BUILDING MUTATIONAL SIGNATURES IN CANCER USING DEEP BAYESIAN NEURAL NETWORKS

A DEEP DIVE INTO CANCER MUTATIONAL SIGNATURES

Jan Alfonso Busker

352905 - j.a.busker@st.hanze.nl
Hanzehogeschool Groningen

ABSTRACT

Cancer, characterized by uncontrolled cell growth, accumulate mutations. The mutations that occur in the context of cancer development are a result of exposure to various DNA-damaging processes and accumulate throughout life. The sources of these DNA-damaging processes include both endogenous and exogenous factors. These genetic variations result in unique "mutational signatures" within the DNA sequence.
This project aims to refine the statistical model and the current representation of mutations by building mutational signatures of cancer using deep bayesian neural networks. Additionally, there is a plan to expand the representation to capture more context. Increasing the context involves subdividing mutations (1). By looking at an extra nucleotide on each side. This expansion aimed to reveal contextual imprints associated with surrounding nucleotides, employing techniques like Latent Dirichlet Allocation (LDA). The methodology centered on curating and quality controlling variant calling samples and replicating mutational signatures using Non-negative Matrix Factorization (NMF). Significant advancements were achieved in this study, particularly in the analysis of Single Base Substitutions (SBS) such as SBS7a, SBS22a, SBS10a, SBS13, and SBS17b, which provided a deeper understanding of their complex etiology through an expanded contextual framework. A novel single-run NMF approach for decomposing mutational signatures was introduced, marking a departure from traditional multi-iteration methods. The future vision involves enhancing the reliability of this single-run approach and integrating user-interactive clustering mechanisms, aiming to connect advanced genomic analysis with personalized cancer treatment strategies.
Not only will this project help us better understand cancer mutational signatures, but it will also have important implications for cancer prevention, treatment, and diagnosis.

ABBREVIATIONS

Latent Dirichlet Allocation (LDA)
Non-negative Matrix Factorization (NMF)
Single Base Substitution (SBS)
Variant Call Format (VCF)

ORGANISATION

This assignment is given by UMCG. The Department of Epidemiology is a major driving force in initiating and conducting life course research and is instrumental to the clinical research within the UMCG's main theme of Healthy Ageing. The research group 'Medical statistics and decision making' is part of the Department of Epidemiology. They focus on developing methods for statistical modeling in clinical and epidemiological studies and analyzing large cohort data. And develop decision analysis techniques to support benefit-risk assessments of medicines and medical decision-making. External guidance is provided by Prof. G.A. Lunter and Dr. Hylke C. Donker.

1. INTRODUCTION

Cancer is a genetic illness characterized by uncontrolled cell proliferation. Cancer cells develop numerous abilities to support this expansion, mostly through genomic alterations. Throughout life, these mutations accumulate as a result of exposure to DNA-damaging mechanisms from both endogenous and external causes [1, 2, 3].
It is becoming increasingly clear that a wide range of mutational pathways contribute to the mutation landscape of cancer. Cancer genomes contain somatic mutations acquired during the normal cell cycle as well as those triggered by cancer-related aberrations of DNA maintenance machinery such as mismatch repair or by carcinogenic exposures such as tobacco smoking, ultraviolet light, and replication stress. Each of these processes frequently results in a particular pattern of changes, known as the mutational signature [4, 5].
The study of mutational changes, specifically the modification of a single letter within the surrounding context, has been the focus of numerous studies. The utilization of this methodology has yielded a meticulously selected collection of mutational signatures as provided by COSMIC [6]. Several approaches have been developed to determine the signatures and infer the attributions.
MutationalPatterns is an R/Bioconductor tool for analyzing single and double base substitutions, as well as small insertions and deletions. It additionally allows the analysis of regional mutation spectra and the discovery of strand asymmetry occurrences [7]. decompTumor2Sig is an R package that can decompose a single tumor genome into a series of Alexandrov- or Shiraishi-type signatures. These signatures represent distinct patterns of somatic mutations in cancer genomes [6]. This enables quantification of the contribution of several mutational processes to the somatic mutations found in a certain tumor [8].
Despite the advancements in understanding mutational signatures and the development of tools to analyze them. There are still several gaps in current knowledge. Most studies have only looked at mutations that include a single letter of context to the left and right. This has hampered the ability to distinguish between other DNA damage sources., Such as ABOPEC3A and ABOPEC3B which require at least one more context letter [9]. These enzymes play a role in DNA editing and mutation processes, with implications in immune defense and cancer development [9]. Large sets of tumor

samples are required for the initial discovery and definition of mutational signatures, which are not always available [8].
Some packages show signature overfitting when determining the contribution of known patterns to a sample, resulting in a disproportionate number of signatures being assigned. It also only allows for the analysis of spectra for mutations in the entire genome, making studying the role of specific genomic elements challenging [7].
Researchers would be able to better distinguish between different mutation origins if they considered additional context around mutations. This could lead to a better understanding of the mutational processes that contribute to cancer. In addition, techniques for mutational signature analysis should be improved. Addressing present techniques' weaknesses such as overfitting and the inability to study specific genetic components. This would significantly improve their accuracy and usefulness. This could allow for more in-depth and advanced studies of mutational phenomena.

2. OBJECTIVE

The previously stated aim will be accomplished of improving the model in order to evaluate mutational signatures that are hierarchically linked, as well as augmenting the context size by incorporating an additional context letter.
The study will involve the curation and quality control of samples and mutations, the replication of mutational signatures through the utilization of NMF, and the subsequent comparison of the outcomes with signatures derived from bayesian neural networks consisting of one and two layers.
Furthermore, the study aims to establish prior probabilities for the expanded mutation representation and compare the identified signatures of higher dimensions with the existing 96 feature representation.

3. THEORY

Mutational signatures are a critical aspect of cancer research, providing a physiological readout of a cancer's biological history [10]. They are the imprints left on the genome by numerous endogenous and exogenous mutagenesis events that have happened throughout a human being's lifetime. These processes can lead to base substitutions, insertions and deletions, or structural modifications, each of which leaves a distinct pattern or signature [10]. SBS7a is a mutational signature with a specific pattern that is easy to identify. C>T mutations and TT dinucleotides on both flanks characterize it [13]. This characteristic is linked to skin malignancies in sun-exposed locations and is most likely caused by UV radiation exposure [12]. Mutational signatures have proven to be a valuable resource in understanding cancer treatment and prevention. For instance, they have been instrumental in studying the molecular processes of DNA damage, DNA repair, and DNA replication [11]. For example, several mutational signatures, such as SBS4 and SBS92, are linked to the same etiology, as are SBS1 and SBS5, both clock-like signatures [4].

$$\sum_{x\in\{A,C,T,G\}}\sum_{y\in\{A,C,T,G\}} P_{\text{SBS1}}xA[C>A]Ay = P_{\text{SBS1}}A[C>A]A \tag{1}$$

This formula expresses the summation over all possible combinations of nucleotides on the left (x) plus the right (y) of the probability of mutation A[C>A]A of SBS1. The approximation of the original probability of mutation A[C>A]A of SBS1 by considering all possible nucleotide combinations in the expanded context.

The context of mutations is crucial in understanding mutational signatures. Increasing the context size is a strategy used to capture complex phenomena. Current SBS signatures are based on max. two flanking nucleotide left and right of the substitution, but this is not enough context to discriminate [9]. This subdivision (1) ensures a more detailed representation, considering the nucleotides on the left and right of the mutations.

4. MATERIALS

The method of analysis is based on an in-depth knowledge of mutational signatures, drawing insights from the COSMIC Mutational Signatures reference collection. The GenomeSigInfer tool `https://github.com/AlfonsoJan/GenomeSigInfer` was developed for processing and analysis. DeepBayesMutSig `https://github.com/AlfonsoJan/DeepBayesMutSig` has detailed analysis and results. All of the libraries and tools used are listed in the appendix (13).
The study "The repertoire of mutational signatures in human cancer" [4] produced Variant Call Format (VCF) files that include both somatic and germline variant calls. VCF files are tab-delimited text files that contain meta-information lines, a header line, and then data lines each containing information about a position in the genome. The study examined 2,658 complete cancer genomes and their corresponding normal tissues from 38 different types of tumors, as part of the Pan-Cancer Analysis of Whole Genomes (PCAWG). The resulting datasets can be accessed on Synapse: `https://www.synapse.org/#!Synapse:syn11801870` The VCF files underwent processing to generate SBS files.
SBS represent a category of mutations characterized by the replacement of one nucleotide with another in the DNA sequence. SBS files function as datasets that encapsulate comprehensive count data pertaining to individual base substitutions within the genomic sequences derived from the mutations recorded in the VCF file.

$$M \epsilon R_{+}^{K\times G} \tag{2}$$

This domain expresses the SBS files of non-negative integers with dimension K × G, where K is the number of mutation types and G is the number of samples.

The COSMIC mutational signatures are a set of mutational properties for each signature. Each row represents the probability of each mutation signature. Each column represents the likelihood of each characteristic mutation [6]. The sum along the mutations classes sum to one.

5. METHODS

The SigProfiler framework has been systematically designed to analyze mutational signatures in genomic data. The methodology is based on the use of NMF, which is essential in the decomposition of high-dimensional mutational matrices into non-negative basis matrices [4]. The framework minimizes a generalized Kullback-Leibler divergence that is strictly bound for non-negativity in each iteration of NMF [4]. This procedure reveals unique mutational signatures and their contributions to each sample [4]. Furthermore, SigProfiler employs a hierarchical de novo extraction technique,

which improves the precision of its analysis [4]. This strategic approach allows for a thorough examination of mutational patterns, allowing for the detection of small alterations in the genomic landscape.

$$C > \{A, G, \text{ or } T\} \quad \text{and} \quad T > \{A, G, \text{ or } C\} \tag{3}$$

The 6 possible pyrimidine single nucleotide variants.

5.1. CREATING SBS FILES

SBS files are currently based on an upper limit of two flanking nucleotides to the left and right of the substitution. GenomeSigInfer can create SBS files with increased context using a function called: 'GenomeSigInfer.sbs.SBSMatrixGenerator()'. This function run a three-step procedure:

1. In the process of crafting SBS files, the initial step involves the filtration of VCF files. This filtration is conducted based on specific criteria, including alignment with a designated reference genome and identification of mutations falling within the category of the six possible pyrimidine variants (3). The reverse complement symmetry mutation was taken if he mutation is not a pyrimidine mutations.
2. The parsed VCF files are organized by chromosome, and a comprehensive SBS file is generated, encompassing all mutations present in the VCF file. This file is structured with the maximum context available.
3. The following step involves the compression of the SBS matrix, gradually reducing the context until it reaches a minimum of three (1). Each resultant matrix is then saved in the efficient Parquet file format.

5.2. OPTIMAL NMF PARAMETERS

In order to identify the most effective NMF initialization and beta loss parameters, a thorough exploration was conducted using the '96' context file as input. A decomposition was performed using the 'SigProfilerAssignment' package on every conceivable combination of NMF initialization and beta loss parameters.
Cosine similarity between the larger context and the '96' were derived based on the decomposition results. In order to assess the effect of these parameter selections on the accuracy and reliability of the mutational signature extraction procedure.

5.3. RUN NMF

After determining the best NMF initialization and beta loss parameters. The SBS files were then subjected to NMF analysis. Important preprocessing stepswere carried out prior to the start of the NMF algorithm. They are as follows:

1. The cutoff value is computed based on the given context, serving as a threshold for data normalization. Specifically, when the context value is '96', the cutoff is determined by multiplying this context value by 1000, resulting in a cutoff of 9600.
2. Normalizing the data by calculating the sum of each column and adjusting small values for stability.

The Preprocessing class plays a crucial role in normalizing the data to ensure its quality and readiness for analysis. Once the data is normalized and brought into an optimal state, it sets the stage for the initiation of the NMF algorithm on this preprocessed dataset.

