Integrated bioinformatics analysis reveals dynamic candidate genes and signaling pathways involved in the progression and prognosis of diffuse large B-cell lymphoma

Alice Charwudzi [Equal first author, 1] , **Ye Meng** [Equal first author, 1] , **Linhui Hu** [1] , **Chen Ding** [1] , **Lianfang Pu** [1] , **Qian Li** [1] , **Mengling Xu** [1] , **Zhimin Zhai** [Corresp., 1] , **Shudao Xiong** [Corresp. 1]

[1] Department of Hematology/Hematological Lab, The Second Hospital of Anhui Medical University, Hefei, Anhui, China

Corresponding Authors: Zhimin Zhai, Shudao Xiong
Email address: zzzm889@163.com, xshdao@ahmu.edu.cn

Background. Diffuse large B-cell lymphoma (DLBCL) is a highly heterogeneous malignancy with varied outcomes. However, the fundamental mechanisms remain to be fully defined. **Aim.** We aimed to identify core differentially co-expressed hub genes and perturbed pathways relevant to the pathogenesis and prognosis of DLBCL. **Methods**. We retrieved the raw gene expression profile and clinical information of GSE12453 from the Gene Expression Omnibus (GEO) database. We used integrated bioinformatics analysis to identify differentially co-expressed genes. The CIBERSORT analysis was also applied to predict tumor-infiltrating immune cells (TIICs) in the GSE12453 dataset. We performed survival and ssGSEA (single-sample Gene Set Enrichment Analysis) (for TIICs) analyses and validated the hub genes using GEPIA 2 and an independent GSE31312 dataset. **Results.** We identified 46 differentially co-expressed hub module genes in the GSE12453 dataset. Gene expression levels and survival analysis found 15 differentially co-expressed core hub genes. The core genes prognostic values and expression levels were further validated in the GEPIA 2 database and GSE31312 dataset to be reliable ($p < 0.01$). The core genes' main KEGG (Kyoto encyclopedia of genes and genomes) pathway enrichments were Ribosome and Coronavirus disease - COVID-19. High expressions of the 15 core hub genes had prognostic value in DLBCL. The core genes showed significant predictive accuracy in distinguishing DLBCL cases from non-tumor controls, with the area under the curve (AUC) ranging from 0.992 to 1.00. Finally, CIBERSORT analysis on GSE12453 revealed immune cells, including activated memory CD4+ T cells and M0, M1, and M2- macrophages as the infiltrates in the DLBCL microenvironment. **Conclusion.** Our study found differentially co-expressed core hub genes and relevant pathways involved in ribosome and COVID-19 disease that may be potential targets for prognosis and novel therapeutic intervention in DLBCL.

1 **Integrated bioinformatics analysis reveals dynamic candidate genes and signaling**

2 **pathways involved in the progression and prognosis of Diffuse Large B-Cell Lymphoma.**

4 Alice Charwudzi[1#], Ye Meng[1#], Linhui Hu[1], Chen Ding[1], Lianfang Pu[1], Qian Li[1], Mengling Xu [1]

5 Zhimin Zhai[1], Shudao Xiong[1]

7 [#] AC and YM should be considered joint first authors; the authors contributed equally to this

8 work.

10 [1]Department of Hematology/Hematological Lab, The Second Hospital of Anhui Medical

11 University, Hefei 230601, Anhui Province, People's Republic of China

13 Corresponding Authors:

14 Prof Shudao Xiong[1]

15 678 Furong Rd, Hefei, Anhui Province, 230601, People's Republic of China

16 Email address: xshdao@ahmu.edu.cn or

17 Prof Zhimin Zhai[1]

18 678 Furong Rd, Hefei, Anhui Province, 230601, People's Republic of China

19 Email address:Email: zzzm889@163.com

23 **Abstract**

24 **Background.** Diffuse large B-cell lymphoma (DLBCL) is a highly heterogeneous malignancy

25 with varied outcomes. However, the fundamental mechanisms remain to be fully defined.

26 **Aim.** We aimed to identify core differentially co-expressed hub genes and perturbed pathways

27 relevant to the pathogenesis and prognosis of DLBCL.

28 **Methods.** We retrieved the raw gene expression profile and clinical information of GSE12453

29 from the Gene Expression Omnibus (GEO) database. We used integrated bioinformatics analysis

30 to identify differentially co-expressed genes. The CIBERSORT analysis was also applied to

31 predict tumor-infiltrating immune cells (TIICs) in the GSE12453 dataset. We performed survival

32 and ssGSEA (single-sample Gene Set Enrichment Analysis) (for TIICs) analyses and validated

33 the hub genes using GEPIA 2 and an independent GSE31312 dataset.

34 **Results.** We identified 46 differentially co-expressed hub module genes in the GSE12453

35 dataset. Gene expression levels and survival analysis found 15 differentially co-expressed core

36 hub genes. The core genes prognostic values and expression levels were further validated in the

37 GEPIA 2 database and GSE31312 dataset to be reliable (p<0.01). The core genes' main KEGG

38 (Kyoto encyclopedia of genes and genomes) pathway enrichments were Ribosome and

39 Coronavirus disease - COVID-19. High expressions of the 15 core hub genes had prognostic

40 value in DLBCL. The core genes showed significant predictive accuracy in distinguishing

41 DLBCL cases from non-tumor controls, with the area under the curve (AUC) ranging from 0.992

42 to 1.00. Finally, CIBERSORT analysis on GSE12453 revealed immune cells, including activated

43   memory CD4$^+$ T cells and M0, M1, and M2- macrophages as the infiltrates in the DLBCL

44   microenvironment.

45   **Conclusion.** Our study found differentially co-expressed core hub genes and relevant pathways

46   involved in ribosome and COVID-19 disease that may be potential targets for prognosis and

47   novel therapeutic intervention in DLBCL.

49

50

51

52

53

54

55

56

57

58

59

60

61
62

63

64

65

66 **Introduction**

67 Diffuse large B-cell lymphoma (DLBCL) is exceptionally heterogeneous and the most common

68 aggressive non-Hodgkin lymphoma (NHL) subtype in adults. It is increasingly appreciated that

69 its varied outcomes depend on the patients' clinical and biological features (Karube et al., 2018;

70 W. Liu et al., 2019; Luo, Gu, Wang, Chen, & Peng, 2018; Naresh et al., 2011). Despite several

71 reports on the mechanism of DLBCL, its pathogenesis characterized by multiple abnormalities at

72 different molecular levels remains unresolved. Its development and progression are multifaceted,

73 comprising various signaling pathways and driver genes. Despite improved clinical outcomes

74 with current therapies, such as rituximab-chemotherapy (R-CHOP), chimeric antigen receptor

75 (CAR)-T cell, and advancement in stem cell transplantation, over 40% of high-risk patients

76 relapse or develop the primary refractory disease. Mortality figures remain high (Karube et al.,

77 2018; Luo et al., 2018). Therefore, an in-depth understanding of disease biology could reveal

78 novel biomarkers of diagnostic and prognostic value. It will also facilitate the design of

79 alternative personalized therapeutic strategies for DLBCL.

80   Advances in gene profiling technologies, high-throughput data, and bioinformatics

81 databases make screening DLBCL for differentially co-expressed genes indispensable,

82 particularly when integrated with personalized genomic profile data (Lui et al., 2015; van Dam,

83 Võsa, van der Graaf, Franke, & de Magalhães, 2018). Recently, Liu et al. identified eleven genes

84 associated with endometrial cancer progression and prognosis by comprehensive bioinformatics

85 analysis (J. Liu et al., 2019). Zhou et used CIBERSORT and other bioinformatics analyses for

86 colon cancer. They found that the tumor microenvironment (TME) was abundantly enriched with

87 M0 and M2 macrophages, activated memory CD4$^+$ T cells, and other immune cells that could

88   play crucial roles as biomarkers (R. Zhou et al., 2019). However, few integrated bioinformatics

89   studies have compared the gene expression profile of DLBCL with non-cancer controls.

90   Thus, we downloaded the Gene Expression Omnibus (GEO) raw dataset of GSE12453 and

91   compared 11 DLBCL cases with 24 non-cancer controls (non-neoplastic B lymphocytes isolated

92   from blood or tonsils). We performed a series of screens and analyses, including filtering off

93   differentially expressed genes (DEGs), enrichment analysis, and co-expression analysis to

94   determine hub genes of clinical significance to DLBCL. We identified 15 differentially co-

95   expressed core hub genes associated with the prognosis of DLBCL (RPS24, RPS21, RPL31,

96   RPL30, RPS17, MRPS28, FAU, RPS25, RPL22L1, NDUFA6, CXCL9, CCL4, MRPL33,

97   HEBP1, and RPL11). The KEGG (Kyoto encyclopedia of genes and genomes) analysis

98   associated most of the genes with Ribosome and Coronavirus disease - COVID-19 pathways.

99   Validation in the GEPIA 2 database and GSE31312 dataset revealed that the core genes had

100 consistent expression levels and were reliable. Receiver operating characteristic (ROC) curves

101 plotted demonstrated that the core genes could be potential diagnostic biomarkers. The identified

102 genes could play critical roles in diagnosis, prognosis and help establish a foundation for

103 developing or identifying novel targeted therapies for DLBCL.

104

105 **Materials and methods**

106 **Data collection**

107 We evaluated and downloaded the raw gene expression profiles from the National Center for

108 Biotechnology Information - Gene Expression Omnibus (NCBI-GEO) database

109 (https://www.ncbi.nlm.nih.gov/geo/). The series was based on GPL570 [HG-U133_Plus_2]

110 Affymetrix Human Genome U133 Plus 2.0 Array. GSE12453 (Brune et al., 2008) was used to

111 identify differentially expressed and co-expressed genes. It was also used to predict tumor-

112 infiltrating immune cells (TIICs). It contained 11 DLBCL cases and 25 normal controls. The

113 controls were non-neoplastic B lymphocytes isolated from healthy donors' blood or tonsils from

114 routine tonsillectomy patients. Similar expression profiles on the selected GPL570 platform

115 contained normal but reactive controls or non-human controls, such as cell lines or limited

116 controls, and were excluded. We used the GSE31312 (Visco et al., 2012) with 498 DLBCL cases

117 for survival analysis and identifying immune cells infiltrating the TME.

118

119 **Study design and data pre-processing**

120 The GSE12453 CEL file was pre-processed using the Affymetrix package (Gautier, Cope,

121 Bolstad, & Irizarry, 2004) in R software (https://www.r-project.org/) version 4.0.2 (Team, 2019).

122 The procedures included background correction, log_2 transformation, followed by quantile

123 normalization. We performed a standard quality assessment, including scaling factors and NUSE

124 plots, and hierarchical clustering to identify outliers. The study's design flowchart is shown in

125 *Fig S1.*

126

127 **Screening of differentially expressed genes (DEGs)**

128 The DEGs between DLBCL cases and non-cancer controls were screened with R package limma

129 (https://www.bioconductor.org/), Release 3.11 (Ritchie et al., 2015). The cutoff criteria were

130 p<0.05, |log_2fold change (FC)| >1.0. The pheatmap package was used to generate hierarchical

131 clustering and ggplot2 (Wickham, 2016) to show volcano plot in R. We used the resulting data

132 output tables that included gene ID, log_2FC, unadjusted and adjusted p-values in subsequent

133 analyses.

134 **Functional and pathway enrichment analysis**

135 We investigated the functional roles and pathway signaling relevance of the DEGs. The gene

136 ontology (GO) and Kyoto encyclopedia of genes and genomes (KEGG) pathway enrichment

137 analysis was performed using the R clusterprofiler (https://bioconductor.org/), Release 3.12 (Yu,

138 Wang, Han, & He, 2012). A p-value < 0.05 was considered significantly enriched. The GO

139 categories included biological process (BP), molecular function (MF), and cellular component

140 (CC).

