Nucleic Acids Research, 2025, **53**, gkaf648
https://doi.org/10.1093/nar/gkaf648
Methods

Comprehensive analysis of mutational processes across 20 000 adult and pediatric tumors

Matteo Villa[1,†], Federica Malighetti[1,†], Luca De Sano[2], Alberto Maria Villa[1], Nicoletta Cordani[1], Andrea Aroldi[1,3], Marco Antoniotti[2,4,5], Giulio Caravagna[6], Alex Graudenzi[2,4,5], Rocco Piazza[1,3], Luca Mologni[1], Daniele Ramazzotti[1,3,4,*]

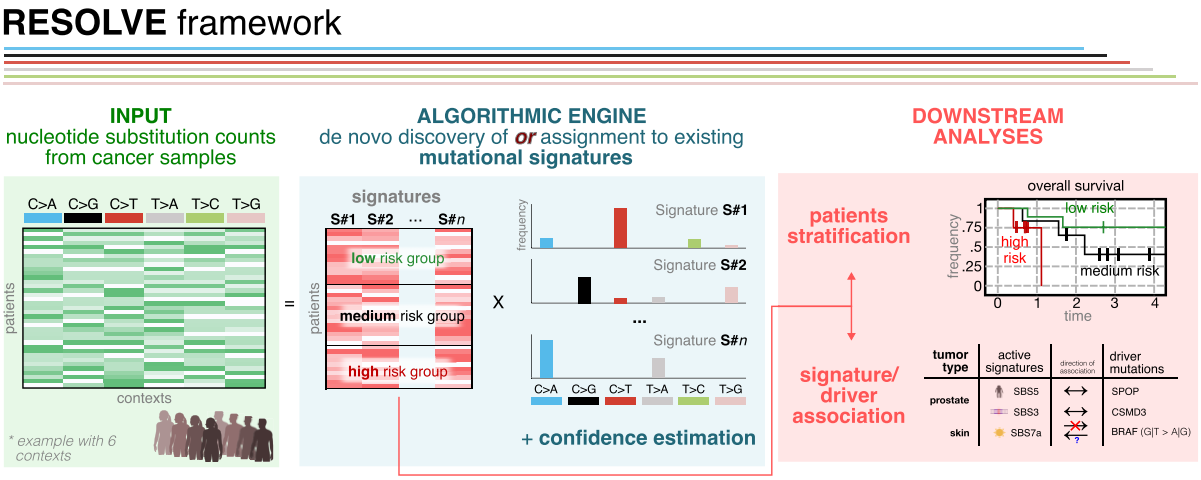
[1]Department of Medicine and Surgery, University of Milano-Bicocca, Monza 20900, Italy
[2]Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan 20126, Italy
[3]Fondazione IRCCS San Gerardo dei Tintori, Monza 20900, Italy
[4]Bicocca Bioinformatics, Biostatistics and Bioimaging Centre (B4), Milan 20854, Italy
[5]Institute of Molecular Bioimaging and Physiology, Consiglio Nazionale delle Ricerche (IBFM-CNR), Segrate, Milan 20090, Italy
[6]Department of Mathematics and Geosciences, University of Trieste, Trieste 34128, Italy

[*]To whom correspondence should be addressed. Email: daniele.ramazzotti@unimib.it
[†]The first two authors should be regarded as Joint First Authors.

Abstract

Despite being a consolidated practice in modern cancer genomics, mutational signature analysis poses various challenges. First, determining the number of signatures is a complex task and depends on heuristics. Second, several signatures lack a clear etiology, raising concerns about whether they result from computational artifacts or actual mutagenic processes. Last, approaches for signature assignment are highly influenced by the catalogue of signatures used for the analysis. To overcome these limitations, we introduce RESOLVE (Robust ESTimation Of mutational signatures Via rEgularization), a framework designed for the efficient extraction, assignment, and confidence estimation of mutational signatures. RESOLVE enables the stratification of cancer genomes according to the active mutational processes, providing insights into those that are consistently shared among various cancer types. We applied RESOLVE to 20 000 samples from adult and pediatric cancers, providing a comprehensive characterization of subtypes associated with specific mutational processes and alterations in driver genes. Our analysis demonstrates that RESOLVE accurately fits observed mutations with a smaller set of signatures compared to existing catalogues, suggesting the existence of a limited number of dominant mutational processes in cancer genomes. Clustering analysis revealed distinct patient groups characterized by specific signatures, with significant associations between certain signatures and patient prognosis. Additionally, we identified strong associations between signatures and driver gene mutations, offering insights into cancer subtype mechanisms and evolution. Our findings highlight the efficiency of RESOLVE and its potential impact on personalized cancer treatment.

Graphical abstract



Received: October 23, 2024. **Revised:** May 31, 2025. **Editorial Decision:** June 13, 2025. **Accepted:** June 19, 2025

Introduction

Cancer is a genetic disease caused by somatic mutations in genes controlling key biological functions such as cellular growth and division [1]. Such mutations may arise both through cell-intrinsic and exogenous processes, generating characteristic mutational patterns over the genome named *mutational signatures* [2]. The study of mutational signatures has become a standard practice of modern cancer genomics, since it can reveal which (environmental and endogenous) mutagenic processes are active in a tumor, and may highlight markers for therapeutic response [3–5].

Mutational signatures can be modeled considering different classes of mutations. The original formulation examined single-base substitutions (SBSs) by incorporating the two flanking bases for each mutation for a total of 96 contexts representing mutational patterns [6, 7]. Such formulation is currently the most widely diffused. Other types of mutational signatures include double-base substitutions (DBSs) classified into 78 contexts [8, 9], indel and rearrangement signatures (IDs) defining 83 categories [8], chromosomal instability signatures (CXs) comprising 43 mixture components [10], and 48-channels copy number signatures (CNs) [11]. These types of mutational signatures represent a newer field of research and currently are not as broadly adopted [2].

Mutational signature computational analyses fall mostly within two categories: (i) *de novo signature extraction* and (ii) *signature exposure estimation*. In the first case, the presence of mutational processes is first assessed from the data, and signatures are identified and extracted and finally assigned to samples [12]. This task is typically performed via non-negative matrix factorization (NMF) [13]. While other approaches have been proposed, NMF-based methods are by far the most used [14]. The estimation of signature exposures is performed by holding a set of signatures fixed (see e.g. the COSMIC mutational signature catalogue [8]) and assigning them to samples by minimizing, for example, the mean squared error between observed and the estimated mutational patterns for each sample. It is crucial to choose a good set of signatures to be used for fitting, since as discussed by Nik-Zainal and colleagues [2], algorithms for estimating exposures can be highly affected by overfitting, which might lead to the misinterpretation of the active mutational processes really present in the tumors [2].

With respect to the *de novo* signature extraction, determining the optimal number of signatures present in a dataset is a complex task, largely dependent on the adopted computational framework, which is typically solved by heuristics or model selection techniques [13, 15–17]. This poses the challenge of determining a trade-off between robustness, usually optimized when considering a low number of signatures, and expressivity, typically optimized when considering a high number of signatures, of the extracted mutational signatures [2, 8]. Thus, it is currently hard to efficiently define how many signatures are actually contributing to the real etiology of a given set of tumors; i.e. how many of them are biologically relevant.

Furthermore, while many signatures have been strongly associated with mutagenic processes [18–20], several other reported signatures present instead a more elusive etiology or are highly similar between each other [8], suggesting possible limitations in the signature extraction framework or overfitting, rather than the existence of associated mutagenic processes [2, 17]. This scenario is also complicated by the fact that mutational processes may cause multiple signatures and, likewise, different processes may generate similar signatures [2].

Lastly, also the methods for signature exposure estimations suffer from several pitfalls, first and foremost the hurdle of choosing the most appropriate catalogue of signatures to be used for the analysis [2]. Most methods typically rely on optimization strategies aimed to fit most of the data, including the patterns that may not be significantly present or biologically relevant for the considered tumors or tissues. In this context, overfitting can lead to the wrong estimation of mutational processes active in the samples. This may also preclude the effective stratification of cancer samples based on signatures, whose characterization could provide valuable information on mechanisms possibly correlated with different prognostic outcomes. In addition, measures of statistical confidence or strength of signature activities in the samples are usually not provided.

To overcome these limitations, we developed a novel computational framework named RESOLVE (Robust ESTimation Of mutationaL signatures Via rEgularization), which allows us to perform (i) the efficient extraction, (ii) the exposure estimation, and (iii) the confidence assessment of mutational signatures from cancer genomes. RESOLVE optimizes the use of signatures toward the best description of mutational processes with the minimal number of signatures extracted. Additionally, RESOLVE enables the effective stratification of cancer genomes based on mutational signatures, which allows us to determine which signatures are significantly contributing to a cancer subtype and may shed light on associations of the discovered mutational processes with distinct mechanisms of tumorigenesis and significant differences in outcomes.

The aim of RESOLVE is to focus on the dominant mutational processes with a statistically significant impact on mutagenesis across patients, offering a balance between biological relevance and computational robustness. Although rare mutational signatures may be present and could offer insights into specific mutational mechanisms or patient subgroups, their computational extraction is often subject to overfitting. Therefore, RESOLVE seeks to address this challenge by focusing on signatures that can be confidently and robustly detected, and by providing a framework that explicitly allows for confidence estimates of the extracted mutational processes.

We compared the performance of RESOLVE with state-of-the-art methods for mutational signature analysis across a wide range of simulations, demonstrating that RESOLVE consistently outperforms competing algorithms in terms of accuracy and robustness. Moreover, we applied RESOLVE to nearly 20 000 whole-genome sequencing samples from adult and pediatric cancers, encompassing most tissues and cancer types [21, 22]. Our analysis revealed a small set of mutational signatures that accurately represent almost all the observed mutations, indicating that a limited number of dominant mutational processes drive cancer development. We identified significant associations between specific signatures and patient prognosis, such as SBS9 correlating with better outcomes in hematopoietic cancers and SBS3 with poorer prognosis. Furthermore, we discovered strong links between mutational signatures and driver gene mutations, exemplified by the association between SBS2 and PIK3CA mutations, and SBS3 and TP53 mutations in breast cancer. These findings highlight the potential of RESOLVE to enhance cancer diagnostics and

personalized therapy design, providing profound insights into the molecular foundations of cancer.

The RESOLVE framework is available as an R package and can be installed from Bioconductor at https://www.bioconductor.org/packages/release/bioc/html/RESOLVE.html or from GitHub at https://github.com/ramazzottilab/RESOLVE.

Materials and methods

Mathematical framework for mutational signature analysis

Let us consider a set of N mutation classes and a set of G cancer genomes, i.e. the tumor samples. In the case of SBSs [6, 7], single nucleotide variants expanded by the nucleotide context are examined leading to $N = 96$ trinucleotides. Instead, e.g. in the case of DBSs [8, 9], the concurrent modification of two consecutive nucleotide bases are considered for a total of $N = 78$ mutation categories. This can be represented in a count matrix M, where in each cell of the matrix we report the number of mutations that have been observed in each cancer sample for each mutation class, with the minimum value being 0:

$$M_{G\times N} = \begin{bmatrix} m_{1,1} & m_{1,2} & \dots & m_{1,N} \\ m_{2,1} & m_{2,2} & \dots & m_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ m_{G,1} & m_{G,2} & \dots & m_{G,N} \end{bmatrix}.$$

Regardless of the considered classes of mutations, mutational signatures and their exposures to samples can be represented by two matrices E and S:

$$E_{G\times K} = \begin{bmatrix} e_{1,1} & e_{1,2} & \dots & e_{1,K} \\ e_{2,1} & e_{2,2} & \dots & e_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ e_{G,1} & e_{G,2} & \dots & e_{G,K} \end{bmatrix},$$

$$S_{K\times N} = \begin{bmatrix} s_{1,1} & s_{1,2} & \dots & s_{1,N} \\ s_{2,1} & s_{2,2} & \dots & s_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ s_{K,1} & s_{K,2} & \dots & s_{K,N} \end{bmatrix}.$$

Above, we indicate with K the number of mutational signatures to be considered, where typically $K < N$. The matrix S is the signature matrix for the K signatures and each row of the matrix is a categorical distribution (summing up to 1) over the N mutation categories, representing the pattern of mutations for each signature. The matrix E is the exposure matrix and represents the activity of the K signatures for each sample in terms of number of mutations; i.e. each cell of E is a positive number, with the minimum value being 0 [13].

