MicroRNA regulation in breast cancer – a Bayesian analysis of expression data

Viljami Aittomäki

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo, 7 October 2016

Thesis supervisor:

Professor Aki Vehtari

Thesis advisor:

Senior Researcher Rainer Lehtonen

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Viljami Aittomäki

Title: MicroRNA regulation in breast cancer – a Bayesian analysis of expression data

Date: 7 October 2016 Language: English Number of pages: 7+65

Department of Computer Science

Professorship: –

Supervisor: Professor Aki Vehtari

Advisor: Senior Researcher Rainer Lehtonen

MicroRNAs are a class of small, non-coding RNAs, which regulate gene expression post-transcriptionally. They downregulate genes by targeting messenger RNA transcripts and causing their degradation and inhibition of translation. Research has revealed microRNAs to participate in diverse cellular functions, such as differentiation and apoptosis, and many pathological processes, including cancer.

Identification of microRNA target genes is crucial in understanding their function in cell biology and disease. A wide range of methods have been proposed for computational prediction of microRNA targets. Early target prediction methods used sequence information, while recent tools have integrated expression measurements of target genes and microRNAs. A limited number of studies have integrated protein, gene and microRNA expression for target prediction.

Breast cancer is the most common cancer in women and a significant cause of morbidity and mortality globally. Analyses of gene expression data have provided insight into the pathogenesis of breast cancer, and intrinsic subtypes correlating with prognosis have been identified. A range of microRNAs have been indicated to contribute to breast cancer pathogenesis.

In this thesis, a recent Bayesian variable selection method was applied for uncovering putative microRNA targets in breast cancer. The proposed model integrated protein, gene and microRNA expression data. Results were compared with another popular prediction method. Analyses showed that the proposed method is applicable to microRNA target prediction. Limitations and refinements of the method and study are discussed, and the importance of an integrative approach is highlighted.

Keywords: Bayesian analysis, breast cancer, gene expression, microarray, microRNA, target prediction

AALTO-YLIOPISTO
SÄHKÖTEKNIIKAN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Viljami Aittomäki

Työn nimi: MikroRNA-säätely rintasyövässä – ekspressiodatan bayesilainen analyysi

Päivämäärä: 7 October 2016 Kieli: Englanti Sivumäärä: 7+65

Tietotekniikan laitos

Professuuri: –

Työn valvoja: Professori Aki Vehtari

Työn ohjaaja: Dosentti Rainer Lehtonen

MikroRNA:t ovat lyhyitä RNA-molekyylejä, jotka säätelevät geenien ekspressiota sitoutumalla lähetti-RNA-molekyyleihin estäen siten niiden translaation proteiiniksi. Aiemmat tutkimukset ovat osoittaneet, että mikroRNA:t osallistuvat monipuolisesti solujen toiminnan säätelyyn, kuten erilaistumiseen ja apoptoosiin, ja ovat osallisena monien tautien, kuten syövän synnyssä.

MikroRNA:n säätelemien kohdegeenien tunnistaminen on olennainen askel mikroRNA:n toiminnan ymmärtämisessä. Kohdegeenien ennustamiseen on kehitetty lukuisia laskennallisia menetelmiä. Varhaiset menetelmät perustuivat RNA-sekvenssien vertailuun. Uudemmat työkalut yhdistävät geeni- ja mikroRNA-ekspressiodataa kohdegeenien tunnistamiseksi. Proteiini-, geeni- ja mikroRNA-ekspressiota yhdistäviä kohdegeenien tunnistamiseen tähtääviä tutkimuksia on julkaistu toistaiseksi suhteellisen vähän.

Rintasyöpä on naisten yleisin syöpä ja merkittävä sairastavuuden ja kuolleisuuden aiheuttaja maailmanlaajuisesti. Geenien ekspressiodatan analysointi on lisännyt tietoa rintasyövän synnystä, ja geenien ekspressioon perustuen on kyetty tunnistamaan rintasyövän alatyyppejä, jotka korreloivat syövän ennusteeseen. MikroRNA:n on todettu olevan osatekijä rintasyövän synnyssä.

Tässä diplomityössä sovellettiin äskettäin julkaistua bayesilaista muuttujavalintamenetelmää mikroRNA-molekyylien kohdegeenien ennustamiseen rintasyövässä. Tähän tarkoitukseen käytettiin proteiini-, geeni- ja mikroRNA- ekspressiodataa. Tulokset osoittivat, että menetelmä soveltuu kohdegeenien ennustamiseen. Työssä esitetään vaihtoehtoja mallin jatkokehittämiseksi.

Avainsanat: bayesilainen analyysi, geenien ekspressio, kohdegeenien ennustaminen, mikroRNA, mikrosiru, rintasyöpä

Preface

This Master's Thesis and the related research were done in the Systems Biology of Drug Resistance in Cancer research group within the Genome-Scale Biology research program of the University of Helsinki. I am grateful to Professor Sampsa Hautaniemi for suggesting this work and providing facilities and also for several years of interesting work in his group.

I want to thank Professor Aki Vehtari for supervision and suggesting the method proposed in this work as well as for advice in implementing it. I would like to thank my advisors Professor Sampsa Hautaniemi and Senior Researcher Rainer Lehtonen for their guidance and trust and the time they have invested into this work. I want to thank Juho Piironen for help with the modeling. I would also like to thank Professor Antti Vaheri for discussions.

I owe a lifetime of gratitude to my parents for their endless love and encouragement. Even in the darkest of moments, they have never stopped believing in me. I also want to thank my mother for discussions relating to genetics.

Finally, I wish to send a thank you and all of my love to all my family and friends for their support during the writing of this thesis and in other endeavors in my life. Special thanks to my friend Heidi for pushing me forward and sharing cookies.

Espoo, 7 October 2016

Viljami Aittomäki

Contents

Symbols and abbreviations

Symbols

β	Regression coefficients for explanatory variables
$\|\beta\|_1$	The 1-norm of β
$\mathrm{E}_\theta(x)$	Expectation of random variable x over parameters θ
n	Number of observations
$\mathrm{N}(\mu, \sigma^2)$	Multivariate normal distribution with mean μ and variance σ^2
p_n	Number of (assumed) true explanatory variables in variable seĺection
$p(y)$	Probability density of x
$p(y, \theta)$	Joint probability of y and θ
$p(y\|\theta)$	Conditional probability of y given θ
θ	Parameters of a probability model
x	Explanatory variable (or microRNA expression vector)
X	Matrix of explanatory variables (or microRNA expression vectors)
y	Outcome variable (or protein expression vector)
$y \sim \mathrm{N}(.)$	Random variable y has probability distribution $N(.)$
$y \propto x$	y is proportional to x, up to a constant factor
z	mRNA expression vector

Abbreviations

bp	Base pairs (as a measure of double-stranded sequence length)
HMM	Hidden Markov model
lasso	Least absolute shrinkage and selection operator
LPD	log predictive density
MCMC	Markov chain Monte Carlo
miRNA	MicroRNA
MLPD	Mean log predictive density
MLR	Multivariate linear regression
mRNA	Messenger RNA
NGS	Next-generation sequencing
nt	Nucleotides (as a measure of sequence length)
PCR	Polymerase chain reaction
PPVS	Projection predictive variable selection
qPCR	Quantitative PCR
RISC	RNA-induced silencing complex
RNAi	RNA interference
RPPA	Reverse-phase protein array
SVM	Support vector machine
UTR	Untranslated region at the beginning (5') or end (3') of a messenger RNA
WHO	World Health Organization

1 Introduction

MicroRNAs (miRNAs) are short single-stranded RNA-molecules, that act in post-transcriptional regulation of gene expression [7]. They have been found in a wide variety of animals and plants, and also in viruses. MicroRNAs are highly conserved in evolution and function in diverse developmental, physiological and pathological processes. miRNAs have also been indicated in the formation of numerous diseases, including several types of cancer [12]. Therefore, the study of microRNAs and their function in cancer can offer insights into tumorigenesis and cancer progression as well as potential new biomarkers and treatments.

Identifying the target genes regulated by microRNAs is key to understanding their function, both in cellular physiology and pathology. Experimental laboratory studies to identify miRNA targets are both laborious and costly. In fact, finding all miRNA targets by experimental studies alone is unfeasible, considering that any gene is potentially targeted by any miRNA, giving rise to tens of millions of potential interactions.

A plethora of methods for computational identification of microRNA target genes have, thus, been developed [77]. Early methods were based on sequence similarity of the microRNA and messenger RNA (mRNA) of putative target genes. More recent methods compare the expression profiles of miRNAs and mRNAs in cell cultures or tissue samples to elucidate interactions transpiring in cells. Most of these expression-based methods rely on variations of correlation or multivariate linear regression, though more complex models have also been proposed. The combinatorial nature of miRNA action, however, makes expression-based strategies difficult, as most transcripts are regulated by several miRNAs simultaneously, the contribution of individual miRNAs may be small, and most miRNAs target a large number of transcripts [6].

Cancer is a genetic disease caused by mutations in the genome of tumor cells [46]. Some of these mutations can be inherited, while some arise during the life-time of an individual. To understand how cancer develops, it is of paramount importance to identify genes which contribute to the tumorigenesis. It is also essential to determine genes, and other factors, that influence cancer aggressiveness as well as treatment sensitivity and resistance so that better and more targeted treatments can be developed.

Breast cancer is the most common of female cancers and causes remarkable morbidity and mortality word-wide. Annually more than 1.5 million women develop breast cancer [34]. Thus, breast cancer constitutes a major global health problem. Previous studies comparing the expression profiles of normal breast tissue and breast cancer tissue have revealed that expression signatures classify breast cancers into distinct subtypes, which are associated with prognosis [94]. Recent studies have suggested that microRNAs can explain some of the heterogeneity and pathology of breast cancer and show promise as prognostic markers [106].

The aim of this thesis was to apply a recently proposed Bayesian variable selection method to microRNA target discovery in breast cancer. Variable selection was applied in the context of Bayesian regression, incorporating protein, mRNA and miRNA expression profiles, to elucidate which microRNAs are relevant in regulating protein expression levels. The prediction results were compared with lasso regression, a popular method for target prediction, and with experimentally validated microRNA targets.

2 Gene expression

Genetic information is encoded in deoxyribonucleic acid (DNA). A gene is a section of DNA that serves as a template for a functional ribonucleic acid (RNA) molecule. Gene expression refers to this process of synthesizing a functional end-product from the information contained in gene. DNA and gene expression serve as the basis of all currently known life [97].

Most of gene expression is dedicated to production of proteins. The Central Dogma of Molecular Biology, postulated by Francis Crick in 1970, describes the general schema of how genetic information flows from genes to proteins; DNA is first transcribed into messenger RNA (mRNA), which is then translated into polypeptides, which ultimately form proteins [20] (illustrated in Figure 1). The flow is not strictly one-directional, though, as reverse transcriptases, a family of enzymes, can synthesize DNA from an RNA template.

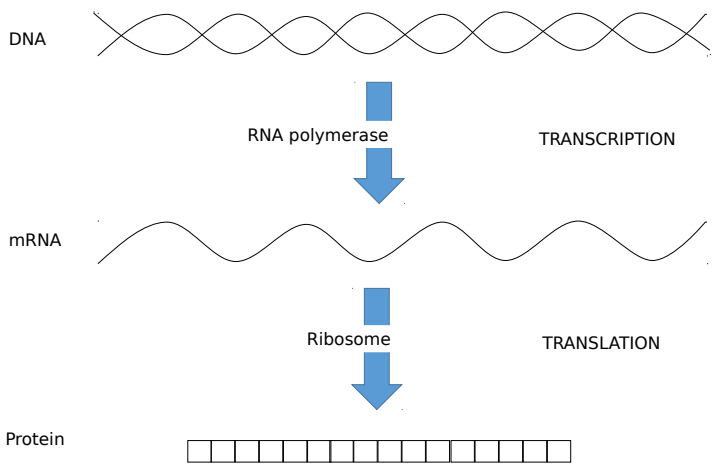


Figure 1: The Central Dogma of Molecular Biology, as postulated by Crick [20]. Most genetic information flows from DNA to RNA to protein.

All genes are transcribed to RNA, but all do not encode proteins. The human genome[1] has been suggested to contain approximately 20 500 protein-coding genes, which encompass only around 1.5% of the whole genetic sequence [19]. The vast majority of the human genome was, thus, previously thought to be without function and referred to as "junk DNA". More recently, however, it has become evident that human DNA is pervasively transcribed, the majority of it appears functional, and many coding and non-coding regions overlap in the DNA [97].

Non-coding genes give rise to non-coding RNA (ncRNA), a class of RNA molecules that both participate in and regulate the expression of other genes. Examples of known ncRNAs and their functions are presented in Table 1. Nevertheless, the

[1]The genome refers to the whole genetic material of an organism or an individual.

