Bioinformatics, 35, 2019, i492–i500
doi: 10.1093/bioinformatics/btz340
ISMB/ECCB 2019

Modeling clinical and molecular covariates of mutational process activity in cancer

Welles Robinson[1,2], Roded Sharan[3] and Mark D. M. Leiserson [1,*]

[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20910, USA, [2]Cancer Data Science Laboratory, National Cancer Institute, NIH, Bethesda, MD 20894, USA and [3]Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 6997801, Israel

*To whom correspondence should be addressed.

Abstract

Motivation: Somatic mutations result from processes related to DNA replication or environmental/lifestyle exposures. Knowing the activity of mutational processes in a tumor can inform personalized therapies, early detection, and understanding of tumorigenesis. Computational methods have revealed 30 validated *signatures* of mutational processes active in human cancers, where each signature is a pattern of single base substitutions. However, half of these signatures have no known etiology, and some similar signatures have distinct etiologies, making patterns of mutation signature activity hard to interpret. Existing mutation signature detection methods do not consider tumor-level clinical/demographic (e.g. smoking history) or molecular features (e.g. inactivations to DNA damage repair genes).

Results: To begin to address these challenges, we present the Tumor Covariate Signature Model (TCSM), the first method to directly model the effect of observed tumor-level covariates on mutation signatures. To this end, our model uses methods from Bayesian topic modeling to change the prior distribution on signature exposure conditioned on a tumor's observed covariates. We also introduce methods for imputing covariates in held-out data and for evaluating the statistical significance of signature-covariate associations. On simulated and real data, we find that TCSM outperforms both non-negative matrix factorization and topic modeling-based approaches, particularly in recovering the ground truth exposure to similar signatures. We then use TCSM to discover five mutation signatures in breast cancer and predict homologous recombination repair deficiency in held-out tumors. We also discover four signatures in a combined melanoma and lung cancer cohort—using cancer type as a covariate—and provide statistical evidence to support earlier claims that three lung cancers from The Cancer Genome Atlas are misdiagnosed metastatic melanomas.

Availability and implementation: TCSM is implemented in Python 3 and available at https://github.com/lrgr/tcsm, along with a data workflow for reproducing the experiments in the paper.

Contact: mdml@cs.umd.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Somatic mutations accumulate over time in normal and cancer cells as a consequence of multiple mutational processes. Measuring and understanding the activity of these mutational process within and across tumors has important applications in modeling tumorigenesis, personalized cancer therapy, early detection and prevention. The large cancer sequencing datasets generated over the past decade have led to the discovery of signatures of mutational processes present in patterns of single base substitutions (Alexandrov *et al.*, 2013a). Discovering and characterizing these *mutation signatures* and their underlying etiology has thus become an important challenge in the field.

The sources of somatic mutations can be broadly classified as due to errors in DNA replication or from environmental or lifestyle exposures (Tomasetti *et al.*, 2017). Errors in DNA replication result both from processes active in healthy cells (e.g. due to spontaneous deamination or reactive oxygen species) and from perturbed DNA damage repair pathways (Tubbs and Nussenzweig, 2017). Clinicians use measures of DNA damage repair deficiency for multiple types of cancer therapy, including chemotherapy (Hegi *et al.*, 2005), synthetic lethal therapy (Farmer *et al.*, 2005), and, more

recently, checkpoint inhibitor immunotherapy (Le *et al.*, 2017). A recent study evaluated mutation signatures of homologous recombination (HR) repair deficiency in breast cancer as a predictive biomarker, and found that the mutation signature-based approach would significantly expand the population of patients eligible for PARP inhibitors (Davies *et al.*, 2017). Mutations also result from environmental or lifestyle exposures, including UV radiation and tobacco smoke (Pfeifer, 2010), as well as many DNA damaging agents used as chemotherapies (Szikriszt *et al.*, 2016). Mutation signatures of these exogenous processes have recently been shown to be prognostic in cutaneous melanomas (Trucco *et al.*, 2018), and revealed pre-cancerous aflatoxin B1 exposure in mice (Chawanthayatham *et al.*, 2017). More generally, these sources of somatic mutations can be thought of as tumor-level covariates where for a given covariate (e.g. smoking status), each tumor is annotated with a specific value (e.g. smoker or non-smoker).

The most widely used methods for discovering mutation signatures are based on non-negative matrix factorization (NMF) of a mutation count matrix (Alexandrov *et al.*, 2013b). To identify signatures in a cohort of N tumors, single base substitutions are first grouped into 96 categories (based on the substitution and its surrounding 5′ and 3′ contexts), yielding an $N \times 96$ matrix M of mutation counts. Then, NMF is applied to decompose M into a $N \times K$ *exposures* matrix E and an $K \times 96$ *signatures* matrix P, and E and P are rescaled so that the rows of P sum to one. Each entry E_{ij} is interpreted as the number of mutations in tumor i generated by signature j, and P_{kj} is the probability signature k generates a mutation of category j. Alexandrov *et al.* (2013a) applied this model to >7000 tumors from 30 different cancer types to identify 20 mutation signatures. Alexandrov and colleagues have since expanded the set to include 30 validated signatures that are widely studied and available from the Catalogue of Somatic Mutations in Cancer (COSMIC) (Forbes *et al.*, 2017) (https://cancer.sanger.ac.uk/cosmic/signatures (8 May 2019, date last accessed)).

Since Alexandrov *et al.* (2013b) first applied NMF to identify mutation signatures, researchers have developed additional NMF algorithms and addressed the problem of inferring exposures in a cohort given a set of active signatures. Kasar *et al.* (2015) introduced the SignatureAnalyzer method that uses a probabilistic formulation of NMF and automatically learns its rank K. Fischer *et al.* (2013) and Rosales *et al.* (2017) both introduced algorithms for NMF that assume that the mutation counts are drawn from a Poisson distribution parameterized by multiplying factors with the latter algorithm using a Gamma prior. Rosenthal *et al.* (2016) introduced several heuristics for computing the exposure matrix E given a signature matrix P, and Huang *et al.* (2017) extended this work to solve the problem optimally.

A handful of researchers have also considered a second type of approaches to inferring mutation signatures that leverages lessons from the natural language processing problem of *topic modeling*. Given a corpus of observed documents, each drawn from the same vocabulary, the goal of topic modeling is to infer latent topics (distributions over words) and to assign each word in each document to its underlying topic (Blei, 2012). Most topic modeling approaches such as the standard latent Dirichlet allocation (LDA) (Blei *et al.*, 2003) are Bayesian and make the 'bag-of-words' assumption that each word in a document is independent given its underlying topic. Applying topic modeling to mutation signatures means interpreting tumors as documents, signatures as topics and mutation categories as the vocabulary. Shiraishi *et al.* (2015) introduced the pmsignatures method that generalizes LDA to enable mutation categorizations that include more than one flanking base. Funnell *et al.* (2018) used a multi-modal topic modeling approach to simultaneously analyze patterns in single base substitutions and structural variations in breast and ovarian cancers.

