

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

1 Yan Lin<sup>1†</sup>, Rong Liang<sup>1†</sup>, Xing Gao<sup>1</sup>, Ziqin He<sup>1</sup>, Lu Lu<sup>1</sup>, Min Luo<sup>1</sup>, Qian Li<sup>1</sup>,  
2 Xiaobo Wang<sup>2</sup>, Yongqiang Li<sup>1</sup>, Guobin Wu<sup>2\*</sup>, Xiaoling Luo<sup>3\*</sup>, Jiazhou Ye<sup>2\*</sup>

3 <sup>1</sup> Department of Medical Oncology, Guangxi Medical University Cancer Hospital,  
4 Nanning, Guangxi, People's Republic of China.

5 <sup>2</sup> Department of Hepatobiliary Surgery, Guangxi Medical University Cancer Hospital,  
6 Nanning, Guangxi, People's Republic of China.

7 <sup>3</sup> Department of Experimental Research, Guangxi Medical University Cancer Hospital,  
8 Nanning, Guangxi, People's Republic of China.

9 <sup>†</sup>These authors have contributed equally to this work and share first authorship

## 10 \* Correspondence:

11 Jiazhou Ye

12 yejiazhou@gxmu.edu.cn

13 Xiaoling Luo

14 luoxiaoling@gxmu.edu.cn

15 Guobin Wu

16 wuguobin@gxmu.edu.cn

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

## 19 Abstract

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

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

29 **NOTE:** This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.  
biological heterogeneity among different subtypes.

## Novel Molecular Classification of HCC

---

30 **Results** The three TP-TME risk subtypes of HCC had significantly different  
31 prognosis and biological characteristics. The TP-TME low risk subtype had the best  
32 prognosis and was characterized by well-differentiated, the TP-TME high risk  
33 subtype had the worst prognosis and was characterized by aberrant activation of  
34 TGF $\beta$  and WNT pathways, and the TP-TME high risk subtype had the moderate  
35 prognosis and was characterized by exhibited activated MYC targets and  
36 proliferation-related gene sets. These three TP-TME risk subtypes may respond  
37 differently to immunotherapy (e.g., immune checkpoint inhibitors and chimeric  
38 antigen receptor-modified T cells) or other drug therapies.

39 **Conclusion** By combining the TP-related PRS and TME-related PRS, we proposed  
40 and validated the TP-TME risk subtyping system to divide patients with HCC into  
41 three subtypes with distinct biological characteristics and prognoses. These findings  
42 highlight the significant clinical implications of the TP-TME risk subtyping system  
43 and provide potential personalized immunotherapy strategies for HCC.

## 44 INTRODUCTION

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

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

## Novel Molecular Classification of HCC

---

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

### 71 **MATERIALS AND METHODS**

#### 72 **Data Processing**

73 We screened the HCCDB database (<http://lifeome.net/database/hccdb/home.html>)(9)  
74 to find the candidate data sets. The inclusion criteria were as follows: 1) the data set  
75 included both gene expression profiles and prognosis of patients with HCC, 2) the  
76 number of patients with a survival of more than 30 days should be more than 100, 3)  
77 and the gene expression profile of the dataset should contain more than 10,000 genes.  
78 According to the inclusion criteria, three data sets (GSE14520\_GPL3921,  
79 TCGA-LIHC, and LIRI-JP) were selected and downloaded from HCCDB database  
80 for our analyses. Data set GSE14520\_GPL3921(10) included the gene expression  
81 profiles based on the GPL3921 platform containing 225 HCC and 220 tumor-adjacent  
82 liver tissue samples were utilized to develop our subtyping systems. TCGA-LIHC  
83 data set containing RNA sequencing (RNA-seq) data and clinical information of 356  
84 HCC belongs to The Cancer Genome Atlas Program (<https://www.cancer.gov/tcga>),  
85 and LIRI-JP data set containing RNA-seq data and clinical information of 212 HCC  
86 from JP Project from International Cancer Genome Consortium (<https://dcc.icgc.org/>)  
87 were used to validated the subtyping systems. The workflow of the present was shown  
88 in **Figure 1**.

#### 89 **Calculation of tumor purity and TME score and identification of differentially** 90 **expressed genes (DEGs) in GSE14520\_GPL3921**

91 The gene expression profiles of GSE14520\_GPL3921 were firstly utilized to calculate  
92 tumor purity (TP) using the ESTIMATE package.(11) Then, GSE14520\_GPL3921  
93 was also used to calculate the TME score using the xCell tools  
94 (<https://xcell.ucsf.edu/>)(12) with the xCell gene signature. The DEGs in HCC  
95 compared to tumor-adjacent liver tissue were identified using limma package.(13)  
96 Genes with fold changes > 1.5 and P (adjusted by false discovery rate) value < 0.05  
97 were considered significant.

#### 98 **Normality test and correlation analysis**

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

## Novel Molecular Classification of HCC

---

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

### 106 **Protein-protein interaction (PPI) networks**

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

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

112 Firstly, for developing the TP-related polygenic risk score (PRS), the overall survival  
113 (OS)-associated TP-related genes were identified using univariate Cox regression  
114 analysis. Secondly, the expression profiles of the OS-associated TP-related genes  
115 were used to carry out the least absolute shrinkage and selection operator (LASSO)  
116 Cox regression model analysis with leave-one-out cross validation using glmnet  
117 package.(16) The genes with non-zero coefficient were considered the optimal  
118 features and subjected to multivariate Cox regression and stepwise regression analysis.  
119 Then the TP-related PRS was developed as the formula: TP-related PRS =  $\sum$   
120 (Expression<sub>i</sub> \* Coefficient<sub>i</sub>). Where the “Coefficient” and “Expression” represent the risk  
121 coefficient and expression of each gene in the multivariate cox regression and  
122 stepwise regression analysis, respectively. The TME-related PRS was also developed  
123 according to the same method as above.

### 124 **The TP-TME subtyping of HCC**

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

### 135 **Gene set enrichment analysis (GSEA)**

136 In order to preliminarily reveal the biological characteristics of these three risk  
137 subtypes, we performed GSEA(17, 18) in the GSE14520\_GPL3921 data set using the  
138 GSEA java software (<http://www.gsea-msigdb.org/gsea/index.jsp>). Hallmark gene

## Novel Molecular Classification of HCC

---

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

### 143 **Analyses of gene mutation and stemness score**

144 Gene mutation data of TCGA-LIHC data set were extracted from mutation annotation  
145 format (MAF) files using the *GDCquery\_Maf* function in the “TCGAbiolinks”  
146 package.(21) Gene mutation frequencies of each risk subtype were visualized as a  
147 waterfall plot using the *oncoplot* function in the “TCGAbiolinks” package. The tumor  
148 mutational burden (TMB) of each sample was obtained from a previous study.(22)  
149 The stemness score(23) was calculated for each individual in the TCGA-LIHC data  
150 set using *TCGAanalyze\_Stemness* function in the “TCGAbiolinks” package.

### 151 **Prediction of efficacy of therapy**

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

### 163 **Statistical analysis**

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

175 **RESULTS**

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

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

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

191 Fifty TP-related genes were identified as OS-associated genes, twenty-two of them  
192 possess non-zero coefficient (**Figure 3A**), and eleven genes (*ALG6*, *ATP5MF*, *CNIH4*,  
193 *ESM1*, *HEY1*, *LANCL1*, *P2RX4*, *PEX11B*, *POP7*, *RCN2*, and *XPO1*) were used to  
194 generate the TP-related polygenic risk score (PRS) (**Supplementary Table 1**). The  
195 TP-related PRS was significantly associated with overall survival (OS) with  $P <$   
196  $0.0001$ , Hazard Ratio (HR) = 2.718 (95%CI for HR = 2.147-3.442) and the area under  
197 curve (AUC) of tROC analysis was stably around 0.8 (**Figure 3B**). The HCC patients  
198 with high TP-related PRS showed shorter OS than those with low TP-related PRS ( $P$   
199  $< 0.0001$ ) (**Figure 3C**). In the TME-related genes, twelve genes were identified as  
200 OS-associated genes, ten of them possess non-zero coefficient (**Figure 3D**), and seven  
201 genes (*ALDH1B1*, *CTSC*, *GUCY1A1*, *MRC1*, *SPRY2*, *TARP*, and *TRIM22*) were used  
202 to generate the TME-related PRS (**Supplementary Table 2**). The TME-related PRS  
203 was also significantly associated with OS with  $P < 0.0001$ , Hazard Ratio (HR) =  
204 2.718 (95%CI for HR = 1.978-3.735) and the AUC of tROC analysis was 0.7-0.8  
205 (**Figure 3E**). The HCC patients with high TME-related PRS showed shorter OS than  
206 those with low TP-related PRS ( $P < 0.0001$ ) (**Figure 3F**). Our TP-TME risk subtype  
207 was generated based the two PRSs, and 34, 52, and 123 patients with HCC were  
208 divided into high-, intermediate-, and low-risk subtypes, respectively. Cases defined  
209 as high-risk subtype had the best poor survival while those in low-risk subtype had the  
210 best survival, and the intermediate-risk cases had a better prognosis than high-risk  
211 subtype and worse than low-risk subtype (**Figure 3G**). A similar trend was also  
212 observed in progression free survival (PFS) (**Figure 3H**). Furthermore, the TP-TME

## Novel Molecular Classification of HCC

213 risk subtyping system showed independent to some routine clinicopathological  
214 features (**Figure 3I**). As it was in the GSE14520\_GPL3921 data set, the TP-related  
215 PRS and the TME-related PRS in the TCGA-LIHC and LIRI-JP data sets were  
216 calculated according to the abovementioned formula, respectively. We found similar  
217 results in the two validation sets. Briefly, the TP-related PRS (**Figure 4A** for  
218 TCGA-LIHC and **Figure 5A** for LIRI-JP) and the TME-related PRS (**Figure 4B** for  
219 TCGA-LIHC and **Figure 5B** for LIRI-JP) were significantly associated with OS, the  
220 TP-TME risk subtyping system was associated with OS (**Figure 4C** for TCGA-LIHC  
221 and **Figure 5C** for LIRI-JP) and PFS (**Figure 4D** for TCGA-LIHC). The prognostic  
222 value of the TP-TME risk subtyping system was independent to the routine  
223 clinicopathological features (**Figure 4E** for TCGA-LIHC and **Figure 5D** for LIRI-JP).  
224 Collectively, the TP-TME risk subtyping is a robust prognosis prediction system.

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

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

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

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

## Novel Molecular Classification of HCC

---

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

### 257 **Potentially effective drugs for the high-risk subtypes**

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

### 265 **DISCUSSION**

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

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



## Novel Molecular Classification of HCC

---

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

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

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

## 322 CONCLUSIONS

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

## Novel Molecular Classification of HCC

---

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

### 329 **ACKNOWLEDGMENTS**

330 Not available

### 331 **COMPETING INTERESTS**

332 The authors declare that they have no competing interests.

### 333 **FUNDING**

334 This research was supported by the National Natural Science Foundation of China  
335 (NO. 81803007, 82060427, 82103297), Guangxi Key Research and Development  
336 Plan (NO. GUIKEAB19245002), Guangxi Scholarship Fund of Guangxi Education  
337 Department, Guangxi Natural Science Foundation (NO. 2020GXNSFAA259080),  
338 Guangxi Medical University Training Program for Distinguished Young Scholars,  
339 Science and Technology Plan Project of Qingxiu District, Nanning (NO. 2020037,  
340 2020038).

### 341 **AUTHORS CONTRIBUTIONS**

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

### 346 **DATA AVAILABILITY**

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

### 349 **CODE AVAILABILITY**

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

### 352 **ETHICS APPROVAL**

353 Not available.

## Novel Molecular Classification of HCC

---

### 354 **CONSENT TO PARTICIPATE**

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

### 356 **CONSENT TO PUBLISH**

357 Not available.

### 358 **REFERENCES**

- 359 1. Forner A, Reig M, Bruix J. Hepatocellular carcinoma. *Lancet*.  
360 2018;391(10127):1301-14.
- 361 2. Schulze K, Nault JC, Villanueva A. Genetic profiling of hepatocellular carcinoma  
362 using next-generation sequencing. *J Hepatol*. 2016;65(5):1031-42.
- 363 3. Minagawa M, Ikai I, Matsuyama Y, Yamaoka Y, Makuuchi M. Staging of  
364 hepatocellular carcinoma: assessment of the Japanese TNM and AJCC/UICC  
365 TNM systems in a cohort of 13,772 patients in Japan. *Ann Surg*.  
366 2007;245(6):909-22.
- 367 4. Nault JC, Martin Y, Caruso S, Hirsch TZ, Bayard Q, Calderaro J, et al. Clinical  
368 Impact of Genomic Diversity From Early to Advanced Hepatocellular Carcinoma.  
369 *Hepatology*. 2020;71(1):164-82.
- 370 5. Wu Y, Liu Z, Xu X. Molecular subtyping of hepatocellular carcinoma: A step  
371 toward precision medicine. *Cancer Commun (Lond)*. 2020;40(12):681-93.
- 372 6. Pinyol R, Montal R, Bassaganyas L, Sia D, Takayama T, Chau GY, et al.  
373 Molecular predictors of prevention of recurrence in HCC with sorafenib as  
374 adjuvant treatment and prognostic factors in the phase 3 STORM trial. *Gut*.  
375 2019;68(6):1065-75.
- 376 7. Hanahan D, Coussens LM. Accessories to the crime: functions of cells recruited to  
377 the tumor microenvironment. *Cancer Cell*. 2012;21(3):309-22.
- 378 8. Pitt JM, Marabelle A, Eggermont A, Soria JC, Kroemer G, Zitvogel L. Targeting  
379 the tumor microenvironment: removing obstruction to anticancer immune  
380 responses and immunotherapy. *Ann Oncol*. 2016;27(8):1482-92.
- 381 9. Lian Q, Wang S, Zhang G, Wang D, Luo G, Tang J, et al. HCCDB: A Database of  
382 Hepatocellular Carcinoma Expression Atlas. *Genomics Proteomics  
383 Bioinformatics*. 2018;16(4):269-75.

