De Novo Mutational Signature Discovery in Tumor Genomes using SparseSignatures

Daniele Ramazzotti[1,2,*], Avantika Lal[1,*,#], Keli Liu[3] , Robert Tibshirani[3,4], and Arend Sidow[1,5]

[1]Department of Pathology, Stanford University; [2]Department of Computer Science, Stanford University; [3]Department of Statistics, Stanford University; [4]Department of Biomedical Data Science, Stanford University; [5]Department of Genetics, Stanford University.

[*] The first two authors should be regarded as joint first authors.

[#] Present address: NVIDIA Corporation, 2788 San Tomas Expy, Santa Clara, CA 95051, USA.

Corresponding author: Arend Sidow, Department of Pathology, 300 Pasteur Drive, Stanford, CA 94305, USA. arend@stanford.edu

Running title: Mutational Signature Discovery by SparseSignatures

Keywords: Mutational Signatures, Cancer Genomics, Mutagenesis Mechanisms

Abstract

Cancer is the result of mutagenic processes that can be inferred from tumor genome sequences by analyzing rate spectra of point mutations, or "mutational signatures". Here we present SparseSignatures, a novel framework to extract signatures from somatic point mutation data. Our approach incorporates DNA replication error as a background, favors sparsity (signatures with few types of mutations) of non-background signatures, uses cross-validation to identify the number of signatures, and is scalable to very large datasets. We show that SparseSignatures outperforms current state-of-the-art methods on simulated data using standard metrics. We then apply SparseSignatures to whole genome sequences of 4476 tumors from 33 cancer types, obtaining 13 signatures in addition to the background. Signatures of known mutagens (e.g., UV light, benzo(a)pyrene, APOBEC dysregulation) occur in the expected tissues and a dominant signature with uncertain etiology is present in liver cancers. Other cancers exhibit a mixture of signatures or are dominated by background and CpG methylation signatures. Apart from cancers that are mostly due to environmental mutagens, there is little correlation between cancer types and signatures, highlighting the idea that any of several mutagenic pathways can be active in any solid tissue.

Introduction

Cancer is caused by somatic mutations in genes that control cellular growth and division (Vogelstein et al. 2013). The chance of developing cancer is massively elevated if mutagenic processes (e.g., defective DNA repair, environmental mutagens) increase the rate of somatic mutations. Due to the specificity of molecular lesions caused by such processes, and the specific repair mechanisms deployed by the cell to mitigate the damage, mutagenic processes generate characteristic point mutation rate spectra ('signatures') (Alexandrov et al. 2013). These signatures can indicate which mutagenic processes are

active in a tumor, reveal biological differences between cancer subtypes, and may be useful markers for therapeutic response (Wang et al. 2018).

Signatures are discovered by identifying common patterns across tumors based on counts of mutations and their sequence context. The original signature discovery method was based on Non-Negative Matrix Factorization (NMF) (Alexandrov et al. 2013). While other approaches have been considered (Gehring et al. 2015; Shiraishi et al. 2015; Fischer et al. 2013), NMF-based methods are by far the most widely used (Bolli et al. 2014; Schulze et al. 2015; Nik-Zainal et al. 2016; Covington et al. 2016; Goncearenco et al. 2017; Gori et al. 2018) and have resulted in a commonly used catalog of 30 signatures across human cancers (Alexandrov et al. 2015), available in the COSMIC database. A recent study (Alexandrov et al. 2018) using two NMF-based methods presented higher numbers (49 and 60) of putative signatures.

While some reported signatures have been associated with mutagenic processes (Alexandrov et al. 2016; Nik-Zainal et al. 2015; Helleday et al. 2014), careful examination reveals that several reported signatures are highly similar, suggesting overfitting rather than distinct mutagenic processes. In addition, there are several 'flat' signatures of uncertain origin (non-specific signatures that include mutations of all types and sequence contexts), and many signatures appear to be distorted by low levels of background noise. As an example, one may consider SBS40, whose etiology is unclear and which has many features in common with SBS5 (Alexandrov et al. 2018). Another example is represented by the four similar signatures attributed to defective DNA mismatch repair (signatures 6, 15, 20, and 26), which share common features and are not clearly separated (COSMIC). Such uncertainty complicates the task of understanding which signatures are active in different patients.

These observations are consistent with critical weaknesses in current signature discovery studies:

(1) State-of-the-art NMF-based methods aim to minimize the residual error after fitting the dataset with the discovered signatures. These methods do not aim to produce well-differentiated

signatures, nor do they minimize noise in the signatures. A method that favors sparsity of the signatures in addition to minimizing residual error would help alleviate these drawbacks. Technically, sparse and dissimilar signatures are advantageous for attributing mutations to individual signatures, inferring the identity of the mutagen, classifying new patients, or correlating signatures with other tumor properties. Biologically, enforcing sparsity has an important rationale on the basis of biochemical mechanism: most mutagens (Pfeifer et al. 2002; Smela et al. 2001) are highly specific in the type of damage they cause, and we therefore expect a majority of somatic mutational signatures to be sparse.

(2) No method incorporates the natural background of 'standard' replication error and repair processes, which occur in the normal course of cell division both in the germline and in somatic cells, including those of a tumor (Ledford 2017). Since we expect it to be present in all samples, and since most tumor cell lineages have undergone very large numbers of cell divisions, it should be considered a constant signature. If unaccounted for, it would likely find its way into other signatures, diminishing their accuracy.

(3) State-of-the-art NMF-based methods require the number of signatures as an input parameter but lack a principled basis for its selection. Discovering more signatures will always tend to reduce the residual error, i.e., fit the observed data better. However, the goal of signature discovery is not to fit the data as well as possible, but to identify signatures that truly reflect separate biological processes. Currently, standard ways to choose the number of signatures are: (1) choosing a number such that more signatures would not significantly reduce residual error (Gehring et al. 2015); (2) choosing a number based on both minimizing residual error and maximizing reproducibility of signatures (Alexandrov et al. 2013); (3) calling signatures hierarchically on subsets of samples, adding more signatures in order to fit every sample (Nik-Zainal et al. 2016). The first two practices are ambiguous and judged by human inspection, while the third aims to

select as many signatures as needed to improve fitting of the data, with little constraint to prevent overfitting. Overfitting can lead to many spurious signatures that actually represent the same process distorted by noise, making it difficult to reliably attribute mutations in a sample to any one signature, leading to misinterpretation of the results and possibly misleading conclusions. A recent method, SignatureAnalyzer, uses automatic relevance determination, which starts with a high number of signatures and attempts to eliminate signatures of low relevance (Tan et al. 2013).

To overcome these drawbacks, we developed SparseSignatures (Figure 1A), a novel framework for mutational signature discovery. Like other NMF-based methods, SparseSignatures both identifies the signatures in a dataset of point mutations and calculates their exposure values (the number of mutations originating from each signature) in each patient.

Results

The SparseSignatures Algorithm. SparseSignatures is implemented in R and is available as a Bioconductor package at https://bioconductor.org/packages/release/bioc/html/SparseSignatures.html. Noteworthy innovations are:

1. It uses the LASSO (Tibshirani 1996) to enhance sparsity and reduce noise in the signatures, except for a fixed background signature. The extent to which sparsity is favored is controlled by a tunable parameter, λ. The value of λ is learned to avoid forcing excessive sparsity.
2. It incorporates an explicit background model (Figure 1B), based on the human germline mutation spectrum (Rahbari et al. 2016), and validated in normal tissue samples (Supplementary Methods). Before using it, we made an empirical adjustment to CpG > TpG mutation rates (see Methods) of this background. This is because CpG > TpG mutations are frequently caused by CpG methylation, which can vary greatly in cancer cells, and are therefore not perfectly correlated with

replication rates in tumors. SparseSignatures fixes the background signature and then discovers additional signatures representing cancer-specific mutagenic processes (including, usually, CpG methylation). We note that our background signature is very similar to 'Signature 5' in COSMIC, which has been found in cancer samples from numerous tissue sites as well as in normal somatic tissue (Alexandrov et al. 2015; Martincorena et al. 2018). This pattern of mutations is a signal of DNA replication error that occurs in all cells, and its clear presence across tissues supports our choice to incorporate a constant background model.

