Characteristic of mutant sequences in human cancers

Ji Hongchen[1, *], Li Junjie[2, *], Zhang Qiong[1, *], Yang Jingyue[1], Wang Xiaowen[1], Bai Shuang[1], Tian Fei[1], Zhang Hongmei[1]

1 Department of oncology, Xijing Hospital, Air Force Military Medical University. No.127 West Changle Road, Xi'an City, Shaanxi Prov. China 710032

2 Department of emergency, Xijing Hospital, Air Force Military Medical University. No.127 West Changle Road, Xi'an City, Shaanxi Prov. China 710032

[*] These authors contributed equally to this work.

Corresponding author: Prof. Zhang Hongmei. E-mail: zhm_fmmu@163.com

Abstract

Mutation processes leave different signatures in genes. Previous studies 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. Here, 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, we obtained 10 classes of mutant sequences (named mutation blots, MBs) from 2,141,527 somatic mutations in The Cancer Genome Atlas (TCGA) database. Differences features were revealed among MBs. MBs were related to clinical features, including age, sex, and cancer stage. Different kinds of MBs for specific genes may affect patient survival. Finally, we clustered the patients into 7 classes by MB composition. Significant differences in survival and clinical features were observed among different patient classes. This study provides a novel method for understanding mutant sequences and revealing the extensive relationships among mutant sequences, clinical features, and cancer patient survival.

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. For example, C:G > A:T transversions predominate in smoking-associated lung cancer[12], and CC:GG > TT:AA double nucleotide substitutions are common in UV light-associated skin cancer[13].

SBSs contribute the largest proportion of genetic mutations. Mathematical methods have been used to decipher mutation signatures from somatic mutation catalogs[2,8,15-21]. Some of the identified mutational signatures have been proposed to be associated with certain etiologies (e.g., in the classification system of Alexandrov et al.[2], SBS2 may be caused by tobacco smoking)[2]. 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.

The application of machine learning, especially neural networks, makes it possible to effectively mine information from large amounts of data. 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[22,23]. LSTM has been used to analysis DNA or RNA sequences information in some studies[24-26]. 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[27,28]. 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.

Results

1. 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. Flanking base sequences were obtained from *Homo sapiens* (Human) genome Assembly GRCh38 (hg38).

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. Units in competitive layers of the SOM are then refreshed to edges closer to the distribution of the input data. In particular, after each iteration of SOM in our LSTM-

SOM model, 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 (Fig. 1). Then, the refreshed input data are used as the labels to train the LSTM model. The above steps were repeated until the LSTM outputs formed clear classifications (Methods).

Vectors were used to represent the bases of mutated sequences for training. In the LSTM process, the influence of unit data on the LSTM output results decreased with increasing distance to the ending unit. As the training data included flanking sequences on both sides of the mutation sites, the LSTM process was carried out on both sides of the mutation site in opposite directions. Thus, the mutation site was placed at the end of both sequences to expand its influence on LSTM output and to reflect the difference between the reference allele and mutant allele.

One hundred samples from different cancers were selected randomly in each iteration of training (batch size = 100). To avoid overfitting of the model, the weight of the vectors in the competitive layer was updated after all input data were trained in one batch. Each iteration of training included 2 LSTM iterations and 2 SOM iterations. By adjusting the parameters of the model, mutated sequences were clustered into 2 types after one stage of training. Thus, 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 (Fig. 1).

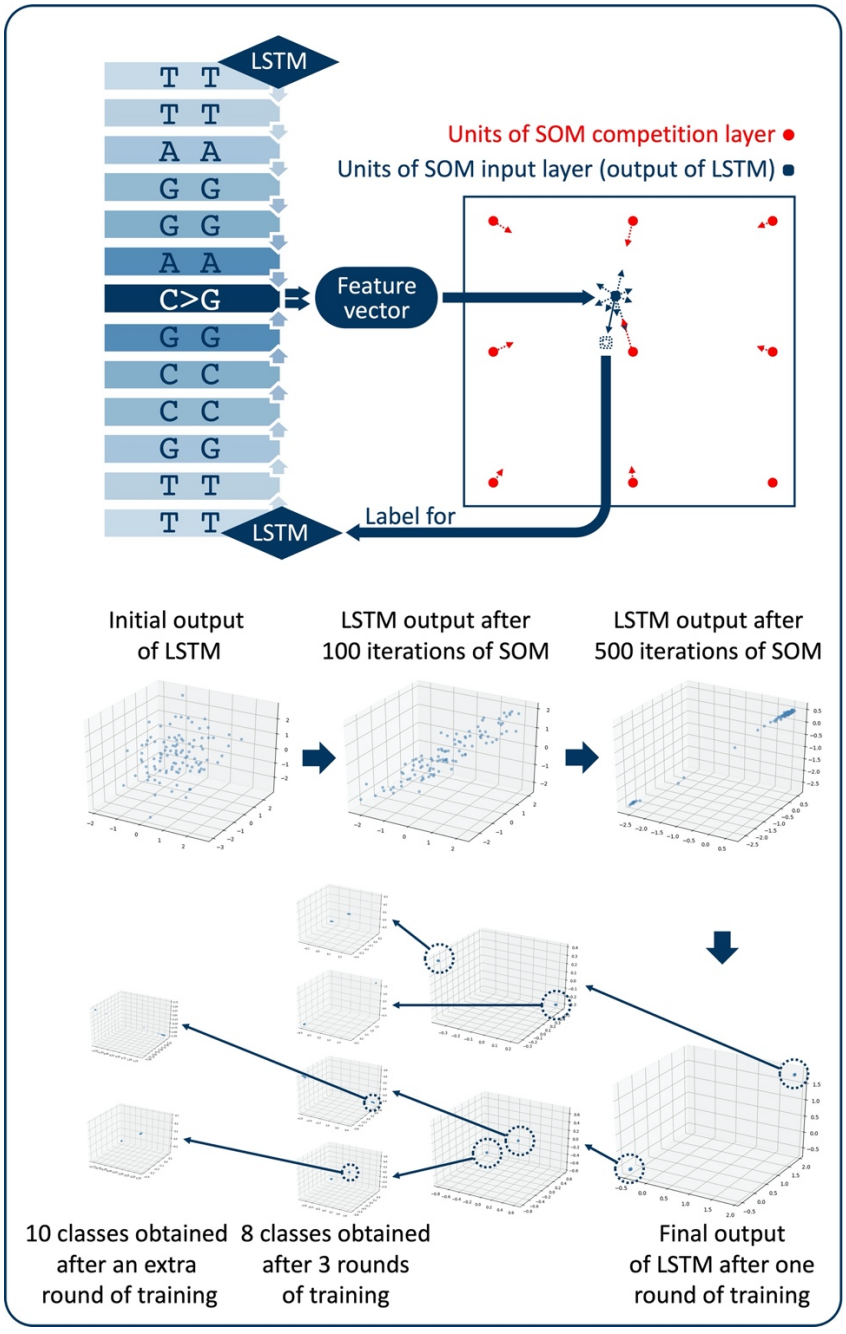


Fig. 1 | Training process of LSTM-SOM model. Details of LSTM-SOM model is described in methods. 2 classification were got in each time of training. 10 classes of mutant sequences were obtained after 3 rounds and an extra round of training, 3 of 8 dimensions are shown in space rectangular coordinate system.

2. 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. Together with the 50 flanking bases on both sides, there are theoretically 6×4^{100} possible classes. 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 5 classes contained multiple types of mutations (MB 1: C>G, T>C, and T>G; MB 2: C>A, C>T, T>A and T>G; MB 3: T>A and T>C; MB 4: C>G and C>T; MB5: C>A, C>G, T>C and T>G; MB 6: C>A, C>T, T>A and T>G). 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 (Fig. 2 and Extended Data Table 1).

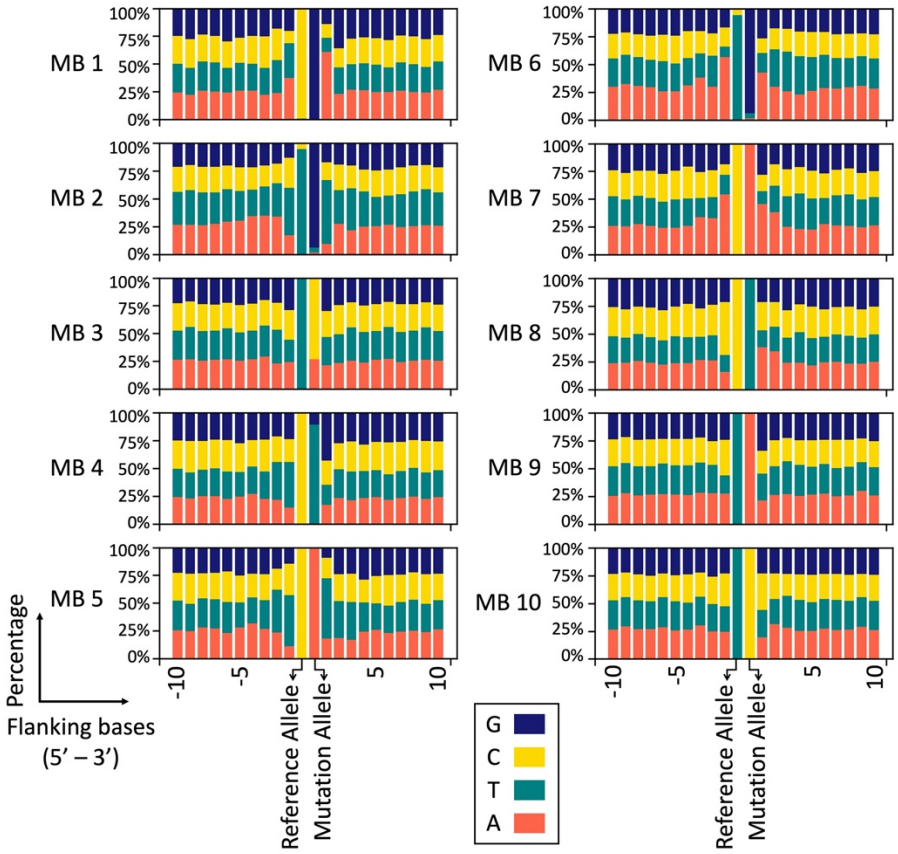


Fig. 2 | Mutation type and composition of flanking bases in different MBs. Each bar, except "Reference Allele" and "Mutation Allele" represents one flanking genetic loci. Bars on the left of "Reference Allele" represent bases on the 5' end of mutation site; bars on the right of "Mutation Allele" represent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

