# Multivariate Correlation of Mutational Signature Exposures and Gene Expression in Single-Cell Breast Cancer

**Traian Dobrin**[1]

**Supervisor(s): Sara Costa, Ivan Stresec**
**Responsible Professor: Joana Gonçalves**

[1]EEMCS, Delft University of Technology, The Netherlands

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

**Abstract**

Understanding the relationship between mutational processes and gene expression patterns is essential for gaining insights into tumor heterogeneity. In this study, we analyze single-cell RNA sequencing data from a breast cancer tumor to investigate associations between mutational signature exposures and gene expression profiles. We propose a scoring method that integrates principal component loadings, canonical correlation analysis (CCA) loadings, and signature contributions to quantify gene-signature associations. Enrichment analysis of the top-ranking genes reveals consistent involvement of extracellular matrix (ECM) receptor interaction, focal adhesion, and immune-related pathways across multiple mutational signatures. These findings suggest that different mutational processes converge on pathways involved in cell adhesion, invasion, and immune modulation. Our approach demonstrates the utility of multivariate statistical methods combined with enrichment analysis to explore the transcriptional consequences of mutational processes in cancer at the single-cell level.

# 1 Introduction

Cancer is a complex and heterogeneous disease, characterized by a vast diversity of genetic alterations and cellular behaviors even within a single tumor. Understanding this heterogeneity is crucial, as it directly influences treatment response, disease progression, and patient outcomes. One emerging approach to characterizing this diversity is the study of mutational signatures-distinctive patterns of somatic mutations that reflect the underlying mutational processes active in a tumor. These processes include environmental exposures (e.g., tobacco smoke or UV light), defects in DNA repair mechanisms, and enzymatic activities such as those driven by APOBEC proteins. Over time, these processes leave characteristic "footprints" in the genome, known as mutational signatures, which can be identified through computational analysis.

Previous studies have explored mutational signatures to infer etiological factors in cancer development [4] and have separately used gene expression data for tumor classification and prognosis. Correlations between mutational signature exposures and biological processes have already been discovered [5], and specific genes are already believed to have their activity impacted by mutational processes [15]. As an example, tobacco smoking has been shown to activate immune responses [2]. However, the processes that cause many of these mutational signatures are still unknown. Mutagenic processes arise from disruptions in normal molecular pathways, and vice-versa they can disrupt normal cell activities. Consequently, uncovering the relationship between these processes and the underlying molecular mechanisms is essential for understanding their origins and biological impact. A great way of monitoring a cell's behavior is through its gene expression data. Therefore, uncovering the relation between mutational processes and the affected cells' gene expression is crucial for understanding their role in carcinogenesis.

This study aims to study the relationships between mutational signature exposure and gene expression for single cell data. More specifically, we start off by employing multivariate correlation analysis through CCA (Canonical Component Analysis) [14]. This enables us to uncover coordinated changes and potential regulatory programs associated with specific mutational processes.

By analyzing the gene groups that are most strongly associated with individual signatures, we aim to gain insights into the biological pathways potentially influenced by different mutational mechanisms. To contextualize these associations, we use pathway enrichment analysis, highlighting functional themes and cellular processes that may be disrupted in the tumor microenvironment.

Overall, this approach contributes to a deeper understanding of how genomic alterations interact with gene expression at single-cell resolution, offering potential insights into the functional consequences of mutational processes in breast cancer.

The remainder of this paper is structured as follows. In Section 2, we describe the necessary preliminaries for preparing and processing the data used in this study. Section 3 details our multivariate analytical approach, including the application of Canonical Correlation Analysis (CCA), the development of a custom scoring technique, and gene set enrichment analysis. Section 4 presents the main results and discusses the biological implications of our findings in the context of breast cancer. In Section 5, we summarize our conclusions and outline potential directions for future work. Finally, Section 6 addresses considerations of responsible research, including ethical aspects and reproducibility.

## 2 Preliminaries

A foundational concept in understanding cancer evolution lies in the idea of *mutational signatures*-distinct patterns of somatic mutations imprinted on the cancer genome by various mutagenic processes. One of the most influential contributions to this field was made by Alexandrov et al., who introduced a computational framework based on Non-negative Matrix Factorization (NMF) to uncover these signatures from large-scale tumor sequencing data [4].

