Oncogene (2020) 39:5031–5041
https://doi.org/10.1038/s41388-020-1343-z

ARTICLE

Decoding whole-genome mutational signatures in 37 human pan-cancers by denoising sparse autoencoder neural network

Guangsheng Pei[1] · Ruifeng Hu[1] · Yulin Dai[1] · Zhongming Zhao[1,2,3,4] · Peilin Jia[1]

Received: 2 December 2019 / Revised: 19 May 2020 / Accepted: 29 May 2020 / Published online: 11 June 2020

Abstract

Millions of somatic mutations have recently been discovered in cancer genomes. These mutations in cancer genomes occur due to internal and external mutagenesis forces. Decoding the mutational processes by examining their unique patterns has successfully revealed many known and novel signatures from whole exome data, but many still remain undiscovered. Here, we developed a deep learning approach, DeepMS, to decompose mutational signatures using 52,671,908 somatic mutations from 2780 highly curated cancer genomes with whole genome sequencing (WGS) in 37 cancer types/subtypes. With rigorous model training and comparison, we characterized 54 signatures for single base substitutions (SBSs), 11 for doublet base substitutions (DBSs) and 16 for small insertions and deletions (Indels). Compared to the previous methods, DeepMS could discover 37 SBS, 5 DBS, and 9 Indel new signatures, many of which represent associations with DNA mismatch or base excision repair and cisplatin therapy mechanisms. We further developed a regression-based model to estimate the correlation between signatures and clinical and demographical phenotypes. The first deep learning model DeepMS on WGS somatic mutational profiles enable us identify more comprehensive context-based mutational signatures than traditional NMF approaches. Our work substantially expands the landscape of the naturally occurring mutational signatures in cancer genomes, and provides new insights into cancer biology.

Introduction

Cancer is linked to the gradual accumulation of somatic mutations [1]. In cancer genomes, somatic mutations are induced by either the intrinsic infidelity of the DNA replication machinery or endogenous mutagen exposures, or both [2]. These processes may generate unique mutational signatures that can be characterized at single or multiple base substitution, small insertion and deletion, genome rearrangement, and copy number variation levels [3]. The observed somatic mutations in each cancer genome are the outcome of multiple mutational processes that have acted during the course as well as other factors including natural selection, drug response, and technical biases in sequencing, errors in variant call, among others. Mutational signatures can be detected by comprehensive analysis of sequence context of these mutations. Recently, many signatures have been identified that could be linked to specific processes [4]. These included several well-established signatures like those associated with tobacco carcinogens [5, 6], ultraviolet light exposure [7], or defective DNA mismatch repair system [8]. Deep investigation of these signatures would not only improve our understanding of the molecular

Supplementary information The online version of this article (https://doi.org/10.1038/s41388-020-1343-z) contains supplementary material, which is available to authorized users.

✉ Zhongming Zhao
zhongming.zhao@uth.tmc.edu

✉ Peilin Jia
peilin.jia@uth.tmc.edu

[1] Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[2] Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[3] MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, TX 77030, USA

[4] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, USA

mechanisms of cancer development but also provide important insights into cancer prevention and therapeutic treatment strategies (e.g., smoking in lung cancer, drug treatment) [9]. However, these signatures are typically discovered by whole exome sequencing (WES) data and considering only the two immediate nucleotides of the somatic mutations.

So far, there has been only limited effort on systematic analysis of sequence-context based mutational signatures, mainly because of the data limitation [2]. In the past several years, large-scale analyses of cancer genomic data across different cancer types (i.e., pan-cancer study) have discovered more than 30 recurrent base substitution patterns [2, 10, 11], but most of these mutational signatures were identified using WES data [2]. Most recently, highly curated somatic mutations from whole genome sequencing (WGS) of tumor samples were made publicly available by International Cancer Genome Consortium (ICGC) Pan-Cancer Analysis of Whole Genomes (PCAWG) working group [3]. These WGS-based somatic mutation datasets dramatically increase both the number of mutations and the coverage of the genomic regions [3]. And sequence-context based analysis of mutational signatures requires both coding (e.g., WES data) and noncoding regions (only available from WGS). Therefore, such new data provides us unprecedented opportunities to further uncover new mutational signatures, distinguish partially correlated signatures, and extract rare mutational signatures, among others [2]. So far, the majority of computational methods to discover mutational signatures has been based on the nonnegative matrix factorization (NMF) algorithm [4], which was first applied in breast cancer in 2013 [11]. Since then, more applications of NMF to cancer somatic mutations have been reported and convenient software packages have been developed, such as *SomaticSignatures* [12], *MutationalPatterns* [13], *MutSpec* [14], and *DeconstructSigs* [15]. All these packages implement the NMF method to conduct signature decomposition. However, NMF related methods often suffer from a high computational cost [16]. With the increase of exome-wide or whole genome-wide mutation data, applications using NMF related methods become more intensive, thus, requiring strong computing facilities. In addition, NMF implements an algorithm to decompose the input matrix as linear combinations of each individual's principal patterns. This is insufficient for capturing the nonlinear and complex inherent structure of somatic mutation profiles.

Recently, artificial neural networks have achieved breakthrough in mining the features from large, but complex, benchmark data sets, including image analysis [17] and natural language processing [16]. Auto-Encoder Neural Network is one of the deep learning methods for learning compact and efficient representations of input data in a nonlinear manner [18]. In this study, we developed a deep learning model based on Auto-Encoder structures (Fig. 1) to identify mutational signatures. We applied the method to the somatic mutations from ICGC PCAWG WGS dataset, currently the largest somatic mutation collection, to identify more comprehensive context-based mutational signatures. From the identified signatures, we linked them with known mutagens and evaluated the contributions of each signature to the spectrum of human neoplasia.

Results

Summary of mutation catalogs in ICGC PCAWG

The number of somatic mutations from the ICGC PCAWG WGS project is currently the largest in cancer research, making it ideal for mutation signature studies. We downloaded 48,276,930 somatic single base substitutions (SBSs), 426,648 doublet base substitutions (DBSs), and 3,968,330 small insertions and deletions (Indels), with the median number of 5260 SBSs, 25 DBSs, and 398 Indels per sample . Here, we called SBS, DBS, and Indel as mutation class. For each mutation class, its detailed mutation types are explained in "Methods" and Supplementary materials. Although most mutation patterns discovered from WGS data are quite similar to those from WES data [2, 3], the median mutation number is at least 50 folds higher than that from the WES data [median numbers were 83 SBSs, 0 DBS, and 4 Indels from The Cancer Genome Atlas (TCGA) project] [2]. We observed a substantial difference in the numbers of somatic mutations across samples and across cancer types (Supplementary Figs. S1, S2). We organized the data into three somatic mutation frequency matrices: $\mathbf{M}_{\mathrm{SBS}}$, $\mathbf{M}_{\mathrm{DBS}}$, and $\mathbf{M}_{\mathrm{Indel}}$. In order to systematically characterize the mutational signatures among different cancer types, we firstly performed a T-distributed Stochastic Neighbor Embedding (t-SNE) analysis [19] for the three mutation classes. As shown in Fig. 2, the mutation frequency matrices failed to distinguish cancer types clearly, except for a few types such as kidney cancer, liver cancer, and melanoma samples. The majority of samples in other cancer types were not well distinguished but rather located together. One possible reason is that the observed mutation frequency matrices were the joint results from multiple mutational processes imposing on each cancer genome. To distinguish these unique driving forces and their resultant mutational signatures, it is thus needed to decompose the mutation matrices.

