Deciphering Cancer Heterogeneity with Machine Learning

Signature fitting analysis on single cells in relation to pseudo-bulk data

Rotar Mircea-Raul
Supervisor(s): Joana Gonçalves, Sara Costa, Ivan Stresec
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: Rotar Mircea-Raul
Final project course: CSE3000 Research Project
Thesis committee: Joana Gonçalves, Sara Costa, Ivan Stresec, Catharine Oertel Genannt Bierbach

1 Abstract

The field of oncology has greatly benefited due to the study of mutational signatures, patterns of mutations that appear within the cancer genome. Previous research has focused its resources on utilizing various mathematical models to uncover and understand these mutational signatures by looking at the genetic information of aggregated cells, typically sequenced from a tumor biopsy, which is referred to as bulk data. However, recent developments in sequencing techniques have provided us with the possibility of investigating the genetic information at the single cell level rather than bulk. Thus, in this paper, we utilized machine learning-based tools to examine the effect of performing signature fitting at the single-cell level in relation to pseudo-bulk.

We found that single cells have a higher degree of expression by contrast to the pseudo-bulk, having the capability to identify a higher number of active mutational signatures. We also saw some single cells achieving better accuracy in the reconstruction of their mutational profile, by comparison to the pseudo-bulk. We identified that the heterogeneity across the single cells could be explained by a small number of clusters, which can potentially elucidate the active signatures found at the level of the pseudo-bulk sample. Finally, we found that some pseudo-bulk samples generated from subpopulations of cells unexpectedly deviate from the single cells which created them.

From the findings, we believe that the study of active mutational signatures at the level of single cells has the potential to enlarge our understanding of cancer by providing us a more in-depth view of this disease. However, further research should be undergone in order to either augment or refute these findings, mainly due to the limiting factor of a relatively small number of mutations characterizing our data, together with the absence of a ground-truth for the bulk data.

Keywords: mutation, mutational profile, exposure, mutational signature, signature fitting, COSMIC, single base substitutions - SBS

2 Introduction

One of the most prevalent issues in the medical sector today is that of cancer. "Cancer is a group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body" [1]. As it turns out, a systematic and generalized way of treatment has yet to be discovered, fact which could be attributed to its heterogeneous nature: cancer is not only different across patients, but also within itself, being composed of cells that had developed differences between one another throughout their life-cycle[2].

During the lifetime of each individual, various endogenous or exogenous factors may alter the genetic information of healthy cells, modifications which are referred to as somatic mutations, that drive the development of the disease. Some exogenous factors include UV-radiation from the sun, active substances like tobacco from cigarettes [3], while endogenous factors are tied to various inadequacies of the inner working of the body machinery itself, e.g. the defective nature of DNA repair processes or the APOBEC family of cytidine deaminases [4]. Studies have found that many of these processes leave behind specific patterns of mutations in the DNA of cells, referred to as mutational signatures [5].

With these, in 2013, Alexandrov et al. have developed a mathematical framework which allowed for the successful extraction of such mutational signatures, by looking at a set of simulated cancer genomes [6]. Besides these, the framework also came with the capability to identify the corresponding number of mutations that can be attributed to each of these signatures, referred to as their exposures. In order to prove the usefulness of their model, they then applied it to almost 5 million mutations among more than 7000 actual cancer genomes, which in turn led to the discovery of more than 20 mutational signatures in various types of cancer [7].

Since the development of the framework, people have contributed to the field in various ways, dividing the problem of analysis of mutational signatures in two distinct approaches, mainly, de novo signature extraction and signature fitting. The technique of refitting, commonly known as signature fitting, the focus of our paper, refers to the process of assigning the signatures, that have been previously discovered and documented in the COSMIC database [8], to each of the samples studied. In this process, signature fitting is trying to identify the number of mutations that can be attributed to each of the signatures found to be active, in order to reconstruct the mutational profile of the sample (mutational profile - a reflection of the mutations found in a sample), such that this reconstruction is as close as possible to the ground-truth mutation profile [9].

Even though the field has been enriched with a variety of computational methods that have improved the accuracy of deciphering and understanding of these mutational signatures, a more in-depth investigation was limited by the sequencing method itself, which only had the capability of representing the genome of a cancer by aggregating the DNA of cells sampled from a biopsy, which we refer to as sequencing the genome in a bulk manner. However, recent advances in the field have been able to overcome this limitation by providing a look at the genetic information of single cells themselves [10], which opens the opportunity for more focused exploration of cancer heterogeneity.

In this paper we will therefore try and contribute to the field by applying the available mathematical frameworks onto the genetic information extracted from cancerous single-cells. We believe that this could enhance our understanding into the processes which lead to the development of cancer. We are interested in relating our findings to pseudo-bulk data that we will generate ourselves from our cells, since there is no ground-truth bulk representing the genome from which the single-cells were sampled from.

2.1 Research question

We will be looking at performing signature fitting rather than de novo extraction, particularly, we'll investigate **what is the effect of performing mutational signature fitting for single-cell by relating it to pseudo-bulk**.

- We will investigate whether applying signature fitting for single-cells can uncover some mutational signatures that are not observed when fitting at the pseudo-bulk level. We believe this will indeed be the case, deciphering in this way a degree of heterogeneity across cells and how during their lifetime, these cells acquire various mutations due to different levels of exposure to a plethora of mutagens. This complexity, we believe to be hard to capture when working with an aggregation of cells, as it is the case in our pseudo-bulk data.

- Q1 - We will look at metrics that try to quantify the accuracy of the reconstructed mutational profile for both single cell data as well as the pseudo-bulk one, and assess whether one provides a better accuracy over the other. We believe that, because the pseudo-bulk represents an aggregation of multiple cells encompassing all across mutations, this will affect the way in which the model fits for the active mutational signatures, driving down its accuracy for the reconstruction of the mutational profile in relationship to the single-cells.

- Q2 - We will then impose the question to whether we can cluster our single-cells based on their reconstructed mutational profiles, in a way that would explain the heterogeneity across the cells, and to whether these clusters can explain the set of active mutational signatures of the pseudo-bulk data.
- Q3 - Finally, we will generate pseudo-bulk samples from subpopulations of cells and investigate whether these are different from the one created from the entire population, and how do these relate back to the cells which generated them. We believe that subpopulations of single-cells will give rise to pseudo-bulk samples that would replicate closely their cells' set of active mutational signatures due to the aggregation process, rather than being a close copy of the pseudo-bulk generated from all cells.