3. It implements repeated bi-cross-validation (Owen et al. 2009) to select the best values for both λ and the number of signatures (K). A randomly chosen subset of data points is held out and signatures are discovered based on the rest of the data. The values of the held-out data points are predicted based on the discovered signatures and their fitted exposure values in each patient, and the mean squared error of the predictions is calculated. This procedure is performed for different values of K and λ, and the values that minimize the error in predicting held-out data points are chosen. The goal is to avoid overfitting, by ensuring that the discovered signatures not only fit the data used for discovery but also predict unseen values with high accuracy. In contrast to several other methods, this is a clear, unambiguous metric to choose the number of signatures.

SparseSignatures accurately detects signatures in simulated data. We compared SparseSignatures to two existing NMF-based methods for signature discovery, mutSignatures (Fantini et al. 2017) and SignatureAnalyzer (Tan et al. 2013). mutSignatures is an R implementation of the widely used WTSI framework (Alexandrov et al. 2013). These two methods were the basis for a recent pan-cancer study (Alexandrov et al. 2018) resulting in 49 and 60 putative signatures. We generated 50 simulated datasets of 500 patients each with 10 underlying mutational signatures (see Methods), and applied all three methods for signature discovery.

On simulated data, SparseSignatures is most effective at discovering the correct number of signatures (Figure 2A). SignatureAnalyzer consistently overfits the data, i.e., it overestimates the number of signatures and discovers many very sparse signatures that fit the data well but do not represent actual biological processes. For mutSignatures, two metrics, reconstruction error and stability are provided to estimate the number of signatures. (Stability is calculated as described (Alexandrov et al. 2013), using cosine similarity to compute the average silhouette over multiple runs). However, these are inconsistent: the minimum reconstruction error is always obtained at a number of signatures higher than 10, while the stability of the signatures is often highest at a number lower than 10.

When comparing the overall residual error obtained by the three methods, we observe that SignatureAnalyzer fits the input matrix best (Figure 2B). However, this is the result of overfitting as the method infers too many signatures. To provide a clearer measure, we assessed how well each method deciphers the input signatures by matching each of the input signatures to the most similar signature produced by the method, and assessing the residual error between these pairs of signatures (we did not include the background signature in this comparison). Compared to other methods, SparseSignatures shows lower error in reconstructing the input signatures (Figure 2C). We also compared the original exposure values for each input signature to the exposure values produced by the method for the closest deciphered signature, and found that SparseSignatures shows lower error in reconstructing the original exposure values (Figure 2D).

The learning and enforcement of sparsity in our approach is one reason for its higher accuracy. The sparsity of signatures deciphered by SparseSignatures closely matches that of the input signatures (Figure 2E), while other methods tend to discover signatures with the addition of considerable noise. We provide an example of a single simulation (Supplementary Figure 1) that illustrates the unambiguous, accurate, and sparse reconstruction of simulated signatures by SparseSignatures.

We performed two additional experiments in order to assess: (1) the effect of different levels of noise to the performance and (2) the capability of the methods to discover dense signatures like the ones described in COSMIC (https://cancer.sanger.ac.uk/cosmic/signatures_v2). These additional simulations showed similar results and are reported as Supplementary Notes (Experiments 2 and 3).

SparseSignatures discovers sparse, differentiated signatures in pan-cancer data. We applied SparseSignatures to a dataset of 33 cancer types from multiple ICGC studies (Supplementary Table 1), containing 59,940,498 point mutations in 4476 whole genomes. Our goal was to discover mutational signatures that are present across human cancers and can be reconstructed with high accuracy and confidence. We therefore limited our analysis to high-quality genomes.

SparseSignatures discovers 13 signatures in addition to the background (Figure 3, Table 1, Supplementary Tables 2 and 3), with diverse exposures for each cancer type (Figure 4, Supplementary Table 4, Supplementary Figures 2 and 3). The exposure values for the background signature have the highest correlation (Pearson rho = 0.28) to age of the patient at diagnosis, and mutation counts for blood cancers are dominated by the background signature. This provides empirical evidence to support our biologically motivated choice of modeling the background independently.

Remarkably, most of the signatures can be associated with a known mutational process (Table 1). For example, signature 5 is caused by deamination of methylated cytosine in CpG contexts. The exposure to this signature has a relatively low correlation (Pearson rho = 0.42) with exposure to the background signature, suggesting that it is additionally influenced by cancer-related changes in DNA methylation (Shen and Laird 2013). Signature 6 is associated with benzo[a]pyrene (tobacco smoking); the proposed mechanism of DNA damage by benzo[a]pyrene is transcription-coupled nucleotide excision repair of bulky DNA adducts on guanine (Alexandrov et al. 2016). Signature 2 is associated with UV light and likely caused by repair of cyclobutane dimers (Brash 2015); it is marked by high exposures in skin

melanomas, and to a lower extent in uveal melanomas. Signature 8 is a pattern of elevated T>C/A>G mutations largely in liver cancer. Although we do not know the cause, we note that its shape largely follows the genomic frequency of trinucleotides containing T in the center, implying that the mutagen modifies A or T to specifically cause T>C / A>G transitions independent of context.

We ran both mutSignatures and SignatureAnalyzer on the same dataset for comparison. mutSignatures is ambiguous (Supplementary Figures 4 and 5) as to the best number of signatures: 3, 10, and 15 are all candidates (Supplementary Figures 6, 7 and 8). In contrast to the other methods, SignatureAnalyzer selected 60 as the best number of signatures (Supplementary Figure 9). To illustrate the results of a simple NMF without incorporating background and sparsity, we also ran NMF on the same input data to obtain 14 signatures (Supplementary Figure 10). Finally, we compared our results to the 30 COSMIC signatures (Alexandrov et al. 2015).

Compared to the solutions produced by mutSignatures, NMF, and the 30 COSMIC signatures, our signatures fit the observed data better, are sparser, and show the lowest similarity between signatures, indicating that they are more clearly differentiated from each other (Supplementary Table 5). Moreover, our signatures show the lowest similarity between the background and the non-background signatures, suggesting that the other sets contain noise due to imperfect separation of the background signature. The mutSignatures solutions also do not contain the well-validated smoking signature caused by benzo[a]pyrene (Alexandrov et al. 2016; Nik-Zainal et al. 2015), which is accurately found by SparseSignatures. While the signatures produced by SignatureAnalyzer are extremely sparse, differentiated, and fit the data best, our simulation results lead us to believe that this method is overfitting by finding too many signatures (60), potentially splitting genuine signatures into multiple parts and/or introducing spurious signatures to fit noisy observations.

Three Signatures dominate the data. Encouraged by the quality and differentiation of the discovered signatures, we asked which signatures contribute most to the different cancer types. Overall, the background signature is responsible for the highest fraction of point mutations (29.1%) followed by signature 5 (methylation, 15.5%) and signature 6 (tobacco smoking, 11.0%). These three signatures alone represent more than 50% of mutations, a trend that is reflected in all of the most represented cancer types (i.e., breast, prostate and pancreatic), with the exception of liver cancer. Here, signature 5 has less impact than signatures specific for this cancer type: signature 8 (unknown aetiology, 13.8%) and signature 12 (aristolochic acid, 9.4%) that is predominant only in a subset of patients with recorded exposure to this mutagen.

Consistent with a previous report (Alexandrov et al. 2018), we find tobacco smoking (signature 6) to be among the predominant signatures in lung, liver and head and neck cancers. However, we also observe a subset of patients with prostate cancer and pediatric cancers to present high exposure to this signature (Supplementary Figure 11). Aggressive prostate cancers have been associated with smoking status (Foerster, Beat, et al., 2018); also, while no mechanistic conclusion can be drawn from this analysis, fetal exposure to smoking during pregnancy has been associated with increased risk of neoplasms (Boffetta, Paolo, et al., 2000).

Finally, we analyzed the exposure to signatures estimated by other approaches. As SignatureAnalyzer discovers too many signatures which are hard to interpret, we focused on the results by mutSignatures (configuration with 15 signatures; Supplementary Figure 8). In particular, we considered the 4 most represented cancer types, i.e., breast cancer, liver cancer, pancreatic cancer and prostate cancer. As observed before, mutSignatures does not infer the smoking signature in this dataset. Compared to SparseSignatures, mutSignatures is more uncertain on the association of exposures to cancer types (Supplementary Figure 12) and 5 signatures are required to cover at least 50% of the observed mutations in the 4 most represented cancer types (namely signatures 5, 9, 12, 13 and 14; Supplementary Figures 8

and 12). While a signature associated with methylation is still one of the predominant signatures, the remaining signatures are of unknown etiology and their overrepresentation in the estimated exposures is most likely an artefact due to the failure to properly estimate background and smoking signatures.

