Mutational Signatures: From Methods to Mechanisms

Yoo-Ah Kim,1,* Mark D.M. Leiserson,2,* Priya Moorjani,3,* Roded Sharan,4,* Damian Wojtowicz,1,* and Teresa M. Przytycka1,*

1National Center of Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA; email: przytyck@ncbi.nlm.nih.gov
2Department of Computer Science and Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland 20742, USA
3Department of Molecular and Cell Biology and Center for Computational Biology, University of California, Berkeley, California 94720, USA
4Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

Keywords
mutational signatures, mutagenic processes, cancer, cancer evolution, population genetics, computational models

Abstract
Mutations are the driving force of evolution, yet they underlie many diseases, in particular, cancer. They are thought to arise from a combination of stochastic errors in DNA processing, naturally occurring DNA damage (e.g., the spontaneous deamination of methylated CpG sites), replication errors, and dysregulation of DNA repair mechanisms. High-throughput sequencing has made it possible to generate large datasets to study mutational processes in health and disease. Since the emergence of the first mutational process studies in 2012, this field is gaining increasing attention and has already accumulated a host of computational approaches and biomedical applications.
1. INTRODUCTION

DNA molecules in our cells are targeted by diverse mutagenic processes. Such mutational processes can act in germ cells, contributing to species evolution (1), or in somatic cells, accumulating with age and contributing to diseases, especially cancer. Recent mutation rate studies of tumors have focussed on deciphering the somatically acquired changes in the DNA of cancer cells to advance our understanding of the relations among mutagenic exposures, DNA damage and repair, and outcomes (such as cancer and uncontrolled cell growth). Cancer genomes accumulate a large number of somatic mutations resulting from various endogenous and exogenous causes, including normal DNA damage and repair, cancer-related aberrations of the DNA maintenance machinery, and mutations triggered by carcinogenic exposures. Most mutations are typically harmless, but they provide a window into mutational processes, as different mutagenic processes result in characteristic mutational patterns in the genome, termed mutational signatures (2–4) (Figure 1). Identifying the mutagenic processes underlying the observed mutational signatures is an important step toward understanding tumor genesis and cancer evolution. Moreover, the understanding of mutational processes acting on a patient’s genome might also help to develop personalized therapies. For example, patients with homologous recombination deficiency (HRD) benefit from PARP [poly(ADP-ribose) polymerase] inhibitor therapy (5). At the same time, HRD leaves a

**Figure 1**

Mutational signatures and their computational analysis. (a) DNA mutations (yellow arrows) are caused by factors that are exogenous (environmental) or endogenous (caused by the natural stochasticity of biological processes and the dysregulation of specific molecular pathways). (b) Mutational signatures are inferred computationally from mutation counts across a set of individuals. Computational methods are used to infer signature activities in individual genomes and their dependencies on the sequential and genomic context. (c) To provide insights into etiology of endogenous signatures, researchers use computational and experimental methods to link mutational signatures to dysregulated genes and pathways. (d) Mutational signatures can be linked to signatures, cancer/genome evolution, and genetic diversity in a population.
characteristic mutational signature in the patient’s genome. Thus, the presence of this signature can be used as a marker for PARP inhibitor therapy (6). However, the etiologies of many signatures are not fully understood, and developing methods facilitating the association of signatures to potential causes has been a subject of intense study. Similarly, there is a growing understanding that the emergence of mutation patterns is often context specific, prompting studies directed to understanding this context dependence.

Access to steadily growing genomic datasets has stimulated the development of computational approaches to address the abovementioned questions. In the past decade, consortia such as The Cancer Genome Atlas (TCGA) (7) and the International Cancer Genome Consortium (ICGC) (8) have produced datasets of millions of somatic mutations from more than 35 cancer types. These datasets have enabled researchers to search for patterns of somatic mutations across thousands of tumors. Nik-Zainal et al. (2) and Alexandrov et al. (3, 4) were the first to model mutations observed in tumors as a mixture of hidden mutational signatures. Their efforts and subsequent work (9) have identified almost 49 validated mutational signatures (10), and mutational signature analysis has now become a standard component of cancer genome analysis pipelines.

Beyond cancer studies, analysis of the mutational signatures of healthy individuals (or non-disease cases) has been also very fruitful in understanding the mechanisms that play a role in embryogenesis, development, and evolution. In this regard, studies of de novo mutations (DNMs) and polymorphisms in humans have been particularly informative about the origin of mutations, their dependence on age and other factors, and heterogeneity in rates and patterns across individuals and species (11–17). However, large gaps remain in connecting exposures to outcomes and in evaluating the similarities and differences in the mutational landscape of the germline and soma.

2. COMPUTATIONAL INFERENCE OF MUTATIONAL SIGNATURES AND THEIR ACTIVITY

Mutational signatures are most commonly modeled as a set $P$ of signatures that are exposed at different frequencies across genomes (2, 4). In this model, each signature is represented as a multinomial distribution over a set of mutation categories. Most commonly, categories are formed based on the mutational change (six choices$^1$: $C > A$, $C > G$, $C > T$, $T > A$, $T > C$, and $T > G$) and the trinucleotide context in which the mutation occurs, yielding 96 mutation categories (e.g., TCC > TAC, CAG > CTG); sometimes an extended context is used, encompassing as many as seven bases on either side of the mutation, as this may explain larger variations in the mutation rate (18, 19). The proportion of different mutation categories is termed the mutation spectrum, and each individual’s genome is then represented as a linear combination of the signatures, where the number of mutations caused by a given signature is called its exposure.

Following the terminology of Omichessan et al. (20), researchers have focused on solving two broad classes of computational problems related to mutational signatures. In the de novo problem, the goal is to both infer the signatures and compute their exposures in the cohort. This was the initial research focus of seminal works in this area, and research on this problem continues apace. In the refitting problem, the goal is to infer the exposures of an existing set of signatures in a new cohort of individuals. The refitting problem became critical after the Catalogue of Somatic Mutations in Cancer (COSMIC) organization assembled an initial catalog of validated mutational signatures, and now refitting methods are arguably more widely used than de novo methods.

$^1$We follow the standard of referring to each substitution’s reference base as the pyrimidine of the base pair, although this also includes the corresponding variant type on the reverse complement strand.
In the rest of this section, we provide an overview of methods for both ideas, focusing on the most widely used approaches but also discussing active areas of research and open questions. We also identify key methodological commonalities and differences. For ease of exposition, we use the following notation throughout the section. We assume that the primary inputs are mutation counts of $m$ individuals across $n$ mutation categories, most commonly given in the $m \times n$ matrix $M$. We assume that the signatures matrix $P$ (which is either inferred or given) is a $k \times n$ matrix where each signature (row) is a probability distribution. We also assume that the exposures matrix $E$ is an $m \times k$ matrix. We note that in order to learn about the mutational processes some researchers compare mutation spectrum or mutation signatures across individuals.