3 Methodology

This section will describe the choices we took throughout the project in order to achieve its completion. We'll thus start by describing the data together with the technical tools used for analyzing it. We'll then portray our reasoning behind generating the pseudo-bulk data from the entire population of cells. Next, we'll introduce the metrics used for assessing the accuracy of reconstruction of mutational profiles between single-cell data and the pseudo-bulk one. Then, we'll explain how was clustering of single-cells performed based on their reconstructed mutational profiles, finishing with how we've chosen to generate pseudo-bulk samples from subpopulations of cells, obtained by grouping the data on the basis of their mutation count.

3.1 Data

There are 688 single-cells that were sequenced from a breast cancer tumor coming from a donor. The data was made available by the supervisor team with no information(with exception for age - 65), on the individual, providing therefore anonymity and privacy. The information for each of the single-cells comes in the format of a VCF file with the following column information: CHROM, POS, FILTER, REF and ALT. Each row represents essentially a position in the genetic information of the cell in which a mutation occurred. The chromosomes fall in the range from 1 to 22 including also the X sex chromosome and M coming from mitochondrial genetic information. It is also relevant to specify that the data contains a single mutation class, more precisely, single base substitution(SBS) which represents the substitution of a single base in the genome of a cell [11].

3.2 Performing signature fitting

Since the scope of the project involves the lookout for relevant patterns in the data, we've chosen Python as the programming language to tackle the challenge because of the available libraries like NumPy [12], Pandas [13], that aid with data processing and analytics, and MatPlotLib [14] for data visualization. Not only this, but the main framework used for signature fitting, SigProfilerAssignment [15], is implemented in Python.

Medo et al. in their paper [16] have performed a comprehensive review of different tools for fitting mutational signatures and found that SigProfilerSingleSample is the best tool on average for fitting when the number of mutations is small, as it is in our case. SigProfilerSingleSample(which is now deprecated) has been fully incorporated in the SigProfilerAssignment [15], making it perhaps the best choice for the current project. The library utilizes a custom forward stagewise algorithm for sparse regression, for fitting the COSMIC signatures that are active in the sample, as well as the technique of non-negative least squares for numerical optimization, for approximating the number of mutations attributed for each active mutational signature [15].

Since we previously stated that the only mutation class in our data are single base substitutions, the SBS mutational signatures are the only ones from the COSMIC library which are being evaluated in the fitting process by SigProfilerAssignment. The SBS mutational signatures are each quantified through a vector with 96 entries, each describing a mutation type within a tri-nucleotide context (the actual base substituted is in the middle, surrounded by the before and after bases in the 5' → 3' direction) [17]. The value for each entry is a float in the range of 0 to 1, representing the number of mutations that can be attributed to a specific mutation type, expressed as a percentage of the total number of mutations attributed to that signature - the exposure of the signature.

Now, some of these SBS signatures from COSMIC are actually similar from the perspective of their mutational profiles. This similarity may affect the way in which these are identified during the signature fitting process, therefore, in our analysis we will compare the found signatures by looking at the cosine similarity of the vectors corresponding to each pair.

3.3 Generation of pseudo-bulk data

One of the aspects that is worth mentioning about the project is that there is no bulk data that would describe the genome of the breast cancer tumor from which the 688 single cells came from.

Instead, this was artificially generated from our single cells by taking inspiration from the variant calling pipeline, the process which is used usually to identify the mutations at the bulk-level data. When identifying mutations at the level of the genome of such a sample, multiple fragments of the sequenced DNA are aligned against a reference genome. If enough of these fragments find a specific mismatch at a position in the DNA, the position is then classified as being a mutation [18]. This pipeline is however prone to errors due to noise captured during the sequencing process and/or misalignment of the fragments against the reference genome. For this reason, several papers showcase the need to choose specific thresholds for classifying a mutation as being significant [19].

We then try and simulate the pipeline process by saying that if a mutation appears in at least a percentage 'x' of the total number of cells, it follows that the specific mutation is part of the pseudo-bulk cancer genome. We choose to generate multiple pseudo-bulk samples by looking at several such percentages for the acceptance of a mutation, to accommodate for the possible noise described before.

3.4 Comparison in the reconstruction accuracy for the mutational profile between single-cells and pseudo-bulk - Q1

In order to look at the performance of signature fitting between the pseudo-bulk and the single-cell data, we chose to analyze the degree of similarity between the original mutational profile against the reconstructed mutational profile for each sample. For these, the mutational profile represents a vector for the true count of mutations that was inferred from the VCF files. Each one of these vectors has exactly 96 entries, each of which representing one of the mutational types within a tri-nucleotide context, for the single-base substitution class, where the value is the actual count of mutations for that type. The reconstructed mutational profile represents the vector that was obtained by performing signature fitting, being a linear superposition of the identified mutational signatures with the weights being their corresponding exposures, hence describing the same 96 entries as the vector discussed previously.

The similarity between these two vectors was quantified by means of two metrics: cosine similarity and Pearson's correlation coefficient. Cosine similarity measures the angle between two vectors and outputs a value between -1 and 1, where 1 indicates a perfect similarity. It was used because of the extent to which it can explain similarity between two vectors by means of directionality, which is essential in our case since we want to see if a profile can be reconstructed by preserving the ratios of the ground-truth count of the mutation types. For the same reason, Pearson's correlation coefficient was chosen to augment the similarity found by the cosine metric, because it outputs a number between -1 and 1 which suggests the linear relationship between two vectors, irrespective of their magnitude. In our analysis, we also wanted to include KL divergence, since it quantifies the deviation from an underlying distribution, which could be considered our mutational profile, however, we will showcase later how this metric was found to be strongly influenced by a confounder in our data.

To have an in-depth analysis for the data, we compare each of the single cells against the pseudo-bulk data.

3.5 Clustering of single-cells - Q2

We chose to cluster our single-cells based on the reconstructed mutational profile corresponding to each sample. This is calculated as the superposition of the found active mutational signatures with their corresponding number of mutations, i.e. the exposure, as the weights, serving thus as a proxy for both. We chose this approach because we wanted to group the cells based on their found exposures, but at the same time, to preserve the fact that two cells, which may appear different in regards to the signatures identified, are actually very similar because of the resemblance between these signatures themselves.

K-medoids will be utilized as the algorithm to perform the clustering, mainly because we're interested in using cosine similarity as the distance metric. Cosine similarity as the distance metric makes more sense in the context because we're interested into clustering our vectors based on directionality, disregarding the magnitude. This is because we want cells with proportionally similar reconstructed mutational profiles to appear in the same cluster because these are encompassing a similar or exactly the same set of active mutational signatures. Due to the difference in number of mutations, a magnitude sensible metric such as euclidean distance is a poor choice exactly because it fails to grasp this proportionality. To make this more explicit with an example, if two single cells have been identified with the same set of mutational signatures, but one has a higher relative count of mutations to the other, the metric sensible to magnitude will pull apart these two points during clustering. To find the hyperparameter of number of clusters in our data, we will investigate different numbers and choose the best one by using the elbow method.

