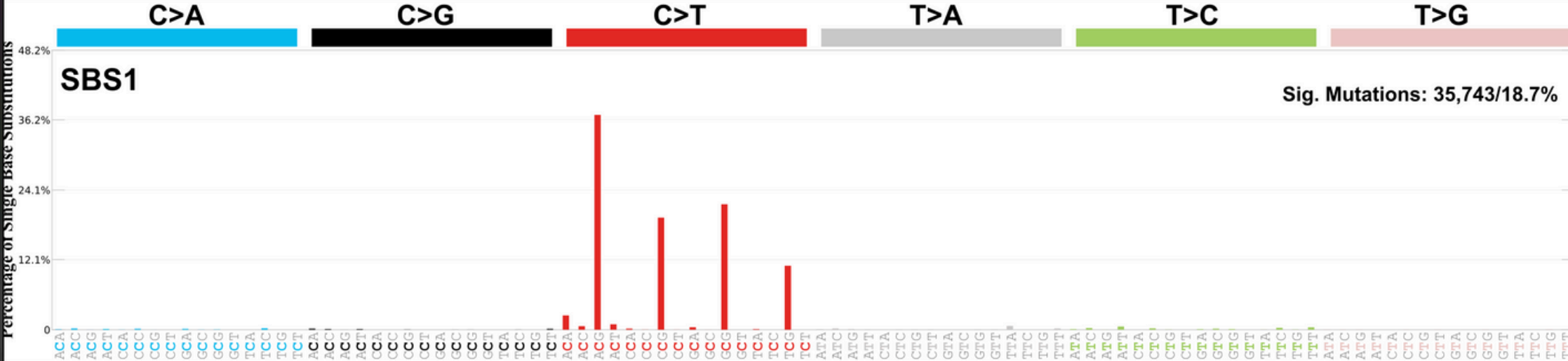**Input:** Mutational matrices (SBS/DBS/ID) from Step 1

Output: Signature activities, stats per sample

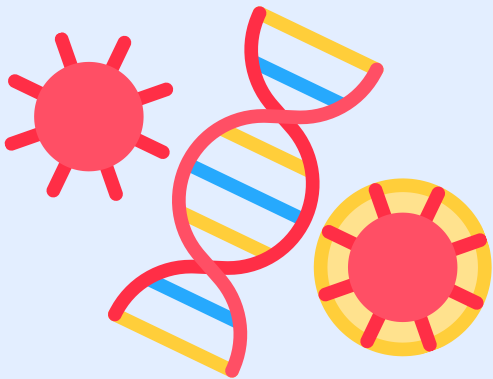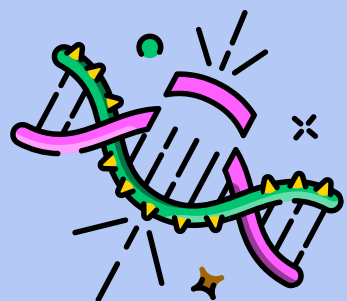**Assignment Solution Activity Plots**

Assignment Solution SBS1 Plot

SBS1 — Sig. Mutations: 35,743/18.7%

Assignment Solution Stats

| Sample Names	Total Mutations	Cosine Similarity	L1 Norm	L1_Norm_%	L2 Norm	L2_Norm_%	KL Divergence	Correlation
TCGA-3L-AA1B	143	0.942	85.036	59.466%	12.741	33.775%	0.43809	0.932
TCGA-4N-A93T	100	0.952	48.999	48.999%	7.173	30.642%	0.36087	0.941
TCGA-4T-AA8H	130	0.939	74.445	57.265%	11.215	34.512%	0.43245	0.926

VISUALIZATION AND ANALYSIS OF MUTATIONAL SIGNATURES

Third step helps with interpretation of COSMIC signature assignments from Step 2. using visualizations

