Identification of key genes affecting disease free survival time of pediatric acute lymphoblastic leukemia based on bioinformatic analysis

Hai-Yan Gao [a], Xin-Guo Luo [b], Xi Chen [a], Jing-Hua Wang [a,*]

[a] Department of Hematology, The Second Affiliated Hospital in Harbin Medical University, Harbin 150086, China
[b] Department of Hematology, Jinhua People's Hospital, Jinhua 321000, China

ARTICLE INFO

ABSTRACT

The poor prognosis of pediatric acute lymphoblastic leukemia (ALL) indicates the existence of key candidate genes that affect pediatric ALL and its prognosis. The limma package in R was applied to screen differentially expressed genes (DEGs), and the Survival package and KMsurv package in R were used to screen disease free survival time related genes (prognosis genes). Then, based on latent pathway identification analysis (LPIA), latent pathways were identified, and pathway–pathway interaction network was constructed and visualized by Cytoscape. Based on the expression values of 8284 genes in 126 chips, 2796 DEGs and 353 prognosis genes were screened out. After overlapping DEGs and prognosis genes, 75 key genes were identified, which were most significantly enriched in 25 GO functions and chronic myeloid leukemia pathway. For the 75 key genes, 27 disease risk sub-pathways were identified, and HK3, HNMT, SULT2B1, KYNU, and PTGS2 were the significant key genes which were enriched in these sub-pathways. Furthermore, based on pathway–pathway interaction analysis, HK3 and PTGS2 were predicted as the most important genes. Through glycolysis and arachidonic acid metabolism, HK3 and PTGS2 might play important roles in pediatric ALL and its prognosis, and thus, might be potential targets for therapeutic intervention to suppress pediatric ALL.

© 2014 Elsevier Inc. All rights reserved.

Introduction

Acute lymphoblastic leukemia (ALL), a cancer of lymphoid progenitor cells, is characterized by the overproduction of cancerous and immature lymphoblasts in the bone marrow, causing death by inhibiting the production of functional blood cells and by infiltrating into other organs [1]. ALL is common in childhood with a peak incidence at ages 2–5, resulting in pediatric ALL. Pediatric ALL represents over a quarter of pediatric cancers, and about 3 quarters of childhood leukemia, attracting much attention of pathologists.

Much effort has been made to understand the pathogenesis of pediatric ALL, and several oncogenes have been found in the past few years, including BAX, PAX5, IKZF1, EBF1, TCF3, LEF1, BCR–ABL, and MLL [2,3]. Among these genes, the BCR–ABL (break-point cluster region gene, Abelson oncogene) fusion gene is formed by chromosome translocation of t(9;22)(q34;q11) [2], coding fusion proteins of 180–210 kD. With enhanced activity of tyrosine kinase, these fusion proteins disturb the functions of signaling pathways, and thus, inhibit the apoptosis of white blood cells, which then results in leukemia [2,4]. Moreover, the

rearrangement of MLL (mixed lineage leukemia) gene is proved to be able to cause pediatric ALL and poor prognosis [5]. The abnormalities of these disease genes have been used as the clinical diagnostic tools for ALL [2], and helped in the development of effective new drugs like imatinib mesylate [6]. Because of the remarkable improvements in clinical diagnosis and new chemotherapeutic agents, the cure rates of pediatric ALL generally exceed 75% [1,7]. However, approximately 25% of the patients still suffer from relapse, indicating the existence of key candidate genes that might play important roles in the pathogenesis of pediatric ALL and its prognosis [8].

With the development of high-throughput technology and data processing, differentially expressed gene (DEG) screening and pathway enrichment have been well established and widely used to identify key genes in pathogenesis [9]. However, complex diseases usually involve various biological pathways that may interact with each other [10]. This phenomenon is ignored by normal pathway enrichment analysis, leading to meaningless predictions. Due to the fact that enriched pathways interact with each other through shared proteins and metabolites, a functionally based approach has been developed to find possible pathway cross-talk by integrating high-throughput data, protein–protein interactions, and biological pathway information [10,11]. The pathway–pathway interaction analysis has been applied to explore insulin resistant in response to a glucose challenge in mouse liver [10].

* Corresponding author at: Department of Hematology, The Second Affiliated Hospital in Harbin Medical University, No.246, Xuefu Road, Nangang District, Harbin 1150086, China. Fax: +86 451 86605269.
E-mail address: JingHuaWangjhw@163.com (J.-H. Wang).

In this study, using the gene expression data of lymphocytes in the bone marrow of pediatric ALL patients and healthy controls, DEGs and disease free survival (DFS) time related genes (prognosis genes) were screened and overlapped, resulting in the identification of disease related key genes. Additionally, pathway and function enrichment, disease risk sub-pathway analysis, and pathway–pathway interaction analysis were performed to identify the most important genes influencing DFS time of pediatric ALL patients, providing a promising platform for further study.

Materials and methods

Microarray data

The gene expression file (ID: E-MTAB-1216) of lymphocytes in the bone marrow of pediatric ALL patients was downloaded from the European Bioinformatics Institute (EBI) Array Express database (http://www.ebi.ac.uk/arrayexpress). EBI Array Express database is an international public gene expression database which stores both raw and normalized data generated by sequencing or array-based technologies [12]. A total of 101 chips were available, including 80 of ALL children aged 0.1667–16.25 (average age: 6.7 ± 4.2), and 21 of neonates with ALL. It should be noted that 80 of these pediatric ALL patients have the records of DFS time. The corresponding platform was Affymetrix GeneChip Human Genome HG-U133A [HG-U133A]. Furthermore, the gene expression file (ID: GSE11504) of healthy controls, was also downloaded from National Center of Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/) [13]. A total of 25 chips were available, and the corresponding platform was GPL570, Affymetrix Human Genome U133 Plus 2.0 [HG-U133_Plus_2].