5.4. DECOMPOSE

Following the completion of the NMF process on the normalized genomic data, the linear sum assignment approach was used to determine optimal column assignments between each NMF result and the COSMIC dataframe. This phase was critical for assigning the correct mutational signature to each row in the NMF dataframe, ensuring exact mutational signature alignment across multiple contexts.
Following this alignment, the Jensen Shannon Distance and Cosine Similarity metrics were generated to quantify the dissimilarity and similarity, respectively, between the 96-context mutational signatures and those produced from larger contexts. These measures were useful in giving a quantitative assessment of the interactions between various genomic contexts, putting insight on the complex differences in the mutational landscape.

5.5. SIGNATURE PLOTS

Barplots were created for each decomposed matrix in different genomic contexts to visually show the correlations and trends identified by comparing mutational signatures. These visualizations presented a thorough and context-specific view, revealing the distribution of mutational types within each genomic context.
This strategy was expanded in the event of additional context files by constructing customized bars for each situation. Bars were color-coded in these enhanced visualizations to effectively reflect the proportion contribution of each nucleotide. This improved graphical depiction enabled a more sophisticated view of the complex mutational landscape, highlighting particular nucleotide contributions within various genomic settings. Using mean field distribution of the letters for the tri-nucleotide context (4, 5).

$$p(M|N, P \to Q, R) = \frac{p(M, N, P \to Q, R)}{p(N, P \to Q, R)} \tag{4}$$

$$\frac{p(M, N, P \to Q, R)}{p(N, P \to Q, R)} = \frac{\sum_{S \in \{A,C,T,G\}} p(M, N, P \to Q, R, S)}{p(N, P \to Q, R)} \tag{5}$$

This formula expresses how the distribution is calculated for the left letter in a 5-context mutation.

5.6. META SIGNATURES

Meta signatures are constructed by combining the topics of different chains to its centroid by repeatedly solving the optimal transport problem for the Jensen-Shannon distance (JSD) using the Hungarian algorithm until the centroid converged in terms of silhouette score [16]. The multinomial belief network is a bayesian deep belief network that uses multinomial-distributed variables as output [16].
In the Multinomial Generative Model, the variable $a_{vj}^{(T+1)}$ represents a specific parameter at level $T + 1$. The parameters $\theta_{vj}^{(t)}$ are distributed according to a Dirichlet distribution, highlighting the model's probabilistic approach. These parameters are crucial at each time step t and contribute to the model's dynamics. The summation variable $a_{vj}^{(t)}$ is a compound parameter, indicating interactions between various model elements. Finally, the observable variables $\{x_{vj}\}_v$ follow a multinomial distribution, emphasizing the model's focus on count data analysis. Each of these components plays a vital role in the model's ability to process and analyze complex datasets.

$$a_{vj}^{T+1} = rv, \tag{6}$$

$$\left\{\theta_{vj}^{(t)}\right\}_v \sim Dir(\left\{c^{t+1}a_{vj}^{(t+1)}\right\}_v), t = T, ..., 1 \tag{7}$$

$$a_{vj}^{(t)} = \sum_{k=1}^{K_t} \phi_{vk}^{(t)}\theta_{kj}^{(t)}, t = T, ..., 1 \tag{8}$$

$$\{X_{vj}\}_v \sim Mult(n_j, \{a_{vj}^1\}_v). \tag{9}$$

These formulas expresses the generative model of multinomial belief networks that uses Dirichlet latent states.

5.7. CLUSTER

In pursuit of simplifying the representation of the genomic context, a method was constructed based on the nucleotide. There are also letters that represent ambiguity which are used when more than one kind of nucleotide could occur at that position [15].
A way to cluster the existing 9-context structure into a more condensed format, resulting in the same number of features as the original 7-context. By employing a clustering approach that combines the 2 nucleotides on the far left and far right into one of the following categories:

- "W" (Weak) or "S" (Strong)
- "M" (aMino) or "K" (Keto)
- "R" (puRine) or "Y" (pYrimidine)

It effectively reduces the dimensionality of the context while still preserving essential structural information.

$$\begin{aligned} &NNN[C > \{A, G, T\}]NNN \text{ or} \\ &NNN[T > \{A, C, G\}]NNN \end{aligned} \tag{10}$$

This formula expresses all the possibilities for the 7-context mutation. 4x4x4x6x4x4x4=24576 total combinations.

$$\begin{aligned} &\{W, S\}\{W, S\}NN[C > \{A, G, T\}]NN\{W, S\}\{W, S\} \text{ or} \\ &\{W, S\}\{W, S\}NN[T > \{A, C, G\}]NN\{W, S\}\{W, S\} \end{aligned} \tag{11}$$

This formula expresses all the possibilities for the clustered 9-context mutation based on strength. 2x2x4x4x6x4x4x2x2=24576 total combinations.

A similar perplexity between the two structures would indicate that the clustering effectively retains the essential information present in the original context, providing a more streamlined yet equally informative representation.

Perplexity serves as a quantitative measure, evaluating how effectively the NMF model captures complex patterns within genomic data. It is mathematically represented as:

$$\mathcal{L} = \exp\left(-\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{x_{ij}\log p_{ij}}{\sum_{j=1}^{n} x_{ij}}\right) \tag{12}$$

How 'per word perplexity' is calculated

In this equation, L is the perplexity, and the exponential function (exp) is applied to the average of the logarithmic likelihoods across all observations. Here, m denotes the number of samples in the dataset, and n represents the number of features. For each observation i and feature j, x_{ij} is the observed data value, and $\log p_{ij}$ is the natural logarithm of the predicted probability or model parameter. The equation's structure, especially the normalization by the sum of x_{ij} across all features for each observation, ensures that the perplexity is a balanced measure of how well the model's predictions align with the observed data. Lower values of L indicate a more accurate model, suggesting a better fit to the complex genomic data patterns.

6. RESULTS

6.1. DATA ACQUISITION AND PREPARATION

Modifications to the PCAWG files were required to obtain a reliable SBS dataset. The VCF files comprises a total of 30,356,419 mutations observed across 2,658 complete cancer genomes and their respective normal tissues spanning 38 different types of tumors. The original VCF file format was not maintained, and there was a lack of a codebook containing additional information on the attributes. Detailed information about the 10 columns in the VCF files can be found in the appendix (table 14).
The dataset requires filtration based on specific criteria. Only mutations classified as 'SNP' and 'SNV' in the 'Mutation Type' column will be retained. Additionally, the reference genome must match the specified value, which is GRCh37 in this study.

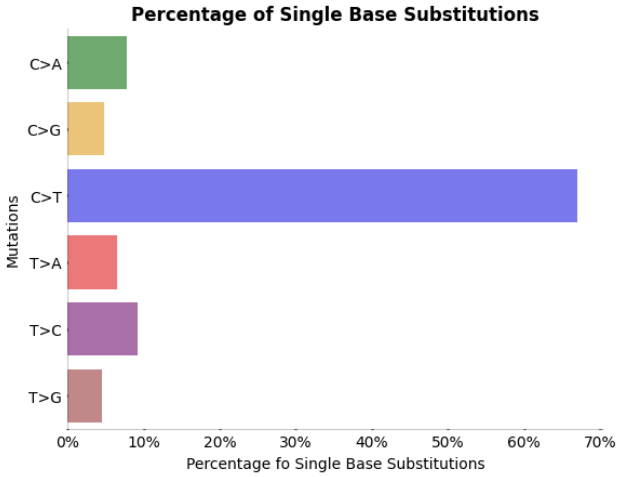


Figure 1: *Barplot of the distribution of nucleotide substitutions in the filtered dataset.*

Among the various mutation types, C>T emerges as the most prevalent, constituting a substantial 67.08% (figure 1) of the total mutations. This finding underscores the significance of C>T transitions in the genomic alterations observed. Furthermore, the observed frequencies of other substitution types, including C>A, C>G, T>A, T>C, and T>G, provide a comprehensive view of the mutational spectrum. The percentages assigned to each mutation type offer a nuanced understanding of their relative prevalence, contributing to a more comprehensive characterization of the mutational patterns in the studied dataset. These results contribute valuable information that may guide further investigations into the underlying biological mechanisms driving these specific nucleotide changes.

6.2. CREATION OF SBS FILES

Transitioning from filtered VCF files to SBS files is a pivotal step. The process involves extracting and organizing information from the VCF files. First a 'maximum context' SBS file created with all zeros for each sample.

Mutation Type	...PTC-73C	...PTC-7C	...PTC-88C
AAA[C>A]AAA	0	0	0
AAA[C>A]AAC	0	0	0
AAA[C>A]AAG	0	0	0
AAA[C>A]AAT	0	0	0
AAA[C>A]ACA	0	0	0

Table 1: The first 5 rows and 4 columns of the initialized SBS dataframe.

This dataframe (table 1) has 4x4x4x6x4x4x4=24576 different mutation types. 64 (4x4x4) possible starting nucleotides x 6 pyrimidine variants x 64 (4x4x4) possible ending dinucleotides = 24576 total combinations (10).

For each mutation in the filtered VCF file, the genomic context, encompassing three nucleotides on each side, was cross-referenced with the reference genome, and the corresponding count in the initialized SBS dataframe was incremented by 1, thereby progressively building a comprehensive record of single base substitution events across the analyzed samples.

Mutation Type	...PTC-73C	...PTC-7C	...PTC-88C
AAA[C>A]AAA	1	4	1
AAA[C>A]AAC	0	0	0
AAA[C>A]AAG	0	1	0
AAA[C>A]AAT	0	1	0
AAA[C>A]ACA	0	0	0

Table 2: The first 5 rows and 4 columns of the SBS dataframe. After counting all the mutations in the VCF files.

This (table 2) encapsulates the cumulative count of mutation occurrences for specific mutation types across three distinct samples: Thy-AdenoCa::PTC-73C, Thy-AdenoCa::PTC-7C, and Thy-AdenoCa::PTC-88C. Each row corresponds to a unique mutation type, signified by the altered nucleotide sequence enclosed in square brackets. The incremented values in the table signify the frequency of each mutation type within the genomic context of the respective samples.

The dataframe has to be compress to a context smaller. Transitioning from the broader A,C,T,GAA[C>A]AAA,C,T,G to the representation AA[C>A]AA (1). This compression involves aggregating the sum of all combinations of nucleotides on both the left and right.

Mutation Type	...PTC-73C	...PTC-7C	...PTC-88C
AA[C>A]AA	3	7	3
AA[C>A]AC	0	1	0
AA[C>A]AG	0	1	0
AA[C>A]AT	0	1	0
AA[C>A]CA	0	2	0

Table 3: The first 5 rows and 4 columns of the SBS dataframe. After compressing the dataframe one context smaller. To the 5-context.

This dataframe (table 3) has 4x4x6x4x4=1536 different mutation types. 16 (4x4) possible starting nucleotides x 6 pyrimidine variants x 16 (4x4) possible ending dinucleotides = 1536 total combinations.

In the pursuit of achieving the smallest possible genomic context, the representation was further compressed from A,C,T,GA[C>A]AA,C,T,G to the most succinct form, A[C>A]A.

Mutation Type	...PTC-73C	...PTC-7C	...PTC-88C
A[C>A]A	8	35	6
A[C>A]C	2	19	2
A[C>A]G	0	20	2
A[C>A]T	2	10	1
A[C>A]C	5	36	6

Table 4: The first 5 rows and 4 columns of the SBS dataframe. After compressing the dataframe to the smallest context. The 3-context.

The refined SBS dataframe (table 4), revealing the cumulative counts of mutation occurrences for specific mutation types across three samples: Thy-AdenoCa::PTC-73C, Thy-AdenoCa::PTC-7C, and Thy-AdenoCa::PTC-88C. This dataframe has 4x6x4=96 different mutation types. 4 possible starting nucleotides x 6 pyrimidine variants x 4possible ending dinucleotides = 96 total combinations.

