A mixture model for signature discovery from sparse mutation data

Itay Sason[1], Yuexi Chen[2], Mark D.M. Leiserson[2], and Roded Sharan[1,⋆]

[1] School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel.
[2] Department of Computer Science and Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, 20740, USA

Abstract. Mutational signatures and their exposures are key to understanding the processes that shape cancer genomes with applications to diagnosis and treatment. Yet current signature discovery or refitting approaches are limited to relatively rich mutation data that comes from whole-genome or whole-exome sequencing. Recently, orders of magnitude sparser data sets from gene panel sequencing have become increasingly available in the clinical setting. Such data has typically less than 10 mutations per sample, making it challenging to deal with using current approaches. Here we suggest a novel mixture model for sparse mutation data. In application to simulated and real gene panel sequences it is shown to outperform current approaches and yield mutational signatures and patient stratifications that are in higher agreement with the literature.

1 Introduction

Each cancer genome is shaped by a combination of processes that introduce mutations over time [14,31]. The incidence and etiology of these mutational processes may provide insight into tumorigenesis and personalized therapy. It is thus important to uncover the characteristic signatures of active mutational processes in patients from their patterns of single base substitutions [2,3,22]. Some such mutation signatures have been linked to exposure to specific carcinogens, such as tobacco smoke [1] and ultraviolet radiation [2]. Other mutation signatures arise from deficient DNA damage repair pathways. By serving as a proxy for the functional status of the repair pathway, mutational signatures provide an avenue around traditional driver mutation analyses. This is important for personalizing cancer therapies, many of which work by causing DNA damage or inhibiting DNA damage response or repair genes [12,17,21,24], because the functional effect of many variants is hard to predict. Indeed, a recent study [8] estimated a >4-fold increase in the number of breast cancer patients with homologous recombination repair deficiency – making them eligible for PARP inhibitors [9] – when using mutational signatures compared to current approaches. Thus, understanding the signatures of mutational processes may lead to the development of many effective diagnostic and treatment strategies.

[⋆] Corresponding author. Email: `roded@tauex.tau.ac.il`.

Statistical models for discovering and characterizing mutational signatures are crucial for realizing their potential as biomarkers in the clinic. A broad catalogue of mutational signatures in cancer genomes was only recently revealed through computational analysis of mutations in thousands of tumors. Alexandrov et al. [2,3] were the first to use non-negative matrix factorization (NMF) to discover mutation signatures. Subsequent methods have used different forms of NMF [7,10,18,27], or focused on inferring the exposures (aka refitting) given the signatures and mutation counts [15,28,5]. A more recent class of approaches borrows from the world of topic modeling, aiming to provide a probabilistic model of the data so as to maximize the model's likelihood [11,29,32,26].

These previous methods are applicable for whole-genome or even whole-exome sequencing. However, they cannot handle very sparse data as obtained routinely in targeted (gene panel) sequencing assays [13]. There is only a single method, SigMA, that attempted to address this challenge [13] by relying on whole-genome training data to interpret sparse samples and predict their homologous recombination deficiency status. Here we present the first model that can handle such sparse data *without pre-training* on rich data. Our model simultaneously clusters the samples and learns the mutational landscape of each cluster, thereby overcoming the sparsity problem. Using simulated and real targeted sequencing data, we show that our method is superior to current non-sparse approaches in signature discovery, signature refitting and patient stratification.

2 Methods

2.1 Preliminaries

We follow previous work and assume that somatic mutations in cancer fall into $M = 96$ categories (denoting the mutation identity and its flanking bases). These mutations are assumed to be the result of the activity of K (a hyperparameter) mutational processes, each of which is associated with a signature $S_i = (e_i(1) \dots e_i(M))$ of probabilities to emit each of the mutation categories. Henceforth, we denote the mutation categories observed in a given tumor n by $O^n = (o_1 \dots o_{T_n})$ and we assume that this sequence was emitted by the (hidden) signature sequence $Z^n = (z_1 \dots z_{T_n})$.

