

RESEARCH

Open Access



neoDL: a novel neoantigen intrinsic feature-based deep learning model identifies IDH wild-type glioblastomas with the longest survival

Ting Sun^{1†}, Yufei He^{1†}, Wendong Li^{1†}, Guang Liu¹, Lin Li¹, Lu Wang¹, Zixuan Xiao¹, Xiaohan Han¹, Hao Wen¹, Yong Liu¹, Yifan Chen¹, Haoyu Wang¹, Jing Li¹, Yubo Fan^{1*}, Wei Zhang^{2,3*} and Jing Zhang^{1*}

*Correspondence: yubofan@buaa.edu.cn; zhangwei_vincent@126.com; jz2716@buaa.edu.cn

[†]Ting Sun, Yufei He and Wendong Li have contributed equally to this work

¹ Key Laboratory for Biomechanics and Mechanobiology of Ministry of Education, Beijing Advanced Innovation Centre for Biomedical Engineering, School of Engineering Medicine, School of Biological Science and Medical Engineering, Beihang University, No.37 Xueyuan Road, Haidian District, Beijing 100083, People's Republic of China

³ Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, No. 119 South Fourth Ring Road West, Fengtai District, Beijing 100070, People's Republic of China
Full list of author information is available at the end of the article

Abstract

Background: Neoantigen based personalized immune therapies achieve promising results in melanoma and lung cancer, but few neoantigen based models perform well in IDH wild-type GBM, and the association between neoantigen intrinsic features and prognosis remain unclear in IDH wild-type GBM. We presented a novel neoantigen intrinsic feature-based deep learning model (neoDL) to stratify IDH wild-type GBMs into subgroups with different survivals.

Results: We first derived intrinsic features for each neoantigen associated with survival, followed by applying neoDL in TCGA data cohort (AUC = 0.988, p value < 0.0001). Leave one out cross validation (LOOCV) in TCGA demonstrated that neoDL successfully classified IDH wild-type GBMs into different prognostic subgroups, which was further validated in an independent data cohort from Asian population. Long-term survival IDH wild-type GBMs identified by neoDL were found characterized by 12 protective neoantigen intrinsic features and enriched in development and cell cycle.

Conclusions: The model can be therapeutically exploited to identify IDH wild-type GBM with good prognosis who will most likely benefit from neoantigen based personalized immunotherapy. Furthermore, the prognostic intrinsic features of the neoantigens inferred from this study can be used for identifying neoantigens with high potentials of immunogenicity.

Keywords: IDH wild-type glioblastoma, Peptide-features, Prognosis, Deep learning, Immunology

Background

Glioblastoma is the most common aggressive primary brain tumor having profound genomic heterogeneity and high recurrence rate [1]. Although the survival of GBMs has improved with the advancement of modern combination therapies, the prognosis of



most GBMs remains poor and varies considerably among patients [2], revealing a dismal median duration of 14 months [3, 4].

Neoantigens are from mutation-containing proteins that generate novel immunogenic epitopes [5]. High nonsynonymous mutation loads harbor more neoantigens presented to CD8+ T cells on restricted HLA-I subtypes [6–8], leading to stronger immunogenicity and better overall survival in melanoma [9], lung cancer [10], and colorectal tumors [11]. However, in gliomas, higher mutational load means increased tumor aggressiveness [12]. Neoantigens are pivotal in personalized immunotherapies, promoting tumor-specific T-cell responses and affecting antitumor immune responses in a number of preclinical models [13, 14]. Although high-quality neoantigen model performed well in identifying IDH wild-type GBMs with the longest survival [15], the number of high quality neoantigens were limited, making clinical application difficult. The occurrence and characterization of neoantigen in pan-cancer showed that all positions in neopeptides containing more hydrophobic residues than the wild-type [16], but the comprehensive features of neoantigens associated with prognosis and immunoreaction in IDH wild-type GBM remain elusive.

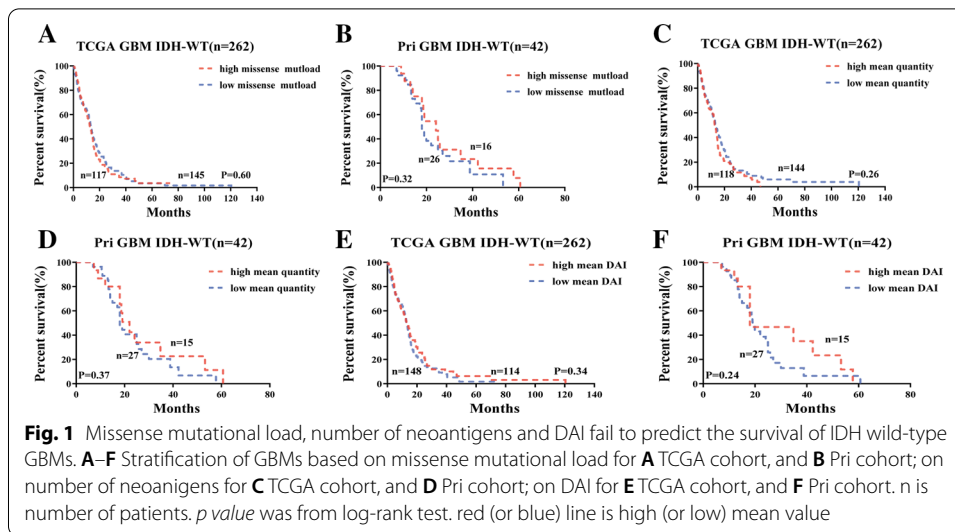
Deep learning models can derive features from noisy and raw data by learning high-level representations [17, 18]. Their flexibility and adaptability lead to their wide application in biomedical imaging [19], showing excellent level-accuracy in precise diagnosis and prognostic stratification of colorectal [20], prostate [21, 22], melanoma [23], and gliomas [24]. Deep learning also demonstrates its strong abilities in predicting Glioma grades [25], Glioma genetic mutation [26] and survival [27]. Recently, neoantigen-based machine learning is reported to predict neoantigen immunogenicity in colon and lung adenocarcinomas [28].