We want to investigate the within cluster similarity which we will quantify through the average cosine similarity between all vectors within the cluster. We're also interested in analyzing the similarity between the clusters which we'll calculate by having for each cluster a representative vector which will represent the 'average' point within the cluster. What it means is that for each cluster, we will average out the reconstructed mutational profiles for all the points within the cluster. Finally, we will compute the cosine similarity between these cluster representatives.

UMAP will be used to be able to visualize the data and make inferences on it. We chose UMAP because it is capable of preserving local and global structure within the data, whilst being a scalable algorithm for dimensionality reduction.

3.6 Generation of pseudo-bulk samples based on subpopulations of cells - Q3

We've chosen to group the single cell data in bins based on the number of mutations. We refer to bins as a set of imaginary baskets where we place each and every single one of our cells. Each of these bins corresponds to a specific interval of mutations. A cell is thus part of a bin if its corresponding number of mutations falls within this bin's interval of mutations.

We've chosen to therefore split our data in these described bins for four different rounds, each of these rounds corresponding to a different range for each bin's interval, which can also be understood as the bin size. More specifically, those bin sizes were 10, 20, 30 and 50. Thus, for each of these bin sizes, we obtained a different number of bins in which all of our data is grouped into. Because each bin contains a subset of the cells in our data, we say that a bin corresponds to a subpopulation of cells. To illustrate this, for a bin size of 10, we would have, for example, bins with mutation ranges 100-109, 110-119 and so on, where cells would be placed in those bins if their corresponding number of mutations would fall within the interval.

For all and each of these bins across rounds, a corresponding pseudo-bulk sample was generated. Unlike the case for the pseudo-bulk sample generated from the entire population of cells, where we chose different thresholds for the acceptance of a mutation, here we will select all mutations appearing in any of the cells within a bin, basically a 0 percent threshold.

We've chosen to perform these groupings based on mutation counts because we were interested in whether there is a significant difference in the exposures of the pseudo-bulk samples generated from bins with smaller counts of mutations, by contrast to ones encompassing a higher count of mutations. Not only this but also, we wanted to investigate how does having more cells in a bucket, a consequence of having multiple bin sizes, affect the way in which the set of active mutational signatures are found for the pseudo-bulk samples.

It is relevant to point that this process is independent from the clustering with K-medoids described before. In that case we do not generate pseudo-bulk samples based on the identified clusters.

4 Results

The current section will delve into the results of the methods described in the previous section. We will describe first the insights that were gained from performing signature fitting on the single-cell data. Next, we will showcase the results of the generation of the pseudo-bulk data which will open the discussion for the difference between the accuracy of reconstruction for the mutational profiles between this data and the single-cells. Next, the results of the clustering of the single-cells be will shown, while the last section will elicit the differences between several pseudo-bulk samples. Each of the subsection will include an interpretation of the findings in order to keep a consistent flow to the document.

4.1 Signature fitting on the single-cells

As mentioned previously, we are working with 688 single-cell files describing the base substitution mutations at the single-cell level.

The number of mutations per cell varies being described by the following:

- Minimum number - 178 mutations
- Maximum number - 458 mutations
- Mean - 321 mutations
- Standard deviation - 90 mutations

Therefore, we can see a high variation with regards to the mutation count, with a relative lower number of mutations to fit for the frameworks available in the field [6]. Having a high variance is sound because of the clonal evolution of cancerous cells. What this means is that cells from advanced stages of cancer development are usually described by a higher mutation count in contrast to ones from incipient phases.

When performing signature fitting, the framework takes each of the single cells and performs the algorithms described in the previous sections in order to find the mutational signatures that are active in the cells, together with the number of mutations that can be attributed to each of these signatures - the exposure vector.

Across all these cells, 6 different exposure profiles were found, with a total of 6 different active mutational signatures: SBS1, SBS5, SBS12, SBS26, SBS40c and SBS54.

We refer to the coined composite term - exposure profile - in order to describe the set of mutational signatures that were found to be active at the single-cell level by performing signature fitting, e.g. if SBS1 and SBS5 were found to be active in a single cell, the exposure profile of that cell is formed of those two signatures.

We found that 676 cells - 98.25%, lay within 3 different exposure profiles:

- 429 of the single-cells have their exposure profile made out of SBS1, SBS26 and SBS40C
- 145 of them have the exposure profile SBS1, SBS5 and SBS26
- 102 have the exposure profile SBS1, SBS12 and SBS40c

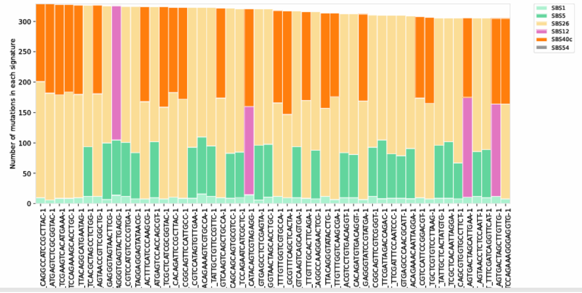


Figure 1: Visualization of signature fitting showcasing a subset of the cells that fall within the 3 exposure profiles with the highest corresponding amount of cells. Each of the cells is represented as a multi-bar entry, where each of these bars represents a mutational signature that is active in that cell sample.

Looking at these profiles more closely, we can understand that the apparent differences between them hide some strong similarities. For example, we can see that SBS40c is never fitted together in a cell with SBS5. The same is true for the pair SBS12 and SBS26. This happens because of the strong similarity between the mutational profiles of these pairs of signatures, as it is showcased in table 1(0.911 cosine similarity for the former pair and 0.929 for the latter).

	SBS1	**SBS5**	**SBS12**	**SBS26**	**SBS40c**	**SBS54**
SBS1	1.000	0.194	0.016	0.052	0.134	0.066
SBS5	0.194	1.000	0.671	0.657	0.911	0.475
SBS12	0.016	0.671	1.000	0.929	0.559	0.535
SBS26	0.052	0.657	0.929	1.000	0.554	0.692
SBS40c	0.134	0.911	0.559	0.554	1.000	0.425
SBS54	0.066	0.475	0.535	0.692	0.425	1.000

Table 1: Cosine similarity between COSMIC mutational signatures found across single-cells

From a computational perspective, in the process of signature fitting, especially in the stagewise algorithm for sparse regression, the algorithm chooses to remove from the set of signatures the one which minimizes the relative increase in error between the mutational profile and the reconstructed mutational profile. For example, assume that SBS26 is active in the cell. Dropping SBS12 is likely to happen in the forward phase because the relative error increases by very little because the reconstructed profile contains SBS26.