2.2 Multinomial mixture model (MMM)

The basic multinomial mixture model we will use was presented in [4,32] and is depicted in Figure 1. The model is parameterized by the signatures $S_1, \dots, S_K$ and their exposure vector π, where π_i is the prior probability for the ith signature to emit any given mutation.

In the following exposition we assume for simplicity a single sample to facilitate the generalization to the `Mix` model presented in the next section. Given the observed mutations O and the unobserved signatures Z, the model's likelihood

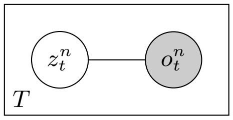


Fig. 1. A plate diagram for MMM

is:

$$\Pr\left[O\right] = \prod_{t=1}^{T} \Pr\left[o_t\right] = \prod_{t=1}^{T}\sum_{i=1}^{K} \Pr\left[o_t, z_t = i\right] = \prod_{t=1}^{T}\sum_{i=1}^{K} \pi_i e_i(o_t)$$

Denoting by $V_j = |\{t|o_t = j\}|$ the number of times the jth category appears in the data, the likelihood can be rewritten as:

$$f(V, \pi, e) := \prod_{j=1}^{M} \left(\sum_{i=1}^{K} \pi_i e_i(j)\right)^{V_j}$$

The likelihood can be maximized using the Expectation Maximization (EM) algorithm. In the E-step we compute the expectation of the model's emissions and exposures under the current assignment to those parameters. Specifically:

- The expected number of times that signature i emitted mutation category j is computed by $E_i(j, V, \pi, e) := \frac{V_j \pi_i e_i(j)}{\sum_{k=1}^{K} \pi_k e_k(j)}$.
- Similarly, the expected number of times signature i was used is computed by $A_i(V, \pi, e) := \sum_{j=1}^{M} E_i(j, V, \pi, e)$.

These expectations are normalized (to probabilities) in the M-step to yield a new set of parameters until convergence.

One obvious weakness of this model is that given a collection of samples, we cannot expect all of them to have the same exposures π. While it is possible to learn a unique exposure vector per sample, the number of parameters then grows linearly with the number of samples, which may lead to overfitting in a sparse data scenario.

2.3 `Mix`: A mixture of MMMs

In order to cope with the problem of sparse data, our approach is to cluster the samples and learn exposures per cluster rather than per sample. To this end, we propose a mixture model and a scheme to optimize its likelihood, leading to simultaneous optimization of sample (soft) clustering and model's parameters (Figure 2). Given a hyper-parameter L indicating the number of clusters, denote by $c^n \in \{1 \dots L\}$ the hidden variables representing the true cluster identity of each sample. Our goal is to learn cluster prior probabilities $w = (w_1 \dots w_L)$, cluster exposures $\pi = (\pi^1 \dots \pi^L)$, and shared signatures e, so as to maximize the model's likelihood:

$$\Pr[V|w,\pi,e] = \prod_{n=1}^{N} \Pr[V^n|w,\pi,e] = \prod_{n=1}^{N}\sum_{\ell=1}^{L} \Pr[c^n=\ell, V^n|w,\pi,e]$$
$$= \prod_{n=1}^{N}\sum_{\ell=1}^{L} \Pr[c^n=\ell]\Pr\left[V^n|\pi^\ell,e\right] = \prod_{n=1}^{N}\sum_{\ell=1}^{L} w_\ell f(V^n,\pi^\ell,e)$$

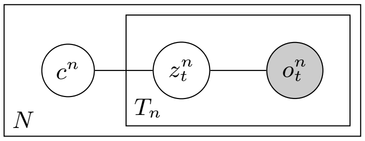


Fig. 2. A plate diagram for `Mix`.

