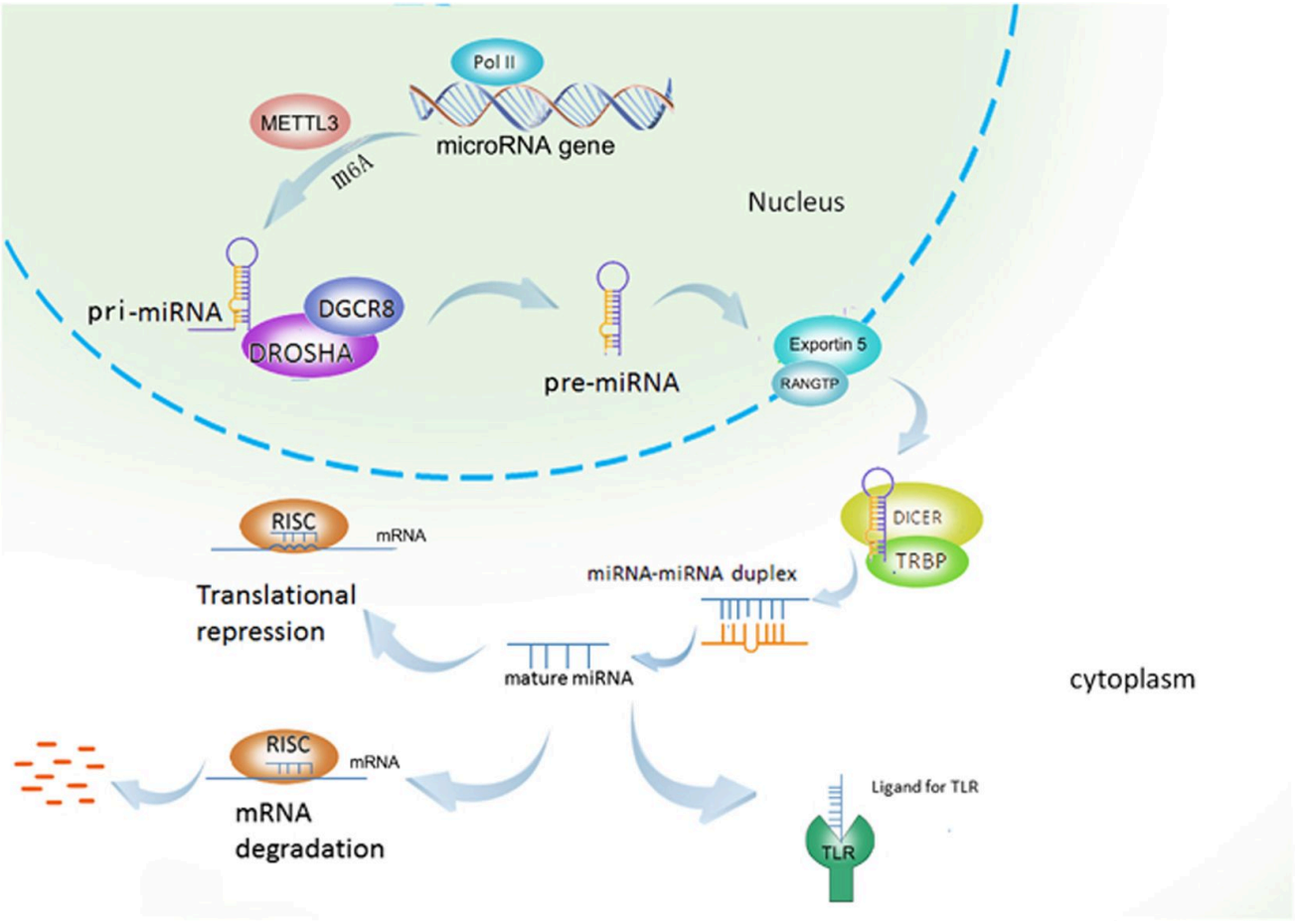
Analysis of microRNA data in relation to their therapeutic potential

Background

What is miRNA?