## Novel Molecular Classification of HCC

---

- 384 10. Roessler S, Jia HL, Budhu A, Forgues M, Ye QH, Lee JS, et al. A unique  
385 metastasis gene signature enables prediction of tumor relapse in early-stage  
386 hepatocellular carcinoma patients. *Cancer Res.* 2010;70(24):10202-12.
- 387 11. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia  
388 W, et al. Inferring tumour purity and stromal and immune cell admixture from  
389 expression data. *Nat Commun.* 2013;4:2612.
- 390 12. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular  
391 heterogeneity landscape. *Genome Biol.* 2017;18(1):220.
- 392 13. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers  
393 differential expression analyses for RNA-sequencing and microarray studies.  
394 *Nucleic Acids Res.* 2015;43(7):e47.
- 395 14. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al.  
396 STRING v11: protein-protein association networks with increased coverage,  
397 supporting functional discovery in genome-wide experimental datasets. *Nucleic  
398 Acids Res.* 2019;47(D1):D607-D13.
- 399 15. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape:  
400 a software environment for integrated models of biomolecular interaction  
401 networks. *Genome Res.* 2003;13(11):2498-504.
- 402 16. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear  
403 Models via Coordinate Descent. 2010. 2010;33(1):22.
- 404 17. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al.  
405 PGC-1alpha-responsive genes involved in oxidative phosphorylation are  
406 coordinately downregulated in human diabetes. *Nat Genet.* 2003;34(3):267-73.
- 407 18. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et  
408 al. Gene set enrichment analysis: a knowledge-based approach for interpreting  
409 genome-wide expression profiles. *Proc Natl Acad Sci U S A.*  
410 2005;102(43):15545-50.
- 411 19. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov  
412 JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics.*  
413 2011;27(12):1739-40.
- 414 20. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The  
415 Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.*  
416 2015;1(6):417-25.

## Novel Molecular Classification of HCC

---

- 417 21. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al.  
418 TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data.  
419 Nucleic Acids Res. 2016;44(8):e71.
- 420 22. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, et al. The  
421 Immune Landscape of Cancer. *Immunity*. 2018;48(4):812-30 e14.
- 422 23. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, et al.  
423 Machine Learning Identifies Stemness Features Associated with Oncogenic  
424 Dedifferentiation. *Cell*. 2018;173(2):338-54 e15.
- 425 24. Kruger S, Ilmer M, Kobold S, Cadilha BL, Endres S, Ormanns S, et al. Advances  
426 in cancer immunotherapy 2019 - latest trends. *J Exp Clin Cancer Res*.  
427 2019;38(1):268.
- 428 25. Kole C, Charalampakis N, Tsakatikas S, Vailas M, Moris D, Gkotsis E, et al.  
429 Immunotherapy for Hepatocellular Carcinoma: A 2021 Update. *Cancers (Basel)*.  
430 2020;12(10).
- 431 26. Paver EC, Cooper WA, Colebatch AJ, Ferguson PM, Hill SK, Lum T, et al.  
432 Programmed death ligand-1 (PD-L1) as a predictive marker for immunotherapy in  
433 solid tumours: a guide to immunohistochemistry implementation and  
434 interpretation. *Pathology*. 2021;53(2):141-56.
- 435 27. Ma S, Li X, Wang X, Cheng L, Li Z, Zhang C, et al. Current Progress in CAR-T  
436 Cell Therapy for Solid Tumors. *Int J Biol Sci*. 2019;15(12):2548-60.
- 437 28. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing  
438 time-dependent areas under receiver operating characteristic curves for censored  
439 event times with competing risks. *Stat Med*. 2013;32(30):5381-97.
- 440 29. Lu LC, Hsu CH, Hsu C, Cheng AL. Tumor Heterogeneity in Hepatocellular  
441 Carcinoma: Facing the Challenges. *Liver Cancer*. 2016;5(2):128-38.
- 442 30. Hinshaw DC, Shevde LA. The Tumor Microenvironment Innately Modulates  
443 Cancer Progression. *Cancer Res*. 2019;79(18):4557-66.
- 444 31. Cancer Genome Atlas Research Network. Electronic address wbe, Cancer  
445 Genome Atlas Research N. Comprehensive and Integrative Genomic  
446 Characterization of Hepatocellular Carcinoma. *Cell*. 2017;169(7):1327-41 e23.

## Novel Molecular Classification of HCC

---

- 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.

## Novel Molecular Classification of HCC

478 42. Wang J, Xiao Q, Chen X, Tong S, Sun J, Lv R, et al. LanCL1 protects prostate  
479 cancer cells from oxidative stress via suppression of JNK pathway. *Cell Death Dis.*  
480 2018;9(2):197.

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

483

### 484 **Figure legends**

485 **Figure 1 The workflow of the present study.**

486 **Abbreviation:** LASSO, least absolute shrinkage and selection operator; TP, tumor  
487 purity; TME, tumor microenvironment; TCGA-LIHC, The Cancer Genome  
488 Atlas-liver hepatocellular carcinoma; ICGC-LIRI-JP: Liver Cancer-RIKEN, JP.

489 **Figure 2 The identification, enrichment analysis, and protein-protein interaction  
490 networks of the tumor purity-related genes and tumor**

491 **microenvironment-related genes.** (A) The volcano plot of the differentially  
492 expressed gene analysis; (B) Hierarchical clustering showed the expression patterns of  
493 tumor purity (TP)-related and tumor microenvironment (TME)-related genes basically  
494 distinguish hepatocellular carcinoma (HCC) and tumor-adjacent liver tissues. (C)  
495 Functional enrichment analysis for the TP-related genes. *Left panel:* GO terms and  
496 pathways involving TP-related genes, and *right panel:* interactions among the GO  
497 terms and pathways. (D) Functional enrichment analysis for the TME-related genes.  
498 *Left panel:* GO terms and pathways involving TME-related genes, and *right panel:*  
499 interactions among the GO terms and pathways. (E) protein-protein interaction  
500 networks of the TP-related genes and TME-related genes. The red represents  
501 up-regulated, and the blue represents down-regulated in hepatocellular carcinoma.  
502 The circle nodes were TP-related genes, and the diamond nodes were TME-related  
503 genes.

504 **Abbreviation:** TME, tumor microenvironment; TP, tumor purity; HCC,  
505 hepatocellular carcinoma; GO, gene ontology.

506 **Figure 3 The development of TP-TME risk subtypes in GSE14520\_GPL3921**

507 **data set.** (A) twenty-two TP-related genes had non-zero coefficients in the LASSO  
508 Cox regression model analysis; (B) Time-dependent ROC curve analysis for the  
509 TP-related PRS; (C) HCC with high TP-related PRS had shorter overall survival than  
510 those with low TP-related PRS. (D) Ten TME-related genes had non-zero coefficients  
511 in the LASSO Cox regression model analysis; (E) Time-dependent ROC curve  
512 analysis for the TME-related polygenic risk score; (F) HCC with high TME-related  
513 PRS had shorter overall survival than those with low TME-related PRS. (G) There are

## Novel Molecular Classification of HCC

514 significantly different overall survivals between the three subtypes in the TP-TME  
515 risk subtypes. **(H)** There are significantly different progression free survivals between  
516 the three subtypes in the TP-TME risk subtypes. **(I)** The TP-TME risk subtype system  
517 was proved to be an independent prognostic factor, after adjusting for other  
518 clinicopathological characteristics.

519 **Abbreviation:** LASSO, least absolute shrinkage and selection operator; TP, tumor  
520 purity; TME, tumor microenvironment; HCC, hepatocellular carcinoma; ROC,  
521 receiver operating characteristic; PRS, polygenic risk score.

522 **Figure 4 The validation of TP-TME risk subtypes in TCGA-LIHC data set.** **(A)**  
523 HCC with high TP-related PRS had shorter overall survival than those with low  
524 TP-related PRS. **(B)** HCC with high TME-related PRS had shorter overall survival  
525 than those with low TME-related PRS. **(C)** There are significantly different overall  
526 survivals between the three subtypes in the TP-TME risk subtypes. **(D)** There are  
527 significantly different progression free survivals between the three subtypes in the  
528 TP-TME risk subtypes. **(E)** The TP-TME risk subtype system was proved to be an  
529 independent prognostic factor, after adjusting for other clinicopathological  
530 characteristics.

531 **Abbreviation:** TP, tumor purity; TME, tumor microenvironment; PRS, polygenic  
532 risk score; AFP, alpha fetoprotein.

533 **Figure 5 The validation of TP-TME risk subtypes in LIRI-JP data set.** **(A)** HCC  
534 with high TP-related PRS had shorter overall survival than those with low TP-related  
535 PRS. **(B)** HCC with high TME-related PRS had shorter overall survival than those  
536 with low TME-related PRS. **(C)** There are significantly different overall survivals  
537 between the three subtypes in the TP-TME risk subtypes. **(D)** The TP-TME risk  
538 subtype system was proved to be an independent prognostic factor, after adjusting for  
539 other clinicopathological characteristics.

540 **Abbreviation:** TP, tumor purity; TME, tumor microenvironment; PRS, polygenic  
541 risk score; NBNC, no hepatitis B virus and no hepatitis C virus; HBV, hepatitis B  
542 virus; HCV, hepatitis C virus; AJCC, American Joint Committee on Cancer.

