bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

De Novo Mutational Signature Discovery in Tumor Genomes using SparseSignatures

Daniele Ramazzotti[1,2,*], Avantika Lal[1,*], Keli Liu[3], Robert Tibshirani[4,3], 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.

Cancer is the result of mutagenic processes that can be inferred from genome sequences by analysis of mutational signatures. Here we present SparseSignatures, a novel framework to extract mutational signatures from somatic point mutation data. Our approach incorporates DNA replication error as a background, enforces sparsity of non-background signatures, uses cross-validation to identify the number of signatures, and is scalable to very large datasets. We apply SparseSignatures to whole genome sequences of 2827 tumors from 20 cancer types and show by standard metrics that our set of signatures is substantially more robust than previously reported ones, having eliminated redundancy and overfitting. Known mutagens (e.g., UV light, benzo(a)pyrene, APOBEC dysregulation) exhibit single signatures and occur in the expected tissues, a dominant signature with uncertain etiology is present in liver cancers, and 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 virtually no correlation between cancer types and signatures, highlighting the idea that any of several mutagenic pathways can be active in any solid tissue.

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Cancer is caused by somatic mutations in genes that control cellular growth and division[1]. 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')[2]. 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[3].

Signatures are discovered by identifying common patterns across many tumors based on counts of mutations and their sequence context. The original signature discovery method was based on Non-Negative Matrix Factorization (NMF)[2]. While other approaches have been developed[4,5,6], NMF-based methods are by far the most widely used[7,8,9,10] and resulted in an initial catalog of 30 signatures across human cancers[11], available in the COSMIC database. Recently, a study[12] using two NMF-based methods (SigProfiler and SignatureAnalyzer) expanded the number of putative signatures to 49 and 60, respectively.

While some reported signatures have been associated with mutagenic processes[13,14], 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, and many signatures appear distorted by low levels of background noise. These observations are consistent with critical weaknesses that remain in the current signature discovery methods:

(1) Several signature discovery studies[5,12,15] are based on whole-exome data. This may introduce bias, as the frequency of trinucleotides differs in genomes and exomes (Supplementary Figure 1), and mutations in exons may be subject to selection. While these biases can be corrected, exomes also

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

contain very few mutations, which makes it difficult to discover reliable signatures and leads to stochastic noise. To illustrate this, we applied two signature discovery methods to a liver cancer dataset[16], using, first, whole-genome data, and second, only mutations in exons (Supplementary Figures 2 and 3). Using only exons, the number of signatures is lower, background noise is higher, and the signatures differ from those obtained using the whole genome. Since there are no criteria defining a sufficient number of mutations, even whole-genome sequences with few mutations may be insufficient for *de novo* signature discovery.

(2) NMF methods aim to minimize the residual error after fitting the dataset with the discovered signatures. This does not necessarily produce well-differentiated signatures, nor does it minimize noise in the signatures. A method that favors sparsity of the signatures in addition to minimizing residual error would help alleviate these drawbacks. In addition, enforcing sparsity has a biological rationale on the basis of biochemical mechanism: Most mutagens[17,18] are highly specific in the type of damage they cause, and we therefore expect a majority of somatic mutational signatures to be sparse.

(3) No method incorporates the natural background of 'standard' replication error, which occurs in the normal course of cell division both in the germline and in somatic cells, including those of a tumor[19]. 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, replication error will likely find its way into other signatures, diminishing their accuracy.

(4) 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 improve the fit, i.e. explain the observed data better. However, the goal of signature discovery is not to fit the data as well as possible, but instead to identify signatures that are truly likely to reflect separate

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

biological processes. Currently, ways to choose the number of signatures include: adding signatures until residual error is no longer significantly reduced (this is decided by human inspection and can be highly ambiguous) and evaluating reproducibility of the signatures[15], and calling signatures hierarchically on subsets of samples in order to fit every sample[9]. SignatureAnalyzer uses automatic relevance determination, starting with a high number of signatures and attempting to eliminate signatures of low relevance[20]. These methods aim to select as many signatures as needed to improve fitting of the data, with no constraint to prevent overfitting. Overfitting can lead to many similar signatures that actually represent the same process distorted by noise; such signatures are therefore limited in their usefulness. Moreover, with multiple similar signatures it is difficult to reliably attribute mutations in a sample to any one signature, leading to misinterpretation of the results and possibly misleading conclusions.

