Bioinformatics, 36, 2020, i309–i316
doi: 10.1093/bioinformatics/btaa433
ISMB 2020

Cancer mutational signatures representation by large-scale context embedding

Yang Zhang[1], Yunxuan Xiao[1,2], Muyu Yang[1] and Jian Ma[1,*]

[1]Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA and [2]Department of Computer Science and Technology, Shanghai Jiao Tong University, Shanghai 200240, China

*To whom correspondence should be addressed.

Abstract

Motivation: The accumulation of somatic mutations plays critical roles in cancer development and progression. However, the global patterns of somatic mutations, especially non-coding mutations, and their roles in defining molecular subtypes of cancer have not been well characterized due to the computational challenges in analysing the complex mutational patterns.

Results: Here, we develop a new algorithm, called MutSpace, to effectively extract patient-specific mutational features using an embedding framework for larger sequence context. Our method is motivated by the observation that the mutation rate at megabase scale and the local mutational patterns jointly contribute to distinguishing cancer subtypes, both of which can be simultaneously captured by MutSpace. Simulation evaluations show that MutSpace can effectively characterize mutational features from known patient subgroups and achieve superior performance compared with previous methods. As a proof-of-principle, we apply MutSpace to 560 breast cancer patient samples and demonstrate that our method achieves high accuracy in subtype identification. In addition, the learned embeddings from MutSpace reflect intrinsic patterns of breast cancer subtypes and other features of genome structure and function. MutSpace is a promising new framework to better understand cancer heterogeneity based on somatic mutations.

Availability and implementation: Source code of MutSpace can be accessed at: https://github.com/ma-compbio/MutSpace.

Contact: jianma@cs.cmu.edu

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

1 Introduction

Cancer is a genetic disease with high degree of heterogeneity in different tissues and cell types (Hanahan and Weinberg, 2011). Even for cancers with the same tissue-of-origin, cancer subtypes have been revealed with distinct histology, molecular phenotypes and responses to therapies, which provide the foundations for more targeted therapies (Perou *et al.*, 2000; Vogelstein *et al.*, 2013). For example, breast cancers can be typically classified into four primary subtypes, i.e. Luminal A (LumA), Luminal B (LumB), Basal and HER2, with different treatments (Perou *et al.*, 2000). Conventionally, cancer subtype identification is achieved by integrating features derived from gene expressions and histopathology. With the advent of high-throughput sequencing, the list of molecular features for cancer subtype classification has been significantly extended (Hoadley *et al.*, 2014). In particular, somatic mutations on protein-coding genes (i.e. coding mutations) emerge as a promising orthogonal feature for uncovering cancer subtypes (Arslanturk and Draghici, 2018; Kuijjer *et al.*, 2018). However, the feasibility of using non-coding mutations for cancer subtype identification is underexplored. Somatic mutations arise from the accumulation of DNA damage and the disruption of the DNA damage repair machinery (Jeggo *et al.*, 2016). Importantly, the magnitude and patterns of somatic mutations are strongly affected by exposures to exogenous and endogenous mutagens, which also serve important features to classify patients (Alexandrov *et al.*, 2013; Martincorena and Campbell, 2015). Previous works have shown that the tissue-of-origin of cancer can be accurately identified using the frequency of non-coding mutations (Jiao *et al.*, 2020; Temiz *et al.*, 2015). However, using non-coding somatic mutations for subtype identification remains challenging due to the sparsity of mutation occurrence in the genome, the complexity of mutation patterns and the difficulty to prioritize the mutations (Fu *et al.*, 2014; Watson *et al.*, 2013).

Several types of methods have been developed to extract somatic mutation features for non-coding mutations in the past. The first approach relies on the regional mutation density (RMD) at the megabase scale. Regional variation of mutation rate is associated with the epigenetic states of the tumour's cell-of-origin. For example, 74–86% of the variance of the mutation rate can be explained by combinatorial patterns of histone modifications (Polak *et al.*, 2015). However, as cancer subtypes are derived from cells with similar tissue-of-origin, whether the regional variation of mutation rate alone can provide accurate cancer subtype identification is unclear. The second approach is based on the mutational spectrum of 96 trinucleotides (immediate 5′ and 3′ bases of each mutated base, named

MS96) in each patient or the decomposed mutational signatures using non-negative matrix factorization (NMF) (Alexandrov *et al.*, 2013). MS96 or its variant mutational signatures are thought to reflect multiple mutational processes caused by the exogenous mutagens such as UV radiation (Brash, 2015) and smoking (Alexandrov *et al.*, 2016), or endogenous mutagens such as defective DNA repair activity (Alexandrov *et al.*, 2013). However, it remains elusive whether such ±1 bp mutation context (i.e. trinucleotides) is sufficient to capture the mutational signatures. Indeed, Stobbe *et al.* (2019) recently used features from recurrent non-coding mutations to cluster cancer patients and revealed a large-scale sequence context on several cancer types, including those samples caused by UV radiation, gastric-acid exposure and deregulated activity of POLE. Unfortunately, current approaches extracting all possible combinations of sequence context are computationally intractable when the mutation sequence context gets longer because the number of possible parameters increases exponentially, causing the *curse of dimensionality* and making the estimation of mutational spectrum unstable. Moreover, other covariates, such as age and sex of patients, and local genomic properties (e.g. mutations in cancer driver genes), may also influence mutation rate and patterns, but have been mostly ignored by current mutational feature representation methods. Therefore, algorithms that can extract mutational features with large-scale sequence context in each patient while also considering different covariates are needed.

Here, we perform a series of computational analysis to approach these questions. In particular, we develop a new machine learning algorithm to effectively extract patient-specific mutational features using an embedding framework for distributed entities in the field of natural language processing (NLP) (see Section 2.2 for an overview and Section 3.2 for details of the algorithm). Our main contribution is threefold: (i) We identify important mutational features that can help distinguish different cancer subtypes. Using melanoma and breast cancer data as examples, we demonstrate that RMD and large-scale nucleotide context beyond trinucleotides are both informative features (see Section 2.1), which serves as the motivation for us to develop a new method for larger-context embedding. (ii) We develop MutSpace, a new method that can effectively capture various types of mutational features by incorporating both large-scale sequence context and genomic location of mutations (see Sections 2.2–2.4). (iii) We evaluate our method using both simulated data and breast cancer patient samples, as a proof-of-principle, to demonstrate that MutSpace can extract mutational features more effectively compared with conventional methods in terms of cancer subtype classification. Overall, our method represents a significant advancement in capturing mutational features, especially for non-coding somatic mutations. The new method distinguishes itself from previous methods by incorporating various features of mutations and patients, which are key to better stratifying distinct cancer subtypes for more targeted potential treatment strategies.

