Ole Fredrik Borgundvåg Berg

Circulating miRNA and lung cancer:
- an analysis of available data

Preparation Project, Fall 2021

Supervisor: Pål Sætrom
Artificial Intelligence Group
Department of Computer and Information Science
Faculty of Information Technology, Mathematics and Electrical Engineering

Abstract

There has been many studies on the diagnostic value of using circulating microRNA to diagnose lung cancer, which is an important research area, as an earlier diagnosis of lung cancer will save many lives. While there has been multiple studies on this research areas, few have tried to combine the results from multiple studies, and none have tried to collect all available datasets, which was the goal of this project. In this project I have collected all available datasets in order to compare them and to look at the results from using one dataset to diagnose lung cancer in another dataset.

The contributions has been that all datasets are collected made into a common format, which would make it easier to do research on this area in the future. Another contribution is the overview of the availability of datasets on the subject, and the properties of the available datasets. In short, the datasets from most studies were not given upon request, which leads to no possibility of replication and limits the possibility for collecting many datasets in order to make research on a larger combined datasets which could have given more statistical power and better estimates.

Finally, trying to diagnose across datasets generally led to quite poor results with low diagnostic value, which suggest that the results from the different studies often don't replicate. However, these results came from very naïve machine learning, and more research are needed in order to see if more advanced machine learning can find patterns across datasets better.

Preface

This is a report for the project in the course ”IT3915 - Master in Informatics, Preparatory Project”, conducted at NTNU and St. Olavs Hospital, supervised by Pål Sætrom. I want to thank friends and family for support.

Ole Fredrik Borgundvåg Berg
Trondheim, December 12, 2021

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The project will primarily be about using methods from machine learning and statistics to look at the diagnostic value of circulating miRNA when it comes to lung cancer.

1.1 Background and Motivation

Lung cancer is common type of cancer with a low survival rate (more statistics in 2.1.1). One of the major reasons for the low survival rate is the late diagnosis of lung cancer. However, several studies indicate that circulating miRNA could be a non-invasive way to diagnose lung cancer [Shen et al., 2013]. This could lead to earlier diagnosis, and thus a higher survival rate.

1.2 Goals and Research Questions

Different studies have pointed to different miRNA sequences for the diagnosis of lung cancer. The point of this project is to collect the datasets from the different studies and create a larger dataset. With that dataset I want to achieve the following overall goal:

Goal: Use algorithms from machine learning on to predict lung cancer from levels of circulating miRNA on a larger dataset

Most studies on the area use simple logistic regression on the data in order to predict lung cancer based on miRNA values (e.g. [Wozniak et al., 2015; Niu et al., 2019]), thus leading to the question:

Research question 1: Are there machine learning algorithms that generally performs better at diagnosing lung cancer based on miRNA values?

Logistic regression is a linear model, and thus is unable to find patterns in the data that are non-linear, which might be the case with the effect of lung cancer on miRNA levels. There have been attempts of using more advanced machine learning methods on miRNA and lung cancer (e.g. [Lopez-Rincon et al., 2020]). My project differs, as I try to collect all available datasets, which gives more statistical power and it will be more of a meta-analysis where datasets from different studies are compared.

The datasets are slightly different in what miRNA that are measured, and what technologies are used to measure miRNA levels. This begs the question:

Research question 2: Will a combined dataset lead have better diagnostic value than each of the datasets alone?

On one hand one might think that the more data, the more information the machine learning algorithm have, and thus, a combined dataset is better. However, it is possible that datasets of lower quality would confuse rather than help a machine learning algorithm.

Other minor questions that might be answered are:

- What are the respective quality of the different datasets?
- Do the same miRNA have the same diagnostic value across different datasets?
- What miRNAs are most important for diagnosing lung cancer?
- What is the effect of lung cancer on the miRNA levels?

However, these questions are more interesting from a medical/biological point of view, than they are from a machine learning point of view, and as such will be a lower priority.

1.3 Research Method

This project is primarily an experimental one, as one need to actually train models on the datasets in order to compare the outcomes. The outcomes of the machine learning model are quantitative, and thus an analytical approach will be used. The main theoretical parts of this project are the parts concerning miRNA and lung cancer, as the outcomes of the machine learning might help in understanding the effect of lung cancer on miRNA, but as these questions are not related to machine learning directly, they are not the main focus.

1.4 Report Structure

Chapter 2 will include some theory around lung cancer and miRNA, together with theory around the machine learning and statistical methods and concepts that are used in this project. Chapter 3 is about how the literature search was done, how the data was processed and how machine learning was applied. Chapter 4 is about how the experiments performed and their results. Finally, chapter 5 is about the conclusions that are made from the results, and their significance.

Chapter 2

Background Theory and Motivation

This project is a cross-disciplinary one, as it combines machine learning and medicine, and as such, some theory from both disciplines are necessary in order to understand the project.

2.1 Biological Theory

The first major part of is the biological/medical part.

2.1.1 Lung Cancer

Lung cancer is the second most common type of cancer worldwide, and the the type of cancer with the highest total mortality worldwide, causing about 1.8 million deaths [Sung et al., 2021]. Lung cancer is also the cancer type leading to the most deaths in Norway, amounting to 1500 deaths per year [Cancer Registry of Norway, 2021]. The most important risk factor related to lung cancer is smoking. Smoking is estimated to explain about 90% of the risk of lung cancer in men, and 70% to 80% of the risk of lung cancer in women [Walser et al., 2008]. Furthermore, about 90% of lung cancer deaths in men, and 79% of lung cancer deaths in women are caused by smoking [Shopland et al., 1991].

There are two main types of lung cancer, Small Cell Lung Cancers (SCLC) and Non-Small Cell Lung Cancers (NSCLC) [Ciupka, 2020]. Of lung cancer cases, about 80-85% are NSCLC, whilst 10-15% of the cases are SCLC, and a few percent are non-mayor types of lung cancer [American Cancer Society, 2019]. NSCLC cancers tend to grow slower than the SCLC cancer types, and thus SCLC has

usually already spread when it is diagnosed [American Cancer Society, 2019]. The NSCLC has three mayor subtypes, namely, adenocarcinoma (30-40%), squamous cell (30%) and large-cell undifferentiated carcinoma (10-15%) [Ciupka, 2020]. The treatment and prognosis for the different NSCLC subtypes are similar [American Cancer Society, 2019].

Lung cancer develops in different stages. According to Bernstein [2019], the main four are:

1. The cancer is only situated in your lung
2. The cancer may have spread to the lymph nodes near the lung
3. The cancer has spread deeper into the lymph nodes and into the middle of your chest
4. Cancer is widespread throughout your body

The main advantage with diagnosing lung cancer early is that the cancer has not yet spread to other parts of the body, which means that it can be removed by surgery [American Cancer Society, 2021]. On the other hand, later stages might require chemotherapy, radiation therapy or immunotherapy, but as the cancer has spread widely, this cure will likely not remove the cancer [American Cancer Society, 2021].

2.1.2 MicroRNA

MicroRNA (miRNA) are short sequences of RNA, about 22 nucleotides each, that regulates the expression of mRNA by binding to the target mRNA sequence, and thus stopping it from being translated. Circulating miRNA has been found to be a biomarker for many diseases, including cancer, infectious diseases and mental illnesses [Correia et al., 2017; Kosaka et al., 2010; Geekiyanage et al., 2012; van den Berg et al., 2020]. miRNA-sequences are usually named with the prefix "miR-" and then a unique number that is incremented for each discovery of a miRNA-sequence. The most commonly used database with known miRNA-sequences is the miRBase database [Griffiths-Jones et al., 2006].

2.1.3 MicroRNA and Lung Cancer

The overall role role of miRNA in relation to lung cancer is not fully understood [Uddin and Chakraborty, 2018]. MicroRNA is thought to be both function as tumor suppressor genes and as oncogenes [Lynam-Lennon et al., 2009]. Anyways, there are multiple studies that report about differential expression of circulating miRNA-sequences in cancer patients compared to healthy controls, which is

results in expression of miRNA being a promising method for diagnosing lung cancer [Uddin and Chakraborty, 2018].

2.1.4 MicroRNA profiling methods

There are several methods for measuring levels of miRNA. The most common ones are qRT-PCR, microarrays and sequencing. Here is a very high level description of the different methods. For more technical details see e.g. Pritchard et al. [2012]. The different technologies typically have different issues.

qRT-PCR

Quantitative Reverse Transcription - Polymerase Chain Reaction (qRT-PCR) is the most common method in the studies used in this project. As the name implies, the process depend on reverse transcription, where miRNA are reverse transcripted, using the enzyme reverse transcriptase, into complementary DNA (cDNA). Then polymerase chain reactions are initiated and monitored in order to measure miRNA levels.

In qRT-PCR, one needs a primer for each miRNA-sequence that should be measured. Therefore it can only measure miRNA-sequences that are decided beforehand. The main advantage of qRT-PCR is that it is the most sensitive method of the different technologies [Pritchard et al., 2012], which means that the results are more accurate, and that it also works well when the concentration of miRNA is low.

Microarrays

Micoroarrays are what is called a hybridization method. It starts out similarly to qRT-PCR, with converting miRNA into cDNA, only that the miRNA in this case are fluorescently labeled. The microarray has several spots, each with single-stranded DNA samples (called probes) that are mounted to the microarray. When the cDNA are added to the microarray, the cDNA will bind to the DNA samples that have the same sequence, in a process called hybridization. Afterwards, the microarray is washed clean, and only the cDNA that has managed to bind will remain. Thus, by checking for the fluorescence of the different spots, one can find which DNA-probes had cDNA bind to it, and which had not. The level of fluorescence can then be used as the concentration of the corresponding miRNA-sequence.

The main advantage of microarrays is that it is the cheapest of the main technologies [Pritchard et al., 2012]. The disadvantages is that it has low sensitivity, and that you have to decide beforehand what miRNA-sequences you

want to measure, as you need to populate the microarray with the corresponding DNA-probes.

Sequencing

Sequencing also starts with converting miRNA into cDNA. A primer is then connected to the cDNA in one direction. The sequencing step works by adding fluorescent bases one by one, and then see if they adds to the sequence starting with the primer. Thus, one can read out the sequence of the cDNA.

The main disadvantage of sequencing is that it is expensive [Pritchard et al., 2012]. It is also less sensitive than qRT-PCR. The main advantage, however, is that you do not need to decide beforehand the miRNA-sequences you want to measure.

2.2 Machine learning theory

The second major part of this project is the machine learning.

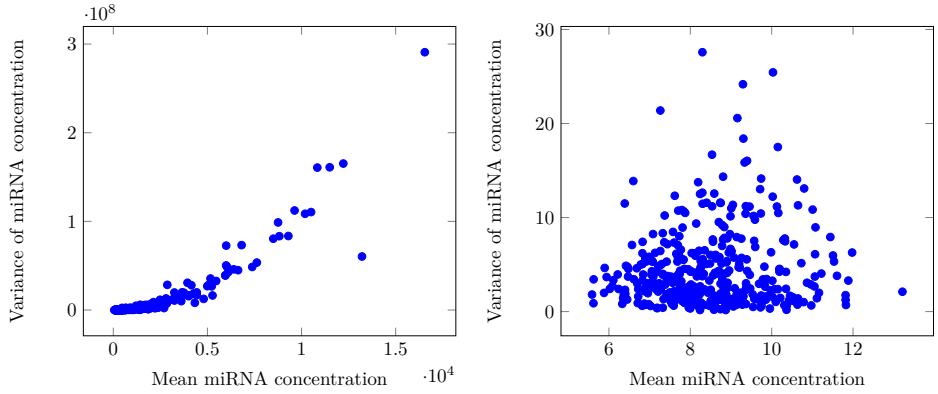
2.2.1 Variance stabilizing transformation

In miRNA measurements, one often see that the variance in miRNA concentration is a function of the mean miRNA concentration. One possible transformation is the log transformation where one takes the logarithm of the data. That can change a curve where $\mathrm{Var}[y] \propto \mathrm{E}[y]^2$ into a curve where the variance of y is independent of the mean of y. One example of this can be seen in Figure 2.1.

Another advantage of a variance stabilizing transformation is to ensure that the data is not skewed. Other statistical tools like explained variance (subsection 2.2.5) assumes that the underlying data has a normal distribution. A normal distribution, however have no skew, therefore unskewing the data is necessary for ensuring that other methods are giving valid results. More formally, if we assume that Y $g(X)$ for some function g and that $X \sim N(\mu, \sigma)$, then doing the transformation $y' = g^{-1}(y)$ ensures that our variables are normally distributed. In particular, if we assume that $g(X) = e^X$, then the log-transformation will ensure that our data is normally distributed.

