A study on expression of genes in various types of leukemia using *in silico* tools - a therapeutic approach

SUGANYA M

17PBI007

Thesis submitted to

DEPARTMENT OF BIOCHEMISTRY, BIOTECHNOLOGY AND BIOINFORMATICS

Avinashilingam Institute for Home Science and Higher Education for Women,

Coimbatore - 641043

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Bioinformatics

April 2019

A study on expression of genes in various types of leukemia using *in silico* tools - a therapeutic approach

SUGANYA M

17PBI007

Thesis submitted to

DEPARTMENT OF BIOCHEMISTRY, BIOTECHNOLOGY AND BIOINFORMATICS

Avinashilingam Institute for Home Science and Higher Education for Women,

Coimbatore - 641043

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Bioinformatics

April 2019

25/4/19

Signature of the

Head of the Department

Signature of the

Supervisor

ACKNOWLEDGEMENT

ACKNOWLEDGEMENT

I owe a special tribute to **God Almighty** for the opportunity given to me to take up this work and also for inner strength to complete my work successfully.

I am grateful to **Dr. P. R. KRISHNA KUMAR,** Chancellor, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore for providing all facilities necessary for the study.

I am thankful to **Dr. PREMAVATHI VIJAYAN,** Vice Chancellor, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore for providing all facilities necessary for the study.

I am thankful to **Dr. S. KOWSALYA**, Registrar, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore for providing all facilities necessary for the study.

I am filled with gratitude to **Dr. P. R. PADMA,** Professor, Dean and Head of the Department of Biochemistry, Biotechnology & Bioinformatics, Avinashilingam Institute for Home Science and Higher Education for Women for her moral support and encouragement to complete the project successfully.

I am humbled and grateful to **Dr. R. NIRMALADEVI,** Assistant Professor, Department of Biochemistry, Biotechnology & Bioinformatics, Avinashilingam Institute for Home Science and Higher Education for Women, for the opulent guidance rendered at every stage of the dissertation. I feel immense pleasure in extending my deep sense of gratitude for her motivation, crucial support, guiding suggestions and encouragement throughout the work.

I heartfully thank to express my gratitude towards **Dr. N. SHANTHI,** Assistant Professor, Department of Biochemistry, Biotechnology & Bioinformatics, Avinashilingam Institute for Home Science and Higher Education for Women, for making the project successful.

I am gently privileged to express my gratitude towards **Dr. M. RAJESWARI,** Assistant Professor, Department of Biochemistry, Biotechnology & Bioinformatics, Avinashilingam Institute for Home Science and Higher Education for Women, for inspiring thoughts rendered for carrying out this dissertation successful.

I am gently privileged to express my gratitude towards **Dr. A. SHOBANA,** Assistant Professor, Department of Biochemistry, Biotechnology & Bioinformatics, Avinashilingam Institute for Home Science and Higher Education for Women, for constant support and tremendous care rendered for carrying this dissertation successful.

I express my deep sense of gratitude to all **STAFF MEMBERS** of Biochemistry, Biotechnology & Bioinformatics, Avinashilingam Institute for Home Science and Higher Education for Women for their immense support.

I express my sincere heart bound thanks and gratitude to my **PARENTS** and **SISTER** for their emotional support to complete the dissertation successful.

SUGANYA M

CONTENTS

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1

INTRODUCTION

A study on expression of genes in various types of leukemia using in silico tools - a therapeutic approach

1. INTRODUCTION

Cancer is the second or third cause of death in developing countries. It has been estimated that 15 million people will die from cancer in 2020. Hematological malignancies are such cancers that initiate in the blood progenitor cells i.e. cells of bone marrow or immune system. The malignances of red blood cells include anemia and polycythemia while the tumors of white blood cells include leukocytosis, leukopenia, leukemia, lymphoma and multiple lymphomas. Leukemia can be characterized as a disorder which is caused due to the uncontrolled proliferation of immature hematopoietic white blood cells in bone marrow (BM). This word leukemia was first coined in mid-19th century by Greek word 'leukos' 'white' + 'haima' 'blood'. Hematological malignancies are caused by certain disturbances in Extracellular Matrix (ECM) (Shahzad *et al*., 2017).

Leukemias are a group of cancers that originate from blood-forming tissues. The name of the disease is derived from the Greek word 'leukos' for 'white blood'. In this disorder an abnormally large number of immature white blood cells (WBCs) is produced by the bone marrow. These leukemic WBCs eventually replace the normal ones, resulting in the clinical manifestation of anemia, leaving the body more susceptible to infection. Leukemia is classified into four main categories or subtypes according to cell type and rate of growth: acute lymphocytic leukemia (ALL) derived from immature T- or B-lymphocytes, most common in children; acute myeloid leukemia (AML) from immature myeloid cells, most common in adults; chronic lymphocytic leukemia (CLL) from mature B-lymphocytes, mostly an adult disorder; and chronic myelogenous leukemia (CML) from granulocyte precursors, most common in adults (Aboul-Soud *et al*., 2016).

According to the World Health Organization, death from cancer is expected to increase 104% worldwide by the year 2020. However, estimated 600,000-700,000 deaths in India were caused by cancer. Leukemia is the 11th most common cancer worldwide, with around 352,000 new cases diagnosed in 2012 [2% of the total]. In 2014, it is estimated that there will be 52,380 new cases of Leukemia and an estimated 24,090 people will die of this disease. The Leukemia is a group of disorder characterized by malignant transformation of blood forming cells. The proliferation of leukemic cells takes place primarily in the bone marrow, and in certain forms, in

the lymphoid tissue. Ultimately, the abnormal cells appear in the peripheral blood raising the total white cell count to high level In addition, feature of bone marrow failure (e.g. anaemia, thrombocytopenia, neutropenia) occurs (Sharma and Porte., 2016).

BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

Bioinformatics involves the integration of computers, software tools, and databases in an effort to address biological questions. Bioinformatics approaches are often used for major initiatives that generate large data sets. Two important large-scale activities that use bioinformatics are genomics and proteomics. Bioinformatics tools aid in the comparison of genetic and genomic data and more generally in the understanding of evolutionary aspects of molecular biology. At a more integrative level, it helps analyze and catalogue the biological pathways and networks that are an important part of systems biology. Bioinformatics is important to genetic research because genetic data has a context. The context is biology. Life forms have certain rules of behavior. The same applies to tissues and cells, genes and proteins (https://www.scq.ubc.ca/what-is-bioinformatics/).

Currently, microarray technology has revealed the guiding principles for the molecular initiation and progression of complex diseases, enabling investigators to explore potential molecular biomarkers for early detection of cervical cancer. Growing evidence has demonstrated that differential expression (DE) analysis and differential coexpression (DC) analysis are powerful tools to explore diagnostic gene signatures and biological processes of complex diseases and contribute greatly to the understanding of gene regulation systems. It is well confirmed that the propensity of many diseases can be reflected in the difference of gene expression levels. In DE analysis, genes that present different expression levels across different conditions are identified, which is conductive to identifying cancer-specific gene signatures for distinguishing cancer patients from normal controls, and screening underlying candidate genes that assist better diagnosis and treatment of diseases at molecular level. Different from DE analysis, DC analysis aims to study the potential interactions among individual genes, and further to reveal altered regulatory mechanisms by analyzing the difference in gene coexpression patterns between disease and normal subjects. In general, both DE and DC analyses are beneficial for gene expression analysis, while they present different performance characteristics,

one for individual genes and one for intrinsic gene interactions. The integration of two types of strategies might improve the testing power and provide new insights on dissecting complex disease mechanism (Fang *et al*, 2018).

The recent exponential decrease in sequencing cost and advancement in microarray technology has resulted in an accumulation of large gene expression datasets; many of which are publically available on the Gene Expression Omnibus (GEO) (Barett *et al*. 2015). Despite having access to numerous datasets, analysing them can be challenging and time-consuming. Typical analyses on gene-expression data includes differential gene expression analysis (DGEA) and Gene-set Analysis (GSA), which are used to find statistically significant differences in expression for specific genes or gene-sets between sample populations. There are a wide variety of R packages that can be used to analyse gene expression data (e.g. Robinson *et al*., 2010; Ritchie *et al*., 2015).

However, the use of these R packages and the subsequent generation of visualisations require extensive coding proficiency in the R programming language (R Core Team, 2013). Existing tools, such as the Geo2R (Barrett *et al*., 2012) and GSEA (Gene-Set Enrichment Analysis). GeoDiver identifies differentially expressed genes by fitting a linear model to each gene which estimates the fold change in expression while accounting for standard errors by applying empirical Bayes smoothing. Genes are then ordered according to the difference in the expression values between the two sets of samples selected. This information is presented as an interactive table, a heatmap and a volcano plot. Upon clicking on a gene, users are provided with an interactive bar chart displaying the gene expression levels for each sample expressing the gene. The Volcano plot has added interactivity showing the gene name, fold-change and p-value of each data point.

A microarray is a laboratory tool used to detect the expression of thousands of genes at the same time. Some of the most known public, curated microarray databases are ArrayTrack, NCI mAdb, ImmGen database, Genevestigator, Gene Expression Omnibus - NCBI, ArrayExpress at EBI, Stanford Microarray database, The Cancer Genome Atlas (TCGA), GeneNetwork system, UNC modENCODE Microarray database, UPSC-BASE, UPenn RAD database, UNC Microarray database, MUSC database and caArray at NCI. Gene expression

analyses study the occurrence or activity of the formation of a gene product from its coding gene. It is a sensitive indicator of biological activity wherein a changing gene expression pattern is reflected in a change of biological process. Differential expression analysis means taking the normalized read count data and performing statistical analysis to discover quantitative changes in expression levels between experimental groups. There are different methods for differential expression analysis such as edgeR and DESeq based on negative binomial (NB) distributions or baySeq and EBSeq which are Bayesian approaches based on a negative binomial model. It is important to consider the experimental design when choosing an analysis method. While some of the differential expression tools can only perform pair-wise comparison, others such as edgeR, limma-voom, DESeq and maSigPro can perform multiple comparisons. Protein-Protein Interaction network analysis is one of the most important tools for interpretation of molecular mechanisms in the process of carcinogenesis (https://www.ebi.ac.uk).

The overall five-year relative survival rate for leukemia has more than quadrupled since 1960. From 1960 to 1963, the five-year relative survival rate among whites (only data available) with leukemia was 14 percent. From 1975 to 1977, the five-year relative survival rate for the total population with leukemia was 34.2 percent, and from 2006 to 2012, the overall relative survival rate was 62.7 percent. From 2006-2012, the five-year relative survival rates overall were,

- CML - 65.9 percent
- CLL - 85.1 percent
- AML - 26.8 percent overall and 66.8 percent for children and adolescents younger than 15 years
- ALL - 70.7 percent overall, 92.3 percent for children and adolescents younger than 15 years, and 94.1 percent for children younger than 5 years.

In 2017, 24,500 people are expected to die from leukemia (14,300 males and 10,200 females). In 2009-2013, leukemia was the fifth most common cause of cancer deaths in men and the sixth most common in women. There are an estimated 363,794 people living with, or in remission from, leukemia in the US. In 2017, 62,130 people are expected to be diagnosed with leukemia (https://www.lls.org/http%3A/llsorg.prod.acquia-sites.com/facts-and-statistics/facts-and-statistics-overview/facts-and-statistics#Leukemia).

Leukemia is classified into four main categories or subtypes according to cell type and rate of growth: acute lymphocytic leukemia (ALL) derived from immature T- or B-lymphocytes, most common in children; acute myeloid leukemia (AML) from immature myeloid cells, most common in adults; chronic lymphocytic leukemia (CLL) from mature B-lymphocytes, mostly an adult disorder; and chronic myelogenous leukemia (CML) from granulocyte precursors, most common in adults. Due to the complex progression, the therapy is particularly challenging.

Since the survival rate of these types of leukemia when diagnosed is less. Therapeutic strategies must be adopted to increase the survival rate. In order to achieve a successful therapy for leukemia, genes involved in causing leukemia and their mechanism of action must be clear For which, the present study aims at understanding the molecular mechanisms of the different genes of various types of leukemia using in silico methods.

In future with this background, the following objectives were set for the present study entitled "**A study on expression of genes in various types of leukemia using i*n silico* tools - a therapeutic approach**".

- To identify and compare the differentially expressed genes in various types of Leukemia data sets
- To analyze Protein-Protein Interactions of Differentially Expressed Genes with significant leukemia pathways
- To understand the biological significance of these DEGs and their modulation for treatment and therapy for leukemia using *in silico* methods

2

REVIEW OF LITERATURE

A study on expression of genes in various types of leukemia using in silico tools - a therapeutic approach

2. REVIEW OF LITERATURE

The review of literature pertaining to the present study "**A study on expression of genes in various types of leukemia using *in silico* tools - a therapeutic approach**" is discussed below.

2.1. LIST OF GENES BASED ON LEUKEMIA TYPE

The following (Table 1) describes the prominent genes involved in various types of Leukemia disease (http://www.bioinformatics.org/legend/leuk_db.htm).

Table 1.
Genes involved in types of leukemia

Type of Leukemia	Gene Name
Acute Lymphoblastic Leukemia	MLLT2, MYC, ZNFN1A1,LAF4
Acute Myeloblastic Leukemia	ARNT
Acute Myelogenous Leukemia	IRF1, RGS2, GMPS
Acute Myeloid Leukemia	AF10, CBFB, NUP98, NUP214, HOXA9, REBBP, ARHGEF12, CDX2, LCP1, CEBPA, DEK, FUS, RUNX1
Acute Promyelocytic Leukemia	PML, THRA, NPM1
Acute Undifferentiated Leukemia	SET
B-cell Chronic Lymphocytic Leukemia	BCL3, BTG1
B-cell (acute) Leukemia	PBXP1
Chronic Lymphocytic Leukemia	DLEU1, DLEU2
Chronic Myelocytic Leukemia	AXL
Murine Myeloid Leukemia	EVI2A, EVI2B
Myeloid Leukemia	CDC23, CLC
pre B-cell Leukemia	PBX1, PBX2, PBX3
T-cell Leukemia	TCL6, TCL1B, TRA@, MTCP1, LDB1
T-cell Acute Lymphoblastic Leukemia	NOTCH1, NOTCH3, LYL1, HOX11, BAX, LMO1, LMO2, TAL1, TAL2
T-cell Leukemia (in lung carcinoma)	CAV1
Cutaneous T-cell Leukemia	NFKB2
Human Monocytic Leukemia	ETS1
Mast cell Leukemia	KIT
Mixed Linkage Leukemia	MLLT3, MLL3, LAF4

2.2. BIOLOGICAL BACKGROUND OF LEUKEMIA

Leukemia is a cancer of the blood or bone marrow. Bone marrow produces blood cells. Leukemia can happen when there is a problem with the production of blood cells. It usually affects the leukocytes, or white blood cells. The classification of Leukemia is principally derived in four types. The current approach to classifying leukemia is based on the 2016 World Health Organization (WHO) system (classification for hematopoietic neoplasms). The WHO classification is based on a combination of clinical, morphologic, immunophenotypic, and genetic features. Other less commonly used classification systems include the French-American-British (FAB) system, which is based on the morphology of the abnormal leukocytes. Leukemias are commonly also categorized as (http://cancer.columbia.edu/leukemia-classifications)

- Acute or chronic: Based on the percentage of blasts or leukemia cells in bone marrow or blood
- Myeloid or lymphoid: Based on the predominant lineage of the malignant cells

2.2.1. ACUTE LYMPHOBLASTIC LEUKEMIA (ALL)

Acute lymphoblastic leukemia is the most common pediatric cancer; it also strikes adults of all ages. Malignant transformation and uncontrolled proliferation of an abnormally differentiated, long-lived hematopoietic progenitor cell results in a high circulating number of blasts, replacement of normal marrow by malignant cells, and the potential for leukemic infiltration of the CNS and testes. ALL has been studied extensively for more than 4 decades. As a result, much is known about the cellular origin of this leukemia and the genetic mechanisms that lead to malignant transformation. Based on the expression of lineage-specific antigens and the presence of lineage-specific gene rearrangements, ALL cells are known to be derived from either B or T-cell precursors. In B-lineage ALL, malignant cells often have additional specific genetic abnormalities, which have a significant impact on the clinical course of the disease (Chiaretti *et al*., 2014).

2.2.2. CHRONIC LYMPHOCYTIC LEUKEMIA (CLL)

Lymphocytes evolve from immature stem cells, growing into T lymphocytes (T cells), B lymphocytes (B cells), or natural killer (NK) cells. Each of these plays a special role in the body's immune defense. In CLL, there are changes in the leukemic B cells that result in an over production of lymphocytes. In rare cases the T cells can be affected as well, resulting in T-cell prolymphocytic leukemia (https://www.genome.jp/kegg/pathway.html).

Little research has been performed to examine the natural clustering of CLL patient samples or to identify subtypes based on gene expression patterns, partly because expression studies in CLL patients have focused on the analysis and comparison of established disease subtypes. However, the identification of CLL patient groups is a current research goal, the realization of which could contribute to the identification of different prognostic subtypes and help to explain the heterogeneity in the clinical behavior of the disease (Yepes *et al*., 2015).

