Mutational signatures in the general population

Population-scale mutational signature analysis of blood-derived genomes from the UK Biobank

CS5000: Thesis Project

Kirsten Timmerman

Delft University of Technology

TU Delft

Mutational signatures in the general population

Population-scale mutational signature analysis of blood-derived genomes from the UK Biobank

by

Kirsten Timmerman

to obtain the degree of Master of Science

at the Delft University of Technology,

to be defended publicly on Monday June 29th, 2026 at 10:00 AM.

Student number:	5013178	
Project duration:	November 1, 2025 – June 29, 2026	
Thesis committee:	Dr. J.S. de Pinho Gonçalves,	TU Delft, supervisor
	Dr. J. Sun,	TU Delft
	Dr. M. Weinmann,	TU Delft
	MSc S. Costa,	TU Delft, daily supervisor

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Mutational signatures in the general population

Kirsten Timmerman[1,*]

[1]Pattern Recognition and Bioinformatics, Department of Intelligent Systems, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands

Abstract

Mutational processes leave characteristic patterns of somatic mutations, traditionally studied in tumour tissue. Far less is known about whether they can be observed in normal tissue, particularly blood. Detecting the mutational imprint of disease-associated processes there could enable earlier detection and intervention. Here, we investigate whether the mutational signal detected in blood can be explained by biological and clinical factors, in particular age, DNA repair deficiencies, and cancer diagnoses. Using whole-genome sequencing of blood-derived DNA from 17,419 UK Biobank participants, we developed a filtering strategy to isolate somatic mutations and analysed four views of the mutational landscape: mutation burden, mutation channel composition, exposures to de novo signatures and exposures to COSMIC signatures. We modelled their relation to these factors using regression analysis. Across all four views, sequencing provider was the dominant predictor, far outweighing other predictors. Among non-technical predictors, BRCA ($p = 0.024$) and POLE ($p = 0.030$) variants were significantly associated with a higher mutation burden. A leukaemia diagnosis was the strongest signal across the remaining views, appearing in both the mutation channel composition and the exposure to the clock-like signature SBS1 ($p = 2.10e\text{-}06$). De novo signature exposures clustered by sequencing provider, and no association survived when extraction was performed separately for each provider. Our results show that some biological and clinical factors do explain part of the mutational signal in blood, but that technical variation between sequencing providers dominates the mutational landscape and must be addressed before blood can serve as a reliable substrate for mutational analysis.

Keywords Mutational Signatures, Population Genomics, Somatic mutations, UK Biobank

Availability and implementation https://gitlab.ewi.tudelft.nl/goncalveslab/master-projects/msc-thesis-2526-kirsten-timmerman

1. Introduction

Understanding human disease is paramount to preventing and treating it. One such disease is cancer, a leading cause of death worldwide [1]. Cancer arises from mutations in the DNA of cells that cause them to evade the normal controls on cell division and cell death. These cancer-driving changes are largely somatic, meaning acquired throughout our lives rather than inherited [2]. Somatic mutations can arise from exogenous processes, originating outside the body, such as tobacco smoking [3] or UV-light exposure [4], or from endogenous processes within the body, such as faulty DNA repair mechanisms [5] or ageing [6]. Collectively, these somatic mutations form characteristic patterns known as mutational signatures: these are genome-wide patterns which describe the frequency of each mutation type [7]. Extracting and studying mutational signatures serves different purposes, among them the following: first, because individual signatures can be linked to specific genetic defects or diseases, their presence may serve as a candidate biomarker [8]. Second, large-scale extraction allows the biological processes behind these signatures to be confirmed, sometimes through experimental validation using experimentally induced signatures [9]. Third, investigating mutational signatures can advance our understanding of how certain processes form and evolve in a cell, such as how a healthy cell develops into a tumour cell [10].

Mutational signatures are characteristically extracted from tumour samples, since tumours not only provide a large number of somatic mutations but also allow for a clear analysis of the processes driving tumour development. Crucially though, the processes that are contributing to cancer can be already active before the development of tumours. In skin cells, for example, normal human skin carries a higher burden of somatic mutations than expected, under pervasive positive selection, with most of these mutations bearing the characteristic UV-induced signature [11]. This shows that some signatures of cancer-associated processes are already detectable before any tumour develops. More broadly, somatic mutations accumulate in all dividing tissues over time [12], and mutational signatures observed in normal tissue overlap substantially with those operating in the corresponding cancers, albeit with documented differences [13]. This raises the question of whether mutational signature analysis can be meaningfully applied outside the tumour context and, in particular, whether signatures can be reliably detected and interpreted in normal, healthy tissue. If so, the potential impact is considerable: detecting the mutational imprint of disease-associated processes in normal tissue could enable earlier detection of disease, at a stage when it can be more effectively monitored, treated, or prevented from progressing.

Blood represents a promising alternative tissue source for mutational signature analysis. Somatic mutations accumulate in

normal haematopoietic stem cells throughout life, leaving a detectable, age-associated mutational imprint in this accessible tissue [14]. More broadly, mutational signatures have been identified in blood and haematopoietic cells across several studies [15; 16], further establishing blood as a viable substrate for this type of analysis. Peripheral blood can be collected non-invasively and at scale, further supporting the need for a population scale study of mutational signatures.

One biological process such a study could probe is DNA repair deficiency. Blood has not been widely used for the detection of faulty DNA repair mechanisms in the way tumour tissue has. Blood has more often been used to identify inherited germline variants in DNA repair genes [17; 18], while tumour tissue is used to characterise the somatic mutational patterns that signal repair deficiency. However, if a germline variant impairs a DNA repair pathway, the downstream mutational consequences of that impairment should, in principle, be detectable in somatic mutations accumulated across blood cells over time. This raises the possibility that mutational signatures linked to DNA repair deficiencies could be recovered from blood samples. We hypothesise that, even if blood lacks circulating tumour DNA, the processes that precede and contribute to tumourigenesis may nonetheless leave detectable signatures in blood collected prior to or concurrent with a cancer diagnosis. Therefore, mutational signatures extracted from blood could potentially serve as proxies for underlying biological processes and be linked to cancer diagnoses of the corresponding individuals.

This study aims to characterise mutational signatures in blood and to relate them to factors such as age, germline DNA repair deficiencies, and cancer diagnoses, including samples collected prior to diagnosis. This study uses whole-genome sequencing of blood-derived DNA from 17,419 participants in the UK Biobank, a large-scale biomedical database containing rich longitudinal health and genetic data from approximately 500,000 participants [19]. From this foundation, the present work addresses three questions. First, can a somatic filtering strategy reliably isolate somatic mutations from blood-derived whole-genome sequencing data? Second, does the somatic mutational landscape of blood, consisting of the mutation burden, the mutation channels and the exposures to the mutational signatures, reflect underlying biological factors, in particular age and germline deficiencies in DNA repair? Third, is this landscape associated with cancer?

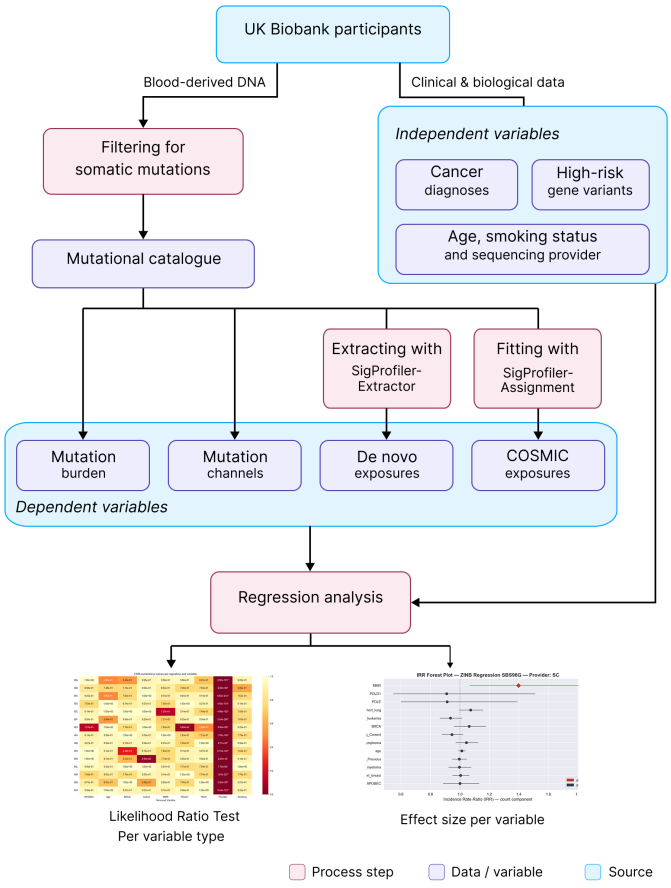


Figure 1 Blood-derived DNA sequencing data and matched clinical, biological, and technical metadata are obtained for UK Biobank participants. From the sequencing data, variants are filtered to retain somatic single base substitutions, which are aggregated by substitution type and trinucleotide context into a 96-channel mutational catalogue. The catalogue yields four dependent variables describing the somatic mutational landscape: mutation burden (total mutation count), mutational channels (the burden-normalised 96-channel composition), and exposures to *de novo* signatures and to reference COSMIC signatures, extracted and fitted with SigProfilerExtractor and SigProfilerAssignment, respectively. The metadata provide the independent variables: cancer diagnosis, high-risk gene variants, age, smoking status, and sequencing provider. Each dependent variable is regressed on the independent variables; association per variable type is assessed with a likelihood ratio test (LRT), and directionality and per-variable contributions are quantified by effect size. Box colour indicates element type: process steps (red), data and variables (purple), and data sources (blue), as in the legend. The two panels at the bottom are representative example outputs: (left) a heatmap of FDR-corrected LRT *p*-values per signature and removed variable, and (right) an incidence rate ratio (IRR) forest plot from a zero-inflated negative binomial (ZINB) regression. Design choices and parameters are described in Sections 2.1–2.5

2. Materials and methods

An overview of the methodology is shown in Figure 1. First, for a set of participants, we obtained DNA sequencing data and associated clinical, biological, and technical metadata from the UK Biobank , comprising cancer type diagnosis, gene variants present, age, smoking status, and sequencing provider. To analyse the influence of these variables on the mutational signature profile, we treated them as the independent variables.

Since mutational signatures capture somatic processes, the DNA was filtered for somatic mutations, from which only single base substitutions, mutations in which one DNA nucleotide is replaced by another, were retained. These substitutions were then aggregated by mutation type and mutational context, where the mutational context refers to the immediately flanking bases. For example, A[C>T]G denotes a C-to-T substitution flanked by an A on one side and a G on the other. Each substitution falls into one of 96 possible mutation types, yielding a 96-element vector that we refer to as the mutational catalogue.

From the mutational catalogue we analysed four different aspects of the data. The first was the mutation burden, the total number of mutations. The second were the mutation channels: the 96-element vector normalised so that we captured only its composition and not the overall burden. Third, we "extracted" *de novo* mutational signatures, meaning we mathematically obtained signatures from the

set of mutational catalogues. A signature is a characteristic mutation pattern, and each sample is described by an *exposure* vector: how many mutations each *de novo* signature contributes to that sample. Finally, we "fitted" the mutational catalogue. This means that rather than discovering signatures from the data, we used a reference mutational signature set, in our case COSMIC signatures: a catalogue of curated reference mutational signatures. Each sample is then again described by an *exposure* vector: how much each COSMIC signature contributes to that sample.

To answer our research questions, we wanted to analyse the effect of the independent variables (i.e. the clinical and biological variables) on the dependent variables representing the mutational landscape of the blood: mutation burden, mutation channels, exposure to *de novo* signatures and exposure to COSMIC signatures. We performed regression analysis and assessed the association per variable type using the likelihood ratio test. We examined the directionality, as well as the specific variables driving this association, by calculating the effect size.

The remainder of this section follows this workflow: we describe the data and predictor variables (Section 2.1), how raw variants are filtered into a somatic mutational catalogue (Section 2.2), how signatures are extracted and fitted to produce the dependent variables (Sections 2.3 and 2.4), and finally the regression models linking predictors to each target (Section 2.5).

2.1. Data collection

The UK Biobank is a large-scale biomedical database containing genetic, lifestyle and health information from half a million UK participants [19]. From these, we selected participants according to their cancer diagnosis and the presence of high-impact germline variants. For each, we collected the variant call files, which are listings of the genetic variants identified in each participant.

2.1.1. Biological and clinical predictor variables

We selected several predictor variables to study their effects on the mutational landscape, grouped into three categories.

The first category concerns biological processes. We wanted to analyse the influence of different biological mechanisms on the exposure to specific COSMIC mutational signatures. To do this, we connected the presence of certain high-impact gene variants to specific processes. These processes are in turn either the direct cause of, or significantly associated with, the presence of specific COSMIC mutational signatures [20; 21]. These connections are shown in Table 1.

The second category concerns cancer diagnoses. We hypothesised that a prevalent cancer case at the time of DNA sequencing (i.e. an existing diagnosis) might be associated with the presence of certain mutational signatures. Furthermore, there is also a large set of patients who, in the years following their DNA sequencing, were diagnosed with cancer [22]. Finding possible associations between mutational signatures and a future cancer diagnosis could aid in understanding the processes involved in developing a tumour. We therefore included several cancer types as predictor variables.

We selected lung cancer, breast cancer, and three subtypes of haematological malignancies: leukaemia, lymphoma, and myeloma. We also included a control cohort comprising samples without a tumour diagnosis. Each cancer type was selected based on its potential connection to other variables in the study or to the nature of the sequencing source. Lung cancer was chosen for its known association with smoking; breast cancer for its potential link to high-impact variants in BRCA genes. The haematological malignancies were selected because the DNA used for sequencing was derived from blood, making these cancers the most likely to leave detectable traces in the source material.

The third category comprises age and smoking status, both hypothesised to influence mutation burden or mutational signature exposures.

Finally, sequencing provider was included as a technical covariate to account for potential batch effects introduced by differences in sequencing protocols or processing pipelines. Although all samples were sequenced on the same Illumina NovaSeq platform, the providers differed in library-preparation protocol, bioinformatics processing pipeline, and quality-control regime, which can introduce systematic, provider-correlated variation [23]. Adjusting for provider therefore guards against such artefacts being mistaken for biological signal.

2.1.2. Metadata

Each cancer type (plus the control group) forms a cohort, the set of participants sharing that diagnosis, and our analyses compare different aspects of the mutational landscape across these cohorts. Because cohorts differ in size and age structure, we applied filters to keep the comparison fair. We enforced a minimum age of 40, since mutation burden accumulates with age and the cohorts differ in age distribution. The control cohort was far larger than the others (n = 146,224), giving it broader distribution tails and more extreme values; to limit the resulting outlier influence, we further restricted it to participants aged 40–70 with a body mass index between 20 and 40 kg/m^2.

2.1.3. Variant Call Format files

The genetic data comes from whole-genome sequencing (WGS), a readout of each participant's complete genome. Genetic variants were called from this sequencing and stored as as Variant Call Format (VCF) files. The DNA itself was extracted from blood, specifically the buffy coat, the fraction richest in white blood cells. Crucially, the variants were identified with the GATK HaplotypeCaller. This is a tool designed to detect germline variants, which are mutations inherited and present in a consistent fraction of the DNA, rather than somatic mutations [24]. Such methods are therefore less sensitive to somatic mutations that appear in only a small fraction of cells. These are mutations with a low variant allele frequency (VAF), i.e. carried by only a small proportion of the sequencing reads at a given position [25].

