A Bayesian framework to infer and cluster mutational signatures leveraging prior biological knowledge

Elena Buscaroli (1,*), Azad Sadr Haghighi (1,*), Riccardo Bergamin (1), Salvatore Milite (1,2), Edith Natalia Villegas Garcia (1,3), Arianna Tasciotti (3), Alessio Ansuini (3), Daniele Ramazzotti (4), Nicola Calonaci (1), Giulio Caravagna (1,3)

(1) Department of Mathematics, Informatics and Geosciences, University of Trieste, Trieste, Italy.
(2) Centre for Computational Biology, Human Technopole, Milan, Italy.
(3) Area Science Park, Trieste, Italy.
(4) University of Milan-Bicocca, Italy.

(*) these authors contributed equally and are listed alphabetically.

Correspondence: gcaravagna@units.it

Abstract. The concept of mutational signatures, with its promising translational potential, provides key insights into the mutational processes underlying cancer. Current efforts are increasingly directed toward developing comprehensive catalogues of signatures linked to various tumour types and therapeutic responses. However, the existence of multiple catalogues, generated by different groups using distinct methodologies, underscores the need for standardisation across the field, and a cohesive framework that integrates established signatures remains to be fully realised. Here, we introduce a set of Bayesian algorithms that merge predefined signature catalogues with newly identified signatures, offering a systematic approach to expanding existing collections. By leveraging a diverse array of mutational signatures, our method also groups patients based on shared mutational patterns, further enhancing the translational relevance of these catalogues. We demonstrate that this approach enables the identification of both known and novel molecular subtypes across nearly 7,000 samples spanning three major cancers: breast, colon, and lung. Building on prior research, we propose a robust strategy to deepen our understanding of mutational processes in cancer.

Introduction

Clonal dynamics in tumours (McGranahan and Swanton, 2015; Turajlic *et al.*, 2019; Black and McGranahan, 2021) are primarily determined by somatic mutations that accumulate during cell division (Greaves and Maley, 2012; Martincorena *et al.*, 2015; Sottoriva *et al.*, 2015). These mutations can arise in various forms. Single Base Substitutions (SBS) and insertions/deletions (indels) are often caused by random DNA replication errors or impaired mismatch repair mechanisms (Nik-Zainal *et al.*, 2012; Li *et al.*, 2020), while more complex structural variants, such as copy number alterations (CNAs), result from chromosome segregation errors, breakage, and rejoining (Sansregret and Swanton, 2017; Lukow and Sheltzer, 2022). Both endogenous and exogenous mutational processes influence the rate of

these mutational events, and substantial efforts are underway to decipher the mechanisms underlying them.

One of the most notable results is the introduction, in 2013, of the concept of SBS signatures, an idea later extended to dinucleotide-base substitutions (DBS), small indels (ID) and eventually copy number (CN) signatures (Alexandrov *et al.*, 2013, 2020; Macintyre *et al.*, 2018; Steele *et al.*, 2022). In an actual tumour, multiple mutational processes co-exist, depending on the molecular features, the patient's genetic background, and environmental factors, including treatment (Cortés-Ciriano *et al.*, 2021). Understanding these mutational processes has significant translational potential, as it can inform treatment strategies and improve patient outcomes. For instance, mutational signatures associated with impaired mismatch repair can guide the use of immune checkpoint inhibitors in patients with microsatellite instability-high tumours (Le *et al.*, 2015). Similarly, mutational profiles linked to homologous recombination deficiency have been used to identify patients who may benefit from PARP inhibitors in breast and ovarian cancers (Lord and Ashworth, 2016). These examples underscore how insights into mutational mechanisms can directly inform therapeutic decisions, advancing precision medicine in oncology (Alexandrov and Stratton, 2014; Helleday, Eshtad and Nik-Zainal, 2014).

Computational tools are increasingly used to detect mutational signatures from mutation count data derived from whole-exome or whole-genome sequencing (WGS). These tools generally fall into two categories (Omichessan, Severi and Perduca, 2019): those that apply established signal deconvolution algorithms to extract mutational signatures and estimate their prevalence across samples, a process known as de novo signature discovery (Gehring *et al.*, 2015; Rosenthal *et al.*, 2016; Huang, Wojtowicz and Przytycka, 2018; Lal *et al.*, 2021; Degasperi *et al.*, 2022), and those that focus on decomposing mutation data using predefined reference signature catalogues, bypassing the need for de novo discovery (Rosenthal *et al.*, 2016; Díaz-Gay *et al.*, 2023). These methodologies have been instrumental in assembling mutational process catalogues in cancer (Omichessan, Severi and Perduca, 2019), including the development of COSMIC (Catalogue of Somatic Mutations in Cancer), which currently lists 144 mutational signatures — 94 SBS, 11 DBS, 18 ID, and 21 CN signatures (Tate *et al.*, 2019). A subset of these signatures has been experimentally linked to specific endogenous and exogenous processes, shedding light on their underlying mechanisms (Zou *et al.*, 2018). For example, distinct SBS and DBS signatures have been associated with exposures such as tobacco smoking, UV light, platinum-based therapies, and asbestos, showing differences in strand bias, GC context, and flanking bases (Helleday, Eshtad and Nik-Zainal, 2014). Similarly, defective homologous recombination in DNA repair and AID/APOBEC-mediated mutagenesis have given rise to SBS and ID signatures, while chromothripsis and genome duplication events have produced CN signatures. A recent analysis by (Degasperi *et al.*, 2022) used primary and metastatic WGS samples to expand COSMIC with 40 additional SBS and 18 DBS signatures (Degasperi *et al.*, 2022).

In this rapidly evolving field, the challenge lies in harmonising multiple mutational catalogues produced by different research groups and computational tools, a process often relying on similarity metrics such as cosine similarity. However, this task requires extensive human oversight and careful parameter optimization. Ideally, more advanced tools would leverage

existing knowledge to streamline this process. A unified catalogue encompassing various signature types could aid in translating mutational signatures into a stratification metric, enabling the identification of molecular subtypes — groups of patients with similar mutational profiles. Unfortunately, no current tool is capable of performing de novo signature discovery while integrating known signature catalogues, nor can any tool stratify patients based on signature exposure.

We introduce BASCULE (Figure 1A), a Bayesian framework that detects mutational signatures by leveraging an existing catalogue, ensuring that newly identified signatures are statistically distinct from known ones (Figure 1B). Additionally, BASCULE groups patients based on similar mutational profiles by analysing exposure across various signature types. Both algorithms are implemented using stochastic variational inference, built in the Pyro probabilistic programming language (Bingham *et al.*, 2019). This design allows BASCULE to run natively on GPUs, enabling seamless scaling to handle large datasets efficiently. After validating the tool through simulations, we applied it to SBS and DBS signatures from 6,923 samples published by (Degasperi *et al.*, 2022). BASCULE successfully identified well-established molecular subtypes in breast, lung, and colorectal cancers, while also generating new hypotheses about previously unreported co-occurrences of SBS and DBS signatures in specific patient groups. Building on prior research, BASCULE offers a powerful approach to enhance our understanding of mutational processes in cancer and uncover novel molecular insights.

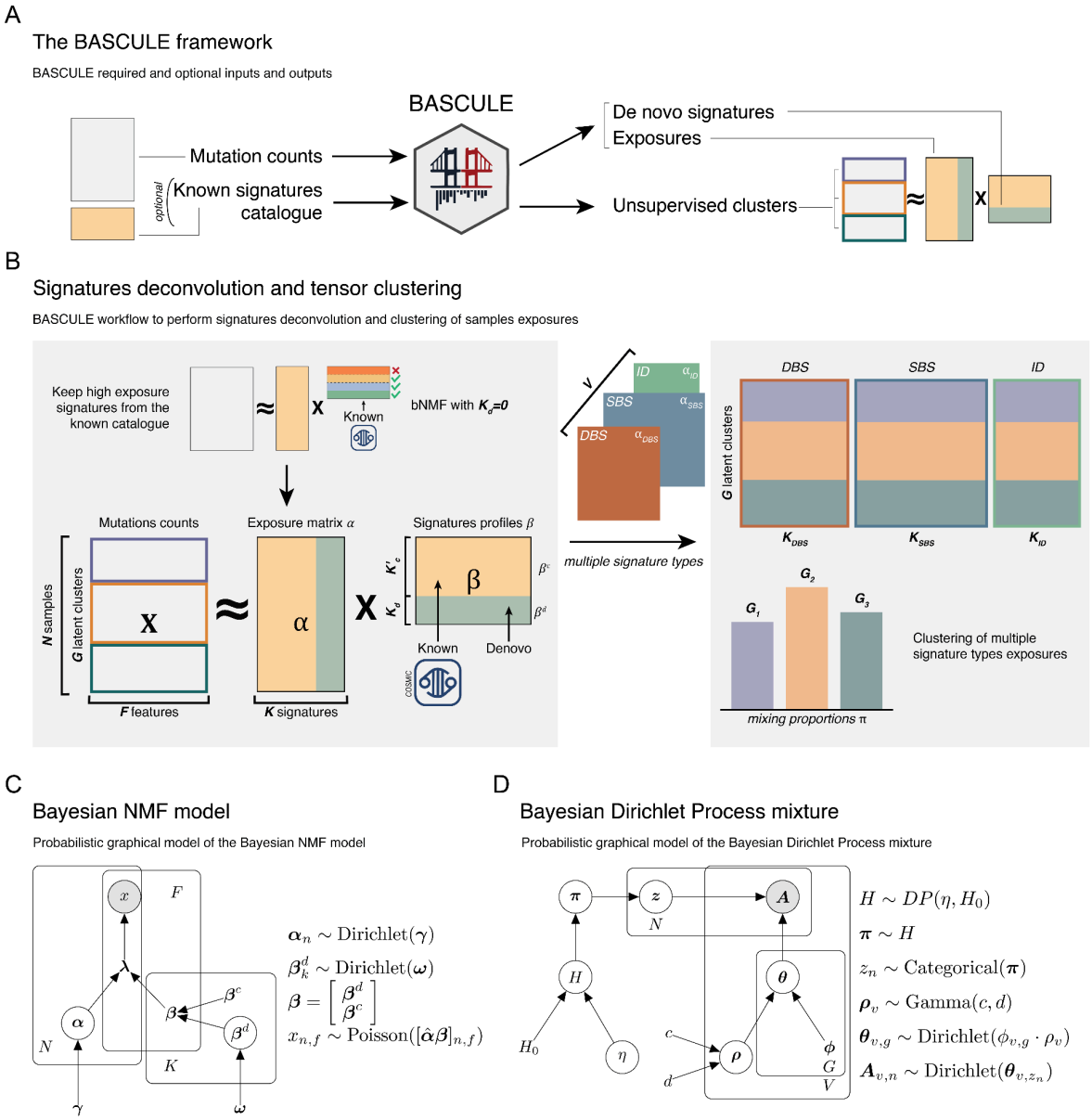


Figure 1. A. BASCULE decomposes a mutation counts matrix into signatures and their exposures while leveraging prior information; then, it clusters samples according to their signatures. B. BASCULE performs signatures deconvolution with a two-step inference process. First, it identifies relevant signatures from a catalogue of known signatures via Bayesian non-negative matrix factorisation. Then, it performs de novo signatures discovery, augmenting the catalogue signatures. BASCULE can deconvolute any signature type (SBS, DBS, IDS, CNS) and use their exposures to cluster the samples. C. The Bayesian non-negative matrix factorisation model: the observed mutation counts matrix X follows a Poisson distribution, whose rate is defined by the product of the latent exposures α and signatures β matrices (Methods). D. BASCULE clusters via the Bayesian Dirichlet Process mixture model. The tensor of exposure matrices is distributed as a Dirichlet, along with its latent group-specific concentration θ. Samples are assigned to the cluster, maximising the posterior distribution (Methods).

Results

Bayesian inference of mutational signatures using a catalogue

BASCULE has a probabilistic model (Figure 1C) that extracts a set $\boldsymbol{\beta}$ of mutational signatures from data of N patients, along with signature exposures $\boldsymbol{\alpha}$. The main novelty is that $\boldsymbol{\beta}$ is a combination of pre-specified catalogued signatures $\boldsymbol{\beta}^c$, and de novo signatures $\boldsymbol{\beta}^d$; signatures in $\boldsymbol{\beta}^d$ differ, probabilistically, from $\boldsymbol{\beta}^c$. The model is general and works with any signature type (SBS, DBS, ID or CN). To extract signatures, we use a Bayesian non-negative matrix factorization (bNMF) approach that factorises $\boldsymbol{X}$, a matrix of non-negative values for the observed counts detected in patients, as

$$(1) \qquad \boldsymbol{X} \approx \boldsymbol{\alpha}\boldsymbol{\beta} = \underbrace{[\boldsymbol{\alpha}^d \quad \boldsymbol{\alpha}^c]}_{\boldsymbol{\alpha} \in \mathbb{R}^{N \times K}} \underbrace{\begin{bmatrix} \boldsymbol{\beta}^d \\ \boldsymbol{\beta}^c \end{bmatrix}}_{\boldsymbol{\beta} \in \mathbb{R}^{K \times F}}$$

such that $K = K^d + K^c$. The matrices of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ report multinomial random variables, that is

$$\sum_{k=1}^{K} \alpha_{n,k} = 1 \wedge \sum_{f=1}^{F} \beta_{k,f} = 1$$

The model likelihood is Poisson

$$(2) \qquad p(\boldsymbol{X}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{n=1}^{N} \prod_{f=1}^{F} \text{Poisson}(x_{n,f}|\lambda_{n,f})$$

with rate defined by the reconstructed count matrix $\boldsymbol{\lambda} = \hat{\boldsymbol{\alpha}}\boldsymbol{\beta}$, where $\hat{\alpha}_{n,k} = \alpha_{n,k} \sum_{f=1}^{F} x_{n,f}$ are the number of mutations linked with signature k in patient n.

In this model, the columns F are features that depend on the type of signature (e.g., for SBS and DBS these are 96 and 78 strand-agnostic substitution contexts, respectively), and $\boldsymbol{X}$ is decomposed into two pairs of exposure and signature matrices, one with K^c signatures from the catalogue, and one with K^d de novo signatures. We note that in the learning process, the signatures from the catalogue $\boldsymbol{\beta}^c$ are not learnt from the data, i.e., they are held fixed based on the catalogue. Instead, the exposure vectors are learnt for both catalogue and de novo signatures. The overall bNMF joint distribution, over signatures and exposures, is

$$(3) \qquad p(\boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{X}|\boldsymbol{\alpha}, \boldsymbol{\beta}) p(\boldsymbol{\alpha}, \boldsymbol{\beta}^d) =$$
$$= \prod_{n=1}^{N} \prod_{f=1}^{F} p(x_{n,f}|\lambda_{n,f}) \prod_{n=1}^{N} p(\boldsymbol{\alpha}_n) \prod_{k=1}^{K} p(\boldsymbol{\beta}_k^d)$$

with $\boldsymbol{\alpha}_n$ and $\boldsymbol{\beta}_k^d$ following a Dirichlet prior with a concentration that ensures sparsity among the de novo and catalogue signatures (Online Methods). Our approach precisely forces the model to sample from features with low density in the catalogue, ensuring that de novo signatures account for the signal unexplained by catalogue signatures. To optimise the number of de novo signatures, we minimise the Bayesian Information Criterion computed using the maximum a posteriori estimates of the model parameters. Moreover, we implement a final heuristic to drop de novo signatures that are linear combinations of other de novo or catalogue signatures. This heuristic helps retain signals that differ from one another, also when considering complex combinations of signatures. Details on the overall formulation are in the Online Methods.