2.2.2 Fold change

Fold change is defined as the ratio of a certain value between two different populations. In this project, the fold change used is typically the ratio levels of a



(a) Before log transformation

(b) After log transformation

Figure 2.1: Mean and variance in different miRNA sequences in [Ma et al., 2011]

certain miRNA-sequence between cases and controls. Log-fold change is the logarithm of the fold change (by convention $\log_2$ is used in this area of research). Furthermore:

$$\text{Fold change} = \frac{a}{b}$$
$$\text{Log-fold change} = \log_2\left(\frac{a}{b}\right) = \log_2 a - \log_2 b$$

In other words, the log-fold change is the difference in miRNA expression when the data are log-transformed.

2.2.3 Loess regression

Loess regression is also sometimes called local regression, and it is a type of regression that is made for smooting scatterplots [Cleveland, 1979]. The regression works by fitting a low degree polynomial for each datapoint. The fitting of each polynomial works by giving weight to nearby points, where more weight is given to points near the original datapoint. The regression value for each datapoint is thus the value of the corresponding polynomial evaluated in this point.

Loess regression is practical when mean and variance still are not independent after a log transformation. Using loess regression can ensure that they become independent as shown in Figure 3.1

2.2.4 Principal component analysis

Principal component analysis (PCA) is a method of data reduction, where a dataset in $\mathbb{R}^n$ is projected down on a lower dimensional vector space $\mathbb{R}^m$. The projection in PCA is the projection that ensures that the most of the variance of the original dataset is kept, whilst ensuring that the projection is not expanding the dataset. One of the main advantages of PCA is that you could project a dataset down to just two or three dimensions, which makes it possible to plot the dataset.

2.2.5 Explained variance

Explained variance is a way of analyzing the sources of variance in a dataset. Using linear regression, one assumes that the dependent variable y, covariates $\mathbf{X}$ and residuals $\epsilon \sim N(0, \sigma)$ have the relationship $y = \mathbf{X}\beta + \epsilon$, for some parameter vector β.

If one creates a linear regression model of the dataset one get a parameter vector $\hat{\beta}$ which is the maximum likelihood estimate of β, and predictions $\hat{y} = \mathbf{X}\hat{\beta}$ for y. Also define $\mathbf{SST} = \sum_i (y_i - \bar{y})^2$ is the total sum of squares, $\mathbf{SSR} = \sum_i (\hat{y}_i - \bar{y})^2$ is the sum of squares due to regression and $\mathbf{SSE} = \sum_i (y_i - \hat{y}_i)^2$ is the sum of squared estimate of errors. Then we have the following relationship:

$$\mathbf{SST} = \mathbf{SSR} + \mathbf{SSE}$$

The proportion of the empirical variance that can be explained by the covariates is thus

$$R^2 = \frac{\mathbf{SSR}}{\mathbf{SST}}$$

2.2.6 Logistic regression

Assume that you have Bernulli trials where each $y_i \sim \text{Bernoulli}(p_i)$ with the relationship:

$$\frac{p_i}{1 - p_i} = e^{x_i^T \beta}$$

for some covariates x_i and some parameter vector β. Then one can show that

$$p_i = \frac{1}{1 + e^{-x_i^T \beta}}$$

A logistic regression model can find $\hat{\beta}$, the maximum likelihood estimate of β.

Logistic regression is a relatively simple classification model, and in the studies used in this project, logistic regression is the most commonly used model for diagnosing lung cancer based on miRNA levels.

2.2.7 XGBoost

XGBoost is a machine learning algorithm that is based on gradient tree boosting [Chen and Guestrin, 2016]. Boosting algorithms are machine learning algorithms that combine weak models into stronger models, by using the combined output of several weak models. Gradient boosting is a type of boosting algorithm that uses an idea similar to gradient descent in order to find optimal weights given to each of the weaker models [Friedman, 2001]. Instead of using the gradient directly, the algorithm just ensure that the weights are updated such that the loss function is lowered in each step. Gradient tree boosting is gradient boosting where the weak models are decision trees. XGBoost's decision trees have default directions of descending in the tree if there are missing data, thus good handling missing values is one of XGBoost's biggest advantages.

XGBoost is a popular machine learning algorithm on the machine learning contest site Kaggle[1], winning 17 of 29 contests in 2015 [Chen and Guestrin, 2016].

2.3 Structured Literature Review Protocol

The point of the literature search was to find studies relevant to miRNA and circulating lung cancer. The main search engine used was PubMed[2], which is a commonly used search engine for medical litterature. The search term used was:

```
(lung OR pulmonary OR NSCLC) and (tumor OR cancer OR carcinoma) and
(microRNA* OR miRNA* OR miR*) and (diagnosis OR biomarker OR detection)
and (serum or plasma or "whole blood")
```

In addition, I search in databases that have public gene expression data as shown in Table 2.1.

Database name	Search term
ArrayExpress[3]	`microrna lung cancer`
Gene Expression Omnibus (GEO)[4]	`(mirna OR microrna) AND "lung cancer" AND (diagnosis OR detection)`
OmicsDI[5]	`"lung cancer" AND TAXONOMY: 9606 AND "breast cancer" AND (mirna OR microrna) AND (serum OR plasma OR "whole blood")`

Table 2.1: Search in public gene expression databases

[1] `https://www.kaggle.com`
[2] `pubmed.ncbi.nlm.nih.gov/`

The inclusion criteria where based on what datasets I thought was relevant to this project:

- The paper is an experiment where circulating miRNA is measured.

Some of the studies measured miRNA levels in the lung tissue or in sputum, rather than measuring circulating miRNA. As the values are somewhat different between lung tissue miRNA and circulating miRNA [Petriella et al., 2016], only the circulating miRNA ones was selected in order to have a consistent dataset. In addition, the research question was to look at the diagnostic value of circulating miRNA, which makes it most reasonable to base on circulating miRNA data.

- The study both have people diagnosed with lung cancer and controls not diagnosed with lung cancer.

The controls in some of the studies are not healthy, but suffers from other kind of lung diseases. Other studies have both have both healthy controls, and controls with other lung illnesses. Both are relevant, as on one hand, one would like to see the difference between healthy controls and patients with lung cancer in order to remove miRNA changes due to other illnesses. On the other hand, people who are getting checked for lung cancer often have lung issues, which is the reason for their checkup.

Some studies were excluded as they did not have a control group like Mitchell et al. [2017].

- At least four different miRNA sequences were measured.

The point of this project is to combine and compare datasets. Having few miRNA sequences measured makes it hard to combine datasets, as there is a high likelihood that there are no overlapping miRNA sequences between the datasets. It is also hard to compare datasets measuring completely different miRNA sequences.

- Meta-analyses were used as source of relevant studies

Some of the studies found were meta-analyses. In that case relevant studies were retrieved from the references of the meta-analysis.

[3] `https://www.ebi.ac.uk/arrayexpress/`
[4] `https://www.ncbi.nlm.nih.gov/gds`
[5] `https://www.omicsdi.org`

Chapter 3

Methodology

The project will be divided in five main phases:

- Literature search
- Preprocessing of datasets
- Statistical analysis of datasets
- Combining the datasets
- Machine learning on datasets

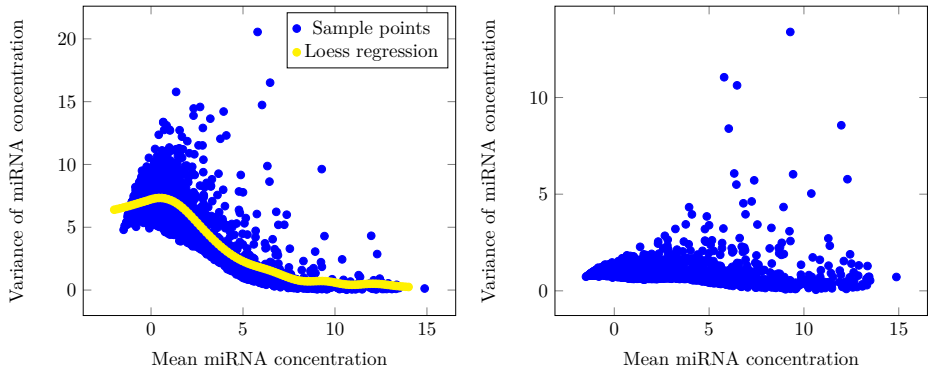
whereas the literature search was described in section 2.3, and the other parts will be described in this chapter.

3.1 Technical setup

In Table 3.1, the main software used in this project is listed.

Software	Version	Usage
Python[1]	3.9.7	Programming language
NumPy[2]	1.20.3	Numerical calculations with vectors and matrices
scikit-learn[3]	0.24.2	Machine learning
XGBoost[4]	1.4.2	XGBoost machine learning algorithm
SciPy[5]	1.7.1	Scientific programming

Table 3.1: Software used in this project



(a) Mean and variance before loess adjustment

(b) Mean and variance after loess adjustment

Figure 3.1: Mean and variance before and after loess in Asakura et al. [2020]

3.2 Preprocessing of the datasets

Preprocessing is done to each of the datasets in order to make the datasets as comparable as possible.

3.2.1 Log-transforming the data

The log-transformation is explained closer in subsection 2.2.1. The point of this is that we see that the variance in miRNA levels are approximately proportional to the square of the mean of the miRNA levels. Then log-transformation will make the variance independent from the mean.

3.2.2 Loess regression on the mean-variance relationship

In some cases, especially if the study uses microarrays, the mean-variance relationship is still not constant after log-transforming the data. In this case, loess regression (see: subsection 2.2.3) is used to adjust the variance of the samples to ensure that the variance is independent of the mean. An example is in Figure 3.1.

[1] `https://www.python.org`
[2] `https://numpy.org`
[3] `https://scikit-learn.org/`
[4] `https://xgboost.readthedocs.io/`
[5] `https://scipy.org`

3.2.3 Adjusting for covariates

Some datasets report information about the patients, including their sex, age and/or pack years[6]. These demographic variables affect the miRNA levels, and we want to make these variables have as little influence as possible on the miRNA levels. Therefore, a linear regression model is fitted with the demographic variables as covariates, and with the miRNA levels as dependent variables. The resulting model will then have an estimate of the effect of the different covariates. By subtracting the effect of these covariates, more of the variance in the miRNA levels will be due to case-control characteristics. Another advantage of doing this adjustment is that the demographics of the different studies will differ, and by removing the effect of demographic variables, the studies become more comparable.

3.2.4 Standardizing miRNA levels

In order to make the measured miRNA levels comparable when they are measured using different technologies, the measured miRNA levels for each miRNA-sequence are standardized. However, one has to take into account that the datasets differ in the relative number of cancer and control samples. To adjust for this, mean and empirical variance was calculated for the cancer and the control samples separately. Let $\hat{\mu}_{ca}$ and $\hat{\sigma}^2_{ca}$ be the mean and empirical variance of the cancer samples, and likewise $\hat{\mu}_{co}$ and $\hat{\sigma}^2_{co}$ for the controls. Then the overall mean $\hat{\mu}$ and overall variance $\hat{\sigma}^2$ are estimated as:

$$\hat{\mu} = \frac{\hat{\mu}_{ca} + \hat{\mu}_{co}}{2}$$
$$\hat{\sigma}^2 = \frac{\hat{\sigma}^2_{ca} + \hat{\sigma}^2_{co}}{2}$$

which are estimates of what the mean and variance would be if the dataset were balanced with an equal number of cancer and control samples.

3.3 Statistical analysis of the datasets

To get an overview of the datasets, different statistical analysis are done on the datasets.

[6]pack years = packs of cigarettes smoked per day * number of years smoked

3.3.1 Explained variance

An interesting question is how much of the variance in the miRNA levels is due to case-control in the different datasets. It is plausible that datasets that have only measured miRNA-sequences assumed to have a correlation with lung cancer have a larger portion of the variance due to case-control. As explained variance is made for the case where there is only one dependent variable, one has to do some adjustments to get an overall estimate. One way is to take the average of the explained variance for each miRNA-sequence (i.e. a uniform weight). Another way will be to to weight by the variance of each miRNA-sequence, as miRNA-sequences showing more variance might do so due to case-control effects (i.e. variance weighted). I will calculate both statistics. The explained variance will be calculated after the log-transformation, to be sure that the distribution of the levels for each miRNA-sequence is approximately normally distributed, as this is one of the prerequisites for the explained variance-analysis to be valid.

Explained variance will be calculated using `LinearRegression` and `explained_variance_score` in scikit-learn.

3.3.2 PCA

I want to run PCA, with two principal components, on the different datasets, as it makes it possible to visualize the dataset in a plot. This serves multiple purposes; first of all it makes it possible to find outliers in the datasets, as they would be far from the other samples in the plot. Secondly, by coloring the samples by whether they are cases or controls, one can visualize how separable the data are between cases and controls. This last point assumes that case-control effects on the data are along one (or both) of the first two principal components, as if they are not, the plot would not be separable, even though the full dataset might be. As the first components represent the main sources of variance in the data, seeing whether the data is separable in case-control in the plot is an alternative to explained variance, for seeing whether case-control effects are the main causes of variance in the dataset.