Table 1: Examples of known major classes of human non-coding RNA and their general functions. This table is not exhaustive and several additional classes have been discovered. Table adapted from [97].
Size: approximate sequence lengths of each class in number of nucleotides. Abbrev.: abbreviations commonly used.

RNA class	Abbrev.	Size (nt)	Function
Ribosomal RNA	rRNA	120–5,000	Components of ribosomes (which perform translation)
Transfer RNA	tRNA	70–80	Transporting peptides and decoding mRNA sequence into peptides
Small nuclear RNA	snRNA	60–360	Intron splicing; regulation of transcription, chromosomal replication and cell cycle etc.
MicroRNA	miRNA	21–24	Post-transcriptional gene regulation
Small interfering RNA	siRNA	21–22	Post-transcriptional regulation
Long non-coding RNA	lncRNA	$>$ 1,000	Gene regulation at several stages

function and significance (if any) of many transcribed and non-coding regions of the genome still remains poorly understood.

2.1 Regulation of gene expression

The proper regulation of gene expression is paramount for cells to respond to external signals, changes in their environment, and to go through different developmental stages. Gene expression is, thus, under complex control mechanisms, which result in tissue and cell-specific expression. This regulation occurs on several stages including transcriptional, post-transcriptional, translational and post-translational regulation [97].

The first step of this regulation is control of transcription. To be transcribed, genes need active initiation of transcription, which occurs in the promotor regions. While the actual transcription is performed by RNA polymerases, many different transcription factors and regulatory proteins participate in its regulation. Transcription is possible only when the chromatin structure of the transcribed genetic region is opened from its tight package around histone proteins, which is controlled by additional factors such as histone acetylation.

The produced mRNA undergoes post-transcriptional modifications, such as capping, polyadenylation and splicing of introns, which all are essential for further translation of the mRNA to protein [97]. All these steps must maintain a certain level of fidelity, as even slight structural changes can render both mRNA and protein to degradation and. The main post-transcriptional control mechanism seems to be RNA interference (RNAi). It causes suppression of gene expression through mRNA degradation and inhibition of translation. MicroRNAs (miRNAs) and small interfering RNAs (siRNAs) are central components of the RNAi pathway; they act as target mRNA recognizing templates [24]. MicroRNAs, which are the focus of this study, are discussed in more detail in the next section.

The translation of mRNA to protein can also be directly regulated, but this seems less prevalent than control of previous phases. Post-translationally, proteins can be modified and degraded to affect their function and cellular expression levels [97].

2.2 Quantification of gene expression

The quantitative measurement of gene expression can be done on either the level of messenger RNA molecules or protein molecules. Although proteins are the eventual effector molecules – at least for protein-coding genes – gene expression is usually thought to be synonymous with mRNA expression. mRNA abundances are significantly easier to measure than protein abundances, due to the chemistry of hybridization and the relative ease of replicating DNA and RNA sequences by exploiting cellular machinery evolved for this purpose.

Quantitative PCR (qPCR) is a DNA/RNA measurement method based on the polymerase chain reaction (PCR). It measures the number of specific DNA or RNA molecules present in a sample during each cycle of the PCR amplification process. These are then extrapolated to obtain gene expression values. qPCR has been considered a golden standard for measuring gene expression and is widely used in research, but also in clinical diagnostics. It is often the method of choice for measuring a moderate number of genes, but does not scale well into large numbers of genes [107].

Gene expression microarrays are based on probes printed on a solid surface, where each probe has been designed to have complementary sequence corresponding to a target mRNA. The amount of mRNA hybridized to the probes provides an estimate of gene expression. The advantage of microarrays is that they allow massively parallel analysis on the whole genome level. They are also relatively inexpensive and easy to use, making them a ubiquitous tool for expression measurements. Detection of expression levels is based on fluorescence and optical sensors and is subject to noise due to imperfect probe design and technical limitations. As microarray data are inherently noisy, proper normalization methods have been shown to be important. Probe designs can also become obsolete as reference genomes are updated and, therefore, reassessment of the true targets of probes is advisable [2].

More recently next-generation sequencing methods have been applied to gene expression profiling. These are not dependent on previous reference sequences, but are relatively expensive and laborious compared to microarrays.

Protein expression can be measured using several different methods. Perhaps most widely used are different variations of mass spectrometry (MS). Application of mass spectrometry is limited by its resource-intensiveness and poor scalability, however. Reverse-phase protein arrays (RPPA) are a platform comparable to microarrays, where samples are fixed to a solid surface and then probed with antibodies binding to specific proteins [13]. This allows measuring a single protein for several samples simultaneously. RPPAs are inexpensive, allow reasonably large-scale analyses, and

analysis of RPPA data is similar to gene expression arrays, making them an attractive choice for studies using multiomics data [71].

The general assumption has been that mRNA expression is representative of gene expression and that changes in mRNA abundances also reflect changes in protein abundances. This assumption has recently been challenged by experiments indicating that correlations between the expression of mRNA and corresponding protein are low, with mRNA expression explaining around 40% of variation in protein expression [112]. Payne recently concluded that "proteome and transcriptome abundances are not sufficiently correlated to act as proxies for each other" and that most of this difference is likely caused by biological regulation and not by measurement technology [82]. Therefore, it is beneficial, even necessary, to integrate measurements from different stages of gene expression – for example mRNA, microRNA and protein abundances – to gain better and novel insight into biological processes.

3 MicroRNAs

MicroRNAs (miRNAs) are a class of endogenous (i.e. synthesized within the cell) non-coding small RNA molecules that function as post-transcriptional regulators of gene expression [3]. In their functional, mature form miRNAs are single stranded and approximately 22 nucleotides long. MicroRNAs are not translated into protein. Instead, they have an important role in regulation of gene expression in a wide range of physiological, developmental and pathological processes [8]. MicroRNAs assert their regulatory function by destabilization and degradation of target messenger RNA (mRNA) molecules and inhibition of mRNA translation [33].

3.1 Discovery of microRNAs

The first known microRNA, lin-4, was discovered in 1993 by two research groups studying the larval development of the nematode *Caenorhabditis elegans*. The researchers noted that lin-4 does not encode a protein, but instead produces a pair of small RNAs, the longer of which was proposed to be a precursor to the shorter one [65]. The RNAs encoded by lin-4 were noted to have conserved antisense complementarity in several sites of an untranslated region of the lin-14 mRNA, and these sites were found to be necessary for the normal repression of lin-14 expression by lin-4 [65, 115].

Let-7, the second microRNA to be discovered, was also first found in *C. elegans*, however, homologues of let-7 were later found in several other species [81]. Soon after, numerous microRNA genes were found across a variety of species, and a registry was set up to serve as a comprehensive knowledge base of published microRNAs and as an independent authority on microRNA nomenclature [42]. This registry later became miRBase, the de facto reference database of known microRNAs, and now provides sequence data, annotations, as well as links to databases of predicted and validated target genes for miRNAs [59].

3.2 MicroRNA genomics

The number of known small RNAs has since vastly expanded and microRNAs have been found in more than 200 organisms, including all studied animals, plants [54] and viruses [43]. The number of records in miRBase has risen exponentially to 35,828 mature miRNAs for 223 species (including 2,588 human miRNAs) in the most recent version (v21, released June 2014 [75]). This illustrates the large number of novel microRNA molecules discovered recently, which has been mainly due to increasing efforts in and availability of sequencing. miRBase lists 2,588 known human miRNAs at the time of writing this thesis.

MicroRNAs are highly conserved in evolution [7]. For instance, approximately 55% of *C. elegans* miRNAs have homologues in humans [52]. Interestingly, the appearance

of multi-cellular organisms appears to co-occur with the appearance of the microRNA machinery. Organism complexity and speciation also seem to correlate with miRNA complexity, together suggesting that microRNAs have had a crucial role in the development of complex organisms [63].

MicroRNAs are found in varying genomic contexts in the DNA. Approximately 50% of mammalian miRNAs are located in close proximity to other miRNAs and form polycistronic miRNA clusters that are transcribed simultaneously. Some miRNAs reside in the genome as dedicated miRNA genes, with their own promotor regions. [57] miRNAs and miRNA clusters can be situated in exons or introns of non-coding genes and some are found in introns of protein-coding genes [24].

MicroRNAs are expressed in all tissues, however, different tissues have differing miRNA expression profiles [61]. Many microRNAs also have differing expression in different developmental stages of an organism, often functioning as molecular switches to move between stages. For instance, let-7 functions to control the transition from late larval to adult stage in *C. elegans* [7].

3.3 MicroRNA biogenesis

The canonical pathway of microRNA biogenesis is illustrated in Figure 3 and is presented here as reviewed by Bartel [7], Melo and Esteller [73], Ha and Kim [45], and many others.

Most microRNAs are transcribed from genomic DNA by RNA polymerase II to form a long primary microRNA (pri-miRNA) molecule [67]. The pri-miRNA molecule contains a hairpin structure, with a 33-bp double-helix stem and a terminal loop, and flanking single-strand sequences, which are several hundreds or thousands of nucleotides long [56].

The pri-miRNA is cut by the ribonuclease Drosha to form a pre-microRNA (pre-miRNA), which consists of the hairpin and is approximately 70 nt long [66]. Examples of typical pre-miRNA structure are shown in Figure 2. Drosha is aided by its cofactor DGRC8 and they form a complex known as the Microprocessor [41]. The hairpin is then exported from the nucleus to the cytoplasm by Exportin 5 (XPO5), a member of the nuclear transport receptor family [70].

In the cytoplasm, the ribonuclease Dicer cleaves out the loop of the hairpin to form a 22-nt-long double-stranded miRNA:miRNA* duplex corresponding to the stem of the hairpin [9]. Dicer associates with a cofactor, in humans TRBP (Tar RNA-binding protein), which is not required for effective dicing of the pre-miRNA, but acts to physically bridge the Dicer to an Argonaute protein [16].

The duplex is then bound by the Argonaute protein, in mammals one of Ago1 through Ago4, forming what is called the RNA-induced silencing complex (RISC). The RISC is a protein complex containing Dicer, TRBP and Ago [40]. Ago, aided by Dicer and TRBP, unwinds the strands of the duplex and retains one of them. The retained

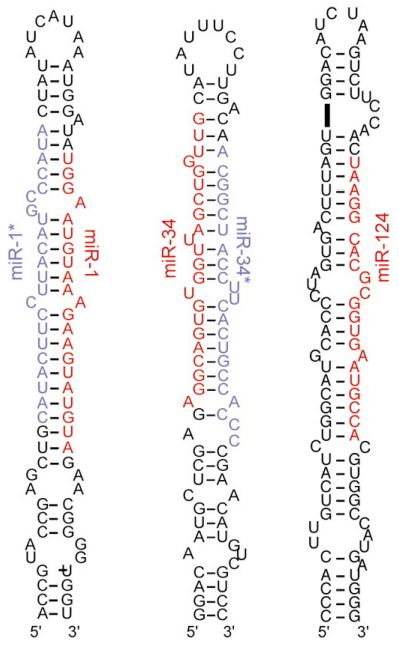


Figure 2: Hairpin structure of three pre-miRNAs from *C. elegans*. Red and gray colors indicate the sequences of mature miRNAs. Reprinted with permission from [7].

strand is known as the guide strand (or miRNA). The other strand, called the passenger (or miRNA*), is released and typically degraded [24]. In some instances, either one of the strands can become the guide or both can be used [21].

Not all miRNAs are generated through this canonical pathway of microRNA biogenesis. Some miRNAs are not dependent on Drosha, such as mirtrons, which are cut into pre-miRNA by the spliceosome, a molecular complex responsible for removing introns (and sometimes exons) from precursor mRNA [89]. The biogenesis of miR-451 is independent of Dicer; miR-451, which has an important role in erythropoiesis, is cleaved by Ago2 [14].

3.4 MicroRNA mechanism of action

RISC is the effector of RNA interference, and Ago functions as its catalytic engine. MicroRNA sequence guides the RISC to target messenger RNAs [35]. Figure 3 illustrates a rough outline of how miRNAs act to regulate mRNA expression.

Target recognition is based on sequence complementarity of the miRNA and mRNA. In animal miRNAs this complementarity is almost always limited [3]. Nucleotides at positions 2-8 of the 5' end of the microRNA have been found crucial to target mRNA matching; these nucleotides are termed the miRNA "seed sequence". miRNA target sequences are mostly located in the 3' UTR (untranslated region) of the mRNA transcript, but in some instances target sites also reside in the coding region or 5' UTR of the mRNA [8].

