

ACUTE MYELOID LEUKEMIA CANCER ANALYSIS USING MAFTOOLS

Submitted in partial fulfillment of the requirements for the award of Bachelor of
Sciences Degree in Bioinformatics And Data Science

By

SUTHARSANAN SAMPATH (40738023)



DEPARTMENT OF BIOINFORMATICS

SCHOOL OF BIO AND CHEMICAL ENGINEERING

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited with Grade "A" by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600119

MAY 2023



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF BIOINFORMATICS

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **SUTHARSANAN SAMPATH (Reg.No: 40738023)** who carried out the project entitled **"ACUTE MYELOID LEUKEMIA CANCER ANALYSIS USING MAFTOOLS"** under my supervision from JANUARY 2023 to MAY 2023 .

Internal Guide

Dr. Jemmy Christy H, M.Tech., Ph.D.

nt

Dr. Jemmy Christy H, M.Tech., Ph.D.

Submitted for Viva voce Examination held on 05/05/23

**Internal Examiner
Examiner**

External

DECLARATION

I, **SUTHARSANAN SAMPATH** hereby declare that the project report entitled **“ACUTE MYELOID LEUKEMIA CANCER ANALYSIS USING MAFTOOLS”** done by me under the guidance of **Dr. Jemmy Christy H, M.Tech., Ph.D.** is submitted in partial fulfillment of the requirements for the award of Bachelor of Science Degree in Bioinformatics and Data science.

DATE:05/05/23

sudharsanan

PLACE: CHENNAI

**SIGNATURE OF THE CANDIDATE
(SUTHARSANAN SAMPATH)**

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dean, School of, Dr. Jemmy Christy H, M.Tech., Ph.D Head of the Department of Bioinformatics and Data science** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. Jemmy Christy H, M.Tech., Ph.D** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Bioinformatics and Data science** who were helpful in many ways for the completion of this.

ABSTRACT

The screening of variant mutation linked to Acute Myeloid Leukemia (AML) is investigated using mutation annotation format tools. The development of AML, a kind of cancer that affects the bone marrow and blood, has been linked largely to abnormalities in specific genes. A standardized file format called MAF (Mutation Annotation Format) contains data on genetic variants and how they affect genes.

According to the analysis's findings FLT3, DNMT3A, NPM1, IDH2, TP53 mutations are present in a sizeable percentage of AML patients and are linked to a worse prognosis. The study also identifies other variants that could potentially serve as AML driver mutations.

This study's objective is to employ MAF techniques to analyze the mutations found in AML patients, with a focus on finding FLT3 gene mutations, which are known to be extremely important in AML. A group of AML patients who underwent next-generation sequencing to collect their genomic data is included in the study.

Finally, using MAF tools offers a thorough and consistent method for examining genomic variations and their effect on genes in AML patients. AML diagnosis and treatment, as well as the creation of individualized medicines based on a patient's unique mutation profile, may be significantly impacted by the discovery of FLT3 mutations and other possible driver mutations.

TABLE OF CONTENTS

CHAPTER NO	TITLE	PAGE NO
	ABSTRACT	
	LIST OF FIGURES	
	LIST OF ABBREVIATION	
1	INTRODUCTION	1-2
2	LITERATURE SURVEY	3-4
3	AIM AND SCOPE	5-6
4	MATERIALS AND METHODS	7-11
	4.1 DATA SETS	7
	4.2 R BIOCONDUCTOR PACKAGES	7
	4.3 VISUALIZATION	8
	4.4 CLINICAL ENRICHMENT ANALYSIS	8
	4.5 SOMATIC INTERACTION AND PATHWAY ANALYSIS	9
	4.6 IDENTIFICATION OF CANCER GENES	10
	4.7 PFAM DOMAINS SUMMARIZATION	11
	4.8 SURVIVAL ANALYSIS	11
5	RESULTS AND DISCUSSION	12-31
	5.1 TO VISUALIZE LARGE SCALE GENOMIC DATA	12
	5.2 ONCOPLOTS	13
	5.3 TRANSITIONS AND TRANSVERSION	14
	5.4 LOLLIPOP PLOTS FOR AMINO ACID CHANGES	15
	5.5 RAINFALL PLOTS	15
	5.6 PLOTTING VAF	16

5.7	READING AND SUMMARIZING GISTIC OUTPUT FILES	17
5.8	VISUALIZING GISTIC RESULTS	17
5.9	ONCOPLOT (2)	18
5.10	VISUALIZING CBS SEGEMENTS	19
5.11	TO ANALYZE MUTATIONAL SIGANUTRES AND ENRICHMENTS	20
5.12	DETECTING CANCER DRIVER GENES BASED ON POSITIONAL CLUSTERING SURVIVAL ANALYSIS	21
5.13	SURVIVAL ANALYSIS COMPARING TWO COHORTS	23
5.14	PREDICT GENESETS WITH SURVIVAL	24
5.15	COMPARING TWO COHORTS	24
5.16	FOREST PLOTS	24
5.17	CO-ONCO PLOTS	25
5.18	LOLLIPOP PLOT-2	26
5.19	CLINICAL ENRICHMENT ANALYSIS	27
5.20	DRUG GENE INTERACTION	27
5.21	ONCOGENIC SIGNALING PATHWAY DISCUSSION	28 32
6	SUMMARY AND CONCLUSION	34-35
	REFERENCES	37-39

LIST OF FIGURES

Figure NO.	TITLE	PAGE NO
5.1	SUMMARY OF MAF FILE	12
5.2	MAF BAR PLOT	13
5.3	ONCOPLOT	14
5.4	TRANSITION AND TRANSVERSION	14
5.5	LOLLIPOP PLOTS FOR DNMT3A	15
5.6	RAINFALL PLOTS	16
5.7	VARIANT ALLELE FREQUENCY	16
5.8	GENOME PLOT	17
5.9	BUBBLE PLOT	18
5.10	ONCOPLOT (2)	19
5.11	CIRCULAR BINARY SEGMENTS	19
5.12	SOMATIC INTERACTION	20
5.13	POSITIONAL CLUSTERING	22
5.14	SURVIVAL ANALYSIS	23
5.15	SURVIVAL ANALYSIS OF GENESETS DNMT3A AND FLT3	24
5.16	FOREST PLOTS	25
5.17	CO-ONCO PLOTS	25
5.18	CO-BAR PLOTS	26
5.19	LOLIPOP PLOT_2	26

5.20	FAB CLASSIFICATION	27
5.21	DRUG GENE INTERACTION	28
5.22	DRUGGABLE CATEGORIES	28
5.23	ONCOGENIC SIGNALING PATHWAYS	29
5.24	RTK-RAS PATHWAYS	29

LIST OF ABBREVIATION

MAF	MUTATION ANNOTATION FORMAT
AML	ACUTE MYELOID LEUKEMIA
DNMT3A	DNA(cytosine-5)- methyltransferase 3A
FLT3	Fms-like tyrosine kinase 3
NPM1	NUCLEOPHOSMIN
IDH2	ISOCITRATE DEHYDROGENASE 2
IDE	INTERGRATED DEVELOPMENT ENVIRONMENT
NRAS	NEUROBLASTOMA RASVIRAL ONCOGENE HOMOLOG
TGCA	THE CANCER GENOME ATLAS
RNA	RIBONUCLEIC ACID
GSEA	GENE SET ENRICHMENT ANALYSIS
SNV	SINGLE NUCLEOTIDE POLYMORPHISM

SNP	SINGLE NUCLEOTIDE POLYMORPHISM
WHO	WORLD HEALTH ORGANISATION
RUNX1	RUNT-RELATED TRANSCRIPTION FACTOR
APL	ACUTE PROMYELOCYTIC LEUKEMIA

CHAPTER 1

INTRODUCTION

Aml (acute myeloid leukemia)

Acute myeloid leukemia, or AML, is a serious and life-threatening form of cancer that affects the blood and bone marrow. It is a type of leukemia that develops rapidly, often over a matter of weeks or months, and can cause a wide range of symptoms and complications. AML occurs when abnormal cells in the bone marrow, known as myeloid cells, begin to grow and divide uncontrollably, crowding out healthy blood cells and interfering with their normal function. This can lead to a range of health problems, including anemia, infections, and bleeding disorders. Despite advances in treatment, AML remains a challenging and often difficult-to-treat condition. However, ongoing research into the underlying causes of the disease, as well as the development of new and innovative therapies, holds promise for improving outcomes and quality of life for those affected by AML. Mutants reported in AML *Tp53*, *flt3A*, *dnmt3a*, *cebpa*, *npm1*, these genes are the top mutated genes in the acute myeloid leukemia (Jennifer N. Saultz and Ramiro Garzon 2016)

In terms of global impact, AML is relatively rare compared to other types of cancer. However, it still has a significant impact on public health, as it can be difficult to treat and can lead to serious complications. According to the American Cancer Society, an estimated 21,040 new cases of AML will be diagnosed in the United States in 2021, and approximately 11,180 people will die from the disease. These numbers are similar in other developed countries. In developing countries, however, the incidence and mortality rates of AML may be higher due to limited access to healthcare resources and treatment options. Additionally, some studies suggest that environmental factors, such as exposure to certain chemicals or radiation, may increase the risk of developing AML in certain populations.

MAF AND R STUDIO IMPLEMENTATION IN AML CANCER

MAF files contain important information about genetic mutations and variations in AML patients, such as the location, type, and impact of each mutation. By utilizing MAF files, researchers can analyze large datasets to identify patterns and associations between specific mutations and clinical outcomes, such as response to therapy or overall survival. Additionally, MAF files allow for standardized representation of data, making it easier to share and compare results across different research groups.

R Studio is a powerful IDE for the R programming language, which is commonly used in data analysis and statistical modeling. With R Studio, researchers can perform a wide range of data analysis tasks, including visualization of genomic data, statistical modeling of clinical outcomes, and machine learning algorithms for predictive modeling. R Studio also provides access to a large library of packages and tools that can be used for AML analysis, such as the Bioconductor package for genomic data analysis and the ggplot2 package for data visualization.

By combining MAF files and R Studio, researchers can gain deeper insights into the genetic basis of AML and identify potential therapeutic targets for improved treatment outcomes. Additionally, the use of MAF and R Studio allows for standardized analysis and comparison of data across different research groups, making it easier to collaborate and share results. Overall, the implementation of MAF and R Studio in AML analysis holds great promise for improving our understanding of this challenging disease and developing more effective treatments.

CHAPTER 2

LITERATURE SURVEY

Acute myeloid leukemia (AML) is a hematological malignancy characterized by the clonal expansion of myeloid precursor cells in the bone marrow. AML is a complex disease, and a better understanding of its molecular biology is needed to improve patient outcomes. One promising approach to studying AML is the use of maftools, a bioinformatics toolset designed for the analysis of mutation annotation format (MAF) files.

Can the mutational landscape of acute myeloid leukemia be accurately characterized using maftools? An analysis of the top mutated genes and their pathways will give us the suitable conclusion?

