UCD School of Medicine
Scoil an Leighis UCD

Decomposition of somatic mutation profiles into mutational signatures

Internship Research Project 10
MEIN40420

Anastasiia Deviataieva

Student Number: 24100519

August 2025

Contents

Abstract

Ampullary adenocarcinoma (AMPAC) is traditionally classified by morphology, which masks underlying biological differences and provides little prognostic value. Zhuravleva et al. (2025) showed that decomposition of somatic mutational spectra can delineate an immunogenic subtype with favourable outcomes and Wnt-associated subtypes with poor survival. The goal was to reproduce this approach using publicly available data and assess whether the same biological subgroups emerge.

Somatic single-nucleotide variant (SNV) data were loaded from the supplementary material of Zhuravleva et al. (2025) ; only SNVs passing the FILTER = PASS criterion were used, and samples with fewer than 15 SNVs were excluded. For each tumour, 96-channel trinucleotide spectra (SBS-96) were generated and normalised. The optimal number of signatures was selected via NMF::nmfEstimateRank (Brunet method, 500 restarts), and the final decomposition with k = 3 was run 1000 times to stabilise signature profiles. Signatures were matched to COSMIC using standard and peak-sensitive cosine similarities, and exposures were hierarchically clustered (row-wise z-normalisation, 1 – Pearson distance, complete linkage). Robustness was evaluated by varying methods/linkages and computing the ARI. For each group, the dominant profile was determined by "majority vote" and by the mean contribution; differences in overall survival were assessed using the Kaplan–Meier method with global and pairwise log-rank tests.

Three stable signatures were recovered. S1 matched COSMIC SBS1/6/15, reflecting a combination of age-related deamination and mismatch-repair deficiency. S2 matched SBS4/29, indicating tobacco carcinogens with SBS95. S3 most closely resembled SBS3 and SBS40a/b, consistent with homologous recombination deficiency. Hierarchical clustering of exposures split the cohort (n = 71) into three groups: Cluster 1 (Sig3-dominant) – average contribution of S3 ≈45%; suggests HR-deficient tumours, potentially harbouring BRCA1/2 mutations; Cluster 2 (Sig2-dominant) – average S2 share ≈54%; indicates tumours exposed to exogenous mutagens such as tobacco smoke; Cluster 3 (Sig1-dominant) – average S1 share ≈74%; consistent with mismatch-repair–deficient (MSI-high) tumours. While these subtypes broadly echo Zhuravleva et al. (2025)'s findings, the presence of SBS95 in the tobacco cluster and the predominance of HRD signatures in Cluster 1 highlight some variability. Clusters had low silhouette scores and ARI, indicating a continuum rather than discrete groups. Survival analysis found no significant differences (global log-rank p ≈0.71; all pairwise p > 0.4), likely due to limited cohort size and overlapping mutational processes.

Key words: Mutational signatures, Ampullary carcinoma, AMPAC, Non-negative matrix factorization (NMF), COSMIC, SomaticSignatures, sigminer, Hierarchical clustering, Homologous recombination deficiency.

1 Introduction

Ampullary carcinoma (AMPAC) is a rare but clinically challenging malignancy that arises from the junction of the biliary, pancreatic and duodenal epithelia at the ampulla of Vater. Although surgical resection offers a potential cure, nearly half of patients present with unresectable disease and the overall five-year survival is only around 30%. The current taxonomy of AMPAC divides tumours into intestinal and pancreatobiliary subtypes according to histomorphology, but this classification has limited prognostic value and does not reliably predict response to therapy. Molecular studies have shown that recurrent driver mutations, particularly in Wnt pathway genes (APC, SMAD4, CTNNB1, ELF3), occur across both histological subtypes and that a substantial proportion of tumours lack clear molecular classification. These observations highlight the need for a more robust molecular taxonomy that can capture the diversity of mutational processes in AMPAC and inform personalised treatments.

One promising approach to refining cancer taxonomy is the analysis of mutational signatures – patterns of base substitutions and sequence contexts that reflect the cumulative activity of endogenous and exogenous mutagenic processes. In the seminal study by Zhuravleva et al. (2025), whole-exome sequences from 170 AMPAC tumours were decomposed into mutational signatures using non-negative matrix factorisation (NMF). Hierarchical clustering of signature exposures identified three patient clusters: C1, enriched for spontaneous deamination at CpG sites and deficient mismatch repair (signatures SBS1 and SBS6), C2, associated with transcription-coupled nucleotide excision repair (SBS40/SBS5), and C3, dominated by polymerase-η-mediated hypermutation (SBS9/SBS5). These clusters correlated with Wnt pathway alterations, DNA methylation patterns and immune infiltration; patients in cluster C3 showed improved survival, while clusters C1 and C2 exhibited Wnt activation and immune exclusion. The study proposed that mutational signature-based classification could identify patients likely to benefit from Wnt inhibitors or immunotherapy.

The present internship project sought to reproduce the mutational signature-based subtyping of ampullary carcinoma (AMPAC) from Zhuravleva et al. (2025) using publicly available data, and explore whether similar biological subgroups can be identified from scratch. Following the project guidelines, we first downloaded the mutation annotation format (MAF) table provided in the supplementary material and implemented a rigorous data-cleaning pipeline to correct distortions introduced by Excel formatting. Only single-nucleotide variants passing quality filters and occurring in samples with ≥ 15 mutations were retained for analysis. Next, I generated 96-trinucleotide mutation matrices for each sample and normalised them to account for differences in mutation burden. I performed NMF across a range of signature numbers ($k = 2...7$) with multiple random restarts, selecting $k = 3$ based on the stability of extracted patterns and their similarity to COSMIC reference signatures. Each de novo signature was compared with COSMIC signatures using both conventional cosine similarity and a peak-sensitive cosine metric that emphasises dominant mutation categories,

enabling robust assignment of signatures to known aetiologies. The resulting signature exposure matrix was subjected to hierarchical clustering; the optimal number of clusters was inferred from dynamic tree cutting and silhouette analysis. Finally, I assessed the biological relevance of the clusters by comparing their signature compositions with those reported by Zhuravleva et al. (2025), examining clinicogenomic features, and exploring potential associations with patient survival.

2 Methods

2.1 Data acquisition and pre-processing

The primary data for analysis were obtained from the electronic supplementary material to the article by Zhuravleva et al. (2025). The table Supplementary material `gutjnl-2024-333368supp011.xlsx` was used, posted on the journal's website as part of the online supplements to the publication. The data were loaded in the original form without changes.

2.1.1 Data loading and recovery

The data are loaded from an Excel file with preservation of the data type of each cell. This is done intentionally, since Excel often automatically formats the contents of cells. For example, if a string looks like a date (3/8 or 01-02), Excel could interpret it as a date and save not a string but a date object. When reading without special handling, such values could be immediately converted into dates or numbers (the Excel identifier of a specific date), which would complicate the detection and correction of errors. This step ensures maximum preservation of the original information from the table and simplifies the subsequent identification and correction of distortions caused by Excel. Further, I performed a number of steps aimed at correcting and restoring distorted values. In the code, an analysis of columns and cells is provided in order to find places where Excel turned text into dates or times. Such cases are especially problematic for genomic data: for example, the gene MARCH1 can become the date "1-Mar", and the numerical coordinates of chromosomes can turn into unrecognizable dates because Excel considers large numbers to be dates (internally, Excel stores dates as the number of days since 1899-12-30). As a result, critically important fields (gene names, positions of mutations) may turn out to be incorrect in further use. In the code (optional block 2.1) a list of columns is determined where there are values of the Date/POSIX class, and such "suspicious" cells are collected for inspection. This is needed in order to understand the scale of the problem and which columns are affected.

Next, I restored the names of genes: the `Hugo_Symbol` column (and its duplicate SYMBOL) is checked. The correction is performed with the help of the custom function: it takes the gene identifier from the column `Entrez_Gene_Id` for each problematic row and, through the annotation database (the `org.Hs.eg.db` package), finds the corresponding gene symbol.

The found gene name is substituted back into `Hugo_Symbol` (and in parallel into the field `SYMBOL`).

The next step was the correction of mutation coordinates – the columns `Start_Position` and `End_Position`. These fields must contain numerical positions on the genome, but it represented large numbers as dates (for example, the number 34189526 could be displayed as the date 95507-09-28 – Excel interprets the number as days since the beginning of the 20th century). To return the correct coordinates, a function is implemented that: leaves already numerical values unchanged; for values of the Date/POSIX type calculates the number of days from the Excel base (1899-12-30) and thereby restores the original number; for strings resembling the YYYY-MM-DD format, also performs the reverse conversion into a number; for the remaining strings extracts only the digits (for example, removes commas/spaces) and converts them to a number.

Next, I corrected the fields `Exon_Number, EXON, INTRON, cDNA_position, CDS_position` and `Protein_position`, which Excel often distorts, taking entries like 3/8, 5/12 or 7/2016 for calendar dates. To fix this, the function goes through the specified columns and touches only those cells that actually became of the Date/POSIX type – everything else remains as is, but is converted to a string for uniformity. For the exon/intron fields a simple heuristic is used: from the "date" cell the two-digit year is taken and according to it the format is chosen – if it is greater than 12, then the original was "month/year," otherwise "day/month." For positions in the transcript (`cDNA_position, CDS_position, Protein_position`) all "date" cells are converted into a string of the form m/YYYY, that is, the month and the full year, which corresponds to the original entries like 2/4425 (instead of Feb-25).

For all the other columns that are not needed for further analysis and already contain many missing values, all distorted values are marked as NA. Additionally, throughout the entire dataset the string marker of missing data " . " is unified to NA, so that missing values are represented uniformly. Next, the table is unpacked from the "list" format of columns to an ordinary vector: for each list-column the first value in the cell is taken (in the original Excel the cell in any case had only one semantic element). After unpacking, R itself selects the common atomic type of the column. The final, cleaned and materialized dataset is saved to the TSV-file `maf_from_excel_sheet4.tsv`.

2.1.2 Data pre-processing

Reading and review of MAF: I loaded the obtained mutational data using `maftools::read.maf()`, and performed their review. A summary plot was built: it shows the number of mutations in each sample and the distribution of types/classifications of mutations across the entire set. In addition, an oncoplot is built for the top-25 frequently mutated genes – this is a "waterfall" diagram, displaying which genes are affected by mutations in which samples, allowing one to immediately see the key genes with the largest number of mutations in the

cohort.

Filtering of variants: At the stage of data cleaning from MAF only single-nucleotide substitutions (SNV/SNP) were selected, since the analysis of mutational signatures usually focuses on single base substitutions. Further, from these SNVs only entries with the filter `FILTER = "PASS"` were retained, that is, variants that passed all threshold quality criteria at variant calling. By filtering, we exclude potentially artefactual or low-quality mutations. Then a culling of samples by the number of mutations is performed: the calculation of the number of SNVs for each sample reveals how many mutations are contained in each sample. Only samples with $\geq$15 SNVs are retained, since too small a number of mutations in a sample makes statistically reliable extraction of signatures difficult; the threshold of 15 mutations increases the robustness of the subsequent analysis (for example, improves the convergence of the NMF model when decomposing into signatures).

QC analysis: Ti/Tv and "rainfall". For rapid quality control, the transition-to-transversion ratio (Ti/Tv) is calculated for each sample and the corresponding plot is built. This is a metric of the quality of genomic variations: the expected Ti/Tv ratio in humans is $\approx 2-3$. (Bridgers et al. 2024) A substantial deviation of Ti/Tv in any sample may indicate a biased composition of mutations or problems with the data (for example, an excess of false-positive calls at a low ratio). Additionally, rainfall plots are built for several samples with the largest number of mutations. The rainfall plot visualizes the positions of all mutations along the genome of a given sample and the distances between neighboring mutations. In such a plot, clusters of very closely located mutations appear as "clusters" of points, testifying to a localized hypermutational event known as kataegis. Identification of similar regions of hypermutations is important for understanding the structure of the data: kataegis can affect the distribution of mutational signatures, therefore its presence is recorded at the QC stage.

Distribution of mutations across samples: Finally, visualization of the distribution of the number of SNVs per sample after all filtrations is performed. A horizontal bar chart is built, where for each remaining sample the number of detected SNPs is shown, ordered by magnitude – this clearly reflects the variation of the mutational burden between samples. A frequency histogram is also built, showing how many samples have a certain range of mutations. These plots make it possible to see the general picture: whether the majority of samples contain a comparable number of mutations, whether there are remaining outliers in mutational burden, what the median value and the range of mutations in the sample are.

2.2 96-trinucleotide matrices preparation

For the analysis of mutational signatures, it is first necessary to represent the profiles of somatic mutations in the form of an SBS-96 matrix – a standardized catalog of single-nucleotide substitutions in 96 contexts. This approach classifies each single-nucleotide mutation into one of 96 categories based on its trinucleotide context: the type of the substituted base is

taken into account (one of six types of substitutions, usually recorded with respect to the pyrimidine base) and the two neighboring nucleotides to the left and right. Thus, all single substitutions (SBS) are distributed across 96 possible contexts (6 types of substitutions × 4 variants of the 5 context × 4 variants of the 3 context). This format corresponds to that adopted in the COSMIC (Catalogue of Somatic Mutations in Cancer), where the reference signatures are specified as probability distributions over the same 96 categories of mutations. (Medo et al. 2024) I implemented two parallel routes for constructing SBS-96 matrices from the cleaned MAF file:

- The first approach uses the package `sigminer::sig_tally()`, which makes it possible, directly on the basis of the MAF file and a reference to the genome, to obtain an SBS-96 frequency matrix for all samples. In this case, sigminer automatically extracts for each SNV its neighboring bases from the reference genome GRCh37/hg19 and assigns the mutation to one of the 96 categories according to the pyrimidine-base rule. The result is a matrix of dimension samples × 96, in which the column names are already brought to the standard SBS format (for example, A[C>T]A denotes a C>T substitution in the context of 5 -A and 3 -A). Such an output is directly compatible with COSMIC signatures and can be used for further decomposition or comparison with known signatures for `sigminer` without additional processing.

- The second approach is based on the `SomaticSignatures` package and is intended for constructing normalized mutational profiles taking into account the occurrence of contexts. At the first stage, the SNVs of each sample were converted into a `VRanges` object and enriched with information on the local trinucleotide context using the function `SomaticSignatures::mutationContext()` and the reference genome hg19. Then, using the function `SomaticSignatures::motifMatrix()`, a matrix of size 96 × N_samples was obtained. Unlike `sigminer`, the `SomaticSignatures` applies its own system of context notation (for example, a context may be denoted in an alternative format different from the A[C>T]A notation), therefore initially the rows of the matrix are labeled differently. This route was provided mainly for reproducing the normalization methodology described in the Zhuravleva et al. (2025) and for obtaining normalized mutation spectra suitable for comparison between samples and clustering.

Next, I considered two strategies for normalizing the obtained profiles.

- Normalization 1 (per-sample frequencies; better for comparison with COSMIC). For each sample, its 96-dimensional spectrum is converted into a probability distribution: each value is divided by the column sum so that the final sum equals 1. This operation eliminates the influence of the overall mutational burden, preserves the relative shape of the spectrum, and brings the profiles to the same scale in which the COSMIC reference signatures are specified.