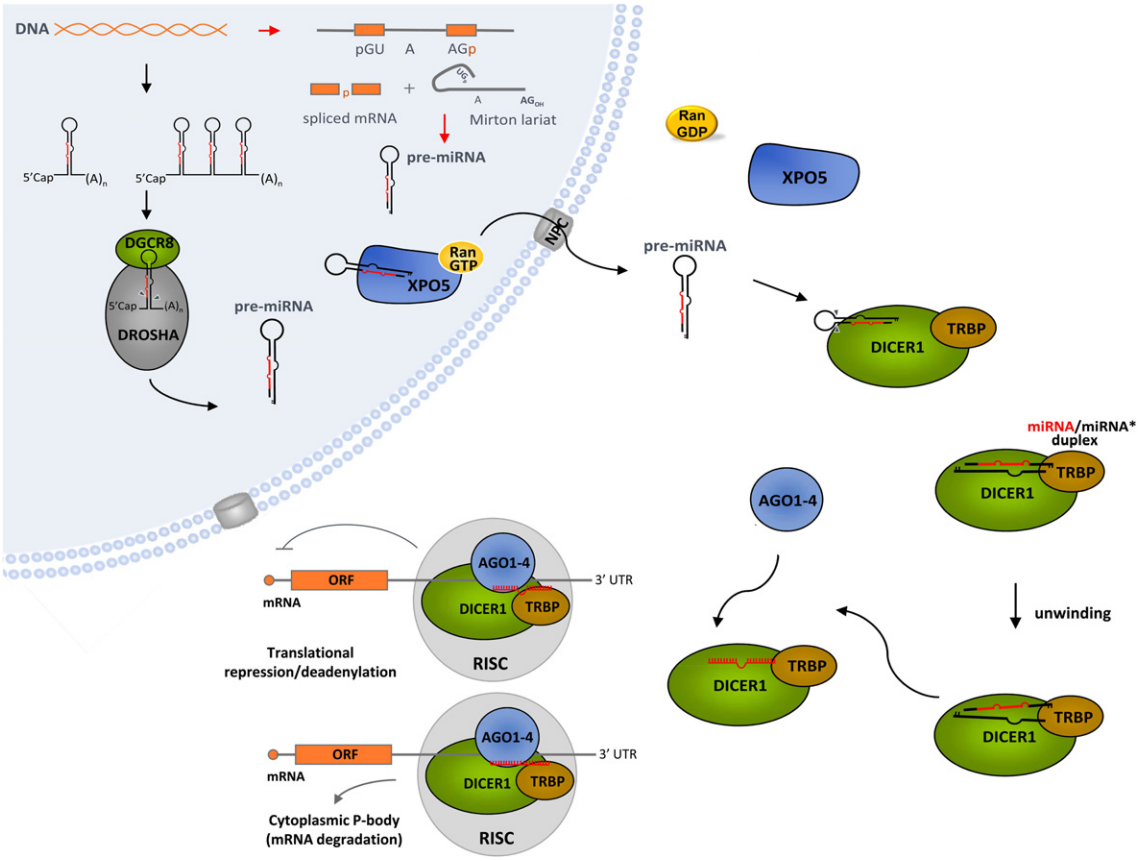


Figure 3: Depiction of the canonical (and mirtron) pathway of microRNA biogenesis and microRNA mechanism of action. miRNA biogenesis begins in the nucleus, where the pri-miRNA is transcribed, then cut by Drosha to the pre-miRNA and exported into the cytoplasm by Exportin 5 (XPO5). Mirtrons do not require Drosha processing. The pre-miRNA is then bound by Dicer, which (aided by TRBP) cuts and unwinds the miRNA into its mature form. RISC is then formed and, guided by the miRNA, regulates gene expression by translational repression or degradation of mRNA. See text for more details. Black arrows depict the movement of miRNA molecules through the process, and gray arrows the action of RISC on the target mRNA. ORF: open reading frame (the protein-coding section) of mRNA. Figure reprinted with permission from Melo and Esteller [73].

MicroRNAs act through inhibition of mRNA translation or destabilization and subsequent degradation of mRNA. The exact mechanisms by which the miRNA and Ago induce translational repression or destabilization of mRNA are unclear [35]. Translational inhibition was earlier believed to be the major form of miRNA action in animals, but recent evidence suggests that mRNA destabilization predominates [44]. Rarely, the mRNA can be directly cleaved by Ago [24].

mRNAs bound to RISC accumulate in so called processing bodies (P-bodies), which are known sites of mRNA catabolism and translational repression in the cytoplasm. The localization in P-bodies, however, appears to be a consequence of RNA silencing, not the cause, and is reversible [32].

Interestingly, several alternative mechanisms of action for microRNA have been reported, illustrating the complexity and diversity of microRNA biology and gene regulation. For example, some miRNAs can increase the translation of target mRNA instead of repressing it [109], miR-373 was found to target DNA promoter areas and act to induce gene transcription [86], and miR-328 targets a protein to prevent inhibition of mRNA translation [29].

3.5 MicroRNA function

MicroRNAs assert extensive control over the transcriptome, and have been found to participate in regulation of almost all studied cellular processes, including embryo development, cell proliferation and differentiation, apoptosis, and metabolism. More than 60% of human mRNA transcripts are predicted to be regulated by miRNAs and most have target sites for several different miRNAs [36]. Furthermore, a single microRNA can have as many as hundreds or thousands of target mRNAs.

The effect of a single miRNA on the expression of its target tends to be subtle [6]. Thus, microRNAs are considered fine-tuners of gene expression. However, the modest effect can be enhanced by multiple binding sites and multiple miRNAs acting on the same target, enabling synergistic interactions [8].

It should be noted, however, that the functional role and importance of many miRNA-mRNA interactions are unknown, even for validated interaction pairs. Uncovering these roles is challenging due to the subtle regulatory effects miRNAs have and, additionally, because of the complexity and robustness of many cellular regulatory networks [8]. Furthermore, experimentally validated targets have been recognized for only a fraction of all known microRNAs. Nonetheless, discovering miRNA targets is a critical step in understanding their function.

3.6 Quantification of microRNA expression

The same methods that are employed for quantifying mRNA expression are generally also applicable to measuring microRNA expression, and the three principal methods used are qPCR, microarrays and next-generation sequencing [50]. However, as Hunt and colleagues in a recent review point out, there are several challenges in detecting miRNA expression in particular [51].

MicroRNAs are very short and typically comprise approximately 0.01% of RNA typically extracted from any tissue sample. This implies that miRNA detection methods must be highly sensitive. Additionally, microRNAs from the same family can differ by only one base, which in turn requires high specificity to distinguish between members of the same miRNA family. On the other hand, variation in miRNA processing can result in slight sequence variations, or isoforms, of a single miRNA, also known as isomiRs [64]. This means high specificity or an incorrect reference sequence (e.g. that of a weakly-expressed isomiR) used for detection can

cause inaccurate measurements. IsomiRs may also have different functions resulting from altered target specificity [18].

These issues mean that miRNA expression data measured with microarrays are often quite noisy. Proper filtering and normalization techniques are, therefore, necessary in analyses of such data. A review of different miRNA microarray platforms and preprocessing methods has been written by Sah et al [91].

Many of these methodological limitations are resolved by next-generation sequencing (NGS) approaches, which are sensitive and reliable in quantifying known miRNAs and enables identification of novel ones [50]. Sequencing can detect variations of single nucleotides and does not depend on previously identified sequences. However, not all identified short RNAs are functional miRNAs, and NGS conveys its own set of problems relating to significant computational complexity and validation efforts to distinguish relevant data from noise [51].

4 Cancer

Cancer is a disease of uncontrolled overgrowth of a population of cells. It is generally viewed as a genetic disease, albeit it is mostly not inherited, as it is caused by mutations in the genome of the tumor. These mutations cause malfunction and dysregulation of the genetic machinery regulating cellular functions, such as cell proliferation, differentiation and apoptosis, resulting in unregulated growth and malignant tumor formation.

There are several classes of genes that influence tumor growth, the two main categories being oncogenes and tumor suppressors. Oncogenes, first identified in retroviruses, promote tumor growth by overexpressing their gene product, leading to for example abnormal cell-cycle control and increased cell division. [108]. Tumor-suppressor genes are often regulators of cell proliferation and inactivation of these genes can lead to tumor progression. The existence of tumor suppressors was first hypothesized by Alfred Knudson [58]. Knudson also formed the "two-hit hypothesis", which suggests that, for cancer to develop, both copies of a tumor suppressor gene should become inactivated and that in inherited cancers one mutation is acquired in the germ-line and the other occurs in somatic cells.

The Hallmarks of Cancer are a set of six features which tumors often acquire to become malignant The hallmarks were suggested Douglas Hanahan and Robert Weinberg [46] in their seminal article in 2000, and consist of sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis. Hanahan and Weinberg postulated that at least three of these six features are required for invasive cancer to develop.

Recently, Hanahan and Weinberg [47] revised the Hallmarks with two new hallmarks and two enabling characteristics. The new hallmarks are deregulation of cellular energetics and avoiding immune destruction. The enabling characteristics of malignant tumors are genome instability and mutation, and tumor-promoting inflammation through recruitment of the immune system. Genome instability and mutation are decidedly important features, as much of cancer research focuses on identifying mutated or dysregulated genes that promote tumor progression.

4.1 Breast cancer

Breast cancer constitutes a significant health issue globally. It is the most common cancer in women and the second most common cancer overall; approximately 1.7 million women develop breast cancer annually world-wide, and in 2012 there were 522,000 breast cancer-related deaths [34]. In Finland there were 5,008 new breast cancer cases and 815 breast cancer-related deaths in 2014 [98].

Most breast cancers are sporadic; only 5-7% of breast cancer cases are of familial type [72]. However, 15-30% of breast cancer patients have a family member or relative

with breast cancer. This is mostly due to the high frequency of breast cancer in many western populations, but also suggests that there are unknown genetic factors and environmental factors that have an impact on breast cancer development. Indeed, breast cancer is a hormone-related disease and hormonal factors, most importantly estrogens, are known to have an impact on breast cancer development.

The most common hereditary forms of breast cancer are related to mutations in the breast cancer genes *BRCA1* and *BRCA2*, which explain about 25% familial breast cancer in many populations. These, however are rare on the population level and familial clustering of beast cancer is multifactorial and caused by moderate risk and low risk genetic variations, which are much more common [72].

4.1.1 Breast cancer classification

Breast cancers are heterogeneous in their nature and classified by morphology (the microscopic structure of cancer tissue). The morphological classification of breast cancer is based on the WHO classification from 2003 and includes altogether 19 histological subtypes of invasive breast cancer [99, 114]. Of these, invasive ductal carcinoma is by far the most common. Additionally, the WHO classification of breast cancer includes the TNM classification (consisting of the size of the primary tumor (T), whether the tumor has spread to lymph nodes (N) or metastasized (M) into other tissues), and grading into well, moderately or poorly differentiated tumors based on microscopical examination, with less differentiated tumors having worse prognosis [99].

In the clinic several other tumor characteristics, such as patient age, and the expression of estrogen (ER) and progesterone receptors (PR) and Her2, are used to group patients into prognostic categories. Treatment regimens can be chosen using for example the St Gallen criteria [39], which classifies tumors into highly endocrine responsive, incompletely endocrine responsive, and endocrine non-responsive based on the expression of estrogen and progesterone receptors in tumor cells and, consequently, the response to endocrine treatment.

One of the problems with morphological classification is that over 50% of tumors do not show any particular features, although tumors in this large group have highly variable outcomes. Additionally, endocrine-responsive tumors can become non-responsive as tumor cells accumulate genomic mutations [78]. Therefore, personalized molecular diagnostics are required to tailor targeted treatments.

More recently, expression profiling has led to a suggestion of new classification of breast cancers. Studying the expression profiles of breast tumors, Perou et al [83] distinguished four subgroups based on gene expression, namely ER+/luminal-like, basal-like, Erb-B2+ and normal breast, The subgroups were shown to correlate with prognosis and a 50-gene test (PAM50) was devised to classify tumors into the subtypes [80].

Other prognostic tests based on expression signatures, including Oncotype DX (a

21-gene recurrence score) and MammaPrint (a 70-gene test), have also been proposed to help to determine the need of adjuvant chemotherapy. These tests have been suggested to be valid and promising, but their utility in clinical decision making remains unclear [5].

4.2 MicroRNAs and cancer

The dysregulation of microRNAs is associated with many human diseases, such as neurological disorders, diabetes and cancer [53]. A disease-promoting role for miRNAs has been implicated in many different cancers [73], including breast cancer, and research has shown microRNAs to have important roles in tumor initiation, progression and metastasis [68].

MicroRNA expression signatures correlate with numerous cancer features, such as tissue of origin, progression, prognosis and treatment response, and all studied cancers have had miRNA expression profiles differing from healthy tissue [12]. In fact, microRNAs appear to be generally underexpressed in cancers [69]. Therefore, it seems clear that microRNAs participate in many of the pathways resulting in the hallmarks of cancer, and examples of miRNAs influencing each of the hallmarks have been found. Similarly to protein-coding genes, microRNAs can function as tumor suppressors or oncogenes [68]. For instance, in a meta-analysis of dysregulation of miRNAs in breast cancer, van Schooneveld et al [106] found five major oncogenic and nine major tumor suppressive microRNAs.