For quality assurance and validation purposes, the $'WGS_{O}ther.96.csv'$ and $'WGS_{P}CAWG.96.csv'$ datasets served as crucial benchmarks. These datasets played a pivotal role in verifying the accuracy of the GenomeSigInfer tool's output by ensuring a precise match with the information contained in the 2 *96.csv files.

6.3. NMF PARAMETER SELECTION

In order to determine the optimal beta loss and NMF initialization parameters, an in-depth analysis was carried out utilizing the $'96'$ context PCAWG file as input. All possible combinations of NMF initialization and beta loss parameters were systematically executed and decomposed using the $'SigProfilerAssignment'$ tool. The output consisted of a dataframe representing the identified signatures for each NMF component.

sigprofiler	None_frobenius	None_kullback-leibler	None_itakura-saito
SBS5	SBS5	SBS5	SBS17b
SBS5	SBS17b, SBS125	SBS39, SBS51	SBS5
SBS17b	SBS5	SBS5, SBS17b	SBS35
SBS5	SBS5, SBS86	SBS1, SBS32	SBS1, SBS5, SBS39
SBS35	SBS7c	SBS7c	SBS39

Table 5: The identified mutational signatures for each component.

The table (table 5) presents the inferred mutational signatures resulting from NMF decomposition, showcasing the association of each signature with specific NMF components under different beta

loss metrics, such as Frobenius, Kullback-Leibler, and Itakura-Saito.
The next step in the analysis involves calculating the cosine similarity scores for various parameter configurations, a crucial measure to assess the degree of resemblance between the decomposed mutational signatures and the original ones.

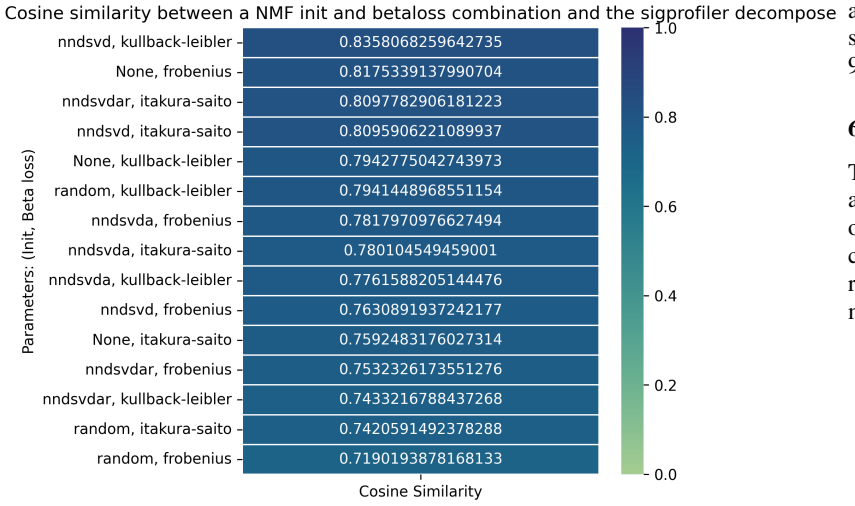


Figure 2: *Heatmap of the average cosine similarity of a distinct parameter setting, achieved for the decomposition of mutational signatures.*

The cosine similarity metric measures the similarity between two vectors, in this case, representing mutational signatures. Higher cosine similarity values indicate a more robust match between the decomposed signatures and the original ones. Higher cosine similarity scores, such as the notable 0.836 achieved with the 'nndsvd, kullback-leibler' configuration, signify a more faithful representation of the underlying mutational patterns.

6.4. RUN NMF

Utilizing the insights gained from the cosine similarity analysis, the next crucial step involves running the NMF algorithm with the identified optimal beta loss and initialization parameters.
A cut-off value is computed based on the shape of the data. For the 96 context the cut-off is 9600. The sum of each sample cannot be larger than this cut-off. Ensuring a robust threshold for subsequent normalization steps.

$$\begin{pmatrix} 0 & 0 & 0 & 8 & 35 & 6 \\ 0 & 0 & 0 & 2 & 26 & 2 \\ 0 & 0 & 0 & 0 & 16 & 2 \\ 0 & 0 & 0 & 3 & 27 & 1 \\ 0 & 0 & 0 & 1 & 29 & 2 \\ 0 & 0 & 0 & 0 & 28 & 1 \end{pmatrix} \tag{13}$$

The first 6 columns and rows of the array before the data preprocessing.

$$\begin{pmatrix} 1.02 \times 10^{-5} & 1.56 \times 10^{-5} & 6.18 \times 10^{-5} & 6.00 \times 10^{0} & 3.90 \times 10^{0} & 7.00 \\ 1.02 \times 10^{-5} & 1.56 \times 10^{-5} & 6.18 \times 10^{-5} & 1.00 \times 10^{0} & 1.90 \times 10^{1} & 1.00 \\ 1.02 \times 10^{-5} & 1.56 \times 10^{-5} & 6.18 \times 10^{-5} & 1.00 \times 10^{-4} & 1.30 \times 10^{4} & 1.00 \\ 1.02 \times 10^{-5} & 1.56 \times 10^{-5} & 6.18 \times 10^{-5} & 2.00 \times 10^{0} & 1.80 \times 10^{1} & 1.00 \\ 1.02 \times 10^{-5} & 1.56 \times 10^{-5} & 6.18 \times 10^{-5} & 1.00 \times 10^{0} & 1.90 \times 10^{1} & 4.00 \\ 1.02 \times 10^{-5} & 1.56 \times 10^{-5} & 6.18 \times 10^{-5} & 1.00 \times 10^{-4} & 2.60 \times 10^{1} & 3.00 \end{pmatrix} \tag{14}$$

The first 6 columns and rows of the array after the data preprocessing.

The initial numpy array (13) exhibits the raw mutational signature data in its unprocessed state, showcasing the diverse range of values across different samples. Following the meticulous preprocessing steps, as illustrated in matrix (14), the same numpy array undergoes normalization, cutoff application, and the introduction of random noise. The resultant array reflects a more refined and robust dataset, where the values have been carefully adjusted to adhere to the specified cutoff limit, ensuring that the sum of each sample does not exceed the established threshold of 9600.

6.5. DECOMPOSE

The NMF process conclusion signifies a crucial step as its results are aligned with the COSMIC dataframe. This alignment is pivotal for evaluating mutational signatures across diverse genomic contexts. Table 6, extracted from the COSMIC 3.4 file, visually represents these outcomes. It provides insights into the distinctive mutational patterns within the analyzed genomic settings.

Type	SBS1	SBS2	SBS3
A[C>A]A	0.000886157	5.80E-07	0.020808323
A[C>A]C	0.002280405	0.000148004	0.016506603
A[C>A]G	0.000177031	5.23E-05	0.0017507
A[C>A]T	0.001280227	9.78E-05	0.012204882
A[C>G]A	0.00186033	2.23E-16	0.019707883

Table 6: The first 5 rows and 4 columns of the COSMIC 3.4 file.

This (table 6) encapsulates the likelihood of various mutation signatures, such as SBS1, SBS2, and SBS3, providing a nuanced understanding of the probability associated with each mutation type. Each row in the COSMIC file signifies the likelihood of a specific mutation signature.

In the pursuit of aligning mutational signatures derived from the smallest genomic context (96) with those in the COSMIC database, a crucial step involves employing linear sum assignment techniques. This process aims to establish optimal column assignments between the 96-context dataframe and the corresponding COSMIC dataset.

Type	SBS7a	SBS4	SBS2
A[C>A]A	0	0.01114092839	0
A[C>A]C	0	0.0616319348	0.00076242742
A[C>A]G	0	0.02172093431304	4.38690545447e-06
A[C>A]T	0	0.0023001320906	0.00059309091
A[C>G]A	0	0.01911412492	0

Table 7: The first 5 rows and 4 columns of the decomposed 96 context dataframe.

This table (7) encapsulates the likelihood of various mutation signatures, such as SBS7a, SBS4, and SBS2, providing a nuanced understanding of the probability associated with each mutation type.

For the larger genomic contexts, a meticulous decompression process was undertaken to align them with the standardized 96-context. This alignment was achieved by employing linear sum assignment techniques, comparing each larger context to

the original 96-context. The aim was to ensure uniformity in the representation of mutational signatures across different genomic contexts, thereby facilitating a consistent and meaningful comparison.

Following the alignment of larger genomic contexts with the standardized 96-context through linear sum assignment, the analysis proceeded to quantify the dissimilarity and similarity between mutational signatures. Two key metrics, namely Jensen Shannon Distance and Cosine Similarity, were employed for this purpose. These metrics provided quantitative assessments of the relationships between the mutational signatures derived from different genomic contexts. Jensen Shannon Distance measured the dissimilarity, offering insights into the unique characteristics of each mutational pattern. In contrast, Cosine Similarity quantified the similarity, highlighting shared features across mutational signatures.

The visual representation (figure 3) underscores the remarkable consistency in cosine similarity across enlarged genomic contexts, specifically highlighting the elevated similarities observed in mutational signatures such as SBS22a, SBS7a, SBS13, SBS10a, and SBS17b. This shared pattern suggests a robust preservation of these mutational features, further emphasizing their stability and prevalence across varying genomic scales. Complementing this observation, (figure 4) provides a nuanced perspective on the dissimilarity aspects, showcasing low Jensen Shannon dissimilarity for mutational signatures such as SBS22a, SBS7a, SBS97, SBS13, and SBS17 across both enlarged contexts. This convergence in dissimilarity metrics aligns with the identified similarities, reinforcing the notion of conserved mutational patterns.

6.6. SIGNATURE PLOTS

For insightful visualization barplots for each decomposed matrix were created. Offering a comprehensive and visually intuitive representation of the correlations and trends identified through the comparison of mutational signatures. These tailored visualizations, specific to different genomic contexts, provided nuanced insights into the distribution and prevalence of distinct mutational types within each setting.
Expanding on this approach to accommodate additional context files, the strategy evolved to construct customized bars for each unique genomic situation. In these refined visualizations, bars were intricately color-coded to vividly represent the proportional contribution of each nucleotide. This color-coded graphical representation elevated the interpretability of the complex mutational landscape, providing a nuanced view of the distinct nucleotide contributions within diverse genomic settings. The use of color-coded bars facilitated a detailed understanding of mutational patterns, allowing for comparative analyses across multiple contexts and enriching insights into the varied and context-specific dynamics of genomic mutations.

SBS2 (figures 5, 6, 7) and SBS13 (figures 8, 9, 10) Are associated with the activation of AID/APOBEC cytidine deaminases in cancer, possibly due to previous viral infection, retrotransposon jumping, or tissue inflammation. SBS13 mutations are likely generated by error-prone polymerases replicating across abasic sites generated by base excision repair removal of uracil [17]. All A3 and AID enzymes deaminate the second cytidine in either CC or TC dinucleotide motifs [18]. However, sequence analysis has determined that not every cognate dinucleotide motif is deaminated, which suggests that the nucleotides flanking the dinucleotide motifs play a role in substrate recognition [18]. The pattern in SBS2 escpecially ion the context of 7, where the majority of the nucleotiudes are thymine, could be explained by the binding residue of A3F-CTD poly T ssDNA [18].

SBS22a (figures 11, 12, 13) is a mutational signature associated with exposure to aristolochic acid (AA), a compound found in certain plants and herbal remedies that has been linked to cancer [6]. Is a signatures characterized by T>A [19]. Aristolochic-acid exposure and similar patterns SBS22 is due to aristolochic acid [19]. The mutational signature of AA-induced DNA damage is characterized by a predominance of A:T-to-T:A transversions, a relatively unusual type of mutation that is infrequently seen in other types of cancer, including those caused by other carcinogens. These mutations concentrate at splice sites, causing the inappropriate inclusion or exclusion of entire exons in the resulting mRNA [20].