Despite this methodological progress, about half of the 30 validated COSMIC signatures have no known etiology. The most common approach to mapping signatures to their underlying causes is to show statistically significant associations between signature exposure and a clinical/demographic features (e.g. a history of smoking and COSMIC Signature 4; Alexandrov *et al.*, 2016) or molecular features (e.g. *BRCA1* inactivations and COSMIC Signature 3; Polak *et al.*, 2017).

Furthermore, even for two signatures with known etiologies, it can be challenging to distinguish their respective exposures with existing methods if the signatures are similar. For example, COSMIC Signature 3 and Signature 5 are highly similar (cosine similarity of 0.83), but Signature 3 is associated with HR repair deficiency (Nik-Zainal *et al.*, 2016; Polak *et al.*, 2017) and Signature 5 is associated with age at diagnosis (Alexandrov *et al.*, 2015) and genetic mutations in the nucleotide excision repair pathway (Kim *et al.*, 2016).

We hypothesize that to overcome these challenges, methods for modeling mutation signatures and tumor-level clinical or molecular covariates are needed. To begin to address this challenge, we present the Tumor Covariate Signature Model (TCSM) to learn how observed tumor-level covariates change signature exposure. We show on simulated and real mutation datasets that, by modeling tumor-level covariates, TCSM outperforms existing NMF- and topic modeling-based approaches that are limited to using only a tumor's mutations as input. We find that the largest differences in performance come when inferring exposures of held-out tumors not used to infer signatures, and that these differences lead to improved performance in downstream analyses, including predicting DNA damage repair deficiency. TCSM is the first method to model mutation signatures and their tumor-level covariates in order to automatically infer signature etiology.

2 Materials and methods

2.1 Tumor-covariate signature model

We present a probabilistic model of mutation signatures and their covariates that builds off of the well-studied area of topic modeling (Blei, 2012; Blei *et al.*, 2003), and the previously observed connection between topic modeling and mutation signatures (Funnell *et al.*, 2018; Shiraishi *et al.*, 2015). Topic models are generative models for text data, and usually encode the 'bag-of-words' assumption that words are independent given their underlying topics. The observed data for topic models are N documents w, where each document w_i consists of n_i words from vocabulary V such that $w_{ij} \in V, 1 \leq j \leq n_i$. Topic modeling seeks to uncover (i) K global latent variables β_k called *topics*, where each topic is a probability distribution over the vocabulary; and (ii) local latent variables including the K *topic mixing proportions* θ_i per document, and the assignment $z_{ij} \in \{1, \ldots, K\}$ of each observed word w_{ij} to a topic. The most common topic modeling approaches such as LDA (Blei *et al.*, 2003) are Bayesian, where both β_k and θ_i are multinomial distributions with Dirichlet priors.

In order to model mutational processes in cancer, we interpret tumors as documents, mutation categories as the vocabulary, signatures as topics, and signature *exposures* as topic mixings. Following earlier work, we categorize mutations into $L = 96$ mutation categories based on its base substitution (C: G > A: T, C: G > G: C, C:

G > T: A, T: A > A: T, T: A > C: G, T: A > G: C) and the 5′ and 3′ flanking bases (four choices each) in the reference genome.

We present TCSM to allow observed tumor covariates to change the per tumor distribution θ_i of signature exposures (Fig. 1). While there is a rich history of topic modeling using document-level covariates (Mimno and McCallum, 2008; Ramage *et al.*, 2009; Roberts *et al.*, 2013), to our knowledge, this is the first time this work has been connected to mutation signatures. Importantly, we do not model the generative process of the observed covariates, but instead take a conditional approach where the D observed covariates $\vec{x_i}$ of the ith tumor change the prior distribution over the signature exposures θ_i. For example, an observed covariate could be a binary indicator for biallelic inactivation of a DNA damage repair gene. The model is flexible enough that the covariates can be any real valued number. The first element of $\vec{x}$ is always set to 1 to model the mean exposure of each signature.

More specifically, we follow the 'topic prevalence' approach of the Structural Topic Model from Roberts *et al.* (2013, 2016b) that combines Dirichlet-multinomial regression (Mimno and McCallum, 2008) and the correlated topic model (Blei and Lafferty, 2005), and describe the model as it relates to mutation signatures. The correlated topic model places a logistic normal prior on θ such that signature exposures can co-vary (correlate), and was previously used to analyze mutation signatures in breast cancer (Funnell *et al.*, 2018). The mean of the logistic normal is set for tumor i as $\vec{x_i}\Gamma$, where Γ is a $D \times (K-1)$ matrix of exposure–covariate coefficients. The full generative process for the TCSM for tumor sample i with n_i mutations is as follows:

$$\theta_i \sim \text{LogisticNormal}(\vec{x_i}\Gamma, \Sigma), \tag{1}$$

$$z_{ij} \sim \text{Multinomial}(\theta_i), 1 \leq j \leq n_i, \tag{2}$$

$$w_{ij} \sim \text{Multinomial}(\beta_{z_{ij}}), 1 \leq j \leq n_i. \tag{3}$$

We place a hyperprior on the exposure–covariate coefficients $\Gamma = [\gamma_1; \ldots; \gamma_{K-1}]$ where

$$\gamma_{d,k} \sim \text{Normal}(0, \sigma_k^2), 1 \leq d \leq D, 1 \leq k \leq K-1, \tag{4}$$

and a Half-Cauchy (1, 1) prior is placed on σ_k to weakly enforce regularization.

2.2 Model training and hyperparameter selection

We train the TCSM to learn the signatures β, signature exposures θ and covariance Σ, and exposure–covariate coefficients Γ using the variational expectation-maximization algorithm from Roberts *et al.* (2016b) and their recommended initialization procedure. The latter is based on a spectral decomposition (via NMF) of the $L \times L$ mutation co-occurrence matrix that was shown to lead to quicker convergence of topic models (Roberts *et al.*, 2016a).

The main hyperparameter of TCSM is the number K of signatures. We set K empirically through 5-fold cross-validation, completely holding out 20% of the tumors in 1-fold. We use the 'document completion' approach of Wallach *et al.* (2009) to compute the likelihood of all of a held-out tumors's mutations w_{test}, i.e. computing $\Pr(w_{\text{test}}|\beta, \cdot)$, where $\cdot$ represents hyperparameters. We choose the K when the likelihood plateaus.