Estimating the course of the disease development at the diagnostic stage has historically relied on clinical staging systems, whereas the mutational status of the immunoglobulin heavy variable gene (IGHV), fluorescence *in situ* hybridization cytogenetics and presence of common somatic mutations contributed to refining of the prognostication later on. However, the former indicators do not fully explain the heterogeneity of the disease evolution before the therapeutic treatment. Essential DNA replication processes have, surprisingly, remained an underexplored source of biomarkers and anti-cancer targets for hematological neoplasms, especially CLL, probably because such an explorative approach was not intuitive for this malignancy. Indeed, CLL has been characterized by accumulation of malignant cells resting in quiescent, mostly G0 and early G1 phase of the cell cycle. In contrast to the non-proliferative peripheral blood compartment that is used for diagnostic purposes, a small actively proliferating fraction of CLL cells residing within the lymph nodes contributes to the daily generation of the leukemic clone. However, according to the international ethical guidelines applied, samples originating from the lymph nodes are not always readily available (Grgurevic *et al*., 2016).

2.2.3. ACUTE MYELOID LEUKEMIA (AML)

Acute myeloid leukemia (AML) is not a single disease but a group of neoplasms with diverse genetic abnormalities and variable responses to treatment. Cytogenetics and molecular analyses can be used to identify subgroups of AML with different prognoses. The pathogenesis of acute myeloid leukemia (AML) in many patients is linked to oncogenic fusion proteins, generated as a consequence of chromosome translocations or inversions. Many different translocations have been described in AML, the most frequent being the t(9;11), t(15;17), t(8;21), and inv(16), which, taken together with their variants, account for ≈20–30% of AML cases, although a recent analysis suggests that the proportion may be closer to 10% (Kumar 2015).

These recurring translocations are now the basis for classification of some patients with AML. Genome-wide gene expression profiling is becoming useful for the classification of many types of cancer, including AML and acute lymphoblastic leukemia. Although AML sub-types can be distinguished by oligonucleotide microarrays, the results of analysis of different translocations between laboratories are not always similar (Lee *et al*., 2016).

The development of high-throughput technologies, such as microarrays and next generation sequencing has contributed to progress in leukemia research. Since 1999, when the first applications of DNA microarrays in leukemia classification and outcome prediction were demonstrated, many publications based on gene expression profiling in hematological malignancies have appeared. Among these, several hundred have focused on AML. Some have shown that certain genetic alterations correspond with specific gene expression signatures. Gene expression profiles have also been correlated with prognosis and treatment outcomes (Handschuh *et al*., 2018).

2.2.4. CHRONIC MYELOID LEUKEMIA (CML)

Chronic myeloid leukemia is a myeloproliferative disorder derived from hematopoietic stem cell transformation and characterized by heterogeneous biological and clinical features (Albano *et al*., 2013). CML is a hematopoietic stem cell disease with distinct biological and clinical features. CML usually presents in chronic phase, in which the clonal expansion of

mature myeloid cells leads to an elevated white blood cell count. Without curative intervention, chronic-phase myeloid leukemia will invariably transform through a phase of "acceleration," often heralded by the appearance of increased immature myeloid cells in the bone marrow and peripheral blood, as well as new cytogenetic changes in addition to the Philadelphia chromosome (Ph). Progression then proceeds to blast crisis, with immature blast cells overwhelming the production of normal hematopoetic elements (Radich *et al.*, 2016).

2.3. PATHWAY ANALYSIS OF LEUKEMIA

The pathway of Leukemia is initiated and maintained by a small number of self-renewing cells called leukemia stem cells (LSCs), which share properties with hematopoietic stem cells (HSCs), the self-renewing cells that produce healthy blood cells (A Pathway to Leukemia 2010). The pathways are retrieved from KEGG (Kyoto Encyclopedia of Genes and Genomes - GenomeNet). KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction, reaction and relation networks for Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes, Organismal Systems, Human Diseases, Drug Development. KEGG PATHWAY is a reference database for Pathway Mapping (https://www.genome.jp/kegg/pathway.html). Among the several types of Leukemia, AML and CML pathway is described as follows.

2.3.1. PATHWAY OF AML

AML is a disease that is characterized by uncontrolled proliferation of clonal neoplastic cells and accumulation in the bone marrow of blasts with an impaired differentiation program. AML accounts for approximately 80% of all adult leukemias and remains the most common cause of leukemia death. Two major types of genetic events have been described that are crucial for leukemic transformation. A proposed necessary first event is disordered cell growth and upregulation of cell survival genes. The most common of these activating events were observed in the RTK Flt3, N-Ras and K-Ras, in Kit, and sporadically in other RTKs. Alterations in myeloid transcription factors governing hematopoietic differentiation provide second necessary event for leukemogenesis. Transcription factor fusion proteins such as AML-ETO, PML-RAR alpha or PLZF-RAR alpha block myeloid cell differentiation by repressing target genes. In other

cases, the transcription factors themselves are mutated (https://www.kegg.jp/dbget-bin/www_bget?hsa05221).

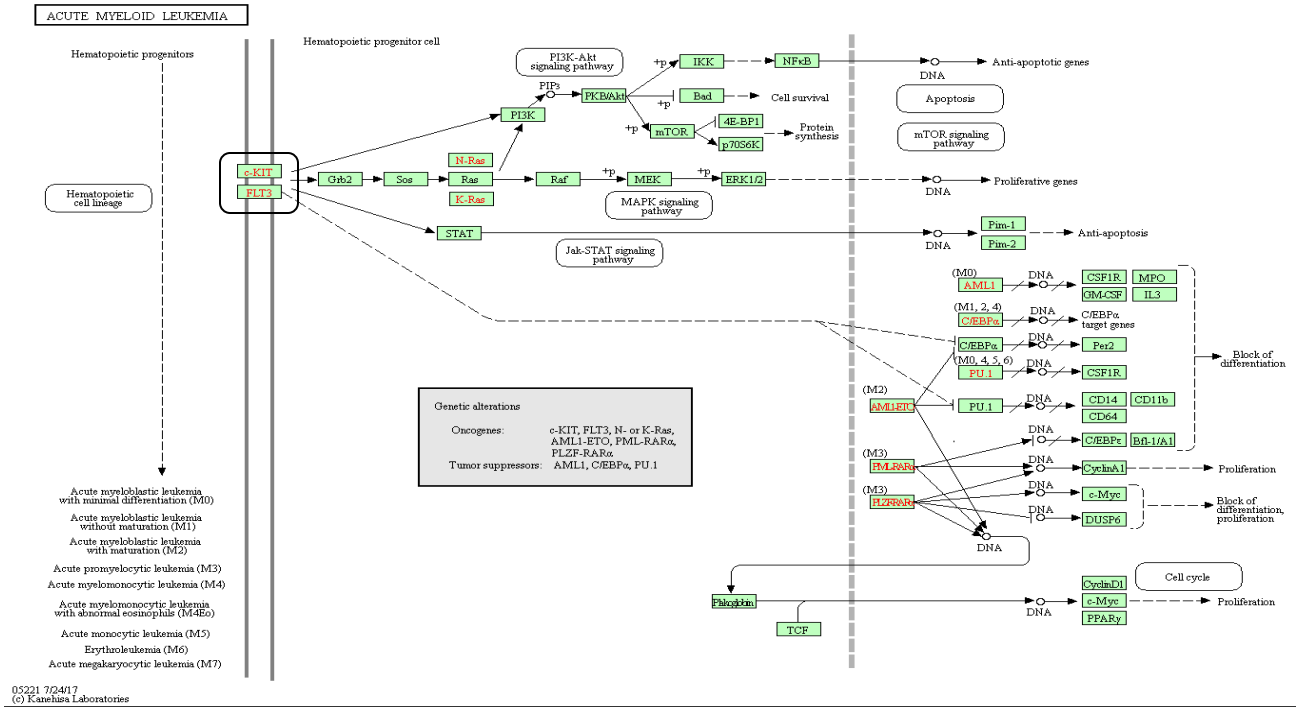


Figure 1. Pathway of AML (https://www.kegg.jp/dbget-bin/www_bget?hsa05221**)**

2.3.2. PATHWAY OF CML

Chronic myeloid leukemia (CML) is a clonal myeloproliferative disorder of a pluripotent stem cell. The natural history of CML has a triphasic clinical course comprising of an initial chronic phase (CP), which is characterized by expansion of functionally normal myeloid cells, followed by an accelerated phase (AP) and finally a more aggressive blast phase (BP), with loss of terminal differentiation capacity. On the cellular level, CML is associated with a specific chromosome abnormality, the t(9; 22) reciprocal translocation that forms the Philadelphia (Ph) chromosome. The Ph chromosome is the result of a molecular rearrangement between the c-ABL proto-oncogene on chromosome 9 and the BCR (breakpoint cluster region) gene on chromosome 22. The BCR/ABL fusion gene encodes p210 BCR/ABL, an oncoprotein, which, unlike the normal p145 c-Abl, has constitutive tyrosine kinase activity and is predominantly localized in the cytoplasm. While fusion of c-ABL and BCR is believed to be the primary cause of the chronic phase of CML, progression to blast crisis requires other molecular changes. Common secondary

abnormalities include mutations in TP53, RB, and p16/INK4A, or overexpression of genes such as EVI1. Additional chromosome translocations are also observed, such as t(3;21)(q26;q22), which generates AML1-EVI1 (https://www.kegg.jp/dbget-bin/www_bget?hsa05221).

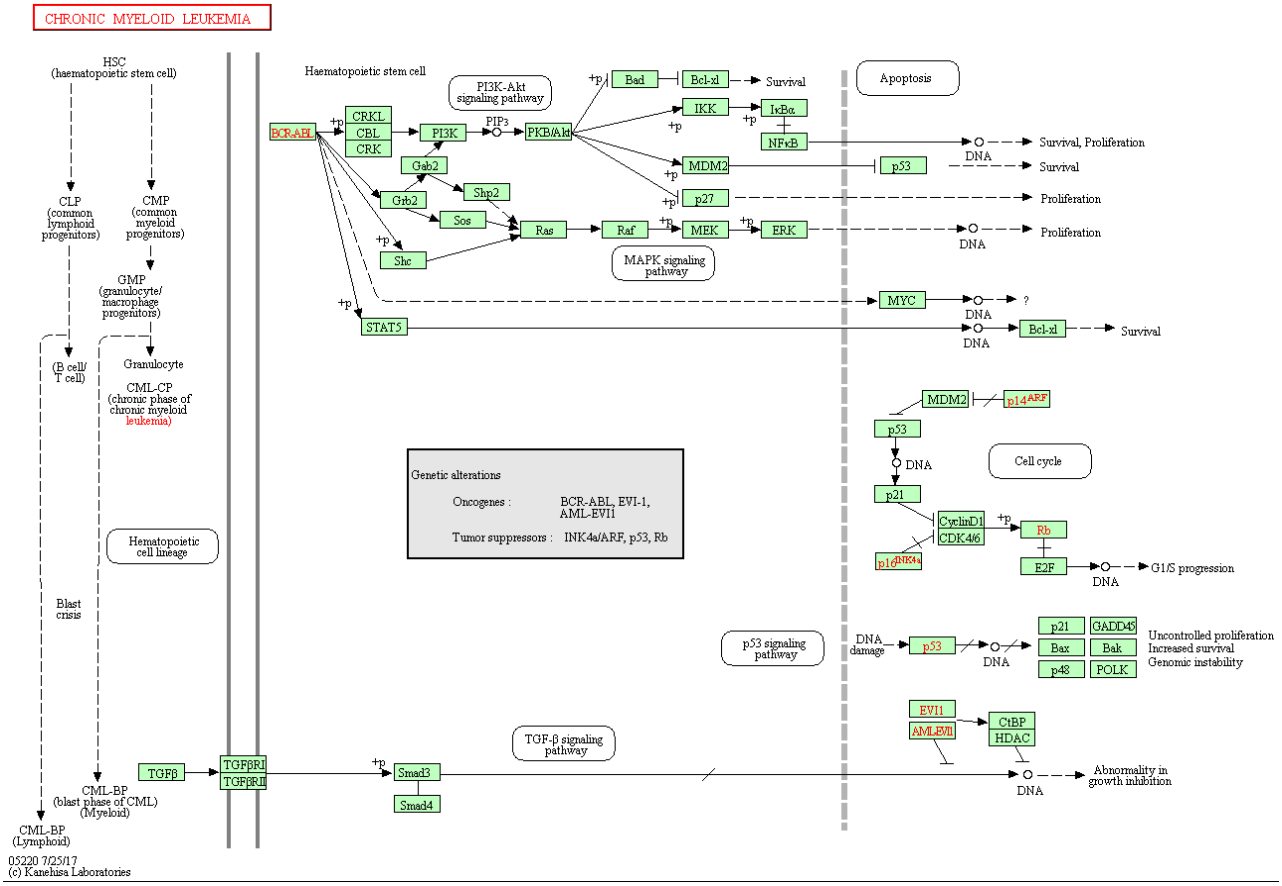


Figure 2. Pathway of CML

2.4. COMPUTATIONAL BACKGROUND

Gene Regulation provides a summary of tools available for a variety of problems in diseases, not just gene regulation, making it a good reference text for computational biologists.

2.4.1. AFFYMETRIX MICROARRAYS

The microarray experiments carried out to get the Affymetrix GeneChip system. Affymetrix probes were using publicly available information. The sequences of the probe sets were selected from GenBank, dbEST, and RefSeq. The sequence clusters were created from the UniGene database (Build 133, April 20, 2001) and then refined by analysis. The comparison

process done with a number of other publicly available databases, including the Washington University EST trace repository and the University of California, Santa Cruz Golden-Path human genome database (April 2001 release). Sequences from these databases were collected and clustered into groups of similar sequences.

The probes are manufactured on the chip using photolithography (a process of using light to control the manufacture of multiple layers of material), which is adapted from the computer chip industry. Each GeneChip contains approximately 1,000,000 features. Each probe is spotted as a pair, one being a perfect match (PM), and the other with a mismatch (MM) at the centre. These probe pairs allow the quantitation and subtraction of signals caused by non-specific cross-hybridization. The differences in hybridization signals between the partners, as well as their intensity ratios, serve as indicators of specific target abundance. Each gene or transcript is described on the GeneChip by eleven probe pairs. The probe sets are given different suffixes to describe their uniqueness and/ or their ability to bind different genes or splice variants (Schinke-Braun *et al*., 2017).

- ✓ "_at" describes probes set that are unique to one gene
- ✓ "_a_at" describes probe sets that recognise multiple transcripts from the same gene
- ✓ "_s_at" describes probe sets with common probes among multiple transcripts from separate genes. The _s_at probe sets can represent shorter forms of alternatively polyadenylated transcripts, common regions in the 3‟ ends of multiple alternative splice forms, or highly similar transcripts. Approximately 90% of the _s_at probe sets represent splice variants. Some transcripts will also be represented by unique _at probe sets.
- ✓ "_x_at" designates probe sets where it was not possible to select either a unique probe set or a probe set with identical probes among multiple transcripts. Rules for cross-hybridisation are dropped in order to design the _x_at probe sets. These probe sets share some probes identically with two or more sequences and therefore, these probe sets may cross-hybridise in an unpredictable manner.

A sample should be registered and an experiment outlined in GCOS (GeneChip Operating Software) before processing a probe array in the fluidics station or scanning. Once the array is scanned, an image file is created called a ".dat" file. The software then computes cell

intensity data (".cel" file) from the image file. It contains a single intensity value for every probe cell delineated by the grid (calculated by the Cell Analysis algorithm). The amount of light emitted at 570nm from stained chip is proportional to the amount of labelled RNA bound to each probe. Each spot correspond to individual probe (either perfect match or mismatch). The probes for each gene are distributed randomly across the chip to nullify any region specific bias. Following this, data analysis algorithms combine the probes to the respective intensity of individual transcripts (Schinke-Braun *et al*., 2017).

2.4.2. MICROARRAY DATA

Microarray technology is used in a wide variety of settings for detecting differential gene expression. Many microarray studies are designed to find genes related to with different phenotypes, for example, the comparison of cancer tumors and normal cells. In some multifactor experiments, genetic networks are perturbed with various treatments to understand the effects of those treatments and their interactions with each other in the dynamic cellular network. For even the best experiments, investigators must consider several issues for appropriate gene selection (Bumgarner, R., 2013).

DNA microarray is a new technique that can analyze genome and characteristic map of gene expression. A variety of DNA microarray and DNA chip devices and systems have currently been developed and commercialized. DNA microarray analysis includes an oligonucleotide chip, cDNA chip, and genomic chip, and is divided into the following two modes: one is to fix the target DNA on the support, which is suitable for the analysis of a large number of different target DNAs, and another involves fixing a large number of probes on the support material, which is suitable for the analysis of various probe sequences of the identical target DNA. There are various platforms available (http://arrayconsortium.tgen. org/np2/home.do;http://www1.amershambiosciences.com/APTR1X/upp01077.nsf/Contentand/c odelink_bioarray_system;http://www.affymetrix.com;http://www.illumina.com/prod_expression. htm; and http://www.nimblegen.com; http://www.xeotron.com).