DeepMS: a DSAE model to decode comprehensive mutational signatures

We named our DSAE model to decode mutational signatures as DeepMS (deep learning of mutational signatures)

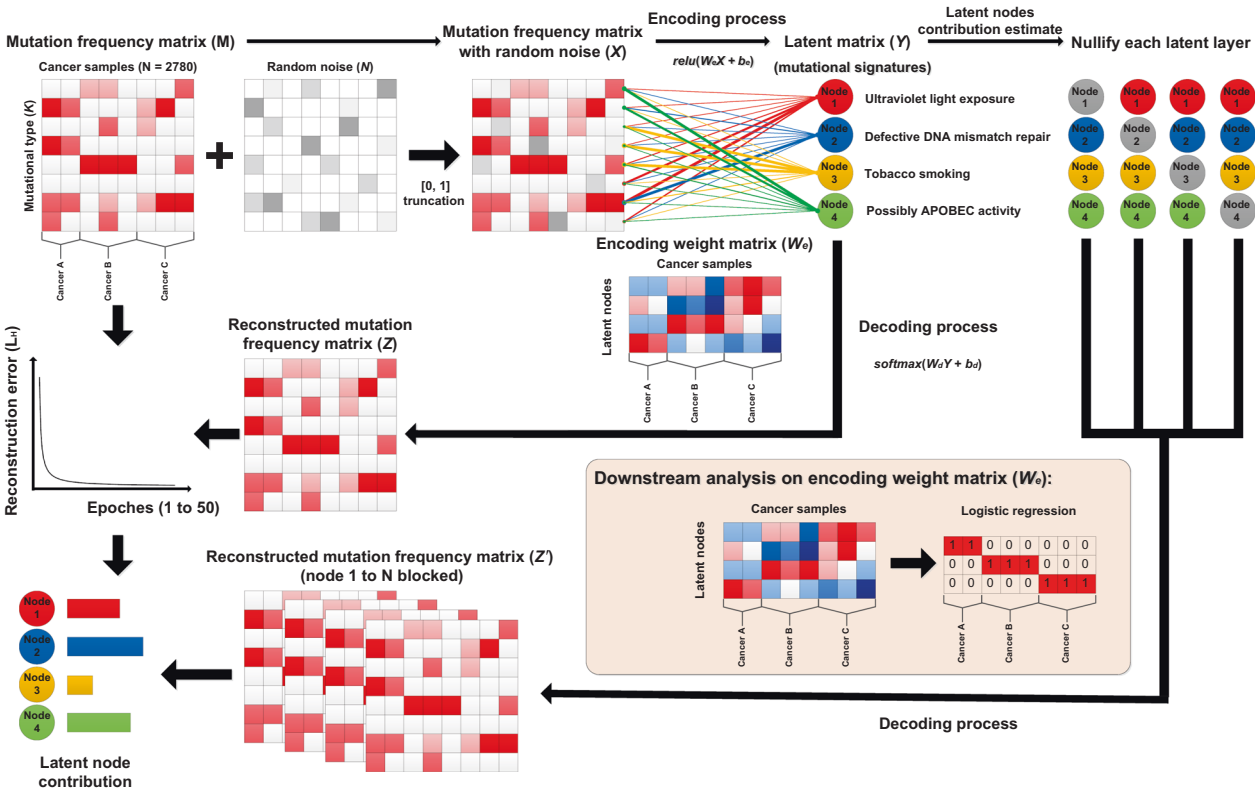

Fig. 1 Framework of Denoising Sparse Auto-Encoder (DSAE) model. Mutation frequency matrices of single base substitution (SBS), double base substitution (DBS), and insertion/deletion (Indel) combined with a random noise factor matrix were used as the input to the model, followed by encoding and decoding processes. For encoding step, we obtained two components: the latent matrix (latent node activities) that represents the compressed information of somatic mutational signatures, and the weight matrix W that reflects the contribution of each cancer sample to the activity of each latent node. For decoding step, the latent node activities derived from these mutation types are decoded back into reconstructed expression values through the decoding weight matrix W'. The reconstructed matrices were compared with the input to estimate the reconstruction error, which was subsequently used to evaluate and terminate the model fitting. For each mutation type, we trained a DSAE model by following this procedure, aiming to minimizing the difference between initial and reconstructed values. After the model training, we nullified each latent layer within the model by setting all output from that latent to zero; this process blocked information flowing through estimating the contribution of each latent layer.

(Fig. 1). The input to the DeepMS is the mutation frequency matrix (**M**) scaled by maximum mutation frequency with add-on noises (***N***), followed by [0, 1] truncation. We trained a DSAE model for each of the three mutation classes independently, with the input being $\boldsymbol{X}_{\mathrm{SBS}}$ (1536 SBS mutations × 2780 cancer genomes), $\boldsymbol{X}_{\mathrm{DBS}}$ (78 DBS mutations × 2780), and $\boldsymbol{X}_{\mathrm{Indel}}$ (84 Indels × 2780). We applied a parameter sweeping and used the reconstruction loss L_H to determine the best parameters to fit the models. This procedure suggested that the best latent layer dimension was 200 for the SBS matrix, 35 for the DBS matrix, and 42 for the Indel matrix, respectively. For example, as shown in Supplementary Table S1, the reconstruction loss L_H tended to be close to 0 at 50 epochs, indicating that our DSAE model could restore the original input data matrix very well. For each mutation class, we added a noise matrix to the input matrix. The noise matrix was generated from the standard normal distribution, followed by scaling to the range of [0,1] [20]. Using these parameters, we obtained two components from each of the trained DeepMS models (Supplementary Table S2). The component of the latent matrix was a compressed representation of somatic mutational signatures (SBS: 1536 × 200; DBS: 78 × 35; Indel: 84 × 42). These signatures can be further interpreted as biologically meaningful features. The component of the weight matrix connected the input matrix to the latent matrix (SBS: 200 × 2780; DBS: 35 × × 2780; Indel: 42 × 2780). This component can be interpreted as the contribution of cancer samples to each signature.