PCA will be computed by the `PCA` function in scikit-learn.

3.3.3 Fold change correlation

Another interesting question is whether the fold changes between cases and controls are the same in the different datasets. To do that, I will calculate the fold changes between the case and control in the different datasets, for different miRNA-sequences. There should be some correlation between fold changes in the different datasets, otherwise, it would seem that the biomarkers for lung cancer

in the different datasets cannot be replicated, and thus one might question the results of the studies.

Correlation will be computed using `pearsonr` in SciPy.

3.4 Combining datasets

Combining the datasets are done by processing the datasets so that they have equal characteristics. The miRNA-sequence representation in the datasets were translated into a common format. The common format is a CSV-file where the columns are the miRNA-sequences together with a last columns that says if the person has lung cancer or not. The rows are the different persons in the dataset. This goes against the convention in miRNA research, where it is more common with miRNA in rows and persons in columns. However, I use the convention used in machine learning, and it is simpler as most machine learning libraries assumes that the data are in this format.

Then a function was made for extracting a subset of the datasets by using the intersection of miRNA-sequences of the extracted datasets, where one takes the intersection of miRNA-sequences in the studies that are collected. In addition, one can choose whether one would have each miRNA-sequence standardized separately or not.

3.5 Machine learning on the datasets

Machine learning on the datasets will be done by using logistic regression, where one takes two datasets and train a logistic regression model on one of the datasets and then tries to predict on the other dataset. The AUC will be used as a metric of the diagnostic value of the model. This would be done for each pair of datasets that has at least 10 different miRNA-sequences in common.

These models will be trained using `LogisticRegression` in scikit-learn, and the AUC will be computed using `roc_auc_score` in scikit-learn.

Chapter 4

Experiments and Results

This section will contain the results from the experiments.

4.1 Literature search

The literature search yielded 123 studies of interest. Of these 25 had raw microRNA public. For the other studies, I sent an email requesting the raw miRNA data. However, only one such dataset were received, leading to an overall 26 datasets that are analyzed in this project ([Asakura et al., 2020], [Bianchi et al., 2011], [Boeri et al., 2011], [Chen et al., 2019][1], [Duan et al., 2021], [Fehlmann et al., 2020], [Halvorsen et al., 2016], [Jin et al., 2017], [Keller et al., 2009], [Keller et al., 2014], [Keller et al., 2020], [Kryczka et al., 2021], [Leidinger et al., 2011], [Leidinger et al., 2014], [Leidinger et al., 2015], [Li et al., 2017], [Marzi et al., 2016], [Nigita et al., 2018], [Patnaik et al., 2012], [Patnaik et al., 2017], [Qu et al., 2017], [Reis et al., 2020], [Wozniak et al., 2015], [Yao et al., 2019], [Zaporozhchenko et al., 2018]).

The distribution of technologies in these different studies are visualized in Figure 4.1, the number of samples are visualized in Figure 4.2 and a table with the characteristics of the different datasets is in Table 4.1.

[1] Chen et al. [2019] is not the study where the dataset originated from, but it is a study using the dataset. The dataset is GSE71661 in the Gene Expression Omnibus, and has no citation listed: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE71661`

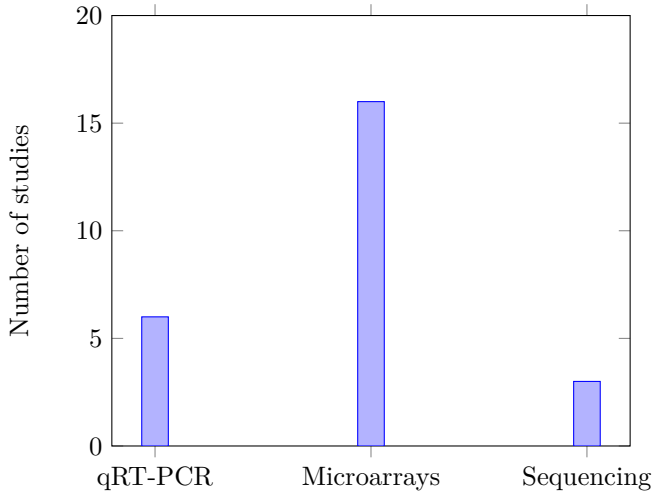


Figure 4.1: Number of studies of each type

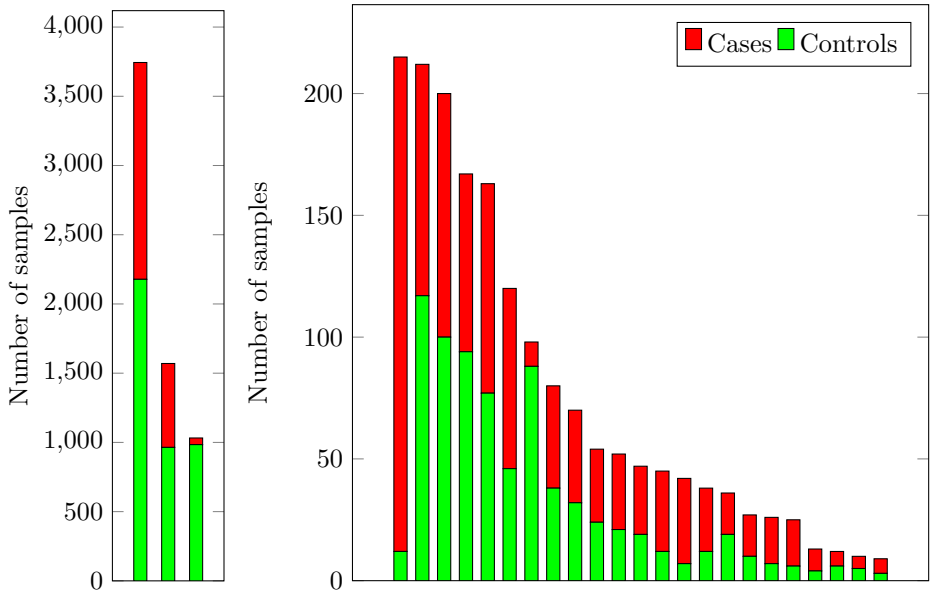
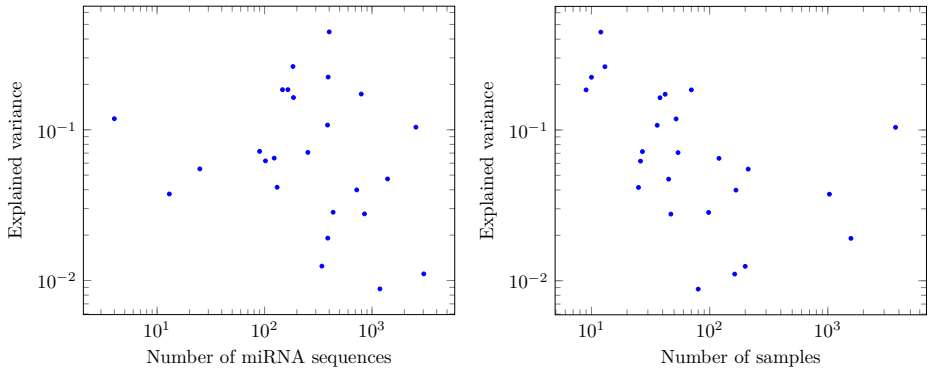


Figure 4.2: The number of samples in the different studies

Study	Technology	EV (uniform)	EV (weighted)	# miRNAs	# Cases	# Controls	# Total
Asakura et al. [2020]	Microarray	0.094	0.104	2565	1566	2178	3744
Bianchi et al. [2011]	Microarray	0.052	0.055	25	95	117	212
Boeri et al. [2011]	Microarray	0.044	0.042	131	19	6	25
Chen et al. [2019]	Sequencing	0.061	0.071	253	30	24	54
Duan et al. [2021]	Microarray	0.319	0.446	400	6	6	12
Fehlmann et al. [2020]	Microarray	0.019	0.019	388	606	964	1570
Halvorsen et al. [2016]	Microarray	0.133	0.185	147	38	32	70
Jin et al. [2017]	qRT-PCR	0.120	0.164	186	26	12	38
Keller et al. [2009]	Microarray	0.095	0.107	386	17	19	36
Keller et al. [2014]	Microarray	0.037	0.040	722	73	94	167
Keller et al. [2020]	Microarray	0.024	0.028	435	10	88	98
Kryczka et al. [2021]	qRT-PCR	0.103	0.118	4	31	21	52
Leidinger et al. [2011]	Microarray	0.027	0.028	852	28	19	47
Leidinger et al. [2014]	Microarray	0.009	0.009	1186	42	38	80
Leidinger et al. [2015]	qRT-PCR	0.060	0.065	123	74	46	120
Li et al. [2017]	Microarray	0.167	0.185	165	6	3	9
Marzi et al. [2016]	qRT-PCR	0.038	0.037	13	48	984	1032
Nigita et al. [2018]	Sequencing	0.058	0.062	102	19	7	26
Patnaik et al. [2012]	Microarray	0.044	0.047	1396	33	12	45
Patnaik et al. [2017]	Microarray	0.011	0.011	3036	86	77	163
Qu et al. [2017]	Microarray	0.204	0.263	184	9	4	13
Reis et al. [2020]	Microarray	0.165	0.173	795	35	7	42
Wozniak et al. [2015]	qRT-PCR	0.012	0.012	342	100	100	200
Yao et al. [2019]	Sequencing	0.176	0.224	391	5	5	10
Zaporozhchenko et al. [2018]	qRT-PCR	0.069	0.072	90	17	10	27

Table 4.1: Characteristics of the studies in this project. EV=Explained Variance. Explained variance is the proportion of variance that is due to case-control characteristics (see subsection 2.2.5).



(a) Scatter plot of explained variance and number of miRNAs.

(b) Scatter plot of explained variance and number of samples.

Figure 4.3: Scatter plots of variance weighted explained variance.

4.2 Processing the datasets

The processing of the datasets went mostly fine, except that there were some issues due to differences in reported information about the patient characteristics. Due to that, not all datasets were adjusted for sex, age and/or packing years. Therefore, one would expect some problems regarding that not all covariates are adjusted for in all datasets, which lead to worse comparability of the datasets.

4.3 Explained variance

The proportion of variance that can be attributed to case-control characteristics is shown in Table 4.1. One interesting question is whether there is some relationship between the number of miRNA-sequences in the datasets and the proportion of variance that is due to case-control characteristics. Intuitively, there might be that studies that are more selective in the number of miRNA-sequences choose miRNA-sequences that are expected to react to case-control characteristics, and thus having a larger portion of variance due to case-control statistics. A scatter plot of the relationship between number of miRNA-sequences and the explained variance, using variance weighted explained variance, is shown in Figure 4.3a. Seemingly, there is no relationship. A correlation test using Pearson's r finds no significant correlation with $r = -0.24$ ($p = 0.25$).

Another possibility is that the proportion of variance is higher when there are fewer samples in the dataset. Which could suggest that the estimated case-

control difference is overestimated when there are few cases, as fewer cases would lead to more overfitting. Figure 4.3b suggests a slightly negative relationship with Asakura et al. [2020] as an outlier. A correlation test using Pearson's r results in $r = -0.50$ ($p = 0.01$), which suggests a negative relationship.

4.4 PCA of the datasets

PCA plots were made for the different datasets to visualize the datasets. PCA plots for all the datasets are in Figure 5.1. The datasets varies in the degrees of separation between cases and controls in the two first principal components. It also shows the general spread and clustering within the dataset. It can also be used to see the effect of the data manipulation as in Figure 4.4.

4.5 Some notes on Asakura et al. [2020]

This is a section with some notes on Asakura et al. [2020]. There are three main reasons why there is a section for this study in particular:

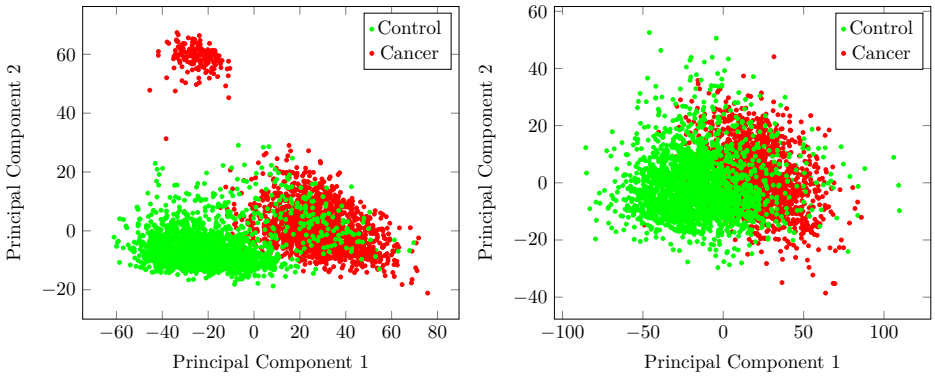
- It is the clearly largest dataset.
- It reports of a very high degree of seperation.
- It was a clear outlier in explained variance (see section 4.3)

Asakura et al. [2020] which reported virtually perfect separation (AUC = 0.996) using two miRNAs. Looking at the dataset in the PCA-plot on unadjusted data also suggested almost perfect seperation. The miRNA-sequence that had best seperation in the study, miR-17-3p, reported a seperation of 0.935 in the discovery set. Using the entire dataset with data preprocessing I found an AUC of 0.88. The AUC of this miRNA in different datasets is shown in Table 4.2. As one can see, the degree of seperation for this microRNA-sequence does not replicate well across studies, which suggest that it might be hard to get a similar seperation in other datasets, if one trains a model on the Asakura et al. [2020] dataset.

