**Integrative analysis of drug response and clinical outcome in acute myeloid leukemia**

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.

**Train the model using waves 1+2 data and predict score for waves 3+4**

**The data and methods they have used**:

Gene expression profiles, cell type, cell stage:

Primary data sources for information regarding drug/target relationships, Targetome and KINOMEscan

* Gene family scores: Rescaled the inhibitor AUC responses to be between 0 and 1 in order to ensure comparability between inhibitors with different concentration ranges. Then used the single sample gene set enrichment analysis (GSEA) (ssGSEA) approach as implemented in gene set variation analysis (GSVA) package to generate a score per family
* Original set of 14 gene expression modules (13 + Mod0 (gray) ‘outlier’ module) from the Beat AML manuscript derived using the weighted gene co-expression network analysis (WGCNA) methodology.
* Gene sets are typically summarized by their ‘eigengene’ which is their first principal component (PC) score. As the mod0 (gray) gene-set was more heterogeneous, we kept the first 5 PC scores for the eigengene.
* Use of WGCNA method
* PCA analysis of genes
* Correlation of gene expression with module eigenegen- kME values
* robust biweight midcorrelation
* WGCNA Module associations
* Cell-type scores
* Approach similar to PLAGE but separate centering and scaling for the specimen types.
* 2 types of data: samples from bone marrow, peripheral blood
* Used the top 30 genes for each of six single-cell AML tumor-derived signatures to compute gene set scores. Based on 1PC and aligned with avg. expression.
* Formation of genomic features
* Cell-type score+expression module
* Continuous correlation
* consensus AML fusions and mutation calls were then added in as binary features
* Random survival forest
* Used the combined set of genomic features
* Partykit package with postdiagnosis survival
* Pearson’s correlation
* Comparisons with the LSC17 signature
* NLP pipeline for automated analysis
* Data from the manually curated GSDS from Waves 1 + 2 were used to evaluate the performance of the NLP system. An NLP DataSet (NLPDS) was assembled that contained as far as possible the source pathology documents set from which the GSDS was originally obtained by manual review supervised by the data manager.
* This was subdivided into five partitions and used in an iterative training development-validation cycle to optimize the NLP and Python code. For evaluation purposes, we treated partitions 1 and 2 as a single training set and partitions 3, 4, and 5 as a single test set.

Tasks to do:

a) Get the list of all drugs used in the paper

1) Convert the drugs SMILES representation into molecular fingerprints using rdkit package in python

2) Convert the drugs SMILES representation into vector representation using TF-LSTM framework of DeepRepurpose

3) Convert the drugs SMILES representation into vector representation using Moltransformer: https://github.com/mpcrlab/MolecularTransformerEmbeddings

(Follow the steps highlighted in this github)

4) For each drug find the target genes using sites such as DrugBank, CheMBL, PubChem

Additional Data Sources:

a) GDSC – Genomics of Drug Sensitivity in Cancer - https://www.sanger.ac.uk/tool/gdsc-genomics-drug-sensitivity-cancer/

b) CCLE – Cancer Cell Line Encyclopedia - https://sites.broadinstitute.org/ccle/

c) DepMap – Contains information about cell lines such as transcriptome, mutation, copy number abberation, drugs which are sensitive for a particular cell line (measured in IC50, AUC etc) - https://depmap.org/portal/

d) Harmonizome – Contains harmonized information from CCLE, GDSC, DepMap about cell lines and sensitivity to various drugs - https://maayanlab.cloud/Harmonizome/download

Please have a look at this synapse link: https://www.synapse.org/#!Synapse:syn21569634/wiki/602437

Action Points:

a) Upload the template to follow for designing the models. – Raghvendra - **Done**

b) For each drug identify the drug targets and save in excel file in the format:

Drug Name, Drug Inchikey, Drug Targets (semi-colon separated) – Siddhi (**Done**, reformatting)

c) Add the list of frequently dysregulated pathways in cancer as genesets and estimate pathway activation for each cell line – Raghvendra - **Done**

d) Put the list of oncogenes in Data folder – Raghvendra - **Done**

e) Identify the list of 1000 most variable genes (based on expression) – Siddhi

f) Convert the mutation information into a matrix where each sample is row and columns are list of genes and each entry is number of times a gene was mutated in a sample – Raghvendra - **Done**

g) Get the list of marker genes for celltype scores and estimate celltype score for each sample- Raghvendra - **Done**

h) Divide the cell lines into training and test set with expression of genes, clinical characteristics and pathway activations in one data frame – Raghvendra - **Done**

I) Get the pathway AUC score for each cell line, pathway combined – Raghvendra

Action Points (21.05.2023):

a) Generate the mapping with <sample id, module class>. - From Raghvendra to Siddhi

b) Identify the optimal set of varying genes using k-means/t-sne with the module class to quantitatively identify the set – From Siddhi to Raghvendra

c) Generate the training and test set for ML models by taking union of oncogenes and varying genes – From Raghvendra

d) Divide the task of building ML models with this new train/test set - Raghvendra/Siddhi