Therefore, a simple framework for mutational signature analysis can be generalized as the decomposition of $M \approx E \cdot S$.

De novo signature extraction by RESOLVE

The *de novo* signature extraction problem can be solved by NMF [13] by minimizing goodness-of-fit metrics such as the Frobenius reconstruction error (other metrics, such as the cosine similarity, are possible targets for the optimization problem), while constraining all elements to be non-negative.

$$\min_{E\geq 0, S\geq 0} \|M - E \cdot S\|_F^2. \quad (1)$$

In RESOLVE, we extended such framework following the ideas discussed by Lal and colleagues [17] by incorporating a background signature during the inference and using LASSO regularization [23] to reduce the impact of overfitting.

To this end, we reframed the inference problem as $M \approx E_0 \cdot S_0 + E \cdot S$, where E_0 is the exposure vector of the background signature and S_0 is the background signature. This new problem can be solved by considering the following optimization task:

$$\min_{E_0\geq 0, S_0\geq 0, E\geq 0, S\geq 0} \|M - (E_0 \cdot S_0 + E \cdot S)\|_F^2 + \lambda_0 \cdot \|E_0\|_1 + \lambda_1 \cdot \|E\|_1 + \lambda_2 \cdot \|S\|_1. \quad (2)$$

For SBSs, S_0 can be derived from mutational signature SBS5 present in the COSMIC catalogue [8, 17]. This signature has been reported to be active in all cancer types and in normal somatic tissues [24, 25] and can be considered as a natural background process [26]. Solving Equation (2) consists intuitively in approximating the matrix of the observed counts M, while fixing S_0 and keeping E_0, E, and S sparse. In particular, λ_0, λ_1, and λ_2 are parameter vectors for the LASSO regularization and represent the optimal degree of sparsification for each regression problem.

The *de novo* signature extraction problem (Equation 2) can be solved by RESOLVE with the following steps:

1. *Input data*: Build the count matrix M by counting the number of mutations for each of the considered category (e.g. SBSs) in each sample. In this step, we suggest to consider when possible only high-quality whole-genome sequencing data, with a given minimum number of mutations, e.g. 1000 [17]. The number of signatures K (positive integer) is an other input for this learning task. We refer to the next paragraph for a discussion on its estimation.
2. *Initialization*: For a given number of signature K, set S_0 to the chosen background signature (see previous discussion) and randomly initialize S with a negative binomial distribution per signature (rows of the matrix). Then, learn initial values of E_0 and E by fixing M, S_0, and S and solving Equation (2). We notice that during each learning step, E_0, E, and S are constrained to be non-negative and each row of S is normalized to sum up to 1 as it represents a categorical distribution.
3. *Learning*: Repeat a learning procedure until convergence, where alternatively M, S_0, and S are held fixed to learn E_0 and E, and M, E_0, E, and S_0 are subsequently held fixed to learn S. Equation (2) is solved by estimating λ_0, λ_1, and λ_2 at each iteration by standard cross-validation, which is applied to each regression task.
4. *Outputs*: As NMF-based estimators are known to be greatly influenced by the initialization step [27], we perform a robust estimation of E_0, E, and S by repeating (e.g. 50–100 times) steps 2 and 3. The best among the inferred models is finally returned.

Optimal number of signatures for *de novo* signature extraction

The estimation of the optimal number K of signatures to be considered as input for the *de novo* signature extraction optimization problem is a complex task. Here, we adopt a strategy to estimate K based on bi-cross-validation [17, 28, 29], implemented as follows. We consider a range of (positive integer) values as candidate number of signatures, e.g. $K = [1, 2, ..., 15]$. For each value, we randomly select a percentage of cells from M (e.g. 1% [17]) and we replace these entries with 0 to generate M'. We then perform the inference by the RESOLVE algorithm for *de novo* signature extraction on M' as described above, and we compute the cosine similarity for the cells that were set to 0 by comparing for these entries the original values in M with the ones predicted by RESOLVE after the inference.

We repeat this procedure multiple times (e.g. at least 50–100 times [17]) and we select as the optimal value of K the one leading to the minimum cross-validation mean squared error estimate over the multiple runs. Intuitively, this aims to choose the value of K that generates a model with good generalization capabilities.

Signature exposure estimation by RESOLVE

In this case, the catalogue of signatures (i.e. S_0 and S) is provided as an input of the optimization task. Therefore, when solving Equation (2), the only unknowns are E_0, E, λ_0, and λ_1. We solve this optimization problem by directly estimating such quantities by regularized regression with the *LASSO* penalty with the same standard cross-validation scheme adopted for the *de novo* signature extraction procedure.

Signature confidence estimation by RESOLVE

RESOLVE implements a procedure based on bootstrap [30] to perform confidence estimation of the activity of signatures in the different samples. To this end, we treat the rows of M for each sample as empirical categorical distributions over the considered mutational categories from which we can sample, with resampling, the same number of mutations observed in M to obtain a set of bootstrapped datasets $M_1, ..., M_B$. We recommend to perform a number of bootstrap resampling B of at least 50–100 iterations.

Given previously estimated S_0 and S, we then perform the signature exposure estimation procedure by RESOLVE as described before, for $M_1, ..., M_B$ to obtain a distribution of B estimated values for E_0 and E. RESOLVE also implements additional quality checks on signature activity by setting (optional) thresholds based on cosine similarity (i.e. consider only the top signatures until reaching a given level of cosine similarity between observed and predicted counts, e.g. 0.95) and minimum number of mutations (e.g. at least 5% of the mutations in a given sample) for a certain signature to be considered in the analysis.

Given these B estimates of E_0 and E, RESOLVE can finally compute P-values measuring the confidence of signature activities for each sample by the Wilcoxon signed-rank test. Entries of E_0 and E with significant P-values (e.g. $<.05$) represent signatures that are active with high confidence in the data. Moreover, a confidence estimate of the global activities of each signature in the considered dataset can be computed as the harmonic mean P-value [31] of the results by the Wilcoxon signed-rank test.

Cosine similarity as evaluation metric to estimate goodness of fit

Cosine similarity was employed throughout this study as the primary evaluation metric to assess the goodness of fit between observed mutational profiles and reconstructed profiles based on the different signature catalogues, including the one from RESOLVE. This metric quantifies the alignment between two non-negative vectors, where a value of 1 indicates perfect alignment and a value of 0 indicates no similarity. For our analysis, cosine similarity was computed between the observed mutation count vector for each sample and the reconstructed vector derived from the weighted combination of mutational signatures assigned to that sample. A high cosine similarity reflects the ability of the signature set to accurately capture the underlying mutational processes in a given sample.

Clustering analysis based on mutational signatures by RESOLVE

RESOLVE can perform clustering based on mutational signatures by considering the estimated E_0 and E. To do so, we first compute a samples $\times$ samples similarity matrix from the exposure matrices as the cosine similarity of E_0 and E. We then solve the clustering problem considering the computed similarity matrix by a clustering algorithm such as k-medoids [32]. The optimal number of clusters can be estimated using a standard approach by considering different numbers of clusters and evaluating the results with metrics such as the silhouette score [33].

Survival analysis based on mutational signatures by RESOLVE

We conducted regularized Cox regression with the LASSO penalty, using overall survival data as response variables and the normalized exposures inferred by RESOLVE as predictors. Hierarchical clustering analysis, based on the risk scores determined by regularized Cox regression, was performed using a hybrid adaptive tree cut to determine the number of clusters [34]. Standard Kaplan–Meier analysis was conducted and log-rank P-values were computed. The presence of statistically different frequencies of gene mutations in the different clusters was assessed using a proportions z-test. P-values were corrected for false discoveries, and a significance threshold of .05 was considered.

Association of the mutational signatures by RESOLVE with genes mutations

We employed an approach utilizing regularized regression techniques to verify the presence of associations between mutational signatures and specific mutations across various cancer types and biological tissues. First, we treated mutation data as response variables, aiming to predict them with the normalized signature exposures. This approach allowed us to quantify the impact of each signature on the probability of observing a specific mutation. Given the binary nature of the response variables (mutation present/absent), we exploited regularized logistic regression for this analysis. Regularized logistic regression is a powerful technique for analyzing data with binary outcomes, incorporating penalty terms to handle multicollinearity and prevent overfitting. We optimized the regularization parameters using cross-validation techniques to ensure robustness. Given the logistic model, we can compute the

expected probability of a mutation given the signature exposures. This allows us to compute a fold change with respect to the intercept, providing insight into how each signature impacts the likelihood of observing a mutation. Such fold change represents the multiplicative effect of a one-unit change in the signature exposure on the odds of observing a mutation, relative to the baseline (intercept).

In the second approach, we considered the normalized exposures generated by RESOLVE as the response variable, with mutations serving as predictors. Here, the objective was to quantify how much a given mutation influenced the activity of various mutational processes. This analysis utilized a general regularized regression model with a Gaussian response. The regularization parameters were optimized using cross-validation to ensure the reliability and stability of our results across different datasets and modeling assumptions. Also in this case, we computed a fold change between the coefficients associated with each mutation and the intercept to quantify the extent of the associations.

Results

The RESOLVE framework

The RESOLVE framework is designed for the analysis of mutational signatures by incorporating LASSO regularization to address critical challenges associated with overfitting [23] (Fig. 1). It provides various metrics for the estimation of the optimal number of signatures such as bi-cross-validation [28] and adopts bootstrap sampling [35] for a robust estimation of mutational processes. Furthermore, RESOLVE extends previous methods [13, 17, 36] by incorporating confidence estimation based on bootstrap for mutational activities, enabling the computation of statistical *P*-values to assess signatures' presence. Additionally, it introduces clustering capabilities to group patients based on mutational profiles and a novel association framework leveraging regularized regression techniques to quantify bidirectional relationships between mutational signatures and driver gene mutations. In the following paragraphs, we describe the key novel features of the framework.

Efficient signature inference with LASSO regularization

RESOLVE redefines the optimization framework to perform the extraction of mutational signatures by employing LASSO regularization [23] and bi-cross-validation [28]. This approach enhances efficiency and precision, yielding signatures that are not only clearer but also better differentiated. The method avoids overfitting, striking a balance that is critical for accurate representation and interpretation of mutational processes. A key innovation in RESOLVE is the application of regularization to both the mutational signature and exposure matrices. This contrasts with the SparseSignatures framework [17], which applies regularization only to the signature matrix. This dual regularization improves robustness and interpretability, particularly in settings with complex or sparse mutation profiles.

Confidence assessment via bootstrapping

RESOLVE enables the assessment of confidence estimates measuring whether the inferred signatures are significantly active in cancer genomes. Leveraging bootstrap sampling [35], the framework iteratively performs the inference process, allowing the computation of *P*-values that quantify the confidence of the identified signatures. This unique capability enhances the robustness of the results, effectively distinguishing genuine mutational processes from potential noise. To the best of our knowledge, no other method provides confidence estimates of mutational signature activities.

Patients' stratification based on mutational processes

RESOLVE allows one to perform robust clustering of the samples, leveraging the identified signatures to explore subsets of patients characterized by similar mutational profiles. This analysis reveals underlying patterns within the data, facilitating the stratification of cancer genomes into mutational signature clusters based on shared mutational processes. These insights can offer a deeper understanding of distinct mechanisms of tumorigenesis, potentially laying the groundwork for personalized treatment strategies.

Association of mutational processes to prognosis or gene mutations

RESOLVE enables the association of specific mutational signatures with clinical data such as overall survival or genomic alterations. For survival data, it employs regularized Cox regression, while for genes mutations, two approaches are exploited: predicting mutations based on signature exposures using regularized logistic regression, or quantifying the influence of mutations on mutational processes through general regularized regression. In both cases, RESOLVE enables the computation of fold changes, providing insights into how much mutational signatures impact the likelihood of observing mutations and the extent to which mutations influence mutational processes. These analyses can offer valuable insights into the molecular mechanisms underlying cancer development and progression.