2.4.3. PATHWAY/ ENRICHMENT ANALYSIS

Comprehensive and insightful characterization of gene sets altered in specific conditions (AGS) is a challenging task. One of the most common approaches is to access the functional associations between a gene set of interest such as differentially expressed genes and known gene sets representing biological processes (e.g. GO terms) or pathways, generally termed as FGS, i.e. lists of genes that were previously assigned a common biological annotation. To identify and rank such associations, a wide range of enrichment analysis tools have been developed in recent years. The term enrichment analysis refers to examination of the list of genes to determine if they are over-represented among any set of certain processes or pathways members. Various enrichment analysis tools such as GSEA, DAVID, GoToolbox, and FATIGO etc address various challenges of functionally analyzing large gene lists. All these methods systematically evaluate the relationships between AGS and FGS, then statistically highlights the most enriched (over/under represented) biological annotations out of thousands of linked terms and contents. The GO and KEGG databases do not encompass all functionally coherent groups and if the information about the differential expressed genes is not present in such databases, then it results in poor overlap between two sets and the analysis gives false negatives (García-Campos *et al*., 2015).

2.5. MICROARRAY DATA ANALYSIS

Microarrays are one of the successful techniques in the field of molecular biology that allow us to monitor the expression levels of even ten thousand of genes simultaneously. Arrays have been applied to studies in gene expression, genetic mapping, and discrimination of SNP, determining transcription factor activity and toxicity, pathogen identification and many other applications (Selvaraj and Natarajan, 2015).

Microarray technologies are in the forefront of managing huge amount of genomic data over several period and have also been evolved. Over the, several variants have been developed and sophisticated measurements have been improved to get valuable insights about the available data. Also much more choices in advances for handling such data has been implemented and improving with accessibility, quality and interpretation. Because the need for augmentation and integration of microarray data is also rapidly increasing due to the churning data (Ruskin, 2016).

Over the last two years, the gene expression database contents have grown to rapid rate. Although high throughput sequencing (HTS)-based experiments account only for 6% of the entire content of the database, the proportion of new HTS submissions has been grown exponentially over the past few year from 2% in 2009 to 6% in ,2010, 7% in 2011 and 15% in 2012. By considering the application of HTS data, 50% of the experiments used RNA-seq only, 32% CHIP-seq data and the remaining experiments uses DNA-seq for genotyping or copy number variations and others (Rustici *et al*, 2013).

Gene Expression Omnibus (GEO) databases has made a priority to support the microarray community by switching to next generation sequence technologies. GEO accepts and stores sequence data for studies that examine gene expression (RNA-seq), gene regulation and epigenomics like CHIP-seq, methyl-seq or other studies including DNase hypersensitivity. GEO hosts processed data files together with sample and study metadata along with the raw data files containing the original sequence reads linked with National Centre for Biotechnology Information (NCBI). To date, GEO has loaded >44 terabases of read data with furthermore several thousand processed data files being incorporated into NCBI's epigenomics database which are curated and available for viewing by genome browsers (Barrett *et al*., 2013).

A large number of experiments in Genome wide expression microarray measurements have been performed over the past few decades, however a very little have been successful in fully identifying the functionally important genes in the pathogenesis of diabetes. Because of the large number of genes often detected as significant in the microarray experiment, it is hard to subset the optimal candidate genes from individual studies. More recently, investigators have applied microarrays to genetics by considering gene expression levels for quantitative to the identification of genes and targets for diabetes (Kodama, 2014).

2.6. IN *SILICO* METHODS FOR THE PREDICTION OF PPI

The yeast two-hybrid (Y2H) system and *in vitro* and *in vivo* approaches resulted in large-scale development of helpful tools for the detection of protein-protein interactions (PPIs) between specific proteins that may occur in different combinations. However, the information of the data generated through these approaches might not be reliable due to non-availability of potential PPIs. In order to know the whole context of potential interactions, it is better to develop

approaches that predict the possible interactions between proteins. A variety of *in silico* ways developed to support the interactions that are detected by experimental approach. The computational ways for *in silico* prediction include sequence-based approaches, structure-based approaches, chromosome proximity, gene fusion, *in silico* 2 hybrid, mirror tree, phylogenetic tree, gene ontology, and gene expression-based approaches. The list of all webservers of *in silico* methods was given in (Table 2).

Table 2.
The list of web servers with their references

WEB SERVER	FUNCTION	REFERENCE
Struct2Net	The Struct2Net server makes structure-based computational predictions of protein-protein interactions (PPIs)	http://groups.csail.mit.edu/cb/struct2net/webserver/
Coev2Net	Coev2Net is a general framework to predict, assess, and boost confidence in individual interactions inferred from a high-throughput experiment	http://groups.csail.mit.edu/cb/coev2net/
PRISM PROTOCOL	PRISM PROTOCOL is a collection of programs that can be used to predict protein-protein interactions using protein interfaces	http://prism.ccbb.ku.edu.tr/prism protocol/
InterPreTS	InterPreTS uses tertiary structure to predict interactions	http://www.russell.embl.de/interprets
PrePPI	PrePPI predicts protein interactions using both structural and nonstructural information	http://bhapp.c2b2.columbia.edu/PrePPI/
iWARP	iWARP is a threading-based method to predict protein interaction from protein sequences	http://groups.csail.mit.edu/cb/iwrap/
PoiNet	PoiNet provides PPI filtering and network topology from different databases	http://poinet.bioinformatics.tw/
PreSPI	PreSPI predicts protein interactions using a combination of domains	http://code.google.com/p/prespi/

WEB SERVER	FUNCTION	REFERENCE
PIPE2	PIPE2 queries the protein interactions between two proteins based on specificity and sensitivity	http://cgmlab.carleton.ca/PIPE2
HomoMINT	HomoMINT predicts interaction in human based on ortholog information in model organisms	http://mint.bio.uniroma2.it/HomoMINT
SPPS	SPPS searches protein partners of a source protein in other species	http://mdl.shsmu.edu.cn/SPPS/
OrthoMCL-DB	OrthoMCL-DB is a graph-clustering algorithm designed to identify homologous proteins based on sequence similarity	http://orthomcl.org/orthomcl/
P-POD	P-POD provides an easy way to find and visualize the orthologs to a query sequence in the eukaryotes	http://ppod.princeton.edu/
COG	COG shows phylogenetic classification of proteins encoded in genomes	http://www.ncbi.nlm.nih.gov/COG/
BLASTO	BLASTO performs BLAST based on ortholog group data	http://oxytricha.princeton.edu/BlastO/
PHOG	PHOG web server identifies orthologs based on precomputed phylogenetic trees	http://phylogenomics.berkeley.edu/phog/
G-NEST	G-NEST is a gene neighborhood scoring tool to identify co-conserved, coexpressed genes	https://github.com/dglemay/G-NEST
InPrePPI	InPrePPI predicts protein interactions in prokaryotes based on genomic context	http://inpreppi.biosino.org/InPrePPI/index.jsp
STRING	STRING database includes protein interactions containing both physical and functional associations	http://string.embl.de
MirrorTree	The MirrorTree allows graphical and interactive study of the coevolution of two protein families and asseses their interactions in a taxonomic context	http://csbg.cnb.csic.es/mtserver/

2.6.1. PROTEIN INTERACTION DATABASES

The massive quantity of experimental PPI data being generated on steady basis has led to the construction of computer readable biological databases in order to organize and to process this data. For example, the biomolecular interaction network database (BIND) is made on an specification system that permits an elaborate description of the way during which the PPI information was derived by experimentation, often including links directly to the concluding evidence from the literature. The database of interacting proteins (DIP) is another database of by experimentation determined by protein protein binary interactions. The biological general repository for interaction datasets (BioGRID) is a database that contains protein and genetic interactions among thirteen different species. Interactions are regularly added through exhaustive curation of the primary literature to the databases. Interaction data is extracted from other databases including BIND and MIPS (Munich Information Center for protein sequences), as well as directly from large-scale experiments. HitPredict is a resource of high confidence protein-protein interactions from which we can get the total number of interactions in a species for a protein and can view all the interactions with confidence scores (Prasad *et al*., 2017).

The Molecular Interaction (MINT) database is another database of experimentally derived PPI data extracted from the literature, with the added element of providing the weight of evidence for each interaction. The Human Protein Interaction Database (HPID) was designed to provide human protein interaction data precomputed from existing structural and experimental data. The information Hyperlinked over Proteins (iHOP) database used to be searched to spot previously reported interactions in PubMed for a protein of interest. IntAct provides an open source database and toolkit for the storage, presentation, and analysis of protein interactions. The web interface provides each textual and graphical representations of protein interactions and allows exploring interaction networks within the context of the GO annotations of the interacting proteins. Recently, the integration has been done and can be explored in the web server called APID (Agile Protein Interaction Data Analyzer) which is an interactive bioinformatics' web tool developed to allow exploration and analysis of presently notable information about protein-protein interactions integrated and unified in an exceedingly common and comparative platform. The Protein Interaction Network Analysis (PINA2.0) platform is a comprehensive web resource, which includes a database of unified protein-protein interaction data integrated from six

manually curated public databases and a set of built-in tools for network construction, filtering, analysis, and visualization. The databases and number of interactions are tabled in (Table 3).

Table 3.
Protein interaction databases

S.No	Database name	Source link
1	BioGrid	http://thebiogrid.org/
2	DIP	http://dip.doe-mbi.ucla.edu/dip/Main.cgi
3	HitPredict	http://hintdb.hgc.jp/htp/
4	MINT	http://mint.bio.uniroma2.it/mint/
5	IntAct	http://www.ebi.ac.uk/intact/
6	APID	http://bioinfow.dep.usal.es/apid/index.htm
7	BIND	http://bind.ca/
8	PINA2.0	http://cbg.garvan.unsw.edu.au/pina/

2.7. STUDY ON DIFFERENTIALLY EXPRESSED GENES

Next generation sequencing (NGS) technologies greatly promote research in genome-wide mRNA expression data. Compared with microarray technologies, NGS provides higher resolution data information and additional precise measurement of levels of transcripts for studying gene expression. single-cell data analysis is important in cancer studies, as differential gene expression analysis between different cells can help to uncover driver genes. Tools developed for differential gene expression analysis on RNAseq data, such as DESeq and edgeR , can be applied to single-cell data. The lack of agreement in finding DE genes by these tools and their limitations in detecting true DE genes and biologically relevant gene sets indicate the need for developing more precise methods for differential expression analysis of scRNAseq data (Wang *et al*., 2019).

Alzheimer's disease (AD) is the most common form of dementia in older adults that damages the brain and results in impaired memory, thinking and behavior. The identification of differentially expressed genes and related pathways among affected brain regions can provide more information on the mechanisms of AD. In the past decade, several studies have reported many genes that are associated with AD. Studies were done to apply a novel combinatorial optimization based meta-analysis approach to identify differentially expressed genes that are associated to AD across brain regions. There are 23 up and down regulated probes which are

differentially expressed between control and AD. Genes related with AD that are consistent with existing studies, and new candidate genes not previously related with AD. This study confirms the up-regulation of *INFAR2* and *PTMA* along with the down regulation of *GPHN, RAB2A, PSMD14* and *FGF*. Novel genes *PSMB2, WNK1, RPL15, SEMA4C, RWDD2A* and *LARGE* are found to be differentially expressed across all brain regions (Puthiyedth *et al*., 2016).

The protein content in Alfalfa (*Medicago sativa*) leaves is the critical factor in determining the quality of Alfalfa. Thus far, the understanding of the molecular mechanism of Alfalfa defoliation traits remains unclear. The transcriptome database created by RNA-Seq is employed to spot critical genes related to defoliation traits. In this study, the transcriptomes of the Zhungeer variety (with easy leaf abscission) and WL319HQ variety (without easy leaf abscission) are sequenced. Among the familiar 66,734 unigenes, 706 differentially expressed genes (DEGs) upregulated, and 392 unigenes downregulated within the Zhungeer vs WL319HQ leaf. KEGG pathway annotations determined that 8,414 unigenes were annotated to 87 pathways and contained 281 DEGs. Six DEGs belonging to the "Carotenoid biosynthesis", "Plant hormone signal transduction" and "Circadian rhythm-plant" pathways involved in defoliation traits were identified and validated by RT-qPCR analyses. This study used RNA-Seq to find genes related to defoliation traits between two alfalfa varieties. This transcriptome data dramatically enriches alfalfa functional genomic studies. In addition, these data provide theoretical guidance for field production practice and genetic breeding, as well as references for future study of defoliation traits in alfalfa (Cheng *et al*., 2018).

The plant genes involved in cellular signaling and metabolism haven't been totally known, whereas the function(s) of the many of these that have are thus far incompletely characterised. Gene expression analysis permits the identification of genes and therefore the study of their relationship with cellular processes. Microarray hybridization technology permits the study of an oversized number of genes from completely different species. A study of wheat cultivars infected with *F. graminearum* was based on a cDNA microarray library of wheat expressed sequence tags (ESTs) obtained using suppressive subtractive hybridization and found 25 differentially expressed wheat UniGenes (Casassola *et al*., 2013).

Microarray technology, that observes thousands of gene expressions directly, is one in all the favored topics in recent decades. When it involves to the analysis of microarray data to find

differentially expressed (DE) genes, several strategies are projected and changed for improvement. A researcher may have to conduct an associate experiment to discover differentially expressed genes between two experimental conditions. For clarification purpose this can be between healthy patients and patients who have a condition of interest like cancer. Microarray analysis can permit the researcher to seek out that genes are expressed differently otherwise between these two groups of patients. Figuring out the foremost efficient factor to find differentially expressed genes under particular data settings can help master the data analysis step in microarray research. An analysis on a real microarray datasets was also performed to evaluate how the methods and the proposed modification would perform in a real situation (Andrew *et al*., 2015).

GEO database (Gene Expression Omnibus) is used to collect datasets from gene expression studies where a tumor tissue (belonging to bladder, colon, kidney or thyroid cancers) was compared against a healthy tissue reference. The initial datasets were derived for six cancer types including bladder (GSE7476), colon (GSE4107), kidney (GSE7023), thyroid (GSE3678), breast (GSE6883) and prostate (GSE3325) cancers. Comparative analysis of PPIs from different cancer types revealed a number of common functions or processes across all these cancers, as well as those that are specific to partial cancers or only to a particular cancer type. The methodology used in this study derives the common functions of protein pairs in PPIs from different tumor tissues and uses this formation as the basis for cross comparison of similarities and differences among various cancer types. The network was found only in colon cancer, which contains two nucleoside diphosphate kinases NDK and Nm23. Out of them, the mitochondrial precursor protein (Nm23) was well studied as a metastasis-associated gene in colon cancer. In addition, comparison of PPI networks across four cancers and identification of experimentally known protein clusters that are common or specific to different cancer types have demonstrated the efficacy of our method in studying similar interaction networks in other disease systems. The similarities and differences observed in the biological processes and molecular functions of PPIs from various cancer types will provide the basis for focused experimental investigations in cancer therapeutics and drug discovery studies. (Guda *et al*., 2019).

Methodology of the *in silico* tools adopted in the present study are given/elaborated in the next chapter.

3

METHODOLOGY

A study on expression of genes in various types of leukemia using in silico tools - a therapeutic approach

3. METHODOLOGY

The experimental design adopted in the study for the analysis of microarray expression profile data is briefed in this section. Microarray techniques provide a platform where one can measure the expression levels of thousands of genes in hundreds of different conditions whereas using traditional methods in molecular biology can only report the expression levels of single genes. Based on the aim of the study, we have chosen four different sets of leukemia data. In this study, we conducted PPI networks of DEGs. The following databases and software tools were used to analyze and compare the microarray datasets for the differential expression of genes.

3.1. DATA REPOSITORIES

Databases are repositories of information, generally store on computers. They contain an organized series of records that can be searched through and displayed in the computer screen, downloaded or emailed to a specific address. A microarray database can be a repository containing microarray gene expression data. Microarray data repositories are large collections of data that are implemented from different array experiments to serve the research community. There are several repositories maintaining huge amount of microarray expression data providing open source access to retrieve and evaluate gene expression. A peer reviewed, public repository that adheres to educational or trade standards and is intended to be utilized by several analytcal applications. A good example of this is the Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) from NCBI or ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) from EBI.

3.1.1. GENE EXPRESSION OMNIBUS (GEO)

GEO is a public functional genomics data information repository supporting MIAME-compliant data information submissions. GEO is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community. The Gene Expression omnibus maintained by NCBI-National Center for Biotechnology Information (http://www.ncbi.nih.gov/geo/) is one of the largest and most widely used database.

Gene expression omnibus is a tool in NCBI where the data was taken. Each sample has its own accession number. The other format is not accepted by the tool. Gene expression omnibus consist of samples in tab delimited form which either opens in note pad or Microsoft excel.

