

MicroRNA Expression and Activity in High-grade Serous Ovarian Cancer



Eleanor A. Howe

Supervisors: Prof Chris Holmes and Prof John Quackenbush

Kellogg College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity Term, 2012

Acknowledgements

I would like to thank my adviser, Chris Holmes, for his support through the last four years and his advice and feedback on my work. I am grateful for the opportunity to work with him and his willingness to support me and my work despite the rather unusual logistics.

John Quackenbush has been my supervisor and mentor for nearly a decade. I would truly not be where I am without his help. I am grateful to him for believing in me for the last ten years.

Dimitris Spentzos has provided invaluable help in the writing of the paper based on the work in Chapter 3. His advice on structuring the paper and his kind but detailed feedback on the writing was incredibly important.

I am grateful to Kellogg College, for their complete acceptance of my situation as an overseas student.

I have many colleagues, both at Oxford and at Dana-Farber, without whom I would be in very, very deep trouble. I would particularly like to thank:

Cecelia Lindgren and Matthias Rantalainen for their helpful comments and critique during the transfer of status exam;

Matthias and Kyle Gaulton for similarly helpful feedback during the confirmation exam;

Stefan Bentink, who began the study of these ovarian cancer data and assisted me with advice and friendship throughout my time of study;

Nicolai Juul for reading/commenting on chapters 2 and 3, and for many lunchtime conversations about the perils of working with miRNA data;

Hugo Aerts for similar lunches and for reading and commenting on chapter 4;

Dan Gusenleitner for many rounds of proofreading and statistical nitpicking;

my colleagues from Dana-Farber: Bose Kochupurakkal, Dan Gusenleitner, Stefan Bentnik, Harold Gomez, Joeseeph White, Fah Sathirapongsasuti and Aedin Culhane;

my colleagues in the Holmes lab: Quinn Wills, Peter Kecskemethy, Chris Yao and Jeremie Becker;

Madeline Mitchell, who was endlessly patient about setting up phone calls and meeting for me.

Finally, I would like to thank my husband Mark, for his patience while I worked and meticulous proofreading while I didn't; Chris and Jen, for many fun times and just the right amount of distraction; and my family, for all of the usual things amazing families do.

Abstract

miRNAs are critical modulators in the development and progression of cancer [56, 81, 32]. Emerging evidence suggests that they are drivers of ovarian cancer [84, 12, 33]. A better understanding of the molecular underpinnings of the development, progression and chemoresistance of the disease is critical for the development of new, more effective therapies. Here we explore the expression patterns of miRNAs as they relate to gene expression, as they differ across molecular subtypes of the disease.

We examine the correlation structure of miRNA expression with mRNA expression in two distinct genomic datasets and report on patterns in correlation structure in several subsets of the data. We find that the datasets show consistency in their correlation structure, and in the specific miRNA-mRNA pairs that are either highly positively or negatively correlated. The data include a larger number of strong positive and strong negative correlations than would be expected by chance, indicating that biological relationships between the types of data are detectable in these datasets. We further find an enrichment for positively-correlated miRNA-mRNA pairs in which the miRNA is encoded in close proximity to the mRNA. The correlation of miRNA and mRNA is apparently unaffected by miRNA and mRNA expression level; similarly the two molecular subtypes do not contain differences in their correlation.

We find that the recently described [15] poorer prognosis, or angiogenic, subtype has a generally lower miRNA activity than the second, non-angiogenic, subtype. The subtypes are characterized by a consistent pattern of differential miRNA expression. We also report on a switch-like relationship between the expression levels of certain miRNAs and the genes that are anticorrelated with them. We propose these miRNAs drive many of the differences in the subtypes both directly, by RISC-mediated repression of target messages and indirectly, by repressing transcription factors that regulate expression in the cell.

We build models of patient survival and time-to-relapse based on these miRNA expression data and inferred miRNA activity scores, using several types of univariate and variable selection models. We find essentially no survival-predictive information provided by the RE score data. While the direct miRNA expression measurements may contain some predictive power, we find that a larger dataset and the segregation of that dataset into distinct molecular phenotypes is likely to be necessary to produce a useful model of survival in ovarian cancer.

Contents

Contents	v
List of Figures	viii
List of Tables	xiv
1 Introduction	2
1.1 Ovarian Cancer	2
1.2 MicroRNAs	4
1.3 MicroRNAs in Cancer	7
1.4 MicroRNAs in Ovarian Cancer	10
1.5 Using target transcript levels to imply miRNA activity	12
1.5.1 microRNA Target Prediction Tools	13
1.6 Analysis tools for the study of genomic data	15
1.7 Identification of an angiogenic subtype in Serous Epithelial Ovarian Cancer	20
1.7.1 ISIS: Identifying Splits with Clear Separation	21
1.7.2 Verification of the subtypes	24
1.8 Ovarian cancer subtypes	25
1.9 The significance of this work	26
2 Exploration of the correlation of miRNA and mRNA expression and miRNA activity in ovarian cancer.	28
2.1 Introduction	28
2.1.1 Datasets	29
2.2 Results	32
2.2.1 Some similarity in miRNA-mRNA partial correlation in the Bentink and TCGA datasets	32
2.2.2 Predicted miRNA-targets have similar partial correlation profile to all possible targets	38

2.2.3	Correlation of miRNA and mRNA expression are driven in part by co-location on the chromosome.	40
2.2.4	Co-located miRNA-mRNA pairs do not account for excess large-positive partial correlations	41
2.2.5	Biological consistency in miRNA-mRNA pairs with large negative correlation coefficients	45
2.2.6	miRNA-mRNA correlation of predicted miRNA-mRNA target pairs is not dependent on median mRNA expression. .	46
2.2.7	Similar patterns of correlation in two molecular subtypes .	53
2.2.8	Testing differentially expressed miRNAs for partial correlation enrichment.	56
2.3	Conclusions and future work	59
3	Characterization of miRNAs in two ovarian subtypes	63
3.1	Introduction	63
3.2	Results	67
3.2.1	Differentially expressed miRNAs.	67
3.2.2	Lower miRNA activity in the angiogenic sample group . .	70
3.2.3	Switch-like activity describes the relationship between 506-514 cluster miRNAs and their anticorrelated genes.	75
3.2.4	Genes anti-correlated with the 506-514 miRNA cluster are enriched for ECM genes and some transcription factors. . .	81
3.2.5	Within the 506-514 Cluster, miR-506 and miR-507 appear to be active.	82
3.2.6	Anti-correlated mRNAs are enriched for a common set of binding sites for cancer-related transcription factors. . . .	83
3.3	Conclusions	85
3.4	Future Work	89
4	Modeling of risk based on miRNA expression and activity	91
4.1	Introduction	91
4.1.1	Cox Proportional Hazards Model with multidimensional data	94
4.1.2	The data	96
4.1.3	Testing the proportionality assumption	97
4.2	Univariate analysis of miRNA expression and activity	99
4.3	Forward stepwise selection	104
4.3.1	Evaluating models using Information Criteria	105
4.3.2	Forward stepwise selection on the Bentink miRNA expression data	107
4.3.3	Forward stepwise selection on the TCGA miRNA expression data	108

4.3.4	Little common signal between TCGA and Bentink miRNA expression datasets	110
4.4	Coefficient shrinkage and variable selection using the Lasso	111
4.4.1	Creating the models.	112
4.4.2	Lasso-based modeling of ovarian cancer survival using miRNA expression data.	115
4.4.3	Examining the models.	119
4.5	Conclusions and Future Work	121
5	Conclusions	124
A	Appendix to Chapter 2	127
B	Appendix to Chapter 3	136
C	Appendix to Chapter 4	153
	Bibliography	164

List of Figures

1.1	The miRNA processing pathway in animals. This diagram is courtesy of Narayanese at English Wikipedia and available under the Creative Commons license.	5
2.1	a) Distribution of partial correlation coefficients of the expression values of each possible miRNA-mRNA pair across all samples in the Bentink dataset. b) Q-Q plot of the partial correlation coefficients in a, against a generated normal distribution of identical mean and standard deviation. c) Distribution of partial correlation coefficients of the expression values of each possible miRNA-mRNA pair across all samples in the TCGA dataset. d) Q-Q plot of the partial correlation coefficients in c, against a generated normal distribution of identical mean and standard deviation. . .	34
2.2	Correlation of miRNA-mRNA pairs measured in both the Bentink and TCGA datasets.	38
2.3	a) Distribution of partial correlation coefficients of expression values in the Bentink dataset, of all PITA-predicted miRNA-mRNA pairs. b) Q-Q plot of distribution in a, against a normal distribution of the same mean and standard deviation. c) Distribution of partial correlation coefficients of expression values in the TCGA dataset of all PITA-predicted miRNA-mRNA pairs. d) Q-Q plot of distribution in c, against a normal distribution of the same mean and standard deviation.	40
2.4	a) Distribution of miRNA-mRNA correlations in the Bentink dataset with miRNA-mRNA pairs separated by at least 10,000kb. b) Q-Q plot of those correlations, against a randomly-generated normal distribution of the same mean and standard deviation. c) Q-Q plot of those correlations against the unfiltered correlations. . . .	42

LIST OF FIGURES

2.5	a) Distribution of miRNA-mRNA correlations in the TCGA dataset with miRNA-mRNA pairs separated by at least 10,000kb. b) Q-Q plot of those correlations, against a normal distribution of the same mean and standard deviation. c) Q-Q plot of those correlations against the unfiltered correlations.	43
2.6	Partial correlations broken down by mRNA expression values for the Bentink dataset. a) Median mRNA expression values vs. partial correlation of miRNAs with their predicted targets. Each point represents a single mRNA-miRNA predicted pair. The majority of the points are shown as a cloud (light blue). b) Comparison of the distribution of partial correlation coefficients for all non-located miRNA-mRNA in the Bentink dataset, against a normal distribution of the same mean and standard deviation. c) Comparison of the distribution of partial correlation coefficients for the high-expression (mean mRNA expression > 8.5) non-located miRNA-mRNA pairs in the Bentink dataset. d) Comparison of the distribution of partial correlations for the low-expression (mean mRNA expression < 8.5) non-located miRNA-mRNA pairs in the Bentink dataset.	47
2.7	Partial correlations broken down by mRNA expression values for the TCGA dataset. a) Median mRNA expression values vs. partial correlation of miRNAs with their predicted targets. Each point represents a single mRNA-miRNA predicted pair. The majority of the points are shown as a cloud (light blue). b) Comparison of the distribution of partial correlation coefficients for all non-located miRNA-mRNA pairs in the TCGA dataset, against a normal distribution of the same mean and standard deviation. c) Comparison of the distribution of partial correlation coefficients for the high-expression (mean mRNA expression > 7.4) non-located miRNA-mRNA pairs in the TCGA dataset. d) Comparison of the distribution of partial correlation coefficients for the low-expression (mean mRNA expression < 7.4) non-located miRNA-mRNA pairs in the TCGA dataset.	48

2.8	Partial correlations broken down by miRNA expression values for the Bentink dataset. a) Median miRNA expression values vs. correlation of miRNAs with their predicted targets. Each point represents a single mRNA-miRNA predicted pair. The majority of the points are shown as a cloud (light blue). b) Correlation of all possible miRNA-mRNA pairs at least 10,000kb apart on the chromosome, compared to a normal distribution of the same mean and standard deviation. c) Comparison of the distribution of correlation coefficients for the high-expression (mean miRNA expression > 10) non-located miRNA-mRNA pairs in the TCGA dataset. d) Comparison of the distribution of correlation coefficients for the low-expression (mean miRNA expression < 10) non-located miRNA-mRNA pairs in the TCGA dataset.	51
2.9	Correlations broken down by miRNA expression values for the TCGA dataset. a) Median miRNA expression values vs. correlation of miRNAs with their predicted targets. Each point represents a single mRNA-miRNA predicted pair. The majority of the points are shown as a cloud (light blue). b) Correlation of all possible miRNA-mRNA pairs at least 10,000kb apart on the chromosome, compared to a normal distribution of the same mean and standard deviation. c) Comparison of the distribution of correlation coefficients for the high-expression (mean miRNA expression > 6) non-located miRNA-mRNA pairs in the TCGA dataset. d) Comparison of the distribution of correlation coefficients for the low-expression (mean miRNA expression < 6) non-located miRNA-mRNA pairs in the TCGA dataset.	52
2.10	Partial correlation distributions for miRNA-mRNA expression profile pairs in the angiogenic subtype of the TCGA and Bentink Datasets. a) Distribution of partial correlations for all possible miRNA-mRNA pairs in the angiogenic subtype, Bentink dataset, located at least 10,000kb apart. b) Q-Q comparison of the same distribution against a normal distribution of the same mean and standard deviation. c) Distribution of partial correlations for all possible miRNA-mRNA pairs in the angiogenic subtype, TCGA dataset, located at least 10,000kb apart. d) Q-Q comparison of the same distribution against a normal distribution of the same mean and standard deviation.	54

LIST OF FIGURES

2.11	Partial correlation distributions for mRNA-miRNA expression profile pairs in the non-angiogenic subtype of the Bentink and TCGA Datasets. a) Distribution of partial correlations for all possible miRNA-mRNA pairs in the non-angiogenic subtype, Bentink dataset, located at least 10,000kb apart. b) Q-Q comparison of the same distribution against a normal distribution of the same mean and standard deviation. c) Distribution of partial correlations for all possible miRNA-mRNA pairs in the non-angiogenic subtype, TCGA dataset, located at least 10,000kb apart. d) Q-Q comparison of the same distribution against a normal distribution of the same mean and standard deviation.	55
2.12	Distribution of median partial correlation coefficients for each subtype, marked with the median partial correlation of differentially expressed miRNAs between the two subtypes. a) Median partial correlation of miRNAs in the Bentink dataset, angiogenic subtype. b) Median partial correlation of miRNAs in the Bentink dataset, non-angiogenic subtype.	57
2.13	Distribution of median partial correlation coefficients for each subtype, marked with the median partial correlation of differentially expressed miRNAs between the two subtypes. a) Median partial correlation of miRNAs in the TCGA dataset, angiogenic subtype. b) Median partial correlation of miRNAs in the TCGA dataset, non-angiogenic subtype.	58
3.1	Expression distribution of each of representative up- and down-regulated miRNAs, in the angiogenic and non-angiogenic samples.	68
3.2	Distribution of t-statistics resulting from test of RE scores for each miRNA between the angiogenic and non-angiogenic subtypes. A t-score of $t < 0$ indicates that a miRNA has a lower RE score in the angiogenic subtype, relative to the non-angiogenic subtype, whereas a score of $t > 0$ indicates that the miRNA has a higher activity in the angiogenic subtype. Each colored line represents t-statistics calculated for RE scores from a different set of target prediction data: red: PITA; green: Pictar; blue: Targetscan; purple: mirSVR. a) Bentink dataset. b) Tothill dataset. c) TCGA dataset.	71

LIST OF FIGURES

3.3	Relationship between miRNA expression and the expression of significantly anti-correlated mRNAs for eight miRNAs in the Bentink (top) and TCGA (bottom) datasets. A typical non-classifier (hsa-miR-100) is shown for contrast. Each point represents one sample. The y-axis represents the measured miRNA expression level. The x-axis represents the average expression level of the mRNAs significantly anti-correlated with the miRNA. Each point represents one sample, colored red in the case of the angiogenic subtype samples, and green in the case of the non-angiogenic subtype samples. . . .	78
3.4	Significantly enriched transcription factors from the 506-514 anti-correlated genes in the TCGA and Bentink datasets. A blue box indicates that the prediction tool predicts a binding site for the miRNA in the transcription factor. Gray boxes indicates no such prediction.	85
4.1	Scaled Schoenfeld residuals of each sample against time, for several example miRNAs in each dataset. The solid line is a smoothing-spline fit to the residuals. The dotted lines are ± 2 -standard-error band around that fit.	98
4.2	a-b) $-\log_{10}$ p-values of permuted univariate models built on miRNA expression and RE scores, respectively. The 5th percentile is marked with a vertical line. c-d) Frequency distribution of the un-permuted p-values, respectively.	100
4.3	a-b) $-\log_{10}$ p-values of permuted univariate models built on TCGA miRNA expression and RE scores, respectively. The 5th percentile is marked with a vertical line. c-d) Frequency distribution of the un-permuted p-values, respectively.	103
4.4	Log-likelihood of the models built with miRNA expression data at varying λ values, for each of 20 resamplings of 10-fold cross-validation on the TCGA dataset. b) Aikake Information Criterion calculated for each of the same models shown in a. c) Bayesian Information Criterion calculated for each model shown in a. . . .	114
4.5	$-\log_{10}$ p-values for univariate models fitted on the miRNA expression in the TCGA dataset (x-axis) vs. the Bentink dataset (y-axis).	120
4.6	Effect sizes for the univariate models fitted on the miRNA expression in the TCGA dataset (x-axis) vs. the Bentink dataset (y-axis).	121
B.1	The top significantly enriched functional annotations from the 506-514 anti-correlated genes in the Bentink dataset. Many more terms were enriched, but too many were available to display here.	146

LIST OF FIGURES

- B.2 The top significantly enriched functional annotations from the 506-514 anti-correlated genes in the TCGA dataset. Many more terms were enriched, but too many were available to display here. . . . 147

List of Tables

2.1	Expected and observed moment ratios for the partial correlation sets of this chapter.	36
3.1	Counts of positive and negative RE-changing miRNAs for the three datasets, q-value of $q < 0.05$	72
4.1	miRNAs included in optimal forward-selected model based on Bentink miRNA expression data.	107
4.2	miRNAs included in optimal forward-selected model based on TCGA expression data.	109
4.3	Summary results for models built in Chapter 4. train IC: average concordance index of training set within the 10-fold cross-validation loop. test IC: average concordance index for predictions on each of the 1000 test sets. Bentink IC: average concordance index for predictions of each of the 1000 models on the Bentink dataset.	116
4.4	miRNAs with non-zero coefficients from TCGA univariate models.	118
A.1	miRNA-mRNA target pairs with the top 30 largest positive partial correlations (Bentink dataset).	128
A.2	miRNA-mRNA target pairs with the top 30 largest positive partial correlations (TCGA Dataset).	129
A.3	miRNA-mRNA pairs with the top 30 largest positive partial correlations, within 10,00kb on the chromosome (Bentink dataset). .	130
A.4	miRNA-mRNA pairs with the top 30 largest positive partial correlations, within 10,00kb on the chromosome (TCGA Dataset). . .	131
A.5	miRNA-mRNA pairs with the top 30 largest negative partial correlations, within 10,00kb on the chromosome (Bentink dataset). .	132
A.6	miRNA-mRNA pairs with the top 30 largest negative partial correlations, within 10,00kb on the chromosome (TCGA Dataset). . .	133
A.7	miRNA-mRNA target pairs with the top 30 largest negative partial correlations (Bentink dataset).	134

A.8	miRNA-mRNA target pairs with the top 30 largest negative partial correlations, within 10,00kb on the chromosome (TCGA Dataset).	135
B.1	Top 30 upregulated miRNAs in Angiogenic subtype of Bentink dataset. d-value: delta value calculated by SAM. miRNA f-test: The p-value resulting from a test of differential variance between sample groups (F-test). RE test: p-value of the t-test comparing Pita-based RE scores between the angiogenic and non-angiogenic subtypes. RE diff: Indicates whether the miRNA had a negative or positive RE change. Pos: RE scores were higher in the angiogenic subtype. Neg: RE scores were lower in the angiogenic subtype.	137
B.2	Top 30 upregulated miRNAs in the angiogenic subtype of the TCGA dataset. d-value: delta value calculated by SAM. miRNA f-test: The p-value resulting from a test of differential variance between sample groups (F-test). RE test: p-value of the t-test comparing Pita-based RE scores between the angiogenic and non-angiogenic subtypes. RE diff: Indicates whether the miRNA had a negative or positive RE change. Pos: RE scores were higher in the angiogenic subtype. Neg: RE scores were lower in the angiogenic subtype.	138
B.3	Top 30 downregulated miRNAs in angiogenic subtype of Bentink dataset. d-value: delta value calculated by SAM. miRNA f-test: The p-value resulting from a test of differential variance between sample groups (F-test). RE test: p-value of the t-test comparing Pita-based RE scores between the angiogenic and non-angiogenic subtypes. RE diff: Indicates whether the miRNA had a negative or positive RE change. Pos: RE scores were higher in the angiogenic subtype. Neg: RE scores were lower in the angiogenic subtype.	139
B.4	Top 30 downregulated miRNAs in the angiogenic subtype of the TCGA dataset. d-value: delta value calculated by SAM. miRNA f-test: The p-value resulting from a test of differential variance between sample groups (F-test). RE test: p-value of the t-test comparing Pita-based RE scores between the angiogenic and non-angiogenic subtypes. RE diff: Indicates whether the miRNA had a negative or positive RE change. Pos: RE scores were higher in the angiogenic subtype. Neg: RE scores were lower in the angiogenic subtype. NA indicates that no target prediction data were available for that miRNA, and therefore no RE score could be calculated.	140

LIST OF TABLES

B.5	Positive RE-changing miRNAs calculated for each of the four miRNA target prediction tools, using the Bentink mRNA expression dataset, $p < 0.05$	141
B.6	Top 15 positive RE-changing miRNAs, TCGA dataset.	141
B.7	To 15 negative RE-changing miRNAs calculated for each of the four miRNA target prediction tools, using the Bentink mRNA expression dataset, $p < 0.05$	142
B.8	Top 15 negative RE-changing miRNAs calculated for each of the four miRNA target prediction tools, using the TCGA mRNA expression dataset, $p < 0.05$	143
B.9	150 of the 2508 genes with expression patterns significantly anticorrelated with the 506-514 cluster of miRNAs in the Bentink dataset.	144
B.10	150 of the 4082 genes with expression patterns significantly anticorrelated with the 506-514 cluster of miRNAs in the TCGA dataset.	145
B.11	GO terms used for selection of genes associated with miRNA processing.	148
B.12	Transcription factors with binding sites enriched in the anticorrelated genes to the 506-514 cluster of miRNAs. Bentink dataset. . .	150
B.13	Transcription factors with binding sites enriched in the anticorrelated genes to the 506-514 cluster of miRNAs. TCGA dataset. . .	151
B.14	miRNA-processing genes downregulated in the angiogenic subtype of the TCGA dataset. Significance Analysis of Microarrays (SAM) with an FDR of 0.05.	152
B.15	Transcription factors differentially expressed in the angiogenic subtype of the TCGA dataset. Significance Analysis of Microarrays (SAM) with an FDR of 0.05.	152
C.1	miRNAs with expression values most predictive of survival using a univariate model. Models calculated using expression data from the Bentink dataset, and based on the p-value calculated with a likelihood ratio test. Models for 98 miRNAs had a calculated p-value of $p < 0.052$. ci: Concordance index of the model.	155
C.2	miRNAs with RE scores most predictive of survival in the Bentink dataset, based on the p-value calculated with a likelihood ratio test. Models for 106 miRNAs had a calculated p-value of $p < 0.051$. ci: Concordance index of the model.	158

LIST OF TABLES

C.3	miRNAs with expression values most predictive of survival in the TCGA dataset, based on the p-value calculated with a likelihood ratio test. Models for 105 miRNAs had a calculated p-value of $p < 0.051$. ci: Concordance index of the model.	161
C.4	miRNAs with RE scores most predictive of survival in the TCGA dataset, based on the p-value calculated with a likelihood ratio test of a univariate model. Models for 88 miRNAs had a calculated p-value of $p < 0.052$. ci: Concordance index of the model.	163

LIST OF TABLES

Chapter 1

Introduction

1.1 Ovarian Cancer

Ovarian cancer is the most lethal of the gynecologic malignancies, responsible for the deaths of 16,000 people per year worldwide. The disease is often diagnosed at a late stage due to the non-specific nature of its symptoms. Epithelial ovarian cancer occurs in patients as one of several subtypes, including serous, endometrioid, clear cell, carcinosarcoma, mucinous, and various mixed and transitional types. By far the most common of these is serous carcinoma, comprising more than 50% of all cases [135].

Serous ovarian cancer is thought to originate with mutations in the fallopian tube epithelium or the ovarian surface epithelium [105]. Mutations in a wide range of genes have been implicated in the development and progression of ovarian tumors. Furthermore, changes in messenger RNA (mRNA) expression profiles are critically important to disease progression [142, 141].

Standard treatment for the disease involves surgery to remove the tumor mass

followed by chemotherapy with platinum-based compounds, such as cisplatin and carboplatin-paclitaxel [122]. Most patients respond well initially, with near total remission of disease. However, most of these patients (75-80%) recur within 12 to 24 months after diagnosis and die of chemotherapy-resistant metastatic tumors.

The chemotherapeutic compounds enter the cell both passively and by active transport and bind with DNA, forming adducts. These adducts trigger the cell's DNA damage response, halting progression of the cell cycle and, if the cell is unable to repair the damage, apoptosis [152]. However, chemotherapy resistance develops in most cases of late-stage ovarian cancer, and is known to be mediated at least in part by genetic changes in the tumor. The exact mechanism or mechanisms of the development of resistance have not been fully determined.

In the last ten years, no improvements have been made in ovarian cancer mortality; it remains the fourth largest source of cancer mortality among women. A better understanding of the mechanism of its development, spread and resistance to chemotherapy is necessary for the development of superior treatments for the disease.

One recent discovery of note is the recent findings supporting the fallopian tube cell-of-origin hypothesis. It has been hypothesized for several years that the original mutations leading to serous ovarian tumors occurred in secretory cells of the fallopian tube, rather than in the epithelium of the ovary itself [44, 100]. Recent studies in mouse models have provided compelling evidence for the fallopian tube origin of the disease [92], resulting in an uncomfortable situation with regards to disease nomenclature. For simplicity's sake, we will continue in this document to refer to this disease as "ovarian" cancer, despite the possibility of a future name change.

1.2 MicroRNAs

microRNAs (miRNAs) are short, non-coding RNAs of 19-24 nucleotides in length that act to repress gene expression of target mRNAs. Mature miRNAs are thought to direct the posttranscriptional regulation of gene expression in animals as part of the RNA-induced silencing complex (RISC). In this complex, they are bound to the protein Argonaute, and base-pair with conserved target sequences in the 3' untranslated regions of messenger RNAs [35]. Nucleotides 2-8 of the miRNA are referred to as the “seed” region, and are critical for miRNA targeting and the downregulation of expression of the target [24]. Most seeds are of approximately 6nt in length, and the majority of these sites are conserved [54]. The majority of these sites are also not covered by current target prediction methods [54]. The mechanism by which the target is downregulated depends at least in part on the degree of complementarity between the miRNA and its target site, with exact matches resulting in Argonaute-mediated cleavage of the mRNA, and less complete base-pairing resulting in translational repression or mRNA destabilization [111, 9, 10, 24].

Binding sites for miRNAs are located within the 3'UTR of genes, with more than 60% of human protein-coding genes thought to be under selective pressure to maintain binding affinity with miRNAs [61]. A single mRNA can be targeted by multiple miRNAs [108, 96, 61].

Most miRNAs are transcribed in capped, polyadenylated, polycistronic primary transcripts [26, 66]. They are then post-processed by the Drosha/Pasha complex into pre-miRNAs, which are transported out of the nucleus. There, they are further processed by the Dicer complex into multiple, functional miRNAs

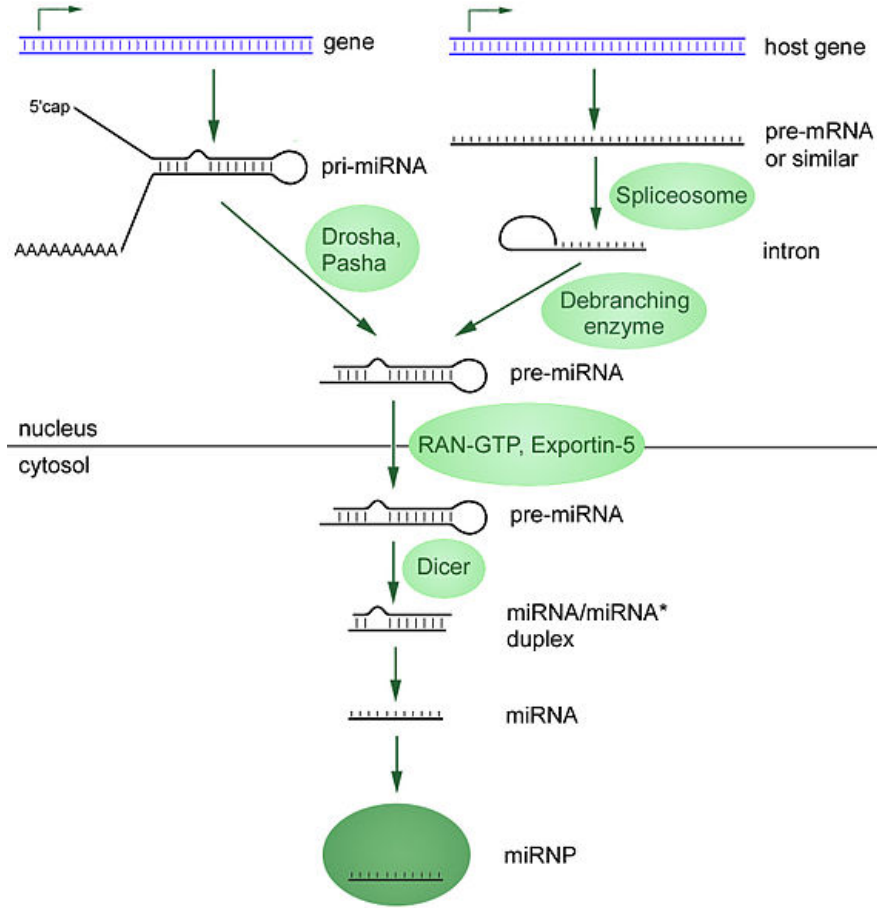


Figure 1.1: The miRNA processing pathway in animals. This diagram is courtesy of Narayanese at English Wikipedia and available under the Creative Commons license.

[104]. These miRNAs are incorporated into the microRNA ribonucleoprotein complex (miRNP), also known as the RNA-Induced Silencing Complex (RISC), the functional assembly that targets messenger RNAs (miRNA processing is summarized in Figure 1.2).

miRNAs are organized into families of related sequences, which can then be grouped into clans if they are determined to be descended from a common ancestral sequence. These molecules are encoded by highly conserved genes across the

animal kingdom and are known to be involved in the regulation of many cellular processes, such as apoptosis [36, 170], development [123, 90] and differentiation [34]. It is thought that the majority of protein-coding genes are miRNA-regulated [107, 136, 7, 18, 61].

Most mammalian mRNAs are targets of miRNAs ([61]). Mukherji *et al.* [119] found that miRNAs can fine-tune expression levels of target mRNAs. Guo *et al.* [67, 79] report that most miRNAs act predominantly to reduce the expression levels of mRNA targets by transcript destabilization, due to a lack of protection from translating ribosomes, rather than by the reduction of protein production by translational repression. Most miRNAs act to reduce the expression level of each of a large number of mRNA transcripts by a small amount [7]. This type of regulation of tissue type is a common theme among miRNAs. Arnold and *et al.* [4] identified a miRNA signature that marks the transition from self-renewing stem cells to differentiating and proliferating progenitor cells that give rise to fully differentiated cells.

The first miRNA described, let-7, was found to be critically important in the developmental timing of the worm *C. elegans*, and later to be similarly important in the development of all other mesoderms. Subsequently discovered miRNAs have important roles to play in many biological processes, including controlling hindlimb development (miR-196; [75]) and B-cell differentiation (miR-150; [168]). One miRNA, miR-203, induces cell-cycle exit by targeting p63 and restricting the proliferative potential of the cell [173].

1.3 MicroRNAs in Cancer

The discovery of the involvement of microRNAs in cancer is opening up many new avenues of research. The molecules are involved in many stages of cancer progression, such as oncogenesis [70], tumor growth [19, 68] and chemoresistance [171, 53, 21]. The first demonstration of miRNA involvement in cancer was with the finding that miR-15a and miR-16-1 function as tumor suppressors, and that deletion of these miRNAs is a common and pathogenic event in chronic lymphocytic leukemia [27, 40, 93]. Further examples of oncogenic or tumor suppressive miRNAs have followed, such as miR-125, which has also been implicated in leukemia [22]. Several miRNAs have been found to be differentially expressed in breast cancer relative to normal tissue [82]. The let-7 family of miRNAs are often downregulated in cancer [23, 130, 125]. The first cancer-linked miRNA discovered was in B-cell chronic lymphocytic leukemia (B-CLL), where Calin *et al.* [27] showed that miR-15 and miR-16 are deleted or downregulated in the majority of patients. In lung cancer, let-7 is up-regulated relative to normal lung tissue and overexpression of the miRNA is prognostic of short survival times [86, 148].

Further investigation of miRNA expression in cancer tissues have demonstrated that miRNA involvement is not restricted to single-molecule expression changes; miRNA expression is widely deregulated in tumor tissue [112, 159]. Expression profiles of miRNAs are informative measurements that can reveal a large amount of information about the character of their tissue of origin. For example, Ferracin *et al.* [58] used miRNA expression profiling to identify the cells-of-origin for metastatic tumors. miRNA profiles have been used to differen-

tiate melanoma subtypes [30]. The miRNA expression profile of tumor samples has also been shown to better classify non-metastatic tumors by tissue type than gene expression data [112]. They are seen as increasingly important players in all stages of cancer development and progression [83, 97]. Many miRNAs map to regions of the genome that are consistently altered in human cancers [29]. They are extensively involved in cancer pathogenesis in solid tumors [159]. miRNAs are deregulated in cancer by many mechanisms, including chromosomal deletion [27], epigenetic alterations[8], transcriptional deregulation, and RISC component disruption[38, 37].

Global changes in miRNA expression have also been linked to carcinogenesis. Lu *et al.* [112] found that overall miRNA expression levels are lower in tumor tissue and tumor cell lines when compared to corresponding normal tissues. They also report that the lineage of each tumor type was reflected in the tumor’s miRNA expression profile, in a way that is not found for messenger RNA expression profiles. Some miRNAs are tumor suppressors — miR-155 has been shown to prevent tumor metastasis from breast to lung, by repressing the transcription factor TCF4, a regulator of the epithelial-mesenchymal transition (EMT), a developmental process characteristic of cells undergoing proliferation [167]. Impaired miRNA processing has been shown to enhance cellular transformation and tumorigenesis [97]. Lu *emphet al.* [112] also found that many of these classifying miRNAs are downregulated in the tumor tissue compared to normal tissue. Knockdown of miRNA processing genes, resulting in a global reduction of miRNA expression, increases tumorigenicity of cancer cell lines and results in greater invasion and cell migration of xenograft tissue in nude mice [98].

Aberrant expression of miRNAs is known to be involved in carcinogenesis, due

to their promotion of the expression of oncogenes or repression of the expression of tumor suppressors. Half of known miRNA genes are found in fragile sites and cancer-related genomic locations [28]. miRNA targets have been found to be generally downregulated in prostate, breast, AML and liver tumors relative to normal tissue [147].

Cancer-promoting miRNAs, or *oncomirs*, are located mainly in amplified regions of cancer genomes, whereas tumor-suppressing miRNAs, or *mirsupps*, are mainly found in deleted regions [161]. Oncomirs generally have functions like promotion of cell growth, inhibition of apoptosis, repression of immune cell development and controlling of the cell cycle. Mirsupps, conversely, perform opposing functions of inhibiting cell growth, inducing apoptosis, regulating immune cell development, and inducing cell cycle arrest. Some miRNAs can function as both tumor suppressors and oncogenes in different cancer contexts [36]. It has been suggested that mirsupps, being more evolutionarily conserved than oncomirs, may be more important to prevention of tumor growth. Oncomirs target tumor suppressor genes, whereas mirsupps target oncogenes. These tumor-associated miRNAs (t-miRNAs) are more likely to act via a direct-cleavage mechanism than the population of animal miRNAs as a whole [161]. This tumor suppressor role for miRNAs and the role of miRNAs in locking down the differentiated state of tissues suggests (reviewed in [143]) that loss of miRNA expression in cancer may result in a cell moving to a less-differentiated state.

1.4 MicroRNAs in Ovarian Cancer

Recent research has uncovered roles for miRNAs in the development, spread and chemotherapeutic response of ovarian cancer. Several studies have identified miRNAs that are differentially expressed in ovarian tumor tissue as compared to normal tissue, whether the normal tissue is cultured Human Ovarian Serous Epithelial (HOSE) cells or primary ovarian tissue samples [120]. Nam *et al.* [120] found that several miRNAs are differentially expressed in serous ovarian carcinoma compared with normal ovarian tissues. Schmeier *et al.* [133] found that experimentally verified miRNA targets are overrepresented among ovarian cancer genes, when compared to all human genes. Chang, *et al.* [31] report that miR-148b is upregulated in ovarian tumors relative to normal tissues.

Higher expression of miR-200, miR-141, miR-18a, miR-93, and miR-429, and lower expression of let-7b and miR-199a in tumor tissue were significantly correlated with a poor prognosis [120]. miR-15a and miR-16 have been found to target Bmi-1, a protein overexpressed in ovarian cancer. Overexpression of these miRNAs results in a downregulation of Bmi-1 protein expression and a reduction in cancer cell proliferation and clonal growth [19].

miRNAs have also been implicated in chemotherapy resistance, an important aspect of ovarian cancer research as most ovarian tumors eventually develop it [21, 140]. Recently, a small group of miRNAs were found to be differentially expressed in platinum sensitive tumor samples, as compared to resistant samples [53]. Lu *et al.* [113] found that patients who responded to platinum and paclitaxel have a lower expression of let-7a than those who did not. miR-125b has been found to confer resistance to cisplatin in ovarian cancer cells [94]. miR-187 is

involved in ovarian cancer progression by targeting Disabled-Homolog-2 [33]. let-7 is a selector for paclitaxel in ovarian cancer [113]. SNPs in a miR-191 target site in the 3'UTR of the MDM4 gene have been shown to be associated with increased tumorigenesis and decreased chemosensitivity in ovarian cancer [166]. Kong *et al.* [94] report that miR-125b promotes cisplatin resistance in ovarian cancer cell lines by downregulating Bak1 and that miR-125b seems to target Bcl-2 antagonist killer 1 (Bak1). Additionally, an ovarian cancer signature predictive of platinum sensitivity has been reported [99]. [181] found that SNPs in miR-193b caused insensitivity to platinum-based chemotherapies in ovarian cancer cell lines. Indeed, let-7 has been proposed as a marker for the selection of paclitaxel treatment in ovarian cancer [113].

The introduction of transfected miRNAs (miR-7 and miR-128) into ovarian cancer cell lines has been shown to affect the expression of hundreds of genes, only about 20% of which were shown to be the direct result of miRNA-mRNA interaction [138]. This indirect measurement technique is justified, in part, because spiking with an endogenous miRNA causes the expression levels of many target mRNA protein products to drop slightly [136]. This broad but small drop in expression should be detectable by a rank-based method like that of Cheng *et al.* [37].

1.5 Using target transcript levels to imply miRNA activity

Because miRNAs are simply RNA molecules, their abundance can be measured in large scale experiments in the same way mRNAs are measured: with DNA microarrays. As such, many of the bioinformatics tools used for gene expression analysis can be applied to them. However the great interest in miRNAs is driven by their regulatory targets: what are they, and how much, and in what circumstances are they regulated?

Recently, Cheng *et al.* [37] have reported a global reduction in miRNA activity in ER positive (ER+) breast cancer tumors when compared to ER negative (ER-) tumors. In reaching this conclusion, they developed a Regulatory Effect (RE) score that is used to compare the relative repressive activity of a given miRNA in a given sample. In this method, the activity of miRNAs is inferred based on the microarray-based expression measurements of predicted-target mRNAs, in comparison to non-targets. For each sample, the mRNAs are ranked by expression level, from highest expression to lowest, and the ranks of predicted-target mRNAs for a given miRNA are averaged. Similarly, the average rank of non-predicted target mRNAs is calculated. The non-target average is subtracted from the target-average, resulting in an RE score for the given miRNA for the given sample. These calculations are performed separately for every miRNA in every sample. Note that direct measurements of miRNA expression levels are not included in the calculation at all.

These RE scores are not informative in an absolute sense, but rather as a comparative measure of miRNA effectiveness among the samples. Cheng *et al.*

used these scores to determine that miRNA targets were more repressed in ER-breast cancer than in ER+; however, miRNA expression did not necessarily correspond to miRNA activity. They showed a global change in miRNA regulatory effect between ER+ and ER- breast cancer tumor samples, based on target gene activity. ER- tumors had an overall lower level of miRNA activity, most likely driven by a deregulation of the miRNA processing genes Ago2 and Dicer.

miRNA repression effect is at least somewhat visible at the mRNA level, especially for highly-repressed mRNAs [67, 7]. These target prediction algorithms have been used for indirect measurement of miRNA activity in the cell by measuring the expression changes of their target genes [37, 174, 5, 162]).

1.5.1 microRNA Target Prediction Tools

Few miRNA-mRNA regulatory relationships have been verified experimentally. Therefore the computational prediction of miRNA targets is an important task and has been undertaken with several tools. Many of these tools make use of RNA secondary structure folding algorithms, such as RNAFold [72], to estimate a maximum free energy of base-pairing of the miRNA to its putative target. Pairing of the 5' end of the miRNA, especially to the 2-8nt seed region, seems to be the most important interaction between a miRNA and its target [108]. This region is treated differently from the rest of the molecule by several of the algorithms.

The TargetScan [108] algorithm begins by selecting regions of the 3'UTR of human genes that are exactly complementary to the seed regions of a miRNA. Matches are evaluated only if they are conserved across five other species (human,

mouse, rat, chicken, and dog). Seeds are extended out into a larger alignment and that alignment is evaluated for folding free energy. Alignments are ranked and assigned a z-score. mRNAs whose alignment meets or exceeds a selected threshold are retained as predicted targets.

Pictar [101] splits miRNA-mRNA interactions into two categories: perfect matches and one-nucleotide mismatches. The algorithm then evaluates the binding free energy of the miRNA-mRNA alignment and retains those pairs that score higher than a user-selected threshold. These pairs are subject to further evaluation; the algorithm identifies highly conserved target sites of 4-5 nucleotides and removes pairs that do not contain a specific number of these anchors. The algorithm then uses a Hidden Markov Model (HMM) to evaluate the likelihood that the mRNA would be targeted by the miRNA in a combinatorial manner. These scores are then combined to make a final score for each mRNA/miRNA pair.

PITA [91] uses a free-energy calculation when considering the binding of a miRNA to a putative target site, but it also considers the accessibility of the target site within the folded structure of the mRNA. PITA calculates the free energy required for a miRNA to access its target site and the flanking nucleotides, and subtracts that score from the free energy gained from the base-pairing interaction of the miRNA with its target site.

Studies are conflicting with regards to which of these methods may be superior to the other [115, 127]. The overlap in target predictions among the tools is small, as each tool focuses on different known features of miRNA repression mechanisms to make its predictions. For instance, a tool that focuses on free energy calculations will result in a different set of predictions than a tool that

focuses on target sequence conservation across species or target sequence count in the 3'UTR of the targeted gene. These tools are not necessarily of different value, but their predictions will always be different. Furthermore, the functional themes of the predicted targets will also differ; only when functional themes are found enriched with very high significance ($p < 0.001$) can we expect different target prediction tools results to have consistency among one another [127].

mirSVR is a newer method that ranks predicted targets by regulatory power, using a method trained on measurements of real downregulation of gene targets by miRNAs transfected into cell culture. The algorithm uses local contextual features, such as seed flanking region AU content, target site position in the 3'UTR, UTR length, and site conservation [17].

1.6 Analysis tools for the study of genomic data

The study of cancer has been greatly advanced by the use of large-scale genomic assays, such as gene expression microarrays. These tools have allowed researchers to collect tens or hundreds of thousands of measurements against each biological sample, a broad survey of gene expression with previous techniques would provide only a few data points. However, this sudden wealth of data has often overwhelmed researchers accustomed to evaluating small datasets, and has required the development of a new suite of tools to analyze it.

One of the first typical analyses to be done after generating large-scale genomic data is to identify genes whose expression values change across experimental conditions. At first, simple statistical methods such as fold-change calculations, t-tests or Analysis of Variance was used to evaluate differential expression

between subgroups of samples. These tools have recently given way to more sophisticated methods such as Significance Analysis of Microarrays (SAM) [154].

SAM is a simple modification of the t-test used for identifying differentially expressed genes between two sample groups. It takes into account the fact that the variance of expression for each gene is different, and calculates a modified t statistic, called d , for each gene, and uses that statistic to test for differential expression.

For each gene i the test statistic d_i is computed that compares the expression levels of i between the two groups of samples. $d(i)$ is calculated as

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}$$

where $\bar{x}_I(i)$ and $\bar{x}_U(i)$ are defined as the average levels of gene expression for gene i in sample groups I and U , respectively. S_0 is selected to minimize the covariance of $d(i)$.

S_0 is computed as the value of the scatter statistic, The scatter statistic, $s(i)$, is calculated for each gene so that

$$s(i) = \sqrt{a \left\{ \sum_m [x_m(i) - \bar{x}_I(i)]^2 + \sum_n [x_n(i) - \bar{x}_U(i)]^2 \right\}}$$

where $\sum_m$ and $\sum_n$ are the sum of expression values in groups I and U, respectively, $a = (1/n_1 + 1/n_2)$ and n_1 and n_2 are the number of samples in groups I and U , respectively.

The coefficient of variation of $d(i)$ is computed as a function of $s(i)$ in moving windows across the data, and the value of $s(i)$ minimizing that coefficient is selected as s_0 .

