Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

Ole Fredrik Borgundvåg Berg

Circulating miRNA and Lung Cancer:

- a More Comprehensive Analysis of Available Data

Master's thesis in Informatics
Supervisor: Pål Sætrom
May 2022

NTNU
Kunnskap for en bedre verden

Ole Fredrik Borgundvåg Berg

Circulating miRNA and Lung Cancer:

- a More Comprehensive Analysis of Available Data

Master's thesis in Informatics
Supervisor: Pål Sætrom
May 2022

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Abstract

Background: This report looks at the possibility of diagnosing lung cancer using circulating miRNA. There has been a lot of research in this field, but little research from a machine learning perspective.

Motivation: Using machine learning to diagnose lung cancer is practical as current methods for diagnosing lung cancer are resource-intensive and the tumor is typically found at a late stage when the survival rate is low.

Experiments: I tried to collect all available datasets on circulating miRNA and lung cancer. Then I tried to find whether there were any patterns in case-control characteristics using different statistical tests. This includes trying to find the correlation in log fold change, looking at the proportion of miRNAs that were differentially expressed in the same way, hierarchical clustering of the datasets and looking at the consistency in differential expression of miRNAs that meta-analyses have found to be predictive of lung cancer. I have done machine learning internally in the different datasets and externally across multiple datasets. There were some attempts at trying to find higher consistency, including setting an RPM threshold for sequencing data and removing principal components conjectured to be noise. I also made a web application for visualizing the data in the different datasets.

Contributions: The main contributions of this project are to make all available datasets on circulating miRNA and lung cancer into a common format so that the work can be built upon by other researchers, and a web application that can be used by researchers to visualize the data.

Results: The result of this project is that I was not able to find any patterns in case-control characteristics that could replicate across datasets, with only a few exceptions. Furthermore, machine learning across different datasets was not able to learn any patterns in most cases, despite good results when using machine learning internally in a dataset. The most important exceptions were that (1) stratification of datasets sometimes gave significant improvement in AUC when using machine learning across datasets, (2) using an RPM threshold on sequencing dataset lead to high AUC across the sequencing datasets, (3) model predictions correlated significantly with case status even when average AUC was close to 0.50 and (4) case status contributed to the plurality of variance in model predictions in a PCA analysis.

Conclusions: Oncology is a field with a low replication rate, which means that it is important to try to replicate results in order to ensure that they are valid. This project tried to do this in connection with diagnosis of lung cancer using circulating miRNAs, and found that findings in single studies rarely have external validity.

Sammendrag

Bakgrunn: Denne rapporten ser på mulighetene for å diagnostisere lungekreft ved hjelp av sirkulerende miRNA. Det har vært mye forskning på dette feltet, men lite forskning fra et maskinlæringsperspektiv.

Motivasjon: Å bruke maskinlæring for å diagnostisere lungekreft er praktisk siden dagens metoder for å diagnostisere lungekreft er ressurskrevende og svulsten blir vanligvis oppdaget ved senstadium når overlevelsesraten er lav.

Eksperimenter: Jeg prøvde å samle alle tilgjengelige datasett om sirkulerende miRNA og lungekreft. Deretter prøvde jeg å finne ut om det var noen mønstre knyttet til kasus/kontroll-status ved å bruke forskjellige statistiske tester. Dette inkluderer å finne korrelasjonen i log2 foldendring, å finne andelen av miRNA-sekvenser som ble differensielt uttrykt likt, hierarkisk klynging av datasettene og å finne hvor konsistent det differensielle uttrykket var for de miRNAene som metaanalysene har funnet kan prediktere lungekreft. Jeg har gjort maskinlæring internt i de ulike datasettene og eksternt på tvers av flere datasett. Det var noen forsøk på å finne høyere konsistens, inkludert å sette en nedre terskel på gjennomsnittlig RPM for sekvenseringsdata og å fjerne prinsipalkomponenter som ble antatt å skyldes støy. Jeg har også laget en webapplikasjon for å visualisere dataene fra de forskjellige datasettene.

Bidrag: Hovedbidraget til dette prosjektet er å omgjøre alle tilgjengelige datasett om sirkulerende miRNA og lungekreft til et felles format slik at dette prosjektet kan bygges videre på av andre forskere, samt en webapplikasjon som kan brukes av forskere for å visualisere dataene.

Resultater: Resultatet av dette prosjektet er at jeg ikke var i stand til å finne noen mønstre i differensielt kasus/kontroll-uttrykk som replikerte på tvers av datasett, med noen få unntak. Videre var maskinlæring på tvers av ulike datasett i de fleste tilfeller ikke i stand til å finne noen mønstre, til tross for gode resultater ved bruk av maskinlæring internt i enkeltdatasett. De viktigste mønstrene som ble funnet var at (1) stratifisering av datasett noen ganger ga signifikant forbedring i AUC ved maskinlæring på tvers av datasett, (2) det å sette en nedre terskel på gjennomsnittlig RPM i sekvenseringsdatasettene førte til høy AUC på tvers av sekvenseringsdatasettene, (3) prediksjoner fra modellene korrelerte signifikant med kasus/kontroll-status selv når gjennomsnittlig AUC var nær 0,50 og (4) kasus/kontroll-status bidro til pluraliteten av variasjon i modellprediksjoner i en PCA-analyse.

Konklusjon: Onkologi er et forskningsfelt med lav replikasjonsrate, noe som betyr at det er viktig å prøve å replikere resultater for å sikre at de er gyldige. Dette prosjektet forsøkte å gjøre dette i forbindelse med diagnostisering av lungekreft ved bruk av sirkulerende miRNA, og fant at resultater i enkeltstudier sjelden har ekstern validitet.

Preface

> Overall the net consequence of hospitals is negative. Now that is just a guess, and it could easily be wrong, but it also could not be wrong.
>
> *Jordan B. Peterson*
> Professor emeritus in psychology
> at the University of Toronto

This is a master's thesis for a master's degree in informatics with a specialization in artificial intelligence, conducted at NTNU with Pål Sætrom as supervisor. I want to thank friends and family for their support.

Ole Fredrik Borgundvåg Berg
Trondheim, May 30, 2022

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This report contains further analysis of the data on circulating miRNA and lung cancer that were collected in Berg [2021].

1.1 Background and Motivation

Lung cancer is a dangerous disease that takes many lives and that has a low survival rate (see subsection 2.1.1). It has been suggested by several studies that circulating miRNA can be used to diagnose lung cancer in humans [Shen et al., 2013]. This can be useful as it is a less resource-intensive way of diagnosing lung cancer compared to e.g. CT scans. In addition, it can be used to diagnose lung cancer earlier, which could lead to a higher survival rate.

There have been done many studies on the connection between circulating miRNA and lung cancer. Some meta-analyses have found some consistency in what miRNAs are up- or down-regulated in lung cancer [Zhong et al., 2021; Huang et al., 2021; Jiang et al., 2018; Yi et al., 2021], whilst Berg [2021] found little consistency when trying to use machine learning to find patterns across datasets. However, the analyses in Berg [2021] were very naïve. The goal of this project is to do a more thorough analysis of the data to try to find more subtle patterns in the datasets.

1.2 Goals and Research Questions

There are some goals that will guide what is done in this project. As this project is a continuation of Berg [2021], the overall goal is the same:

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

There were some attempts to achieve this goal in Berg [2021], but the results were generally poor. However, the experiments had huge limitations, and there is thus a need for more thorough experiments in this project.

One of the reasons for collecting all the datasets in Berg [2021] was that it would be easy for other researchers to build on top of the work, as they could use the transformed datasets and thus save time and effort. I would like to formalize this into a goal: make the data easily available to third parties.

I will do some revisions on the research questions in Berg [2021] as the results in that project often revealed some implicit assumptions in the questions that were either taken for granted or not considered. The first research question was "Are there machine learning algorithms that generally perform better at diagnosing lung cancer based on miRNA values?". The question is interesting, but one thing that is taken for granted in the question is that it is possible to diagnose lung cancer based on miRNA data at all. It would be a reasonable assumption, as there is plenty of literature on the subject. On the other hand, almost all of the literature is studies that are based on single datasets, and Berg found that the datasets were very different and that it was hard for a machine learning model that was trained on one dataset to do well on another dataset. Berg also found that there was little consistency in the fold change of the different miRNAs.

Even if there is no reproducible differential expression of miRNA in cases versus controls, one could compare machine learning algorithms' internal performance in the different datasets. If one algorithm has generally better internal performance in the different datasets, one might say that this algorithm is better at diagnosing lung cancer based on miRNA-levels. However, this would not be very helpful, as there is no external diagnostic value. Thus, I will rather change the question to be more cautious.

Research question 1: To what extent can one find patterns in case-controls characteristics that generalize across different datasets?

The second research question in Berg [2021] was "Will a combined dataset lead have better diagnostic value than each of the datasets alone?". This question is also hard to answer given the results in Berg [2021]. One generally gets better results when doing training and testing inside one dataset, but what is the diagnostic value if this internal model has no external validity?

I think this research question needs some revision. A better research question would be:

Research question 2: Using multiple datasets, to what extent can you find a model that is able to diagnose lung cancer in a new dataset, given that the new dataset has a reasonable quality?

It is hard to say whether that is possible, and it is a less fundamental question than the first question in that it requires the first question to be true. However, not finding any evidence does not imply that it is false. Therefore, a lack of evidence after researching the first question does not mean that I will not research the second question. However, due to the dependencies of the questions, I will research the first question first.

There are other minor research questions that might be answered through this project, but they are more interesting from a medical than a machine learning point of view:

- What is the level of quality of the different datasets?
- Do the miRNAs 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?

These minor questions are the same that were asked in Berg [2021].

1.3 Research Method

This project is primarily an experimental one, as one needs 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 miRNAs and lung cancer, as the outcomes of the machine learning might help in understanding the effect of lung cancer on miRNAs, 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, and technical details concerning some of the experiments. Chapter 4 is about how the experiments were performed and their results. Finally, chapter 5 is about the conclusions that are made from the results.

Chapter 2

Background Theory

This project is a cross-disciplinary one, as it combines machine learning and medicine, and as such, some theory from both disciplines is necessary in order to understand the project. Most of the subsections here were originally found in Berg [2021], as the necessary preliminaries are mostly the same as in the specialization project.

2.1 Biological Theory

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

2.1.1 Lung Cancer

Lung cancer is the second most common type of cancer worldwide, and the type of cancer with the highest total mortality worldwide, causing about 1.8 million deaths per year [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 minor 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 major subtypes: adenocarcinoma (30-40% of NSCLC cases), 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 of 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 completely [American Cancer Society, 2021].

2.1.2 MicroRNA

MicroRNAs (miRNAs) are short sequences of RNA, about 22 nucleotides each, that regulate 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 "miR" as prefix and a unique number as suffix. 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 roles of miRNAs in relation to lung cancer are not fully understood [Uddin and Chakraborty, 2018]. MicroRNAs are thought to function both as tumor suppressor genes and as oncogenes, and tumor miRNA expression profiles can distinguish tumors from normal tissue, distinguish tumor subtypes and predict survival [Lynam-Lennon et al., 2009]. Moreover, multiple studies report differential expression of circulating miRNA-sequences in cancer patients compared to

healthy controls, which 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 advantages and disadvantages.

2.1.4.1 qRT-PCR

Quantitative Reverse Transcription - Polymerase Chain Reaction (qRT-PCR) is a common method of measuring miRNA-levels. As the name implies, the process depends on reverse transcription, where miRNA is 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 is going to 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.

2.1.4.2 Microarrays

Microarrays are what is called a hybridization method. It starts similarly to qRT-PCR, by converting miRNA into cDNA, but the miRNA is fluorescently labeled in this case. The microarray has several spots, each with single-stranded DNA samples (called probes) that are mounted to the microarray. When the cDNA is added to the microarray, the cDNA will bind to the DNA samples that have the same sequence, in a process called hybridization. Afterward, 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 is then a measure of 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 are 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.

2.1.4.3 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 add 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.

2.1.5 Types of blood fractions

When doing blood profiling, there are different ways to process and filter the blood depending on what parts of the blood one wants in the final sample. There are three different main types.

2.1.5.1 Whole blood

Whole blood is the simplest type, as it is the blood that runs through your body without any filtering. It might be added anticoagulant, as blood will naturally start clotting outside of the human body unless something is done.

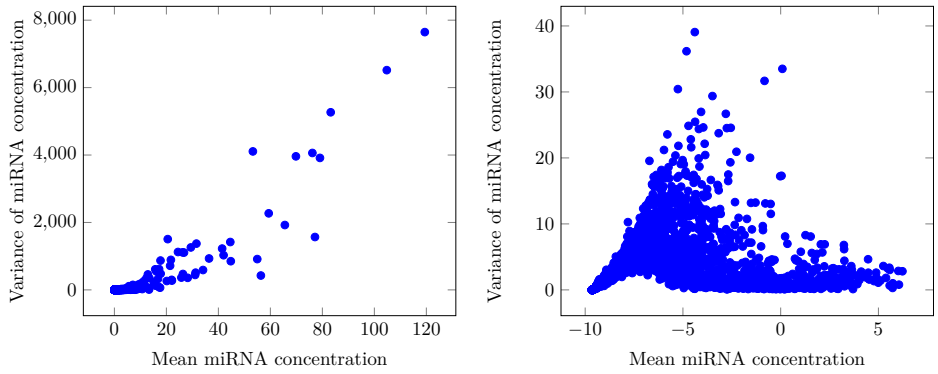
2.1.5.2 Serum

Here you let the blood clot, which means that no anticoagulant is added. When you let the blood clot you end up naturally with a liquid and a solid part, whereas the serum is the liquid part. The solid part has cells, including red blood cells that are thus filtered out. To separate the liquid and solid parts, centrifugation is often used.

2.1.5.3 Plasma

Here you add anticoagulant and then apply centrifugation. Again, the sample will separate into two parts, but here both parts are liquid. The heavy part with all the blood cells will fall to the bottom, and you get a clear liquid at the top. The clear liquid on the top is what is the plasma. One main difference from serum is that plasma contains fibrinogen, which is a protein that is converted to fibrin in blood clotting.

One important thing to note is that serum and plasma are very similar liquids, while whole blood has a different consistency as it contains many cells, especially red blood cells.



(a) Before log transformation

(b) After log transformation

Figure 2.1: Mean and variance in different miRNA-sequences in Duan et al. [2021]

2.2 Machine learning/statistical theory

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

2.2.1 Variance stabilizing transformation

In miRNA measurements, one often sees 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) assume that the underlying data has a normal distribution. A normal distribution, however, has no skew, therefore unskewing the data is necessary for ensuring that other methods are giving valid results. More formally, if we assume that $Y \sim g(X)$ 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 of levels of a 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 smoothing scatter plots [Cleveland, 1979]. The regression works by fitting a low degree polynomial for each data point. The fitting of each polynomial works by giving weight to nearby points that are used for fitting the polynomial, where more weight is given to points near the original data point. The regression value for each data point is thus the value of the corresponding polynomial evaluated in this point.

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

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 most of the variance of the original dataset is kept in the $1 \leq k \leq m$ first principal components, 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 $y = [y_1, y_2 \dots y_n]$,

covariates $\mathbf{X}$ and residuals $\epsilon \sim N(0, \sigma)$ have the relationship $y = \mathbf{X}\beta + \epsilon$, for some parameter vector β. Furthermore, let $\bar{y}$ be the mean of the y's, i.e $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

If one creates a linear regression model of the dataset, one gets 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 $\mathbf{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ as the total sum of squares, $\mathbf{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ as the sum of squares due to regression and $\mathbf{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ as the sum of the squared estimates 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 Wilcoxon signed-rank test

Wilcoxon signed-rank test is a statistical test to find the location of a distribution [Wilcoxon, 1945; Pratt and Gibbons, 1981]. More formally, if $\mathbf{X}_1$ and $\mathbf{X}_2$ are independently distributed with cumulative distribution function F, and

$$p = P\left(\frac{1}{2}(\mathbf{X}_1 + \mathbf{X}_2) > 0\right)$$

then the test is testing the null hypothesis $p = \frac{1}{2}$ against the alternative hypothesis $p \neq \frac{1}{2}$. The median of $\frac{1}{2}(\mathbf{X}_1 + \mathbf{X}_2)$ is called the pseudomedian of F. F is called symmetric about μ if $f(\mu + x) = f(\mu - x)$. Furthermore, F is called symmetric if there exists such a μ. If F is symmetric, the median and the pseudomedian are the same. Thus, if F is assumed to be symmetric, the null hypothesis is equivalent to that the median of F is $\mu = 0$, against the alternative hypothesis that $\mu \neq 0$.

If one has pairs $(\mathbf{X}, \mathbf{Y})$ where $\mathbf{X} \sim F_X$ and $\mathbf{Y} \sim F_Y$, with the null hypothesis $F_X = F_Y$, then under the null hypothesis $\mathbf{X} - \mathbf{Y}$ has a symmetric distribution (as the distribution is the same as $\mathbf{Y} - \mathbf{X}$ as $\mathbf{X}$ and $\mathbf{Y}$ are interchangeable) and it has a median of 0. Let $\mathbf{X} - \mathbf{Y}$ have a cumulative distribution function F_{X-Y}. If $F_X \neq F_Y$, it cannot be guaranteed that F_{X-Y} is symmetric. However, one can instead test the null hypothesis that F_{X-Y} is symmetric around 0 against the alternative hypothesis that F_{X-Y} has a median not equal to 0 (and is not necessarily symmetric).

2.2.7 Bonferroni correction

A Bonferroni correction is a correction that is done when doing multiple testing to avoid false positives [Bonferroni, 1936]. The correction is in order to control

the family-wise error rate (FWER). FWER is the probability of at least one type I error when testing several hypotheses simultaneously. The correction is based on Boole's inequality:

$$P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$$

for all possible events A_i. Letting A_i be the event that one makes a type I error in hypothesis i, and letting n be the number of hypotheses gives that FWER $= P(\bigcup_{i=1}^{n} A_i)$. Setting a significance level α_0 for each hypothesis gives that $P(A_i) \leq \alpha_0$. Then:

$$\text{FWER} = P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i) \leq n \cdot \alpha_0$$

If we want FWER $\leq \alpha$, this can be achieved by setting $\alpha_0 = \frac{\alpha}{n}$.

2.2.8 Interval using t-distribution

Given that I have n random variables $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ where $\mathbf{X}_i \sim N(\mu, \sigma^2)$, then what is the conditional distribution of $\mathbf{X}_k$, $k \in \{1 \ldots n\}$, if we are given the sample mean ($\bar{\mathbf{X}}$) and the sample variance ($S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)^2$)?

We have that $\mathbf{E}[\mathbf{X}_k - \bar{\mathbf{X}}] = \mathbf{E}[\mathbf{X}_k] - \mathbf{E}[\bar{\mathbf{X}}] = \mu - \mu = 0$. Furthermore:

$$\begin{aligned}
&\mathbf{Var}[\mathbf{X}_k - \bar{\mathbf{X}}] \\
=&\mathbf{Var}[\mathbf{X}_k - \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i] \\
=&\mathbf{Var}[-\frac{n-1}{n}\mathbf{X}_k + \frac{1}{n}\sum_{\substack{i=1\\ i\neq k}}^{n}\mathbf{X}_i] \\
=&\mathbf{Var}[-\frac{n-1}{n}\mathbf{X}_k] + \mathbf{Var}[\frac{1}{n}\sum_{\substack{i=1\\ i\neq k}}^{n}\mathbf{X}_i] \\
=&\frac{(n-1)^2}{n^2}\mathbf{Var}[\mathbf{X}_k] + \frac{1}{n^2}\sum_{\substack{i=1\\ i\neq k}}^{n}\mathbf{Var}[\mathbf{X}_i] \\
=&\frac{(n-1)^2}{n^2}\sigma^2 + \frac{n-1}{n^2}\sigma^2 \\
=&\frac{n-1}{n}\sigma^2
\end{aligned}$$

Thus:

$$\frac{\mathbf{X}_k - \bar{\mathbf{X}}}{\sqrt{\frac{n-1}{n}\sigma^2}} \sim N(0,1)$$

We know that $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$. And if $\mathbf{Z} \sim N(0,1)$ and $\mathbf{V} \sim \chi^2_v$ are independent then:

$$\frac{Z}{\sqrt{\frac{\mathbf{V}}{v}}} \sim T_v$$

Thus, this will be an approximation of a such interval:

$$\frac{\frac{\mathbf{X}_k - \bar{\mathbf{X}}}{\sqrt{\frac{n-1}{n}\sigma^2}}}{\sqrt{\frac{(n-1)\frac{S^2}{\sigma^2}}{n-1}}} = \frac{\mathbf{X}_k - \bar{\mathbf{X}}}{\sqrt{\frac{n-1}{n}S^2}} \sim T_{n-1}$$

It is only an approximation as $\mathbf{X}_k - \bar{\mathbf{X}}$ is not independent of S^2.
Finally, an approximate α-interval for $\mathbf{X}_k$ is:

$$-t_{1-\alpha/2,n-1} \leq \frac{\mathbf{X}_k - \bar{\mathbf{X}}}{\sqrt{\frac{n-1}{n}S^2}} \leq t_{1-\alpha/2,n-1}$$

$$\bar{\mathbf{X}} - t_{1-\alpha/2,n-1}\sqrt{\frac{n-1}{n}S^2} \leq \mathbf{X}_k \leq \bar{\mathbf{X}} + t_{1-\alpha/2,n-1}\sqrt{\frac{n-1}{n}S^2}$$

2.2.9 Logistic regression

Assume that you have Bernoulli 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.

2.2.10 Support Vector Machine

A support vector machine (SVM) is a machine learning algorithm that works by creating boundaries in high dimensional vector space [Cortes and Vapnik, 1995]. The easiest form of SVM is a binary linear SVM. Then the algorithm creates a hyperplane in the data space that separates the two classes in an optimal way, where different loss functions would lead to different hyperplanes being optimal. If the data is linearly separable, then a linear SVM would create a perfect separation.

However, the data is often not linearly separable, but might be separable with a more complex boundary. In these cases, one can use a kernel SVM. A kernel SVM would map the data onto a higher-dimensional space, where the data is linearly separable [Patle and Chouhan, 2013]. A common kernel, which is default in `scikit-learn`, is the radial basis function (RBF) kernel [Chang et al., 2010]. RBF is a kernel that is based on radial distances between points, where points have exponentially less influence on each other as the distance between points grows.

2.2.11 Random Forest

A decision tree is a tree-like model, where at each internal node the next node is further down the tree. Which node is next is based on the decision criterion in the current node. The decision criterion is a condition on your data point. Finally, leaf nodes have the classification of the data point. Decision tree learning is when one learns these criteria based on data to find an optimal classification [Hastie et al., 2009]. A random forest is a classifier where one trains multiple trees, where each three is trained on a subset of the training set. Having each tree being trained on only a subset of the training set is a way of reducing overfitting. The overall classification is then given by aggregating the results from the classification given by the different trees [Hastie et al., 2009].

2.2.12 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 to find optimal weights given to each of the weaker models [Friedman, 2001]. Instead of using the gradient directly, the algorithm just ensures 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 of 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].

[1] `https://www.kaggle.com`

Chapter 3

Methodology

For a description of the data processing, see Berg [2021].

This project consists of the following steps:

1. Checking if there are properties of the datasets (like the type of technology used for measuring miRNA) that contribute to the log fold change correlation.
2. Finding whether there is evidence that there are patterns in miRNA expression that are consistently found to be related to case-control characteristics.
3. Finding whether there are structural differences between the datasets by seeing whether a machine learning algorithm can distinguish samples from different datasets.
4. Finding datasets that are similar to each other using hierarchical clustering algorithms.
5. Do machine learning internally in each dataset, as it would be close to an upper threshold on the accuracy that can be found across datasets.
6. Try to find one or more miRNA-sequences that are consistent biomarkers for lung cancer using meta-analyses and the datasets.
7. PCA analysis using several datasets at once.
8. Machine learning on multiple datasets that are joined in different ways.
9. Looking for noise in PCA components.
10. Checking for optimal RPM thresholds in sequencing data.

11. Checking for red blood cells in the data.

12. Creating a web application for visualizing the data.

The methodology only includes experiments where additional details are needed. All experiments are described in the results.

3.1 Structured Literature Review Protocol

The literature review was done in Berg [2021] and the same explanation can be found there, but is also added here for completeness.

The point of the literature search was to find studies relevant to circulating miRNA and lung cancer. The main search engine used was PubMed[1], which is a commonly used search engine for medical literature. 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 searched databases that have public gene expression data, as described in Table 3.1.

Table 3.1: Search in public gene expression databases. The first column is the name of the database. The second column is the search term that was used to search the database.

Database name	**Search term**
ArrayExpress[2]	`microrna lung cancer`
Gene Expression Omnibus (GEO)[3]	`(mirna OR microrna) AND "lung cancer" AND (diagnosis OR detection)`
OmicsDI[4]	`"lung cancer" AND TAXONOMY: 9606 AND "breast cancer" AND (mirna OR microrna) AND (serum OR plasma OR "whole blood")`

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

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

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

Some of the studies measured miRNA levels in the lung tissue or in the 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 were 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 reasonable to only use circulating miRNA data.

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

The controls in some of the studies are not healthy, but suffer from other kinds of lung diseases. Other studies 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 find what miRNA changes are due to the lung cancer. On the other hand, people who are getting checked for lung cancer often have lung issues, which is the reason for their checkup, so distinguishing lung cancer from other illnesses is important.

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.

- Meta-analyses were used as a 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.

3.2 Technical setup

In Table 3.2, the main software used in this project is listed.

Table 3.2: Software used in this project. The first column is the name of the software, the second column is the version of the software and the last column is the usage of the software.

Software	Version	Usage
Python[5]	3.9.7	Programming language
NumPy[6]	1.20.3	Numerical calculations with vectors and matrices
scikit-learn[7]	0.24.2	Machine learning
XGBoost[8]	1.4.2	XGBoost machine learning algorithm
SciPy[9]	1.7.1	Scientific programming

3.3 Evidence for consistently differentially expressed miRNA-sequences

There are multiple ways to find out whether there are miRNAs that are consistently differentially expressed.

3.3.1 Paired sign test

A paired sign test is in this case an experiment trying to estimate the probability that a miRNA-sequence is differentially expressed in the same direction (up- or down-regulated in cancer) in two different datasets. This will be done by finding all pairs of studies where both have a given miRNA, and then find the differential expression of the given miRNA in the two studies. This will be done for all miRNAs, resulting in one pair for each time two studies have measured the same miRNA. By looking at all these pairs, it is possible to calculate the wanted probability by looking at the proportion of such pairs that have differential expressions in the same direction. One question is whether one should only consider pairs where both miRNAs are significantly differentially expressed, i.e. a p-value less than 0.05 on a t-test of the log fold change, or not. One advantage of only considering significantly differentially expressed miRNAs is that when the difference is not significant, it is more likely that the sign of the difference is only due to chance. On the other hand, if a miRNA is significantly up-regulated in one study, but not significantly regulated in another study, this lowers the consistency

[5] `https://www.python.org`
[6] `https://numpy.org`
[7] `https://scikit-learn.org/`
[8] `https://xgboost.readthedocs.io/`
[9] `https://scipy.org`

of the differential expression. Due to the advantages and disadvantages of both options, I will report using (1) only pairs where both are significant, (2) pairs where at least one is significant and (3) all pairs, and corresponding p-values that the different probabilities are larger than 0.50 using a binomial test.

3.3.2 Signed-rank test with cross validation

Another way to find whether there is any consistency in the differential expression of miRNA is to use Wilcoxon signed-rank test (see subsection 2.2.6). This will be done by looking at the log fold change in a miRNA-sequence across different studies, and then use the signed-rank test to find whether the miRNA-sequence is significantly up- or down-regulated across studies, by looking at the signed-rank test of median differential expression of the miRNA-sequence, and whether this median is positive or negative.

3.3.2.1 Using t-test results as values for signed-rank test

As seen in Berg [2021], there is a large difference in the number of samples in the different datasets, which means that using raw log fold chance might lead to small datasets having a big impact on the signed-rank test due to chance. Therefore, I will also do an experiment where instead of using log fold change in the signed-rank test, I will use the p-value of a t-test instead. Then, datasets with more samples get a larger impact as they have more statistical power. More formally, I will do a two-sided t-test of the log fold change and use $\frac{\text{sgn(t-value)}}{\text{p-value}}$ as the value for the signed-rank test. Notice that the sign is the same as the log fold change, and that the absolute value of the fraction is inverse proportional to the p-value. As the signed-rank test only considers the rank of the value, and not the absolute value, any function decreasing by increasing p-values would work, including this.

