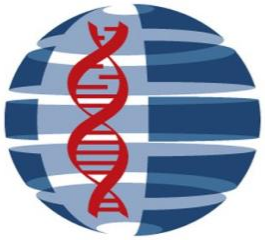
AN INTEGRATIVE EXPLORATORY ANALYSIS OF –OMICS DATA FROM THE ICGC CANCER GENOMES LUNG ADENOCARCINOMA STUDY



International Cancer Genome Consortium

SINJINI SIKDAR

DEPARTMENT OF BIOINFORMATICS AND BIOSTATISTICS

UNIVERSITY OF LOUISVILLE

CAMDA 2014

INTRODUCTION

- All agents that cause cancer (carcinogens) also cause a change in the DNA sequence.

- Many ways of observing biological data related to the DNA sequence - measuring gene expression, miRNA expression, protein expression, somatic copy number variation, and methylation profiles.

DATASETS

Gene →

No. of subject ids	No. of genes (after filtration)
131	15916

miRNA →

No. of subject ids	No. of miRNAs (after filtration)
379	709

Protein →

No. of subject ids	No. of proteins
237	139

DATASETS

Copy Number →

No. of subject ids	No. of chromosomes
383	24

Clinical →

No. of subject ids	Disease status	
395	Complete remission	progression

Methylation →

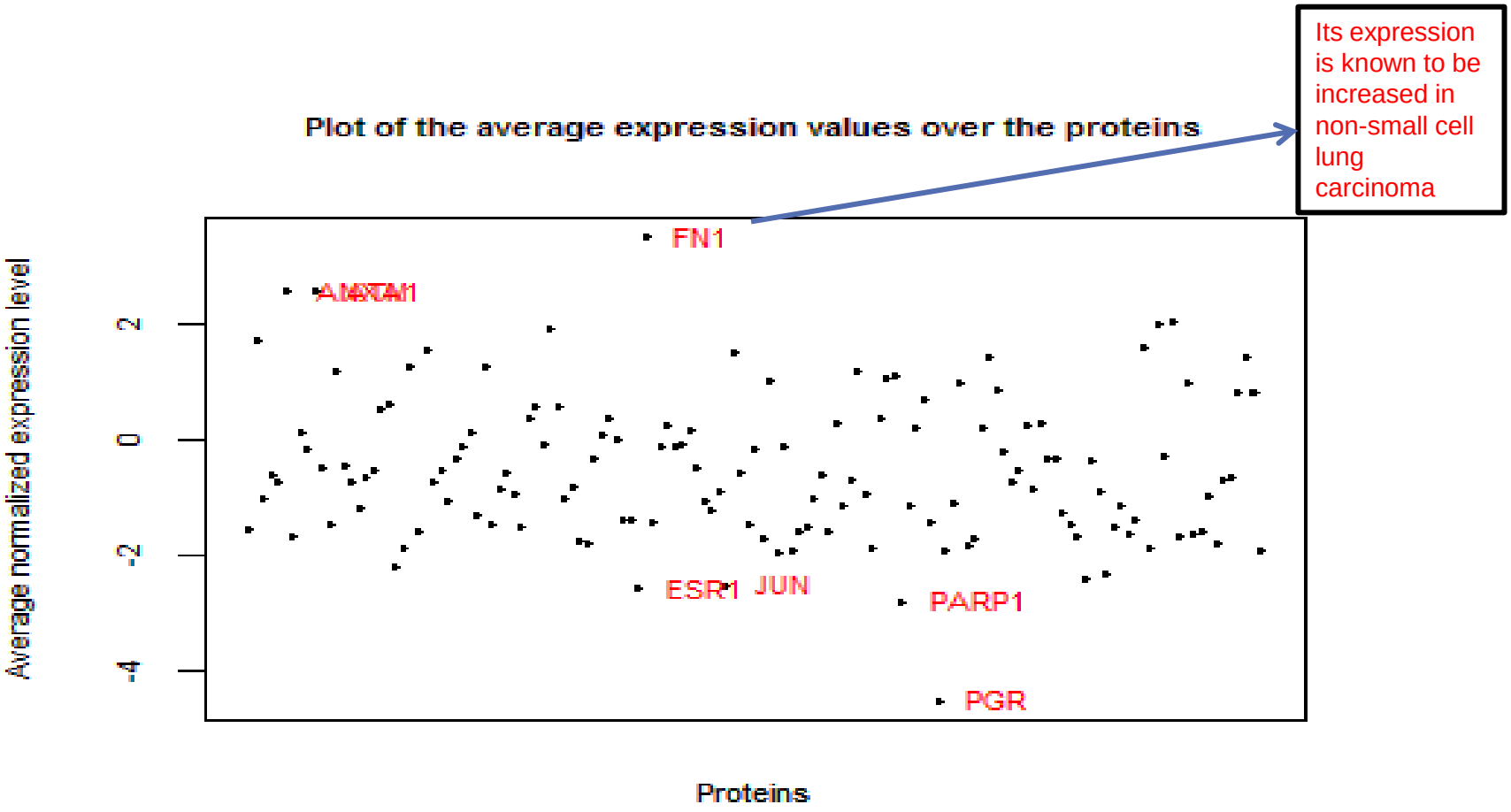
No. of subject ids	No. of chromosomes
382	24

SOME USEFUL PLOTS

For the CAMDA 2014 lung adenocarcinoma challenge data,

Initial exploratory plots of normalized expression values for genes, miRNAs, and proteins are made for each data set to obtain overall summaries of the data sets and to identify extreme values.