6.7. META SIGNATURE

Meta-mutational signatures, or meta-signatures for short, are signatures of signatures. That is, they describe mutational signatures that typically go together in a sample. Those reported below are based on the COSMIC v3.4 single base substitution signatures and are inferred using the multinomial belief network.

The outcomes of the MBN model applied to the mutation dataset. Showcase its proficiency in deriving resilient consensus metasignatures that effectively encapsulate the concurrent presence of mutational signatures across patients. Designated as $k = M_1, ..., M_4$, these four identified metasignatures offer valuable insights into the fundamental mutagenic processes underlying cancer.

The statistics (table 15) shows the mean, standard deviation, and highest density interval (HDI) for four identified metasignatures. The attributions (for both signature and meta signatures) are normalised to one, denoted as M_1, M_2, M_3, and M_4. The mean values for M_1, M_3, and M_4 are 2.97, 2.95, and 3.04, respectively, while the mean value for M_2 is 2.58. The HDI for M_1 ranges from 2.90 to 3.11, for M_2 ranges from 1.99 to 3.06, for M_3 ranges from 2.85 to 3.06, and for M_4 ranges from 2.95 to 3.15.

M_1 (figures 14a, 14b) describes the co-occurrence of SBS2 and SBS13 are linked to the APOBEC family of cytidine deaminases, which are recognized for instigating distinctive mutational patterns, notably involving C-to-T and C-to-G substitutions [18], within the genomic DNA of diverse organisms. SBS7a is linked to ultraviolet (UV) light exposure, which catalyzes the formation of cyclobutane pyrimidine dimers [17].
Meta signature M_2 (figures 14c, 14d) captures SBS5. This is present in all cancer types and is typically one of the most prominent signatures [1]. Meta signature 2 describes signatures with a strong transcriptional strand bias (SBS5, SBS8, SBS12, SBS16, SBS92, and SBS22) with the exception of SBS40 [6]. Its primary constituent SBS12 is believed to be related to transcription-coupled nucleotide excision repair. The second

largest contributor, SBS40, is a spectrally flat, latereplicating, signature with spectral similarities to SBS5 (both are related to age) and is believed to be linked to SBS8 [6].
Meta signature M_3 (figures 14e, 14f) describes the co-occurrence of several, seemingly disparate, mutational signatures of known and unknown aetiology. Of known cause are, SBS7b, linked to ultraviolet light. SBS4 is associated with tobacco exposure and is found in liver. SBS8 is characterized by C > T and C > G mutations and is found at TpCpN trinucleotides.
Finally meta signature M_4 (figures 14g, 14h) captures multiple signatures. SBS10a is a mutational signature that is often associated with large numbers of somatic mutations, leading to samples with these signatures being termed hypermutators. The etiology of SBS10a is possibly driven by the POLE (polymerase epsilon) exonuclease domain mutations [17, 21]. SBS10a split from SBS7, SBS10a is characterized by C>A mutation and TT dinucleotides [4].

6.8. CLUSTER

A unique technique for representing genomic contexts was developed, focused on the nucleotide level and includes characters that signify ambiguity when numerous nucleotide options exist at a particular site [15]. The objective was to cluster the existing 9-context structure into a more condensed format, ultimately aligning with the original 7-context structure while preserving essential structural information. This clustering approach involves combining the two nucleotides on the far left and far right into distinct categories: "W" (Weak) or "S" (Strong), "M" (aMino) or "K" (Keto), and "R" (puRine) or "Y" (pYrimidine). By doing so, the dimensionality of the context is effectively reduced, offering a more simplified yet equally informative representation.

Mutation Type	...PTC-73C	...:PTC-7C	...PTC-88C
AAAA[C>A]AAAA	0	2	1
AAAA[C>A]AAAC	0	0	0
AAAA[C>A]AAAG	0	0	0
AAAA[C>A]AAAT	0	0	0
AAAA[C>A]AACA	0	0	0

Table 8: The first 5 rows and 4 columns of the 9-context SBS dataframe.

This table (table 8) illustrates the mutation counts for different mutation types within the 9-context SBS dataframe across three specific samples (Thy-AdenoCa::PTC-73C, Thy-AdenoCa::PTC-7C, and Thy-AdenoCa::PTC-88C) in the first 5 rows and 4 columns. Each row corresponds to a specific mutation type, and each column represents a distinct sample, providing a snapshot of the mutational landscape within this condensed context.

Mutation Type	...PTC-73C	...PTC-7C	...PTC-88C
KKAA[C>A]AAKK	0	0	0
KKAA[C>A]AAKM	0	0	1
KKAA[C>A]AAMK	0	0	0
KKAA[C>A]AAMM	0	0	0
KKAA[C>A]ACKK	0	0	0

Table 9: The first 5 rows and 4 columns of the clustered 9-context SBS dataframe based on pyrimidine/purine.

Mutation Type	...PTC-73C	...:PTC-7C	...PTC-88C
SSAA[C>A]AASS	1	0	1
SSAA[C>A]AASW	0	0	0
SSAA[C>A]AAWS	0	1	0
SSAA[C>A]AAWW	0	0	0
SSAA[C>A]ACSS	0	0	0

Table 10: The first 5 rows and 4 columns of the clustered 9-context SBS dataframe based on strength.

Mutation Type	...PTC-73C	...:PTC-7C	...PTC-88C
RRAA[C>A]AARR	1	3	1
RRAA[C>A]AARY	0	0	0
RRAA[C>A]AAYR	0	0	0
RRAA[C>A]AAYY	0	0	0
RRAA[C>A]ACRR	0	0	0

Table 11: The first 5 rows and 4 columns of the clustered 9-context SBS dataframe based on structure.

These three tables (tables 9, 10, 11) provide insights into the mutational landscape within the 9-context SBS dataframe, each employing a different clustering approach to condense the data while retaining essential structural information.
Table 9, based on pyrimidine/purine clustering, showcases the mutation counts for specific mutation types in the clustered 9-context SBS dataframe across three samples (Thy-AdenoCa::PTC-73C, Thy-AdenoCa::PTC-7C, and Thy-AdenoCa::PTC-88C). The rows represent mutation types, while each column corresponds to a distinct sample, offering a condensed perspective on the mutational signatures.
Table 10 utilizes a clustering method based on nucleotide strength, demonstrating mutation counts in the clustered 9-context SBS dataframe. The rows represent mutation types clustered by nucleotide strength, and columns represent distinct samples. This clustering provides a nuanced view of mutational patterns based on the strength of nucleotides.
Table 11, based on structural clustering, highlights mutation counts in the clustered 9-context SBS dataframe. The rows represent mutation types clustered by structural features, and columns represent samples. This approach condenses the data based on the structural characteristics of the mutation types, offering a simplified yet informative representation of the mutational landscape.

Context	Perplexity
7-context	$652^{+0.03}_{-0.03}$
9-context amino type clustered	$667^{+0.02}_{-0.03}$
9-context nucleotide strength clustered	$654^{+0.03}_{-0.03}$
9-context structure type clustered	$667^{+0.03}_{-0.03}$

Table 12: The perplexity of the 7-context files and the clustered 9-context files.

A lower perplexity generally indicates a better fit of the model to the dataset. The 7-context perplexity values is $652^{+0.03}_{-0.03}$ (table 12), revealing diverse levels of model performance in representing patterns within this context. The 9-context amino type clus-

tered scenario exhibits a perplexity $667^{+0.02}_{-0.03}$. Signifying a heightened complexity in capturing patterns specific to amino acid types. The same for the 9-context structure type clustered scenario. The perplexity for the 9-context amino type clustered the perplexity is $654^{+0.03}_{-0.03}$.

7. DISCUSSION & CONCLUSION

This study focused on improving the statistical model for analysing cancer mutational signals using deep bayesian neural networks. The goal was to improve the representation of mutations by increasing the context around them and use advanced computational techniques like NMF. The findings showed an improvement in identifying and interpreting complex mutational signatures. Notably, adding additional context to mutations resulted in more exact distinction between diverse DNA damage sources. This improved accuracy and depth of analysis has important implications for cancer research, diagnosis, and treatment options.

The study successfully created SBS matrices tailored to a user-defined context size. kullback-leibler divergence, in line with the SigProfiler technique [4], is the best method for NMF. The model excelled at decomposing SBS files into separate mutational signatures by linear sum assignment. The SBS22a, SBS7a, SBS10a, SBS13, and SBS17b signatures show a higher degree of cosine similarity in higher context environment. Similarly, SBS22a, SBS7a, SBS97, SBS13, and SBS17 have reduced Jensen-Shannon dissimilarities. Furthermore, study of signature plots and meta signatures offered more information. Investigating the SBS7a, SBS22a, SBS10a, SBS13, and SBS17b produced signatures has uncovered important information regarding their genesis. SBS22a, SBS10a, SBS13, and SBS17b are known to benefit from higher-level analysis, as this study has successfully proved. SBS7a was not previously reported to benefit from having a larger context. This signature is characterized C>T mutations and TT dinucleotides [13]. In both context the most dominant nucleotide is thymine.

While cluster analysis found that bigger context data presented increased perplexity, indicating complexity and diversity, 9-context nucleotide strength clustered data had the lowest perplexity across clustered types. I suggest that you do not lose too much information. Comparatively, the method of decomposing mutational signatures in this study presents a distinct approach from that used by SigProfiler. SigProfiler normally runs NMF 1024 times to average the findings [4]. This study used a single run of NMF for decomposition. Although it brings a new angle to mutational signature research, this methodological difference also adds a possible flaw. The averaging process in SigProfiler helps in stabilizing the results against random fluctuations in the data, a feature that the single-run method may lack.

The aim of this research was to refine the analysis of cancer mutation signatures through advanced statistical modeling. This goal has been achieved, resulting in a model that provides more accurate and detailed insights into mutation patterns in cancer genomes. In conclusion, while this study represents a step forward in mutation signature analysis, it also sets the stage for further research. The pursuit of more refined, efficient, and reliable methods will continue to be crucial in the rapidly evolving landscape of cancer genomics, with the ultimate goal of translating these scientific advancements into tangible clinical benefits.

8. REFERENCES

[1] Hanahan, D., & Weinberg, R. (2000). "The Hallmarks of Cancer." Cell,100(1), 57-70.

[2] Hanahan, D., & Weinberg, R. (2011). "Hallmarks of Cancer: The Next Generation." Cell, 144(5), 646-674.

[3] Hanahan, D. (2022). "Hallmarks of Cancer: New Dimensions." Cancer Discovery, 12(1), 31-46.

[4] Alexandrov, L., et al. Stratton, M. R. (2020). "The repertoire of mutational signatures in human cancer." Nature, 578(7793), 94-101.

[5] Huang, X., Wojtowicz, D., & Przytycka, T. M. (2018). "Detecting presence of mutational signatures in cancer with confidence." Bioinformatics, 34(2), 330-337.

[6] COSMIC. (n.d.). Wellcome Sanger Institute. https://cancer.sanger.ac.uk/cosmic.

[7] Kim, Y., Leiserson, M. D. M., Moorjani, P., Sharan, R., Wojtowicz, D., & Przytycka, T. M. (2021). "Mutational Signatures: From Methods to Mechanisms." Annual Review of Biomedical Data Science, 4, 189-206.

[8] Krüger, S., & Piro, R. M. (2019). "decompTumor2Sig: identification of mutational signatures active in individual tumors." BMC Bioinformatics, 20(4), 152.

[9] Chan, K., et al. (2015). "An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers." Nature Genetics, 47(9), 1067-1072.

[10] Helleday, T., Eshtad, S., & Nik-Zainal, S. (2014). "Mechanisms underlying mutational signatures in human cancers." Nature Reviews. Genetics, 15(9), 585-598.