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

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



## Novel Molecular Classification of HCC

---

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

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

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

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

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

GSE14520

Tumor purity and microenvironment score analysis

Screening differentially expressed genes

Spearman correlation analysis

Tumor purity-related genes

Tumor microenvironment-related genes

Protein-protein interaction network

Univariate Cox regression analysis

Enrichment analysis

LASSO Cox regression analysis

Multivariate Cox and stepwise regression analysis

Malignant cell-related PRS

Tumor microenvironment-related PRS

TP-TME risk subtypes

TCGA-LIHC

ICGC-LIRI-JP

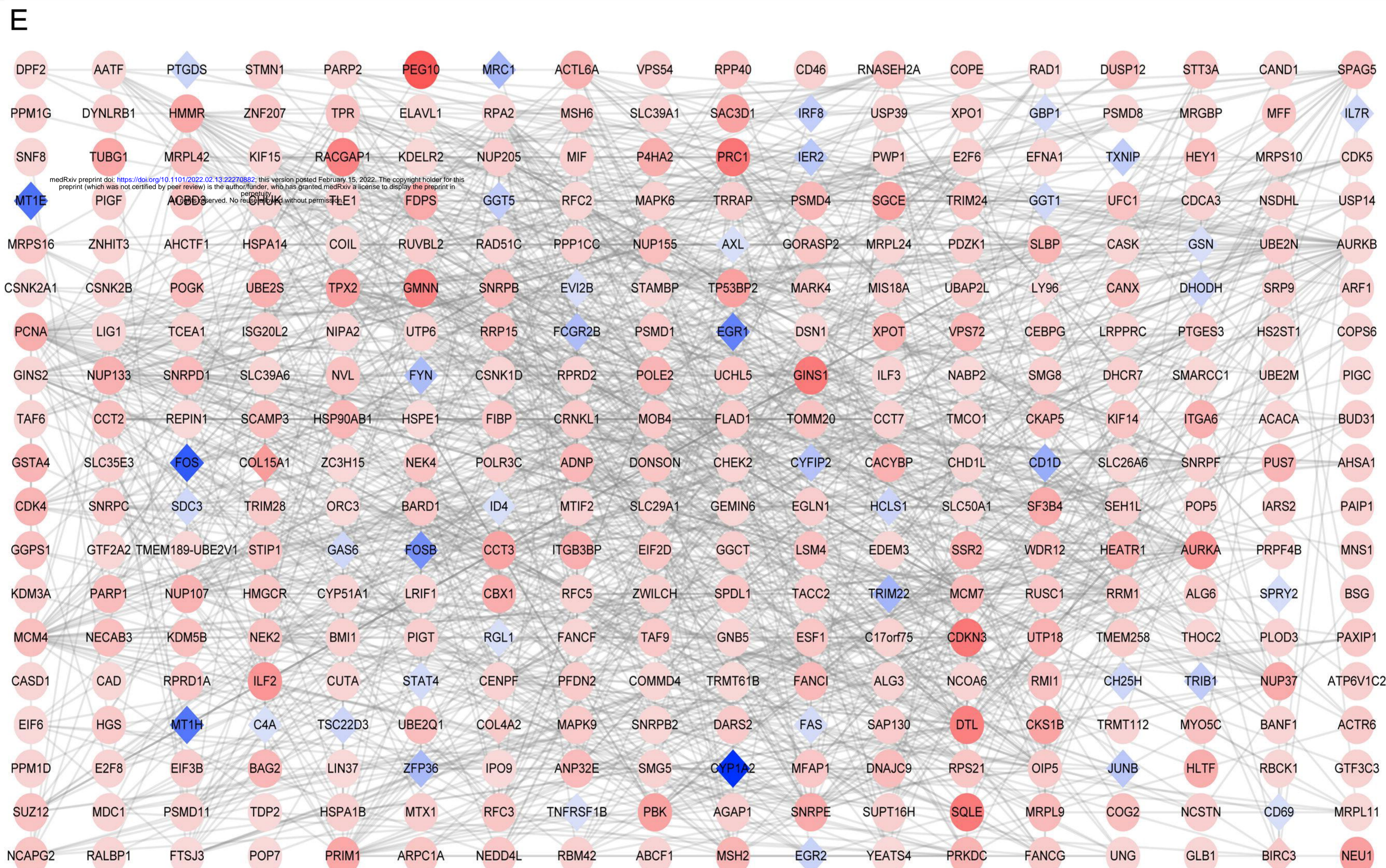
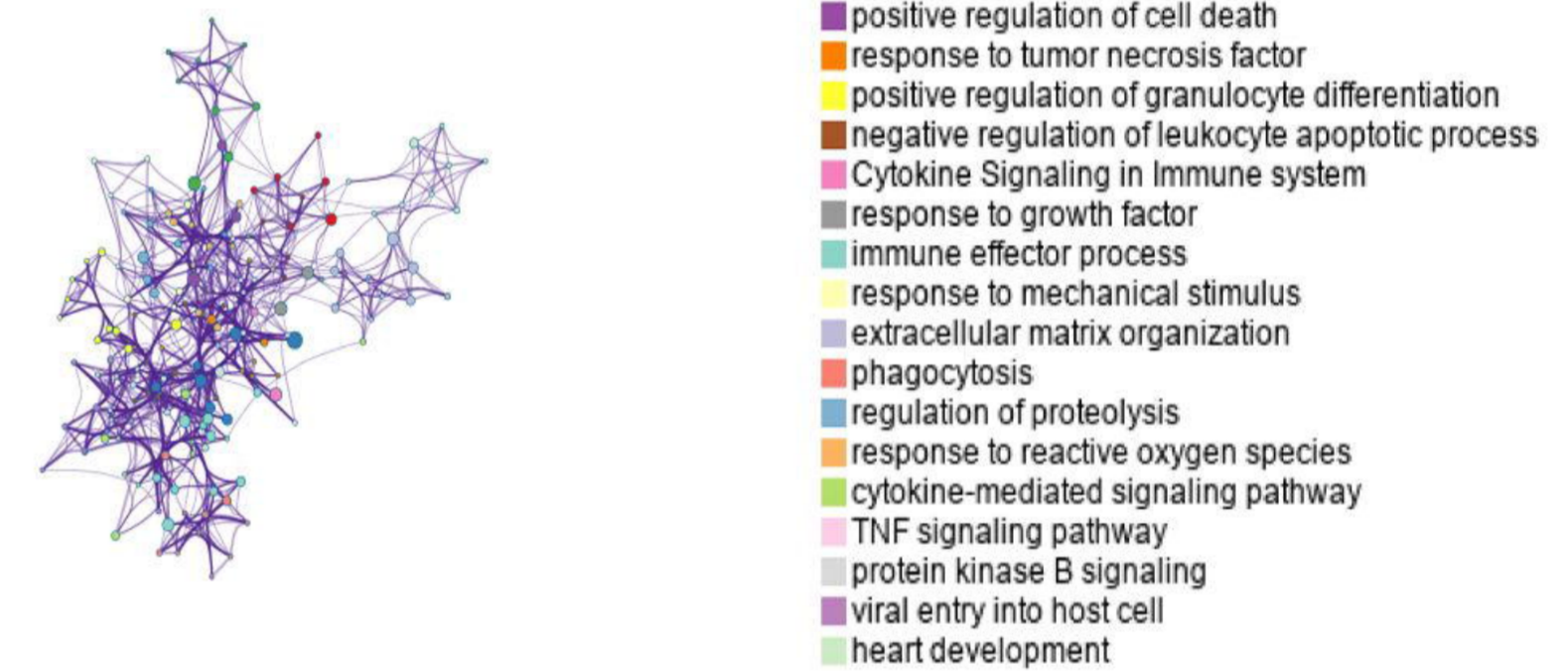
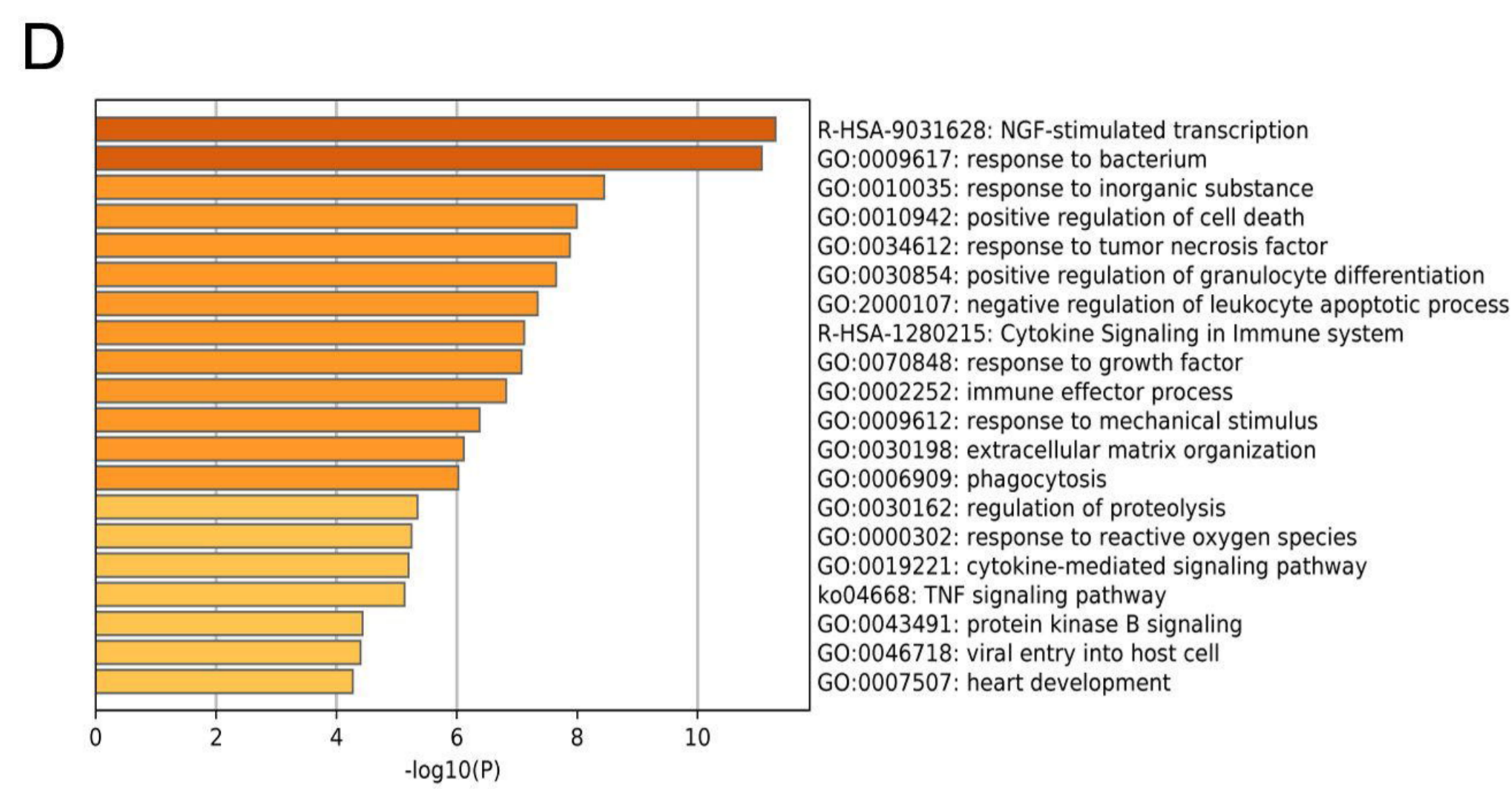
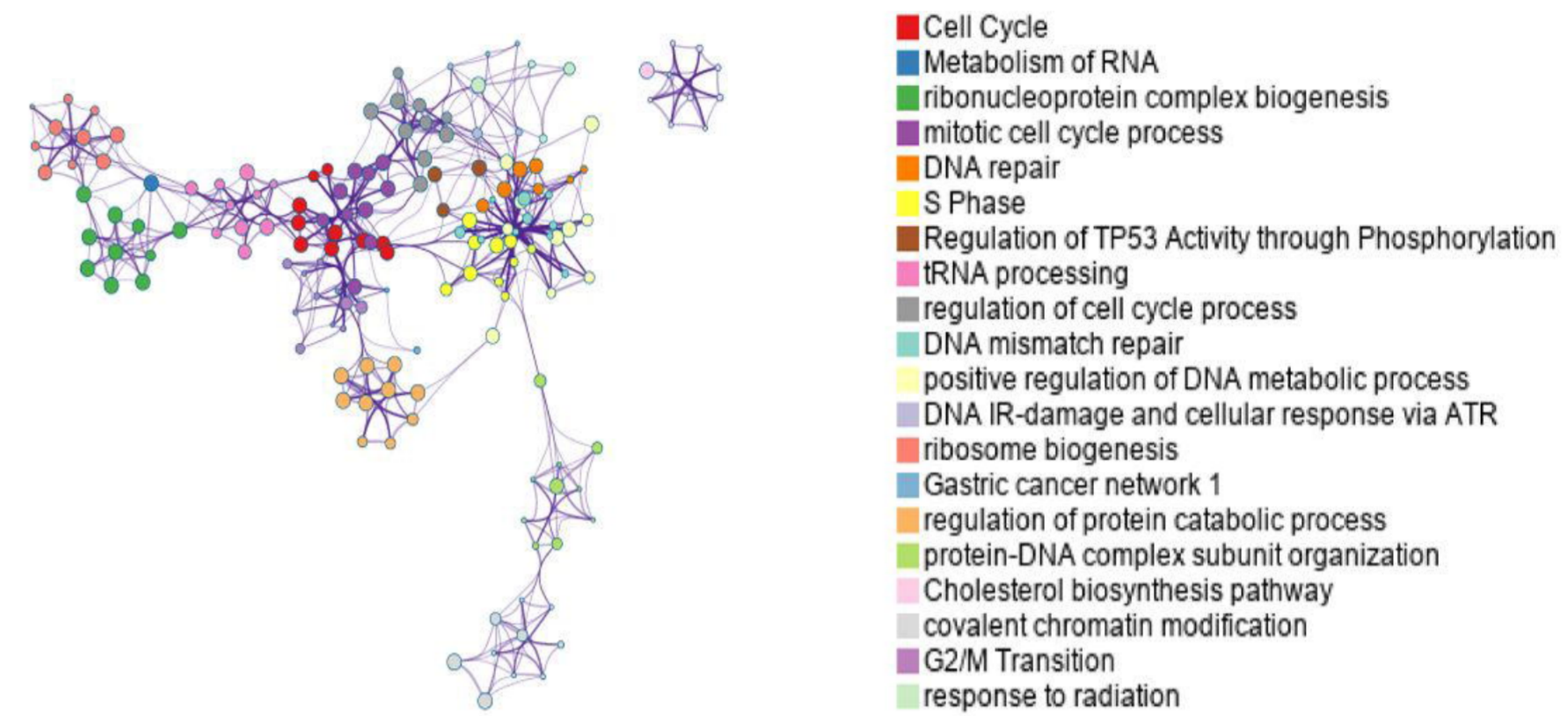
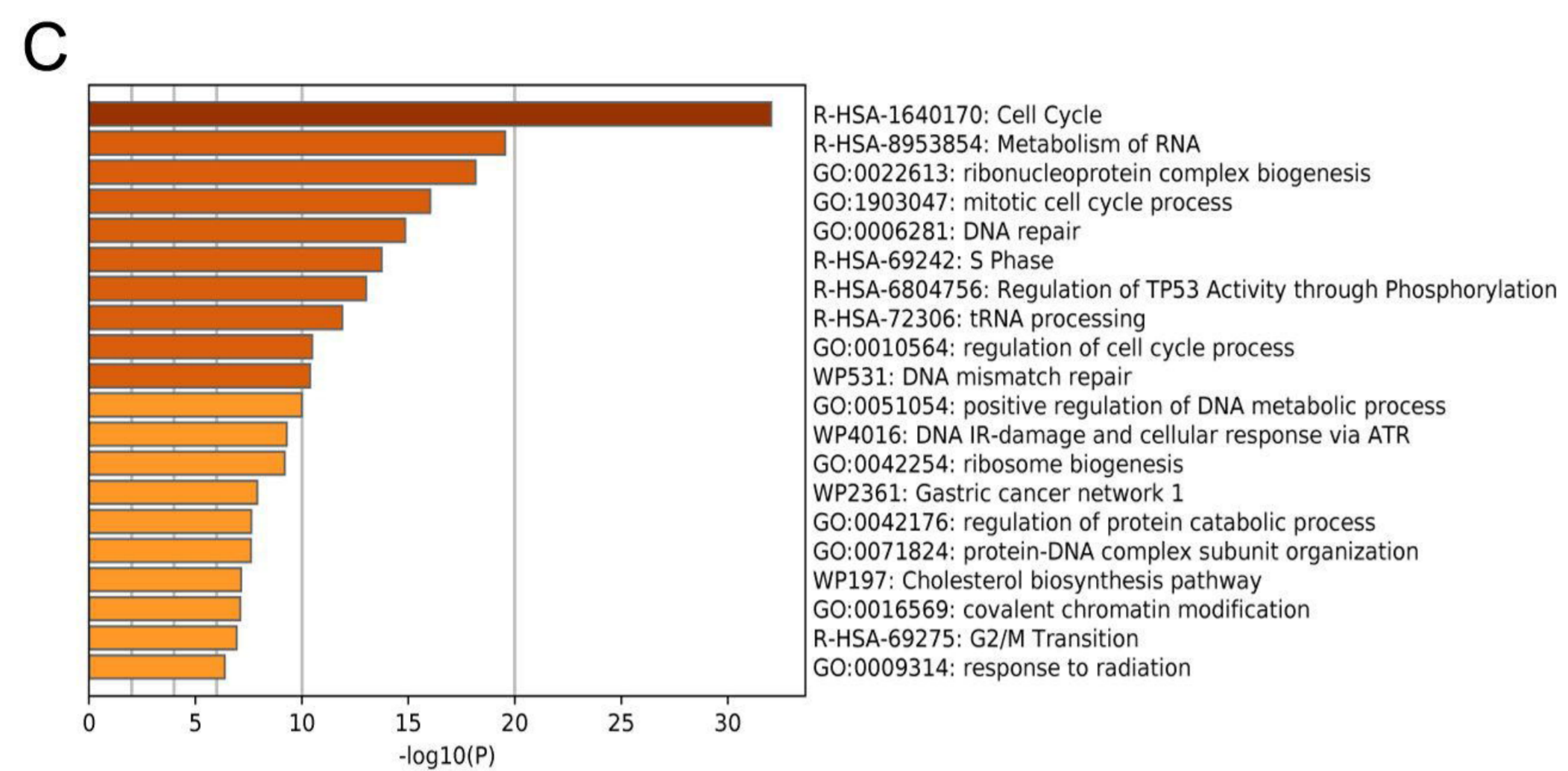
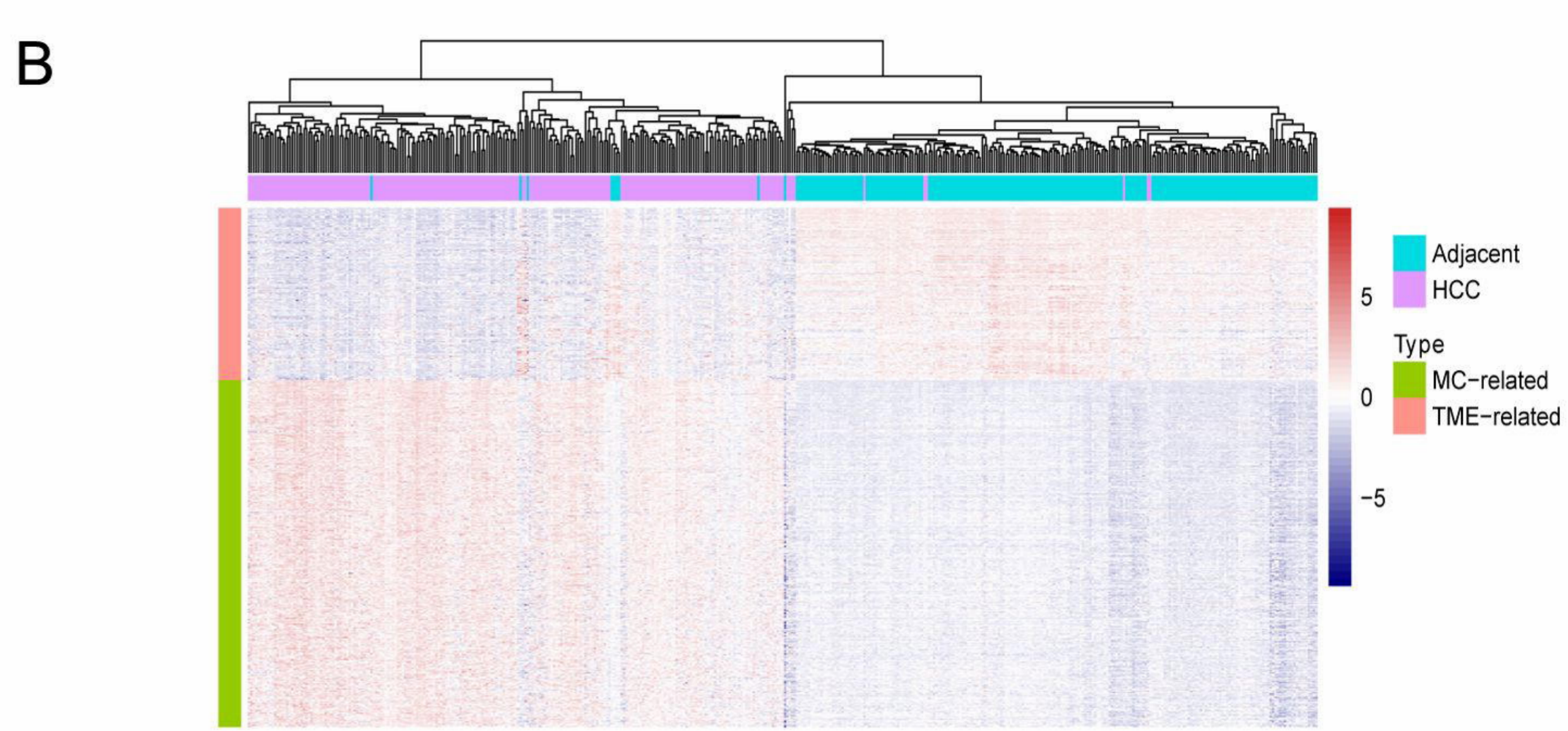
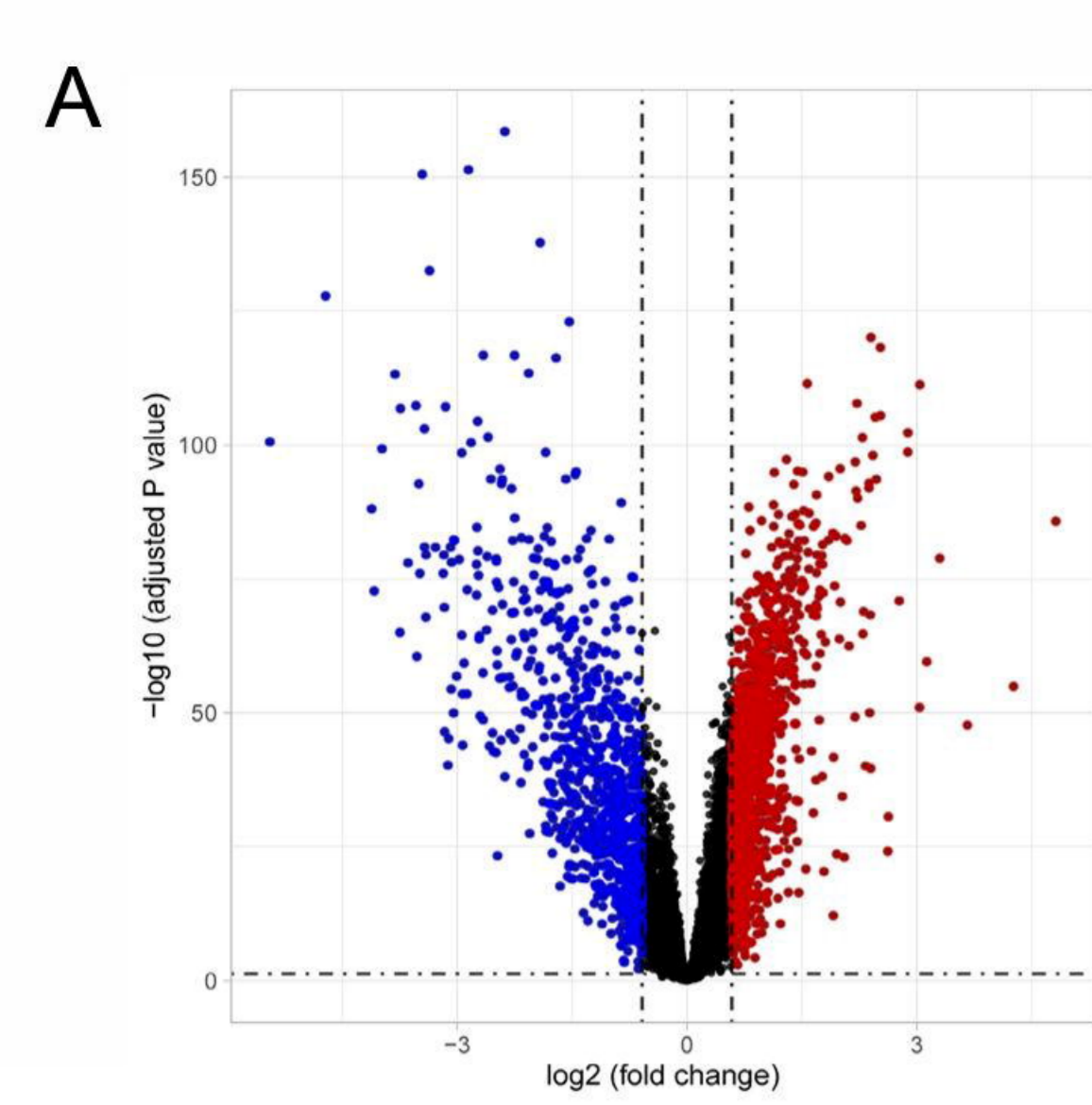
TCGA-LIHC