SOME USEFUL PLOTS

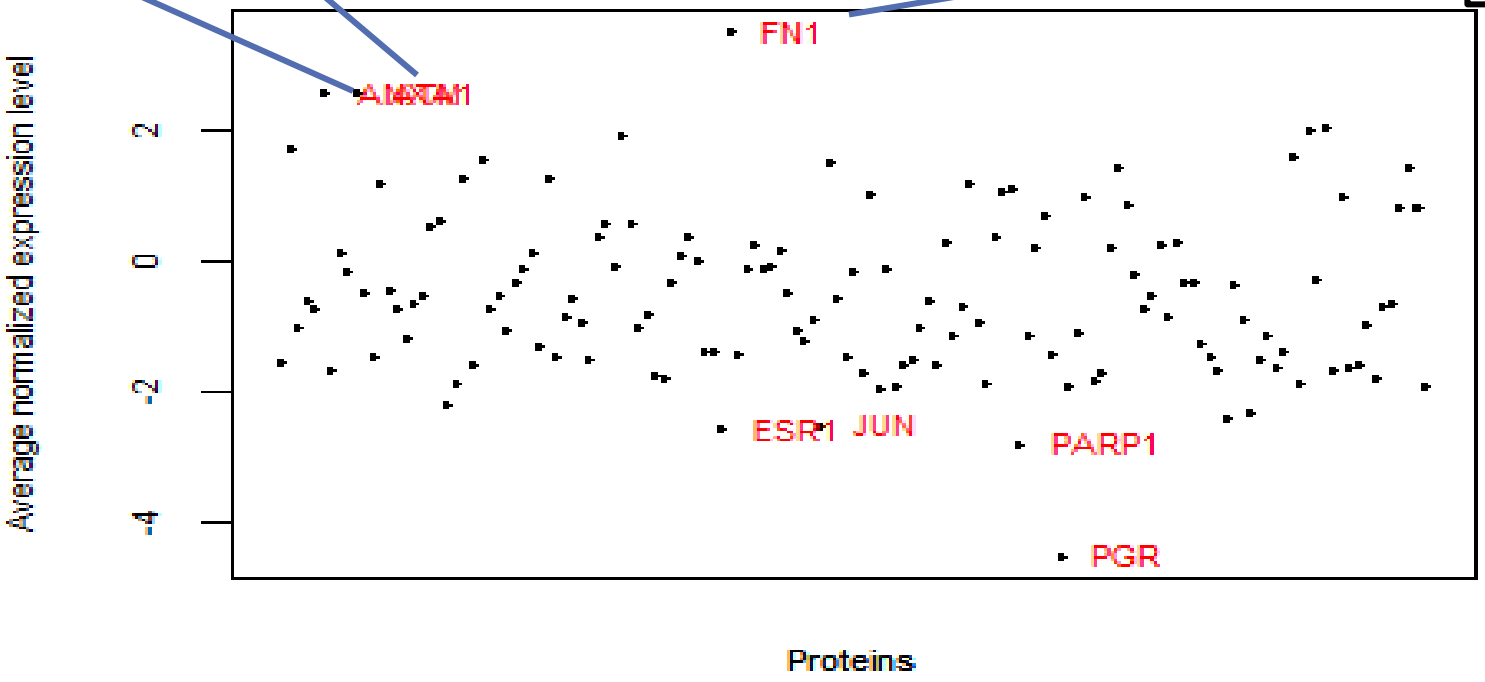


SOME USEFUL PLOTS

ANXA1 is upregulated in hairy cell leukemia

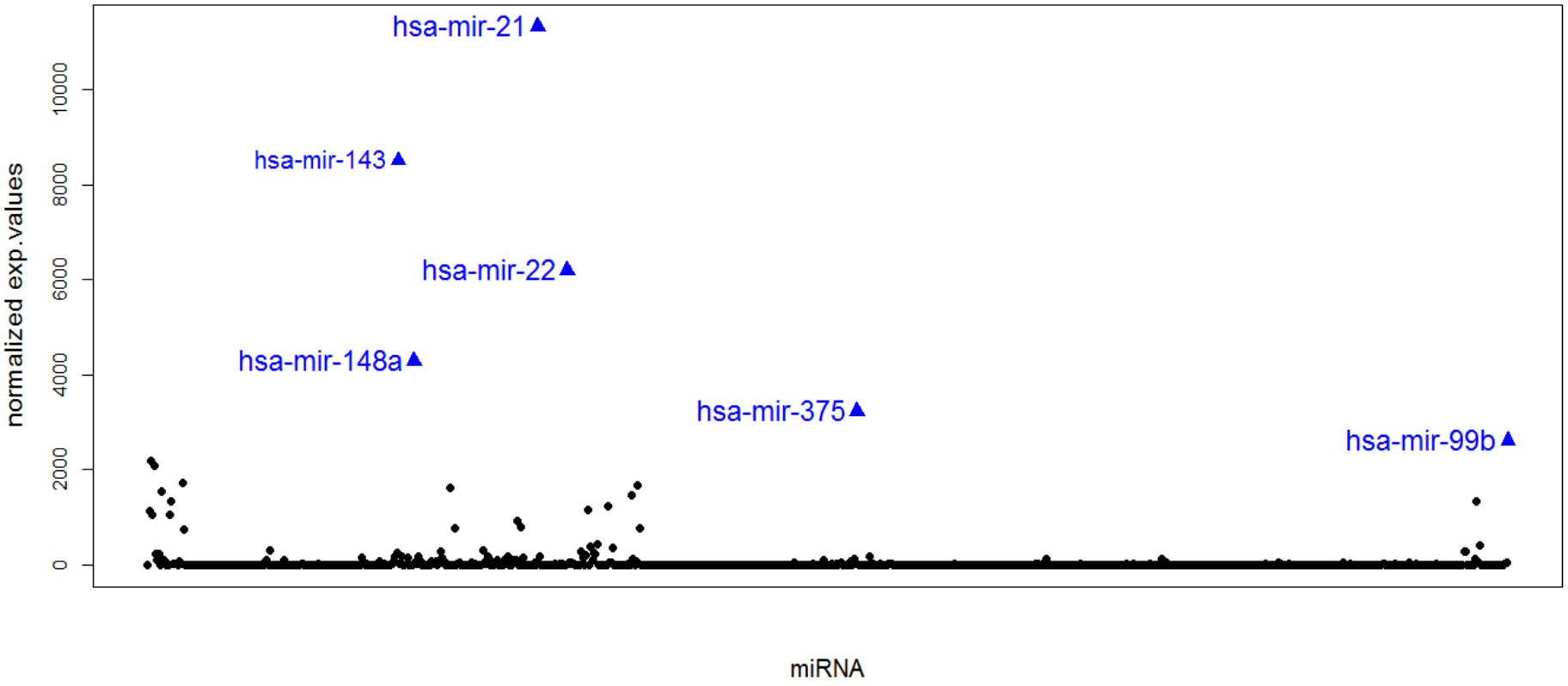
ATM is associated with Certain kinds of leukemias and lymphomas