2 Results

2.1 Non-coding somatic mutations show regional variations across cancer subtypes

We started by calculating the RMD (i.e. the number of mutations per million base pairs per patient) to explore its connection with different cancer types and their subtypes. Note that here we only consider non-coding mutations to mitigate bias from coding mutations affected by selective pressure during tumour progression (see Section 3.1). Previous studies have shown associations between variations of mutations in megabase scale and functional genomic data derived from tumour cell-of-origin, including replication timing and chromatin accessibility (Gonzalez-Perez *et al.*, 2019; Schuster-Böckler and Lehner, 2012; Stamatoyannopoulos *et al.*, 2009). As shown in Figure 1, we found that genome-wide RMD profiles not only exhibit distinct patterns between melanoma and breast cancer but also show different variations within their subtypes. For example, cutaneous melanoma, the most severe subtype of melanoma, has a much higher mutation rate than other subtypes of melanoma (Fig. 1A). This pattern may be caused by the elevated C→T mutations at dipyrimidines induced by the UV radiation in cutaneous melanoma [Fig. 1B; also revealed in prior studies (Akbani *et al.*, 2015; Hayward *et al.*, 2017)]. However, despite the difference in the absolute values of mutation rate, we observed that different melanoma subtypes show a consistent trend between mutational variation (Fig. 1A) and chromatin structures. In particular, heterochromatic regions marked as B compartment based on Hi-C data tend to have a higher mutation rate than more active regions in A compartment. Also, RMD profiles in breast cancer subtypes are quite different from melanoma. For example, Basal and Her2 subtypes have a higher mutation rate than LumA and LumB subtypes, consistent with previous reports based on different patient cohorts (Network *et al.*, 2012). Except for a few regions, different subtypes of breast cancer have fewer coherence patterns in terms of the correlation between mutation rate and chromatin structure, though the compositions of six mutation types are similar across different subtypes (Fig. 1D and E). Overall, these results suggest connections between cancer subtypes and mutation rate at the megabase scale.

We next explored the sequence context of mutations in different cancers and their subtypes. The trinucleotides context (i.e. the immediate nucleotides flanking the mutated base) around mutations are widely used for learning mutational signatures and characterizing different patient cohort (Alexandrov *et al.*, 2013). Recently, the analysis of recurrent mutations revealed that local sequence contexts beyond trinucleotides are hallmarks of recurrent mutations in certain cancer types, which may be related to context-specific mutagenesis (Stobbe *et al.*, 2019). We therefore sought to ask whether there exist subtype-specific sequence patterns between cancer subtypes. In Figure 1C and F, we plotted the relative enrichment of DNA sequence up to ±5 bp around the mutated base for subtypes in melanoma and breast cancer. We found that there is a strong enrichment of TTT[C→]CTT mutational pattern for C→T mutation in cutaneous melanoma, which cannot be observed in mucosal melanoma. Similarly, C→T mutations in LumB breast cancer tend to have a GGCCT sequence in the 5′ of the mutated C, which cannot be observed in the Basal subtype. We also observed that the sequence context of mutation has associations with the locations of mutations. Specifically, in cutaneous melanoma, there is a slight enrichment of C in the +1 bp of the 5′ for C→T mutations in A compartment (Fig. 1C). In addition, in LumB, there is an enrichment of C/G at the ±4th and 5th nucleotides from the mutated base for mutations in A compartment compared with mutations in B compartment (Fig. 1F). Altogether, these observations suggest that the extended context beyond trinucleotides is an informative feature related to molecular subtypes of cancer that may provide important predictive power in addition to RMD.

2.2 Overview of MutSpace—learning mutation patterns by embedding larger context

Motivated by the observations from the previous section, we developed a novel machine learning approach, named MutSpace, to effectively extract mutational features from larger context based on neural embedding algorithms (Bengio *et al.*, 2003), which can then be used for downstream analysis such as cancer subtype identification (see Sections 2.3 and 2.4). As illustrated in Figure 2A, the input of MutSpace contains two types of data: somatic mutations and cancer patients, which are naturally related by the fact that a somatic mutation is observed in one patient. The output of MutSpace is a set of vectors in high-dimensional space for cancer patients and mutational patterns, which is derived by maximizing the dot-product similarities between the embeddings of related mutation–patient pairs while minimizing similarities between those incompatible pairs. Therefore, the similarities between vectors of patients in the embedding space reflect the closeness of patients' mutational patterns, as shown in Figure 2B. The algorithm details of MutSpace are discussed in Section 3.2.

Without loss of generality, although the input data of MutSpace are in different formats, we consider both of them as

