Original Article

Insight of characteristic of mutation sequences in human cancers via an unsupervised neural network approach.

H. C. Ji[1 *], J. J. Li, Q[2 *]. Zhang[1 *], J. Y. Yang[1], F. Tian[1], X. W. Wang, W. Pan[1], X. X. Wang[1], H. M. Zhang[1]

1 Department of oncology, Xijing Hospital, Fourth Military Medical University. No.127 West Changle Road, Xi'an, China.

2 Department of emergency, Xijing Hospital, Fourth Military Medical University. No.127 West Changle Road, Xi'an, China.

* These authors contributed equally to this work.

Corresponding author: Prof. Hongmei Zhang. Department of oncology, Xijing Hospital, Air Force Military Medical University. No.127 West Changle Road, Xi'an, 710032, China. Tel: (+86) 13991293309, E-mail: zhm@fmmu.edu.cn

Abstract

Background: Mutation processes leave different signatures in genes. Previous studies have suggested that both the mutated and flanking bases influence somatic mutation characteristics. However, the understanding of how flanking sequences influence somatic mutation characteristics is limited. *Materials and methods:* We constructed a long short-term memory – self organizing map (LSTM-SOM) unsupervised neural network. By extracting mutated sequence features via LSTM and clustering similar features with SOM, somatic mutations in The Cancer Genome Atlas database were clustered according to their mutation type and flanking sequences. The relationship between MB and cancer characteristics was then analyzed. At last, we clustered the patients into different classes according to the composition of MB by K-means method, and then studied the differences in clinical features and survival between classes. *Results:* Ten classes of mutant sequences (named mutation blots, MBs) were obtained from 2,141,527 somatic mutations. Different features in mutation bases and flanking sequences were revealed among MBs. MB reflect both the site and pathological features of cancers. MBs were related to clinical features, including age, sex, and cancer stage. Class of MB in a given gene is associated with survival. Finally, patients were clustered into 7 classes according to MB composition. Significant differences in survival and clinical features were observed among different patient classes. *Conclusions:* Our study provides a novel method for analyzing the information of mutant sequences and reveals the extensive relationships among mutant sequences, clinical features, and cancer patient survival.

Keywords: mutation sequence; unsupervised learning; cancer; clinical feature; prognosis

Highlights:

1. An unsupervised learning approach to analysis information of somatic mutation sequences in cancers.

2. Mutation sequences reflect the characteristics of cancers.

3. The composition of mutation sequences affects the clinical features and survival of cancer patients.

Introduction

The stability of the cell genome is continually threatened by endogenous and exogenous factors that may lead to DNA damage.[1,2] If not repaired properly, DNA damage may result in genetic mutations.[3,4] The development of cancers involves a series of genetic mutations.[5] A number of internal and external factors underlying genetic mutations have been identified, such as smoking, alcohol consumption and mismatch repair deficiency.[5,6] In some kinds of cancers, such as colon cancer and breast cancer, there has been a great deal of research elucidating the relationship between genetic mutations and cancer-related processes.[7] However, in most cases, the role of genetic mutations in tumor progression is still poorly understood.

Genetic mutations include single-base substitutions (SBSs), small insertions and deletions (indels), genome rearrangement and chromosome copy-number changes.[8] In clinical studies, patients with mutations in a given gene show differences in survival and drug susceptibility.[9-11] With the development of sequencing technology, large amounts of mutation data from cancer patients have been obtained and made available in relevant databases, such as The Cancer Genome Atlas (TCGA) database. In the context of increasing sample sizes, a number of mutation signatures that are correlated with certain mutation processes have been identified[12, 13].

SBSs contribute the largest proportion of genetic mutations. Mathematical methods have been used to decipher mutation signatures from somatic mutation catalogs.[2,8,14-20]. The clustering methods applied in some studies have included 1-2 bases next to mutated bases, and the results have suggested that flanking bases influence mutation signatures[2,8]. However, the inclusion of adjacent genes in such analyses leads to an exponential increase in the number of possible classifications, which makes it difficult to analyze the effect of flanking sequences on mutation signatures.