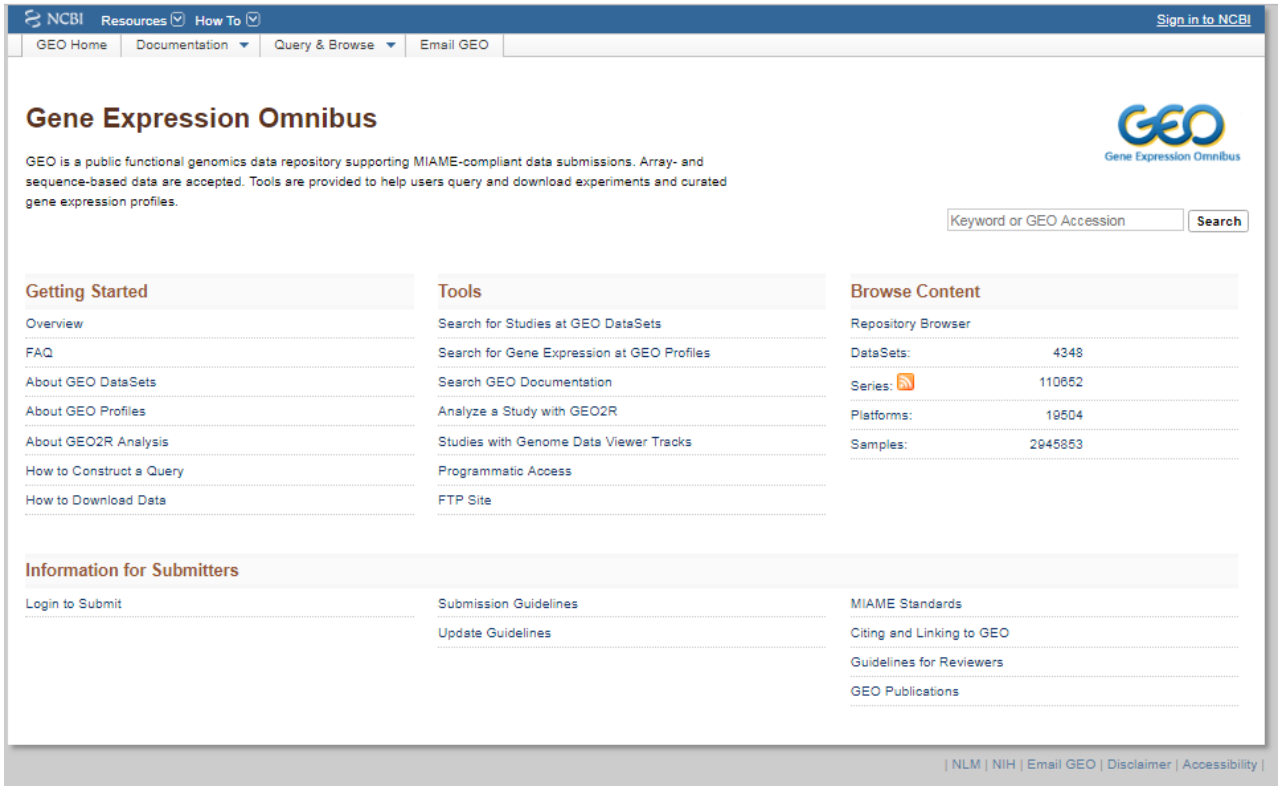


Figure 3. Gene expression omnibus homepage

3.1.2. GEO DATASETS

The GEO DataSets database stores original submitter-supplied records (Series, Samples and Platforms) likewise as curated DataSets. The GEO DataSets database information will be searched using several attributes together with keywords, organism, DataSet type and authors. DataSets records contain additional resources, including cluster tools and differential expression queries (http://www.ncbi.nlm.nih.gov/geo/info/datasets.html).

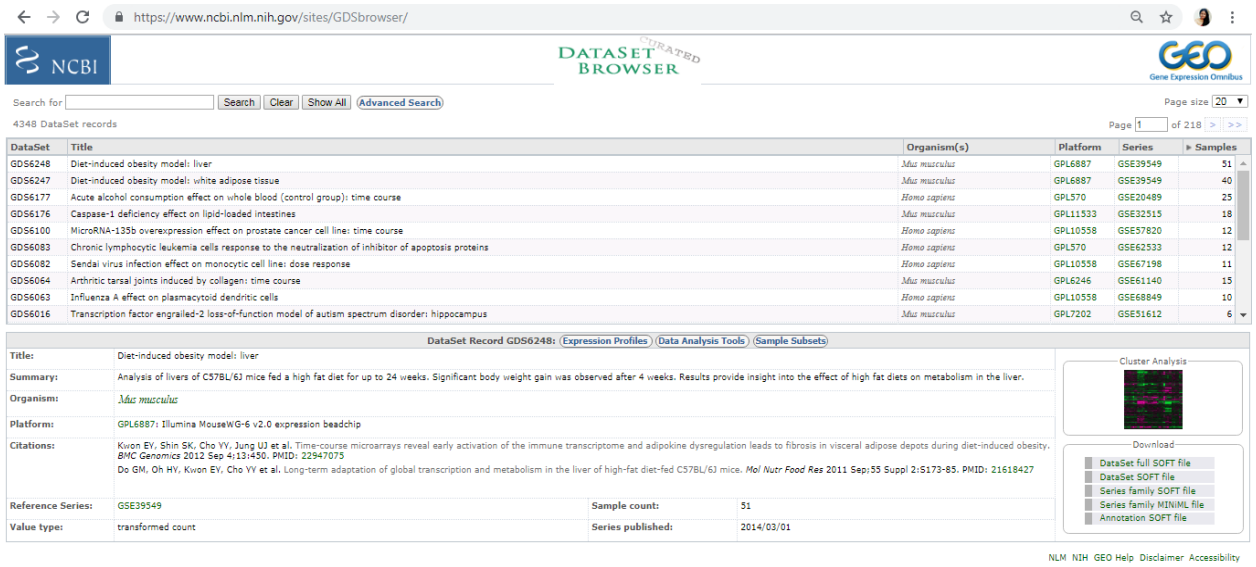


Figure 4. GEO DataSets webpage

3.1.3. GEO PROFILES

The GEO Profiles database stores gene expression profiles derived from curated GEO DataSets. Each Profile is presented as a chart that displays the expression level of one gene across all Samples within a DataSet. Experimental context is provided in the bars along the bottom of the charts making it possible to see at a glance whether a gene is differentially expressed across different experimental conditions. Profiles have various types of links including internal links that connect genes that exhibit similar behaviour, and external links to relevant records in other NCBI databases (https://www.ncbi.nlm.nih.gov/geo/info/profiles.html).

3.1.4. FUNCTIONAL (GENE ONTOLOGY) AND PROTEIN-PROTEIN INTERACTION (PPI) NETWORK

Protein–protein interaction networks are the networks of protein complexes formed by biochemical events and/or electrostatic forces and that serve a distinct biological function as a complex. The protein interactome describes the full repertoire of a biological system's protein–protein interactions (PPIs). For functional analysis and construction of protein-protein interactions (PPI) network, the Search Tool for the Retrieval of Interacting Genes (STRING) database is used. STRING offers integrative tools for providing besides constructing PPI networks and also functional and pathway enrichment analysis.

STRING (https://string-db.org/) is a database of known and predicted protein-protein interactions. The interactions include direct (physical) and indirect (functional) associations; they stem from computational prediction, from knowledge transfer between organisms, and from interactions aggregated from other (primary) databases. Interactions in STRING are derived from five main sources such as Genomic Context Predictions, High-throughput Lab Experiments, (Conserved) Co-Expression, Automated Text mining and Previous Knowledge in Databases. String database was employed in https://string-db.org. Visualization of all networks together was done in Cytoscape version 3.7.1 (http://www.cytoscape.org). The differentially expressed genes were further subjected to gene ontology studies using the online source program GENECODIS (www.genecodis.cnb.csic.es).

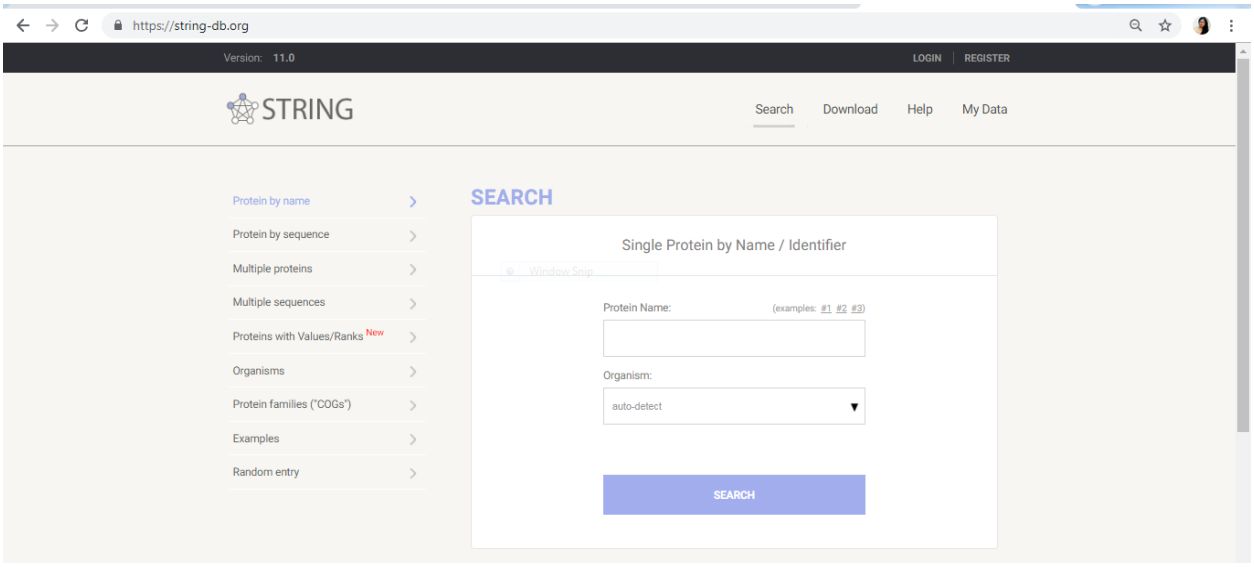


Figure 5. STRING webpage

3.2. DATA PROCESSING METHODOLOGY

3.2.1. RETRIEVAL OF DATASETS

A total of four samples with the common array platform (Affymetrix) with the annotation type HG-U133_Plus_2 array was selected and downloaded from the NCBI Gene Expression Omnibus (GEO). The following shows the sample details about the datasets collected for the present study.

Source File I:

GO Accession : GSM331660, GSM331661, GSM331662

Source name : Non-Leukemia and Healthy Bone Marrow

Source File II:

GO Accession : GSM329843, GSM329856, GSM329903

Source name : T-ALL

Source File III:

GO Accession : GSM330417, GSM330418, GSM330419

Source name : AML with t(8;21)

Source File IV:

GO Accession : GSM330930, GSM330931, GSM330932

Source name : CLL

Source File V:

GO Accession : GSM331392, GSM331394, GSM331398

Source name : CML

Each of the source file is having three different samples. In this study three control samples and three disease samples were downloaded. These files were downloaded as .txt format.

3.2.2. META-ANALYSIS OF MICROARRAY DATASETS

A total of 15 samples from the selected dataset were pooled into a single file containing all the .cel files of raw data. R programming environment was used to carry out the different layers of microarray data analysis.

Due to its data handling and modeling capabilities as well as its flexibility, R is becoming the most widely used software in bioinformatics. R Programming for Bioinformatics explores the programming skills required to use this software package tool for the answer of bioinformatics and computational biology problems. The datasets were downloaded and normalized in R language (https://www.r-project.org/). Differentially expressed genes (DEGs) were identified

using limma package in R for each datasets. The control group and diseased group were compared using moderated t-test and linear model for microarray data analysis with the cut-off conditions were set to adjusted p value<0.05 and absolute value of fold change>2.

The differentially expressed genes among the groups were clustered and ordered by hierarchical clustering algorithm. The clustering was performed using "ggplot" package in R and the heat map was constructed. Using this heat map, the expressed genes were represented graphically by coloring each cell on the basis of fluorescence ratio with fold change and p-value. The volcono plot is obtained that indicates up-regulated genes in red and down-regulated genes in blue.

3.2.3. FUNCTIONAL ANALYSIS AND PPI NETWORK

The differentially expressed genes were further subjected to Gene Ontology (GO) studies using the online source program GENECODIS (www.genecodis.cnb.csic.es) and the pathway analysis was carried out using STRING database to find the significant positive and negative regulating pathways of the DEGs. The results of the bioinformatics parameters carried out in the present study are discussed in the next chapter.

4

RESULTS AND DISCUSSION

A study on expression of genes in various types of leukemia using in silico tools - a therapeutic approach

4. RESULTS AND DISCUSSION

The results obtained from the present study entitled **"A study on expression of genes in various types of leukemia using *in silico* tools - a therapeutic approach**" are discussed in this section.

Cancer is caused by the accumulation of genetic and epigenetic changes resulting from the altered sequence or expression of cancer-related genes, such as oncogenes or tumor suppressor genes, as well as genes involved in cell cycle control, apoptosis, adhesion, DNA repair, and angiogenesis. Because gene expression profiles provide a snapshot of cell functions and processes at the time of sample preparation, comprehensive combinatorial analysis of the gene expression patterns of thousands of genes in tumor cells and comparison to the expression profile obtained with healthy cells should provide insights concerning consistent changes in gene expression that are associated with tumor cellular dysfunction and any concomitant regulatory pathways. Microarray technology has been widely used in the past 3 years to investigate tumor classification, cancer progression, and chemotherapy resistance and sensitivity. Many researchers are very hopeful about the future of cancer treatments based on the specific gene changes found in cancer cells, and this remains a very active area of research.

4.1. OVERVIEW OF THE DATASETS AND SOFTWARES INCLUDED IN THE STUDY

In this study, a total of two microarray DataSets of GSE13159 and GSE13164 with more than 2000 samples together. From this Affymetrix Human Genome U133 Plus 2.0 Array [HG-U133_Plus_2] data, is grouped under conditions, in which 3 samples of Non-leukemia and healthy bone marrow, 3 samples of T-cell acute lymphoblastic leukemia (ALL), 3 samples of acute myeloid leukemia (AML), 3 samples of chronic lymphoblastic leukemia (CLL), 3 samples of acute myeloid leukemia (AML) were taken for differential expressed genes analysis. The study used the software R statistical Language and for the protein-protein interactions STRING was used.

4.2. DATA PROCESSING AND NORMALIZATION

The preprocessing of microarray data includes the normalization of the data in which the expression ratios are log transformed into a reasonable measure prior to the detection of differentially expressed genes. It is the process of eliminating variations in the dataset that allows appropriate comparison of data. The average raw intensity signals in each array samples of the dataset against their prominent intensity based on the mean quantile variations were compared before and after normalization and represented in the form of a box plot. (Table 4) depicts the differentially expressed genes among the samples analyzed in the present study.

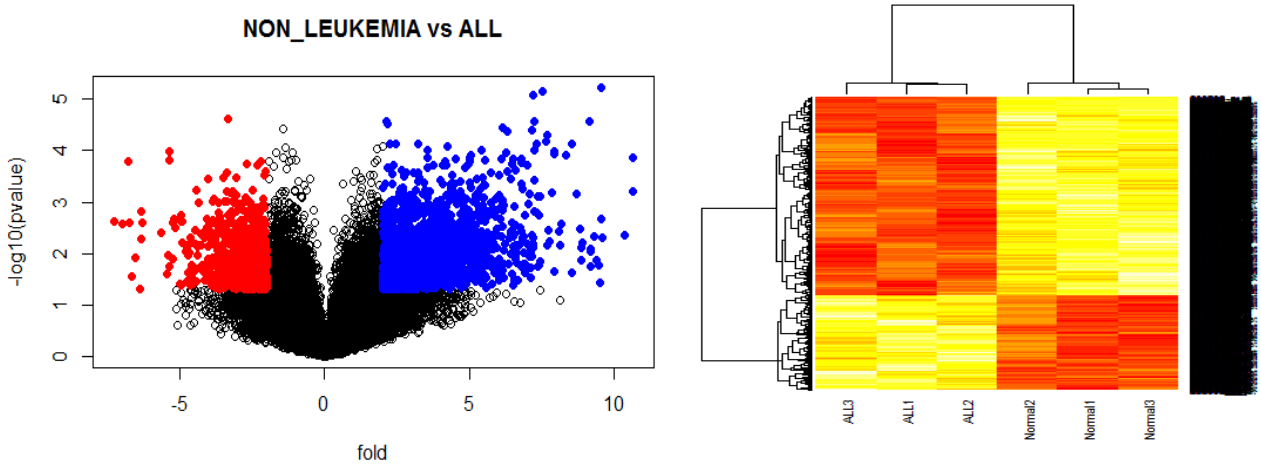
Table 4.
DEGs of microarray data samples

SAMPLES TAKEN (54675 genes)		DEGs	UP-REGULATED	DOWN-REGULATED
Non-Leukemia and Healthy Bone Marrow	ALL	10351	5251	5100
Non-Leukemia and Healthy Bone Marrow	AML	4533	1928	2605
Non-Leukemia and Healthy Bone Marrow	CLL	10975	5296	5679
Non-Leukemia and Healthy Bone Marrow	CML	7619	3477	4142

DEGs - Differentially Expressed Genes, ALL - Acute Lymphocytic Leukemia, AML - Acute Myeloid Leukemia, CLL - Chronic Lymphocytic Leukemia, CML - Chronic Myeloid Leukemia.

4.2.1. NON-LEUKEMIA AND ALL

In the analysis of Non-Leukemia and Acute Lymphocytic Leukemia samples, 10351 genes were identified differentially expressed. Among this, 5251 genes were up-regulated and 5100 genes were down-regulated. Plot of the average raw intensities of each samples before normalization and after normalization of this samples were done using R program (R version 3.5.2). In the volcano Plot shown in Figure 6A. is highlighting up-regulated genes in red and down-regulated genes in blue color. Heatmap generated against Differentially Expressed Genes (DEGs) shown in Figure 6B.