2.1.4. COSMIC and supplementary reference signatures

We compared the extracted signatures to the COSMIC signatures. We restricted our scope to single base substitution (SBS) signatures, as these are the most prevalent mutation type in both adult (97%) [26] and paediatric (93%) [27] cancer. We used COSMIC version 3.5, the most recent release at the time of analysis. Many of these signatures have a listed aetiology, which can be used to biologically interpret the extracted signatures.

Furthermore, we added several mutational signatures that were found in the literature, but not present in the COSMIC set. First, we added the following set of mutational signatures [N9, N13, N15] from a study on multiple myeloma [28]. These signatures did not meet the 0.85 cosine similarity threshold required for a match with

Table 1 Genes, the DNA-repair processes they affect, and the associated COSMIC mutational signatures. Each gene set is linked to a DNA-repair process whose disruption produces a characteristic set of COSMIC signatures. Samples carrying a high-impact variant in any gene of a set are annotated with the corresponding process label (BRCA, MMR, APOBEC, POLE or POLD1).

Genes	Biological process	COSMIC signatures
BRCA1, BRCA2	Homologous recombination deficiency (HRD)	SBS3, SBS8
MLH1, PMS2, MSH2, MSH6	Mismatch Repair (MMR) deficiency	SBS6, SBS14, SBS15, SBS20, SBS26, SBS44, SBS97
APOBEC1, APOBEC2, APOBEC3A, APOBEC3B, APOBEC3C, APOBEC3D, APOBEC3F, APOBEC3G, APOBEC3H, APOBEC4, AICDA	APOBEC activity	SBS2, SBS13
POLE	DNA Pol ϵ dysfunction	SBS10a, SBS10b, SBS14
POLD1	DNA Pol δ dysfunction	SBS10c, SBS10d, SBS20

COSMIC signatures, while also not marked as possible sequencing artefacts. Another added reference signature is a signature coined "SBSBlood", found by several investigations studying lymphocytes in blood [29; 30]. Finally, a large study [20] investigating 18,640 cancers developed a framework allowing for the extraction of rare mutational signatures[1]. Here we only selected signatures that were extracted independently multiple times and/or reported in orthogonal studies, and found in at least 1 sample with lymphoid or myeloid cancer.

2.2. Somatic filtering strategy of VCF files

Before the mutational catalogue can be built, the raw variants in each VCF file must be filtered to isolate true somatic mutations, removing germline variants, technical artefacts and low-quality variant calls. However, since signatures are extracted as patterns across a large number of samples, overly stringent filtering is counterproductive if removing false positives comes at the cost of many false negatives. We hypothesized that random false positives are largely absorbed as noise rather than distorting the extracted signatures. The real problem is non-random errors: a patterned artefact can form a signature indistinguishable from one reflecting a biological process. Some are recognisable, as several COSMIC signatures are already annotated as artefactual, but artefacts arising from our own methodology have no reference label, which is an important caveat when interpreting the extraction results.

Variant filtering followed the steps below, adapted from a study based on a comparable biological source: whole-genome sequencing of peripheral blood at an average depth of 38× [31]. Unless otherwise cited, all steps derive from this study.

- **Quality control:**
  - We perform several quality control steps, such as a minimum sequencing depth of 20 and a maximum sequencing depth of 100 (as this can indicate complex mapping regions).
  - We only keep variants that have at least two reads.
- **Removing artefacts:**
  - First of all, we restrict mutation calls to regions where a read length of 100 base pairs was uniquely mappable to GRCh38
  - Added to this, we filter out several regions with a higher likelihood of artefacts. These are low complexity regions, segmental duplications, centromeres and contigs that differed in sequence between hg19 and GRCh38.
  - Finally, we remove all variants that are in a standardised Panel of Normals[2], which contains a list of common artefacts.
- **Removing germline variants:**
  - All variants from the set that are known to be common germline variants are removed. These are variants that appear in the gnomAD database with an allele frequency of at least 0.05 [32].
  - Furthermore, a one-sided binomial test was performed. This test is based on the assumption that all heterozygous germline mutations follow a binomial distribution centred around a VAF of 0.50. This means that of the different reads, half of them are one DNA base, the other half of the reads are another DNA base. With a p-value of 0.05, only variants that violate the null hypothesis i.e. have a significantly lower number of reads and are thus unlikely to be a germline variant, are kept, and the other variants are removed. This test is one-sided because in normal blood (as opposed to tumour cells), somatic mutations are very unlikely to be present in such a high percentage, and therefore are unlikely to be present in the VAF range of 0.50 to 1.00. [33].

2.3. De novo signature extraction with SigProfilerExtractor

This extraction step produces the *de novo* signature exposures, while the fitting step that follows (Section 2.4) produces the COSMIC signature exposures; together these form the signature-based dependent variables of our analysis. Mutational signature analysis assumes that an observed mutational catalogue can be expressed as a combination of underlying signatures:

$$C \approx S \times E \tag{1}$$

Here, S is a matrix of mutational signatures, where each column defines a signature as a probability distribution over mutation types. E is the exposure matrix, where each entry represents the number of mutations in a given genome attributable to a given signature. C is the resulting mutational catalogue: a matrix of observed absolute mutation counts per mutation type per genome.

De novo mutational signature extraction refers to the unsupervised discovery of mutational signatures directly from the mutational catalogue. Given C, the goal is to find matrices S and E such that $C \approx S \times E$, where both are unknown. This is typically solved using non-negative matrix factorisation (NMF).

[1] https://signal.mutationalsignatures.com/

[2] https://gatk.broadinstitute.org/hc/en-us/articles/360035890631-Panel-of-Normals-PON

Once we had a mutational catalogue containing the aggregated mutations of all samples, we used *SigProfilerExtractor* to perform this extraction [34]. The tool works as follows:

1. The catalogue is arranged as a matrix C with n rows (mutation channels, e.g. A[A>T]G) and m columns (samples), each entry giving the absolute mutation count for that channel in that sample (Equation 1). This matrix is the input to the non-negative matrix factorisation (NMF).
2. $l = 100$ replicates of C are created, each subjected to independent Poisson resampling so that no two inputs are identical, which increases the stability of the final result
3. Each replicate C_l is factorised independently by NMF into a signature matrix S_l and an exposure matrix E_l, using the multiplicative update algorithm with random initialisation and the generalised Kullback-Leibler divergence as objective.
4. This is repeated for each candidate number of signatures k (from 1 to 25, standard parameters). For each k, the signatures from all replicates are clustered around k centroids by similarity, iterating until convergence; the final centroids are the consensus signatures, and each signature's stability is the silhouette value of its cluster.
5. The optimal k is chosen from the stability and reconstruction error. Solutions are stable if their signatures have an average stability above 0.80 and a minimum of 0.20 (standard parameters). Among stable solutions, we select the one with the fewest signatures whose increase in reconstruction error, relative to the most complex stable solution, is non-significant ($p > 0.05$, Wilcoxon rank-sum test).

Once the signatures and the optimal number of signatures have been determined, the *de novo* signatures are decomposed into COSMIC signatures. Decomposition refers to the post-hoc interpretation of *de novo* extracted signatures by mapping them to known reference signatures, typically using cosine similarity or non-negative least squares (NNLS). Unlike fitting (Section 2.4), where exposures are estimated directly from reference signatures, decomposition first extracts signatures *de novo* and then identifies which COSMIC signatures they most closely resemble. SigProfilerExtractor performs this in the following way:

6. Iteratively, COSMIC signatures are added to the list of signatures that explain the *de novo* signature, based on the NNLS algorithm. This means that the signatures that cause the greatest decrease in the L2 error are added, with a minimum threshold of 5% decrease to avoid overfitting. After the addition step, a removal step is performed, where the same process occurs, but then signatures causing the least decrease of L2, while also having a decrease under 1%, are removed. This process continues until no signatures are either added or removed.
7. Finally, for the exposures of the individual samples, the same process as the previous step is performed. First the presence of *de novo* signatures is determined, and then these signatures are decomposed into COSMIC signatures. These decomposed COSMIC signatures are used as a reference set, and are iteratively added and removed using NNLS.

2.4. Fitting COSMIC signatures to samples with SigProfilerAssignment

Fitting is the process of estimating exposures to a fixed, pre-defined set of signatures. Given C and a known S (Equation 1), which is in practice often the COSMIC reference signatures[3], the task is finding E such that $C \approx S \times E$, subject to constraints such as non-negativity. Because S is fixed, fitting is a simpler and more constrained problem than extraction.

Fitting can be a less convoluted process than decomposing signatures. As explained in the previous section, decomposing signatures consists of first extracting *de novo* signatures and then decomposing those *de novo* signatures into a reference signature set. This process is convoluted due to it involving a chain of transformations: mutations are first compressed into a *de novo* signature, which is then decomposed into reference signatures, so the final interpretation sits two abstraction layers away from the data and each step can introduce error that propagates downward. Furthermore, each of these steps involves its own parameters and thresholds, and tweaking them can significantly change the outcome.

Moreover, this chain of transformations also comes into play when performing regression analysis and investigating the results. If we observe an association between a variable and a *de novo* signature, we want to investigate this association by linking the *de novo* signature to a reference signature with an existing aetiology. To do this, we need to take several metrics into account. First, the stability of the *de novo* signature (its silhouette value; see section 2.3, step 4). Second, the cosine similarity of the *de novo* signature to its decomposition: a low value means the reference combination reconstructs the signature poorly. Third, the percentage of the signatures present in the decomposition: if the decomposition is spread thinly across many low-percentage references, no single process dominates, and attributing the association to one named signature overstates the case. Furthermore, even a stable, well-reconstructed decomposition leaves open which of its constituent signatures an association actually reflects. For example, suppose the variable POLD1 is associated with a *de novo* signature that decomposes into an artefact signature (83%) and a signature associated with POLD1 (17%). The fact that the decomposition mainly consists of an artefact signature, casts doubt on the biological validation between the POLD1 variable and the POLD1 COSMIC signature.

This is why we will also fit the COSMIC signatures per sample, and use those activities as dependent variables to regress against. We do this by applying SigProfilerAssignment [35]. SigProfilerAssignment works as follows:

1. First, per sample, it determines the original set of reference signatures by using the NNLS algorithm, by calculating for which combination of reference signatures, the error is minimized. This provides the best possible explanation of the data, but also results in overfitting.
2. To regularise the number of signatures, the same step is performed as step 6 of the extracting process. Iteratively signatures are added and removed, until the signature set converges.

[3] https://cancer.sanger.ac.uk/signatures/

2.5. Regression modelling of mutation burden, channels and signature exposures

To investigate the clinical and demographic determinants of mutation burden and mutational signature exposures, a series of regression models were fitted. The following sections describe the preprocessing steps applied to prepare the data for analysis, the regression models employed and their associated feature importance measures, and finally the structure of the regression targets and analyses performed.

2.5.1. Data cleaning and variable encoding

The dataset was cleaned and transformed prior to analysis. Multi-label cohort entries were converted into binary indicator variables, with the control cohort (indicating samples without a cancer diagnosis) omitted as the reference category. Smoking status entries labelled "Prefer not to answer" were removed, and the remaining categories were encoded into dummy variables with "Never" as the reference. The variable "sex" was omitted from the regression analysis due to its high correlation with the breast cancer diagnosis and the presence of a high-impact BRCA gene.

Sequencing provider was one-hot encoded with the most frequent category used as the reference. Boolean variables (which indicated the presence of a high-impact gene variant) were converted to integer format. Finally, the continuous variable age was standardised using z-score scaling to ensure comparability across features.

All of the samples ($n = 17,419$) were used as the source for the extraction process, but several of these samples were omitted from regression analysis due to missing data on predictor variables. This resulted in a dataset consisting of $14,489$ samples.

2.5.2. Regression models and feature importance measures

We performed regression analysis to investigate the role different variables play in determining the outcome. This means that given a result i.e. the dependent variables Y_i, we wanted to investigate how much the independent predictor variables $X_0, X_1..X_n$ influenced the total result.

Negative binomial regression

Negative binomial (NB) regression is a type of generalized linear model in which the dependent variable Y is a count of the number of times an event occurs [36; 37]. It is commonly used for overdispersed count data, where the variance exceeds the mean.

The model assumes each count Y_i follows a negative binomial distribution with mean μ_i and a dispersion parameter α. The link function, the function that connects the model to the response variable is the log function:

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} \quad (2)$$

Somatic mutation counts are overdispersed across samples and genomic regions, due to factors such as inter-patient heterogeneity [38]. Poisson models, which assume mean equals variance, therefore generate inflated false-positive rates in this setting, and NB regression has become the standard alternative for modelling mutation burden [39; 40]. Because of this overdispersion, we used NB regression to model the mutation burden in our data.

Zero-inflated negative binomial regression

When count data contain a higher proportion of zero observations than expected under the NB distribution, a zero-inflated negative binomial (ZINB) model can be more appropriate [36]. ZINB models the response as a mixture of two processes: a logistic component generating structural zeros (samples in which the process cannot occur) and a negative-binomial component generating the remaining counts. This is biologically motivated for exposures to signatures, where a substantial fraction of samples have zero contribution from a given signature because the underlying mutational process was inactive, distinct from samples with low but non-zero exposure.

To determine whether to use the negative binomial model or the zero-inflated negative binomial model, the Akaike information criterion (AIC), was used. This measures the quality of the model by combining the goodness of the fit and the number of parameters [41]. The lower the AIC, the better. We compared the two models, NB and ZINB, and selected the one with the lower AIC. In practice, this meant that for the majority of the signature exposures, we used ZINB.

Feature importance: Incidence rate ratio

In negative binomial regression, the influence of each predictor can be interpreted through the model coefficients in equation 2 by exponentiating them to obtain the incidence rate ratio (IRR). For example, a coefficient of $\beta_1 = 0.0279$ yields $e^{0.0279} \approx 1.03$, meaning that a one-unit increase in x_1 is associated with a 3% increase in the dependent variable Y, holding all other variables constant [36]. For categorical variables encoded as binary indicator variables, the IRR is always interpreted relative to the reference category. For instance, as described in Section 2.5.1, the cohort variable was encoded into a set of binary indicators with the Control cohort as the reference. An IRR of $IRR_{\text{lymphoma}} \approx 1.06$ therefore indicates that, compared to the Control cohort, a lymphoma diagnosis is associated with a 6% increase in the dependent variable.

Beta regression

Beta regression is used for data that can be modelled by a beta probability distribution [42]. It is specifically designed for dependent variables that represent rates, proportions, or percentages and are strictly bounded between 0 and 1. The model assumes each observation Y_i follows a beta distribution with mean μ_i and a precision parameter ϕ, where larger ϕ implies less dispersion around the mean.