- Normalization 2 (opportunity-normalization + subsequent conversion to fractions; closer to the methodology of the article). This approach includes an additional adjustment for the uneven occurrence of various trinucleotide contexts in the genome. First, for each sample a vector of absolute frequencies of 96 types of mutations is constructed. Then each element of this vector is scaled inversely proportional to the frequency of the corresponding triplet in the reference genome. In other words, the mutation counts are divided by the expected probability of the given trinucleotide in the human genome (for hg19 the frequencies of all 3-mers were calculated using the function `SomaticSugnatures::kmerFrequency()`). Such a correction takes into account that, for example, CpG motifs occur less often or more often than others, and reduces the influence of the background distribution of bases on the mutation profile. After this, the corrected profiles are also normalized by columns to sums = 1, in order to obtain distributions over 96 contexts comparable to each other. This closely corresponds to the procedure for preparing profiles described in the article by Zhuravleva et al. (2025). However, this approach also has significant drawbacks. First, division by the frequency of rare triplets can disproportionately amplify statistical noise in cases when there are very few mutations in a rare context. Second, normalization to the frequencies of contexts from the genome can introduce a systematic difference of the profile if the real sequencing data do not cover the entire genome uniformly (for example, exome or targeted panels with a different distribution of triplets). As a result, the obtained "opportunistically" normalized spectra may be shifted relative to the standard COSMIC signatures, reducing the value of cosine similarity in the comparison. Thus, if the goal of the analysis is the identification and comparison of signatures with COSMIC patterns, it is preferable to limit oneself to the first normalization strategy.

It should be noted that signature decomposition and comparison with COSMIC were performed using the first approach along with SBS-96 alignment.

2.2.1 Normalization & SBS-96 Alignment

To ensure the correct use of the obtained profiles in subsequent steps, it was necessary to take into account differences in the formats of context representation between the two instrumental routes. The matrix formed by `sigminer` initially has the SBS-96 format, fully compatible with COSMIC – its columns are ordered and labeled (e.g., A[C>A]A) according to the standardized context notation. At the same time, the alternative matrix from `SomaticSignatures` contains the same 96 channels, but labeled differently (e.g., "CA A.A") and arranged in a different order. To compare these two representations and integrate the normalized profiles into further signature analysis (via `sigminer` functions), alignment of contexts is necessary. It's reduces to establishing a one-to-one correspondence between the 96 rows of the `SomaticSignatures` matrix and the 96 columns of the `sigminer` matrix.

This is achieved by pairwise comparison of context profiles: both matrices are brought to a comparable form (column-wise normalization), a correlation matrix is computed between the profile of each context from the first matrix and each context from the second, after which the problem of the best partition into pairs (the assignment problem) is solved, maximizing the total correspondence. Applying this mapping brings the format of the alternative matrix to the SBS-96 standard. This makes it possible to safely use the normalized profiles in the sigminer ecosystem.

2.3 Signature count selection & COSMIC matching

To decompose the mutation spectra into constituent mutational signatures, nonnegative matrix factorization (NMF) was applied. To choose the optimal number of signatures, a rank survey of NMF was performed over the range k = 2...7. For this, the standard rank estimation tool was used (`NMF::nmfEstimateRank()`, brunet method, 500 multiple restarts), which simultaneously returns diagnostics of the quality of the decomposition: cophenetic correlation (stability of clustering of solutions), residual errors (RSS/Residuals), proportion of explained variability, cluster silhouette, etc. The optimum was considered to be the smallest number of signatures at which the model still adequately describes the data and the metrics reach a plateau or begin to degrade.

Next, I extracted k=2…7 signatures and aggregated a stable solution (using `sigminer::sig_extract()`): we obtain the signature profiles themselves in the SBS-96 basis, and the exposure matrix across samples. Each extracted NMF signature represents a vector of 96 relative frequencies of mutations in trinucleotide contexts. Matching with COSMIC was performed using two complementary similarity metrics. First, the standard cosine (takes into account all 96 bins; `compute_cosine_matrix()`, inside – `sigminer::get_sig_similarity()`). Second, the peak-sensitive cosine metric (analogous to the Zhuravleva et al. (2025)), which compares only the most significant mutational categories (peaks) in the profiles, rather than the entire spectrum as a whole. In algorithm, from the signature profile positions whose values are less than 1% of the maximum height of this profile were discarded, and at the same time positions that are not major peaks in the reference COSMIC signature were ignored (in the COSMIC profiles the threshold for a peak is taken as 70% of their maximum). Put simply, the metric focuses on the key distinctive mutations of each pattern, masking random "noise" over rare bases. After such filtering, the cosine similarity is computed only over the remaining coordinates. Peak-sensitive similarity requires a match precisely on the main peaks of the distribution, which is why it is a more stringent criterion: if the extracted signature lacks a pronounced peak that is present in the candidate COSMIC signature, then the similarity measure drops sharply, even when the remainder of the profile matches. For example, for a confident assignment of a new signature to SBS1 it is required that its spectrum clearly contains a high peak of C→T mutations in the CpG context (characteristic

of SBS1). Such an approach increases the specificity of the comparison, not allowing background components (for example, the ubiquitous SBS5 with numerous small contributions) to create artificially high similarity.

Comparison of de novo signatures (k=2…7 with 500 runs) with the COSMIC catalog serves not only for naming processes, but also as an additional criterion for choosing the optimal number of signatures k. Within each tested decomposition, I calculated the indicator the mean value (for the ordinary cosine and for peak-sensitive) of the best similarity of each of the extracted signatures with some COSMIC signature. In other words, for a given k the maximal similarity coefficients (the best "hit") are taken for each of the k signatures and averaged. This indicator reflects how fully the set of k signatures covers the known patterns.

For transparency of the reporting comparison I formed two tables. First, a wide table with both metrics on the same columns: for each of our signatures it is visible the standard and peak-sensitive matches with each COSMIC-SBS. Second, summaries of top matches – Top3 ranked lists for quick viewing of "candidate analogs" Sets of files for each model k are saved in batches (`save_k_bundle()`) together with profiles, exposures, and summary tables; this fixes reproducibility and facilitates biological interpretation.

2.4 Final signature extraction

The final factorization of the mutational spectra matrix data was performed (using same options and functions as in signature count selection process) with a fixed number of components k = 3, which had been selected at the previous step as optimal. For this, NMF was applied, as a result of which the original mutation spectrum was decomposed into three characteristic components – three mutational signatures – and their corresponding matrix weights (exposures across samples). To ensure the reliability of the result and the reproducibility of the obtained signatures, the NMF algorithm was run 1000 times with different random initializations. Multiple restarts made it possible to confirm that the detected patterns are stable: the same three signatures are identified in the vast majority of runs, which increases the robustness of clustering and the stability of the model components. After extracting the signatures, they were compared with known reference COSMIC signatures. Two cosine-similarity metrics were used for the comparison: the standard cosine similarity and a special measure sensitive to differences in the main peak values of the profile (peak-sensitive cosine similarity).

2.5 Hierarchical clustering

In this analysis, samples are grouped by the similarity of their mutational profiles using hierarchical clustering. Before clustering, each of the 96 categories of triplet mutations is row-normalized: the Z-score (deviation from the mean in units of the standard deviation) is calculated for each context and sample. This Z-normalization eliminates differences in

scale between contexts, allowing the algorithm to focus on relative deviations of mutation frequencies instead of absolute values. Thus, features with initially high values will not dominate, and the clustering will reveal precisely the characteristic regularities of the distribution of mutations in different samples. To assess similarity between samples, different distance measures can be applied: Euclidean distance (takes into account direct numerical differences), Spearman rank correlation distance (1 − Spearman coefficient, reflects monotonic similarity of distributions), and cosine distance (1 − cosine similarity, characterizes the angle between feature vectors). In the analysis, Pearson correlation distance was used, defined as 1 − Pearson correlation coefficient between the profile vectors of the samples. This metric emphasizes the similarity of the shape of mutational profiles and weakly depends on the total number of mutations in the sample. For the agglomerative merging of samples into clusters, the complete linkage method was chosen. The distance between two clusters in this case is defined as the maximum distance between any pair of samples taken from different clusters. Such an approach is prone to forming relatively compact clusters, avoiding premature merging of heterogeneous samples.

Next, automatic determination of the optimal number of clusters is implemented on the basis of a combination of two approaches: the DynamicTreeCut algorithm and analysis of the silhouette coefficient. First, after constructing the dendrogram, the DynamicTreeCut algorithm automatically identifies clusters by analyzing the structure of the branches without a pre-specified k. Constraints are taken into account, for example the minimum cluster size and the splitting depth parameter. This method finds natural groups on the dendrogram, guaranteeing that branches that are too small do not form a separate cluster. In parallel, an evaluation of partitions is carried out through silhouette analysis for different numbers of clusters. The silhouette coefficient shows how well each sample corresponds to its cluster and is separated from other clusters. In our analysis, the average value of the silhouette is calculated when cutting the dendrogram into k = 2...8 clusters. The k at which the average silhouette is maximal is chosen – this corresponds to clustering with the best density of clusters and separation between them. Having obtained two partition variants (dynamic and by the optimal silhouette), we compare their quality. If the DynamicTreeCut solution gives an average silhouette practically no worse than the optimal (the difference does not exceed a threshold 0.02), then the Dynamic Tree Cut partition is chosen. Otherwise, the partition with the maximal silhouette is adopted. Such a criterion ensures a balance between the strict statistical optimum and the structure revealed on the dendrogram. As a result, the final number of clusters is determined and cluster labels are assigned for all samples.

2.5.1 Normalised three-nucleotide mutational frequency profiles heatmap

In Fig.12 the heatmap is presented, illustrating the profiles of mutational contexts for all samples taking into account the identified clusters. Along the rows of the plot are laid out

the 96 categories of triplet mutation contexts, along the columns – the samples, ordered according to the clustering dendrogram. The color value in each cell reflects the Z-score of the frequency of the corresponding context in the given sample. So as not to overload the plot, only about 20% of the contexts with the largest average deviation in the absolute value of the Z-score are labeled.

2.5.2 Clustering robustness

For the assessment of the stability of the obtained clusters, an analysis of clustering was performed under different combinations of metrics and agglomeration methods. Standard similarity metrics are used (Pearson's, Spearman's, cosine distance and the Euclidean metric) and linkage types (complete, average, ward.D2). It is noted separately that the "ward.D2" method is correct only with Euclidean distance. For each such pair, hierarchical clustering is built, and then the optimal partition into clusters is automatically selected in two ways. The first is the algorithm of dynamic cutting of the dendrogram (`DynamicTreeCut::cutreeDynamic()`), which by the shape of the dendrogram itself determines the number and composition of clusters without prior fixing of "k." The second is the classical enumeration of the number of clusters using the silhouette index: for each possible number of clusters the average silhouette index (a measure of the quality of the partition) is computed, and that number of clusters is chosen at which the silhouette is maximal. Thus one approach removes the necessity of specifying in advance the "correct" number of clusters, and the other guarantees the maximum quality of the partition according to the criterion "within-cluster compactness – between-cluster separation."

Further, both obtained partitions are compared by the value of the average silhouette: if the DynamicTreeCut algorithm gives a silhouette practically not inferior to the best one (within an allowable margin), then its result is chosen, otherwise – the usual variant with the optimal "k" by silhouette. As a result, for each pair (metric-linkage) the final partition, the number of clusters, the method used, and the corresponding values of the silhouette index are fixed. Such a scheme guarantees that the final clusters are of high quality and do not critically depend on the random choice of parameters. For visualization and numerical analysis of stability, a pairwise comparison of all obtained partitions is carried out using the Adjusted Rand Index (ARI). ARI reflects the degree of coincidence of two partitions (1 means complete coincidence), and its matrix is displayed in the form of a heatmap. High ARI values between different combinations of metrics/methods indicate the stability of the data structure: the clustering scheme is stable if the partitions are similar regardless of the metric and structural assumptions used. Thus, this block checks the reliability of the discovered clusters and provides confidence in their further analysis.

2.6 Determining dominant signatures for clusters

At this stage of analysis, for each cluster it is determined which mutational signature predominates. First, the dominant signature is calculated for each individual sample. For this, the exposure profiles of signatures are normalized: the share (fraction) of the contribution of each signature to the total profile of the sample is calculated. Thus, the influence of the total number of mutations is leveled out, and it becomes possible to compare the relative contribution of different signatures. Each sample is assigned its leading signature and the magnitude of its contribution is recorded. This makes it possible to reduce a complex profile of many signatures to one main feature for each sample, which simplifies interpretation. Further, the result is generalized at the level of clusters. For each group of samples the average fractions of all signatures present in the cluster are computed – this shows the typical composition of signatures in the given group. By the maximum average value, the signature is determined that on average is the strongest in the cluster (the so-called "dominant by the average" signature). In parallel, another indicator is also calculated: the most frequent dominant signature (majority vote) – that is, the signature that most often turned out to be the leading one among the samples of the given cluster. Both approaches complement each other and, as a rule, point to the same signature characteristic of the cluster. Such an analysis guarantees that the determination of the key signature of the cluster does not depend on single outliers: we take into account both the overall contribution of the signature and the prevalence of its dominance among the samples.

Finally, for clarity, a visualization of the results was performed. A heatmap of the average fractions of signatures by clusters was constructed, where along the rows clusters are laid out, and along the columns – signatures; the intensity of the color reflects the average contribution of the signature in the given cluster. In addition, a stacked bar chart was created, showing the composition of signatures in each sample, ordered by clusters. In this chart the samples are grouped by clusters, and within each cluster the distribution of the fractions of signatures in each sample is shown by colored segments of the bars.

2.7 KM survival curves by clusters

For the analysis of the association between cluster membership and patient survival, the clinical data (follow-up time and patient status – alive/dead) are combined with the clustering results. Next, Kaplan–Meier survival curves are constructed for each cluster. To formally assess the significance of differences between the curves, the log-rank test is used. With `survival::survdiff()`, a global p-value is computed, reflecting the statistical significance of differences in survival between all clusters at once. In addition, the code further iterates over all pairs of clusters and for each pair separately performs a log-rank test (with p computed via the chi-square statistic). Such pairwise comparisons make it possible to find out between which clusters exactly there are statistically significant differences in survival. For

visualization of the results, the package `survminer::ggsurvplot()` is applied.

2.8 Clinicogenomic heatmap

For the visualization of complex relationships between clinical and genomic data we build a composite heatmap. In it the columns represent patient samples, and the rows – two types of information. In the upper part, color annotations of categorical clinical features are placed (e.g., morphological subtyping of the tumor, sex, MSI/MSS status, activity of the Wnt signaling pathway). Such an approach facilitates the identification of patterns, such as the search for correlations between groups of high mutational burden and certain clinical characteristics. In the lower block of the heatmap, numerical indicators characterizing genomic changes are presented: mutation burden (mutations per Mb), the number of deletions and amplifications. For clarity, all metrics are z-normalized by row, which makes it possible to compare variations of different quantities on a single color scale. Then the data are "trimmed" by the extreme quantiles (the 5% and 95% percentiles) – this prevents the dominance of rare extreme values and makes the map readable, while preserving the diversity of signals. As a result, a single comprehensive visualization is obtained, where the sample labels are aligned between the annotations and the numerical data.