Figure 4.4 shows that a lot of the separation between cases and controls in the dataset is due to different demographics. This makes the data harder to make inference on, as the two groups are not directly comparable. If the effect of the demographic variables is not linear, which is a reasonable assumption, the adjustment done will not control for demographic effects fully, leading to worse results than what would be the case if cases and controls were more similar demographically.

Study	AUC
Asakura et al. [2020]	0.883
Keller et al. [2009]	0.633
Keller et al. [2014]	0.607
Leidinger et al. [2014]	0.539
Yao et al. [2019]	0.520
Fehlmann et al. [2020]	0.473
Patnaik et al. [2017]	0.464
Leidinger et al. [2011]	0.342

Table 4.2: The AUC when using miR-17-3p to diagnose lung cancer in studies where miR-17-3p is measured



(a) PCA of Asakura et al. [2020] before covariates were adjusted for and before post operation samples were removed

(b) PCA of Asakura et al. [2020] after covariates have been adjusted for and with post operation samples removed

Figure 4.4: The PCA plots show the effect of data cleaning in Asakura et al. [2020]

4.6 Log-fold-change correlation

Log-fold-change correlations were calculated between the datasets in order to see to what degree differences between cases and controls replicate across datasets. To ensure that the correlations were significant, p-values were calculated. As controls, p-values were also calculated the columns representing the different miRNA-sequences were shuffled and when samples were randomly assigned to case and control. This is visualized in Figure 4.5.

As one can see, the correlations were much more significant when the miRNA-sequences were not shuffled, which means there are at least some consistencies across the datasets. However, that correlations have similar significance whether I use real or random case-control suggests that the case-control differences might not replicate across studies.

The size of the correlations were, however, not very promising as the correlations were poor, and in many cases negative. The correlations of the log-fold-changes are mostly small, and the correlations are often negative, as shown in Figure 4.6. Even as these correlations are small and centered around zero, they are not spurious as shown with the p-values in Figure 4.5. However, as p-values were similar when case-control characteristics were randomly assigned, this might suggest that the correlations are due to covariance between different miRNA-sequences rather than having to do with case-control characteristics.

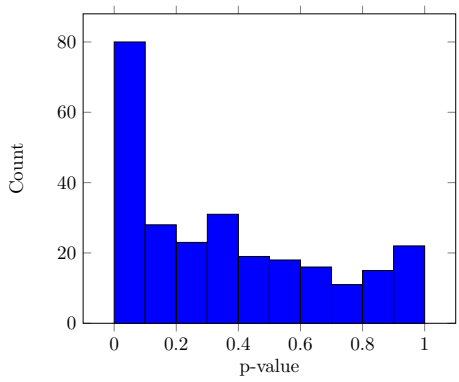
An overview over all the log-fold-change correlations is in Table 5.1.

4.6.1 Graph analysis of log-fold-change correlations

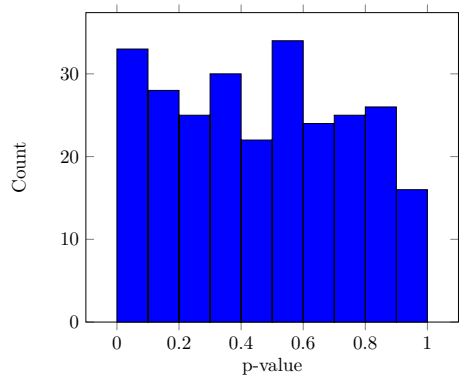
It is interesting to see whether there are any pattern in what studies correlates with other studies. I plotted a graph where there is an edge between two datasets iff they have at least 10 miRNA-sequences in common and the log-fold-change correlation is significant at a 0.05 significance level. The graph is in Figure 4.7. There are clearly some studies that correlates with many other studies, and others that are outliers. Studies that are outliers would probably be difficult to use to diagnose in other datasets.

4.7 Machine learning on the datasets

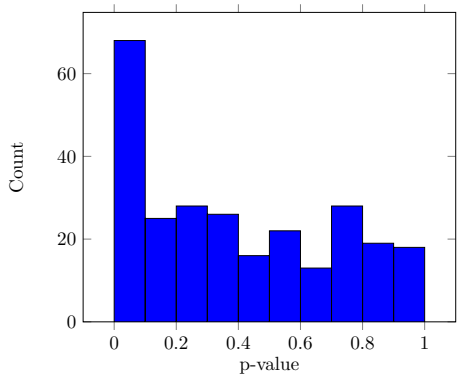
Machine learning was done where a logistic regression model was trained on one dataset and used to predict on another dataset. In Figure 4.8 the AUC values are shown, but only for pair of datasets with at least 10 different miRNA-sequences in common. The AUCs are close to 0.5, which means that there is little predictive power in general. Table 5.2 contains all the AUCs for the different datasets.



(a) p-values for log-fold-change correlation for when doing no manipulations



(b) p-values for log-fold-change correlation when the order of the columns with miRNA-sequences are shuffled



(c) p-values for log-fold-change correlation when case-control is randomly assigned

Figure 4.5: p-values for the correlations when doing different manipulations of the data

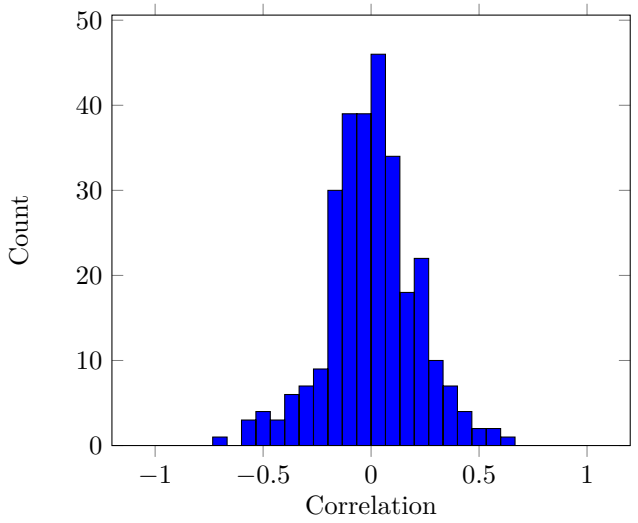


Figure 4.6: Histogram of the correlation in log-fold-change between the studies.

4.7.1 Clique analysis of machine learning results

Post hoc, it seemed interesting to see if there is some transitivity in whether datasets can be used to predict each other. I.e. if a model trained on dataset A can diagnose well in a dataset B, and a model trained on B can diagnose well in dataset C, does that imply that a model trained on A can predict well on C? For this analysis a graph was made that had a edge between dataset A and B iff they had at least 10 miRNA-sequences in common and the AUC of the model trained on A predicting on B was greater than 0.6 and same with the model trained on B predicting on A. This graph is shown in Figure 4.9. Then the problem became to find the maximal cliques in the resulting graph. The maximum clique that was found in the graph consisted of Duan et al. [2021], Keller et al. [2020], Jin et al. [2017] and Halvorsen et al. [2016].

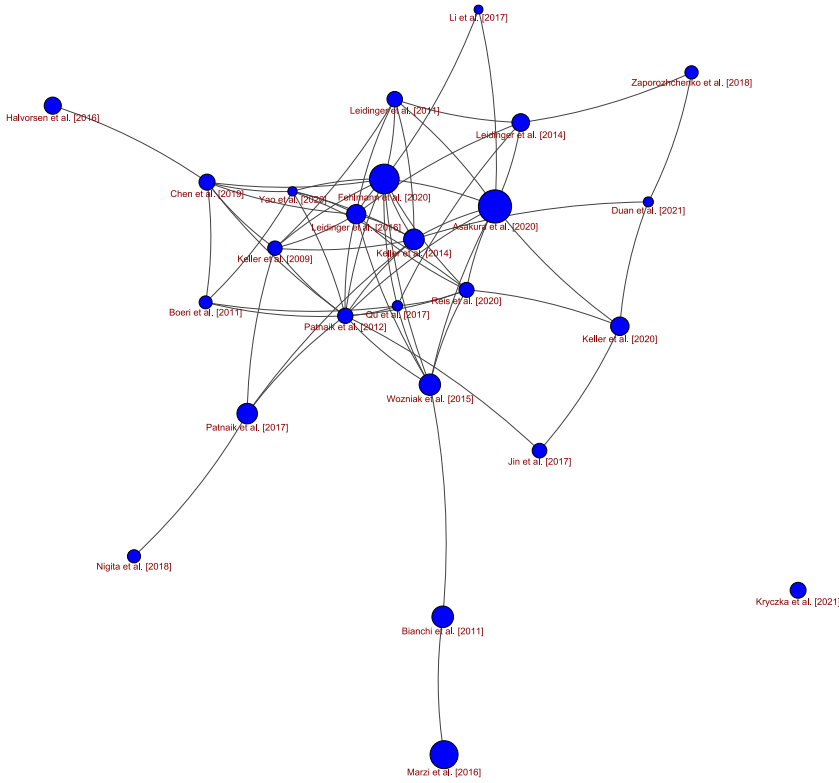


Figure 4.7: Graph over the log-fold-change correlation. The size of the nodes is proportional to the logarithm of number of cases in the datasets.

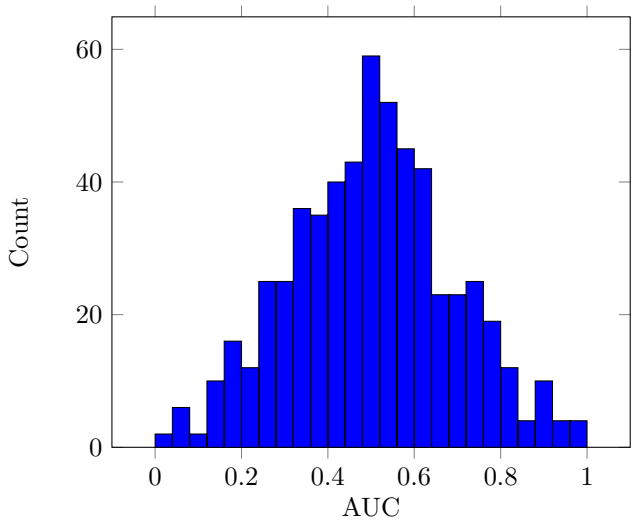


Figure 4.8: Histogram of the AUC when training on one study and predicting on another study.

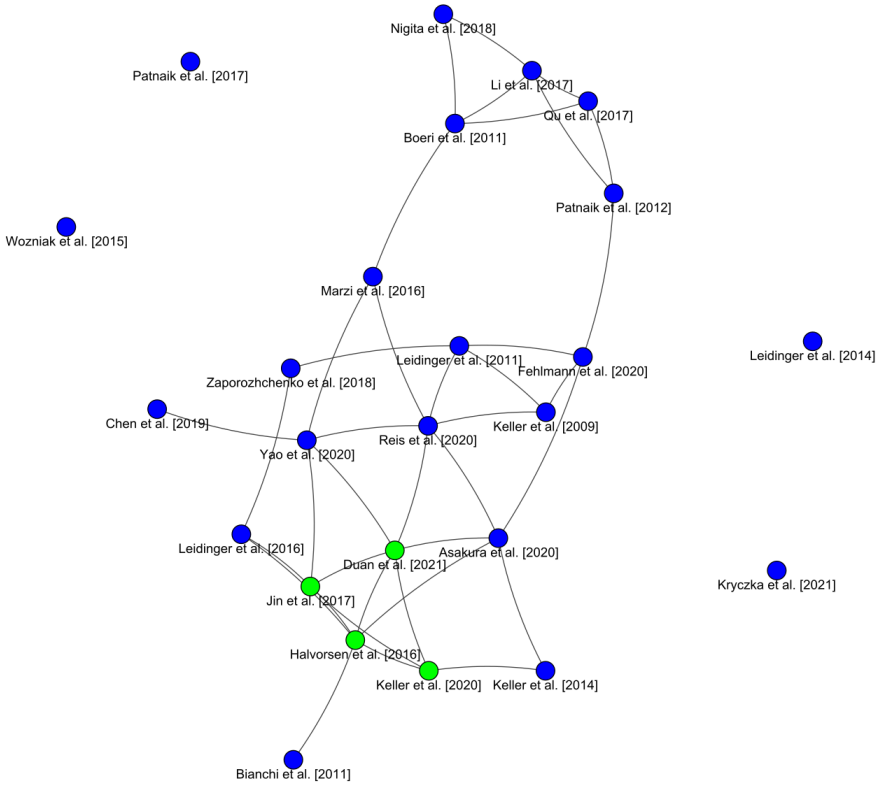


Figure 4.9: Graph over the datasets. The maximum clique is marked in green.