Its expression is known to be increased in non-small cell lung carcinoma

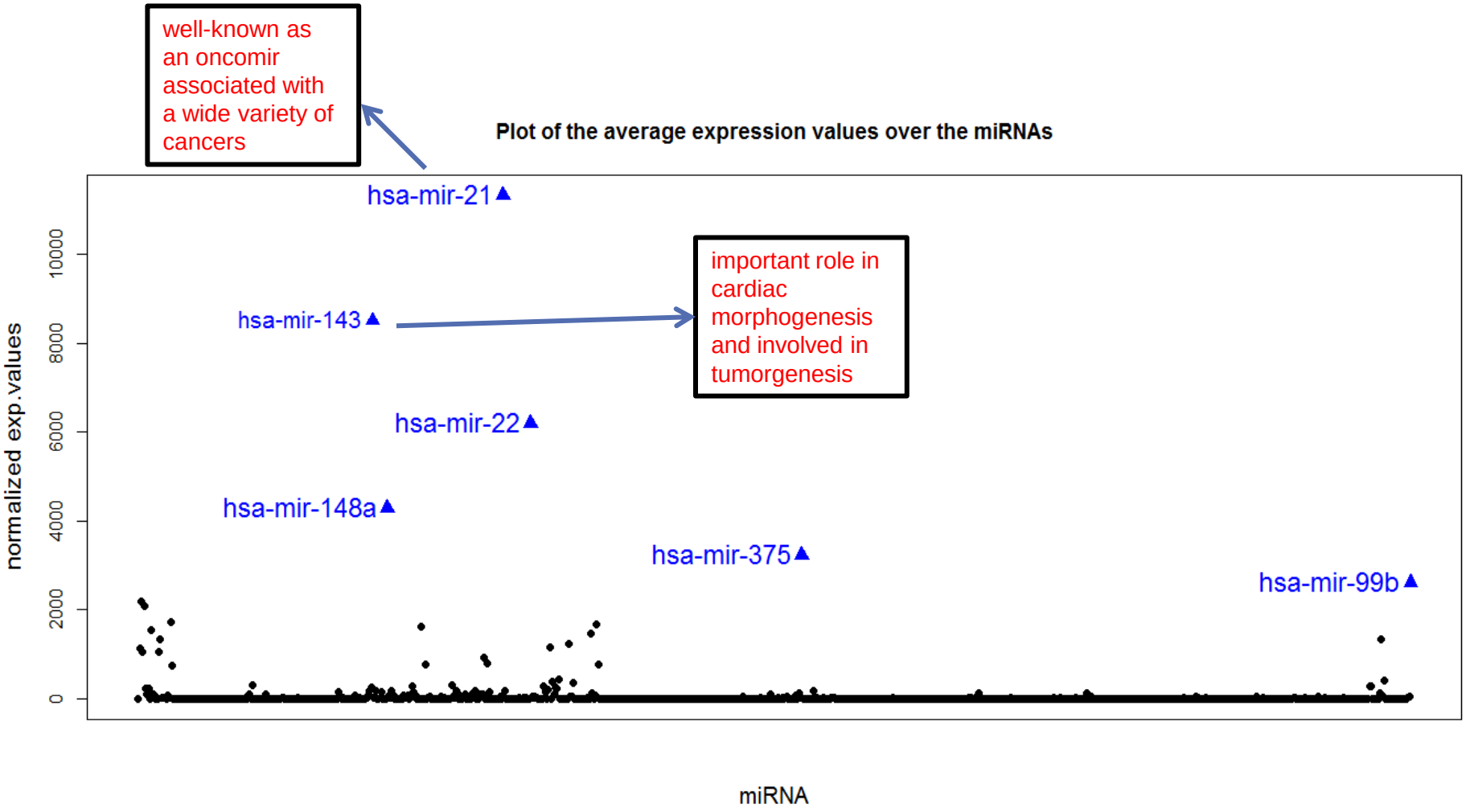


SOME USEFUL PLOTS

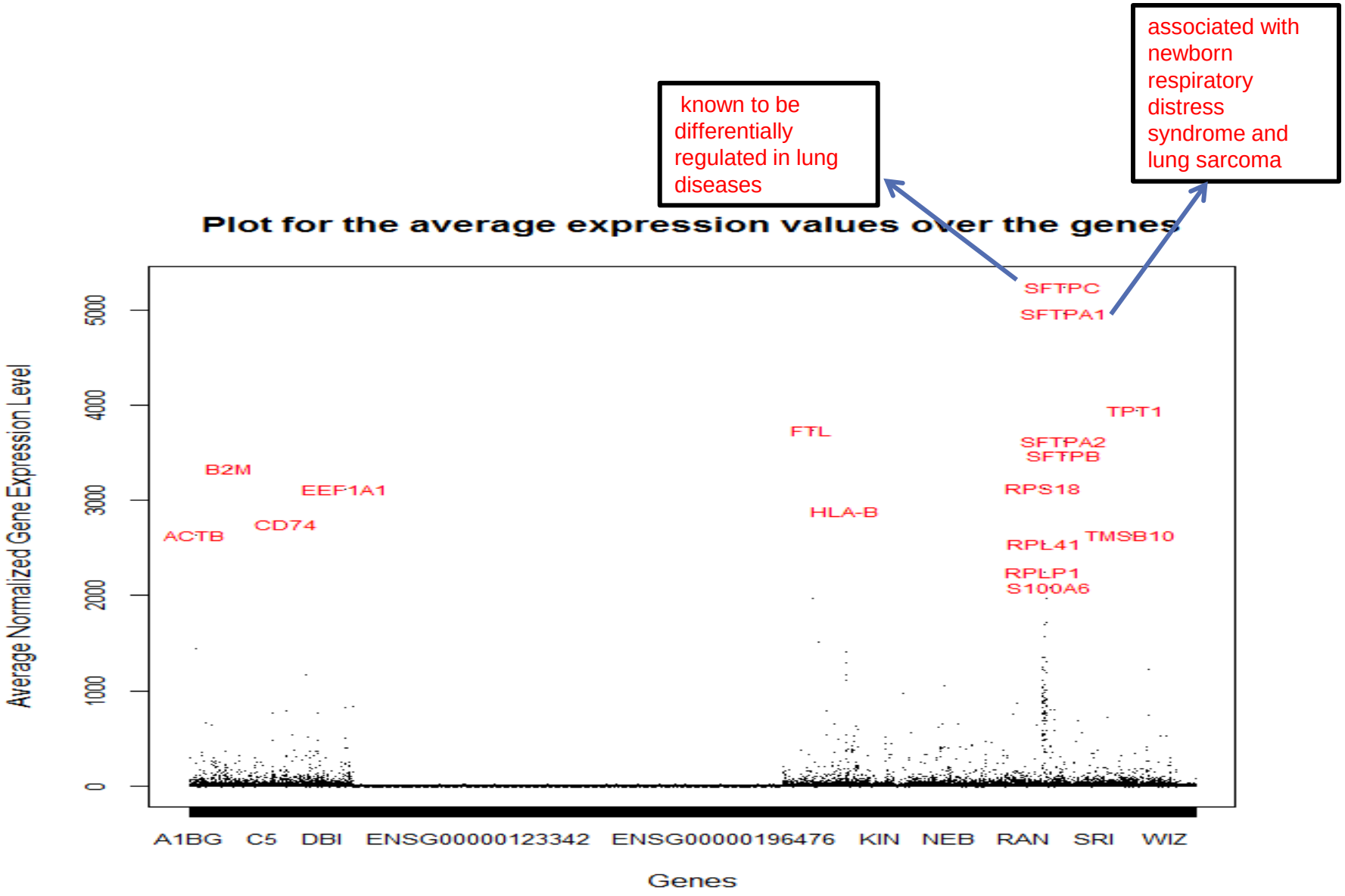
well-known as an oncomir associated with a wide variety of cancers