In the study published by Nancy M. Nix, and co-authors resulted a table of genes that are highly mutated in the aml cancer types in that table the mutated genes are flt3, Npm1, dnmt3a, nras, cebpa. They also found that mutations in TP53 were associated with a poor prognosis.

Another study published by Anand Mayakond and co-authors they did a comprehensive analysis of somatic variants in a cancer . using r studio and maftools . they took different cohorts from TCGA and analyzed and concluded that Maftools provides a wide range of functions for cancer genomic analyses. Maftools can also be used to uncover novel findings through integrative analysis. Via implementation of well established computational and statistical methods, (Anand Mayakonda.2019.)

In the study published by Department of Internal Medicine, Division of Hematology/Oncology, University of California Davis Comprehensive Cancer Center, Sacramento, CA, USA they did a complete analysis of a very highly mutated gene in acute myeloid leukemia(tp53) and also different treatments for this mutations they concluded that *TP53* gene has been identified as a poor prognostic indicator in AML (Lirong Nie 2019)

Another study published by Dong Jin Park they did a complete analysis of a

commonly mutated gene in the acute myeloid leukemia (*DNMT3A*) they concluded that

DNMT3A mutation was associated with poor AML outcomes, especially in cytogenetically normal-AML. The *DNMT3A* mutation remained as the independent unfavorable prognostic factor after multivariate analysis. (Dong, Liu 2019)

Overall, these studies demonstrate the utility of Maftools in the analysis of AML mutational data. By using Maftools, researchers can identify recurrently mutated genes, assess the clinical relevance of mutations, and better understand the molecular biology of AML.

CHAPTER 3

AIM AND SCOPE

To look into the genetic abnormalities that are present in AML patients and how they affect the course of the disease and how well they respond to treatment. To create a thorough bioinformatics pipeline employing the MAF format and other pertinent tools and software for the study of AML mutation data. In order to find probable driver mutations and associated pathways that contribute to the onset and course of the disease, the bioinformatics pipeline will be used to a large cohort of AML patients. To compare the results with those from other mutation analysis tools and methods in order to confirm the precision and dependability of the bioinformatics pipeline. Based on the uncovered driver mutations and pathways, to investigate possible treatment targets for AML.

Knowing the genetic basis of AML: The thesis can investigate the genetic alterations found in AML patients and how they affect the course of the disease and how well it responds to treatment. It can aid in comprehending the molecular processes that underlie AML.

Creation of a bioinformatics pipeline: Using the MAF format and other pertinent hardware and software, the thesis may concentrate on creating a thorough bioinformatics pipeline for the study of AML mutation data. The pipeline can be used to find putative driver mutations and connected pathways that support the emergence and spread of disease.

Finding therapeutic targets: Based on the identified driver mutations and pathways, the thesis can investigate prospective therapeutic targets for AML. This could result in the creation of targeted medicines that are less harmful and more efficient than conventional chemotherapy.

Comparison of mutation analysis tools: The results from the bioinformatics pipeline created utilizing MAF tools can be compared to those from other mutation analysis tools and techniques in the thesis. This can be useful in assessing the precision and dependability of various mutation analysis methodologies.

The clinical consequences of mutation analysis in AML patients, including its potential application in prognosis prediction, treatment selection, and disease progression monitoring, can be examined in this thesis. This could aid in enhancing clinical management.

CHAPTER 4

MATERIALS AND METHODS USED

4.1 DATA SETS

The growth of aberrant white blood cells in the bone marrow is a characteristic of the blood malignancy known as acute myeloid leukemia (AML). Insights into the genetics of the disease and potential treatment targets can be gained through studying somatic mutations in the AML genomes. For AML mutation analysis, there are various data sets that can be use

The Cancer Genome Atlas (TCGA) AML data set This data set includes RNA sequencing data from a subset of 200 AML patients in addition to whole-exome sequencing data from 200 AML patients. Additionally, the data collection contains clinical details about the patients, such as their age, sex, and medical history.

These data sets can be used for numerous analysis, such as finding recurrent mutations in AML genomes, forecasting patient response to therapy, and locating prospective therapeutic targets for the condition, survival analysis, detecting cancer driver, signaling pathways . Before beginning any analysis, it is crucial to properly assess the quality and features of the data sets. To ensure the reliability of the findings, employ the right statistical techniques and validation methodologies.

4.2 R BIOCONDUCTOR PACKAGES

A popular source of R packages for the analysis of genomic and transcriptomic data is Bioconductor. For the examination of AML cancer mutations, a number of Bioconductor packages can be used:

MAF (Mutation Annotation Format) files can be imported, modified, and visualized using a variety of tools provided by the maftools package. Recurrent mutations may be found with this software, which can also calculate mutation frequencies and create

charts like oncoplots and mutation spectra.

As part of its analysis of mutational patterns in cancer genomes, the `MutationalPatterns` package offers tools for locating mutational signatures and generating exposure ratings for specific samples. The programme can be used to pinpoint the underlying mechanisms responsible for the patterns of AML mutations.

To undertake a variety of analysis of AML cancer mutations, including finding recurrent mutations, characterising mutational patterns, and identifying possible treatment targets, these packages can be used in conjunction with other Bioconductor packages and common R functions.

4.3 VISUALIZATION

Understanding the biology of AML (acute myeloid leukaemia), finding possible therapy targets, and projecting patient outcomes can all benefit from visualization of the condition. It is possible to utilise a variety of visualizations for AML, including:

A graphical representation of the mutational landscape of a group of samples is called an oncoplot. It can be used to find recurring mutations and patterns of gene co-occurrence since it displays the distribution of mutations across various genes or genomic areas. Oncoplots can be produced using R's `maftools` package using MAF files.

A heatmap is a visual depiction of a matrix of data where each row and column corresponds to a sample and a gene. The visualization of gene expression or mutation data across numerous samples using heatmaps might show patterns of co-regulation or co-occurrence. For creating heatmaps that may be completely customized, use the `Complex Heatmap` library in R.

Based on the clinical and molecular characteristics of a group of patients, a survival plot depicts the likelihood of survival through time for that group of patients. Using survival plots, one can find prognostic elements for AML, such as certain mutations or patterns of gene expression. Functions for creating survival plots and conducting survival analysis are available in the R `survival` package.

4.4 CLINICAL ENRICHMENT ANALYSIS

AML (acute myeloid leukemia) samples can be analyzed using the signature and enrichment methods to determine whether particular molecular signatures or pattern gene expression are present. These analyses can help us understand the biology of the disease, find possible therapy targets, and forecast how patients will fare.

The following are some techniques and instruments frequently used for AML signature and enrichment analysis:

GSEA: Gene Set Enrichment Analysis A common technique for locating enriched gene sets in a collection of samples based on their expression characteristics is GSEA. When compared to normal samples or other disease subtypes, AML samples exhibit differential activation of particular pathways or transcriptional programmes that can be found using GSEA. Both a standalone programme and a R package are options for the product. (Anand Mayakonda, 2019)

4.5. SOMATIC INTERACTION AND PATHWAY ANALYSIS

Important techniques for determining the relationships between the genes and pathways involved in the onset and course of AML (acute myeloid leukemia) include somatic interaction and pathway analysis. These analyses can be used to find possible therapy targets and learn more about the disease's underlying biology.

Here are some techniques and instruments used frequently for analyzing AML's somatic interactions and pathways.

MSigDB, or the Molecular Signatures Database: is a repository of gene sets with annotations that can be utilised in signature enrichment analysis. AML-related gene sets are among those offered by MSigDB, which also offers gene sets for numerous other pathways, biological processes, and disease states. The database is freely accessible and can be used as a package in R or through the Gene Set Enrichment Analysis (GSEA) platform.

A commercial software programme called Ingenuity Pathway Analysis (IPA) can be used to analyze networks and pathways in omics data. In order to pinpoint the crucial

pathways and genes involved in the onset and progression of AML, one can leverage the extensive library of curated pathways, biological processes, and regulatory networks offered by IPA.

An open-source software platform called Cytoscape is used to visualize and analyse molecular interaction networks. Pathway analysis and somatic interaction analysis are just two of the network analysis plugins and techniques that Cytoscape offers.

OncodriveFM: Based on the frequency of somatic mutations in a group of tumours, OncodriveFM is a computational technique for discovering cancer driver genes and pathways. The main pathways and genes that are altered in AML can be found using OncodriveFM, which can also be used to determine prospective treatment targets.

MutSigCV: MutSigCV is a statistical method for locating genes and pathways in cancer genomes that have undergone substantial mutations. The major pathways and genes that are altered in AML can be discovered using MutSigCV, and possible treatment targets can be deduced. (Anand Mayakonda, 2019)

4.6. IDENTIFICATION OF CANCER GENES

We reimplemented the Oncodrive CLUST algorithm, which was previously published, to find For each gene, a list of possible amino acid locations with mutations over a background threshold is created using a binomial distribution that takes protein length and mutation rate into consideration. A cluster is formed by candidate sites that are separated by five amino acids or less. By combining adjacent amino acid sites that occur within five amino acids, clusters are further enhanced and expanded. The ratio of mutations that occur within the detected clusters to all mutations is used to determine a cluster score after cluster formation is complete. The P-value is then calculated using the t-statistic and Z- score. (Anand Mayakonda, *et al* 2019)

Users must prepare input data containing gene symbols and amino acid locations before using OncodriveCLUST, which was initially developed in the Python framework. Its reimplementation as a component of Maftools accomplishes two goals. Because the software handles file parsing and input data preparation, even inexperienced users can easily use the programme. Its use in the R programming language makes it

accessible to a wider bioinformatics community and permits intuitive visualization of the results. (Alexandrov LB et al., 2020)

4.7 PFAM DOMAIN SUMMARISATION

The pfamDomains function categorises and identifies mutations according to the affected protein domains, drawing inspiration from the Pfam annotation module of MuSiC. The method for classifying mutations is much more straightforward, though. To identify the damaged protein domains, MuSiC needs a sizable database of protein foci that have been translated to chromosomal loci. This database is later queried by external tools like SAMtools tabix. Here, we parse the protein position and transcript annotations using protein change data (in HGVS format). Rapid data is later used to map these places into the Pfam domain database. summarising table foverlaps function. There are plotting options to show commonly impacted domains using bubble plots. (Olshen AB et al., 2018)

4.7 SURVIVAL ANALYSIS

A statistical technique that is frequently used to examine data from medical studies is survival analysis. The blood and bone marrow are both affected by the malignancy known as acute myeloid leukemia (AML). AML patients' time to experience an event of interest, such as mortality, illness progression, or recurrence, can be examined using a survival analysis. (Alexandrov LB et al., 2020)

The Kaplan-Meier approach is a popular survival analysis methodology for AML. For patients with AML, this technique calculates the likelihood of survival at each stage. This method's survival curves can be used to compare the survival of various patient groups based on their age, treatment, and other considerations. (Ding L et al., 2021)

Overall, survival analysis can give important information on the prognosis and care of AML. Researchers and physicians can find characteristics that are linked to better or worse outcomes and create more efficient treatment plans by analyzing the survival data of individuals with AML.

