Analysis of mRNA biomarker predicting progression of acute lymphoblastic leukaemia by big data mining

Jianzhi Deng¹, Xiaohui Cheng¹, Yuehan Zhou²,*

¹ Guilin University of Technology, Guilin 541004, China
² Guilin Medical University, Guilin 541004, China
*Corresponding author: yuehanzhou2012@163.com

Abstract. Acute lymphoblastic leukaemia (ALL) is a hematologic malignancy. In this study, we focus on the research of the progressed biomarker of the ALL disease based on the ALL phase II to phase III mRNA expression data from Therapeutically Applicable Research to Generate Effective Treatment (TARGET). 204 differentially expressed mRNAs (DEmRNAs) were screened from the mRNA matrix (P-value < 0.01 and |LogFoldChange| > 4). And the DEmRNAs were enriched in 16 MF, 15 CC and 49 BP groups of the gene ontology (GO) terms (Count > 2 and P-value < 0.1) and 3 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (P-value < 0.05, q-value < 0.05) by the GO and KEGG pathway analysis. The survival analysis done by Kaplan-meier method was shown that the DEmRNAs and their enriched GOterms were closely related with the high risk of ALL progression. And the DEmRNAs would be the progressed biomarker of ALL when they were proved by the receiver operating characteristic (ROC) analysis.

1. INTRODUCTION
Acute lymphoblastic leukaemia is a high aggressive disease which progress fast and mortally in a few weeks if without treatment [1-3]. In this research, we aim to find out progressed biomarker of ALL disease. Our work was based on the comparison of ALL phase II and phase III RNA-seq data downloaded from TARGET database. According to the DEmRNAs screened from the mRNA matrix, the GO and KEGG pathway enrichment analysis was done by online tools and R software. The GO enrichment analysis was focus on the three affects, biological process (BP), cellular component (CC), molecular function (MF). After the survival analysis by the random selected DEmRNAs which were enriched in the highest enrichment GOterms, we find that the relationship between DEmRNAs, enriched GO terms and survival status. And then, the ROC curves were proved that the selected DEmRNAs would be the progressed biomarker of ALL.

2. MATERIALS AND METHODS

2.1 TARGET data and DEmRNA clustering
The phase II and phase III gene expression data of ALL patients were downloaded from TARGET database (https://ocg.cancer.gov /programs/target) and shown in Table 1. The mRNA expression matrix was merged by the downloaded RNA-seq txt files. The over-all survival time data and survival status data was extracted from the downloaded clinical data excel files. The differentially expressed mRNAs compared from phase II to phase III were screened from the mRNA expression matrix by edgeR package with standards of P-value < 0.01 and |LogFoldChange| > 4 [4, 5].
Table 1. Data number

| stage    | patient sample | mRNA | clinical data |
|----------|----------------|------|---------------|
| Phase II | 216            | 15906| 791           |
| Phase III| 96             | 15906| 116           |

2.2 GO and KEGG pathway enrichment analysis

GO enrichment analysis of the DEmRNAs was focused on 3 groups of GO terms, including BP, CC and MF by DAVID online tool (https://david.ncifcrf.gov/) with cut-off value of count > 2 and P-value < 0.1 [6]. Meanwhile, signaling pathway analysis of the DEmRNAs was enriched by KEGG (https://www.kegg.jp) and clusterProfiler R package (organism=hsa, P-value Cutoff = 0.05, q-value Cutoff = 0.05) [7, 8]. Both GO and KEGG pathway enrichment were analyzed in homo species.

2.3 KM Survival analysis and ROC analysis

According to the result of GO and pathway analysis, survival and ROC analysis were accomplished by survival R package. The DEmRNAs of survival analysis was selected from the highest enriched GO terms [9-11]. The survival curve was drawn by Kaplan-meier method [12]. And the receiver operating characteristic (ROC) curves were drawn based on the DEmRNAs. The area under the ROC curve (AUC) was shown the accurate of the prediction.

3. RESULTS

3.1 DEmRNA and clinical data

With the common standards of P-value < 0.05 and |LogFoldChange| > 2, over 1500 DEmRNAs were found. So we set the other cut-off standards of P-value < 0.01 and |LogFoldChange| > 4, and screened 204 more significant DEmRNAs including 34 high expressed mRNAs and 170 low expressed mRNAs. 192 effective clinical data were extracted from the downloaded clinical data excel files after delete the incomplete data.