Non-parametric clustering using multiple types of mutational signatures

Patterns of similar $\boldsymbol{\alpha}$ across patients can unravel distinct molecular subtypes, suggesting that patients could be clustered by exposures. The main difference from signature detection, however, is that subtypes should be detected by considering distinct types of signatures (i.e., SBS, DBS, etc.), and not independent bNMF tasks. BASCULE provides a Bayesian non-parametric model (Figure 1D) to cluster N patients into $G \geq 1$ groups with similar exposure patterns; our model is the first model to use a tensor-based clustering algorithm that leverages a Dirichlet process (Blei and Jordan, 2006) to i) combine $V \geq 1$ signature types and ii) it automatically determines the number of clusters.
BASCULE represents the input as a tensor $\boldsymbol{A}$ with dimensions $N \times V \times K'$, where $K' = \max_{v=1,\cdots,V} K^{(v)}$ for $K^{(v)}$ the number of columns (i.e., signatures) of the v-th type of signature. A zero-padding strategy that flattens out the difference in the number of dimensions per signature type is required to assemble $\boldsymbol{A}$. Similarly to the bNMF step, the exposures are modelled with a Dirichlet, leading to a mixture likelihood

$$(4) \qquad p(\boldsymbol{A}|\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\theta}) = \prod_{v=1}^{V}\prod_{n=1}^{N}\sum_{g=1}^{G}\pi_g \mathrm{Dirichlet}(\boldsymbol{A_{v,n}}|\boldsymbol{\theta_{v,g}})$$

and the factored joint probability

$$(5) \qquad p(\boldsymbol{A},\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\rho}) = p(\boldsymbol{A}|\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\rho})p(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\rho}) =$$
$$= p(\boldsymbol{A}|\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\rho})p(\boldsymbol{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})\prod_{v=1}^{V}\prod_{g=1}^{G}p(\boldsymbol{\theta}_{v,g}|\rho_v)\prod_{v=1}^{V}p(\rho_v)$$

BASCULE defines the distribution over the mixture components and their weights as a Dirichlet Process $DP(\eta, H)$, with concentration η and base distribution H such that $\boldsymbol{\theta}_g$ is sampled from a Dirichlet. In the implementation, we use a stick-breaking construction to sample from the Dirichlet Process (Ishwaran and James, 2001). In so doing, the model automatically determines the optimal number of clusters based on the concentration η. After clustering, BASCULE merges clusters with similar centroids; this heuristic helps retain

clusters with different exposures to one another. The derivations of the model are available as Online Methods.

Synthetic simulations

We conducted realistic simulations by sampling from the generative model of BASCULE (Figure 1C,D) and measured performance for various combinations of the number of patients N, the number of groups G (molecular subtypes, and mutational processes K (example dataset in Figure 2A). Our datasets have 150, 500 or 1000 samples, 1, 3, or 6 groups, and incorporate up to 13 SBS COSMIC signatures. For each configuration, we generated 30 independent datasets.

We compared the performance of the bNMF model to SigProfiler (Islam *et al.*, 2022), the most widely used tool for de novo signature discovery, and to SparseSignatures (Lal *et al.*, 2021), a method that uses LASSO regularisation with a background signature. Notably, SigProfiler has been used to develop COSMIC, and SparseSignatures, like BASCULE, utilises prior biological information to achieve a parsimonious explanation of the data (Online Methods).

Results from model recall (Figure 2B), i.e., the proportion of all actual positives classified correctly as positives, indicate that the model can accurately identify the true number of signatures better than competitors. Moreover, the quality of the deconvolution was high, as measured with the mean squared error (MSE) of the reconstructed mutation count matrix (Figure 2D), and of the cosine similarity (CS) of the estimated signatures and exposures (Figure 2E,F). In our tests, all methods reconstruct the original counts with similar accuracy (BASCULE MSE $6.8e^{-3}$, SigProfiler MSE $6.3e^{-3}$, SparseSignatures MSE $6.9e^{-3}$). Still, BASCULE achieves higher CS and lower variability than SigProfiler and SparseSignatures, for signatures (BASCULE CS 0.97, SigProfiler CS 0.93, SparseSignatures CS 0.95) and exposures (BASCULE CS 0.89, SigProfiler CS 0.79, SparseSignatures CS 0.77). This indicates that our model is precise in reconstructing the signature profile and determining the contribution of each signature to a sample. As expected, performance increases across all analyses and methods as the number of samples increases and the number of signatures decreases (Supplementary Figures S1 and S2).

Moreover, as BASCULE is the only method to cluster signature exposures across multiple signature types, we evaluated its performance against a baseline k-means (Ahmed, Seraj and Islam, 2020) algorithm applied to the exposure vectors (Online Methods). Clustering performance was assessed using the normalised mutual information (NMI) between simulated groups and retrieved cluster labels (Figure 2C). BASCULE demonstrated superior performance than k-means, achieving an NMI of 0.96 compared to 0.67 and supporting the practical advantage of employing a Dirichlet-based model.

Finally, we measured the runtime of BASCULE on these synthetic tests and compared it to existing methods, at least for signature deconvolution (bNMF step) (Figure 2G and Supplementary Figure S3). In this case, thanks to GPU deployment, BASCULE was significantly faster than SigProfiler and to SparseSignatures, which required from 15 to 90

times more time than our approach. In particular, for large cohorts with 1000 samples, SigProfiler took 45 times longer (standard deviation 10), while SparseSignatures required 92 times more time (standard deviation 58) for inference. These results suggest that BASCULE is more likely to scale efficiently to analyse large cohorts.

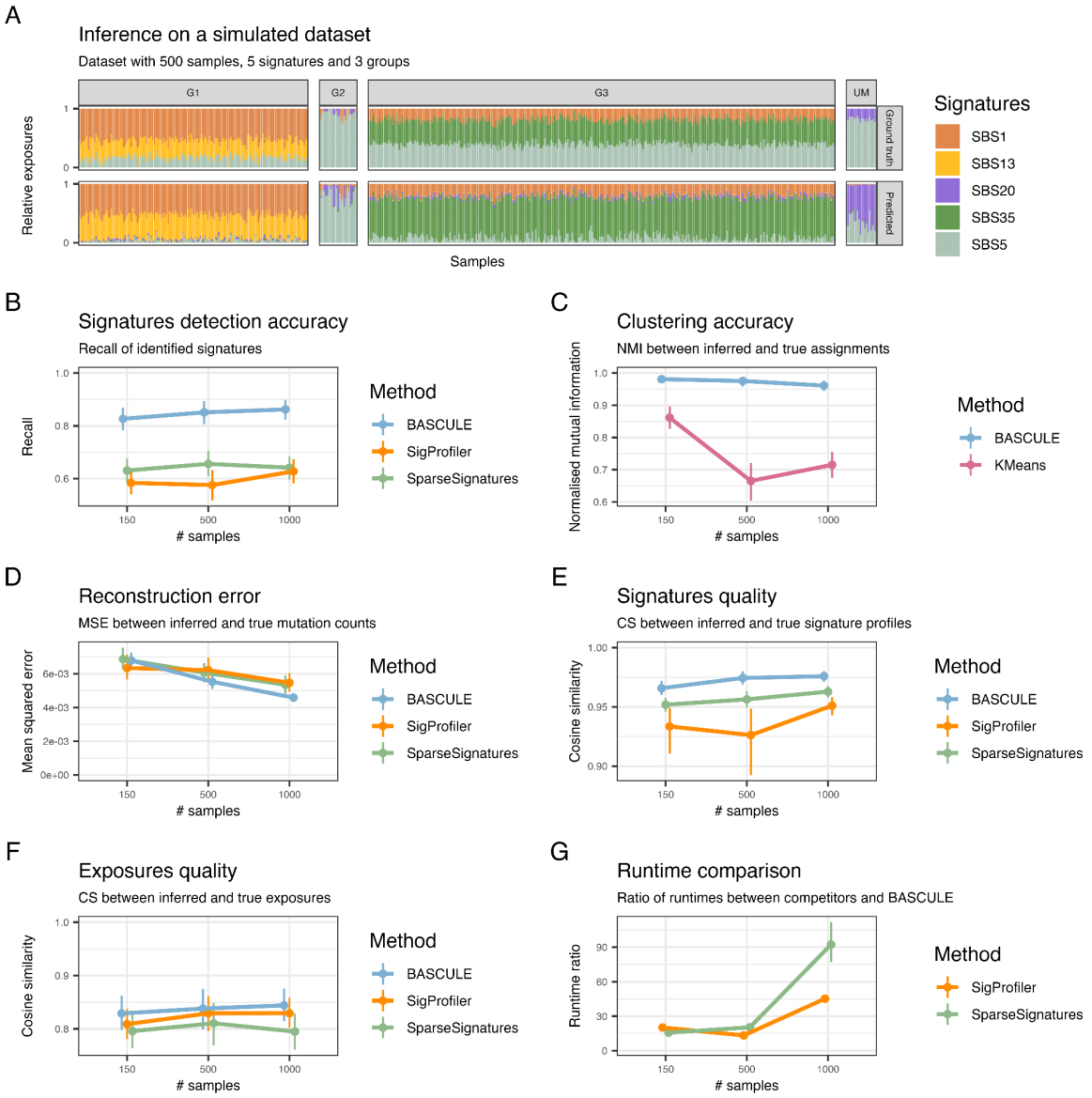


Figure 2. A. Example of inference on a simulated dataset. The generated dataset (top panel) consists of 500 samples, 3 groups, and 5 SBS signatures, including the profiles of COSMIC SBS13, SBS20, and SBS36 as de novo. The results (bottom panel) demonstrate that BASCULE successfully retrieves the signatures and learns similar exposures (cosine similarity = 0.94). Moreover, BASCULE can group the samples in G1, G2 and G3 with high accuracy (NMI = 0.93), with only a few unmatched samples (UM). B. Performance evaluation of BASCULE on synthetic datasets with an increasing number of input samples, compared to SigProfiler and SparseSignatures. Accuracy was assessed using recall, the number of true positives divided by the number of true positives plus the number of false negatives. C. Clustering accuracy was evaluated using the normalised mutual information, compared to K-means clustering of the exposure vectors. D. Reconstruction error on the input mutation counts matrix. E, F. Quality of signatures and exposures measured by cosine similarity between true and inferred parameters. G. Runtime ratio for SigProfiler and SparseSignatures relative to BASCULE during inference.

Molecular subgroups across three main human cancers

We used BASCULE to analyse WGS data recently released by (Degasperi *et al.*, 2022), which carried out one of the most significant attempts to establish a catalogue of SBS and DBS signatures beyond COSMIC (Tate *et al.*, 2019). One of the critical advances of (Degasperi *et al.*, 2022)et al. is to establish, using multi-step heuristics, which signatures are common across patients, as opposed to rare ones. Degasperi *et al.* used 12,222 WGS cancer samples collected through the Genomics England 100,000 Genomes Project (Turnbull *et al.*, 2018) and available at Genomics England (Sosinsky *et al.*, 2024) to discover 40 SBS and 18 DBS mutational signatures in addition to COSMIC, for a total of 135 SBS and 180 DBS signatures in 19 tumour types. Additionally, Degasperi *et al.* compared their results with 18640 samples from two external cohorts: the International Cancer Genome Consortium (ICGC/TCGA PCAWG Consortium, 2020) and the Hartwig Medical Foundation (Martínez-Jiménez *et al.*, 2023). In this paper, we used all data from Degasperi *et al.* in combination with the COSMIC and the Degasperi *et al.* catalogues, focusing on 6923 samples from breast, lung, and colorectal cancers.

For each tumour type, we ran BASCULE with a minimal catalogue curated from the common signatures reported by Degasperi *et al.,* using the COSMIC version if available (details in the Online Methods). In brief, we selected 10 SBS and 3 DBS signatures for breast, 11 SBS and 4 DBS for lung, and 13 SBS and 4 DBS for colorectal cancers. After bNMF for up to 25 de novo signatures, we used cosine similarity to map de novo signatures to the full COSMIC and Degaspari *et al.* catalogues (Supplementary Figure S4); this ideally would allow us to detect less common signatures in our data. Ultimately, we performed clustering and used the centroid and patient exposures to derive meaningful insights from clustering assignments.

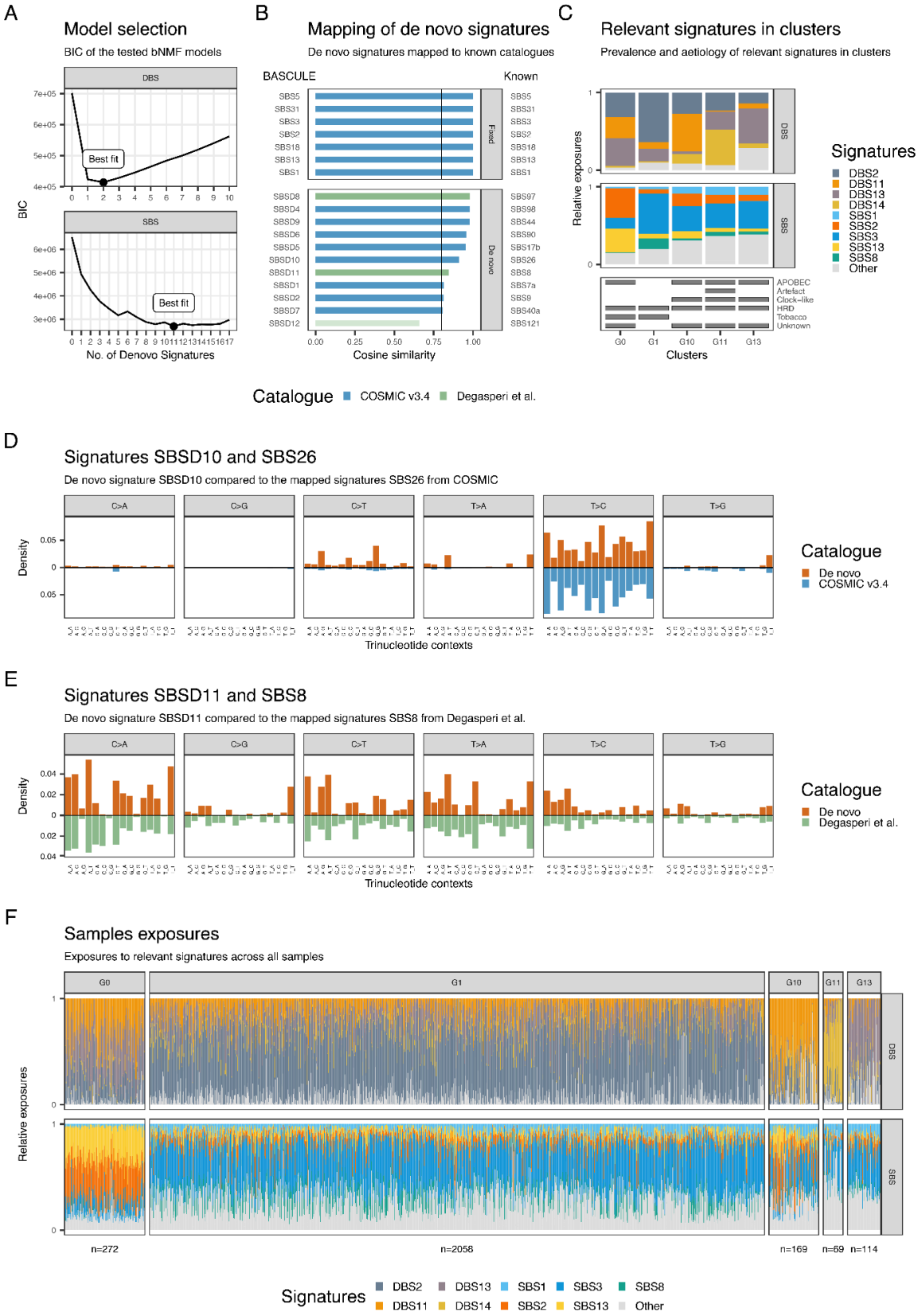


Figure 3. A. Optimal number of de novo SBS and DBS signatures based on BIC criteria, running BASCULE on 2682 breast cancers with a starting catalogue of known signatures (SBS1, SBS2, SBS3, SBS5, SBS8, SBS13, SBS17, SBS18, SBS31, and SBS127 for SBS and DBS11, DBS13, and DBS20 for DBS) from COSMIC. B. Mapping for the inferred de novo SBS signatures to known catalogues. SBSD4 and SBSD8 are mapped to SBS98 and SBS97,

respectively. C. The exposure of the centroids for the clusters where only the most relevant signatures in each cluster are present. It includes clusters with more than 20 patients with their corresponding aetiology at the bottom. The full plot is accessible at (Supplementary Figure S5). D. De novo signature SBSD10 compared to the mapped signature SBS26 from COSMIC v3.4. E. De novo signature SBSD11 compared to the mapped signature SBS8 from Degasperi et al. F. Exposures of relevant signatures for clusters with more than 20 samples.