[11] Steele, C. D., Pillay, N., & Alexandrov, L. B. (2022). "An overview of mutational and copy number signatures in human cancer." The Journal of Pathology, 257(4), 454-465.

[12] Jialong, L., Jing, B., Tao, P., Haozhe, Z., Yueying, G., Jing, G., Qi, X., Juan, X., Yongsheng, L., & Xia, L. (2022). "Clinical and genomic characterization of mutational signatures across human cancers." International Journal of Cancer, 152(8), 1613-1629.

[13] Machado, H. E., etal. (2022). "Diverse mutational landscapes in human lymphocytes." Nature, 608(7924), 724-732.

[14] J.A. Busker. GenomeSigInfer. (2024). GenomeSigInfer.

[15] Nomenclature Committee of the International Union of Biochemistry (NC-IUB). (1986). Proceedings of the National Academy of Sciences of the United States of America, 83(1), 4-8.

[16] H. C. Donker, D. Neijzen, G. A. Lunter. (2023). " Multinomial belief networks." Cornell University.

[17] Boot, A., Ng, A. W. T., Chong, F. T., Ho, S., Yu, W., Tan, D. S. W., Iyer, N. G., & Rozen, S. G. (2020). "Characterization of colibactin-associated mutational signature in an Asian oral squamous cell carcinoma and in other mucosal tumor types." Genome Research, 30(6), 803-813.

[18] Silvas, T. V., & Schiffer, C. A. (2019). "APOBEC3s: DNA-editing human cytidine deaminases." Protein Science: A Publication of the Protein Society, 28(9), 1552-1566.

[19] Degasperi, A. et al. (2022). "Substitution mutational signatures in whole-genome-sequenced cancers in the UK population." Science (New York, N.Y.), 376(6591).

[20] Poon, S. et al. (2013). "Genome-Wide Mutational Signatures of Aristolochic Acid and Its Application as a Screening Tool". Science. 5(197), 197ra101.

[21] Huang H., et al. (2022). "Gene Mutational Clusters in the Tumors of Colorectal Cancer Patients with a Family History of Cancer". frontiersin. (22).

APPENDICES

Tool	Version	Reference	Used for
Python	>=3.10	https://www.python.org/downloads/release/python-3100/	To run all the scripts
urllib.request	3.9	https://docs.python.org/3/library/urllib.request.html	Download the VCF gzip files
gzip	3.12	https://docs.python.org/3/library/gzip.html	Unzip the VCF files
requests	2.31.0	https://requests.readthedocs.io/en/latest/	Download and save the reference genome
numpy	1.23.1	https://numpy.org/	For preprocessing the SBS file. To calculate a cutoff value for data normalization. And used for multinomial randomization
pandas	1.5.0	https://pandas.pydata.org/	Used for all the matrices operations
pyarrow	14.0.1	https://arrow.apache.org/docs/python/index.html	To write the large SBS and NMF files as parquet files to save space.
fastparquet	2023.10.1	https://fastparquet.readthedocs.io/en/latest/	To write the large SBS and NMF files as parquet files to save space.
scikit-learn	1.3.1	https://scikit-learn.org/stable/	Calculating optimal column assignments between two DataFrames based on the linear sum assignment and calculating Jensen-Shannon distance.
seaborn	0.13.0	https://seaborn.pydata.org/	Used to create all the plots/figures.
matplotlib	3.7.1	https://matplotlib.org/	Used for the backend of all the plots/figures.
statkit	0.2.3	https://gitlab.com/hylkedonker/statkit	To calculate the perplexity of the NMF result.
pdoc	14.1.0	https://pdoc.dev/	Auto-generates documentation.
mubelnet	Commit 7ed677cd	https://gitlab.com/hylkedonker/mubelnet	Performs the deep Bayesian unsupervised clustering models and trains them.

Table 13: A table of all the used libraries/tools/plugins for this research. And what they were used for.

Name	Fullname	Datatypes
Type	Represents the type of mutation.	str
Gene	Indicates the specific gene associated with the mutation.	str
PMID	Refers to the PubMed ID of the associated research paper.	str
Genome	Specifies the genome version used for mapping.	str
Mutation Type	Describes the type of mutation.	str
Chromosome	Represents the chromosome number where the mutation occurs	str
Start Position	Indicates the starting position of the mutation on the chromosome.	str
End Position	Represents the ending position of the mutation on the chromosome.	str
Reference Allele	Denotes the original allele at the mutation site.	str
Mutant Allele	Represents the altered allele resulting from the mutation.	str
Method	Describes the method used for mutation detection	str

Table 14: The codebook of the VCF files.

	Mean	SD	HDI 3%	HDI 97%	Mcse mean	Mcse sd	Ess bulk	Ess tail	R hat
M1	2.97	0.07	2.90	3.11	0.04	0.03	5.41	12.87	2.13
M2	2.58	0.34	1.99	3.06	0.16	0.12	4.91	27.58	2.61
M3	2.95	0.06	2.85	3.06	0.03	0.02	5.05	25.61	2.47
M4	3.04	0.06	2.95	3.15	0.02	0.01	9.50	135.20	1.37

Table 15: Statistic about the 4 chains of the multinomial belief network.

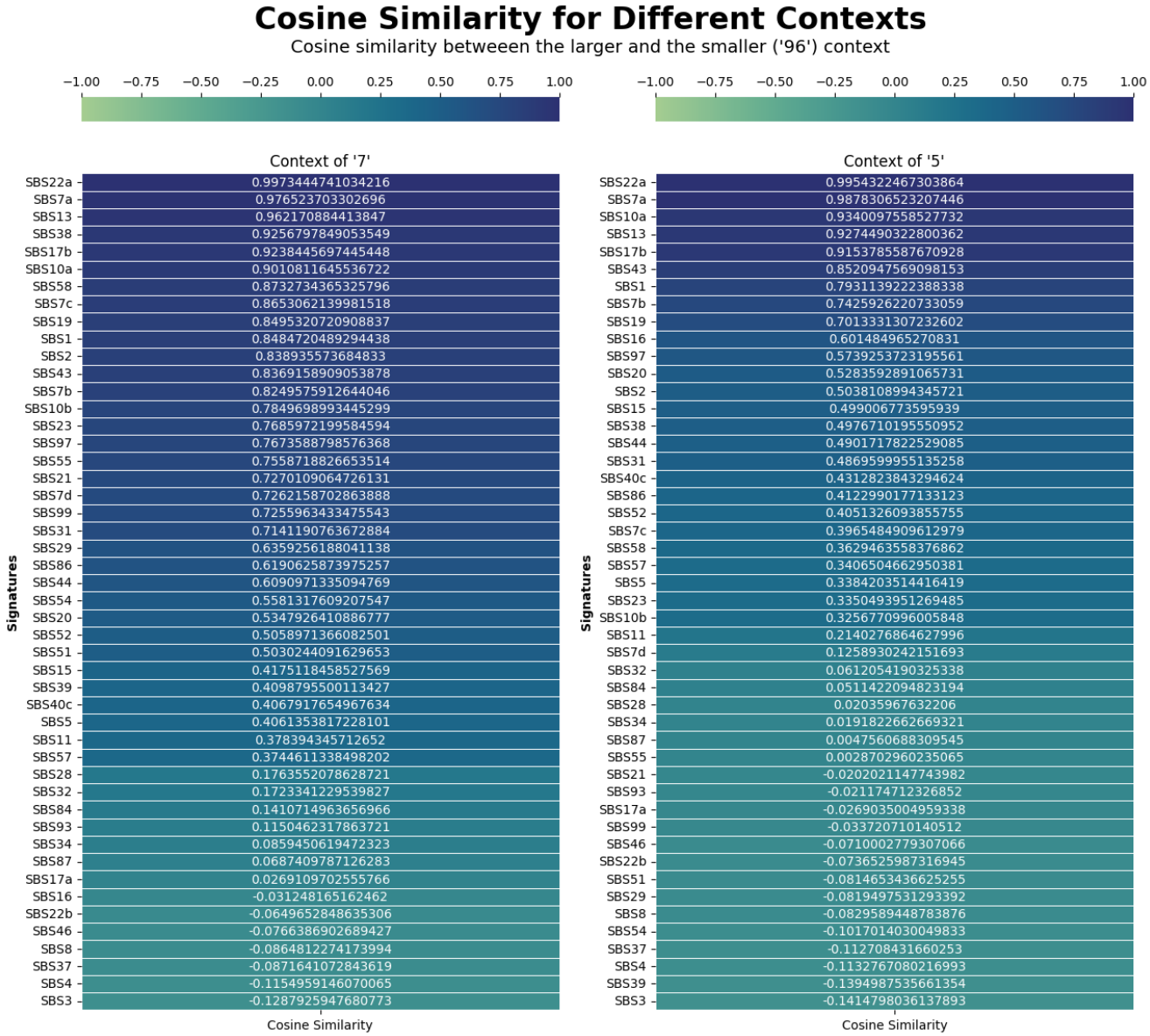


Figure 3: *Heatmap of the cosine similarity between each compressed signature and the 96 signature.*

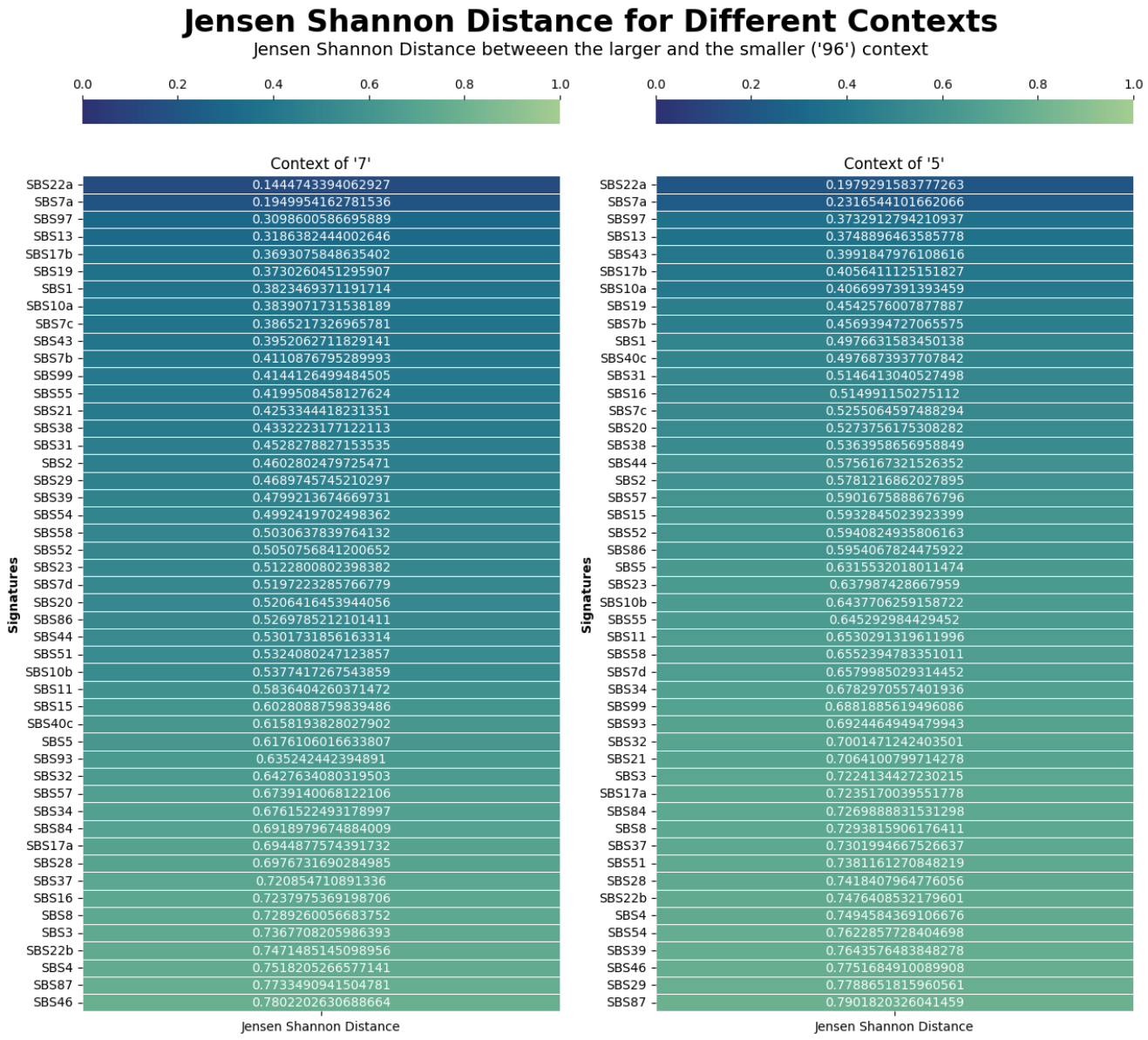


Figure 4: *Heatmap of the Jensen Shannon Distance between each compressed signature and the 96 signature.*