In Alexandrov's approach, the observed mutational catalog of a tumor cohort is represented as a non-negative matrix $\mathbf{V} \in R^{m \times n}$, where $m$ is the number of mutation types (e.g., 96 trinucleotide contexts), and $n$ is the number of tumor samples. NMF factorizes this matrix into two non-negative matrices: $\mathbf{W} \in R^{m \times k}$, which contains the mutational signatures, and $\mathbf{H} \in R^{k \times n}$, which represents the exposure of each sample to each signature, such that $\mathbf{V} \approx \mathbf{WH}$. Here, $k$ denotes the number of extracted mutational signatures. The columns of $\mathbf{W}$ can be interpreted as probabilistic profiles over mutation types, and the rows of $\mathbf{H}$ describe how strongly each signature is expressed in a given tumor.

These mutational signatures provide a molecular record of past mutational processes, such as exposure to UV light, tobacco smoke, or deficiencies in DNA repair mechanisms. Signatures like SBS1, associated with age-related accumulation of C>T mutations at CpG sites, or SBS5, a common signature with unknown etiology, have been consistently observed across cancer types. Alexandrov's NMF-based framework laid the groundwork for cataloging these signatures and inspired further efforts to refine their biological interpretation.

In the context of single-cell data, mutational signature exposures can vary between cells, potentially reflecting intratumoral heterogeneity in DNA damage and repair processes. Understanding how these exposures relate to transcriptional programs is crucial for linking somatic evolution to phenotypic diversity in cancer. In this study, we build on Alexandrov's model by treating the signature exposure matrix (analogous to $\mathbf{H}$) as the starting point for multivariate correlation analysis, seeking to uncover systematic associations between mutational processes and gene expression profiles at the single-cell level.

## 3 Materials and methods

This study investigates the relationship between mutational signature exposure and gene expression in single cells derived from a single breast cancer tumor. The analysis is conducted on a multi-modal dataset containing both gene expression profiles and mutational signature exposures for single cells.

## 3.1 Dataset

The dataset used in this study includes single-cell RNA sequencing (scRNA-seq) data and mutational signature exposure estimates for individual cells. All cells originate from the same breast tumor sample, enabling analysis of intra-tumor heterogeneity. The gene expression data is derived from a dataset of 750 sorted cells from a human invasive ductal carcinoma sample provided by 10x Genomics [1]. Gene expression captures the transcriptomic activity of each cell, while mutational signature exposures quantify the contribution of different mutagenic processes-such as SBS1, SBS5, SBS26, and others-to each cell's mutation profile. The mutational signature exposures were inferred using the Non-negative Matrix Factorization (NMF) framework described by Alexandrov et al. [4]. After performing this step, the datasets contains entries for a total of 87 different mutational signatures, 36601 genes and 687 cells. Each cell has a certain exposure to each signature and a certain value for each gene, representing the expression of that gene.

## 3.2 Preprocessing

The initial dataset consists of single-cell gene expression profiles and associated mutational signature exposures for cells from a single breast cancer tumor. Before proceeding with any analysis, a preprocessing step is applied to select informative features.

First, mutational signatures with zero exposure across all cells are discarded. Since these signatures do not vary and therefore cannot provide any meaningful signal for downstream correlation or classification tasks, they are excluded from the analysis. After the removal of the absent signatures, we are left with 6 others: SBS1, SBS5, SBS12, SBS26, SBS40c and SBS54. It is worth noting that, although all these signatures are detected in at least a few cells, SBS26 and SBS40c show the highest overall exposure. In contrast, SBS54 displays very low total exposure. Previous reports have suggested that SBS54 may represent a sequencing artifact or contamination with germline variants [6], which could explain its limited presence in the data.

In order to stabilize and reduce the impact of outliers or genes with consistently high expression values, the gene expression data is transformed using the log transform.

Basic filtering is also applied to the gene expression data to remove genes with low variance or no expression across all cells, reducing noise and computational complexity in the analyses that follow.

## 3.3 Statistical Association Analysis

To investigate the relationship between mutational signature exposures and gene expression, we first apply *Canonical Correlation Analysis* (CCA) [14]. CCA identifies pairs of canonical variables-linear combinations of features from each dataset-that are maximally correlated. In our context, this reveals coordinated patterns between the transcriptomic landscape and mutational processes across single cells. Unlike univariate approaches, which examine one gene or one signature at a time, multivariate statistical techniques such as CCA are well-suited to uncover high-dimensional relationships where groups of genes may jointly associate with multiple mutational signatures. This holistic view allows us to account for dependencies and covariation within and across modalities, offering a more integrated understanding of how mutational processes shape gene expression programs.