Breast cancer patients (n = 2682)

From 2682 breast cancer samples, we determined 23 signatures (13 de novo) and 5 clusters (Figure 3A; Supplementary Figure S5). Upon post-inference mapping (Supplementary Figure S4,S6), 6 out of 11 de novo signatures were mapped to COSMIC, and 3 to the Degasperi *et al.* catalogue, with two signatures flagged as rare in the original study (Figure 3B).

We found two major biological interpretations for the 5 clusters, joining SBS and DBS signatures to explain all groups (Figure 3C; Supplementary Figure S7). These interpretations are facilitated by the potential alignment of some de novo signatures with already established known signatures (Figure 3D,E). None of the clusters in breast cancer is featured with de novo signatures. Cluster G1 (n=2058, Figure 3F), the largest among all, showed a strong presence of SBS3, SBSD11 (mapped to SBS8 from Degasperi et al.), DBS2 and DBS13. SBS3 and DBS13 are caused by homologous recombination deficiency (HRD); DBS2, instead, is linked to smoking in some cancers but is often observed also in cancers not directly linked with tobacco exposure. Based on these prevalent signatures, cluster G1 can be associated with triple-negative breast cancers (Li *et al.*, 2024). In cluster G0 (n=272, Figure 3F), the dominant mutational signatures (SBS2 and SBS13, and evenly DBS11, DBS13 and DBS2) are associated with APOBEC activity, linking this group to HER2 breast cancers (DiMarco *et al.*, 2022). Clusters G10 (n=169), G11 (n=69), and G13 (n=114) share similar SBS patterns, including exposure to SBS1, SBS2, SBS3, SBS11, and SBS13, pointing to a mix of APOBEC activity, HRD, and ageing-related mutational processes. BASCULE can differentiate these three clusters based on DBS exposure. In particular, G10 is strongly linked to DBS2 and DBS11, G11 primarily to DBS14 and G13 to DBS13 and DBS20. Therefore, G10 can be associated with APOBEC activity, G13 with HRD-related processes, and G11, the smallest cluster, with DBS14, a signature reported as artifactual in COSMIC. However, the presence of DBS13 in G11, suggests that this cluster is also linked to HRD. We remark that these clusters would not have been detectable by SBS patterns alone.

Lung cancer patients (n = 1396)

From 1396 lung cancer samples, we identified 7 distinct clusters characterised by 21 mutational signatures (9 de novo) (Figure 4A; Supplementary Figure S8). Following the mapping stage (Figure 4B; Supplementary Figure S4,S9), five de novo signatures were found in COSMIC and two in the Degasperi *et al.* study (one flagged as rare). As for breasts, we observed that by leveraging both SBS and DBS (Figure 4C), as well as mapping some de novo signatures to known signatures (Figure 4D,E), we uncover potential biological differences missing from SBS data alone. None of the clusters in lung cancer is featured with de novo signatures (Supplementary Figure S9). The exposure patterns in lung cancer (Figure 4F; Supplementary Figure S10) seemed more complex than in breast cancer , with many clusters sharing a significant number of signatures and similar exposures.

Cluster G1 (n=1204), the largest cluster, was associated with heavy smokers based on moderate exposure to SBS4, and pronounced exposure to DBS2 (Chen *et al.*, 2020). Another cluster, G3 (n=31), harboured SBS31 and DBS5, two signatures linked to chemotherapy exposure. Chemotherapy was also linked to cluster G13 (n=8), but this time because of SBS17b, presenting a unique pattern among the clusters. Both G3 and G13 contained, therefore, patients exposed to chemotherapy. Cluster G14 (n=23), compared to others, showed a slightly elevated presence of SBS5 and significant exposure to DBS1. As for breast cancers, the causative process for DBS1 remains unknown, whereas in other tumours, this signature can be linked with UV light exposure (Mata *et al.*, 2022). Finally, clusters G0 (n=119), G12 (n=4), and G8 (n=7) were similar in terms of SBS patterns but differed for DBS profiles; namely, G0 was characterised by DBS13, related to HRD, while G12 exhibited DBS6 and G8 DBS20, which have unknown aetiology (Online Methods).

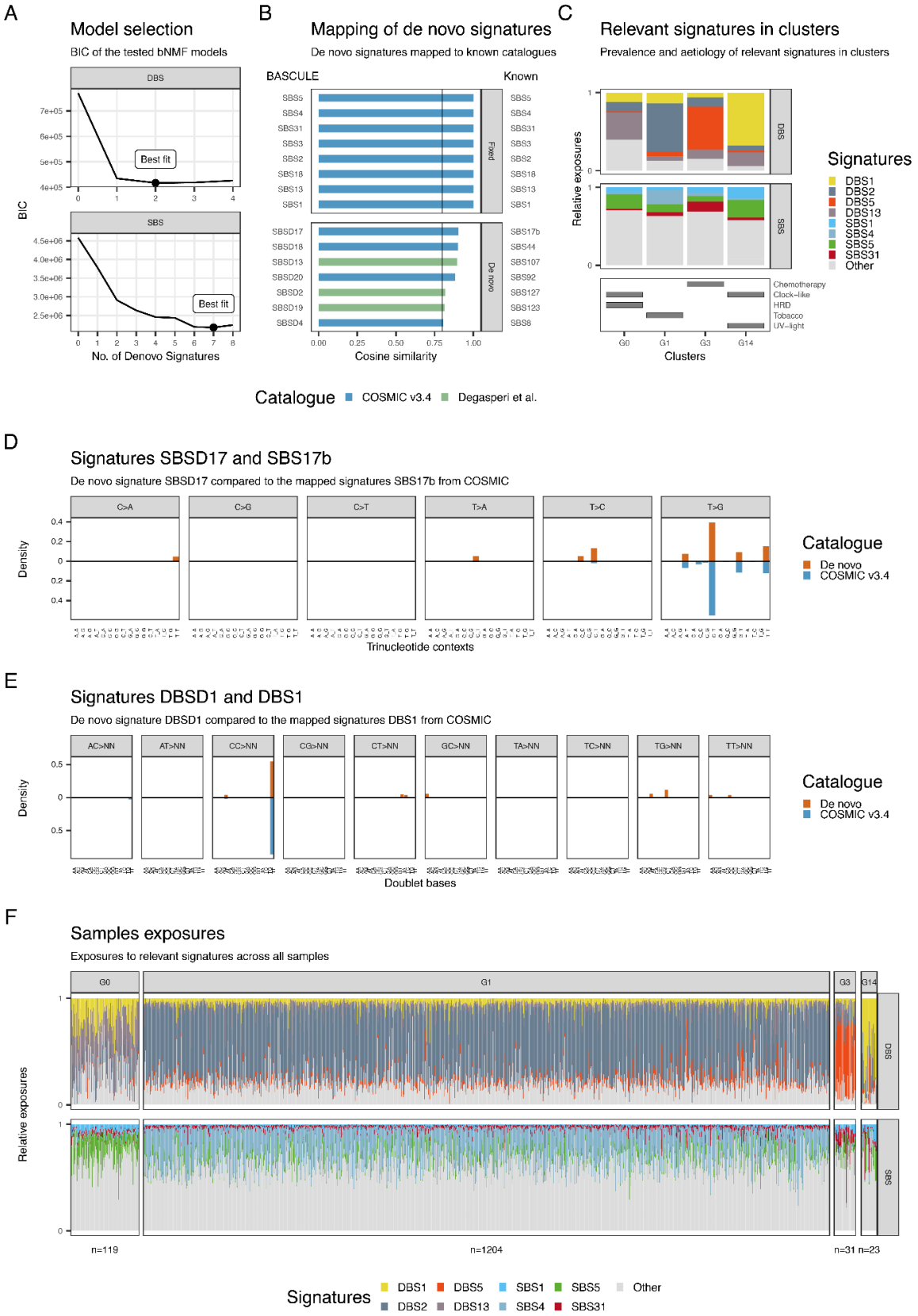


Figure 4. A. Optimal number of de novo SBS and DBS signatures based on BIC criteria, running BASCULE on 1396 lung cancers with a starting catalogue of known signatures (SBS2, SBS4, SBS1, SBS5, SBS3, SBS8, SBS13, SBS17, SBS18, SBS31, and SBS92 for SBS and DBS2, DBS5, DBS13, and DBS20 for DBS) from COSMIC. B. Mapping for the inferred de novo SBS signatures to known catalogues. SBSD13 and SBSD18 are mapped to

SBS107 and SBS44, respectively. C. The exposure of the centroids for the clusters where only the most relevant signatures in each cluster are present in the plot. It includes clusters with more than 20 patients with their corresponding aetiology at the bottom. The full plot is accessible at (Supplementary Figure S8). D. De novo signature SBSD17 compared to the mapped signature SBS17b from COSMIC v3.4. E. De novo signature DBSD1 compared to the mapped signature DBS1 from COSMIC v3.4. F. Exposures of relevant signatures for clusters with more than 20 samples.

Colorectal cancer patients (n = 2845)

From n=2845 colorectal cancer patients, BASCULE identified 32 signatures and 8 clusters of patients (Figure 5A; Supplementary Figure S11). Twenty signatures were de novo, but nine were mapped to COSMIC, and four to the Degasperi *et al.* study, with three rare signatures (Figure 5B; Supplementary Figure S4, S12). Among our case studies, this dataset has the highest number of clusters and signatures (Figure 5C) and mapping some de novo signatures to known signatures provides additional insights into the explanation of the clusters identified by our model (Figure 5D,E). The remaining unmapped de novo signatures (Supplementary Figure S12) did not feature in defining clusters.

At baseline (Supplementary Figure S13), we could retrieve the canonical split among microsatellite stable (MSS) and unstable (MSI) colorectal cancers, including polymerase-eta (POLE) subtypes (Díaz-Gay and Alexandrov, 2021; Battuello *et al.*, 2024; Cornish *et al.*, 2024). Clusters G1 (n=2,041), G3 (n=112), G6 (n=95), and G12 (n=130) displayed similar patterns of mutational signatures and exposures (Figure 5F). These clusters prominently featured a BASCULE de novo SBSD8 signature. Quality control on all de novo signatures revealed that this signature is neither a linear combination of other signatures nor similar to any known signature (minimum tested cosine similarity 0.8). In addition, we found ageing-related signatures SBS1 and SBS5 alongside with, which are expected in most patients. Additional signatures were SBS44 (mismatch repair deficiency), SBS35 (chemotherapy), SBS3 (HRD), and SBS18 (reactive oxygen species). The tensor-based clustering approach distinguished these clusters by DBS signatures. Cluster G6 was characterised by hyper-exposure to the DBS8 signature, G3 by the unique DBS5 signature linked to chemotherapy, G12 by DBS13 associated with HRD, and G1 by the predominant DBS25 signature of unknown aetiology. Cluster G10 (n=412) reported exposure to the BASCULE de novo signatures SBSD12 and SBSD7. SBSD12 is mapped to SBS57 related to sequencing artefacts, and SBSD7 is mapped to signature SBS44 related to MSI status. In the DBS context, this cluster shows hyper-exposure to DBS14, which is considered artifactual in COSMIC. Cluster G9 (n=28) showed high exposure to SBS10a and DBS3, both indicative of the POLE status. Cluster G0 (n=23) featured the SBS18 signature and little of SBS3, two signatures linked with reactive oxygen species aetiology. In cluster G5 (n=4), the most relevant signature was SBS15, linked to mismatch repair deficiency, alongside notable exposure to the de novo SBSD7 signature identified by BASCULE which is mapped to SBS44 from COSMIC catalogue and related to MSI status (Online Method).

