frontiers | Frontiers in Oncology

ORIGINAL RESEARCH
published: 08 June 2022
doi: 10.3389/fonc.2022.927426

Predicting Tumor Mutational Burden From Lung Adenocarcinoma Histopathological Images Using Deep Learning

Yi Niu[1†], Lixia Wang[2†], Xiaojie Zhang[1†], Yu Han[1†], Chunjie Yang[1†], Henan Bai[1†], Kaimei Huang[2], Changjing Ren[2], Geng Tian[2,3], Shengjie Yin[1], Yan Zhao[1], Ying Wang[4], Xiaoli Shi[2,3] and Minghui Zhang[1*]*

[1] Department of Oncology, Municipal Hospital of Chifeng, Chifeng, China, [2] Geneis Co., Ltd., Beijing, China, [3] Qingdao Geneis Institute of Big Data Mining and Precision Medicine, Qingdao, China, [4] Department of Oncology, Inner Mongolia Medical University, Hohhot, China

OPEN ACCESS

Edited by:
Min Tang,
Jiangsu University, China

Reviewed by:
Yushuang Li,
Yanshan University, China
Wenyan Wang,
Anhui University of Technology, China

***Correspondence:**
Minghui Zhang
cfzhangminhui@163.com
Xiaoli Shi
shixl@geneis.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

Received: 24 April 2022
Accepted: 16 May 2022
Published: 08 June 2022

Citation:
Niu Y, Wang L, Zhang X, Han Y, Yang C, Bai H, Huang K, Ren C, Tian G, Yin S, Zhao Y, Wang Y, Shi X and Zhang M (2022) Predicting Tumor Mutational Burden From Lung Adenocarcinoma Histopathological Images Using Deep Learning. Front. Oncol. 12:927426. doi: 10.3389/fonc.2022.927426

Tumor mutation burden (TMB) is an important biomarker for tumor immunotherapy. It plays an important role in the clinical treatment process, but the gold standard measurement of TMB is based on whole exome sequencing (WES). WES cannot be done in most hospitals due to its high cost, long turnaround times and operational complexity. To seek out a better method to evaluate TMB, we divided the patients with lung adenocarcinoma (LUAD) in TCGA into two groups according to the TMB value, then analyzed the differences of clinical characteristics and gene expression between the two groups. We further explored the possibility of using histopathological images to predict TMB status, and developed a deep learning model to predict TMB based on histopathological images of LUAD. In the 5-fold cross-validation, the area under the receiver operating characteristic (ROC) curve (AUC) of the model was 0.64. This study showed that it is possible to use deep learning to predict genomic features from histopathological images, though the prediction accuracy was relatively low. The study opens up a new way to explore the relationship between genes and phenotypes.

Keywords: tumor mutation burden, lung cancer, digital pathology, immunotherapy, deep learning, difference analysis

INTRODUCTION

At present, the most advanced treatment for non-small cell lung cancer is the combination of immunotherapy and chemotherapy. Among them, PD-L1 expression is a common biomarker in immunotherapy response, but a large number of patients with low PD-L1 expression and some patients with h7igh PD-L1 expression are suitable for this treatment plan (1–3). Therefore, searching for new predictors of immunotherapy response is crucial.

TMB is an important biomarker for predicting immunotherapy (4–6). TMB is a measure of the total number of non-synonymous somatic mutations per megabase in the coding region of the tumor genome (7). Tumors with high TMB are thought to express a variety of neoantigens. A number of studies have shown that the response of patients with advanced solid tumors to immunotherapy is related to high TMB (4, 7, 8). Therefore, researchers will pay more attention to the existence of TMB when implementing immunotherapy. Therefore, there is an urgent need for a low-cost, fast and reliable TMB detection method.

Whole-exome sequencing is the gold standard for measuring TMB. However, due to technical limitations and high costs, whole-exome sequencing has not been promoted in the field of clinical oncology (9). Therefore, clinicians usually use low-cost next-generation targeted gene sequencing. When the gene-directed therapy guided by the target panel is used and tested, the turnaround time is usually about three weeks. At present, clinicians are trying to re-adjust the purpose of targeted sequencing analysis to facilitate the prediction of TMB when using whole-exome sequencing, and use the default technical method to normalize the number of mutant genes found in the sequencing area (10). Therefore, to obtain a robust normalized TMB, paired normal samples and a larger panel size (at least about 2 megabases) are required (11). The increase in panel size is directly proportional to the cost, which forces a trade-off between the depth of sequencing and the number of patients for each sequencing run. In most clinical treatment processes, the turnaround time of targeted sequencing analysis usually exceeds the prescribed time due to the limitations of various problems. Therefore, it is very beneficial to develop an alternative and convenient method to assess TMB.

Although early histopathologists have recognized the connection between individual genetic mutations and certain cancer morphological phenotypes, they did not consider the application of TMB. Previous machine learning used manual features in histopathological images (12, 13) to distinguish subtypes (14) and predict recurrence (15, 16) and survival outcomes (17). With the rise of deep learning, it is expected to achieve more robust and accurate predictions in biomedical images (18). Deep convolutional networks have shown good results in tumor detection (19–21) and distinguishing subtypes of non-small cell lung cancer (22, 23) and other cancers (24, 25).

