Bioinformatics analysis and identification of hub genes and immune-related molecular mechanisms in chronic myeloid leukemia

Fangyi Yao[1], Cui Zhao[1], Fangmin Zhong[1], Tingyu Qin[1], Shuqi Li[1], Jing Liu[1], Bo Huang[1], Xiaozhong Wang[1]

[1]Jiangxi Province Key Laboratory of Laboratory Medicine, Department of Clinical Laboratory, the Second Affiliated Hospital of Nanchang University, Nanchang, Jiangxi, China.

Corresponding Author:
Xiaozhong Wang[1]
Jiangxi Province Key Laboratory of Laboratory Medicine, Department of Clinical Laboratory, the Second Affiliated Hospital of Nanchang University, Nanchang, Jiangxi 330006, China.
Email address: wangxiaozhong@ncu.edu.cn

Abstract

Background

Chronic myeloid leukemia (CML) is a malignant hyperplastic tumor of the bone marrow originating from pluripotent hematopoietic stem cells. The advent of tyrosine kinase inhibitors (TKIs) has greatly improved the survival rate of patients with CML. However, TKI-resistance leads to the disease recurrence and progression. This study aimed to identify immune-related genes (IRGs) associated with CML progression.

Methods

We extracted the gene's expression profiles from the Gene Expression Omnibus (GEO). Bioinformatics analysis was used to determine the differentially expressed IRGs of CML and normal peripheral blood mononuclear cells (PBMCs). Functional enrichment and gene set enrichment analysis (GSEA) were used to explore its potential mechanism. Hub genes were identified using Molecular Complex Detection (MCODE) and the CytoHubba plugin. The hub genes' diagnostic value was evaluated using the receiver operating characteristic (ROC). The relative proportions of infiltrating immune cells in each CML sample were evaluated using CIBERSORT. Quantitative real-time PCR (RT-qPCR) was used to validate the hub gene expression in clinical samples.

35 **Results**

36 A total of 31 differentially expressed IRGs were identified. GO analyses revealed that the

37 modules were typically enriched in the receptor ligand activity, cytokine activity, and

38 endopeptidase activity. KEGG enrichment analysis of IRGs revealed that CML involved Th17

39 cell differentiation, the NF-kappa B signaling pathway, and cytokine-cytokine receptor

40 interaction. A total of 10 hub genes were selected using the protein-protein interaction (PPI)

41 network. GSEA showed that these hub genes were related to the gamma-interferon immune

42 response, inflammatory response, and allograft rejection. ROC curve analysis suggested that six

43 hub genes may be potential biomarkers for CML diagnosis. Further analysis indicated that

44 immune cells were associated with the pathogenesis of CML. The RT-qPCR results showed that

45 proteinase 3 (PRTN3), cathepsin G (CTSG), matrix metalloproteinase 9 (MMP9), resistin

46 (RETN), eosinophil derived neurotoxin (RNase2), eosinophil cationic protein (ECP, RNase3)

47 were significantly elevated in CML patients' PBMCs compared with healthy controls.

48 **Conclusions**

49 These results improved our understanding of the functional characteristics and immune-related

50 molecular mechanisms involved in CML progression and provided potential diagnostic

51 biomarkers and therapeutic targets.

52

53 # Introduction

54 Chronic myeloid leukemia (CML) is a malignant, clonal, and proliferative disease originating

55 from hematopoietic stem cells (Nash 1999). CML is characterized by the presence of a BCR-

56 ABL1 fusion gene. This disease is classified into a chronic phase (CP), accelerated phase (AP),

57 and a blast phase (BP). Tyrosine kinase inhibitors (TKIs), especially Imatinib (IM), have been

58 used to treat CML with favorable outcomes (Lugo et al. 1990). However, TKI use may lead to

59 side effects and drug resistance. This in turn may decrease TKI efficacy in some patients

60 (Hehlmann 2012). Thus, it is important to understand the mechanisms underlying CML and

61 determine potential biomarkers.

62   Studies have proposed that gene expression-based characteristics may be used in CML

63 diagnosis and prognosis (Vinhas et al. 2017), but they have not been made a part of routine

64 clinical care due to insufficient validation cohorts. A number of new biomarkers have emerged

65 following the validation of large-scale gene expression datasets (Feng et al. 2020). The

66   establishment of multiple gene expression profiles makes it possible to identify reliable CML
67   biomarkers. New evidence has suggested that the immune system, consisting of immune-related
68   genes (IRGs) and tumor-infiltrating immune cells, is vital during cancer initiation and
69   progression (Zou & Hu 2020). A number of recent studies have constructed prognostic immune
70   signals for use in the diagnosis and prognosis of a variety of cancers (Geng et al. 2021). Hu *et al*.
71   (2021) constructed a seven-IRG signature using data from the Gene Expression Omnibus (GEO)
72   database to predict glioblastoma multiform prognosis. This signature also revealed the potential
73   functions of these genes in glioblastoma occurrence and development (Hu *et al*. 2021). Zhu *et al*.
74   (2020) predicted the survival rate of patients with acute myeloid leukemia with a normal
75   karyotype by constructing a six-IRG signature to determine their prognostic characteristics.
76   However, the clinical role of IRGs in CML has yet to be determined.

77   In this study, we identified IRGs using the data from the GEO database and performed
78   network analyses to assess the molecular mechanisms underlying CML progression. Further, the
79   biological processes involved were analyzed using gene ontology (GO) enrichment and kyoto
80   encyclopedia of genes and genomes (KEGG) pathways for the differentially expressed genes.
81   Moreover, the top 10 hub genes screened via protein-protein interaction (PPI) network were
82   selected for their functional similarity, and their diagnostic value was assessed. Our results
83   established the first comprehensive network of IRGs and provided a useful framework for
84   explaining CML's molecular mechanisms at the systems biology level.

85

Materials & Methods

Data collection

88   The gene expression dataset GSE100026 (Li et al. 2020) was downloaded from the GEO
89   database (GPL18573 Illumina NextSeq 500) and annotated using R software equipped with
90   annotation files. This dataset included five CML chronic-phase (CML-CP) peripheral blood
91   mononuclear cell (PBMCs) samples, five CML blast-period (CML-BP) PBMCs samples, and
92   five normal PBMCs samples. We acquired 2,498 genes from the ImmPort database
93   (https://immport.niaid.nih.gov) (Bhattacharya et al. 2018). Hallmark gene sets were extracted
94   from the MSigDB database (http://software.broadinstitute.org/gsea/msigdb/genesets.jsp). The
95   overview of the workflow is shown in Fig. 1.

Identifying IRGs

97 We converted the RNA sequencing (RNA-seq) data from RNA-Seq by Expectation
98 Maximization (RSEM) into transcripts per kilobase million (TPM) expression profiles. Data
99 analysis was performed using the limma package in R (Ritchie et al. 2015). The cutoffs were set
100 as |log FC| > 1, P < 0.05, and the false discovery rate (FDR) < 0.05 to screen for differentially
101 expressed genes. The intersection of the differentially expressed genes and the immune genes
102 was used to find the IRGs. These results were displayed on a volcano map and a heat map (Xue
103 et al. 2020).

104 **PPI network construction and module analysis**

105 We used the Metascape database (https://metascape.org/gp/index.html#/main/step1) to perform
106 GO and KEGG pathway enrichment analyses. We identified the most enriched biological
107 pathways and functions related to the IRGs (Zhou et al. 2019). We then constructed a protein-
108 protein interaction (PPI) network of IRGs using the Search Tool for the Retrieval of Interacting
109 Genes (STRING) online database (http://string-db.org, version 10). The network was visualized
110 in Cytoscape (version 3.6.1) software (Szklarczyk et al. 2017). We used the Molecular Complex
111 Detection (MCODE) and cytoHubba plug-ins to identify the merged network's hub genes. These
112 genes serve as diagnostic markers (Chin et al. 2014).

113 **Verifying diagnostic markers**

114 We performed receiver operating characteristic (ROC) curve analysis on each screened
115 diagnostic marker in the GSE100026 dataset to verify its accuracy. R's pROC package was used
116 for ROC curve analysis. Regarding the interpretation of AUC results, a test with an area > 0.9
117 indicates high accuracy, 0.7-0.9 as moderate accuracy, 0.5-0.7 as low accuracy, and 0.5 as a
118 chance result (Akobeng 2007). Genes with an AUC value greater than 0.7 were retained.

119 **Gene set enrichment analysis (GSEA)**

120 GSEA is a statistical method used to determine whether genes from a particular pathway or other
121 predefined gene sets are differentially expressed in different phenotypes. Reactome pathways
122 were analyzed with GSEA using clusterProfiler to define each functional cluster.
123 C2.all.v6.2.symbols.gmt was selected as the reference gene set. A false discovery rate <0.1 and
124 P-value < 0.01 were set as the cut-off criteria.

125 **Estimating immune cell type fractions**

126 CIBERSORT is a bioinformatics method used to evaluate the composition of immune cells
127 through standardized gene expression transformation (Chen et al. 2018). The CIBERSORT

128   algorithm (CIBERSORT R script v1.03; http://cibersort.stanford.edu/) was used to analyze the

129   immune infiltration of the CML-CP-Normal, CML-BP-Normal, and CML-CP-CML-BP datasets,

130   respectively. This analysis helped us assess differences in immune cell infiltration among CML-

131   CP, CML-BP, and healthy samples.

132   **Analysis of the relationship between hub genes and immune cells**

133   The relationship between the six hub genes and the infiltrating immune cells was analyzed in

134   three datasets. Hub genes expression levels were divided into low- or high-expression groups

135   based on the genes' median expression levels. The Wilcoxon rank-sum test was used to analyze

136   the difference of immune cell infiltration between the high and low levels of hub genes. Box

137   plots were used to visualize the difference in infiltrating immune cell expression between both

138   groups.

139   **Sample collection**

140   Peripheral blood samples were collected from patients in the Second Affiliated Hospital of

141   Nanchang University from March 2020 to February 2021. CML patients were pathologically

142   diagnosed and categorized into CML-CP (n=25) and CML-BP (n=25) groups. The normal

143   samples (n=25) were obtained from healthy individuals. The characteristics of CML patients and

144   healthy donors are presented in Table 1. PBMCs were isolated by density gradient centrifugation

145   on a Ficoll-Paque (Sigma, USA) according to the manufacturer's protocol. Our study was

146   approved by the Ethics Committee of the Second Affiliated Hospital of Nanchang University

147   (approval no. 2017096) and all aspects of the study followed the guidelines set forth by the

148   Helsinki Declaration. All participants provided signed informed consent.

149   **RNA isolation and quantitative real-time PCR (RT-qPCR)**

150   Total RNA was extracted using TRIzol reagent (Invitrogen). Reverse transcription was

151   performed using a PrimeScript RT Reagent Kit (TaKaRa, Dalian, China). We performed RT-

152   qPCR using SYBR Premix Ex Taq (TaKaRa) on a ABI 7500 Real-Time PCR System (Applied

153   Biosystems, Foster City, CA, USA). All primers were designed and synthesized by Nanjing

154   Genscript Biotech Co., Ltd. GAPDH was used as the internal reference. The primers used in RT-

155   qPCR assay are listed in Table 2. Indicated genes' fold changes were calculated using the $2^{-\Delta\Delta Ct}$

156   method.