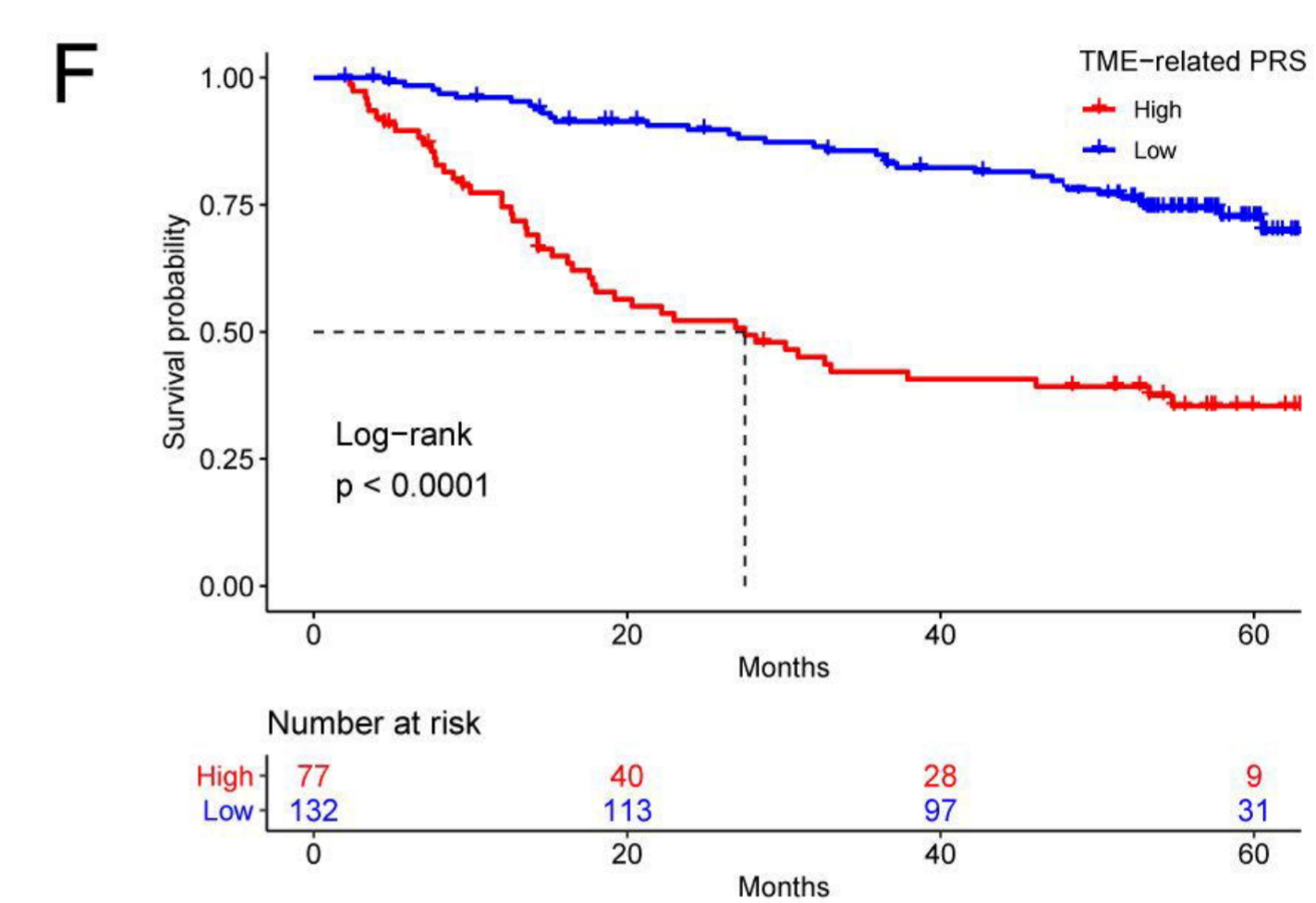
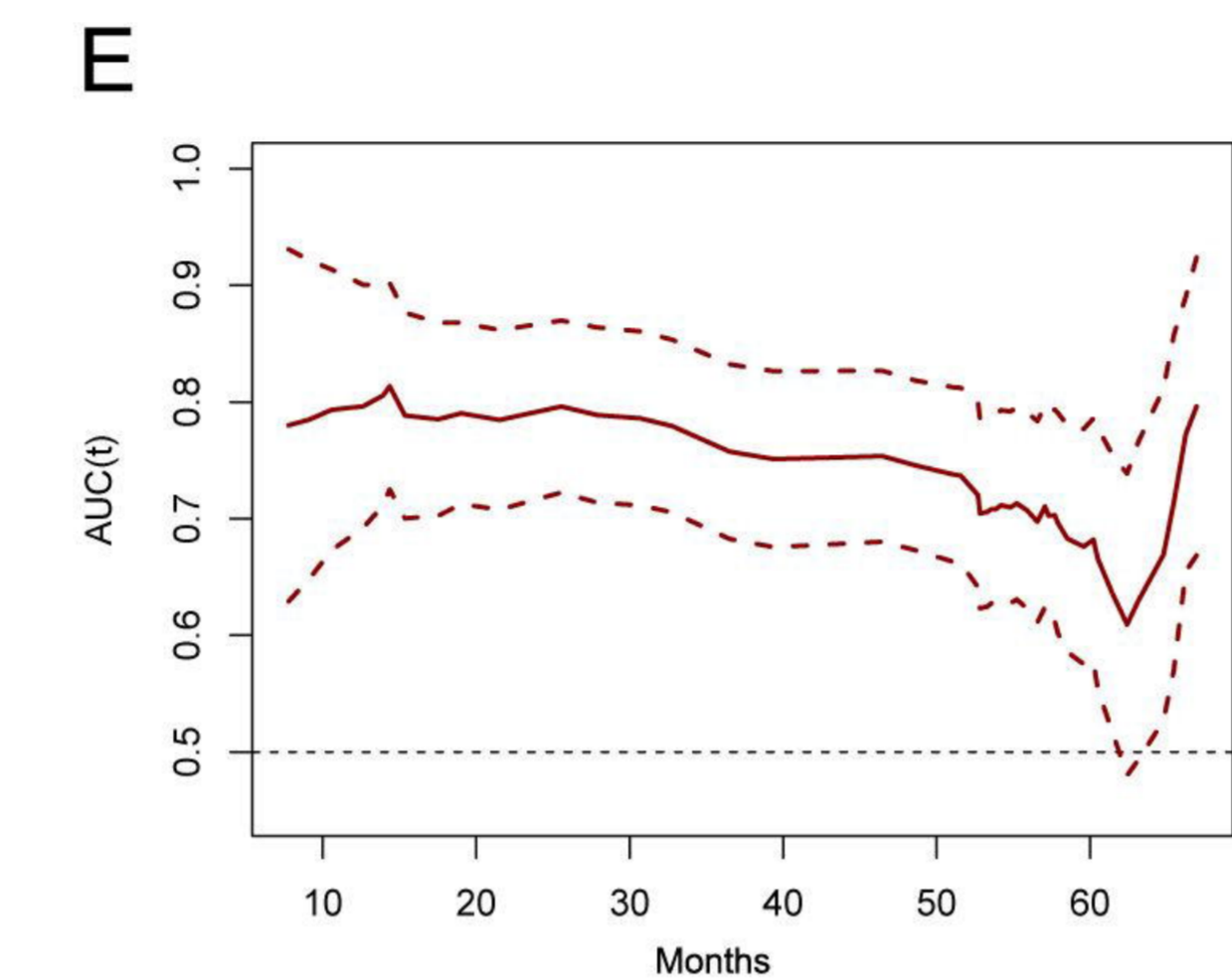
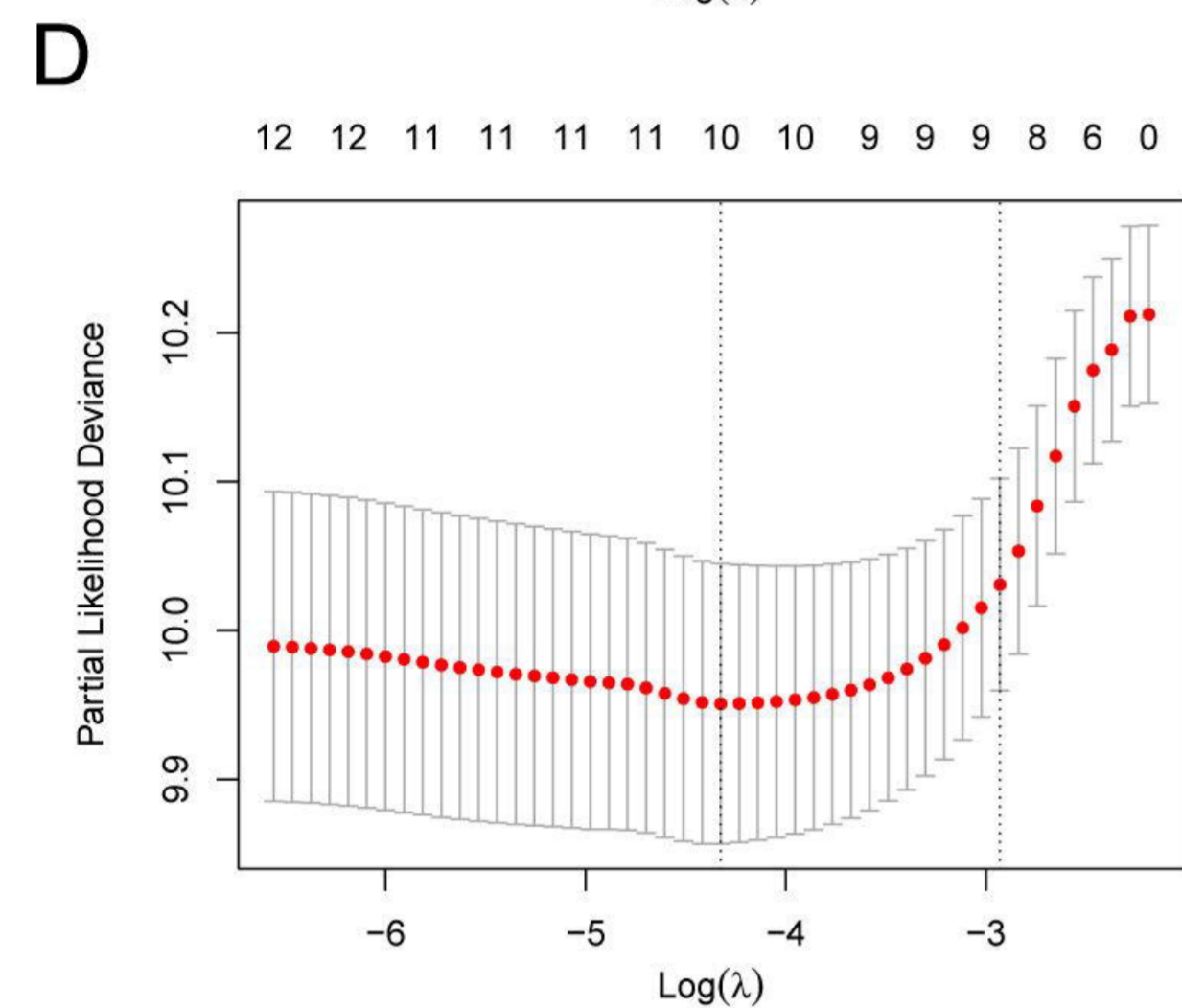
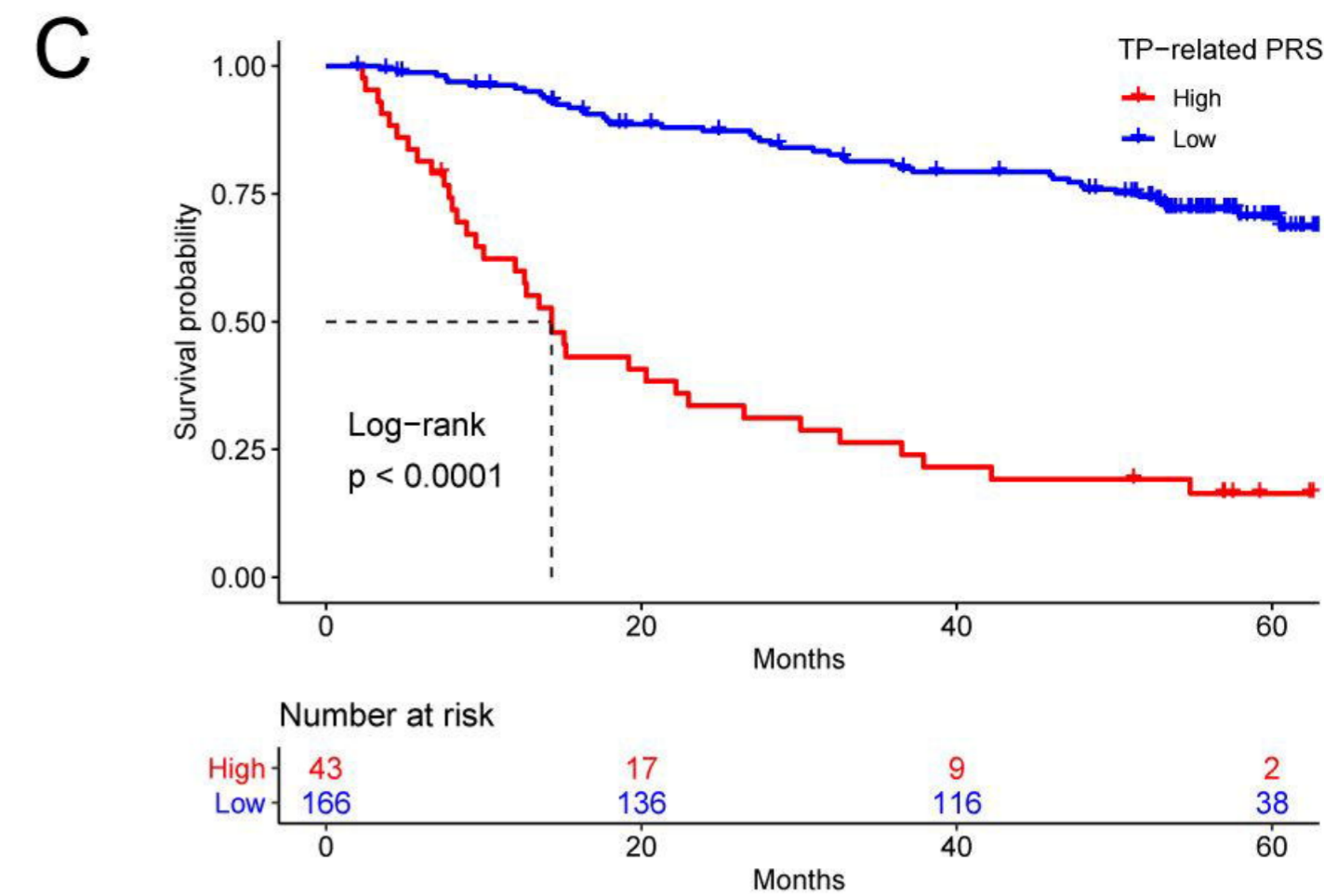
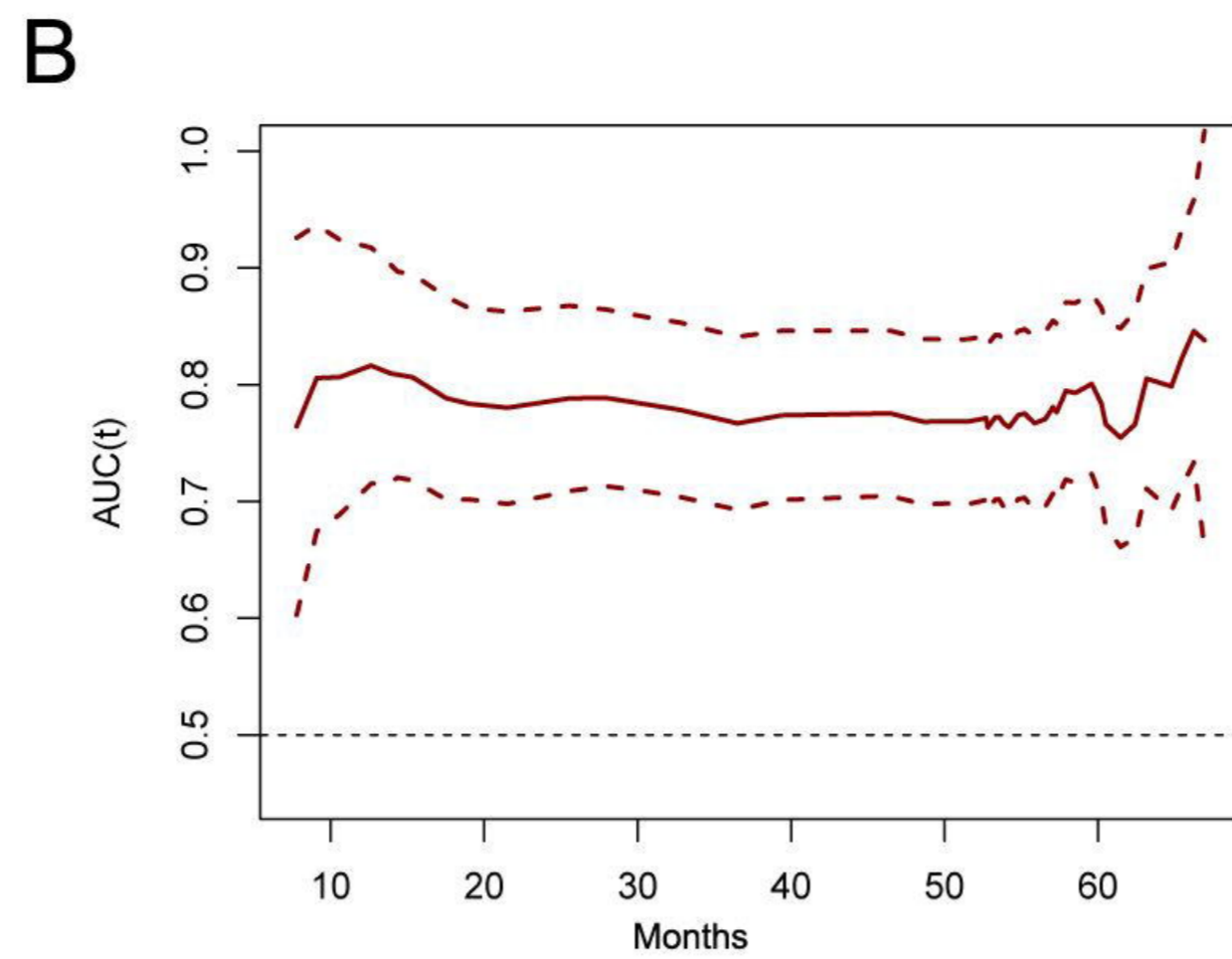
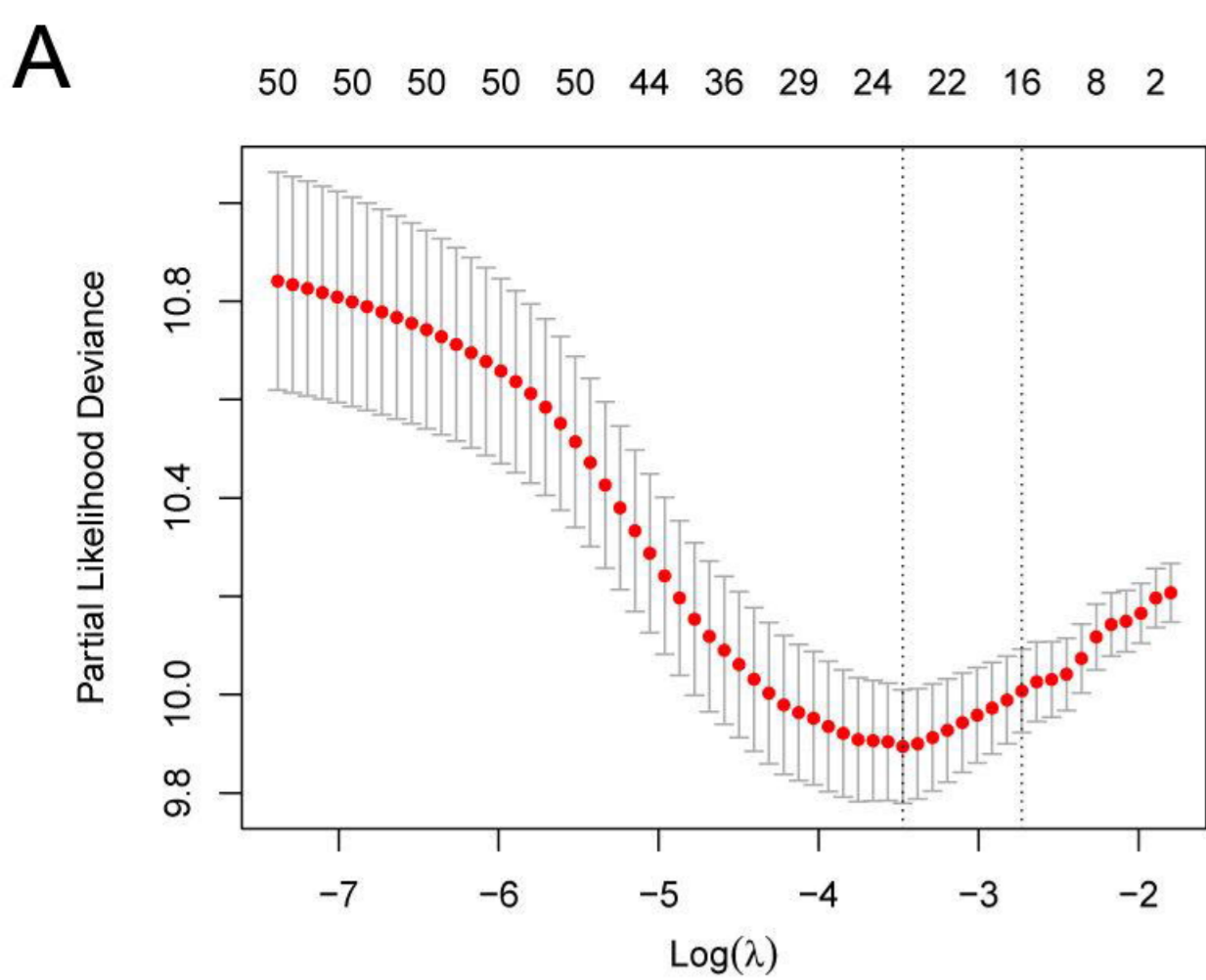
Mutation, stemness, and drug efficacy analysis