3.3.2.2 Cross validation

Firstly, to ensure external validity of the results of the signed-rank test, I will do a test where I do a signed-rank test on all studies, except two that are exempted. Then I will look at the 10 most and 10 least consistently differentially expressed miRNA-sequences based on the signed-rank test, using only miRNAs that are in at least ten of the studies, where these 20 miRNA-sequences are also in the two exempted studies. If two exempted studies do not have at least 20 miRNA-sequences in common that are in at least ten of the other datasets, that pair of studies will not be exempted together. Otherwise, all pairs of two datasets will be tried as exempted datasets. If there is a larger consistency in the two exempted datasets in the expression of the miRNAs that had the most consistency

in the signed-rank test, that would suggest that the signed-rank test has external validity. The consistency in the two exempted datasets will be calculated similarly to subsection 3.3.1, i.e. the proportion of miRNAs that have the same direction of differential expression is compared between the 10 most and 10 least consistently differentially expressed miRNA-sequences in the signed-rank test.

3.3.2.3 Finding most consistently differentially expressed miRNAs

By using the signed-rank test on all the datasets, one can find the miRNA-sequences that are the most consistently differentially expressed in the datasets. This will both be done using log fold change and using t-test results as the value in the signed-rank test. Thereafter, I will find the 10 most consistently differentially expressed miRNAs using each of the two possible metrics for the change in miRNA expression. The p-values will be adjusted using a Bonferroni correction, to adjust for the multiple testing.

3.4 Hierarchical clustering of datasets

The clustering will be computed using `scipy.cluster.hierarchy.linkage` in SciPy with "ward" as method. The distance will be the mean of the squared difference in log fold change for each miRNA-sequence that the two datasets have in common. I.e. if x_i and y_i are the log fold changes in miRNA i in the two datasets, and there are n miRNA-sequences in common between these datasets. Then the distance is

$$\text{dist(x, y)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2$$

The results will be visualized in a dendrogram.

3.5 Machine learning on single datasets

I will train four different models on each dataset using logistic regression, SVM, random forest and XGBoost. The models will be tested using AUC, and the AUC will be calculated using cross validation where the dataset is split into $c = \min(5, \#\text{Cases}, \#\text{Controls})$ equal parts and for each of the c parts, there will be a round where the model is trained on the $c-1$ other parts of the dataset and tested on the last part. The resulting AUC will be the average over the c rounds.

3.6 Machine learning using multiple datasets

There are multiple ways to do machine learning using multiple datasets.

3.6.1 Using the most replicated miRNA-sequences from the meta-analyses

I will select the datasets that have all the miRNA-sequences that were most replicated in the meta-analyses, and train a logistic regression model using leave-one-out cross validation. One dataset is chosen as the test dataset in each iteration, whilst the model is trained on the other datasets. The samples will be weighted so that the sum of weights in each dataset is the same, and the weights of all samples in the same dataset are the same.

3.6.2 Training on two datasets

I will train different machine learning models on two datasets and try to predict on a third dataset, and then compare the results to the results that are found by training the model on only one of the datasets. The results will only be considered if the three datasets have at least 10 miRNA-sequences in common, to ensure the datasets are similar enough. The samples will be weighted so that the sum of weights in each dataset is the same, and the weights of all samples in the same dataset are the same.

3.7 Finding RPM threshold for sequencing data

More precisely, I will make cutoffs on 0, 1, 10, 100 and 1000 mean RPM (reads per million) in the sequencing datasets, where all miRNA-sequences that have a lower average read than the threshold are filtered out.

I will train on the sequencing datasets using leave-one-out cross validation, i.e. using all datasets except one test set, and take the average AUC when the different datasets are used as test datasets. This will be done for all the different thresholds. I will use two different models. Firstly, I will train a logistic regression model, using the miRNAs that the datasets have in common. Afterward, I will train an XGBoost model using the union of the miRNAs in the different datasets, and setting missing values to NaN.

3.8 Creating a web app for visualizing data

One goal of this project was to make data easily available for other researchers to explore. Therefore, would make a web application where one can visualize the data easily. The web application was made using the front-end framework

React[10] for the main application and Plotly.js for graphs[11].

3.8.1 Pairwise machine learning

The pairwise machine learning between two pairs of datasets is calculated as follows: First, the intersection of miRNA-sequences between the two datasets is computed. If the size of this intersection is smaller than four, then the pair is skipped. Otherwise, the following AUCs are calculated:

- The mean AUC when using a min(5, #cases, #controls)-fold cross validation in the first dataset, only using the miRNAs that are in common for the two datasets.
- The mean AUC when using a min(5, #cases, #controls)-fold cross validation in the second dataset, only using the miRNAs that are in common for the two datasets.
- The AUC when you train on the first dataset and test on the second dataset.
- The AUC when you train on the second dataset and test on the first dataset.

This calculation is done once for each of these four different machine learning models: logistic regression, SVM, random forest and XGBoost.

[10] `https://reactjs.org/`
[11] `https://plotly.com/javascript/`

Chapter 4

Experiments and Results

The main goal of this work was to use machine learning algorithms to identify individuals with lung cancer from their levels of circulating miRNA. This chapter describes the experiments done toward this goal. Specifically, I present the studies included in the experiments (section 4.2), investigate pairwise correlation in case-control differences between the studies (section 4.3), evaluate to what extent there is evidence that individual miRNAs are consistently differentially expressed across studies (section 4.4), find to what extent machine learning algorithms can distinguish samples from different datasets (section 4.5), find a hierarchical clustering of the datasets based on their differential expression in miRNAs between cases and controls (section 4.6), do machine learning on case-control status internally in each dataset (section 4.7), find a single miRNA-sequence to use as baseline across datasets (section 4.8), find if any principal components in the biggest datasets have any association with case-control status (section 4.9), do machine learning across different datasets (section 4.10), group datasets based on characteristics and do machine learning in each group (section 4.11), find whether the largest principal components are noise and whether their removal is beneficial (section 4.12), find whether an RPM threshold for miRNAs in sequencing datasets is beneficial (section 4.13), explore whether samples are contaminated by red blood cells (section 4.14), create a web application for visualizing the data (section 4.15) and do some further exploration based on the visualizations in the web application (section 4.16).

4.1 Code and data availability

All code and results used in this project can be found on GitHub (`https://github.com/OleFredrik1/masterthesis`). The normalized data is available at

https://doi.org/10.5281/zenodo.6568981. The code and calculation for the web application is available in the GitHub repository, and at time of publication a live demo is available at https://mirna-visualizer.netlify.app/. A preprint of an article based on this project is found in appendix A.

4.2 Studies included

Current literature is replete with studies investigating the potential of circulating miRNAs for lung cancer diagnosis, but for such studies to be useful for machine learning analyses and for replication purposes, the data from individual miRNAs and individuals should be available. To identify a large and unbiased set of studies that had investigated and reported the blood expression profiles of multiple miRNAs in multiple individuals, including both lung cancer patients and controls, I did a structured literature review (see section 3.1).

The review identified 123 studies. However, most datasets that were requested by email were not received. The 26 studies whose datasets that were either received or were publicly available are: Abdollahi et al. [2019], 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. [2016], 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]. A basic overview of the different studies is found in Table 4.1. A more detailed overview of the studies is found in Berg [2021][2].

4.3 Log fold change correlation

Berg [2021] showed that there is little log fold change correlation in the data. Furthermore, it showed that even though the correlation direction was arbitrary, there was a significant correlation between the datasets. However, the significance of the correlation was similar when randomizing the column corresponding to case-control, which means that the correlation could partially be a result of the covariance between different miRNA-sequences rather than due to case-control characteristics. The lack of correlation could be due to the characteristics of the

[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

[2] Abdollahi et al. [2019] is not described in Berg [2021] as the data was received too late for it to be included.

Table 4.1: Characteristics of the studies in this project. The columns are as follows: *Study*: The study the row is describing, *Technology*: The technology used to measure miRNA in that study, *Blood fraction*: What blood fraction was used for measuring miRNAs, *# miRNAs*: The number of different miRNA-sequences that were measured in the study, *# Cases*: The number of samples from cancer patients in the study, *# Controls*: The number of healthy controls in the study, *Total*: The total number of samples in the study. EV = Extracellular Vehichle, Ex = Exsomal

Study	**Technology**	**Blood fraction**	**# miRNAs**	**# Cases**	**# Controls**	**# Total**
Abdollahi et al. [2019]	qRT-PCR	Whole blood	4	43	43	86
Asakura et al. [2020]	Microarray	Serum	2565	1566	2178	3744
Bianchi et al. [2011]	qRT-PCR	Serum	29	95	69	164
Boeri et al. [2011]	Microarray	Plasma	131	19	6	25
Chen et al. [2019]	Sequencing	Plasma	253	30	24	54
Duan et al. [2021]	Microarray	Serum	1998	6	6	12
Fehlmann et al. [2020]	Microarray	Whole blood	689	606	2440	3046
Halvorsen et al. [2016]	qRT-PCR	Serum	254	38	16	54
Jin et al. [2017]	Sequencing	Plasma	527	26	12	38
Keller et al. [2009]	Microarray	Blood cells	386	17	19	36
Keller et al. [2014]	Microarray	P. Blood	722	73	94	167
Keller et al. [2020]	Microarray	Serum	435	10	90	100
Kryczka et al. [2021]	qRT-PCR	Serum EV	4	31	21	52
Leidinger et al. [2011]	Microarray	Whole blood	852	28	19	47
Leidinger et al. [2014]	Microarray	Whole blood	1186	42	38	80
Leidinger et al. [2016]	qRT-PCR	Whole blood	205	74	46	120
Li et al. [2017]	Microarray	Plasma	165	6	3	9
Marzi et al. [2016]	qRT-PCR	Serum	13	48	984	1032
Nigita et al. [2018]	Sequencing	Plasma Ex	102	19	7	26
Patnaik et al. [2012]	Microarray	Whole blood	1396	33	12	45
Patnaik et al. [2017]	Microarray	Whole blood	3036	86	77	163
Qu et al. [2017]	Microarray	Plasma	184	9	4	13
Reis et al. [2020]	Microarray	Plasma	795	35	7	42
Wozniak et al. [2015]	qRT-PCR	Plasma	342	100	100	200
Yao et al. [2019]	Sequencing	Plasma EV	569	5	5	10
Zaporozhchenko et al. [2018]	qRT-PCR	Plasma	175	17	10	27

different studies, like what particular technology was used for measuring and what blood fraction was measured. Therefore, a relevant experiment would be to see whether there is a larger correlation between datasets using the same technology and blood fraction, contrasted with the correlation when the datasets differ in these characteristics.

Due to the limited amount of datasets in this project, grouping based on both characteristics would give too low statistical power to make any conclusions. Therefore, I will first group by technology, and group by blood fraction afterward. I will use a t-test to compare the calculated correlations when both datasets are in the same group to the correlation when only one of the datasets is in a certain group. To avoid spurious correlations, only pairs of datasets with at least 10 miRNAs in common are considered. The correlation is calculated using Pearson's r.

The results from when checking the log fold change correlation where both datasets are in the same category (the in-group), contrasted with when only one of the datasets is in the category (the out-group), are shown in Table 4.2. There was no significant change between the in-groups and the out-groups in any of the cases if one adjust for multiple testing. This might be due to a lack of statistical power as there are few datasets in this project. That said, regardless of the significance, the mean correlation is low. One should also note that, according to Berg [2021], the correlation seemed to be due to covariance between miRNA expressions rather than due to case-control characteristics. Therefore, I will replicate the case-control randomization done in Berg [2021], but only for the in-groups.

The results from randomly assigning case-control status, and then calculating the pairwise log fold change correlation are shown in Table 4.3. As there is no significant difference in the correlations, it seems that there was no significant correlation that could be separated from the effect of covariance between the miRNAs. There is a difference between the experiment done here and the experiment in Berg [2021], namely that here I look at the direction of the correlation, while Berg primarily looked at the significance of the correlation. It might be that the case-control characteristics are the cause of the direction of the correlation, whilst covariance between the miRNA-sequences is the cause of the significance of the correlation.

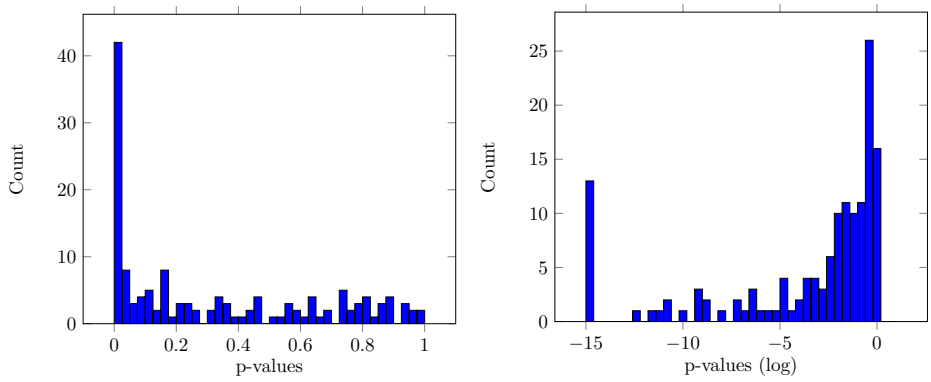
It is hard to test the last hypothesis as the p-values do not have a known distribution. They are not uniformly distributed as there is a significant correlation between the datasets. They are also far from normally distributed, due to a very strong left skew, which would make a t-test give misleading results. One possibility might be to log transform the p-values. The results are shown in Figure 4.1. Neither are normally distributed, but the distribution of the log transformed p-values seems closer to a normal distribution than the non-transformed p-values.

Table 4.2: Pearson's r of the log fold change between pairs of datasets. The first column is what group of datasets is selected. The second column is the mean log fold change correlation for pairs of datasets inside the group. The third column is the mean log fold change correlation for pairs of datasets where one of the datasets is inside the group and the other dataset is outside the group. The fourth and the fifth columns are the result of a t-test where the correlation coefficients in the in-group and the out-group were compared, with the t-statistic and the corresponding two-sided p-value.
Note: IG = in-group, OG = out-group

Group	**Mean IG**	**Mean OG**	**t-value**	**p-value**
Microarray	0.049	−0.001	1.980	0.049
Sequencing	0.086	−0.033	1.650	0.103
qRT-PCR	−0.007	0.024	−0.519	0.605
Plasma	−0.072	0.007	−1.981	0.049
Whole blood	0.116	0.039	1.453	0.149
Serum	0.048	0.013	0.651	0.516

Table 4.3: Pearson's r of the log fold change between pairs of datasets inside each group when case-control status is shuffled and not shuffled. The first column is what group of datasets is analyzed for the row. The second column is the mean log fold change correlation coefficient when the case-control characteristics are not shuffled. The third column is the mean log fold change correlation coefficient when the case-control characteristics are shuffled. The fourth and fifth columns are the result of a t-test where the correlation coefficients are compared when shuffled and when not shuffled, with the corresponding t-value and a two-sided p-value.

Group	**Mean Non-shuffled**	**Mean Shuffled**	**t-value**	**p-value**
Sequencing	0.086	−0.020	1.256	0.245
qRT-PCR	−0.007	−0.049	0.492	0.627
Microarray	0.049	0.022	1.029	0.305
Serum	0.048	−0.044	0.875	0.390
Whole blood	0.116	0.029	1.333	0.193
Plasma	−0.072	0.020	−1.582	0.119



(a) Non-transformed (b) Log transformed

Figure 4.1: p-values of the log fold change correlation between each pair of studies with the same technology or the same blood fraction

The experiment thus becomes to look at the log transformed p-values when randomizing the case-control assignment and the log transformed p-values without randomization. Then a t-test will be performed to check for a possible difference in the p-values between the randomized and the non-randomized case. The t-test showed that the correlation was more significant when the case-control assignment was not randomized ($p = 0.015$), which suggests that case-control characteristics are the cause of some of the significance in the correlations, rather than it being only due to covariance between miRNA expressions.

None of the differences in Table 4.2 were significant. It might be that some of the differences would be significant with more statistical power. One possible test is to test all correlations that are in in-groups to all correlations that are only in out-groups. This would have more statistical power, with the disadvantage that it does not say anything about which groups have more internal consistency, as it seems like it varies between the groups.

The result was that there was still no significant difference between the in-groups and the out-groups when aggregating over all categories ($p = 0.287$), which suggests that the increased correlation in in-groups is either non-existent or too small to be found with the current number of datasets. Either way, the mean correlation in the in-group was $r = 0.030$, which is a very small correlation that would suggest that case-control characteristics' effect on log fold change is either much smaller than the other effects, or the effects cannot be replicated across datasets. Indeed, Berg [2021] shows using linear regression that the proportion of variance in the miRNA expression that is due to case-control characteristics

Table 4.4: Pearson's r of the log fold change between pairs of datasets when only stages 3 and 4 are considered compared to when only stages 1 and 2 are considered, and the result from a t-test between the early and late stage log fold change correlation coefficients, with a t-value and a two-sided p-value.

Mean Late	**Mean Early**	**t-value**	**p-value**
−0.059	0.004	−0.651	0.521

varies and is generally quite small. The highest proportion is 0.446 in Duan et al. [2021] and the smallest is 0.009 in Leidinger et al. [2014]. The proportion found by linear regression is probably overstating the actual proportion due to overfitting to the data, as the proportion of explained variance was smaller in datasets with larger sample sizes [Berg, 2021].

4.3.1 Using stages

There is some evidence that suggests that the diagnostic accuracy of microRNA is somewhat higher in later stages of lung cancer [Yu et al., 2019]. Thus, it might be valuable to check whether there is a higher log fold change correlation when only using cancer samples with advanced stages, in this case, stage 3 and 4. The log fold change is in this case the difference between the mean expression in the cancer samples in advanced stages and the controls. If the correlation is better when considering later stages, that might suggest that later stages have more consistent expression, and thus are easier to diagnose. The test will be a t-test of the correlation coefficients when only stage 3 and 4 are considered, compared to the correlation coefficients when only stage 1 and 2 are considered.

The datasets where stage is marked are Abdollahi et al. [2019]; Bianchi et al. [2011]; Zaporozhchenko et al. [2018]; Duan et al. [2021]; Boeri et al. [2011]; Leidinger et al. [2011]; Qu et al. [2017]; Li et al. [2017]; Nigita et al. [2018]. However, Duan et al. [2021] only have samples in stages 1 and 2, and is therefore not used in this analysis. The results when only using advanced stages or only using early stages are shown in Table 4.4. This shows that there is no significant difference in the log fold change correlation when only considering late stages compared to when only considering early stages. The sample size is small, though, which makes it hard to conclude anything with certainty. However, the results suggest that there is no large improvement in log fold change correlation when using only late stage cancer, and if anything the correlation is higher in earlier stages. Indeed, Yu et al. [2019] suggested that the improvement in the diagnostic value of miRNA in late stage cancer was relatively small.

4.4 Evidence for consistently differentially expressed miRNA-sequences

One question that has to be considered, especially given how Berg [2021] found that sign of the log fold change correlation was virtually arbitrary, is whether there is evidence that there exists any consistently differentially expressed miRNA-sequences at all. By consistently, I mean that it is differentially expressed in the same direction (up- or down-regulated) across studies. Section 4.8 shows that some miRNA-sequences are often differentially expressed, but that the direction is not consistent, which would make diagnosis hard and lead one to question whether the differential expression was primarily due to case-control characteristics. It is difficult to rule out that there exist any consistently differentially expressed miRNA-sequences, especially as many of the miRNA-sequences are only present in a few datasets, which would mean that it would be hard to say whether the differential expression is due to chance, study characteristics or case-control characteristics.

4.4.1 Paired sign test

The calculated probabilities that a miRNA-sequence is differentially expressed in the same way in two different studies are shown in Table 4.5, where the pairs are filtered on whether the differential expression is statistically significant in the different studies. The experiment is described in more detail in subsection 3.3.1. None of the probabilities are higher than 0.50, which suggests that there is no consistency in the differential expression of single miRNAs. However, there is still a need to check if there is a confounder in the technology or blood fraction used in the different studies.

4.4.1.1 Stratification of the datasets

If the differences are due to technology or blood fraction, one might assume that the consistency is higher when only checking against datasets where these characteristics are similar. The results from an analysis checking only significant pairs where both studies share a characteristic are shown in Table 4.6. Whole blood is the only group that gives significantly better than chance levels. Still, it was only slightly better than chance level, which suggests that neither technology nor blood fraction is the cause of the poor consistency.

4.4.1.2 Possible significance levels

One issue that remains is the significance level. I sat a significance level of $p = 0.05$, but if the number of miRNAs that are differentially expressed due to

Table 4.5: The results from an experiment where one takes a pair of datasets that have measured the same miRNA. Then one checks whether the signs of the fold change are equal or not equal. The first column is what pairs are used, where significant means the log fold change was significantly different from zero using a two-sided t-test and a significance level of 0.05. One or both refers to whether the differential expression was significant in one or both datasets in the pair. The second column is the portion of the pairs that have the same sign, or if you know the sign of one of the datasets in the pair, then it is the probability that the other dataset has the same sign. Finally, the last column contains a p-value, which is the resulting p-value from a one-sided binomial test on whether the portion of pairs with the same sign is larger than 0.50.

Pairs	**Probability**	**p-value**
All pairs	0.495	0.992
One significant	0.497	0.911
Both significant	0.494	0.862

Table 4.6: The results from an experiment where one takes pairs of datasets that share a certain characteristic and that have measured the same miRNA. Then one checks whether the signs of the fold change are equal or not equal, only using pairs where the log fold change was significantly different from zero, using a two-sided t-test. The first column is what characteristic the datasets in that row share. The second column is the portion of the pairs that have the same sign, or if you know the sign of one of the datasets in the pair, then it is the probability that the other dataset has the same sign. Finally, the last column contains a p-value, which is the resulting p-value from a one-sided binomial test on whether the portion of pairs with the same sign is larger than 0.50.

Group	**Probability**	**p-value**
Microarray	0.495	0.912
Sequencing	0.521	0.357
qRT-PCR	0.527	0.164
Whole blood	0.528	0.009
Serum	0.292	1.000
Plasma	0.442	1.000

Table 4.7: The results from an experiment where one takes a pair of datasets that have measured the same miRNA. Then one checks whether the signs of the fold change are equal or not equal, when there is significant differential expression in both datasets using a two-sided t-test. The first column is the significance level that is used for the t-test. The second column is the portion of the pairs that have the same sign, or if you know the sign of one of the datasets in the pair, then it is the probability that the other dataset has the same sign. Finally, the last column contains a p-value, which is the resulting p-value from a one-sided binomial test on whether the portion of pairs with the same sign is larger than 0.50.

Significance level	**Probability**	**p-value**
5×10^{-2}	0.494	0.862
5×10^{-3}	0.479	0.987
5×10^{-4}	0.488	0.796
5×10^{-5}	0.528	0.0855
5×10^{-6}	0.519	0.223
5×10^{-7}	0.529	0.161
5×10^{-8}	0.534	0.157
5×10^{-9}	0.543	0.112
5×10^{-10}	0.534	0.205
5×10^{-11}	0.516	0.376

case-control characteristics is low, this will lead to a large portion of false positives among the miRNAs found to be significantly differentially expressed. Due to that, I tried with different significance levels, and the results are shown in Table 4.7. There seems to be a general upward trend where a more stringent significance level results in a higher consistency. However, none of the significance levels result in a consistency significantly better than chance levels, and the highest probability is still only very slightly better than chance. Thus, the lack of consistency was not due to a high significance level or similarly a high share of false positives.

4.4.2 Signed-rank test with cross validation

Here are the results from the signed-rank test including when cross validation was used, which is another method to check for consistent differential expression in miRNA.

Table 4.8: The proportion of pairs that had the same direction of differential expression in the two excluded datasets, among the miRNAs that were shown to be most and least consistently differentially expressed in the signed-rank test as described in subsubsection 3.3.2.2. The t-value is the t-value for the difference between the two proportions, and the p-value is the corresponding p-value.

Most significant	**Least significant**	**t-value**	**p-value**
0.510	0.479	2.12	0.0337

4.4.2.1 Cross validation

This is an experiment where a signed-rank test is used to find the most and least consistently differentially expressed miRNAs in a set of datasets, and then one uses two external datasets to see whether the consistency generalizes. The experiment is described in more detail in subsubsection 3.3.2.2. The results are shown in Table 4.8. The results suggest that the miRNA-sequences that were the most consistently differentially expressed in the signed-rank test were somewhat more consistently differentially expressed in the two excluded datasets. The difference is significant at a 0.05-level, but there have been done many statistical tests in this chapter, and adjusted for this multiple testing it is not significant. What causes this poor consistency? One hint may lay in subsection 4.4.1, which suggests that the consistency in might be better if only looking at pairs where both are significantly differentially expressed.

Only pairs with significantly differentially expressed miRNAs
Now I will only consider pairs where the miRNA is significantly differentially expressed in the two excluded datasets using a t-test and a significance level of $p = 0.05$. The results are shown Table 4.9. The results show that there is no significant difference in the proportion of the pairs with the same direction of differential expression. This suggests that the lack of improvement in the proportion in subsubsection 4.4.2.1 was not the result of insignificance in the differential expression in the pairs.

Using t-test results instead of log fold change in signed-rank test
One possible reason for the results in subsubsection 4.4.2.1 could be that small studies have a big log fold change due to chance. Therefore, I will also try using the p-value of a t-test in the signed-rank test as explained in subsubsection 3.3.2.1. The results are shown in Table 4.10. Neither here were there any signs of external validity in the results from the signed-rank test. Thus, either the signed-rank

Table 4.9: The proportion of pairs that had the same direction of differential expression in the two excluded datasets, among the miRNAs that were shown to be most and least consistently differentially expressed in the signed-rank test as described in subsubsection 3.3.2.2. The miRNA-sequence had to be significantly differentially expressed in the two excluded datasets in a t-test. The t-value in the table is the t-value for the difference between the two proportions, and the p-value is the corresponding p-value.

Most significant	**Least significant**	**t-value**	**p-value**
0.502	0.534	−0.640	0.522

Table 4.10: The proportion of pairs that had the same direction of differential expression in the two excluded datasets, among the miRNAs that were shown to be most and least consistently differentially expressed in the signed-rank test as described in subsubsection 3.3.2.2, with the difference that the p-value of a t-test of the log fold change was used instead of the log fold change. The t-value in the table is the t-value for the difference between the two proportions, and the p-value is the corresponding p-value.

Most significant	**Least significant**	**t-value**	**p-value**
0.486	0.467	1.24	0.215

test is a subpar way to find what miRNAs are the most consistently differentially expressed, or there are no consistently differentially expressed miRNAs.

4.4.2.2 Signed-rank test

A signed-rank test was done using all datasets. The signed-rank test was done both using log fold change and t-test results. Using t-test results is explained in subsubsection 3.3.2.1, and this experiment is explained in more detail in subsubsection 3.3.2.3. The results are in Table 4.11. As subsubsection 4.4.2.1 suggests that this test has low external validity, one should be careful when using this data for conclusions. None of the miRNA-sequences were significantly consistently differentially expressed in the datasets when adjusted for multiple testing, which could either be because there is no consistently differentially expressed miRNA-sequence, or there are too few datasets included in this study to get enough statistical power to find any.

Table 4.11: The most consistently differentially expressed miRNA-sequences according to a signed-rank test as described in subsubsection 3.3.2.3. p-values are adjusted using Bonferroni correction. The direction is whether the miRNA is up- or down-regulated in cancer. The headers of the subtables tell whether it is the log fold change or the results from t-tests that is the input to the signed-rank test.