Mutational signatures define tumor clusters that span multiple tissue types. We then clustered patients to identify common patterns of mutagenic processes within and across cancer types. Our sparse and well-differentiated signatures provide much higher confidence in attributing mutations to signatures (exposure values) and in differentiating between individual samples (patients) on that basis. Using CIMLR (Wang et al. 2017; Ramazzotti et al. 2018) to perform clustering on the fitted exposure values for all patients, we separated our pan-cancer dataset into 24 well-separated clusters (Figure 5, Supplementary Table 6). Surprisingly, the clusters are only moderately associated (NMI=0.32) with the tissue of origin. Barring a few clusters linked to a single tissue and mechanism (such as, e.g., cluster 23, which is predominantly composed of breast cancers and dominated by signatures 9 and 11, i.e., APOBEC dysregulation), the majority of clusters show distinct patterns of signatures but span several cancer types. For example, almost all esophageal and many gastric cancers fall into two clusters: cluster 14, which is dominated by signature 13 (associated in (Dulak et al. 2013) to gastroesophageal reflux), and cluster 5, which shows high contributions from both background (DNA replication error) and signature 5 (cytosine methylation). In particular, cluster 5 is a mixed cluster, also including pancreatic and prostate tumors. Interestingly, skin melanomas also fall largely into two clusters: cluster 20, which presents high exposure for signature 2 (UV light) and cluster 12, which is more diverse, with high contributions from both the background and signature 1. These results highlight the idea that any of several mutagenic pathways can be active in any solid tissue.

The clustering results also confirm that most of the total observed point mutations are contributed by background and methylation signatures. In fact, we identify 8 clusters (clusters 1, 5, 9, 11, 16, 18, 21

and 24) where these two signatures alone explain more than 50% of mutations. We show in Supplementary Figure 13 the averaged observed counts for the patients of each of these clusters. Of these, of particular interest are cluster 21, enriched for clear cell renal cell cancer patients, and cluster 24 composed mostly of lymphomas and where signature 10 (unknown aetiology) is also observed. Interestingly, this signature is present in chronic lymphocytic leukemia and in a small subset of skin cancers (Supplementary Figures 14 and 15).

Discussion

SparseSignatures is a novel approach designed to discover the best number of clearly differentiated signatures with minimal background noise, which have robust statistical support by repeated cross-validation on unseen data points and are not likely due to overfitting.

Analyses of simulated data show that SparseSignatures outperforms current state-of-the-art methods. It provides the most accurate and least ambiguous estimation of the number of signatures, and reconstructs the original signatures most accurately. In comparison, other methods tend to discover too many signatures or retain noise in the discovered signatures.

Complementing its methodological innovations, SparseSignatures models a universal biological process in the form of a constant background signature. The biological motivation for incorporating a background signature is the fact that all cells are subject to replication error, which is the result of misincorporation of nucleotides and subsequent failure of the proofreading mechanisms of the DNA polymerases. Other processes such as transcription-coupled repair also contribute to the 'normal' mutation burden (Green et al. 2013). Although cell culture mutation spectra have been estimated (Milholland et al. 2017), we chose to base our background on the human germline signature, which is currently the most robust estimate of a non-cancer in vivo mutation spectrum. Our result that the

background signature is the most dominant signature overall provides empirical evidence that it is in fact prevalent in our data.

Aside from weaknesses in the methods themselves, previous signature discovery studies, aiming to obtain a comprehensive repertoire of signatures, have used large numbers of exomes, performed serial discovery of signatures on small subsets of samples, and carried out *post hoc* manual processing of the resulting signatures, leading to extensive and complex catalogs of signatures, such as the 49 and 60 signatures presented in an extensive pan-cancer analysis (Alexandrov et al. 2018). In this paper we aim to identify high-confidence, accurate mutational signatures with a single robust and statistically transparent method. For this reason, we considered only high quality whole genomes in order to avoid large numbers of weak or spurious signatures that detract from attention to the most important signals. Future work could be directed at incorporating indels and doublet base substitutions (Alexandrov et al. 2018), especially when larger datasets become available to support analyses of these comparatively rarer events.

We have also applied our method to a larger breast cancer dataset (Lal et al. 2019). On this dataset, we were able to discover additional signatures such as a signature associated with germline and somatic BRCA1 and BRCA2 mutations, some of which contributed only to a small subset of samples. We anticipate that the availability of larger datasets comprising curated, uniformly processed whole genome sequences may allow us to validate those signatures and discover new ones.

Using SparseSignatures on data from 33 cancer types, we obtain 13 signatures in addition to the background. Compared to other methods, we obtain a better fit to the observed data, with signatures that are sparse, differentiated, and have reduced noise, while at the same time preventing overfitting. Consistent with biological expectation, the contribution of DNA replication error (the 'background' signature) is the predominant cause of point mutations in 15 of the 24 analyzed cancer types (Figure 4) with at least 15 patients. In five, CpG methylation is one of the main causes, suggesting that it is a major contributor to, perhaps a driver of, the etiology of these tumors. Known mutagens (e.g., UV light or

smoking) contribute in expected ways (e.g., melanoma and lung cancers, respectively). Remarkably, none of the signatures are similar to one another, highlighting the potential significance of signatures 1 and 8, which do not have known etiologies, but which, due to their sparsity, suggest highly specific chemical or cell biological mechanisms.

Signature 8 seems particularly important to understand, as it is the largest non-background contributor to liver cancer, a usually aggressive disease. Similarly, clustering of the samples (Figure 5) suggests strongly that signature 8 is the main force behind a distinct liver etiology, as clusters 3, 7 and 22, where signature 8 provides high contribution, comprise most of the liver samples. Also of note is signature 13, which defines esophageal and a subset of stomach cancers, and which has been associated with acid reflux, but for which the actual mutagen is unknown. This sparse signature, which is enriched in for T>G / A>C mutations in the CTT / AAG context, suggests a specific mutagen, as opposed to a more general mechanism.

The small number of highly specific, differentiated signatures leads us to predict that whole genome sequencing of individual cancers and their classification on the basis of signatures, including the background, may become much more easily interpretable and possibly useful in a clinical context. For example, strong contribution of CpG methylation versus background in a patient suggests that methylation changes have been more important for the growth of the cancer and that overall cellular turnover (associated with background) may have been modest, suggesting that DNA replication inhibitors may be less effective than gene regulatory therapy for such patients. We suggest that future work be directed at greater numbers of patients for whole genome sequencing and the simultaneous collection of other omic data to connect mutagenesis with molecular phenotype and eventually mechanistic cause.

Methods

Mathematical Framework for Mutational Signature Discovery. The mathematical framework developed for signature extraction (Alexandrov et al. 2013) is as follows. First, all point mutations are classified into 6 groups (C>A, C>G, C>T, T>A, T>C, T>G; the original pyrimidine base is listed first). Then, these are subdivided into 16 x 6 = 96 categories based on the 16 possible combinations of 5' and 3' flanking bases. Each tumor sample is described by the count of mutations in each of the 96 categories. This forms a count matrix M, where the rows are the tumor samples and the columns are the 96 categories.

Signature extraction aims to decompose M into the multiplication of two low-rank matrices: the exposure matrix α and the signature matrix β.

$$M \approx \alpha\beta \quad \text{(Equation 1)}$$

Here, α is the exposure matrix with one row per tumor and K columns, and β is the signature matrix with K rows and 96 columns. K is the number of signatures. Each row of β represents a signature, and each row of α represents the exposure of a single tumor to all K signatures, i.e. the number of mutations contributed by each signature to that tumor. In NMF, this equation is solved for α and β by minimizing the squared residual error (some methods use Kullback–Leibler divergence instead) while constraining all elements of α and β to be non-negative.

$$min\ ||M - \alpha\beta||^2_F \ \ subject\ to\ \alpha \geq 0,\ \beta \geq 0$$

Improvements to the NMF framework in SparseSignatures. In SparseSignatures, we incorporate a background signature by modifying Equation (1) as follows:

$$M \approx \alpha_0\beta_0 + \alpha\beta \quad \text{(Equation 2)}$$

Here, β_0 is the known 'background' signature of point mutations caused by replication errors during cell division, and α_0 is the vector of exposures of all tumors to that signature. The dimensions of α_0 are (number of tumors x 1) and the dimensions of β_0 are 1 x 96.