Chapter 5

Evaluation and Conclusion

This chapter contains the conclusions infered from the results in this project.

5.1 Evaluation

The results of the project were disappointing for two major reasons:

- I received virtually none of the requested datasets.
- The correlation in log-fold-change was poor, which suggest that replicating fold changes are hard, and that conclusions made from one dataset generally cannot be generalized.
- Using one dataset to try to predict on another dataset generally led to poor results.

On the other hand there are some important positive points to note:

- Null findings are also important.
- The work with transforming dataset is done so that other researchers can benefit.
- There is still possibilities for improving data modelling as this project have just done naive machine learning.

When it comes to the research questions, the two major one remains unanswered due to not enough time to do the machine learning part, as especially the literature search and data collection and processing work were exhaustive.

However, some useful information about the properties of the different datasets and the main questions will be considered more in the main thesis.

In retrospect, it might seem that a reasonable research question would be whether one dataset has diagnostic value for another dataset at all, which I took mostly for granted as I formulated the research questions. Especially because there are fold changes that seem replicated across studies (see e.g. Zhong et al. [2021]). There might of course be publication bias or similar issues at play, which means that one should be cautious in concluding with anything with certainty. Zhong et al. [2021] also found conflicting results regarding whether a certain miRNA-sequence was up- or down-regulated in many cases. In addition, ontology has generally a low replication rate. Errington et al. [2021] looked at 50 experiments from 23 high-inpact papersin cancer biology with a total of 158 effects. They found that positive effects could only be replicated in 43% of the cases, while 49% yielded null results and 7% resulted in significant results in the opposite direction. In total, the correlation between the original and the new effect sizes were $r = 0.47$.

Everything considered, I think that the project has been mostly successful in achieving most of what was planned in this project, though there was not much time for machine learning, which is for future work.

5.2 Discussion

The fact that I received few of the requested datasets is problematic, as it makes replication of the findings in the different studies hard. Walters et al. [2019] have also found a lack of transparency and data availability in oncology, and list some problems with this. The first one is that the cost of data collecting is typically high in cancer research, and in some cases can be affecting cancer patients negatively, which means that one would like to not have to collect more data than necessary. Thus having data available allows one to do cancer research more cost effectively. One example could be that one could use data from a study and do analysis of certain subsets of patients based on e.g. age, sex etc. Another point is that having data available leads to the possibility for researchers to replicate the statistical findings in the studies. There could be problems with p-hacking, spurious results etc. that would be hard for independent researchers to find without having the dataset available.

5.3 Contributions

This was the first project the try to collect all datasets on circulating microRNA and diagnosis of lung cancer. Trying to collect all the datasets gave an esti-

mate of the data availability that are in this area of research. Now that all the datasets are collected and processed into a common format, it would be easier for future research to build upon this data, as all the work doing data collection and processing is already done.

This project tried to look at different statistical properties of the different miRNA datasets that were available. This has led to an overview of the different available datasets, which is practical for reference.

There are also some preliminary results concerning machine learning across datasets which can be built upon in future studies to have something to measure against for trying to find improvements.

5.4 Future Work

There are several possibilities for building onto this work. The first major way is to try to make the datasets more comparable. The datasets were made using different technologies and different patient groups. This project tried to compensate by standardizing the data and adjusting the data using linear estimates of demographic effects, where demographic data was reported for the patients. However, it is possible that other adjustments to the datasets will lead to better correlations between the datasets.

Another possibility for future work is to combine different datasets and try to learn a model on this combined dataset, in hopes that this would lead to better ability for the model to generalize, so that the model would only use case-control effects that are reproduced between different datasets. It is possible that this would lead to better results. We already know that there are significant correlation in the log-fold-change, which means that it might be possible to get better results if preprocessing and machine learning in other ways, but it is highly uncertain as the correlations were similar when case-control status was randomly assigned.

It is also possible to try different machine learning models, as some models might be better than others when it comes to generalizing across datasets. This will be the focus in the main thesis.

Bibliography

American Cancer Society (2019). What Is Lung Cancer? | Types of Lung Cancer. https://www.cancer.org/cancer/lung-cancer/about/what-is.html.

American Cancer Society (2021). Non-small Cell Lung Cancer Treatment by Stage. https://www.cancer.org/cancer/lung-cancer/treating-non-small-cell/by-stage.html.

Asakura, K., Kadota, T., Matsuzaki, J., Yoshida, Y., Yamamoto, Y., Nakagawa, K., Takizawa, S., Aoki, Y., Nakamura, E., Miura, J., Sakamoto, H., Kato, K., Watanabe, S.-i., and Ochiya, T. (2020). A miRNA-based diagnostic model predicts resectable lung cancer in humans with high accuracy. *Communications Biology*, 3(1):1–9.

Bernstein, S. (2019). Lung Cancer Stages: Why and How Your Cancer Is Staged. https://www.webmd.com/lung-cancer/guide/lung-cancer-stages.

Bianchi, F., Nicassio, F., Marzi, M., Belloni, E., Dall'Olio, V., Bernard, L., Pelosi, G., Maisonneuve, P., Veronesi, G., and Di Fiore, P. P. (2011). A serum circulating miRNA diagnostic test to identify asymptomatic high-risk individuals with early stage lung cancer. *EMBO Molecular Medicine*, 3(8):495–503.

Boeri, M., Verri, C., Conte, D., Roz, L., Modena, P., Facchinetti, F., Calabrò, E., Croce, C. M., Pastorino, U., and Sozzi, G. (2011). MicroRNA signatures in tissues and plasma predict development and prognosis of computed tomography detected lung cancer. *Proceedings of the National Academy of Sciences*, 108(9):3713–3718.

Cancer Registry of Norway (2021). Cancer in Norway 2020 - Cancer incidence, mortality, survival and prevalence in Norway.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Chen, X., Jin, Y., and Feng, Y. (2019). Evaluation of Plasma Extracellular Vesicle MicroRNA Signatures for Lung Adenocarcinoma and Granuloma With Monte-Carlo Feature Selection Method. *Frontiers in Genetics*, 10:367.

Ciupka, B. (2020). Small Cell Lung Cancer vs. Non-small Cell Lung Cancer: What's the Difference?

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Correia, C. N., Nalpas, N. C., McLoughlin, K. E., Browne, J. A., Gordon, S. V., MacHugh, D. E., and Shaughnessy, R. G. (2017). Circulating microRNAs as Potential Biomarkers of Infectious Disease. *Frontiers in Immunology*, 8:118.

Duan, X., Qiao, S., Li, D., Li, S., Zheng, Z., Wang, Q., and Zhu, X. (2021). Circulating miRNAs in Serum as Biomarkers for Early Diagnosis of Non-small Cell Lung Cancer. *Frontiers in Genetics*, 12:987.

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., and Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, 10:e71601.

Fehlmann, T., Kahraman, M., Ludwig, N., Backes, C., Galata, V., Keller, V., Geffers, L., Mercaldo, N., Hornung, D., Weis, T., Kayvanpour, E., Abu-Halima, M., Deuschle, C., Schulte, C., Suenkel, U., von Thaler, A.-K., Maetzler, W., Herr, C., Fähndrich, S., Vogelmeier, C., Guimaraes, P., Hecksteden, A., Meyer, T., Metzger, F., Diener, C., Deutscher, S., Abdul-Khaliq, H., Stehle, I., Haeusler, S., Meiser, A., Groesdonk, H. V., Volk, T., Lenhof, H.-P., Katus, H., Balling, R., Meder, B., Kruger, R., Huwer, H., Bals, R., Meese, E., and Keller, A. (2020). Evaluating the Use of Circulating MicroRNA Profiles for Lung Cancer Detection in Symptomatic Patients. *JAMA oncology*, 6(5):714–723.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Geekiyanage, H., Jicha, G. A., Nelson, P. T., and Chan, C. (2012). Blood serum miRNA: Non-invasive biomarkers for Alzheimer's disease. *Experimental Neurology*, 235(2):491–496.

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., and Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(Database issue):D140–D144.

Halvorsen, A. R., Bjaanæs, M., LeBlanc, M., Holm, A. M., Bolstad, N., Rubio, L., Peñalver, J. C., Cervera, J., Mojarrieta, J. C., López-Guerrero, J. A., Brustugun, O. T., and Helland, Å. (2016). A unique set of 6 circulating microRNAs for early detection of non-small cell lung cancer. *Oncotarget*, 7(24):37250–37259.

Jin, X., Chen, Y., Chen, H., Fei, S., Chen, D., Cai, X., Liu, L., Lin, B., Su, H., Zhao, L., Su, M., Pan, H., Shen, L., Xie, D., and Xie, C. (2017). Evaluation of Tumor-Derived Exosomal miRNA as Potential Diagnostic Biomarkers for Early-Stage Non–Small Cell Lung Cancer Using Next-Generation Sequencing. *Clinical Cancer Research*, 23(17):5311–5319.

Keller, A., Fehlmann, T., Backes, C., Kern, F., Gislefoss, R., Langseth, H., Rounge, T. B., Ludwig, N., and Meese, E. (2020). Competitive learning suggests circulating miRNA profiles for cancers decades prior to diagnosis. *RNA Biology*, 17(10):1416–1426.

Keller, A., Leidinger, P., Borries, A., Wendschlag, A., Wucherpfennig, F., Scheffler, M., Huwer, H., Lenhof, H.-P., and Meese, E. (2009). miRNAs in lung cancer - Studying complex fingerprints in patient's blood cells by microarray experiments. *BMC Cancer*, 9(1):353.

Keller, A., Leidinger, P., Vogel, B., Backes, C., ElSharawy, A., Galata, V., Mueller, S. C., Marquart, S., Schrauder, M. G., Strick, R., Bauer, A., Wischhusen, J., Beier, M., Kohlhaas, J., Katus, H. A., Hoheisel, J., Franke, A., Meder, B., and Meese, E. (2014). miRNAs can be generally associated with human pathologies as exemplified for miR-144*. *BMC Medicine*, 12(1):224.

Kosaka, N., Iguchi, H., and Ochiya, T. (2010). Circulating microRNA in body fluid: A new potential biomarker for cancer diagnosis and prognosis. *Cancer Science*, 101(10):2087–2092.

Kryczka, J., Migdalska-Sęk, M., Kordiak, J., Kiszałkiewicz, J. M., Pastuszak-Lewandoska, D., Antczak, A., and Brzeziańska-Lasota, E. (2021). Serum Extracellular Vesicle-Derived miRNAs in Patients with Non-Small Cell Lung Cancer—Search for Non-Invasive Diagnostic Biomarkers. *Diagnostics*, 11(3):425.

Leidinger, P., Backes, C., Dahmke, I. N., Galata, V., Huwer, H., Stehle, I., Bals, R., Keller, A., and Meese, E. (2014). What makes a blood cell based miRNA expression pattern disease specific? - A miRNome analysis of blood cell subsets in lung cancer patients and healthy controls. *Oncotarget*, 5(19):9484–9497.

Leidinger, P., Brefort, T., Backes, C., Krapp, M., Galata, V., Beier, M., Kohlhaas, J., Huwer, H., Meese, E., and Keller, A. (2015). High-throughput qRT-PCR validation of blood microRNAs in non-small cell lung cancer. *Oncotarget*, 7(4):4611–4623.

Leidinger, P., Keller, A., Borries, A., Huwer, H., Rohling, M., Huebers, J., Lenhof, H.-P., and Meese, E. (2011). Specific peripheral miRNA profiles for distinguishing lung cancer from COPD. *Lung Cancer*, 74(1):41–47.

Li, L.-L., Qu, L.-L., Fu, H.-J., Zheng, X.-F., Tang, C.-H., Li, X.-Y., Chen, J., Wang, W.-X., Yang, S.-X., Wang, L., Zhao, G.-H., Lv, P.-P., Zhang, M., Lei, Y.-Y., Qin, H.-F., Wang, H., Gao, H.-J., and Liu, X.-Q. (2017). Circulating microRNAs as novel biomarkers of ALK-positive non-small cell lung cancer and predictors of response to crizotinib therapy. *Oncotarget*, 8(28):45399–45414.