Here, we used the ResNet18 deep learning model and used the formalin-fixed paraffin-embedded (FFPE) hematoxylin and eosin (H&E) stained lung adenocarcinoma (LUAD) histopathology images from the Cancer Genome Atlas (TCGA, https://www.cancer.gov/tcga) to predict TMB status. We tried to develop a deep learning method for LUAD, because compared with other cancers of TCGA, we can obtain a large number of sequence and image data of patients with non-small cell lung cancer, and used TMB as a biomarker for tumor treatment. ResNet18 is a convolutional neural network (CNN) that achieves advanced performance on ImageNet. We convinced that with the further development of deep learning and clinical verification, deep learning can provide a potential alternative detection method to determine TMB, while reducing the diagnosis cycle and cost consumption. Our results showed that we could use deep learning techniques to detect previously unexplored features in histopathological images that have been clinically proven to be useful.

METHODS

Data Processing

In this study, we used whole-slide images (WSIs) of LUAD from TCGA (https://portal.gdc.cancer.gov/repository/). In addition, the clinical information and next generation sequencing results of these patients were downloaded and analyzed. All LUAD WSIs were stored in SVS format and adjusted to 0.5 um per pixel at the same magnification (40x). These WSIs were then marked with tumor areas by a professional pathologist (**Figure 1A**). We set a TMB threshold (10, corresponding to a maftools count of 400) to mark each patient's TMB status.

Considering that the image pixels of WSIs were too large, it cannot be directly used as the input of the deep learning network. Therefore, we divided WSIs into 512*512 pixels slides, each WSI can be divided into tens of thousands image slides. And then deleted slides with less information (e.g. blank rate over 30%) (15). The error of the manual production process and the difference between stains and scanners will produce a color difference between digital sections, which will cause errors in the subsequent analysis work. Therefore, we performed color normalization using the Macenko method in the Tia toolbox software package (**Figure 1B**) (26).

In the pipeline, we promised each patient and its related slides were allocated to a training or testing dataset to ensure that there was no overlap problem and to ensure the accuracy of the final test results (**Figure 1C**). The final model predicted the TMB status according to the features of each patient, so as to assist the doctor in giving the corresponding immunity treatment plan.

Deep Learning on Histopathological Images

We used a deep neural network based on the ResNet18 architecture to predict the probability of TMB status for each tile (**Figure 1D**). ResNet18 builds a 18-layer deep convolutional neural network by repeatedly using two residual blocks, Conv Block and Identity Block. The last of the ResNet18 architecture is a fully connected layer for soft-max operations. The output produced two normalized probability predictions. We choose a TMB value of ten as the segmentation threshold. The model parameters were initialized using pre-trained weights form the ImageNet competition (27). Using backpropagation to train all parameters of the model during model training. The loss function was defined as the cross entropy between the true label and the predicted probability.

Patient-Level Prediction

As shown in **Figure 1**, the TMB status of each image block can be predicted by our prediction model, and these predicted image blocks were divided into two categories: TMB high and TMB low by the soft-max function. The image patches of each patient were then aggregated together to predict the probability of high or low TMB for each patient. Specifically, we divided the number of patches predicted to be high TMB per patient by the total number of patches per patient. If its probability was greater than 0.5, the LUAD cancer patient was predicted to have high

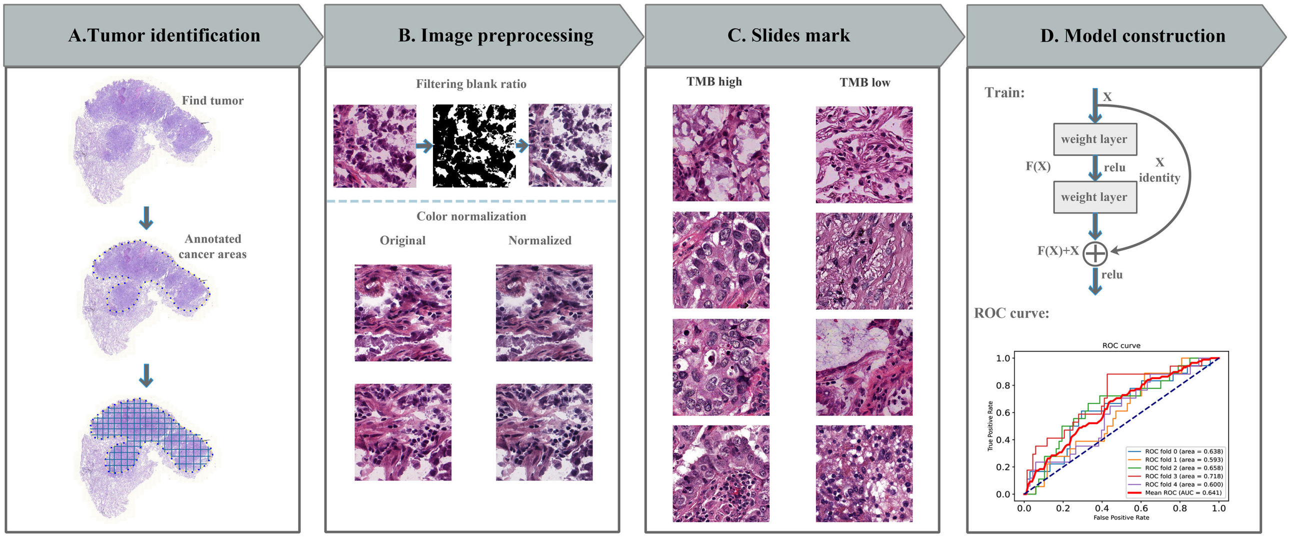

FIGURE 1 | Computational pipeline for predicting TMB in lung cancer. **(A)** The tumor area was annotated by a professional pathologist and cut into 512*512 image blocks. **(B)** Image processing includes noise reduction (discarding image blocks with a blank rate greater than 30%) and color normalization (Macenko method). **(C)** Divide TMB high and low with 10 as the threshold. **(D)** The prediction model was constructed with residual network, the model was tested with 5-fold cross-validation, and receiver operating characteristic (ROC) curve was used to evaluate the model.

