NCBI DATASET ON T-ALL CANCER

Exploring the dataset:

GSE63602 is a dataset from the national s center for biotechnology information (NCBI)

*. The dataset is about a disease called T-cell acute lymphoblastic leukemia (T-ALL)

T-ALL : It is a type of cancer of the blood & bone marrow the spongy tissue inside bones where blood cells are made.

*. This develops from white blood cells which is called lymphoblasts

*. The primary cause of the disease is unknown but reports suggest that it is caused by genetic mutation.

How was the dataset created:

The researchers brought a total of 9 T-ALL patients & 4 normal persons. They collected the blood & bone marrow samples from each patient.

From the collected cells, scientists extract total RNA (The RNA includes mRNA, miRNA) these were then sent to a high-throughput sequencing & finally we got the dataset.

There are in total of 14,000 rows & 7 columns in the dataset.

Each row is a id & kind of the RNA that is present).

(3) (4)

What can be found inside the dataset.

i) ID: These are code names for the miRNA's. Each identifies which miRNA's the rows represent.

ii) Basemean: This is the average normalized expression of the miRNA across all samples.

iii) Base mean A: Average normalized expression of a miRNA in group A.

iv) Base mean B: Average normalized expression of a miRNA in group B.

v) Foldchange: The ratio of expression between two groups (often T-ALL vs normal)

$$\text{foldchange} = \frac{\text{base mean B}}{\text{base mean A}}$$

vi) log2fold change: Logarithmic base 2 of foldchange
$\log_2 FC = \log_2(\text{foldchange})$.

vii) Pvalue: Probability that the observed difference happend by chance, assuming there is no real difference between groups.

viii) Padj: corrected P.value for multiple testing using FDR.

If log2 foldchange closer to 1 (the cells are more active) closer to -1 (the cells are less active)

Padj tells if the change is caused by the gene activity or is it just a random noise

Some insightful research paper level questions that can be solved with this dataset.

i) Can we build a model to distinguish T-ALL samples from healthy control samples based on gene or miRNA expression profiles?

ii) Are there molecular subtypes of T-ALL that can be identified using unsupervised ML techniques..

How ML models have actual impact on predicting diseases.

* There are few challenges that are faced by ML models
  i) Heterogenous datasets
  ii) Class imbalance
  iii) Scalability

Project continuation:

1. Types of Questions:

"Predicting Leukemia subtypes & Discovering Bio marker miRNA's using ML on RNA-Seq Data".

Therefore Resume:

"Developed an end-to-end ML pipeline to classify T-cell leukemia using RNA-Seq expression data, identified candidate microRNA biomarker regulating pathways

i) Identifying key RNA Biomarkers linked to leukmia.
ii) Distinguish cancer vs normal Expression Signatures.
iii) Link Dysregulated RNA's to cancer Pathways
iv) Construct RNA co-expression networks.
v) Develop a minimal Biomarker Signature Panel.

STEPS:

1. Data Acquisition & quality control.
2. Exploratory Data Analysis
3. Feature Selection & Statistical Testing
4. Build Interpretable models.
5. Biological interpretation.
6. Validation & Visualization
7. Documentation & Sharing.

1. DATA ACQUISITION & QUALITY CONTROL:

There are totally 14000 rows in the file. In these 14000 rows there is a lot of unwanted genes which has to be removed based on certain conditions. The conditions that I used to remove the rows are:

i) If Padj > 0.1 - we should remove these because these are not statistically significant.

ii) Filtering genes with baseMean <10, Why do we have to do that.

*. Genes with very low counts produce unstable dispersion estimates which can distort ML results.

This 2 are the main methods that I used to remove the data from the dataset.

2. Exploratory DATA Analysis:

EDA gives us a sense of how the RNA expression data behave across samples & can reveal biological structure, batch effects, or outliers before you start ML modeling.

i) Check Expression Distributions Across samples

Best EDA practices for Biological Data:

A) Summary Statistics
B) Data Transformation & Normalization
C) Visualizing Distributions
D) Sample clustering
E) Gene correlation
F) outlier detection
G) Functional Annotation.

What is Box Plot?

Box Plots are used in statistics to graphically display various parameters at a glance. In a boxplot the median, the interquartile range and the outliers can be read.

Concept of upregulation & downregulation.

If the log_2Foldchange > 0 - the gene is more active in T-ALL (up regulation)

If the log_2Foldchange < 0 - the gene is less active in T-ALL (down regulation)

Using Interquartile Range:

Formula:

Q1 = sig_genes['log_2Foldchange'].quantile(0.25)

Q2 = sig_genes['log_2Foldchange'].quantile(0.75)

IQR = Q3 - Q1

upperbound = Q3 + 1.5 * IQR

upper_outliers = sig_genes[sig_genes['log_2Foldchange'] > upper_bound]

The above formula gives the Top 10 genes that are responsible for the cancer.

Why volcano Plot?

We can use volcano plot to identify the genes which are very highly significant & also probabilitially true. We take two columns for comparison - log10 Padj & log2Foldchange.

(X) If the log2FoldChange is above 0, that means a gene is upregulated in leukemia samples compared to control

From the volcanoPlot we can find the cells which are more active in control & cells which are more active in Patient, & also what is the probability of these cells are showing actual difference rather than being a noise.