The genetic mechanisms for microRNA involvement in cancer are varied, including mutations in miRNA or target mRNA sequence, chromosomal rearrangements of the miRNA-encoding DNA regions and epigenetic changes in DNA methylation or histones, leading to aberrant miRNA expression [12, 73]. For example, a single-nucleotide polymorphism in the microRNA miR-196a2 has been found to be associated with breast cancer risk [37]. A mutation in the sequence of estrogen receptor alpha, in the target site of miR-453, has been suggested to be associated with a lower breast cancer risk [100], an example of a mutation in a target transcript affecting miRNA function. MicroRNA function can also be altered by abnormalities in the miRNA-processing machinery. For instance, a mutation in the Dicer gene causes a tumor predisposition syndrome known as DICER1 syndrome [93]. Another example of this is apparent dysregulation of Dicer and Drosha in breast cancer [116].

The different subtypes of breast cancer, explained above, reflect the genetic background of the tumor and, accordingly, the subtypes differ in their gene expression profiles. This also applies for miRNA expression, the different intrinsic subtypes have different miRNA expression profiles, suggesting their importance in breast cancer evolution [11]. de Rinaldis et al [23] identified a 46-miRNA signature that could be used in differentiating the intrinsic subtypes from each other. In addition to tumor development, many miRNAs have been found to modulate the response to breast cancer therapies. These include chemotherapy, antiendocrine therapy, radiotherapy

and targeted therapies.

Accordingly, miRNAs have been studied as biomarkers for diagnosing cancer and cancer prognosis. Emmadi et al [30] recently found let-7 expression to be negatively correlated with the Oncotype DX recurrence score in breast cancer. This corroborated with the earlier finding of let-7 being downregulated in breast cancer stem cells (tumor cells possessing the ability of self-renewal) [117] and later research suggesting let-7 to act as a tumor suppressor. Several miRNAs have also been associated with breast cancer metastasis [15].

MicroRNAs also show promise as a novel therapeutic tool. Several studies have tested miRNA-based cancer treatments in animal models with encouraging results [105]. However, more research in this area is needed before microRNA treatments are ready for the clinical setting.

5 Computational identification of miRNA targets

Recognizing the targets of microRNAs is essential in understanding their biological function and role in disease. Many target interactions have been found in experimental laboratory studies. Common methods for such studies include using cell lines and introducing exogenous miRNAs by transfection or suppressing endogenous ones and measuring the effect on mRNA or protein expression. For a detailed review of experimental methods, see Thomson et al [101].

Several public databases list currently known experimentally validated microRNA targets. Examples include DIANA-TarBase [111] and mirTarBase [17], which are both manually curated from published literature, and MiRWalk [27], which combines data from several other databases using text mining.

Although recent advances in high-throughput methodologies, such as CLIP-seq, have significantly increased the scale of experimental studies, experimental identification of microRNA targets remains laborious and costly, and many methods still rely on computational processing of results [111]. To this end, a wide range of computational tools have been developed to aid in miRNA target discovery.

Computational approaches to target prediction can be roughly classified into solely sequence-based tools and tools based on analysis of expression data (which often incorporate sequence-based predictions). This section presents an overview of published methods developed for target prediction. Examples of these methods are shown in Table 2. For more in-depth reviews, see references [77, 118].

5.1 Sequence-based target prediction

Sequence-based prediction methods focus on finding miRNA-mRNA pairs that have complementary sequences, as sequence complementarity is the primary determinant of miRNA targeting. From a machine learning perspective, prediction of miRNA targets is essentially a *classification* problem, where the goal is to identify a set features (both of the miRNA and mRNA) that allows classifying mRNAs as either a target or a non-target of any given miRNA.

Most sequence-based approaches are essentially rule-based filters, where features of both the miRNA and mRNA sequence are used to narrow down candidate target lists [118]. These features are derived from earlier experimental knowledge, and commonly used features include: (i) sequence matches between the seed region of the miRNA and 3' UTR of the mRNA, (ii) sequence matches outside the seed region (in the 3' UTR), (iii) sequence matches in the 5' UTR or coding sequence of the mRNA, (iv) free energy of the bound miRNA-mRNA duplex, and (v) evolutionary conservation of matches between species. Rule-based prediction methods are unsupervised, i.e. no training data is used to form the classifier. Instead, the relevance of the used features is decided by the method's authors. An example of a rule-based algorithm is depicted in Figure 4.

Table 2: Examples of tools for computational prediction of miRNA targets. All listed methods, except MAGIA, account only of suppression by miRNAs.
Method: type of inference method used for predictions.
SVM: support vector machine. HMM: hidden Markov model. MI: mutual information.
Seq. used: sequence features considered (sequence-based methods) or use of previous sequence-based predictions (expression-based methods); see text for more details.
i: sequence matches between the seed region and 3' UTR, ii: sequence matches outside the seed region (in the 3' UTR), iii: sequence matches in the 5' UTR or coding sequence of the mRNA, iv: free energy of the bound miRNA-mRNA duplex, v: evolutionary conservation of matches between species.
prefilter: sequence-based predictions used as a filter prior to analysis.
prior: sequence-based predictions included in prior distributions

Name	Method	Seq. used	Additional notes
Sequence-based methods			
TargetScan [1]	rule based	i,ii,iv,v	Originally the first published target prediction tool.
miRanda [10]	rule based	i,ii,v	Aligns whole miRNA to mRNA 3' UTR.
mirTarget [113]	SVM	i,ii,iii,iv,v	
rna22 [74]	Markov chain and rule	i,ii,iv	Uses a Markov chain to identify potential regions in mRNA 3' UTR and then sequence-rule filtering.
PicTar [60]	rule and HMM	(i, iv,v)	Semi-supervised-like approach; uses strict sequence rules^{i,iv,v} to obtain a training set for a HMM classifier.
TargetBoost [90]	genetic programming	learned	Learns sequence features and classifier from training data.
Expression-based methods			
MAGIA [92]	correlation, MI	prefilter	Also produces a bipartite network of miRNA-mRNA interactions.
TaLasso [76]	lasso regression	prefilter	
Engelmann et al [31]	least angle regression	none/prefilter	Least-angle regression is a specific implementation of lasso.
miRNAmRNA [104]	global test	prefilter	Uses mRNA expression profiles to predict miRNA expression.
GenMir++/3 [48, 49]	Bayesian regression	prefilter/i,iv,v	GenMir3 can incorporate sequence features into the model.
Stingo et al [96]	Bayesian variable selection	prior	Effectively a spike-and-slab variable selection approach, scores from any sequence-based method can be used as prior information.

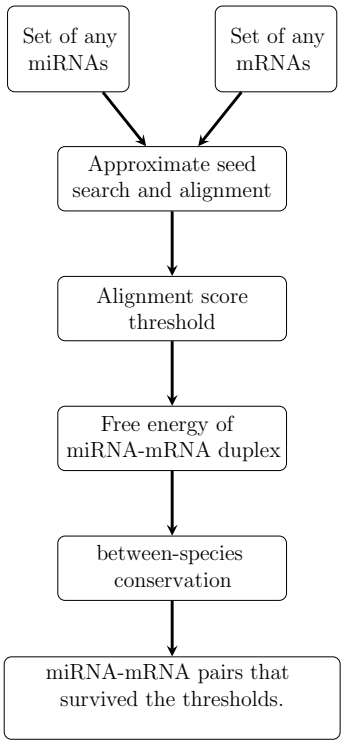


Figure 4: A schematic of the miRanda [10] algorithm for miRNA target prediction. The algorithm takes as input a set of miRNA and mRNA sequences. It then searches the mRNA 3'UTRs for the seed sequence of each miRNA and performs an alignment of the miRNA and mRNA if the seed is found. Then thee free energy of the aligned miRNA-mRNA duplex is computed, and finally a conservation score between several species is computed for the aligned section of the mRNA. At each step a score threshold is used for filtering miRNA-mRNA pairs to the final set of candidate pairs. Modified with permission from [55].

Supervised machine-learning approaches has also been employed, where a training data set consisting of experimentally validated targets and non-targets (often obtained from expression data sets) is used to train a classifier for classifying mRNAs as miRNA targets. Support vector machines (SVM) are the most common choice for classifier. The features used for classification are similar to rule-based tools, i.e. mostly derived from sequences, but supervised learning allows the inclusion of much more features. For example mirTarget uses a set of 113 features including seed region matches, conservation and a range of different sequence features from different parts of the miRNA sequence [113]. More complex approaches have also been applied, such as using genetic programming to learn sequence features, and using Markov chains or hidden Markov models (HMM) as a sequence generative model to estimate targeting probability.

The advantage of sequence-based methods is that they are based on experimentally derived knowledge on molecular mechanisms and, thus, are likely to represent causal relationships. As such, the predictions are easy to interpret. Disadvantages of sequence-based methods include: considering only pair-wise interactions cannot capture combinatorial effects; using sequence conservation misses poorly conserved

species-specific targets; requiring seed region matches cannot identify miRNA targets without seed matches (while these appear rare, they should not be discounted altogether [8]); a sequence match does not always confer repression and could be functionally inactive; and, finally, rule-based methods are static and do not account for differing miRNA and mRNA expression profiles in various tissues and disease states. There is also a general lack of overlap between predictions from different sequence-based approaches, suggesting that many results are spurious false positives [77].

5.2 Expression-data-based target prediction

Integrating expression data with sequence-based target prediction helps combat the high false-positive rate of sequence-only methods and, importantly, enables tissue and disease specific support for target predictions in real-world data. Recent evidence indicates that miRNAs act predominantly through mRNA degradation [44]. Thus, it is feasible to use mRNA or protein expression data to infer target relationships, since the regulatory effect of miRNAs should be reflected in mRNA and protein abundances. Sequence-based predictions are often incorporated as a preliminary filter step to limit the potential interactions examined.

Various mathematical approaches, ranging from correlation to complex Bayesian models, have been proposed for expression-based prediction. Most methods limit the studied relationship to repression by the miRNA. This has been suggested to improve performance [76], but has the limitation of not being able to detect positive regulation, both direct and indirect (mediated through regulation of other mRNAs) [31]. Notably, the majority of published efforts use either protein or mRNA expression together with miRNA expression, very few have combined all three.

Let us henceforth define $y_k = [y_{1k}, \dots y_{nK}]$ as a vector of expression values of mRNA $k(k = 1, 2, \dots, K)$ and $X_{n\times p} = [x_j] = [x_{ij}]$ as the matrix of expression values of miRNAs $j(j = 1, 2, \dots, p)$ for observations $i(i = 1, \dots, n)$.

Correlation Several methods and publications use a straightforward approach to identifying miRNA targets by finding miRNA-mRNA pairs whose expression patterns are similar across observations. This is achieved with simple measures of variable association. Pearson correlation is widely used because of its simplicity and intuitive interpretation. Pearson correlation between mRNA k and miRNA j is defined as:

$$\rho_{kj} = (y_k)^T_{\mu_k=0|\sigma_k=1} \cdot (x_j)_{\mu_j=0|\sigma_j=1}, \tag{1}$$

where $\mu = 0|\sigma = 1$ indicates normalization to zero mean and unit variance. Significantly correlated miRNA-mRNA pairs are classified as putative target interactions. Other measures used include Spearman correlation and mutual information (MI). A crucial limitation of correlation analysis is being restricted to studying pair-wise associations. Single miRNAs often have a small effect on mRNA expression, which

leads to weak associations and, therefore, low power to identify miRNA targets. This issue is worsened by a large multiple-hypothesis testing problem when considering all possible miRNA-mRNA pairs. Some approaches have used additional information, such as sequence-based prediction or differential expression analysis, for limiting examined miRNA-mRNA pairs to alleviate this to some extent [77].

Multivariate linear regression Many proposed expression-based methods use some form of multivariate linear regression (MLR) to examine the relationship between miRNAs and mRNAs. Expression profiles of miRNAs are commonly used to predict the expression of a single mRNA. Recently, Engelmann and Spang [31] reported that miRNA expression can indeed be used to predict mRNA expression. In the context of regression, target prediction essentially becomes a *variable selection* problem, where the goal is to choose a set of miRNAs that best predict mRNA expression without overfitting.