However, gene expression data is inherently high-dimensional, typically with far more genes than mutational signatures. This imbalance can lead to overfitting in CCA. To address this, we first apply *Principal Component Analysis* (PCA) to the gene expression matrix to reduce its dimensionality while preserving most of its variance. The number of principal components retained is determined using the *elbow method*, which identifies the point at which adding more components yields diminishing returns in explained variance. In this study, we retained enough components to account for 90% of the total variance, which resulted in 158 principal components being selected for downstream analysis. Both PCA and CCA were performed using the `scikit-learn` library [12].

After dimensionality reduction, each cell is represented by its mutational signature exposure profile and its corresponding principal component scores from the gene expression data. We then apply CCA to these two sets of variables. After applying CCA, we obtained pair of canonical components, representing linear combinations of the exposure data and the principal components that are maximally correlated. For each of these pairs, by examining the coefficients (loadings) within these components, we can assess the contribution of individual mutational signatures or gene expression components to each correlated pattern, providing insights into how specific mutational signatures are linked to transcriptional programs.

## 3.4 Differential scoring for gene-signature pairs

To further highlight genes that are strongly associated with specific mutational signatures, we introduce a scoring approach that integrates the structure of the canonical components. The goal of this approach is to identify pairs of genes and mutational signatures that correlate. For each pair of the canonical components obtained from CCA, we analyze which *mutational signatures* and *genes* contribute most by examining the corresponding *canonical loadings*, as explained before.

Since PCA was applied to the gene expression matrix before CCA, the gene-side canonical components exist in a reduced space. Therefore, to trace contributions back to individual genes, we also consider each gene's contribution to each principal component. In order to differentiate between the genes contributions to correations, we propose the following approach: We assign a score $S_{g,s}$ to each gene $g$ and mutational signature $s$ pair, initialized to zero. Then, for each pair of canonical components, we examine the loadings on both sides. The magnitude of these loadings reflects how strongly each principal component or mutational signature contributes to the canonical correlation. To map this back to individual genes, we proceed as follows: For each canonical pair, we iterate over all principal components and mutational signatures, and for each gene-signature pair, we increment its score by the product of:

1. the gene's contribution to a selected principal component (PCA loading),
2. the principal component's contribution to the gene-side canonical component (CCA loading),
3. the mutational signature's loading on the corresponding signature-side canonical component, and
4. the correlation between the two canonical components under consideration.

A visual representation of the contribution for one mutational signature and one Principal Component can be seen in Figure 1.

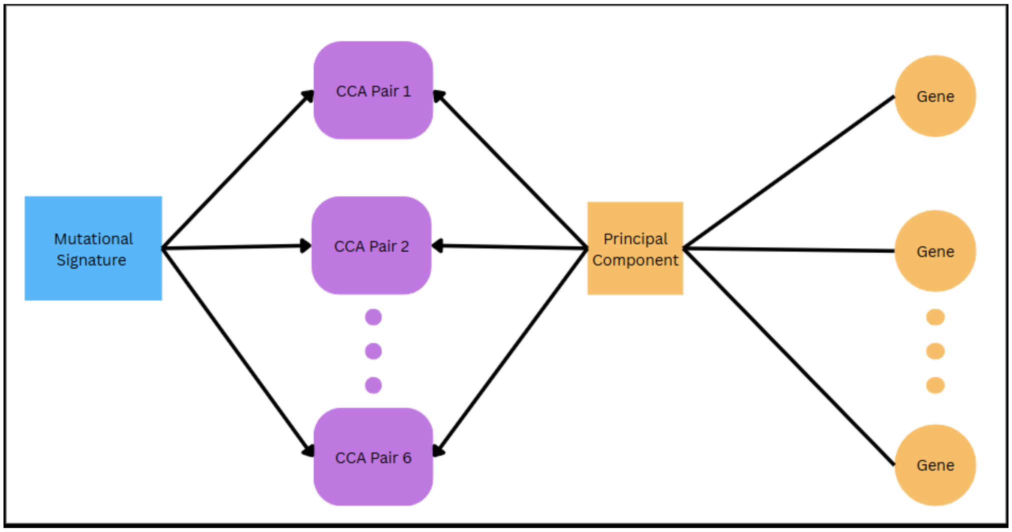

Figure 1: Contribution to the CCA of one Mutational Signature and one Principal Component.

