



## OPEN ACCESS

EDITED BY  
Federico Zambelli,  
University of Milan, Italy

REVIEWED BY  
Rosario Michael Piro,  
Politecnico di Milano, Italy  
Shilpa Garg,  
University of Copenhagen, Denmark

\*CORRESPONDENCE  
Tiziana Sanavia,  
✉ tiziana.sanavia@unito.it  
Piero Fariselli,  
✉ piero.fariselli@unito.it

SPECIALTY SECTION  
This article was submitted to  
Computational Genomics,  
a section of the journal  
Frontiers in Genetics

RECEIVED 20 September 2022  
ACCEPTED 07 December 2022  
PUBLISHED 04 January 2023

CITATION  
Pancotti C, Rollo C, Birolo G,  
Benevenuta S, Fariselli P and Sanavia T  
(2023), Unravelling the instability of  
mutational signatures extraction *via*  
archetypal analysis.  
*Front. Genet.* 13:1049501.  
doi: 10.3389/fgene.2022.1049501

COPYRIGHT  
© 2023 Pancotti, Rollo, Birolo,  
Benevenuta, Fariselli and Sanavia. This is  
an open-access article distributed  
under the terms of the [Creative  
Commons Attribution License \(CC BY\)](#).  
The use, distribution or reproduction in  
other forums is permitted, provided the  
original author(s) and the copyright  
owner(s) are credited and that the  
original publication in this journal is  
cited, in accordance with accepted  
academic practice. No use, distribution  
or reproduction is permitted which does  
not comply with these terms.

# Unravelling the instability of mutational signatures extraction *via* archetypal analysis

Corrado Pancotti, Cesare Rollo, Giovanni Birolo,  
Silvia Benevenuta, Piero Fariselli\* and Tiziana Sanavia\*

Department of Medical Sciences, University of Torino, Torino, Italy

The high cosine similarity between some single-base substitution mutational signatures and their characteristic flat profiles could suggest the presence of overfitting and mathematical artefacts. The newest version (v3.3) of the signature database available in the Catalogue Of Somatic Mutations In Cancer (COSMIC) provides a collection of 79 mutational signatures, which has more than doubled with respect to previous version (30 profiles available in COSMIC signatures v2), making more critical the associations between signatures and specific mutagenic processes. This study both provides a systematic assessment of the *de novo* extraction task through simulation scenarios based on the latest version of the COSMIC signatures and highlights, through a novel approach using archetypal analysis, which COSMIC signatures are redundant and more likely to be considered as mathematical artefacts. 29 archetypes were able to reconstruct the profile of all the COSMIC signatures with cosine similarity >0.8. Interestingly, these archetypes tend to group similar original signatures sharing either the same aetiology or similar biological processes. We believe that these findings will be useful to encourage the development of new *de novo* extraction methods avoiding the redundancy of information among the signatures while preserving the biological interpretation.

## KEYWORDS

archetypal analysis, mutational signatures, matrix factorization, COSMIC, cancer genomics

## Introduction

Somatic mutations in cancer are caused by a wide range of endogenous (i.e. genome instability or deficiency in a DNA repair mechanism) or exogenous (environmental exposures such as ultraviolet radiation or tobacco smoking) mutagenic agents, which stratify over time. It has been hypothesised that a mutational pattern in the genome can be deconvolved considering different generative processes, each of them associated with a specific mutational signature represented by 96 somatic mutation frequencies of six single nucleotide variants ( $C > A: G > T, C > G: G > C, C > T: T > A, T > A: A > T, T > C: A > G, T > G: A > C$ ) flanked by

one nucleotide on each side (Alexandrov et al., 2013). The latest version of the Catalogue Of Somatic Mutations In Cancer (COSMIC) (Bamford et al., 2004) hosts 79 single-base substitution (SBS) signatures extracted from 2,780 genomes of the Pan Cancer Analysis of Whole Genomes (PCAWG) (as described in <https://cancer.sanger.ac.uk/signatures/sbs/>) using SigProfilerExtractor (Islam et al., 2022), an updated version of the original method based on Non-Negative Matrix Factorization (NMF) proposed by Alexandrov et al. (2020).

Although signature extraction analysis is becoming a routine, there are some issues that should be further investigated. Recently, some studies have highlighted caveats and warnings in using these representations for clinical applications (Maura et al., 2019; Koh et al., 2021). Omichessan et al. (2019) did an empirical evaluation of the main *de novo* extraction tools showing that the identification of signatures is more difficult for tumours characterised by multiple signatures having a small contribution, and pointing out that different signatures might have very close cosine similarity, as it was observed between COSMIC signatures (cosine similarity > 0.8). Huang et al. used two complementary approaches to assess the confidence and stability of the resulting decomposition, showing that some mutagenic signatures (e.g., the signatures related to APOBEC activity) are more stable with respect to others, whose instability can be explained, in part, by the fact that these signatures can be decomposed into a linear combination of other signatures with a very small error (Huang et al., 2018). This was further tested by Schumann et al. (2019), who clearly showed that COSMIC signatures are, by construction, non-orthogonal, and this aspect affects the stability of the exposures. In particular, they computed the error between each COSMIC signature profile and its reconstruction using all the other signatures, finding four signatures with cosine similarity higher than 0.95 with respect to their reconstructed profile.

Lal et al. (2021) pointed out that state-of-the-art NMF-based methods aim to minimise the residual error after fitting the data with the discovered signatures to fit the data perfectly, which may generate overfitting issues by including stochastic noise in the data as part of the signatures, or multiple similar signatures for the same underlying process. Indeed, the goal of the signature discovery is not only to fit the data as well as possible, but also to identify signatures that truly reflect separate biological processes and the current version (3.3) of COSMIC database shows several signatures with no experimentally-validated aetiological causes associated yet, suggesting possible mathematical artefacts due to overfitting rather than distinct mutagenic processes (Koh et al., 2021). These issues become more critical when the signature extraction is highly dependent on the number of samples available, which complicates the correct identification

of the true components, jeopardising the stability of the results. In addition, the studies of Maura et al. (2019) and Lal et al. (2021) highlighted that the presence of *flat signatures*, showing similar frequencies across all the 96 mutational classes, could represent a source of background noise and collinearities, making the *de novo* signature extraction task difficult and ambiguous.

These issues are expected to become more critical in the newest version of COSMIC catalogue, which considers 79 signatures compared to the 30 signatures of the previous versions investigated in the above studies. Therefore, the proposed study will focus on two main goals: 1) to provide a systematic approach to assess to which extent the extraction of the newest version of COSMIC signatures can be affected by the high similarity among the signatures in the same catalogue, the presence of flat signatures and the number of available samples; 2) to provide a compact representation of the current catalogue by prioritising the identification of those profiles representing extreme patterns in the data so that all the observations can be reproduced as mixtures of their extremes. To this aim, Archetypal Analysis (Cutler and Breiman, 1994) was applied to represent how the information from COSMIC can be projected into a reduced number of dimensions and to explain potential instability issues in specific extraction scenarios. Figure 1 displays a summary of the workflow analysis proposed in this study.