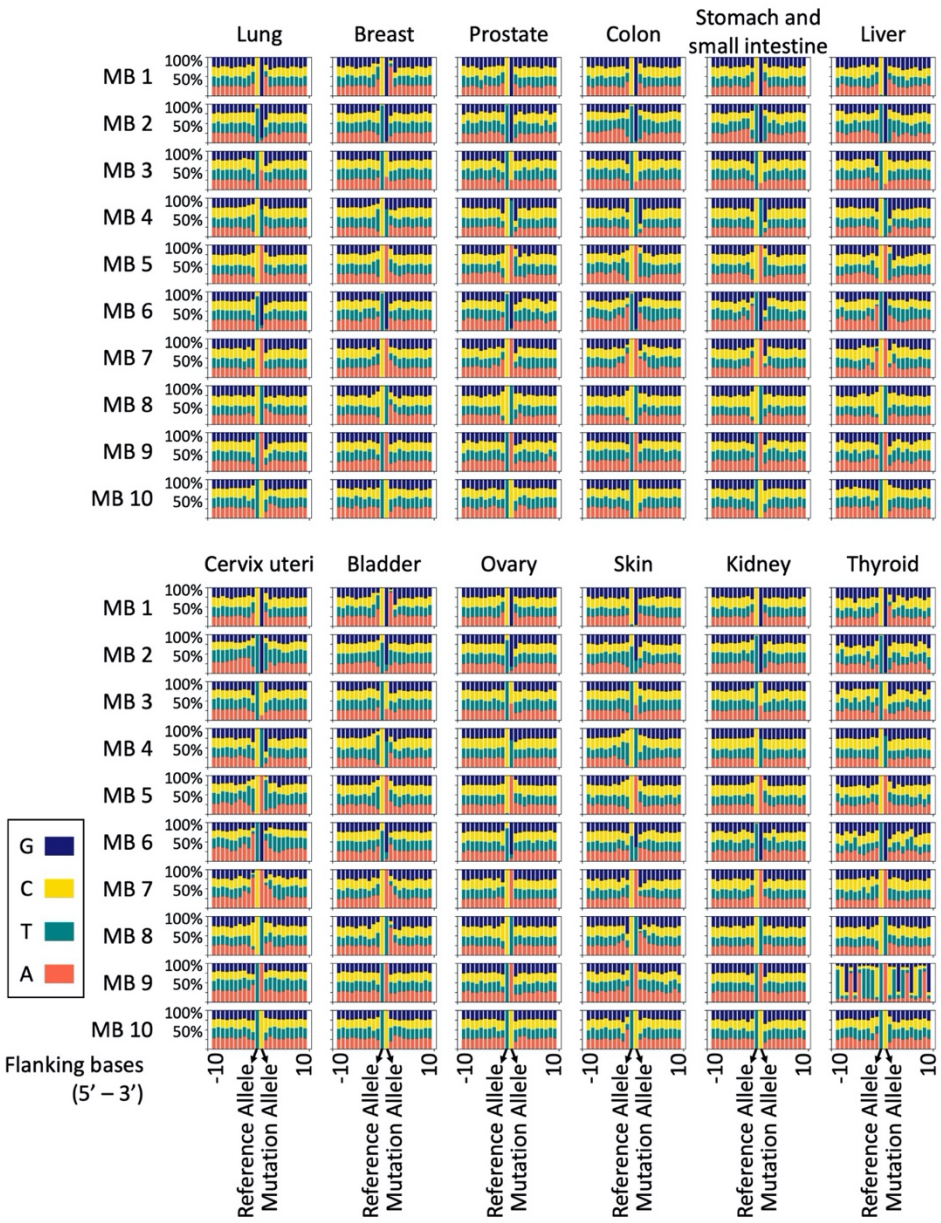
Extended Data Table 1 | Mutation type in each kind of MB.

	Mutation type					
	C>A	**C>G**	**C>T**	**T>A**	**T>C**	**T>G**
MB 1	0	68877	0	0	346	152
MB 2	559	0	2595	239	0	53892
MB 3	0	0	0	41995	114665	0
MB 4	0	70276	587893	0	0	0
MB 5	205678	313	0	0	403	126
MB 6	552	1	2645	210	0	52238

MB 7	201605	0	0	0	0	0
MB 8	0	0	583264	0	0	0
MB 9	0	0	0	40582	0	0
MB 10	0	0	0	0	112421	0

The composition of the bases flanking the mutation sites differed considerably. Generally, units located far from the endpoint had less influence on the LSTM output[23]. This characteristic was reflected in the flanking bases of the mutation site. In all kinds of MBs, the proportions of A, T, C, and G were quite different among the bases near the mutation site of each MB. 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. For example, both MB 5 and MB 7 exhibited C>A mutations, and the flanking bases of MB 5 were dominated by T bases, but MB 7 was dominated by A bases. Differences in flanking bases could also be observed in other classes of MBs with similar mutation features, such as MB 2 and MB 6, MB 4 and MB 8 (Fig. 2). In the analysis of cancers with a high incidence (lung, breast, prostate, colon, stomach, bladder, ovary, cervix uteri, liver, thyroid, skin and kidney cancers), the composition of the bases in the mutation site and the flanking sites of each MB basically followed that in the entire sample (Extended Data Fig. 1). We then counted mutated sequences with high frequency in different kinds of MBs. The results showed differences in the mutated sequences between MBs. For example, when the 3 flanking bases were taken into consideration, the mutated sequences with the highest frequency were TTTCTTT>TTTATTT, ATTCTTT>ATTATTT and ATTCTCT>ATTATCT in MB 5 but AAACAAA>AAAAAAA, AAACAAA>AAAAAAT and AGACAAT>AGAAAAT in MB 7, although the 2 classes of MBs exhibited the same mutation signature. Such differences are common among different kinds of MBs (Extended Data Table 2).



Extended Data Fig. 1 | Mutation type and composition of flanking bases of each MB in different cancers.

Extended Data Tabel 2 | Mutant sequence with high frequency in different class of MBs. 3 flanking bases on each end are calculated.

MB 1		**MB 2**		**MB 3**		**MB 4**		**MB 5**	
Mutant sequence	Count	Mutant sequence	Count	Mutant sequence	Count	Mutant sequence	Count	Mutant sequence	Count
tttCGaga	314	aatTGtat	473	tttTAaaa	263	tttCTctt	1522	tttCAttt	3165
aaaCGaaa	272	tttTGttt	404	tttTCccc	217	cttCTgtg	1375	attCAttt	2704

gaaCGagg	256	aatTGttt	400	aaaTAttt	212	cttCTgag	1298	attCAtct	2079
cctCGagg	252	tttTGtat	336	tttTCttt	204	tttCTctg	1191	tttCAtct	1843
cttCGagg	247	attTGtat	287	aggTCgag	172	tttCTgaa	1142	gttCAttt	1560
tttCGagg	245	aatTGtct	271	atgTCcat	168	cttCTgaa	1127	tttCAttc	1424
ccaCGagg	243	attTGttt	233	ttcTCttt	164	cttCTggg	1040	cttCAttt	1324
tttCGatg	242	aatTGttg	232	tttTCcca	160	attCTgaa	1012	attCAttc	1238
cctCGaga	238	aatTGtgt	224	atgTCcca	159	cttCTctg	1002	tttCAttg	1103
cttCGaga	231	aatTGttc	222	ctgTCcct	152	cagCTgcc	972	attCAttg	978

MB 6		**MB 7**		**MB 8**		**MB 9**		**MB 10**	
Mutant sequence	Count	Mutant sequence	Count	Mutant sequence	Count	Mutant sequence	Count	Mutant sequence	Count
ataTGatt	502	aaaCAaaa	3091	aagCTaaa	1442	ttcTActg	531	cacTCtca	242
aaaTGatt	378	aaaCAaat	2551	cacCTaag	1391	tttTAaaa	278	gggTCaaa	225
ataTGaaa	330	agaCAaat	1867	ctcCTaag	1188	aaaTAttt	214	tggTCcat	168
ataTGaat	300	agaCAaaa	1769	ttcCTaaa	1107	ttaTAaaa	123	atgTCcat	159
aaaTGaaa	280	aaaCAaac	1400	ttcCTaag	1083	aaaTAaaa	100	aaaTCcaa	153
aaaTGaat	233	aaaCAaag	1394	cagCTaaa	1025	taaTAttt	69	aaaTCcat	153
agaTGatt	222	gaaCAaaa	1355	ttcCTaat	1019	tccTAgga	66	ctcTCcct	145
caaTGatt	219	gaaCAaat	1146	cagCTaag	961	taaTAaaa	62	aatTCcat	144
ttaTGatt	205	caaCAaaa	1064	cccCTaag	931	tgcTAggt	61	tggTCcac	141
gaaTGatt	203	caaCAaat	942	gacCTaag	927	gaaTAttt	61	aggTCcag	141

3. MBs in different cancers

Significant differences in the composition of MBs existed among cancers with different pathologies. 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. Cancers of germ cells and the glomus (paragangliomas) exhibited a high proportion of MB 10. An obvious feature of melanomas was the dominance of MB 4 and MB 8. This finding suggested that these classes of MBs may be correlated with ultraviolet light exposure. 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 squamous cell carcinoma (SCC), transitional cell carcinoma, malignant mesothelioma and complex epithelial carcinoma (Fig. 3).

The components of MBs varied in different cancers, and some cancers presented distinct features (Fig. 3). 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. This was consistent with the MB composition in the two pathological types. However, for the same pathological type, differences in the MB composition could be observed in different cancers. 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 MB5 and MB 7 than SCC in the head and neck (Extended Data Fig. 2). Therefore, MB may comprehensively reflect the difference in cancers in both locations and pathological types.