Here, we present a neoantigen intrinsic feature based deep learning model (neoDL), successfully stratifying IDH wild-type GBMs of TCGA into different prognostic subgroups (Additional file 1: Figure S1). Our model was further validated in an independent data from Asian population, even demonstrating its strong predictive power in some higher-grade gliomas, including Classical, Classical-like, Glioblastoma, IDH wild-type, Mesenchymal-like. GBMs identified by neoDL with better prognosis enriched in development, and cell cycle. Our neoDL has important implications in diagnosis and prognosis of IDH wild-type GBMs, and helps identify GBMs who most likely benefit from neoantigen based personalized immunotherapy.

Results and discussion

Identification of neoantigen intrinsic features associated with the overall survival of IDH wild-type GBMs

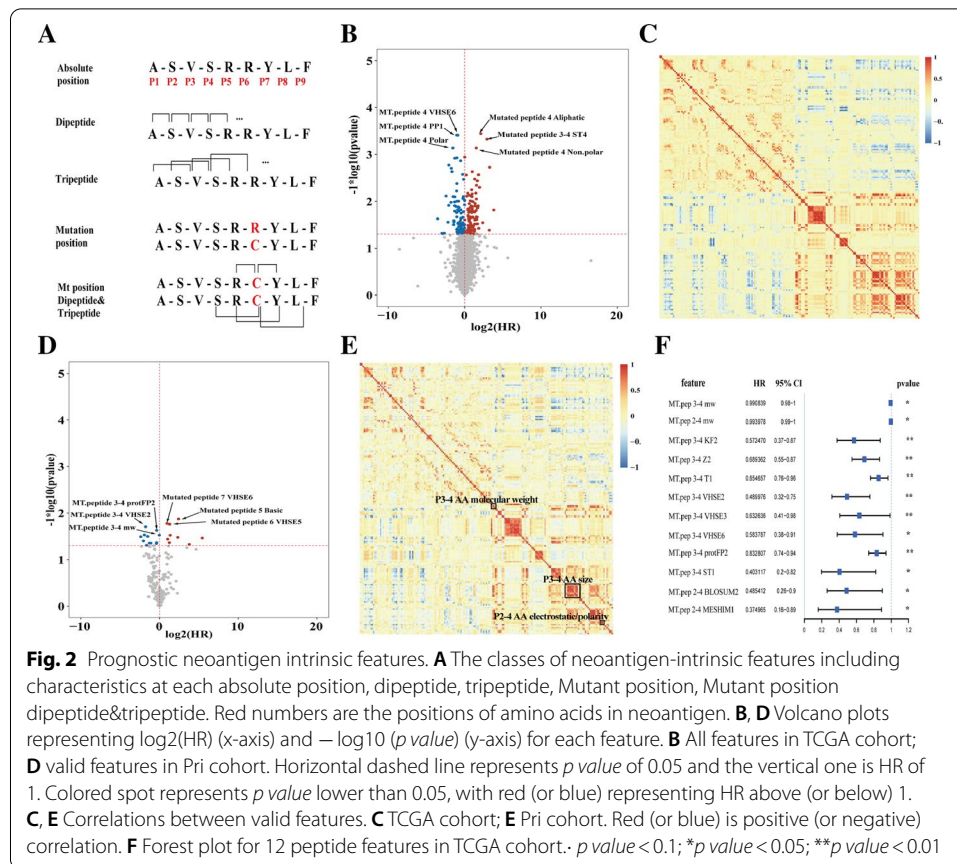
Tumor mutational burden has been described as a predictor of tumor behavior and immunological response [29], with improved survival and immunotherapy response in melanomas [30], ovarian [31], and bladder carcinoma [32]. We calculated missense mutational load for 262 and 42 IDH wild-type GBMs in TCGA and Pri-cohort, respectively, finding no statistically significantly different overall survival between higher and lower mutation loads (Fig. 1A, B), consistent with the previous research [15]. Similarly, mutation loads were found either not prognostic or related to worse survival in 16 different glioma subtypes (Additional file 1: Figure S2). High missense mutational load



harbored more neoantigens, rendering them more susceptible T-cell targets [33]. The neoantigen quantity also failed to predict the survival of IDH wild-type GBMs (Fig. 1C, D) and 16 different glioma subgroups (Additional file 1: Figure S3). DAI, defined as difference between binding affinity of wildtype and mutant-type peptides for MHC class I, was reported to be a better predictor of survival and immunogenicity in advanced lung cancer and melanoma [34]. We calculated the average DAI of each sample in both TCGA and Pri cohort, finding that DAI model failed in predicting the overall survival of IDH wild-type GBMs (Fig. 1E, F) and 16 different glioma subgroups (Additional file 1: Figure S4).

An immunogenic neoantigen must possess structural and physical properties distinct enough to promote efficient recognition by T cells [35]. We calculated a total of 2928 features for each neoantigen and its wild-type peptide, including physical–chemical properties, AA (amino acid) features, and AA descriptors at each absolute position, composed-dipeptide and tripeptide at the site of mutation, and the dipeptides and tripeptides related to the mutation site, and complete sequence (Fig. 2A). The Shannon entropy and the AA composition were also calculated. We then performed Cox regression to estimate the association between the feature values and overall survival in IDH wild-type GBMs of TCGA, finding 189 prognostic features (termed as valid features) (Fig. 2B), among which the most significant positive associations were aliphatic AA in the absolute site 4 (Mutated peptide 4 Aliphatic), ST-scales4 descriptors of site 3 and 4 compose-dipeptide (Mutated peptide 3–4 ST4), and Nonpolar AA in the absolute site 4 (Mutated peptide 4 Non.polar). The most significant negative associations were the VHSE-scales6 descriptors, PP1 descriptors, and polar AA at the absolute position 4 (MT.peptide 4 VHSE6, MT.peptide 4 PP1, MT.peptide 4 polar). After calculating the correlation of valid features, we discovered that correlated feature modules were consistent across IDH wild-type GBM (Fig. 2C) and 16 different glioma subtypes in TCGA cohort (Additional file 1: Figure S5).