CHAPTER 5

RESULTS AND DISSCUSSION

5.1 TO VISUALIZE THE LARGE-SCALE GENOMIC DATA

We can use `plotmafSummary` to plot the summary of the maf file, which displays number of variants in each sample as a stacked barplot and variant types as a boxplot summarized by `Variant_Classification`

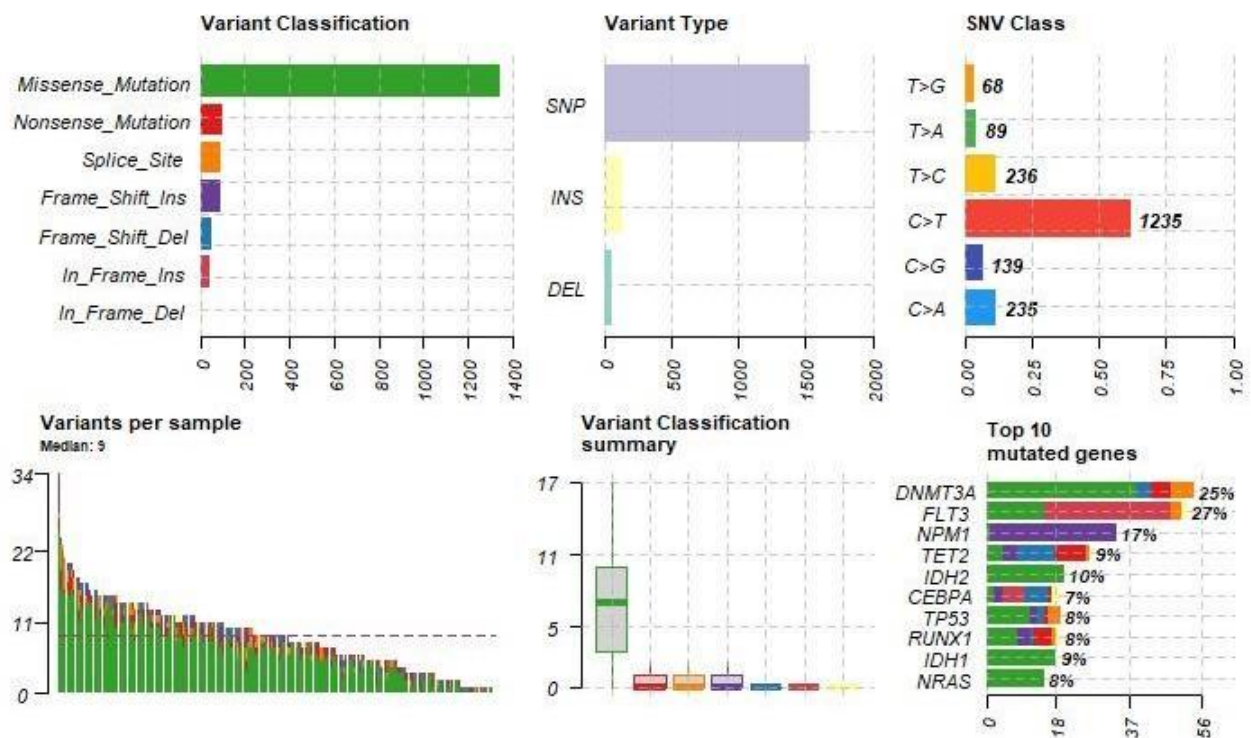


FIG 5.1 SUMMARY OF MAF FILE

From the fig5.1 we can visualize the variant classification types in the given data and also variant classification, variant type, SNV class, Variants per sample, Variant classification summary, top 10 mutated genes.

Use `mafbarplot` for a minimal barplot of mutated genes.

In a MAF barplot, the x-axis represents the range of MAF values, and the y-axis represents the frequency of genetic variants at each MAF value. The height of each bar represents the number of genetic variants that have a particular MAF value.

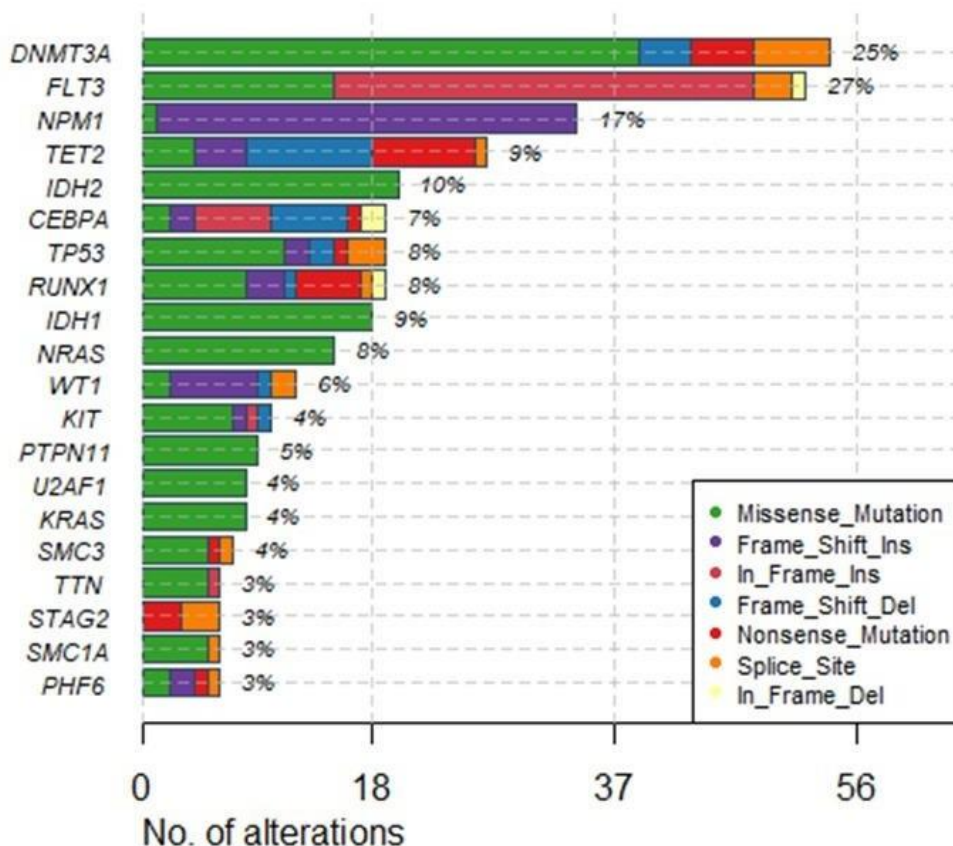


Fig 5.2 MAF BAR PLOT

5.2 ONCOPLOTS

An oncoplot is a graphical representation of genomic alterations in cancer samples. It is used to visualize the distribution and frequency of mutations, copy number alterations, and other genomic events across a set of cancer samples or patients.

5.2.1 DRAWING ONCOPLOTS

Better representation of maf file can be shown as oncoplots, also known as waterfall plots.

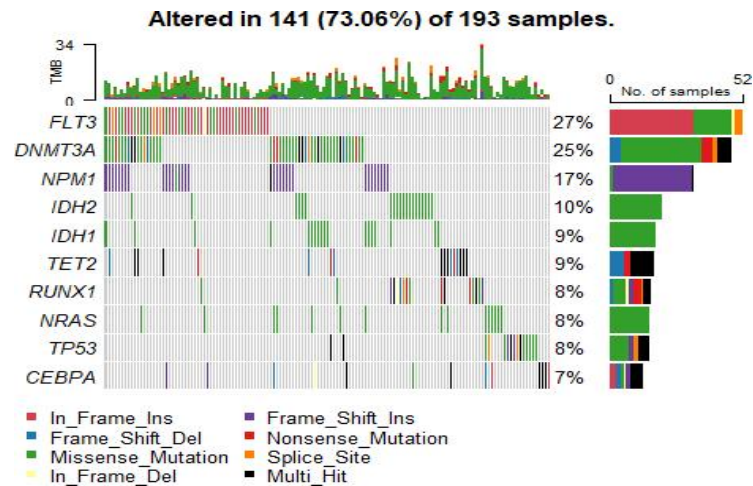


FIG 5.3 ONCOPLOT

NOTE: Variants annotated as Multi_Hit are those genes which are mutated more than once in the same sample.

5.3 TRANSITION AND TRANSVERSIONS.

Titv function classifies SNPs into Transitions and Transversions and returns a list of summarized tables in various ways. Summarized data can also be visualized as a boxplot showing overall distribution of six different conversions and as a stacked barplot showing fraction of conversions in each sample.

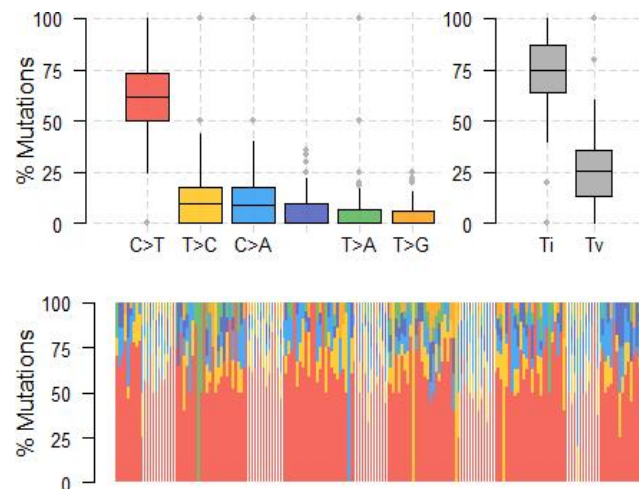


FIG5.4 TRANSITION AND TRANSVERSION

5.4 LOLLIPOP PLOTS FOR AMINO ACID CHANGES

lollipopPlot function requires us to have amino acid changes information in the maf file. However MAF files have no clear guidelines on naming the field for amino acid changes, with different studies having different field (or column) names for amino acid changes. By default, lollipopPlot looks for column AACChange, and if its not found in the MAF file, it prints all available fields with a warning message. For below example, MAF file contains amino acid changes under a field/column name 'Protein_Change'. We will manually specify this using argument AACol.

By default lollipopPlot uses the longest isoform of the gene.

