Acta Biochim Biophys Sin 2026, 58(8): 1905–1909
https://cstr.cn/32197.14.abbs.2026037
https://doi.org/10.3724/abbs.2026037
New Phenomenon

New Phenomenon

Identification and experimental validation of core genes associated with breast cancer brain metastasis via machine learning

Zhaoda Duan, Chunjiao Yu, Wenjie Yang, Qiaoling Ruan, Rui Zhang, Yongfang Zhao*, and Shan Yan*

Yunnan Key Laboratory of Breast Cancer Precision Medicine, Academy of Biomedical Engineering, Kunming Medical University, Kunming 650000, China
*Correspondence address. E-mail: yanshan@kmmu.edu.cn (S.Y.); yongfang.zhao@outlook.com (Y.Z.)

Received 1 October 2025 Accepted 16 January 2026 Published 16 April 2026

Breast cancer (BC) is the most common malignancy among women, with approximately 2.3 million new cases diagnosed annually, accounting for approximately 11.6% of all cancer cases worldwide [1]. Distant metastasis is the primary cause of mortality in BC patients, with nearly 50% of patients ultimately developing metastatic disease. The predominant metastatic sites of BC include the lung, liver, brain, and bone, each exhibiting distinct biological characteristics that drive the organ-specific tropism of cancer cells [2,3]. Among these, brain metastasis represents a significant cause of mortality in BC patients and is particularly prevalent in those with human epidermal growth factor receptor 2 (HER2)-positive or triple-negative breast cancer (TNBC) subtypes [4]. Breast cancer brain metastasis (BCBM) can manifest in three forms: choroid plexus metastasis (rare), leptomeningeal metastasis (approximately 8%), and parenchymal metastasis, the most common presentation, with multiple lesions in 78% of cases and solitary lesions in 14%. Distinct anatomical regions of the brain provide different microenvironments, which in turn shape epidemiological patterns, biological behaviors, and therapeutic vulnerabilities of metastatic cancer [5]. With the continuous advancement of systemic therapies and imaging surveillance, brain metastases from BC have become increasingly prevalent, accounting for approximately 10%–30% of all metastatic breast cancer (MBC) cases. The continuous progression of BCBM often compromises patients' cognitive and sensory functions, leading to neurological impairment and severely limiting quality of life (QOL). Notably, the mortality rate within one year after diagnosis remains at 80% [6]. Current therapeutic strategies for BCBM primarily include surgery, whole-brain radiotherapy (WBRT), stereotactic radiosurgery (SRS), chemotherapy, or combinations thereof [7]. Although these approaches provide some clinical benefit, the efficacy remains limited due to the blood-brain barrier (BBB), which restricts drug penetration and contributes to chemoresistance [8]. Therefore, elucidating the molecular mechanisms underlying BCBM is imperative to identify novel diagnostic biomarkers and therapeutic targets, with the ultimate goal of improving treatment efficacy and patient prognosis.

Bioinformatics provides a powerful platform and data foundation for exploring the mechanisms of tumor initiation and progression. High-throughput platforms for gene expression analysis have gained significant popularity, with next-generation sequencing (NGS) and microarray analysis now widely applied as essential tools in medical oncology. These techniques have diverse clinical applications, including molecular cancer classification, prediction of therapeutic response, prognostic assessment, molecular diagnostics, and the discovery of novel drugs and therapeutic targets [9,10]. Weighted gene coexpression network analysis (WGCNA) has been widely applied in studies of gene regulatory networks, biomarker discovery, and elucidation of the molecular mechanisms underlying complex phenotypes. In this study, we utilized the BCBM microarray dataset GSE43837. We performed differential expression analysis and WGCNA clustering using the R packages limma and WGCNA to identify potential gene modules and candidate targets. GSE43837 consists of 19 nonmetastatic primary breast tumor samples and 19 breast cancer brain metastasis samples. Differential expression analysis, with thresholds set at $|\log FC| > 1$ and $P < 0.05$, identified 245 upregulated and 188 downregulated genes (Supplementary Table S1 and Supplementary Figure S1A). WGCNA further confirmed that the constructed network satisfied the scale-free topology criterion, with the optimal soft-threshold power determined to be 14 based on model fit and mean connectivity (Supplementary Figure S1B). Using the dynamic tree cut method, we clustered genes into multiple modules, each representing a group of coexpressed genes with varying degrees of correlation among modules (Supplementary Figure S1C,D). Notably, the midnightblue and black modules showed stronger correla-