A linear regression model for the expression of mRNA k is defined as

$$y_k = \sum_{j=0}^{p} (\beta_{kj} \cdot x_j) + \epsilon_k = X\beta_k + \epsilon_k, \tag{2}$$

where $\beta_k = [\beta_{k0}, \ldots, \beta_{kp}]$ is the vector of regression coefficients, β_{k0} is the intercept term, ϵ_k is the error term, and X is the matrix of covariates, i.e. the miRNA expression vectors (where a constant column vector of $x_0 = 1$ has been added for the intercept). The parameter of interest is β_k, which determines to the contribution of each miRNA to the response variable y_k, i.e. mRNA expression. The regression error ϵ_k represents noise and fitting error caused by variation not captured by the included covariates. ϵ_k is commonly assumed to be normally distributed, with equal variance and no correlation between observations, giving the *normal linear model*. It is straightforward to incorporate previous sequence-based predictions by adding an indicator variable $y_k = Xc_k\beta_k + \epsilon_k$, where $c_{kj} = 1$ if mRNA k is a potential target of miRNA j, and $c_{kj} = 0$ otherwise.

The advantage of using regression for target prediction is the ability to model the effect of several miRNAs on one gene simultaneously. Simple MLR is not applicable in cases, where the number covariates is larger than the number of observations (here $p > n$) [2], because the linear model is undetermined and a single solution cannot be obtained. Furthermore, simple MLR cannot solve the problem of variable selection as the model fit improves asymptotically by adding more covariates, leading to overfitting.

Regularized regression Both the dimensionality problem and overfitting can be overcome using regularized regression. The most common approach is to apply regularized least squares, where a penalty depending on the magnitude of the coefficients

[2] A characteristic that is very common in analysis of high-throughput biological data, for example microarray expression data.

β is applied to force them small. This entails minimizing the expression

$$min\{\|y_k - X\beta_k\|_2 + \lambda R(\beta_k)\}, \tag{3}$$

where the first term corresponds to fitting error (the sum of squared residuals), $R(\beta_k)$ is the penalty function and λ is a tuning parameter that controls the amount of regularization. The 1-norm ($R(\beta_k) = \|\beta_k\|_1 = \sum_{j=0}^{j} |\beta_{kj}|$) is frequently used for regularization; this is referred to as lasso regression (shorthand for *least absolute shrinkage and selection operator*). Lasso regression in effect forces the number of non-zero coefficients in β_k to be small, leading to a sparse solution that chooses seemingly important covariates. While regularization solves the dimensionality problem and improves interpretability, it has several important limitations. Firstly, regularization may remove covariates highly associated with and functionally regulating the response, instead retaining an unimportant covariate that correlates with actual regulators [31]. Secondly, only a limited number of covariates may be included in the model, and thus some relevant associations can be missed by number of included covariates alone. Relating to both limitations, van Iterson et al [104] showed that lasso did not consistently select highly correlated miRNA-mRNA pairs.

Other approaches Other suggested approaches used include the global test [104], which is a generalization for testing the global null hypothesis ($H_0 : \beta = 0$) of a linear regression model when $p >> n$, and approaches similar to gene-set enrichment analysis, where the over-representation of sequence-based target genes in differentially-expressed gene sets is considered indicative of a target relationship in the studied condition. Le at al [62] have proposed an ensemble method, which combines predictions from several separate algorithms to build on the advantages and compensate for the drawbacks of each. Several Bayesian approaches have also been proposed; these are discussed in the next section.

6 Bayesian analysis

Bayesian data analysis is a modeling framework that is based on the principle of quantifying uncertainty as probability. Current knowledge about unknown model parameters, variables and future observations is described in terms of probability statements [38]. This provides a framework which is inherently suited to dealing with noisy real-world data, as measurement noise is naturally incorporated into probability distributions. This section provides a brief introduction into Bayesian analysis and discusses Bayesian regression and variable selection as applicable to the problem of microRNA target prediction.

6.1 Basics of Bayesian analysis

Bayesian analysis begins by defining a joint probability model $p(y, \theta)$ for observed data y and unknown model parameters θ. The joint distribution can be written as a product of two probability distributions

$$p(y, \theta) = p(\theta)p(y \mid \theta), \tag{4}$$

which are referred to as the *prior distribution* $p(\theta)$ and the data distribution or *likelihood* $p(y \mid \theta)$. The prior conveys information on the presumed values a parameter may take and the likelihood represents the likeliness of observed data for given parameter values (in the context of the chosen data model). Applying the Bayes' theorem we obtain the *posterior distribution* for θ given the known values of the observations y:

$$p(\theta \mid y) = \frac{p(y, \theta)}{p(y)} = \frac{p(\theta)p(y \mid \theta)}{p(y)}, \tag{5}$$

where $p(y) = \int_\theta p(\theta) \cdot p(y \mid \theta) d\theta$. Noting that $p(y)$ does not depend on θ, we can write Equation (5) as

$$p(\theta \mid y) \propto p(\theta)p(y \mid \theta). \tag{6}$$

The latter is referred to as the unnormalized posterior distribution. The Bayes' theorem forms the heart of Bayesian inference and illustrates the core concept of updating prior beliefs to account for observed evidence. The posterior provides a probability assessment of the possible values of a parameter, and represents a compromise between prior knowledge and information obtained from observations. As the number of observations increases, the data have increasing influence on the posterior [38].

Hierarchical models – where parameters of prior distributions have their own priors, called *hyperpriors* – allow significant flexibility in Bayesian modeling. In situations where model parameters are related to each other, a common hyperprior may be used for several parameters. Hierarchical models are particularly appropriate in settings where data are sparse (as available information is shared between parameters) or the

data are naturally structured into several levels, such as similar measurements from different hospitals or schools.

The virtue of modeling uncertainty as probabilities – in addition to naturally dealing with noise – is that they are conceptually easy to grasp and often allow common-sense interpretations of conclusions to be made.[3] Perhaps the foremost advantage of Bayesian modeling, however, is its flexibility and extensibility; it can cope with complex problems and data with relative ease. Prior knowledge about the parameters of interest can be embedded in prior distributions, priors can be assigned freely to each parameter, hierarchical models can be used to model layered data, and new observations can be added sequentially to update previous conclusions.

The challenge in Bayesian analysis is choosing proper probability models for the parameters and observations, including prior distributions as well as the likelihood [38]. In fact, Bayesian methodology has been criticized for the subjectivity related to choosing suitable priors. One could, however, argue that the choice of any model is always subjective to a certain degree, irrespective of chosen methodology. Additionally, weakly informative or non-informative priors can be used to decrease the effect of subjective information (or in cases where no prior knowledge is available) and conclusions from inferences using non-informative priors often coincide with classical analyses.

6.2 Bayesian inference

The goal of Bayesian inference is to make conclusions about unknown parameters θ or unknown observations $\tilde{y}$, given the observed data y. These are formulated as posterior distributions or features describing them, such as point or interval estimates. In simple cases, the posterior $p(\theta \mid y)$ can be derived in analytical form. In practice, however, it is often not possible to obtain explicit forms of posteriors or analytical solutions to integrals involved in inference, especially with complex and hierarchical models. Therefore, numerical estimation or simulation methods, in the form of sampling from probability distributions, are frequently used to approximate the posterior.

The simplest approaches to simulation include sampling directly from the posterior distribution $p(\theta \mid y)$, when possible, or from a simpler distribution proportional to $p(\theta \mid y)$ using e.g. rejection sampling [38]. More sophisticated methods are often needed when dealing with complex models. Markov chain Monte Carlo (MCMC) is a general approach to simulation that is based on drawing samples of θ from an approximating distribution. The draws are corrected at each iteration so that the approximating distribution becomes closer to $p(\theta \mid y)$. Each draw $\theta^{(t)}$ is conditional

[3] This is especially true compared to classical frequentist inference, which is defined within the context of repeated sampling (and inference) from a fixed but unknown process generating the observations. For example, frequentist confidence intervals strictly do not indicate that the true value of the parameter is contained within with high probability – a common misconception – where as Bayesian posterior intervals do (subject to modeling assumptions).

(only) on $\theta^{(t-1)}$, the previous draw .[4] MCMC methods are applicable to arbitrary posterior distributions and a range of programs for running simulations of full Bayesian inferences are available.

A key issue with iterative methods, such as MCMC, is running the simulation long enough, so that the distribution for drawing samples has converged close enough to the target distribution. Basic solutions to this include discarding a burn-in period of samples from the beginning of each simulation (to assure that the samples arise from a converged state), and running several separate simulations (chains) with different starting points to improve coverage of the posterior. Various approaches to measuring convergence have been proposed.

6.3 Bayesian regression

Bayesian regression analysis aims to infer the posterior distributions for the regression coefficients of covariates and other model parameters, such as the variance (i.e. the error term) of the observation model. Within the Bayesian framework, the *normal linear regression* defined in Eq. (2) can be expressed as

$$y \mid \beta, \sigma, X \sim N(X\beta, \sigma^2 I), \tag{7}$$

where N is the multivariate normal distribution, and I is the $n \times n$ identity matrix (the gene index k has been suppressed for clarity). The mean of y is then the familiar linear sum of x_k

$$\mathrm{E}(y \mid \beta, X) = X\beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{8}$$

The posterior distribution for the regression coefficients (up to a normalizing constant) is obtained from the marginal posterior

$$p(\beta \mid \sigma, y, X) \propto \int p(y \mid \beta, \sigma, X) p(\beta \mid \sigma) p(\sigma) d\sigma, \tag{9}$$

with the joint prior $p(\beta, \sigma) = p(\beta \mid \sigma)p(\sigma)$. It is relatively straightforward to extend this simple model, for instance by allowing unequal variances or correlation between observations, choosing a different data distribution to represent the error term, or including hyperparameters to construct a hierarchical model.

As mentioned in Section 5.2, microRNA target prediction using regression analysis of expression data is effectively a variable selection problem. For Bayesian regression, several different priors that provide model shrinkage have been proposed, including the Laplace prior (which is closely related to lasso regression), the horseshoe prior, and the hierarchical shrinkage prior (a generalization of the horseshoe). A hierarchical shrinkage prior for regression weight β_j can be formulates as

$$\begin{aligned} \beta_j \mid \ \lambda_j, \tau &\sim N(0, \lambda_j^2 \tau^2) \\ \lambda_j &\sim t_\nu^+(0, 1), \end{aligned} \tag{10}$$

[4] This is essentially the definition of a Markov chain; a sequence of random variables, where the probability density of each one is dependent on only the previous one.

where t_ν^+ denotes the half-Student-t prior with ν degrees of freedom [84]. The λ_j correspond to a local scale parameter and τ controls the amount of global shrinkage. As an example, in a very sparse model with many irrelevant covariates, the model would ideally have small τ (so that $p(\beta)$ is mostly shrunk close to zero), but allow some λ_j to be large to escape the shrinkage.

A weakly informative half-Cauchy distribution is often the suggested choice of prior for τ, but van der Pas et al have also proposed (for the horseshoe prior) using a fixed value of $\tau = \frac{p_n}{n}\sqrt{\log(n/p_n)}$, where p_n is the assumed number of relevant covariates and n is the number of observations [103]. Bayesian shrinkage priors, however, do not lead to a sparse solution as there remains uncertainty in the posterior distribution and no coefficient can be considered exactly zero.

6.4 Bayesian variable selection

In order to find a small set of relevant predictive variables, a model selection approach needs to be applied. To this end, a range of methods applicable in Bayesian analysis have been proposed; examples include using cross validation, different information criteria, and projection methods to determine the submodel giving the best compromise between prediction accuracy and model size. A detailed review of these falls outside the scope of this thesis, however, a comprehensive one has been recently written by Vehtari and Ojanen [110].

In the context of variable selection for regression, Piironen and Vehtari [85] recently suggested that, for problems where data are scarce and the number of candidate variables high, using projection predictive variable selection is effective. The idea, proposed by Dupuis and Robert [25], is to fit a full reference model M_{ref} encompassing all candidate variables, and then project the information in the reference posterior onto a submodel $M_\perp$ so that the predictions are as similar as possible.

Given the reference model parameters θ_{r}, the projection $\theta_\perp$ in the parameter space of $M_\perp$ is obtained by solving

$$\theta_\perp = \operatorname*{argmin}_\theta \frac{1}{n}\sum_i^n \mathrm{KL}\left(p(\tilde{y} \mid x_i, \theta_{\text{r}}, M_{\text{ref}}) \parallel p(\tilde{y} \mid x_i, \theta, M_\perp)\right), \tag{11}$$

where $\mathrm{KL}\left(P \parallel Q\right)$ is the Kullback-Leibler divergence between probability distributions P and Q. The discrepancy between the reference model M_{ref} and submodel $M_\perp$ is then defined as the expectation of the divergence over the posterior of the reference model:

$$\delta(M_{\text{ref}} \parallel M_\perp) = \frac{1}{n}\sum_i^n \mathrm{E}_{\theta_{\text{r}}|D,M_{\text{ref}}}\left[\mathrm{KL}\left(p(\tilde{y} \mid x_i, \theta_{\text{r}}, M_{\text{ref}}) \parallel p(\tilde{y} \mid x_i, \theta_\perp, M_\perp)\right)\right]. \tag{12}$$

The posterior expectation in (12) can, in practice, be estimated by drawing samples from the reference posterior (using e.g. MCMC). It can be shown, that in the case of

normal linear regression, the minimization in (11) can be solved analytically and the discrepancy (12) has a simple form depending only on the reference and projected model variances σ^2 [84].