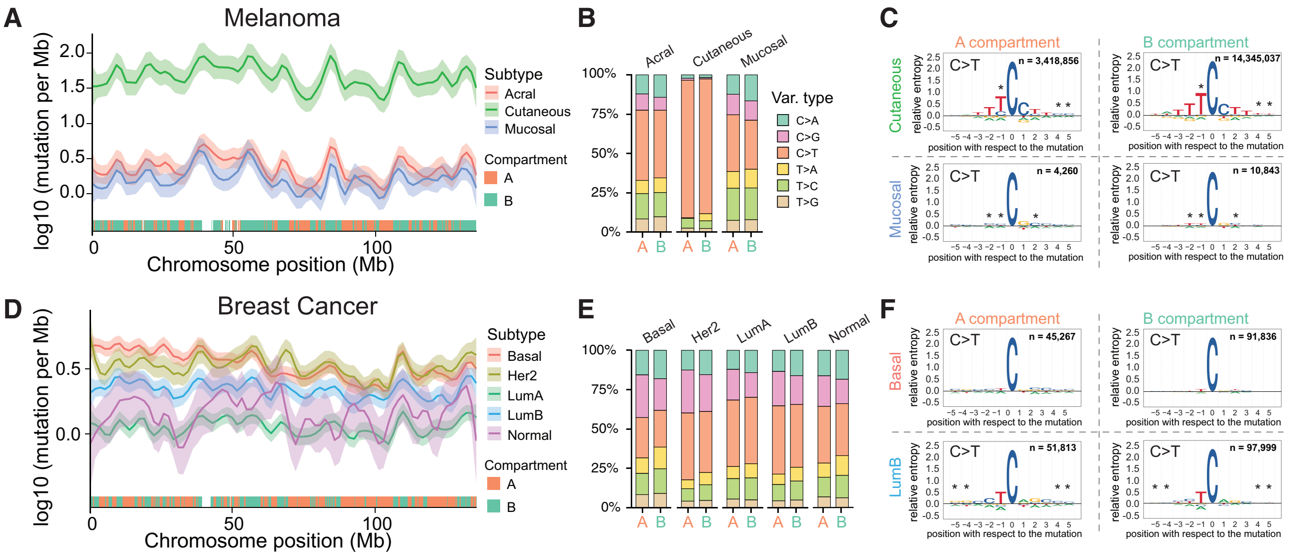

Fig. 1. Mutational patterns vary across cancers and their subtypes. (**A**) An example of RMD profile on Chromosome 10 for different subtypes in melanoma. *Y*-axis represents log10 of number of mutations per Mb per patient. (**B**) The fraction of six mutation types in different subtypes in melanoma. Mutations are separated into two groups based on Hi-C A/B compartments. (**C**) The sequence logos show the sequence context of ±5 bp around the mutated bases. *Y*-axis indicates the relative entropy representing the enrichment of nucleotides compared with genome-wide background. Each motif logo is calculated using mutations from different cancer subtypes and the location of mutations (see Section 3.1). (**D–F**) Similar plots as (A–C) using data from breast cancer

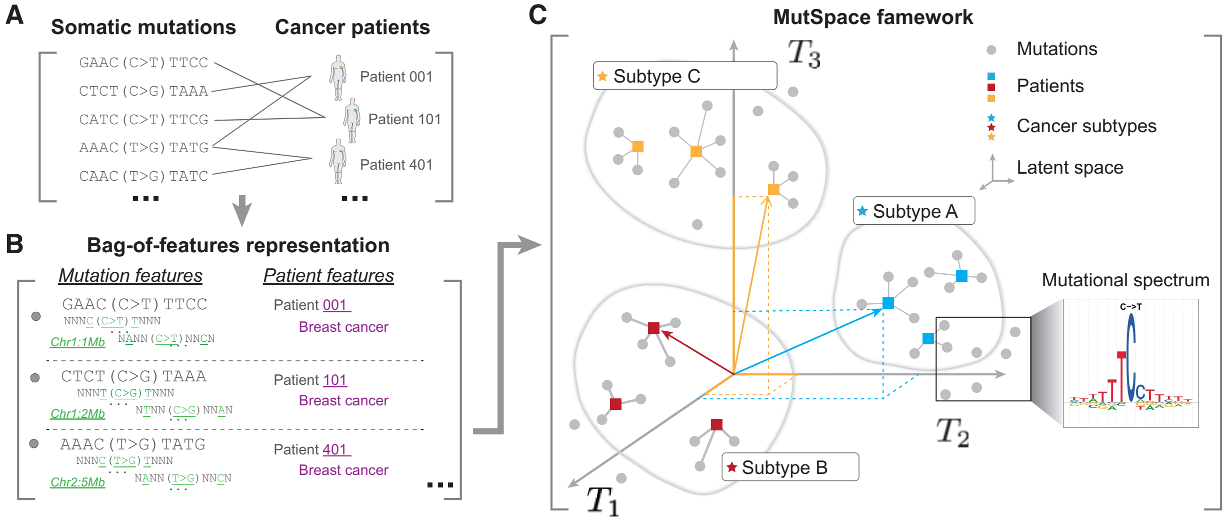

Fig. 2. Overview of the MutSpace framework. (**A**) Cancer patients, somatic mutations and their local sequence context are represented as bags-of-features (see example in Section 3.2). (**B**) Examples of mutations and their decomposed features are shown in green. Features associated with patients such as patients' ID and cancer type are shown in purple. (**C**) MutSpace jointly learns embeddings of mutations, patients and their features in the same high-dimensional space. Patients with similar mutational patterns are expected to be close to each other in the latent space

abstract entities consisting of a set of discrete features (bag-of-features) (Fig. 2B). For example, features of mutations contain the mutation type (one of the six substitution patterns, see Section 3.1), the genomic location by assigning mutations into genomic bins with fixed size (e.g. 1 Mb), and a series of discrete features for each mutation type representing the flanking nucleotides with variable distances from the mutated base (see examples in Section 3.2 and Fig. 2B). Decomposing sequence context into separate discrete features is preferred to reduce the space complexity of the algorithm especially when we model the large-scale sequence context of mutations (Shiraishi *et al.*, 2015). For patients, relevant discrete features include the cancer type of patients, unique patients' ID, and possibly other covariates of patients, such as sex and age of diagnosis. The key concept of MutSpace is to integrate these heterogeneous entities through a unified embedding framework adapted from StarSpace (Wu *et al.*, 2018), a recent embedding framework developed for NLP. In particular, MutSpace embeds mutations, patients and their features into a common high-dimensional space such that the comparisons between them can be achieved by calculating the similarity of between two vectors. MutSpace then learns embeddings of mutations and patients such that mutation patterns observed in the same patient stay close to each other but away from mutations belonging to other patients (Fig. 2C, Section 3.2). Consequently, patients containing similar mutational information tend to be embedded close in the high-dimensional space, whereas more distinct patients are further away. The model is trained to maximize the similarities between true pairs of mutations and patients while minimizing the similarities between randomly sampled noisy pairs (see Section 3.2). Finally, patients' embedded vectors can be used as patient-specific features extracted from mutations for the downstream subtype classification.