(a) Using log fold change

MiRNA	p-value	Direction
miR-663a	0.891	Up
miR-17	1.06	Down
miR-625-3p	2.78	Up
miR-93	4.09	Down
miR-374a-5p	4.09	Down
miR-106b	5.51	Down
miR-20	6.32	Down
miR-202	7.57	Up
miR-106a-5p	8.24	Down
let-7d	8.24	Down

(b) Using t-test results

MiRNA	p-value	Direction
miR-663a	1.34	Up
miR-625-3p	3.67	Up
miR-425-3p	4.68	Up
miR-1224-5p	4.79	Up
miR-296-5p	5.68	Up
miR-211	6.23	Up
miR-17	6.98	Down
miR-205-5p	7.79	Up
miR-518f-3p	8.91	Up
miR-483	10.6	Up

4.5 Are datasets separable from each other?

One question arises when the consistency between the datasets is as poor as it has been shown to be in Berg [2021], namely can one recognize what dataset a sample is from? Given that there are differences between the datasets, can one use these differences to recognize a dataset? The way I will test this is using logistic regression on a pair of datasets where one-third of each dataset is used for testing and two-thirds is used for training. The model will be trained to separate samples from the two datasets from each other. Only pairs of datasets with at least 10 miRNA-sequences in common are considered. The metric to evaluate the separation is AUC.

Learning a logistic regression model to separate samples from two datasets lead to a mean AUC of 0.229 and a standard deviation of 0.210. The results were very poor, and somewhat suspicious, as a random separation would give an AUC of 0.50. On the other hand, an AUC of 0.229 is good in some sense, because it separates well, but it mixes up the two categories. I do not know what caused this result, but there was strong evidence that the datasets could be separated, and therefore I tried using XGBoost instead. XGBoost gave a mean AUC of 0.934 with a standard deviation of 0.121. This suggests that the datasets can be separated from each other, but far from perfectly in general. However, looking at the AUC values, several had AUC > 0.99, which means that there were

datasets that were easier to distinguish than others. One question that remains is whether this knowledge can be used to adjust the datasets so that they are more comparable. One possibility would be to use linear regression to find expression patterns that are characteristic of that dataset and then adjust for it. However, this would not work as the miRNA expressions are already standardized to a mean of 0, so the linear regression would not find any mean effects of any dataset. Furthermore, logistic regression performed poorly compared to XGBoost when trying to separate datasets, which suggests that what distinguishes the datasets are non-linear patterns, which are hard to adjust for.

4.6 Hierarchical clustering of datasets

Hierarchical clustering of miRNA expressions is a common analysis in this field. As I want to find patterns in the comparability of the datasets, I will try to cluster the datasets. This would not only give information about what datasets are more comparable, but also whether there are clusters of datasets that are closer to each other, and in that case, what characterizes them. This is somewhat similar to the analysis in Berg [2021] where Berg created a graph of what datasets were similar. The difference is that the hierarchical clustering will give clusters of datasets that are similar to each other, rather than just comparing pairs of datasets.

The results from doing hierarchical clustering of the datasets are shown in Figure 4.2. The clustering was based on the difference in log fold change as described in section 3.4. There seem to be a close cluster that includes Wozniak et al. [2015]; Leidinger et al. [2014]; Patnaik et al. [2017]; Leidinger et al. [2011]; Keller et al. [2014]. It might be that these datasets are close to each other, and that models trained on one of the datasets would do well on the other datasets. Testing this hypothesis by using pairs of datasets where one train on one dataset and test on the other using logistic regression resulted in a mean AUC of 0.520 with a standard deviation of 0.144. Trying to test on one dataset while training on the others using XGBoost resulted in AUCs with a mean of 0.594 and a standard deviation of 0.171, which was not significantly higher than 0.500, plausibly due to the low sample size. Overall, even though this is a cluster, the diagnostic value across these datasets is relatively low, at least compared to the internal diagnostic value found in section 3.5.

4.7 Machine learning on single datasets

One question that was asked in section 1.2 was whether more advanced machine learning algorithms are better at diagnosing lung cancer based on miRNA-levels. Therefore, I have chosen to do machine learning on single datasets. As seen

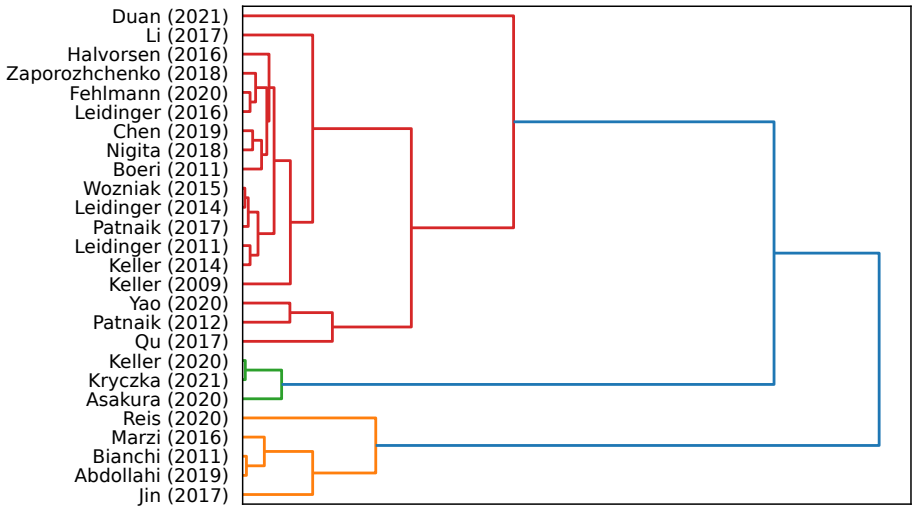


Figure 4.2: Hierarchical clustering of the datasets, where the distance between the datasets is equal to the mean squared difference in log fold change. Coloring is for aesthetic reasons.

in Berg [2021], the results when doing machine learning across datasets were mostly poor. Furthermore, if the connection between miRNA expression and lung cancer is very sensitive to study characteristics, machine learning across arbitrary different datasets might not be the best idea, compared to training on data where the characteristics are known to be the same. I have chosen four different types of machine learning algorithms to test.

- Logistic regression: It is natural to use as a baseline model to compare against, as it has been used in many of the studies that are included in this project.
- Support vector machine: If the data are nearly linearly separable (which the PCA-plots in Berg [2021] suggest is often the case), this will find such a separation.
- Random forest: It is a powerful algorithm that is able to generalize well also on small datasets, as it is an ensemble method.
- XGBoost: Has had the most success in tabled data with limited samples in Kaggle competitions (see subsection 2.2.12), and is thus a reasonable algorithm to test.

The results from machine learning on single datasets are shown in Table 4.12. The results are from using cross validation on the datasets with the given machine learning algorithms. A more detailed explanation is found in section 3.5. Random forest performed best, while XGBoost performed worst. One question is whether these differences are statistically significant or not. Therefore, I performed t-tests on the differences in AUC values. The results are in Table 4.13, which shows that none of the differences were significant. Thus one cannot say that one algorithm performed better than another in general.

4.7.1 Using stages

Another question is whether there is any difference in AUC between early and late stage cancer. Therefore, I will train and test a logistic regression classifier on datasets where stage is labeled, using either only late stage cancer or only early stage cancer. The training and testing will follow the same cross validation strategy as in the experiment above. The resulting AUCs when training and testing a logistic regression model on only late stage cancer or only on early stage cancer are shown in Table 4.14. The results might suggest that there might be a higher AUC when only using late stage cancer, however, the statistical power here is too small to conclude with any certainty.

4.8 Baseline miRNA-sequence

One question, as asked in section 1.2, was what miRNA-sequences would be most successful in diagnosing lung cancer. This has not only clinical relevance, but is also important to have as a comparison for a machine learning model, as a complicated machine learning model would be worthless if one single miRNA-sequence had the same diagnostic value. In addition, there will always be additional costs associated with measuring more miRNA-sequences in a blood test, therefore ideally one would like to find the simplest model possible.

There are two main types of methods possible for finding such miRNA-sequences, each with some pros and cons:

1. Look at meta-analyses for finding miRNA-sequences that are found to diagnose lung cancer well across studies.

   Pros:

   - The miRNA-sequences found would be based on more data, and thus they are likely better.

Table 4.12: The mean AUC when using cross validation on the given studies with the given models, as described in section 3.5. The first column says which dataset is used, and the rest of the columns have a column name that represents the model used. LR = Logistic Regression, RF = Random Forest

Study	LR	SVM	RF	XGBoost
Abdollahi et al. [2019]	0.670	0.814	0.933	0.853
Asakura et al. [2020]	0.734	0.913	0.968	0.939
Bianchi et al. [2011]	0.795	0.852	0.823	0.843
Boeri et al. [2011]	0.783	0.950	0.950	0.575
Chen et al. [2019]	0.882	0.787	0.793	0.623
Duan et al. [2021]	0.900	0.900	0.900	0.800
Fehlmann et al. [2020]	0.977	0.980	0.960	0.980
Halvorsen et al. [2016]	0.985	0.993	0.983	0.933
Jin et al. [2017]	1.000	1.000	1.000	0.980
Keller et al. [2009]	0.950	0.900	0.931	0.800
Keller et al. [2014]	0.847	0.888	0.864	0.815
Keller et al. [2020]	0.956	0.944	0.925	0.967
Kryczka et al. [2021]	0.749	0.658	0.646	0.606
Leidinger et al. [2011]	0.162	0.752	0.705	0.411
Leidinger et al. [2014]	0.160	0.286	0.365	0.528
Leidinger et al. [2016]	0.916	0.903	0.948	0.936
Li et al. [2017]	0.333	0.167	0.833	0.500
Marzi et al. [2016]	0.976	0.969	0.950	0.968
Nigita et al. [2018]	0.700	0.300	0.183	0.233
Patnaik et al. [2012]	0.698	0.883	0.777	0.763
Patnaik et al. [2017]	0.573	0.481	0.476	0.543
Qu et al. [2017]	1.000	1.000	1.000	0.479
Reis et al. [2020]	1.000	1.000	0.957	0.943
Wozniak et al. [2015]	0.494	0.565	0.663	0.689
Yao et al. [2019]	0.800	0.600	1.000	0.400
Zaporozhchenko et al. [2018]	0.242	0.800	0.842	0.642
Mean	0.742	0.780	0.822	0.721

Table 4.13: The p-values of the difference in mean AUC in Table 4.12 using a two-sided t-test. The row and column labels represent what algorithms we are comparing. LR = Logistic Regression, RF = Random Forest

	LR	**SVM**	**RF**	**XGBoost**
LR		0.585	0.227	0.761
SVM	0.585		0.503	0.357
RF	0.227	0.503		0.092
XGBoost	0.761	0.357	0.092	

Table 4.14: The mean AUC when using cross validation on the given studies when only using early stage cancer samples or only using late stage cancer samples. Empty fields mean that there were less than two cancer samples in the dataset, thus any inference would be impossible.

Study	**Mean early**	**Mean late**
Abdollahi et al. [2019]	0.586	0.726
Bianchi et al. [2011]	0.735	0.634
Zaporozhchenko et al. [2018]	0.500	0.250
Duan et al. [2021]	1.000	*
Boeri et al. [2011]	0.850	1.000
Leidinger et al. [2011]	0.083	0.525
Qu et al. [2017]	*	1.000
Li et al. [2017]	*	0.333
Nigita et al. [2018]	0.350	0.600
Mean	0.586	0.634

- The miRNA-sequences are nearly[3] independent of the datasets used in this project, and are therefore mostly unbiased.

Cons:

- The miRNA-sequences that are reported in these meta-analyses are often not in many of the datasets used in this project.

2. Look at the datasets used in this project to find miRNA-sequences that can separate cases and controls across the different datasets.

   Pros:

   - It is easier to limit the search to miRNA-sequences that are found in many of the datasets in this project.

   Cons:

   - The miRNA-sequences are biased, as the baseline is the ability of these miRNA-sequences to diagnose cancer on the datasets, but the miRNA-sequences were chosen because they diagnosed well on said datasets.

I want to try a hybrid strategy, in order to mitigate the cons of each method. That is, I want to try to find an intersection between microRNA-sequences that have been found in meta-analyses to be consistently good at diagnosing lung cancer and microRNA-sequences that separate well in the studies used in this project.

There are several ways to measure to which degree a miRNA-sequence can be used to separate cases from controls. One possibility would be to use the t-statistic. The advantage of the t-statistic is that it has a known distribution (given the null hypothesis), and thus one could get to know whether a difference is plausibly a result of chance or not. The disadvantage of the t-statistic is that it does not only measure to which degree the miRNA-sequence separates well in the dataset, but also the statistical power of each dataset. Therefore, large datasets would be given more weight, and the value could hide to what degree the miRNA-sequence diagnoses correctly in the dataset.

Another alternative is to use Cohen's d. The advantage of Cohen's d is that it tells to what degree cases and controls are separated independently of the number of samples in the dataset. The disadvantage of Cohen's d is that it does not consider the statistical power at all, and thus one might expect a number of spurious results when using Cohen's d. A final statistic is to use AUC. The advantages and disadvantages are similar to Cohen's d, with the difference that

[3] After all, the meta-analyses might be based on some datasets used in this project.

AUC has the advantage that it is the metric that the results will be measured against in the end. However, Cohen's d has the advantage that it also looks at the size of the difference in expression, and not just whether there is a separation like AUC does.

After consideration of the different statistics, I found that Cohen's d and AUC would be the most appropriate statistics for this purpose, as the t-statistic would give too much power to the large datasets (which might not be representative at all) and not tell the actual degree of separation.

4.8.1 Meta-analyses

Meta-analyses gave an overview of the possible miRNA-sequences that can be used as baselines in this project [Zhong et al., 2021; Huang et al., 2021; Jiang et al., 2018; Yi et al., 2021], whereas Zhong et al. [2021] was the most thorough of the meta-analyses. These meta-analyses suggest that the miRNA-sequences that have been shown to be able to diagnose lung cancer in the most studies are miR-21 and miR-210, with Zhong et al. [2021] suggesting that miR-182, miR-155 and miR-17 are in third, fourth and fifth place, respectively. All of these miRNA-sequences were reported to be up-regulated in cases compared to controls. However, these results were not representative of the studies used in this project.

Zhong et al. [2021] found that of all studies that they went through miR-21 was significantly up-regulated in cases in 48 studies, and down-regulated in two studies. However, among the studies used in this project, Patnaik et al. [2012]; Jin et al. [2017]; Leidinger et al. [2016]; Fehlmann et al. [2020] all reported that miR-21 was down-regulated in cases compared to controls, which suggests that miR-21 might not be as good of a biomarker for lung cancer as the meta-analyses suggest. An overview of the reported up- and down-regulation of the aforementioned miRNA-sequences in the studies in this project is shown in Table 4.15.

Table 4.15 shows that none of the five miRNA-sequences were consistently up-regulated. However, miR-17 was consistently down-regulated in the sample, which contrasts with Zhong et al. [2021] which reported that miR-17 had been up-regulated in 7 studies and down-regulated in one study, if one only looks at the studies using circulating miRNA. Everything considered, this points to very inconsistent results across datasets, which suggests that there might be little consistency, and hard to replicate results. Indeed, Berg [2021] found little consistency in the datasets that are considered in this study.

4.8.2 Using datasets

The meta-analyses gave some candidate miRNA-sequences that can be used as baselines in this project, namely miR-21, miR-210, miR-182, miR-155 and miR-

Table 4.15: Whether the miRNA-sequences were reported to be significantly up- or down-regulated ($p < 0.05$) in the studies.
Note: Reis et al. [2020] only report miR-210 and miR-182 to be up-regulated in adenocarcinoma. In Wozniak et al. [2015] and Keller et al. [2014] abu-miR-155 was measured instead of hsa-miR-155.

Study	miR-21	miR-210	miR-182	miR-155	miR-17
Abdollahi et al. [2019]	Up				
Asakura et al. [2020]					
Bianchi et al. [2011]					Down
Boeri et al. [2011]	Up	Up			
Chen et al. [2019]					
Duan et al. [2021]					
Fehlmann et al. [2020]	Down	Up	Down		Down
Halvorsen et al. [2016]		Down			
Jin et al. [2017]	Down			Down	
Keller et al. [2009]		Up	Up		Down
Keller et al. [2014]				Down	Down
Keller et al. [2020]					
Kryczka et al. [2021]					
Leidinger et al. [2011]					Down
Leidinger et al. [2014]	Up				
Leidinger et al. [2016]	Down				Down
Li et al. [2017]					
Marzi et al. [2016]					
Nigita et al. [2018]					
Patnaik et al. [2012]	Down	Down			Down
Patnaik et al. [2017]					
Qu et al. [2017]					
Reis et al. [2020]		Up	Up	Up	
Wozniak et al. [2015]			Down	Up	
Yao et al. [2019]					
Zaporozhchenko et al. [2018]		Up			

Table 4.16: Cohen's d of the different miRNAs in the different datasets. Difference in miRNA expression: case - controls

Study	miR-21	miR-210	miR-182	miR-155	miR-17
Abdollahi et al. [2019]	-0.784				
Asakura et al. [2020]	0.496	0.719	0.427	0.592	0.690
Bianchi et al. [2011]					-0.811
Boeri et al. [2011]	0.300			0.004	0.158
Chen et al. [2019]	0.165		0.610		0.147
Duan et al. [2021]	-1.723	-0.976	-3.535	-1.631	-0.816
Fehlmann et al. [2020]	-0.453	0.002	-0.290	-0.054	-0.542
Halvorsen et al. [2016]	-0.311	0.499		0.007	-1.410
Jin et al. [2017]	-0.128		-0.132	-1.536	-0.385
Keller et al. [2009]	0.321	1.499	0.617		-0.678
Keller et al. [2014]	0.067	0.208	-0.008		-1.303
Keller et al. [2020]	-0.097				
Kryczka et al. [2021]					
Leidinger et al. [2011]	0.165	-0.102	0.221		-0.663
Leidinger et al. [2014]	0.140	-0.006	-0.035	-0.033	0.131
Leidinger et al. [2016]	-0.804		-0.442		-0.756
Li et al. [2017]		-0.318			-0.242
Marzi et al. [2016]					
Nigita et al. [2018]	-0.215	-0.341			-0.309
Patnaik et al. [2012]	-1.009	-0.836			-1.044
Patnaik et al. [2017]	-0.044	-0.070	0.217	0.254	-0.211
Qu et al. [2017]	-1.183				-0.954
Reis et al. [2020]	-0.374	1.436	1.265		
Wozniak et al. [2015]	0.221	0.013	-0.357		0.429
Yao et al. [2019]					
Zaporozhchenko et al. [2018]	-0.404	-0.097	-0.456	-0.042	-0.503
Average	-0.269	0.109	-0.135	-0.271	-0.454

17. The Cohen's d and AUC of the miRNA-sequences in the different datasets are shown in Table 4.16 and Table 4.17 respectively.

Interestingly, the average Cohen's d of four of the miRNA-sequences was negative, even though Zhong et al. [2021] found that they were consistently up-regulated in cancer compared to healthy controls, which again suggests that these miRNA-sequences are not as good biomarkers for cancer as Zhong et al. [2021] suggest. Overall miR-210 was the only one that the datasets and the meta-analyses agree on being up-regulated, which is why I chose that miRNA as my baseline.

Table 4.17: AUC when using the expression of the different miRNAs to diagnose lung cancer in the different datasets

Study	miR-21	miR-210	miR-182	miR-155	miR-17
Abdollahi et al. [2019]	0.345				
Asakura et al. [2020]	0.630	0.742	0.601	0.660	0.690
Bianchi et al. [2011]					0.256
Boeri et al. [2011]	0.579			0.579	0.588
Chen et al. [2019]	0.506		0.608		0.443
Duan et al. [2021]	0.083	0.389	0.083	0.111	0.417
Fehlmann et al. [2020]	0.359	0.488	0.400	0.472	0.343
Halvorsen et al. [2016]	0.332	0.891		0.641	0.112
Jin et al. [2017]	0.417		0.554	0.141	0.399
Keller et al. [2009]	0.418	0.814	0.672		0.296
Keller et al. [2014]	0.509	0.554	0.499		0.167
Keller et al. [2020]	0.461				
Kryczka et al. [2021]					
Leidinger et al. [2011]	0.564	0.481	0.598		0.323
Leidinger et al. [2014]	0.548	0.501	0.518	0.479	0.541
Leidinger et al. [2016]	0.218		0.334		0.253
Li et al. [2017]		0.444			0.333
Marzi et al. [2016]					
Nigita et al. [2018]	0.425	0.421			0.421
Patnaik et al. [2012]	0.217	0.260			0.230
Patnaik et al. [2017]	0.471	0.477	0.541	0.576	0.449
Qu et al. [2017]	0.194				0.250
Reis et al. [2020]	0.441	0.910	0.918		
Wozniak et al. [2015]	0.541	0.533	0.421		0.639
Yao et al. [2019]					
Zaporozhchenko et al. [2018]	0.388	0.359	0.256	0.324	0.224
Average	0.412	0.551	0.500	0.443	0.369

4.9 PCA analysis across datasets

PCA-plots of all the datasets are found in Berg [2021]. However, a PCA analysis with multiple datasets is still to be done. Doing a PCA-analysis, there might be several principal components that separate cases and controls well, but they might not be possible to replicate across datasets, as there are study-specific reasons that the principal components separate well. For exploration purposes and to get good statistical power, I will first do a study where I only look at Asakura et al. [2020] and Fehlmann et al. [2020], as they have the most samples. I will calculate the ten largest principal components in each dataset, using the miRNA-sequences that are in both datasets. Then I will project every sample along the ten principal components, and using a t-test I will find whether there is a significant difference between cases and controls along the principal components for each dataset.

The results when finding the 10 largest principal components in Asakura et al. [2020] and projecting Asakura et al. [2020] and Fehlmann et al. [2020] onto the principal components can be found in Table 4.18. The t-values and the p-values are from the t-test of projecting cases and controls along the principal component. Similarly, the results when finding the 10 largest principal components in Fehlmann et al. [2020] are found in Table 4.19. Many of the components with significant separation in one dataset also have good separation in the other dataset. Interestingly, the components sometimes separate well, but in different directions in the two datasets. One candidate principal component as a separator is the third principal component in Asakura et al. [2020] which separates well in both datasets, and it separates in the right direction in both datasets. What remains is to test this principal component in other datasets to see whether it separates well beyond Asakura et al. [2020] and Fehlmann et al. [2020].

Unfortunately, few datasets have all the miRNAs that are in the chosen principal component. Therefore, I will use only datasets that have at least half of the miRNAs in the principal component, and replace missing values with 0. The results are shown in Table 4.20. Among the three other datasets, the t-test was only significant in Duan et al. [2021], and the sign differs among the three. One might ask why this component separates well in Asakura et al. [2020], Fehlmann et al. [2020] and Duan et al. [2021], but not in any of the two other datasets? Of course, there is selection bias at play, but it is still noticeable that the component seemingly represents something that separates cases and controls well in three datasets, but not in the two others. The fact that not all miRNAs are represented might be one factor, but as there is generally low consistency between the datasets, one should be cautious about attributing all the lack of consistency along this principal component to the lack of overlap in miRNAs.

Table 4.18: The results from a PCA analysis looking at the 10 largest principal components in Asakura et al. [2020]. The t-values are the result of a t-test along the given principal component in the two datasets.
Note: A = Asakura et al. [2020], F = Fehlmann et al. [2020], PVE = proportion of variance explained (i.e. the proportion of variance in Asakura et al. [2020] that is explained by the principal component)

#	**PVE**	**t-value A**	**p-value A**	**t-value F**	**p-value F**
1	0.292	51.894	0.0	0.381	7.03×10^{-1}
2	0.043	11.562	2.1×10^{-30}	-2.647	8.17×10^{-3}
3	0.031	-9.670	7.26×10^{-22}	-14.707	2.36×10^{-47}
4	0.030	-1.905	5.68×10^{-2}	10.430	4.74×10^{-25}
5	0.024	-9.919	6.61×10^{-23}	-10.136	9.10×10^{-24}
6	0.019	-5.869	4.76×10^{-9}	-4.925	8.88×10^{-7}
7	0.015	-3.955	7.80×10^{-5}	-0.909	3.63×10^{-1}
8	0.013	-8.427	4.99×10^{-17}	1.111	2.67×10^{-1}
9	0.011	3.554	3.84×10^{-4}	-7.662	2.44×10^{-14}
10	0.010	4.115	3.95×10^{-5}	12.563	2.51×10^{-35}

Table 4.19: The results from a PCA analysis looking at the 10 largest principal components in Fehlmann et al. [2020]. The t-values are the result of a t-test along the given principal component in the two datasets
Note: A = Asakura et al. [2020], F = Fehlmann et al. [2020], PVE = proportion of variance explained (i.e. the proportion of variance in Fehlmann et al. [2020] that is explained by the principal component)

#	**PVE**	**t-value F**	**p-value F**	**t-value A**	**p-value A**
1	0.612	-42.075	3.98×10^{-317}	-2.453	1.42×10^{-2}
2	0.107	-30.224	9.56×10^{-180}	12.344	3.42×10^{-34}
3	0.063	16.724	1.32×10^{-60}	1.626	1.04×10^{-1}
4	0.043	40.553	2.34×10^{-298}	2.835	4.61×10^{-3}
5	0.025	-7.447	1.17×10^{-13}	-14.315	4.95×10^{-45}
6	0.017	17.219	5.24×10^{-64}	8.515	2.58×10^{-17}
7	0.012	-13.395	5.33×10^{-40}	3.394	6.99×10^{-4}
8	0.011	-14.752	6.49×10^{-48}	3.969	7.37×10^{-5}
9	0.009	-7.459	1.08×10^{-13}	-9.434	7.57×10^{-21}
10	0.007	-0.527	5.98×10^{-1}	-5.442	5.68×10^{-8}

Table 4.20: The results from t-tests when projecting cases and controls along the third largest principal component in Asakura et al. [2020]. The "proportion miRNA" is the proportion of miRNA-sequences in the principal component that was in the dataset.

Study	t-value	p-value	Proportion miRNA
Asakura et al. [2020]	−9.670	7.26×10^{-22}	1.000
Duan et al. [2021]	−3.248	8.74×10^{-3}	0.740
Fehlmann et al. [2020]	−14.707	2.36×10^{-47}	1.000
Leidinger et al. [2014]	0.438	0.663	0.560
Patnaik et al. [2017]	0.103	0.918	0.752

4.10 Machine learning based on several datasets

One interesting question is whether combining multiple datasets will result in better diagnostic accuracy than using a single dataset. The result of training on one dataset and predicting on another dataset was done in Berg [2021] with subpar results. However, it is possible that training on multiple datasets will help the machine learning algorithm to find case-control patterns that transcend the patterns that are found internally in one dataset, leading to better generalizability.

4.10.1 Using the most replicated miRNA-sequences from the meta-analyses

One option is to take the miRNA-sequences that were found in the meta-analyses to be the best biomarkers for lung cancer across studies, and then train a model using these miRNA-sequences. The most replicated miRNA-sequences from the meta-analyses were miR-21, miR-210, miR-182, miR-155 and miR-17 (see subsection 4.8.1). Furthermore, the datasets that have all these miRNA-sequences are Asakura et al. [2020]; Fehlmann et al. [2020]; Leidinger et al. [2014]; Patnaik et al. [2017]; Yao et al. [2019], which means that they are the studies that were used in the leave-one-out cross validation using logistic regression. More details are in subsection 3.6.1. The resulting AUC values are in Table 4.21. Seemingly, the results are poor except when using Yao et al. [2019] as the test set. It should be noted, however, that Yao et al. [2019] have only 10 samples, which means that one should be careful about concluding based on that AUC value, especially as it is an outlier.

Table 4.21: AUC when training a logistic regression model on all the datasets in this table except the test set, using only the most replicated miRNA-sequences (miR-21, miR-210, miR-182, miR-155 and miR-17) as described in subsection 3.6.1

Test set	AUC
Asakura et al. [2020]	0.502
Fehlmann et al. [2020]	0.468
Leidinger et al. [2014]	0.509
Patnaik et al. [2017]	0.471
Yao et al. [2019]	0.720

4.10.2 Training on two datasets

The results from training on one or two datasets and testing on a third. More details are in subsection 3.6.2.

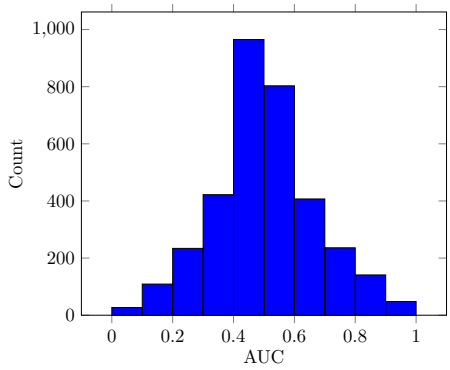
4.10.2.1 Logistic Regression

The first model I will try is logistic regression as it is a basic classification model. It is often used in the studies that try to predict cancer based on miRNA, and it therefore serves well as a baseline. The model will be trained on the miRNA-sequences that all the three datasets have in common.