We can generalize the EM algorithm of the previous MMM model to `Mix` as follows:

E-step : Compute for every i, j, n, ℓ:

- $f^{n,\ell} = \Pr[c^n=\ell|V^n,w,\pi,e] = \frac{w_\ell f(V^n,\pi^\ell,e)}{\sum_{\ell'=1}^{L} w_{\ell'} f(V^n,\pi^{\ell'},e)}$
- $E_i(j) = \sum_{n=1}^{N}\sum_{\ell=1}^{L} f^{n,\ell}E_i(j,V^n,\pi^\ell,e)$
- $A_i^\ell = \sum_{n=1}^{N} f^{n,\ell}A_i(V^n,\pi^\ell,e)$
- $W_\ell = \sum_{n=1}^{N} f^{n,\ell}$

M-step : Compute for every i, j, ℓ:

- $\pi_i^\ell = \frac{A_i^\ell}{\sum_{i'=1}^{K} A_{i'}^\ell}$
- $e_i(j) = \frac{E_i(j)}{\sum_{j'=1}^{M} E_i(j')}$
- $w_\ell = \frac{W_\ell}{\sum_{\ell'=1}^{L} W_{\ell'}}$

Each EM iteration can be completed in $\mathcal{O}(NLK)$ time for N samples, L clusters and K signatures.

2.4 Data

Mutational signatures. We present below both *de-novo* experiments in which we learn mutational signatures as well as *refitting* experiments in which we assume the signatures are given. In the latter cases, we restrict our analyses to Single Base Substitution (SBS) mutation signatures in COSMIC[1] that are known to be active in the cancer type being analyzed.

Mutation datasets. We applied Mix to analyze mutational signatures in three datasets.

1. **MSK-IMPACT [6,33] Pan-Cancer.** We downloaded mutations for a cohort of patients with Memorial Sloan Kettering Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT) targeted sequencing data from https://www.cbioportal.org/. The MSK-IMPACT dataset contains 11,369 pan-cancer patients' sequencing samples across 410 target genes. We restrict our analysis to the 18 cancer types with more than 100 samples, which results in a dataset of 5931 samples and about 7 mutations per sample. According to COSMIC there are 11 mutational signatures that are active in those cancer types and are associated with more than 5% of the mutations (where a mutational signature is associated with any mutation in a sample in which the signature is active according to COSMIC). Five of those signatures are associated with 30% or more of the mutations.
2. **ICGC breast cancer.** We downloaded mutations for 560 breast cancer patients [23] with whole-genome sequencing data from the International Cancer Genome Consortium. There are about 6214 mutations per sample in this collection and 12 active COSMIC signatures are associated with it.
3. **TCGA ovarian cancer**. We downloaded mutations from whole-exome-sequencing data of 411 ovarian cancer patients from The Cancer Genome Atlas [30]. There are about 113 mutations per sample in this collection and 3 active signatures are associated with it.

2.5 Performance evaluation in a refitting scenario

In order to evaluate Mix and other algorithms on their ability to infer accurate mutational signature exposures on sparse data, we focus on the whole-genome or whole-exome data where we have information about active signatures. The evaluation procedures require generating sparse, downsampled datasets to imitate the target sequencing data. In this section, we first describe the downsampling procedure, and then describe the way we use the downsampled datasets for evaluation.

Downsampling cancer sequencing datasets. In order to test our methods, we seek to simulate targeted sequencing panels such as MSK-IMPACT from higher coverage datasets. Ideally, we would simply take any mutation in the higher coverage dataset that occurs in any targeted location in the panel, but Gulhan et al. [13] found that approximately 50% of the mutations in the MSK-IMPACT dataset were not in target regions. Instead, these mutations were most often in

[1] https://cancer.sanger.ac.uk/cosmic/signatures/SBS/

non-coding regions of the genes. Because the complete set of genomic locations in the panel was not available, given a *full* dataset of mutation counts, we generate a *downsampled* version of the dataset by selecting a proportion d of mutations uniformly at random. We set d in order to approximate the number of mutations observed in targeted sequencing panels. Specifically, for ICGC breast cancer, $d = 0.4\%$ and 0.2%, and the average mutation count per sample is 25 and 13, respectively; for TCGA ovarian cancer, $d = 20\%$ and 10%, and the average mutation count per sample is 23 and 11, respectively.