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

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

1. It incorporates an explicit background model (Figure 1b) based on the human germline mutation spectrum[21], with an empirical adjustment to CpG > TpG mutation rates. 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

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the background signature and then discovers additional signatures representing cancer-specific mutagenic processes (including, usually, CpG methylation).

2. It uses the LASSO[22] to enhance sparsity and reduce noise in the signatures, except for the 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.
3. It implements repeated bi-cross-validation[23] 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.

We applied SparseSignatures to a pan-cancer dataset[12], which after eliminating exomes and genomes with extreme numbers of mutations (see Methods), comprises 22,380,733 point mutations from 2827 whole genomes, belonging to 20 cancer types. SparseSignatures discovers 9 signatures in addition to the background (Figure 2, Table 1, Supplementary Tables 1 and 2), with diverse exposures for each cancer type (Figure 3, Supplementary Table 3). The exposure values for the background signature have the highest correlation (Pearson rho = 0.26) 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 replication error independently.

Remarkably, most of the signatures can be associated with a known mutational process (Table 1), and there is only one signature for each process. For example, signature 7 is caused by deamination of

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

methylated cytosine in CpG contexts. The exposure to this signature has a relatively low correlation (Pearson rho = 0.20) with exposure to the background signature, suggesting that it is additionally influenced by cancer-related changes in DNA methylation and likely reflects gene deregulatory mechanisms[24]. Signature 8 is associated with UV light[25] and marked by high exposures in skin melanomas, and to a lower extent in uveal melanomas. Signature 1 is a pattern of elevated T>C/A>G mutations largely in liver cancer; though 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 compared our 10 signatures to the 30 COSMIC signatures[11] and to the sets of 49 and 60 signatures previously proposed[12] (Supplementary Table 4). Our signatures are considerably sparser than the other sets, and also show the lowest similarity between signatures, indicating that they are more clearly differentiated from each other. Moreover, our signatures show the lowest similarity between background (replication error) signature and the non-background signatures, suggesting that the other sets contain noise in the signatures due to improper separation from the DNA replication errors. These results demonstrate the value of sparsity and of explicitly separating the background.

While our approach is not the first to emphasize sparsity, it is the first to combine sparsity with a fixed background and principled discovery of the number of signatures. Without a fixed background, increasing sparsity may prevent detection of the background replication error signature due to its dense nature. We ran other methods for sparse signature discovery[10,20] on our dataset; none detected any signature resembling the background. Instead, replication error seems to be distributed among several other signatures (Supplementary Figures 4, 5 and 6). This illustrates the importance of a model that is not only statistically sound but also grounded in the underlying biology of mutations. We also note that SignatureAnalyzer[20] selected 49 as the number of signatures in this dataset, suggesting overfitting once again.

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 SIMLR[26] to perform clustering on the fitted exposure values for all patients, we separated our pan-cancer dataset into 19 well-separated clusters (Figure 4, Supplementary Table 5). Surprisingly, the clusters are only moderately associated (NMI=0.39) with the tissue of origin; barring a few clusters linked to a single tissue and mechanism (such as cluster 9, which is composed of skin melanomas dominated by signature 8, i.e., UV light), 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 8, which is dominated by signature 9 (tentatively linked to gastroesophageal reflux[27]), and cluster 9, which shows high contributions from both signature 7 (cytosine methylation) and signature 9. However, many gastric cancer cases also fall into cluster 16, which is a mixed cluster, including pancreatic and prostate tumors, that is dominated by the methylation signature. Interestingly, skin melanomas also fall largely into two clusters: cluster 9, which is dominated by signature 8 (UV light) and cluster 10, which is more diverse, with high contributions from both the background and signature 2.