Permuted d statistics are also calculated for each gene, wherein samples are shuffled between the groups, resulting in a permuted d statistic, $d_p(i)$ for some number P of permutations. An expected relative difference, $d_E(i)$, is calculated as the average expected difference in the permuted expression values over all P , or

$$d_E(i) = \frac{\sum_p d_p(i)}{P}.$$

Significance of differential expression is determined by the difference between d_E and $d(i)$, with non-significant genes having $d_E(i) = d(i)$ and significant genes having the difference between $d_E(i)$ and $d(i)$ greater than some threshold Δ . A false discovery rate (FDR) is set by selecting a value for Δ that results in the desired proportion of genes in the permuted dataset to be designated as significantly differentially expressed. That is, for those permuted gene expression values,

$$d_E(i) - d_p(i) > \Delta.$$

This method takes into account the fact that genes expressed at a low level are less affected by noise, and so smaller changes in their expression can be

more easily detected than highly-expressed genes. Other methods of differential expression detection that use simple fold-change cutoffs miss these genes.

After a candidate list of differentially expressed genes is identified, the challenge is to interpret that list in a biological context. Often the goal is to examine what, if any, biological themes are represented by that list of genes. The Gene Ontology consortium began this work by creating a set of annotations called Gene Ontology terms (GO terms) [6]. These terms are arranged in hierarchical trees, in which each term may have one or more parent or child terms. Three directed acyclic graphs, or DAGs, are rooted in three terms: Biological Process, Molecular Function and Cellular Component. Genes are manually annotated with terms from each of the three DAGs in an attempt to assign biological meaning to the genes. Other classification schemes are available as well, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway system [87, 89, 88].

These ontologies can be used to examine lists of genes for functional themes. One tool that is of particular use for characterizing long gene lists is the Expression Analysis Systematic Explorer (EASE), developed by the DAVID group [47, 76]. The examination of a given list of genes begins with the identification of all of the GO terms (or other functional classifiers) associated with any gene in the list, and the number of genes each GO term is associated with. This list of GO terms "hits" is compared to the number of similar "hits" for each term found in a background list of genes, in many cases the complete list of genes found on the DNA microarray from which the gene list was derived.

Another tool that makes use of functional classification is Gene Set Enrichment Analysis [146, 85]. This method ranks genes based on their ability to distinguish between two sample classes and then searches for over-represented

functional classes assigned to the genes at either end of the rank-ordered list. Functional classes can include GO term associations or KEGG pathway membership lists or other, specialized gene signatures contained in databases like MSigDb [146] or GeneSigDb [45].

Another task researchers set themselves is the prediction of clinical outcome or disease subtypes from a set of sample predictors, in this case gene expression measurements. Linear regression modeling and classification is a standard approach to this problem. However, the mathematics underlying these models relies on the number of predictors, or measurements, being roughly equal to or less than the number of samples. The vastly greater number of measurements available in genomic datasets leads to the problem referred to as the “curse of dimensionality”, which causes a problem with overfitting in linear models, and leads to instability of the resulting predictions.

Variable selection is one method of compensating for the curse of dimensionality [55]. Only a relative few of the most important variables are retained as part of the model. The choice of which variables to retain is sometimes accomplished through shrinkage, or applying a penalty term to the variable’s coefficient. Only those variables whose coefficients surpass a given threshold value are retained [150, 151]. Furthermore, until recently very few miRNA expression datasets were available for study. mRNA expression data, however, are abundant and easily accessible in public repositories.

Because of the lack of relationship between expression of miRNAs and their predicted targets, we decided to use the RE score method of Cheng, *et al.* [37] to examine miRNA activity indirectly. This method calculates a Regulatory Effect, or RE score for each miRNA at each sample, based on the expression of

its predicted mRNA targets. These scores can then be compared across samples to examine the differences in implied miRNA activity. A score is calculated for each miRNA, for each sample. This score is not informative in isolation or in comparing one miRNA to another; it is only useful as a means of comparing miRNA activity across samples.

1.7 Identification of an angiogenic subtype in Serous Epithelial Ovarian Cancer

Recently, Bentink *et al.* [15] have reported the discovery of two subtypes of serous ovarian cancer, having different survival times and being distinguished by a list of classifier genes enriched for angiogenesis-related Gene Ontology (GO) terms. These two subtypes, referred to hereafter as the angiogenic subtype and the non-angiogenic subtype, differ not only in survival time and gene expression profile, but in miRNA expression profile as well. A set of eight classifier miRNAs are able to reliably segregate the two subtypes. This subtype is similar to the C1 subtype described by Tothill *et al.* [153].

To identify these subtypes, Bentink *et al.* used the Identifying Splits with Clear Separation (ISIS) method to create candidate bipartitions in a set of gene expression data from ovarian tumors. They then validated significant bipartitions with clinical data and miRNA expression profiles.

The larger dataset, published by the public consortium The Cancer Genome Atlas (TCGA) and referred to here as the TCGA dataset, comprises expression measurements from 502 patients with high-grade (Grade 2 and 3) late-stage

(Stages III and IV) serous ovarian adenocarcinomas. The data were collected as described in [149], on an Agilent miRNA expression arrays and Affymetrix GeneChip whole-genome gene expression arrays. A more detailed description of these datasets is presented in the introduction to Chapter 2.

The Consortium found that mRNA expression and miRNA expression could be used to divide the patients into clusters: 4 for mRNA, and 3 for miRNA. These clusters showed some overlap, especially between miRNA cluster 1 and the Proliferative mRNA cluster, and miRNA cluster 2 and the Mesenchymal mRNA cluster. They also identified a prognostic list of genes that predicted good and poor survival. They also identified common mutations in ovaca, including TP53 (96% of samples) and BRCA1 and 2 (22%).

1.7.1 Identifying Splits with Clear Separation

ISIS is an unsupervised classification method whereby samples are split into two classes which are clearly distinguished by the differential expression of a small set of genes [160]. Therefore, ISIS is both a method for finding classes and also for determining the genes whose expression values distinguish them. The method can identify multiple independent bipartitions in the data, defined by different lists of differentially expressed genes. The method works by identifying candidate bipartitions and scoring them with a diagonal linear discriminant (DLD) score. These candidates are then used as starting points for a greedy search of a network of bipartitions, as measured by DLD score. The network is traversed following the path of highest DLD score until a local maximum is reached. A complete description of the method follows.

Identification of candidate bipartitions. An exhaustive search and scoring of all possible bipartitions (2^{p-1} , where p is the number of samples) would be computationally prohibitive. Instead, candidate bipartitions are generated using a fast heuristic. mRNA expression profiles are agglomeratively clustered with a Pearson correlation similarity metric, resulting in a rooted dendrogram wherein each node except the root has only one parent and each non-leaf node has two children. Each of the nodes (including leaf nodes) is used to define a cluster of expression profiles — the profiles belonging to that node and all of its descendants. This averaging results in a centroid expression vector for each of $2k - 1$ clusters, where k is the number of mRNA profiles used to create the dendrogram. Each centroid is then examined to determine whether it suggests a bipartition of the samples into two groups based on expression level. To do this, each sample in the centroid is used as a cutpoint, and those samples with higher expression levels than the cutpoint are assigned to one group, and those lower to another. A two-sided t-test is performed on those two groups and the test statistic compared to quantiles of the corresponding null distribution generated by Monte-Carlo simulation. The bipartition defined by this cutpoint is selected as a candidate for the next step of the algorithm if its t-statistic is higher than 95% of the values in the simulated null distribution. Those high-scoring bipartitions are used as starting points for the local maxima search. No other significance is associated with this bipartition and the membership of the gene cluster used to generate it is not retained.

Rank mRNAs for use in DLD scoring. For each candidate bipartition, the $n = 100$ mRNAs that are most differentially expressed between the two classes are identified using a two-sided t-test on the entire set of mRNA expression data.

The resulting p-values are used to rank the mRNAs. The 100 mRNAs with lowest scores are used to score the candidate bipartition with the DLD scoring method described below. This low-p-value group of mRNAs may or may not include the mRNAs whose expression profiles made up the centroid used in the heuristic step of this method. The DLD score evaluates the extent of the separation between the two sample groups.

Score bipartitions with DLDA. Diagonal linear discriminant analysis (DLDA) provides a normal vector defining the discriminant axis between the samples in the two candidate classes based on the expression of n top differentially expressed genes [52]. The linear combination $s = a_1x_1 + a_2x_2 + \dots + x_n$ aggregates the expression of the $n = 100$ top differentially expressed genes into a single expression score s for each individual sample. The vector $x_1, x_2, \dots, x_n$ denotes the expression levels of the top n differentially-expressed genes in one sample. The DLD score is the t-statistic comparing the aggregated univariate expression scores s in the two candidate classes.

Greedy search of complete bipartition network. The candidate bipartitions are used only as a starting point for finding bipartitions with locally-maximized DLD-scores in their neighborhoods. Consider the neighborhood as a graph whose nodes are all possible bipartitions of the sample set. This graph will be of size 2^{k-1} . An edge exists between two nodes (they are neighbors) if they differ only by the group assignment of one sample. Each candidate bipartition is used as a starting point and the graph is traversed in a greedy manner, where the DLD score is calculated for all adjacent nodes and the node with the highest DLD score is chosen. This step is repeated until the selected node has the highest score of all adjacent nodes.

1.7.2 Verification of the subtypes

Bentink *et al.* used ISIS to identify four bipartitions yielding class predictions consistent among a set of perturbed binary classifiers on an independent validation set of samples. Those four bipartitions, each supported by a different set of mRNAs (modules 1-4), were examined for clinical relevance and also examined in two other clinical datasets of serous ovarian cancer [50, 153]. One of the bipartitions (module one) demonstrated a difference in patient survival between the two groups, though this survival difference was only significant in one of the datasets examined (Bentink, $p = 0.0715$, Tothill, $p = 0.0013$, Dressman $p = 0.1142$).

Module one splits a smaller group of patients ($n = 33$) associated with a shorter survival time from the remainder ($n = 96$). Bentink *et al.* used GSEA to identify gene signatures associated with the differences in gene expression between the samples in this group. GSEA is a linear model-based method of testing for enrichment of pre-defined gene signatures within the differentially expressed genes across two groups of samples [146]. They found strong associations with angiogenesis and extracellular matrix proteins, including GO categories describing genes involved in vascular development and regulation of cell adhesion, which showed relatively high expression levels in the smaller of the two classes. The larger sample class within this module showed relatively high expression for genes involved in single stranded DNA binding and structure specific DNA binding. The smaller of the two groups will be referred to here as the angiogenic class, and the other as the non-angiogenic class. The mRNAs comprising Module one are hereafter referred to as the *classifier mRNAs*.

The class membership of Module one was verified using miRNA expression

values. PAM analysis of miRNA expression data collected from the 129 samples resulted in a classifier consisting of six miRNAs, which successfully classified samples into the two groups with low error rates (< 3 misclassifications for the angiogenic type, and a rate of $< 20\%$ for the non-angiogenic type). The six miRNAs chosen by PAM as classifiers fall into two miRNA families: miR-202 (miR-202 and 202*) and miR-506 (miR-506, 508, 509, and 514).

1.8 Ovarian cancer subtypes

The genes involved in the subtype differences are related to angiogenesis and the extracellular matrix, but little else is known about the driving forces behind the differences in the two subtypes; for instance, little is known about the relationship between subtype and chemotherapeutic sensitivity.

Further, the work of Bentink *et al.* corresponds well with that of other researchers investigating subtypes of ovarian cancer; the angiogenic subtype overlaps strongly with the C1 or “stromal reactive” subtype of Tothill [153] and the mesenchymal and immunoreactive subtypes identified by The Cancer Genome Atlas (TCGA) consortium [149]. What remains to be discovered is the relationship among these differentially expressed miRNAs and mRNAs, and their biological roles in the subtypes.

The TCGA consortium published gene expression and sample-matched miRNA expression data for a set of 500 serous ovarian cancer samples, and claiming identification of four robust cancer subtypes, (“differentiated”, “immunoreactive”, “mesenchymal” and “proliferative”). However, these subtypes show no difference in survival time while the Bentink subtypes do. The angiogenic subtype identi-

fied by Bentink *et al.* are significantly overlapping with two of these groups, the immunoreactive and mesenchymal subtypes.

1.9 The significance of this work

The remainder of this document is a report of the work done to investigate the role of miRNAs in ovarian cancer.

Chapter Two is a brief survey of the relationships between miRNA and mRNA expression profiles in the Bentink and TCGA datasets. We examine the partial correlation structure of miRNA expression with mRNA expression and report on patterns in this structure in several subsets of the data. We find that the datasets show consistency in their correlation structure, and in the specific miRNA-mRNA pairs that are either highly positively or negatively correlated. The data include a larger number of strong positive and strong negative correlations than would be expected by chance, indicating that biological relationships between the types of data are detectable in these datasets. We further find an enrichment for positively-correlated miRNA-mRNA pairs in which the miRNA is encoded in close proximity to the mRNA. The correlation of miRNA and mRNA is apparently unaffected by miRNA and mRNA expression level; similarly the two molecular subtypes do not contain differences in their correlation.

In Chapter Three, we describe our analysis of the two ovarian cancer subtypes discovered by Bentink *et al.* [15]. We find that the poorer prognosis, or angiogenic, subtype has a generally lower miRNA activity than the second, non-angiogenic, subtype. The subtype is characterized by a consistent pattern of differential miRNA expression. We also report on a switch-like relationship

between the expression levels of certain miRNAs and genes that are anticorrelated with them. We propose these miRNAs drive many of the differences in the subtypes both directly, by RISC-mediated repression of target messages, and indirectly, by repressing transcription factors that regulate expression in the cell.

Chapter 4 seeks to build a model of survival based on miRNA expression. We progress from a simple univariate analysis of miRNA expression and RE score measurements, through simple variable selection, to a more sophisticated variable shrinkage method: the Lasso. We find essentially no survival-predictive information provided by the RE score data. Furthermore; while direct miRNA expression measurements may contain some predictive power, a large dataset and the segregation of that dataset into distinct molecular phenotypes is likely to be necessary to produce a useful model of survival in ovarian cancer.

Chapter 2

Exploration of the correlation of miRNA and mRNA expression and miRNA activity in ovarian cancer.

2.1 Introduction

Our interest in miRNAs in ovarian cancer prompted an analysis of the connection between miRNA and mRNA expression observed in ovarian tumor tissue. MicroRNAs (miRNAs) are known to affect the protein production of their target genes through both translational repression and transcript destabilization. While it is true that changes in protein production due to message sequestration are not directly detectable via microarray technology, the destabilization of target transcripts is easily measured and has been reported to account for as much as

84% of the suppressive effect of a miRNA on its target [67, 7]. Given this phenomenon, it should be possible to detect miRNA-mRNA interaction in cells by transcriptional profiling alone.

In this chapter, we examined the expression patterns of miRNAs and mRNAs in two datasets, published respectively by Bentink *et al.* at the Dana-Farber Cancer Institute [15], and by The Cancer Genome Atlas (TCGA) consortium [149]. These datasets were collected from high-grade serous ovarian tumors and hybridized to DNA and miRNA microarrays, resulting in sample-matched expression profiles of genes and miRNAs.

2.1.1 Datasets

The Bentink dataset was published in the ArrayExpress data repository (ArrayExpress Experiment Accession E-MTAB-386) and is described in full in [15]. Briefly, tumor samples were acquired from patients undergoing debulking surgery for high-grade serous endometrial cancer at the Dana-Farber Cancer Institute between 2005 and 2009. Both total mRNA and miRNA were extracted from FFPE blocks. Total mRNA was hybridized to an Illumina DASL BeadArray containing approximately 12,000 selected mRNAs, or 9,000 genes (ArrayExpress Array Design Accession A-MEXP-931). miRNA was hybridized to a prototype DASL-based miRNA expression profiling BeadArray containing probes to 743 miRNAs (ArrayExpress Array Design Accession A-MEXP-1678). Data were normalized using variance stabilizing and robust spline normalization as implemented in the Bioconductor lumi package [51, 110]. Each sample was clinically annotated with information regarding tumor stage, patient age, survival time post-diagnosis, and

chemotherapy treatment status.

The second dataset was published by The Cancer Genome Atlas [149] and includes matched mRNA and miRNA expression data from 502 clinical samples of serous ovarian cystadenocarcinoma. 723 miRNAs and 22,000 expression profiles for 12,000 genes were measured for each sample. Gene expression data were collected with the Affymetrix HG-U133A expression array at the Broad Institute. The miRNA data was collected at University of North Carolina using Agilent arrays. All data are quantile normalized and available on the TCGA website at <http://tcga-data.nci.nih.gov/tcga/>. While TCGA is a multi-site consortium, the data used here were each collected from only one of the member institutions.

One concern about using data collected with two different expression platforms involves the difference in the genes measured. The Affymetrix GeneChip platform used for the TCGA study is a whole-genome array, whereas the Illumina custom array used in the Bentink study is more limited, measuring only 9,000 genes. However, there is substantial overlap between the gene sets measured by the two chips - when mapped to Entrez Gene IDs, they contain 8465 identical mRNAs. Of the miRNAs in the two platforms used, 368 miRNAs are common between the two datasets.

The initial analysis of these datasets was aimed at determining whether the measurements would be useful for investigating the role of miRNAs in ovarian cancer biology. Before models could be constructed using these data, they must be shown to be informative. It would be helpful to know if predicted miRNA-target interactions were detectable at the expression level. Blower *et al.* [20] found that there is no difference in correlation between verified miRNA-mRNA pairs and randomly-selected miRNA-mRNA pairs. However, Creighton *et al.*

[43] found an enrichment for predicted target genes in high-correlation transcriptional profiles. In this analysis we tested for the presence of inverse patterns of expression between miRNAs and the mRNAs they had been predicted to target. We also examined patterns differing from random in the correlation of all miRNA and mRNA expression profiles.

We were also interested in the relevance of these data to the biological processes that impact the development and progression of ovarian cancer. Given the important role of miRNAs in controlling aspects of cell differentiation and growth, connections between miRNA and mRNA expression profiles may inform current models of regulatory controls within these tumors. Our interest is in finding miRNAs with strong regulatory connections to mRNAs in these data, hopefully leading to a more informed model of gene regulation in ovarian tumors. This would illuminate the mechanisms of disease and potential targets for its eventual cure. In our preliminary analysis we survey the miRNAs and mRNAs with strong expression profile relationships for relevance in the fields of ovarian cancer, or cancer in general.

Lastly, we were concerned with consistency between the two datasets. Though they were taken from similar tissues and processed in a similar manner, it was necessary to evaluate whether the datasets were comparable and consistent. If they showed great differences in correlation profile, or few common high- or low-correlation miRNA-mRNA pairs, the data may not be useful for a bioinformatic analysis.

In this chapter we explore the data that we will use for subsequent analyses and model-building. We verify that some connection between miRNA and mRNA expression profiles exists; more high and low correlations exist in these data than

would be expected by chance. We verify the findings of Blower *et al.* [20] that the expression patterns of miRNAs are no more correlated with their predicted targets' expression patterns than randomly-selected sets of miRNA and mRNA expression profiles. We find an enrichment for large positive miRNA:mRNA correlation coefficients that cannot be explained by simple miRNA-gene co-location. We also find a pattern of consistent enrichment for large positive miRNA-mRNA pairs in the angiogenic subtype that is not present in the non-angiogenic subtype, and a consistent set of miRNA-mRNA pairs with large negative correlations across the two datasets.

2.2 Results

2.2.1 Some similarity in miRNA-mRNA partial correlation in the Bentink and TCGA datasets

We approached our first analysis simply by examining the distribution of correlation coefficients of each miRNA expression profile with each mRNA expression profile measured in the Bentink and TCGA datasets (Figure 2.1a and c). We expected a large proportion of these correlations to be near zero; a randomly chosen mRNA-miRNA pair is unlikely to have a biological interaction in the cell leading to a measurable correlation, positive or negative. miRNA action is tissue- and target- specific, and the number of true miRNA-mRNA actions, while possibly large, is still likely to be much smaller than the complete set of possible miRNA-mRNA combinations we examined. To reduce the indirect effects that must appear in such correlations of large networked systems such as gene ex-

pression networks, we calculated partial correlations using the shrinkage method of Schafer and Strimmer, as implemented in the R package **corpcor** [131, 126]. This method is well-suited to large datasets; its runtime is very low (a few hours on a standard desktop computer, for our data) when compared to other methods like glasso [60], with a runtime of multiple days. This method calculates the partial correlation of each pair of miRNA or mRNA molecules conditioned on the remaining variables. The partial correlations $\text{corr}(X_i, X_j | \text{rest})$ is the correlation between X_i and X_j conditioned on the remaining variables. It equals the sign-reversed entries of the off-diagonal entries of the precision matrix, standardized by the the squared root of the associated inverse partial variances.

The matrix of partial correlations $P = (p_{ij})$ is related to the inverse of the covariance matrix Σ [163]. This leads to the estimator of covariance

$$r_{ij} = -\hat{\omega} \sqrt{\hat{\omega}_{ii} \hat{\omega}_{jj}},$$

where

$$\hat{\Omega} = (\hat{\omega}_{ij}) = \hat{\Sigma}^{-1}.$$

The method shrinks the empirical correlation matrix R_{ij} towards the identity matrix, using the shrinkage intensity of λ . This shrinkage intensity is estimated using an analytic formula from Shaefer and Strimmer [131];

$$\lambda^* = \frac{\sum_{i \neq j} \text{Var}(r_{ij})}{\sum_{i \neq j} r_{ij}^2}.$$

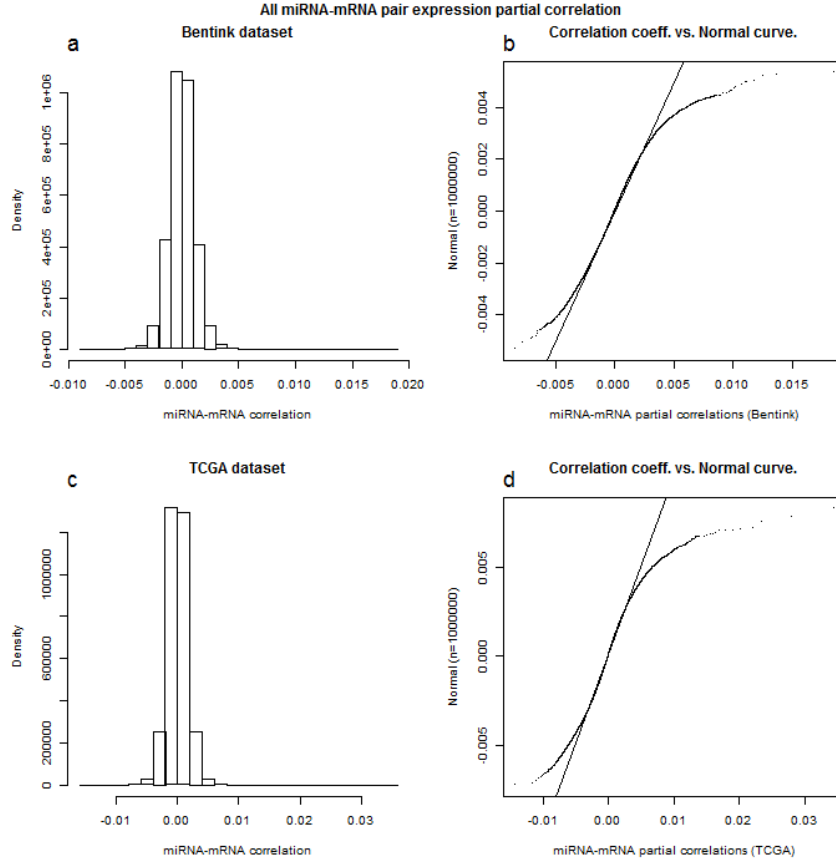


Figure 2.1: a) Distribution of partial correlation coefficients of the expression values of each possible miRNA-mRNA pair across all samples in the Bentink dataset. b) Q-Q plot of the partial correlation coefficients in a, against a generated normal distribution of identical mean and standard deviation. c) Distribution of partial correlation coefficients of the expression values of each possible miRNA-mRNA pair across all samples in the TCGA dataset. d) Q-Q plot of the partial correlation coefficients in c, against a generated normal distribution of identical mean and standard deviation.

We calculated partial correlations between all possible pairwise combinations of mRNA and miRNA expression values. We retained only the miRNA-mRNA pairs, rather than mRNA-mRNA or miRNA-miRNA pairs, resulting in 9,249,607 correlation coefficients for the Bentink dataset and 16,106,271 for the TCGA

dataset. The distribution of these correlations (Figure 2.1a and c), with a mean of $\mu = 0.00092$ and standard deviation of $\sigma^2 = 0.13$, appears normal, so we compared it to a true normal distribution of identical mean and standard deviation. A Q-Q plot of these correlations vs. the normal distribution (Figure 2.1b and d) has a characteristic S-shaped curve, indicating that this distribution has heavier tails than the normal distribution. That is, more large positive and large negative correlation coefficients were present in our calculated distribution than in the one predicted by a normal distribution.

We quantified these tails by calculating moment ratios and comparing those moment ratios to the expected value for a moment ratio in a normal distribution. A moment is a quantitative measure of a distribution, used to characterize the shape of that distribution. A normal distribution has certain characteristics, such as a first moment (μ_1) of zero and a second moment (μ_2) equal to the variance. Normal distributions have an expected relationship between the values of the various higher moments, which can be used to determine whether the tails of the distribution are heavier than expected. A normal distribution will have a ratio of the fourth moment (μ_4) to the square of the second moment equal to $\mu_4/(\mu_2)^2 = 3$. Distributions with heavier tails will have a larger ratio than expected. All of the observed moment ratios for distributions investigated in this Chapter are summarized in Table 2.1.

We tested these moment ratios for significant difference from the expected value by simulation. We generated sets of normally-distributed data with the same mean and standard deviation as the observed data. We calculated moment ratios for these randomly-generated datasets and calculated empirical p-values for our observed ratios by counting the number of random datasets that had

moment ratios more extreme than the observed data.

	$4th : 2nd^2$ Ratio	$6th : 2nd^4$ Ratio	$8th : 2nd^6$ Ratio
Normal Curve (Expected)	3.00	15.00	105.00
Bentink partial correlations	4.07	61.45	4979.00
TCGA partial correlations	4.96	191.96	50966.34
Bentink predicted partial	4.06	60.19	3274.24
TCGA predicted partial	5.21	246.93	44348.73
Bentink > 10,000kb	4.04	60.04	4719.79
TCGA > 10,000kb	4.62	137.42	33503.11
Bentink High mRNA Expr	3.70	39.37	2081.02
Bentink Low mRNA Expr	4.15	67.55	5472.00
TCGA High mRNA Expr	4.51	130.49	28403.45
TCGA Low mRNA Expr	4.43	116.90	26927.76
Bentink High miRNA Expr	3.65	32.51	1229.60
Bentink Low miRNA Expr	4.21	71.51	5896.11
TCGA High miRNA Expr	4.73	344.92	146761.34
TCGA Low miRNA Expr	4.46	77.52	9010.76
Bentink Angiogenic	3.43	23.37	277.87
TCGA Angiogenic	3.88	40.99	1371.14
Bentink Nonangiogenic	3.93	57.99	4346.14
TCGA Nonangiogenic	4.31	79.72	10137.50

Table 2.1: Expected and observed moment ratios for the partial correlation sets of this chapter.

We find here that the fourth moment ratio of the partial correlations calculated on the Bentink dataset is $r = 4.07$. For the TCGA dataset, the ratio is higher, $r = 4.96$. Higher moments, the sixth and eighth, have proportionally even higher ratios than the expected values of 15 and 105 (Table 2.1). Each of these observed ratios is larger than expected by chance ($p < 0.05$).

These 'heavy tails' could represent true interactions between miRNAs and mRNAs. Negative correlations could result from a miRNA repressing its target directly or indirectly, whereas positive correlations could represent only indirect regulatory relationship — an undetected true mRNA target message that pro-

duces a protein product repressing transcription of another gene, for instance. A strong positive correlation could also represent the regulation of the mRNA and miRNA by a common third actor, such as co-transcription.

This excess of positive and negative correlations could result from biologically relevant interactions among the molecules. It is encouraging that both the TCGA and Bentink datasets show similar trends in correlation coefficient distribution: they both include a larger number of large positive correlations than expected in a normal distribution. The TCGA dataset also shows a larger number of large-negative correlations than would be expected by chance, however the Bentink dataset does not. This may be simply because the Bentink dataset is smaller, or because those correlations are not detectable in the dataset. These findings show that the two datasets have some consistency, and that there is meaningful information in the data beyond noise; there is a relationship between the miRNA and mRNA expression values and that relationship is easier to detect (the moment ratios are larger) in the larger TCGA dataset.

In order to determine whether one dataset simply had larger partial correlations, overall, than the other, we matched miRNA-mRNA pairs from each dataset and plotted the partial correlations (Figure 2.2). We find that the correlations are clustered around the line of identity, indicating, again, that the distributions of correlations are similar and that pairs tend to have similar correlations in both datasets. This clustering, however, is not perfect; even qualitatively we can see that the TCGA dataset has a wider spread of partial correlations.

The sheer number of miRNA-mRNA pairs visualized here makes examining these relationships difficult, so next we sought to reduce the search space of possible combinations by using miRNA target prediction tools.

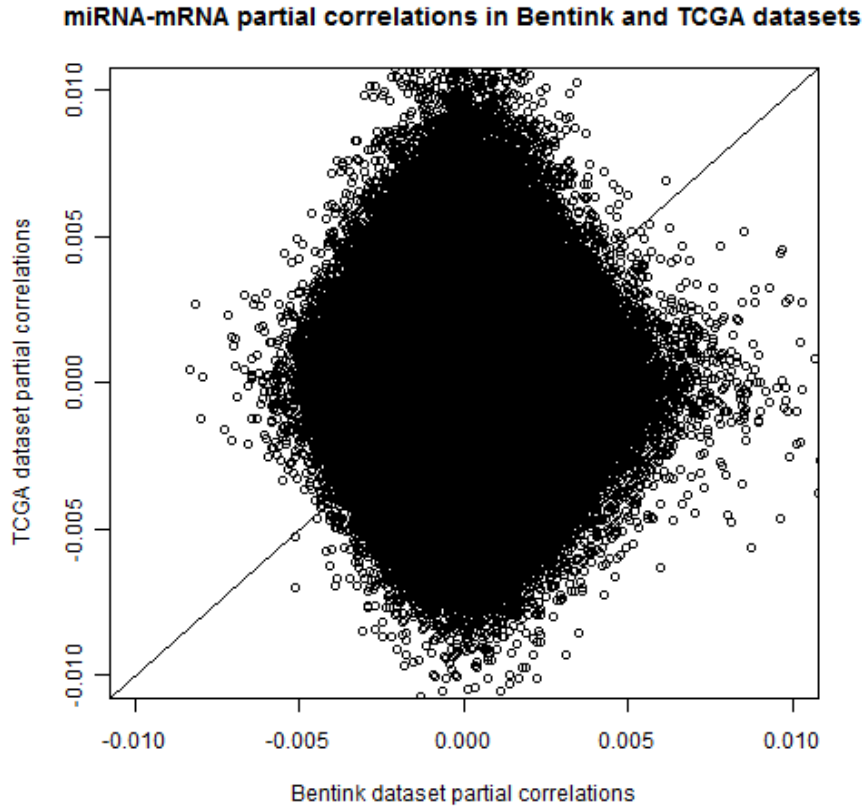


Figure 2.2: Correlation of miRNA-mRNA pairs measured in both the Bentink and TCGA datasets.

2.2.2 Predicted miRNA-targets have similar partial correlation profile to all possible targets

We expect a relatively small number of valid molecular interactions occurring within the cell. Of the 9,249,607 possible combinations of species interactions, only a portion are likely to be occurring at any given time. In addition, most miRNAs are thought to target only a subset of messages in the cell (between 6 and 2191, as reported by the PITA target prediction tool). To somewhat shrink

our search space, we restricted our correlation calculations to predicted target interaction pairs. We used the PITA target-prediction tool, described in Chapter 1 and by Kertesz *et al.* [91], to identify 85,886 (Bentink) and 197,223 (TCGA) miRNA-mRNA interaction pairs. We calculated partial correlation coefficients for the expression profiles of these pairs (Figure 2.3a and c). Again, the characteristic S-shaped curve appears; this correlation distribution also includes more large positive and large negative correlations than a normal distribution of the same mean and standard deviation.

We see again that while the Bentink dataset has only a very small curve to its Q-Q plot, the TCGA dataset has a more distinct S-curve. In assessing both of these correlation sets — the complete all vs. all partial correlations (Figure 2.1) and the filtered set of only PITA-predicted correlations (Figure 2.3) — we observe the same pattern: more large positive and negative partial correlations are present than would be expected in a normal distribution of the same mean and standard deviation, in both datasets. The moment ratios for these distributions are similar to the non-filtered partial correlation distributions (Bentink unfiltered 4.07 vs. filtered 4.06, TCGA unfiltered ratio: 4.96 vs. filtered 5.21). We conclude, therefore that retaining only predicted target pairs is not particularly helpful for our analysis. Further work in this chapter is not filtered based on target prediction algorithms.

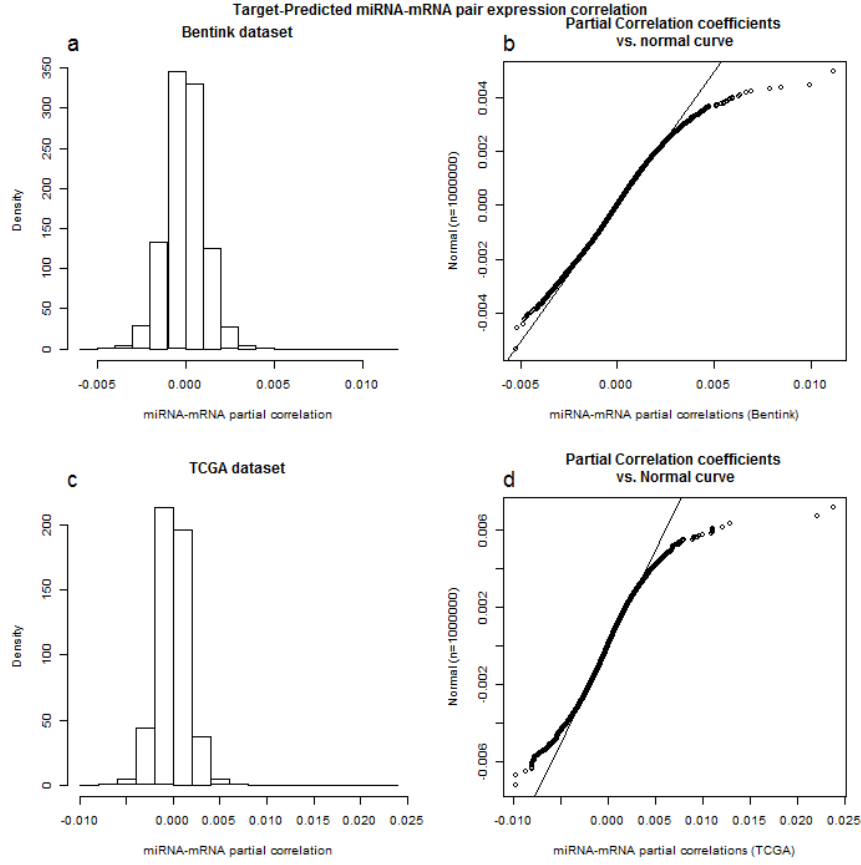


Figure 2.3: a) Distribution of partial correlation coefficients of expression values in the Bentink dataset, of all PITA-predicted miRNA-mRNA pairs. b) Q-Q plot of distribution in a, against a normal distribution of the same mean and standard deviation. c) Distribution of partial correlation coefficients of expression values in the TCGA dataset of all PITA-predicted miRNA-mRNA pairs. d) Q-Q plot of distribution in c, against a normal distribution of the same mean and standard deviation.

2.2.3 Correlation of miRNA and mRNA expression are driven in part by co-location on the chromosome.

The large number of highly positively correlated mRNA-miRNA pairs attracted our attention. An informal look at these high-correlation pairs (Tables A.1 and

A.2) results in a few observations. First, several of the mRNA-miRNA pairs are co-located on the chromosome; miR-196b is encoded within the HOXA9 gene, miR-224 is encoded in an intron of GABRE on the X chromosome, and miR-10a is encoded within the HOXB3 gene’s intron 1 on Chromosome 17. It has previously been shown that the expression profiles of miRNAs encoded within an intron are positively correlated with the expression profile of that gene [20]; some proportion of the large positively-correlated pairs may be due to an enrichment of miRNAs co-located with a mRNA. This is a common arrangement for miRNAs, leading to simultaneous transcription of both miRNA and mRNA and highly correlated expression profiles [20]. miRNAs encoded near a gene but not within its intron could be affected by locus-sensitive effects, such as methylation, histone acetylation or chromosomal lesions.

These common themes across the two datasets are encouraging; they indicate that biological information is present in these data, and that the information is consistent across the two datasets. This is an important characteristic for future analyses.

2.2.4 Co-located miRNA-mRNA pairs do not account for excess large-positive partial correlations

As we are interested in the regulatory activities of miRNAs, we removed from our analysis those miRNA-mRNA pairs that are located close together on the chromosome. If co-located pairs are the source of the large number of positive correlations, removing these pairs would reduce the magnitude of the right-hand-side tail in our correlation distribution while preserving the correlation coefficients of

many miRNA-mRNA interactions mediated by direct miRNA repression.

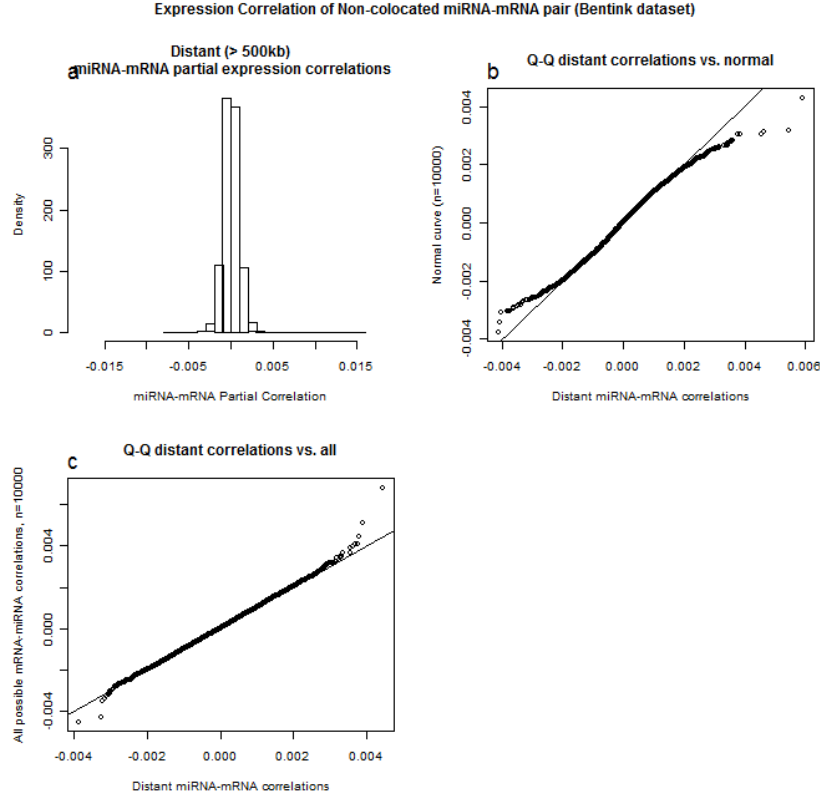


Figure 2.4: a) Distribution of miRNA-mRNA correlations in the Bentink dataset with miRNA-mRNA pairs separated by at least 10,000kb. b) Q-Q plot of those correlations, against a randomly-generated normal distribution of the same mean and standard deviation. c) Q-Q plot of those correlations against the unfiltered correlations.

All correlations calculated for the miRNA-mRNA pairs located within 10,000kb of one another were removed from the analysis (miRNA annotation from mirbase, [2, 65, 66]; mRNA annotation from Bioconductor [62]). The cutoff of 10,000kb was chosen because it is far greater than the length of the largest human genes (~2,000kb, [157]), and is long enough to include large swaths of upstream pro-

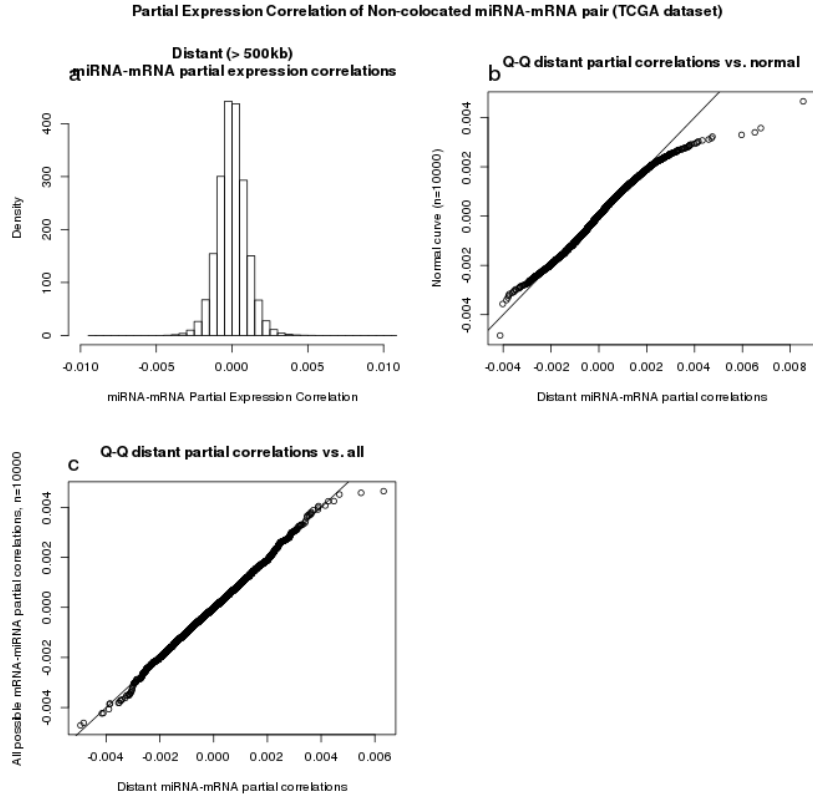


Figure 2.5: a Distribution of miRNA-mRNA correlations in the TCGA dataset with miRNA-mRNA pairs separated by at least 10,000kb. b) Q-Q plot of those correlations, against a normal distribution of the same mean and standard deviation. c) Q-Q plot of those correlations against the unfiltered correlations.

moter sequences as well. Therefore this length would ensure that pairs including a mRNA and a miRNA encoded within its intronic region would be removed. This window will remove pairs that share a methylation state as well; methylation depletion domains are typically around 4kb in size [128] and CpG islands of enriched methylation are even smaller, around 300bp [1].

Further, miRNA-mRNA pairs containing one or more members with missing locus information were also removed, leaving only miRNA-mRNA pairs anno-

tated with locations separated by at least 10,000kb. A total of 9,876 pairs were removed from the Bentink dataset and 14,723 from the TCGA dataset. This filtered set of correlation coefficients (Figures 2.4a/b and 2.5a/b) has an apparently identical distribution to the unfiltered set, an unsurprising finding given the small number of miRNA-mRNA pairs removed, compared to the total number of pairs.

We see that the fourth moment ratio is reduced by the removal of the closely-located miRNA-mRNA pairs, from 4.1 to 4 in the Bentink dataset, and from 5 to 4.6 in the TCGA dataset. This reduction in moment ratio is also apparent when the sixth and eighth moment ratios are calculated, (Bentink 6th moment changes from 61 to 60, Bentink 8th: 4979 to 4720, TCGA 6th: 192 to 137, TCGA 8th: 50966 to 33503, Table 2.1). While these ratios are lower than the ratios for the unfiltered partial correlations, they are still higher than the expected values, indicating heavier tails than a normal distribution ($p < 0.05$).

The top positively- and negatively-correlated pairs removed from the datasets are summarized in Tables A.3 through A.6. We find the large positively correlated pairs to be dominated by homeobox (HOX) genes paired with miR-10a/b and miR-196b. These pairs are, indeed, encoded in overlapping regions, miR-10a being encoded within the intron of HOX6B, for instance. The negatively correlated pairs are not dominated by such genes, and have less extreme partial correlations than the positively-correlated pairs (Bentink largest positive: 0.011, TCGA largest positive: 0.02, Bentink largest negative: -0.0054 , TCGA largest negative: -0.0052).

While the removal of these pairs does somewhat reduce the size of the heavy positive tail, it does not remove it entirely. We conclude that the excess large-

positive correlation coefficients are not mainly due to miRNA-mRNA pairs located within 10,000kb of each other on the chromosome, which are presumably affected by location-specific effects, such as co-translation, localized methylation or deletions/amplifications.

2.2.5 Biological consistency in miRNA-mRNA pairs with large negative correlation coefficients

As we are most interested in direct regulation of mRNAs by miRNAs, we looked more closely at the strongly negatively correlated miRNA-mRNA pairs in the left-hand tail of the correlation distribution. We examined the pairs of miRNAs and mRNAs with large negative expression profile correlation (Tables A.7 and A.8). We found a common set of miRNAs, miR-141, miR-200c, miR-629, miR-553, miR-606, and miR-302b (Tables A.7 and A.8) highly anti-correlated with mRNAs in both datasets. The genes were not common across datasets.

miRNAs 141 and 200c are members of a well-known family with a role in the epithelial mesenchymal transition, implicated in ovarian and and other female reproductive cancers [116, 132, 41, 13]. miR-200c is anticorrelated in the TCGA dataset with BCAS1, a gene amplified in a variety of tumor types [11]. The miRNA is anticorrelated with a puzzling gene in the Bentink dataset, however; NPPB, a cardiac hormone implicated in cardiovascular homeostasis [71]. miR-141 is also anticorrelated with NPPB in the Bentink dataset, as well as TM4SF4, a trans-membrane signaling protein implicated in liver metastasis [78] and with Fibroblast Growth Factor (FGF2), a well-known player in cancer invasion [103], in the TCGA dataset.

miRNAs 629 and 606 were discovered in a colorectal cell line, but little else is known about them[46]. In our data, miR-606 is anticorrelated with HR in the Bentink dataset and IKZF1 in the TCGA dataset. miR-629 is anticorrelated with MFSD11, a mostly uncharacterized gene, in the Bentink dataset and TLE1, implicated in synovial sarcoma [145], in the TCGA dataset.

