EARLY BREAST CANCER DETECTION USING LOGISTIC REGRESSION MODELS

ALYSSON DOS SANTOS

EARLY BREAST CANCER DETECTION USING LOGISTIC REGRESSION MODELS

Dissertação apresentada ao Programa de Pós-Graduação em Bioinformática do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Bioinformática.

Orientador: Marcos Augusto dos Santos

Belo Horizonte

Outubro de 2017

ALYSSON DOS SANTOS

EARLY BREAST CANCER DETECTION USING LOGISTIC REGRESSION MODELS

Dissertation presented to the Graduate Program in Bioinformática of the Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Master in Bioinformática.

Advisor: Marcos Augusto dos Santos

Belo Horizonte

October 2017

dos Santos, Alysson

D1234p Early Breast Cancer Detection Using Logistic Regression Models / Alysson dos Santos. — Belo Horizonte, 2017
xxii, 57 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Marcos Augusto dos Santos

1.miRNA 2.Câncer. 3.Regressão logística. 4.AdHoc.

CDU 519.6*82.10

[Folha de Aprovação]

Quando a secretaria do Curso fornecer esta folha,
ela deve ser digitalizada e armazenada no disco em formato gráfico.

Se você estiver usando o `pdflatex`,
armazene o arquivo preferencialmente em formato PNG
(o formato JPEG é pior neste caso).

Se você estiver usando o `latex` (não o `pdflatex`),
terá que converter o arquivo gráfico para o formato EPS.

Em seguida, acrescente a opção `approval={`*nome do arquivo*`}`
ao comando `\ppgccufmg`.

Se a imagem da folha de aprovação precisar ser ajustada, use:
`approval=[`*ajuste*`][`*escala*`]{`*nome do arquivo*`}`
onde *ajuste* é uma distância para deslocar a imagem para baixo
e *escala* é um fator de escala para a imagem. Por exemplo:
`approval=[-2cm][0.9]{`*nome do arquivo*`}`
desloca a imagem 2cm para cima e a escala em 90%.

I dedicate this work to those who do not give up

Acknowledgments

Agradeço a família por apoiarem e terem paciência. A minha esposa Mariana pela ajuda. A prima Marlene pela ajuda. Ao tio Lilio Mestre em Engenharia de Materiais pela ajuda. A minha filha Rachel pela ajuda. The Document Foundation pelo LibreOffice. A PRODABEL pelo tempo cedido. A UFMG.

"Maior que a tristeza de não haver vencido
é a vergonha de não ter lutado !."
(Ruy Barbosa)

Abstract

MicroRNAs (miRNAs) play a central role in gene expression and have remarkable abundance in body fluids. They are candidate diagnostics for a variety of conditions and diseases, including breast cancer. Their main objective is to identify miRNAs for the discrimination of cancer and their intrinsic molecular subtypes in order to recognize potential biomarkers.

More and more linear algebra and statistics methods are used to address issues in gene expression literature. RNAseq technology is one of the extended use tool for overall analysis of miRNAs expression allowing simultaneus investigation of hundreds or thousands of miRNAs in a sample and is characterized by a low sample size and a large number of characteristics (miRNAs) that impair measures of similarity and classification performance. To avoid the problem of "curse dimensionality" many authors have carried out the selection of characteristics or reduced the size of data matrix.

We present new predictive models to classify breast cancer tumor samples in early stage. The methodologies allowed correct classification of early stage breast cancer data set GSE58606 from NCBI with sensibility and specificity greater than 0.95. Also, as a sub-product of the methodology we are able to identify a set of biomarkers already known in others types of cancer.

Keywords: MicroRNA, breast cancer classification, logistic regression .

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Motivation

In recent years, there has been a growing interest in the application of linear algebra and statistical methods in data mining, social networking, machine learning, bioinformatics, information, data retrieval etc [Berry et al., 1995; Eldén, 2006; Horn and Axel, 2003; Koren et al., 2009; Wall et al., 2003]. Among these methods, logistic regression has been shown to be effective for classification using gene expression data for predicting diseases with good results for the classification of cancer [Eilers et al., 2001; Fort and Lambert-Lacroix, 2004; Nguyen and Rocke, 2002; Shen and Tan, 2005; Zhou et al., 2004; Zhu and Hastie, 2004].

Microarray is the technology of choice since the 1990s for the global analysis of gene expression allowing simultaneous occurrence of hundreds or thousands of genes in a sample [Brentani et al., 2005]. Although this genomic tool is not new [Schena et al., 1995], it has matured in the last fifteen years, with the emergence of standardized hybridization protocols producing high quality arrays, precise scanning technology, and robust computational methodology [Powell et al., 2015]. This technology has several limitations and a new powerful technology named RNA sequencing (RNAseq) is predicted to replace microarrays for transcriptome profiling by avoiding some technical issues in microarray studies related to probe performance such as limited detection range of individual probes, cross-hybridization and non-specific hybridization [Zhao et al., 2014]. However, RNAseq is still facing some challenges that are currently limiting its potential utilization: higher cost that makes its use almost impractical for large studies, high data-storage requirements as data produced by an RNAseq experiment is orders of magnitude greater than microarrays data, and the analysis is quite complex for example, a significant number of sequence reads in RNAseq are multireads (reads that have high-scoring alignments to multiple positions in a reference genome or transcript set) and the way to assign multireads to genes is still a problem in reads mapping. Microarray and RNAseq studies are characterized by a low sample number and a large feature (miRNAs) number, which adversely impact similarity measurements and classification performance, since many of these features are irrelevant to specific traits of interest, and therefore contain no discrimination power. If we would project our sam-

ples in the features space, we would have a thousand-dimensional space and we could talk about the "*curse of dimensionality*", coined by Richard E. Bellman [Bellman, 2015] and that in general terms is the widely observed phenomenon that data analysis techniques frequently perform poorly as the dimensionality of the analyzed data increases. Conceptually, the samples are lost in the features space as the dimensionality increases and we would need an enormous number of samples to obtain a satisfactory estimate of, for example, which miRNA have altered expression patterns in a specific tumor type. Many algorithms have been developed to deal with the high-dimensionality problem including the ones that are based on distance functions, clustering or dimensionality reduction [Fort and Lambert-Lacroix, 2004; Giancarlo et al., 2010]. It is possible to have an idea of this problem looking at figure 1.1: there isn't a clear separation between normal breast tissues ones and the molecular cancer sub-types.

Cancer incidence and mortality statistics reported by the American Cancer Society [Siegel et al., 2015; ACS, 2017] and by the United Kingdom Office for National Statistics [ONS, 2015] indicate breast cancer as one of the four most common cancer types, along with lung, colorectal, and prostate. Breast cancer alone is expected to account for 30% of all new cancer diagnoses in women in 2017, being the most frequently diagnosed cancer in women [ACS, 2017].

Breast cancer is a very heterogeneous disease [da Cunha et al., 2013; Zhao et al., 2009] with significant variability between patients. Breast tumors can be grouped in four molecular subtypes, which have major implications for determining treatment (Luminal A, Luminal B, TNBC/Basal-like and HER2) [Irvin and Carey, 2008; CDC, 2015].

In this work we analyze the GSE58606 miRNAs data sets (www.ncbi.nlm.nih.gov) of patients with breast cancer distributed in cancer sub-types and we introduce new logistic regression-based models to classify breast cancer tumor samples based on miRNAs measure data. These new model allows the assignment of values to the parameters of logistic regression that are associated with the presence of a specific miRNA. Scrutinizing these parameters unveiled that some of the parameters topologically located further away from the majority of the parameters are associated with known cancer related miRNAs and flagged some others for further investigation.

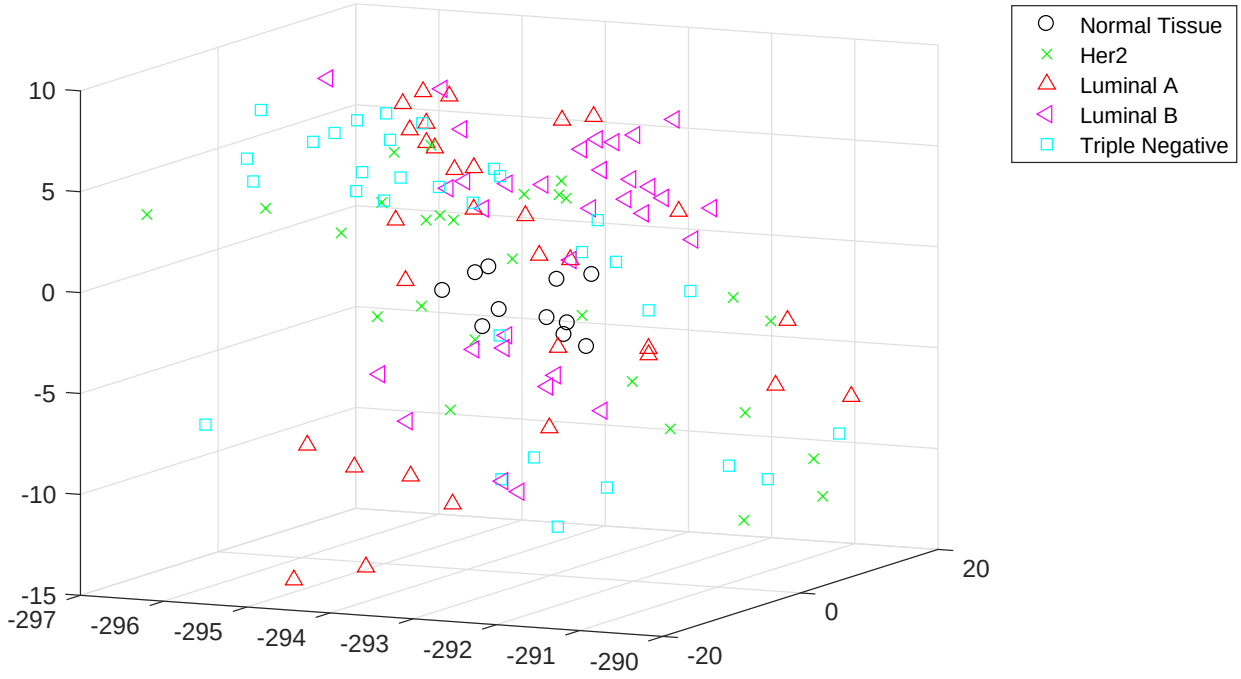

Figure 1.1. Types of tissues obtained using 1926 miRNAs.

Visualization of the distribution of tissue types analyzed using all 1926 miRNAs. The vectors in space with dimension 1926 where projected in space R^3 using the method described by [Marcolino et al., 2010]

1.2 Data Mining Methods

There are several proposed data mining algorithms and an even greater number of implementations of these algorithms. By observing the main characteristics of these algorithms we can separate them into 8 groups: *Bayes, Functions, Tree, Lazy, Rules, Meta, Multi-instance, Miscellaneous*. The WEKA [George-Nektarios, 2013] is a well known prototyping environment in data mining. As an example, we cite for each group, one implementation present in this software.

The algorithms of the *Bayes* group are based on the Bayes Theorem, an example is the Naive Bayes algorithm. The *Functions* group is based on the use of mathematical functions, an example is the SVM implemented in WEKA with the names LibSVM and SMO. The *Tree* group uses a tree-shaped data structure, an example is the C4.5 algorithm implemented with the name J48. The *Lazy* group has the characteristic to postpone the training until the time the query is made an example is or KStar. *Rules* are algorithms that work with discrete and non-ordered values such as Decision Table and FURIA. The *Meta* group is based on the use of other mining algorithms combined one or more times. The *Multi-instance* group are new implementations of mining algorithms for improvement or specialization. The *Miscellaneous* group can

make use of all groups and does the filtering and or selection of data and classification algorithms.

The main characteristic of these methods is the necessity of determination: the number of features have to be lower than the number of samples. To use any of the cited implementations of WEKA, some kind of feature selection must be performed before hand. This is an open research area in data mining. But our approach doesn't use any a priori feature selection algorithm; as a sub-product of then, it is possible to select features to be used by others methods. This is a remarkable aspect; as discussed in the Results section, it is possible to configure good classifiers with only 4 miRNAs out of 1926 ones.

1.3 Objectives

The main objective of this dissertation was to develop new logistic regression based model for early breast cancer detection with good classification performance.

1.3.1 Specific objectives

- Classify GSE58606 miRNAs profile using all miRNAs
- Apply new models using a limited set of profiles and all features
- Flag potential biomarkers for breast cancer in early stage
- Use the flagged biomarkers to classify GSE508606 early breast cancer profiles
- Compare our results with some benchmark in literature

1.4 Conclusion

Breast cancer is a public healthy concern and the RNAseq technology is establishing outstanding problems for their effective use in cancer studies. In this work we are interested in developing tools to aggregate to the data mining methods for detection early breast cancer using RNAseq data sets profiles.

Chapter 2

Material and Methods

In this chapter we present methods to accomplish our specific objectives shown in section 1.3.1. The Adhoc algorithm uses only a subset of patients to classify some given query. It selects the most similar ones from the data base using singular value decomposition (SVD). To circumvent the intrinsic mathematical indetermination of the miRNAs profiles, we use the modified logistic regression method that prevents from pruning the set of miRNAs.

We are testing our methodology using only one data set. To our best knowledge, at present, there is not another data set in the literature to study early breast cancer using miRNAs.

2.1 Data Collection and Generation

A collection of data set containing miRNAs profiles of patients with early breast cancer samples, together with others from healthy patients, was used to demonstrate the usefulness of the proposed methodology. Data sets were downloaded from NCBI [GEO, 2015] with the identifier GSE58606, that was acquired using miRCURY LNA microRNA Array 7th generation. The data set consists of 1926 measurements of miRNA gene expression proflings from 133 samples, grouped into 31 Luminal A, 33 Luminal B, 27 HER2 and 31 TNBC and 11 normal breast tissues.

A set of five systems were created: system Normal distinguishes between normal breast tissues and all other types; system Her2 discriminates HER2 and all other types; system Luminal A discriminates Luminal A against and all other types; system Luminal B distinguishes between Luminal B and all other types; system TNBC distinguishes TNBC from all other types.

2.2 Singular Value Decomposition

SVD has proven ability to establish non-obvious and relevant relationships in a vector space model, providing a deterministic method for grouping related items (vectors). The logic behind the SVD is that a matrix A can be represented by a set of derived matrices (2.1), in the same way that a number can be derived in factors. One can also

think of SVD as a set of matrices that provide numerically different representations of data without loss of semantic meaning, such as representation in different base numbers. To understand the mathematical concept of SVD, suppose that $\boldsymbol{A}$ is a set of arrays of real numbers or complex numbers composed of m rows by n columns. A matrix with a singular value decomposition of matrix $\boldsymbol{A}$ can be made:

$$A = U\Sigma V^T \tag{2.1}$$

.

where $\boldsymbol{U}$ is an array m rows by m orthogonal columns, and Σ is an $m \times n$ diagonal matrix, with real and non-negative numbers. The matrix V^T is known as conjugate transposition. Since the diagonal values of Σ are sorted in descending order, Σ is a direct function of matrix A and characterizes the singular values of that matrix, ordering them from the most significant values to the least significant ones. Considering a subset of singular values of size $k < n$, we can obtain A_k an approximate matrix of the matrix A:

$$A_k = U_k\Sigma_k V_k^T \tag{2.2}$$

.

Thus, the approximation of data depends on how many singular values are used (2.2). In this case, the number of singular values k is also the rank of the matrix A_k, indicating how many rows and columns in the matrix A_k are linearly independent. The possibility of extracting information based on less data is part of the reason for the success of this technique, since it allows the data compression / decompression, with a runtime that does not increase exponentially with the increase of the matrix size, making the analysis viable. A data set represented by a smaller number of unique values than the original size data set has a tendency to group data items that would not be grouped if we used the original data set (2.1). This could explain why clusters derived from SVD can expose nontrivial relationships between the original data set items (2.3). In this work we do not use the matrix A_k, the product factorization by SVD; with only two SVD matrices, the matrix D_k (2.2) is represented in the context of the matrix

$$A_k = U_k\Sigma_k V_k^T = U_k(\Sigma_k V_k^T) = U_k D_k \tag{2.3}$$

.

The justification for using only D_k is that it has k lines instead of m lines of A_k, so D_k is composed of linear combinations of U_k columns, which in turns gives the relation $A \approx A_k$ which is represented by D_k.

2.3 Modified Logistic Regression

We associate a function $P(x)$ for each individual of the model given by

$$P_i(x) = \frac{e^{\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n}}{1 + e^{\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n}} \tag{2.4}$$

.

The logistic regression consists of finding a vector $\alpha = (\alpha_1, ..., \alpha_n)^T$ to adjust the set of equations (2.4). We note that when $e^{\alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_n x_n}$ falls to zero, $P_i(x)$ also drops to zero. On the other hand, if it goes to infinite, $P_i(x)$ approaches to one. Considering $P_i(x)$ as probability, the odds $C_i(x)$ are given by:

$$C_i = \frac{P_i(x)}{1 - P_i(x)} \tag{2.5}$$

.

Expressing the equation (2.5) using (2.4) ,we have:

$$C_i = e^{\alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_n x_n} \tag{2.6}$$

To implement the method, we adopt $\hat{C}_i(x) \approx C_i(x) = \frac{0,9999}{1-0,9999}$ instead of $C_i(x)$, when the odds are related to $P_i(x) = 1$; when $P_i(x) = 0$, we use $\hat{C}_i(x) \approx C_i(x) = \frac{0,0001}{1-0,0001}$. Taking the logarithm on both sides of (2.6), we obtain a system of linear equations to determine α :

$$b_i = \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_n x_n \tag{2.7}$$

.

where $b_i = log(\hat{C}_i)$, for $i = 1, 2, ..., m$. Let $b = (b_1, b_2, ..., b_m)^T$, them the system of linear equations (2.7) can be represented by:

$$A\alpha = b \tag{2.8}$$

.

