Discovery of Novel Gene Biomarker For Acute Myeloid Leukemia Through Differential Gene Expression Analysis

Sweta P Tripathi*, Himanshu A Pandya

Department of Bioinformatics, Gujarat University, Gujarat, India

Keywords: Acute Myeloid Leukemia Disease, Microarray Analysis, Differentially Expressed Genes, Pathway Analysis.

ABSTRACT

Acute Myeloid Leukemia (AML) is the major cause of death in developing countries. It is more prevalent in males as compared to females. Various cytogenetic detection methods are currently available, but their results are varied due to chromosomal changes. To overcome this, multiple microarray studies have been conducted on transcriptional profiling of AML patients. In depth reanalysis of microarray studies was performed. It was found that 21 genes involved in the seven pathways. The most significant pathway is antigen presentation and processing and genes are CD4, CD8a, and CD8b. On the basis of this result, we suggested that genes CD4, Neuraminidase (NA), Androgen-dependent TFPI-regulating (ADTRP) are helpful in discovering novel biomarkers and also as a therapeutic target of AML.

*Corresponding author:
Sweta Tripathi, Department of Bioinformatics, Gujarat University, Gujarat, India
Phone: +91 9722641135, 380024
E-mail: sweta.tripathi2012@gmail.com
Introduction
Leukemia is cancer of the blood and bone marrow. It can be either acute or chronic [1]. Further, it is classified on the basis of origin into myeloid and lymphoblast [2]. Acute myeloid leukemia (AML) is the diversified disorder which is caused by the accumulation of abnormal white blood cells in the bone marrow [3] or when the blood cells are not able to produce granulocyte or monocytes [4] resulting in the formation of myeloid cells. These myeloid cells differentiate into haematopoietic progenitor cells clump together and form the blast cells. The accumulated blast cells in the bone marrow lead to diseases like anaemia, blood clotting and severe infection caused the death from AML [5]. In human T-cell three types of oncogenes HTLV-1, HTLV-2 and simian T-cell are responsible for the AML [6]. These oncogenes forms Natural killer T cell complex (NKT) which kills the target cells (MHC-I, MHC-II). Besides this, genes involved in inhibition of T-cell proliferation are also known to cause AML [7].

The microarray technology based study on AML patients is very useful in identification genomic mutations. Microarray gene expression profiling has identified various upregulated and down regulated genes PML, TNFα, JUNB, FOS, TP63, VEGFA and ID2 [8]. Cytogenetic analysis of certain patients with AML led to identification of 133 genes [9].

This study aims to identify biomarkers of AML in differentially expressed genes involved in the different pathways by reanalysis of transcription profile of AML patients available in ArrayExpress database.

Materials and Methods
Data mining from the microarray databases: The microarray data is available in the European Molecular Biology Laboratory Array Express database [10]. It consists of 47068 experiments and 1336717 bio assays. There are two types of data: Minimum Information about Microarray Experiments [11] and Minimum Information about a high-throughput SEQuencing Experiment. The data is available for download in two formats viz. CEL (Affymetrix) file and MAGE-ML format. E-GEOD-14924 the transcriptional profiling of human T-cell in the AML CEL files was downloaded. The file contains 41 samples which constitute four types viz.: AML with CD4 (n=10) and CD8 (n=10) and two healthy samples with CD4 (n=10), and CD8 (n=11). All these samples are taken from the patients of leukemia and healthy persons.

The steps involved in data mining and processing are summarized in figure1

Pre-processing and Processing of Raw Data: The Bioconductor software [12] and its packages [13] were downloaded and installed. It was used to load the downloaded CEL files and further processing of raw data. The RcolourBrewer package was used for the colour of probes, after the loading step we check the quality control of the raw data through box plot and find its standard errors, removed all the variability between each of the sample file. The AffyPLM package [14, 15] was used to generate pseudo CEL file image. The germa package was used to normalize the raw data [16, 17, 18]. This step involves three results probe PM and MM background correction, adjusts the binding and non-binding specificity and expression calculation on the basis of GC% content then normalized the sample data. These data is transformed into the log2 based transformation in the excel file. After normalization step a quality control check was carried out using arrayQualityMetrics [19].