2.1. Methods for Inferring Mutational Signatures De Novo

The standard methods for inferring mutational signatures de novo are easiest to understand as a typical latent variable inference problem. The observed variables are the mutation counts per patient, i.e., $M$. The latent variables (parameters) are the signatures $P$ (global, i.e., shared by all genomes) and the exposures $E$ (local, i.e., differing by the individual). Approaches for modeling $M$ have largely fallen into two camps. The original and most common approach is non-negative matrix factorization (NMF) (21, 22). More recently, researchers have begun exploring hierarchical probabilistic graphical models, in part because they allow the addition of observed and latent variables without making inference algorithms significantly more complicated.

2.1.1. Non-negative matrix factorizations. NMF methods for the de novo problem are the most widely used and come in several flavors. In its simplest form, NMF involves solving an optimization problem of minimizing the reconstruction error of observed matrix $M$ given inferred values for $P$ and $E$ and a divergence function $d$:

$$\arg\min_{E,P} d(M, E, P),$$

where $M$, $E$, and $P$ are all non-negative. While solving this optimization problem in its exact and approximate forms is NP-complete (23, 24) and there is no guarantee of a single optimal solution, there are several heuristics that seem to do well in practice. In particular, the multiplicative update method of Lee & Seung (21) is the most commonly used. Most forms of NMF have at least one hyperparameter, namely, the rank $k$ of the latent matrices $E$ and $P$.

NMF also admits different probabilistic interpretations or extensions. One advantage of the simple form of NMF is that it can be interpreted as a probabilistic model, depending on the choice of divergence $d$. In the case where $d$ is the Frobenius norm, minimizing the reconstruction error of $M$ is optimal assuming Gaussian noise (25). In the case where $d$ is the Kullback–Leibler divergence (KLD), minimizing the reconstruction error of $M$ is equivalent to finding the maximum likelihood solution where $M$ is Poisson distributed given $E$ and $P$ (26), which is a natural approach for count data. NMF can also be solved as a Bayesian inference problem where priors are placed on the latent variables or where the solution is constrained by regularization factors (e.g., for sparsity).

For the specific application to mutational signatures, a wide variety of NMF methods have been introduced. The original and one of the most commonly used methods is SigProfiler from Alexandrov et al. (4), which solves the problem in Equation 1 where the divergence is the Frobenius norm. More recently, SigProfiler has begun using the KLD divergence (9). The other most commonly used method is SignatureAnalyzer, first used in Kasar et al. (27) and introduced by Kim et al. (28). SignatureAnalyzer uses a Bayesian form of NMF called automatic relevance determination NMF, which, in addition to inferring the parameters $E$ and $P$, automatically infers the rank $k$.  

Kim et al.
of the latent matrices (29). Forms of these two methods were both used for identifying mutational signatures in ICGC’s Pan-Cancer Analysis of Whole Genomes (PCAWG) project, which are now reported in the COSMIC database, version 3 (9, 30).

There are other, less widely used methods for mutational signatures that employ different forms of NMF. Fischer et al. (31) introduced EMu, which uses a statistical form of NMF that is solved as a maximum likelihood problem. Rosales et al. (32) introduced signeR, a Bayesian NMF where the observed matrix $M$ is Poisson distributed with rates set by the latent matrices $E$ and $P$, which have Gamma priors, and a Markov chain Monte Carlo expectation-maximization algorithm is used for inference. Critically, the Bayesian form allows sampling from the posterior, both for data-driven selection of the number $k$ of signatures and in order to test the significance of differences in inferred parameters (e.g., whether two groups have significantly different exposure to a given signature). Both Fischer and Rosales also admit additional information in the form of the trinucleotide composition of each sequenced region (which differ for, e.g., whole-genome versus whole-exome studies), which can bias the inferred signatures. Covington et al. (33), Goncearenco et al. (34) and Ramazzotti et al. (35) each used forms of NMF that encourage sparsity in the signatures or the exposures.

2.1.2. Other hierarchical probabilistic graphical models. Researchers have also considered probabilistic graphical models with a different form than NMF for inferring mutational signatures de novo. The key difference comes in as an additional layer of hierarchy in the generative story: Instead of modeling counts, they model each individual mutation. In other words, each mutation has a latent variable that indicates which signature generated it. This additional resolution can be important for modeling phenomena that vary by mutation within the same individual’s genome. Another advantage of these models is that it is simple to add or expand the hierarchy in the generative process, either to change how signatures or mutations are modeled. A final advantage of these methods is that they can leverage decades of research in related fields, such as natural language processing. The field of topic model research is particularly relevant. In a classical topic model such as latent Dirichlet allocation (LDA) (36), a corpus of documents’ word counts are modeled as a combination of topics with per document activations. In the case of mutational signatures, the topics are signatures, the words are mutation categories, and the activations are exposures.

Such hierarchical probabilistic graphical models for inferring mutational signatures fall into four different categories, based on their purpose. In the first class is the first such model, pmsignatures, which was introduced in 2015 with the purpose of changing how signatures are represented to reduce the parameter explosion that would happen if researchers wanted to model a greater number of flanking bases per mutation (37). In the second class are methods that integrate additional observed data. Robinson et al. (38) adapted a structural topic model (39) to model associations between observed covariates (such as cancer type or DNA damage repair pathway inactivation status) and per patient exposures. Funnell et al. (40) adapted a multimodal, correlated topic model (41) to infer signatures and exposures for both single-base substitution and structural variation data. Additional methods that take genomic location data into account are detailed below in Section 5. In the third class are methods with an additional layer of hierarchy for distinguishing among groups of individuals with similar exposures. The models of Yang et al. (42) and Sason et al. (43) both fall into this category. Finally, the fourth class includes methods that use different optimization algorithms. For example, Matsutani et al. (44) used a variational Bayes form of LDA to better select the number $k$ of signatures active in a cohort.

2.1.3. Hyperparameter selection and other practical considerations. One challenge shared by nearly all de novo methods is hyperparameter selection. The most common hyperparameter
is the number $k$ of signatures. Approaches for inferring $k$ range from cross-validation (e.g., 38, 45) and using the Bayesian information criterion (e.g., 32) to Alexandrov et al.’s approach, which combines bootstrapping with a measure of stability and reconstruction error (4). Most commonly, researchers only search for $k$ within a relatively narrow range. Even methods such as Signature-Analyzer that automatically infer the rank have other hyperparameters that must be selected.

Another challenge is that the number of mutations per tumor can vary greatly and that this mutation rate and signature activity vary greatly by cancer type. Further, even if the mutation rate is fixed, the number of mutations reported varies by sequencing method, with generally 100 times more mutations in whole-genome samples than whole-exome samples. Consequently, a key challenge all the methods face is how to weigh individuals or population/cancer types such that the ones with the different mutation rates or active signatures do not overwhelm the signal of rarer signatures. Alexandrov et al. (3) took the approach of running their method on each cancer type individually and then all cancer types together, reporting a consensus of signatures across cancer types. Kim et al. (28) sought to address the high variance in mutation rate within endometrial cancer by splitting samples with extremely high mutation rates into multiple rows within the $M$ matrix. Versions of both of these approaches were used in the mutational signatures project of ICGC's PCAWG (9), where different NMF analyses were performed by cancer type, sequencing type, and hypermutator status.