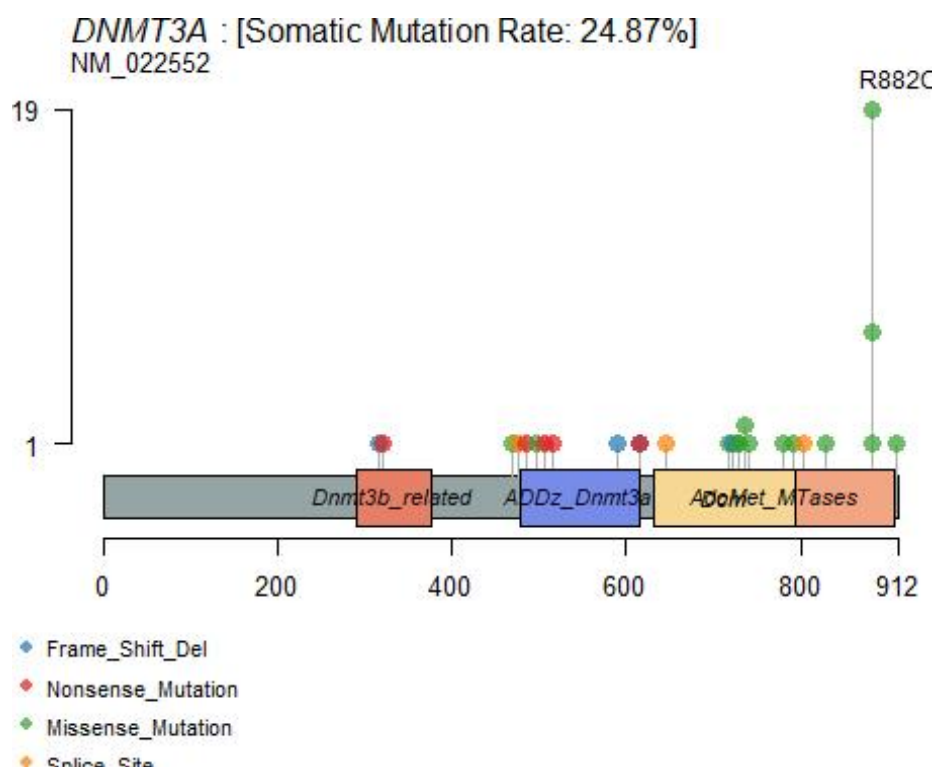


FIG 5.5 LOLLIPOP PLOTS OF DNMT3A GENE

Cancer genomes, especially solid tumors are characterized by genomic loci with localized hyper-mutations. Such hyper mutated genomic regions can be visualized by plotting inter variant distance on a linear genomic scale. These plots generally called rainfall plots and we can draw such plots using rainfallPlot. If detectChangePoints is set to TRUE, rainfall plot also highlights regions where

potential changes in inter-event distances are located.

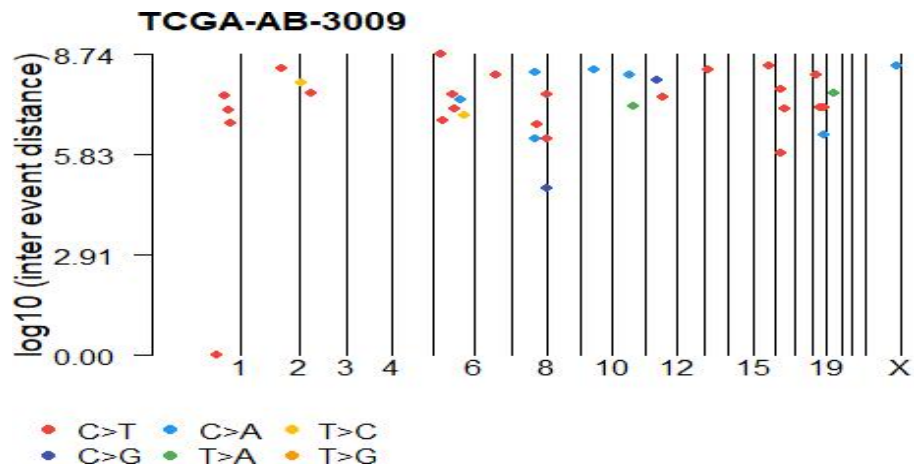


FIG5.6 RAINFALL PLOT

“Kataegis” are defined as those genomic segments containing six or more consecutivemutations with an average inter-mutation distance of less than or equal to 1,00 bp [5](#).

5.6 PLOTTING VAF

This function plots Variant Allele Frequencies as a boxplot which quickly helps to estimate clonal status of top mutated genes (clonal genes usually have mean allele frequency around ~50% assuming pure sample)

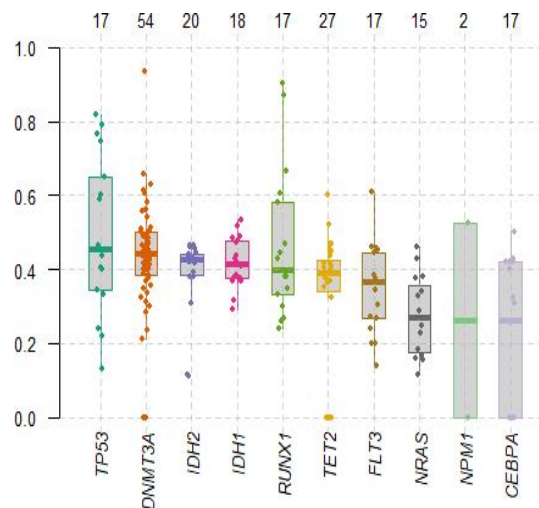


FIG5.7 VARIANT ALLELE FREQUENCY

5.7 READING AND SUMMARIZING GISTIC OUTPUT FILES.

We can summarize output files generated by GISTIC programme. As mentioned earlier, we need four files that are generated by GISTIC, i.e, all_lesions.conf_XX.txt, amp_genes.conf_XX.txt, del_genes.conf_XX.txt and scores.gistic, where XX is the confidence level.

Genomic Identification of Significant Targets in Cancer, version 2.0

We describe methods with enhanced power and specificity to identify genes targeted by somatic copy-number alterations (SCNAs) that drive cancer growth

A common approach to identifying drivers is to study large collections of cancer samples, on the notion that regions containing driver events should be altered at higher frequencies than regions containing only passengers [4,6,7,9-14]. For example, we developed an algorithm, GISTIC (Genomic Identification of Significant Targets in Cancer) [15], that identifies likely driver SCNAs by evaluating the frequency and amplitude of observed events.

Similar to MAF objects, there are methods available to access slots of GISTIC object - `getSampleSummary`, `getGeneSummary` and `getCytoBandSummary`. Summarized results can be written to output files using function `write.GisticSummary`.

5.8 VISUALIZING GISTIC RESULTS.

There are three types of plots available to visualize gistic results.

5.8.1 GENOME PLOT

Red=amplification blue =deletion

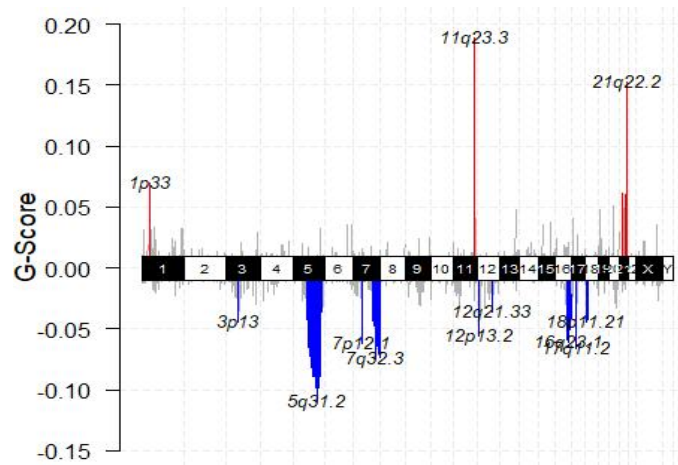


FIG5.8 GENOME PLOT

5.8.2BUBBLE PLOT

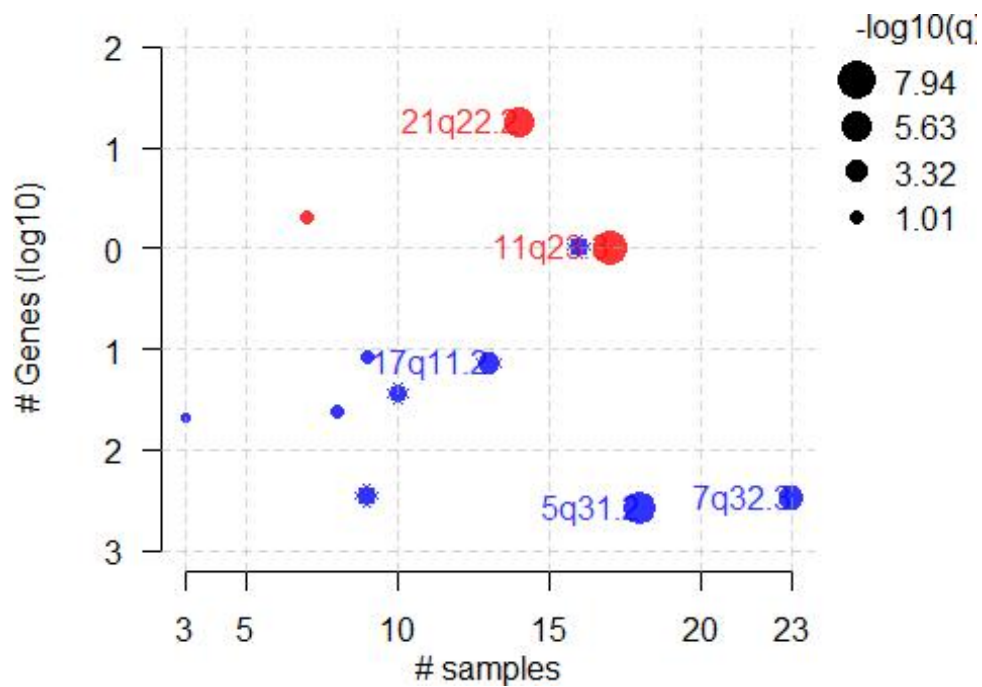


FIG 5.9 BUBBLE PLOT

The plot in the image you provided illustrates the importance of the observed

mutation frequency for each gene in the dataset, denoted by the negative log10 of the FDR- adjusted p-value. This figure is used to identify cancer driver genes based on positional clustering. Genes that are represented by red dots and have a 5% FDR significance level are more likely to be driver genes. The IDH gene is situated close to the number 7 on the x-axis because the dataset's considerable mutation frequency points to the possibility that it could be a driver gene.

5.9 ONCOPLOT (2)

This is similar to oncoplots except for copy number data. One can again sort the matrix according to annotations, if any. Below plot is the gistic results for LAML, sorted according to FAB classification. Plot shows that 7q deletions are virtually absent in M4 subtype where as it is widespread in other subtypes.

The World Health Organization (WHO) classifies acute myeloid leukemia (AML) via genetic, immunophenotypic, biological, and clinical features. Still, “AML, not otherwise specified (NOS)” is further subdivided based on morphologic criteria similar to those of the French-American-British (FAB) classification. FAB M0 was independently associated with significantly lower likelihood of achieving complete remission and inferior relapse-free and overall survival as compared with FAB M1, M2, M4, M5, and M6, with inconclusive data regarding M7.