SOME USEFUL PLOTS



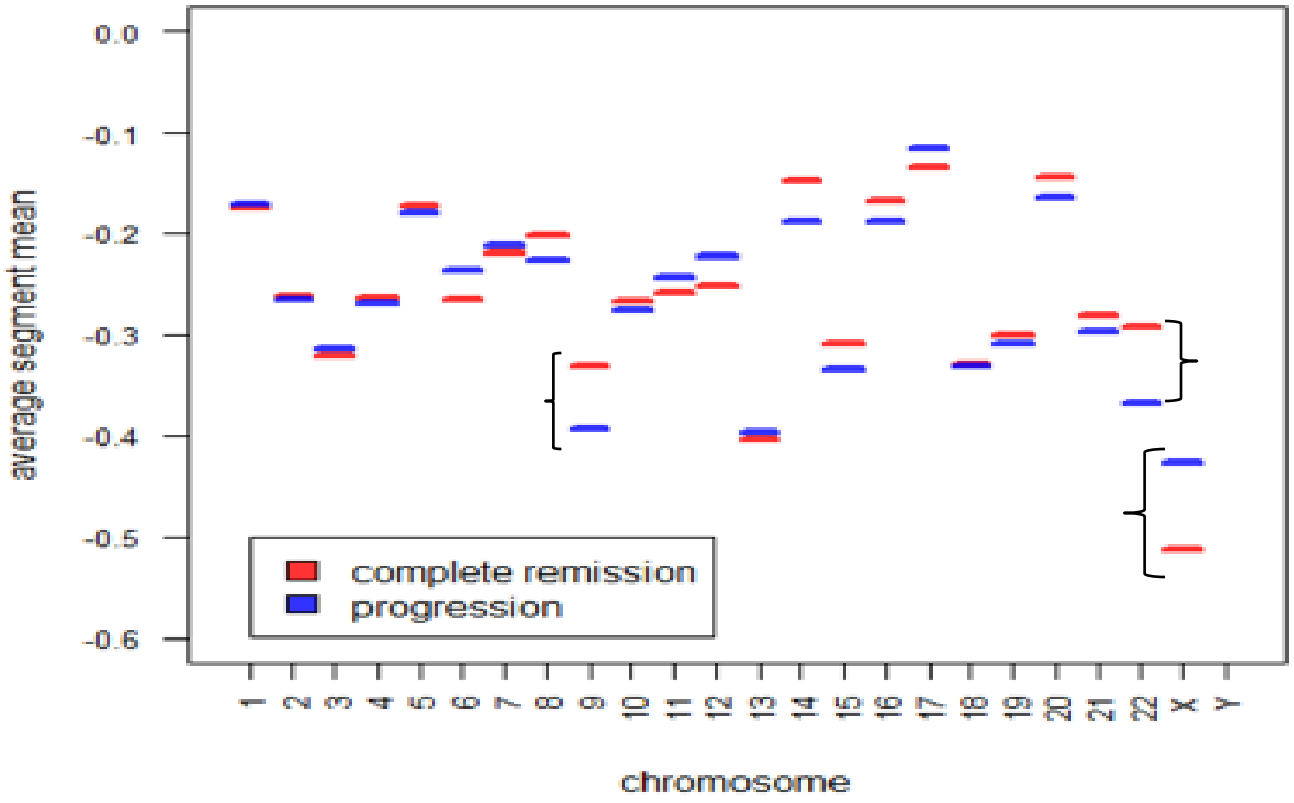
SOME USEFUL PLOTS



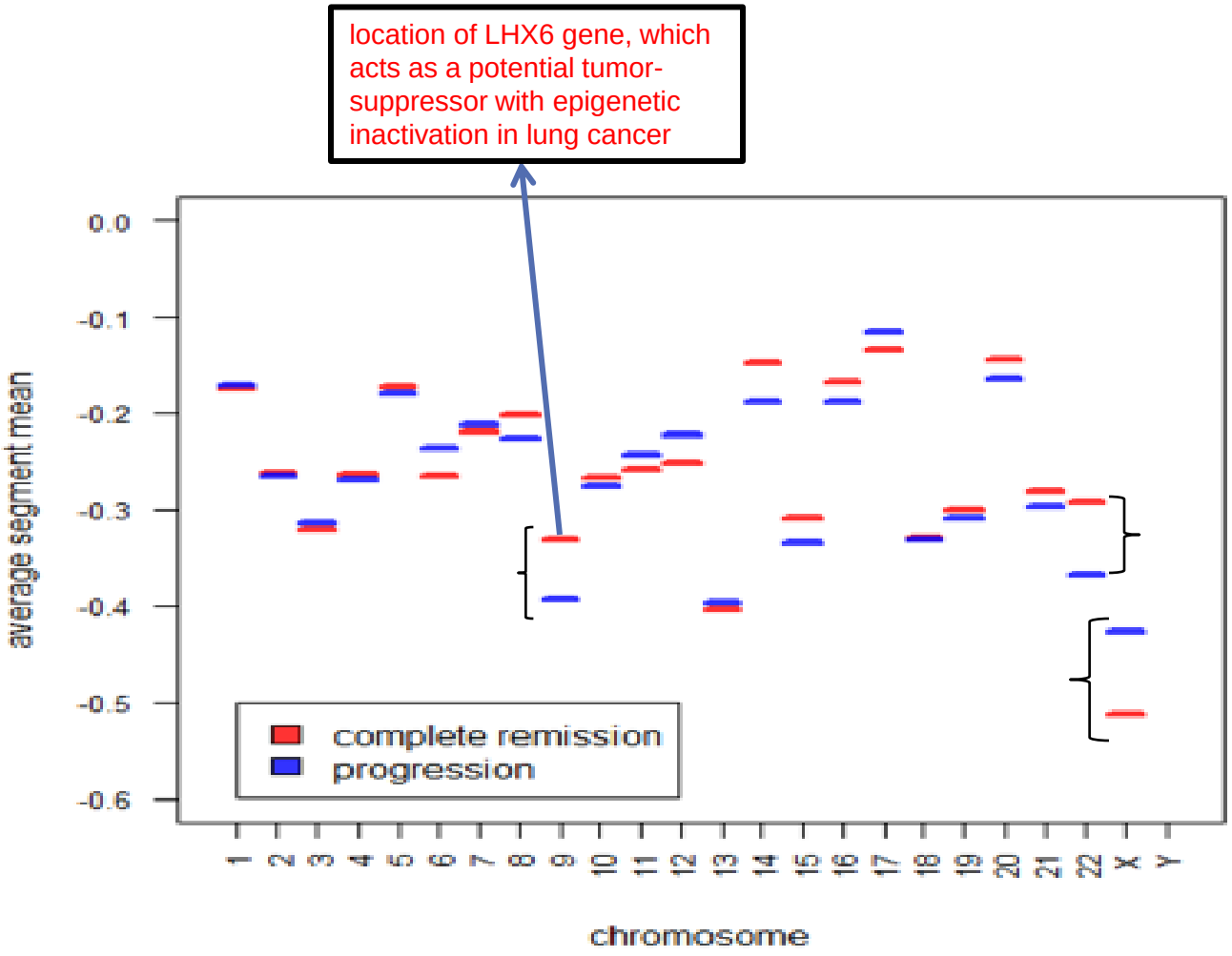
SOME USEFUL PLOTS

A plot of the average segment mean by chromosome for the progression and complete remission groups is also done.

SOME USEFUL PLOTS



SOME USEFUL PLOTS



MATERIALS AND METHODS

- For the CAMDA 2014 lung adenocarcinoma challenge data, we considered several methods of analyzing matched data on genes, miRNA, proteins, and copy number variation and also explored the methylation patterns.