2.2.1 Learning exposures in held-out samples

When the signatures β are given (e.g. from learning on a training cohort), we learn the exposures θ for additional, held-out samples by maximum a posteriori probability estimation.

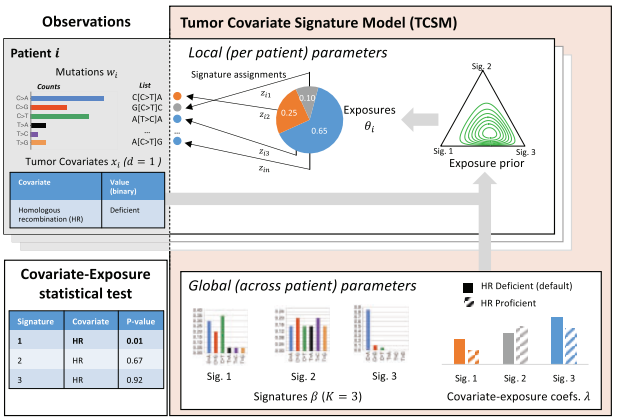

Fig. 1. Overview of the Tumor Covariate Signature Model (TCSM) with an illustrative example of $d = 1$ covariate and $K = 3$ signatures. For ease of illustration, the model shown is a simplified version of TCSM that does not model correlations between signatures. Given the observed mutations in a cohort of patients (top left), TCSM learns per patient exposures and assignments of each mutation to a signature (top right), and a global set of signatures and covariate–exposure coefficients (bottom right). The associations between covariates and exposures are then tested for statistical significance (bottom left). Parts of the design of the figure are inspired by Blei (2012) and Alexandrov *et al.* (2013b)

2.3 Imputing binary covariates in held-out samples

One advantage of TCSM is that it enables probabilistic imputation of held-out (or missing) covariates, including for previously unseen tumors. For example, for a single binary covariate in tumor x_{id}, we compute the log-likelihood ratio (LLR) of the tumor's mutations under the model with $x_{id} = 1$ and $x_{id} = 0$:

$$\text{LLR} = \log \frac{\Pr(w|x_{i1=1}, \beta, \Sigma, \Gamma, \cdot)}{\Pr(w|x_{i1=0}, \beta, \Sigma, \Gamma, \cdot)}, \tag{5}$$

where $\cdot$ is the hyperparameters of the model. A positive LLR indicates that the tumor's mutations are better fit when $x_{id} = 1$. After imputing held-out or missing covariates in this way, we then report the exposures θ estimated from the model with higher likelihood for downstream analysis.

2.4 Statistical significance of covariates on signature exposure

After applying variational EM to infer the latent variables of TCSM, we perform a statistical test for the significance of a covariate with respect to signature exposure. In this work, we only perform the test for a single binary covariate. For each signature k and binary covariate d, we generate 10 000 random exposures to signature k, half setting $x_d = 1$ and half setting $x_d = 0$, according to Equation (1). We then generate an empirical distribution by repeating these steps for TCSM trained on data where the covariates are permuted among samples uniformly at random. We compute a P-value for a signature-covariate pair by counting how often the mean differences in exposure of any signature-covariate pairs on the permuted datasets are greater than the mean difference of exposures on real data. We specifically test for an increase in exposure and only report the cases where the mean exposure when the covariate is present is greater than the mean exposure without the covariate; the parameterization of the Dirichlet (or Logistic Normal approximation) necessarily means that increasing the exposure of one signature will decrease the exposure of at least one of the others. We report Benjamini Hochberg-corrected P-values (Benjamini and Hochberg, 1995).

2.5 Benchmarking of mutation signature methods

It is challenging to compare mutation signature methods on real data because the true signatures and exposures are unknown. For that reason, we perform comparisons on both simulated and real data.

2.5.1 Simulated mutation datasets

We generate simulated mutation datasets from a simplified version of TCSM with known ground truth parameters and hyperparameters based on real cancer datasets and previous mutation signatures studies. The simulation process is simplified in that we do not allow correlations between signature exposures, so instead sample each tumor's exposures θ from a Dirichlet (as in Dirichlet-multinomial regression; Mimno and McCallum, 2008) instead of the logistic normal. As a case study, we generate data to reflect HR repair deficiency in breast cancer, using a single binary covariate. We use four of the validated COSMIC signatures (Forbes *et al.*, 2017) found to be active in breast cancer (Signatures 1, 2, 3 and 5). For each sample, we generate a single binary covariate x_i, representing HR deficiency, that increases the prior probability of exposure to Signature 3 (the COSMIC HR deficiency signature). We then generate θ_i of tumor i from a Dirichlet distribution with parameter vector $\eta_{ik} = \exp\{\lambda_{0,k} + \lambda_{1,k}x_i\}$. We use $\lambda_0 = [-2, -2, -5, -2]$ and $\lambda_1 = [0, 0, 4, 0]$. Thus, simulated tumors with HR deficiency have a much greater prior probability of high Signature 3 exposure, while the other signatures prior probabilities remain unchanged. We note that Signatures 3 and 5 have a high cosine similarity of 0.83 to each other, making it challenging to distinguish between Signature 3 mutations resulting from HR deficiency and Signature 5 mutations.

2.5.2 Evaluation methods

To quantify the importance of tumor covariates in modeling mutation signatures, we compare the TCSM with and without covariates. We also compare the models to NMF, using the popular SomaticSignatures implementation of NMF for mutation signature analysis (Gehring *et al.*, 2015).

Recovery of ground truth parameters. On simulated data, we compare the models on their learned signatures (using average cosine similarity) and exposures (using mean squared error). Note that these are in-sample comparisons.

Held-out log-likelihood. We compare TCSM with and without covariates using average log-likelihood per mutation of held-out data. Since NMF is non-probabilistic, we cannot compare it to TCSM using likelihood.

Prediction tasks using estimated exposures. To compare between probabilistic and non-probabilistic models, we compare the prediction power of the inferred exposures for a target binary covariate that is known to be associated with mutation signatures. First, we learn the mutation signature model on the training dataset. Then, we use the model to estimate the exposures of the test dataset to the identified signatures. Importantly, while the covariate is used when training TCSM, we hold it out completely in the testing dataset. For TCSM, we first impute the covariate in held-out samples before computing exposures (as described in Section 2.2). For NMF, we estimate the exposures in held-out samples using SignatureEstimation (Huang *et al.*, 2017). Next, a Support Vector Classification (SVC) model with a linear kernel is trained using the normalized exposures of the training dataset and the target covariate and evaluated on the test dataset. When the distribution of the target covariate is unbalanced, we set the class weight parameter of the SVC method to balanced and evaluate the performance using area under the precision-recall curve (AUPRC).