TMB, otherwise, the patient was defined as low TMB. And the patient's TMB status is the standard for subsequent provision of corresponding treatment plans.

Performance Evaluation

After the model was trained, cross-validation was used for model testing and performance evaluation. We used the percentage of correctly classified slices to aggregate the probability of each slide. We used the scikit-learn in the python library to calculate the ROC curve and the corresponding AUC value in the case of prediction.

In addition, we performed survival analysis of true and predicted TMB status of these 427 lung adenocarcinoma patients. Overall survival time was calculated from the date of surgery to the date of death or last follow-up contact. Survival curves were estimated using the Kaplan-Meier multiplicative limit method (E. L. 28). Differences in predicted survival outcomes between high and low TMB groups were compared by log-rank test.

Hyperparameter and Model Selection

Model selection and all hyperparameters were based on the performance of the validation dataset, including the use of the ResNet18 model, optimization of parameters, and so on. The validation dataset is only used after the model is developed and used to introduce all the results obtained in this research. In the process of model evaluation, the methods and models we use are not modified in any way to ensure independence assumptions between datasets. To avoid the effects of the algorithm falling into local optimal solutions and data noise, we employed the SGD + momentum optimizer, where momentum assigns a value of 0.9. For every seven epochs, the learning rate of the parameters decayed by 0.1 times. Moreover, 30 epochs were trained throughout the process.

Statistical Analysis

All statistical analysis was conducted using R software. We used statistical methods to analyze the differences of clinical characteristics in 427 cases of TCGA lung adenocarcinoma patients with high and low TMB data.

In clinical, for continuous clinical characteristic variables such as age and the number of cigarettes per day, we used the Wilcoxon rank sum test method to analyze (29). For non-continuous variables such as tumor stage, we used the fisher's exact test method to analyze. P-values less than 0.05 were considered statistically significant.

DESeq2 method of R package was used to analyze the difference of mRNA in TMB high and low groups. DEseq2 requires the input data to be an unnormalized matrix of integers (30). At the same time, we also performed GO (gene ontology) enrichment and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway enrichment analysis on mRNA, and annotated from three aspects: BP (biological process), MF (molecular function), and CC (cellular component). The GO enrichment and KEGG calculation method and formula are the same, and they are calculated using the hypergeometric test:

$$P = 1 - \sum_{i=0}^{m} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Among them, 'N' is the number of genes with Pathway annotation in all genes. 'n' is the number of differentially expressed genes in N. 'M' is the number of genes annotated as a specific pathway in all genes. 'm' is the number of differentially expressed genes annotated as a specific Pathway. The calculated P value is further corrected for multiple testing to obtain the corrected p-value (that is, the Q value). Usually we take Q value ≤ 0.05 as a threshold, and a pathway that satisfies this condition is defined as a pathway that is significantly enriched in differentially expressed genes (31).

RESULTS

Smoking and Age Were Significantly Correlated With TMB Status

In this study, 468 patients' sequencing results were downloaded from TCGA and 427 H&E-stained WSIs of formalin-fixed paraffin-embedded tumor tissue sections of all available LUAD cases were obtained. Therefore, after matching the patient information, 427 patients were finally included in this study. The clinical information of these 427 patients was also obtained from TCGA and performed statistical analysis, and some of the clinical information was shown in **Table 1**. According to the TMB threshold, the number of TMB high was 88, and the number of TMB low was 339 among 427 patients with LUAD.

Then, we analyzed the relationship between these clinical characteristics and TMB status, the results were shown in **Figure 2**. Daily cigarette consumption was significant different (p=0.014) in patients between TMB-H group and TMB-L group (**Figure 2A**). The daily smoking volume of TMB-H group was significantly higher than that of TMB-L group. Besides, the age of patients in TMB-H group was significantly lower than that in TMB-L group (**Figure 2B**). Whereas, there was no significant correlation between tumor stage and TMB status, with a P value of 0.09, which was shown in **Supplementary Figure 1**.

Many Differential Genes Were Identified Between TMB-H and TMB-L Samples

The mRNA sequencing data of cancer tissue samples from 427 patients with lung adenocarcinoma were downloaded from TCGA. And then these patients were divided into TMB-H group and TMB-L group according to the value of TMB. To explore whether the changes of high or low expression of some key genes will directly lead to the increase of the total number of gene mutations, that is, affect TMB status, the gene expression differences between the two groups were analyzed using DESeq2. With log2 |fold change| ≧ 1 and p value ≤ 0.05 as the threshold, we got 2140 significantly differentially expressed genes in TMB-H group compared with TMB-L group. Among these genes, 960 genes were up-regulated and 1180 genes were down-regulated in TMB-H group (**Figure 3A**). The top 10 up-regulated genes and down-regulated genes in TMB-H group compared to TMB-L group were shown respectively in **Figure 3B**.