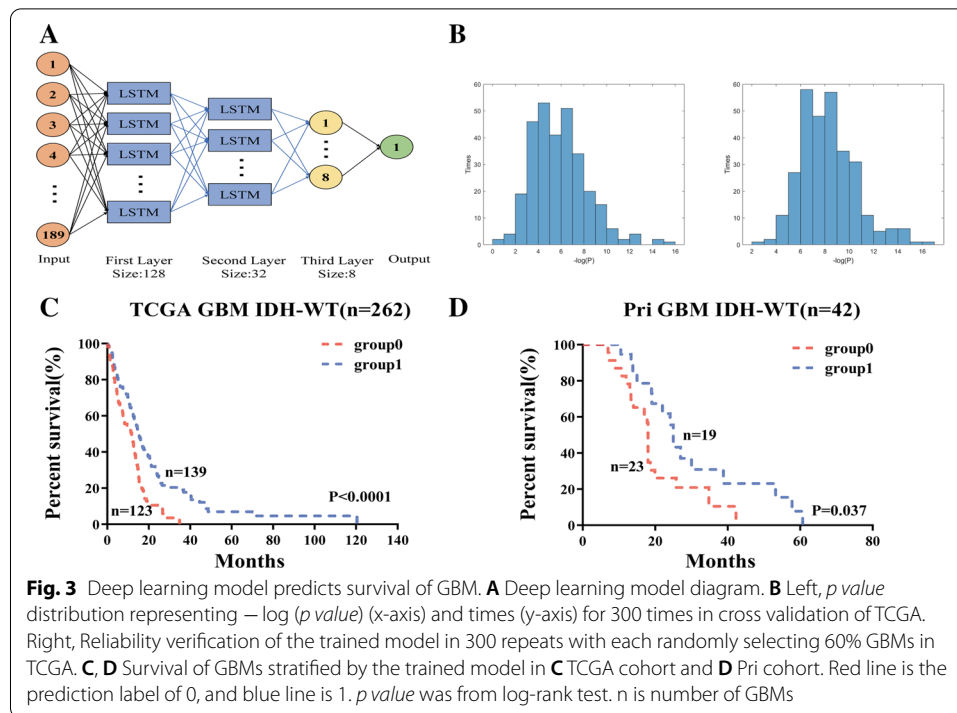
To further evaluate the prognostic value of valid features, we conducted Cox regression analysis in an independent data of Pri cohort, revealing 22 valid features significantly



associated with the overall survival (Fig. 2D). The most significant positive associations were VHSE-scales6 at the 7 sites (Mutated peptide 7 VHSE6), basic AA at the site 5 (Mutated peptide 5 basic), VHSE-scales5 at the 6 site (Mutated pep 6 VHSE5). The most significant negative associations were mainly related to the characteristics of the positions 3 and 4 composed-dipeptide, including protFP2, VHSE-scales2, and molecular weight. Particularly, 12 features had shown strong mutual correlation, mainly associated with the molecular weight and molecular size/volume of the position 3,4 composed-dipeptide, and molecular electrostatic of the position 2–4 composed-tripeptide (Fig. 2E). Moreover, the 12 features were protective factors ($\text{HR} < 1$) in both TCGA cohort (Fig. 2F) and Pri cohort (Additional file 1: Figure S6).

Deep learning model using neoantigen intrinsic features predicted IDH wild-type GBMs with better survival

Deep learning methods learn high-level representations with multilayer computational models, and are advantageous in learning high-dimensional datasets [17]. LSTM can avoid the problem of vanishing gradient [36], and has the ability to remember all previous data. To stratify IDH wild-type GBMs, we constructed a valid feature-based deep learning model including three hidden layers (two LSTM layers and one fully connected layer) with 128, 32, 8 nodes, respectively (Fig. 3A). We chose the Sigmoid function as neuron activation function for fully connected layer, MSE as the loss function and Adam