These findings are encouraging in that they suggest a reasonable biological story may exist in the correlation data of these samples. A further examination of the individual genes and miRNAs will occur in Chapter 3. For the remainder of this chapter this set of correlation calculations for all possible miRNA-mRNA pairs separated by at least 10,000kb is investigated.

2.2.6 miRNA-mRNA correlation of predicted miRNA-mRNA target pairs is not dependent on median mRNA expression.

Having determined that some miRNA-mRNA transcription profile relationships are detectable at the transcriptional level, we hope to uncover additional patterns in miRNA-mRNA correlation. One possible pattern that could be uncovered in the data is a connection between correlation of miRNA-mRNA pairs and the expression levels of the molecules. For instance, highly-expressed miRNAs may be more effective at regulating their targets than lowly-expressed miRNAs. In addition, it may be the case that miRNA-based regulation of gene expression is not detectable if concentrations of the target mRNA are too high; an over-saturation of target molecules may overwhelm the capacity of miRNAs to downregulate their targets. The miRNA may not, in that case, be present in sufficient quan-

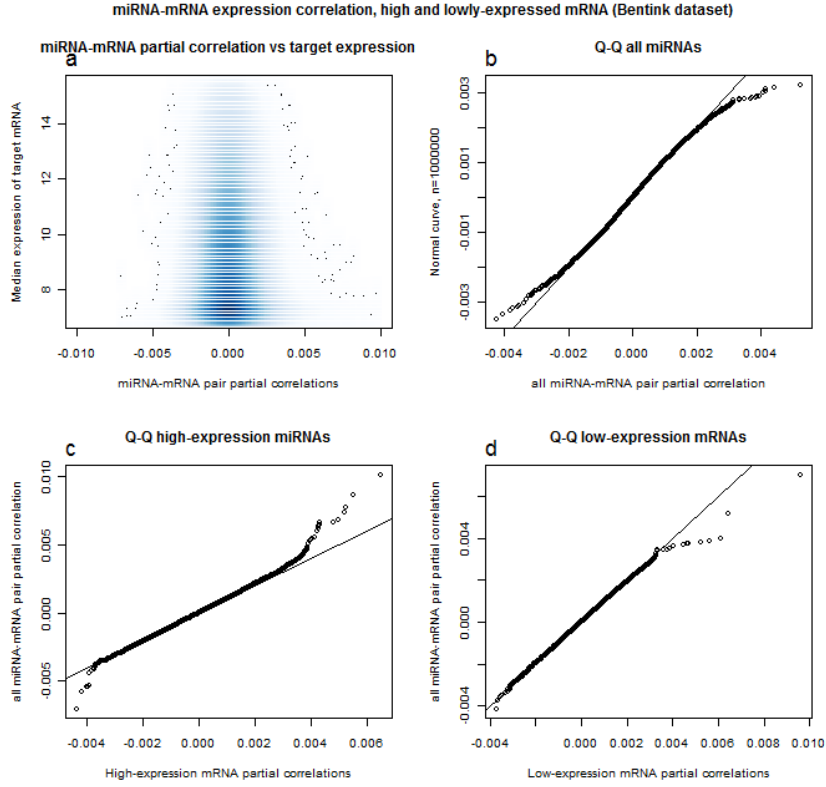


Figure 2.6: Partial correlations broken down by mRNA expression values for the Bentink dataset. a) Median mRNA expression values vs. partial correlation of miRNAs with their predicted targets. Each point represents a single mRNA-miRNA predicted pair. The majority of the points are shown as a cloud (light blue). b) Comparison of the distribution of partial correlation coefficients for all non-located miRNA-mRNA in the Bentink dataset, against a normal distribution of the same mean and standard deviation. c) Comparison of the distribution of partial correlation coefficients for the high-expression (mean mRNA expression > 8.5) non-located miRNA-mRNA pairs in the Bentink dataset. d) Comparison of the distribution of partial correlations for the low-expression (mean mRNA expression < 8.5) non-located miRNA-mRNA pairs in the Bentink dataset.

tity to affect the concentration of target messages in the cell. In the next two sections, we examine the change in distribution of correlation coefficients for

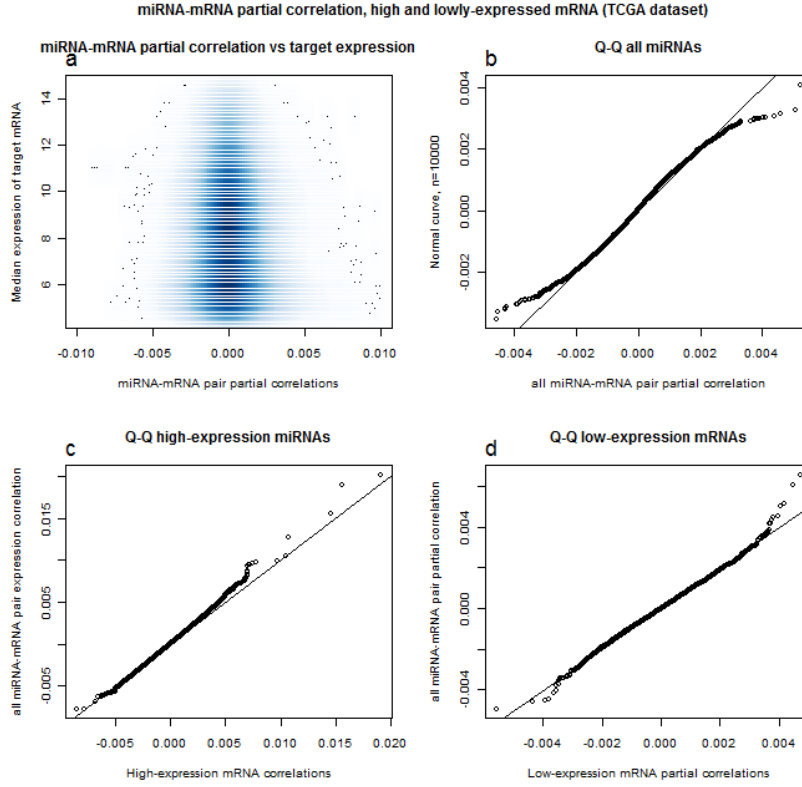


Figure 2.7: Partial correlations broken down by mRNA expression values for the TCGA dataset. a) Median mRNA expression values vs. partial correlation of miRNAs with their predicted targets. Each point represents a single mRNA-miRNA predicted pair. The majority of the points are shown as a cloud (light blue). b) Comparison of the distribution of partial correlation coefficients for all non-located miRNA-mRNA pairs in the TCGA dataset, against a normal distribution of the same mean and standard deviation. c) Comparison of the distribution of partial correlation coefficients for the high-expression (mean mRNA expression > 7.4) non-located miRNA-mRNA pairs in the TCGA dataset. d) Comparison of the distribution of partial correlation coefficients for the low-expression (mean mRNA expression < 7.4) non-located miRNA-mRNA pairs in the TCGA dataset.

miRNA-mRNA pairs when one member of that pair is expressed at higher or lower levels.

Multiple miRNAs often bind the 3'UTR of target messages, with more bound copies resulting in greater repression [96, 74]. We suppose that highly-expressed mRNAs may be affected less by miRNA-mediated repression than lowly-expressed mRNAs, since an mRNA species with few copies present in the cell could be bound by an miRNA in multiples with fewer miRNAs.

We examined mRNA target expression vs. partial correlation of profiles in the Bentink and TCGA datasets (Figures 2.6a and 2.7a, respectively). Each point represents a single miRNA-mRNA pair. The points center around $r = -1.6e - 05$ (Bentink) and $r = 0.00031$ (TCGA) and are approximately symmetrically distributed. In these figures the slightly larger number of points on the right-hand side of the plot seems to be evenly distributed across all mRNA expression levels.

Q-Q plots comparing high-expression mRNAs (above the median, Figures 2.6c and 2.7c) versus the whole population have heavier tails than the normal distribution (fourth moment ratio 3.7 and 4.5, respectively), but have lighter tails than the distribution of all partial correlations, which are 4 and 4.6, respectively (Table 2.1). A similar pattern is evident for low-expression mRNAs (below the median, Figures 2.6d and 2.7d). The fourth moment ratio for low-expression Bentink pairs is 4.2 and the ratio for TCGA is 4.4.

mRNA/miRNA expression correlation is not dependent upon miRNA expression level

Having found no difference in miRNA-mRNA expression correlation between high-expression and low-expression target mRNAs, we looked at the influence of miRNA expression level on miRNA-mRNA expression correlation. There may

be several reasons why miRNA concentration affects miRNA-mRNA expression correlation profiles in these data. It may be that lowly-expressed miRNAs are unable to exert a detectable effect upon gene expression. They may not function unless present in sufficient quantity to affect the concentration of target message in the cell. We may expect miRNA with higher expression levels to have a greater activity; more miRNAs in the cell can bind and suppress more targets.

We again compared partial correlation coefficients for miRNA-mRNA pairs located at least 10,000kb apart to the median expression level of the miRNA (Figures 2.8a and 2.9a). To determine whether high miRNA expression caused more repressive effects, we examined the miRNA-mRNA partial correlation distribution for highly- and lowly- expressed miRNAs, as measured by their median expression level (Figures 2.8a and 2.9a). The median expression values are centered around $\mu = -0.0011$ with a standard deviation of $\sigma^2 = 0.03$. The TCGA data have an average partial correlation of $\mu = -8.5e-06$ and standard deviation of $\sigma^2 = 3.4e-05$.

We see little change in the distribution of partial correlations when the dataset is filtered to retain only high- or low- expression miRNAs. The fourth moment ratios for these distributions are changed somewhat, from 4 to 3.7 for the high miRNAs in the Bentink dataset (Figure 2.8c), and from 4.6 to 4.7 in the TCGA dataset (Figure 2.9c). The low-expression miRNAs are similar, moving from 4 in the entire set of partial correlations to 4.2 in the Bentink dataset (Figure 2.8d) and from 4.6 to 4.5 in the TCGA dataset (Figure 2.9d). All moment ratios for these datasets are summarized in Table 2.1. Again, further restricting the subset of miRNAs examined to those of more extreme expression level (the top or bottom 5%, for instance) may reveal more consistent patterns. This, as well,

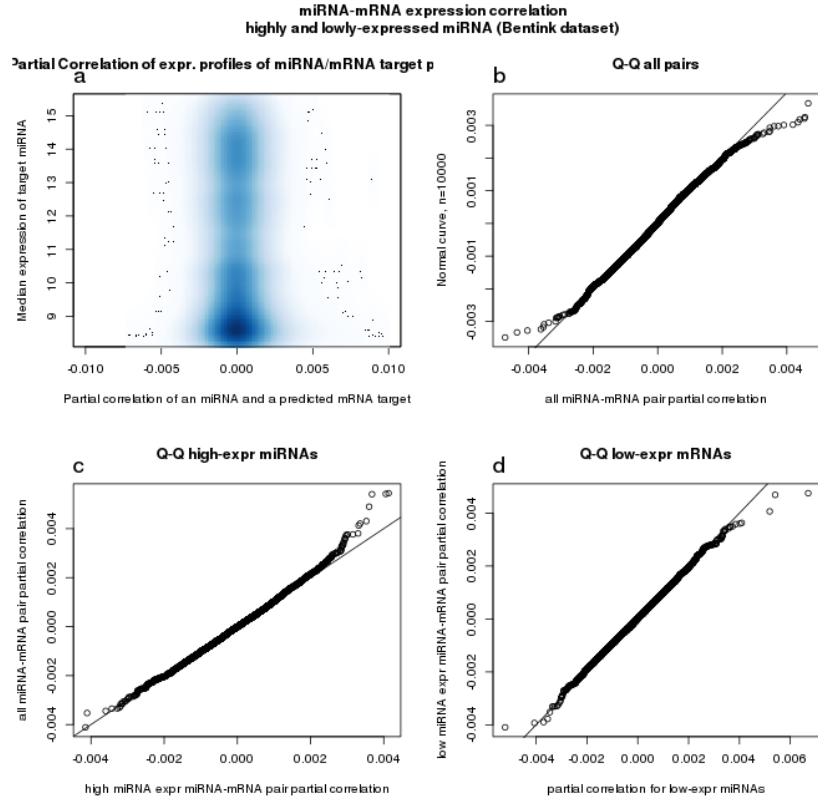


Figure 2.8: Partial correlations broken down by miRNA expression values for the Bentink dataset. a) Median miRNA expression values vs. correlation of miRNAs with their predicted targets. Each point represents a single mRNA-miRNA predicted pair. The majority of the points are shown as a cloud (light blue). b) Correlation of all possible miRNA-mRNA pairs at least 10,000kb apart on the chromosome, compared to a normal distribution of the same mean and standard deviation. c) Comparison of the distribution of correlation coefficients for the high-expression (mean miRNA expression > 10) non-located miRNA-mRNA pairs in the TCGA dataset. d) Comparison of the distribution of correlation coefficients for the low-expression (mean miRNA expression < 10) non-located miRNA-mRNA pairs in the TCGA dataset.

is left for future research.

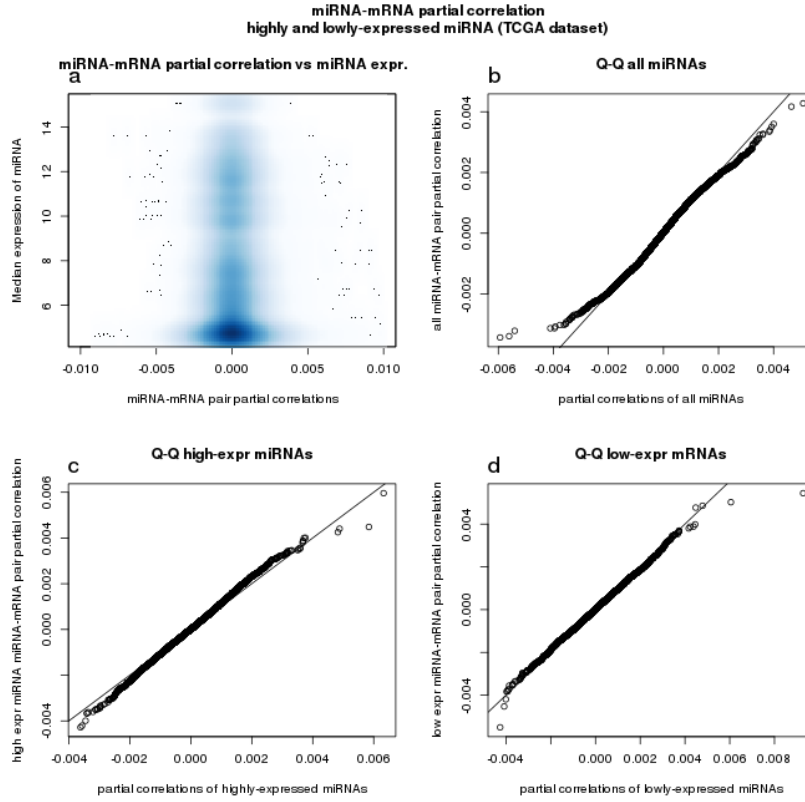


Figure 2.9: Correlations broken down by miRNA expression values for the TCGA dataset. a) Median miRNA expression values vs. correlation of miRNAs with their predicted targets. Each point represents a single mRNA-miRNA predicted pair. The majority of the points are shown as a cloud (light blue). b) Correlation of all possible miRNA-mRNA pairs at least 10,000kb apart on the chromosome, compared to a normal distribution of the same mean and standard deviation. c) Comparison of the distribution of correlation coefficients for the high-expression (mean miRNA expression > 6) non-located miRNA-mRNA pairs in the TCGA dataset. d) Comparison of the distribution of correlation coefficients for the low-expression (mean miRNA expression < 6) non-located miRNA-mRNA pairs in the TCGA dataset.

2.2.7 Similar patterns of correlation in two molecular subtypes

The two ovarian molecular subtypes described by Bentink *et al.* [15] and summarized in Chapter 1 have distinct survival profiles and molecular phenotypes. The smaller of the two groups, consisting of about one-third of the total population of samples (33 samples from the Bentink dataset and 181 from the TCGA dataset) are characterized by an up-regulation of angiogenesis and extracellular matrix-associated genes, and a poorer prognosis. The non-angiogenic subtype comprises the remaining two-thirds of the samples (96 and 321, respectively) and has a relative upregulation of DNA repair and transcription-related genes. These two subtypes are discussed further in Chapter 3, however we briefly examine their correlation structure here.

If these subtypes have genuinely different molecular mechanisms of disease or tumor progression, those mechanisms are likely to include differences in miRNA activity or expression, as well. miRNA repression of target genes leading to anticorrelation of expression profiles in only one subtype may be lost when the samples of that subtype are pooled with others that lack this relationship. Therefore, we will examine the two subtypes separately.

We divided each dataset into two groups, one consisting of the samples from the angiogenic subtype and the other containing the non-angiogenic subtype. We calculated partial correlation coefficients for the miRNA and mRNA profiles within each subtype and examined their distribution. Figures 2.10a and 2.10c show the partial correlation across the samples of the angiogenic subtype in both the Bentink and TCGA datasets. In the angiogenic subtype for both datasets we

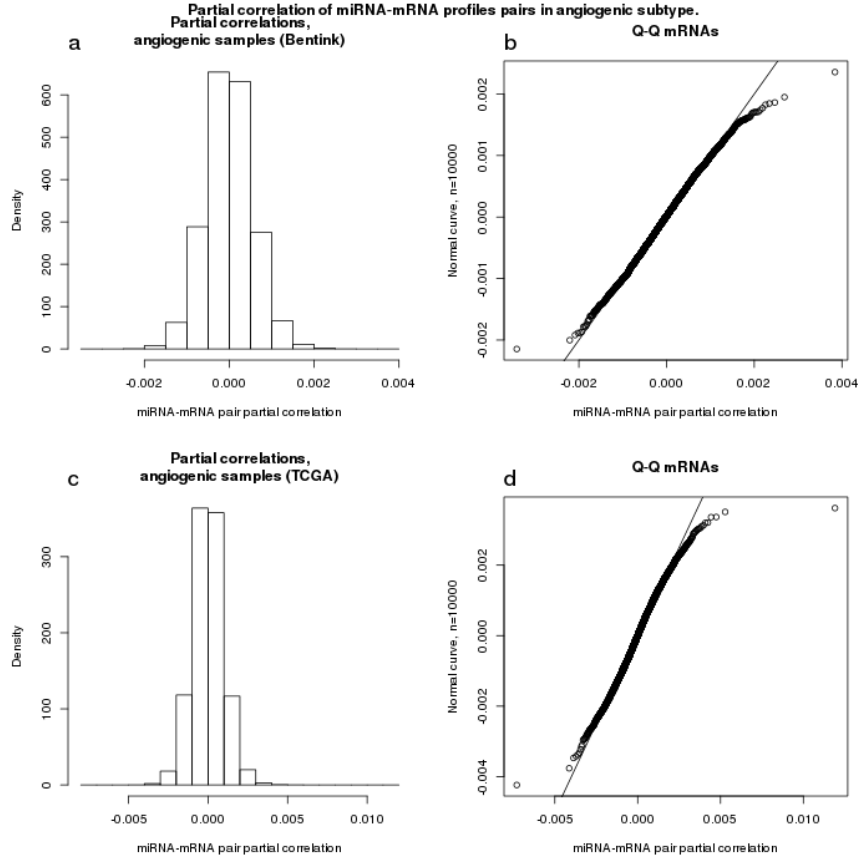


Figure 2.10: Partial correlation distributions for miRNA-mRNA expression profile pairs in the angiogenic subtype of the TCGA and Bentink Datasets. a) Distribution of partial correlations for all possible miRNA-mRNA pairs in the angiogenic subtype, Bentink dataset, located at least 10,000kb apart. b) Q-Q comparison of the same distribution against a normal distribution of the same mean and standard deviation. c) Distribution of partial correlations for all possible miRNA-mRNA pairs in the angiogenic subtype, TCGA dataset, located at least 10,000kb apart. d) Q-Q comparison of the same distribution against a normal distribution of the same mean and standard deviation.

observe an excess of large-positive and large-negative partial correlations relative to a normal distribution. The moment ratios of these distributions support this conclusion, each being larger than the expected value by a significant margin (Table 2.1, $p < 0.05$).

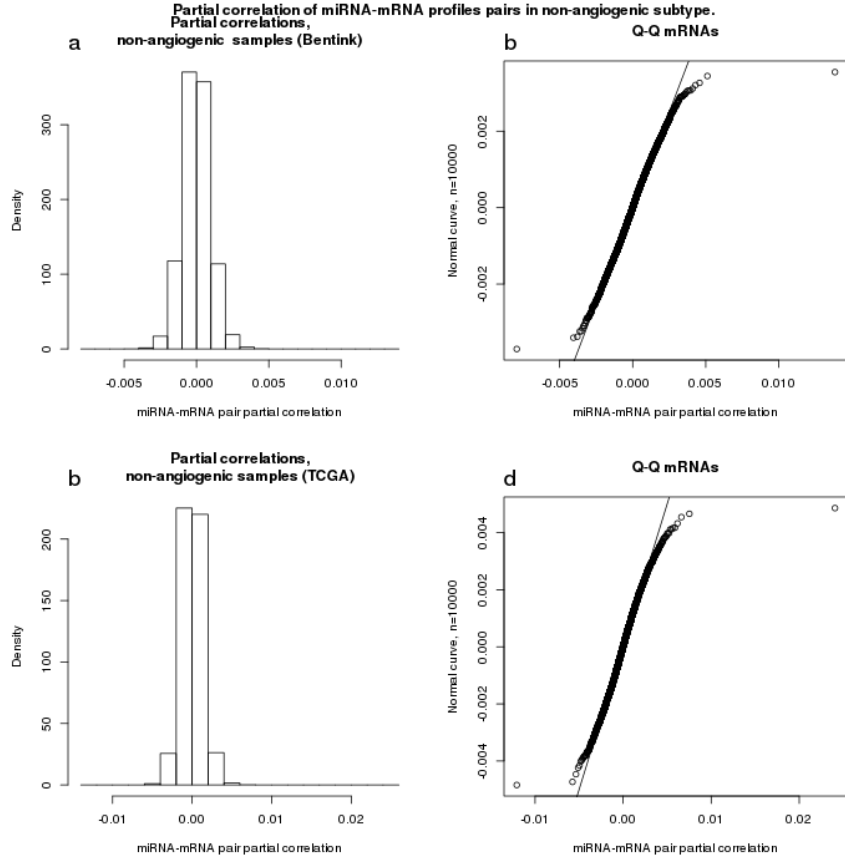


Figure 2.11: Partial correlation distributions for mRNA-miRNA expression profile pairs in the non-angiogenic subtype of the Bentink and TCGA Datasets. a) Distribution of partial correlations for all possible miRNA-mRNA pairs in the non-angiogenic subtype, Bentink dataset, located at least 10,000kb apart. b) Q-Q comparison of the same distribution against a normal distribution of the same mean and standard deviation. c) Distribution of partial correlations for all possible miRNA-mRNA pairs in the non-angiogenic subtype, TCGA dataset, located at least 10,000kb apart. d) Q-Q comparison of the same distribution against a normal distribution of the same mean and standard deviation.

Figures 2.10b and 2.10d show the correlation distribution compared to a normal distribution of the same mean and standard deviation. The correlation distribution and Q-Q plot comparison to a normal distribution for the non-angiogenic subtype are shown in Figures 2.11a and b and 2.11c and d.

The fourth moment ratio of the angiogenic subtype is 3.4 in the Bentink dataset and 3.9 in the TCGA dataset. Again, these and the higher moment ratios are significantly higher than expected by chance ($p < 0.05$). The non-angiogenic dataset is similar, with larger fourth, sixth and eighth moment ratios than expected for a normal distribution (Table 2.1, $p < 0.05$).

2.2.8 Testing differentially expressed miRNAs for correlation enrichment.

One of the important characteristics of these two reported molecular subtypes is the specific set of miRNAs that are differentially expressed between them. We presumed that these differentially-expressed miRNAs were possible candidates for functional contributors to the differences in molecular phenotype. If they are, indeed, regulatory controllers of some aspect of the molecular difference between the two subtypes they may have a different correlation profile between the two subtypes. Therefore, we have added markers for the differentially expressed miRNAs.

We recalculated the differentially expressed miRNAs using Significance Analysis of Microarrays (SAM) using an FDR of 5%. This resulted in a list of 57 miRNAs. We compared the distribution of median correlation coefficients of the differentially expressed miRNAs to the distribution of partial correlation coefficients of each of the two molecular subtypes (Figures 2.12 and 2.13). The vertical lines mark the correlation coefficients of the differentially expressed miRNAs with their predicted targets. They do not appear to be normally distributed around $r = 0$, and are statistically offset relative to the background of all coefficients in

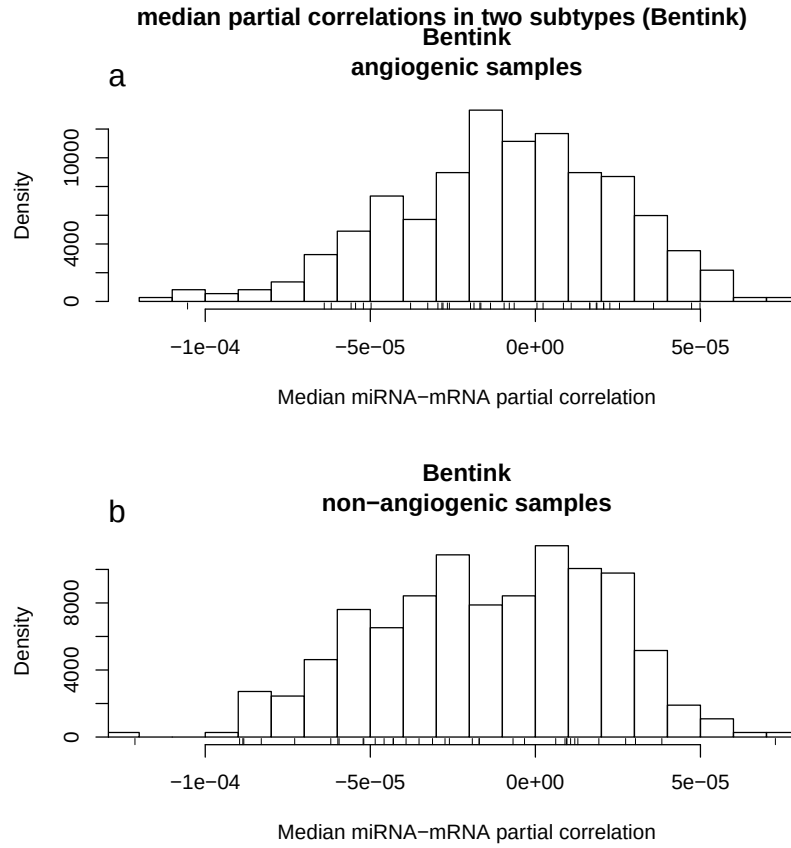


Figure 2.12: Distribution of median partial correlation coefficients for each subtype, marked with the median partial correlation of differentially expressed miRNAs between the two subtypes. a) Median partial correlation of miRNAs in the Bentink dataset, angiogenic subtype. b) Median partial correlation of miRNAs in the Bentink dataset, non-angiogenic subtype.

the angiogenic subtype (Kendall rank test, Bentink: $p = 0.003$, $\tau = 0.12$, TCGA: $p = 0.0033$, $\tau = -0.19$).

When comparing the distribution in the non-angiogenic subtype, however, we find no significant association between the ranked list of differentially expressed miRNAs and the ranking of miRNAs anticorrelated with genes (Kendall rank test, Bentink: $p = 0.19$, $\tau = -0.053$, TCGA: $p = 0.53$, $\tau = 0.04$), indicating

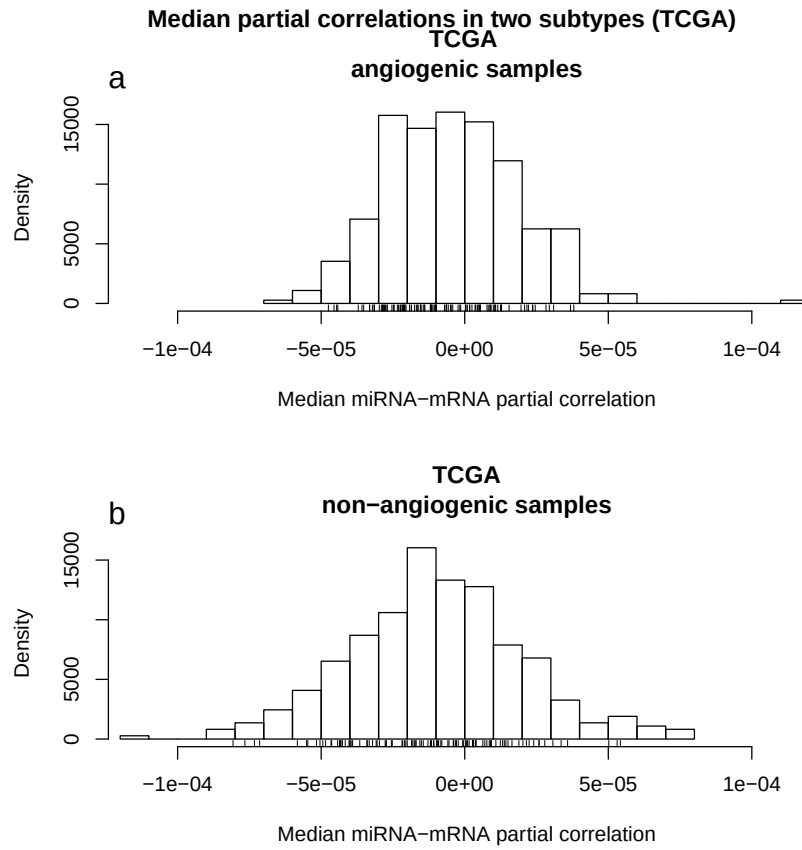


Figure 2.13: Distribution of median partial correlation coefficients for each subtype, marked with the median partial correlation of differentially expressed miRNAs between the two subtypes. a) Median partial correlation of miRNAs in the TCGA dataset, angiogenic subtype. b) Median partial correlation of miRNAs in the TCGA dataset, non-angiogenic subtype.

that the expression profiles of the classifier miRNAs are not more significantly correlated with mRNAs than the non-classifier miRNAs are.

2.3 Conclusions and future work

Here we have examined the relationship between mRNA and miRNA expression profiles in two high-grade serous ovarian cancer datasets. We had hoped to find a pattern of miRNA-mRNA pair expression that suggested some regulation of gene expression by miRNAs observable at the transcriptional level. We find that some relationships do exist and are detectable in these microarray datasets. We were also reassured that the datasets had similarity in their distribution of miRNA-mRNA correlation: we found similarities in the targets of the anti-correlated miRNA-mRNA pairs, for instance. If these pairs were dramatically different we would be concerned that their results are not compatible and that findings in one dataset would not be generalizable to the other, and certainly not to the larger population of ovarian cancer patients.

We have observed several patterns in the relationship of miRNA and mRNA expression profiles: first, that a larger number of large positive correlations exist than would be expected by chance in both datasets. A larger number of negative correlations than would be expected in a normal distribution are present in both datasets, as well. Some of these large positive correlations are accounted for by mRNA-miRNA pairs encoded on the same chromosome, and therefore likely to be either translated together, in the case of intronic miRNAs, or affected by the same chromosomal lesions. Once the collocated miRNA-mRNA pairs are removed, the remainder of the correlation distribution retains a larger number of large positive correlations than is expected in a normal distribution.

We have found that co-location with a range of 10,000kb is not the only reason for this large positive tail. Expanding that range may reveal that the

large tail is a result of larger location-specific effects other than co-translation — large deletions, for example.

We have found some consistency in the biological characteristics of the two datasets. The top positively correlated mRNA-miRNA pairs in the two datasets are similar; a few are identical and the functional enrichments of the genes is comparable. Many of the same miRNAs and mRNAs appear in the top-ranked positive and negative correlation pairs.

We hypothesized that miRNAs with higher expression may show a stronger regulatory effect on targets than lowly expressed miRNAs. Upon examination of expression correlation profiles for highly-expressed and lowly-expressed miRNAs, however, we found no such difference in correlation. We also hypothesized that lowly-expressed mRNAs may be more affected by miRNA regulation than highly-expressed mRNAs. We found little evidence for this, as well. Either of these observations may not have been possible, again, because of the large number of measurements at hand compared to the likely number of true miRNA-mRNA suppression relationships present in the data. It seems reasonable that further analysis of perhaps the top five or ten percent of the most highly-expressed or low-expressed miRNAs may reveal differences. It is also worth remembering that some miRNA regulatory activity will occur at the translational level, and therefore will not be detectable in our transcript-only data.

We have verified in these data the observation of Blower *et al.* [20]; predicted mRNA-miRNA pairs do not have a different correlation distribution than random pairs, or in our case, all possible pairs. This observation may simply be due to the poor performance of miRNA target prediction technology. Another contributing factor may be that, whatever the strength of the predicted interaction between

miRNA and mRNA, if one of the molecules is not produced in the tissue sample being measured, correlation between them would be near zero. We could look for further relationships in predicted-target pairs by restricting the set of predicted targets examined. For instance, the PITA prediction algorithm provides a score for each mRNA-miRNA pair; choosing a score cutoff that results in a very small (≈ 50) number of predicted pairs would ensure that only extremely strongly predicted interactions would be included in the analysis. This may result in a greater connection between expression and profiles. Other prediction tools may also prove to select highly anti-correlated pairs better. A possible approach to addressing the limitations of the currently available prediction tools would be to limit our analysis to only verified miRNA-gene interactions. However, the currently validated set of these interactions is very small, at 1094 verified pairs, as of this writing [137].

The findings of Blower and colleagues are further verified by our finding that a “long tail” of highly positively correlated mRNA-miRNA pairs includes miRNAs and mRNAs encoded in close proximity on the chromosome. This proximity could lead to related expression patterns for a variety of reasons. For instance, chromosomal lesions could affect the loci together, leading to a common pattern of repression. Similarly, duplications, hyper- or hypo-methylation or histone activation are location-specific events that would affect closely-located genes in similar ways. While we cannot determine from these data what type of location-specific effect is at work, we have shown that there is such an effect.

These analyses have met the goals set out at the beginning of this chapter; we find consistency between the datasets and interesting, biologically relevant information in the relationships between their miRNA and mRNA expression

profiles. We have established that meaningful consistency exists between the datasets and that they are therefore suitable for further analysis.

We are particularly encouraged by the consistency of correlation distributions in the angiogenic subtype in these two separate datasets. A strong pattern of positive and negative partial correlation exists in both subtypes, and is consistent between datasets. We find, however, that only in the angiogenic subtype is the ranking of differentially expressed miRNAs significantly associated with the partial correlation rank of those miRNAs with mRNA partners. As discussed further in Chapters 3 and 4, it seems that this subtype is a homogeneous and distinct molecular phenotype, whereas the non-angiogenic subtype represents, instead, a heterogeneous collection of “all the rest” of the subtypes.

We find some inconsistencies in the data as well; no miRNA-mRNA pairs are consistent across the top 30 pairs of the two datasets. The two datasets were collected with different array platforms; their probe types differ, as well as the miRNA populations the arrays were designed to measure.

In Chapter 3, we continue to investigate the relationship between miRNA expression and gene expression and how that relationship differs between the two subtypes of serous ovarian cancer.

Chapter 3

Characterization of miRNAs in two ovarian subtypes

3.1 Introduction

Several groups have attempted to segregate serous ovarian cancer into molecular subtypes. Tothill *et al.* [153] first used gene expression measurements to describe subtypes of serous ovarian cancer. Their C1 subtype, identified with k-means clustering of gene expression data, is characterized by poor survival and an up-regulation of stromal and immune gene signatures relative to the other subtypes. The TCGA consortium later reported an additional sample classification of ovarian cancers using gene and microRNA expression data [149]. Non-negative matrix factorization of the gene expression data resulted in four subtypes, labeled by the consortium as immunoreactive, differentiated, proliferative and mesenchymal.

The work of Bentink *et al.*, covered in Chapter 1, built upon these prior

subtype classifications to divide high grade serous carcinoma into two molecular phenotypes - the angiogenic and the non-angiogenic subtypes. The angiogenic subtype is characterized by an up-regulation of genes related to angiogenesis, cell adhesion and the extracellular matrix, and a relative down-regulation of DNA binding-related genes [15]. The subtype is largely concordant with the C1 or "reactive stroma" subtype described by Tothill et al. [153] and with the "mesenchymal" and "immunoreactive" subtypes described by the TCGA consortium [149]. The 100-gene classifier used to identify these subtypes was validated in ten additional ovarian cancer datasets and is significantly associated with overall survival.

Bentink *et al.* used miRNA expression data to further validate their angiogenic and non-angiogenic subtypes, creating a miRNA signature that classified the subtypes with low misclassification rates (0.091% misclassification rate for the angiogenic subtype, 0.17% for non-angiogenic) [15]. This classifier consists of six miRNAs, hsa-miR-202, hsa-miR-202*, hsa-miR-506, hsa-miR-508, hsa-miR-509, and hsa-miR-514. Because miRNAs seem to be involved in all phases of cancer development and progression, we expect to find that the molecules play a causal role in the differential development, spread, or metastasis of these two molecular subtypes. Here we further investigate the role of miRNAs in the molecular biology of these two subtypes by investigating the patterns of expression and activity and the relationship of those measurements to gene expression profiles.

Direct measurement of miRNA expression is not the only means we use to study the effects of miRNA action on the cell; in tumors, it is sometimes the reduction in mature, processed miRNA molecules. The disruption of miRNA action in the cell seems to be due to miRNA processing systems (proteins such as Dicer1,

DGCR8 and Drosha) [97]. Therefore, direct measurement of miRNA expression may not be all that useful in determining whether miRNA action is broken. Since approximately 80% of protein down-regulation caused by miRNAs happens as a result of transcript destabilization rather than translational repression [67], the disruption of miRNA maturation should be observable by measurement of target transcript levels. This is why the RE score method proposed by Cheng *et al.* is so useful.

Cheng *et al.* [37] reported a global reduction in miRNA activity in ER positive breast cancer tumors when compared to ER negative tumors. They came to this conclusion by inferring miRNA activity using measurements of mRNA expression. Their Regulatory Effect (RE) score compares the averaged ranked expression levels of predicted miRNA targets, and is described more fully in Chapter 1.

Our investigation of the involvement of miRNAs in these two subtypes was again centered on the datasets described previously; two datasets that included sample-matched miRNA and mRNA expression measurements (TCGA and Bentink), and the Tothill dataset, which includes only gene expression data. The PITA, Pictar, Targetscan and mirSVR miRNA target prediction tools, described in Chapter 1, were used to calculate RE scores for miRNAs on the mRNA expression data of all three datasets.

We begin our analysis by comparing differences in abundance and regulatory activity of miRNAs in these two subtypes. We hope to characterize what, if any, differences in miRNA expression and action exist between the two groups and to investigate what these differences may tell us about the molecular mechanisms giving rise to the two subtypes. We hope that this information will lead to a

better understanding of the molecular underpinnings of ovarian cancer pathology. We approach the subtypes as if the angiogenic subtype was the true subtype of interest - the subtype is in concordance with other described subtypes in the literature, whereas the non-angiogenic subtype seems to be more heterogeneous, consisting of the "remainder" of all the other samples.

We find that the angiogenic subtype has lower overall miRNA activity than the non-angiogenic subtype, and a larger number of miRNAs are downregulated in the angiogenic subtype relative to the non-angiogenic subtype. The expression of many miRNAs is downregulated in the angiogenic subtype, and suggests a complete loss of several miRNAs. This cluster of severely down-regulated miRNAs, the 506-514 cluster, shows strong anticorrelation with a set of genes enriched for extracellular matrix and angiogenesis functional annotation, and transcription factor binding sites for several important cancer-related transcription factors. Further, the relationship of expression of this miRNA cluster with its anticorrelated miRNAs suggests a thresholding control pattern common to miRNAs. We finish the chapter by proposing a mechanism by which the down-regulation of this miRNA cluster results in the downregulation of miRNA activity across the angiogenic subtype and the progression of the EMT, leading to the described phenotype of the angiogenic subtype.

3.2 Results

3.2.1 Differentially expressed miRNAs.

We first examined the two subtypes by testing the miRNAs for differential expression. We used Significance Analysis of Microarrays [154], as described in Chapter 2, to identify a complete list of 19 differentially expressed miRNAs (Tables B.1 through B.4), using a false discovery rate of $FDR = 0.05$.

No overall difference in global miRNA variance was found between the two sample groups (F-test, $p = 0.91$). miR-22 was upregulated in the angiogenic subtype in both datasets (Tables B.1 and B.2). This miRNA is a known tumor suppressor acting on a regulator of the oncogene c-MYC [169] and has been shown to suppress proliferation in hepatocellular carcinoma [176]. We do not further investigate this miRNA in this thesis, but its status as a consistently differentially expressed miRNA and its known links to cancer make it a good target for future analysis.

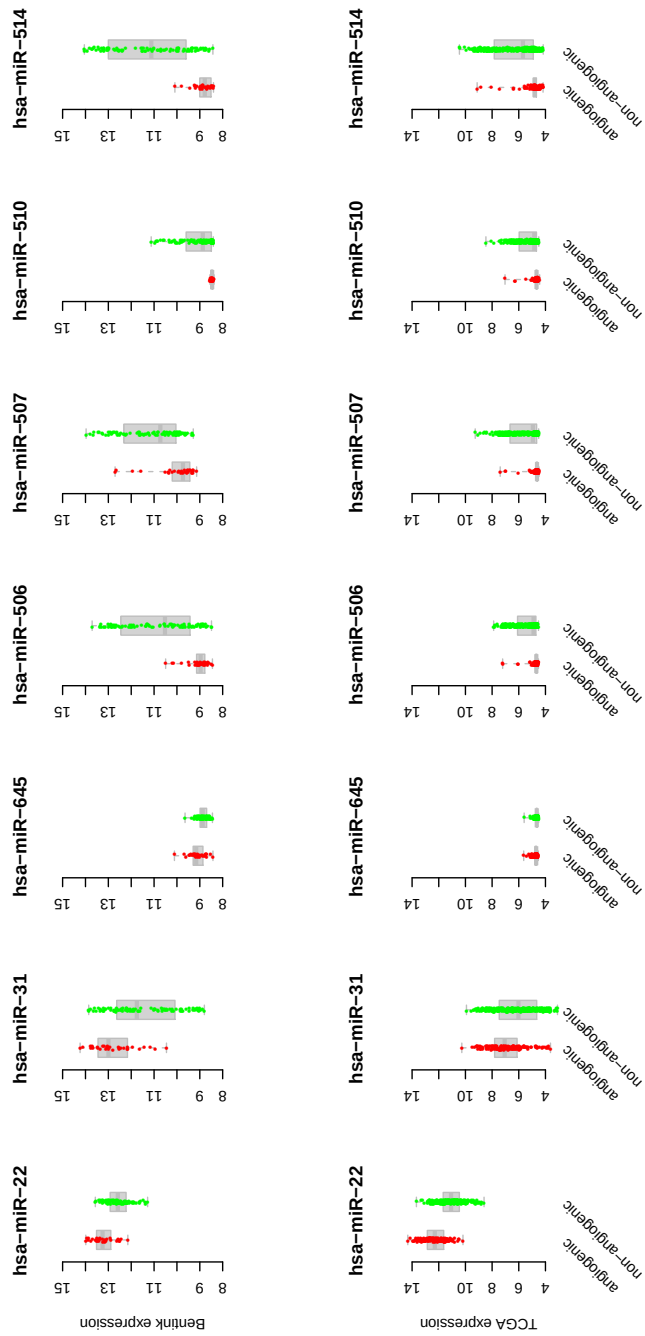


Figure 3.1: Expression distribution of each of representative up- and down-regulated miRNAs, in the angiogenic and non-angiogenic samples.

All six of the classifier miRNAs reported by Bentink *et al.* (miR-202, miR-202*, miR-506, miR-507, miR-510 and miR-514) have lower expression in the angiogenic sample group of the Bentink dataset. Additionally, they have significantly lower variance in the angiogenic group (F-test, $p < 0.05$, Figure 3.1, Table B.3). A similar set of miRNAs is downregulated in the TCGA dataset (Tables B.3 and B.4). The downregulated miRNAs includes members of the 506-514 cluster. The 506-514 cluster is encoded within separate transcripts in the $\approx 100\text{kb}$ region of band Xq27.3 [16]). More members of the 506 family, miR-510, miR-513, and miR-507, are found among the differentially expressed miRNAs in Table B.3. The nearest known protein-encoding gene to this location is the Fragile X gene (FMR1). They do not, however, include the 202 and 202* miRNAs. We note that other members of the hsa-miR-506 family of miRNAs, miR-513, miR-510 and miR-507, are differentially expressed between the subtypes and also have a significantly lower variance, though they do not appear in the classifier reported by Bentink *et al.*. The low microarray expression measurements and their low variance across the samples of the angiogenic group point to a loss of expression of the miR-202 and miR-506 miRNA family members.

This cluster of miRNAs is interesting because of its consistent pattern of deregulation between the subtypes (Figure 3.1), and its location on the X-chromosome, a common point of lesions in gynecologic cancers [57, 73]. We will return to these miRNAs later in the chapter. First, however we continue to examine global miRNA activity.

