# Unsupervised Cancer Subtype Classification using SNV-derived Differential Cumulative Mutation Signatures and Soft Clustering

Nikhil Anand (BE20B022), Rajagopal Subramaniam C (BE20B026)

May 2024

**Abstract**

We aim to identify robust cancer subtypes and demonstrate statistical and biological validity. Current methods for identifying these subtypes rely on expression data and less interpretable models. We propose a novel approach integrating functionally deleterious mutations, Differential Cumulative Mutation (DCM) signatures, and miRNA data. Our methodology involves two strategies: generating differentially mutated gene subnetworks and miRNA profiles for K-means clustering and assessing functionally deleterious mutations to create a consensus matrix for soft clustering. These methods identify subtypes driven by network dysregulation and specific mutations, respectively. Validation using gold standard labels with ARI and NMI metrics shows that our model performs below benchmarks for Breast invasive Carcinoma but exceeds them for Colon Adenocarcinoma, demonstrating the potential validity of our approach.

## 1 Introduction

Cancer subtypes are specific, identifiable variations of cancer defined by common characteristics, including molecular, genetic, and cellular differences. Identifying these subtypes is crucial for several reasons. Firstly, it enables tailored treatment strategies that improve patient prognosis and personalize therapy, ensuring that treatments are specifically targeted to the unique characteristics of each subtype. Secondly, subtypes vary widely and can be categorized by tissue of origin, such as carcinoma, sarcoma, or lymphoma, or by genetic mutations driving tumor growth, such as HER2-positive breast invasive carcinoma or KRAS-mutant colon adenocarcinoma. Understanding these distinctions allows for developing more effective treatment protocols and identifying potential therapeutic targets.

### 1.1 Definitions

Single Nucleotide Variations (SNVs) are mutated genes that differ in a specific nucleotide relative to the original genome. By PPI (Protein-Protein Interaction), we refer to the functional protein interaction network provided by the STRING database. miRNA stands for micro-RNA, non-coding RNA fragments that bind and inhibit mRNA translation to proteins. Soft clustering assigns samples to multiple clusters with varying degrees of likelihood. Biclustering identifies subsets of samples and genes sharing similar patterns. DCM signatures refer to Differential Cumulative Mutation signatures, a concept introduced by this paper, referring to a type of sample mutation profile formulated by adding up the number of mutations within each subnetwork and getting the difference between those values and the mean number of mutations of that sample across subnetworks.

### 1.2 Problem Statement, Objectives and Justification of our approach

Our research aims to identify cancer heterogeneity by identifying robust subtypes using Single Nucleotide Variation (SNV) mutation data. We intend to integrate this genomic information with relevant biological data, such as shared pathways, to form the bedrock of our approach. Our central hypothesis is that cancers within the same subtype exhibit similar dysregulations in common pathways, suggesting a shared mechanism of oncogenesis. This hypothesis is grounded in the understanding that cancers, despite their diverse origins, often converge on a limited number of disrupted biological processes. We aim to uncover the genetic signatures that define each subtype by utilizing these shared pathways.

The outcome we propose is driven by the belief that any cancer sample can be comprehensively characterized by three distinct aspects: functionally deleterious single nucleotide variants (SNVs), gene subnetworks with higher DCM, and microRNA profiles.

- **Functionally Deleterious SNVs:** Functionally deleterious SNVs are a crucial aspect of cancer characterization because they directly impact protein function, leading to cellular dysfunctions that can drive tumorigenesis. These mutations often result in loss of function in tumor suppressor genes or gain of function in oncogenes, playing pivotal roles in cancer development and progression.

- **Gene Subnetworks with High Mutation Rates:** Analyzing gene subnetworks with elevated mutation rates offers insights into the dysregulation of entire biological pathways. Cancer is rarely caused by a single gene mutation; instead, it often results from disruptions in multiple genes within a pathway. By identifying subnetworks with a high density of mutations, we can pinpoint key pathways likely contributing to cancer.
- **MicroRNA Profiles:** MicroRNAs (miRNAs) are small non-coding RNAs that regulate gene expression post-transcriptionally. Their role in cancer is well-documented, as they can act as oncogenes or tumor suppressors. Dysregulated miRNA expression can simultaneously lead to the aberrant regulation of multiple genes, contributing to cancer progression. Profiling miRNA expression provides a layer of regulatory information that complements the genomic and network-based analyses, offering a more holistic view of the molecular mechanisms driving cancer.

By integrating these three aspects—functionally deleterious SNVs, mutational patterns in gene subnetworks, and miRNA profiles—we aim to capture a holistic view of the dysregulation within a sample. Thus, we aim to implement a robust algorithm to detect this, describe the clusters formed, and statistically validate the clusters.

### 1.2.1 Advantages of soft clustering

Soft clustering offers three key advantages, particularly in classifying cancer subtypes. First, it provides a measure of confidence in cluster assignments, addressing the limitations of hard clustering, which lacks this capability. Second, it accommodates the complex nature of biological data, recognizing that some samples may exhibit characteristics of multiple clusters, thus reflecting the true variability within the data. Third, it enhances understanding of clustering outcomes, capturing the inherent uncertainty and overlap.