A long short-term memory (LSTM) network is a special kind of recurrent neural network (RNN). Compared with a naive RNN, LSTM performs better in extracting features from long sequences, such as sentences.[21,22] LSTM has been used to analyze DNA or RNA sequence information[23-25]. A self-organizing map (SOM) algorithm is an unsupervised clustering algorithm. The method of "competitive learning" can identify interconnections between samples and present their categories in a lower-dimensional form.[26,27] The use of LSTM to extract the features of mutated sequences and the identification of similar features with the SOM algorithm provided an approach for analyzing the characteristics of mutated sequences and their relationship with cancer development.

Methods

Data availability

SBS data and clinical data of patients involved in this study were obtained from the TCGA database. In the LSTM-SOM model, 100 flanking bases were included in the analysis, and the flanking sequence was obtained from the Genome Reference Consortium human genome build 38(GRCh38) based on the mutation sites of SBSs in TCGA data. Reference bases provided by TCGA were compared with GRCh38 to further ensure accuracy.

LSTM-SOM model building

Our LSTM model works via a cycle of 3 steps: 1. extraction of the feature vector of the mutant sequence by LSTM; 2. clustering of feature vectors by SOM, and feature vectors are

updated at the same time to bring vectors with similar features closer together; and 3. use of the updated feature vectors for the labeling and training of the LSTM model.

Step 1. Obtaining feature vectors by LSTM

Mutant sequences are represented in the form of a matrix. A 1×2 vector is used to represent different bases (A: [0, 0]; T:[0, 1]; C: [1, 0]; G: [1, 1]; N:[-1, -1]). In this way, a mutated sequence can be represented by an n×4 matrix.

As the "forgetting" mechanism of LSTM[21, 22], the unit closer to the end of the sequences has a greater influence on the output of LSTM. In our model, LSTM is designed to read from both ends of the mutated sequence. In this way, the mutation site is placed at the ends of both sequences to reinforce its influence on the LSTM output.

We used the torch.nn package in PyTorch to construct a neural network. The LSTM procedure that we used consists of two hidden layers, each with 64 nodes. The data subsequently entered a full connection layer, and a 1×8 vector was finally output as the feature vector of a single mutated sequence.

Step 2. Clustering by SOM

The SOM consists of two kinds of layers: an input layer and a competition layer[27]. In the SOM process of the LSTM-SOM model, the feature vector obtained from the LSTM process is used as the input.

The settings included 200 units in the SOM competition layer. For each input vector, the Euclidean distance between it (x) and each unit in the competition layer (w_j) was calculated as follows:

$$d_j(x) = \sqrt{\sum_{i=1}^{D} \left(x_i - w_{j_i}\right)^2}$$

The unit closest to x is recorded as w_{min}, and the distance between w_{min} and each other competition layer unit is calculated as follows:

$$d_j(w_{min}) = \sqrt{\sum_{i=1}^{D} \left(w_{j_i} - w_{min_i}\right)^2}$$

A threshold of S was set in the process of training. If $d_j(w_{min}) \leq S$, w_j will move in the direction of x; otherwise, w_j will move in the opposite direction. The transportation distance decays with an increase in $d_j(w_{min})$. The neighborhood function refers to the Gaussian function:[28]

$$D\left(w_j\right) = e^{-\frac{d_j(w_{min})^2}{2\pi\sigma^2}}$$

In the neighborhood function, σ is a constant that affects the amplitude of transportation distance decay. The update vector is as follows (where L is the learning rate of SOM):

$$\Delta\left(w_j\right) = \begin{cases} L \times D(w_j) \times (w_j - x) & d_j(w_{min}) \leq S \\ -L \times D(w_j) \times (w_j - x) & d_j(w_{min}) > S \end{cases}$$

To avoid overfitting, the units in the SOM competition layer are updated after each training batch of 100 samples. The samples in each batch are selected randomly from different cancers. To change the discrete status of the input vectors and cause similar input vectors to aggregate, the input units are updated in the opposite direction (x is the input vector):

$$x(new) = x + \sum_{j=1}^{200} \Delta(w_j)$$

Step 3. Train the LSTM model

The updated $x(new)$ is used as the label to train the LSTM network. In this way, the output feature vectors of LSTM with similar features can be gradually closed.