On (2.8) there are fewer equations than unknowns, and the system is indeterminate, with an infinite number of solutions. The classical approach in linear algebra minimizes α subject to $A\alpha = b$, which requires the complete rank of $A^T A$ - a property not expected from the matrix A. It is usually to circumvent this difficulty by discarding a subset of the original variables, keeping only a subset of the original variables. This procedure resembles the feature selection in data mining - an open research area. We propose the use of a stabilizing term in the logistic regression model, found in the works of [Linnik et al., 1961], [Golub, 1965] and [Menard, 2010], which allows the assignment of values to the parameters α minimizing the sum of squares of residues $(A\alpha - b)$, added to the squares of α. Thus, to assign a solution to (2.9), we solve an unrestricted quadratic optimization problem given by

$$Minimize f(\alpha) = \alpha^T \alpha + (A\alpha - b)^T (A\alpha - b) \tag{2.9}$$

.

As $f(\alpha)$ Is convex, the argument that minimizes (2.9) is given by the derivation of $f(\alpha)$ about α and matching the result to zero:

$$(I + A^T A)\alpha = A^T b \tag{2.10}$$

Where I is an identity matrix of dimension n.

The optimal solution for α (2.9) is obtained by the solution for (2.10) and it is unique. Then, given a query $q = [q_1, q_2..., q_n]$ with the microRNA profile, the probability of q to belong to a class associated to the related system is given by:

$$P(q) = \frac{e^{q\alpha}}{1 + e^{q\alpha}} \tag{2.11}$$

2.4 AdHoc Algorithm

The adhoc algorithm is another proposed improvement to the logistic regression method. We use a subset of the expression profiles with the objective of creating a more precise model.

We express the AdHoc using 2 extra parameters, $k1$ and $k0$, that indicates how many positive and negative elements will be used in the model. They are the $k0$ negative ($k1$ positive) closest to the query. The proximity is determined by the Euclidean distance calculated in a reduced space given by SVD.

To determine $k0$ and $k1$, 5 cross-validations are performed. The details are in Algorithm 1 in the following.

Algorithm 1: Ad Hoc

L_0: Matrix with unseen expression profiles;
L_1: Matrix with searched expression profiles;
c: Expression profile to be analyzed;
v: Cutoff value;
k_0: Number of unseen expression profiles;
k_1: Number of searched expression profiles;
$LR_0 \leftarrow$ Reduces the rank of L_0 using SVD;
$LR_1 \leftarrow$ Reduces the rank of L_1 using SVD;
$ic \leftarrow$ Projection of c in the space defined by LR_0;
$ac \leftarrow$ Projection of c in the space defined by LR_1;
$IM_0 \leftarrow$ Get the k_0 unseen expression profiles closest to ic in LR_0;
$IM_1 \leftarrow$ Get the k_1 searched expression profiles closest to ac in LR_1;
Now with the IM_0 and IM_1 matrix, use the our modified logistic regression method to calculate the p of c;
if $p \geq v$ **then**
  The expression profile is positive
end
else
  Expression profile is negative
end

2.5 Conclusions

The SVD is an up to date technique commonly used in search engines that retrieves information not *prima facie* related. On the other hand, Modified Logistic Regression

prevent the use of some previous feature selection technique. These two methods are the core of the Adhoc algorithm.

Unfortunately it was possible to use only one data set in our study. We didn't find another one to complete our study.

Chapter 3

Results

In this section we present experiments with three models: Full Logistic Regression (**FLR**), Reduced Logistic Regression (**RLR**) and Ad Hoc Logistic Regression (**ALR**). To assess how accurately these predictive models will be in practice, we perform cross-validations. In each round we did a partition of the data set in two complementary sub data sets (training and test). We subject all models to exhaustive experiments; as supplementary material we registered results from 50 rounds where 20% of the data set were randomly assigned to test and the 80% used in our experiment as training.

Some authors in small sets propose 50-50 partition (training-test) to perform the cross-validations. We chose the 80-20 partition along 50 rounds to eventually detect data set distortions.

The GSE58606 consists of 122 breast cancer samples, grouped into 4 major subtypes (31 TNBC, 27 HER2, 31 Luminal A e 33 Luminal B) and 11 non-tumor breast cancer. We create a set of five systems: Normal, Hers2, Luminal A, Luminal B and TNBC, that distinguishes each type of tissue from the others.

The results of a specific classification system along the 50 rounds of cross-validations are presented in just one figure. All values computed of $P(x)$ in each test set are plotted together to give a clear image of the performance.

In the following, for each model (**FLR**, **RRL** and **ARL**) we present in one table the results for all the 5 systems. Other figures for are presented here only for the Normal system; one of then presents the performance along all cross-validations and in the others are shown the most important miRNAs from the point of view of the classifier. These same figures for the other four systems are shown in the Appendix.

3.1 FRL model

This model uses the modified logistic regression method with all 1926 miRNAs (the matrix A has dimension (1926×133)). The results are summarized in Table 3.1, where we would like to point the excellent result obtained from the Normal system.

Table 3.1. Results for Full Logistic Regression (**FLR**) model

	Performance			
System	sensitivity	specificity	F_1	AUC
Normal	0.92	1	0.95	0.96
TNBC	0.91	0.85	0.88	0.94
Luminal A	0.84	0.77	0.80	0.85
Luminal B	0.67	0.74	0.69	0.74
Her2	0.85	0.70	0.76	0.82

To each one of the 50 cross-validations in the Normal system we randomly select 3 non-breast tissue data and 25 of all other cancer profile tissues for test. The other 8 normal tissues profiles and the remaining 97 cancer profiles are used to construct the model in each of the 50 rounds. Figure 3.1 shows all the values of $P(x)$ computed in each folder. The most relevant miRNAs from the point of view of the values of α are shown in Figure 3.3.

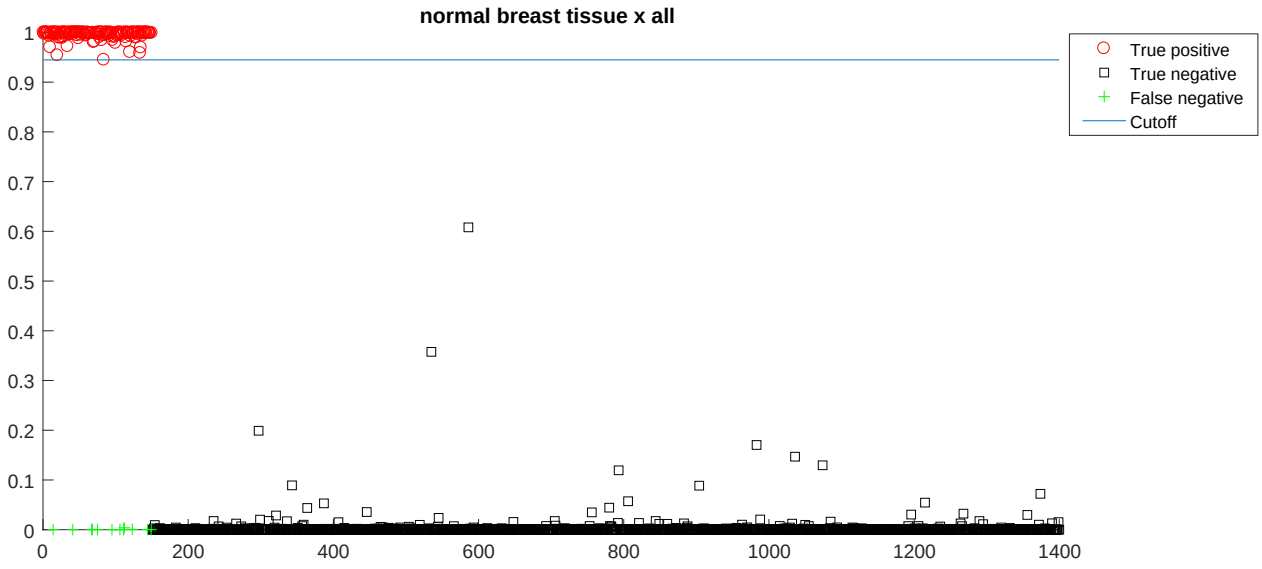

Figure 3.1. $P(x)$ values computed along all 50 cross-validations

3.2 RLR model

This model uses a subset of 20 features selected from the **FLR** model (see Table 3.3). In table 3.4 we show the references in the literature of this set of miRNAs. The results are in Table 3.2. Again, the Normal system, that distinguishes non-tumor breast cancer from early breast cancer tissues had the best performance. The result for Luminal B system was improved from $F_1 = 0.69$ to $F_1 = 0.83$, when compared with the **FLR** model. In Figure 3.2 we show the performance of the model RLR in the Normal system.

Table 3.2. Results for Reduced Logistic Regression (**RLR**) model

	Performance			
System	sensitivity	specificity	F_1	AUC
Normal	0.99	0.92	0.95	0.97
TNBC	0.94	0.85	0.88	0.94
Luminal A	0.90	0.80	0.85	0.92
Luminal B	0.81	0.83	0.81	0.88
Her2	0.74	0.75	0.74	0.79

Table 3.3. Selected miRNAs for **RLR** model

miRNAs with $\alpha > 0$	miRNAs with $\alpha < 0$
hsa-miR-147b	hsa-miR-4726-5p
hsa-miR-125a-5p	hsa-miR-4419b
hsa-miR-4475	hsa-miR-135a-3p
hsa-miR-4421	hsa-miR-4764-3p
hsa-miR-4667-5p	hsa-miR-491-3p
hsa-miR-4507	hsa-miR-1908
hsa-miR-3621	hsa-miR-1973
hsa-miR-3124-3p	hsa-miR-21-5p
hsa-miR-125b-5p	hsa-miR-720
hsa-miR-4695-3p	hsa-miR-4456

Table 3.4. References found in literature of the selected 20 miRNAs used in **RLR** model

miRNA	Type of Cancer	References
hsa-miR-147b	ovarian, Colon	[Kleemann et al., 2017] [Bertero et al., 2012; Omrane et al., 2014]
hsa-miR-125a-5p	Lung, gastric	[Jiang et al., 2010; Leotta et al., 2014] [Hashiguchi et al., 2012]
hsa-miR-4475	Esophagus	[Drahos et al., 2015]
hsa-miR-4421	Mucosa	[Slattery et al., 2016]
hsa-miR-4667-5p	Bone marrow (plasmocyte)	[Ronchetti et al., 2016]
hsa-miR-4507	Pancreas	[Schreiber et al., 2016; Xun et al., 2015]
hsa-miR-3621	Rectal , Colorectal	[Mullany et al., 2016; Jung et al., 2016]
hsa-miR-3124-3p	breast	[Wang et al., 2017]
hsa-miR-125b-5p	Hepatocellular, Lymphomas	[Giray et al., 2014] [Manfe et al., 2013]
hsa-miR-4695-3p	Colorectal	[Moshammer et al., 2014]
hsa-miR-4726-5p	Colorectal	[Mullany et al., 2016]
hsa-miR-4419b	Esophageal, Gastric and hepatic	[Okumura et al., 2015] [Hibino et al., 2014]
hsa-miR-135a-3p	Gallbladder	[Fukagawa et al., 2017; Zhou et al., 2014]
hsa-miR-4764-3p	Breast	[Wang et al., 2017]
hsa-miR-491-3p	Tongue, hepatocellular carcinoma	[Zheng et al., 2015] [Zhao et al., 2017]
hsa-miR-1908	Glioblastoma (brain)	[Xia et al., 2015]
hsa-miR-1973	Classic Hodgkin's Lymphoma	[Jones et al., 2013]
hsa-miR-21-5p	Rectum	[Lopes-Ramos et al., 2014]
hsa-miR-720	Breast, Colorectal Cervical cancer	[Li et al., 2013c; Wang et al., 2015] [Tang et al., 2015]
hsa-miR-4456		[Ge et al., 2017]

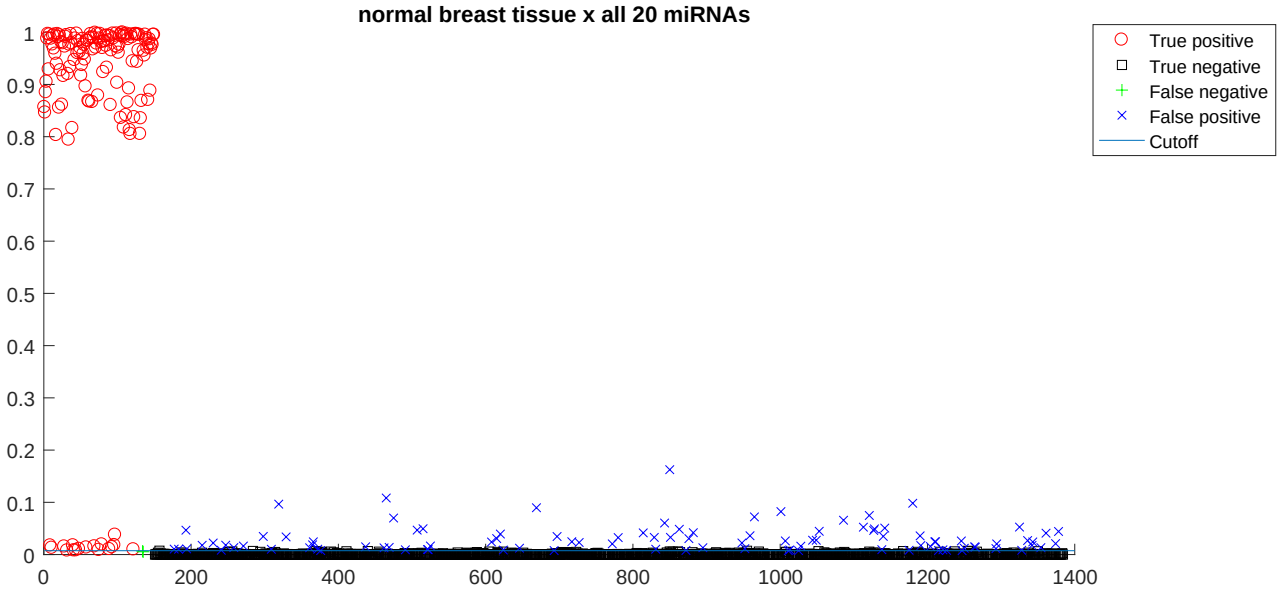

Figure 3.2. $P(x)$ values computed along all 50 cross-validations

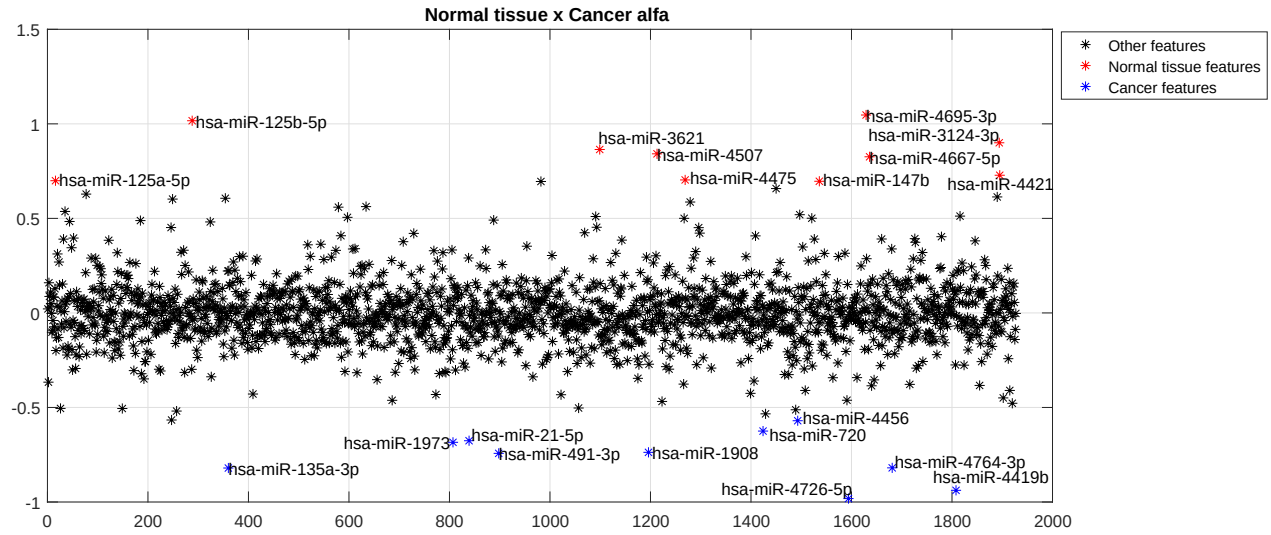

Figure 3.3. Values of α computed for the Normal system

3.3 ALR model

In the Normal system discussed here is present the same 20 features shown in Table 3.3 . But only a subset of samples are used to construct the model. In other words, for each query in the test set, the model is built from selected profiles of 3 non-tumor breast tissues and 59 from other cancer tissues. All profiles are retrieved among the most similar ones using SVD technique. We achieved an exceptional result for this system ($F_1 = 0.98$).