141

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

143 We used the normalized expression dataset of GSE12453 for the GSEA (www.gsea-

144 msigdb.org/gsea/index.jsp), version 4.1.0. We followed the recommended protocol. This

145 included "gct" and "cls" file formats for the expression dataset and phenotype labels, respectively

146 (Subramanian et al., 2005). Significant gene sets had false discovery rate (FDR) <0.25 and a

147 nominal p-value < 0.05.

148

149 **Protein-protein interaction (PPI) network of DEGs and hub modules selection**

150 We applied the STRING database (https://string-db.org), version 11.0 (Szklarczyk et al., 2019),

151 to assess and map the identified DEGs into a human PPI network. Then, we uploaded the

152 resulting data output into Cytoscape software (https://cytoscape.org/), version 3.8.2 (Shannon et

153 al., 2003), and used MCODE with default parameters (Bader & Hogue, 2003) and the Cytohubba

154 (Chin et al., 2014) plug-ins to select significant modules and top-ranked genes. Nodes and edges

155 represented genes and their interactions.

156

157   **Predicting tumor-infiltrating immune cells (TIICs)**

158   We applied the Cibersort algorithm (Newman et al., 2019) in R to predict TIICs using the

159   normalized GSE12453 dataset, according to the CIBERSORT instructions. We used data with a

160   p-value < 0.05 for further analysis.

161

162   **Co-expression network construction and identification of modules related to DLBCL**

163   We conducted gene co-expression analysis on the processed GSE12453 data using the weighted

164   correlation network analysis (WGCNA) (Peter Langfelder & Horvath, 2008; P. Langfelder &

165   Horvath, 2012) in R (Team, 2019). We followed the standard protocols, including quality control

166   procedures. We performed a power β transformation on the computed Pearson correlation matrix

167   to ensure a scale-free topology. The minimum number of module genes was set at 30. The

168   WGCNA R package then generated a co-expression network from the resulting adjacency

169   matrix. We applied the dynamic Tree Cut package (Team, 2019) to create the co-expression

170   modules from the color-coded hierarchical clustering dendrogram. We assessed clinical module-

171   trait relationships with Pearson's correlation. Gene significance (GS) and module membership

172   (MM) were also analyzed for their correlation in modules. Statistically significant modules were

173   defined as p<0.05.

174

175   **Identifying differentially co-expressed hub module genes**

176   The hub genes were identified from the intersection between the DEGs and genes significant in

177   the WGCNA modules. We then analyzed the hub genes with the STRING and Cytoscape

178   databases. We further conducted functional and pathway analysis on the hub genes to determine

179   the relevant genes that impact DLBCL.

180 **Validation and survival analysis of core hub genes**

181 We validated the expression levels of the hub genes in the GEPIA2 (gepia2.cancer-

182 pku.cn/#index) database (Tang, Kang, Li, Chen, & Zhang, 2019). We retrieved the pre-processed

183 quantile normalized series matrix file (GSE31312, Affymetrix HG-U133 Plus 2.0 GeneChips)

184 containing 498 de-novo adults DLBCL from the NCBI-GEO database. We used it for the

185 prognostic value analysis (overall survival (OS) and progression-free survival (PFS)). We plotted

186 Kaplan-Meier survival curves with ggplot2 in R. The genes with p-value < 0.05 were selected as

187 core hub genes. Also, receiver operating characteristic (ROC) curves were plotted with the

188 pROC R package (Team, 2019) to validate the diagnostic value of the core genes. Then, the

189 human protein atlas (proteinatlas.org/) was used to confirm their protein expressions in some

190 selected lymph node samples (Uhlen et al., 2017).

191

192 We used the ssGSEA (single-sample Gene Set Enrichment Analysis) package (Subramanian et

193 al., 2005) in R on the GSE31312 dataset to investigate immune-associated core hub genes in the

194 TME.

195

196 In addition, we also employed the GEPIA 2 and Oncomine (Rhodes et al., 2004) databases to

197 determine the expression levels and relevance of the core hub genes in other tumors.

198

199 **Results**

200 **Differentially expressed genes (DEGs)**

201 According to standard protocols, one poor quality control sample (GSM312887) was excluded

202 from the GSE12453 dataset after pre-processing. We identified 1260 DEGs between the 11

203   DLBCL and 24 non-tumor controls. They comprised 1014 up-regulated and 246 down-regulated

204   genes (p < 0.05 and |log$_2$FC|>1). The gene list, Affymetrix probe ID, and log FC are shown in

205   (Excel S1). The expression patterns of the DEGs are shown in *Fig 1*. *Figure 1A* shows the

206   volcano plot of all expressed genes. The DLBCL cases showed a distinctive gene expression

207   profiling. As shown in *Fig. 1B*, the heatmap of the top 169 DEGs with |log$_2$FC| > 2 suggested

208   that the identified DEGs expression levels could differentiate DLBCL from non-tumor samples.

209   We utilized the CytoHubba application in Cytoscape, employing five calculation methods: the

210   Maximal Clique Centrality (MCC), Maximum Neighborhood Component (MNC), Degree, Edge

211   Percolation Component (EPC), and EcCentricity to rank the top 250 DEGs. The genes from the 5

212   methods were intersected using the Venn diagram software

213   (http://bioinformatics.psb.ugent.be/webtools/Venn/). Most of the intersecting genes (*Fig. 1C*)

214   were associated with significant and valuable pathways such as the proteasome, spliceosome,

215   and viral protein interaction with cytokine and cytokine receptor (*Fig. S2*). Intersecting genes are

216   common genes with a high degree of interconnection and are more likely to represent key

217   candidate genes with important biological regulatory functions.

218

219   **Functional and pathway analysis**

220   We applied the enrichGO or enrichKEGG function of the clusterProfiler package (Yu et al.,

221   2012) in R to investigate the biological functions of all the DEGs (p < 0.05). The GO and KEGG

222   pathway analysis results are shown in *Fig. 2*. In GO analysis, the DEGs were primarily enriched

223   in ATP synthesis coupled electron transport and mitochondria ATP synthesis coupled electron

224   transport for biological processes (BP). Their cellular components (CC) were mainly related to

225   the mitochondria protein complex, mitochondria inner membrane, etc. Their molecular functions

226 (MF) consisted of structural constituents of ribosomes and NADH dehydrogenase activity (*Fig.*

227 *2A*). The KEGG pathway analysis showed significant enrichment in ribosome and oxidative

228 phosphorylation (*Fig. 2B*). Strikingly, the DEGs involved in the KEGG pathway and GO

229 enrichment analysis were mainly upregulated genes, with few down-regulated genes.

230 Further analysis showed that the down-regulated genes were enriched in B cell activation (adjust

231 p-value = 0.003) for GO-BP; our data showed no other significant GO and KEGG enrichment for

232 the down-regulated genes. The GO and KEGG enrichment analysis for the upregulated genes

233 was similar to the analysis we performed earlier. The GO and KEGG analysis classifies genes

234 into functional categories to help understand their functions and regulatory pathways. Hence,

235 with additional investigations, the genes in these pathways might throw more light on the

236 pathogenesis of DLBCL.

237

238 **Gene Set Enrichment Analysis (GSEA) of DLBCL expression dataset (GSE12453)**

239 We performed a GSEA analysis to compare the DLBCL and non-tumor controls' expression

240 profiles to understand better the biological functions of the relevant genes discovered. We

241 analyzed all the qualified genes in the GSE12453 expression dataset. The KEGG output from the

242 GSEA was similar to our previous pathway analysis and confirmed our earlier results. The

243 Hallmark gene sets showed immune and metabolic-related signaling predominance, including

244 estrogen response late, epithelial-mesenchymal transition, and UV response up (*Table S1*).

245 Interestingly, genes defining late response to estrogen, epithelial-mesenchymal transition (such

246 as in fibrosis and metastasis), and genes upregulated in response to ultraviolet (UV) radiation

247 have not been fully elucidated in DLBCL. However, evidence suggests these genes have

248   essential roles in oncogenesis. These findings provide evidence that can drive future research

249   with therapeutic implications.

250

251   **Predicting tumor-infiltrating immune cells (TIICs)**

252   Our GO, KEGG, and GSEA analyses showed that the DEGs were enriched in some immune-

253   related biological functions. Immune cell infiltration into tumors plays an essential role in

254   tumorigenesis and metastasis. So, we applied the CIBERSORT algorithm to the GSE12453

255   dataset to predict immune cells infiltrating the TME. Among the immune subsets analyzed,

256   activated memory CD4$^+$ T cells, CD8$^+$ T cells, and M0, M1, and M2 macrophages were the most

257   represented cell fractions within the DLBCL microenvironment (*Fig. 3 A, B*). The correlations

258   among the TIICs ranged from high to negligible (*Fig. S3*). Regulatory T cells (Tregs) showed a

259   moderate negative correlation with activated memory CD4$^+$ T cells. However, M0 macrophages

260   had a high positive correlation with activated memory CD4$^+$ T cells and M1 macrophages. The

261   varied infiltrating immune cell types could be reflective of the complexity and the unusual

262   behavior of DLBCL. Uncommitted macrophages (M0) can polarize into the M1 (considered

263   tumoricidal) and M2 (pro-tumorigenic) to show paradox effects on tumor prognosis (Dancsok et

264   al., 2020). Memory B cells and activated dendritic cells were the most represented fractions in

265   the non-tumor controls (*Fig. 3 A*). The findings suggest that TIICs may be closely associated

266   with clinical outcomes. Future studies, including their correlation to DLBCL disease stages, will

267   be meaningful, particularly for immunotherapy.

268

269   **Protein-protein interaction (PPI) network and hub modules establishment**

270   We constructed the PPI network of all DEGs using the STRING database's multiple proteins

271   function to determine genes likely to perform biological functions together. The highest

272   confidence of 0.9 was set with unconnected nodes taken out. As shown in *Fig. 4A,* it yielded 609

273   nodes and 8583 edges. These genes were highly inter-connected than expected (PPI enrichment

274   p-value <1.0e-16). We observed 6 important clusters with a k-score > 10 when we analyzed the

275   tab-separated values (tsv) file with Cytoscape's MCODE. The largest cluster (#1) had the highest

276   score of 33.17 (*Fig. 4B*); it mainly comprised Ribosome genes (FDR= 3.49e-83) in KEGG

277   analysis. Cluster 2 (*Fig. 4C*) was enriched in genes associated with oxidative phosphorylation

278   (FDR, 7.23e-56) and Parkinson's disease (FDR, 2.62e-55). Cluster 3 (*Fig. 4D*) genes were

279   mainly involved in chemokine signaling (FDR, 1.06e-22). We identified 109 highly connected

280   DEGs (Excel S1) from these top 3 clusters. These DEGs could play essential roles in DLBCL, so

281   they were selected for further hub gene screening.

282

283   In this study, the down-regulated DEGs were not part of the constructed PPI, so we determined

284   their relevance in DLBCL. We ranked the top 60 DEGs (30 up- and 30 down-regulated) by

285   log_2FC and analyzed them in the STRING database (*Fig. S4A*). The only down-regulated DEG in

286   the network built was identified upregulated when verified in the GEPIA2 database. To further

287   investigate the down-regulated DEGs' functional relatedness, a PPI was constructed for all the

288   down-regulated DEGs (*Fig. S4B*). We found that the down-regulated genes were enriched in B

289   cell activation for GO-BP, shown in (*Fig. S4C*). Finally, we used the degree method of the

290   CytoHubba application to predict the top 250 important genes (*Fig. 4E*). Genes with a high

291   degree of centrality are vital since they have many direct interacting gene partners. If confirmed,

292   these critical findings could improve the general understanding and the potential causes of

293   variation in the clinical prognosis of DLBCL.