© The Author(s) 2026. This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License (https://creativecommons.org/licenses/bync-nd/4.0/).

tions, and a significant positive relationship was observed between gene significance (GS) and module membership (MM) within these modules (Supplementary Figure S1E). This finding suggests that the core genes in these modules are highly representative and stable within the coexpression network. A total of 89 BCBM-related candidate genes were extracted from these key modules (Supplementary Table S2).

To further identify key feature genes associated with BCBM, we applied two machine learning methods, LASSO regression and random forest (RF), to the 29 overlapping genes obtained from the intersection of DEGs and hub module genes (Figure 1A and Supplementary Table S3). In the LASSO regression analysis, the optimal penalty parameter λ was determined by cross-validation, yielding a set of candidate genes with nonzero regression coefficients (Figure 1B). Concurrently, in the RF model, 500 decision trees were constructed, and the classification error rate gradually stabilized. Based on the Gini importance index, the top 10 most important genes were identified (Figure 1C). By intersecting the results of the two methods, four core genes were ultimately identified: *B3GNT9*, *SERPINF1*, *LUM*, and *CILP* (Figure 1D), which were subsequently selected for modeling and validation. We next evaluated the discriminatory power of these core genes in the GSE43837 dataset. ROC curve analysis demonstrated that *SERPINF1* achieved an area under the curve (AUC) of 0.914 (95% CI: 0.781–1.000, $P = 0.000274$), *LUM* an AUC of 0.914 (95% CI: 0.772–1.000, $P = 0.000274$), *CILP* an AUC of 0.914 (95% CI: 0.781–1.000, $P = 0.000274$), and *B3GNT9* an AUC of 0.879 (95% CI: 0.738–1.000, $P = 0.001150$), indicating strong predictive performance in distinguishing BC from BCBM samples (Figure 1E). Furthermore, violin plots revealed that B3GNT9 expression was significantly higher in BCBM samples than in primary BC tissues, whereas SERPINF1, LUM, and CILP were significantly downregulated in BCBM samples ($P < 0.01$; Figure 1F). Collectively, these findings suggest that SERPINF1, LUM, CILP, and B3GNT9 may play critical roles in the development of BCBM and hold potential as novel biomarkers.

Using the external validation dataset GSE125989, we further assessed the expression of the identified genes. GSE125989 consists of 16 primary breast cancer samples and 16 brain metastasis samples. *B3GNT9* was excluded due to the absence of a corresponding probe annotation; therefore, only *SERPINF1*, *LUM*, and *CILP* (Supplementary Materials) were validated. The results demonstrated that all three genes were significantly downregulated in the brain metastasis group compared with the breast cancer group ($P < 0.001$; Figure 1G). ROC curve analysis showed strong discriminatory power for each gene, with AUC values of 0.902 for *SERPINF1*, 0.947 for *LUM*, and 0.938 for *CILP* (Figure 1H). Notably, the logistic regression model combining the three genes achieved superior performance, with an AUC of 0.984 (Figure 1I), outperforming any single gene alone, thus underscoring the enhanced diagnostic value of their combined use in predicting brain metastasis. To further visualize individualized prediction, a nomogram was constructed based on the three-gene model (Figure 1J). The calibration curve showed a high degree of agreement between predicted probabilities and observed outcomes (Figure 1K). Decision curve analysis (DCA) demonstrated that the three-gene model provided consistently higher net clinical benefit across a wide range of threshold probabilities (Figure 1L). Collectively, these findings indicate that the three-gene model exhibits excellent diagnostic efficacy and has strong potential for clinical application in distinguishing breast cancer from brain metastasis.