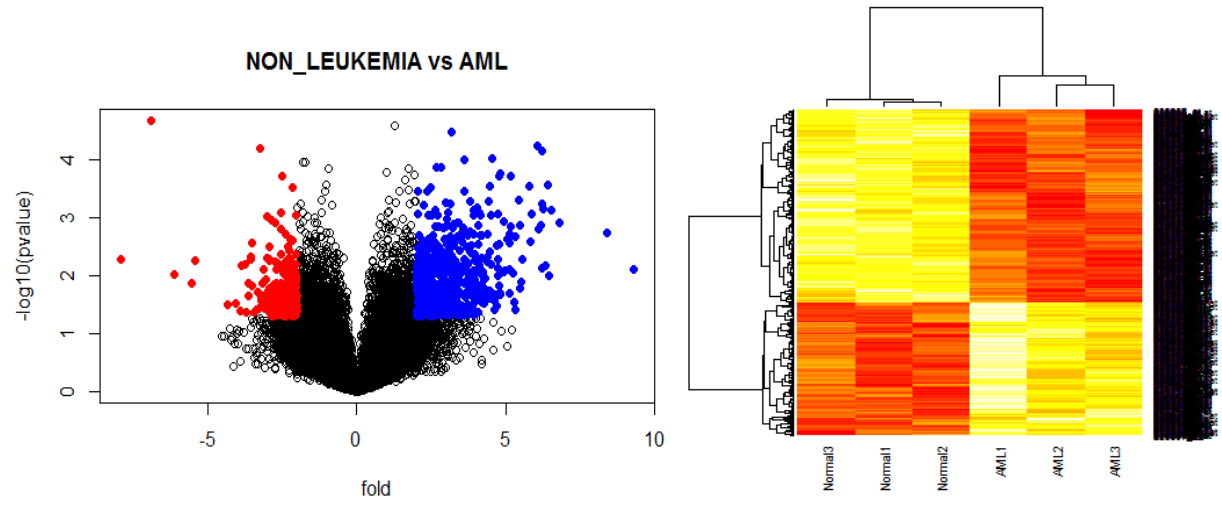


(A) **(B)**

Figure 6. (A) VOLCANO PLOT (B) Heatmap

4.2.2. NON-LEUKEMIA AND AML

When the Acute Myeloid Leukemia (AML) samples were analyzed, a total of 4533 genes were differentially expressed in comparison with the expression of Non-Leukemia samples. Among this, 1928 genes were up-regulated and 2605 genes were down-regulated. Plot of the average raw intensities of each samples before normalization and after normalization of this samples were done using R program. In the volcano Plot shown in Figure 7A. is highlighting up-regulated genes in red and down-regulated genes in blue color. Heatmap generated against Differentially Expressed Genes (DEGs) shown in Figure 7B.

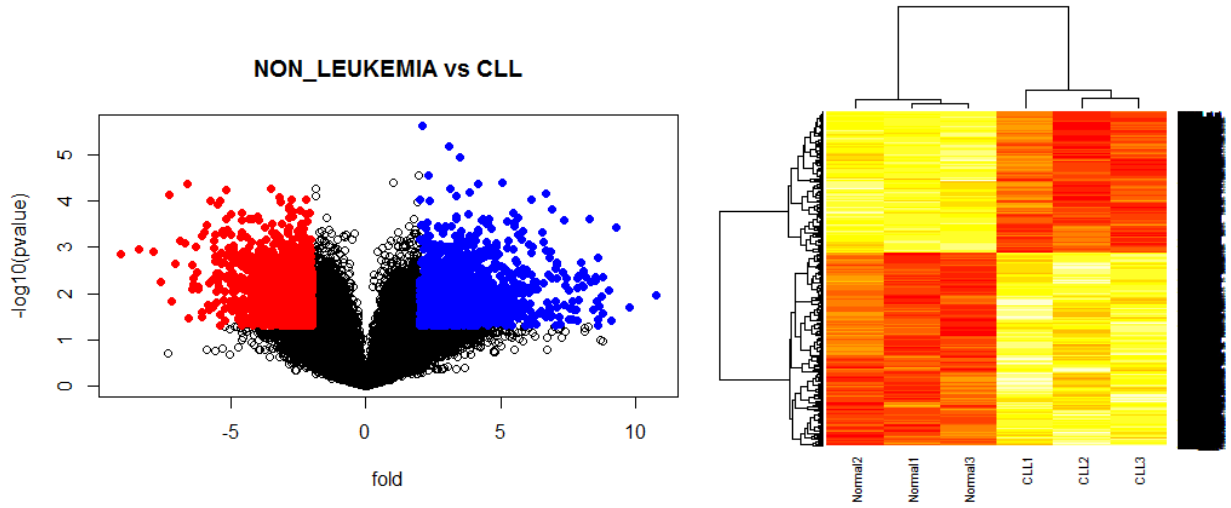


(A) **(B)**

Figure 7. (A) VOLCANO PLOT (B) Heatmap

4.2.3. NON-LEUKEMIA AND CLL

In the analysis of Non-Leukemia and Chronic Lymphocytic Leukemia samples, 10975 genes were identified as differentially expressed. Among this, 5296 genes were up-regulated and 5679 genes were down-regulated. Plot of the average raw intensities of each samples before normalization and after normalization of this samples were done using R program. In the volcano Plot shown in Figure 8A. is Highlighting up-regulated genes in red and down-regulated genes in blue. Heatmap generated against Differentially Expressed Genes (DEGs) shown in Figure 8B.

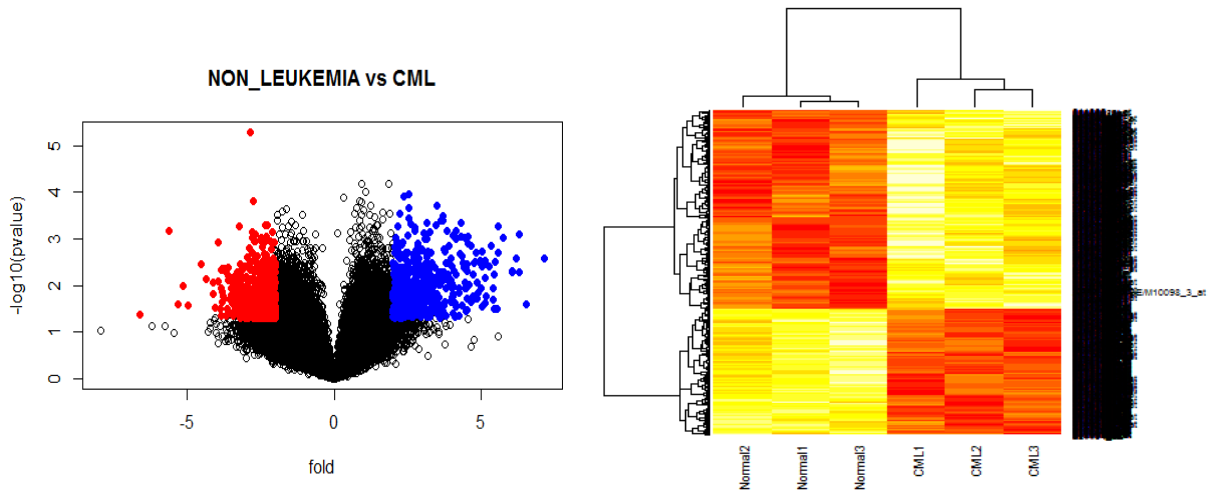


(A) **(B)**

Figure 8. (A) VOLCANO PLOT (B) Heatmap

4.2.4. NON-LEUKEMIA AND CML

In the analysis of Non-Leukemia and Chronic Myeloid Leukemia (CML) samples, 7619 genes were identified differentially expressed. Among this, 3477 genes were up-regulated and 4142 genes were down-regulated. Plot of the average raw intensities of each samples before normalization and after normalization of this samples were done using R program. In the volcano Plot shown in Figure 9A. is highlighting up-regulated genes in red and down-regulated genes in blue color. Heatmap generated against Differentially Expressed Genes (DEGs) shown in Figure 9B.



(A) **(B)**

Figure 9. (A) VOLCANO PLOT (B) Heatmap

The various types of Leukemia expression microarray datasets samples were downloaded from GEO datasets. The data were calibrated, standardized, and log2 transformed. All the dataset was screened by the limma package (corrected P-value - 0.05, logFC - 2).

From the results of DEGs, maximum number of altered gene expression was found to be in CLL and then ALL. However, Myeloid Leukemia samples showed lesser modulation in gene expression compared to Lymphocytic Leukemia.

4.2.5. INTERSECTION OF DEGs

The common genes were identified using Draw venn diagram in Bioinformatics & Evolutionary Genomics webpage of http://bioinformatics.psb.ugent.be/webtools/Venn/ is shown in Figure 10.

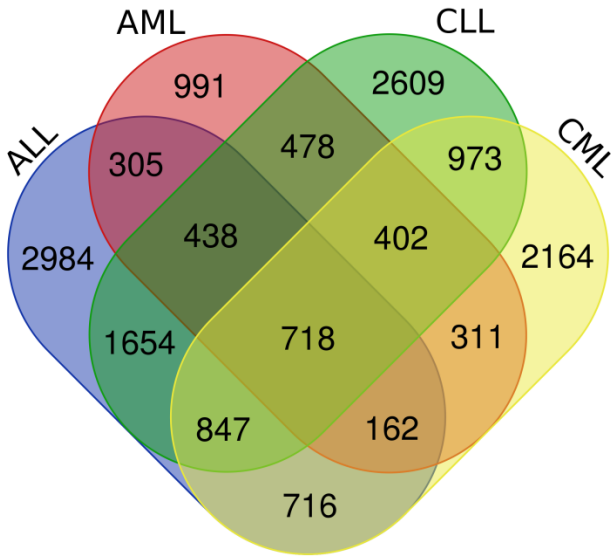


Figure 10. Venn Diagram of clustering DEGs of ALL (Non-Leukemia and healthy bone marrow vs ALL), AML (Non-Leukemia and healthy bone marrow vs AML), (Non-Leukemia and healthy bone marrow vs CLL) and (Non-Leukemia and healthy bone marrow vs CML)

In order to investigate resulted DEGs, we overlapped the genes in each comparison, to obtain the unique set of genes characteristic for each comparison given in (Table 5). By grouping the DEGs, 718 genes were commonly identified DEGs among ALL (Non-Leukemia and healthy bone marrow vs ALL), AML (Non-Leukemia and healthy bone marrow vs AML), (Non-Leukemia and healthy bone marrow vs CLL) and (Non-Leukemia and healthy bone marrow vs CML). Each cell indicates the common DEGs among the particular sample type. Thus ALL is compared with AML, CLL, CML and vice versa.

Table 5.
List of most common DEGs

SAMPLES	TOTAL DEGs	SAMPLES	TOTAL DEGs
ALL AML CLL CML	718	ALL AML	305
ALL AML CLL	438	ALL CLL	1654
ALL AML CML	162	ALL CML	716
ALL CLL CML	847	AML CLL	478
AML CLL CML	402	AML CML	311
ALL	2984	CLL CML	973
CLL	2609	AML	991
CML	2164		

A total of 718 genes were found to be differentially expressed in all the four types of leukemia. A maximum of 847 DEGs were found to be common in ALL, CLL and CML, indicating DEGs associated with AML are lesser compared to all other types tested. Also, DEGs that are restricted to only one type of leukemia is also found to be lesser in AML (991). However, nearly 2984 genes were DE in ALL.

The unique DEGs of each category is 2984 of ALL (Non-Leukemia and healthy bone marrow vs ALL), 991 of AML (Non-Leukemia and healthy bone marrow vs AML), 2609 of (Non-Leukemia and healthy bone marrow vs CLL) and 2164 of (Non-Leukemia and healthy bone marrow vs CML).

4.3. PROTEIN-PROTEIN INTERACTION (PPI) NETWORKS

To further evaluate the biological significance for the differentially expressed genes, KEGG pathway enrichment analysis was carried out using online software STRING. The PPI network was constructed on the basis of STRING database and visualized using Cytoscape software.

4.3.1. PPIs of DEGs of Non-Leukemia and Leukemia individually

The total DEGs were classified according to the samples taken. Each DEGs are analyzed with their protein-protein interaction network. The pathways of ABC transporters, Arrhythmogenic right ventricular cardiomyopathy (ARVC), Cell cycle, Cellular senescence, DNA replication, Fatty acid degradation, Hematopoietic cell lineage, Mismatch repair, p53 signaling pathway and Progesterone-mediated oocyte maturation is the common PPI networks of DEGs of ALL, AML, CLL, CML which are compared with Non-Leukemia.

4.3.2. PPIs of unique DEGs of Non-Leukemia and Leukemia

There are unique DEGs were also identified using clustering of DEGs. PPIs of KEGG pathway analysis were also carried out for these DEGs. While comparing Acute type of Leukemia DEGs, observed in AML and ALL were found to be not involved in any of the pathways. Interestingly, the DEGs of Chronic Leukemia both CLL and CML involves in

metabolic process, cellular process. According to the comparison of process of unique DEGs, Chronic Leukemia has similar process.

4.3.3. PPIs of common DEGs of Non-Leukemia and Leukemia

Also the common DEGs (718) of all the samples with Non-Leukemia was analyzed for their PPIs. Within these 718 DEGs, we further shortlisted under the condition of those which has PPIs scores more than 0.9 with other genes. In the whole network, there are 240 genes as show in Table 6. were having higher and significant interactions with other DEGs. Figure 11 shows the STRING interactions of 240 DEGs. These 240 DEGs were found to be active participants of the pathways of ABC transporters, Arrhythmogenic right ventricular cardiomyopathy (ARVC), Cell cycle, Cellular senescence, DNA replication, Fatty acid degradation, Hematopoietic cell lineage, Mismatch repair, p53 signaling pathway and Progesterone-mediated oocyte maturation.

Comparative analysis of protein protein interactions of individual DEGs and common DEGs protein protein interactions, the process Cell cycle, DNA replication, Hematopoietic cell lineage, p53 signaling pathway and Progesterone-mediated oocyte maturation. According to the comparative study of DEGs in Leukemia, the most of the DEGs are significantly involved in the pathways of Cell cycle and p53 signaling pathway.

Among the 240 DEGs, the following 20 genes BUB1, CCNA2, CCNB1, CCNB2, CCNE1, CDC20, CDC25A, CDC25C, CDC27, CDC45, CDK1, CHEK1, DP2, E2F2, MAD2L1, MCM4, ORC1, PTTG1, TFDP1, TTK were involved in cell cycle pathways with higher inteaction. about As well as CCNB1, CCNB2, CCNE1, CDK1, CHEK1, GTSE1 genes were involved in p53 signaling pathway. CD36, CD59, CSF2RA, HLA-DOA, IL1R1, ITGA2B, ITGA4, ITGB3, TFRC were involved in Hematopoietic cell lineage. In the pathway of Progesterone-mediated oocyte maturation, the genes involved were found to be AURKA, BUB1, CCNA2, CCNB1, CCNB2, CDC25A, CDC25C, CDC27, CDK1, MAD2L1, PIK3R3.