Specifically, for each canonical component pair $k$, we update the score as:

$$S_{g,s} += \rho_k \cdot \left( \sum_p P_{g,p} \cdot U_{p,k} \right) \cdot V_{s,k}$$

where:

- $\rho_k$ is the absolute value of the canonical correlation coefficient for canonical component pair $k$,
- $P_{g,p}$ is the absolute value of the loading of gene $g$ on principal component $p$ (PCA loading),
- $U_{p,k}$ is the absolute value of the loading of principal component $p$ on gene-side canonical component $k$ (CCA loading),
- $V_{s,k}$ is the absolute value of the loading of signature $s$ on signature-side canonical component $k$ (CCA loading).

Hence, the final score of each pair would be:

$$S_{g,s} = \sum_k \left( \rho_k \cdot \left( \sum_p P_{g,p} \cdot U_{p,k} \right) \cdot V_{s,k} \right)$$

This composite score quantifies the strength of association between a gene and a mutational signature, aggregating contributions across both latent spaces. Further analysis of these scores' formula is described in the following sections.

### 3.5 Pathway and Enrichment Analysis

To interpret the biological relevance of genes associated with mutational signatures, we perform pathway and enrichment analysis using gene set enrichment tools. Specifically, genes are ranked based on their association scores with each mutational signature, as defined in the previous section. These ranked gene lists were analyzed using the GSEApy library to perform pre-ranked Gene Set Enrichment Analysis (GSEA) [13], leveraging curated gene sets obtained from Enrichr [10], specifically the *KEGG 2021 Human* [8] collection.

The analysis identifies pathways that are significantly overrepresented among genes highly correlated with specific mutational processes. This enables the mapping of distinct mutational signatures to relevant transcriptional programs and cellular processes, potentially revealing the functional consequences of underlying mutational mechanisms. To ensure statistical robustness, we apply multiple testing correction using the False Discovery Rate (FDR) and Family-Wise Error Rate (FWER) procedures, and consider pathways significant when both values are smaller than 0.01, a widely used threshold in enrichment analysis. The most enriched pathways for each signature are reported and discussed in the results section.

Together, these methods form a pipeline for linking mutational processes to transcriptomic consequences at the single-cell level, providing a basis for understanding how genomic alterations shape cellular behavior in cancer.

## 4 Results and Discussion

### 4.1 Dimensionality Reduction and statistical correlations

To reduce the high dimensionality of the gene expression data and mitigate overfitting in subsequent analyses, Principal Component Analysis (PCA) was applied. Using the elbow method to select the optimal number of components, we retained 158 principal components, which together explain over 90% of the total variance in gene expression. This dimensionality reduction balances data complexity and interpretability while preserving most of the biological variability (see Figure 2).

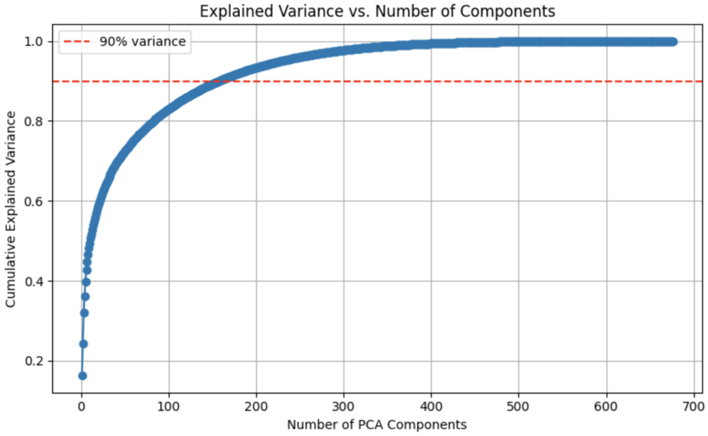

Figure 2: Elbow plot used to determine the optimal number of principal components in gene expression data. The chosen cutoff retains over 90% of variance with 158 components.

Our analysis focused on six mutational signatures: SBS1, SBS5, SBS40c, SBS26, SBS12, and SBS54. Since there are six mutational signatures, the CCA produced six pairs of canonical components linking the reduced gene expression data and the mutational signature exposures across single cells. The correlations of these pairs can be seen in Figure 4.