Table 2. Differentially expressed mRNAs

| expression    | DEmRNA (Significant 30 ones) |
|---------------|------------------------------|
| High-expression| NDST3, SUCNR1, CNTNAP5, IQC1, MNX1, OR52K2, MAMDC2, CLCA1, PRODH, PREX2, LIN28B, SLC24A3, ZNF521, PEX5L, HOXA13, ILDR2, CNTN4, HPGDS, INHBA, TTN, CYP4F2, BDNF, GALNT13, SCUBE1, TAS2R3, TRIM71, NEUROG2, P2RY1, KRT73, ATP8B4 |
| Low-expression| SEMA5B, SOX1, GABRA5, LLRC15, CCDC60, SERPINB13, SEMA6D, KCNF1, TMSB15A, SLC6A15, SHOX, PBX1, SLITRK2, SH3GL3, ATP5EP2, OR4K2, MUC6, RPS17, APELA, RP11-152F13.10, FAM205A, AHSP, HBA1, TOX3, WNT16, SHISA9, SALL1, HBA2, MYF6, PTPRZ1 |

3.2 GO and KEGG pathway enrichment

Based on the screened DEmRNAs, it was enriched in 80 GO terms (Count > 2 and Pvalue < 0.1) including 16 MF, 15 CC and 49 BP.
The 12 highest enriched GO terms (count >= 10) were extracellular exosome (GO:0070062), extracellular region (GO:0005576), integral component of plasma membrane (GO:0005887), extracellular space (GO:0005615), positive regulation of transcription from RNA polymerase II promoter (GO:0045944), sequence-specific DNA binding (GO:0043565), cell junction (GO:0030054), cell-cell signaling (GO:0007267), heme binding (GO:0020037), nervous system development (GO:0007399), cell adhesion (GO:0007155) and transcription from RNA polymerase II promoter (GO:0006366). Highest GO term was enriched in CC group, extracellular exosome. And there were 38 DEmRNAs enriched in this CC group.

The DEmRNA were mainly enriched in 3 KEGG pathways (Figure 3), Taste transduction (hsa04742), Malaria (hsa05144) and African trypanosomiasis (hsa05143) by KEGG pathway analysis.

| ID    | Description          | p.adjust  | GeneRatio |
|-------|----------------------|-----------|-----------|
| 05144 | Malaria              | 0.008059  | 5/74      |
| 04742 | Taste transduction   | 0.008059  | 6/74      |
| 05143 | African trypanosomias | 0.016766  | 4/74      |
3.3 KM Survival analysis

Table 4. The DE-mRNAs in the highest enriched GO terms

| Term | Genes |
|------|-------|
| CC GO:0070062 | SLC36A2, IFITM3, HBM, UCHL1, GYPA, LRRC15, TTN, KRT33B, RPS29, PVALB, SUCNR1, RTN4L2, RPL11, SERPINB13, SCNN1G, HBB, SMIM1, SH3GL3, KIF12, TRHDE, LGALS7, RPL27, KRT10, HBA2, IGF2, HBA1, AMBP, KRT73, VWF, CDH13, RPS18, LGALS7B, PKP1, RPS17, CNFN, CTSE, CA1, ATP6V0A4 |
| BP GO:0045944 | MYF6, UTF1, FOXL2, SOX1, TBX5, BEX1, PGR, INHBA, CDH13, BARX2, SFRP2, GATA6, SALL1, PAX7, P2RY1, FOXF2, HIF3A, PBX1, ID4, TFAP2C |
| MF GO:0043565 | PGR, FOXL2, HOXA13, TBX5, IRX2, PAX7, SALL1, FOXF2, MNX1, SHOX, PBX1, SNAI2 |

It can be known from the GO analysis above that the screened DE-mRNAs were enriched in 16 MF, 15 CC and 49 BP with the cut-off value of Count > 2 and P-value < 0.1. And the highest enriched GO terms of CC, BP and MF groups were listed in Table 4.

Based on the 3 GO terms in Table 4, 5 DE-mRNAs were selected in every GO term at random for the later survival analysis by survival R package. In GO:0070062, SLC36A2, IFITM3, HBM, UCHL1 and GYPA were selected. Meanwhile, MYF6, UTF1, FOXL2, SOX1 and TBX5 were selected in BP group, GO:0045944. Then, PGR, FOXL2, HOXA13, TBX5 and IRX2 were selected in GO:0043565. The survival kmplot curves of the 3 GO terms were shown in Figure 3(a), Figure 4(a) and Figure 5(a). Figure 3(b), Figure 4(b) and Figure 5(b) were the ROC curves that analyzed by the same DE-mRNAs.

It could be known from the survival curves in Figure 3(a) (p = 4.144e-05), Figure 4(a) (p = 1.972e-01) and Figure 5(a) (p = 1.808e-03) that the dys-regulated mRNAs were closely related to the high risk of ALL.