157   **Statistical analysis**

158 Statistical analysis was performed with R Software (version 3.6.2). A *P*-value < 0.05 was
159 considered to be statistically significant.

160

161 **Results**

162 **Identification of IRGs**

163 We used the limma package to analyze the differences between the CML-CP-Normal (CML-
164 CP=5, Normal=5), CML-BP-Normal (CML-BP=5, Normal=5), and CML-CP-CML-BP (CML-
165 CP=5, CML-BP=5) datasets. We ascertained where the differential genes and IRGs in each
166 dataset intersected. From the three datasets, we identified 336, 159, and 376 differential genes,
167 respectively (Fig. 2G, 2H, 2I). CML-CP-Normal's clustering heat map is shown in Fig. 2A and
168 its volcano map is shown in Fig. 2D. CML-BP-Normal's clustering heat map is shown in Fig. 2B
169 and its volcano map is shown in Fig. 2E. CML-CP-CML-BP's clustering heat map is shown in
170 Fig. 2C and its volcano map is shown in Fig. 2F. Thirty-one IRGs were obtained from the
171 intersection of the differential genes and the immune genes in the three datasets (Fig. 2J).

172 **Functional enrichment analysis of IRGs**

173 Thirty-one IRGs were enriched and analyzed using the Metascape database. The top GO terms
174 related to biological processes (BC) were neutrophil activation, neutrophil degranulation, and
175 neutrophil activation involved in immune response (Fig. 3A). The top GO terms related to
176 cellular components (CC) included secretory granule lumen, cytoplasmic vesicle lumen, and
177 vesicle lumen. The top GO terms related to molecular functions (MF) included receptor ligand
178 activity, cytokine activity, and endopeptidase activity. KEGG analysis indicated that these IRGs
179 were mainly enriched in Th17 cell differentiation, NF-kappa B signaling pathway, and cytokine-
180 cytokine receptor interaction (Fig. 3B).

181 **PPI network analysis of IRGs**

182 STRING was used to generate the PPI network, and the Cytoscape tool was used to visualize
183 interactions between 31 IRGs (Fig. 4A and 4B). The cytoHubba and MCODE plug-ins were used
184 to analyze the hub genes with maximum correlation criterion (MCC), and genes with the top 10
185 scores were identified as hub genes (Fig. 4C and 4D). Among these genes, *AZU1*, *CAMP*, *CCL5*,
186 *CTSG*, *MMP9*, *MPO*, *PRTN3*, *RETN*, *RNASE2*, and *RNASE3* showed the highest node scores,
187 suggesting that they may play causative roles in CML progression.

188 **GSEA analysis of hub genes**

189 GSEA were performed on the 10 hub genes obtained from the GSE10002 dataset. The median of

190 each gene expression value was used to categorize the patients into high or low groups. The

191 msigDB database's hallmark gene set was used as a reference in our analysis. We identified

192 different significant pathways between the higher- and lower-gene-expression groups. $P<0.05$

193 and $|NES|>1.0$ was considered to be significant. As shown in Fig. 5, most pathways enriched in

194 the hub genes high-expression cohort were immune- related, including interferon gamma

195 response, IL2-STAT5 signaling pathway, complement, allograft rejection.

196 **Hub genes for CML diagnosis**

197 To explore the accuracy of the top 10 hub genes as the diagnostic biomarkers for CML, the ROC

198 curves were plotted, respectively (Fig. 6). Six hub genes with an AUC value greater than 0.7

199 were used as diagnostic markers. Notably, the AUC for *CTSG* reached 0.96, and was the largest

200 among the 10 hub genes. The other AUC values were 0.82, 0.74, 0.88, 0.84, and 0.88 for *MMP9*

201 , *PRTN3*, *RETN*, *RNASE2*, and *RNASE3*, respectively. These results suggest that the 6 hub genes

202 may serve as diagnostic biomarkers for CML.

203 **Immune infiltration analysis**

204 We performed immune infiltration analysis on three datasets (Figs. 7-9). The distribution of the

205 22 types of infiltrating immune cells in each sample showed there to be immunological

206 differences between the two groups (Fig. 7A, 8A, 9A). The abundance of immune cells in each

207 sample is illustrated as a heatmap in Fig. 7B, 8B, 9B; the correlation between each of the

208 infiltrated immune cell types is shown in Fig. 7C, 8C, 9C; and the violin plot in Fig. 7D, 8D, 9D

209 visualizes the differences in each type of immune cell between the 2 groups.

210 **Relationships between the hub genes and immune cells**

211 We analyzed the relationship between hub genes *CTSG*, *MMP9*, *PRTN3*, *RETN*, *RNASE2* and

212 *RNASE3* and immune cells in the three datasets, respectively (Figs. 10-12). In the CML-BP-

213 Normal datasets, *CTSG*, *MMP9*, *PRTN3*, *RNASE2* and *RNASE3* were significantly associated

214 with the infiltration of monocytes, mast cells resting and NK cells (Fig. 10). In the CML-CP-

215 CML-BP datasets, the six hub genes were significantly associated with the infiltration of CD8[+]T

216 cells, CD4[+]T cells, NK cells resting, macrophages M0, eosinophils (Fig. 12). The results suggest

217 that there is some correlation between the six hub genes and immune response in CML patients.

218 **Validating hub genes expression in clinical samples**

219 To further assess the *CTSG*, *MMP9*, *PRTN3*, *RETN*, *RNASE2*, and *RNASE3* expression levels in

220 CML, we perform RT-qPCR to search the mRNA expression of the hub genes in PBMCs of the

221 healthy individuals and CML-CP/BP patients (Fig. 13). The results showed that the expression of

222 *CTSG*, *MMP9*, *PRTN3*, *RETN*, *RNASE2*, and *RNASE3* were dramatically upregulated in CML-

223 CP comparing to the normal cells (*P*<0.001). Compared with CML-CP samples, the expression

224 of *CTSG*, *RETN*, *RNASE2*, and *RNASE3* were significantly lower in the CML-BP samples. The

225 expression trend was consistent with the GSE100026 dataset.

226 In the present study, the correlation between clinical features of CML and the expression

227 levels of hub genes were analyzed. As shown in Table 3, the expression of *CTSG* was positively

228 associated with white blood cell (WBC) number (*r*=0.4827, *P*=0.0007) and haemoglobin

229 (*r*=0.3135, *P*=0.0241). The expression of *PRTN3* was positively associated with haemoglobin

230 (*r*=0.3326, *P*=0.0128). The expression of *RETN* was positively associated with WBC number

231 (*r*=0.3832, *P*=0.0027). Furthermore, the expression levels of *MMP9*, *RNASE2*, and *RNASE3* in

232 PBMCs from patients with CML did not correlate with WBC number, haemoglobin, and platelet

233 (Table 3).

234

Discussion

236 Approximately 15% of all new leukemia cases are CML according to American Cancer Society

237 estimates (Apperley 2015). The use of TKIs such as imatinib and nilotinib significantly

238 improved the efficacy of CML treatment. TKI treatment over 12 months led 66% of CML

239 patients to have a complete cytogenetic response. Major molecular biological responses were

240 noted in 40% of patients and the 10-year overall survival rate reached 85%-90% (Melo & Barnes

241 2007). However, drug resistance has emerged with diversified and individual characteristics,

242 seriously affecting TKI efficacy (Deininger et al. 2000). Therefore, it is important to develop

243 therapeutic targets aside from ABL kinase.

244 Recent immunotherapy research has brought to light the relationship between tumors and

245 immunity. The positive response to immunotherapy typically depends on the interaction between

246 tumor cells and immune regulation. Previous studies have investigated the role of immunity in

247 various cancers and have found that immune effector cells such as T and NK cells can effectively

248 eliminate tumor cells (Sconocchia et al. 2005). However, immune cells also confer resistance

249 when the oncogenic signaling pathway is activated or the protective mechanism against cell

250   death is active (Held et al. 2016). These findings are ready to be applied clinically. For example,

251   the National Institute for Health and Care Excellence (NICE) recommends that some patients

252   using National Healthcare System (NHS) adopt new CAR-T treatments for lymphoma (Mahase

253   2021). We focused on immune molecule and immune cell changes during CML progression and

254   attempted to locate IRGs to extract key genes and explore effective diagnostic markers. We

255   identified 31 IRGs by comparing the transcriptional expression profiles of CML-CP, CML-BP,

256   and healthy samples. GO and KEGG enrichment analysis revealed that these enriched modules

257   and pathways are closely related to immune cell activation and the immune response observed in

258   CML. The top 10 hub genes associated with CML were identified in the PPI network. Of these

259   ten, six hub genes (*CTSG*, *MMP9*, *PRTN3*, *RETN*, *RNASE2* and *RNASE3*) may be used as CML

260   diagnostic markers. Their expressions were significantly associated with a variety of immune

261   cell disorders by immune infiltration analysis.

262   We performed GO enrichment analysis to investigate the 31 IRGs' BPs. Of the MF

263   annotations, receptor ligand activity, cytokine activity, endopeptidase activity were found to be

264   significantly associated with CML occurrence and development. We performed KEGG analysis

265   to determine the biological functions of the 31 immune-related genes associated with CML.

266   Th17 cell differentiation, NF-kappa B signaling pathway, and cytokine-cytokine receptor

267   interaction were the three most significantly enriched pathways. Interestingly, some enriched

268   pathways were associated with CML. For instance, NF-κB is activated through the canonical

269   IKK pathway and plays distinct roles in the pathogenesis of myeloid and lymphoid leukemias

270   induced by BCR-ABL1. This confirms that NF-κB and IKKs are targets for Ph(+) leukemias

271   therapy (Hsieh & Van Etten 2014). Previous studies demonstrated that IL-2-activated NK

272   (ANK) with a potent major histocompatibility complex that had unrestricted cytotoxic activity

273   and   suppressed   malignant   hematopoiesis.   These   observations   support   the   use   of

274   autologous ANK therapy to prevent the relapse of CML after autologous marrow transplantation

275   or the use of ANK to purge CML marrow for autologous transplantation (Cervantes et al. 1996).