2.6 Implementation and software

We implemented TCSM in Python 3. We perform model training and inference using a wrapper of the Structural Topic Models R package (Roberts *et al.*, 2018). We provide a workflow for reproducing the experiments in the paper using Snakemake (Köster and Rahmann, 2012). The source code is publicly available at https://github.com/lrgr/tcsm.

2.7 Data

We analyze mutations in breast cancer exomes processed and standardized by The Cancer Genome Atlas PanCanAtlas (Hoadley *et al.*, 2018) and downloaded from the Genomic Data Commons (https://gdc.cancer.gov/about-data/publications/pancanatlas (8 May 2019, date last accessed)). To investigate the relationship between breast cancer and HR repair deficiency, we restrict our analysis to 760 tumors with called biallelic inactivations in 82 HR genes and counts of large-scale state transitions (LST; a measure of HR deficiency; Rieunier *et al.*, 2012) from Riaz *et al.* (2017). We obtain biallelic inactivation calls for the 82 HR genes by combining epigenetic silencing calls from Knijnenburg *et al.* (2018) with germline and somatic mutation and loss of heterozygosity (LOH) calls from Riaz *et al.* (2017).

We also analyze 466 melanoma exomes and 485 lung squamous cell carcinoma (LUSC) tumors from The Cancer Genome Atlas PanCanAtlas dataset (Hoadley *et al.*, 2018). We exclude 48 melanoma samples that were annotated as either acral melanomas or metastatic samples with unknown primary tumor origin by Trucco *et al.* (2018) (list of excluded samples obtained via personal correspondence). We download CC > TT dinucleotide polymorphism counts for these samples from both Firehose (https://doi.org/10.7908/C11G0KM9 (8 May 2019, date last accessed)) and (Alexandrov *et al.*, 2018). We combine these data sources by taking the average CC → TT count for samples that appear in both sources.

3 Results

3.1 Comparison on simulated data

We first compare the Tumor Covariate Signatures Model (TCSM) on simulated data with known ground truth to two baseline methods: NMF and TCSM using no covariates. To better understand how a single signature with changes in exposure due to tumor covariates affects the performance of TCSM and existing methods, we perform this comparison using simple simulated datasets with a single binary covariate that changes the prior probability of exposure for a single signature. The remaining parameters are set using previously discovered mutation signatures or are derived from real mutation datasets.

We randomly generate 50 simulated datasets (see Section 2.5.1), varying the number of samples from 50 to 250 and sampling with replacement the number of mutations per sample from real breast cancer exomes from The Cancer Genome Atlas PanCanAtlas dataset (Hoadley *et al.*, 2018). We then compare the output of our model to NMF as implemented by the SomaticSignatures R package (Gehring *et al.*, 2015). We apply TCSM with and without covariates to directly quantify the importance of incorporating tumor covariates. We evaluate the models in terms of the log-likelihood of held-out samples for $K = 2$–8. We compute the average held-out log-likelihood

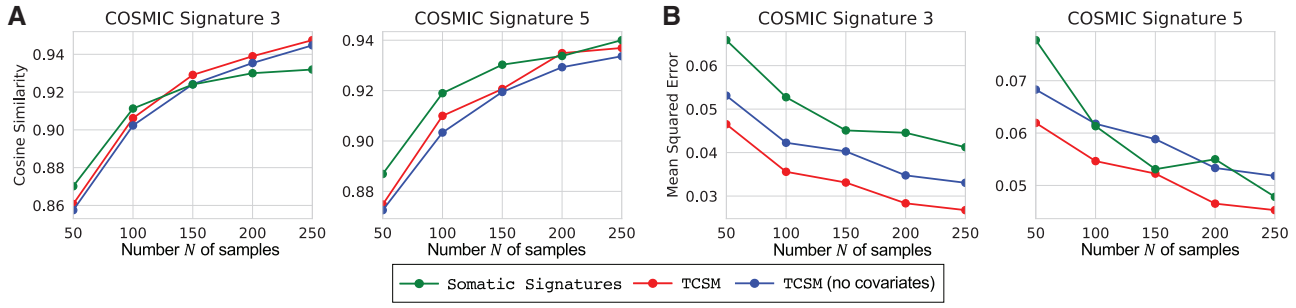

Fig. 2. Benchmark of TCSM with (red) and without (blue) covariates and NMF-based SomaticSignatures (green) on synthetic data. (**A**) Cosine similarity of inferred signatures (β) to hidden Signatures 3 and 5 using the true $K=4$ averaged across 50 datasets, varying the number of samples. (**B**) Mean-squared error of the inferred exposures (θ) for the same datasets as in (A)

using Monte Carlo cross-validation with 50 train/test splits, holding out 20% of the samples. We also report each model's in-sample accuracy at identifying the hidden signature and exposure parameters.

In terms of model selection (identifying the true K), we find that TCSM with covariates consistently outperforms TCSM without covariates and SomaticSignatures. While none of the models are able to consistently learn the true number of signatures ($K=4$) in datasets with only 50 samples, TCSM identifies the true K more often than the other methods (7/50 times compared with 2 and 1 for TCSM without covariates and SomaticSignatures, respectively) (We used the residual sum-of-squares and explained variances for model selection for SomaticSignatures, as suggested by the authors (Gehring *et al.*, 2015)). When we use 250 samples, we find that TCSM with covariates identifies the true number of signatures ($K=4$) for 35 of the datasets (compared with 3 and 19 for TCSM without covariates and SomaticSignatures, respectively). We also find that covariates provide additional signal, as TCSM with covariates achieves higher held-out likelihood than the TCSM without covariates on the majority of the synthetic datasets when $K=4$ for $N=50$ (28/50) and nearly all datasets when $N=250$ (49/50). All models identified the signatures with relatively high accuracy (cosine similarity >0.90; Fig. 2A) for $N>100$. However, TCSM with covariates was better able to distinguish between mutations caused by Signatures 3 and mutations caused by Signature 5, with higher accuracy in identifying the true exposures across all datasets (Fig. 2B).

3.2 HR repair deficiency in breast cancer

After establishing the utility of our model on simulated data, we turn to test it on real data. As an initial case study, we apply TCSM to study HR repair deficiency in breast cancer. Understanding HR deficiency in breast cancer is particularly important because of the clinical importance of identifying patients who might respond to PARP inhibitors (Farmer *et al.*, 2005). We use the TCGA BRCA cohort and divide the samples stratified by the biallelic HR covariate (described below) into (i) a training/validation dataset (75%) for choosing the encoding of the covariate, model selection and benchmarking TCSM with/without covariates; and (ii) a completely held-out test dataset (25%) for evaluation with a prediction task.