To further validate the expression changes of the core genes during the process of BCBM, we employed both a zebrafish migration model and a mouse carotid artery injection model. Zebrafish provide a BBB structure highly comparable to mammals and allow real-time visualization of tumor cell extravasation, while the mouse model better mimics the mammalian brain microenvironment. In the zebrafish model, MDA-MB-231 BC cells (Pricella, Wuhan, China) were microinjected into the yolk sac of embryos, and tumor cells migrating to the brain region were sequentially isolated to obtain BM1 cells. After three and six rounds of iterative migration and selection, distinct brain metastatic subpopulations (BM3 and BM6) were established. qRT-PCR analysis revealed that, compared with the parental MDA-MB-231 cells (BM0), the expression levels of CILP, LUM, and SERPINF1 were significantly reduced in BM3 and BM6 cells ($P < 0.001$; Figure 2A,B). Consistently, western blot analysis further confirmed the marked downregulation of these three proteins in BM3 and BM6 cells relative to BM0 cells, demonstrating that the transcriptional alterations were reflected at the protein level (Figure 2C,D). In the mouse (Hunan SJA Laboratory Animal Co., Ltd, Changsha, China) brain metastasis model established via carotid artery injection of E0771 BC cells (Pricella), brain tissues were harvested on day 5 and day 14 post-injection to examine gene expression. The results showed that CILP, LUM, and SERPINF1 were significantly downregulated in brain tissues at both time points compared with the sham group ($P < 0.05$; Figure 2E,F). Western blot analysis of mouse brain tissues similarly demonstrated reduced protein expression of the three genes at both day 5 and day 14, further supporting their downregulation during brain metastatic colonization (Figure 2G,H). These findings suggest that the downregulation of these genes may be strongly associated with the enhanced brain metastatic potential of BC cells, supporting their role as key molecules in the development and progression of brain metastasis.

In summary, this study integrated transcriptomic data with machine learning approaches to identify potential key genes associated with BCBM, ultimately highlighting CILP, LUM, and SERPINF1 as core candidate molecules. These genes demonstrated robust diagnostic performance in both the training and external validation cohorts (AUC > 0.85) and were further validated in zebrafish serial selection and mouse carotid artery injection brain metastasis models, where they consistently exhibited a downregulated trend during metastatic progression. These findings suggest that CILP, LUM, and SERPINF1 may play vital roles in the development of BCBM and show potential as novel molecular biomarkers. Nevertheless, several limitations should be acknowledged. First, the sample sizes of the public datasets and the validation cohort were relatively small, which may introduce potential bias. Second, this study primarily focused on transcriptomic analyses and expression validation in animal models, and an in-depth investigation of the underlying molecular mechanisms was not performed. Future studies should incorporate more systematic *in vitro* and *in vivo* functional experiments to elucidate the mechanistic roles of these core genes in BCBM and strengthen the rationale for their potential as therapeutic targets.

Supplementary Data

Supplementary data are available at *Acta Biochimica et Biophysica Sinica* online.

