A novel molecular subtype of hepatocellular carcinoma based on the tumor purity and tumor microenvironment-related polygenic risk scores

Yan Lin[1†], Rong Liang[1†], Xing Gao[1], Ziqin He[1], Lu Lu[1], Min Luo[1], Qian Li[1], Xiaobo Wang[2], Yongqiang Li[1], Guobin Wu[2*], Xiaoling Luo[3*], Jiazhou Ye[2*]

[1] Department of Medical Oncology, Guangxi Medical University Cancer Hospital, Nanning, Guangxi, People's Republic of China.

[2] Department of Hepatobiliary Surgery, Guangxi Medical University Cancer Hospital, Nanning, Guangxi, People's Republic of China.

[3] Department of Experimental Research, Guangxi Medical University Cancer Hospital, Nanning, Guangxi, People's Republic of China.

[†]These authors have contributed equally to this work and share first authorship

*** Correspondence:**
Jiazhou Ye
yejiazhou@gxmu.edu.cn
Xiaoling Luo
luoxiaoling@gxmu.edu.cn
Guobin Wu
wuguobin@gxmu.edu.cn

Keywords: Hepatocellular carcinoma, molecular classification, tumor microenvironment, immunotherapy, precision medicine.

Abstract

Purpose The purpose of the present study was to use malignant cell-related and tumor microenvironment (TME)-related molecules to develop a novel molecular subtype of hepatocellular carcinoma (HCC).

Methods The tumor purity (TP)-related and TME-related genes were identified and separately used to construct the TP-related and TME-related polygenic risk score (PRS). According to the two PRSs, we developed the TP-TME risk classification which was validated in two external data sets from The Cancer Genome Atlas Program and International Cancer Genome Consortium database. We also performed functional enrichment and drug repositioning analysis to reveal the potential biological heterogeneity among different subtypes.

Results The three TP-TME risk subtypes of HCC had significantly different prognosis and biological characteristics. The TP-TME low risk subtype had the best prognosis and was characterized by well-differentiated, the TP-TME high risk subtype had the worst prognosis and was characterized by aberrant activation of TGFβ and WNT pathways, and the TP-TME high risk subtype had the moderate prognosis and was characterized by exhibited activated MYC targets and proliferation-related gene sets. These three TP-TME risk subtypes may respond differently to immunotherapy (e.g., immune checkpoint inhibitors and chimeric antigen receptor-modified T cells) or other drug therapies.
Conclusion By combining the TP-related PRS and TME-related PRS, we proposed and validated the TP-TME risk subtyping system to divide patients with HCC into three subtypes with distinct biological characteristics and prognoses. These findings highlight the significant clinical implications of the TP-TME risk subtyping system and provide potential personalized immunotherapy strategies for HCC.

INTRODUCTION

Hepatocellular carcinoma (HCC) is the most common primary liver cancer and ranks as the sixth most common neoplasm and the third leading cause of cancer death.(1) The development of HCC is a complex multistep process that involves the accumulation of somatic genomic alterations in driver genes in addition to epigenetic modifications, which lead to its huge molecular heterogeneity.(2) The current systems for HCC staging or subtyping are mainly according to radiologic, serologic, and pathologic-based tumor burden evaluations.(3) However, a previous study(4) indicated that HCCs at the same stage have diverse molecular characteristics. Thus, it is imperative to propose more precise subtyping system for predicting prognosis and treatment effects. Given the advances in sequencing technology, several molecular subtyping systems were developed according to multi-omics features of HCC, nevertheless, the differences of the previous studies in technological platforms, preparation, and processing of samples make it difficult to explore a common method for typing HCC.(5) Thus, none of these molecular classifications so far are recommended to predict disease progression or prognosis.(6)

In addition, heterogeneous tumor microenvironment (TME) in HCC is also a crucial part of tumor heterogeneity. It is now increasingly accepted that tumor cells are not working alone but interact closely with the TME.(7) The heterogeneous TME affects tumor response to various treatments.(8) Targeting the TME was proposed as a strategy for removing obstruction to anticancer immune responses and immunotherapy.(8) Intriguingly, few molecular classifications of HCC have so far taken into account both the related molecules of malignant cells and TME. Thus, in the present study, we tried to develop a novel molecular classification for HCC

according to the expression patterns of malignant cell (tumor purity)- and TME-related genes. These unique risk factor patterns may provide a new frame to study cancer heterogeneity.

MATERIALS AND METHODS

Data Processing

We screened the HCCDB database (http://lifeome.net/database/hccdb/home.html)(9) to find the candidate data sets. The inclusion criteria were as follows: 1) the data set included both gene expression profiles and prognosis of patients with HCC, 2) the number of patients with a survival of more than 30 days should be more than 100, 3) and the gene expression profile of the dataset should contain more than 10,000 genes. According to the inclusion criteria, three data sets (GSE14520_GPL3921, TCGA-LIHC, and LIRI-JP) were selected and downloaded from HCCDB database for our analyses. Data set GSE14520_GPL3921(10) included the gene expression profiles based on the GPL3921 platform containing 225 HCC and 220 tumor-adjacent liver tissue samples were utilized to develop our subtyping systems. TCGA-LIHC data set containing RNA sequencing (RNA-seq) data and clinical information of 356 HCC belongs to The Cancer Genome Atlas Program (https://www.cancer.gov/tcga), and LIRI-JP data set containing RNA-seq data and clinical information of 212 HCC from JP Project from International Cancer Genome Consortium (https://dcc.icgc.org/) were used to validated the subtyping systems. The workflow of the present was shown in **Figure 1**.

Calculation of tumor purity and TME score and identification of differentially expressed genes (DEGs) in GSE14520_GPL3921

The gene expression profiles of GSE14520_GPL3921 were firstly utilized to calculate tumor purity (TP) using the ESTIMATE package.(11) Then, GSE14520_GPL3921 was also used to calculate the TME score using the xCell tools (https://xcell.ucsf.edu/)(12) with the xCell gene signature. The DEGs in HCC compared to tumor-adjacent liver tissue were identified using limma package.(13) Genes with fold changes > 1.5 and P (adjusted by false discovery rate) value < 0.05 were considered significant.

Normality test and correlation analysis

The tumor purity and TME score were separately performed Shapiro-Wilk test. Spearman or Pearson correlation analyses were performed to calculate the correlation between DEGs and TP and TME score. A DEG showed a positive correlation with TP and a negative correlation with TME score was considered a TP-related gene, while a DEG showed a negative correlation with TP and a positive correlation with TME

score was considered a TME-related gene. In addition, the TME-related genes do not include mark genes of TME cells in xCell signature.

Protein-protein interaction (PPI) networks

The PPI networks of TP-related and TME-related genes were obtained from STRING database (version 11.5)(14) to preliminarily reveal the crosstalks between the tumor cells and TME. The interactions with the high confidence (>0.7) were included in our present study and visualized using the Cytoscape software (version 3.8.0).(15)

Development of the TP- and TME-related gene-based polygenic risk scores

Firstly, for developing the TP-related polygenic risk score (PRS), the overall survival (OS)-associated TP-related genes were identified using univariate Cox regression analysis. Secondly, the expression profiles of the OS-associated TP-related genes were used to carry out the least absolute shrinkage and selection operator (LASSO) Cox regression model analysis with leave-one-out cross validation using glmnet package.(16) The genes with non-zero coefficient were considered the optimal features and subjected to multivariate Cox regression and stepwise regression analysis. Then the TP-related PRS was developed as the formula: TP-related PRS = Σ ($Expression_i$ * $Coeffient_i$). Where the "Coeffient" and "Expression" represent the risk coefficient and expression of each gene in the multivariate cox regression and stepwise regression analysis, respectively. The TME-related PRS was also developed according to the same method as above.

The TP-TME subtyping of HCC

The optimal cutoffs of the TP-related and TME-related PRS were identified using the surv_cutpoint function from survminer package (https://CRAN.R-project.org/package=survminer) to separately divide patients into high and low TP- and TME-related PRS groups. Each individual got a TP- and a TME-related PRS level, and we developed the TP-TME subtype according to the TP- and TME-related PRS levels. Patients possessing high TP- and TME-related PRS were considered as the high-risk subtype, those possessing low TP- and TME-related PRS were considered the low-risk subtype, and the remaining patients possessing a high TP-related and low TME-related PRS or a low TP-related and high TME-related PRS were considered the intermediate-risk subtype.

Gene set enrichment analysis (GSEA)

In order to preliminarily reveal the biological characteristics of these three risk subtypes, we performed GSEA(17, 18) in the GSE14520_GPL3921 data set using the GSEA java software (http://www.gsea-msigdb.org/gsea/index.jsp). Hallmark gene

sets and canonical pathway gene sets derived from the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database were download from The Molecular Signatures Database (MSigDB)(18-20) and used as the reference gene sets. The threshold was set to nominal P (NOM P) value < 0.05 or FDR q value < 0.25.

Analyses of gene mutation and stemness score

Gene mutation data of TCGA-LIHC data set were extracted from mutation annotation format (MAF) files using the *GDCquery_Maf* function in the "TCGAbiolinks" package.(21) Gene mutation frequencies of each risk subtype were visualized as a waterfall plot using the *oncoplot* function in the "TCGAbiolinks" package. The tumor mutational burden (TMB) of each sample was obtained from a previous study.(22) The stemness score(23) was calculated for each individual in the TCGA-LIHC data set using *TCGAanalyze_Stemness* function in the "TCGAbiolinks" package.

Prediction of efficacy of therapy

Immunotherapy has achieved tremendous successes in treatment of various cancers,(24) including HCC.(25) Treatment of immune checkpoint inhibitors (ICIs) and chimeric antigen receptor-modified T cell (CAR-T) were currently the two most widely studied immunotherapies. The expression of immune checkpoints was associated with the efficacy of ICIs.(26) The therapeutic effect of CAR-T cells is related to the expression of target genes in the tumor cells. For the TP-TME risk subtypes, we compared the expression levels of two immune checkpoints (PDL1 and CTLA4) and five antigens (CD133, EPCAM, GCP3, MSLN, and MUC1)(27) to predict the potential response to these treatments. In addition, we performed drug repositioning analysis for the high-risk subtype using the PATHOME-Drug (http://statgen.snu.ac.kr/software/pathome/) web tools.

Statistical analysis

In our present study, unless otherwise stated, all these analyses were performed in R (version 4.0.2). We identified DEGs using unpaired t-tests provided by limma package. Shapiro-Wilk test was used for the nrmality test. Time-dependent receiver operating characteristic curve (tROC) analysis was carried out using the tROC package.(28) Kaplan–Meier survival curves for overall survival (OS) and progression free survival (PFS) were compared in different subtypes using the log-rank method in the "survival" package (https://CRAN.R-project.org/package=survival) and "survminer" package (https://CRAN.R-project.org/package=survminer). Intergroup differences in continuous variables were assessed for significance using Wilcoxon, Kruskal–Wallis, or unpaired t-tests. All tests were two-sided and unless otherwise stated, we set P value < 0.05 to be statistically significant.

RESULTS

The biological functions and interactions of tumor purity (TP)-related and TME-related genes

We identified a total of 2263 DEGs in HCC compared to tumor-adjacent liver tissue (**Figure 2A**). Both the TP and TME score showed non-normal distribution with Shapiro-Wilk $P < 0.001$. A total of 451 TP-related and 121 TME-related genes were identified by Spearman correlation analysis, and the bidirectional hierarchical clustering showed the expression patterns of them could basically distinguish HCC and tumor-adjacent liver tissue (**Figure 2B**). Unsurprisingly, the TP-related genes are mainly involved in cancer-related gene ontology terms or pathways **(Figure 2C**), such as cell cycle and mismatch repair; while the TME-related genes are mainly involved in immune system (**Figure 2D**). The PPI networks of the TP-related and TME-related genes contain 342 nodes and 1177 edges (**Figure 2E**). Red indicates upregulated and blue indicates downregulated in the metastasis group, while circular nodes represent the TP-related genes and rhombic nodes represent the TME-related genes.