The above three steps are repeated until a clear, stable classification is obtained.

Obtain the classification

We adjusted the parameters to optimize the LSTM-SOM model. The units in the competition layer of the SOM were sorted according to the distance to w_{min}. S was set as the distance of unit rank 40 (5% of entire units) to w_{min}. The updated input data were used as labels to train the LSTM model for 2 iterations. The LSTM learning rate was set as 0.001. The SOM learning rate was set as 0.005.

Through the adjustment of parameters, 2 classes could be obtained after one round of training. After 3 rounds of training, a total of 8 clustered classes were obtained. It was observed that there were 2 classes showing significantly larger sample sizes than the other classes. Therefore, an additional round of clustering was carried out in the 2 classes. Finally, we obtained 10 classes of mutated sequences.

Analysis of clinical features

In the analysis of clinical features, measurement data were expressed as the mean $\pm$ standard deviation. In the analysis of differences between groups, an independent-samples T test (number of groups = 2) or analysis of variance (ANOVA) (number of groups > 2) was used. Chi-square analysis was used for difference testing of enumeration data. $P<0.05$ was considered to indicate a statistically significant difference.

The log-rank test was used to analyze the difference in survival between different groups. In some cases, there were many groups of patients involved in the survival analysis between groups, so a heat map was used to show differences in survival between groups. The difference in survival was reflected in the color. In the survival analysis of different MBs in a single mutant gene with a high incidence, some patients exhibited multiple mutations in the same gene and could be grouped into multiple groups. Such patients were excluded in the survival analysis between groups but were included in the survivorship curve.

Clustering of patients according to the MB composition

Patients were clustered according to their MB composition. Each kind of MB was reflected as the percentage of the entire MB in one patient. The K-means method was used for clustering performed by the K-means method in the scikit-learn package. An "elbow method" was used to evaluate the K value (number of clustered groups).[29, 30] The K value evaluated in different cancers, and the entire sample was generally between 5-8. After comparing the clustering results, K=7 was selected as the class number for K-means clustering.

Code available

All mathematical methods were performed with Python. The code for the pretreatment of TCGA data and the construction, training and testing of the model is stored at *https://github.com/FruedDolce/SBS_CLUSTER/*. For clinical data analysis, patient clustering, survival analysis and drawing, the code is stored at *https://github.com/FruedDolce/SATA/*. All the code is open source and freely available.

Results

SBS clustering via the LSTM-SOM model

A total of 2,141,527 somatic SBS data points from 9596 patients were collected from the TCGA database. For each SBS sample, 100 flanking bases (50 bases at the 5' end and 50 at the 3' end) were included in the LSTM training data.

In brief, our LSTM-SOM model functions by extracting the features of mutant sequences via the LSTM network and then taking the generated feature vector as the input data for the SOM. In particular, not only will the units in the competitive layer of SOM be refreshed, but the input data generated by LSTM will also be adjusted in the opposite direction. Then, the refreshed input data are used as the labels to train the LSTM model (Figure 1A). The above steps were repeated until the LSTM outputs formed clear classifications.

Mutated sequences were clustered into 2 types after one round of training. We obtained 8 classes of mutated sequences (for easy understanding, mutated sequences with different features clustered by LSTM-SOM are referred to as mutation blots, MBs) after 3 rounds of training. Then, an additional round of training was performed for 2 classes of MB with a significantly larger number of samples and ultimately revealed 10 classes of MBs, recorded as MB 1-MB 10 (Figure 1B).