Next we can move with PCA:

What is PCA & why?

PCA is a powerful dimensionality reduction technique widely used in bioinformatics.

But PCA cannot be done in this dataset because unlike traditional dataset where we have a lot of samples here in this dataset we do not have those samples. So, I have decided to SKIP PCA.

In the future works that I do I would love to use PCA, because PCA gives a really good visualisation for a large amount of samples.

10

Heatmaps:

A heatmap is a visual way to display how strongly genes are expressed across samples or conditions.

What we can tell is
1. which genes are upregulated & which genes are down regulated.
2. visual grouping or clustering
3. Magnitude of change

MA Plot:

MA plot is a scatter plot which is commonly used in gene expression analysis to visualize the relationship between the average expression level and the magnitude of expression change between two conditions.

MA plots visualize the relationship between log2 foldchange (M values) & mean of normalized counts (A values) for differential expression analysis results.

NEXT IS FUNCTIONAL ENRICHMENT & PATHWAY ANALYSIS:

Gene ontology (GO) & other Enrichment analysis.

It is a Biological framework used to determine (or) describe the functions of genes and their products in a consistent, structured, & hierarchical way across different species.

Gene ontology provides a standardised vocabulary to describe gene function. There are 3 main ontology categories:

i) Biological Process
ii) Molecular function
iii) Cellular component.

Gene ontology is a Pre-built framework.

What in general Functional Enrichment analysis & Pathway analysis mean:

Functional enrichment & Pathway analysis are very critical for interpreting which Biological Process, molecular functions, or signaling pathways are affected.

By performing this I have found out that in GO-cellular component-2023 collagen-containing Extracellular matrix.

The collagen-containing extracellular matrix has really a lot of genes value representation. A statistically significant portion of the genes in the filtered list (the ones that we selected based on p-value & foldchange) are annotated to the "collagen-containing Extracellular matrix".

KEGG Pathway enrichment:

KEGG is a database of manually curated pathways, covering metabolism, cell signaling, disease mechanisms & more.

RESULTS THAT I GOT FROM THE FOLLOWING

1. GO RESULTS:

i) GO-Biological Process-2023 → collagen-containing Extracellular matrix

ii) GO-Cellular-component-2023 → collagen-containing Extracellular matrix

iii) GO-Molecular-Function-2023 → collagen-containing Extracellular Matrix

2. KEGG-2021_Human
   i) Extracellular matrix organization

3. Reactome-2022
   ii) Extracellular matrix organization.

What can we do next?

What I did is that I have downloaded a CSV file which has the gene list which are common in my CSV file & enrichGO Database, KEGG Database

i) Statistical strength
ii) Separate upregulated Vs Downregulated genes
iii) Functional Network/Pathway Visualization.
iv) Validate Biological relevance.
v) ML/Predictive modeling.

TOOLS TO USE:

*. Cytoscape.
*. string-db for Protein-Protein interaction
*. Cluster Profiler/Enrichr network plots.

Now that I have the genes which are specific to ECM.

Based on the GO Interpretation most of the genes belong to ECM category & all the remaining genes are in the category of granule & there are lot of granule types.

Plan:

Phase 1:

i) There are a total of 89 Granule types & some genes are repeating in granule types. So the first process is to find all the unique Granule genes & keep them as one master Granule list.

ii) If there is a particular granule gene which appears 5 times then we can note that down.

iii) We have to determine if the granule genes are upregulated or downregulated.

Phase 2:

In Phase 2 we will dive into the visualization of the files, here are the files which are going to be used for visualization

i) matched_ecm_genes: This is the file which have ecm-genes out of the 4k genes

ii) Granule unique-genes - This file contains the unique granule genes out of 9 granule file.

I have performed the visualization & here are the results that I found.

Image 1: T-ALL vs Normal.Png

This is a heatmap representation of the genes which are downregulated & upregulated.

Yellow (4.0): very low expression.
Light Green (8.0): Moderate expression
Dark Blue (14.0): Extremely high expression

Ex: COL3A1 (The top):

T-ALL → 14.7 (Dark Blue)
Normal → 10.4 (medium Blue)

This is a massive upregulation. The leukemia cells are producing huge amounts of Type 3 collagen.

ELANE (The Enzyme):

T-ALL -> 7.6 (Green)
Normal -> 12.7 (Dark Blue).

This gene is Downregulated. In a healthy person, ELANE is active to Protect the body.

COL4A2 (The structural switch):

T-ALL -> 12.2 (Blue)
Normal -> 4.4 (Yellow)

In healthy people, the gene is nearly silent (Yellow). In T-ALL, it "wakes up" and becomes highly active. This is a Perfect Biomarker.

Image 2: clustermap_granule_comp.png

0.0 (Yellow): gene at normal avg level
positive (Red): The gene is "high"
Negative (Blue): The gene is "Low"

The Bottom red block: These are the genes are the "Leukemia Team". They are always "ON" in T-ALL & always "OFF" in normal.

The TOP Blue Block, These are usually "Healthy Team". The leukemia has actively silenced these genes. Since these genes (like MMP8 usually destroy collagen, silencing them allows the "Red Block" to grow out of control.