3.2.2 Lower miRNA activity in the angiogenic sample group

The regulatory power of miRNAs depends not only on the amount of miRNA present in the cell, but also on the functions of other molecules, such as the miRNA processing machinery and the RISC components. A disruption in these molecules can mean that miRNAs, however abundant, cannot function to repress target messages. Measurements of miRNA target expression can be useful to gauge a miRNA's regulatory effect at the transcript level. Furthermore, while miRNA-based regulation is sometimes detectable only at the protein, or translational level, many miRNAs exert their effects on their targets by transcript destabilization or deadenylation [67, 79].

The Regulatory Effect scoring method of Cheng *et al.*, described in Chapter 1, is one method of using target mRNA expression measurements to indirectly characterize miRNA activity [37]. The activity of miRNAs is inferred based on the expression values of predicted-target mRNAs, in comparison to non-targets. Briefly, the RE score represents the difference in average ranks between target and non-target genes within a given sample. RE scores are calculated for each miRNA in each sample, and can be used to compare the relative activity of miRNAs between samples. The scores are not informative in an absolute sense, but rather as a comparative measure of miRNA regulation across samples.

We used target prediction data from four target prediction algorithms to calculate the scores: PITA, Pictar, Targetscan and mirSVR [91, 96, 108, 17]. To identify miRNAs with differential activity between the two groups, we then compared the scores between the angiogenic and non-angiogenic sample groups,

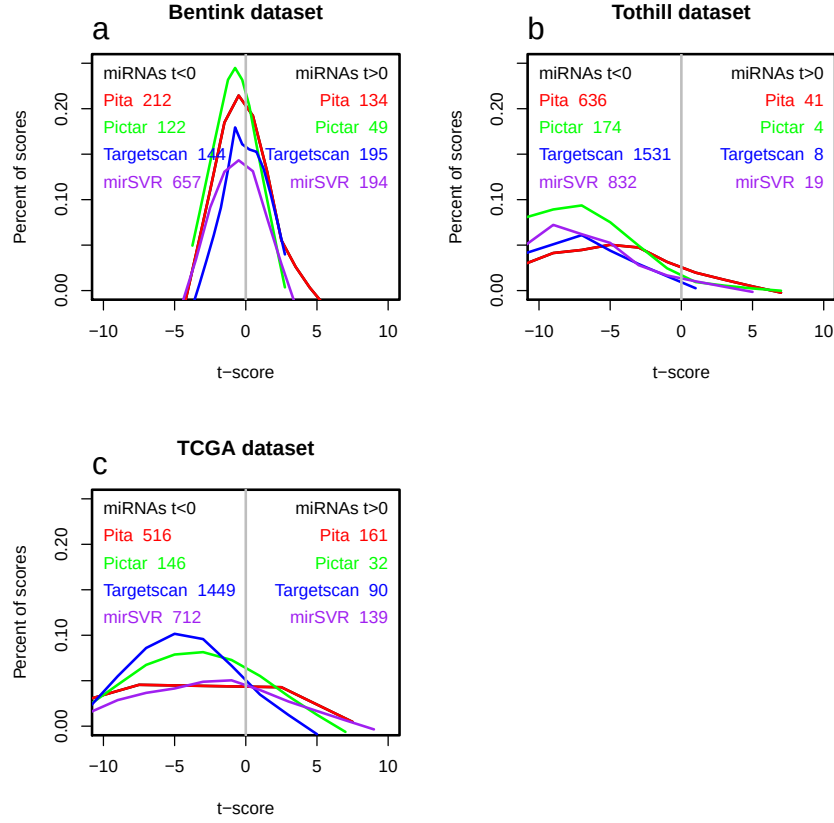


Figure 3.2: Distribution of t-statistics resulting from test of RE scores for each miRNA between the angiogenic and non-angiogenic subtypes. A t-score of $t < 0$ indicates that a miRNA has a lower RE score in the angiogenic subtype, relative to the non-angiogenic subtype, whereas a score of $t > 0$ indicates that the miRNA has a higher activity in the angiogenic subtype. Each colored line represents t-statistics calculated for RE scores from a different set of target prediction data: red: PITA; green: Pictar; blue: Targetscan; purple: mirSVR. a) Bentink dataset. b) Tothill dataset. c) TCGA dataset.

using the latter as the reference. We calculated a t-score as a measurement of the difference in RE scores between groups (Figure 3.2). A negative t-score indicates that the miRNA has a lower RE score in the angiogenic group, compared to the non-angiogenic group, whereas a positive t-score indicates a higher RE score in the angiogenic group.

We calculated an empirical p -value for each miRNA by permuting the sample class membership and repeating the t-test comparing RE scores between the randomized sample groups, and corrected these p -values with the method of Benjamini and Hochberg [14]. We retained only miRNAs with a corrected p -value of $p < 0.05$. Adopting the terminology of Cheng *et al.*, we call miRNAs whose RE scores were higher in the angiogenic group *positively RE-changing*, and those whose scores were lower in the angiogenic group *negatively RE-changing*.

We expanded our analysis by recalculating RE scores using three additional miRNA target prediction tools: Targetscan [107], Pictar [101] and mirSVR [17]. We expected to find differing results, as the predictions made by these tools do not greatly overlap. Indeed, little agreement exists; no individual miRNAs were found to be RE-changing when scores were calculated using all four prediction sets using miRNA expression from the Bentink (Tables B.5 and B.7) and TCGA datasets (Tables B.6 and B.8). As expected, the miRNAs individually detected as RE-changing were not consistent across prediction tools; no miRNA is significantly RE-changing across all tools, for either dataset.

	Bentink Pos	Neg	TCGA Pos	Neg	Tothill Pos	Neg
PITA	1	4	62	389	19	602
Pictar	0	15	9	107	2	169
Targetscan	0	0	14	1255	1	1481
mirSVR	0	10	72	614	6	804

Table 3.1: Counts of positive and negative RE-changing miRNAs for the three datasets, q -value of $q < 0.05$.

We observe many fewer RE-changing miRNAs in the Bentink dataset, compared to the TCGA and Tothill datasets. Two characteristics of the datasets contribute to this fact; one, the Bentink dataset contains fewer samples than

either of the other two datasets (129 vs. 502 and 285, respectively). This lower number of samples means that less statistical power is available for detecting changes in predicted target expression across the groups of samples. Secondly, the TCGA and Tothill datasets were collected using Affymetrix whole-genome arrays, each mapping to approximately 12,000 genes, compared to the 9,000 genes measured by the Illumina arrays used in the Bentink experiments. This larger number of gene measurements makes it more likely that an miRNA's effect on gene expression is observed in the two Affymetrix-based datasets.

We would expect to see miRNAs with a strong difference in expression to have differing RE scores, assuming they did have some repressive functions in the cell. We tested the differentially expressed miRNAs separately for RE-changing status (reported in the column *REtest* in Tables B.1 - B.4) and found several to be significantly RE-changing across the subtypes. Few of the miRNAs with greater expression in the angiogenic subtype had significantly differing RE scores; only miR-22 had a RE score lower in the angiogenic subtype than in the non-angiogenic subtype, and only in one of the two datasets.

In contrast, several of the miRNAs with lower expression in the angiogenic subtype had in addition a significantly lower RE score. miR-514 and 506 had a lower RE score in the angiogenic subtype of the Bentink dataset, and miR-507, 508 and 509 had lower RE scores in the angiogenic subtype of the TCGA dataset. These lower activity scores suggest that members of the 506-514 cluster do, in fact, repress mRNA expression in these samples, and that their repression in the angiogenic subtype does have some differential effect on gene expression.

In a more global view, one consistent theme among the RE score data is that for all four target prediction tools and all three datasets, more miRNAs are

determined to be negatively RE-changing than positively RE-changing (Table 3.1). Furthermore, the number of miRNAs with negative t-scores outnumbers the count of miRNAs with positive t-scores, in all three datasets and using all four target prediction tools. These findings are a strong indication that the miRNAs were, on average, less active in the angiogenic subtype. We do not expect that this effect is caused by an overall lower expression of the genes in the non-angiogenic group, as the data were quantile normalized before this analysis.

Cheng *et al.* observed a similar skew in the distribution of the RE score *t*-statistic; in their work, ER+ breast cancer samples had a lower RE score than the ER- samples overall, regardless of the prediction tool used. They found that the miRNA processing genes DICER and AGO2 were downregulated in the ER+ samples, and attributed the global lower miRNA activity to the lowered action of these genes.

We supposed that miRNA-processing genes may be similarly downregulated in the angiogenic subtype, and looked for evidence of this downregulation in the differential expression results, above. Genes annotated with miRNA-processing GO terms (listed in Table B.11) were checked for differential expression using SAM, in the same manner used for differential expression analysis of the miRNAs ($FDR = 0.05$). Several of these genes (ZCCHC11) were found to be downregulated in the angiogenic subtype, including DGCR8, a part of the miRNA-processing machinery (Table B.14). These six genes were only found to be differentially expressed in the TCGA data set. It is possible that differential expression is more difficult to detect in the Bentink dataset, due to its much smaller sample size. This downregulation of miRNA-processing machinery is one possible explanation for the repression of miRNA activity in the angiogenic

subtype.

The complete miRNA processing and RISC-based translational repression machinery has not yet been completely characterized, so other components of the pathway may also be affected. Lowered miRNA activity is intriguing, as low miRNA activity due to impaired miRNA processing has been shown to cause increased transformation in cancer [97]. This impairment is a potential cause for the poorer prognosis of the angiogenic subtype.

3.2.3 Switch-like activity describes the relationship between 506-514 cluster miRNAs and their anticorrelated genes.

When first examining the repression of the 506-514 cluster of miRNAs, we hypothesized that the cluster was repressed by some location specific means, such as chromosomal lesions or methylation. Our hypothesis was validated when Bag-noli *et al.* [8] reported that the cluster is indeed downregulated by methylation in ovarian tumors, that the repression of these miRNAs led to an increased invasiveness of ovarian cancer cell lines, and that the loss of this cluster was predictive of shortened time to relapse in ovarian cancer patients.

The 506 family of miRNAs are encoded as multiple separate transcripts in ChXq27.3, and are expressed in normal testis and conserved as a whole across primates, and in part in dog, mouse, and rat [16]. A few cancer-related copy number variations are present in this region, but none include known disease associations [80]. The family has undergone rapid recent evolution in the primate family [109, 69, 177].

One member of the family, miR-514, was found to be associated with ovarian cancer resistance to chemotherapy [21]. This miRNA is the only family member normally expressed in ovary, and the only one of the classifier miRNAs having the property of a higher RE score in the angiogenic subtype, relative to the non-angiogenic subtype. In addition, miR-514 down-regulation is important for metastatic colonization in melanoma cells [118].

All of the members of the 506-514 cluster are downregulated in the angiogenic subtype, relative to the non-angiogenic subtype, though only four (506, 508, 509 and 514) were included in the subtype classifier developed by Bentink *et al.* To learn more about the possible effects of the 506-514 cluster of miRNAs in these samples we revisited the partial correlations calculated in Chapter 2. mRNAs with expression profiles highly anti-correlated with the expression profiles of the 506-514 cluster miRNAs were tested for significant difference from a background partial correlations calculated using sample label permutation (one thousand permutations). We calculated a background distribution of correlations by permuting sample labels, recalculating partial correlations one thousand times, then calculating an empirical p -value using these permutations. These empirical p -values were corrected using the Benjamini-Hochberg method. Between 7 and 128 mRNAs mRNAs from the Bentink dataset and between 454 and 1170 from the TCGA dataset were significant at this cutoff. We presume that the large difference in the number of large negative partial correlation mRNAs in the Bentink dataset is, again, due to the smaller size of this dataset.

We examined the relationship between miRNAs and the anticorrelated mRNAs¹.

¹The initial examination was done by Stefan Bentink. He created an earlier version of the plots shown in Figure 3.3, in which only the averaged expression value of the top twenty most anti-correlated mRNAs were compared to mRNA expression values.

It was not simply the correlation that was important, but the shape of that correlation. We observed a striking relationship between the expression of classifier miRNAs and the expression of mRNAs that are closely anti-correlated with those miRNAs. Figure 3.3 illustrates this relationship for four members of the 506-514 cluster, plus one other miRNA unrelated to this study, miR-100, for contrast. In each graph the expression of one miRNA (y-axis) is plotted against the average expression of the mRNAs significantly anti-correlated with that miRNA (x-axis). Each point represents one sample, color-coded according to subtype.

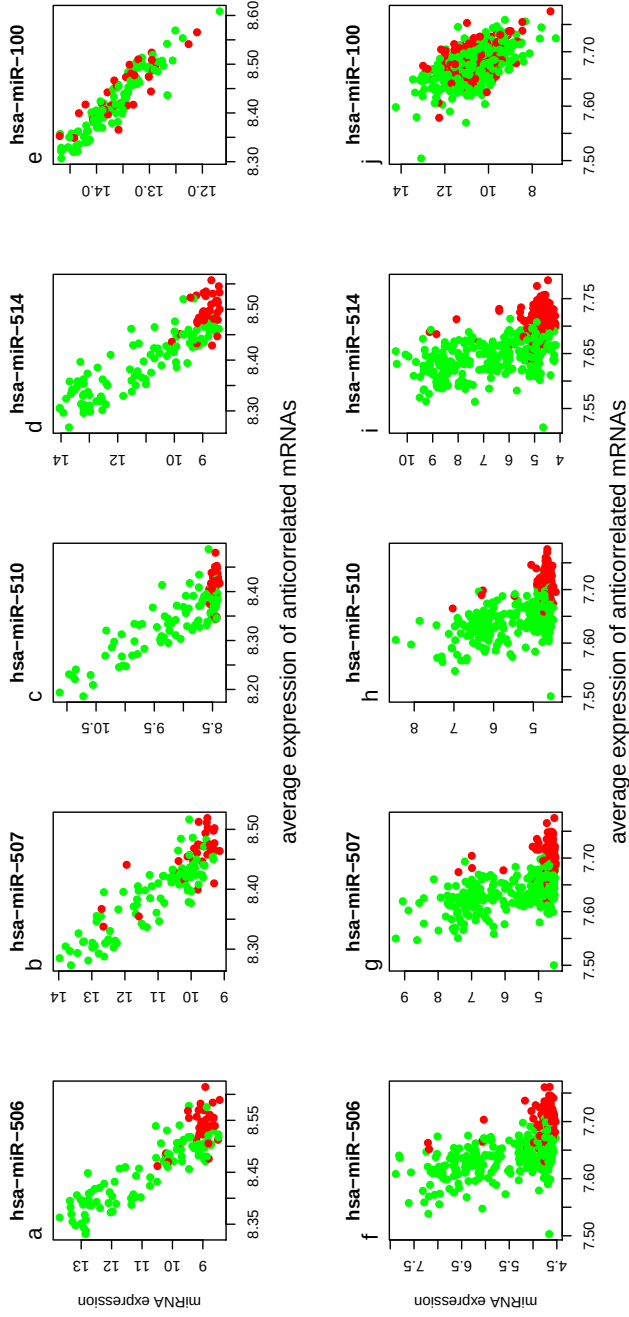


Figure 3.3: Relationship between miRNA expression and the expression of significantly anti-correlated mRNAs for eight miRNAs in the Bentink (top) and TCGA (bottom) datasets. A typical non-classifier (hsa-miR-100) is shown for contrast. Each point represents one sample. The y-axis represents the measured miRNA expression level. The x-axis represents the average expression level of the mRNAs significantly anti-correlated with the miRNA. Each point represents one sample, colored red in the case of the angiogenic subtype samples, and green in the case of the non-angiogenic subtype samples.

The relationship between the 506-514 cluster of miRNAs and the mRNAs most anti-correlated with them follow a striking pattern. When a miRNA is relatively highly-expressed, the mRNAs that are anticorrelated with that miRNA are expressed at relatively low levels. However, when the miRNAs expression falls below a certain threshold, the expression level of anticorrelated mRNAs varies according to whatever conditions prevail in the cell. Those samples with high expression of these anticorrelated mRNAs belong to the angiogenic subtype.

Low expression of the 506 cluster of miRNAs is not sufficient to allow mRNA expression to vary along its full range, but it does seem to be a requirement; only samples with low miRNA expression have high mRNA expression. Those patients in which the anticorrelated genes are expressed at high levels seem to fall within the angiogenic subtype. The presence of the classifier miRNAs may lock down the expression of some genes, either directly or indirectly. However, when the miRNAs are not present, those mRNAs are not regulated and are free to be expressed according to other cellular mechanisms.

One potential weakness of this analysis is that the switch-like effect observed in Figure 3.3 may be only a coincidental correlation. However, simple anticorrelation appearing by chance is much more likely to appear as an oblong cloud in such a graph, like that observed in Figure 3.3e and j, not a hockey-stick shape as we observe in Figures 3.3a-d and f-i. If the data were, indeed, distributed randomly along both x and y- axis, (miRNA and mRNA expression, respectively) and anti-correlated only by chance, we would expect a linear relationship between the mRNA and miRNA expression values scattering along the main diagonal of the plot with a negative slope, as is observed in 3.3e and j. Here we observe instead a discrete change in miRNA expression between the two groups, related

to a similar discrete change in mRNA expression.

This observation translates well into a model for miRNA regulation of these mRNAs; when miRNA expression is sufficiently high, mRNA expression is repressed. However, when miRNA expression is low, mRNA expression is free to vary in response to the prevailing conditions in the cell. This observation is clear when raw correlation coefficients are used to select the most anticorrelated mRNAs as well (data not shown). Raw correlations would be a far less computationally demanding calculation to use, and would perhaps be the better choice for this method.

The relationship also suggests a few possible directions in which this work could be taken, both based on using the regulatory control of the 506-514 cluster explored in Chapter 3. The first possibility is one of supervised learning — discovering miRNAs with strong activity differences in two classes of samples. Given two known groups of samples, a simple visual examination of graphs similar to those shown in Figure 3.3 for all differentially expressed miRNAs would be enough to allow discovery of miRNAs with switch-like relationships between those miRNAs and anticorrelated genes. A simple tool such as this could serve as a hypothesis generator for any set of paired miRNA-mRNA expression data.

A more complex, unsupervised approach would simultaneously allow the discovery of miRNAs with significant potential activity differences and the two groups across which those differences occurred. This method would first need to identify miRNAs with the characteristic distribution, perhaps based on the skew in the distribution of miRNA expression. The samples must then be divided into two groups: one in which the miRNAs are apparently strongly repressing mRNA expression (directly or not) and one in which the miRNAs are not doing

so. The expression level of miRNA is the discriminant used to distinguish these two groups. This expression level would be set by identifying an inflection point in the distribution of mRNA expression. One of the challenges of building this method would be determining how to find that inflection point. Perhaps the skewed distributions could be modeled as two distinct normal distributions, and the point of their crossover used as the mRNA expression threshold.

The identification of miRNA switches like this may be of interest for two reasons. First, we would like to learn which miRNAs have a regulatory effect on genes within the cell. Note that this regulatory activity need not be direct or through classical miRNA repression mechanisms; indirect methods of regulation may be identified in this way. Second, this method would allow classification of samples into groups based on the action and non-action of one or more miRNAs, rather than mRNAs. A miRNA-based classifier may uncover new subtypes in ovarian cancer, and may be generalizable to other diseases. Therefore, we next investigated these anti-correlated mRNAs themselves.

3.2.4 Genes anti-correlated with the 506-514 miRNA cluster are enriched for ECM genes and some transcription factors.

The significantly anti-correlated mRNAs were tested for enrichment of functional annotations using the DAVID algorithm [47]. The Bentink anti-correlated mRNAs map to 2508 gene symbols and the TCGA to CDH2 (Tables B.9 and B.10). These genes overlap considerably; 52% of the Bentink genes appear in the TCGA list as well. The anti-correlated mRNAs in both datasets are enriched for

extracellular matrix and cell adhesion terms such as *extracellular region*, *extracellular space*, *cell adhesion*, *glycoprotein*, and *cell membrane* (Tables B.1 and B.2). Also included were cell motility terms such as *localization of cell*, *cell motility* and *positive regulation of cell motion*. These terms agree well with those associated with the 100 classifier genes used by Bentink *et al.* to distinguish the angiogenic subtype from the non-angiogenic subtype.

As X-linked genes are often important to ovarian cancer and other gynecologic cancers, we sought more connections to the X chromosome. We found no representation of genes on the X chromosome among the anticorrelated mRNAs, nor enrichment of X-chromosome located genes in the differentially expressed mRNAs, the predicted PITA targets of the selected miRNAs, or the Pita-predicted RE-changing miRNAs.

3.2.5 Within the 506-514 Cluster, miR-506 and miR-507 appear to be active.

We tested the Pita, Targetscan and mirSVR-predicted targets of each member of the 506-514 cluster for greater anti-correlation than non-predicted targets (the Pictar tool does not provide predictions for the 506-514 cluster). Of the cluster members, only miR-506 and miR-507 showed significant differences in anti-correlation between target and non-target genes, for some of the target prediction tools: miR-506 targets were significantly more anti-correlated in the Bentink dataset (t-test, PITA $p = 5.0e - 5$, Targetscan $p = 0.02$) but not in the TCGA dataset. No target predictions for miR-506 were available in the mirSVR dataset.

miR-507 targets were significantly more anti-correlated in both datasets (Bentink: Targetscan, $p = 0.04$, mirSVR $p = 0.03$, TCGA: Targetscan, $p = 0.011$). None of the other miRNAs showed significant enrichment for predicted targets when tested with any of the prediction tools. This suggests that it is these miRNAs, rather than the cluster as a whole, that have a functional effect on gene expression in these data. The other miRNAs may then show their switch-like effect simply because of their strong correlation with the other miRNAs in the cluster.

3.2.6 Anti-correlated mRNAs are enriched for a common set of binding sites for cancer-related transcription factors.

We cannot simply explain the anticorrelation of gene expression and miRNA expression with direct miRNA-based repression alone; approximately half of the anticorrelated genes are not predicted targets of the 506-514 cluster. We investigated the possibility that transcription factors may be responsible for the indirect regulation of the remainder of the anticorrelated genes. A transcription factor repressed by the 506-514 family would be translated in greater amount in samples with low 506-514 expression, leading to the up-regulation of its targets. Transcription factor binding sites are detectable by computational and experimental means and have been cataloged extensively in databases like JASPAR and TRANSFAC.

The anti-correlated mRNAs were tested for enrichment of transcription factor binding sites (Pscan [175], JASPAR database). Significant enrichment of binding

sites in the anti-correlated mRNAs in the Bentink (9) and TCGA (19) overlapped by 23 transcription factors (Tables B.12 and B.13).

Several of these transcription factors are predicted to be targeted by members of the 506-514 cluster of miRNAs, including NFKB1, INSM1, PLAG1, SP1 and RREB1. The enriched transcription factors and the 506-514 cluster miRNAs predicted to target them are summarized in Figure 3.4.

We examined the lists of differentially expressed genes returned by the SAM algorithm for these enriched transcription factors, supposing that those factors that are targeted by a member of the 506-514 cluster may be derepressed. Differential expression was not detected in the Bentink dataset, while NFKB1 is upregulated in the angiogenic subtype of the TCGA dataset and MZF1 and TBP are downregulated (Table B.15). The fold-change of these differentially expressed genes is low - transcription factors are transcribed at low levels under normal conditions, and therefore changes in their expression is difficult to detect.

If NFKB1 is, in fact, directly regulated by a member of the 506-514 miRNA cluster, the effect of that cluster's differential expression could be propagated across an entire set of genes transcriptionally enhanced by the protein. That is, when a miRNA fails to repress the translation of a powerful transcription factor, that transcription factor may be free to enhance transcription of a wide range of other genes. NFKB1 is known to be activated in many carcinomas (the role of NF-kB in cancer is reviewed in [49] and [124]).

Predicted targets of 506–514 miRNAs

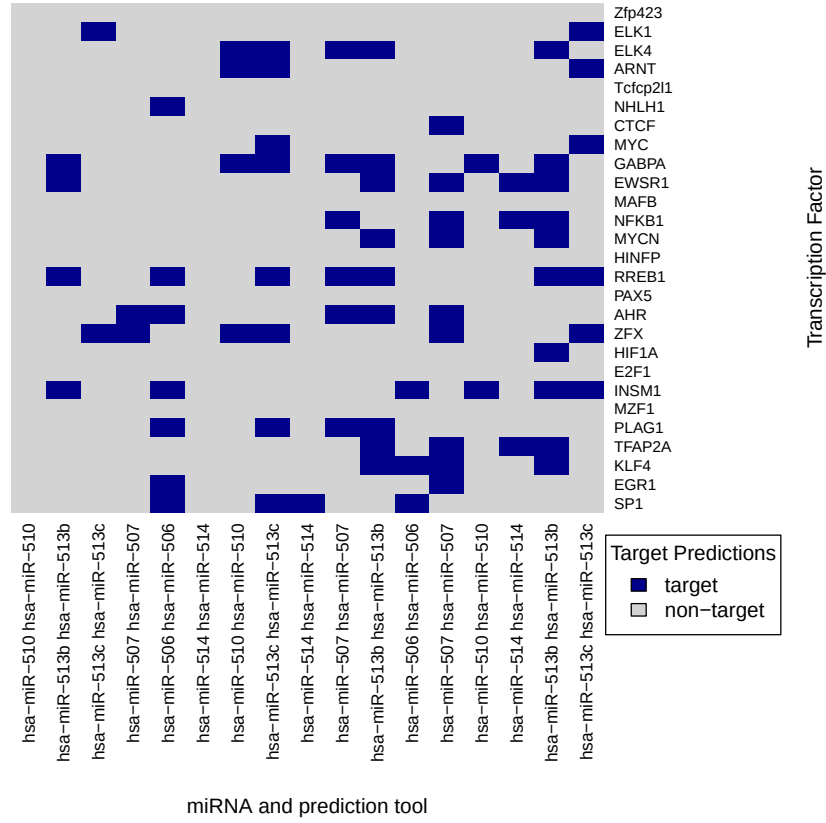


Figure 3.4: Significantly enriched transcription factors from the 506-514 anti-correlated genes in the TCGA and Bentink datasets. A blue box indicates that the prediction tool predicts a binding site for the miRNA in the transcription factor. Gray boxes indicates no such prediction.

3.3 Conclusions

Our study indicates substantial differences in miRNA expression and activity between two recently discovered subtypes of high-grade serous ovarian cancer. As measured by RE score, miRNAs are generally less active in the poor-prognosis angiogenic subtype. Most miRNAs are less active in that subtype and more miRNAs are negatively RE-changing than positively-RE changing. miRNA ex-

pression is affected as well; more miRNAs are underexpressed in the angiogenic subtype than the reverse. Therefore it seems that there exist broad-based differences in the miRNA disposition in the two subtypes. Previous studies have attributed different miRNA activity patterns to different levels of the DICER1 protein [37]; we also observe lowered expression of miRNA processing genes. Deregulation of miRNA processing machinery is known to increase cellular transformation [97, 98]. We propose that one of the characteristics of the angiogenic subtype is decreased global miRNA activity, possibly due to disruption of the miRNA processing machinery.

In addition to this global reduction of miRNA activity, we find a consistent set of miRNAs to be differentially expressed between the angiogenic and non-angiogenic subtypes, including a set of miRNAs involved in early relapse in ovarian cancer: the 506-514 cluster. These miRNAs, whose methylation-induced repression is predictive of early relapse in ovarian cancer patients, are strongly repressed in the angiogenic subtype.

The relationship between expression of this miRNA cluster and the mRNAs significantly anticorrelated with them suggests a causal relationship between the two; miRNAs in the cluster seem to repress, directly or indirectly, a set of genes that are upregulated in the angiogenic subtype. When the miRNAs are expressed at a low level, the expression level of anti-correlated mRNAs can vary greatly, subject to other regulators in the cell. However, when the miRNA is highly-expressed, the mRNAs that are affected are expressed at very low levels.

The genes anti-correlated with the 506-514 cluster members are functionally associated with extracellular signaling. These genes are also enriched for a set of transcription factor binding sites for known oncogenes such as insulinoma-

associated protein 1 (INSM1), a zinc-finger DNA binding protein involved in cell-cycle arrest in neuroendocrine cancers (reviewed in [102]). Another of the enriched binding sites is targeted by NF- κ B, a transcription factor often constitutively activated in cancers and implicated in early relapse and drug resistance in ovarian cancer [106, 114, 180]. The miR-506-514 cluster is also predicted to target several of these enriched transcription factors, including NF- κ B, suggesting that the miRNAs influence gene expression indirectly, by inhibiting transcription factors that would normally activate gene expression. Therefore the molecular differences between the angiogenic and non-angiogenic subtypes may be partially mediated by the interaction between certain miRNAs and the transcription factors reported here.

The 506-514 cluster is involved in carcinogenesis and metastasis in other tissues, including renal cancer [179] and melanoma [118, 144]). This switch-like relationship between the 506-514 miRNAs and anti-correlated genes suggests that the 506-514 cluster down-regulation is not only a marker of the subtype, but may be a cause of the molecular differences between the two subtypes. Bagnoli *et al.* [8] have shown the down-regulation of the 506-514 cluster of miRNAs to be a predictor of early relapse in ovarian cancer in the same TCGA dataset examined here. miR-506, specifically, when knocked down, leads to increased proliferation, seemingly due to loss of the G1 cell-cycle checkpoint. Bagnoli and colleagues also report that this cluster of miRNAs are down-regulated with a methylation-based mechanism in ovarian tumors. While we were unable to test the Bentink dataset for methylation at the 506-514 cluster, (no methylation data were available at this locus) we expect that downregulation of the cluster proceeds by the same mechanism. These characteristics can be used to propose a

mechanism driving some of the molecular differences between the angiogenic and non-angiogenic subtypes.

The extracellular matrix signature of the anti-correlated genes suggests a potential route by which the downregulation of the cluster could result in this phenomenon. This suggests that the 506-514 cluster may drive the difference between the two subtypes by affecting the tumor's interaction with the stromal environment. The enrichment of extracellular matrix associated annotations in the list of genes anticorrelated with the 506 cluster supports this.

This mechanism would begin with an unknown event causing excess methylation of the 506-514 miRNA cluster at chX27.3. This methylation represses transcription of the miRNA cluster, reducing the amount of processed miRNA in the cell. The absence of miR-506 and -507 miRNAs, as the apparently functional miRNAs in these dataset, allows de-repression of a set of both anti-correlated and predicted-target genes, including transcription factors such as NFkB1. These transcription factors are transcribed or translated at a higher level, allowing them to enhance the transcription of their targets more. The combination of miRNA-based derepression and transcriptional enhancement causes an entire set of genes to be upregulated, including a set of extracellular matrix genes. These genes are not all turned on; other conditions in the cell cause some of the samples to retain low expression of the anticorrelated genes. These patients remain in the non-angiogenic subtype. However, patients with low 506 expression are at risk of developing a tumor of the angiogenic subtype.

3.4 Future Work

This explanation for differences in the subtypes brings further insight into the molecular mechanisms of tumor progression in ovarian cancer. These are important insights into the angiogenic subtype, and further work will be needed to better elucidate their full biological context as well as how they may be used as markers or targets for designing novel and tailored therapies for the disease.

We still lack a comprehensive model that describes the characteristics of the angiogenic phenotype. Specifically, we lack a link between the general downregulation of miRNA activity and the suppression of the 506-514 miRNA cluster. Further investigation of the genes targeted by the cluster may reveal a regulator of miRNA activity and explain these two characteristics.

One experiment that would help validate our findings is suggested by the cell line experiments of Bagnoli, et al. They forced expression of miR-506 finding that the expression of the miRNA decreased cancer cell line proliferation. If NF-KB is the key intermediate in the miRNA's effects on the cell, then manipulating the target regions in its 3'UTR should mitigate the effect of the miRNA. Gene expression profiling of this experiment would also be informative as to which genes are directly affected by miR-506 and which indirectly by NF-KB. This experiment could be expanded to the other potential transcription factor targets identified here. Hopefully this type of experiment would provide a direct observation of the switch-like effect illustrated in Figure 3.3.

That switch-like effect also suggests a pattern: a relationship between mRNA and miRNA that may be found in other settings. It may be possible to identify other meaningful relationships or sample classes by looking for the characteristic

expression pattern of anti-correlated expression. Quantifying this observation of switch-like behavior could lead to a method of identifying miRNAs with real regulatory effects on target genes, or possibly for classifying samples into subtypes based on the miRNA-mRNA expression relationship.

The discovery and understanding of the molecular differences in the cancer subtypes has led to effective differential treatment of breast cancer. We hope that this work has helped contribute to a similar understanding of subtypes in ovarian cancer. A better understanding of these subtypes can inform the search for drugs that effectively target the growth of these tumors. For instance, if NFkB is proven to be an exacerbating effect on ovarian cancer, we can target that protein with drugs already in development in other diseases.

We have established here a distinct difference in the function of miRNAs between these two molecular subtypes. Establishing that miRNAs are expressed in a different pattern, regardless of whether they do in fact cause the differences in the subtypes, means that they might also allow us to make meaningful predictions about the samples. Since Bentink *et al.* have already established that a miRNA signature can be used to classify the samples into two groups with differential survival, we next attempt to use the miRNA expression data to inform predictions of patient survival time.

Chapter 4

Modeling of risk based on miRNA expression and activity

4.1 Introduction

In Chapter 3 we established the involvement of miRNAs in the differences in the molecular phenotype of two subtypes of ovarian cancer. As these two phenotypes include a difference in overall survival, we hope to find that miRNAs are directly predictive of survival. It is often helpful, in the course of treating cancer patients, to have some knowledge of a patient's projected survival time. If a patient is known by some means to have a low likelihood of recovery, that patient may decide to forgo the costly and difficult chemotherapy treatments. Each disease requires a different set of decisions by doctor and patient; better information about a patient's likely outlook is extremely important.

We wish to build predictive models with time to event data: data in which each patient, or sample, is associated with either a death or censoring event,

and a time associated with that event. In the case of cancer survival, this time measurement refers to the time from initial diagnosis to death (uncensored) or last reported visit to a care provider (censored).

Clinical information has long been used to make this type of prediction. Tumor stage and grade, as well as patient age and other health indicators have been used to tell physicians which patients will respond to treatment and which patients may not find it worth the stress and pain of difficult cancer therapies. Classification of patients using these survival models is useful for similar reasons. However, gene expression data is powerful and has been used to classify patients into survival classes in other cancers, including breast cancer.

miRNA expression has been shown to be predictive of survival in ovarian cancer [95]. The ratio of miR-221/miR-222 expression was shown to predict survival in individual ovarian cancer patients [164]. In addition, miR-30b has been shown to be associated with survival in ovarian cancer [165]. The TCGA group showed that miRNA expression values can segregate patients into three groups differing in survival time [149]. Similarly, the classifier miRNAs described by Bentink *et al.* [15] were used to distinguish the angiogenic and non-angiogenic subtypes, with differing survival profiles.

Shih *et al.* [139] report that a signature of miRNAs is predictive of survival time in ovarian cancer. They used the Cox proportional hazards [42] model to identify a group of miRNAs whose expression values were predictive of overall survival in ovarian cancer patients and verified these microarray-based transcriptional measurements with RT-PCR. They found that two miRNAs, miR-410 and miR-645, were negatively associated with survival. Their method involved splitting the expression values into tertiles (high, middle, low expression) and using

those discretized values to compare patient survival using a log-rank test and Cox regression.

Bovelstad, *et al.* [25] did a comparative analysis of gene expression based modeling with several methods of model-building, including stepwise selection, ridge regression and the Lasso. However, as of this writing no such comparison has been done using miRNA expression data. We hope that miRNA expression data will prove as useful as gene expression data, or more so, given the role of miRNAs as master regulators in the cell and the power that miRNA expression data has in distinguishing tumor cell of origin [129, 59].

This thesis attempts to build models of survival using miRNA expression as predictors. We first establish that univariate Cox models fitted on individual miRNA expression profiles are not informative for survival. We then attempt to combine miRNA expression profiles into multivariate models using the forward stepwise selection method. Finding these models insufficient, we then use a more sophisticated model, the Lasso variable selector, to build a complete survival model. We attempt to reduce the dimensionality of the data by averaging miRNA expression profiles by miRNA family and build a refined model from this data. We also attempt to filter the input data by including only miRNAs known to be involved in gynecologic and prostate cancers. We find that miRNA activity score profiles are not useful for predicting survival, and that while direct measurements of miRNA expression are probably informative about survival time, the mainly hypothesis-free approaches used here are not sufficient to build useful models. Good predictive models most likely await larger miRNA expression datasets.

4.1.1 Cox Proportional Hazards Model with multidimensional data

The Proportional Hazards Model is the most widely used method of survival analysis. It is a semi-parametric linear regression method often used for creating predictive models of survival from a set of predictors, such as clinical covariates, demographic data, or gene expression measurements. The model, described by Cox in [42], assumes an unknown underlying hazard function dependent on time, and makes no assumptions about the shape of that function.

The risk of an event (in the case of these data, death or disease progression) at time t is given by

$$h(t|x) = h_0(t)e^{\mathbf{x}\beta},$$

where $h_0(t)$ is the baseline hazard function, the shape of which is unknown, β is a vector of regression parameters and $\mathbf{x}$ is a matrix of gene or miRNA expression values, having p measurements values over n observations .

The probability that a given case j within the risk set R_i will fail at time t_i is given by

$$Pr(t_j = T_i | R(t_i)) = \frac{e^{\mathbf{x}_i\beta}}{\sum_{j \in R(t_i)} e^{\mathbf{x}_j\beta}}.$$

Taking the product of these conditional probabilities yields the partial likelihood function,

$$L_p = \prod_{i=1}^K \left[\frac{e^{\mathbf{x}_i\beta}}{\sum_{j \in R(t_i)} e^{\mathbf{x}_j\beta}} \right]^{\delta_i}.$$

where $\delta_i = 0$ if the case is censored and $\delta_i = 1$ if not. The regression parameters are solved for, in the case of $n < p$, by maximizing the log-likelihood, given by

$$l_p = \log(L_p) = \sum_{i=1}^K \delta_i \mathbf{x}_i \beta - \log \sum_{j \in R(t_i)} e^{\mathbf{x}_j \beta}.$$

An estimate of β , denoted $\hat{\beta}$, can be found by maximizing the log likelihood as

$$\hat{\beta} = \arg \max[l_p(\beta)].$$

The assumption of proportionality is critical to the Cox proportional hazards model, stating that the predictors contribute to the hazard in a time-independent way. Predictors can be tested for proportionality by testing for time-dependence of the Schoenfeld residual [134]. This residual is defined as the observed covariate value for the individual that failed minus its expected value conditioned on the risk set R_i or

$$x_{ij} - \hat{E}(\mathbf{x}_{ij} | R_i).$$

The expected value can be calculated with

$$\hat{E}(X_{ij} | R_i) = \sum_{k \in R_i} \mathbf{x}_{kj} e^{\mathbf{x}_k \hat{\beta}} / \sum_{k \in R_i} e^{\mathbf{x}_k \hat{\beta}}.$$

If these values are relatively constant for all values of time t , the proportionality assumption holds.

This semi-parametric method is useful when the number of observations equals or exceeds the number of measurements. In the case of genomic-scale data, however, the linear model cannot be solved due to the high dimensionality of the data. The standard microarrays used to measure gene expression data contain probes allowing measurement of hundreds or thousands of molecules per sample. In contrast, collecting biological samples is an expensive process; reagents for array experiments are costly, running into the hundreds of dollars per sample. Obtaining appropriate consent and legal approval to use human tissue is similarly expensive, as well as time-consuming and heavily restricted by Institutional Review Boards. Therefore collecting human samples in batches of greater than a few hundred is prohibitively expensive. These two factors result in data sets with a much greater number of predictors (mRNA or miRNA expression measurements) compared to observations (tumor samples). This condition, where $n \ll p$, is referred to as the "curse of dimensionality". This is one challenge we must address when building these models.

4.1.2 The data

For these analyses we used two datasets, the Bentink and TCGA datasets described in previous chapters [15, 149]. The Bentink dataset includes measurements of 12449 mRNAs and 474 miRNAs on 129 tumor samples. The TCGA dataset is larger, consisting of 22277 mRNAs and 723 miRNA measurements on 502 samples.

Each of these samples is annotated with two censored-survival markers. The first is overall survival (OS), the measure of time between initial cancer diagnosis and death. The second is time to progression (TTP), measuring the time from initial diagnosis to disease recurrence. For most of this work we focus on the overall survival measure.

For each of these datasets we calculated Regulatory Effect, or RE scores, as previously described in Chapter 1. These scores are an indirect measure of the effect of a miRNA upon mRNA expression in the cell. Developed by Cheng *et al.* [37] and described in Chapter 1, this score is a means of inferring miRNA activity in a sample based on a comparison of the ranks of predicted target mRNAs vs. non-targets. Here we have RE scores calculated for all miRNAs included in the PITA target prediction database (677 miRNAs), as described in Chapter 1.

4.1.3 Testing the proportionality assumption

We tested the assumption of proportionality by examining the Schoenfeld residual of each miRNA for each sample [134]. This residual is defined as the covariate value for the individual that failed minus its expected value. If these values are relatively constant for all values of time t , the proportionality assumption holds. We test whether a line of regression drawn through these points has a slope of zero, using a linear regression t-test. We evaluated each miRNA individually, using a univariate Cox proportional hazard model. Plots of a few example miRNAs are shown in Figure 4.1. We conclude that the proportionality assumption holds for these data and proceed with our analysis.

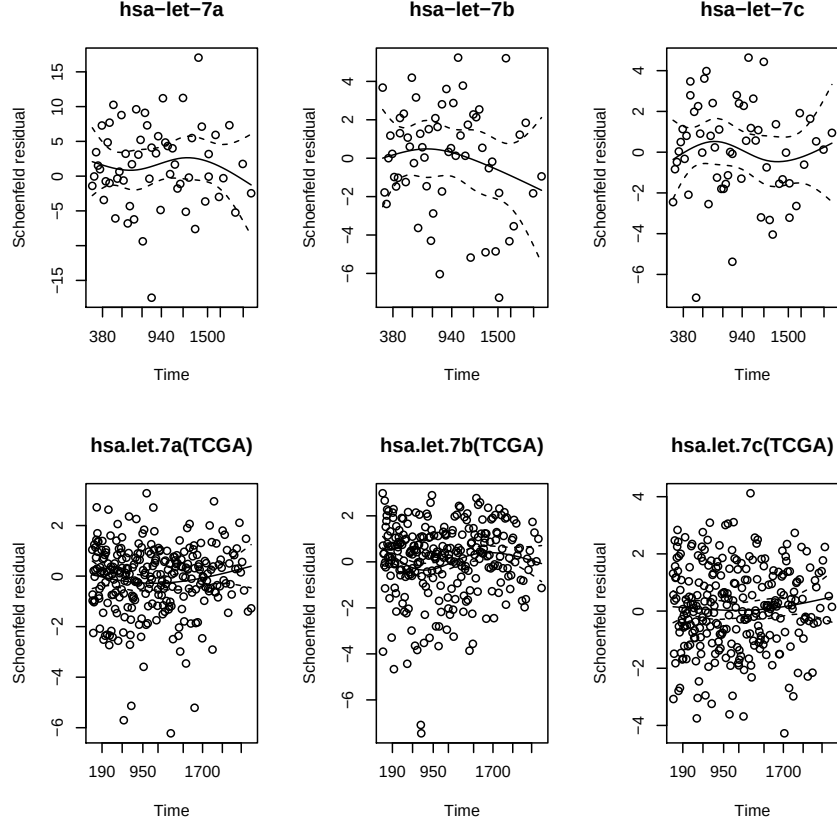


Figure 4.1: Scaled Schoenfeld residuals of each sample against time, for several example miRNAs in each dataset. The solid line is a smoothing-spline fit to the residuals. The dotted lines are ± 2 -standard-error band around that fit.

For the majority of miRNAs (Bentink: 96%, TCGA: 97%), the slope of a line of regression drawn through the residuals did not significantly deviate from zero (linear regression t-test, $p < 0.05$). As fewer than 5% of the miRNAs had a p -value lower than this, we assume that they are significant by chance, and so do not discard them from our analysis.

4.2 Univariate analysis of miRNA expression and activity

We examined the predictive power of miRNA expression values by building a series of univariate Cox proportional hazard models of survival using expression profiles from the Bentink dataset ($n = 474$). Similarly, 677 models were created using the RE scores calculated in Chapter 3, using sample-matched gene expression data from the Bentink data set.

The R package *coxph* [3] was used to create models of risk predicted by miRNA expression values or RE scores. The predictive power of each model was evaluated with a likelihood ratio test, in which the log likelihood of the null model, including only clinical covariate data and labeled L_0 , is subtracted from the log likelihood of the candidate model, containing clinical covariate data and miRNA expression data and labeled L , resulting in the statistic s . A p -value is calculated with a chi-squared test on s with one degree of freedom. This p -value represents the likelihood that the model including the miRNA predictor is predictive of survival over and above the clinical data by chance alone.

$$s = 2 * (L - L_0).$$

We calculated concordance indices (CIs) for each model as well, to report the fidelity with which the model predicted response. An average concordance index of $CI > 75\%$ would indicate a very strong predictor, whereas $CI > 65\%$ would indicate a weaker but still predictive model. The closer the concordance index

came to $CI = 0.50\%$, an indication that the model predicted outcome with a similar accuracy to random assignment, the poorer the model.