Table 3.5. Results for Ad hoc Logistic Regression (**ALR**) model

	Performance			
System	sensitivity	specificity	F_1	AUC
Normal	0.98	0.98	0.98	0.99
TNBC	0.89	0.87	0.88	0.92
Luminal A	0.91	0.86	0.88	0.94
Luminal B	0.86	0.82	0.83	0.90
Her2	0.79	0.78	0.77	0.82

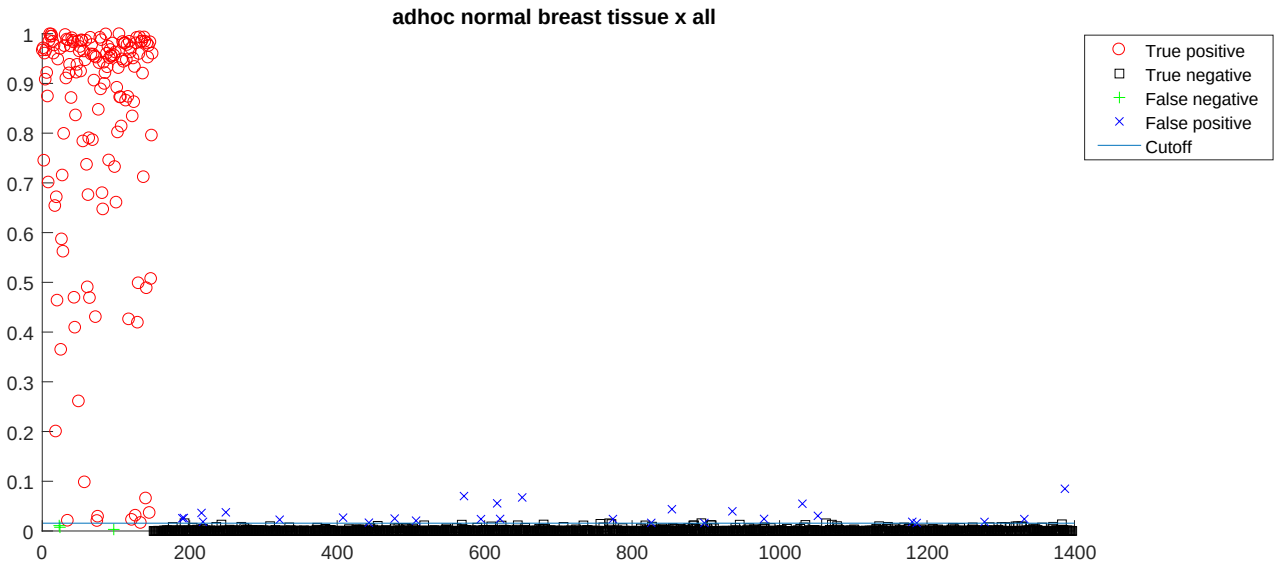

Figure 3.4. $P(x)$ values computed along all 50 cross-validations

3.4 On the number of features

In the **RLR** and **ALR** methods, only a subset of 20 from all 1926 features are present in the models. This number of features (20) was assigned without any rational criteria. We have some insights that it is possible to use a lower number of then. In the following we report results using 10 and 6 features in the **RLR** algorithm and 4 features in the **ALR**.

Using 10 features (see Table 3.6), the **RLR** in Normal system achieves $F_1 = 0.94$ $sensitivity = 0.99$, $specificity = 0.91$ and $AUC = 0.97$.

Table 3.6. Selected 10 miRNAs for **RLR** model

miRNAs with $\alpha > 0$	miRNAs with $\alpha < 0$
hsa-miR-4507	hsa-miR-491-3p
hsa-miR-3621	hsa-miR-4764-3p
hsa-miR-3124-3p	hsa-miR-135a-3p
hsa-miR-125b-5p	hsa-miR-4419b
hsa-miR-4695-3p	hsa-miR-4726-5p

Table 3.7. References found in literature of the selected 10 miRNAs used in **RLR** model

miRNA	Type of Cancer	References
hsa-miR-4507	Pancreas	[Schreiber et al., 2016; Xun et al., 2015]
hsa-miR-3621	Rectal , Colorectal	[Mullany et al., 2016; Jung et al., 2016]
hsa-miR-3124-3p	breast	[Wang et al., 2017]
hsa-miR-125b-5p	Hepatocellular, Lymphomas	[Giray et al., 2014] [Manfe et al., 2013]
hsa-miR-4695-3p	Colorectal	[Moshammer et al., 2014]
hsa-miR-4726-5p	Colorectal	[Mullany et al., 2016]
hsa-miR-4419b	Esophageal, Gastric and hepatic	[Okumura et al., 2015] [Hibino et al., 2014]
hsa-miR-135a-3p	Gallbladder	[Fukagawa et al., 2017; Zhou et al., 2014]
hsa-miR-4764-3p	Breast	[Wang et al., 2017]
hsa-miR-491-3p	Tongue, hepatocellular carcinoma	[Zheng et al., 2015] [Zhao et al., 2017]

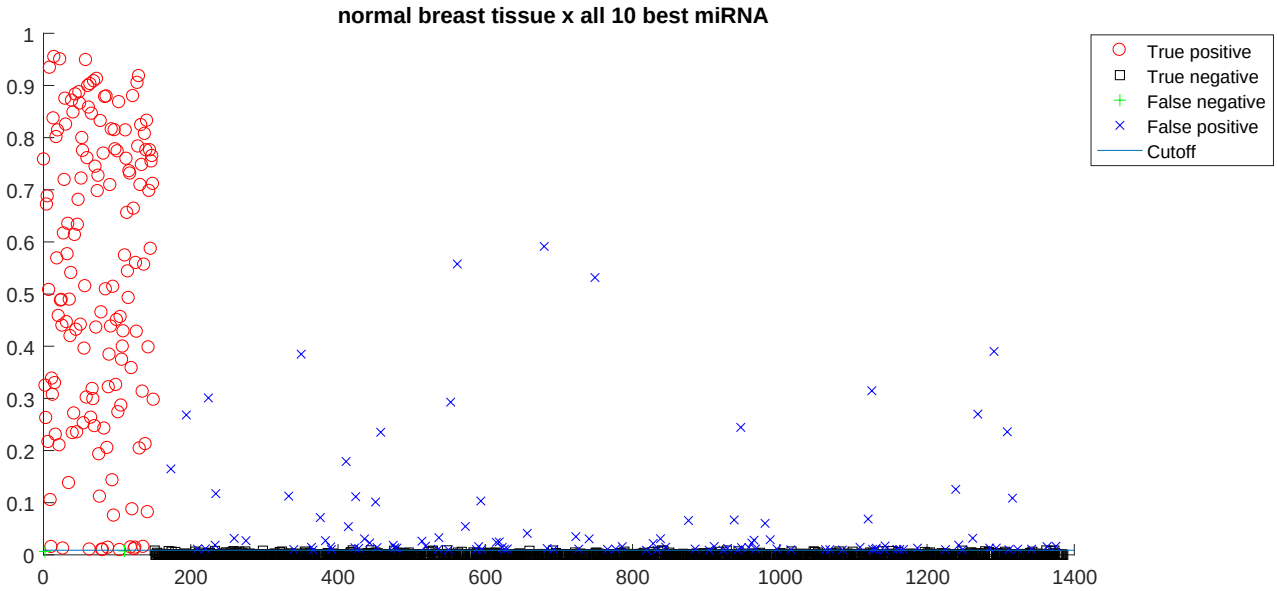

Figure 3.5. RLR applied to the the Normal system using 10 miRNAs $P(x)$ computed along 50 cross-validations

The figures using using 6 miRNs are: $F_1 = 0.85$, $sensibility = 0.89$, $specificity = 0.83$ and $AUC = 0.75$.

Table 3.8. Selected miRNAs for Normal System in **RLR** model

miRNAs with $\alpha > 0$	miRNAs with $\alpha < 0$
hsa-miR-3124-3p	hsa-miR-135a-3p
hsa-miR-125b-5p	hsa-miR-4419b
hsa-miR-4695-3p	hsa-miR-4726-5p

Table 3.9. References found in literature of the selected 6 miRNAs used in **RLR** model

miRNA	Type of Cancer	References
hsa-miR-3124-3p	breast	[Wang et al., 2017]
hsa-miR-125b-5p	Hepatocellular, Lymphomas	[Giray et al., 2014] [Manfe et al., 2013]
hsa-miR-4695-3p	Colorectal	[Moshammer et al., 2014]
hsa-miR-4726-5p	Colorectal	[Mullany et al., 2016]
hsa-miR-4419b	Esophageal, Gastric and hepatic	[Okumura et al., 2015] [Hibino et al., 2014]
hsa-miR-135a-3p	Gallbladder	[Fukagawa et al., 2017; Zhou et al., 2014]

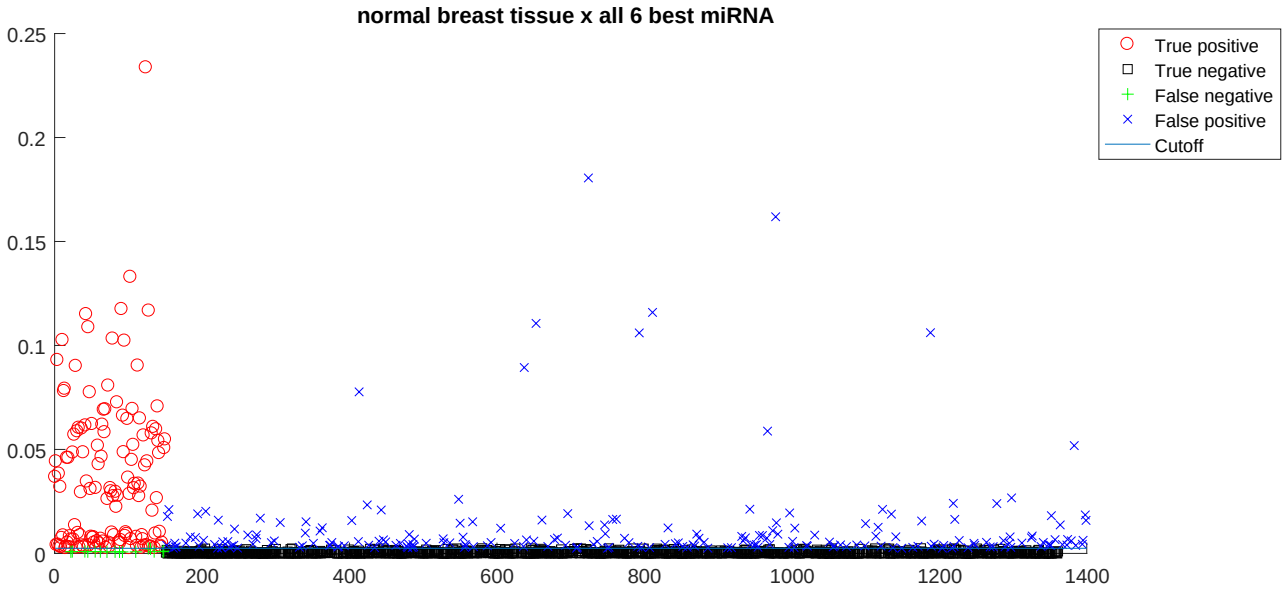

Figure 3.6. $P(x)$ computed for test set of Normal system in **RLR** model along 50 cross-validations

Using only 4 out of 1926 miRNAs present in the profile of each sample(**ALR** algorithm with $k0 = 1$ and $k1 = 1$), we have the following result: $F_1 = 0.80$, $sensivity = 0.89$ $specificity = 0.75$ and $AUC = 0.88$.

Table 3.10. Selected miRNAs for Normal System in **ALR** model

miRNAs with $\alpha > 0$	miRNAs with $\alpha < 0$
hsa-miR-125b-5p	hsa-miR-4419b
hsa-miR-4695-3p	hsa-miR-4726-5p

Table 3.11. References found in literature of the selected 4 miRNAs used in **ALR** model

miRNA	Type of Cancer	References
hsa-miR-125b-5p	Hepatocellular, Lymphomas	[Giray et al., 2014] [Manfe et al., 2013]
hsa-miR-4695-3p	Colorectal	[Moshammer et al., 2014]
hsa-miR-4726-5p	Colorectal	[Mullany et al., 2016]
hsa-miR-4419b	Esophageal, Gastric and hepatic	[Okumura et al., 2015] [Hibino et al., 2014]

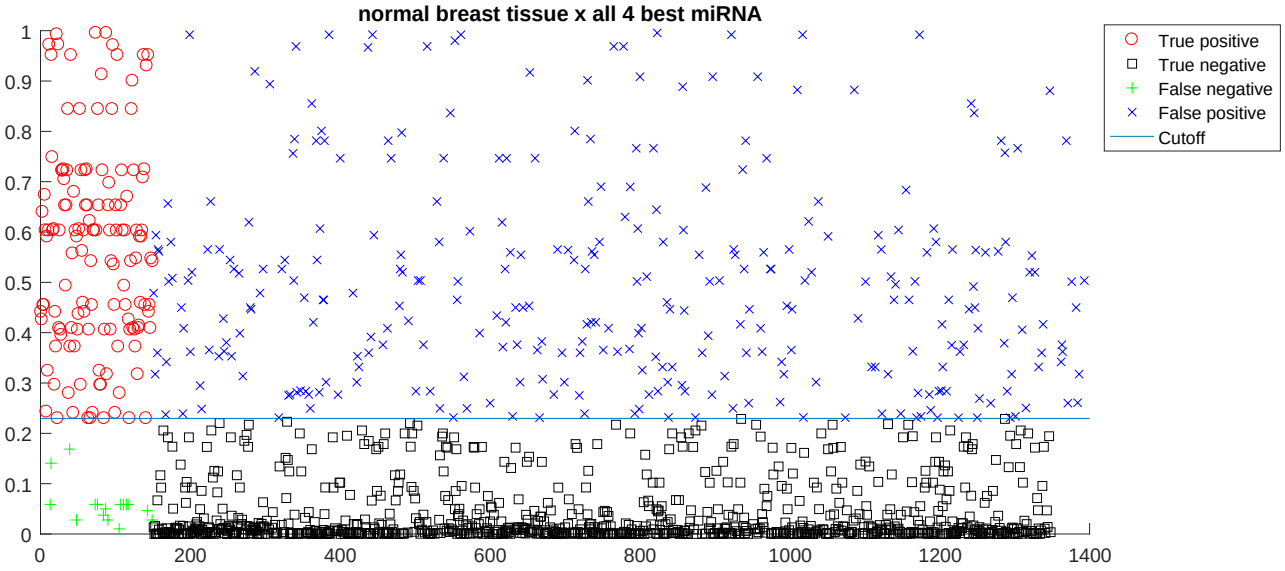

Figure 3.7. $P(x)$ computed for test set of Normal system in **ALR** model along 50 cross-validations

3.5 Conclusions

We presented tree efficient methods (**FRL,RLA** and **ALR**) to classify breast cancer in early stage that had remarkable results. We are able not only to detect early breast cancer in tissues from their miRNA profile but it was also possible to classify them in the four possible molecular subtypes (TNBC, Luminal A, Luminal B and Her2). Our main result was attained with the **ARL** in the case that distinguishes healthy tissues from tumor tissues using a set of only 10 miRNAs (sensitivity 0.99 and specificity 0.91).

In the **RLR** model we verify that some miRNAs are common in some systems (see Figure 3.8). We used a naive algorithm to select them; the set of 20 miRNAs was selected picking the 10 miRNAs with the greater α and the 10 miRNAs with lowest

values of α. We think some heuristic should be used based in a criteria that maximizes the usage of the miRNA along the five systems should be used.

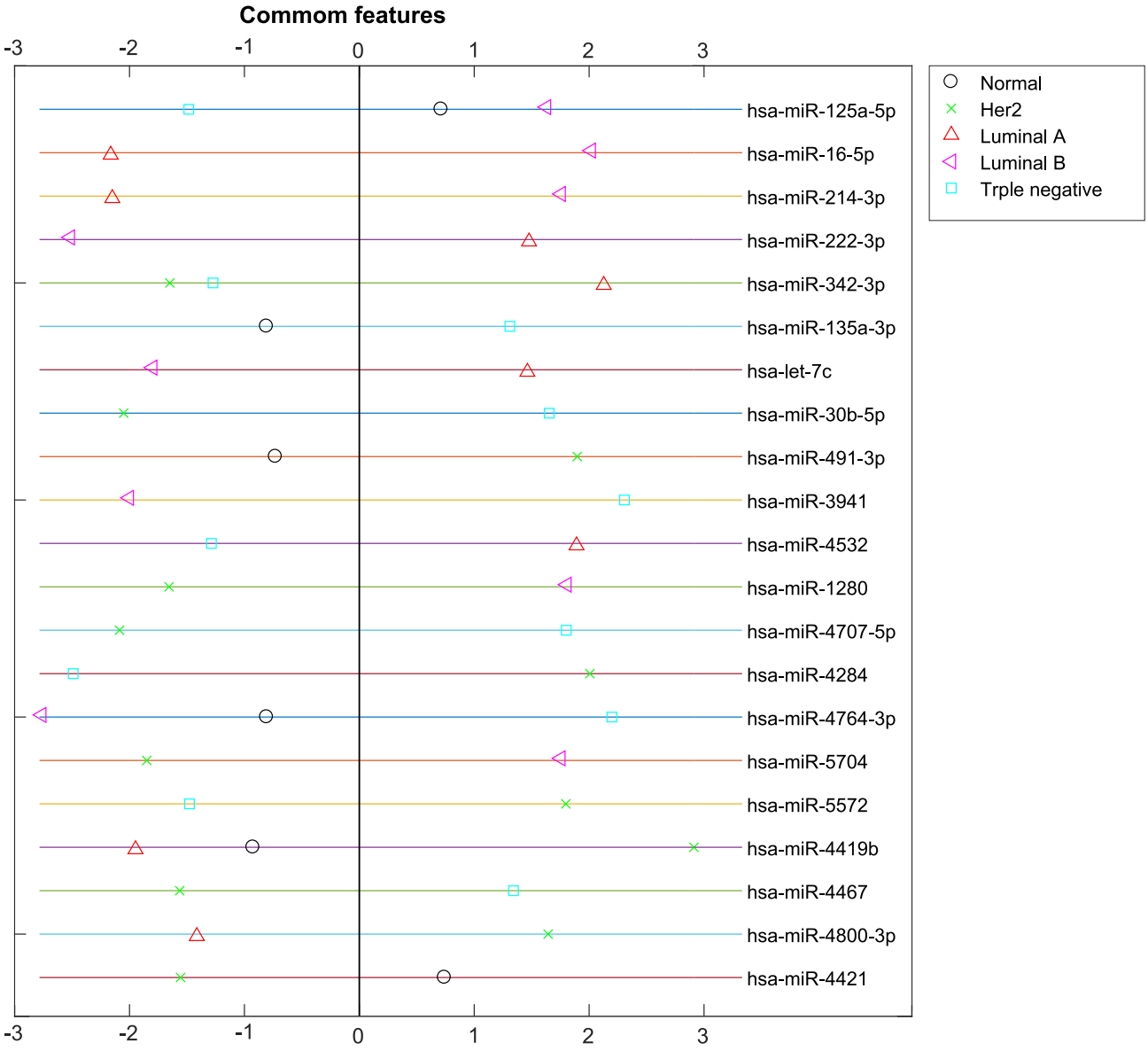

Figure 3.8. Common miRNAs in all 5 **RLR** systems with 20 features extracted from **FRL** models

Chapter 4

Conclusions

MiRNAs are emerging as promising diagnostic tool and has a potential to play a fundamental role in cancer prevention and treatment. Our study signs their remarkable discrimination power. With only a small number of miRNAs; 4 from 1926; we are able to classify healthy tissues from early breast cancer with $F_1 = 0.80$;. With 10 this figure jumps to 0.94.