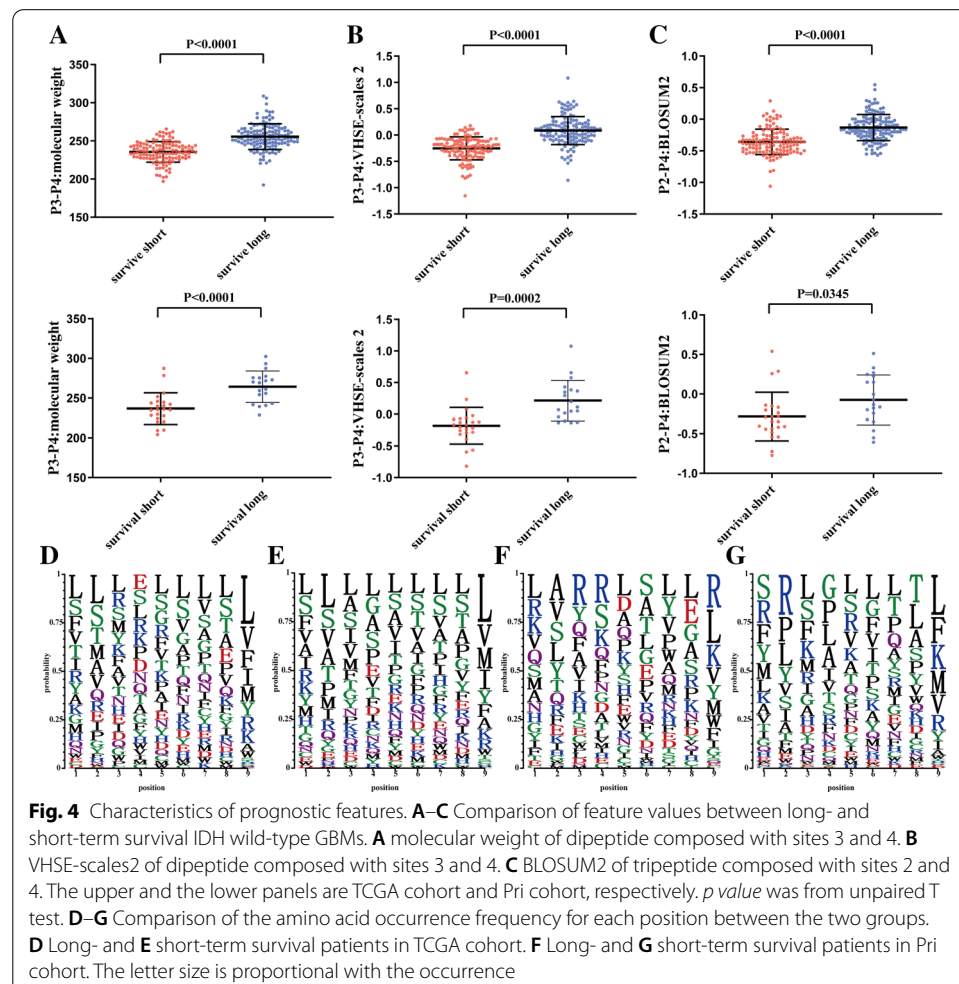
as the iterative optimizer with the number of iterations set as 1000. When setting 1000 epochs when training the model, loss approaches to zero and accuracy approached to 100%. Predicting accuracy in cross validation continuously remained at a high level (over 90%), showing that the model was not over-fitting (Additional file 1: Figure S7). The samples in TCGA cohort (containing 262 labeled samples) were used as training data, while the samples in Pri cohort (containing 42 unlabeled samples) as external testing data. TCGA cohort was labeled based on the result of hierarchical k-means clustering, which stratified the data into a short-term survival group (cluster = 1, $n = 126$) and a long-term survival group (cluster = 2, $n = 136$).

To validate the reliability of the deep learning model, we performed 300 random trials with each splitting the samples into training set and testing set at the ratio of six vs four. The two sets were extracted separately from short- and long-term survival group with the specific ratio, thus the training set contains 60% of cluster 1 samples ($n = 76$) and 60% cluster 2 samples ($n = 82$). In each trial, the parameters learned in the training set were applied in the testing set. In 275 out of 300 trials, IDH wild-type GBMs in TCGA were successfully separated into two significantly different prognostic subgroups (p value < 0.05) (Fig. 3B left). The optimal parameter settings were determined and applied to randomly selected 60% of IDH wild-type GBMs in TCGA. In 299 of 300 randomly selected 60% of IDH wild-type GBMs in TCGA, our trained deep learning model successfully separated patients into two subgroups with significantly different overall survival (Fig. 3B right), demonstrating the stability and reliability of our model. We then applied the trained model to stratify all IDH wild-type GBMs in TCGA into two prognostic subgroups (AUC = 0.988, p value < 0.0001 , Fig. 3C, Additional file 1: Table S7). As an independent validation, we successfully applied the trained model to separate IDH wild-type GBMs in an independent data (Pri GBM cohort) into two prognostic

subgroups (p value = 0.037, Fig. 3D). We also successfully applied the trained model to divide patients into two different prognostic subgroups for GBM, IDH wildtype, Classical, Classical-like, Mesenchymal-like subtypes in TCGA pan-glioma cohort (p value < 0.05 for all subtypes) (Additional file 1: Figure S8). The flow chart of the neoDL model was visualized (Additional file 1: Figure S1).

The prognostic characteristics of 12 protective intrinsic features

To characterize the 12 protective intrinsic features in the molecular weight, molecular size of dipeptide, and molecular electrostatic potential of tripeptide, we compared their distributions in the short- and long-term survival IDH wild-type GBMs. Compared with the short-term survival GBMs, the long-term survival patients exhibited statistically significantly higher molecular weight of dipeptide at the site 3 and 4 (p value < 0.05; Fig. 4A; Additional file 1: Figure S9a), molecular size-related features (Kidera Factors 2, Z-scale 2, T-scale 1, protFP2, VHSE-sclae 2, VHSE-sclae 3, VHSE-sclae 6, ST-scale 1) (p value < 0.05; Fig. 4B; Additional file 1: Figure S9b) and the electrostatic potential related features (BLOSUM2 and MESHIM1) (p value < 0.05; Fig. 4C; Additional file 1: Figure S9c) in both TCGA and Pri-cohort.



Univariate and multi-variate Cox regression [37] analysis demonstrated that two of 12 features (VHSE2 and protFP2) were associated with the overall survival in the two cohorts (Additional file 1: Table S1, Additional file 1: Table S2, Additional file 1: Table S3 and Additional file 1: Table S4). Kaplan Meier analysis demonstrated statistically significantly different overall survival between the low-value (below mean) and high-value (above mean) groups of IDH wild-type GBMs stratified by the two features. The patients with high-value (above mean) had a significantly longer overall survival (for protFP2: p value = 0.002 in TCGA cohort and p value = 0.03 in Pri cohort; for VHSE2: p value = 0.018 in TCGA cohort and p value = 0.11 in Pri cohort) (Additional file 1: Figure S10a–b). Furthermore, the two feature-based stratification of the IDH wild-type GBMs were found independent of age and mutational load. The two features also exhibited strong correlations ($R=0.87$, p value $<2.2e-16$ for TCGA; $R=0.91$, p value $<2.2e-16$ for Pri Cohort) (Additional file 1: Figure S10c).

The distributions of amino acid residue for neoantigens between long- and short-term survival groups were examined, revealing that the ratios of amino acid residues at positions 3 and 4 were significantly different (Fig. 4D–G). At the site 3, the patients with neoantigens containing a lower frequency of L and S amino acids and a higher frequency of R amino acid survived longer than those with the opposite frequencies in both cohorts. The enrichment of residues R and S at site 4 of neoantigens were evident in the long-term survival of IDH wild-type GBMs. The ratios of L and G at site 4 of neoantigens increased in the short-term survival patients.