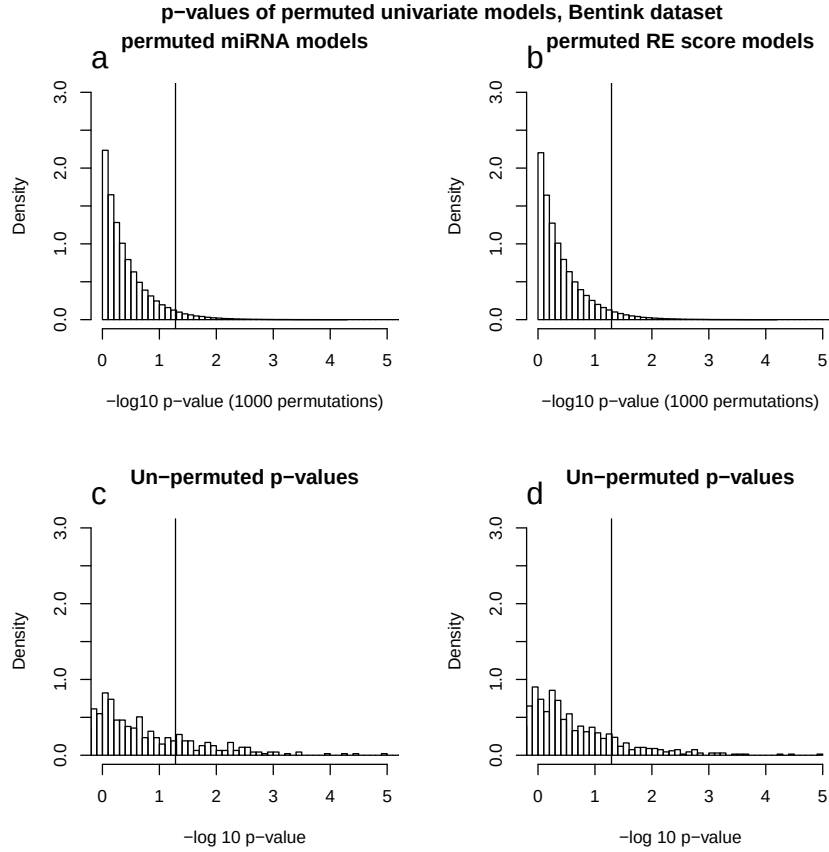


Figure 4.2: a-b) $-\log_{10}$ p-values of permuted univariate models built on miRNA expression and RE scores, respectively. The 5th percentile is marked with a vertical line. c-d) Frequency distribution of the un-permuted p-values, respectively.

We simulated randomly significant models by shuffling the survival values among the samples and re-calculating the models. This resulted in a distribution of p -values that represent significant predictive power by chance alone (Figures 4.2a and b). We used the 5th percentile p -value for the simulated models as our new p -value cutoff for significance in the true models ($p < 0.052$). p -values for the univariate models are plotted in Figure 4.2a.

Of the models calculated with miRNA expression data from the Bentink dataset, 98 (21%) were predictive. We also calculated similar univariate models using RE scores (Figure 4.2b and d). For these models, 106 resulted in a p -value of $p < 0.05$. The empirical p -value was similar, at $p < 0.051$ (Figure 4.2b).

Models created from the expression values of 98 miRNAs (21%) resulted in p -values below the empirical p -value cutoff of $p = 0.052$ (Table C.1). We informally examined the miRNAs with significant predictive power for known associations with cancer biology, either by a published molecular mechanism or by an expression-based association with survival time or recurrence. We observed no strong association with an appropriate prognostic value: good prognosis miRNAs were not necessarily associated with a negative coefficient (indicating reduced risk), nor bad-prognosis miRNAs with a positive one (increased risk). miR-let-7g, for instance, is associated here with a negative coefficient, indicating that it predicts decreased risk. However, the miRNA is a member of the known oncomiR let-7 family, and so would be expected to associate with increased risk. Some miRNAs do accord well in these models, however. The miR-200 and miR-29 family miRNAs are a well-known poor prognosis markers in cancer [178, 13, 77, 155].

Of course, miRNA biology is complex. It is possible that these miRNAs have different functions in different contexts, and that they are predictive in the way that the models indicate. In addition, our examination of miRNA function is far from rigorous. A true analysis of functional enrichment of these miRNAs compared to the total population of miRNAs modeled here awaits the development of an appropriate miRNA functional database, such as the GO and KEGG databases available for use with gene-based data.

The models fitted on activity (RE score) profiles resulted in substantially lower coefficients associated with each miRNA, with coefficients reaching on the order of $1e - 5$, rather than the order of $1e - 01$, as is the case for the miRNA expression-based models (Table C.2). 106 of 677 (16%) miRNA RE score profiles resulted in models with significant ($p < 0.051$) p -values.

We expanded our analysis to the TCGA dataset of 502 patient-matched samples of high-grade serous ovarian tumor. RE scores were calculated from matched mRNA data in a manner identical to that used with the Bentink dataset.

Univariate models were built and an empirical p -value cutoff calculated at 5%, yielding a p -value cutoff of $p = 0.05067$ for the miRNA expression data and $p = 0.05173$ for the activity data (Figure 4.3a-d). 105 models (15%) were predictive. The miRNAs whose expression profiles result in $p < 0.051$ are listed in Table C.3. Again, more models have significant p -values than are expected by chance. This excess of significant p -values suggests that there is, in fact, useful information about survival in the miRNA expression data.

When we examine the models built on RE score data calculated from the TCGA mRNA expression dataset, we find a much smaller list of models with significant scores (Table C.4). Again, the RE scores result in models with lower coefficients. 105 expression models (15%) were predictive, whereas 88 activity score models (13%) were predictive.

We expected that, because of previously reported consistencies between the datasets, we would find that at least some of the miRNAs with good predictive power — miRNAs that resulted in univariate models with large effect sizes and small p -values — would have these predictive properties in both the TCGA and Bentink datasets. Indeed, that is the case. 17 of the miRNAs found to

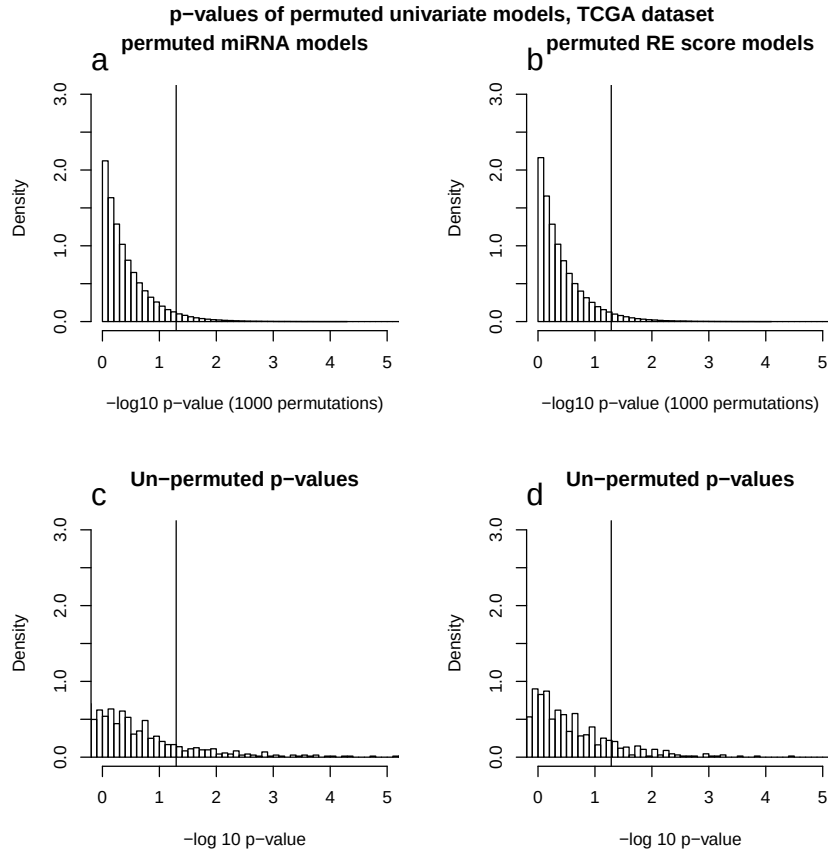


Figure 4.3: a-b) $-\log_{10}$ p-values of permuted univariate models built on TCGA miRNA expression and RE scores, respectively. The 5th percentile is marked with a vertical line. c-d) Frequency distribution of the un-permuted p-values, respectively.

be predictive in the TCGA dataset are also predictive in the Bentink dataset. This small amount of commonality in the strong predictors could be because the miRNAs whose expression values were measured were somewhat different; the microarrays that were used in the assay included a different set of probes.

The far smaller coefficients associated with the RE-score-based models indicates that direct measurement of miRNA expression, rather than indirect measure of miRNA activity via mRNA expression calculations, is more informative

with regards to survival prediction. We will focus our efforts in building future models using only the miRNA expression data. It is possible that miRNAs are not individually predictive of survival, but may be useful when aggregated together into a multivariate model. We attempt to build such a multivariate model next.

4.3 Forward stepwise selection

Having determined that miRNA expression values are minimally informative of risk in the Bentink and TCGA datasets, we wish to build combined, multivariate models that predict risk. A combined model may allow multiple, weakly-predictive covariates to combine together to result in a more powerful model. To prevent over-fitting, we wish to include in those models an optimal number of predictors, rather than the entire set of all predictors.

Another advantage of this reduction of dimensionality is that the resulting model may be interpretable; that is, the subset of miRNAs included in a high-quality model predicting survival of ovarian cancer patients are likely to be biologically meaningful; they may be involved in the etiology or progress of the disease. Further investigation of these miRNAs may elucidate mechanisms of metastasis, cancer development, angiogenesis, or other components of the cancer.

We could simply select the subset of predictors having the most power and combine them into a model. However, this simple method of building models (called best-subset selection) suffers from a severe drawback: the need to test all combinations of predictors, such that the model with the smallest residual sum of squares can be chosen. Given our count of $p = 474$ predictors for the

smallest dataset (RE scores for the Bentink dataset), we would need to test $2^p = 4.9e + 142$ combinations of predictors. Efficient algorithms are available that allow calculation of these types of combinations up to $p = 30$, but that is not sufficient in our case.

We chose instead to employ another subset selection method: forward stepwise selection. This method is similar to best-subset selection in that it produces a model predicting outcome using only part of the set of predictors available. However, the variables are selected differently. The method begins with the intercept, a simple linear model, in this case fitted with clinical covariate annotations already known to be predictive of survival (tumor stage and grade, and patient age). The method then sequentially adds other predictors, in this case miRNA expression measurements or activity scores, ordered by their ability to improve the model. The result is a set of p models, each a better predictor of outcome than the last. The number of predictors is chosen by optimizing a metric of goodness of fit of a statistical model, the Aikake Information Criterion.

4.3.1 Evaluating models using Information Criteria

The Aikake Information Criterion, or AIC, is a measure of a model's complexity against its predictive power. It can be used when selecting among a series of similar models with varying numbers of predictors, like Forward Stepwise Selection. One benefit of using an Information Criterion like this one for model comparison, rather than the log likelihood alone, is that this criterion penalizes models proportionate to the number of predictors included; complex models with many predictors are penalized more, allowing the selection of a model balanced

between good predictive power and the tendency toward overfitting.

The AIC is calculated for a model by subtracting the log likelihood of the model from the number of predictions. It is defined as

$$AIC = -2\ln(L) + 2p$$

where p is the number of coefficients in that model and $\ln(L)$ is the log likelihood for the model. Models with smaller AIC values compare favorably over models with higher values.

Another method of choosing a model is the Bayesian Information Criteria,

$$BIC = -2\ln(L) + \ln(n)p$$

where, again, p is the number of non-zero coefficients for that model, $\ln(L)$ is the log likelihood for the model and n is the number of observations. The BIC is sometimes considered a better measure of the fitness/bias tradeoff because it takes into account the dimensionality of the model by including the number of observations.

Either of these criteria can be used to compare the different models resulting from different variable selection or coefficient-penalizing techniques. Forward stepwise selection, as implemented in the `stepAIC` package used here, uses the AIC calculation to optimize model selection.

The `stepAIC` subroutine of the `MASS` package was used to do this selection step [156]. This subroutine uses the Aikake Information Criteria to determine the optimal number of predictors to use in the final model.

4.3.2 Forward stepwise selection on the Bentink miRNA expression data

We first attempted to build a model using miRNA expression data from the Bentink dataset (Table 4.1). Our initial model was simply the outcome censored survival time predicted by a subset of the clinical covariates available as part of the dataset. For the Bentink dataset, these covariates are: tumor stage, optimal debulking status and platinum-based chemotherapy status, and patient age at diagnosis.

	coefficient	AIC	p.value
hsa.miR.507	-0.66	462.74	3.12e-03
hsa.miR.196a	0.24	459.93	2.20e-02
hsa.miR.506	0.73	459.10	3.66e-02
hsa.miR.514	-0.37	456.99	1.88e-01
hsa.miR.34c	1.26	456.80	7.16e-02
hsa.miR.34b	-0.98	456.31	1.17e-01

Table 4.1: miRNAs included in optimal forward-selected model based on Bentink miRNA expression data.

In examining the model, we note first that the effects of the clinical covariates are much greater than the effect of most of the miRNAs. The exception is miR-34b, one of the strongest predictors of lower risk and even stronger than an optimal surgical outcome. miR-514 and 507, members of the 506 cluster of miRNAs described in the previous chapter, are also prognostic of lower risk. These miRNAs and other members of their miRNA family were found to be downregulated in the poorer-prognosis subtype (the angiogenic subtype) in the work of Bentink *et al.* and in Chapter 3. In addition, Bagnoli *et al.* [8] report that the 506 cluster is downregulated in early-relapse ovarian cancer patients.

The p -value for this model was calculated in the same manner as the calculation for the univariate models: a likelihood ratio test, resulting in a p -value of $p = 0.00046$.

One particularly interesting observation here is the occurrence of two members of the same family occurring in the model with opposing coefficients. miR-34b has negative coefficient, whereas miR-34c, a miRNA with an identical seed region and, presumably, similar target genes and therefore similar physiological effects, has a positive one. We can only speculate as to why this is the case. It may be that the two miRNAs (34b and c, for instance) form a negative-feedback loop, and serve to regulate each other. When one is highly expressed, it inhibits the other, and so both miRNAs are significantly associated with whatever survival effect the highly expressed miRNA may have. Another possibility is that the targets of these miRNAs differ more than we would expect, given their similar seed regions. For instance, they may target similar messages with greater or lesser affinity, and therefore have different effects in the cell.

4.3.3 Forward stepwise selection on the TCGA miRNA expression data

We performed a similar analysis using the miRNA expression data collected by TCGA. These data include similar clinical covariates to those in the Bentink dataset: patient age at diagnosis, tumor grade and optimal debulking status (tumor residual disease). Absent from this data set is the chemotherapy status, as all patients in this study were treated with chemotherapy. This dataset also includes one additional annotation: the tumor grade.

	coefficient	AIC	p.value
hsa.miR.144	0.32	2451.42	7.32e-03
hsa.miR.449a	0.12	2449.39	3.51e-04
hsa.miR.494	0.18	2446.34	3.55e-02
hsa.miR.10a	0.15	2443.24	2.94e-03
hsa.miR.151.3p	0.08	2442.71	3.17e-02
hsa.miR.223	0.24	2441.94	8.23e-03
hsa.miR.142.3p	-0.18	2439.91	3.22e-02
hsa.miR.196b	0.12	2438.91	2.35e-02
hsa.miR.7	-0.14	2438.26	2.30e-02
hsa.miR.135a	0.14	2437.75	5.94e-02
hsa.miR.224	0.13	2436.80	2.11e-02
hsa.miR.96	0.12	2435.41	4.35e-02
hsa.miR.135b	-0.10	2434.15	6.83e-02
hsa.miR.451	-0.20	2434.09	5.86e-02
hsa.miR.1225.5p	0.17	2433.07	3.63e-02
hsa.miR.10b	0.10	2432.35	6.56e-02
hsa.miR.509.3.5p	-0.09	2432.10	1.41e-01

Table 4.2: miRNAs included in optimal forward-selected model based on TCGA expression data.

Table 4.2 has the stepAIC results from the TCGA miRNA expression data ($p = 2.6e - 08$). 35 covariates were selected for this model (the top 20 are shown in Table 4.2). miR-10a's positive coefficient indicates an association with higher risk. NPM1 mutant AML samples were shown to markedly overexpress miR-10a, compared to healthy normal controls [121]. miRNA 509-3-5p, a negative coefficient miRNA here, is a member of the 506 family so strongly differentially expressed between the angiogenic and non-angiogenic subtypes in Chapter 3. This low-risk association is one consistency between the TCGA data and the Bentink dataset.

4.3.4 Little common signal between TCGA and Bentink miRNA expression datasets

There is little overlap in the membership of the selected predictors for these two models. Further, some of the selected predictors contradict one another; miR-196b is selected as a negative predictor of risk by the method when run on the Bentink data, and as a positive predictor of risk when the method is run on the TCGA data. One of the few commonalities is the association of the miR-10a/b family with high risk. The connection with the miR-506 family is also promising. But many other miRNAs show no overlap.

The orientation of the coefficients assigned to the selected miRNAs is plausible, with tumor suppressors being assigned negative coefficients indicating low risk when expression is high, and oncogenes and other proliferation related genes being assigned positive coefficients. However, there doesn't seem to be an expression signal consistently strong enough across these two datasets to be picked up by this method.

Many criticisms have been levied against stepwise selection. Subset selection methods like these can suffer from high variance due to the fact that coefficients are either used in the model or not; they are never shrunk to reduce their influence on the model. This variance could be contributing to the lack of consistency in the miRNAs selected from the two datasets. Rather than attempt to fix this method, for instance, by cross-validation, we hope to reduce this variance by way of coefficient shrinkage — a method of reducing the impact a given variable has on the model without removing it entirely.

4.4 Coefficient shrinkage and variable selection using the Lasso

Coefficient shrinkage methods reduce the impact of predictors on a model by decreasing them toward zero. The amount by which the coefficients are shrunk depends on the type of method and on the shrinkage coefficient minimizing the predicted error of the model. Shrinkage methods reduce the overfitting typical of linear regression without the destabilizing effects of removing entire coefficients, as is the case with pure feature selection methods like forward stepwise selection. Shrinkage methods have some advantages over subset selection methods in that shrunk models can have lower variance than purely feature-selected models.

The Lasso maximizes the log likelihood of the model subject to a penalty term applied to each predictor's coefficient. The penalty term consists of a user-selected tuning parameter, λ , times the L_1 norm of the coefficients.

Recall the log likelihood optimization used to select values for $\hat{\beta}$ in the Cox model,

$$\hat{\beta} = \arg \max[l_p(\beta)],$$

where l_p is the log likelihood of the model and β is the vector of coefficients applied to the predictors in the model. The Lasso estimates β' with a penalty term based on the L1 norm of the vector of predictors, multiplied by a tuning parameter, λ ,

$$\hat{\beta}^{Lasso} = \arg \max(l(\beta) - \lambda \sum_{i=1}^p |\beta_i|)$$

This penalty results in the shrinkage of many coefficients to zero, effectively combining variable selection with shrinkage.

These penalized variable selection methods have been previously used to build models based on microarray data. A comparative analysis of the Lasso and ridge penalties, among others, by Bovelstad *et al.* [25] found that ridge regression resulted in somewhat better classification than the Lasso. However, the ridge penalty does not result in shrinkage to zero, and so does not result in variable selection. This retention of all variables is a disadvantage to the model-builder interested in an interpretable model. A method like the Lasso, which shrinks less-informative predictors to zero and results in a short, interpretable list of miRNAs, could suggest biological mechanisms underlying the model. Bovelstad *et al.* did find that the Lasso performed well — outperforming subset selection — on gene expression data, resulting in small selected sets of genes, in some situations as few as five. We expect this method to perform well on the similarly-structured miRNA expression data as well.

4.4.1 Creating the models.

For this analysis, the TCGA and Bentink miRNA expression data were filtered to retain only those 368 miRNAs that were measured in both studies. With this restriction in place, models built with one dataset could be used to predict outcome in the other. Most of the models described here are trained on the TCGA dataset and then validated with the Bentink dataset. The models were built using the `penalized` R package by Jelle Goeman [63], which supports the application of the Lasso penalty to models built with censored data.

Selection of an appropriate value for the tuning parameter λ is a critical component of building useful models with the Lasso. Too large a value for λ results in a restricted set of predictors and a model with little predictive power. Too small a value for λ results in overfitting. To choose an optimal value of λ we first profiled the log-likelihood of models created using the TCGA dataset with varying values of the parameter. Log-likelihoods were calculated for 10-fold cross validated models, across multiple values of λ .

The k -fold cross-validated partial log-likelihood (*cvpl*) was calculated according to the method of Verweij and van Houwelingen [158, 117]

$$cvpl(\lambda) = \sum_{K=1}^k L_p(\hat{\beta}_{-K}^\lambda) - L_{p(-K)}(\hat{\beta}_{-K}^\lambda).$$

This calculation was done 20 times, resulting in a set of 20 log-likelihood profiles. An example of this profiling, produced when building the model based on the 368 miRNAs, from the expression measurements in the TCGA dataset, is shown in Figure 4.4a.

We also calculated Aikake and Bayesian information criteria at each value of λ , for each cross-validation (Figure 4.4b and 4.4c). While the AIC and BIC values increase monotonically with decreasing λ , the unadjusted log-likelihood has a trend towards a global maximum at λ values of between 20 and 30.

The TCGA dataset was used to build models using resampled cross-validation, repeated 1000 times. For each of these iterations, 10% (50) of the samples were randomly selected to be withheld from the model-creation process (test set). The remaining 90% (452) was used as the training set. These 452 samples were further subdivided into 10 folds, and used in an additional 10-fold cross-validation,

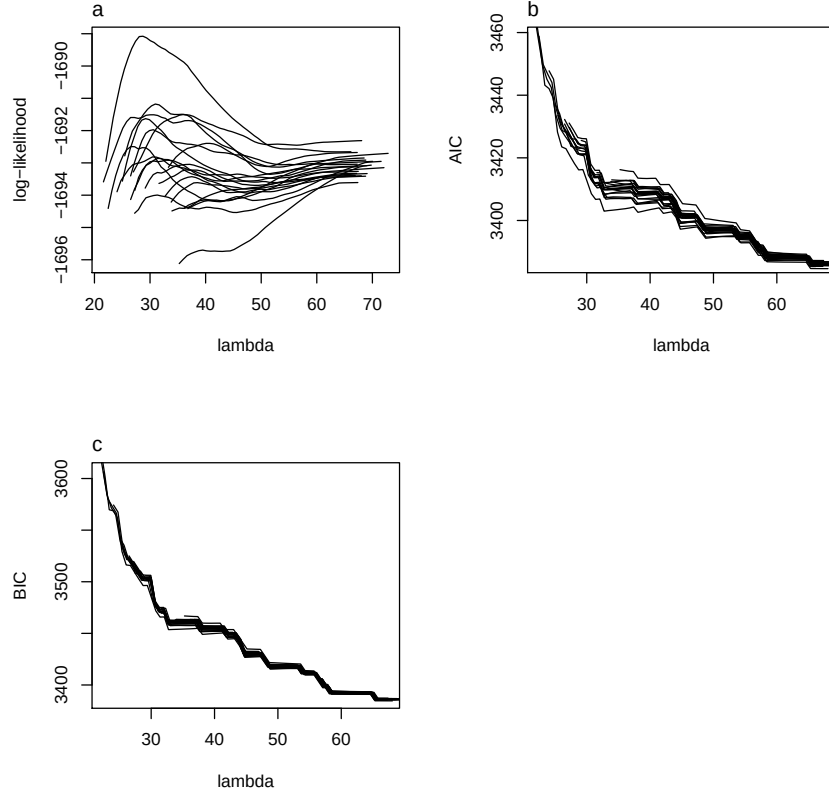


Figure 4.4: Log-likelihood of the models built with miRNA expression data at varying λ values, for each of 20 resamplings of 10-fold cross-validation on the TCGA dataset. b) Aikake Information Criterion calculated for each of the same models shown in a. c) Bayesian Information Criterion calculated for each model shown in a.

the only purpose of which was to select the most optimal value for λ , subject to $10 < \lambda < 30$. A final model was then built, using this optimal value of λ and the training set of expression data. This model, one of the 1000 created, was then used to predict the outcome for each of the remaining 10% of samples from the TCGA data and from the full Bentink data set.

The resulting 1000 models were evaluated for their ability to predict outcome for their test dataset and a concordance index was calculated for each. This mean

concordance index was used to evaluate the predictive power of the models.

4.4.2 Lasso-based modeling of ovarian cancer survival using miRNA expression data.

We hoped to find that these models would have some amount of predictive power, both within the training (TCGA) dataset and in the test (Bentink) dataset. We would expect that the CI of the cross-validated models in the training set to be stronger than the CI of the model on the test set. Results for each model in this section include average concordance indices for each model, cross-validated on the training data, tested on the test data and tested again on the external validation (Bentink) dataset. They are summarized in Table 4.3.

The mean CI of these models on the training datasets was a disappointing $CI = 0.53$, indicating that the model predicted outcome for patient survival essentially no better than chance. Each model was also used to predict outcome for the patients in the Bentink dataset, resulting in an even poorer set of predictions, with a mean CI of $CI = 0.51$. These values are disappointing as they indicate nearly no predictive power in the models. One possible factor that may weaken these models is that the angiogenic and non-angiogenic subtypes of ovarian cancer, described in Chapter 1 and further studied in Chapter 3, have differing molecular characteristics. If the miRNAs that are associated with survival in one are not in the other, the power of any model built on data taken from both subtypes would be reduced.

We created separate models built with the data for each of the two subtypes described by Bentink *et al.*, the angiogenic and non-angiogenic subtypes. The

Model	train IC	test IC	Bentink IC
All Common miRNAs	0.62	0.53	0.51
Angiogenic	0.68	0.52	0.54
Non-angiogenic	0.66	0.54	0.54
miRNA family averaging	0.62	0.54	0.5
Known Cancer miRNAs	0.6	0.52	0.49
Families with Known Cancer miRNAs	0.6	0.55	0.48

Table 4.3: Summary results for models built in Chapter 4. train IC: average concordance index of training set within the 10-fold cross-validation loop. test IC: average concordance index for predictions on each of the 1000 test sets. Bentink IC: average concordance index for predictions of each of the 1000 models on the Bentink dataset.

models were created in an identical manner to the models above. Profiling of the log-likelihood of the models indicated that a range of $\lambda = 15 - 20$ would be optimal. For the angiogenic model, the average CI on the training sets was $CI = 0.68$ and the average CI of the test sets was $CI = 0.52$. These angiogenic models were used to predict outcome for the Bentink dataset samples, resulting in a mean of $CI = 0.54$. Models built using the expression measurements for the non-angiogenic subtype resulted in an average training set CI of $CI = 0.66$ and an average test set CI of $CI = 0.54$. Testing on the Bentink data resulted in a mean of $CI = 0.54$. These poor results may be due to the small size of the angiogenic dataset — only 181 samples of the TCGA dataset are a part of this smaller subtype. However, this slight improvement in CI over the full dataset is not enough to justify separating the data by subtype.

The poor predictive power of the models generated here is possibly a result of the curse of dimensionality — this hypothesis-free approach, where the model-building process would inform us of which miRNAs were predictive, does not seem to work. We thought to reduce the dimensionality of the dataset by introducing

some biological information.

While the classification power of these models is not strong, the feature selection built into them provides a resource for investigating the most informative miRNAs with regards to patient survival time. We reexamined the first Lasso models, built using the entire set of miRNAs measured in the TCGA and Bentink datasets. To investigate these miRNAs, we selected all of the miRNAs selected as features for each of the 1000 models. That is, we simply counted, for each miRNA, how many times it had been assigned a non-zero coefficient. Presumably, miRNAs selected often are more informative than miRNAs selected only a few times.

One thing that this type of multi-model creation allows is the investigation of groups of miRNAs with redundant predictive power. The Lasso should select only one miRNA out of a group of highly correlated predictors, and if those predictors are similarly powerful, the selection of the particular one should be random (Table 4.4).

If the models are, indeed, picking up on some information in the data but are unable to select miRNAs reliably due to the large number of predictors available, perhaps some sort of dimension reduction before model-building would result in a better model.

	percent
hsa.miR.148a	93
hsa.miR.592	91
hsa.miR.652	82
hsa.miR.505	82
hsa.miR.604	67
hsa.miR.504	63
hsa.miR.146a	62
hsa.miR.320	57

hsa.miR.326	56
hsa.miR.198	53
hsa.miR.657	50
hsa.miR.204	36
hsa.miR.569	33
hsa.miR.150	30
hsa.miR.422a	30

Table 4.4: miRNAs with non-zero coefficients from TCGA univariate models.

miRNAs are grouped into families – sets of miRNAs related by sequence and therefore often having similar target genes. We supposed that averaging individual miRNA expression profiles across these families may decrease the redundancy in the dataset — miRNAs belonging to the same family may be performing the same functions and, therefore, have similar expression profiles. We used the miR-Base database [64, 2, 65, 66] to choose miRNAs belonging to the same family and average their expression profiles. This resulted in 200 profiles.

We further restricted our model by selecting miRNAs already known to be involved in cancer development and progression. We proposed that using this restricted set of predictors as input into Lasso would cut down on the problem of overconstraining, and perhaps result in a better model.

The Database of Differentially Expressed miRNAs in human Cancers (db-DEMC) is a manually-curated catalog of miRNAs with measurable differential expression between human cancer tissues and matched normal tissues [172]. These changes in expression were measured using both large-scale expression measurement methods, such as microarrays, and low-throughput methods, such as qRT-PCR. Many cancer types are cataloged, including prostate, breast, pancreatic, and non-solid tumors such as lymphomas. Ovarian cancer data is not

included in the database. We chose to use the gynecologic cancers recorded in the database: breast and uterine cancer.

We used these validated miRNAs to filter our miRNA expression data (again, only miRNAs common between the TCGA and Bentink datasets) down to 210 validated, cancer-related miRNAs. The model was created in the same way as the others; trained and cross-validated on the TCGA dataset and tested on the Bentink dataset.

Again, the predictive power of the resulting model is near chance: the average test dataset concordance index of 0.49 indicates that the model has essentially no predictive power.

We restricted the family model further, to retain only miRNA families that included one or more known gynecologic cancer miRNAs, resulting in 83 profiles. Again, this model resulted in a poor concordance index ($CI = 0.48$), indicating little to no predictive power.

4.4.3 Examining the models.

In order to determine why these models weren't working, we reexamined the univariate analysis results from both datasets. We had originally found little concordance in the membership of the miRNAs with top-ranked p -values. We next examined the pattern of univariate models further by comparing the p -values for each of the 368 miRNAs for which a model was created using the data in both datasets. We would expect that, if the miRNAs measured in the two datasets had similar predictive power across the datasets, we would find a strong relationship between a miRNA's univariate model p -value in each dataset; if a

model based on miR-100 was very low in one dataset, it should be low in the other, and vice versa.

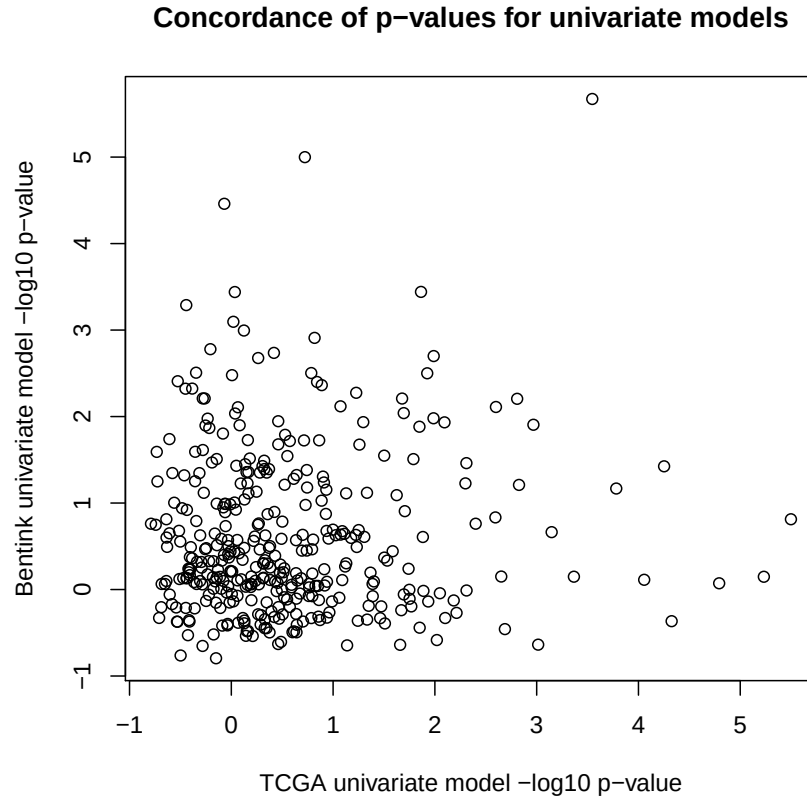


Figure 4.5: $-\log_{10}$ p-values for univariate models fitted on the miRNA expression in the TCGA dataset (x-axis) vs. the Bentink dataset (y-axis).

However when the p -values for these univariate models were compared, we found little relationship between them (correlation $r = -0.033$, Figure 4.5). Similarly, the effect sizes have a low correlation (correlation $r = -0.0037$, Figure 4.6).

Other consistencies between these two datasets, reported in chapters 2 and 3 and by other authors, such as Bentink *et al.*, indicate that this inconsistency is

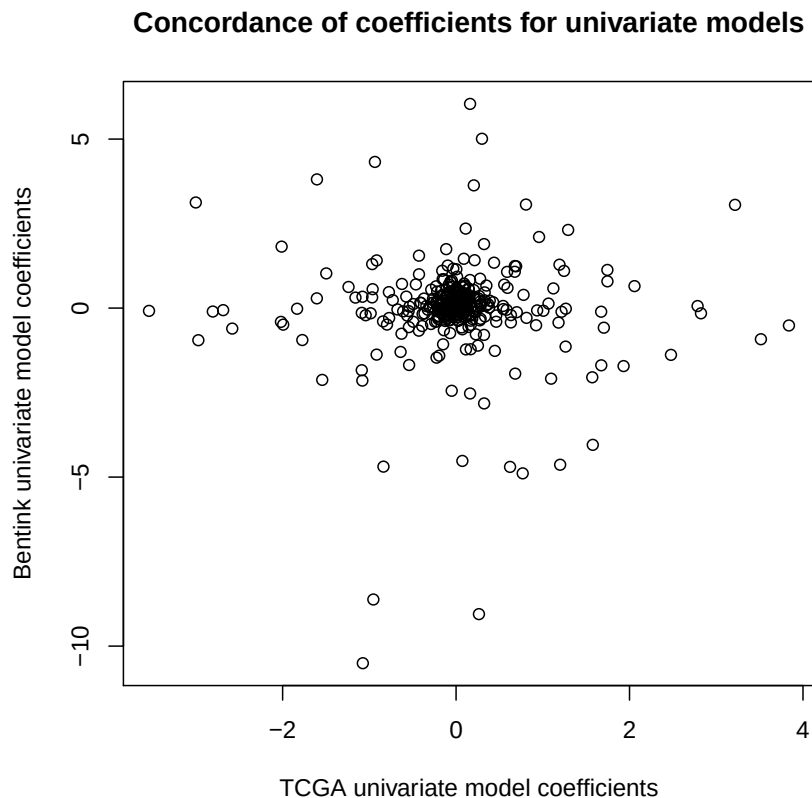


Figure 4.6: Effect sizes for the univariate models fitted on the miRNA expression in the TCGA dataset (x-axis) vs. the Bentink dataset (y-axis).

not representative of a general lack of consistency between the datasets. It seems that the expression level of individual miRNAs is not closely-enough related to survival to be detectable in these data.

4.5 Conclusions and Future Work

We have shown here that, at least for datasets of this size, miRNA expression data and activity score data are not particularly useful for predicting survival, at least not when approached in a hypothesis-free manner.

We found that the miRNA expression data was likely to contain some predictive information, as the number of univariate models with significant p -values was greater than the number expected by chance. The miRNA activity score models also had a greater number of significant p -values than expected by chance, but the coefficients for these scores were not nearly so high, and so we focused our further model-building on direct expression measurements. Activity scores are an aggregate measurement of the expression profiles of many genes, and as such likely include more noise than direct miRNA measurements.

We continued to try building simple multivariate models with miRNA expression data, using forward stepwise selection. We had some success, in that the models we built included some sensible miRNAs and had low p -values. We attempted to improve on these models by moving to a variable shrinkage method, the Lasso.

The Lasso-based models had some small predictive power when cross-validated on the training dataset, but when tested on the test-subset, or the external validation (Bentink) dataset, all predictive power was lost ($CI \approx 0.5$). Many variations of the Lasso model resulted in poor predictive performance. The best model seemed to be build on the samples from the angiogenic subtype alone, though its performance on the test dataset was still poor ($CI = 0.54$).

Our attempts to improve the models by subsetting the data according to molecular phenotype, or to use time-to-progression as the endpoint rather than overall survival, did not result in noticeably better concordance indices on the test datasets, though the training sets did improve a small amount.

One possibility is that the datasets presented here are too small to allow a variable selection method like the Lasso to pick up weak predictive signals from

a great deal of noise among the many possible predictors. Splitting the data into subtypes improves the model somewhat, but is not sufficient. Given more samples, perhaps a more informative model could be built. It could also be that there are small, weak results masked out by the large number of non-informative miRNAs. A much larger sample size would allow us to find these results.

Models built with expression data from the angiogenic subtype alone have particular potential. This subtype, concordant as it is with subtypes described by several other groups in the literature, seems to represent a clear molecular phenotype. Given more data, this subtype may be the most promising for further modeling. Perhaps combining the angiogenic subtype data from the TCGA and Bentink datasets into a single dataset and building a model on that would produce better predictions.

One possible means to improve these models, given the data available today, would be to combine the Bentink and TCGA datasets together to provide a larger sample size, thereby reducing the problem of overconstraining. Combining these datasets would mean that we would no longer have a test set with which to validate, though future studies of miRNA expression in ovarian cancer may soon provide additional datasets.

A potential problem with these data may simply be that the datasets were collected on different platforms: the Bentink dataset on a custom Illumina miRNA array, and the TCGA dataset on a commercially available Affymetrix array. As the latter is more common, we can expect more miRNA expression data collected on Affymetrix arrays to become available in the future. Perhaps these types of hypothesis-free models will become more useful as more data is collected.

Chapter 5

Conclusions

MicroRNAs are important actors in cancer and many other diseases as well as being critical components of many biological systems including development and tissue differentiation [39, 59, 29, 97, 48]. Developing an understanding of the activity of these molecules in ovarian cancer is critical to a more complete understanding of the disease and the eventual development of preventative measures and treatments. Furthermore, the development of drugs that target miRNAs or miRNA targets or regulators could provide powerful new therapies for a disease that has not seen treatment improvements in the last ten years.

The goal of this thesis was to better understand the functionality of miRNAs in ovarian cancer, both to provide a better understanding of the underlying biological mechanisms driving the disease, and to search for potential biomarkers that would predict patient prognosis. We first presented a synopsis of known miRNA activity in ovarian cancer and a summary of miRNA mechanisms and common bioinformatics tools used to understand miRNA and mRNA expression data.

We explored the relationships between miRNA and gene expression in ovarian cancer datasets. We examined the partial correlation between the expression profiles of these molecules across datasets and in two ovarian subtypes. We found that the distribution of coefficients had heavier tails than a normal distribution. Some of the strong positive correlation miRNA-mRNA pairs were determined to be the result of miRNAs encoded either within the intron of the respective gene or within a large gene's length (500kb) of that gene. This sort of positive correlation is expected, as miRNAs transcribed with their "host" gene would be expected to have a strongly positive correlation with that host. miRNAs encoded within a short distance of another gene could also be expected to have related transcription profiles — they could be affected by local events such as methylation and chromosomal lesions.

The discovery and understanding of molecular phenotypes of cancer has led to tailored treatments that are more effective in diseases like breast cancer. Subtypes of serous ovarian cancer have only recently been discovered, and so an understanding of their molecular phenotypes is still under development. We examined the different miRNA expression and activity patterns in two subtypes of ovarian cancer with the goal of better understanding the molecular mechanisms that drive the disease, and in hope of potentially uncovering mechanisms that could be targeted by drug therapies. We investigated the patterns of miRNA expression in the two subtypes of ovarian cancer described by Bentink, *et al.* Indeed, strong patterns of differential expression — both miRNA and gene expression - distinguish the two groups. We established an overall decrease in miRNA activity in the angiogenic subtype, possibly due to a slightly lower expression of miRNA processing genes in that subtype. We further determined

that a small group of miRNAs, comprising the 506-514 cluster, were a salient difference between the subtypes and could be driving large changes in gene expression between the two groups, both through direct mRNA suppression and indirectly through one or more transcription factor intermediates. These results generate specific hypotheses that can be tested in the lab, and would shed light on potential mechanisms for ovarian cancer development and spread. A better understanding of these mechanisms can uncover potential drug targets that may reduce the impact of the disease.

While models of survival have been made using gene expression data in the past, no unsupervised models of miRNA expression data using variable selection have been made in ovarian cancer. We developed several types of unsupervised models, using univariate, variable-selected and variable shrinkage and selection methods. Inconsistent selection of variables, plus low concordance indices indicate poor performance when used to predict outcome on a test dataset. We establish that, at the very least, larger datasets will be necessary to build useful models.