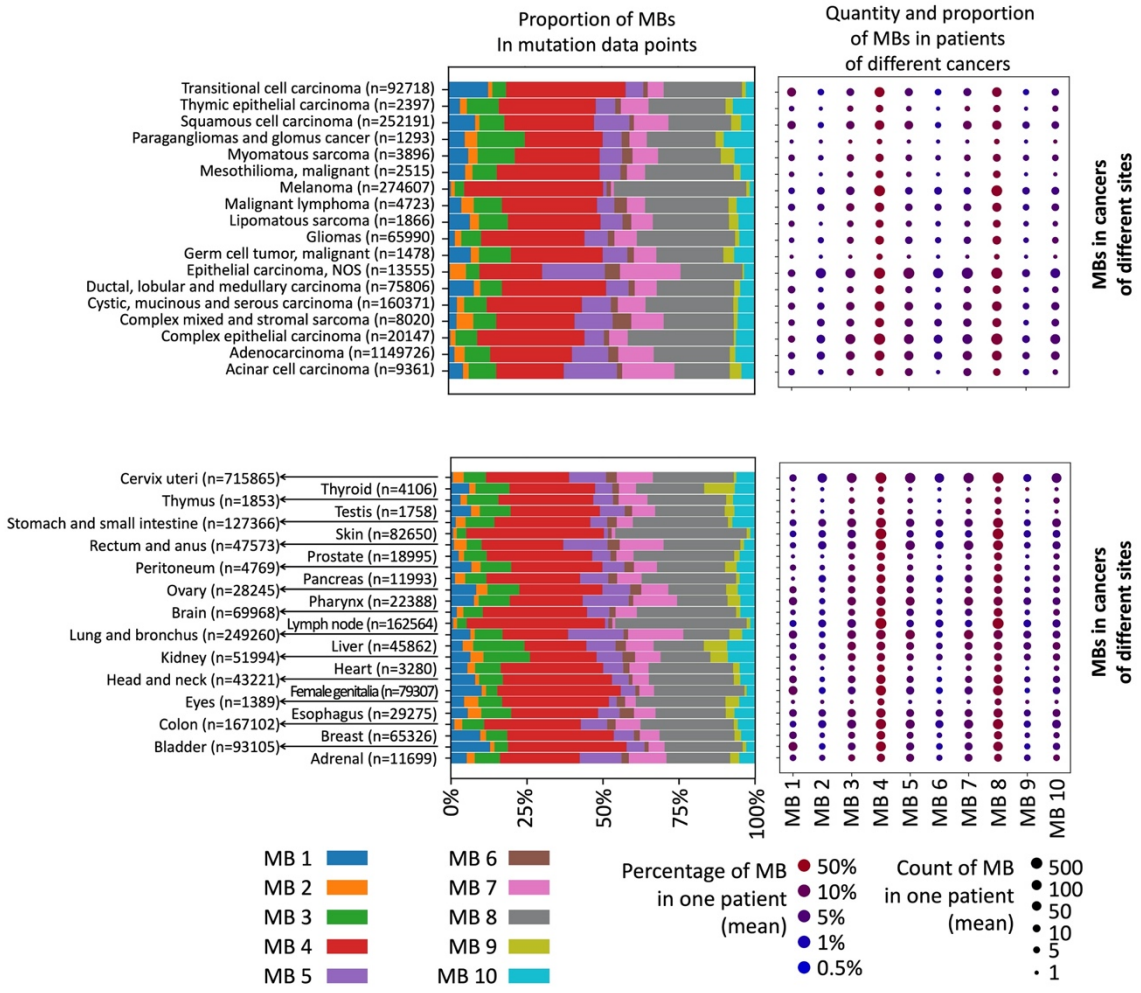
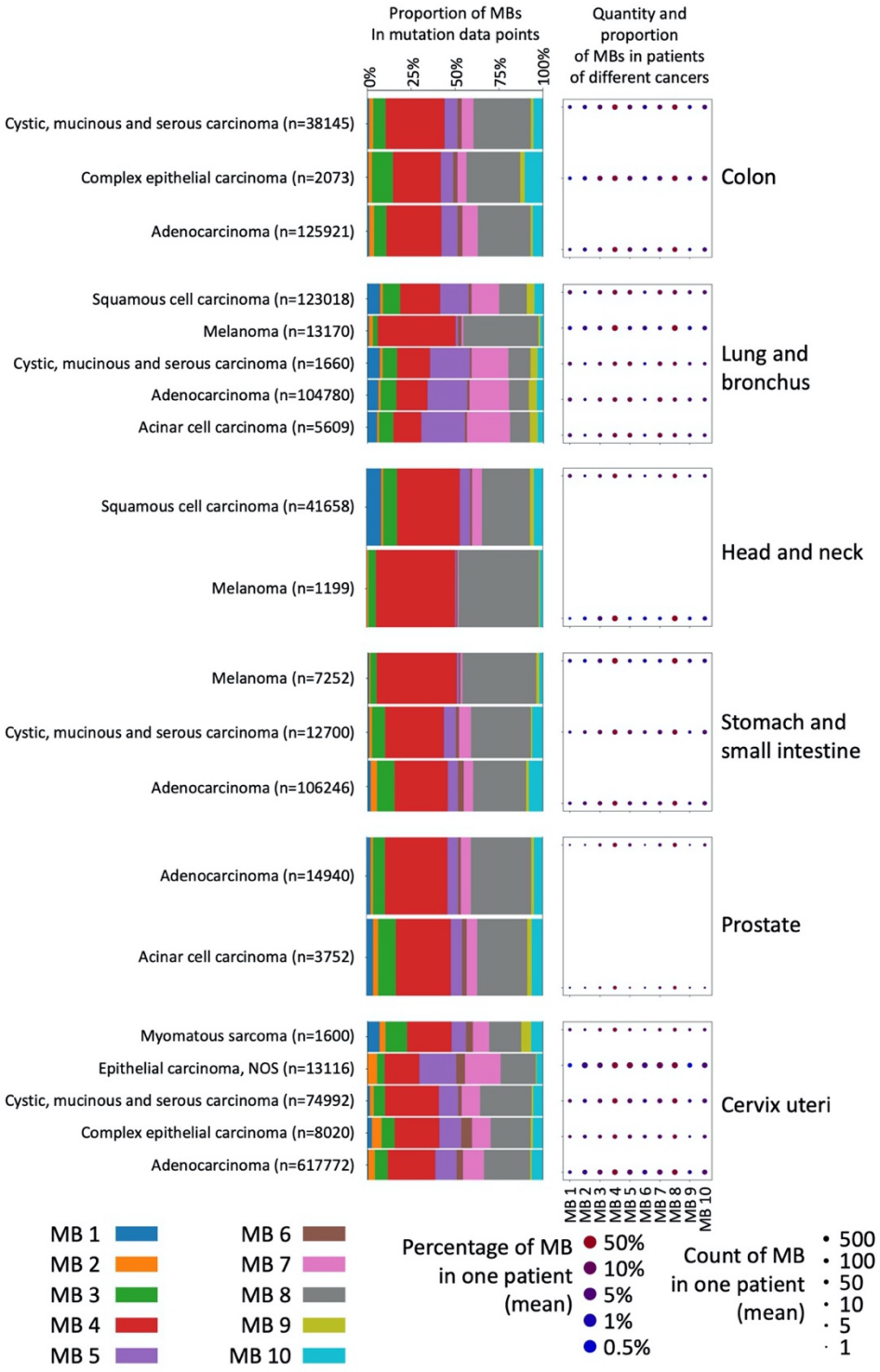


Fig. 3 | Quantity and proportion of MBs in different cancers. The left subgraph shows the proportion of different MBs in total SBS mutation data points of different kinds of cancers. The right subgraph shows the quantity and proportion of different MBs in patients. Differences in quantity are reflected in size of points and differences in proportion are reflected in color of points.



Extended Data Fig. 2 | Quantity and proportion of MBs in cancers with high incidence and consist of multiple types of pathology. For each cancer, the left subgraph shows the proportion of different MBs in SBS mutation data points, the right subgraph shows the quantity and proportion of different MBs in patients. (Differences in quantity are reflected in size of points and differences in proportion are reflected in color of points).

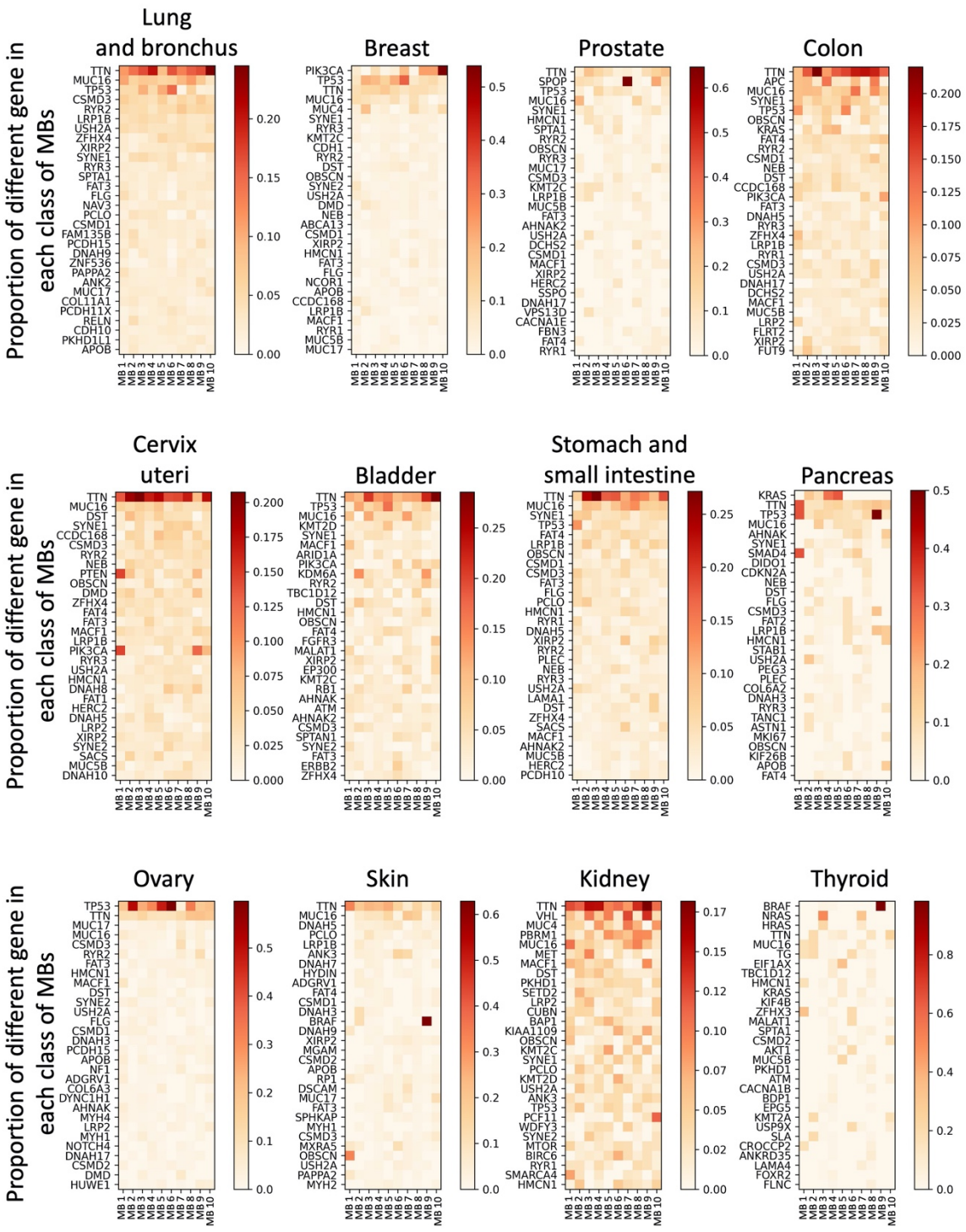
We then counted the high-frequency mutated genes in different classes of MBs. In most classes of MBs, the frequency of genes that were commonly mutated in malignant tumors, such as TTN, TP53, and MUC16, was relatively high. Distinct features existed in some classes of MB. The proportion of TP53 mutations was generally high, but it was relatively low in MB 7 and MB 10. Remarkably, BRAF was the most common mutated gene in MB 9 and takes a high proportion. A higher proportion of PIK3CA mutations was observed in MB 8 and MB 10 than in the other classes of MBs (Extended Data Table 3). More distinct features could be observed when considering specific cancers. For example, in pancreatic cancer, MB 4 and MB 5 contained a higher frequency of KRAS mutations (>1%) than did the other classes of MBs. In kidney cancer (mainly clear cell carcinoma), the frequency of VHL mutations ranked high in MB 3, MB5, MB7 and MB 9. In cancer of skin and thyroid, BRAF mutations were common in MB 9 but not in the other classes of MBs (Extended Data Fig. 3). The differences in the mutated gene composition of different MBs suggested that attention should be paid to the effect on the characteristics of cancer when different MBs occur in the same gene.

Extended Data Table 3 | Gene with high frequency in different classes of MBs (Genes with top 15 mutation rate were shown).