Characteristics of different MBs

Following the principle of complementary base pairing, 4 kinds of bases form 6 classes of base substitutions: C>A, C>G, C>T, T>A, T>C, and T>G, where base substitutions are represented by the pyrimidine residue of the base pair. Among the 10 classes of MBs clustered

by LSTM-SOM, 4 contained a single kind of mutation (MB 7: C>A; MB 8: C>T; MB 9: T>A; MB 10: T>C). The other 6 classes contained multiple types of mutations. The dominant mutations in some classes of MB were similar (MB 4 and MB 8; MB 5 and MB 7; MB 2 and MB 6). No one kind of mutation was contained in a single class of MBs (Figure 2 and Table S1). With an increase in the distance from the mutation site, the proportions of the four bases tended to become balanced.

The clustering results were strongly influenced by the flanking bases of the mutation site. Differences in flanking bases could be observed in other classes of MBs with similar mutation features, such as MB 2 and MB 6, MB 5 and MB 7, MB 4 and MB 8 (Figure 2). In the analysis of cancers with a high incidence, the composition of the bases in the mutation site and the flanking sites of each MB basically followed that in the entire sample (Figure S1).

MBs in different cancers

Significant differences in the composition of MBs existed among cancers with different pathologies (Figure 3A). Overall, MB 4, MB 5, MB 7 and MB 8 accounted for much greater percentages of the MBs than the other classes of MBs, especially MB 4 and MB 8. Malignant mesenchymal tumors (sarcomas) seemed to present a higher percentage of MB 2 and MB 6 than epithelial malignant tumors (carcinomas). Transitional cell carcinoma of the urinary tract showed a distinctly higher MB 1 incidence than other cancers. In general, a high incidence of MB 4 and MB 8 was also observed in other pathologic types of cancers that are considered more likely to be caused by external mutagenic exposure, such as melanoma, squamous cell carcinoma (SCC), and transitional cell carcinoma.

The components of MBs varied in different cancers, and some cancers presented distinct features (Figure 3B). The proportions of MBs in different cancers were influenced by the pathological type to some extent. For example, cancers of the skin and lymph nodes showed extraordinarily high proportions of MB 4 and MB 8 but small proportions of other MBs. In both types of cancers, melanoma is the major pathologic type. Lung cancer presented high proportions of MB 5 and MB 7. Among the 2 major pathological types of lung cancer, adenocarcinoma (AC) exhibited much higher proportions of MB 5 and MB 7 than did SCC (Figure S2). However, for the same pathological type, differences in the MB composition could be observed in different cancers (Figure S2). For example, AC of the colon presented higher proportions of MB 4 and MB 8 than did AC of the lung. SCC of the lung exhibited more MB 5 and MB 7 than SCC in the head and neck.

We then counted the high-frequency mutated genes in different classes of MBs. In most classes of MBs, the frequency of genes that were commonly mutated in malignant tumors was relatively high. Distinct features existed in some classes of MB. The proportion of TP53 was relatively low in MB 7 and MB 10. Remarkably, BRAF was the most common mutated gene in MB 9. A higher proportion of PIK3CA mutations was observed in MB 8 and MB 10 than in the other classes of MBs (Table S2). More distinct features could be observed when considering specific cancers. For example, in kidney cancer, the frequency of VHL mutations ranked high in MB 3, MB 5, MB 7 and MB 9. In skin and thyroid cancers, BRAF mutations were common in MB 9 but not in the other classes of MBs (Figure S3).

To further study the influence of certain genes with different MBs on survival, we analyzed the survival of patients who exhibited mutations in genes with high mutation frequencies (TTN, MUC16, TP53, etc.). Patients were grouped according to the MB

classification of specific genes. The results of survival analysis showed a significant correlation between survival and MB for specific genes. Patients carrying genes of MB 4 and MB 8 usually showed better survival. In contrast, MB 1, 6, and 9 in a gene could predict worse survival (Figure 4, S4 and S5).