Reconstruction error. To compare methods in their ability to learn mutational signature exposures on sparse datasets, we compare the reconstruction error (RE) obtained by each method on a *full* dataset using relative exposures inferred on a *downsampled* dataset. For ease of comparison, we fix the signature matrix S to be known signatures from COSMIC. Since the full and downsampled datasets have different numbers of mutations, we compare them only on their *relative* exposures. Let V be an $N \times M$ matrix where V_{ij} is the number of times mutation category j is observed in tumor i in the full dataset (minus the downsampled one), and let $\tilde{V}$ be the matrix V normalized so that each row sums to one. Then, given the $N \times K$ relative exposure matrix E_d computed on the downsampled data, we define the reconstruction error as

$$RE = \frac{\left\| \tilde{V} - E_d \cdot S \right\|_F}{\left\| \tilde{V} \right\|_F} \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm. We explore two ways of computing the row of E_d corresponding to some sample n: (i) as the exposure vector of the cluster associated with n (i.e., with maximum posterior probability); and (ii) as the weighted sum of exposure vectors of all clusters, where each vector is weighted by the posterior probability of the corresponding cluster.

2.6 Software

`Mix` is implemented in Python 3 and is available at `https://github.com/itaysason/Mix-MMM`. The data generation and processing workflow is managed by Snakemake [19].

3 Results

We tested our algorithm on both gene panel data and downsampled whole-genome/whole-exome data, and compared its performance to existing approaches. First, we applied `Mix` to the MSK-IMPACT Pan-Cancer targeted sequencing data. We tested its success in discovering mutational signatures and in clustering the patients to their cancer types. Second, we applied `Mix` to reconstruct mutation profiles for downsampled ICGC breast cancer and TCGA ovarian cancer data.

3.1 Learning signatures and patient classes from MSK-IMPACT

We applied `Mix` to analyze 5931 samples from the MSK IMPACT targeted sequencing dataset. As a first test, we assessed the algorithm's ability to recover signatures known to be active in each cancer type according to COSMIC. We trained `Mix` with ten random initializations on number L of clusters ranging from 1 to 15 and number K of signatures ranging from 1 to 11 (5-11 signatures are associated with these data according to COSMIC, see Methods). We used the Bayesian information criterion (BIC) to weigh the tradeoff between model fit and the number of parameters, and found that $L = 10$ and $K = 6$ had the lowest BIC score (Figure 3).