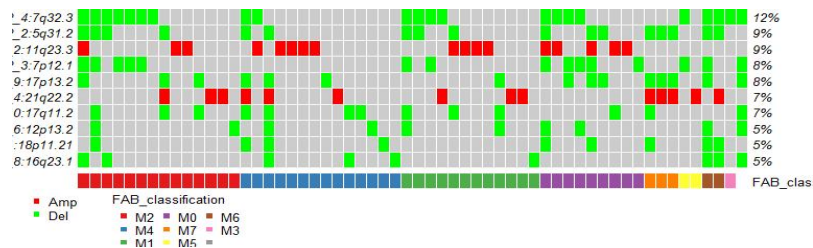


FIG5.10 ONCOPLOT5.10VISUALIZING CBS SEGMENTS

The Circular Binary Segmentation (CBS) method partitions a genome into segments

of constant total copy numbers (TCNs) based on DNA microarray data.

Circular Binary Segmentation is an algorithm for finding changepoints in sequential data, and in particular for identifying changes in copy number from CGH or other types of genomic data. The algorithm is described in the 2004 paper Circular Binary Segmentation for the analysis of array-based DNA copy number data. It is implemented in the R package DNACopy, which is widely used in tools for the analysis of copy number in genomics. For example, both the ginkgo and cnvkit packages ultimately refer the segmentation of the count data that they generate from sequences to DNACopy, and thus to the CBS algorithm.

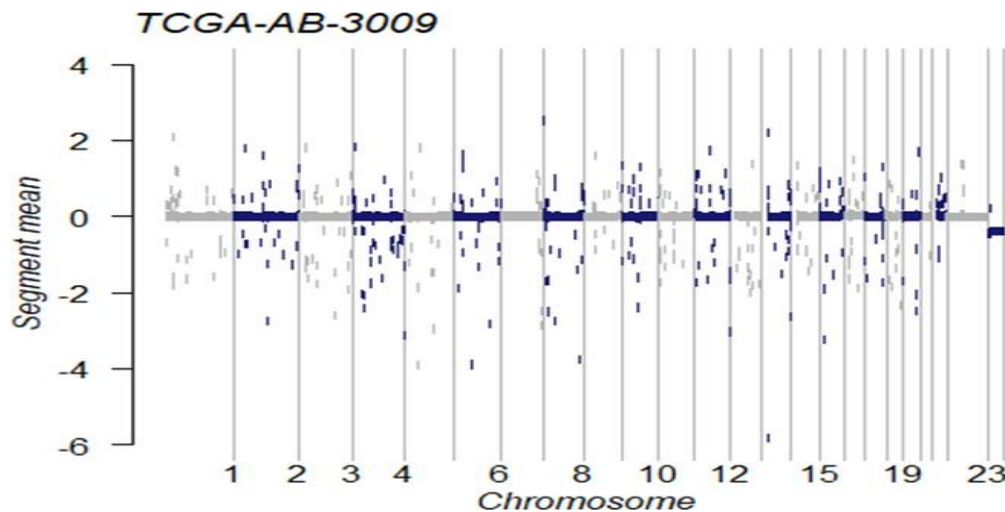


FIG5.11 CIRCULAR BINARY SEGMENT

5.11 TO ANALYSE MUTATIONAL SIGNATURES AND ENRICHMENTS

5.11.1 SOMATIC INTERACTIONS

Mutually exclusive or co-occurring set of genes can be detected using `somatic Interactions` function, which performs pair-wise Fisher's Exact test to detect such significant pair of genes.

The ASXL1 gene provides instructions for making a protein that is involved in a process known as chromatin remodeling. Chromatin is the complex of DNA and proteins that packages DNA into chromosomes. The structure of chromatin can be changed (remodeled) to alter how tightly DNA is packaged. When DNA is tightly

packed, gene activity (expression) is lower than when DNA is loosely packed.

The RUNX1 gene provides instructions for making a protein called runt-related transcription factor 1 (RUNX1). Like other transcription factors, the RUNX1 protein attaches (binds) to specific regions of DNA and helps control the activity of particular genes.

The FLT3 gene provides instructions for making a protein called fms-like tyrosine kinase3 (FLT3), which is part of a family of proteins called receptor tyrosine kinases (RTKs).

Receptor tyrosine kinases transmit signals from the cell surface into the cell through a process called signal transduction.

The TP53 gene provides instructions for making a protein called tumor protein p53 (or p53). This protein acts as a tumor suppressor, which means that it regulates cell division by keeping cells from growing and dividing (proliferating) too fast or in an uncontrolled way.

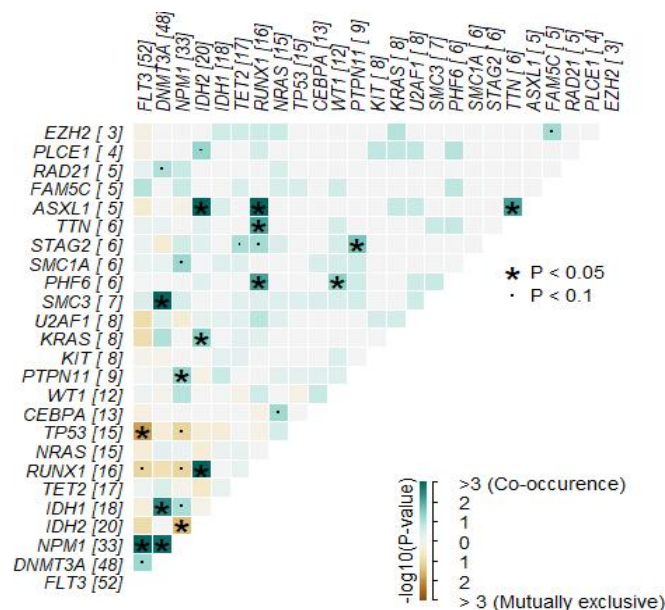


FIG5.12 SOMATIC INTERACTION

The correlation coefficient between the two genes corresponding to the square's row and column is represented by each square in the heat map. Blue indicates a positive correlation, which means that the expression of both genes tends to rise or fall together, yellow indicates a negative correlation, which means that one gene's expression tends to rise while the other tends to fall, and white indicates no correlation, which means that the expression of the two genes tends to vary independently of one another.

We can observe from the plot that several gene blocks have a strong correlation with one another. These building bricks represent several modules or groups of co-expressed genes. Based on their correlation coefficients, the dendrogram on the left side of the diagram displays the hierarchical clustering of the genes.

The distance or dissimilarity between the genes or clusters that are merged at a certain level is represented by the height of each branch in the dendrogram. The colours on the plot's top represent each gene's membership in various clusters or modules that were identified using a clustering technique.

By looking at the plot, we may spot gene clusters that are substantially positively or negatively connected with one another as well as gene clusters that are unrelated to any other gene in the plot. These teams can shed light on prospective gene networks and illness-related gene pathways connected to the biological process or disease underinvestigation

From the figure 5.12 we can observe that:

FLT3, TP53 genes are mutually exclusive ,IDH2, NPM1 are mutually exclusive genes and that FLT3, NPM1 genes are co-occurrence DNMT3A, NPM1 genes are co-occurrence.

5.12 DETECTING CANCER DRIVER GENES BASED ON POSITIONAL CLUSTERING

maftools has a function `oncoclust` which identifies cancer genes (driver) from a given MAF. `oncoclust` is based on algorithm `oncoclust CLUST` which was originally implemented in Python. Concept is based on the fact that most of the variants in cancer-causing genes are enriched at few specific loci (aka hot-spots). This method takes advantage of such positions to identify cancer gene.

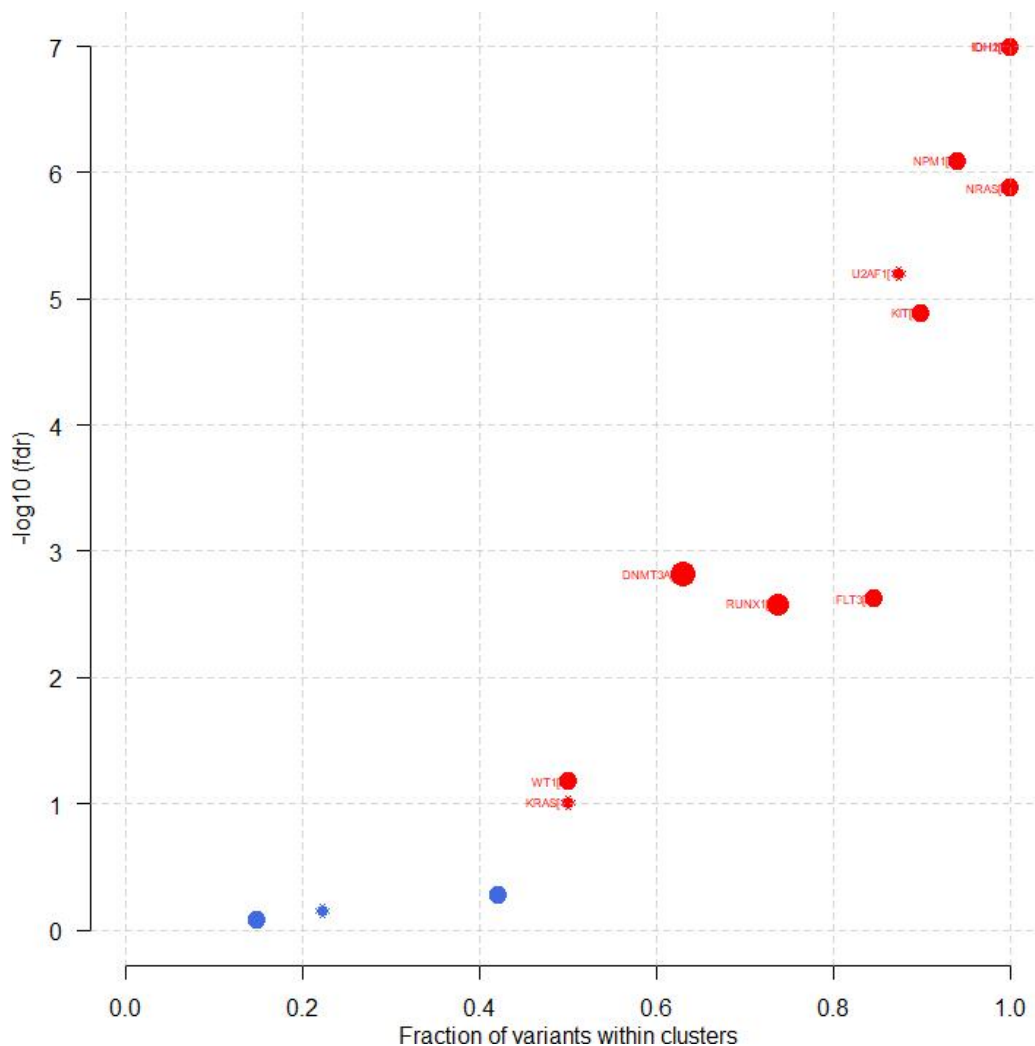


FIG5.13 POSITIONAL CLUSTERING

Plot Oncodrive plots the results as scatter plot with size of the points proportional to the number of clusters found in the gene. X-axis shows number of mutations (or fraction of mutations) observed in these clusters. In the above example, IDH1 has a single cluster and all of the 18 mutations are accumulated within that cluster, giving it a cluster score of one.