276   Autology-activated NK cells can inhibit primitive CML progenitors in long term cultures. CML

277   cells can avoid immune escape *in vivo*, resulting in NK cell-mediated immune destruction

278   (Cervantes et al. 1996). IFN-γ released by activated T or NK cells modulate the sensitivity of

279   CML cells to TKI, thereby interfering with the therapeutic effect of TKI on CML. This effect

280 suggests that inflammation-mediated BCR-ABL1 independent resistance mechanisms are
281 significant (Held et al. 2016).

282   In the PPI network identified in the present study, 10 IRGs were determined to be the most
283 significant hub genes, with multiple interactions in the network. The 10 hub genes' AUCs were
284 analyzed with regard to diagnostic value. An AUC value greater than 0.7 was the screening
285 criteria. Six hub genes (*CTSG, MMP9, PRTN3, RETN, RNASE2, RNASE3*) were determined to
286 be diagnostic markers for CML. The AUC values ranged from 0.74-0.96. The RT-qPCR results
287 showed that CTSG, MMP9, PRTN3, RETN, RNASE2, RNASE3 were significantly elevated in
288 PBMCs of CML patients compared with healthy controls. GSEA showed that these hub genes
289 were predominantly associated with the interferon gamma response, inflammatory response, and
290 allograft rejection.

291   Cathepsin G (CTSG) plays an important role in host defense and can escape intracellular
292 monitoring systems by preventing degradation of foreign proteins. CTSG is broadly expressed in
293 acute myeloid leukemia (AML) blasts and leukemia stem cells (Alatrash 2013). Recent studies
294 have shown that *CTSG* is an effective target for the immunotherapy of AML and acute lymphoid
295 leukemia (ALL) (Groborz et al. 2019). We found that CTSG was highly-expressed in CML-CP
296 and CML-BP. We also found that the AUC of CTSG was 0.96, indicating that it has high
297 diagnostic value. Matrix metalloproteinase 9 (MMP9) upregulation in leukemic cells promotes
298 the blast crisis in CML (Nakahara et al. 2014). In chronic lymphocytic leukemia (CLL), MMP9
299 contributes to CLL pathology by regulating cell survival and migration, promoting angiogenesis,
300 and is associated with poor prognosis (Aguilera-Montilla et al. 2019). It has been shown that the
301 ID1 transcriptional inhibitor-MMP9 axis can enhance the invasiveness of BCR-ABL1
302 transformed leukemia cells (Nieborowska-Skorska et al. 2006). Human proteinase 3 (PRTN3) is
303 a leukemia-associated antigen. The identified PRTN3(235)-epitope can be used to study the role
304 of CD4+Th- and Treg-cells in immune responses against PRTN3 in leukemia patients (Berlin et
305 al. 2015). Resistin (RETN), a proinflammatory cytokine, is elevated in a number of pathological
306 disorders, including leukemia. The serum RETN level in all patients was elevated and may serve
307 as a potential clinical diagnostic marker to detect the recurrence of leukemia (El-Baz et al. 2013).
308 Human eosinophil derived neurotoxin (EDN, RNase2) and eosinophil cationic protein
309 (ECP, RNase3) sequences may have up to a 92% identity in their promoter regions (Wang et al.
310 2009). It produces host defense effects by promoting leukocyte activation, maturation, and

311 chemotaxis. Niini *et al*. (2002) demonstrated that RNase2 was abnormally expressed in
312 childhood acute lymphoblastic leukemia (a cDNA array) and was associated with prognosis.
313 These findings are consistent with our work, which suggests that there may be a close
314 relationship between CML and immune composition.

315 Our study had some limitations. First, the sample size was small; a larger sample size is
316 needed to verify our results. Second, some clinical data are lacking due to the heterogeneity, long
317 duration, and difficulty of follow-up of CML. It is difficult for us to assess the relationship
318 between risk indicators and patient stratification based on the severity of CML. In order to
319 determine the value of hub genes related to CML as diagnostic and therapeutic biomarkers, more
320 clinical variables should be used for further external verification. Third, the potential mechanism
321 of the six hub genes in CML progression is still not well understood and additional studies may
322 be needed to determine the six hub genes' underlying molecular mechanisms in CML.

323

324 Conclusions

325 Ours is the first study to comprehensively identify IRGs in CML. In total, 31 IRGs and 10 hub
326 genes were identified and may serve as novel potential targets for the diagnosis and treatment of
327 CML. However, more molecular experiments are needed to validate our findings.

328

329 References

330 Aguilera-Montilla N, Bailón E, Ugarte-Berzal E, Uceda-Castro R, Prieto-Solano M, García-
331   Martínez E, Samaniego R, Van den Steen P, Opdenakker G, García-Marco J, and García-
332   Pardo A. 2019. Matrix metalloproteinase-9 induces a pro-angiogenic profile in chronic
333   lymphocytic leukemia cells. *Biochemical and biophysical research communications*
334   520:198-204. 10.1016/j.bbrc.2019.09.127
335 Akobeng AK. 2007. Understanding diagnostic tests 3: Receiver operating characteristic curves.
336   *Acta Paediatr* 96:644-647. 10.1111/j.1651-2227.2006.00178.x
337 Alatrash G. 2013. Targeting cathepsin G in myeloid leukemia. *Oncoimmunology* 2:e23442.
338   10.4161/onci.23442
339 Apperley JF. 2015. Chronic myeloid leukaemia. *Lancet* 385:1447-1459. 10.1016/s0140-
340   6736(13)62120-0
341 Berlin C, Kowalewski D, Schuster H, Mirza N, Walz S, Handel M, Schmid-Horch B, Salih H,
342   Kanz L, Rammensee H, Stevanović S, and Stickel J. 2015. Mapping the HLA ligandome
343   landscape of acute myeloid leukemia: a targeted approach toward peptide-based
344   immunotherapy. *Leukemia* 29:647-659. 10.1038/leu.2014.233
345 Bhattacharya S, Dunn P, Thomas C, Smith B, Schaefer H, Chen J, Hu Z, Zalocusky K, Shankar
346   R, Shen-Orr S, Thomson E, Wiser J, and Butte A. 2018. ImmPort, toward repurposing of

347   open access immunological assay data for translational and clinical research. *Scientific*
348   *data* 5:180015. 10.1038/sdata.2018.15
349   Cervantes F, Pierson B, McGlave P, Verfaillie C, and Miller J. 1996. Autologous activated
350   natural killer cells suppress primitive chronic myelogenous leukemia progenitors in long-
351   term culture. *Blood* 87:2476-2485.
352   Chen B, Khodadoust M, Liu C, Newman A, and Alizadeh A. 2018. Profiling Tumor Infiltrating
353   Immune Cells with CIBERSORT. *Methods in molecular biology (Clifton, NJ)* 1711:243-
354   259. 10.1007/978-1-4939-7493-1_12
355   Chin C, Chen S, Wu H, Ho C, Ko M, and Lin C. 2014. cytoHubba: identifying hub objects and
356   sub-networks from complex interactome. *BMC systems biology*:S11. 10.1186/1752-0509-
357   8-s4-s11
358   Deininger MW, Goldman JM, and Melo JV. 2000. The molecular biology of chronic myeloid
359   leukemia. *Blood* 96:3343-3356.
360   El-Baz HA, Mosa TE, Elabd EM, Ramadan A, Elharoun AS, Elmorsy EA, and Fouda MI. 2013.
361   Serum adiponectin and resistin levels in de novo and relapsed acute lymphoblastic
362   leukemia children patients. *Iran J Public Health* 42:504-510.
363   Feng XQ, Nie SM, Huang JX, Li TL, Zhou JJ, Wang W, Zhuang LK, and Meng FJ. 2020.
364   Circular RNA circHIPK3 serves as a prognostic marker to promote chronic myeloid
365   leukemia progression. *Neoplasma* 67:171-177. 10.4149/neo_2018_181129N908
366   Geng Q, Shen Z, Li L, and Zhao J. 2021. COL1A1 is a prognostic biomarker and correlated with
367   immune infiltrates in lung cancer. *PeerJ* 9:e11145. 10.7717/peerj.11145
368   Groborz K, Kołt S, Kasperkiewicz P, and Drag M. 2019. Internally quenched fluorogenic
369   substrates with unnatural amino acids for cathepsin G investigation. *Biochimie* 166:103-
370   111. 10.1016/j.biochi.2019.05.013
371   Hehlmann R. 2012. How I treat CML blast crisis. *Blood* 120:737-747. 10.1182/blood-2012-03-
372   380147
373   Held S, Heine A, Kesper A, Schönberg K, Beckers A, Wolf D, and Brossart P. 2016. Interferon
374   gamma modulates sensitivity of CML cells to tyrosine kinase inhibitors.
375   *Oncoimmunology* 5:e1065368. 10.1080/2162402x.2015.1065368
376   Hsieh M, and Van Etten R. 2014. IKK-dependent activation of NF-κB contributes to myeloid
377   and lymphoid leukemogenesis by BCR-ABL1. *Blood* 123:2401-2411. 10.1182/blood-
378   2014-01-547943
379   Hu L, Han Z, Cheng X, Wang S, Feng Y, and Lin Z. 2021. Expression Profile Analysis Identifies
380   a Novel Seven Immune-Related Gene Signature to Improve Prognosis Prediction of
381   Glioblastoma. *Front Genet* 12:638458. 10.3389/fgene.2021.638458
382   Li SQ, Liu J, Zhang J, Wang XL, Chen D, Wang Y, Xu YM, Huang B, Lin J, Li J, and Wang XZ.
383   2020. Transcriptome profiling reveals the high incidence of hnRNPA1 exon 8 inclusion
384   in chronic myeloid leukemia. *J Adv Res* 24:301-310. 10.1016/j.jare.2020.04.016
385   Lugo TG, Pendergast AM, Muller AJ, and Witte ON. 1990. Tyrosine kinase activity and
386   transformation potency of bcr-abl oncogene products. *Science* 247:1079-1082.
387   10.1126/science.2408149
388   Mahase E. 2021. NICE recommends new CAR-T treatment for lymphoma for some patients on
389   NHS. *BMJ (Clinical research ed)* 372:n184. 10.1136/bmj.n184
390   Melo JV, and Barnes DJ. 2007. Chronic myeloid leukaemia as a model of disease evolution in
391   human cancer. *Nat Rev Cancer* 7:441-453. 10.1038/nrc2147