3.2.1 Covariate selection

The first key challenge in applying TCSM to real data is choosing the events or measures to use as covariates. Ideally, the covariates should be associated with changes in signature exposure and be easy to interpret biologically in order to reveal signature etiology. We begin by examining traditional markers of HR deficiency, including the biallelic inactivation of specific genes in the HR pathway (Riaz *et al.*, 2017) and the number of LST, which are chromosomal breakages that generate fragments of at least 10 Mb (Rieunier *et al.*, 2012).

We first compare TCSM using LST count to TCSM using biallelic inactivations of HR genes as covariates in terms of held-out log-likelihood for $K=2$–10 (Supplementary Fig. S1). We encode the biallelic inactivations as a single binary covariate where a 1 indicates the tumor has a biallelic inactivation in one of the seven genes (*ATM, BRCA1, BRCA2, CHEK2, FANCM, FANCF, RAD51C*) in the HR pathway inactivated in at least five samples in our cohort. We find that LST gives consistently better performance as measured in held-out log-likelihood, which makes intuitive sense as it is designed to be a direct readout of the functional status of the HR pathway. However, even though TCSM can use continuous variables as covariates, binary covariates—such as whether a gene has a biallelic inactivation—are more interpretable and easier to analyze downstream, e.g. when inferring the true value in a previously unseen sample. Therefore, we search for a subset of the HR genes whose biallelic inactivation maximizes the mutual information with the number of LSTs. More specifically, we use a greedy algorithm that adds the HR gene whose inactivations increase the mutual information with LST the most, halting when the mutual information stops increasing. The genes in the identified set, *BRCA1*, *BRCA2* and *RAD51C*, exhibit almost perfect mutual exclusivity (1/57 tumors have co-occurring mutations), a pattern expected for genes in the same pathway (Vandin *et al.*, 2012). Furthermore, TCSM trained using a single covariate for these three genes achieves superior performance than TCSM trained using a single covariate for all seven genes and nearly the same performance as TCSM using LST count as the covariate (Supplementary Fig. S1). In subsequent sections, we refer to TCSM with a single covariate—the biallelic inactivation of either *BRCA1*, *BRCA2* or *RAD51C*—as TCSM with the biallelic HR covariate.

3.2.2 Automated discovery of mutation signatures and etiology

After selecting the covariate to use, we perform model selection over the range $K=2$–10 using the TCSM with the biallelic HR covariate. We select $K=5$ as that is where the held-out log-likelihood plateaus, and show the resulting signatures in Supplementary Figure S3. All five signatures have cosine similarity >0.8 to COSMIC signatures with known etiologies (Alexandrov *et al.*, 2013a) (Supplementary Fig. S2); specifically, TCSM Signature 1 maps to the APOBEC signatures (COSMIC Signatures 2 and 13), TCSM Signature 2 maps to the HR deficiency signature (COSMIC Signature 3), TCSM Signature 3 maps to the polymerase epsilon signature (COSMIC Signature 10), TCSM Signature 4 maps to the mismatch repair

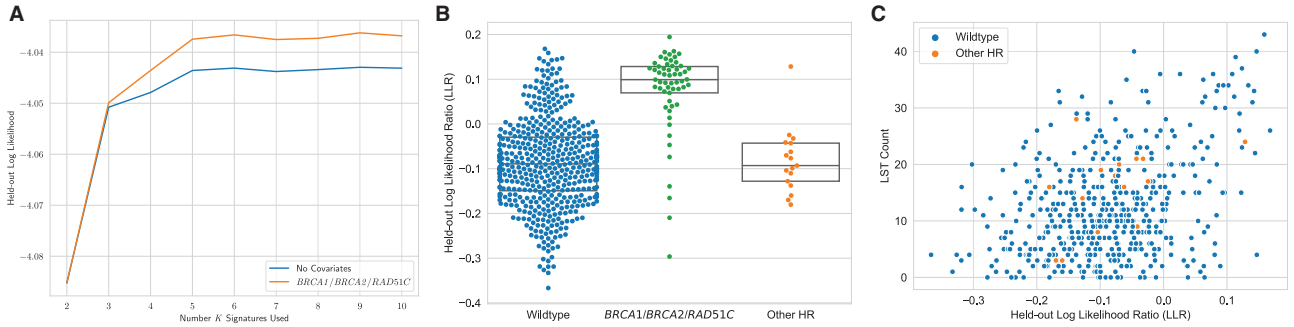

Fig. 3. (**A**) Comparison of the log-likelihood of held-out samples across $K = 2$–10 between TCSM with the biallelic HR covariate (inactivations of *BRCA1, BRCA2* or *RAD51C*) and TCSM without covariates. (**B**) The log-likelihood ratio (LLR) of samples with the biallelic HR covariate hidden where LLR>0 indicates the mutations of a sample are more likely under the biallelic HR covariate inactivation model. (**C**) After excluding tumors with known biallelic inactivations in *BRCA1, BRCA2* or *RAD51C*, the plot of a tumor's LLR against its LST count

(MMR) deficiency signature (COSMIC Signature 6) and TCSM Signature 5 maps to the aging signature (COSMIC Signature 1). Reassuringly, our covariate significance test identifies statistically significant increases in exposure to one TCSM signature, the TCSM signature that resembles COSMIC Signature 3, in the presence of the biallelic HR covariate (HR-proficient mean: 0.200, HR-deficient mean: 0.418, Benjamini Hochberg-corrected $P < 0.001$).

Next, we evaluate the ability of the TCSM to impute a hidden biallelic covariate value given a held-out tumor's mutations. We impute each tumor's biallelic covariate when it is in the test fold during 5-fold cross-validation. The log-likelihood ratio of tumors with known HR inactivations—including inactivations in the three HR genes used in training (orange) and four other HR genes (green)—is significantly greater than the ratio of the samples without known HR inactivations (blue; Fig. 3C; Wilcoxon rank-sum $P = 7e^{-22}$). Moreover, the tumors predicted to be HR deficient (i.e. those with LLR >0) without known HR inactivations have significantly more LSTs than the tumors predicted to be HR proficient (Wilcoxon rank sum $P = 8e^{-10}$; Fig. 3C), possibly indicating that they may have some form of HR deficiency due to some other event. Together, these results demonstrate the use of TCSM for automated discovery of mutation signatures and their etiology.