Along this project was developed efficient algorithms to solve a class of indeterminate data mining problems where the number of features are greater than the number of samples. The contribution in Logistic Regression field generates interesting models that can be applied in a broad range of problems. At first, **FRL** uses all 1926 features in five models to distinguish healthy tissues and each molecular breast cancer sub-type. As a sub-product using only the k (20, 10, 6 and 4) features with more discrimination power indicated by **FRL** we propose the **RLR** method. Finally we build ad hoc models (**ARL**) specific for each query, that uses only a subset of the most similar profiles from the samples to classify the query.

Our models not only classify breast cancer in early stage from healthy tissues; they also classify each molecular sub-type in early stage from the other sub-types and healthy tissues. In particular, the TNBC system had a performance of $F_1 = 0.88$ (sensitivity=0.94, specificity=0.85); TNBC is reported as difficulty to classify [Cetin and Topcul, 2014].

To our best knowledge, at present there isn't another miRNA profiles repository reported for early breast cancer. So, unfortunately, we are not able to confront our results with another data base.

Our project was based only in miRNAs in tissues, although miRNAs in body fluids are candidate diagnostics for a variety of conditions and diseases, including breast cancer. One premise for using extracellular miRNAs to diagnose disease is the notion that the abundance of the miRNAs in body fluids reflects their abundance in the cells causing the disease. As a result, the search for such diagnostics in body fluids has focused on miRNAs that are abundant in the cells of origin. There is a report that released miRNAs do not necessarily reflect the abundance of miRNA in the cell of origin ([Pigati et al., 2010]). So, as a future work, we would like to research profiles from fluids for the identification of circulating miRNAs as biomarkers of disease.

Appendix A

Results of the models FRL, RLR and ALR in the cancer sub-types systems

A.1 FRL Models

System TNBC

Figures A.1 and A.2 present the results for **FRL** model. As a test set was used in 7 TNBC samples and 21 tissues from other sub-types of cancer and healthy tissues. The training was built with 105 tissues from healthy and cancer profiles. The **FRL** in the system TNBC attained $sensitivity = 0.91$, $specificity = 0.85$, $F_1 = 0.88$ and $AUC = 0.94$.

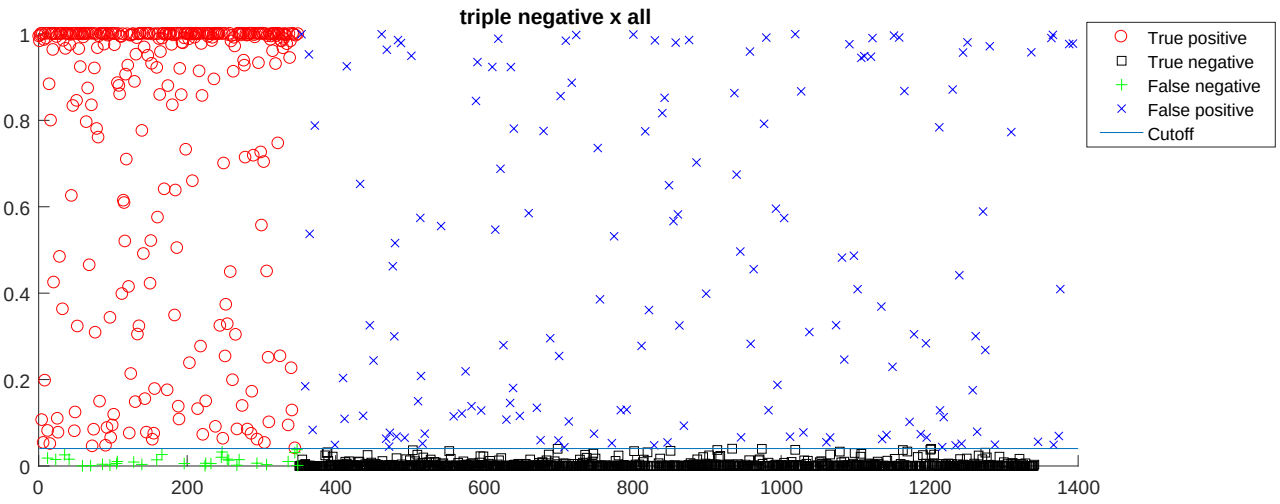

Figure A.1. $P(x)$ values computed along all 50 cross-validations (**FRL**)

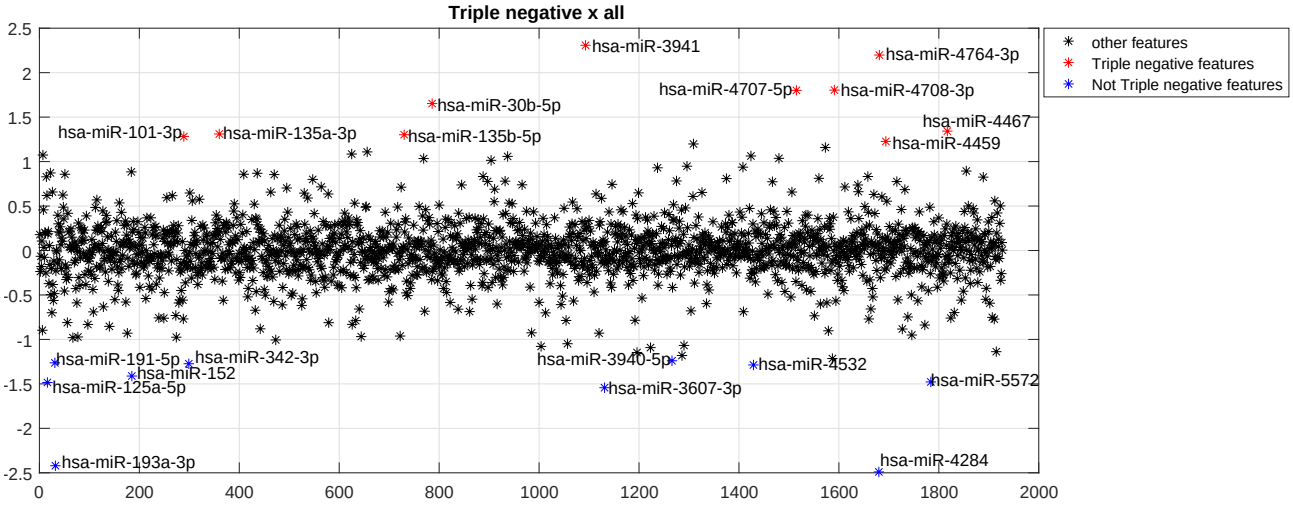

Figure A.2. Values of α computed for the TNBC system

System Luminal A

Figures A.3 and A.4 present the results for **FRL** model. As a test set was used in 7 Luminal A samples and 21 tissues from other sub-types of cancer and healthy tissues. The training was built with 105 tissues from healthy and cancer profiles. The **FRL** in the system Luminal A attained $sensitivity = 0.84$, $specificity = 0.77$, $F_1 = 0.80$ and $AUC = 0.85$.

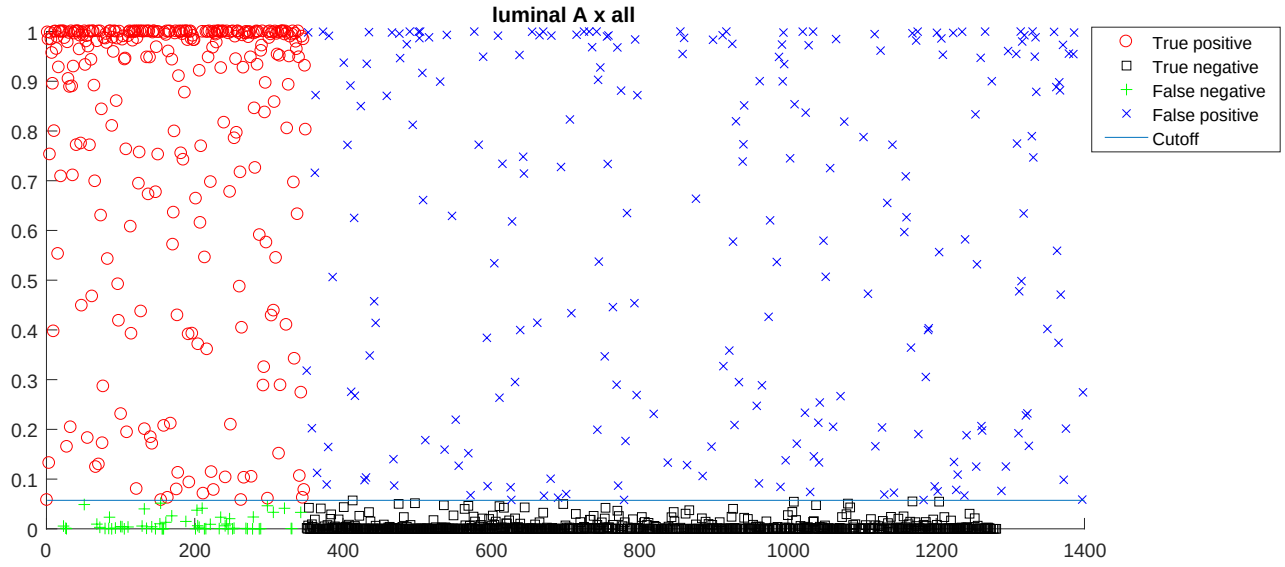

Figure A.3. $P(x)$ values computed along all 50 cross-validations (**FRL**)

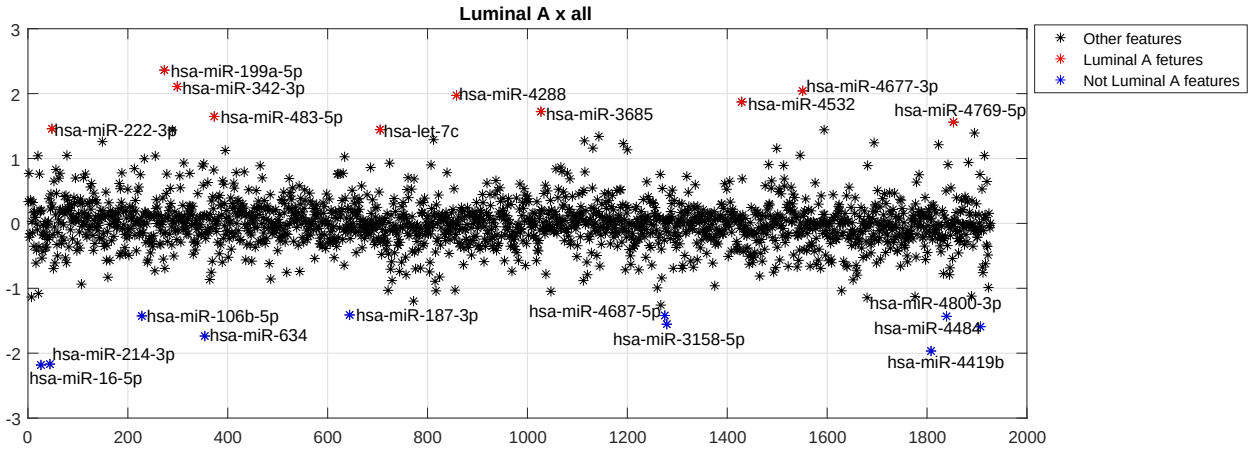

Figure A.4. Values of α computed for the Luminal A system

System Luminal B

Figures A.5 and A.6 present the results for **FRL** model. As a test set was used in 7 Luminal B samples and 20 tissues from other sub-types of cancer and healthy tissues. The training was built with 106 tissues from healthy and cancer profiles. The **FRL** in the system Luminal B attained $sensitivity = 0.67$, $specificity = 0.74$, $F_1 = 0.69$ and $AUC = 0.74$.

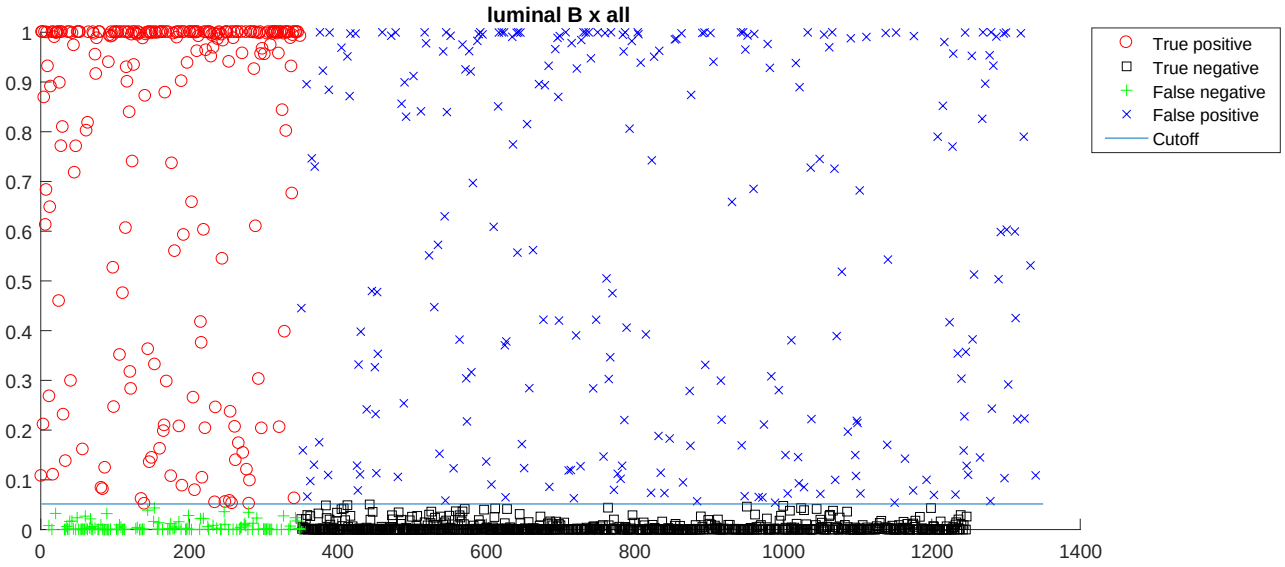

Figure A.5. $P(x)$ values computed along all 50 cross-validations (**FRL**)

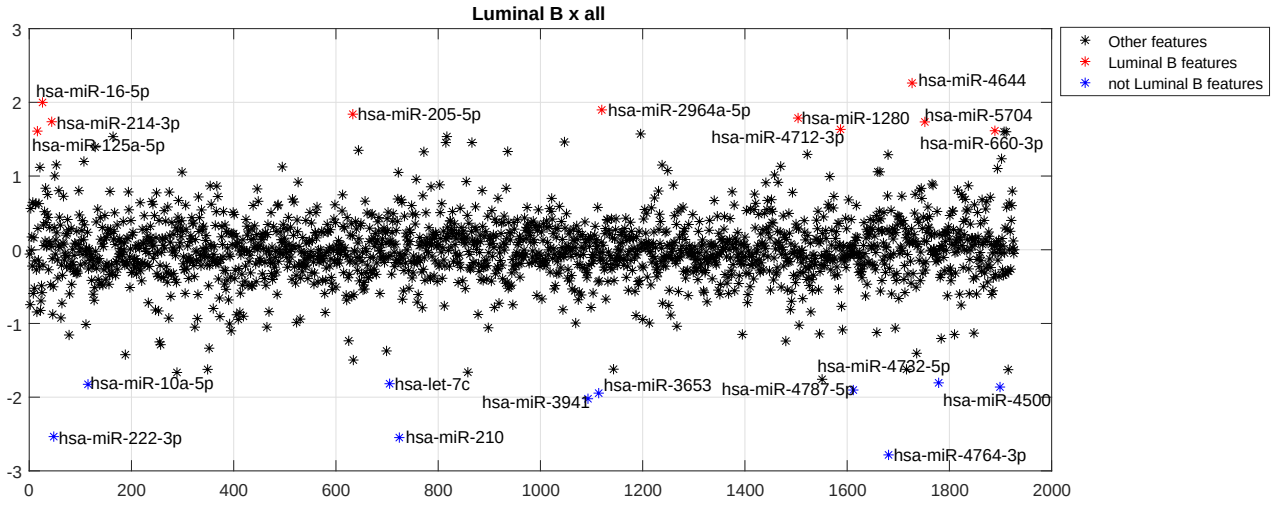

Figure A.6. Values of α computed for the Luminal B system

System HER2

Figures A.7 and A.8 present the results for **FRL** model. As a test set was used in 6 HER2 samples and 22 tissues from other sub-types of cancer and healthy tissues. The training was built with 105 tissues from healthy and cancer profiles. The **FRL** in the system HER2 attained $sensitivity = 0.85$, $specificity = 0.71$, $F_1 = 0.76$ and $AUC = 0.82$.