As expected, we found that both the latent matrix and the encoding weight matrix were informative for mutation signature discovery. For the latent matrix, each column could be considered as one candidate signature. For each signature, most values were close to 0 while a few formed "sharp" peaks (Supplementary Fig. S3), which resembled what genuine mutational signatures would look like in actual data and greatly reduced the probability of relatively featureless signatures [3]. Some latent vectors were highly

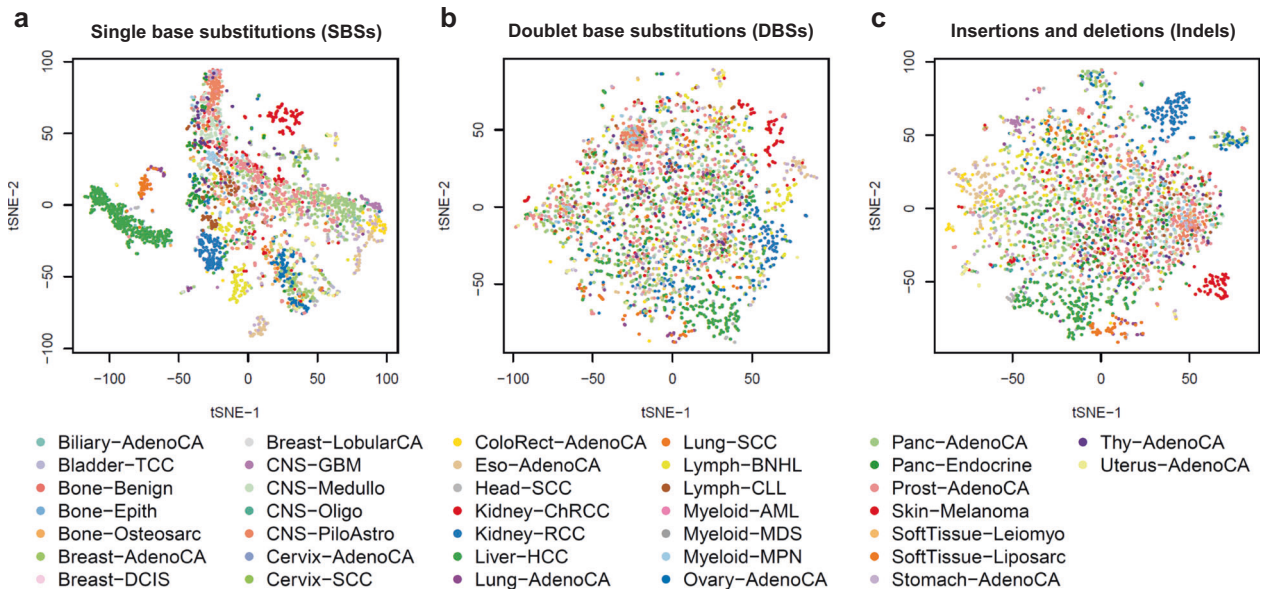

Fig. 2 t-SNE plots for the signatures of three somatic mutational classes. a SBS. **b** DBS. **c** Indel. Each dot represents a cancer sample. Color of dots indicates cancer types or subtypes.

correlated and were combined to represent one single mutation signature (see Supplementary Tables S2, S3). For the encoding weight matrix, the values represented the contribution of each sample to a latent vector. These values approximately formed a uniform distribution (Supplementary Fig. S3). Both matrices were used to examine the relationship between the discovered mutational signatures and clinical information.

SBS mutational signatures

The number of SBS mutations varied dramatically among samples and cancer types, ranging from hundreds to millions (Supplementary Fig. S1) [2, 21]. For the DSAE model for SBS, there were 200 latent vectors. After grouping those with cosine similarity (cos) ≥ 0.8 (Supplementary Fig. S4, Supplementary Table S3), we obtained a total of 54 SBS mutational signatures that were regarded independent (Fig. 3a and Supplementary Fig. S5). After comparing with the previous reports [3, 22], we found 17 out of these 54 signatures were reported before (cos ≥ 0.8) and the majority of them (16/17) were linked to known mutational processes (Fig. 3b). Hereafter, we adopted the format D-SBS-Sx (D for the method DeepMS; SBS: mutation class; Sx: signature number x; similarly D-DBS-Sy and D-Indel-Sz for DBS and Indel signatures) to refer the signatures found by our DeepMS models, like the format C-SBSx, C-DBSy, and C-IDz (ID represents Indel) used in Alexandrov et al. study [3]. For example, D-SBS-S1 (peak at TCT>A), D-SBS-S8 (peak at TCG>T), and D-SBS-S53 (peaks at TCN>T and TCT>A) had high similarity with two previous signatures (C-SBS10a and C-SBS10b), which were associated with putative polymerase epsilon defection [23, 24]. D-SBS-S2 (peaks at CC [A|C|T]>T and TCN>T) and D-SBS-S48 (peaks at CC [A|C|T]>T, and TCC>T) showed high similarity with previous signatures C-SBS7a and C-SBS7b that were associated with the mutations at two adjacent pyrimidines (thymines, TT, or cytosines, CC) likely resulted from ultraviolet light exposure [3]. D-SBS-S5, S12, S15, S25, S30, and S45 (peak at GCN>T) were associated with defective DNA mismatch repair [3].

Among 37 new signatures identified only by DeepMS (Fig. 3a), we found some of them could be represented by combinations of constituent signatures. For example, D-SBS-S4 could be split into two constituent signatures C-SBS13 (activation of APOBEC cytidine deaminase, peak at TC [A|T]>G) and C-SBS18 (reactive oxygen species, peak at NCA>A and TCT>A, where N can be any of the four nucleotides) [3]. D-SBS-S7 could be split into two constituent signatures: C-SBS36 (defective base excision repair due to *MUTYH* mutation) and C-SBS44 (defective DNA mismatch repair due to *MLH1* inactivation) [3, 25]. We speculated that these are likely resulted from the samples whose genomes were influenced by multiple factors [3]. In addition, a relatively featureless ("flat") signature, i.e., D-SBS-S3, was identified by our model. This signature was associated with failure of DNA double-strand break-repair by homologous recombination [10]. Among the remaining signatures, some could be explained with the relationship to disease. For example, D-SBS-S44 and D-SBS-S52 (peak at CC [C|T]>T) had 0.719 and 0.730 cos with C-SBS31, a signature associated with prior platinum compound