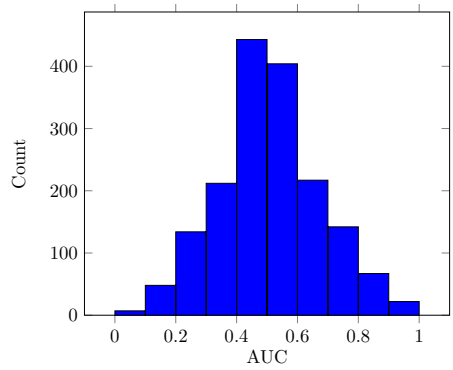
The AUC values from training on one of the datasets and testing on the third dataset using logistic regression are shown in Figure 4.3a. The AUC values from training on two datasets and testing on a third are shown in Figure 4.3b. The histograms are very similar, and this can be confirmed by other statistical measures. When training on just one of the datasets the mean AUC was 0.501 and the standard deviation was 0.168. When training on both datasets, the mean AUC was 0.508 and the standard deviation was 0.169. This is worse than the baseline miR-210, which had a mean AUC of 0.551 (see Table 4.17).

4.10.2.2 XGBoost

It is plausible that a model like XGBoost will perform better on the datasets, as it has methods for handling missing data, and it can handle non-linear relationships in the data. In addition, it is a boosting algorithm, which usually performs well when data is sparse, as in this case. Here, I will make use of the way XGBoost handles missing data and therefore train the model on all the miRNA-sequences that the two training datasets have in common.



(a) Histogram over AUC values when training a logistic regression model on one dataset and test on another according to subsection 3.6.2



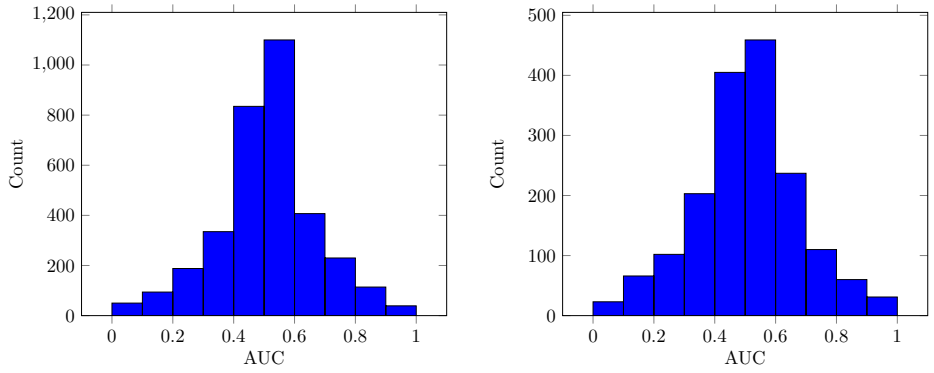
(b) Histogram over AUC values when training a logistic regression model on two datasets and test on a third dataset according to subsection 3.6.2

Figure 4.3: AUC values when training on one or two datasets using logistic regression

The AUC values from training on one of the datasets and testing on the third dataset using XGBoost are shown in Figure 4.4a. The AUC values from training on two datasets and testing on a third are shown in Figure 4.4b. The mean and standard deviation in AUC values when training on one dataset were 0.504 and 0.162 respectively. The mean and standard deviation when training on two datasets were 0.505 and 0.172 respectively. Again both the histograms and statistics are similar in the two cases, which suggests that combining two datasets have little to no effect. Furthermore, the results were very similar to the ones achieved with logistic regression, which suggests that the problem is not the model.

4.10.3 Merging all datasets

No miRNA-sequence is in all datasets. Thus, if one tries to merge all the datasets, there would not be any miRNA-sequences in the intersection of miRNAs. This would be a problem with logistic regression, but XGBoost has a way of handling missing data, which means that it can handle this. However, merging all datasets would not be useful, as one wants to see whether one can train on some collection of datasets and test on another dataset to check external validity. Therefore, I will have a strategy similar to subsection 3.6.1, where I leave one dataset out which is used as a test set. Here, every sample will have the same weight, as



(a) Histogram over AUC values when training on one dataset and test on another according to subsection 3.6.2

(b) Histogram over AUC values when training on two datasets and test on a third dataset according to subsection 3.6.2

Figure 4.4: AUC values when training on one or two datasets using XGBoost

there is not any problem with having one dataset dominating as there are many datasets in the training set.

The AUC values from merging all datasets except one which is used for testing are shown in Table 4.22. The mean of the AUC values is 0.518 and the standard deviation of the AUC values is 0.205, which is lower than the AUC of miR-210 (0.551), which means that this was a poor way to learn a model to diagnose lung cancer.

4.10.4 Maximal training sets

Unfortunately, trying to train a model on all possible subsets of datasets is computationally infeasible. There are 26 datasets in this project, leading to $2^{26} = 67108864$ possible subsets of datasets. Thus, there has to be some sort of selection of what subsets are interesting to look at. There are some pros and cons associated with merging more datasets:

Pros:

- Having more datasets leads to more samples, which gives greater statistical power.
- Having more datasets might lead to better generalizability for the algorithm, as it has to learn what is common across more datasets.

Cons:

Table 4.22: AUC when merging all datasets except one which is used for testing, as described in subsection 4.10.3

Test set	AUC
Abdollahi et al. [2019]	0.617
Asakura et al. [2020]	0.466
Bianchi et al. [2011]	0.688
Boeri et al. [2011]	0.404
Chen et al. [2019]	0.386
Duan et al. [2021]	0.083
Fehlmann et al. [2020]	0.588
Halvorsen et al. [2016]	0.783
Jin et al. [2017]	0.340
Keller et al. [2009]	0.632
Keller et al. [2014]	0.836
Keller et al. [2020]	0.258
Kryczka et al. [2021]	0.346
Leidinger et al. [2011]	0.865
Leidinger et al. [2014]	0.566
Leidinger et al. [2016]	0.794
Li et al. [2017]	0.389
Marzi et al. [2016]	0.629
Nigita et al. [2018]	0.429
Patnaik et al. [2012]	0.576
Patnaik et al. [2017]	0.515
Qu et al. [2017]	0.833
Reis et al. [2020]	0.441
Wozniak et al. [2015]	0.440
Yao et al. [2019]	0.320
Zaporozhchenko et al. [2018]	0.241

- The more datasets are merged, the larger is the problem that the datasets measure different miRNA-sequences. If one takes the intersection of miRNA-sequences this intersection quickly becomes small, and if one takes the union one would end up with many NaN-values.

- Using small groups of datasets can be used to find properties of that group, e.g., one can try to find the level of consistency in sequencing datasets.

One way to balance these conflicting concerns would be to find a compromise. I want the algorithm to be trained on at least 10 miRNA-sequences (similar to subsection 3.6.2 and Berg [2021]) to ensure that the algorithm has some different miRNA-sequences to consider. On the other hand, I want to merge as many datasets as feasible. One possibility then is to generate all subsets that satisfy the following two criteria:

1. The datasets have at least 10 different miRNA-sequences in common.

2. If you add another dataset to the subset, you would end up with less than 10 miRNA-sequences in common.

These subsets might be called maximal subsets. For each maximal subset, there will be a leave-one-out cross validation where each of the datasets is left out. The machine learning will be done using logistic regression. The resulting mean AUC value was 0.525 and the standard deviation was 0.178, which is still worse than the baseline of miR-210. A histogram of the AUC values is shown in Figure 4.5. In short, there was little diagnostic accuracy achieved by using maximal subsets.

4.11 Stratification of the datasets

There are several possibilities as to why the datasets are incompatible. One possibility is that some factors like what technology was used for measuring miRNA-levels play a role. There are other factors as well that differ between the datasets, like cancer stage and what blood fraction was measured (plasma, serum, whole blood, etc.). If these factors play a role, one would expect to see more consistency in datasets that are similar in these characteristics. One way to test this hypothesis is to stratify the datasets based on these characteristics, and see if one sees a larger consistency between the datasets when the datasets are stratified in this way.

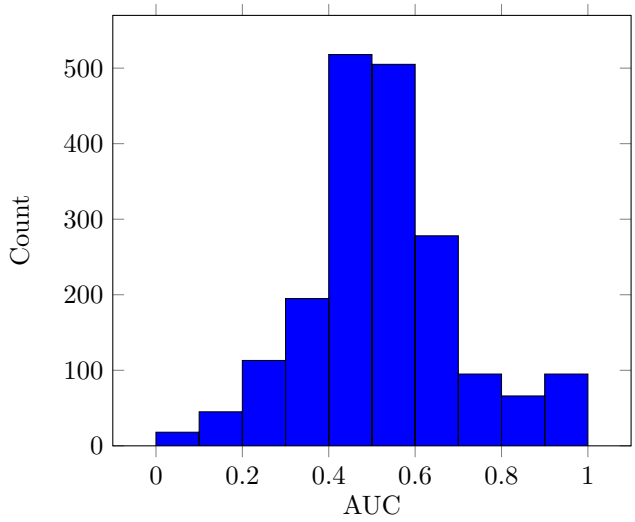


Figure 4.5: Histogram over AUC values when training maximal datasets in a leave-one-out cross validation using logistic regression

4.11.1 Training and testing on pairs of datasets, in-group vs. out-group

Here I will use pairs of datasets and train a model on one of the datasets and test on the other dataset, only that the AUC will be compared when the datasets have the same characteristics to when they have different characteristics. E.g., I will compare the AUC when two datasets are using qRT-PCR to when one is using qRT-PCR and the other study is using a different technology. I will do this stratification for technology and for the type of blood fraction. Here I will use logistic regression and only do pairs of datasets that have at least 10 miRNA-sequences in common.

4.11.1.1 Technology

The results from training on one dataset and testing on another dataset when stratifying using technology are shown in Table 4.23. The in-group is when both datasets use the given technology, and the out-group is when only one of the datasets uses the given technology. As the table shows, the AUC was generally somewhat better in in-groups than out-groups. However, the improvement in AUC was only significant for microarrays. Still, the category “microarray” is hiding heterogeneity, as the microarrays in the studies varied a lot.

Table 4.23: The results when training a logistic regression model on one dataset and testing on another, when stratifying by technology. The in-group is when both datasets have the technology that is listed in the first column. The out-group is when exactly one of the two datasets has the technology that is listed in the first column.
Note: IG = in-group, OG = out-group, mean and standard deviation are of AUC values, t-values are in-group minus out-group and p-values correspond to the t-values

Technology	**Mean IG**	**Std. IG**	**Mean OG**	**Std. OG**	**t-value**	**p-value**
Sequencing	0.535	0.180	0.452	0.165	1.545	0.124
qRT-PCR	0.512	0.153	0.500	0.155	0.416	0.678
Microarray	0.529	0.208	0.477	0.162	2.967	0.003

To check the hypothesis that the problems are due to heterogeneity in the microarray technology, I wanted to do an experiment with a finer stratification of the microarray technology. Of the microarray technologies that have been used multiple times, there were three that used *Exiqon microarrays* (Duan et al. [2021]; Patnaik et al. [2012, 2017]), three that used *Agilent microarrays* (Fehlmann et al. [2020]; Qu et al. [2017]; Li et al. [2017]), three that used *Geniom microarrays* (Keller et al. [2009]; Leidinger et al. [2011]; Keller et al. [2014]) and two that used *SurePrint microarrays* (Leidinger et al. [2014]; Keller et al. [2020]). This experiment will only consider pair of studies where both use microarrays. The in-group here is when the pair of studies have the same type of microarray, and the out-group is when they use different types of microarrays. The results were that the in-group had a mean AUC of 0.612 while the out-group had a mean of 0.518. The difference was not significant using a t-test ($p = 0.056$). It might be that this is due to a low sample size, but even in the in-group the consistency is relatively low compared to the internal consistency in the datasets found in Table 4.12.

4.11.1.2 Blood fraction

The results from training on one dataset and testing on another dataset when stratifying using blood fraction are shown in Table 4.24. The in-group is when both datasets measure the given blood fraction, and the out-group is when only one of the datasets measures the given blood fraction. In contrast to when stratifying by technology, it seems that there is no use in stratifying by blood fraction. None of the changes in AUC are significant, and one of the changes is even negative. It might suggest that technology contributes to more variance in the resulting data than what blood fraction does.

Table 4.24: The results when training a logistic regression model on one dataset and testing on another, when stratifying by blood fraction. The in-group is when both datasets have the blood fraction that is listed in the first column. The out-group is when exactly one of the two datasets has the blood fraction that is listed in the first column.
Note: IG = in-group, OG = out-group, mean and standard deviation are of AUC values, and t-values are in-group minus out-group and p-values correspond to the t-values

Blood fraction	Mean IG	Std. IG	Mean OG	Std. OG	t-value	p-value
Plasma	0.451	0.178	0.497	0.176	−1.766	0.078
Whole blood	0.538	0.109	0.517	0.166	0.659	0.511
Serum	0.549	0.228	0.494	0.185	1.386	0.167

Table 4.25: The AUC values when training a logistic regression model on one dataset and testing on another, when using only late or only early stage cancer samples from datasets where stage is labeled.
Note: mean and standard deviation are of AUC values, the t-value is late minus early and the p-value correspond to the t-value

Mean Late	Std. Late	Mean Early	Std. Early	t-value	p-value
0.528	0.187	0.460	0.140	1.291	0.205

4.11.1.3 Using stages

Cancer stage might be a covariate that hinders the replicability of the datasets. To check this hypothesis, I will do an analysis where I only use the datasets where samples are labeled, and compare the results when only using the early stages to when only using the late stages. If there is higher consistency in the late stages, it would suggest that some of the lack of replicability is due to cancer stage. The results from training on one dataset and testing on another dataset, when only using late or only using early stage cancer from datasets where stage is labeled, are in Table 4.25. There is no significant difference between the AUCs in the two cases, and both are close to 0.50, which suggests that stage does not explain the low AUC scores in the previous results.

4.11.2 Combining all except one

Another attempt will be to take all datasets with a certain characteristic, like technology or blood fraction, and then train on all datasets except one that will

Table 4.26: The results when training an XGBoost model on all datasets except one in a certain category and doing testing on the last dataset, when stratifying by technology. The t-value and the corresponding p-value is for the t-test checking whether the expected AUC is larger than 0.50.

Technology	Mean AUC	Std. AUC	t-value	p-value
Sequencing	0.625	0.089	2.797	0.034
Microarray	0.505	0.262	0.077	0.470
qRT-PCR	0.493	0.219	−0.086	0.533

be used for testing, and use AUC as the metric. For checking whether the AUC values are better than chance levels I checked a one-sided hypothesis of AUC > 0.50 using a t-test. This is similar to subsection 4.10.3, only with stratification of the datasets. I will use the union of the miRNAs in the datasets in each category to train on. To ensure that missing values will not be a problem, I will use XGBoost as the model as it handles missing values by default. I will also try to do this using datasets where cancer stage is labeled, and try both using only early cancer samples and using only late cancer samples.

4.11.2.1 Technology

The results from stratifying by technology are shown in Table 4.26. Sequencing is an outlier where the AUC was better than the other categories. Notably, an AUC of 0.625 is higher than any of the other AUCs achieved so far when testing on a different dataset than training on, but it might be due to chance as the AUC is not significantly higher than 0.50 when adjusting for multiple testing. It does not seem like technology is the main reason for the low consistency between the datasets.

Also, here I want to see whether stratifying by subtypes of microarrays will be beneficial. The training will be done on maximally two datasets, as the subcategories are small with the largest ones having three datasets. The resulting mean AUC was 0.567 and the resulting standard deviation was 0.282, which was not significantly better than 0.50 ($p = 0.225$). This suggests that neither here heterogeneity in the microarray technology was the reason for the poor results for the microarrays. Even an AUC of 0.567 is much lower than the internal consistency found in Table 4.12.

Table 4.27: The results when training an XGBoost model on all datasets except one in a certain category and doing testing on the last dataset, when stratifying by blood fraction. The t-value and the corresponding p-value are for the t-test checking whether the expected AUC is larger than 0.50.

Blood fraction	Mean AUC	Std. AUC	t-value	p-value
Serum	0.531	0.222	0.337	0.375
Whole blood	0.583	0.079	2.773	0.016
Plasma	0.376	0.184	−1.908	0.951

4.11.2.2 Blood fraction

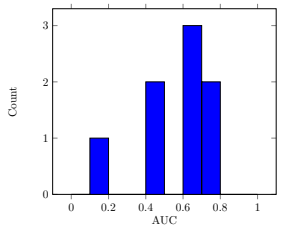
The results from stratifying by blood fraction are shown in Table 4.27. None of the AUCs were significantly larger than 0.50 when adjusted for multiple testing. This suggests that the lack of consistency is not due to blood fraction either.

4.11.2.3 Distribution of AUC values

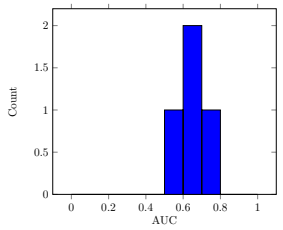
In the subsections above, only summary statistics were reported. However, mean and variance can hide a lot of information about the distribution, e.g. whether the distribution is unimodal or bimodal. As t-values have been used to check for statistical significance, there has been an implicit assumption that AUC values have been approximately normally distributed. To be sure, some checks have been done. I am not including the results from stratifying by cancer stage here as those values do not use full datasets, and are thus less comparable. Histograms of the AUCs are shown in Figure 4.6. It is hard to judge the normality of the plots due to the low sample sizes. Therefore, I have also plotted a histogram and a Q-Q plot combining all the AUC values from the different categories. Those are in Figure 4.7. The Q-Q plot shows that the distribution of AUC values follows the normal distribution quite nicely, except for slight deviations in the tails of the distribution. Thus, the normality assumption seems to hold.

4.11.2.4 Using stages

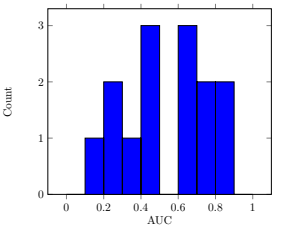
The results from stratifying using cancer stages can be found in Table 4.28. The mean AUCs, both when only using early stage cancer and only using late stage cancer, were only slightly higher than 0.50 and the differences were not significant, which suggests that there is no improvement in AUC by stratifying by stage.



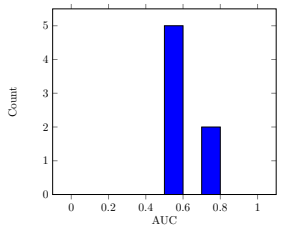
(a) Histogram over AUC values when training on all datasets except one in the qRT-PCR category and testing on the last dataset.



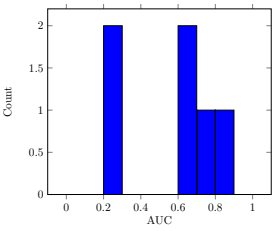
(b) Histogram over AUC values when training on all datasets except one in the sequencing category and testing on the last dataset.



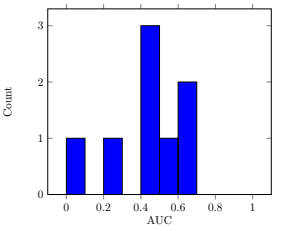
(c) Histogram over AUC values when training on all datasets except one in the microarray category and testing on the last dataset.



(d) Histogram over AUC values when training on all datasets except one in the whole blood category and testing on the last dataset.



(e) Histogram over AUC values when training on all datasets except one in the serum category and testing on the last dataset.

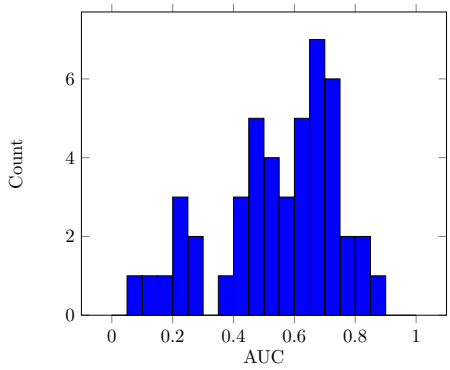


(f) Histogram over AUC values when training on all datasets except one in the plasma category and testing on the last dataset.

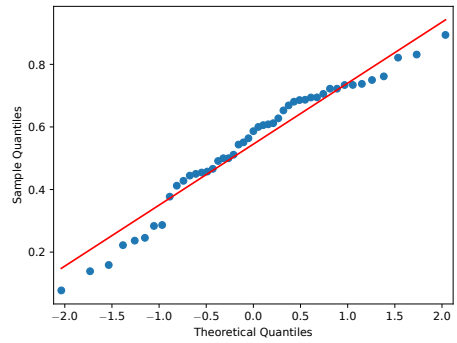
Figure 4.6: Histogram over AUC values when training on all datasets except one in a category and testing on the last dataset

Table 4.28: The results when training on all datasets except one, using datasets where cancer stage is labeled, when stratifying by cancer stage. The mean and standard deviation are of AUC values, and the t-values and the corresponding p-values correspond to a two-sided t-test with a null hypothesis of AUC = 0.50.

Cancer stage	**Mean**	**Std.**	**t-value**	**p-value**
Early	0.523	0.171	0.354	0.736
Late	0.509	0.194	0.125	0.904



(a) Histogram over AUC values when training on all datasets except one in one category and testing on the last dataset, aggregated over all categories.



(b) Q-Q plot over AUC values when training on all datasets except one in one category and testing on the last dataset, aggregated over all categories.

Figure 4.7: Histogram and Q-Q plot over AUC values when training on all datasets except one in a category and testing on the last dataset

4.12 PCA for removing artifacts

The measured miRNA-levels will have some noise that is due to the technology that is used for measurement. One possible way to remove technical noise is to remove the first two principal components from the data, with an assumption that the removed principal components correspond to technical noise rather than biology. It is difficult to say whether this is the case or not. If the datasets are more comparable when the principal components are removed than when they are not, then it would seem plausible that these principal components correspond to technical noise, or at least have little to no connection with lung cancer.

4.12.1 Check comparability using PCA

One way to check if the datasets are more comparable is to check their joint PCA-plot. There are some problems with that. For once, the miRNA-sequences are not the same in the different datasets, which means that the principal components will not represent the datasets fully. Another problem is that there are many datasets. The more datasets that are plotted in one PCA-plot, the more chaotic the plot becomes, the fewer miRNA-sequences there are in common, and the less weight each dataset would have on the PCA on the joint dataset. Furthermore, I cannot plot all datasets against all datasets, as that would lead to too many

PCA-plots. Therefore, I will plot two and two datasets in PCA-plots and see whether they became more comparable, using only the largest datasets.

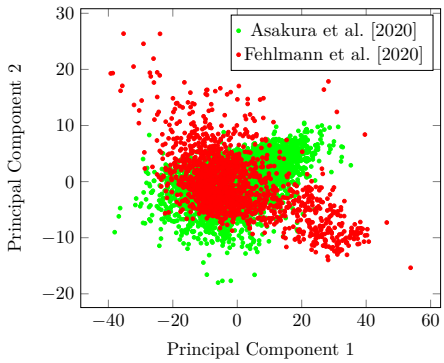
The resulting PCA-plots when the two first principal components are removed are shown in Figure 4.8. It is hard to say whether the PCA adjustments made the datasets more similar using these PCA-plots, but it seems like the spreads are more equal after the removal of the principal components, with the exception of Asakura et al. [2020] and Fehlmann et al. [2020].

4.12.2 Using machine learning models

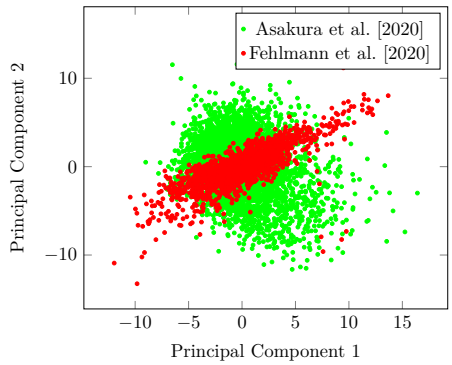
The results from combining all datasets using sequencing were relatively good (see subsubsection 4.11.2.1), and will therefore be used to check if removing the principal components lead to better comparability, as you would assume that a dataset where noise is removed would be a dataset that would be more comparable to the sequencing datasets, as the sequencing datasets seem to have some external validity when compared to other sequencing datasets.

The way this will be done is for each dataset that is not using sequencing, I will find the miRNAs that they have in common with all the sequencing datasets. Thereafter, using these miRNAs I will do a leave-one-out cross validation on the sequencing datasets, similarly to subsection 4.11.2. I will also do a cross validation on the other dataset, similar to section 4.7. Finally, I will train a model on the other dataset and test on the sequencing datasets, and vice versa. All the machine learning will be done using XGBoost, as it was that model that performed well in subsubsection 4.11.2.1.

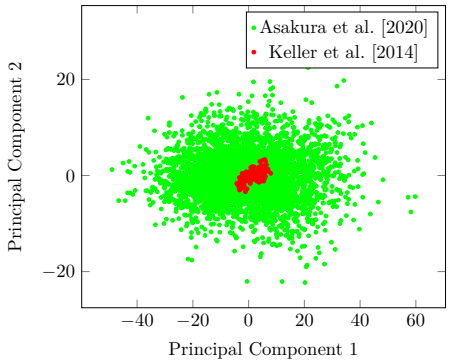
The resulting AUC-values from doing the experiment are shown in Table 4.29 and Table 4.30. The internal results were poorer both in the sequencing datasets and in the non-sequencing datasets, which means that some information about case-control was lost when removing the two first principal components. This does not mean that these principal components were due to biological factors connected to cancer, as it might be technical artifacts like batch effects or be due to demographic differences between cases and controls. The external validity, here represented by to which degree it is possible to get good results when training or testing on sequence data, was low in both cases. Everything considered, the data suggest that removing the two first principal components did not have any effect on the comparability. However, it is important to recognize that this was a one-size-fits-all solution, and that it is possible that the two first principal components correspond to technical noise in some datasets, but not in others, and that this effect did not show up in aggregate. Indeed, looking at the PCA plots in Berg [2021] it seems that some datasets have clustering that looks like batch effects, while other datasets have not.