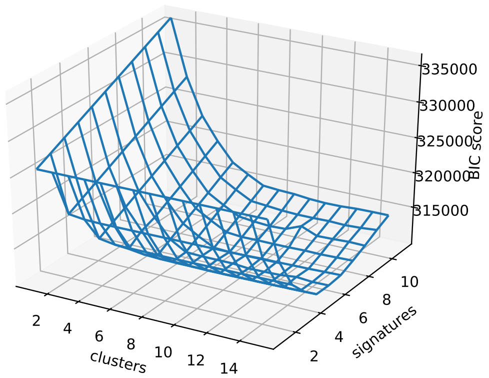
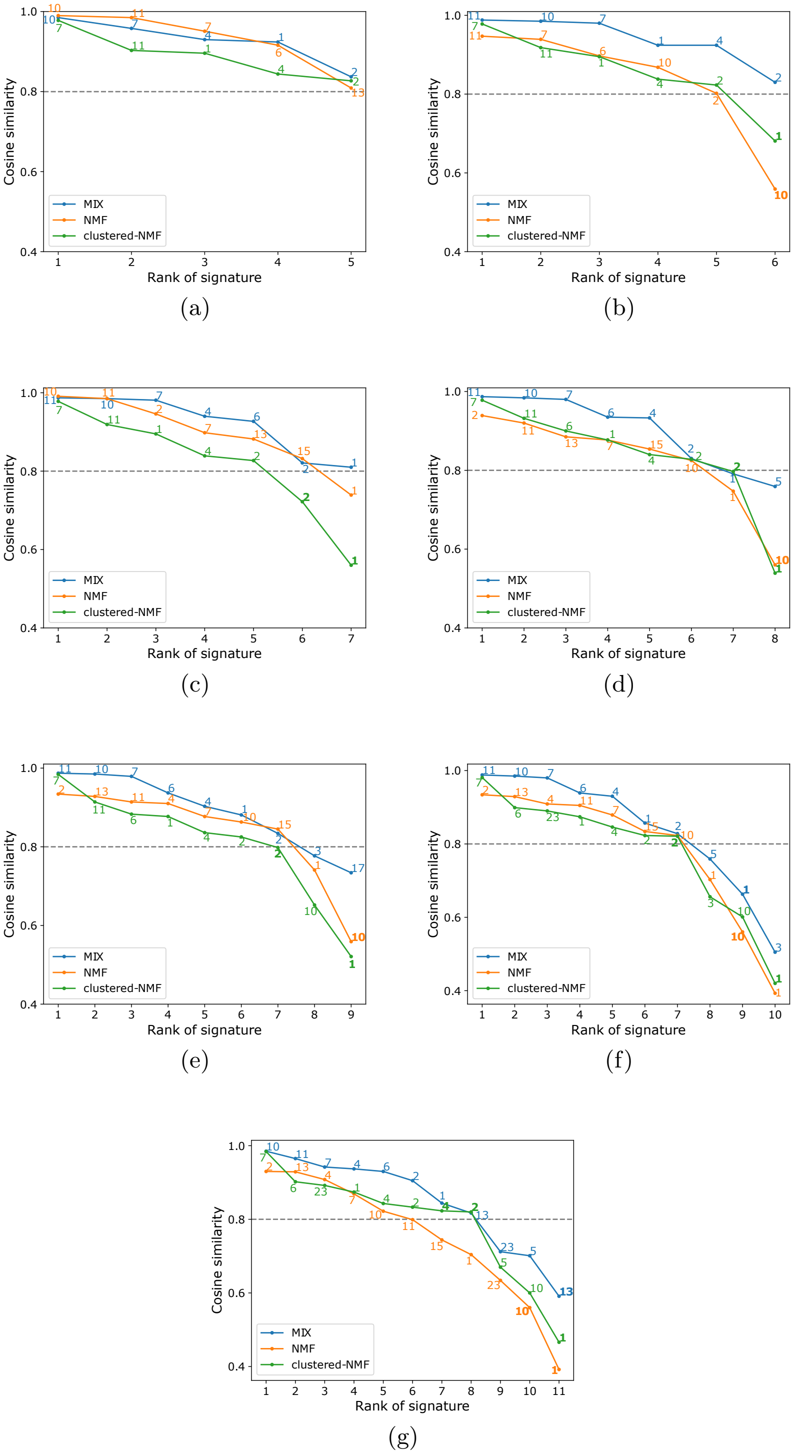


Fig. 3. Hyper-parameter selection in `Mix`. A plot of BIC as a function of the number of signatures (z-axis) and the number of clusters (x-axis).

To evaluate the learned signatures, we compared them to the COSMIC signatures using the cosine similarity measure (Figure 4). `Mix` accurately reconstructed 5-8 known signatures with cosine similarity > 0.8 as commonly required [18]. We compared our performance to that of the standard NMF algorithm as well as to a clustered variant where we first form meta-samples by summing together all samples within cancer types, and only then apply the NMF. For these additional applications we varied the number of signatures from 5 to 11; for `Mix` we optimized the number of clusters in each application using BIC as described above. Further, we executed each algorithm ten times with different (random) initializations and chose the run that yielded the best score (likelihood for `Mix` or approximation error for NMF). Evidently, `Mix` dominates the other approaches across the explored range, yielding a larger number of highly accurate and distinct signatures.