Lopez-Rincon, A., Mendoza-Maldonado, L., Martinez-Archundia, M., Schönhuth, A., Kraneveld, A. D., Garssen, J., and Tonda, A. (2020). Machine Learning-Based Ensemble Recursive Feature Selection of Circulating miRNAs for Cancer Tumor Classification. *Cancers*, 12(7):1785.

Lynam-Lennon, N., Maher, S. G., and Reynolds, J. V. (2009). The roles of microRNA in cancer and apoptosis. *Biological Reviews of the Cambridge Philosophical Society*, 84(1):55–71.

Ma, L., Huang, Y., Zhu, W., Zhou, S., Zhou, J., Zeng, F., Liu, X., Zhang, Y., and Yu, J. (2011). An Integrated Analysis of miRNA and mRNA Expressions in Non-Small Cell Lung Cancers. *PLOS ONE*, 6(10):e26502.

Marzi, M. J., Montani, F., Carletti, R. M., Dezi, F., Dama, E., Bonizzi, G., Sandri, M. T., Rampinelli, C., Bellomi, M., Maisonneuve, P., Spaggiari, L., Veronesi, G., Bianchi, F., Di Fiore, P. P., and Nicassio, F. (2016). Optimization and Standardization of Circulating MicroRNA Detection for Clinical Application: The miR-Test Case. *Clinical Chemistry*, 62(5):743–754.

Mitchell, K. A., Zingone, A., Toulabi, L., Boeckelman, J., and Ryan, B. M. (2017). Comparative Transcriptome Profiling Reveals Coding and Noncoding RNA Differences in NSCLC from African Americans and European Americans. *Clinical Cancer Research*, 23(23):7412–7425.

Nigita, G., Distefano, R., Veneziano, D., Romano, G., Rahman, M., Wang, K., Pass, H., Croce, C. M., Acunzo, M., and Nana-Sinkam, P. (2018). Tissue and exosomal miRNA editing in Non-Small Cell Lung Cancer. *Scientific Reports*, 8(1):10222.

Niu, Y., Su, M., Wu, Y., Fu, L., Kang, K., Li, Q., Li, L., Hui, G., Li, F., and Gou, D. (2019). Circulating Plasma miRNAs as Potential Biomarkers of Non–Small Cell Lung Cancer Obtained by High-Throughput Real-Time PCR Profiling. *Cancer Epidemiology and Prevention Biomarkers*, 28(2):327–336.

Patnaik, S. K., Kannisto, E. D., Mallick, R., Vachani, A., and Yendamuri, S. (2017). Whole blood microRNA expression may not be useful for screening non-small cell lung cancer. *PLOS ONE*, 12(7):e0181926.

Patnaik, S. K., Yendamuri, S., Kannisto, E., Kucharczuk, J. C., Singhal, S., and Vachani, A. (2012). MicroRNA Expression Profiles of Whole Blood in Lung Adenocarcinoma. *PLOS ONE*, 7(9):e46045.

Petriella, D., De Summa, S., Lacalamita, R., Galetta, D., Catino, A., Logroscino, A. F., Palumbo, O., Carella, M., Zito, F. A., Simone, G., and Tommasi, S. (2016). miRNA profiling in serum and tissue samples to assess noninvasive biomarkers for NSCLC clinical outcome. *Tumor Biology*, 37(4):5503–5513.

Pritchard, C. C., Cheng, H. H., and Tewari, M. (2012). MicroRNA profiling: Approaches and considerations. *Nature Reviews Genetics*, 13(5):358–369.

Qu, L., Li, L., Zheng, X., Fu, H., Tang, C., Qin, H., Li, X., Wang, H., Li, J., Wang, W., Yang, S., Wang, L., Zhao, G., Lv, P., Lei, Y., Zhang, M., Gao, H., Song, S., and Liu, X. (2017). Circulating plasma microRNAs as potential markers to identify EGFR mutation status and to monitor epidermal growth factor receptor-tyrosine kinase inhibitor treatment in patients with advanced non-small cell lung cancer. *Oncotarget*, 8(28):45807–45824.

Reis, P. P., Drigo, S. A., Carvalho, R. F., Lopez Lapa, R. M., Felix, T. F., Patel, D., Cheng, D., Pintilie, M., Liu, G., and Tsao, M.-S. (2020). Circulating miR-16-5p, miR-92a-3p, and miR-451a in Plasma from Lung Cancer Patients: Potential Application in Early Detection and a Regulatory Role in Tumorigenesis Pathways. *Cancers*, 12(8):2071.

Shen, Y., Wang, T., Yang, T., Hu, Q., Wan, C., Chen, L., and Wen, F. (2013). Diagnostic Value of Circulating microRNAs for Lung Cancer: A Meta-Analysis. *Genetic Testing and Molecular Biomarkers*, 17(5):359–366.

Shopland, D. R., Eyre, H. J., and Peachacek, T. F. (1991). Smoking-Attributable Cancer Mortality in 1991: Is Lung Cancer Now the Leading Cause of Death Among Smokers in the United States? *JNCI: Journal of the National Cancer Institute*, 83(16):1142–1148.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249.

Uddin, A. and Chakraborty, S. (2018). Role of miRNAs in lung cancer. *Journal of Cellular Physiology*.

van den Berg, M. M. J., Krauskopf, J., Ramaekers, J. G., Kleinjans, J. C. S., Prickaerts, J., and Briedé, J. J. (2020). Circulating microRNAs as potential biomarkers for psychiatric and neurodegenerative disorders. *Progress in Neurobiology*, 185:101732.

Walser, T., Cui, X., Yanagawa, J., Lee, J. M., Heinrich, E., Lee, G., Sharma, S., and Dubinett, S. M. (2008). Smoking and Lung Cancer. *Proceedings of the American Thoracic Society*, 5(8):811–815.

Walters, C., Harter, Z. J., Wayant, C., Vo, N., Warren, M., Chronister, J., Tritz, D., and Vassar, M. (2019). Do oncology researchers adhere to reproducible and transparent principles? A cross-sectional survey of published oncology literature. *BMJ Open*, 9(12):e033962.

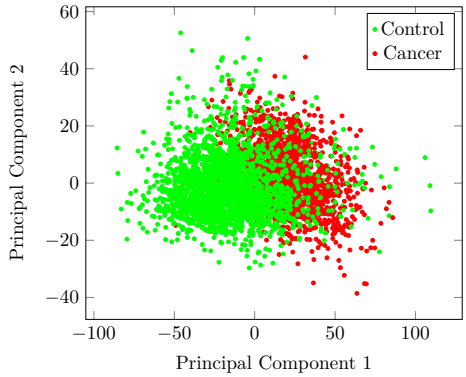
Wozniak, M. B., Scelo, G., Muller, D. C., Mukeria, A., Zaridze, D., and Brennan, P. (2015). Circulating MicroRNAs as Non-Invasive Biomarkers for Early Detection of Non-Small-Cell Lung Cancer. *PLOS ONE*, 10(5):e0125026.

Yao, B., Qu, S., Hu, R., Gao, W., Jin, S., Liu, M., and Zhao, Q. (2019). A panel of miRNAs derived from plasma extracellular vesicles as novel diagnostic biomarkers of lung adenocarcinoma. *FEBS Open Bio*, 9(12):2149–2158.

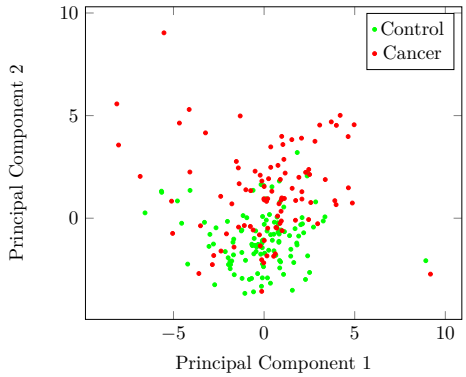
Zaporozhchenko, I. A., Morozkin, E. S., Ponomaryova, A. A., Rykova, E. Y., Cherdyntseva, N. V., Zheravin, A. A., Pashkovskaya, O. A., Pokushalov, E. A., Vlassov, V. V., and Laktionov, P. P. (2018). Profiling of 179 miRNA Expression in Blood Plasma of Lung Cancer Patients and Cancer-Free Individuals. *Scientific Reports*, 8(1):6348.

Zhong, S., Golpon, H., Zardo, P., and Borlak, J. (2021). miRNAs in lung cancer. A systematic review identifies predictive and prognostic miRNA candidates for precision medicine in lung cancer. *Translational Research*, 230:164–196.

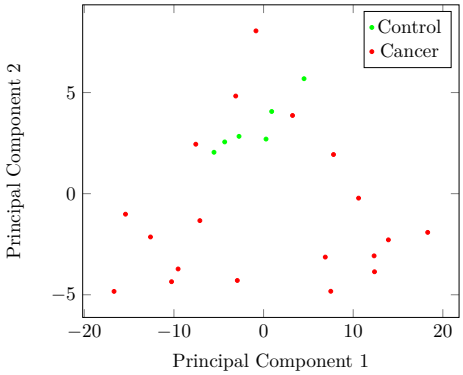
Appendices



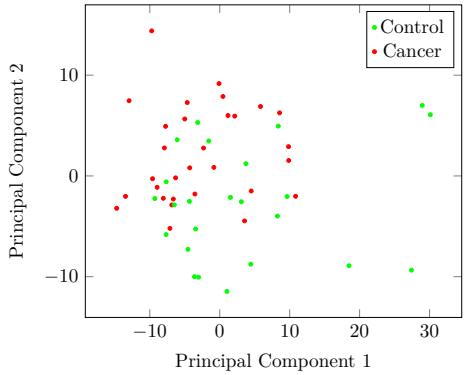
(a) PCA of Asakura et al. [2020]



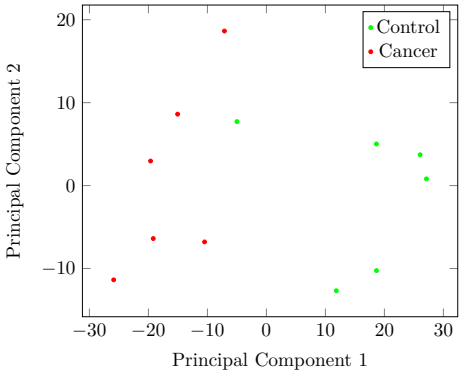
(b) PCA of Bianchi et al. [2011]



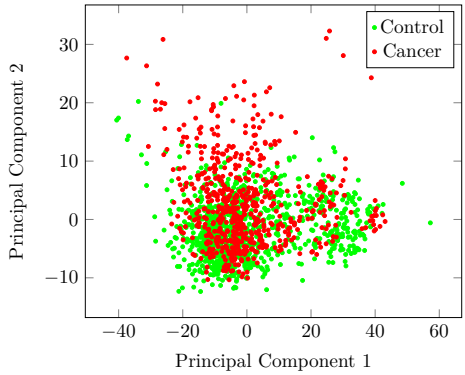
(c) PCA of Boeri et al. [2011]



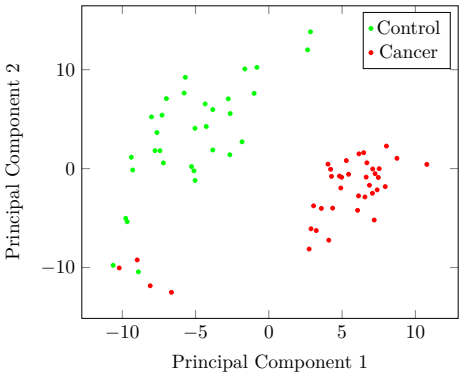
(d) PCA of Chen et al. [2019]



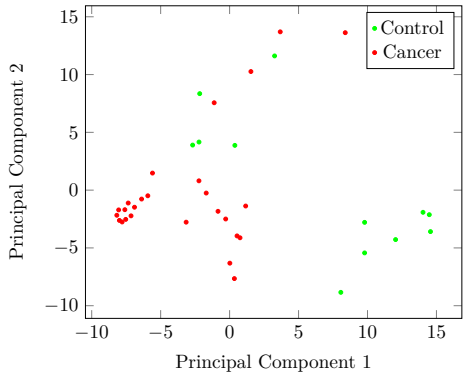
(e) PCA of Duan et al. [2021]



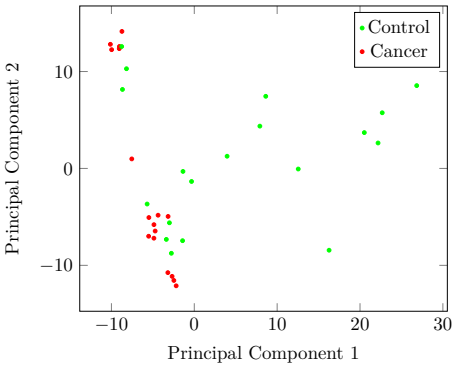
(f) PCA of Fehlmann et al. [2020]