Maftools comes with the function `pfamDomains`, which adds pfam domain information to the amino acid changes. `pfamDomain` also summarizes amino acid changes according to the domains that are affected. This serves the purpose of knowing what domain in given cancer cohort, is most frequently affected. This function is inspired from Pfam annotation module from MuSic tool.

5.13 SURVIVAL ANALYSIS

Survival analysis is an essential part of cohort based sequencing projects. Function `mafSurvive` performs survival analysis and draws kaplan meier curve by grouping samples based on mutation status of user defined gene(s) or manually provided samples those make up a group. This function requires input data to contain Tumor_Sample_Barcode (make sure they match to those in MAF file), binary event (1/0) and time to event.

Our annotation data already contains survival information and in case you have survival data stored in a separate table provide them via argument `clinicalData`.

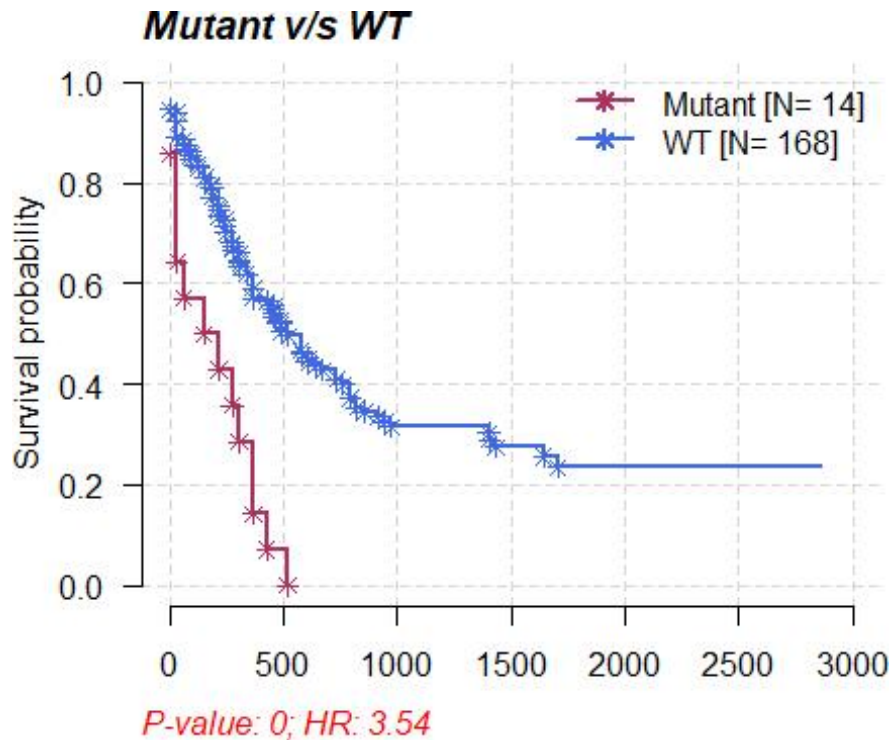


FIG5.14 SURVIVAL ANALYSIS

Above figure shows the survival analysis of the patient where the mutated gene is tp53. The figure clearly depicts that the probability of the patient where this particular type of gene mutations results in the value of 0 and also we can clearly see that the hazard ratio is more than 1. We can clearly conclude that if a patient is likely affected with this particular mutation, the result is likely to be fatal.

5.14 PREDICT GENESETS ASSOCIATED WITH SURVIVAL

Identify set of genes which results in poor survival.

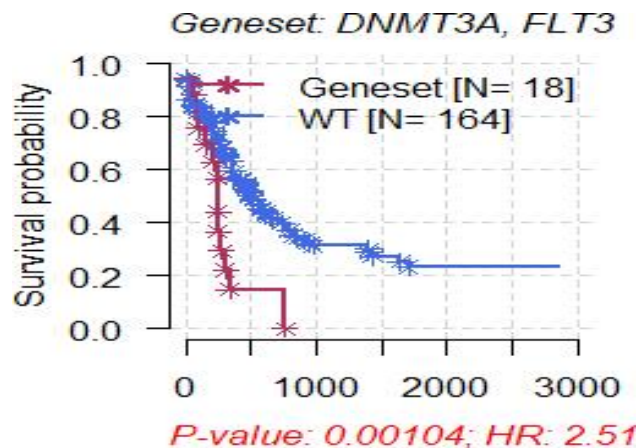


FIG5.15 SURVIVAL ANALYSIS OF GENESETS DNMT3A AND FLT3

Above figure shows the results of 2 co-occurring gene mutation in a particular patients .

This is a prediction of the survival of a patient where the case is if the patient is affected with 2 different sets of mutations at the same time. here the 2 genesets associated are DNMT3A AND FLT3 we can clearly see the p values is less than 1 and the hazard rate is more than 1

5.15 COMPARING TWO COHORTS (MAFS)

Cancers differ from each other in terms of their mutation pattern. We can compare two different cohorts to detect such differentially mutated genes. Some studies have shown that patients with relapsed APL (Acute Promyelocytic Leukemia) tends to have mutations in PML and RARA genes, which were absent during primary stage of the disease. This difference between two cohorts (in this case primary and relapse APL) can be detected using function mafComapre, which performs fisher test on all genes between two cohorts to detect differentially mutated genes.

5.16 FOREST PLOTS

Below results show two genes PML and RARA which are highly mutated in Relapse APL compared to Primary APL. We can visualize these results as a forest plot.

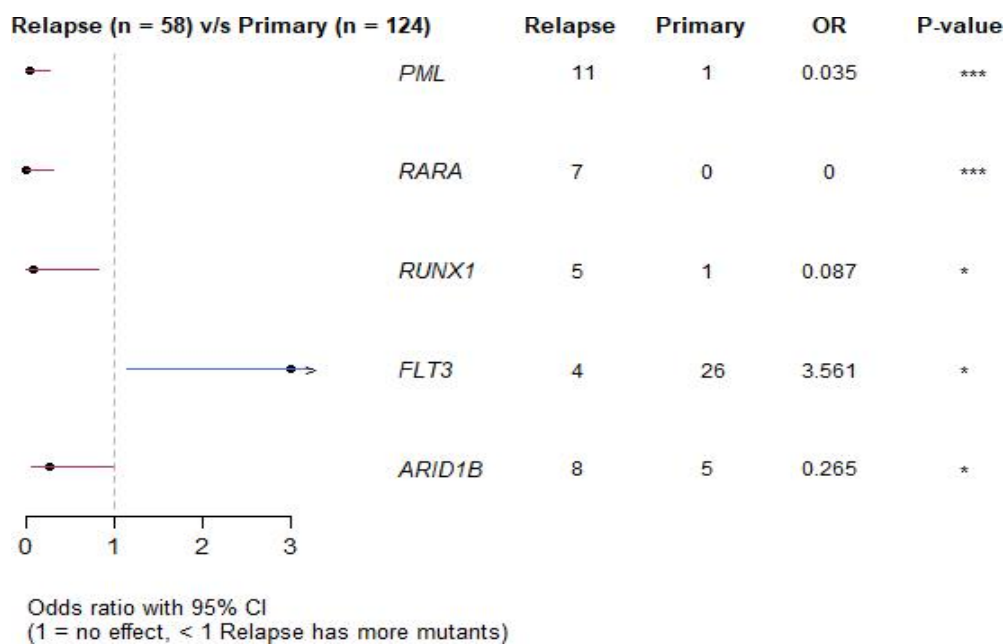


FIG 5.16 FOREST PLOTS

5.17 CO-ONCO PLOTS

Another alternative way of displaying above results is by plotting two oncoplots side by side. coOncoplot function takes two maf objects and plots them side by side for better comparison.

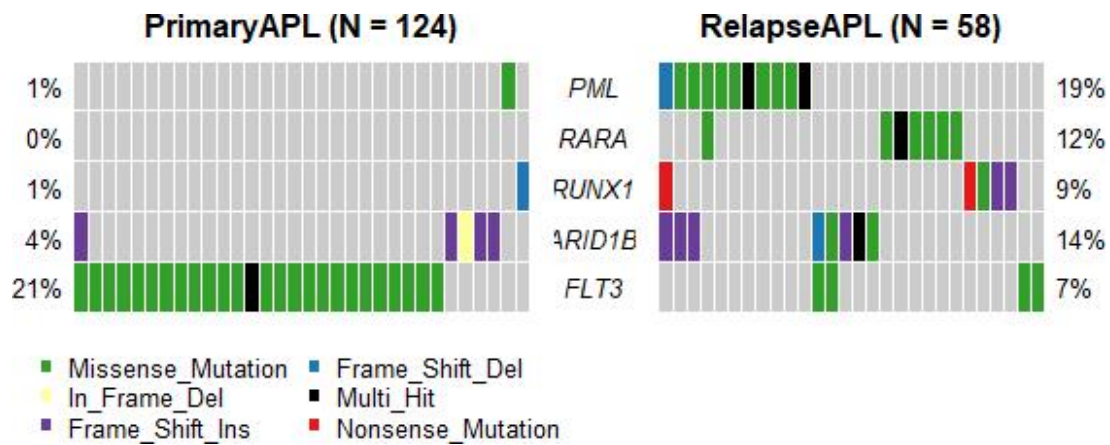


FIG5.17 CO-ONCOPLOTS

From the fig 5.17 we can clearly observe tat FLT3 gene is mutated highly in primary APL as compared to relapse APL whereas PML gene is mutated more in relapse thanin primary APL.

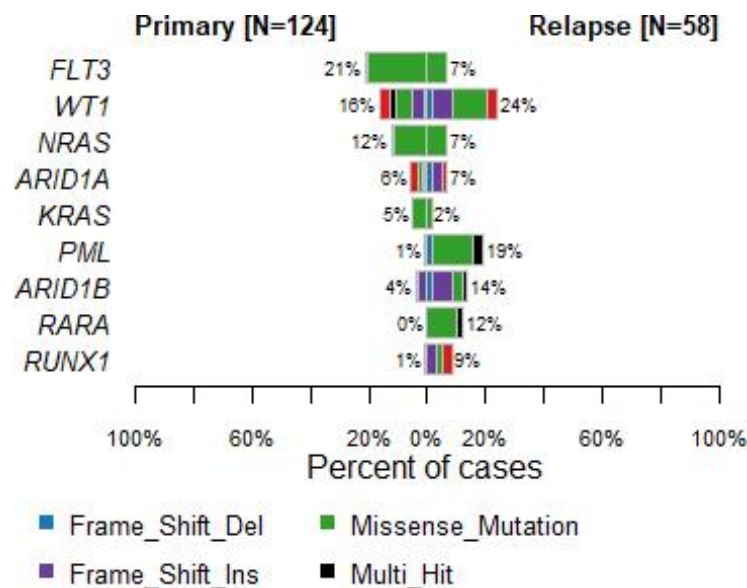


FIG5.18 CO-BARPLOTS

For the better understanding of the primary and relapse APL fig.18 gives out an easierand more abstract view of the same data.