The link function is:

$$\text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \quad (3)$$

Mutational catalogues are inherently compositional: each tri-nucleotide channel represents a proportion of the total mutation burden, and the channels sum to one across each sample [43]. Modelling raw counts ignores this constraint and conflates absolute burden with relative composition. Beta regression directly respects the (0,1) support of each channel's proportional contribution. Furthermore, unlike Dirichlet-multinomial models, which fit all channels at once, beta regression fits each channel separately, so every channel gets its own predictor effects [44]. This is why we used beta regression to independently model the mutation channels. The drawback is that fitting channels separately ignores how they are linked: because the proportions sum to one, they are inherently negatively correlated, and per-channel models cannot capture this or

recover signature-level structure on their own. We accept this trade-off here because our aim at the channel level is the interpretability of per-channel effects; joint, signature-level structure is captured separately through the *de novo* and COSMIC exposures.

The beta distribution fails if values are exact zeroes, which can occur when a sample has no mutations in a given channel. Therefore, per channel, we selected a version of the model. Channels with no zeroes used standard beta regression; those with a small zero fraction (below 5%) used beta regression after the data was transformed with the Smithson–Verkuilen transformation which compresses the data into $(0, 1)$ [45]. Channels with at least 5% zeros used a zero-inflated beta (ZIB) model [46], fit in two independent parts: a logistic component for the probability of an exact zero, and a beta component for the proportion among non-zero observations. As with the zero-inflated negative binomial model used for signature exposures, this separates structural zeros from small non-zero contributions.

Feature importance: Odds ratio

The influence of each predictor in beta regression can be interpreted through the model coefficients in equation 3 by exponentiating them to obtain the odds ratio (OR). For example, a coefficient of $\beta_1 = 0.0296$ yields $e^{0.0296} \approx 1.03$, meaning that a one-unit increase in x_1 is associated with a 3% increase in the odds $\mu_i/(1 - \mu_i)$, holding all other variables constant [42]. Just like the IRR, the OR is interpreted relative to the reference category for categorical variables. Unlike the IRR, which directly scales the expected count, the OR scales the odds of the proportion.

Feature importance: Likelihood Ratio Test

To quantify the individual contribution of each covariate to overall model fit, a series of likelihood ratio tests (LRTs) were conducted [47]. For each predictor or, in the case of categorical variables, the full set of categorical binary variables, a reduced model was estimated with the variable of interest omitted. The likelihood ratio statistic, defined as $2(l_{full}$ - $l_{reduced})$, where l is the log-likelihood, was computed. The p-value was calculated by using a χ^2distribution, with degrees of freedom corresponding to the difference in the number of free parameters between the full and reduced models. Variables with a p-value below 0.05 were considered to contribute significantly to explaining the predicted variable.

Multiple testing correction

For experiments where many different dependent variables were tested, p-values were adjusted using the Benjamini-Hochberg procedure to control the False Discovery Rate (FDR) at 5%. This method ranks the m observed p-values and declares the largest k significant for which $P_{(k)} \leq \frac{k}{m}\alpha$, ensuring that the expected proportion of false discoveries among significant results does not exceed α [48]. Multiple testing correction was applied to the LRT p-values, since these test whether each covariate explains variation in the outcome and form the relevant family of hypotheses across the 96 channels or the set of signatures.

Because certain signatures are active in only a small subset of samples, individual variable effects can become statistically inestimable. For instance, if a rare genetic variant never co-occurs with exposure to a specific signature, the model coefficient cannot be identified. To guard against this, we applied an events per variable (EPV) heuristic [49], requiring a minimum of 10 samples where the predictor variable was positive and the target signature was active. Any variable × signature pair that did not satisfy this threshold was removed from the analysis and excluded from the downstream family of hypotheses. This exclusion is a statistical safeguard to ensure coefficients can be estimated reliably; it does not imply the association is biologically meaningless. On the contrary, the absence of co-occurrence that makes a coefficient inestimable, for example, a variant that is never observed alongside a given signature, may itself be biologically informative, even if it cannot be captured as a regression effect within this framework.

2.5.3. Regression targets and analysis structure

While the independent variables and their processing remained the same across all experiments, the dependent variable differed between analyses.

First, negative binomial regression was performed on the total **mutation burden**. Given a mutational catalogue C in which each sample is represented as a vector of mutation counts per mutation type, the mutation burden is the sum over all entries of that vector. This follows the established use of NB regression to relate per-sample mutation counts to clinical and demographic covariates [39; 40].

Second, we analysed the relative contribution of each **mutation channel**. Since mutation burden was already modelled directly, we normalised the channel counts by dividing each channel's mutation count by the sample's mutation burden, thereby focusing on the compositional profile rather than the absolute count. We used beta regression to model the proportions. As there are 96 channels, 96 independent models were fitted, and multiple testing correction was applied.

Third, zero-inflated negative binomial and standard negative binomial regression was performed with the **exposure to each *de novo* signature** as dependent variable. Since 15 *de novo* signatures were identified, multiple testing correction was again applied.

Finally, zero-inflated negative binomial and standard negative binomial regression was performed with the **exposure to each COSMIC signature** as dependent variable. Multiple testing correction was applied.

3. Results and discussion

We present our findings following the four views of the mutational landscape introduced in the methodology: total mutation burden, the relative composition of mutation channels, exposure to *de novo* extracted signatures, and exposure to fitted COSMIC signatures. For each, we assess how the clinical and biological predictors shape the outcome, and interpret the results in light of known mutational processes. However, we first investigate the somatic filtering which produces the mutational catalogue, and is thus the source of the four analyses. A recurring theme across all four analyses is that sequencing provider, a technical variable, accounts for a substantial share of the variation. This observation motivates the provider-stratified re-extraction in Section 3.4.3, in which we try to separate biological signal from this technical confound.

3.1. Somatic variant filtering reveals residual C>A and T>G artefacts

Because every downstream view (burden, channels, and signatures) is computed from the filtered catalogue, we first ask how clean that

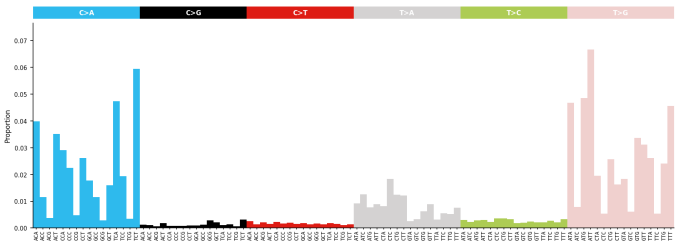


Figure 2 Average mutational catalogue across all filtered samples (n = 17,419). For each of the 96 single-base-substitution channels (the six substitution types C>A, C>G, C>T, T>A, T>C, T>G, each split by 5′ and 3′ flanking base), the mutation count is averaged across samples and the resulting 96-element profile is normalised to sum to one. The x-axis shows the mutation channel, grouped and coloured by substitution type; the y-axis shows the proportion each channel contributes to the total.

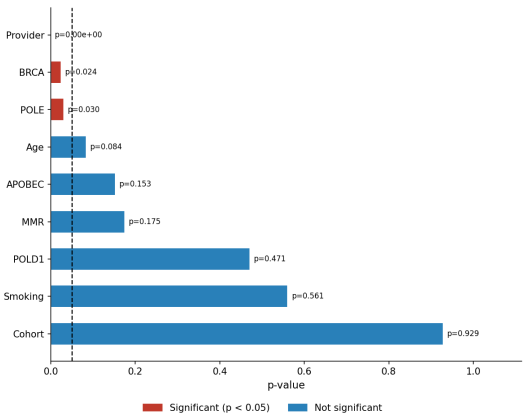


Figure 3 Likelihood ratio test (LRT) p-values for each variable in the negative binomial mutation-burden model, with sequencing provider included. Each variable (or, for categorical variables, the full set of levels) is dropped from the full model and the reduced model compared by LRT. Bars show the resulting p-value; the dashed line marks the $\alpha = 0.05$ threshold, and bars are coloured by significance (red, $p < 0.05$; blue, not significant). BRCA, MMR, POLE, POLD1 and APOBEC denote samples carrying a high-impact variant in the corresponding gene set.

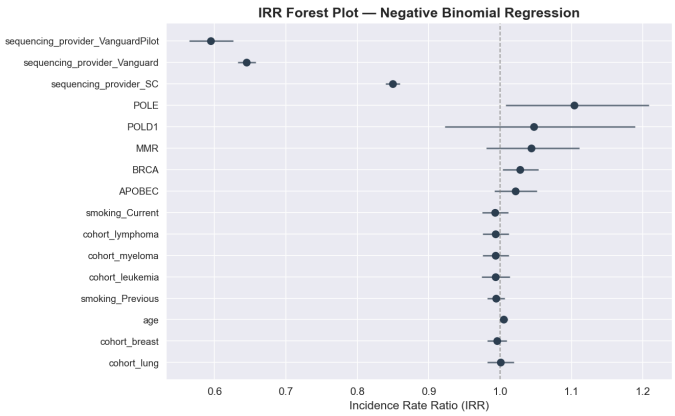


Figure 4 Incidence rate ratios (IRRs) with 95% confidence intervals for each covariate in the negative binomial mutation-burden regression. Points are IRR point estimates and horizontal bars the 95% confidence interval; the dashed line at IRR = 1 marks no effect (values >1 indicate increased burden). Categorical variables (sequencing provider, smoking status, cohort) are shown as individual levels relative to their reference category [provider = DECODE, smoking = Never, cohort = Control (no cancer diagnosis)].

catalogue is: to what extent do the retained variants plausibly reflect real somatic mutations rather than residual germline calls or technical artefacts? After the VCF Filtering Process, we end up with 17,419 filtered VCF files. In Figure 2, the average mutational catalogue can be seen. The average cosine similarity between all of the mutational catalogues is 0.91, with the maximum similarity 0.9997 and the minimum similarity 0.28.

The high similarity is consistent with a large shared component across samples. This could partly reflect ubiquitous clock-like somatic processes, which are processes which accumulate mutations at a roughly constant rate with age. However, the specific channels enriched in our case (C>A and T>G) point to residual technical artefacts rather than biology alone. To rule out a germline origin for these channels, we compared them against germline mutational spectra: the most common de novo germline mutations occur in the C>T and T>C channels [50; 51], whereas C>A and T>G each constitute only around 10% and 8% of germline variation, respectively [50]. The enrichment we observe is therefore unlikely to be driven by germline variants, consistent with a technical rather than germline source.

When looking at the common mutational context of artefacts, we can relate this to our findings. Specifically, many C>A mutations with a VAF of $< 20\%$ can be attributed to the oxidation of guanine to 8-oxoguanine during library preparation [52]. For the T>G mutations, we can look closer at the sequencing pipeline used by the UK Biobank. The UK Biobank uses Illumina NovaSeq sequencing machines [53]. Research has found that these machines introduce artefactual T>G mutation calls, which can confound the detection of low-VAF somatic variants in high-depth sequencing samples, particularly in studies of mosaic mutations in normal tissues, where variants have low read support and are called without a matched normal [54].

3.2. Mutation burden: BRCA and POLE variants predict increased burden

We begin with the first of the four aspects defined in Section 2.5.3: the mutation burden, the total number of somatic mutations per sample, before decomposing the catalogue into its compositional and signature-level structure.

To identify which biological and technical factors drive mutation burden, a negative binomial regression was fitted and evaluated using both a LRT and the IRRs, shown in Figures 3 and 4. In Figure 3, the results of the LRT are shown, with provider included. The LRT identifies BRCA and POLE mutations as the only non-technical variables reaching statistical significance ($p = 0.024$ and $p = 0.030$), both associated with an increased burden. POLE shows a larger point estimate (IRR ≈ 1.11) but considerably wider confidence intervals than BRCA, reflecting the small nature of the POLE-mutant subgroup. APOBEC and MMR fail to reach significance, which could be caused by reduced power from small group sizes. Cohort, smoking and POLD1 contribute minimally to model fit and have IRRs clustered tightly around 1.0. Sequencing provider is a strong technical confounder: it contributes significantly to model fit ($p < .000001$), exceeding all biological variables.

3.3. Mutation channels: associations with leukaemia and smoking

Having modelled the absolute burden, we turn to the second view: the relative composition of the catalogue. Normalising each channel to sum to one removes the influence of total burden and isolates the compositional profile, which we model with beta regression rather than counts (Section 2.5). Figure 5 shows the per-channel LRT test after applying Benjamini-Hochberg correction, and Figure S3 in the supplementary shows the coefficients of the model.

Sequencing provider was the strongest predictor in the model, with the largest coefficient magnitudes and near-zero FDR-corrected p-values across nearly all channels. This reflects technical variation between sequencing centres rather than biology and should be treated as a confound.

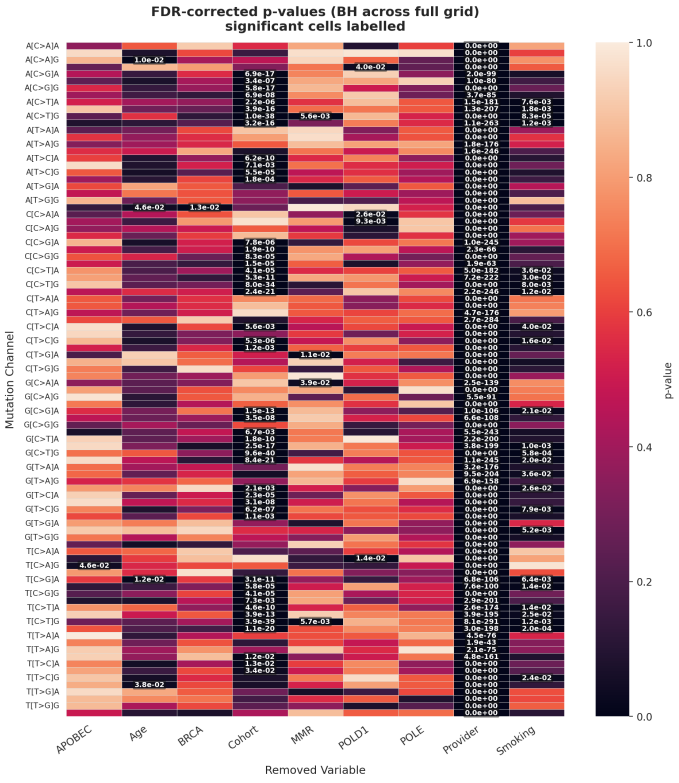


Figure 5 Per-channel likelihood ratio test results for the beta regression of mutation-channel composition. Rows are the 96 substitution channels (in standard SBS-96 order). Only alternate labels shown (A/G contexts); intervening rows are C/T contexts. Columns are the variables removed from the full model. Cell colour shows the FDR-corrected p-value (Benjamini-Hochberg, applied across the full channel × variable grid); darker cells indicate smaller p-values. Cells reaching significance ($p < 0.05$) are labelled with their corrected p-value.