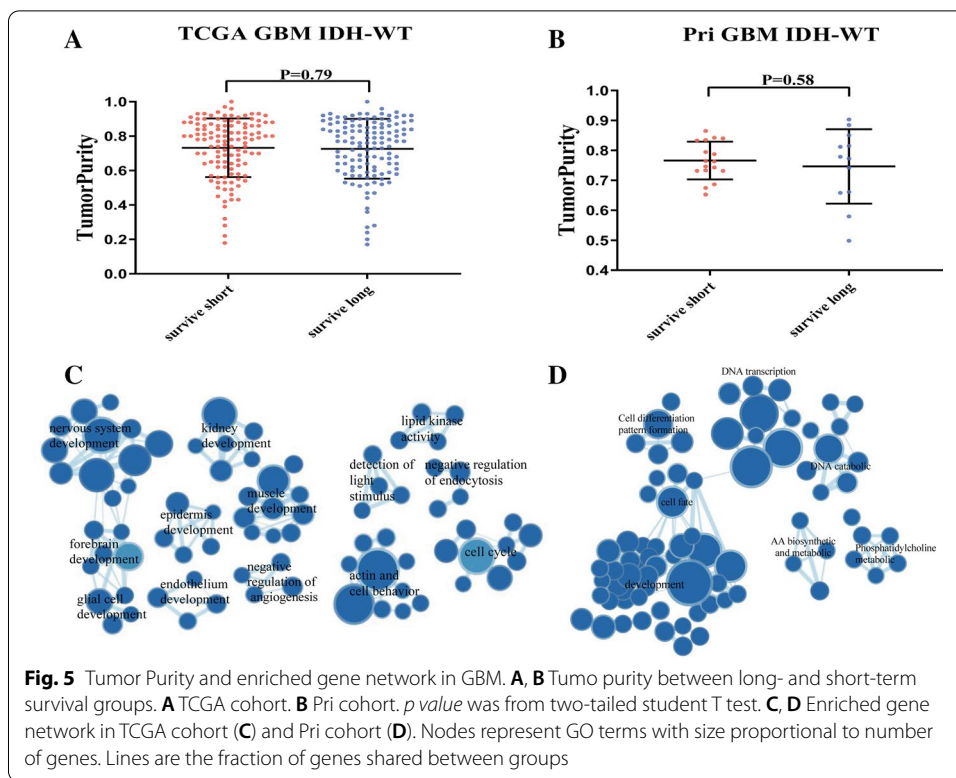
Tumor purity and functional annotation of gene expression in GBM

We calculated the tumor purity, immune score, and stromal score using gene expression data for each patient in both TCGA and Pri cohorts. No significant differences were observed between long- and short-term survival of IDH wild-type GBMs (Fig. 5A, B for tumor purity, Additional file 1: Figure S11a for immune scores and S11b for stromal scores). No correlations were discovered between purity levels and mutational burden (Additional file 1: Figure S11C).

To understand the mechanisms in transcriptomic architecture, we conducted GSEA [38, 39], an algorithm for determining whether a set of genes differs between two biological states, between long- and short-term survival groups of IDH wild-type GBMs in both TCGA and Pri cohorts, respectively. Enrichment map analysis of deregulated GO terms in TCGA data demonstrated that GO terms related to development and cell cycle were enriched in long-term survival patients (Fig. 5C, Additional file 1: Table S5 and Additional file 1: Table S6). In Pri cohort, the most significant biological processes enriched in longer-survived GBMs were development associated GO terms such as epidermis development, cell cycle, which were also identified in TCGA cohort (Fig. 5D).

Conclusion

In this paper, we presented a prognostic prediction deep learning model based on neoantigen intrinsic features. Although several survival prediction models have been reported based on the expression of several genes [40–42] or medical images [43, 44], they are not related to neoantigens and immune response. As neoantigens are associated with tumor-specific T-cell responses and anti-tumor immune responses, the method we



provided can help predict the prognosis of IDH wild-type GBM patients who will likely benefit from neoantigen based personalized immunotherapy.

Our model achieved good predictive performances in two independent data cohorts of IDH wild-type GBMs (KM: log rank *p* value < 0.0001 in TCGA cohort; 0.037 in Pri cohort) and even in some other high-grade glioma subtypes. Currently, the vast majority of deep learning models (such as DeepLearningModel [45] and PASNet [46]) are based on gene expression, clinical information and medical image data for learning modeling, and there are few predictions of GBM patient survival based on the nature of neoantigens. We compared our neoDL with them and found that neoDL performed better than DeepLearningModel and PASNet (Additional file 1: Table S7). GBMs predicted by our model to have better survival enriched in development and cell cycle. Two correlated neoantigen features (VHSE2 and protFP2) were identified to stratify GBMs into a high- and low-value subgroup with significant different survival independent of other clinical and pathological characteristics.

Of 189 valid features, 12 protective features associated with survival in both cohorts were amino acid molecular weight, molecular size/volume, and electrostatic potential/polarity, which were characterized by close relation with the amino acid properties at the positions 3 and 4 of the neoantigen, confirmed by the amino acid distributions between different survival groups. The features at the site 3 and 4 of the neoantigen may have potential effects on the survival of GBMs and immunotherapy response, and they are worthy of further investigation.

In this study, we focused on sequence structure in this study, but not on secondary and tertiary protein structure. More features may be integrated into the model to improve

predictive power, which shall be resolved in the future. The deep learning methods (such as DeepCoxPH [47] and FuzzyDeepCoxPH [48]) reported to be effective in other scenario can also be used to augment the prognostic evaluation and improve decision-making in glioma. To predict the patients' outcome, more studies related to generalizability test are still in need.