To enforce sparsity in the discovered signatures, we use the LASSO (Tibshirani 1996). This is done by adding an additional regularization term to the cost function to be minimized:

$$min\ ||M - (\alpha_0\beta_0 + \alpha\beta)||_F^{\ 2} + \lambda||\beta||_1\ subject\ to\ \alpha \geq 0,\ \beta \geq 0,\ \alpha_0 \geq 0$$

The parameter λ controls the extent to which sparsity is encouraged in the signature matrix β. If the value of λ is set too low, it is ineffective, whereas if it is set too high, the signatures are forced to be too sparse and no longer accurately fit the data.

It should be noted that unlike the standard LASSO, the objective function we minimize here is non-convex. But it is bi-convex (convex in α with β fixed and vice-versa). Hence the alternating algorithm described below is natural and yields good solutions. A standard issue with all NMF algorithms is non-identifiability: if we scale β by c and α by $1/c$, the product $\alpha\beta$ remains unchanged. One can change the relative magnitudes of α and β at convergence by changing their relative magnitudes at initialization. To remove this ambiguity, we initialize β so that each row (signature) sums to 1. The choice of 1 is not important: if we had instead initialized β so that each row sums to c, the signatures we obtain at algorithm convergence would be equivalent (up to proportionality) to those obtained by initializing β with all rows summing to 1 and λ set to λ/c.

Implementation of SparseSignatures. SparseSignatures discovers mutational signatures by following the steps below.

Step 1: Build the Count Matrix M by counting the number of mutations of each of the 96 categories in each sample.

Step 2: Remove samples with less than a minimum number of mutations. In the analysis described in this paper, we have used a minimum number of 1000 mutations per tumor genome.

Step 3: Choose a range of values to test for K (number of signatures) and λ (level of sparsity).

Step 4: For each value of K in the chosen range, obtain a set of K initial signatures using repeated NMF (Brunet et al. 2004) to obtain a more robust estimation. This is an initial value for the matrix β. We use these NMF results as a starting point (although other starting points such as randomly generated signatures may also be chosen) and further refine the signatures. In practice, the final discovered signatures are often very different from those produced by the initial NMF.

Step 5: For each pair of parameter values (K and λ), perform cross-validation as follows:

5a. Randomly select a given percentage of cells from M. Based on simulations (Supplementary Methods, Supplementary Table 7), we currently use 1% of the points in the dataset for cross-validation; however, the method appears robust to large variations in this value.

5b. Replace the values in those cells with 0.

5c. Consider the NMF results for the chosen value of K as an initial value of β. Add the background signature (β_0). Then use an iterative approach to discover signatures with sparsity. Each iteration involves two steps:

5c(i). While keeping fixed the values of β_0 and β, fit α_0 and α by minimizing:

$$min\ ||M - (\alpha_0\beta_0 + \alpha\beta)||_F^{\ 2}\ subject\ to\ \alpha \geq 0,\ \alpha_0 \geq 0$$

5c(ii). While keeping fixed the values of β_0, α_0 and α, fit β by minimizing:

$$min\ ||M - (\alpha_0\beta_0 + \alpha\beta)||_F^{\ 2} + \lambda||\beta||_1\ subject\ to\ \beta \geq 0$$

These steps are repeated for a number of iterations (set to 20 by default; in all our experiments we found that this was sufficient to reach convergence).

5d. Use the obtained signatures to predict the values for the cells that were set to 0 (we do this by calculating the matrix $\alpha_0\beta_0 + \alpha\beta$ and taking the entries corresponding to the cross-validation cells).

Then replace the values in these cells with the predicted values and repeat step 5c. We repeat step 5c a number of times (set to 5 by default), each time discovering signatures and then replacing the values of the cross-validation cells by the predicted values. After each iteration, the predictions improve, as the algorithm converges, making the mean squared errors used in the next step more stable.

5e. At the last iteration of step 5d, measure the mean squared error (MSE) of the prediction.

5f. Repeat the entire cross-validation procedure (steps 5a-5d) a number of times (set to 10 by default) and calculate the MSE for all cross-validations. Since we randomly select a different set of cells for cross-validation each time, this allows us to obtain a robust measure of MSE.

Step 6: Choose the values of K and λ that correspond to the lowest MSE in most of the cross-validations.

Step 7: Using the selected values for K and λ, repeat sparse signature discovery (step 5c) on the complete matrix M (without replacing any cells with 0). This generates the final values of α_0, α and β.

Background signature. We used the germline mutation spectrum calculated by (Rahbari et al 2016) as our background signature. To validate this, we independently calculated the germline mutational spectrum using whole-genome sequencing data from normal tissue samples (see Supplementary Methods for details), and the spectrum thus obtained had a high cosine similarity of 0.98 with that calculated by (Rahbari et al 2016). We then adjusted the rates of ACG>ATG, CCG>CTG, GCG>GTG and TCG>TTG mutations to be equal to the rates of ACA>ATA, CCA>CTA, GCA>GTA and TCA>TTA mutations respectively, in order to separate the effects of DNA methylation from the background signature. We also compared our background signature with the COSMIC signature 5, which has been associated with aging, and found a high cosine similarity of 0.93.

Definition of the λ parameter. This parameter tunes the desired level of sparsity to be obtained by LASSO. For any analysis by LASSO, one can compute a maximal value of the LASSO penalty after

which all the coefficients of the regression get shrunk to zero (Friedman et al. 2010). As this maximal value can vary depending of the problem, our λ parameter represents the fraction of the actual maximal value to be used. Values closer to 1 result in sparser signatures.

Simulations. We generated 50 sets of simulated signatures, each one consisting of our background signature, a 'methylation' signature including four peaks for mutations caused by CpG methylation, and 8 randomly generated signatures. From each set of signatures, we generated a simulated dataset of 500 patients, resulting in 50 such simulated datasets. We ran three methods for *de novo* signature discovery (SparseSignatures, SignatureAnalyzer, and mutSignatures) on each of the 50 datasets and evaluated their performance. To evaluate the accuracy with which discovered signatures reconstructed the original signatures, we matched each input signature to its closest discovered signature and evaluated the match by mean squared error. We then also measured the mean squared error between the exposure values of the input signature and the discovered exposure values for its most similar discovered signature. Further details are given in the Supplementary Methods (Experiment 1).

Pan-cancer dataset. We obtained a dataset of point mutations from multiple ICGC studies (see Supplementary Table 1 for the full list of samples). We selected only whole-genome sequencing data and removed samples with less than 1000 point mutations. Finally, we removed samples with >50,000 mutations so that the signature extraction process is not biased toward these outliers. After this preprocessing, a total of 4476 samples from 33 different cancer types remained.

Software. The experiment carried out in this paper were performed using the SparseSignatures v1.0.1 R package and R version 3.4.3. The software is available for download on Bioconductor at https://bioconductor.org/packages/release/bioc/html/SparseSignatures.html. This package in its current

version makes use of external R packages NMF v0.21.0 (Gaujoux and Seoighe 2010), nnls v1.4 and nnlasso v0.3. Clustering of exposure values was carried out using CIMLR (Wang et al. 2017; Ramazzotti et al. 2018) MATLAB implementation (Wang et al. 2018). CIMLR is a recently developed approach based on multiple kernel learning and k-means clustering for dimension reduction and clustering, that has shown high performance on a variety of datasets.

Data Access

The whole-genome sequencing data used in this study is publicly available and was downloaded from https://dcc.icgc.org/search and https://www.synapse.org/#!Synapse:syn11726601/files/.

Acknowledgments

This work was supported by an R01 grant to A.S. (NIH/NCI) and gift funding from the BRCA Foundation. A.L. is supported by a Young Investigator Award from the BRCA Foundation. The results published here are based in part upon data generated by the TCGA Research Network (http://cancergenome.nih.gov/).

Disclosure declaration

The authors declare that there are no conflicts of interest.