294

295   **Weighted gene co-expression network (WGCNA) analysis**

296   To identify co-expression modules that could share similar biological functions or regulatory

297   mechanisms with clinical relevance to DLBCL, we applied the WGCNA package (Peter

298   Langfelder & Horvath, 2008; P. Langfelder & Horvath, 2012) in R (Team, 2019). The

299   GSE12453 dataset we processed was used. We carried out quality control procedures, including

300   inspecting good genes and sample hierarchical clustering to detect potential outliers but no

301   obvious outliers (*Fig. 5 A*). The 35 samples yielded two main clusters. We applied the WGCNA

302   on the top 25% of the 21654 expressed genes ranked by the largest variance. To satisfy a scale-

303   free network topology, we choose the soft-threshold power β of 8 with R2 = 0.86 (*Fig. 5 B, C*).

304   Hierarchical clustering and the dynamic tree-cutting yielded 18 modules of co-expressed genes

305   (*Fig. 5D*). Finally, we visualized the top 1000 significantly expressed genes with a heatmap (*Fig.*

306   *5E*); they represent interesting genes for further analysis.

307

308   To investigate the molecular mechanisms of the traits, we correlated each Module Eigengenes

309   (ME) to disease status (DLBCL and non-tumor controls). The results are shown in *Fig. 6*. The

310   ME turquoise and green (*Fig. 6 A, B*) containing 292 and 72 genes, respectively, strongly

311   correlated with DLBCL. ME dark magenta with 5 genes had the strongest negative correlation.

312   The cut-off was set at gene significance (GS) value > 0.8, and absolute Module Membership

313   (MM) value ≥ 0.7. Besides, the GS versus MM plots for these 3 modules were highly correlated

314   (*Fig. 6C*), reflecting their high association with DLBCL. We selected these 3 clinically

315   significant modules with the 369 high connectivity genes (gene list shown in Excel S1) for

316   further analysis. The highly connected genes are often the most important (central) elements of

317   the respective modules and tend to play key roles in the biological processes. The ME genes are

318 listed in *Excel S1*. Altogether, these co-expressed genes might provide new clues to understand

319 the biology of DLBCL in the future.

320

321 **PPI and functional enrichment analysis of the WGCNA relevant modules**

322 The 369 high connectivity genes from the 3 relevant modules were filtered in the STRING

323 database followed by the Cytoscape; the network yielded a PPI with 195 nodes and 1579 edges.

324 The PPI and gene list are detailed in *Fig. S5 and Excel S1*. These 195 genes were considered

325 functionally important. As presented in (*Fig. 7*), functional annotation revealed that these genes

326 were involved in viral transcription and viral gene expression in the BP category. In KEGG

327 analysis, the genes were primarily enriched in ribosome, coronavirus disease - COVID-19, and

328 oxidative phosphorylation. The identified pathways were roughly consistent with that of the

329 DEGs. These processes and signaling pathways are usually disrupted in cancer and could provide

330 an insight into the pathogenesis of DLBCL.

331

332 **Identification of hub genes and pathways**

333 Eventually, we identified 46 important differentially co-expressed genes by the Venn diagram

334 software, as shown in (*Fig. 8A*). These 46 genes were common between the DEGs and WGCNA

335 hub module genes and were regarded as hub genes. We re-analyzed the 46 genes with the

336 STRING and Cytoscape databases, and the PPI is shown in (*Fig. 8B*). Their GO and KEGG

337 enrichments in the R software (*Table 1*) were similar to the other analyses. The KEGG common

338 genes in the ribosome, COVID-19, and oxidative phosphorylation pathways are shown in *Table*

339 *2*. The above pathway genes play essential roles in metabolic reprogramming and tumor-

340 promoting inflammation of cancer and warrant further studies.

341

342 **Validation of expressions and prognostic analysis of core hub genes**

343 We applied GEPIA 2 to validate the reliability and authenticity of the 46 hub genes in the cancer

344 genome atlas (TCGA) dataset. We identified 44 prognostic genes with higher expression

345 (consistent with that in the GSE12453 dataset) in DLBCL tissues than the non-tumor control

346 tissues (p<0.01) (*Fig. 9*) (15 genes shown) and (*Excel S1*). Kaplan-Meier survival analysis on the

347 GSE31312 showed 15 of these genes (p<0.05) correlated with patient outcomes (*Fig. 10 & Table*

348 *S2*). Except for RPL11, the patients with high expressions had significantly shorter 5-year OS

349 and PFS, suggesting these genes are potential oncogenes and have a role in DLBCL development

350 and/or progression. The 15 genes (*Fig. 11A*) were considered as the core hub genes. Moreover,

351 ROC curve analysis for their diagnostic potentials obtained AUCs ranging from 0.992 to 1.00,

352 indicating optimal performance to accurately differentiate DLBCL from non-tumor control cases

353 (*Fig. S6*). Also, immunohistochemistry data from the human protein atlas (HPA) database

354 demonstrated the protein expressions of some of the genes in some lymph node samples with

355 cytoplasmic/ membranous localization (*Fig.* S7, 4 genes shown); data were retrieved from

356 https://www.proteinatlas). The genes included RPS21, MRPS28, RPL31, and RPL30. As

357 expected, they would be involved in metabolic pathways such as glycolysis and processes

358 including signal transduction and cell division.

359 The core hub genes' functional annotation was mainly associated with Ribosome and

360 Coronavirus disease - COVID-19 (*Fig. 11 B, C; Table 3*). To assess the tumorigenic potentials of

361 the COVID-19 genes regarding immune cells infiltrating the TME, the ssGSEA analysis was

362 applied (GSE31312). As shown in *Fig. S8*, 7 out of the 9 COVID-19 pathway genes negatively

363 correlated with mast cells, 5 with immature dendritic cell (iDC), and 3 genes negatively

364   correlated with plasmacytoid DC. RPL30, RPL31, RPS25, and FAU positively correlated with

365   tumor-infiltrating lymphocytes (TIL) and macrophages. RPL30, RPL31 correlated with Tregs.

366   These infiltrating immune cells may be involved in regulating tumor proliferation, dormancy,

367   and drug resistance.

368

369   Finally, we determined whether the core genes were upregulated in other tumors. The GEPIA 2

370   database revealed that all the 15 core genes were upregulated in thymoma (THYM), and 11

371   genes were upregulated in testicular germ cell tumors (TGCT). Notably, RPL30 and FAU genes

372   were consistently upregulated in all 6 different cancer types identified (Table S3). In the

373   Oncomine database, some of the core genes were upregulated in various lymphoma datasets and

374   other cancers, including Sarcoma (*Fig. S9*). The results suggest that the upregulation of these 15

375   hub genes may not be limited to DLBCL.

376

377   **Discussion**

378   Diffuse large B-cell lymphoma (DLBCL) remains a significant clinical challenge; over 30% of

379   patients are not cured (Pasqualucci & Dalla-Favera, 2018; Yi et al., 2020). So far, no functional

380   assays capable of screening exit, so effective management is required once diagnosed. Hence,

381   identifying unique gene signatures and regulatory pathways related to its pathogenesis and

382   prognosis is meaningful. Here, we examined the gene expression profile of GSE12453 to find

383   dysregulated common core hub genes and pathways to help further understand DLBCL

384   pathogenesis and provide potential biomarkers.

385

386   Integrated bioinformatics analysis, gene expression levels, and survival analysis identified 15

387   differentially co-expressed core hub genes linked to DLBCL pathogenesis. The genes included

388 RPS24, RPS21, RPL31, RPL30, RPS17, MRPS28, FAU, RPS25, RPL22L1, NDUFA6, CXCL9,

389 CCL4, MRPL33, HEBP1, and RPL11. Their primary KEGG enrichment was ribosome and

390 coronavirus disease - COVID-19, which was in line with the other analyses. The construction of

391 ROC curves yielded very high AUC values suggesting the genes could accurately distinguish

392 between DLBCL and non-tumor control cases and might be potential biomarkers. In addition,

393 experimentally derived data from the HPA by IHC indicated the protein expression of some of

394 the genes in some lymph node samples. RPS21, MRPS28, RPL31, and RPL30 showed relatively

395 higher protein expressions in some DLBCL and other malignant lymphoma tissues than the

396 averaged expressions in normal tissues, though not significant. The HPA database sample size

397 was limited; however, the HPA experimental findings can be extended to DLBCL and other

398 lymphomas, thus providing a valuable basis for medical and biological research.

399

400 Lastly, most of the core genes were upregulated in different cancer types. Cancer is a complex

401 disease, so the genes might have similar or different prognostic roles in these tumors. The overall

402 survival data from GEPIA 2 demonstrated that low levels of FAU, RPS17, and RPS 24 were

403 significantly associated with shorter survival, while high CCL4 was significantly associated with

404 shorter survival in thymoma patients. Nonetheless, the genes' potential biological and clinical

405 relevance is not restricted to only DLBCL. These genes could be prognostic markers and

406 therapeutic targets across different tumor types, particularly for patients with multiple coexisting

407 tumors.

408

409 Little is known experimentally about the roles of most of the core genes proteins in DLBCL.

410 However, dysregulated ribosomal proteins have been reported to play various critical roles in

411   other tumors (Wang et al., 2015). Among our ribosome genes, over-expressed RPS21 promoted

412   prostate cancer (PCa) cell proliferation, migration, and invasion, inhibited PCa cell apoptosis,

413   and was suggested as a promising biomarker, with a potential application in diagnosis or

414   treatment (Liang et al., 2019). The 8q-mapped RPL30 gene was associated with adverse

415   outcomes in Medulloblastoma patients (De Bortoli et al., 2006). RPS24 significantly promoted

416   colorectal cancer (CRC) cells' proliferation rate and increased CRC risk in patients (Zou et al.,

417   2020). The knockdown of RPS24 inhibited cell proliferation and cell migration in human CRC

418   cell lines and was recommended as a biomarker (Wang et al., 2015). A study also implicated

419   MRPS28 in the molecular pathogenesis of bladder cancer (Liu et al., 2021).

420

421   The coronavirus disease - COVID-19 and viral transcription enrichments agree with recent

422   studies implicating various viruses (Gandhi et al., 2020) (Fedoriw et al., 2020) in the

423   development and progression of DLBCL subtypes. Most DLBCL patients have an underlying

424   immune dysfunction and can easily get viral infections. Viruses such as COVID 19 could

425   manipulate the function of the COVID 19 related genes in the TME. Besides, the Gene Cards

426   database (https://www.genecards.org/) demonstrated all the covid 19 pathway genes (RPS24,

427   RPS21, RPL31, RPL30, RPS17, FAU, RPS25, RPL22L1, RPL11) are related to viral mRNA

428   translation (Stelzer et al., 2016). Additionally, the DAVID database (Protein interactions)

429   associated 5 COVID 19 genes (RPL30, RPS17, RPS 24, RPS 25, and FAU) with HIV

430   interactions (Huang da, Sherman, & Lempicki, 2009).