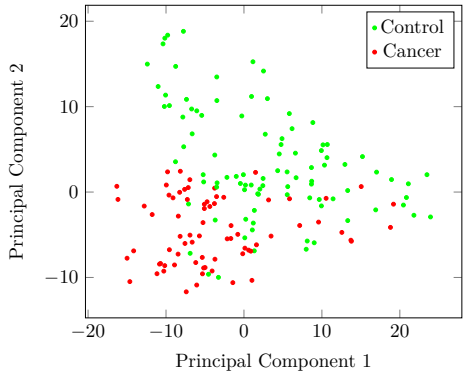
(g) PCA of Halvorsen et al. [2016]



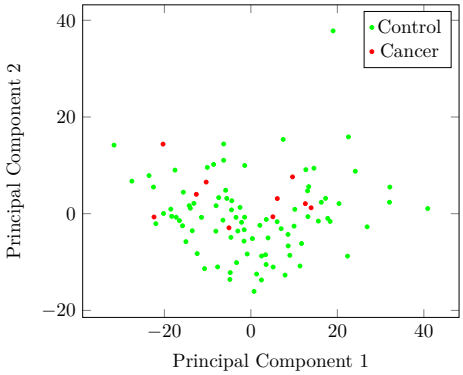
(h) PCA of Jin et al. [2017]



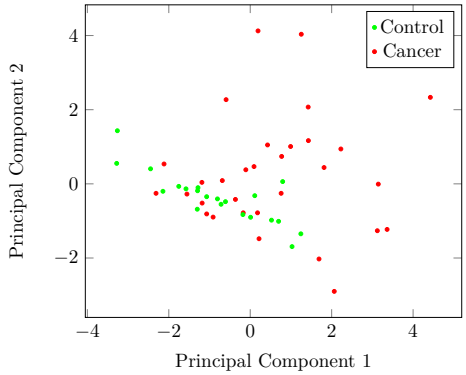
(i) PCA of Keller et al. [2009]



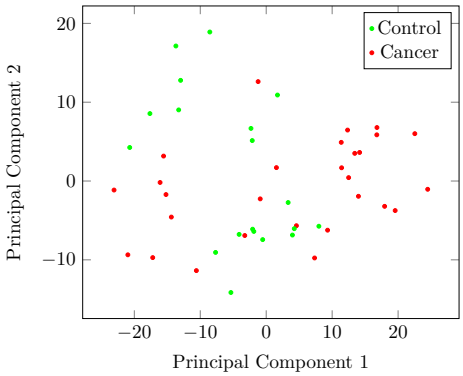
(j) PCA of Keller et al. [2014]



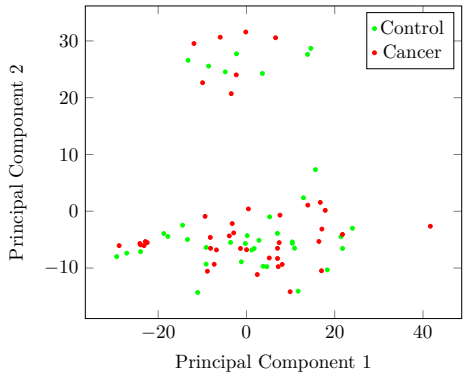
(k) PCA of Keller et al. [2020]



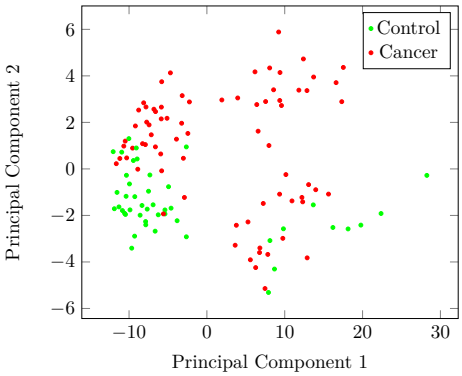
(l) PCA of Kryczka et al. [2021]



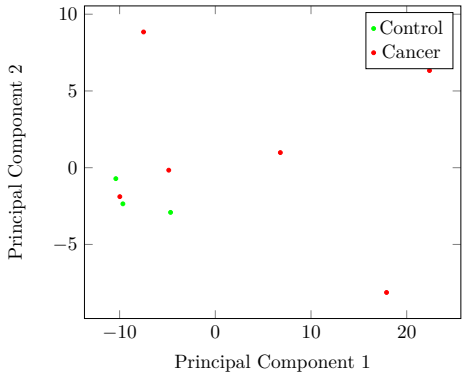
(m) PCA of Leidinger et al. [2011]



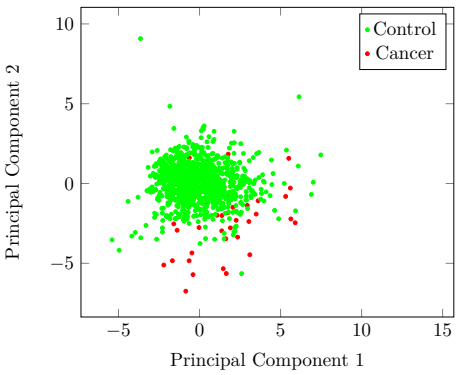
(n) PCA of Leidinger et al. [2014]



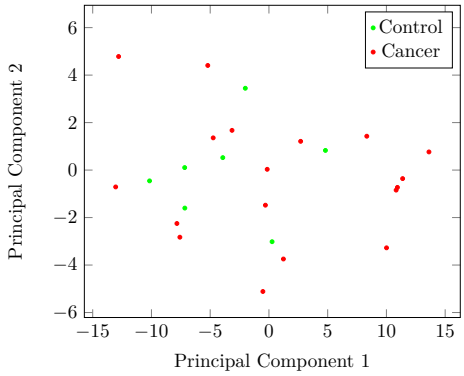
(o) PCA of Leidinger et al. [2015]



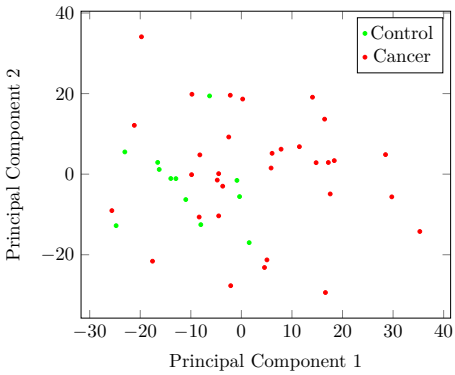
(p) PCA of Li et al. [2017]



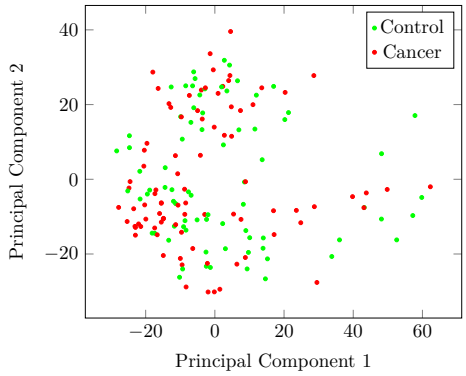
(q) PCA of Marzi et al. [2016]



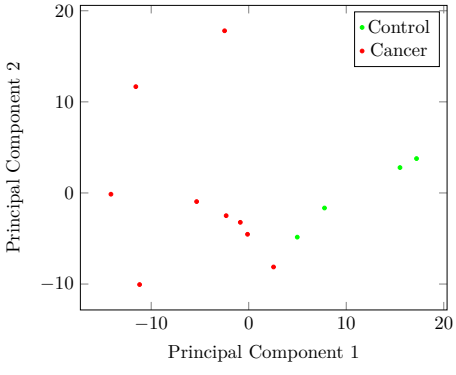
(r) PCA of Nigita et al. [2018]



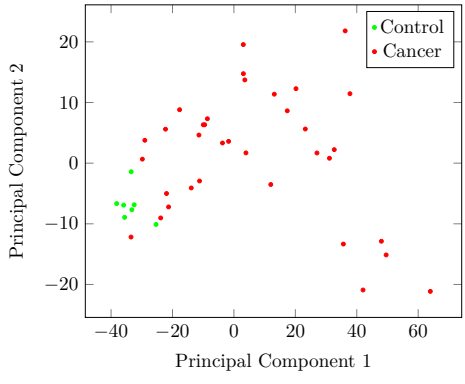
(s) PCA of Patnaik et al. [2012]



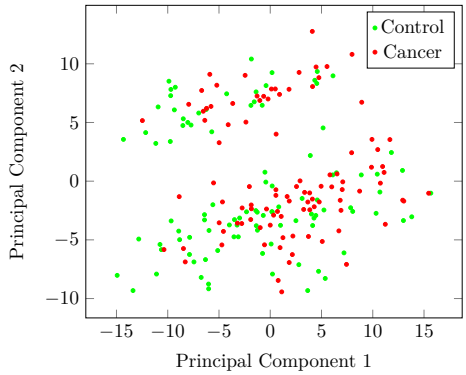
(t) PCA of Patnaik et al. [2017]



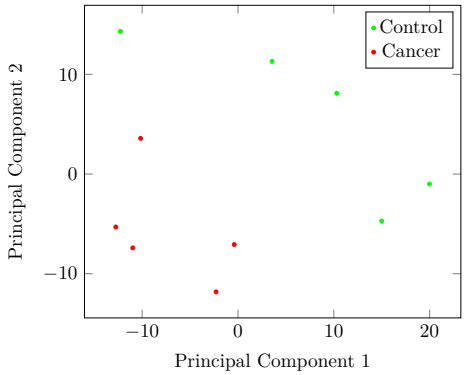
(u) PCA of Qu et al. [2017]



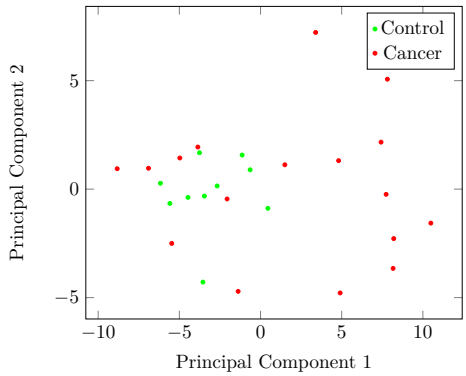
(v) PCA of Reis et al. [2020]



(w) PCA of Wozniak et al. [2015]



(x) PCA of Yao et al. [2019]



(y) PCA of Zaporozhchenko et al. [2018]

Figure 5.1: PCA plots

Table 5.1: The log-fold-change correlation between studies that have at least 10 miRNA-sequences in common