Plots for SBS96
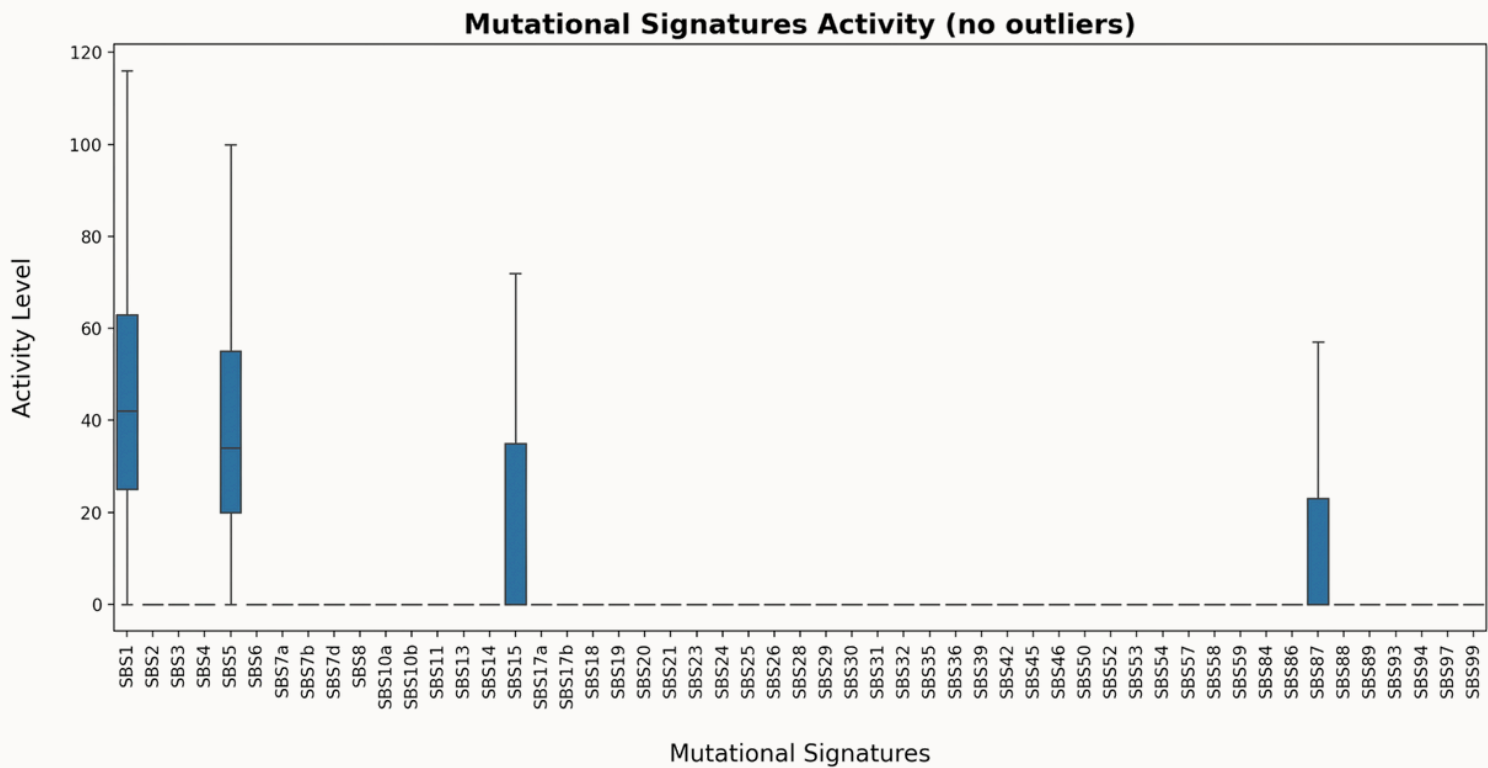
Input: Assignment_Solution_Activities.txt (SigProfiler output)

Output: Publication-ready plots + summary report

Number of Patients with Active Signatures

Etiology Prevalence - All Patients

Etiologies
- Clock-like (418 patients, 37.4%)
- dMMR (263 patients, 23.5%)
- Chemotherapy (143 patients, 12.8%)
- Polymerase mutation (88 patients, 7.9%)
- Unknown (79 patients, 7.1%)
- ROS (38 patients, 3.4%)
- UV (30 patients, 2.7%)
- AID/APOBEC activity (22 patients, 2.0%)
- Tobacco (15 patients, 1.3%)
- Aflatoxin (11 patients, 1.0%)
- Haloalkanes (4 patients, 0.4%)
- dBER (3 patients, 0.3%)
- HRD (3 patients, 0.3%)
- Colibactin (1 patients, 0.1%)

Hierarchical Clustering of Mutational Signatures