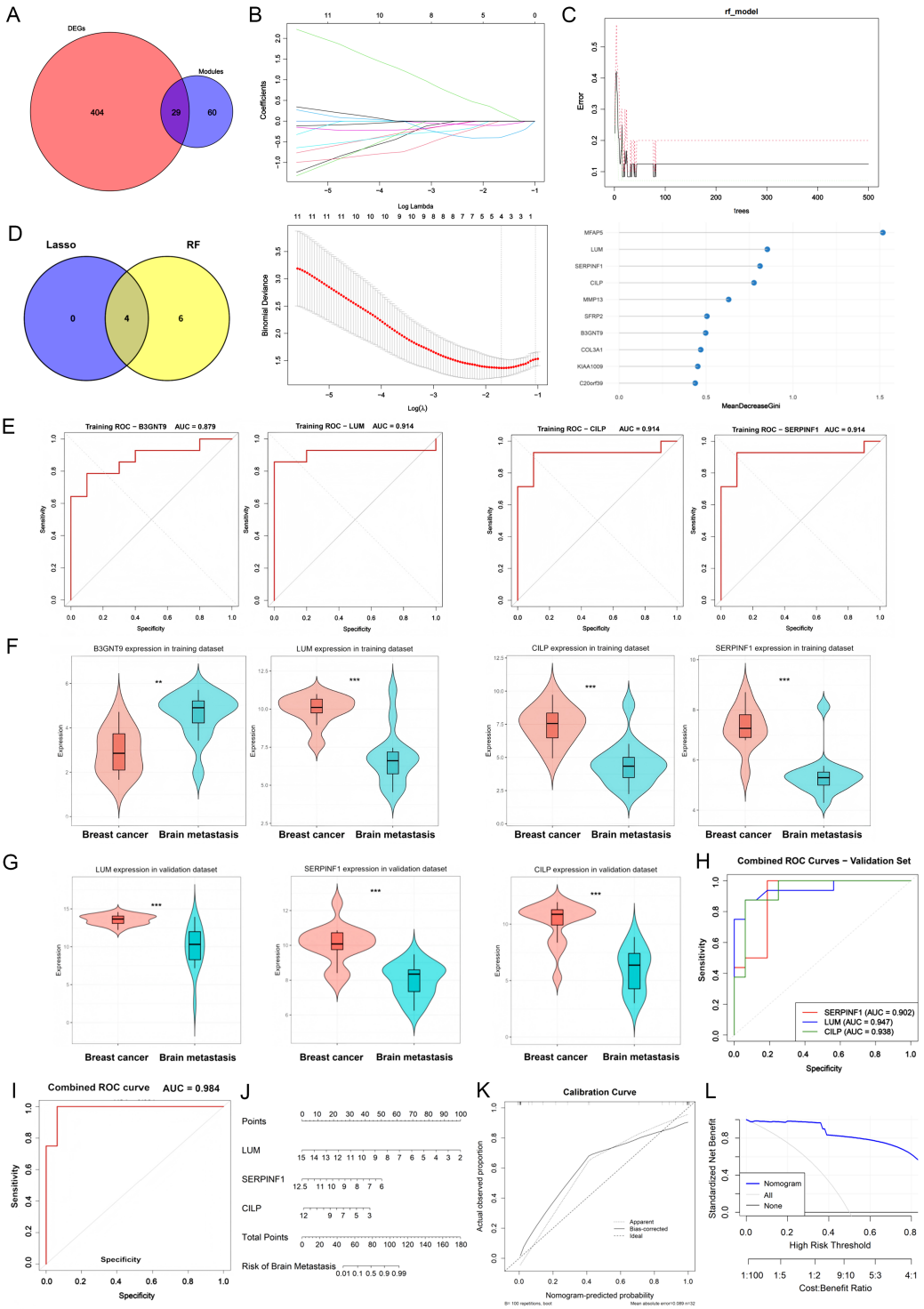


Figure 1. Identification and validation of core genes in breast cancer brain metastasis (A) Venn diagram showing the overlap between differentially expressed genes (DEGs) and key module genes identified by weighted gene coexpression network analysis WGCNA. (B) Distribution of LASSO coefficients for the overlapping genes and cross-validation plot used to determine the optimal tuning parameter log(λ) in the LASSO regression model. (C) Training error curve of the random forest (RF) model and the top 10 most important candidate genes ranked by the Gini index. (D) Venn diagram of genes selected by both LASSO and RF analyses. (E) Receiver operating characteristic (ROC) curves of the core genes (*B3GNT9, LUM, CILP,* and *SERPINF1*) in the training cohort. (F) Expression differences of the core genes between primary breast cancer and brain metastasis samples in the training set. (G) Expression differences of the core genes (*LUM, SERPINF1,* and *CILP*) between breast cancer and brain metastasis samples in the validation cohort. (H) Receiver operating characteristic (ROC) curves of the individual core genes in the validation cohort. (I) Combined ROC curve constructed based on the three core genes. (J) Nomogram model for predicting the risk of brain metastasis constructed from the three core genes. (K) Calibration curve of the nomogram model. (L) Decision curve analysis (DCA) of the three-gene model. $^{**}P < 0.01$, $^{***}P < 0.001$.