References

1. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., & Stratton, M. R. (2013). Deciphering signatures of mutational processes operative in human cancer. Cell reports, 3(1), 246-259.

2. Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., & Stratton, M. R. (2015). Clock-like mutational processes in human somatic cells. Nature genetics, 47(12), 1402.

3. Alexandrov, L. B., Ju, Y. S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., ... & Campbell, P. J. (2016). Mutational signatures associated with tobacco smoking in human cancer. Science, 354(6312), 618-622.

4. Alexandrov, L., Kim, J., Haradhvala, N. J., Huang, M. N., Ng, A. W., Boot, A., ... & Klimczak, L. J. (2018). The Repertoire of Mutational Signatures in Human Cancer. bioRxiv, 322859.

5. Bolli, N., Avet-Loiseau, H., Wedge, D. C., Van Loo, P., Alexandrov, L. B., Martincorena, I., ... & Hinton, J. W. (2014). Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. Nature communications, 5, 2997.

6. Boffetta, P., Trédaniel, J., & Greco, A. (2000). Risk of childhood cancer and adult lung cancer after childhood exposure to passive smoke: A meta-analysis. Environmental Health Perspectives, 108(1), 73-82.

7. Brash, D. E. (2015). UV signature mutations. Photochemistry and photobiology, 91(1), 15-26.

8. Brunet, J. P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the national academy of sciences, 101(12), 4164-4169.

9. Covington, K., Shinbrot, E., & Wheeler, D. A. (2016). Mutation signatures reveal biological processes in human cancer. bioRxiv, 036541.

10. Dulak, A. M., Stojanov, P., Peng, S., Lawrence, M. S., Fox, C., Stewart, C., ... & McKenna, A. (2013). Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nature genetics, 45(5), 478.

11. Fantini, D., Huang, S., Asara, J. M., Bagchi, S., & Raychaudhuri, P. (2017). Chromatin association of XRCC5/6 in the absence of DNA damage depends on the XPE gene product DDB2. Molecular biology of the cell, 28(1), 192-200.

12. Fischer, A., Illingworth, C. J., Campbell, P. J., & Mustonen, V. (2013). EMu: probabilistic inference of mutational processes and their localization in the cancer genome. Genome biology, 14(4), R39.

13. Foerster, B., Pozo, C., Abufaraj, M., Mari, A., Kimura, S., D'Andrea, D., ... & Shariat, S. F. (2018). Association of smoking status with recurrence, metastasis, and mortality among patients with localized prostate cancer undergoing prostatectomy or radiotherapy: a systematic review and meta-analysis. JAMA oncology, 4(7), 953-961.

14. Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1), 1.

15. Gaujoux, R., & Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. BMC bioinformatics, 11(1), 367.

16. Gehring, J. S., Fischer, B., Lawrence, M., & Huber, W. (2015). SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics, 31(22), 3673-3675.

17. Goncearenco, A., Rager, S. L., Li, M., Sang, Q. X., Rogozin, I. B., & Panchenko, A. R. (2017). Exploring background mutational processes to decipher cancer genetic heterogeneity. Nucleic acids research, 45(W1), W514-W522.

18. Gori, K., & Baez-Ortega, A. (2018). sigfit: flexible Bayesian inference of mutational signatures. bioRxiv, 372896.

19. Green, P., Ewing, B., Miller, W., Thomas, P. J., Green, E. D., & NISC Comparative Sequencing Program. (2003). Transcription-associated mutational asymmetry in mammalian evolution. Nature genetics, 33(4), 514.

20. Helleday, T., Eshtad, S., & Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. Nature Reviews Genetics, 15(9), 585.

21. Lal, A., Ramazzotti, D., Weng, Z., Liu, K., Ford, J. M., & Sidow, A. (2019). Comprehensive genomic characterization of breast tumors with BRCA1 and BRCA2 mutations. BMC Medical Genomics, 12(1), 84.

23. Ledford, H. (2017). DNA typos to blame for most cancer mutations. Nature News.

23. Martincorena, I., Fowler, J. C., Wabik, A., Lawson, A. R., Abascal, F., Hall, M. W., ... & Fitzgerald, R. C. (2018). Somatic mutant clones colonize the human esophagus with age. Science, eaau3879.

24. Milholland, B., Dong, X., Zhang, L., Hao, X., Suh, Y., & Vijg, J. (2017). Differences between germline and somatic mutation rates in humans and mice. Nature communications, 8, 15183.

25. Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., ... & Van Loo, P. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature, 534(7605), 47.

26. Nik-Zainal, S., Kucab, J. E., Morganella, S., Glodzik, D., Alexandrov, L. B., Arlt, V. M., ... & Phillips, D. H. (2015). The genome as a record of environmental exposure. Mutagenesis, 30(6), 763-770.

27. Owen, A. B., & Perry, P. O. (2009). Bi-cross-validation of the SVD and the nonnegative matrix factorization. The annals of applied statistics, 3(2), 564-594.

28. Shen, H., & Laird, P. W. (2013). Interplay between the cancer genome and epigenome. Cell, 153(1), 38-55.

29. Pfeifer, G. P., Denissenko, M. F., Olivier, M., Tretyakova, N., Hecht, S. S., & Hainaut, P. (2002). Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. oncogene, 21(48), 7435.

30. Rahbari, R., Wuster, A., Lindsay, S. J., Hardwick, R. J., Alexandrov, L. B., Al Turki, S., ... & Stratton, M. R. (2016). Timing, rates and spectra of human germline mutation. Nature genetics, 48(2), 126.

31. Ramazzotti, D., Lal, A., Wang, B., Batzoglou, S., & Sidow, A. (2018). Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. Nature communications, 9(1), 4453.

32. Schulze, K., Imbeaud, S., Letouzé, E., Alexandrov, L. B., Calderaro, J., Rebouissou, S., ... & Calatayud, A. L. (2015). Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nature genetics, 47(5), 505.

33. Shiraishi, Y., Tremmel, G., Miyano, S., & Stephens, M. (2015). A simple model-based approach to inferring and visualizing cancer mutation signatures. PLoS genetics, 11(12), e1005657.

34. Smela, M. E., Currier, S. S., Bailey, E. A., & Essigmann, J. M. (2001). The chemistry and biology of aflatoxin B1: from mutational spectrometry to carcinogenesis. Carcinogenesis, 22(4), 535-545.

35. Tan, V. Y., & Fevotte, C. (2013). Automatic relevance determination in nonnegative matrix factorization with the/spl beta/-divergence. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(7), 1592-1605.

36. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

37. Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer genome landscapes. science, 339(6127), 1546-1558.

38. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., & Batzoglou, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nature methods, 14(4), 414.

39. Wang, B., Ramazzotti, D., De Sano, L., Zhu, J., Pierson, E., & Batzoglou, S. (2018). SIMLR: A Tool for Large-Scale Genomic Analyses by Multi-Kernel Learning. Proteomics, 18(2), 1700232.

40. Wang, S., Jia, M., He, Z., & Liu, X. S. (2018). APOBEC3B and APOBEC mutational signature as potential predictive markers for immunotherapy response in non-small cell lung cancer. Oncogene.

Tables

Signature	Proposed etiology	Basis for proposed etiology	Tumor type(s) in which signature is prominent
Background	DNA replication error	Experimental evidence (Rahbari et al. 2016)	Leukemias
Signature 1	Unknown	N/A	N/A
Signature 2	UV light	Experimental evidence (Brash 2015; Nik-Zainal et al. 2015)	Skin melanoma
Signature 3	POLE mutation	Hypothesised (Alexandrov et al. 2018)	N/A
Signature 4	Defective base excision repair	Hypothesised (Alexandrov et al. 2018)	N/A
Signature 5	Cytosine methylation / deamination	Experimental evidence (Helleday et al. 2014)	Glioblastoma Multiforme, Colorectal cancers
Signature 6	Benzo(a)pyrene (Tobacco smoking)	Experimental evidence (Nik-Zainal et al. 2015; Alexandrov et al. 2016)	Lung cancers
Signature 7	Defective DNA mismatch repair	Hypothesised (Alexandrov et al. 2018)	N/A

Signature 8	Unknown	N/A	Subset of liver cancers
Signature 9	APOBEC dysregulation	Experimental evidence (Helleday et al. 2014)	Bladder cancer, head and neck cancers, subset of breast cancers
Signature 10	Unknown	N/A	N/A
Signature 11	APOBEC dysregulation	Experimental evidence (Helleday et al. 2014)	Bladder cancer, head and neck cancers, subset of breast cancers
Signature 12	Aristolochic acid	Experimental evidence (Nik-Zainal et al. 2015)	Subset of liver cancers
Signature 13	Gastroesophageal reflux	Hypothesised (Dulak et al. 2013)	Esophageal cancer, Gastric cancer

Table 1. 10 signatures (including background) discovered by SparseSignatures and their proposed etiology.

