Benchmarking state-of-the-art unsupervised learning methods for multi-omics data

Shamieraah Jamal
Computational Biology Department
Carnegie Mellon University
shamierj@andrew.cmu.edu

Anushka Sinha
Computational Biology Department
Carnegie Mellon University
anushka3@andrew.cmu.edu

Jen Yi Wong
Computational Biology Department
Carnegie Mellon University
jenyiw@andrew.cmu.edu

1 Introduction

Acute Myeloid Leukemia (AML) is a blood disorder characterized by the rapid proliferation of myeloid stem cells in the bone marrow [**Nix**]. It constitutes 15-20% of leukemias in children [1]. Patient stratification is typically done through cytogenetic profiling and determines the best treatment for patients. While there are known genetic and chromosomal abnormalities in AML, stratification by epigenetic profiles of patients is not well-studied and could offer novel insights for children with Type II mutations [1]. Mutations in *DNMT3A* are fairly common in AML [**Nix**] and there were also differences in miRNAs expressed between pediatric and adult AML patients [1], suggesting that DNA methylation and miRNA profiles could offer new patient stratification strategies. Thus, this project aims to identify novel patient classifications based on a combination of altered methylation patterns and miRNA transcriptomes of pediatric acute myeloid leukemia (AML) cases from the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) AML project [2], sourced from the NCI Genomic Data Commons (GDC) [3].

Since this requires combining large -omic datasets with different modalities, this project aimed to compare two existing approaches to unsupervised learning in the literature. Firstly, Eckardt et al. (2023) [4] employed meta-clustering to identify inherent patterns withing an acute myeloid leukemia (AML) dataset which comprised of a diverse cohort of patients from multiple medical centers. This was done by applying a combination of 11 dimensionality reduction techniques with 11 clustering algorithms, resulting in a new feature matrix whereby each patient had a cluster label as a feature from each combination. This resulting matrix was itself clustered to produce a final aggregate clustering that would overcome the biases and heterogeneity in each method. Secondly, Wu et al. utilized Multiple Kernel k-means clustering (MKKC) in their study to identify clinically relevant subgroups. In their application, the MKCC method successfully identified asthma subphenotypes exhibiting distinct responses to a specific treatment.

In this report, we compare the effectiveness of ensemble-based meta-clustering and kernel-based MKKC approaches in clustering methylomic and miRNA transcriptomic data from pediatric AML patients. As a result, we identify 3 distinct clusters and characterized biological differences through Gene Ontology (GO) analysis to suggest stratification strategies for the patient population.

2 Methods.

2.1 Data Pre-processing

The methylation data was obtained from the Illumina Infinium 27K platform, while the miRNA data was sequenced using the Illumina Hi-Seq 2000 platform. Data preprocessing was conducted to ensure consistency and address sparsity in the methylation dataset obtained from the TARGET-AML project. The methylation 27 dataset initially comprised 380 samples with 27579 features, while the miRNA dataset consisted of 321 samples with 486427 features. Subsequently, only samples with both methylation 27 and miRNA data were retained. Features with over 90 % missing values were eliminated. Remaining missing values were imputed using the mean. Samples corresponding to patients who underwent more than one timepoint were excluded as well. This yielded a final methylation 27 dataset comprising 169 samples with 26150 features and a miRNA dataset containing 262 samples with 1881 features.

Submitted to Computational Biology Department, Carnegie Mellon University. 02-718 Computational Medicine, Fall '23.

2.2 Feature Selection

Highly variable features were identified by calculating the variance of each feature (Appendix Figure 5) and selecting the top 1000 features with the most variation for the Methylation 27K dataset and the top 250 features for the miRNA dataset. The data was then standardized using `Scikit-learn`'s `StandardScaler`.

2.3 Meta-clustering

Meta-clustering was performed closely following the methods outlined in Eckardt et al (2023) [4]. The aim was to produce a good clustering result in an unsupervised manner by trying different combinations of dimensionality reduction and clustering algorithms. Briefly, 9 different dimensionality reduction methods were used in combination with 10 different clustering methods. The data was reduced to 2, 4, 6, 8 or 10 dimensions. This produced a 450-feature vector for each sample. The 9 different dimensionality reduction methods and their abbreviations are shown in Table 1. The 10 different clustering methods were k-means, agglomerative clustering with Ward linkage, spectral clustering, agglomerative clustering with average linkage, Gaussian Mixture Model (GMM), BIRCH clustering, Ordering Points To Identify the Clustering Structure (OPTICS), Mean-Shift clustering, DBScan and Affinity Propagation. The feature matrix was then clustered again using k-means clustering to obtain the final predicted labels for each cluster. The number of clusters for the k means was determined using an elbow plot of the objection function for k-means.