431

432   Recently, hematological malignancies (HM) patients were reported to have a more severe

433   COVID-19 trajectory than patients with solid organ tumors (Lee et al., 2020). A significant

434   number of the COVID 19 related genes were upregulated in various lymphomas and some

435   multiple myeloma datasets *(Fig. S9)*. Most of our COVID-19 pathway genes showed some

436   correlations with immune infiltrates such as TIL and macrophages. An anti-viral immune

437   response can have protective effects with improved survival in coronavirus infection, but

438   excessive inflammation can be harmful ("cytokine storm"). High pro-inflammatory macrophage

439   (M1) and low CD8[+] T cells were observed in the microenvironment of severe/critical COVID-19

440   patients (Liao et al., 2020). Higher expression of CXCL9 in COVID-19 patients than healthy

441   controls and higher levels of CCL4 in severe COVID-19 patients were also found (Liao et al.,

442   2020). These are partly consistent with our data and previous knowledge on various cancers

443   (Brune et al., 2008; Chang et al., 2013; De la Fuente López et al., 2018). Not much is known

444   about COVID-19 and most cancers, including DLBCL pathogenesis. However, it is tempting to

445   speculate that COVID-19 infection, together with the COVID-19 related genes, could increase

446   macrophage polarization to M1 (hyper-inflammatory response) to worsen prognosis (He et al.,

447   2020; Passamonti et al., 2020; Shah et al., 2020). But, the mechanism is unclear, and limited data

448   on the topic did not permit detailed discussion. However, extensive genome-related studies are

449   required to verify this association between COVID 19 and DLBCL; these genes could provide a

450   basis to identify effective preventive and therapeutic strategies.

451

452   Clinical implication analysis in the GEPIA2 database showed that the core hub genes were

453   significantly overexpressed in DLBCL. The high expressions of 14 (93%) were negatively

454   associated with prognostic outcomes (worse OS and PFS times). These emphasize their potential

455   role as oncogenes and could be utilized as prognostic indicators for DLBCL. The high

456   expression of RPL11 might be associated with a favorable clinical outcome (Kawahata et al.,

457   2020; Kayama et al., 2017). They offer unique opportunities for further investigation.

458   In addition to the ribosome (translation), the 46 hub genes were significantly represented in

459   oxidative phosphorylation (OxPhos) and mitochondria inner membrane. Growing evidence

460   suggests cancer is primarily a mitochondrial metabolic disease that exhibits altered energy

461   production and dysregulated metabolic crosstalk (Yin et al., 2019) (Norberg et al., 2017). Their

462   inhibition has demonstrated anti-cancer efficacy (Martínez-Reyes et al., 2020) (Norberg et al.,

463   2017). Thus, with further studies, these metabolic genes could be rational targets, especially for

464   the metabolically coupled and OxPhos-DLBCL subsets, and help understand the metabolic

465   differences in DLBCL.

466

467   The few bioinformatics analyses on DLBCL focused on the subtypes (Huang, Liu, & Shen,

468   2019; L. Zhou et al., 2020) or clinical features (Xiao, Wang, & Bai, 2020). However, some

469   dysregulated genes specific to DLBCL versus non-tumor controls cannot distinguish the

470   subtypes (Huang et al., 2019). Moreover, most related studies that focused on DLBCL and non-

471   cancer controls were based on DEGs (Huang et al., 2019; Luo et al., 2018) and discovered

472   entirely different core hub genes. However, the complexity of DLBCL and the emergence of

473   novel targeted therapies warrants more predictive personalized biomarkers for precision

474   medicine. To our knowledge, no integrated bioinformatics analysis on DLBCL and non-tumor

475   controls has so far been reported on the common core hub genes found and immune cells

476   associations. Thus, our finding is novel.

477   One potential limitation of this study is the lack of experimental validation. However, this

478   analysis provides a theoretical basis for our future work, which will focus on experimental

479   verification. Second, an individual study with limited DLBCL cases was used, but the hub genes,

480 pathways, and immune cells infiltrate identified are relevant to the pathogenesis of DLBCL and

481 cannot be ignored. However, our results should be interpreted with caution.

482

483 **Conclusions**

484 We used the integrated bioinformatics method to highlight the critical roles of differentially co-

485 expressed core hub genes and relevant pathways in DLBCL. We identified some immune-related

486 core hub genes linked to DLBCL pathogenesis. The core genes' main KEGG pathway

487 enrichments were Ribosome and Coronavirus disease - COVID-19. Their verification in GEPIA

488 2 showed they were reliable. Nevertheless, most of the core genes were upregulated in different

489 cancer types and hold potential biological and clinical relevance in cancers. Thus, the identified

490 genes could be potential targets for prognosis and therapeutic intervention in DLBCL and may

491 provide insight into the pathogenetic mechanisms in DLBCL.

492

493 **References**

494 Bader, G. D., & Hogue, C. W. (2003). An automated method for finding molecular complexes in large
495   protein interaction networks. *BMC Bioinformatics, 4*, 2. doi:10.1186/1471-2105-4-2
496 Brune, V., Tiacci, E., Pfeil, I., Döring, C., Eckerle, S., van Noesel, C. J. M., . . . Küppers, R. (2008). Origin and
497   pathogenesis of nodular lymphocyte-predominant Hodgkin lymphoma as revealed by global
498   gene expression analysis. *The Journal of experimental medicine, 205*(10), 2251-2268.
499   doi:10.1084/jem.20080809
500 Chang, K. P., Wu, C. C., Fang, K. H., Tsai, C. Y., Chang, Y. L., Liu, S. C., & Kao, H. K. (2013). Serum levels of
501   chemokine (C-X-C motif) ligand 9 (CXCL9) are associated with tumor progression and treatment
502   outcome in patients with oral cavity squamous cell carcinoma. *Oral Oncol, 49*(8), 802-807.
503   doi:10.1016/j.oraloncology.2013.05.006
504 Chin, C.-H., Chen, S.-H., Wu, H.-H., Ho, C.-W., Ko, M.-T., & Lin, C.-Y. (2014). cytoHubba: identifying hub
505   objects and sub-networks from complex interactome. *BMC systems biology, 8 Suppl 4*(Suppl 4),
506   S11-S11. doi:10.1186/1752-0509-8-S4-S11
507 Choesmel, V., Fribourg, S., Aguissa-Touré, A. H., Pinaud, N., Legrand, P., Gazda, H. T., & Gleizes, P. E.
508   (2008). Mutation of ribosomal protein RPS24 in Diamond-Blackfan anemia results in a ribosome
509   biogenesis disorder. *Hum Mol Genet, 17*(9), 1253-1263. doi:10.1093/hmg/ddn015
510 Dancsok, A. R., Gao, D., Lee, A. F., Steigen, S. E., Blay, J. Y., Thomas, D. M., . . . Demicco, E. G. (2020).
511   Tumor-associated macrophages and macrophage-related immune checkpoint expression in
512   sarcomas. *Oncoimmunology, 9*(1), 1747340. doi:10.1080/2162402x.2020.1747340

513   De Bortoli, M., Castellino, R. C., Lu, X. Y., Deyo, J., Sturla, L. M., Adesina, A. M., . . . Kim, J. Y. (2006).
514   Medulloblastoma outcome is adversely associated with overexpression of EEF1D, RPL30, and
515   RPS20 on the long arm of chromosome 8. *BMC Cancer, 6*, 223. doi:10.1186/1471-2407-6-223
516   De la Fuente López, M., Landskron, G., Parada, D., Dubois-Camacho, K., Simian, D., Martinez, M., . . .
517   Hermoso, R. M. (2018). The relationship between chemokines CCL2, CCL3, and CCL4 with the
518   tumor microenvironment and tumor-associated macrophage markers in colorectal cancer.
519   *Tumour Biol, 40*(11), 1010428318810059. doi:10.1177/1010428318810059
520   Fedoriw, Y., Selitsky, S., Montgomery, N. D., Kendall, S. M., Richards, K. L., Du, W., . . . Gopal, S. (2020).
521   Identifying transcriptional profiles and evaluating prognostic biomarkers of HIV-associated
522   diffuse large B-cell lymphoma from Malawi. *Mod Pathol, 33*(8), 1482-1491. doi:10.1038/s41379-
523   020-0506-3
524   Gandhi, M. K., Hoang, T., Law, S. C., Brosda, S., O'Rourke, K., Tobin, J. W. D., . . . Keane, C. (2020). EBV-
525   tissue positive primary CNS lymphoma occurring after immunosuppression is a distinct
526   immunobiological entity. *Blood*. doi:10.1182/blood.2020008520
527   Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). affy—analysis of Affymetrix GeneChip data at
528   the probe level. *Bioinformatics, 20*(3), 307-315. doi:10.1093/bioinformatics/btg405
529   He, W., Chen, L., Chen, L., Yuan, G., Fang, Y., Chen, W., . . . Gale, R. P. (2020). COVID-19 in persons with
530   haematological cancers. *Leukemia, 34*(6), 1637-1645. doi:10.1038/s41375-020-0836-7
531   Huang da, W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene
532   lists using DAVID bioinformatics resources. *Nat Protoc, 4*(1), 44-57. doi:10.1038/nprot.2008.211
533   Huang, Q., Liu, F., & Shen, J. (2019). Bioinformatic validation identifies candidate key genes in diffuse
534   large-B cell lymphoma. *Per Med, 16*(4), 313-323. doi:10.2217/pme-2018-0068
535   Karube, K., Enjuanes, A., Dlouhy, I., Jares, P., Martin-Garcia, D., Nadeu, F., . . . Campo, E. (2018).
536   Integrating genomic alterations in diffuse large B-cell lymphoma identifies new relevant
537   pathways and potential therapeutic targets. *Leukemia, 32*(3), 675-684.
538   doi:10.1038/leu.2017.251
539   Kawahata, T., Kawahara, K., Shimokawa, M., Sakiyama, A., Shiraishi, T., Minami, K., . . . Furukawa, T.
540   (2020). Involvement of ribosomal protein L11 expression in sensitivity of gastric cancer against
541   5-FU. *Oncol Lett, 19*(3), 2258-2264. doi:10.3892/ol.2020.11352
542   Kayama, K., Watanabe, S., Takafuji, T., Tsuji, T., Hironaka, K., Matsumoto, M., . . . Fujita, M. (2017).
543   GRWD1 negatively regulates p53 via the RPL11-MDM2 pathway and promotes tumorigenesis.
544   *EMBO Rep, 18*(1), 123-137. doi:10.15252/embr.201642444
545   Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis.
546   *BMC Bioinformatics, 9*(1), 559. doi:10.1186/1471-2105-9-559
547   Langfelder, P., & Horvath, S. (2012). Fast R Functions for Robust Correlations and Hierarchical Clustering.
548   *J Stat Softw, 46*(11).
549   Lee, L. Y. W., Cazier, J. B., Starkey, T., Briggs, S. E. W., Arnold, R., Bisht, V., . . . Kerr, R. (2020). COVID-19
550   prevalence and mortality in patients with cancer and the effect of primary tumour subtype and
551   patient demographics: a prospective cohort study. *Lancet Oncol, 21*(10), 1309-1316.
552   doi:10.1016/s1470-2045(20)30442-3
553   Liang, Z., Mou, Q., Pan, Z., Zhang, Q., Gao, G., Cao, Y., . . . Feng, W. (2019). Identification of candidate
554   diagnostic and prognostic biomarkers for human prostate cancer: RPL22L1 and RPS21. *Med
555   Oncol, 36*(6), 56. doi:10.1007/s12032-019-1283-z
556   Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., . . . Zhang, Z. (2020). Single-cell landscape of
557   bronchoalveolar immune cells in patients with COVID-19. *Nat Med, 26*(6), 842-844.
558   doi:10.1038/s41591-020-0901-9