The Differential Genes Between TMB-H and TMB-L Are Significantly Enriched in Many GO Terms and KEGG Pathways

Go enrichment analysis and KEGG pathway enrichment analysis were further carried out to clarify the functions and signaling pathways involved in these differential expressed genes. GO enrichment analyses include three aspects: BP, CC and MF. The top ten Go terms in the biological process were displayed in **Figure 4A**, including hormone metabolic process, sodium ion transport, carboxylic acid transport, regulation of blood pressure, digestion, digestive system process, neuron fate commitment, G protein-coupled receptor signaling pathway etc. KEGG pathway analysis results indicated that differentially expressed genes are mainly enriched in the neuroactive ligand-receptor interaction pathway, Calcium signalling pathway, drug metabolism-cytochrome P450 pathway and nicotine addiction pathway (**Figure 4B**). Genes in these pathways may regulate the mutation and repairment of DNA.

Histopathological Images Could be Used to Predict TMB

After preprocessing, the histological images of H&E staining in TMB high group were labeled as 1, meaning positive sample, and those in TMB low was labeled as 0. Small tiles of pathological images from the same patient are divided into the same dataset to ensure that information is not leaked. Then the ResNet18 model

TABLE 1 | Summary of the general clinical information of patients with lung adenocarcinoma.

Clinicopathologic variable	Category	TCGA LUAD
Sample type	H&E stained sections	427
Age	<=50	33
	>50	379
	NA	15
Cigarettes_per day	<2	140
	>=2	157
	NA	130
Tumor stage	I	238
	II	102
	III	55
	IV	25
	NA	7

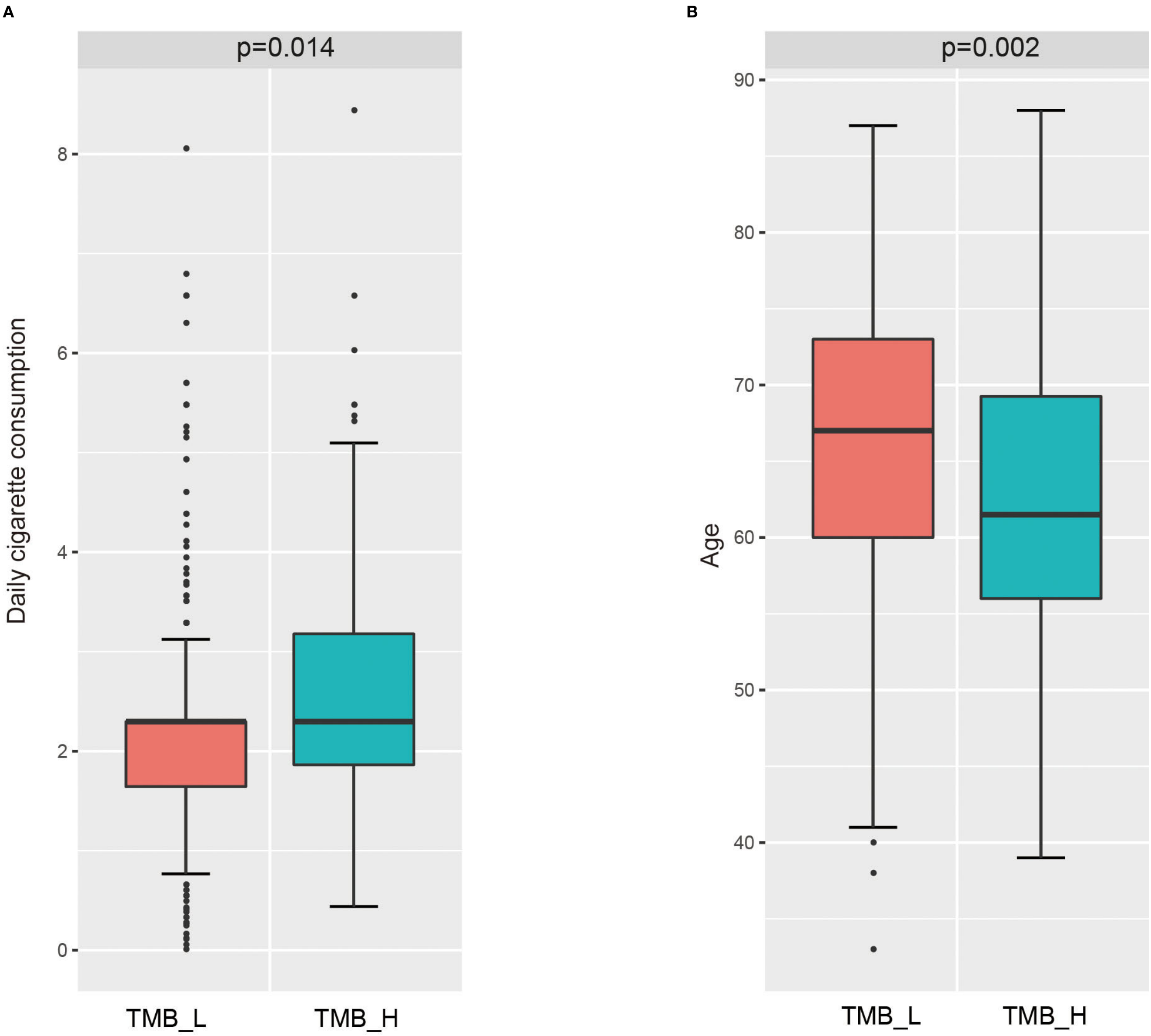

FIGURE 2 | Correlation of several clinical features with TMB status. **(A)** Correlation analysis between daily cigarette consumption and TMB status. **(B)** Relationship between ages and TMB status.

was employed to train the samples, and 5-fold cross validation was used to split the samples and verify the results.