3 Results

3.1 Mutation Landscape Summary

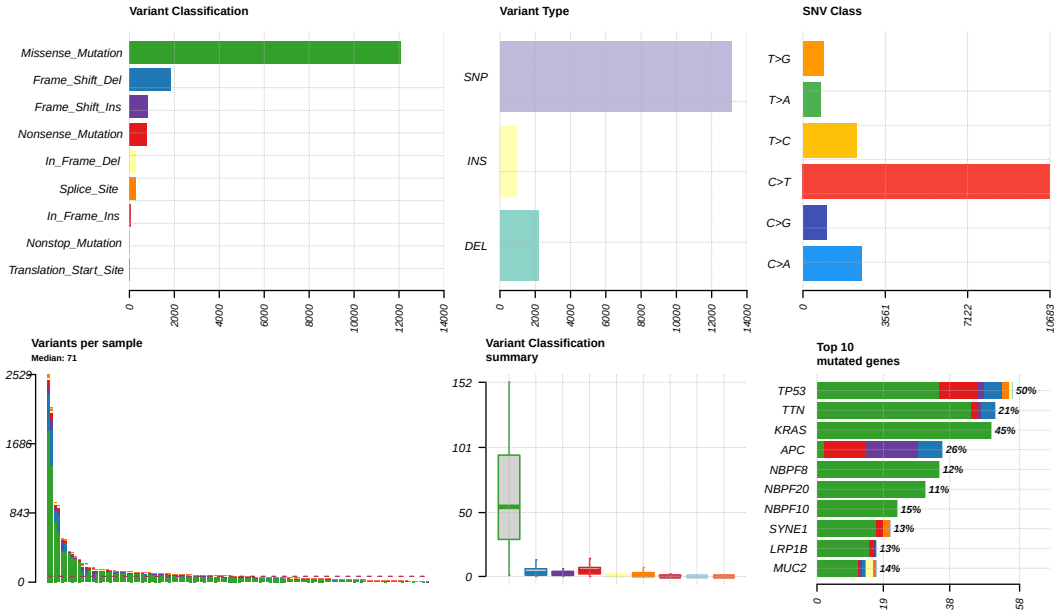

Figure 1. A summary dashboard for the entire MAF matrix that depicts the mutational landscape of the cohort.

On the summary panel plotmafSummary (Fig. 1) it is shown that the overwhelming majority of the detected variants are point substitutions (SNP), and not insertions or deletions. According to the classification of mutations, missense variants predominate – single nucleotide substitutions leading to the replacement of one amino acid (missense variant), whereas significant losses of the reading frame, splice-site and nonsense mutations occur much more rarely. On the SNV-type diagram, C→T transitions dominate; such transitions are often due to spontaneous deamination of 5-methylcytosine in CpG dinucleotides. (Olivier et al. 2010) In the Olivier et al. (2010) treatise on TP53 mutations it is noted that the majority of mutations in the DNA-binding domain are missense (87.9%) and approximately a quarter of single substitutions are precisely C:G→T:A transitions, which coincides with the predominance of C→T on the graph. The distribution of the number of variants across samples has a pronounced right-sided "heavy tail" – the median amounts to 71 variants, but in some samples the number of mutations exceeds a thousand. The summary diagram of classifications shows that missense variants sharply predominate, and the other types of mutations give only small outliers.

From the oncoplot (Fig. 2) it is evident that most (≈92.7%) of the tumors studied contain mutations in at least one of the 25 genes shown, however the mutation burden (TMB) varies greatly between samples: in the overwhelming majority, dozens of alterations are recorded, whereas several tumors demonstrate hundreds and even thousands of SNV

and indel mutations, which may reflect the presence of microsatellite instability or repair deficiency. (Center n.d.)

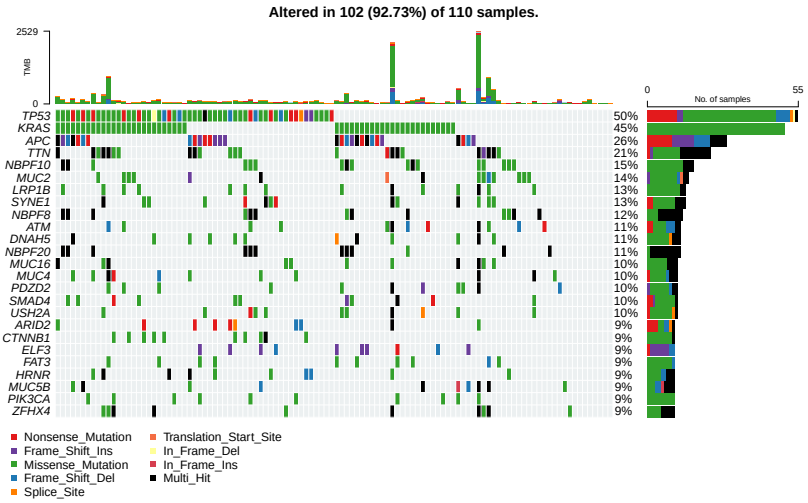

Figure 2. Oncoplot of the top 25 genes: columns represent individual tumours, rows represent genes; mutation types are color-coded. Tumour mutational burden (TMB) for each sample is shown at the top, while the proportion of patients with mutations in each gene is displayed on the right.

The graph shows that the most frequently mutating genes are TP53 (≈50% of samples), KRAS (≈45%) and APC (≈26%); such a triad is characteristic of ampullary carcinoma: in the pancreatobiliary subtype the frequency of mutations of KRAS and TP53 reaches 67%, whereas in the intestinal subtype APC and CTNNB1 are more often encountered. The appearance of SMAD4 among ten percent of the samples corresponds to reports of its more frequent inactivation in the pancreatobiliary form, whereas CTNNB1, PIK3CA and ACVR2A are associated with the intestinal form. (Rizzo et al. 2021; Hong 2021) ELF3, ARID2 and PDZD2 also enter the list of mutated genes; it is precisely ELF3 that is considered a potential driver of AC, since it plays a role in the regulation of epithelial differentiation. (Rizzo et al. 2021) In addition to classical drivers, the oncoplot shows that the mucin genes MUC2, MUC4, MUC5B, MUC16 are mutated in 9–14% of cases. Mucin proteins are high-molecular-weight glycoproteins that form a protective layer on the epithelium. (Kulkarni et al. 2017) In AMPAC, the expression of mucins clearly correlates with subtypes: all intestinal tumors express MUC2, whereas most pancreatobiliary tumors express MUC1 and MUC5AC, which is used in histological diagnosis and suggests that mucin genes may be mutated or activated in different ways in these subgroups. Although mutations of mucin genes in our set may be passenger due to their large length, the membrane mucins themselves possess an important biological function: MUC1 serves as a platform for interaction with signaling proteins, whereas MUC4 stabilizes and activates the growth receptor ErbB2; MUC16 (known as CA125) participates in adhesion and dissemination of tumor cells through interaction with mesothelin. (Bafna et al. 2010) Consequently, it is quite expedient to take into account the expression of mucins in the classification of the tumor. TTN, SYNE1, USH2A, DNAH5 and LRP1B are “giants” of the genome, therefore they often mutate by virtue of the length and architecture of the loci; their signal value for AMPAC is limited and more often reflects the overall mutational burden rather than driver

status. (Cho et al. 2018) The NBPF family (NBPF8/10/20) is located in zones of segmental duplications of chromosome 1 and gives many "noisy" findings in exome profiling. ((RDDC) n.d. Genetics 2020) On the contrary, ATM is a biologically significant marker of the pathway of response to DNA double-strand breaks; its inactivation is compatible with increased TMB and may accompany subgroups with a repair defect. (Jiang et al. 2020; Sinha et al. 2025) Single mutations of FAT3, HRNR and ZFHX4 are interpreted as markers of genomic instability/modifiers of the phenotype and more rarely determine the subtype. (Zhu et al. 2022; Arnoff and El-Deiry 2021)

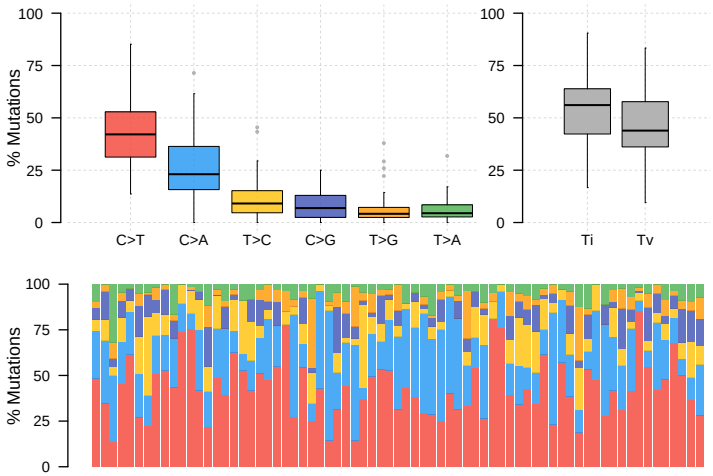

Figure 3. Summary of the spectrum of SNV substitutions across the cohort. At the top left, box plots are shown for the six classes of SNV: on the Y axis – the percentage of all SNVs in each sample; the line – the median, the box – the interquartile range, the "whiskers" – the range, the points – outliers. At the top right – the same box plots for transitions (Ti) and transversions (Tv). At the bottom – one column normalized to 100% per sample (color segments show the share of the same classes); the notation is pyrimidine-centered (for example, "C→A" includes the complementary G→T).

Based on the graph (Fig. 3) that classifies SNVs into transitions and transversions, it is evident that mutations of the C→T type markedly predominate. As a result, the fraction of transitions exceeds the fraction of transversions, although the TiTv ratio is moderate. (Son et al. 2017) Such a profile (Ti »Tv with an excess of C→T) is typical of endogenous, "background" mutational processes, which indicates a possible connection with the widely distributed mutational signature SBS1 in the COSMIC catalog. Signature SBS1 is regarded as age-related – the number of such mutations accumulates with age due to endogenous processes of DNA replication and repair. (COSMIC Mutational Signatures 2023) The presence of a substantial percentage of C→A transversions may indicate the influence of exogenous mutagens. A classic example is tobacco smoke: tobacco carcinogens (for example, benzopyrene) cause characteristic damage leading to G→T transversions in DNA, which on the complementary strand are recorded as C→A mutations. (Alexandrov et al. 2016) In the COSMIC database, such a pattern corresponds to the SBS4 signature associated with smoking. The lower part of the graph (stacked bar charts for each case) demonstrates a similar SNV profile in all tumor samples. Such uniformity may imply common background mutational processes in AMPAC.

Analysis of inter-event distances of SNVs across chromosomes (the so-called rainfall plot)

(Fig. 4) for the sample with the largest number of SNPs (TUMOR_16_dbgap), shows that mutations are distributed across the genome relatively uniformly.

Most distances between consecutive mutations are large, and no groups of dense, closely spaced SNVs are observed. In other words, no signs of mutation clustering have been identified – the local "squalls" of mutations are absent. In this sample there is no evidence of kataegis – a phenomenon of localized hypermutation. In kataegis, mutations form clusters within a limited segment of a chromosome, usually characterized by specific types of substitutions. It is likely that there is a predominance of a diffuse, random distribution of mutational events throughout the genome, characteristic of background (endogenous) processes rather than focal DNA damage.

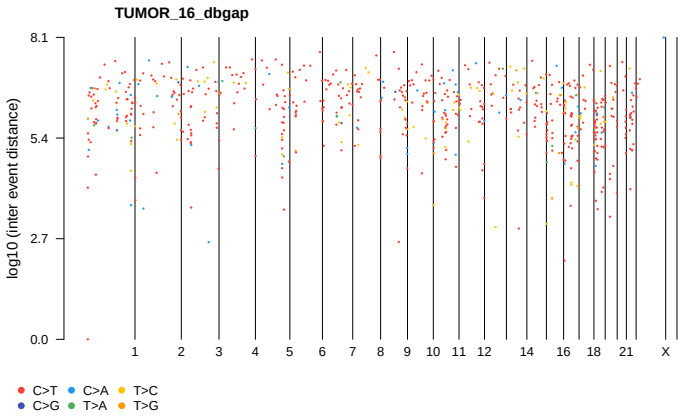

Figure 4. Rainfall plot for sample TUMOR_16_dbgap: the X-axis shows the position of SNVs along the chromosomes, and the Y-axis represents the log_{10} of the distance to the previous mutation. The color of each point corresponds to the nucleotide substitution class.

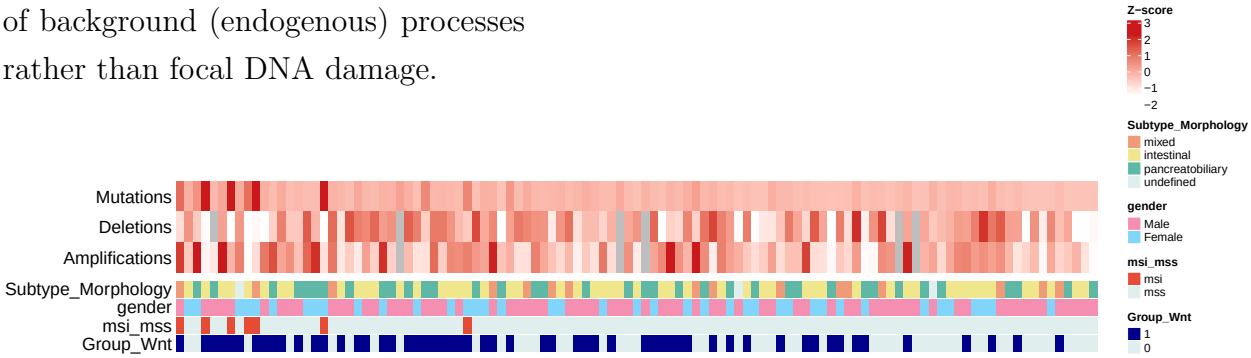

Figure 5. Clinicogenomic heatmap

A heatmap (Fig. 5) shows the Z-scores of the number of irregularities in the patients' genomes. It can be seen that a part of the patients stands out by a high mutational load: they have bright-red columns in the Mutations row. These same samples often have an MSI label (a red cell in the msi_mss row) and an activated Wnt pathway (Group_Wnt=1, dark-blue labels). (Zhuravleva et al. 2025; Tsagkalidis et al. 2023) This corresponds to the known "hypermutagenic" phenotype in deficiency of the MMR system (microsatellite instability), when an excessive number of C>T mutations in CpG islands leads to a sharp increase in the number of point mutations. Such tumors usually have the intestinal subtype and a pronounced immune response (an immunogenic microenvironment). The most noticeable regularity is the connection between the morphological subtype, Wnt activation, and MSI. Patients with the intestinal subtype (yellow labels) much more often show both Wnt activation (dark-blue

labels, Group_Wnt=1) and MSI than patients with the pancreatobiliary subtype (green labels). In published studies, the intestinal subtype of AMPAC in 67% of cases had mutations in the Wnt pathway, whereas in the pancreatobiliary subtype – only in ≈30%. (Gingras et al. 2016) This means that intestinal tumors are prone to combined hypermutagenesis and Wnt activation: defects of MMR and accumulation of C>T mutations are combined with mutations in APC/CTNNB1 and disruption of the Wnt signaling pathway. In contrast, the pancreatobiliary subtype more rarely has MSI and Wnt mutations, and most of their genetic drivers are associated with other pathways (for example, KRAS, TP53, PI3K/RTK). The patient's sex on the map is distributed almost randomly and clearly does not correlate with the profile of mutations or CNA. This reflects the absence of a strong gender effect on the molecular profile of AMPAC: although in the population men are diagnosed more often, the very genetic mechanisms of tumor formation in men and women look similar. (Ramai et al. 2019)

By depicting combinations of alterations, the map makes it possible to draw conclusions about mutagenesis pathways. In total, the combination of observations points to two main scenarios of oncogenesis: one – MMR-deficient hypermutagenesis with an immune response and Wnt activation, the other – genomic CNV disturbances (amplifications/deletions) leading to tumor growth without a high mutational load. These patterns are consistent with Zhuravleva et al. (2025) and help to understand the mechanisms of tumor formation embedded in them.