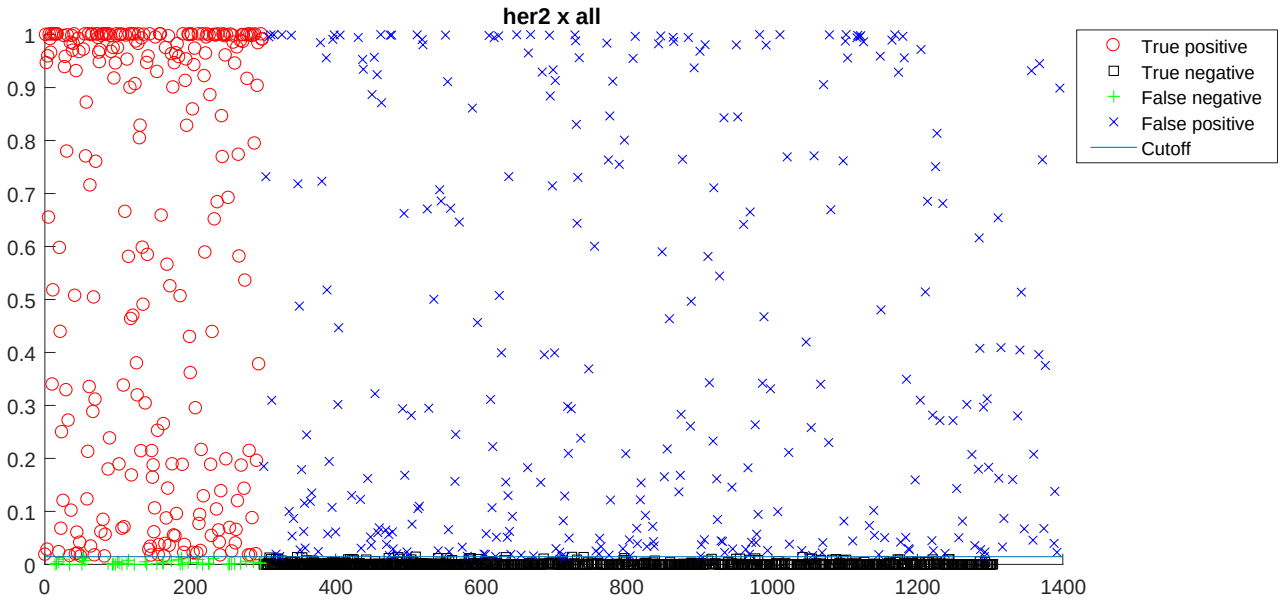

Figure A.7. $P(x)$ values computed along all 50 cross-validations (**FRL**)

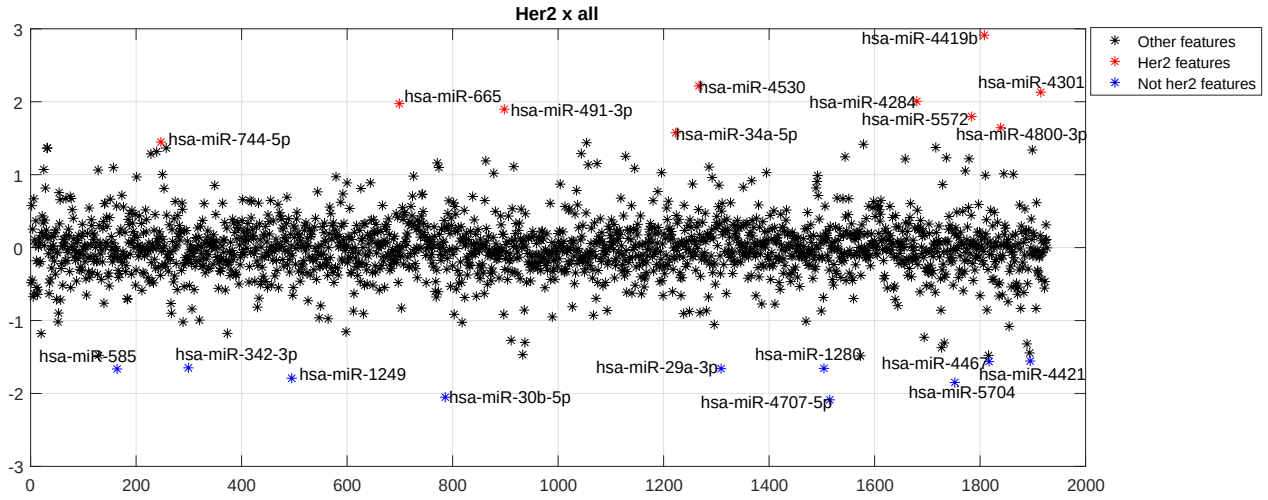

Figure A.8. Values of α computed for the Her2 system

A.2 RLR models

System TNBC

Table A.1. Selected miRNAs for **RLR**

miRNAs with $\alpha > 0$	miRNAs with $\alpha < 0$
hsa-miR-4459	hsa-miR-4284
hsa-miR-101-3p	hsa-miR-193a-3p
hsa-miR-135b-5p	hsa-miR-3607-3p
hsa-miR-135a-3p	hsa-miR-125a-5p
hsa-miR-4467	hsa-miR-5572
hsa-miR-30b-5p	hsa-miR-152
hsa-miR-4707-5p	hsa-miR-4532
hsa-miR-4708-3p	hsa-miR-342-3p
hsa-miR-4764-3p	hsa-miR-191-5p
hsa-miR-3941	hsa-miR-3940-5p

System Luminal A

Table A.2. Selected miRNAs for **RLR**

miRNAs with $\alpha > 0$	miRNAs with $\alpha < 0$
hsa-let-7c	hsa-miR-16-5p
hsa-miR-222-3p	hsa-miR-214-3p
hsa-miR-4769-5p	hsa-miR-4419b
hsa-miR-483-5p	hsa-miR-634
hsa-miR-3685	hsa-miR-4484
hsa-miR-4532	hsa-miR-3158-5p
hsa-miR-4288	hsa-miR-4800-3p
hsa-miR-4677-3p	hsa-miR-106b-5p
hsa-miR-342-3p	hsa-miR-4687-5p
hsa-miR-199a-5p	hsa-miR-187-3p

System Luminal

Table A.3. Selected miRNAs for **RLR**

miRNAs with $\alpha > 0$	miRNAs with $\alpha < 0$
hsa-miR-125a-5p	hsa-miR-4764-3p
hsa-miR-660-3p	hsa-miR-210
hsa-miR-4712-3p	hsa-miR-222-3p
hsa-miR-5704	hsa-miR-3941
hsa-miR-214-3p	hsa-miR-3653
hsa-miR-1280	hsa-miR-4787-5p
hsa-miR-205-5p	hsa-miR-4500
hsa-miR-2964a-5p	hsa-miR-10a-5p
hsa-miR-16-5p	hsa-let-7c
hsa-miR-4644	hsa-miR-4732-5p

System HER2

Table A.4. Selected miRNAs for **RLR**

miRNAs with $\alpha > 0$	miRNAs with $\alpha < 0$
hsa-miR-744-5p	hsa-miR-4707-5p
hsa-miR-34a-5p	hsa-miR-30b-5p
hsa-miR-4800-3p	hsa-miR-5704
hsa-miR-5572	hsa-miR-1249
hsa-miR-491-3p	hsa-miR-585
hsa-miR-665	hsa-miR-29a-3p
hsa-miR-4284	hsa-miR-1280
hsa-miR-4301	hsa-miR-342-3p
hsa-miR-4530	hsa-miR-4467
hsa-miR-4419b	hsa-miR-4421

A.3 ALR models

System TNBC

Figures A.9 present the results for **ARL** model. As a test set was used in 7 TNBC samples and 21 tissues from other sub-types of cancer and healthy tissues. The training was built with 105 tissues from healthy and cancer profiles. The **ARL** in the system TNBC attained $sensitivity = 0.89$, $specificity = 0.87$, $F_1 = 0.88$ and $AUC = 0.92$.

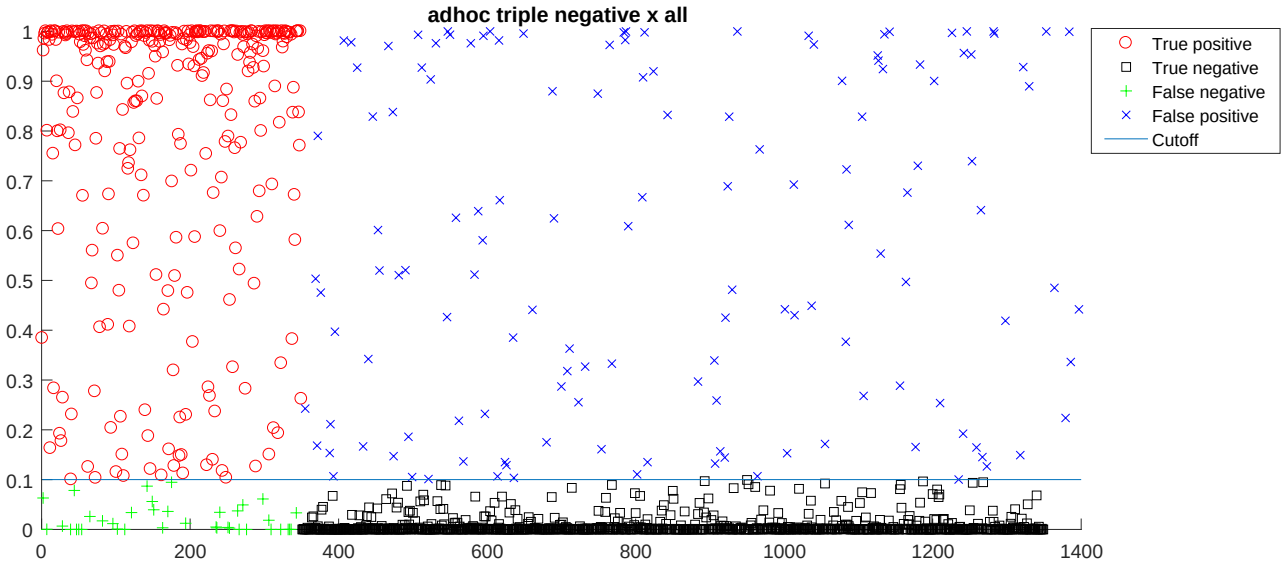

Figure A.9. $P(x)$ values computed along all 50 cross-validations (**ARL**)

System Luminal A

Figures A.10 present the results for **ARL** model. As a test set was used in 7 Luminal A samples and 21 tissues from other sub-types of cancer and healthy tissues. The training was built with 105 tissues from healthy and cancer profiles. The **ARL** in the system Luminal A attained $sensitivity = 0.91$, $specificity = 0.87$, $F_1 = 0.88$ and $AUC = 0.94$.

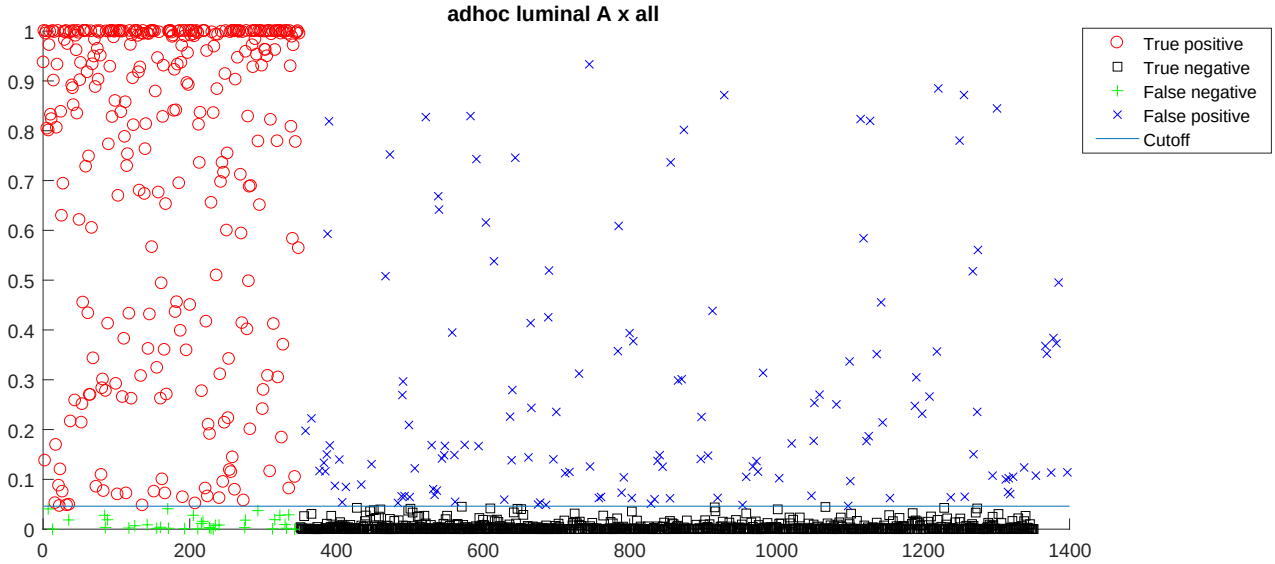

Figure A.10. $P(x)$ values computed along all 50 cross-validations (**ARL**)

System Luminal B

Figures A.11 present the results for **ARL** model. As a test set was used in 7 Luminal B samples and 20 tissues from other sub-types of cancer and healthy tissues. The training was built with 106 tissues from healthy and cancer profiles. The **ARL** in the system Luminal B attained $sensitivity = 0.86$, $specificity = 0.82$, $F_1 = 0.82$ and $AUC = 0.90$.

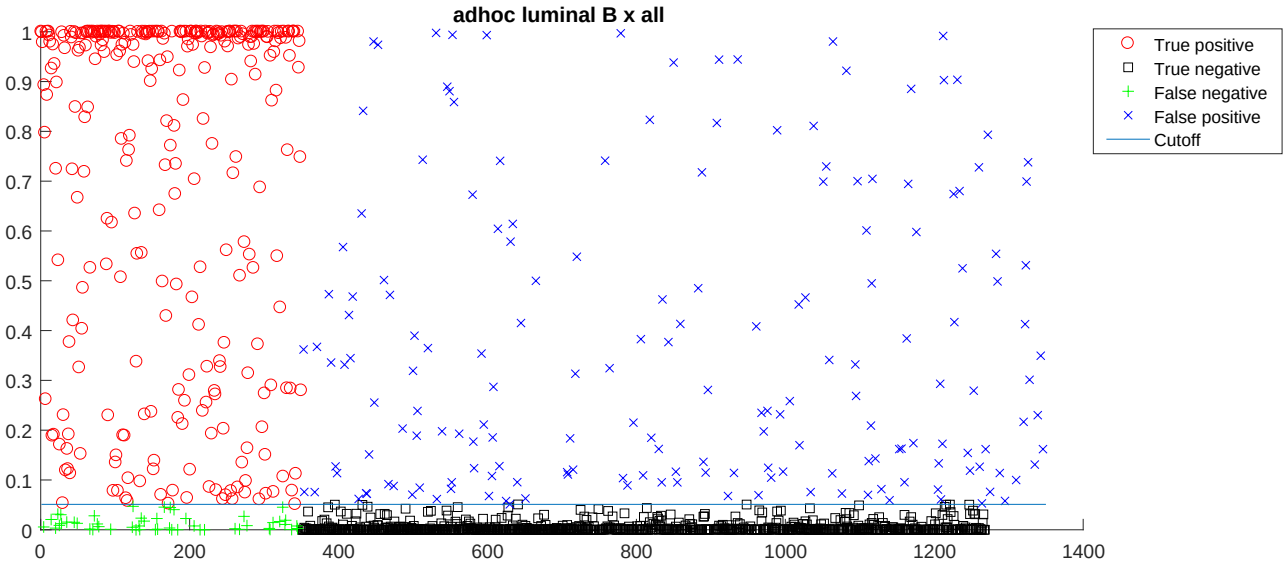

Figure A.11. $P(x)$ values computed along all 50 cross-validations (**ARL**)

System HER2

Figures A.12 present the results for **ARL** model. As a test set was used in 6 HER2 samples and 22 tissues from other sub-types of cancer and healthy tissues. The training was built with 105 tissues from healthy and cancer profiles. The **ARL** in the system HER2 attained $sensitivity = 0.79$, $specificity = 0.78$, $F_1 = 0.77$ and $AUC = 0.82$.