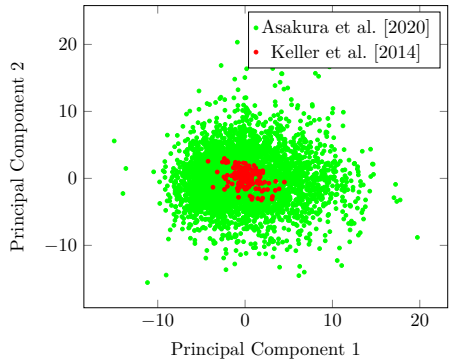
(a) PCA of Asakura et al. [2020] and Fehlmann et al. [2020] using the two first principal components of the joint dataset, without any principal components removed



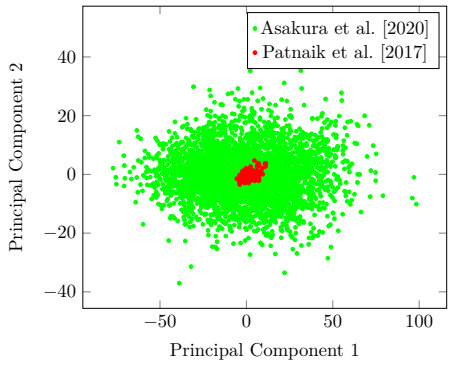
(b) PCA of Asakura et al. [2020] and Fehlmann et al. [2020] using the two first principal components of the joint dataset, after the two first principal components of each dataset are removed



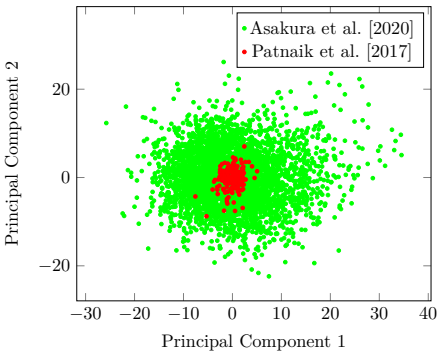
(c) PCA of Asakura et al. [2020] and Keller et al. [2014] using the two first principal components of the joint dataset, without any principal components removed



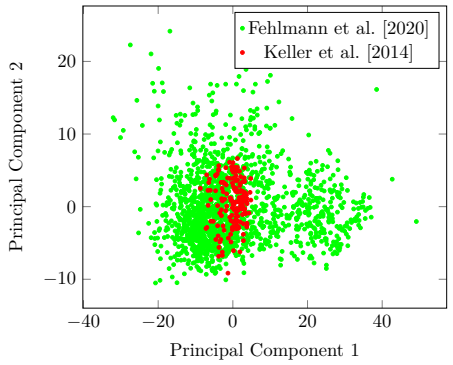
(d) PCA of Asakura et al. [2020] and Keller et al. [2014] using the two first principal components of the joint dataset, after the two first principal components of each dataset are removed



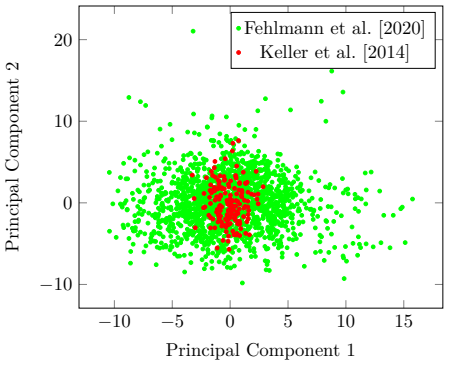
(e) PCA of Asakura et al. [2020] and Patnaik et al. [2017] using the two first principal components of the joint dataset, without any principal components removed



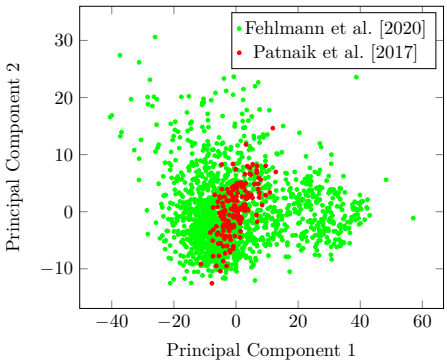
(f) PCA of Asakura et al. [2020] and Patnaik et al. [2017] using the two first principal components of the joint dataset, after the two first principal components of each dataset are removed



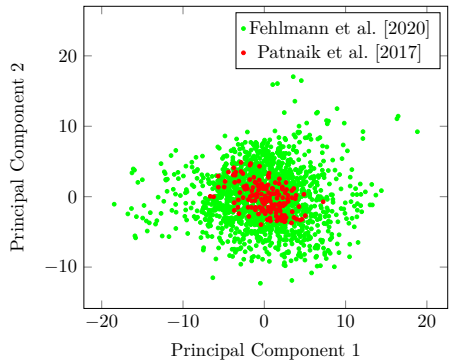
(g) PCA of Fehlmann et al. [2020] and Keller et al. [2014] using the two first principal components of the joint dataset, without any principal components removed



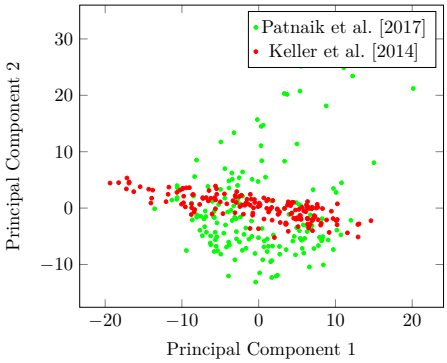
(h) PCA of Fehlmann et al. [2020] and Keller et al. [2014] using the two first principal components of the joint dataset, after the two first principal components of each dataset are removed



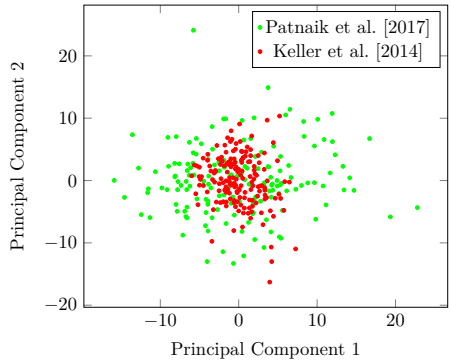
(i) PCA of Fehlmann et al. [2020] and Patnaik et al. [2017] using the two first principal components of the joint dataset, without any principal components removed



(j) PCA of Fehlmann et al. [2020] and Patnaik et al. [2017] using the two first principal components of the joint dataset, after the two first principal components of each dataset are removed



(k) PCA of Patnaik et al. [2017] and Keller et al. [2014] using the two first principal components of the joint dataset, without any principal components removed



(l) PCA of Patnaik et al. [2017] and Keller et al. [2014] using the two first principal components of the joint dataset, after the two first principal components of each dataset are removed

Figure 4.8: PCA of datasets with and without removing the two first principal components in each individual dataset. OBS: the axes differ between the plots.

Table 4.29: The resulting AUC-values when using XGBoost and doing cross validation internally in the given study, doing cross validation inside the sequencing datasets, doing training on the given study and testing in the sequencing datasets, and doing training on the sequencing datasets and testing in the given study, all without removing the two first principal components.
Note: I. = internal, Seq = sequencing, To seq = training model on the study and testing on the sequencing datasets, From seq = training model on the sequencing datasets and testing on the study

Study	**I. study**	**I. seq**	**To seq**	**From seq**
Wozniak et al. [2015]	0.597	0.719	0.514	0.554
Bianchi et al. [2011]	0.779	0.716	0.541	0.497
Zaporozhchenko et al. [2018]	0.808	0.401	0.335	0.555
Leidinger et al. [2016]	0.950	0.699	0.287	0.506
Reis et al. [2020]	0.943	0.715	0.661	0.553
Asakura et al. [2020]	0.926	0.736	0.536	0.563
Duan et al. [2021]	0.650	0.550	0.750	0.560
Leidinger et al. [2014]	0.468	0.736	0.521	0.466
Keller et al. [2009]	0.942	0.735	0.350	0.509
Patnaik et al. [2017]	0.549	0.736	0.366	0.292
Patnaik et al. [2012]	0.684	0.718	0.523	0.513
Fehlmann et al. [2020]	0.974	0.743	0.410	0.505
Marzi et al. [2016]	0.892	0.610	0.250	0.440
Halvorsen et al. [2016]	0.967	0.698	0.461	0.524
Boeri et al. [2011]	0.500	0.649	0.088	0.542
Leidinger et al. [2011]	0.604	0.741	0.479	0.417
Qu et al. [2017]	0.854	0.642	0.611	0.606
Li et al. [2017]	0.500	0.711	0.222	0.525
Keller et al. [2014]	0.852	0.708	0.330	0.543
Keller et al. [2020]	0.535	0.555	0.686	0.663
Kryczka et al. [2021]	0.625	0.684	0.320	0.365
Abdollahi et al. [2019]	0.851	0.516	0.382	0.433
Average	0.748	0.669	0.437	0.506

Table 4.30: The resulting AUC-values when using XGBoost and doing cross validation internally in the given study, doing cross validation inside the sequencing datasets, doing training on the given study and testing in the sequencing datasets, and doing training on the sequencing datasets and testing in the given study, all while removing the two first principal components.
Note: I. = internal, Seq = sequencing, To seq = training model on the study and testing on the sequencing datasets, From seq = training model on the sequencing datasets and testing on the study

Study	**I. study**	**I. seq**	**To seq**	**From seq**
Wozniak et al. [2015]	0.621	0.530	0.552	0.613
Bianchi et al. [2011]	0.716	0.512	0.446	0.570
Zaporozhchenko et al. [2018]	0.542	0.625	0.447	0.617
Leidinger et al. [2016]	0.803	0.620	0.404	0.353
Reis et al. [2020]	0.586	0.516	0.514	0.595
Asakura et al. [2020]	0.931	0.511	0.403	0.453
Duan et al. [2021]	0.250	0.535	0.500	0.584
Leidinger et al. [2014]	0.559	0.511	0.540	0.598
Keller et al. [2009]	0.800	0.483	0.759	0.531
Patnaik et al. [2017]	0.522	0.511	0.373	0.321
Patnaik et al. [2012]	0.360	0.483	0.485	0.442
Fehlmann et al. [2020]	0.961	0.504	0.367	0.508
Marzi et al. [2016]	0.828	0.530	0.493	0.425
Halvorsen et al. [2016]	0.914	0.633	0.559	0.481
Boeri et al. [2011]	0.683	0.624	0.430	0.417
Leidinger et al. [2011]	0.541	0.494	0.586	0.534
Qu et al. [2017]	0.583	0.538	0.528	0.598
Li et al. [2017]	0.500	0.557	0.611	0.367
Keller et al. [2014]	0.564	0.500	0.559	0.552
Keller et al. [2020]	0.600	0.391	0.581	0.560
Kryczka et al. [2021]	0.564	0.590	0.488	0.641
Abdollahi et al. [2019]	0.715	0.441	0.443	0.606
Average	0.643	0.529	0.503	0.517

Table 4.31: The resulting AUC-values when doing the experiment as described in section 3.7. The threshold is the mean RPM needed for a miRNA-sequence to be included in the dataset. Intersection (I) and union (U) represent whether the model was trained on the intersection of the miRNAs (logistic regression) or the union of the miRNAs (XGBoost). # miRNA is the number of miRNAs in the intersection or the union of the datasets, when filtered according to the thresholds.

Threshold	**AUC (I)**	**# miRNA (I)**	**AUC (U)**	**# miRNA (U)**
0	0.625	68	0.742	1042
1	0.626	68	0.734	839
10	0.640	68	0.643	608
100	0.536	33	0.684	326
1000	0.718	7	0.716	148

4.13 Finding RPM threshold for sequencing data

Not all miRNA-sequences have been found in all samples in the sequencing datasets. One question is what to do when a sequence is barely read in a dataset. One might remove it or not, depending on whether one considers the levels of the miRNA-sequence relevant or not. As these sequences are barely read, they might be a source of noise rather than a source of information about case-control characteristics. One way to analyze whether they are noise or they have relevant information is to check the consistency across the sequencing datasets when they are removed and when they are not removed.

Here I have taken leave-one-out cross validation on the sequencing datasets, filtering out low expressed miRNAs at different thresholds for mean RPM, as described in section 3.7. The results are shown in Table 4.31. There are a few things to note. The first thing is that the 0, 1 and 10 thresholds give the same number of miRNAs in the intersection, but still the results differ. This is because the normalization of the data was done after filtering out miRNAs based on RPM. When setting the mean variance to 1 in a dataset, this will lead to a different denominator depending on which miRNAs were filtered out. Thus, the numerical values differ slightly at the different thresholds, even if the miRNAs in the intersection are the same. Secondly, it is interesting to see that one gets a mean AUC of 0.718 with only 7 different miRNAs.

To check whether these 7 miRNAs actually predict case-control status, I did an experiment where I trained logistic regression models on the 1000 RPM thresholded sequencing datasets using intersection between the 7 miRNAs and

Table 4.32: The resulting AUC-values when doing logistic regression training on either the 1000 RPM thresholded or the non-thresholded sequencing datasets, using all miRNAs that the datasets have in common in the respective cases

Study	AUC no threshold	AUC threshold
Asakura et al. [2020]	0.564	0.769
Boeri et al. [2011]	0.281	0.316
Fehlmann et al. [2020]	0.386	0.426
Halvorsen et al. [2016]	0.408	0.424
Keller et al. [2009]	0.582	0.548
Keller et al. [2014]	0.338	0.331
Leidinger et al. [2011]	0.372	0.434
Leidinger et al. [2014]	0.365	0.357
Patnaik et al. [2012]	0.414	0.293
Patnaik et al. [2017]	0.434	0.458
Qu et al. [2017]	0.639	0.583
Reis et al. [2020]	0.922	0.898
Wozniak et al. [2015]	0.510	0.530
Average	0.478	0.490

the miRNAs in the different other datasets if there were at least 4 miRNAs in common. I compared that to the same experiment using the non-thresholded sequencing datasets and all miRNAs that were in common between the sequencing datasets and the other datasets, where the models thus generally could use more miRNA-sequences. The other datasets were used as test sets. The results are shown in Table 4.32. There was no significant difference between the experiment when using the thresholded datasets and when using the non-thresholded datasets ($p = 0.864$). This suggests that either the non-sequencing datasets generally have poor quality or these miRNAs are not as predictive of case-control characteristics as it would seem from looking at results only from the sequencing datasets.

4.14 Checking for red blood cells

One thing that can go wrong when doing experiments where serum and plasma are extracted (see: subsection 2.1.5) is that red blood cells might burst, and their contents will then be spread into the serum and plasma. Then one can end up with different miRNA-levels in this serum or plasma, in a way that would make

the resulting sample more similar to a typical whole blood sample rather than a plasma or serum sample.

There are ways to trace whether red blood cells appear in the sample or not, as they have a known footprint in miRNA-levels. According to Sun et al. [2020], two of the most characteristic miRNAs for red blood cells are miR-486-5p and miR-451a. Thus, high levels of these miRNAs would suggest that miRNA from red blood cells are found in the sample.

Till now, all comparison of data between datasets has been done on standardized data. However, this standardization sets the mean expression for each miRNA to zero, which removes any differences in mean expression between datasets. As I want to compare levels of the miRNA between datasets I have to use raw values, but they are hard to compare directly as the raw values are found using different technologies and thus are on different scales.

The way to come around these issues is to find some miRNA-sequences to use as controls that one can calculate the expression of miR-486-5p and miR-451a relative to. In the studies reviewed by Donati et al. [2019] miR-16 and miR-93 were the most common ones. Thus, they are the ones I will use here. The reason I chose to use two different miRNA-sequences both for footprint and control is that it will hopefully paint a more accurate picture, and not all datasets have all sequences, therefore having more sequences lead to the possibility of getting more information on more datasets.

The relative expression of miR-486-5p and miR-451a is in Table 4.33. There is a lot of variation between the datasets. One interesting experiment is to group them by what blood fraction is used and then take the average, to see whether blood fraction actually matters. As miR-451a with miR-93 as control is the one with the most values, it is the one that will be used. A table of the mean for each blood fraction group is found in Table 4.34. Surprisingly, whole blood seemingly has less miR-451a than either the serum or the plasma group. This renders this experiment somewhat meaningless, as miR-451a was meant to find samples with a high level of red blood cells.

Another way to look at it is to see to which degree miR-486-5p and miR-451a vary internally in a dataset. Intuitively, if there are many red blood cells in the dataset, there will be more variance in these miRNAs, as the red blood cells are a main contributor to the level of expression of these miRNAs. A table with the relative variances is in Table 4.35. Whole blood datasets had the highest relative variance in miR-486-5p but not in miR-451a according to Table 4.36. Furthermore, the high relative variance in miR-486-5p in whole blood is largely driven by Leidinger et al. [2014], which was a big outlier in Table 4.35. Excluding Leidinger et al. [2014] gives a mean relative variance in the whole blood datasets of 1.103 in miR-486-5p and 2.295 in miR-451a, which are lower even though it is still the highest for miR-486-5p. Given the earlier results and the uncertainty

Table 4.33: The relative expression of miR-486-5p and miR-451a to miR-16 and miR-93 in the different datasets. The values are calculated as $\frac{\text{footprint expression} - \text{control expression}}{\text{control expression}}$. C = control

Studies	miR-486-5p C: miR-16	miR-486-5p C: miR-93	miR-451a C: miR-16	miR-451a C: miR-93
Asakura et al. [2020]	-36.403%	143.096%	232.049%	1169.239%
Boeri et al. [2011]	-35.767%	162.134%	19.454%	387.492%
Chen et al. [2019]	-15.134%	38.145%	99.049%	224.012%
Duan et al. [2021]	-75.000%	-108.333%	300.000%	-233.333%
Fehlmann et al. [2020]	-136.022%	-139.201%	-107.105%	-107.732%
Halvorsen et al. [2016]	-1096.492%			
Jin et al. [2017]	490.574%	-266.217%	681.561%	-319.970%
Keller et al. [2009]	-62.340%	3.987%	16.361%	221.294%
Keller et al. [2014]	-121.880%	-110.952%	-40.278%	-70.108%
Keller et al. [2020]	-4587.592%		-8524.127%	
Leidinger et al. [2011]	-218.204%	-63.009%	-83.217%	-105.252%
Leidinger et al. [2014]	413.936%	-1103.786%	823.077%	-1902.894%
Leidinger et al. [2016]				-32.045%
Nigita et al. [2018]				-188.699%
Patnaik et al. [2012]	-89.745%	-90.109%	-52.739%	-54.417%
Patnaik et al. [2017]	-238.871%	-109.828%	-1638.338%	-208.875%
Qu et al. [2017]	-128.963%	-121.686%	-85.323%	-89.010%
Reis et al. [2020]			25.021%	312.228%
Wozniak et al. [2015]		-286.667%		33.333%
Zaporozhchenko et al. [2018]	-98.858%	-98.623%	19.444%	43.960%

Table 4.34: The mean relative expression of miR-451a to miR-93 (with similar formula as in Table 4.33) when grouped by blood fraction. P. Blood = Peripheral blood, Ex = exosomal

Blood fraction	Mean
Serum	467.953%
Plasma	84.578%
Whole blood	-401.869%
Blood cells	221.294%
P. Blood	-70.108%
Plasma Ex	-188.699%

Table 4.35: The relative variance of miR-486-5p and miR-451a in the different datasets. The relative variance is the variance in the expression of these miRNAs, divided by the mean variance among all miRNAs.

Study	miR-486-5p	miR-451a
Asakura et al. [2020]	0.964	2.998
Boeri et al. [2011]	0.989	0.730
Chen et al. [2019]	0.819	1.047
Duan et al. [2021]	0.894	2.486
Fehlmann et al. [2020]	0.851	0.955
Halvorsen et al. [2016]	0.299	
Jin et al. [2017]	1.150	1.617
Keller et al. [2009]	0.609	1.157
Keller et al. [2014]	0.595	1.461
Keller et al. [2020]	0.984	2.601
Leidinger et al. [2011]	0.673	1.294
Leidinger et al. [2014]	6.141	5.183
Leidinger et al. [2016]		0.500
Li et al. [2017]		2.866
Nigita et al. [2018]		1.027
Patnaik et al. [2012]	1.607	3.425
Patnaik et al. [2017]	1.280	3.507
Qu et al. [2017]	1.177	3.849
Reis et al. [2020]		3.357
Wozniak et al. [2015]	0.625	0.463
Zaporozhchenko et al. [2018]	0.224	0.634

around the current results, it is hard to say whether relative variance has any external validity when it comes to whether samples are contaminated.

A final way to look at it is to see whether there are many outliers in miR-486-5p or miR-541a expression. One might assume that the expressions of the miRNAs are normally distributed. In that case, one can fit a t-distribution to the data and see whether 95% of the data is inside the interval where 95% of the data should lie given by the mean and the standard deviation of the data (see subsection 2.2.8 for how to compute such an interval). The portion in each dataset that is inside such an interval is shown in Table 4.37. The dataset with the lowest portion of samples inside the interval was Leidinger et al. [2014], which also had the highest relative variance of these miRNAs in Table 4.35. However, Leidinger et al. [2014] is a whole blood-dataset, so neither the high variance nor

Table 4.36: The mean relative variance of miR-486-5p and miR-451a in the datasets in the different groups. The relative variance is the variance in the expression of these miRNAs, divided by the mean variance among all miRNAs. Note: P. blood = Peripheral blood, Ex = exosomal

Group	miR-486-5p	miR-451a
Serum	0.785	2.695
Plasma	0.831	1.820
Whole blood	2.110	2.477
Blood cells	0.609	1.157
P. Blood	0.595	1.461
Plasma Ex		1.027

the relatively high portions of outliers are due to contamination of red blood cells.

4.15 Web application for visualizing data

A web application was made in this project in order to make it possible to visualize the data that has been collected. A live demo of the web application is available at `https://mirna-visualizer.netlify.app/`, and code and computations are available at `https://github.com/OleFredrik1/masterthesis`.

4.15.1 Some considerations that were made during the project

How should computations, especially computationally expensive computations like PCA, be done? There were three main options:

1. Precompute all possible computations
2. Do computations server-side on demand
3. Do computations client-side on demand

All three options had their advantages and disadvantages. The main advantage of precomputing all possible computations was to save computation time during the use of the application. One disadvantage was that very many computations had to be done despite the results might not being used. There is also a large amount of resulting data that has to be stored. Another disadvantage that it shares with the "compute server-side on demand" is that new data has to be fetched when the user wants to visualize a new result. Finally, one disadvantage of the "client-side

Table 4.37: The portion of the samples which has an expression of the given miRNAs inside a 95% interval given by the sample mean and standard deviation, calculated as in subsection 2.2.8.

Study	**miR-486-5p**	**miR-451a**
Asakura et al. [2020]	0.966	0.940
Boeri et al. [2011]	0.920	1.000
Chen et al. [2019]	0.926	0.981
Duan et al. [2021]	0.917	1.000
Fehlmann et al. [2020]	0.947	0.950
Halvorsen et al. [2016]	0.944	
Jin et al. [2017]	0.921	0.921
Keller et al. [2009]	0.944	0.917
Keller et al. [2014]	0.940	0.928
Keller et al. [2020]	0.950	0.970
Leidinger et al. [2011]	0.936	0.979
Leidinger et al. [2014]	0.825	0.938
Leidinger et al. [2016]		0.950
Li et al. [2017]		1.000
Nigita et al. [2018]		1.000
Patnaik et al. [2012]	0.978	0.956
Patnaik et al. [2017]	0.957	0.957
Qu et al. [2017]	0.923	0.923
Reis et al. [2020]		0.952
Wozniak et al. [2015]	0.960	0.960
Zaporozhchenko et al. [2018]	0.926	0.926

on demand" is that the raw data files (especially of Asakura et al. [2020]) are quite large. Thus, having to load those before the application can start leads to a long loading time for the application. I tried both the "precompute" and the "client-side" options, and decided in favor of the "precompute"-option as it had a low loading time when one visits the web application, and empirically it was on average faster to fetch new results than to compute them client-side, with the biggest difference when doing heavy computations (like PCA) on large datasets (like Asakura et al. [2020]).

All results shown in the dashboard have been precomputed using Python scripts and stored in JSON-files that are loaded into the application on demand.

4.15.2 PCA Single Dataset

The first module that was made for this web application was a module for visualizing the PCA plot for a single dataset. The possible options are (1) to select the dataset whose PCA plot is shown, (2) the principal component for the x-axis, (3) the principal component for the y-axis, (4) the size of the markers (i.e. the dots in the chart representing each sample), (5) the opacity of the markers and (6) whether to color cancer samples differently from control samples. The proportion of variance explained by the principal components is shown in the axis labels.

4.15.3 PCA Two Datasets

Similar to *PCA Single Dataset* with the difference that you also select a second dataset to be plotted. You can also select whether the principal components are computed using only the first dataset, only the second dataset or both datasets. In any way, all the computations only use the miRNAs that are in both datasets.

4.15.4 PCA Two Datasets (Matrix)

Here, there are several rows and several columns, where each row and each column represents a single dataset. In the intersection between a row and a column, there is a PCA plot similar to *PCA Two Datasets* using the pair of datasets represented by the row and the column. The options are similar to *PCA Two Dataset*, with the difference that here one selects subsets of datasets to be represented by the rows and the columns respectively. In addition, one selects whether the principal components are calculated based on the row-dataset, the column-dataset or both. Similarly to *PCA Two Datasets*, all computations only use the miRNAs that are in both datasets (here: both the column-dataset and the row-dataset). A screenshot is shown in Figure 4.9.

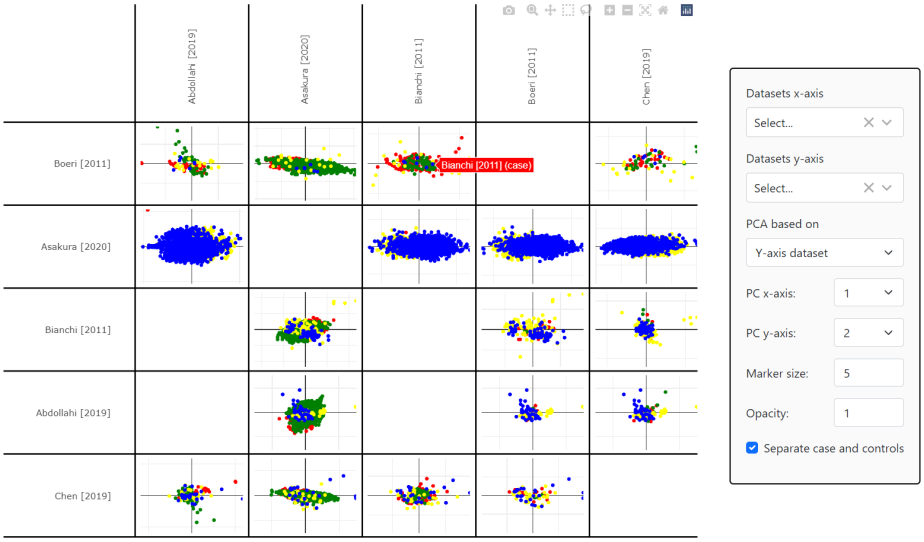


Figure 4.9: Screenshot of *PCA Two Datasets (Matrix)*

4.15.5 Boxplot

The boxplot is a plot that shows the expression of a certain miRNA in the different datasets in cases and controls. In addition, a p-value is computed for the separation between cases and controls using a t-test. The p-value is adjusted using Bonferroni, where the p-value for a miRNA-sequence is multiplied by the number of miRNA-sequences and the p-value for a single combination of miRNA and dataset is multiplied by the number of miRNA and dataset combinations. There are three options here: (1) the ordering of the datasets in the boxplot, where one can sort the datasets alphabetically or on either the size or p-value of the separation between cases and controls, (2) the ordering of the miRNA selector (i.e. (3)), where one can either sort the miRNAs alphabetically or based on the p-value of the separation when using all datasets, and finally (3) a selector where one can choose what miRNA will be shown in the boxplot. A screenshot of the boxplot is shown in Figure 4.10.

4.15.6 Log Fold Change correlation

The log fold change correlation plot is a plot where each data point in a scatterplot is the log fold change of a miRNA-sequence in two datasets, where the x-coordinate is the log fold change in the first dataset and the y-coordinate is

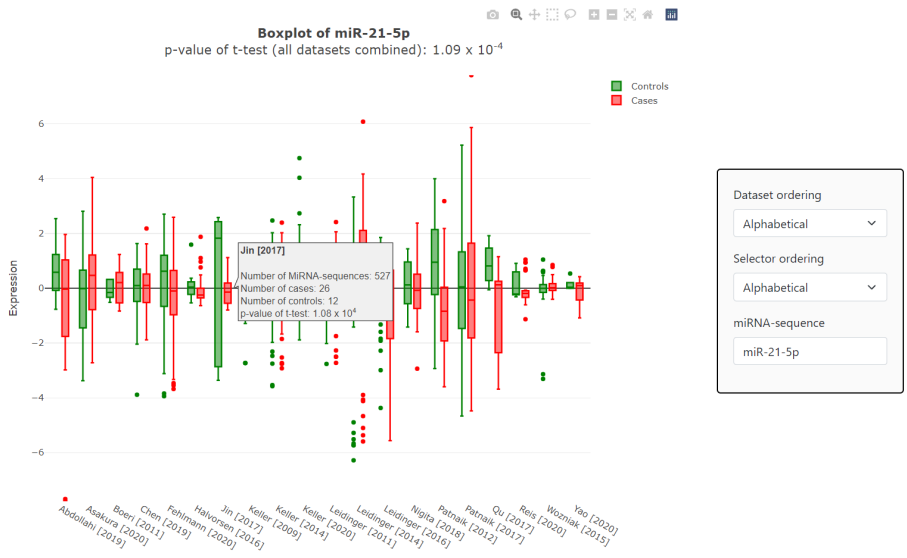


Figure 4.10: Screenshot of *Boxplot*

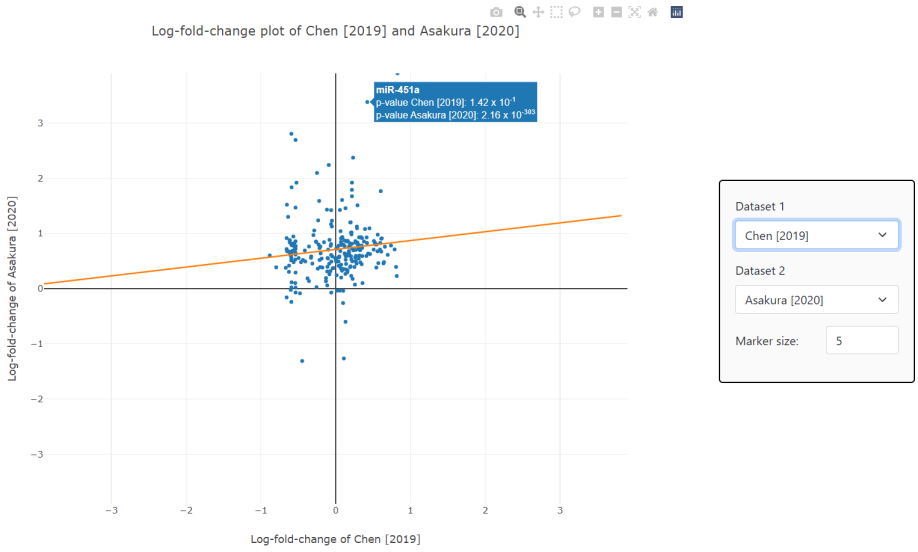


Figure 4.11: Screenshot of *Log Fold Change correlation plot*

the log fold change in the second dataset. Each data point is labeled with the miRNA-sequence of the data point and the p-value of the separation between cases and controls in the two datasets using a t-test. There is also a regression line, which is labeled with the correlation in log fold change and the p-value of the correlation. The options here are (1) to select the two datasets for the plot and (2) the size of the markers. A screenshot of the log fold change correlation plot is shown in Figure 4.11.