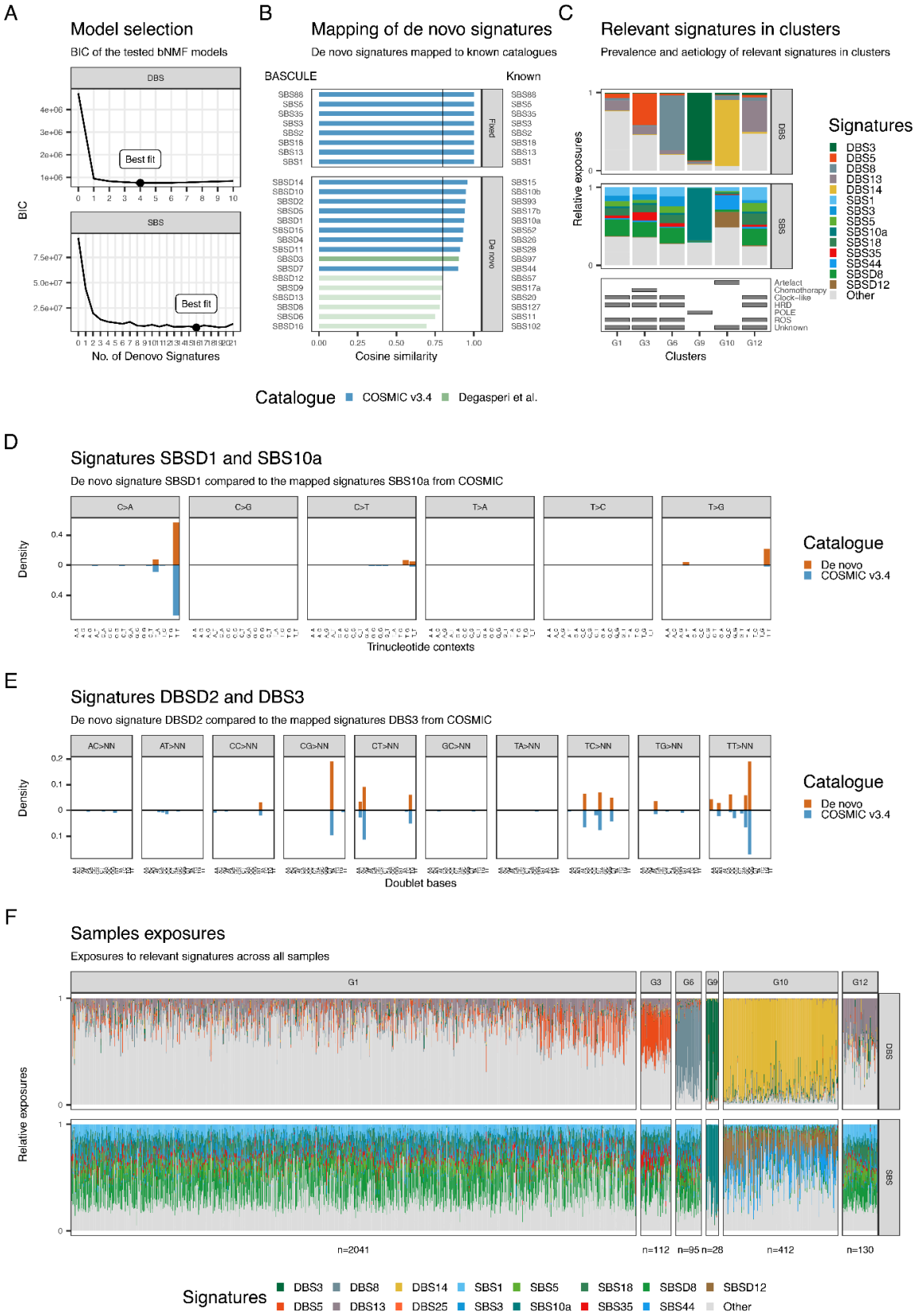


Figure 5. A. Optimal number of de novo SBS and DBS signatures based on BIC criteria, running BASCULE on 2845 colorectal cancers with a starting catalogue of known signatures (SBS1, SBS2, SBS3, SBS5, SBS8, SBS13, SBS17, SBS18, SBS35, SBS88, SBS93, SBS121, SBS157 for SBS and DBS2, DBS5, DBS13, DBS20 for DBS) from COSMIC. B. Mapping for the inferred de novo SBS signatures to known catalogues. SBSD3 and SBSD14 are mapped to

SBS97 and SBS15, respectively. C. The exposure of the centroids for the clusters where only the most relevant signatures in each cluster are present in the plot. It includes clusters with more than 20 patients with their corresponding aetiology at the bottom. The full plot is accessible at (Supplementary Figure S11). D. De novo signature SBSD1 compared to the mapped signature SBS10a from COSMIC v3.4. E. De novo signature DBSD2 compared to the mapped signature DBS3 from COSMIC v3.4. F. Exposures of relevant signatures for clusters with more than 25 samples.

Discussion

Over the last few years, large-scale analyses revealed ubiquitous mutational signatures across distinct cancer types and, for this reason, signatures have become a pillar concept of modern cancer genomics. The possibility of linking signatures with endogenous and exogenous processes that have a known aetiology can revolutionise our understanding of how mutations drive cancers (Zou *et al.*, 2018; Kucab *et al.*, 2019), and translate (Le *et al.*, 2015; Lord and Ashworth, 2016; Rogozin *et al.*, 2018). The field is quickly advancing toward developing signature catalogues that enable the stratification of human tumours into distinct molecular subtypes, with COSMIC standing as the leading example of this progress.

BASCULE addresses two major limitations of existing frameworks (Omichessan, Severi and Perduca, 2019): the inability to incorporate known catalogues in the search for new signatures, and the lack of clustering capabilities based on signature data. By building on top of existing catalogues, our approach can be used to create new biological knowledge incrementally. By identifying groups of patients with similar signatures, our model enhances the translation potential of signatures. We created a minimal version of two known catalogues to test these ideas on three widespread human cancers, leveraging a GPU-based implementation of the Bayesian models of BASCULE. From about 7000 WGS samples, our approach retrieved off catalogue signatures that were never shown to our tool, but that were present in the extended version of the input catalogues. Notably, some were flagged as rare using dedicated heuristics in (Degasperi *et al.*, 2022), suggesting that BASCULE is sensitive enough to detect signatures even when involving few patients.

On the same data, using the non-parametric clustering model of BASCULE, we retrieved molecular subtypes that are common in breast, lung and colorectal cancers. Our model autonomously detected groups that include triple-negative and hormone receptor-positive breast cancers, microsatellite-stable and unstable colorectal cancers, and lung cancers associated with tobacco exposure. In this analysis we observed a key feature of our tool, which is the possibility of leveraging multiple type of signatures for clustering. By jointly leveraging SBS and DBS signatures, BASCULE did uncover new biological subtypes proving that combined signatures stratify groups at a resolution that is finer than commonly considered. While in the future these results would require further experimental validation, this work shows the advantage of a principled method capable of including existing biological knowledge and offering advanced downstream analysis, such as clustering, which can be used to assemble and translate to the clinic reliable catalogues of mutational signatures.

Author's contribution

EB, ASH, RB, SM and GC conceptualised and developed the model. EB, ASH, RB and SM implemented the tool. EB, ASH and EV carried out simulations and analysed real data. EB, ASH, DR, NC and GC interpreted results. EB, ASH and GC drafted the manuscript, which all authors approved in final form. GC supervised and coordinated this project.

Acknowledgements

The research leading to these results has received funding from AIRC under MFAG 2020 - ID. 24913 project – P.I. Caravagna Giulio. We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 1409 published on 14.9.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union – NextGenerationEU– CUP J53D23015060001.

Competing interests

The authors declare no competing interests.

Data availability

The open-source BASCULE R package can be accessed at

https://github.com/caravagnalab/bascule.

The input datasets utilised in this manuscript, which were derived from the Genomics England (GEL), International Cancer Genome Consortium (ICGC), and Hartwig Medical Foundation (HMF) cohorts, are publicly accessible at (Degasperi *et al.*, 2022).

Additionally, all source code used for the analysis and reproduction of all the results described in this manuscript is available at

https://doi.org/10.5281/zenodo.13273336

These resources provide researchers with the tools and data to replicate this study.

References

Ahmed, M., Seraj, R. and Islam, S.M.S. (2020) 'The k-means Algorithm: A Comprehensive Survey and Performance Evaluation', *Electronics*, 9(8), p. 1295.

Alexandrov, L.B. *et al.* (2013) 'Signatures of mutational processes in human cancer', *Nature*, 500(7463), pp. 415–421.

Alexandrov, L.B. *et al.* (2020) 'The repertoire of mutational signatures in human cancer', *Nature*, 578(7793), pp. 94–101.

Alexandrov, L.B. and Stratton, M.R. (2014) 'Mutational signatures: the patterns of somatic mutations hidden in cancer genomes', *Current opinion in genetics & development*, 24(100), pp. 52–60.

Battuello, P. *et al.* (2024) 'Mutational signatures of colorectal cancers according to distinct computational workflows', *Briefings in bioinformatics*, 25(4). Available at: https://doi.org/10.1093/bib/bbae249.

Bingham, E. *et al.* (2019) 'Pyro: Deep Universal Probabilistic Programming', *Journal of machine learning research: JMLR*, 20(28), pp. 1–6.

Black, J.R.M. and McGranahan, N. (2021) 'Genetic and non-genetic clonal diversity in cancer evolution', *Nature reviews. Cancer*, 21(6), pp. 379–392.

Blei, D.M. and Jordan, M.I. (2006) 'Variational inference for Dirichlet process mixtures', *Bayesian analysis* , 1(1), pp. 121–143.

Chen, Z. *et al.* (2020) 'From tobacco smoking to cancer mutational signature: a mediation analysis strategy to explore the role of epigenetic changes', *BMC cancer*, 20(1), p. 880.

Cornish, A.J. *et al.* (2024) 'The genomic landscape of 2,023 colorectal cancers', *Nature*, 633(8028), pp. 127–136.

Cortés-Ciriano, I. *et al.* (2021) 'Computational analysis of cancer genome sequencing data', *Nature reviews. Genetics* [Preprint]. Available at: https://doi.org/10.1038/s41576-021-00431-y.

Degasperi, A. *et al.* (2022) 'Substitution mutational signatures in whole-genome–sequenced cancers in the UK population', *Science*, 376(6591), p. abl9283.

Díaz-Gay, M. *et al.* (2023) 'Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment', *Bioinformatics* , 39(12). Available at: https://doi.org/10.1093/bioinformatics/btad756.

Díaz-Gay, M. and Alexandrov, L.B. (2021) 'Unraveling the genomic landscape of colorectal cancer through mutational signatures', *Advances in cancer research*, 151, pp. 385–424.

DiMarco, A.V. *et al.* (2022) 'APOBEC Mutagenesis Inhibits Breast Cancer Growth through Induction of T cell-Mediated Antitumor Immune Responses', *Cancer immunology research*, 10(1), pp. 70–86.

Gehring, J.S. *et al.* (2015) 'SomaticSignatures: inferring mutational signatures from single-nucleotide variants', *Bioinformatics* , 31(22), pp. 3673–3675.

Greaves, M. and Maley, C.C. (2012) 'Clonal evolution in cancer', *Nature*, 481(7381), pp. 306–313.

Helleday, T., Eshtad, S. and Nik-Zainal, S. (2014) 'Mechanisms underlying mutational signatures in human cancers', *Nature reviews. Genetics*, 15(9), pp. 585–598.

Huang, X., Wojtowicz, D. and Przytycka, T.M. (2018) 'Detecting presence of mutational signatures in cancer with confidence', *Bioinformatics* , 34(2), pp. 330–337.

ICGC/TCGA PCAWG Consortium (2020) 'Pan-cancer analysis of whole genomes', *Nature*, 578(7793), pp. 82–93.

Ishwaran, H. and James, L.F. (2001) 'Gibbs Sampling Methods for Stick-Breaking Priors', *Journal of the American Statistical Association*, 96(453), pp. 161–173.

Islam, S.M.A. *et al.* (2022) 'Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor', *Cell genomics*, 2(11), p. None.

Kucab, J.E. *et al.* (2019) 'A Compendium of Mutational Signatures of Environmental Agents', *Cell*, 177(4), pp. 821–836.e16.

Lal, A. *et al.* (2021) 'De novo mutational signature discovery in tumor genomes using SparseSignatures', *PLoS computational biology*, 17(6), p. e1009119.

Le, D.T. *et al.* (2015) 'PD-1 Blockade in Tumors with Mismatch-Repair Deficiency', *The New England journal of medicine*, 372(26), pp. 2509–2520.

Li, R.Q. *et al.* (2024) 'Genomic characterization reveals distinct mutational landscapes and therapeutic implications between different molecular subtypes of triple-negative breast cancer', *Scientific reports*, 14(1), p. 12386.

Li, Y. *et al.* (2020) 'Patterns of somatic structural variation in human cancer genomes', *Nature*, 578(7793), pp. 112–121.

Lord, C.J. and Ashworth, A. (2016) 'BRCAness revisited', *Nature reviews. Cancer*, 16(2), pp. 110–120.

Lukow, D.A. and Sheltzer, J.M. (2022) 'Chromosomal instability and aneuploidy as causes of cancer drug resistance', *Trends in cancer research*, 8(1), pp. 43–53.

Macintyre, G. *et al.* (2018) 'Copy number signatures and mutational processes in ovarian carcinoma', *Nature genetics*, 50(9), pp. 1262–1270.

Martincorena, I. *et al.* (2015) 'Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin', *Science*, 348(6237), pp. 880–886.

Martínez-Jiménez, F. *et al.* (2023) 'Pan-cancer whole-genome comparison of primary and metastatic solid tumours', *Nature*, 618(7964), pp. 333–341.

Mata, D.A. *et al.* (2022) 'Prevalence of UV Mutational Signatures Among Cutaneous Primary Tumors', *JAMA network open*, 5(3), p. e223833.

McGranahan, N. and Swanton, C. (2015) 'Biological and therapeutic impact of intratumor heterogeneity in cancer evolution', *Cancer cell*, 27(1), pp. 15–26.

Nik-Zainal, S. *et al.* (2012) 'Mutational processes molding the genomes of 21 breast cancers', *Cell*, 149(5), pp. 979–993.

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(9), p. e0221235.

Rogozin, I.B. *et al.* (2018) 'Mutational signatures and mutable motifs in cancer genomes', *Briefings in bioinformatics*, 19(6), pp. 1085–1101.

Rosenthal, R. *et al.* (2016) 'DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution', *Genome biology*, 17, p. 31.

Sansregret, L. and Swanton, C. (2017) 'The Role of Aneuploidy in Cancer Evolution', *Cold Spring Harbor perspectives in medicine*, 7(1). Available at: https://doi.org/10.1101/cshperspect.a028373.

Sosinsky, A. *et al.* (2024) 'Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme', *Nature medicine*, 30(1), pp. 279–289.

Sottoriva, A. *et al.* (2015) 'A Big Bang model of human colorectal tumor growth', *Nature genetics*, 47(3), pp. 209–216.

Steele, C.D. *et al.* (2022) 'Signatures of copy number alterations in human cancer', *Nature*, 606(7916), pp. 984–991.

Tate, J.G. *et al.* (2019) 'COSMIC: the Catalogue Of Somatic Mutations In Cancer', *Nucleic acids research*, 47(D1), pp. D941–D947.

Turajlic, S. *et al.* (2019) 'Resolving genetic heterogeneity in cancer', *Nature reviews. Genetics*, 20(7), pp. 404–416.

Turnbull, C. *et al.* (2018) 'The 100 000 Genomes Project: bringing whole genome sequencing to the NHS', *BMJ* , 361, p. k1687.

Zou, X. *et al.* (2018) 'Validating the concept of mutational signatures with isogenic cell models', *Nature communications*, 9(1), p. 1744.

Online Methods

The BASCULE framework

We developed a method, BASCULE, that can extract a broad spectrum of mutational signatures in a dataset, while leveraging the existing knowledge established from previous studies, expressed as a reference catalogue. The model will use Bayesian inference to search for de novo signatures that are statistically distinct from the reference ones, allowing to augment a catalogue with minimum overlap among signatures, the number of which is optimised by a likelihood criterion.

Bayesian Non-negative Matrix Factorization

Signature deconvolution is carried out via Bayesian non-negative matrix factorization (bNMF), starting from a data matrix X with N rows (patients) and F columns (features). The entries of X are non-negative values for the observed counts, and F depends on the type of signatures we wish to extract. BASCULE supports signatures from the simplest SBS, to DBS, ID and CN. For instance, when dealing with SBS and DBS, the features will be 96 and 78 strand-agnostic substitution contexts, respectively. In the general NMF step, X is factorised as

$$(1) \qquad \boldsymbol{X} \approx \boldsymbol{\alpha\beta}$$

One key feature of BASCULE is its ability to embed prior knowledge of known signatures into the model. To this extent, the matrices α and β can be defined as follows

$$(2) \qquad \underbrace{[\boldsymbol{\alpha}^d \quad \boldsymbol{\alpha}^c]}_{\boldsymbol{\alpha} \in \mathbb{R}^{N \times K}} \qquad \text{and} \qquad \underbrace{\begin{bmatrix} \boldsymbol{\beta}^d \\ \boldsymbol{\beta}^c \end{bmatrix}}_{\boldsymbol{\beta} \in \mathbb{R}^{K \times F}}$$

where the matrix β is defined as the concatenation of de novo signatures (β^d) to the reference catalogue (β^c), and the matrix α as the concatenation of exposures to de novo (α^d) and reference (β^c) signatures.