The same follows in the backward phase, when the algorithm adds back to the set of active mutational signatures the ones which reduce significantly the relative error. Going back to our example, the algorithm will not add back SBS12 in the active set because the addition will not decrease significantly this relative error, due to the inclusion of the signature SBS26.

The remaining of the cells (688 - 676 = 12 cells) have different combinations of the 6 mentioned signatures, with the specification that only 8 of these have been fitted with 4 active signatures, out of the entire single-cell data. SBS54 appears in all and only those 8 cells. One might think that this category is relevant in our data, however, when looking at the COSMIC library, the proposed etiology for the SBS54 signature says that this signature appears as a possible sequencing artifact. This is likely to be the case considering the relatively small number of cells wherein this signature is identified, compared to the entire population. Because of such, we believe that this signature appears in these cells exactly due to noise in the process of sequencing, rather than being something significant in our data.

Having now discussed this, we can go back to the remaining 4 cells that haven't been talked about, which have the exposure profile composed of the signatures SBS1, SBS5, and SBS12. Due to the similarity between SBS5 with SBS40c, these cells are very similar to the group SBS1, SBS12 and SBS40c, which constitute 102 of the cells. This will be also shown in the clustering phase, where these two groups will be clustered together.

As a final note, we included in the appendix a section where we tried to interpret these results from a more biological perspective(6). It was not included in the discussion of this section since we believe it lies slightly outside of the scope of the project.

4.2 Comparison between pseudo-bulk and single-cell - Q1

We'll start this section by firstly describing the generation of the pseudo-bulk data. We generated this data by choosing different thresholds expressed as a minimum number of cells in which a mutation occurs, in order to account for the noise effect present in the variant calling pipeline process. However, we observed that each of the mutations from the single-cell data appears in at least 23% of the cells. Since the maximum threshold for our data was 15%, thus, even though these thresholds were chosen, every single one of them yielded to the same pseudo-bulk sample. Therefore, this single sample that was generated represents the only one that we'll relate to the single cells. We want to specify that this sample has 459 mutations with the exposure profile formed from the signatures SBS1, SBS26 and SBS40c.

Accuracy of reconstruction of mutational profile across metrics

We find 75 of the cells, 10.9% of the data, achieving higher accuracies than the pseudo-bulk sample. Out of these 75, 6 have different exposure profiles than the pseudo-bulk sample, whereas the rest of them have the same one. There is only a single exposure profile type that doesn't appear in any of these 75 cells from the 6 different ones identified by signature fitting, mainly, the one with the active mutational signatures SBS1, SBS5 and SBS12. The following table shows the min, max, standard deviation and mean values for these 75 cells across the metrics together with the mutation count:

Metric	Cosine Similarity	Correlation	Total Mutations
Mean	0.877627	0.829653	346.4
Standard Deviation	0.006426	0.008377	93.8
Minimum	0.872000	0.822000	188
Maximum	0.898000	0.858000	456

Table 2: Cosine similarity, Pearson's correlation and mutation counts for single cells with strictly higher accuracy than pseudo-bulk for the reconstruction of mutational profile

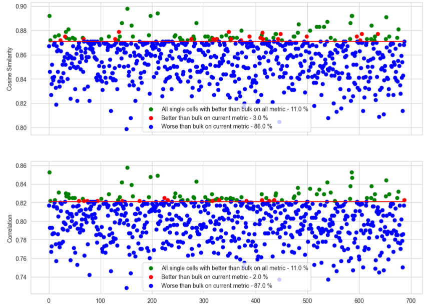
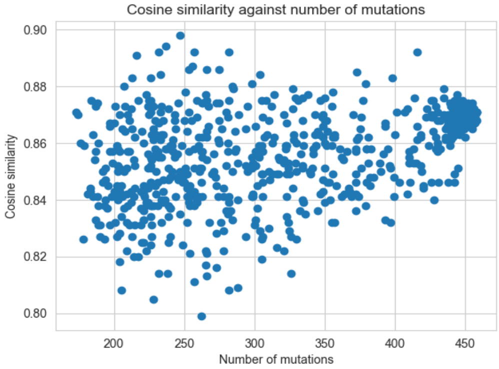
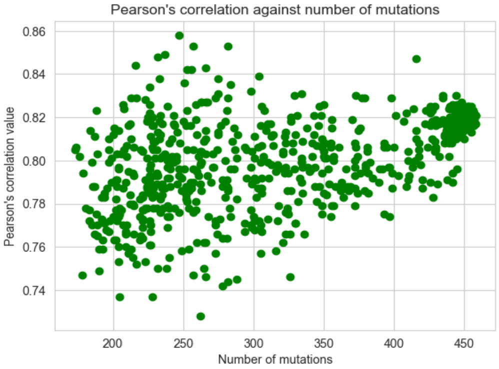


Figure 2: Accuracy of reconstructed mutational profiles. Each point corresponds to a single-cell, the x-axis represents the index of the cell in our collection of cells, while the y-axis represents the value of the accuracy of the reconstructed mutational profile. The red line represents the value for the pseudo-bulk.

Contrary to our initial hypothesis, when looking at the reconstruction accuracy for the mutational profiles across these metrics for pseudo-bulk against single-cells, it seems that single-cells performing better than pseudo-bulk constitutes the exception from the rule, rather than the generality, as it can be seen from the previous illustration. In many of the cases, such can be attributed to the relative lower number of mutations for the single-cells in comparison to the one of the pseudo-bulk sample, as it was also expressed in the paper by Alexandrov, wherein the authors specify that the increase in the number of mutations for a sample yields usually to a higher accuracy in the way the model fits the signatures with their corresponding exposures [6]. Plotting the mutation count against the corresponding cosine similarity or Pearson's correlation value for each cell may contribute to this narrative as the following two graphs show:



Cosine similarity



Pearson's correlation

However, the total count of mutation can only be considered a moderate confounder for these two due to the fact that the correlation between these variables is relatively low (0.41 for cosine and 0.47 for Pearson's correlation) by contrast, for example, to KL divergence, which we briefly mentioned in the methods section as not being used. We choose not to include this metric exactly because it can largely be explained by the total mutation count (-0.93 correlation).

Nonetheless, some of the cells(75) have been found to indeed achieve a better accuracy for the metrics despite the relative smaller number of mutations. We will not make a strong assumption on why we think this is the case. We believed at first that because of the aggregation of mutations describing the pseudo-bulk, these mutations would dilute the influence of the actual mutational signatures contributing to the sample. However, we can't make such a case because there are many cells with a very close number of mutations to the pseudo-bulk, some of them missing only a single mutation by comparison. Accepting such an interpretation would mean that we would classify these cells' genetic information essentially as noise. Only from a mathematical stand, this difference in mutations most likely leads to a profile that aligns better with the SBS COSMIC signature vectors, achieving a better reconstruction as a consequence.