(a)

(b)

(c)

(d)

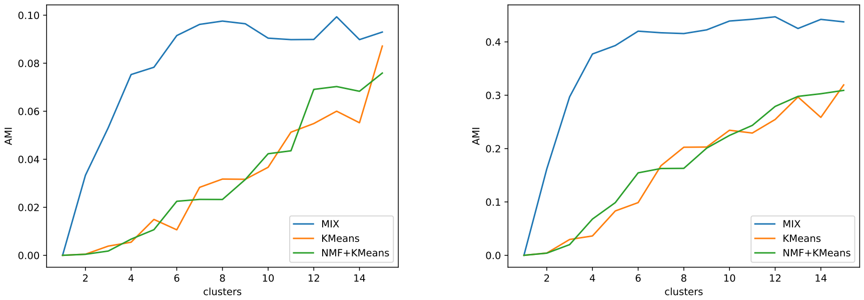
(e)

(f)

(g)

Fig. 4. De-novo signature discovery from MSK-IMPACT panel data. Shown are sorted cosine similarities between learned signatures and most similar COSMIC signature (denoted next to the plot) for `Mix`, NMF and clustered NMF across a range of number of signatures (5-11 corresponding to (a)-(g), respectively). Repeating signatures of the same model are in bold.

Next, we used `Mix` to cluster samples: for each sample we chose the cluster with the highest posterior probability. We scored the resulting clustering against a benchmark clustering of the samples according to their cancer type with the adjusted mutual information (AMI) score. We note that in addition to validating our results, predicting cancer type from targeted sequencing panels has potential clinical relevance, as approximately 3% of tumors are of unknown primary origin [25] and there has been a recent focus on developing methods to predict cancer type using mutations [16,20]. We compared our results to those obtained by KMeans clustering of the original mutation count vectors as well as to a refined variant where we first apply NMF to the data (in a de-novo setting) and then cluster the resulting exposures using KMeans. For all methods we report results with $K = 6$ signatures and $L = 1 - 15$ clusters. As the clustering of specific samples depends on their sparsity, we also report AMI scores when focusing on samples with at least 10 mutations. The results appear in Figure 5 and demonstrate the superiority of `Mix` over the alternative approaches.



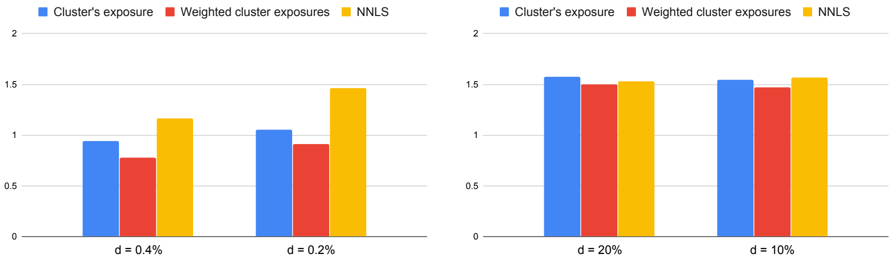
(a) (b)

Fig. 5. AMI score as a function of the number of clusters for each model across (a) all samples, or (b) samples with 10 or more mutations.

3.2 Reconstructing mutation profiles from sparse data for breast and ovarian cancer patients

Most methods for mutation signature analysis rely on the rich information contained in whole-genome or whole-exome data. Such reliance limits their use in the clinical practice, where the most common scenario is of sparse targeted gene panel sequencing data. In this section we evaluate `Mix` in reconstructing mutation profiles under sparse data and compare to the widely-used non-negative least squares (NNLS) approach.

In these experiments, we focus on learning the exposures and fix the COSMIC signatures to be those active in the given cancer type. We train `Mix` with downsampled data from 50% of samples, compute exposures on downsampled data for the remaining 50% of (test) samples, and report the reconstruction error on the rest of the mutations in the test samples (i.e., those mutations hidden from us due to the downsampling, see Methods for complete details). We repeat these experiments for two realistic downsampling proportions d. We compare `Mix` to NNLS, which is also applied to the downsampled portions of the test samples and evaluated on the rest of the test sample mutations. The results are shown in Figure 6, exemplifying the power of `Mix` also in this refitting setting. We can also observe that a weighted refitting by `Mix` yields the lowest reconstruction error in all cases.