### 2.2. Methods for Refitting Mutational Signatures

In the refitting approaches, it is assumed that the set of mutational signatures is given (matrix $P$), in addition to the count matrix $M$, and the goal is to infer the activity of each signature in every sample (exposure matrix $E$). The signature matrix can consist of either the full set of COSMIC signatures, a subset thereof, or signatures inferred from a specific cancer cohort using a de novo method described above. The refitting methods are especially useful when the analyzed set of mutations is too small for de novo signature inference (3), for example, in the case of small sample size, targeted sequencing panels, or samples with few mutations such as in healthy populations or in slowly growing tumors, or in the analysis of mutations located only in the specific genomic region of interest. This allows extending the applicability of validated mutational signatures in small targeted studies and even in clinical settings for individual patients.

There is a wide variety of refitting methods that have been introduced. Here, we present selected representative approaches; see also a review by Omichessan et al. (20). Rosenthal et al. (46) developed an approach called deconstructSigs, which determines a linear combination of the predefined signatures that best reconstructs the mutational profile of a single tumor sample. It is a heuristic method based on the iterative application of the multiple linear regression and removal of signatures with little exposure. With any decomposition problem, it is important to verify how stable the solution is and confidently establish which mutational signatures are present in a given sample. Huang et al. (47) addressed this problem from the perspective of input data perturbation and suboptimal solutions in a tool called SignatureEstimation. They showed that some mutational signatures, such as APOBEC (apolipoprotein B mRNA editing catalytic polypeptide-like) signatures, are very stable, while others, especially so-called flat signatures, are not. This emphasizes the importance of analyzing the confidence and stability of signature decomposition results. Li et al. (48) proposed a framework called SigLASSO that jointly optimizes the signature refitting and the likelihood of sampling and provides a sparse and high-confidence solution. Moreover, many of the de novo methods can be adapted for refitting. For example, SigProfiler offers a single-sample mode called SigProfilerSingleSample that identifies the activity of each predefined signature in
the sample and assigns the probability for each signature to cause a specific mutation type in the sample (4, 9). In the subsequent sections, we describe other refitting approaches presented in the context of different applications.

3. ASSOCIATION WITH GENOMIC FEATURES AND SEQUENCE CONTEXT

Methods for mutational signatures typically assume that even though the distribution of sites vulnerable to mutations, known as mutation opportunity, can vary along the genome due to genomic features like GC content, it is similar between different cancer genomes. Locally, mutation rates are shaped by genomic or epigenomic features, as well as specific properties of the DNA damage and repair mechanisms. In recent years, it has been shown that the activity of mutational processes along the genome can be influenced by large-scale features such as GC content (49), chromatin organization (50), transcription level, orientation (2, 51), and replication timing and direction (51–55), as well as local chromatin features such as transcription factor binding sites (56, 57), nucleosomes (58), gene structure (59), and noncanonical DNA motifs (60, 61), among others reviewed in References 62 and 63. For example, replication and transcriptional mutational asymmetries have been found for most signatures across different cancers (51, 54). In addition, the activity of the APOBEC enzyme family selectively deaminates single-stranded cytosines exposed on the lagging strand during DNA replication. Sometimes an activity of a mutational process is specifically localized; for example, ultraviolet (UV)-induced mutations are preferentially found in the DNA minor groove facing away from nucleosomes due to the abundance of UV-induced pyrimidine dimers, leading to the formation of CC > TT mutations at these sites (58). As another example, C > T mutations at CpG sites have a 10– to 20-fold-higher mutation rate due to the hypermutability of methylated CpG sites. However, CpG islands that are enriched for CpG sites tend to have a lower rate of CpG transitions, as most CpG sites in CpG islands are hypomethylated (64). All these global and local features are major determinants of mutational processes, but in most cases, such explanations remain to be established.

There have been many approaches to analyze context dependencies. The most straightforward solution relies on the partition of the observed mutations into categories of particular interest based on their genomic location or features (e.g., exome, promoter, CpG island, heterochromatin, repetitive region). Then each category of mutations can be analyzed separately using an NMF-based method by scaling the number of observed mutations to account for trinucleotide composition differences between the specific genomic category and the whole genome (46, 65). Alternatively, these mutational opportunities can be included directly into the statistical model, as in EMu (31) and signeR (32). Vöhinger, Gerstung, and colleagues proposed the TensorSignatures method (66), which allows for simultaneous inference of mutational signatures across different genomic features. RepairSig (67) adopted a similar approach, which helps to identify genomic determinants of DNA damage and repair processes. In another approach, Alexandrov et al. (3) expanded the set of mutational categories by incorporating the information on the transcriptional strand on which each mutation took place. This doubles the number of mutational categories because a mutation in a transcribed region, annotated as a pyrimidine base substitution, can be either on the transcribed or nontranscribed strand. Then the transcriptional strand-specific signatures can be extracted using the original signature inference method, SigProfiler. Such feature-specific signatures can be inferred in an analogous way for other features as well (54). Other approaches assign each individual mutation in a given sample a mutational process or signature that is most likely responsible for
causing the mutation (47, 51, 53). Then the dependency between mutational processes and their genomic context is studied based on the specific signature assignments and genomic features of the analyzed mutations.

Recently, it was observed that some signatures operate in sequential manner where consecutive mutations tend to be the result of the same mutation signature (2, 53, 68). Morganella et al. (53) identified groups of mutations of the same reference allele and on the same strand believed to come from the same signature and termed them processive groups. Supek & Lehner (68) performed a systematic analysis of clustered mutations and identified nine mutational signatures that are specifically linked to local increased mutation rates. SigMa (69) is a hidden Markov–based method that incorporates such properties of the mutational signatures into the model. It captures sequential dependencies between close-by mutations and allows for an accurate assignment of mutations to signatures. Following a similar reasoning, the StickySig method (43) accounts for the stickiness (i.e., the tendency of a certain signature to operate on consecutive mutations) and strand coordination of mutational signatures. It models consecutive, although not necessarily close-by, mutations that occur on the same strand, as well as independent mutations. In summary, mutational signatures have their origin in the interplay between the DNA damage caused by mutagenic agents and processes, DNA repair mechanisms, and global and local genomic features. For a given mutational process, different combinations of these factors can lead to varying mutation opportunities and drastically different mutation patterns among individual genomes. Consequently, the number of newly inferred mutational signatures is growing with the number of genomes being sequenced (9). To untangle the dependencies between mutational processes and genomic contexts they are acting in, researchers need new methods that go beyond the current paradigm of modeling mutational signatures (see Section 5).

4. LINKING MUTATIONAL SIGNATURES TO MOLECULAR CAUSES

While mutational signatures may arise due to environmental factors, some signatures are linked to genetic aberrations such as mutations or perturbed expression of DNA repair pathways. Both computational and experimental approaches have been utilized to identify such associations and shed light on the endogenous etiology.