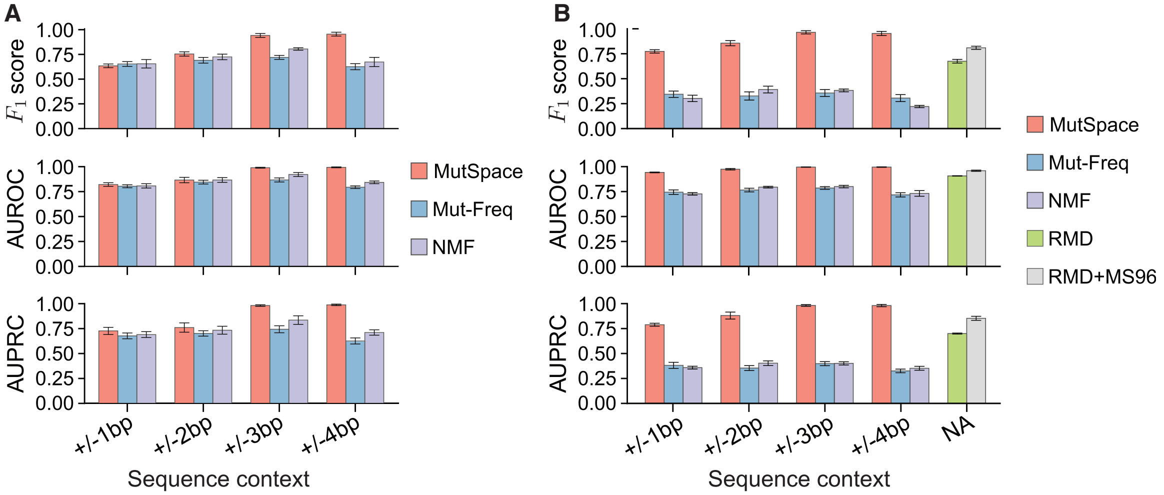

Fig. 3. Evaluation using simulated datasets. (**A**) Evaluation of MutSpace, Mut-Freq and NMF on simulated data SS-I in terms of F_1 score, AUROC and AUPRC. (**B**) Evaluation of MutSpace, Mut-Freq, NMF and RMD on simulated data SS-II in terms of F_1 score, AUROC and AUPRC

MutSpace provides a generic framework to more effectively incorporate mutational features including large-scale sequence context around mutations and other covariates for stratifying cancer patients.

2.3 Simulation demonstrates the accuracy of MutSpace in identifying cancer subtypes

To assess whether the mutational features obtained by MutSpace can accurately identify cancer subtypes, we evaluated MutSpace using synthetic datasets in two types of simulation studies. We assessed the performance based on F_1 score, AUROC and AUPRC by comparing the predicted labels of subtypes with the ground truth labels (see Section 3.5). In simulation study I (SS-I), mutational patterns of each subtype were generated from different mixtures of synthetic mutational signatures up to ±4 bp, whereas in simulation study II (SS-II), we further allowed the distribution of mutational density on the genome to vary across subtypes. Both simulation studies contain three to six subtypes and 100 patients in each subtype, respectively (see details in Section 3.4).

We compared MutSpace with other methods that extract patient-specific mutational features, including the frequency of the occurrence for each mutational pattern with different lengths of sequence context (Mut-Freq), NMF decomposition of Mut-Freq and regional mutational density (RMD) at megabase resolution (see descriptions of these methods in Section 3.3). Among different types of classifiers, we found that the SVM had the best performance and was therefore chosen as the classifier for further analysis. The average performance is reported based on fivefold nested cross validation with hyperparameters selected using the grid-search strategy (see Section 3.5). Performances are calculated for mutational patterns with different lengths of context to provide a comprehensive evaluation. We found that MutSpace is able to make accurate predictions with different lengths of sequence context in both simulated studies (Fig. 3A). In particular, MutSpace shows significant advantages in reaching a higher F_1 score than other methods as a longer sequence context is considered (Fig. 3A). In SS-II, we further added the genomic location of mutations (after discretization into the ID of 1 Mb genomic bin in the human genome) as features of mutations in MutSpace to capture the variation of mutational density in different subtypes. As the difference of mutational density cannot be captured by methods that only examine at the frequency of mutational pattern per patient, Mut-Freq and NMF consistently show poor performance with different lengths of sequence context (Fig. 3B). We therefore concatenated features extracted by RMD and Mut-Freq with ±1 bp (MS96) and used those combined features (RMD+MS96) for predictions. However, MutSpace still outperforms the combined method when ±2 bp sequence context is considered (average F_1 score 0.809 for RMD+MS96 versus 0.856 for MutSpace). Importantly, in MutSpace, only the ID of the genomic bins instead of the count of mutations per bin is used as part of the input. Together, these simulation evaluations strongly suggest that MutSpace can effectively extract features by jointly modelling mutational patterns and patients. Our method even outperforms approaches that integrate both regional mutational density and frequencies of mutation patterns.

2.4 The embeddings from MutSpace can accurately classify breast cancer subtypes

To further assess the performance of MutSpace for delineating patient-specific mutational features, we applied MutSpace to reveal the heterogeneity of breast cancer subtypes from mutational patterns. We first evaluated MutSpace on its ability to classify subtypes of breast cancer. Similar to evaluations for the simulated datasets, we trained an SVM model based on the mutational features from MutSpace and compared its performance with other methods. Figure 4A shows the best prediction results using these methods when different lengths of sequence context are considered. We found that MutSpace can accurately identify three primary subtypes in breast cancer. In particular, in the basal subtype, the average F_1 score achieved by MutSpace is 0.922, which is higher than the method using features by the concatenation of RMD and MS96 (0.896). In LumA and LumB, MutSpace outperforms other methods with a higher F_1 score, AUROC and AUPRC. Notably, we observed that MutSpace achieves its highest performance for encoding ±5 bp sequence context, consistent with our observations that breast cancer subtypes have distinct patterns in the large-scale sequence context around mutations.

To better demonstrate the ability of MutSpace in representing patients with similar mutational patterns, we projected patients' embedding from MutSpace into two-dimensional space with t-SNE (van der Maaten and Hinton, 2008) (Fig. 4B). Here each data point represents a patient with colours showing subtype information. We found that patients belonging to the same subtype are clearly clustered together. We also noticed that some patients from three subtypes tend to form a unique cluster in the t-SNE. We therefore applied HDBSCAN (McInnes *et al.*, 2017), an unsupervised clustering method to separate patients into four groups based on mutational patterns alone (Fig. 4C and D). Using the homologous recombination deficiency (HRD) score as an orthogonal clinical feature, we found that these four groups have distinct distributions of HRD score. Further