Statistical Analysis: The Genefilter [20] package was used to filter the microarray data and limma package [21] for the differential gene expression identification. Annotationdbi package was used for the annotation of those differentially expressed genes. On the basis of these a heat map was generated of differentially expressed genes.

Pathway Enrichment Analysis: DAVID functional annotation tool was used to determine the pathways involved [22].

Result and Discussion
Normalization and Quality Control Checks: A boxplot was generated for the raw and normalized data and further transformed into log2 based values (Figure 2.A and B). This box plot was also based on the interquartile range (IQR) values which lie between 5.5 to 8.2. Figure 2.A illustrates the raw data boxplot indicating that array 29 has a high level probe intensity value as compared to other arrays [23]. Probe intensity can be adjusted through normalization. Hence the normalized boxplot shows equal probe level intensity for all the arrays (Figure 2.B).

Heatmap for the differentially expressed gene
The hierarchical clustering based heatmap is a useful method for microarray analysis [25, 26]. The heat map was designed for differentially expressed genes based on their statistical values (Figure 3). The heat map shows probe intensities for each gene probe_id indicating the up regulated (yellow) and down regulated (red) genes. These expressions are based on logFC value and P-value. The logFC value for up regulated genes lies in the range 4 to 8 and for down regulated genes it lies between -4 to -5 whereas the P-value should be between P<0.05 to P<5*10-8. The probe_ids for up regulated genes are 205758_s_at, 207795_s_at, 1555691_a_at, 215332_s_at. These genes...
Fig. 1: Schematic of microarray analysis

Fig. 1: Quality control check (A) Raw data box plot (B) Normalized data box plot

http://www.pacificjournals.com/aabs
Fig. 3: Heat map for differentially expressed genes

Fig. 4: Pathways summary for the Acute myeloid leukemia

Figure 4: Pathways summary for the Acute myeloid leukemia

Fig. 4: Pathways summary for the Acute myeloid leukemia
have higher intensity as compared to rest of the genes. The Probe_id for down regulated genes are 232584_at, 203547_at, 229070_at have low intensity based expression.

Limma package is used for differential gene expression analysis. It takes all the different experimental condition for microarray data as input. The result is given in a tabulated format which contains 24 unique differentially expressed genes(Table.1).Amongst the 24, 3 genes are not identified and properties for the remaining 21 (listed above) including gene symbol, name, function and expression have been shown. The gene expression can be identified based on P-value and logFC values. If the P-value is less than P< 0.05 and logFC <1.5 – 4 or more, it is considered significant [27, 28]. Out of the 21 genes; XCL1,GPR56,GZMB,GZMH,CCL4 have been reported in AML disease gene expression, S1PR5,CD160,SLAMF7,NEG7 have not been reported in AML disease patients. On the basis of gene ontology, 7 genes have been selected for differentiation studies between AML and normal tissues.

**CD8 B molecule:** CD8 molecule is involved in the T cell mediated killing process [29] and was up regulated (logFc 7.10) in the present study. The CD8 T cells have an abnormality that is instead of binding to the target cell, it kills the target cell (MHC class I) molecules. As a result, the antigen specific activation and cytokine production process is stopped. However, reports have been published showing relationship between CD8 and AML [30].