The TP-TME risk subtyping is a robust prognosis prediction system

Fifty TP-related genes were identified as OS-associated genes, twenty-two of them possess non-zero coefficient (**Figure 3A**), and eleven genes (*ALG6*, *ATP5MF*, *CNIH4*, *ESM1*, *HEY1*, *LANCL1*, *P2RX4*, *PEX11B*, *POP7*, *RCN2*, and *XPO1*) were used to generate the TP-related polygenic risk score (PRS) (**Supplementary Table 1**). The TP-related PRS was significantly associated with overall survival (OS) with $P < 0.0001$, Hazard Ratio (HR) = 2.718 (95%CI for HR = 2.147-3.442) and the area under curve (AUC) of tROC analysis was stably around 0.8 (**Figure 3B**). The HCC patients with high TP-related PRS showed shorter OS than those with low TP-related PRS ($P < 0.0001$) (**Figure 3C**). In the TME-related genes, twelve genes were identified as OS-associated genes, ten of them possess non-zero coefficient (**Figure 3D**), and seven genes (*ALDH1B1*, *CTSC*, *GUCY1A1*, *MRC1*, *SPRY2*, *TARP*, and *TRIM22*) were used to generate the TME-related PRS (**Supplementary Table 2**). The TME-related PRS was also significantly associated with OS with $P < 0.0001$, Hazard Ratio (HR) = 2.718 (95%CI for HR = 1.978-3.735) and the AUC of tROC analysis was 0.7-0.8 (**Figure 3E**). The HCC patients with high TME-related PRS showed shorter OS than those with low TP-related PRS ($P < 0.0001$) (**Figure 3F**). Our TP-TME risk subtype was generated based the two PRSs, and 34, 52, and 123 patients with HCC were divided into high-, intermediate-, and low-risk subtypes, respectively. Cases defined as high-risk subtype had the best poor survival while those in low-risk subtype had the best survival, and the intermediate-risk cases had a better prognosis than high-risk subtype and worse than low-risk subtype (**Figure 3G**). A similar trend was also observed in progression free survival (PFS) (**Figure 3H**). Furthermore, the TP-TME

risk subtyping system showed independent to some routine clinicopathological features (**Figure 3I**). As it was in the GSE14520_GPL3921 data set, the TP-related PRS and the TME-related PRS in the TCGA-LIHC and LIRI-JP data sets were calculated according to the abovementioned formula, respectively. We found similar results in the two validation sets. Briefly, the TP-related PRS (**Figure 4A** for TCGA-LIHC and **Figure 5A** for LIRI-JP) and the TME-related PRS (**Figure 4B** for TCGA-LIHC and Figure 5B for LIRI-JP) were significantly associated with OS, the TP-TME risk subtyping system was associated with OS (**Figure 4C** for TCGA-LIHC and **Figure 5C f**or LIRI-JP) and PFS (**Figure 4D** for TCGA-LIHC). The prognostic value of the TP-TME risk subtyping system was independent to the routine clinicopathological features (**Figure 4E** for TCGA-LIHC and **Figure 5D** for LIRI-JP). Collectively, the TP-TME risk subtyping is a robust prognosis prediction system.

The subtype-specific curated gene sets of the TP-TME risk subtypes

Compared to the TP-TME intermediate- and high-risk subtypes, the liver function-related hallmark (**Figure 6A**) and metabolism-related Kyoto Encyclopedia of Genes and Genomes (KEGG) (**Figure 6D**) gene sets were significantly enriched in the TP-TME low-risk subtype. This suggests the TP-TME low-risk subtype HCC were well-differentiated. The TP-TME intermediate-risk subtype was charactered by enriching transcription factor *E2F* and *MYC* targets (**Figure 6B**) and cell cycle pathways (**Figure 6E**), while the TP-TME high-risk subtype was charactered by enriching hypoxia and Wntβ catenin signaling (**Figure 6C**), and Notch signaling pathway and TGFβ signaling pathway (**Figure 6F**). These results indicated that there is significant biological heterogeneity between these three subtypes.

These three TP-TME risk subtypes may respond differently to immunotherapy

The **Figures 7A-C** show the top 30 mutation genes in the low-, intermediate-, and high-risk subtypes, respectively. However, few known drugs so far targeted these genes for HCC. The tumor mutational burden (TMB) in HCC was low, and no significantly different among these three subtypes (**Figure 7D**). This suggests TMB may not be an efficient biomarker for selecting patients with HCC for ICI treatment. In addition, our analysis also found no significant difference in the stemness score among these three risk subtypes (**Figure 7E**). Though the CD274 (also known as PDL1) expression of the three risk subtypes show no significantly differentially (**Figure 7F**), the expressions of CTLA4 (**Figure 7G**) and PDCD1 (also known as PD1) (**Figure 7H**) were higher in the intermediate-risk subtypes than the low-risk subtypes. Thus, the intermediate-risk subtype may possess a higher response rate to treatment with ICIs more than the low-risk subtype. In addition, the three TP-TME risk subtypes have distinct expression levels of the five cancer antigens which were used as targets in the chimeric antigen receptor-modified T cell (CAR-T) therapy. The

GPC3 express higher in the intermediate-risk subtype than the low-risk subtype (**Figure 7I**), the MSLN express higher in the intermediate- and high-risk subtypes than the low-risk subtypes (**Figure 7J**). Impressively, the high-risk subtypes have highest expression of MUC1 (**Figure 7K**), EPCAM (**Figure 7L**), and PROM1 (**Figure 7M**). Thus, theoretically, the three TP-TME risk subtypes may respond differently to corresponding CAR-T therapies.

Potentially effective drugs for the high-risk subtypes

Many agents which tried to treat HCC evaluated in phase 3 trials got the disappointing results,(1) the response is usually observed in a small subset of individuals. Through the PATHOME-Drug analysis, we constructed the drug-target networks to identify the potentially effective drugs for the high-risk subtypes (**Figure 8**). The potentially effective drugs include recommended agents such as sorafenib, regorafenib, and cabozantinib, and some drugs which are used to treat other diseases (**Supplementary Table 3**).

DISCUSSION

The heterogeneity of HCC is attributed to the presence of various etiologies, such as infections of virus or parasitic, chemical carcinogens, cigarette smoking, excess alcohol intake, and dietary factors.(1) Accordingly, the host's response to various pathogenic factors leads to the formation of their own unique microenvironment. One hypothesis is that different TMEs and pathogenic factors may evoke distinct molecular alteration to independently initiate HCC progress, which results in extensive inter-tumor molecular heterogeneity. One of the essential efforts for improving the poor outcome of HCC is to provide a subtyping system that is capable of accurately defining tumor risk subtypes, each displaying unique molecular characteristics linked to potentially druggable driver genes in order to provide personalized treatment choices based on the subtyping system. Though many efforts, mainly focusing on the malignant cells, were paid to elaborate the inter-tumor heterogeneity and propose various single- or multi-omics-based molecular typing systems,(5, 29) their effectiveness remain limited for providing precision treatment. In this scope, given that the crucial role of TME in cancers is confirmed,(30) TME-related molecules should be contributed to the subtyping for HCC. Another challenge of previous molecular typing methods is cost effectiveness, due to they need hundreds of genes or even multiple omics data types.

In our present study, we firstly identified the related genes of TP and TME and subsequently generated a TP-related PRS and a TME-related PRS according to the expression patterns of these types of genes, and furtherly proposed a novel risk subtyping that could successfully divide patients with HCC into three risk subtypes.

Similar to other molecular typing systems,(31-34) our subtypes have distinct prognosis and were validated in two independent external data sets. As reviewed by Wu et al.(5) a few subtypes were repeatedly uncovered in different studies, indicating that different HCC subtypes derived from different omics technologies may share common molecular characteristics. Our TP-TME low-risk subtype may be well-differentiated and enrich gene sets related to liver function (e.g., bile acid metabolism), similar to the Hoshida's S3 subtype(35) and TCGA's ICluster2. The TP-TME intermediated-risk subtype exhibited activated MYC targets and proliferation-related gene sets (e.g., cell cycle and G2M checkpoint), corresponding to the Hoshida's S2 subtype and Chaisaingmongkol et al.'s C1 subtype.(36) The TP-TME high-risk subtype characterized by aberrant activation of TGFβ and WNT pathways, displayed multiple similarities with Hoshida's S1 subtype and Kurebayashi et al.'s immune-high subtype.(37) Compared with these typing methods, low cost is the potential advantage of our TP-TME risk subtyping system. In addition, though the further study was required, we also proposed the potential immunotherapy and drugs for the high-risk subtype, which may help for decision-making in clinical practice.

Unsurprisingly, some of these the candidate eleven TP-related and seven TME-related genes have been found to be associated with HCC or other types of cancers in previous studies. *ESM1* was found as a biomarker of macrotrabecular-massive HCC.(38) *HEY1* plays a critical role in hypoxia-related regulation of mitochondrial activity in HCC.(39) *RCN2* can enhance HCC proliferation via modulating the EGFR-ERK pathway.(40) The interactions between *CTSC* and TNF-α/p38 MAPK signaling pathway are associated with proliferation and metastasis in HCC.(41) LANCL1 was reported that can protect prostate cancer cells from oxidative stress.(42) XPO1-dependent nuclear export was proposed as a target for cancer therapy.(43) However, the associations of some TP-related or TME-related genes and HCC have been not noted in the previous study. According to our current analysis, we provide potential candidate molecules for further research on HCC.

Although our current study provided a novel molecular classification system, it has some noted limitations. Firstly, the roles of some candidate genes in HCC remain elusive, it is not clear whether these genes are causal or merely prognostic markers for HCC. Secondly, the TP-TME risk subtyping system was generated from a retrospective analysis, it should be validated or improved by the prospective trials before being used in clinical practice.

CONCLUSIONS

In conclusion, by combining separately constructed TP-related PRS and TME-related PRS, we proposed and validated a novel molecular classification system, the TP-TME

risk subtyping system, to divide patients into three subtypes with distinct biological characteristics and prognosis. These findings highlight the significant clinical implications of the TP-TME risk subtyping system and provide potential personalized immunotherapy strategies for patients with HCC.

ACKNOWLEDGMENTS

Not available

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

This research was supported by the National Natural Science Foundation of China (NO. 81803007, 82060427, 82103297), Guangxi Key Research and Development Plan (NO. GUIKEAB19245002), Guangxi Scholarship Fund of Guangxi Education Department, Guangxi Natural Science Foundation (NO. 2020GXNSFAA259080), Guangxi Medical University Training Program for Distinguished Young Scholars, Science and Technology Plan Project of Qingxiu District, Nanning (NO. 2020037, 2020038).

AUTHORS CONTRIBUTIONS

YJZ, LXL, and WGB designed the study and revised the manuscript. LY and LR performed the analyses and wrote the manuscript. GX, HZQ, LL, LM, LQ, WXB, and LYQ assisted with analyzing the data and writing the manuscript. All authors read and approved the final manuscript.

DATA AVAILABILITY

The datasets used during the current study are available in the HCCDB repository, [http://lifeome.net/database/hccdb/home.html]

CODE AVAILABILITY

The code used during the current study is available from the corresponding author upon request.

ETHICS APPROVAL

Not available.

CONSENT TO PARTICIPATE

Informed consent was obtained from all individual participants included in the study.

CONSENT TO PUBLISH

Not available.

REFERENCES

1. Forner A, Reig M, Bruix J. Hepatocellular carcinoma. Lancet. 2018;391(10127):1301-14.

2. Schulze K, Nault JC, Villanueva A. Genetic profiling of hepatocellular carcinoma using next-generation sequencing. J Hepatol. 2016;65(5):1031-42.

3. Minagawa M, Ikai I, Matsuyama Y, Yamaoka Y, Makuuchi M. Staging of hepatocellular carcinoma: assessment of the Japanese TNM and AJCC/UICC TNM systems in a cohort of 13,772 patients in Japan. Ann Surg. 2007;245(6):909-22.

4. Nault JC, Martin Y, Caruso S, Hirsch TZ, Bayard Q, Calderaro J, et al. Clinical Impact of Genomic Diversity From Early to Advanced Hepatocellular Carcinoma. Hepatology. 2020;71(1):164-82.

5. Wu Y, Liu Z, Xu X. Molecular subtyping of hepatocellular carcinoma: A step toward precision medicine. Cancer Commun (Lond). 2020;40(12):681-93.

6. Pinyol R, Montal R, Bassaganyas L, Sia D, Takayama T, Chau GY, et al. Molecular predictors of prevention of recurrence in HCC with sorafenib as adjuvant treatment and prognostic factors in the phase 3 STORM trial. Gut. 2019;68(6):1065-75.

7. Hanahan D, Coussens LM. Accessories to the crime: functions of cells recruited to the tumor microenvironment. Cancer Cell. 2012;21(3):309-22.

8. Pitt JM, Marabelle A, Eggermont A, Soria JC, Kroemer G, Zitvogel L. Targeting the tumor microenvironment: removing obstruction to anticancer immune responses and immunotherapy. Ann Oncol. 2016;27(8):1482-92.