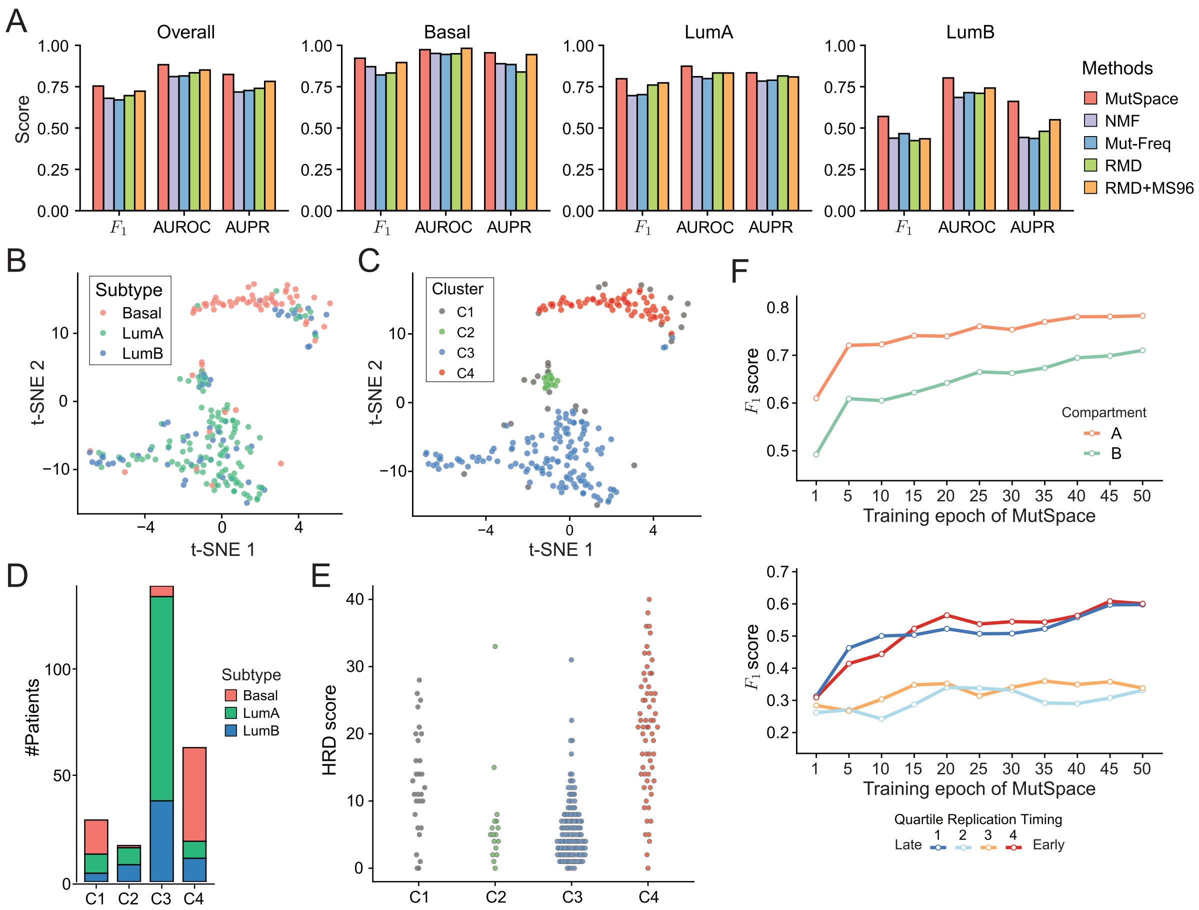

Fig. 4. (**A**) Evaluation of MutSpace, Mut-Freq, NMF and RMD for identifying breast cancer subtypes. (**B**) Visualization of patients' embeddings using t-SNE with colours indicating subtypes. (**C**) Unsupervised clustering of embedded vectors of patients using HDBSCAN identifies four subgroups of breast cancer. (**D**) Barplot shows the number of patients in each cluster with colours indicating subtypes. (**E**) The distribution of HDR score in each *de novo* identified cluster. (**F**) Curves show the trend of F_1 scores for predicting A/B compartments (upper panel) and replication timing signals (bottom panel) as different epochs of embedded vectors for genomic bins are used as input

analysis could be performed to assess whether those four groups of patients have different clinical outcomes.

Besides patients' and mutations' embeddings, we also sought to demonstrate the utility of other jointly learned feature embeddings, such as the embeddings of genomic bins. Specifically, we asked if the embedding of a genomic bin reflects useful features related to the organization and function of the corresponding genomic location. We tested this hypothesis by extracting the embeddings of genomic bins on each epoch during the MutSpace training process and using them as input to predict corresponding orthogonal features of genome structure and function. Figure 4F shows the F_1 score of the prediction on Hi-C A/B compartment annotations and quartile of replication timing signals, respectively. Notably, we found that as the training proceeds, embeddings of genomic bins become more correlated with A/B compartmentalization and replication timing, suggesting that MutSpace can successfully learn the underlying information of genomic location (covariate of mutations) and predict related genome structure and function.

Taken together, these results highlight the advantage of MutSpace in representing mutation patterns.

3 Materials and methods

3.1 Data collection

We collected tumour somatic mutations derived from whole-genome sequencing (WGS) across multiple resources (see Supplementary Table S1). We only used single-nucleotide substitutions in autosomes and combined substitutions that are reverse-complement to each other (e.g. G→T is treated the same as C→A) into six mutation types: C→A, C→G, C→T, T→A, T→C and T→G. We discarded mutations located in the protein-coding regions according to the human GENCODE annotation release 19 as sequence context of those somatic coding mutations are likely under stronger selection pressure and may show distinct sequence preference compared with non-coding mutations. We further removed mutations that are located in the human blacklisted regions defined by the ENCODE project (accession: ENCSR636HFF) or overlapped with known common SNP track (snp151Common) downloaded from the UCSC Genome Browser. Flanking nucleotides around mutations (reverse complement the DNA sequence if necessary, see above for the definition of six mutation types) were extracted from the human reference assembly hg19 downloaded from the UCSC Genome Browser. Clinical information including molecular subtypes of cancer patients was retrieved from various previous literatures investigating each cancer sequencing dataset. Supplementary Table S2 shows the number of patients in each subtype from different cancer datasets. We collected functional genomics data from public resources. The source of data is shown in Supplementary Table S3. Sequence logos of mutational pattern around mutations were generated by adapting scripts provided by Stobbe *et al.* (2019) on GitHub.