Performance evaluation on synthetic data

We conducted a comprehensive comparison of RESOLVE with the most used state-of-the-art methods through extensive testing on simulated data, encompassing a total of 16 distinct configurations. Specifically, we generated synthetic datasets with varying parameters, including 5 or 10 signatures, 100 or 300 samples, and 4 distinct generative models. During the dataset generation process, we ensured the inclusion of COSMIC SBS signatures SBS1 and SBS5 for all configurations. Additionally, we incorporated the following four signature selection strategies: (i) randomly selecting signatures from the remaining set in COSMIC; (ii) randomly selecting signatures from the dense subset within COSMIC; (iii) randomly selecting signatures from the sparse subset within COSMIC; and (iv) randomly generating signatures according to methods proposed in previous studies [17]. Count matrices were then inflated with uniform additive noise by introducing a number between 0 and 25 of random mutations [17]. For each configuration, we conducted 50 independent simulations, resulting in a total of 800 simulated datasets. Further details on the simulated data generation can be found in the "Materials and methods" section and Supplementary Materials.

The performance of RESOLVE was assessed in comparison to various methods for *de novo* signature extraction, including the standard NMF [27], nsNMF [37], SignatureAnalyzer [36], SigProfiler [13], and SparseSignatures [17] algorithms. Additionally, we evaluated RESOLVE's performance in estimating signature exposure by comparing it with

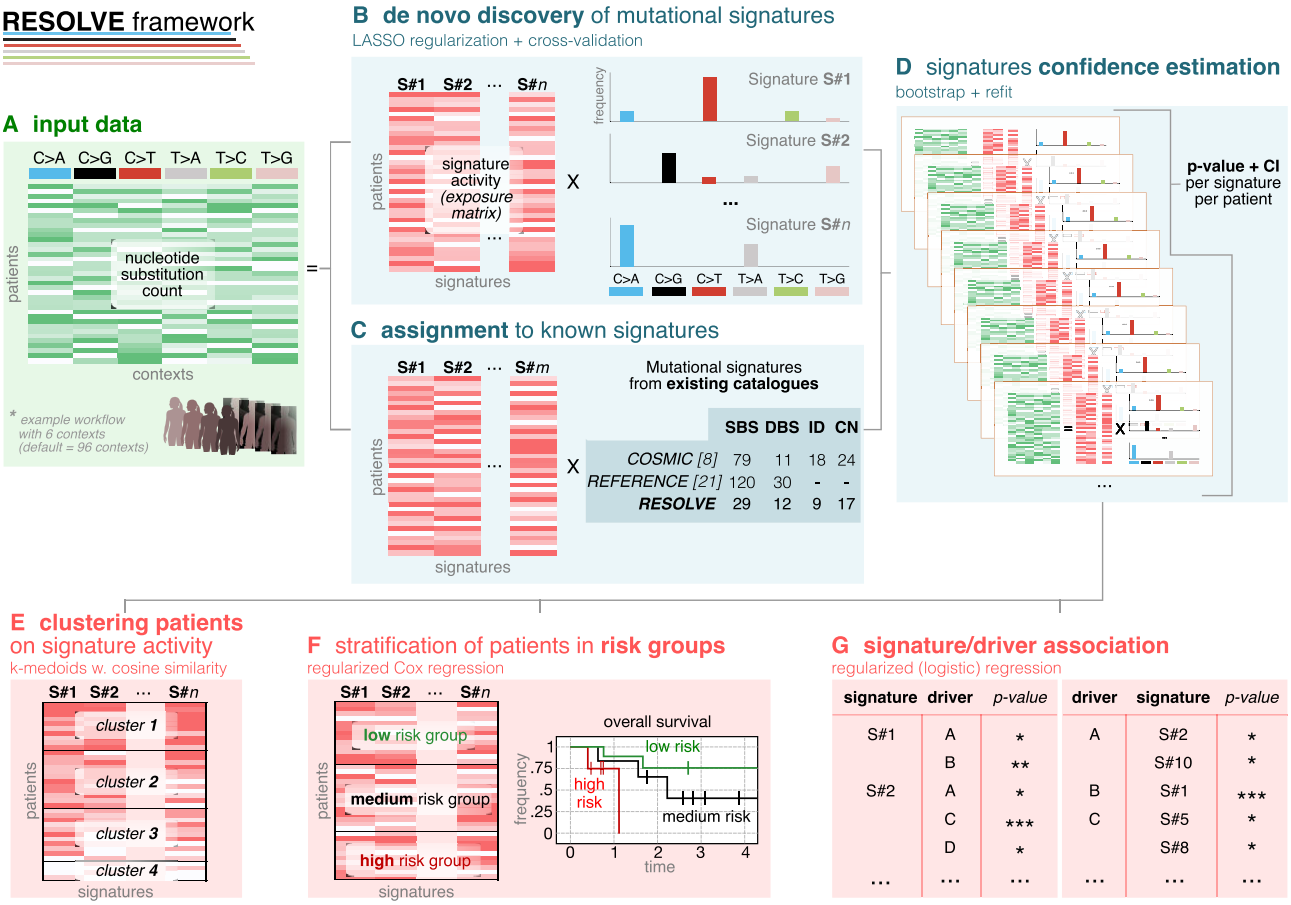


Figure 1. The RESOLVE framework, a comprehensive approach designed for the efficient extraction of mutational signatures from large-scale datasets. (**A**) The input data consist of mutation counts categorized based on nucleotide substitution contexts. (**B**) Subsequently, *de novo* discovery of mutational signatures is performed using LASSO regularization and cross-validation, enabling the identification of novel mutation patterns. (**C**) These signatures can then be associated with known catalogues, facilitating biological interpretation. (**D**) To ensure robustness, confidence estimation is performed using a bootstrapping and refitting procedure, refining the reliability of the detected signatures. (**E**) Patients are clustered based on their mutational signature activity using *k*-medoids clustering, revealing distinct patient subgroups. (**F**) Patients can be further stratified into risk groups (e.g. low, intermediate, and high) via regularized Cox regression, enabling prognostic assessment. (**G**) Finally, a signature–driver association analysis is conducted using regularized regression, uncovering potential links between mutational signatures and key driver mutations. Together, these steps provide a structured and statistically rigorous methodology to decipher the role of mutational signatures in cancer genomics.

`deconstructSigs` [38]. For all algorithms, default parameters were utilized, and we employed multiple metrics for comparison (see Supplementary Materials for details). Detailed results are provided as Supplementary Figs S1-- S11.

In particular, we show in Fig. 2 results for each method averaged over the 800 simulated datasets, created considering the four signature selection strategies described above. We evaluated three tasks: (i) *de novo* signature extraction, where both the exposure and signature matrices are estimated; (ii) signature extraction given the true number of signatures, with estimation of both the exposure and signature matrices; and (iii) signature exposure estimation, where the true signature matrix is provided as input, and the exposure matrix is then estimated. The performance of each algorithm was assessed in terms of the absolute error in the number of signatures compared to the ground truth. This metric aims to evaluate the quality of the different methods in estimating the optimal number of signatures. Subsequently, we analyzed the estimations of the exposure and signature matrices considering cosine similarity [39] and Euclidean distance. In Supplementary Figs S2-- S12, we also provide a detailed visualization of the results based on different metrics and selection criteria for the optimal number of signatures, averaged over the 800 simulated datasets. Additionally, we also report the results of running RESOLVE disabling the LASSO regularization step, which show a significant performance drop. This clearly highlights the relevance of regularization in the signature extraction task. The improved performance of RESOLVE over the other baseline methods, including `SparseSignatures` [17] (Fig. 2), can be partly attributed to its joint regularization strategy, which better captures both mutational signatures and their sample-specific contributions.

These results demonstrate that RESOLVE consistently outperforms the competing methods showing very high and stable accuracy across all settings.

Analysis of 20 000 whole-genome samples from adult and pediatric cancers

We performed *de novo* extraction of SBS, DBS, ID, and CN signatures on pan-cancer datasets encompassing both adult and pediatric cancers. Specifically, for adult tumors, we utilized SBS and DBS signature counts obtained from high-quality whole-genome sequencing data provided by Genomics

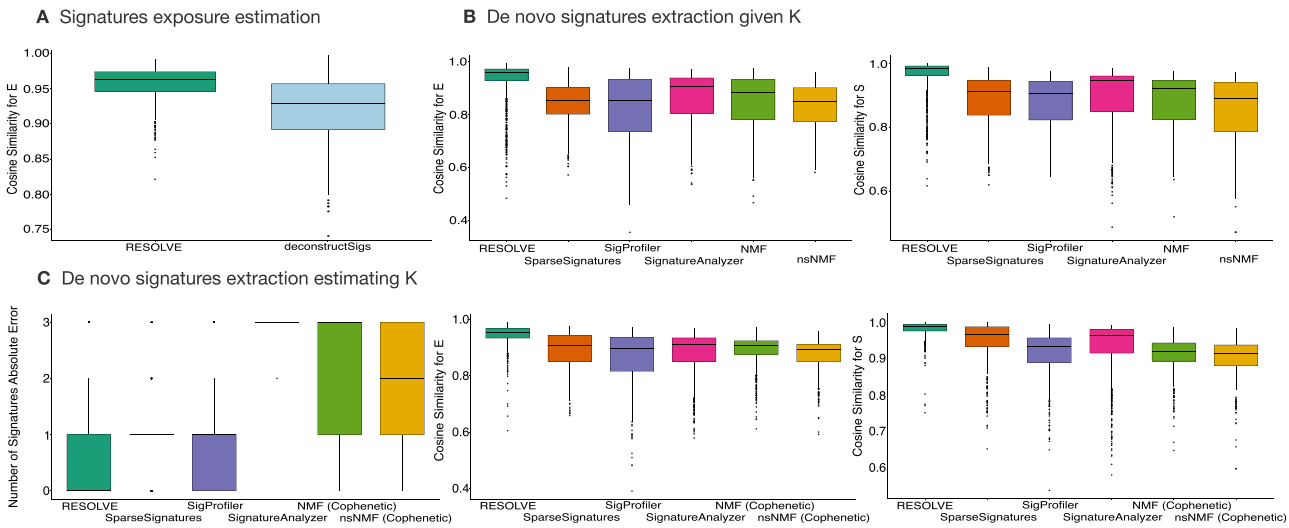


Figure 2. Overview of the results of simulations comparing RESOLVE with competing methods for *de novo* signature extraction (NMF [27], nsNMF [37], SignatureAnalyzer [36], SigProfiler [13], and SparseSignatures [17]) and signature exposure estimation (deconstructSigs [38]). In panel (**A**), we present the results of simulations for the signature exposure estimation problem, comparing RESOLVE to deconstructSigs [38]. The comparison is made in terms of cosine similarity for the exposure with respect to the generative models. In panel (**B**), we showcase the performance for the *de novo* signature extraction problem when the number of signatures (*K*) is provided as input, reporting the cosine similarity for both the exposure and signature matrices with respect to the generative models. Finally, in panel (**C**), we illustrate the performance for the *de novo* signature extraction problem when the number of signatures (*K*) is not given as input. This is assessed in terms of absolute error in the estimation of the number of signatures and cosine similarity for the exposure and signature matrices concerning the generative models. As the NMF and nsNMF methods do not provide a unique criterion to estimate the number of signatures, we report here results obtained considering the commonly used cophenetic correlation coefficient [27] heuristic.

England [21], Hartwig [40], and ICGC [41]. This resulted in a total of 18 640 samples with SBS counts spanning 21 distinct tissues and 13 962 samples with DBS counts across 20 different tissues. ID counts were derived from the ICGC projects providing whole-genome sequencing samples, yielding a total of 4360 adult samples with ID counts representing 20 tissues. Lastly, CN counts for adult cancers were derived from CN segments obtained from the TCGA studies [11], including a total of 10 444 samples from 30 distinct cancer types. The counts data and the signatures extracted by RESOLVE are provided as Supplementary Tables S1 and S3.

