King Saud University

Saudi Journal of Biological Sciences

www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

Identifying new targets in leukemogenesis using computational approaches

Archana Jayaraman [a,b], **Kaiser Jamil** [a,*], **Haseeb A. Khan** [c]

[a] *Centre for Biotechnology and Bioinformatics, School of Life Sciences, Jawaharlal Nehru Institute of Advanced Studies (JNIAS), Secunderabad, Telangana, India*
[b] *Center for Biotechnology, Jawaharlal Nehru Technological University (JNTUH), Kukatpally, Hyderabad, Telangana, India*
[c] *Department of Biochemistry, College of Sciences, Bldg. 5, King Saud University, P.O. Box 2455, Riyadh, Saudi Arabia*

Abstract There is a need to identify novel targets in Acute Lymphoblastic Leukemia (ALL), a hematopoietic cancer affecting children, to improve our understanding of disease biology and that can be used for developing new therapeutics. Hence, the aim of our study was to find new genes as targets using *in silico* studies; for this we retrieved the top 10% overexpressed genes from Oncomine public domain microarray expression database; 530 overexpressed genes were short-listed from Oncomine database. Then, using prioritization tools such as ENDEAVOUR, DIR and TOPPGene online tools, we found fifty-four genes common to the three prioritization tools which formed our candidate leukemogenic genes for this study. As per the protocol we selected thirty training genes from PubMed. The prioritized and training genes were then used to construct STRING functional association network, which was further analyzed using cytoHubba hub analysis tool to investigate new genes which could form drug targets in leukemia. Analysis of the STRING protein network built from these prioritized and training genes led to identification of two hub genes, *SMAD2* and *CDK9*, which were not implicated in leukemogenesis earlier. Filtering out from several hundred genes in the network we also found *MEN1*, *HDAC1* and *LCK* genes, which re-emphasized the

* Corresponding author. at: Centre for Biotechnology and Bioinformatics, School of Life Sciences, Jawaharlal Nehru Institute of Advanced Studies (JNIAS), Buddha Bhawan, 6th Floor, M.G. Road, Secunderabad 500003, Telangana, India. Tel.: + 91 9676872626; fax: +91 40 27541551.
  E-mail address: kaiser.jamil@gmail.com (K. Jamil).

ELSEVIER | **Production and hosting by Elsevier**

1. Introduction

A key focus of cancer research is the identification of driver genes in the tumorigenesis pathway as tumor specific signature genes, for use as drug targets or biomarkers, which could be possible from microarray databases (Ma et al., 2013). The recent advancement in bioinformatics techniques has made it possible to search for therapeutic targets for specific diseases in a systematic and comprehensive manner (Desany and Zhang, 2004). Acute Lymphoblastic Leukemia (ALL) is a blood cancer that targets B and T-lymphocyte cells, affecting their differentiation and leading to the loss of regulation of cell division (Khalid et al., 2010). Even with numerous advances in therapeutic efficacy, 20–40% of patients still relapse, especially children and young adults (Smith et al., 2010). Research studies have implicated alterations in several pathways that mediate crucial biological processes to play a role in disease progression and particularly in relapse (Bhojwani et al., 2006; Pui et al., 2011). These studies suggest that an interconnected network of many genes and their products are altered in carcinogenesis and may contribute to leukemia pathogenesis (Bhojwani et al., 2006; Pui et al., 2011).

A study by Kang et al. (2012) reported a correlation between event free survival and expression levels of *NEGR1*, *IRX2*, *EPS8* and *TPD52*. Lin et al. (2012) reported that point mutations in *NOTCH1* led to increased expression of this gene which might contribute to pathogenesis in T-ALL. In recent years, meta-analysis studies have led to the identification of novel genetic markers that might play crucial roles in the neoplastic process and in other diseases, as demonstrated through our previous studies (Khan and Jamil, 2008; Shaik et al., 2009; Jamil and Sabeena, 2011). Understanding the evolutionary relationship of these genes could also help to investigate the mechanisms of neoplastic transformation observed in leukemic cells (Jayaraman et al., 2011; Jayaraman and Jamil, 2012). Further, our previous studies using bioinformatics approaches have helped in highlighting the significance of protein networks in ALL (Jayaraman and Jamil, 2013) and identified important amino acid residues that may be useful in therapeutic targeting of cell cycle proteins (Jayaraman and Jamil, 2014). In recent years, several research studies have applied a systems biology approach to understand ALL leukemogenesis. Maiorov et al. (2013) identified a set of non-differential putative biomarkers in T-ALL based on network analysis of expression data. Gao et al. (2014) analyzed differentially expressed genes, screened for prognostic genes and identified latent pathway genes. Their analysis identified *HK3* and *PTGS2*, two key metabolic pathway genes as possible prognostic genes in pediatric ALL. Chaiboonchoe et al. (2014) used an integrated bioinformatics approach to identify glucocorticoid regulated genes in Childhood ALL. Many studies have shown that various bioinformatics and computational biology approaches, such as PseKNC (Chen et al., 2014) or Chou's

PseAAC (Chou, 2001), can be successfully used to identify modifications in the genome such as recombination spots of DNA (Chen et al., 2013), various PTM (posttranslational modification) sites (Xu et al., 2014), anticancer peptides (Hajisharifi et al., 2014), interactions between drugs and target proteins in cellular networking (Xiao et al., 2014), providing very useful information and insights for both basic research and drug development, and hence are widely welcomed by the scientific community, both experimental and theoretical. Here, we have used computational approaches to identify new targets in leukemogenesis in the hope to provide useful information for stimulating the development of new and effective drugs to treat leukemia.

Understanding the interactions of disease genes is essential as dynamic networking of genes could be correlated with clinical informatics, including therapeutic and imaging profile and other parameters and this correlation could help in a better understanding of the disease in relation to each patient (Wang, 2011). To meet such challenges our objective was to retrieve overexpressed genes from Oncomine expression database (Rhodes et al., 2007), to perform gene prioritization analysis using bioinformatics software. Further, we have also analyzed protein interactions of the prioritized proteins as studies investigating Protein–Protein interactions have provided key insights on the biological functioning of many proteins and have also been effective in identifying novel genes that play a role in pathogenesis of various diseases such as cancer (Huang et al., 2011; Li et al., 2012). Our hypothesis is based on our belief that a large amount of data generated through expression studies in previous reports which contribute to leukemogenesis may have been missed due to the varied detection methods. Hence, our research combines the use of expression data, gene prioritization analysis and a network based approach to identify genes of significance in ALL and the use of well validated datasets, prioritization based approach, using rigorous network analysis suggests that the results from our study may be replicable *in vivo* as well. The use of these combined bioinformatics approaches enhances the validity of our results and has led to the identification of few novel genes in this study.

2. Materials and methods

An overview of the analysis workflow for the study is represented in Fig. 1.

2.1. Microarray expression data analysis using Oncomine database

In the current study we queried the Oncomine database to obtain only those datasets which have reported differentially expressed genes between normal and leukemic tissues. Oncomine database 3.0 (Rhodes et al., 2007) is a comprehensive