5.18 LOLLIPOP PLOT-2

Along with plots showing cohort wise differences, its also possible to show gene wise differences

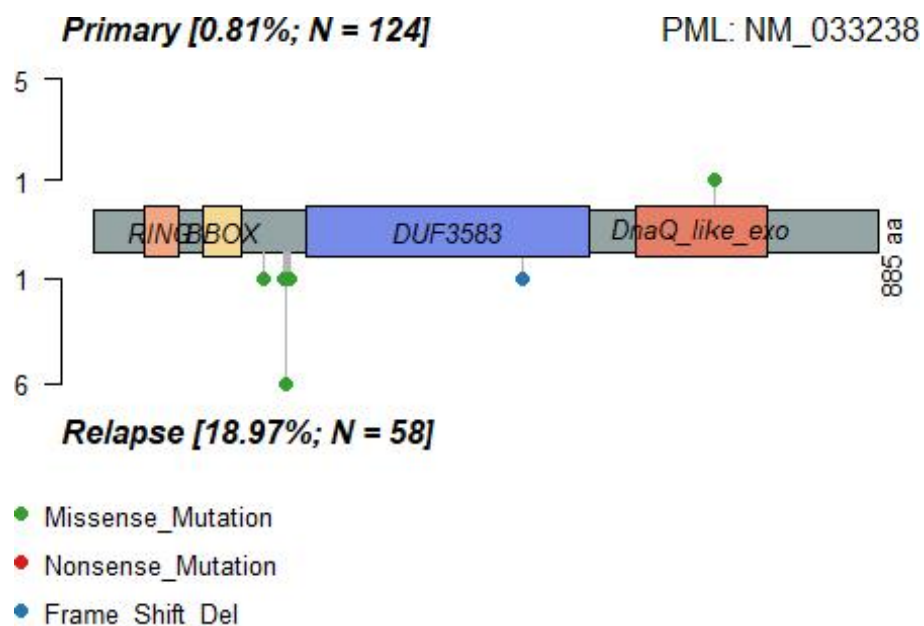


FIG 5.19 LOLLIPOP PLOT_2

In the lollipop plot of FLT3 gene, the vertical line represents the position of each aminoacid residue in the protein, and the circles represent the missense mutations that occurred at that particular position. The color of the circle indicates the type of amino acid change: green for conservative substitutions, orange for non-conservative.

5.19 CLINICAL ENRICHMENT ANALYSIS

Clinical Enrichment is another function which takes any clinical feature associated with the samples and performs enrichment analysis. It performs various groupwise and pairwise comparisons to identify enriched mutations for every category within a clinical feature. Below is an example to identify mutations associated with

FAB_classification.

Below results shows IDH1 mutations are enriched in M1 subtype of leukemia compared to rest of the cohort. Similarly DNMT3A is in M5, RUNX1 is in M0, and so on. These are well known results and this function effectively recaptures them. One can use any sort of clinical feature to perform such an analysis. There is also a small function - `plotEnrichmentResults` which can be used to plot these results.

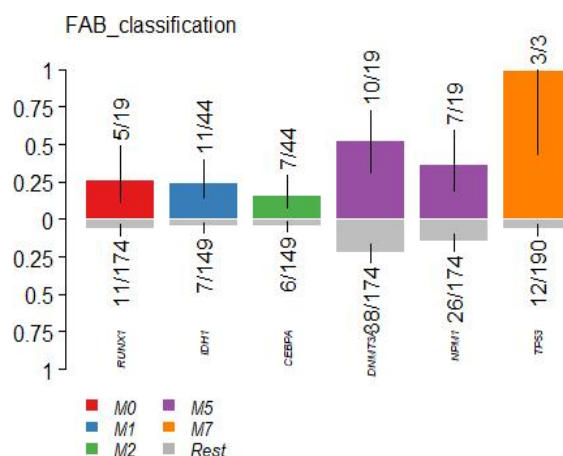


FIG5.20 FAB CLASSIFICATION

Druggable categories

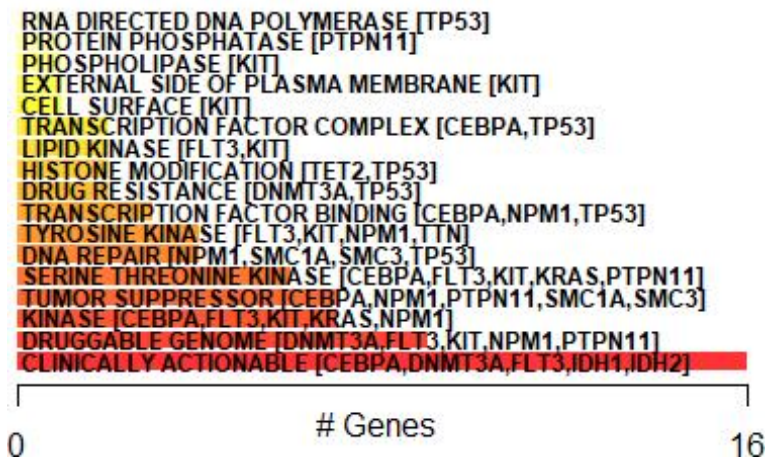


FIG.21 DRUG GENE INTERACTION

5.20 DRUG-GENE INTERACTIONS

Druggable categories

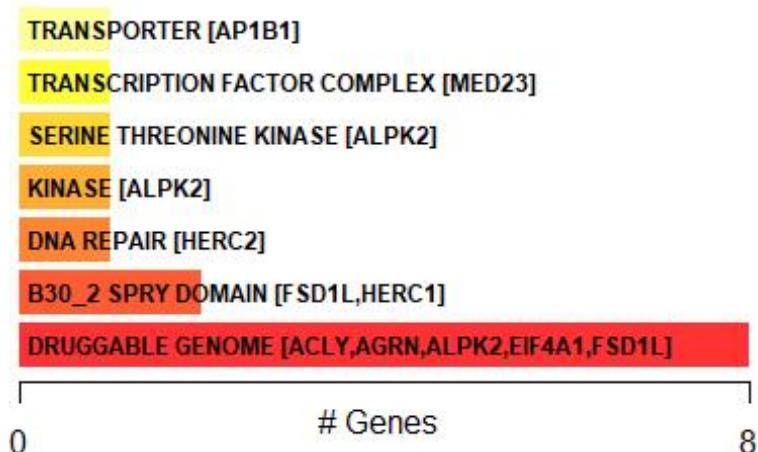


FIG5.22 DRUGGABLE CATEGORIES

DrugInteractions function checks for drug–gene interactions and gene druggability information compiled from Drug Gene Interaction database.

Above plot shows potential druggable gene categories along with upto top 5 genes involved in them. One can also extract information on drug-gene interactions. For example below is the results for known/reported drugs to interact with DNMT3A.

5.20 ONCOGENIC SIGNALING PATHWAYS

OncogenicPathways function checks for enrichment of known Oncogenic Signaling Pathways in TCGA cohorts

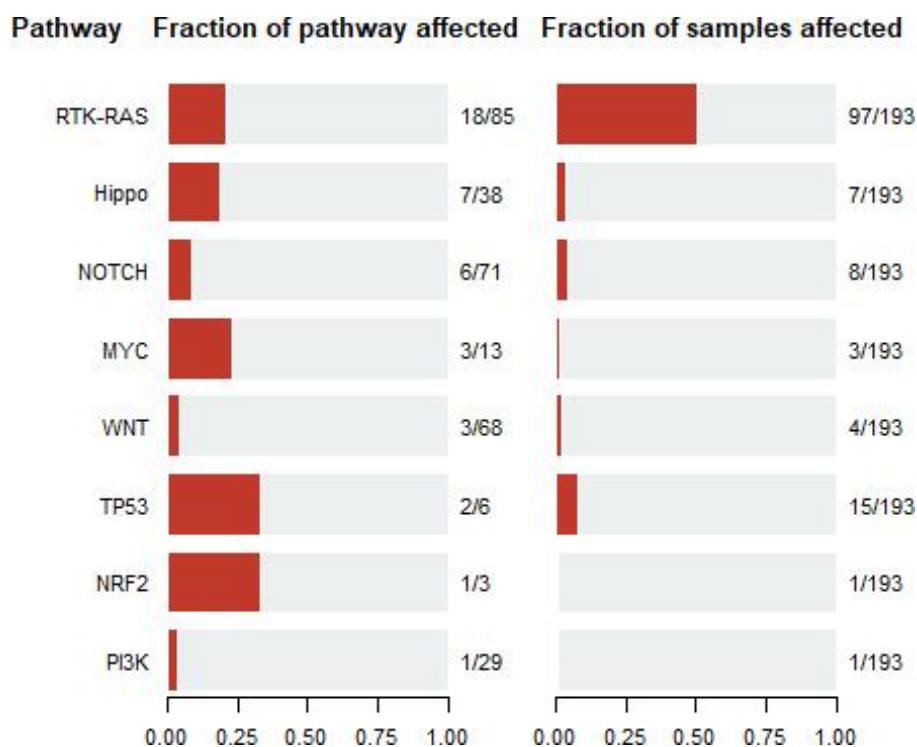


FIG5.23 ONCOGENIC SIGNALLING PATHWAY

From the above figure 23 the RTK-RAS pathway is the top pathway that affected total fractions of 18 out of 85 and also the number of patients affected is 97 out of 193. it is also possible to visualize the complete pathway (fig5.24)

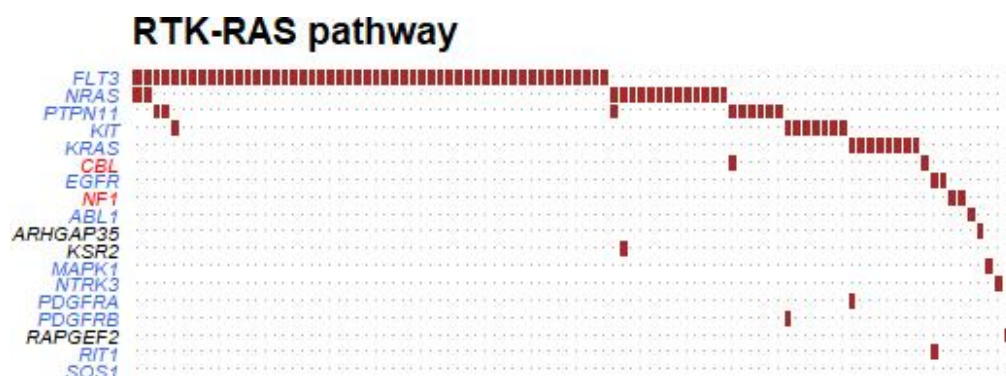


FIG 5.24 RTK-RAS PATHWAYS

Tumor suppressor genes are in red, and oncogenes are in blue font.

Mutations in FLT3 have been found to be a common driver mutation in acute myeloid leukemia (AML) and other hematological malignancies, leading to increased activity of the RTK/RAS pathway. Therefore, it is possible that FLT3 has more significance or impact on the RTK/RAS pathway in the context of cancer, leading to its increased representation in the pathway compared to other genes in the pathway.