The classification model we developed will be used to predict TMB status of the entire slide image of a given patient, which will provide a certain basis for pathologists' later diagnosis and treatment. The ROC curve and AUC of the predict model were shown in **Figure 5A** and the confusion matrix was illustrated in **Figure 5B**. The TMB predict model achieved a relatively good performance, with an area under the curve (AUC) of 0.641.

We further evaluated the performance of the deep learning model by survival analysis (**Figure 6**). As shown in **Figure 6A**, there was no significant difference in the survival time between TMB-H group and TMB-L group from the real clinical statistics. The survival analysis results of the two groups were not significantly different based on the predict TMB status by H&E-stained histological images, which was consistent with the real statistical data (**Figure 6B**).

DISCUSSION

TMB plays an important role in immunotherapy response, and it is an immunotherapeutic biomarker recommended by National

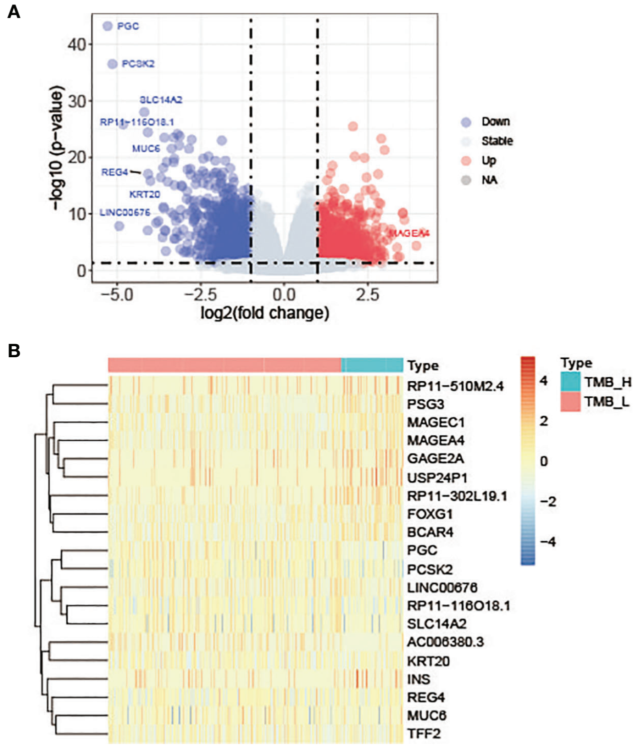

FIGURE 3 | Differentially expressed genes in TMB-H group compared with TMB-L group. **(A)** Volcano plots of the differentially expressed genes. Red dots represent up-regulated genes, blue dots represent down-regulated genes and gray dots represent genes with no significant expression differences. **(B)** Heatmap of top 10 up-regulated genes and down-regulated genes in TMB-H group compared to TMB-L group.

Comprehensive Cancer Network (NCCN) guidelines. In this study, we proved that TMB can be evaluated using digitized FFPE histopathological images in LUAD, though the prediction accuracy was relatively low. This may be related to the small number of training datasets or the threshold division of TMB. In addition, the inconsistent calculation method of TMB may also lead to inaccurate data division, which will have an impact on the results.

At present, the generalization ability of the model is affected by the dataset. In the course of clinical treatment, patients with advanced malignant tumors have relatively small diagnostic biopsy specimens, which may come from many different potential sites, including the liver, lymph nodes and other sites. In order to test the practical application of our method to clinical samples, it is necessary for us to train and test in these different scenarios. In addition, TCGA FFPE images can be highly enriched in tumor cells, but this does not reflect the real tissue samples used in the biopsy process. However, experts can manually mark the tumor area on the slide image and apply ResNet18 to the area of interest. In view of the results obtained using FFPE section samples, it was feasible to use frozen sections for research in the future.

Our deep learning model was only trained to determine whether TMB was above a selected threshold, rather than predicting specific TMB values. And, it would be better if we had an independent validation set to validate our model. As a clinically relevant measurement standard, TMB usually requires additional and laborious testing. We can predict TMB from the H&E-stained histopathology images of the LUAD datasets in TCGA. In the future, we can predict TMB from H&E images of lung squamous cell carcinoma (LUSC). In addition, our method has certain advantages compared with other diagnostic tools. We can predict the TMB value of each area of the image. This method can represent the heterogeneity of TMB itself or the heterogeneity of histological characteristics related to high TMB.