392 Nakahara F, Kitaura J, Uchida T, Nishida C, Togami K, Inoue D, Matsukawa T, Kagiyama Y,
393   Enomoto Y, Kawabata K, Chen-Yi L, Komeno Y, Izawa K, Oki T, Nagae G, Harada Y,
394   Harada H, Otsu M, Aburatani H, Heissig B, Hattori K, and Kitamura T. 2014. Hes1
395   promotes blast crisis in chronic myelogenous leukemia through MMP-9 upregulation in
396   leukemic cells. *Blood* 123:3932-3942. 10.1182/blood-2013-01-476747
397 Nash I. 1999. Chronic myeloid leukemia. *N Engl J Med* 341:765.
398   10.1056/nejm199909023411016
399 Nieborowska-Skorska M, Hoser G, Rink L, Malecki M, Kossev P, Wasik M, and Skorski T.
400   2006. Id1 transcription inhibitor-matrix metalloproteinase 9 axis enhances invasiveness
401   of the breakpoint cluster region/abelson tyrosine kinase-transformed leukemia cells.
402   *Cancer research* 66:4108-4116. 10.1158/0008-5472.Can-05-1584
403 Ritchie M, Phipson B, Wu D, Hu Y, Law C, Shi W, and Smyth G. 2015. limma powers
404   differential expression analyses for RNA-sequencing and microarray studies. *Nucleic*
405   *acids research* 43:e47. 10.1093/nar/gkv007
406 Sconocchia G, Lau M, Provenzano M, Rezvani K, Wongsena W, Fujiwara H, Hensel N,
407   Melenhorst J, Li J, Ferrone S, and Barrett A. 2005. The antileukemia effect of HLA-
408   matched NK and NK-T cells in chronic myelogenous leukemia involves NKG2D-target-
409   cell interactions. *Blood* 106:3666-3672. 10.1182/blood-2005-02-0479
410 Szklarczyk D, Morris J, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva N, Roth
411   A, Bork P, Jensen L, and von Mering C. 2017. The STRING database in 2017: quality-
412   controlled protein-protein association networks, made broadly accessible. *Nucleic acids*
413   *research* 45:D362-D368. 10.1093/nar/gkw937
414 Vinhas R, Cordeiro M, Pedrosa P, Fernandes AR, and Baptista PV. 2017. Current trends in
415   molecular diagnostics of chronic myeloid leukemia. *Leuk Lymphoma* 58:1791-1804.
416   10.1080/10428194.2016.1265116
417 Wang H, Ho P, Lan C, and Chang M. 2009. Transcriptional regulation of human eosinophil
418   RNase2 by the liver-enriched hepatocyte nuclear factor 4. *Journal of cellular*
419   *biochemistry* 106:317-326. 10.1002/jcb.22008
420 Xue F, Yang L, Dai B, Xue H, Zhang L, Ge R, and Sun Y. 2020. Bioinformatics profiling
421   identifies seven immune-related risk signatures for hepatocellular carcinoma. *PeerJ*
422   8:e8301. 10.7717/peerj.8301
423 Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi A, Tanaseichuk O, Benner C, and Chanda S.
424   2019. Metascape provides a biologist-oriented resource for the analysis of systems-level
425   datasets. *Nature communications* 10:1523. 10.1038/s41467-019-09234-6
426 Zou Y, and Hu C. 2020. A 14 immune-related gene signature predicts clinical outcomes of
427   kidney renal clear cell carcinoma. *PeerJ* 8:e10183. 10.7717/peerj.10183
428

Clinical characteristics

Table 1. Clinical characteristics			
Characteristic	CML-CP (n = 25)	CML-BP (n = 25)	Normal (n = 25)
Age (years), median (range)	49 (19-69)	55 (30-73)	52 (21-70)
Male/female (n/n)	13/12	15/10	14/11
WBC count, ×10^9/L, median (range)	138.3(14.8-440.5)	72.6 (2.5-381.9)	6.4 (4.8-9.7)
Haemoglobin level (g/L)	102 (50-143)	79 (43-135)	139 (120-156)
Platelet count, ×10^9/L, median (range)	395 (6-1309)	50 (10-183)	223 (105-318)

1   BP, blast phase; CML, chronic myeloid leukemia; CP, chronic phase; WBC, white blood cells.

RT-qPCR primers

Table 2. RT-qPCR primers	
Primers	Sequences (5'-3')
GAPDH	Forward: ATGGTGAAGGTCGGTGTGAA
	Reverse: GAGTGGAGTCATACTGGAAC
CTSG	Forward: GAGTCAGACGGAATCGAAACG
	Reverse: CGGAGTGTATCTGTTCCCCTC
MMP9	Forward: CGCAGACATCGTCATCCAGT
	Reverse: CGCAGACATCGTCATCCAGT
PRTN3	Forward: CCCTGATCCACCCGAGATTC
	Reverse: GGTTCTCCTCGGGGGTTGTAA
RETN	Forward: GTGTGCCGGATTTGGTTAGC
	Reverse: GAGGGAACCAAGAGACCCAC
RNASE2	Forward: ATCAACGACGAGACCCTCCA
	Reverse: AGGAGCTTGGCAGATGAGTG
RNASE3	Forward: GATCCACGGGATTCTCCACG
	Reverse: GGAGCTTGGCAGATGAGTGA

Correlation of the expression of 6 hub genes with clinical characteristics of CML

Table 3. Correlation of the expression of 6 hub genes with clinical characteristics of CML												
Categories	CTSG		MMP9		PRTN3		RETN		RNASE2		RNASE3	
	r	P-value	r	P-value	r	P-value	r	P-value	r	P-value	r	P-value
Age	0.0328	0.8476	0.0215	0.8041	0.0755	0.6429	0.0952	0.5738	0.0415	0.7928	0.0617	0.6952
WBC	0.4827	0.0007	0.1913	0.2267	0.1625	0.3009	0.3832	0.0027	0.2376	0.1443	0.1954	0.2217
Haemoglobin	0.3135	0.0241	0.0952	0.6673	0.3326	0.0128	0.1884	0.5246	0.1268	0.4572	0.1668	0.3058
Platelet	0.1769	0.2745	0.1364	0.4115	0.2328	0.1494	0.2159	0.1776	0.1675	0.3894	0.0364	0.8215

Figure 1

Flow chart of methodologies applied in the study

GSE100026

Blast crisis   Chronic phase   Normal

Immport

Differentially expressed
immune-related genes

GO and KEGG pathway
enrichment analyses.

PPI and GSEA analysis
(Screening hub genes)

CIBERSORT evaluate the
composition of immune cells

RNA isolation and quantitative
real-time PCR
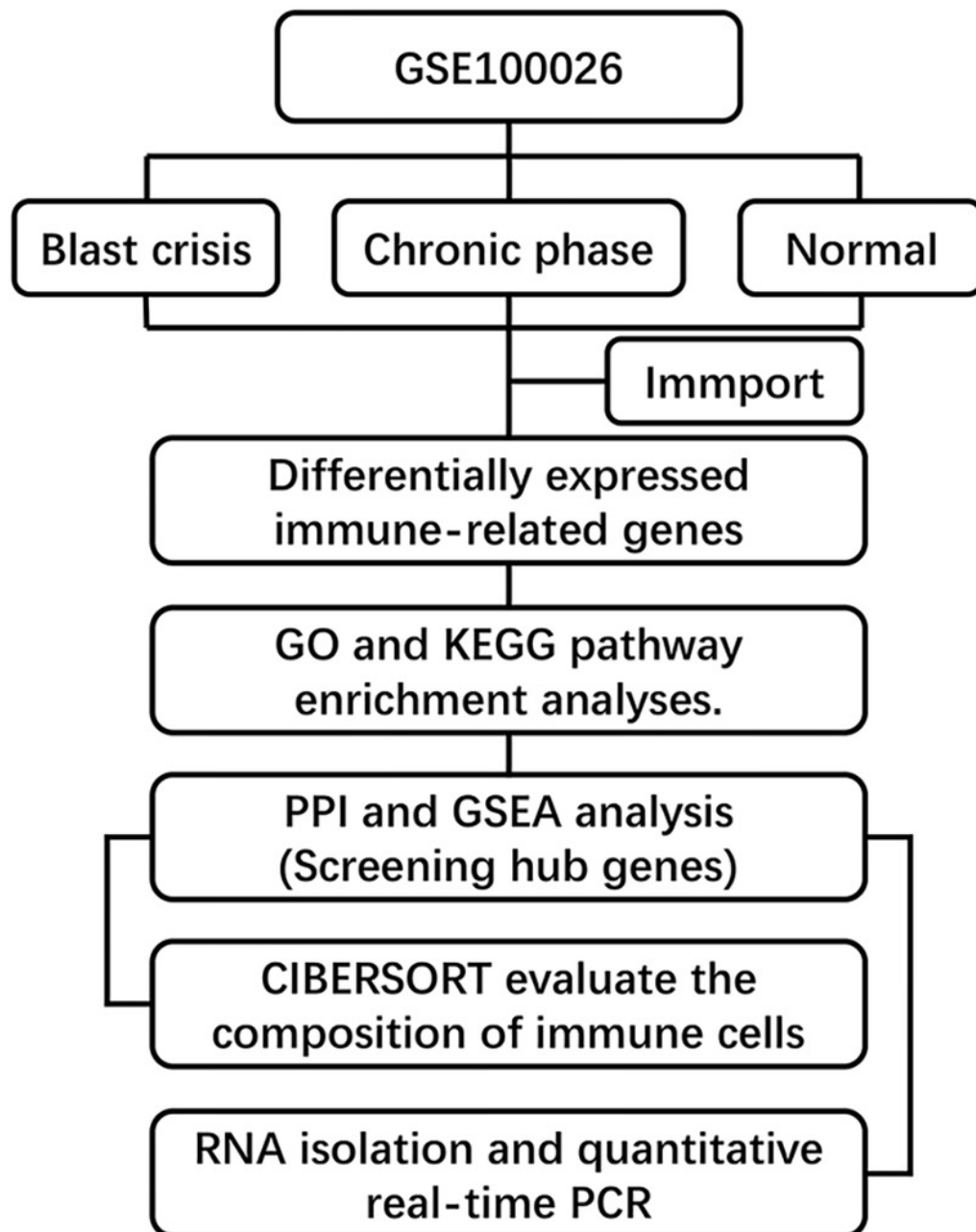
Figure 2