## 1.3 Related Works

In our methodology, we focus on unsupervised learning to help us differentiate cancer samples without relying on a ground truth that is often unavailable. We do this for two reasons, as observed by Park et al. [1]: firstly, the inability to discover new cancer subtypes prevents our understanding of cancer expansion. Additionally, existing literature indicates that supervised approaches have frequently demonstrated limited efficacy in subtype classification. In literature, the exploration of unsupervised approaches is diverse, encompassing several sophisticated methods: Bayesian theory, such as a method proposed by Lock et al. [2], which provides a probabilistic framework for modeling the uncertainties inherent in cancer data; matrix factorization techniques, including a method proposed by Zhang et al. [3], which decompose complex data into simpler, interpretable structures; statistical-based methods, observed in Mo et al.[4], As seen in Vaske et al., offering robust tools for identifying clusters within data without prior labeling and network-based approaches.[5], which considers the interconnectedness of genes and pathways. Advances in network-based approaches, such as by Chuang et al. [6], helped uncover that using gene subnetworks as markers was much more efficient than determining single-gene markers (SGMs) for cross-platform evaluations. Furthermore, the advent of deep learning models has allowed us to leverage neural networks for subtype identification, as leveraged by Guo et al. [7]. There is a lot of heterogeneity regarding the combinations of datasets used. The most common datasets used are mRNA expression data, followed by miRNA expression data, DNA methylation data, CNV data, expression and mutation data, PPI, protein, and other data in that order. However, most papers use a combination of these datasets for these models.

In 2010, P. Dao et al. [8] investigated the utility of cancer subnetwork markers in detecting cancer metastasis. The approach used a density-based clustering method to detect subnetworks that showed high similarity in differential expression profiles among a small group of samples, and they evaluated their results on data determining whether or not cancer would metastasize. This paper mainly talks about how biclustering approaches have been used in classification tasks (for example, deciding whether or not a cancer sample may metastasize). Furthermore, it addresses the phenotypic complexity of cancer and the various pathways that may be affected in any given cancer. The methodology presented was twofold – first, they clustered the genes into density-constrained subnetworks using a threshold $\alpha$ with the PPI network edges as edge weights. They only considered strongly connected subnetworks that exceed the threshold for further analysis. Finally, only those differential $L$-bicluster networks were considered the final clusters. Differential $L$-biclusters are clusters that remain relatively consistent in their differential expressions among all elements of a group of at least L samples. Thus, these samples will likely have similar dysregulated pathways and must be classified together. The paper notes that the benefit of the given approach over simply correlating single gene markers is that those approaches tend to be inconsistent across platforms and studies. The given approach is much more consistent. The paper also evaluated the resultant biclusters found by using the clusters to classify colon cancers into metastatic or non-metastatic and classifying breast cancers into mutant and wild-type.

Multiple papers have reviewed the existing methodologies for identifying subtypes. We benchmarked our results by referencing the review paper by Park et al. (2023) [1], which analyzed BRCA and COAD. This review compared metrics

such as Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) across different feature selection methods, feature numbers, and algorithms. They used the identified molecular subtypes as the gold standard. We selected the best-performing algorithm as the benchmark to measure our model's performance.

### 1.4 Gaps Identified

- Limited focus on SNV data compared to expression data
- Insufficient analysis of gene subnetwork dysregulation using PPI data, especially on the basis of SNV data
- Failure of hard clustering methods to address sample heterogeneity within clusters

### 1.5 Methodology and Findings in Brief

To address these issues, we propose a novel approach to subtype identification by integrating functionally deleterious mutations, differentially mutated gene subnetworks, and miRNA data.

Our methodology involves two complementary strategies. First, we generate DCM signatures and miRNA profiles, applying K-means clustering to identify potential subtypes. This approach leverages the collective impact of network-level dysregulation and miRNA expression differences. Second, we assess functionally deleterious mutations, create a consensus matrix from best-performing predictive models using Bipartite and BIRCH clustering algorithms, and perform soft clustering to capture the confidence of subtype assignments.

The resulting subtypes from each method are carefully analyzed and validated. The first method reveals subtypes driven by network and miRNA differences, while the second method highlights those influenced by specific deleterious mutations. We validate our approach using gold standard labels and metrics such as Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). Our model performs below benchmarks for Breast Invasive Carcinoma but exceeds benchmarks for Colon Adenocarcinoma, indicating the model's utility.

## 2 Methodology

### 2.1 Data Description

We obtained our data from The Cancer Genome Atlas (TCGA) and utilized multi-omic data from micro-RNA data and Single Nucleotide Variation (SNV) data. The SNV data was provided for about 936 BRCA (Breast Cancer) and 362 COAD samples (Colon Adenocarcinoma) across more than 18000 genes.

### 2.2 Method 1: Subnetwork Analysis-based Retrieval of DCM signatures for K-Means Clustering

#### 2.2.1 Protein-Protein Interaction (PPI) Network Data Extraction and Relevance

Protein-protein interaction (PPI) networks are networks encoding information about the interactions between various proteins produced within cells. Proteins are studied in terms of their interactions with other proteins in various wet lab studies, and the results of several such studies published in the literature are averaged to get a probability of interaction between any two proteins. The STRING database collects all this information and generates a graph with nodes representing various proteins and edges representing the probability of interaction between the two nodes that the edge connects.

To access this data, we need to run API calls to the STRING database by passing a list of proteins of interest as arguments to the API call and expecting the function to return a table showing a list of all the edges, with information on the node endpoints of each edge, the weight of the edge (probability of interaction), and some more information which we don't need.