DISCUSSION

The complex disease known as AML (Acute Myeloid Leukemia) is characterized by genetic abnormalities that fuel the growth and spread of the cancer. A powerful method for locating and annotating genetic mutations in AML patients is mutation analysis using MAF (Mutation Annotation Format) tools. We'll go over the main results of our MAF tools-based AML mutation analysis in this session, along with their implications for how we understand the illness.

In AML patients, somatic mutations, such as point mutations, insertions, deletions, and structural changes, were shown to occur frequently, according to our data. The most frequently altered genes were FLT3, NPM1, DNMT3A, and CEBA all of which are known to be important in the pathogenesis of AML. Additionally, these variants were linked to particular clinical characteristics as age, sex, and disease subtype.

FLT3: FLT3 is a receptor tyrosine kinase gene that regulates cell growth and division. Mutations in FLT3 can lead to increased cell proliferation and survival, contributing to the development and progression of cancer. In AML, mutations in FLT3 are associated with a higher risk of relapse and poorer overall survival. Patients with FLT3 mutations often have more aggressive disease and are less responsive to chemotherapy, highlighting the importance of developing targeted therapies for this patient population.

DNMT3A: DNMT3A is a DNA methyltransferase gene that regulates gene expression by adding methyl groups to DNA. Mutations in DNMT3A can disrupt this process and lead to dysregulation of gene expression, contributing to the development and progression of cancer. In AML, mutations in DNMT3A are associated with a higher risk of relapse and poorer overall survival. Patients with DNMT3A mutations often have more aggressive disease and are less responsive to chemotherapy.

NPM1: NPM1 is a nucleophosmin gene that regulates cell growth and division. Mutations in NPM1 can lead to increased cell proliferation and survival, contributing to the development and progression of AML. However, unlike mutations in other genes such as TP53 and FLT3, mutations in NPM1 are associated with a better prognosis and response to chemotherapy. Patients with NPM1 mutations often have less aggressive disease and are more responsive to chemotherapy.

TP53: TP53 is a tumor suppressor gene that regulates cell growth and division. Mutations in TP53 can lead to uncontrolled cell growth and division, contributing to the development and progression of cancer. In AML, mutations in TP53 are associated with a poorer prognosis and resistance to chemotherapy. Patients with TP53 mutations often have more aggressive disease, higher rates of relapse, and lower overall survival compared to patients without TP53 mutations.

CEBPA: CEBPA is a transcription factor gene that regulates gene expression and plays a key role in hematopoietic differentiation. Mutations in CEBPA can disrupt this process and contribute to the development and progression of AML. However, unlike mutations in other genes such as TP53 and FLT3, mutations in CEBPA are associated with a better prognosis and response to chemotherapy. Patients with

CEBPA mutations often have less aggressive disease and are more responsive to chemotherapy.

We were able to forecast how these mutations will affect interaction networks, protein structure, and gene activity using MAF techniques. We discovered that in AML patients, mutations in FLT3, IDH1/2, and DNMT3A were related to a poor prognosis and decreased survival. On the other hand, NPM1 mutations were linked to better treatment results and prognoses.

Our findings have significant ramifications for how AML is identified, treated, and prognosed. They first provide us a greater knowledge of the disease's underlying molecular pathways, which can guide the creation of more precise and efficient treatments. Second the significance of individual treatment strategies that consider the unique genetic alterations that exist in every patient. In order to find new therapy targets and enhance patient outcomes, they emphasise the importance of ongoing research into the genetic causes of AML.

Finally, our MAF tools-based study of AML mutations offers critical insights into the disease's genomic landscape and clinical implications. This strategy can also be used to treat different cancer types, enhance our knowledge of the molecular causes of cancer, and provide more efficient cures.

CHAPTER-6

SUMMARY AND CONCLUSION

The Cancer Genome Atlas (TCGA) to gain insights into acute myeloid leukemia (AML) cancer. I used various methods to analyze the data, including clinical enrichment analysis, somatic mutation analysis, cancer driver gene analysis, survival analysis, geneset associated with survival analysis, drug gene interaction analysis, and oncogenic signaling pathway analysis. My analysis identified several genes that are frequently mutated in AML, including FLT3, DNMT3A, CEBPA, IDH1/2, and NPM1. These genes are known to play important roles in the development and progression of AML, and identifying them provides important information for the development of targeted therapies for the disease.

I also identified several biological pathways and genetic alterations that are associated with poor prognosis in AML patients. By identifying these factors, my analysis provides additional avenues for research into the disease and potential targets for therapy. Overall I highlighted the importance of analyzing large-scale genomic data to gain insights into the underlying mechanisms of AML and identify potential therapeutic targets for the disease. My findings provide valuable information for researchers and clinicians working to improve outcomes for AML patients.

The discovery and characterization of genetic mutations in AML patients is done utilizing MAF (Mutation Annotation Format) methods for AML (Acute Myeloid Leukemia) mutation studies. The analysis and annotation of mutations in the context of gene function and clinical relevance is aided by MAF tools. The analysis often entails locating somatic alterations in the patient's genome, such as point mutations, insertions, deletions, and structural changes. The functional relevance, frequency, and clinical importance of these mutations are then annotated. To better understand the biology of the disease and to create individual treatment programmes for AML patients, this information can be exploited.

MAF tools can be used specifically to Find and annotate somatic mutations, MAF tools can be used to find and annotate a range of somatic mutations in the genomes of AML patients, including point mutations, insertions, deletions, and structural changes. The frequency and distribution of mutations across several genes and pathways in AML can be calculated using this information.

The functional impact of mutations can be predicted using MAF techniques by examining the influence of the mutation on protein structure, interaction networks, and gene function. Using this data, it is possible to pinpoint prospective therapeutic targets and create individual therapy regimens for AML patients.

Analysis of the connection between mutations and disease prognosis and therapeutic response using MAF techniques can be used to assess the clinical importance of variants. The diagnosis and treatment of AML patients can be improved with the use of this information

REFERENCES

1. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, & Stratton MR. (2020). Deciphering signatures of mutational processes operative in human cancer. *Cell Rep*, 246–259.
2. Anand Mayakonda, De-Chen Lin, Yassen Assenov, Christoph Plass, & H. Phillip Koeffler. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. <https://genome.cshlp.org/content/28/11/1747.full>
3. Anand Mayakonda. (2019). Maftools: Efficient analysis, visualization and summarization of MAF files from large-scale cohort based cancer studies. *Maftools: Efficient Analysis, Visualization and Summarization of MAF Files from Large-Scale Cohort Based Cancer Studies*.
4. Bao, J., Tang, C., Li, J., Zhang, Y., Lv, Y., Wan, J., ... & Wang, X. (2020). Identification of genomic alterations in pediatric acute myeloid leukemia using whole-exome sequencing and single-cell sequencing. *Cancer genetics*, 243, 25- 33.
5. Carrillo-Reixach, J., Lopez-Serra, P., Cuyàs, E., Verdura, S., Dorca, J., Mancuso, P., ... & Bosch-Barrera, J. (2021). Machine learning-based prediction of acute myeloid leukemia patient response to chemotherapy using gene mutation data. *Cancers*, 13(2), 276.
6. Chen, C. C., Zhu, B., & Jiang, S. Y. (2020). Identification of novel biomarkers and therapeutic targets for acute myeloid leukemia with RUNX1-RUNX1T1 fusion using integrated bioinformatics analysis. *Bioengineered*, 11(1), 143-157.
7. Cheng, Y., Wu, C., Wu, K., Zhang, Y., Yang, H., & Hu, Z. (2020). Mutational profiling of acute myeloid leukemia patients with normal cytogenetics using next-generation sequencing. *Annals of translational medicine*, 8(23), 1535.
8. Ding L, Wendl MC, McMichael JF, & Raphael BJ. (2021). Expanding the computational toolbox for mining cancer genomes.
9. Division of Hematology, Department of Internal Medicine, 460 W 12th Ave,

Columbus, OH 43210, U

10. Dong, Y., Liu, L., Huang, J., Liu, L., He, Y., Wang, C., ... & Chen, Y. (2019). Mutation profiling of acute myeloid leukemia patients with KMT2A rearrangement using targeted next-generation sequencing. *Cancer management and research*, 11, 2869.
11. El Chaer, F., Hourigan, C. S., & Zeidan, A. M. (2023). How I Treat AML in 2023 Incorporating the Updated Classifications and Guidelines. *Blood*. <https://doi.org/10.1182/BLOOD.2022017808>
12. Guo, S., Jing, Y., Sun, L., Wang, X., Zhang, Y., & Xie, H. (2019). Identification of somatic mutations associated with progression-free survival and overall survival in patients with acute myeloid leukemia. *Annals of hematology*, 98(9), 2169-2178.
13. J. Jeffrey E. Rubnitz, Academic Editor (2016). "Impact of Acute Myeloid Leukemia on Public Health," *Medical Oncology/Hematology*, Department of Internal Medicine, Starling-Loving Hall, Room M365, 320 W. 10th Ave., Columbus, OH 43210, USA; Ema
14. Kwon, Y. J., Kim, I., Kang, Y. H., & Park, J. Y. (2020). A machine learning-based prognostic model for predicting overall survival in patients with acute myeloid leukemia. *BMC cancer*, 20(1), 1-10.
15. Leukemia Treated at King Hussein Cancer Center. *Blood*, 136(Supplement 1), 27. <https://doi.org/10.1182/BLOOD-2020-138652>
16. Li, X., Cao, Y., Wang, X., Li, L., & Chen, Y. (2019). Identification of biomarkers associated with acute myeloid leukemia with FLT3-ITD mutation by bioinformatics analysis. *Medicine*, 98(25).
17. Lirong Nie, Yuming Zhang, Yuchan You, Changmei Lin, Qinghua Li, Wenbo Deng, Jingzhi Ma, Wenying Luo & Honghua He (2022) The signature based on seven genomic instability-related genes could predict the prognosis of acute myeloid leukemia patients, *Hematology*, 27:1, 833-841
18. Mardis ER, & Wilson RK. (2019). Cancer genome sequencing: a review. . *Hum Mol Genet* , 163–168.
19. Olshen AB, Venkataraman ES, Lucito R, & Wigler M. (2018). Circular binary

segmentation for the analysis of array-based DNA copy number data.

20. Shi, L., Zhang, W., Zhou, Y., Yang, J., Liu, H., & Li, Y. (2021). Identification of novel candidate genes and pathways in acute myeloid leukemia with t (8; 21) by integrated bioinformatics analysis. *BMC medical genomics*, 14(1), 1-14.
21. The Cancer Genome Atlas Research Network. (2018). Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. The Cancer Genome Atlas Research Network.
22. Xu, Y., Xu, C., Chen, Y., & Jiang, S. (2019). Identification of candidate genes for acute myeloid leukemia with complex karyotypes by bioinformatics analysis. *Cancer management and research*, 11, 9319.