MB 1		**MB 2**		**MB 3**		**MB 4**		**MB 5**	
Gene	Frequency	Gene	Frequency	Gene	Frequency	Gene	Frequency	Gene	Frequency
TTN	0.287%	TTN	0.553%	TTN	0.585%	TTN	0.481%	TTN	0.430%
MUC16	0.248%	TP53	0.232%	TP53	0.242%	MUC16	0.165%	TP53	0.279%
TP53	0.173%	MUC16	0.178%	MUC16	0.218%	TP53	0.137%	MUC16	0.191%
MACF1	0.110%	DST	0.138%	LRP1B	0.122%	DNAH5	0.100%	CSMD3	0.147%
CSMD3	0.094%	PCLO	0.133%	SYNE1	0.120%	OBSCN	0.089%	LRP1B	0.126%
FLG	0.092%	DMD	0.122%	DST	0.113%	SYNE1	0.077%	DST	0.117%
SYNE2	0.089%	LRP1B	0.120%	CSMD3	0.110%	LRP1B	0.076%	KRAS	0.115%
RYR2	0.086%	PTEN	0.108%	PCLO	0.106%	FAT3	0.073%	SYNE1	0.106%
USH2A	0.078%	SYNE1	0.108%	USH2A	0.097%	RYR2	0.070%	RYR2	0.099%
OBSCN	0.075%	CSMD3	0.096%	RYR2	0.096%	PCLO	0.070%	ZFHX4	0.099%
SYNE1	0.072%	SACS	0.094%	FAT4	0.083%	FAT4	0.064%	CCDC168	0.091%
PIK3CA	0.072%	NEB	0.091%	NEB	0.082%	RYR1	0.064%	DNAH5	0.090%
FAT1	0.071%	TSPAN15	0.086%	DMD	0.073%	ZFHX4	0.064%	NEB	0.088%
LRP1B	0.069%	XIRP2	0.082%	NRAS	0.072%	NEB	0.063%	DMD	0.082%
ADGRV1	0.063%	ADGRG4	0.080%	DNAH5	0.070%	FLG	0.061%	MUC17	0.081%

MB 6	**MB 7**	**MB 8**	**MB 9**	**MB 10**

Gene	Frequency	Gene	Frequency	Gene	Frequency	Gene	Frequency	Gene	Frequency
TTN	0.385%	TTN	0.420%	TTN	0.379%	BRAF	1.264%	TTN	0.471%
TP53	0.325%	MUC16	0.253%	MUC16	0.230%	TTN	0.530%	PIK3CA	0.263%
MUC16	0.138%	RYR2	0.146%	TP53	0.118%	MUC16	0.234%	MUC16	0.132%
SYNE1	0.135%	CSMD3	0.133%	PIK3CA	0.101%	CSMD3	0.165%	XIRP2	0.102%
XIRP2	0.102%	SYNE1	0.104%	SYNE1	0.092%	TP53	0.155%	CSMD3	0.095%
CSMD3	0.101%	USH2A	0.101%	OBSCN	0.091%	RYR2	0.145%	LRP1B	0.093%
ABCA13	0.097%	FAT4	0.094%	CSMD1	0.074%	PIK3CA	0.133%	MACF1	0.092%
DNAH8	0.090%	LRP1B	0.094%	RYR2	0.071%	MYBBP1A	0.131%	MUC4	0.091%
USH2A	0.084%	ZFHX4	0.090%	DNAH5	0.069%	XIRP2	0.126%	RYR2	0.082%
LRP1B	0.084%	APOB	0.087%	FAT4	0.067%	LRP1B	0.111%	SYNE2	0.079%
FAT4	0.084%	XIRP2	0.083%	PCLO	0.067%	PCDH15	0.106%	FAT4	0.076%
BIRC6	0.083%	ABCA13	0.082%	DNAH17	0.067%	ABCA13	0.099%	TP53	0.076%
SYNE2	0.081%	RYR3	0.079%	DNAH8	0.066%	ZFHX4	0.091%	ADGRV1	0.075%
DST	0.079%	FLG	0.078%	RYR1	0.064%	MACF1	0.091%	SYNE1	0.074%
UNC13C	0.077%	OBSCN	0.078%	MUC5B	0.063%	FAT3	0.089%	DNAH5	0.073%



Extended Data Fig. 3 | Genes with high mutation frequency of different MBs in cancers with high incidence.

In a given gene, SBSs may occur at different bases with different features and present as different kinds of MBs. Each gene with a high mutation frequency contained multiple

kinds of MBs. In these mutated genes with a high frequency, the composition of MBs varied between different kinds of cancers. Such differences reflected the overall MB composition of each cancer (Fig. 4). 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, DNAH5, USH2A, PIK3CA, SYNE1, 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 with MB 4 and MB 8 mutations usually showed better survival. In contrast, MB 1, 6, and 9 in a gene could predict worse survival (Fig. 4 and Extended Data Fig. 4, 5).

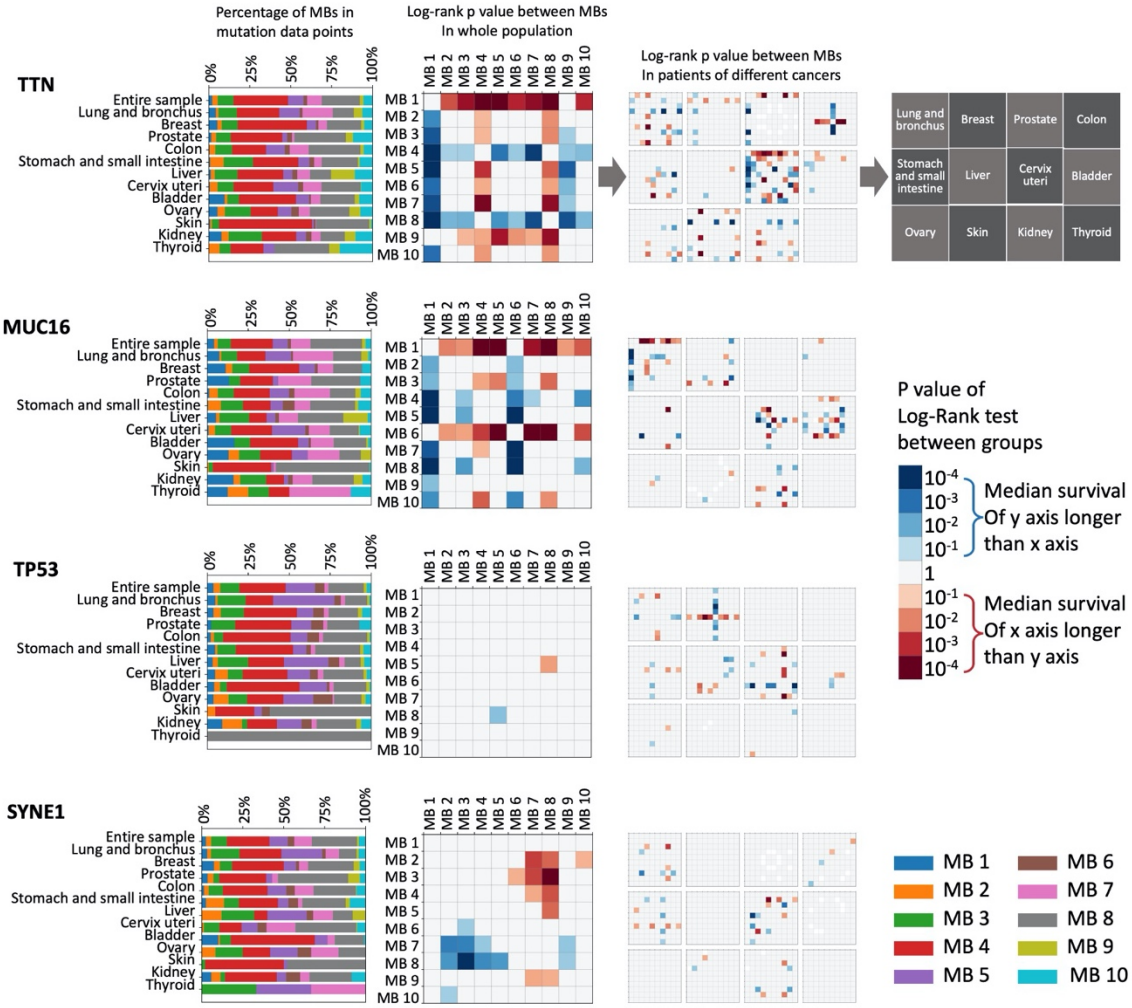
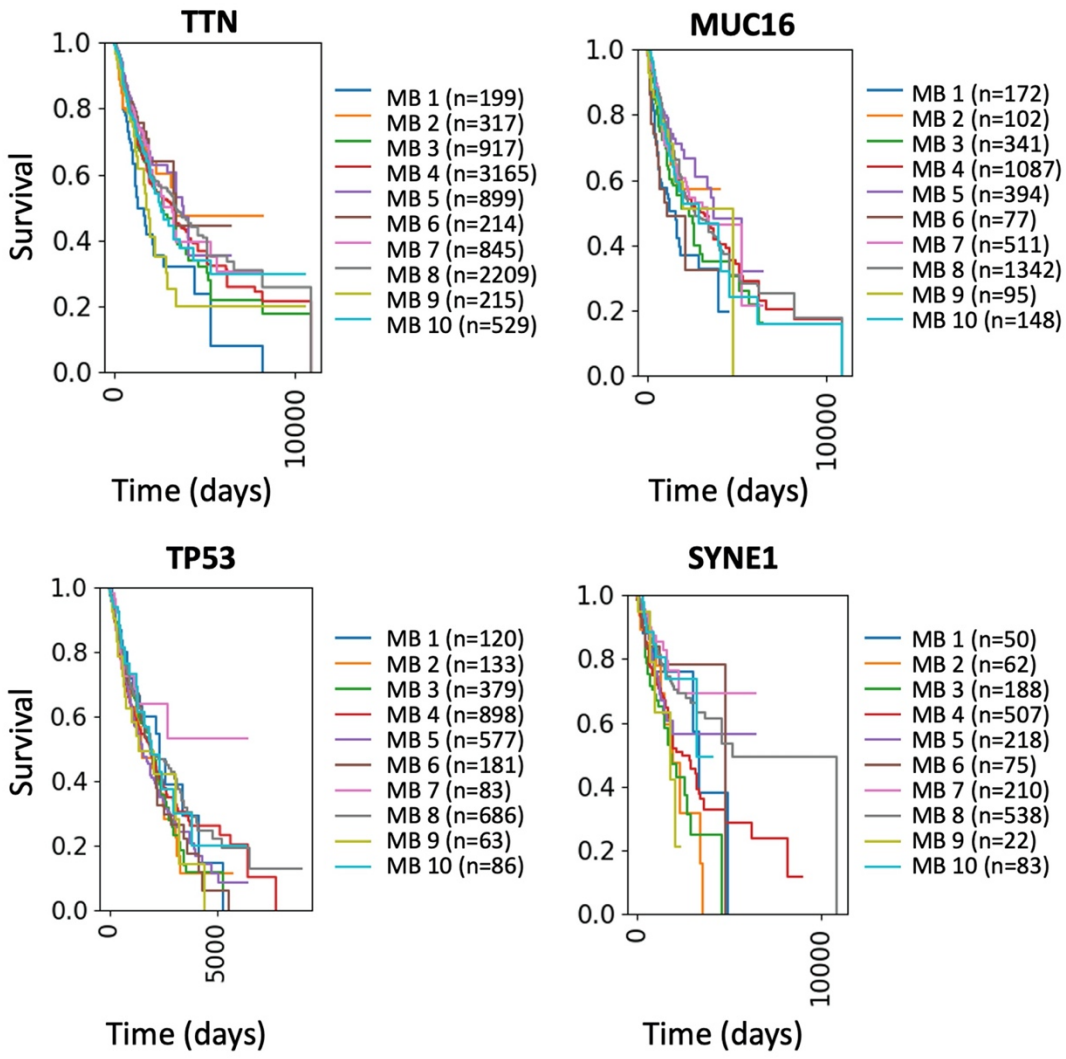
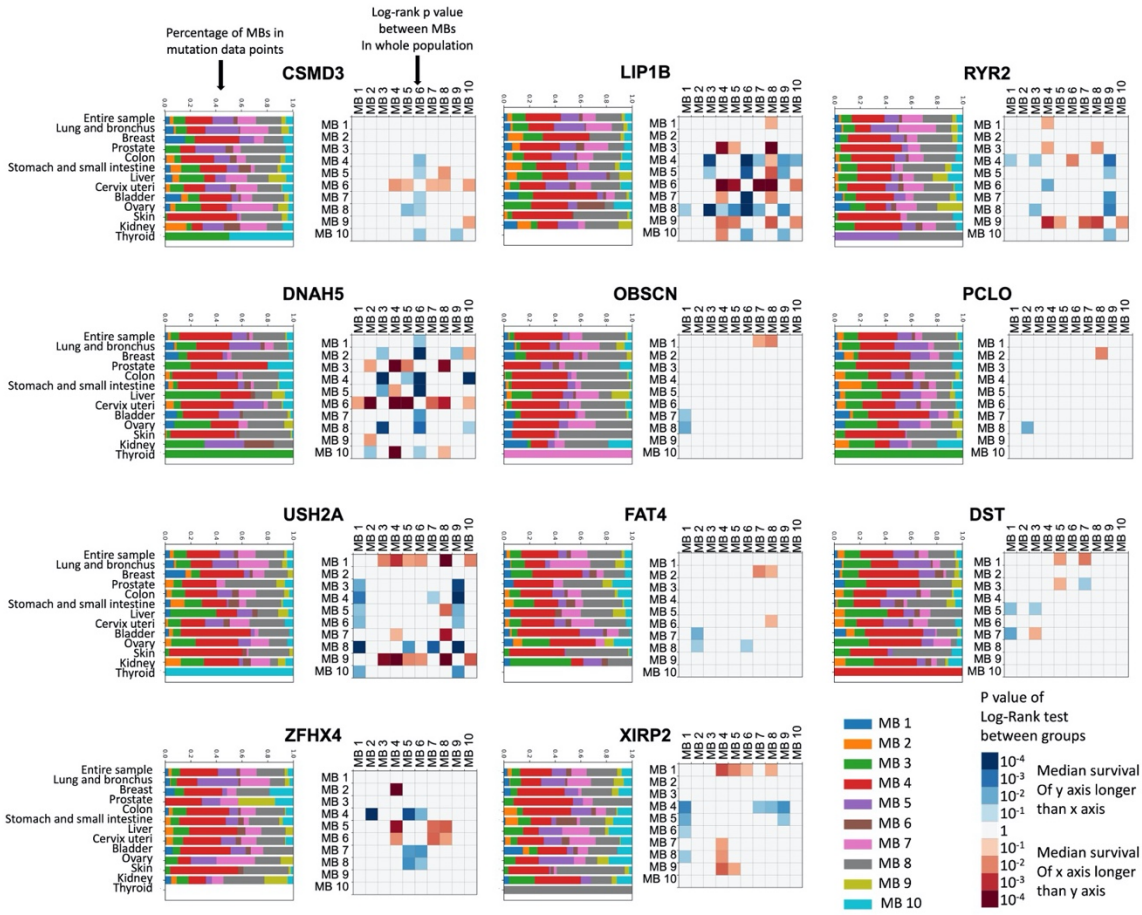