- The methods included:

    1. Exploratory Cluster Analysis
    2. Prediction of Clinical outcome
    3. Correlation Analysis

1. EXPLORATORY CLUSTER ANALYSIS

We clustered the subject ids using

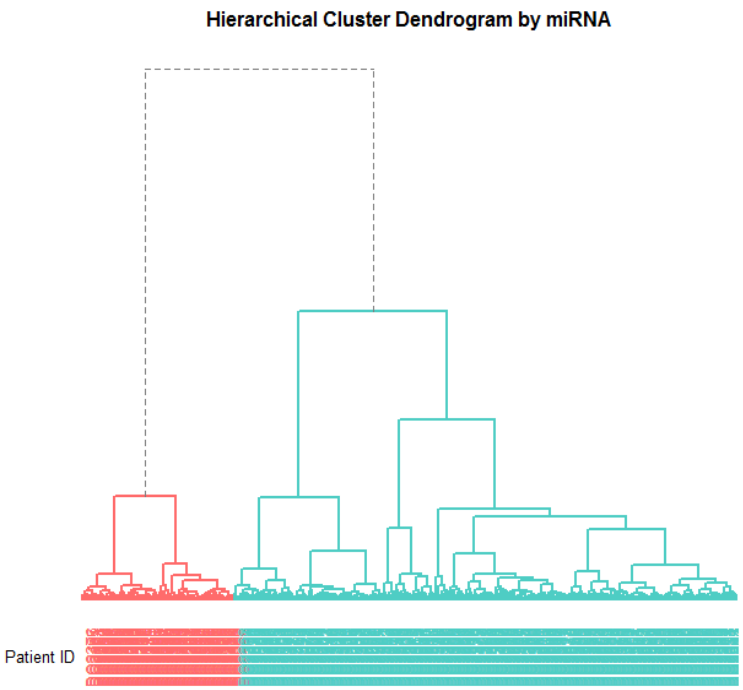
- Internal cluster validation with validation measures like connectivity, Dunn index and silhouette width.
- Clustering algorithms like hierarchical with ward linkage, k-means and PAM.

We combined the measures and determined the optimal clustering algorithm using Rank Aggregation (Pihur et al., Bioinformatics,2007; Pihur et al.,BMC Bioinformatics,2009).