There are a few limitations of this work. First, we used deep learning to link genome-wide features with histopathological images, which helps to study the spatial heterogeneity of tumors and the relationship between cancer phenotypes and genotypes. Therefore, the use of deep learning is a useful method to improve the current large number of ready-made histopathological images, and helps to prioritize and screen patient samples and

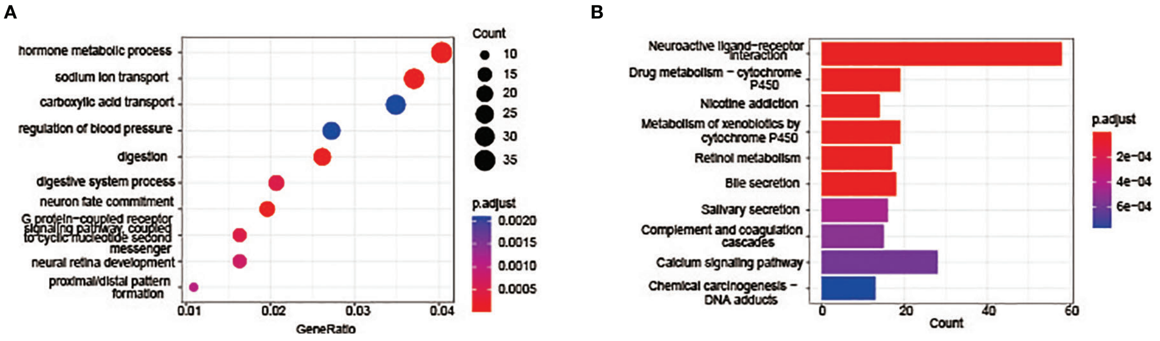

FIGURE 4 | Functional enrichment and pathway analysis results of differentially expressed genes. **(A)** Top 10 enriched GO terms annotated in biological process. **(B)** Top 10 KEGG pathways of the differentially expressed genes.

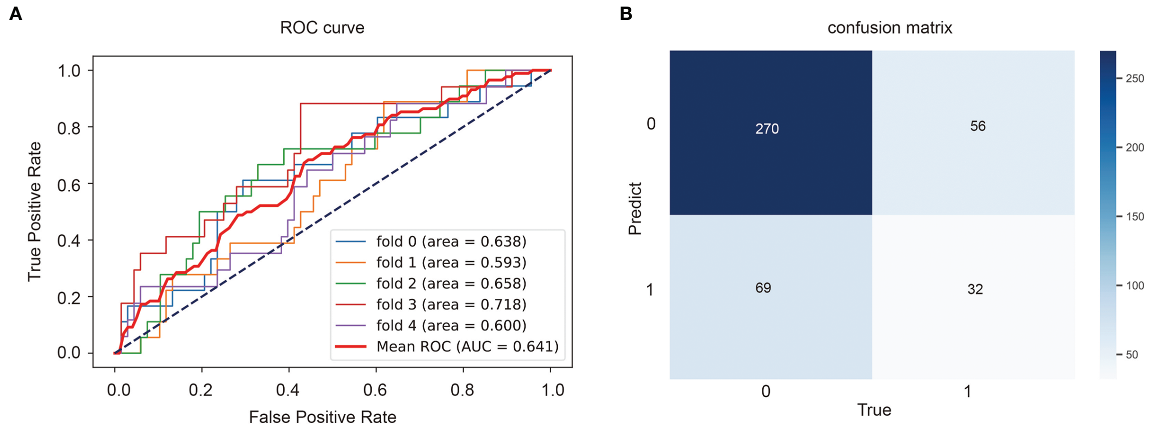

FIGURE 5 | Performance of the TMB predict model based on histopathological images. Area under the receiver operator curves (ROCs) **(A)** and the confusion matrix **(B)** of the deep learning model.

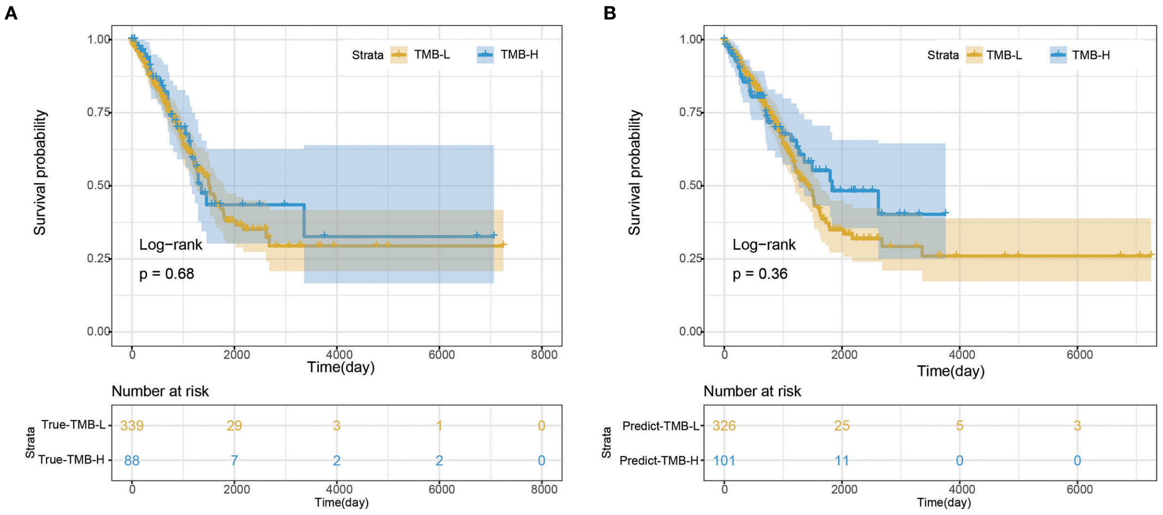

FIGURE 6 | Survival analysis. **(A)** Survival analysis based on the true clinical data. **(B)** Survival analysis based on the predict result of TMB status by H&E-stained histological images.