Identification of IRGs in CML

(A)-(C) Difference Analysis between CML-CP-Normal,CML-BP-Normal and CML-CP-CML-BP Cluster Heat Map. (D)-(F) Difference Analysis between CML-CP-Normal, CML-BP-Normal and CML-CP-CML-BP Volcano Map. (G)-(I) CML-CP-Normal, CML-BP-Normal, CML-CP-CML-BP and Immune Gene Intersection Venn diagram. (J) Intersection venn diagram of the IRGs signifificantly associated with CML which were short-listed for the cross-validation.
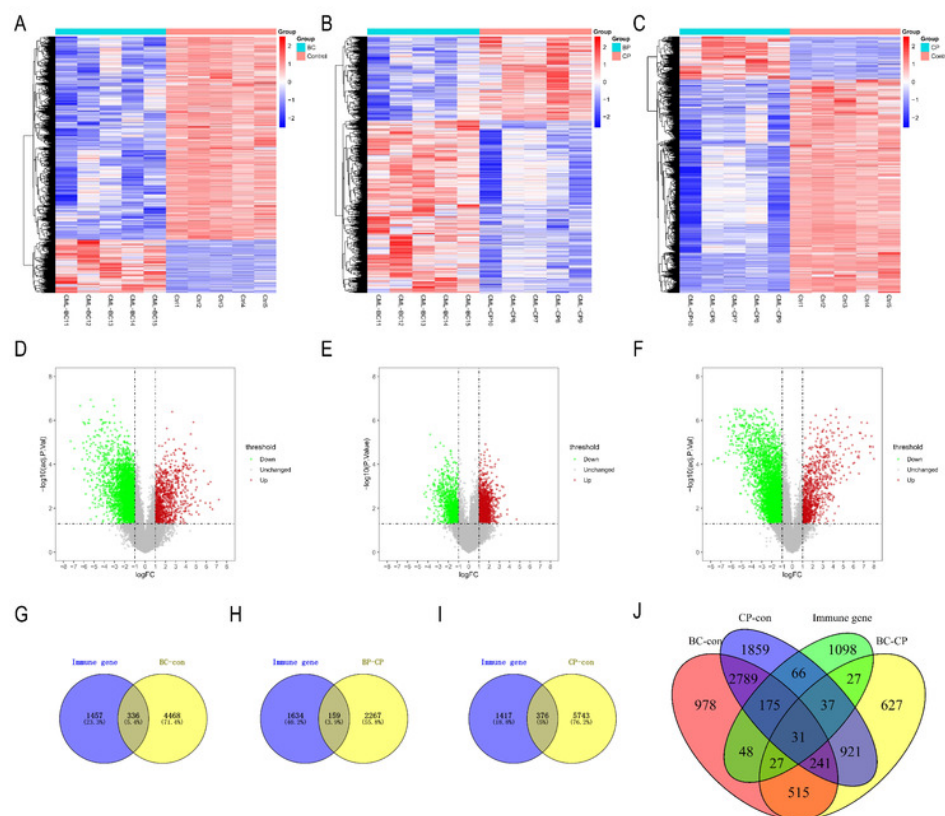
Figure 3

GO term and KEGG pathway analysis for 31 IRGs

(A) GO enrichment analysis. (B) KEGG enrichment analysis.
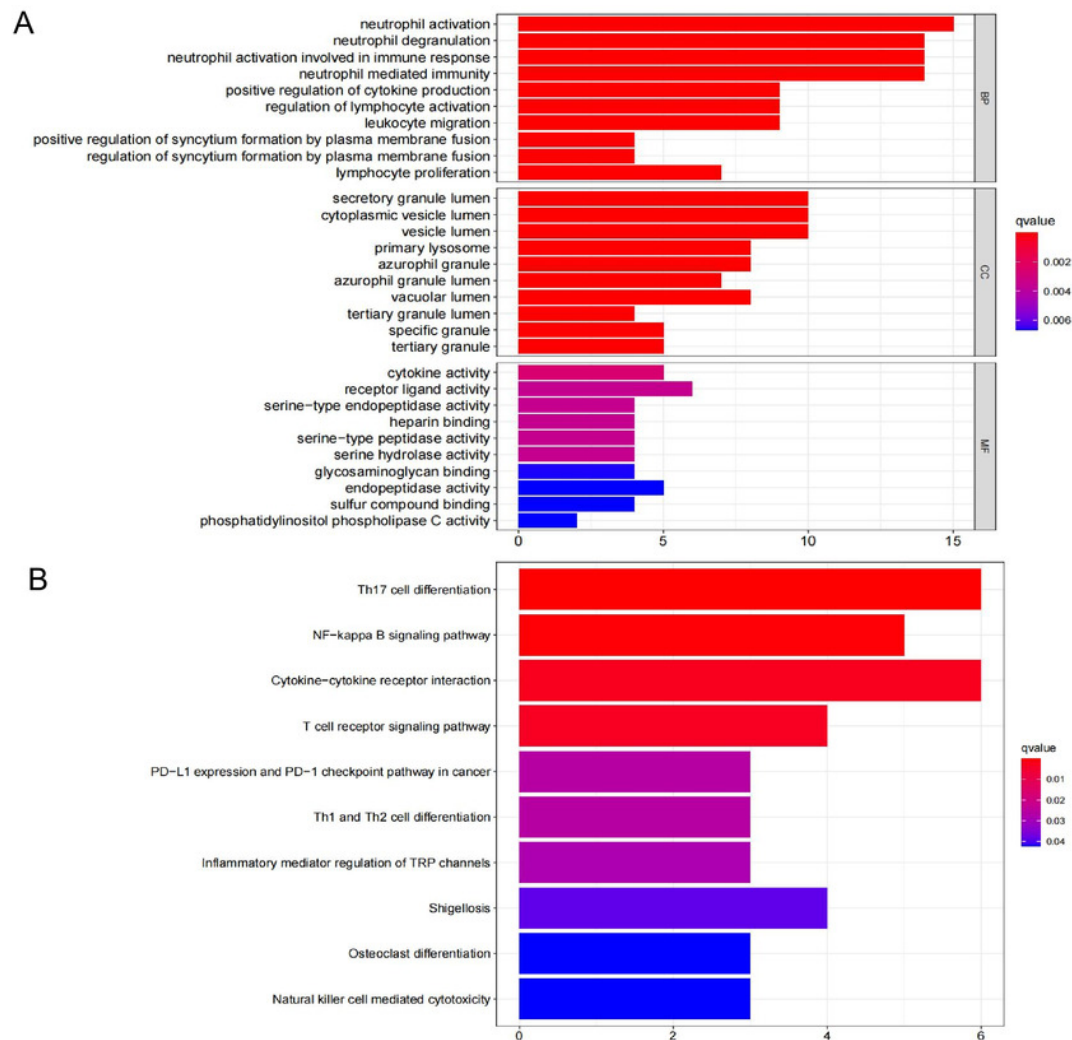
Figure 4

Significant modular analysis based on PPI network

(A)-(B) PPI network was constructed using a total of 31 IRGs. (C) 10 hub genes recognized by cytoHubba plug-in. (D) 10 hub gene recognized by MCODE plug-in.
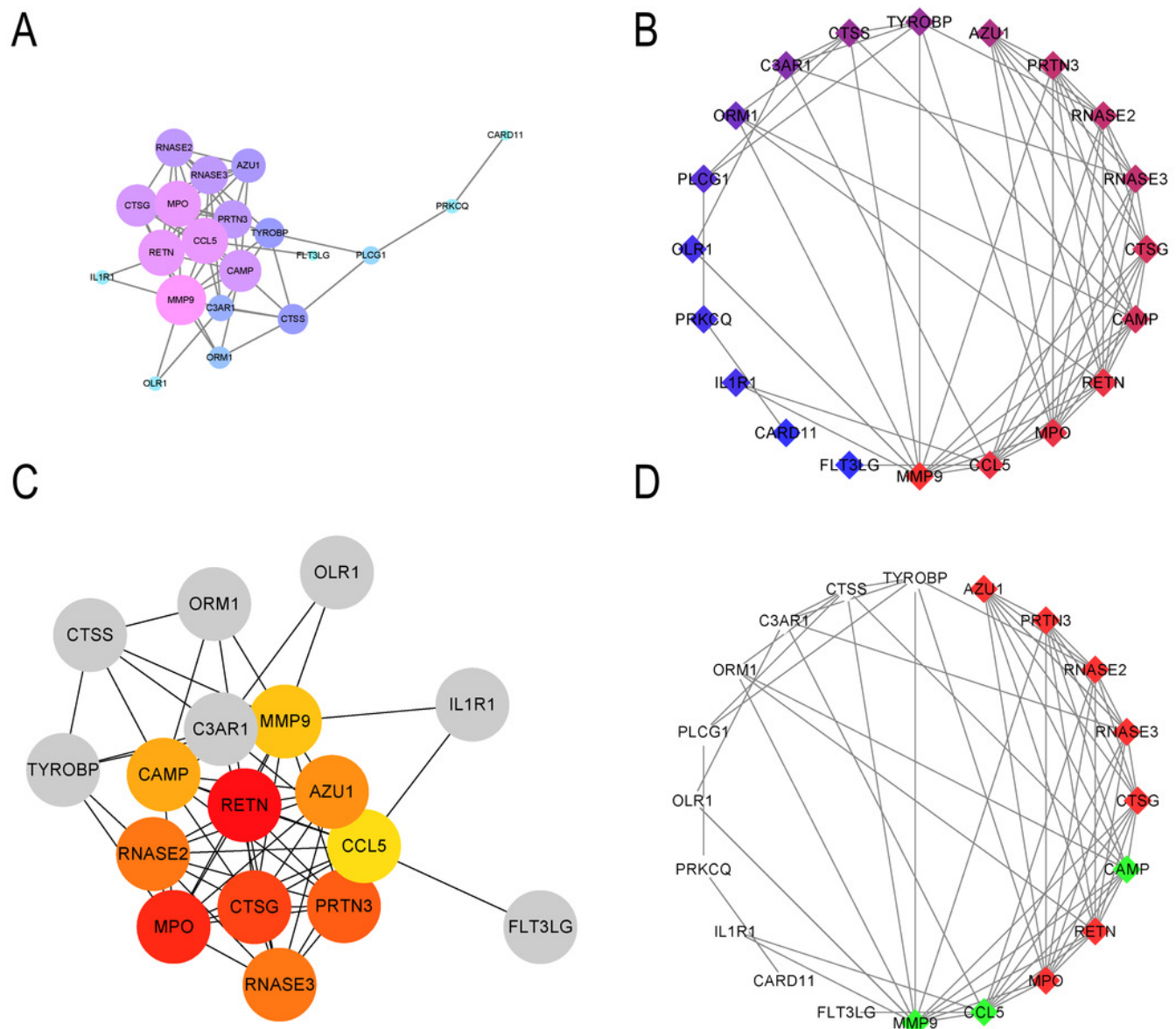
Figure 5