9. Lian Q, Wang S, Zhang G, Wang D, Luo G, Tang J, et al. HCCDB: A Database of Hepatocellular Carcinoma Expression Atlas. Genomics Proteomics Bioinformatics. 2018;16(4):269-75.

10. Roessler S, Jia HL, Budhu A, Forgues M, Ye QH, Lee JS, et al. A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. Cancer Res. 2010;70(24):10202-12.

11. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612.

12. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18(1):220.

13. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.

14. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1):D607-D13.

15. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498-504.

16. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. 2010. 2010;33(1):22.

17. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003;34(3):267-73.

18. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, 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.

19. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739-40.

20. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417-25.

21. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016;44(8):e71.

22. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, et al. The Immune Landscape of Cancer. Immunity. 2018;48(4):812-30 e14.

23. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. Cell. 2018;173(2):338-54 e15.

24. Kruger S, Ilmer M, Kobold S, Cadilha BL, Endres S, Ormanns S, et al. Advances in cancer immunotherapy 2019 - latest trends. J Exp Clin Cancer Res. 2019;38(1):268.

25. Kole C, Charalampakis N, Tsakatikas S, Vailas M, Moris D, Gkotsis E, et al. Immunotherapy for Hepatocellular Carcinoma: A 2021 Update. Cancers (Basel). 2020;12(10).

26. Paver EC, Cooper WA, Colebatch AJ, Ferguson PM, Hill SK, Lum T, et al. Programmed death ligand-1 (PD-L1) as a predictive marker for immunotherapy in solid tumours: a guide to immunohistochemistry implementation and interpretation. Pathology. 2021;53(2):141-56.

27. Ma S, Li X, Wang X, Cheng L, Li Z, Zhang C, et al. Current Progress in CAR-T Cell Therapy for Solid Tumors. Int J Biol Sci. 2019;15(12):2548-60.

28. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med. 2013;32(30):5381-97.

29. Lu LC, Hsu CH, Hsu C, Cheng AL. Tumor Heterogeneity in Hepatocellular Carcinoma: Facing the Challenges. Liver Cancer. 2016;5(2):128-38.

30. Hinshaw DC, Shevde LA. The Tumor Microenvironment Innately Modulates Cancer Progression. Cancer Res. 2019;79(18):4557-66.

31. Cancer Genome Atlas Research Network. Electronic address wbe, Cancer Genome Atlas Research N. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell. 2017;169(7):1327-41 e23.

447 32. Fujimoto A, Furuta M, Totoki Y, Tsunoda T, Kato M, Shiraishi Y, et al.
448 Whole-genome mutational landscape and characterization of noncoding and
449 structural mutations in liver cancer. Nat Genet. 2016;48(5):500-9.

450 33. Lee JS, Chu IS, Heo J, Calvisi DF, Sun Z, Roskams T, et al. Classification and
451 prediction of survival in hepatocellular carcinoma by gene expression profiling.
452 Hepatology. 2004;40(3):667-76.

453 34. Jiang Y, Sun A, Zhao Y, Ying W, Sun H, Yang X, et al. Proteomics identifies new
454 therapeutic targets of early-stage hepatocellular carcinoma. Nature.
455 2019;567(7747):257-61.

456 35. Hoshida Y, Nijman SM, Kobayashi M, Chan JA, Brunet JP, Chiang DY, et al.
457 Integrative transcriptome analysis reveals common molecular subclasses of
458 human hepatocellular carcinoma. Cancer Res. 2009;69(18):7385-92.

459 36. Chaisaingmongkol J, Budhu A, Dang H, Rabibhadana S, Pupacdi B, Kwon SM, et
460 al. Common Molecular Subtypes Among Asian Hepatocellular Carcinoma and
461 Cholangiocarcinoma. Cancer Cell. 2017;32(1):57-70 e3.

462 37. Kurebayashi Y, Ojima H, Tsujikawa H, Kubota N, Maehara J, Abe Y, et al.
463 Landscape of immune microenvironment in hepatocellular carcinoma and its
464 additional impact on histological and molecular classification. Hepatology.
465 2018;68(3):1025-41.

466 38. Calderaro J, Meunier L, Nguyen CT, Boubaya M, Caruso S, Luciani A, et al.
467 ESM1 as a Marker of Macrotrabecular-Massive Hepatocellular Carcinoma. Clin
468 Cancer Res. 2019;25(19):5859-65.

469 39. Kung-Chun Chiu D, Pui-Wah Tse A, Law CT, Ming-Jing Xu I, Lee D, Chen M, et
470 al. Hypoxia regulates the mitochondrial activity of hepatocellular carcinoma cells
471 through HIF/HEY1/PINK1 pathway. Cell Death Dis. 2019;10(12):934.

472 40. Ding D, Huang H, Jiang W, Yu W, Zhu H, Liu J, et al. Reticulocalbin-2 enhances
473 hepatocellular carcinoma proliferation via modulating the EGFR-ERK pathway.
474 Oncogene. 2017;36(48):6691-700.

475 41. Zhang GP, Yue X, Li SQ. Cathepsin C Interacts with TNF-alpha/p38 MAPK
476 Signaling Pathway to Promote Proliferation and Metastasis in Hepatocellular
477 Carcinoma. Cancer Res Treat. 2020;52(1):10-23.

42. Wang J, Xiao Q, Chen X, Tong S, Sun J, Lv R, et al. LanCL1 protects prostate cancer cells from oxidative stress via suppression of JNK pathway. Cell Death Dis. 2018;9(2):197.

43. Azizian NG, Li Y. XPO1-dependent nuclear export as a target for cancer therapy. J Hematol Oncol. 2020;13(1):61.

Figure legends

Figure 1 The workflow of the present study.
Abbreviation: LASSO, least absolute shrinkage and selection operator; TP, tumor purity; TME, tumor microenvironment; TCGA-LIHC, The Cancer Genome Atlas-liver hepatocellular carcinoma; ICGC-LIRI-JP: Liver Cancer-RIKEN, JP.

Figure 2 The identification, enrichment analysis, and protein-protein interaction networks of the tumor purity-related genes and tumor microenvironment-related genes. (A) The volcano plot of the differentially expressed gene analysis; **(B)** Hierarchical clustering showed the expression patterns of tumor purity (TP)-related and tumor microenvironment (TME)-related genes basically distinguish hepatocellular carcinoma (HCC) and tumor-adjacent liver tissues. **(C)** Functional enrichment analysis for the TP-related genes. *Left panel:* GO terms and pathways involving TP-related genes, and *right panel*: interactions among the GO terms and pathways. **(D)** Functional enrichment analysis for the TME-related genes. *Left panel:* GO terms and pathways involving TME-related genes, and *right panel*: interactions among the GO terms and pathways. **(E)** protein-protein interaction networks of the TP-related genes and TME-related genes. The red represents up-regulated, and the blue represents down-regulated in hepatocellular carcinoma. The circle nodes were TP-related genes, and the diamond nodes were TME-related genes.
Abbreviation: TME, tumor microenvironment; TP, tumor purity; HCC, hepatocellular carcinoma; GO, gene ontology.

Figure 3 The development of TP-TME risk subtypes in GSE14520_GPL3921 data set. (A) twenty-two TP-related genes had non-zero coefficients in the LASSO Cox regression model analysis; **(B)** Time-dependent ROC curve analysis for the TP-related PRS; **(C)** HCC with high TP-related PRS had shorter overall survival than those with low TP-related PRS. **(D)** Ten TME-related genes had non-zero coefficients in the LASSO Cox regression model analysis; **(E)** Time-dependent ROC curve analysis for the TME-related polygenic risk score; **(F)** HCC with high TME-related PRS had shorter overall survival than those with low TME-related PRS. **(G)** There are

significantly different overall survivals between the three subtypes in the TP-TME risk subtypes. **(H)** There are significantly different progression free survivals between the three subtypes in the TP-TME risk subtypes. **(I)** The TP-TME risk subtype system was proved to be an independent prognostic factor, after adjusting for other clinicopathological characteristics.
Abbreviation: LASSO, least absolute shrinkage and selection operator; TP, tumor purity; TME, tumor microenvironment; HCC, hepatocellular carcinoma; ROC, receiver operating characteristic; PRS, polygenic risk score.

Figure 4 The validation of TP-TME risk subtypes in TCGA-LIHC data set. (A) HCC with high TP-related PRS had shorter overall survival than those with low TP-related PRS. **(B)** HCC with high TME-related PRS had shorter overall survival than those with low TME-related PRS. **(C)** There are significantly different overall survivals between the three subtypes in the TP-TME risk subtypes. **(D)** There are significantly different progression free survivals between the three subtypes in the TP-TME risk subtypes. **(E)** The TP-TME risk subtype system was proved to be an independent prognostic factor, after adjusting for other clinicopathological characteristics.
Abbreviation: TP, tumor purity; TME, tumor microenvironment; PRS, polygenic risk score; AFP, alpha fetoprotein.

Figure 5 The validation of TP-TME risk subtypes in LIRI-JP data set. (A) HCC with high TP-related PRS had shorter overall survival than those with low TP-related PRS. **(B)** HCC with high TME-related PRS had shorter overall survival than those with low TME-related PRS. **(C)** There are significantly different overall survivals between the three subtypes in the TP-TME risk subtypes. **(D)** The TP-TME risk subtype system was proved to be an independent prognostic factor, after adjusting for other clinicopathological characteristics.
Abbreviation: TP, tumor purity; TME, tumor microenvironment; PRS, polygenic risk score; NBNC, no hepatitis B virus and no hepatitis C virus; HBV, hepatitis B virus; HCV, hepatitis C virus; AJCC, American Joint Committee on Cancer.

Figure 6 Gene set enrichment analysis for identifying the subtype-specific curated gene sets of the TP-TME risk subtypes. The hallmark gene sets enriched in the **(A)** TP-TME low-risk subtype, **(B)** TP-TME intermediate-risk subtype, and **(C)** TP-TME high-risk subtype. The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway gene sets enriched in the **(D)** TP-TME low-risk subtype, **(E)** TP-TME intermediate-risk subtype, and (F) TP-TME high-risk subtype.

Figure 7 Mutation, stemness, and immunotherapeutic efficacy analysis. Top 30 mutant genes in the (A) TP-TME low-risk subtype, **(B)** TP-TME intermediate-risk

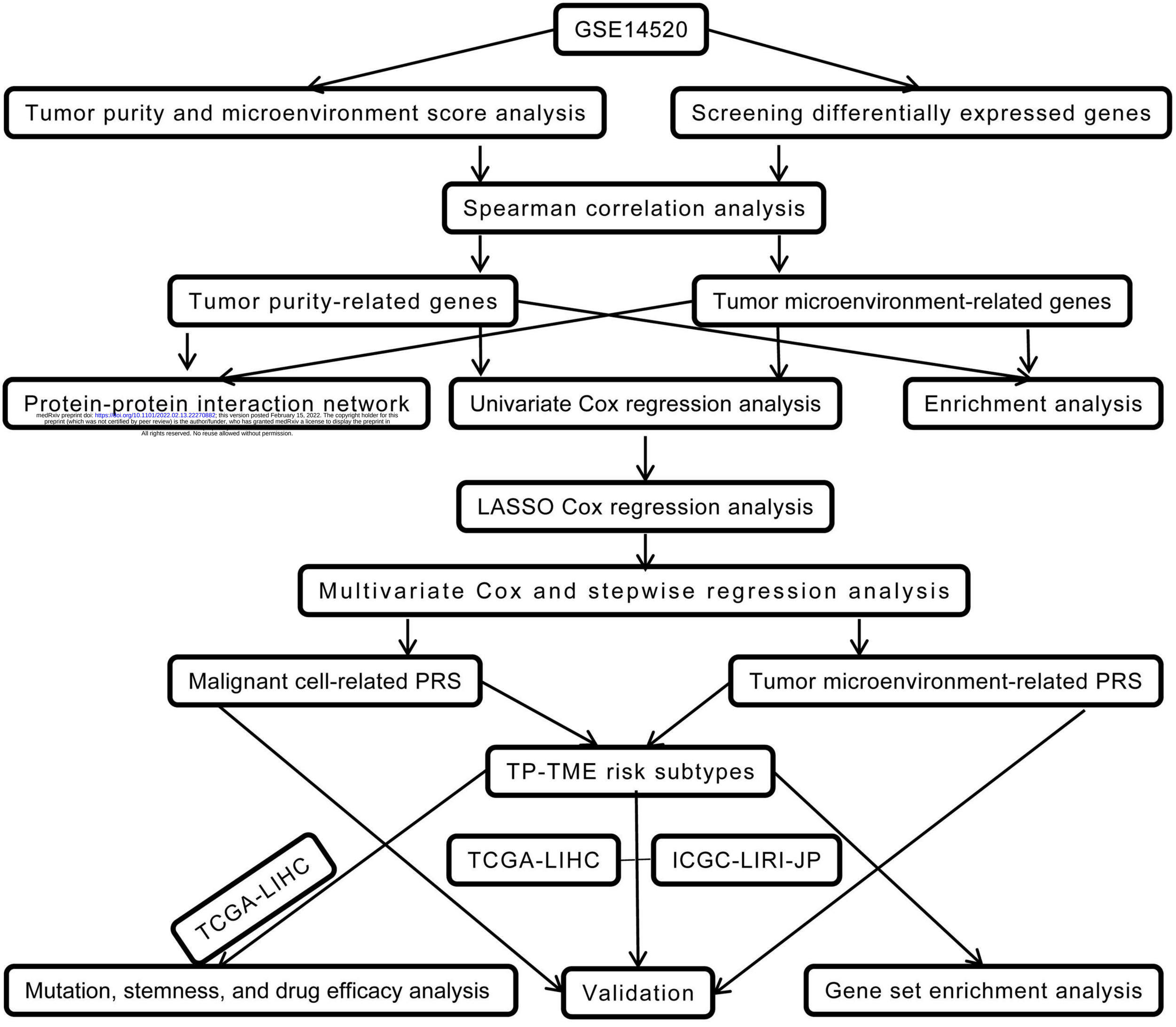
subtype, **(C)** and TP-TME high-risk subtype. **(D)** Tumor mutation burden in the three TP-TME high-risk subtypes. **(E)** Stemness score in the three TP-TME high-risk subtypes. The expression of (F) CD274, (G) CTLA4, **(H)** PDCD1, (I) GPC3, **(J)** MSLN, (K) MUC1, **(L)** EPCAM, and **(M)** PROM1 in the three TP-TME high-risk subtypes.