3.2.3 Comparison to other methods

We compare the performance of TCSM with the biallelic HR covariate to TCSM without covariates (Fig. 3A) for $K = 2$–10. We find that using covariates leads to an increase in held-out log-likelihood for all $K > 2$.

Next, we add NMF to the comparison. Since NMF is not probabilistic, we compare the estimated exposures of the three methods. We use the SomaticSignature R package implementation of NMF (Gehring *et al.*, 2015), using the SomaticSignatures model selection process. We choose $K = 5$ because the model selection yields a range from $K = 3$–6 (Supplementary Fig. S4) and $K = 5$ enables the most fair comparison between the models. The five signatures extracted by SomaticSignatures map with cosine similarity >0.8 to the same five COSMIC signatures as TCSM.

We compare how well the estimated exposures of each method for *held-out* tumors correspond with standard measures of HR deficiency. We train a linear model to classify tumor HR deficiency from the tumor's signature exposures. Davies *et al.* (2017) recently demonstrated the potential of a similar approach using NMF-based exposures to expand treatment with PARP inhibitors to a broader class of patients. As ground truth HR deficiency, we use biallelic inactivations in *BRCA1*, *BRCA2* or *RAD51C*. We then train the model on exposures from TCSM with the biallelic HR covariate, TCSM without covariates and SomaticSignatures (see Section 2.5.2 for details). To enable a fair comparison, TCSM is not provided with the true value for the biallelic HR covariate for the held-out tumors but instead infers the covariate value before estimating the exposure (see Section 2.3). We evaluate the models in terms of the AURPC on held-out cohorts not used when training the classifier.

We first compare within the cross-validation framework used for model selection. TCSM with the biallelic HR covariate (mean AUPRC = 0.62 across the 5-fold) outperforms both TCSM without covariates (mean AUPRC = 0.57) and the NMF approach (mean AUPRC = 0.56). We then compare on the completely held-out 25% samples not used for model selection or choosing the encoding for covariates. Again, we find that TCSM with the biallelic HR covariate (AUPRC = 0.64) outperforms both TCSM without covariates (AUPRC = 0.59) and the NMF approach (AUPRC = 0.58).

3.3 Simultaneously learning signatures in melanomas and lung cancers

Next, we investigate mutation signatures in cutaneous melanomas (SKCM) and lung squamous cell carcinomas (LUSC), two cancer types where mutational processes relating to environmental or lifestyle exposures are predominant. We examine whole-exome sequences of 418 SKCM and 485 LUSC tumors from TCGA PanCanAtlas (see Section 2.7 for details). One advantage of TCSM is the ability to encode cancer type in the model while performing a pan-cancer analysis. In contrast, previous work searched for a consensus set of signatures from a pan-cancer run and individual cancer type runs (Alexandrov *et al.*, 2013a, 2018).

We investigate using multiple covariates for TCSM: cancer type, smoking history (expected for many lung cancers) and exposure to UV radiation (expected for many melanomas). For cancer type, we use one binary covariate for SKCM and one binary covariate for LUSC. For smoking history, we set to one if the patient has a history of smoking and zero for never-smokers. Note that smoking history data are missing for SKCM patients, so we set their history of smoking covariates to zero. For UV radiation, we use the number of CC > TT mutations in the tumor, which has long been known as a marker of UV radiation exposure (Miller, 1985). Note that these dinucleotide mutations are excluded from the traditional 96 single base substitution categories analyzed by mutation signature methods, and are thus not included in the observations.

We first perform model selection using TCSM and compare the held-out log-likelihood using all four covariates (cancer type, smoking history and UV radiation exposure), using only the cancer type

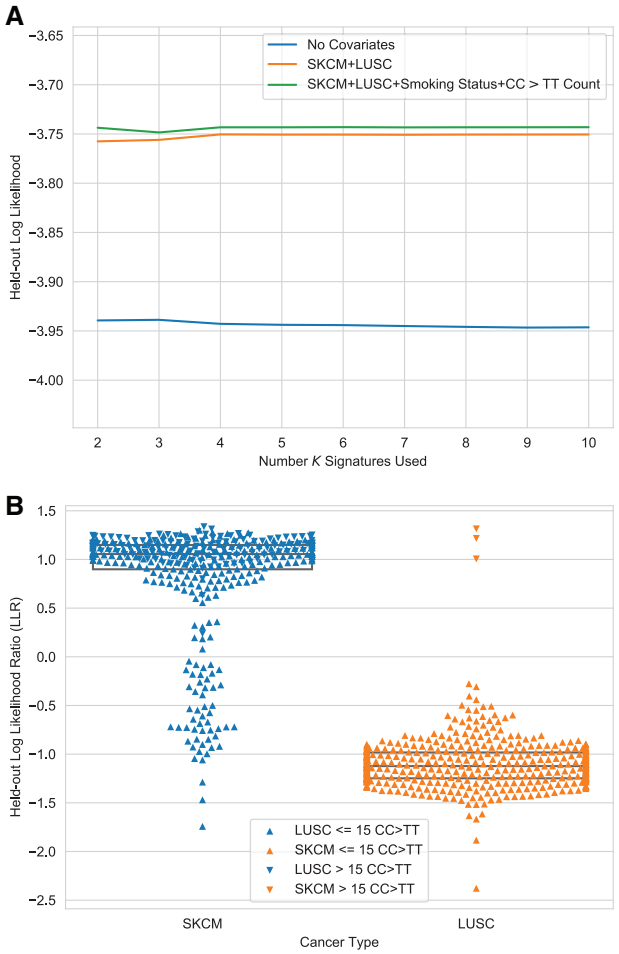

Fig. 4. (**A**) The heldout log-likelihood plot used for model selection to obtain $K = 4$. (**B**) The log-likelihood ratio (LLR) of the cancer type covariate for tumors where LLR <0 means the mutations of the tumor are more likely under LUSC and LLR >0 means the mutations of the tumor are more likely under SKCM

and using no covariates (Fig. 4A). We find that using the cancer type covariates results in a large improvement in held-out likelihood across K compared with using no covariates (Fig. 4A). In contrast, we find that using all four covariates results in a much smaller improvement in held-out likelihood across K compared with using only cancer type. We hypothesize that the additional covariates yield minimal improvement because they are strongly associated with the cancer type. To simplify downstream analysis, we remove the smoking status and UV radiation exposure covariates and use only the cancer type covariate. To further simplify the model, we use a single cancer type covariate with two possible values (LUSC and SKCM), instead of using one binary covariate for each cancer type as these two models have identical held-out likelihood performance (Supplementary Fig. S5). Using TCSM with the single cancer type covariate, we select $K = 4$ as the optimal number of signatures and show the resulting signatures in Supplementary Figure S7.