Validation

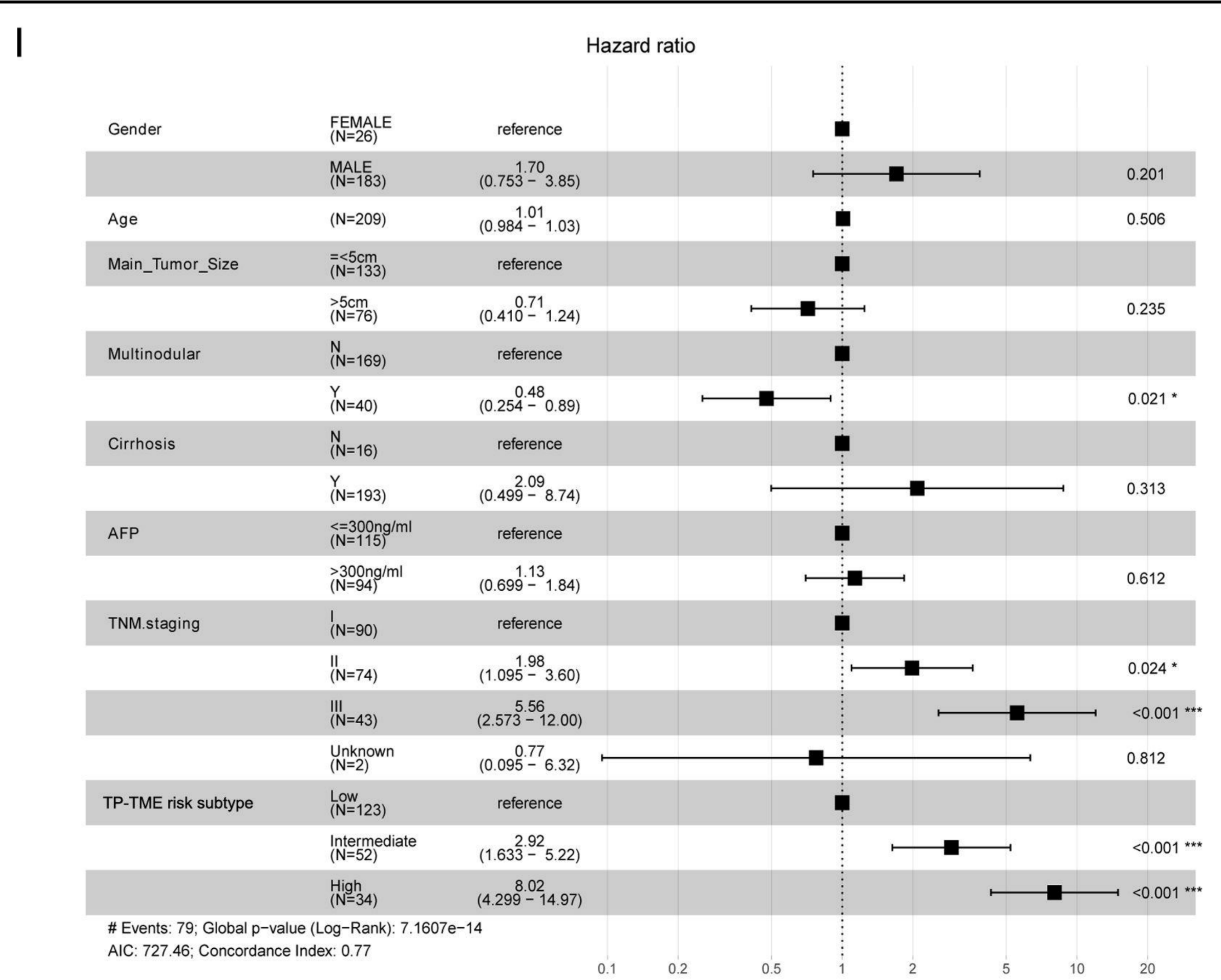
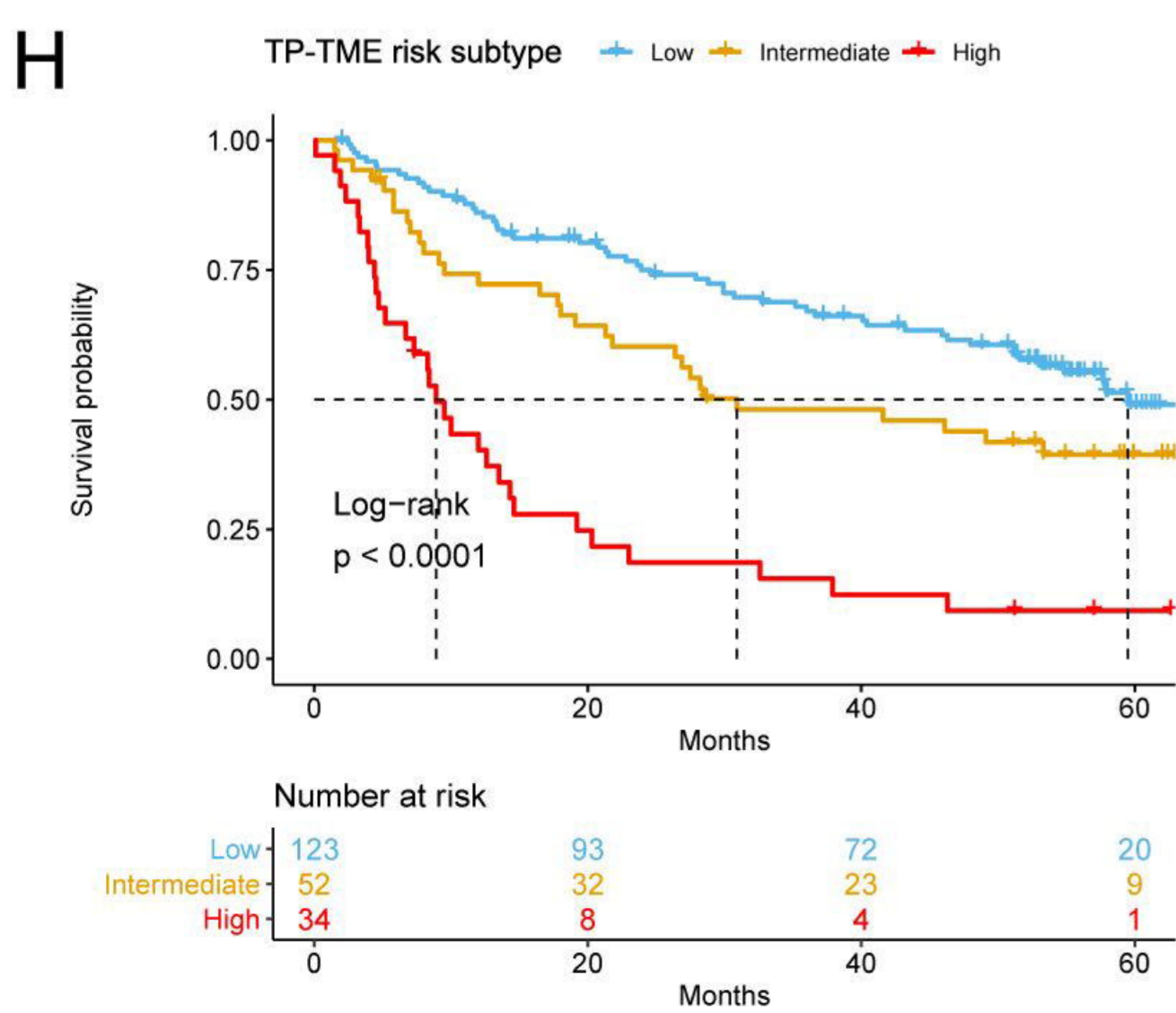
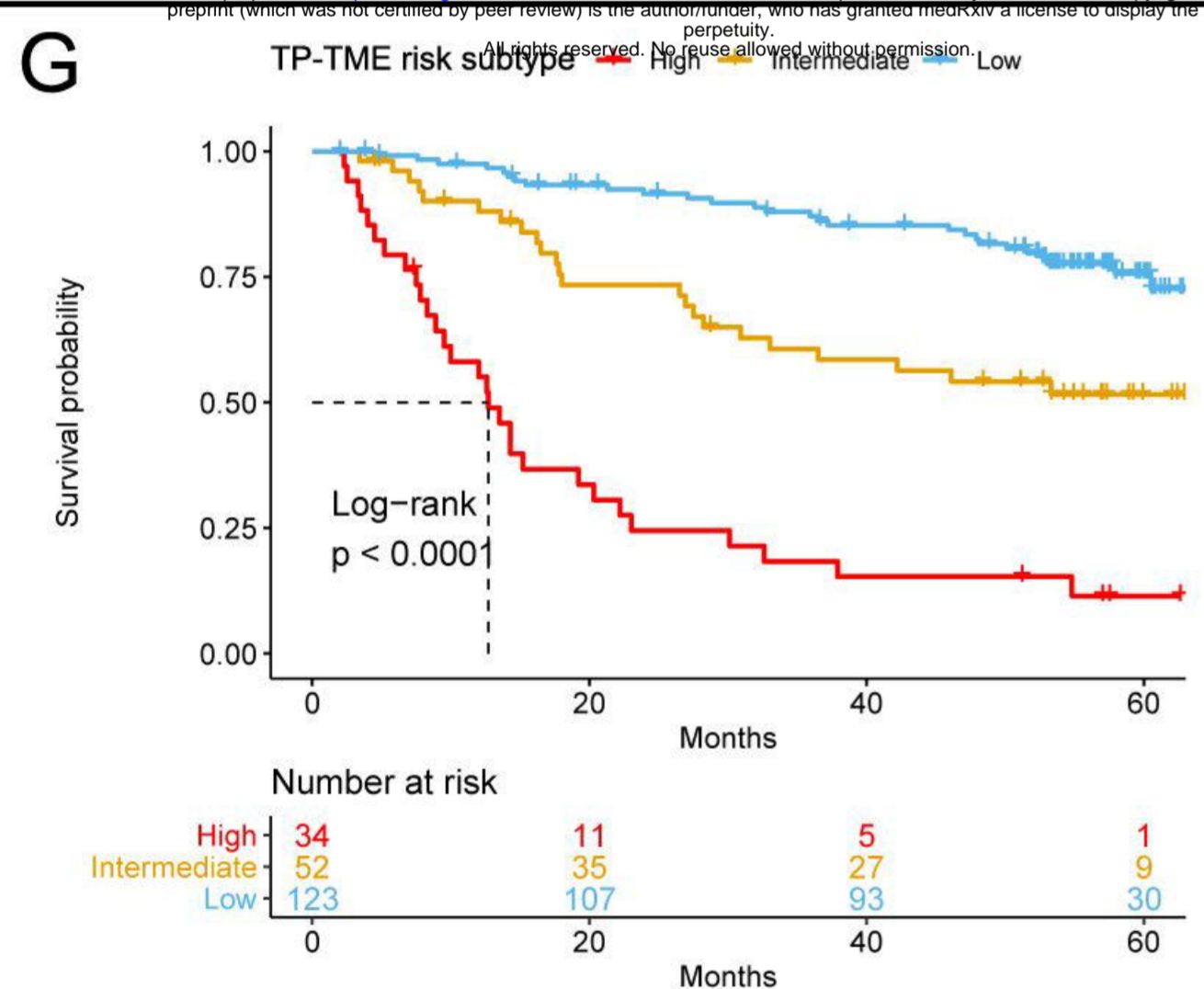
Gene set enrichment analysis

medRxiv preprint doi: <https://doi.org/10.1101/2022.02.13.22270882>; this version posted February 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