Appendix A

Chapter 2 Additional Tables

	mRNA	mRNAChr	miRNA	miRNAChr	corr
1	SCTR	2	hsa-miR-448	X	0.012
2	HTR2C	X	hsa-miR-448	X	0.011
3	POU3F4	X	hsa-miR-448	X	0.011
4	TF	3	hsa-miR-620	12	0.011
5	TF	3	hsa-miR-580	5	0.011
6	ASGR1	17	hsa-miR-448	X	0.0098
7	NR5A2	1	hsa-miR-620	12	0.0096
8	NR5A2	1	hsa-miR-580	5	0.0096
9	HOXA9	7	hsa-miR-196b	7	0.0094
10	CALML3	10	hsa-miR-620	12	0.0094
11	HOXB5	17	hsa-miR-10a	17	0.009
12	NNAT	20	hsa-miR-648	22	0.0089
13	HBB	11	hsa-miR-451	17	0.0089
14	CALML3	10	hsa-miR-580	5	0.0088
15	TF	3	hsa-miR-597	8	0.0087
16	NR5A2	1	hsa-miR-597	8	0.0086
17	CALML3	10	hsa-miR-597	8	0.0084
18	HOXA5	7	hsa-miR-196b	7	0.0084
19	HOXA4	7	hsa-miR-196b	7	0.0083
20	HBB	11	hsa-miR-144	17	0.0082
21	HOXD3	2	hsa-miR-10b	2	0.0082
22	SPAG6	10	hsa-miR-648	22	0.0082
23	TF	3	hsa-miR-644	20	0.0081
24	DLX5	7	hsa-miR-217	2	0.008
25	FOXI1	5	hsa-miR-580	5	0.0079
26	HOXB2	17	hsa-miR-10a	17	0.0079
27	SLITRK5	13	hsa-miR-611	11	0.0077
28	SLITRK3	3	hsa-miR-611	11	0.0077
29	CALML3	10	hsa-miR-640	19	0.0076
30	TF	3	hsa-miR-640	19	0.0075

Table A.1: miRNA-mRNA target pairs with the top 30 largest positive partial correlations (Bentink dataset).

	mRNA	mRNAChr	miRNA	miRNAChr	corr
1	HOXB2	17	hsa.miR.10a	17	0.02
2	DDX3X	X	hsa.miR.646	20	0.019
3	HBB	11	hsa.miR.144	17	0.016
4	HBB	11	hsa.miR.451	17	0.015
5	HOXB6	17	hsa.miR.10a	17	0.015
6	HOXD9	2	hsa.miR.10b	2	0.014
7	HOXA5	7	hsa.miR.196b	7	0.014
8	HOXB5	17	hsa.miR.10a	17	0.014
9	HOXA1	7	hsa.miR.196b	7	0.013
10	HOXB7	17	hsa.miR.10a	17	0.013
11	CTAG2	X	hsa.miR.217	2	0.013
12	CCNO	5	hsa.miR.449b	5	0.012
13	HOXA4	7	hsa.miR.196b	7	0.012
14	HOXD3	2	hsa.miR.10b	2	0.011
15	MYH11	16	hsa.miR.133b	6	0.011
16	OPRK1	8	hsa.miR.661	8	0.011
17	HOXA6	7	hsa.miR.196b	7	0.011
18	SKAP2	7	hsa.miR.196b	7	0.011
19	MYH11	16	hsa.miR.143	5	0.011
20	HOXD1	2	hsa.miR.10b	2	0.011
21	AKT2	19	hsa.miR.641	19	0.0099
22	EREG	4	hsa.miR.384	X	0.0099
23	PAH	12	hsa.miR.646	20	0.0098
24	CALY	10	hsa.miR.489	7	0.0097
25	EREG	4	hsa.miR.626	15	0.0097
26	CSN2	4	hsa.miR.675	11	0.0095
27	HTN3	4	hsa.miR.675	11	0.0093
28	MYH11	16	hsa.miR.145	5	0.0092
29	SSX5	X	hsa.miR.504	X	0.0092
30	PPBP	4	hsa.miR.144	17	0.0091

Table A.2: miRNA-mRNA target pairs with the top 30 largest positive partial correlations (TCGA Dataset).

	mRNA	mRNAChr	miRNA	miRNAChr	corr
1	HTR2C	X	hsa-miR-448	X	0.011
2	HOXA9	7	hsa-miR-196b	7	0.0094
3	HOXB5	17	hsa-miR-10a	17	0.009
4	HOXA5	7	hsa-miR-196b	7	0.0084
5	HOXA4	7	hsa-miR-196b	7	0.0083
6	HOXD3	2	hsa-miR-10b	2	0.0082
7	HOXB2	17	hsa-miR-10a	17	0.0079
8	SKAP2	7	hsa-miR-196b	7	0.0068
9	HOXB7	17	hsa-miR-10a	17	0.0066
10	HOXD1	2	hsa-miR-10b	2	0.0065
11	IHH	2	hsa-miR-375	2	0.0063
12	HOXB8	17	hsa-miR-10a	17	0.0058
13	HOXD9	2	hsa-miR-10b	2	0.0057
14	CYCS	7	hsa-miR-148a	7	0.0056
15	GPC1	2	hsa-miR-149	2	0.0055
16	HOXA6	7	hsa-miR-196b	7	0.0052
17	STRN3	14	hsa-miR-624	14	0.005
18	PYGB	20	hsa-miR-644	20	0.0049
19	AKT2	19	hsa-miR-641	19	0.0048
20	HOXA11	7	hsa-miR-196b	7	0.0048
21	PVALB	22	hsa-miR-659	22	0.0047
22	CCDC130	19	hsa-miR-181d	19	0.0045
23	PPP2R1A	19	hsa-miR-99b	19	0.0043
24	CPA2	7	hsa-miR-335	7	0.0043
25	DVL2	17	hsa-miR-324-5p	17	0.0042
26	ATP5H	17	hsa-miR-636	17	0.0042
27	KIAA0753	17	hsa-miR-324-3p	17	0.0041
28	TULP1	6	hsa-miR-586	6	0.0041
29	DVL2	17	hsa-miR-324-3p	17	0.0041
30	GMPS	3	hsa-miR-15b	3	0.004

Table A.3: miRNA-mRNA pairs with the top 30 largest positive partial correlations, within 10,00kb on the chromosome (Bentink dataset).

	mRNA	mRNAChr	miRNA	miRNAChr	corr
1	HOXB2	17	hsa.miR.10a	17	0.02
2	HOXB6	17	hsa.miR.10a	17	0.015
3	HOXD9	2	hsa.miR.10b	2	0.014
4	HOXA5	7	hsa.miR.196b	7	0.014
5	HOXB5	17	hsa.miR.10a	17	0.014
6	HOXA1	7	hsa.miR.196b	7	0.013
7	HOXB7	17	hsa.miR.10a	17	0.013
8	CCNO	5	hsa.miR.449b	5	0.012
9	HOXA4	7	hsa.miR.196b	7	0.012
10	HOXD3	2	hsa.miR.10b	2	0.011
11	HOXA6	7	hsa.miR.196b	7	0.011
12	SKAP2	7	hsa.miR.196b	7	0.011
13	HOXD1	2	hsa.miR.10b	2	0.011
14	AKT2	19	hsa.miR.641	19	0.0099
15	SLIT3	5	hsa.miR.585	5	0.0088
16	USP25	21	hsa.miR.99a	21	0.0087
17	CTAG2	X	hsa.miR.767.5p	X	0.008
18	TLE3	15	hsa.miR.629	15	0.0079
19	HOXA11	7	hsa.miR.196b	7	0.0078
20	RNF130	5	hsa.miR.340	5	0.0077
21	BFAR	16	hsa.miR.484	16	0.0076
22	CRKL	22	hsa.miR.185	22	0.0076
23	DVL2	17	hsa.miR.324.5p	17	0.0076
24	BFAR	16	hsa.miR.193b	16	0.0075
25	MKL2	16	hsa.miR.193b	16	0.0074
26	GABARAP	17	hsa.miR.324.5p	17	0.0073
27	CXADR	21	hsa.miR.99a	21	0.0073
28	MED15	22	hsa.miR.185	22	0.0073
29	SNAP29	22	hsa.miR.185	22	0.0073
30	SREBF1	17	hsa.miR.33b	17	0.0071

Table A.4: miRNA-mRNA pairs with the top 30 largest positive partial correlations, within 10,00kb on the chromosome (TCGA Dataset).

	mRNA	mRNAChr	miRNA	miRNAChr	corr
1	PUM1	1	hsa-miR-552	1	-0.0054
2	ZNF586	19	hsa-miR-527	19	-0.0039
3	DZIP3	3	hsa-miR-567	3	-0.0034
4	QDPR	4	hsa-miR-573	4	-0.0032
5	GFER	16	hsa-miR-662	16	-0.0032
6	IL2	4	hsa-miR-302a	4	-0.0031
7	ST6GALNAC2	17	hsa-miR-636	17	-0.0031
8	STK10	5	hsa-miR-585	5	-0.0031
9	MYH8	17	hsa-miR-212	17	-0.003
10	PLOD3	7	hsa-miR-653	7	-0.003
11	SPIN2A	X	hsa-miR-98	X	-0.0029
12	PDHX	11	hsa-miR-610	11	-0.0029
13	LDLR	19	hsa-miR-640	19	-0.0029
14	APBA3	19	hsa-miR-637	19	-0.0029
15	ZNF480	19	hsa-miR-520e	19	-0.0028
16	APOC2	19	hsa-miR-527	19	-0.0028
17	CH25H	10	hsa-miR-607	10	-0.0028
18	MAN1B1	9	hsa-miR-602	9	-0.0027
19	HBB	11	hsa-miR-210	11	-0.0027
20	IGLL1	22	hsa-miR-648	22	-0.0027
21	MSMB	10	hsa-miR-605	10	-0.0027
22	BPGM	7	hsa-miR-593	7	-0.0027
23	TMEM53	1	hsa-miR-552	1	-0.0027
24	CCDC49	17	hsa-miR-144	17	-0.0026
25	TSPAN32	11	hsa-miR-210	11	-0.0026
26	TTYH1	19	hsa-miR-519d	19	-0.0026
27	WBP2	17	hsa-miR-657	17	-0.0026
28	ATP1B1	1	hsa-miR-557	1	-0.0025
29	USP29	19	hsa-miR-526b	19	-0.0025
30	ZNF227	19	hsa-miR-520e	19	-0.0025

Table A.5: miRNA-mRNA pairs with the top 30 largest negative partial correlations, within 10,00kb on the chromosome (Bentink dataset).

	mRNA	mRNAChr	miRNA	miRNAChr	corr
1	WNT8B	10	hsa.miR.608	10	-0.0052
2	TMEM11	17	hsa.miR.144	17	-0.0046
3	MXD4	4	hsa.miR.571	4	-0.0045
4	CYP3A4	7	hsa.miR.489	7	-0.0042
5	SRSF6	20	hsa.miR.644	20	-0.0041
6	CMA1	14	hsa.miR.624	14	-0.0041
7	BDKRB1	14	hsa.miR.655	14	-0.0041
8	PTPN1	20	hsa.miR.646	20	-0.0039
9	LILRA2	19	hsa.miR.527	19	-0.0039
10	LETM1	4	hsa.miR.571	4	-0.0039
11	SEC24D	4	hsa.miR.302b	4	-0.0037
12	KL	13	hsa.miR.621	13	-0.0035
13	ARHGAP35	19	hsa.miR.372	19	-0.0034
14	ABCB4	7	hsa.miR.591	7	-0.0034
15	MATN1	1	hsa.miR.552	1	-0.0033
16	TMEM53	1	hsa.miR.552	1	-0.0033
17	WHSC2	4	hsa.miR.571	4	-0.0033
18	UBTD1	10	hsa.miR.608	10	-0.0033
19	LMAN2	5	hsa.miR.585	5	-0.0033
20	PCGF3	4	hsa.miR.571	4	-0.0033
21	LDLR	19	hsa.miR.639	19	-0.0032
22	SLC22A7	6	hsa.miR.133b	6	-0.0032
23	TMEM11	17	hsa.miR.451	17	-0.0031
24	MMP9	20	hsa.miR.645	20	-0.0031
25	PTDSS2	11	hsa.miR.675	11	-0.0031
26	GNAI3	1	hsa.miR.553	1	-0.0031
27	SLC15A2	3	hsa.miR.567	3	-0.0031
28	ZNF586	19	hsa.miR.520f	19	-0.0031
29	ADRBK2	22	hsa.miR.650	22	-0.0031
30	FRS3	6	hsa.miR.586	6	-0.003

Table A.6: miRNA-mRNA pairs with the top 30 largest negative partial correlations, within 10,00kb on the chromosome (TCGA Dataset).

	mRNA	mRNAChr	miRNA	miRNAChr	corr
1	CCDC25	8	hsa-miR-648	22	-0.0072
2	H1F0	22	hsa-miR-302a	4	-0.007
3	ADH5	4	hsa-miR-633	17	-0.0067
4	NPAS1	19	hsa-miR-648	22	-0.0067
5	RFPL1	22	hsa-miR-302a	4	-0.0065
6	ADH5	4	hsa-miR-611	11	-0.0061
7	MORC2	22	hsa-miR-607	10	-0.006
8	FGF2	4	hsa-miR-200c	12	-0.006
9	ARFGAP3	22	hsa-miR-553	1	-0.0057
10	IL24	1	hsa-miR-302a	4	-0.0056
11	KRT12	17	hsa-miR-335	7	-0.0055
12	SCTR	2	hsa-let-7d	9	-0.0054
13	PUM1	1	hsa-miR-552	1	-0.0054
14	ITGB1BP3	19	hsa-miR-567	3	-0.0054
15	BCAS1	20	hsa-miR-200c	12	-0.0053
16	NPAS1	19	hsa-miR-657	17	-0.0053
17	ADH5	4	hsa-miR-602	9	-0.0052
18	XPA	9	hsa-miR-623	13	-0.0052
19	FGF2	4	hsa-miR-141	12	-0.0052
20	SQRDL	15	hsa-miR-555	1	-0.0052
21	BHLHB3	12	hsa-miR-448	X	-0.0052
22	GRB14	2	hsa-miR-363	X	-0.0051
23	SCTR	2	hsa-miR-98	X	-0.0051
24	AZGP1	7	hsa-miR-200c	12	-0.0051
25	RFPL1	22	hsa-miR-302b	4	-0.005
26	MAGEA8	X	hsa-miR-570	3	-0.005
27	MFSD11	17	hsa-miR-629	15	-0.005
28	PTPN22	1	hsa-miR-640	19	-0.005
29	CLTA	9	hsa-miR-302a	4	-0.005
30	HR	8	hsa-miR-606	10	-0.0049

Table A.7: miRNA-mRNA target pairs with the top 30 largest negative partial correlations (Bentink dataset).

	mRNA	mRNAChr	miRNA	miRNAChr	corr
1	PTBP1	19	hsa.miR.600	9	-0.009
2	PTBP1	19	hsa.miR.614	12	-0.0088
3	NPTN	15	hsa.miR.137	1	-0.0084
4	NPPB	1	hsa.miR.141	12	-0.0078
5	CUL2	10	hsa.miR.489	7	-0.0074
6	TM4SF4	3	hsa.miR.141	12	-0.0069
7	NPPB	1	hsa.miR.200c	12	-0.0068
8	PTBP1	19	hsa.miR.549	15	-0.0066
9	ZNF343	20	hsa.miR.599	8	-0.0066
10	ZMYM4	1	hsa.miR.651	X	-0.0065
11	VSNL1	2	hsa.miR.604	10	-0.0064
12	TAX1BP1	7	hsa.miR.661	8	-0.0063
13	TLE1	9	hsa.miR.629	15	-0.0062
14	AP3D1	19	hsa.miR.646	20	-0.0062
15	KPNA3	13	hsa.miR.591	7	-0.0062
16	GCKR	2	hsa.miR.553	1	-0.0062
17	COX5B	2	hsa.miR.573	4	-0.0061
18	FRG1	4	hsa.miR.217	2	-0.0061
19	HOXD3	2	hsa.miR.10a	17	-0.006
20	GTPBP1	22	hsa.miR.589	7	-0.006
21	CMPK1	1	hsa.miR.10b	2	-0.006
22	PPCS	1	hsa.miR.217	2	-0.006
23	IKZF1	7	hsa.miR.606	10	-0.006
24	CTAGE1	18	hsa.miR.635	17	-0.006
25	ITGB5	3	hsa.miR.562	2	-0.006
26	ZNF467	7	hsa.miR.621	13	-0.0059
27	KIAA0391	14	hsa.miR.646	20	-0.0059
28	MLL4	19	hsa.miR.571	4	-0.0059
29	RPL10	X	hsa.miR.585	5	-0.0058
30	ARNT	1	hsa.miR.302b	4	-0.0058

Table A.8: miRNA-mRNA target pairs with the top 30 largest negative partial correlations, within 10,00kb on the chromosome (TCGA Dataset).

Appendix B

Chapter 3 Additional Tables

	miRNA	d-value	f-test	RE test	RE diff
1	hsa-miR-22	6.11090384286018	4.4e-01	0.37	Pos
2	hsa-miR-152	5.63634808136578	1.3e-01	0.58	Neg
3	hsa-miR-31	4.83792231438214	1.2e-03	0.64	Neg
4	hsa-miR-23a	4.46874869977806	4.9e-02	0.26	Neg
5	hsa-miR-645	4.33474087080049	1e+00	0.25	Pos
6	hsa-miR-603	4.24626063585685	1.5e-01	0.49	Neg
7	hsa-miR-409-3p	3.75149035776316	4.4e-01	0.47	Neg
8	hsa-miR-155	3.52864903161192	4.4e-03	0.75	Neg
9	hsa-miR-494	3.519958208806	1.9e-03	0.75	Neg
10	hsa-miR-132	3.51791690457135	9.6e-01	0.69	Neg
11	hsa-miR-199a*	3.46541859282364	1.7e-03	NA	NA
12	hsa-miR-199b	3.33171987323433	2.5e-02	NA	NA
13	hsa-miR-452*	3.30394443060902	1e+00	NA	NA
14	hsa-miR-181a	3.29119805530226	7.3e-02	0.68	Pos
15	hsa-miR-223	3.28530193441059	8.8e-01	0.19	Pos
16	hsa-miR-213	3.05566118907489	8.8e-01	NA	NA
17	hsa-miR-618	3.03643902223845	8.9e-01	0.26	Neg

Table B.1: Top 30 upregulated miRNAs in Angiogenic subtype of Bentink dataset. d-value: delta value calculated by SAM. miRNA f-test: The p-value resulting from a test of differential variance between sample groups (F-test). RE test: p-value of the t-test comparing Pita-based RE scores between the angiogenic and non-angiogenic subtypes. RE diff: Indicates whether the miRNA had a negative or positive RE change. Pos: RE scores were higher in the angiogenic subtype. Neg: RE scores were lower in the angiogenic subtype.

	miRNA	d-value	f-test	RE test	RE diff
1	hsa.miR.22	14.5759914487857	1.6e-01	1.8e-02	Pos
2	hsa.miR.22.	13.5249969296589	1e+00	NA	NA
3	hsa.miR.152	12.7668394139521	9.8e-01	9.5e-04	Neg
4	hsa.miR.21.	10.8201868895572	3.7e-03	NA	NA
5	hsa.miR.214	9.19825342854156	5.9e-03	4.1e-01	Pos
6	hsa.miR.21	9.03382174336964	2.3e-09	5.3e-01	Neg
7	hsa.miR.409.3p	8.69248981290253	2.2e-01	6.1e-04	Neg
8	hsa.miR.181a.	8.40948825316703	6.2e-02	NA	NA
9	hsa.miR.199b.5p	8.37852218525244	2.4e-01	5.8e-01	Neg
10	hsa.miR.199a.5p	8.24585556784546	8e-03	5.8e-01	Neg
11	hsa.miR.370	7.82873940709167	1e+00	2.7e-03	Pos
12	hsa.miR.214.	7.80740785241898	2.5e-01	NA	NA
13	hsa.miR.132	7.53823330226452	2.7e-01	4.3e-10	Neg
14	hsa.miR.142.5p	7.40909134765141	6.8e-03	9.6e-06	Neg
15	hsa.miR.382	7.32184893555379	1.6e-01	4.3e-09	Neg
16	hsa.miR.199b.3p	7.02267392080627	1.4e-04	8.1e-11	Neg
17	hsa.miR.142.3p	6.73438407792585	4.2e-02	5.8e-06	Neg
18	hsa.miR.379	6.5480476626936	2.5e-02	4.3e-06	Neg
19	hsa.miR.127.3p	6.47116045447919	6.2e-03	9.0e-05	Pos
20	hsa.let.7b	6.43215142796517	1.5e-03	1.4e-01	Neg
21	hsa.miR.150	6.19356359472626	2.9e-01	7.6e-01	Pos
22	hsa.miR.31	6.02696145707174	1.6e-04	7.5e-01	Pos
23	hsa.miR.24	5.72303397951046	1.3e-07	1.7e-03	Neg
24	hsa.miR.452	5.65036780466417	2e-03	4.1e-11	Neg
25	hsa.miR.645	5.59764782710086	1e+00	3.5e-01	Pos
26	hsa.miR.193a.5p	5.59131307158961	6.3e-01	6.5e-03	Neg
27	hsa.miR.155	5.56947337893767	1.6e-04	2.6e-04	Neg
28	hsa.miR.342.3p	5.29191553243484	1.2e-03	9.8e-07	Neg
29	hsa.miR.630	5.25535688376085	1e+00	5.5e-01	Neg
30	hsa.miR.27b	5.23617770799762	3.6e-04	1.6e-10	Neg

Table B.2: Top 30 upregulated miRNAs in the angiogenic subtype of the TCGA dataset. d-value: delta value calculated by SAM. miRNA f-test: The p-value resulting from a test of differential variance between sample groups (F-test). RE test: p-value of the t-test comparing Pita-based RE scores between the angiogenic and non-angiogenic subtypes. RE diff: Indicates whether the miRNA had a negative or positive RE change. Pos: RE scores were higher in the angiogenic subtype. Neg: RE scores were lower in the angiogenic subtype.

	miRNA	d-value	f-test	RE test	RE diff
1	hsa-miR-514	-7.62262982978304	1.2e-14	0.0322	Pos
2	hsa-miR-202*	-7.50335189045396	< 1e-16	NA	NA
3	hsa-miR-509	-7.18034797347884	1e-03	NA	NA
4	hsa-miR-508	-6.55506990784591	3.5e-07	NA	NA
5	hsa-miR-506	-6.21247481904752	4.4e-11	0.0432	Neg
6	hsa-miR-202	-6.16921382703114	9.5e-10	0.1732	Neg
7	hsa-miR-532	-5.50253856407225	4.6e-01	NA	NA
8	hsa-miR-510	-4.9401297643018	< 1e-16	0.9856	Pos
9	hsa-miR-660	-4.92007418146522	1.1e-01	0.8459	Pos
10	hsa-miR-500	-4.80153574588705	2.2e-02	0.8901	Pos
11	hsa-miR-513	-4.61123034244555	6.6e-14	NA	NA
12	hsa-miR-30e-5p	-4.59988368047642	7.9e-01	NA	NA
13	hsa-miR-507	-4.46360303507421	1.5e-02	0.0868	Neg
14	hsa-miR-188	-4.42316572507479	8.6e-01	NA	NA
15	hsa-miR-10b	-4.24206486311524	8.9e-01	0.2175	Pos
16	hsa-miR-497	-4.12085160381904	6.5e-01	0.5251	Pos
17	hsa-miR-362	-4.10555874283819	7.8e-03	NA	NA
18	hsa-miR-101	-3.96670729141043	2e-01	0.4266	Neg
19	hsa-miR-502	-3.83265409942297	1e-02	NA	NA
20	hsa-miR-29c	-3.49913127513713	3.3e-01	0.0053	Neg
21	hsa-miR-592	-3.33423688498934	8.9e-03	0.0268	Pos
22	hsa-miR-328	-3.21308504281908	2.5e-02	0.9486	Pos
23	hsa-miR-195	-3.18724768232318	6.1e-01	0.5251	Pos
24	hsa-miR-186	-3.17934884511477	1e+00	0.8286	Pos
25	hsa-miR-99a	-3.16823596985077	9.3e-01	0.0235	Neg
26	hsa-miR-488	-3.15063455115632	2e-07	0.9942	Pos
27	hsa-miR-130a	-3.1049627301132	9.8e-01	0.4338	Neg
28	hsa-miR-345	-3.05379541192924	7.9e-01	0.2726	Neg
29	hsa-miR-30e-3p	-3.04632424695345	9.8e-01	NA	NA
30	hsa-miR-29b	-3.01189584330326	9.3e-01	0.0053	Neg

Table B.3: Top 30 downregulated miRNAs in angiogenic subtype of Bentink dataset. d-value: delta value calculated by SAM. miRNA f-test: The p-value resulting from a test of differential variance between sample groups (F-test). RE test: p-value of the t-test comparing Pita-based RE scores between the angiogenic and non-angiogenic subtypes. RE diff: Indicates whether the miRNA had a negative or positive RE change. Pos: RE scores were higher in the angiogenic subtype. Neg: RE scores were lower in the angiogenic subtype.

	miRNA	d-value	f-test	RE test	RE diff
1	hsa.miR.514	-11.0744803824979	< 1e-16	5.4e-01	Pos
2	hsa.miR.509.3p	-10.1827702348673	< 1e-16	9.4e-03	Pos
3	hsa.miR.508.3p	-9.91581126305263	< 1e-16	5.1e-12	Neg
4	hsa.miR.507	-9.72807539865771	< 1e-16	1.9e-02	Neg
5	hsa.miR.509.3.5p	-9.69325234983162	< 1e-16	6.2e-03	Neg
6	hsa.miR.509.5p	-9.64652324502946	< 1e-16	6.2e-03	Neg
7	hsa.miR.506	-9.31159026633211	< 1e-16	5.0e-12	Neg
8	hsa.miR.510	-8.86317262804193	< 1e-16	5.5e-05	Neg
9	hsa.miR.219.5p	-7.36597760916123	7.4e-04	5.9e-03	Neg
10	hsa.miR.513c	-7.16165521980113	2.2e-16	7.4e-01	Pos
11	hsa.miR.20a.	-6.44558381098497	4.2e-02	NA	NA
12	hsa.miR.20b	-6.16587037961964	3.3e-01	2.9e-11	Neg
13	hsa.miR.362.3p	-6.14042674300895	1.3e-01	3.9e-01	Pos
14	hsa.miR.26b	-6.09000202374861	1e+00	7.1e-09	Neg
15	hsa.miR.30e.	-5.98461383236935	3.3e-01	NA	NA
16	hsa.miR.513a.5p	-5.9286787501011	7.7e-10	2.4e-01	Pos
17	hsa.miR.513b	-5.72384134997389	4.2e-08	2.6e-06	Neg
18	hsa.miR.20a	-5.68934314656094	9.4e-01	2.9e-11	Neg
19	hsa.miR.335	-5.63019838505151	4.7e-03	7.8e-02	Neg
20	hsa.miR.141.	-5.59807365643323	3.3e-01	NA	NA
21	hsa.miR.96	-5.54329186471013	3.5e-01	4.0e-11	Neg
22	hsa.miR.30c	-5.50124336578829	1.5e-01	1.5e-07	Neg
23	hsa.miR.25	-5.43089551434303	9.1e-02	1.6e-12	Neg
24	hsa.miR.148b	-5.29737312736283	3.1e-01	9.5e-04	Neg
25	hsa.miR.106b	-5.17071517266145	3.5e-01	2.9e-11	Neg
26	hsa.miR.660	-5.02755940993347	3.4e-01	2.5e-03	Neg
27	hsa.miR.19b.1.	-4.98238623573534	5.2e-03	NA	NA
28	hsa.miR.9.	-4.94842910115994	< 1e-16	NA	NA
29	hsa.miR.153	-4.93786612880719	3.1e-05	2.1e-01	Neg
30	hsa.miR.183	-4.87069679073173	1.3e-02	4.5e-03	Neg

Table B.4: Top 30 downregulated miRNAs in the angiogenic subtype of the TCGA dataset. d-value: delta value calculated by SAM. miRNA f-test: The p-value resulting from a test of differential variance between sample groups (F-test). RE test: p-value of the t-test comparing Pita-based RE scores between the angiogenic and non-angiogenic subtypes. RE diff: Indicates whether the miRNA had a negative or positive RE change. Pos: RE scores were higher in the angiogenic subtype. Neg: RE scores were lower in the angiogenic subtype. NA indicates that no target prediction data were available for that miRNA, and therefore no RE score could be calculated.

	Pita	Pictar	Targetscan	mirSVR
1	hsa-miR-639			

Table B.5: Positive RE-changing miRNAs calculated for each of the four miRNA target prediction tools, using the Bentink mRNA expression dataset, $p < 0.05$.

	Pita	Pictar	Targetscan	mirSVR
1	hsa-miR-574-5p	hsa-miR-34b		hsa-miR-18a*
2	hsa-miR-1207-5p	hsa-miR-198		hsa-miR-25*
3	hsa-miR-939	hsa-miR-369-5p		hsa-miR-93*
4	hsa-miR-663	hsa-miR-331		hsa-miR-129*
5	hsa-miR-671-5p	hsa-miR-133b		hsa-miR-181a*
6	hsa-miR-608	hsa-miR-133a		hsa-miR-122*
7	hsa-miR-1268	hsa-miR-34c		hsa-miR-129-3p
8	hsa-miR-1254	hsa-miR-195		hsa-miR-29c*
9	hsa-miR-885-3p	hsa-miR-152		hsa-miR-361-3p
10	hsa-miR-708	hsa-miR-215		hsa-miR-342-5p
11	hsa-miR-602	hsa-miR-107		hsa-miR-151-5p
12	hsa-miR-661	hsa-miR-103		hsa-miR-196b*
13	hsa-miR-1293	hsa-miR-16		hsa-miR-423-5p
14	hsa-miR-637	hsa-miR-15a		hsa-miR-423-3p
15	hsa-miR-423-5p	hsa-miR-217		hsa-miR-409-5p

Table B.6: Top 15 positive RE-changing miRNAs, TCGA dataset.

	Pita	Pictar	Targetscan	mirSVR
1	hsa-miR-675	hsa-miR-29c		hsa-miR-425*
2	hsa-miR-644	hsa-miR-29b		hsa-miR-602
3	hsa-miR-614	hsa-miR-29a		hsa-miR-612
4	hsa-miR-521	hsa-let-7i		hsa-miR-1538
5		hsa-miR-32		hsa-miR-4296
6		hsa-miR-92		hsa-miR-4319
7		hsa-miR-9		hsa-miR-4322
8		hsa-let-7b		hsa-miR-4260
9		hsa-let-7f		hsa-miR-4254
10		hsa-miR-129		hsa-miR-4265
11		hsa-miR-191		
12		hsa-miR-151		
13		hsa-miR-33		
14		hsa-miR-125b		
15		hsa-let-7g		

Table B.7: To 15 negative RE-changing miRNAs calculated for each of the four miRNA target prediction tools, using the Bentink mRNA expression dataset, $p < 0.05$.

	Pita	Pictar	Targetscan	mirSVR
1	hsa-miR-765	hsa-miR-299-3p		hsa-let-7a*
2	hsa-miR-339-5p	hsa-miR-124a		hsa-let-7a-2*
3	hsa-miR-638	hsa-miR-137		hsa-let-7b*
4	hsa-miR-1299	hsa-miR-324-3p		hsa-let-7c*
5	hsa-miR-663b	hsa-miR-22		hsa-let-7d*
6	hsa-miR-1825	hsa-miR-182-star		hsa-let-7e*
7	hsa-miR-1253	hsa-miR-326		hsa-let-7f-1*
8	hsa-miR-499-5p	hsa-miR-299-5p		hsa-let-7f-2*
9	hsa-miR-1236	hsa-miR-136		hsa-miR-16-1*
10	hsa-miR-767-5p	hsa-miR-218		hsa-miR-17*
11	hsa-miR-1226	hsa-miR-328		hsa-miR-19a*
12	hsa-miR-493	hsa-miR-30e-5p		hsa-miR-19b-1*
13	hsa-miR-616	hsa-miR-30d		hsa-miR-19b-2*
14	hsa-miR-612	hsa-miR-30c		hsa-miR-20a*
15	hsa-miR-103	hsa-miR-30b		hsa-miR-21*

Table B.8: Top 15 negative RE-changing miRNAs calculated for each of the four miRNA target prediction tools, using the TCGA mRNA expression dataset, $p < 0.05$.

	X1	X2	X3	X4	X5
1	NAT2	ALG3	IKZF1	GPNMB	PDLIM5
2	ADA	PSME3	RRAGB	BAIAP2	TRIM3
3	CDH2	MPHOSPH9	SIRPB1	CDX4	HEXIM1
4	MED6	ADAM10	EMC8	TACC3	STAMBP
5	ZBTB33	CALCRL	CDKN3	COG5	TGOLN2
6	HDAC5	USPL1	TFG	PFDN6	CENPE
7	CHAF1A	FLOT1	NPM3	SLC9A6	POLR3C
8	CDH6	SSX3	KLF2	EIF3M	POSTN
9	MICA	PRG4	CITED2	VTI1B	GAS2L1
10	SMC4	CD96	SEMA3A	CRTAP	EXOC5
11	DNM1L	PLIN3	TUBA1B	SYNCRIP	IGF2BP3
12	DPP3	MFSD10	TUBB3	VAT1	IGF2BP2
13	ATP9A	RASGRP2	SYCP2	SEMA6C	CAMKK2
14	TSFM	HNRNPR	SCML2	CEBPB	CELF1
15	TSPAN1	CDK8	ANAPC10	FBLN5	LBX1
16	CTDSP2	ABCC4	DLC1	IPO8	KLF1
17	KIF20A	FSTL3	NDRG1	NEBL	GNA13
18	HIPK3	NET1	NDC80	BATF	GNB5
19	ARL4A	UBE4B	BASP1	ARL6IP5	EBP
20	CELA3A	EIF1B	IFITM3	AGR2	CLDN16
21	TOB1	SF3A1	SPON2	AGPAT2	GMEB1
22	G3BP1	TRAIP	TESK2	SPTLC1	CCT6B
23	GPR64	APC2	ARIH2	SLC19A2	TBR1
24	MBNL2	PAK4	CDC42EP2	CXCL13	TCFL5
25	CEBPZ	MARCH6	OLFM1	MRPL28	PLK4
26	FARP1	ADAR	CDX1	CCT4	STAG2
27	LPCAT3	ZNF267	TIMM17A	SORBS1	PHTF1
28		CDKN2C	N4BP2L2	ERN2	MAP3K2
29	DHRS9	DSCR3	MCRS1	AHSA1	CTSC
30	TNK2	CDKN2D	LRRN2	CDC42EP3	NUP50

Table B.9: 150 of the 2508 genes with expression patterns significantly anticorrelated with the 506-514 cluster of miRNAs in the Bentink dataset.

	X1	X2	X3	X4	X5
1	CDH2	ACTR3	CDK2	CDK8	TCIRG1
2	AKT3	ACTR2	DHRS9	MRPS31	TNIP1
3	GNPDA1	TSPAN3	CDK3	CALCOCO2	IKZF1
4	TANK	NUBP2	RBM6	GPHN	CRISP3
5	HDAC6	TSPAN1	LHFP	RABEPK	SMYD5
6	BCL2L11	CTDSP2	TNK2	SLC17A2	KLHL41
7	GNE	SGK2	CDK4	HRSP12	RRAGB
8	TROAP	RAD50	RNF41	CDK9	SIRPB1
9	MED16	KIF20A	MPHOSPH9	SRRM1	AKR1A1
10	CHAF1A	PREB	CALCRL	CNKSR1	EMC8
11	TOM1L1	ACTR1A	NUTF2	CDKN1A	CDKN3
12	TOM1	ARL4C	MPZL2	SF3B4	CNPY2
13	SH2D3A	ARL4A	USPL1	CDK2AP2	TLR6
14	DNAJB6	DNAL4	EIF1	IRX5	TFG
15	OCN	FRY	FLOT1	ZMPSTE24	TRIM22
16	SMC4	TRAP1	DDX39A	FSTL3	HMG20B
17	SAE1	BCAP31	PSMD14	NET1	KLF2
18	ABCB6	CELA3A	SSX3	PRSS16	CACNG3
19	DNM1L	RBM12	OLIG2	CDKN1C	CACNG2
20	CDH8	ARFRP1	PRG4	RGS19	CITED2
21	NR1H3	TOB1	CTDSPL	LILRB2	TUBA1B
22	COX17	C4orf6	CDK7	EIF1B	CDR1
23	DPP3	FAM13A	GPA33	DNAJA2	TUBB4A
24	TSPAN32	G3BP1	CD96	APC2	TUBB4B
25	CDH10	SUGP2	MFSD10	PAK4	BTN2A2
26	USH1C	ABI2	STX6	MARCH6	CORO2B
27	EDIL3	CEBPZ	RCAN2	KATNB1	ANAPC10
28	UST	FARP1	LRRC17	SNAPC5	DLC1
29	ARPC5	LPCAT3	SLC35B1	APBB3	NDRG1
30	ARPC3	WASF2	AP3S2	CCNO	CDS1

Table B.10: 150 of the 4082 genes with expression patterns significantly anticorrelated with the 506-514 cluster of miRNAs in the TCGA dataset.

Annotation Cluster 1 Enrichment Score: 18.032855642537324									
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Benjamini
UP_SEQ_FEATURE	disulfide bond	166	30.57	1.35e-21	542	1546	10361	2.053	2.30e-18
SP_PIR_KEYWORDS	disulfide bond	171	31.49	2.01e-21	543	1621	10365	2.014	2.46e-18
UP_SEQ_FEATURE	signal peptide	188	34.62	4.27e-20	542	1924	10361	1.868	8.82e-19
SP_PIR_KEYWORDS	signal	188	34.62	5.14e-20	543	1924	10365	1.865	9.82e-19
SP_PIR_KEYWORDS	glycoprotein	206	37.94	7.12e-16	543	2375	10365	1.656	7.38e-17
UP_SEQ_FEATURE	glycosylation site-N-linked (GlcNAc...)	189	34.81	1.49e-13	542	2220	10361	1.627	1.26e-17
Annotation Cluster 2 Enrichment Score: 16.20844730095251									
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Benjamini
UP_SEQ_FEATURE	signal peptide	188	34.62	4.27e-20	542	1924	10361	1.868	8.82e-17
SP_PIR_KEYWORDS	signal	188	34.62	5.14e-20	543	1924	10365	1.865	9.82e-17
GOTERM_CC_FAT	GO:0044421~extracellular region part	98	18.05	4.20e-18	434	693	7605	2.478	1.28e-17
GOTERM_CC_FAT	GO:0005576~extracellular region	132	24.31	7.07e-17	434	1148	7605	2.015	1.47e-15
SP_PIR_KEYWORDS	Secreted	111	20.44	5.22e-16	543	958	10365	2.212	5.74e-15
GOTERM_CC_FAT	GO:0005615~extracellular space	67	12.34	1.65e-10	434	515	7605	2.280	1.93e-14
Annotation Cluster 3 Enrichment Score: 11.615177057667342									
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Benjamini
GOTERM_BP_FAT	GO:0009611~response to wounding	64	11.79	3.58e-13	463	432	8423	2.695	4.97e-10
GOTERM_BP_FAT	GO:0006952~defense response	63	11.60	1.95e-12	463	437	8423	2.623	1.81e-09
GOTERM_BP_FAT	GO:0006954~inflammatory response	45	8.29	2.04e-11	463	264	8423	3.101	3.49e-09
Annotation Cluster 4 Enrichment Score: 7.840774441472784									
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Benjamini
SP_PIR_KEYWORDS	extracellular matrix	34	6.26	2.83e-11	543	168	10365	3.863	1.38e-08
GOTERM_CC_FAT	GO:0031012~extracellular matrix	42	7.73	1.34e-10	434	240	7605	3.067	2.76e-09
GOTERM_CC_FAT	GO:0005578~proteinaceous extracellular matrix	40	7.37	1.89e-10	434	223	7605	3.143	1.56e-08
GOTERM_MF_FAT	GO:0005201~extracellular matrix structural constituent	15	2.76	7.59e-06	437	65	8019	4.235	1.32e-08
GOTERM_CC_FAT	GO:0044420~extracellular matrix part	16	2.95	1.15e-04	434	88	7605	3.186	5.08e-03
Annotation Cluster 5 Enrichment Score: 5.96045725765667									
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Benjamini
GOTERM_BP_FAT	GO:0001775~cell activation	38	7.00	1.29e-09	463	225	8423	3.072	3.99e-03
GOTERM_BP_FAT	GO:0042110~T cell activation	24	4.42	1.48e-09	463	97	8423	4.501	2.76e-06
GOTERM_BP_FAT	GO:0045321~leukocyte activation	33	6.08	9.68e-09	463	190	8423	3.160	6.86e-07
GOTERM_BP_FAT	GO:0046649~hemopoietic or lymphoid organ development	29	5.34	1.11e-08	463	152	8423	3.471	3.36e-06
GOTERM_BP_FAT	GO:0048534~hemopoietic or lymphoid organ development	30	5.52	8.20e-06	463	219	8423	2.492	3.42e-06
GOTERM_BP_FAT	GO:0002520~immune system development	31	5.71	1.01e-05	463	233	8423	2.420	7.10e-04
GOTERM_BP_FAT	GO:0030217~T cell differentiation	13	2.39	1.44e-05	463	51	8423	4.637	1.80e-02
GOTERM_BP_FAT	GO:0030097~hemopoiesis	26	4.79	7.69e-05	463	198	8423	2.389	7.78e-04
GOTERM_BP_FAT	GO:0002521~leukocyte differentiation	17	3.13	2.20e-04	463	107	8423	2.890	1.92e-01
GOTERM_BP_FAT	GO:0030098~lymphocyte differentiation	14	2.58	6.02e-04	463	84	8423	3.032	9.19e-03

Figure B.1: The top significantly enriched functional annotations from the 506-514 anti-correlated genes in the Bentink dataset. Many more terms were enriched, but too many were available to display here.

Annotation Cluster 1										
Enrichment Score: 35.169782664245865										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Benjamini	FDR
UP_SEQ_FEATURE	disulfide bond	410	28.69	7.51e-43	1423	1879	12139	1.861	2.87e-39	1.39e-39
SP_PIR_KEYWORDS	disulfide bond	420	29.39	1.61e-41	1424	1967	12139	1.821	1.06e-38	2.41e-38
UP_SEQ_FEATURE	signal peptide	449	31.42	5.77e-37	1423	2248	12139	1.704	2.20e-33	1.07e-33
SP_PIR_KEYWORDS	signal	449	31.42	6.18e-37	1424	2248	12145	1.703	4.06e-34	9.25e-34
SP_PIR_KEYWORDS	glycoprotein	514	35.97	6.03e-30	1424	2868	12145	1.529	3.95e-27	9.01e-27
UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	483	33.80	3.68e-27	1423	2709	12139	1.521	4.68e-24	6.81e-24
Annotation Cluster 2										
Enrichment Score: 23.635210669213503										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Benjamini	FDR
UP_SEQ_FEATURE	signal peptide	449	31.42	5.77e-37	1423	2248	12139	1.704	2.20e-33	1.07e-33
SP_PIR_KEYWORDS	signal	449	31.42	6.18e-37	1424	2248	12145	1.703	4.06e-34	9.25e-34
GOTERM_CC_FAT	GO:0005576-extracellular region	295	20.64	3.31e-22	1155	1363	8856	1.660	1.76e-19	4.81e-19
SP_PIR_KEYWORDS	Secreted	241	16.86	4.77e-21	1424	1155	12145	1.780	6.26e-19	7.13e-18
GOTERM_CC_FAT	GO:0044421-extracellular region part	192	13.44	5.22e-19	1155	800	8856	1.840	1.39e-16	7.58e-16
GOTERM_CC_FAT	GO:0005615-extracellular space	131	9.17	5.25e-10	1155	596	8856	1.685	2.80e-07	7.63e-07
Annotation Cluster 3										
Enrichment Score: 21.831966369785977										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Benjamini	FDR
GOTERM_BP_FAT	GO:0006952-defense response	149	10.43	1.81e-24	1218	518	9805	2.316	7.48e-21	3.38e-21
GOTERM_BP_FAT	GO:0009611-response to wounding	136	9.52	1.11e-22	1218	469	9805	2.334	4.59e-19	2.07e-19
GOTERM_BP_FAT	GO:0006954-inflammatory response	96	6.72	1.59e-20	1218	289	9805	2.674	6.56e-17	2.96e-17
Annotation Cluster 4										
Enrichment Score: 13.635701379948442										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Benjamini	FDR
SP_PIR_KEYWORDS	extracellular matrix	66	4.62	2.47e-17	1424	184	12145	3.059	1.62e-14	3.69e-14
GOTERM_CC_FAT	GO:0031012-extracellular matrix	84	5.88	4.26e-15	1155	267	8856	2.412	2.25e-12	6.13e-12
GOTERM_CC_FAT	GO:0005578-proteinaceous extracellular matrix	79	5.53	5.31e-15	1155	244	8856	2.483	2.84e-12	7.74e-12
GOTERM_CC_FAT	GO:0044420-extracellular matrix part	37	2.59	5.13e-10	1155	94	8856	3.018	2.74e-07	7.46e-07
Annotation Cluster 5										
Enrichment Score: 12.562813777850412										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Benjamini	FDR
GOTERM_BP_FAT	GO:0007155-cell adhesion	141	9.87	1.31e-15	1218	584	9805	1.944	5.50e-12	2.49e-12
GOTERM_BP_FAT	GO:0022610-biological adhesion	141	9.87	1.31e-15	1218	584	9805	1.944	5.50e-12	2.49e-12
SP_PIR_KEYWORDS	cell adhesion	78	5.46	1.19e-08	1424	345	12145	1.928	7.79e-06	1.78e-05
Annotation Cluster 6										
Enrichment Score: 11.727423997928085										
Category	Term	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	Benjamini	FDR
SP_PIR_KEYWORDS	glycoprotein	514	35.97	6.03e-30	1424	2868	12145	1.529	3.95e-27	9.01e-27
UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	483	33.80	3.68e-27	1423	2709	12139	1.521	4.68e-24	6.81e-24
UP_SEQ_FEATURE	topological domain:Extracellular	314	21.97	7.09e-21	1423	1652	12139	1.621	2.70e-17	1.31e-17
GOTERM_CC_FAT	GO:0005886-plasma membrane	477	33.38	2.37e-17	1155	2700	8856	1.355	1.26e-14	3.44e-14
UP_SEQ_FEATURE	topological domain:Cytoplasmic	348	24.35	4.57e-12	1423	2143	12139	1.385	1.74e-08	8.44e-09
SP_PIR_KEYWORDS	receptor	181	12.67	8.06e-12	1424	948	12145	1.628	5.29e-09	1.20e-08
GOTERM_CC_FAT	GO:0044459-plasma membrane part	327	22.88	3.90e-10	1155	1870	8856	1.341	2.08e-07	5.68e-07
GOTERM_CC_FAT	GO:0005887-integral to plasma membrane	209	14.63	1.64e-09	1155	1097	8856	1.461	8.77e-07	2.39e-06
GOTERM_CC_FAT	GO:0031226-intrinsic to plasma membrane	211	14.77	2.11e-09	1155	1113	8856	1.454	1.12e-06	3.06e-06
SP_PIR_KEYWORDS	membrane	598	39.05	3.03e-07	1424	4031	12145	1.181	1.99e-04	4.53e-04
SP_PIR_KEYWORDS	transmembrane	422	29.53	3.32e-06	1424	2991	12145	1.203	2.18e-03	4.96e-03
UP_SEQ_FEATURE	transmembrane region	419	29.32	3.67e-06	1423	2971	12139	1.203	1.39e-02	6.78e-03
GOTERM_CC_FAT	GO:0031224-intrinsic to membrane	481	33.66	2.55e-03	1155	3350	8856	1.101	7.43e-01	3.64e+00
GOTERM_CC_FAT	GO:0016021-integral to membrane	463	32.40	3.76e-03	1155	3227	8856	1.100	8.66e-01	5.33e+00

Figure B.2: The top significantly enriched functional annotations from the 506-514 anti-correlated genes in the TCGA dataset. Many more terms were enriched, but too many were available to display here.

	term
GO:0060965	negative regulation of gene silencing by miRNA
GO:0010586	miRNA metabolic process
GO:0035280	miRNA loading onto RISC involved in gene silencing by miRNA
GO:0010587	miRNA catabolic process
GO:0035196	production of miRNAs involved in gene silencing by miRNA
GO:0031053	primary microRNA processing
GO:0035279	mRNA cleavage involved in gene silencing by miRNA
GO:0035195	gene silencing by miRNA
GO:0035278	negative regulation of translation involved in gene silencing by miRNA
GO:0031054	pre-microRNA processing
GO:0060964	regulation of gene silencing by miRNA
GO:0035281	pre-microRNA export from nucleus

Table B.11: GO terms used for selection of genes associated with miRNA processing.