4.3 Clustering based on the reconstructed mutational profiles - Q2

As said previously, we've used the reconstructed mutational profiles in order to cluster the single-cells instead of their exposures in order to have a higher degree of expressivity when comparing the cells, mainly to include in this comparison not only the exposure counts for each active mutational signature, but also the mutational signatures themselves.

We chose the number of clusters that appear in the data by applying the elbow method, plotting against each number of clusters, the inertia got back from performing K-medoids.

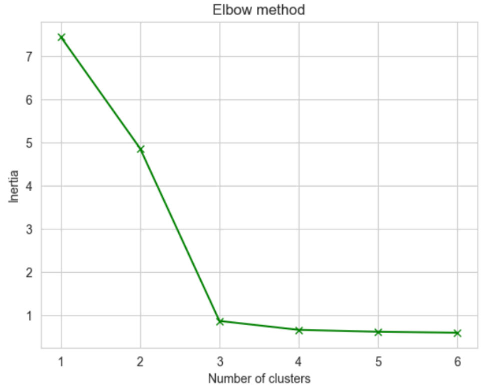


Figure 3: Elbow method performed to find the number of clusters in the data

This way, we've chosen 3 as the number of clusters that can group the data in the most representative way as it can be seen from figure 3.

Having performed K-medoids, we're interested in visualizing these results. As said, we've chosen UMAP in order to reduce the data so that it can be visualized on a two-dimensional grid. Once having reduced the data - we assigned to each point the corresponding cluster found by K-medoids and colored the groups correspondingly. Figure 4 showcases the visualization through UMAP.

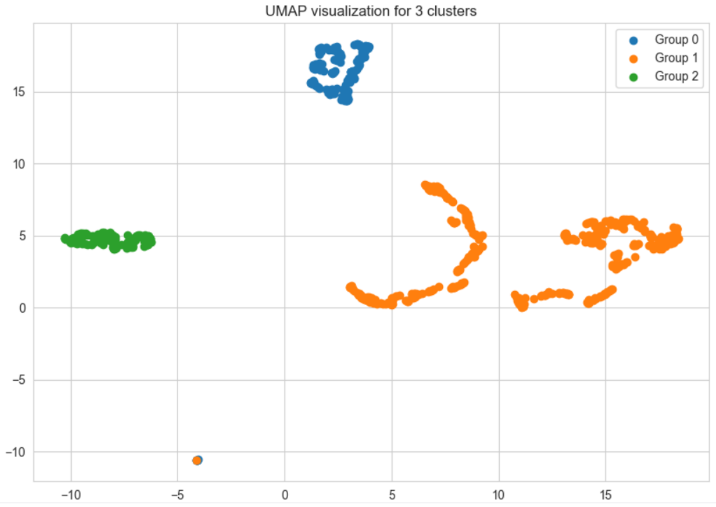


Figure 4: UMAP visualization for the single-cell

- Group 0 - blue color - 152 cells. From these, 145 have the exposure profile SBS1, SBS5 and SBS26 and 7 cells have the exposure profile SBS1, SBS5, SBS12 and SBS54

- Group 1 - orange color - 430 cells; 429 cells - exposure profile SBS1, SBS26, SBS40c and one cell - exposure profile SBS1, SBS12, SBS40c and SBS54
- Group 3 - green color - 106 cells; 102 cells - exposure profile SBS1, SBS12, SBS40c and 4 cells - SBS1, SBS5 and SBS12

As previously pointed, for the within cluster similarity, we calculate the average cosine similarity between each of the single cells within the clusters and every other point within the same cluster. For these, group 0 has an average of 0.9963, group 1 has 0.9982, while the last group has 0.9961.

For the similarity between clusters, we motivated how we created an aggregate of the reconstructed mutational profile of the cells of each group and then compared these points with one another. With this we obtain that the similarity between clusters 0 and 1 is 0.983, between 0 and 2 is 0.933, while the one between 1 and 2 is 0.960. Such a high value for inter-cluster similarity was expected because even though the clusters encompass cells with different active mutational signatures, those signatures are very similar to one another, as showcased before with the similarity between these signatures.

It is of no surprise that the cluster with the most amount of cells, group 1, has, for almost all of the cells, exactly the same exposure profile as the pseudo-bulk sample, possibly indicating to why this sample has this particular exposure. The cells with the highest amount of mutations are in this cluster, which differ very slightly by contrast to the pseudo-bulk.

In the UMAP graph we can see that in the bottom of the screen, there are a few points that are pulled apart from the three clusters. These points are actually the ones in which the signature SBS54 has been identified, showcasing again how these cells can be considered outliers/noise in the data.

4.4 Generation of pseudo-bulk samples based on subpopulations of cells - Q3

As specified before, we have grouped all our single cells in bins based on the their number of mutations, across several ranges of mutations, 10, 20, 30 and 50 - bin sizes - in order to generate, based on subpopulations of cells, pseudo-bulk samples.

In the case of all those mentioned ranges, there is a recurrent pattern in the way that the exposures are being fit, starting from the bins which contain cells with a relatively smaller amount of mutations, compared to the ones with a higher amount(see all bin sizes in appendix).

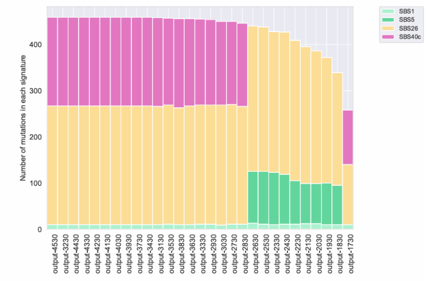


Figure 5: Pseudo-bulk samples corresponding to bins of cells with range of 10

The pseudo-bulks corresponding to bins containing cells with a relatively smaller amount of mutations are always fitted with the active mutational signatures SBS1, SBS5 and SBS26. However, around the count mark of 270, across all bin sizes, the bins containing cells with a higher or equal amount of mutations compared to this threshold, generate pseudo-bulk samples which have exactly the same exposure profile as the one generated from the entire population, that are the active mutational signatures SBS1, SBS26 and SBS40c.

There appears to be a a single exception across all the four ranges for the above described pattern, in that for the first bin corresponding to the range of 10, more specifically the bin having the interval of 173-182 mutations, the pseudo-bulk's exposure profile is also the same as the original pseudo-bulk sample (see figure from above).

Another relevant finding is that for the bin size of 10, contrary to the expected behavior, some of the bins, encompassing cells with smaller counts of mutations, generate pseudo-bulk samples with a higher amount of mutations than others that have been generated from bins of cells with a relative higher mutation count. To illustrate this, the bin with mutation range 283-292 generates a pseudo-bulk with a total mutation count of 446, whereas the bin with a relative smaller count, 273-282, generates a pseudo-bulk sample with 450 mutations. This was not found to be the the case for the rest bin sizes - 20, 30 and 50.