**NK2G (natural killer gene family):** This family includes KLRK1 (Killer cell lectin-like receptor subfamily K, member1), KLRD1(Killer cell lectin-like receptor subfamily D, member 1), KLRC3(Killer cell lectin-like receptor subfamily C, member 3), KLRC4(Killer cell lectin-like receptor subfamily C, member 4), KLRF1(Killer cell lectin-like receptor subfamily F, member 1) genes. They form the natural killer complex and are present on chromosome 12.In the present study, expression of these genes was shown to be upregulated (logFc 5.40 to 7.45). The KLRK1 gene is involved in increased production of multiple myeloma cells (plasma cells) .Multiple myeloma stage II and III are involved in acute lymphoblastic leukemia [30]. The KLRD1 gene is infected by cytomegalovirus resulting in inhibition of major histocompatibility complex class II by the interuption of the Jak/Stat pathway[31].The KLRC3 gene is involved in the cytoytic process of natural killer T cells. The KLRC4 gene activates or inhibits natural killer T cells. The KLRF1 gene stimulates the natural killer cell mediated cytotoxicity and cytokine release [32]. Hence, the NK2G gene family involved in the NKT celllymphoblastic leukemia, breast cancer disease [33].

**CRTAM (Cytotoxic and regulatory T cell molecule):** It belongs to nectin family of proteins. It is involved in tumour immunosurveillance and act as a supressor for the NKT cellreceptor surface proteins. In the present study, it was shown to be upregulated (logFc 5.40).Also, reports have been published showing relationship between CRTAM and lung cancer [34], breast cancer [35] , AML and chronic leukemia [36].

**CD4 molecule:** It is used for lymphocyte count and present on chromosome 7. It form neoplasms in lymphoma and hence has been shown to cause AML. In this study it was downregulated (logFc -4.44).

**ENPP5 (Ectonucleotide phosphodiesterase5):** It is involved in the neuronal cell communication and was upregulated (logFc 4.16) in the present study. Though its pathway is not reported, it has been shown to cause chronic myeloid leukemia disease.

**NA(Neuraminidase):** It is involved in neuroacanthocytosis disorder related to the brain tissue and spiculated blood cells. In the present study it was downregulated (logFc -5.600) and reported in AML disease (T lymphocytes) and carcinoma (B lymphocytes) gene expression [37].

**ADTRP (androgen-dependent TFPI-regulating):** It is involved in the tissue factor inhibitor pathway, it regulates the cell associated anticoagulating activity of TFPI inhibitor in endothelial cells [38] and was down regulated (logFc -4.30) in this study. Its cancer related study and pathway is not reported yet.

**Pathway Enrichment Analysis**

For differentially expressed genes, KEGG pathway based analysis was used which showed 7 differentially regulated pathways.CD4 molecules are involved in antigen processing and presentation, primary immunodeficiency, T-cell receptor signalling pathway, hematopoetic lineage and cell adhesion molecule. CD8 molecules are involved in all CD4 molecule pathways and additionally in Graft-Versus-Host disease pathway. A new pathway was derived from these studies i.e. Vascular tissue factor inhibitor pathway. This pathway and the gene (ADTRP) involved in it have not been reported yet.

The significance of the pathways is indicated based on FDR (False discovery rate) and P-value (<0.05). It was found that the most significant pathway is antigen processing and presentation pathway (Table 2) which involves 7 genes: CD4, CD8a, CD8b, KLRC1, KLRC3, KLRC4, KLRD. Since CD4, CD8a and CD8b genes are involved in 5 pathways, they are considered the most significant genes. On the basis of this study, a flowchart of the pathways involved in acute myeloid leukemia was made.
Conclusion
We analyzed the transcriptional profiling of human T-cell involved in the AML disease. We found the differential expressed genes CD8B, KLRK1, CRTAM, KLRD1, ENPP5, NA, CD8A, KLRC3, KLRC4, KLRF1, S1PR5, XCL1, S1PR5, CD160, SLAMF7, GPR56, ADTRP, GZMB, FGFBP2, NKG7, CCL4, GNLY have varied expression in AML patients. We suggested that CD4, NA, ADTRP genes may be helpful to discover the novel diagnostic biomarker, therapeutic targets and gene therapy for the AML disease.

Funding Source
None

Conflict of Interest
None

References
leukemias”. Leukemia, 18(4), 665–9. doi:10.1038/sj.leu.2403317


38. Gentleman R, Carey V, Huber W, Hahne F:Genefilter:method for filtering from microarray experiments. R package version1.44.0