A practical issue with this method is deciding how many variables should be included in the submodel. This depends on what is considered acceptable loss of prediction performance compared to the submodel. Piironen and Vehtari suggest using cross-validation to guide the variable selection process and give a practical guideline for stopping the selection. This is discussed Section 7.2.5. In this thesis, projection predictive variable selection was used for inferring putative microRNA targets from breast cancer expression data using Bayesian regression. Further details of the used method are presented in Section 7.2.5.

6.5 Bayesian microRNA target-prediction methods

One of the earliest tools to use expression data for target prediction was GenMir++ [48]. It takes as input a candidate set of miRNA targets (the authors have used TargetScanS) and uses mRNA and miRNA expression data across multiple tissues to predict whether a given candidate miRNA-target interaction is real. GenMir is based on a Bayesian regression model, where mRNAs are assumed to share a tissue-specific common background expression, which is downregulated by regulating miRNAs. Let μ_t represent the background mRNA expression in tissue t, the probability model for mRNA expression y_{kt} is defined in GenMir as:

$$y_{kt} \mid X, S_k, \Gamma, \Lambda, \mu_t, \Sigma \sim \mathrm{N}\left(\mu_t - \gamma_t \sum_j^p \lambda_j s_{kj} x_k, \Sigma\right), \tag{13}$$

where γ_t is a tissue-specific scaling factor (modeling differences in mRNA and miRNA measurements and normalization) and λ_j are regulatory weights of the miRNAs (irrespective of candidate target k), and s_{kj} an indicator variable of target interaction. Compared to Eq. (7), the degree of regulation of a miRNA becomes $\beta_k j = \gamma_t \lambda_j s_k j$. The goal is to infer the posterior $p(s_{kj} \mid c_{kj} = 1, D, M)$, that is, the probability of a candidate interaction being true, where c_{kj} is an indicator variable of the input putative interactions. A log-odds score is given for each miRNA-mRNA interaction.

The latest version of GenMir (GenMir3) defines Gamma priors for γ_t and λ_t and Bernoulli priors for s_{kj} (leading to negative interactions only) and allows including sequence features in a logit hyperprior for the prior $p(s_{kj})$. The model is solved simultaneously for all genes and tissues, using a variational Bayes method, to obtain an approximate posterior for S. The authors note, that the method can be easily extended to add protein expression.

Another Bayesian approach to target prediction is a Bayesian network method published by Stingo et al [96]. The approach is essentially an implementation of the spike-and-slab variable selection method [110] applied to a Bayesian regression model

equivalent to Eq. (7). Sequence features are included in the (Bernoulli) prior of the covariate inclusion variable S. The posterior for S is obtained with MCMC methods. A time-dependent coefficients version is also presented for data measured for several time points. Additionally, many published analyses of different expression data for target prediction have utilized various probabilistic or Bayesian approaches.

7 Materials and methods

7.1 Research material

Analyses performed in this thesis used protein, mRNA and microRNA expression data previously published by Aure et al [4]. The miRNA and mRNA data are publicly available on the Gene Expression Omnibus [28] (accessions GSE8210 and GSE8212, respectively) and the protein data is provided as supplementary data in [4].

The data consist of 283 tumor samples collected from 280 breast cancer patients treated in two Norwegian hospitals. The data are part of the larger Oslo2 cohort, which consists of breast cancer patients with primarily operable disease [4]. Collection of the cohort started in 2006 and is still ongoing. Therefore, no survival data were available for analysis. Patient ages ranged from 30 to 83 years with a median of 55 years. The vast majority of tumors in the data were ductal carcinomas, which is the most common type of breast cancer, however, in general the dataset consisted of a heterogeneous collection of different histological types and stages of breast cancer tumors. No matched control samples of healthy breast tissue were available.

The mRNA and microRNA expression were measured using Agilent Technologies (Santa Clara, CA, USA) SurePrint G3 Human GE 8x60K and Human miRNA Microarray Kit (V2) microarrays, respectively. These microarrays measure 27958 genes and 887 miRNAs, according to manufacturer annotation. Protein expression was measured with reverse-phase protein arrays (RPPA) [102] for a selected set of 105 proteins relevant in cancer. Most of the proteins are found on the PI3K/AKT intracellular pathway, which plays a central role in cell-cycle regulation.

7.2 Methods

A Bayesian regression model of protein, mRNA and miRNA expression data was constructed, and projection predictive variable selection was used to predict microRNA targets in breast cancer data. Details of the methods used are presented in this section.

All computational analyses were performed in R [87] and workflow management was handled with Anduril [79]. Monte-Carlo simulations for the Bayesian regression models were computed with RStan [95] using the No-U-Turn variant of a Hamiltonian Monte Carlo algorithm for sampling posterior distributions. The simulations were performed using computer resources within the Aalto University School of Science "Science-IT" project.

7.2.1 Preprocessing and quality control

The preprocessed miRNA and mRNA data[5] were downloaded from GEO using the GEOquery [22] R package. The protein data were downloaded and extracted from a Microsoft Excel file. All of the data had been transformed to log2 scale, which is useful for making the distribution of expression values closer to a normal distribution.[6] No further preprocessing of the expression values was done. For details on the preprocessing process, the reader is referred to the supplementary data of Aure et al [4]. In regression analyses, all variables (miRNA, mRNA and protein) were further scaled to have zero mean and unit variance; this transformation is commonly used for regression.

The mRNA expression data were available as probe-level measurements, these were summarized to gene-level using manufacturer probe annotations by taking the mean of all probes targeting the same gene. Out of 421 miRNAs present in the data, eleven had been deleted from miRBase. These miRNAs (namely hsa-miR-1274a, hsa-miR-1274b, hsa-miR-1280, hsa-miR-1308, hsa- miR-1826, hsa-miR-1974, hsa-miR-1975, hsa-miR-1977, hsa-miR-1979, hsa-miR-720, hsa-miR-886-3p) were therefore from all analyses.

Members of the *AKT* and *GSK* gene families (namely *AKT1*, *AKT2*, *AKT3* and *GSK3A*, *GSK3B*) were not distinguishable in the protein assay. Therefore, the protein data included only a single set of measurements for each family. For the analyses presented here, each of these genes was considered separately using the same expression values for all family members.

For assessing the quality of the data, distributions of expression values for each tumor sample and each variable (miRNA, mRNA and protein) were visualized using boxplots. A principal component analysis (PCA) and hierarchical clustering of samples were performed separately for each data type to assess possible bias introduced by the data having been measured at two separate hospitals.

7.2.2 Validated target reference

The predicted miRNA targets obtained in the analyses were compared to validated targets in DIANA-TarBase v7.0 [111] (referred to as TarBase from here on) and miRTarBase release 6.0 [17]. Data from both databases were downloaded and a union of the databases was used as a reference set of validated interactions. The resulting set contained 328825 validated miRNA-mRNA interactions, out of which only 4082 were between a gene and miRNA present in the analyzed data, however.

[5] Raw array data produced from Agilent array readers are also available, but were not used in this thesis.

[6] Raw expression values are approximately log-normal.

7.2.3 Correlation analysis

To assess dependencies between variables from the different expression data types (protein, mRNA, miRNA) and the validity of correlation as a target prediction tool, Pearson's correlations were computed between matched and unmatched protein-mRNA pairs (where matched refers to both corresponding to the same gene) as well as validated and random protein-miRNA and mRNA-miRNA pairs, where validated pairs were ones present in the reference set described above. The random correlations consisted of 5000 randomly picked pairs (with replacement).

7.2.4 Regression models

For predicting protein expression from mRNA and miRNA expression, a similar regression model to Aure et al was used:

$$y = \beta_0 + z\beta_g + X\beta + \epsilon, \tag{14}$$

where y is the protein expression and z the mRNA expression for gene k (k is suppressed for clarity), w_g is the regression coefficient for the mRNA, X is the matrix of miRNA expression vectors, and w_0 is the intercept term (for a justification of this equation, see [4]). A separate model was fitted for each gene. A model with only the mRNA expression covariate (called the *gene-only model*), defined as $y = \beta_0 + z\beta_g + \epsilon$, was used as a baseline. A normally distributed error term with equal errors and no correlation between observations was assumed for all models.

The likelihood for Bayesian regression was therefore defined as

$$y|\beta, \sigma, z, X \sim N(\beta_0 + z\beta_g + X\beta, \sigma^2 I), \tag{15}$$

where $\beta = [\beta_0, \beta_g, \beta]$ for convenience. The intercept and mRNA coefficient were given diffuse Gaussian priors and σ a uniform prior:

$$\beta_0 \sim N(0, 5^2) \tag{16a}$$
$$\beta_g \sim N(0, 5^2) \tag{16b}$$
$$\sigma \propto 1. \tag{16c}$$

A hierarchical shrinkage prior was applied to the miRNA coefficients β, as defined in Equation (10). The degrees of freedom for the λ_j priors was set at $\nu = 3$ (similar to Piironen and Vehtari [84]). The prior for τ was defined as:

$$\tau \sim \text{half-Cauchy}\left(0, \frac{p_n}{n}\sqrt{\log(n/p_n)}\right), \tag{17}$$

combining the previous suggestions of half-Cauchy and fixed *tau*. The assumed number of relevant miRNAs, p_n, was estimated as follows. Ensembl (release 86) gene ID's were downloaded for all protein-coding genes in the human genome using biomaRt [26]. From these, a sample of 1000 genes was taken, and known validated

microRNA interaction partners for each sampled gene were downloaded from miRWalk [27]. Genes for which there were no validated miRNA interactors were assumed to have zero. The mean number of miRNA interactors per gene was used as the estimate, giving $\hat{p_n} = 13.75$.[7]

7.2.5 Variable selection

Projection predictive variable selection (as described in Section 6.4 and as applied by Piironen and Vehtari [84]), was used to obtain the relevant set of microRNAs for each gene. A full reference model was fitted by drawing 2000 samples from the posterior using RStan (with 4 chains, 1000 samples each and the first half discarded as burn-in). A random sample of $S = 1000$ simulation samples from the full posterior was used to increase projection speed.

A series of projected submodels was obtained using a forward search strategy, that is, the search started from a model including only the intercept, the mRNA expression z was always added as the first covariate, and at each subsequent step, the miRNA covariate x_j giving the largest decrease in discrepancy between the reference and projected models (Eq. (12)) was chosen. The forward search was continued up to 200 variables.

For choosing the model size, 10-fold cross validation was used, as proposed by Piironen and Vehtari [85]. This means that the above model selection process was repeated $K = 10$ times, each time leaving out n/K observations for model evaluation. For judging the appropriate model size, estimation of model predictive performance was performed as explained below.

Model predictive performance The predictive performance of each submodel was evaluated using a log predictive density. Given submodel $M_\perp$ with the posterior predictive distribution $p(\tilde{y}|\tilde{z}, \tilde{X}_\perp, \beta_\perp, \sigma_\perp, D_\perp)$, where $D_\perp$ is the observed data in the current submodel and cross-validation fold and $(\beta_\perp, \sigma_\perp)$ the projected model parameters, the logarithm of the predictive density (LPD) was computed at each of the left-out observations (y_*, z_*, X_*). The LPD was estimated by averaging over the S simulated posterior samples:

$$\mathrm{LPD}_*(M_\perp) \approx \log \frac{1}{S} \sum_s^S p(y_*|z_*, X_*, \beta_\perp, \sigma_\perp, D_\perp).$$

The LPD values from each cross-validation fold were pooled and their mean over the full set of data (MLPD) was used as a summary. To compare the predictive performance of a submodel to the full reference model, the difference in MLPD (ΔMLPD) was computed. Bayesian bootstrap [88] (with 5000 samples) was used to

[7] Note that this is probably an underestimate of the true p_n, as the set of experimentally validated miRNA targets is very likely to underrepresent the true *in vivo* set of miRNA targets.

estimate a distribution for ΔMLPD as:

$$\Delta\text{MLPD}^b(M_\perp, M_{\text{ref}}) = \sum_i^n w_i^{(b)} \left[\text{LPD}_i(M_\perp) - \text{LPD}_i(M_{\text{ref}})\right],$$

where $w_i^{(b)}, i = 1, \ldots, n$, are the bootstrap weights for the b'th bootstrap sample (subject to $\sum_i w_i^{(b)} = 1$). The *bayesboot* R package was used for computing the bootstrap. The $\Delta\text{MLPD}(M_\perp, M_{\text{ref}})$ was then used as an estimate of the predictive performance of submodel $M_\perp$.