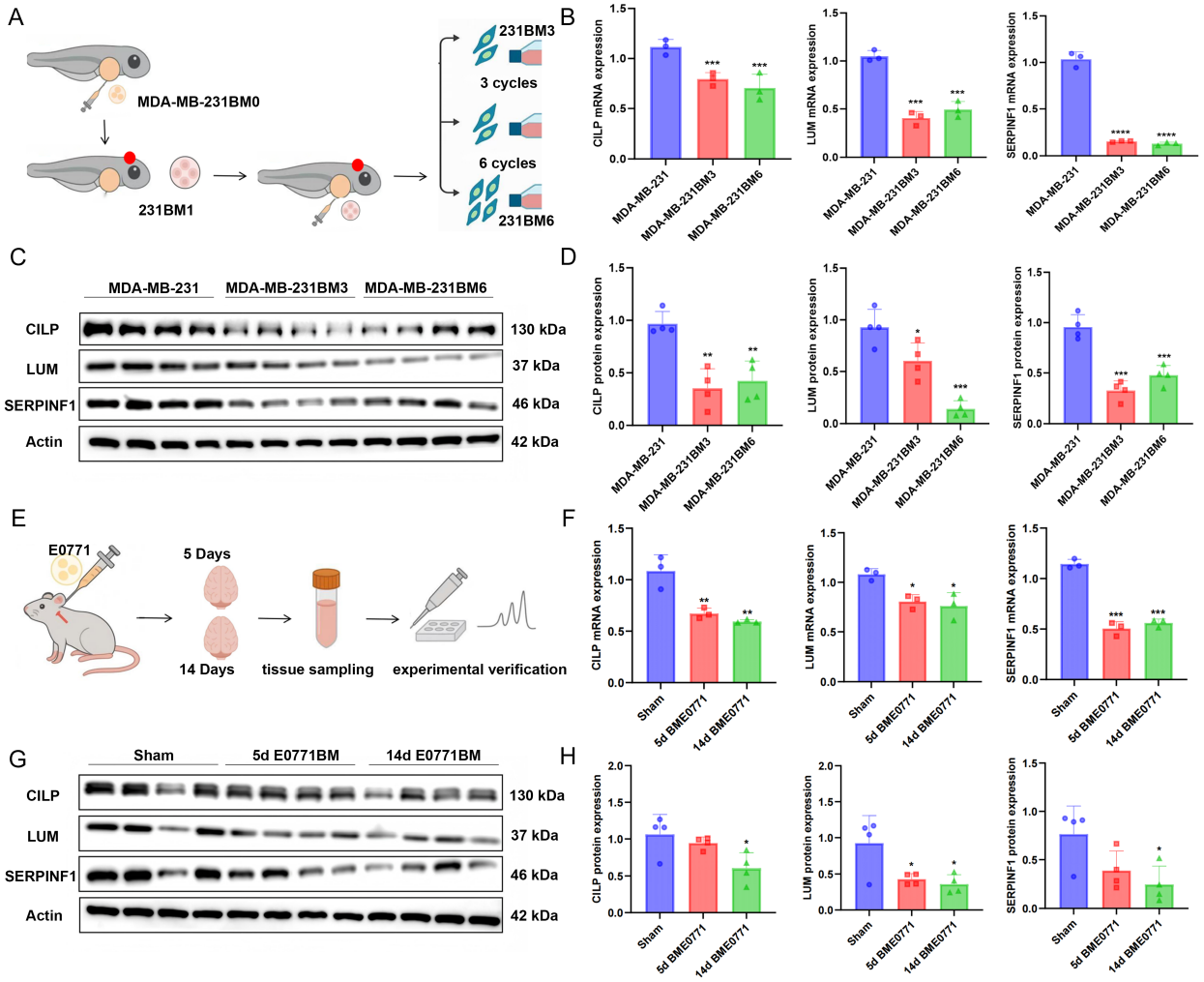


Figure 2. Validation of core gene expression in zebrafish and mouse brain metastasis models (A) Establishment of brain metastatic subpopulations from the MDA-MB-231 cell line using a zebrafish brain metastasis screening model, with stable metastatic subpopulations obtained after three and six rounds of iterative selection. (B) mRNA expression levels of core genes in the zebrafish brain metastasis model ($n = 3$). (C,D) Western blot analysis and quantitative densitometry of CILP, LUM, and SERPINF1 protein expression in BM0, BM3, and BM6 cells derived from the zebrafish screening model ($n = 4$). (E) Mouse brain metastasis model established by carotid artery injection of E0771 breast cancer cells, with brain tissues collected at days 5 and 14 post-injection. (F) mRNA expression levels of core genes in the mouse brain metastasis model ($n = 3$). (G,H) Western blot analysis and quantitative densitometry of CILP, LUM, and SERPINF1 protein expression in mouse brain tissues collected at days 5 and 14 post-injection ($n = 4$). $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

Data Availability

Data can be accessed upon reasonable request from the corresponding authors.

Funding

The work was supported by the grants from the National Natural Science Foundation of China (Nos. 82260537 and 82460478), the Association Foundation Program of Department of Yunnan Science and Technology and Kunming Medical University (No. 202501AY070001-006), and the Biomedical Projects of Yunnan Key Science and Technology Program (No. 202302AA310046).

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Nicolás-Morales ML, Luna-Pineda VM, Serrano-Bello CA, Guerrero-Macedonio MD, Rodríguez-Nava C, Parra-Rojas I, Espinoza-Rojo M, *et al.* The administration of passive and active immunotherapy against Syntenin-1 decreased the tumoral growth and pulmonary metastasis in a murine model of triple-negative breast cancer. *Sci Rep* 2025, 15: 32134