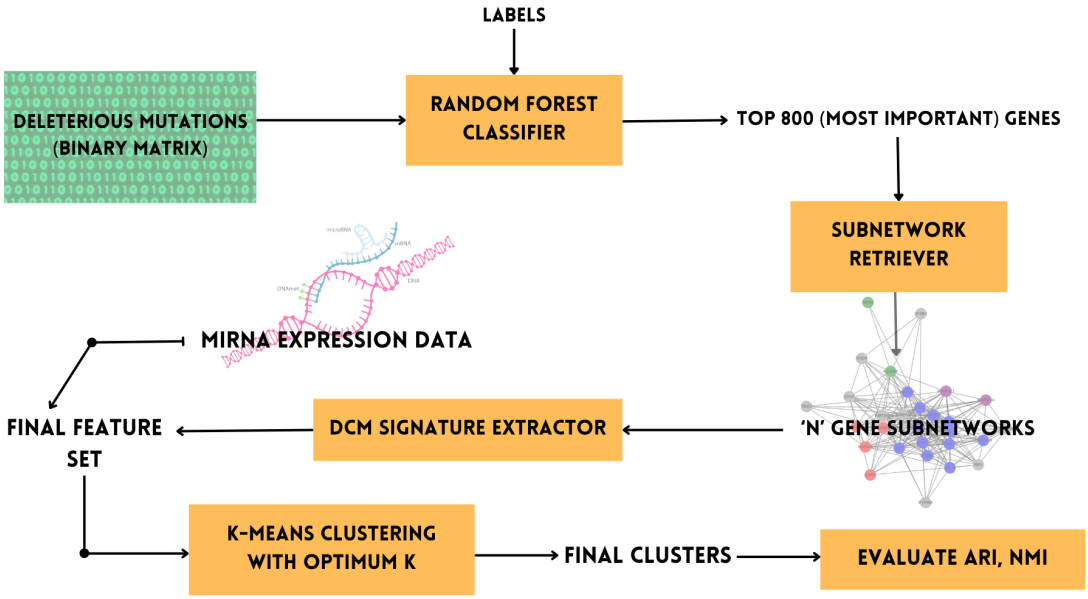


Figure 1: In the first approach, we use supervised feature selection, followed by analysis of subnetworks and extraction of the DCM signatures of the samples. After combining this with the miRNA expression data, we used the final feature set for K-Means clustering and evaluated the ARI and NMI for the clusters obtained.

The PPI data turns out to be quite relevant since it has been found in several literature studies that while single gene markers have proven extremely potent in most disease studies, cancer tends to have huge numbers of mutations, and studying properties at a subnetwork level is much more sensible. While there have been studies incorporating differential expression data at a subnetwork level, studying mutational signatures at a subnetwork level is a relatively new idea that we propose.

### 2.2.2 Data Preprocessing

The first step is the preprocessing of data. We retrieved the SIPS scores, telling us the likelihood that the mutation is deleterious by analyzing how often the same mutation occurs in the majority of healthy genomes. We marked the samples labeled as deleterious with a value of 1 and those that were either non-deleterious or lacked a score with a value of 0. Thus, we generated a binary value for each sample-gene pair and a matrix with samples in the rows and genes in the columns.

### 2.2.3 Supervised Feature Selection

Next, the matrix was fed into a set of supervised learning algorithms. We had the subtype labels (the "gold standard"), and we used these subtypes to analyze the most important genes that might influence the determination of the specific subtype. Through a decision-tree-based approach, it was easy to visualize the impact of the TP53 and PIK3CA genes that helped purify out some of the subtypes in the earlier nodes of the tree. This essential line of analysis shows us that mutations in certain genes significantly correlate with deciding the specific cancer subtype. Furthermore, these results were consistent with the top genes obtained using the Random Forest Classifier, where the top 2 genes were TP53 and PIK3CA, followed by several other genes. Due to certain computational restrictions in the next section, we had to reduce the space of the genes, and we reduced it to the top 800 most important features deduced by the Random Forest classifier. It can be noted that the accuracy of classification improved significantly (by about 2%) after only incorporating the top 20 features into the model, showing that a small set of genes (and not all 18000) might be encoding a majority of the information. Fig. 2 represents the Decision Tree, and Fig. 3 shows the Random Forest feature importance plot. These features are then exported to the next step.

Furthermore, we analyzed micro-RNA (miRNA) features using the Decision Tree and Random Forest approach. Through the decision tree approach, the accuracy of our classifier improved by a whopping 10% by including the micro-RNA data, and the accuracy of our Random Forest improved by a similar amount. Furthermore, the Decision Tree showed that a particular miRNA feature heavily correlated to some cancer subtypes. Thus, it is evident that micro-RNA data was as important as the gene mutation data in our clustering approach. In the BRCA dataset, five miRNA features