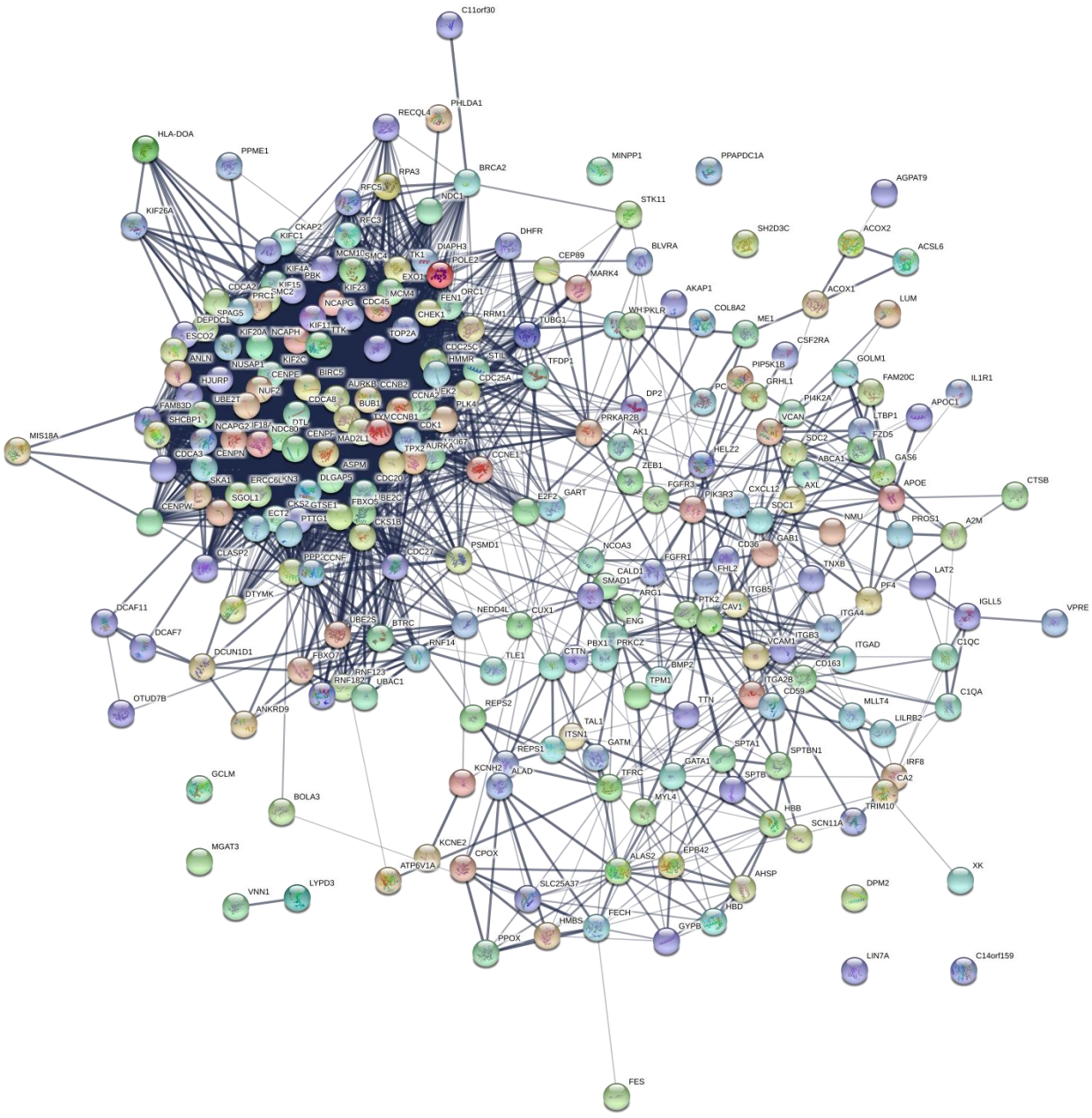


Figure 11. PPIs of selected DEGs

The PPI string analysis revealed that these 240 DEGs were common to all the samples, actively interacting with other genes which shows interaction score more than 0.9. Among these, CDK1 shows higher interactions having 14. Alterations in cell cycle pathways and retinoic acid signaling are implicated in leukemogenesis. However, little is known about the roles of cyclin-dependent kinases (CDKs) in treatment response of leukemia. Cyclin-dependent kinases (CDKs) and their associated regulatory cyclins are required for cell cycle progression and DNA

replication. It has been proposed that tumor cells with defective CDK signaling can escape the anti-mitogenic signals induced by chemotherapeutic drugs (Hedblom, 2013). The oncogenic role of FLT3 mutants has been attributed to the abnormal activation of several downstream signaling pathways, such as STAT3, STAT5, ERK1/2, and AKT. The cyclin-dependent kinase 1 (CDK1) pathway is also affected by internal tandem duplication mutations in FLT3. Inhibiting either FLT3 receptor, MEK1 kinase, or CDK1 can restore the activity of C/EBPα and induce myeloid maturation of leukemic blasts (Radomska, 2014).

Table 6.
DEGs and their PPIs

A2M	C1QA	CKS1B	FBXO5	ITGA4	MKI67	PPP2R1B	SPTA1	C14orf159
ABCA1	C1QC	CKS2	FBXO7	ITGAD	MLLT4	PRC1	SPTB	CKAP2
ACOX1	CA2	CLASP2	FECH	ITGB3	MYL4	PRKAR2B	SPTBN1	FAM83D
ACOX2	CALD1	COL8A2	FEN1	ITGB5	NCAPG	PRKCZ	STIL	ITGA2B
ACSL6	CAV1	CPOX	FES	ITSN1	NCAPG2	PROS1	STK11	MIS18A
AGPAT9	CCNA2	CSF2RA	FGFR1	KCNE2	NCAPH	PSMD1	TAL1	PPOX
AHSP	**CCNB1**	CTSB	FGFR3	KCNH2	NCOA3	PTK2	TFDP1	SPAG5
AK1	**CCNB2**	CTTN	FHL2	KIF11	NDC1	PTTG1	TFRC	ZEB1
AKAP1	**CCNE1**	CUX1	FZD5	KIF15	NDC80	RECQL4	TK1	C11orf30
ALAD	CCNF	CXCL12	GAB1	KIF18A	NEDD4L	REPS1	TLE1	**CHEK1**
ALAS2	CD163	DCAF11	GART	KIF20A	NEK2	REPS2	TNXB	FAM20C
ANKRD9	CD36	DCAF7	GAS6	KIF23	NMU	RFC3	TOP2A	IRF8
ANLN	CD59	DCUN1D1	GATA1	KIF26A	NUF2	RFC5	TPM1	MINPP1
APOC1	CDC20	DEPDC1	GATM	KIF2C	NUSAP1	RNF123	TPX2	PPME1
APOE	CDC25A	DHFR	GCLM	KIF4A	ORC1	RNF14	TRIM10	SMC4
ARG1	CDC25C	DIAPH3	GOLM1	KIFC1	OTUD7B	RNF182	TTK	XK
ASPM	CDC27	DLGAP5	GRHL1	LAT2	PBK	RPA3	TTN	BUB1
ATP6V1A	CDC45	DP2	**GTSE1**	LILRB2	PBX1	RRM1	TUBG1	CEP89
AURKA	CDCA2	DPM2	GYPB	LIN7A	PC	SCN11A	TYMS	EXO1
AURKB	CDCA3	DTL	HBB	LTBP1	PF4	SDC1	UBAC1	IL1R1
AXL	CDCA8	DTYMK	HBD	LUM	PHLDA1	SDC2	UBE2C	MGAT3
BIRC5	**CDK1**	E2F2	HELZ2	LYPD3	PI4K2A	SGOL1	UBE2S	PPAPDC1A
BLVRA	CDKN3	ECT2	HJURP	MAD2L1	PIK3R3	SH2D3C	UBE2T	SMC2
BMP2	CENPE	ENG	HLA-DOA	MARK4	PIP5K1B	SHCBP1	VCAM1	WHSC1
BOLA3	CENPF	EPB42	HMBS	MCM10	PKLR	SKA1	VCAN	
BRCA2	CENPN	ERCC6L	HMMR	MCM4	PLK4	SLC25A37	VNN1	
BTRC	CENPW	ESCO2	IGLL5	ME1	POLE2	SMAD1	VPREB1	

4.4. COMPARATIVE ANALYSIS OF PPIs OF DEGs

4.4.1. PPIs Vs LYMPHOCYTIC LEUKEMIA

While comparing of DEGs of Lymphocytic Leukemia both Acute and Chronic, there are 97 pathways in KEGG are found in both Lymphocytic Leukemia DEGs. This study focuses on cancer and leukemia and hence only the top enriched KEGG pathways are given below in Table 7.

Table 7.
Top enriched KEGG pathways of Lymphocytic Leukemia

ABC transporters	Insulin resistance
African trypanosomiasis	Insulin signaling pathway
AGE-RAGE signaling pathway in diabetic complications	Intestinal immune network for IgA production
Alcoholism	Kaposi's sarcoma-associated herpesvirus infection
Aldosterone-regulated sodium reabsorption	Leishmaniasis
Allograft rejection	Long-term depression
Amoebiasis	Lysosome
AMPK signaling pathway	Malaria
Apelin signaling pathway	MAPK signaling pathway
Apoptosis	Metabolic pathways
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	MicroRNAs in cancer
Asthma	Mismatch repair
Autoimmune thyroid disease	Morphine addiction
Axon guidance	Non-small cell lung cancer
Breast cancer	Nucleotide excision repair
Carbon metabolism	Osteoclast differentiation
Cell cycle	p53 signaling pathway
Central carbon metabolism in cancer	Pathways in cancer
Choline metabolism in cancer	Pertussis
Cholinergic synapse	Phagosome
Circadian entrainment	Phosphatidylinositol signaling system
Cytokine-cytokine receptor interaction	Phospholipase D signaling pathway
DNA replication	PI3K-Akt signaling pathway
Dopaminergic synapse	Platelet activation
ECM-receptor interaction	Progesterone-mediated oocyte maturation
EGFR tyrosine kinase inhibitor resistance	Protein digestion and absorption
Endocytosis	Proteoglycans in cancer
ErbB signaling pathway	Purine metabolism

Fatty acid degradation	Rap1 signaling pathway
Fc gamma R-mediated phagocytosis	Ras signaling pathway
GABAergic synapse	Regulation of actin cytoskeleton
Gap junction	Renal cell carcinoma
Glioma	Retrograde endocannabinoid signaling
Glucagon signaling pathway	Rheumatoid arthritis
Glutamatergic synapse	Serotonergic synapse
Glutathione metabolism	Signaling pathways regulating pluripotency of stem cells
Glycosaminoglycan degradation	Sphingolipid signaling pathway
Glycosphingolipid biosynthesis - globo and isoglobo series	Staphylococcus aureus infection
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	Starch and sucrose metabolism
Graft-versus-host disease	Synaptic vesicle cycle
Hematopoietic cell lineage	Systemic lupus erythematosus
HIF-1 signaling pathway	Th1 and Th2 cell differentiation
Human papillomavirus infection	Th17 cell differentiation
Inflammatory bowel disease (IBD)	Thyroid hormone signaling pathway
Inflammatory mediator regulation of TRP channels	Toxoplasmosis
Influenza A	Transcriptional misregulation in cancer
Inositol phosphate metabolism	Tuberculosis
VEGF signaling pathway	Type I and Type II diabetes mellitus

4.4.2. PPIs Vs MYELOID LEUKEMIA

While comparing of DEGs of Chronic Leukemia both Acute and Chronic, there are 8 pathways in KEGG are found in both Myeloid Leukemia DEGs. This study focusing on cancer pathways. Thus the top enriched KEGG pathways are given below in Table 8.

Table 8.
Top enriched KEGG pathways of Myeloid Leukemia

ABC transporters
Cell cycle
DNA replication
Hematopoietic cell lineage
Mismatch repair
Oocyte meiosis
p53 signaling pathway
Progesterone-mediated oocyte maturation

4.4.3. LYMPHOCYTIC Vs MYELOID LEUKEMIA

When comparison with significant metabolic pathways in KEGG, Lymphocytic Leukemia both Acute and Chronic includes certain genes that are found in miRNA's in Cancer, Apoptosis and Pathways in Cancer. However, no such genetic signatures are identified in Myeloid Leukemia both Acute and Chronic.

4.5. GENETIC SIGNATURES IN LEUKEMIA

We further analyzed DEGs of ALL, AML,CLL and CML with the non-leukemia genes which are found to be overlapped with different leukemia pathways.

Table 9.
Significant Genes of different Leukemia pathways

ALL	AML	CLL	CML
BCR-ABL	FLT3	Bcl-2	BCR-ABL
MLL-AF4	c-KIT	p53	EVI1
E2A-PBX1	N-ras	ATM	AML1
TEL-AML1	K-ras	Fas	p16/INK4A
c-MYC	PML-RARalpha		p53
CRLF2	AML1-ETO		RB1
PAX5	PLZF-RARalpha		
NOTCH1	AML1		
TAL1	C/EBPalpha		
LYL1	PU1		
MLL-ENL			
HOX11			
MYC			
LMO2			
HOX11L2			
PICALM-MLLT10			

In ALL DEGs, PAX5 and NOTCH1 are up-regulated whereas TAL1 is down-regulated gene. In AML DEGs, FLT3 gene is up-regulated. In CLL, ATM is up-regulated and Fas is down-regulated during expression. In CML DEGs, RB1 gene is down-regulated.

4.5.1. SIGNIFICANT GENES OF ALL

From the above results, we found PAX5, NOTCH1 and TAL1 are the genes expressed in Acute Lymphocytic Leukemia pathway.

PAX5 (Paired Box 5) is a Protein Coding gene. Diseases associated with PAX5 include Leukemia, Acute Lymphoblastic and Leukemia. Using integrated genomic analysis of leukemic cells from 1,988 childhood and adult cases of B-progenitor acute lymphoblastic leukemia (B-ALL), described a revised taxonomy of B-ALL incorporating 23 subtypes. Two subtypes were characterized by distinct gene expression profiles and different types of PAX5 alterations. Lesions in the PAX5 and IKZF1 genes, encoding B-lymphoid transcription factors, occur in over 80% of cases of pre-B-cell acute lymphoblastic leukemia (ALL). By combining studies using chromatin immunoprecipitation with sequencing and RNA sequencing, identified a novel B-lymphoid program for transcriptional repression of glucose and energy supply. The metabolic analyses revealed that PAX5 and IKZF1 enforce a state of chronic energy deprivation, resulting in constitutive activation of the energy-stress sensor AMPK (Zhaohui, *et al*., 2019).

Pax5 transcription factor, also known as B-cell specific activator protein (BSAP), plays a dual role in the hematopoietic system. Pax5 expression is essential in B-cell precursors for normal differentiation and maturation of B-cells. On the other hand, it inhibits the differentiation and progress toward other lineages. The expression of this factor is involved in several aspects of B-cell differentiation, including commitment, immunoglobulin gene rearrangement, BCR signal transduction and B-cell survival, so that the deletion or inactivating mutations of Pax5 cause cell arrest in Pro-B-cell stage. In recent years, point mutations, deletions and various rearrangements in Pax5 gene have been reported in several types of human cancers. However, no clear relationship has been found between these aberrations and disease prognosis. Specific expression of Pax5 in B-cells can raise it as a marker for the diagnosis and differentiation of B-cell leukemias and lymphomas as well as account for remission or relapse (Shahjahani, *et al*., 2015).

PAX5 is found to be DEGs in Non-leukemia and ALL samples where it is up-regulated. This could be modulated to be down-regulated which may present leukemogenesis.

Notch signaling pathway regulates many different events of embryonic and adult development; among them, Notch plays an essential role in the onset of hematopoietic stem cells

and influences multiple maturation steps of developing lymphoid and myeloid cells. Deregulation of Notch signaling determines several human disorders, including cancer. In the last decade it became evident that Notch signaling plays pivotal roles in the onset and development of T- and B-cell acute lymphoblastic leukemia by regulating the intracellular molecular pathways involved in leukemia cell survival and proliferation. On the other hand, bone marrow stromal cells are equally necessary for leukemia cell survival by preventing blast cell apoptosis and favoring their reciprocal interactions and cross-talk with bone marrow microenvironment. Quite surprisingly, the link between Notch signaling pathway and bone marrow stromal cells in acute lymphoblastic leukemia has been pointed out only recently. In fact, bone marrow stromal cells express Notch receptors and ligands, through which they can interact with and influence normal and leukemia T- and B-cell survival (Kamdje and Krampera, 2015).

NOTCH1 gene modulation therefore may help to prevent the onset of development of T and B cell Acute lymphoblastic leukemia.

T-cell acute leukemia (TAL1) is a lineage-specific oncogene, as its forced expression in T-cells lead to T-ALL, whereas overexpression under a ubiquitous promoter induces bony abnormalities but not cancers in other tissues. Moreover, TAL1, GATA3, RUNX1 and MYB gene loci are associated with super-enhancers in T-ALL cells (Sanda and Leong, 2017).

Finally, identification of targets of PAX5, NOTCH1 and TAL1 is crucial for understanding the molecular pathogenesis of T-ALL as well as PAX5, NOTCH1 could be down-regulated and TAL1 could be up-regulated for developing novel therapeutic strategies.

4.5.2. SIGNIFICANT GENES OF AML

The FMS-like tyrosine kinase 3 (FLT3) pathway has an important role in cellular proliferation, survival, and differentiation. Acute myeloid leukemia patients with mutated FLT3 have a large disease burden at presentation and a dismal prognosis. A number of FLT3 inhibitors have been developed over the years. The first-generation inhibitors are largely non-specific, while the second-generation inhibitors are more specific and more potent. These inhibitors are used to treat patients with FLT3-mutated AML in virtually all disease settings

including induction, consolidation, maintenance, relapse, and after hematopoietic cell transplantation (HCT) (Fakih, *et al*., 2018).

Several studies have shown that internal tandem duplication (ITD) of FMS-like tyrosine kinase 3 (*FLT3*) can result in the failure of leukemia treatment and contribute to a poor prognosis. By analyzing Gene Expression Omnibus and The Cancer Genome Atlas data, it was found that genetic alterations and modification of DNA methylation increased the expression of *FLT3* in leukemia (Cheng *et al*., 2018).

FLT3 is up-regulated in the AML dataset. The modification of FLT3 as down-regulated could be the effective change during leukemia therapy. Finally, identification of targets of FLT3 is crucial for understanding the molecular pathogenesis of AML as well as for developing novel therapeutic strategies.

4.5.3. SIGNIFICANT GENES OF CLL

Ataxia Telangiectasia Mutated (ATM) is a DNA-damage response gene that is commonly mutated in cancer. Inherited mutations in the ATM (Ataxia-Telangiesctasia mutated) gene are associated with increased risk of certain cancers. People who inherit a mutated copy of ATM from one parent are at increased risk of female breast cancer (up to 52% life time risk), and possibly pancreatic, prostate and other cancers (ATM gene alterations in chronic lymphocytic leukemia patients induce a distinct gene expression profile and predict disease progression). Inherited biallelic mutations of the ATM (ataxia-telangiectasia mutated) gene cause ataxia-telangiectasia, a rare autosomal recessive disorder associated with a high incidence of childhood leukemias and lymphomas, suggesting that *ATM* gene alterations may be involved in lymphomagenesis (Gumy-Pause *et al*., 2014).