Among the cohort variables, the strongest signal was at *[C>T]G channels, which is among the most frequent and well-characterized mutations seen across human genomes [55]. These are also the defining mutation types of SBS1 (in Figure 6), which is the clock-like signature caused by deamination of 5-methylcytosine to thymine [6; 10]. This was driven predominantly by the leukaemia cohort. Interestingly, MMR showed OR coefficients of similar

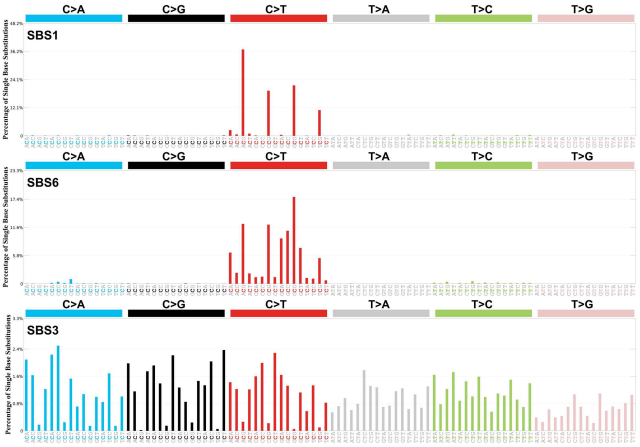


Figure 6 Reference profiles of three COSMIC SBS signatures, shown to contrast peaked and flat signatures. SBS1 (top), the clock-like signature from 5-methylcytosine deamination, is sharply peaked at [C>T]G; SBS6 (middle), associated with mismatch-repair deficiency, is concentrated in C>T channels; SBS3 (bottom), associated with homologous recombination deficiency, is flat, with signal spread across most channels. The x-axis shows the 96 substitution channels grouped by substitution type (C>A, C>G, C>T, T>A, T>C, T>G); the y-axis shows the percentage of single base substitutions per channel. Signature definitions from COSMIC v3.5.

magnitude at these same channels (e.g. T[C>T]G: MMR $\beta = 0.128$, cohort_leukaemia $\beta = 0.111$). This can be explained by the fact that the signatures SBS6 (in Figure 6) and SBS15, both associated with MMR mechanisms [20], are also enriched for [C>T]G.

POLD1 showed its highest positive coefficients at C[C>A]* channels, though FDR significance was only reached for C[C>A]A and C[C>A]C (p-value = 0.026 and 0.009) but not the G and T flanking contexts, most likely due to limited numbers of POLD1-mutant cases in the cohort. The picture is mixed: two of the four POLD1-significant channels (C[C>A]A, C[C>A]C) are prominent in the POLD1-associated SBS20, whereas T[C>A]T, which is prominent in SBS10c/SBS10d, shows no POLD1 association (p = 0.33).

POLE was not significantly associated with any mutation channel (minimum p = 0.084 at T[T>G]C). For POLE, just like POLD1, there was no direct connection between mutation channels with high coefficients and the mutational signatures associated with POLE: SBS10a, SBS10b and SBS14.

BRCA reached significance at only one channel after FDR correction (A[T>G]T, p = 0.013). This is not consistent with the signatures associated with BRCA: SBS3 (in Figure 6) and SBS8 [56]. However, these signatures are more difficult to detect due to their flat mutational profile, meaning that mutations are spread across different mutation types, instead of in few small peaks. This is especially limiting in our model, which uses normalised mutation proportions: because the channels must sum to one, a gain in one channel forces a compensating loss in others. A flat signature like SBS3, whose effect is spread thinly across many channels, is therefore difficult to distinguish from compositional rearrangement of more peaked signatures. Furthermore, these signatures are only indirectly associated with BRCA-ness. SBS3, and to a lesser extent SBS8, have been associated with homologous recombination deficiency (HRD), but not all HRD cases are caused by germline BRCA1/2 variants [8], and conversely not all germline BRCA1/2 carriers develop HRD tumours, since this requires biallelic loss of the wild-type allele [57].

Signature	SBSA	SBSB	SBSC	SBSD	SBSE	SBSF	SBSG	SBSH	SBSI	SBSJ	SBSK	SBSL	SBSM	SBSN	SBSO
Stability	0.98	0.98	0.91	0.99	1.0	0.98	1.0	1.0	1.0	0.93	0.91	0.98	0.75	0.99	0.98
Active in samples	16641	17081	15932	16998	16853	16239	16150	15978	15961	15436	15293	15715	14963	14213	15145
Average exposure (incl. zeros)	4776.7	3923.1	3926.8	3750.3	3195.4	3162.6	3138.9	2988.6	2701.8	2401.0	2233.8	2190.8	1912.2	1566.5	1503.0
Average exposure (excl. zeros)	5000.1	4000.7	4293.3	3843.2	3302.7	3392.4	3385.5	3258.1	2948.6	2709.4	2544.3	2428.4	2226.1	1919.9	1728.7

Table 2 Metrics for *de novo* signatures. SBS refers here to SBS96 signatures, which is the mutational profile using the conventional 96 mutation type classification

For smoking, 25 mutation channels were elevated, of which the majority were [C>T] mutations. This is notable, as the canonical tobacco signature SBS4 is dominated by C>A. The absence of a clear C>A signal is itself unsurprising: signatures related to tobacco smoking tend to be broader across trinucleotide contexts than the sharply peaked POLE/POLD1 signatures, with much of their signal distributed across many contexts at modest per-channel contribution rather than concentrated in a few, which may make them harder to identify from a small set of enriched channels in a per-channel framework. The C>T enrichment is harder to attribute, however. The smoking effect is estimated with age, provider, and cohort held constant, so it is unlikely to be a simple artefact of those variables; but a statistical association with smoking need not reflect a direct tobacco mutagenic mechanism, and a C>T signal is inconsistent with SBS4. It is therefore more compatible with smoking covarying with a C>T-generating process, than with recovery of a tobacco signature.

3.4. *De novo* signatures are dominated by sequencing provider

We now turn to the third view: the latent signature structure of the catalogue. Having characterised the overall burden (Section 3.2) and its compositional profile (Section 3.3), we ask whether the somatic mutations in blood organise into a coherent, biologically interpretable set of de novo signatures, and what processes those signatures reflect. We approach this in three steps. First, we extract the signatures and characterise them descriptively, examining their stability, prevalence across samples, and similarity to known COSMIC signatures (Section 3.4.1). This raises the possibility that the dominant axis of variation is technical rather than biological. Second, because provider differences rather than biology may drive most of the variation in signature exposure, we test this formally through regression, asking whether any biological covariate explains signature exposure once provider is accounted for (Section 3.4.2). Third, because the pooled extraction is dominated by this batch effect, we re-extract signatures separately within each provider, reasoning that removing the between-provider variance should allow weaker biological signal to surface (Section 3.4.3).

3.4.1. Extraction yields 15 provider-driven signatures

When running the extraction of the *de novo* signatures on the mutational catalogue, SigProfilerExtractor determines that the optimal number of signatures is 15, according to the process described in section 2.3, step 5. This can be seen in Figure 7. The metrics for these *de novo* signatures can be found in Table 2, while the exposure distribution per signature can be seen in Figure 8. With the exception of SBS96M (stability = 0.75), all signatures have a stability of at least 0.91. All of the signatures are detected in at least 14,213 samples, meaning at least 81.6% of the samples. This suggests that, while the extraction process is able to find processes that are common among many samples, signatures that are very localised i.e. those

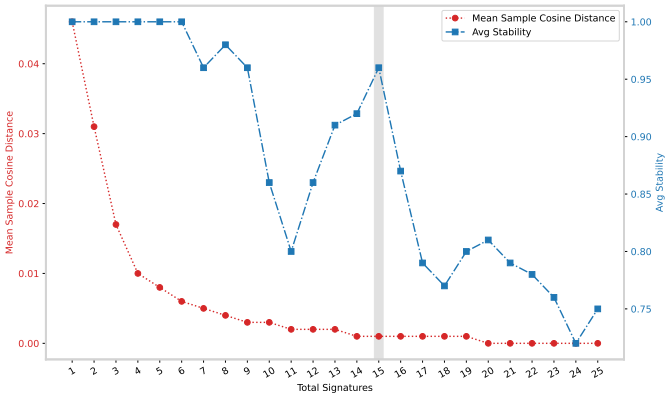


Figure 7 Signature-number selection plot from SigProfilerExtractor (Section 2.3, step 5). For each candidate number of *de novo* signatures (x-axis), the red line shows the mean sample cosine distance between each reconstructed and original sample (lower indicates better reconstruction), and the blue line the average stability of the solution (the mean silhouette value of the clustered signature replicates; higher is more reproducible). The highlighted column marks the chosen solution of 15 signatures.

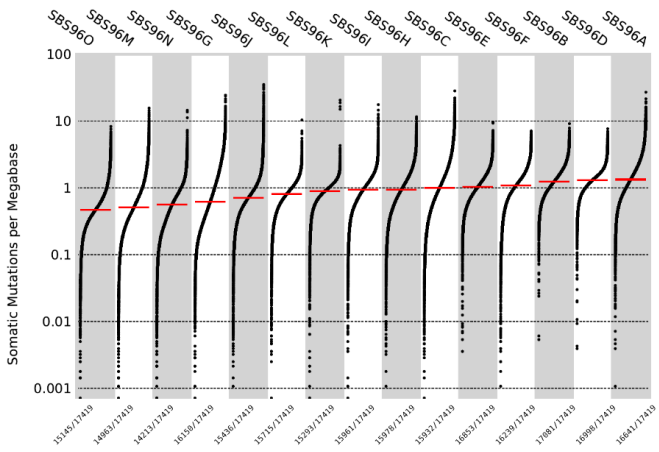
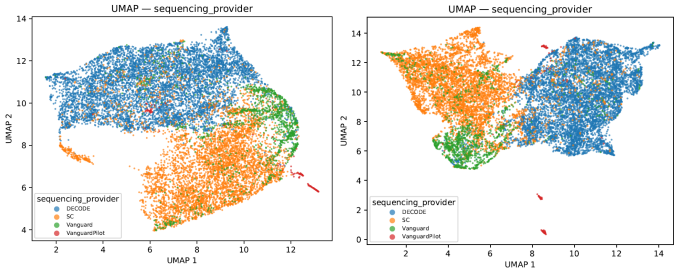


Figure 8 Distribution of *de novo* signature activity across samples. Each vertical strip corresponds to one of the 15 *de novo* signatures; within a strip, each point is one sample, positioned by its activity (somatic mutations per megabase, where one megabase is 10^6 base pairs) on a log-scaled y-axis. The red dash marks the median activity per signature. The fraction beneath each strip gives the number of samples in which the signature is detected (non-zero activity) out of 17,419 total. Strips are ordered by the median.

that are only present in a subset of the samples, are not captured by the extraction process. If so, this likely means that *de novo* signatures that we find are either caused by ubiquitous background biological processes, such as clock-like processes, or signatures related



(a) Absolute exposure (b) Relative exposure

Figure 9 Two-dimensional UMAP embedding of the per-sample *de novo* signature exposure vectors, with samples coloured by sequencing provider (DECODE, SC, Vanguard, VanguardPilot). (a) Absolute exposures (mutation counts per signature); (b) relative exposures (per-sample exposures normalised to sum to one).

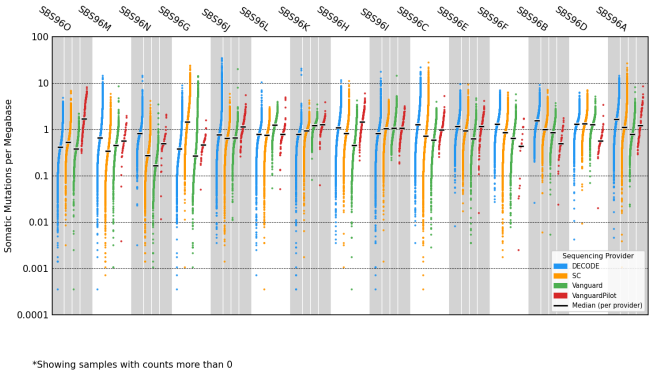


Figure 10 Per-provider activity of each *de novo* signature. For each of the 15 signatures, samples are split into four strips by sequencing provider (DECODE, SC, Vanguard, VanguardPilot); each point is one sample, positioned by its activity (somatic mutations per megabase) on a log-scaled y-axis. Black bars mark the per-provider median. Only samples with non-zero activity for a given signature are shown, so distributions reflect activity among carriers rather than all samples. Provider medians differ only modestly within each signature, in contrast to the clear provider separation seen in the UMAP embedding.

to artefacts, though limited power to detect rarer signatures cannot be excluded.

When investigating a higher number of signatures, such as 20 or 25, we still observe *de novo* signatures that are present in at least 14,799 or 15,126 samples, respectively. Furthermore, based on visual inspection, the actual distributions of the activities of these signatures are very similar to each other, even when increasing the number of signatures. This suggests that increasing the number of signatures will not necessarily force the tool to extract signatures present from a smaller subset.

Uniform Manifold Approximation and Projection (UMAP) was used to investigate possible clusters in the sample absolute exposure data. The input for UMAP was a set of vectors, where each vector denoted one sample, and each value in the vector denoted the exposure to one *de novo* signature. Once this UMAP model was created and visualised, we coloured the samples based on different variables. This can be seen in the supplementary (Figure S9). The sequencing provider stood out in explaining the different clusters (Figure 9a).

This suggests that, at least in part, the SigProfilerExtractor tool extracts artefactual background signatures associated with the sequencing provider rather than purely biological processes. At the level of individual signatures the dependence is weaker as seen in the per-provider exposure plot (Figure 10), each signature is active across all providers and the per-provider medians differ only modestly, but these small, consistent per-signature shifts accumulate across the 15-dimensional exposure profile into the clear provider separation seen in the UMAP. Because sequencing provider is essentially uncorrelated with tumour type and gene-variant status ($|r| \leq 0.03$), this clustering is unlikely to reflect differences in the underlying biology of the samples sequenced by each provider, and is therefore most plausibly technical in origin. When investigating whether part of this clustering could be caused by the average mutation burden differing per sequencing provider, UMAP was also applied to the normalised exposure vectors. We observe the same thing. Other biological variables show no visible clusters, and sequencing provider variables still cluster together, as can be seen in Figure 9b.

Cosine similarity to COSMIC signatures

To analyse whether *de novo* signatures are similar to COSMIC signatures, we use the cosine similarity. While this differs among literature, the common threshold used for denoting a *de novo* signature as corresponding to a COSMIC signature is 0.85 [58; 59; 60]. The results can be seen in Table 3.

SBS96G is the most credible match: it has the highest similarity and maps to a single COSMIC signature (SBS45). However, SBS45 is an artefact signature attributed to 8-oxo-guanine introduced during sequencing [52]. The other two each map to multiple COSMIC signatures and are thus less credible, particularly as several are artefactual: both SBS56 and SBS46 are possible artefacts, while SBS10d reflects defective POLD1 proofreading, SBS36 defective base excision repair, and SBS12 an unknown aetiology. We also compared it to the custom set of signatures we created, as referenced in section 2.1.4, but no signatures proved to be similar.

3.4.2. Regression confirms provider as the dominant predictor

As shown in Section 3.4.1, the UMAP projection indicated that samples appeared to cluster by sequencing provider in their exposure profiles. We now test this formally and ask whether any biological covariate survives once provider is accounted for. Mutational signature exposures were modelled using (Zero-inflated) Negative Binomial regression.

Analysis of significant *de novo* signature × variable associations

In Figure 11, the heatmap of FDR-corrected p-values reveals a pattern across signatures: sequencing provider is overwhelmingly the dominant predictor, reaching extreme significance (FDR p-values

Table 3 *De novo* signatures matching COSMIC above the 0.85 threshold.