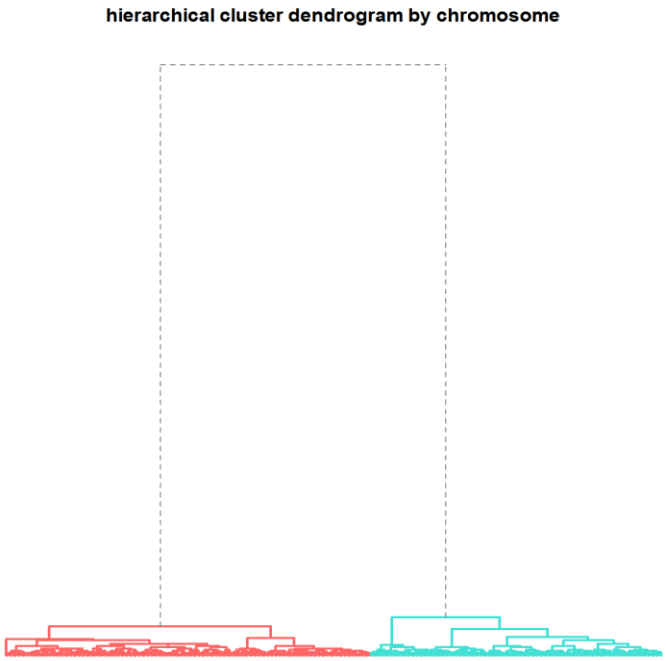
1. EXPLORATORY CLUSTER ANALYSIS

RESULTS

miRNA



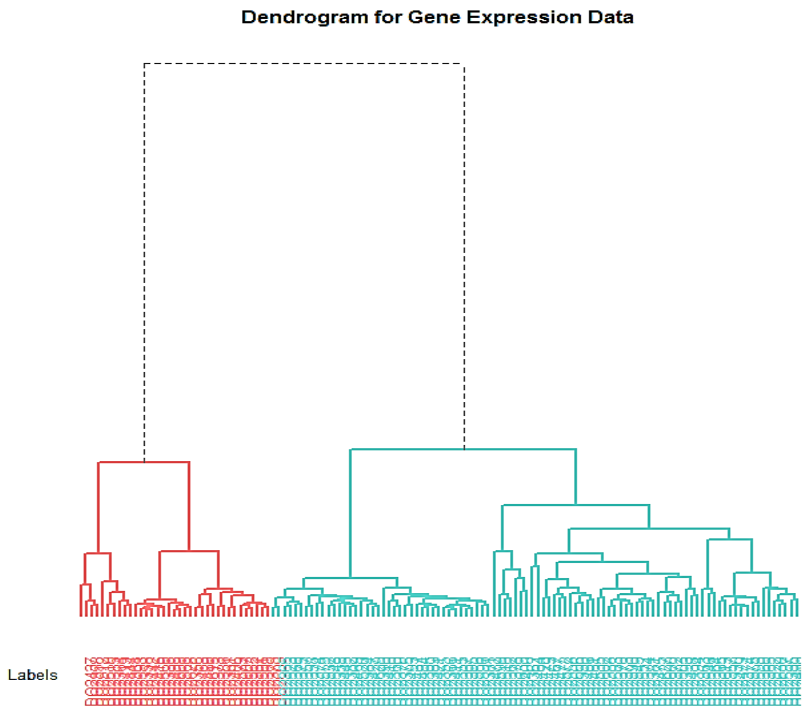
copy number



1. EXPLORATORY CLUSTER ANALYSIS

RESULTS

gene



protein

kmeans-2

1. EXPLORATORY CLUSTER ANALYSIS

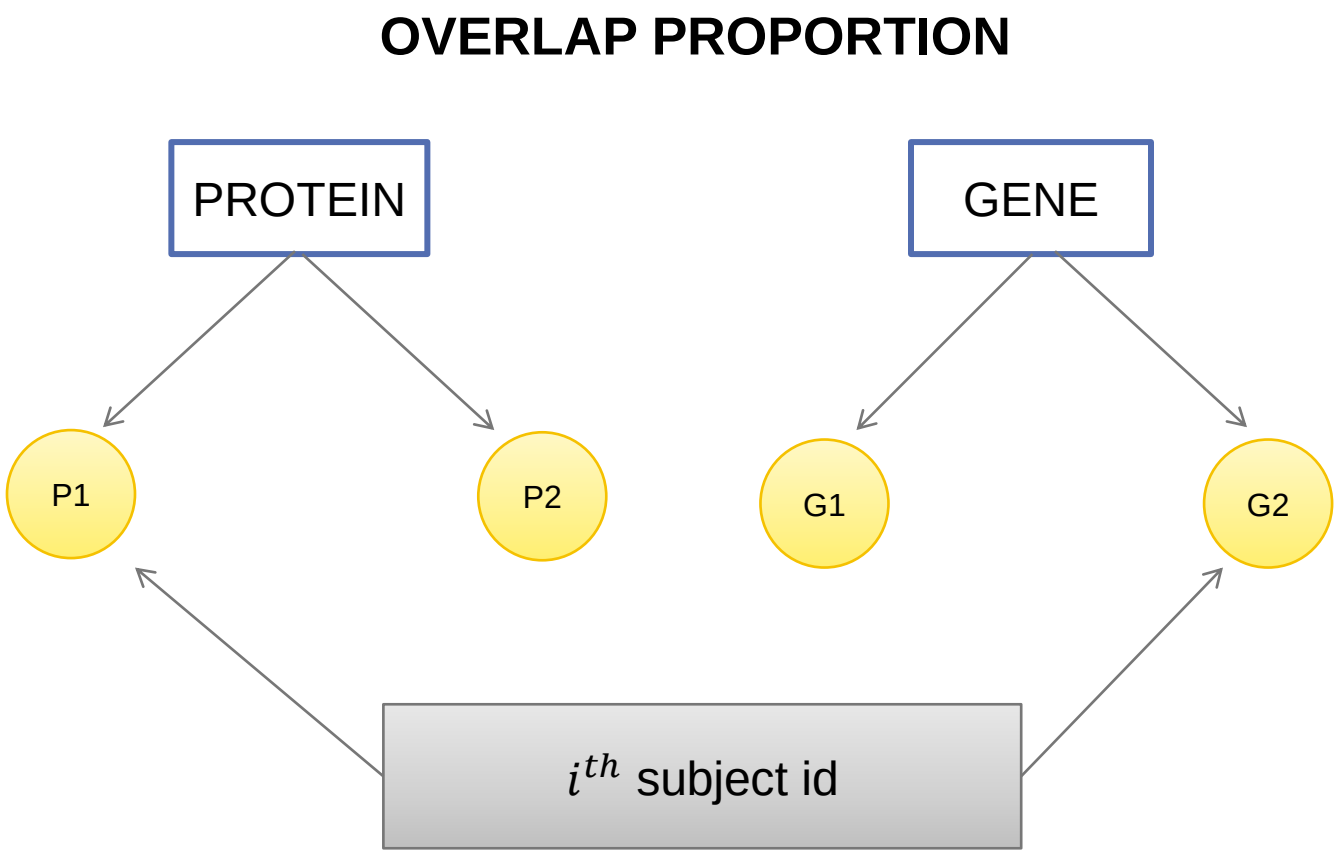
OVERLAP PROPORTION

To measure the similarity between the cluster of subjects, we calculated the overlap proportion for each pair of profiles using the formula

$$R_{j,k} = \frac{1}{n_{j,k}} \sum_{i=1}^{n_{j,k}} \frac{\left| C_{(i)}^{j} \cap C_{(i)}^{k} \right|}{\left| (C_{(i)}^{j} \cup C_{(i)}^{k}) \cap I_{j,k} \right|}$$

where $C_{(i)}^{j}$ is the set of subjects in the cluster containing the *i*th subject based on the *j*th profile, $I_{j,k}$ is the set of common subjects, $n_{j,k}$ is the number of common subjects in the two profiles.

1. EXPLORATORY CLUSTER ANALYSIS



1. EXPLORATORY CLUSTER ANALYSIS

OVERLAP PROPORTION

Table 1: Overlap proportions between the data sets using clusters				
	Gene	**miRNA**	**Protein**	**Chromosome**
miRNA	0.5239			
Protein	0.4350	0.3873		
Chromosome	0.3736	0.3821	0.3612	
Clinical data	0.4677	0.4353	0.3937	0.3576

1. EXPLORATORY CLUSTER ANALYSIS

OVERLAP PROPORTION

Empirically (as well as mathematically), it can be seen that the overlap proportion is expected to be roughly 1/3 if the group assignments are made randomly.

2. PREDICTION OF CLINICAL OUTCOME