Data preprocessing and DEG screening

The probe-level data were converted into expression values, which were averaged for every gene. The preprocess Core package in R [14] was used to standardize the expression values. The limma package in R [15] was applied to identify genes that were significantly differentially expressed between the case group and the control group. The p values were adjusted by the Benjamin–Hochberg (BH) method [16], and

Table 1
The pathway and function enrichment of 16 positive prognosis genes which were also down-regulated DEGs.

Category	Term	Count	p value
GOTERM_BP_FAT	GO:0006983 ~ ER overload response	2	0.009571419
GOTERM_BP_FAT	GO:0010165 ~ response to X-ray	2	0.014325351
GOTERM_BP_FAT	GO:0034976 ~ response to endoplasmic reticulum stress	2	0.032198852
GOTERM_BP_FAT	GO:0006984 ~ ER-nuclear signaling pathway	2	0.033131223
GOTERM_CC_FAT	GO:0031981 ~ nuclear lumen	5	0.01298958
GOTERM_CC_FAT	GO:0070013 ~ intracellular organelle lumen	5	0.02633158
GOTERM_CC_FAT	GO:0043233 ~ organelle lumen	5	0.028442367
GOTERM_CC_FAT	GO:0031974 ~ membrane-enclosed lumen	5	0.030382961
GOTERM_CC_FAT	GO:0016363 ~ nuclear matrix	2	0.038758381
GOTERM_MF_FAT	GO:0003697 ~ single-stranded DNA binding	2	0.049688384
KEGG_PATHWAY	hsa05220:Chronic myeloid leukemia	3	0.005674829
GOTERM_MF_FAT	GO:0004252 ~ serine-type endopeptidase activity	4	0.024888368
GOTERM_MF_FAT	GO:0008233 ~ peptidase activity	7	0.028719128
KEGG_PATHWAY	hsa04610:Complement and coagulation cascades	3	0.047839857

DEGs: differentially expressed genes; KEGG: Kyoto Encyclopedia of Genes and Genomes; GO: gene ontology.

only the genes with |log fold change (FC)| larger than 1 and adjusted p value less than 0.05 were selected as DEGs.

Prognosis gene and key gene screening

In the case group, 80 chips with records of DFS time were selected for prognosis gene screening. Firstly, the 80 samples were divided into high- and low-expression groups according to the average expression values of genes. Secondly, survival analysis was performed based on the Survival package and KMsurv package in R [17,18]. The Log-Rank test [19] was conducted to investigate the significance of the difference between the two separated groups. Only the genes with p value less than 0.01 were selected as prognosis genes, which were divided into positive and negative prognosis genes. Positive prognosis genes cause longer DFS time with higher expression, while negative prognosis
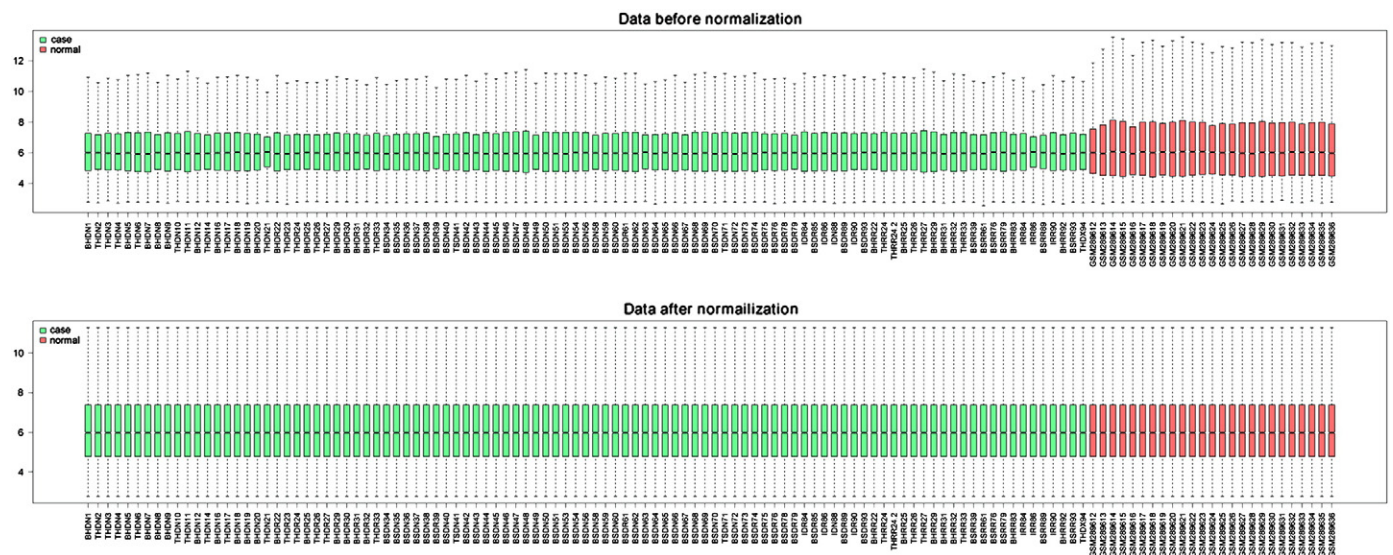
Fig. 1. Expression values before and after standardization. Green: pediatric ALL patients; red: healthy controls.