Methods

Data description

Mutations and clinical information were from the ATLAS-TCGA pan-glioma study [49]. Gene expression data (G4502A) at level 3 were from TCGA Data portal. We termed the data from TCGA as TCGA cohort. Mutations, RNAseq data, and clinical information in Asian population were from a recently published cohort [50], designated as Pri cohort. The samples that not diagnosed as IDH wild-type GBM or have clinical information lost were removed, resulting in 268 and 46 samples in the two cohorts, respectively.

A neoepitope with strong affinity for MHC ($IC_{50} \leq 500$ nM) may be a more robust neoantigen candidate if the paired wild-type epitope has a poor affinity for MHC ($IC_{50} > 500$ nM) [51]. The neoantigens and their corresponding wild-type peptides for each sample in TCGA cohort and Pri cohort were from our previous study [15], which used missense mutations to generate all possible 9-mer peptides and defined the mutant 9-mer peptides as neoantigens when the IC_{50} of mutant-type peptides was < 500 nM and the corresponding wild-type binder > 500 nM.

Feature calculation for neoantigens

262 (TCGA cohort) and 42 samples (Pri cohort) with detected neoantigens remained in the downstream analysis. 2928 features (Additional file 1: Table S8) were extracted from 2263 neoantigens (2081 for TCGA cohort; 182 for Pri cohort) using R: 'Peptides'(v2.4.2) for 66 amino acid descriptors and 10 physical–chemical properties, aaComp for amino acid composition of neoantigens, and custom scripts for features from Shannon entropy (Additional file 1).

Prognostic feature selection

The features were calculated for all neoantigens and wild-type peptides, followed by averaging all feature values in each patient. Univariate Cox regression analysis was to predict the prognostic impact of each feature. 189 features with p value ≤ 0.05 were termed as valid features (Additional file 1: Table S9). Correlation matrix of the valid features were visualized through heatmaps using R: 'pheatmap'.

Hierarchical k-means clustering

Hierarchical k-means clustering was applied upon Z-Score-transformed valid features to stratify patients into two clusters using the "hkmeans" command of the R: 'factoextra' (version 1.0.7).

Deep-learning model construction

The valid features in TCGA cohort were used to train deep learning model. The groups from hierarchical k-means clustering were used as labels. Z-Score-transformed were

applied upon feature values of valid features to avoid gradient disappearance problem. The LSTM (Long short-term memory) deep learning model was built with three hidden layers (two LSTM layers and one fully connected layer), with each containing 128, 32, and 8 nodes, respectively. We chose the Sigmoid function as neuron activation function for fully connected layer, since we wanted to map the original statistics to a single number with domain of 0–1 through learning, which referred to the final classification result. The original data were normalized using z-score, therefore no serious gradient vanishing problem would be caused when using Sigmoid function as activation function. For hyper-parameters, we chose MSE as the loss function and Adam as the iterative optimizer with the number of iterations set as 1000. MSE is a commonly used loss function in regression problem, thus we utilized such function to calculate the preference of a sample. The initial connection weights and biases of each layer were randomly generated, and end up reaching stable parameters through training iterations.

Leave one out cross validation (LOOCV)

Cross validation was performed as follows. TCGA cohort was randomly separated into training and test sets at the ratio of six to four. To obtain the optimal model, the above randomizations were conducted 300 times. For each randomization trial, the model parameters were trained in the training sets. The trained model was applied to stratify the test set into two subgroups, followed by Kaplan–Meier survival analysis. *p* value ≤ 0.05 were regarded as statistically significant. The optimal parameter settings were determined from 300 randomization trials. To evaluate the reliability, the trained model were then applied to randomly selected 60% of IDH wild-type GBMs in TCGA, which were repeated 300 times.

Independent validation

Pri cohort was used as an external test data to test the performance of the trained model, which divided patients into long- and short-term survival clusters. Other glioma subtypes from TCGA were also used to test the trained model, including Astrocytoma, Classical-like, Classical, Codel, Glioblastoma, G-CIMP-high, IDH-MT-codel, IDH-MT-noncodel, IDH-MT, IDH-WT, Mesenchymal-like, Mesenchymal, Neural, Oligodendroglioma, Proneural and OligoAstrocytoma.

Tumor purity estimation

Tumor purities were estimated by ESTIMATE [52] using R: 'estimate'(version 1.6.7). There were 242 (TCGA cohort) and 29 IDH wild-type GBMs (Pri cohort) with gene expression profiles available.

GO enrichment analysis

GO enrichment analysis was conducted using Gene Set Enrichment Analysis (GSEA 4.0.3). The GO terms were from the Molecular Signatures Database (c5.all.v6.2.symbols.gmt). Gene sets with $FDR < 0.05$ were considered as differentially expressed, and visualized using Cytoscape [53]. The GSEA results were shown in Additional file 1: Table S5 and Additional file 1: Table S6.

Statistical analysis

Variables between groups were compared by the unpaired T test, a Parametric test method which compares two different subjects. Correlations were evaluated by Pearson correlations. Kaplan–Meier survival and Cox regression analyses were performed using R: "survminer" and "survival". p value ≤ 0.05 was determined as significance in all tests. All analyses were conducted in R and Python.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-021-04301-6>.

Additional file 1. Description of neoDL and supplementary results.

Acknowledgements

The authors would like to thank Wei Shi at Beihang University for her valuable advice in general.

Authors' contributions

Conceptualization: JZ, WZ, YBF; Methodology: TS, YFH, WDL, GL, JZ, WZ, YBF; Data curation: GL, LL, WDL, LW, ZXX, XHH, HW, YL, YFC, HYW, and JL; Writing-review and editing: TS, YFH, ZXX, WZ, YBF, JZ; Supervision: WZ, YBF, JZ; Funding acquisition: WZ, YBF, JZ. All authors read and approved the final manuscript.