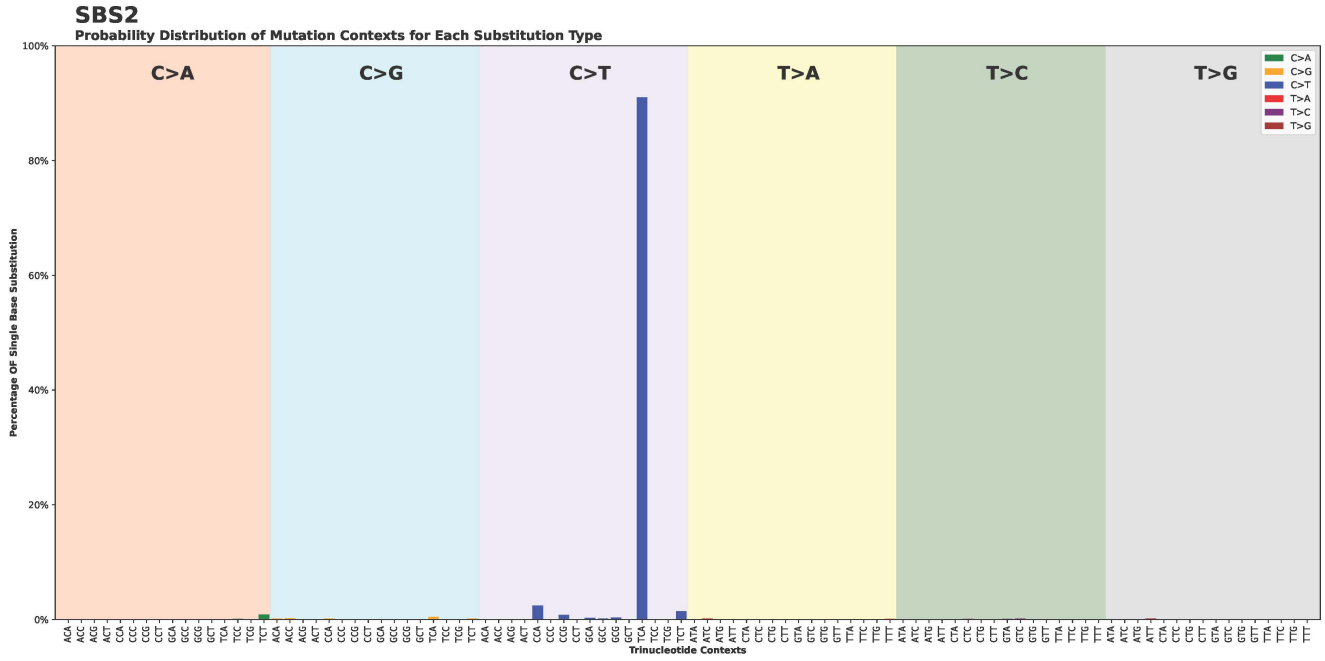


Figure 5: *'96' context signature plot of the whole mutation spectrum of SBS2.*

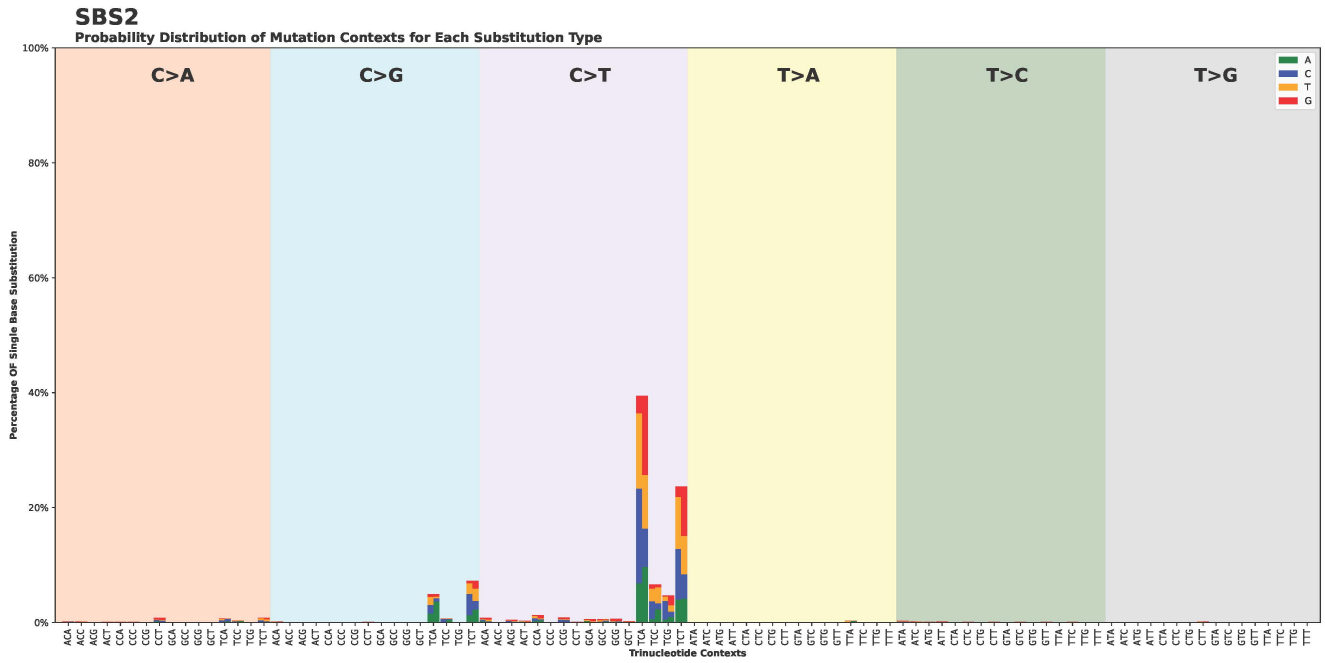


Figure 6: *'1536' context signature plot of the whole mutation spectrum of SBS2. This also shows for each extra position the nucleotide distribution.*

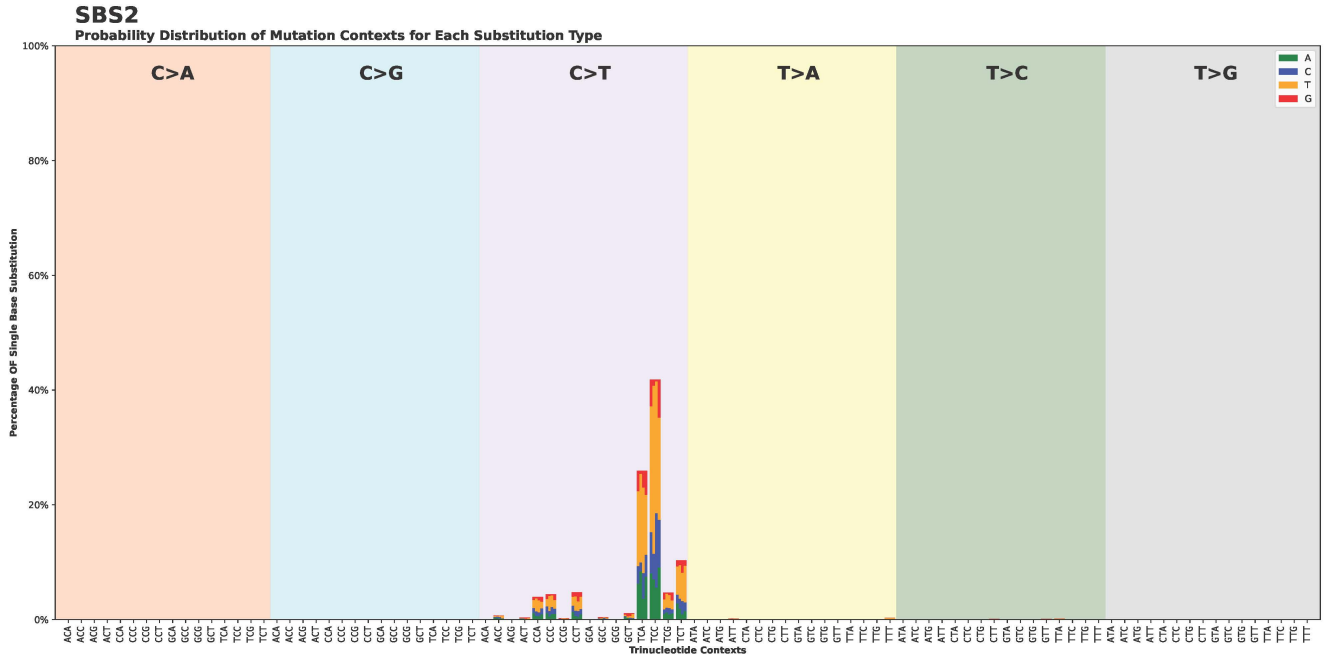


Figure 7: *'24756' context signature plot of the whole mutation spectrum of SBS2. This also shows for each extra position the nucleotide distribution.*