Relationship between MBs and clinical features of cancer patients

Analysis was performed to determine the relationship between MBs and the clinical features of patients. The change in MBs showed a nonmonotonic trend with patient age. The percentages of MB 2, MB 5, and MB 7 in single patients increased with age within the first interval but decreased when age exceeded the threshold. This trend was reversed for MB 4 and MB 8. Female patients were likely to show higher percentages of MB 2, MB 3, MB 5, MB 6 and MB 10, while male patients exhibited higher percentages of MB 4, MB 8 and MB 9. No apparent rule regarding the relationship between the weight and MB composition of a patient was observed (Figure 5).

Correlations with cancer stages could be observed in most classes of MBs. The proportions of MB 3, MB 7 and MB 9 showed a decreasing trend with increasing T and N stages. In contrast, MB 4 and MB 8 had a positive relationship with T and N stages. MB 2 and MB 6 exhibited a remarkably high prevalence in N3 patients. Interestingly, although carcinoma in situ (Tis) is considered to be a relatively less malignant diagnosis,[31] the proportions of MBs in Tis seemed to be inconsistent with the MB change trends in T stages. This finding may have resulted from the small sample size of Tis ($N = 8$) (Figure 5).

In most kinds of cancers, the MB composition at different ages basically followed the pattern shown in the total samples. This suggests that the similarity of the cancer biology of

young and old patients requires further study. Regarding cancer staging, T and M stages showed obvious tendencies basically consistent with those for the total sample. Stomach cancer and colon cancer, in particular, showed opposite MB tendencies in T and N stages compared with the entire sample and with other cancers with high incidence. This suggests that local and lymph node progression in gastrointestinal cancers may exhibit distinct mechanisms (Figure S6-S9).

MBs and survival of cancer patients

A K-means clustering method was used to classify patients according to MB composition. Patients were clustered into 7 classes and designated as Classes 1-7. Each class of patients presented distinct characteristics of the composition of MBs. High proportions of MB 5 and MB 7 were observed in Class 1 patients. Extraordinarily high percentages of MB 4 and few other kinds of MBs are were found in Class 2 patients. In Class 3 patients, the percentages of all kinds of MBs were relatively balanced. Class 4 patients exhibited an extraordinarily high percentage of MB 8 but relatively lower proportions of other MBs. The proportions of MB 3 and MB 10 were higher in patients of Class 5 than in other classes of patients. An obvious feature of Class 6 patients was a high proportion of MB 4 and MB 1. Patients of Class 7 presented high proportions of MB 4 and MB 8, and the percentages of MB 4 and MB 8 were relatively balanced (Figure 6A).

Significant differences in survival were observed in different classes of patients (Figure 6B). In the pairwise comparison of survival, patients with Classes 2, 4, and 5 showed better survival, and patients with Classes 1, 3, 6, and 7 showed worse survival (Figure 6C). In the analysis of specific cancers, survival in different classes of patients generally followed the results obtained for the total sample (Figure S10). Class 3 patients, in particular, seemed to show poor

survival for most of the analyzed cancers. These results suggested that a balanced MB composition may predict poor survival in patients.

Patients of different classes showed distinct clinical features (Figure 6D and E). According to AJCC staging, a significantly lower proportion of stage IV patients and a higher proportion of stage I patients were observed in Classes 4 and 5, which may be related to the better survival of these 2 classes of patients. Class 6 patients showed the highest percentage of AJCC stage 4 and lowest percentage of AJCC stage I, which may be the reason for the poor survival of these patients. Patients of Class 3 were found to present significantly greater ages and higher weights. These factors may be partly responsible for the poor survival of Class 3 patients. Class 1 patients exhibited a high percentage of AJCC stage 1 and a low percentage of stage IV. Therefore, further study is still needed to determine the mechanism causing Class 1 patients to show poor survival.

Discussion

Previous studies have identified a variety of mutation signatures that may be associated with different triggers involved in various mutation processes and result in differing biological behaviors of cancers[2,8,14-20,32-35]. It is suggested that both the mutated and flanking bases influence somatic mutation characteristics. However, the relationship between sequence feature and its influence on cancer characteristics is still not adequately explained.