Fig. 4 | Relationship between survival and MB in genes with high mutation frequency. The top 4 high-frequency mutated genes are shown (other genes with high mutation frequency are shown in Extended Data Fig. 5). For each gene, the left subgraph shows proportion of MB in total mutation data points of different cancers; the middle subgraph shows the p value of log-rank test between groups in whole population; the right subgraph shows the p value of log-rank test between groups in different cancers with high incidence, respectively. Only p values less than 0.05 are shown in heat map.



Extended Data Fig. 4 | Survivorship curve of patients with different MBs in TTN, MUC16, TP53, and SYNE1.



Extended Data Fig. 5 | Relationship between survival and MB in genes with high mutation frequency. Genes rank 5-15 in mutation frequency are shown. For each gene, the left subgraph shows proportion of MB in total mutation data points of different cancers; the right subgraph shows the p value of log-rank test between groups in whole population. Only p values less than 0.05 are shown in heat map.

4. Relationship between MB and clinical features of cancer patients

Analysis was performed to determine the relationship between MBs and the clinical features of tumor patients, including their age, sex, weight, AJCC stage and TNM stage. 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 (<70 for MB 2; <75 for MB 5 and MB 7) but decreased when age exceeded the threshold. This trend was reversed for MB 4 and MB 8. An exception was observed for MB 9, whose proportion in single patients decreased monotonically with age. The proportion of MBs in a single patient generally varied between the sexes. 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. The difference was not significant in MB 1 and MB 7. No apparent rule regarding the relationship between the weight and MB composition of a patient was observed (Fig. 5).

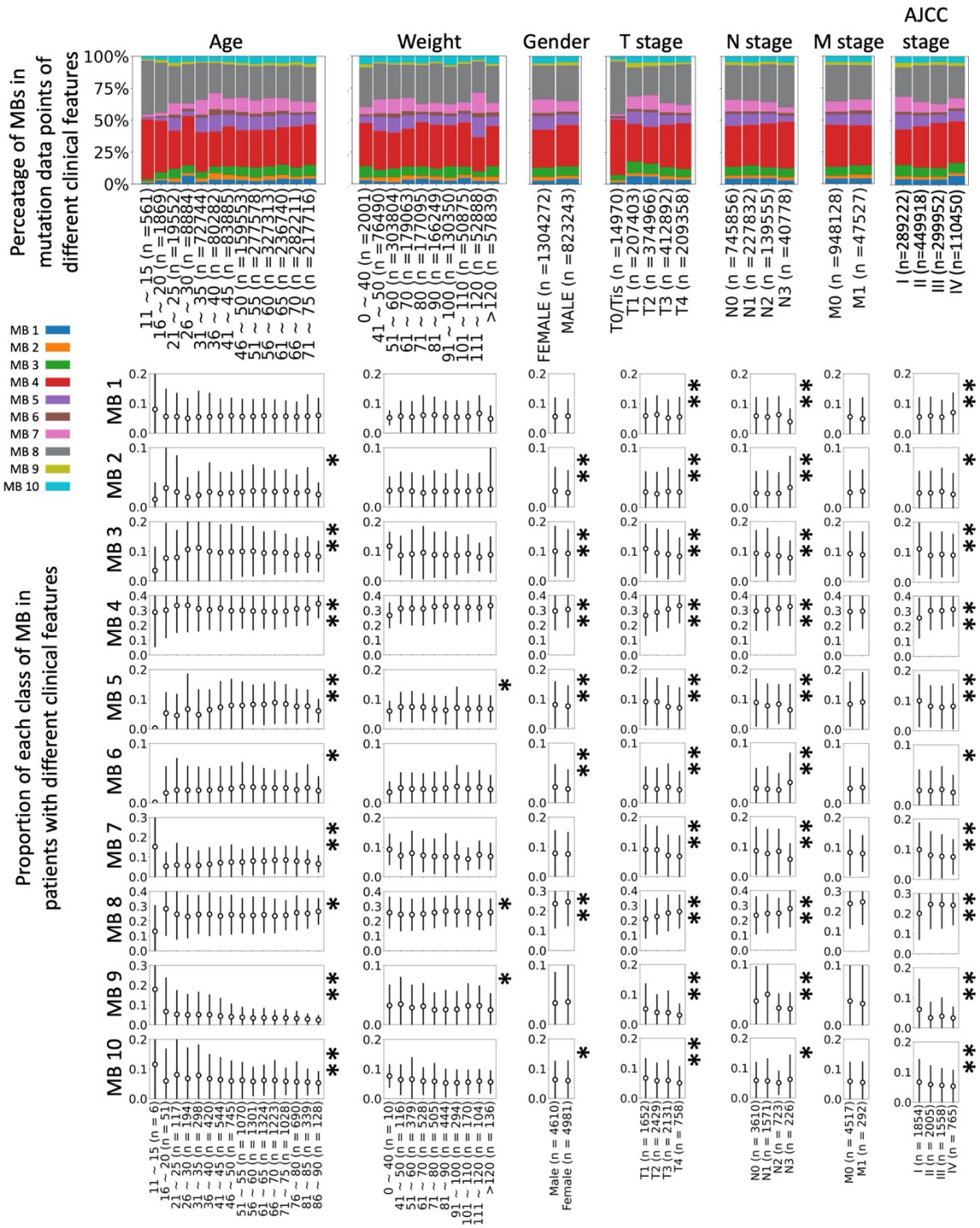
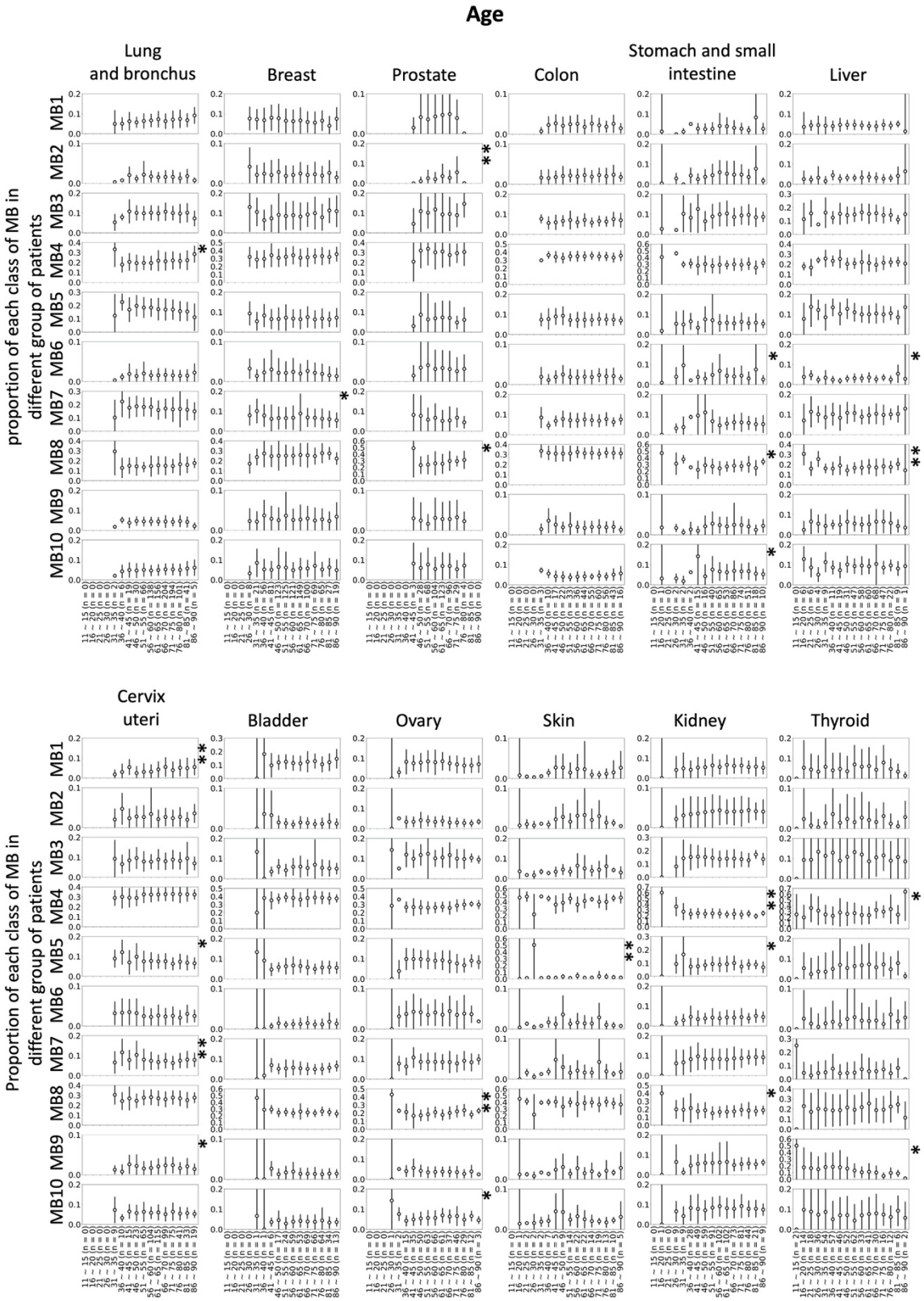