Figure Legends

Figure 1. A) Schematic of the SparseSignatures method. N represents the number of tumors in the dataset, K the number of signatures. B) Background signature based on the human germline mutation spectrum. Vertical bars represent the probability of mutation in each of 96 categories. These are based on 6 possible mutation types (upper gray labels) and 16 possible combinations of 5' and 3' flanking bases (x-axis labels).

Figure 2. Comparison between SparseSignatures and other methods on simulated data. mutSignatures (1) is the solution obtained by selecting the number of signatures that minimize residual error, while (2) is the solution obtained by maximizing stability. A) Bar and line plot showing, for each method, the number of times it selected each value of K (number of signatures). The x-axis shows values of K and the y-axis shows the number of times each value was selected. Each method was run on 50 simulated datasets. In all cases, the correct value of K was 10. B) Box plots showing the residual error for the solutions produced by each method, over 50 simulations. Residual error was measured as the mean squared error (MSE) in reconstructing the original count matrix. C) Box plots showing the mean squared error in reconstructing the 9 non-background input signatures, over 50 simulations. D) Box plots showing the mean squared error in reconstructing the exposure values for the 9 non-background input signatures, over 50 simulations. E) Box plots showing the sparsity of the signatures produced by each method, over 50 simulations. The gray box shows the sparsity of the input signatures. Sparsity was measured as the fraction of cells in the signature matrix whose values is $<10^{-3}$. The dotted line on all box plots represents the median value for SparseSignatures.

Figure 3. The 14 mutational signatures obtained by applying SparseSignatures to a pan-cancer dataset of 4476 whole genomes. We report the number and correlation of the most similar (correlation higher than 0.70) corresponding signature from COSMIC (https://cancer.sanger.ac.uk/cosmic/signatures_v2).

Figure 4. Fitted values for exposure to each of the 14 signatures obtained by SparseSignatures for the 24 cancer types with at least 15 patients. Each panel shows boxplots representing the fraction of mutations per tumor (on the y-axis) contributed by the given signature (on the x-axis). Each panel represents a different cancer type.

Figure 5. A) Scatter plot of 24 clusters obtained by CIMLR (color-coding by cluster) on the fitted exposure values for the 24 cancer types with at least 15 patients in the pan-cancer dataset. B) Distribution of tumors from each cancer type in each of the 24 clusters. c) Box plots showing the exposure values for each signature in each of the 24 clusters, in terms of the fraction of mutations in each sample that are attributed to a given signature.

Supplementary Material Legends

Supplementary Figure 1. Example of result for a single simulation obtained by SparseSignatures. A) cross-validation results selecting 10 as the best number of signatures and $\lambda = 0.05$. B) Signatures deciphered by SparseSignatures. Black dots represent the input signatures for the simulation.

Supplementary Figure 2. Exposures to Background signatures and signatures 1 to 6 for each cancer type.

Supplementary Figure 3. Exposures to signatures 7 to 13 for each cancer type.

Supplementary Figure 4. Reconstruction error obtained by mutSignatures on pan-cancer data.

Supplementary Figure 5. Stability error obtained by mutSignatures on pan-cancer data.

Supplementary Figure 6. 3 signatures obtained by mutSignatures on pan-cancer data.

Supplementary Figure 7. 10 signatures obtained by mutSignatures on pan-cancer data.

Supplementary Figure 8. 15 signatures obtained by mutSignatures on pan-cancer data.

Supplementary Figure 9. 60 signatures obtained by SignatureAnalyzer on pan-cancer data.

Supplementary Figure 10. 14 signatures obtained by NMF on pan-cancer data.

Supplementary Figure 11. Average trinucleotide counts for the patients with high exposure to signature 6 (tobacco smoking) in lung cancer, pediatric brain cancer and prostate cancer. The Figure shows the distribution of observed trinucleotide counts for different cancer types averaged for the subset of patients where signature 6 is contributing at least 25% of the total point mutations.

Supplementary Figure 12. Exposures to signatures for the configuration with 15 signatures (shown in Supplementary Figure 8) by mutSignatures for the 4 most represented cancer types.

Supplementary Figure 13. Average trinucleotide counts for the patients in clusters 1, 5, 9, 11 16, 18, 21 and 24.

Supplementary Figure 14. Average trinucleotide counts for the patients with high exposure (above 25% of the observed counts) to signature 10 (unknown aetiology) in chronic lymphocytic leukemia, malignant lymphoma and a subset of skin cancers.

Supplementary Figure 15. Average trinucleotide counts for the skin cancer patients in clusters 12, 20 and 24.

Supplementary Table 1. List of samples within the pan-cancer dataset of 4476 whole genomes used for signatures discovery.

Supplementary Table 2. Results of cross-validation to choose the best values of K and λ on pan-cancer data, using 1% of the cells in the matrix for cross-validation. We tested values of K ranging from 2 to 16

and values of lambda of 0.01, 0.025, 0.05, 0.075 and 0.1. Cross-validation was repeated 500 times with 5 restarts each. The entries in the table represent the median mean square error (MSE) in fitting the unseen data points across the 500 repetitions.

Supplementary Table 3. 14 signatures (including the background signature) discovered by applying SparseSignatures to pan-cancer data.

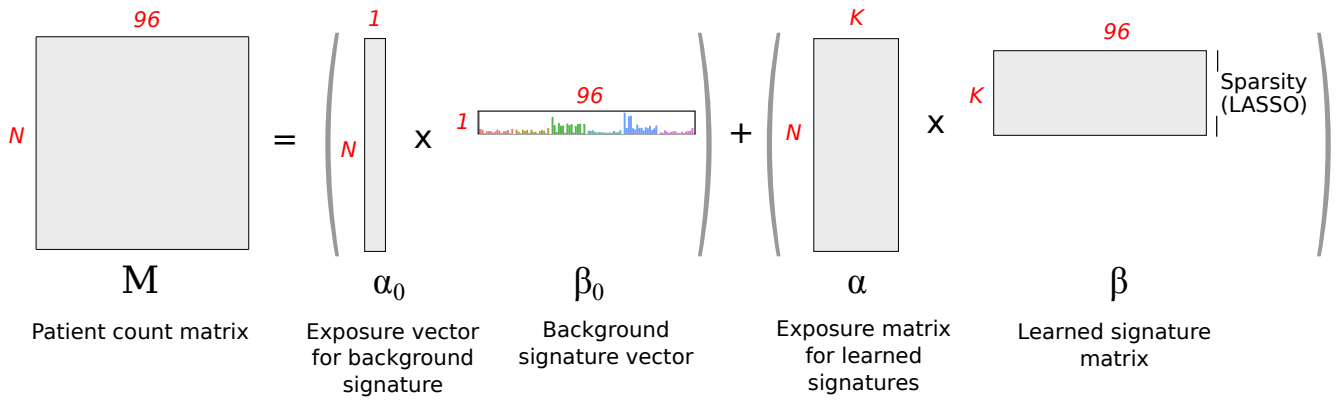
Supplementary Table 4. Fitted values for exposure to each of the 14 signatures (including the background signature) discovered by applying SparseSignatures to pan-cancer data, of each of the 4476 whole genomes in the pan-cancer dataset.

Supplementary Table 5. Comparison of the signatures discovered by SparseSignatures on pan-cancer data with the signatures from COSMIC (https://cancer.sanger.ac.uk/cosmic/signatures) and those obtained on the same data using SignatureAnalyzer, mutSignatures, and NMF.