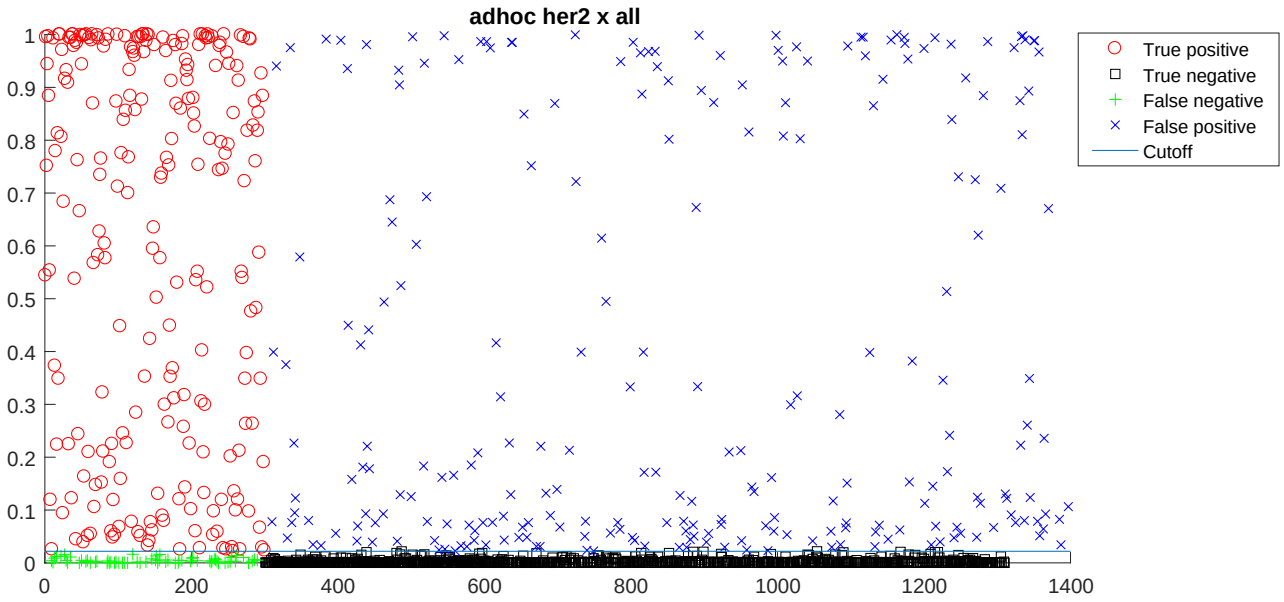

Figure A.12. $P(x)$ values computed along all 50 cross-validations (**ARL**)

Appendix B

References found in literature of the selected miRNAs

Table B.1. References found in literature of the selected 20 miRNAs used in **RLR** model for Her2 system

miRNA	Type of Cancer	References
hsa-miR-744-5p	Hepatocellular Carcinoma	[Shen et al., 2016]
hsa-miR-34a-5p	Osteosarcoma	[Pu et al., 2017]
hsa-miR-4800-3p	Colon	[Zhang et al., 2017b]
hsa-miR-4800-3p	Colon, Breast, colorectal	[Zhang et al., 2017b] [Xun et al., 2015] [Wang et al., 2017]
hsa-miR-5572		[Li et al., 2017b]
hsa-miR-491-3p	Osteosarcoma, Hepatocellular carcinoma	[Duan et al., 2017] [Zhao et al., 2017]
hsa-miR-665	Breast	[Nygren et al., 2014]
hsa-miR-4284	Prostate	[McDermott et al., 2017]
hsa-miR-4301	Colorectal	[Moshammer et al., 2014]
hsa-miR-4530	Breast	[Zhang et al., 2017a]
hsa-miR-4419b	Esophagus	[Okumura et al., 2015] [Hibino et al., 2014]
hsa-miR-4707-5p	Esophageal Squamous Cell	[Qin et al., 2016]
hsa-miR-30b-5p	Gastric	[Qiao et al., 2014]
hsa-miR-5704	Adenomas colon	[Zhang et al., 2017b]
hsa-miR-1249	Hepatocellular	[Ye et al., 2017]
hsa-miR-585	Myeloma	[Adamia et al., 2009]
hsa-miR-29a-3p	Gastric cancer	[Zhao et al., 2015]
hsa-miR-1280	Colorectal Cancer	[Huang et al., 2017]
hsa-miR-342-3p	Colon, cervical câncer	[Tao et al., 2014; Li et al., 2014]
hsa-miR-4467	Glioma	[Shou et al., 2015]
hsa-miR-4421	Colorectal	[Slattery et al., 2016]

Table B.2. References found in literature of the selected 20 miRNAs used in **RLR** model for Luminal A system

miRNA	Type of Cancer	References
hsa-let-7c	Hepatocellular	[Lan et al., 2011]
hsa-miR-222-3p	Osteoclastogenesis	[Takigawa et al., 2016]
hsa-miR-4769-5p	Breast	[Kim et al., 2016]
hsa-miR-483-5p	Lung	[Song et al., 2014]
hsa-miR-3685	Lung, Vulvar squamous cell	[Yang and Wu, 2016]
hsa-miR-4532	Gastric	[Bibi et al., 2016]
hsa-miR-4288	Colorectal	[Naccarati et al., 2012]
hsa-miR-4677-3p	Colorectal	[Wu et al., 2015]
hsa-miR-342-3p	Lung	[Tai et al., 2015]
hsa-miR-199a-5p	Breast	[Chen et al., 2016]
hsa-miR-16-5p	Bone	[Sang et al., 2017]
hsa-miR-214-3p	Bladder	[Ecke et al., 2017]
hsa-miR-4419b	Esophagus	[Okumura et al., 2015]
hsa-miR-634	Ovarian	[van Jaarsveld et al., 2015]
hsa-miR-4484	Large, B-Cell Lymphoma	[Tamaddon et al., 2016]
hsa-miR-3158-5p		[Galtsidis et al., 2017]
hsa-miR-4800-3p	Colon	[Zhang et al., 2017b]
hsa-miR-106b-5p	Glioma	[Liu et al., 2014b]
hsa-miR-4687-5p	Breast	[Maltseva et al., 2014]
hsa-miR-187-3p	Breast	[Nygren et al., 2014]

Table B.3. References found in literature of the selected 20 miRNAs used in **RLR** model for Luminal B system

miRNA	Type of Cancer	References
hsa-miR-125a-5p	Lung	[Jiang et al., 2010]
hsa-miR-660-3p	Hepatocellular	[Fu et al., 2017]
hsa-miR-4712-3p	glioblastoma	[Dong et al., 2014]
hsa-miR-5704	colon	[Zhang et al., 2017b]
hsa-miR-214-3p	Oral squamous cell carcinoma	[Yoon et al., 2014]
hsa-miR-1280	Lung	[Xu et al., 2015]
hsa-miR-205-5p	Breast	[De Cola et al., 2015]
hsa-miR-2964a-5p	Pancreatic	[Chijiiwa et al., 2016]
hsa-miR-16-5p	Gastric	[Zhang et al., 2015a]
hsa-miR-4644	Pancreatobiliary tract	[Machida et al., 2016]
hsa-miR-4764-3p	Oropharyngeal	[Mirghani et al., 2016]
hsa-miR-210	Breast	[Camps et al., 2008]
hsa-miR-222-3p	Endometrial	[Liu et al., 2014a]
hsa-miR-3941	Lung	[Sato et al., 2017]
hsa-miR-3653	Breast	[Li et al., 2013b]
hsa-miR-4787-5p	Pancreatic	[Mody et al., 2016]
hsa-miR-4500	Lung	[Zhang et al., 2014]
hsa-miR-10a-5p	Gastric	[Lu et al., 2017]
hsa-let-7c	HepG2 cancer cells	[Wen et al., 2009]
hsa-miR-4732-5p	Breast	[Persson et al., 2011]

Table B.4. References found in literature of the selected 20 miRNAs used in **RLR** model for TNBC system

miRNA	Type of Cancer	References
hsa-miR-4459	Lung	[Zhou et al., 2016]
hsa-miR-101-3p	Hepatocellular	[Li et al., 2017a]
hsa-miR-135b-5p	Osteosarcoma	[Lauvrak et al., 2013]
hsa-miR-135a-3p	Ovarian	[Palmenberg, 1987]
hsa-miR-4467	Endometrial	[Canlorbe et al., 2016]
hsa-miR-30b-5p	Renal	[Liu et al., 2017]
hsa-miR-4707-5p	Esophageal Squamous Cell	[Qin et al., 2016]
hsa-miR-4708-3p	Pancreatic	[Madhavan et al., 2015]
hsa-miR-4764-3p	Oropharyngeal	[Mirghani et al., 2016]
hsa-miR-3941	Lymphoblastic Leukemia	[Lu et al., 2015]
hsa-miR-4284	Renal cell	[Munari et al., 2014]
hsa-miR-193a-3p	Colorectal	[Yong et al., 2013]
hsa-miR-3607-3p	Gastric	[Xie et al., 2014]
hsa-miR-125a-5p	Hepatocellular	[Kim et al., 2013]
hsa-miR-5572		[Kondou et al., 2015]
hsa-miR-152		[Liu et al., 2016]
hsa-miR-4532	Gastric	[Bibi et al., 2016]
hsa-miR-342-3p	Lung	[Tai et al., 2015]
hsa-miR-191-5p	Colorectal	[Zhang et al., 2015b]
hsa-miR-3940-5p	Lung	[Ren et al., 2017]

Table B.5. Other miRNA references

miRNA	Type of Cancer	References
miR-125b-5p	Hepatocelular, Cutaneous T-cell lymphomas	[Giray et al., 2014] [Manfe et al., 2013]
miR-3613-3p	Neuroblastoma	[Zheng et al., 2016]
miR-4668-5p	Colorretal	[Moshammer et al., 2014]
miR-3656	Gastric	[Piazera, 2016]
miR-5704	Colon Adenomas, Breast	[Zhang et al., 2017b] [Wang et al., 2017]
miR-3676-3p	Gastric	[Estrêla, 2014]
miR-3196	Lung	[Andäng et al., 2008] [Xu et al., 2016]
miR-3941	Lung	[Sato et al., 2017]
miR-585	Lung	[Ding et al., 2017]
miR-1264	Larynx	[Xu et al., 2013]
miR-200a-3p	Hepatocelular	[Li et al., 2016b]
miR-1273g-3p	Glioblastoma	[Dong et al., 2014]
miR-5581-3p	Colon Adenomas, Breast	[Zhang et al., 2017b] [Wang et al., 2017]
miR-877-5p	Leukemia	[Feng et al., 2013]
miR-96-5p	Breast, Colorectal	[Calvano Filho et al., 2014] [Ress et al., 2015]
miR-744-3p	Laryngealsquamouscell carcinoma	[Li et al., 2016a]
miR-2276	Breast	[Torkashvand et al., 2016]
miR-342-5p	Breast	[Cittelly et al., 2010]
miR-760	Breast	[Lv et al., 2015]
miR-203	Breast	[Zhang et al., 2011]
miRPlus-A1086	Prostate	[McDermott et al., 2017]
miR-185-5p	Breast	[Wang et al., 2014]
miR-20b-5p	Breast	[Li et al., 2013a]
miR-4521	leukemia / lymphoma 1	[Pekarsky et al., 2016]
miR-4692	EsophagealSquamousCell	[Qin et al., 2016]

Appendix C

Visualization of the samples projected from their representation using 20 miRNA

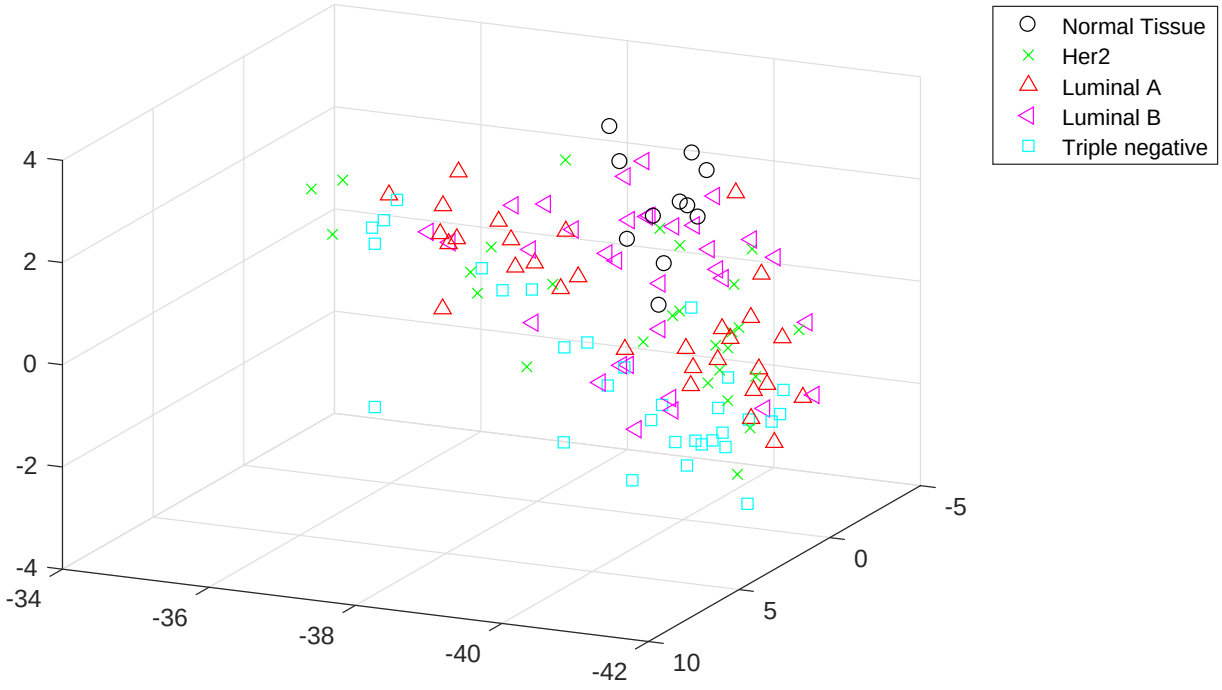

Figure C.1. Types of tissues represented by selected 20 miRNAs from 1929

Visualization of the distribution of tissue types analyzed using selected 20 miRNAs. The vectors in space with dimension 6 where projected in space R^3 using the method descibed by [Marcolino et al., 2010]

Bibliography

(2015). *Cancer Registration Statistics, England; 2015*. Office for National Statistics.

(2015). *GEO: the Gene Expression Omnibus; Series GSE58606*. National Center for Biotechnology Information.

(2015). *New Analysis of Breast Cancer Subtypes Could Lead to Better Risk Strati?cation*. Centers for Disease Control and Prevention.

(2017). *Cancer Facts and Statistics; 2017*. American Cancer Society.

Adamia, S., Fulciniti, M., Avet-Loiseau, H., Amin, S. B., Shah, P., Carrasco, D. R., Minvielle, S., Moreau, P., Anderson, K. C., and Munshi, N. C. (2009). Biological and therapeutic potential of mir-155, 585 and let-7f in myeloma in vitro and in vivo.

Andäng, M., Hjerling-Leffler, J., Moliner, A., Lundgren, T. K., Castelo-Branco, G., Nanou, E., Pozas, E., Bryja, V., Halliez, S., Nishimaru, H., et al. (2008). Histone h2ax-dependent gabaa receptor regulation of stem cell proliferation. *Nature*, 451(7177):460.

Bellman, R. E. (2015). *Adaptive control processes: a guided tour*. Princeton university press.

Berry, M. W., Dumais, S. T., and O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM review*, 37(4):573--595.

Bertero, T., Grosso, S., Robbe-Sermesant, K., Lebrigand, K., Hénaoui, I.-S., Puisségur, M.-P., Fourre, S., Zaragosi, L.-E., Mazure, N. M., Ponzio, G., et al. (2012). "seed-milarity" confers to hsa-mir-210 and hsa-mir-147b similar functional activity. *PLoS One*, 7(9):e44919.

Bibi, F., Naseer, M. I., Alvi, S. A., Yasir, M., Jiman-Fatani, A. A., Sawan, A., Abuzenadah, A. M., Al-Qahtani, M. H., and Azhar, E. I. (2016). microrna analysis of gastric cancer patients from saudi arabian population. *BMC genomics*, 17(9):751.

Brentani, R. R., Carraro, D. M., Verjovski-Almeida, S., Reis, E. M., Neves, E. J., de Souza, S. J., Carvalho, A. F., Brentani, H., and Reis, L. F. (2005). Gene expression arrays in cancer research: methods and applications. *Critical reviews in oncology/hematology*, 54(2):95--105.

Calvano Filho, C. M. C., Calvano-Mendes, D. C., Carvalho, K. C., Maciel, G. A., Ricci, M. D., Torres, A. P., Filassi, J. R., and Baracat, E. C. (2014). Triple-negative and luminal a breast tumors: differential expression of mir-18a-5p, mir-17-5p, and mir-20a-5p. *Tumor Biology*, 35(8):7733--7741.

Camps, C., Buffa, F. M., Colella, S., Moore, J., Sotiriou, C., Sheldon, H., Harris, A. L., Gleadle, J. M., and Ragoussis, J. (2008). hsa-mir-210 is induced by hypoxia and is an independent prognostic factor in breast cancer. *Clinical cancer research*, 14(5):1340--1348.

Canlorbe, G., Wang, Z., Laas, E., Bendifallah, S., Castela, M., Lefevre, M., Chabbert-Buffet, N., Daraï, E., Aractingi, S., Méhats, C., et al. (2016). Identification of microrna expression profile related to lymph node status in women with early-stage grade 1–2 endometrial cancer. *Modern Pathology*, 29(4):391--401.

Cetin, I. and Topcul, M. (2014). Triple negative breast cancer. *Asian Pac J Cancer Prev*, 15(6):2427--2431.

Chen, J., Shin, V. Y., Siu, M. T., Ho, J. C., Cheuk, I., and Kwong, A. (2016). mir-199a-5p confers tumor-suppressive role in triple-negative breast cancer. *BMC cancer*, 16(1):887.

Chijiiwa, Y., Moriyama, T., Ohuchida, K., Nabae, T., Ohtsuka, T., Miyasaka, Y., Fujita, H., Maeyama, R., Manabe, T., Abe, A., et al. (2016). Overexpression of microrna-5100 decreases the aggressive phenotype of pancreatic cancer cells by targeting podxl. *International journal of oncology*, 48(4):1688--1700.

Cittelly, D. M., Das, P. M., Spoelstra, N. S., Edgerton, S. M., Richer, J. K., Thor, A. D., and Jones, F. E. (2010). Downregulation of mir-342 is associated with tamoxifen resistant breast tumors. *Molecular cancer*, 9(1):317.

da Cunha, J. P. C., Galante, P. A. F., de Souza, J. E. S., Pieprzyk, M., Carraro, D. M., Old, L. J., Camargo, A. A., and de Souza, S. J. (2013). The human cell surfaceome of breast tumors. *BioMed research international*, 2013.

De Cola, A., Volpe, S., Budani, M., Ferracin, M., Lattanzio, R., Turdo, A., D'agostino, D., Capone, E., Stassi, G., Todaro, M., et al. (2015). mir-205-5p-mediated downregulation of erbb/her receptors in breast cancer stem cells results in targeted therapy resistance. *Cell death & disease*, 6(7):e1823.