## Materials and methods

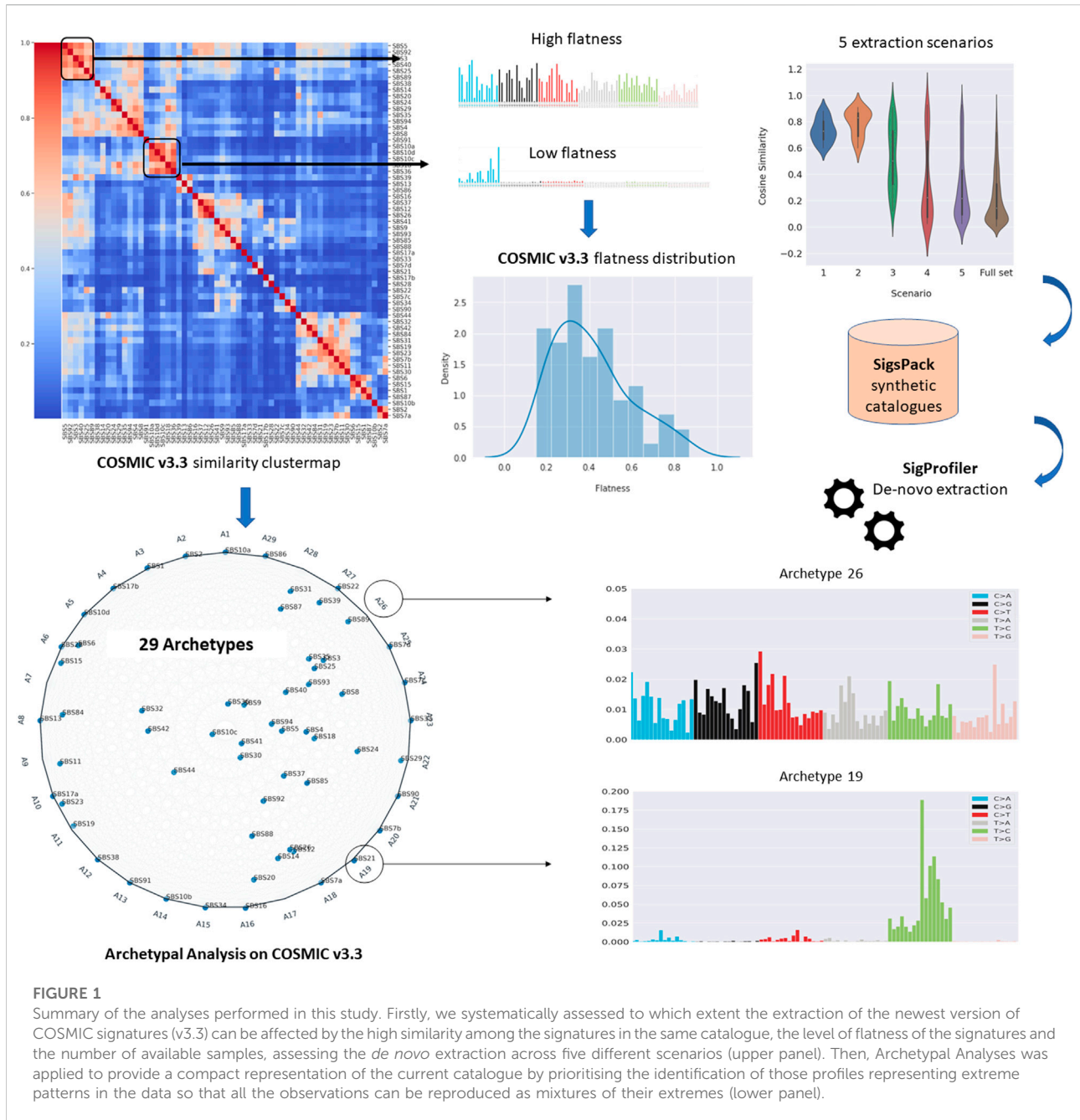
### Similarity of COSMIC signatures

Analyses were performed on COSMIC catalogue v3.3 considering the 79 SBS mutational signature profiles on the reference genome GRCh37 identified by SigProfilerExtractor (Islam et al., 2022). Among these, we removed 19 signatures classified by the catalogue as sequencing artefacts (<https://cancer.sanger.ac.uk/signatures/sbs/>). Of the remaining 60 signatures considered, 19 neither have a direct association with an experimentally validated mutagenic process nor are they supported by statistical association with a specific process.

We first quantified the pairwise level of similarity in the signature catalogue. Therefore, for each pair of signatures  $\mathbf{s}_i$  and  $\mathbf{s}_j$  we calculated the cosine similarity:

$$\cos(\mathbf{s}_i, \mathbf{s}_j) = \frac{\mathbf{s}_i \cdot \mathbf{s}_j}{\|\mathbf{s}_i\| \|\mathbf{s}_j\|} \quad (1)$$

obtaining a pairwise cosine distance matrix  $D \in \mathbb{R}^{d \times d}$ , where  $D(i,j) = 1 - \text{sim}(i,j)$  and  $d = 60$  represents the number of signatures considered. We then built a cluster map to simulate different *de novo* extraction scenarios by applying to the  $D$  distance matrix a Hierarchical Clustering (Johnson, 1967) with average linkage.



**FIGURE 1**

Summary of the analyses performed in this study. Firstly, we systematically assessed to which extent the extraction of the newest version of COSMIC signatures (v3.3) can be affected by the high similarity among the signatures in the same catalogue, the level of flatness of the signatures and the number of available samples, assessing the *de novo* extraction across five different scenarios (upper panel). Then, Archetypal Analyses was applied to provide a compact representation of the current catalogue by prioritising the identification of those profiles representing extreme patterns in the data so that all the observations can be reproduced as mixtures of their extremes (lower panel).

### Flatness of COSMIC signatures

SBS3, SBS5, SBS40 and SBS8 are often referred to as flat signatures given their relative featureless profile, almost uniformly distributed across the 96 mutational classes. However, to the best of our knowledge, no quantitative definition of flatness has ever been provided. To fill this gap we formulated a simple definition of signature flatness by calculating the cosine similarity between the signature and the

uniform distribution. Therefore the flatness of a signature  $s_i$  can be defined as:

$$flatness(s_i) = \cos(s_i, s_u) \tag{2}$$

Where  $s_u$ , in the case of SBS mutational signatures, consists of a signal uniformly distributed over the 96 mutational classes. Hence, the flatness is a score ranging from 0 to 1, where 1 represents a perfectly flat profile. Since the presence of flat signatures in a catalogue can complicate the extraction task, a

**TABLE 1** Summary of the *de novo* extraction scenario. For each simulated scenario, the number of active signatures, the cosine similarity level, the flatness level and the n° of samples.

Scenario	No of latent Signatures	Median Similarity	Median Flatness	No of Samples
1	6	0.73	0.76	200,500
2	5	0.83	0.34	200,500
3	11	0.50	0.64	200,500,1000,3000,5000,10000
4	11	0.22	0.44	200,500,1000,3000,5000
5	20	0.22	0.45	1000,3000,5000,10000

quantitative definition of flatness can be useful to build robust *de novo* extraction methods and to test their capabilities to correctly extract multiple signatures with different levels of flatness. In this regard, we designed scenarios with different levels of similarity and flatness.

## De novo extraction scenarios

To assess both the reliability and the feasibility of the *de novo* extraction procedure, several synthetic catalogues were generated considering COSMIC mutational signatures as underlying generative processes. The function `create_mut_catalogues` from SigsPack R package (<https://github.com/bihealth/SigsPack>) was used to generate 10 synthetic mutational catalogues for each extraction scenario to take into account statistical fluctuations (Schumann et al., 2019). In particular, `create_mut_catalogues` takes *samples*, *mutations* (per sample) and *signatures* as input and it generates mutational catalogues with exposures to the specified signatures by sampling the mutations from a distribution of those signature profiles. The contribution of each signature (exposure) is randomly drawn from a uniform distribution for each sample. Different ranges between a minimum of 200 and a maximum of 10,000 samples were set according to the chosen scenario and the number of signatures considered. The number of mutations in each tumour sample was set to 5,000 for each scenario, based on the PCAWG median number of sample mutations. All the simulated scenarios are summarised in Table 1 and the generated catalogues are available at the Github repository <https://github.com/compbio-med-univ/archetypal-analysis-cosmic>.

*De novo* extraction analysis was applied to each scenario using the gold-standard approach SigProfilerExtractor (Islam et al., 2022) (<https://github.com/AlexandrovLab/SigProfiler>), with the aim of evaluating the extraction performance from catalogues with groups of similar latent signatures, varying in number and flatness score. To choose the optimal number of latent signatures, SigProfilerExtractor performs a repeated NMF for a range of  $k$  operative signatures. For each value of  $k$ , this algorithm applies a custom partition clustering based on the

Hungarian algorithm to the signature matrices resulting from the repeated NMF, so that a final consensus signature matrix is obtained. The number of NMF repetitions can be specified by the user. In our experiments this value was set to 30. All the other default parameters were used. To efficiently run SigProfilerExtractor, computational resources from HPC4AI center (Aldinucci et al., 2018) and Occam cluster (Aldinucci et al., 2017) were used, for a total of 64 CPUs and 6 GPUs.