Choosing model size For choosing the model size, the following condition was used:

$$\Pr(\Delta\text{MLPD}(M_\perp, M_{\text{ref}}) > U) \geq \alpha, \tag{18}$$

where $U = \gamma\Delta\text{MLPD}(M_0, M_{\text{ref}})$, and M_0 refers to the intercept-only model. The number of covariates in the smallest model satisfying this condition was chosen as the final model size. This means that model size was chosen such, that the probability of the projected model improving performance by a constant factor (γ) over the intercept-only model was at least α. The choice of values for γ and α reflects accepted loss in predictive performance (and related uncertainty) compared to the reference model; several values for both were tested.

Final model selection The final projected model was obtained by reapplying the projection search up to the chosen number of variables, using all data for each gene. miRNAs included in each gene's model were considered putative target interactors. miRNAs for which the 95% posterior interval did not include the origo were considered *significant*. For some genes the condition in (18) was not met after including 200 covariates. In these cases it was concluded that the miRNA covariates provided no additional information on the protein expression and, thus, none of them were deemed as targeting the gene. In some cases the condition was met already by the model with only the mRNA covariate, and the same conclusion was made.

Lasso regression A lasso regression model was also fitted for each gene using the *glmnet* R package. In this case the mRNA variable was treated equal to the miRNA variables and subjected to the lasso regularization. For choosing the regularization parameter λ, the default algorithm of *glmnet* was used. That is, models were fit for a decreasing sequence of λ values, and 10-fold cross validation was used to compute a mean square prediction error (MSE) at each λ. The largest value of λ (i.e. sparsest model) that had a MSE within one standard error of the lowest MSE attained in the cross validation was used. The covariates included in the model using the chosen λ were considered putative target interactors. For some genes, this criterion was met by the intercept-only model, and again, in these cases none of the miRNAs were deemed as targeting the gene.

7.2.6 Measuring model fit

To assess convergence of simulations, the potential scale reduction measure $\hat{R}$ proposed by Gelman et al was used [38]. The coefficient of determination R^2 was used as a measure of regression model performance. R^2 is defined as

$$R^2 = \frac{SS_{\text{residuals}}}{SS_{\text{total}}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{19}$$

where $\hat{y}_i$ are the predictions made by the model and $\bar{y}$ is the mean of the outcome variable y. R^2 corresponds to the proportion of variance of the outcome variable that is explained by a statistical model, and can be used as a measure of model fit. R^2 has the property of being invariant to variable scaling, which makes it suitable for use with expression data, as expression data do not have a well defined scale.

An important caveat of R^2 is that in linear regression it often increases monotonically by adding more explanatory variables. The adjusted R^2, defined as $\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p}$ (where n is the number of observations and p the number of explanatory variables) adjusts for the number of regressors relative to the number of observations, thus penalizing inclusion of additional variables. Therefore, $\bar{R}^2$ was used for comparing the projected model with the gene only model.

8 Results

Quality control

The protein and mRNA expression data appeared reasonably uniform and mRNA data were approximately normally distributed, but several protein variables had long tails and a few had significant outliers. 28 of the protein variables had values that were further than 5 standard deviations away from the mean (highly unlikely assuming normal distribution) and proteins CDK1, ERRFI1, and PIK3CA had one value over 8 standard deviations away from the mean. These are visually most apparent in the scaled data shown in Figure .

miRNA microarrays had strongly bimodal distributions with a gap between the modes, many miRNA variables were highly skewed towards very small expression values, and there appeared to be some quantification artifacts in the miRNA data (see Figures C6 and C7). This raises suspicion of significant noise in the miRNA data, actual miRNA abundances should not have such a clear gap in their distribution, and values below the gap possibly corresponded to miRNAs not actually expressed in the data. miRNA variables that had bimodal distributions could signify that these miRNAs are not expressed in some of the breast tumor types in the data.

No significant hospital batch effect was apparent in the data. The samples collected at the different hospitals did not cluster into separate groups in principal component analysis (shown in Figure C1) or hierarchical clustering (not shown) of any of the three data types. All quality control plots are available in Appendix C.

Correlation analysis

Correlation between protein and mRNA expression was low on average. Protein-mRNA correlation was clearly higher for gene-matched pairs (that is protein and mRNA expression corresponding to the same gene) than unmatched pairs, however, even the gene-matched correlations were quite low, with mean $\bar{\rho} \approx 0.37$. Using the squared correlation coefficient ρ^2 as a measure of explained variance, the amount of protein variance explained by the matched mRNA ranged from 0% to 82% with a mean of 21%. Distributions of Pearson correlation for matched and unmatched protein-mRNA pairs are shown in Figure .

Correlations between miRNAs and their validated target genes were not significantly different from random gene-miRNA pairs (data shown in Figure). The validated target pairs showed no preference towards negative (or positive) correlation. There was virtually no difference in correlations between validated targets and randomly picked gene-miRNA pairs, contrary to what might be expected. These finding were replicated when comparing miRNA and protein data (Figure).

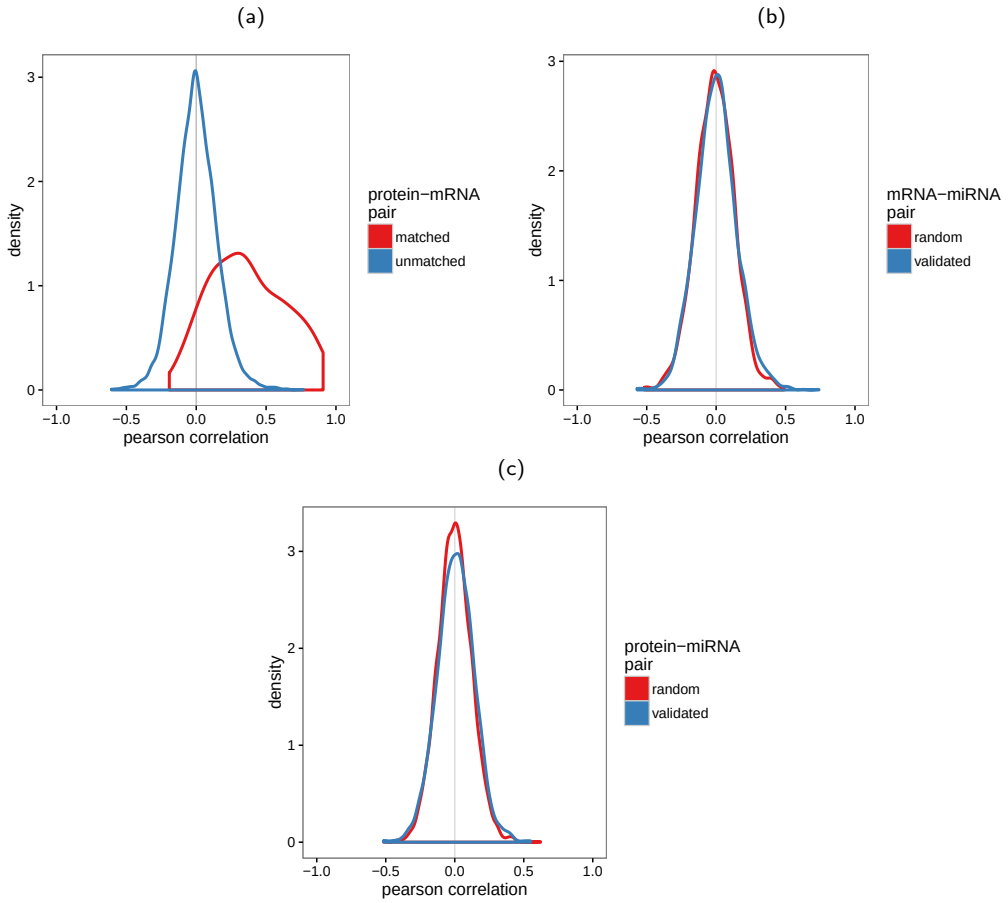


Figure 5: Distributions of Pearson correlations between variables of different expression data types. The distributions are trimmed at the smallest and largest values. (a) Correlation between protein and mRNA pairs, where "matched" refers to correlating protein and mRNA from the same gene. Matched pairs tended to have higher correlations, but the correlations were low, nonetheless. (b) Correlation between protein and miRNA pairs, where "validated" refers to the gene being a validated target of the miRNA (according to TarBase), and "random" to a randomly picked group of protein-miRNA pairs. (c) Correlation between mRNA and miRNA pairs, where grouping is the same as in (b). There was no difference between validated target pairs and random pairs in (b) and (c).

Model simulations

Simulations for the cross validation (to determine model size) lasted between 4 to 6 hours (on a computing-cluster node with 2 12-core Xeon E5 2680 2.50GHz processors), and simulations for the final projected models lasted between 2 to 45 minutes, depending on the chosen model size. All parameters for all simulations had low potential scale reduction $\hat{R}^2 < 1.1$, indicating good convergence of simulation chains [38].

The parameters for the model-size criterion, α and γ, had a significant effect on the resulting model sizes and the number of models found. Model-size distributions for

several parameter values are shown in Appendix B. Values $\alpha = 0.50$ and $\gamma = 0.2$ were chosen in order to get a projected model for most genes and to keep models relatively sparse. This choice was, however, largely subjective.

A projected model was found for 74 genes out of the 105 genes in the data. For the rest, no miRNA variables were included in the model for 27 genes, and the model-size criterion was met in under 200 variables for 4 genes. Figure 6 shows the predictive performance during the forward search in cross validation for one of the best performing models (*CDH3*) and one for which no model was found (*PIK3CA*). A table of properties for the final projected models is available in Appendix A.

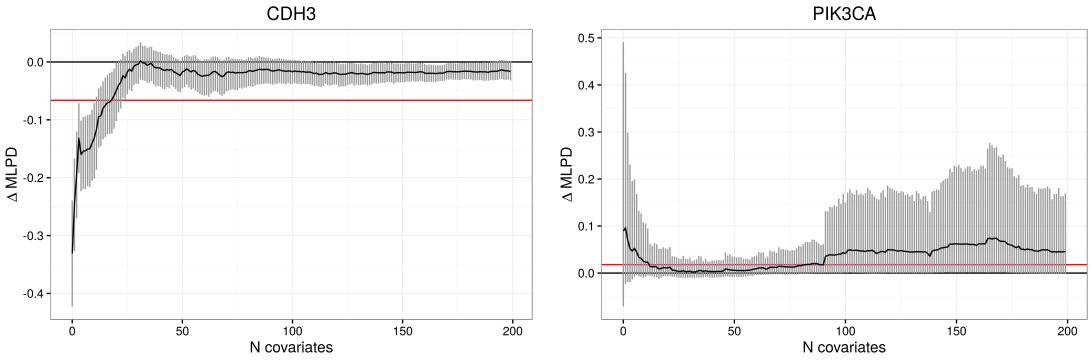


Figure 6: Predictive performance (ΔMLPD) of projected submodel $M_\perp$ during each step of the forward search in the model-size-selection phase. Two of the 105 models are shown. The black curve shows the median (corresponding to $\alpha = 0.50$) and gray lines the 95% interval of ΔMLPD computed with Bayesian bootstrap. The red line depicts the model-size selection threshold U (with $\gamma = 0.2$). The model size was chosen as the point where the black curve crosses the red line. The final *CDH3* model (shown left) was one of the best performing ones, no model was found for *PIK3CA* (shown right), as the intercept-only model already performed better than the full model.

Projected models larger than approximately 28 miRNA variables performed increasingly poorly as the model size increased. This is illustrated in Figure 7. A similar trend was observed by trying different model-size threshold parameters α and γ (data not shown).

Lasso regression produced models with at least one miRNA for 74 genes, out of which only 57 were common to the ones found by projection prediction (PPVS). For 18 of the lasso models, the mRNA expression variable was excluded from the chosen model. Figure 8 shows a comparison of model size distribution and $\Delta\bar{R}^2$ between PPVS and lasso.

Target prediction performance

Considering each selected miRNA variable as a putative miRNA-mRNA target interaction, PPVS generated a total 945 target predictions, out of which 253 were significant (on a 95% posterior interval). Table A1 lists the significant predictions.