Ding, X., Yang, Y., Sun, Y., Xu, W., Su, B., and Zhou, X. (2017). Microrna-585 acts as a tumor suppressor in non-small-cell lung cancer by targeting hsmg-1. *Clinical and Translational Oncology*, 19(5):546--552.

Dong, L., Li, Y., Han, C., Wang, X., She, L., and Zhang, H. (2014). mirna microarray reveals specific expression in the peripheral blood of glioblastoma patients. *International journal of oncology*, 45(2):746--756.

Drahos, J., Schwameis, K., Orzolek, L. D., Hao, H., Birner, P., Taylor, P. R., Pfeiffer, R. M., Schoppmann, S. F., and Cook, M. B. (2015). Microrna profiles of barrett9s esophagus and esophageal adenocarcinoma: Differences in glandular non-native epithelium. *Cancer Epidemiology and Prevention Biomarkers*, pages cebp--0161.

Duan, J., Liu, J., Liu, Y., Huang, B., and Rao, L. (2017). mir-491-3p suppresses the growth and invasion of osteosarcoma cells by targeting tspan1. *Molecular Medicine Reports*, 16(4):5568--5574.

Ecke, T. H., Stier, K., Weickmann, S., Zhao, Z., Buckendahl, L., Stephan, C., Kilic, E., and Jung, K. (2017). mir-199a-3p and mir-214-3p improve the overall survival prediction of muscle-invasive bladder cancer patients after radical cystectomy. *Cancer Medicine*.

Eilers, P. H., Boer, J. M., van Ommen, G.-J., and van Houwelingen, H. C. (2001). Classification of microarray data with penalized logistic regression. In *BiOS 2001 The International Symposium on Biomedical Optics*, pages 187--198. International Society for Optics and Photonics.

Eldén, L. (2006). Numerical linear algebra in data mining. *Acta Numerica*, 15:327--384.

Estrêla, M. S. (2014). Análise do perfil de expressão gênica de metiltransferases proteicas no câncer colorretal.

Feng, D.-Q., Huang, B., Li, J., Liu, J., Chen, X.-M., Xu, Y.-M., Chen, X., Zhang, H.-B., Hu, L.-H., and Wang, X.-Z. (2013). Selective mirna expression profile in chronic myeloid leukemia k562 cell-derived exosomes. *Asian Pacific journal of cancer prevention*, 14(12):7501--7508.

Fort, G. and Lambert-Lacroix, S. (2004). Classification using partial least squares with penalized logistic regression. *Bioinformatics*, 21(7):1104--1111.

Fu, L., Yao, T., Chen, Q., Mo, X., Hu, Y., and Guo, J. (2017). Screening differential circular rna expression profiles reveals hsa_circ_0004018 is associated with hepatocellular carcinoma. *Oncotarget*, 8(35):58405.

Fukagawa, S., Miyata, K., Yotsumoto, F., Kiyoshima, C., Nam, S. O., Anan, H., Katsuda, T., Miyahara, D., Murata, M., Yagi, H., et al. (2017). Microrna-135a-3p as a promising biomarker and nucleic acid therapeutic agent for ovarian cancer. *Cancer Science*.

Galtsidis, S., Logotheti, S., Pavlopoulou, A., Zampetidis, C. P., Papachristopoulou, G., Scorilas, A., Vojtesek, B., Gorgoulis, V., and Zoumpourlis, V. (2017). Unravelling a p73-regulated network: The role of a novel p73-dependent target, mir3158, in cancer cell migration and invasiveness. *Cancer letters*, 388:96--106.

Ge, Q.-M., Huang, C.-M., Zhu, X.-Y., Bian, F., and Pan, S.-M. (2017). Differentially expressed mirnas in sepsis-induced acute kidney injury target oxidative stress and mitochondrial dysfunction pathways. *PloS one*, 12(3):e0173292.

George-Nektarios, T. (2013). Weka classifiers summary. *Athens University of Economics and Bussiness Intracom-Telecom, Athens*.

Giancarlo, R., Bosco, G. L., and Pinello, L. (2010). Distance functions, clustering algorithms and microarray data analysis. *LION*, 4:125--138.

Giray, B. G., Emekdas, G., Tezcan, S., Ulger, M., Serin, M. S., Sezgin, O., Altintas, E., and Tiftik, E. N. (2014). Profiles of serum microrna; mir-125b-5p and mir223-3p serve as novel biomarkers for hbv-positive hepatocellular carcinoma. *Molecular biology reports*, 41(7):4513--4519.

Golub, G. (1965). Numerical methods for solving linear least squares problems. *Numerische Mathematik*, 7(3):206--216.

Hashiguchi, Y., Nishida, N., Mimori, K., Sudo, T., Tanaka, F., Shibata, K., Ishii, H., Mochizuki, H., Hase, K., Doki, Y., et al. (2012). Down-regulation of mir-125a-3p in human gastric cancer and its clinicopathological significance. *International journal of oncology*, 40(5):1477--1482.

Hibino, S., Saito, Y., Muramatsu, T., Otani, A., Kasai, Y., Kimura, M., and Saito, H. (2014). Inhibitors of enhancer of zeste homolog 2 (ezh2) activate tumor-suppressor micrornas in human cancer cells. *Oncogenesis*, 3(5):e104.

Horn, D. and Axel, I. (2003). Novel clustering algorithm for microarray expression data in a truncated svd space. *Bioinformatics*, 19(9):1110--1115.

Huang, B., Yang, H., Cheng, X., Wang, D., Fu, S., Shen, W., Zhang, Q., Zhang, L., Xue, Z., Li, Y., et al. (2017). trf/mir-1280 suppresses stem cell–like cells and metastasis in colorectal cancer. *Cancer Research*, 77(12):3194--3206.

Irvin, W. J. and Carey, L. A. (2008). What is triple-negative breast cancer? *European journal of cancer*, 44(18):2799--2805.

Jiang, L., Huang, Q., Zhang, S., Zhang, Q., Chang, J., Qiu, X., and Wang, E. (2010). Hsa-mir-125a-3p and hsa-mir-125a-5p are downregulated in non-small cell lung cancer and have inverse effects on invasion and migration of lung cancer cells. *BMC cancer*, 10(1):318.

Jones, K., Nourse, J. P., Keane, C., Bhatnagar, A., and Gandhi, M. K. (2013). Plasma microrna are disease response biomarkers in classical hodgkin lymphoma. *Clinical cancer research*.

Jung, C. K., Jung, S.-H., Yim, S.-H., Jung, J.-H., Choi, H. J., Kang, W.-K., Park, S.-W., Oh, S.-T., Kim, J.-G., Lee, S. H., et al. (2016). Predictive micrornas for lymph node metastasis in endoscopically resectable submucosal colorectal cancer. *Oncotarget*, 7(22):32902.

Kim, B. G., Kang, S., Han, H. H., Lee, J. H., Kim, J. E., Lee, S. H., and Cho, N. H. (2016). Transcriptome-wide analysis of compression-induced microrna expression alteration in breast cancer for mining therapeutic targets. *Oncotarget*, 7(19):27468.

Kim, J. K., Noh, J. H., Jung, K. H., Eun, J. W., Bae, H. J., Kim, M. G., Chang, Y. G., Shen, Q., Park, W. S., Lee, J. Y., et al. (2013). Sirtuin7 oncogenic potential in human hepatocellular carcinoma and its regulation by the tumor suppressors mir-125a-5p and mir-125b. *Hepatology*, 57(3):1055--1067.

Kleemann, M., Bereuther, J., Fischer, S., Marquart, K., Hänle, S., Unger, K., Jendrossek, V., Riedel, C. U., Handrick, R., and Otte, K. (2017). Investigation on tissue specific effects of pro-apoptotic micro rnas revealed mir-147b as a potential biomarker in ovarian cancer prognosis. *Oncotarget*, 8(12):18773.

Kondou, S., Nobumasa, H., Kozono, S., Hiroko, S., Kawauchi, J., Ochiya, T., and Kosaka, N. (2015). Prostate cancer detection kit or device, and detection method. US Patent App. 15/317,882.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8).

Lan, F.-F., Wang, H., Chen, Y.-C., Chan, C.-Y., Ng, S. S., Li, K., Xie, D., He, M.-L., Lin, M. C., and Kung, H.-F. (2011). Hsa-let-7g inhibits proliferation of hepatocellular carcinoma cells by downregulation of c-myc and upregulation of p16ink4a. *International Journal of Cancer*, 128(2):319--331.

Lauvrak, S., Munthe, E., Kresse, S., Stratford, E., Namløs, H., Meza-Zepeda, L., and Myklebost, O. (2013). Functional characterisation of osteosarcoma cell lines and identification of mrnas and mirnas associated with aggressive cancer phenotypes. *British journal of cancer*, 109(8):2228.

Leotta, M., Biamonte, L., Raimondi, L., Ronchetti, D., Di Martino, M. T., Botta, C., Leone, E., Pitari, M. R., Neri, A., Giordano, A., et al. (2014). A p53-dependent tumor suppressor network is induced by selective mir-125a-5p inhibition in multiple myeloma cells. *Journal of cellular physiology*, 229(12):2106--2116.

Li, C.-Y., Pang, Y.-Y., Yang, H., Li, J., Lu, H.-X., Wang, H.-L., Mo, W.-J., Huang, L.-S., Feng, Z.-B., and Chen, G. (2017a). Identification of mir-101-3p targets and functional features based on bioinformatics, meta-analysis and experimental verification in hepatocellular carcinoma. *American journal of translational research*, 9(5):2088.

Li, D., Ilnytskyy, Y., Kovalchuk, A., Khachigian, L. M., Bronson, R. T., Wang, B., and Kovalchuk, O. (2013a). Crucial role for early growth response-1 in the transcriptional regulation of mir-20b in breast cancer. *Oncotarget*, 4(9):1373.

Li, J.-Y., Jia, S., Zhang, W.-H., Zhang, Y., Kang, Y., and Li, P.-S. (2013b). Differential distribution of micrornas in breast cancer grouped by clinicopathological subtypes. *Asian Pacific Journal of Cancer Prevention*, 14(5):3197--3203.

Li, J. Z.-H., Gao, W., Lei, W.-B., Zhao, J., Chan, J. Y.-W., Wei, W. I., Ho, W.-K., and Wong, T.-S. (2016a). Microrna 744-3p promotes mmp-9-mediated metastasis by simultaneously suppressing pdcd4 and pten in laryngeal squamous cell carcinoma. *Oncotarget*, 7(36):58218.

Li, L.-Z., Zhang, C. Z., Liu, L.-L., Yi, C., Lu, S.-X., Zhou, X., Zhang, Z.-J., Peng, Y.-H., Yang, Y.-Z., and Yun, J.-P. (2013c). mir-720 inhibits tumor invasion and migration in breast cancer by targeting twist1. *Carcinogenesis*, 35(2):469--478.

Li, P., Zhai, P., Ye, Z., Deng, P., Fan, Y., Zeng, Y., Pang, Z., Zeng, J., Li, J., and Feng, W. (2017b). Differential expression of mir-195-5p in collapse of steroid-induced osteonecrosis of the femoral head. *Oncotarget*, 8(26):42638.

Li, S.-P., Xu, H.-X., Yu, Y., He, J.-D., Wang, Z., Xu, Y.-J., Wang, C.-Y., Zhang, H.-M., Zhang, R.-X., Zhang, J.-J., et al. (2016b). Lncrna hulc enhances epithelial-mesenchymal transition to promote tumorigenesis and metastasis of hepatocellular carcinoma via the mir-200a-3p/zeb1 signaling pathway. *Oncotarget*, 7(27):42431.

Li, X.-r., Chu, H.-j., Lv, T., Wang, L., Kong, S.-f., and Dai, S.-z. (2014). mir-342-3p suppresses proliferation, migration and invasion by targeting foxm1 in human cervical cancer. *FEBS letters*, 588(17):3298--3307.

Linnik, U. V., Elandt-Johnson, R. C., and Johnson, N. L. (1961). Method of least squares and principles of the theory of observations.

Liu, B., Che, Q., Qiu, H., Bao, W., Chen, X., Lu, W., Li, B., and Wan, X. (2014a). Elevated mir-222-3p promotes proliferation and invasion of endometrial carcinoma via targeting erα. *PloS one*, 9(1):e87563.

Liu, F., Gong, J., Huang, W., Wang, Z., Wang, M., Yang, J., Wu, C., Wu, Z., and Han, B. (2014b). Microrna-106b-5p boosts glioma tumorigensis by targeting multiple tumor suppressor genes. *Oncogene*, 33(40):4813.

Liu, W., Li, H., Wang, Y., Zhao, X., Guo, Y., Jin, J., and Chi, R. (2017). Mir-30b-5p functions as a tumor suppressor in cell proliferation, metastasis and epithelial-to-mesenchymal transition by targeting g-protein subunit α-13 in renal cell carcinoma. *Gene*.

Liu, X., Li, J., Qin, F., and Dai, S. (2016). mir-152 as a tumor suppressor microrna: Target recognition and regulation in cancer. *Oncology letters*, 11(6):3911--3916.

Lopes-Ramos, C. M., Habr-Gama, A., de Souza Quevedo, B., Felício, N. M., Bettoni, F., Koyama, F. C., Asprino, P. F., Galante, P. A., Gama-Rodrigues, J., Camargo, A. A., et al. (2014). Overexpression of mir-21-5p as a predictive marker for complete tumor regression to neoadjuvant chemoradiotherapy in rectal cancer patients. *BMC medical genomics*, 7(1):68.

Lu, L., Wang, F., He, L., Xue, Y., Wang, Y., Zhang, H., Rong, L., Wang, M., Zhang, Z., Fang, Y., et al. (2015). Interaction between igf1 polymorphisms and the risk of acute lymphoblastic leukemia in chinese children. *Cellular Physiology and Biochemistry*, 36(4):1346--1358.

Lu, Y., Wei, G., Liu, L., Mo, Y., Chen, Q., Xu, L., Liao, R., Zeng, D., and Zhang, K. (2017). Direct targeting of mapk8ip1 by mir-10a-5p is a major mechanism for gastric cancer metastasis. *Oncology Letters*, 13(3):1131--1136.

Lv, J., Fu, Z., Shi, M., Xia, K., Ji, C., Xu, P., Lv, M., Pan, B., Dai, L., and Xie, H. (2015). Systematic analysis of gene expression pattern in has-mir-760 overexpressed resistance of the mcf-7 human breast cancer cell to doxorubicin. *Biomedicine & Pharmacotherapy*, 69:162--169.

Machida, T., Tomofuji, T., Maruyama, T., Yoneda, T., Ekuni, D., Azuma, T., Miyai, H., Mizuno, H., Kato, H., Tsutsumi, K., et al. (2016). mir-1246 and mir-4644 in salivary exosome as potential biomarkers for pancreatobiliary tract cancer. *Oncology reports*, 36(4):2375--2381.

Madhavan, B., Yue, S., Galli, U., Rana, S., Gross, W., Müller, M., Giese, N. A., Kalthoff, H., Becker, T., Büchler, M. W., et al. (2015). Combined evaluation of a panel of protein and mirna serum-exosome biomarkers for pancreatic cancer diagnosis increases sensitivity and specificity. *International journal of cancer*, 136(11):2616--2627.

Maltseva, D. V., Galatenko, V. V., Samatov, T. R., Zhikrivetskaya, S. O., Khaustova, N. A., Nechaev, I. N., Shkurnikov, M. U., Lebedev, A. E., Mityakina, I. A., Kaprin, A. D., et al. (2014). mirnome of inflammatory breast cancer. *BMC research notes*, 7(1):871.

Manfe, V., Biskup, E., Willumsgaard, A., Skov, A. G., Palmieri, D., Gasparini, P., Lagana, A., Woetmann, A., Ødum, N., Croce, C. M., et al. (2013). cmyc/mir-125b-5p signalling determines sensitivity to bortezomib in preclinical model of cutaneous t-cell lymphomas. *PloS one*, 8(3):e59390.

Marcolino, L., Couto, B., and dos Santos, M. (2010). Genome visualization in space. *Advances in Bioinformatics*, pages 225--232.

McDermott, N., Meunier, A., Wong, S., Buchete, V., and Marignol, L. (2017). Profiling of a panel of radioresistant prostate cancer cells identifies deregulation of key mirnas. *Clinical and Translational Radiation Oncology*, 2:63--68.

Menard, S. (2010). *Logistic regression: From introductory to advanced concepts and applications*. Sage.

Mirghani, H., Ugolin, N., Ory, C., Goislard, M., Lefèvre, M., Baulande, S., Hofman, P., Guily, J. L. S., Chevillard, S., and Lacave, R. (2016). Comparative analysis of micrornas in human papillomavirus–positive versus–negative oropharyngeal cancers. *Head & neck*, 38(11):1634--1642.

Mody, H., Hung, S. W., Al-Saggar, M., Griffin, J., and Govindarajan, R. (2016). Inhibition of s-adenosylmethionine-dependent methyltransferase attenuates tgf-beta1-induced emt and metastasis in pancreatic cancer: Putative roles of mir-663a and mir-4787-5p. *Molecular Cancer Research*, pages molcanres--0083.

Moshammer, M. I., Kalipciyan, M., Offner, F., Sterlacci, W., Steger, G. G., Mader, R. M., and Sedivy, R. (2014). microrna expression profiles distinguish colorectal cancer patients in two regions of austria. *International journal of clinical pharmacology and therapeutics*, 52(1):85--86.

Mullany, L. E., Herrick, J. S., Wolff, R. K., Stevens, J. R., and Slattery, M. L. (2016). Association of cigarette smoking and microrna expression in rectal cancer: Insight into tumor phenotype. *Cancer epidemiology*, 45:98--107.

Munari, E., Marchionni, L., Chitre, A., Hayashi, M., Martignoni, G., Brunelli, M., Gobbo, S., Argani, P., Allaf, M., Hoque, M. O., et al. (2014). Clear cell papillary renal cell carcinoma: micro-rna expression profiling and comparison with clear cell renal cell carcinoma and papillary renal cell carcinoma. *Human pathology*, 45(6):1130--1138.