## Evaluation metrics

Four metrics were considered for the performance evaluations:

- Frequency ( $F$ ) of simulation runs where all the signatures are correctly identified:

$$F = \frac{\text{N° of successful runs}}{\text{N° of total runs}} \quad (3)$$

- Mean square error (MSE) between simulated and reconstructed catalogues:

$$MSE = \frac{\sum_{i=1}^n \sum_{j=1}^m (x_{i,j} - \hat{x}_{i,j})^2}{n \cdot m} \quad (4)$$

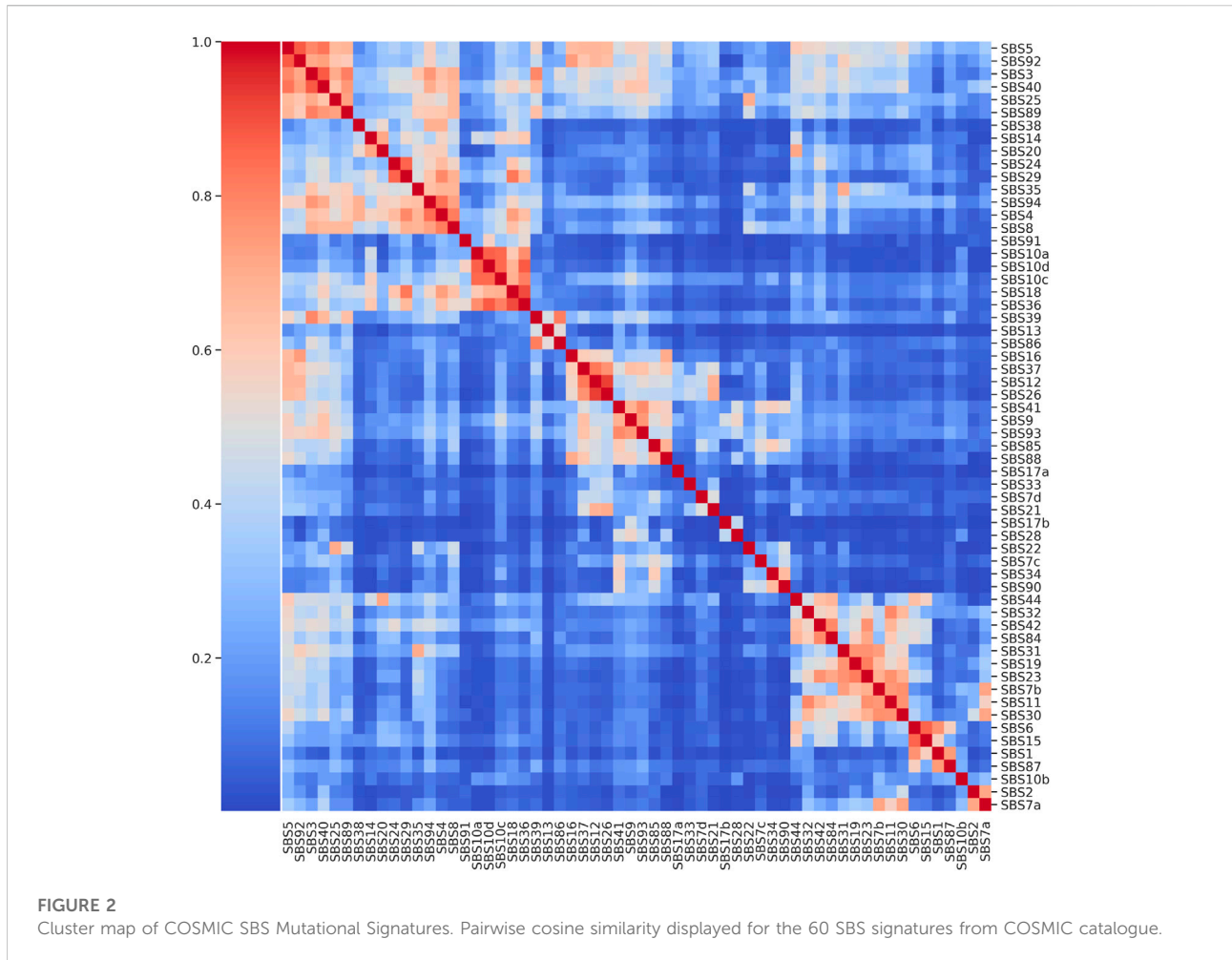
where  $x_{i,j}$  and  $\hat{x}_{i,j}$  are the matrix elements of the original  $X \in \mathbb{R}^{n \times m}$  and the reconstructed  $\hat{X} \in \mathbb{R}^{n \times m}$  catalogues, respectively.

- Average stability ( $C_{mean}$ ) measured by the mean silhouette coefficient score of the signature clusters generated by SigProfilerExtractor:

$$C_{mean} = \frac{\sum_{k=1}^K \sum_{i=1}^N c_{ik}}{K \cdot N} = \frac{\sum_{k=1}^K C_k}{K} \quad (5)$$

$c_{ik}$  is the silhouette coefficient of the  $i$ -th sample which belongs to cluster  $k$ .  $N$  is the number of NMF runs performed by SigProfilerExtractor,  $K$  is the number of cluster labels, and  $C_k$  is the mean silhouette score of the  $k$ -th cluster.

$c_{ik}$  is given by:



$$c_{ik} = \frac{b_{ik} - a_{ik}}{\max(a_{ik}, b_{ik})} \tag{6}$$

where  $a_{ik}$  is the mean intra-cluster distance and  $b_{ik}$  is the mean nearest-cluster distance of the  $i$ -th sample which belongs to cluster  $k$ .

- Minimum stability ( $C_{min}$ ), represented by the minimum silhouette coefficient score of the signature clusters generated by SigProfilerExtractor:

$$C_{min} = \min\{C_k\}, \quad \text{with } k = 1, \dots, K \tag{7}$$

### Archetypal analysis

Archetypal analysis (AA) is an unsupervised learning method that aims to represent data points as sparse convex combinations of their extreme elements. More formally, let  $X \in \mathbb{R}^{p \times q}$  be a matrix whose row vectors are  $\mathbf{x}_i \in \mathbb{R}^q$  and let

$Z \in \mathbb{R}^{r \times q}$  be another matrix, whose column vectors  $\mathbf{z}_k \in \mathbb{R}^q$  represent the archetypes. AA reconstructs  $X$  through a linear combination of archetypes  $\mathbf{z}_k$ , which are themselves convex combinations of the  $X$  rows  $\mathbf{x}_i$ . Therefore, AA solves the following constrained equation:

$$\text{minimize } \sum_{i=1}^p \|\mathbf{x}_i - \sum_{k=1}^r \alpha_{ik} \mathbf{z}_k\|^2 \quad \text{with } \mathbf{z}_k = \sum_{j=1}^q \beta_{kj} \mathbf{x}_j \tag{8}$$

where  $\alpha_{ik} \geq 0$ ;  $\sum_{k=1}^r \alpha_{ik} = 1$  and  $\beta_{kj} \geq 0$ ;  $\sum_{j=1}^q \beta_{kj} = 1$  ensure that the archetypes fall on the convex hull of  $X$ .

In this study, AA was applied directly to the COSMIC signature matrix  $M \in \mathbb{R}^{60 \times 96}$  where, as reported above, 60 is the number of COSMIC SBS mutational signatures and 96 are the mutation contexts. Since AA is a matrix decomposition technique, the number of basis elements has to be chosen, which in our case corresponds to the number of archetypes. This number was set in order to explain the 95% of the variance. AA was performed using the Python-based *Archetypal Analysis Package* freely available at <https://data.csiro.au/collection/csiro>:



40600v1 (Motevalli Soumehsaraei and Barnard, 2019). The code and the archetypal profiles are available at the Github repository <https://github.com/compmiomed-unito/archetypal-analysis-cosmic>.

## Results

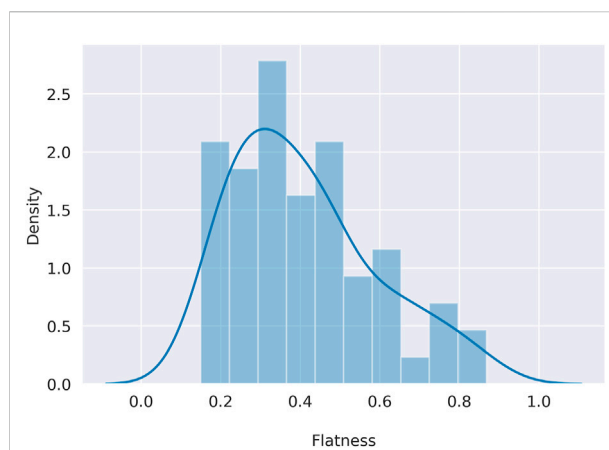
### COSMIC cluster map

The cluster map on COSMIC v3.3 catalogue revealed that there are several groups of signatures showing pairwise cosine similarity  $> 0.8$  (Figure 2). The first group in the top-left corner consists of six signatures and it includes those signatures which are commonly referred to as *flat*, i.e. presenting a relatively featureless profile distributed over all the 96 mutational contexts, as the well-known SBS3 experimentally associated with defective homologous recombination-based DNA damage repair and the clock-like SBS5, statistically associated with age. The median pairwise similarity is 0.73, with a maximum equal to 0.88 for the pairs SBS3-SBS40 and SBS5-SBS92, while the median flatness, calculated through Eq. 2, is 0.76. These six signatures were used to build up the first extraction scenario presented in Table 1.