Figure 8 Drug-target networks for potentially effective drugs for the TP-TME high-risk subtypes. The red represents up-regulated, and the blue represents down-regulated in the TP-TME high-risk subtypes.

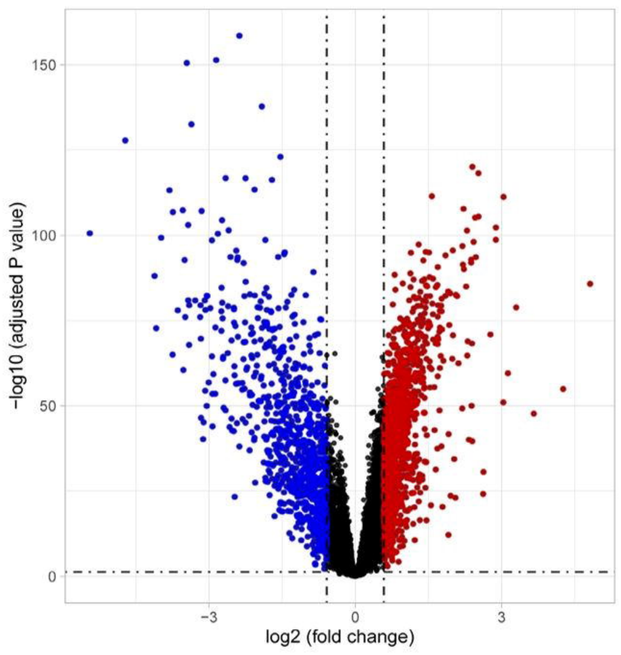
Supplementary Table 1 The overall survival-associated tumor purity-related genes in hepatocellular carcinoma

Supplementary Table 2 The overall survival-associated tumor microenvironment-related genes in hepatocellular carcinoma

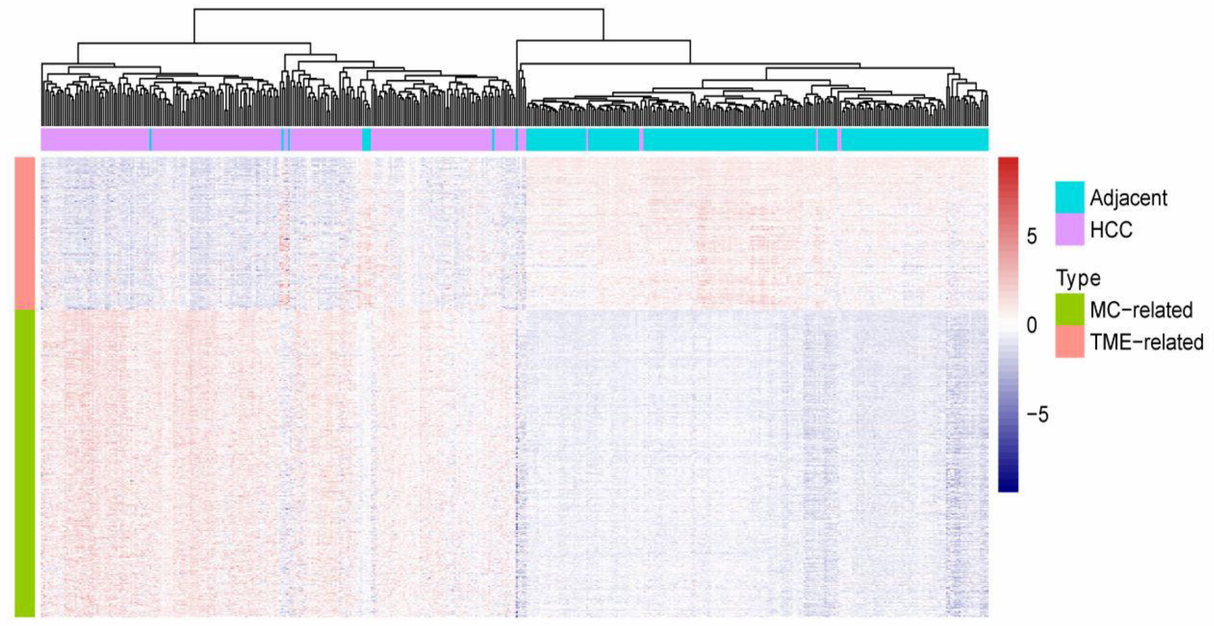
Supplementary Table 3 Potentially effective drugs and relevant targets for the high-risk subtypes