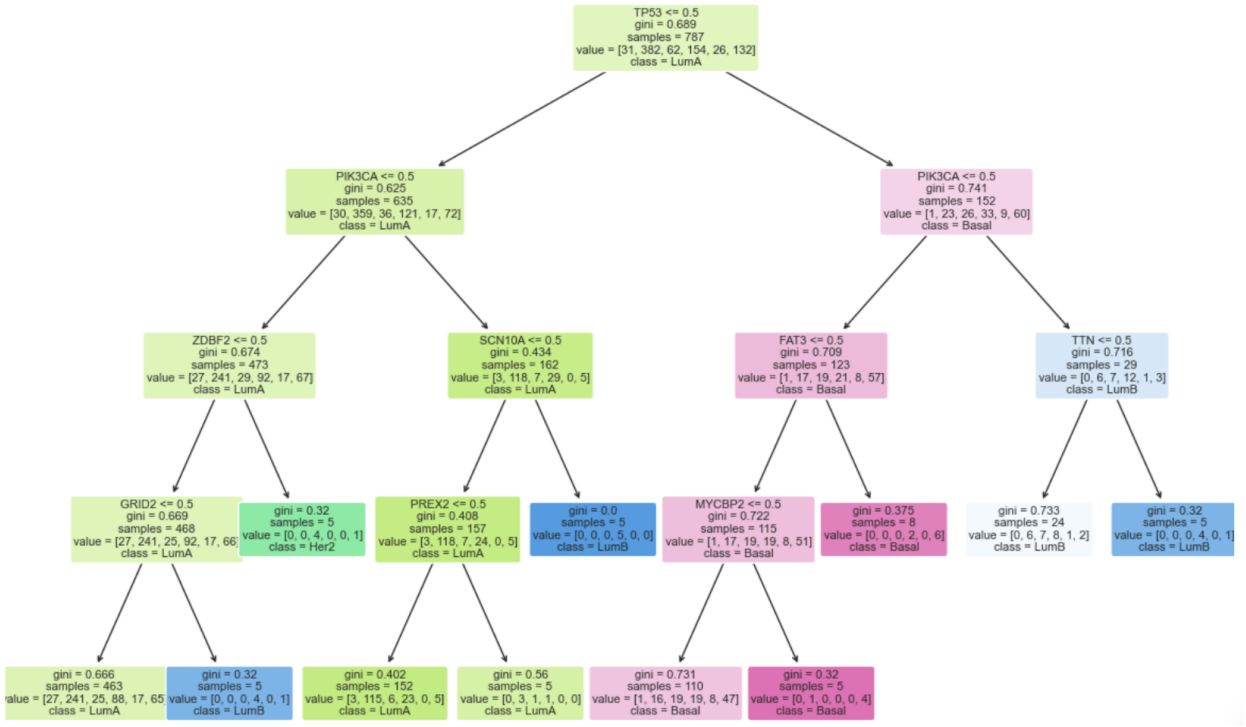


Figure 2: We constructed a decision tree showing the most critical genetic mutations determining cancer subtypes. The list of numbers in each cell shows the number of elements in each subtype at that node. We see that TP53 and PIK3CA are significant markers.

were available, whereas ten features were made available in the COAD dataset. These features were standardized using a StandardScaler (shown in Fig. 4) and passed into the next step.

#### 2.2.4 Retrieval of Gene Subnetworks

We analyzed several algorithms for retrieving gene subnetworks, including a density-based ranking algorithm to access the $\alpha$-dense subnetworks of genes and extract them. We decided to go ahead with the Louvain algorithm, which uses a greedy heuristic approach that fetches distinct subnetworks of genes. One of the drawbacks of the Louvain approach was that it could not be applied to all 18000 genes since the set of interactions would become exponentially large, making it take up too much space to process all the edges. To combat this issue, we retrieved the top 800 most important genes, as per the Random Forest classifier, and fetched their interactions. Next, we implemented the Louvain algorithm, where we found subnetworks for these 800 genes. We retrieved twenty-eight subnetworks for the BRCA dataset and sixteen subnetworks for the COAD dataset.

#### 2.2.5 Extraction of DCM Signatures for each Sample

We hypothesize that the specific mutational signatures that samples have across subnetworks of genes play a crucial role in determining the subtype of cancer. Thus, rather than looking at "binary" mutations in genes, we analyze the variance of the number of mutations (a more continuous feature) across different subnetworks. Unlike other diseases, often caused by a single point mutation leading to the condition, cancer tends to be caused by several unpredictable mutations metastasizing in unpredictable ways. We found that certain patients had many more mutations than others across subnetworks. Therefore, we took the mean number of mutations that a specific patient suffered across all subnetworks and subtracted that number from the number of mutations in each subnetwork, giving us a standardized measure of the variance of each subnetwork from the mean number of mutations for that sample, which we call the Differential Cumulative Mutation (DCM) signature of that sample. Finally, we appended the miRNA data as additional columns. Fig. 6 shows the final heatmap of the differential cumulative mutation signatures for the BRCA dataset, where the Y-axis is the sample in question, and the X-axis is the specific ID of the subnetwork (in the first 18 columns) and miRNA expression data in the last five columns. We visualized the same heatmap using the COAD dataset in Fig. 5, where the last ten columns represent miRNA expression data. As can be seen, the patient samples can be clustered quite naturally by analyzing these features.

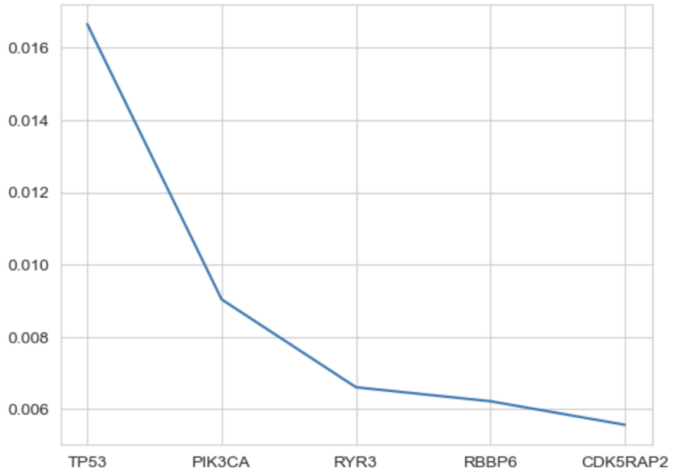


Figure 3: Plot of the most important features on the X-axis and their corresponding feature importances on the Y-axis obtained through the Random Forest Classifier (BRCA Dataset)