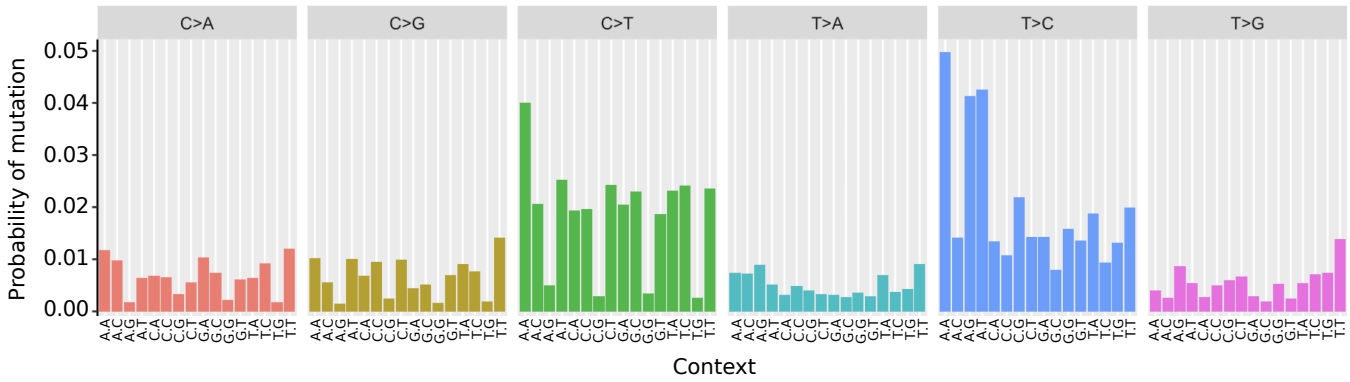
Supplementary Table 6. Cluster assignments generated by CIMLR for each sample.

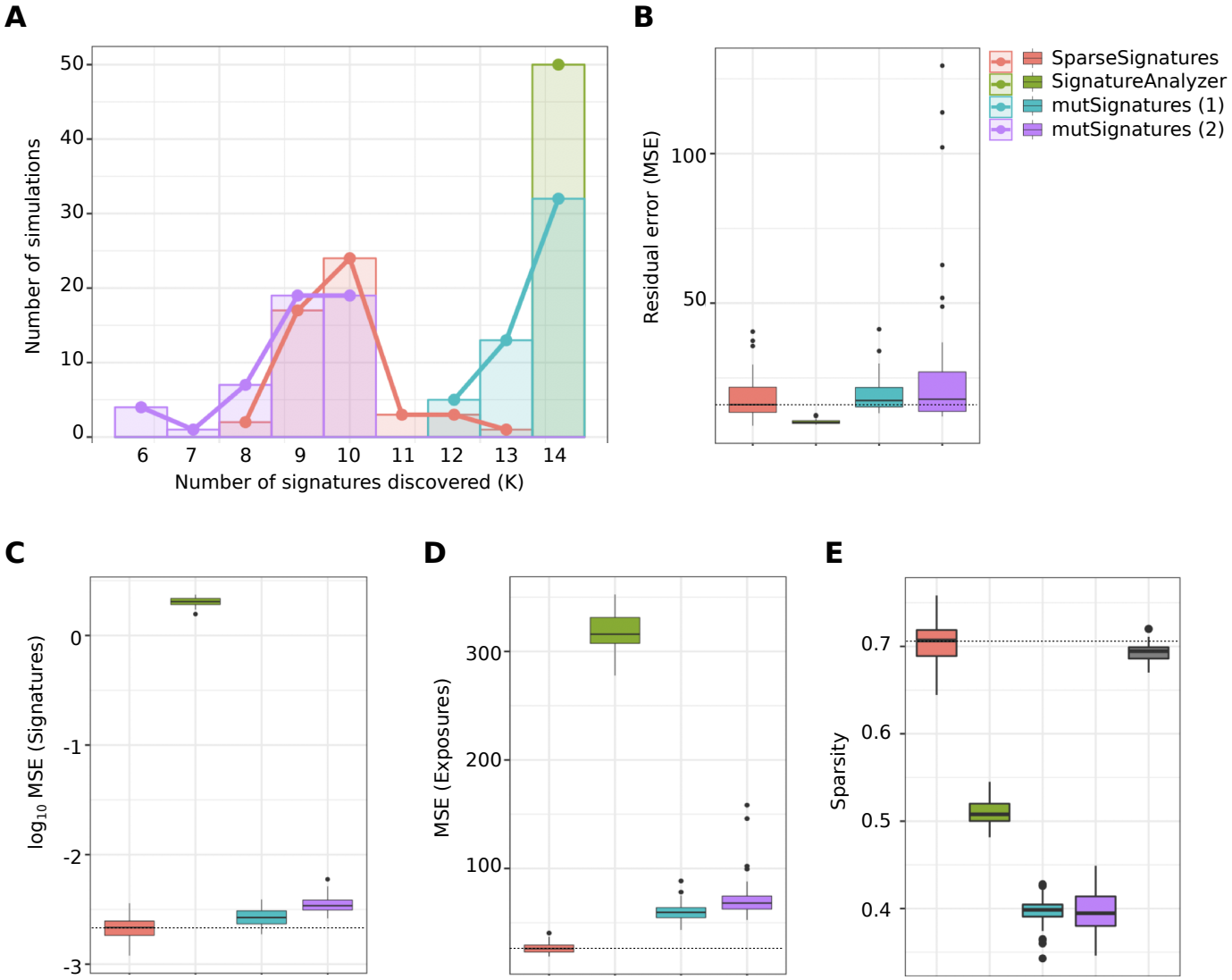
Supplementary Table 7. Results of cross-validation to choose the best values of K on simulated data, using 0.1%, 1%, and 10% of the cells in the matrix M for cross-validation. Cross-validation was repeated 100 times for each percentage of cells. The entries in the table represent the median mean square error (MSE) in fitting the unseen data points across the 100 repetitions.