Fig. 5 | MBs in patients with different clinical features. *: $p < 0.05$ in t test or ANOVA test between groups; ** $p < 0.005$ in t test or ANOVA test between groups. Proportion is shown as mean ± standard deviation, error bars represent standard deviation

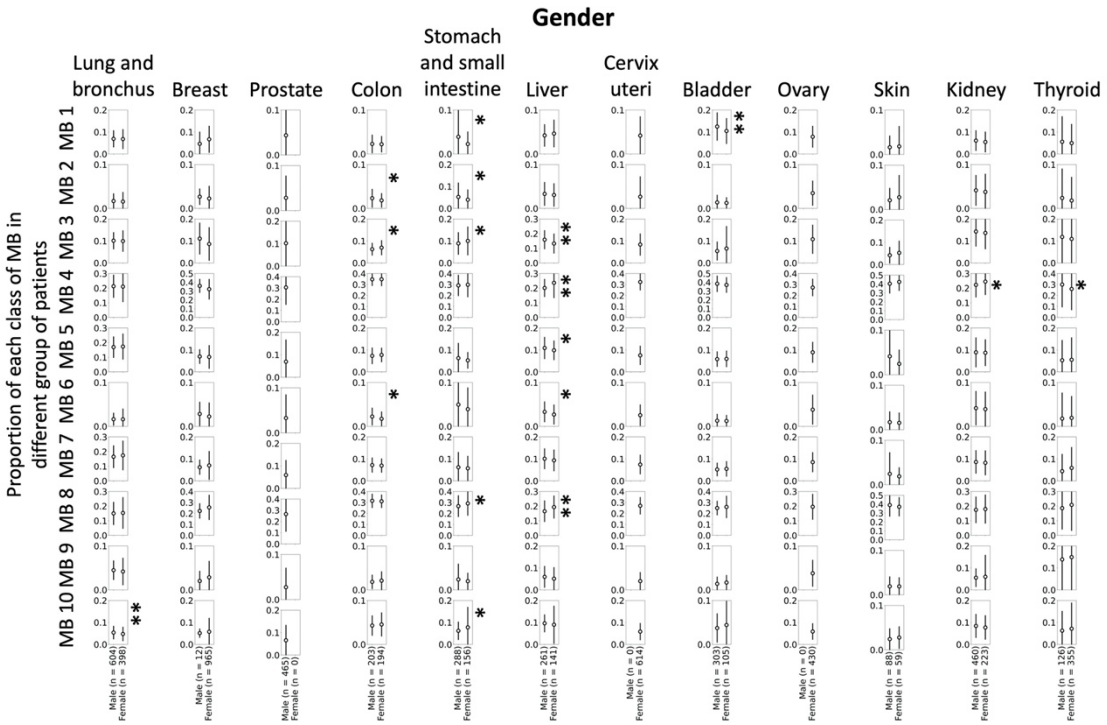
In the analysis of cancer stage, we included AJCC stage and T, N, and M stage. Although the detailed methods of AJCC staging in different cancers are not the same, they

generally follow similar principles. For example, cancers with distal metastasis are usually classified as stage M1 or AJCC stage IV, while AJCC stage I usually refers to local lesions. Therefore, we merged the subdivisions of the stages in some cancers (i.e., T3a and T3b are recorded as T3 in stomach cancer; AJCC IIIA, IIIB, and IIIC stages in lung cancer are recorded as AJCC stage III)[29]. 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 of with T and N stages. For some MBs, their relationship with cancer staging was complicated. MB 5 decreased with the progression of T and N stages, but M1 patients presented more MB 5 than M0 patients. 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[30], 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) (Fig. 5).

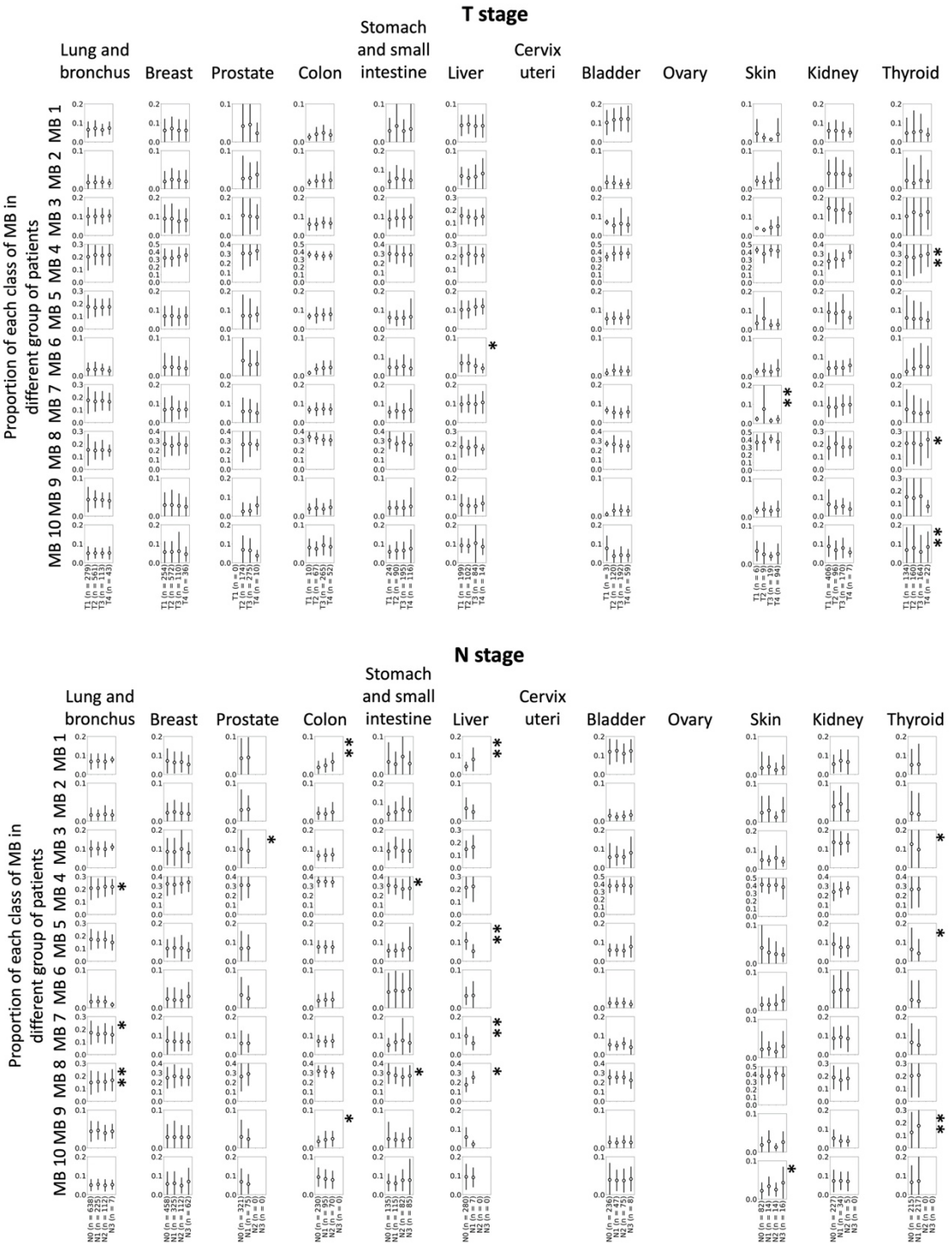
Considering the differences in the clinical significance of staging in different cancers, further analysis was performed for each cancer with high incidence. In most cancers, the MB composition at different ages basically followed the pattern shown in the total samples (i.e., presenting a nonmonotonic relationship). Younger and older patients showed a similar feature of the MB composition in the form of a conic structure in the bar graph. This suggests that the similarity of the cancer biology of young and old patients requires further study. The proportion of MB 2 in most cancers was significantly higher in males than in females. Regarding cancer staging, T and M stages showed obvious tendencies in most kinds of cancers, and their trends were 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 (Extended Data Fig. 6 - 9). Generally, although the results showed a clear relationship between MB and clinical features, the details of the relationship as well as its mechanism still require further study.



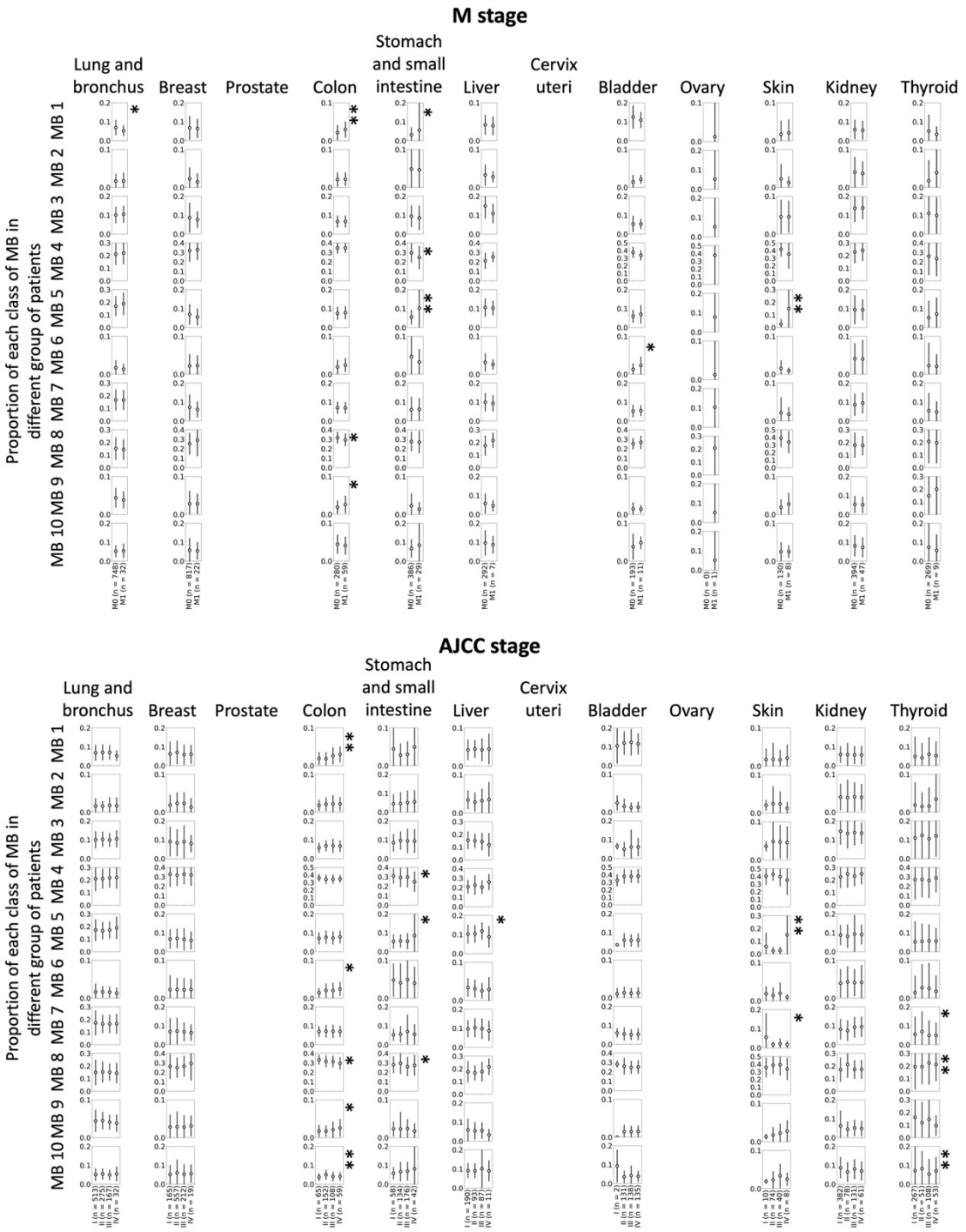
Extended Data Fig. 6 | Statistics of proportion of each MB by age in cancers with high incidence.
(Proportion is shown as mean ± standard deviation, error bars represent standard deviation)