Figure 3. K-mplot and ROC curve of the highest CC GOTer
4. DISCUSSION

In order to find out the relationship between mRNA and the progression of ALL disease, we downloaded the gene expression data and the clinical data from the TARGET database. According to the stage, the mRNA data were divided into phase II group and phase III group. Compared to the phase II group, 34 mRNAs were significantly high expressed in phase III. Besides, 170 mRNAs were low expressed. In the GO analysis, the 204 DEmRNAs were enriched in 80 GOterms (Count > 2 and Pvalue < 0.1), especially in 12 GOterms (Count > 10), GO:0070062, GO:0005576, GO:0005887, GO:0005615, GO:0045944, GO:0043565, GO:0030054, GO:0007267, GO:0020037, GO:0007399, GO:0007155 and GO:0006366. There were 38DEmRNAs enriched in extracellular exosome (GO:0070062). By KEGG pathway analysis with the clusterProfiler R package, the 204 DEmRNAs were mainly enriched in 3 KEGG pathways, Taste transduction, Malaria and African trypanosomiasis. The survival and ROC analysis of the random selected DEmRNAs which were enriched in the highest level CC(GO:0070062), BP(GO:0043565) and MF (GO:0045944) GOterms. From the kmplot and ROC curves, it can be shown that the GOterms and their enriched DEmRNAs were closely related with the high risk of the ALL progression. And the DEmRNAs screened in this research would be the progressed biomarker of the ALL disease.

Figure 4. Kmplot and ROC curve of the highest BP GOterm

Figure 5. Kmplot and ROC curve of the highest MF GOterm
ACKNOWLEDGMENTS
This work was supported by National Natural Science Foundation of China (No. 81660031, 81360090), Guangxi Natural Science Foundation of China (2014GXNSFBA 118151), and Guangxi key Laboratory Foundation of Embedded Technology and Intelligent System.

REFERENCES
[1] H. Beesley et al., The gene expression signature of relapse in paediatric acute lymphoblastic leukaemia: implications for mechanisms of therapy failure, Br J Haematol, vol. 131, no. 4, pp. 447-56, Nov 2005.
[2] F. W. van Delft et al., Prospective gene expression analysis accurately subtypes acute leukaemia in children and establishes a commonality between hyperdiploidy and t(12;21) in acute lymphoblastic leukaemia, Br J Haematol, vol. 130, no. 1, pp. 26-35, Jul 2005.
[3] L. Tian et al., The downregulation of miR-3173 in B-cell acute lymphoblastic leukaemia promotes cell invasion via PTK2, Biochem Biophys Res Commun, vol. 494, no. 3-4, pp. 569-574, Dec 16 2017.
[4] C. W. Law et al., RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR, F1000Res, vol. 5, 2016.
[5] H. Varet, L. Brillet-Gueguen, J. Y. Coppee, and M. A. Dillies, SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data, PLoS One, vol. 11, no. 6, p. e0157022, 2016.
[6] G. Dennis, Jr. et al., DAVID: Database for Annotation, Visualization, and Integrated Discovery, Genome Biol, vol. 4, no. 5, p. P3, 2003.
[7] J. Wixon and D. Kell, The Kyoto encyclopedia of genes and genomes--KEGG, Yeast, vol. 17, no. 1, pp. 48-55, Apr 2000.
[8] G. Yu, L. G. Wang, Y. Han, and Q. Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters, OMICS, vol. 16, no. 5, pp. 284-7, May 2012.
[9] J. Z. Deng, Meng Yao, Xiaohui Cheng, Yuehan Zhou, PRLR is a target of oncogenic miR-204 in Low Grade Glioma and involved in a ceRNA network by Computational Biology, presented at the AMMSO2019, guilin, 2019.4.22, 2019.
[10] J. Z. Deng, Xiaohui Cheng, Yuehan Zhou, Gene clustering, enrichment and survival analysis of differentially expressed genes in Low Grade Glioma between different genders by big data analysis, presented at the ICDMML2019, Hong Kong, 2019.4.28, 2019.
[11] G. Xu, Y. Zhou, and F. Zhou, Development and validation of an immunity-related classifier of nine chemokines for predicting recurrence in stage I-III patients with colorectal cancer after operation, Cancer Manag Res, vol. 10, pp. 4051-4064, 2018.
[12] W. N. Dudley, R. Wickham, and N. Coombs, An Introduction to Survival Statistics: Kaplan-Meier Analysis, J Adv Pract Oncol, vol. 7, no. 1, pp. 91-100, Jan-Feb 2016.