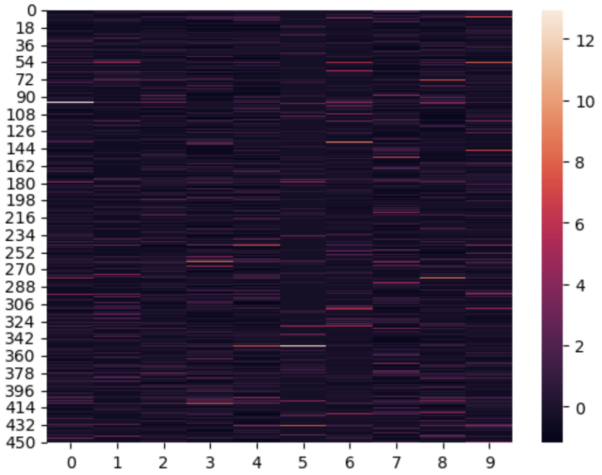

Figure 4: Heatmap Analysis of 10 miRNA Expression Patterns in COAD Dataset

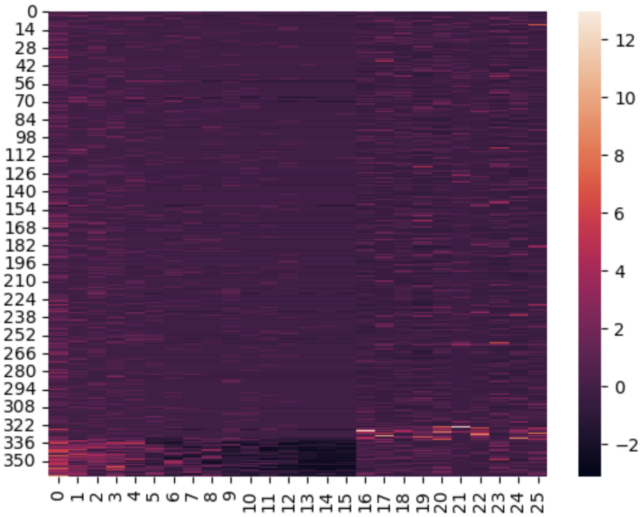

Figure 5: Heatmap-Based Integrated Analysis of Subnetwork Mutational Signatures and miRNA Expression Variability in COAD Dataset

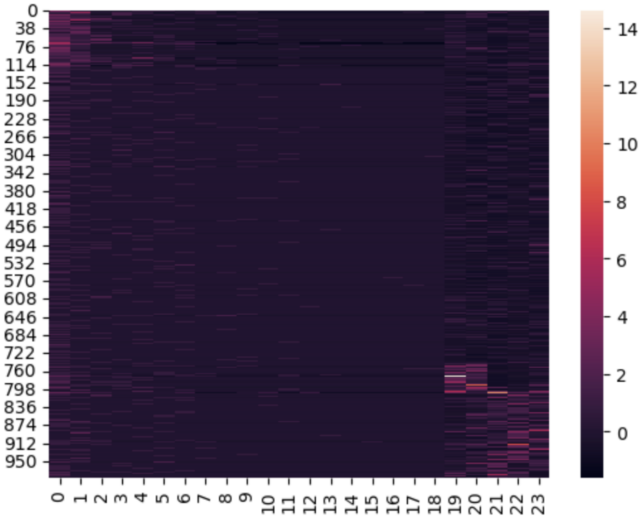

Figure 6: Heatmap-Based Integrated Analysis of Subnetwork Mutational Signatures and miRNA Expression Variability in BRCA Dataset

### 2.2.6 Unsupervised K-Means Clustering

A basic distance-based clustering approach seems feasible for the given dataset, given that the points are quite distinct. The parameter $K$ representing the number of clusters was varied to obtain the knee point of the graph where the Y-axis represented the variance of each cluster. Furthermore, we used the silhouette scores to determine the optimal K. It was found that the optimal K was 3 for the COAD dataset and seemed to be unstable for the BRCA dataset (in the range of 5 to 10), so we took it as 5, equal to the number of classes in the gold standard. Using the method, clustering predictions were made for the various points.

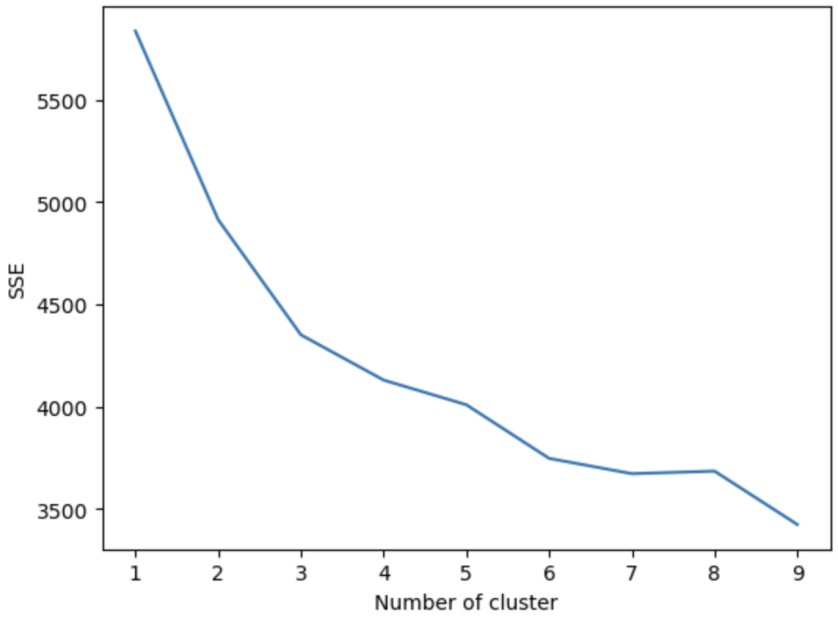


Figure 7: The SSE metric plotted against the hyperparameter $K$ in the K-means algorithm (COAD Dataset)