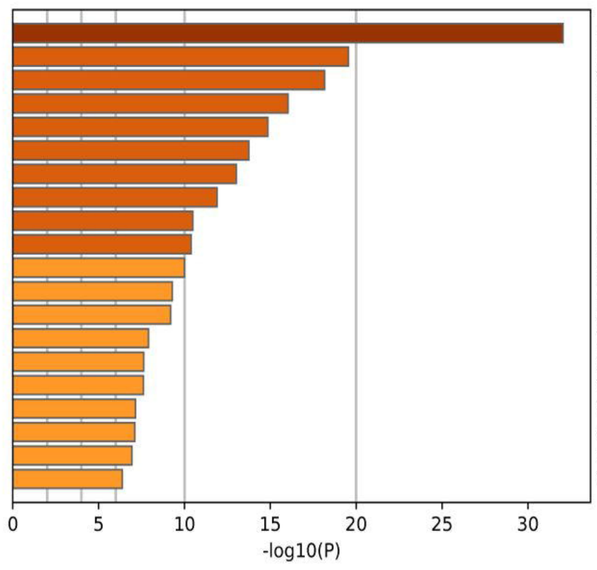
A

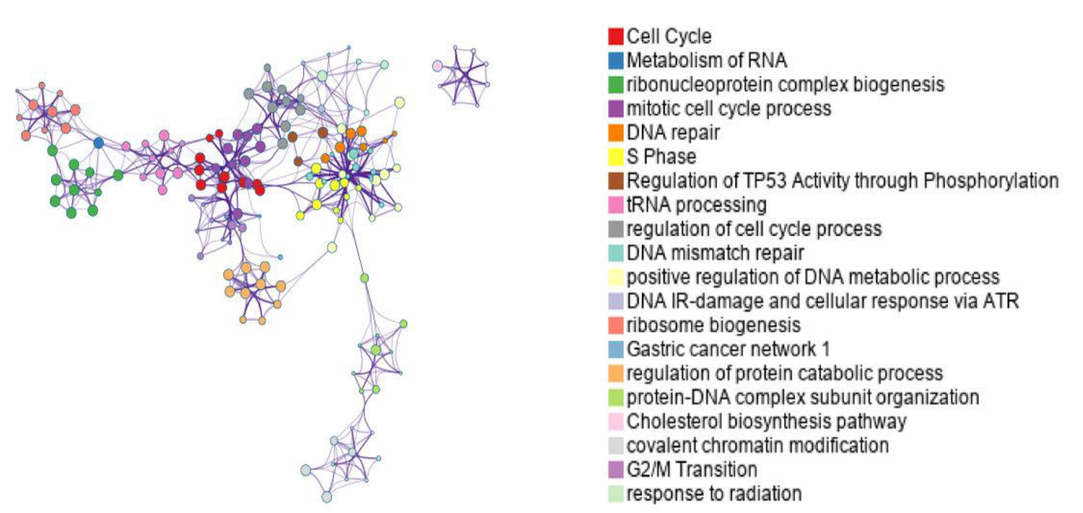


B

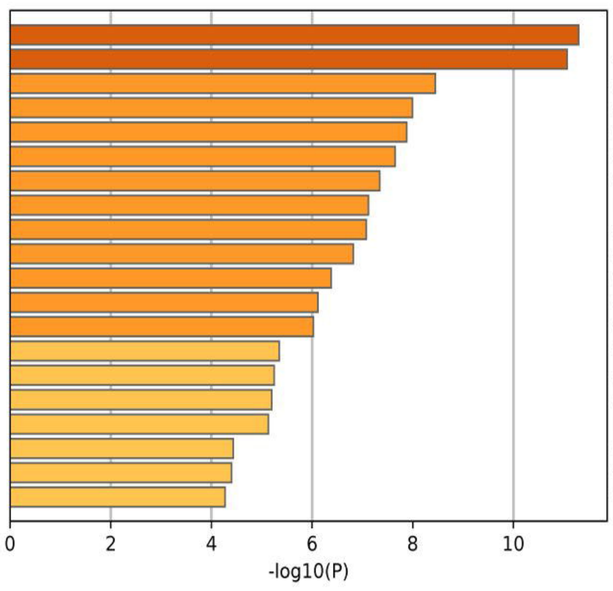


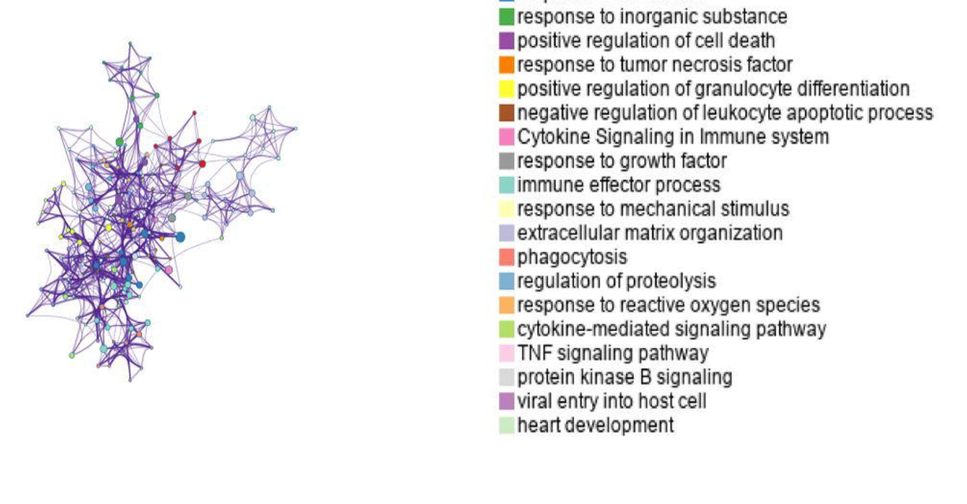
C



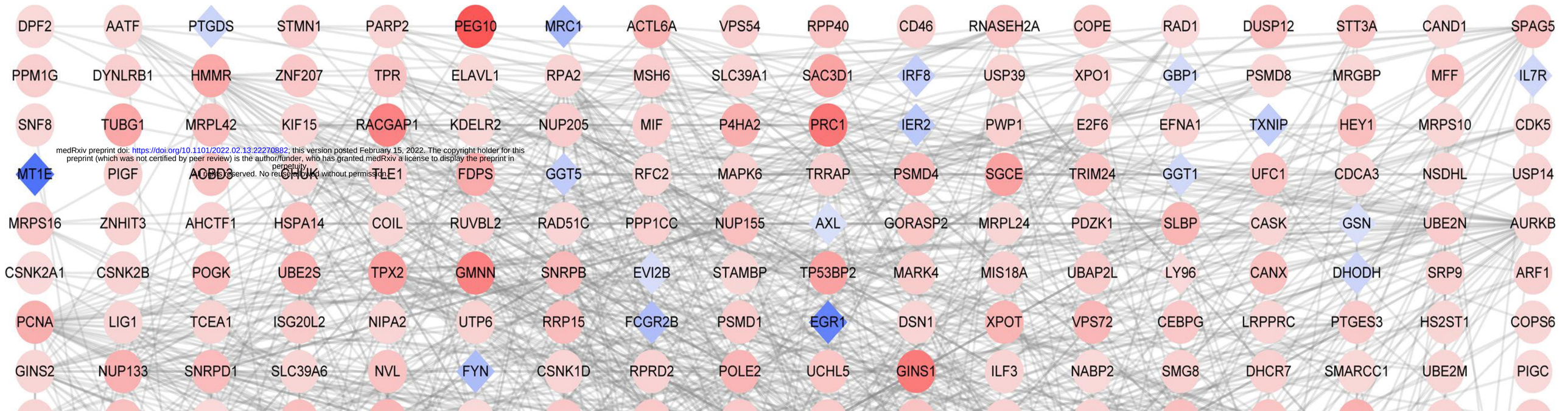


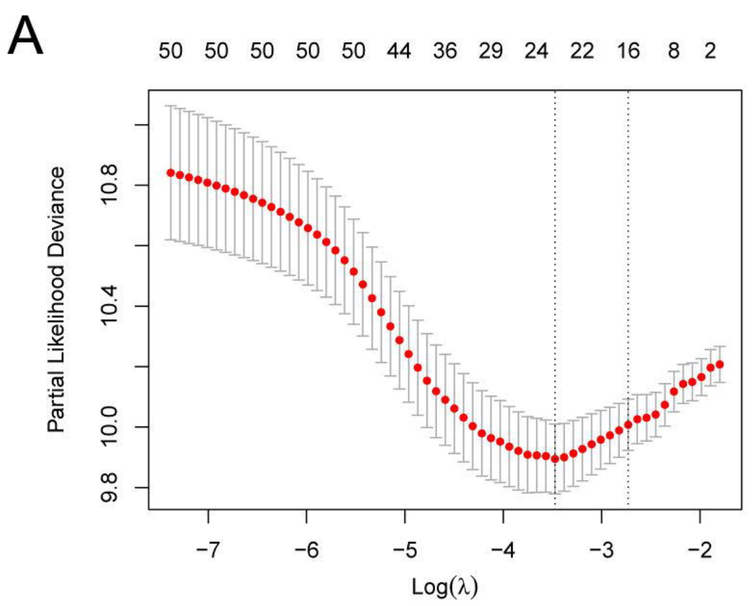
D

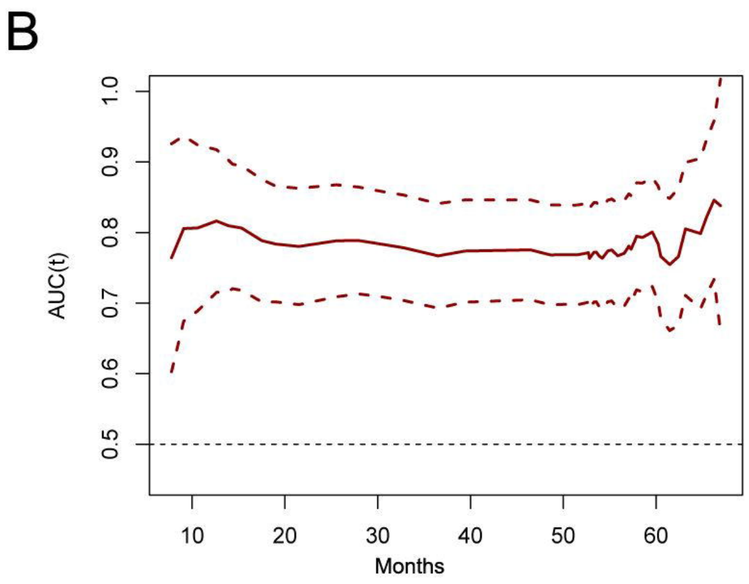


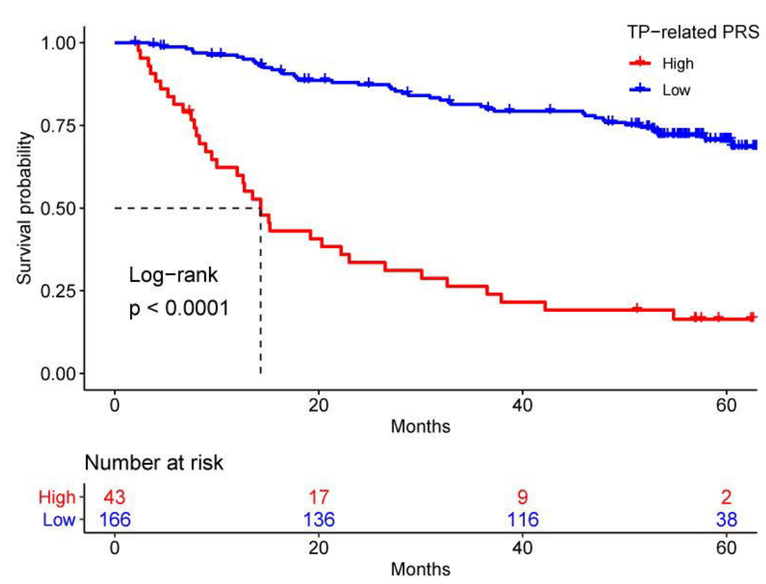


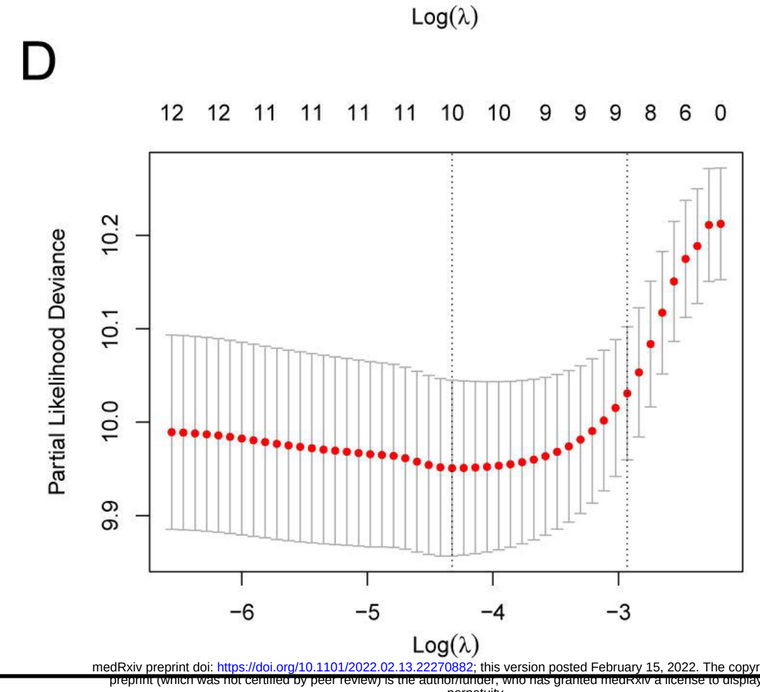
E

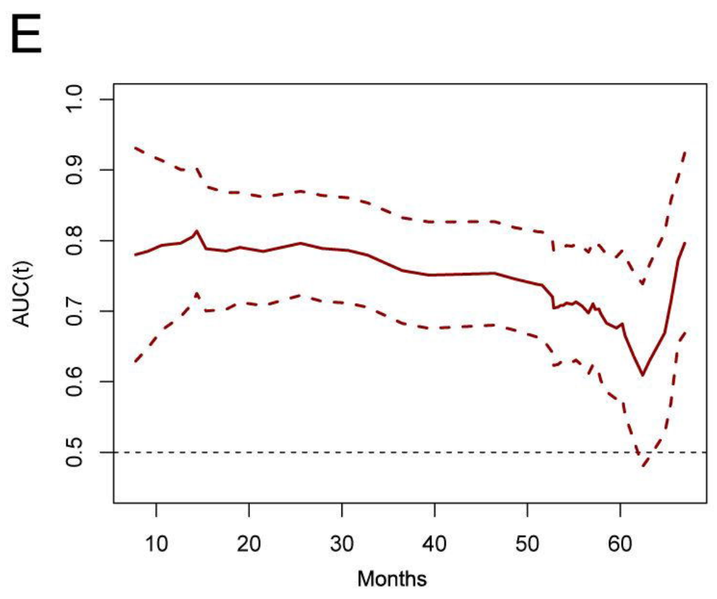


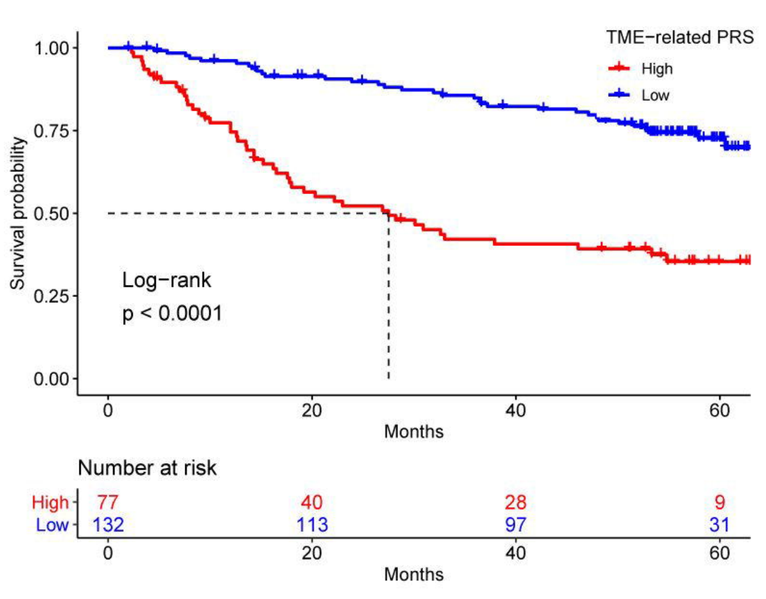


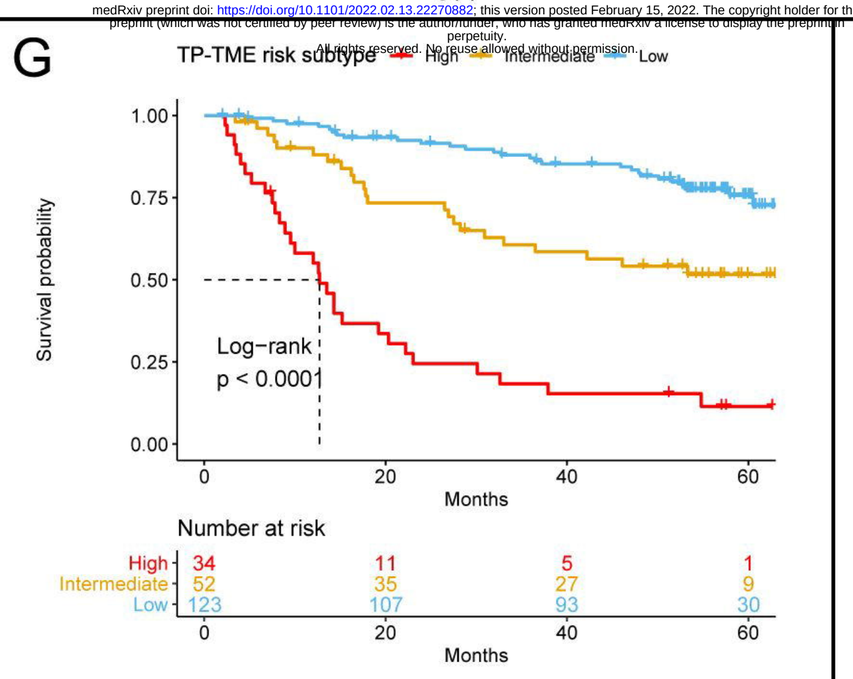


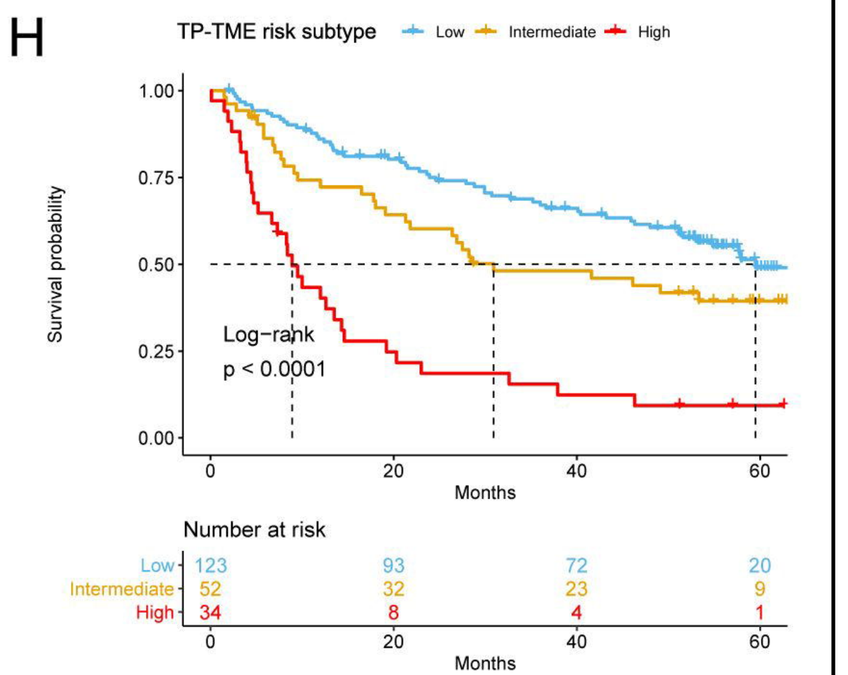


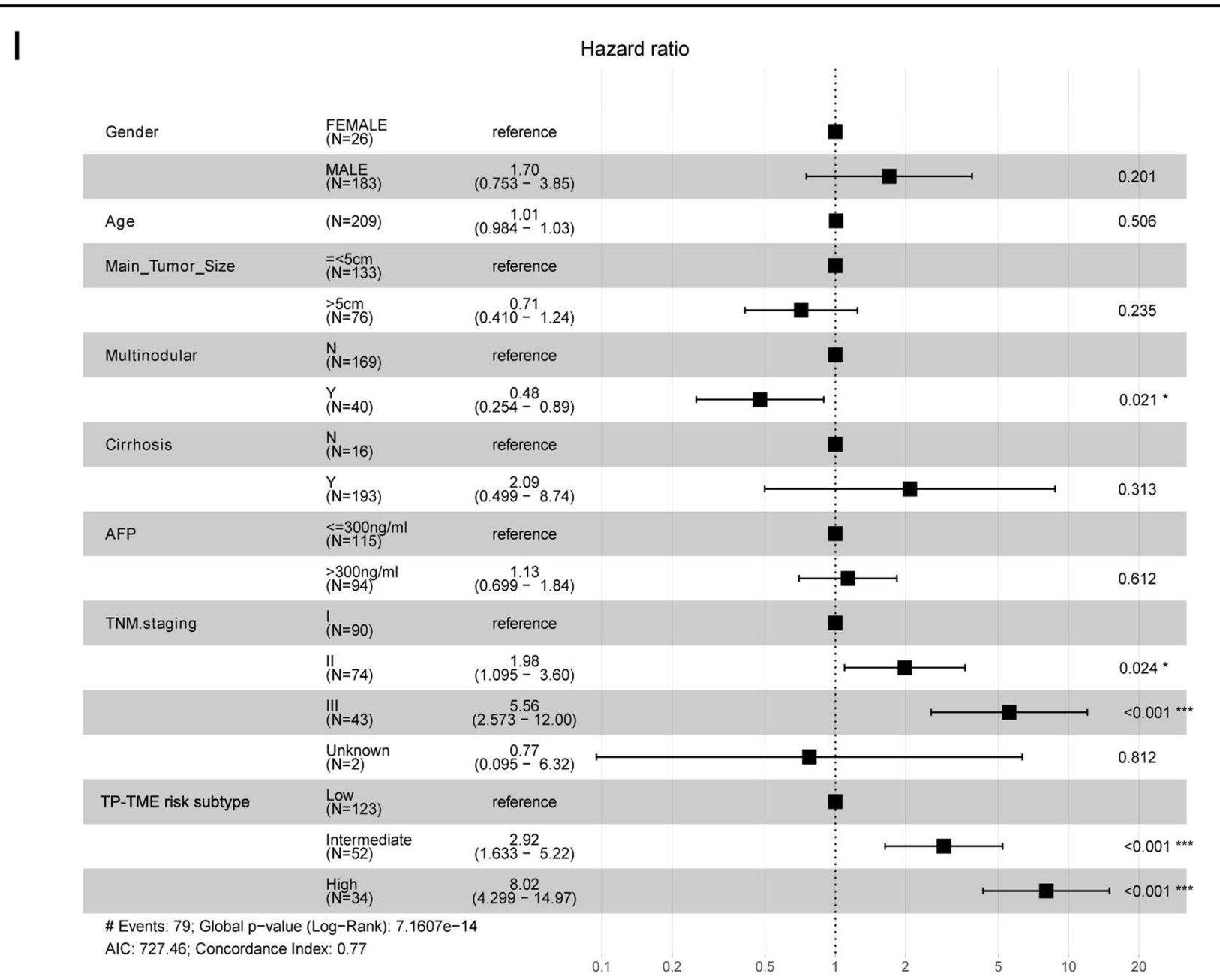




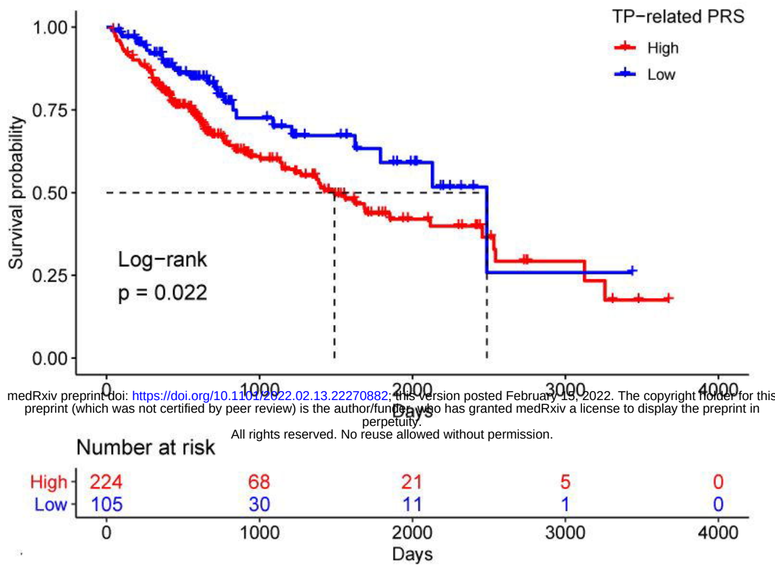




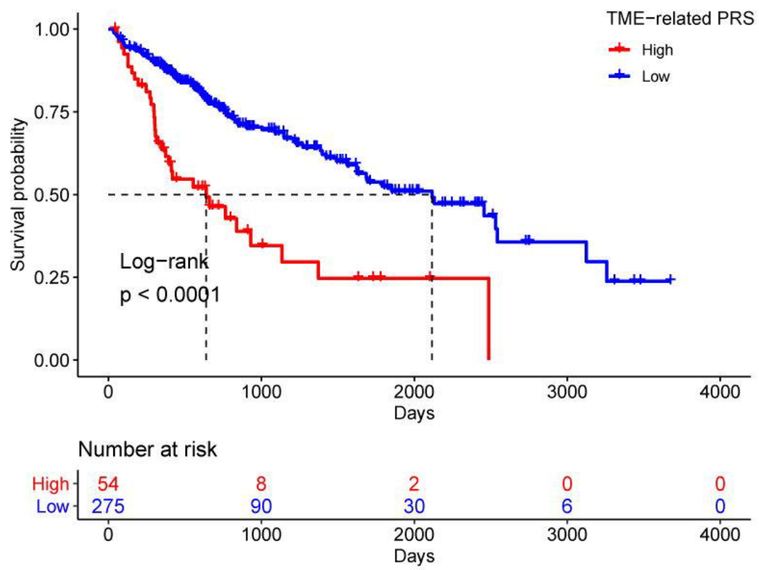




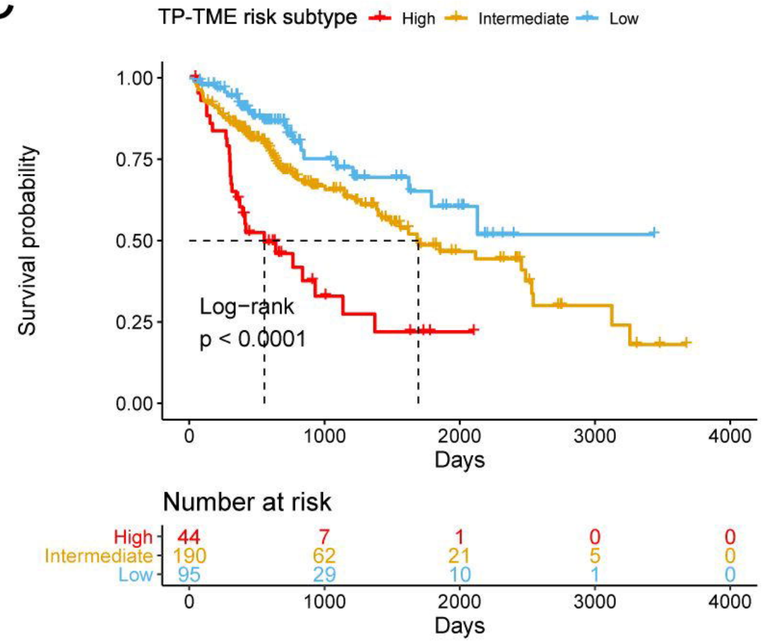