559   Liu, B. B., Ma, T., Sun, W., Gao, W. Y., Liu, J. M., Li, L. Q., . . . Guo, Y. Y. (2021). Centromere protein U
560   enhances the progression of bladder cancer by promoting mitochondrial ribosomal protein s28
561   expression. *Korean J Physiol Pharmacol, 25*(2), 119-129. doi:10.4196/kjpp.2021.25.2.119
562   Liu, J., Zhou, S., Li, S., Jiang, Y., Wan, Y., Ma, X., & Cheng, W. (2019). Eleven genes associated with
563   progression and prognosis of endometrial cancer (EC) identified by comprehensive
564   bioinformatics analysis. *Cancer cell international, 19*, 136-136. doi:10.1186/s12935-019-0859-1
565   Liu, W., Liu, J., Song, Y., Zeng, X., Wang, X., Mi, L., . . . Union for China Lymphoma Investigators of the
566   Chinese Society of Clinical, O. (2019). Burden of lymphoma in China, 2006-2016: an analysis of
567   the Global Burden of Disease Study 2016. *Journal of hematology & oncology, 12*(1), 115-115.
568   doi:10.1186/s13045-019-0785-7
569   Lui, T. W. H., Tsui, N. B. Y., Chan, L. W. C., Wong, C. S. C., Siu, P. M. F., & Yung, B. Y. M. (2015). DECODE:
570   an integrated differential co-expression and differential expression analysis of gene expression
571   data. *BMC Bioinformatics, 16*(1), 182. doi:10.1186/s12859-015-0582-4
572   Luo, B., Gu, Y.-y., Wang, X.-d., Chen, G., & Peng, Z.-g. (2018). Identification of potential drugs for diffuse
573   large b-cell lymphoma based on bioinformatics and Connectivity Map database. *Pathology -*
574   *Research and Practice, 214*(11), 1854-1867. doi:https://doi.org/10.1016/j.prp.2018.09.013
575   Martínez-Reyes, I., Cardona, L. R., Kong, H., Vasan, K., McElroy, G. S., Werner, M., . . . Chandel, N. S.
576   (2020). Mitochondrial ubiquinol oxidation is necessary for tumour growth. *Nature, 585*(7824),
577   288-292. doi:10.1038/s41586-020-2475-6
578   Naresh, K. N., Raphael, M., Ayers, L., Hurwitz, N., Calbi, V., Rogena, E., . . . Leoncini, L. (2011).
579   Lymphomas in sub-Saharan Africa--what can we learn and how can we help in improving
580   diagnosis, managing patients and fostering translational research? *Br J Haematol, 154*(6), 696-
581   703. doi:10.1111/j.1365-2141.2011.08772.x
582   Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., . . . Alizadeh, A. A.
583   (2019). Determining cell type abundance and expression from bulk tissues with digital
584   cytometry. *Nature Biotechnology, 37*(7), 773-782. doi:10.1038/s41587-019-0114-2
585   Norberg, E., Lako, A., Chen, P. H., Stanley, I. A., Zhou, F., Ficarro, S. B., . . . Danial, N. N. (2017).
586   Differential contribution of the mitochondrial translation pathway to the survival of diffuse large
587   B-cell lymphoma subsets. *Cell Death Differ, 24*(2), 251-262. doi:10.1038/cdd.2016.116
588   Pasqualucci, L., & Dalla-Favera, R. (2018). Genetics of diffuse large B-cell lymphoma. *Blood, 131*(21),
589   2307-2319. doi:10.1182/blood-2017-11-764332
590   Passamonti, F., Cattaneo, C., Arcaini, L., Bruna, R., Cavo, M., Merli, F., . . . Corradini, P. (2020). Clinical
591   characteristics and risk factors associated with COVID-19 severity in patients with
592   haematological malignancies in Italy: a retrospective, multicentre, cohort study. *Lancet*
593   *Haematol, 7*(10), e737-e745. doi:10.1016/s2352-3026(20)30251-9
594   Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., . . . Chinnaiyan, A. M. (2004).
595   ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia, 6*(1),
596   1-6. doi:10.1016/s1476-5586(04)80047-2
597   Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers
598   differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids*
599   *Research, 43*(7), e47-e47. doi:10.1093/nar/gkv007
600   Shah, V., Ko Ko, T., Zuckerman, M., Vidler, J., Sharif, S., Mehra, V., . . . Kulasekararaj, A. G. (2020). Poor
601   outcome and prolonged persistence of SARS-CoV-2 RNA in COVID-19 patients with
602   haematological malignancies; King's College Hospital experience. *Br J Haematol, 190*(5), e279-
603   e282. doi:10.1111/bjh.16935
604   Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., . . . Ideker, T. (2003). Cytoscape:
605   a software environment for integrated models of biomolecular interaction networks. *Genome*
606   *research, 13*(11), 2498-2504. doi:10.1101/gr.1239303

607   Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., . . . Lancet, D. (2016). The
608   GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc*
609   *Bioinformatics, 54*, 1.30.31-31.30.33. doi:10.1002/cpbi.5
610   Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P.
611   (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-
612   wide expression profiles. *Proceedings of the National Academy of Sciences, 102*(43), 15545.
613   doi:10.1073/pnas.0506580102
614   Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., . . . Mering, C. V. (2019).
615   STRING v11: protein-protein association networks with increased coverage, supporting
616   functional discovery in genome-wide experimental datasets. *Nucleic Acids Res, 47*(D1), D607-
617   d613. doi:10.1093/nar/gky1131
618   Tang, Z., Kang, B., Li, C., Chen, T., & Zhang, Z. (2019). GEPIA2: an enhanced web server for large-scale
619   expression profiling and interactive analysis. *Nucleic Acids Res, 47*(W1), W556-w560.
620   doi:10.1093/nar/gkz430
621   Team, R. C. (2019). R: A language and environment for statistical computing. R Foundation for Statistical
622   Computing, Vienna, Austria. *URL https://www.R-project.org/.*
623   Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., . . . Ponten, F. (2017). A pathology
624   atlas of the human cancer transcriptome. *Science, 357*(6352), eaan2507.
625   doi:10.1126/science.aan2507
626   van Dam, S., Võsa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2018). Gene co-expression
627   analysis for functional classification and gene-disease predictions. *Briefings in bioinformatics,*
628   *19*(4), 575-592. doi:10.1093/bib/bbw139
629   Visco, C., Li, Y., Xu-Monette, Z. Y., Miranda, R. N., Green, T. M., Li, Y., . . . Young, K. H. (2012).
630   Comprehensive gene expression profiling and immunohistochemical studies support application
631   of immunophenotypic algorithm for molecular subtype classification in diffuse large B-cell
632   lymphoma: a report from the International DLBCL Rituximab-CHOP Consortium Program Study.
633   *Leukemia, 26*(9), 2103-2113. doi:10.1038/leu.2012.83
634   Wang, Y., Sui, J., Li, X., Cao, F., He, J., Yang, B., . . . Pu, Y. D. (2015). RPS24 knockdown inhibits colorectal
635   cancer cell migration and proliferation in vitro. *Gene, 571*(2), 286-291.
636   doi:10.1016/j.gene.2015.06.084
637   Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York*, ISBN 978-
638   973-319-24277-24274.
639   Xiao, J., Wang, X., & Bai, H. (2020). Clinical Features and Prognostic Impact of Coexpression Modules
640   Constructed by WGCNA for Diffuse Large B-Cell Lymphoma. *BioMed research international,*
641   *2020*, 7947208. doi:10.1155/2020/7947208
642   Yi, J. H., Yoon, S. E., Ryu, K. J., Ko, Y. H., Kim, W. S., & Kim, S. J. (2020). Pre-treatment serum IL-10
643   predicts the risk of secondary central nervous system involvement in patients with diffuse large
644   B-cell lymphoma. *Cytokine, 129*, 155048. doi:10.1016/j.cyto.2020.155048
645   Yin, Z., Bai, L., Li, W., Zeng, T., Tian, H., & Cui, J. (2019). Targeting T cell metabolism in the tumor
646   microenvironment: an anti-cancer therapeutic strategy. *Journal of Experimental & Clinical*
647   *Cancer Research, 38*(1), 403. doi:10.1186/s13046-019-1409-3
648   Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). clusterProfiler: an R package for comparing biological
649   themes among gene clusters. *Omics, 16*(5), 284-287. doi:10.1089/omi.2011.0118
650   Zhou, L., Ding, L., Gong, Y., Zhao, J., Xin, G., Zhou, R., & Zhang, W. (2020). Identification of hub genes
651   associated with the pathogenesis of diffuse large B-cell lymphoma subtype one characterized by
652   host response via integrated bioinformatic analyses. *PeerJ, 8*, e10269. doi:10.7717/peerj.10269

653   Zhou, R., Zhang, J., Zeng, D., Sun, H., Rong, X., Shi, M., . . . Liao, W. (2019). Immune cell infiltration as a
654   biomarker for the diagnosis and prognosis of stage I-III colon cancer. *Cancer immunology,*
655   *immunotherapy : CII, 68*(3), 433-442. doi:10.1007/s00262-018-2289-7
656   Zou, D., Zhang, H., Ke, J., Li, J., Zhu, Y., Gong, Y., . . . Miao, X. (2020). Three functional variants were
657   identified to affect RPS24 expression and significantly associated with risk of colorectal cancer.
658   *Arch Toxicol, 94*(1), 295-303. doi:10.1007/s00204-019-02600-9

659

660

Figure 1

Statistics for the differentially expressed genes.

(A) Volcano plot highlighting significant genes in DLBCL and non-tumor tissues. UP represents upregulated; DOWN, downregulated; NOT, not significant. (B) Heatmap of the top 169 DEGs between DLBCL cases and controls (NORMAL); |log2FC|> 2, p-value <0.05; the range of the colors corresponds with the range of expression values. (C) Venn diagram shows 118 overlapping (common) DEGs screened using 5 Cytohubba centrality methods. DEGs, differentially expressed genes.

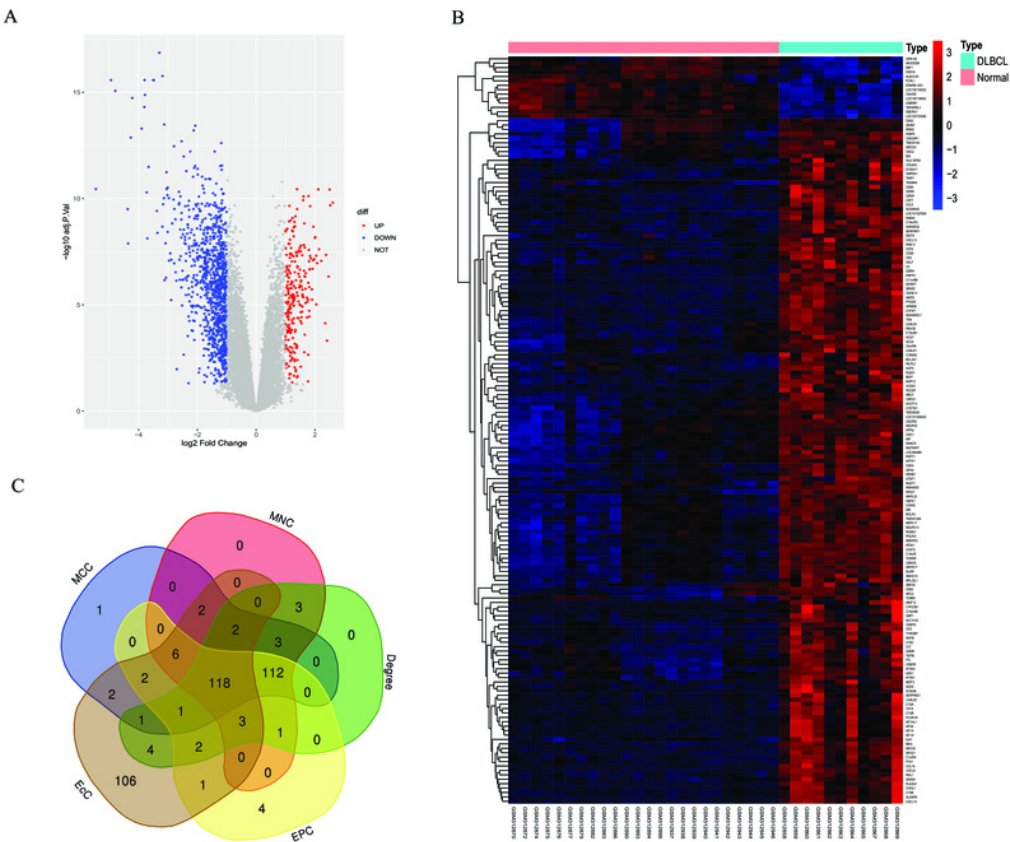



Figure 2

Gene ontology (GO) and KEGG enrichment analysis of all differentially expressed genes (DEGs).