There seems to also be a specific threshold in the data from where the pseudo-bulk samples generated from those bins onward achieve exactly the same amount of mutations as the one that is generated from the entire population, which is usually when the bin contains cells with above 320 mutations.

Switching now to the pseudo-bulks which come from cells with a relatively higher mutational counts, these have the same exposure profile as the one generated from the entire population of cells, having the active mutational signatures SBS1, SBS26 and SBS40c. This finding is expected because of the fact that in our data, when increasing the number of mutations per cell, the percentage of cells identified with this same exposure profile becomes more and more dominant in each of the bins. This closely replicates the insights gathered from the previous section, where the cluster with the majority of the cells have in almost all cases the same exposure profile as the pseudo-bulk sample representing the entire population.

However, this doesn't appear to be the case for bins encompassing cells with a smaller count of mutations. Take for example the bin of cells with range 173-182, which is composed of 6 cells. Out of these, there is a single cell that has the same exposure profile as the pseudo-bulk generated, with the active mutational signatures of SBS1, SBS26 and SBS40c. The rest five have the active signatures SBS1, SBS12 and SBS40C. Not only that, but it also seems that the pseudo-bulk is fitted sometimes with a mutational signature that was not fitted in any of the cells - the ranges 183-192 till 233-242 are mostly composed from a majority of cells with either the exposure profile encompassing SBS1, SBS12 and SBS40c, or SBS1, SBS26 and SBS40c. However, for all these bins, the pseudo-bulk sample that is generated from these cells has the exposure profile with signatures SBS1, SBS5 and and SBS26. We believe this happens because of the acquired mutations that describe the aggregate pseudo-bulk. We believe that in the aggregation phase, the combination of mutations is done in such a way that will drive the model towards fitting another signature that describes better the aggregated mutations. We believe this is the reason for why we see SBS5 fitted in these pseudo-bulks.

5 Responsible research and limitations

In our paper we examined single-cell data coming from a breast cancer tumor from a donor. As specified earlier, the data was provided by the supervisor team, with no information about the donor herself, other than the age of the woman - 65. The data we worked with specifically was pre-processed by our supervisor team, therefore the exact VCF files are not public. However, the raw data is [1].

We do acknolewdge that due to the unavailability of the actual VCF files that we worked on, it is hard to reproduce the setup for the project, reason for which we tried to document extensively the methods and results to aid for at least the replicability of this project.

One of the main challenge of this project came from the need to generate a pseudo-bulk sample that would be representative of an actual bulk sample coming from the breast cancer tumor. Even though we believe that in our approach we tried to replicate the reality of the method done usually in this process, there is no ground-truth in this manner. This being said, we tried to only make suppositions over the found results, instead of strong claims.

Another limitation comes from the relative small number of mutations from the single cells. The technique of signature fitting has been documented to be affected by the number of mutations that characterize a sample, which is also the case in our study.

In regards to the accuracy of the reconstruction of the mutational profiles between single-cell and pseudo-bulk, we do account for the possible confounding effect of total number of mutation, which is however moderate in our metrics used.

6 Conclusion

In this paper, we investigated the effect of performing signature fitting on single-cell data in contrast to a pseudo-bulk sample that was generated from it. We found that signature fitting is able to uncover more mutational signatures at the level of the single-cells in comparison to the pseudo-bulk sample. We've also seen that some of the cells perform better in the reconstruction of the mutational profiles across some accuracy metrics. We've showcased that we can cluster the single-cells based on their reconstructed mutational profiles in a way that reduces the heterogeneity across the cells to a few clusters and that the majority cluster is representative of the pseudo-bulk sample. However, the generation of the pseudo-bulk samples from smaller populations have showcased that this is not always the case, and indicated that some underlying biological interpretation might otherwise explain these observations.

This research had however the limitation in that no ground-truth bulk data was available as a benchmark comparison with our single-cells, together with the fact that the number of mutations describing our cells was relatively low, reason for which future research should investigate on different cancer types described by cells with a higher mutation count, as well as the availability of the ground-truth bulk describing the tumor from where the cells have come from.

[1] https://www.10xgenomics.com/datasets/750-sorted-cells-from-human-invasive-ductal-carcinoma-3-lt-v-3-1-3-1-low-6-0-0

References

[1] Wikipedia contributors, "Cancer – Wikipedia, The Free Encyclopedia," https://en.wikipedia.org/wiki/Cancer, accessed: 2025-06-20.

[2] N. McGranahan and C. Swanton, "Clonal heterogeneity and tumor evolution: Past, present, and the future," *Cell*, vol. 168, no. 4, pp. 613–628, Feb. 2017, doi: 10.1016/j.cell.2017.01.018.

[3] P. Pfeifer, G. "Environmental exposures and mutational patterns of cancer genomes," *Genome Medicine*, vol. 2, no. 8, p. 54, Aug. 2010, doi: 10.1186/gm175.

[4] M. Di Noia, J. and S. Neuberger, M. "Molecular mechanisms of antibody somatic hypermutation," *Annual Review of Biochemistry*, vol. 76, pp. 1–22, 2007, doi: 10.1146/annurev.biochem.76.061705.090740.

[5] COSMIC, "Mutational signatures," https://cancer.sanger.ac.uk/signatures/, 2024, accessed: 2025-06-16.

[6] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton, "Deciphering signatures of mutational processes operative in human cancer," *Cell Reports*, vol. 3, no. 1, pp. 246–259, 2013, doi: 10.1016/j.celrep.2012.12.008.

[7] L. Alexandrov, S. Nik-Zainal, D. Wedge *et al.*, "Signatures of mutational processes in human cancer," *Nature*, vol. 500, pp. 415–421, 2013, doi: 10.1038/nature12477.

[8] J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, P. Fish, B. Harsha, C. Hathaway, S. C. Jupe, C. Y. Kok, K. Noble, L. Ponting, C. C. Ramshaw, C. E. Rye, H. E. Speedy, R. Stefancsik, S. L. Thompson, S. Wang, S. Ward, P. J. Campbell, and S. A. Forbes, "COSMIC: the Catalogue Of Somatic Mutations In Cancer," *Nucleic Acids Research*, vol. 47, no. D1, pp. D941–D947, Jan 2019, doi: 10.1093/nar/gky1015.

[9] A. Baez-Ortega and K. Gori, "Computational approaches for discovery of mutational signatures in cancer," *Briefings in Bioinformatics*, vol. 20, no. 1, pp. 77–88, Jan 2019, doi: 10.1093/bib/bbx082.