3.2 MutSpace—joint modelling of mutational patterns and patients

In MutSpace, we jointly learn an embedding for mutational patterns (i.e. one of the six mutation types and its flanking nucleotides) observed in cancer cohort and cancer patients into a continuous vector space $\mathbb{R}^d$. In $\mathbb{R}^d$, mutational patterns and patients are both entities represented as d dimensional continuous vectors (i.e. embedding vectors) induced in a bag-of-features manner. This bag-of-features embedding framework is developed based on a recent semantic embedding model, StarSpace (Wu *et al.*, 2018), which has been proven to be a general-purpose method for various tasks in the field of NLP and also the prediction of transcription factor binding in genomics (Yuan *et al.*, 2019). In StarSpace, heterogeneous entities are embedded in the same semantic space, and each entity is described by a set of discrete features (i.e. bag-of-features). For example, in NLP a document entity can be described by a set of words, and a reader can be described by a set of documents the reader likes.

The bag-of-features representation provides flexibility to the choice of entities' features and makes it possible to incorporate heterogeneous covariates. In our cases, the embedded vector of the ith patient is represented by embedded vectors of this patient's features ($\mathcal{F}_{i,1}$, $\mathcal{F}_{i,2}$, ..., $\mathcal{F}_{i,m_i}$), where m_i is the number of discrete features of the patient. For example, features associated with patients include cancer type, patient's IDs, etc. The embedding of the ith patient is thus induced by the summation of embeddings of all features from this patient.

$$\boldsymbol{p}_i = \frac{1}{m_i^p}\sum_{j=1}^{m_i}\mathcal{F}_{i,j} \tag{1}$$

Here p is a hyperparameter and we keep it the same as the default value (0.5) used in StarSpace.

Similarly, the embedding of a mutation with a large-scale local context is represented by the summation of embedded vectors of its sub-components and other features associated with this mutation. Here, we define a set of sub-components of a mutational pattern as a series of non-overlapping flanking nucleotides of length w with different distances from the central mutated base. The maximum distance of the flanking nucleotides from the central mutated base determines the size of sequence context we consider (represented by l). For instance, AACC[C→A]TTCG is one example of an observed ±4 bp mutational pattern. When w equals to 1, we split this mutational pattern into four sub-components: ANNN[C→A]NNNG, NANN[C→A]NNCN, NNCN[C→A]NTNN and NNNC[C→A]TNNN (here N indicates any nucleotide from A, C, G and T). This representation of mutation patterns by sub-components can significantly reduce the number of features associated with a mutation, allowing a long flanking sequence of the mutation to be included in the model. Considering a mutational pattern of length $2l+1$, where l is the length of sequence on either up- or downstream of the mutated base, the number of sequence patterns given the sub-components with length w is $\lceil l/w \rceil \times 4^{2w}$ because the number of sub-component is $\lceil l/w \rceil$, and the number of the possible DNA sequence is 4^{2w} because the number of nucleotides in each sub-component is $2w$ (considering both up- and down-stream nucleotides). Therefore, the total number of sub-components associated with six types of substitutions is $6 \times \lceil l/w \rceil \times 4^{2w}$, with a spatial complexity significantly lower than the entire combination of the flanking sequence (6×4^{2l}). Based on our observations on real mutation data, when l is greater than 6, the nucleotide compositions are very similar to the genome-wide average. Therefore, we only considered l smaller than 6 bp in both simulation and real data application. We can also encode other covariates of mutations as discrete features. In particular, the genomic location of the mutation is encoded by splitting the genome into fixed-sized bins and assigning mutations to these bins by the bins' ID. Therefore, the final embedding of the ith mutation is induced by the summation of embeddings of all its associated sub-components and other discrete features ($\mathcal{C}_{i,1}$, $\mathcal{C}_{i,2}$, ..., $\mathcal{C}_{i,n_i}$), where n_i is the number of features associated with each mutation.

$$\boldsymbol{s}_i = \frac{1}{n_i^p}\sum_{j=1}^{n_i}\mathcal{C}_{i,j} \tag{2}$$

To learn the embedded vectors, we developed a similar approach used in the StarSpace by maximizing the similarities calculated from positive mutation–patient pairs while minimizing those from negative ones. Briefly, we define a positive pair of latent embedded vectors in mutation–patient relationship as $(\boldsymbol{s}_{ij}, \boldsymbol{p}_i^+)$, which represents the jth mutation $\boldsymbol{s}_{ij}$ observed in the ith patient $\boldsymbol{p}_i$. Negative pairs of latent embedded vectors between mutations and patients $\{(\boldsymbol{s}_{ij}, \boldsymbol{p}^-_{k=1,2,\ldots,K}) | k \neq i\}$ are generated by negative sampling (Mikolov *et al.*, 2013), through which the jth mutation observed in the ith patient is paired with those from the rest of the patients other than patient i and this process is repeated for K times. The model is then trained to maximize the similarities between positive pairs while minimizing those in negative pairs through optimizing the following hinge loss function:

$$L = \sum_{(\boldsymbol{s}_{ij},\boldsymbol{p}_i^+)}\frac{1}{K}\sum_{k=1}^{K}\max\left(0, \mu - \mathrm{sim}\left(\boldsymbol{s}_{ij}, \boldsymbol{p}_i^+\right) + \mathrm{sim}\left(\boldsymbol{s}_{ij}, \boldsymbol{p}_k^-\right)\right) \tag{3}$$

where μ is a constant margin parameter, which is 1.0 by default. Also, to measure the similarity between entities, we calculated the scaled dot-product similarity $\mathrm{sim}(\cdot,\cdot)$ between mutation embeddings and patient embeddings as shown below:

$$\mathrm{sim}(a, b) = \frac{ab^\top}{\sqrt{d}}, a, b \in \mathbb{R}^d \tag{4}$$

We also applied a constraint to embedded vectors to enforce an upper bound of the ℓ^2-norm of embeddings. In each training epoch of MutSpace, patients' embeddings and the embeddings of patients' features, mutations' embeddings and the embedding of mutations' sub-components are all optimized based on the gradient of this loss using the Adam optimization algorithm (Kingma and Ba, 2014). By minimizing the defined loss shown above, in the latent vector space, the mutational patterns will gather around their corresponding patients, while the patients with similar mutational patterns will also become closer to each other.