Funding

This work was supported by Youth Thousand Scholar Program of China (J.Z.) and Program for High-Level Overseas Talents, Beihang University (J.Z.), National Natural Science Foundation of China (NSFC No. 11421202, and 11827803 to YBF, No. 81672479 to W.Z.), National Natural Science Foundation of China (NSFC)/Research Grants Council (RGC) Joint Research Scheme (81761168038) (W.Z.), Beijing Municipal Administration of Hospitals' Mission Plan (SML20180501) (W.Z.).

Availability of data and materials

All data are from original researches properly cited in Material and methods. neoDL and the intrinsic features of neoantigens calculated for both TCGA cohort and Pri cohort are at github (<https://github.com/zhangbig/neoDL>).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Key Laboratory for Biomechanics and Mechanobiology of Ministry of Education, Beijing Advanced Innovation Centre for Biomedical Engineering, School of Engineering Medicine, School of Biological Science and Medical Engineering, Beihang University, No.37 Xueyuan Road, Haidian District, Beijing 100083, People's Republic of China. ²Department of Molecular Neuropathology, Beijing Neurosurgical Institute, Capital Medical University, Beijing 100070, People's Republic of China. ³Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, No. 119 South Fourth Ring Road West, Fengtai District, Beijing 100070, People's Republic of China.

Received: 26 January 2021 Accepted: 7 July 2021

Published online: 23 July 2021

References

1. Fabian D, Guillermo Prieto Eibl MDP, Alnahhas I, Sebastian N, Giglio P, Puduvali V, Gonzalez J, Palmer JD. Treatment of glioblastoma (GBM) with the addition of tumor-treating fields (TTF): a review. *Cancers (Basel)*. 2019;11(2):174.
2. Mahlokoza T, Vellimana AK, Li T, Mao DD, Zohny ZS, Kim DH, Tran DD, Marcus DS, Fouke SJ, Campian JL, et al. Biological and therapeutic implications of multisector sequencing in newly diagnosed glioblastoma. *Neuro Oncol*. 2018;20(4):472–83.
3. Buckner JC. Factors influencing survival in high-grade gliomas. *Semin Oncol*. 2003;30(6 Suppl 19):10–4.
4. Van Meir EG, Hadjipanayis CG, Norden AD, Shu HK, Wen PY, Olson JJ. Exciting new advances in neuro-oncology: the avenue to a cure for malignant glioma. *CA Cancer J Clin*. 2010;60(3):166–93.
5. Gubin MM, Artyomov MN, Mardis ER, Schreiber RD. Tumor neoantigens: building a framework for personalized cancer immunotherapy. *J Clin Investig*. 2015;125(9):3413–21.

6. McGranahan N, Furness AJ, Rosenthal R, Ramskov S, Lyngaa R, Saini SK, Jamal-Hanjani M, Wilson GA, Birkbak NJ, Hiley CT, et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*. 2016;351(6280):1463–9.
7. McGranahan N, Rosenthal R, Hiley CT, Rowan AJ, Watkins TBK, Wilson GA, Birkbak NJ, Veeriah S, Van Loo P, Herrero J, et al. Allele-specific HLA loss and immune escape in lung cancer evolution. *Cell*. 2017;171(6):1259–1271.e1211.
8. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. *Science*. 2015;348(6230):69–74.
9. Lennerz V, Fatho M, Gentilini C, Frye RA, Lifke A, Ferel D, Wolfel C, Huber C, Wolfel T. The response of autologous T cells to a human melanoma is dominated by mutated neoantigens. *Proc Natl Acad Sci U S A*. 2005;102(44):16013–8.
10. Zeneyedpour L, Dekker LJM, van Sten-vant THJJM, Burgers PC, Ten Hacken NHT, Luider TM. Neoantigens in chronic obstructive pulmonary disease and lung cancer: a point of view. *Proteomics Clin Appl*. 2019;13(2):e1800093.
11. Giuseppe Rospo AL, Amirouchene-Angelozzi N, et al. Evolving neoantigen profiles in colorectal cancers with DNA repair defects. *Genome Med*. 2019;11(1):42.
12. Draaisma KWMMJ, Weenink B, et al. PI3 kinase mutations and mutational load as poor prognostic markers in diffuse glioma patients. *Acta Neuropathol Commun*. 2015;3(1):88.
13. Castle JC, Kreiter S, Diekmann J, Lower M, van de Roemer N, de Graaf J, Selmi A, Diken M, Boegel S, Paret C, et al. Exploiting the mutanome for tumor vaccination. *Cancer Res*. 2012;72(5):1081–91.
14. Kranz LM, Diken M, Haas H, Kreiter S, Loquai C, Reuter KC, Meng M, Fritz D, Vascotto F, Hefesha H, et al. Systemic RNA delivery to dendritic cells exploits antiviral defence for cancer immunotherapy. *Nature*. 2016;534(7607):396–401.
15. Zhang J, Caruso FP, Sa JK, Justesen S, Nam DH, Sims P, Ceccarelli M, Lasorella A, Iavarone A. The combination of neoantigen quality and T lymphocyte infiltrates identifies glioblastomas with the longest survival. *Commun Biol*. 2019;2:135.
16. Teku GN, Vihinen M. Pan-cancer analysis of neoepitopes. *Sci Rep*. 2018;8(1):12735.
17. Kalinin AA, Higgins GA, Reamaroon N, Soroushmehr S, Allyn-Feuer A, Dinov ID, Najarian K, Athey BD. Deep learning in pharmacogenomics: from gene regulation to patient stratification. *Pharmacogenomics*. 2018;19(7):629–50.
18. Lecun YBY, Hinton G. Deep learning. *Nature*. 2015;521(7553):436.
19. Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Brief Bioinform*. 2017;18(5):851–69.
20. Bychkov D, Linder N, Turkki R, Nordling S, Kovanen PE, Verrill C, Walliander M, Lundin M, Haglund C, Lundin J. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci Rep*. 2018;8(1):3395.
21. Nagpal K, Foote D, Liu Y, Chen PC, Wulczyn E, Tan F, Olson N, Smith JL, Mohtashamian A, Wren JH, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit Med*. 2019;2:48.
22. Tolkach YDT, Toma M, et al. High-accuracy prostate cancer pathology using deep learning. *Nat Mach Intell*. 2020;2(7):411–8.
23. Harder N, Schonmeyer R, Nekolla K, Meier A, Briue N, Vanegas C, Madonna G, Capone M, Botti G, Ascierto PA, et al. Automatic discovery of image-based signatures for ipilimumab response prediction in malignant melanoma. *Sci Rep*. 2019;9(1):7449.
24. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. *Nat Commun*. 2015;6:8971.
25. Yang Y, Yan LF, Zhang X, Han Y, Nan HY, Hu YC, Hu B, Yan SL, Zhang J, Cheng DL, et al. Glioma grading on conventional MR images: a deep learning study with transfer learning. *Front Neurosci*. 2018;12:804.
26. Chang P, Grinband J, Weinberg BD, Bardis M, Khy M, Cadena G, Su MY, Cha S, Filippi CG, Bota D, et al. Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. *AJNR Am J Neuroradiol*. 2018;39(7):1201–7.
27. Lao J, Chen Y, Li ZC, Li Q, Zhang J, Liu J, Zhai G. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Sci Rep*. 2017;7(1):10353.
28. Smith CC, Chai S, Washington AR, Lee SJ, Landoni E, Field K, Garness J, Bixby LM, Selitsky SR, Parker JS, et al. Machine-learning prediction of tumor antigen immunogenicity in the selection of therapeutic epitopes. *Cancer Immunol Res*. 2019;7(10):1591–604.
29. Goodman AMKS, Bazhenova L, et al. Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Mol Cancer Ther*. 2017;16(11):2598–608.
30. Gupta S, Artomov M, Goggins W, Daly M, Tsao H. Gender disparity and mutation burden in metastatic melanoma. *J Natl Cancer Inst*. 2015;107(11):dvj221.
31. Birkbak NJ, Kochupurakkal B, Izzarzugaza JM, Eklund AC, Li Y, Liu J, Szallasi Z, Matulonis UA, Richardson AL, Iglehart JD, et al. Tumor mutation burden forecasts outcome in ovarian cancer with BRCA1 or BRCA2 mutations. *PLoS ONE*. 2013;8(11):e80023.
32. Klebanov N, Artomov M, Goggins WB, Daly E, Daly MJ, Tsao H. Burden of unique and low prevalence somatic mutations correlates with cancer survival. *Sci Rep*. 2019;9(1):4848.
33. Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, Schrock A, Campbell B, Shlien A, Chmielecki J, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med*. 2017;9(1):34.
34. Ghorani ERR, McGranahan N, et al. Differential binding affinity of mutated peptides for MHC class I is a predictor of survival in advanced lung cancer and melanoma. *Ann Oncol*. 2018;29(1):271–9.
35. Riley TP, Keller GLJ, Smith AR, Davancaze LM, Arbuiso AG, Devlin JR, Baker BM. Structure based prediction of neoantigen immunogenicity. *Front Immunol*. 2019;10:2047.
36. Munir KEH, Ayub A, et al. Cancer diagnosis using deep learning: a bibliographic review. *Cancers*. 2019;11(9):1235.
37. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B Methodol*. 1972;34:187–202.
38. Mootha V, Lindgren C, Eriksson KF, et al. PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*. 2003;34:267–73.
39. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–50.