GSEA of the 10 hub gene expression profiles of the hallmark gene set

(A) AZU1. (B) CAMP. (C) CCL5. (D) CTSG. (E) MMP9. (F) MPO. (G) PRTN3. (H) RETN. (I) RNASE2. (J) RNASE3.
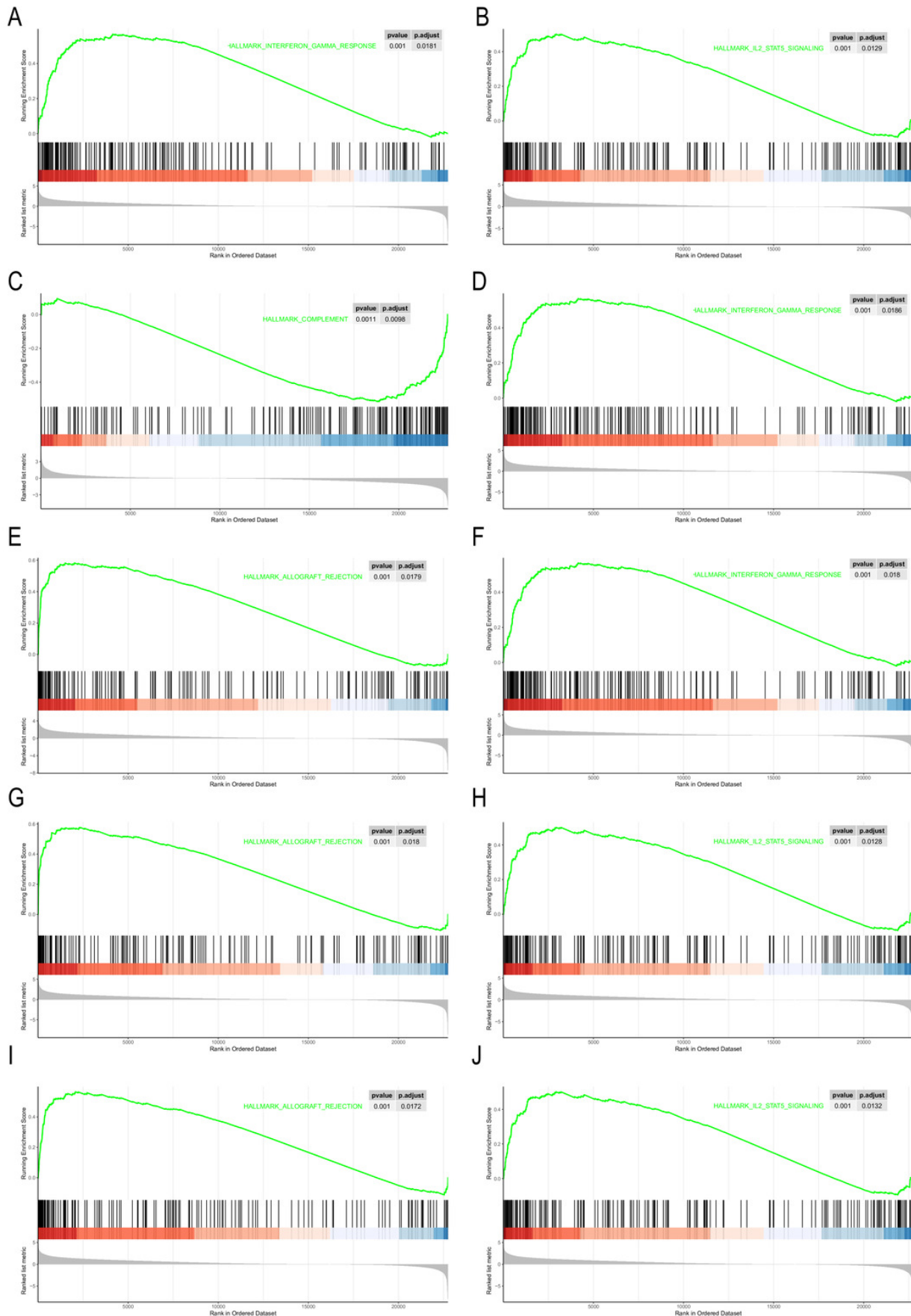
Figure 6

ROC curves were generated to validate the ability of diagnostic value of the 10 hub genes for CML

(A) AZU1. (B) CAMP. (C) CCL5. (D) CTSG. (E) MMP9. (F) MPO. (G) PRTN3. (H) RETN. (I) RNASE2. (J) RNASE3.
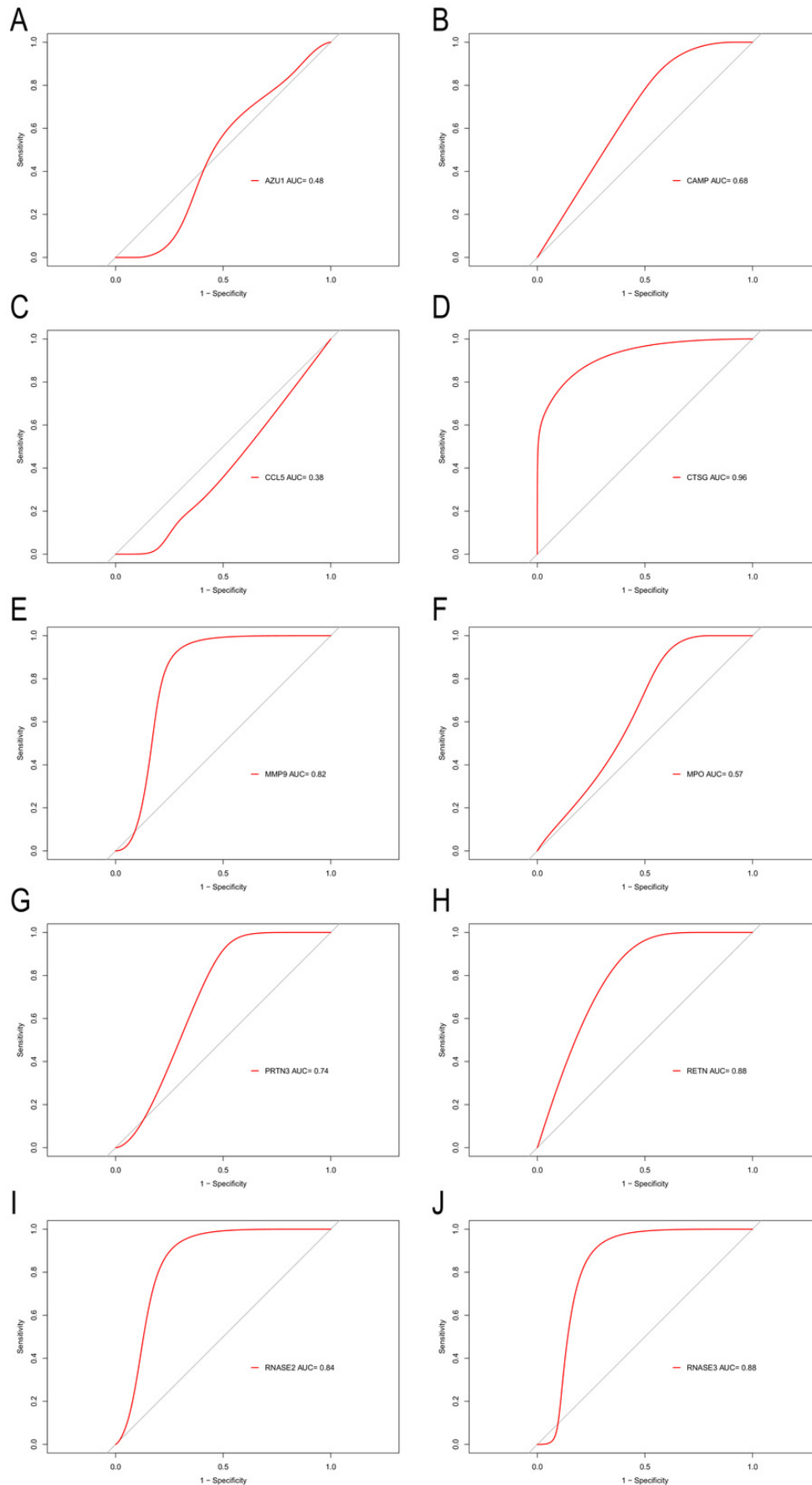
Figure 7

Immune infiltration analysis of CML-CP and Normal

(A) 22 Immune cell composition of each sample. (B) Heat map of the 22 immune cell proportions in CML-CP and Normal. (C) The co-expression patterns among fractions of immune cells. (D) The violin graph shows the difference of immune infiltration between CML-CP and Normal.
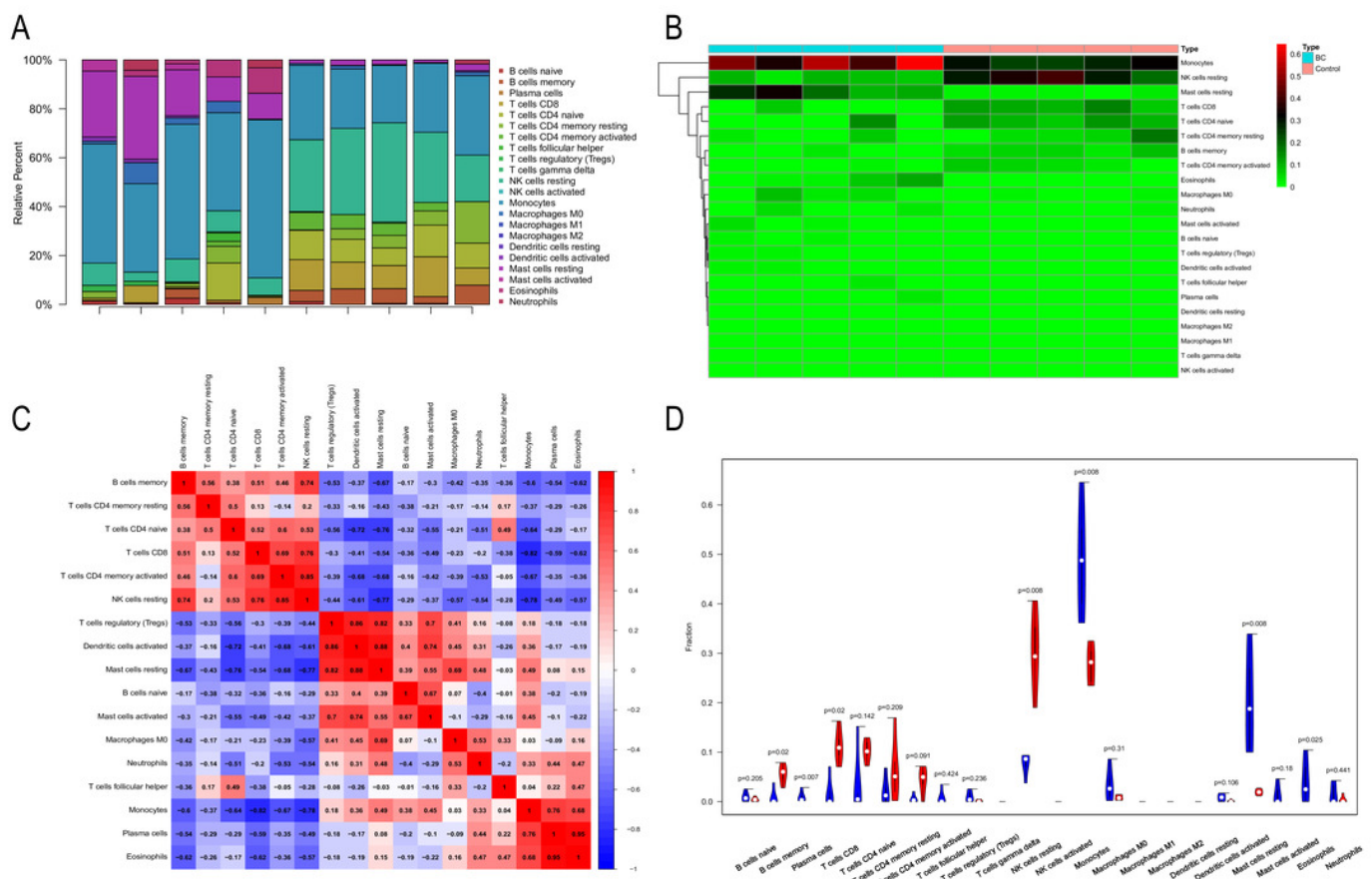
Figure 8