Fas (Fas Cell Surface Death Receptor antigen) (Apo-1/CD95) is an apoptosis-signaling cell surface receptor belonging to the tumor necrosis factor receptor superfamily. Adult T cell leukemia (ATL) cells express Fas antigen and show apoptosis after treatment with an anti-Fas monoclonal antibody. The Fas antigen (Apo-1/CD95) is a 45-kD transmembrane protein that is a member of the TNF receptor superfamily that can induce programmed cell death, i.e., apoptosis, when cross-linked by natural ligand or specific antibodies. The Fas/Fas ligand system of

apoptosis appears to play an important role in the normal development of T lymphocytes in the thymus by eliminating self-reactive lymphocytes (Sanséau and Legembre, 2015).

Finally, identification of targets of ATM and Fas is crucial for understanding the molecular pathogenesis of CLL. As well as for developing novel therapeutic strategies the ATM could be down-regulated and Fas could be up-regulated. This modulation could help in adopting therapeutic strategy for leukemia.

4.5.4. SIGNIFICANT GENES OF CML

The retinoblastoma protein (RB1) is a tumor suppressor protein that is dysfunctional in several major cancers. One function of Rb is to prevent excessive cell growth by inhibiting cell cycle progression until a cell is ready to divide. In particular, the p53 and RB1 tumor-suppressor genes have been found to be involved in a variety of human cancers, including both solid tumors and hematopoietic malignancies (Gaidano *et al.*, 2014).

RB1 (RB Transcriptional Corepressor 1) is a Protein Coding gene. Diseases associated with RB1 include Retinoblastoma and Small Cell Cancer Of The Lung. The protein encoded by this gene is a negative regulator of the cell cycle and was the first tumor suppressor gene found. The encoded protein also stabilizes constitutive heterochromatin to maintain the overall chromatin structure. The active, hypophosphorylated form of the protein binds transcription factor E2F1. Defects in this gene are a cause of childhood cancer retinoblastoma (RB), bladder cancer, and osteogenic sarcoma (https://www.genecards.org/cgi-bin/carddisp.pl?gene=RB1).

Finally, identification of targets of RB1 is crucial for understanding the molecular pathogenesis of CML as well as for developing novel therapeutic strategies. RB1 is found to be down-regulated in CML datasets whereas it could be up-regulated gives the effective change in leukemia therapy.

4.6. FUNCTIONAL ENRICHMENT ANALYSIS

To gain insight about the biological significance of the differentially expressed genes. Gene Ontology (GO) enrichment analysis was performed using web based software GENECODIS. Gene Ontology studies provide a brief descriptive framework on the functional

annotation and biological classification of the gene sets in three different categories: Biological Process, Cellular Component and Molecular Functions.

4.6.1. ENRICHMENT ANALYSIS OF GO BIOLOGICAL PROCESS

The initial process in functional enrichment analysis is the biological process of each differentially expressed genes. Figure 12 shows the differentially expressed genes of Non-Leukemia compared to ALL with significantly enriched GO terms in Biological Process (BP). Figure 13 shows the differentially expressed genes of Non-Leukemia compared to AML with significantly enriched GO terms in Biological Process (BP). Figure 14 shows the differentially expressed genes of Non-Leukemia compared to CLL with significantly enriched GO terms in Biological Process (BP). Figure 15 shows the differentially expressed genes of Non-Leukemia compared to CML with significantly enriched GO terms in Biological Process (BP).

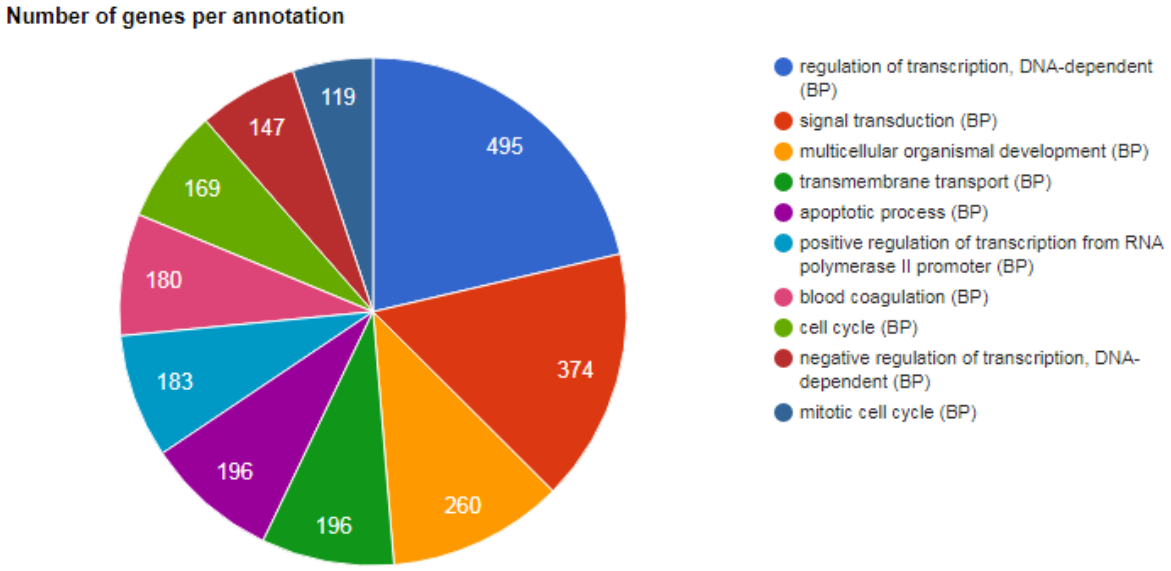


Figure 12. Biological Process of DEGs in Non-Leukemia compared to ALL

From the figure above mentioned, the more significantly enriched BP were mainly associated with regulation of transcription (DNA dependent) and signal transduction.

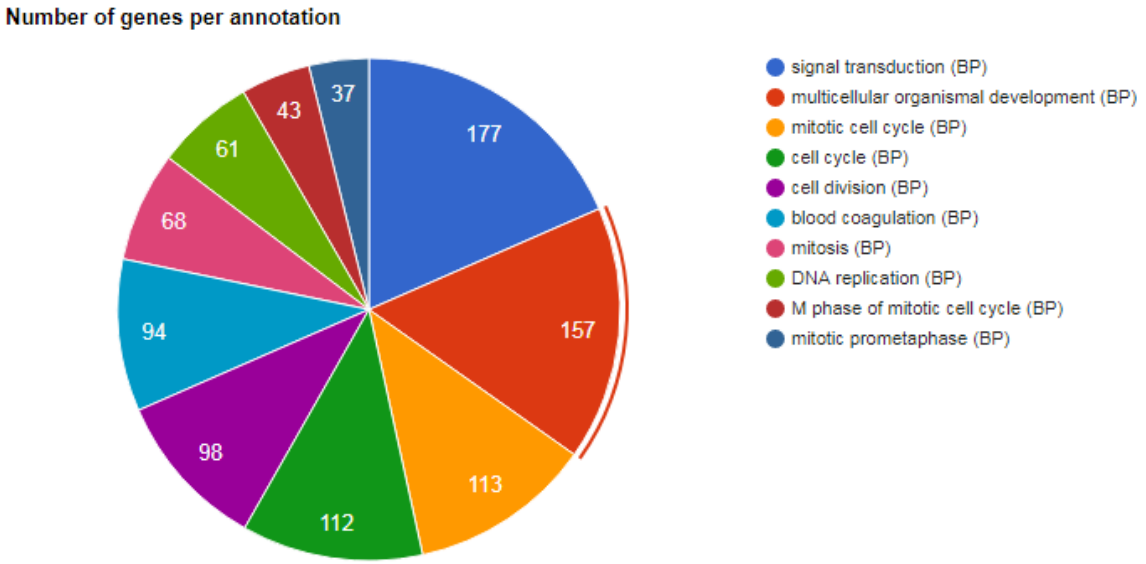


Figure 13. Biological Process of DEGs in Non-Leukemia compared to AML

Significal biological activity rendered by the DEGs of AML type includes signal transduction, multicellular organismal development.

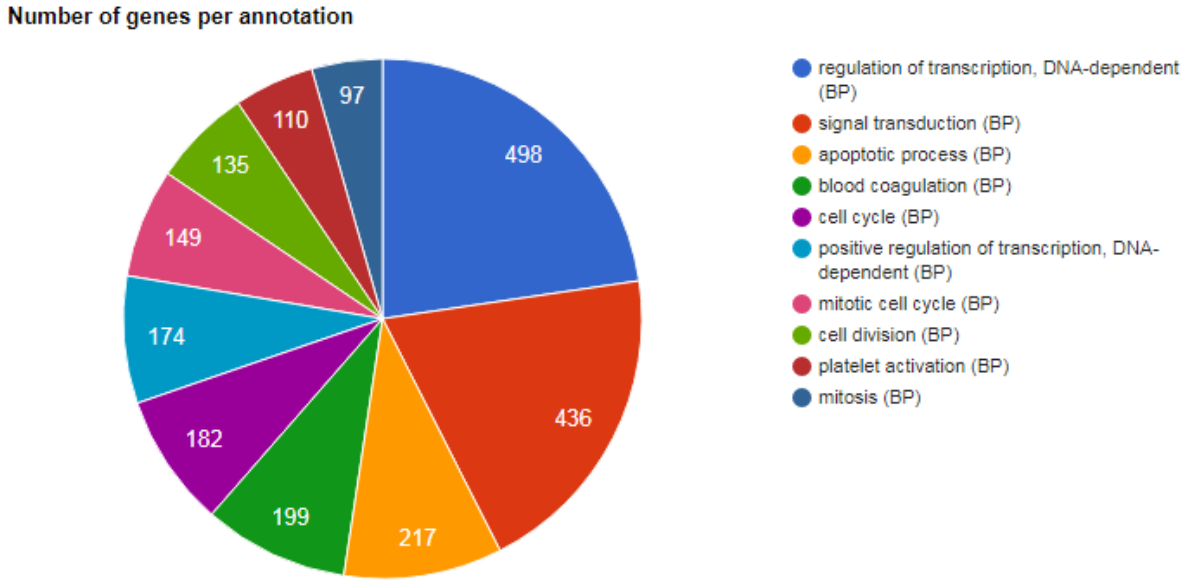


Figure 14. Biological Process of DEGs in Non-Leukemia compared to CLL

From the figure above mentioned, the more significantly enriched BP were mainly associated with regulation of transcription (DNA dependent) and signal transduction.

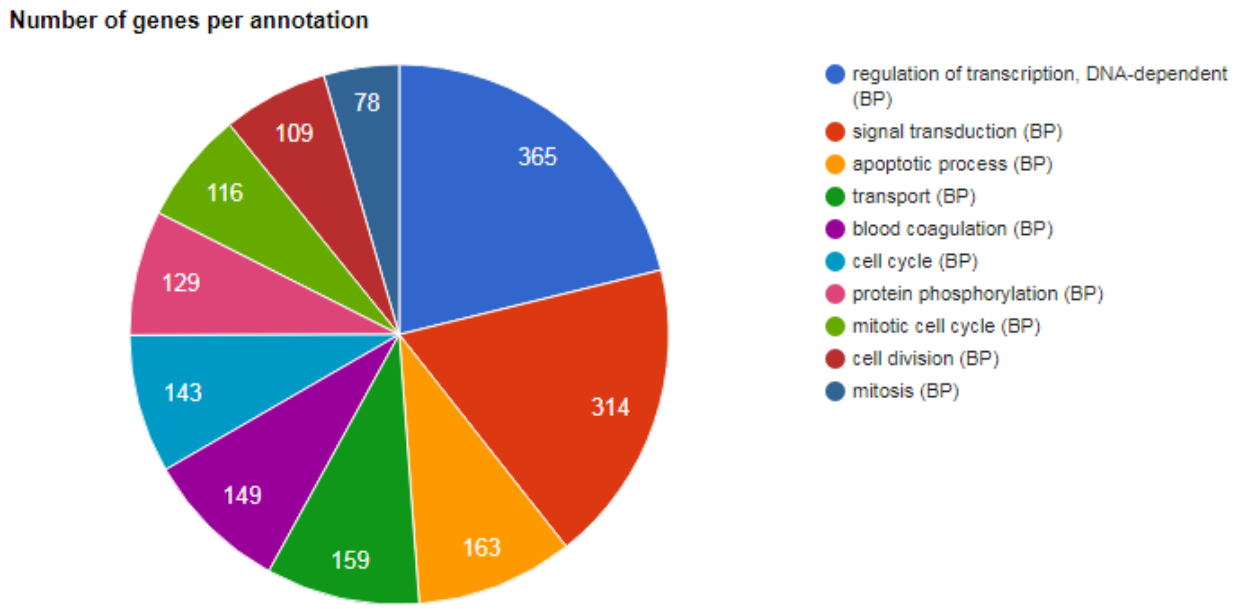


Figure 15. Biological Process of DEGs in Non-Leukemia compared to CML

From the figure above mentioned, the more significantly enriched BP were mainly associated with regulation of transcription (DNA dependent) and signal transduction.

Genes involved in signal transduction were differentially expressed in all the four types of leukemia. Also genes involved in transcription regulation were also modulated in all the three types of leukemia ALL, CLL, CML excepting AML. Interestingly, AML has more DEGs involved in multicellular organisimal development reiterating the results observed from PPIs and venn diagram constructed, that AML has specific characteristic features that less coindices with the other types of leukemia analyzed.

4.7. DISCUSSION

The primary objective of the study was to identify differentially expressed genes among Non-Leukemia and four types of Leukemia (ALL, AML, CLL, CML) with the gene expression data obtained from peripheral blood (ALL, AML,CLL, CML) and healthy bone marrow (Non-Leukemia). Since, the selection of data was solely based on Non-Leukemia and Leukemia, there were many exclusions incorporated in retrieving the datasets for the study. But in spite of all limit, the study revealed certain features characteristic of different types of Leukemia.

In this study, the microarray data obtained from databases pertaining to the conditions: Non-Leukemia and Leukemia were integrated and successfully adopted for Meta-analysis. Firstly, the datasets were processed and normalized to gain knowledge about the quality of the data. The average intensities of each sample were viewed using a box plot before and after normalization. And the distribution of expression values in terms of conversion based on log2 transformation was also plotted and visualized.

Primarily, the differential gene expression analysis were carried out by comparing the two conditions: Non-Leukemia and four types of Leukemia (ALL, AML, CLL, CML). The results obtained from the expression analysis suggested a list of DEGs between the conditions with the regulation of fold change value > 2.0 and p-value < 0.05. Among the DEGs, they were further classified as up-regulated and down-regulated genes.

4.7.1. DEGs INVOLVED IN COMMON PATHWAYS

The differentially expressed genes between the Non-Leukemia and Leukemia (ALL, AML, CLL, CML) were shown in Table 5. which has significant response to pathways such as ABC transporters, Arrhythmogenic right ventricular cardiomyopathy (ARVC), Cell cycle, Cellular senescence, DNA replication, Fatty acid degradation, Hematopoietic cell lineage, Mismatch repair, p53 signaling pathway and Progesterone-mediated oocyte maturation.

Cellular senescence is a state of irreversible cellular arrest and can be triggered by a number of factors, such as telomere shortening, oncogene activation, irradiation, DNA damage and oxidative stress. Blood-cell development progresses from a hematopoietic stem cell (HSC), which can undergo either self-renewal or differentiation into a multilineage committed progenitor cell: a common lymphoid progenitor (CLP) or a common myeloid progenitor (CMP). A CLP gives rise to the lymphoid lineage of white blood cells or leukocytes-the natural killer (NK) cells and the T and B lymphocytes. A CMP gives rise to the myeloid lineage, which comprises the rest of the leukocytes, the erythrocytes (red blood cells), and the megakaryocytes that produce platelets important in blood clotting. Cells undergoing these differentiation process express a stage- and lineage-specific set of surface markers. Therefore cellular stages are identified by the specific expression patterns of these genes. p53 activation is induced by a

number of stress signals, including DNA damage, oxidative stress and activated oncogenes. (https://www.genome.jp/kegg-bin/show_pathway?hsa05206).

MiRNAs are also involved in resisting cell death by regulating components of extrinsic apoptotic pathway, such as the Fas ligand/Fas receptor (Peng., 2016). The upregulation (overexpression) of specific miRNAs could lead to the repression of tumor suppressor gene expression, and conversely the downregulation of specific miRNAs could result in an increase of oncogene expression; both these situations induce subsequent malignant effects on cell proliferation, differentiation, and apoptosis that lead to tumor growth and progress.

Several pathways were found to be in correlation with the DEGs of four types of leukemia. However, in the present study we have chosen only the following pathways based on their relation to cancer biology.

Apoptosis is a genetically programmed process for the elimination of damaged or redundant cells by activation of caspases (aspartate-specific cysteine proteases). The onset of apoptosis is controlled by numerous interrelating processes. The 'extrinsic' pathway involves stimulation of members of the tumor necrosis factor (TNF) receptor subfamily, such as TNFRI, CD95/Fas or TRAILR (death receptors), located at the cell surface, by their specific ligands, such as TNF-alpha, FasL or TRAIL, respectively.