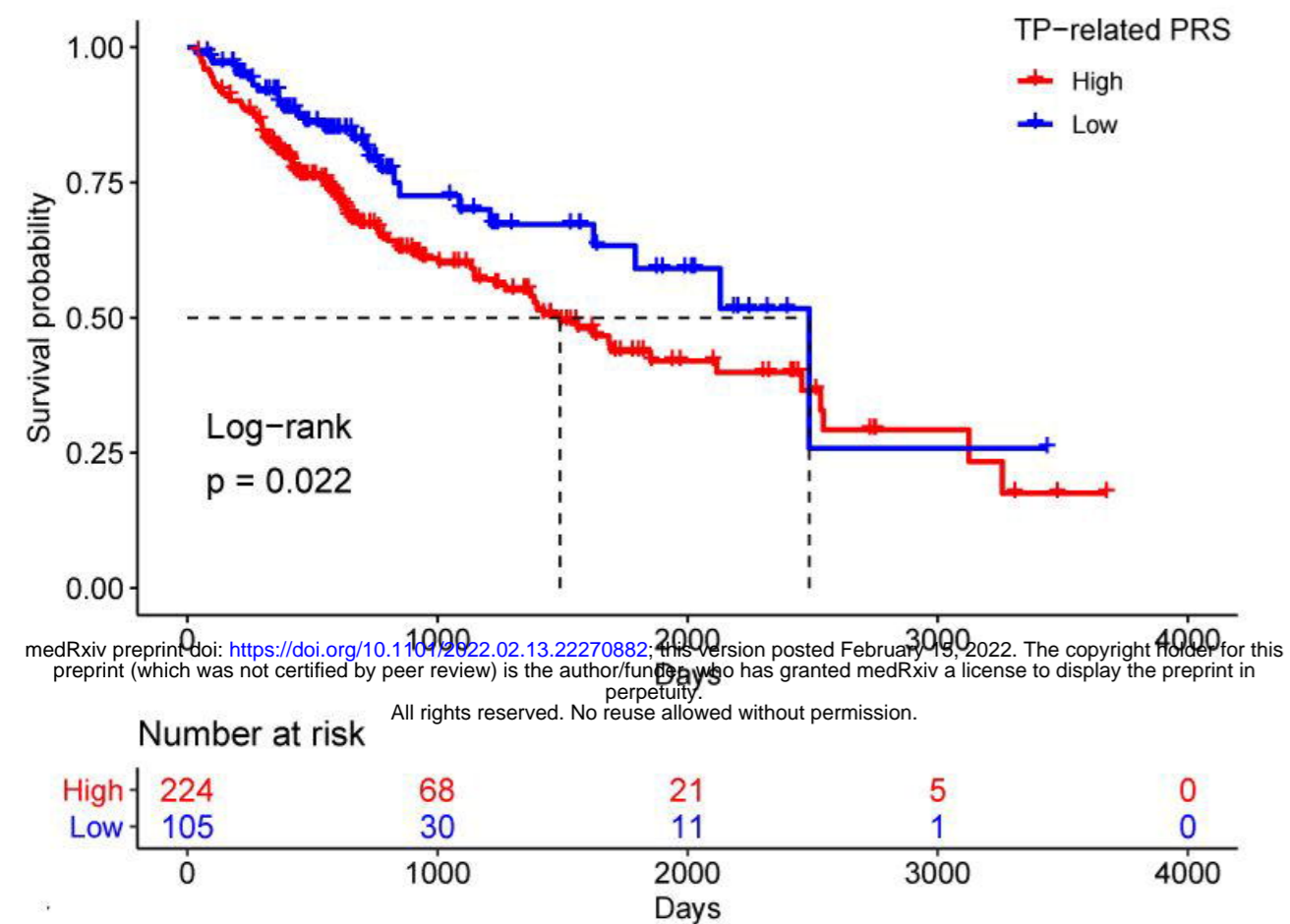




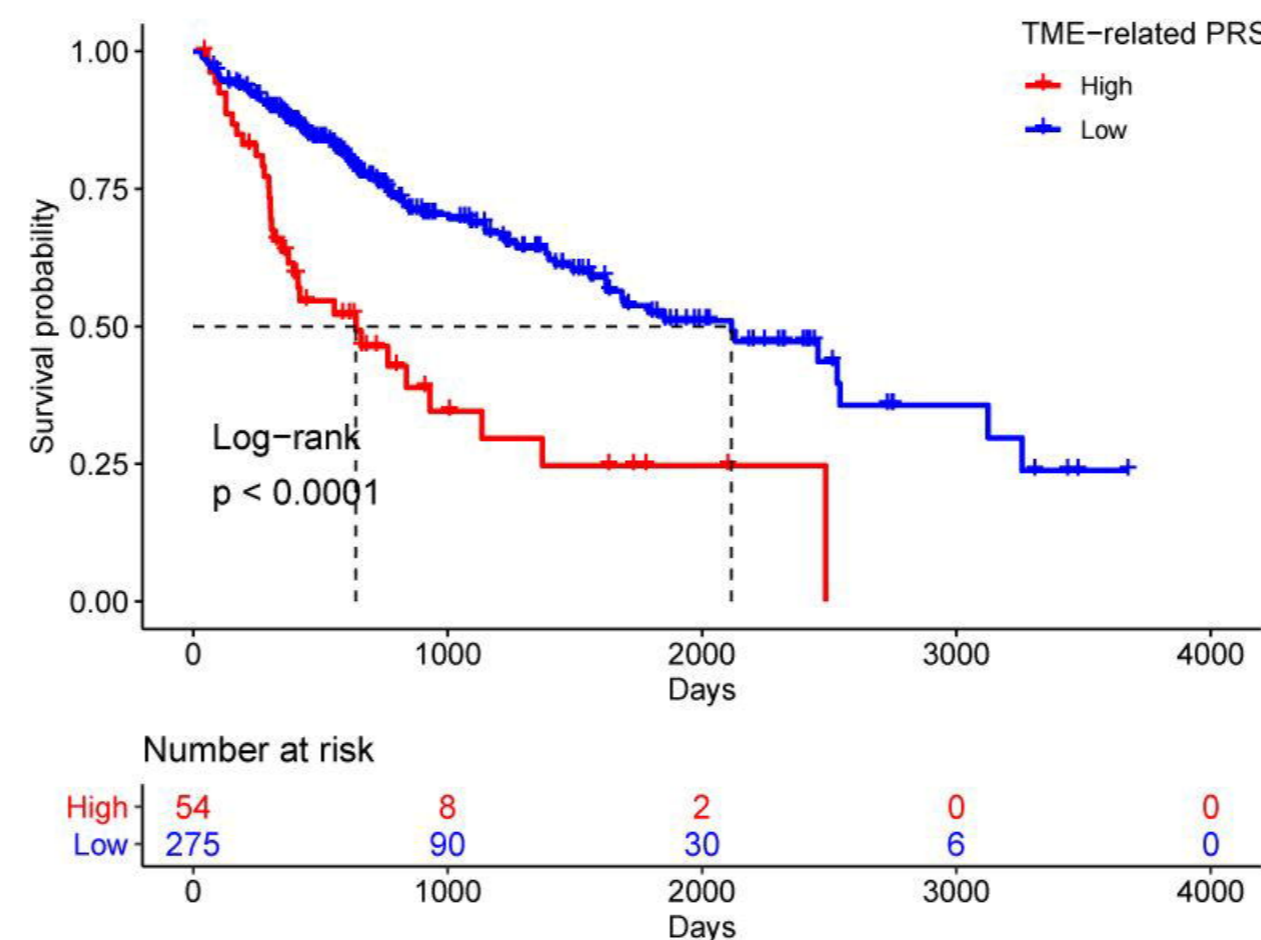
medRxiv preprint doi: <https://doi.org/10.1101/2022.02.13.22270882>; this version posted February 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.