(a) (b)

Fig. 6. Reconstruction error on downsampled data. Shown are reconstruction errors (RE, y-axis) for `Mix` (based on cluster exposures) and NNLS across two datasets, breast cancer (a) and ovarian cancer (b), and two downsampling proportions.

4 Conclusions

Sparse mutation data, as characteristic of targeted sequencing assays, is becoming increasingly available in the clinical setting with important applications in diagnosis and therapy. In this paper we have presented a novel algorithm to model such data and derive the underlying mutational signatures, exposures and clinically-relevant predictions. Our model is the first to capture sparse data directly without the need for pre-training on rich datasets. We have shown its utility in a range of tasks as well as its favorable performance in comparison to existing methods. Of special interest may be the principled way in which we optimize the hyper-parameters of our model using the Bayesian information criterion. In addition, our method may enhance the SigMA pipeline [13], which uses NMF and NNLS for discovering signatures and computing exposures, respectively, as part of a larger pipeline that includes panel simulations and training classifiers for predicting homologous recombination deficiency. We plan to investigate whether using `Mix` instead of NMF and NNLS leads SigMA to more accurate, clinically-relevant predictions.

Acknowledgements

This study was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. RS was supported by a research grant from the Israel Science Foundation (grant no. 715/18).

References

1. L. B. Alexandrov, Y. S. Ju, K. Haase, P. V. Loo, I. Martincorena, S. Nik-Zainal, Y. Totoki, A. Fujimoto, H. Nakagawa, T. Shibata, P. J. Campbell, P. Vineis, D. H. Phillips, and M. R. Stratton. Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312):618–622, 2016.
2. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. Aparicio, S. Behjati, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013.
3. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports*, 3(1):246–259, 2013.
4. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.
5. F. Blokzijl, R. Janssen, R. van Boxtel, and E. Cuppen. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Medicine*, 10:33, 2018.
6. D. T. Cheng, T. N. Mitchell, A. Zehir, R. H. Shah, R. Benayed, A. Syed, R. Chandramohan, Z. Y. Liu, H. H. Won, S. N. Scott, et al. Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (msk-impact): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *The Journal of molecular diagnostics*, 17(3):251–264, 2015.
7. K. Covington, E. Shinbrot, and D. A. Wheeler. Mutation signatures reveal biological processes in human cancer. *bioRxiv*, page 036541, 2016.
8. H. Davies, D. Glodzik, S. Morganella, L. R. Yates, J. Staaf, X. Zou, M. Ramakrishna, S. Martin, S. Boyault, A. M. Sieuwerts, P. T. Simpson, T. A. King, K. Raine, J. E. Eyfjord, G. Kong, Åke Borg, E. Birney, H. G. Stunnenberg, M. J. van de Vijver, A.-L. Børresen-Dale, J. W. Martens, P. N. Span, S. R. Lakhani, A. Vincent-Salomon, C. Sotiriou, A. Tutt, A. M. Thompson, S. Laere, A. L. Richardson, A. Viari, P. J. Campbell, M. R. Stratton, and S. Nik-Zainal. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nature Medicine*, 23(4):517–525, 2017.
9. H. Farmer, N. McCabe, C. J. Lord, A. N. Tutt, D. A. Johnson, T. B. Richardson, M. Santarosa, K. J. Dillon, I. Hickson, C. Knights, N. Martin, S. P. Jackson, G. C. Smith, and A. Ashworth. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature*, 434(7035):917–921, 2005.
10. A. Fischer, C. J. Illingworth, P. J. Campbell, and V. Mustonen. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biology*, 14(4):1–10, 2013.
11. T. Funnell, A. Zhang, Y.-J. Shiah, D. Grewal, R. Lesurf, et al. Integrated single-nucleotide and structural variation signatures of DNA-repair deficient human cancers. *bioRxiv*, page 267500, 2018.
12. N. S. Gavande, P. S. VanderVere-Carozza, H. D. Hinshaw, S. I. Jalal, C. R. Sears, K. S. Pawelczak, and J. J. Turchi. DNA repair targeted therapy: The past or future of cancer treatment? *Pharmacology Therapeutics*, 160:65–83, 2016.