Extended Data Fig. 7 | Statistics of proportion of each MB by gender in cancers with high incidence.
(Proportion is shown as mean ± standard deviation, error bars represent standard deviation)



Extended Data Fig. 8 | Statistics of proportion of each MB by T stage and N stage in cancers with high incidence. (Proportion is shown as mean ± standard deviation, error bars represent standard deviation)



Extended Data Fig. 9 | Statistics of proportion of each MB by M stage and AJCC stage in cancers with high incidence. (Proportion is shown as mean ± standard deviation, error bars represent standard deviation)

5. MBs and survival of cancer patients

Usually, there are multiple kinds of MBs in a single patient. To further analyze the influence of the MB composition on the clinical features of patients, a K-means clustering method was used to classify patients according to MB composition. Different kinds of MBs were recorded according to their proportion rather than their number in a single patient. After analysis via the "elbow method" and the examination of clustering results, K=7 was selected as the number of classes to be distinguished. Clustered patients were 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 MB4 and MB 1. Patients of Class 7 presented high proportions of MB 4 and MB 8 but exhibited few other classes of MBs, and the percentages of MB 4 and MB 8 were relatively balanced (Fig. 6a).

In the survival analysis, significant differences in survival curves were observed in different classes of patients (Fig. 6b). We then carried out a pairwise comparison of survival by applying the log-rank test between different classes of patients. Patients with Classes 2, 4, and 5 showed better survival, and patients with Classes 1, 3, 6, and 7 showed worse survival (Fig. 6c). In the analysis of specific cancers, survival in different classes of patients generally followed the results obtained for the total sample (Extended Data Fig. 10), but with some discrepancies and not that significant. Class 3 patients, in particular, seemed to show poor survival for most of the analyzed cancers. Considering the relatively balanced composition of MBs in patients of Classes 1, 3 and 5, especially Class 3, these results suggested that a balanced MB composition may predict poor survival in patients.

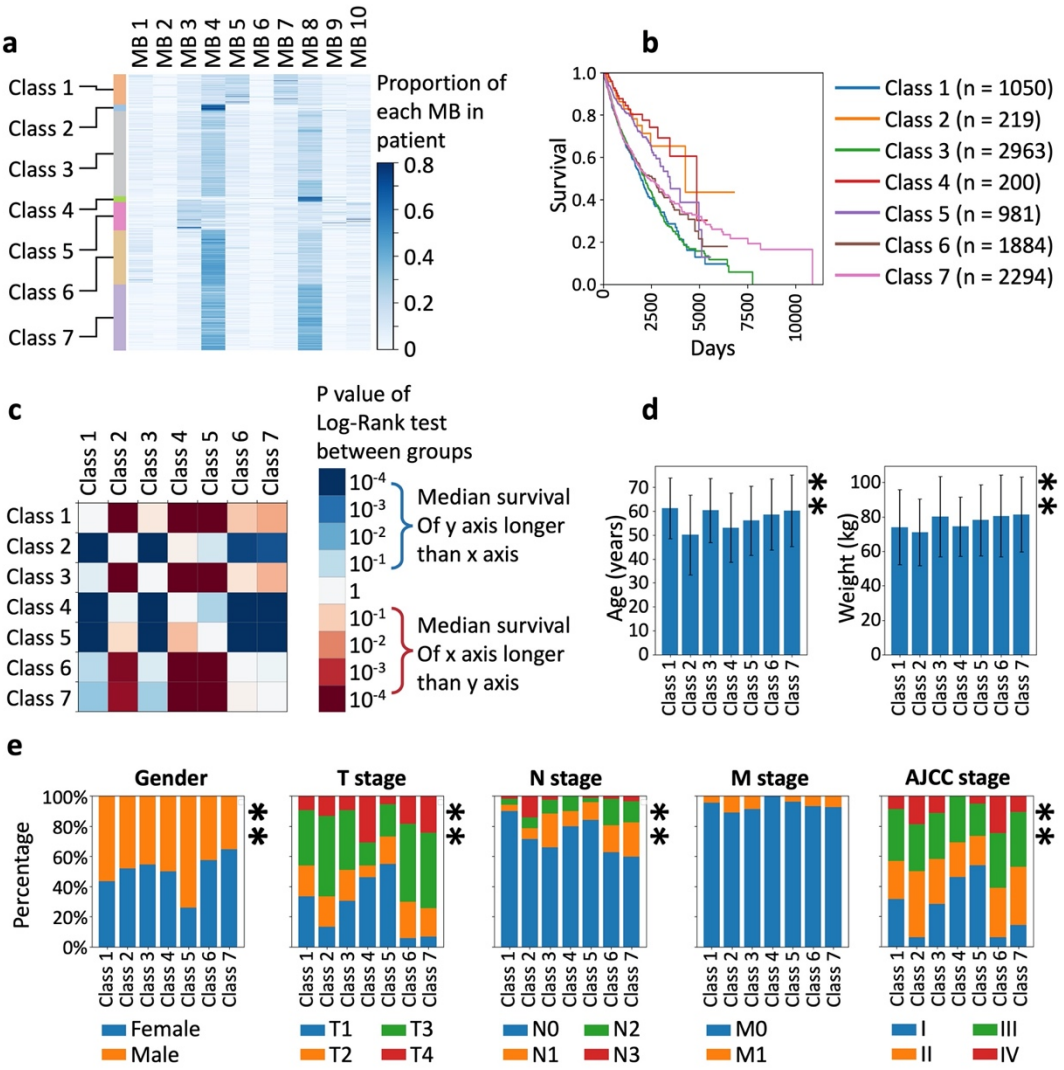
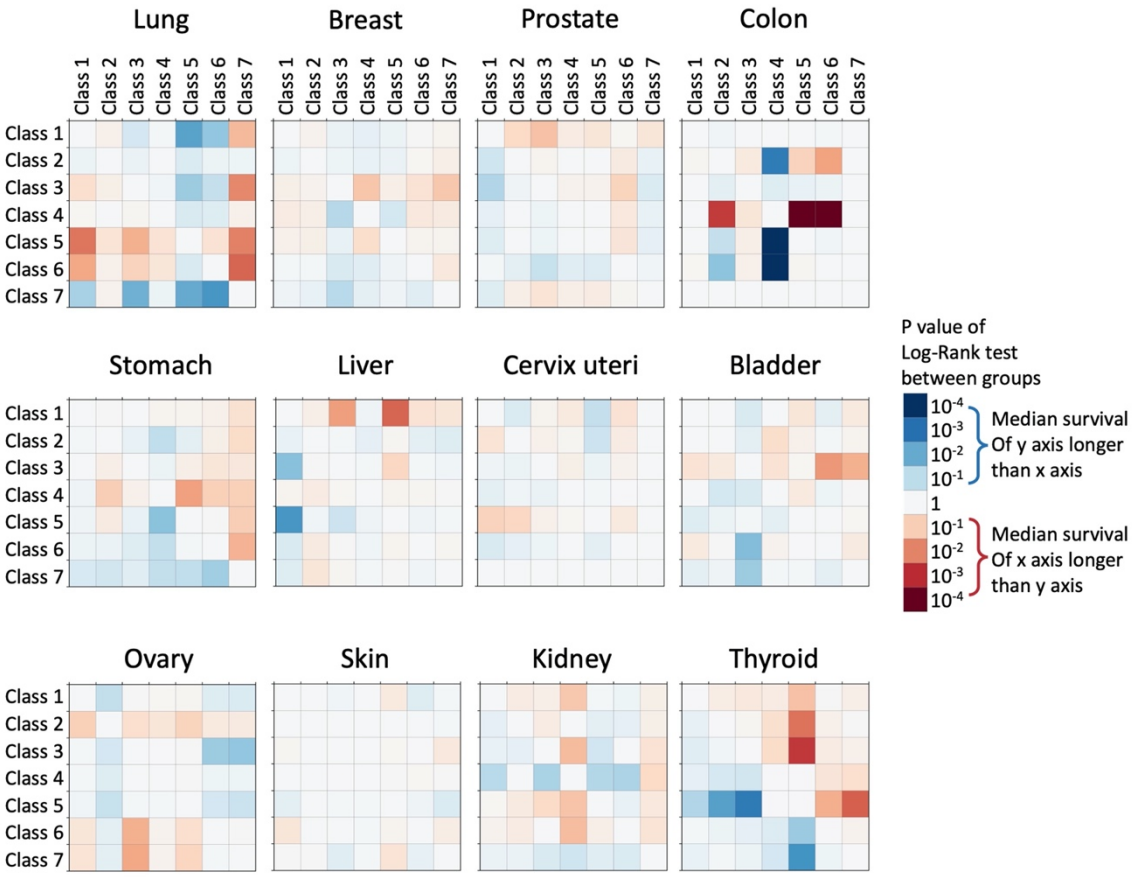


Fig. 6 | Cluster of patients by proportion of MBs and differences in survival and clinical features between classes. a: characteristic of MB composition in patients of 7 classes clustered by K-means method, each line represents one patient. b: survivorship curve of each class of patients. c: log-rank test between classes, differences in p value are reflected in color. d, e: clinical features in different classes of patients (*: $p < 0.05$ in ANOVA test or chi-square test; **: $p < 0.005$ in ANOVA test or chi-square test. Error bar stand for standard deviation).



Extended Data Fig. 10 | Log-rank test between classes in different cancers. Differences in p value are reflected in color.