Assuming K signatures, we define K^d and K^c as the number of de novo and reference signatures, respectively. Hence, the total number of signatures is given by $K = K^d + K^c$. X is decomposed in a $K \times F$ matrix of signature profiles β, and a $N \times K$ matrix of exposures α, such that

$$(3) \qquad \sum_{k=1}^{K} \alpha_{n,k} = 1 \quad \forall \quad n = 1, \cdots N \quad \text{and} \qquad \sum_{f=1}^{F} \beta_{k,f} = 1 \quad \forall \quad k = 1, \cdots K$$

The non-negativity constraint is such that all entries of α and β are positive. To extend to model to be Bayesian, we use a Poisson likelihood

$$(4) \qquad p(\boldsymbol{X}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{n=1}^{N}\prod_{f=1}^{F} \mathrm{Poisson}(x_{n,f}|\lambda_{n,f})$$

assuming the mutation classes to be independent, with the Poisson rate defined as the reconstructed mutation counts matrix $\boldsymbol{\lambda} = \hat{\boldsymbol{\alpha}}\boldsymbol{\beta}$, where $\hat{\alpha}_{n,k} = \alpha_{n,k}\sum_{f=1}^{F} x_{n,f}$ reports for each patient n the number of mutations linked with signature k. This is computed as the product of the normalised exposure by the total number of mutations observed in a patient.

The matrix $\boldsymbol{\beta}$ entries learnt during the inference are those regarding $\boldsymbol{\beta}^d$, while $\boldsymbol{\beta}^c$ is held fixed to reflect the reference. If K^d is set to 0, the model solely infers the latent exposures for the input catalogue of reference signatures. Otherwise, the model additionally learns de novo signature profiles β^d and their respective exposures.

The bNMF joint distribution can be factored as follows:

$$(5) \qquad p(\boldsymbol{X},\boldsymbol{\alpha},\boldsymbol{\beta}) = p(\boldsymbol{X}|\boldsymbol{\alpha},\boldsymbol{\beta})p(\boldsymbol{\alpha},\boldsymbol{\beta}^d) = \\ = \prod_{n=1}^{N}\prod_{f=1}^{F} p(x_{n,f}|\lambda_{n,f}) \prod_{n=1}^{N} p(\boldsymbol{\alpha}_n) \prod_{k=1}^{K} p(\boldsymbol{\beta}_k^d)$$

with priors

- $p(\boldsymbol{\alpha}_n) = \mathrm{Dirichlet}(\boldsymbol{\gamma})$
- $p(\boldsymbol{\beta}_k^d) = \mathrm{Dirichlet}(\boldsymbol{\omega})$.

Since $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are probability vectors, their priors are Dirichlet distributions with concentration γ (K-dimensional) and ω (F-dimensional). The value of ω is computed to maximise the distance between each de novo and reference signature. Specifically, we generate a vector with low concentration in the features widely represented in the reference catalogue. This approach forces the model to sample from contexts with low density in the reference, ensuring that de novo signatures will be used to account for the signal that remains unexplained by reference signatures.

The model is parametric, and the number of signatures is required as input. The method selects the optimal number of de novo signatures as the one minimising the Bayesian Information Criterion (BIC), defined as $\mathrm{BIC} = -2\log\hat{\mathcal{L}} + \log N \cdot p$. Here, N denotes the number of samples in the input data, p represents the model complexity, i.e., the number of inferred parameters, and $\hat{\mathcal{L}}$ is the likelihood function (3) calculated using the Maximum A Posteriori (MAP) values of the parameters.

In the current formulation of the method, we use a two-steps inference process. First, we run the model to infer the exposures of reference signatures, without introducing any de novo signatures, i.e., with $K^d = 0$, and filter out signatures from the catalogue with low exposures, resulting in a reduced catalogue $\beta^{c'}$. In the second step, we include the significant reference

signatures ($\beta^{c'}$) and infer any new de novo one (β^{d}). This approach minimises noise from the reference catalogue and maximises the signal attributed to already validated signatures.

Post-fit heuristics. After inferring de novo signatures, the method can further evaluate their similarity relative to linear combinations of other signatures. This helps determine if one signature is actually a combination of other known signatures (or other de novo signatures). Specifically, we decompose each de novo signature profile into a linear combination of the remaining ones, as

$$\boldsymbol{w}_k = \sum_{i \neq k} \nu_i \boldsymbol{\beta}_i. \tag{6}$$

If the vector $\boldsymbol{w}_k$ closely matches – based on a cutoff on the cosine similarity – the original signature β_k, the method will remove the de novo signature and redistribute its exposures according to the weights $\boldsymbol{\nu}$ of this linear combination. This process reduces the number of signatures, enhancing the sparsity of the final results. We discuss this heuristic in the application to real data, together with specific parameters we used to refine our fits.

Nonparametric Dirichlet mixture model

BASCULE can perform deconvolution of multiple signature types and cluster a cohort of N samples based on these multiple signals. Here, we are clustering a tensor of V exposure matrices into G latent groups through a Dirichlet Process Dirichlet Mixture Model.

Assume to have V exposure matrices $\boldsymbol{\alpha}^{(1)}, \cdots, \boldsymbol{\alpha}^{(V)}$ of V distinct types of signatures (i.e., SBS, DBS, IDS, etc.). Each matrix $\boldsymbol{\alpha}$ has N rows (i.e., samples) and $K^{(1)}, \cdots, K^{(V)}$ columns (i.e., signatures). To cluster the N samples based on multiple signature types, we embed the $\boldsymbol{\alpha}$ matrices in a tensor $\boldsymbol{A}$ of dimension $V \times N \times K'$. Since the deconvolution of distinct signature types might result in different optimal numbers of signatures, we need to standardise the dimensions to embed the exposure matrices into a tensor. Here, we define $K' = \max_{v=1,\cdots,V} K^{(v)}$ and perform zero-padding in the columns of $\boldsymbol{\alpha}^{(v)}$ if $K^{(v)} < K', \quad \forall \quad v \in 1, \cdots, V$.

Similarly to the bNMF step, since the exposures can be modelled with a Dirichlet, the overall mixture likelihood can be written as

$$p(\boldsymbol{A}|\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{v=1}^{V} \prod_{n=1}^{N} \sum_{g=1}^{G} \pi_g \mathrm{Dirichlet}(\boldsymbol{A_{v,n}}|\boldsymbol{\theta_{v,g}}) \tag{7}$$

and the joint probability can be factored as

$$\begin{aligned} p(\boldsymbol{A}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\rho}) &= p(\boldsymbol{A}|\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\rho}) p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\rho}) = \\ &= p(\boldsymbol{A}|\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\rho}) p(\boldsymbol{Z}|\boldsymbol{\pi}) p(\boldsymbol{\pi}) \prod_{v=1}^{V} \prod_{g=1}^{G} p(\boldsymbol{\theta}_{v,g}|\rho_v) \prod_{v=1}^{V} p(\rho_v) \end{aligned} \tag{8}$$

with

- $p(\boldsymbol{\theta}_{v,g}) = \mathrm{Dirichlet}(\boldsymbol{\phi}_{v,g} \cdot \rho_v)$,
- $p(\rho_v) = \mathrm{Gamma}(c, d)$.

The model defines a distribution over mixture components and weights as a Dirichlet Process $DP(\eta, H)$ with concentration η and base distribution H. We adopt the stick-breaking formulation of the Dirichlet Process to automatically determine the optimal number of clusters, up to a fixed maximum value G. $\boldsymbol{\theta}_{v,g}$ is a K-dimensional vector reporting the component-specific exposure centroids for each signature k. This is distributed as a Dirichlet whose concentration depends on $\boldsymbol{\phi}_{v,g}$, a K-dimensional concentration vector scaled by ρ_v. In the model, $\boldsymbol{\phi}_{v,g}$ might account for prior information on each retrieved signature and ρ_v is a signature type-specific scaling factor used to lower the variance of the Dirichlet distribution and to allow sparsity in the centroids.

The inference has been carried out via Stochastic Variational Inference in Pyro (Bingham *et al.*, 2019). The parameters were initialised with K-means to determine the component centroids (θ). The Dirichlet distribution has a characteristic where low concentration values lead to high variance, and high concentration values result in low variance. To address this, we introduced a scaling factor (ρ) to reduce the variance of the θ samples. After running K-means, we calculated an upper bound for the empirical variance of exposures for each cluster g and signature type v. We then chose the value of ρ to ensure that samples drawn from the θ prior would approximate the observed variance.
The model infers the maximum a posteriori (MAP) set of parameters and assigns each data point to the cluster maximising the posterior distribution.

Post-fit heuristics. After inferring the latent clustering assignments, the method can evaluate whether some clusters should be merged based on the inferred centroids (θ). Specifically, the method iteratively calculates the cosine similarity between the centroids. If this similarity exceeds a custom threshold, the clusters are merged, and the centroids are updated. This process continues until all clusters are distinct according to the defined similarity measure. We discuss this heuristic in the application to real data, together with specific parameters we used to refine our fits.

Simulated data

We evaluated BASCULE's performance by examining the error in reconstructing mutation counts, the accuracy of signature retrieval, the quality of signature exposures and the accuracy of clustering. To do this, we generated synthetic datasets of increasing complexity from the BASCULE generative bNMF model. Specifically, we considered cohorts of 150, 500 and 1000 samples, composed of 1, 3 or 6 groups. As the number of groups increases, the number of signatures (SBS and DBS) also increases, ranging from 1 to 13. In all fits, certain signatures are shared among all clusters. In contrast, others are either private to a single cluster or shared by a specific combination of clusters, enabling the unique identification of each group of patients. For each configuration, we generated 30 datasets. Each fit has been

run with SBS1, SBS5, DBS3 and DBS5 as reference catalogues, testing five different de novo signature values around the true one.

BASCULE performance. We evaluated BASCULE performance for the bNMF and clustering models. We first assessed, in terms of recall and precision, the similarity among the number of retrieved de novo signatures and the simulated one (Supplementary Figure S1A,B). The recall results demonstrate that BASCULE effectively retrieves a precise number of de novo signatures, with only a few signatures missed in the analysis. The precision results indicate that only a few signatures were incorrectly identified. As expected, in both analyses, increasing the number of signatures results in a lower accuracy, whereas increasing the number of samples, thus providing a stronger signal for the signatures, improves the accuracy. Additionally, the performance is higher when working with SBS, which could be attributed to the greater sparsity of DBS signatures than SBS ones.

We then assessed the quality of the mutation counts reconstruction, as well as the signatures and exposure learning, in terms of mean squared error (MSE) and cosine similarity (CS), respectively. These are standard metrics in the field (Omichessan, Severi and Perduca, 2019). The input mutation counts matrix was reconstructed with high accuracy (Supplementary Figure S1C), resulting in low values of MSE. The CS computed between true and estimated signatures (Supplementary Figure S1D) and between true and estimated exposures (Supplementary Figure S1E) show great concordance between our results and the simulated datasets. Similar to recall and precision analysis, increasing the number of signatures adds complexity to the input data, resulting in reduced quality. However, increasing the number of samples leads to an improvement in the quality of the fit results.

Finally, we assessed the performance of the clustering model using the adjusted Rand index (ARI) and normalised mutual information (NMI) (Supplementary Figure S1F). The ARI compares how well the clusters predicted by a model match with the true clusters (ground truth), adjusting for the chance of random clustering, with a score ranging from -1 (completely incorrect clustering) to 1 (perfect clustering). NMI measures how much information is shared between the predicted clusters and the true clusters. The values range from 0 to 1, where 0 means the clustering assignments are independent, and 1 means identical (it is normalised to ensure comparability across different datasets, regardless of the size or number of clusters). In our tests both metrics indicate that BASCULE performs well in accurately retrieving the correct sample assignments, with accuracy decreasing as the number of signatures increases and improving as the number of samples increases.

Comparison to other methods. Simulated counts were also used to compare BASCULE to two other methods for de novo mutational signature extraction: SigProfiler and SparseSignatures. Both methods were run on 10 CPU cores of an AMD EPYC 7H12 computer, performing NMF with 10 repetitions.

We compared the performance of the three methods to perform de novo signature extraction on SBS through the same metrics described in the previous paragraph. BASCULE outperformed the other methods in identifying the correct number of signatures (Supplementary Figure S2A,B), mainly when the number of signatures was low. Notably, all

methods showed low reconstruction error for the input datasets (Supplementary Figure S3C).

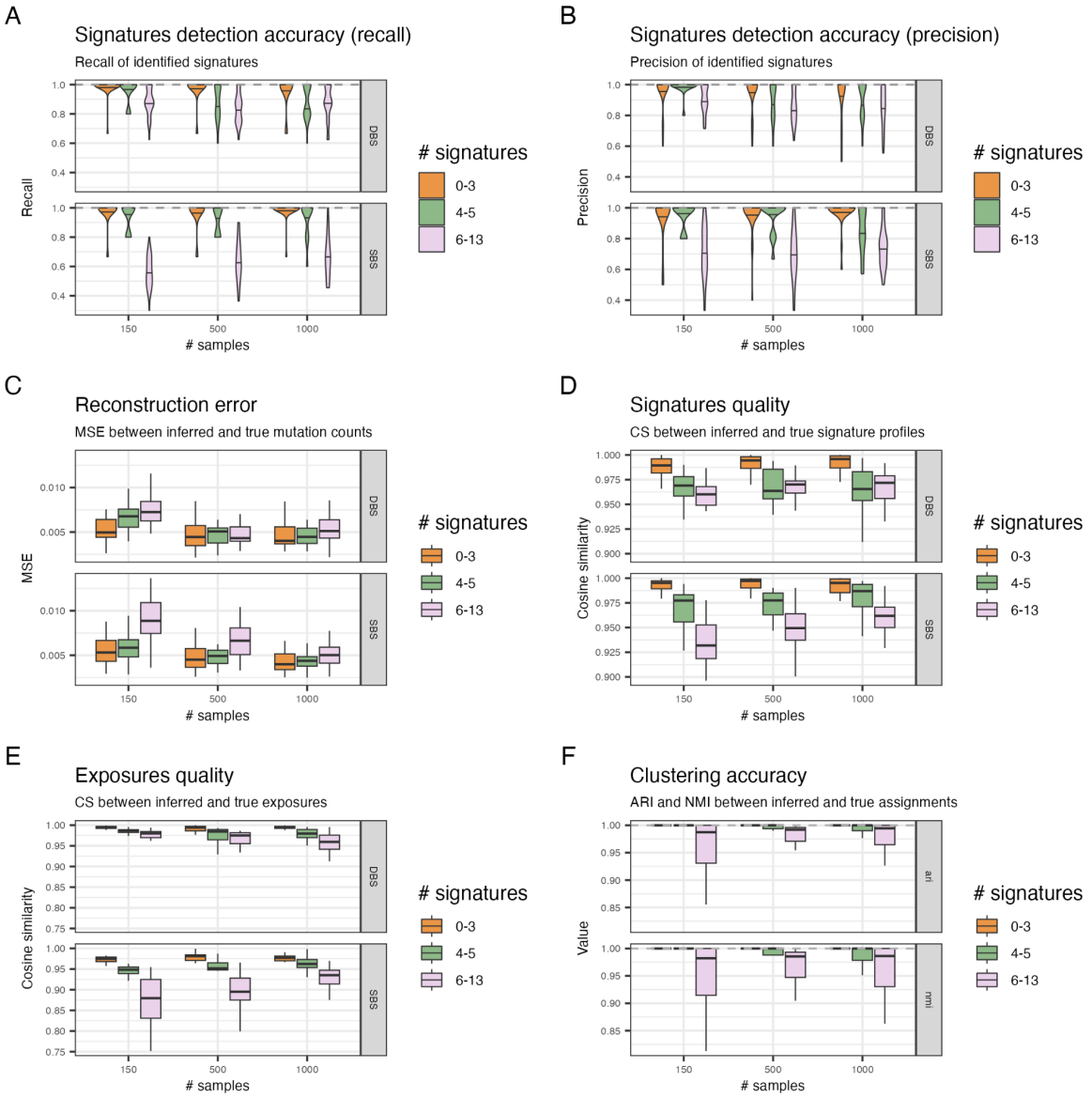
For the correctly retrieved signatures, we assessed their quality based on cosine similarity (CS) (Supplementary Figure S2D). BASCULE achieved higher CS scores than other methods when the number of signatures was low, but all methods performed similarly as the number of signatures increased.