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 to be the result of overfitting.

Using SparseSignatures on data from 20 cancer types, we obtain 9 signatures in addition to the background. The dramatic difference in number compared to previous methods and studies[2,12] reflects the perennial issue of the balance between sensitivity and specificity. It is possible that our method does not find some signatures that make very small contributions to the dataset. However, while overfitting may

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

capture weakly represented signatures, it can and does lead to a proliferation of misleading results that detract from attention to the most important signals. SparseSignatures selects the signatures that perform best at fitting unseen data points, allowing us to focus on high-confidence signatures. This also allows us to avoid *post hoc* processing of the discovered signatures which introduces ambiguity and bias. We suggest, on the basis of our methodological innovations that prevent overfitting and utilize best practices in inference, that there may be less complexity in the repertoire of human cancer mutational signatures than previously thought.

Consistent with biological expectation, the contribution of DNA replication error (the 'background' signature) is the predominant cause of point mutations in 13 of the 20 analyzed cancer types (Figure 2). In five, CpG methylation is the predominant cause, suggesting that gene deregulation 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 adenocarcinoma, respectively). Remarkably, none of the signatures are similar to one another, highlighting the potential significance of signatures 1 and 2, which do not have known etiologies, but which, due to their sparsity, suggest highly specific chemical or cell biological mechanisms. Signature 1 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 4) suggests strongly that signature 1 is the main force behind a distinct liver etiology, as clusters 3-5, which are dominated by signature 1, contain most of the liver samples.

Also of note is signature 9, 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 a very specific manner for T>G / A>C mutations in the CTT / AAG context, suggests a specific mutagen, as opposed to a more general mechanism. We suggest that this lead will spark interest in both epidemiological (for associations) and biochemical (for mechanism) communities to understand the cause.

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Finally, the small number of highly specific 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 global gene deregulation (associated with methylation) has 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[2] 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

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

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

Modifications of 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[22]. 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.

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

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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[28] 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 6), 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).

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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[21]. We validated this independently using whole-genome sequencing data from normal tissue samples (see Supplementary Methods for details). 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.

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[29]. As this maximal value can vary

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

Pan-cancer dataset. We obtained a dataset of point mutations from Alexandrov et al.[12] that includes samples from PCAWG, ICGC and TCGA. We selected only whole-genome sequencing data and removed samples with less than 1000 point mutations. We also removed cancer types with less than 10 samples. 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 2827 samples from 20 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[30], nnls v1.4 and nnlasso v0.3. Clustering of exposure values was carried out using SIMLR[26,31] MATLAB implementation[32]. SIMLR 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.

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/).

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. Vogelstein, Bert, et al. "Cancer genome landscapes." science 339.6127 (2013): 1546-1558.
2. Alexandrov, Ludmil B., et al. "Deciphering signatures of mutational processes operative in human cancer." Cell reports 3.1 (2013): 246-259.
3. Wang, Shixiang, et al. "APOBEC3B and APOBEC mutational signature as potential predictive markers for immunotherapy response in non-small cell lung cancer." Oncogene (2018).
4. Gehring, Julian S., et al. "SomaticSignatures: inferring mutational signatures from single-nucleotide variants." Bioinformatics 31.22 (2015): 3673-3675.
5. Shiraishi, Yuichi, et al. "A simple model-based approach to inferring and visualizing cancer mutation signatures." PLoS genetics 11.12 (2015): e1005657.
6. Fischer, Andrej, et al. "EMu: probabilistic inference of mutational processes and their localization in the cancer genome." Genome biology 14.4 (2013): R39.
7. Bolli, Niccolo, et al. "Heterogeneity of genomic evolution and mutational profiles in multiple myeloma." Nature communications 5 (2014): 2997.
8. Schulze, Kornelius, et al. "Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets." Nature genetics 47.5 (2015): 505.
9. Nik-Zainal, Serena, et al. "Landscape of somatic mutations in 560 breast cancer whole-genome sequences." Nature 534.7605 (2016): 47.
10. Covington, Kyle, Eve Shinbrot, and David A. Wheeler. "Mutation signatures reveal biological processes in human cancer." bioRxiv (2016): 036541.
11. Alexandrov, Ludmil B., et al. "Clock-like mutational processes in human somatic cells." Nature genetics 47.12 (2015): 1402.
12. Alexandrov, Ludmil, et al. "The Repertoire of Mutational Signatures in Human Cancer." bioRxiv (2018): 322859.