[10] C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, I. Hellmann, and W. Enard, "Comparative analysis of single-cell rna sequencing methods," *Molecular Cell*, vol. 65, no. 4, pp. 631–643.e4, Feb 2017, doi: 10.1016/j.molcel.2017.01.023.

[11] COSMIC, "Single base substitutions (sbs)," https://cancer.sanger.ac.uk/signatures/sbs, 2024, accessed: 2025-06-12.

[12] C. R. Harris, K. J. Millman, S. J. van der Walt *et al.*, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020, doi: 10.1038/s41586-020-2649-2.

[13] J. Reback, W. McKinney, jbrockmendel, J. Van Den Bossche, T. Augspurger, S. Cloud, K. Hawkins, Gfyoung, Sinhrks, A. Klein *et al.*, "pandas-dev/pandas: Pandas 1.0.3," 2020, doi: 10.5281/zenodo.3509134.

[14] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[15] M. Díaz-Gay, R. Vangara, M. Barnes, X. Wang, S. M. A. Islam, I. Vermes, S. Duke, N. B. Narasimman, T. Yang, Z. Jiang, S. Moody, S. Senkin, P. Brennan, M. R. Stratton, and L. B. Alexandrov, "Assigning mutational signatures to individual samples and individual somatic mutations with sigprofilerassignment," *Bioinformatics*, vol. 39, no. 12, p. btad756, Dec. 2023, doi: 10.1093/bioinformatics/btad756.

[16] M. Medo, C. K. Y. Ng, and M. Medová, "A comprehensive comparison of tools for fitting mutational signatures," *Nature Communications*, vol. 15, no. 1, p. 9467, Nov. 2024, doi: 10.1038/s41467-024-53711-6.

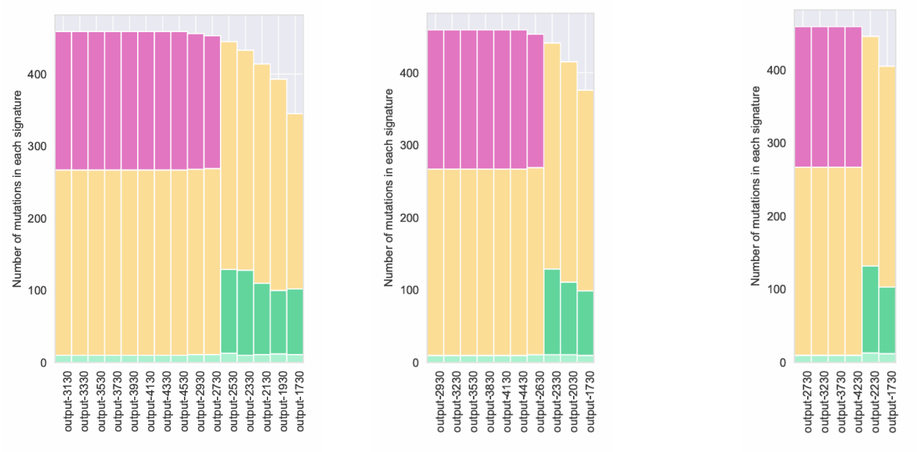
[17] E. N. Bergstrom, M. N. Huang, U. Mahto, M. Barnes, M. R. Stratton, S. G. Rozen, and L. B. Alexandrov, "Sigprofilermatrixgenerator: a tool for visualizing and exploring patterns of small mutational events," *BMC Genomics*, vol. 20, no. 1, p. 685, 2019, doi: 10.1186/s12864-019-6041-2.

[18] Melbourne Bioinformatics, "Advanced variant detection background," https://www.melbournebioinformatics.org.au/tutorials/tutorials/var_detect_advanced/var_detect_advanced_background/#bcf-file-format, n.d., accessed: 2025-05-23.

[19] K. O. Wrzeszczynski, V. Felice, A. Abhyankar, L. Kozon, H. Geiger, D. Manaa, F. London, D. Robinson, X. Fang, D. Lin, M. F. Lamendola-Essel, D. Khaira, E. Dikoglu, A. K. Emde, N. Robine, M. Shah, K. Arora, O. Basturk, U. Bhanot, A. Kentsis, M. M. Mansukhani, G. Bhagat, and V. Jobanputra, "Analytical validation of clinical whole-genome and transcriptome sequencing of patient-derived tumors for reporting targetable variants in cancer," *Journal of Molecular Diagnostics*, vol. 20, no. 6, pp. 822–835, Nov. 2018, doi: 10.1016/j.jmoldx.2018.06.007; Epub 2018 Aug 21.

Appendix

The following images represent the pseudo-bulk samples generated from subpopulations of cells, for all the four bin sizes - 10, 20, 30, 50. Each plot contains, for each bucket of cells from that range, the exposure profile corresponding to the pseudo-bulk sample generated from that bucket.



(a) Bucket size of 20

(b) Bucket size of 30

(c) Bucket size of 50

Figure 6: VCF Ranges for Buckets

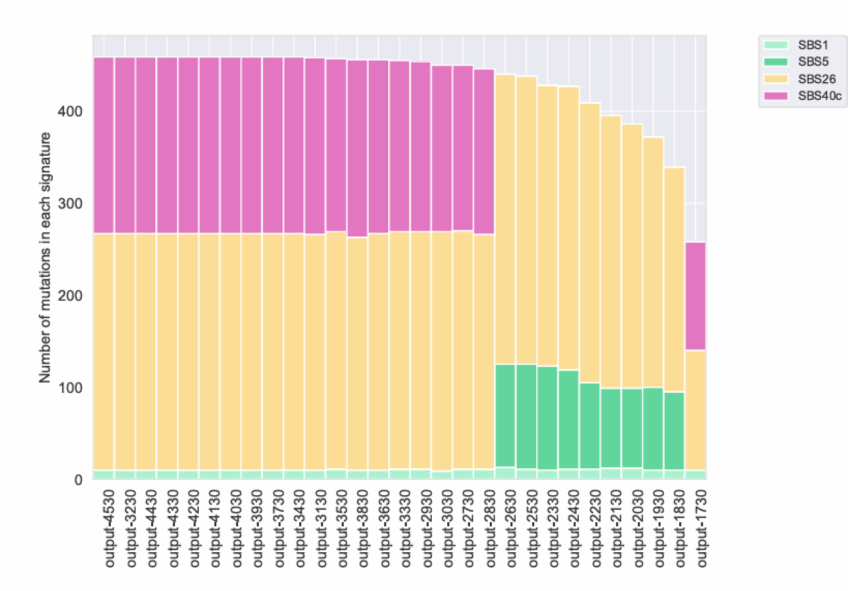


Figure 7: Bucket size of 10

Hypothesis for biological interpretation of signature fitting across cells

This section of the appendix tries to provide a logical reasoning in order to give our findings some biological interpretation. It was left out from the results because of it's scope and it's degree of interpretation. We ground the discussion in the following assumptions:

1. The single-cell sequencing has identified almost all/all mutations for all the single cells.
2. Cells with relatively smaller mutation counts have appeared earlier in the development of the cancer, in contrast to others.