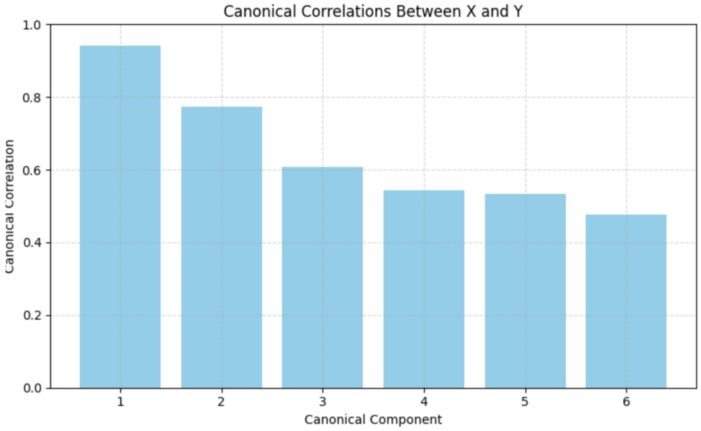

Figure 3: Plot of canonical correlation analysis (CCA) showing the correlations between gene expression and mutational signature exposures.

Figure 4 illustrates the relative contributions of each mutational signature to the canonical correlation components, highlighting how different signatures are associated with the multivariate gene expression patterns captured by CCA. These distinct signature-component relationships underscore the capacity of CCA to disentangle complex, multivariate associations between mutational exposures and gene expression. Understanding these patterns provides insights into how mutational processes may influence tumor heterogeneity and gene regulation at the single-cell level. Although not explored in this study, examining which signatures contribute simultaneously to the same components could provide insights into potential relationships or co-occurrence patterns among mutational processes. Such analyses could be extended by comparing mutational profiles or investigating biological overlaps.

## 4.2 Scoring formula for each gene-signature pair

Using a novel scoring formula that combines contributions from gene PCA loadings, CCA loadings, and the correlation of canonical components, we derived association scores for each gene-mutational signature pair. This approach enabled us to quantify the strength and direction of the relationship between individual genes and mutational signatures.

The scoring formula reflects a decomposition of the multivariate association between gene expression and mutational signature exposures into interpretable components.

First, dimensionality reduction via PCA expresses each gene $g$ as a linear combination of principal components, with weights $P_{g,p}$ indicating the gene's contribution to each PC $p$. Since Canonical Correlation Analysis (CCA) operates on the reduced PC space, the gene-side canonical component $k$ is itself a weighted combination of PCs, with weights $U_{p,k}$. Therefore, the total contribution of gene $g$ to canonical component $k$ can be written as $\sum_p P_{g,p} \cdot U_{p,k}$, capturing how much each gene contributed to each principal component and

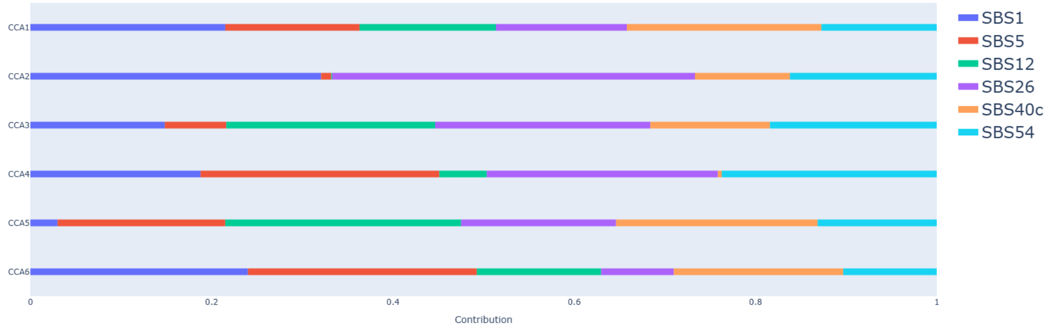

Figure 4: Magnitudes of each signature's contributions to each canonical component(considering absolute values of contributions). CCA components are numbered 1-6, with 1 being the highest correlated with the gene expression data and 6 the lowest correlated.

how much each specific principal component contributed to the correlation of the canonical pair.

Similarly, we also want to account for the contribution the signatures have on the canonical pair, since we are considering gene-signature pairs. This can be quantified by the canonical loading of that signature in canonical component $k$.

Rather than using raw loadings, we take their absolute values to emphasize the magnitude of association, disregarding directionality, which can be investigated separately if relevant. This approach is motivated by the fact that canonical components across different pairs are uncorrelated [14], meaning each pair captures distinct and potentially diverse patterns of association. Including signed loadings in the aggregation could lead to cancellation effects, where strong but oppositely signed contributions from different canonical components negate each other, yielding artificially low gene-signature scores. By focusing on absolute loadings, our scoring method robustly reflects the overall strength of association between genes and mutational signatures across all canonical components, enabling more reliable identification of biologically relevant relationships, but losing the direction of correlation.