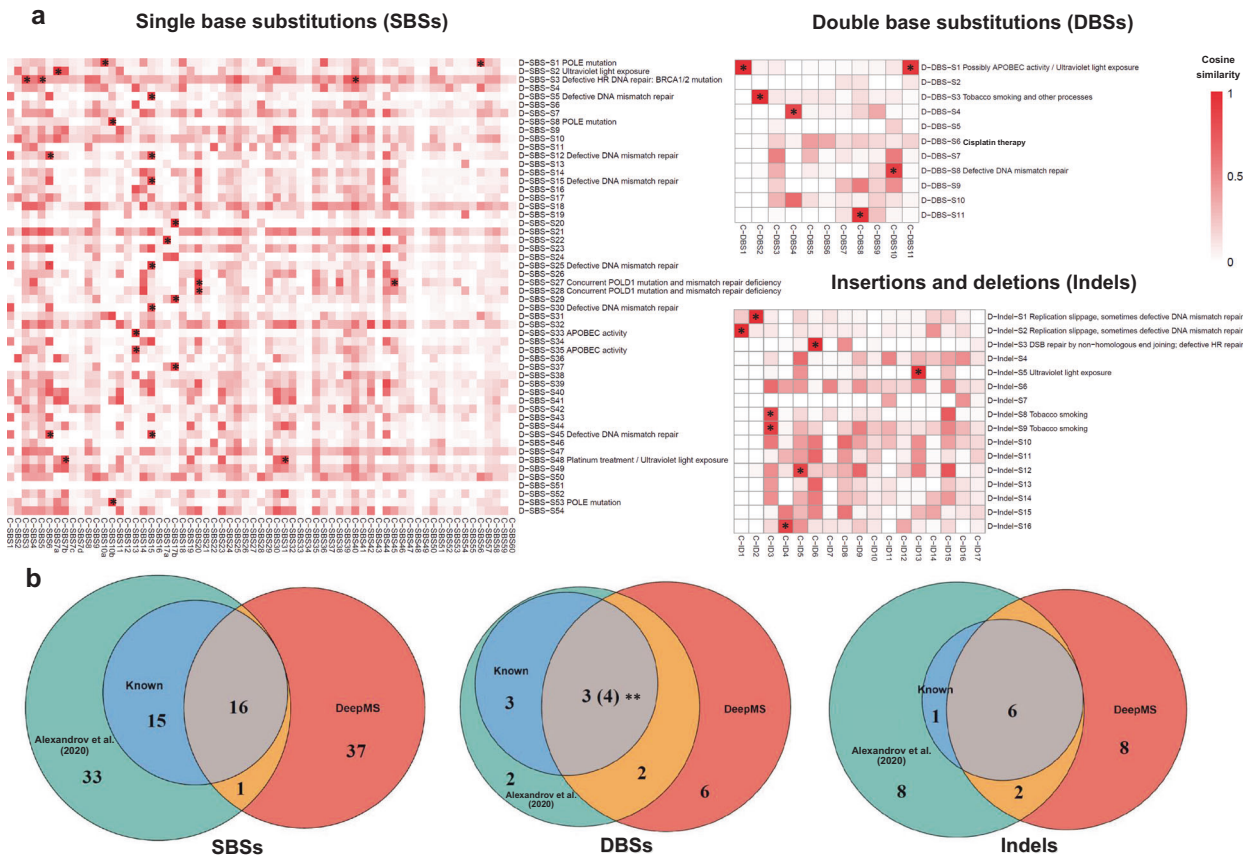

Fig. 3 Comparison of somatic mutational signatures with Alexandrov et al. study. a A heatmap showing the cosine similarity values between our mutatinal signatures (*y*-axis) and those that were previously reported [3] (*x*-axis). The color is proportional to the cosine similarity. Two signatures with cosine similarity ≥ 0.8 were labeled with *. **b** Venn diagrams showing the overlap of the detected mutatinal signature in this study with the Alexandrov et al. study [3]. **: D-DBS-S1 shows high similarity with two signatures (C-DBS1 and 11), so it was labeled "3(4)".

chemotherapy [26]. In another example, D-SBS-S4, D-SBS-S7, and D-SBS-S50 could be further split into a signature (peaks at NCT>A) associated with 8-oxoguanine exposure or base repair gene *MUTYH* mutation. Normally, the MUTYH DNA glycosylase can restrain 8-oxoguanine-related mutagenesis outcome by excising the incorporated adenine, while MUTYH mutations impair this enzymatic function [27, 28]. Furthermore, D-SBS-S21, D-SBS-S23, D-SBS-S32, and D-SBS-S39 (peaks at [C|G]AN>T and CCT>A) had 0.730–0.796 cos with C-SBS44, a signature associated with defective DNA mismatch repair due to *MLH1* inactivation [25]. However, there are many signatures with unknown causes, which were probably derived from random mistakes during normal DNA replication [29]. Mutation frequency in some specific DNA sequences may be higher than expected by chance (e.g., hotspots) [30]. In addition to the well-known CpG dinucleotide hotspot associated with the C>T mutation, there are other sequences with higher mutation rate, such as the CpHpG trinucleotide, where H stands for A, C or T [31], and the GTAAGT motif [32]. It was observed that a sequence of ±2 nucleotides around a mismatch site has an influence on the relative rates of SBSs and may lead to inherited disorders [33, 34]. Details in Supplementary information. However, such sequence context had much weaker effect than that of −1 and +1 bases [3, 11].

DBS mutational signatures

Tandem DBS and multiple base substitutions at immediately adjacent bases were observed with ~1% of the total SBS number [3], but the number of DBSs varied dramatically among cancer samples and cancer types (Supplementary Fig. S1) [3]. Our application of DeepMS to the DBS matrix discovered 35 latent vectors. After grouping those having cos ≥ 0.8 (Supplementary Fig. S4), we obtained 11 unique DBS mutational signatures (Supplementary Fig. S6). We further compared these signatures with those reported in previous studies [3] (Fig. 3b).

D-DBS-S1, which was highly similar to C-DBS1 (cos = 0.993), was characterized almost exclusively by CC > TT mutations. C-DBS1 is associated with ultraviolet light induced DNA damage and predominantly occurring in malignant melanomas [35]. D-DBS-S3 (C-DBS2, cos= 0.985) was composed of CC>AA mutations, with smaller number of CC>AG and CC>AT mutations. This signature is associated with tobacco smoking and mainly occurs in lung cancer [3]. D-DBS-S8 (C-DBS10, cos= 0.878) was composed of CG>TA and has been associated with defective DNA mismatch repair [3]. The remaining six signatures seem to be newly found, none of which had cosine similarity ≥ 0.8 with the previously reported signatures. For these signatures, we compared them with the compendium of environmental agents mutational signatures [22]. As a result, we found D-DBS-S6, which was featured with CT>TA and CT>AA, was associated with cisplatin therapy [22]. Interestingly, D-DBS-S6 also showed cosine similarity of 0.43 with C-DBS5, a signature associated with platinum treatment [3].