For pediatric cancers, we analyzed data from 27 distinct cancer types [22]. We processed a total of 783 samples with SBS counts, 570 samples with DBS counts, and 779 samples with ID counts. The counts data and the signatures extracted by RESOLVE are provided as Supplementary Tables S2 and S3.

For both adult and pediatric cancers, raw data were converted into count matrices using the functions provided by the RESOLVE R package.

Since cancer from different tissues can present very heterogeneous exposures, we conducted the inference by considering both individual tissue data and pan-cancer data independently for both adult and pediatric cancer datasets, as previously suggested [8, 21]. In all these configurations, we applied the RESOLVE *de novo* extraction framework and estimated the optimal number of signatures by bi-cross-validation [28]. This analysis led to the identification of a total of 29 SBS, 12 DBS, 9 ID, and 17 CN signatures. We compared the signatures extracted by RESOLVE and most of them resembled the known signature catalogues from [21] (SBS and DBS *reference signatures*) and COSMIC [8] (SBS, DBS, ID, and CN *COSMIC signatures* v3). The extracted signatures, along with their cosine similarities to known catalogues and assignments to both adult and pediatric cancers, are provided in Supplementary Tables S3 and S4.

As for DBS in pediatric cancers we only had 30 samples with at least 10 mutations, mostly being due to either COSMIC DBS1 (associated with the exposure to ultraviolet light) or DBS2 (associated with the exposure to tobacco smoking and other endogenous and/or exogenous mutagens such as cetaldehyde), we did not consider this configuration for further analyses.

To assess the impact of rare versus common mutational processes on the acquisition of somatic mutations in cancer genomes, we compared in terms of cosine similarity predicted versus observed mutational counts for the signatures by RESOLVE, and those in the reference and COSMIC catalogues. This was performed by comparing the observed mutation count vector for each sample and the reconstructed vector derived from the weighted combination of mutational signatures assigned to that sample.

For SBS, we obtained a total of 29 signatures by RESOLVE, and considered the 79 signatures by the COSMIC catalogue and the 120 signatures by the reference catalogue (see Table 1). In adult cancers, this led to a mean cosine similarity of the predicted counts of 0.98447, 0.99036, and 0.99241, in the three sets of signatures, respectively, and percentages of samples with cosine similarity >0.95 in their predictions of 97.205%, 98.680%, and 99.383%, respectively. In pediatric cancers, we observed mean cosine similarities of 0.93869, 0.94801, and 0.94986, and percentages of samples with cosine similarity >0.95 of 66.284%, 72.797%, and 76.884%, respectively. Notably, we identified 23 distinct mutational processes in the COSMIC catalogue based on the annotations from COSMIC v3.4, and several signatures with unknown etiology. Of these mutational processes, 17 were also detected

Table 1. Overview of mutational signatures identified in the RESOLVE catalogue and their correspondence to COSMIC and reference catalogues, along with associated etiologies as reported in COSMIC

RESOLVE	COSMIC	Reference	Associated etiology
S1	SBS1/0.92	SBS1/0.99	Spontaneous deamination of 5-methylcytosine
S2	SBS2/0.98	SBS2/0.99	AID/APOBEC activity
S3	SBS3/0.93	SBS3/0.82	HR deficiency
S4	SBS4/0.98	SBS4/0.97	Tobacco exposure
S5	SBS5/0.98	SBS5/0.93	Aging/tobacco smoking/NER deficiency
S6	SBS7a/0.99	SBS7a/1.00	UV light exposure
S7	SBS7b/0.96	NA	UV light exposure
S8	SBS8/0.82	SBS8/0.92	HR + NER deficiency
S9	SBS9/0.94	SBS9/0.86	Polymerase eta
S10	SBS10a/0.92	SBS10a/1.00	POLE mutations
S11	SBS10d/0.98	SBS10d/0.98	POLD1 mutations
S12	SBS11/0.99	SBS11/0.98	Temozolomide chemotherapy
S13	SBS13/0.99	SBS13/0.96	AID/APOBEC activity
S14	SBS14/0.65	SBS14/0.98	MMR deficiency + POLE mutation
S15	SBS15/0.92	SBS15/0.97	MMR deficiency
S16	SBS17a+b/0.94	SBS17/0.99	Damage by ROS/5FU chemotherapy
S17	SBS18/0.97	SBS18/0.92	Damage by ROS
S18	SBS19/0.95	SBS19/0.95	Unknown
S19	SBS20/0.90	SBS20/0.98	MMR deficiency + POLD1 mutation
S20	SBS22a/b/0.99	SBS22/0.99	Aristolochic acid exposure
S21	SBS23/0.93	SBS23/0.94	Unknown
S22	SBS26/0.94	SBS26/0.89	MMR deficiency
S23	SBS28/0.96	SBS28/0.87	POLE exonuclease domain mutation
S24	SBS31/0.96	SBS31/0.98	Platinum chemotherapy
S25	SBS32/0.94	SBS32/0.91	Azathioprine exposure
S26	SBS44/0.94	SBS44/0.97	MMR deficiency
S27	SBS88/0.86	SBS88/0.92	Colibactin exposure
S28	SBS92/0.87	SBS92/0.95	Tobacco exposure
S29	NA	SBS97/0.95	Unknown

by RESOLVE (Supplementary Table S14), with the missing 6 processes being either very rare—typically identified in specific and uncommon cancer subtypes—or cell lines not included in our datasets—or observed exclusively in samples that were not available for our analysis. These results highlight two main findings: (i) most SBS counts in samples can be fully explained with the extracted mutational signatures, even though we observe a slightly lower cosine similarity in pediatric cancer, likely due to the lower average number of mutations in these samples, reducing predictive power; and (ii) many mutational signatures contribute minimally to SBS counts, resulting in marginal improvement in the predicted counts between the RESOLVE signatures and those in the reference and COSMIC catalogues (Supplementary Figs S12 and S13).

For DBS, we extracted a total of 12 signatures by RESOLVE, 11 signatures were derived from the COSMIC catalogue, and 39 from the reference catalogue. In adult cancers, this resulted in mean cosine similarities of the predicted counts of 0.78880, 0.76850, and 0.85655, for the three sets of signatures, respectively. The percentages of samples with cosine similarity >0.95 in their predictions were 26.594%, 19.761%, and 34.272%, respectively. In this case, although most of the additional signatures present in the two signature catalogues did not provide a significant improvement in the goodness of fit of samples, we observe a much lower number of samples with cosine similarity >0.95. While this might be associated, in part, with the fact that DBS mutations are, on average, much less frequent compared to SBS mutations leading to a poorer fit, it might also be attributed to the possibility that DBS mutations may only partially result from mutational processes effectively mutating pairs of contiguous bases. In fact, some mutations might occur by chance in adjacent bases when mutational processes are active in physically close regions of the cancer genome (Supplementary Fig. S14).

For ID, we have a total of 9 signatures by RESOLVE and 18 by the COSMIC catalogue (ID signatures not available in the reference catalogue). In this case, in adult cancers we observe a mean cosine similarity of 0.8114 in the predictions by RESOLVE and 0.69346 in the ones considering the ID signatures by COSMIC. Moreover, 49.266% of samples had a prediction cosine similarity >0.95 by RESOLVE and 38.968% by COSMIC. Despite considering fewer signatures (9 versus 18), the signatures by RESOLVE exhibited higher predictive power in this scenario. Finally, in pediatric cancers, we observe a mean cosine similarity of 0.80276 for the signatures by RESOLVE and 0.82998 for the ones from COSMIC, with a percentage of samples with cosine similarity >0.95 of 46.147% and 62.901%, respectively. Also in this case, we observe a lower average prediction for all settings, possibly suggesting that not all ID mutations might be explained by ID mutational processes (Supplementary Figs S15 and S16).

Finally, for CN signatures, we extracted 17 signatures by RESOLVE and considered the 24 signatures provided by COSMIC, resulting for adult cancers in mean cosine similarities of the predicted CN counts of 0.92783 and 0.94053, respectively. Moreover, RESOLVE could fully explain (cosine similarity >0.95) 58.311% samples, while COSMIC signatures explained 66.057% samples (Supplementary Fig. S17).

All in all, these results suggest that only a relatively small set of mutational processes strongly contributes to the mutation patterns observed in cancer genomes, with the remaining processes being either rare (observed in a small set of patients) or

contributing only to a few mutations. We further investigate these findings in the next sections.

Robust estimation of signature activities and confidence assessment

In order to investigate the mutational processes that significantly contribute to the acquisition of somatic alterations in human cancers, we employed a bootstrap approach to assess which signatures were generating at least 5% of the observed mutations in each sample (see the "Materials and methods" section). The results are presented in Fig. 3 and as Supplementary Figs S12–S17. Signature assignments by bootstrap are provided as Supplementary Table S5.

Across all settings, the signatures generated by RESOLVE demonstrated comparable fitting to the samples when compared to those from published catalogues, despite RESOLVE employing a smaller set of signatures. Specifically, in the case of SBS, the RESOLVE set was almost one order of magnitude smaller compared to the published SBS catalogues. Additionally, even upon further reduction of signatures within the RESOLVE set contributing to the fit through bootstrap, there was no significant decrease in the goodness of fit. This supports the hypothesis that only a small subset of mutational processes primarily drive mutations in cancer genomes, while other signatures either are private of specific patients or contribute to a low number of mutations.

Interestingly, while the goodness of fit is consistently high for SBS across all catalogues, it notably appears lower for the other types of signatures for all catalogues, especially for DBS. DBS characterizes the concurrent modification of two consecutive nucleotide bases. The lower fitting of these types of mutational events by signature catalogues might be attributed to the lower number of mutations contributed by these processes, resulting in poor sampling of the actual mutational processes. However, another possibility is that only a few mutational processes mechanistically induce mutations in consecutive nucleotides within cancer genomes. Because of this, the observed DBS mutations to some extent could reflect stochasticity in processes affecting single bases, which, by chance, mutate consecutive positions in the genome of specific cancers. In both cases, caution must be taken when evaluating the presence or impact of these types of mutational signatures in human cancers.

Furthermore, it is interesting to observe that even within the smaller set of RESOLVE signatures, the goodness of fit quickly reaches a plateau when including additional mutational processes into the inference, with additional signatures not significantly improving the fit. For instance, for SBS in adult cancers (Supplementary Fig. S12), 10 signatures are sufficient to perfectly recapitulate mutations in over 80% of the samples. These signatures can be strongly associated with known mutational processes such as aging, methylations, APOBEC activity, exposure to UV light, etc.

These findings suggest that, despite the existence of a vast array of mutational processes potentially capable of causing mutations in cancer, patients' genomes are largely influenced by only a small subset of these processes. However, this observation might be biased by the enrichment of primary or early-stage samples in cancer datasets, including our cohorts. Indeed, we cannot rule out the possibility that at later stages, with increased genomic instability, additional mutational processes may come into play and contribute to the occurrence of mutations. Furthermore, our data are mostly derived from patients treated in the United States or Europe, which may result in potential underrepresentation of certain geographical areas. Therefore, mutational processes derived from exposure to specific environmental conditions might not be properly represented. These limitations, along with the impact of specific therapeutic approaches on mutational signatures, require further research.

Signature-based clustering of human cancers

RESOLVE allows us to investigate the existence of clusters of patients characterized by the same mutational processes. The goal of this analysis is to stratify cancer samples based on their signature exposure profiles, without including genomic or clinical covariates. This allows us to identify patient subgroups characterized by recurrent patterns of mutational processes, independently of specific driver mutations or outcomes. To this end, we performed *k*-medoids clustering based on the extracted mutational signatures (see the "Materials and methods" section). The results are presented in Supplementary Figs S18-- S27. Clustering assignments are provided as Supplementary Table S6.

For SBS, we identified 10 clusters (Fig. 4 and Supplementary Figs S18 and S19). Over 60% of the patients fall into clusters 1, 2, or 3, primarily characterized by three mutational processes: methylation (SBS1), likely aging (SBS5), and defective homologous recombination-based DNA damage repair (SBS3) [8]. The remaining clusters predominantly exhibit a single mutational signature attributed to either exogenous factors, such as tobacco smoking or UV light exposure, or endogenous factors like AID/APOBEC activity or defective DNA mismatch repair. In pediatric cancers, the landscape is simpler, with only two clusters, where methylation-induced mutations prevail (Supplementary Fig. S20).