## 2.3 Method 2: Development of a Soft Clustering Algorithm using SNV Data via Generation of a Consensus Matrix

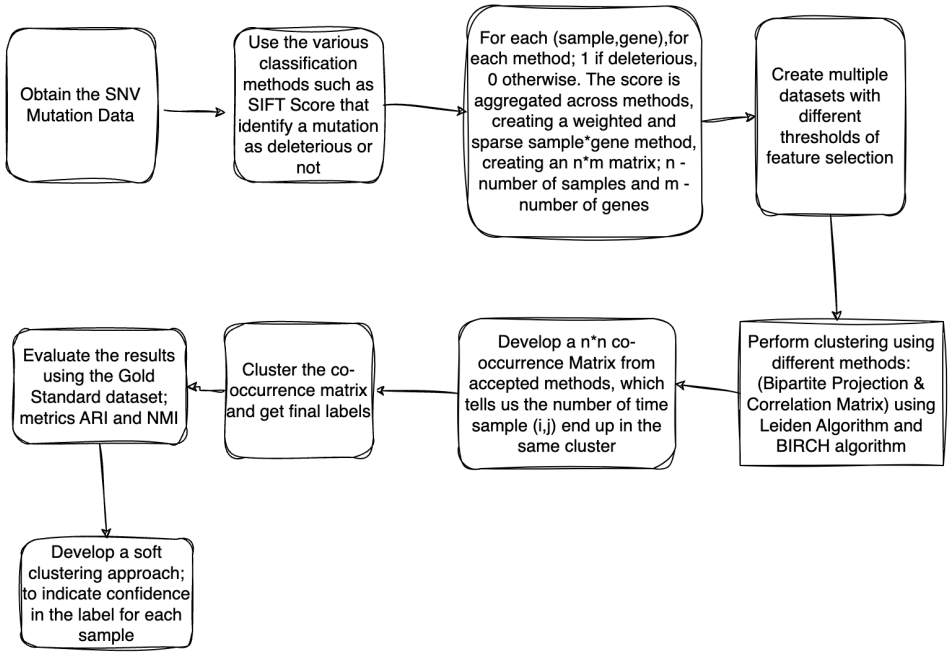


Figure 8: Overview of Method 2

### 2.3.1 Generation of feature matrix

Our goal is to construct the Feature Matrix using multiple predictions of the nature of the mutation. Our rationale is that genes predicted to be deleterious by more methods should have higher weights. The idea is to aggregate predictions from various methods, assigning a score of 1 if a gene is predicted deleterious and 0 if not, then summing these scores. This approach results in a weighted sample-gene matrix, where each gene's weight reflects the consensus of multiple predictive methods.

### 2.3.2 Feature Selection

We utilized both unsupervised feature selection by quantifying mutual information across different thresholds and supervised feature selection using the random forest method. The rationale was to observe the impact of varying degrees of feature selection on the clustering output. As a result, we generated multiple datasets with differing numbers of features, enabling a comprehensive analysis of how feature selection affects clustering performance.

### 2.3.3 Adjacency Matrix Generation

We employed two methods to generate adjacency matrices: bipartite projection and a correlation matrix via the Tanimoto coefficient. The rationale was to evaluate the impact of different data manipulation techniques on the resulting network structure. This approach resulted in two distinct adjacency matrices, providing a basis for comparing how each method influences the connectivity and clustering outcomes.

### 2.3.4 Clustering Algorithm

We employed the Leiden algorithm, a graph clustering method, and BIRCH, a hierarchical clustering algorithm. The rationale was to leverage the strengths of different clustering techniques to uncover diverse patterns within the data. This approach provided a comprehensive perspective on clustering outcomes. As a result, we obtained clusters from different methods across various datasets.

### 2.3.5 Creation of a consensus matrix

The goal is integrating information from well-performing clustering methods to achieve a robust consensus. By selecting the best-performing methods, we aim to track the frequency with which each pair of samples $(i, j)$ appears in the same cluster across different clustering outcomes. A higher frequency indicates a greater co-existence, suggesting a stronger likelihood that samples $i$ and $j$ belong to the same subtype. This process results in an $n$n×n co-occurrence matrix, where each element $(i, j)$ represents the number of times samples $i$ and $j$ are clustered together. We then perform clustering on this consensus matrix.

### 2.3.6 Soft Clustering Metric

We formulate the following metric.

$$\text{score}_{i,k} = \frac{\sum_{j \in k} A[i,j]}{\sum_{j} A[i,j]} \tag{1}$$

**Definitions:**

- $A$ - The co-occurrence matrix
- $i$ - The sample of interest
- $k$ - The cluster of interest on clustering the co-occurrence matrix
- $j$ - Any other sample

The formula calculates the score for a sample i belonging to cluster k. The numerator sums the co-occurrence counts of sample i with all samples j in cluster k. The denominator sums the co-occurrence counts of sample i with all other samples. This score reflects how frequently sample i co-occurs with cluster k, normalized by its overall co-occurrence frequency.