Indel mutational signatures

The number of Indels was within a small range in most cancer types although variation was observed. Some cancer types showed more deletions while others more insertions [3]. In most cancer genomes, the numbers of Indels were ~10% of the numbers of SBSs (Supplementary Fig. S1). The application of DeepMS to the Indel data unveiled 42 latent vectors representing candidate signatures. After grouping those that had cos ≥ 0.8, we had a total of 16 Indel mutational signatures (Supplementary Fig. S4). Eight of them have been previously reported [3], including six with known mutational processes (Fig. 3b).

D-Indel-S1 (C-ID2) and D-Indel-S2 (C-ID1) were mainly composed of deletions and insertions of thymine at long (≥5) thymine mononucleotide repeats and were likely due to DNA mismatch repair deficiency [3]. D-Indel-S3 (C-ID6) was characterized by deletions with ≥5 nt (Supplementary Fig. S7) and exhibited overlapping microhomology at the deletion boundaries with a mode of 2 nt and often longer stretches. These mutation patterns were attributed to defective homologous recombination repair [3, 36]. D-Indel-S5 (C-ID13, cosine similarity= 0.962) was characterized by deletions of thymine at thymine-thymine dinucleotides. Alexandrov et al. postulated ultraviolet perhaps predominantly induced thymine than cytosine dimers, although the underlying mechanism is unclear [37]. In addition, D-Indel-S8 and D-Indel-S9 (C-ID3, associated with tobacco smoking, cos= 0.801 and 0.870) were both characterized predominantly by short cytosine deletions (≤5 nt) with small member of short cytosine (≤5 nt) at mononucleotide cytosine repeats.

Associating mutational signatures with clinical variables

Somatic mutations found in cancer genomes are the consequence of a combination of multiple factors. Under the hypothesis that each mutational factor would leave its own characteristic mark on the genome [13], we fitted regression models to investigate the association between our discovered mutational signatures with clinical variables representing patient environmental exposure histories. We hypothesized that a signature having a stronger association with samples with a particular exposure would have a higher chance to be causally linked to that exposure.

We conducted the analyzes to the latent vectors (candidate signatures before merging) obtained for SBS, DBS, and Indel, respectively. In total, we obtained 15, 7, and 8 latent vectors with significant SBS, DBS and Indel association respectively (Fig. 4). For example, D-SBS-latent 32 was significantly associated with alcohol and tobacco smoking intensity and history; D-SBS-latent 42 and latent 117, were significantly associated with solar ultraviolet exposure. In addition, D-DBS-latent 29 (mainly CC > AA) were associated with alcohol history; D-DBS-latent 27 (CC > TT) was associated with ultraviolet exposure. These findings were consistent with both previous studies [3, 35] and our results described in the previous subsections.

Associating mutational signatures and cancer types

To examine the correlation between the number of mutations attributable to each signature and a specific cancer type, we further fitted a logistic regression model to estimate how samples of a given cancer type contributed to each signature. As shown in Fig. 5 and Supplementary Fig. S8, we observed a number of signatures had significant correlations with certain cancer types. For instance, D-SBS-S1 was significantly associated with colorectal adenocarcinomas, while D-SBS-S2 was significantly associated with skin melanoma. In addition, we observed D-DBS-S10 was significantly associated with head squamous cell carcinoma, and D-DBS-S8 was significantly associated with kidney renal cell carcinoma. Moreover, D-DBS-S1 and D-DBS-S3 might be weakly related to skin melanomaand lung squamous cell carcinoma, respectively. Simultaneously, D-Indel-S2 [mainly insertions of thymine at long (≥5) thymine mononucleotide repeats] was significantly associated with myelodysplastic syndromes, and D-Indel-S14 was significantly associated with head squamous cell carcinoma. Interestingly, we observed some signatures presented negative association pattern with cancer type. For example, bone osteosarc is negatively associated with D-SBS-S12, S35, and S44, indicating that these mutational signatures were significantly depleted from bone osteosarcoma