LSTM provided us with a method for extracting the features of mutated sequences across a wider spatial scope[36]. A follow-up SOM method can then be used to discover internal relationships between the extracted features and ultimately obtain different categories of mutant sequences. In this way, we explore the interconnections between mutant bases and flanking

sequences. Results of our study show that the same mutation type can be clustered into different classes according to their contexts. MB may comprehensively reflect the difference in cancers in both locations and pathological types. These results suggest that different MB may reflect different mutation processes involving different mechanisms. As the mechanism of machine learning models is difficult to explain.[37] It is meaningful to explore the mechanism whereby LSTM-SOM functions, improving the interpretability of the LSTM-SOM model and explaining the formation of different classes of MB to determine how sequences of bases affect the characteristics of cancers. Moreover, molecular biology methods are helpful for explaining the differences in the characteristics of MBs. Different MB may also be involved in complex changes in chromosomal three-dimensional conformation.

Our study suggests that even among patients with different MBs in the same gene, significant differences may exist in their clinical features and survival. In the analysis of MB composition in patient level, clinical features and survival vary among patients of different classes. Due to the natural differences in cancer process, a further analysis of the relationship between MBs and distinctive clinical features in specific kinds of cancer may provide more information about how MBs are related to cancer etiology, processes, prognosis and drug susceptibility.

There are still some constraints and limitations of this study. The clustering results obtained with the LSTM-SOM model are largely dependent on the selection of SOM parameters (especially the neighborhood function parameter), and the threshold value determines whether units in the SOM competition layer move to the target. There exists the possibility that when training with other parameters, the classification obtained may be related to clinical features that were not included in this study. In summary, this study provided a method for classifying the

characteristics of mutant sequences and found the influences of mutation sequences on cancer characteristics. Further study of the mechanism of MBs related to cancer characteristics is suggested.

Acknowledgements

Funding

This study is supported by Xijing Hospital science foundation (XJZT19ML38)

Declaration of potential conflicts of interest

To the best of our knowledge, the named authors have no conflict of interest, financial or otherwise.

References

1 Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011; 144: 646-674.

2 Alexandrov LB, Nik-Zainal S, Wedge DC et al. Deciphering signatures of mutational processes operative in human cancer. Cell Rep 2013; 3: 246-259.

3 Cooke MS, Evans MD, Dizdaroglu M et al. Oxidative DNA damage: mechanisms, mutation, and disease. FASEB J 2003; 17: 1195-1214.

4 Pfeifer GP. Environmental exposures and mutational patterns of cancer genomes. Genome Med 2010; 2: 54.

5 Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009; 458: 719-724.

6 Peña-Diaz J, Bregenhorn S, Ghodgaonkar M et al. Noncanonical mismatch repair as a source of genomic instability in human cells. Mol Cell 2012; 47: 669-680.

7 Cappell MS. Pathophysiology, clinical presentation, and management of colon cancer. Gastroenterol Clin North Am 2008; 37: 1-24, v.

8 Alexandrov LB, Kim J, Haradhvala NJ et al. The repertoire of mutational signatures in human cancer. Nature 2020; 578: 94-101.

9 Remon J, Steuer CE, Ramalingam SS et al. Osimertinib and other third-generation EGFR TKI in EGFR-mutant NSCLC patients. Ann Oncol 2018; 29: i20-i27.

10 Stintzing S, Wirapati P, Lenz HJ et al. Consensus molecular subgroups (CMS) of colorectal cancer (CRC) and first-line efficacy of FOLFIRI plus cetuximab or bevacizumab in the FIRE3 (AIO KRK-0306) trial. Ann Oncol 2019; 30: 1796-1803.

11 von Minckwitz G, Huang CS, Mano MS et al. Trastuzumab emtansine for residual invasive HER2-positive breast cancer. N Engl J Med 2019; 380: 617-628.

12 Sawrycki P, Domagalski K, Cechowska M et al. Relationship between CYP1B1 polymorphisms (c.142C > G, c.355G > T, c.1294C > G) and lung cancer risk in Polish smokers. Future Oncol 2018; 14: 1569-1577.