Dimensionality reduction method	Abbreviation
Principal component analysis	PCA
Incremental principal component analysis	IPCA
Singular Value Decomposition	SVD
Gaussian Random Projection	GRP
Sparse Random Projection	SRP
Multidimensional Scaling	MDS
IsoMAP	ISOMAP
Linear Local Embedding	LLE
Mini-Batch Dictionary Learning	MBDL

Table 1: Table showing dimensionality reduction methods for meta-clustering their abbreviations.

2.4 Robust Multi-Kernel K-means clustering (MKKC)

The robust Multi-Kernel K-means clustering (MKKC) method was performed according to the paper by Wu et al. (2019) [5]. The code was written in Python. Briefly, PCA was used to reduced the methylation27 data and the miRNA data into 20 components each. The Radial Gaussian Function kernel was used to fit Methylation 27 and miRNA data separately to produce two views. MKKC was used to optimize the kernel weights on these two kernels and k-means was then used to cluster the combined weighted kernel data and produce the sample labels.

2.5 Clustering metrics and feature perturbation

The quality of clustering for the two methods were assessed using the Calinski-Harabasz Index and the Davies-Bouldin Score. The Calinski-Harabasz Index is calculated as a ratio of the between-cluster separation to the within-cluster dispersion. A higher score usually means better separation. The Davies-Bouldin Score is calculated by comparing the ratio of the inter and intra-cluster distances for a cluster with another cluster most similar to it. A lower score indicates better separated clusters. To explore the robustness of the MKKC method, we randomly chose different proportion of features (0.1, 0.2, 0.5, 0.75, 1.0) and replaced the values with random values drawn from a density function fitted on the data from each feature.

2.6 Feature enrichment and gene ontology analysis

The differential features between the clusters were found using an ANOVA test after correcting the p-values using the Benjamini-Hochberg method. Further analysis was conducted for differentially-enriched miRNA. The miRNA target genes were searched using TargetScanHuman [6], a prediction-tool that finds targets based on conserved k-mer sites complementary to the seed regions of miRNAs. The genes for all the miRNAs were combined and fed to Gene Ontology Biological Process (BP) [7] [8] and PANTHER Pathway [9] searches. Genes with log fold-enrichment higher than 1 and negative log FDR-corrected p-value greater than 10 were used to rank the most enriched BPs. Genes with log fold-enrichment higher than 1 and negative log FDR-corrected p-value greater than 5 were used to rank the most enriched PANTHER Pathway terms.

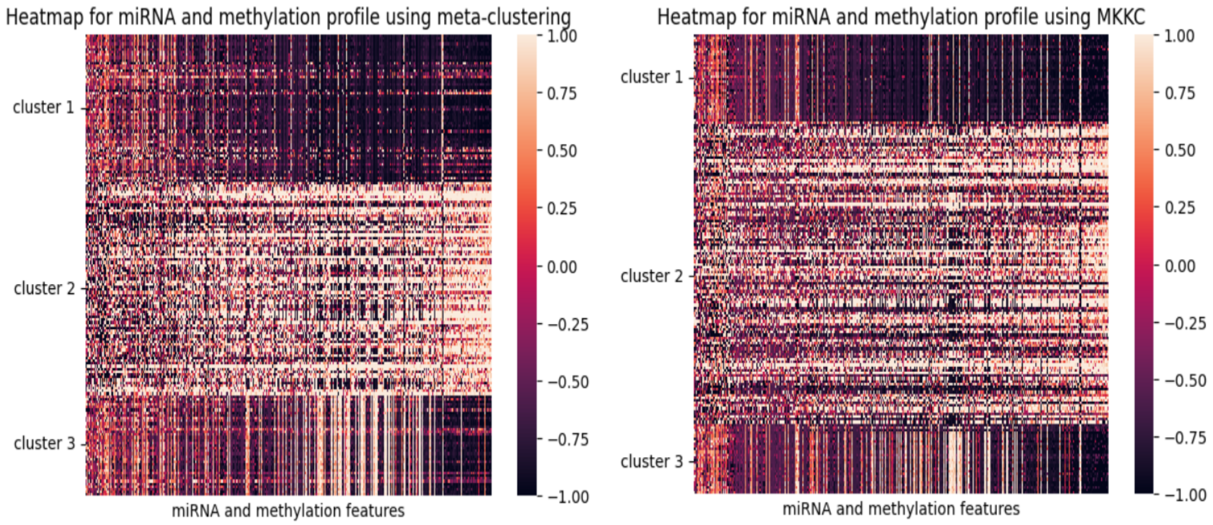

Figure 1: Heatmap showing cluster labels using meta-clustering (left) and MKKC (right).