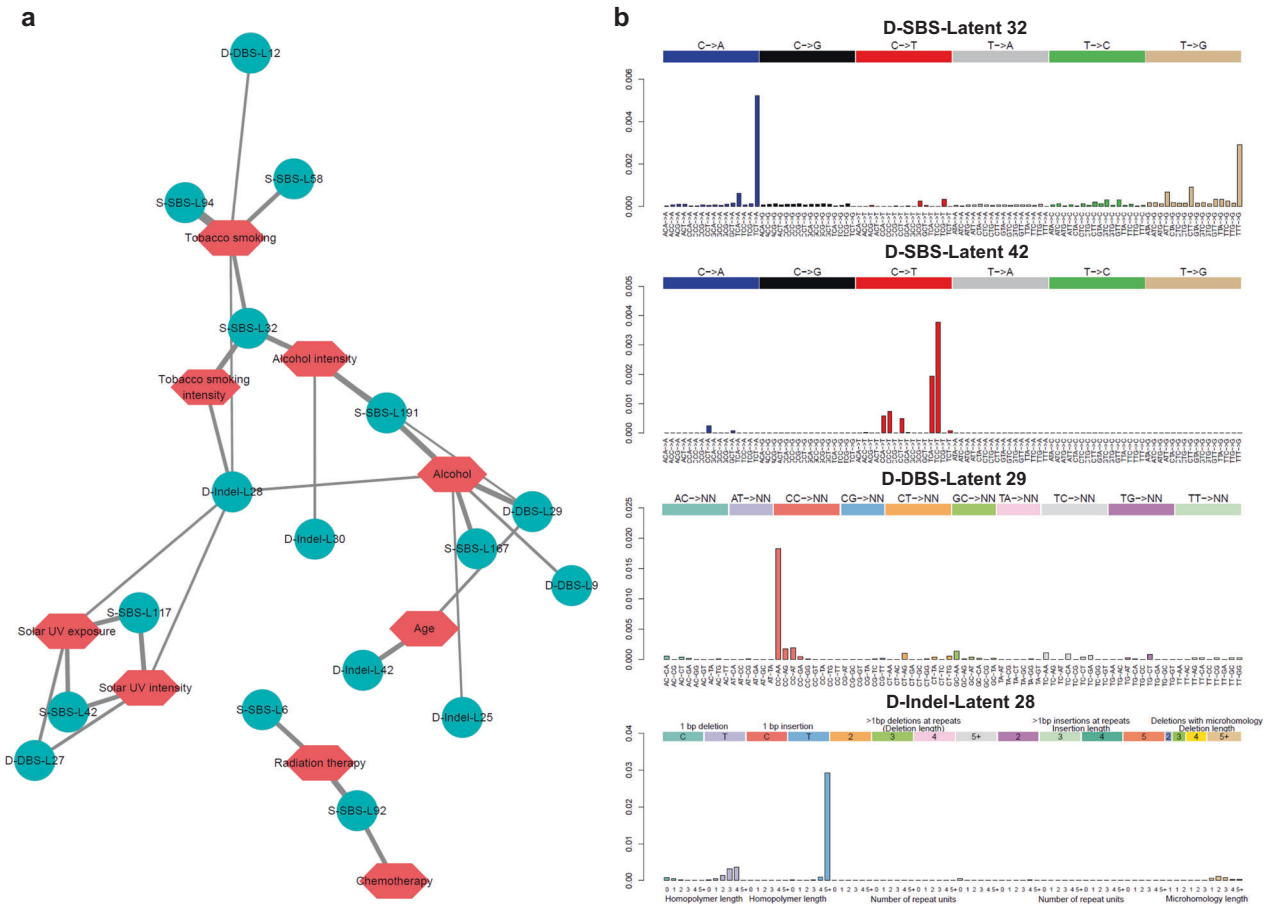

Fig. 4 Association between latent vector and patient environmental exposure based on regression models. a The network only shows the relationship between latent vector and clinical data with significant association [regression with $-\log_{10}$ (Benjamini and Hochberg [43] adjusted p value) < 0.05 for SBS mutations and <0.2 for DBS and Indel mutations]. The network layout is based on a force-directed graph. Edge width is proportional to $-\log_{10}$(adjusted p value). **b** Four specific latent vectors associated with enviromental exposure are highlighted with detailed visulization.

Discussion

Cancer is a genetic disease where somatic mutations contribute significantly to the tumorigenesis and progression, as well as clinical outcome such as drug response. In this work, we developed a constructive approach based on the Denoising Sparse Auto-Encoder Neural Network to decompose somatic mutation profiles from WGS of cancer samples in 37 cancer types/subtypes. Our proposed DSAE approach is superior to the traditional NMF approaches by two features: the denoising feature enables the intermediate representation of signatures to be robust for the small number of random variants in the input samples [18, 38], and the sparse feature allows appropriate signature representation [39]. By applying DSAE to matrices of different mutation classes derived from the 52,671,908 somatic mutations covering 2780 cancer whole genomes in 37 cancer types/subtypes, the comprehensive mutation signature catalogs we discovered represent the most comprehensive landscape of mutational signatures in cancer, and will provide new insights into understanding mutational processes linking to various environmental and genetic factors in cancer.

Our primary goal was to use a non-linear model to identify mutational signatures from cancer mutational profiles. As validation, many mutational signatures from our work across three mutation classes were close to previous NMF approaches [3]: 30 (17 + 5 + 8) out of 81 (54 SBS + 11 DBS + 16 Indel) of our signatures showed high similarity with those identified by NMF approaches [3]. Interestingly, we found 26 (16 + 4 + 6) of them were associated with known mutational processes or other factors. The high annotation frequency (26/30, 86.7%) implies that DeepMS can identify genuine common signatures and can dig out major information with high reliability. Among the remaining newly identified mutational signatures, some showed moderate level of association with previous constituent signatures. However, there are other signatures without significant correlation with any clinical information available in this study. One possible reason is the lack of the

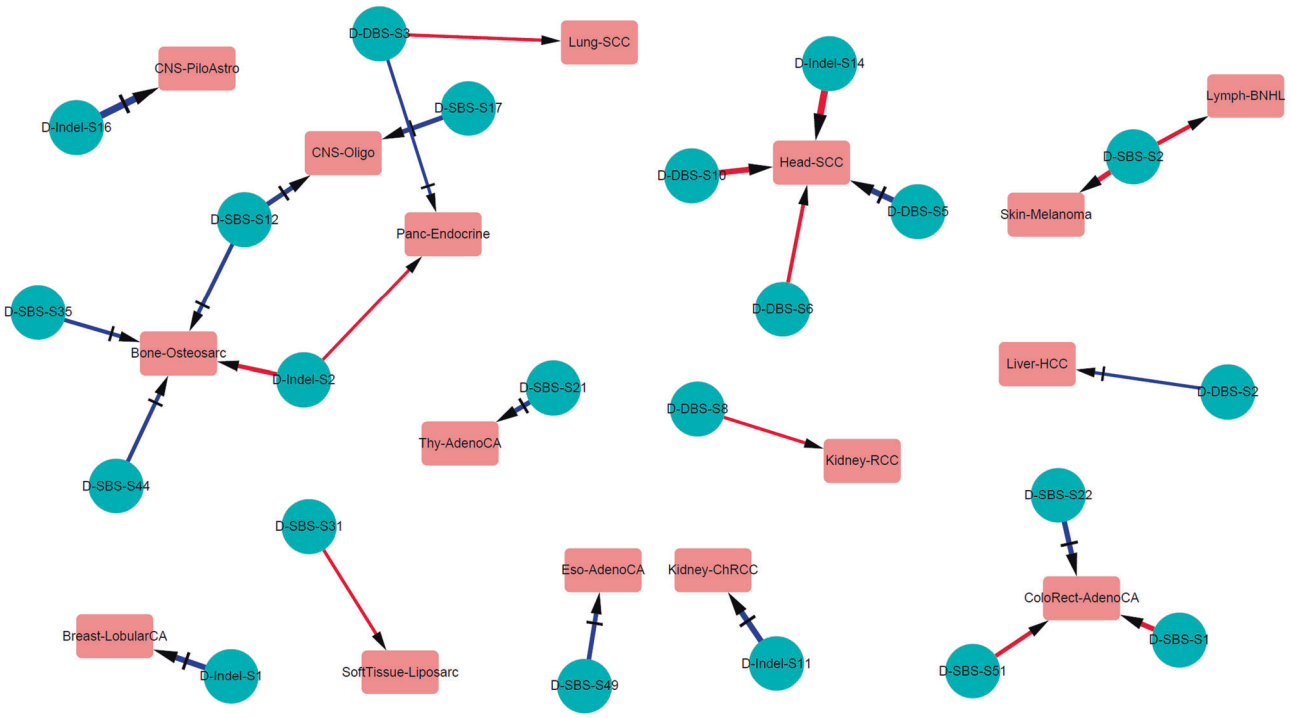

Fig. 5 Association between mutational signatures and 37 cancer types/subtypes through logistic regression models. The network shows the relationship between mutational signatures and cancer types/subtypes with significant association [regression with $-\log_{10}$ (Benjamini and Hochberg [43] adjusted p value) < 0.2]. A circle node represents a mutational signature and a rectangle node represents a cancer type or subtype. A red edge indicates an enrichment pattern while a blue edge indicates a depeted pattern.

necessary clinical data, such as mutagen exposure, previous medicine and treatment history, diet habit, and other deleterious environmental factors. Another reason is that the contributions from environmental factors are likely not strong compared to the accumulation of random errors during normal DNA replication [29]. Nevertheless, this study represents a novel deep learning approach to assess the contribution from each signature to the burden of mutational catalogs of individual cancer using the most comprehensive cancer data at the WGS level. Our regression models identified a number of associations between mutational signatures and cancer types, and these findings were consistent with previous studies [3, 35], supporting our approach to be reliable. Importantly, our DeepMS approach revealed a number of associations that have never been reported before. These new signatures will expand our knowledge in cancer biology.