13 Zerp SF, van Elsas A, Peltenburg LT et al. p53 mutations in human cutaneous melanoma correlate with sun exposure but are not always involved in melanomagenesis. Br J Cancer 1999; 79: 921-926.

14 Nik-Zainal S, Alexandrov LB, Wedge DC et al. Mutational processes molding the genomes of 21 breast cancers. Cell 2012; 149: 979-993.

15 Poon SL, Pang ST, McPherson JR et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. Sci Transl Med 2013; 5: 197ra101.

16 Alexandrov LB, Jones PH, Wedge DC et al. Clock-like mutational processes in human somatic cells. Nat Genet 2015; 47: 1402-1407.

17 Nik-Zainal S, Davies H, Staaf J et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 2016; 534: 47-54.

18 Petljak M, Alexandrov LB. Understanding mutagenesis through delineation of mutational signatures in human cancer. Carcinogenesis 2016; 37: 531-540.

19 Mimaki S, Totsuka Y, Suzuki Y et al. Hypermutation and unique mutational signatures of occupational cholangiocarcinoma in printing workers exposed to haloalkanes. Carcinogenesis 2016; 37: 817-826.

20 Polak P, Kim J, Braunstein LZ et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. Nat Genet 2017; 49: 1476-1486.

21 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997; 9: 1735-1780.

22 Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. Neural Comput 2000; 12: 2451-2471.

23 Tayara H, Chong KT. Improving the quantification of DNA sequences using evolutionary information based on deep learning. Cells 2019; 8: 1635.

24 Liu Q, Fang L, Yu G et al. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. Nat Commun 2019; 10: 2449.

25 Zhou J, Lu Q, Xu R et al. EL_LSTM: prediction of DNA-binding residue from protein sequence by combining long short-term memory and ensemble learning. IEEE/ACM Trans Comput Biol Bioinform 2018; 17: 124-135.

26 Markey MK, Lo JY, Tourassi GD et al. Self-organizing map for cluster analysis of a breast cancer database. Artif Intell Med 2003; 27: 113-127.

27 Furukawa T. SOM of SOMs. Neural Netw 2009; 22: 463-478.

28 Kolasa M, Długosz R, Pedrycz W et al. A programmable triangular neighborhood function for a Kohonen self-organizing map implemented on chip. Neural Netw 2012; 25: 146-160.

29 Fukuoka Y, Zhou M, Vittinghoff E et al. Objectively measured baseline physical activity patterns in women in the mPED trial: cluster analysis. JMIR Public Health Surveill 2018; 4: e10.

30 Iuliia Pavlova, Dmytro Zikrach, Dariusz Mosler et al. Determinants of anxiety levels among young males in a threat of experiencing military conflict-Applying a machine-learning algorithm in a psychosociological study. PLoS One. 2020; 7: e0239749.

31 Fosså SD, Aass N, Heilo A et al. Testicular carcinoma in situ in patients with extragonadal germ-cell tumours: the clinical role of pretreatment biopsy. Ann Oncol 2003; 14: 1412-1418.

32 Meier B, Cooke SL, Weiss J et al. C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. Genome Res 2014; 24: 1624-1636.

33 Huang MN, Yu W, Teoh WW et al. Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors. Genome Res 2017; 27: 1475-1486.

34 Nik-Zainal S, Kucab JE, Morganella S et al. The genome as a record of environmental exposure. Mutagenesis 2015; 30: 763-770.

35 Kucab JE, Zou X, Morganella S et al. A compendium of mutational signatures of environmental agents. Cell 2019; 177: 821-836.e816.

36 Sahin S, Kozat S. Nonuniformly sampled data processing using LSTM networks. IEEE Trans Neural Netw Learn Syst 2018; 30: 1452-1461.

37 McCloskey K, Taly A, Monti F et al. Using attribution to decode binding mechanism in neural network models for chemistry. Proc Natl Acad Sci USA 2019; 116: 11624-11629.