(A) In GO analysis, the top 10 significantly enriched DEGs. The x-axis is the number of DEGs involved in the GO terms; the y-axis is the significantly enriched GO terms. (B) In KEGG analysis, the top 15 significantly enriched pathways of the DEGs. KEGG, Kyoto encyclopedia of genes and genomes.

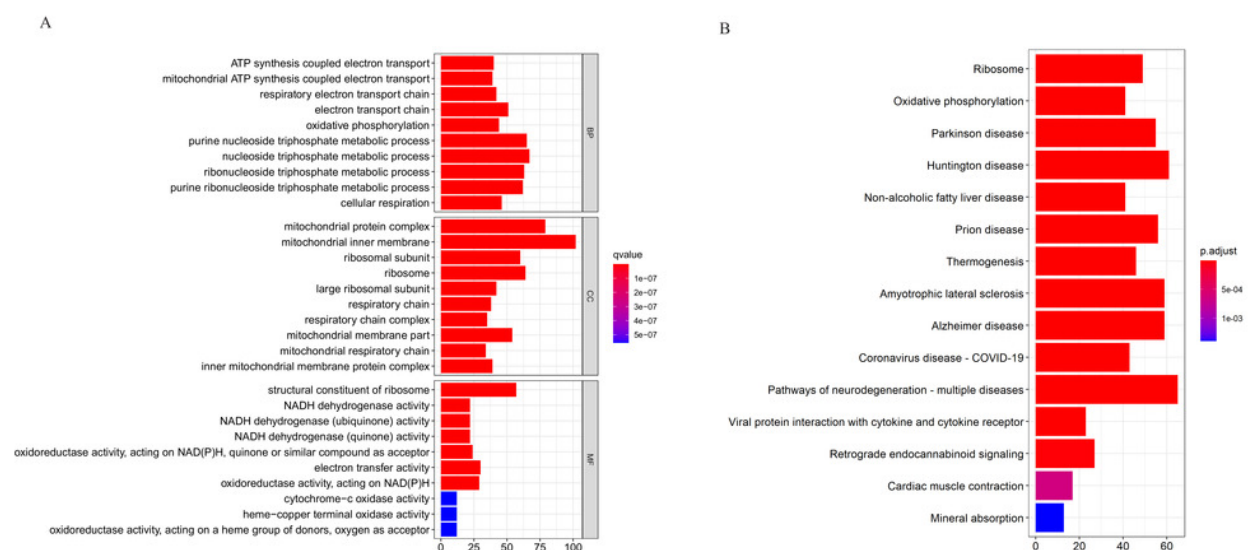



Figure 3

The prediction of tumor-infiltrating immune cells (TIICs) using the GSE12453 dataset.

Violin plot comparing the proportions of TIICs between non-tumor controls (in blue) and DLBCL (in red). The x and y axes represent TIICs and their relative percentages, respectively. There was no T cell CD4 memory resting. (B) Bar plots for 24 non-tumor controls and 11 DLBCL samples (x-axis) and the percentages of immune cell subsets (y-axis).

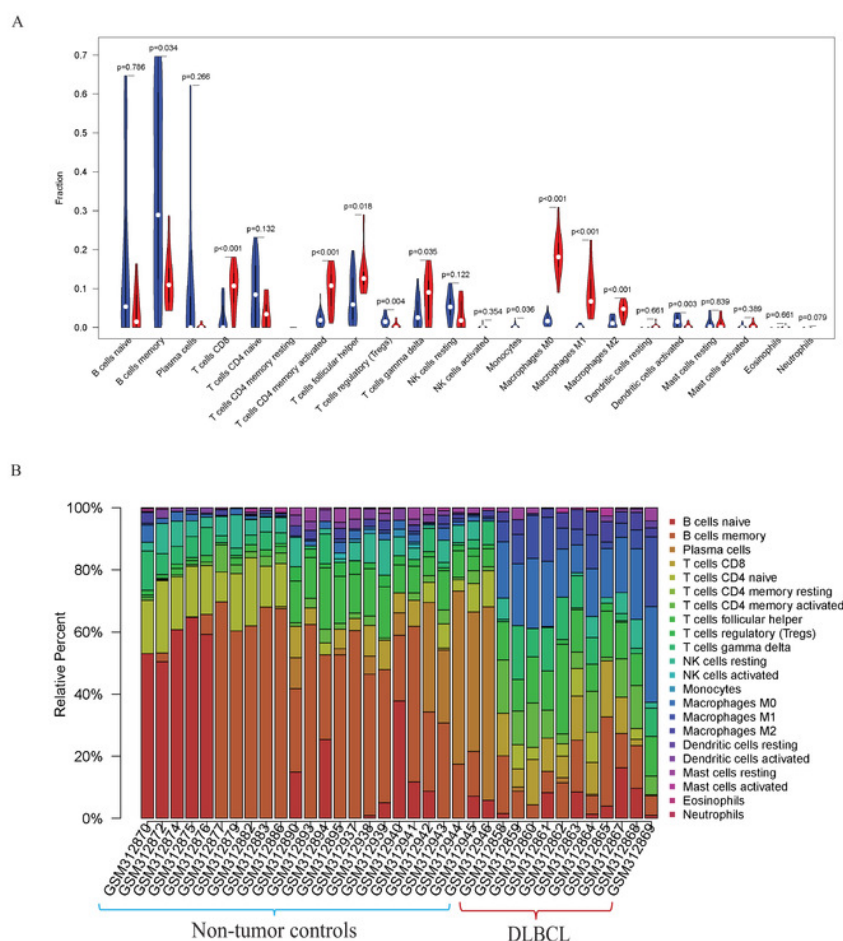



Figure 4

The protein-protein interaction (PPI) networks using the STRING and Cytoscape databases.

(A) The PPI network for the 109 highly connected DEGs; confidence score, 0.9. (B) Cluster 1 consisted of 61 nodes, 995 edges with the highest k-score of 33.17. (C) Cluster 2 had 28 nodes, 359 edges, and a k-score of 26.59. (D) Cluster 3 had 20 nodes, 190 edges, and a k-score of 20.00. (e) The top 250 ranked DEGs.

Figure 5

Construction of gene co-expression network.

(A) Sample clustering to detect outliers, no obvious outliers. (B and C) Determination of soft-threshold power. When β is set at 8, the log-log plot of the network connectivity distribution produces a straight line. (D) Hierarchical clustering dendrograms (top modules). Each color band (bottom) represents a color-coded module that contains a group of highly connected genes. The Dynamic Tree Cut identified 18 modules. (E) A heatmap showing the topological overlap matrix (TOM) among the top 1000 genes selected from all genes. The color intensity indicates the correlation strength between pairs of modules: the left side (gene dendrogram) and the top side (module assignment).

A

Sample dendrogram and trait heatmap

B

Scale independence

Mean connectivity

Soft Threshold (power)

Soft Threshold (power)

C

Histogram of k

Check scale free topology
scale R^2= 0.86 , slope= -1.78

D

Gene dendrogram and module colors

Dynamic Tree Cut

E

Heatmap of top 1000 genes

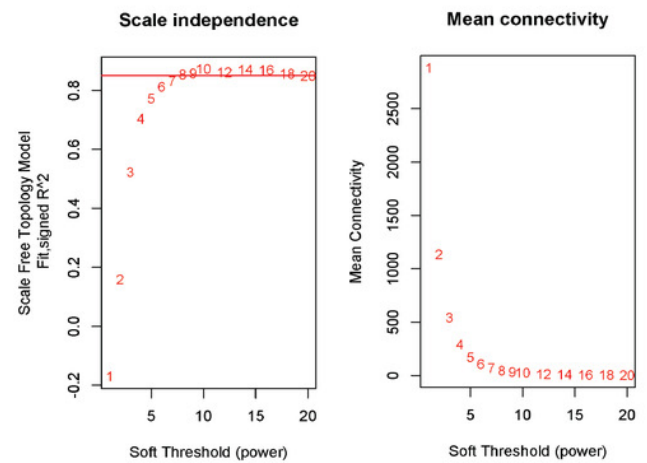
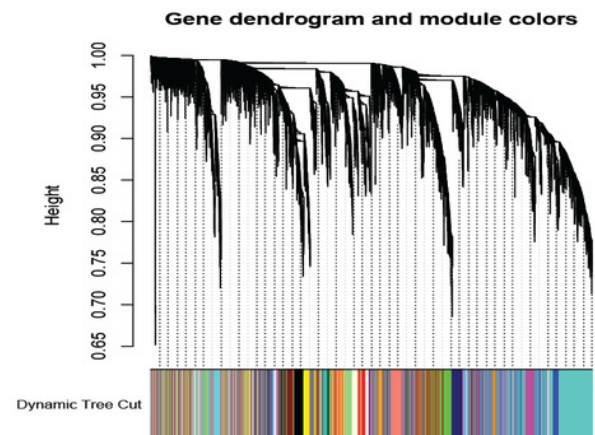



Figure 6

Identifying modules of clinical relevance from the GSE12453 dataset.

(A) Module trait relationship showing correlation coefficients between module eigengenes (row) and disease status (column), with the corresponding p-values in brackets. The degree of correlation is based on a color legend: red, strong positive and blue, strong negative correlation. (B) Heatmap plot of the adjacencies in the eigengene network, including the relationship with DLBCL trait. The top panel is the hierarchical clustering dendrogram of the eigengenes. The bottom panel shows the eigengene adjacency. (C) Scatter plots of gene significance (GS) versus module membership (MM) for the DLBCL related modules (turquoise, green and dark magenta, respectively).