4.15.7 Log Fold Change correlation (Matrix)

This is a matrix of *Log Fold Change correlation* plots, made in a similar manner to *PCA Two Datasets (Matrix)* as shown in Figure 4.9. Here the options are similar to *Log Fold Change correlation*, with the difference that here one can select two subsets of the datasets to be represented by the columns and by the rows respectively.

4.15.8 Pairwise machine learning

The point of this page is to show how well a machine learning algorithm can generalize across two datasets. However, one needs a baseline to compare a model to when assessing how well a machine learning model performs on a task. Section 4.7 shows that a machine learning model generally performs well internally in a dataset, whilst section 4.10 shows that a machine learning model generally performs poorly across datasets.

Comparing an internal machine learning model with a machine learning model across studies might be a good comparison. This is because the internal machine learning model gives a lower bound on how well it is possible to separate cases and controls in a dataset. However, it is not in any way a perfect bound as (1) the internal model might use miRNA expression patterns that are correlated with case-control characteristics in the dataset, but are due to other factors than case-control characteristics and are thus impossible to replicate in other datasets and (2) there might be other patterns of case-control characteristics that an internal model might not recognize due to low sample size, but if these patterns are also present in larger datasets, then a model trained on a larger dataset might outperform the internal model. Anyway, a suboptimal baseline was considered better than none in this case.

The plot is a bar plot of AUC values resulting from training a machine learning algorithm on one of the datasets in a pair of datasets, and then either testing on the same dataset or the other dataset in the pair. The calculation and the meaning of the AUC values in this plot are explained in detail in subsection 3.8.1. The options here are (1) which pair of datasets to use and (2) what machine

learning algorithm to use (with the options: logistic regression, SVM, random forest and XGBoost).

4.15.9 Pairwise machine learning (Matrix)

This is a plot with a matrix of *Pairwise machine learning* plots, made in a similar manner to *PCA Two Datasets (Matrix)*. The options are similar to *Pairwise machine learning*, with the difference that one selects two subsets of the dataset to be represented by the rows and the columns respectively instead of only selecting a pair of datasets.

4.15.10 Sample p-value PCA (single)

One important question is how to identify outlier samples, as outlier samples might be the result of e.g. technical issues, which would mean that removal of these samples would lead to better results. One way is to see how the models classify a certain sample. If the sample has a similar classification pattern to other samples it would be reasonable to assume that the sample is not an outlier. The way I am going to do this is by first taking a dataset as a test set. Then I would find all datasets that have at least four miRNA in common with the test dataset. These datasets would be used as training sets. Each training set is then used for training a machine learning model. The model will then give a probability that each sample in the test set is a cancer sample.

After the probability is calculated for each sample in a test set using each of the training datasets, PCA is conducted on these probabilities.

The machine learning models used here are logistic regression, SVM, random forest and XGBoost; and the whole experiment is conducted once with each of these models.

The proportion of variance explained by the principal components is shown in the axis labels. One can choose to see the PCA loadings, which in this case are the eigenvectors of the PCA. There have been no adjustments on the eigenvectors based on the variance explained. The options are (1) whose test dataset's PCA is shown, (2) whose datasets' loadings are shown, (3) which machine learning algorithm is used, (4) which principal components are plotted along the axes, (5) a scaling for the loadings (as the loadings are unadjusted one might want some scalar scaling as it would change the vector lengths in the plot), (6) marker size, (7) marker opacity, (8) option of whether samples should be shown in the plot and (9) whether to color case and control samples differently. A screenshot of this plot is shown in Figure 4.12.

4.15.11 Sample p-value PCA (combined)

This is similar to *Sample p-value PCA (single)* with the difference that here the PCA is calculated over all samples in all studies. This results in some issues. First of all, the set of training sets can be different for each test set, as it varies which datasets a certain dataset has at least four miRNAs in common with. This is solved by naively setting all probabilities to 0.50 if the training set does not have at least four miRNAs in common with the test set. Another issue is that the test set also has to be a part of the training sets, as all samples across the datasets are supposed to have the same types of probabilities. This is also naively solved by training and testing on the same dataset. The disadvantage here is that the sample is tested on a model that the sample did also train. Other methods would be to set the prediction probabilities to 0.50 in that case, with the disadvantage of information getting lost. A last method would be to use some kind of internal cross validation to find a probability, but the disadvantage is that the probability would be calculated in a different way, which would make the numbers less comparable across the datasets.

The plot and the options are similar to *Sample p-value PCA (single)*, with some differences. One difference is that several datasets can be plotted at once (as here the PCA values are comparable across test datasets). Thus, one selects a subset of datasets to be plotted instead of a single dataset. Selecting no datasets is an option, which means that the "show cases"-option is removed as it is redundant.

4.15.12 AUC PCA

One final question is whether there are any patterns in model performance in a certain test set. One way to check this is to do PCA on the AUC values from testing on a certain dataset while training on all other datasets. Here also, I only train on a dataset if it has at least four miRNAs in common with the test dataset. Thus, the same problem arose as in *Sample p-value PCA (combined)*, with the solution here that I sat 0.50 as the AUC if the datasets did not have four miRNAs in common. If the training and test datasets were the same I did a $\min(5, \#\text{cases}, \#\text{controls})$-fold cross validation and used the mean AUC.

The plot is similar to *Sample p-value PCA (combined)*, with the difference that the markers represent different datasets rather than samples. In addition, there is an option for choosing a color coding for the markers. The options are no color coding, color the datasets based on technology used (e.g. qRT-PCR, microarray or sequencing) or color the datasets based on the blood fraction where the miRNA-levels were measured (e.g. whole blood, serum or plasma). There is also an option that toggles whether there is a label with the corresponding dataset name near each marker. A screenshot is shown in Figure 4.13.

4.15.13 Pairwise Multi Plot

In some cases, it might be an advantage to be able to compare two datasets. There are three different ways to compare two datasets in this web application, namely: *PCA Two Datasets*, *Log Fold Change correlation* and *Pairwise Machine Learning*. This plot is a 2x2-matrix with those three plots as subplots. The options here are (1) the pair of datasets; (2) the machine learning algorithm (for *Pairwise Machine Learning*); (3) whether the PCA is based on the first, the second or both datasets; (4) which principal components are plotted; (5) marker size; (6) marker opacity (for *PCA Two Datasets*) and (7) whether to have different colors for cases and controls (for *PCA Two Datasets*).

4.16 Results from web application

The web application allowed for doing analysis with less effort than would otherwise be required. As such, there were also some results from using the web application.

4.16.1 Sample p-values PCA (single)

In general, the models performed poorly when trying to predict across studies, with mean AUCs close to 0.5. However, by doing PCA on the prediction probabilities one found that the cases and controls differed in how they were predicted by the different models. Here, I want to focus on one example, Asakura et al. [2020], but other datasets show similar patterns.

The plot is shown in Figure 4.12. Two immediate observations would be:

1. The cases and controls are separable in the PCA plot.

2. The loadings are going in very different directions.

The first point is somewhat surprising, because as the mean AUC is around 0.50 one would assume that the models would not be able to separate cases from controls, but the plot suggests that the first principal component (which explains the plurality of the variance in the prediction probability) is at least partially due to case-control characteristics. In other words, despite the models doing no better than chance at predicting cancer status, cancer status is a main contributor to the variance in the predictions of the models.

The second point explains how this can be the case. We see in the plot that giving a high prediction probability from e.g. Yao et al. [2019] or Reis et al. [2020] leads to a high value along PC 1. On the other hand, a high prediction probability from e.g. Leidinger et al. [2016] or Fehlmann et al. [2020] leads to

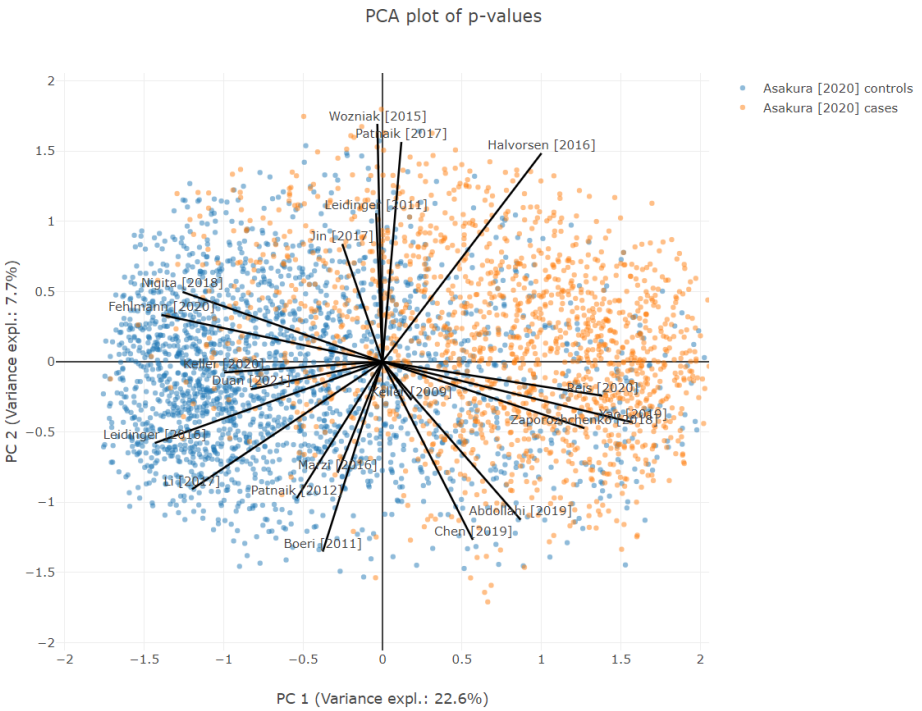


Figure 4.12: Screenshot of *Sample p-value PCA (single)* using Asakura et al. [2020] and logistic regression

a low value along PC 1. As cases are generally high along PC 1, this means that one would believe that Yao et al. [2019] and Reis et al. [2020] are predicting well, while Leidinger et al. [2016] and Fehlmann et al. [2020] are predicting very poorly as they give high probabilities to controls along PC 1. In aggregate these differences even out as there is little bias for loadings along PC 1, which lead to a mean AUC of 0.50. Thus the poor results in aggregate hide differences in how the models predict when trained on different datasets. One interesting result would be the correlations in prediction probabilities across studies.

The correlations are plotted in Table 4.38. There is little correlation between the predictions from the models trained on Reis et al. [2020] and Yao et al. [2019]. And while Reis et al. [2020] had a large positive correlation with case-characteristics, Yao et al. [2019] had a slight negative correlation. There was a slight negative correlation between the predictions of Leidinger et al. [2016] and Fehlmann et al. [2020], and while Leidinger et al. [2016] had a slight negative correlation with case status, Fehlmann et al. [2020] had a moderate positive correlation. Thus, even though cases and controls separate well along PC 1, it is important to remember that PC 1 only explains 22.6% of the variance in predictions. The results from the correlation table indeed show that other factors dominate as e.g. Fehlmann et al. [2020] performed moderately well in its predictions overall, even though it predicts in the wrong direction along PC 1. Thus, even though cases and controls separate well along PC 1, this effect almost disappears on an aggregate level as other sources of variance dominate.

Another interesting observation from the table is that of all the datasets that have been used for training, only Boeri et al. [2011], Jin et al. [2017], Keller et al. [2020] and Patnaik et al. [2012] did not have significant correlation with case status. It is interesting because as the mean AUC was around 0.50 one might think that predictions were in general independent of case-control characteristics, but this shows that that was not the case. Many datasets have a significant negative correlation with case-control characteristics, which means that there are features in the data that separate cases and controls to some degree, but those features lead the models to predict wrongly. It is difficult to say what is the reason for this. One hypothesis would be that some confounding variables are correlated with case characteristics in one dataset, but are correlated with control characteristics in other datasets. As the Asakura dataset was adjusted for sex and age, either the adjustment using a linear predictor was not sufficient, or the confounding variables are due to other characteristics.

4.16.2 AUC PCA

The AUC PCA-plot gave some insight into what created the difference between the datasets. A screenshot is shown in Figure 4.13, when coloring is based on

Table 4.38: The correlation between the prediction probabilities when training on the different datasets given by the column and row names and testing in Asakura et al. [2020] using logistic regression. In addition, the correlation with case status is calculated (i.e. 1 for cancer, 0 for control). Green or red color means significant (positive and negative, respectively) correction at a 0.05 level with Bonferroni correction.

	Abdollahi et al. [2019]	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]	Leidinger et al. [2011]	Leidinger et al. [2014]	Leidinger et al. [2016]	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]	Case status
Abdollahi et al. [2019]	1.00	−0.02	0.18	−0.06	−0.35	0.13	−0.16	0.04	0.09	−0.19	−0.07	−0.18	−0.14	−0.02	−0.31	−0.08	0.01	0.05	0.25	−0.15	0.38	0.23	−0.07	0.11	0.17
Bianchi et al. [2011]	−0.02	1.00	0.07	0.06	0.19	−0.13	−0.11	0.02	−0.26	−0.02	−0.02	0.13	0.14	−0.07	0.12	0.28	−0.11	0.01	−0.14	−0.23	−0.12	−0.00	0.05	−0.12	−0.18
Boeri et al. [2011]	0.18	0.07	1.00	−0.11	−0.25	0.03	−0.04	0.17	−0.07	−0.21	−0.11	−0.24	−0.10	−0.06	−0.17	0.09	−0.13	−0.07	0.16	−0.20	0.22	0.23	−0.12	0.20	0.04
Chen et al. [2019]	−0.06	0.06	−0.11	1.00	0.34	−0.17	−0.05	−0.09	−0.03	0.15	−0.01	0.24	0.21	0.06	0.21	0.08	0.01	0.02	−0.31	−0.04	−0.29	−0.21	0.05	−0.20	−0.22
Duan et al. [2021]	−0.35	0.19	−0.25	0.34	1.00	−0.28	0.10	−0.16	−0.22	0.35	0.04	0.53	0.45	0.01	0.63	0.24	0.01	−0.08	−0.67	0.05	−0.83	−0.45	0.10	−0.32	−0.55
Fehlmann et al. [2020]	0.13	−0.13	0.03	−0.17	−0.28	1.00	−0.07	0.00	0.08	−0.25	0.04	−0.28	−0.23	−0.13	−0.27	−0.28	0.17	−0.01	0.28	0.15	0.33	0.21	−0.08	0.09	0.41
Halvorsen et al. [2016]	−0.16	−0.11	−0.04	−0.05	0.10	−0.07	1.00	0.01	0.00	0.02	−0.06	−0.04	−0.01	−0.04	0.20	0.12	−0.04	−0.05	−0.10	0.25	−0.21	−0.16	−0.06	0.10	−0.07
Jin et al. [2017]	0.04	0.02	0.17	−0.09	−0.16	0.00	0.01	1.00	−0.12	−0.10	−0.10	−0.14	−0.03	−0.11	−0.12	−0.00	0.01	−0.11	0.12	−0.21	0.13	0.15	−0.02	0.13	−0.02
Keller et al. [2009]	0.09	−0.26	−0.07	−0.03	−0.22	0.08	0.00	−0.12	1.00	0.08	−0.01	0.03	−0.07	0.22	−0.15	−0.11	−0.10	0.18	0.15	0.08	0.15	0.00	0.06	0.08	0.11
Keller et al. [2014]	−0.19	−0.02	−0.21	0.15	0.35	−0.25	0.02	−0.10	0.08	1.00	0.01	0.47	0.26	0.16	0.29	0.11	−0.05	0.08	−0.28	0.01	−0.38	−0.30	0.17	−0.16	−0.32
Keller et al. [2020]	−0.07	−0.02	−0.11	−0.01	0.04	0.04	−0.06	−0.10	−0.01	0.01	1.00	−0.01	−0.03	0.04	0.05	−0.14	0.18	−0.01	−0.06	0.11	−0.02	0.02	0.07	−0.04	−0.05
Leidinger et al. [2011]	−0.18	0.13	−0.24	0.24	0.53	−0.28	−0.04	−0.14	0.03	0.47	−0.01	1.00	0.38	0.18	0.34	0.20	−0.05	0.09	−0.39	−0.09	−0.47	−0.36	0.27	−0.20	−0.39
Leidinger et al. [2014]	−0.14	0.14	−0.10	0.21	0.45	−0.23	−0.01	−0.03	−0.07	0.26	−0.03	0.38	1.00	0.13	0.33	0.21	−0.02	−0.00	−0.31	−0.16	−0.39	−0.28	0.18	−0.14	−0.36
Leidinger et al. [2016]	−0.02	−0.07	−0.06	0.06	0.01	−0.13	−0.04	−0.11	0.22	0.16	0.04	0.18	0.13	1.00	0.02	0.06	−0.02	0.07	−0.01	−0.10	−0.02	−0.11	0.17	0.03	−0.10
Li et al. [2017]	−0.31	0.12	−0.17	0.21	0.63	−0.27	0.20	−0.12	−0.15	0.29	0.05	0.34	0.33	0.02	1.00	0.23	−0.00	−0.03	−0.50	0.12	−0.65	−0.36	0.05	−0.21	−0.42
Marzi et al. [2016]	−0.08	0.28	0.09	0.08	0.24	−0.28	0.12	−0.00	−0.11	0.11	−0.14	0.20	0.21	0.06	0.23	1.00	−0.16	0.06	−0.22	−0.10	−0.30	−0.18	0.01	−0.01	−0.30
Nigita et al. [2018]	0.01	−0.11	−0.13	0.01	0.01	0.17	−0.04	0.01	−0.10	−0.05	0.18	−0.05	−0.02	−0.02	−0.00	−0.16	1.00	−0.04	0.01	0.11	0.06	0.05	0.05	−0.10	0.13
Patnaik et al. [2012]	0.05	0.01	−0.07	0.02	−0.08	−0.01	−0.05	−0.11	0.18	0.08	−0.01	0.09	−0.00	0.07	−0.03	0.06	−0.04	1.00	−0.01	0.03	0.04	−0.05	0.01	−0.02	−0.00
Patnaik et al. [2017]	0.25	−0.14	0.16	−0.31	−0.67	0.28	−0.10	0.12	0.15	−0.28	−0.06	−0.39	−0.31	−0.01	−0.50	−0.22	0.01	−0.01	1.00	−0.02	0.61	0.36	−0.08	0.25	0.46
Qu et al. [2017]	−0.15	−0.23	−0.20	−0.04	0.05	0.15	0.25	−0.21	0.08	0.01	0.11	−0.09	−0.16	−0.10	0.12	−0.10	0.11	0.03	−0.02	1.00	−0.13	−0.13	−0.12	0.01	0.18
Reis et al. [2020]	0.38	−0.12	0.22	−0.29	−0.83	0.33	−0.21	0.13	0.15	−0.38	−0.02	−0.47	−0.39	−0.02	−0.65	−0.30	0.06	0.04	0.61	−0.13	1.00	0.47	−0.08	0.23	0.55
Wozniak et al. [2015]	0.23	−0.00	0.23	−0.21	−0.45	0.21	−0.16	0.15	0.00	−0.30	0.02	−0.36	−0.28	−0.11	−0.36	−0.18	0.05	−0.05	0.36	−0.13	0.47	1.00	−0.15	0.10	0.29
Yao et al. [2019]	−0.07	0.05	−0.12	0.05	0.10	−0.08	−0.06	−0.02	0.06	0.17	0.07	0.27	0.18	0.17	0.05	0.01	0.05	0.01	−0.08	−0.12	−0.08	−0.15	1.00	−0.05	−0.15
Zaporozhchenko et al. [2018]	0.11	−0.12	0.20	−0.20	−0.32	0.09	0.10	0.13	0.08	−0.16	−0.04	−0.20	−0.14	0.03	−0.21	−0.01	−0.10	−0.02	0.25	0.01	0.23	0.10	−0.05	1.00	0.15
Case status	0.17	−0.18	0.04	−0.22	−0.55	0.41	−0.07	−0.02	0.11	−0.32	−0.05	−0.39	−0.36	−0.10	−0.42	−0.30	0.13	−0.00	0.46	0.18	0.55	0.29	−0.15	0.15	1.00

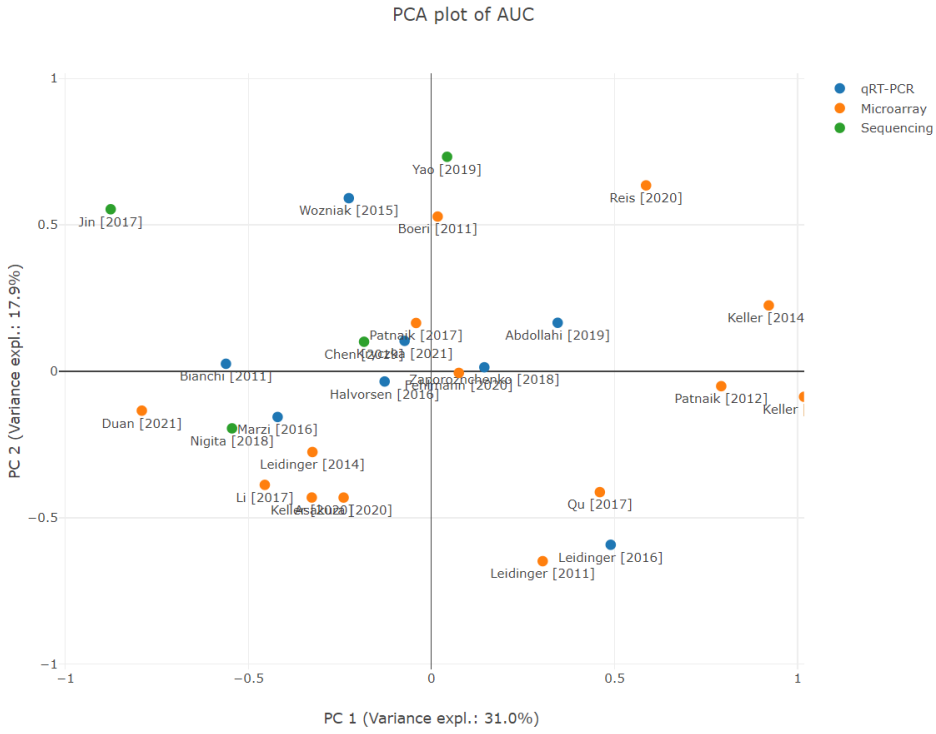


Figure 4.13: Screenshot of *AUC PCA* when coloring based on technology

technology. There is some clustering in the plot based on what technology is used for measuring miRNA-levels. Thus, there is more evidence that heterogeneity in technology is one reason for low consistency across datasets. However, to be sure, it would be useful to do a statistical test to see whether there actually are any differences. In particular, taking a t-test along the first principal component would show whether there actually are any significant differences along that axis.

The mean value along the first principal component was -0.389 for the sequencing datasets, -0.052 for the qRT-PCR datasets and 0.141 for the microarray datasets. The difference between qRT-PCR and microarrays was not significant ($p = 0.391$). Likewise, the difference between sequencing and microarrays ($p = 0.097$), and the difference between qRT-PCR and sequencing ($p = 0.174$) were not significant.

Chapter 5

Evaluation and Conclusion

This chapter contains the conclusions inferred from the results of this project. Parts of the conclusions are taken directly from Berg [2021] as there has been an overlap in questions that are discussed.

5.1 Evaluation

Overall, this project has been a success as the research questions have been thoroughly examined. The results have been mixed with some positive results and many null results. The overall conclusion is that if there are any patterns in differential case-control expression of miRNA, these patterns have to be relatively small compared to other sources of variance in the dataset. As a result, the patterns are hard to find, and results that are found in one dataset would not replicate in other datasets in general.

There are issues with data availability in research. It is important that data is available in order for third parties to be able to replicate the statistical findings in papers. Despite, out of 97 datasets requested by email I only received 2, which suggests that the data is not as available as it should be. This is especially worrisome as this field seems to have trouble with findings not replicating across different datasets, as found in this project. Furthermore, 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 collection is typically high in cancer research, and in some cases, the data collection 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 an 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.

This project also shows the value of trying to replicate findings, as the results show that there was a vast difference in the diagnostic accuracy you could get internally in a dataset and the accuracy you could get across datasets.

5.2 Discussion

Despite vast resources invested into cancer research globally, the field suffers from a low replication rate. Errington et al. [2021] looked at 50 experiments from 23 high-impact papers in 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. The correlation between the original and the new effect sizes was $r = 0.47$ using Spearman's r. For positive results, the median effect size was 85% smaller in the replication than in the original research. According to one survey, around 50% of cancer researchers have been unable to reproduce a published result [Mobley et al., 2013]. As a result, it is important to check what findings do replicate across studies and what findings do not. There have been many studies on circulating miRNA and lung cancer, but none has collected all available datasets for comparison.

There was an inconsistency between what was reported in the meta-analyses and what was found in the studies and the datasets in this project regarding what miRNAs had a consistent expression across studies. This suggests that the consistency found in other meta-analyses might not be universal, and some skepticism is warranted. On one hand, these meta-analyses had more studies in their analyses than this project had, which should point to one having more trust in their results than in the results of this project. On the other hand, there might be a publication bias where one is more likely to publish, note or report results that are consistent with the existing literature. In addition, in e.g. Zhong et al. [2021] only studies showing significant differential expression in a miRNA-sequence were noted, meaning that studies finding no differential expression were not taken into account. In this project, I tried to find whether there was a differential expression in the miRNAs in the datasets regardless of whether the authors had reported them as a result, which might paint a more representative picture.

Indeed, trying to find significant patterns in differential case-control expression was futile as the distribution of pairwise log fold correlations using Pearson's r had a mean close to zero. If there were linear effects of case-control expres-

sion, one would expect that it would result in a higher correlation coefficient. As there were many datasets that were tested against each other, one can assume that there indeed was no large correlation in log fold change in general. The small correlations that were indeed found seem to be partially due to covariance between different miRNAs that was unrelated to case-control status.

The datasets have different structures as a machine learning algorithm is able to distinguish between samples from different datasets quite well. This means that each study has some kind of footprint in the miRNA expressions that can be used for a machine learning algorithm to recognize a certain dataset. It would be hard to adjust for these effects that depend on which dataset is used, as effects that are linear on single miRNAs are already adjusted for. Effects thus have to be on a multi-miRNA level, but to find if such an effect is only in one dataset, one has to look at multiple datasets, and thus remove the independence between the datasets. As shown in section 4.12, removing the first two principal components is not sufficient to ensure consistency between datasets.

It is still an open question what causes the lack of reproducibility in the datasets. This project tried to check for some obvious answers like technology, blood fraction or cancer stage. However, none seemed to explain the lack of reproducibility completely, as there was still a lack of reproducibility when adjusting for these differences or when only considering different subgroups. There is a limitation here, as it was not possible to group on a finer level due to having few datasets.