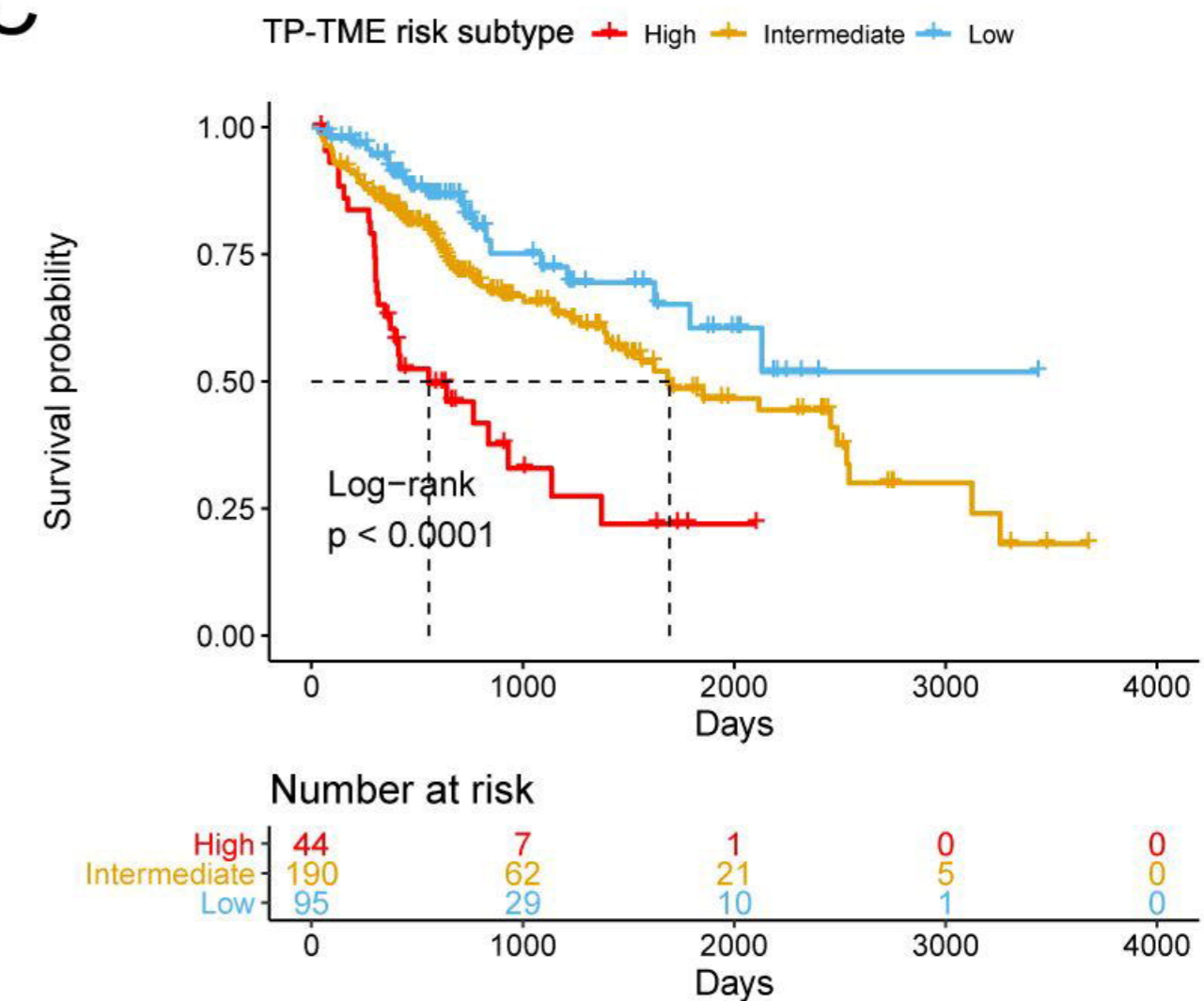
A



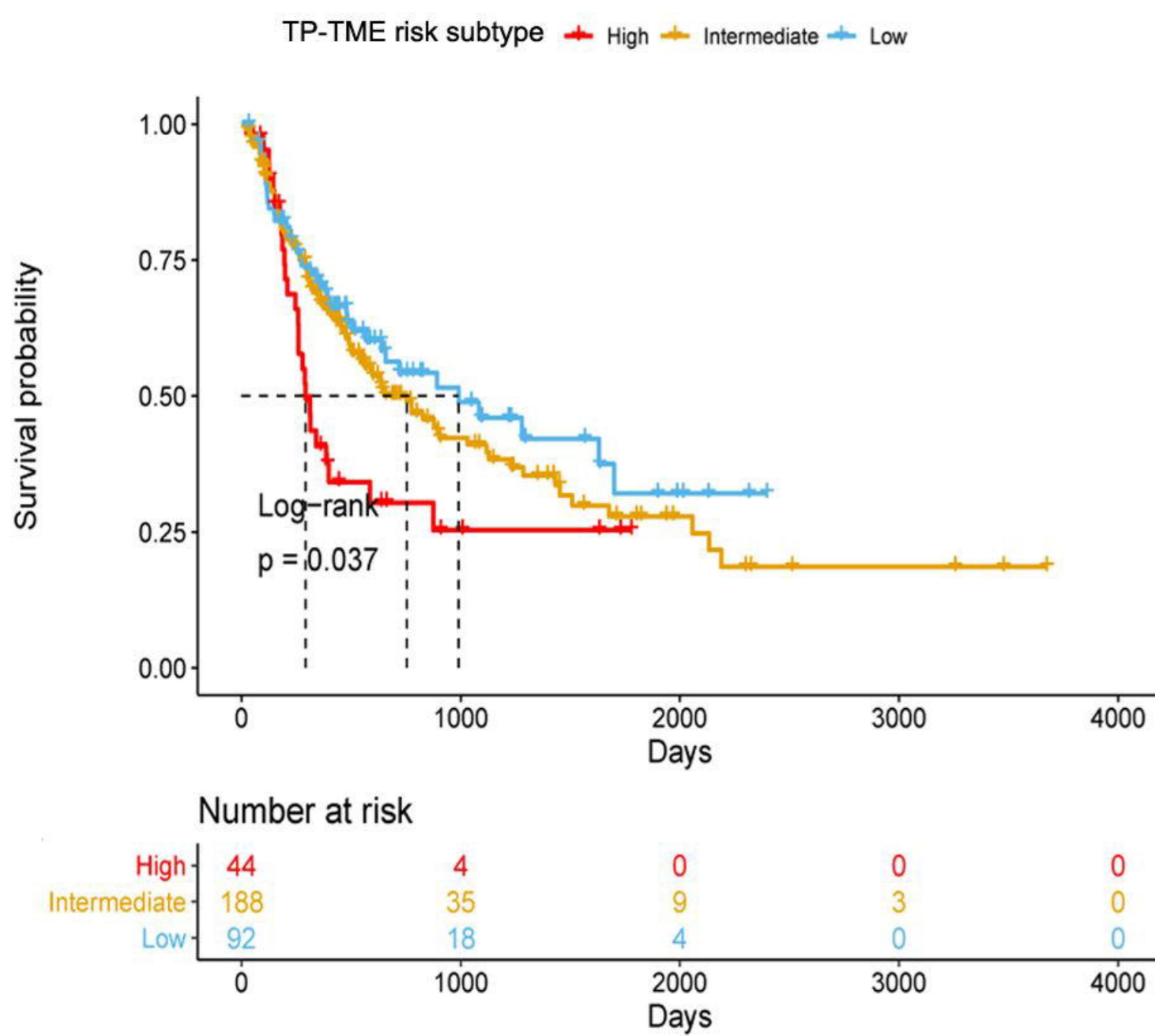
B



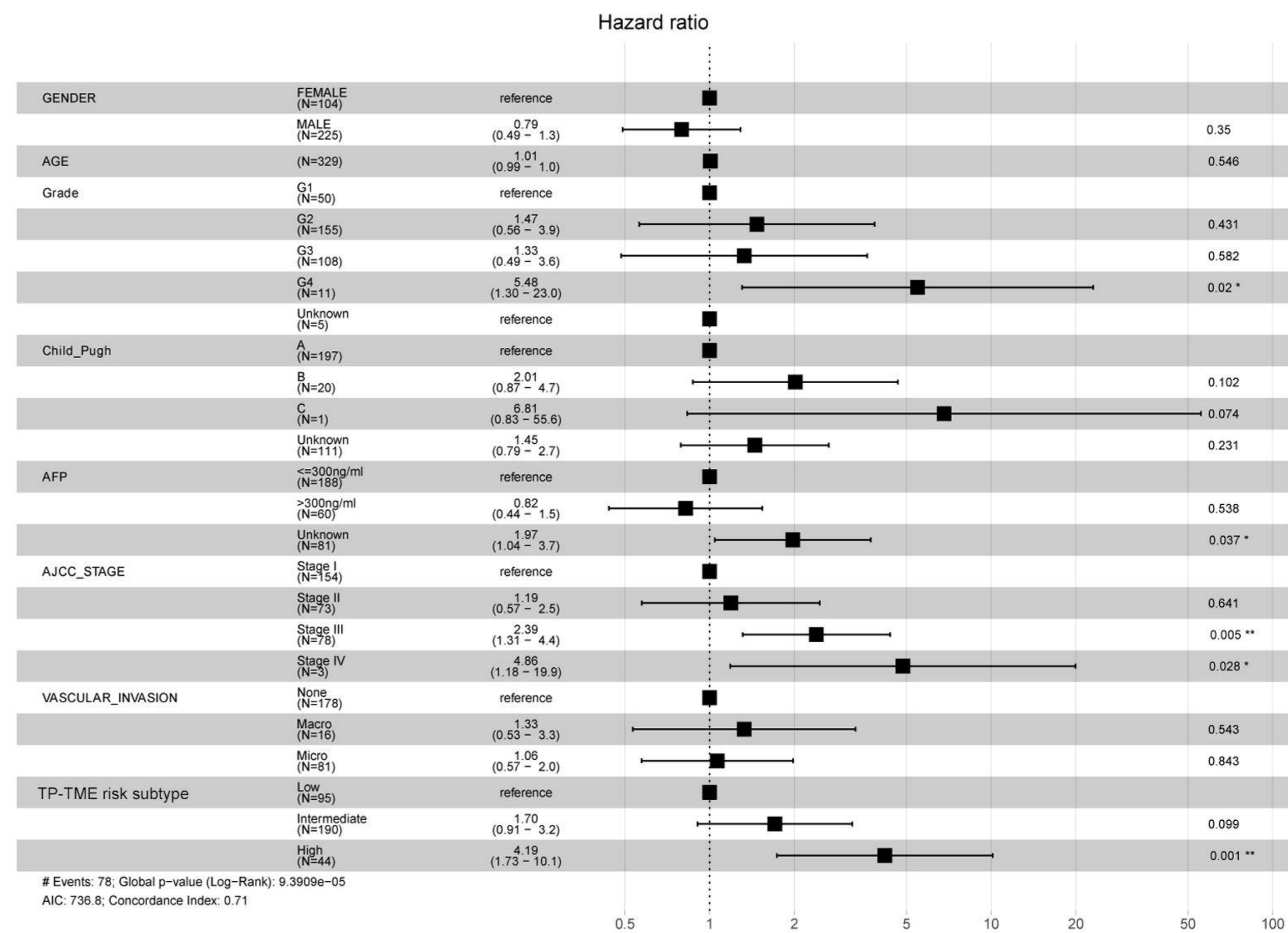
C



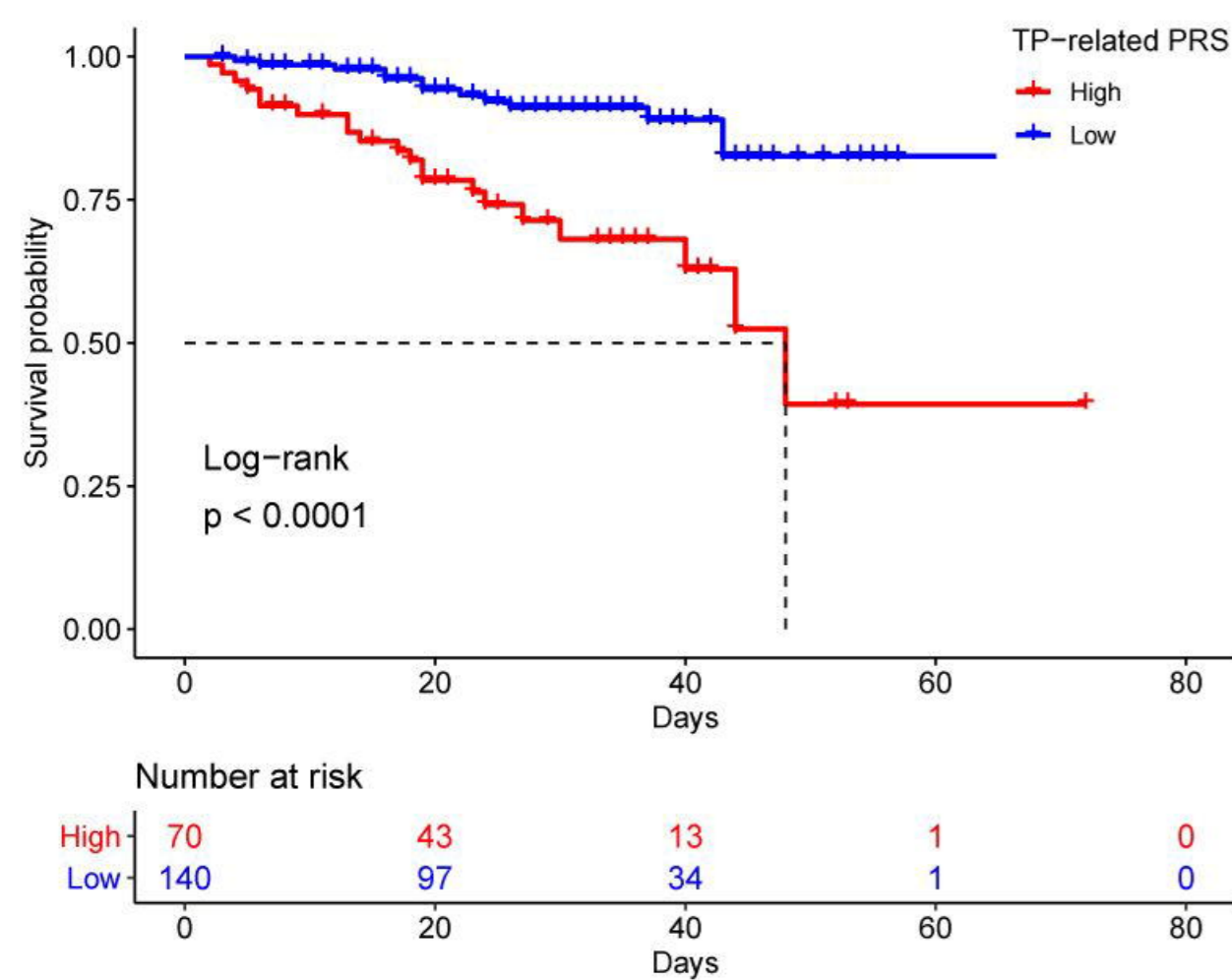
D



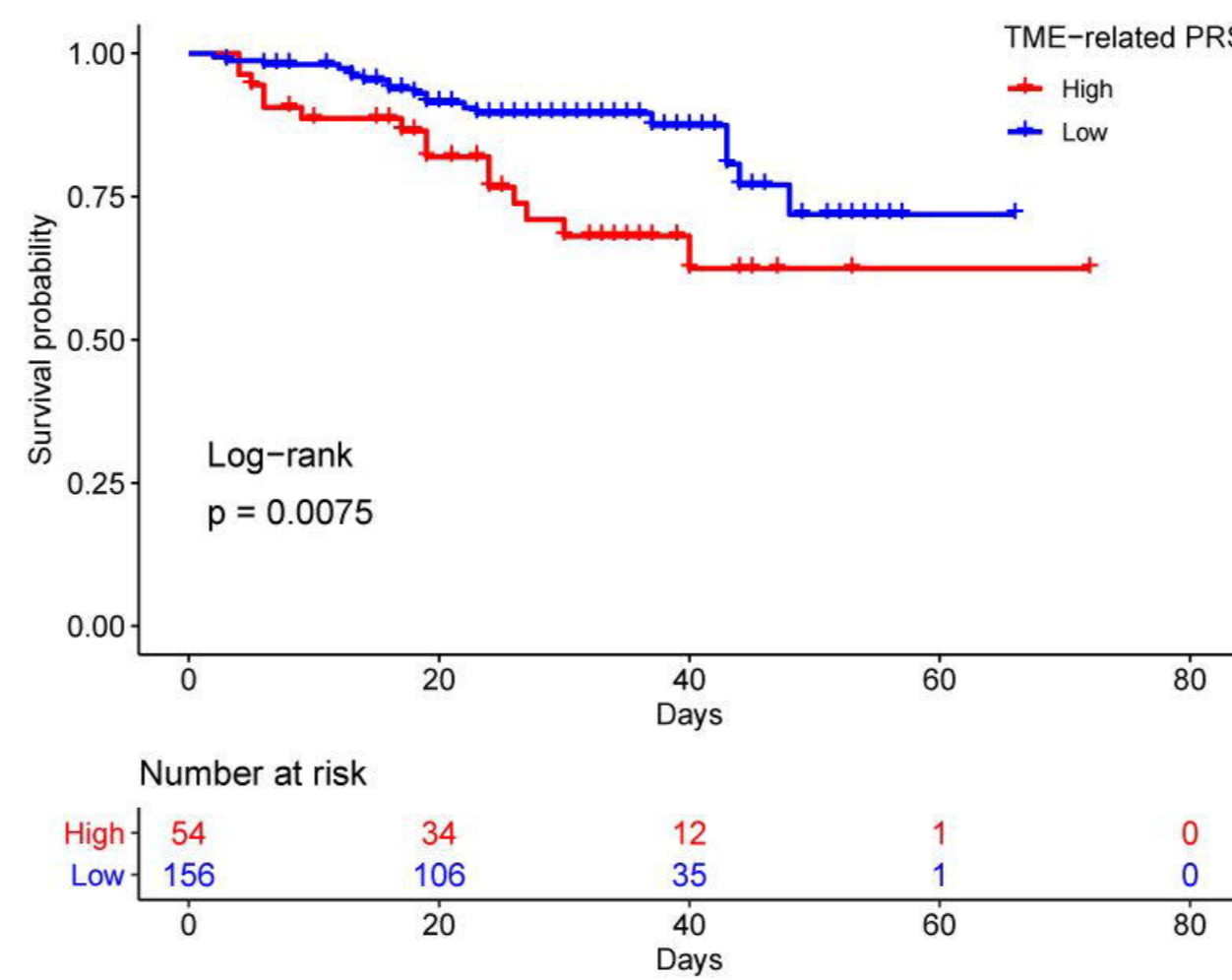
E



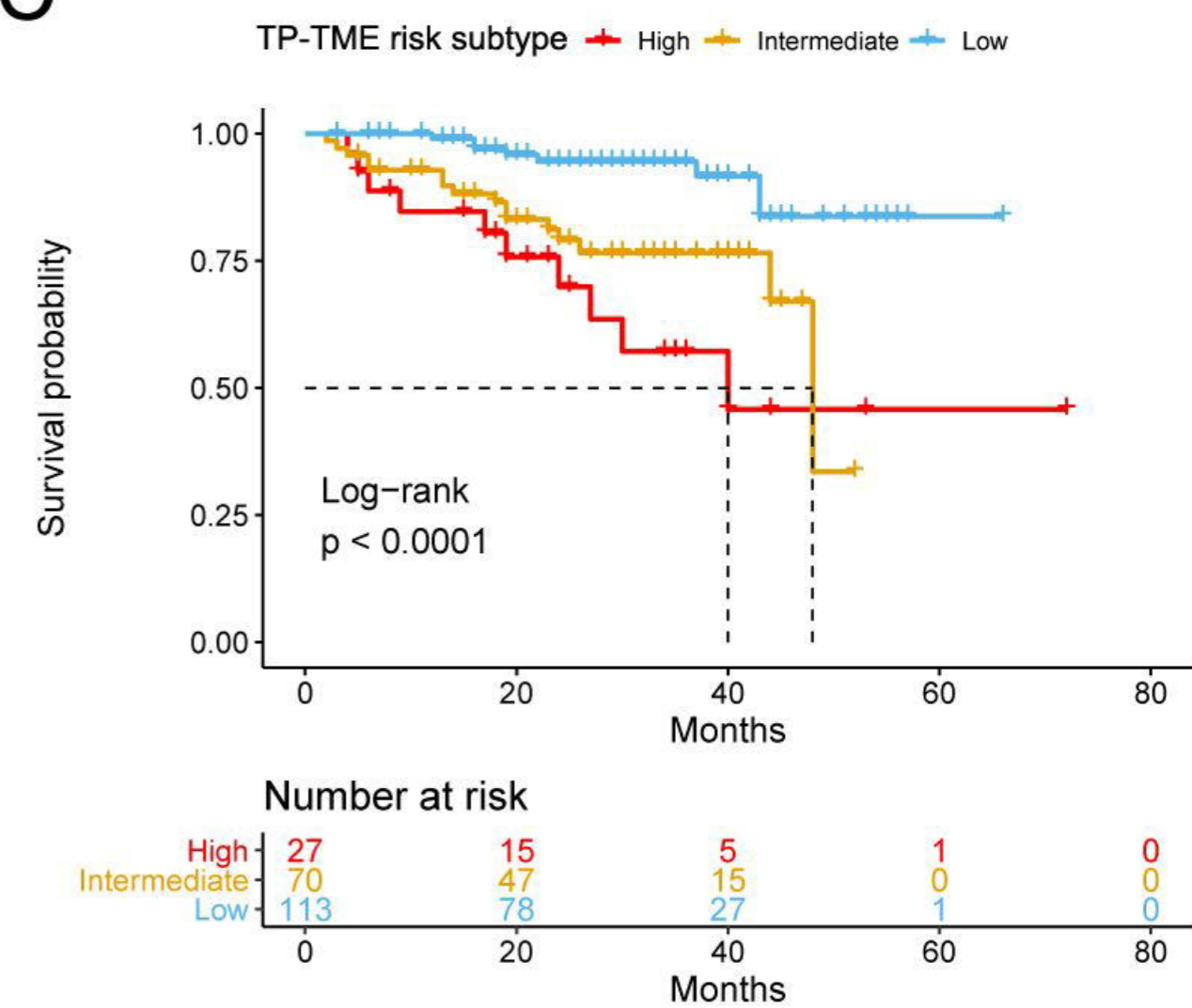
A



B



C



D