We also calculated the CS for the exposures of matched signatures (Supplementary Figure S2E), where BASCULE's performance decreased relative to other methods as dataset complexity increased. Finally, we computed the CS of exposures for matched and unmatched signatures, considering zero exposures for false negatives (FN) and false positives (FP) in the inferred and true exposures. As dataset complexity increased, BASCULE outperformed the other methods, indicating that FN and FP are not necessarily signatures with low exposure levels.

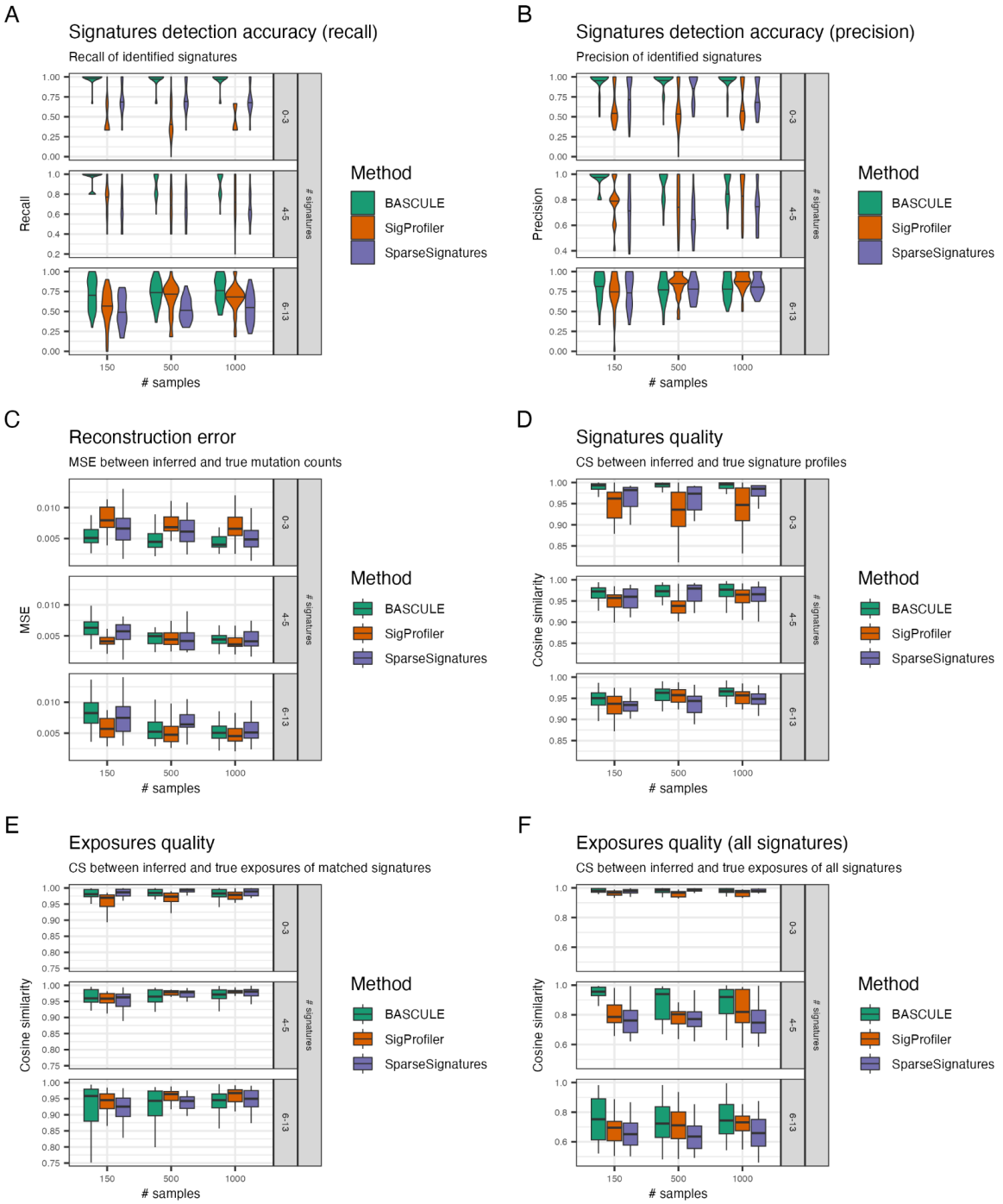
We finally compared the time required to fit each dataset for the three methods (Supplementary Figure S3). BASCULE, implemented in Pyro, uses Stochastic Variational Inference (SVI), which enables faster inference compared to standard NMF methods. Additionally, BASCULE supports GPU-based inference, further accelerating computationally intensive calculations.

SigProfiler. Version 1.1.21 of SigProfiler was run with the default values for the stability thresholds, using a minimum of 1000 and a maximum of 10000 NMF iterations, with a value of 1000 iterations for the convergence test. All other parameters were set to the default values.

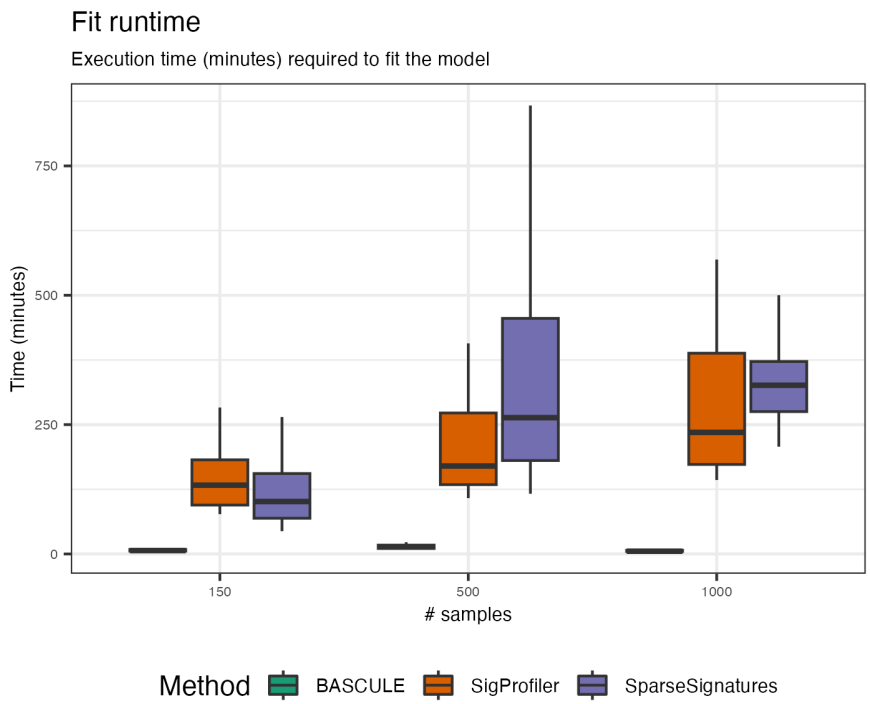
SparseSignatures. Version 2.10.0 of SparseSignatures was run with the “normalize_counts” parameter set to TRUE, a value of 20 iterations and 10000 maximum lasso iterations. We used the “background2” signature, which is derived from COSMIC SBS5, as the background. For the cross-validation parameters, we used 5 entries, 10 repetitions, and a percentage of 0.01 entries to be masked by zeros. For sparsification, we tested λ values of 0 and 0.05 for α and values of 0, 0.01, 0.05 and 0.1 for β.



Supplementary Figure S1. BASCULE performance assessment with varying the number of signatures in the input dataset. A. Recall computed on the number of retrieved de novo signatures. B. Precision computed on the number of retrieved de novo signatures. C. Reconstruction error (mean squared error) on the input mutation counts matrix. D. Cosine similarity of inferred and true signatures. E. Cosine similarity of inferred and true exposures. F. Adjusted Rand index and normalised mutual information of inferred and true clustering assignments.



Supplementary Figure S2. BASCULE performance compared to SigProfiler and SparseSignatures, with varying the number of signatures in the input dataset. A. Recall computed on the number of retrieved de novo signatures. B. Precision computed on the number of retrieved de novo signatures. C. Reconstruction error (mean squared error) on the input mutation counts matrix. D. Cosine similarity of inferred and true signatures. E. Cosine similarity of inferred and true exposures considering matched signatures. F. Cosine similarity of inferred and true exposures considering matched and unmatched signatures.



Supplementary Figure S3. Runtime (in minutes) of BASCULE compared to SigProfiler and SparseSignatures.

Real data

Construction of the starting catalogue. We curated a simple version of a catalogue by combining our current understanding of mutational processes for SBS and DBS – the two types of mutational signatures for which we had data from (Degasperi *et al.*, 2022). The aim was to use this catalogue as input for our tool. To assemble it, we selected the common signatures from each tumour type as documented by Degasperi et. al. Subsequently if these particular signatures were also found in COSMIC, the corresponding COSMIC version of the signature was adopted for our input reference catalogue. In the final step, we enhance the signatures so that each one accurately reflects a distinct pattern of mutational signatures. This is achieved by eliminating signals from 96 contexts whose values fall below a specified threshold.

Input data. To perform the large-scale analysis discussed in the Main Text, we collected data from supplementary materials by Degasperi et al. Initially, we identified the organ-specific common signatures from Tables 9 and 10. Subsequently, we located their corresponding reference signatures in Tables 21 and 22, utilising the conversion matrix provided in Tables 25 and 26.

Fitting off-catalogue signatures. In the initial stage, we ran our tool, searching for a maximum of 25 de novo signatures (beyond the input catalogue). The model underwent the learning phase, consisting of 3000 iterations and a learning rate set at 0.005, and the most suitable number of de novo signatures was determined by evaluating the BIC score.

After inferring the latent variables, including the mutational signatures and exposure matrices, we applied a refinement step to the de novo signatures. This process aimed to

identify and remove any de novo signatures that could be adequately explained by a linear combination of the other de novo and input catalogue signatures. The refinement method attempted to reconstruct each de novo signature using the coefficients of the corresponding linear combination. If the cosine similarity between the reconstructed signature and the original de novo signature exceeded 0.9 - an arbitrary threshold denoting very high similarity - the de novo signature was removed, and its exposure was redistributed among the signatures in the linear combination, weighted by their respective coefficients. By applying this refinement step, we ensured that the final set of de novo signatures represented mutational processes that could not be sufficiently captured by combining the existing signatures, either from the input catalogue or the newly identified de novo signatures. This approach helped to minimise redundancy and improve the interpretability of the mutational signature analysis.

Clustering patients by signature exposures. After this step, we ran the Dirichlet Process clustering. To optimise those results, we examined the clusters of signatures and merged those that exhibited some similarity. This process was conducted iteratively, merging clusters based on their similarity with their cluster centroids. The merging process continued until the cosine similarity between all pairs of clusters fell below 0.8, a predetermined cutoff value that denoted low similarity.

To interpret our clusters, we additionally assessed whether our de novo signatures could be accounted for by any signatures from the COSMIC database and the catalogue of Degaspari et al. To achieve this, we calculated the cosine similarity between all possible pairs of de novo signatures versus signatures from both extended catalogues. If the similarity exceeded 0.9, we replaced the de novo signature with the corresponding known signature from these catalogues, as discussed in the Main Text.

In this analysis phase, we examined the clusters of signatures and merged those that exhibited similarity. This process was conducted iteratively, merging clusters based on the similarity of their centroids. The merging process continued until the cosine similarity between all pairs of clusters fell below a predetermined cutoff value. Additionally, we assessed whether our de novo signatures could be accounted for by any signature known in COSMIC or the catalogue by Degaspari et al. To achieve this, we calculated the cosine similarity between all possible pairs of de novo signatures and the target catalogues, and, if the similarity exceeded 0.9, we replaced our de novo signature with the corresponding known one.

Breast cancer analysis. We analysed 2,682 breast tumour samples from the GEL, ICGC, and HMF cohorts. For the SBS context, we selected signatures SBS1, SBS2, SBS3, SBS5, SBS8, SBS13, SBS17, SBS18, SBS31, and SBS127, and for the DBS context, we selected signatures DBS11, DBS13, and DBS20, collectively forming the reference catalogue for breast tumour data.

The bNMF analysis identified 19 SBS (12 de novo) and 5 DBS (2 de novo) signatures. After a refinement step, we identified 18 SBS (11 de novo) and 5 DBS (2 de novo) signatures. The refinement process revealed that the de novo signature SBSD3 could be explained by the

input catalogue signature SBS1, with a cosine similarity of 0.99 between the reconstructed and original signatures. As a result, one de novo signature was removed from the inferred set.

Following refinement, we proceeded with the merging step. Initially, the tool detected 14 clusters among the breast cancer samples. We then applied an iterative merging function, resulting in 5 clusters (Supplementary Figure S5), using a cutoff value of 0.8.

Later, we mapped the de novo signatures to known signatures from the COSMIC catalogue and the findings from Degasperi et al.; within the SBS context, 6 de novo signatures corresponded to COSMIC signatures SBS7a, SBS17b, SBS26, SBS40a, SBS90, and SBS98. Additionally, one de novo signature aligned with the common signature SBS8, and two signatures matched rare signatures SBS44 and SBS97, both from the Degasperi et al. study. In the DBS context, we mapped two de novo signatures to COSMIC signatures DBS14 and DBS20. Ultimately, two de novo signatures remained unmapped (Supplementary Figure S6), one of which, SBSD12, exhibits a pronounced pattern of C>G and C>T transitions. The detailed signature exposures across the various clusters are presented in (Supplementary Figure S7).

Lung cancer analysis. We analysed 1,396 lung tumour samples obtained from the GEL, ICGC, and HMF cohorts. For our study, we selected the following SBS signatures: SBS2, SBS4, SBS1, SBS5, SBS3, SBS8, SBS13, SBS17, SBS18, SBS31, and SBS92, as well as the DBS signatures DBS2, DBS5, DBS13, and DBS20. These signatures collectively formed the reference catalogue for our lung data.