While, analysis of Lymphocytic Leukemia both Acute and Chronic were involved in specific pathways such as microRNA's in Cancer, Apoptosis and Pathways in Cancer whereas no such genetic signatures were identified in Myeloid Leukemia both Acute and Chronic.

4.7.2. DEGs INVOLVED IN SELECTED/SPECIFIC PATHWAYS OF CANCER

We further anlayzed DEGs of ALL, AML,CLL, CML compared to Non_Leukemia, with their specific pathways such ALL, AML, CLL, CML. From this analysis we retrieved that the significant genes which expresses in our comparative analysis.

Finally, identification of targets of PAX5, NOTCH1 and TAL1 of T-ALL, FLT3 of AML, ATM and Fas CLL, FB1 of CML are crucial for understanding the molecular pathogenesis of specified Leukemia as well as for developing novel therapeutic strategies. These genes were further investigate for its PPIs results given in Table 10.

Table 10.
Significant Genes and metabolic pathways

#term ID	term description	false discovery rate	matching proteins in your network (labels)
hsa05165	Human papillomavirus infection	0.00013	ATM,FAS,NOTCH1,RB1
hsa05200	Pathways in cancer	0.00045	FAS,FLT3,NOTCH1,RB1
hsa05202	Transcriptional misregulation in cancer	0.00045	ATM,FLT3,PAX5
hsa01524	Platinum drug resistance	0.0037	ATM,FAS
hsa04115	p53 signaling pathway	0.0037	ATM,FAS
hsa01522	Endocrine resistance	0.0047	NOTCH1,RB1
hsa04110	Cell cycle	0.0067	ATM,RB1
hsa04210	Apoptosis	0.007	ATM,FAS
hsa04218	Cellular senescence	0.007	ATM,RB1
hsa05161	Hepatitis B	0.007	FAS,RB1
hsa05206	MicroRNAs in cancer	0.007	ATM,NOTCH1
hsa05224	Breast cancer	0.007	NOTCH1,RB1
hsa05167	Kaposi's sarcoma-associated herpesvirus infection	0.0078	FAS,RB1
hsa05166	HTLV-I infection	0.0133	ATM,RB1
hsa04060	Cytokine-cytokine receptor interaction	0.0137	FAS,FLT3
hsa04010	MAPK signaling pathway	0.0158	FAS,FLT3

It can be concluded that DEGs of Myeloid leukemia type were found to have lesser interactions when compared to lymphocytic leukemia.

5

SUMMARY AND CONCLUSION

A study on expression of genes in various types of leukemia using in silico tools - a therapeutic approach

5. SUMMARY AND CONCLUSION

Leukemia is a form of cancer of the blood or bone marrow. The blood/marrow produces an abnormal amount of immature white blood cells known as blasts. Leukemia is treatable through chemotherapy, radiation, bone marrow transplants and hormone treatments. Complete remission is more likely among children than adults. Bioinformatics tools are used to help facilitate scientific exploration of related genes, diseases and pathways based on co-citations.

Leukemia is classified into four main categories or subtypes according to cell type and rate of growth: acute lymphocytic leukemia derived from immature T- or B-lymphocytes, most common in children; acute myeloid leukemia from immature myeloid cells, most common in adults; chronic lymphocytic leukemia from mature B-lymphocytes, mostly an adult disorder; and chronic myelogenous leukemia from granulocyte precursors, most common in adults. Due to the complex progression, the therapy is particularly challenging. This study aimed to find key genes and pathways related with Leukemia and their significant role in causing the disease.

In this study, the microarray datasets pertaining to metabolically healthy bone marrow (Non_Leukemia) from four types of Leukemia (ALL, AML, CLL, CML) which are obtained from peripheral blood samples. These datasets were extracted from gene expression databases namely GEO and ArrayExpress. Upon completion of extraction data with the selected criteria, the datasets were grouped into two as metabolically healthy (Non_Leukemia) and metabolically unhealthy (Leukemia) in order to carry out the identification of differential of gene expression in between both the conditions.

The significant up-regulated and down-regulated genes among the differentially expressed genes DEGs between the conditions were identified. The DEGs were further analyzed for their biological process only. To gain knowledge about the mechanism or role of Leukemia, DEGs were annotated for identifying the pathway using KEGG data source. Interesting inferences were made from the pathway analysis of DEGs which showed ABC transporters, Arrhythmogenic right ventricular cardiomyopathy (ARVC), Cell cycle, Cellular senescence, DNA replication, Fatty acid degradation, Hematopoietic cell lineage, Mismatch repair, p53 signaling pathway and Progesterone-mediated oocyte maturation.

This study focuses on Leukemia and Cancer pathways. The most DEGs were involved in Cellular senescence, Hematopoietic cell lineage, p53 signaling pathway and Progesterone-mediated oocyte maturationpathways. These pathways are common to all the resulted DEGs. Further comparison between Lymphocytic Leukemia (Acute and Chronic), as well as Myeloid Leukemia (Acute and Chronic), we found the DEGs involved in miRNA's in Cancer, Apoptosis and Pathways in Cancer. However, no such genetic signatures are identified in Myeloid Leukemia both Acute and Chronic. From this we concluded that the genes involved in Myeloid Leukemia doesn't involved in any interactions with the genes participated in above mentioned pathways.

Further we analyzed DEGs, with their specific pathways such ALL, AML, CLL, CML individually. Significant genes may play vital role in causing leukemia includes PAX5, NOTCH1 and TAL1 of T-ALL, FLT3 of AML, ATM and Fas CLL, FB1 of CML which expresses in our comparative analysis. Thus, ATM and FAS were involved in pathways of microRNA's in cancer. As well as ATM, FAS and NOTCH1 were involved in pathways of Apoptosis. From this, also we concluded that no such genetic signatures are identified in Myeloid Leukemia type both Acute and Chronic.

Finally, identification of targets of PAX5, NOTCH1 and TAL1 of T-ALL, FLT3 of AML, ATM and Fas CLL, FB1 of CML are crucial for understanding the molecular pathogenesis of specified Leukemia as well as for developing novel therapeutic strategies. By gene modulation i.e up-regulating TAL1, Fas and RB1 as well as down-regulating Pax5, NOTCH1, FLT3 and ATM could be the effective alteration that would bring significant outcomes in the leukemia therapy.

BIBLIOGRAPHY & WEBOGRAPHY

BIBLIOGRAPHY

- Aboul-Soud, M.A., El-Shemy, H.A., Aboul-Enein, K.M., Mahmoud, A.M., Al-Abd, A.M. and Lightfoot, D.A., (2016) Effects of plant-derived anti-leukemic drugs on individualized leukemic cell population profiles in Egyptian patients, *Oncology letters*, *11*(1), 642-648.
- Albano, F., Zagaria, A., Anelli, L., Coccaro, N., Impera, L., Minervini, C.F., Minervini, A., Rossi, A.R., Tota, G., Casieri, P. and Specchia, G., (2013) Gene expression profiling of chronic myeloid leukemia with variant t (9; 22) reveals a different signature from cases with classic translocation, *Molecular cancer*, *12*(1), 36.
- Andrew, H., Florence, G. and Kibria, G.B., (2015) Methods for identifying differentially expressed genes: An empirical comparison, *Journal of Biometrics & Biostatistics*, *6*(5), 1.
- Barrett, T., Suzek, T.O., Troup, D.B., Wilhite, S.E., Ngau, W.C., Ledoux, P., Rudnev, D., Lash, A.E., Fujibuchi, W. and Edgar, R., (2015) NCBI GEO: mining millions of expression profiles—database and tools, *Nucleic acids research*, *33*(suppl_1), D562-D566.
- Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M. and Yefanov, A., (2013) NCBI GEO: archive for functional genomics data sets—update, *Nucleic acids research*, *41*(D1), D991-D995.
- Casassola, A., Brammer, S.P., Chaves, M.S., Martinelli, J.A., Grando, M.F. and Denardin, N.D.Á., (2013) Gene expression: a review on methods for the study of defense-related gene differential expression in plants, *American Journal of Plant Sciences*, *4*(12), 64.
- Cheng, J., Qu, L., Wang, J., Cheng, L. and Wang, Y., (2018) High expression of FLT3 is a risk factor in leukemia, *Molecular medicine reports*, *17*(2), 2885-2892.
- Cheng, Q., Bai, S., Ge, G., Li, P., Liu, L., Zhang, C. and Jia, Y., (2018) Study on differentially expressed genes related to defoliation traits in two alfalfa varieties based on RNA-Seq, *BMC genomics*, *19*(1), 807.
- Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J. and Foa, R., (2014) Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival, *Blood*, *103*(7), 2771-2778.

- El Fakih, R., Rasheed, W., Hawsawi, Y., Alsermani, M. and Hassanein, M., (2018) Targeting FLT3 mutations in acute myeloid leukemia, *Cells*, *7*(1), 4.
- Fang, S.Q., Gao, M., Xiong, S.L., Chen, H.Y., Hu, S.S. and Cai, H.B., (2018) Combining differential expression and differential coexpression analysis identifies optimal gene and gene set in cervical cancer, *Journal of cancer research and therapeutics*, *14*(1), 201.
- Gaidano, G., Serra, A., Guerrasio, A., Rege-Cambrin, G., Mazza, U. and Saglio, G., (2014) Genetic analysis of p53 and RB1 tumor-suppressor genes in blast crisis of chronic myeloid leukemia, *Annals of hematology*, *68*(1), 3-7.
- García-Campos, M.A., Espinal-Enríquez, J. and Hernández-Lemus, E., (2015) Pathway analysis: state of the art, *Frontiers in physiology*, *6*, 383.
- Grgurevic, S., Berquet, L., Quillet-Mary, A., Laurent, G., Récher, C., Ysebaert, L., Cazaux, C. and Hoffmann, J.S., (2016) 3R gene expression in chronic lymphocytic leukemia reveals insight into disease evolution, *Blood cancer journal*, *6*(6), 429.
- Gu, Z., Churchman, M.L., Roberts, K.G., Moore, I., Zhou, X., Nakitandwe, J., Hagiwara, K., Pelletier, S., Gingras, S., Berns, H. and Payne-Turner, D., (2019) PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia, *Nature genetics*, *51*(2), 296.
- Guda, P., Chittur, S.V. and Guda, C., (2019) Comparative analysis of protein-protein interactions in cancer-associated genes, *Genomics, proteomics & bioinformatics*, *7*(1-2), 25-36.
- Gumy-Pause, F., Wacker, P. and Sappino, A.P., (2004) ATM gene and lymphoid malignancies, *Leukemia*, *18*(2), 238.
- Handschuh, L., Kaźmierczak, M., Milewski, M.C., Góralski, M., Łuczak, M., Wojtaszewska, M., Uszczyńska-Ratajczak, B., Lewandowski, K., Komarnicki, M. and Figlerowicz, M., (2018) Gene expression profiling of acute myeloid leukemia samples from adult patients with AML-M1 and-M2 through boutique microarrays, real-time PCR and droplet digital PCR, *International journal of oncology*, *52*(3), 656-678.
- Hedblom, A., Laursen, K., Miftakhova, R., Sarwar, M., Anagnostaki, L., Bredberg, A., Mongan, N., Gudas, L.J. and Persson, J., (2013) CDK1 interacts with RARγ and plays an important role in treatment response of acute myeloid leukemia, *Cell Cycle*, *12*(8), pp.1251-1266.

- Kamdje, A.H.N. and Krampera, M., (2015) Notch signaling in acute lymphoblastic leukemia: any role for stromal microenvironment?, *Blood*, *118*(25), 6506-6514.
- Kodama, K., Horikoshi, M., Toda, K., Yamada, S., Hara, K., Irie, J., Sirota, M., Morgan, A.A., Chen, R., Ohtsu, H. and Maeda, S., (2014) Expression-based genome-wide association study links the receptor CD44 in adipose tissue with type 2 diabetes, *Proceedings of the National Academy of Sciences*, *109*(18), 7049-7054.
- Kumar, C.C., (2015) Genetic abnormalities and challenges in the treatment of acute myeloid leukemia, *Genes & cancer*, *2*(2), 95-107.
- Lee, S., Chen, J., Zhou, G., Shi, R.Z., Bouffard, G.G., Kocherginsky, M., Ge, X., Sun, M., Jayathilaka, N., Kim, Y.C. and Emmanuel, N., (2016) Gene expression profiles in acute myeloid leukemia with common translocations using SAGE, *Proceedings of the National Academy of Sciences*, *103*(4), 1030-1035.
- Prasad, B.D., Sahni, S., Jha, V.K., Pal, A.K., Kumar, P., Ranjan, T. and Sharma, V., (2017) Protein–Protein Interaction Detection: Methods and Analysis. In *Plant Biotechnology, Volume 1* (pp. 391-411), Apple Academic Press.
- Puthiyedth, N., Riveros, C., Berretta, R. and Moscato, P., (2016) Identification of differentially expressed genes through integrated study of Alzheimer's disease affected brain regions, *PloS one*, *11*(4), 0152342.
- Radich, J.P., Dai, H., Mao, M., Oehler, V., Schelter, J., Druker, B., Sawyers, C., Shah, N., Stock, W., Willman, C.L. and Friend, S., (2016) Gene expression changes associated with progression and response in chronic myeloid leukemia, *Proceedings of the National Academy of Sciences*, *103*(8), 2794-2799.
- Radomska, H.S., Alberich-Jordà, M., Will, B., Gonzalez, D., Delwel, R. and Tenen, D.G., (2014) Targeting CDK1 promotes FLT3-activated acute myeloid leukemia differentiation through C/EBPα, *The Journal of clinical investigation*, *122*(8), pp.2955-2966.
- Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K., (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies, *Nucleic acids research*, *43*(7), e47-e47.
- Robinson, M.D., McCarthy, D.J. and Smyth, G.K., (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics*, *26*(1), 139-140.

- Ruskin, H., (2016) Computational Modeling and Analysis of Microarray Data: New Horizons, 26
- Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Ison, J., Keays, M. and Kurbatova, N., (2013) ArrayExpress update—trends in database growth and links to data analysis tools, *Nucleic acids research*, *41*(D1), D987-D990.
- Sanda, T. and Leong, W.Z., (2017) TAL1 as a master oncogenic transcription factor in T-cell acute lymphoblastic leukemia, *Experimental hematology*, *53*, 7-15.
- Sanséau, D. and Legembre, P., (2015) FAS (Fas cell surface death receptor), *Atlas of Genetics and Cytogenetics in Oncology and Haematology*.
- Schinke-Braun, M. and Couget, J.A., (2017) Expression Profiling Using Affymetrix GeneChip® Probe Arrays. In *Cardiac Gene Expression* (13-40), Humana Press.
- Selvaraj, S. and Natarajan, J., (2015) Microarray data analysis and mining tools. *Bioinformation*, *6*(3), 95.
- Shahjahani, M., Norozi, F., Ahmadzadeh, A., Shahrabi, S., Tavakoli, F., Asnafi, A.A. and Saki, N., (2015) The role of Pax5 in leukemia: diagnosis and prognosis significance, *Medical Oncology*, *32*(1), 360.
- Shahzad Farooq, M., Malik, A., Shaheen, B. and Waquar, S., (2017) LEUKEMIA; DIFFERENTIAL EXPRESSION OF PROPHETIC VARIABLES OF MEDICAL IMPORTANCE AND THEIR POTENTIAL ROLE IN THE PATHOGENESIS, *Professional Medical Journal*, *24*(10).
- Sharma, M. and Porte, S.M., 2016. Role of ayurvedain management of leukemia (Raktarbuda). *Int J Pharm Sci Res*, *7*, 520-530.
- Wang, T., Li, B., Nelson, C.E. and Nabavi, S., 2019. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC bioinformatics*, *20*(1), 40.
- Yepes, S., Torres, M.M. and Andrade, R.E., 2015. Clustering of expression data in chronic lymphocytic leukemia reveals new molecular subdivisions. *PloS one*, *10*(9), 0137132.

WEBOGRAPHY

1) http://bioinformatics.psb.ugent.be/webtools/Venn/
2) http://cancer.columbia.edu/leukemia-classifications
3) http://www.bioinformatics.org/legend/leuk_db.htm
4) http://www.cytoscape.org
5) http://www.ncbi.nlm.nih.gov/geo/info/datasets.html
6) https://string-db.org/
7) https://www.ebi.ac.uk
8) https://www.ebi.ac.uk/arrayexpress/
9) https://www.genecards.org/cgi-bin/carddisp.pl?gene=RB1
10) https://www.genome.jp/kegg/pathway.html
11) https://www.genome.jp/kegg-bin/show_pathway?hsa05206
12) https://www.kegg.jp/dbget-bin/www_bget?hsa05221
13) https://www.lls.org/http%3A/llsorg.prod.acquia-sites.com/facts-and-statistics/facts-and-statistics-overview/facts-and-statistics#Leukemia
14) https://www.ncbi.nlm.nih.gov/geo/
15) https://www.ncbi.nlm.nih.gov/geo/info/profiles.html
16) https://www.r-project.org/
17) https://www.scq.ubc.ca/what-is-bioinformatics/
18) www.genecodis.cnb.csic.es