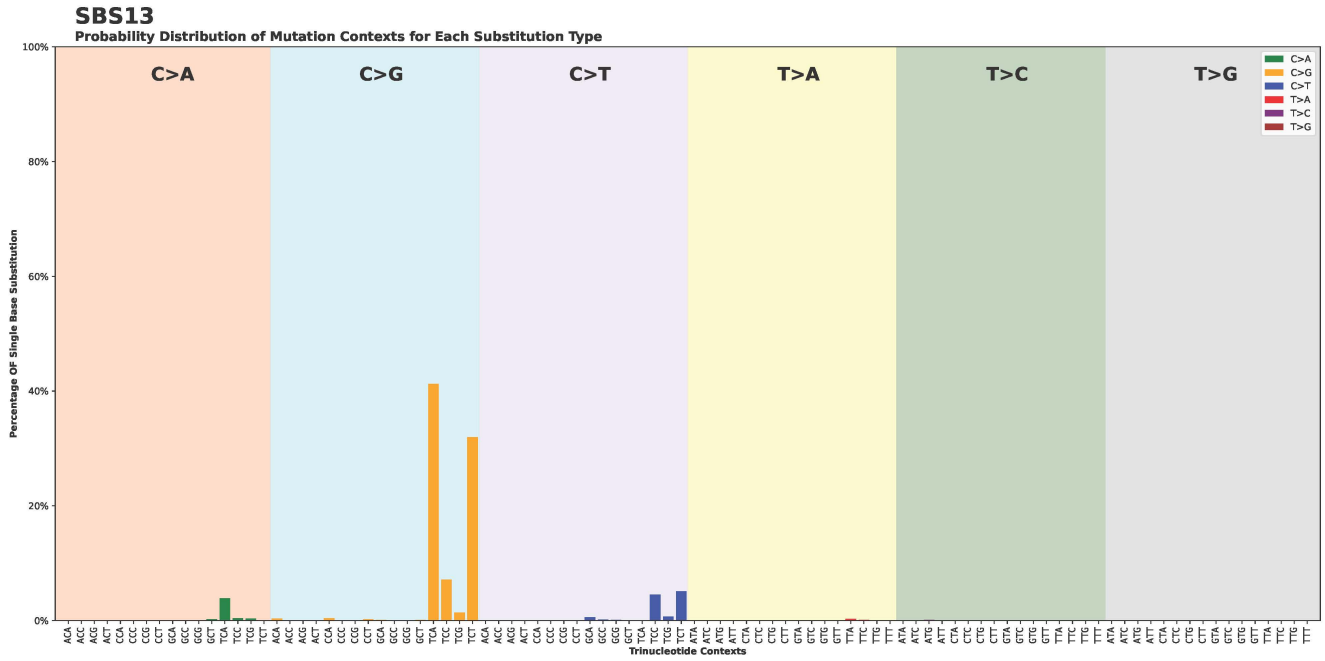


Figure 8: *'96' context signature plot of the whole mutation spectrum of SBS13.*

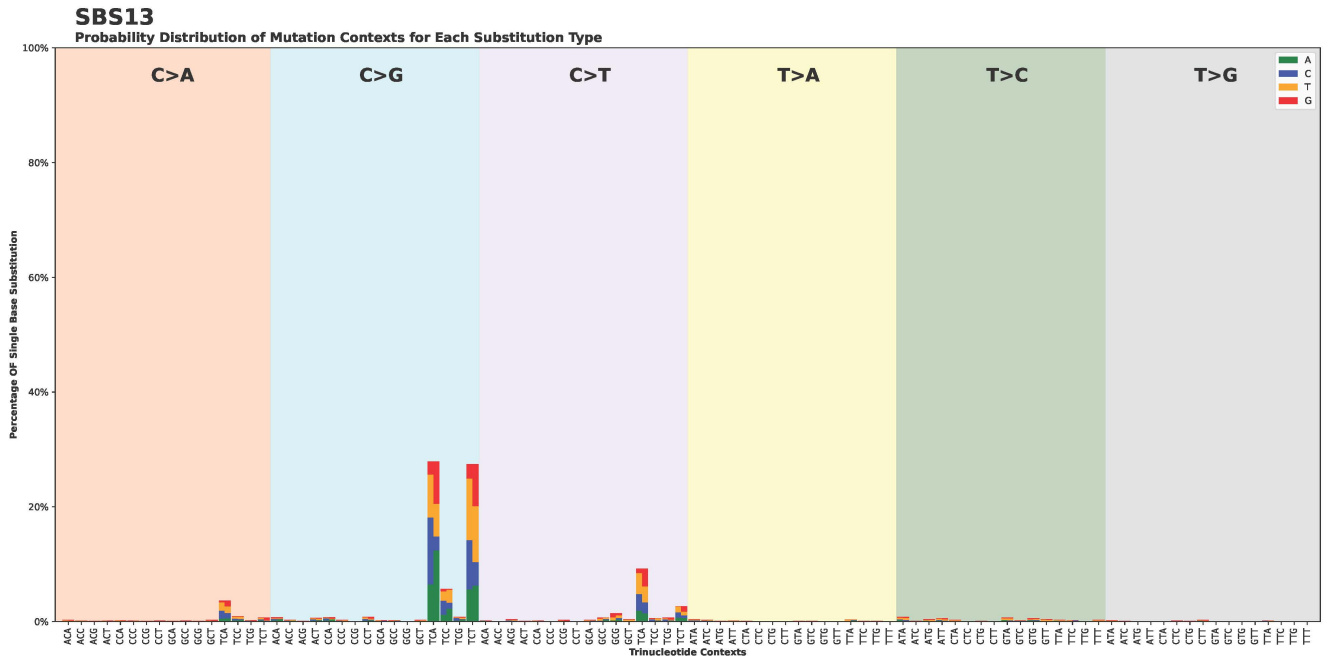


Figure 9: *'1536' context signature plot of the whole mutation spectrum of SBS13. This also shows for each extra position the nucleotide distribution.*

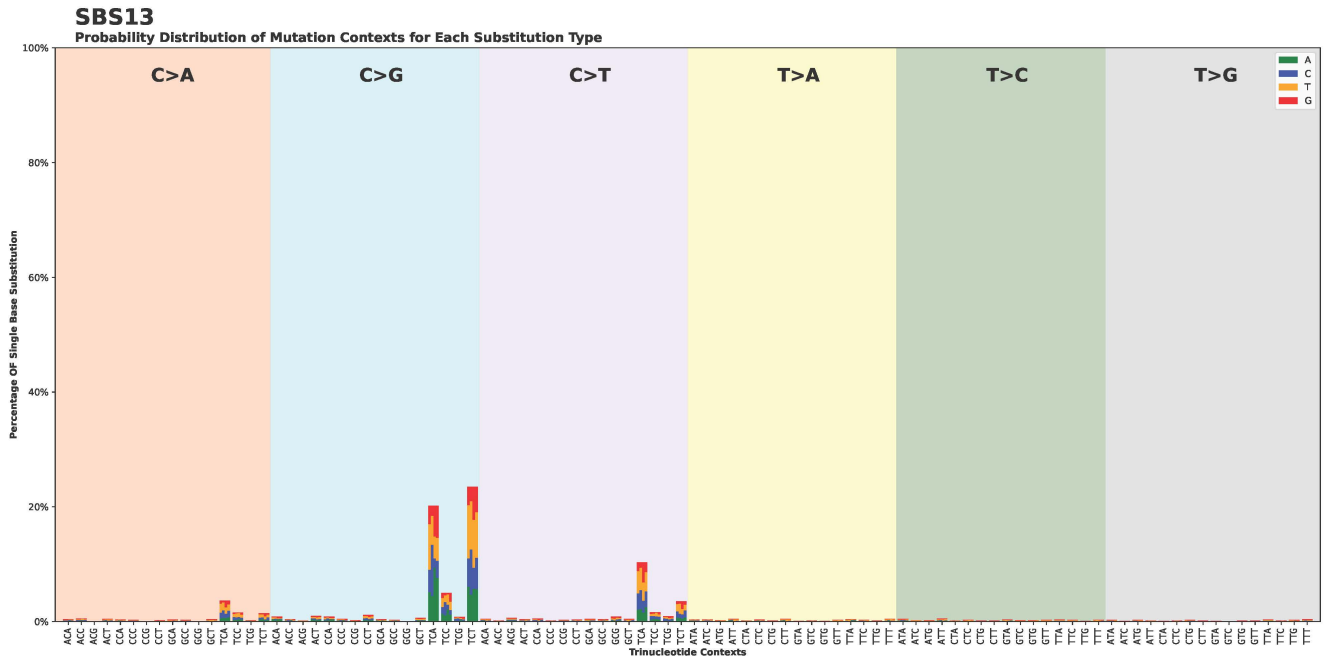


Figure 10: *'24756' context signature plot of the whole mutation spectrum of SBS13. This also shows for each extra position the nucleotide distribution.*

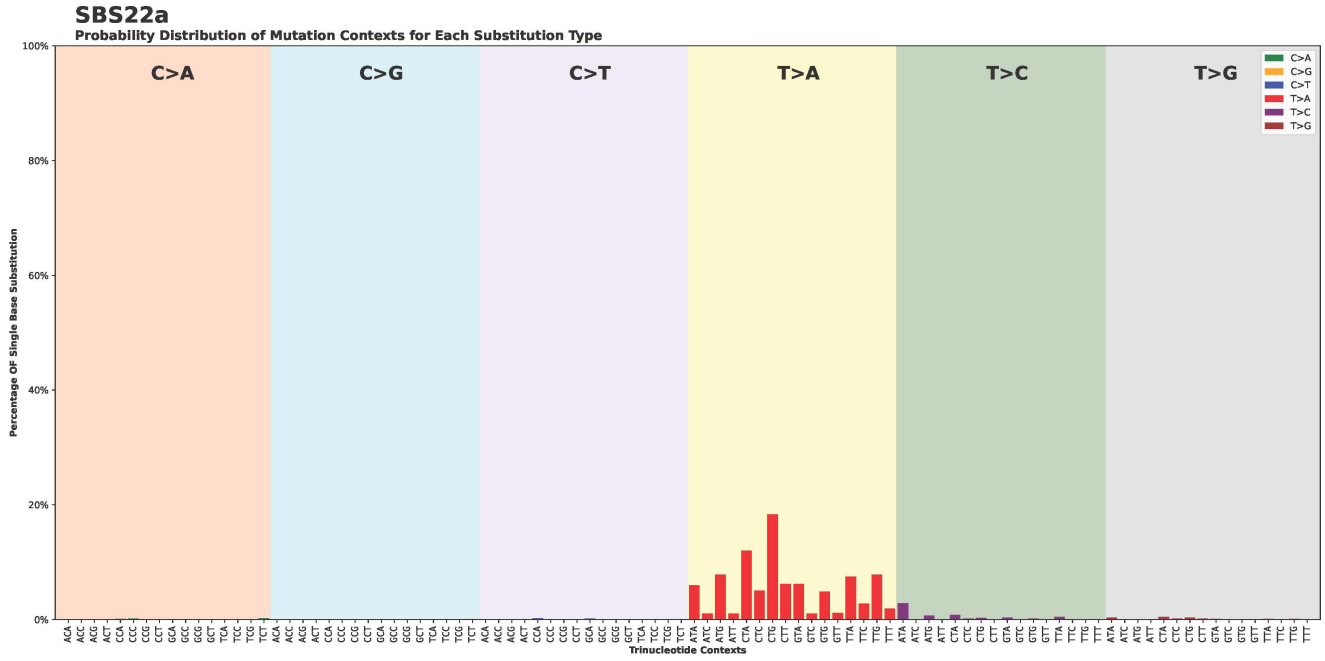


Figure 11: *'96' context signature plot of the whole mutation spectrum of SBS22a.*

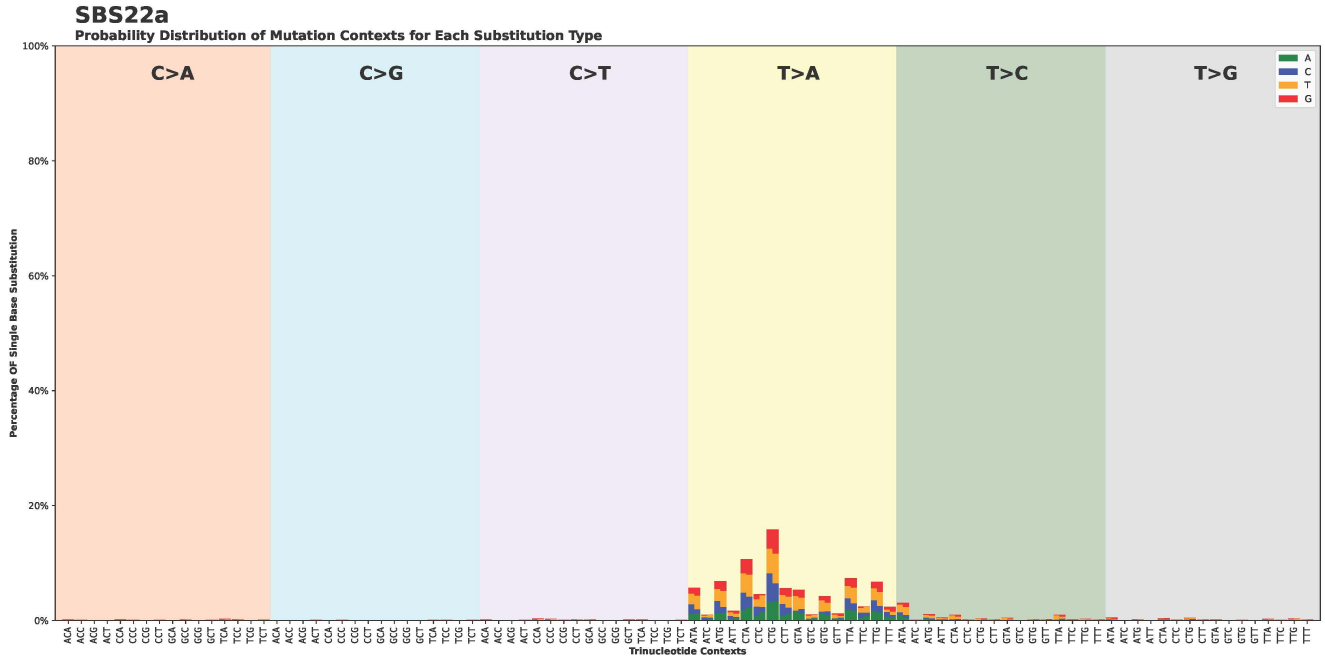


Figure 12: *'1536' context signature plot of the whole mutation spectrum of SBS22a. This also shows for each extra position the nucleotide distribution.*