13. D. C. Gulhan, J. J.-K. Lee, G. E. Melloni, I. Cortés-Ciriano, and P. J. Park. Detecting the mutational signature of homologous recombination deficiency in clinical samples. Technical report, Nature Publishing Group, 2019.
14. T. Helleday, S. Eshtad, and S. Nik-Zainal. Mechanisms underlying mutational signatures in human cancers. *Nature Reviews Genetics*, 15(9):585–598, 2014.
15. X. Huang, D. Wojtowicz, and T. M. Przytycka. Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics (Oxford and England)*, 2017.
16. W. Jiao, G. Atwal, P. Polak, R. Karlic, E. Cuppen, et al. A deep learning system can accurately classify primary and metastatic cancers based on patterns of passenger mutations. bioRxiv, 2017.
17. A. Khanna. DNA Damage in Cancer Therapeutics: A Boon or a Curse? *Cancer Research*, 75(11):2133–2138, 2015.
18. J. Kim, K. W. Mouw, P. Polak, L. Z. Braunstein, A. Kamburov, et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nature Genetics*, 48(6):600–606, 2016.
19. J. Köster and S. Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.
20. K. Kübler, R. Karlić, N. J. Haradhvala, K. Ha, J. Kim, et al. Tumor mutational landscape is a record of the pre-malignant state. bioRxiv, 2019.
21. K. W. Mouw, M. S. Goldberg, P. A. Konstantinopoulos, and A. D. D'Andrea. DNA Damage and Repair Biomarkers of Immunotherapy Response. *Cancer Discovery*, 7(7):675–693, 2017.
22. S. Nik-Zainal, L. B. Alexandrov, D. C. Wedge, P. Van Loo, C. D. Greenman, et al. Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell*, 149(5):979–993, 2012.
23. S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, X. Zou, I. Martincorena, L. B. Alexandrov, S. Martin, D. C. Wedge, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47, 2016.
24. M. J. O'Connor. Targeting the DNA Damage Response in Cancer. *Molecular Cell*, 60(4):547–560, 2015.
25. N. Pavlidis, H. Khaled, and R. Gaafar. A mini review on cancer of unknown primary site: A clinical puzzle for the oncologists. *Journal of Advanced Research*, 6:375–382, 2015.
26. W. Robinson, R. Sharan, and M. D. Leiserson. Modeling clinical and molecular covariates of mutational process activity in cancer. *Bioinformatics*, 35(14):i492–i500, 2019.
27. R. A. Rosales, R. D. Drummond, R. Valieris, E. Dias-Neto, and I. T. da Silva. signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics*, 33(1):8–16, 2016.
28. R. Rosenthal, N. McGranahan, J. Herrero, B. S. Taylor, and C. Swanton. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology*, 17(1):31, 2016.
29. Y. Shiraishi, G. Tremmel, S. Miyano, and M. Stephens. A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures. *PLOS Genetics*, 11(12):e1005657, 2015.
30. K. Tomczak, P. Czerwińska, and M. Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A):A68, 2015.
31. A. Tubbs and A. Nussenzweig. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell*, 168(4):644–656, 2017.

32. D. Wojtowicz, I. Sason, X. Huang, Y.-A. Kim, M. D. M. Leiserson, T. M. Przytycka, and R. Sharan. Hidden Markov models lead to higher resolution maps of mutation signature activity in cancer. *Genome Medicine*, 11:49, 2019.
33. A. Zehir, R. Benayed, R. H. Shah, A. Syed, S. Middha, H. R. Kim, P. Srinivasan, J. Gao, D. Chakravarty, S. M. Devlin, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nature medicine*, 23(6):703, 2017.