Comparing De Novo and COSMIC Mutational Signatures in Single-Cell Sequencing Data

Fedde de Haas
Supervisors: Joana de Pinho Gonçalves, Ivan Stresec, Sara Costa
EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Fedde de Haas
Final project course: CSE3000 Research Project
Thesis committee: Joana de Pinho Gonçalves, Ivan Stresec, Sara Costa, Catharine Oertel Genannt Bierbach

Abstract

Understanding mutational processes active in cancer at the single-cell level is essential for characterizing intra-tumor heterogeneity. Previous studies extracted these processes, called mutational signatures, and the known signatures can be found in the Catalogue of Somatic Mutations in Cancer (COSMIC) database. These signatures were derived based on bulk sequencing data of thousands of whole genomes. This study proposes and applies a systematic method to compare single-cell-derived de novo mutational signatures to the COSMIC signatures. Using two single-cell cancer datasets (breast and neck cancer), two stable signatures were extracted per dataset. Within each dataset, the de novo signatures were extremely similar (cosine similarity > 0.96), suggesting uniform mutational processes within individual tumors. No direct one-to-one matches were found between de novo and COSMIC signatures. However, the de novo signatures can be interpreted as combinations of known mutational processes. These results demonstrate the feasibility of extracting de novo signatures based on single-cell data, while also highlighting limitations due to possible overfitting. Future work should include simulation experiments, analysis of additional tumors, and evaluation of alternative signature extraction methods.

Introduction

Cancer is caused by somatic mutations in the DNA that allow the cell to start reproducing rapidly. These mutations include, for example, single base substitutions (SBS) where one base is replaced by another, and deletions or insertions of DNA. Mutations are caused by mutagens, e.g. UV light and tobacco smoke, and high exposure to the mutagens will result in a higher likelihood that a mutation occurs and is not repaired. After a cell becomes malignant, it can start to acquire many more mutations that cause the tumor to become heterogeneous within itself. Due to the evolutionary nature of cancer, some mutations can allow a cell to increase its likelihood of survival during chemotherapy, resulting possibly in a relapsing tumor (Stratton, Campbell, and Futreal 2009).

Mutagens leave behind characteristic patterns in the somatic mutations, called mutational signatures. For example, one of the signatures that results from smoking causes predominantly C:G>A:T substitutions, while one of the signatures caused by UV light exposure results in mainly C:G>T:A substitutions. Finding the signatures responsible for the mutations in cancer leads to a better understanding of the cause of the tumor. This can, in some cases, be used to determine the best treatment options and predict the effectiveness of the treatment (Abbasi and Alexandrov 2021). Knowing the underlying signatures behind cancer is therefore an important area of cancer research.

Somatic mutations in a sample can be summarized by a mutational profile. This is a count vector of SBS and the mutated base's immediate 5' and 3' base neighbors, resulting in 96 total mutation types, which can later be normalized to yield a probability distribution over mutation types. Each mutation is categorized into one of 96 types, where all mutations are added together in one vector. Mutational signatures are inferred from these profiles.

The combined somatic mutations of the genome are the result of exposures to different mutagens. Each cancer genome, represented by a mutational profile, can be expressed as a linear combination of mutational signatures and the corresponding number of mutations caused by this process, reflecting the exposure to different mutagens.

The best method available for extracting de novo mutational signatures is the SigProfilerExtractor algorithm (Islam et al. 2022; Alexandrov, Nik-Zainal, et al. 2013). Here, the mutational signatures are found by making use of non-negative matrix factorization (NMF) (Lee and Seung 1999), where the whole mutational catalog is decomposed into a matrix representing the signatures and another matrix representing the signature exposures for each cell (Alexandrov, Nik-Zainal, et al. 2013).

The COSMIC (Catalogue Of Somatic Mutations In Cancer) database provides a comprehensive collection of known mutational signatures derived from large-scale bulk sequencing data, such as the Pan-Cancer Analysis of Whole Genomes (PCAWG) dataset, which contains mutational data of more than 2,500 donors ("Pan-cancer analysis of whole genomes" 2020). Here, a total of 67

SBS signatures were identified, many of which do not have a known process that causes them (Alexandrov, Kim, et al. 2020; Sondka et al. 2024).

The mutational signature extraction process has so far been based on bulk sequencing data, which aggregates mutations from many cells ("Pan-cancer analysis of whole genomes" 2020). However, individual cells gather many more mutations after becoming cancerous, resulting in a diverse spectrum of mutations. A better understanding of the mutations of individual cells results in a better understanding of tumor heterogeneity. With most research so far having been conducted based on bulk sequencing data, research on single-cell sequencing data remains relatively new and less validated, making it unclear how well the existing methodologies translate when using single-cell data.