follow-up treatment. However, the prediction accuracy of model is relatively low. More advanced machine learning models might be able to improve the accuracy like some recent classification models used in other biological problems (32–34). Second, we used 5-fold cross validation to evaluate the model accuracy. It might be better to find some more independent datasets. However, it is infeasible to find a new dataset with both histopathological images and WES. Finally, TMB is used as an indicator for immunotherapy. It might be more direct to predict the outcomes of immunotherapy directly. However, this is listed as our future work provided that we can find some appropriate datasets.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

MZ and XS conceived the project; SY and LW implemented the experiments and analyzed the data; CY, YZ, XZ, YW, HB, YN, and YH prepared the data and performed literature search; KH, CR, and XS wrote the manuscript; all authors revised and approved the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.927426/full#supplementary-material

REFERENCES

1. Garon EB, Rizvi NA, Hui R, Leighl N, Balmanoukian AS, Eder JP, et al. Pembrolizumab for the Treatment of non-Small-Cell Lung Cancer. *N Engl J Med* (2015) 372:2018–28. doi: 10.1056/NEJMoa1501824
2. Song Z, Chen X, Shi Y, Huang R, Wang W, Zhu K, et al. Evaluating the Potential of T Cell Receptor Repertoires in Predicting the Prognosis of Resectable Non-Small Cell Lung Cancers. *Mol Ther Methods Clin Dev* (2020) 18:73–83. doi: 10.1016/j.omtm.2020.05.020
3. Yang J, Hui Y, Zhang Y, Zhang M, Ji B, Tian G, et al. Application of Circulating Tumor DNA as a Biomarker for Non-Small Cell Lung Cancer. *Front Oncol* (2021) 11:725938. doi: 10.3389/fonc.2021.725938
4. Goodman AM, Kato S, Bazhenova L, Patel SP, Frampton GM, Miller V, et al. Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. *Mol Cancer Ther* (2017) 16:2598–608. doi: 10.1158/1535-7163.MCT-17-0386
5. Gandara DR, Paul SM, Kowanetz M, Schleifman E, Zou W, Li Y, et al. Blood-Based Tumor Mutational Burden as a Predictor of Clinical Benefit in Non-Small-Cell Lung Cancer Patients Treated With Atezolizumab. *Nat Med* (2018) 24:1441–8. doi: 10.1038/s41591-018-0134-3
6. Chan TA, Yarchoan M, Jaffee E, Swanton C, Quezada SA, Stenzinger A, et al. Development of Tumor Mutation Burden as an Immunotherapy Biomarker: Utility for the Oncology Clinic. *Ann Oncol* (2019) 30:44–56. doi: 10.1093/annonc/mdy495
7. Rizvi NA, Mazières J, Planchard D, Stinchcombe TE, Dy GK, Antonia SJ, et al. Activity and Safety of Nivolumab, an Anti-PD-1 Immune Checkpoint Inhibitor, for Patients With Advanced, Refractory Squamous non-Small-Cell Lung Cancer (CheckMate 063): A Phase 2, Single-Arm Trial. *Lancet Oncol* (2015) 16:257–65. doi: 10.1016/S1470-2045(15)70054-9
8. Samstein RM, Lee CH, Shoushtari AN, Hellmann MD, Shen R, Janjigian YY, et al. Tumor Mutational Load Predicts Survival After Immunotherapy Across Multiple Cancer Types. *Nat Genet* (2019) 51:202–6. doi: 10.1038/s41588-018-0312-8
9. Steuer CE, Ramalingam SS. Tumor Mutation Burden: Leading Immunotherapy to the Era of Precision Medicine? *J Clin Oncol* (2018) 36:631–2. doi: 10.1200/JCO.2017.76.8770
10. Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, et al. Analysis of 100,000 Human Cancer Genomes Reveals the Landscape of Tumor Mutational Burden. *Genome Med* (2017) 9:34. doi: 10.1186/s13073-017-0424-2
11. Buchhalter I, Rempel E, Endris V, Allgäuer M, Neumann O, Volckmar AL, et al. Size Matters: Dissecting Key Parameters for Panel-Based Tumor Mutational Burden Analysis. *Int J Cancer* (2019) 144:848–58. doi: 10.1002/ijc.31878
12. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, et al. CellProfiler: Image Analysis Software for Identifying and Quantifying Cell Phenotypes. *Genome Biol* (2006) 7:R100. doi: 10.1186/gb-2006-7-10-r100
13. Mcquin C, Goodman A, Chernyshev V, Kamentsky L, Cimini BA, Karhohs KW, et al. CellProfiler 3.0: Next-Generation Image Processing for Biology. *PLoS Biol* (2018) 16:e2005970. doi: 10.1371/journal.pbio.2005970
14. Yu KH, Berry GJ, Rubin DL, Ré C, Altman RB, Snyder M. Association of Omics Features With Histopathology Patterns in Lung Adenocarcinoma. *Cell Syst* (2017) 5:620–27.e623. doi: 10.1016/j.cels.2017.10.014