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

13. Alexandrov, Ludmil B., et al. "Mutational signatures associated with tobacco smoking in human cancer." Science 354.6312 (2016): 618-622.
14. Helleday, Thomas, Saeed Eshtad, and Serena Nik-Zainal. "Mechanisms underlying mutational signatures in human cancers." Nature Reviews Genetics 15.9 (2014): 585.
15. Alexandrov, Ludmil B., et al. "Signatures of mutational processes in human cancer." Nature 500.7463 (2013): 415.
16. Goldman, Mary, et al. "Online resources for PCAWG data exploration, visualization, and discovery." bioRxiv (2018): 163907.
17. Pfeifer, Gerd P., et al. "Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers." Oncogene 21.48 (2002): 7435.
18. Smela, Maryann E., et al. "The chemistry and biology of aflatoxin B1: from mutational spectrometry to carcinogenesis." Carcinogenesis 22.4 (2001): 535-545.
19. Ledford, Heidi. "DNA typos to blame for most cancer mutations." Nature News (2017).
20. Tan, Vincent YF, and Cedric Fevotte. "Automatic relevance determination in nonnegative matrix factorization with the/spl beta/-divergence." IEEE Transactions on Pattern Analysis and Machine Intelligence 35.7 (2013): 1592-1605.
21. Rahbari, Raheleh, et al. "Timing, rates and spectra of human germline mutation." Nature genetics 48.2 (2016): 126.
22. Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological) (1996): 267-288.
23. Owen, Art B., and Patrick O. Perry. "Bi-cross-validation of the SVD and the nonnegative matrix factorization." The annals of applied statistics 3.2 (2009): 564-594.
24. Shen, Hui, and Peter W. Laird. "Interplay between the cancer genome and epigenome." Cell 153.1 (2013): 38-55.

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

25. Brash, Douglas E. "UV signature mutations." Photochemistry and photobiology 91.1 (2015): 15-26.

26. Wang, Bo, et al. "Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning." Nature methods 14.4 (2017): 414.

27. Dulak, Austin M., et al. "Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity." Nature genetics 45.5 (2013): 478.

28. Brunet, Jean-Philippe, et al. "Metagenes and molecular pattern discovery using matrix factorization." Proceedings of the national academy of sciences 101.12 (2004): 4164-4169.

29. Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. "Regularization paths for generalized linear models via coordinate descent." Journal of statistical software 33.1 (2010): 1.

30. Gaujoux, Renaud, and Cathal Seoighe. "A flexible R package for nonnegative matrix factorization." BMC bioinformatics 11.1 (2010): 367.

31. Ramazzotti, Daniele, et al. "Multi-omic tumor data reveal diversity of molecular mechanisms underlying survival." bioRxiv (2018): 267245.

32. Wang, Bo, et al. "SIMLR: A Tool for Large-Scale Genomic Analyses by Multi-Kernel Learning." Proteomics 18.2 (2018): 1700232.