For each data set, we have fitted a penalized logistic regression model for predicting clinical outcomes for disease status based on

1. Age
2. Gender
3. expression values or chromosomal segment means

The model is

$$\text{logit}(p_{j,i}) = \beta_{j,0} + \beta_{j,A} Age_{j,i} + \beta_{j,G} Gender_{j,i} + \beta_{j,1} X_{j,i,1} + \cdots + \beta_{j,m} X_{j,i,m} + penalty$$

where, for the ith subject based on the jth profile data set,

$p_{j,i}$ is the probability for progression of the disease

$X_{j,i,k}$ is the kth expression value or chromosomal segment mean

2. PREDICTION OF CLINICAL OUTCOME

- Fitted the model on each of the four profiles using elastic net regression where the parameters are selected through the 0.632+ Bootstrap method.

- Covariates are selected using the bootstrap samples with optimal values of the parameters.

- This process is repeated 1000 times, and a few top covariates, for each profile, are selected.

2. PREDICTION OF CLINICAL OUTCOME

The 20 most significant genes, miRNAs, and proteins are used for the correlation analysis in the next section.

3. CORRELATION ANALYSIS

Spearman's rank correlation coefficients are computed among the most important genes, proteins, miRNAs and chromosomes using the formula

$$\rho_{j,k} = 1 - (6 \sum_{i=1}^{n_{j,k}} d_{j,k,i}^2) / (n_{j,k}^3 - n_{j,k})$$

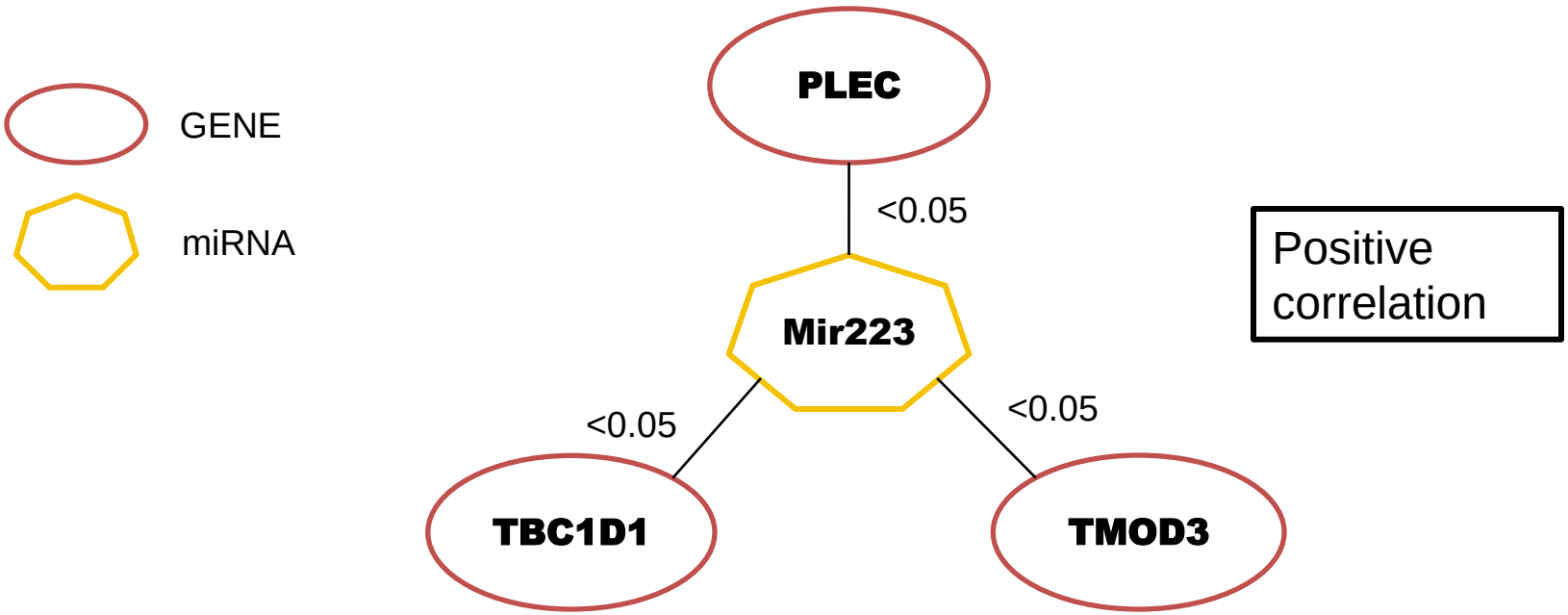
where $d_{j,k,i}$ is the difference between the ranks of the *i*th subject in the *j*th and *k*th profile data sets.