The four extracted signatures resemble known COSMIC signatures (Supplementary Fig. S6): the ultraviolet (UV) radiation-associated signature (Signature 7), the smoking-associated signature (COSMIC Signature 4), the APOBEC-associated signature (Signatures 2 and 13) and a signature that resembles both the aging-associated signature (Signature 1) and the mismatch repair deficient signature (Signature 6), which is likely a composite of the two COSMIC signatures that share a high cosine similarity to each other (cosine similarity = 0.84). Reassuringly, TCSM finds an association between the SKCM cancer type and an increase in the exposure to the TCSM signature most similar to COSMIC Signature 7 (LUSC mean: 0.113, SKCM mean: 0.808, Benjamini Hochberg-corrected $P < 0.001$). TCSM finds an association between the LUSC cancer type and an increase in the smoking signature (LUSC mean: 0.448, SKCM mean: 0.054, Benjamini Hochberg-corrected $P < 0.001$), the APOBEC signature (LUSC mean: 0.180, SKCM mean: 0.013, Benjamini Hochberg-corrected $P < 0.001$) and the mismatch repair/aging signature (LUSC mean: 0.260, SKCM mean: 0.125, Benjamini Hochberg-corrected $P < 0.001$).

We then investigate imputing a tumor's cancer type from its mutations. Campbell *et al.* (2016) examined 660 lung adenocarcinomas (LUAD) and 484 LUSC from TCGA and identified three LUSC tumors whose molecular profile resembled melanomas. They hypothesized that these three LUSC tumors might represent metastases from the skin and noted that one of these patients was previously diagnosed with basal cell carcinoma. Campbell *et al.* (2017) reported a related result in a targeted sequencing dataset, such that 35% of hypermutated lung cancers had high COSMIC Signature 7 exposure. Motivated by these reports, we use TCSM to re-examine the TCGA LUSC tumors to quantify the probability each primary tumor was correctly classified as LUSC.

We find that the cancer types imputed by TCSM are the same as the classified cancer type in the vast majority of cases (Fig. 4B). All but three LUSC have negative log-likelihood ratios, and the three outliers all have LLRs >1 (indicating that they strongly resemble melanomas). Indeed, these three outliers are the same as those Campbell *et al.* (2016) identified as having high UV radiation signature exposure. The number of CC > TT mutations in these tumors further supports the hypothesis that they are misclassified melanomas, as they are the only three tumors in the LUSC cohort with at least 15 CC > TT mutations (Fig. 4B). This analysis confirms and expands upon the conclusions of Campbell *et al.* (2016), and demonstrates the use of TCSM for probabilistically reasoning about cancer type classification.

TCSM identifies several SKCM tumors as likely LUSC (LLR > 0) that are less likely to be true misclassifications. One explanation is that SKCM tumors with LLR < 0 have very few mutations and almost no CC > TT mutations, especially when compared with SKCM tumors with LLR > 0 (mean number of mutations: 70 versus 1032, $P = 5e^{-27}$ Wilcoxon rank sum; mean number of CC > TT mutations: 0 versus 23, $P = 1e^{-27}$). However, many SKCM tumors with very few or no CC > TT mutations are still correctly classified as SKCM tumors, which demonstrates the importance of using the entire mutation spectrum, instead of a single feature.

4 Discussion

We presented the first probabilistic model, TCSM, of mutation signatures and their tumor-level clinical/demographic and molecular covariates. We found that TCSM outperformed NMF- and topic modeling-based approaches on both simulated and real mutation datasets, particularly in distinguishing between exposures of similar signatures. We then modeled mutation signatures of HR repair deficiency in breast cancers, demonstrating an approach for selecting interpretable covariates and predicting HR deficiency in held-out tumors. We also modeled mutation signatures in melanomas and lung cancers simultaneously. By including cancer type as a covariate, we were able to provide statistical support for earlier claims that

three lung cancers in our cohort from The Cancer Genome Atlas are misdiagnosed metastatic melanomas.

The key advantage of TCSM over existing methods is in inferring exposures, particularly in distinguishing exposures of similar signatures. For example, we found that a linear model trained on exposures from TCSM was better able to predict HR deficiency than linear models trained on exposures from methods that do not model covariates. While not the focus of the applications in this study, we hypothesize that by modeling the effects of tumor covariates on signature exposures, TCSM may be more sensitive than existing methods in discovering rare signatures. To do so may require explicit modeling of the number of mutations per tumor.

While modeling tumor covariates of mutation signatures brings clear advantages, it also raises the challenge of encoding and selecting covariates for the model. Encoding a particular covariate requires considering its sparseness and interpretability. Consider the covariate representing HR deficiency. We reasoned that biallelic inactivations in HR genes are more interpretable than existing HR indices—even if the HR indices may be a more direct encoding of the covariate—and that because each HR gene's inactivations are sparse and approximately mutually exclusive, they could be combined into a single event. Selecting covariates also brings challenges, particularly when the mutational processes active in a cohort are not well understood, there are multiple covariates related to the same process, there is population structure or batch effects correlated with exposure, or for discovering new signatures. In this case, it may be important to add a covariate selection component to the model.

Certain aspects of TCSM are computationally expensive and can be improved. For example, choosing the value of K, the number of signatures, requires multiple runs of TCSM for each potential value of K. One future extension is to model K as a draw from a Dirichlet Process, a version of which is popular for topic modeling (Teh *et al.*, 2005). Another computationally expensive step is our statistical test, which requires sampling 10 000 random exposures from the model because the mean of the logistic normal distribution is parameterized by a vector of $K-1$ coefficients, which does not lend itself to an easy interpretation of the significance of exposure–covariate associations. Substituting the Dirichlet distribution for the logistic normal distribution, such as in Mimno and McCallum (2008), would improve the direct interpretability of the parameters, which would enable a fully Bayesian approach for evaluating the significance of the exposure–covariate associations.

Finally, one direction we plan to explore in future work is modeling the effect of covariates on the signatures themselves, rather than their exposure. This is analogous to topic models of regional variation in language usage per topic (Eisenstein *et al.*, 2010, 2011; Roberts *et al.*, 2016b). There are multiple cases of researchers reporting multiple different signatures of the same mutational process, though it is not always clear what each of the distinct signatures represents. Learning how covariates change the signature themselves may help uncover these relationships.

Acknowledgements

M.D.M.L. gratefully acknowledges Jennifer Listgarten for first introducing him to the connection between mutation signatures and topic models. M.D.M.L. and W.R. gratefully acknowledge Jordan Boyd-Graber for discussions on topic modeling, Mark Keller for help in processing the TCGA datasets, and Mark Keller and Jason Fan for help with Figure 1.