### 2.4 Integration of methods

Each of the two methods provides distinct sets of features that characterize a sample. One method focuses on functionally deleterious SNVs and soft clustering, while the other emphasizes miRNA profiles and DCMs. The goal was always to combine the subtype characteristics of both methods to achieve a more comprehensive classification. We would also highlight that both sets of approaches complement each other, offering differing levels of nuance. Method 1 provides a broad overview of dysregulated subnetworks, while Method 2 identifies specific genes dysregulated within those subnetworks. For example, fine-grained differences in mutations between genes cannot be detected in a subnetwork-based method such as Method 1, while Method 2's approach allows us to compare mutational patterns of PIK3CA and TP53 genes even though they belong to the same subnetwork. We created a consensus matrix to integrate these methodologies, giving equal weight to clusters from both approaches. This consensus matrix allows us to leverage the strengths of each method, potentially resulting in a more robust and accurate identification of cancer subtypes.

## 3 Results

### 3.1 ARI and NMI Discussion

The cluster prediction classes were compared with the ground truth data distribution using the ARI (Adjusted Rand Score) and NMI (Normalised Mutual Information Score). For the COAD dataset, the performance of our clustering approach exceeded benchmarks, with an ARI score of 0.354 and an NMI score of 0.384. On the BRCA dataset, the ARI was 0.13, and the NMI was 0.114 (poor with respect to benchmarks).

| Comparison of Methods | | | |
|---|---|---|---|
| Metric | Method 1 | Method 2 | Benchmark |
| ARI (BRCA) | 0.13 | **0.14** | 0.45 |
| NMI (BRCA) | 0.114 | **0.13** | 0.34 |
| ARI (COAD) | **0.354** | 0.28 | 0.32 |
| NMI (COAD) | **0.384** | 0.37 | 0.25 |

Table 1: Comparison of methods ARI and NMI vs Benchmarks

### 3.2 Description of Clusters by Method 1 for BRCA

- **Cluster 1:** Represents the *Her2* class, characterized by a variation in the DCM profile in the Subnetworks 0 and 2. Some of the important genes of these subnetworks include *TP53*, *PIK3CA*, *RYR3* and *TBX15*.
- **Cluster 2:** Represents the *LumA* class, with the maximum number of elements.
- **Cluster 3:** Represents the *Basal* class, characterized by a variation in the expression profiles of miRNAs *hsa-mir-301b* and *hsa-mir-429*.
- **Cluster 4:** Represents the *Normal* class, with the fewest data points.
- **Cluster 5:** Represents the *LumB* class, characterized by differences in the expression profiles of miRNAs *hsa-mir-190b*, *hsa-mir-375* and *hsa-mir-592*.

### 3.3 Description of Clusters by Method 1 for COAD

- **Cluster 1:** Represents the CIN class with the maximum number of data points.
- **Cluster 2:** Represents the GS class and is particularly characterized by variances in miRNA expression data.
- **Cluster 3:** Largely correlates to the *MSI* class, characterized by variations in the DCM signatures across all subtypes, but also characterized by positive variations among subnetworks 0-4 and negative variations among the subnetworks 6-16. The exact genes present in each of these subnetworks have been provided in the code repository.

### 3.4 Description of Clusters by Method 2 for BRCA

- **Cluster 1:** Characterized by *TP53* mutations; along with *TTN*, *MUC16*, *USH2A*, *DMD*, *FAT3*
- **Cluster 2:** Characterized by *PIK3CA* mutations, along with *TTN*, *CDH1*, *MUC16*, *MAP3K1*, *RYR2*

- **Cluster 3:** Characterized by an absence of *TP53* and *PIK3CA* mutations; *TTN*, *MUC16*, *KMT2C*, *CDH1*, *RYR2*, *AKT1* and *SYNE1*
- **Cluster 4:** Characterized by the co-occurrence of *TP53* and *PIK3CA* mutations

The observation highlights that the discrimination based on *TP53* and *PIK3CA* mutations aligns well with the supervised feature importance analysis. This suggests that these mutations play a significant role in defining the mutational subtypes of BRCA.

### 3.5 Description of Clusters by Method 2 for COAD

- **Cluster 1:** Dominated by *APC* mutations; also present are *TP53*, *PIK3CA*, *KRAS*, and *TTN*; *ZHFX4*, *SYNE1*, *MUC16*, *FAT4* - unique high-frequency mutations.
- **Cluster 2:** Dominated by *TTN* mutations, with *APC* and *KRAS* at lower frequencies.
- **Cluster 3:** Dominated by *TP53* mutations; *APC* and *KRAS* at lower frequencies; *DNAH5*, *FBXW7*, and *CACNA1E* - unique high-frequency mutations.
- **Cluster 4:** Characterized by a very limited number of traditional oncogenes such as *TP53*, *KRAS*, and *APC* mutations; includes *TTN*, *MACF1*, *MUC16*, *PCLO*, *OBSCN*, *CSMD3*, *KMT2D*, *ANKRD11*, *KMT2B*.

Cluster analysis of COAD highlights distinct genetic profiles. Cluster 1 is notable for high frequency of traditional oncogene mutations like *APC*, *TP53*, and *PIK3CA*, and with unique mutations. Cluster 2 primarily has *TTN* mutations, while Cluster 3 is dominated by *TP53* mutations and unique high-frequency mutations. Cluster 4 shows limited traditional oncogenes but includes a variety of other significant mutations, indicating diverse genetic pathways in COAD.