	MATRIX_ID	Z_SCORE	P_VALUE	SAMP_AVE	BG_AVE	SAMPLE_DEVSTD	SAMP_SIZE	adjpvalue
SP1	MA0079.2	10.1246	1.20e-26	0.932206	0.921702	0.0518854	3236	1.56e-24
	Egr1	MA0162.1	8.24902	0.862186	0.854058	0.0514077	3236	1.16e-16
Klf4	MA0039.2	7.43504	2.90e-15	0.914826	0.90572	0.061981	3236	1.25e-13
	TFAP2A	MA0003.1	7.0375	0.952133	0.945971	0.0441295	3236	2.43e-12
PLAG1	MA0163.1	6.49935	3.83e-12	0.808432	0.802299	0.0490257	3236	9.95e-11
	MZF1_1-4	MA0056.1	5.78779	0.974655	0.970387	0.0363925	3236	1.25e-08
MZF1_5-13	MA0057.1	5.74021	6.72e-10	0.914918	0.910872	0.0375354	3236	1.25e-08
	INSM1	MA0155.1	4.92981	0.861603	0.857782	0.0410032	3236	1.60e-06
E2F1	MA0024.1	4.81814	1.83e-07	0.835719	0.830619	0.0561422	3236	2.64e-06
	MA0259.1	4.59662	6.25e-07	0.897335	0.892276	0.0578951	3236	8.12e-06
Zfx	MA0146.1	4.43507	1.52e-06	0.86328	0.858064	0.0607455	3236	1.74e-05
	Arnt::Ahr	MA0006.1	4.41265	0.919877	0.914593	0.0633826	3236	1.74e-05
Pax5	MA0014.1	4.33335	2.50e-06	0.838593	0.836212	0.0285935	3236	2.50e-05
	RREB1	MA0073.1	4.18444	0.799964	0.796261	0.0471448	3236	4.60e-05
MIZF	MA0131.1	4.13634	6.31e-06	0.813958	0.810462	0.0448559	3236	5.47e-05
	Mycn	MA0104.2	3.96498	0.842807	0.838828	0.0532902	3236	1.15e-04
NFKB1	MA0105.1	3.66545	5.63e-05	0.817214	0.813419	0.0539528	3236	4.31e-04
	Mafb	MA0117.1	3.58483	0.924411	0.922029	0.0352002	3236	5.40e-04
EWSR1-FLI1	MA0149.1	3.57993	7.89e-05	0.692644	0.688983	0.0541787	3236	5.40e-04
	GABPA	MA0062.2	3.48809	0.846908	0.843222	0.0558217	3236	7.57e-04
Myc	MA0147.1	3.18902	3.83e-04	0.844684	0.841806	0.0476602	3236	2.37e-03
	CTCF	MA0139.1	2.95016	0.78242	0.780042	0.0429906	3236	5.41e-03
NHLH1	MA0048.1	2.65173	2.58e-03	0.8261	0.82363	0.0491447	3236	1.46e-02
	NF-kappaB	MA0061.1	2.41104	0.844446	0.842125	0.0515092	3236	2.84e-02
Tcfcp2l1	MA0145.1	2.41229	5.46e-03	0.838507	0.836732	0.038893	3236	2.84e-02
	Arnt	MA0004.1	2.38194	0.865155	0.862142	0.0684786	3236	2.93e-02

Table B.12: Transcription factors with binding sites enriched in the anticorrelated genes to the 506-514 cluster of miRNAs. Bentink dataset.

	MATRIX_ID	Z_SCORE	P_VALUE	SAMP_AVE	BG_AVE	SAMPLE_DEVSTD	SAMP_SIZE	adjpvalue
SP1	MA0079.2	10.3131	5.55e-27	0.933989	0.921702	0.0520694	2397	7.21e-25
	MA0039.2	9.68396	5.57e-24	0.919312	0.90572	0.0606135	2397	3.62e-22
Egr1	MA0162.1	8.19494	9.29e-18	0.863297	0.854058	0.0497597	2397	4.02e-16
	MA0003.1	6.63735	3.16e-12	0.952632	0.945971	0.0430928	2397	1.03e-10
E2F1	MA0024.1	5.81603	6.52e-10	0.837688	0.830619	0.0572874	2397	1.70e-08
	MA0163.1	5.5478	4.23e-09	0.808294	0.802299	0.0480448	2397	9.16e-08
MIZF	MA0131.1	5.44133	6.74e-09	0.815745	0.810462	0.046108	2397	1.25e-07
	MA0146.1	5.23925	2.55e-08	0.865137	0.858064	0.0612571	2397	4.04e-07
MZF1_5-13	MA0057.1	5.21587	2.80e-08	0.915084	0.910872	0.0372886	2397	4.04e-07
	MA0014.1	5.17108	3.61e-08	0.839476	0.836212	0.0292015	2397	4.70e-07
MZF1_1-4	MA0056.1	5.00967	1.10e-07	0.974626	0.970387	0.0358833	2397	1.29e-06
	MA0104.2	4.77419	3.28e-07	0.844321	0.838828	0.0534569	2397	3.34e-06
HIF1A::ARNT	MA0259.1	4.77362	3.34e-07	0.898303	0.892276	0.0582675	2397	3.34e-06
	MA0105.1	4.73234	4.21e-07	0.819044	0.813419	0.0545544	2397	3.91e-06
INSM1	MA0155.1	4.61387	7.57e-07	0.861886	0.857782	0.0415867	2397	6.56e-06
	MA0147.1	4.32312	3.30e-06	0.846285	0.841806	0.0484569	2397	2.68e-05
RREB1	MA0073.1	4.11286	9.21e-06	0.800431	0.796261	0.0469013	2397	7.04e-05
	MA0149.1	3.90325	2.34e-05	0.693567	0.688983	0.0553221	2397	1.69e-04
GABPA	MA0062.2	3.8457	3.12e-05	0.84788	0.843222	0.0557673	2397	2.14e-04
	MA0004.1	3.14695	5.01e-04	0.866712	0.862142	0.0700971	2397	3.26e-03
ELK4	MA0076.1	2.83894	1.54e-03	0.827961	0.824416	0.0585784	2397	9.17e-03
	MA0061.1	2.83689	1.55e-03	0.845256	0.842125	0.0516411	2397	9.17e-03
Tcfcp2l1	MA0145.1	2.51587	4.39e-03	0.838855	0.836732	0.039138	2397	2.48e-02

ELK1	MA0028.1	2.34985	7.18e-03	0.899308	0.897219	0.041279	2397	3.89e-02
Zfp423	MA0116.1	2.27971	8.87e-03	0.828998	0.826905	0.0418209	2397	4.54e-02
Arnt::Ahr	MA0006.1	2.26556	9.09e-03	0.917707	0.914593	0.0643648	2397	4.54e-02
NHLH1	MA0048.1	2.2416	9.96e-03	0.826018	0.82363	0.0475723	2397	4.80e-02

Table B.13: Transcription factors with binding sites enriched in the anticorrelated genes to the 506-514 cluster of miRNAs. TCGA dataset.

	Gene Symbol	Fold-change
1	ZCCHC11	0.828797785484468
2	TARBP2	0.859199121208351
3	DGCR8	0.893944194553101
4	SRRT	0.903452443991202
5	LIMD1	0.885246870396264
6	SMAD3	0.929470804122887
7	SRRT	0.917074609017892

Table B.14: miRNA-processing genes downregulated in the angiogenic subtype of the TCGA dataset. Significance Analysis of Microarrays (SAM) with an FDR of 0.05.

	Gene Symbol	Fold-change
1	MZF1	0.847375700753309
2	MZF1	0.86856610168371
3	ELK1	0.909555823477458
4	ELK1	0.912496599918915
5	CTCF	0.93190084059775
6	MAFB	1.65743194757325
7	EGR1	1.42766591591708
8	EGR1	1.41325065192716
9	HIF1A	1.11561212885967
10	NFKB1	1.11835865833132

Table B.15: Transcription factors differentially expressed in the angiogenic subtype of the TCGA dataset. Significance Analysis of Microarrays (SAM) with an FDR of 0.05.

Appendix C

Chapter 4 Additional Tables

	p-value	coefficient	ci
hsa-miR-373*	1.89e-02	-6.54e-01	0.63
hsa-miR-609	4.51e-02	-5.63e-01	0.64
hsa-miR-216	3.73e-02	-4.07e-01	0.63
hsa-miR-184	4.44e-02	-1.80e-01	0.64
hsa-miR-190	3.39e-02	-1.40e-01	0.64
hsa-miR-548b	5.40e-03	-1.24e-01	0.64
hsa-miR-373	6.17e-03	-1.08e-01	0.63
hsa-miR-613	1.26e-02	-9.65e-02	0.64
hsa-miR-27a	4.78e-02	-9.26e-02	0.63
hsa-miR-147	1.92e-02	-7.86e-02	0.63
hsa-miR-563	4.05e-02	-6.73e-02	0.63
hsa-miR-602	1.83e-03	-6.71e-02	0.63
hsa-miR-330	3.89e-02	-6.37e-02	0.62
hsa-miR-517c	1.57e-02	-6.31e-02	0.63
hsa-miR-565	3.10e-02	-6.05e-02	0.63
hsa-let-7g	4.76e-03	-5.60e-02	0.64
hsa-miR-144	3.90e-03	-5.33e-02	0.63
hsa-miR-516-3p	2.62e-03	-4.13e-02	0.64
hsa-miR-554	9.22e-03	-3.96e-02	0.63
hsa-miR-638	1.06e-02	-3.78e-02	0.64
hsa-miR-422b	2.62e-02	-3.57e-02	0.64
hsa-miR-675	3.76e-02	-3.12e-02	0.63
hsa-miR-213	1.49e-02	-2.94e-02	0.63

hsa-miR-492	3.63e-04	-2.91e-02	0.64
hsa-miR-605	1.01e-03	-2.84e-02	0.63
hsa-miR-519d	1.13e-02	-2.83e-02	0.64
hsa-miR-522	3.05e-02	-2.28e-02	0.62
hsa-miR-576	6.29e-03	-1.86e-02	0.63
hsa-miR-551a	7.81e-03	-1.85e-02	0.63
hsa-miR-128a	5.08e-03	-1.84e-02	0.63
hsa-miR-614	3.32e-03	-1.81e-02	0.64
hsa-miR-202	1.82e-02	-1.69e-02	0.64
hsa-miR-624	2.00e-03	-1.58e-02	0.63
hsa-miR-501	3.66e-03	-1.53e-02	0.64
hsa-miR-196b	6.19e-03	-1.13e-02	0.64
hsa-miR-192	8.02e-04	-8.41e-03	0.64
hsa-miR-219	1.22e-04	-5.49e-03	0.64
hsa-miR-130b	3.62e-04	-4.34e-03	0.64
hsa-miR-140	5.31e-05	-1.89e-03	0.64
hsa-miR-378	3.46e-05	-1.21e-03	0.64
hsa-miR-635	2.13e-06	-1.13e-03	0.64
hsa-miR-93	1.00e-05	1.63e-03	0.64
hsa-miR-218	1.23e-03	6.93e-03	0.64
hsa-miR-560	9.81e-04	7.51e-03	0.63
hsa-miR-486	2.59e-03	7.68e-03	0.64
hsa-miR-189	1.38e-03	8.03e-03	0.64
hsa-miR-768-5p	2.11e-03	9.85e-03	0.64
hsa-miR-19b	1.66e-03	1.18e-02	0.64
hsa-miR-512-5p	4.33e-03	1.42e-02	0.63
hsa-miR-223	7.61e-03	2.00e-02	0.64
hsa-miR-338	1.30e-02	2.13e-02	0.63
hsa-miR-29a	3.14e-03	2.21e-02	0.63
hsa-miR-517b	1.05e-02	2.38e-02	0.64
hsa-miR-204	1.87e-02	2.53e-02	0.63
hsa-miR-594	8.08e-03	2.53e-02	0.63
hsa-miR-625	1.35e-02	2.65e-02	0.63
hsa-miR-34b	4.44e-02	2.75e-02	0.62
hsa-miR-598	1.24e-02	2.78e-02	0.64
hsa-miR-595	5.14e-04	3.02e-02	0.63
hsa-miR-19a	3.24e-02	3.09e-02	0.63
hsa-miR-671	1.27e-02	3.30e-02	0.63
hsa-miR-182*	4.74e-02	3.74e-02	0.63
hsa-miR-661	6.23e-03	3.78e-02	0.63
hsa-miR-96	1.31e-02	3.90e-02	0.63

hsa-miR-106b	7.75e-03	4.03e-02	0.64
hsa-miR-132	2.86e-02	4.04e-02	0.63
hsa-let-7c	1.27e-02	4.22e-02	0.63
hsa-miR-652	2.56e-02	4.27e-02	0.64
hsa-miR-99a	3.74e-02	4.32e-02	0.64
hsa-miR-23b	3.98e-03	4.45e-02	0.63
hsa-miR-484	6.18e-03	4.63e-02	0.64
hsa-miR-22	2.55e-02	4.79e-02	0.64
hsa-miR-200a	1.88e-02	4.82e-02	0.64
hsa-miR-345	4.38e-02	5.47e-02	0.63
hsa-miR-494	2.44e-02	5.55e-02	0.64
hsa-miR-518b	2.10e-02	6.11e-02	0.63
hsa-miR-616	4.75e-03	6.15e-02	0.64
hsa-miR-18b	3.56e-02	6.31e-02	0.64
hsa-miR-520b	3.11e-03	6.36e-02	0.63
hsa-miR-610	3.71e-02	6.87e-02	0.63
hsa-miR-25	2.11e-02	7.60e-02	0.63
hsa-miR-141	3.45e-02	7.88e-02	0.64
hsa-miR-525*	3.17e-03	8.10e-02	0.64
hsa-miR-767-5p	4.15e-02	8.51e-02	0.63
hsa-miR-587	3.16e-03	8.91e-02	0.63
hsa-miR-520g	5.31e-03	1.12e-01	0.63
hsa-miR-644	2.83e-02	1.17e-01	0.63
hsa-miR-662	4.39e-02	1.22e-01	0.64
hsa-miR-197	4.50e-02	1.33e-01	0.64
hsa-miR-518c	3.10e-02	1.53e-01	0.63
hsa-miR-589	9.12e-03	1.56e-01	0.63
hsa-miR-580	1.17e-02	2.60e-01	0.63
hsa-miR-520f	1.16e-02	2.77e-01	0.63
hsa-miR-557	3.10e-02	3.13e-01	0.64
hsa-miR-593	4.05e-02	3.36e-01	0.63
hsa-miR-562	1.63e-02	6.25e-01	0.63
hsa-miR-516-5p	4.02e-02	8.35e-01	0.63
hsa-miR-367	4.94e-02	1.25e+00	0.63

Table C.1: miRNAs with expression values most predictive of survival using a univariate model. Models calculated using expression data from the Bentink dataset, and based on the p-value calculated with a likelihood ratio test. Models for 98 miRNAs had a calculated p-value of $p < 0.052$. ci: Concordance index of the model.

	p-value	coefficient	ci
RE-hsa-miR-525-5p	3.04e-02	-4.01e-04	0.64
RE-hsa-miR-520a-5p	3.04e-02	-4.01e-04	0.64
RE-hsa-miR-592	4.52e-02	-3.87e-04	0.63
RE-hsa-miR-654-5p	4.18e-02	-3.86e-04	0.64
RE-hsa-miR-708	4.91e-02	-3.83e-04	0.62
RE-hsa-miR-324-5p	4.88e-02	-3.73e-04	0.63
RE-hsa-miR-646	4.13e-02	-3.71e-04	0.64
RE-hsa-miR-28-5p	4.63e-02	-3.69e-04	0.63
RE-hsa-miR-541	3.82e-02	-3.65e-04	0.63
RE-hsa-miR-34c-3p	4.12e-02	-3.27e-04	0.63
RE-hsa-miR-657	4.98e-02	-3.22e-04	0.64
RE-hsa-miR-365	2.07e-02	-2.86e-04	0.64
RE-hsa-miR-1299	3.29e-02	-2.76e-04	0.64
RE-hsa-miR-453	2.77e-02	-2.74e-04	0.64
RE-hsa-miR-616	4.19e-02	-2.70e-04	0.64
RE-hsa-miR-554	4.53e-02	-2.64e-04	0.64
RE-hsa-miR-636	2.67e-02	-2.54e-04	0.64
RE-hsa-miR-329	1.31e-02	-2.49e-04	0.63
RE-hsa-miR-362-3p	1.31e-02	-2.49e-04	0.63
RE-hsa-miR-335	4.63e-02	-2.49e-04	0.63
RE-hsa-miR-127-5p	2.71e-02	-2.34e-04	0.64
RE-hsa-miR-28-3p	4.54e-02	-2.10e-04	0.63
RE-hsa-miR-1245	4.06e-02	-2.03e-04	0.64
RE-hsa-miR-1208	1.65e-02	-2.01e-04	0.64
RE-hsa-miR-595	3.23e-02	-2.00e-04	0.64
RE-hsa-miR-433	2.33e-02	-1.95e-04	0.64
RE-hsa-miR-532-5p	2.34e-02	-1.91e-04	0.64
RE-hsa-miR-328	1.05e-02	-1.89e-04	0.63
RE-hsa-miR-556-3p	2.87e-02	-1.83e-04	0.64
RE-hsa-miR-188-5p	9.77e-03	-1.71e-04	0.64
RE-hsa-miR-149	1.22e-02	-1.70e-04	0.63
RE-hsa-miR-590-5p	1.94e-02	-1.65e-04	0.64
RE-hsa-miR-1185	1.67e-02	-1.50e-04	0.64
RE-hsa-miR-412	1.23e-02	-1.48e-04	0.64
RE-hsa-miR-610	1.84e-02	-1.48e-04	0.63
RE-hsa-miR-573	1.13e-02	-1.36e-04	0.64
RE-hsa-miR-639	2.64e-02	-1.36e-04	0.63
RE-hsa-miR-342-5p	1.38e-02	-1.28e-04	0.64
RE-hsa-miR-549	9.98e-03	-1.25e-04	0.64

RE-hsa-miR-562	1.25e-02	-1.21e-04	0.64
RE-hsa-miR-518e	1.37e-02	-1.19e-04	0.64
RE-hsa-miR-640	7.54e-03	-1.16e-04	0.64
RE-hsa-miR-1227	6.40e-03	-1.10e-04	0.64
RE-hsa-miR-21	6.97e-03	-9.99e-05	0.63
RE-hsa-miR-147b	2.09e-02	-9.96e-05	0.63
RE-hsa-miR-548m	5.25e-03	-8.82e-05	0.64
RE-hsa-miR-663b	4.70e-03	-8.79e-05	0.64
RE-hsa-miR-664	3.62e-03	-8.38e-05	0.64
RE-hsa-miR-142-3p	4.23e-03	-8.30e-05	0.63
RE-hsa-miR-208a	7.11e-03	-8.10e-05	0.64
RE-hsa-miR-208b	7.11e-03	-8.10e-05	0.64
RE-hsa-miR-491-3p	3.93e-03	-7.70e-05	0.64
RE-hsa-miR-379	2.75e-03	-5.02e-05	0.64
RE-hsa-miR-518a-3p	1.84e-03	-4.83e-05	0.64
RE-hsa-miR-518d-3p	1.84e-03	-4.83e-05	0.64
RE-hsa-miR-518c	1.84e-03	-4.83e-05	0.64
RE-hsa-miR-518f	1.84e-03	-4.83e-05	0.64
RE-hsa-miR-518b	1.84e-03	-4.83e-05	0.64
RE-hsa-miR-222	1.47e-03	-4.64e-05	0.64
RE-hsa-miR-937	1.42e-02	-3.65e-05	0.63
RE-hsa-miR-126	8.47e-03	-3.35e-05	0.63
RE-hsa-miR-221	5.60e-04	-2.87e-05	0.64
RE-hsa-miR-886-5p	3.38e-03	-2.53e-05	0.64
RE-hsa-miR-720	2.50e-04	-1.40e-05	0.64
RE-hsa-miR-1254	5.45e-05	-1.40e-05	0.64
RE-hsa-miR-671-3p	3.39e-05	-3.63e-06	0.64
RE-hsa-miR-1281	1.24e-05	-3.14e-06	0.64
RE-hsa-miR-663	5.24e-04	1.79e-05	0.64
RE-hsa-miR-885-5p	2.94e-04	2.00e-05	0.64
RE-hsa-miR-1296	3.30e-04	2.06e-05	0.63
RE-hsa-miR-598	1.27e-03	4.36e-05	0.63
RE-hsa-miR-451	2.30e-03	4.89e-05	0.64
RE-hsa-miR-302f	7.59e-04	5.20e-05	0.63
RE-hsa-miR-206	9.48e-04	5.63e-05	0.63
RE-hsa-miR-1	9.48e-04	5.63e-05	0.63
RE-hsa-miR-582-3p	2.43e-03	6.64e-05	0.63
RE-hsa-miR-220c	7.84e-04	6.71e-05	0.63
RE-hsa-miR-1304	3.48e-03	7.35e-05	0.64
RE-hsa-miR-299-5p	2.33e-03	7.64e-05	0.63
RE-hsa-miR-496	5.51e-03	8.20e-05	0.63

RE-hsa-miR-642	3.43e-03	8.68e-05	0.63
RE-hsa-miR-615-3p	2.85e-02	9.53e-05	0.63
RE-hsa-miR-409-5p	5.46e-03	9.61e-05	0.63
RE-hsa-miR-645	4.75e-03	1.01e-04	0.64
RE-hsa-miR-614	3.08e-02	1.04e-04	0.63
RE-hsa-miR-875-5p	1.16e-02	1.05e-04	0.63
RE-hsa-miR-886-3p	2.19e-02	1.06e-04	0.63
RE-hsa-miR-1233	4.31e-03	1.20e-04	0.63
RE-hsa-miR-939	8.28e-03	1.20e-04	0.64
RE-hsa-miR-768-5p	8.33e-03	1.34e-04	0.63
RE-hsa-miR-921	1.33e-02	1.34e-04	0.64
RE-hsa-miR-524-3p	1.73e-02	1.39e-04	0.63
RE-hsa-miR-525-3p	1.73e-02	1.39e-04	0.63
RE-hsa-miR-613	9.77e-03	1.81e-04	0.63
RE-hsa-miR-522	1.38e-02	1.82e-04	0.63
RE-hsa-miR-100	3.86e-02	1.92e-04	0.63
RE-hsa-miR-99a	3.86e-02	1.92e-04	0.63
RE-hsa-miR-99b	3.86e-02	1.92e-04	0.63
RE-hsa-miR-1288	2.55e-02	2.24e-04	0.63
RE-hsa-miR-382	3.20e-02	2.59e-04	0.63
RE-hsa-miR-1324	1.66e-02	2.67e-04	0.63
RE-hsa-miR-486-3p	4.44e-02	2.75e-04	0.63
RE-hsa-miR-584	4.41e-02	2.88e-04	0.63
RE-hsa-miR-1302	3.62e-02	3.30e-04	0.63
RE-hsa-miR-342-3p	2.82e-02	3.31e-04	0.63
RE-hsa-miR-205	4.63e-02	3.32e-04	0.63

Table C.2: miRNAs with RE scores most predictive of survival in the Bentink dataset, based on the p-value calculated with a likelihood ratio test. Models for 106 miRNAs had a calculated p-value of $p < 0.051$. ci: Concordance index of the model.

	p-value	coefficient	ci
hsa.miR.552	3.98e-02	-4.63e-01	0.63
hsa.miR.325	2.19e-02	-3.91e-01	0.63
hsa.miR.589	2.02e-02	-2.71e-01	0.63
hsa.miR.371.3p	4.81e-02	-2.07e-01	0.63
hsa.miR.580	8.05e-03	-2.01e-01	0.63
hsa.miR.658	1.16e-02	-1.94e-01	0.63

hsa.miR.216b	5.68e-03	-1.50e-01	0.63
hsa.miR.499.3p	1.01e-02	-1.49e-01	0.63
hsa.miR.524.3p	1.01e-02	-1.15e-01	0.63
hsa.miR.519c.3p	3.13e-02	-1.02e-01	0.63
hsa.miR.520h	4.29e-02	-9.62e-02	0.63
hsa.miR.135b.	4.29e-03	-9.35e-02	0.63
hsa.miR.629	2.02e-02	-8.43e-02	0.63
hsa.miR.519b.3p	1.95e-02	-7.35e-02	0.63
hsa.miR.556.5p	1.49e-03	-6.52e-02	0.63
hsa.miR.485.5p	7.92e-03	-5.78e-02	0.63
hsa.miR.518e	1.78e-02	-5.03e-02	0.63
hsa.miR.501.3p	3.32e-02	-4.02e-02	0.63
hsa.miR.10a.	4.56e-02	-3.90e-02	0.63
hsa.miR.129.	3.73e-02	-3.85e-02	0.63
hsa.miR.137	4.09e-02	-3.79e-02	0.63
hsa.miR.369.5p	8.93e-03	-3.22e-02	0.63
hsa.miR.371.5p	2.12e-02	-2.66e-02	0.63
hsa.miR.154.	2.14e-02	-2.35e-02	0.63
hsa.miR.635	2.84e-04	-2.06e-02	0.63
hsa.miR.141.	1.19e-02	-1.96e-02	0.63
hsa.miR.196a	4.63e-02	-1.85e-02	0.63
hsa.miR.509.5p	3.68e-02	-1.70e-02	0.63
hsa.miR.517b	1.03e-02	-1.59e-02	0.63
hsa.miR.508.3p	2.01e-02	-1.45e-02	0.63
hsa.miR.22.	4.22e-03	-1.44e-02	0.63
hsa.miR.193b	1.30e-02	-1.38e-02	0.63
hsa.miR.624.	2.66e-04	-1.37e-02	0.63
hsa.miR.142.5p	2.94e-02	-1.29e-02	0.63
hsa.miR.19b.1.	1.44e-03	-1.06e-02	0.63
hsa.miR.185	3.97e-03	-1.01e-02	0.63
hsa.miR.196b	2.10e-02	-9.79e-03	0.63
hsa.miR.141	4.90e-03	-8.90e-03	0.63
hsa.miR.106b	2.51e-03	-7.62e-03	0.63
hsa.let.7e.	1.75e-04	-5.79e-03	0.63
hsa.miR.29c	2.54e-03	-5.61e-03	0.63
hsa.miR.205	1.97e-02	-5.51e-03	0.63
hsa.miR.380.	7.55e-06	-5.19e-03	0.63
hsa.miR.598	1.08e-03	-4.41e-03	0.63
hsa.miR.30e	9.88e-04	-4.40e-03	0.63
hsa.miR.487b	9.66e-04	-4.10e-03	0.63
hsa.miR.510	7.13e-04	-3.40e-03	0.63

hsa.miR.338.3p	1.29e-03	-2.81e-03	0.63
hsa.miR.500	1.65e-04	-2.54e-03	0.63
hsa.miR.361.5p	1.15e-04	-1.62e-03	0.63
hsa.miR.509.3p	4.35e-04	-1.52e-03	0.63
hsa.miR.200a.	5.86e-06	3.27e-04	0.63
hsa.miR.181b	4.72e-05	6.81e-04	0.63
hsa.miR.551b	4.31e-04	1.58e-03	0.63
hsa.miR.24.1.	3.18e-04	2.50e-03	0.63
hsa.miR.515.3p	8.78e-05	2.72e-03	0.63
hsa.miR.31.	1.87e-03	3.21e-03	0.63
hsa.miR.544	3.18e-06	3.71e-03	0.63
hsa.miR.514	4.89e-03	4.12e-03	0.63
hsa.miR.100	6.13e-03	6.35e-03	0.63
hsa.miR.509.3.5p	1.03e-02	7.13e-03	0.63
hsa.miR.675	5.60e-05	7.24e-03	0.63
hsa.miR.432	2.05e-03	7.42e-03	0.63
hsa.miR.181c	9.56e-03	7.61e-03	0.63
hsa.miR.96	1.42e-02	7.77e-03	0.63
hsa.miR.203	1.31e-02	8.11e-03	0.63
hsa.miR.200b.	2.64e-03	8.51e-03	0.63
hsa.miR.27b	5.01e-03	8.79e-03	0.63
hsa.miR.302b.	1.61e-05	8.92e-03	0.63
hsa.miR.125b	1.41e-02	1.13e-02	0.63
hsa.miR.130b	1.37e-02	1.32e-02	0.63
hsa.miR.507	3.37e-02	1.65e-02	0.63
hsa.miR.195	4.08e-02	1.68e-02	0.63
hsa.miR.181d	1.81e-02	1.82e-02	0.63
hsa.miR.34a.	6.57e-03	2.17e-02	0.63
hsa.miR.892b	1.41e-02	2.18e-02	0.63
hsa.miR.182	2.60e-02	2.28e-02	0.63
hsa.miR.376b	3.10e-02	2.55e-02	0.63
hsa.miR.126	3.14e-02	2.59e-02	0.63
hsa.miR.330.5p	2.22e-04	2.65e-02	0.63
hsa.miR.30c.2.	2.81e-02	3.88e-02	0.63
hsa.miR.543	2.99e-02	4.27e-02	0.63
hsa.miR.661	1.56e-03	4.85e-02	0.63
hsa.miR.376a.	3.45e-02	5.19e-02	0.63
hsa.miR.875.5p	3.42e-03	7.38e-02	0.63
hsa.miR.518c	1.63e-02	7.92e-02	0.63
hsa.miR.191.	2.37e-02	7.98e-02	0.63
hsa.miR.30c.1.	1.37e-02	8.51e-02	0.63

hsa.miR.511	1.49e-03	9.08e-02	0.63
hsa.miR.20b.	4.20e-03	9.76e-02	0.63
hsa.miR.377.	2.03e-02	1.04e-01	0.63
hsa.miR.643	2.23e-03	1.14e-01	0.63
hsa.miR.525.3p	6.91e-03	1.29e-01	0.63
hsa.miR.624	1.03e-02	1.37e-01	0.63
hsa.miR.26a.2.	6.23e-03	1.52e-01	0.63
hsa.miR.527	6.53e-03	1.68e-01	0.63
hsa.miR.363.	1.77e-02	1.69e-01	0.63
hsa.miR.888	4.89e-02	2.09e-01	0.63
hsa.miR.621	4.47e-02	2.64e-01	0.63
hsa.miR.587	1.19e-02	2.78e-01	0.63
hsa.miR.612	1.71e-02	3.22e-01	0.63
hsa.miR.523	4.94e-02	3.49e-01	0.63
hsa.miR.644	3.15e-02	3.69e-01	0.63
hsa.miR.626	4.65e-02	4.63e-01	0.63

Table C.3: miRNAs with expression values most predictive of survival in the TCGA dataset, based on the p-value calculated with a likelihood ratio test. Models for 105 miRNAs had a calculated p-value of $p < 0.051$. ci: Concordance index of the model.

	p-value	coefficient	ci
hsa-miR-1301	4.76e-02	-9.27e-05	0.63
hsa-miR-592	4.95e-02	-8.82e-05	0.63
hsa-miR-768-3p	4.72e-02	-8.62e-05	0.63
hsa-miR-885-3p	4.43e-02	-8.49e-05	0.63
hsa-miR-484	4.77e-02	-8.37e-05	0.63
hsa-miR-938	4.95e-02	-7.99e-05	0.63
hsa-miR-125b	1.81e-02	-7.89e-05	0.63
hsa-miR-125a-5p	1.81e-02	-7.89e-05	0.63
hsa-miR-197	3.59e-02	-7.73e-05	0.63
hsa-miR-627	4.30e-02	-7.66e-05	0.63
hsa-miR-188-3p	2.72e-02	-7.19e-05	0.63
hsa-miR-769-5p	4.17e-02	-6.98e-05	0.63
hsa-miR-760	2.06e-02	-6.66e-05	0.63
hsa-miR-1260	2.97e-02	-6.58e-05	0.63
hsa-miR-623	3.30e-02	-6.21e-05	0.63
hsa-miR-515-3p	2.60e-02	-6.07e-05	0.63

hsa-miR-920	2.82e-02	-5.90e-05	0.63
hsa-miR-875-5p	2.97e-02	-5.75e-05	0.63
hsa-miR-770-5p	3.59e-02	-5.71e-05	0.63
hsa-miR-596	1.60e-02	-5.53e-05	0.63
hsa-miR-220a	2.66e-02	-5.18e-05	0.63
hsa-miR-492	1.62e-02	-5.06e-05	0.63
hsa-miR-339-3p	4.14e-02	-5.04e-05	0.63
hsa-miR-519e	1.71e-02	-4.89e-05	0.63
hsa-miR-628-5p	1.48e-02	-4.79e-05	0.63
hsa-miR-9	8.56e-03	-4.51e-05	0.63
hsa-miR-219-2-3p	1.51e-02	-4.47e-05	0.63
hsa-miR-379	2.83e-02	-4.29e-05	0.63
hsa-miR-362-5p	8.88e-03	-3.61e-05	0.63
hsa-miR-940	5.03e-03	-3.56e-05	0.63
hsa-miR-450a	4.37e-02	-3.51e-05	0.63
hsa-miR-553	1.67e-02	-3.31e-05	0.63
hsa-miR-652	1.77e-02	-3.25e-05	0.63
hsa-miR-1243	7.66e-03	-3.22e-05	0.63
hsa-miR-378	5.89e-03	-3.11e-05	0.63
hsa-miR-422a	5.89e-03	-3.11e-05	0.63
hsa-miR-524-3p	1.35e-02	-3.00e-05	0.63
hsa-miR-525-3p	1.35e-02	-3.00e-05	0.63
hsa-miR-668	7.22e-03	-2.48e-05	0.63
hsa-miR-1265	3.55e-03	-2.42e-05	0.63
hsa-miR-1184	3.67e-03	-2.38e-05	0.63
hsa-miR-140-5p	5.25e-03	-2.31e-05	0.63
hsa-miR-542-5p	1.13e-02	-2.26e-05	0.63
hsa-miR-188-5p	2.83e-03	-2.01e-05	0.63
hsa-miR-122	4.33e-03	-1.93e-05	0.63
hsa-miR-1281	1.70e-02	-1.91e-05	0.63
hsa-miR-1296	2.24e-03	-1.55e-05	0.63
hsa-miR-1233	1.24e-03	-1.26e-05	0.63
hsa-miR-1288	5.36e-04	-6.13e-06	0.63
hsa-miR-1180	1.28e-04	-2.51e-06	0.63
hsa-miR-127-3p	3.49e-05	-9.11e-07	0.63
hsa-miR-1227	5.17e-06	-8.22e-07	0.63
hsa-miR-100	2.89e-06	4.17e-07	0.63
hsa-miR-99a	2.89e-06	4.17e-07	0.63
hsa-miR-99b	2.89e-06	4.17e-07	0.63
hsa-miR-210	5.90e-04	4.80e-06	0.63
hsa-miR-1307	1.04e-03	5.39e-06	0.63

hsa-miR-630	2.89e-04	6.43e-06	0.63
hsa-miR-629	7.91e-04	9.07e-06	0.63
hsa-miR-663	8.18e-04	9.51e-06	0.63
hsa-miR-598	1.62e-03	9.70e-06	0.63
hsa-miR-573	1.17e-03	1.17e-05	0.63
hsa-miR-886-3p	4.88e-03	1.20e-05	0.63
hsa-miR-1273	4.19e-03	1.26e-05	0.63
hsa-miR-1280	8.86e-03	1.50e-05	0.63
hsa-miR-1181	4.49e-02	2.68e-05	0.63
hsa-miR-487b	2.04e-02	2.90e-05	0.63
hsa-miR-199a-5p	8.55e-03	3.39e-05	0.63
hsa-miR-199b-5p	8.55e-03	3.39e-05	0.63
hsa-miR-329	8.19e-03	3.39e-05	0.63
hsa-miR-362-3p	8.19e-03	3.39e-05	0.63
hsa-miR-139-3p	3.73e-02	3.51e-05	0.63
hsa-miR-28-3p	6.26e-03	3.78e-05	0.63
hsa-miR-24	6.28e-03	3.82e-05	0.63
hsa-miR-487a	1.27e-02	4.05e-05	0.63
hsa-miR-193a-3p	1.82e-02	4.38e-05	0.63
hsa-miR-7	1.30e-02	4.68e-05	0.63
hsa-miR-342-3p	1.81e-02	4.81e-05	0.63
hsa-miR-129-3p	1.46e-02	4.93e-05	0.63
hsa-miR-296-5p	3.72e-02	4.96e-05	0.63
hsa-miR-193b	2.70e-02	5.35e-05	0.63
hsa-miR-483-3p	2.64e-02	5.73e-05	0.63
hsa-miR-1289	4.34e-02	6.27e-05	0.63
hsa-miR-490-3p	3.87e-02	6.89e-05	0.63
hsa-miR-874	3.83e-02	7.80e-05	0.63
hsa-miR-1236	3.42e-02	7.95e-05	0.63
hsa-miR-574-5p	4.47e-02	7.99e-05	0.63
hsa-miR-220b	4.69e-02	9.16e-05	0.63

Table C.4: miRNAs with RE scores most predictive of survival in the TCGA dataset, based on the p-value calculated with a likelihood ratio test of a univariate model. Models for 88 miRNAs had a calculated p-value of $p < 0.052$. ci: Concordance index of the model.

Bibliography

- [1] Stein Aerts, Gert Thijs, Michal Dabrowski, Yves Moreau, and Bart De Moor. Comprehensive analysis of the base composition around the transcription start site in metazoa. *BMC Genomics*, 5:34, June 2004. PMID: 15171795 PMCID: PMC436054. 43
- [2] Victor Ambros, Bonnie Bartel, David P. Bartel, Christopher B. Burge, James C. Carrington, Xuemei Chen, Gideon Dreyfuss, Sean R Eddy, Sam Griffiths-Jones, Mhairi Marshall, Marjori Matzke, Gary Ruvkun, and Thomas Tuschl. A uniform system for microRNA annotation. *RNA*, 9(3):277 –279, March 2003. 42, 118
- [3] P. K. Andersen and R. D. Gill. Cox’s regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4):1100–1120, December 1982. ArticleType: research-article / Full publication date: Dec., 1982 / Copyright Â© 1982 Institute of Mathematical Statistics. 99
- [4] Christopher P. Arnold, Ruoying Tan, Baiyu Zhou, Si-Biao Yue, Steven Schaffert, Joseph R. Biggs, Regis Doyonnas, Miao-Chia Lo, John M. Perry, ValÃ©rie M. Renault, Alessandra Sacco, Tim Somervaille, Patrick Viatour, Anne Brunet, Michael L. Cleary, Linheng Li, Julien Sage, Dong-Er Zhang,

- Helen M. Blau, Caifu Chen, and Chang-Zheng Chen. MicroRNA programs in normal and aberrant stem and progenitor cells. *Genome Research*, March 2011. 6
- [5] Amit Arora and David A. C Simpson. Individual mRNA expression profiles reveal the effects of specific microRNAs. *Genome Biology*, 9:R82+, May 2008. 13
- [6] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1):25–29, May 2000. PMID: 10802651. 18
- [7] Daehyun Baek, Judit Vill  n, Chanseok Shin, Fernando D. Camargo, Steven P. Gygi, and David P. Bartel. The impact of microRNAs on protein output. *Nature*, 455(7209):64–71, September 2008. PMID: 18668037 PMCID: 2745094. 6, 13, 29
- [8] Marina Bagnoli, Loris De Cecco, Anna Granata, Roberta Nicoletti, Edoardo Marchesi, Paola Alberti, Barbara Valeri, Massimo Libra, Mattia Barbareschi, Francesco Raspagliesi, Delia Mezzanzanica, and Silvana Canevari. Identification of a chrXq27.3 microRNA cluster associated with early relapse in advanced stage ovarian cancer patients. *Oncotarget*, 2(12):1265–1278, December 2011. PMID: 22246208. 8, 75, 87, 107

- [9] David P Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, January 2004. PMID: 14744438. 4
- [10] David P Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–233, January 2009. PMID: 19167326. 4
- [11] Dillon I Beardsley, David Kowbel, Tamara A Lataxes, Jennifer M Mannino, Hong Xin, Wan-Ju Kim, Colin Collins, and Kevin D Brown. Characterization of the novel amplified in breast cancer-1 (NABC1) gene product. *Experimental cell research*, 290(2):402–413, November 2003. PMID: 14567997. 45
- [12] Jennifer L Bearfoot, David Y H Choong, Kylie L Gorringer, and Ian G Campbell. Genetic analysis of cancer-implicated MicroRNA in ovarian cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 14(22):7246–7250, November 2008. PMID: 19010840. iii
- [13] Ausra Bendoraite, Emily C Knouf, Kavita S Garg, Rachael K Parkin, Evan M Kroh, Kathy C O’Briant, Aviva P Ventura, Andrew K Godwin, Beth Y Karlan, Charles W Drescher, Nicole Urban, Beatrice S Knudsen, and Muneesh Tewari. Regulation of miR-200 family microRNAs and ZEB transcription factors in ovarian cancer: evidence supporting a mesothelial-to-epithelial transition. *Gynecologic Oncology*, 116(1):117–125, January 2010. PMID: 19854497. 45, 101

- [14] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995. 72
- [15] Stefan Bentink, Benjamin Haibe-Kains, Thomas Risch, Jian-Bing Fan, Michelle S. Hirsch, Kristina Holton, Renee Rubio, Craig April, Jing Chen, Eliza Wickham-Garcia, Joyce Liu, Aedin Culhane, Ronny Drapkin, John Quackenbush, and Ursula A. Matulonis. Angiogenic mRNA and micro-RNA gene expression signature predicts a novel subtype of serous ovarian cancer. *PLoS ONE*, 7(2):e30269, February 2012. iii, 20, 26, 29, 53, 64, 92, 96
- [16] Isaac Bentwich, Amir Avniel, Yael Karov, Ranit Aharonov, Shlomit Gilad, Omer Barad, Adi Barzilai, Paz Einat, Uri Einav, Eti Meiri, Eilon Sharon, Yael Spector, and Zvi Bentwich. Identification of hundreds of conserved and nonconserved human microRNAs. *Nat Genet*, 37(7):766–770, July 2005. 69, 75
- [17] Doron Betel, Anjali Koppal, Phaedra Agius, Chris Sander, and Christina Leslie. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biology*, 11(8):R90, 2010. PMID: 20799968. 15, 70, 72
- [18] Doron Betel, Manda Wilson, Aaron Gabow, Debora S Marks, and Chris Sander. The microRNA.org resource: targets and expression. *Nucleic Acids Research*, 36(Database issue):D149–153, January 2008. PMID: 18158296. 6

BIBLIOGRAPHY

- [19] Resham Bhattacharya, Milena Nicoloso, Rochelle Arvizo, Enfeng Wang, Angelica Cortez, Simona Rossi, George A Calin, and Priyabrata Mukherjee. MiR-15a and MiR-16 control bmi-1 expression in ovarian cancer. *Cancer Research*, 69(23):9090–9095, December 2009. PMID: 19903841. 7, 10
- [20] Paul E Blower, Joseph S Verducci, Shili Lin, Jin Zhou, Ji-Hyun Chung, Zunyan Dai, Chang-Gong Liu, William Reinhold, Philip L Lorenzi, Eric P Kaldjian, Carlo M Croce, John N Weinstein, and Wolfgang Sadee. MicroRNA expression profiles for the NCI-60 cancer cell panel. *Molecular Cancer Therapeutics*, 6(5):1483–1491, May 2007. PMID: 17483436. 30, 32, 41, 60
- [21] Todd Boren, Yin Xiong, Ardeshir Hakam, Robert Wenham, Sachin Apte, Gina Chan, Siddharth G Kamath, Dung-Tsa Chen, Holly Dressman, and Johnathan M Lancaster. MicroRNAs and their target messenger RNAs associated with ovarian cancer response to chemotherapy. *Gynecologic Oncology*, 113(2):249–255, May 2009. PMID: 19237188. 7, 10, 76
- [22] Marina Bousquet, Marian H Harris, Beiyan Zhou, and Harvey F Lodish. MicroRNA miR-125b causes leukemia. *Proceedings of the National Academy of Sciences of the United States of America*, November 2010. PMID: 21118985. 7
- [23] Benjamin Boyerinas, Sun-Mi Park, Annika Hau, Andrea E Murmann, and Marcus E Peter. The role of let-7 in cell differentiation and cancer. *Endocrine-Related Cancer*, 17(1):F19–36, March 2010. PMID: 19779035. 7

- [24] Julius Brennecke, Alexander Stark, Robert B Russell, and Stephen M Cohen. Principles of microRNA-target recognition. *PLoS Biology*, 3(3):e85, March 2005. PMID: 15723116. 4
- [25] H M BÄyvelstad, S NygÄerd, H L StÄyrvold, M Aldrin, ÄB Borgan, A Frigessi, and O C Lingjaerde. Predicting survival from microarray data—a comparative study. *Bioinformatics (Oxford, England)*, 23(16):2080–2087, August 2007. PMID: 17553857. 93, 112
- [26] Xuezhong Cai, Curt H Hagedorn, and Bryan R Cullen. Human micro-RNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA (New York, N. Y.)*, 10(12):1957–1966, December 2004. PMID: 15525708. 4
- [27] George Adrian Calin, Calin Dan Dumitru, Masayoshi Shimizu, Roberta Bichi, Simona Zupo, Evan Noch, Hansjuerg Aldler, Sashi Rattan, Michael Keating, Kanti Rai, Laura Rassenti, Thomas Kipps, Massimo Negrini, Florencia Bullrich, and Carlo M Croce. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences of the United States of America*, 99(24):15524–15529, November 2002. PMID: 12434020. 7, 8
- [28] George Adrian Calin, Chang-Gong Liu, Cinzia Seignani, Manuela Ferracin, Nadia Felli, Calin Dan Dumitru, Masayoshi Shimizu, Amelia Cimmino, Simona Zupo, Mariella Dono, Marie L Dell’Aquila, Hansjuerg Alder, Laura Rassenti, Thomas J Kipps, Florencia Bullrich, Massimo Negrini, and

- Carlo M Croce. MicroRNA profiling reveals distinct signatures in b cell chronic lymphocytic leukemias. *Proceedings of the National Academy of Sciences of the United States of America*, 101(32):11755–11760, August 2004. PMID: 15284443. 9
- [29] George Adrian Calin, Cinzia Sevignani, Calin Dan Dumitru, Terry Hyslop, Evan Noch, Sai Yendamuri, Masayoshi Shimizu, Sashi Rattan, Florencia Bullrich, Massimo Negrini, and Carlo M Croce. Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2999–3004, March 2004. PMID: 14973191. 8, 124
- [30] Elcie Chan, Rajeshvari Patel, Sunitha Nallur, Elena Ratner, Antonella Bacchiocchi, Kathleen Hoyt, Sebastian Szpakowski, Sirie Godshalk, Stephan Ariyan, Mario Sznol, Ruth Halaban, Michael Krauthammer, David Tuck, Frank J Slack, and Joanne Barnes Weidhaas. MicroRNA signatures differentiate melanoma subtypes. *Cell Cycle (Georgetown, Tex.)*, 10(11), June 2011. PMID: 21543894. 8
- [31] Hua Chang, Xin Zhou, Zhen-Ning Wang, Yong-Xi Song, Fang Zhao, Peng Gao, Yeunpo Chiang, and Hui-Mian Xu. Increased expression of miR-148b in ovarian carcinoma and its clinical significance. *Molecular Medicine Reports*, February 2012. PMID: 22344713. 10
- [32] Tsung-Cheng Chang, Duonan Yu, Yun-Sil Lee, Erik A. Wentzel, Dan E. Arking, Kristin M. West, Chi V. Dang, Andrei Thomas-Tikhonenko, and

- Joshua T. Mendell. Widespread microRNA repression by myc contributes to tumorigenesis. *Nature Genetics*, 40(1):43–50, 2008. iii
- [33] A Chao, C-Y Lin, Y-S Lee, C-L Tsai, P-C Wei, S Hsueh, T-I Wu, C-N Tsai, C-J Wang, A-S Chao, T-H Wang, and C-H Lai. Regulation of ovarian cancer progression by microRNA-187 through targeting disabled homolog-2. *Oncogene*, July 2011. PMID: 21725366. iii, 11
- [34] Chang-Zheng Chen, Ling Li, Harvey F Lodish, and David P Bartel. Micro-RNAs modulate hematopoietic lineage differentiation. *Science (New York, N.Y.)*, 303(5654):83–86, January 2004. PMID: 14657504. 6
- [35] Thimmaiah P Chendrimada, Richard I Gregory, Easwari Kumaraswamy, Jessica Norman, Neil Cooch, Kazuko Nishikura, and Ramin Shiekhattar. TRBP recruits the dicer complex to ago2 for microRNA processing and gene silencing. *Nature*, 436(7051):740–744, August 2005. PMID: 15973356. 4
- [36] Angie M Cheng, Mike W Byrom, Jeffrey Shelton, and Lance P Ford. Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. *Nucleic Acids Research*, 33(4):1290–1297, 2005. PMID: 15741182. 6, 9
- [37] Chao Cheng, Xuping Fu, Pedro Alves, and Mark Gerstein. mRNA expression profiles show differential regulatory effects of microRNAs between ER+ and ER- breast cancer. *Genome Biology*, 10(9):R90, 2009. 8, 11, 12, 13, 19, 65, 70, 86, 97

- [38] Chao Cheng and Lei M Li. Inferring microRNA activities by combining gene expression with microRNA target prediction. *PLoS ONE*, 3(4):e1989, 2008. PMID: 18431476. 8
- [39] Foteini Christodoulou, Florian Raible, Raju Tomer, Oleg Simakov, Kalliopi Trachana, Sebastian Klaus, Heidi Snyman, Gregory J Hannon, Peer Bork, and Detlev Arendt. Ancient animal microRNAs and the evolution of tissue identity. *Nature*, 463(7284):1084–1088, February 2010. PMID: 20118916. 124
- [40] Amelia Cimmino, George Adrian Calin, Muller Fabbri, Marilena V Iorio, Manuela Ferracin, Masayoshi Shimizu, Sylwia E Wojcik, Rami I Aqeilan, Simona Zupo, Mariella Dono, Laura Rassenti, Hansjuerg Alder, Stefano Volinia, Chang-Gong Liu, Thomas J Kipps, Massimo Negrini, and Carlo M Croce. miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13944–13949, September 2005. PMID: 16166262. 7
- [41] Dawn R Cochrane, Erin N Howe, Nicole S Spoelstra, and Jennifer K Richer. Loss of miR-200c: a marker of aggressiveness and chemoresistance in female reproductive cancers. *Journal of Oncology*, 2010:821717, 2010. PMID: 20049172. 45
- [42] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. 92, 94
- [43] Chad J Creighton, Anadulce Hernandez-Herrera, Anders Jacobsen, Douglas A Levine, Parminder Mankoo, Nikolaus Schultz, Ying Du, Yiqun

- Zhang, Erik Larsson, Robert Sheridan, Weimin Xiao, Paul T Spellman, Gad Getz, David A Wheeler, Charles M Perou, Richard A Gibbs, Chris Sander, D Neil Hayes, and Preethi H Gunaratne. Integrated analyses of microRNAs demonstrate their widespread influence on gene expression in high-grade serous ovarian carcinoma. *PloS One*, 7(3):e34546, 2012. PMID: 22479643. 31
- [44] Christopher P Crum, Ronny Drapkin, Alexander Miron, Tan A Ince, Michael Muto, David W Kindelberger, and Yonghee Lee. The distal fallopian tube: a new model for pelvic serous carcinogenesis. *Current opinion in obstetrics & gynecology*, 19(1):3–9, February 2007. PMID: 17218844. 3
- [45] Aedin C Culhane, Thomas Schwarzl, Razvan Sultana, Kermshlise C Picard, Shaita C Picard, Tim H Lu, Katherine R Franklin, Simon J French, Gerald Papenhausen, Mick Correll, and John Quackenbush. GeneSigDB—a curated database of gene expression signatures. *Nucleic Acids Research*, 38(Database issue):D716–725, January 2010. PMID: 19934259. 19
- [46] Jordan M Cummins, Yiping He, Rebecca J Leary, Ray Pagliarini, Jr Diaz, Luis A, Tobias Sjoblom, Omer Barad, Zvi Bentwich, Anna E Szafranska, Emmanuel Labourier, Christopher K Raymond, Brian S Roberts, Hartmut Juhl, Kenneth W Kinzler, Bert Vogelstein, and Victor E Velculescu. The colorectal microRNAome. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10):3687–3692, March 2006. PMID: 16505370. 46

- [47] G. Dennis, B. T Sherman, D. A Hosack, J. Yang, W. Gao, H. C Lane, and R. A Lempicki. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol*, 4(5), 2003. 18, 81
- [48] Gianpiero Di Leva and Carlo M. Croce. Roles of small RNAs in tumor formation. *Trends in Molecular Medicine*, In Press, Corrected Proof, 2010. 124
- [49] Joseph A DiDonato, Frank Mercurio, and Michael Karin. NF- κ B and the link between inflammation and cancer. *Immunological reviews*, 246(1):379–400, March 2012. PMID: 22435567. 84
- [50] Holly K Dressman, Andrew Berchuck, Gina Chan, Jun Zhai, Andrea Bild, Robyn Sayer, Janiel Cragun, Jennifer Clarke, Regina S Whitaker, Lihua Li, Jonathan Gray, Jeffrey Marks, Geoffrey S Ginsburg, Anil Potti, Mike West, Joseph R Nevins, and Johnathan M Lancaster. An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 25(5):517–525, February 2007. PMID: 17290060. 24
- [51] Pan Du, Warren A Kibbe, and Simon M Lin. lumi: a pipeline for processing illumina microarray. *Bioinformatics (Oxford, England)*, 24(13):1547–1548, July 2008. PMID: 18467348. 29
- [52] S Dudoit and T P Speed. A score test for the linkage analysis of qualitative and quantitative traits based on identity by descent data from sib-pairs.