We estimated the p-values for these correlation coefficients, using the asymptotic t-approximation.

3. CORRELATION ANALYSIS

RESULTS

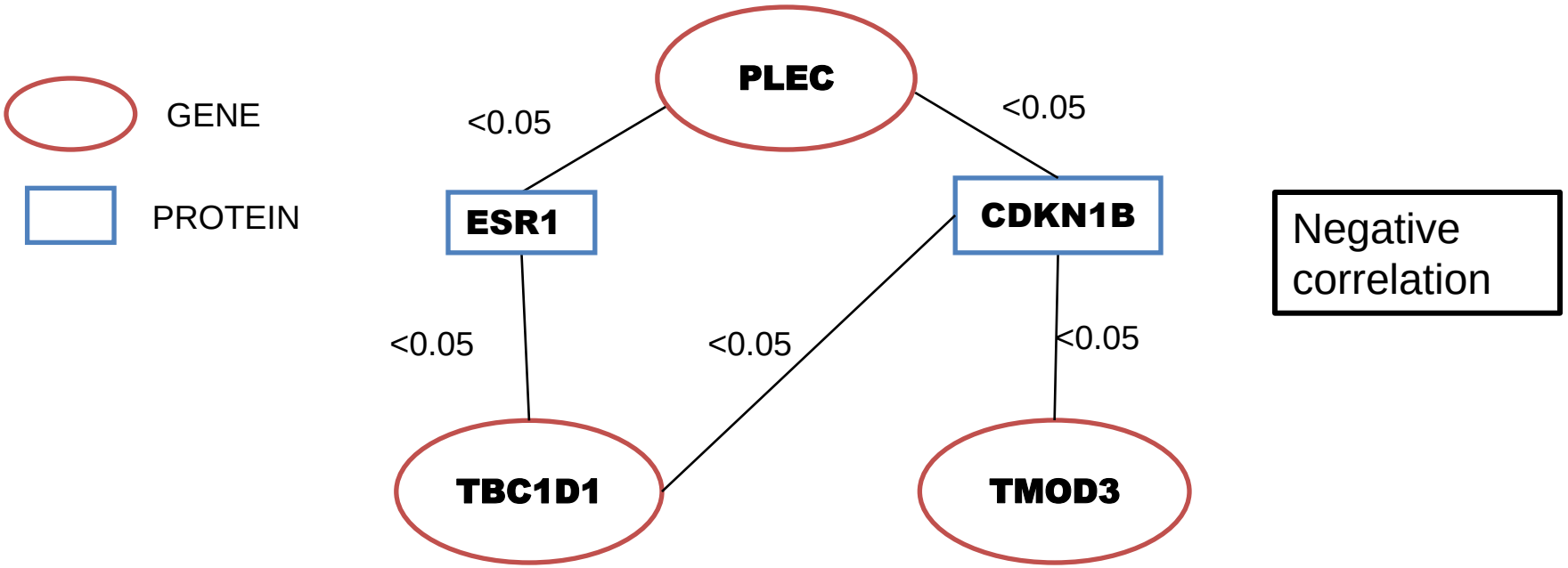
miRNA & Gene



3. CORRELATION ANALYSIS

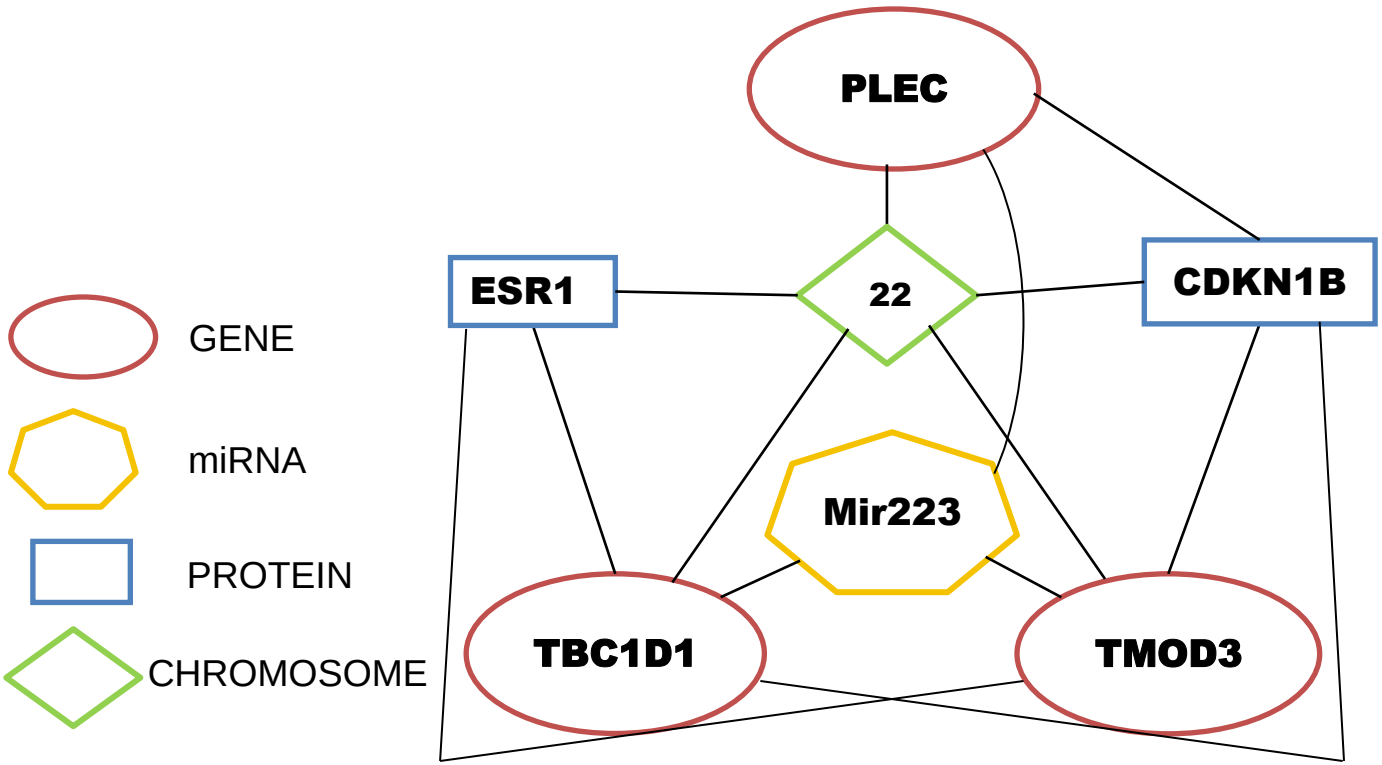
RESULTS

Protein & Gene



3. CORRELATION ANALYSIS

RESULTS

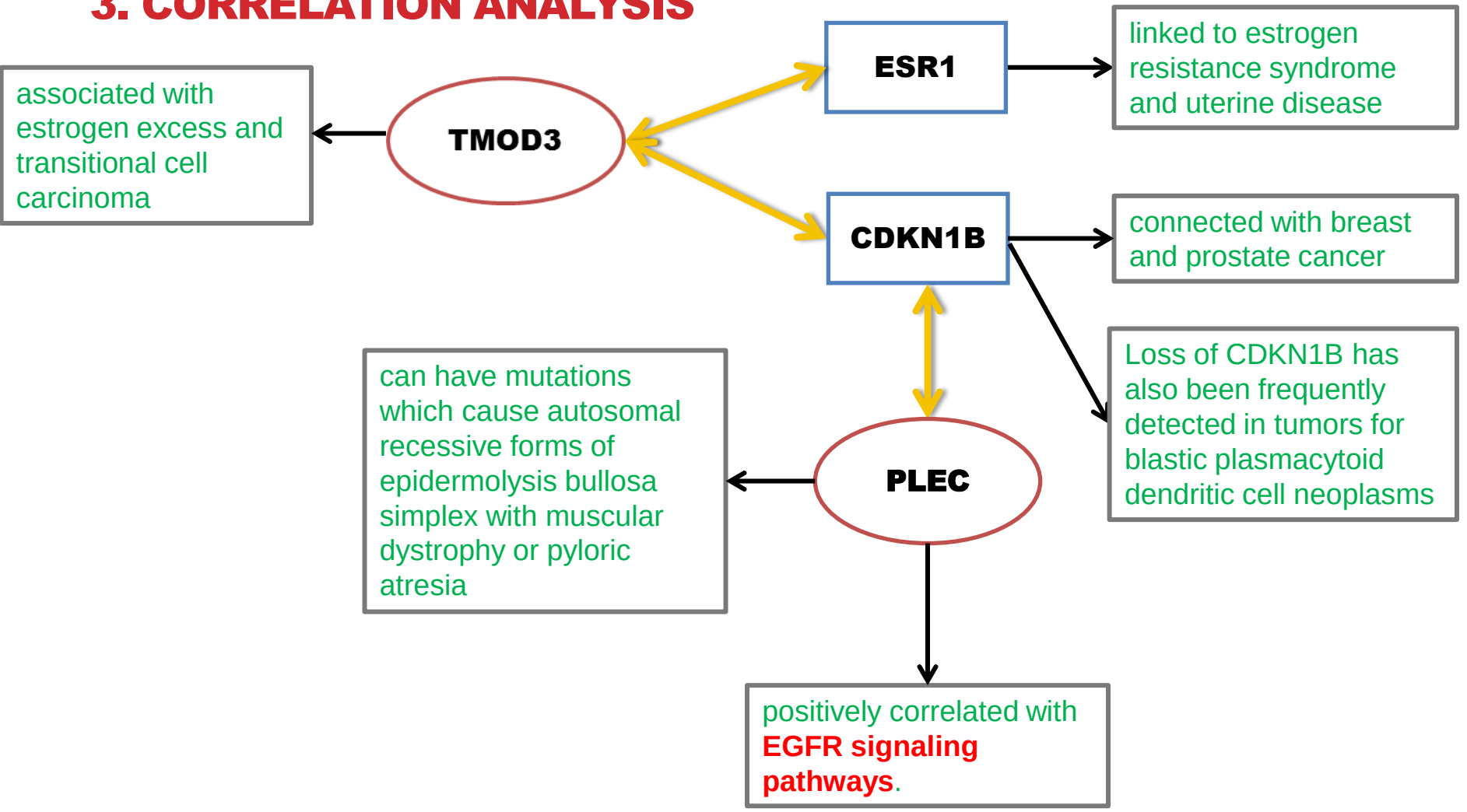


3. CORRELATION ANALYSIS

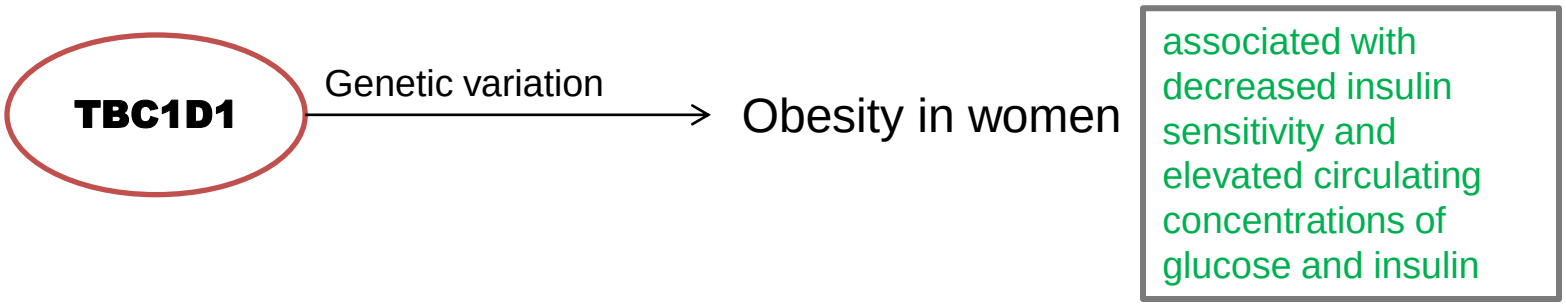
For chromosome 22, we have tested the difference between the means of the methylation values for the complete remission and progression groups using a two-sample t-test.

The p-value for the test is <0.0001.

3. CORRELATION ANALYSIS



3. CORRELATION ANALYSIS



- **Estrogen signaling** plays a role in this process.

CONCLUSION

- Important biologically meaningful differences between the –omics profiles of the two groups of patients.

- Estrogen signaling pathway and epidermal growth factor receptor (EGFR) signaling pathways are two of the significantly differentiating pathways between the two groups of patients.

ACKNOWLEDGEMENT

FACULTY COLLABORATORS

Dr. Susmita Datta

Dr. Somnath Datta

Dr. Ryan Gill

STUDENT COLLABORATORS

Sandipan Dutta

Hyoyoung Choo-Wosoba

Younathan Abdia

This work was partially funded by NIH/NCI Grant CA170091-01A1 to Susmita Datta


THANK
YOU