3.2 Distribution of SNPs

The histogram (Fig. 6) of the distribution of the number of single-nucleotide polymorphisms (SNPs) across the 71 analyzed tumor samples shows a pronounced asymmetry. In most patients the number of detected SNPs is relatively small, concentrating at the level of several tens of mutations, however there is a long "tail" of the distribution due to several samples with a very high number of SNPs. In total, 3805 SNPs were found in these 71 samples, the median value amounts to 26 SNPs per sample. This means that more than half of the tumors contain on the order of several tens of point mutations, whereas individual tumors substantially stand out from the general picture with a number of SNPs an order of magnitude

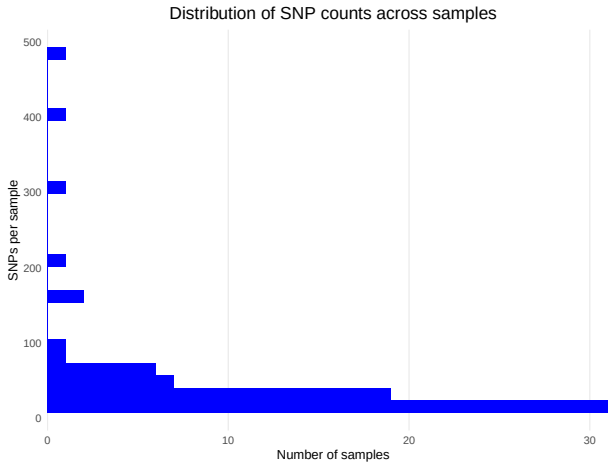

Figure 6. Histogram summarizing SNP burden across 71 samples (Total SNPs = 3,805; median = 26)

higher than the median.

The bar chart (Fig. 7) displaying the number of SNPs in each sample clearly illustrates the indicated variability. Only a small subgroup of tumors demonstrates numbers of SNPs in the range of hundreds (in individual cases more than 400 mutations per sample), whereas in most samples the values are significantly lower. The observed spread of the number of mutations between samples may be due to biological differences of the tumors – for example, the presence of a hypermutational phenotype, the degree of genomic instability or the action of various mutagenic factors – as well as to some extent to technical factors (different sequencing depth or sample quality). In the context of the analysis of the mutational profile of tumors, the obtained data indicate a significant heterogeneity of the mutational burden: the main mass of tumors has a relatively low number of somatic SNPs, and rare cases with an extremely high number of mutations stand out as potentially special subgroups with increased mutational activity.

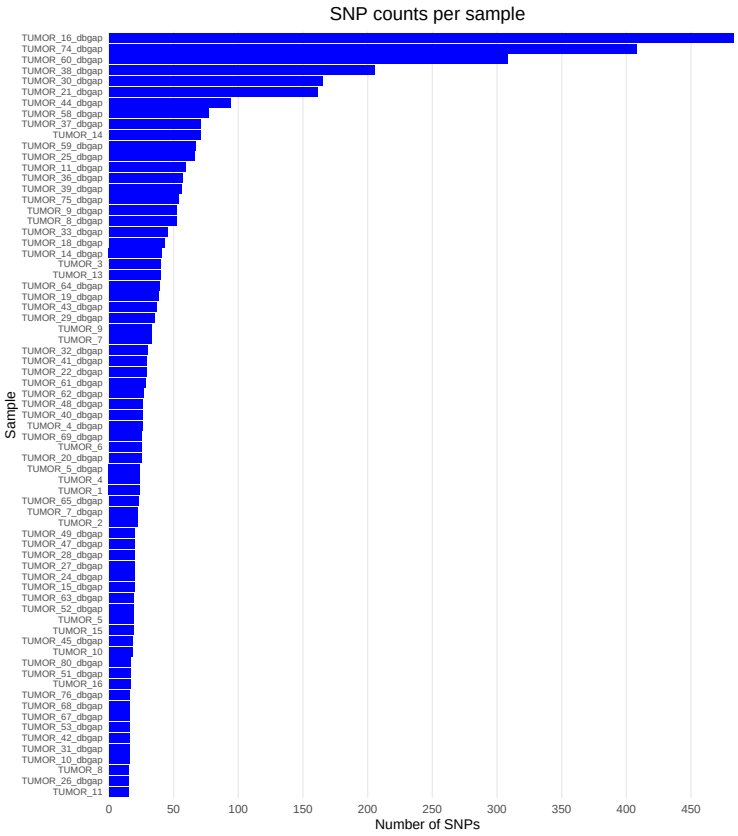

Figure 7. Horizontal bar chart showing SNP burden for each sample, sorted in descending order

3.3 Signature count selection

A multi-panel plot (Fig. 8) of decomposition quality (NMF rank survey) for ranks k=2...7 (500 NMF restarts for robustness) shows summary indicators that make it possible to assess the stability and accuracy of the model. The cophenetic correlation coefficient of the consensus matrix characterizes the stability of the clustering of samples at a given number of signatures (a value of 1 corresponds to ideal cluster stability). It rapidly increases and reaches a maximum at k=3, after which it noticeably falls. The rank at which the cophenetic correlation begins to decrease is considered optimal, which points to k=3 as the first cycle of deterioration in cluster concordance. The dispersion coefficient increases monotonically, after k = 3 the growth rate is insignificant. Silhouette metrics, which assess the separability of clusters (values near 1 indicate well-separated groups), generally decrease as k increases.

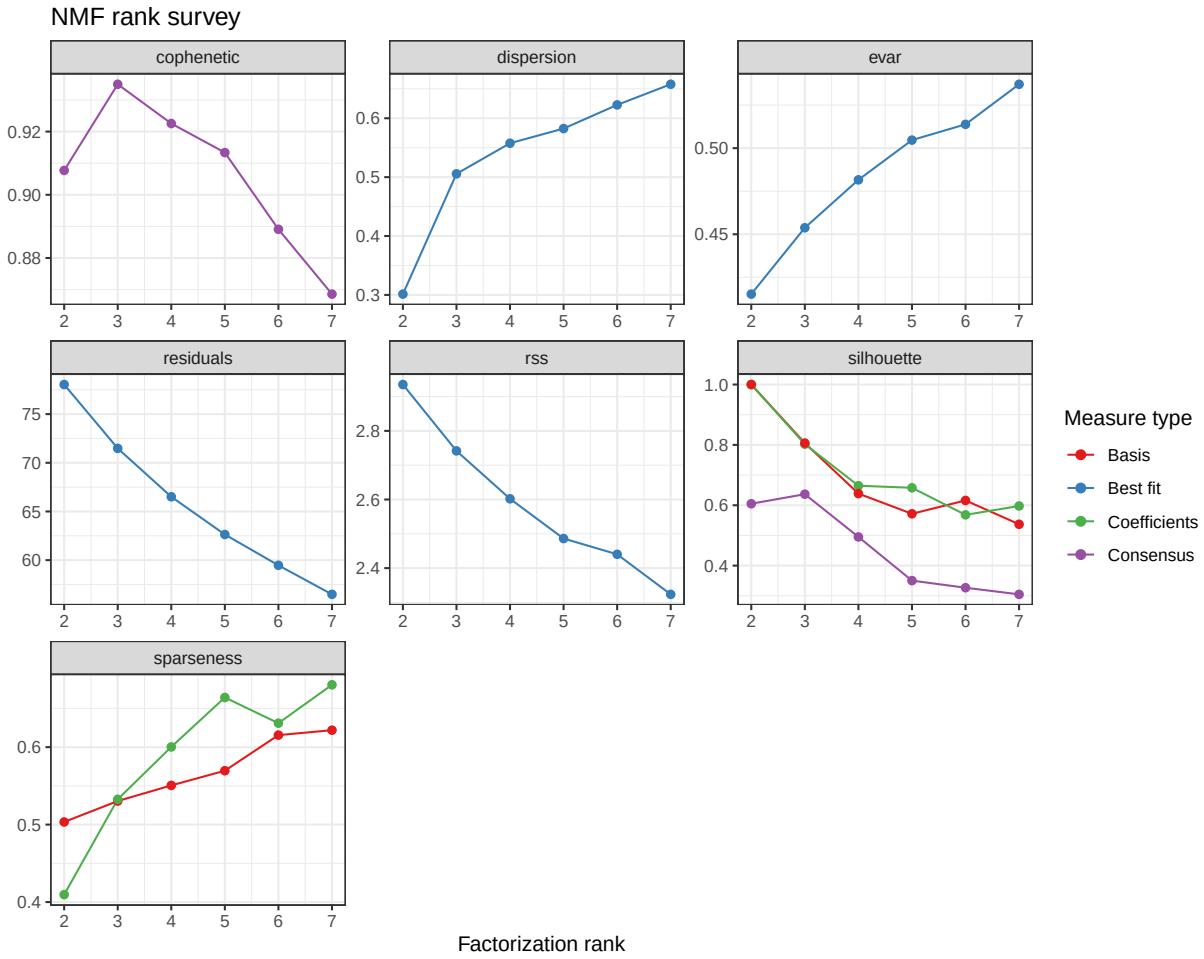

Figure 8. NMF rank survey. For each candidate factorization rank k on the x-axis, 500 NMF runs are performed and summary quality metrics are computed together with the consensus matrix. Panels report: cophenetic correlation of the consensus matrix (cluster stability; higher is better); Kim & Park's dispersion coefficient of the consensus matrix (reproducibility of clusters; higher indicates sharper consensus); evar – explained variance $= 1 - \frac{\text{RSS}}{\sum v_{ij}^2}$ (higher is better); residual approximation error and residual sum of squares of the best fit (lower is better); average silhouette widths computed from the basis W, coefficients H, and the consensus clustering (higher is better); Hoyer's sparseness for W and H (0–1, larger values indicate more localized, sparse factors) (Gaujoux and Seoighe 2025b; Gaujoux and Seoighe 2024a; Gaujoux and Seoighe 2025a; Gaujoux and Seoighe 2024b)

In this survey, the consensus silhouette shows a small local rise at k=3 before declining; coefficients exhibit a local maximum at k=5; basis has a minor uptick at k=6; and best fit decreases nearly monotonically. These patterns indicate a brief improvement in separability for consensus at k=3 and for coefficients around k=5, with overall degradation as k grows. (Zitnik 2025) Reconstruction error metrics – residuals and RSS (Residual Sum of Squares)– decrease as k increases, reflecting a better fit of the data. Across k = 2...7, both curves fall nearly linearly. The proportion of explained variance (evar) increases from k=2 to k=7. From about $k \approx 3$–4 onward, additional signatures yield only incremental increases. The sparsity indicator (sparseness) for the matrices W (basis signature profiles) and H (coefficients of sig-

natures in samples) increases with increasing k. Sparsity takes the value 1 when there is only one non-zero component in a vector, and 0 when all components are equal. Sparseness for the W matrix rises monotonically with k. For the H matrix it is non-monotonic: it increases up to k=5, dips at k=6, then rises again at k=7. Thus, signatures become more specialized, but sample coefficients do not increase strictly monotonically with rank. (Gaujoux et al. 2024)

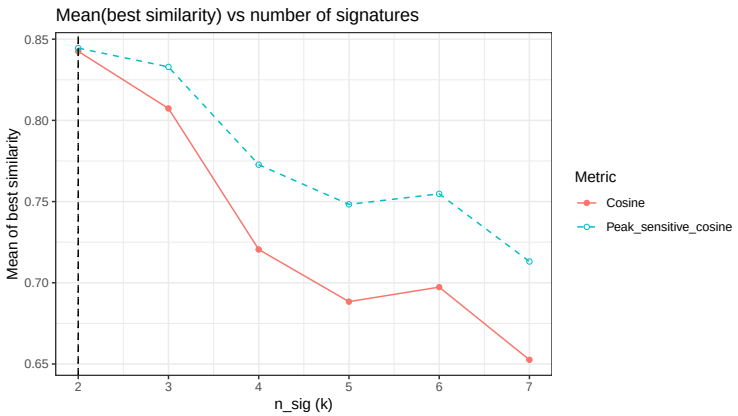

Figure 9. Mean (best) similarity vs. number of signatures for cosine (red) and peak-sensitive (blue) cosine metrics.

The "Mean (best) similarity vs. number of signatures" plot (Fig. 9) shows how the average "best" similarity of the extracted signatures (500 NMF restarts for robustness) to any COSMIC signature changes for different numbers of factors k. Two metrics are shown: the standard cosine and the peak-sensitive cosine. The latter differs in that it compares only the informative peaks of the profiles (masking the low-amplitude background), and is therefore more robust to noise and the shape of the "tails." It declines steadily with increasing k: the peak is at k=2, remains relatively high at k=3, and then falls, which is consistent with "over-decomposition" and fragmentation of signals that do not have direct analogs in the catalog. Importantly, the peak-sensitive metric is consistently higher than the standard one at all k: focusing on the peak bins yields a slightly higher similarity estimate when the dominant features of the profiles coincide, even if the background differs. Tables comparing the signatures with COSMIC for the different metrics can be found in the appendix (Tables 4, 5, 6, 7, 8, 9).

Based on the aggregate of metrics, I consider k=3 optimal – this is consistent with the Zhuravleva et al. (2025) decisions.