To further characterize the variability of the scores, Figure 5 shows a boxplot of the distribution of the scores of the genes for each signature. The scores for SBS54 are the lowest, while the scores for signatures like SBS26 or SBS40c get the highest values. This pattern likely reflects the relative presence of these signatures in our dataset: SBS54 has a very low presence, limiting the strength and number of detectable gene associations and introducing more noise, not allowing genes to acquire particularly high scores. On the other hand, SBS26 and SBS40c are more prevalent, resulting in stronger gene-signature associations.

While the scoring formula currently equally integrates contributions across canonical components, it is flexible and can be adjusted in future work. For instance, one could introduce weighting schemes that give greater emphasis to canonical components with higher canonical correlations, under the assumption that this would reflect more accurate scores. We chose this unweighted approach for the formula in order to avoid introducing subjective bias into the scoring process. By doing so, we aim to provide an unbiased initial map of gene-signature associations, which can serve as a foundation for more targeted analyses in

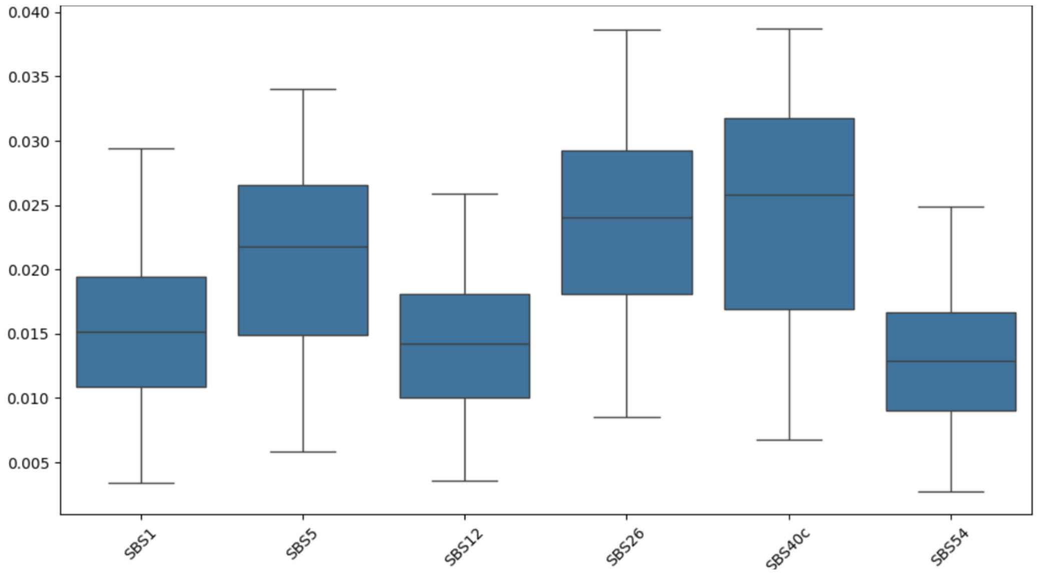

Figure 5: Boxplot of gene scores per signature

subsequent studies.

## 4.3 Enrichment Analysis of Signature-Associated Genes

To interpret the biological significance of the gene sets associated with each mutational signature, we performed gene set enrichment analysis (GSEA) against the KEGG 2021 Human pathways database. Genes were ranked according to their association scores, normalized by their total score sum, and subjected to pre-ranked GSEA.

There are however certain limitations that apply to the GSEA framework. Enrichment results can be highly dependent on the underlying pathways dataset that was used. Depending on the dataset, specific biological pathways may be underrepresented or absent. Furthermore, despite testing corrections being used, they may not completely eliminate the risk of false positives, especially when dealing with correlated gene sets. However, GSEA remains a valuable exploratory tool to generate hypotheses and highlight possible pathways for biological investigation.

Our results revealed distinct pathway enrichments for each signature, highlighting potential functional consequences of mutational processes on cellular transcriptional programs. Table 1 shows the most relevant pathways for the signatures. Due to the very low presence of SBS54 and the consequently high p-values observed, we excluded this signature from the final analyses.