Our work has several limitations. First, despite that the median number of mutations from current WGS (PCAWG: 5740) was much larger than that from previous WES data (TCGA: 89), most of these mutations were located in the noncoding regions and lacked strand bias information [40]. Therefore, all mutational signatures in this study had limitation of without considering strand specificity. Second, we optimized different dimension of latent nodes in the hidden layer to minimize the reconstruction error. However, in this situation, some latent nodes were not independent. We had to merge similar latent nodes to representative signature, thereby bringing difficulty in evaluating each signature's contribution. Despite these challenges, our work provides a complementary way to first time utilize deep learning model to study somatic mutational signatures at large-scale, and at WGS level (both coding and non-coding). With the accumulation of data and improvement of artificial intelligence technology, neural networks will become a promising strategy to discover novel somatic mutational signatures with more WGS data being released in future.

Taken together, we presented the first deep learning approach to explore the mutational signatures from the largest ever somatic mutation dataset: whole genome sequencing of 2780 samples in 37 cancer types/subtypes. Our method could effectively capture non-linear relationships among mutation patterns. It provides an alternative, but powerful, approach to the traditional methods based on nonnegative matrix factorization. We identified a total of 54 SBS, 11 DBS and 16 Indel mutational signatures. These results included both the previously reported ones ($n = 30$) and novel signatures ($n = 51$), representing a substantially expanded landscape of the naturally occurring mutational signatures in pan-cancer genomes. Moreover, the regression-based approaches allowed us to quantitatively link the mutational signatures with environmental exposures

and cancer types, supporting the practice of precision medicine. These associations between signatures and clinical/demographical phenotypes are also potential indicators for cancer prevention and therapeutic treatment strategies.

Materials and methods

Pan-cancer whole genome somatic mutation data

Genome-wide somatic mutation data was generated and curated by the ICGC PCAWG group. We downloaded the data from Synapse (syn11726620, November 8, 2018). In total, there were 52,671,908 somatic mutations from 2780 cancer genomes covering 37 cancer types/subtypes. The list of cancer names and related information is summarized in Supplementary Table S4. Throughout this work, we refer single base substitutions (SBSs), doublet base substitutions (DBSs), and small insertions and deletions (Indels) as mutation classes and each single nucleotide change as a mutation type.

Preparation of mutation profiles

To obtain the input matrix for our model, we first built a mutation frequency matrix for each of the three mutation classes: SBS, DBS and Indel. Details can be found in Supplementary information and Supplementary Fig. S9. Overall, we included 1536, 78, and 84 mutation types for the SBS DBS, and Indel class, respectively [3]. We constructed three mutation frequency matrices: $\mathbf{M}_{\mathrm{SBS}}$, $\mathbf{M}_{\mathrm{DBS}}$, and $\mathbf{M}_{\mathrm{Indel}}$. Each matrix was formatted as mutation types on rows and samples on columns, i.e., $\mathbf{M} = \{m_{ij}\}$, $i = 1,\ldots K$, $j = 1,\ldots N$, where m_{ij} represented the frequency of mutation type i in sample j, K was the total number of mutation types ($K_{\mathrm{SBS}} = 1536$, $K_{\mathrm{DBS}} = 78$, $K_{\mathrm{Indel}} = 84$), and N was sample size ($N = 2780$). Due to the difference of mutation frequency among the three mutation classes (SBS, DBS and Indel), we re-scaled the mutation frequency to the [0, 1] interval by dividing the maximum mutation frequency for further deciphering. t-SNE analysis was conducted using the R package *tSNE* [19] to explore the data distribution.

Framework of DSAE model for mutational signature discovery

We designed our DSAE model with three layers: an input layer, a latent layer, and an output layer (Fig. 1). For each dataset, we randomly selected 80% of the input matrix as the training data and the remaining 20% as the testing data. The encoding process includes a linear transformation of the input matrix followed by a nonlinear Rectified Linear Units transformation.

The decoding process aims at reconstructing the input by transforming the latent matrix Y using the decoding weight matrix W_d and the hidden bias vector b_d, followed by applying a *Softmax* classification.

We defined the loss function L_H based on the difference between the input matrix (X) and the reconstructed mutational profiles on the output layer (Z). L_H, also called the reconstruction error, takes the format of mean squared error (MSE). To avoid flat signatures, we further included a L_1 regularization to minimize L_H, defined {71}as:

$$\mathrm{minimize}\left(L_H(X,Z)\right) = \mathrm{minimize}\left(\underbrace{\frac{1}{K}\sum_{i=1}^{K}(z_i - x_i)^2}_{} + \lambda \times \underbrace{\frac{1}{K}\|Y\|_{L_1}}_{}\right)$$

where K is the total number of mutation types and λ is the parameter to balance the relative contribution between the mean squared error (the left part) and the mean of absolute value of latent matrix Y (the right part).

To accelerate the training process, we trained the DSAE model in sample batches. Training processes stopped once the specified number of epochs was reached. DSAE models were implemented using the *Keras* python library with a TensorFlow backend (version 1.0.1).

Model hyperparameter optimization and latent contribution evaluation

Several parameters in the model could impact the performance, such as dimension of the latent layer, number of epochs, batch size, and learning rate. To reach the appropriate performance of the model, we carried out parameter optimization with a 10-fold cross validation for each mutation class (SBS, DBS, and Indel), respectively. After parameter sweeping, we selected the optimized parameters with the best performance: the latent vector dimensions were determined as 200, 35, and 42 for SBS, DBS, and Indel, respectively; L_1 regularization was determined as $\lambda = 1 \times 10^{-12}$; the batch size was determined as 32 over 50 training epochs with a learning rate of 0.001, and the noisy factor was 0.01 for DBS or 0 for SBS and Indel. To allow the manual interpretation of nodes, we named each node in the hidden layer as "latent i" based on the order appeared.

To assess the impact and contribution of each latent layer, we nullified each latent layer within the model by setting all output from that latent to zero to block information flow, instead of removing it followed by re-training [41]. The new reconstruction error L_H' was compared to the original L_H to represent the contribution of the corresponding latent layer, which was later used for signature weight calculation.

Mutational signature comparison

We used the cosine similarity implemented in the R package *MutationalPatterns* [13] to compare two mutational signatures A and B: $cos = similarity(A, B) = \frac{\sqrt{\sum_{i=1}^{K} A_i B_i}}{\sqrt{\sum_{i=1}^{K} A_i^2}\sqrt{\sum_{i=1}^{K} B_i^2}}$. A cos of 1 indicates the two signatures are identical and 0 indicates the two signatures being independent.