Single-cell data allows for more precise mutation analysis, as each rare mutation can be analyzed rather than being averaged out by bulk data. De novo mutational signatures, patterns of mutations inferred directly from the data without prior assumptions, can be obtained from this data. Leading to the potential discovery of novel signatures that would not be found from just the bulk data. Interpretation of de novo extracted signatures requires the comparison to known signatures, such as those cataloged in the COSMIC database, allowing researchers to validate their findings and the biological phenomena they correlate with.

In this study, we address the gap between mutational signature analysis in bulk sequencing and single-cell sequencing data. We present a method to systematically compare de novo signatures derived from single-cell data with reference signatures in the COSMIC database. This approach allows us to evaluate whether known mutagenic processes can be detected at the single-cell level and to identify novel, previously unobserved mutational signatures. This method is then applied to two single-cell sequenced datasets of breast and neck cancer, containing 688 and 53065 samples, respectively.

Methodology

This methodology explains the steps required to compare de novo signatures based on single-cell data. The purpose is to classify de novo mutational signatures based on their relation with a set of reference signatures. Here, the mutational signatures are first extracted de novo. Then the samples are fitted to a known set of mutational signatures to find the reconstruction based on known signatures. Based on the found signatures, a method combining a direct one-to-one comparison between de novo and reference signatures and a decomposition of de novo signatures into reference signatures. Lastly, Statistics on the diversity of the samples are then calculated.

analysis of the mutational profiles

Knowing the diversity of mutational profiles across single-cell samples helps determine the expected number of signatures. Moreover, single-cell datasets can contain thousands of samples per donor, with highly variable sample sizes across different datasets. To avoid a bias towards datasets with disproportionately large sample counts, we reduce the number of cells analyzed per dataset. This also ensures computational feasibility, as having more samples primarily increases the computation time without improving the quality of signatures found. Previous work has shown that 200 cancer genomes are sufficient to accurately decipher up to 20 signatures (Alexandrov, Nik-Zainal, et al. 2013).

For the mutational profile clustering, K-medoids is applied on the mutational catalogs using cosine similarity. K-means has been used before to cluster mutational signatures (Alexandrov, Nik-Zainal, et al. 2013); however, it requires Euclidean distances. This cannot be used for non-normalized mutational profiles, as mutation counts vary substantially between samples. K-medoids works similarly to K-means; they differ in that K-medoids selects an existing data point as a centroid rather than using the cluster mean. This also allows other distance metrics to be used. The number of clusters is determined by making use of the elbow method in combination with the silhouette score.

To visualize the data, UMAP is applied based on cosine similarity, which is a nonlinear method for dimensionality reduction. UMAP has a similar performance to other nonlinear methods on

single-cell data; however, it is much faster (Becht et al. 2019). Data points are then labeled based on the cluster assignment. The visualization shows the general structures of the data and how well the cluster assignments are separated.

To reduce the sample space, 30 samples are randomly selected from each cluster. This paper focuses on de novo signature extraction based on a single cancer tissue, and having 30 samples per cluster allows for sufficient data to capture the various signatures (Alexandrov, Nik-Zainal, et al. 2013). If there are not enough samples in a cluster, resampling with replacement is applied to ensure balanced representation of various processes. When multiple single-cell datasets are analyzed, this number can be decreased since multiple samples from a single cluster tend to have similar mutational profiles and therefore contribute less unique information.

Extraction of de novo mutational signatures

NMF-based Extraction

De novo signatures are extracted using `SigProfilerExtractor`, which applies Non-negative matrix factorization (NMF) as is described by Alexandrov et al (Alexandrov, Nik-Zainal, et al. 2013). Samples are represented by a mutation count matrix M (mutation types x samples). This matrix is approximated using two matrices: P (mutation types x signatures), representing all the mutational signatures, and E (signatures x samples), representing each sample's exposure to those signatures, as can be seen in Equation 1. The number of signatures k needs to be much smaller than the number of samples n, leading to the discovery of general patterns in the P and E matrices. Using NMF, the matrices P and E can then be estimated.

$$\underbrace{\begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,n} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ m_{m,1} & m_{m,2} & \cdots & m_{m,n} \end{bmatrix}}_{\substack{\text{Mutational catalog } M \\ (m \times n)}} \approx \underbrace{\begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,k} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m,1} & p_{m,2} & \cdots & p_{m,k} \end{bmatrix}}_{\substack{\text{Signatures Matrix } P \\ (m \times k)}} \times \underbrace{\begin{bmatrix} e_{1,1} & e_{1,2} & \cdots & e_{1,n} \\ e_{2,1} & e_{2,2} & \cdots & e_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ e_{k,1} & e_{k,2} & \cdots & e_{k,n} \end{bmatrix}}_{\substack{\text{Exposures Matrix } E \\ (k \times n)}} \tag{1}$$

The NMF is performed on bootstrapped data across multiple iterations. The extracted signatures are clustered, and the centroids of each cluster are taken as the final de novo signatures. The performance of the de novo solution is evaluated using the average silhouette width of each cluster, referred to as stability. This represents the consistency of the solutions across iterations. Additionally, average cosine similarity between original samples and their reconstruction is used to assess the reconstruction quality (Alexandrov, Nik-Zainal, et al. 2013).