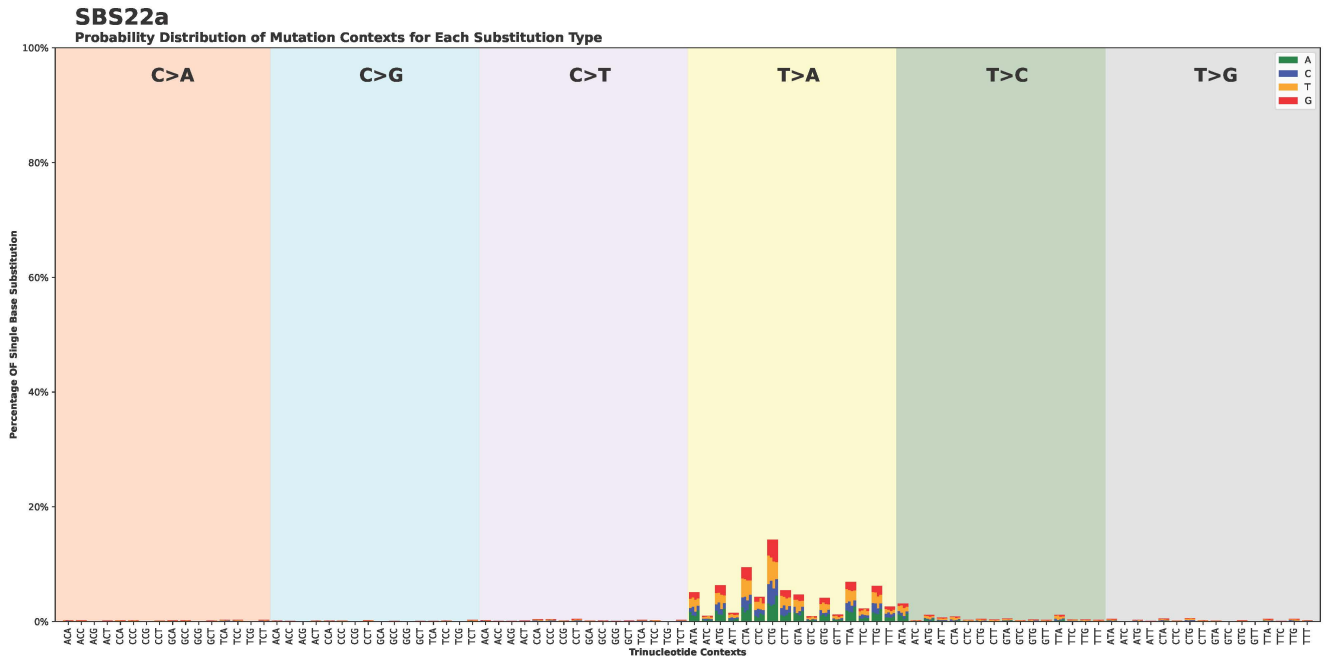
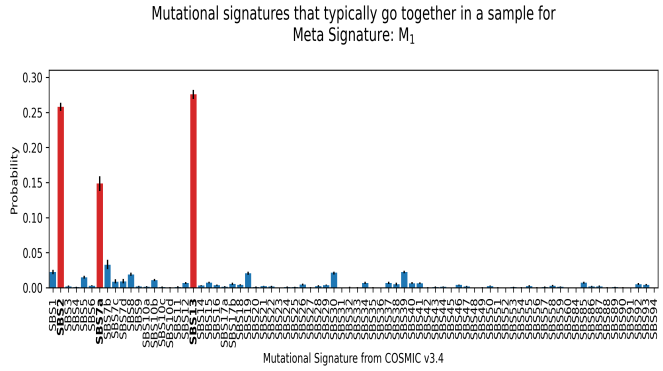
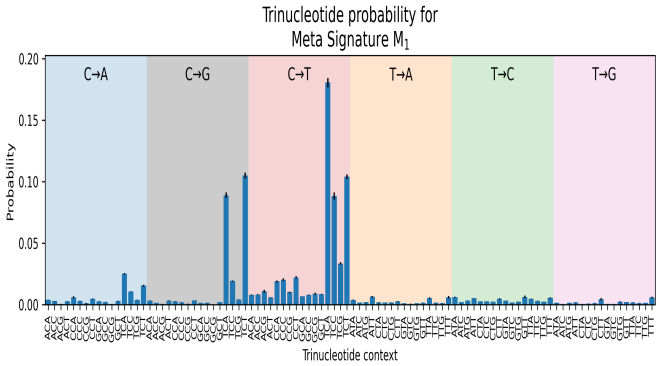


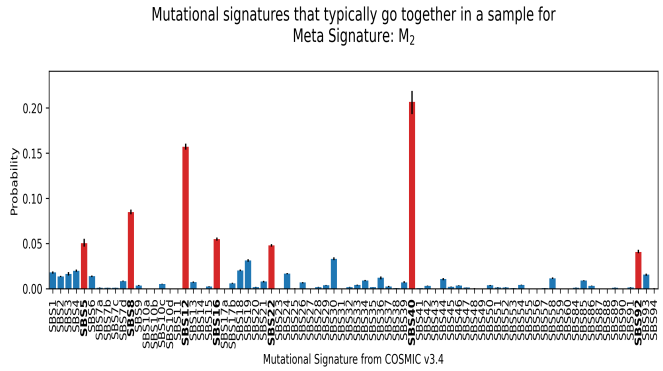
Figure 13: *'24756' context signature plot of the whole mutation spectrum of SBS22a. This also shows for each extra position the nucleotide distribution.*



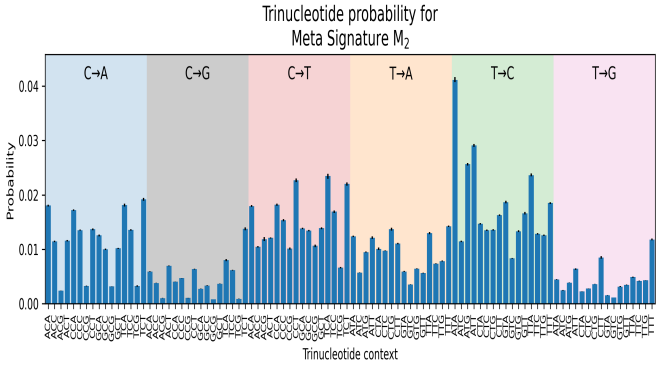
(a) Mutational signatures that typically go together in a sample for meta signature M_1



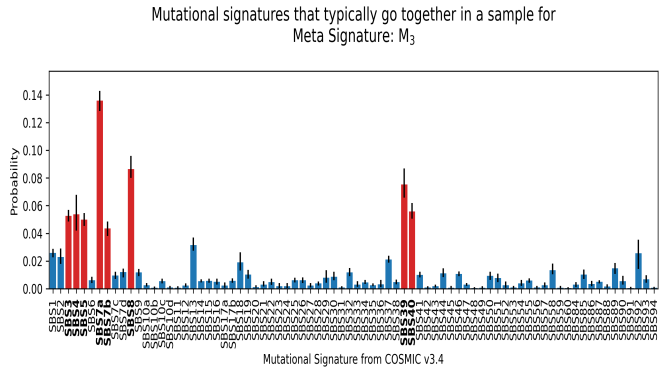
(b) Trinucleotide probability for meta signature M_1



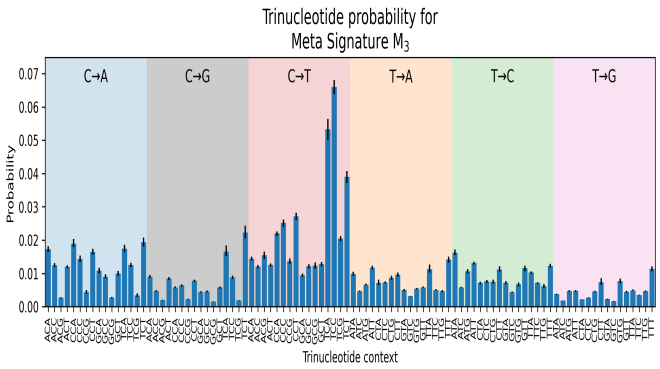
(c) Mutational signatures that typically go together in a sample for meta signature M_2



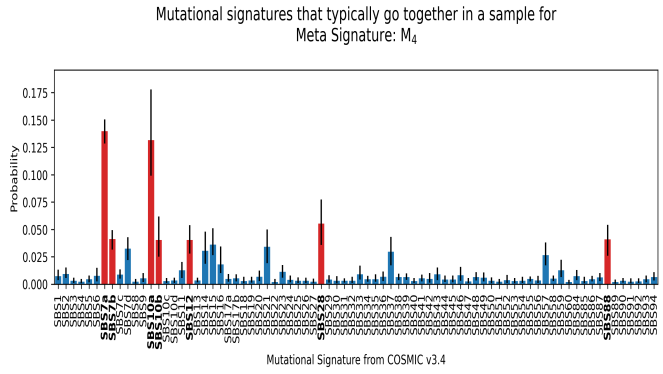
(d) Trinucleotide probability for meta signature M_2



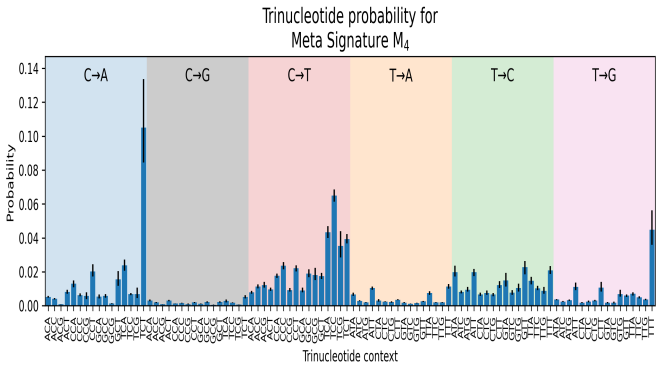
(e) Mutational signatures that typically go together in a sample for meta signature M_3



(f) Trinucleotide probability for meta signature M_3



(g) Mutational signatures that typically go together in a sample for meta signature M_4



(h) Trinucleotide probability for meta signature M_4

Figure 14: Posterior of four meta-mutational signatures $\phi^{(2)}_{vk}$ (labelled $k = M_1, ..., M_4$) in terms of COSMIC v3.4 mutational signatures $v = SBS_1, ..., SBS_{94}$ (left column) and its projection $\sum_{v=SBS_1}^{SBS_{94}} \phi^{(1)}_{lv} \phi^{(2)}_{vk}$ onto tri-nucleotide single base substitutions l (right column). Bars indicate the average and 95% quantile range of the posterior samples. On the left, mutational signatures exceeding three times the uniform probability have been marked in boldface and red.