The results from the bNMF analysis revealed 19 SBS signatures (11 of which were identified de novo) and 6 DBS signatures (2 of which were de novo). Following a refinement step, we identified 15 SBS signatures (7 de novo) and maintained 6 DBS signatures (2 de novo). During the refinement process, it was determined that 4 de novo signatures could be attributed to existing signatures (both de novo and input reference catalogue) based on a cosine similarity threshold exceeding 0.9 between the reconstructed and original signatures. After the refinement, we proceeded to the merging phase, where the tool initially identified 15 clusters among the lung cancer samples. We then applied a merging function that iteratively combined these clusters until we arrived at 7 distinct clusters (Supplementary Figure S8), using a cutoff value of 0.8.

Later, we aligned the de novo mutational signatures with known signatures from the COSMIC database and the findings of Degasperi et al. In the context of SBS, we identified three de novo signatures that corresponded to COSMIC signatures SBS3, SBS5, and SBS17b; one de novo signature that matched common signature SBS8 from Degasperi et al., and another signature that aligned with rare signature SBS44 reported by Degasperi et al. For DBS, we successfully mapped two signatures to COSMIC signatures DBS1 and DBS6. Two de novo signatures SBSD2 and SBSD19 (Supplementary Figure S9) remained unmapped. Most of the density in SBSD19 is concentrated in T>C mutations, while SBSD2 primarily exhibits C>T and T>A transitions. The detailed signature exposures across the various clusters in lung cancer type are presented in (Supplementary Figure S10).

Colorectal cancer analysis. Our study examined 2,845 colorectal tumour samples from the GEL, ICGC, and HMF cohorts. We utilised an input reference catalogue comprising 13 SBS signatures SBS1, SBS2, SBS3, SBS5, SBS8, SBS13, SBS17, SBS18, SBS35, SBS88, SBS93, SBS121, SBS157 and 4 DBS signatures; DBS2, DBS5, DBS13, DBS20 for the analysis of colorectal tumour data.

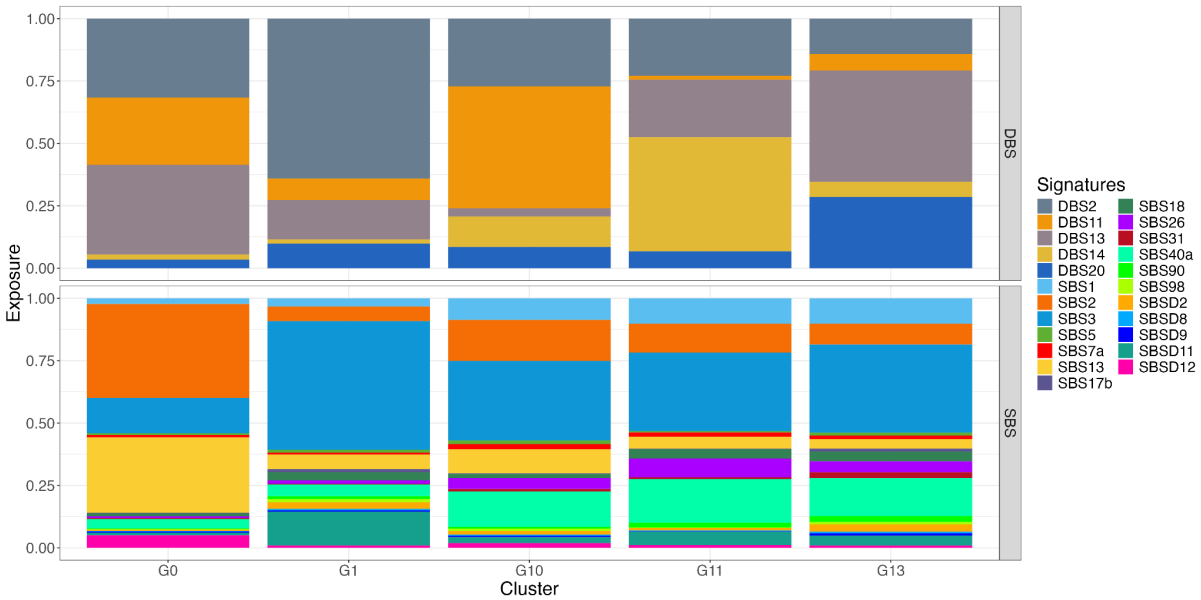
Initial bNMF yielded 32 SBS signatures (24 de novo) and 8 DBS signatures (4 de novo). Subsequent refinement reduced these to 24 SBS signatures (16 de novo) and 8 DBS signatures (4 de novo), as 8 de novo signatures were found to be explicable by other signatures, with cosine similarities exceeding 0.9. Post-refinement the algorithm initially identified 15 clusters, which were reduced to 8 distinct clusters through iterative merging with a cutoff value of 0.8 (Supplementary Figure S11).

We compared the de novo mutational signatures identified in our analysis to known signatures from the COSMIC database and the work of Degasperi et al. In the context of SBS, we found six de novo signatures aligned with COSMIC signatures SBS10a, SBS10b, SBS15, SBS17b, SBS26, and SBS28. Additionally, one de novo signature matched signature SBS93 from the common signatures in Degasperi et al., and two signatures corresponded to signatures SBS44 and SBS97 from the rare signatures group reported by Degasperi et al. For DBS context, we successfully mapped three signatures to COSMIC signatures DBS3, DBS8, and DBS14, and one de novo to signature DBS25 from the rare category in the Degasperi et al. study.

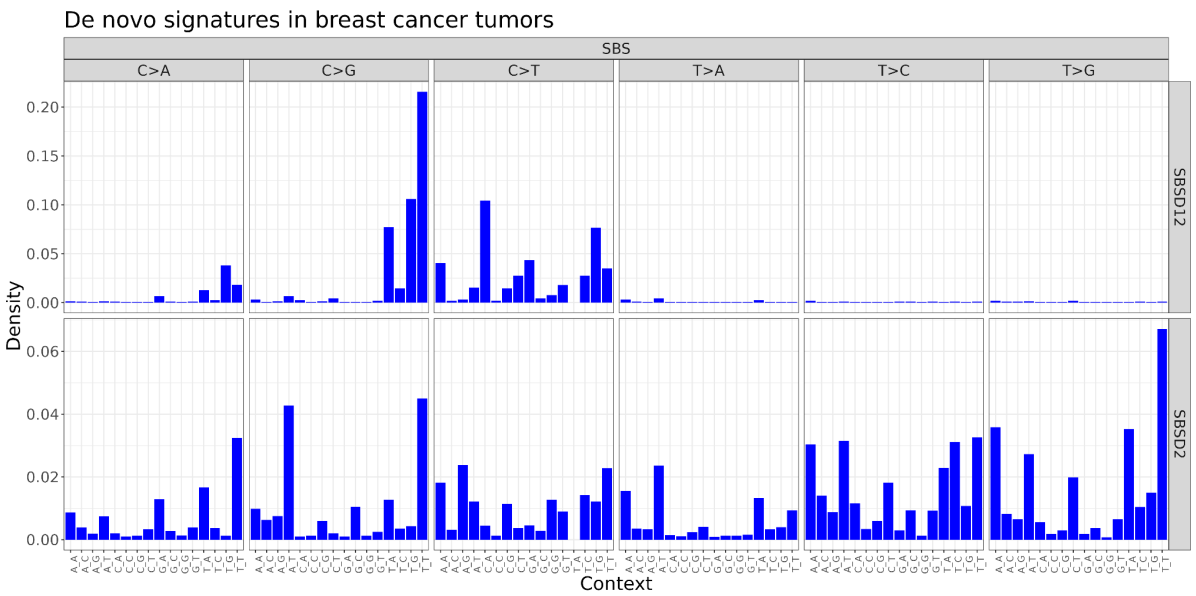
Ultimately, seven de novo signatures remained unmapped (Supplementary Figure S12). SBSD12 exhibits a higher density of C>T and T>C mutations, while SBSD15 shows a pronounced peak in C>A transitions. SBSD16 primarily displays density in C>T, T>A, and T>C mutations. SBSD6 has its density distributed across C>A and C>T transitions, and SBSD7 predominantly features mutations in C>A and C>T. SBSD8 is the most widely distributed signature among the others, with higher densities in C>A, C>T, and T>A. Finally, SBSD9 has the majority of its density concentrated in C>A, T>C, and T>G transitions. The detailed signature exposures across the various clusters in this cohort are presented in (Supplementary Figure S13).

Unmapped de novo signatures. The de novo signatures had no mutual signatures with a cosine similarity higher than 0.8 compared to known signatures from the Degasperi et al. study and the COSMIC catalogue. This threshold value is based on heuristics and may yield different results with varying values. To assess the difference between de novo and known signatures, we subsequently mapped these de novo signatures using a threshold lower than 0.8 (Supplementary Figure S4). The results indicate that some known signatures still exhibit similarities close to 0.8 with specific de novo signatures. This comparison highlights the need for further investigation into selecting an appropriate threshold.

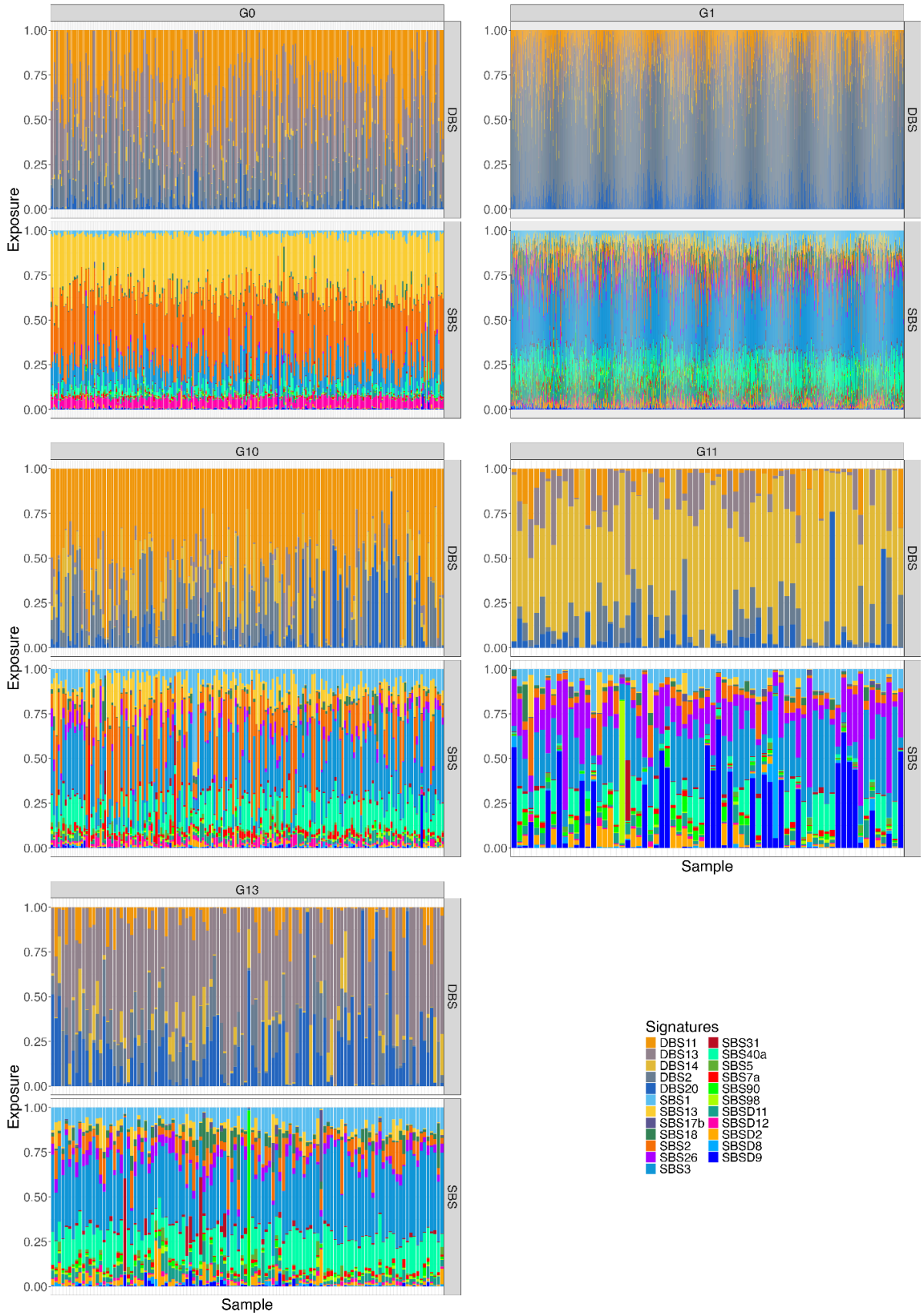
Supplementary Figure S4. De novo signatures, detected from breast, lung and colorectal tumours, mapped to signatures from the Degasperi et al. study and the COSMIC catalogue with cosine similarity less than 0.8.



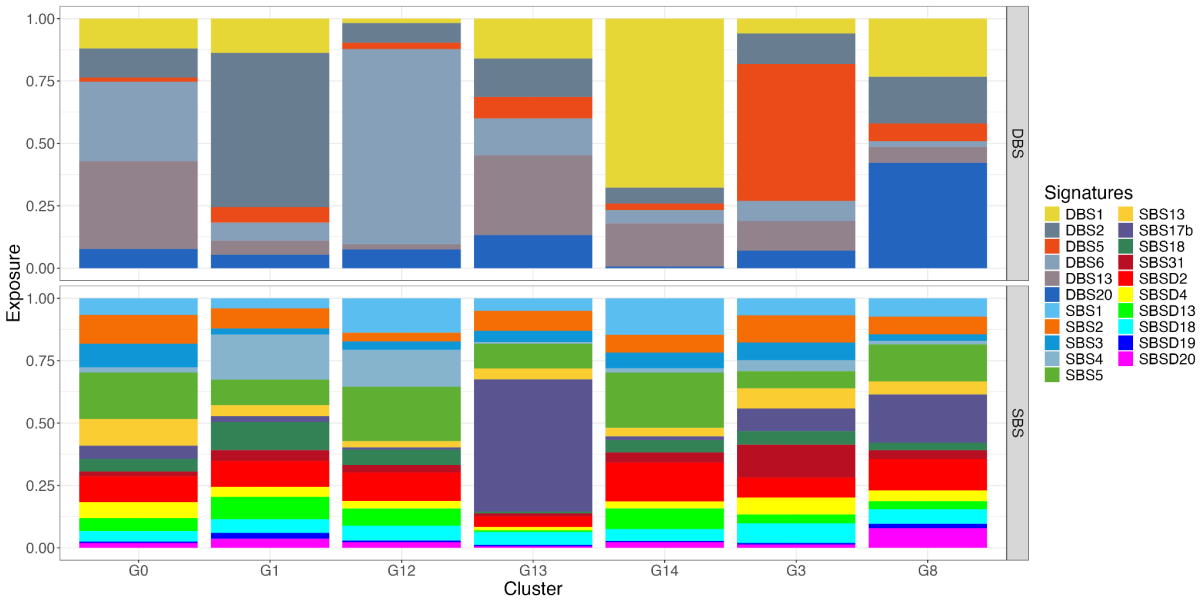
Supplementary Figure S5. Clustering centroids for breast cancers, with 5 clusters and 23 SBS and DBS signatures.