Immune infiltration analysis of CML-BP and Normal

(A) Immune cell composition of each sample. (B) Heat map of the 22 immune cell proportions in CML-BP and Normal. (C) The co-expression patterns among fractions of immune cells. (D) The violin graph shows the difference of immune infiltration between CML-BP and Normal.
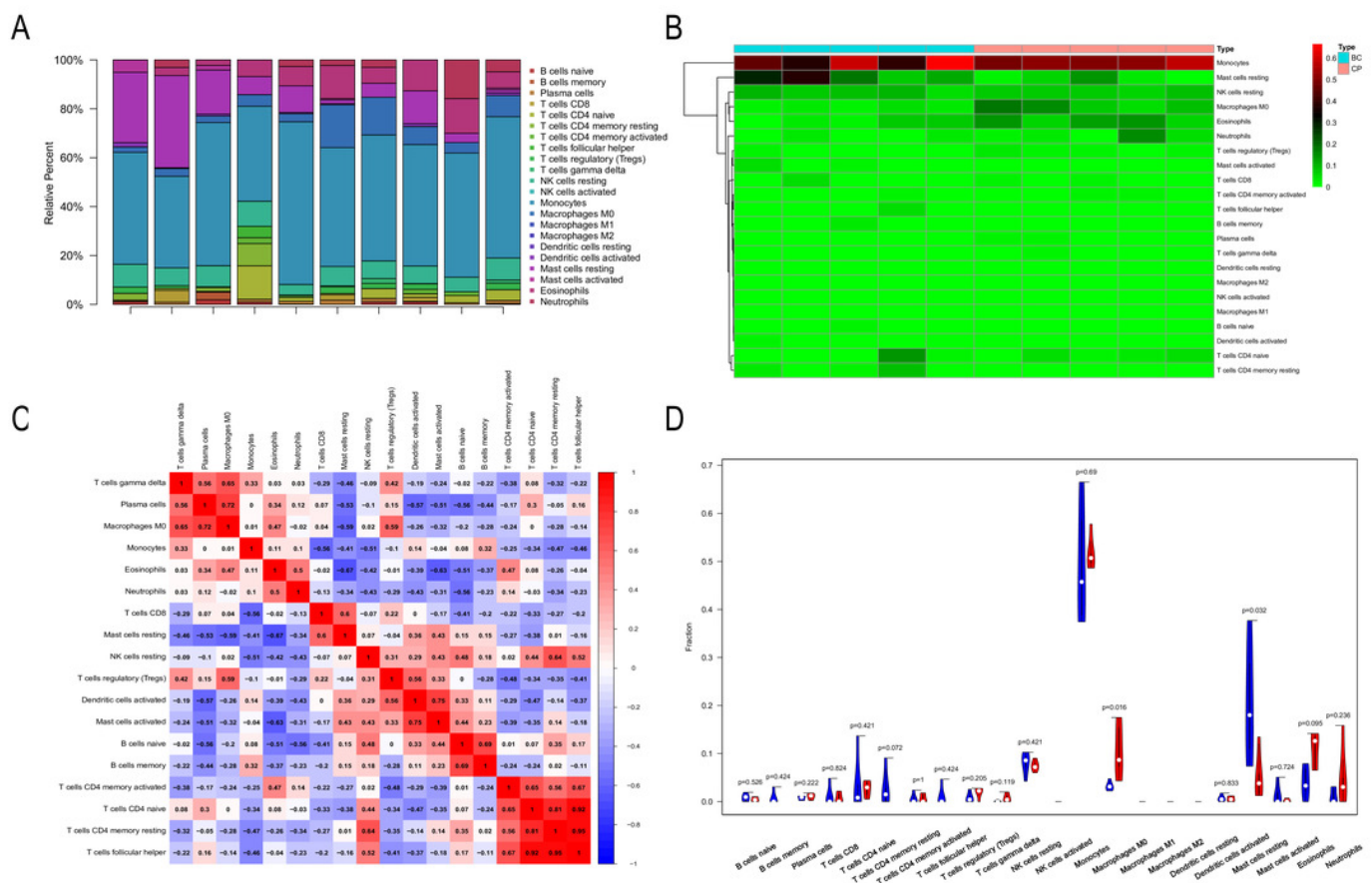
Figure 9

Immune infiltration analysis of CML-CP and CML-BP

(A) Immune cell composition of each sample. (B) Heat map of the 22 immune cell proportions
in CML-CP and CML-BP. (C) The co-expression patterns among fractions of immune cells. (D)
The violin graph shows the difference of immune infiltration between CML-CP and CML-BP.
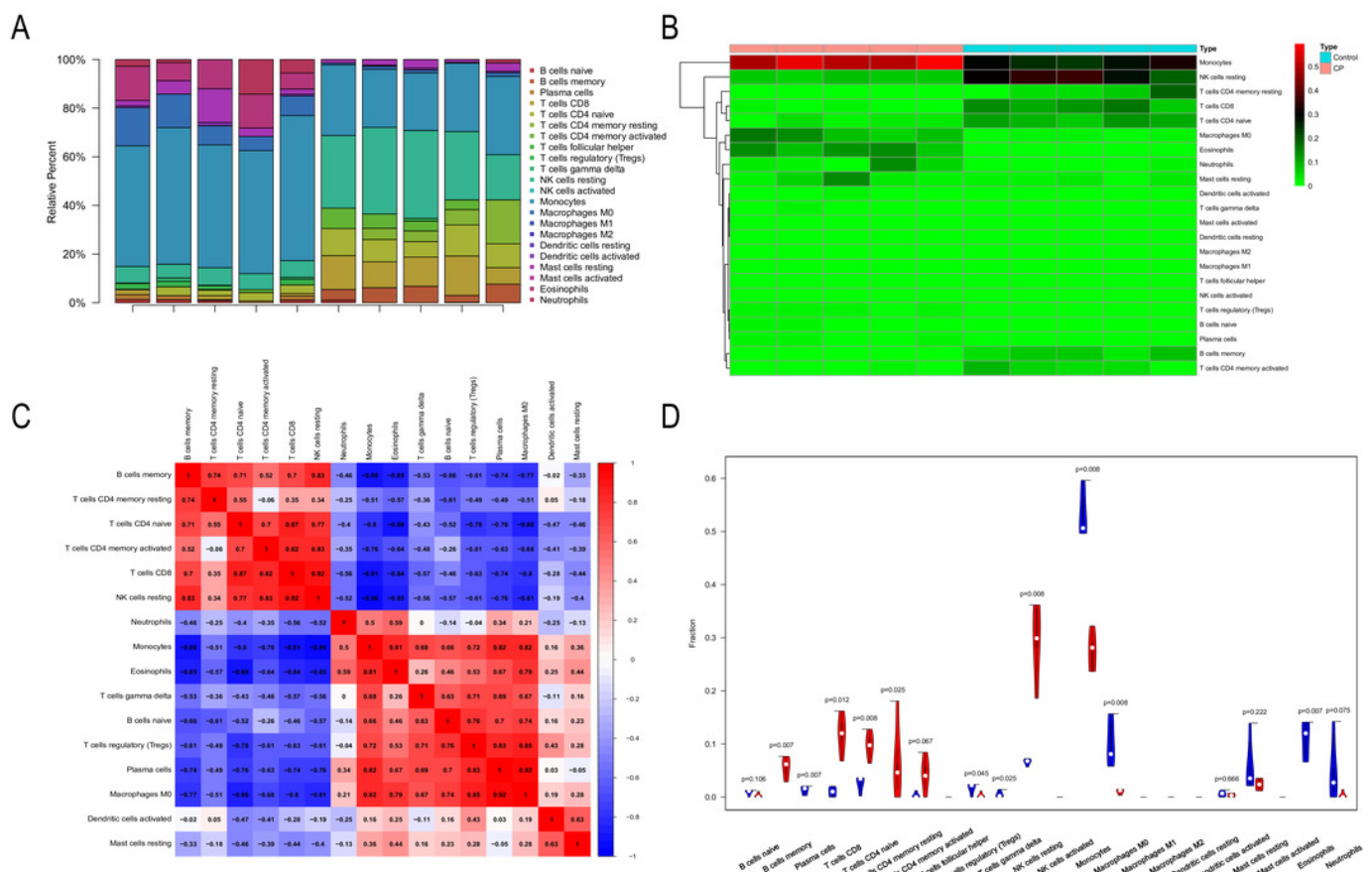
Figure 10