In adult tumors, we identified 10 DBS clusters (Supplementary Figs S21 and S22). DBS2 (CC>AA), likely attributed to tobacco smoking exposure along with other endogenous and/or exogenous mutagens, emerges as the predominant signature, characterizing cluster 4 (1795 samples), cluster 1 (alongside DBS5, associated with chemotherapy treatment using platinum drugs), and cluster 2 (alongside DBS4, with an unknown etiology) [8]. Once again, most clusters are dominated by a single mutational signature.

When considering ID (Supplementary Figs S23–S25), we observe six clusters in adult cancers and four in pediatric cancers. In both scenarios, ID1 and ID2 emerge as the most prevalent signatures across the clusters, previously linked to slippage during DNA replication [8]. Among the signatures represented in adult cancers, we also note ID5 (of unknown etiology), ID6 (associated with defective homologous recombination-based DNA damage repair), ID8 (related to the repair of DNA double-strand breaks and TOP2A mutation), and ID9 (also of unknown etiology).

Finally, the analysis of CN signatures in adult cancers revealed nine clusters (Supplementary Figs S26– S27). The larger clusters, clusters 1 and 2, are associated with CN1 and CN2, representing diploidy and tetraploidy genomes, respectively [11]. Other observed signatures are CN7 (linked to chromothripsis), as well as CN9 and CN12 (both associated with loss of heterozygosity). Representing a rarer signature but strongly impacting cluster 9 is CN17, associated with homologous recombination deficiency [11].

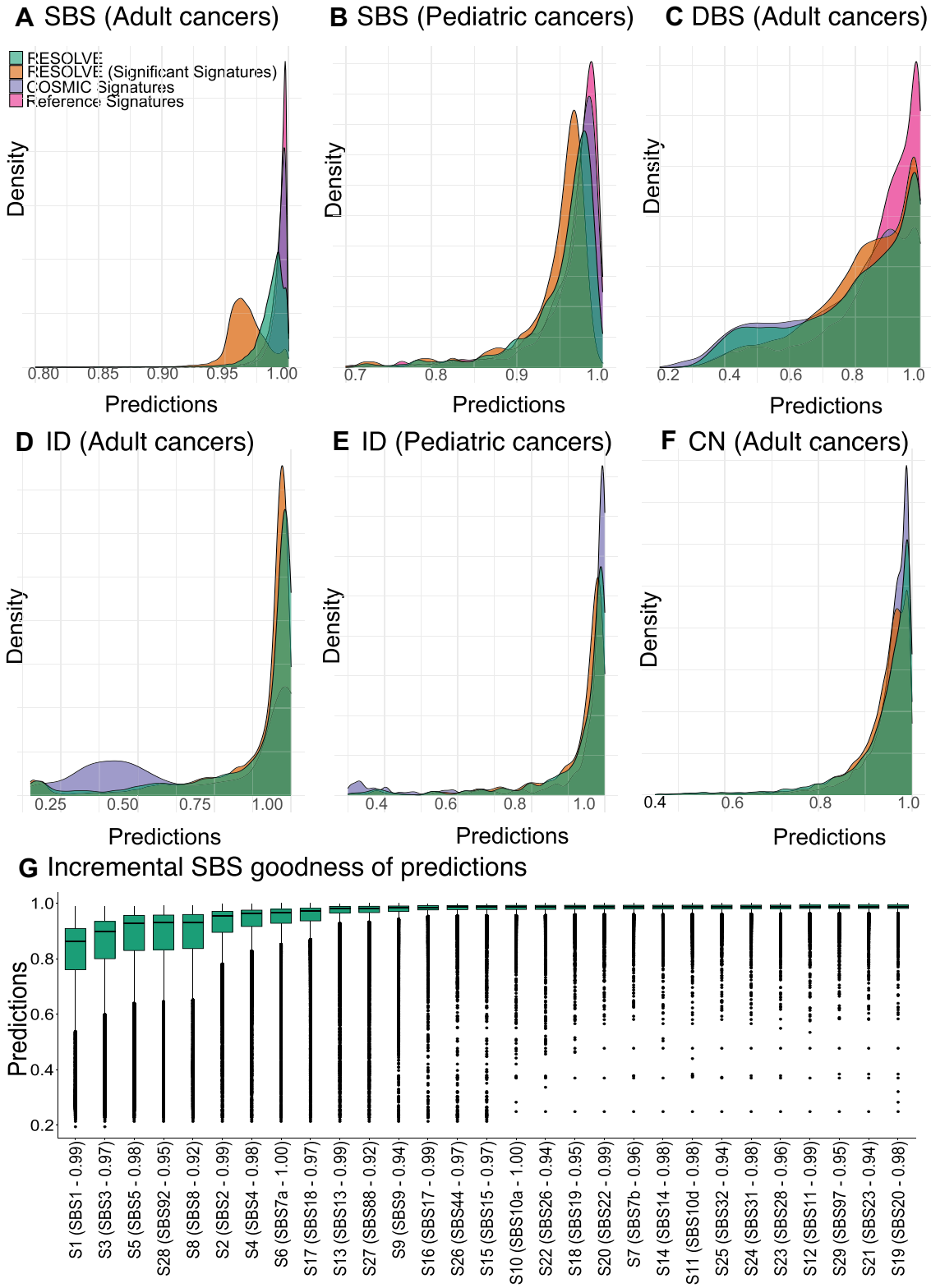


Figure 3. We show the goodness of fit in terms of cosine similarity between observed and predicted counts, comparing the results from RESOLVE (with and without robust estimation by bootstrap) with those from the COSMIC and the reference signature catalogues. (**A**) The results for SBS signatures in adult cancers. (**B**) The results for SBS signatures in pediatric cancers. (**C**) The results for DBS signatures in adult cancers. (**D**) The results for ID signatures in adult cancers. (**E**) The results for ID signatures in pediatric cancers. (**F**) The results for CN signatures in adult cancers. Finally, in panel (**G**), we show the incremental goodness of fit of SBS given the signatures extracted by RESOLVE.

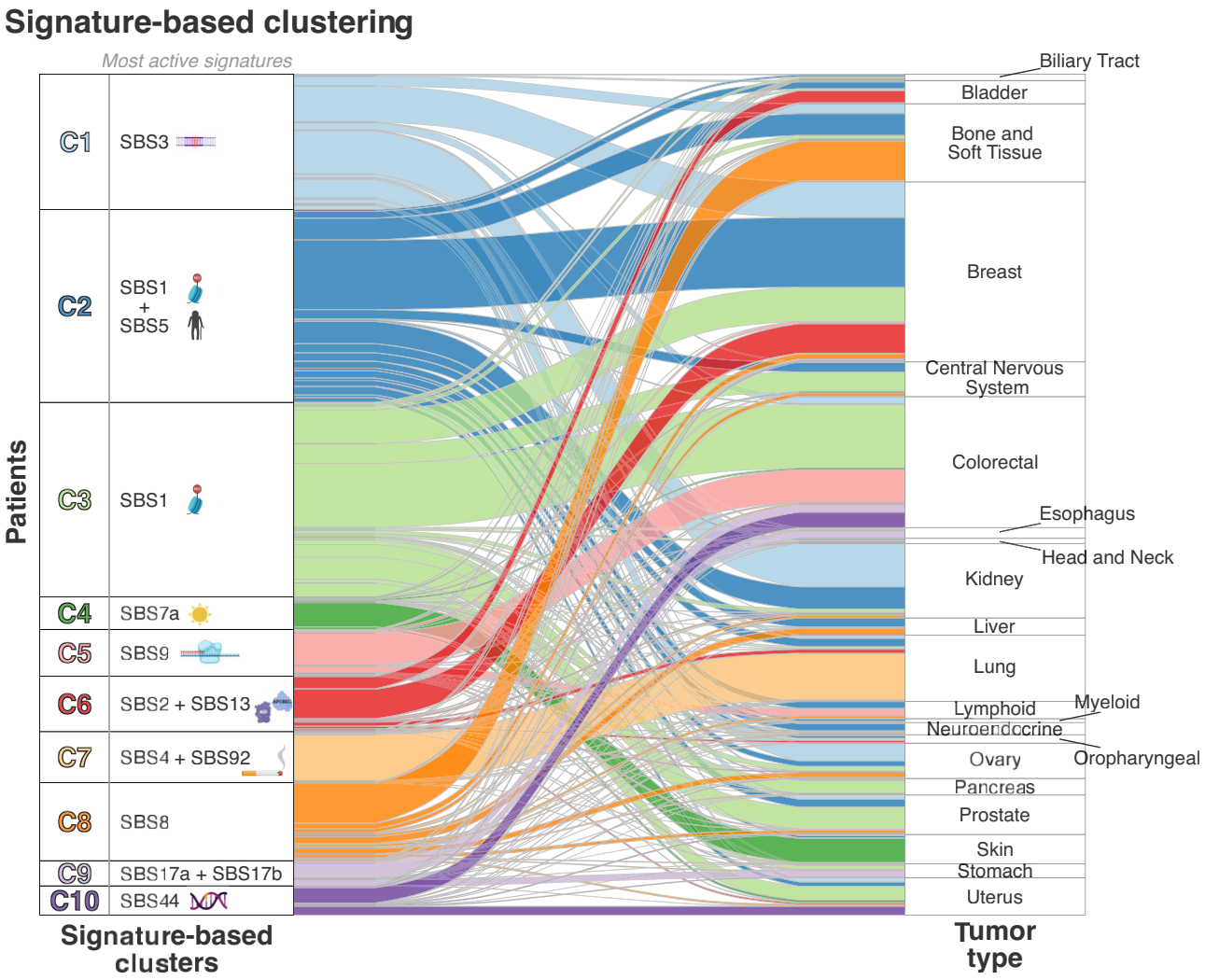


Figure 4. The RESOLVE signature-based clustering analysis revealed that the majority of adult cancer patients (>60% from clusters 1, 2, or 3) are primarily characterized by three mutational processes: methylation (SBS1), likely aging (SBS5), and defective homologous recombination-based DNA damage repair (SBS3). This indicates that a limited number of mutational processes significantly contribute to tumorigenesis. The remaining clusters predominantly exhibit a single mutational signature attributed to either exogenous factors, such as tobacco smoking or UV light exposure, or endogenous factors like AID/APOBEC activity or defective DNA mismatch repair. Created in BioRender. Mologni, L. (2025) https://BioRender.com/7obl4lq.

In conclusion, our clustering analysis unveils a complex landscape portraying various mutational processes of diverse origins, encompassing both endogenous and exogenous factors. These processes potentially induce mutational events across varying scales within cancers. However, it is noteworthy that the mutational processes accounting for a significant number of mutations constitute only a small fraction of the signatures documented in published catalogues. This indicates the predominance of a small number of mutational processes potentially driving tumorigenesis in most patients. These clusters summarize shared patterns of mutational processes across cancer types and may serve as a framework for future investigations into their clinical or molecular relevance.

Association of mutational signatures with prognosis

RESOLVE enables the identification of significant associations between mutational signatures and prognosis, offering insights not previously reported due to the comprehensive nature of the framework and to the scale of the analyzed datasets. To this end, we considered a set of 2231 samples from 15 distinct tissues with complete survival information and a goodness-of-signatures fit >0.95 cosine similarity. Given the reduced set of samples with complete data with respect to the full cohort, we focused on SBS. We performed survival association between normalized exposures and overall survival data, first at the pan-cancer level, and then considering cancer types/tissues with at least 100 samples.

The association between SBS and overall survival was performed with regularized Cox regression. Samples were clustered based on the risk predicted by the inferred regularized Cox models and the predicted risk groups were confirmed to show significant differences in prognosis via standard Kaplan–Meier analysis (P-value <.05). Then, we performed enrichment analysis to verify the presence of genes mutated at significantly different frequencies in the determined clusters (see the "Materials and methods" section). Novel findings include the identification of distinct clusters characterized by unique prognostic outcomes linked to specific mutational signatures, such as SBS9 correlating with better

prognosis in hematopoietic cancers and SBS3 correlating with poorer outcomes in multiple cancers, including esophageal and pancreatic cancers. Full results are presented in Fig. 5 and Supplementary Figs S28-- S39. Enrichment analyses are provided in Supplementary Table S7.