	Asakura et al. [2020]	Bianchi et al. [2011]	Boeri et al. [2011]	Chen et al. [2019]	Duan et al. [2021]	Fehlmann et al. [2020]	Halvorsen et al. [2016]	Jin et al. [2017]	Keller et al. [2009]	Keller et al. [2014]	Keller et al. [2020]	Kryczka et al. [2021]	Leidinger et al. [2011]	Leidinger et al. [2014]	Leidinger et al. [2015]	Li et al. [2017]	Marzi et al. [2016]	Nigita et al. [2018]	Patnaik et al. [2012]	Patnaik et al. [2017]	Qu et al. [2017]	Reis et al. [2020]	Wozniak et al. [2015]	Yao et al. [2019]	Zaporozhchenko et al. [2018]
Asakura et al. [2020]	*	-0.13	0.09	0.03	-0.08	-0.12	0.09	-0.11	-0.01	0.19	-0.12	*	-0.13	-0.09	0.12	0.16	0.18	0.09	-0.07	-0.01	-0.09	-0.09	0.15	0.00	-0.01
Bianchi et al. [2011]	-0.13	*	-0.26	0.03	*	0.24	0.21	0.24	0.42	-0.28	*	*	0.35	0.04	-0.28	*	-0.71	0.31	0.14	0.09	*	0.00	-0.49	0.25	*
Boeri et al. [2011]	0.09	-0.26	*	0.31	-0.04	-0.05	-0.06	-0.00	-0.18	-0.01	0.31	*	-0.02	0.09	0.00	0.51	0.25	0.11	-0.27	0.03	0.56	0.04	-0.04	0.22	0.13
Chen et al. [2019]	0.03	0.03	0.31	*	0.11	-0.35	0.00	-0.24	-0.07	-0.22	0.05	0.18	-0.00	0.05	0.37	0.00	-0.20	0.03	-0.19	-0.00	-0.17	-0.11	0.10	0.20	0.28
Duan et al. [2021]	-0.08	*	-0.04	0.11	*	0.00	0.16	0.00	0.04	-0.16	0.20	*	0.03	0.00	-0.12	-0.17	*	0.25	-0.09	-0.03	-0.04	0.11	0.06	0.15	-0.43
Fehlmann et al. [2020]	-0.12	0.24	-0.05	-0.35	0.00	*	-0.17	0.00	0.24	-0.36	0.01	*	0.25	0.07	-0.56	-0.38	0.36	-0.13	0.32	0.09	0.36	0.31	-0.19	0.14	-0.07
Halvorsen et al. [2016]	0.09	0.21	-0.06	-0.24	0.16	-0.17	*	0.12	0.10	0.11	0.35	*	-0.14	-0.07	0.24	-0.04	0.28	-0.17	-0.07	0.05	-0.32	-0.08	0.01	-0.11	-0.00
Jin et al. [2017]	-0.11	0.24	-0.00	-0.07	0.00	0.00	0.12	*	0.07	0.05	0.40	*	-0.03	0.03	0.00	0.03	-0.24	-0.01	-0.19	-0.06	-0.11	-0.13	0.03	0.08	-0.22
Keller et al. [2009]	-0.01	0.42	-0.18	-0.22	0.04	0.24	0.10	0.07	*	-0.33	-0.07	*	0.30	-0.04	-0.55	-0.16	-0.07	-0.09	0.13	0.12	0.04	0.09	-0.09	0.01	0.05
Keller et al. [2014]	0.19	-0.28	-0.01	0.05	-0.16	-0.36	0.11	0.05	-0.33	*	0.11	*	-0.56	0.03	0.53	0.14	-0.03	0.11	-0.09	-0.08	-0.09	0.01	0.01	-0.16	0.03
Keller et al. [2020]	-0.12	*	0.31	0.18	0.20	0.01	0.35	0.40	-0.07	0.11	*	*	-0.00	0.09	-0.18	-0.02	*	-0.20	0.12	-0.05	-0.01	0.19	-0.16	0.21	-0.33
Kryczka et al. [2021]	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
Leidinger et al. [2011]	-0.13	0.35	-0.02	-0.00	0.03	0.25	-0.14	-0.03	0.30	-0.56	-0.00	*	*	-0.14	-0.43	-0.02	-0.13	-0.01	0.01	0.03	0.19	-0.06	0.04	0.11	-0.04
Leidinger et al. [2014]	-0.09	0.04	0.09	0.05	0.00	0.07	-0.07	0.03	-0.04	0.03	0.09	*	-0.14	*	0.38	0.10	-0.47	-0.15	-0.01	0.02	0.22	0.05	0.09	-0.08	0.33
Leidinger et al. [2015]	0.12	-0.28	0.00	0.37	-0.12	-0.56	0.24	0.00	-0.55	0.53	-0.18	*	-0.43	0.38	*	0.31	0.24	0.12	-0.40	-0.15	-0.01	-0.39	0.22	-0.21	0.18
Li et al. [2017]	0.16	*	0.51	0.00	-0.17	-0.38	-0.04	0.03	-0.16	0.14	-0.02	*	-0.02	0.10	0.31	*	*	-0.14	-0.09	0.08	-0.06	-0.19	0.24	-0.00	*
Marzi et al. [2016]	0.18	-0.71	0.25	-0.20	*	0.36	0.28	-0.24	-0.07	-0.03	*	*	-0.13	-0.47	0.24	*	*	*	0.18	-0.21	*	0.42	0.01	0.32	*
Nigita et al. [2018]	0.09	0.31	0.11	0.03	0.25	-0.13	-0.17	-0.01	-0.09	0.11	-0.20	*	-0.01	-0.15	0.12	-0.14	*	*	-0.03	-0.19	0.10	-0.00	-0.15	0.08	0.01
Patnaik et al. [2012]	-0.07	0.14	-0.27	-0.19	-0.09	0.32	-0.07	-0.19	0.13	-0.09	0.12	*	0.01	-0.01	-0.40	-0.09	0.18	-0.03	*	-0.12	0.54	0.24	-0.31	-0.24	-0.08
Patnaik et al. [2017]	-0.01	0.09	0.03	-0.00	-0.03	0.09	0.05	-0.06	0.12	-0.08	-0.05	*	0.03	0.02	-0.15	0.08	-0.21	-0.19	-0.12	*	-0.00	-0.05	0.02	-0.09	-0.07
Qu et al. [2017]	-0.09	*	0.56	-0.17	-0.04	0.36	-0.32	-0.11	0.04	-0.09	-0.01	*	0.19	0.22	-0.01	-0.06	*	0.10	0.54	-0.00	*	0.43	*	0.22	*
Reis et al. [2020]	-0.09	0.00	0.04	-0.11	0.11	0.31	-0.08	-0.13	0.09	0.01	0.19	*	-0.06	0.05	-0.39	-0.19	0.42	-0.00	0.24	-0.05	0.43	*	-0.34	0.20	-0.18
Wozniak et al. [2015]	0.15	-0.49	-0.04	0.10	0.06	-0.19	0.01	0.03	-0.09	0.01	-0.16	*	0.04	0.09	0.22	0.24	0.01	-0.15	-0.31	0.02	-0.51	-0.34	*	-0.05	0.09
Yao et al. [2019]	0.00	0.25	0.22	0.20	0.15	0.14	-0.11	0.08	0.01	-0.16	0.21	*	0.11	-0.08	-0.21	-0.00	0.32	0.08	-0.24	-0.09	0.22	0.20	-0.05	*	-0.07
Zaporozhchenko et al. [2018]	-0.01	*	0.13	0.28	-0.43	-0.07	-0.00	-0.22	0.05	0.03	-0.33	*	-0.04	0.33	0.18	*	*	0.01	-0.08	-0.07	*	-0.18	0.09	-0.07	*

	Asakura et al. [2020]	Bianchi et al. [2011]	Boeri et al. [2011]	Chen et al. [2019]	Duan et al. [2021]	Fehlmann et al. [2020]	Halvorsen et al. [2016]	Jin et al. [2017]	Keller et al. [2009]	Keller et al. [2014]	Keller et al. [2020]	Kryczka et al. [2021]	Leidinger et al. [2011]	Leidinger et al. [2014]	Leidinger et al. [2015]	Li et al. [2017]	Marzi et al. [2016]	Nigita et al. [2018]	Patnaik et al. [2012]	Patnaik et al. [2017]	Qu et al. [2017]	Reis et al. [2020]	Wozniak et al. [2015]	Yao et al. [2019]	Zaporozhchenko et al. [2018]
Asakura et al. [2020]	*	0.51	0.34	0.27	0.78	0.77	0.76	0.60	0.72	0.74	0.24	*	0.34	0.31	0.67	0.21	0.59	0.36	0.32	0.81	0.46	0.83	0.66	0.53	0.69
Bianchi et al. [2011]	0.33	*	0.47	0.52	*	0.67	0.61	0.52	0.68	0.37	*	*	0.75	0.46	0.57	*	0.38	0.72	0.49	0.72	*	0.44	0.35	0.66	*
Boeri et al. [2011]	0.69	0.20	*	0.52	0.48	0.14	0.37	0.40	0.39	0.35	0.59	*	0.44	0.72	0.31	0.76	0.71	0.70	0.45	0.55	0.60	0.43	0.31	0.62	0.48
Chen et al. [2019]	0.28	0.53	0.63	*	0.64	0.32	0.34	0.47	0.48	0.36	0.52	*	0.61	0.55	0.49	0.45	0.41	0.58	0.53	0.25	0.46	0.41	0.37	0.61	0.54
Duan et al. [2021]	0.66	*	0.58	0.22	*	0.19	0.66	0.88	0.38	0.30	0.80	*	0.5	0.80	0.33	0.05	*	0.58	0.02	0.83	0	0.88	0.61	0.94	0.27
Fehlmann et al. [2020]	0.65	0.42	0.52	0.36	0.49	*	0.38	0.43	0.63	0.34	0.59	*	0.62	0.58	0.43	0.44	0.69	0.46	0.63	0.67	0.58	0.54	0.41	0.57	0.47
Halvorsen et al. [2016]	0.73	0.67	0.56	0.20	0.64	0.60	*	0.70	0.52	0.89	0.69	*	0.40	0.49	0.83	0.24	0.59	0.16	0.79	0.53	0.30	0.39	0.59	0.33	0.29
Jin et al. [2017]	0.31	0.53	0.35	0.59	0.65	0.32	0.90	*	0.57	0.49	0.72	*	0.41	0.42	0.87	0.49	0.25	0.46	0.06	0.35	0.16	0.35	0.24	0.69	0.45
Keller et al. [2009]	0.40	0.50	0.37	0.43	0.72	0.79	0.74	0.62	*	0.05	0.27	*	0.97	0.53	0.14	0.24	0.57	0.24	0.57	0.61	0.48	0.85	0.47	0.65	0.49
Keller et al. [2014]	0.71	0.44	0.35	0.44	0.41	0.42	0.56	0.57	0.13	*	0.76	*	0.19	0.39	0.55	0.61	0.45	0.56	0.35	0.28	0.54	0.36	0.73	0.31	0.48
Keller et al. [2020]	0.35	*	0.52	0.45	0.62	0.28	0.62	0.77	0.19	0.83	*	*	0.38	0.74	0.40	0.53	*	0.57	0.49	0.38	0.35	0.20	0.38	0.69	0.29
Kryczka et al. [2021]	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
Leidinger et al. [2011]	0.36	0.43	0.58	0.53	0.55	0.66	0.41	0.57	0.77	0.13	0.55	*	*	0.39	0.45	0.42	0.53	0.60	0.51	0.51	0.55	0.70	0.30	0.56	0.62
Leidinger et al. [2014]	0.54	0.50	0.55	0.52	0.53	0.37	0.42	0.55	0.48	0.51	0.51	*	0.48	*	0.47	0.55	0.46	0.49	0.51	0.55	0.51	0.51	0.52	0.50	0.51
Leidinger et al. [2015]	0.54	0.27	0.30	0.65	0.55	0.44	0.73	0.71	0.25	0.71	0.30	*	0.35	0.43	*	0.35	0.55	0.38	0.27	0.57	0.27	0.57	0.76	0.59	0.75
Li et al. [2017]	0.66	*	0.88	0.22	0.22	0.5	0.33	0.5	0.33	0.55	0.66	*	0.5	0.61	0.33	*	*	0.61	0.88	0.72	0.88	0.16	0.5	0.38	*
Marzi et al. [2016]	0.78	0.21	0.71	0.47	*	0.57	0.53	0.33	0.63	0.41	*	*	0.30	0.78	0.79	*	*	*	0.60	0.34	*	0.81	0.63	0.86	*
Nigita et al. [2018]	0.46	0.54	0.69	0.62	0.42	0.19	0.31	0.41	0.19	0.75	0.57	*	0.57	0.30	0.45	0.63	*	*	0.54	0.45	0.68	0.30	0.25	0.59	0.24
Patnaik et al. [2012]	0.56	0.32	0.29	0.28	0.22	0.62	0.37	0.13	0.68	0.19	0.60	*	0.60	0.54	0.39	0.76	0.51	0.51	*	0.38	0.81	0.35	0.19	0.14	0.38
Patnaik et al. [2017]	0.51	0.47	0.48	0.42	0.48	0.51	0.53	0.46	0.56	0.46	0.39	*	0.51	0.54	0.46	0.56	0.52	0.41	0.43	*	0.46	0.44	0.51	0.44	0.51
Qu et al. [2017]	1	*	1	0.16	0.33	0.52	0.19	0.22	0.58	0.63	0.80	*	0.94	0.44	0.25	0.61	*	0.55	0.77	0.63	*	0.5	0.22	0.19	*
Reis et al. [2020]	0.84	0.75	0.24	0.46	0.92	0.47	0.34	0.50	0.97	0.10	0.41	*	0.93	0.80	0.19	0.04	0.79	0.05	0.88	0.75	0.60	*	0.34	0.89	0.57
Wozniak et al. [2015]	0.42	0.51	0.48	0.58	0.57	0.44	0.56	0.56	0.46	0.57	0.39	*	0.50	0.56	0.50	0.53	0.46	0.43	0.40	0.44	0.35	0.52	*	0.49	0.54
Yao et al. [2019]	0.88	0.4	0.56	0.64	0.72	0.56	0.24	0.76	0.6	0.04	0.36	*	0.72	0.24	0.2	0.48	0.72	0.68	0.12	0.6	0.68	0.72	0.76	*	0.16
Zaporozhchenko et al. [2018]	0.47	*	0.51	0.75	0.37	0.6	0.53	0.44	0.68	0.72	0.32	*	0.64	0.61	0.60	*	*	0.38	0.27	0.51	*	0.60	0.7	0.39	*

Table 5.2: The AUC when training logistic regression on the study in the row and doing inference on the study in the column when they have at least 10 miRNA-sequences in common