3.4 Final k = 3 signature extraction

After applying NMF (k=3, 1000 iterations), three stable signatures were obtained (Fig. 10). Each of the extracted signatures was matched with the known mutational signatures in the COSMIC v3 database (see the correspondence Table 1) by calculating cosine similarity (standard and "peak-sensitive"). Below is the interpretation of each signature taking into account the best matches:

1. **Signature 1** S1 demonstrates the greatest similarity to COSMIC SBS1 by both metrics (0.878). The COSMIC signature SBS1 corresponds to spontaneous deamination of

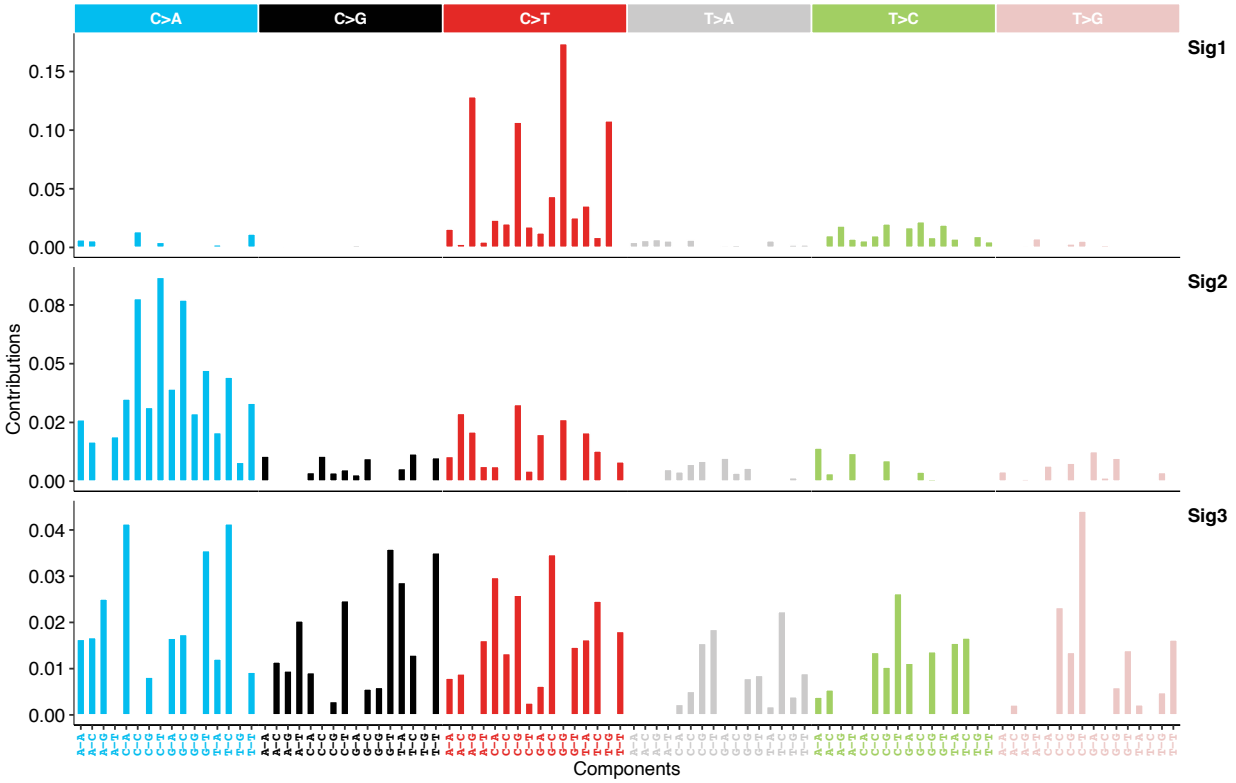

Figure 10. Three de novo single-base substitution (SBS) mutational signatures (Sig1–Sig3) plotted in COSMIC style. Each panel shows a normalized 96-channel trinucleotide spectrum grouped by the six substitution classes (C>A, C>G, C>T, T>A, T>C, T>G); the x-axis lists the 96 contexts, the y-axis is the relative contribution – taller bars mark sequence motifs preferentially targeted by that signature. Such profiles are interpreted by comparing their shapes to the COSMIC reference compendium to infer underlying processes. (Alexandrov et al. 2020)

Signature	Rank	COSMIC (cosine)	Cosine similarity	COSMIC (peak-sensitive)	Peak-sensitive similarity
Sig1	1	SBS1	0.878	SBS1	0.878
Sig1	2	SBS6	0.873	SBS6	0.873
Sig1	3	SBS15	0.776	SBS15	0.783
Sig2	1	SBS4	0.833	SBS95	0.847
Sig2	2	SBS29	0.827	SBS4	0.846
Sig2	3	SBS95	0.822	SBS29	0.833
Sig3	1	SBS3	0.711	SBS3	0.774
Sig3	2	SBS40a	0.633	SBS40a	0.669
Sig3	3	SBS40b	0.592	SBS40b	0.645

Table 1. Comparison of k = 3 signatures with COSMIC: cosine and peak-sensitive cosine metrics.

5-methylcytosine – the so-called “clocklike” signature that accumulates with age. The high similarity of S1 to SBS1 indicates that in this signature mutations of the C→T type in the CpG context, characteristic of age-related mutagenesis, predominate. In

addition, S1 substantially correlates with COSMIC SBS6 (0.873) and SBS15 (cosine 0.776, peak-sensitive cosine 0.783). Both of these signatures are associated with defects of the DNA mismatch repair system (MMR) and are often detected in microsatellite instability. Such a distribution of similarities indicates that signature S1 reflects the combined influence of two processes: background accumulation of age-related mutations and a hypermutational process in the case of a defect of Mismatch Repair. Equally high scores by the standard and the peak-sensitive cosine similarity indicate that the dominant mutational pattern (for example, the abundance of CC→T in CpG) consistently coincides with the reference signature SBS1. Consequently, S1 can be interpreted as a signature characterizing a cluster of tumors with age-related changes of DNA against the background of a defective MMR process, which coincides with the known etiological contribution of SBS1 (spontaneous DNA damage) and SBS6/15 (inability to repair replication errors). (Alexandrov et al. 2020; COSMIC 2023b)

2. **Signature 2.** S2 by ordinary cosine similarity is the closest to COSMIC SBS4 (0.833) – the classic "tobacco" signature associated with the effect of carcinogens of tobacco smoke. A close value is also the similarity with SBS29 (0.827), which is known as the signature in chewing tobacco. Both results indicate that S2 reflects mutagenesis characteristic of the exogenous effect of nitrosamines and polycyclic aromatic hydrocarbons of tobacco, leading predominantly to transitions of the C→A type. However, the peak-sensitive metric revealed the highest similarity of S2 with COSMIC SBS95 (0.847), whereas for SBS4 it is only slightly lower (0.846). The COSMIC signature SBS95 is not associated with a known biological agent, and is considered an artefact signature (a possible technical artifact of sequencing). (COSMIC 2023a) The superiority of SBS95 by the peak-sensitive metric implies that the most pronounced mutational "peaks" in the spectrum of S2 coincide with the artefact pattern. At the same time, the overall profile of S2 by the standard cosine is closer to the tobacco signature SBS4. Such duality hints that S2 may reflect the contribution of tobacco carcinogens (SBS4/SBS29) against the background of a specific artefact pattern (SBS95), possibly arising due to features of sample processing or sequencing (for example, oxidative DNA damage during sequencing preparation). (COSMIC 2023b; Alexandrov et al. 2020)

3. **Signature 3.** S3 shows moderate similarity to COSMIC SBS3 (cosine 0.711; peak-sensitive 0.774), which indicates correspondence to the signature of homologous recombination (HR) defect. COSMIC SBS3 is known as a mutational signature associated with impairment of repair of double-strand breaks of the BRCA1/2-pathway type, often accompanied by a specific spectrum of substitutions and large deletions. A high value of the peak-sensitive similarity for SBS3 implies that the most characteristic mutations of S3 (probably certain triplets enriched, for example, in T→G and others) coincide well with the peak picture of the HR-deficient signature. In addition to this, S3 has

a noticeable similarity with the signatures COSMIC SBS40a/SBS40b (ordinary cosine 0.633 and 0.592; peak-sensitive 0.669 and 0.645). The signatures of the SBS40 family do not have an established etiology and are considered "flat," background mutational profiles. It is known that SBS40, similar to SBS5, manifests a clocklike (age-related) character of accumulation of mutations and correlates with the age of the tumor. (Hwang et al. 2025) Overall, the similarity metrics indicate that the main driver of S3 is processes associated with loss of function of DNA HR repair, whereas the accompanying uniform contribution of SBS40 reflects an undefined background (a possible "clocklike" component). Such a combination is consistent with the fact that in some tumors (for example, with mutations of a BRCA-like pathway) a unique spectrum of substitutions is observed against the background of general age-related mutagenesis. (COSMIC 2023b)

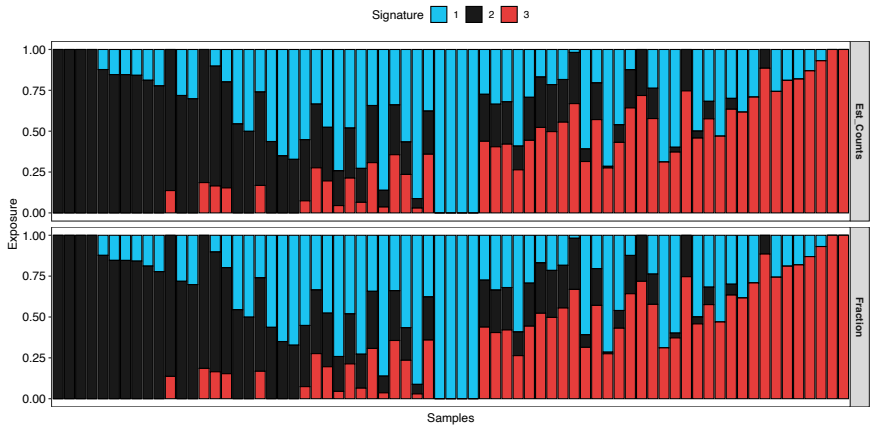

Figure 11. Visualization of the exposure of the three signatures across the selected samples shows that the contribution of each signature varies substantially between tumors. The upper panel (normally) reflects the absolute number of mutations attributed to the signatures (Est. Counts), and the lower one – their share in the total mutational burden of the sample (Fraction). This makes it possible to assess which signature predominates in each tumor and to detect "pure" cases (dominance of one signature) or a mixed pattern of mutagenesis.

Visualization of the three signatures across all selected samples (Fig. 11) shows pronounced heterogeneity of composition. In the left part of the plot, a segment of almost "pure" profiles is noticeable, where Signature 2 dominates; this means that the majority of mutations in these tumors are attributed to a single subprocess. Further along the sample axis, a transition zone is traced: mixed signature profiles appear, and in a number of cases Signature 1 becomes the largest source by Fraction. In the right part of the cohort, a cluster of samples is formed with an almost homogeneous predominance of Signature 3, and the share of the other subprocesses is minimal. Thus, taken together, three stable patterns are encountered: monosignature profiles (almost entirely the 2nd or 3rd and several samples with the 1st), two-signature mixtures with the 1st as the leading one, and balanced mixtures

of all three. It should also be noted that Est_Counts coincides with Fraction, since earlier we performed normalization to 1.

3.4.1 Comparison of S1–S3 with signatures from Zhuravleva et al. (2025)

The signatures obtained in our analysis partially coincide with the signatures described in the original article by Zhuravleva et al. (2025), however there are also substantial differences. Signature 1 in the original was characterized by the predominance of age-related mutagenesis signatures and defective MMR – namely COSMIC SBS1 and SBS6 with very high similarity (cosine ≈ 0.99 and 0.96). Our signature S1 demonstrates a similar profile: it closely corresponds to SBS1 and signatures associated with MMR defect (SBS6, SBS15), indicating that we managed to reproduce this biological subtype (hypermutational, MSI-associated background). In contrast to this, signature 2 of the original was associated with a different combination of signatures – predominantly with SBS40 (unknown, background) and SBS5 (A-nucleotide damage and transcription-coupled NER). In our analysis, however, signature S2 does not show similarity to SBS40/5, and instead is correlated with tobacco signatures (SBS4, SBS29) and the artefact SBS95. This means that our detected S2 diverges from the original signature S2. Similarly, signature 3 of the original work was defined by a combination of SBS5 and SBS9 – a background clocklike signature plus a mutational process induced by DNA polymerase η (for example, as in somatic hypermutation in lymphoid cells). Our S3, on the contrary, is correlated with SBS3 and SBS40 – that is, with a defect of homologous recombination and an undefined background component, not present in the description of the original S3. Thus, signature S3 in the reproduced analysis does not substantially coincide.

3.5 Cluster analysis

Sig1	Sig2	Sig3	cluster	Majority signature
7	6	18	1	Sig3
2	14	8	2	Sig2
16	0	0	3	Sig1

cluster	Sig1	Sig2	Sig3	Dominant by mean
1	0.27	0.28	0.45	Sig3
2	0.19	0.54	0.26	Sig2
3	0.74	0.11	0.15	Sig1

Table 2. Cluster majority dominant counts (left) and cluster mean exposures (right).

Normalised three-nucleotide mutational frequency profiles heatmap (Fig. 12) shows the separation of samples into three clusters by the similarity of mutational patterns. The clustering results are supported by signature distribution Tables 2. In the majority signature table by clusters it is shown which mutational signature gives the largest contribution in each sample, and it is counted which signature dominates among the samples of each cluster. It is seen that cluster 1 is distinguished by the predominance of signature S3, cluster 2 – of signature S2, and cluster 3 – of signature S1. The same is reflected by the table of mean

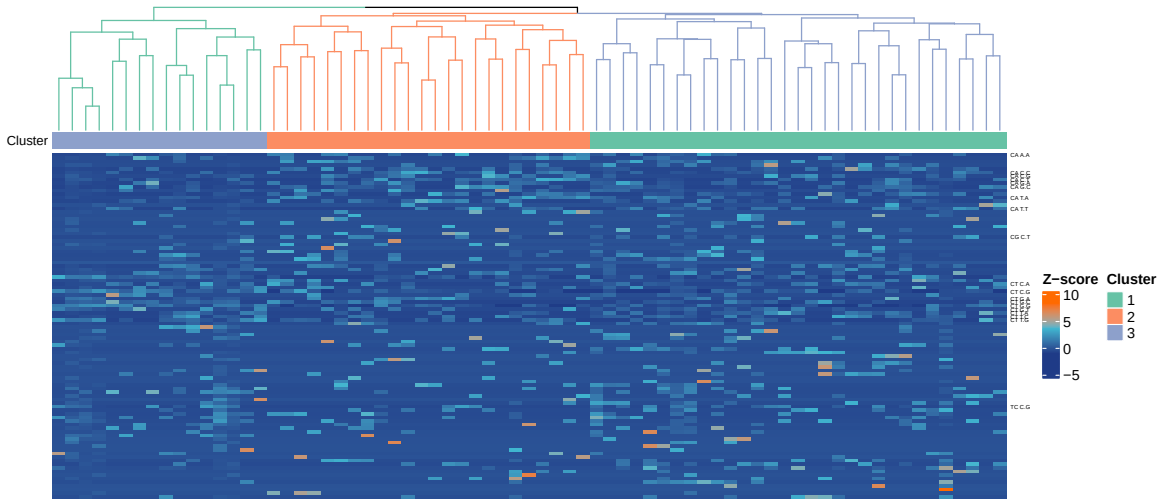

Figure 12. Mutational signature-associated clusters in AMPAC (n=71). Normalised three-nucleotide mutational frequency profiles. Annotation shows the top 20% contexts by mean absolute Z-score.

exposures (the diagram of signature exposures per sample by clusters can be seen in Fig. 13): for each cluster the share of the corresponding signature is maximal (for example, in cluster 3 the average contribution of S1 is ≈74%, which is significantly higher than the shares of S2 and S3). Both approaches – the qualitative assessment of the dominant signature in samples and the quantitative assessment of average contributions – give consistent results. Each cluster is clearly characterized by its own leading signature, which indicates a strict separation of tumor groups by the prevailing mutational process.