The pathways are ranked according to the FDR p-values, ascendingly. In particular, the strong enrichment of ECM-receptor interaction and focal adhesion pathways (seen for all 5 of the signatures) is highly consistent with prior analyses of breast cancer. Naba et al. (2014) systematically profiled extracellular matrix (ECM) proteins across breast tumors and demonstrated that distinct ECM signatures are associated with tumor progression and metastasis [11]. Their analysis identified multiple ECM components (e.g., LTBP3, SNED1, COL10A1) whose dysregulation contributes to invasive potential. The fact that multiple

| **Signature** | **Term** |
|---|---|
| SBS1 | ECM-receptor interaction |
| SBS5 | Focal adhesion |
| | Tuberculosis |
| | Systemic lupus erythematosus |
| SBS12 | ECM-receptor interaction |
| | Focal adhesion |
| | Systemic lupus erythematosus |
| SBS26 | Focal adhesion |
| | ECM-receptor interaction |
| | Protein digestion and absorption |
| SBS40c | Phagosome |
| | Systemic lupus erythematosus |
| | Focal adhesion |

Table 1: Top 3 enriched terms per mutational signature

mutational signatures in our data point toward ECM and adhesion pathways suggests that the accumulation of somatic mutations-whether due to aging (SBS1, SBS5), defective repair (SBS26), or other processes (SBS40c)-can impact key components of the tumor microenvironment, potentially contributing to invasive and metastatic behavior. Furthermore, Yeh et al. (2018) demonstrated that overexpansion of several ECM-receptor interaction genes (including HMMR, SDC1, and ITGA5) is correlated with poor survival in breast ductal carcinoma [16], supporting the idea that mutations in ECM and adhesion pathways are central features of breast cancer progression.

In SBS1, a signature correlated with age [3], we observed enrichment in the ECM-receptor interaction pathway. This is consistent with the notion that age-related accumulation of somatic mutations may impact genes regulating cell-matrix interactions Previous studies have demonstrated that ECM remodeling is a key feature of breast tumor progression, with specific ECM components implicated in metastasis and poor clinical outcome [11].

SBS5 is a ubiquitous, clock-like signature whose underlying mutational mechanism is still incompletely understood [3]. The repeated enrichment of immune-related pathways (tuberculosis, SLE) may reflect chronic inflammatory processes in the tumor microenvironment, while the focal adhesion pathway again underscores the central role of cell-ECM interactions in tumor progression [16].

Although SBS12's etiology remains less well-characterized, recent studies suggest links to defective DNA repair processes, potentially involving mismatch repair or transcription-coupled repair [3]. The enrichment of the SLE pathway may reflect immune dysregulation, consistent with the known infiltration of immune cells in the tumor environment [7], correlatinig this signature with such processes.

SBS26, associated with mismatch repair deficiency [3], showed enrichment for focal adhesion, ECM-receptor interaction, and protein digestion pathways. These results are also consistent with increased ECM remodeling and altered cell adhesion, which facilitate invasion and metastasis in mismatch repair-deficient tumors [9].

SBS40c, of uncertain etiology, was enriched for phagosome, systemic lupus erythematosus, and focal adhesion. These pathways suggest possible involvement of immune processes and tumor-immune interactions [7], which are known to influence tumor progression and therapeutic response in breast cancer.

Together, these results demonstrate the power of multivariate correlation analyses combined with enrichment testing to uncover coordinated relationships between mutational processes and gene expression at the single-cell level.

## 5 Conclusions and Future Work

This study investigated the relationship between mutational signature exposures and gene expression profiles in single-cell RNA sequencing data from a breast cancer tumor. The central research question was: *(1) Can specific gene expression patterns be linked to mutational processes?*.

To address this, we applied dimensionality reduction via PCA, followed by Canonical Correlation Analysis (CCA) to identify multivariate associations between gene expression and mutational signatures. We introduced a novel gene-scoring approach that aggregates PCA loadings, CCA loadings, and canonical correlations, enabling the quantification of gene contributions to each mutational signature. This score served as a ranking metric for gene set enrichment analysis (GSEA), revealing multiple biologically relevant pathways across the identified signatures.

Our results highlighted enrichment of ECM-receptor interaction and focal adhesion pathways across multiple signatures, aligning with prior evidence of their role in breast cancer progression and metastasis. Additionally, immune-related pathways such as systemic lupus erythematosus (SLE), tuberculosis, and phagosome were enriched for signatures like SBS5, SBS12, and SBS40c, suggesting a link between mutational processes and tumor-immune microenvironment interactions.

The scoring approach proved useful in translating multivariate CCA results into interpretable gene-level associations, but has limitations. It currently disregards the direction of associations and assumes equal contribution of all canonical components. Future work could refine the scoring by introducing component weighting or polarity to better capture the biological complexity. Furthermore, validation across larger datasets and tumor types, as well as integration of complementary omics data, may strengthen the generalizability of these findings.

Overall, this work provides a foundation for studying how mutational processes shape tumor cell identity, with potential applications in cancer diagnosis, prognosis, and therapy development.