Extraction of de novo signatures

For this, the `SigProfilerExtraction` package for Python is used. The number of NMF replicates is set to 25, which corresponds to the NMF iterations mentioned in the previous section. The number of candidate signatures is explored, ranging from 1 up to a maximum determined by the number of mutational clusters identified during sub-sampling, typically capped at one above the cluster count. The number of clusters approximates the upper bound of distinct mutational processes present, therefore being a natural constraint for the number of signatures. The maximum number of signatures is capped at 25 to reduce computation time. The solution right before the stability drops below 0.8 is selected. When the stability does not drop below 0.8, additional signatures are extracted until the solution is no longer stable.

Fitting of known mutational signatures

Explanation of the method

To decompose the samples into a set of reference mutational signatures, `SigProfilerAssignment` is used. Let $\vec{v}$ be a column vector of matrix M, representing the mutational profile of a sample, and let S be a matrix containing k reference signatures. The objective is to estimate the exposure

vector $\vec{a}$ of length k as described by Equation 2, where the goal is to estimate $\vec{a}$ by finding a solution that both minimizes the reconstruction error and the number of signatures fitted.

$$\vec{v} \approx S \times \vec{a} \tag{2}$$

`SigProfilerAssignment` solves this using an iterative method based on nonnegative least squares (NNLS). First, the NNLS solution is computed using all signatures; then, each signature's contribution is evaluated by removing that signature and calculating the reconstruction error increase for the new NNLS solution. The signature causing the smallest increase, provided it is below a predefined threshold, is removed. This process is repeated until no signature removal can be done without exceeding the threshold. This process is then reversed, continually reintroducing signatures that reduce the error beyond a second threshold. These removal and reintroduction steps are repeated until convergence (Díaz-Gay et al. 2023).

Fitting reference signatures

Decomposition of the samples into a set of reference signatures reveals the known active signatures present in the samples, which sets an expectation for the de novo signatures extracted. Additionally, the cosine reconstruction similarity provides insight into how well-known the mutational processes behind the samples are. The `SigProfilerAssignment` Python package is used, using Cosmic version 3.4.

Comparison between signatures

Similarity metric

Signatures are compared using cosine similarity, as defined in Equation 3. Cosine similarity is chosen because it measures the degree of alignment between two vectors, independent of their magnitude. This property also makes it well-suited for comparing mutational catalogs with signatures, as it effectively accounts for differences in mutation counts by only taking the angle between vectors into account. Furthermore, cosine similarity has been widely used in prior studies for mutational signature comparison, including foundational work by Alexandrov et al. (Alexandrov, Nik-Zainal, et al. 2013; Alexandrov, Kim, et al. 2020), which supports its applicability in this context.

$$similarity(A, B) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\|\|\vec{B}\|} \tag{3}$$

Direct comparison

To accommodate a systematic comparison between de novo signatures and a set of reference signatures, a one-to-one matching between both sets of signatures is employed. This method is particularly appropriate when it is expected that the de novo signatures are closely related to the existing set of reference signatures. This will function as a first check to see what signatures have a strong match (cosine similarity above 0.85) to existing signatures and what signatures could represent either novel or a combination of known processes.

For this problem, we try to maximize the total sum of cosine similarities between the matched signatures. Here let $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ be the set of de novo signatures, and $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$ be the set of COSMIC signatures. Note that these sets of signatures are not necessarily the same size. Let $M \subseteq \mathcal{D} \times \mathcal{C}$ be the set of matched signature pairs in a one-to-one matching. Then, the 1-to-1 matching problem can be formulated as the following optimization:

$$\max_{M} \sum_{(d_i, c_j) \in M} \cos(d_i, c_j) \tag{4}$$

Subject to the following constraints that enforce strict one-to-one matching:

$$\forall d_i \in \mathcal{D}, \quad \sum_{c \in \mathcal{C}} (d_i, c) \in M \leq 1 \tag{5}$$

$$\forall c_i \in \mathcal{D}, \quad \sum_{d \in \mathcal{D}} (d, c_i) \in M \leq 1 \tag{6}$$