- **Cluster 1** – Sig3-dominant (SBS3, SBS40a/b). Cluster 1 can be characterized as a subgroup of HR-deficient AMPAC. It is likely that these tumors have mutations in genes responsible for homologous recombination (BRCA1/2, RAD51C, etc.), which leads to a high overall mutational burden and genomic instability. (Zemet et al. 2023) A high level of mutations is often accompanied by the formation of neoantigens and an influx of immune cells. In the Zhuravleva et al. (2025) study, cluster C3 was described as being due to the activity of polymerase η (SBS9) against the background of the NER process. We, on the other hand, observe in cluster 1 a clear predominance of the SBS3 signature associated with an HR defect. These two descriptions do not contradict each other: HR defects (for example, a BRCA1/2 mutation) can lead to an ultra-mutated phenotype with the recruitment of alternative replicative polymerases (such as η) to bypass DNA damage. (Tomlinson et al. 2024) Thus, our cluster 1 can be considered the

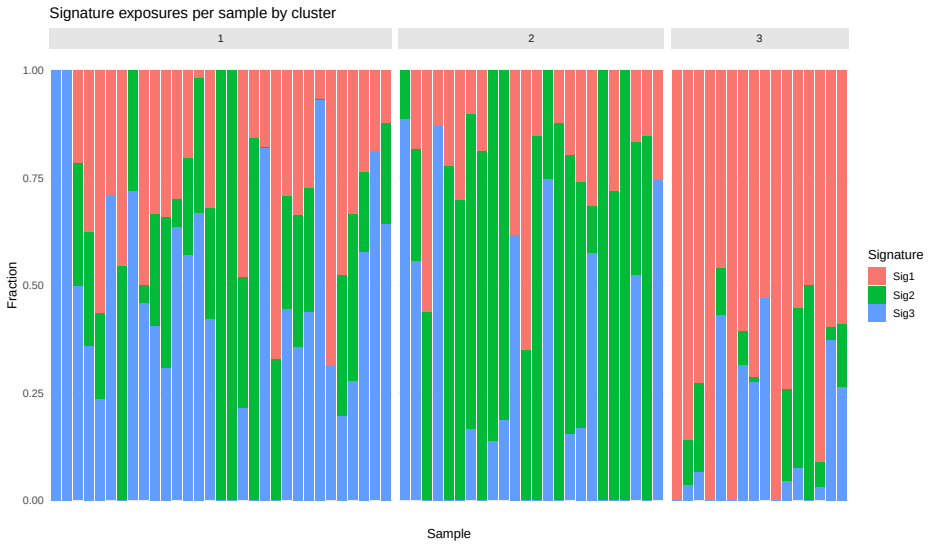

Figure 13. Signature exposures per sample by cluster.

equivalent of C3, but emphasizing precisely the BRCA-like mechanism of mutagenesis.

- **Cluster 2** – Sig2-dominant (SBS4, SBS29, SBS95). Cluster 2 in its characteristics corresponds to a group of tumors exposed to external carcinogens. In the context of AMPAC the most likely candidate is the influence of smoking. It is known that smoking is a risk factor for ampullary/pancreatobiliary tumors, and our cluster 3 probably reflects this connection. In the Zhuravleva et al. (2025), clusters C2 and C3 were associated with processes of transcription-coupled excision repair (TC-NER), with cluster C2 not having additional features of polymerase errors. Our cluster 3 shows precisely the tobacco signatures SBS4/29, which is consistent with the TC-NER mechanism (removal of benzoapyrene adducts) and indicates that it is closer to C2. A new observation in our analysis is the presence of the signature SBS95 in cluster 3, however it probably does not reflect a new biological process, but relates to variations of the tobacco spectrum or technical effects.

- **Cluster 3** –Sig1-dominant (SBS1, SBS6 and SBS15). SBS1 is an age-related "clock-like" signature (C→T in CpG in spontaneous deamination of methylcytosine), usually background for all tumors, but in this cluster it is especially pronounced. SBS6 and SBS15 are classic signatures of microsatellite instability (deficiency of the MMR system): SBS6 indicates failures in the DNA mismatch repair system, and SBS15 is often found together with it in MMR-deficient tumors. The totality of these signatures reflects loss of function of MMR genes (for example, MLH1, MSH2) with accumulation of unrepaired point mutations and indels throughout the genome – the phenomenon of hypermutation. Cluster 3 in fact corresponds to the MSI-high (MMR-deficient) molecular subtype of AMPAC (analogous to cluster C1 according to Zhuravleva et al. (2025)) with an extremely high mutational burden. These microsatellite-unstable tu-

mors usually respond well to PD-1/PD-L1 inhibitors in other types of cancer, although in AMPAC the effectiveness of immunotherapy is not yet fully clear. Overall, cluster 3 is a hypermutational MMR-deficient type of AMPAC.

3.6 Clustering robustness

metric	linkage	k_auto	chooser	mean_silhouette	k_dtc	sil_dtc	k_sil	sil_best
euclidean	average	2	DynamicTreeCut	0.2108	2	0.2108	2	0.2108
spearman	average	2	DynamicTreeCut	0.1902	2	0.1902	2	0.1902
spearman	complete	2	Silhouette	0.1394	4	0.0508	2	0.1394
euclidean	complete	3	Silhouette	0.1219	5	0.0091	3	0.1219
pearson	average	4	DynamicTreeCut	0.0638	4	0.0638	8	0.0719
cosine	average	3	DynamicTreeCut	0.0615	3	0.0615	8	0.0700
euclidean	ward.D2	2	Silhouette	0.0493	4	0.0156	2	0.0493
pearson	complete	4	DynamicTreeCut	0.0489	4	0.0489	2	0.0585
cosine	complete	5	DynamicTreeCut	0.0380	5	0.0380	2	0.0560

Table 3. Clustering robustness summary

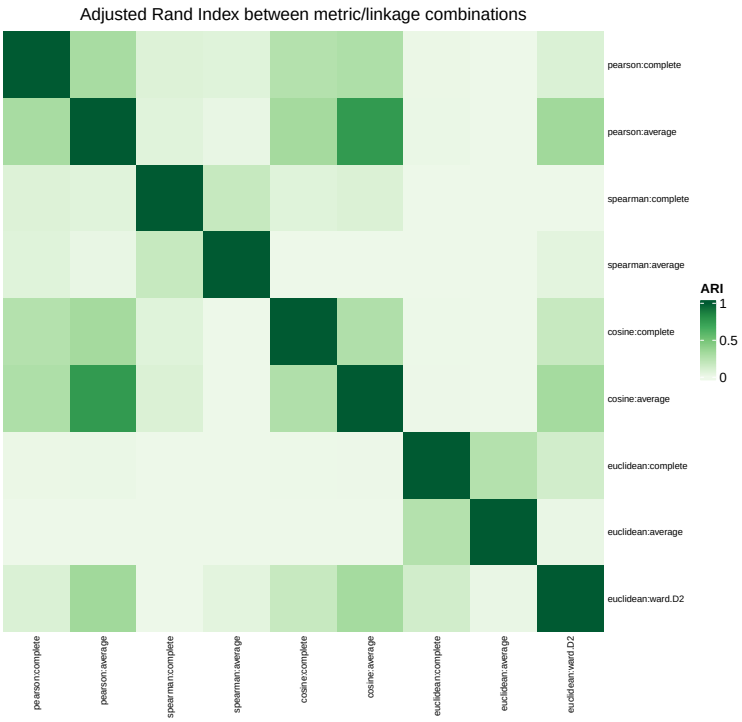

Figure 14. Heatmap of the consistency of cluster auto-partitions. Dark green → high consistency (almost the same clusters), light → low (almost random).

The table (Tab. 3) with silhouette statistics shows extremely low values of the average silhouette width for all methods ($0.038-0.211$). Even the best case (Euclidean distance + average linkage, k=2) gives a silhouette ≈ 0.211, which by generally accepted criteria is considered a very weak cluster structure. For most methods, the average silhouette is significantly below 0.15, which practically indicates the absence of clear groups (silhouette < 0.25 indicates the absence of a substantial cluster structure).DynamicTreeCut generates more clusters (k_dtc up to $4-5$), but in these cases the silhouette falls almost to zero (for example, Euclid/complete: k_dtc=5, sil_dtc$\approx$0.009). This indicates that splitting into a larger number of groups leads to excessive fragmentation without a clear improvement in

cluster quality. In addition, in a number of cases the choice by the silhouette (silhouette method) outputs more clusters (k_sil=8) with a slightly larger silhouette ($\leq$0.07), but the absolute values of the silhouette remain small.

The ARI heat map between all combinations of metrics and linkage (Fig. 14) confirms extremely weak consistency of the clustering solutions. Almost all ARI values are close to zero (white cells), that is, the various clustering solutions practically do not agree with each other (ARI $\approx$ 0 means closeness to random correspondence). Only within a single metric are moderately high ARIs observed (pearson:average $\leftrightarrow$ cosine:average $\approx$0.79). This indicates that with an unchanged metric the clustering is somewhat similar (in the composition of clusters), but even in these cases the ARI is far from 1. Between different metrics (for example, Euclidean vs Pearson) the ARI is practically equal to zero, indicating completely different partitions. That is, no clustering method reproduces the structure of another, which once again underscores the low stability of the results.

The most probable reasons for low-quality clustering:

- Normalization of profiles. If the data are brought to frequencies (the sum of components is the same for all samples), then classical distance measures (especially Euclidean) cease to take into account the difference in the scale of mutational burden. This additionally smooths differences between samples and can conceal even moderate groupings. When using normalization 2 (see section 96-trinucleotide matrices preparation) the Euclidean results are significantly higher (best values around 0.7). Metrics based on similarity (Cosine, Pearson) also reflect only the correlation of the shapes of profiles, not taking into account absolute differences, which in these data does not lead to a clear separation.

- High dimensionality and sparsity of the data. Arrays of trinucleotide profiles have a large number of features (96 contexts), and many components can be small or zero. It is known that in a high-dimensional space the distances between points tend to equal values, worsening the separability of clusters. In addition, the very nature of mutational data – "high dimensionality and sparsity" – complicates analysis and leads to weak clustering.

- Biological heterogeneity. Normalized mutational profiles often lie on a continuum of changes without pronounced gaps between groups. Cancer diseases possess pronounced heterogeneity in the mutational background, therefore the samples rather differ smoothly. In such conditions statistical methods "see" a weak or absent cluster structure, which is reflected in small silhouettes and unstable partitions.

- Noise and variability of the data. Small biological differences between samples can be "drowned out" by noise, especially with small differences between profiles. As a result, agglomeration algorithms split the data arbitrarily, not finding clear groups.

3.7 Cluster survival analysis

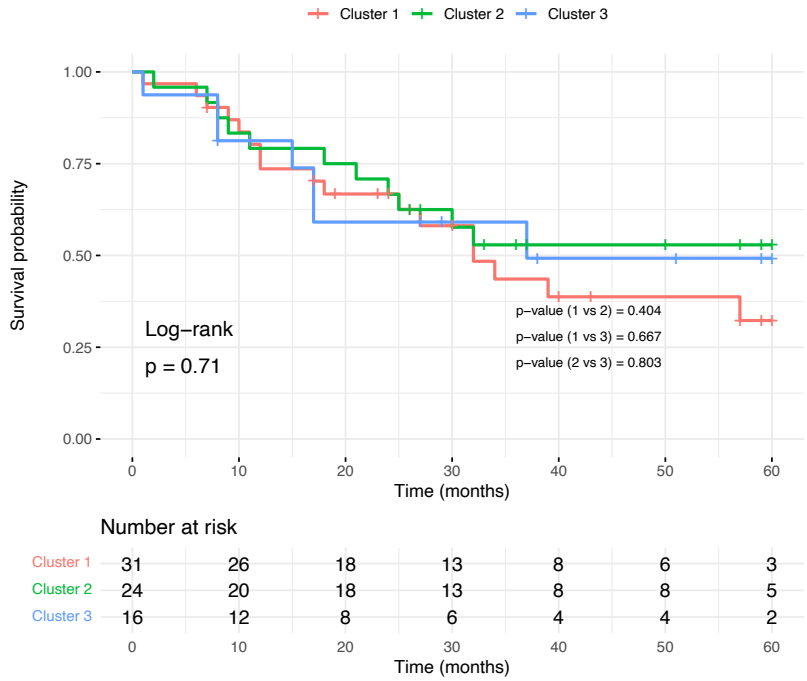

Figure 15. Kaplan-Meier curves of overall survival stratified by clusters. Tick marks on the curves denote censored observations. Below is the "number at risk" table showing the count of patients in each group at key time points, and the log-rank p-value is displayed in the lower-left corner.

Kaplan–Meier survival analysis did not reveal statistically significant differences between the three signature clusters of ampullary carcinoma (log-rank p=0.71, all pairwise comparisons p>0.4) – the curves practically coincide. Probable reasons: limited statistical power due to the small cohort size, incorrect clustering, overlapping biological mechanisms (for example, one patient may combine signs of HRD and MMR deficiency). In addition, in ampullary carcinoma the choice of therapy usually does not take into account the mutational profile (there is no targeted therapy for HRD or MSI), therefore biological differences are not reflected in survival. According to the literature, some signatures are associated with response to treatment: for example, the HRD signature SBS3 is associated with increased sensitivity to platinum drugs and PARP inhibitors. In studies of gastrointestinal tumors it has been shown that patients with dominant SBS3 receive a better response to platinum therapy. (Ghareyazi et al. 2021; Tsang et al. 2023) Signatures associated with MMR deficiency (for example, SBS15) are characteristic of MSI-high gastrointestinal tumors, which usually have a more favorable prognosis and a good response to immune checkpoint inhibitors. (Farmanbar et al. 2023) Thus, despite the absence of differences in survival in our sample, the literature data confirm that HRD- and MMR-associated signatures can influence sensitivity to therapy and prognosis in related gastrointestinal tumors.

4 Conclusions

This project reproduced the core steps of the mutational signature – based classification of ampullary carcinoma and yielded several biologically insights. From the cleaned mutation data, NMF identified three robust signatures. Sig1 showed a strong match to COSMIC

signatures SBS1, SBS6 and SBS15, consistent with spontaneous deamination of methylated cytosines and defective mismatch repair. Sig2 corresponded primarily to SBS4 and SBS29 (tobacco smoking and aflatoxin exposure) and included SBS95, a recently described signature of tobacco carcinogens. Sig3 matched SBS3, a hallmark of homologous-recombination deficiency, and was reminiscent of the polymerase-η–associated SBS9 in the original study. Comparing signature exposures across samples revealed three clusters with distinct aetiological themes:

- **Cluster 3 (Sig1-dominant)** – characterised by high levels of SBS1/SBS6/SBS15, this cluster reflects a hypermutational, mismatch-repair–deficient phenotype. The enrichment of SBS6 and SBS15 indicates microsatellite instability, and the accumulation of unrepaired point mutations and indels is analogous to the MSI-high subtype described in the Zhuravleva et al. (2025) study. In other cancers, tumours with MMR deficiency respond well to PD-1/PD-L1 blockade; whether this applies to AMPAC remains to be tested.

- **Cluster 2 (Sig2-dominant)** –enriched for SBS4, SBS29 and SBS95, this cluster likely represents tumours exposed to exogenous carcinogens, most plausibly tobacco smoke. SBS4 arises from bulky DNA adducts produced by benzoapyrene in tobacco, while SBS29 and SBS95 appear to capture variations of this smoking signature. The correspondence between this cluster and the TC-NER-associated C2/C3 clusters in the original article suggests that environmental mutagens are important drivers of AMPAC in a subset of patients.

- **Cluster 1 (Sig3-dominant)** – dominated by SBS3, this cluster points to homologous-recombination deficiency (HRD), a defect in DNA double-strand-break repair often caused by BRCA1/2 mutations. HRD can lead to an ultra-mutated phenotype and recruitment of translesion polymerases such as polymerase , linking this cluster to the SBS9-driven C3 group reported by Zhuravleva et al. (2025). Tumours with HRD signatures may be sensitive to platinum drugs and PARP inhibitors; thus, identifying this subgroup has therapeutic implications.