Naccarati, A., Pardini, B., Stefano, L., Landi, D., Slyskova, J., Novotny, J., Levy, M., Polakova, V., Lipska, L., and Vodicka, P. (2012). Polymorphisms in mirna-binding sites of nucleotide excision repair genes and colorectal cancer risk. *Carcinogenesis*, 33(7):1346--1351.

Nguyen, D. V. and Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1):39--50.

Nygren, M., Tekle, C., Ingebrigtsen, V., Mäkelä, R., Krohn, M., Aure, M., Nunes-Xavier, C., Perälä, M., Tramm, T., Alsner, J., et al. (2014). Identifying micrornas regulating b7-h3 in breast cancer: the clinical impact of microrna-29c. *British journal of cancer*, 110(8):2072.

Okumura, T., Shimada, Y., Omura, T., Hirano, K., Nagata, T., and Tsukada, K. (2015). Microrna profiles to predict postoperative prognosis in patients with small cell carcinoma of the esophagus. *Anticancer research*, 35(2):719--727.

Omrane, I., Kourda, N., Stambouli, N., Privat, M., Medimegh, I., Arfaoui, A., Uhrhammer, N., Bougatef, K., Baroudi, O., Bouzaienne, H., et al. (2014). Micrornas 146a and 147b biomarkers for colorectal tumor's localization. *BioMed research international*, 2014.

Palmenberg, A. (1987). A vaccine for the common cold? *Nature*, 329:668--669.

Pekarsky, Y., Balatti, V., Palamarchuk, A., Rizzotto, L., Veneziano, D., Nigita, G., Rassenti, L. Z., Pass, H. I., Kipps, T. J., Liu, C.-g., et al. (2016). Dysregulation of a family of short noncoding rnas, tsrnas, in human cancer. *Proceedings of the National Academy of Sciences*, 113(18):5071--5076.

Persson, H., Kvist, A., Rego, N., Staaf, J., Vallon-Christersson, J., Luts, L., Loman, N., Jonsson, G., Naya, H., Hoglund, M., et al. (2011). Identification of new micrornas in paired normal and tumor breast tissue suggests a dual role for the erbb2/her2 gene. *Cancer research*, 71(1):78--86.

Piazera, F. Z. (2016). Análise de expressão de genes da família setd em leucemia linfocítica crônica.

Pigati, L., Yaddanapudi, S. C., Iyengar, R., Kim, D.-J., Hearn, S. A., Danforth, D., Hastings, M. L., and Duelli, D. M. (2010). Selective release of microrna species from normal and malignant mammary epithelial cells. *PloS one*, 5(10):e13515.

Powell, J. R., Bennett, M. R., Evans, K. E., Yu, S., Webster, R. M., Waters, R., Skinner, N., and Reed, S. H. (2015). 3d-dip-chip: a microarray-based method to measure genomic dna damage. *Scientific reports*, 5.

Pu, Y., Zhao, F., Li, Y., Cui, M., Wang, H., Meng, X., and Cai, S. (2017). The mir-34a-5p promotes the multi-chemoresistance of osteosarcoma via repression of the agtr1 gene. *BMC cancer*, 17(1):45.

Qiao, F., Zhang, K., Gong, P., Wang, L., Hu, J., Lu, S., and Fan, H. (2014). Decreased mir-30b-5p expression by dnmt1 methylation regulation involved in gastric cancer metastasis. *Molecular biology reports*, 41(9):5693--5700.

Qin, H.-D., Liao, X.-Y., Chen, Y.-B., Huang, S.-Y., Xue, W.-Q., Li, F.-F., Ge, X.-S., Liu, D.-Q., Cai, Q., Long, J., et al. (2016). Genomic characterization of esophageal squamous cell carcinoma reveals critical genes underlying tumorigenesis and poor prognosis. *The American Journal of Human Genetics*, 98(4):709--727.

Ren, K., Li, Y., Lu, H., Li, Z., and Han, X. (2017). mir-3940-5p functions as a tumor suppressor in non–small cell lung cancer cells by targeting cyclin d1 and ubiquitin specific peptidase-28. *Translational oncology*, 10(1):80--89.

Ress, A. L., Stiegelbauer, V., Winter, E., Schwarzenbacher, D., Kiesslich, T., Lax, S., Jahn, S., Deutsch, A., Bauernhofer, T., Ling, H., et al. (2015). Mir-96-5p influences cellular growth and is associated with poor survival in colorectal cancer patients. *Molecular carcinogenesis*, 54(11):1442--1450.

Ronchetti, D., Manzoni, M., Todoerti, K., Neri, A., and Agnelli, L. (2016). In silico characterization of mirna and long non-coding rna interplay in multiple myeloma. *Genes*, 7(12):107.

Sang, S., Zhang, Z., Qin, S., Li, C., and Dong, Y. (2017). Microrna-16-5p inhibits osteoclastogenesis in giant cell tumor of bone. *BioMed Research International*, 2017.

Sato, T., Shiba-Ishii, A., Kim, Y., Dai, T., Husni, R. E., Hong, J., Kano, J., Sakashita, S., Iijima, T., and Noguchi, M. (2017). mir-3941: A novel microrna that controls igbp1 expression and is associated with malignant progression of lung adenocarcinoma. *Cancer science*, 108(3):536--542.

Schena, M., Shalon, D., Davis, R. W., Brown, P. O., et al. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *SCIENCE-NEW YORK THEN WASHINGTON-*, pages 467--467.

Schreiber, R., Mezencev, R., Matyunina, L., and McDonald, J. (2016). Evidence for the role of microrna 374b in acquired cisplatin resistance in pancreatic cancer cells. *Cancer gene therapy*, 23(8):241.

Shen, L. and Tan, E. C. (2005). Dimension reduction-based penalized logistic regression for cancer classification using microarray data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2(2):166--175.

Shen, S., Lin, Y., Yuan, X., Shen, L., Chen, J., Chen, L., Qin, L., and Shen, B. (2016). Biomarker microrna for diagnosis, prognosis and treatment of hepatocellular carcinoma: a functional survey and comparison. *Scientific reports*, 6.

Shou, J., Gu, S., and Gu, W. (2015). Identification of dysregulated mirnas and their regulatory signature in glioma patients using the partial least squares method. *Experimental and therapeutic medicine*, 9(1):167--171.

Siegel, R. L., Miller, K. D., and Jemal, A. (2015). Cancer statistics, 2015. *CA: a cancer journal for clinicians*, 65(1):5--29.

Slattery, M. L., Herrick, J. S., Pellatt, D. F., Stevens, J. R., Mullany, L. E., Wolff, E., Hoffman, M. D., Samowitz, W. S., and Wolff, R. K. (2016). Microrna profiles in colorectal carcinomas, adenomas and normal colonic mucosa: variations in mirna expression and disease progression. *Carcinogenesis*, 37(3):245--261.

Song, Q., Xu, Y., Yang, C., Chen, Z., Jia, C., Chen, J., Zhang, Y., Lai, P., Fan, X., Zhou, X., et al. (2014). mir-483-5p promotes invasion and metastasis of lung adenocarcinoma by targeting rhogdi1 and alcam. *Cancer research*, 74(11):3031--3042.

Tai, M. C., Kajino, T., Nakatochi, M., Arima, C., Shimada, Y., Suzuki, M., Miyoshi, H., Yatabe, Y., Yanagisawa, K., and Takahashi, T. (2015). mir-342-3p regulates myc transcriptional activity via direct repression of e2f1 in human lung cancer. *Carcinogenesis*, 36(12):1464--1473.

Takigawa, S., Chen, A., Wan, Q., Na, S., Sudo, A., Yokota, H., and Hamamura, K. (2016). Role of mir-222-3p in c-src-mediated regulation of osteoclastogenesis. *International journal of molecular sciences*, 17(2):240.

Tamaddon, G., Geramizadeh, B., Karimi, M. H., Mowla, S. J., and Abroun, S. (2016). Mir-4284 and mir-4484 as putative biomarkers for diffuse large b-cell lymphoma. *Iranian journal of medical sciences*, 41(4):334.

Tang, Y., Lin, Y., Li, C., Hu, X., Liu, Y., He, M., Luo, J., Sun, G., Wang, T., Li, W., et al. (2015). Microrna-720 promotes in vitro cell migration by targeting rab35 expression in cervical cancer cells. *Cell & bioscience*, 5(1):56.

Tao, K., Yang, J., Guo, Z., Hu, Y., Sheng, H., Gao, H., and Yu, H. (2014). Prognostic value of mir-221-3p, mir-342-3p and mir-491-5p expression in colon cancer. *American journal of translational research*, 6(4):391.

Torkashvand, S., Damavandi, Z., Mirzaei, B., Tavallaei, M., Vasei, M., and Mowla, S. J. (2016). Decreased expression of bioinformatically predicted piwil2-targetting micrornas, mir-1267 and mir-2276 in breast cancer. *Archives of Iranian medicine*, 19(6):420.

van Jaarsveld, M. T., van Kuijk, P. F., Boersma, A. W., Helleman, J., van IJcken, W. F., Mathijssen, R. H., Pothof, J., Berns, E. M., Verweij, J., and Wiemer, E. A.

(2015). mir-634 restores drug sensitivity in resistant ovarian cancer cells by targeting the ras-mapk pathway. *Molecular cancer*, 14(1):196.

Wall, M. E., Rechtsteiner, A., and Rocha, L. M. (2003). Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91--109. Springer.

Wang, B., Li, J., Sun, M., Sun, L., and Zhang, X. (2014). mirna expression in breast cancer varies with lymph node metastasis and other clinicopathologic features. *IUBMB life*, 66(5):371--377.

Wang, X., Jiang, D., Xu, C., Zhu, G., Wu, Z., and Wu, Q. (2017). Differential expression profile analysis of mirnas with her-2 overexpression and intervention in breast cancer cells. *Int J Clin Exp Pathol*, 10(5):5039--5062.

Wang, X., Kuang, Y., Shen, X., Zhou, H., Chen, Y., Han, Y., Yuan, B., Zhou, J., Zhao, H., Zhi, Q., et al. (2015). Evaluation of mir-720 prognostic significance in patients with colorectal cancer. *Tumor Biology*, 36(2):719--727.

Wen, X.-Y., Wu, S.-Y., Li, Z.-Q., Liu, Z.-q., Zhang, J.-J., Wang, G.-F., Jiang, Z.-H., and Wu, S.-G. (2009). Ellagitannin (bja3121), an anti-proliferative natural polyphenol compound, can regulate the expression of mirnas in hepg2 cancer cells. *Phytotherapy research*, 23(6):778--784.

Wu, X., Li, S., Xu, X., Wu, S., Chen, R., Jiang, Q., Li, Y., and Xu, Y. (2015). The potential value of mir-1 and mir-374b as biomarkers for colorectal cancer. *International journal of clinical and experimental pathology*, 8(3):2840.

Xia, X., Li, Y., Wang, W., Tang, F., Tan, J., Sun, L., Li, Q., Sun, L., Tang, B., and He, S. (2015). Microrna-1908 functions as a glioblastoma oncogene by suppressing pten tumor suppressor pathway. *Molecular cancer*, 14(1):154.

Xie, J., Tan, Z.-H., Tang, X., Mo, M.-S., Liu, Y.-P., Gan, R.-L., Li, Y., Zhang, L., and Li, G.-Q. (2014). Mir-374b-5p suppresses reck expression and promotes gastric cancer cell invasion and metastasis. *World Journal of Gastroenterology: WJG*, 20(46):17439.

Xu, C., Zhang, L., Duan, L., and Lu, C. (2016). Microrna-3196 is inhibited by h2ax phosphorylation and attenuates lung cancer cell apoptosis by downregulating puma. *Oncotarget*, 7(47):77764.

Xu, C.-Z., Xie, J., Jin, B., Chen, X.-W., Sun, Z.-F., Wang, B.-X., and Dong, P. (2013). Gene and microrna expression reveals sensitivity to paclitaxel in laryngeal cancer cell line. *International journal of clinical and experimental pathology*, 6(7):1351.

Xu, L.-M., Li, L.-Q., Li, J., Li, H.-W., Shen, Q.-B., Ping, J.-L., Ma, Z.-H., Zhong, J., and Dai, L.-C. (2015). Upregulation of mir-1280 expression in non-small cell lung cancer tissues. *Chinese medical journal*, 128(5):670.

Xun, M., Ma, C.-F., Du, Q.-L., Ji, Y.-H., and Xu, J.-R. (2015). Differential expression of mirnas in enterovirus 71-infected cells. *Virology journal*, 12(1):56.

Yang, X. and Wu, X. (2016). mirna expression profile of vulvar squamous cell carcinoma and identification of the oncogenic role of mir-590-5p. *Oncology reports*, 35(1):398--408.

Ye, Y., Wei, Y., Yunxiuxiu, X., Li, Y., Wang, R., Chen, J., Zhou, Y., Fu, Z., Chen, Y., Wang, X., et al. (2017). Induced mir-1249 expression by aberrant activation of hedegehog signaling pathway in hepatocellular carcinoma. *Experimental cell research*, 355(1):9--17.

Yong, F. L., Law, C. W., and Wang, C. W. (2013). Potentiality of a triple microrna classifier: mir-193a-3p, mir-23a and mir-338-5p for early detection of colorectal cancer. *BMC cancer*, 13(1):280.

Yoon, A. J., Wang, S., Shen, J., Robine, N., Philipone, E., Oster, M. W., Nam, A., and Santella, R. M. (2014). Prognostic value of mir-375 and mir-214-3p in early stage oral squamous cell carcinoma. *American journal of translational research*, 6(5):580.

Zhang, J., Song, Y., Zhang, C., Zhi, X., Fu, H., Ma, Y., Chen, Y., Pan, F., Wang, K., Ni, J., et al. (2015a). Circulating mir-16-5p and mir-19b-3p as two novel potential biomarkers to indicate progression of gastric cancer. *Theranostics*, 5(7):733.

Zhang, L., Qian, J., Qiang, Y., Huang, H., Wang, C., Li, D., and Xu, B. (2014). Down-regulation of mir-4500 promoted non-small cell lung cancer growth. *Cellular Physiology and Biochemistry*, 34(4):1166--1174.

Zhang, T., Jing, L., Li, H., Ding, L., Ai, D., Lyu, J., and Zhong, L. (2017a). Microrna-4530 promotes angiogenesis by targeting vash1 in breast carcinoma cells. *Oncology Letters*, 14(1):111--118.

Zhang, X.-F., Li, K.-k., Gao, L., Li, S.-Z., Chen, K., Zhang, J.-B., Wang, D., Tu, R.-F., Zhang, J.-X., Tao, K.-X., et al. (2015b). mir-191 promotes tumorigenesis of human colorectal cancer through targeting c/ebpβ. *Oncotarget*, 6(6):4144.

Zhang, Y., Li, M., Ding, Y., Fan, Z., Zhang, J., Zhang, H., Jiang, B., and Zhu, Y. (2017b). Serum microrna profile in patients with colon adenomas or cancer. *BMC medical genomics*, 10(1):23.

Zhang, Z., Zhang, B., Li, W., Fu, L., Fu, L., Zhu, Z., and Dong, J.-T. (2011). Epigenetic silencing of mir-203 upregulates snai2 and contributes to the invasiveness of malignant breast cancer cells. *Genes & cancer*, 2(8):782--791.

Zhao, Q., Caballero, O. L., Levy, S., Stevenson, B. J., Iseli, C., De Souza, S. J., Galante, P. A., Busam, D., Leversha, M. A., Chadalavada, K., et al. (2009). Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. *Proceedings of the National Academy of Sciences*, 106(6):1886--1891.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. *PloS one*, 9(1):e78644.

Zhao, Y., Qi, X., Chen, J., Wei, W., Yu, C., Yan, H., Pu, M., Li, Y., Miao, L., Li, C., et al. (2017). The mir-491-3p/sp3/abcb1 axis attenuates multidrug resistance of hepatocellular carcinoma. *Cancer Letters*.

Zhao, Z., Wang, L., Song, W., Cui, H., Chen, G., Qiao, F., Hu, J., Zhou, R., and Fan, H. (2015). Reduced mir-29a-3p expression is linked to the cell proliferation and cell migration in gastric cancer. *World journal of surgical oncology*, 13(1):101.

Zheng, G., Jia, X., Peng, C., Deng, Y., Yin, J., Zhang, Z., Li, N., Deng, M., Liu, X., Liu, H., et al. (2015). The mir-491-3p/mtorc2/foxo1 regulatory loop modulates chemo-sensitivity in human tongue cancer. *Oncotarget*, 6(9):6931.

Zheng, L., Jiao, W., Mei, H., Song, H., Li, D., Xiang, X., Chen, Y., Yang, F., Li, H., Huang, K., et al. (2016). mirna-337-3p inhibits gastric cancer progression through repressing myeloid zinc finger 1-facilitated expression of matrix metalloproteinase 14. *Oncotarget*, 7(26):40314.

Zhou, H., Guo, W., Zhao, Y., Wang, Y., Zha, R., Ding, J., Liang, L., Yang, G., Chen, Z., Ma, B., et al. (2014). Microrna-135a acts as a putative tumor suppressor by directly targeting very low density lipoprotein receptor in human gallbladder cancer. *Cancer science*, 105(8):956--965.

Zhou, R., Zhou, X., Yin, Z., Guo, J., Hu, T., Jiang, S., Liu, L., Dong, X., Zhang, S., and Wu, G. (2016). Microrna-574-5p promotes metastasis of non-small cell lung cancer by targeting ptpru. *Scientific reports*, 6.

Zhou, X., Liu, K.-Y., and Wong, S. T. (2004). Cancer classification and prediction using logistic regression with bayesian gene selection. *Journal of Biomedical Informatics*, 37(4):249--259.

Zhu, J. and Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(3):427--443.