After the training, the learned embedded vectors of entities can then be used for downstream tasks, such as classifying patients into subtypes. To retrieve mutational patterns that are closer to one patient, we can use the learned embedded vectors to measure the similarity between two entities (e.g. mutational patterns with patients, patients with other patients). We can select top q compatible mutational patterns for the ith patient by sorting $\mathrm{sim}(\boldsymbol{s}, \boldsymbol{p}_i)$ over all possible mutational patterns $\boldsymbol{s}$. The MutSpace algorithm was implemented using PyTorch (Paszke *et al.*, 2019).

3.3 Mutational features extraction using other methods

There are other methods that can also extract mutational features for each patient. Here we compared the performance of MutSpace with several other methods, including calculating the frequencies of occurrence for mutational patterns in each patient (Mut-Freq), decomposing the frequency matrix of mutational patterns using the NMF algorithm, and regional mutational density (RMD) (Jiao *et al.*, 2020; Salvadores *et al.*, 2019).

We obtained the frequency of mutational patterns (Mut-Freq) by calculating the relative frequencies of them in each patient. Regarding mutational patterns, we considered different lengths of sequence context around the mutated base. As the length of the sequence context grows, the number of mutational patterns grows exponentially, which makes the training process of the classifier unstable. Therefore, we applied principal component analysis (PCA) on the mutation frequency matrix and kept the top 50 principal components as the mutational features for each patient. Evaluations on both simulated and real data showed that this strategy can improve the performance when a larger sequence context (>±2 bp) is considered. We also applied the NMF algorithm to the raw frequency matrix of mutational patterns to reduce the number of mutational features. We used scikit-learn for the implementation of NMF (Pedregosa *et al.*, 2011). We set the number of components decomposed by NMF also to be 50. We found that choosing different parameters will not noticeably affect the performance. We calculated RMD by counting the number of non-coding mutations in each 1 Mb window in the human genome. We further normalized it by dividing the count by the effective size of each bin after we removed those regions belonging to protein-coding genes or unmappable regions according to the annotation of Umap (Karimzadeh *et al.*, 2018).

3.4 Simulation of mutation dataset

We used two types of settings for data simulation, corresponding to SS-I and SS-II. In SS-I, we assume that cancer subtypes are different from each other only by their mutational patterns rather than the distributions of mutation density on the genome. Specifically, we first built a series of synthetic mutational signatures such that each signature represents a set of 393 216 mutation patterns depicting all possible combinations of nucleotides within ±4 bp of the central mutated base. Here we only consider six types of single nucleotide mutations as mentioned above. Each mutational pattern is then associated with a frequency indicating the probability of observing such mutational pattern and the sum of all frequencies in each synthetic signature is 1. To make the synthetic signatures analogous to what we observed in the real data, we utilized the COSMIC mutational signatures (https://cancer.sanger.ac.uk/cosmic/signatures_v2) as a reference to build the margin frequency of trinucleotides (immediate nucleotides flanking the mutated base) in synthetic signatures by mixing frequency of different COSMIC signatures. The frequency of flanking nucleotides with distance >1 bp from the mutated base is derived by ignoring the central mutated base when mixing COSMIC trinucleotides signatures. Finally, the frequencies of mutational patterns in each subtype were constructed by a mixture model with weights as the contribution of those ±4 bp synthetic signatures. To generate synthetic data, we randomly sampled mutations according to the frequencies in each subtype. In each subtype, we simulated mutation data for 100 patients. The number of mutations in each patient (N) is determined by parameter θ ($N = 10^{\theta}$), where the value of θ is randomly sampled from a Gaussian distribution with a mean of 3 and a standard deviation of 0.3. In SS-II, we assume that cancer subtypes are also varied by their mutational density profile along the genome. For simplicity, we assume all the mutations are from one chromosome and their locations on the chromosome are sampled from a beta distribution with shape parameters α and β (we assume the start of chromosome is 0 and the end of chromosome is 1). We further discretized the locations of mutations by binning mutations into 100 bins. In SS-II, cancer subtypes are different from each other by either the weight of mixtures of the synthetic signatures or the underlying distribution of locations determined by the shape parameters.

3.5 Classification of cancer subtypes using mutational features

To demonstrate the power of the embedded vectors extracted by MutSpace on identifying cancer subtypes, we performed supervised classification using kernelized support-vector machine (SVM) (Vapnik 2000). The kernelized SVM takes mutational features extracted from each patient as input and returns the predicted subtype assignment and probability. To apply SVM in multi-class setting, we applied one-versus-rest strategy. For each subtype, we treated patients with subtype A as positive and patients with other subtypes as negative, and build a single SVM classifier to learn whether a patient belongs to A subtype. We trained our model with nested cross validation using different representations of mutational features. The inner fivefold cross-validation conducts grid-search to select hyperparameters (e.g. kernel, regularization penalty, etc.) that leads to the highest F_1 score, and the outer fivefold cross validation evaluates model performance on testing dataset by randomly splitting data into 80% training and 20% testing for five times. To account for the imbalanced number of instances in each subtype, we modulated the weight of class for a subtype to increase the penalty for mis-classifying minor classes by the inverse of their frequency.

We evaluated the performance using F_1 score, area under receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC). F_1 score is the harmonic mean of precision and recall for each binary classification task, i.e. $F_1 = (2 \times \text{precision} \times \text{recall})/(\text{precision} + \text{recall})$. AUPRC measures how well the model is able to identify all positive instances without mistakenly labelling negative instances as positive. ROC curve is plotted as true positive rate versus false positive rate, and the area under this curve measures the ability to separate two classes. Since all of the metrics described above are only applicable to binary cases, we first calculated the scores for each subtype and then took the average weighted by the class frequency as the overall score. We report the mean and variance of these metrics calculated from our nested cross validation. Model training and evaluation were implemented using Python package scikit-learn version 0.22.

4 Conclusion and discussion

Although the integration of somatic mutations for cancer subtype identification has received much attention recently, it remains unclear whether non-coding somatic mutations can offer unique characterizations of molecular subtypes of cancer. One of the advantages of using somatic mutations as input data is that the mutations provide a cumulative record of the evolutionary process caused by exogenous and endogenous mutagens (Jeggo *et al.*, 2016). In this work, we developed MutSpace to specifically address the computational challenges to consider large-scale context for mutational signatures. MutSpace effectively extracts patient-specific features by jointly modelling mutations and patients in a latent high-dimensional space. We demonstrated the feasibility of using non-coding somatic mutations for cancer subtype identification by simultaneously considering mutational density and mutational patterns. Both simulations and proof-of-principle real data application confirmed that the mutational features captured by MutSpace are capable of stratifying cancer subtypes.