A second notable group is characterised by a high pairwise cosine similarity among signatures but with a low level of flatness and it includes SBS36, SBS18 and the three signatures SBS10a, SBS10c and SBS10d associated with an altered activity of polymerase (polymerase epsilon exonuclease domain mutations and defective POLD1 proofreading), which were considered for the second extraction scenario. The median pairwise similarity is 0.83 with a maximum equal to 0.91 between SBS36 and SBS18 while the median flatness is 0.34. In the third extraction scenario the synthetic catalogues were generated from the signatures used in the first and the second scenario together (11 signatures). Finally, in the fourth and fifth extraction scenarios, 11 and 20 signatures with a low flatness score were considered, respectively, where each signature has at least another similar one (cosine similarity  $> 0.8$ ). The complete list of signatures used in each scenario and the list of those signatures with a cosine similarity  $> 0.8$  were reported in Supplementary Tables S1 and S2, respectively. In Supplementary Figure S1 we reported the pairwise similarity distribution for each scenario compared with the full set of non-artefactual COSMIC signatures. The first and second scenarios have a very high similarity level as they were built by taking the two largest clusters of the similarity-based cluster map. The others are gradually less similar since the considered number of signatures increases but the number of similar signatures in each cluster decreases.

### Flatness analysis

To overcome the qualitative description of flatness, in Eq. (2) (Methods section) we defined a simple way to quantitatively



**FIGURE 3**

Distribution of the COSMIC flatness. On the x axis the flatness defined in Eq. 2, on the y axis the density for each flatness level.

assess the flatness of the signatures, being in line with the qualitative description. Indeed, as shown in Supplementary Table S3, the known flat signatures SBS3, SBS40 and SBS5 show the highest degree of flatness, but a similar level to SBS5 can be found for SBS25 and SBS89. In addition, this definition of flatness appears to be well distributed within COSMIC from a minimum of 0.15 (SBS1) to a maximum of 0.87 (SBS3), showing that this metric can emphasise the differences in shape between the various signatures in COSMIC, as shown in Figure 3. As mentioned in the previous section, the extraction scenarios, built to highlight possible issues in the *de novo* extraction process, differ in the number of signatures involved, the pairwise similarity between profiles, and the level of flatness. In Supplementary Figure S2, the flatness distribution for each scenario is shown.

### De novo signature extraction

The SigProfilerExtractor performance for each scenario is shown in Table 2. MSE,  $C_{mean}$  and  $C_{min}$  are reported as their corresponding median values across 10 repetitions, together with their inter-quartile range. When considered separately, signatures involved in scenarios 1 and 2 were almost always correctly extracted at 200 samples ( $F = 0.9$ ), regardless of the high level of similarity in each group.

However, when the extraction was performed by combining these two scenarios (scenario 3), SigProfilerExtractor was never able to identify the correct number of signatures up to a high number of samples (5,000) and only at 10,000 samples it succeeded 80% of the times ( $F = 0.8$ ). In this scenario it is worth noticing that, as the number of available samples increases, while the MSE decreases and the average stability decreases but remaining relatively high, the minimum stability dramatically

**TABLE 2** *De novo* signature extraction performance. For each simulated scenario, the frequency of runs with all the signatures correctly identified  $F$ , the mean square error (MSE) between simulated and reconstructed catalogues, the average  $C_{mean}$  and minimum  $C_{min}$  stability scores of signature clusters are displayed.

Scenario	Number of samples	F	MSE	$C_{mean}$	$C_{min}$
1	200	0.9	45.59 (45.45, 46.34)	0.88 (0.86–0.89)	0.78 (0.74–0.82)
	500	1	46.64 (46.41, 46.89)	0.93 (0.9–0.94)	0.86 (0.75–0.88)
2	200	0.9	27.75 (27.62, 28.05)	0.84 (0.83–0.86)	0.67 (0.54–0.69)
	500	1	28.11 (27.85, 28.45)	0.87 (0.86–0.89)	0.69 (0.65–0.72)
3	200	0.0	124.36 (120.66, 127.48)	0.98 (0.97–0.98)	0.96 (0.94–0.97)
	500	0.0	129.08 (126.64, 130.17)	0.99 (0.99–0.99)	0.98 (0.97–0.99)
	1,000	0.0	90.49 (41.68, 127.92)	0.90 (0.81–0.99)	0.80 (0.58–0.98)
	3,000	0.0	41.05 (40.89, 41.28)	0.81 (0.8–0.82)	0.47 (0.41–0.5)
	5,000	0.0	40.18 (38.99, 41.38)	0.81 (0.80–0.82)	0.35 (0.26–0.49)
4	200	0.0	101.93 (85.45, 103.0)	0.88 (0.84, 0.9)	0.76 (0.68–0.81)
	500	0.1	40.38 (38.24, 41.01)	0.82 (0.82–0.83)	0.38 (0.3–0.45)
	1,000	0.9	33.20 (32.81, 33.42)	0.83 (0.81–0.85)	0.52 (0.47–0.58)
	3,000	1	32.58 (32.56, 32.69)	0.86 (0.85–0.87)	0.58 (0.53–0.62)
	5,000	1	32.54 (32.36, 32.69)	0.89 (0.87–0.89)	0.67 (0.6–0.7)
5	1,000	0.0	68.98 (67.01, 70.04)	0.83 (0.81–0.86)	0.49 (0.33–0.59)
	3,000	0.0	37.42 (37.16, 37.68)	0.81 (0.81–0.82)	0.33 (0.26–0.39)
	5,000	0.0	37.01 (36.94, 37.12)	0.81 (0.80–0.82)	0.34 (0.29–0.39)
	10,000	0.1	36.63 (36.44, 36.77)	0.81 (0.81–0.82)	0.38 (0.32–0.4)

decreases. As expected, by further increasing the number of samples, both the mean and minimum stability values rise again. However, given that obtaining 10,000 tumour samples is often unfeasible in practice, this scenario highlights well a limitation of the NMF-based extraction process. Indeed, this scenario is particularly complex since it considers two main subgroups, highly similar internally (median pairwise cosine similarity 0.73 and 0.83, respectively) but one at high and the other one at low flatness score (0.76 and 0.34, respectively, as shown in Table 1). Therefore, this difference in the flatness levels makes the extraction process much more difficult if there is not a very large number of samples. As shown in Supplementary Figure S3, the algorithm starts to differentiate similar signatures inside each of these two groups at 3,000 samples, but still failing at differentiating them well even at 5,000 samples.