5.2.1 Is there consistent differential expression?

As there was little to no consistency in the differential expression of the miRNA between cases and controls, one might ask whether there exists a consistent differential expression at all? On one hand, there was limited consistency between the datasets in this study, and one might wonder to what degree the positive results in this project are statistical artifacts rather than valid results, as there seemed not to be a general pattern in the positive results. It might seem like most of the evidence in this field is based on single studies that look at one specific dataset. As shown in this project, one might not extrapolate the results from these kinds of studies and conclude that the results are valid in general.

On the other hand, there often seemed to be a significant correlation between whether a machine learning model predicted that a certain sample was a cancer sample and whether it was an actual cancer sample (see Table 4.38). There also seems to be a clear pattern in the AUC PCA-plot in Figure 4.13, however, there were no significant differences along the first principal component. In addition, one should note that this is a project with only 26 datasets, while other meta-analyses like Zhong et al. [2021] have a larger sample size of studies they are

based on. However, these results have issues as explained above including that the degree of consistency in the differential expression of the miRNAs that was reported seems unlikely given the results in this project.

Anyhow, I think the results from this project make a good case to actually compare the data between studies, and not only the results, when doing meta-analyses in this field.

5.2.2 Limitation

There are limitations to this report. For once, there was heterogeneity in how miRNAs were measured, which is a source of noise in the data. It is plausible that a similar report as this that had available datasets that were homogenous in technology and blood fraction would indeed have more consistent patterns and more significant results. There were no adjustments for different handling of samples or different lung cancer types, which might explain some of the lack of consistency.

There was also a limitation in that few of the requested datasets were received, which means that the analysis is less thorough than it would otherwise have been.

Another limitation is on the machine learning. This project looked at some possible machine learning models, but there are more available that might lead to better results, but that is out of the scope of this project and would be future work.

5.3 Contributions

The main contribution of this project has been to collect all the available datasets on circulating miRNAs and lung cancer. I converted all the datasets into a common format, thus making it easier for other researchers to use them if they want to compare different datasets or show that their findings replicate across studies.

I have done a simplified meta-analysis where I looked at different meta-analyses and saw what miRNAs were reported to be consistently differentially expressed in the meta-analyses and saw whether these results replicated in the datasets that I collected.

Furthermore, I tried to compare the different datasets in different ways to find what patterns in case-control characteristics can be replicated across datasets. I have also tried different methods to find the effect technology, blood fraction and cancer stage had on the comparability of the datasets.

I have tried to group different datasets and use different machine learning algorithms to try to find subsets of datasets where there are patterns in case-control characteristics that a machine learning algorithm can find across studies.

Finally, I made a visualization tool for the data so that other researchers can explore the data easily.

5.4 Future Work

The main goal should be to try to find the reason for the lack of reproducibility, as it might lead to methods that would adjust for these problems, thus making circulating miRNAs a valid diagnostic marker for lung cancer that is not very sensitive to study design.

The process of data collection in this project can be built on by finding more datasets to add to the collection, or by using the already collected datasets to do analysis. Furthermore, the data visualization tool is available if someone wants a high-level exploration of the data without having to deal with the raw files.

There are methods that might lead to higher reproducibility. One way would be to try to manipulate the raw data differently in hope that it would result in more consistent case-control patterns. However, I would argue that it would be hard to find such a data manipulation. That is because the log fold change correlation is close to zero, and many types data manipulations would keep relative rank across the miRNAs, thus making it unlikely these data manipulations would lead to higher correlations. Of course, other types of data manipulations that do not preserve rank might work, e.g. by using principal components. Another way to try to get more consistency between datasets would be to collect datasets using the same technology and blood fraction, as there were some cases in this project where datasets had a higher consistency when using the same technology or the same blood fraction.

Another idea would be to try different machine learning tools to find consistent patterns. There are still many possible models to choose from that might find patterns in the datasets that are replicable, and one might try to explore other possible models to see if they give any improvement in diagnostic accuracy. However, given that it was hard to find consistent patterns here, it would be unlikely that other models work. It is also important to ensure that the model actually performs better when testing different models, rather than concluding based on some statistical coincidence. This is especially important as the results from this project suggest a low prior for having a model perform well across datasets and one would like to avoid the problem described by Ioannidis [2005].

Finally, I would request people working in this area to make sure that one's findings replicate across different datasets. This would ensure that the findings are general and not spurious, which this project shows is rare.

Bibliography

Abdollahi, A., Rahmati, S., Ghaderi, B., Sigari, N., Nikkhoo, B., Sharifi, K., and Abdi, M. (2019). A combined panel of circulating microRNA as a diagnostic tool for detection of the non-small cell lung cancer. *QJM: An International Journal of Medicine*, 112(10):779–785.

American Cancer Society (2019). What Is Lung Cancer? | Types of Lung Cancer.

American Cancer Society (2021). Non-small Cell Lung Cancer Treatment by Stage.

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.

Berg, O. F. B. (2021). Circulating miRNA and lung cancer: - an analysis of available data.

Bernstein, S. (2019). Lung Cancer Stages: Why and How Your Cancer Is Staged.

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.

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilità*. Seeber.

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

Chang, Y.-W., Hsieh, C.-J., Chang, K.-W., Ringgaard, M., and Lin, C.-J. (2010). Training and Testing Low-degree Polynomial Data Mappings via Linear SVM. *The Journal of Machine Learning Research*, 11:1471–1490.

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.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Donati, S., Ciuffi, S., and Brandi, M. L. (2019). Human Circulating miRNAs Real-time qRT-PCR-based Analysis: An Overview of Endogenous Reference Genes Used for Data Normalization. *International Journal of Molecular Sciences*, 20(18):4353.

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.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Science & Business Media.

Huang, H., Zhu, J., Lin, Y., Zhang, Z., Liu, J., Wang, C., Wu, H., and Zou, T. (2021). The potential diagnostic value of extracellular vesicle miRNA for human non-small cell lung cancer: A systematic review and meta-analysis. *Expert Review of Molecular Diagnostics*, 21(8):823–836.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, 2(8):e124.

Jiang, M., Li, X., Quan, X., Li, X., and Zhou, B. (2018). Clinically Correlated MicroRNAs in the Diagnosis of Non-Small Cell Lung Cancer: A Systematic Review and Meta-Analysis. *BioMed Research International*, 2018:e5930951.

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. (2016). 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.

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.

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.

Mobley, A., Linder, S. K., Braeuer, R., Ellis, L. M., and Zwelling, L. (2013). A Survey on Data Reproducibility in Cancer Research Provides Insights into Our Limited Ability to Translate Findings from the Laboratory to the Clinic. *PLoS ONE*, 8(5):e63221.

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.

Patle, A. and Chouhan, D. S. (2013). SVM kernel functions for classification. In *2013 International Conference on Advances in Technology and Engineering (ICATE)*, pages 1–9.

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.

Pratt, J. W. and Gibbons, J. D. (1981). One-Sample and Paired-Sample Inferences Based on Signed Ranks. In Pratt, J. W. and Gibbons, J. D., editors, *Concepts of Nonparametric Theory*, Springer Series in Statistics, pages 145–202. Springer, New York, NY.

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.

Sun, L., Yu, Y., Niu, B., and Wang, D. (2020). Red Blood Cells as Potential Repositories of MicroRNAs in the Circulatory System. *Frontiers in Genetics*, 11.

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.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83.

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.

Yi, M., Liao, Z., Deng, L., Xu, L., Tan, Y., Liu, K., Chen, Z., and Zhang, Y. (2021). High diagnostic value of miRNAs for NSCLC: Quantitative analysis for both single and combined miRNAs in lung cancer. *Annals of Medicine*.

Yu, H., Guan, Z., Cuk, K., Zhang, Y., and Brenner, H. (2019). Circulating MicroRNA Biomarkers for Lung Cancer Detection in East Asian Populations. *Cancers*, 11(3):415.

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 Article based on project (preprint)

RESEARCH

On the reproducibility of circulating miRNA in lung cancer diagnosis

Ole Fredrik B. Berg* and Pål Sætrom

*Correspondence:
ole.fredrik.berg@gmail.com
Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway
Full list of author information is available at the end of the article

Abstract

Background: There have been many studies on circulating miRNA and lung cancer, but few studies have compared datasets from different studies. In this paper, we collected all available datasets in the field in order to compare them and try to find whether patterns in case-control characteristics could replicate across different datasets.

Results: A machine learning model could separate cases and controls well internally in one dataset. However, in general there was little consistency in case-control patterns across datasets.

Keywords: miRNA; lung cancer; machine learning; NSCLC; circulating miRNA; serum; plasma; whole blood

Background

Lung cancer is the second most common type of cancer worldwide, and the type of cancer with the highest total mortality worldwide, causing about 1.8 million deaths per year [1]. 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 [2]. Furthermore, about 90% of lung cancer deaths in men, and 79% of lung cancer deaths in women are caused by smoking [3].

There are two main types of lung cancer, Small Cell Lung Cancers (SCLC) and Non-Small Cell Lung Cancers (NSCLC) [4]. Of lung cancer cases, about 80-85% are NSCLC, whilst 10-15% of the cases are SCLC, and a few percent are minor types of lung cancer [5]. NSCLC cancers tend to grow slower than the SCLC cancer types, and thus SCLC has usually already spread when it is diagnosed [5]. The NSCLC has three major subtypes: adenocarcinoma (30-40% of NSCLC cases), squamous cell (30%) and large-cell undifferentiated carcinoma (10-15%) [4]. The treatment and prognosis for the different NSCLC subtypes are similar [5].

The main advantage of 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 [6]. In contrast, later stages might require chemotherapy, radiation therapy or immunotherapy, but as the cancer has spread widely, this cure will likely not remove the cancer completely [6].

MicroRNAs (miRNAs) are short sequences of RNA, about 22 nucleotides each, that regulate 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 [7, 8, 9, 10].

The overall roles of miRNAs in relation to lung cancer are not fully understood [11]. MicroRNAs are thought to function both as tumor suppressor genes and as oncogenes, and tumor miRNA expression profiles can distinguish tumors from normal tissue, distinguish tumor subtypes, and predict survival [12]. Moreover, multiple studies report differential expression of circulating miRNA-sequences in cancer patients compared to healthy controls, which suggests expression of circulating miRNAs is a promising method for diagnosing lung cancer [11].

Statement of the problem

We want to collect all available datasets on circulating miRNA and lung cancer in order to be able to do machine learning on a larger combined dataset to see whether that leads to better diagnostic results. We also want to see to what degree machine learning algorithms can generalize across different datasets.

Methods

The methods section includes a description of the literature review, how machine learning was done on single datasets and how training was done when using two datasets as the training set. This paper contains a subcollection of results from [13] and [14].

Structured Literature Review Protocol

The point of the literature search was to find studies relevant to circulating miRNA and lung cancer. The main search engine used was PubMed, which is a commonly used search engine for medical literature. 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 searched databases that have public gene expression data, as described in Table 1.

The inclusion criteria were based on what datasets I thought were 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 [15], only the circulating miRNA ones were 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 reasonable to only use circulating miRNA data.

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

The controls in some of the studies are not healthy, but suffer from other kinds of lung diseases. Other studies 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 find what miRNA changes are due to the lung cancer. On the other hand, people who are getting checked for lung cancer often have lung issues, which is the reason for their checkup, so distinguishing lung cancer from other illnesses is important.

Some studies were excluded as they did not have a control group like [16].

- 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.

- Meta-analyses were used as a 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.

Machine learning on single datasets

We will train four different models on each dataset using logistic regression, SVM, random forest and XGBoost. The models will be tested using AUC, and the AUC will be calculated using cross validation where the dataset is split into $c = \min(5, \#\text{Cases}, \#\text{Controls})$ equal parts and for each of the c parts, there will be a round where the model is trained on the $c - 1$ other parts of the dataset and tested on the last part. The resulting AUC will be the average over the c rounds.

Training on two datasets

We will train different machine learning models on two datasets and try to predict on a third dataset, and then compare the results to the results that are found by training the model on only one of the datasets. The results will only be considered if the three datasets have at least 10 miRNA-sequences in common, to ensure the datasets are similar enough. The samples will be weighted so that the sum of weights in each dataset is the same, and the weights of all samples in the same dataset are the same.

Results

Studies included

Current literature is replete with studies investigating the potential of circulating miRNAs for lung cancer diagnosis, but for such studies to be useful for machine learning analyses and replication purposes, the data from individual miRNAs and individuals should be available. To identify a large and unbiased set of studies that had investigated and reported the blood expression profiles of multiple miRNAs in multiple individuals, including both lung cancer patients and controls, we did a structured literature review (see the section on literature review in the methodology).

The review identified 123 studies. However, most datasets that were requested by email were not received. The 26 studies whose datasets that were either received or were publicly available are: [17], [18], [19], [20], [21][1], [22], [23], [24], [25], [26],

[1] [21] 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

[27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41] and [42]. A basic overview of the different studies is found in Table 2.

Machine learning on single datasets

One important question is whether more advanced machine learning algorithms are better at diagnosing lung cancer based on miRNA-levels. Therefore, I have chosen to do machine learning on single datasets. As seen in [13], the results when doing machine learning across datasets were mostly poor. Furthermore, if the connection between miRNA expression and lung cancer is very sensitive to study characteristics, machine learning across different datasets might not be the best idea, compared to training on data where the characteristics are known to be the same. I have chosen four different types of machine learning algorithms to test.

- Logistic regression: Is natural to use as a baseline model to compare against, as it has been used in many of the studies that are included in this project
- Support vector machine: If the data is nearly linearly separable, this will find such a separation.
- Random forest: It is a powerful algorithm that is able to generalize well also on small datasets, as it is an ensemble method.
- XGBoost: Has had the success in tabled data with limited samples in Kaggle competitions [43], and is thus a natural algorithm to test.

The results from machine learning on single datasets are shown in Table 3. The results are from using cross validation on the datasets with the given machine learning algorithms. A more detailed explanation is found under methodology. Random forest performed best while XGBoost performed worst. One question is whether these differences are statistically significant or not. Therefore I performed t-tests on the differences in AUC values, which showed that none of the differences were significant. Thus one cannot say that one algorithm performed better than another in general.

Baseline miRNA-sequence

One important question is what miRNA-sequences would be most successful in diagnosing lung cancer. This has not only clinical relevance, but is also important to note as a machine learning model would be more powerful than using one single miRNA-sequence. There will always be additional costs associated with measuring more miRNA-sequences, and therefore I need to show that a machine learning model will perform better than a model based on a single miRNA-sequence.

There are two main types of methods possible for finding such miRNA-sequences, each with some pros and cons:

1 Look at meta-analyses for finding miRNA-sequences that are found to diagnose lung cancer well across studies.
  Pros:
  - The miRNA-sequences found would be based on more data, and thus they are likely better.
  - The miRNA-sequences are nearly[2] independent of the datasets used in this project and are therefore mostly unbiased.

[2] After all, the meta-analyses might be based on some datasets used in this project.

Cons:

- The miRNA-sequences that are reported in these meta-analyses are often not in many of the datasets used in this project.

2 Look at the datasets used in this project to find miRNA-sequences that can separate cases and controls across the different datasets.

Pros:

- It is easier to limit the search to miRNA-sequences that are found in many of the datasets in this project.

Cons:

- The miRNA-sequences are biased, as the baseline is the ability of these miRNA-sequences to diagnose cancer on the datasets, but the miRNA-sequences were chosen because they diagnosed well on said datasets.

I want to try a hybrid strategy, in order to mitigate the cons of each method. That is, I want to try to find an intersection between microRNA-sequences that have been found in meta-analyses to be consistently good at diagnosing lung cancer and microRNA-sequences that separate well in the studies used in this project.

There are several ways to measure to which degree a miRNA-sequence can be used to separate cases from controls. One possibility would be to use the t-statistic. The advantage of the t-statistic is that it has a known distribution (given the null hypothesis), and thus one could get to know whether a difference is plausibly a result of chance or not. The disadvantage of the t-statistic is that it does not only measure to which degree the miRNA-sequence separates well in the dataset, but also the statistical power of each dataset. Therefore, large datasets would be given more weight, and the value could hide to what degree the miRNA-sequence diagnoses correctly in the dataset.

Another alternative is to use Cohen's d. The advantage of Cohen's d is that it tells to what degree cases and controls are separated independently of the number of samples in the dataset. The disadvantage of Cohen's d is that it does not consider the statistical power at all, and thus one might expect many spurious results when using Cohen's d. A final statistic is to use AUC. The advantages and disadvantages are similar to Cohen's d, with the difference that AUC has the advantage that it is the metric that the results will be measured against in the end. However, Cohen's d has the advantage that it also looks at the size of the difference in expression, and not just whether there is a separation like AUC does.

After a consideration of the different statistics, I found that Cohen's d and AUC would be the most appropriate statistic for this purpose, as the t-statistic would give too much power to the large datasets (which might not be representative at all), and not tell the actual degree of separation.

Meta-analyses

Meta-analyses gave an overview of the possible miRNA-sequences that can be used as baselines in this project [44, 45, 46, 47], where [44] was the most thorough of the meta-analyses. These meta-analyses suggest that the miRNA-sequences that have been shown to be able to diagnose lung cancer in most studies are miR-21 and miR-210, with [44] suggesting that miR-182, miR-155 and miR-17 are in third, fourth and fifth place, respectively. All of these miRNA-sequences were reported

to be up-regulated in cases compared to controls. However, these results were not representative of the studies used in this project.

[44] found that of all studies that they went through miR-21 was significantly up-regulated in cases in 48 studies, and down-regulated in two studies. However, among the studies used in this project, [36, 25, 32, 23] all reported that miR-21 was down-regulated in cases compared to controls, which suggests that miR-21 might not be as good of a biomarker for lung cancer as the meta-analyses suggest. An overview of the reported up- and down-regulation of the aforementioned miRNA-sequences in the studies in this project is shown in Table 4.

Table 4 shows that none of the five miRNA-sequences were consistently up-regulated. However, miR-17 was consistently down-regulated in the sample, which contrasts with [44] which reported that miR-17 had been up-regulated in 7 studies and down-regulated in one study, if one only looks at the studies using circulating miRNA. Everything considered, this points to very inconsistent results across datasets, which suggests that there might be little consistency, and hard to replicate results.

Using datasets

The meta-analyses gave some candidate miRNA-sequences that can be used as baselines in this project, namely miR-21, miR-210, miR-182, miR-155 and miR-17. The Cohen's d and AUC of the miRNA-sequences in the different datasets are shown in Table 5 and Table 6 respectively.

Interestingly, the average Cohen's d of four of the miRNA-sequences was negative, even though [44] found that they were consistently up-regulated in cancer compared to healthy controls, which again suggests that these miRNA-sequences are not as good biomarkers for cancer as [44] suggests. Overall miR-210 was the only one that my datasets and the meta-analyses agree on being up-regulated, which is why we chose that miRNA as our baseline.

Training on two datasets

One of the goals of this project is to find whether combining multiple datasets will result in better diagnostic accuracy than using a single dataset. The result of training on one dataset and predicting on another dataset was done in the exploratory phase of this project with subpar results. However, it is possible that training on multiple datasets will help the machine learning algorithm to find case-control patterns that transcend the patterns that are found internally in one dataset, leading to better generalizability. In this experiment, we will find sets of three datasets, where one of them is a test set, and the two others are training sets. We will compare results when only training on one of the training sets to when training on both the data sets. More details are in the methodology.

Logistic Regression

The first model we will try is logistic regression as it is a basic classification model, and it is often used in the studies that try to predict cancer based on miRNA, it therefore serves well as a baseline. The model will be trained on the miRNA-sequences that all the three datasets have in common.

When training on just one of the datasets the mean AUC was 0.501 and the standard deviation was 0.168. When training on both datasets, the mean AUC was 0.508 and the standard deviation was 0.169. This is worse than the baseline miR-210, which had a mean AUC of 0.551 (see Table 6).

XGBoost

It is plausible that a model like XGBoost will perform better on the datasets, as it has methods for handling missing data, and it can handle non-linear relationships in the data. In addition, it is a boosting algorithm, which usually performs well when data is sparse, as in this case. Here, we will make use of the way XGBoost handles missing data and therefore train the model on all the miRNA-sequences that the two training datasets have in common.

The mean and standard deviation in AUC values when training on one dataset were 0.504 and 0.162 respectively. The mean and standard deviation when training on two datasets were 0.505 and 0.172 respectively. The results suggest that combining two datasets have little to no effect. Furthermore, the results were very similar to the ones achieved with logistic regression, which suggests that the problem is not the model.

Stratification of the datasets

There are several possibilities as to why the datasets are incompatible. One possibility is that some factors like what technology was used for measuring miRNA-levels play a role. There are other factors as well that differ between the datasets, like cancer stage and what blood fraction was measured (plasma, serum, whole blood, etc.). If these factors play a role one would expect to see more consistency in datasets that are similar in these characteristics. One way to test this hypothesis is to stratify the datasets based on these characteristics, and see if one sees a larger consistency between the datasets when the datasets are stratified in this way.

Training and testing on pairs of datasets, in-group vs. out-group

Here we will use pairs of datasets and train a model on one of the datasets and test on the other dataset, only that the AUC will be compared when the datasets have the same characteristics to when they have different characteristics. E.g., we will compare the AUC when two datasets are using qRT-PCR to when one is using qRT-PCR and the other study is using a different technology. We will do this stratification for technology and for type of blood fraction. Here we will use logistic regression and only do pairs of datasets that have at least 10 miRNA-sequences in common.

Stratifying by technology: The results when training on one dataset and testing on another dataset when stratifying using technology are shown in Table 7. The in-group is when both datasets use the given technology, and the out-group is when only one of the datasets uses the given technology. As the table shows, the AUC was generally somewhat better in in-group than out-groups. However, the improvement in AUC was only significant for microarrays. Still, the category "microarray" is hiding heterogeneity, as the microarrays in the studies varied a lot.

To check the hypothesis that the problems are due to heterogeneity in the microarray-technology, we wanted to do an experiment with a finer stratification of the microarray technology. Of the microarray-technologies that have been used multiple times, there were three that used *Exiqon microarrays* ([22, 36, 37]), three that used *Agilent microarrays* ([23, 38, 33]), three that used *Geniom microarrays* ([26, 30, 27]) and two that used *SurePrint microarrays* ([31, 28]). This experiment will only consider pair of studies where both use microarrays. The in-group here is when the pair of studies have the same type of microarray, and the out-group is when they use different types of microarrays. The results were that the in-group had a mean AUC of 0.612 while the out-group had a mean of 0.518. The difference was not significant using a t-test ($p = 0.056$). It might be that this is due to the low sample size, but even in the in-group, the consistency is relatively low compared to the internal consistency in the datasets found in Table 3.

Stratifying by blood fraction: The results when training on one dataset and testing on another dataset when stratifying using blood fraction are shown in Table 8. The in-group is when both datasets measure the given blood fraction, and the out-group is when only one of the datasets uses the given blood fraction. In contrast to when stratifying by technology, it seems that there is no use in stratifying by blood fraction. None of the changes in AUC are significant, and one of the changes is even negative. It might suggest that technology contributes to more variance in the resulting data than what blood fraction does.

Stratifying by cancer stage: Cancer stage may be a covariate that hinders the replicability of the datasets. To check this hypothesis, I will do an analysis where I only use the datasets where samples are labeled, and compare the results when only using the early stages to when only using the late stages. If there is higher consistency in the late stages, it would suggest that some of the lack of replicability is due to cancer stages. The result from training on one dataset and testing on another dataset, when only using late stage cancer was a mean AUC of 0.528 with a standard deviation of 0.187. Using only early stage cancer gave a mean AUC of 0.460 with a standard deviation of 0.140. There is no significant difference between the AUCs in the two cases given by a t-test ($p = 0.204$), and both mean AUCs are close to 0.50, which suggests that stage does not explain the low AUC scores in the previous results.

Combining all except one

Another attempt will be to take all datasets with a certain characteristic, like technology or blood fraction, and then train on all datasets except one that will be used for testing, and using AUC as the metric. For checking whether the AUC values are better than chance levels we took a one-sided hypothesis of AUC > 0.50 using a t-test. We will use the union of the miRNAs in the datasets in each category to train on. To ensure that missing values will not be a problem, we will use XGBoost as the model as it handles missing values by default. We will also try to do this using the datasets where cancer stage is labeled, and try both using only early cancer samples and using only late cancer samples.

Stratifying by technology: The results from stratifying by technology are shown in Table 9. Sequencing is an outlier where the AUC was better than the other categories. Notably, an AUC of 0.625 is higher than any of the other AUCs achieved so far when testing on a different dataset than testing on, but it might be due to chance as the AUC is not significantly higher than 0.50 when adjusting for multiple testing. It does not seem like technology is the main reason for the low consistency between the datasets.

Also here I want to see whether stratifying by subtypes of microarrays will be beneficial. The subcategories are small, with the largest ones having three datasets, meaning that training will be done on maximally two datasets. The resulting mean AUC was 0.567 and the resulting standard deviation was 0.282, which was not significantly better than 0.50 ($p = 0.225$). This suggests that neither here heterogeneity in the microarray-technology was the reason for the poor results for the microarrays. Even an AUC of 0.567 is much lower than the internal consistency found in Table 3.

Stratifying by blood fraction: The results from stratifying by blood fraction are shown in Table 10. None of the AUCs were significantly larger than 0.50 when adjusted for multiple testing. This suggests that the lack of consistency is not due to blood fraction either.

Distribution of AUC values: In the subsections above, only summary statistics were reported. However, mean and variance can hide a lot of information about the distribution, e.g. whether the distribution is unimodal or bimodal. As t-values have been used to check for statistical significance, there has been an implicit assumption that AUC values have been approximately normally distributed. I have plotted a histogram and a Q-Q plot combining all the AUC values from the different categories (Figure 1), as it is too few values to do inference based on any of these categories alone. I am not including the results from stratifying by cancer stage here as those values do not use full datasets, and are thus less comparable. The Q-Q plot shows that the distribution of AUC values follows the normal distribution quite nicely, except for a slight deviation in the tails of the distribution. Thus, the normality assumption seems to hold.

Stratifying by cancer stage: Here all datasets with labeled cancer stages are used. The training on all datasets except one using only early stage cancer samples and controls with leave-one-out cross validation results in a mean AUC of 0.523 with a standard deviation of 0.171. A t-test shows that this is not significantly better than 0.50 ($p = 0.738$). The same experiment using late stage cancer results in a mean AUC of 0.509 with a standard deviation of 0.194. Neither this is significantly better than 0.50 ($p = 0.904$). As the mean AUCs, both when only using early stage cancer and only using late stage cancer, were only slightly higher than 0.50 and the differences were not significant, it seems like there is no improvement in AUC by stratifying by stage.

Conclusions

In short, the results from this project suggest that there is low consistency between different datasets on circulating miRNA and lung cancer.

Evaluation

Overall, this project has been a success as we have done what we intended in the statement of this problem. The results have been mixed with some positive results, but mostly null results. The overall conclusion is that if there is any pattern in differential case-control expression of miRNA, this pattern has to be relatively small compared to other sources of variance in the dataset. As a result, the patterns are hard to find, and results that are found in one dataset do not replicate in other datasets in general.

There are issues with data availability in research. It is important that data is available in order for third parties to be able to replicate the statistical findings in papers. Despite, out of 97 datasets requested by email, I only received 2, which suggests that the data is not as available as it should be. This is especially worrisome as this field seems to have trouble with findings not replicating across different datasets, as found in this project. Furthermore, [48] 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 an 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.

This project also shows the value of trying to replicate findings, as the results show that there was a vast difference in the diagnostic accuracy you could get internally in a dataset and the accuracy you could get across datasets.

Discussion

Despite vast resources invested into cancer research globally, the field suffers from a low replication rate. [49] looked at 50 experiments from 23 high-impact papers in 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. The correlation between the original and the new effect sizes was $r = 0.47$ using Spearman's r. For positive results, the median effect size was 85% smaller in the replication than in the original research. Around 50% of cancer researchers have been unable to reproduce a published result, according to one survey [50]. As a result, it is important to check what findings do replicate across studies and what findings do not. There have been many studies on circulating miRNA and lung cancer, but none has collected all available datasets for comparison.

There was an inconsistency between what was reported in the meta-analyses and what was found in the studies and the datasets in this project, regarding what miRNAs had a consistent expression across studies. This suggests that the consistency found in other meta-analyses might not be universal, and some skepticism is warranted. On one hand, these meta-analyses had more studies in their analyses

than this project had, which should point to one having more trust in their results than in the results of this project. On the other hand, there might be a publication bias where one is more likely to publish, note or report results that are consistent with the existing literature. In addition, in e.g. [44] only studies showing significant differential expression in a miRNA-sequence were noted, meaning that studies finding no differential expression were not taken into account. In this project, I tried to find whether there was differential expression of these miRNAs in the datasets regardless of whether the authors had reported them as a result, which might paint a more representative picture.