Firstly, even though very similar between one another, we believe that SBS12 and SBS26 appear as a manifestation of different processes that contribute to the development of cancer, rather than being a consequence of mathematical limitations. One might be tempted to draw the latter conclusion, by witnessing that SBS12 appears in cells with a relatively lower count of mutations, whereas SBS26 generally appears in ones with a higher count, when the two have a high similarity in their proposed profiles.

From statistics, we know that realizations of distributions with a smaller sample count are usually noisier by contrast to the ones with a relatively higher sample count. This could be a potential explanation for why they are fitted they way they are, however, we showcase through the following graph how the mutations accumulated for SBS26 across cells with increasing number of mutations seem to grow with a higher pace by contrast to the ones of SB12 (fig 8). This difference can elicit the under-pinning of different processes manifesting as a consequence of clonal evolution which gives the cells that appear further in the cancer development to be more prone to deficiencies of endogenous mechanisms like DNA repair. SBS26 is associated with defective DNA mismatch repair and microsatellite instability which in turn lead to a higher accumulation of new mutations [2].

By contrast, such a strong case can't be made for the pair SBS5 and SBS40c, regarding different underlying processes since there is no clear separation in the cells that are fitted with one against the other. There is also no difference in the increase of mutations as it is the case for the previous pair (fig 9). Thus, for this pair, we can't bring forward the idea of different underlying processes, case also made in current studies [3].

[2] K. R. Loeb and L. A. Loeb, "Genetic instability and the mutator phenotype. Studies in ulcerative colitis," Am. J. Pathol., vol. 154, no. 6, pp. 1621–1626, Jun. 1999, doi: 10.1016/S0002-9440(10)65415-6.

[3] T. Hwang, L. K. Sitko, R. Khoirunnisa, F. Navarro-Aguad, D. M. Samuel, H. Park, B. Cheon, L. Mutsnaini, J. Lee, B. Otlu, S. Takeda, S. Lee, D. Ivanov, and A. Gartner, "Comprehensive whole-genome sequencing reveals origins of mutational signatures associated with aging, mismatch repair deficiency and temozolomide chemotherapy," Nucleic Acids Res., vol. 53, no. 1, Jan. 2025, Art. no. gkae1122, doi: 10.1093/nar/gkae1122.

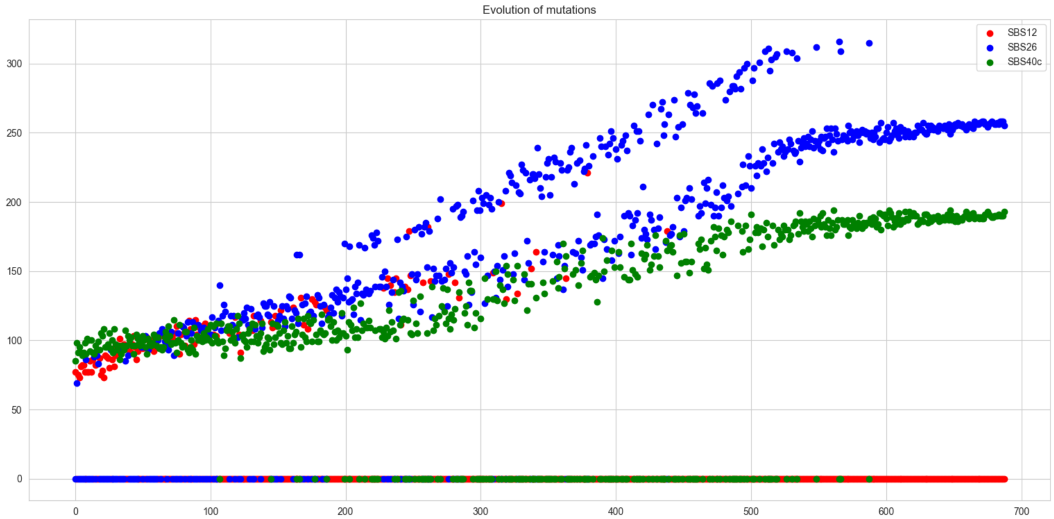


Figure 8: Difference across cells in mutations attributed to SBS40c, SBS12 and SBS26

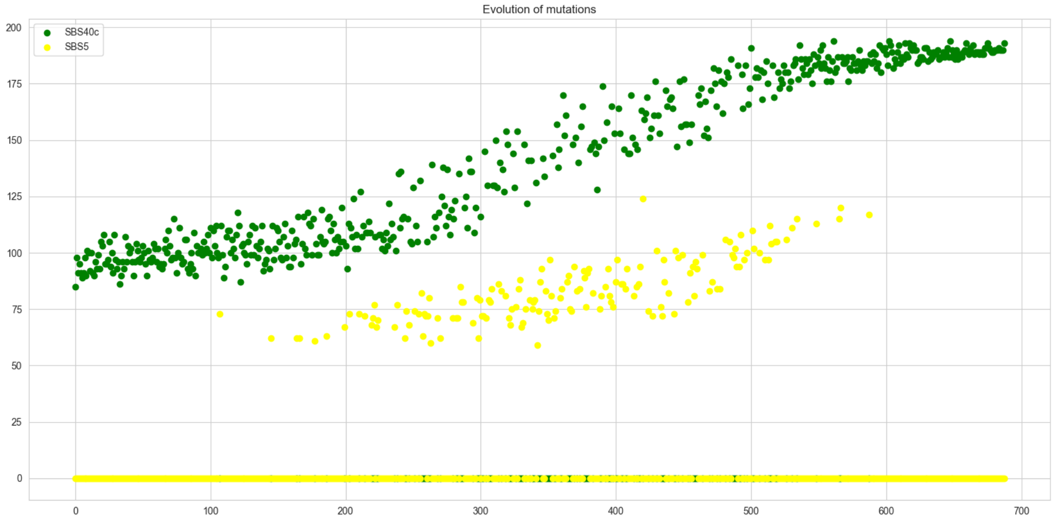


Figure 9: Difference across cells in mutations attributed to SBS5 vs SBS40c

We also believe that the driving factor for the development of this cancer in particular is the age. SBS1, SBS5 have been previously showcased to be correlated with this factor. Also, tumors in older individuals are more prone to DNA repair defects, thus SBS26 can also sustain this idea. We believe in particular that the process under-pinning SBS26 is perhaps the main driver to the rapid development of cancer in this patient, as we can see from the increase in mutations as this signature starts to manifest more actively in the cells.