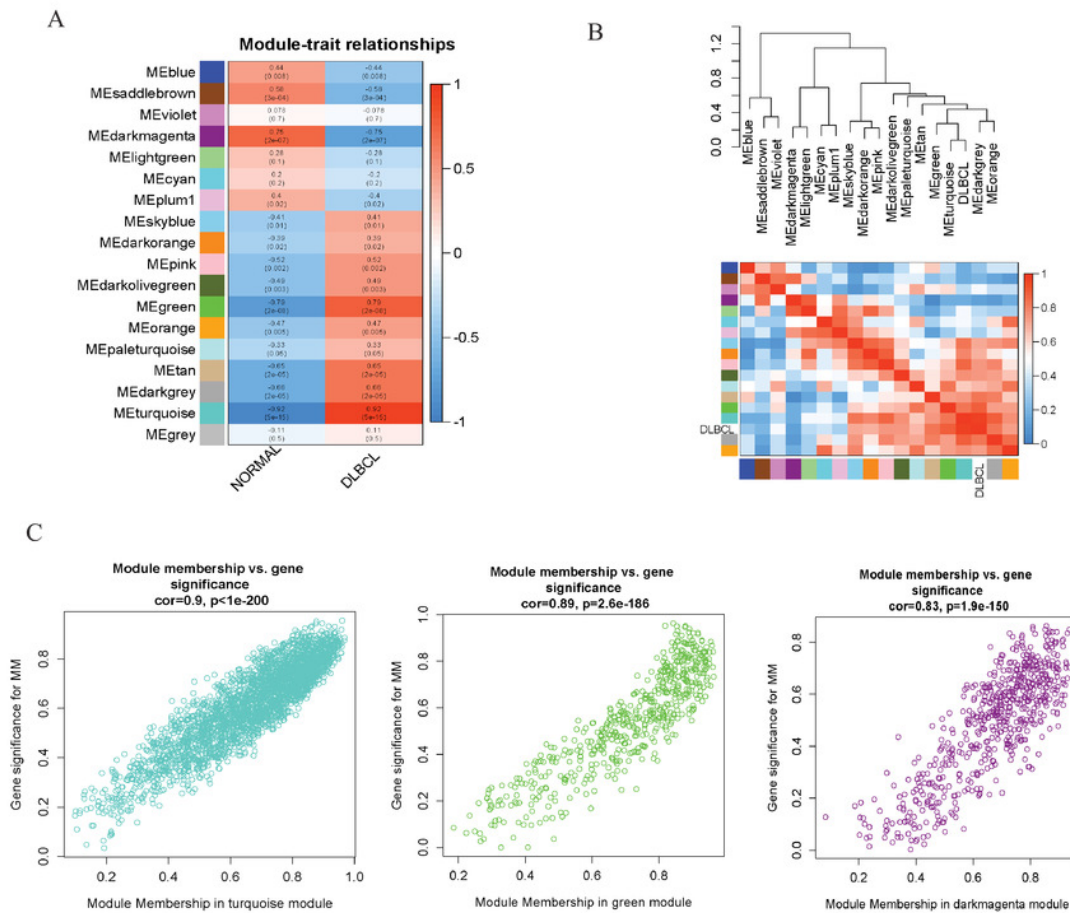



Figure 7

Top 10 enrichment results of the 195 WGCNA (weighted correlation network analysis) genes.

(A) Gene ontology (GO) functional analysis. (B) Kyoto encyclopedia of genes and genomes (KEGG) pathway analysis.

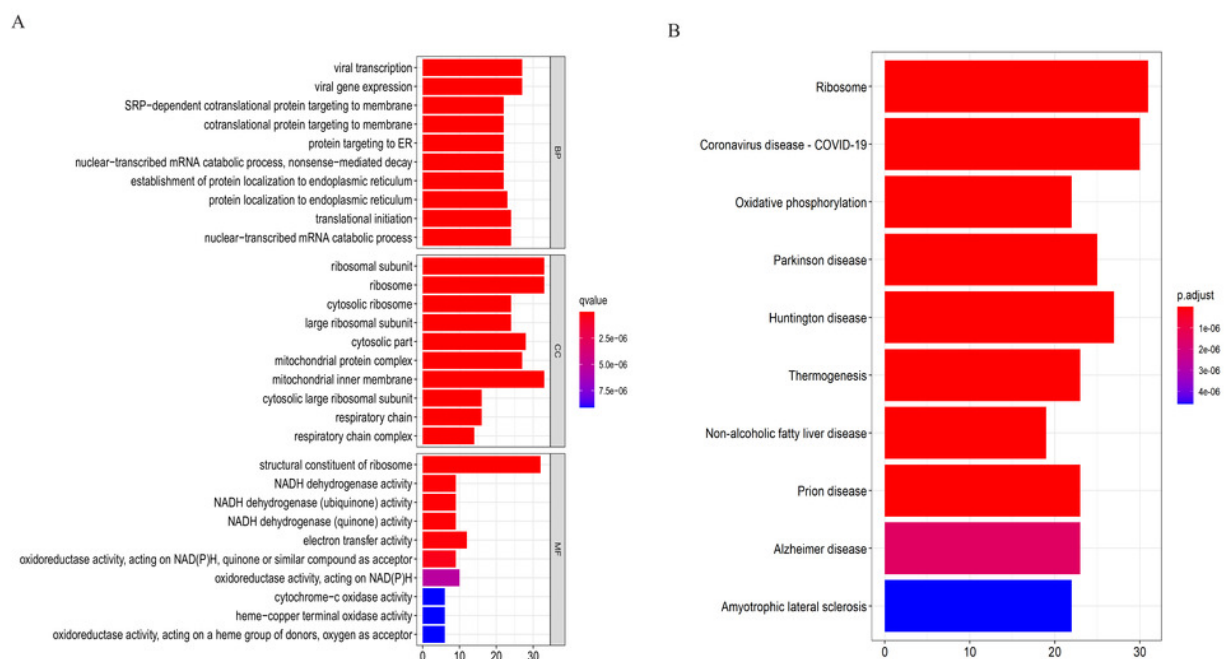



Figure 8

Differentially co-expressed hub genes.

(A) Venn diagram indicating the 46 common genes. (B) Protein-protein interaction (PPI) network of the 46 hub genes (STRING and Cytoscape databases).

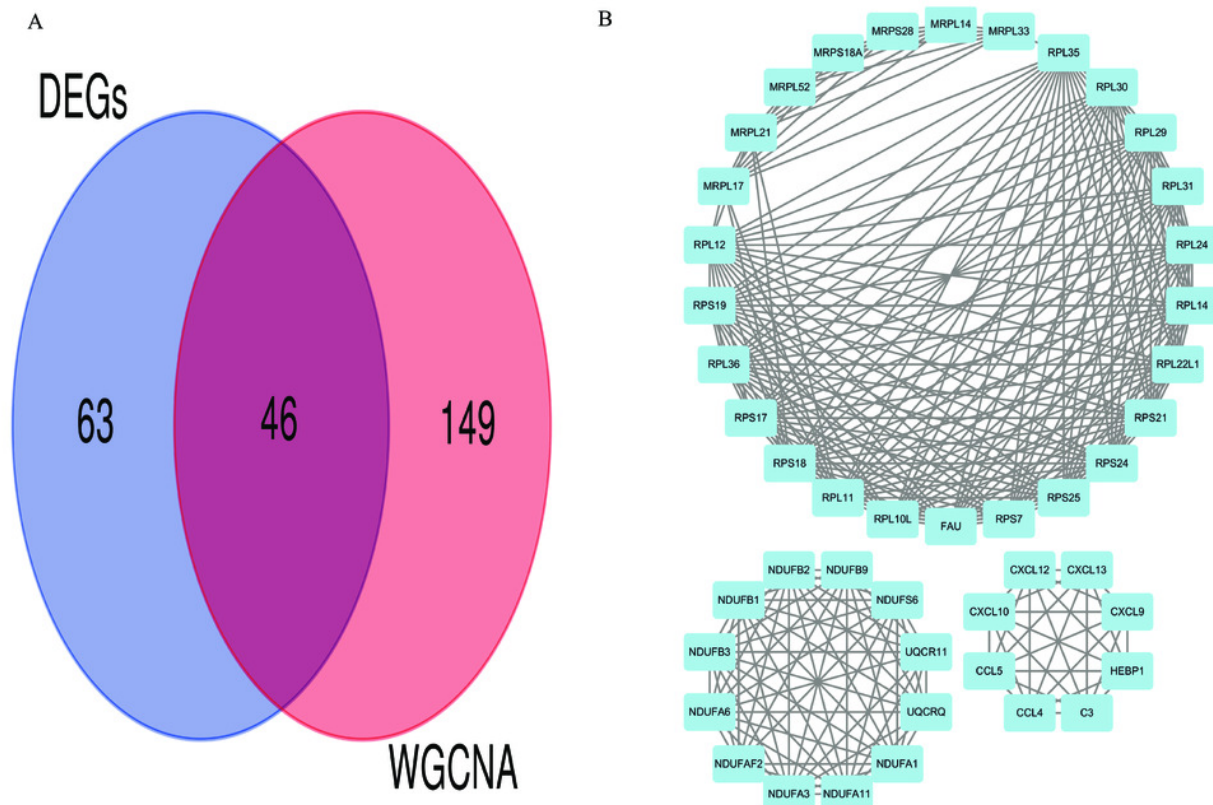



Figure 9

The expressions of the hub genes in the GEPIA2 database.

*(p<0.01). The data were retrieved from the GEPIA 2 database
(http://gepia2.cancer-pku.cn/#index).

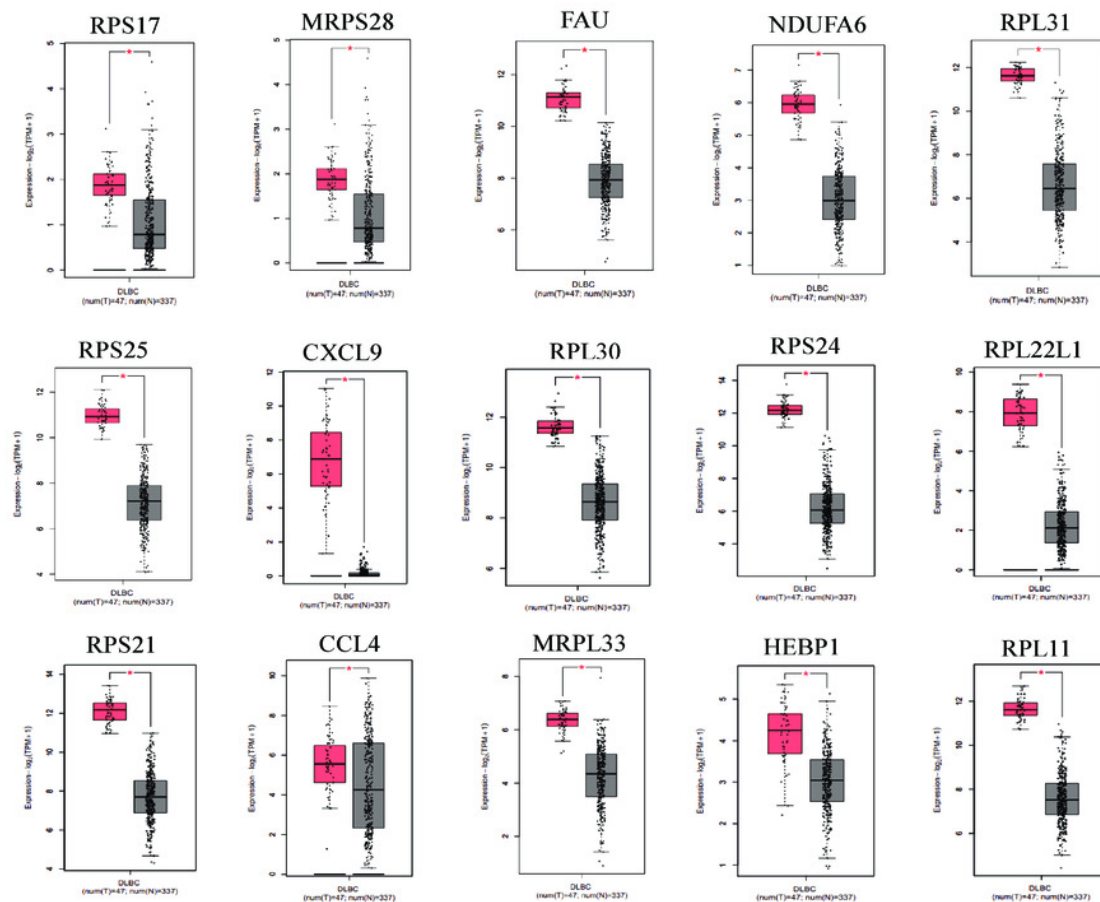



Figure 10

The Kaplan–Meier estimates for the overall survival (OS) of the 15 core hub genes in GSE31312 (p < 0.05).