De novo signature	COSMIC match	Cosine
SBS96D	SBS10d / SBS36 / SBS56	0.96 / 0.92 / 0.96
SBS96G	SBS45	0.96
SBS96N	SBS12 / SBS46	0.88 / 0.88

ranging from 10^{-43} to $< 10^{-268}$) for every one of the fifteen *de novo* signatures. In contrast, of the remaining covariates (age, smoking, cohort, APOBEC, MMR, BRCA, POLD1, or POLE), only three survive multiple-testing correction. These are: SBS96A × age, SBS96K × cohort and SBS96N × cohort.

To further investigate these significant signature × covariate combinations, we need to look at the respective forest plots. For SBS96K, seen in Figure 12, provider effects again dominate, with Vanguard samples showing an IRR of 1.7 and SC samples 1.2; the only biological covariate reaching significance is the leukaemia cohort indicator (IRR ≈ 1.1), and its effect size is small. The decomposition of SBS96K can be seen in Table 4. This is notable, as SBS1 and SBS5, while they are clock-like signatures and correlated with age, they are also signatures typically found in acute myeloid leukaemia (AML) [2]. However, any biological reading is tentative: the leukaemia link could relate to the clock-like SBS5/SBS1 components (the latter only 6%), but the dominant component, SBS89 (40%), has no established aetiology, so the association cannot be cleanly attributed to a named process.

SBS96N resembles SBS96K: Provider effects dominate. Here, SBS96N has a significant positive association with the myeloma cohort, and a negative association with the breast cancer cohort, causing this significant association between cohort and this signature to remain after multiple testing correction. SBS96N's decomposition can be seen in Table 4, but except for the two clock-like signatures (SBS1 & SBS5), SBS12 has no association with myeloma found in literature [28]. SBS46 is artefactual.

SBS96A is not similar to any COSMIC signature (either directly COSMIC or decomposition), so no biological aetiology can be deduced. Together, these analyses demonstrate that the *de novo* SBS96 signatures extracted from this cohort are heavily confounded by sequencing provider, and that the apparent biological signal in individual signatures is, in most cases, small relative to the technical between-provider variation. Any downstream biological interpretation of these signatures must therefore explicitly account for sequencing provider as a covariate, and ideally be carried out on signature attributions in which this batch effect has been modelled out or corrected for.

To check whether the covariate effects from the pooled analysis were real or driven by between-provider differences, the same negative binomial regression was refit separately within each provider. VanguardPilot was excluded because of its small size (n = 165), leaving three strata: DECODE (reference, n = 7,441), SC (n = 5,572) and Vanguard (n = 1,311). Provider dummies were dropped within each stratum since they are constant there. Age, smoking, cohort, APOBEC, MMR, BRCA, POLD1 and POLE were each tested by leave-one-out likelihood-ratio test, with multiple testing correction applied within each stratum. As described previously in section 2.5.2, the variables POLE and POLD1 did not have enough positive samples in the Vanguard dataset, and were therefore omitted from the analysis.

After stratification, only two covariates reached $p < 0.05$ for any signature in any provider. This was SBS96N × smoking ($p = 0.027$) and SBS96K × cohort ($p = 1.14 * 10^{-10}$) in DECODE. For the significant predictors seen in the combined model fit, age for SBS96A, cohort for SBS96K, and cohort for SBS96N, only SBS96K × cohort reproduced within a single provider. No covariates were significant among all sequencing providers. However, because the datasets differ substantially in size, p-values are not comparable across providers: an identical underlying effect will appear significant in DECODE yet not in Vanguard purely through statistical power. We therefore

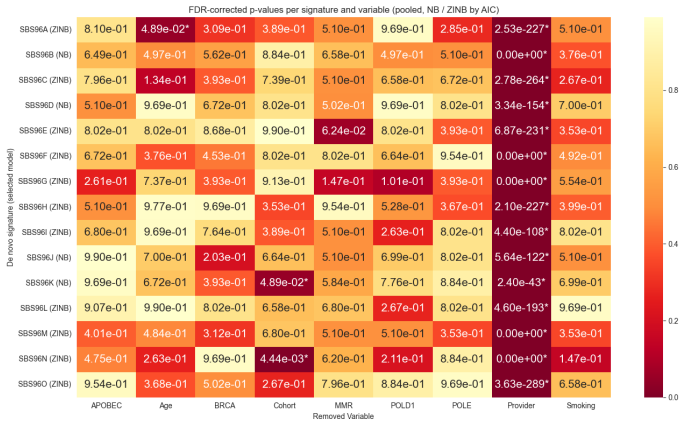


Figure 11 Likelihood ratio test results for the regression of *de novo* signature exposures. Rows are the 15 *de novo* signatures, with the count model selected per signature by AIC (NB, negative binomial; ZINB, zero-inflated negative binomial) given in parentheses; columns are the variables removed from the full model. Cell colour and the printed value show the FDR-corrected p-value (Benjamini-Hochberg); darker indicates smaller p-values. Cells reaching significance ($p < 0.05$) are marked with an asterisk.

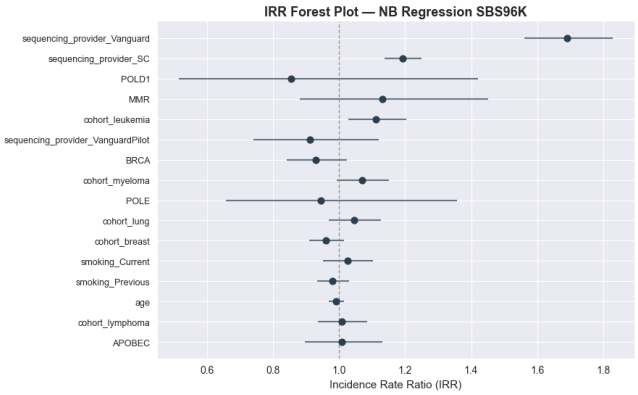


Figure 12 Incidence rate ratios (IRRs) with 95% confidence intervals for each covariate in the negative binomial regression of SBS96K exposure. Points are IRR point estimates and bars the 95% confidence interval; the dashed line at IRR = 1 marks no effect (values >1 indicate higher signature activity). Categorical variables (sequencing provider, smoking status, cohort) are shown as individual levels relative to their reference category [provider = DECODE, smoking = Never, cohort = Control (no cancer diagnosis).]

Table 4 Decomposition of the *de novo* signatures into COSMIC signatures, with the cosine similarity of the decomposition and the contribution of each COSMIC signature.

De novo signature	Cosine similarity	Decomposition
SBS96A	0.77	SBS28 (59%) + SBS43 (41%)
SBS96K	0.91	SBS89 (40%) + SBS5 (34%) + SBS43 (12%) + SBS10a (8%) + SBS1 (6%)
SBS96N	0.96	SBS46 (40%) + SBS5 (31%) + SBS12 (29%) + SBS1 (0.12%)

compared the three stratified datasets based on the scale of effect size, examining whether the per-provider incidence rate ratios (IRRs) and their confidence intervals were consistent. For SBS96A × age, the per-provider IRRs were 1.01, 1.03 and 1.02 (DECODE, SC, Vanguard) with confidence intervals that overlapped almost completely. For SBS96K × cohort, the results were similar, except for the lung cohort. Here, DECODE showed an IRR of 1.07, SC had an IRR of 0.99 and Vanguard 1.04. For the SBS96N × cohort association, Vanguard consistently showed a lower IRR compared to the other providers across all five cohort levels (IRRs = 0.75–0.94 versus = 1.0). In summary, stratification supports SBS96A × age as the most consistent biological association, while the cohort associations reproduced only partially and SBS96N × cohort in particular remained entangled with provider-specific variation. These results underline the need to model sequencing provider explicitly before drawing aetiological conclusions from *de novo* signature exposures.

3.4.3. Provider-stratified extraction to recover biological signal

Our analysis revealed that the *de novo* exposure vectors clustered by sequencing provider (Figure 9). When examining the exposure plot, it was seen that each individual signature was active across all providers. However, differences could be observed in the median exposure *de novo* signature, per sequencing provider (Figure 10). We also observed that in the pooled (containing all the sequencing providers) dataset, the influence of the sequencing provider significantly exceeds the influence of the biological variables. When the dataset was stratified by sequencing provider, no strong biological signal was observed. To better resolve biological signal, we performed *de novo* signature extraction stratified by sequencing provider, reasoning that this would prevent provider-associated variance from dominating the decomposition and thereby allow the underlying biological processes to emerge.

There were four different sequencing providers: DECODE (n = 7475), SC (n = 5597), Vanguard (n = 1319) and VanguardPilot (n = 165). We removed samples with VanguardPilot as provider, as the sample size was too small to reliably extract *de novo* signatures. We used SigProfilerExtractor to extract the signatures. We used the same settings for these 3 datasets, as for the dataset that contained all of the samples combined.

Characterising the stratified extracted *de novo* signatures

Table 5 shows the extraction metrics for each provider-stratified dataset alongside those of the initial, unstratified extraction. The stratified datasets have a lower average and minimum stability than the combined dataset, which can be attributed to two related factors. First, stability is closely tied to sample size: the larger the dataset, the more stable the extracted solution, and the stratified subsets are by definition smaller than the combined dataset. Second, this effect is compounded by the nature of the signal itself. In the combined dataset, provider-associated patterns contribute strong, consistent variance across many samples, which the extraction algorithm captures reliably and reproducibly. Once stratification removes this source of variance, the remaining biological signals are weaker and more heterogeneous, and therefore harder to recover stably.

Applying UMAP to each stratified dataset revealed no well-defined distinct clusters. Colouring the samples by each metadata variable likewise showed no visually discernible structure.

Table 5 Metrics of the extraction process, separated by sequencing provider. The last row shows the original (unstratified) data, which includes samples from all sequencing providers as well as samples without a listed provider.

	Number of samples	Optimal number of signatures	Average stability	Minimum stability
DECODE	7,475	12	0.90	0.62
SC	5,597	15	0.87	0.43
Vanguard	1,319	13	0.84	0.50
Combined	17,419	15	0.96	0.75

Comparing signatures across datasets

In Figure 13, the signatures cluster by cosine similarity into several easily discernible groups that each contain signatures from all three sequencing providers, with low cosine distance (< 0.1). These are clusters 1, 3, 5, 9, 10, 11 and 12. From these clusters, we can hypothesize that these represent either biological processes or artefactual background signatures unrelated to the sequencing providers.

There are some clusters consisting of two signatures, namely clusters 2, 4, 6, 7 and 8. For these clusters, these could be artefactual signatures that are shared among two sequencing providers.

There are also signatures that are dissimilar to every other signatures. These could be a reflection of a strong artefactual signal specific to the sequencing provider. However, it could also be the case that it concerns a biological process, but that the signal was only strong enough in one of the sequencing providers.

Decomposing the clustered signatures into COSMIC signatures reveals a few patterns; only the clusters with the most notable aetiologies are highlighted here. Cluster 9 comprises three signatures all with cosine similarity >0.813 to SBS55 (an artefactual signature), showing that artefacts common to all providers persist: the sequencing providers differ, but the underlying (artefact-prone) methodology does not. In cluster 7, both signatures decompose well (cosine similarity 0.856 and 0.903) and contain SBS100 (tobacco smoking), with one also showing SBS29 (tobacco chewing). Clusters 3–8 likewise decompose with relatively high similarity (>0.80), whereas clusters 10 and 1 do not decompose into COSMIC signatures at all (cosine similarity 0.56 and 0.63). Whether these represent novel biology or artefacts cannot be determined here. Comparing the *de novo* signatures to the combined reference set (COSMIC plus blood-associated signatures, Section 2.1.4) yielded no notable differences. The only custom signature referenced was RefSig SBS169 [20], found in 34 lymphoid tumour samples; however, its decompositions had low cosine similarities to the corresponding *de novo* signatures (0.708, 0.712, 0.713), indicating they are unlikely to genuinely represent SBS169.

Comparing signatures to the combined dataset

Comparing the *de novo* signatures extracted from the stratified datasets to those from the original extraction confirms that the clusters defined earlier correspond well to the original *de novo* signatures, with most clusters mapping to a single original signature at high similarity (>0.9 in the majority of cases). The notable exceptions are two singletons that do not correspond to any original signature: SC::SBS96O (highest similarity 0.64) and Vanguard::SBS96L (highest similarity 0.80). These most likely

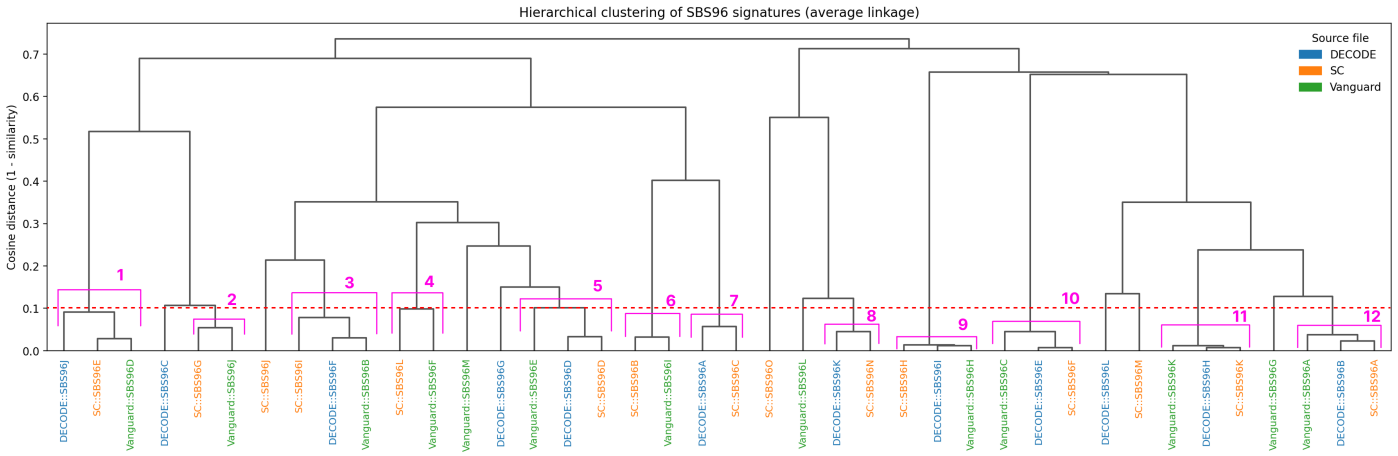


Figure 13 Hierarchical clustering of *de novo* signatures extracted separately per sequencing provider, based on cosine similarity (average linkage). Each leaf is one provider-specific signature, labelled provider::signature and coloured by source provider (DECODE, blue; SC, orange; Vanguard, green). Signatures merging below a cosine distance of 0.1 (red dashed line) form a cluster; the resulting clusters are numbered 1–12, and singleton signatures (clusters of one) are left unnumbered.

Table 6 Stratified *de novo* signatures matching COSMIC above the 0.85 threshold.

De novo signature	COSMIC match	Cosine
DECODE::SBS96F	SBS10d / SBS36 / SBS56	0.92 / 0.90 / 0.92
DECODE::SBS96K	SBS12 / SBS46	0.86 / 0.87
SC::SBS96B	SBS45	0.97
SC::SBS96J	SBS10d / SBS36	0.88 / 0.95
Vanguard::SBS96B	SBS10a/10d/36/56	0.86 / 0.89 / 0.86 / 0.91
Vanguard::SBS96I	SBS45	0.96

represent provider-specific signals for SC and Vanguard, respectively.