Although the extracted signatures broadly mirrored those described by Zhuravleva et al. (2025), several differences emerged. First, our cluster assignments differed: the HRD-like Sig3-dominant cluster corresponded to the original study's C3, while our Sig1-dominant cluster aligned with their C1 (MMR-deficient). Second, the presence of SBS95 in the smoking-related cluster suggests either previously unrecognised variation within tobacco-associated processes or technical artefacts. Third, hierarchical clustering of signature exposures produced low silhouette scores and poor reproducibility across distance metrics, indicating that the mutational profiles form a continuum rather than discrete groups. This instability,

together with the small cohort size (71 samples) and potential technical noise, likely contributed to the lack of significant survival differences between clusters (log-rank p $\approx$0.71).

Despite these limitations, the study underscores the potential value of mutational signatures in understanding AMPAC biology. By carefully cleaning and analysing public data, we showed that distinct mutational processes – MMR deficiency, HRD and exposure to tobacco carcinogens – can shape the AMPAC mutational landscape. Future work should integrate larger cohorts, whole-genome sequencing and additional omics layers (transcriptomics, methylation, immune profiling) to refine classification and validate therapeutic associations.

Code Availability

All analysis scripts, processed data, code outputs are available at: GitHub

References

(RDDC), Rare Disease Data Center (n.d.). *NBPF8 gene*. English. Database gene page. Rare Disease Data Center (RDDC). URL: https://rddc.tsinghua-gd.org/gene/728841 (visited on 09/03/2025).

Alexandrov, Ludmil B, Young Seok Ju, Kerstin Haase, Peter Van Loo, Iñigo Martincorena, Serena Nik-Zainal, Yasushi Totoki, Akihiro Fujimoto, Hidewaki Nakagawa, Tatsuhiro Shibata, Peter J Campbell, Paolo Vineis, David H Phillips, and Michael R Stratton (Nov. 4, 2016). "Mutational signatures associated with tobacco smoking in human cancer". In: *Science* 354.6312, pp. 618–622. ISSN: 0036-8075. DOI: 10.1126/science.aag0299. URL: https://doi.org/10.1126/science.aag0299 (visited on 09/06/2025).

Alexandrov, Ludmil B., Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, Kyle R. Covington, Dmitry A. Gordenin, Erik N. Bergstrom, S. M. Ashiqul Islam, Nuria Lopez-Bigas, Leszek J. Klimczak, John R. McPherson, Sandro Morganella, Radhakrishnan Sabarinathan, David A. Wheeler, Ville Mustonen, Gad Getz, Steven G. Rozen, and Michael R. Stratton (2020). "The repertoire of mutational signatures in human cancer". English. In: *Nature* 578.7793. Original research; PMCID: PMC7054213; PMID: 32025018; Author Correction: Nature 614(7948):E41 (2023), doi:10.1038/s41 022-05600-5, pp. 94–101. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-020-1943-3. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC7054213/ (visited on 09/08/2025).

Arnoff, Taylor E. and Wafik S. El-Deiry (2021). "CDKN1A/p21WAF1, RB1, ARID1A, FLG, and HRNR mutation patterns provide insights into urinary tract environmental exposure carcinogenesis and potential treatment strategies". English. In: *American Journal of Cancer Research* 11.11. Original research, pp. 5452–5471. ISSN: 2156-6976. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC8640812/ (visited on 09/03/2025).

Bafna, S., S. Kaur, and S. K. Batra (2010). "Membrane-bound mucins: the mechanistic basis for alterations in the growth and survival of cancer cells". English. In: *Oncogene* 29.20. Review, pp. 2893–2904. ISSN: 0950-9232, 1476-5594. DOI: 10.1038/onc.2010.87. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC2879972/ (visited on 09/03/2025).

Bridgers, Joshua, Kenyon Alexander, and Aly Karsan (2024). "Operationalizing Quality Assurance for Clinical Illumina Somatic Next-Generation Sequencing Pipelines". English. In: *The Journal of Molecular Diagnostics* 26.2. Article; free full text; Epub 2023-12-10, pp. 96–105. ISSN: 1525-1578, 1943-7811. DOI: 10.1016/j.jmoldx.2023.11.006. URL:

https://www.jmdjournal.org/article/S1525-1578(23)00288-X/fulltext (visited on 09/07/2025).

Center, Memorial Sloan Kettering Cancer (n.d.). *MMRd, MSI-H, and TMB-H Tumors: What They Are and Why They Matter for Cancer Immunotherapy*. English. Patient education webpage. Memorial Sloan Kettering Cancer Center. URL: https://www.mskcc.org/cancer-care/diagnosis-treatment/cancer-treatments/immunotherapy/mmrd-msi-h-and-tmb-h-tumors (visited on 09/03/2025).

Cho, William C. S., Kien Thiam Tan, Victor W. S. Ma, Jacky Y. C. Li, Roger K. C. Ngan, Wah Cheuk, Timothy T. C. Yip, Yi-Ting Yang, and Shu-Jen Chen (2018). "Targeted next-generation sequencing reveals recurrence-associated genomic alterations in early-stage non-small cell lung cancer". English. In: *Oncotarget* 9.91. Original research, pp. 36344–36357. ISSN: 1949-2553. DOI: 10.18632/oncotarget.26349. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC6284742/ (visited on 09/03/2025).

COSMIC (2023a). *SBS95 — Mutational Signatures*. Human Cancer Signatures v3.4 (October 2023); COSMIC v102. Catalogue Of Somatic Mutations In Cancer (COSMIC). URL: https://cancer.sanger.ac.uk/signatures/sbs/sbs95/ (visited on 09/08/2025).

COSMIC (2023b). *Single Base Substitution (SBS) Signatures — Mutational Signatures*. Human Cancer Signatures v3.4; page lists proposed aetiologies (e.g., SBS1: spontaneous deamination of 5-methylcytosine). Catalogue Of Somatic Mutations In Cancer (COSMIC). URL: https://cancer.sanger.ac.uk/signatures/sbs/ (visited on 09/10/2025).

COSMIC Mutational Signatures (Oct. 2023). *SBS1 — Mutational Signatures*. Database page; version v3.4 (COSMIC v102). URL: https://cancer.sanger.ac.uk/signatures/sbs/sbs1/ (visited on 09/06/2025).

Farmanbar, Amir, Robert Kneller, and Sanaz Firouzi (2023). "Mutational signatures reveal mutual exclusivity of homologous recombination and mismatch repair deficiencies in colorectal and stomach tumors". English. In: *Scientific Data* 10.1. Analysis; Open access; PMCID: PMC10314920; PMID: 37393385, p. 423. ISSN: 2052-4463. DOI: 10.1038/s41597-023-02331-8. URL: https://www.nature.com/articles/s41597-023-02331-8 (visited on 09/11/2025).

Gaujoux, Renaud and Cathal Seoighe (Sept. 11, 2024a). *cophcor: Cophenetic correlation coefficient*. R package documentation mirrored at RDrr.io; page built on 2024-09-11. URL: https://rdrr.io/cran/NMF/man/cophcor.html (visited on 09/08/2025).

Gaujoux, Renaud and Cathal Seoighe (Aug. 19, 2024b). *sparseness: Sparseness*. R package documentation on RDocumentation; NMF version 0.28. URL: https://www.rdocumentation.org/packages/NMF/versions/0.28/topics/sparseness (visited on 09/08/2025).

Gaujoux, Renaud and Cathal Seoighe (2025a). *dispersion: Dispersion of a matrix*. R package documentation; R-Forge site; version 0.17.6. URL: https://nmf.r-forge.r-project.org/dispersion.html (visited on 09/08/2025).

Gaujoux, Renaud and Cathal Seoighe (2025b). *nmfEstimateRank — Estimate rank for NMF models*. R package documentation; NMF devel site; version 0.23. URL: https://renozao.github.io/NMF/devel/nmfEstimateRank.html (visited on 09/08/2025).

Gaujoux, Renaud, Cathal Seoighe, and Nicolas Sauwen (Aug. 22, 2024). *Algorithms and framework for nonnegative matrix factorization (NMF)*. Version 0.28. R package reference manual. Comprehensive R Archive Network (CRAN). URL: https://cran.r-project.org/web/packages/NMF/NMF.pdf (visited on 09/08/2025).

Genetics, Blueprint (2020). *Blueprint Genetics' approach to pseudogenes and other duplicated genomic regions*. English. Technical explainer webpage. Blueprint Genetics. URL: https://blueprintgenetics.com/pseudogene/ (visited on 09/03/2025).

Ghareyazi, Amin, Amir Mohseni, Hamed Dashti, Amin Beheshti, Abdollah Dehzangi, Hamid R. Rabiee, and Hamid Alinejad-Rokny (2021). "Whole-Genome Analysis of De Novo Somatic Point Mutations Reveals Novel Mutational Biomarkers in Pancreatic Cancer". English. In: *Cancers* 13.17. Original research, p. 4376. ISSN: 2072-6694. DOI: 10.3390/cancers13174376. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC8431675/ (visited on 09/11/2025).

Gingras, Marie-Claude, Kyle R. Covington, David K. Chang, Lawrence A. Donehower, Anthony J. Gill, Michael M. Ittmann, Chad J. Creighton, Amber L. Johns, Eve Shinbrot, Ninad Dewal, William E. Fisher, Australian Pancreatic Cancer Genome Initiative, Christian Pilarsky, Robert Grützmann, Michael J. Overman, Nigel B. Jamieson, II Van Buren George, Jennifer Drummond, Kimberly Walker, Oliver A. Hampton, Xi Liu, Donna M. Muzny, Harsha Doddapaneni, Sandra L. Lee, Michelle Bellair, Jianhong Hu, Yi Han, Huyen H. Dinh, Mike Dahdouli, Jaswinder S. Samra, Peter Bailey, Nicola Waddell, John V. Pearson, Ivon Harliwong, Huamin Wang, Daniela Aust, Karin A. Oien, Ralph H. Hruban, Sally E. Hodges, Amy McElhany, Charupong Saengboonmee, Fraser R. Duthie, Sean M. Grimmond, Andrew V. Biankin, David A. Wheeler, and Richard A. Gibbs (2016). "Ampullary Cancers Harbor ELF3 Tumor Suppressor Gene Mutations and Exhibit Frequent WNT Dysregulation". In: *Cell Reports* 14.4. Epub 2016-01-21; PMCID: PMC4982376; PMID: 26804919, pp. 907–919. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2015.12.005. URL: https://doi.org/10.1016/j.celrep.2015.12.005 (visited on 09/11/2025).

Hong, Seung-Mo (2021). "Histologic subtyping of ampullary carcinoma for targeted therapy". English. In: *Journal of Pathology and Translational Medicine* 55.3. Editorial, p. 235. ISSN: 2383-7837, 2383-7845. DOI: 10.4132/jptm.2021.04.28. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC8141967/ (visited on 09/03/2025).

Hwang, Taejoo, Lukasz Karol Sitko, Ratih Khoirunnisa, Fernanda Navarro-Aguad, David M. Samuel, Hajoong Park, Banyoon Cheon, Luthfiyyah Mutsnaini, Jaewoong Lee, Burçak Otlu, Shunichi Takeda, Semin Lee, Dmitri Ivanov, and Anton Gartner (2025). "Comprehensive whole-genome sequencing reveals origins of mutational signatures associated

with aging, mismatch repair deficiency and temozolomide chemotherapy". In: *Nucleic Acids Research* 53.1. Article number; Advance access publication date: 5 December 2024, gkae1122. ISSN: 1362-4962. DOI: 10.1093/nar/gkae1122. URL: https://academic.oup.com/nar/article/53/1/gkae1122/7917113 (visited on 09/10/2025).

Jiang, Minlin, Keyi Jia, Lei Wang, Wei Li, Bin Chen, Yu Liu, Hao Wang, Sha Zhao, Yayi He, and Caicun Zhou (2020). "Alterations of DNA damage repair in cancer: from mechanisms to applications". English. In: *Annals of Translational Medicine* 8.24. Review, p. 1685. ISSN: 2305-5839, 2305-5847. DOI: 10.21037/atm-20-2920. URL: https://atm.amegroups.org/article/view/58380/html (visited on 09/03/2025).

Kulkarni, Maithili Mandar, Siddhi Gaurish Sinai Khandeparkar, Avinash R. Joshi, Aniket Kakade, Lokesh Fegade, and Ketan Narkhede (2017). "Clinicopathological Study of Carcinoma of the Ampulla of Vater with Special Reference to MUC1, MUC2 and MUC5AC Expression". English. In: *Journal of Clinical and Diagnostic Research* 11.5. Original research, EC17–EC20. ISSN: 0973-709X, 2249-782X. DOI: 10.7860/JCDR/2017/26842.9830. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC5483668/ (visited on 09/03/2025).

Medo, Matúš, Charlotte K. Y. Ng, and Michaela Medová (Nov. 2, 2024). "A comprehensive comparison of tools for fitting mutational signatures". In: *Nature Communications* 15.1, p. 9467. ISSN: 2041-1723. DOI: 10.1038/s41467-024-53711-6. URL: https://doi.org/10.1038/s41467-024-53711-6 (visited on 09/07/2025).

Olivier, Magali, Monica Hollstein, and Pierre Hainaut (2010). "TP53 mutations in human cancers: origins, consequences, and clinical use". English. In: *Cold Spring Harbor Perspectives in Biology* 2.1. Review, a001008. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a001008. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC2827900/ (visited on 09/03/2025).

Ramai, Daryl, Andrew Ofosu, Jameel Singh, Febin John, Madhavi Reddy, and Douglas G. Adler (2019). "Demographics, tumor characteristics, treatment, and clinical outcomes of patients with ampullary cancer: a Surveillance, Epidemiology, and End Results (SEER) cohort study". In: *Minerva Gastroenterologica e Dietologica* 65.2. Epub 2018-11-27, pp. 85–90. ISSN: 1121-421X. DOI: 10.23736/S1121-421X.18.02543-6. URL: https://www.minervamedica.it/en/journals/gastroenterology/article.php?cod=R08Y2019N02A0085 (visited on 09/11/2025).

Rizzo, Alessandro, Vincenzo Dadduzio, Lucia Lombardi, Angela Dalia Ricci, and Gennaro Gadaleta-Caldarola (2021). "Ampullary Carcinoma: An Overview of a Rare Entity and Discussion of Current and Future Therapeutic Challenges". English. In: *Current Oncology* 28.5. Review, pp. 3393–3402. ISSN: 1718-7729. DOI: 10.3390/curroncol28050293. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC8482111/ (visited on 09/03/2025).

Sinha, Sonali, Victor Ng, Ardijana Novaj, Yingjie Zhu, Shu Yazaki, Xin Pei, Fatemeh Derakhshan, Fresia Pareja, Jeremy Setton, Flavie Naulin, Manuel Beltrán-Visiedo, Ethan

Shin, Ana Leda F. Longhini, Rui Gardner, Jennifer Ma, Kevin Ma, Anne Roulston, Stephen Morris, Maria Koehler, Simon Powell, Ezra Rosen, Lorenzo Galluzzi, Jorge Reis-Filho, Atif Khan, and Nadeem Riaz (2025). "The cold immunological landscape of ATM-deficient cancers". English. In: *Journal for ImmunoTherapy of Cancer* 13.5. Original research; PMCID: PMC12067784; PMID: 40350205, e010548. ISSN: 2051-1426. DOI: 10.1136/jitc-2024-010548. URL: https://jitc.bmj.com/content/13/5/e010548 (visited on 09/03/2025).