Mutational signatures can accumulate due to the malfunction of DNA repair mechanisms when mutations in related pathways lead to genetic inactivation. For instance, the pattern of mutations attributed to Signature 3 is associated with biallelic inactivation of \( BRCA1 \) or \( BRCA2 \), two core homologous recombination (HR) genes (3, 70, 71). HR is a high-fidelity DNA repair mechanism for double-strand breaks (DSBs). Other HR related defects such as epigenetic silencing and somatic mutations in \( RAD51C \) can also yield characteristic mutational signature similar to Signature 3 (70).

Several other mutational signatures were found to be caused by genetic mutations. One study found that Signature 5 in urothelial tumors is significantly associated with somatic mutations in \( ERCC2 \), which is a member of the nucleotide excision repair pathway (72). Another example is the association found between Signature 18 and mutations in \( MUTYH \), a member of the base excision repair pathway (73). In addition, as shown in Section 5, a mutational signature can be shaped jointly by two different mutations. For example, Haradhvala et al. (74) showed that composite signatures arise from a concurrent loss of proofreading (\( POLE \) or \( POLD1 \)) and mismatch repair (MMR) function.

Kim et al. (71) studied the associations of mutational signature strength with genetic mutations using network-based approaches, investigating if a pathway inactivation due to genetic alterations can lead to the accumulation of specific mutational signatures. Utilizing a network-based
optimization algorithm named NETPHIX (75), they uncovered several subnetworks whose genetic alterations are associated with mutational signatures in breast cancer. In particular, they studied the differences between clustered and disperse APOBEC mutations. The proteins encoded by APOBEC gene family deaminate cytosines in single-stranded DNA (ssDNA). Such deamination, if not properly repaired, can lead to C > T or C > G mutations depending on how the resulting lesion is repaired. The strength of APOBEC signatures depends not only on availability of the enzyme but also on the presence of ssDNA. APOBEC signatures (Signatures 2 and 13) may arise as an immune response in cancer, and understanding the etiology is critical to understanding tumor progression (76, 77). Although both APOBEC signatures are known to be associated with APOBEC activities, several studies have reported that clustered and dispersed mutations may have different etiologies (68, 69, 76). Consistent with the previous studies, the network-based analysis found that dispersed mutations attributed to Signature 2 are associated with the alterations in a very different subnetwork than the remaining APOBEC-related signatures (71). Note that the cause-effect relationship can be in either direction—a mutation in a DNA repair gene can cause a specific mutational signature, or the mutagenic processes may generate uncontrolled mutations or cancer drivers (see Section 6). Interestingly, the subnetwork associated with dispersed Signature 2 includes PIK3CA mutations, which are considered as resulting from APOBEC-related mutational signatures. However, the association remains significant even after removing the patterns of APOBEC activities in the genomic region of PIK3CA, suggesting the possibility of opposite relationships.

Some of the computationally identified associations of genetic alterations in human tumors were also validated in experimental studies. The validations can be conducted via genetic manipulation techniques in various model systems (78). Using CRISPR (clustered regularly interspaced short palindromic repeats)-modified human stem cell organoids, Drost et al. (79) reproduced the mutational signatures driven by the deficiency of MMR gene MLH1 and the cancer predisposition gene NTHL1. Zou et al. (80) also recreated the mutational signatures observed in tumors by performing knockouts of several DNA repair genes in an isogenic human cell system. Furthermore, Volkova et al. (81) investigated the interplay of genotoxic exposure and DNA repair deficiency by a systematic screening of mutant Caenorhabditis elegans exposed to various genotoxic factors and characterized mutational patterns induced by environmental treatments. Their study experimentally demonstrated that mutational signatures are joint products of DNA damage and repair mechanisms.

Using putative causes as additional covariates in the model can help identify mutational signatures and the associations simultaneously. Robinson et al. (38) developed a probabilistic topic model-based method, named Tumor Covariate Signature Model (TCSM), to learn mutational signatures and automatically infer how observed covariates (such as DNA damage repair gene inactivations, cancer type, or demographic or lifestyle factors) are associated with signature exposure. Robinson et al. performed two proof-of-concept experiments. With a breast cancer dataset, they demonstrated that TCSM can be used to predict HR-deficient tumors and uncover the associated signature. In a lung cancer and melanoma dataset, they used TCSM to impute cancer type from observed mutations, finding supporting evidence for earlier studies that reported several TCGA lung cancers may be misdiagnosed metastatic melanomas.

While most studies have focused on the genetic aberration of DNA repair genes, some mutational signatures have also been linked with differential gene expression activities. For example, MGMT expression level (a DNA repair gene involved in cellular defense against mutagenesis and toxicity) may be associated with unique patterns of mutations, as MGMT silencing affects the direct repair mechanism via the gene (81, 82). Another example is the expression levels of APOBEC
family genes related to immune response activities, which are correlated with the accumulation of mutations attributed to APOBEC-related signatures (53, 68, 69).

To identify associations of mutational signatures with gene expression activities at a pathway level, Kim et al. (71) performed a correlation analysis and subsequently clustered the genes using consensus clustering. The analysis revealed several interesting network-level associations. In particular, the different patterns between two clock-like signatures, Signatures 1 and 5, were observed. The two signatures have been shown to correlate with the patient's age in many cancer types and thus are known as clock-like signatures (83). However, these two signatures are rarely correlated with each other, suggesting that they have distinct etiologies. Indeed, the aforementioned association analysis indicated that the magnitude of Signature 1 is positively correlated with the expression activity of cell cycle genes and, thus, corresponds to the biological clock. In contrast, Signature 5 shows correlation patterns consistent with continuous exposure to environmental mutagens such as reactive oxidative species (72, 83, 84). The mutations arising due to exogenous factors accumulate over time independent of cell cycle events. Grasping the etiologies of clock-like signatures can provide an important foundation for studying cancer evolution, as these provide a direct measure of the genomic timescale of exposure.

Linking mutational signatures to molecular features can help understand cancer etiology and develop personalized cancer therapies. However, due to the complex and dynamic nature of tumor evolution, untangling the cause and effect relationship can be challenging and requires further integrated and comprehensive analyses.

5. TOWARD DECONVOLUTING COMPLEX MULTIWAY RELATIONS BETWEEN MUTAGENIC FACTORS AND MUTATIONAL SIGNATURES

Traditional methods to infer mutational signatures assume that the signatures represent additive processes. However, there is a growing understanding that mutagenic processes are not necessarily additive (74, 81, 85). Instead, the mutational landscape of the cancer genome should be seen as the end effect of several interacting factors: the nature of DNA damage, the distribution of sites that are vulnerable to the damage, and potential deficiencies of the mechanism responsible for repairing the initial damage. In particular, it should be noted that DNA repair processes act by modifying the outcome of other mutagens. Under an additive model, different compositions of DNA damage and repair deficiencies must be modeled with different signatures to account for such dependencies. This can introduce a very large number of signatures and hamper their interpretability. For example, the current set of COSMIC signatures contains eight signatures associated with deficiency of MMR, a DNA repair process for recognizing and correcting mismatched nucleotides in complementary DNA strands. A recent study revealed that two of these signatures are in fact composite signatures, where two different types of DNA damage, caused by mutations in polymerases POLE and POLD1, are accompanied by MMR deficiency (74). This suggests that many other signatures, especially those known to be related to DNA repair deficiency, might also be composite. This recognition prompted the question of whether it is possible to decompose such complex signatures into their contributing factors (74).