It is still an open question what causes the lack of reproducibility in the datasets. This project tried to check for some obvious answers like technology, blood fraction or cancer stage. However, none seemed to explain the lack of reproducibility completely, as there was still a lack of reproducibility when adjusting for these differences or when only considering different subgroups. There is a limitation here, as it was not possible to group on a finer level, due to having few datasets and a finer grouping would have too low statistical power when looking at the accuracy inside a subgroup.

Limitations

There are limitations to this report. For once, there was heterogeneity in how miRNAs were measured, which is a source of noise in the data. It is plausible that a similar report as this that had available datasets that were homogenous in technology and blood fraction would indeed have higher consistency patterns and more significant results. There were no adjustments for different handling of samples or different lung cancer types, which might explain some of the lack of consistency. There was also a limitation in that few of the requested datasets were received, which means that the analysis is less thorough than it would otherwise have been.

Another limitation is on the machine learning. This project looked at some possible machine learning models, but there are more available that might lead to better results, but that is out of the scope of this project and would be future work.

Contributions

The main contribution of this project has been to collect all the available datasets on circulating miRNAs and lung cancer and doing an analysis where all datasets are used and compared to each other.

We have done a simplified meta-analysis where we looked at different meta-analyses and saw what miRNAs were reported to be consistently differentially expressed in the meta-analyses and saw whether these results replicated in the datasets that we collected.

We have tried to group different datasets and use different machine learning algorithms to try to find subsets of datasets where the pattern in case-control characteristics is such that a machine learning algorithm can find it across studies.

Future Work

The main goal should be to try to find the reason for the lack of reproducibility, as it might lead to methods that would adjust for these problems, thus making circulating

miRNAs a valid diagnostic marker for lung cancer that is not very sensitive to study design.

The process of data collection in this project can be built on by finding more datasets to add to the collection, or by using the already collected datasets to do analysis.

There are methods that might lead to higher reproducibility. One way would be to try to manipulate the raw data differently in hope that it would result in more consistent case-control patterns. However, I would argue that such data manipulation probably does not exist. That is because the log fold change correlation is close to zero, and most reasonable data manipulations would keep relative rank across the miRNAs, thus making it unlikely that another data manipulation would lead to higher correlations. Another way to try to get more consistency between datasets would be to collect datasets using the same technology and blood fraction, as there are several cases in this project where datasets had a higher consistency when using the same technology or the same blood fraction.

Another idea would be to try different machine learning tools to find consistent patterns. There are still many possible models to choose from that might find patterns in the datasets that are replicable, and one might try to explore other possible models to see if they give any improvement in diagnostic accuracy.

Finally, I would request people working in this area to make sure that one's findings replicate across different datasets. This would ensure that the findings are general and not spurious, which this project shows is rare.

Acknowledgements
Text for this section...

Abbreviations
Text for this section...

Availability of data and materials
All code used in this project is available at https://github.com/OleFredrik1/masterthesis. The normalized miRNA data is available at https://doi.org/10.5281/zenodo.6568981.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
We consent to publication of this paper.

Authors' contributions
Berg did virtually everything, but Sætrom needs some citations for career purposes.

Author details
Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway.

References

1. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: 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 (2021). doi:10.3322/caac.21660. Accessed 2021-09-22
2. Walser, T., Cui, X., Yanagawa, J., Lee, J.M., Heinrich, E., Lee, G., Sharma, S., Dubinett, S.M.: Smoking and Lung Cancer. Proceedings of the American Thoracic Society **5**(8), 811–815 (2008). doi:10.1513/pats.200809-100TH. Accessed 2021-09-22
3. Shopland, D.R., Eyre, H.J., Peachacek, T.F.: 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 (1991). doi:10.1093/jnci/83.16.1142. Accessed 2021-09-22
4. Ciupka, B.: Small Cell Lung Cancer vs. Non-small Cell Lung Cancer: What's the Difference? (2020). https://www.nfcr.org/blog/small-cell-lung-cancer-vs-non-small-cell-lung-cancer-whats-the-difference/ Accessed 2021-09-23

5. American Cancer Society: What Is Lung Cancer? | Types of Lung Cancer (2019). https://www.cancer.org/cancer/lung-cancer/about/what-is.html Accessed 2021-09-23
6. American Cancer Society: Non-Small Cell Lung Cancer Treatment by Stage (2021). https://www.cancer.org/cancer/lung-cancer/treating-non-small-cell/by-stage.html Accessed 2021-09-23
7. Correia, C.N., Nalpas, N.C., McLoughlin, K.E., Browne, J.A., Gordon, S.V., MacHugh, D.E., Shaughnessy, R.G.: Circulating microRNAs as Potential Biomarkers of Infectious Disease. Frontiers in Immunology **8**, 118 (2017). doi:10.3389/fimmu.2017.00118. Accessed 2021-09-23
8. Kosaka, N., Iguchi, H., Ochiya, T.: Circulating microRNA in body fluid: A new potential biomarker for cancer diagnosis and prognosis. Cancer Science **101**(10), 2087–2092 (2010). doi:10.1111/j.1349-7006.2010.01650.x
9. Geekiyanage, H., Jicha, G.A., Nelson, P.T., Chan, C.: Blood serum miRNA: Non-invasive biomarkers for Alzheimer's disease. Experimental Neurology **235**(2), 491–496 (2012). doi:10.1016/j.expneurol.2011.11.026. Accessed 2021-09-23
10. van den Berg, M.M.J., Krauskopf, J., Ramaekers, J.G., Kleinjans, J.C.S., Prickaerts, J., Briedé, J.J.: Circulating microRNAs as potential biomarkers for psychiatric and neurodegenerative disorders. Progress in Neurobiology **185**, 101732 (2020). doi:10.1016/j.pneurobio.2019.101732. Accessed 2021-09-23
11. Uddin, A., Chakraborty, S.: Role of miRNAs in lung cancer. Journal of Cellular Physiology (2018). doi:10.1002/jcp.26607
12. Lynam-Lennon, N., Maher, S.G., Reynolds, J.V.: The roles of microRNA in cancer and apoptosis. Biological Reviews of the Cambridge Philosophical Society **84**(1), 55–71 (2009). doi:10.1111/j.1469-185X.2008.00061.x
13. Berg, O.F.B.: Circulating miRNA and Lung Cancer: - An Analysis of Available Data. NTNU Open (2021)
14. Berg, O.F.B.: Circulating miRNA and Lung Cancer: - a More Comprehensive Analysis of Available Data. NTNU Open (2022)
15. Petriella, D., De Summa, S., Lacalamita, R., Galetta, D., Catino, A., Logroscino, A.F., Palumbo, O., Carella, M., Zito, F.A., Simone, G., Tommasi, S.: miRNA profiling in serum and tissue samples to assess noninvasive biomarkers for NSCLC clinical outcome. Tumor Biology **37**(4), 5503–5513 (2016). doi:10.1007/s13277-015-4391-1. Accessed 2021-09-24
16. Mitchell, K.A., Zingone, A., Toulabi, L., Boeckelman, J., Ryan, B.M.: Comparative Transcriptome Profiling Reveals Coding and Noncoding RNA Differences in NSCLC from African Americans and European Americans. Clinical Cancer Research **23**(23), 7412–7425 (2017). Chap. Biology of Human Tumors. doi:10.1158/1078-0432.CCR-17-0527. Accessed 2021-10-05
17. Abdollahi, A., Rahmati, S., Ghaderi, B., Sigari, N., Nikkhoo, B., Sharifi, K., Abdi, M.: A combined panel of circulating microRNA as a diagnostic tool for detection of the non-small cell lung cancer. QJM: An International Journal of Medicine **112**(10), 779–785 (2019). doi:10.1093/qjmed/hcz158. Accessed 2022-02-05
18. 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., Ochiya, T.: A miRNA-based diagnostic model predicts resectable lung cancer in humans with high accuracy. Communications Biology **3**(1), 1–9 (2020). doi:10.1038/s42003-020-0863-y. Accessed 2021-10-24
19. Bianchi, F., Nicassio, F., Marzi, M., Belloni, E., Dall'Olio, V., Bernard, L., Pelosi, G., Maisonneuve, P., Veronesi, G., Di Fiore, P.P.: A serum circulating miRNA diagnostic test to identify asymptomatic high-risk individuals with early stage lung cancer. EMBO Molecular Medicine **3**(8), 495–503 (2011). doi:10.1002/emmm.201100154. Accessed 2021-10-24
20. Boeri, M., Verri, C., Conte, D., Roz, L., Modena, P., Facchinetti, F., Calabrò, E., Croce, C.M., Pastorino, U., Sozzi, G.: 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 (2011). Chap. Biological Sciences. doi:10.1073/pnas.1100048108. Accessed 2021-10-24
21. Chen, X., Jin, Y., Feng, Y.: Evaluation of Plasma Extracellular Vesicle MicroRNA Signatures for Lung Adenocarcinoma and Granuloma With Monte-Carlo Feature Selection Method. Frontiers in Genetics **10**, 367 (2019). doi:10.3389/fgene.2019.00367. Accessed 2021-10-24
22. Duan, X., Qiao, S., Li, D., Li, S., Zheng, Z., Wang, Q., Zhu, X.: Circulating miRNAs in Serum as Biomarkers for Early Diagnosis of Non-small Cell Lung Cancer. Frontiers in Genetics **12**, 987 (2021). doi:10.3389/fgene.2021.673926. Accessed 2021-10-04
23. 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., Keller, A.: Evaluating the Use of Circulating MicroRNA Profiles for Lung Cancer Detection in Symptomatic Patients. JAMA oncology **6**(5), 714–723 (2020). doi:10.1001/jamaoncol.2020.0001
24. 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., Helland, Å.: A unique set of 6 circulating microRNAs for early detection of non-small cell lung cancer. Oncotarget **7**(24), 37250–37259 (2016). doi:10.18632/oncotarget.9363. Accessed 2021-10-24
25. 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., Xie, C.: 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 (2017). Chap. Biology of Human Tumors. doi:10.1158/1078-0432.CCR-17-0577. Accessed 2021-10-24
26. Keller, A., Leidinger, P., Borries, A., Wendschlag, A., Wucherpfennig, F., Scheffler, M., Huwer, H., Lenhof,

H.-P., Meese, E.: miRNAs in lung cancer - Studying complex fingerprints in patient's blood cells by microarray experiments. BMC Cancer **9**(1), 353 (2009). doi:10.1186/1471-2407-9-353. Accessed 2021-10-24

27. 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., Meese, E.: miRNAs can be generally associated with human pathologies as exemplified for miR-144*. BMC Medicine **12**(1), 224 (2014). doi:10.1186/s12916-014-0224-0. Accessed 2021-10-24
28. Keller, A., Fehlmann, T., Backes, C., Kern, F., Gislefoss, R., Langseth, H., Rounge, T.B., Ludwig, N., Meese, E.: Competitive learning suggests circulating miRNA profiles for cancers decades prior to diagnosis. RNA Biology **17**(10), 1416–1426 (2020). doi:10.1080/15476286.2020.1771945. Accessed 2021-10-24
29. Kryczka, J., Migdalska-Sęk, M., Kordiak, J., Kiszałkiewicz, J.M., Pastuszak-Lewandoska, D., Antczak, A., Brzeziańska-Lasota, E.: Serum Extracellular Vesicle-Derived miRNAs in Patients with Non-Small Cell Lung Cancer—Search for Non-Invasive Diagnostic Biomarkers. Diagnostics **11**(3), 425 (2021). doi:10.3390/diagnostics11030425. Accessed 2021-10-24
30. Leidinger, P., Keller, A., Borries, A., Huwer, H., Rohling, M., Huebers, J., Lenhof, H.-P., Meese, E.: Specific peripheral miRNA profiles for distinguishing lung cancer from COPD. Lung Cancer **74**(1), 41–47 (2011). doi:10.1016/j.lungcan.2011.02.003. Accessed 2021-10-24
31. Leidinger, P., Backes, C., Dahmke, I.N., Galata, V., Huwer, H., Stehle, I., Bals, R., Keller, A., Meese, E.: 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 (2014). doi:10.18632/oncotarget.2419. Accessed 2021-10-24
32. Leidinger, P., Brefort, T., Backes, C., Krapp, M., Galata, V., Beier, M., Kohlhaas, J., Huwer, H., Meese, E., Keller, A.: High-throughput qRT-PCR validation of blood microRNAs in non-small cell lung cancer. Oncotarget **7**(4), 4611–4623 (2016). doi:10.18632/oncotarget.6566. Accessed 2021-10-24
33. 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., Liu, X.-Q.: 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 (2017). doi:10.18632/oncotarget.17535. Accessed 2021-10-24
34. 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., Nicassio, F.: Optimization and Standardization of Circulating MicroRNA Detection for Clinical Application: The miR-Test Case. Clinical Chemistry **62**(5), 743–754 (2016). doi:10.1373/clinchem.2015.251942
35. Nigita, G., Distefano, R., Veneziano, D., Romano, G., Rahman, M., Wang, K., Pass, H., Croce, C.M., Acunzo, M., Nana-Sinkam, P.: Tissue and exosomal miRNA editing in Non-Small Cell Lung Cancer. Scientific Reports **8**(1), 10222 (2018). doi:10.1038/s41598-018-28528-1. Accessed 2021-10-24
36. Patnaik, S.K., Yendamuri, S., Kannisto, E., Kucharczuk, J.C., Singhal, S., Vachani, A.: MicroRNA Expression Profiles of Whole Blood in Lung Adenocarcinoma. PLOS ONE **7**(9), 46045 (2012). doi:10.1371/journal.pone.0046045. Accessed 2021-10-24
37. Patnaik, S.K., Kannisto, E.D., Mallick, R., Vachani, A., Yendamuri, S.: Whole blood microRNA expression may not be useful for screening non-small cell lung cancer. PLOS ONE **12**(7), 0181926 (2017). doi:10.1371/journal.pone.0181926. Accessed 2021-10-24
38. 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., Liu, X.: 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 (2017). doi:10.18632/oncotarget.17416. Accessed 2021-10-24
39. Reis, P.P., Drigo, S.A., Carvalho, R.F., Lopez Lapa, R.M., Felix, T.F., Patel, D., Cheng, D., Pintilie, M., Liu, G., Tsao, M.-S.: 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 (2020). doi:10.3390/cancers12082071. Accessed 2021-10-24
40. Wozniak, M.B., Scelo, G., Muller, D.C., Mukeria, A., Zaridze, D., Brennan, P.: Circulating MicroRNAs as Non-Invasive Biomarkers for Early Detection of Non-Small-Cell Lung Cancer. PLOS ONE **10**(5), 0125026 (2015). doi:10.1371/journal.pone.0125026. Accessed 2021-09-22
41. Yao, B., Qu, S., Hu, R., Gao, W., Jin, S., Liu, M., Zhao, Q.: A panel of miRNAs derived from plasma extracellular vesicles as novel diagnostic biomarkers of lung adenocarcinoma. FEBS Open Bio **9**(12), 2149–2158 (2019). doi:10.1002/2211-5463.12753. Accessed 2021-10-24
42. 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., Laktionov, P.P.: Profiling of 179 miRNA Expression in Blood Plasma of Lung Cancer Patients and Cancer-Free Individuals. Scientific Reports **8**(1), 6348 (2018). doi:10.1038/s41598-018-24769-2. Accessed 2021-10-24
43. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794 (2016). doi:10.1145/2939672.2939785. 1603.02754. Accessed 2021-10-05
44. Zhong, S., Golpon, H., Zardo, P., Borlak, J.: miRNAs in lung cancer. A systematic review identifies predictive and prognostic miRNA candidates for precision medicine in lung cancer. Translational Research **230**, 164–196 (2021). doi:10.1016/j.trsl.2020.11.012. Accessed 2021-11-07
45. Huang, H., Zhu, J., Lin, Y., Zhang, Z., Liu, J., Wang, C., Wu, H., Zou, T.: The potential diagnostic value of extracellular vesicle miRNA for human non-small cell lung cancer: A systematic review and meta-analysis. Expert Review of Molecular Diagnostics **21**(8), 823–836 (2021). doi:10.1080/14737159.2021.1935883. Accessed 2022-01-13

46. Jiang, M., Li, X., Quan, X., Li, X., Zhou, B.: Clinically Correlated MicroRNAs in the Diagnosis of Non-Small Cell Lung Cancer: A Systematic Review and Meta-Analysis. BioMed Research International **2018**, 5930951 (2018). doi:10.1155/2018/5930951. Accessed 2022-01-13
47. Yi, M., Liao, Z., Deng, L., Xu, L., Tan, Y., Liu, K., Chen, Z., Zhang, Y.: High diagnostic value of miRNAs for NSCLC: Quantitative analysis for both single and combined miRNAs in lung cancer. Annals of Medicine (2021). Accessed 2022-01-13
48. Walters, C., Harter, Z.J., Wayant, C., Vo, N., Warren, M., Chronister, J., Tritz, D., Vassar, M.: Do oncology researchers adhere to reproducible and transparent principles? A cross-sectional survey of published oncology literature. BMJ Open **9**(12), 033962 (2019). doi:10.1136/bmjopen-2019-033962. Accessed 2021-11-06
49. Errington, T.M., Mathur, M., Soderberg, C.K., Denis, A., Perfito, N., Iorns, E., Nosek, B.A.: Investigating the replicability of preclinical cancer biology. eLife **10**, 71601 (2021). doi:10.7554/eLife.71601. Accessed 2021-12-11
50. Mobley, A., Linder, S.K., Braeuer, R., Ellis, L.M., Zwelling, L.: A Survey on Data Reproducibility in Cancer Research Provides Insights into Our Limited Ability to Translate Findings from the Laboratory to the Clinic. PLoS ONE **8**(5), 63221 (2013). doi:10.1371/journal.pone.0063221. Accessed 2022-04-07

Figures

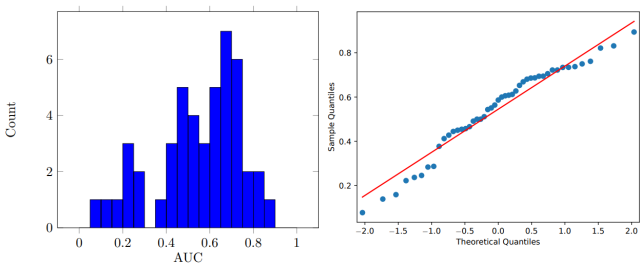


Figure 1 Histogram and QQ-plot over AUC values when training on all datasets except one in one category, and test on the last dataset, aggregated over all categories.

Table 1 Search in public gene expression databases. The first column is the name of the database. The second column is the search term that was used to search the database.

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

Table 2 Characteristics of the studies in this project. The columns are as follows: *Study*: The study the row is describing, *Technology*: The technology used to measure miRNA in that study, *Blood fraction*: What fluid is used for measuring miRNAs, *# miRNAs*: The number of different miRNA-sequences that are measured in the study, *# Cases*: The number of samples from cancer patients in the study, *# Controls*: The number of healthy controls in the study, *Total*: The total number of samples in the study. EV = Extracellular Vehicle, Ex = Exosomal

Study	Technology	Blood fraction	# miRNAs	# Cases	# Controls	Total
[17]	qRT-PCR	Whole blood	4	43	43	86
[18]	Microarray	Serum	2565	1566	2178	3744
[19]	qRT-PCR	Serum	29	95	69	164
[20]	Microarray	Plasma	131	19	6	25
[21]	Sequencing	Plasma	253	30	24	54
[22]	Microarray	Serum	1998	6	6	12
[23]	Microarray	Whole blood	689	606	2440	3046
[24]	qRT-PCR	Serum	254	38	16	54
[25]	Sequencing	Plasma	527	26	12	38
[26]	Microarray	Blood cells	386	17	19	36
[27]	Microarray	P. Blood	722	73	94	167
[28]	Microarray	Serum	435	10	90	100
[29]	qRT-PCR	Serum EV	4	31	21	52
[30]	Microarray	Whole blood	852	28	19	47
[31]	Microarray	Whole blood	1186	42	38	80
[32]	qRT-PCR	Whole blood	205	74	46	120
[33]	Microarray	Plasma	165	6	3	9
[34]	qRT-PCR	Serum	13	48	984	1032
[35]	Sequencing	Plasma Ex	102	19	7	26
[36]	Microarray	Whole blood	1396	33	12	45
[37]	Microarray	Whole blood	3036	86	77	163
[38]	Microarray	Plasma	184	9	4	13
[39]	Microarray	Plasma	795	35	7	42
[40]	qRT-PCR	Plasma	342	100	100	200
[41]	Sequencing	Plasma EV	569	5	5	10
[42]	qRT-PCR	Plasma	175	17	10	27

Table 3 The mean AUC when using cross validation on the given studies with the given models. The first column says which dataset is used, and the rest of the columns have a column name that represents the model used. LR = Logistic Regression, RF = Random Forest

Study	LR	SVM	RF	XGBoost
[17]	0.670	0.814	0.933	0.853
[18]	0.734	0.913	0.968	0.939
[19]	0.795	0.852	0.823	0.843
[20]	0.783	0.950	0.950	0.575
[21]	0.882	0.787	0.793	0.623
[22]	0.900	0.900	0.900	0.800
[23]	0.977	0.980	0.960	0.980
[24]	0.985	0.993	0.983	0.933
[25]	1.000	1.000	1.000	0.980
[26]	0.950	0.900	0.931	0.800
[27]	0.847	0.888	0.864	0.815
[28]	0.956	0.944	0.925	0.967
[29]	0.749	0.658	0.646	0.606
[30]	0.162	0.752	0.705	0.411
[31]	0.160	0.286	0.365	0.528
[32]	0.916	0.903	0.948	0.936
[33]	0.333	0.167	0.833	0.500
[34]	0.976	0.969	0.950	0.968
[35]	0.700	0.300	0.183	0.233
[36]	0.698	0.883	0.777	0.763
[37]	0.573	0.481	0.476	0.543
[38]	1.000	1.000	1.000	0.479
[39]	1.000	1.000	0.957	0.943
[40]	0.494	0.565	0.663	0.689
[41]	0.800	0.600	1.000	0.400
[42]	0.242	0.800	0.842	0.642
Mean	0.742	0.780	0.822	0.721

Table 4 Whether the miRNA-sequences were reported to be significantly up- or down-regulated ($p < 0.05$) in the studies.
Note: [39] only reports miR-210 and miR-182 to be up-regulated in adenocarcinoma. In [40] and [27] abu-miR-155 was measured instead of hsa-miR-155.

Study	miR-21	miR-210	miR-182	miR-155	miR-17
[17]	Up				
[18]					
[19]					Down
[20]	Up	Up			
[21]					
[22]					
[23]	Down	Up	Down		Down
[24]		Down			
[25]	Down			Down	
[26]		Up	Up		Down
[27]				Down	Down
[28]					
[29]					
[30]					Down
[31]	Up				
[32]	Down				Down
[33]					
[34]					
[35]					
[36]	Down	Down			Down
[37]					
[38]					
[39]		Up	Up	Up	
[40]			Down	Up	
[41]					
[42]		Up			

Table 5 Cohen's d of the different miRNAs in the different datasets. Difference in miRNA expression: case - controls

Study	miR-21	miR-210	miR-182	miR-155	miR-17
[17]	-0.784				
[18]	0.496	0.719	0.427	0.592	0.690
[19]					-0.811
[20]	0.300			0.004	0.158
[21]	0.165		0.610		0.147
[22]	-1.723	-0.976	-3.535	-1.631	-0.816
[23]	-0.453	0.002	-0.290	-0.054	-0.542
[24]	-0.311	0.499		0.007	-1.410
[25]	-0.128		-0.132	-1.536	-0.385
[26]	0.321	1.499	0.617		-0.678
[27]	0.067	0.208	-0.008		-1.303
[28]	-0.097				
[29]					
[30]	0.165	-0.102	0.221		-0.663
[31]	0.140	-0.006	-0.035	-0.033	0.131
[32]	-0.804		-0.442		-0.756
[33]		-0.318			-0.242
[34]					
[35]	-0.215	-0.341			-0.309
[36]	-1.009	-0.836			-1.044
[37]	-0.044	-0.070	0.217	0.254	-0.211
[38]	-1.183				-0.954
[39]	-0.374	1.436	1.265		
[40]	0.221	0.013	-0.357		0.429
[41]					
[42]	-0.404	-0.097	-0.456	-0.042	-0.503
Mean	-0.269	0.109	-0.135	-0.271	-0.454

Table 6 AUC of when using the expression of the different miRNAs to diagnose lung cancer in the different datasets

Study	miR-21	miR-210	miR-182	miR-155	miR-17
[17]	0.345				
[18]	0.630	0.742	0.601	0.660	0.690
[19]					0.256
[20]	0.579			0.579	0.588
[21]	0.506		0.608		0.443
[22]	0.083	0.389	0.083	0.111	0.417
[23]	0.359	0.488	0.400	0.472	0.343
[24]	0.332	0.891		0.641	0.112
[25]	0.417		0.554	0.141	0.399
[26]	0.418	0.814	0.672		0.296
[27]	0.509	0.554	0.499		0.167
[28]	0.461				
[29]					
[30]	0.564	0.481	0.598		0.323
[31]	0.548	0.501	0.518	0.479	0.541
[32]	0.218		0.334		0.253
[33]		0.444			0.333
[34]					
[35]	0.425	0.421			0.421
[36]	0.217	0.260			0.230
[37]	0.471	0.477	0.541	0.576	0.449
[38]	0.194				0.250
[39]	0.441	0.910	0.918		
[40]	0.541	0.533	0.421		0.639
[41]					
[42]	0.388	0.359	0.256	0.324	0.224
Mean	0.412	0.551	0.500	0.443	0.369

Table 7 The results when training a logistic regression model on one dataset and testing on another, when stratifying by technology. The in-group is when both datasets have the technology that is listed in the first column. The out-group is when exactly one of the two datasets has the technology that is listed in the first column.
Note: IG = in-group, OG = out-group, mean and standard deviation are of AUC values, t-values are in-group minus out-group and p-values correspond to the t-values

Technology	Mean IG	Std. IG	Mean OG	Std. OG	t-value	p-value
Sequencing	0.535	0.180	0.452	0.165	1.545	0.124
qRT-PCR	0.512	0.153	0.500	0.155	0.416	0.678
Microarray	0.529	0.208	0.477	0.162	2.967	0.003

Table 8 The results when training a logistic regression model on one dataset and testing on another, when stratifying by blood fraction. The in-group is when both datasets have the blood fraction that is listed in the first column. The out-group is when exactly one of the two datasets has the blood fraction that is listed in the first column.
Note: IG = in-group, OG = out-group, mean and standard deviation are of AUC values, t-values are in-group minus out-group and p-values correspond to the t-values

Blood fraction	Mean IG	Std. IG	Mean OG	Std. OG	t-value	p-value
Plasma	0.451	0.178	0.497	0.176	-1.766	0.078
Whole blood	0.538	0.109	0.517	0.166	0.659	0.511
Serum	0.549	0.228	0.494	0.185	1.386	0.167

Table 9 The results when training an XGBoost model on all datasets except one in a certain category and doing testing on the last dataset, when stratifying by technology. The t-value and the corresponding p-value are for the t-test checking whether the expected AUC is larger than 0.50

Technology	Mean AUC	Std. AUC	t-value	p-value
Sequencing	0.625	0.089	2.797	0.034
Microarray	0.505	0.262	0.077	0.470
qRT-PCR	0.493	0.219	-0.086	0.533

Table 10 The results when training an XGBoost model on all datasets except one in a certain category and doing testing on the last dataset, when stratifying by blood fraction. The t-value and the corresponding p-value are for the t-test checking whether the expected AUC is larger than 0.50

Technology	Mean AUC	Std. AUC	t-value	p-value
Serum	0.531	0.222	0.337	0.375
Whole blood	0.583	0.079	2.773	0.016
Plasma	0.376	0.184	-1.908	0.951

Ole Fredrik Borgundvåg Berg

Circulating miRNA and Lung Cancer: - a More Comprehensive Analysis of Available Data

NTNU
Kunnskap for en bedre verden