Cosine similarity of stratified *de novo* signatures to COSMIC signatures

Six stratified signatures exceeded the 0.85 cosine threshold against COSMIC (Table 6); notably, cluster co-membership was preserved in their matches (e.g. DECODE::SBS96F and Vanguard::SBS96B both map to the SBS10d/SBS36/SBS56 group), supporting that clustered signatures represent the same process across providers.

Stratified regression analysis

We performed regression analysis separately on the stratified extraction dataset, with the absolute exposure to the *de novo* signatures extracted from that dataset, as dependent variables. Per *de novo* signature, an independent test was performed. Multiple testing correction was applied. None of the models showed significant ($p < 0.05$) results after applying multiple testing correction, but still we can investigate the variables that are closest to significant, and compare them across different datasets.

Per-provider regression results are summarised in Table 7; no association reached significance in any stratum. The full heatmaps can be viewed in the supplementary as Figures S4, S5 and S6. In DECODE, the SBS96G decomposition is likely spurious: SBS96G is negatively associated with lung cancer (can be seen in the supplementary, Figure S7) yet decomposes partly into the tobacco-related SBS100. Both SBS96G and SBS96L are singleton clusters

Table 7 Per-provider regression results for the *de novo* signatures closest to significance. For each provider, the signature × variable association closest to significance is shown, with its FDR-corrected p-value and the decomposition into COSMIC signatures (cosine similarity in parentheses). Artefactual COSMIC signatures in the table: SBS43, SBS45, SBS51, SBS55, SBS56

Provider	Signature × variable	p-value	Decomposition (cosine similarity)
DECODE	SBS96G × BRCA	0.36	SBS100 (49%) + SBS51 (34%) + SBS56 (17%) (0.82)
DECODE	SBS96I × POLE	0.36	SBS55 (100%) (0.836)
DECODE	SBS96L × MMR	0.36	No good match (0.573)
SC	SBS96A × age	0.073	SBS28 (56%) + SBS55 (42%) + SBS5 (2%) (0.736)
SC	SBS96L × cohort	0.073	SBS89 (49%) + SBS10d (16%) + SBS43 (15%) + SBS5 (15%) + SBS1 (5%) (0.879)
Vanguard	SBS96I × APOBEC	0.12	SBS45 (83%) + SBS10d (17%) (0.978)

(suggesting that they represent a provider-specific process) and decompose poorly, marking them as probably artefactual.

In SC, SBS96A belongs to a three-member cluster, making a biological origin more plausible; its small SBS5 component (a clock-like, age-correlated signature) is consistent with the age association, though it accounts for only 2% of a weak (0.736) decomposition. SBS28's aetiology is unknown but linked to POLE germline variants [61]. SBS96L shows a positive association with leukaemia and a negative one with breast cancer (can be seen in the supplementary, Figure S8); none of its COSMIC components explain this directly, except possibly SBS1 and SBS5, which occur in AML [2]. SBS89 has no established aetiology but is found in normal colorectal epithelium, and SBS10d is associated with POLD1.

In Vanguard, POLD1 and POLE were dropped owing to too few positive samples (4 and 8). The low sample size (n = 1,319) limits power, but SBS96I belongs to a multi-member cluster and decomposes with high similarity (0.978). However, it decomposes into SBS45 (artefactual) and SBS10d (POLD1-linked), neither of which is APOBEC-related. The APOBEC association therefore lacks a clear mechanistic basis despite the clean decomposition.

Comparing effect size of cluster members

To test whether signatures in the same cluster share similar associations, we compared effect sizes (IRRs) rather than p-values, since statistical power differs across providers. For each provider we took its closest to significant associations and asked whether the other members of the same cluster behaved similarly (Table 8).

For DECODE, only one near-significant association fell within a cluster: SBS96I (cluster 9), associated with POLE. The SC member did not reproduce this effect, and POLE was dropped for Vanguard, which had too few POLE-positive samples for a reliable estimate.

SC contributed two near-significant associations: SBS96A × age (cluster 12) was mirrored by the DECODE member but not by Vanguard, while SC::SBS96L × cohort (cluster 4, with Vanguard::SBS96F) was driven by the breast-cancer diagnosis and not shared by its Vanguard partner.

Vanguard's estimates carry wide confidence intervals owing to its smaller sample size; even so, its SBS96I × APOBEC association (cluster 6) remained strong across the interval, whereas its cluster partner SC::SBS96B showed no comparable APOBEC effect.

Overall, cluster members did not consistently share associations. Effect sizes frequently diverged in magnitude or direction across providers, and no cluster produced a comparable signal in all three datasets, with the exception of SBS96A × age, where DECODE and SC agreed.

Table 8 Effect sizes for near-significant associations and their cluster partners.

Cluster	Signature	Covariate	IRR [95% CI]
9	DECODE::SBS96I	POLE	1.32 [1.06, 1.65]
9	SC::SBS96H	POLE	0.86 [0.65, 1.14]
9	Vanguard::SBS96H	POLE	removed (low n)
12	SC::SBS96A	age	1.04 [1.01, 1.06]
12	DECODE::SBS96B	age	1.02 [1.00, 1.03]
12	Vanguard::SBS96A	age	0.97 [0.94, 1.02]
4	SC::SBS96L	cohort (breast)	0.93 [0.89, 0.98]
4	Vanguard::SBS96F	cohort (breast)	0.98 [0.90, 1.07]
6	Vanguard::SBS96I	APOBEC	1.79 [1.21, 2.65]
6	SC::SBS96B	APOBEC	0.99 [0.88, 1.13]

3.5. Fitting COSMIC signatures: SBS1 tracks leukaemia

Finally, we examine the fourth view: exposures to the fixed COSMIC reference set. The de novo analysis (Section 3.4) sits two abstraction layers from the data and is dominated by the provider batch effect; fitting COSMIC signatures directly to each sample sidesteps the de novo decomposition and gives a more interpretable, if more constrained, signature-level readout to regress against. When fitting signatures to the samples, as described in section 2.4, 32 different

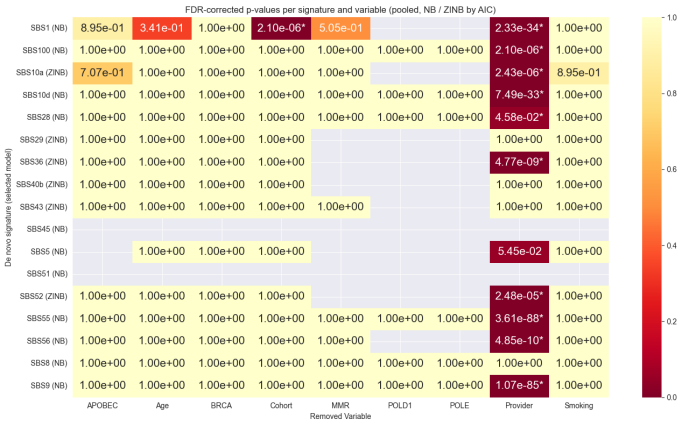


Figure 14 Likelihood ratio test results for the regression of COSMIC signature exposures (fitted, not extracted). Rows are COSMIC signatures, with the per-signature count model (NB or ZINB, selected by AIC) in parentheses; columns are the removed variables. Cell colour and printed value give the FDR-corrected (Benjamini-Hochberg) p-value; darker is smaller, and significant cells ($p < 0.05$) are marked with an asterisk. Greyed cells were not estimated, either because the signature × variable combination had too few samples or because the model did not converge (SBS45, SBS51).

COSMIC signatures are attributed to the samples. Out of these signatures, 17 signatures are present in more than 1% of the samples (> 145). The average cosine similarity of the reconstructed sample to the original sample is 0.832. On average, 3.4 signatures are fitted to one sample (min = 1, median = 3, max = 7). First, we investigated the Pearson correlation between the presence of signatures and our variables. Again, we saw the highest correlation for the sequencing providers, with the highest SBS51 × Vanguard. The highest non-technical variable correlation is leukaemia × SBS1.

We performed the same analysis as in the previous section: for each signature we fit both a zero-inflated negative binomial and a standard negative binomial regression, selected between them by AIC, used likelihood ratio tests to quantify the influence of each variable on the exposure to that signature, and applied multiple-testing correction. We only selected signatures that were present in at least 1% of the samples.

This procedure, however, raised several issues related to sample size. Unlike the *de novo* signatures, were the COSMIC signatures active in small subsets of the samples. This can make individual variable effects inestimable due to complete separation (e.g., no POLD1-positive sample showing SBS1 exposure). As specified by our model constraints in Section 2.5.2, any variable × signature combination not meeting the minimum threshold of 10 co-occurring events was removed from the analysis and excluded from the multiple-testing correction to ensure model stability.

Results of the LRT can be seen in Figure 14. Signature × variable combinations with not enough samples are greyed out. SBS45 and SBS51 did not converge for both NB and ZINB. The many $1.00e+00$ p-values are due to the multiple testing correction, which pushes many raw p-values into the maximum raw p-value, which was $1.00e+00$. The only non-technical variable surviving multiple testing correction is SBS1 × cohort. The strongest variable driving this association is the leukaemia cohort, with an IRR of 3.34 with a CI 95%[1.89, 5.90], meaning that holding all other values constant,

a leukaemia diagnosis is associated with 234% increase in SBS1 compared to the control cohort.

4. Conclusion

4.1. Summary of findings

This was an exploratory study of whether mutational signatures reflecting biological processes can be recovered from normal blood, an accessible tissue in which the per-sample somatic signal is weaker than in tumours. From 17,419 filtered samples we found that, despite the filtering strategy, residual germline variants and technical artefacts persisted, most visibly as enrichment of the C>A and T>G channels attributable to guanine oxidation during library preparation and to Illumina NovaSeq chemistry. *De novo* extraction yielded 15 signatures, each active in the large majority of samples and clustering primarily by sequencing provider, indicating that the extraction captured ubiquitous background and provider-specific artefactual patterns rather than localised biological processes.

This pattern was reflected in the regression analysis across all four target variables: Sequencing provider outweighed every biological variable. For total mutation burden, negative binomial regression identified BRCA and POLE variants as the only biological predictors reaching significance, both associated with increased burden. For the relative composition of mutation channels, beta regression again placed sequencing provider as the dominant predictor across nearly all 96 channels; the strongest biological signal was at *[C>T]G channels, driven by the leukaemia cohort and mirrored by MMR, consistent with the clock-like SBS1 and the MMR-associated SBS6/SBS15. For *de novo* signature exposures, sequencing provider was overwhelmingly significant for all 15 signatures, and only three biological associations survived correction (SBS96A × age, SBS96K × cohort, SBS96N × cohort). We then assessed these associations in two ways.

First, the regression for the combined-dataset signatures was refit separately within each provider stratum; from the three significant assocations, only SBS96K × cohort reproduced in a single provider. Second, signatures were re-extracted independently within each provider and the regression rerun on each provider's own signatures; after correction no signature–covariate association was significant in any stratum, the closest being SBS96A × age and SBS96L × cohort in SC ($p = 0.073$). Clustered signatures shared across providers did not share their covariate associations. Finally, fitting COSMIC signatures directly to each sample is a more interpretable readout that sidesteps the de novo decomposition, again placed sequencing provider as the strongest predictor. However, here the one biological association surviving correction was SBS1 × cohort, driven by a leukaemia diagnosis; this echoes the [C>T]G channel result and is consistent with the clock-like aetiology of SBS1.

4.2. Limitations of data access, variant calling and study design

This study has several limitations. The first concerns data access. The investigation relied on the UK Biobank resource throughout its duration; however, on 23 April 2026, access to the UK Biobank was suspended following a major security incident, and at the time of writing it has not been reinstated. Although the VCF files had already been filtered and *de novo* mutational signatures extracted before the suspension, downstream predictor variables for a subset of samples could no longer be retrieved. For these samples, only the presence of high-impact gene variants, mutation burden, mutation channels, and exposure to *de novo* signatures were available; demographic, clinical, and lifestyle covariates required for the regression models were not.

This had three consequences for the analysis. First, samples with incomplete predictor data had to be excluded from the regression, reducing the effective sample size from 17,419 to 14,489 (a reduction of 17%) and correspondingly lowering statistical power to detect small-to-moderate effects. Second, and more importantly, the exclusion was not missing-at-random: every incomplete sample carried at least one high-impact variant, meaning that the cases most informative for the outcome of interest were disproportionately removed. This introduced class imbalance in the regression cohort and is likely to have biased effect-size estimates toward the null. Third, the same exclusions also left certain signature × variable combinations with very low or zero sample sizes, forcing us to drop those combinations to avoid unreliable estimates.

The suspension also limited the scope of follow-up experiments. The initial analyses suggested that batch effects played a larger role in the results than expected. Natural next steps would have been to include additional batch variables as covariates in the regression, or to apply stricter artefact filters to a subset of the VCF files and compare the outcomes. Neither is possible without renewed access to the data. Another valuable check would be to distinguish incident from prevalent cancer diagnoses. To illustrate, we had a set of 1,287 participants with a leukaemia diagnosis. Due to not having access to the UK Biobank, it was not possible to investigate whether these were participants where leukaemia was diagnosed before or after the DNA was sequenced of this participant. This distinction matters for interpretation: a signature associated with a prevalent diagnosis may reflect an established disease process, whereas an association with an incident diagnosis would point to a detectable mutational imprint preceding clinical presentation, which is closer to the early-detection aim motivating this study.

A further limitation concerns the suitability of the available data. We relied on VCF files generated by GATK HaplotypeCaller because these were the files accessible through the UK Biobank. However, as noted earlier, HaplotypeCaller is designed for germline variant calling. As a result, the data was less sensitive to somatic mutations, and true somatic variants make up a smaller proportion of the total calls, with the remainder likely to include artefacts.

This, in combination with the fact that we study blood instead of tumours, exacerbates the problem. First, somatic variant calling in tumour studies typically relies on a paired tumour-normal design, in which tumour DNA is compared against matched non-tumour DNA and only the differences are retained as somatic. We did not have access to such a paired design, since our samples come from blood rather than tumour tissue, so somatic variants could not be separated from germline variants in this way. Second, HaplotypeCaller performs poorly on low-VAF variants: one study found that it failed to detect variants with a VAF below 10% [62]. This is particularly relevant here, because somatic mutations in blood typically occur at much lower VAFs than somatic mutations in tumours, where a dominant clone can drive the VAF much higher.

4.3. Recommendations for future work

The first recommendation is to improve artefact filtering by generating a Panel of Normals from the dataset itself, rather than

relying on a public Panel of Normals. This could be taken a step further by stratifying the data by sequencing provider and generating a separate Panel of Normals for each stratum, which would help control for provider-specific technical variation.

A second recommendation is to replace GATK HaplotypeCaller with Mutect2, which is purpose-built for somatic variant calling and should both reduce artefacts and improve sensitivity for true somatic mutations [63]. We chose HaplotypeCaller because its VCF files were directly available through the UK Biobank, whereas using Mutect2 would have required calling variants ourselves from the CRAM files: an additional processing step for which we did not have the time or compute resources at the full sample size. One alternative would have been to apply Mutect2 to a smaller subset of samples, trading sample size for calling accuracy. We chose to prioritise sample size, but a future study with similar resource constraints could reasonably make the opposite choice.