As a first step in this direction, Wojtowicz et al. (85) introduced a new descriptor of mutational signatures called RePrint. RePrint takes as input a signature obtained via an additive model and, for each triple, computes the conditional probability of each of three possible mutations under the assumption that the triple is mutated. Specifically, recall that a mutational signature is a vector describing the probability distribution of mutation categories. In contrast, the RePrint of a signature is a vector of the same length but describing the conditional probability of each mutation category under the assumption that a mutation of the middle nucleotide in the specific triple
occurred. By the definition, conditional probabilities of the three possible mutations for each individual triple sum to one. Wojtowicz et al. showed that the similarity of RePrint signatures can indicate signatures that are likely to share common DNA repair deficiency mechanisms (85).

While RePrint provides a way to identify signatures that might share DNA repair deficiency mechanisms, the approach was not designed to provide a decomposition of composite signatures into contributing mutagenic factors. The first biologically realistic, nonadditive model to capture such decomposition is a recently proposed method called RepairSig (67). RepairSig explicitly models the composition of DNA damage processes and defective DNA repair processes. Its authors used the model to infer the signature of defective MMR process in breast cancer. The inferred signature was in good correspondence with the experimentally derived signature (80). In addition, by modeling the mutational landscape as a composition of DNA damage and repair, these authors have been able to use a single MMR signature to explain mutation data in TCGA (7) breast cancers, as opposed to several MMR deficiency signatures inferred by NMF-type models.

While additive models have provided many important insights, it is expected that new, more complex models will continue to emerge, providing more complete perspectives on mutational processes.

6. MUTATIONAL SIGNATURES AND CANCER EVOLUTION

The process of accumulating mutations is dynamic. Some types of mutations accumulate steadily over a lifetime, and some occur as a consequence of exogenous processes such as smoking and depend on the time and duration of the exposure. Yet other mutations emerge in response to cancer-related endogenous processes occurring in the cell, such as DNA repair deficiency or specific cancer-driver mutations (see Section 4). This has prompted the interest in studies on the dynamics of mutational processes across time and linking them to cancer evolution.

Mutations due to continuous exposure to mutagenic processes are expected to accumulate with age. In particular, one of the clock signatures, Signature 1 (discussed in Section 4), is assumed to be the result of spontaneous deamination of 5-methylcytosine, which can occur during replication, suggesting that the strength of this signature should reflect the number of past DNA replications (86, 87). Thus, Signature 1 could potentially be used as a clock for estimating the time of other mutagenic processes in cancer. However, the calibration of such a genetic clock remains challenging. Under the assumption that the strength of Signature 1 is related to the number of replications, mutations due to this signature would first accumulate with a rate dependent on the tissue renewal rate and then might accelerate during tumor growth. There are many unknown parameters in this process, including the time of the emergence of the tumor. Despite these obstacles, a recent study was able to utilize the basic principles behind this concept to estimate the timing of the whole-genome duplication events relative to the time of patient diagnosis (87).

The activities of mutational signatures can also change over time. Methods to infer such time dependencies typically rely on specific cancer-related reference events that can be used for ordering the mutations as occurring before or after the event. For example, a recent study used such reference events to divide mutations into multiple stages: early, late, and subclonal (87). The division between early and late is based on the relation to the whole-genome duplication event, as determined by assessing whether the mutation is present in both allelic copies. In contrast, the late/subclonal timing is based on the ability to make a distinction between subclonal and clonal mutations. Using such time partitioning, the study found, for example, that APOBEC mutagenesis tends to be higher in the late clonal stage than the early stage, while the exposures of signatures of defective MMR often increase from clonal to subclonal stages (87).

An alternative approach to study the dynamics of mutational signatures has been proposed in a recently developed method called TrackSig. TrackSig orders mutations by their inferred
population frequency—the cancer cell fraction (CCF). Given this ordering, it reconstructs the trajectories of signature exposures by identifying intervals of constant signature activities separated by computationally inferred change points (88). TrackSigFreq improves on TrackSig by additionally utilizing variant allele frequencies to obtain an optimal partition into segments of constant signature activities (89). Researchers have recently introduced two algorithms for analyzing mutational signatures and cancer evolution while moving beyond the approach from TrackSig. Abécassis et al. (90) introduced the first model, CloneSig, for simultaneously refitting mutational signatures and inferring the subclonal composition of the tumor and each subclone’s relative frequency. Christensen et al. (91) introduced a refitting algorithm, PhySigs, for inferring exposures in cancer phylogenies. Given an inferred phylogeny and set of active mutational signatures, PhySigs identifies which edges of the phylogeny included exposure shifts, thus inferring (subtrees of) clones with identical exposures. PhySigs has the advantage of not assuming a linear order of mutations, as mutations in different branches of a cancer phylogeny could have similar CCFs, although obtaining high-quality cancer phylogenies at scale remains a challenge. Christensen et al. analyzed one such dataset of high-quality phylogenies from Jamal-Hanjani et al. (92) and found evidence of exposure shifts in 20 out of 91 analyzed lung cancers (91).

Despite significant progress, the evolutionary dynamics of mutational signatures is still not fully understood. Challenges include the difficulty in inferring dynamics from typically static data and complex dependencies between mutagenic processes.

7. BEYOND CANCER: MUTATIONAL SIGNATURE ANALYSIS IN GERMLINE AND POLYMORPHISM DATASETS

In parallel to the study of somatic mutations in cancer, researchers are investigating mutational spectra in healthy populations to understand genetic diversity and genome evolution. In this regard, recent sequencing of thousands of genomes from family trios of Europeans and smaller surveys of non-European populations have identified genomic and nongenomic factors that impact the rate of new mutations, DNM (13, 14, 17, 93). These studies have shown (a) that the age of both the father and the mother are positively correlated with the number of DNMs in an offspring, with the effect size of paternal age being larger, and (b) that the parental age effects differ by mutation type (13, 14). In accordance, application of mutational signature analysis has shown that DNMs mainly comprise the two clock-like mutational signatures (Signatures 1 and 5) and represent the impact of aging on DNA (14, 94). Moreover, studies of parent-of-origin specific signatures have shown that as fathers age, the C > T mutations at CpG sites increase at a faster rate than other mutation types, and that increasing the mother’s age leads to more C > G mutations (13). The enrichment of paternal CpG transitions accords with the temporal dynamics of methylation in germ cells, consistent with the expectation that remethylation takes place early during embryogenesis in males, but very late (shortly before ovulation) in females (16, 95). Furthermore, the spatial distribution of maternal C > G mutations in genomic regions that also have elevated rates of noncrossover gene conversions highlights DSBs as an important source of these mutations in aging oocytes (13, 14, 16, 93).