At the pan-cancer level, we inferred four clusters of samples showing different prognoses (Supplementary Fig. S28). Cluster 1 is enriched with hematopoietic and lymphoid cancers. Cluster 2 is composed mostly of hematopoietic and lymphoid cancers, liver cancers, prostate, and skin cancers. Clusters 3 and 4 are composed mostly of esophageal and pancreatic cancers (Supplementary Fig. S29). SBS9 (associated with mutations induced during replication by polymerase eta as part of somatic hypermutation in lymphoid cells [8]) is the predominant signature associated with lower risk and is enriched in cluster 1, while SBS5 (clock-like signature), also associated with lower risk, is highly present in clusters 1 and 2. The most frequent signatures associated with worse prognosis were SBS1 (spontaneous deamination of 5-methylcytosine [8]), SBS3 (defective homologous recombination-based DNA damage repair [8]), and SBS17 (unknown etiology) (Supplementary Fig. S30). Notably, clock-like signatures have been associated with variable risk in different cancers [42, 43] possibly due to an ambiguous resolution of the real aging processes, e.g. disentangling SBS1 (clock-like CpG methylation) from an SBS5 background. The ability to better disambiguate individual mutational signatures in patients allows RESOLVE to strongly assign SBS5 to better prognosis at the pan-cancer level, which is in line with its suggested baseline mutational rate, compared with stronger environmental processes. Interestingly, our analysis attributed SBS1 to a set of patients showing high risk.

In terms of differential genes, cluster 1 showed higher frequencies of BCL2, CREBBP, and EZH2. The other clusters (clusters 2–4) showed higher frequencies of long genes, which might be due to an overall higher mutation burden of cancers belonging to these clusters, associated with stronger mutagenic processes. Finally, clusters 3 and 4, with worse prognosis, showed higher frequencies of mutations in KRAS and TP53 (Supplementary Table S7). TP53 pathway genes mutations have been previously associated with SBS1 and high risk, in line with our observations [44].

Esophageal cancers with better prognosis present a higher incidence of SBS5 and SBS18 (possibly caused by damage from reactive oxygen species [8]) (Supplementary Figs S31 and S32) and mutations in ATM and NTRK3. SBS9 was confirmed to be associated with better prognosis in hematopoietic and lymphoid cancers, while SBS3 was associated with worse overall survival (Supplementary Figs S33 and S34).

In pancreatic cancers, SBS5 is found to be strongly associated with lower risk (Supplementary Figs S35 and S36), while patients in the higher-risk group show significantly higher frequencies of mutations in KRAS, SMAD4, and TP53 (Supplementary Table S7). In prostate cancers, SBS3 is the signature associated with worse prognosis (Supplementary Figs S37 and S38), with mutations in FAT4 and LRP1B showing much higher frequencies in these patients (Supplementary Table S7).

Furthermore, although skin cancers, especially melanomas, are dominated by the UV light signature (SBS7a and SBS7b [8]), the tumors showing higher risk present an enrichment of SBS7b, which is usually rarer compared to SBS7a (Supplementary Figs S39 and S40). In these cancers with worse prognosis, we find significantly higher frequencies of mutations in CBLB, CHD2, CREBBP, CSF3R, ESRP1, JAK3, PDGFRB, and STIL (Supplementary Table S7).

Finally, we validated some of the gene mutations associated with specific risk groups using the cBioPortal platform [45, 46], verifying that the overall survival of patients carrying the identified mutations consistently correlated with prognosis. These results are shown in Supplementary Fig. S41.

Association of mutational signatures with the acquisition of mutations in driver genes

RESOLVE's robust statistical framework can provide novel insights into the relationships between mutational signatures and driver mutations. Unlike previous methods, RESOLVE can quantify both how mutational signatures influence the likelihood of specific driver mutations and how mutations impact mutational processes, providing fold-change estimates to highlight the strength of these associations.

For this purpose, we examined a dataset comprising 3842 samples from 18 different tissues, each with complete data and a goodness-of-signatures fit exceeding 0.95 cosine similarity. Due to the reduced sample size with complete data in comparison to the full dataset, also in this case we focused our analysis on SBS. We conducted association analyses between normalized exposures and mutations, initially at the pan-cancer level, and subsequently for the cancer types/tissues with at least 100 samples. These included breast cancer, central nervous system cancers, esophageal cancer, hematopoietic and lymphoid cancers, liver cancer, pancreatic cancer, prostate cancer, and skin cancer. The association analysis employed regularized regression as follows (see the "Materials and methods" section). First, we treated the mutation data as response variables, to be predicted by the normalized signature exposures. This allowed us to quantify the impact of each signature on the probability of observing a mutation, utilizing regularized logistic regression, given the binary nature of the response variables (mutation present/absent). In the second approach, we considered the normalized exposures by RESOLVE as the response variables, with mutations acting as predictors. Here, the aim was to quantify how much a given mutation increases the activity of the different mutational processes. This analysis utilized a general regularized regression with Gaussian response. We repeated the analysis for both synonymous and nonsynonymous mutations. An overview of the results is shown in Fig. 6. Full results are provided as Supplementary Tables S8--S11 and Supplementary Figs S42–S57.

At the pan-cancer level, we observe strong associations of SBS7a (UV light [8]) and SBS17 (unknown etiology [8]) with most of the considered genes, but excluding prominent oncogenes like KRAS and PIK3CA. These associations reflect the increased mutational burden associated with these mutational processes. Interestingly, KRAS mutations were associated with SBS1 (linked to methylation [8]) and SBS3 (defective homologous recombination-based DNA damage repair [8]), while PIK3CA mutations were associated with SBS1 and SBS5 (clock-like process [8]). This might be due to subtype-specific mechanisms.

Considering the two most mutated genes in breast cancer, namely TP53 and PIK3CA, we observe an association of TP53 mutations with SBS3 and SBS13 (AID/APOBEC activity [8]), and PIK3CA mutations with SBS5 and SBS2 (AID/APOBEC activity [8]). The associations, respectively, of TP53 with SBS3

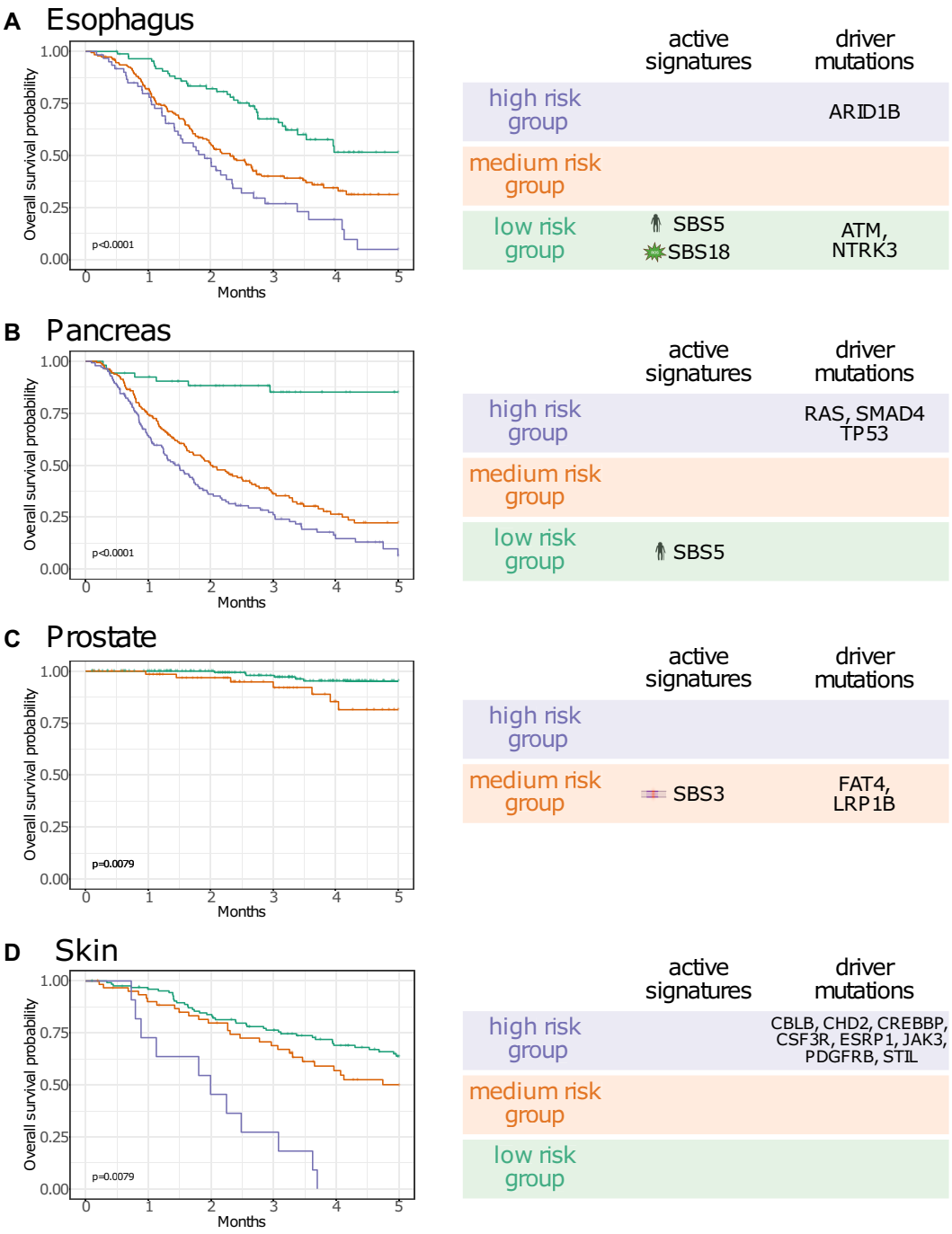


Figure 5. The RESOLVE framework enables the association of mutational processes with prognosis. We considered a set of 2231 samples from 15 distinct tissues with complete survival information and a goodness-of-signatures fit >0.95 cosine similarity. We show risk groups stratified by overall survival probabilities (e.g. low, intermediate, and high) in several cancer types: (**A**) esophagus, (**B**) pancreas, (**C**) prostate, and (**D**) skin. Survival curves highlight the impact of active mutational signatures and associated driver mutations on patient prognosis. Statistical significance is indicated, with *P*-values denoting differences between groups. Key mutational signatures (e.g. SBS5, SBS18, and SBS3) and their corresponding driver mutations (e.g. ATM, SMAD4, and TP53) are linked to distinct risk categories, illustrating their potential role in tumor progression and patient outcomes. Created in BioRender. Mologni, L. (2025) https://BioRender.com/ax6zh8s.

and PIK3CA with SBS5 are somewhat expected, given that breast tumors exhibiting SBS3 are typically of the triple-negative subtype, which often harbor TP53 mutations, and PIK3CA mutations are enriched in ER-positive breast tumors where SBS5 prevalence is higher. However, the intriguing aspect is the association of APOBEC mutations, attributed to either SBS2 or SBS13, with either TP53 or PIK3CA mutations. Thus, TP53 and PIK3CA might not follow the canonical breast cancer signatures for a subset of patients. The higher incidence of PIK3CA mutations associated with SBS2 may suggest a causal relationship: the predominant activating PIK3CA mutation in our dataset is a T[C>T]A, constituting 28.8% of the reported mutations, which aligns with the mutation profile of SBS2. This novel finding, if validated in additional cohorts,