**GitHub Link:** github.com/nikhilanand03/CS6024CancerSubtypesClustering

## 4 Discussion

### 4.1 Improvements on previous works

While Dao et al.'s approach [10] used a density-based clustering method to detect subnetworks that showed high similarity in differential expression profiles among a small group of samples, they only evaluated their results on data determining whether or not cancer would metastasise rather than a holistic analysis of commonly known cancer subtypes. Furthermore, while they focused on differential expression data, we used SNV data to get results. Finally, we implemented an existing Louvain algorithm for labelling subnetwork markers and incorporated a reasonably different approach for determining subnetworks. Overall, the dataset used and the methodology followed for the first method were inspired by this paper. However, it was heavily modified since this research aims to perform unsupervised clustering of data points efficiently.

### 4.2 Progress made after Project Presentation

Earlier, we presented some initial findings on a consensus-based clustering approach, algorithms for determining subnetworks in a PPI graph, and supervised learning algorithms for identifying relevant features (genes). However, we still needed to cover the proper integration of these approaches. In the report, we have therefore made some significant additions. Firstly, the Louvain algorithm has been implemented to detect subnetworks within graphs. Furthermore, the computation of DCM signatures has been introduced in the report to properly integrate mutational information, subnetwork information, and supervised feature extraction to create an optimal clustering approach, as outlined in Method 1. Finally, we explored various interpretability aspects in further detail and benchmarked our method's results against existing methods.

## 5 Conclusion

In this work, we developed a method to classify cancer subtypes based on three key aspects: functionally deleterious SNVs, differential cumulative mutations, and miRNA profiles, along with soft clustering techniques. The results for BRCA, while not aligning well with traditional gold standard labels, reveal significant mutational signatures in clusters formed by p53 and PIK3CA mutations. Our approach meets or exceeds literature benchmarks for COAD, indicating that incorporating SNV and subnetwork dysregulations enhances our ability to identify true labels. This supports our working hypothesis and provides a robust biological explanation for the same.

# 6 Future Work

We decided to use supervised learning to predict the most important genes due to the graph algorithm's constraints for determining subnetworks. When the set of genes goes up to 18000, the edges increase exponentially, making it impossible to process. However, a more efficient algorithm can be developed, such as recursively processing subsets of genes and extracting subnetworks using the Louvain algorithm, as mentioned earlier in the report. We would take 900 nodes at a time and find subnetworks within the subset of 900 nodes, and we can finally draw relationships to obtain subnetworks for the entire 18000-node space. Doing this, we could potentially use the entire set of 18000 genes to predict subnetworks and generate the number of mutations for each sample in each subnetwork before proceeding to later stages of the algorithm. Since a lot more data is taken into consideration, we could obtain better final predictions. While a preliminary effort is available in the code repository, we hope to complete this approach for a more holistic analysis in future works.

# 7 Acknowledgements and Author Contributions

We thank Dr. Manikandan Narayanan for allowing us to pursue this project through the CS6024 course. We also thank him for his continuous mentorship, support, and guidance at all project phases. The clarity we received after talking to him helped us break through various stalemates we encountered during the project.

### 7.0.1 Author Contributions

- **Problem Statement Formulation** - Equal Contribution
- **Literature Review** - Equal Contribution
- **Methodology Ideation & Formulation** - Equal Contribution
- **Method 1 Implementation** - Nikhil
- **Method 2 Implementation** - Rajagopal
- **Final Report Writing** - Equal Contribution
- **Project Presentation** - Equal Contribution

# 8 Extra Credit - The Blend of Biology and Algorithms

We were fortunate that this project allowed us to work at the intersection of our degrees. We are pursuing an Integrated Disciplinary Dual Degree; our degree will consist of a Bachelor's in **Biological Engineering** and a Master's in **Data Science**. Our learnings while pursuing our degree thus allowed us to have a firm footing on both a biological aspect and an algorithm development aspect, allowing us to attempt this project from a truly interdisciplinary perspective.

# References

[1] Park J, Lee JW, Park M. Comparison of cancer subtype identification methods combined with feature selection methods in omics data analysis. BioData Min. 2023 Jul 7;16(1):18. Doi: 10.1186/s13040-023-00334-0. PMID: 37420304; PMCID: PMC10329370.

[2] Lock EF, Dunson DB. Bayesian consensus clustering. Bioinformatics 2013;29(20):2610–6

[3] Zhang S, Liu CC, Li W, et al. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res 2012;40:9379–9391.

[4] Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proceedings of the National Academy of Sciences. 2013;110: 4245–4250. doi: 10.1073/pnas.1208949110

[5] Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010;26: i237–i245. doi: 10.1093/bioinformatics/btq182

[6] Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007;3:140. doi: 10.1038/msb4100180. Epub 2007 Oct 16. PMID: 17940530; PMCID: PMC2063581.

[7] Guo Y, Shang X, Li Z. Identification of cancer subtypes by integrating multiple types of transcriptomics data with deep learning in breast cancer. Neurocomputing. 2018;324: 20–30. doi: 10.1016/j.neucom.2018.03.072

[8] Dao P, Colak R, Salari R, Moser F, Davicioni E, Schönhuth A, Ester M. Inferring cancer subnetwork markers using density-constrained biclustering. Bioinformatics. 2010 Sep 15;26(18):i625-31. doi: 10.1093/bioinformatics/btq393. PMID: 20823331; PMCID: PMC2935415.