Mutational signatures across cancer types

We used the encoding weight matrix W_e to determine mutational signatures in association with cancer types or cancer clinical variables. For each cancer type, we defined a group label, g (a vector in length $N = 2780$), to denote samples from the cancer type ($g_i = 1$) and samples from other cancer types ($g_i = 0$). We fitted a logistic regression as follows: $logit(g) \sim (W_e)^T$. The clinical data was downloaded from the ICGC website [42].

Data and code availability

All data and the code in the paper are present in Supplementary materials and GitHub https://github.com/bsml320/DeepMS.

Acknowledgements We thank Drs Chen Wang and Wei Xie for insightful discussion.

Funding Cancer Prevention and Research Institute of Texas (CPRIT RP180734), National Institutes of Health (R01LM012806).

Author contributions PJ, GP, and ZZ conceived the study. GP performed data analysis. GP and RH constructed the models. GP, YD, PJ, and ZZ interpreted the results. GP, PJ, and ZZ wrote the manuscript. All authors read and approved the final manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Garraway LA, Lander ES. Lessons from the cancer genome. Cell. 2013;153:17–37.
2. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500:415–21.
3. Alexandrov L, Kim J, Haradhvala NJ, Huang MN, Ng AWT, Boot A, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578:94–101.
4. Jia P, Pao W, Zhao Z. Patterns and processes of somatic mutations in nine major cancers. BMC Med Genom. 2014;7:11.
5. Hainaut P, Pfeifer GP. Patterns of p53 G–>T transversions in lung cancers reflect the primary mutagenic signature of DNA-damage by tobacco smoke. Carcinogenesis. 2001;22:367–74.
6. Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene. 2002;21:7435–51.
7. Pfeifer GP, You YH, Besaratinia A. Mutations induced by ultraviolet light. Mutat Res. 2005;571:19–31.
8. Pena-Diaz J, Bregenhorn S, Ghodgaonkar M, Follonier C, Artola-Boran M, Castor D, et al. Noncanonical mismatch repair as a source of genomic instability in human cells. Mol Cell. 2017;47:669–80.
9. Alexandrov LB, Nik-Zainal S, Siu HC, Leung SY, Stratton MR. A mutational signature in gastric cancer suggests therapeutic strategies. Nat Commun. 2015;6:8683.
10. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45:D777–83.
11. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. Cell Rep. 2013;3:246–59.
12. Gehring JS, Fischer B, Lawrence M, Huber W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics. 2015;31:3673–5.
13. Blokzijl F, Janssen R, van Boxtel R, Cuppen E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med. 2018;10:33.
14. Ardin M, Cahais V, Castells X, Bouaoun L, Byrnes G, Herceg Z, et al. MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. BMC Bioinforma. 2016;17:170.
15. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. 2016;17:31.
16. Thurau C, Kersting K, Wahabzada M, Bauckhage C. Convex non-negative matrix factorization for massive datasets. Knowl Inform Syst. 2011;29:457–78.
17. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems*, vol. 1. Lake Tahoe, Nevada: Curran Associates Inc.; 2012, p. 1097–105.
18. Rifai S, Vincent P, Muller X, Glorot X, Bengio Y. Contractive auto-encoders: explicit invariance during feature extraction. In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. Bellevue, Washington, USA: Omnipress; 2011, p. 833–40.
19. Lvd Maaten, Hinton GE. Visualizing high-dimensional data Using t-SNE. J Mach Learn Res. 2008;9:2579–605.
20. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol P-A. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res. 2010;11:3371–408.
21. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499:214–8.

22. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, et al. A compendium of mutational signatures of environmental agents. Cell. 2019;177:821–36.
23. Hatakeyama K, Ohshima K, Nagashima T, Ohnami S, Serizawa M, Shimoda Y, et al. Molecular profiling and sequential somatic mutation shift in hypermutator tumours harbouring POLE mutations. Sci Rep. 2018;8:8700.
24. Palles C, Cazier JB, Howarth KM, Domingo E, Jones AM, Broderick P, et al. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. Nat Genet. 2013;45:136–44.
25. Drost J, van Boxtel R, Blokzijl F, Mizutani T, Sasaki N, Sasselli V, et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. Science. 2017;358:234–8.
26. Boot A, Huang MN, Ng AWT, Ho SC, Lim JQ, Kawakami Y, et al. In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. Genome Res. 2018;28:654–65.
27. Viel A, Bruselles A, Meccia E, Fornasarig M, Quaia M, Canzonieri V, et al. A specific mutational signature associated with DNA 8-oxoguanine persistence in MUTYH-defective colorectal cancer. EBioMedicine. 2017;20:39–49.
28. Pilati C, Shinde J, Alexandrov LB, Assie G, Andre T, Helias-Rodzewicz Z, et al. Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. J Pathol. 2017;242:10–5.
29. Tomasetti C, Li L, Vogelstein B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science. 2017;355:1330–4.
30. Jia P, Wang Q, Chen Q, Hutchinson KE, Pao W, Zhao Z. MSEA: detection and quantification of mutation hotspots through mutation set enrichment analysis. Genome Biol. 2014;15:489.
31. Cooper DN, Mort M, Stenson PD, Ball EV, Chuzhanova NA. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. Hum Genom. 2010;4:406–10.
32. Chuzhanova NA, Anassis EJ, Ball EV, Krawczak M, Cooper DN. Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. Hum Mutat. 2003;21:28–44.
33. Ollila J, Lappalainen I, Vihinen M. Sequence specificity in CpG mutation hotspots. FEBS Lett. 1996;396:119–22.
34. Krawczak M, Ball EV, Cooper DN. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet. 1998;63:474–88.
35. Brash DE. UV signature mutations. Photochem Photobiol. 2015;91:15–26.
36. Ceccaldi R, Rondinelli B, D'Andrea AD. Repair pathway choices and consequences at the double-strand break. Trends Cell Biol. 2016;26:52–64.
37. Pfeifer GP. Formation and processing of UV photoproducts: effects of DNA sequence and chromatin environment. Photochem Photobiol. 1997;65:270–83.
38. Vincent P, Larochelle H, Bengio Y, Manzagol P-A. Extracting and composing robust features with denoising autoencoders. In: *Proceedings of the 25th International Conference on Machine Learning*. Helsinki, Finland: ACM; 2008, p. 1096–103.
39. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35:1798–828.
40. Haradhvala NJ, Polak P, Stojanov P, Covington KR, Shinbrot E, Hess JM, et al. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. Cell. 2016;164:538–49.
41. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016;26:990–9.
42. Zhang J, Bajari R, Andric D, Gerthoffert F, Lepsa A, Nahal-Bose H, et al. The International Cancer Genome Consortium data portal. Nat Biotechnol. 2019;37:367–9.
43. Benjamini Y, Hochberg Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. J R Stat Soc: Ser B. 1995;57:289–300.