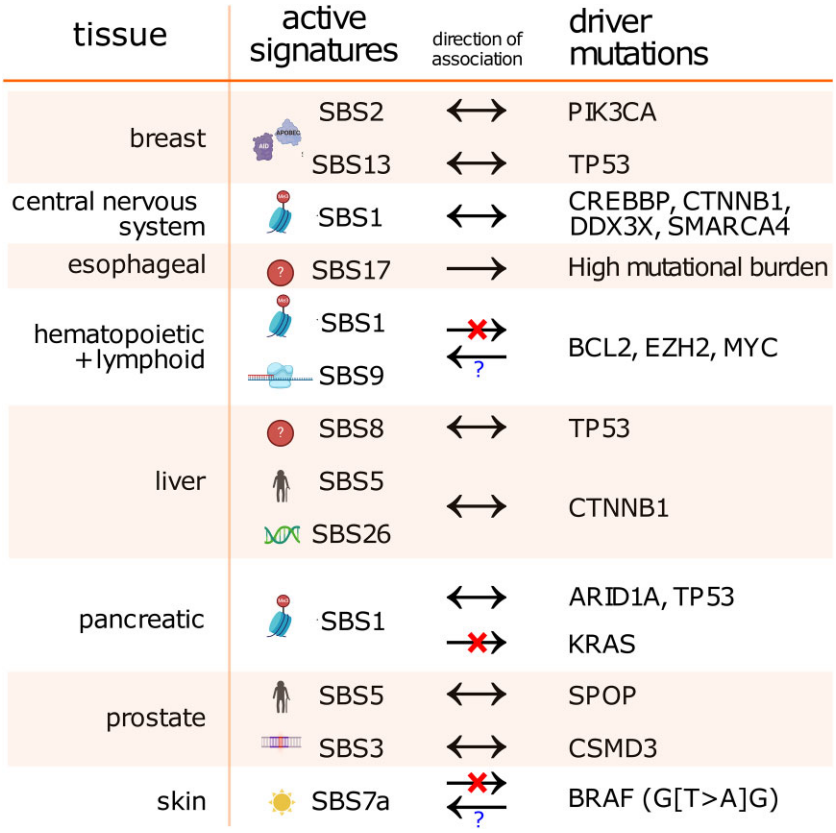


Figure 6. We show the relationships between active mutational signatures and key driver mutations in various cancer types, including liver, breast, central nervous system, esophageal, hematopoietic and lymphoid, pancreatic, prostate, and skin cancers. Specific mutational signatures (e.g. SBS1, SBS2, SBS5, and SBS7a) are linked to driver mutations (e.g. PIK3CA, TP53, KRAS, and BRAF), with arrows indicating the direction of association. Some mutations are associated with increased mutational burden, while others have uncertain effects. This visualization highlights potential mechanistic links between mutational processes and oncogenic drivers in different tissues. Created in BioRender. Mologni, L. (2025) https://BioRender.com/9o4u4ot.

has important implications, as it implies a potential direct association between one of the main driver genes and a specific mutational process active in breast cancers. Other notable associations include AKAP9 mutations with SBS2 and SBS3, KMT2C mutations with SBS13, consistent with the observed association of KMT2C deficiency with increased APOBEC expression and promotion of its associated mutational process [47], and MUC16 mutations with SBS2.

In central nervous system tumors, we observe an association of mutations in CREBBP, CTNNB1, DDX3X, and SMARCA4 with SBS1, while SBS5 is linked to tumors exhibiting a generally higher mutational burden and increased mutation rates in long genes such as MUC16 and MUC4, in both nonsynonymous and synonymous mutations. Also, in esophageal cancers, we observe an increase in mutations of long genes such as CSMD1, CSMD3, GLI3, LRP1B, MUC16, and SPTA1 associated with SBS17 (unknown etiology). SBS17 is generally associated with a high mutational burden, which explains the enrichments of long genes. Additionally, we also observe a significant increase in mutations in PEG3, both synonymous and nonsynonymous. TP53 mutations are also associated with SBS17.

In hematopoietic and lymphoid tumors, we observed associations between long genes such as CSMD3, LRP1B, and MUC16 with signatures SBS5, SBS9 (mutations induced during replication by polymerase eta [8]), or SBS17, potentially indicating a higher mutational burden associated with these signatures. Additionally, we noted significant association of mutations in EZH2 and MYC with tumors dominated by SBS1 signature, and BCL2 and EZH2 with tumors bearing SBS9. However, these oncogenes had mutations inconsistent with the related tumor signature. In other words, these mutations are not highly likely to have originated from these signatures. This lack of association suggests either that the mutations precede the signatures during cancer evolution or that these oncogenes have somehow been protected during evolution from the most probable mutagenic factors blood cells might be exposed to in the blood.

The most frequent mutations in liver cancers, namely TP53 and CTNNB1, are, respectively, associated with SBS8 (unknown etiology), SBS5, and SBS26 (defective DNA mismatch repair [8]), as already documented in colorectal cancer, where mutated CTNNB1 may be due to mutational events favored by defective DNA mismatch [48].

In pancreatic cancers, we observe an association of long genes (APOB, MUC4) with SBS3, indicating a high mutational burden. SBS1, on the other hand, is associated with mutations in ARID1A, KRAS, and TP53. Interestingly, here again, the strongest driver, i.e. the oncogene KRAS, does not show consistent mutations with SBS1, with 38.5% of mutations being of type A[C>T]C, 28.7% A[C>A]C, and 16.4% C[C>G]A. In prostate cancer, mutations in SPOP are associated with SBS5, while mutations in CSMD3 are associated with SBS3.

Finally, in skin cancers, the UV light signature (SBS7a), as expected, is associated with mutations in long genes such

Table 2. Significant associations between mutational signatures and main driver mutations identified by RESOLVE

Driver gene	Signature	Cancer type	Reported associations
ARID1A	S1/SBS1	Pancreatic	Novel
BCL2	S9/SBS9	Hematopoietic and lymphoid	Novel
BRAF	S6/SBS7a	Skin	[50, 51]
CREBBP	S1/SBS1	Central nervous system	Novel
CSMD3	S3/SBS3	Prostate	Novel
CTNNB1	S1/SBS1	Central nervous system	Novel
CTNNB1	S5/SBS5	Liver	Novel
CTNNB1	S22/SBS26	Liver	Novel
DDX3X	S1/SBS1	Central nervous system	Novel
EZH2	S1/SBS1	Hematopoietic and lymphoid	Novel
EZH2	S9/SBS9	Hematopoietic and lymphoid	Novel
KRAS	S1/SBS1	Pancreatic	Novel
KRAS	S3/SBS3	Pan-cancer	Novel
MYC	S1/SBS1	Hematopoietic and lymphoid	Novel
PIK3CA	S1/SBS1	Pan-cancer	Novel
PIK3CA	S2/SBS2	Breast	[50, 51]
PIK3CA	S5/SBS5	Breast	Novel
SMARCA4	S1/SBS1	Central nervous system	Novel
SPOP	S5/SBS5	Prostate	Novel
TP53	S1/SBS1	Pancreatic	Novel
TP53	S3/SBS3	Breast	[52]
TP53	S8/SBS8	Liver	Novel
TP53	S13/SBS13	Breast	Novel

Associations previously discussed in the literature are reported.

as APOB, FAT3, LRP1B, MUC16, and SPTA1, indicating that it is generating a lot of mutations. SBS7a is also associated with tumors carrying BRAF mutations, but like we reported for other oncogenes, the observed BRAF mutations in skin cancers are mostly G[T>A]G (50%), which is inconsistent with the distribution of mutations attributed to SBS7a. This once again suggests that either BRAF mutations occur early in the cancer evolution, possibly before SBS7a, or, alternatively, BRAF has acquired protection from C>T mutations due to high exposure to UV light in the skin, during species evolution. This observation is consistent with previous reports suggesting that driver mutations may be subject to selective pressures that deviate from the dominant mutational processes in a given tumor type [49].

Our analysis revealed a trend across various cancer types. Strong oncogenes associated with specific tissues display activating mutations that contradict the alterations expected from the most significant mutational signature in those tissues. This might suggest that either the oncogenes precede the signature in the evolution of the specific cancer or human evolution has shielded cells in specific tissues from the more potent exogenous or endogenous mutational processes, or possibly both. Finally, we repeated the analysis considering COSMIC signatures to assess whether using this catalogue led to additional associations. However, this analysis yielded similar results, confirming RESOLVE signatures as the ones significantly associated with driver acquisition (Supplementary Tables S12 and S13).

We finally assessed which of the observed associations were previously reported. To this end, we cross-referenced our findings with three recent major studies on driver–signature interactions [50–52]. A summary of the main significant associations identified by RESOLVE, along with their novelty status, is provided in Table 2.

Discussion

This study presents RESOLVE, a robust computational framework for analyzing mutational signatures in cancer genomes. RESOLVE introduces multiple advancements over previous methods, including statistical confidence estimation, robust clustering, and a novel framework for association analysis. These features fill critical gaps in current methodologies and significantly enhance the biological and clinical utility of mutational signature analysis.

Our findings demonstrate that RESOLVE can accurately fit mutational profiles with a significantly reduced number of signatures compared to existing catalogues. This efficiency underscores the notion that a few key mutational processes primarily drives cancer development. This does not indicate that rare signatures do not exist, but they likely represent a non-significant contribution to cancer etiology in most patients. Therefore, RESOLVE prioritizes statistically significant signatures affecting a broad range of patients, which aligns with the goal of identifying robust mutational drivers. In other words, we may be able to describe the ongoing processes with simpler models, without losing accuracy. Nevertheless, rare signatures might be relevant for understanding certain rare DNA repair deficiencies or small patient subgroups. Future developments could aim to integrate complementary approaches to address this specific scenario.

Despite the high goodness of fit for SBS signatures, the lower fit observed for DBS and other signatures highlights the complexity of these mutational events. This discrepancy could be due to the lower frequency of mutations from these processes or inherent stochasticity in their occurrence.

Our clustering analysis reveals distinct groups of patients characterized by specific mutational processes. In SBS, three clusters dominate, driven by methylation (SBS1), aging (SBS5), and defective homologous recombination (SBS3). These findings suggest that a limited number of processes significantly

contribute to tumorigenesis, aligning with the hypothesis that only a small subset of mutational signatures are critical in most cancers. The identified clusters provide a framework for exploring relationships between mutational profiles and cancer subtypes, offering opportunities for targeted therapeutic interventions and further mechanistic studies. Future studies could, for example, utilize this resource to investigate tissue-specific mutational mechanisms or explore therapeutic vulnerabilities in cluster-defined patient subgroups.

Importantly, we discovered significant associations between specific mutational signatures and patient prognosis. For instance, SBS9 is associated with better outcomes in hematopoietic and lymphoid cancers, while SBS3 correlates with poorer prognosis. These associations provide valuable prognostic markers that could guide treatment strategies.

Furthermore, our analysis links mutational signatures to driver gene mutations, shedding light on subtype-specific mechanisms and the evolutionary timeline of mutations in cancer. Some strong observed associations, such as the one between SBS2 and PIK3CA mutations in breast cancer, suggest potential causal relationships and underline the importance of understanding the interplay between mutational processes and oncogenic drivers. Interestingly, we found a previously unrecognized anomaly in the mutational processes involving strong driver oncogenes, such as KRAS, BRAF, and MYC, in some tissues: our results suggest that some key driver genes do not seem to adhere to the general signature observed in the same tumor. Although this requires further investigations, it may suggest that strong oncogenes follow a private evolution with respect to the global mutational landscape, either because they are hit at a pre-cancerous stage or because their sequence evolved to protect them from common mutational mechanisms.

Our findings have several implications. The ability of RESOLVE to reduce the number of required signatures without compromising fit accuracy offers an interpretable approach to mutational signature analysis. This could enhance the precision of cancer diagnostics and the development of targeted therapies.

Future research should explore the dynamics of mutational processes in advanced cancer stages and under therapeutic pressures. Understanding how these processes evolve and interact could provide deeper insights into cancer biology and inform more effective treatment strategies.

In conclusion, RESOLVE represents a significant advancement in mutational signature analysis, offering robust and efficient estimation of signature activities with profound implications for cancer research and personalized medicine. Our work highlights the dominant role of a few key mutational processes in cancer development and their potential as prognostic markers and therapeutic targets.

Acknowledgements

The graphical abstract was created in BioRender with an academic license. Mologni, L. (2025) https://BioRender.com/9o4u4ot. Figures 4 and 6 were partially created with BioRender.com.