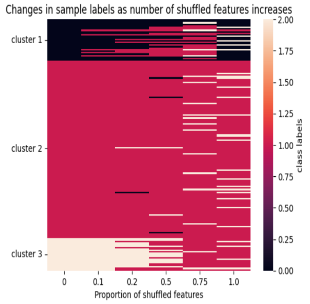

Figure 2: Heatmap showing how feature perturbation affects MKKC class labels.

Proportion of shuffled features	Proportion of samples with different labels
0.1	0.030
0.2	0.083
0.5	0.178
0.75	0.296
1.0	0.249

Table 3: Table showing proportion of samples with different class labels after shuffling.

3 Results

3.1 Comparing cluster quality between meta-clustering and MKKC

The results from metaclustering and MKKC showed similar profiles on the heatmap. Clusters 1 and 3 are distinct whereas Cluster 2 appeared to be more heterogenous. It was found that meta-clustering had 949 significantly different features while MKKC had 946 significantly different features using ANOVA with Benjamini-Hochberg p-value correction. However, the number of patients in each cluster was different. Metaclustering reported clusters with 54, 78 and 37 patients in Clusters 1, 2 and 3 respectively, whereas MKKC reported clusters of 28, 119 and 22 respectively.

Method	Calinski-Harabasz Index	Davies-bouldin Index
MKKC	0.070	14.32
Metaclustering	1.53	5.58

Table 2: Table showing evaluation of the MKKC and metaclustering methods.

As shown in Table 2, the MKKC method produced a better clustering of the samples compared to the meta-clustering method as it has a lower Calinski-Harabasz score and a higher Davies-bouldin Index. This may be because the meta-clustering method is able to choose the best clustering based on the aggregate vote, none of the methods available to it performs as well as the MKKC. Hence, meta-clustering is unable to perform as well as the MKKC in general. In terms of runtime, the MKKC also runs much faster than the meta-clustering method as it does not need to iterate over every model. As seen in Figure 2, the changes in class labels increases as the proportion of features are shuffled. The MKKC model seemed to be quite robust as most of the samples retained the same label even when up to 20 % of the features were random values as seen in 3.

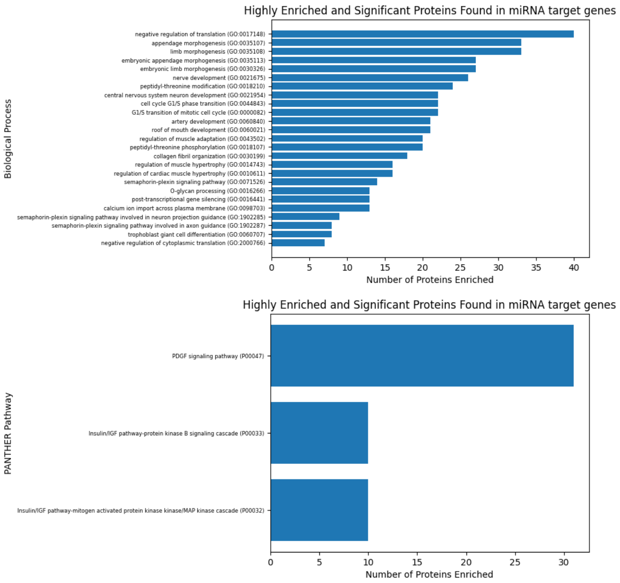

Figure 3: Gene Ontology (GO) terms enriched using Biological Processes (top) and PANTHER Pathway (bottom)

3.2 Biological processes affected by differential features.

Since MKKC produced a better clustering, it was chosen for further analysis. Features that differentiated the clusters were identified by performing a one-way ANOVA between clusters for every feature. Corrected p-values smaller than 0.05 were then used to pick the top differentially-enriched features, giving 2809 CpG loci and 27 miRNAs. Target genes for miRNAs were searched using TargetScanHuman and the gene list was used to search for enriched GO terms for Biological Processes and PANTHER Pathways. From Figure 3, we can see that negative regulation of translation, morphogenesis and cell cycle terms are enriched in BP, while PDGF and IGF pathways terms are enriched in PANTHER. There is evidence to support the role of these processes in AML. Negative regulation of translation could indicate the suppression of cytokines released by immune cells, hence allowing tumour cells to evade death [10]. Morphogenesis could indicate the abnormal differentiation of myeloid stem cells into other lineages, which is a common feature of cancers, as well as the formation of new blood vessels which promotes the spread the cancerous cells [11]. Furthermore, PDGF isoforms have been show to affect the proliferation and cytokine secretion behaviour [12]. The exact effect of the miRNAs identified on these pathways could be further investigated experimentally.