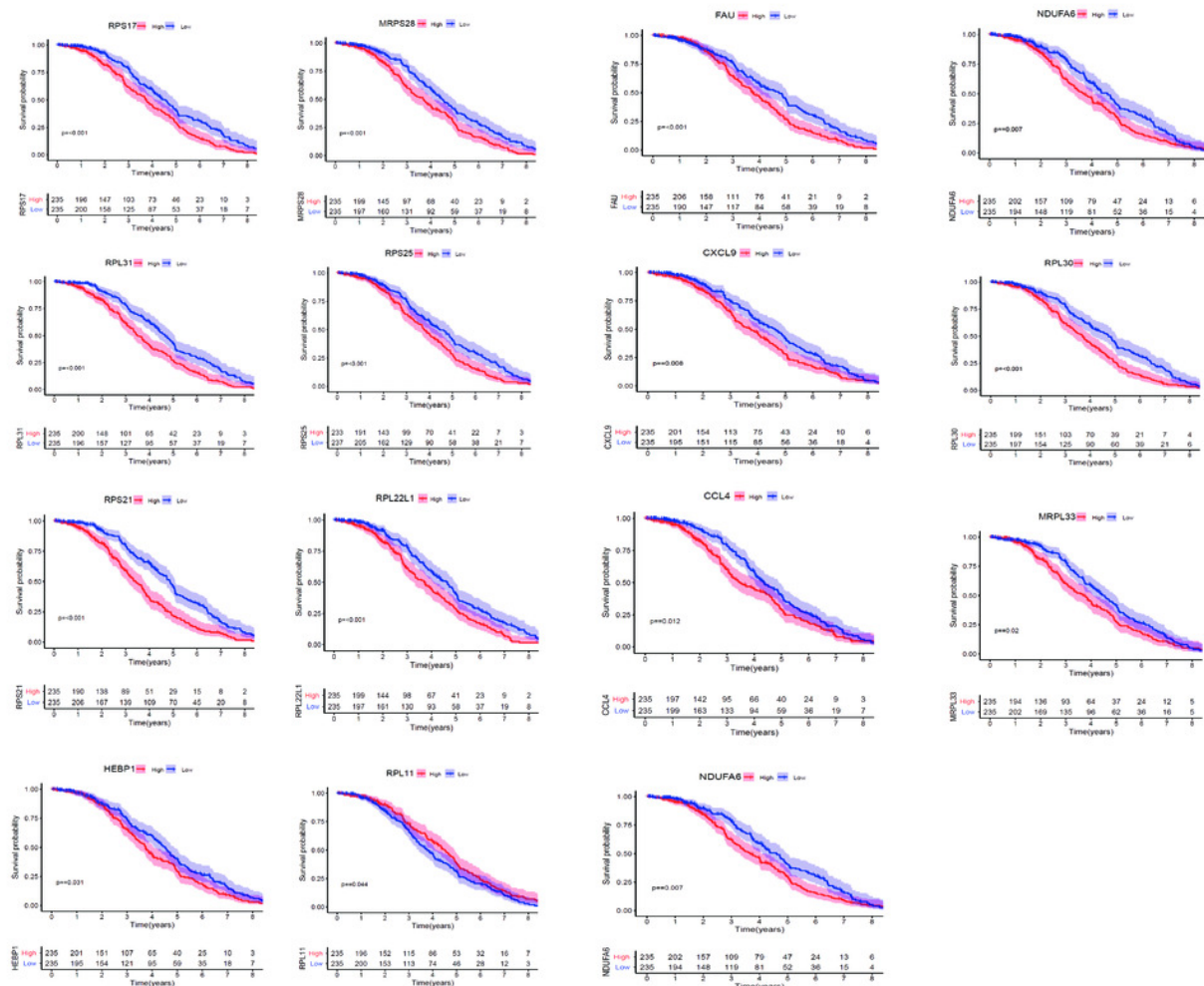
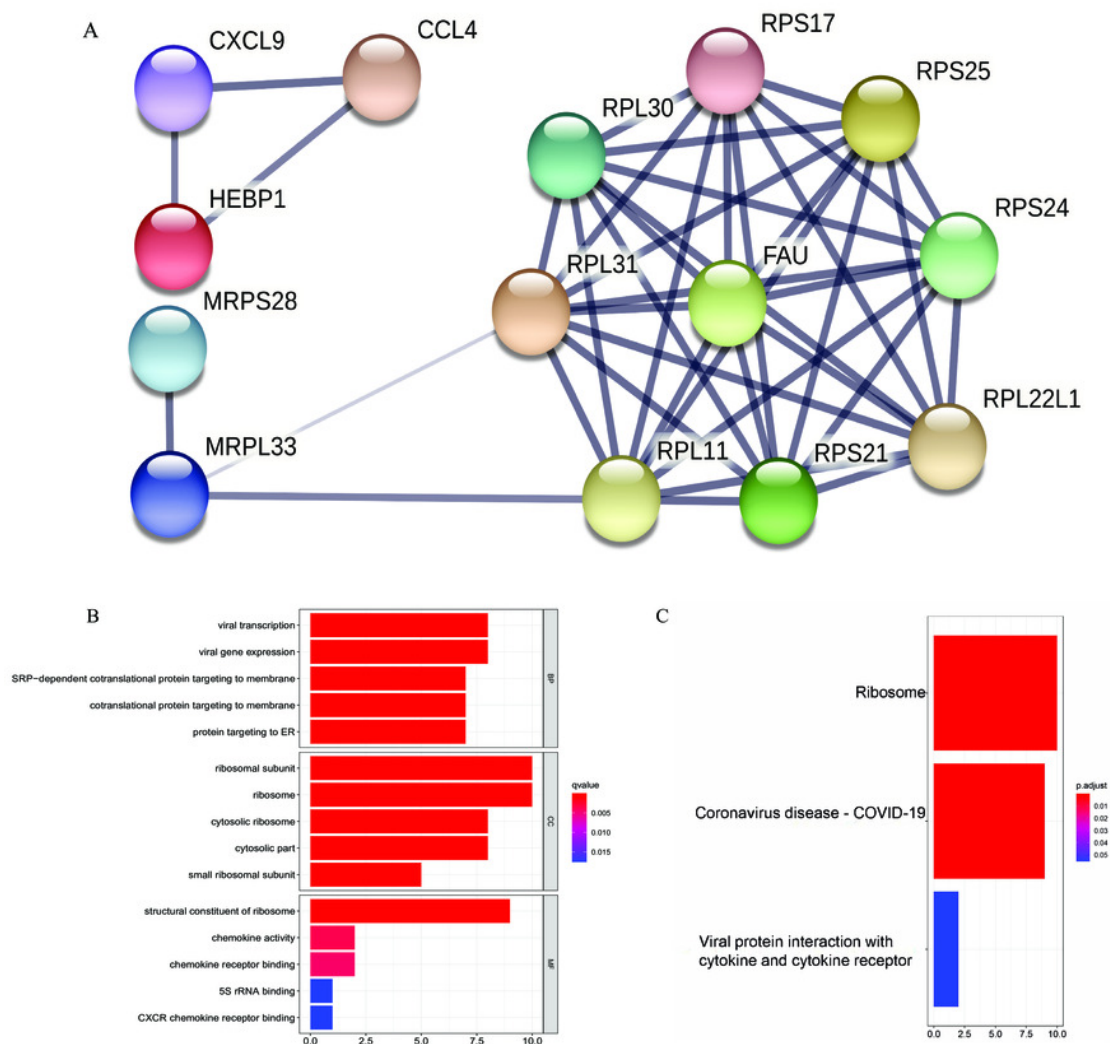



Figure 11

PPI and enrichment analysis of the 15 core hub genes.

(A) Cluster analysis. Genes in the circle represent covid 19 pathway genes. (B) The top 5 gene ontology (GO) terms. (C) The top 3 KEGG (Kyoto encyclopedia of genes and genomes) pathways.

Gene ontology and KEGG pathway analysis of the 46 differentially co-expressed genes.

Selection of terms enriched in the categories based on the most significant adjusted p-value (p. adjust). Count: the number of genes enriched in each term or pathway.

Term	Description	Count	p. adjust
Biological process			
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	16	1.29E-22
GO:0019083	Viral transcription	18	1.29E-22
GO:0006613	Cotranslational protein targeting to membrane	16	1.29E-22
GO:0019080	Viral gene expression	18	2.85E-22
GO:0045047	Protein targeting to ER	16	2.98E-22
GO:0000184	Nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	16	3.30E-22
GO:0072599	Establishment of protein localization to endoplasmic reticulum	16	3.74E-22
GO:0070972	Protein localization to endoplasmic reticulum	16	7.45E-21
Cellular component			
GO:0044391	Ribosomal subunit	25	2.39E-37
GO:0005840	Ribosome	25	1.45E-33
GO:0022626	Cytosolic ribosome	18	3.96E-28
GO:0015934	Large ribosomal subunit	17	1.54E-25
GO:0044445	Cytosolic part	18	6.76E-22
GO:0098798	Mitochondrial protein complex	17	8.16E-20
GO:0005743	Mitochondrial inner membrane	19	2.01E-18
GO:0022625	Cytosolic large ribosomal subunit	11	1.14E-17
Molecular function			
GO:0003735	Structural constituent of ribosome	24	1.15E-33
GO:0003954	NADH dehydrogenase activity	9	3.12E-14
GO:0008137	NADH dehydrogenase (ubiquinone) activity	9	3.12E-14
GO:0050136	NADH dehydrogenase (quinone) activity	9	3.12E-14
GO:0016655	Oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor	9	3.26E-13
GO:0016651	Oxidoreductase activity, acting on NAD(P)H	9	6.05E-11
GO:0008009	Chemokine activity	6	2.25E-08
KEGG pathway			
hsa03010	Ribosome	24	4.36E-29
hsa05171	Coronavirus disease - COVID-19	21	2.35E-20
hsa00190	Oxidative phosphorylation	11	9.15E-10

List of the hub genes involved in the 3 KEGG pathways

DCE: differentially co-expressed

S/N	24 differentially co-expressed (DCE) ribosome genes	21 DCE coronavirus disease - COVID-19	11 differentially co-expressed oxidative phosphorylation genes
1	MRPS18A	RPL10L	UQCRQ
2	RPL10L	RPL24	NDUFB1
3	MRPL17	RPS17	NDUFB2
4	RPL24	CXCL10	NDUFB3
5	RPS17	RPL30	NDUFS6
6	RPL30	RPS18	NDUFA1
7	RPS18	C3	NDUFA3
8	RPS19	RPS19	NDUFA11
9	RPL29	RPL29	NDUFB9
10	RPL31	RPL31	NDUFA6
11	RPL36	RPL36	UQCR11
12	MRPL33	RPS21	
13	RPS21	FAU	
14	MRPL21	RPS24	
15	FAU	RPL14	
16	RPS24	RPS25	
17	RPL14	RPL22L1	
18	RPS25	RPL11	
19	RPL22L1	RPL35	
20	RPL35	RPL12	
21	MRPL14	RPS7	
22	RPL11		
23	RPL12		
24	RPS7		

List of the core hub genes involved in the 3 KEGG pathways.

KEGG pathway	Number of genes	List of genes	p. adjust
hsa03010: Ribosome	10	RPS24/RPS21/RPL31/RPL30/RPS17/FAU/RPS25/RPL22L1/MRPL33/RPL11	3.12E-14
hsa05171: Coronavirus disease - COVID-19	9	RPS24/RPS21/RPL31/RPL30/RPS17/FAU/RPS25/RPL22L1/RPL11	6.89E-11
hsa04061: Viral protein interaction with cytokine and cytokine receptor	2	CXCL9/CCL4	0.0552