This problem can be efficiently solved by modeling it as a minimum cost maximum flow (MCMF) algorithm. For this algorithm, we model each signature $d \in \mathcal{D}$ and signature $c \in \mathcal{C}$ as a separate node in a flow graph. Edges are added between each d and c, where the capacity is 1 with the edge having a cost of $1 - cos(d, c)$. The source is then connected exactly once to each of the $d \in \mathcal{D}$ with a capacity of 1 and a zero cost; similarly, each $c \in \mathcal{C}$ is connected to the sink with capacity 1 and zero cost. Having only a single edge between the source, sink, and the other nodes ensures that Equations 5 and 6 are enforced. In this study, we implement this approach using the NetworkX Python library.

Analysis of composite signatures

Some de novo signatures may not match any COSMIC signature with high similarity, above 0.85. The remaining de novo signatures could then be either composite, based on multiple known processes, or novel signatures. To analyze this, we additionally decompose the de novo signatures into a set of reference signatures using the `SigProfilerAssignment` algorithm, which is also used to determine the contributions of known mutational processes in a sample.

The reconstruction cosine similarity of the signatures can then be used to determine the type of signature found; A cosine similarity above 0.85 for the direct comparison suggests it is well-explained by a known signature, whereas a decomposition similarity above 0.85 indicates the signature is likely composite, if neither is true, it hints at a possible novel signature.

Results

This study analyzes two single-cell mutational datasets to evaluate the proposed methodology. The first dataset consists of 688 mutational profiles from a breast cancer sample. The second dataset includes 53,065 single-cell samples from a Laryngeal Squamous Cell Carcinoma (neck cancer), collected across four spatial regions of the tumor. These datasets illustrate both the heterogeneous nature of cancer and the large scale of single-cell sequencing data.

Data comparison

The optimal number of clusters for the breast cancer sample was determined based on Figure 1c, which shows the change in cluster assignment cost and silhouette score as the number of clusters increases. No elbow is observed in the cost curve, while the silhouette score starts to decline more rapidly beyond three clusters. Based on this, three clusters are chosen. Having no clear inflection point indicates that the data might not contain distinct substructures. Likewise, figure 1f illustrates the equivalent plot for the neck cancer dataset. Both a clear peak in the silhouette score and an elbow in the total cost are observed at 5 clusters, resulting in five clusters being chosen.

Figures 1d and 1g show the UMAP projection of the mutational profile data, colored by k-medoids cluster assignment. No distinct visual groupings are observed in the breast cancer dataset, although the assigned clusters show moderate separation in UMAP space. In contrast, the neck cancer samples show five distinct groups. However, the K-medoids assignment does not align perfectly with the visually apparent groups. This may be due to the different objectives of both algorithms; UMAP preserves local structures in high-dimensional space, whereas k-medoids assigns each sample to the closest medoid.

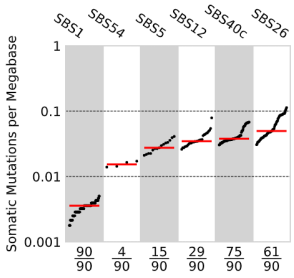
Figures 1e and 1h show the heatmap of cosine similarities between the medoids of each group. In both datasets, the minimum pairwise cosine similarity exceeds 0.97, suggesting that there might not be distinct processes behind the mutational profiles. Though the clusters are more clearly distinguished for neck cancer in the UMAP projection, the similarity between the medoids is also very high. The groupings might be caused by subtle regional differences within the tumor, as the samples were collected across four spatially distinct regions.

Figures 1a and 1b show the contributions of COSMIC mutational signatures in the mutational profiles. Some signatures appear consistently across all samples, for example, SBS1 in the breast cancer dataset and both SBS1 and SBS5 in the neck cancer dataset. While the proportions of these

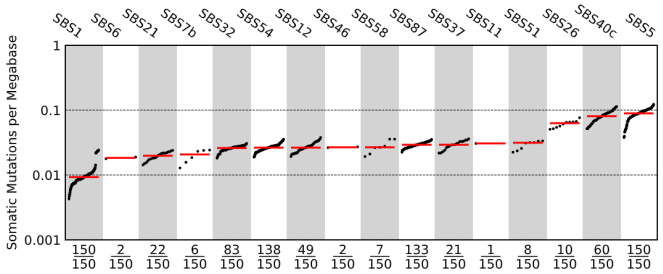
COSMIC signatures differ slightly between samples, no major differences are observed. Therefore, it is expected to extract superpositions of these processes rather than isolate individual, directly correlated COSMIC signatures.