Moreover, we examined the effects of specific miRNAs to elucidate the differences between clusters (Figure 4). For example, hsa-miR-148b has high expression in clusters 1 and 3 and low expression in cluster 2. Hsa-miR-148b comes from the miR-148/152 family of miRNAs, which was found to be correlated with high remission rate in a cohort of adult AML patients [13]. This provides some preliminary validation that the identified miRNAs may have both clinical and biological significance, though this should be supported with further analysis.

4 Limitations and Future work

In this report, pediatric AML data consisting of methylation sites and miRNA transcript profiles were clustered using either meta-clustering or MKKC. It was found that MKKC produced better clustering when comparing Calinski-Harabasz Index and Davies-bouldin Index, so the cluster labels from this method was used for further analysis. To elucidate the biological differences between the clusters, the most differentiating features were identified using one-way ANOVA tests with corrected p-values. The resulting features were then used for gene set enrichment analysis (GSEA) to investigate the biological pathways affected by the target genes of the methylation and miRNA sites. Negative regulation of translation was a top identified pathway and miR-148b was a crucial differentiator of the clusters. Potential limitations of this project include the size of the data set and the selection of clustering algorithms in the meta-clustering method. With 169 samples, the relatively small sample size may hinder generalization to a broader and more diverse population. This constrained sample size did not allow the allocation of a subset of subjects for cluster analysis, a strategy employed in Wu et al.'s work [5] to validate the clusters. Future work should continue GSEA for methylation sites using `methylGSA`, a R package designed for methylomic gene set analysis. As a result, we have demonstrated the effectiveness of unsupervised learning methods to integrate multi-omic data for patient stratification in a pediatric AML population.

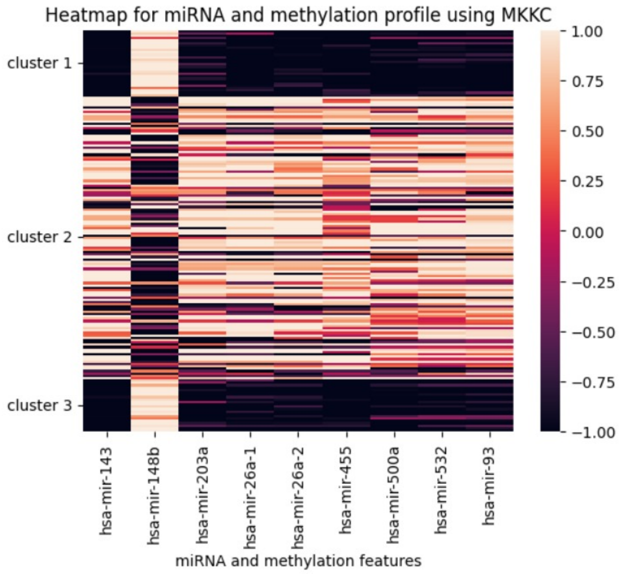

Figure 4: Differential expression of miRNAs among clusters

References

[1] Jasmijn de Rooij, C. Zwaan, and Marry van den Heuvel-Eibrink. “Pediatric AML: From biology to clinical management”. In: *Journal of Clinical Medicine* 4.1 (2015), pp. 127–149. DOI: 10.3390/jcm4010127.

[2] URL: https://www.cancer.gov/ccg/research/genome-sequencing/target/studied-cancers/acute-myeloid-leukemia.

[3] Robert L. Grossman et al. “Toward a shared vision for cancer genomic data”. In: *New England Journal of Medicine* 375.12 (2016), pp. 1109–1112. DOI: 10.1056/nejmp1607591.

[4] Jan-Niklas Eckardt et al. “Unsupervised meta-clustering identifies risk clusters in acute myeloid leukemia based on clinical and genetic profiles”. In: *Communications Medicine* 3.1 (2023). DOI: 10.1038/s43856-023-00298-6.

[5] Wei Wu et al. “Multiview cluster analysis identifies variable corticosteroid response phenotypes in severe asthma”. In: *American Journal of Respiratory and Critical Care Medicine* 199.11 (2019), pp. 1358–1367. DOI: 10.1164/rccm.201808-1543oc.

[6] Vikram Agarwal et al. “Predicting effective microrna target sites in mammalian mrnas”. In: *eLife* 4 (2015). DOI: 10.7554/elife.05005.