A



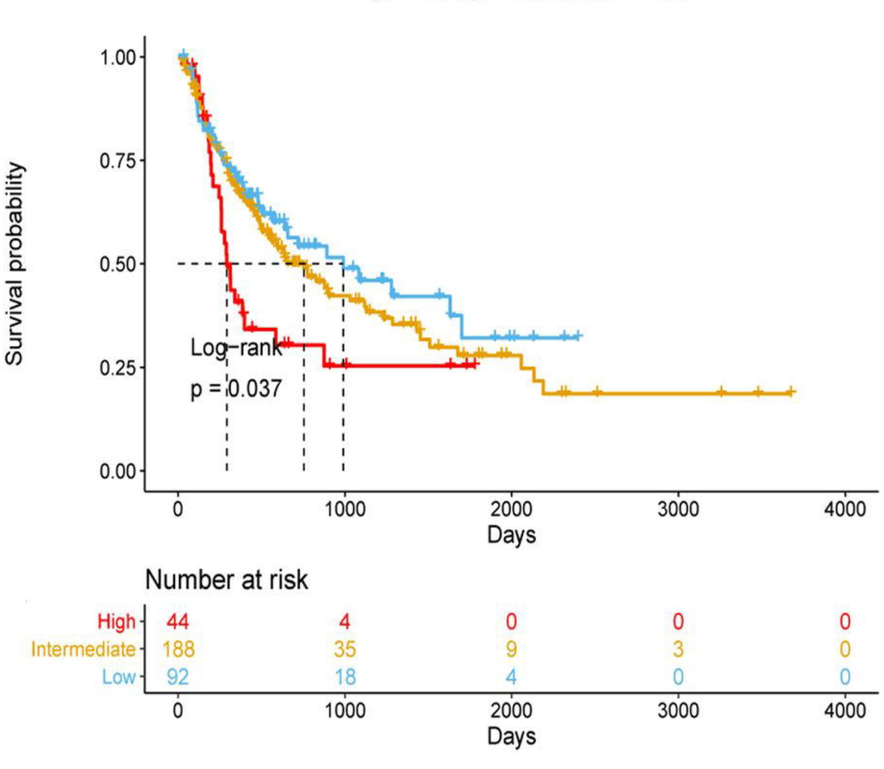
B



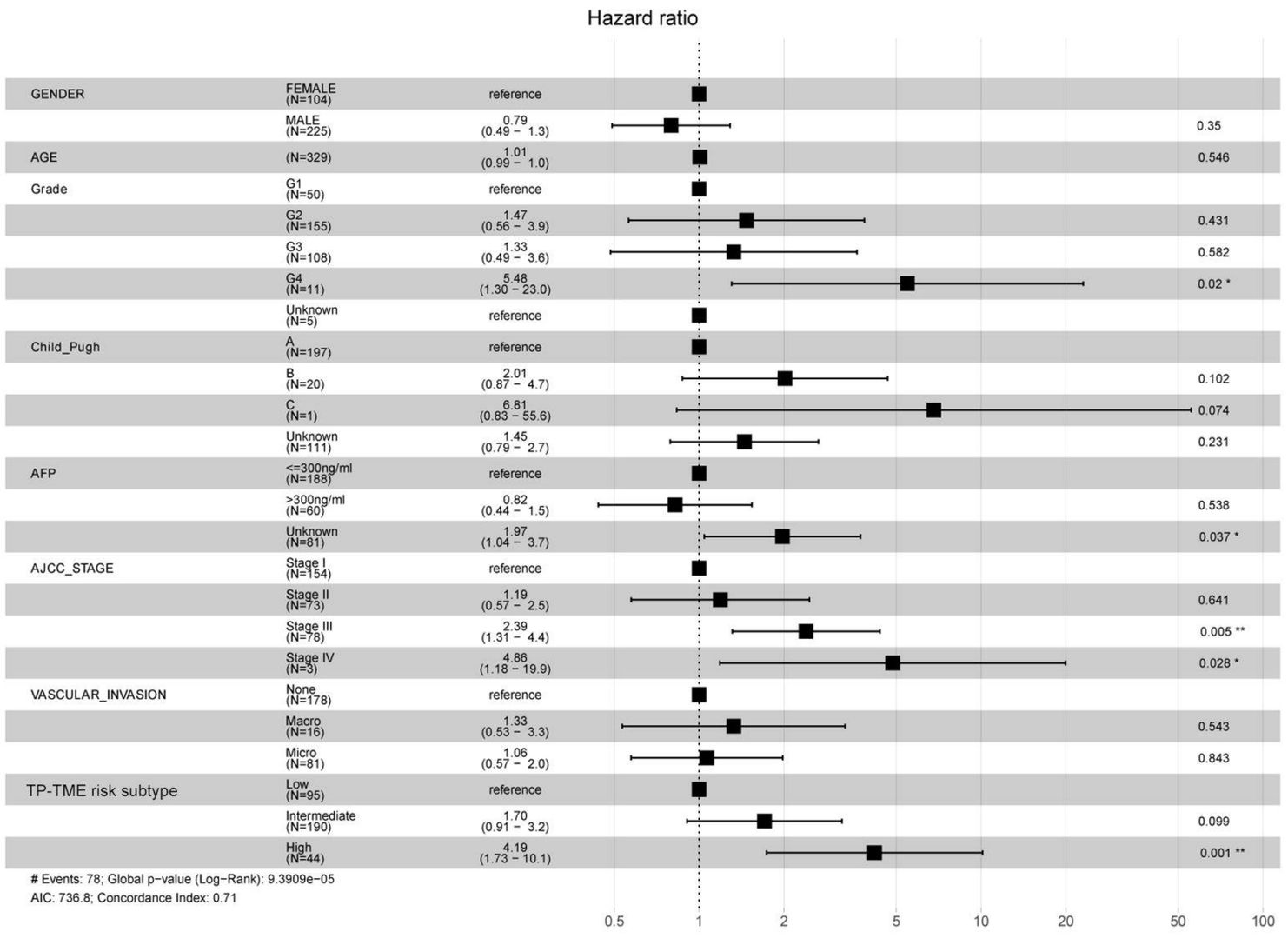
C



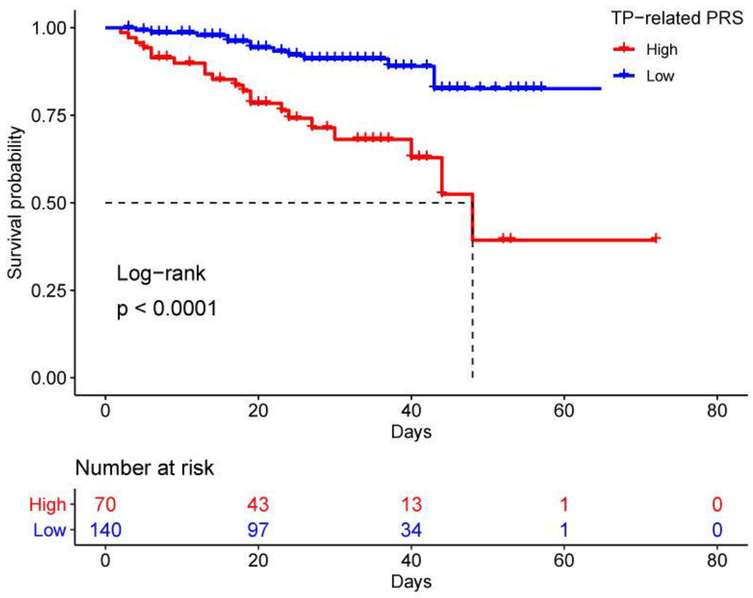
D



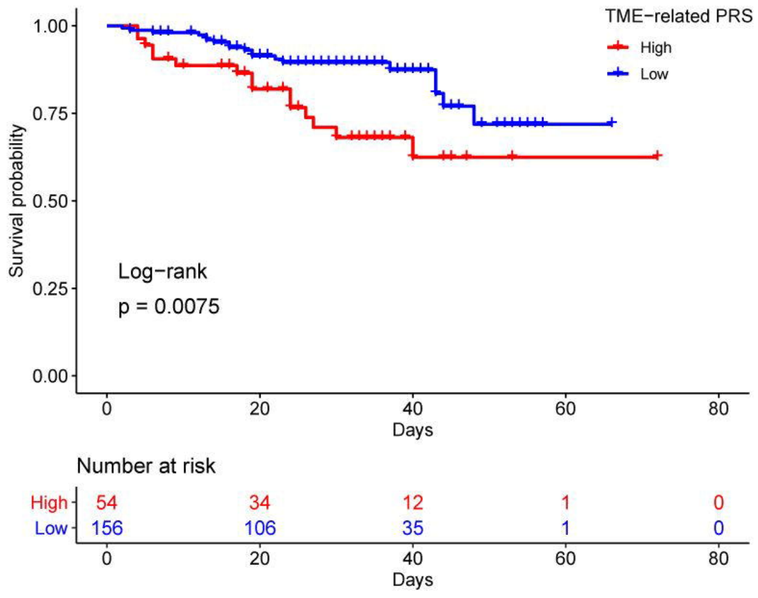
E



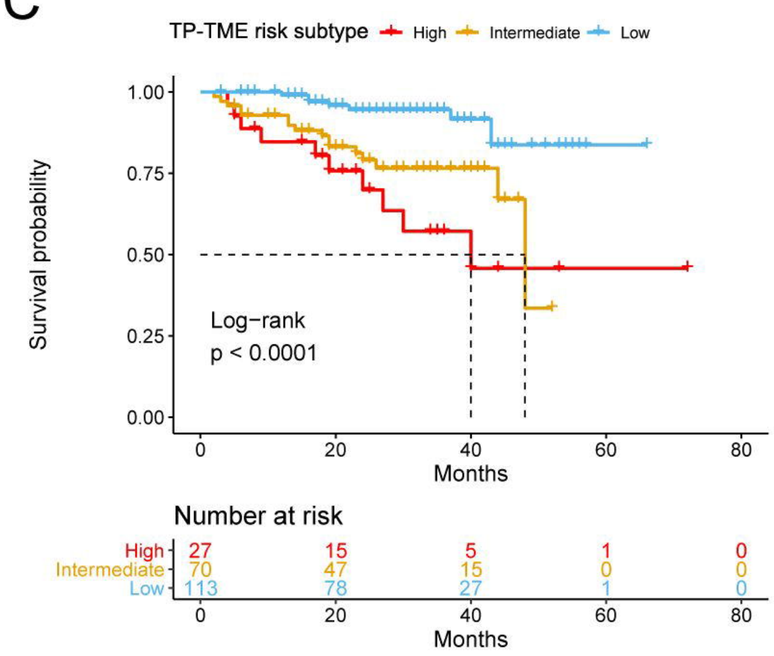
A



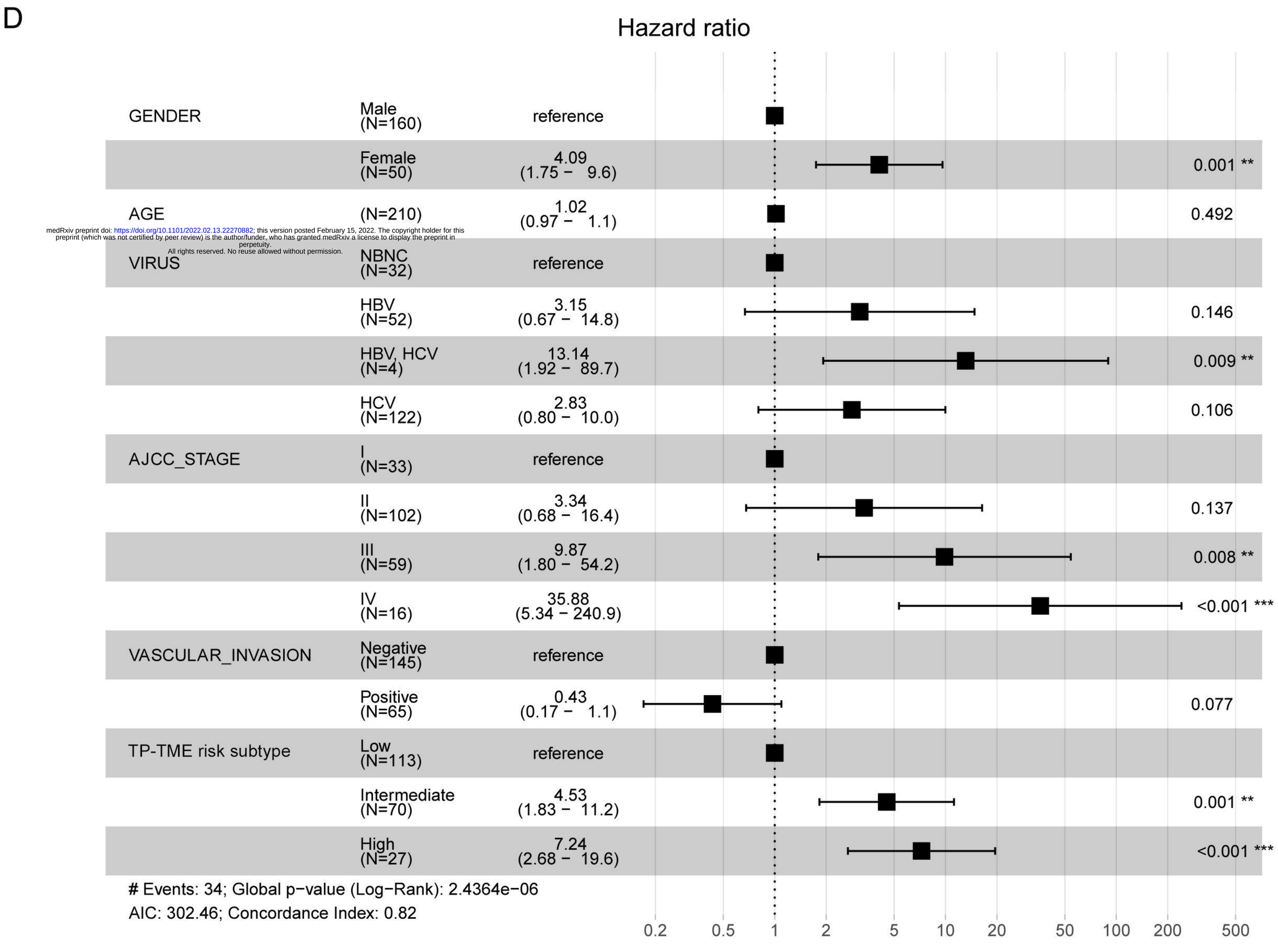
B

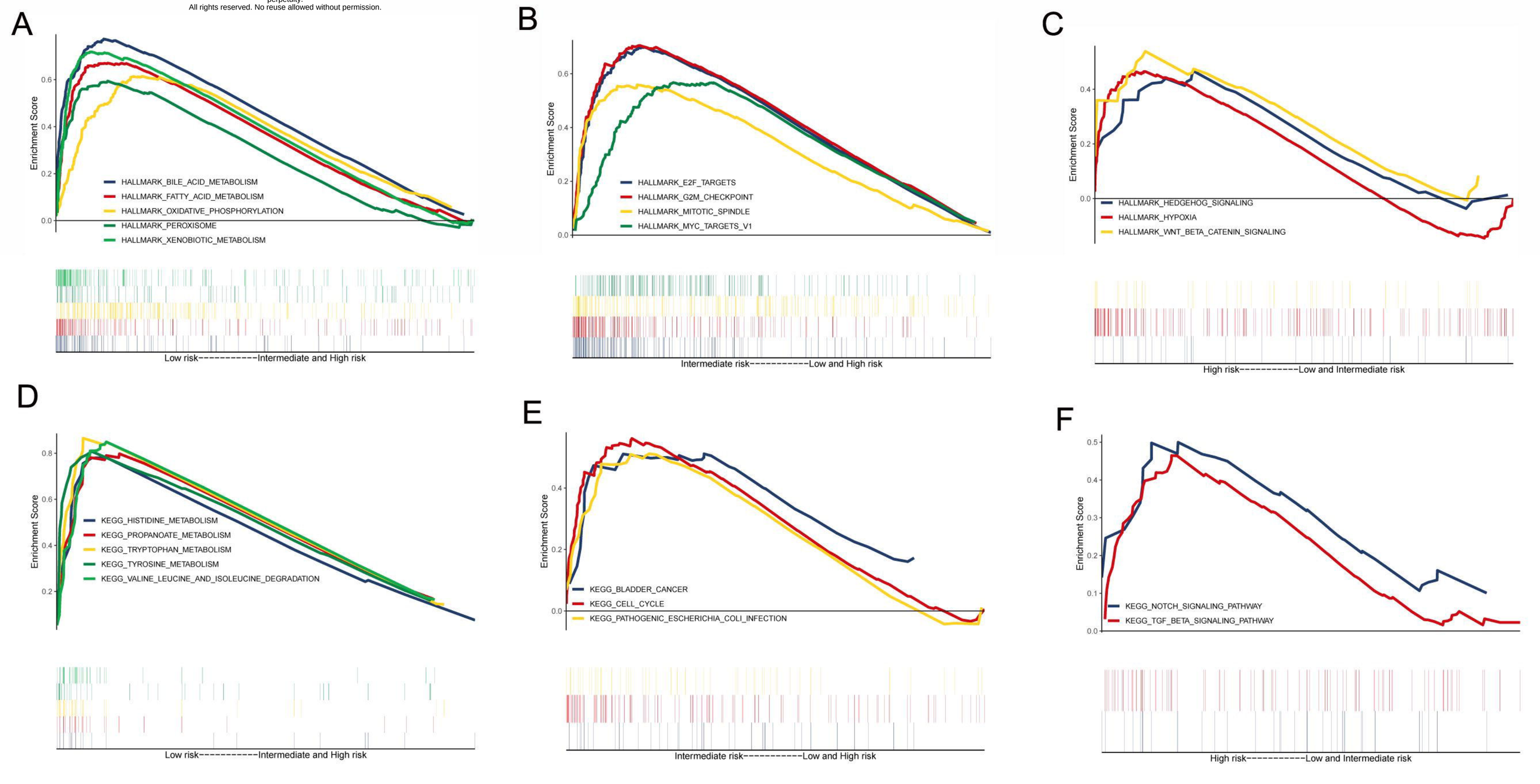


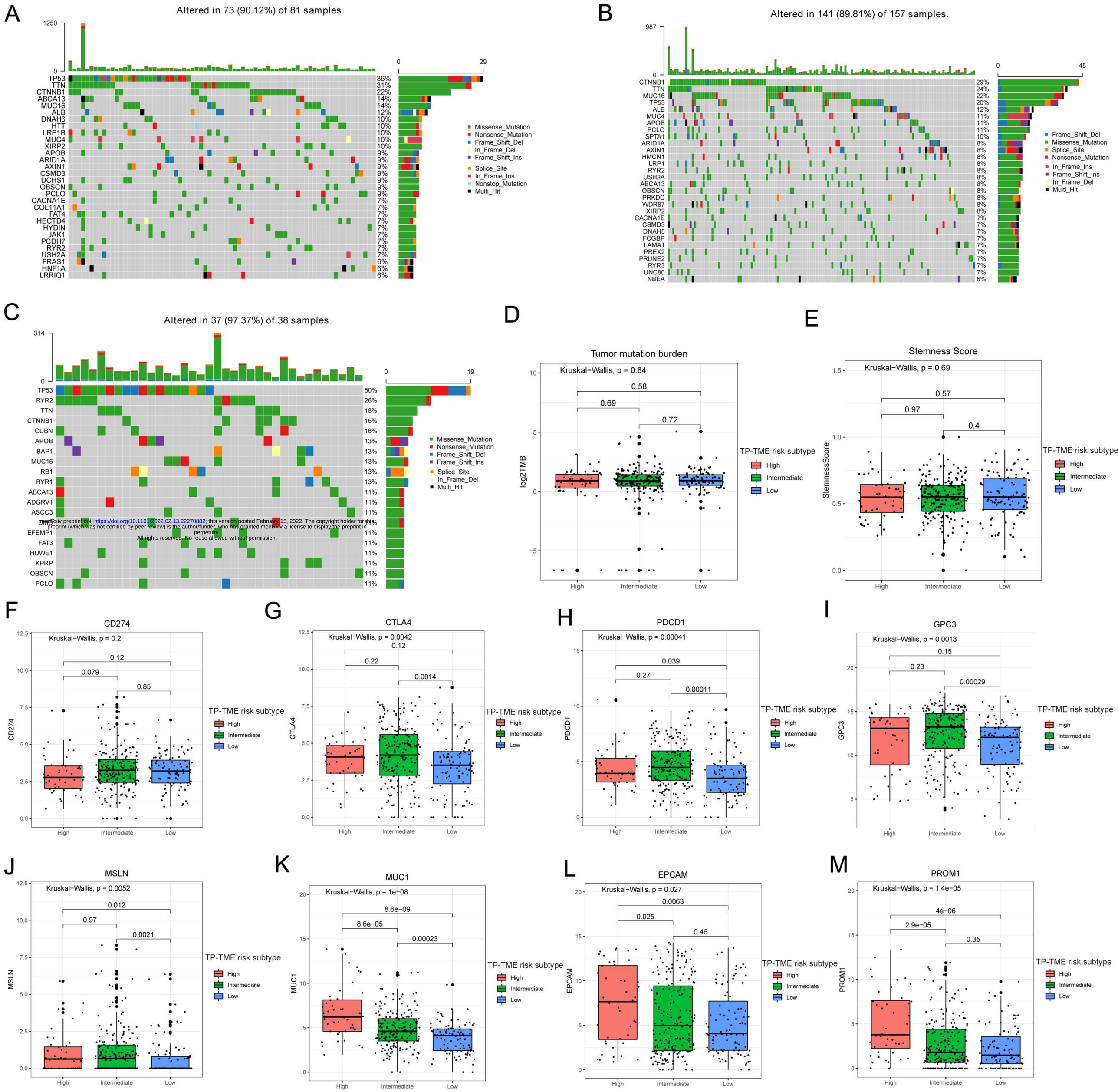
C

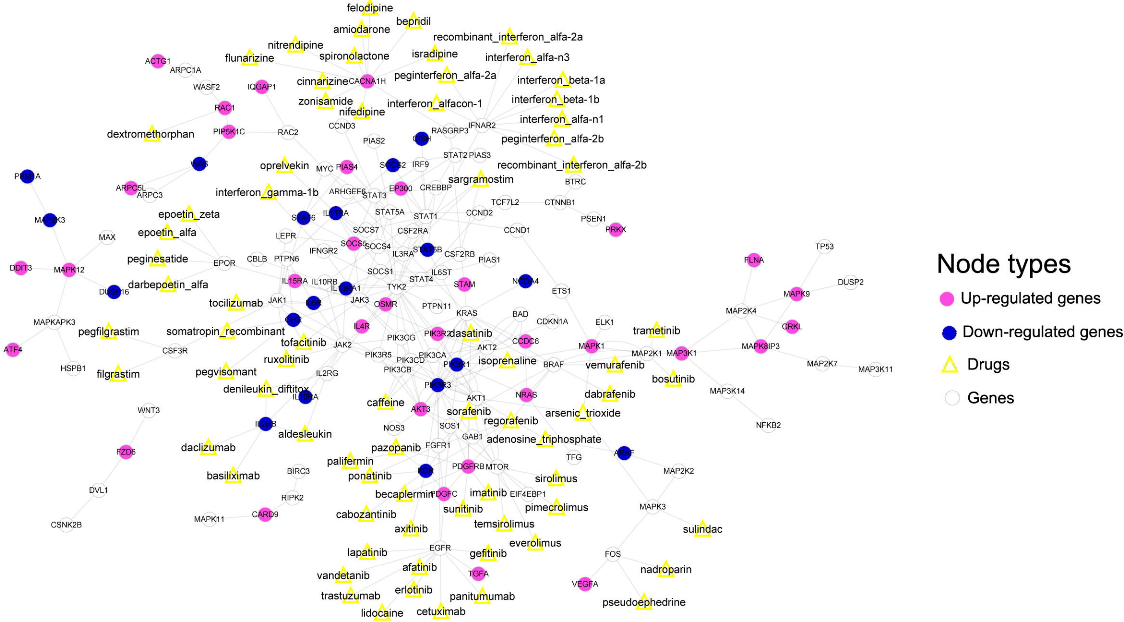


D








felodipine
bepridil
amiodarone
recombinant_interferon_alfa-2a
nitrendipine
flunarizine
spironolactone
isradipine
interferon_alfa-n3
ACTG1
ARPC1A
cinnarizine
CACNA1H
peginterferon_alfa-2a
interferon_beta-1a
WASF2
IQGAP1
zonisamide
interferon_alfacon-1
interferon_beta-1b
RAC1
nifedipine
interferon_alfa-n1