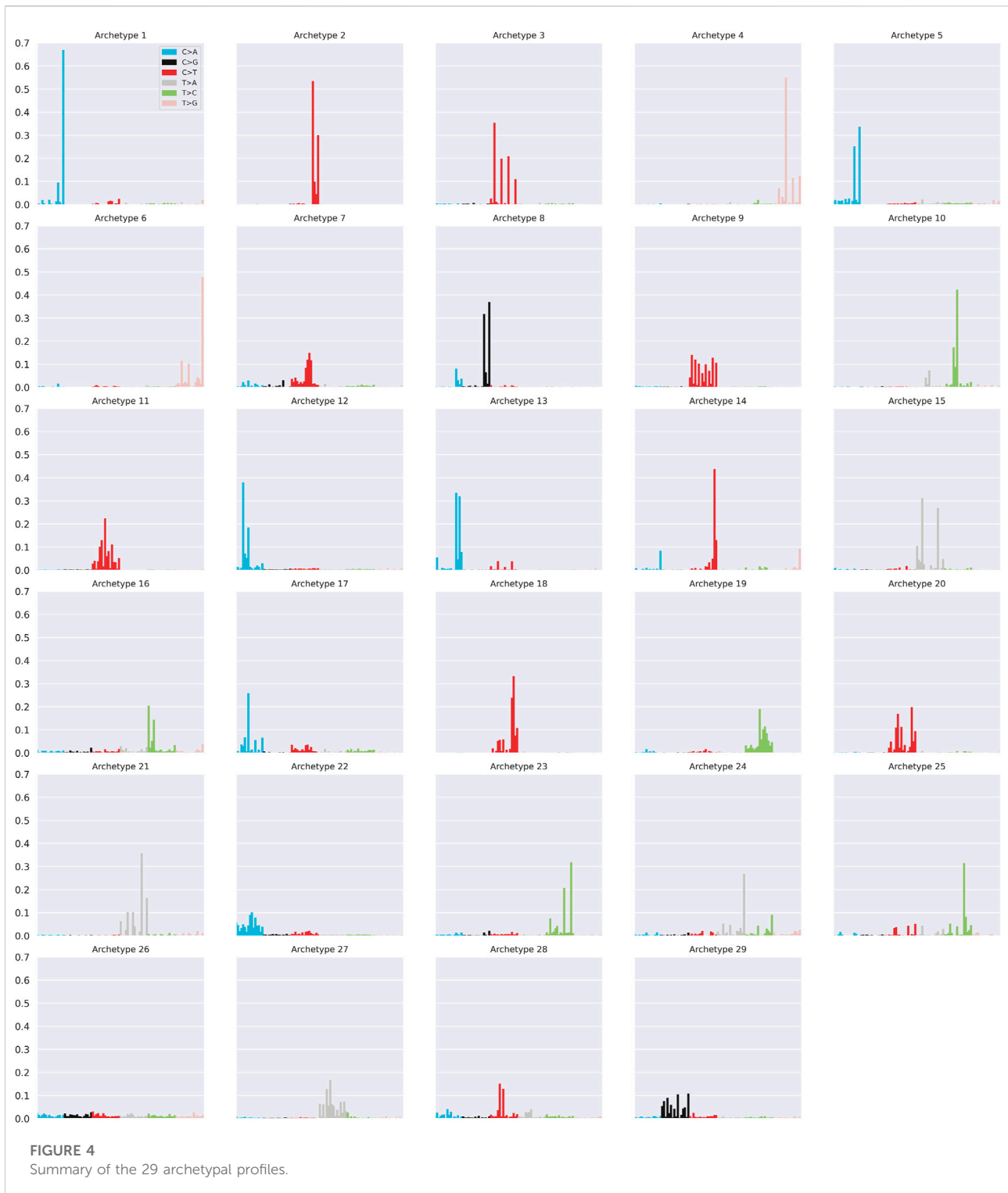
On the other hand, considering again 11 signatures but with a lower level of similarity and flatness (scenario 4), the algorithm required at least 1,000 samples to identify the signatures with  $F = 0.9$  (Supplementary Figure S3). Finally, when 20 signatures were considered (scenario 5), the algorithm always failed even at 5,000 samples, and it only succeeded 10% of the times at 10,000 samples. It is worth

highlighting that the maximum number considered is significantly higher compared to the 2,780 genomes from PCAWG used to build the gold-standard catalogue of mutational signatures available in COSMIC.

## Archetypal analysis

The application of AA to the COSMIC SBS mutational signature matrix  $M \in \mathbb{R}^{60 \times 96}$ , revealed that 29 archetypes  $\mathbf{z}_k \in \mathbb{R}^{96}$  ( $k = 1, \dots, 29$ ) were able to explain the 95% of the variance (Supplementary Figure S4) and that through a combination of them it is possible to reconstruct each COSMIC signature profile with cosine similarity  $> 0.8$  (Supplementary Figure S5).

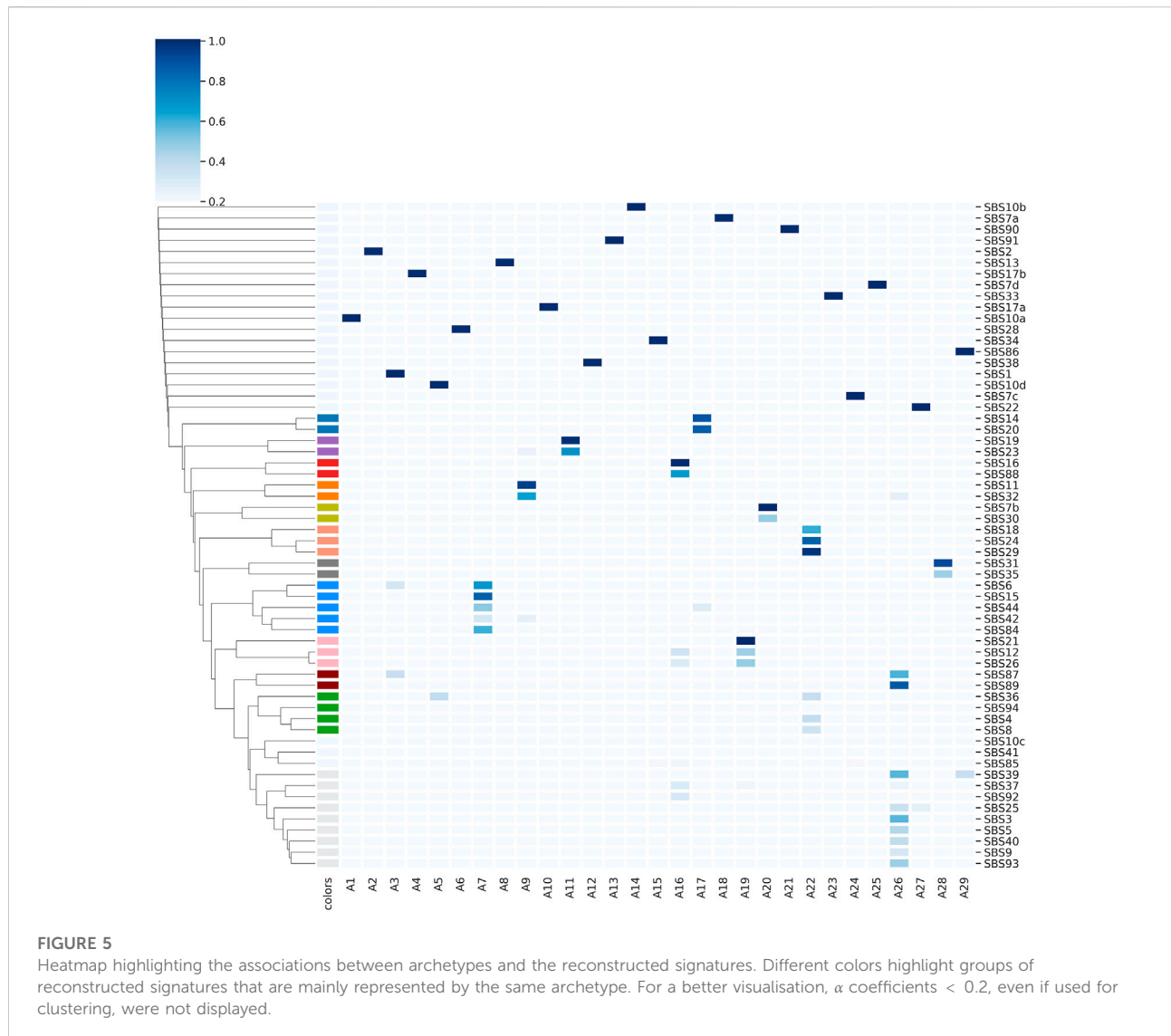
The archetypal profiles are summarised in Figure 4. Most of the archetypal profiles coincide almost perfectly with some of the COSMIC signatures. Specifically, 26 out of 29 archetypes correspond to at least one COSMIC signature with cosine similarity of at least 0.97 (Supplementary Figure S6). These results suggest that a subset of signature profiles represents extreme patterns of the catalogue and that a combination of them is capable of reconstructing the entire



catalogue with a high level of accuracy. The relationship between signatures and archetypes can be better understood considering the  $\alpha$  coefficients of Eq (8). In particular, the coefficients  $a_{ik}$  represent the weights that each archetype  $\mathbf{z}_k$  has in the reconstruction of the  $i$ -th signature  $\mathbf{x}_i$ .

Figure 5 shows the association between the COSMIC signatures and the archetypes through the  $\alpha$  coefficients. The heatmap was consequently clustered to find those signatures which share a common reconstruction pattern through the archetypes. It can be seen that 19 archetypes reconstruct only





one signature, indicating a one-to-one relationship between them. Others were found to contribute in more than one signature at different weights, as well as there are groups of reconstructed signatures that are mainly represented by the same archetype, highlighted by different colours in Figure 5.

