The Beat AML Master Clinical Trial is a groundbreaking precision medicine approach led by the Leukemia & Lymphoma Society (LLS) to improve treatment for acute myeloid leukemia (AML) patients. The trial uses advanced technology to examine the genetic makeup of each patient's cancer and then matches them to the most promising targeted treatment.

This study presents a dataset of 805 Acute Myeloid Leukemia (AML) patients, which was harmonized with a previous report to yield a cumulative cohort of 942 specimens. The study showed strong cross-cohort concordance and identified features of drug response in AML patients. The transcriptomic data was deconvoluted to show that drug sensitivity is broadly governed by AML cell differentiation state and sometimes affects other response correlates. The study also found that a single gene, PEAR1, is a strong predictor of patient survival, especially for young patients. This report expands the existing functional genomic resource for AML and provides avenues for further exploration and drug development. The methods used in this study include:

* Ex vivo drug sensitivity testing
* Genomic and transcriptomic analysis
* Clinical annotations of the patients
* Harmonization of the dataset with a previous report
* Cross-cohort concordance analysis
* Deconvoluting transcriptomic data
* Modeling of clinical outcome

It combined multiple methods to expand the existing functional genomic resource for AML and provide avenues for further exploration and drug development.

The study highlights the importance of comprehensive analysis in drug sensitivity testing using large patient specimens. This includes understanding the connection between drug sensitivity and tumor cell differentiation, as well as the relationship between drug/mutation and cell type. Assessing cell type is crucial in evaluating drug sensitivity differences in clinical settings and considering interactions between cell type and other disease or patient features can help identify factors that may affect drug response. The analysis also revealed a promising prognostic and potentially targetable feature in AML that requires further investigation.

The interactions referred to in the study are between cell type and other disease or patient features. These interactions can help identify potential confounders or modifiers of drug response, meaning they may have an impact on the relationship between cell type and drug sensitivity. By considering these interactions, the study suggests that it is possible to gain a more complete understanding of drug sensitivity in clinical settings.

Confounders are factors that can distort the true relationship between an exposure (e.g. cell type) and an outcome (e.g. drug sensitivity). Confounders are often related to both the exposure and the outcome and can bias the results of an analysis if not properly accounted for. In the context of the study, confounders could be other disease or patient characteristics that are related to both cell type and drug sensitivity, and if not controlled for, could lead to incorrect conclusions about the relationship between cell type and drug sensitivity. By considering interactions between cell type and other disease or patient features, the study aims to identify potential confounders that may affect drug response.

**#Tasks for Siddhi**

**Material and Methods section points:**

0. Where the data was downloaded from and how it was processed. See the ‘analyze\_cell\_lines\_rnaseq.R’ script for this.

1. Names of all samples in clinical data from Wave 1+2 and Wave 3+4 match with the names of all samples in gene expression data from Wave 1+2 and Wave 3+4, otherwise remove samples. Need numbers for these !

1. What are the different clinical traits for the samples in Wave 1+2 and Wave 3+4 in line with Figure 1.

2. How many genes were expressed and the total number of samples in Wave 1+2 and Wave 3+4 in line with Figure 1.

3. A t-sne is performed on gene expression profiles of Wave 1+2 and Wave 3+4 patients referred as train and test set in the original Cancer Cell manuscript and t-SNE shows no segregation and uniform distribution of samples i.e. no batch effects and good quality control (Supp. Figure 1A)

4. Using different number of varying genes tried to identify optimal number of clusters [look at ‘gene\_expression\_data\_analysis.R’]. Best clustering result for 150 genes (Supp. Figure 1B). Best number of clusters is 3 (Supp. Figure 1C) and PCA plot of gene expression with 150 most varying gene highlighted in Supp. Figure 1D.

4. 50 hallmark pathways for cell survival were used. Rephrase ‘Gene set enrichment analysis’ section of <https://jitc.bmj.com/content/8/1/e000617.long>

5. Define the marker genes for different cell types

6. Gene set enrichment analysis performed for pathway activities and cell type enrichment (see Figure 1).

7. Marker gene for different modules identified in original Cancer Cell paper highlighted.

8. Gene set enrichment analysis performed for the modules to get module enrichment scores.

9. Oncogenes identified [refer to source] and 150 most varying genes considered and only these genes are considered for gene expression profiles.

10. Interpret the steps in the `analyze\_cell\_lines\_dnaseq.R’ script

a) Select genes with largest standard deviation in mutation profile sd>0.1, 374 genes in total

b) Convert mutation matrix to mutation type [Frameshift, Indel, etc.] matrix

c) Save both the mutation matrix and mutation type matrix for Wave 1+2, 3+4.

11. Join the cell line RNA-Seq and DNA-seq information as in `combined\_drug\_cell\_info.py’

12. Process all the drugs from `beataml\_probit\_curve\_fits\_v4\_dbgap.txt’ file.

13. Save the drugs and use the pubchempy package to get drug attributes such as:

a) InchiKey, CanonicalSMILES, XlogP, MolecularWeight

b) Convert SMILES into a format that is ingestible for the SMILES encoder from DrugRepurpose paper [Mall et al Bioinformatics, 2020] [Figure 3A]

14. For each drug SMILES representation, we also get molecular fingerprints using the rdkit package in python and vector representation using DrugRepurpose paper [Mall et al Bioinformatics, 2020].

15. Drug targets are obtained from Drugbank, PubChem combined and include direct targets, enzymes, transporters, etc.

16. Perform join operations on drug dataset, cell line dataset and drug-cell line response dataset to build the full dataset.

17. Downloaded the StringPPI database with strong interactions (interaction score > 0.7) to get protein-protein interactions.

18. Starting with StringPPI and drug targets, we perform random walk with restart to get probability of diffusion for all proteins. We multiply this with expression of each protein in a sample to take into account whether a protein is over-expressed or repressed. We have a weighted probability distribution for all proteins in a given cell line. We next use the hallmark gene sets, to estimate the area under the curve based on AUCell package (it gives higher AUC if the genes in the pathway are ranked higher and lower AUC if the genes in the pathway are randomly ranked). Thus, the AUC acts as a surrogate metric for similarity with higher AUC meaning lower distance of the pathway from the drug. Hence 1-AUC represents distance of the pathway from a drug. With higher distance meaning less impact of the drug on that pathway and lower distance meaning more impact of the drug on that pathway [Figure 3B].

19.This process is performed for the data from Wave 1+2 [Training Set] and Wave 3+4 [Test Set] in the ‘analyze\_cell\_lines\_drug\_combinations.R’ script.

20. The set of features used in the big data include:

a) Expression of oncogenes + variable genes

b) mutation profile of genes [370]

c) Mutation type profiles for all samples

d) Cell type enrichment profiles

e) Module enrichment profiles

f) Pathway enrichment profiles

g) Pathway distance for each sample

h) SMILES encoding or Morgan Fingerprint encoding for drugs

21. Four step process followed: a) Data curation b) Feature engineering (Enrichment scores and Pathway distances); c) Model building and d) Experimental analysis [see Figure 2].

22. Different suite of methods benchmarked in a regression framework to predict IC50 values. These method include:

a) Ridge Regression

b) Linear Regression

c)