Table 1 presents the cosine similarity between the mutational profiles and their COSMIC-based reconstruction. In the breast cancer dataset, average similarity is approximately 0.85, but nearly half of the samples fall below this threshold, indicating poor reconstruction by known COSMIC signatures. In contrast, the neck cancer reconstruction shows higher similarity values, averaging at 0.927, with no samples falling below the 0.85 threshold. This suggests known mutational processes are behind the mutations in the neck cancer.

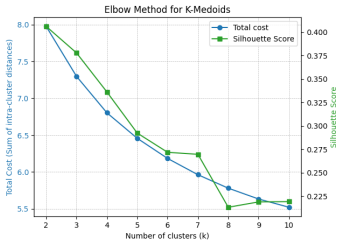
The poorer reconstruction of the breast cancer dataset may reflect differences between single-cell and bulk sequencing, resulting in a different set of underlying signatures. Alternatively, low mutation counts per cell could cause greater stochastic variability of mutational profiles, which may appear novel. However, with more mutations, they may converge towards known processes.



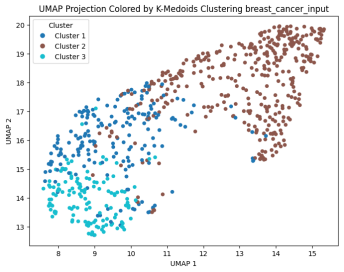
(a) Breast cancer: COSMIC signature occurrence in the samples, the x-axis shows the number of cells with at least one mutation attributed to this signature. All signatures occur in similar proportions throughout the samples.



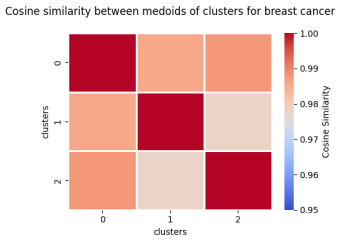
(b) Neck cancer: same format as 1a, similar behavior is observed here, with no large difference in proportions of various processes.



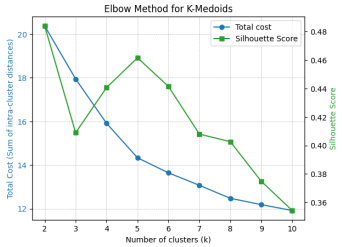
(c) Breast cancer: k-medoids cluster quality assessment based on silhouette score (green) and clustering cost (blue). The silhouette score starts to drop slightly more after three clusters, thus three clusters are chosen.



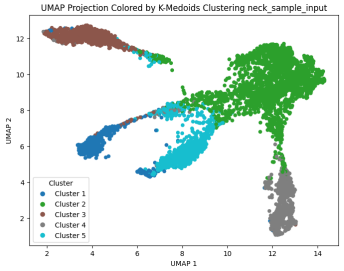
(d) Breast cancer: UMAP visualization of the mutational profiles based on cosine similarity. Each point represents a single cell, colored by its cluster assignment, revealing no well-separated structures.



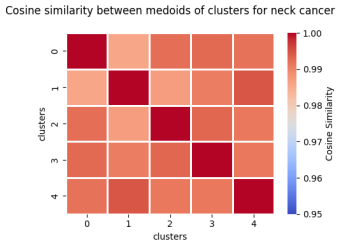
(e) Breast cancer: heatmap of cosine similarity between cluster centroids derived from the single-cell mutational profiles. High similarities are observed, all above 0.97, suggesting poor separation of the mutational profiles.



(f) Neck cancer: showing same plot as 1c. The peak in silhouette score and a slight elbow in total cost at five clusters results in five clusters being chosen.



(g) Neck cancer: same format as 1d. Distinct visual clusters are observed in the UMAP space. Illustrating distinct mutational processes present in the samples.



(h) Neck cancer: same plot as 1e. Even though clear clusters are illustrated in 1g, high similarity is observed between the medoids, suggesting low diversity of mutational processes present in the samples.

Figure 1: Visualization of mutational profile diversity and clustering for breast and neck cancer datasets. Figures show COSMIC sample reconstruction, UMAP projections, and cosine similarity heatmaps based on k-medoids clustering of mutational profiles.

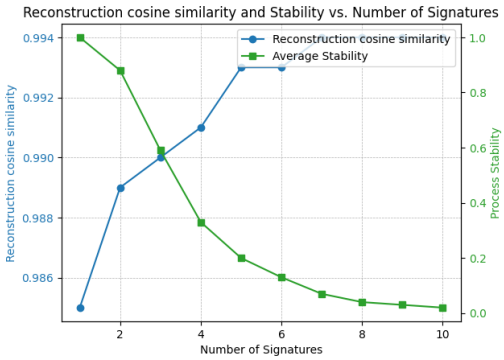
Table 1: Summary statistics of cosine similarities between COSMIC-based reconstructions and original mutational profiles.