Interestingly, AA tends to group similar profiles together fairly well, since the signatures belonging to the same group usually share either the same aetiology or similar biological processes. In Supplementary Figure S7 we further explored the relationship between the mutational signatures and the archetypes by plotting the pairwise cosine similarity distribution of the  $\alpha$  coefficients for different categories of pairwise cosine similarity between the original signatures. It is possible to clearly observe that, as the pairwise cosine similarity between the signatures increases, the cosine similarity between  $\alpha$  coefficients increases. This confirms that, while providing a more

compact representation of the COSMIC signatures, the archetypal analysis is able to maintain a good consistency with the original profiles.

Table 3 summarises some of the qualitative information that can be extracted from the heatmap, showing the relationships between the reconstructed signatures and the archetype that contributed most to them. Each signature was reported with its aetiology, and whether it had been validated experimentally or by statistical association (i.e. unclear evidence for real signature, as reported in COSMIC).

It is possible to observe that seven signatures (SBS6, SBS14, SBS15, SBS20, SBS21, SBS26, and SBS44), experimentally associated with mismatch repair (MMR) deficiency, are divided into three groups: Blue, Silver Blue and Pink. The Blue group includes two signatures associated with the concurrent effect of MMR deficiency and DNA polymerase (POLD1 and POLE), showing a profile

**TABLE 3** Aetiological information related to each archetype. For each archetype, the corresponding reconstructed signatures and, when available, their associated aetiologies are reported, indicating the validation studies supporting the biological interpretation.

Color	Main Archetype	Signature	Aetiology	Validation
Blue	A17	SBS14	MMR deficiency + POLE	Experimental Hodel et al. (2020)
		SBS20	MMR deficiency + POLD1	Experimental Meier et al. (2018)
Purple	A11	SBS19	Unknown	-
		SBS23	Unknown	-
Red	A16	SBS16	Unknown	-
		SBS88	Colibactin Exposure	Experimental Pleguezuelos et al. (2020)
Orange	A9	SBS11	Temozolomide treatment	Experimental Kucab et al. (2019)
		SBS32	Azathioprine treatment	Statistical Inman et al. (2018)
Yellow	A20	SBS7b	UV exposure	Experimental Nik-Zainal et al. (2015)
		SBS30	BER deficiency	Experimental Drost et al. (2017)
Salmon	A22	SBS18	Damage by ROS	Experimental Kucab et al. (2019)
		SBS24	Aflatoxin exposure	Experimental Chawanthayatham et al. (2017)
		SBS29	Tobacco chewing	Statistical Alexandrov et al. (2015)
Grey	A28	SBS31	Platinum chemotherapy	Experimental Boot et al. (2018)
		SBS35	Platinum chemotherapy	Experimental Boot et al. (2018)
Silver Blue	A7	SBS6	MMR deficiency	Experimental Meier et al. (2018)
		SBS15	MMR deficiency	Experimental Meier et al. (2018)
		SBS44	MMR deficiency	Experimental Drost et al. (2017)
		SBS84	AID activity	Statistical Kasar et al. (2015)
Pink	A19	SBS12	Unknown	-
		SBS21	MMR deficiency	Experimental Meier et al. (2018)
		SBS26	MMR deficiency	Experimental Meier et al. (2018)
Brown	A26	SBS87	Thiopurine treatment	Experimental Li et al. (2020)
		SBS89	Unknown	-
Green	A22	SBS4	Tobacco smoking	Experimental Nik-Zainal et al. (2015)
		SBS8	HR/NER deficiency	Statistical Alexandrov et al. (2013)
		SBS36	BER deficiency (ROS damage)	Experimental Pilati et al. (2017)
		SBS94	Unknown	-
Light grey	A26	SBS3	HR deficiency	Experimental Zamborszky et al. (2017)
		SBS5	Aging/Tobacco/NER deficiency	Statistical Alexandrov et al. (2013)
		SBS9	Polymerase eta hypermutation	Statistical Alexandrov et al. (2013)
		SBS25	Unknown Chemotherapy	Statistical Alexandrov et al. (2015)
		SBS39	Unknown	-
		SBS40	Unknown	-
SBS93	Unknown	-		

mainly polarised on C > A mutations, whereas the Silver Blue and Pink groups have mutational peaks at C > T and T > C, respectively. Thus, although all these seven signatures are involved in MMR deficiency, they probably refer to different types of deficiency in MMR genes. The Silver Blue group, in addition to MMR deficiency associated signatures, also includes SBS84, which is statistically associated with AID activity, found in the immunoglobulin genes and other regions in lymphoid cancers. Although the MMR pathway is generally involved in repairing errors that can occur in DNA during the replication and the recombination, it was found to cooperate with the AID enzyme to generate DNA mutations as part of the antibody diversification process (Martin and Scharff, 2002; Liu and Schatz, 2009; Zanotti and Gearhart, 2016). Thus, MMR pathway and AID mechanism are closely related. In this context, it is very interesting to note that SBS6, although it is implicated in several tumour types, was mainly found in B-Cell Non-Hodgkin Lymphoma samples <https://cancer.sanger.ac.uk/signatures/sbs/sbs6/> and that SBS84 is associated with the AID activity in B-Cell Non-Hodgkin Lymphoma. This might suggest that SBS6 relates to an MMR deficiency for genes involved in the development of antibody specificity that cooperates with AID activity. On the other hand, in the Pink group, in addition to the two signatures SBS21 and SBS26 associated with MMR deficiency, there is also SBS12 that was mainly found in liver cancer-related tissue; the high similarity (0.93, Table 2) between the signatures SBS26 and SBS12 and the unknown aetiology could suggest that either SBS12 is also related to MMR deficiency or that these two could actually correspond to the same signature. A cluster formed by these two signatures was also highlighted by Degasperis et al. (2020) by performing an organ-wise mutational signature extraction. Another interesting group is the Grey one, including SBS31 and SBS35, both referring to platinum chemotherapy. The Orange group (SBS11 and SBS32) refers to two treatments as well: Temozolomide, an alkylating agent used as treatment for high-grade brain tumors and melanoma, and Azathioprine, an immunosuppressant. Both these treatments were shown to induce myelosuppression (Connell et al., 1993; Sylvester et al., 2011). The associated signatures are mainly represented by archetype A9, whose profile is characterised by high frequency of C > T mutations. The Yellow group includes SBS7b and SBS30, which are both characterised by C > T mutations and linked with ultraviolet (UV) radiation and base excision repair (BER) deficiency. Recently, it was found that BER increases cellular tolerance to UV independently of nucleic excision repair (Saha et al., 2020). The Salmon group consisted in three signatures (SBS18, SBS24 and SBS29) associated with reactive oxygen species (ROS) damage, aflatoxin exposure and tobacco chewing, respectively. These three signatures are linked together by oxidative stress processes, since it is known that aflatoxin biosynthesis is linked with ROS (Shen et al., 1996; An et al., 2017; Dey and Kang, 2020; Huang et al., 2020), as well as the cytotoxic effects from tobacco chewing are mediated by ROS production (Stich and Anders, 1989; Bagchi et al., 2002). The Green group is composed of SBS4, SBS8, SBS36 and SBS94. These signatures are mainly represented by the

archetype A22 as the Salmon group. Indeed SBS29 (Salmon group) associated with tobacco chewing seems to be “complementary” to SBS4 since it was found in some liver and lung cancers where SBS4, related to tobacco smoking, has not been detected. In addition, also SBS36 is associated with BER deficiency including DNA damage due to ROS, as SBS18 in the Salmon group. SBS8 is statistically associated with HR/NER deficiency. Purple, Red and Brown groups include mainly signatures with unknown aetiology and therefore it was not possible to establish qualitative associations between their signatures. Finally, the Light Grey group includes mainly signatures with a high level of flatness (SBS3, SBS5, SBS25 and SBS40) or at least with mutations distributed over all the 96 mutational contexts. Indeed, these signatures are mostly represented by A26, whose profile is in turn homogeneously distributed over all the contexts.