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tables

Signature	**Proposed etiology**	**Tumor type(s) in which this signature is prominent**
Background signature	DNA replication error	Leukemias, lymphomas, liver cancer
Signature 1	Unknown	Liver cancer
Signature 2	Unknown	N/A
Signature 3	Defective base excision repair	N/A
Signature 4	Aristolochic acid	N/A
Signature 5	Benzo(a)pyrene (Tobacco smoking)	Lung adenocarcinoma
Signature 6	APOBEC dysregulation	Head and neck cancer, subset of breast cancer
Signature 7	Cytosine methylation / deamination	Prostate cancer, pediatric brain cancer, pancreatic cancer
Signature 8	UV light	Skin melanomas
Signature 9	Gastroesophageal reflux	Esophageal cancer, subset of gastric cancers

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

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure Legends

Figure 1. a) 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). b) Schematic of the SparseSignatures method. N represents the number of tumors in the dataset, K the number of signatures.

Figure 2. The 10 mutational signatures obtained by applying SparseSignatures to a pan-cancer dataset of 2827 whole genomes.

Figure 3. Fitted values for exposure to each of the 10 signatures obtained by SparseSignatures, for 20 cancer types. 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 4. a) Scatter plot of 19 clusters obtained by SIMLR on the fitted exposure values for the pan-cancer dataset. b) Distribution of tumors from each cancer type in each of the 19 clusters. c) Box plots showing the exposure values for each signature in each of the 19 clusters, in terms of the fraction of mutations in each sample that are attributed to a given signature.

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Material Legends

Supplementary Figure 1. a) Frequency of trinucleotides in the whole human genome b) Frequency of trinucleotides in the human exome c) Ratio of the frequencies of trinucleotides in the exome to those in the whole genome. These figures are plotted with the same axis labels as mutational signatures for ease of comparison.

Supplementary Figure 2. Signature discovery using NMF on 323 liver cancer samples. a) Plot of reconstruction error with increasing number of signatures using NMF on whole-genome data. b) 7 signatures obtained by NMF on whole-genome data c) Plot of reconstruction error with increasing number of signatures using NMF on only the mutations in exons from the same dataset. d) 5 signatures obtained by NMF on exon mutations from the same dataset.

Supplementary Figure 3. Signature discovery using SignatureAnalyzer on 323 liver cancer samples. a) Plot of reconstruction error with increasing number of signatures using SignatureAnalyzer on whole-genome data. b) 22 signatures obtained by SignatureAnalyzer on whole-genome data c) Plot of reconstruction error with increasing number of signatures using SignatureAnalyzer on only the mutations in exons from the same dataset. d) 4 signatures obtained by SignatureAnalyzer on exon mutations from the same dataset.

Supplementary Figure 4. 10 signatures obtained by SignatureAnalyzer on pan-cancer data.

Supplementary Figure 5. 10 signatures obtained by nsNMF on pan-cancer data.

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Figure 6. 10 signatures obtained by applying SparseSignatures without the fixed background on pan-cancer data. No cross-validation was performed to choose parameters; the same parameters used in to produce the signatures in Figure 2 were used.

Supplementary Table 1. 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 2. 10 signatures (including the background signature) discovered by applying SparseSignatures to pan-cancer data.

Supplementary Table 3. Fitted values for exposure to each of the 10 signatures (including the background signature) discovered by applying SparseSignatures to pan-cancer data, of each of the 2827 whole genomes in the pan-cancer dataset.

Supplementary Table 4. Comparison of the signatures discovered by SparseSignatures on pan-cancer data with the signatures from COSMIC (https://cancer.sanger.ac.uk/cosmic/signatures) and (Alexandrov, Ludmil, et al. "The Repertoire of Mutational Signatures in Human Cancer." bioRxiv (2018): 322859).