Statistic	Breast cancer	Neck cancer
Mean	0.852	0.927
Minimum	0.814	0.896
Maximum	0.885	0.953
Average number of mutations	272	614
Sample count	90	150
Poorly explained sample count (< 0.85)	42	0

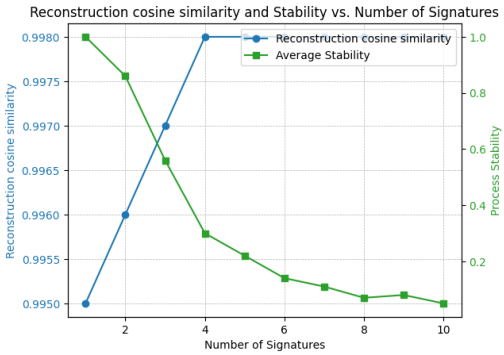
De novo signatures

Figure 2a and 2b display the reconstruction similarity and stability with increasing numbers of extracted signatures. In both datasets, the final solution with a stability above 0.8 corresponds to two signatures, after which, a steep decline in stability is observed. The similarity between the two de novo signatures is 0.968 and 0.965 for the breast and neck cancer, respectively, indicating overlapping processes.

Reconstruction cosine similarity remains high for both samples at low signature counts. When extracting a single signature, the reconstruction similarity exceeds 0.985 for both datasets. This is much higher than reported in Table 1 for the COSMIC reconstruction. This, in combination with the immediate steep stability decline, could imply that the extraction process is overfitting the data by finding more than one signature. Given the high similarity between clusters, it could be that the only difference in processes observed is due to slight variations caused by the random nature of somatic mutations rather than different underlying mutational processes.



(a) Breast cancer: Plot of average stability (green) and reconstruction similarity (blue) versus number of extracted signatures. Stability declines sharply after two signatures, resulting in the selection of two signatures.

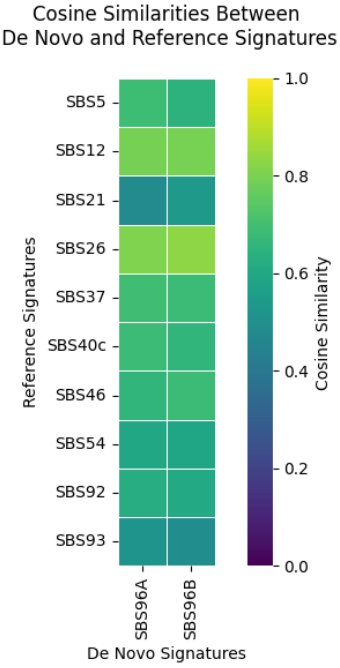


(b) Neck cancer: Similar trend as in 2a. Stability drops beyond two signatures, confirming the choice of two as the optimal count.

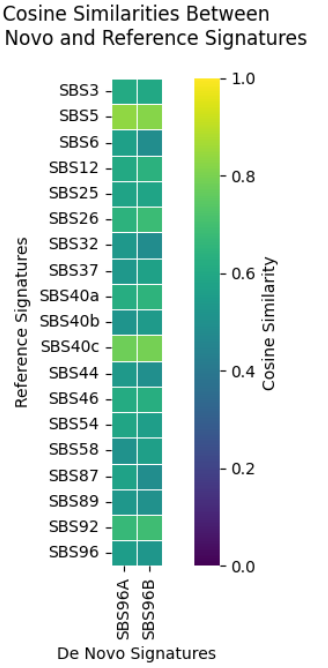
Direct comparison

To see the potential processes that correspond to the de novo signatures, Figures 4a and 4b present heatmaps of the pairwise cosine similarities between COSMIC and de novo signatures. As the de novo signatures within both datasets are closely related, the 2 signatures show high similarity to the same COSMIC signatures.

Tables 2 and 3 show the direct one-to-one matching of the de novo signatures with COSMIC reference signatures. For both datasets, all direct matches fall below the set threshold. This suggests that, when viewed in isolation, the signatures are either novel or represented by a mixture of multiple known processes.



(a) Breast cancer: Cosine similarity between the two de novo signatures and COSMIC references. No match exceeds the threshold of 0.85, indicating no one-to-one correspondence.



(b) Neck cancer: Cosine similarity between the two de novo signatures and COSMIC references. Same as for 3a, no matches exceed the threshold of 0.85.

Figure 3: Heatmaps show cosine similarity between each de novo signature and COSMIC reference signatures, helping to assess whether the extracted signatures represent known processes. Only COSMIC signatures with similarity > 0.5 to any de novo signature are shown.