CCND3
IFNAR2
PIP5K1C
RAC2
RASGRP3
dextromethorphan
PIAS2
peginterferon_alfa-2b
STAT2
PIAS3
oprelvekin
MYC
PIAS4
IRF9
recombinant_interferon_alfa-2b
sargramostim
BTRC
ARPC5L
ARPC3
interferon_gamma-1b
ARHGEF6
STAT3
EP300
CREBBP
TCF7L2
CTNNB1
epoetin_zeta
STAT5A
STAT1
CCND2
PSEN1
PRKX
MAX
epoetin_alfa
LEPR
SOCS7
CSF2RA
CCND1
TP53
SOCS5
SOCS4
IFNGR2
IL3RA
CSF2RB
PIAS1
FLNA
peginesatide
EPOR
CBLB
PTPN6
SOCS1
IL6ST
DUSP2
DDIT3
MAPK12
darbepoetin_alfa
IL15RA
IL10RB
TYK2
STAT4
STAM
ETS1
MAPK9
tocilizumab
JAK1
JAK3
OSMR
PTPN11
KRAS
BAD
MAP2K4
MAPKAPK3
IL4R
CDKN1A
ELK1
CRKL
pegfilgrastim
somatropin_recombinant
dasatinib
trametinib
ATF4
CSF3R
tofacitinib
JAK2
PIK3CG
PIK3R2
CCDC6
MAPK1
MAP8IP3
AKT2
MAP2K1
MAP3K1
HSPB1
ruxolitinib
PIK3R5
PIK3CD
PIK3CA
isoprenaline
BRAF
vemurafenib
MAP2K7
MAP3K11
filgrastim
pegvisomant
IL2RG
PIK3CB
bosutinib
denileukin_diftitox
AKT1
NRAS
dabrafenib
MAP3K14
caffeine
AKT3
WNT3
sorafenib
arsenic_trioxide
NOS3
regorafenib
NFKB2
SOS1
aldesleukin
GAB1
adenosine_triphosphate
daclizumab
pazopanib
FGFR1
FZD6
TFG
palifermin
BIRC3
basiliximab
ponatinib
PDGFRB
MTOR
MAP2K2
sirolimus
DVL1
becaplermin
PDGFC
imatinib
EIF4EBP1
RIPK2
pimecrolimus
sunitinib
MAPK3
MAPK11
CARD9
cabozantinib
temsirolimus
CSNK2B
axitinib
sulindac
everolimus
EGFR
lapatinib
gefitinib
FOS
afatinib
nadroparin
vandetanib
TGFA
VEGFA
erlotinib
trastuzumab
panitumumab
pseudoephedrine
lidocaine
cetuximab
Node types
Up-regulated genes
Down-regulated genes
Drugs
Genes