## Discussion

This study investigates the extraction stability issues among the SBS mutational signatures of the most recent version of COSMIC catalogue (3.3). Through a series of simulations considering different scenarios, we showed that high levels of similarity combined with some peculiar (e.g., showing high levels of flatness) signature profiles considerably complicate the *de novo* extraction. Most of the previous studies evaluated stability issues on COSMIC signatures version 2, which includes 30 signatures. However, here we showed that these issues are becoming more critical in the newest version by evaluating 60 non-artefactual signatures. Although SigProfilerExtractor has been proven to be a robust method for signature extraction, even when the number of samples was high (i.e. up to 5,000), it failed in identifying the correct number of signatures and it succeeded 80% of times with 10,000 samples when we simulated a combined set of six similar signatures at high level of flatness and five similar signatures at a lower level of flatness (scenario 3). Similarly, in scenario 5, considering a higher number of latent signatures (20) with at least each signature highly similar (i.e. pairwise cosine similarity > 0.8) to another one, it always failed up to 5,000 samples and with 10,000 samples it succeeded only 10% of the times.

Although the mutational signatures are not orthogonal by definition, the presence of highly similar signatures, together with the fact that some have a very high level of flatness and there is a lack of an aetiology for many of them, cast some doubts on the real existence of some of these, suggesting that they may be the result of overfitting and hence a mathematical artefact. Several studies already pointed to this issue (Maura et al., 2019; Koh et al., 2021). However, to the best of our knowledge, the most recent assessment of the signature stability observed among COSMIC signatures was performed by Schumann et al. (2019), where they considered the second version of this database, therefore working on half the number of signatures compared to our study and without exploring different scenarios in terms of number of samples, cosine similarity and flat vs non-flat signatures. A limitation of the catalogues used in this work, realised with

SigsPack functions, is represented by the random exposure assigned to each latent signature to create the count matrices, subsequently extracted by SigProfilerExtractor. Hence, simulated catalogues may not represent realistic cancer samples. However, this does not affect the technical evaluations of the limitations in the extraction process highlighted by our simulations.

A novelty introduced by this study was the application of AA to investigate whether the information contained in the COSMIC catalogue could be represented more compactly. AA was shown to be an intuitive and straightforward approach to interpret the data like the clustering, but including the flexibility of the matrix factorization (Mørup and Hansen, 2012; Chen et al., 2014). In contrast to the common distance-based approaches, archetypes characterise extremal rather than average properties of the given data and therefore lead to a more compact representation (Abrol and Sharma, 2020). AA is a type of decomposition where convex combinations of extremal points lie on the convex hull of the data and are themselves restricted to being convex combinations of individual observations (Cutler and Breiman, 1994; Mørup and Hansen, 2012). In our study, by applying AA to the COSMIC catalogue, it was possible to identify 29 archetypes able to explain 95% of the variance. Interestingly, it emerged that most of the archetypes correspond almost perfectly (similarity > 0.97) to some signatures and that, through a combination of them, it is possible to reconstruct with a certain degree of accuracy the other signatures of the COSMIC catalogue. As further validation of the reconstruction process, Supplementary Table S4 shows the refitting performance for the simulated catalogues in Scenario 1 with 500 samples, comparing the archetypes to the original COSMIC signatures using MutationalPatterns (Blokzijl et al., 2018). As expected, since the simulated catalogues are generated based on the original signatures, these latter showed high cosine similarity and low mean absolute error (MAE), i.e. 0.967 and 6.2 on average, respectively. However, keeping in mind that the archetypes explain the 95% of the variance in the simulated catalogue, they were able to perform well by achieving average cosine similarity and MAE equal to 0.958 and 9.91, respectively.

However, it is worth highlighting that archetypes do not substitute the COSMIC signatures, but emphasise the importance of considering alternative approaches able to reduce redundant information. These observations, together with the lack of known aetiology and experimental validation for many signatures, suggest the need to reformulate the COSMIC catalogue using representations including sparsity constraints in latent vectors during the extraction procedure without loss of information. In the future, archetypal analysis can be also considered to evaluate sparse representations of signatures not only in the context of single base substitutions but also for other types of variants like copy number variations and structural variants (Heller et al., 2020; Steele et al., 2022).

## Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://github.com/complbiomed-unito/archetypal-analysis-cosmic>.

## Author contributions

Manuscript concept and design: CP and TS; analyses: CP with the support of CR, TS, PF, GB, and SB; manuscript preparation: CP; revision and editing: TS and PF.

## Funding

This study was supported by the Italian Ministry for Education, University and Research (Ministero dell'Istruzione, dell'Università e della Ricerca- MIUR) under the programmes “Ricerca Locale ex-60%” and “Dipartimenti di Eccellenza 2018–2022” (Project code D15D18000410001).

## Acknowledgments

The authors thank the European Union's Horizon 2020 projects GenoMed4All (Grant Agreement ID: 101017549) and Brainteaser (Grant Agreement ID: 101017598). We thank PNRR M4C2 HPC – 1.4 “CENTRI NAZIONALI”- Spoke 8 for fellowship support.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2022.1049501/full#supplementary-material>