A third recommendation, given greater computational resources, would be to increase the sample size. The UK Biobank contains whole-genome sequencing data for roughly half a million participants, of which only a subset was used in this study. This is particularly important for our study, as we study blood, where the per-sample mutational signal is weaker than in tumour sequencing, so larger sample sizes are needed for reliable estimates. Analysing the full cohort would improve statistical power and may allow rarer *de novo* signatures to be extracted, or new signatures to emerge that were too infrequent to detect at the current sample size.

A further recommendation would be to expand the set of predictor variables, both to capture additional demographic, clinical, and biological information and to include further covariates that may act as confounders. Our regression analysis focused on identifying and quantifying feature importance, but the overall predictive accuracy of the models remained low. Adding more variables could improve this, and may help uncover additional relationships. This is particularly motivated by our observation that sequencing provider had the strongest association with the outcome in every experiment, contributing more to the model than any biological variable. Including further batch-effect variables, such as shipment batch number and sample plate ID, would allow this technical signal to be characterised more thoroughly and separated from biological effects. Other covariates that are commonly included, but that we could not add, are genetic principal components (PC) [64], which can be used to adjust for population structure or ancestry differences. The UK Biobank provides the first 40 genetic PC based on genotype data of about 805,000 markers [65]. Finally, relating mutational signatures of samples to polygenic risk scores (PRS) would be an interesting avenue to research. Polygenic risk scores are personalised scores indicating a person's genetic liability to disease by combining genetic risk information from across the genome. The UK Biobank already stores the PRS of each person [66].

To conclude, this exploratory study found that the somatic mutational landscape of blood-derived WGS was dominated by technical rather than biological variation. Somatic filtering did not cleanly isolate somatic mutations: residual germline and artefactual signal persisted. Sequencing provider outweighed every biological variable across all four views of the data, and the germline DNA-repair-deficiency hypothesis was not clearly borne out, with associations for BRCA, MMR, POLE and POLD1 either weak, not reproducing across providers, or not reliably estimable.

Yet even against this strongly confounded background, one biological signal survived across different views: a leukaemia-associated elevation of C>T mutations and the clock-like signature SBS1, recovered independently in the per-channel and COSMIC-fitting analyses. This result is notable for its timing as well as its consistency. Incident and prevalent status could not be retrieved for our 1,287 leukaemia samples after access was suspended; however, the entire UK Biobank contains only 511 prevalent leukaemia cases [22], so at least 776 (60%) of our leukaemia samples are necessarily incident, so sequenced before clinical diagnosis. The SBS1 association therefore cannot be attributed to established disease alone, and plausibly reflects, at least in part, a mutational imprint preceding diagnosis: precisely the signal most relevant to early detection. That such an association emerged at all, despite the technical noise, and within a predominantly pre-diagnosis cohort, is a promising result. Taking into account the recommendations provided earlier, blood may still prove a reliable substrate for population-scale mutational-signature analysis, and a candidate source of disease-associated biomarkers.

References

1. Freddie Bray et al. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3):229–263, 2024.
2. Iñigo Martincorena and Peter J. Campbell. Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489.
3. Ludmil B Alexandrov et al. Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312):618–622, November 2016.
4. Douglas E Brash. UV signature mutations. *Photochem. Photobiol.*, 91(1):15–26, January 2015.
5. Xueqing Zou et al. A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. *Nat. Cancer*, 2(6):643–657, June 2021.
6. Ludmil B Alexandrov et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.*, 47(12):1402–1407, December 2015.
7. Serena Nik-Zainal et al. Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell*, 149(5):979–993.
8. Paz Polak et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.*, 49(10):1476–1486, October 2017.
9. Jill E Kucab et al. A compendium of mutational signatures of environmental agents. *Cell*, 177(4):821–836.e16, May 2019.
10. Gunnar Boysen et al. Investigating the origins of the mutational signatures in cancer. *Nucleic Acids Research*, 53(1):gkae1303, 01 2025.
11. Iñigo Martincorena et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348(6237):880–886, May 2015.
12. Sipontina Faienza et al. Reconstructing the lifelong history of cells and tissues via somatic mutation analysis. *Cellular and Molecular Life Sciences*, 82(1):436.
13. Aik Seng Ng and Dedrick Kok Hong Chan. Commonalities and differences in the mutational signature and somatic driver mutation landscape across solid and hollow viscus organs. *Oncogene*, 42(37):2713–2724.
14. Emily Mitchell et al. Clonal dynamics of haematopoiesis across the human lifespan. *Nature*, 606(7913):343–350.
15. Fernando G. Osorio et al. Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. *Cell Reports*, 25(9):2308–2316.e4.
16. Felipe de Almeida Sartori et al. Mutational signatures in hematological malignancies. *Einstein (Sao Paulo)*, 24:eRW1961.
17. Miriam Elbracht et al. Germline variants in DNA repair genes, including BRCA1/2, may cause familial myeloproliferative neoplasms. *Blood Advances*, 5(17):3373–3376.
18. Erin M. Parry et al. Germline mutations in DNA repair genes in lung adenocarcinoma. *Journal of Thoracic Oncology*, 12(11):1673–1678.
19. Cathie Sudlow et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, 12(3):e1001779, March 2015.
20. Andrea Degasperi et al. Substitution mutational signatures in whole-genome-sequenced cancers in the UK population. *Science*, 376(6591), April 2022.
21. Ludmil B Alexandrov et al. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, February 2020.
22. Megan C. Conroy et al. UK Biobank: A globally important resource for cancer research. *Br J Cancer*, 128(4):519–527.
23. The UK Biobank Whole-Genome Sequencing Consortium et al. Whole-genome sequencing of 490,640 UK Biobank participants. *Nature*, 645(8081):692–701, September 2025.
24. Ryan Poplin et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, 2017.
25. Alexander J. Cole et al. Comprehensive analyses of somatic TP53 mutation in tumors with variable mutant allele frequency. *Scientific Data*, 4(1):170120, September 2017.
26. Michael S. Lawrence et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501, Jan 2014.
27. Xiaotu Ma et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature*, 555(7696):371–376, Mar 2018.
28. Jean-Baptiste Alberge et al. Genomic landscape of multiple myeloma and its precursor conditions. *Nat Genet*, 57(6):1493–1503.
29. Heather E. Machado et al. Diverse mutational landscapes in human lymphocytes. *Nature*, 608(7924):724–732.
30. Cédric van der Ham et al. Mutational mechanisms in multiply relapsed pediatric acute lymphoblastic leukemia. *Leukemia*, 38(11):2366–2375.
31. Joshua S. Weinstock et al. The genetic determinants of recurrent somatic mutations in 43,693 blood genomes. *Science Advances*, 9(17):eabm4945, April 2023.
32. Alessandro Laganà, editor. *Computational Methods for Precision Oncology*, volume 1361 of *Advances in Experimental Medicine and Biology*. Springer International Publishing.
33. Tim H. H. Coorens et al. Inherent mosaicism and extensive mutation of human placentas. *Nature*, 592(7852):80–85.
34. S.M. Ashiqul Islam et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genomics*, 2(11):100179.
35. Marcos Díaz-Gay et al. Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment. *Bioinformatics*, 39(12):btad756, December 2023.
36. Joseph M. Hilbe. *Negative Binomial Regression*. Cambridge University Press, 2 edition.
37. P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989.
38. M S Lawrence et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, July 2013.
39. Nikolai Klebanov et al. Burden of unique and low prevalence somatic mutations correlates with cancer survival. *Sci Rep*, 9(1):4848.
40. Jing Zhang et al. NIMBus: A negative binomial regression based Integrative Method for mutation Burden Analysis. *BMC Bioinformatics*, 21(1):474.
41. J. deLeeuw. Introduction to akaike (1973) information theory and an extension of the maximum likelihood

principle. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics: Foundations and Basic Theory*, pages 599–609. Springer New York.

42. Silvia Ferrari and Francisco Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815, 2004.
43. Ludmil B Alexandrov et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, August 2013.
44. Jacob C. Douma and James T. Weedon. Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. *Methods in Ecology and Evolution*, 10(9):1412–1430.
45. Michael Smithson and Jay Verkuilen. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11(1):54–71, March 2006.
46. Raydonal Ospina and Silvia L. P. Ferrari. A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, 56(6):1609–1623, June 2012. arXiv:1103.2372 [stat.ME].
47. Fraser Lewis et al. A unified approach to model selection using the likelihood ratio test. *Methods in Ecology and Evolution*, 2(2):155–162, 2011.
48. Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.
49. Peter Peduzzi et al. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379, December 1996.
50. Mizuki Ohno. Spontaneous de novo germline mutations in humans and mice: Rates, spectra, causes and consequences. *Genes Genet. Syst.*, 94(1):13–22.
51. UK10K Consortium et al. Timing, rates and spectra of human germline mutation. *Nat Genet*, 48(2):126–133.
52. Maura Costello et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Research*, 41(6):e67–e67.
53. Bjarni V. Halldorsson et al. The sequences of 150,119 genomes in the UK Biobank. *Nature*, 607(7920):732–740, July 2022.
54. Beverly J. Fu et al. A recurrent sequencing artifact on Illumina sequencers with two-color fluorescent dye chemistry and its impact on somatic variant detection. *bioRxiv*, page 2025.09.27.678978.
55. Marketa Tomkova et al. Human DNA polymerase is a source of C>T mutations at CpG dinucleotides. *Nature Genetics*, 56(11):2506–2516, November 2024.
56. Helen Davies et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat Med*, 23(4):517–525.
57. Luan Nguyen et al. Pan-cancer landscape of homologous recombination deficiency. *Nat Commun*, 11(1):5584.
58. Susanne N. Gröbner et al. The landscape of genomic alterations across childhood cancers. *Nature*, 555(7696):321–327, March 2018.
59. Laura Riva et al. The mutational signature profile of known and suspected human carcinogens in mice. *Nature Genetics*, 52(11):1189–1197, November 2020.
60. David Liu et al. Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma. *Nature Medicine*, 25(12):1916–1927, December 2019.
61. Pilar Mur et al. Recommendations for the classification of germline variants in the exonuclease domain of POLE and POLD1. *Genome Medicine*, 15(1):85.
62. Borahm Kim et al. Next-generation sequencing with comprehensive bioinformatics analysis facilitates somatic mosaic APC gene mutation detection in patients with familial adenomatous polyposis. *BMC Med Genomics*, 12(1):103.
63. David Benjamin et al. Calling Somatic SNVs and Indels with Mutect2. *bioRxiv*.
64. Caitlin E. Carey et al. Principled distillation of UK Biobank phenotype data reveals underlying structure in human variation. *Nat Hum Behav*, 8(8):1599–1615.
65. Clare Bycroft et al. Genome-wide genetic data on ˜500,000 UK Biobank participants. *bioRxiv*, page 166298.
66. Deborah J. Thompson et al. A systematic evaluation of the performance and properties of the UK Biobank Polygenic Risk Score (PRS) Release. *PLOS ONE*, 19(9):e0307270.
67. Christopher Glen Thompson et al. Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results. *Basic and Applied Social Psychology*, 39(2):81–90, 2017.
68. John Fox and Georges Monette. Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417):178–183, 1992.
69. Sana Naderi et al. Within-host genetic diversity of sars-cov-2 across animal species. *Virus Evolution*, 11(1):veae117, 01 2025.
70. Joseph M. Hilbe. *Overdispersion*, page 141–184. Cambridge University Press, 2011.

A. Supplementary Information

A.1. Verification of model assumptions

Before interpreting the fitted models, we verified the two assumptions on which they rely: absence of multicollinearity among predictors, and overdispersion of the count targets.

A.1.1. Multicollinearity

A regression model assumes its predictors are not linearly dependent [67]. We quantified this with the Variance Inflation Factor, which measures how much the variance of an estimated coefficient is inflated by collinearity. Because several predictors are categorical (and thus one-hot encoded), we used the generalised VIF (GVIF) [68]:

$$GVIF_j = \frac{\det(R_{11})\,\det(R_{22})}{\det(R)} \tag{4}$$

where R_{11} is the correlation submatrix of the predictors in group j, R_{22} the correlation submatrix of all other predictors, and R the full predictor correlation matrix. As the raw GVIF is not comparable across groups of differing size, we report the adjusted form $GVIF^{\frac{1}{2\,df}}$, where df is the degrees of freedom of the group [68]. Adjusted values above 1.75 are conventionally regarded as moderate and above 2.23 as high [69].

A.1.2. Overdispersion

The negative-binomial model is preferred over Poisson precisely when the data are overdispersed (variance exceeding the mean), as motivated in Section 2.5.2. We confirmed this empirically using the ratio of the Pearson χ^2 statistic to the degrees of freedom, where a value above 1.0 indicates overdispersion [70], and verified that it held for each count target before applying NB or ZINB.

Verification of regression model assumptions

Both model assumptions (collinearity and overdispersion, see Section A.1) held. Adjusted GVIF values for all predictors ranged between 1.001 and 1.054, far below the moderate threshold of 1.75 [69], indicating no multicollinearity. All count targets were strongly overdispersed: the Pearson χ^2/df ratio was 6776 for mutation burden, ranged from 39,306 to 504,619 for the *de novo* signature exposures and ranged from 235 (SBS1, the mildest) to 36,300 (SBS40b) for the COSMIC signature exposures, exceeding the threshold of 1.0 and justifying the negative-binomial family over Poisson throughout.

A.2. Correlation matrix

We also investigated potential correlations between variables, shown in Figure S1. Because we used one-hot encoding for the categorical variables, the correlations between the binary indicators of a single categorical variable (cohort, smoking status, or sequencing provider) are structural: they reflect the relative group sizes rather than any genuine association between categories.

Among the remaining correlations, we observe a positive correlation between a lung cancer diagnosis and current-smoker status (r = 0.28). There is also a negative correlation (r = -0.26) between patient age and the absence of a cancer diagnosis, which is expected, since the likelihood of a cancer diagnosis increases with age. A weaker positive correlation (r = 0.11) links the absence of a cancer diagnosis with never having smoked. Finally, there is a correlation between a breast cancer diagnosis and the presence of a high-impact BRCA variant (r = 0.27); this is by design, as our study intentionally selected samples with BRCA variants and breast cancer to investigate the relationship between them.