15. Yang J, Ju J, Guo L, Ji B, Shi S, Yang Z, et al. Prediction of HER2-Positive Breast Cancer Recurrence and Metastasis Risk From Histopathological Images and Clinical Information *via* Multimodal Deep Learning. *Comput Struct Biotechnol J* (2022) 20:333–42. doi: 10.1016/j.csbj.2021.12.028
16. Ye Z, Zhang Y, Liang Y, Lang J, Zhang X, Zang G, et al. Cervical Cancer Metastasis and Recurrence Risk Prediction Based on Deep Convolutional Neural Network. *Curr Bioinf* (2022) 17:164–73. doi: 10.2174/1574893616666210708143556
17. Yu KH, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, et al. Predicting non-Small Cell Lung Cancer Prognosis by Fully Automated Microscopic Pathology Image Features. *Nat Commun* (2016) 7:12474. doi: 10.1038/ncomms12474
18. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-Level Classification of Skin Cancer With Deep Neural Networks. *Nature* (2017) 542:115–8. doi: 10.1038/nature21056
19. Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, et al. Classification of Breast Cancer Histology Images Using Convolutional Neural Networks. *PLoS One* (2017) 12:e0177544. doi: 10.1371/journal.pone.0177544
20. Ehteshami Bejnordi B, Veta M, Johannes Van Diest P, Van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *Jama* (2017) 318:2199–210. doi: 10.1001/jama.2017.14585
21. Ma X, Xi B, Zhang Y, Zhu L, Sui X, Tian G, et al. A Machine Learning-Based Diagnosis of Thyroid Cancer Using Thyroid Nodules Ultrasound Images. *Curr Bioinf* (2020) 15:349–58. doi: 10.2174/1574893614666191017091959
22. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and Mutation Prediction From non-Small Cell Lung Cancer Histopathology Images Using Deep Learning. *Nat Med* (2018) 24:1559–67. doi: 10.1038/s41591-018-0177-5
23. Khosravi P, Kazemi E, Imielinski M, Elemento O, Hajirasouliha I. Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images. *EBioMedicine* (2018) 27:317–28. doi: 10.1016/j.ebiom.2017.12.026
24. He B, Lang J, Wang B, Liu X, Lu Q, He J, et al. TOOme: A Novel Computational Framework to Infer Cancer Tissue-Of-Origin by Integrating Both Gene Mutation and Expression. *Front Bioeng Biotechnol* (2020) 8:394. doi: 10.3389/fbioe.2020.00394
25. Liu H, Qiu C, Wang B, Bing P, Tian G, Zhang X, et al. Evaluating DNA Methylation, Gene Expression, Somatic Mutation, and Their Combinations in Inferring Tumor Tissue-Of-Origin. *Front Cell Dev Biol* (2021) 9:619330. doi: 10.3389/fcell.2021.619330
26. Macenko M, Niethammer M, Marron JS, Borland D, T. Woosley J, Guan X, et al. *A Method for Normalizing Histology Slides for Quantitative Analysis*. IEEE International Symposium on Biomedical Imaging: From Nano to Macro (2009). 1107–10 p
27. He K, Zhang X, Ren S, Sun J. *Deep Residual Learning for Image Recognition*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016). 770–8 p.
28. Kaplan EL, Meier P. Nonparametric Estimation From Incomplete Observations. *J Am Stat Assoc* (2012) 53:457–81. doi: 10.1080/01621459.1958.10501452
29. Natarajan S, Lipsitz SR, Fitzmaurice GM, Sinha D, Ibrahim JG, Haas J, et al. An Extension of the Wilcoxon Rank-Sum Test for Complex Sample Survey Data. *J R Stat Soc Ser C Appl Stat* (2012) 61:653–64. doi: 10.1111/j.1467-9876.2011.01028.x
30. Love MI, Huber W, Anders S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data With Deseq2. *Genome Biol* (2014) 15:550. doi: 10.1186/s13059-014-0550-8
31. Chen L, Zhang YH, Wang S, Zhang Y, Huang T, Cai YD. Prediction and Analysis of Essential Genes Using the Enrichments of Gene Ontology and KEGG Pathways. *PLoS One* (2017) 12:e0184129. doi: 10.1371/journal.pone.0184129
32. Liu C, Wei D, Xiang J, Ren F, Huang L, Lang J, et al. An Improved Anticancer Drug-Response Prediction Based on an Ensemble Method Integrating Matrix Completion and Ridge Regression. *Mol Ther Nucleic Acids* (2020) 21:676–86. doi: 10.1016/j.omtn.2020.07.003

33. Xu J, Cai L, Liao B, Zhu W, Yang J. CMF-Impute: An Accurate Imputation Tool for Single-Cell RNA-Seq Data. *Bioinformatics* (2020) 36:3139–47. doi: 10.1093/bioinformatics/btaa109

34. Meng Y, Lu C, Jin M, Xu J, Zeng X, Yang J. A Weighted Bilinear Neural Collaborative Filtering Approach for Drug Repositioning. *Brief Bioinform* (2022) 23(2):bbab581. doi: 10.1093/bib/bbab581

Conflict of Interest: The authors LW, KH, CR, GT, and XS are employed by Geneis Beijing Co. Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.