Cosmic decomposition

Figures 4a and 4b show the decomposition of the de novo signatures into linear combinations of known COSMIC signatures. In the breast cancer dataset, both de novo signatures consist of the same COSMIC signatures, with minor variations in proportions. In contrast, the neck cancer de novo signatures are comprised of distinct COSMIC processes.

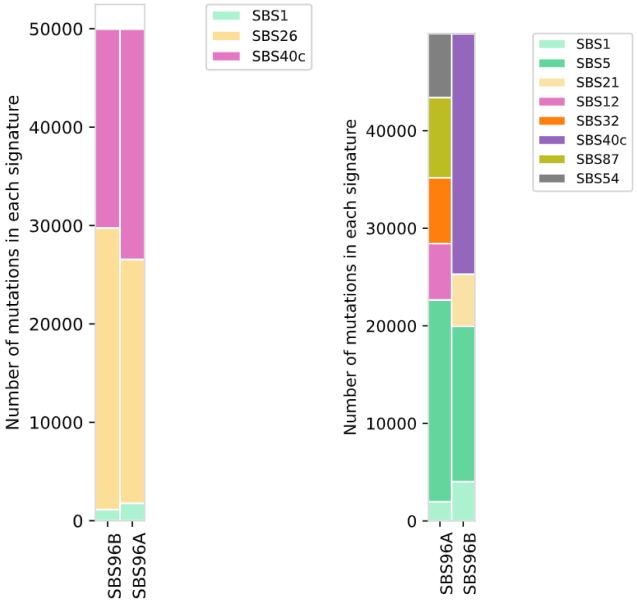
Despite the lack of strong direct matches, the COSMIC-based reconstruction of the de novo signatures yields high cosine similarities, all exceeding 0.85, as can be seen in Tables 2 and 3. This suggests that the de novo signatures likely reflect mixtures of known mutational signatures rather than novel processes.

Table 2: Comparison of de novo signatures from the breast cancer dataset to COSMIC reference signaturess.

Statistic	SBS96A	SBS96B
Direct match	SBS12	SBS26
Direct matching cosine similarity	0.790	0.831
COSMIC decomposition signatures	SBS1, SBS26, SBS40c	SBS1, SBS26, SBS40c
Decomposition cosine similarity	0.859	0.865
Signature type	Composite	Composite

Table 3: Comparison of de novo signatures from the neck cancer dataset to COSMIC reference signatures.

Statistic	SBS96A	SBS96B
Direct match	SBS5	SBS40c
Direct matching cosine similarity	0.833	0.790
COSMIC decomposition signatures	SBS1, SBS5, SBS12, SBS32, SBS54, SBS87	SBS1, SBS5, SBS21, SBS40c
Decomposition cosine similarity	0.946	0.888
Signature type	Composite	Composite



(a) Breast cancer: both de novo signatures are reconstructed from the same set of COSMIC reference signatures. Reconstruction similarities are above 0.85, indicating a likely composite origin for the signatures.

(b) Neck cancer: contrary to 4a, both signatures are recomposed of a different set of COSMIC signatures. Where SBS1 and SBS5 occur in both signatures. SBS96B has a significantly lower reconstruction similarity at 0.888 compared to 0.946 for SBS96A.

Figure 4: Stacked bar plots illustrate how each de novo signature can be expressed as a linear combination of known COSMIC signatures using non-negative least squares fitting.

conclusion

This study proposed and applied a systematic method to compare single-cell-derived de novo signatures to COSMIC reference signatures, using two single-cell sequenced datasets, breast and neck cancer. Each de novo signature was classified as either a known COSMIC signature, a composite of multiple COSMIC signatures, or potentially a novel signature.

Analysis of the mutational profiles of the data shows low cosine distances between cell clusters. For both datasets, two stable signatures were consistently identified. These de novo signatures show high similarity within each dataset (cosine similarity > 0.96). This suggests that, despite somatic heterogeneity at the mutational level, the driving mutational processes may be largely uniform within each tumor. This is further supported by the high reconstruction similarity with the extraction of a single de novo signature.

Direct one-to-one matching of de novo signatures to individual COSMIC signatures yielded no strong alignments. However, decomposition into combinations of COSMIC signatures resulted in high reconstruction similarities (all > 0.85), indicating that the signatures are likely composed of multiple mutational mechanisms rather than entirely novel ones. Whether this pattern is shared among other single-cell samples remains unclear and warrants further study.