Figure captions

Figure 1 Training process of the LSTM-SOM model. A: method of LSTM-SOM unsupervised neural network. Details of the LSTM-SOM model are described in the methods. B: cluster result in different stage. Three of the eight dimensions in LSTM output vectors are shown in the space rectangular coordinate system.

Figure 2 Mutation type and composition of flanking bases in different MBs. Each bar except for "Reference Allele" and "Mutation Allele" represents one flanking genetic locus. Bars on the left of "Reference Allele" represent bases on the 5' end of the mutation site, and bars on the right of "Mutation Allele" represent bases on the 3' end of the mutation site.

Figure 3 Quantity and proportion of MBs in different cancers. A: quantity and proportion of each MB in cancers of different pathology; B: quantity and proportion of each MB in cancers of different site. The left subgraph shows the proportion of each MB in data points from different kinds of cancers. The right subgraph shows the quantity and proportion of different MBs in patients. Differences in quantity are reflected in the size of the point, and differences in proportion are reflected in the color of the point.

Figure 4 Relationship between patient survival and MB in genes with high mutation frequencies. The top 4 most frequently mutated genes are shown (other genes with high mutation frequencies are shown in Figure S5). For each gene, the left subgraph shows the proportion of MB in all mutation data points from different cancers; the middle subgraph shows the *P* value of the log-rank test between groups in the whole population; and the right subgraph shows the *P* value of the log-rank test between cancers with high incidence. Only *P* values less than 0.05 are shown in the heat map.

Figure 5 MBs in patients with different clinical features. *: $P < 0.05$ in the t test or ANOVA between groups. **: $P < 0.005$ in the t test or ANOVA between groups.

Figure 6 Clusters of patients by the proportion of MBs and differences in survival and clinical features between classes. A: characteristics of MB composition in patients of 7 classes clustered by the K-means method; each line represents one patient. B: survivorship curve of each class of patients. C: log-rank test between classes; differences in the *P* value are reflected in color. D, E: clinical features of patients in different classes. *: $P < 0.05$ in the ANOVA or the chi-square test; **: $P < 0.005$ in the ANOVA or the chi-square test.

Figure S1 Mutation type and composition of flanking bases of each MB in cancers with high incidence.

Figure S2 Quantity and proportion of MBs in cancers with high incidence and multiple types of pathology. For each cancer, the left subgraph shows the proportion of different MBs in SBS mutation data points, and the right subgraph shows the quantity and proportion of different MBs in patients. Differences in quantity are reflected in the size of the point, and differences in proportion are reflected in the color of the point.

Figure S3 Genes with high mutation frequency in different MBs in cancers with high incidence.

Figure S4 Survivorship curve of patients with different MBs in TTN, MUC16, TP53, and SYNE1.

Figure S5 Relationship between patient survival and MB in genes with high mutation frequencies. Genes ranked 5-15 in mutation frequency are shown. For each gene, the left subgraph shows the proportion of MB in all mutation data points from different cancers; and the

right subgraph shows the *P* value of the log-rank test between groups in the whole population. Only *P* values less than 0.05 are shown in the heat map.

Figure S6 Statistics of the proportion of each MB by age in cancers with high incidence. *: $P < 0.05$ in the t test or ANOVA between groups; ** $P < 0.005$ in the t test or ANOVA between groups.

Figure S7 Statistics of the proportion of each MB by sex in cancers with high incidence. *: $P < 0.05$ in the t test or ANOVA between groups; ** $P < 0.005$ in the t test or ANOVA between groups.

Figure S8 Statistics of the proportion of each MB by T stage and N stage in cancers with high incidence. *: $P < 0.05$ in the t test or ANOVA between groups; ** $P < 0.005$ in the t test or ANOVA between groups.

Figure S9 Statistics of the proportion of each MB by M stage and AJCC stage in cancers with high incidence. *: $P < 0.05$ in the t test or ANOVA between groups; ** $P < 0.005$ in the t test or ANOVA between groups.

Figure S10 Log-rank test between classes of different cancers. Differences in the *P* value are reflected in color.