Funding

This research was supported in part by the Intramural Research Program of the National Institutes of Health, NCI. W.R.'s contribution to this research was supported (in part) by NSF award DGE-1632976. R.S. was supported by Len Blavatnik and the Blavatnik Family foundation.

Conflict of interest: M.D.M.L. is a paid consultant for Microsoft.

References

Alexandrov,L. *et al.* (2018) The repertoire of mutational signatures in human cancer. *bioRxiv*, doi:10.1101/322859.

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

Alexandrov,L.B. *et al.* (2013) Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.*, **3**, 246–259.

Alexandrov,L.B. *et al.* (2015) Clock-like mutational processes in human somatic cells. *Nat. Genet.*, **47**, 1402–1407.

Alexandrov,L.B. *et al.* (2016) Mutational signatures associated with tobacco smoking in human cancer. *Science*, **354**, 618–622.

Benjamini,Y., and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, **57**, 289–300.

Blei,D.M. (2012) Probabilistic topic models. *Commun. ACM*, **55**, 77–84.

Blei,D.M., and Lafferty,J.D. (2005) Correlated topic models. In *Proceedings of the 18th International Conference on Neural Information Processing Systems, NIPS'05*, MIT Press, Cambridge, MA, USA, pp. 147–154.

Blei,D.M. *et al.* (2003) Latent Dirichlet allocation. *J. Mach. Learn. Res.*, **3**, 993–1022.

Campbell,B.B. *et al.* (2017) Comprehensive analysis of hypermutation in human cancer. *Cell*, doi:10.1016/j.cell. 2017.09.048.

Campbell,J.D. *et al.* (2016) Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.*, **48**, 607–616.

Chawanthayatham,S. *et al.* (2017) Mutational spectra of aflatoxin B1 in vivo establish biomarkers of exposure for human hepatocellular carcinoma. *Proc. Natl. Acad. Sci. USA*, doi:10.1073/pnas.1700759114.

Davies,H. *et al.* (2017) HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.*, **23**, 517–525.

Eisenstein,J. *et al.* (2010) A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP'10, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1277–1287.

Eisenstein,J. *et al.* (2011) Sparse additive generative models of text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, Madison, WI USA, Omnipress, pp. 1041–1048.

Farmer,H. *et al.* (2005) Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature*, **434**, 917–921.

Fischer,A. *et al.* (2013) EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol.*, **14**, 1–10.

Forbes,S.A. *et al.* (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.*, **45**, D777–D783.

Funnell,T. *et al.* (2018) Integrated single-nucleotide and structural variation signatures of DNA-repair deficient human cancers. *bioRxiv*, doi: 10.1101/267500.

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

Hegi,M.E. *et al.* (2005) MGMT gene silencing and benefit from temozolomide in glioblastoma. *New Eng. J. Med.*, **352**, 997–1003.

Hoadley,K.A. *et al.* (2018) Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, **173**, 291–304.e6.

Huang,X. *et al.* (2017) Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics (Oxford, England)*, **34**, 330–337.

Kasar,S. *et al.* (2015) Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.*, **6**, 8866.

Kim,J. *et al.* (2016) Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.*, **48**, 600–606.

Knijnenburg,T.A. *et al*. (2018) Genomic and molecular landscape of DNA damage repair deficiency across The Cancer Genome Atlas. *Cell Rep*., **23**, 239–254.e6.

Köster,J., and Rahmann,S. (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics (Oxford, England)*, **28**, 2520–2522.

Le,D.T. *et al*. (2017) Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science*, **357**, 409–413.

Miller,J.H. (1985) Mutagenic specificity of ultraviolet light. *J. Mol. Biol*., **182**, 45–65.

Mimno,D. and McCallum,A. (2008) Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, Helsinki, Finland, July 9–12, 2008. UAI'08, pp. 411–418.

Nik-Zainal,S. *et al*. (2016) Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, **534**, 47–54.

Pfeifer,G.P. (2010) Environmental exposures and mutational patterns of cancer genomes. *Genome Med*., **2**, 54.

Polak,P. *et al*. (2017) A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet*., **49**, doi:10.1038/ng.3934.

Ramage,D. *et al*. (2009) Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, August 6–7, 2009, EMNLP'09, pp. 248–256.

Riaz,N. *et al*. (2017) Pan-cancer analysis of bi-allelic alterations in homologous recombination DNA repair genes. *Nat. Commun*., **8**, doi: 10.1038/s41467-017-00921-w.

Rieunier,G. *et al*. (2012) Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Res*., **72**, 5454–5463.

Roberts,M.E. *et al*. (2018) *stm: R Package for Structural Topic Models*. http://www.structuraltopicmodel.com.

Roberts,M.E. *et al*. (2013) The structural topic model and applied social science. In *Proceedings of the Neural Information Processing Systems 2013 Workshop on Topic Models: Computation, Application, and Evaluation*, NIPS'13 Workshop on Topic Models.

Roberts,M. *et al*. (2016a) *Navigating the Local Modes of Big Data: The Case of Topic Models*, Cambridge University Press, New York, pp. 51–97.

Roberts,M.E. *et al*. (2016b) A model of text for experimentation in the social sciences. *J. Am. Stat. Assoc*., doi:10.1080/01621459.2016.1141684.

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

Rosales,R.A. *et al*. (2017) signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics (Oxford, England)*, **33**, 8–16.

Shiraishi,Y. *et al*. (2015) A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLoS Genet*., **11**, e1005657.

Szikriszt,B. *et al*. (2016) A comprehensive survey of the mutagenic impact of common cancer cytotoxics. *Genome Biol*., **17**, 99.

Teh,Y.W. *et al*. (2005) Sharing clusters among related groups: Hierarchical Dirichlet processes. In: *Advances in neural information processing systems*, pp. 1385–1392.

Tomasetti,C. *et al*. (2017) Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*, **355**, 1330–1334.

Trucco,L.D. *et al*. (2018) Ultraviolet radiation-induced DNA damage is prognostic for outcome in melanoma. *Nat. Med*., doi:10.1038/s41591-018-0265-6.

Tubbs,A., and Nussenzweig,A. (2017) Endogenous DNA damage as a source of genomic instability in cancer. *Cell*, **168**, 644–656.

Vandin,F. *et al*. (2012) De novo discovery of mutated driver pathways in cancer. *Genome Res*., **22**, 375–385.

Wallach,H.M. *et al*. (2009) Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML'09, ACM, New York, NY, USA, pp. 1105–1112.