These results demonstrate both the promises and current challenges of single-cell data for mutational signature analysis. While de novo extraction is feasible, the low mutation counts per cell may introduce stochastic variation, potentially inflating apparent heterogeneity. In the neck cancer dataset, mutational profiles appeared less distinct despite a larger sample size, likely due to higher mutation counts per sample. However, as this study is limited to two datasets, broader validation across a wider range of tumors is essential to assess the generalization of these findings and to gain deeper insights into intra-tumor mutational process diversity.

Future work

To explore the limits of single-cell de novo mutational signature extraction, further research should include simulation studies. Simulated single-cell datasets allow precise control over the mutational processes, the number of mutations per cell, and the total sample count. Varying these parameters enables systematic benchmarking of de novo signature extraction. This way, the limitations and sensitivity of the extraction and comparison methods used in this study can be assessed.

Future studies should also perform de novo extraction across multiple single-cell datasets. Ideally, tumors would also have corresponding bulk sequencing data available. This would enable a direct comparison between the signatures extracted from both data types and would help to clarify whether single-cell data provides additional resolution or introduces greater variability. These samples can also be analyzed using the same methodology as this paper to validate and expand upon the conclusions of this study.

This study has been conducted based on the `SigProfiler` tools for signature extraction and comparison. However, single-cell data presents unique challenges, such as lower mutation counts and risk of overfitting. Other tools may be more suited to these conditions, and additional methods could be evaluated. Alternatively, new tools, specifically designed for the challenges of single-cell data, could be developed.

Responsible Research

All data used in this study are anonymized and publicly available. No personally identifiable information was accessed or processed. The source for the breast cancer is: https://www.10xgenomics.com/datasets/750-sorted-cells-from-human-invasive-ductal-carcinoma-3-lt-v-3-1-3-1-low-6-0-0, and for the Laryngeal Squamous Cell Carcinoma (neck cancer): https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE206332. Software tools and parameter settings are fully described in the Methodology section to ensure reproducibility. The analysis was performed using Python 3.12.4 with the following package versions:

- `SigProfilerAssignment` v0.2.1 - for fitting known COSMIC signatures to samples and de novo signatures.
- `SigProfilerMatrixGenerator` v1.3.3 - for generating of the mutational profiles.

- `SigProfilerExtractor` v1.2.0 - for de novo extraction of mutational signatures.
- `umap-learn` v0.5.7 - for umap dimensionality reduction.
- `networkx` v3.4.2 - for implementation of the minimum cost maximum flow algorithm.
- `pyclustering` v0.10.1.2 - for k-medoids clustering.

All parameters, thresholds, and decision criteria (e.g., stability cutoff, similarity thresholds) are specified in the text, allowing others to replicate the pipeline.

Due to time constraints, only two datasets were analyzed. These datasets may not be representative of tumors in other patients. Therefore, the results are not generalized beyond these specific datasets. As no simulated data was used, there is no ground truth to the data, making it unclear whether the various signatures found are caused by distinct combinations of COSMIC processes or are the results of overfitting on the data. Future work involving additional, more diverse, or simulated datasets is necessary to draw broader conclusions about the differences between single-cell-derived de novo signatures and COSMIC reference signatures.

References

Abbasi, Ammal and Ludmil B Alexandrov (2021). "Significance and limitations of the use of next-generation sequencing technologies for detecting mutational signatures". In: *DNA repair* 107, p. 103200.

Alexandrov, Ludmil B, Jaegil Kim, et al. (2020). "The repertoire of mutational signatures in human cancer". In: *Nature* 578.7793, pp. 94–101.

Alexandrov, Ludmil B, Serena Nik-Zainal, et al. (2013). "Deciphering signatures of mutational processes operative in human cancer". In: *Cell reports* 3.1, pp. 246–259.

Becht, Etienne et al. (2019). "Dimensionality reduction for visualizing single-cell data using UMAP". In: *Nature biotechnology* 37.1, pp. 38–44.

Díaz-Gay, Marcos et al. (2023). "Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment". In: *Bioinformatics* 39.12, btad756.

Islam, SM Ashiqul et al. (2022). "Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor". In: *Cell genomics* 2.11.

Lee, Daniel D and H Sebastian Seung (1999). "Learning the parts of objects by non-negative matrix factorization". In: *nature* 401.6755, pp. 788–791.

"Pan-cancer analysis of whole genomes" (2020). In: *Nature* 578.7793, pp. 82–93.

Sondka, Zbyslaw et al. (2024). "COSMIC: a curated database of somatic variants and clinical data for cancer". In: *Nucleic Acids Research* 52.D1, pp. D1210–D1217.

Stratton, Michael R, Peter J Campbell, and P Andrew Futreal (2009). "The cancer genome". In: *Nature* 458.7239, pp. 719–724.