Author contributions: Conceptualization and methodology: D.R. Software: L.D. and D.R. Investigation: M.V., F.M., A.M.V., N.C., A.A., and D.R. Visualization: F.M., A.G., and D.R. Funding acquisition: M.A., G.C., A.G., R.P., L.M., and D.R. Supervision: L.M. and D.R. Writing—original draft: D.R. All authors read and approved the final manuscript.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was partially supported by a Bicocca 2020 Starting Grant to D.R. R.P. acknowledges support by the Italian Ministry of University and Research (MIUR)—Department of Excellence project PREMIA (PREcision MedIcine Approach: bringing biomarker research to the clinic), European Union—NextGenerationEU through the Italian Ministry of University and Research under PNRR—M4C2-I1.3 Project PE_00000019 'HEAL ITALIA' and Italian MUR Dipartimenti di Eccellenza 2023–2027 (l. 232/2016, art. 1, commi 314–337) and from the Italian Foundation for Cancer Research (AIRC) under IG 2023, ID 29341 project. Support was also provided by the Cancer Research UK and Associazione Italiana per la Ricerca sul Cancro (CRUK/AIRC) "Accelerator Award" (22790) "Single-cell Cancer Evolution in the Clinic" to A.G., G.C., and M.A., and by the European Commission Program PPPA2027, PPPA-2021-AIPC #LC-01815952/101052609, "Towards an UNIque approach for artificial intelligence data-driven solutions to fight Childhood cAncer FOR Europe" (UNICA4EU) and the 2021 FAQC program of the Università degli Studi di Milano-Bicocca to M.A. G.C. acknowledges funding from the Italian Foundation for Cancer Research (AIRC) under MFAG 2020, ID 24913 project. L.M. acknowledges funding from the Italian Foundation for Cancer Research (AIRC) under IG 2020, ID 24828 project. Partial support to A.G. and M.A. was also provided by the MUR under the grant "Dipartimenti di Eccellenza 2023–2027" of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Italy.

Data availability

The datasets used in this study are publicly available. Whole-genome sequencing data from adult cancer samples obtained from Genomics England, Hartwig Medical Foundation, and the International Cancer Genome Consortium (ICGC) were previously published in [21]. Pediatric cancer data were sourced and are available from [22]. Processed data, including mutational count matrices and extracted signatures, are provided as supplementary materials accompanying this publication. RESOLVE is available as an R package and can be installed from Bioconductor at https://www.bioconductor.org/packages/release/bioc/html/RESOLVE.html or from GitHub at https://github.com/ramazzottilab/RESOLVE. The R scripts to replicate all the results are available at https://github.com/danro9685/RESOLVE-UTILITIES. A permanent snapshot of the code associated with this publication has been deposited in Zenodo and is available at https://doi.org/10.5281/zenodo.15683389.

References

1. Vogelstein B, Papadopoulos N, Velculescu VE *et al.* Cancer genome landscapes. *Science* 2013;**339**:1546–58. https://doi.org/10.1126/science.1235122
2. Koh G, Degasperi A, Zou X *et al.* Mutational signatures: emerging concepts, caveats and clinical applications. *Nat Rev Cancer* 2021; **21**:619–37.
3. Secrier M, Li X, De Silva N *et al.* Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nat Genet* 2016;**48**:1131–41. https://doi.org/10.1038/ng.3659
4. Wang S, Jia M, He Z *et al.* APOBEC3B and APOBEC mutational signature as potential predictive markers for immunotherapy response in non-small cell lung cancer. *Oncogene* 2018;**37**:3924–36. https://doi.org/10.1038/s41388-018-0245-9
5. Pich O, Muiños F, Lolkema MP *et al.* The mutational footprints of cancer therapies. *Nat Genet* 2019;**51**:1732–40. https://doi.org/10.1038/s41588-019-0525-5
6. Nik-Zainal S, Alexandrov LB, Wedge DC *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* 2012;**149**:979–93. https://doi.org/10.1016/j.cell.2012.04.024
7. Koh G, Zou X, Nik-Zainal S. Mutational signatures: experimental design and analytical framework. *Genome Biol* 2020;**21**:1–13. https://doi.org/10.1186/s13059-020-1951-5
8. Alexandrov LB, Kim J, Haradhvala NJ *et al.* The repertoire of mutational signatures in human cancer. *Nature* 2020;**578**:94–101. https://doi.org/10.1038/s41586-020-1943-3
9. Kucab JE, Zou X, Morganella S *et al.* A compendium of mutational signatures of environmental agents. *Cell* 2019;**177**:821–36. https://doi.org/10.1016/j.cell.2019.03.001
10. Drews RM, Hernando B, Tarabichi M *et al.* A pan-cancer compendium of chromosomal instability. *Nature* 2022;**606**:976–83. https://doi.org/10.1038/s41586-022-04789-9
11. Steele CD, Abbasi A, Islam SA *et al.* Signatures of copy number alterations in human cancer. *Nature* 2022;**606**:984–91. https://doi.org/10.1038/s41586-022-04738-6
12. Maura F, Degasperi A, Nadeu F *et al.* A practical guide for mutational signature analysis in hematological malignancies. *Nat Commun* 2019;**10**:1–12.
13. Alexandrov LB, Nik-Zainal S, Wedge DC *et al.* Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* 2013;**3**:246–59. https://doi.org/10.1016/j.celrep.2012.12.008
14. Baez-Ortega A, Gori K. Computational approaches for discovery of mutational signatures in cancer. *Brief Bioinform* 2019;**20**:77–88. https://doi.org/10.1093/bib/bbx082
15. Fischer A, Illingworth CJ, Campbell PJ *et al.* EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol* 2013;**14**:1–10. https://doi.org/10.1186/gb-2013-14-4-r39
16. Rosales RA, Drummond RD, Valieris R *et al.* signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics* 2017;**33**:8–16. https://doi.org/10.1093/bioinformatics/btw572
17. Lal A, Liu K, Tibshirani R *et al. De novo* mutational signature discovery in tumor genomes using SparseSignatures. *PLoS Comput Biol* 2021;**17**:e1009119. https://doi.org/10.1371/journal.pcbi.1009119
18. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet* 2014;**15**:585–98. https://doi.org/10.1038/nrg3729
19. Nik-Zainal S, Davies H, Staaf J *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 2016;**534**:47–54. https://doi.org/10.1038/nature17676
20. Alexandrov LB, Ju YS, Haase K *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* 2016;**354**:618–22. https://doi.org/10.1126/science.aag0299
21. Degasperi A, Zou X, Dias Amarante T *et al.* Substitution mutational signatures in whole-genome–sequenced cancers in the UK population. *Science* 2022;**376**:abl9283. https://doi.org/10.1126/science.abl9283
22. Thatikonda V, Islam SA, Autry RJ *et al.* Comprehensive analysis of mutational signatures reveals distinct patterns and molecular processes across 27 pediatric cancers. *Nat Cancer* 2023;**4**:276–89. https://doi.org/10.1038/s43018-022-00509-4
23. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser A Methodol* 1996;**58**:267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
24. Alexandrov LB, Jones PH, Wedge DC *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet* 2015;**47**:1402–7. https://doi.org/10.1038/ng.3441
25. Martincorena I, Fowler JC, Wabik A *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* 2018;**362**:911–7. https://doi.org/10.1126/science.aau3879
26. Riva L, Pandiri AR, Li YR *et al.* The mutational signature profile of known and suspected human carcinogens in mice. *Nat Genet* 2020;**52**:1189–97. https://doi.org/10.1038/s41588-020-0692-4
27. Brunet JP, Tamayo P, Golub TR *et al.* Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci USA* 2004;**101**:4164–9. https://doi.org/10.1073/pnas.0308531101
28. Owen AB, Perry PO. Bi-cross-validation of the SVD and the nonnegative matrix factorization. *Ann Appl Stat* 2009;**3**:564–94. https://doi.org/10.1214/08-AOAS227
29. Mazumder R, Hastie T, Tibshirani R. Spectral regularization algorithms for learning large incomplete matrices. *J Mach Learn Res* 2010;**11**:2287–322.
30. Efron B. Bootstrap methods: another look at the jackknife. In: *Breakthroughs in Statistics*. Springer, 1992, 569–93. https://doi.org/10.1007/978-1-4612-4380-9_41
31. Wilson DJ. The harmonic mean *p*-value for combining dependent tests. *Proc Natl Acad Sci USA* 2019;**116**:1195–200. https://doi.org/10.1073/pnas.1814092116
32. Kaufman L, Rousseeuw PJ. Partitioning around medoids (program pam). In: *Finding Groups in Data*. Vol. **344**, 1990, 68–125. https://doi.org/10.1002/9780470316801.ch2
33. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987;**20**:53–65. https://doi.org/10.1016/0377-0427(87)90125-7
34. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 2008;**24**:719–20. https://doi.org/10.1093/bioinformatics/btm563
35. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. CRC Press, 1994. https://doi.org/10.1201/9780429246593
36. Tan VY, Févotte C. Automatic relevance determination in nonnegative matrix factorization with the β-divergence. *IEEE Trans Pattern Anal Mach Intell* 2012;**35**:1592–605. https://doi.org/10.1109/TPAMI.2012.240
37. Pascual-Montano A, Carazo JM, Kochi K *et al.* Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Trans Pattern Anal Mach Intell* 2006;**28**:403–15. https://doi.org/10.1109/TPAMI.2006.60
38. Rosenthal R, McGranahan N, Herrero J *et al.* DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol* 2016;**17**:1–11. https://doi.org/10.1186/s13059-016-0893-4
39. Tan PN, Steinbach M, Kumar V. *Introduction to Data Mining*. India: Pearson Education, 2016.
40. Priestley P, Baber J, Lolkema MP *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* 2019;**575**:210–216. https://doi.org/10.1038/s41586-019-1689-y
41. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* 2020;**578**:82–93. https://doi.org/10.1038/s41586-020-1969-6
42. Karihtala P, Kilpivaara O, Porvari K. Mutational signatures and their association with cancer survival and gene expression in

multiple cancer types. *Int J Cancer* 2025;**156**:114–29. https://doi.org/10.1002/ijc.35148
43. Chong W, Wang Z, Shang L *et al.* Association of clock-like mutational signature with immune checkpoint inhibitor outcome in patients with melanoma and NSCLC. *Mol Ther Nucleic Acids* 2021;**23**:89–100. https://doi.org/10.1016/j.omtn.2020.10.033
44. Hakobyan A, Meyenberg M, Vardazaryan N *et al.* Pan-cancer analysis of the interplay between mutational signatures and cellular signaling. *iScience* 2024;**27**:109873. https://doi.org/10.1016/j.isci.2024.109873
45. Cerami E, Gao J, Dogrusoz U *et al.* The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012;**2**:401–4. https://doi.org/10.1158/2159-8290.CD-12-0095
46. Gao J, Aksoy BA, Dogrusoz U *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013;**6**:pl1. https://doi.org/10.1126/scisignal.2004088
47. Hu X, Biswas A, De S. KMT2C-deficient tumors have elevated APOBEC mutagenesis and genomic instability in multiple cancers. *NAR Cancer* 2022;**4**:zcac023. https://doi.org/10.1093/narcan/zcac023
48. Arnold A, Tronser M, Sers C *et al.* The majority of β-catenin mutations in colorectal cancer is homozygous. *BMC Cancer* 2020;**20**:1–10.
49. Supek F, Lehner B. Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. *Cell* 2017;**170**:534–47. https://doi.org/10.1016/j.cell.2017.07.003
50. Temko D, Tomlinson IP, Severini S *et al.* The effects of mutational processes and selection on driver mutations across cancer types. *Nat Commun* 2018;**9**:1857. https://doi.org/10.1038/s41467-018-04208-6
51. Poulos RC, Wong YT, Ryan R *et al.* Analysis of 7,815 cancer exomes reveals associations between mutational processes and somatic driver mutations. *PLoS Genet* 2018;**14**:e1007779.
52. Sørensen SG, Shrikhande A, Poulsgaard GA *et al.* Pan-cancer association of DNA repair deficiencies with whole-genome mutational patterns. *eLife* 2023;**12**:e81224.

Received: October 23, 2024. **Revised:** May 31, 2025. **Editorial Decision:** June 13, 2025. **Accepted:** June 19, 2025