Patients of different classes showed distinct clinical features (Fig. 6d, 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. Interestingly, Class 4 included significantly more T4 patients but hardly any M1 patients. This suggests that the MB composition of Class 4 may be associated with the local progression of cancers. 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. In the analysis of age, patients of Class 3 were found to present significantly greater ages. At the same time, the weight of Class 3 patients was also high. These factors may be partly responsible for the poor survival of Class 3 patients. In contrast to the findings for Class 3, although the patients of Class 1 exhibited a similarly higher age, their weight was not very high. Class 1 patients exhibited a high percentage of AJCC stage 1 and a low percentage of stage IV. Moreover, the proportion of N0 patients in Class 1 was significantly higher than that in other classes, and the proportion of M1 patients in Class 1 was average. Therefore, further study is still needed to determine the mechanism causing Class 1 patients to show poor survival.

Discussion

Previous studies have proven that different mutation characteristics may be associated with different triggers involved in various mutation processes and result in differing biological behaviors of cancers. Some mathematical methods are now used for clustering the signature of mutation. A variety of mutation signatures that may be related to the biology and etiology of cancer have been identified[2,8, 15-21,31-34]. LSTM is a machine learning approach that is good at extracting the features of long sequences[35]. This approach provided us with a method for extracting the features of mutated sequences across a wider spatial scope. 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.

There are still some constraints and limitations of this study. The clustering results obtained with the LSTM-SOM model is 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). We adjusted the parameters so that we could divide the samples into two classes in one round of training and obtained 10 classes after 3 rounds and an extra round of training. However, there exist the possibility that when training with other parameters, the classification obtained may be related to clinical features that were not included in this study, which needs further study.

The results of our study showed that even among patients with mutations in the same gene, significant differences may exist in their clinical features and survival. The mechanism of machine learning models is difficult to explain[36]. It is meaningful to use a mathematical method for exploring 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.

The TCGA database includes a large sample of mutation data. Due to the natural differences in cancer incidence, large differences exist between different cancers. In different cancers, MB may be involved in different kinds of cancer-related processes. Therefore, the analysis of the relationship between MBs and distinctive clinical features in specific kinds of cancer can provide more information about how MBs are related to cancer etiology, processes, prognosis and drug susceptibility. In summary, this study provided a method for classifying the characteristics of mutant sequences and explored their relationships with the clinical features and prognosis. Further study of the mechanism of MBs related to cancer characteristics is suggested.

References

1. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011).
2. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).

3. Cooke, M. S., Evans, M. D., Dizdaroglu, M. & Lunec, J. Oxidative DNA damage: mechanisms, mutation, and disease. *FASEB J.* **17**, 1195-1214 (2003).
4. Pfeifer, G. P. Environmental exposures and mutational patterns of cancer genomes. *Genome Med.* **2**, 54 (2010).
5. Stratton, M.R, Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
6. Pena-Diaz, J. et al. Noncanonical mismatch repair as a source of genomic instability in human cells. *Mol. Cell.* **47**, 669–680 (2012).
7. Cappell, M. S. Pathophysiology, clinical presentation, and management of colon cancer. *Gastroenterol Clin North Am.* **37**, 1-24 (2008).
8. Alexandrov L. B. et al. The repertoire of mutational signatures in human cancer. *Nature.* **578**, 94-101 (2020).
9. Remon, J. et al. Osimertinib and other third-generation EGFR TKI in EGFR-mutant NSCLC patients. *Ann Oncol.* **29(suppl_1)**, i20-i27 (2018).
10. Stintzing, S. 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.* **30**, 1796-1803 (2019).
11. von Minckwitz, G. et al. Trastuzumab Emtansine for Residual Invasive HER2-Positive Breast Cancer. *N Engl J Med.* **380**, 617-628 (2019).
12. Sawrycki, P. et al. Relationship between CYP1B1 polymorphisms (c.142C > G, c.355G > T, c.1294C > G) and lung cancer risk in Polish smokers. *Future Oncol.* **14**, 1569-1577 (2018).
13. Zerp, S. F. et al. p53 mutations in human cutaneous melanoma correlate with sun exposure but are not always involved in melanomagenesis. *Br J Cancer.* **79**, 921-926 (1999).
14. Petljak, M. et al. Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell.* **176**, 1282-1294.e20 (2019).
15. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
16. Poon, S. L. et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci. Transl. Med.* **5**, 197ra101 (2013).
17. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
18. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
19. Petljak, M. & Alexandrov, L. B. Understanding mutagenesis through delineation of mutational signatures in human cancer. *Carcinogenesis* **37**, 531–540 (2016).
20. Mimaki, S. et al. Hypermutation and unique mutational signatures of occupational cholangiocarcinoma in printing workers exposed to haloalkanes. *Carcinogenesis* **37**, 817–826 (2016).
21. Polak, P. et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.* **49**, 1476–1486 (2017).
22. Hochreiter, S. & Schmidhuber, J. Long short-term memory. Neural Comput. **15**, 1735-1780 (1997).
23. Gers, F. A. et al. Learning to forget: continual prediction with LSTM. *Neural Comput.* **12**, 2451-2471 (2000).
24. Tayara, H. & Chong K. T. Improving the Quantification of DNA Sequences Using Evolutionary Information Based on Deep Learning. *Cells* **8**, 1635 (2019).
25. Liu, Q. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nat. Commun.* **10**, 2449 (2019).
26. Zhou, J. 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.* **17**, 124-135 (2020).

27. Markey, M. K. et al. Self-organizing map for cluster analysis of a breast cancer database. *Artif. Intell. Med.* **27**, 113-27 (2003).
28. Furukawa T. SOM of SOMs. *Neural Netw.* 22, 463-478 (2009).
29. Edge, S. B. & Compton, C. C.The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Ann. Surg. Oncol.* **17**, 1471-1474 (2010).
30. Fossa, S. D. Testicular carcinoma in situ in patients with extragonadal germ-cell tumours: the clinical role of pretreatment biopsy. *Ann. Oncol.* **14**, 1412-1418 (2003).
31. Meier, B. et al. C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res.* **24**, 1624–1636 (2014).
32. Huang, M. N. et al. Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors. *Genome Res.* **27**, 1475–1486 (2017).
33. Nik-Zainal, S. et al. The genome as a record of environmental exposure. *Mutagenesis* **30**, 763–770 (2015).
34. Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. *Cell* **177**, 821–836 (2019).
35. Sahin, S. O. et al. Nonuniformly Sampled Data Processing Using LSTM Networks. *IEEE Trans. Neural. Netw. Learn Syst.* **30**, 1452-1461 (2019).
36. McCloskey, K. Using attribution to decode binding mechanism in neural network models for chemistry. *Proc. Natl. Acad. Sci. U S A.* **116**, 11624-11629 (2019).

Methods

Data availability

SBS data and clinical data of patients involved in this study were obtained from the TCGA database. All data used in this study are public data from the TCGA downloaded from https://portal.gdc.cancer.gov. There are 2 kinds of files. SBS data are stored in .maf files, and the clinical data of each patient are stored in an .xml data file. First, the SBS information includes the sample barcode, chromosomal location, mutant allele, reference allele, Hugo gene symbol, etc. Clinical data, including age, sex, weight, cancer stage, and survival or time to last follow-up, were extracted from .xml files according to the sample barcode. In the LSTM-SOM model, 100 flanking bases were included in the analysis, and the flanking sequence was obtained from Genome Reference Consortium Human genome build 38(GRCh38, downloaded from ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/) based on the mutation sites of SBSs in TCGA data. Reference bases provided by TCGA were compared with bases at corresponding sites in GRCh38 to further ensure accuracy.

LSTM-SOM model building

The LSTM-SOM model consists of two components: LSTM is used to extract the features of sequences, and SOM is used to cluster samples. In brief, 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]). When placing the reference sequence in the corresponding position, mutated bases can be recorded as a 1×4 vector. For example, C>T can be expressed as [1, 0, 0, 1]. When the flanking bases are included, a mutated sequence can be represented by an n×4 matrix. For example, CATTG > CACTG can be expressed as follows:

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

RNNs have long been used in the analysis of sequence data, for example, for speech recognition. A naive RNN effectively analyzes short sequences. When used for long sequences, a naive RNN may show the problems of gradient disappearance and gradient explosion. An LSTM network is based on the network structure of RNNs[22]. The LSTM approach introduces the mechanisms of "forgetting" and "memory". Thus the capacity of the LSTM network to analyze long sequences is improved by controlling the long-term state[23]. As the "forgetting" mechanism of LSTM, the unit closer to the end of the sequences has a greater influence on the output of LSTM. For the mutation site to have the greatest impact on the output result, LSTM is designed to read from both ends of the mutated sequence. The sequence upstream of the mutation site is read from 5' to 3' (forward sequence), and the sequence downstream of the mutation site is read from 3' to 5' (reverse 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. The randomized units in the competition layer were trained to describe the distribution of units in the input layer via the mechanism of "competitive learning"[28]. In the SOM process of LSTM-SOM model, the feature vector obtained from the LSTM process is used as the input. Units in the competition layer are adjusted continuously according to their distance to the input unit. For one input unit, the unit in the competition layer nearest to it is regarded as the "winning unit", which will move the maximal distance to the input unit (target), and for the other units, their travel distance to the target decreases with the increase in the distance to the winning unit. To avoid an excessive concentration of the results, we set a threshold value in the model. When the distance between the competition layer unit and the target is over the threshold value, the unit will move in the opposite direction to the target. In particular, not only will units in the competition layer be updated in our SOM model, but the input unit will also be updated in the opposite

direction of the vector sum of the competition layer unit movement. Then, the updated input unit will be used as a label to train the LSTM model.

First, we obtained feature vectors of 100 samples from LSTM in one batch, and they were used as the input units of the SOM. 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 is referred to the Gaussian function[37]:

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

In the decay 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)：

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

It can be seen that when the distance between w_j and the target x is less than S, they will approach each other. Otherwise, they will pull away from each other. Due to the existence of the decay function, the influence of distant units on each other is very small, and no excessive dispersion of units was observed in training. 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. During training, 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 competition layer units) to w_{min}. After each iteration of SOM analysis, 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 the classification of samples was obtained via the method of logistic regression classification, another round of training was performed to obtain the further classification of each class of samples. 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. Enumeration data were expressed as count data, and chi-square analysis was used for difference testing. A sample was removed if the data of an item required for statistics were missing. $P<0.05$ was considered to indicate a statistically significant difference.

In the survival analysis, 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. Only significant differences between groups were shown in the heat map. 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 included in the TCGA database usually exhibit multiple kinds of MBs. 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)[38]. 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. All the code is saved at *https://github.com/FreudDolce/folder.* 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.

37. Kolasa, M. A. programmable triangular neighborhood function for a Kohonen self-organizing map implemented on chip. *Neural Netw.* **25**, 146-60 (2012).
38. Fukuoka, Y. Objectively Measured Baseline Physical Activity Patterns in Women in the mPED Trial: Cluster Analysis. *JMIR Public Health Surveill*. **4**, e10 (2018).

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

Author Contributions J. H. C., L. J. J., and Z. H. M. designed this study. J. H. C. and Z. Q. designed the mathematical method used in this study. J. H. C. and Z. H. M. wrote the manuscript. W. X. W., B. S., and T. F. collected and prepared data for analysis. Y. J. Y. created figures and tables. Z. H. M. directed the overall research.