## 6 Responsible Research

This research was conducted entirely on publicly available, deidentified breast cancer single-cell RNA sequencing data. No new data collection or patient interaction was performed, ensuring full compliance with ethical standards related to human subjects research, privacy, and informed consent, as originally managed by the data generators. Consequently, no ethical risks to participants arise directly from this work.

We recognize that working with real biological data carries the risk of overinterpretation. Although we applied careful statistical controls, including dimensionality reduction and multiple testing corrections, the results remain exploratory and should not be interpreted as conclusive evidence of causal biological mechanisms without further experimental validation. In particular, the novel gene-signature scoring approach proposed here is an initial methodological contribution that could inform future work, but it may be sensitive to parameter choices and eventual changes in the importance given to each term in the formula.

To promote reproducibility, we employed widely used and well-documented analysis techniques such as PCA, canonical correlation analysis, and gene set enrichment analysis, which are described in sufficient detail to allow reimplementation. All methodological steps are fully described in this paper, and all datasets used are accessible to the research community.

To assist with writing and literature review, we used generative AI tools during the drafting phase to summarize related work, suggest phrasing and help formatting of the paper. Prompts used include "summarize papers on [topic]", "help me rephrase this paragraph". However, the authors conducted the analyses, methodological design, data processing, and final interpretations.

Overall, this study contributes an initial methodological framework for linking mutational signatures and gene expression at the single-cell level.

# References

[1] 10x Genomics. 750 sorted cells from human invasive ductal carcinoma (3' v3.1), 2020. Accessed: 2025-06-16.

[2] Ludmil B Alexandrov, Young Seok Ju, Kerstin Haase, Peter Van Loo, Inigo Martincorena, Serena Nik-Zainal, Yasushi Totoki, Akihiro Fujimoto, Hidewaki Nakagawa, Tatsuhiro Shibata, et al. Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312):618–622, 2016.

[3] Ludmil B. Alexandrov, Juhee Kim, Niloy Haradhvala, and et al. The repertoire of mutational signatures in human cancer. *Cell*, 182(4):812–827, 2020.

[4] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel A J R Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolò Bolli, Ãke Borg, Anne-Lise Borresen-Dale, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013.

[5] Bayarbaatar Amgalan, Damian Wojtowicz, Yoo-Ah Kim, and Teresa M Przytycka. Influence network model uncovers relations between biological processes and mutational signatures. *Genome Medicine*, 15(1):15, 2023.

[6] Catalogue Of Somatic Mutations In Cancer (COSMIC). Cosmic mutational signatures v3.3, 2024. Accessed June 2025.

[7] Wolf Hervé Fridman, Franck Pagès, Catherine Sautès-Fridman, and Jérôme Galon. The immune contexture in human tumours: impact on clinical outcome. *Nature Reviews Cancer*, 12(4):298–306, 2012.

[8] Minoru Kanehisa, Yoko Sato, Mika Kawashima, Miho Furumichi, and Mao Tanabe. Kegg: integrating viruses and cellular organisms. *Nucleic Acids Research*, 49(D1):D545–D551, 2021.

[9] Kristin Kessenbrock, Vicki Plaks, and Zena Werb. Matrix metalloproteinases: regulators of the tumor microenvironment. *Cell*, 141(1):52–67, 2010.

[10] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Natividad F Fernandez, Qiaonan Duan, Zichen Wang, Sergey Koplev, Sarah L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97, 2016.

[11] Alexandra Naba, Karl R. Clauser, John M. Lamar, Steven A. Carr, and Richard O. Hynes. Extracellular matrix signatures of human mammary carcinoma identify novel metastasis promoters. *eLife*, 3:e01308, 2014.

[12] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Mathieu Brucher, Matthieu Perrot, and Ãdouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[13] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[14] Bruce Thompson. *Canonical Correlation Analysis*. American Psychological Association, Washington, DC, 2000.

[15] Alessandra Viel, Alessandro Bruselles, Elisabetta Meccia, Mauro Fornasarig, Massimo Quaia, Vincenzo Canzonieri, Eleonora Damiana Urso, Marco Agostini, Nicola Clemente, Antonio Masi, et al. A specific mutational signature associated with dna 8-oxoguanine persistence in mutyh-defective colorectal cancer. *EBioMedicine*, 20:39–49, 2017.

[16] Ming-Hsin Yeh, Yau-Jin Tzeng, Ting-Ying Fu, et al. Extracellular matrix-receptor interaction signaling genes associated with inferior breast cancer survival. *Anticancer Research*, 38(8):4593–4605, 2018.