MicroRNA (**miRNA**) is a small non-coding RNA (~20-25 nucleotides) that regulates the expression of multiple target genes. As target recognition does not require full complementarity, a single miRNA can regulate multiple messenger RNAs (mRNAs). This combined effect by a single miRNA results in significant changes that can be measured.



miRNAs in Disease

Alteration in miRNA expression is associated with various diseases such as heart failure, cancer, and atherosclerosis to name a few. When a miRNA is differentially expressed in a specific disease, it can be considered as a "marker" for that disease. The table below provides a list of miRNAs that were found to be differentially expressed for various diseases:

Table 1: Differentially expressed microRNAs (signature microRNAs) in various diseases

Disease	Signature miRNA	Role
Hepatitis C	miR-122	Replication of HCV
Heart failure	miR-208	Necessary for cardiomyocyte hypertrophy
Inflammatory disease	miR-155	miR-155 regulates T-cell differentiation by regulating cytokine production
Cardiac fibrosis	miR-21	Promotes fibroblast survival and growth factor secretion
Neoangiogenesis	miR-92a	Negative regulator of endothelial cell proliferation, angiogenesis, and vascular repair
Metabolic disease	miR-33a	Regulates pathways controlling three of the risk factors of metabolic syndrome, namely levels of HDL, triglycerides, and insulin signaling
Myeloproliferative disease	miR-451	Upregulated during terminal erythroid differentiation and maturation
Cardiac injury	miR-15	Upregulated in response to ischemic damage
HCC	miR-21	Regulate MAP2K3 in HCC pathogenesis
Cancer	*Let-7*	Downregulated in several cancers and acts as a tumor suppressor and a regulator of terminal differentiation and apoptosis
Glioblastoma	miR-10b	miRNA not expressed in human brain and strongly upregulated in both low-grade and density-grade gliomas
Atherosclerosis	miR-33	Regulates HDL biogenesis and RCT via posttranscriptional repression of cholesterol efflux genes (ABCA1, ABCG1, Npc1). ABCA1 mediates the transport of cholesterol from peripheral tissues to apolipoprotein-1 and it is also important in the RCT pathway, where cholesterol is delivered from peripheral tissue to the liver, where it can be excreted into bile or converted to bile acids prior to excretion
Vascular disease	miR-145	Specific for VSMCs and determine the phenotype of VSMCs. miR-145 levels increase and are released into the plasma in response to vascular injury
Peripheral artery disease	miR-92	Overexpressed during ischemic injury, which in turn blocks angiogenesis and vessel formation
Kidney fibrosis	miR-21	Anti-apoptotic and that apoptosis leads to loss of tubular epithelial cells, decreased re-epithelialization, and sustained inflammation, thereby promoting kidney interstitial fibrosis

miRNA=MicroRNAs, RCT=Reverse cholesterol transport, ABC=Adenosine triphosphate-binding cassette, Npc1=Niemann-Pick C1, VSMCs=Vascular smooth muscle cells, HDL=High-density lipoprotein, HCC=Hepatocellular carcinoma, HCV=HCV=Hepatitis C virus

miRNAs as Therapeutics

Due to their role in various diseases, miRNAs are being investigated as therapeutics. Similar to drug development, this process requires several steps:

1. Identify a "marker" miRNA for disease of interest (via miRNA profiling)
2. Validation of "marker" miRNA (via loss/gain of function studies *in vitro* and *in vivo*)
3. Pharmacological analysis (via delivery studies and pharmacokinetics/pharmacodynamics)
4. Clinical trials (evaluation of drug efficacy and safety)

Exploratory Analysis of miRNA Data

For today's lecture, we will be analyzing expression levels of 18 miRNAs collected from 70 lung tissue samples. This dataset includes samples from both healthy and diseased individuals, with the disease being lung cancer. Our main tasks for today will be to:

1. Determine the optimal number of clusters for this data using k-means clustering
2. Visualize this data as a 2D scatter plot using principal component analysis
3. Identify which miRNAs are differentially expressed between healthy and diseased individuals using hierarchical clustering

Load and examine the data

To start, load the dataset (*miRNA_data.xlsx*) as a table and see what information is available. Based on the information provided, create a variable to store the patient IDs, a variable to store the health status for each patient, a variable to store the miRNA names, and a variable to store the miRNA expression data as a matrix:

```
% Load the dataset

% Create a variable for patient IDs

% Create a variable for patient health status

% Create a variable for miRNA names

% Create a variable for miRNA expression data
```

Plot a histogram (use the *hist* function) of the expression data for the first five miRNAs (make sure to include axis labels and a title):

```
% Histogram of miRNA expression data
```

Questions: What do you notice about the expression level for these miRNAs?