By studying DNMs in a diverse set of human populations, Kessler et al. (17) showed that there is no significant difference in the DNM rate between individuals of different ancestries (European, African, and Latinos), although this is expected from small sample sizes, as the effects are likely to be subtle. A significant decrease was, however, observed in the proportion of C > A and T > C mutations in Amish individuals compared to Europeans, even after accounting for parental age and other confounding factors. Kessler et al. hypothesized that this might be due to the fact that Amish are exposed to fewer environmental mutagens, leading to lower rates of DNA damage and

200 Kim et al.
hence lower mutation rates (17). In sum, these analyses suggest an underappreciated role of DNA damage, in addition to replication, as the source of new mutations in the germline.

Comparisons of mutational spectra in polymorphism datasets have revealed a multitude of differences across human populations. The strongest signal detected in humans is the enrichment of TCC > TTC variants in Europeans and South Asians, relative to Africans or Asians (11, 12). Application of NMF suggests that this mutation is related to COSMIC Signature 11, which is also enriched in melanoma cancers and may be related to UV exposure. However, pyrimidine dimers leading to CC > TT mutations generated by UV radiation are not seen in Europeans (12, 96), and it remains unclear how UV exposure could impact the germline. Several other mutation types have also been shown to significantly differ among human populations; however, the magnitude of these effects is typically small (<20%). A recent analysis of Neanderthal ancestry segments recovered from 27,566 Europeans revealed differences in Neanderthal mutational spectra compared to modern humans, with a higher rate of C > G mutations and a lower rate of T > C and CpG > TpG mutations in introgressed segments compared to the nonintrogressed regions of the genome (97). DNM studies have shown that these mutations track parental age effects, highlighting a role of life history traits underlying some of the differences in the mutational spectra of modern humans and Neanderthals (97).

While there is emerging evidence of rapid evolution of mutational signatures across individuals and populations, there is no clear mechanistic explanation for the observed patterns. These differences could be due to several factors, such as demography (12); selection (in particular, biased gene conversion) (98), life history traits (such as mean age of reproduction) (13, 97, 99), environmental exposures (17), or even technical artifacts due to sequencing technologies (100). Moreover, like cancer and somatic studies, mutations in DNA polymerases or repair enzymes could in turn induce changes in the mutation rate or spectrum, acting as modifiers of mutation rate. Unlike cancer studies, it is difficult to directly measure historical exposures and relate the observed variation to the molecular mechanisms. DNM data from more diverse populations and, in particular, large sample sizes are further needed to assess the impact of various factors in contributing to de novo and population mutational differences.

8. CONCLUSION

Steadily increasing collections of genomic data provide an unprecedented opportunity for discovering and studying patterns of mutations across tumors and populations. These patterns have proven to be informative about mutagenic processes acting on genomes and, in some cancer-related cases, suggestive of potential interventions. The importance of understanding these processes has motivated the development of new computational methods to identify mutational signatures, the static and dynamic relations between them, their dependence on genomic context, and their relation to biological processes within cells, the environment, disease progression, aging, and evolution. Recent years have witnessed an explosion of new computational approaches and experimental studies, leading to steady progress in this area. However, many questions remain open, promising that computational studies of mutational patterns will continue to provide exciting results in coming years.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.
ACKNOWLEDGMENTS

This study was partially supported by the Intramural Research Programs of the National Library of Medicine of the US National Institutes of Health. P.M. and R.S. were supported by the Koret–UC Berkeley–Tel Aviv University Initiative in Computational Biology and Bioinformatics.