Son, Hyeonju, Hyundeok Kang, Hyun Seok Kim, and Sangwoo Kim (Oct. 27, 2017). "Somatic mutation driven codon transition bias in human cancer". In: *Scientific Reports* 7.1, p. 14204. ISSN: 2045-2322. DOI: 10.1038/s41598-017-14543-1. URL: https://doi.org/10.1038/s41598-017-14543-1 (visited on 09/06/2025).

Tomlinson, Jennifer L., Binbin Li, Jingchun Yang, Emilien Loeuillard, Hannah E. Stumpf, Hendrien Kuipers, Ryan Watkins, Danielle M. Carlson, Jessica Willhite, Daniel R. O'Brien, Rondell P. Graham, Xin Chen, Rory L. Smoot, Haidong Dong, Gregory J. Gores, and Sumera I. Ilyas (2024). "Syngeneic murine models with distinct immune microenvironments represent subsets of human intrahepatic cholangiocarcinoma". In: *Journal of Hepatology* 80.6. Epub 2024-03-07; PMCID: PMC11141161, pp. 892–903. ISSN: 0168-8278. DOI: 10.1016/j.jhep.2024.02.008. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11141161/ (visited on 09/11/2025).

Tsagkalidis, Vasileios, Russell C. Langan, and Brett L. Ecker (2023). "Ampullary Adenocarcinoma: A Review of the Mutational Landscape and Implications for Treatment". In: *Cancers* 15.24. Article number; PMCID: PMC10741460, p. 5772. ISSN: 2072-6694. DOI: 10.3390/cancers15245772. URL: https://www.mdpi.com/2072-6694/15/24/5772 (visited on 09/11/2025).

Tsang, Erica S., Veronika Csizmok, Laura M. Williamson, Erin Pleasance, James T. Topham, Joanna M. Karasinska, Emma Titmuss, Intan Schrader, Stephen Yip, Basile Tessier-Cloutier, Karen Mungall, Tony Ng, Sophie Sun, Howard J. Lim, Jonathan M. Loree, Janessa Laskin, Marco A. Marra, Steven J. M. Jones, David F. Schaeffer, and Daniel J. Renouf (2023). "Homologous recombination deficiency signatures in gastrointestinal and thoracic cancers correlate with platinum therapy duration". English. In: *npj Precision Oncology* 7. Original research; Open access, p. 31. ISSN: 2397-768X. DOI: 10.1038/s41698-023-00368-x. URL: https://www.nature.com/articles/s41698-023-00368-x (visited on 09/11/2025).

Zemet, Roni, Haowei Du, Tomasz Gambin, James R. Lupski, Pengfei Liu, and Paweł Stankiewicz (2023). "SNV/indel hypermutator phenotype in biallelic RAD51C variant: Fanconi anemia". In: *Human Genetics* 142.6. Epub 2023-04-09, pp. 721–733. ISSN: 1432-1203. DOI: 10.1007/s00439-023-02550-4. URL: https://link.springer.com/article/10.1007/s00439-023-02550-4 (visited on 09/11/2025).

Zhu, Mingyu, Lu Zhang, Haiyan Cui, Qiang Zhao, Hao Wang, Baochao Zhai, Richeng Jiang, and Zhansheng Jiang (2022). "Co-Mutation of FAT3 and LRP1B in Lung Adenocarcinoma Defines a Unique Subset Correlated With the Efficacy of Immunotherapy". English. In: *Frontiers in Immunology* 12. Original research; eCollection 2021, p. 800951. ISSN: 1664-3224. DOI: 10.3389/fimmu.2021.800951. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC8770854/ (visited on 09/03/2025).

Zhuravleva, Ekaterina, Monika Lewinska, Colm J. O'Rourke, Antonio Pea, Asif Rashid, Ann W. Hsing, Andrzej Taranta, David Chang, Yu-Tang Gao, Jill Koshiol, Rui Caetano Oliveira, and Jesper B. Andersen (Apr. 7, 2025). "Mutational signatures define immune and Wnt-associated subtypes of ampullary carcinoma". In: *Gut* 74.5, pp. 804–814. ISSN: 0017-5749. DOI: 10.1136/gutjnl-2024-333368. URL: https://doi.org/10.1136/gutjnl-2024-333368 (visited on 09/07/2025).

Zitnik, Marinka (2025). *Nmf (models.nmf) — Nimfa documentation*. Library documentation; Nimfa v1.3.4. URL: https://ai.stanford.edu/~marinka/nimfa/nimfa.models.nmf.html (visited on 09/08/2025).

Appendix

Signature	Rank	COSMIC (cosine)	Cosine similarity	COSMIC (peak-sensitive)	Peak-sensitive similarity
Sig1	1	SBS95	0.814	SBS95	0.818
Sig1	2	SBS4	0.802	SBS4	0.806
Sig1	3	SBS94	0.759	SBS94	0.797
Sig2	1	SBS1	0.871	SBS1	0.871
Sig2	2	SBS6	0.860	SBS6	0.861
Sig2	3	SBS15	0.776	SBS15	0.783

Table 4. Comparison of $k = 2$ signatures with COSMIC: cosine and peak-sensitive cosine metrics.

Signature	Rank	COSMIC (cosine)	Cosine similarity	COSMIC (peak-sensitive)	Peak-sensitive similarity
Sig1	1	SBS1	0.878	SBS1	0.878
Sig1	2	SBS6	0.873	SBS6	0.873
Sig1	3	SBS15	0.776	SBS15	0.783
Sig2	1	SBS4	0.833	SBS95	0.847
Sig2	2	SBS29	0.827	SBS4	0.846
Sig2	3	SBS95	0.822	SBS29	0.833
Sig3	1	SBS3	0.711	SBS3	0.774
Sig3	2	SBS40a	0.633	SBS40a	0.669
Sig3	3	SBS40b	0.592	SBS40b	0.645

Table 5. Comparison of $k = 3$ signatures with COSMIC: cosine and peak-sensitive cosine metrics.

Signature	Rank	COSMIC (cosine)	Cosine similarity	COSMIC (peak-sensitive)	Peak-sensitive similarity
Sig1	1	SBS1	0.900	SBS1	0.900
Sig1	2	SBS6	0.887	SBS6	0.890
Sig1	3	SBS15	0.790	SBS15	0.802
Sig2	1	SBS4	0.778	SBS4	0.810
Sig2	2	SBS95	0.744	SBS94	0.788
Sig2	3	SBS29	0.741	SBS95	0.773
Sig3	1	SBS24	0.617	SBS40a	0.701
Sig3	2	SBS40a	0.604	SBS3	0.684
Sig3	3	SBS3	0.598	SBS24	0.657
Sig4	1	SBS3	0.587	SBS3	0.680
Sig4	2	SBS17b	0.532	SBS9	0.627
Sig4	3	SBS9	0.519	SBS40c	0.590

Table 6. Comparison of $k = 4$ signatures with COSMIC: cosine and peak-sensitive cosine metrics.

Signature	Rank	COSMIC (cosine)	Cosine similarity	COSMIC (peak-sensitive)	Peak-sensitive similarity
Sig1	1	SBS1	0.917	SBS1	0.917
Sig1	2	SBS6	0.88	SBS6	0.884
Sig1	3	SBS15	0.774	SBS15	0.782
Sig2	1	SBS4	0.829	SBS4	0.868
Sig2	2	SBS95	0.765	SBS94	0.796
Sig2	3	SBS29	0.746	SBS95	0.793
Sig3	1	SBS5	0.569	SBS3	0.660
Sig3	2	SBS40a	0.564	SBS40a	0.647
Sig3	3	SBS3	0.552	SBS5	0.631
Sig4	1	SBS40a	0.555	SBS40a	0.651
Sig4	2	SBS15	0.514	SBS25	0.635
Sig4	3	SBS39	0.5	SBS6	0.612
Sig5	1	SBS17b	0.572	SBS3	0.645
Sig5	2	SBS3	0.539	SBS9	0.632
Sig5	3	SBS9	0.517	SBS17b	0.577

Table 7. Comparison of $k = 5$ signatures with COSMIC: cosine and peak-sensitive cosine metrics.

Signature	Rank	COSMIC (cosine)	Cosine similarity	COSMIC (peak-sensitive)	Peak-sensitive similarity
Sig1	1	SBS1	0.903	SBS1	0.905
Sig1	2	SBS6	0.875	SBS6	0.899
Sig1	3	SBS15	0.776	SBS15	0.791
Sig2	1	SBS4	0.81	SBS4	0.853
Sig2	2	SBS95	0.729	SBS94	0.818
Sig2	3	SBS29	0.714	SBS95	0.759
Sig3	1	SBS40a	0.537	SBS40a	0.615
Sig3	2	SBS29	0.482	SBS6	0.578
Sig3	3	SBS1	0.481	SBS3	0.571
Sig4	1	SBS24	0.637	SBS24	0.725
Sig4	2	SBS95	0.562	SBS29	0.644
Sig4	3	SBS29	0.559	SBS94	0.641
Sig5	1	SBS2	0.66	SBS2	0.755
Sig5	2	SBS7a	0.641	SBS40a	0.729
Sig5	3	SBS40a	0.581	SBS7a	0.688
Sig6	1	SBS17b	0.637	SBS9	0.676
Sig6	2	SBS9	0.553	SBS17b	0.641
Sig6	3	SBS3	0.514	SBS3	0.611

Table 8. Comparison of $k = 6$ signatures with COSMIC: cosine and peak-sensitive cosine metrics.

5 Discussion about differences in clustering/Figure 3A-style heatmap

I suppose, the first reason for the differences between my Fig 12 and Figure 3A by Zhuravleva et al. (2025) is the composition of the data. In the work by Zhuravleva et al. (2025), 103 cases of ampullary carcinoma were used for clustering by hierarchical classification. I worked with a table from the supplementary material, restored from Excel, where some values had to be manually reverted (because of automatic conversion of some values to the date type),

Signature	Rank	COSMIC (cosine)	Cosine similarity	COSMIC (peak-sensitive)	Peak-sensitive similarity
Sig1	1	SBS1	0.899	SBS1	0.900
Sig1	2	SBS6	0.850	SBS6	0.856
Sig1	3	SBS15	0.768	SBS15	0.781
Sig2	1	SBS4	0.782	SBS4	0.842
Sig2	2	SBS95	0.709	SBS94	0.793
Sig2	3	SBS94	0.697	SBS95	0.742
Sig3	1	SBS42	0.583	SBS6	0.631
Sig3	2	SBS6	0.576	SBS5	0.624
Sig3	3	SBS5	0.527	SBS42	0.623
Sig4	1	SBS39	0.571	SBS3	0.686
Sig4	2	SBS3	0.570	SBS40a	0.646
Sig4	3	SBS40a	0.569	SBS25	0.640
Sig5	1	SBS95	0.504	SBS40a	0.573
Sig5	2	SBS14	0.491	SBS25	0.563
Sig5	3	SBS29	0.487	SBS40c	0.561
Sig6	1	SBS24	0.565	SBS24	0.690
Sig6	2	SBS94	0.563	SBS94	0.677
Sig6	3	SBS53	0.556	SBS24	0.644
Sig7	1	SBS17b	0.664	SBS17b	0.669
Sig7	2	SBS9	0.538	SBS9	0.669
Sig7	3	SBS3	0.485	SBS3	0.582

Table 9. Comparison of $k = 7$ signatures with COSMIC: cosine and peak-sensitive cosine metrics.

and I filtered out samples with fewer than 15 SNVs (according to the criterion from the article's supplement). Any errors in cell conversion (for example, automatic substitution of dates or loss of leading zeros in coordinates) could change the triplet context and thereby distort the profile. In addition, about 70 samples remained in my dataset after filtering, whereas in the paper the heatmap was built on a more complete cohort; this inevitably affects channel frequencies and the structure of the dendrogram. There are also likely other reasons for preprocessing and filtering that are implied but not mentioned in the methods of the article.

The second important point is normalization. In the supplement to the article it is described that the frequencies were additionally divided by the occurrence of the corresponding triplets in the hg19 genome, that is, "opportunity" normalization was applied. In my code I tested two options: (1) converting each SBS-96 catalog into fractions (sum per sample = 1) and (2) opportunity normalization. Practice showed that the second approach greatly increases noise – the rare contexts in exome data yield large fractions, the spectrum shifts, and the cosine similarity to a COSMIC signature decreases. Therefore I used the first option for signature decomposition and comparison with COSMIC. In the article, by contrast, opportunity normalization was used at the stage of building the profiles, therefore the relative weights of triplets in the authors' work and in mine differ. When comparing with COSMIC it makes sense to analyze raw counts, since for samples with a low number of SNVs raw matrices do not "inflate" rare contexts and make it possible to understand whether the sig-

nature is really present. Normalized fractions, on the contrary, are important for an honest cosine comparison, because the reference COSMIC signatures are given as distributions. In this work I initially chose normalized profiles for the decomposition and the heatmap, since they were fairly stable, introduced variability subsequently to clustering (for the dominant signatures), and also in order to maintain compatibility with the methodology of the article. There are also possible shortcomings in the alignment of the normalized matrix with SomaticSignatures to SBS-96 format.

For signature extraction, the paper used the SomaticSignatures package with 1000 NMF repetitions; the number of signatures ($k = 3$) was chosen based on aggregate statistics, and samples with <15 mutations were excluded. I used the sigminer package and scanned $k = 2...7$ with 500 NMF runs, then selected the optimal k by the cophenetic correlation coefficient, RSS, and similarity to COSMIC. At the final stage for $k = 3$ I ran 1000 iterations. Different NMF implementations and differences in the choice of initial matrices yield slightly different profiles: in my work the third signature resembled SBS3 (homologous recombination), whereas in the authors' work the third signature corresponded to SBS9 (polymerase η). Differences in the profiles themselves lead to different exposures, which in turn result in differences in the clusters.

The key difference is clustering. The paper states that after decomposing the profiles they applied hierarchical clustering and identified three clusters, C1–C3; details of the method (distance, linkage, selection of contexts) are not provided. In Figure 3A it is evident that the scale is centered around zero and symmetric; therefore, the authors centered and scaled the series. However, in addition to z-scaling they may have used additional steps that are not described—for example, retaining only the most variable contexts, trimming extreme values, tuning the palette, or manually fixing $k = 3$.In my code I construct the heatmap as follows: the normalized profile is converted to Z-scores by rows; a distance of $1-$Pearson is computed between columns; the complete linkage method is used; then the number of clusters is determined automatically with DynamicTreeCut or by the maximum silhouette. In my case, clustering based on normalized and z-standardized profiles performed poorly: the structure turned out to be unstable, with low silhouette values and weak correspondence to the expected biological subtypes. This contrasts with Figure 3A in the paper, where the clusters appear clearly delineated.

Thus, differences in the source data and their restoration, a different normalization choice, different NMF tools and parameters, and a different clustering algorithm likely explain why my Fig 12 mainly differs from Figure 3A in the work by Zhuravleva et al. (2025).