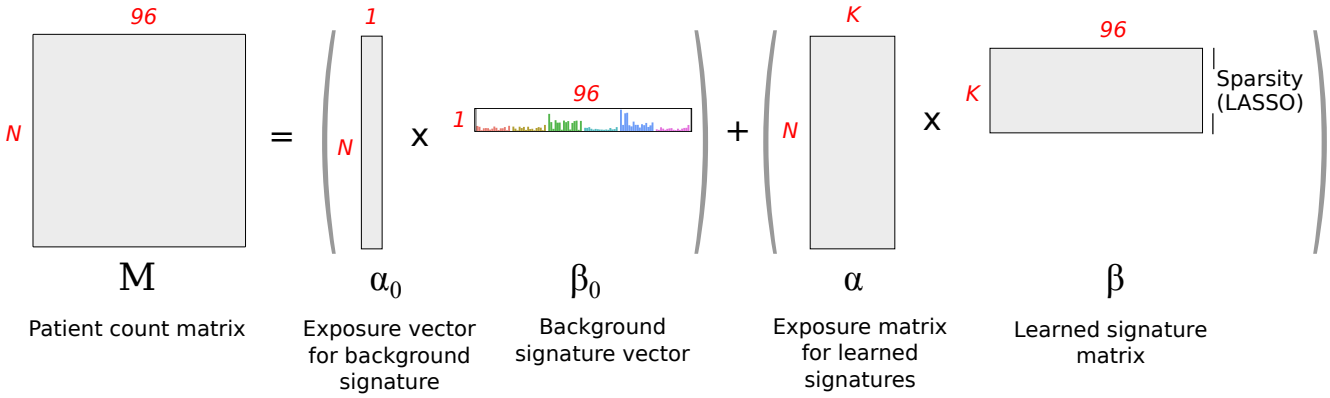
Supplementary Table 5. Cluster assignments generated by SIMLR for each sample.

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

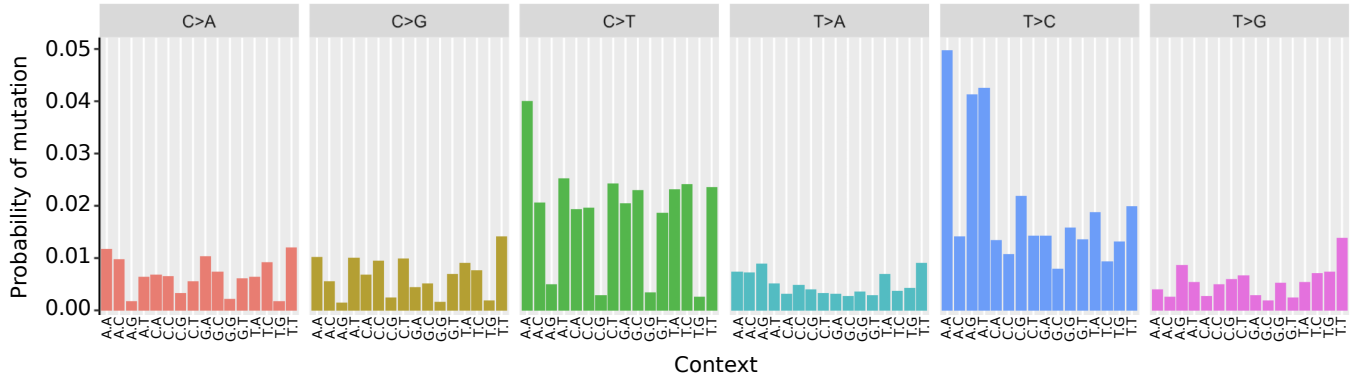
Supplementary Table 6. Results of cross-validation to choose the best values of K on simulated data, using 0.1%, 1%, and 10% of the cells in the matrix M for cross-validation. Cross-validation was repeated 100 times for each percentage of cells. The entries in the table represent the median mean square error (MSE) in fitting the unseen data points across the 100 repetitions.

bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

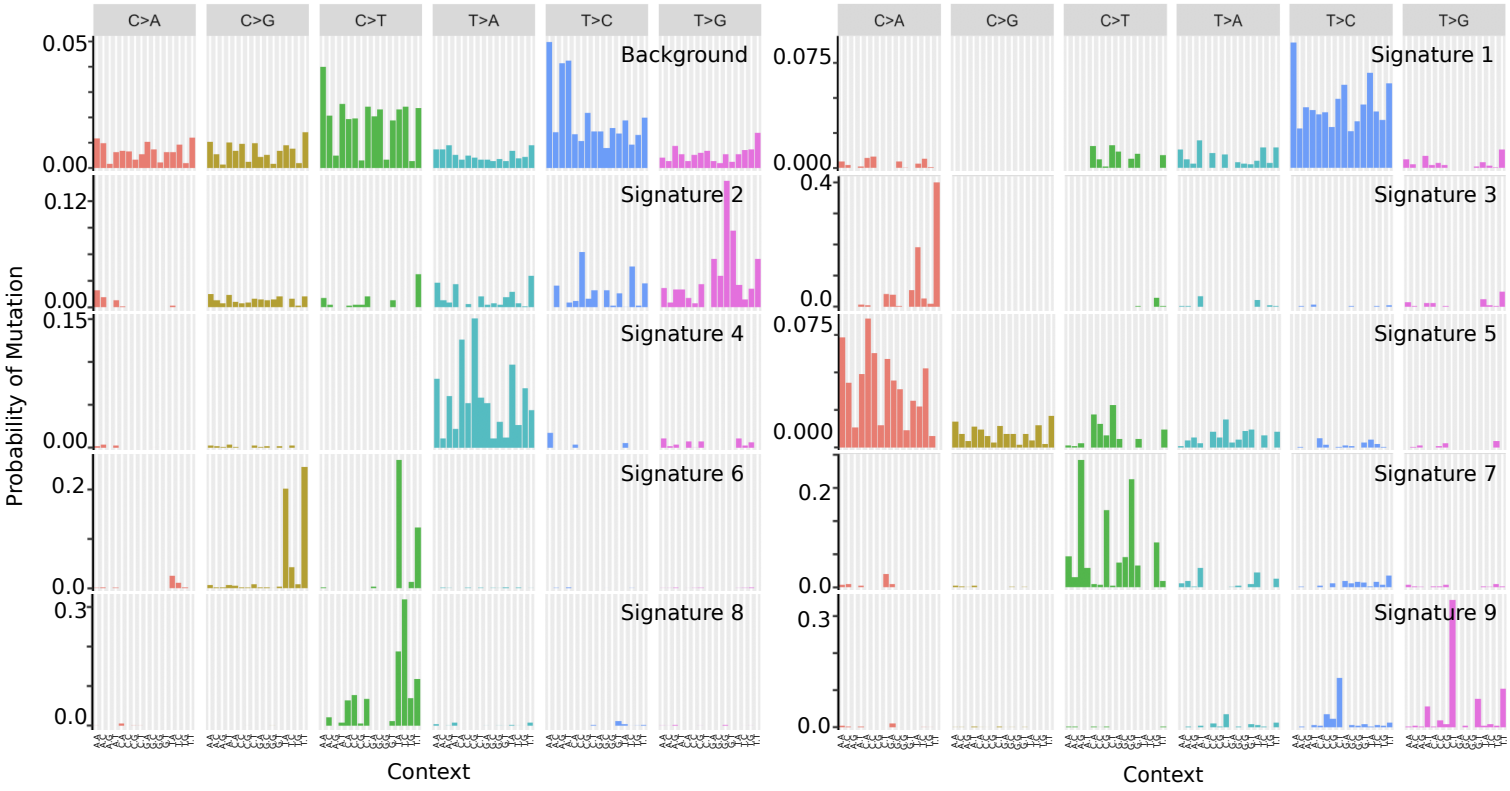
a



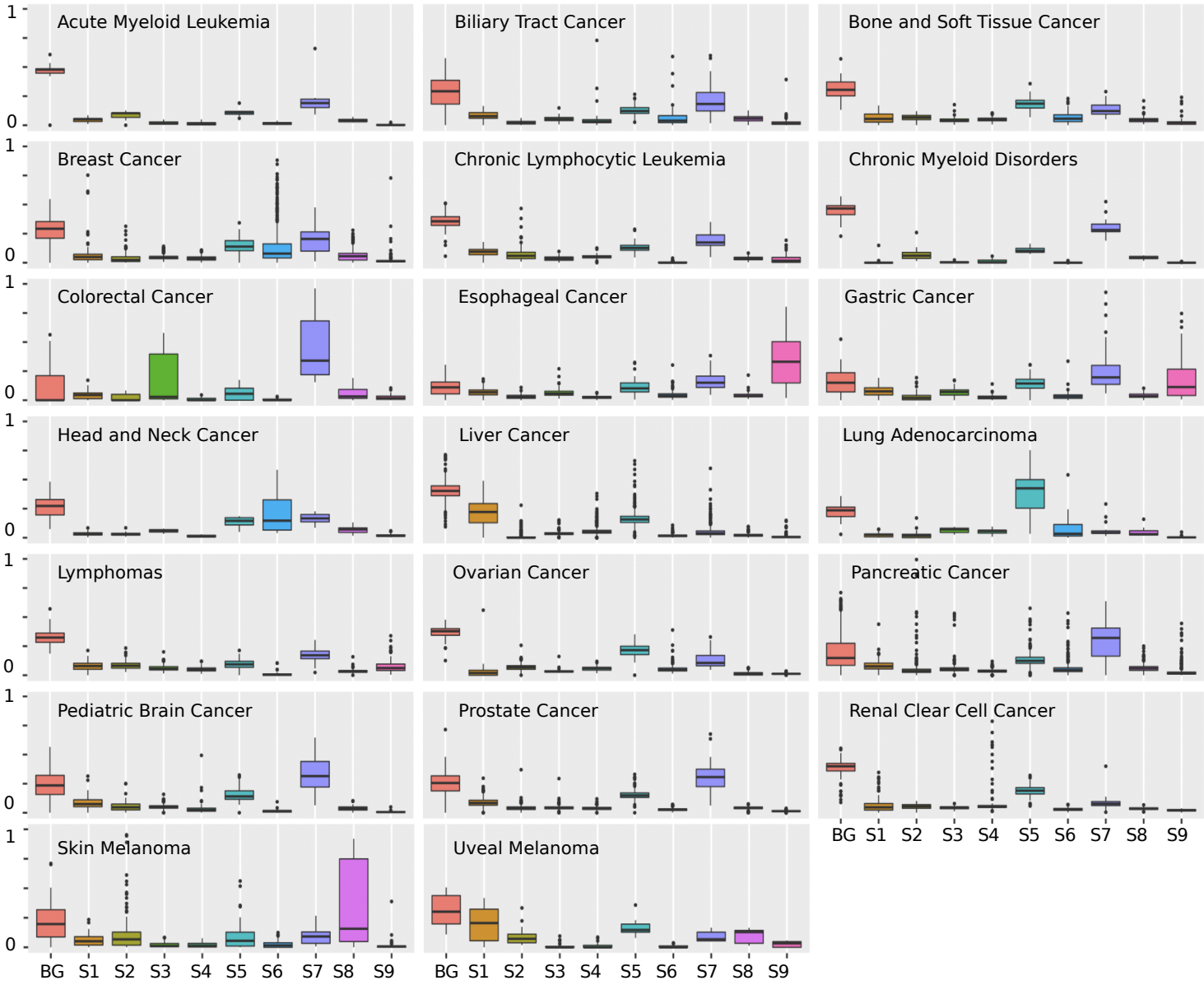
b



bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



bioRxiv preprint doi: https://doi.org/10.1101/384834; this version posted August 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.