Relationships between the hub genes and immune cells in CML-CP-Normal

(A) CTSG. (B) MMP9. (C) PRTN3. (D) RETN. (E) RNASE2. (F) RNASE3.
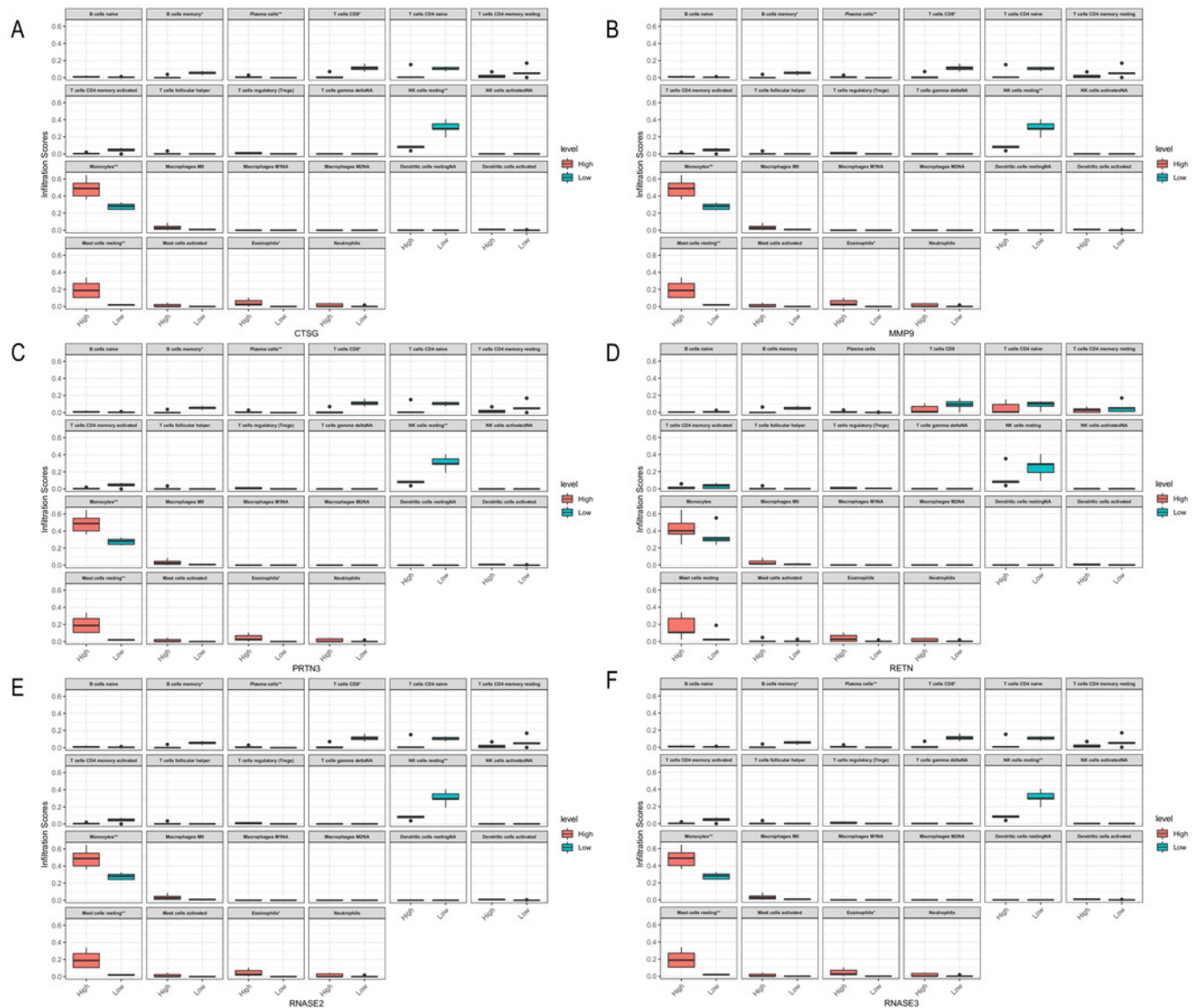
Figure 11

Relationships between the hub genes and immune cells in CML-BP-Normal

(A) CTSG. (B) MMP9. (C) PRTN3. (D) RETN. (E) RNASE2. (F) RNASE3.
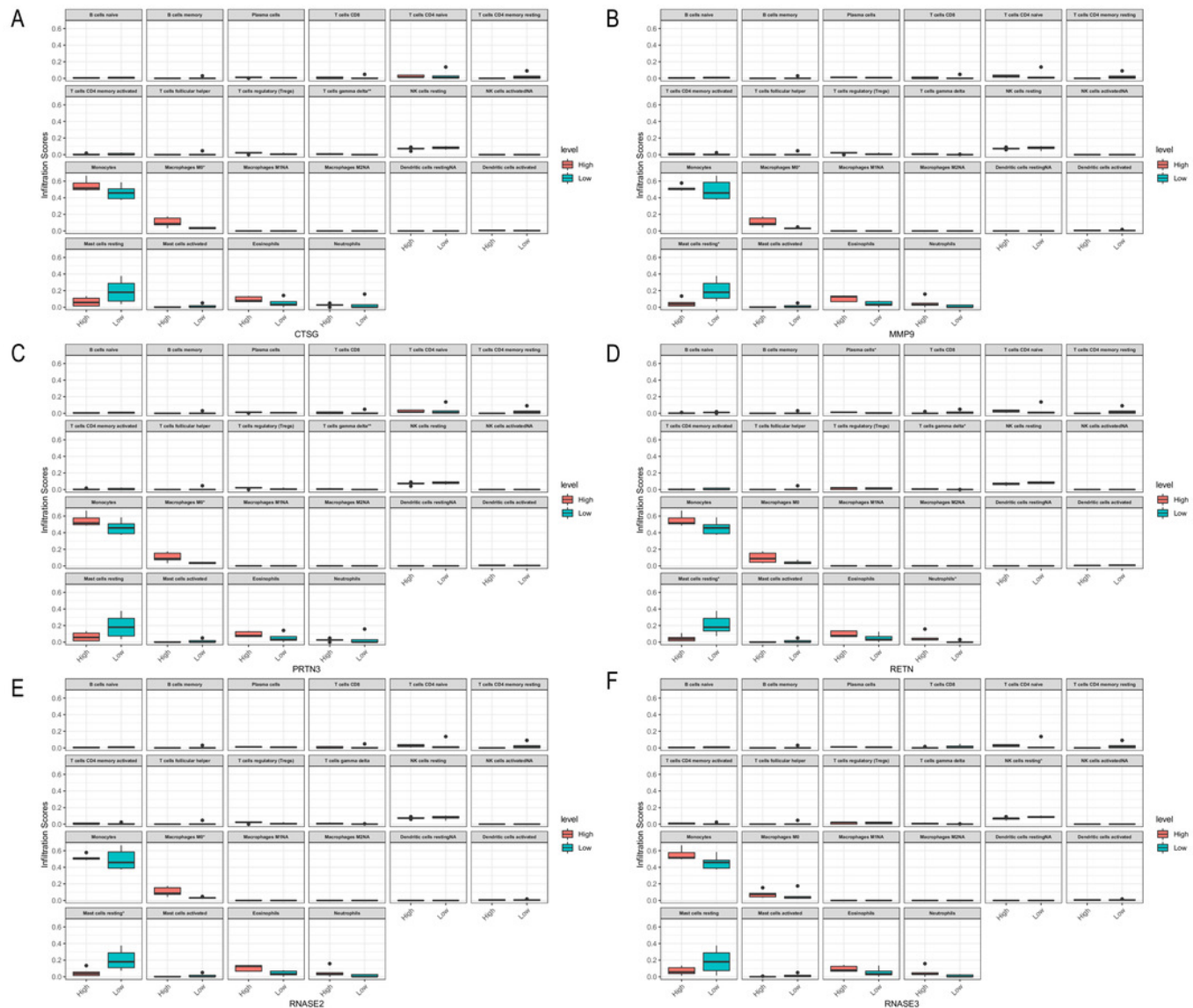
Figure 12

Relationships between the hub genes and immune cells in CML-CP-CML-BP

(A) CTSG. (B) MMP9. (C) PRTN3. (D) RETN. (E) RNASE2. (F) RNASE3.
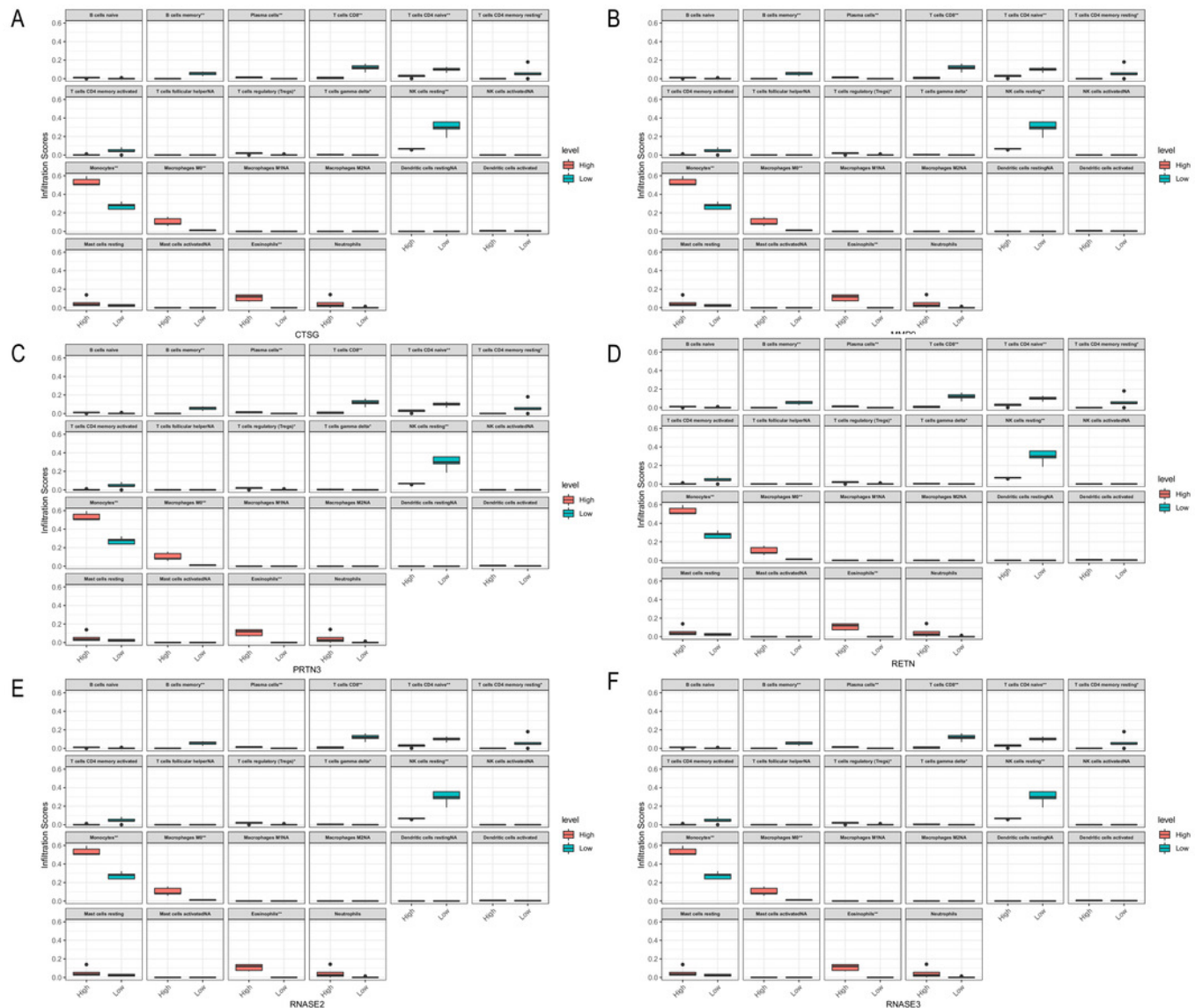
Figure 13

Related expression levels of hub genes

(A) CTSG. (B) MMP9. (C) PRTN3. (D) RETN. (E) RNASE2. (F) RNASE3. Significant differences were supposed at $^*P<0.05$; $^{**}P<0.01$; $^{***}P<0.001$.