[7] Michael Ashburner et al. *Gene ontology: Tool for the unification of biology*. URL: https://www.nature.com/articles/ng0500_25/.

[8] ; *The Gene Ontology Resource: Enriching a gold mine*. URL: https://pubmed.ncbi.nlm.nih.gov/33290552/.

[9] Huaiyu Mi et al. “Panther version 14: More genomes, a new panther go-slim and improvements in enrichment analysis tools”. In: *Nucleic Acids Research* 47.D1 (2018). DOI: 10.1093/nar/gky1038.

[10] Arati Khanna-Gupta. “Regulation and deregulation of mrna translation during myeloid maturation”. In: *Experimental Hematology* 39.2 (2011), pp. 133–141. DOI: 10.1016/j.exphem.2010.10.011.

[11] Olga Blau. “Bone Marrow Microenvironment in the pathogenesis of AML”. In: *Myeloid Leukemia - Basic Mechanisms of Leukemogenesis* (2011). DOI: 10.5772/25889.

[12] Brynjar Foss, Elling Ulvestad, and Øystein Bruserud. “Platelet-derived growth factor (PDGF) in human acute myelogenous leukemia: PDGF receptor expression, endogenous PDGF release and responsiveness to exogenous PDGF isoforms by in vitro cultured acute myelogenous leukemia blasts”. In: *European Journal of Haematology* 67.4 (2001), pp. 267–278. DOI: 10.1034/j.1600-0609.2001.0430a.x.

[13] Xiao-Xue Wang, Rui Zhang, and Yan Li. “Expression of the mir-148/152 family in acute myeloid leukemia and its clinical significance”. In: *Medical Science Monitor* 23 (2017), pp. 4768–4778. DOI: 10.12659/msm.902689.

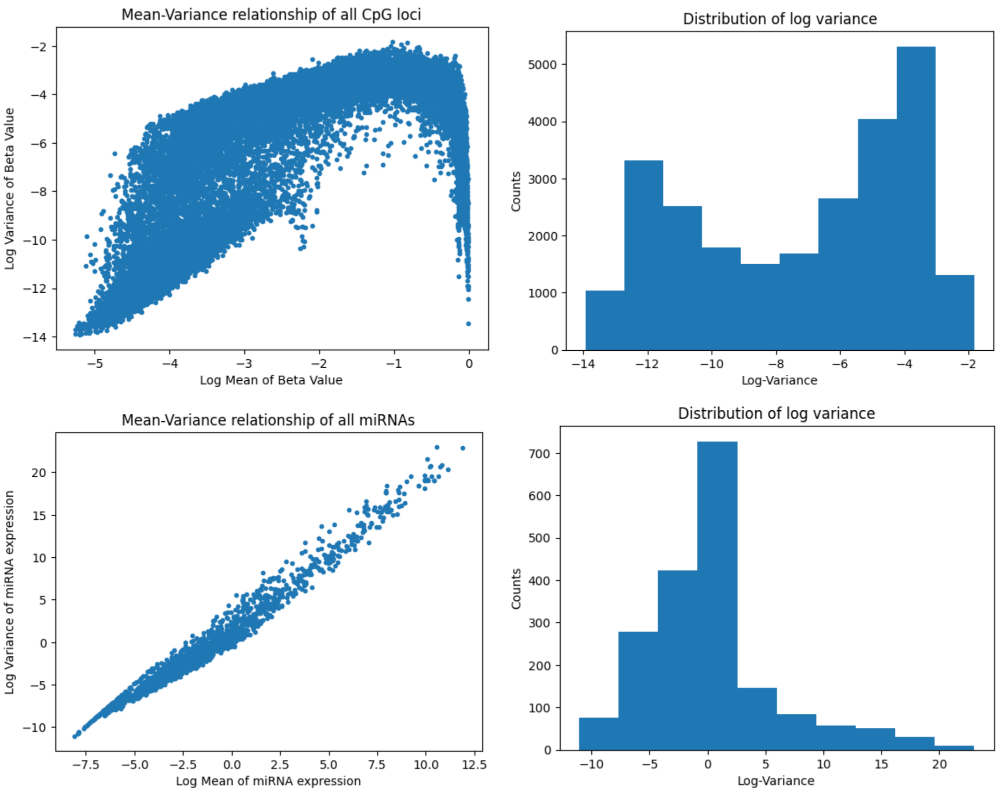

Figure 5: Selection of highly variable genes (HVGs) of Methylation 27K features (top row) and miRNA features (bottom row) was done using mean-variance relationship (left) and distribution of log-variance (right).

5 Appendix

A Highly Variable Gene Selection