## References

- Abrol, V., and Sharma, P. (2020). "A geometric approach to archetypal analysis via sparse projections," in Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria. Editors H. D. III and A. Singh (PMLR), 42–51.
- Aldinucci, M., Bagnasco, S., Lusso, S., Pasteris, P., Rabellino, S., and Vallerio, S. (2017). Occam: a flexible, multi-purpose and extendable HPC cluster. *J. Phys. Conf. Ser.* 898, 082039. IOP Publishing. doi:10.1088/1742-6596/898/8/082039
- Aldinucci, M., Rabellino, S., Pironti, M., Spiga, F., Viviani, P., Drocco, M., et al. (2018). "Hpc4ai: an ai-on-demand federated platform endeavour," in Proceedings of the 15th ACM International Conference on Computing Frontiers, Ischia, Italy, 279–286.
- Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., et al. (2013). Signatures of mutational processes in human cancer. *Nature* 500, 415–421. doi:10.1038/nature12477
- Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., et al. (2015). Clock-like mutational processes in human somatic cells. *Nat. Genet.* 47, 1402–1407. doi:10.1038/ng.3441
- Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., et al. (2020). The repertoire of mutational signatures in human cancer. *Nature* 578, 94–101. doi:10.1038/s41586-020-1943-3
- An, Y., Shi, X., Tang, X., Wang, Y., Shen, F., Zhang, Q., et al. (2017). Aflatoxin b1 induces reactive oxygen species-mediated autophagy and extracellular trap formation in macrophages. *Front. Cell. Infect. Microbiol.* 7, 53. doi:10.3389/fcimb.2017.00053
- Bagchi, M., Balmoori, J., Bagchi, D., Stohs, S. J., Chakrabarti, J., and Das, D. K. (2020). Role of reactive oxygen species in the development of cytotoxicity with various forms of chewing tobacco and pan masala. *Toxicology* 179, 247–255. doi:10.1016/s0300-483x(02)00357-8
- Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., et al. (2004). The cosmic (catalogue of somatic mutations in cancer) database and website. *Br. J. Cancer* 91, 355–358. doi:10.1038/sj.bjc.6601894
- Blokzijl, F., Janssen, R., van Boxtel, R., and Cuppen, E. (2018). Mutational patterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* 10, 33–11. doi:10.1186/s13073-018-0539-0
- Boot, A., Huang, M. N., Ng, A. W., Ho, S.-C., Lim, J. Q., Kawakami, Y., et al. (2018). In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome Res.* 28, 654–665. doi:10.1101/gr.230219.117
- Chawanthayatham, S., Valentine, C. C., III, Fedeles, B. I., Fox, E. J., Loeb, L. A., Levine, S. S., et al. (2017). Mutational spectra of aflatoxin b1 *in vivo* establish biomarkers of exposure for human hepatocellular carcinoma. *Proc. Natl. Acad. Sci. U. S. A.* 114, E3101–E3109. doi:10.1073/pnas.1700759114
- Chen, Y., Mairal, J., and Harchaoui, Z. (2014). "Fast and robust archetypal analysis for representation learning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus (Ohio), USA, 1478–1485.
- Connell, W. R., Kamm, M. A., Ritchie, J. K., and Lennard-Jones, J. E. (1993). Bone marrow toxicity caused by azathioprine in inflammatory bowel disease: 27 years of experience. *Gut* 34, 1081–1085. doi:10.1136/gut.34.8.1081
- Cutler, A., and Breiman, L. (1994). Archetypal analysis. *Technometrics* 36, 338–347. doi:10.1080/00401706.1994.10485840
- Degasperi, A., Dias Amarante, T., Czarnecki, J., Shooter, S., Zou, X., Glodzik, D., et al. (2020). A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. *Nat. Cancer* 1, 249–263. doi:10.1038/s43018-020-0027-5
- Dey, D. K., and Kang, S. C. (2020). Aflatoxin b1 induces reactive oxygen species-dependent caspase-mediated apoptosis in normal human cells, inhibits allium cepa root cell division, and triggers inflammatory response in zebrafish larvae. *Sci. Total Environ.* 737, 139704. doi:10.1016/j.scitotenv.2020.139704
- Drost, J., Van Boxtel, R., Blokzijl, F., Mizutani, T., Sasaki, N., Sasselli, V., et al. (2017). Use of crispr-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science* 358, 234–238. doi:10.1126/science.aao3130
- Heller, D., Vingron, M., Church, G., Li, H., and Garg, S. (2020). Sdip: A novel graph-based approach to haplotype-aware assembly based structural variant calling in targeted segmental duplications sequencing. *BioRxiv*. doi:10.1101/2020.02.25.964445
- Hodel, K. P., Sun, M. J., Ungerleider, N., Park, V. S., Williams, L. G., Bauer, D. L., et al. (2020). Pole mutation spectra are shaped by the mutant allele identity, its abundance, and mismatch repair status. *Mol. Cell* 78, 1166–1177. doi:10.1016/j.molcel.2020.05.012
- Huang, X., Wojtowicz, D., and Przytycka, T. M. (2018). Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics* 34, 330–337. doi:10.1093/bioinformatics/btx604
- Huang, B., Chen, Q., Wang, L., Gao, X., Zhu, W., Mu, P., et al. (2020). Aflatoxin b1 induces neurotoxicity through reactive oxygen species generation, DNA damage, apoptosis, and s-phase cell cycle arrest. *Int. J. Mol. Sci.* 21, 6517. doi:10.3390/ijms21186517
- Inman, G. J., Wang, J., Nagano, A., Alexandrov, L. B., Purdie, K. J., Taylor, R. G., et al. (2018). The genomic landscape of cutaneous scc reveals drivers and a novel azathioprine associated mutational signature. *Nat. Commun.* 9, 3667–3714. doi:10.1038/s41467-018-06027-1
- Islam, S. M. A., Diaz-Gay, M., Wu, Y., Barnes, M., Vangara, R., Bergstrom, E. N., et al. (2022). Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genomics* 2, 100179. doi:10.1016/j.xgen.2022.100179
- Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika* 32, 241–254. doi:10.1007/BF02289588
- Kasar, S., Kim, J., Improgo, R., Tiao, G., Polak, P., Haradhvala, N., et al. (2015). Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* 6, 8866–8912. doi:10.1038/ncomms9866
- Koh, G., Degasperi, A., Zou, X., Momen, S., and Nik-Zainal, S. (2021). Mutational signatures: emerging concepts, caveats and clinical applications. *Nat. Rev. Cancer* 21, 619–637. doi:10.1038/s41568-021-00377-7
- Kucab, J. E., Zou, X., Morganello, S., Joel, M., Nanda, A. S., Nagy, E., et al. (2019). A compendium of mutational signatures of environmental agents. *Cell* 177, 821–836. doi:10.1016/j.cell.2019.03.001
- Lal, A., Liu, K., Tibshirani, R., Sidow, A., and Ramazzotti, D. (2021). De novo mutational signature discovery in tumor genomes using sparsesignatures. *PLoS Comput. Biol.* 17, e1009119. doi:10.1371/journal.pcbi.1009119
- Li, B., Brady, S. W., Ma, X., Shen, S., Zhang, Y., Li, Y., et al. (2020). Therapy-induced mutations drive the genomic landscape of relapsed acute lymphoblastic leukemia. *Blood* 135, 41–55. doi:10.1182/blood.2019002220
- Liu, M., and Schatz, D. G. (2009). Balancing AID and DNA repair during somatic hypermutation. *Trends Immunol.* 30, 173–181. doi:10.1016/j.it.2009.01.007
- Martin, A., and Scharff, M. D. (2002). AID and mismatch repair in antibody diversification. *Nat. Rev. Immunol.* 2, 605–614. doi:10.1038/nri858
- Maura, F., Degasperi, A., Nadeu, F., Leongamornlert, D., Davies, H., Moore, L., et al. (2019). A practical guide for mutational signature analysis in hematological malignancies. *Nat. Commun.* 10, 2969. doi:10.1038/s41467-019-11037-8
- Meier, B., Volkova, N. V., Hong, Y., Schofield, P., Campbell, P. J., Gerstung, M., et al. (2018). Mutational signatures of DNA mismatch repair deficiency in *C. elegans* and human cancers. *Genome Res.* 28, 666–675. doi:10.1101/gr.226845.117
- Mørup, M., and Hansen, L. K. (2012). Archetypal analysis for machine learning and data mining. *Neurocomputing* 80, 54–63. doi:10.1016/j.neucom.2011.06.033
- Motevalli Soumehsaraei, B., and Barnard, A. (2019). *Archetypal analysis package*. Canberra, Australia: Commonwealth Scientific and Industrial Research Organisation.
- Nik-Zainal, S., Kucab, J. E., Morganello, S., Glodzik, D., Alexandrov, L. B., Arlt, V. M., et al. (2015). The genome as a record of environmental exposure. *Mutagenesis* 30, 763–770. doi:10.1093/mutage/gev073
- Omichessan, H., Severi, G., and Perduca, V. (2019). Computational tools to detect signatures of mutational processes in DNA from tumours: a review and empirical comparison of performance. *PLoS one* 14, e0221235. doi:10.1371/journal.pone.0221235
- Pilati, C., Shinde, J., Alexandrov, L. B., Assié, G., André, T., Hélias-Rodzewicz, Z., et al. (2017). Mutational signature analysis identifies mutyh deficiency in colorectal cancers and adrenocortical carcinomas. *J. Pathol.* 242, 10–15. doi:10.1002/path.4880
- Pleguezuelos, C., Puschhof, J., Rosendahl Huber, A., van Hoeck, A., Wood, H. M., Nomburg, J., et al. (2020). Mutational signature in colorectal cancer caused by genotoxic pks+ *e. coli*. *Nature* 580, 269–273. doi:10.1038/s41586-020-2080-8



- Saha, L. K., Wakasugi, M., Akter, S., Prasad, R., H Wilson, S., Shimizu, N., et al. (2020). Topoisomerase  $\alpha$ -driven repair of uv-induced damage in *ner*-deficient cells. *Proc. Natl. Acad. Sci. U. S. A.* 117, 14412–14420. doi:10.1073/pnas.1920165117
- Schumann, F., Blanc, E., Messerschmidt, C., Blankenstein, T., Busse, A., and Beule, D. (2019). SigsPack, a package for cancer mutational signatures. *BMC Bioinforma.* 20, 450–459. doi:10.1186/s12859-019-3043-7
- Shen, H.-M., Shi, C.-Y., Shen, Y., and Ong, C.-N. (1996). Detection of elevated reactive oxygen species level in cultured rat hepatocytes treated with aflatoxin b1. *Free Radic. Biol. Med.* 21, 139–146. doi:10.1016/0891-5849(96)00019-6
- Steele, C. D., Abbasi, A., Islam, S. M. A., Bowes, A. L., Khandekar, A., Haase, K., et al. (2022). Signatures of copy number alterations in human cancer. *Nature* 606, 984–991. doi:10.1038/s41586-022-04738-6
- Stich, H. F., and Anders, F. (1989). The involvement of reactive oxygen species in oral cancers of betel quid/tobacco chewers. *Mutat. Res.* 214, 47–61. doi:10.1016/0027-5107(89)90197-8
- Sylvester, R. K., Steen, P., M Tate, J., Menta, M., J Petrich, R., Berg, A., et al. (2011). Temozolomide-induced severe myelosuppression: analysis of clinically associated polymorphisms in two patients. *Anticancer. Drugs* 22, 104–110. doi:10.1097/CAD.0b013e3283407e9f
- Zámborszky, J., Szikriszt, B., Gervai, J. Z., Pipek, O., Póti, Á., Krzystanek, M., et al. (2017). Loss of *brca1* or *brca2* markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions. *Oncogene* 36, 746–755. doi:10.1038/ncr.2016.243
- Zanotti, K. J., and Gearhart, P. J. (2016). Antibody diversification caused by disrupted mismatch repair and promiscuous DNA polymerases. *DNA repair* 38, 110–116. doi:10.1016/j.dnarep.2015.11.011