LITERATURE CITED

1. Pinto Y, Gabay O, Arbiza L, Sams AJ, Keinan A, Levanon EY. 2016. Clustered mutations in hominid genome evolution are consistent with APOBEC3G enzymatic activity. *Genome Res.* 26:579–87
2. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, et al. 2012. Mutational processes molding the genomes of 21 breast cancers. *Cell* 149:979–93
3. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio S, Behjati S, et al. 2013. Signatures of mutational processes in human cancer. *Nature* 500:415–21
4. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. 2013. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* 3:246–59
5. Bryant HE, Schultz N, Thomas HD, Parker KM, Flower D, et al. 2005. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature* 434:913–17
6. Davies H, Glodzik D, Morganella S, Yates LR, Staaf J, et al. 2017. HRDetect is a predictor of *BRCA1* and *BRCA2* deficiency based on mutational signatures. *Nat. Med.* 23:517–25
7. Cancer Genome Atlas Res. Netw. 2020. *Genomic data commons data portal.* Data Portal, Natl. Cancer Inst., Rockville, MD. https://portal.gdc.cancer.gov/
8. ICGC (Int. Cancer Genome Consort.). 2020. *ICGC dataportal.* Data Portal, Int. Cancer Genome Consort. https://dcc.icgc.org/
9. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, et al. 2020. The repertoire of mutational signatures in human cancer. *Nature* 578:94–101
10. COSMIC (Cat. Somat. Mutat. Cancer). 2020. *Mutational signatures (v3.1—June 2020).* Web Resour., Cat. Somat. Mutat. Cancer. https://cancer.sanger.ac.uk/cosmic/signatures
11. Harris K, Pritchard JK. 2017. Rapid evolution of the human mutation spectrum. *eLife* 6:e24284
12. Mathieson I, Reich D. 2017. Differences in the rare variant spectrum among human populations. *PLOS Genet.* 13:e1006581
13. Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, et al. 2017. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* 549:519–22
14. Goldmann J, Velman J, Gilissen C. 2019. De novo mutations reflect development and aging of the human germline. *Trends Genet.* 35:828–39
15. Chintalapati M, Moorjani P. 2020. Evolution of the mutation rate across primates. *Curr. Opin. Genet. Dev.* 62:58–64
16. Gao Z, Moorjani P, Sasani TA, Pedersen BS, Quinlan AR, et al. 2019. Overlooked roles of DNA damage and maternal age in generating human germline mutations. *PNAS* 116:9491–500
17. Kessler M, Loesch D, Perry J, Heard-Costae N, Talium D, et al. 2020. De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population. *PNAS* 117:2560–69
18. Aggarwala V, Voight BF. 2016. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat. Genet.* 48:349–55
19. Carlson J, Locke AE, Flickinger M, Zawistowski M, Levy S, et al. 2018. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nat. Commun.* 9:3753
20. Omichessan H, Severi G, Perduca V. 2019. Computational tools to detect signatures of mutational processes in DNA from tumours: a review and empirical comparison of performance. *PLOS ONE* 14:e0221235
21. Lee DD, Seung HS. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401:788–91
22. Lee D, Seung HS. 2001. Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing Systems, ed. T Leen, T Dietterich, V Tresp, pp. 556–62. https://papers.nips.cc/paper/2000/file/9d1152547c0bde01830b7e8bd6d0024c-Paper.pdf
23. Vavasis SA. 2009. On the complexity of nonnegative matrix factorization. SIAM J. Optimiz. 20(3):1364–77
24. Arora S, Ge R, Kannan R, Moitra A. 2012. Computing a nonnegative matrix factorization—provably. In Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of Computing, pp. 145–62. New York: Assoc. Comput. Mach.
25. Cichocki A, Zdunek R, Amari S. 2006. Csiszár’s divergences for non-negative matrix factorization: family of new algorithms. In Independent Component Analysis and Blind Signal Separation, ed. J Rosca, D Erdogmus, JC Principe, S Haykin, pp. 32–39. Berlin: Springer-Verlag
26. Cemgil A. 2009. Bayesian inference for nonnegative matrix factorisation models. Comput. Intell. Neurosci. 2009:785152
27. Kasar S, Kim J, Improgo R, Tiao G, Polak P, et al. 2015. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukemia evolution. Nat. Commun. 6:8866
28. Kim J, Mouw KW, Polak P, Braunstein LZ, Kamburov A, et al. 2016. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. Nat. Genet. 48:600–6
29. Tan YV, Fvotte C. 2013. Automatic relevance determination in nonnegative matrix factorization with the β-divergence. IEEE Trans. Pattern Anal. Mach. Intell. 35:1392–605
30. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, et al. 2017. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 45:D777–83
31. Fischer A, Illingworth CJ, Campbell PJ, Mustonen V. 2013. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. Genome Biol. 14:R39
32. Rosales RA, Drummond RD, Valieris R, Dias-Neto E, da Silva IT. 2016. signeR: an empirical Bayesian approach to mutational signature discovery. Bioinformatics 33:8–16
33. Covington K, Shinbrot E, Wheeler DA. 2016. Mutation signatures reveal biological processes in human cancer. bioRxiv 036541. https://doi.org/10.1101/036541
34. Goncêarenco A, Rager SL, Li M, Sang QX, Rogozin IB, Panchenko AR. 2017. Exploring background mutational processes to decipher cancer genetic heterogeneity. Nucleic Acids Res. 45:W514–22
35. Ramazzotti D, Lal A, Liu K, Tibshirani R, Sidow A. 2018. De novo mutational signature discovery in tumor genomes using SparseSignatures. bioRxiv 384834. https://doi.org/10.1101/384834
36. Blei DM, Ng AY, Jordan MI. 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3:993–1022
37. Shiraiishi Y, Tremmel G, Miyano S, Stephens M. 2015. A simple model-based approach to inferring and visualizing cancer mutation signatures. PLOS Genet. 11:e1005657
38. Robinson W, Sharan R, Leiserson MDM. 2019. Modeling clinical and molecular covariates of mutational process activity in cancer. Bioinformatics 35:492–500
39. Roberts ME, Stewart BM, Airoldi EM. 2016. A model of text for experimentation in the social sciences. J. Am. Stat. Assoc. 111(515):988–1003
40. Funnell T, Zhang AW, Grewal D, McKinney S, Bashashati A, et al. 2019. Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. PLOS Comp. Biol. 15:e1006799
41. Salomatin K, Yang Y, Lad A. 2009. Multi-field correlated topic modeling. In Proceedings of the 2009 International Conference on Data Mining, ed. C Apte, H Park, K Wang, MJ Zaki, pp. 628–37. Philadelphia: SIAM
42. Yang Z, Pandey P, Shibata D, Conti DV, Marjoram P, Siegmund KD. 2019. HiLDA: a statistical approach to investigate differences in mutational signatures. PeerJ 7:e7557
43. Sason I, Wojtowicz D, Robinson W, Leiserson MDM, Przytycka TM, Sharan R. 2020. A sticky multinomial mixture model of strand-coordinated mutational processes in cancer. iScience 23:100900
44. Matsutani T, Ueno Y, Fukunaga T, Hamada M. 2019. Discovering novel mutation signatures by latent Dirichlet allocation with variational Bayes inference. Bioinformatics 35:4543–52
45. Yang P, Liu K, Zhou J, Ke Y, Liu X, et al. 2018. CoMD: a computational model of mutational process differentiation in cancer. iScience 20:141–53
46. Yang Z, Pandey P, Shibata D, Conti DV, Marjoram P, Siegmund KD. 2019. HiLDA: a statistical approach to investigate differences in mutational signatures. PeerJ 7:e7557
47. Sason I, Wojtowicz D, Robinson W, Leiserson MDM, Przytycka TM, Sharan R. 2020. A sticky multinomial mixture model of strand-coordinated mutational processes in cancer. iScience 23:100900
48. Matsutani T, Ueno Y, Fukunaga T, Hamada M. 2019. Discovering novel mutation signatures by latent Dirichlet allocation with variational Bayes inference. Bioinformatics 35:4543–52
49. Yang P, Liu K, Zhou J, Ke Y, Liu X, et al. 2018. CoMD: a computational model of mutational process differentiation in cancer. iScience 20:141–53
50. Matsutani T, Ueno Y, Fukunaga T, Hamada M. 2019. Discovering novel mutation signatures by latent Dirichlet allocation with variational Bayes inference. Bioinformatics 35:4543–52
45. Gilad G, Sason I, Sharan R. 2020. An automated approach for determining the number of components in non-negative matrix factorization with application to mutational signature learning. *Mach. Learn. Sci. Technol.* 2:015013

46. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. 2016. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* 17:31

47. Huang X, Wojtowicz D, Przytycka TM. 2018. Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics* 34:330–37

48. Li S, Crawford FW, Gerstein MB. 2020. Using sigLASSO to optimize cancer mutation signatures jointly with sampling likelihood. *Nat. Commun.* 11:3575

49. Fryxell KJ, Moon WJ. 2005. CpG mutation rates in the human genome are highly dependent on local GC content. *Mol. Biol. Evol.* 22:650–58

50. Schuster-Bckler B, Lehner B. 2012. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* 488:504–7

51. Haradhvala N, Polak P, Stojanov P, Covington K, Shinbrot E, et al. 2016. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* 164:538–49

52. Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nat. Genet.* 41:393–95

53. Morganella S, Alexandrov LB, Glodzik D, Zou X, Davies H, et al. 2016. The topography of mutational processes in breast cancer genomes. *Nat. Commun.* 7:11383

54. Tomkova M, Tomek J, Kriaucionis S, Schuster-Bckler B. 2018. Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biol.* 19:129

55. Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, et al. 2012. Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.* 91:1033–40

56. Chan CWY, Gu Z, Bieg M, Eils R, Herrmann C. 2019. Impact of cancer mutational signatures on transcription factor motifs in the human genome. *BMCMed.Genom.* 12:64

57. Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. 2016. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* 532:264–67