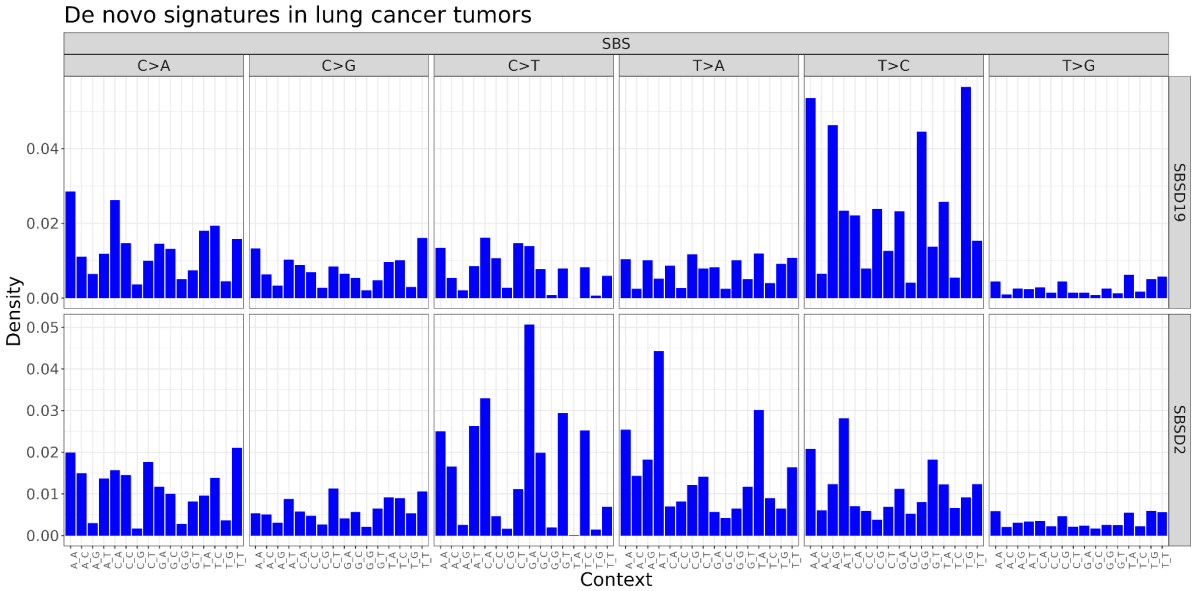
Supplementary Figure S6. Breast cancer de novo signatures that are not mapped to any known signatures from COSMIC or Degasperi et al., based on cosine similarity. See also Supplementary Figure S4.



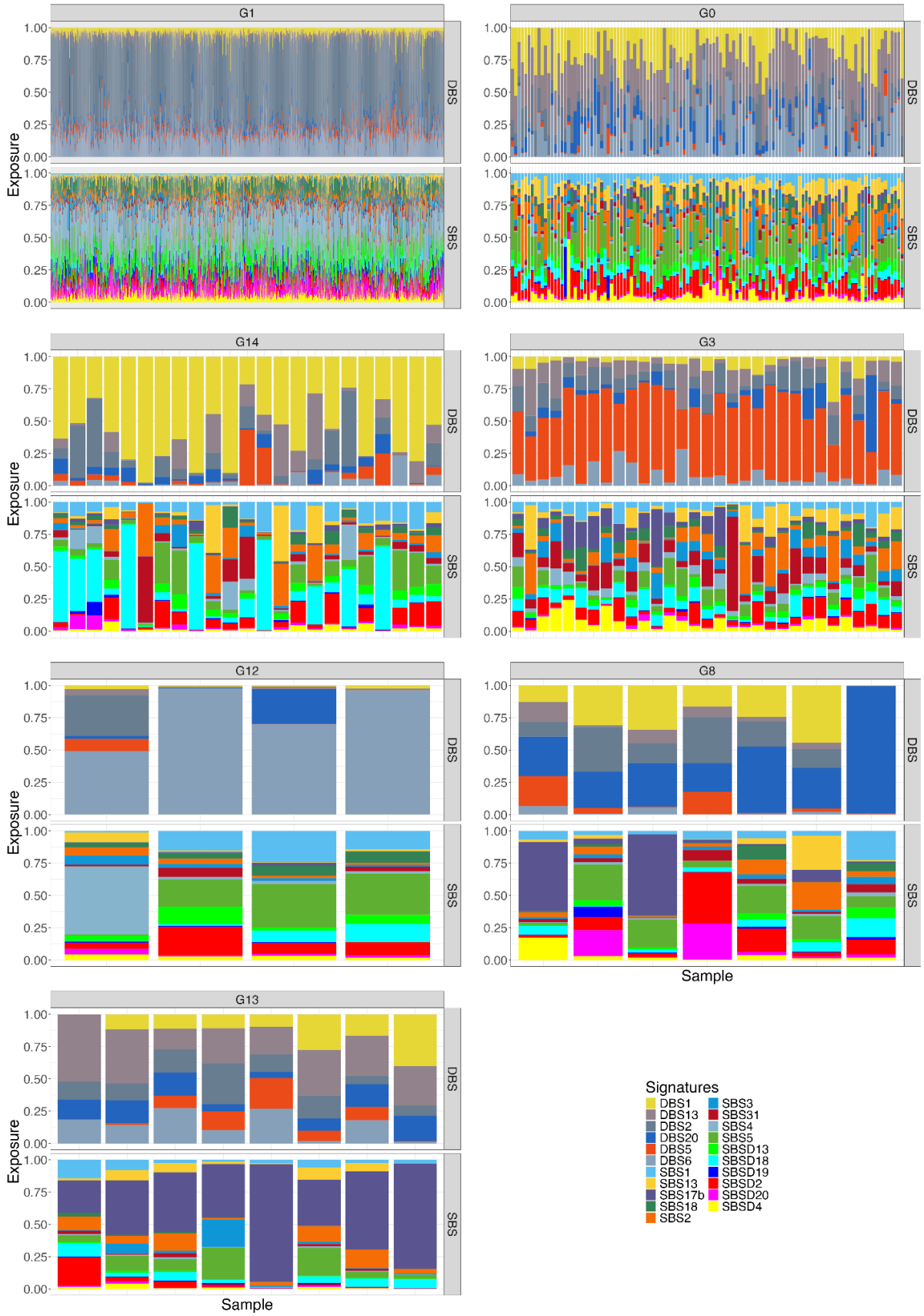
Supplementary Figure S7. The exposure plot of all 5 clusters detected in breast cancers, including both SBS and DBS signatures.



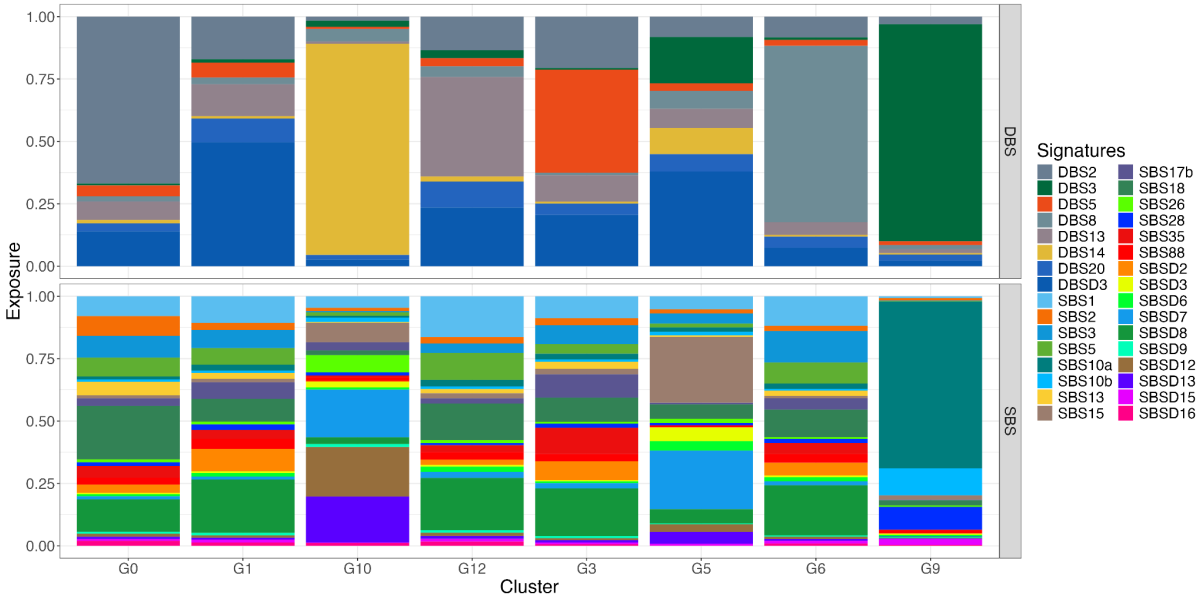
Supplementary Figure S8. Clustering centroids for lung cancers, with 7 clusters and 21 SBS and DBS signatures.



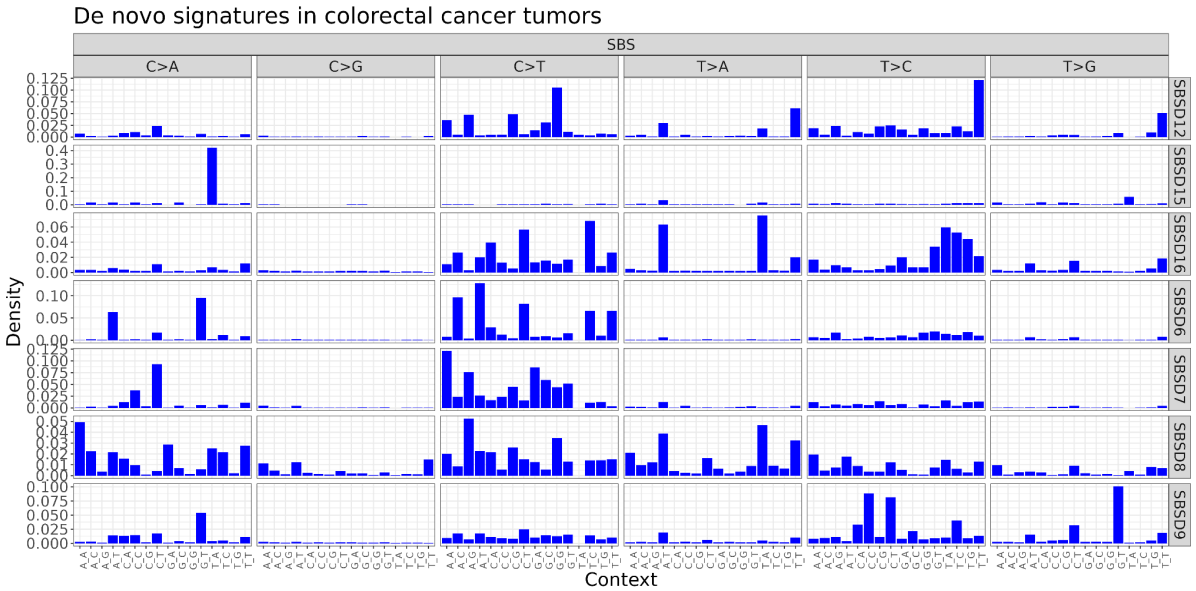
Supplementary Figure S9. Lung cancer de novo signatures that are not mapped to any known signatures from COSMIC or Degasperi et al., based on cosine similarity.



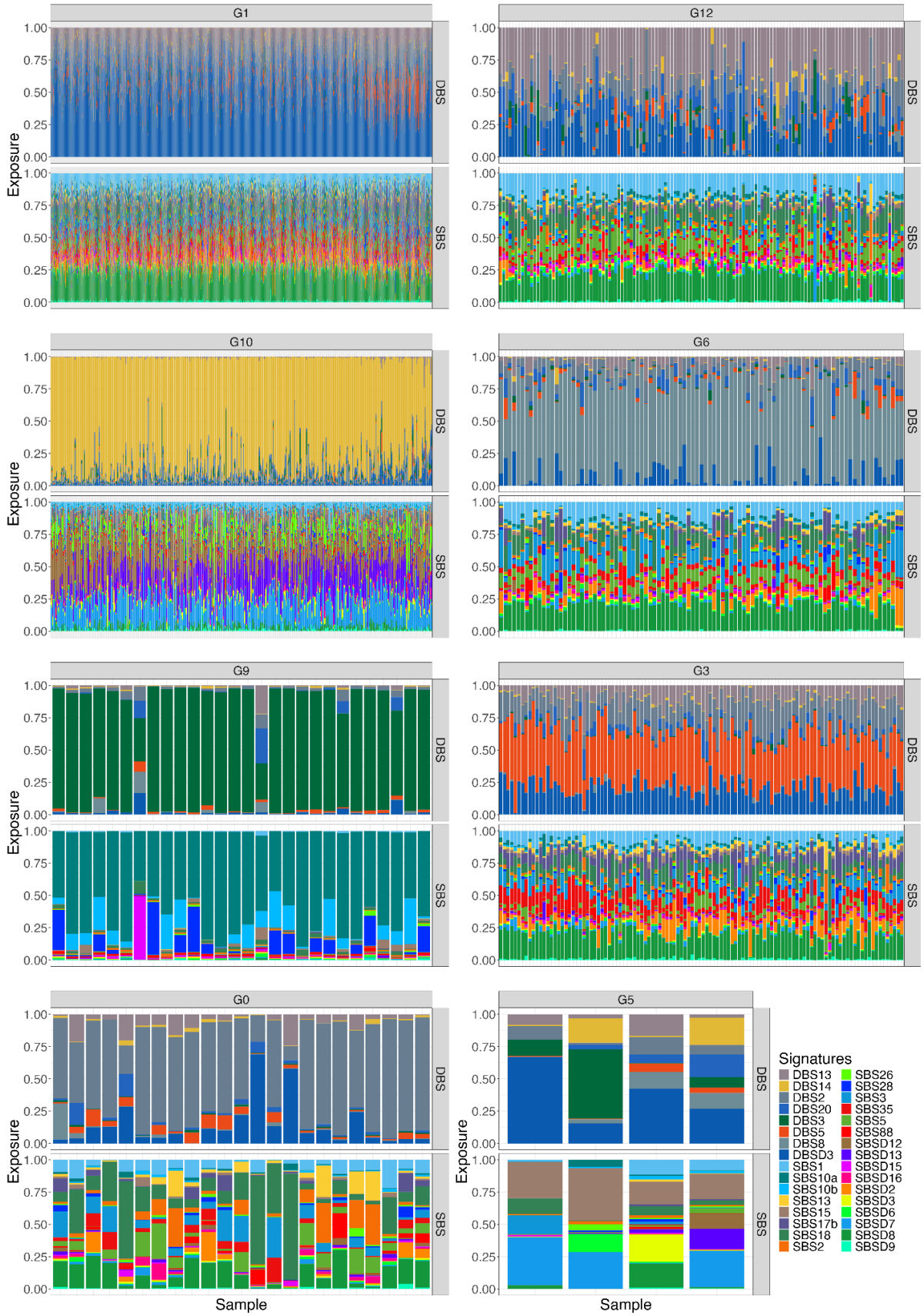
Supplementary Figure S10. The exposure plot of all 7 clusters detected in lung cancers, including both SBS and DBS signatures.



Supplementary Figure S11. Clustering centroids for colorectal cancers, with 8 clusters and 32 SBS and DBS signatures.



Supplementary Figure S12. Colorectal cancer de novo signatures that are not mapped to any known signatures from COSMIC or Degasperi et al., based on cosine similarity.



Supplementary Figure S13. The exposure plot of all 8 clusters detected in colorectal cancers, including both SBS and DBS signatures.