Determine the optimal k value using k-means clustering

I will first show a short demo of what we will be doing for this section:

```
% Demo for k-means clustering
idx = kmeans({miRNA data array},2);     % identify clusters with k-means clustering
s = silhouette({miRNA data array},idx); % determine silhouette values
s_score = mean(s)                       % calculate silhouette score
```

Now I want you to carry out these steps for k values ranging from 2 to 10 and store the silhouette scores you calculate. This will help us to quantitatively determine the best estimate for how many clusters the data separates into. Below is a base template to help you get started (you'll only need to modify the parts in curly brackets):

```
% Create a vector of k values ranging from 2 to 10
{vector of k values}

% For-loop to calculate the silhouette statistic for different k values
n = length({vector of k values});   % number of k values
s_score = zeros(n,1);               % variable to store silhouette statistic values
for i = 1:n
    % Use the kmeans function to cluster patients into k clusters
    idx = kmeans({miRNA data array},{k value});
    % Use the silhouette function to determine silhouette values
    s = silhouette({miRNA_data array},idx);
    % Calculate the silhouette score by taking the mean
    s_score(i) = mean(s);
end
```

```
table({vector of k values}',s_score)
```

Question: What is the value for the max silhouette statistic? Which k value does this correspond with?

Based on your answer to the question above, generate a silhouette plot:

```
% Silhouette plot based on best k value
```

Question: Are there any negative silhouette values? If so, what does this imply?

Using your variable for patient health status, assess how patients cluster based on your *idx* variable from k-means clustering:

```
% Cluster patients by health status using idx variable from kmeans
```

Question: What do you notice from this assessment?

Visualize the data in lower dimension using PCA

Now we'll apply principal component analysis (PCA) to visualize our data within a lower-dimensional space. Use the *pca* function to determine the PC matrix, the transformed data matrix, and a vector for the variance explained. Then generate a scatter plot of the transformed data for PC1 vs. PC2, using patient health status for group labeling (Note: be sure to normalize the data before doing PCA):

```
% Do PCA (determine PC matrix, transformed data matrix, and variance explained)

% Visualize transformed data on PC1 vs. PC2 plot (label = patient health status)
```

Question: Based on the plot of PC1 vs. PC2, how many distinct clusters can you identify? How much variance do PC1 and PC2 account for?

Create a table of miRNA gene names matched with the coefficients of the principal component that best separates our data:

```
% Create table of genes matched to coefficients of best PC
```

Question: Which miRNA has the greatest weight for this PC?

Identify differentially expressed miRNAs using hierarchical clustering

Create a clustergram of the expression data, using patient health status to label the rows and the miRNA gene names to label the columns. Also be sure to standardize all values by column:

```
% Create clustergram (label rows, label columns, and standardize by
% columns)
```

Question: Based on how the rows (patients) cluster, does this match our results from k-means clustering and PCA?

Generate the clustergram again, but this time use correlation as the distance metric for both rows and columns:

```
% Generate clustergram (label rows, label columns, standardize by columns,
% and use correlation as the distance metric for both rows and columns)
```

Question: How did the clustering change? Does this match our results from k-means clustering and PCA?

Question: Based on the patterns seen in the clustergram, which miRNAs are differentially expressed between healthy and diseased individuals?

Final Comments

The exploratory analysis that we did for today shows how unsupervised learning methods can be applied to assess miRNA data. Though the dataset we used was small, we were able to assess how some miRNAs were differentially expressed between healthy and diseased individuals. When considering miRNAs for therapeutic potential, one would typically compare thousands of miRNA profiles and carry out statistically rigorous steps to identify marker miRNAs. For the purposes of this lesson, only exploratory and visual assessment of the miRNA data was covered.