40. Cao M, Cai J, Yuan Y, Shi Y, Wu H, Liu Q, Yao Y, Chen L, Dang W, Zhang X, et al. A four-gene signature-derived risk score for glioblastoma: prospects for prognostic and response predictive analyses. *Cancer Biol Med*. 2019;16(3):595–605.
41. Prasad B, Tian Y, Li X. Large-scale analysis reveals gene signature for survival prediction in primary glioblastoma. *Mol Neurobiol*. 2020;57(12):5235–46.
42. Zuo S, Zhang X, Wang L. A RNA sequencing-based six-gene signature for survival prediction in patients with glioblastoma. *Sci Rep*. 2019;9(1):2615.
43. Lao J, Chen Y, Li ZC, et al. A deep learning-based radiomics model for prediction of survival in glioblastoma multi-forme. *Sci Rep*. 2017;7(1):1–8.
44. Luo H, Zhuang Q, Wang Y, et al. A novel image signature-based radiomics method to achieve precise diagnosis and prognostic stratification of gliomas. *Lab Invest*. 2020;101:1–13.
45. Wong KK, Rostomily R, Wong STC. Prognostic gene discovery in glioblastoma patients using deep learning. *Cancers (Basel)*. 2019;11(1):53.
46. Hao J, Kim Y, Kim TK, Kang M. PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data. *BMC Bioinform*. 2018;19(1):510.
47. Yang CH, Moi SH, Ou-Yang F, Chuang LY, Hou MF, Lin YD. Identifying risk stratification associated with a cancer for overall survival by deep learning-based CoxPH. *IEEE Access*. 2019;7:67708–17.
48. Yang CH, Moi SH, Hou MF, Chuang LY, Lin YD. Applications of deep learning and fuzzy systems to detect cancer mortality in next-generation genomic data. *IEEE Trans Fuzzy Syst*. 2020;99:1.
49. Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, Morozova O, Newton Y, Radenbaugh A, Pagnotta SM, et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell*. 2016;164(3):550–63.
50. Wang J, Cazzato E, Ladewig E, Frattini V, Rosenbloom DI, Zairis S, Abate F, Liu Z, Elliott O, Shin YJ, et al. Clonal evolution of glioblastoma under therapy. *Nat Genet*. 2016;48(7):768–76.
51. Wood MA, Paralkar M, Paralkar MP, et al. Population-level distribution and putative immunogenicity of cancer neoepitopes. *BMC Cancer*. 2018;18(1):414.
52. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, Trevino V, Shen H, Laird PW, Levine DA, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun*. 2013;4:2612.
53. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498–504.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