2. Ruan X, Yan W, Cao M, Daza RAM, Fong MY, Yang K, Wu J, *et al.* Breast cancer cell-secreted miR-199b-5p hijacks neurometabolic coupling to promote brain metastasis. *Nat Commun* 2024, 15: 4549
3. Rehman AU, Khan P, Maurya SK, Siddiqui JA, Santamaria-Barria JA, Batra SK, Nasser MW. Liquid biopsies to occult brain metastasis. *Mol Cancer* 2022, 21: 113
4. Wang Q, Wu H, Zhai Z, Fang D, Yang C, Liu L, Jia X, *et al.* SPANXB1 drives brain metastasis in breast cancer via MMP1 regulation: potential therapeutic insights with metformin. *Cell Death Discov* 2025, 11: 418
5. Arvanitis CD, Ferraro GB, Jain RK. The blood-brain barrier and blood-tumour barrier in brain tumours and metastases. *Nat Rev Cancer* 2020, 20: 26–41

6. Mills MN, Figura NB, Arrington JA, Yu HHM, Etame AB, Vogelbaum MA, Soliman H, *et al.* Management of brain metastases in breast cancer: a review of current practices and emerging treatments. *Breast Cancer Res Treat* 2020, 180: 279–300
7. Wang Y, Ye F, Liang Y, Yang Q. Breast cancer brain metastasis: insight into molecular mechanisms and therapeutic strategies. *Br J Cancer* 2021, 125: 1056–1067
8. Ren D, Cheng H, Wang X, Vishnoi M, Teh BS, Rostomily R, Chang J, *et al.* Emerging treatment strategies for breast cancer brain metastasis: from translational therapeutics to real-world experience. *Ther Adv Med Oncol* 2020, 12: 1758835920936151
9. Rezaei Z, Ranjbaran J, Safarpour H, Nomiri S, Salmani F, Chamani E, Larki P, *et al.* Identification of early diagnostic biomarkers via WGCNA in gastric cancer. *Biomed Pharmacother* 2022, 145: 112477
10. Wang YC, Tian ZB, Tang XQ. Bioinformatics screening of biomarkers related to liver cancer. *BMC Bioinformatics* 2021, 22: 521