Biostatistics (Oxford, England), 1(1):1–26, March 2000. PMID: 12933522.

23

- [53] Ram Eitan, Michal Kushnir, Gila Lithwick-Yanai, Miriam Ben David, Moshe Hoshen, Marek Glezerman, Moshe Hod, Gad Sabah, Shai Rosenwald, and Hanoch Levavi. Tumor microRNA expression patterns associated with resistance to platinum based chemotherapy and survival in ovarian cancer patients. *Gynecologic Oncology*, 114(2):253–259, August 2009. PMID: 19446316. 7, 10

- [54] Daniel C Ellwanger, Florian A BÄijttner, Hans-Werner Mewes, and Volker StÄijmpfen. The sufficient minimal set of miRNA seed types. *Bioinformatics (Oxford, England)*, March 2011. PMID: 21441577. 4

- [55] David Engler and Yi Li. Survival analysis with large dimensional covariates: An application in microarray studies. Technical Report 68, Harvard University, 2007. 19

- [56] Aurora Esquela-Kerscher and Frank J Slack. Oncomirs - microRNAs with a role in cancer. *Nature Reviews. Cancer*, 6(4):259–269, April 2006. PMID: 16557279. iii

- [57] M F Evans, I W McDicken, and C S Herrington. Numerical abnormalities of chromosomes 1, 11, 17, and x are associated with stromal invasion in serous and mucinous epithelial ovarian tumours. *The Journal of pathology*, 189(1):53–59, September 1999. PMID: 10451488. 69

- [58] Manuela Ferracin, Massimo Pedriali, Angelo Veronese, Barbara Zagatti, Roberta GafÄä, Eros Magri, Maria Lunardi, Gardenia Munerato, Giulia

- Querzoli, Iva Maestri, Linda Ulazzi, Italo Nenci, Carlo M Croce, Giovanni Lanza, Patrizia Querzoli, and Massimo Negrini. MicroRNA profiling for the identification of cancers with unknown primary tissue-of-origin. *The Journal of Pathology*, April 2011. PMID: 21630269. 7
- [59] Jeffrey M. Friedman and Peter A. Jones. MicroRNAs: critical mediators of differentiation, development and disease. *Swiss medical weekly*, 139(33-34):466–472, August 2009. PMID: 19705306 PMCID: 2854010. 93, 124
- [60] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostat*, December 2007. 33
- [61] Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105, January 2009. PMID: 18955434. 4, 6
- [62] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004. PMID: 15461798. 42

- [63] Jelle J Goeman. L(1) penalized estimation in the cox proportional hazards model. *Biometrical Journal*, 52(1):70–84, February 2010. PMID: 19937997. 112
- [64] S. Griffiths-Jones. The microRNA registry. *Nucleic Acids Research*, 32(90001):109D–111, January 2004. 118
- [65] S. Griffiths-Jones. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(90001):D140–D144, January 2006. 42, 118
- [66] S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Research*, 36(Database):D154–D158, December 2007. 4, 42, 118
- [67] Huili Guo, Nicholas T. Ingolia, Jonathan S. Weissman, and David P. Bartel. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308):835–840, 2010. 6, 13, 29, 65, 70
- [68] Li-Min Guo, Yong Pu, Zhe Han, Tao Liu, Yi-Xuan Li, Min Liu, Xin Li, and Hua Tang. MicroRNA-9 inhibits ovarian cancer cell growth through regulation of NF-kappaB1. *The FEBS Journal*, August 2009. PMID: 19702828. 7
- [69] Xuejiang Guo, Bing Su, Zuomin Zhou, and Jiahao Sha. Rapid evolution of mammalian x-linked testis microRNAs. *BMC Genomics*, 10:97, 2009. PMID: 19257908. 75

BIBLIOGRAPHY

- [70] Lin He, J Michael Thomson, Michael T Hemann, Eva Hernando-Monge, David Mu, Summer Goodson, Scott Powers, Carlos Cordon-Cardo, Scott W Lowe, Gregory J Hannon, and Scott M Hammond. A microRNA polycistron as a potential human oncogene. *Nature*, 435(7043):828–833, June 2005. PMID: 15944707. 7
- [71] J Hino, H Tateyama, N Minamino, K Kangawa, and H Matsuo. Isolation and identification of human brain natriuretic peptides in cardiac atrium. *Biochemical and biophysical research communications*, 167(2):693–700, March 1990. PMID: 2138890. 45
- [72] Ivo L Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13):3429–3431, July 2003. PMID: 12824340. 13
- [73] Estrid V S Hogdall, Andy Ryan, Susanne K Kjaer, Jan Blaakaer, Lise Christensen, Johannes E Bock, Eva Glud, Ian J Jacobs, and Claus K HÃ¸gdall. Loss of heterozygosity on the x chromosome is an independent prognostic factor in ovarian carcinoma: from the danish ”MALOVA” ovarian carcinoma study. *Cancer*, 100(11):2387–2395, June 2004. PMID: 15160342. 69
- [74] Lawrence S Hon and Zemin Zhang. The roles of binding site arrangement and combinatorial targeting in microRNA repression of gene expression. *Genome Biology*, 8(8):R166, 2007. PMID: 17697356 PMCID: PMC2374997. 49
- [75] Eran Hornstein, Jennifer H Mansfield, Soraya Yekta, Jimmy Kuang-Hsien Hu, Brian D Harfe, Michael T McManus, Scott Baskerville, David P Bartel,

- and Clifford J Tabin. The microRNA miR-196 acts upstream of *hoxb8* and *shh* in limb development. *Nature*, 438(7068):671–674, December 2005. PMID: 16319892. 6
- [76] Douglas A Hosack, Glynn Dennis, Brad T Sherman, H Clifford Lane, and Richard A Lempicki. Identifying biological themes within lists of genes with EASE. *Genome Biology*, 4(10):R70, 2003. PMID: 14519205. 18
- [77] Xiaoxia Hu, Dusten M Macdonald, Phyllis C Huettner, Zhihui Feng, Issam M El Naqa, Julie K Schwarz, David G Mutch, Perry W Grigsby, Simon N Powell, and Xiaowei Wang. A miR-200 microRNA cluster as prognostic marker in advanced ovarian cancer. *Gynecologic Oncology*, 114(3):457–464, September 2009. PMID: 19501389. 101
- [78] Ming-Yii Huang, Hwei-Ming Wang, Hui-Jen Chang, Chao-Peng Hsiao, Jaw-Yuan Wang, and Shiu-Ru Lin. Overexpression of *S100B*, *TM4SF4*, and *OLFM4* genes is correlated with liver metastasis in taiwanese colorectal cancer patients. *DNA and cell biology*, 31(1):43–49, January 2012. PMID: 22011044. 45
- [79] Eric Huntzinger and Elisa Izaurralde. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet*, 12(2):99–110, February 2011. 6, 70
- [80] A John Iafrate, Lars Feuk, Miguel N Rivera, Marc L Listewnik, Patricia K Donahoe, Ying Qi, Stephen W Scherer, and Charles Lee. Detection of large-scale variation in the human genome. *Nature Genetics*, 36(9):949–951, September 2004. PMID: 15286789. 75

- [81] Marilena V Iorio and Carlo M Croce. microRNA involvement in human cancer. *Carcinogenesis*, April 2012. PMID: 22491715. iii
- [82] Marilena V Iorio, Manuela Ferracin, Chang-Gong Liu, Angelo Veronese, Riccardo Spizzo, Silvia Sabbioni, Eros Magri, Massimo Pedriali, Muller Fabbri, Manuela Campiglio, Sylvie MÃnard, Juan P Palazzo, Anne Rosenberg, Piero Musiani, Stefano Volinia, Italo Nenci, George A Calin, Patrizia Querzoli, Massimo Negrini, and Carlo M Croce. MicroRNA gene expression deregulation in human breast cancer. *Cancer Research*, 65(16):7065–7070, August 2005. PMID: 16103053. 7
- [83] Ariel Israel, Roded Sharan, Eytan Ruppin, and Eithan Galun. Increased MicroRNA activity in human cancers. *PLoS ONE*, 4(6):e6045, June 2009. 8
- [84] Wei Jia, Juliet O Eneh, Supriya Ratnaparkhe, Molly K Altman, and Mandi M Murph. MicroRNA-30c-2* expressed in ovarian cancer cells suppresses growth factor induced cellular proliferation and downregulates the oncogene BCL9. *Molecular Cancer Research: MCR*, October 2011. PMID: 22024689. iii
- [85] Zhen Jiang and Robert Gentleman. Extensions to gene set enrichment. *Bioinformatics (Oxford, England)*, 23(3):306–313, February 2007. PMID: 17127676. 18
- [86] Steven M Johnson, Helge Grosshans, Jaclyn Shingara, Mike Byrom, Rich Jarvis, Angie Cheng, Emmanuel Labourier, Kristy L Reinert, David Brown,

- and Frank J Slack. RAS is regulated by the let-7 microRNA family. *Cell*, 120(5):635–647, March 2005. PMID: 15766527. 7
- [87] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, January 2000. PMID: 10592173. 18
- [88] Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, and Mika Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38(Database issue):D355–360, January 2010. PMID: 19880382. 18
- [89] Minoru Kanehisa, Susumu Goto, Masahiro Hattori, Kiyoko F. Aoki-Kinoshita, Masumi Itoh, Shuichi Kawashima, Toshiaki Katayama, Michihiro Araki, and Mika Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34(Database issue):D354–D357, January 2006. PMID: 16381885 PMCID: 1347464. 18
- [90] Xantha Karp and Victor Ambros. Developmental biology. encountering microRNAs in cell fate signaling. *Science (New York, N.Y.)*, 310(5752):1288–1289, November 2005. PMID: 16311325. 6
- [91] Michael Kertesz, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. The role of site accessibility in microRNA target recognition. *Nature Genetics*, 39(10):1278–1284, October 2007. PMID: 17893677. 14, 39, 70
- [92] Jaeyeon Kim, Donna M Coffey, Chad J Creighton, Zhifeng Yu, Shannon M Hawkins, and Martin M Matzuk. High-grade serous ovarian cancer arises from fallopian tube in a mouse model. *Proceedings of the National Academy*

of Sciences of the United States of America, 109(10):3921–3926, March 2012. PMID: 22331912. 3

- [93] Ulf Klein, Marie Lia, Marta Crespo, Rachael Siegel, Qiong Shen, Tongwei Mo, Alberto Ambesi-Impiombato, Andrea Califano, Anna Migliazza, Govind Bhagat, and Riccardo Dalla-Favera. The DLEU2/miR-15a/16-1 cluster controls b cell proliferation and its deletion leads to chronic lymphocytic leukemia. *Cancer Cell*, 17(1):28–40, January 2010. PMID: 20060366.

7

- [94] Fanfei Kong, Chaoyang Sun, Zhongxian Wang, Lingfei Han, Danhui Weng, Yunping Lu, and Gang Chen. miR-125b confers resistance of ovarian cancer cells to cisplatin by targeting pro-apoptotic bcl-2 antagonist killer 1. *Journal of Huazhong University of Science and Technology*, 31(4):543–549, August 2011. PMID: 21823019. 10, 11

- [95] Panagiotis A. Konstantinopoulos, Stephen A. Cannistra, Helen Fountzilas, Aedin Culhane, Kamana Pillay, Bo Rueda, Daniel Cramer, Michael Seiden, Michael Birrer, George Coukos, Lin Zhang, John Quackenbush, and Dimitrios Spentzos. Integrated analysis of multiple microarray datasets identifies a reproducible survival predictor in ovarian cancer. *PLoS ONE*, 6(3):e18202, March 2011. 92

- [96] Azra Krek, Dominic GrÄijn, Matthew N. Poy, Rachel Wolf, Lauren Rosenberg, Eric J. Epstein, Philip MacMenamin, Isabelle da Piedade, Kristin C. Gunsalus, Markus Stoffel, and Nikolaus Rajewsky. Combinatorial micro-

- RNA target predictions. *Nature Genetics*, 37(5):495–500, May 2005. 4, 49, 70
- [97] Madhu S Kumar, Jun Lu, Kim L Mercer, Todd R Golub, and Tyler Jacks. Impaired microRNA processing enhances cellular transformation and tumorigenesis. *Nat Genet*, 39(5):673–677, May 2007. 8, 65, 75, 86, 124
- [98] Madhu S Kumar, Ryan E Pester, Cindy Y Chen, Keara Lane, Christine Chin, Jun Lu, David G Kirsch, Todd R Golub, and Tyler Jacks. Dicer1 functions as a haploinsufficient tumor suppressor. *Genes & Development*, 23(23):2700–2704, December 2009. PMID: 19903759. 8, 86
- [99] Smriti Kumar, Arooshi Kumar, Parag P Shah, Shesh N Rai, Siva K Panguluri, and Sham S Kakar. MicroRNA signature of cis-platin resistant vs. cis-platin sensitive ovarian cancer cell lines. *Journal of Ovarian Research*, 4:17, 2011. 11
- [100] Robert J Kurman and Ie-Ming Shih. The origin and pathogenesis of epithelial ovarian cancer: a proposed unifying theory. *The American journal of surgical pathology*, 34(3):433–443, March 2010. PMID: 20154587. 3
- [101] Sabbi Lall, Dominic GrÄijn, Azra Krek, Kevin Chen, Yi-Lu Wang, Colin N Dewey, Pranidhi Sood, Teresa Colombo, Nicolas Bray, Philip Macmenamin, Huey-Ling Kao, Kristin C Gunsalus, Lior Pachter, Fabio Piano, and Nikolaus Rajewsky. A genome-wide map of conserved microRNA targets in *c. elegans*. *Current Biology: CB*, 16(5):460–471, March 2006. PMID: 16458514. 14, 72

BIBLIOGRAPHY

- [102] Michael S Lan and Mary B Breslin. Structure, expression, and biological function of INSM1 transcription factor in neuroendocrine differentiation. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, 23(7):2024–2033, July 2009. PMID: 19246490. 87
- [103] Man-Tat Lau, Wai-Kin So, and Peter C K Leung. Fibroblast growth factor 2 induces e-cadherin down-regulation via PI3K/Akt/mTOR and MAPK/ERK signaling in ovarian cancer cells. *PloS one*, 8(3):e59083, 2013. PMID: 23554977. 45
- [104] Yoontae Lee, Kipyoun Jeon, Jun-Tae Lee, Sunyoung Kim, and V.Narry Kim. MicroRNA maturation: stepwise processing and subcellular localization. *The EMBO Journal*, 21(17):4663–4670, September 2002. PMID: 12198168 PMCID: 126204. 5
- [105] F Legge, G Ferrandina, V Salutari, and G Scambia. Biological characterization of ovarian cancer: prognostic and therapeutic implications. *Annals of Oncology*, 16 Suppl 4:iv95–101, May 2005. PMID: 15923439. 2
- [106] Aliza Leizer, Ayesha B. Alvero, Han Hsuan Fu, Jennie C. Holmberg, Yung-Chi Cheng, Dan-Arin Silasi, Thomas Rutherford, and Gil Mor. Regulation of inflammation by the NF- κ B pathway in ovarian cancer stem cells. *American journal of reproductive immunology (New York, N.Y. : 1989)*, 65(4):438–447, April 2011. PMID: 20825380 PMCID: PMC3006007. 87
- [107] Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of hu-

BIBLIOGRAPHY

- man genes are microRNA targets. *Cell*, 120(1):15–20, January 2005. PMID: 15652477. 6, 72
- [108] Benjamin P Lewis, I-hung Shih, Matthew W Jones-Rhoades, David P Bartel, and Christopher B Burge. Prediction of mammalian microRNA targets. *Cell*, 115(7):787–798, December 2003. PMID: 14697198. 4, 13, 70
- [109] Jingjing Li, Yu Liu, Dong Dong, and Zhaolei Zhang. Evolution of an x-linked primate-specific micro RNA cluster. *Molecular Biology and Evolution*, 27(3):671–683, March 2010. PMID: 19933172. 75
- [110] Simon M Lin, Pan Du, Wolfgang Huber, and Warren A Kibbe. Model-based variance-stabilizing transformation for illumina microarray data. *Nucleic acids research*, 36(2):e11, February 2008. PMID: 18178591. 29
- [111] Jidong Liu, Marco Antonio Valencia-Sanchez, Gregory J Hannon, and Roy Parker. MicroRNA-dependent localization of targeted mRNAs to mammalian p-bodies. *Nature Cell Biology*, 7(7):719–723, July 2005. PMID: 15937477. 4
- [112] Jun Lu, Gad Getz, Eric A Miska, Ezequiel Alvarez-Saavedra, Justin Lamb, David Peck, Alejandro Sweet-Cordero, Benjamin L Ebert, Raymond H Mak, Adolfo A Ferrando, James R Downing, Tyler Jacks, H Robert Horvitz, and Todd R Golub. MicroRNA expression profiles classify human cancers. *Nature*, 435(7043):834–838, June 2005. PMID: 15944708. 7, 8
- [113] Lingeng Lu, Peter Schwartz, Luca Scarampi, Thomas Rutherford, Emilie Marion Canuto, Herbert Yu, and Dionyssios Katsaros. MicroRNA let-

- 7a: a potential marker for selection of paclitaxel in ovarian cancer management. *Gynecologic Oncology*, 122(2):366–371, August 2011. PMID: 21571355. 10, 11
- [114] Seiji Mabuchi, Masahide Ohmichi, Yukihiro Nishio, Tadashi Hayasaka, Akiko Kimura, Tsuyoshi Ohta, Maki Saito, Jun Kawagoe, Kazuhiro Takahashi, Namiko Yada-Hashimoto, Masahiro Sakata, Teiichi Motoyama, Hirohisa Kurachi, Keiichi Tasaka, and Yuji Murata. Inhibition of NFkappaB increases the efficacy of cisplatin in in vitro and in vivo ovarian cancer models. *The Journal of biological chemistry*, 279(22):23477–23485, May 2004. PMID: 15026414. 87
- [115] Manolis Maragkakis, Panagiotis Alexiou, Giorgio L Papadopoulos, Martin Reczko, Theodore Dalamagas, George Giannopoulos, George Goumas, Evangelos Koukis, Kornilios Kourtis, Victor A Simossis, Praveen Sethupathy, Thanasis Vergoulis, Nectarios Koziris, Timos Sellis, Panagiotis Tsanakas, and Artemis G Hatzigeorgiou. Accurate microRNA target prediction correlates with protein repression levels. *BMC Bioinformatics*, 10:295, 2009. PMID: 19765283. 14
- [116] Sergio Marchini, Duccio Cavalieri, Robert Fruscio, Enrica Calura, Daniela Garavaglia, Ilaria Fuso Nerini, Costantino Mangioni, Giorgio Cattoretti, Luca Clivio, Luca Beltrame, Dionyssios Katsaros, Luca Scarampi, Guido Menato, Patrizia Perego, Giovanna Chiorino, Alessandro Buda, Chiara Romualdi, and Maurizio D’Incalci. Association between miR-200c and the survival of patients with stage i epithelial ovarian cancer: a retrospective

- study of two independent tumour tissue collections. *The Lancet Oncology*, February 2011. PMID: 21345725. 45
- [117] Rosa J. Meijer and Jelle J. Goeman. Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. *Biometrical Journal*, 55(2):141–155, March 2013. 113
- [118] Daniel W Mueller, Michael Rehli, and Anja K Bosserhoff. miRNA expression profiling in melanocytes and melanoma cell lines reveals miRNAs associated with formation and progression of malignant melanoma. *The Journal of Investigative Dermatology*, 129(7):1740–1751, July 2009. PMID: 19212343. 76, 87
- [119] Shankar Mukherji, Margaret S Ebert, Grace X Y Zheng, John S Tsang, Phillip A Sharp, and Alexander van Oudenaarden. MicroRNAs can generate thresholds in target gene expression. *Nat Genet*, advance online publication, 2011. 6
- [120] Eun Ji Nam, Heejei Yoon, Sang Wun Kim, Hoguen Kim, Young Tae Kim, Jae Hoon Kim, Jae Wook Kim, and Sunghoon Kim. MicroRNA expression profiles in serous ovarian carcinoma. *Clinical Cancer Research*, 14(9):2690–2695, May 2008. PMID: 18451233. 10
- [121] Dmitriy Ovcharenko, Friedrich StÄlzel, David Poitz, Fernando Fierro, Markus Schaich, Andreas Neubauer, Kevin Kelnar, Timothy Davison, Carsten MÄjller-Tidow, Christian Thiede, Martin BornhÄuser, Gerhard Ehninger, David Brown, and Thomas Illmer. miR-10a overexpression is associated with NPM1 mutations and MDM4 downregulation in

- intermediate-risk acute myeloid leukemia. *Experimental Hematology*, July 2011. PMID: 21784052. 109
- [122] R. F. Ozols. Treatment goals in ovarian cancer. *International Journal of Gynecological Cancer*, 15(s1):3–11, 2005. 3
- [123] A E Pasquinelli, B J Reinhart, F Slack, M Q Martindale, M I Kuroda, B Maller, D C Hayward, E E Ball, B Degan, P M Åijller, J Spring, A Srinivasan, M Fishman, J Finnerty, J Corbo, M Levine, P Leahy, E Davidson, and G Ruvkun. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–89, November 2000. PMID: 11081512. 6
- [124] Neil D Perkins. The diverse and complex roles of NF-ÎB subunits in cancer. *Nature reviews. Cancer*, 12(2):121–132, February 2012. PMID: 22257950. 84
- [125] Marcus E Peter. Let-7 and miR-200 microRNAs: guardians against pluripotency and cancer progression. *Cell Cycle (Georgetown, Tex.)*, 8(6):843–852, March 2009. PMID: 19221491. 7
- [126] Rainer O Rhein and Korbinian Strimmer. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Systems Biology*, 1(1), 2007. 33
- [127] William Ritchie, Stephane Flamant, and John E J Rasko. Predicting microRNA targets and functions: traps for the unwary. *Nature Methods*, 6(6):397–398, June 2009. PMID: 19478799. 14, 15

BIBLIOGRAPHY

- [128] Robert A. Rollins, Fatemeh Haghighi, John R. Edwards, Rajdeep Das, Michael Q. Zhang, Jingyue Ju, and Timothy H. Bestor. Large-scale structure of genomic methylation patterns. *Genome Research*, 16(2):157–163, February 2006. PMID: 16365381 PMCID: PMC1361710. 43
- [129] Nitzan Rosenfeld, Ranit Aharonov, Eti Meiri, Shai Rosenwald, Yael Spector, Merav Zepeniuk, Hila Benjamin, Norberto Shabes, Sarit Tabak, Asaf Levy, Danit Lebanony, Yaron Goren, Erez Silberschein, Nurit Targan, Alex Ben-Ari, Shlomit Gilad, Netta Sion-Vardy, Ana Tobar, Meora Feinmesser, Oleg Kharenko, Ofer Nativ, Dvora Nass, Marina Perelman, Ady Yosepovich, Bruria Shalmon, Sylvie Polak-Charcon, Eddie Fridman, Amir Avniel, Isaac Bentwich, Zvi Bentwich, Dalia Cohen, Ayelet Chajut, and Iris Barshack. MicroRNAs accurately identify cancer tissue origin. *Nature biotechnology*, 26(4):462–469, April 2008. PMID: 18362881. 93
- [130] Sarah Roush and Frank J Slack. The let-7 family of microRNAs. *Trends in Cell Biology*, 18(10):505–516, October 2008. PMID: 18774294. 7
- [131] J. Schafer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biol*, 4, 2005. 33
- [132] Robert Schickel, Sun-Mi Park, Andrea E. Murmann, and Marcus E. Peter. mir-200c regulates induction of apoptosis through CD95 by targeting FAP-1. *Molecular Cell*, 38(6):908–915, June 2010. 45
- [133] Sebastian Schmeier, Ulf Schaefer, Magbubah Essack, and Vladimir B Bajic. Network analysis of microRNAs and their regulation in human ovarian

cancer. *BMC Systems Biology*, 5(1):183, November 2011. PMID: 22050994.

10

- [134] David Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241, April 1982. 95, 97
- [135] Jeffrey D Seidman and Robert J Kurman. Pathology of ovarian carcinoma. *Hematology/Oncology Clinics of North America*, 17(4):909–925, vii, August 2003. PMID: 12959182. 2
- [136] Matthias Selbach, Björn Schwanhüsser, Nadine Thierfelder, Zhuo Fang, Raya Khanin, and Nikolaus Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63, September 2008. PMID: 18668040. 6, 11
- [137] Praveen Sethupathy, Benoit Corda, and Artemis Hatzigeorgiou. TarBase: a comprehensive database of experimentally supported animal microRNA targets. *RNA*, 12(2):192–197, February 2006. 61
- [138] Shubin W Shahab, Lilya V Matyunina, Christopher G Hill, Lijuan Wang, Roman Mezencev, L Deette Walker, and John F McDonald. The effects of MicroRNA transfections on global patterns of gene expression in ovarian cancer cells are functionally coordinated. *BMC medical genomics*, 5(1):33, August 2012. PMID: 22853714. 11
- [139] Karin K Shih, Li-Xuan Qin, Edward J Tanner, Qin Zhou, Maria Bisogna, Fanny Dao, Narciso Olvera, Agnes Viale, Richard R Barakat, and Douglas A Levine. A microRNA survival signature (MiSS) for advanced ovarian cancer. *Gynecologic Oncology*, February 2011. PMID: 21354599. 92

BIBLIOGRAPHY

- [140] Antonio Sorrentino, Chang-Gong Liu, Antonio Addario, Cesare Peschle, Giovanni Scambia, and Cristiano Ferlini. Role of microRNAs in drug-resistant ovarian cancer cells. *Gynecologic Oncology*, 111(3):478–486, December 2008. PMID: 18823650. 10
- [141] Dimitrios Spentzos, Douglas A Levine, Shakirahmed Kolia, Hasan Otu, Jeff Boyd, Towia A Libermann, and Stephen A Cannistra. Unique gene expression profile based on pathologic response in epithelial ovarian cancer. *J Clin Oncol*, 23(31):7911–7918, November 2005. 2
- [142] Dimitrios Spentzos, Douglas A Levine, Marco F Ramoni, Marie Joseph, Xuesong Gu, Jeff Boyd, Towia Libermann, and Stephen A Cannistra. Gene expression signature with independent prognostic significance in epithelial ovarian cancer. *J Clin Oncol*, 22(23):4700–4710, December 2004. 2
- [143] Giovanni Stefani and Frank J Slack. Small non-coding RNAs in animal development. *Nature Reviews. Molecular Cell Biology*, 9(3):219–230, March 2008. PMID: 18270516. 9
- [144] K L Streicher, W Zhu, K P Lehmann, R W Georgantas, C A Morehouse, P Brohawn, R A Carrasco, Z Xiao, D A Tice, B W Higgs, L Richman, B Jallal, K Ranade, and Y Yao. A novel oncogenic role for the miRNA-506-514 cluster in initiating melanocyte transformation and promoting melanoma growth. *Oncogene*, August 2011. PMID: 21860416. 87
- [145] Le Su, Arthur V Sampaio, Kevin B Jones, Marina Pacheco, Angela Goytain, Shujun Lin, Neal Poulin, Lin Yi, Fabio M Rossi, Juergen Kast, Mario R Capecchi, T Michael Underhill, and Torsten O Nielsen. Decon-

- struction of the SS18-SSX fusion oncoprotein complex: insights into disease etiology and therapeutics. *Cancer cell*, 21(3):333–347, March 2012. PMID: 22439931. 46
- [146] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, October 2005. 18, 19, 24
- [147] Ruping Sun, Xuping Fu, Yao Li, Yi Xie, and Yumin Mao. Global gene expression analysis reveals reduced abundance of putative microRNA targets in human prostate tumours. *BMC Genomics*, 10:93, 2009. PMID: 19245699. 9
- [148] Junichi Takamizawa, Hiroyuki Konishi, Kiyoshi Yanagisawa, Shuta Tomida, Hirotaka Osada, Hideki Endoh, Tomoko Harano, Yasushi Yatabe, Masato Nagino, Yuji Nimura, Tetsuya Mitsudomi, and Takashi Takahashi. Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Research*, 64(11):3753–3756, June 2004. PMID: 15172979. 7
- [149] TCGA. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, June 2011. 21, 25, 29, 30, 63, 64, 92, 96

BIBLIOGRAPHY

- [150] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. 19
- [151] Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395, 1997. 19
- [152] Ryan C. Todd and Stephen J. Lippard. Inhibition of transcription by platinum antitumor compounds. *Metallomics : integrated biometal science*, 1(4):280–291, January 2009. PMID: 20046924 PMCID: 2752884. 3
- [153] Richard W Tothill, Anna V Tinker, Joshy George, Robert Brown, Stephen B Fox, Stephen Lade, Daryl S Johnson, Melanie K Trivett, Dariush Etemadmoghadam, Bianca Locandro, Nadia Traficante, Sian Fereday, Jillian A Hung, Yoke-Eng Chiew, Izhak Haviv, Dorota Gertig, Anna DeFazio, and David D L Bowtell. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical Cancer Research*, 14(16):5198–5208, August 2008. PMID: 18698038. 20, 24, 25, 63, 64
- [154] V G Tusher, R Tibshirani, and G Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–5121, April 2001. PMID: 11309499. 16, 67
- [155] Marijn T.M. van Jaarsveld, Jozen Helleman, Els M.J.J. Berns, and Erik A.C. Wiemer. MicroRNAs in ovarian cancer biology and therapy

BIBLIOGRAPHY

- resistance. *The International Journal of Biochemistry & Cell Biology*, 42(8):1282–1290, August 2010. 101
- [156] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0. 106
- [157] J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, S Broder, A G Clark, J Nadeau, V A McKusick, N Zinder, A J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea, A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington, J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C Evangelista, A E Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K A Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Milshina, H M Moore, A K Naik, V A Narayan, B Neelam, D Nusskern, D B Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport, R De-

BIBLIOGRAPHY

silets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy, B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Rear-don, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R GuigÃs, M J Campbell, K V Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, and X Zhu. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507):1304–1351, February 2001. PMID: 11181995. 42

- [158] P J Verweij and H C Van Houwelingen. Cross-validation in survival anal-

- ysis. *Statistics in Medicine*, 12(24):2305–2314, December 1993. PMID: 8134734. 113
- [159] Stefano Volinia, George A Calin, Chang-Gong Liu, Stefan Ambs, Amelia Cimmino, Fabio Petrocca, Rosa Visone, Marilena Iorio, Claudia Roldo, Manuela Ferracin, Robyn L Prueitt, Nozumu Yanaihara, Giovanni Lanza, Aldo Scarpa, Andrea Vecchione, Massimo Negrini, Curtis C Harris, and Carlo M Croce. A microRNA expression signature of human solid tumors defines cancer gene targets. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2257–2261, February 2006. PMID: 16461460. 7, 8
- [160] A von Heydebreck, W Huber, A Poustka, and M Vingron. Identifying splits with clear separation: a new class discovery method for gene expression data. *Bioinformatics (Oxford, England)*, 17 Suppl 1:S107–114, 2001. PMID: 11472999. 21
- [161] Dong Wang, Chengxiang Qiu, Haijun Zhang, Juan Wang, Qinghua Cui, and Yuxin Yin. Human microRNA oncogenes and tumor suppressors show significantly different biological patterns: from functions to targets. *PloS One*, 5(9), 2010. PMID: 20927335. 9
- [162] Xiaowei Wang and Xiaohui Wang. Systematic identification of microRNA functions by combining target prediction and expression profiling. *Nucleic Acids Research*, 34(5):1646–1652, 2006. PMID: 16549876 PMCID: 1405822. 13

BIBLIOGRAPHY

- [163] J Whittaker. *Graphical models in applied multivariate statistics*. Wiley, Chichester [England]; New York, 1990. 33
- [164] Kaitlyn Wurz, Rochelle L Garcia, Barbara A Goff, Patrick S Mitchell, Jun Haeng Lee, Muneesh Tewari, and Elizabeth M Swisher. MiR-221 and MiR-222 alterations in sporadic ovarian carcinoma: Relationship to CDKN1B, CDKN1C and overall survival. *Genes, Chromosomes & Cancer*, 49(7):577–584, July 2010. PMID: 20461750. 92
- [165] Stacia K Wyman, Rachael K Parkin, Patrick S Mitchell, Brian R Fritz, Kathy O'Briant, Andrew K Godwin, Nicole Urban, Charles W Drescher, Beatrice S Knudsen, and Muneesh Tewari. Repertoire of microRNAs in epithelial ovarian cancer as determined by next generation sequencing of small RNA cDNA libraries. *PloS One*, 4(4):e5311, 2009. PMID: 19390579. 92
- [166] Jessika Wynendaele, Anja BÄhnke, Eleonora Leucci, SÄren Jensby Nielsen, Irina Lambertz, Stefanie Hammer, Nadja Sbrzesny, Dana Kubitza, Anja Wolf, Elise Gradhand, Katharina Balschun, Ioana Braicu, Jalid Sehouli, Silvia Darb-Esfahani, Carsten Denkert, Christoph Thomssen, Stefan Hauptmann, Anders Lund, Jean-Christophe Marine, and Frank Bartel. An illegitimate microRNA target site within the 3' UTR of MDM4 affects ovarian cancer progression and chemosensitivity. *Cancer Research*, 70(23):9641–9649, December 2010. PMID: 21084273. 11
- [167] X Xiang, X Zhuang, S Ju, S Zhang, H Jiang, J Mu, L Zhang, D Miller, W Grizzle, and H-G Zhang. miR-155 promotes macroscopic tumor forma-

- tion yet inhibits tumor dissemination from mammary fat pads to the lung by preventing EMT. *Oncogene*, 30(31):3440–3453, August 2011. PMID: 21460854. 8
- [168] Changchun Xiao, Dinis Pedro Calado, Gunther Galler, To-Ha Thai, Heide Christine Patterson, Jing Wang, Nikolaus Rajewsky, Timothy P Bender, and Klaus Rajewsky. MiR-150 controls b cell differentiation by targeting the transcription factor c-myb. *Cell*, 131(1):146–159, October 2007. PMID: 17923094. 6
- [169] J Xiong, Q Du, and Z Liang. Tumor-suppressive microRNA-22 inhibits the transcription of e-box-containing c-myc target genes by silencing c-myc binding protein. *Oncogene*, 29(35):4980–4988, September 2010. PMID: 20562918. 67
- [170] Peizhang Xu, Ming Guo, and Bruce A Hay. MicroRNAs and the regulation of cell death. *Trends in Genetics: TIG*, 20(12):617–624, December 2004. PMID: 15522457. 6
- [171] Nuo Yang, Sippy Kaur, Stefano Volinia, Joel Greshock, Heini Lassus, Kosei Hasegawa, Shun Liang, Arto Leminen, Shan Deng, Lori Smith, Cameron N Johnstone, Xian-Ming Chen, Chang-Gong Liu, Qihong Huang, Dionyssios Katsaros, George Adrian Calin, Barbara L Weber, Ralf BÄijtzow, Carlo M Croce, George Coukos, and Lin Zhang. MicroRNA microarray identifies let-7i as a novel biomarker and therapeutic target in human epithelial ovarian cancer. *Cancer Research*, 68(24):10307–10314, December 2008. PMID: 19074899. 7

BIBLIOGRAPHY

- [172] Zhen Yang, Fei Ren, Changning Liu, Shunmin He, Gang Sun, Qian Gao, Lei Yao, Yangde Zhang, Ruoyu Miao, Ying Cao, Yi Zhao, Yang Zhong, and Haitao Zhao. dbDEMC: a database of differentially expressed miRNAs in human cancers. *BMC Genomics*, 11(Suppl 4):S5, December 2010. PMID: 21143814 PMCID: PMC3005935. 118
- [173] Rui Yi, Matthew N Poy, Markus Stoffel, and Elaine Fuchs. A skin microRNA promotes differentiation by repressing 'stemness'. *Nature*, 452(7184):225–229, March 2008. PMID: 18311128. 6
- [174] Zhenbao Yu, Zhaofeng Jian, Shi-Hsiang Shen, Enrico Purisima, and Edwin Wang. Global analysis of microRNA target gene expression reveals that miRNA targets are lower expressed in mature mouse and drosophila tissues than in the embryos. *Nucleic Acids Research*, 35(1):152–164, 2007. PMID: 17158157. 13
- [175] F. Zambelli, G. Pesole, and G. Pavesi. Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Research*, 37(Web Server):W247–W252, May 2009. 83
- [176] J Zhang, Y Yang, T Yang, Y Liu, A Li, S Fu, M Wu, Z Pan, and W Zhou. microRNA-22, downregulated in hepatocellular carcinoma and correlated with prognosis, suppresses cell proliferation and tumourigenicity. *British Journal of Cancer*, 103(8):1215–1220, October 2010. PMID: 20842113 PMCID: PMC2967065. 67

BIBLIOGRAPHY

- [177] Rui Zhang, Yi Peng, Wen Wang, and Bing Su. Rapid evolution of an x-linked microRNA cluster in primates. *Genome Research*, 17(5):612–617, May 2007. PMID: 17416744. 75
- [178] Jian-Jun Zhao, Jianhong Lin, Tint Lwin, Hua Yang, Jianping Guo, William Kong, Sophie Dessureault, Lynn C. Moscinski, Dorna Rezanian, William S. Dalton, Eduardo Sotomayor, Jianguo Tao, and Jin Q. Cheng. microRNA expression profile and identification of miR-29 as a prognostic marker and pathogenetic factor by targeting CDK6 in mantle cell lymphoma. *Blood*, 115(13):2630–2639, April 2010. PMID: 20086245. 101
- [179] Liang Zhou, Jiahao Chen, Zhizhong Li, Xianxin Li, Xueda Hu, Yi Huang, Xiaokun Zhao, Chaozhao Liang, Yong Wang, Liang Sun, Min Shi, Xiaohong Xu, Feng Shen, Maoshan Chen, Zujing Han, Zhiyu Peng, Qingna Zhai, Jing Chen, Zhongfu Zhang, Ruilin Yang, Jiongxian Ye, Zhichen Guan, Huanming Yang, Yaoting Gui, Jun Wang, Zhiming Cai, and Xiquing Zhang. Integrated profiling of microRNAs and mRNAs: microRNAs located on xq27.3 associate with clear cell renal cell carcinoma. *PloS One*, 5(12):e15224, 2010. PMID: 21253009. 87
- [180] Yamei Zhou, Christina Yau, Joe W. Gray, Karen Chew, Shanaz H. Dairkee, Dan H. Moore, Urs Eppenberger, Serenella Eppenberger-Castori, and Christopher C. Benz. Enhanced NF κ B and AP-1 transcriptional activity associated with antiestrogen resistant breast cancer. *BMC Cancer*, 7(1):59, April 2007. 87

BIBLIOGRAPHY

- [181] Dana Ziliak, Eric R Gamazon, Bonnie Lacroix, Hae Kyung Im, Yujia Wen, and Rong Stephanie Huang. Genetic variation that predicts platinum sensitivity reveals the role of miR-193b* in chemotherapeutic susceptibility. *Molecular cancer therapeutics*, June 2012. PMID: 22752226. 11