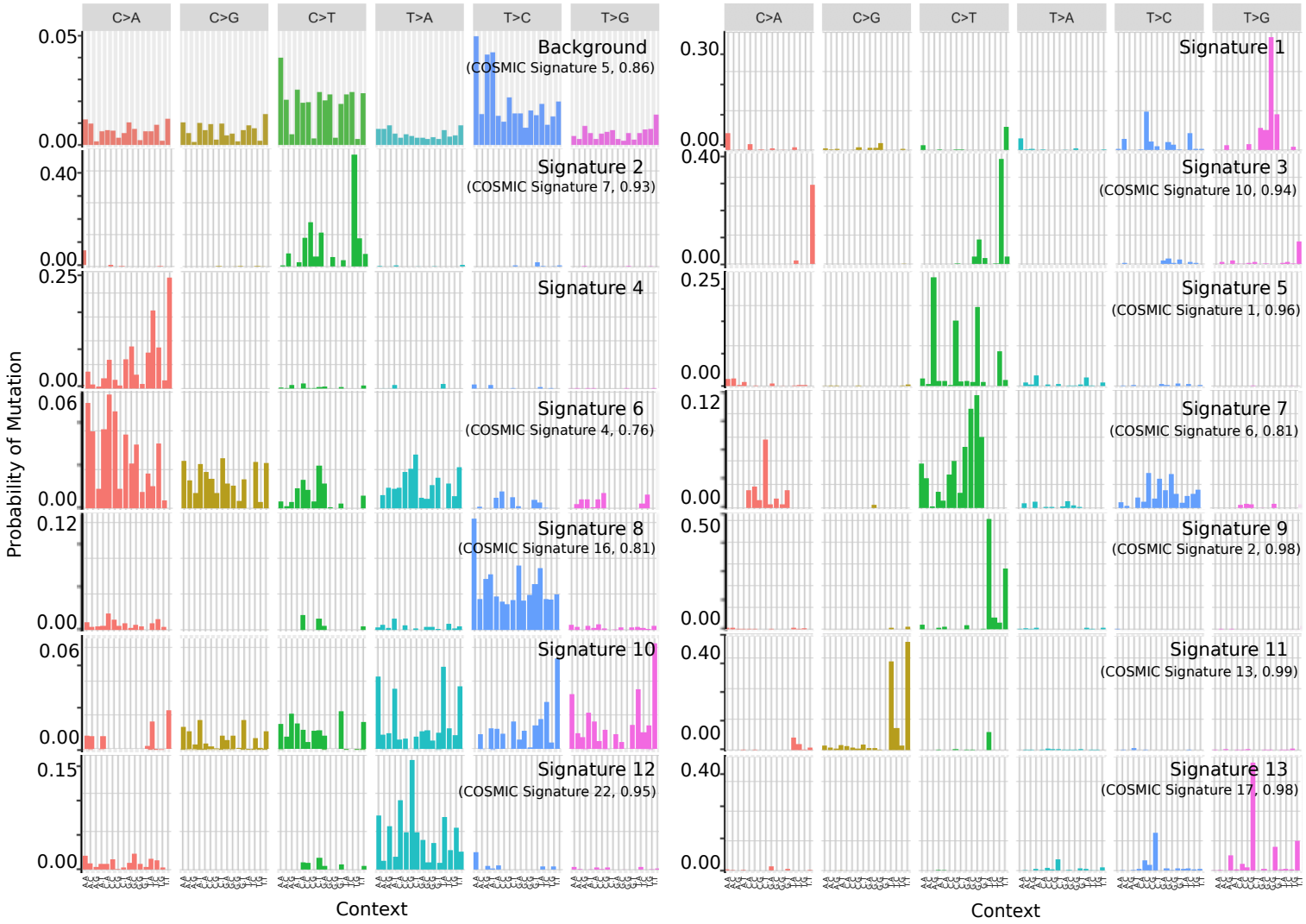
A


96
N
=
1
N
x
1
96
+
K
N
x
K
96
Sparsity
(LASSO)
M
Patient count matrix
α0
Exposure vector
for background
signature
β0
Background
signature vector
α
Exposure matrix
for learned
signatures
β
Learned signature
matrix

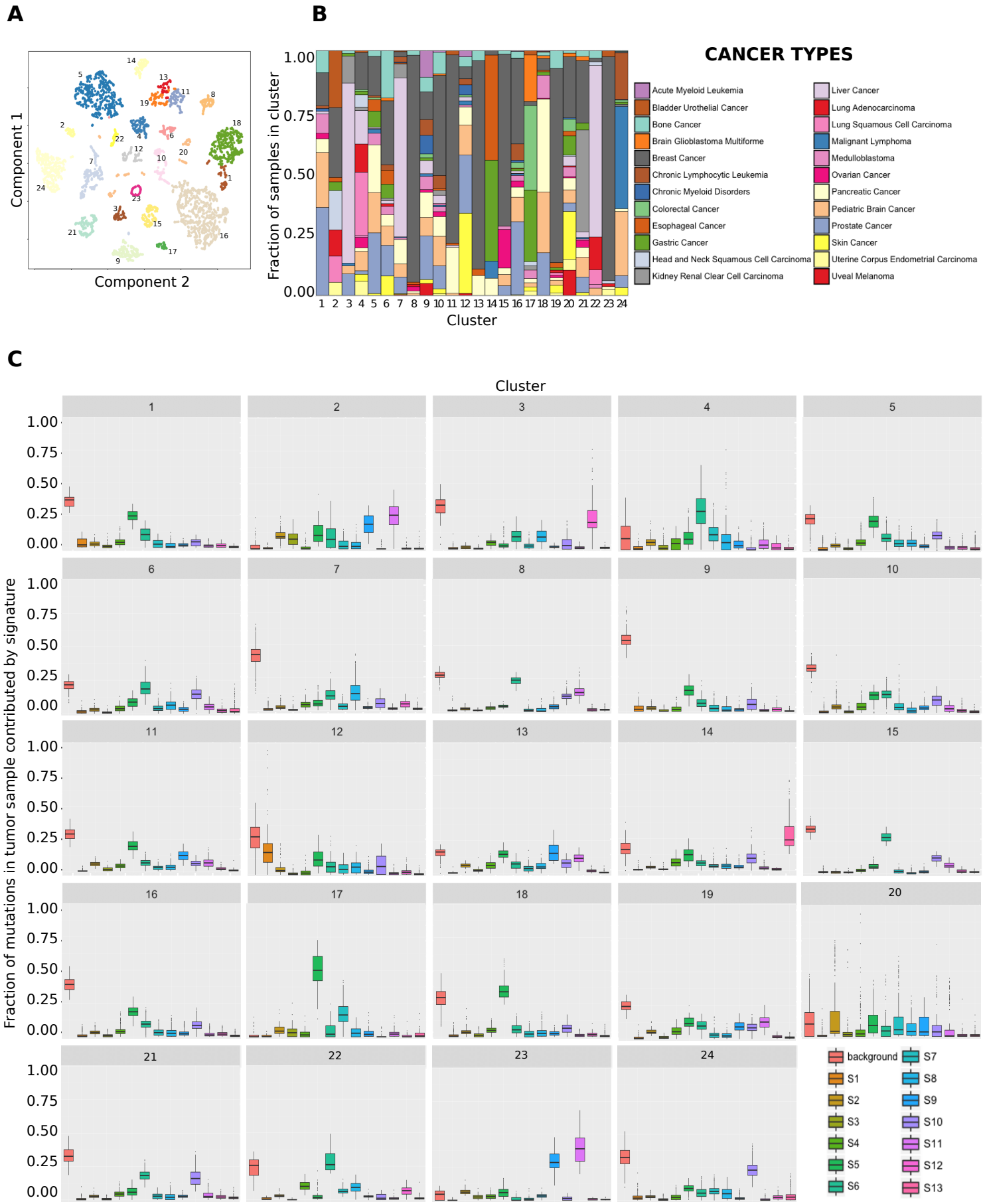
B


C>A
C>G
C>T
T>A
T>C
T>G
Probability of mutation
0.05
0.04
0.03
0.02
0.01
0.00
Context


A
Number of simulations
Number of signatures discovered (K)
B
Residual error (MSE)
SparseSignatures
SignatureAnalyzer
mutSignatures (1)
mutSignatures (2)
C
log10 MSE (Signatures)
D
MSE (Exposures)
E
Sparsity


C>A
C>G
C>T
T>A
T>C
T>G
Background
(COSMIC Signature 5, 0.86)
Signature 1
Signature 2
(COSMIC Signature 7, 0.93)
Signature 3
(COSMIC Signature 10, 0.94)
Signature 4
Signature 5
(COSMIC Signature 1, 0.96)
Signature 6
(COSMIC Signature 4, 0.76)
Signature 7
(COSMIC Signature 6, 0.81)
Signature 8
(COSMIC Signature 16, 0.81)
Signature 9
(COSMIC Signature 2, 0.98)
Signature 10
Signature 11
(COSMIC Signature 13, 0.99)
Signature 12
(COSMIC Signature 22, 0.95)
Signature 13
(COSMIC Signature 17, 0.98)
Probability of Mutation
Context
Context

Acute Myeloid Leukemia
Bladder Urothelial Cancer
Bone Cancer
Brain Glioblastoma Multiforme
Breast Cancer
Chronic Lymphocytic Leukemia
Chronic Myeloid Disorders
Colorectal Cancer
Esophageal Cancer
Gastric Cancer
Head and Neck Squamous Cell Carcinoma
Kidney Renal Clear Cell Carcinoma
Liver Cancer
Lung Adenocarcinoma
Lung Squamous Cell Carcinoma
Malignant Lymphoma
Medulloblastoma
Ovarian Cancer
Pancreatic Cancer
Pediatric Brain Cancers
Prostate Cancer
Skin Cancer
Uterine Corpus Endometrial Carcinoma
Uveal Melanoma
1
0
BG S1 S2 S3 S4 S5 S6 S7 S8 S9 S10 S11 S12 S13


A
Component 1
Component 2
B
Fraction of samples in cluster
Cluster
CANCER TYPES
Acute Myeloid Leukemia
Bladder Urothelial Cancer
Bone Cancer
Brain Glioblastoma Multiforme
Breast Cancer
Chronic Lymphocytic Leukemia
Chronic Myeloid Disorders
Colorectal Cancer
Esophageal Cancer
Gastric Cancer
Head and Neck Squamous Cell Carcinoma
Kidney Renal Clear Cell Carcinoma
Liver Cancer
Lung Adenocarcinoma
Lung Squamous Cell Carcinoma
Malignant Lymphoma
Medulloblastoma
Ovarian Cancer
Pancreatic Cancer
Pediatric Brain Cancer
Prostate Cancer
Skin Cancer
Uterine Corpus Endometrial Carcinoma
Uveal Melanoma
C
Cluster
Fraction of mutations in tumor sample contributed by signature
background
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
S11
S12
S13