58. Pich O, Muiños F, Sabarinathan R, Reyes-Salazar I, Gonzalez-Perez A, López-Bigas N. 2018. Somatic and germline mutation periodicity follow the orientation of the DNA minor groove around nucleosomes. *Cell* 175:1074–87

59. Frigola J, Sabarinathan R, Mularoni L, Muiños F, Gonzalez-Perez A, López-Bigas N. 2017. Reduced mutation rate in exons due to differential mismatch repair. *Nat. Genet.* 49:1684–92

60. Zou X, Morganella S, Glodzik D, Davies H, Li Y, et al. 2017. Short inverted repeats contribute to localized mutability in human somatic cells. *Nucleic Acids Res.* 45:11213–21

61. Georgakopoulos-Soares I, Morganella S, Jain N, Hemberg M, Nik-Zainal S. 2018. Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Res.* 28:1264–71

62. Gonzalez-Perez A, Sabarinathan R, López-Bigas N. 2019. Local determinants of the mutational landscape of the human genome. *Cell* 177:101–14

63. Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* 12:756–66

64. Takai D, Jones PA. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *PNAS* 99:3740–45

65. Gehring JS, Fischer B, Lawrence M, Huber W. 2015. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* 31:3673–75

66. Vöhringer H, van Hoeck A, Cuppen E, Gerstung M. 2020. Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. *bioRxiv* 850453. [https://doi.org/10.1101/850453](https://doi.org/10.1101/850453)

67. Wojtowicz D, Hoinka J, Angalan B, Kim YA, Przytycka TM. 2020. RepairSig: deconvolution of DNA damage and repair contributions to the mutational landscape of cancer. *bioRxiv* 2020.11.21.392878. [https://doi.org/10.1101/2020.11.21.392878](https://doi.org/10.1101/2020.11.21.392878)
68. Supek F, Lehner B. 2017. Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. Cell 170:534–47

69. Wojtowicz D, Sason I, Huang X, Kim YA, Leiserson MDM, et al. 2019. Hidden Markov models lead to higher resolution maps of mutation signature activity in cancer. Genome Med. 11:49

70. Polak P, Kim J, Braunstein LZ, Karlic R, Haradhvala NJ, et al. 2017. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. Nat. Genet. 49:1476–86

71. Kim YA, Wojtowicz D, Sarto Basso R, Sason I, Robinson W, et al. 2020. Network-based approaches elucidate differences within APOBEC and clock-like signatures in breast cancer. Genome Med. 11:49

72. Polak P, Kim J, Maruvka YE, Kim J, Rossbrock D, et al. 2018. Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. Nat. Commun. 9:1746

73. Viel A, Bruselles A, Meccia E, Fornasarig M, Quaia M, et al. 2017. A specific mutation signature associated with DNA 8-oxoguanine persistence in MUTYH-defective colorectal cancer. eBioMedicine 20:39–49

74. Haradhvala NJ, Kim J, Maruva YE, Polak P, Rosebrock D, et al. 2018. Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. Nat. Commun. 9:1746

75. Kim YA, Sarto Basso R, Wojtowicz D, Liu AS, Hochbaum DS, et al. 2020. Identifying drug sensitivity subnetworks with NETPHIX. iScience 23:101619

76. Swanton C, McGranahan N, Starrett G, Harris RS. 2015. APOBEC enzymes: mutagenic fuel for cancer evolution and heterogeneity. Cancer Discov. 5:704–12

77. Petljak M, Alexandrov LB, Brammeld JS, Price S, Wedge DC, et al. 2019. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. Cell 176:1282–94

78. Koh G, Zou X, Nik-Zainal S. 2020. Mutational signatures: experimental design and analytical framework. Genome Biol. 21:37

79. Drost J, van Boxtel R, Blokzijl F, Mizutani T, Sasaki N, et al. 2017. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. Science 358:234–38

80. Zou X, Owusu M, Harris R, Jackson SP, Loizou JI, Nik-Zainal S. 2018. Validating the concept of mutational signatures with isogenic cell models. Nat. Commun. 9:1744

81. Volkova NV, Meier B, Gonzalez-Huici V, Bertolini S, Gonzalez S, et al. 2020. Validating the concept of mutational signatures with isogenic cell models. Nat. Commun. 9:1744

82. Ma J, Setton J, Lee NY, Riaz N, Powell SN. 2018. The therapeutic significance of mutational signatures from DNA repair deficiency in cancer. Nat. Commun. 9:3292

83. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, et al. 2015. Clock-like mutational processes in human somatic cells. Nat. Genet. 47:1402–7

84. Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, et al. 2016. Mutational signatures associated with tobacco smoking in human cancer. Science 354:618–22

85. Wojtowicz D, Leiserson MDM, Sharan R, Przytycka TM. 2020. DNA repair footprint uncovers contribution of DNA repair mechanism to mutational signatures. PLoS Comput. Biol. 25:262–73

86. Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, et al. 2020. The evolutionary history of 2,658 cancers. Nature 578:122–28

87. Rubanova Y, Shi R, Harrigan CF, Li R, Wintersinger J, et al. 2020. Reconstructing evolutionary trajectories of mutation signature activities in cancer using TrackSig. Nat. Commun. 11:731

88. Harrigan CF, Rubanova Y, Morris Q, Selega A. 2020. TrackSigFREQ: subclonal reconstructions based on mutation signatures and allele frequencies. PLoS Comput. Biol. 25:238–49

89. Abcassis J, Reyal F, Vert JP. 2019. CloneSig: joint inference of intra-tumor heterogeneity and signature deconvolution in tumor bulk sequencing data. bioRxiv 825778. https://doi.org/10.1101/825778

90. Christensen S, Leiserson MDM, El-Kebir M. 2020. PhySigS: phylogenetic inference of mutational signature dynamics. PLoS Comput. Biol. 25:226–37

91. Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins T, et al. 2017. Tracking the evolution of non–small-cell lung cancer. N. Engl. J. Med. 376(22):2109–21
93. Halldorsson B, Palsson G, Stefansson O, Jonsson H, Hardarson M, et al. 2019. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* 363:eaau1043
94. Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, et al. 2016. Timing, rates and spectra of human germline mutation. *Nat. Genet.* 48:126–33
95. Reik W, Dean W, Walter J. 2001. Epigenetic reprogramming in mammalian development. *Science* 293:1089–93
96. Miller JH. 1985. Mutagenic specificity of ultraviolet light. *J. Mol. Biol.* 182:45–65
97. Skov L, Macià MC, Sveinbjörnsson G, Mafessoni F, Lucotte EA, et al. 2020. The nature of Neanderthal introgression revealed by 27,566 Icelandic genomes. *Nature* 582:78–83
98. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538:201–6
99. Agarwal I, Przeworski M. 2019. Signatures of replication timing, recombination, and sex in the spectrum of rare variants on the human X chromosome and autosomes. *PNAS* 116:17916–24
100. Anderson-Trocmé L, Farouni R, Bourgey M, Kamatani Y, Higasa K, et al. 2020. Legacy data confound genomics studies. *Mol. Biol. Evol.* 37:2–10