There are several directions that we can further improve MutSpace. First, one of the promising future directions is to integrate non-coding somatic mutations with coding mutations, in particular, those residing in the known cancer driver genes. Kumar *et al.* (2020) have recently discovered that the aggregated effect of non-coding mutations may play a more important role in tumorigenesis than coding mutations. In the framework of MutSpace, coding mutations can be treated as covariates of patients and thus modulate the final embeddings of patients. Future research is needed to evaluate whether this integrative approach can provide a better stratification for different types of cancers. Second, we have a limited ability to interpret mutational features extracted from MutSpace, which is a challenge for embedding-based methods. Projecting the mutational features to lower dimensions by approaches such as t-SNE is a possible way of exploring the enrichment of external annotations on embedded space and has been widely used in other areas of computational genomics such as single-cell transcriptomics. A hierarchical approach to further cluster patients based on their embeddings using localized diffusion folders (LDF) (David *et al.*, 2010) may provide potential representations of patients' relationship. Another possible alternative is to utilize synthetic mutation data with known aetiology. For example, a recent study (Kucab *et al.*, 2019) generated mutation data by exposing cells to dozens of known environmental carcinogens. Integrating these types of data together with somatic mutation data from cancer patients in the same latent space may have the potential to interpret patient-specific mutation patterns based on mutations related to known carcinogens. Third, choosing the right covariates as features of mutations and patients requires prior knowledge. Future work is needed to more efficiently determine the best set of covariates as input features. Finally, with the development of single-cell-based technologies, somatic mutation data from longitudinal cancer studies are becoming available. It would be important to further extend the framework of MutSpace by specifically considering the temporal patterns of the mutational signatures in the latent space. This may provide key insights into the evolutionary forces that drive somatic mutation accumulation. Nevertheless, MutSpace provides a new framework that would allow us to further pursue these exciting questions.

Acknowledgements

The authors thank Ruochi Zhang for helpful discussions that improved the manuscript.

Funding

This work was supported in part by an ASPIRE Award from the Mark Foundation for Cancer Research (Grant# 19-043-ASP) (to J.M.) and a Rhino-Bird Gift Award from Tencent AI Lab (to J.M.).

Conflict of Interest: none declared.

References

Akbani,R. *et al.* (2015) Genomic classification of cutaneous melanoma. *Cell*, **161**, 1681–1696.

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

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

Arslanturk,S. and Draghici,S. (2018) Disease subtyping using somatic variant data. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 277–282. ACM, New York, NY.

Bengio,Y. *et al.* (2003) A neural probabilistic language model. *J. Mach. Learn. Res.*, **3**, 1137–1155.

Brash,D.E. (2015) UV signature mutations. *Photochem. Photobiol.*, **91**, 15–26.

David,G. *et al.* (2010) Hierarchical clustering via localized diffusion folders. In: *2010 AAAI Fall Symposium Series*. AAAI, Menlo Park, CA.

Fu,Y. *et al.* (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.*, **15**, 480.

Gonzalez-Perez,A. *et al.* (2019) Local determinants of the mutational landscape of the human genome. *Cell*, **177**, 101–114.

Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.

Hayward,N.K. *et al.* (2017) Whole-genome landscapes of major melanoma subtypes. *Nature*, **545**, 175–180.

Hoadley,K.A. *et al.* (2014) Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, **158**, 929–944.

Jeggo,P.A. *et al.* (2016) DNA repair, genome stability and cancer: a historical perspective. *Nat. Rev. Cancer*, **16**, 35–42.

Jiao,W. *et al.* (2020) A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat. Commun.*, **11**, 1–12.

Karimzadeh,M. *et al.* (2018) Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Res.*, **46**, e120.

Kingma,D.P. and Ba,J. (2014) Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Kucab,J.E. *et al.* (2019) A compendium of mutational signatures of environmental agents. *Cell*, **177**, 821–836.

Kuijjer,M.L. *et al.* (2018) Cancer subtype identification using somatic mutation data. *Br. J. Cancer*, **118**, 1492–1501.

Kumar,S. *et al.* (2020) Passenger mutations in more than 2,500 cancer genomes: overall molecular functional impact and consequences. *Cell*, **180**, 915–927.e16.

Martincorena,I. and Campbell,P.J. (2015) Somatic mutation in cancer and normal cells. *Science*, **349**, 1483–1489.

McInnes,L. *et al.* (2017) hdbscan: hierarchical density based clustering. *J. Open Source Softw.*, **2**, 205.

Mikolov,T. *et al.* (2013) Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Network,C.G.A. *et al.* (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61.

Paszke,A. *et al.* (2019) PyTorch: an imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems*, pp. 8024–8035. Curran Associates, Inc., Red Hook, NY.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Perou,C.M. *et al.* (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.

Polak,P. *et al.* (2015) Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*, **518**, 360–364.

Salvadores,M. *et al.* (2019) Passenger mutations accurately classify human tumors. *PLoS Comput. Biol.*, **15**, e1006953.

Schuster-Böckler,B. and Lehner,B. (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, **488**, 504–507.

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

Stamatoyannopoulos,J.A. *et al.* (2009) Human mutation rate associated with DNA replication timing. *Nat. Genet.*, **41**, 393–395.

Stobbe,M.D. *et al.* (2019) Recurrent somatic mutations reveal new insights into consequences of mutagenic processes in cancer. *PLoS Comput. Biol.*, **15**, e1007496.

Temiz,N.A. *et al.* (2015) The somatic autosomal mutation matrix in cancer genomes. *Hum. Genet.*, **134**, 851–864.

van der Maaten,L. and Hinton,G. (2008) Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.

Vapnik, V. (2000) The nature of statistical learning theory. Springer-Verlag, *New York*, NY.

Vogelstein,B. *et al.* (2013) Cancer genome landscapes. *Science*, **339**, 1546–1558.

Watson,I.R. *et al.* (2013) Emerging patterns of somatic mutations in cancer. *Nat. Rev. Genet.*, **14**, 703–718.

Wu,L.Y. *et al.* (2018) StarSpace: embed all the things! In: *Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI, Palo Alto, CA.

Yuan,H. *et al.* (2019) BindSpace decodes transcription factor binding signals by large-scale sequence embedding. *Nat. Methods*, **16**, 858–861.