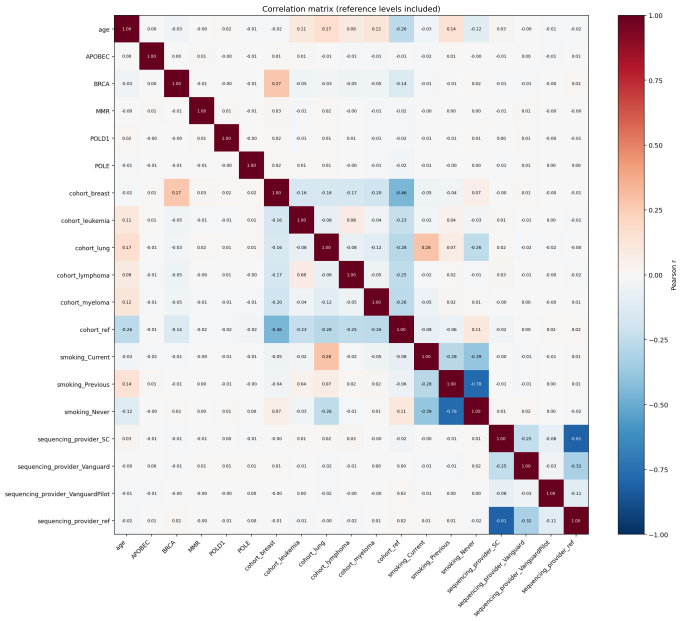
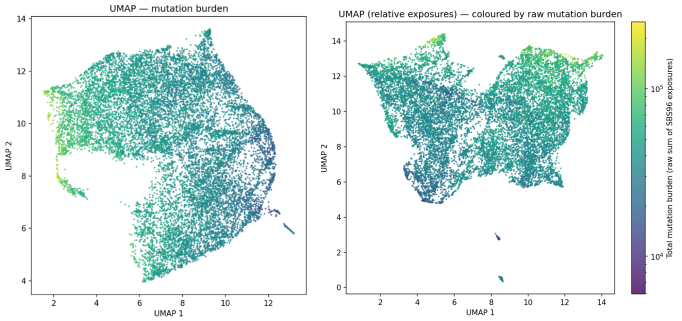


Figure S1 Correlation matrix showing the correlation between the different variables. Here, cohort_ref indicates the control cohort: the cohort of samples without cancer diagnosis.



(a) Absolute exposure (b) Relative exposure

Figure S2 2 components of the UMAP created by clustering the vector of exposures. Samples are coloured based on the mutation burden.

A.3. Mutation burden in UMAP of exposure vectors

When investigating whether part of the sequencing provider clustering could be caused by the average mutation burden differing per sequencing provider, the samples were also coloured by the total mutation burden (so the elements of the exposure vector aggregated). This can be seen in Figure S2a. As expected, the dominant axis of variation is total mutation count. When clustering on relative exposure level, we can still observe that the mutation burden has some influence on the clustering. This can be seen in Figure S2b, and can be quantified by correlating burden with each UMAP axis. On the raw embedding, burden was strongly tied to one axis (Spearman $\rho = -0.88$ on UMAP 1), confirming magnitude as the dominant axis of variation. After normalisation this weakens but persists, now spread across both axes ($\rho = 0.35$ on UMAP 1, $\rho = 0.59$ on UMAP 2). Normalisation removed most, but not all, of the burden structure, with the residual signal consistent with a genuine correlation between composition and total burden rather than a magnitude artifact.

B. Supplementary Figures

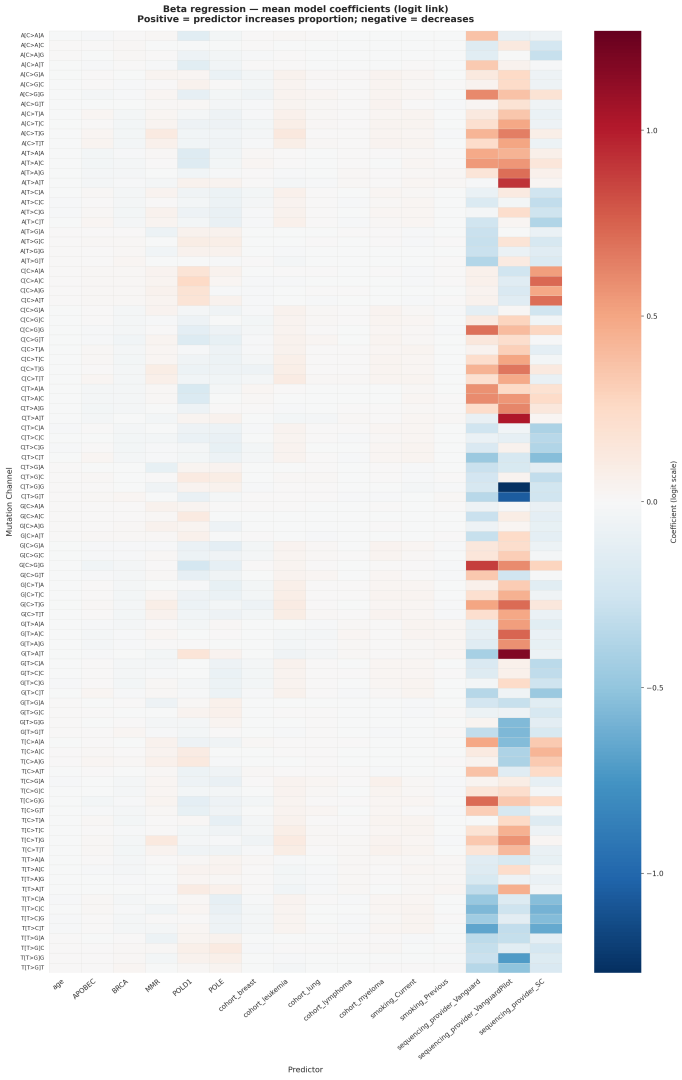


Figure S3 Heatmap showing the coefficients of each predictor for each mutation channel. The coefficients are the OR (the odds ratio) of each variable. This means that it scales the odds of each proportion, instead of scaling the expected count. For categorical variables, the OR should be interpreted relative to the reference category.

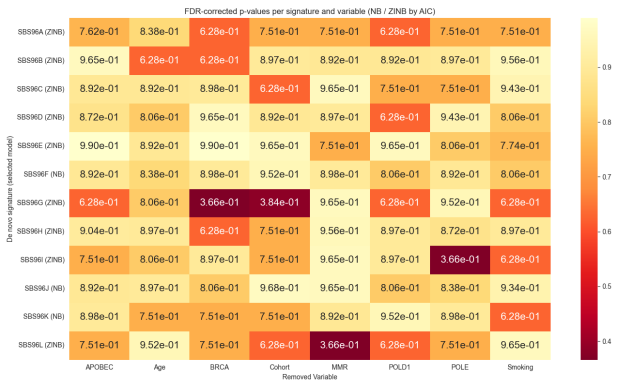


Figure S4 Regression analysis of *de novo* mutational signatures extracted from the DECODE dataset. The heatmap shows the results from the Likelihood Ratio Test. It shows the p-values of the removed variables and the *de novo* signatures

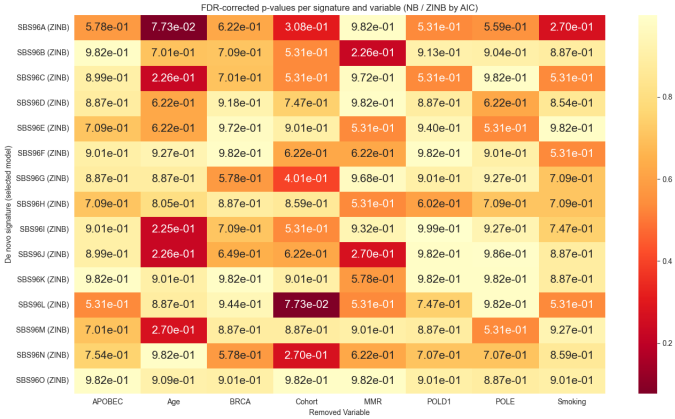


Figure S5 Regression analysis of *de novo* mutational signatures extracted from the SC dataset. The heatmap shows the results from the Likelihood Ratio Test. It shows the p-values of the removed variables and the *de novo* signatures

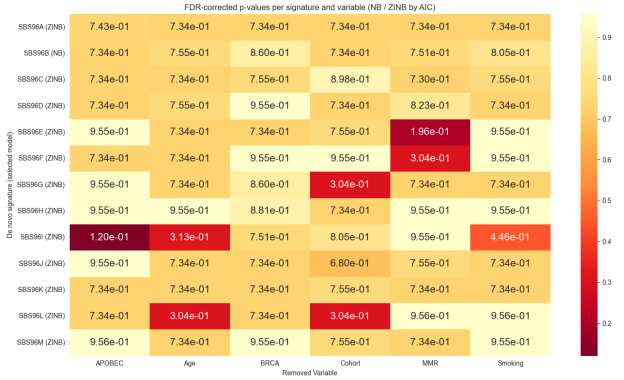


Figure S6 Regression analysis of *de novo* mutational signatures extracted from the Vanguard dataset. The heatmap shows the results from the Likelihood Ratio Test. It shows the p-values of the removed variables and the *de novo* signatures

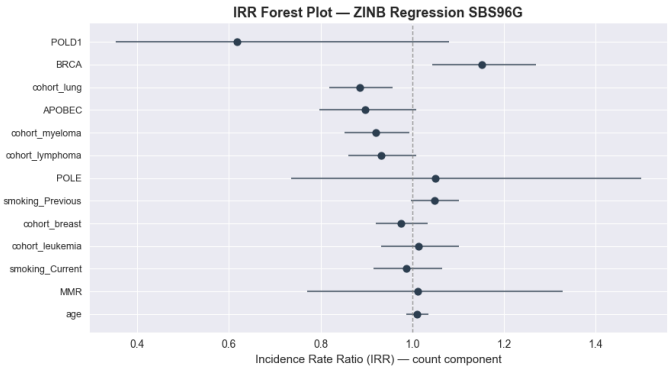


Figure S7 Count-component IRRs (95% CI) from the ZINB regression of SBS96G exposure, DECODE only. Dashed line: IRR = 1 (no effect); Categorical levels relative to reference.

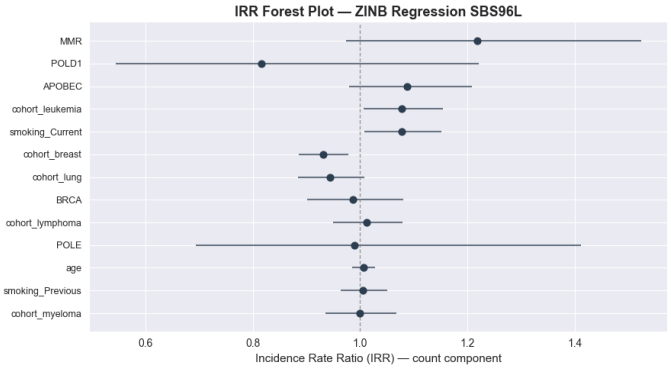


Figure S8 Count-component IRRs (95% CI) from the ZINB regression of SBS96L exposure, SC only. Dashed line: IRR = 1 (no effect); Categorical levels relative to reference.

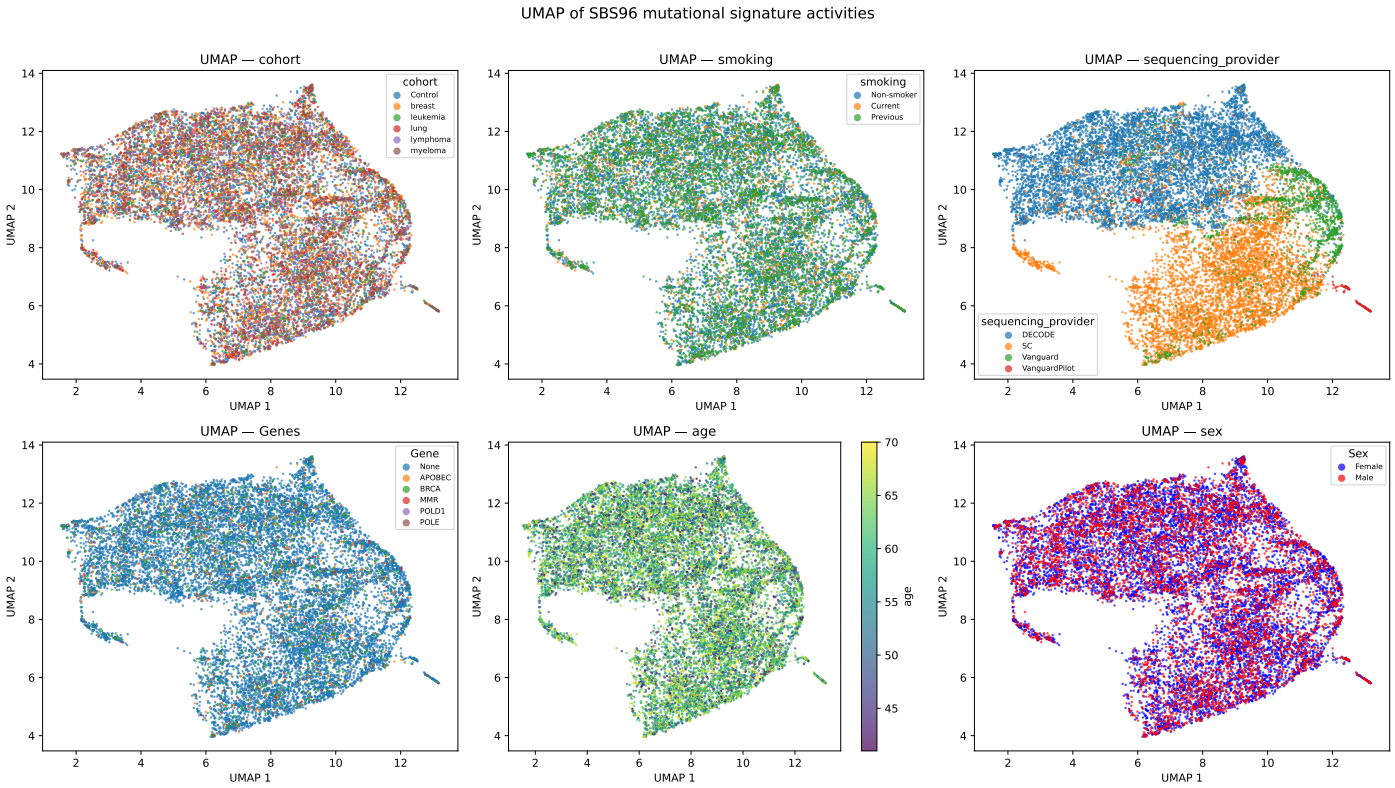


Figure S9 UMAPs created by analysing the exposure vectors of the *de novo signatures*. Samples are coloured based on the different variables.

Acknowledgements

Everyone keeps reminding me that a thesis is not a sprint, but a marathon. And, like any marathon, I didn't get to the finish line alone. I would like to thank everyone who has supported me along the way. First of all, I would like to thank thesis committee members Jing Sun and Michael Weinmann, for taking time out of their (indubitably) busy lives to read my thesis and attend my defence. I would like to thank my supervisors Joana and Sara. Discussing the material with you every week made the subject matter come alive for me, and also made me realise that science is not just about creating ideas, but also about discussing, dissecting (and criticising) those ideas together. I would like to thank Joana for keeping me mindful not to get lost in the small details, but to look at the big picture. I would like to thank Sara for her kind and supportive messages throughout the entire process.

I want to thank my parents for supporting me and cheering me on. I want to thank my mother especially, for being there for me, and I hope I was able to be there for her during our turbulent June. I want to thank my friends, for supporting me, but especially for pushing me away from my laptop and providing distraction and fun for me when I clearly needed it. Finally, I want to thank my boyfriend, David, for providing me with in-depth feedback, but also knowing when not to give feedback, and just to listen to my stress-filled rants.

Kirsten Timmerman
Delft, June 2026