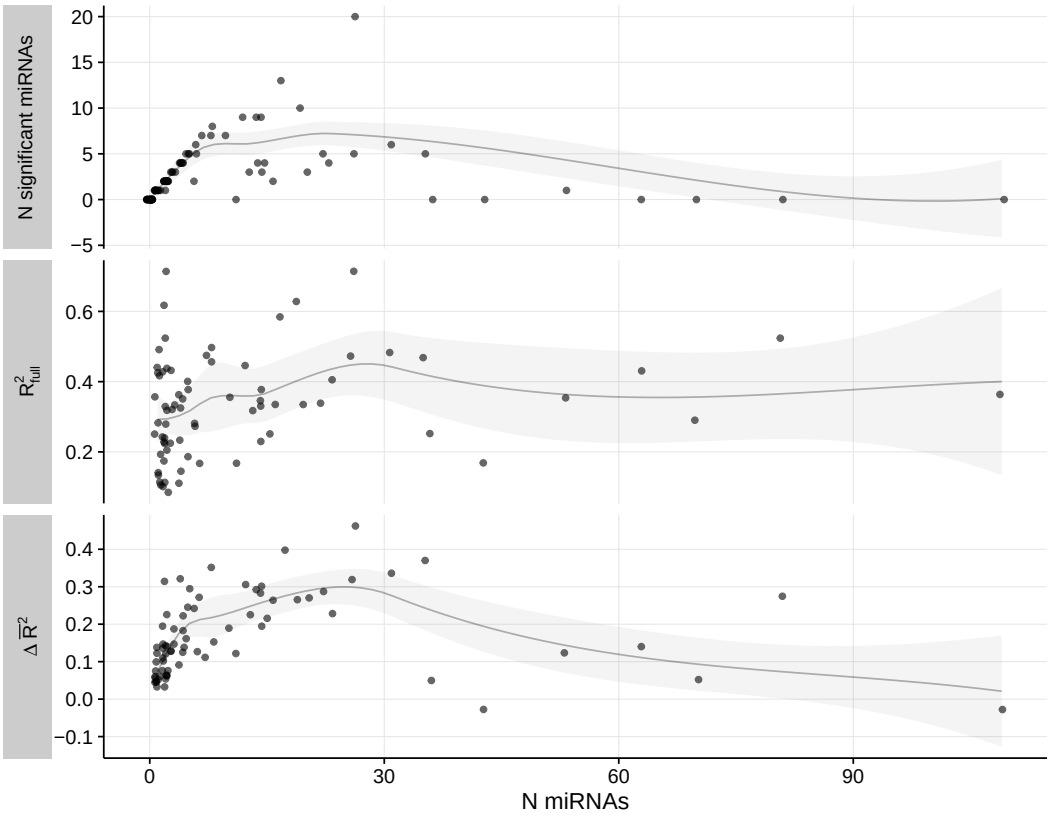


Figure 7: Size of final model compared to model goodness-of-fit. The x-axis corresponds to the total number of chosen miRNA variables in each final projected model (N miRNAs). The three y-axes show the number of significant miRNA variables (N significant miRNAs), R^2 of the projected model (R^2_{full}), and the difference in $\bar{R}^2$ between the projected and gene-only models ($\Delta\bar{R}^2$). Higher values for the y-axes are better. Each point represents one model fitted for one gene. A small jitter has been added to the points on the x-axis to help visualize all of them. Curves were fitted with locally weighted scatter plot smoothing (LOESS) and shaded areas represent 95% confidence interval. A trend can be seen, where models larger than approximately 28 included miRNAs perform increasingly poorly.

Lasso regression generated all together 650 target predictions. Figure 9 shows the overlap of predictions by PPVS and lasso, and also that of significant predictions from PPVS and the same number of top predictions from lasso, where lasso predictions were ranked by the absolute value of the regression coefficient.[8]

PPVS regression coefficients for miRNA covariates had larger magnitude (i.e. absolute value) on average than lasso, for target interactions predicted by both, implying stronger miRNA effects on protein expression. Correlation of the coefficients was fairly high (0.86), and there were only three predicted targets for which the methods did not agree on the sign of the coefficient. A scatter plot comparing the coefficients of the common predictions is shown in Figure 10.

[8]Lasso regression does not provide any rank measures or tests of significance.

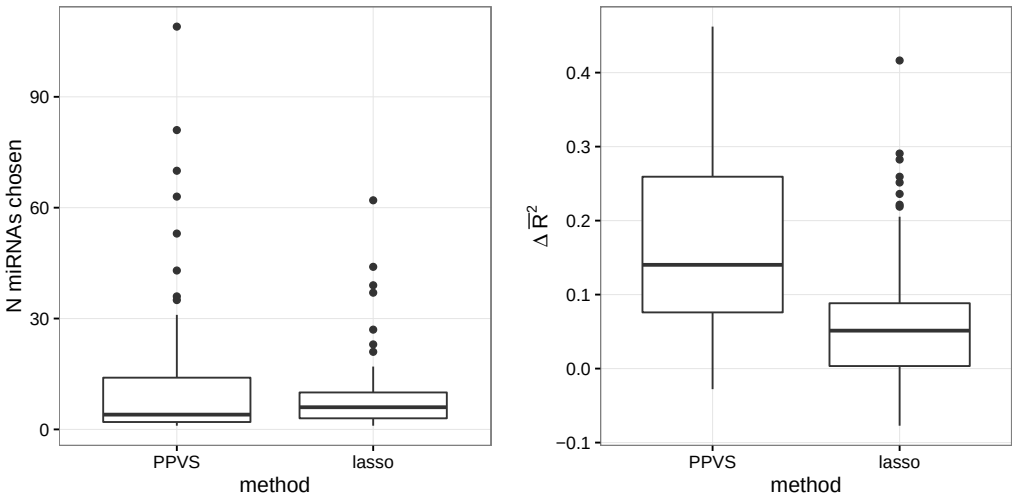


Figure 8: Comparison of final model sizes (shown left) and increase in proportion of protein variance explained compared to the gene-only model ($\Delta\bar{R}^2$, shown right) between projection prediction (PPVS) and lasso regression. PPVS models were slightly smaller on average (though the difference was not significant). The predictive performance of PPVS was better (as measured with $\Delta\bar{R}^2$).

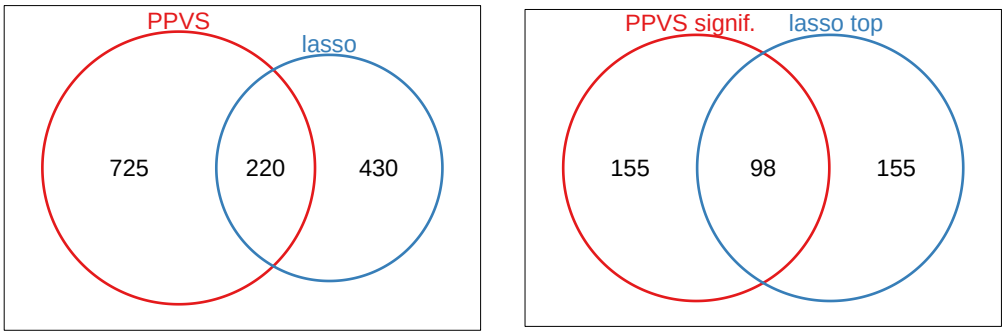


Figure 9: Venn diagrams showing the overlap of all target predictions from PPVS and lasso (shown left) and the overlap of significant (on the 95% posterior) predictions from PPVS and the same number of top predictions from lasso (ranked by absolute value of regression coefficient). There is very little overlap between the predictions made by the two methods. The proportion of overlap is slightly larger for the significant and top predictions.

PPVS and lasso had similar performance in regards of discovering validated targets; limiting predictions to the ones with most confidence did not significantly increase the proportion of validated targets for either method. Approximately 12% of PPVS and 14% of lasso predictions were present in the union of TarBase and miRTarBase. These proportions were 13% and 14% for the significant PPVS and top lasso predictions,

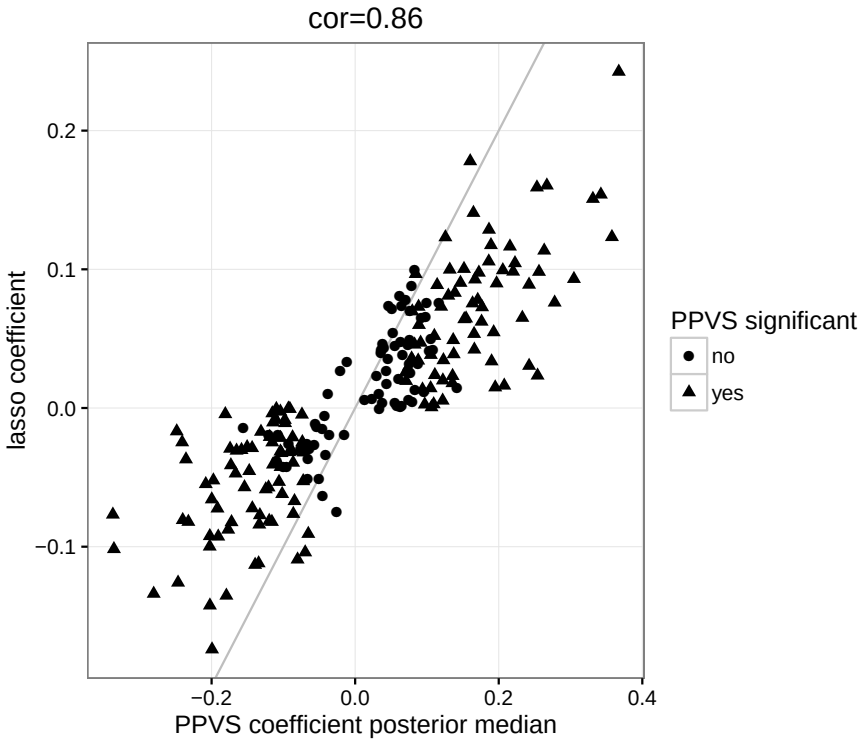


Figure 10: Comparison of miRNA regression coefficients from projection prediction (PPVS, x-axis, posterior median) and lasso regression (y-axis) for the 145 targets predicted by both methods. Triangles are significant coefficients in PPVS. The gray line ($y = x$) corresponds to equal coefficients from the two methods. The coefficients from PPVS had larger magnitude on average, implying stronger miRNA effects on protein expression. cor: Pearson correlation between the coefficients.

respectively.

For both methods, approximately half of the coefficients for both all predicted and validated targets were positive, indicating upregulation by the miRNA. In contrast, in TarBase, all of the discovered validated interactions were classified as suppressive. miRTarBase does not provide data on the nature of the regulation.

9 Discussion

This thesis presents a review of the basics of gene expression, microRNAs and the computational prediction of microRNA target genes. A modern Bayesian variable selection method was applied in the context of regression, to predict protein expression from mRNA and miRNA expression in breast cancer tumor samples, with the goal of identifying putative miRNA target. The Bayesian method was compared to lasso regression, a popular method for target prediction from expression data.

Quality control plots suggested the presence of noise in the data, especially in the miRNA microarrays. More strict filtering of miRNA data could lead to better overall results. The choice of data model can also alleviate the effects of noise. The student-t distribution, which has heavier tails than the normal distribution, would likely be a better choice for the regression model likelihood. This, however, would require a different solution to the projection Eq. (11) than the one derived by Piironen and Vehtari [84] and, therefore, was not tried in this work.

Correlation between mRNA and protein measurements for the same gene were mostly low, a finding supported by previous studies [82]. Interestingly, there was practically no difference between correlations of validated miRNA-target pairs or randomly chosen ones. This supports the view that modeling single interaction pairs individually is unlikely to be sufficient for effective target prediction using expression data.

It is possible that low correlation of miRNA-target pairs was mostly due to the dataset used and measurement noise it contained, however, correlation has low power to detect weak relationships and cannot capitalize on combinatorial effects. The low correlations also suggest that simple (Pearson) correlation of expression data is possibly inadequate for finding miRNA targets. Poor prediction results using correlation have been reported before (see for example [76]).

There could be several explanations for the increasingly poor performance of larger models, evident in Figure 7. The largest models could be a result of the covariates not explaining the outcome variable well. This results in a large model, as the inclusion of each covariate can provide only a minute improvement to submodel performance and, therefore, a large number of covariates is required to satisfy the criterion for choosing model size. The assumption that the reference model represents the best current knowledge could also be false, as seemed to be the case for some genes where the submodels performed consistently better than the full model. This likely means that the reference model has overfit the training data (in the cross- validation).

From a biological perspective, poor performance means that the microRNAs did not provide additional information for predicting protein expression after accounting for gene expression. This could be due to relevant miRNAs missing from the dataset, the number of observations being too small leading to low statistical power (unable to capture the often small effect that miRNAs have), or the biological heterogeneity of breast cancer. The proposed regression model could also be inadequate for capturing

the actual biological effect of miRNAs, though previous research seems to suggest otherwise.

Compared to lasso regression PPVS achieved better model fit, yet from a target prediction perspective, performance of the two methods was similar. There was little overlap of predictions made by the two methods, a common issue in microRNA target prediction [104]. Only a small fraction of predicted targets were validated according to TarBase and miRTarBase, however, this is probably true of all miRNA targets in general; only a limited number of validation studies have been published. The computational cost of PPVS was significantly higher than that of lasso: MCMC simulations took hours compared to less than a minute per model for lasso.

Approximately half of the regression coefficients for miRNAs were positive, suggesting that those miRNAs increase gene expression. Some of these could indicate indirect regulation. However, this proportion seems too high, as the vast majority of known microRNA interactions are suppressive. In fact, of all the experimentally validated human miRNA targets listed in TarBase, only approximately 0.2% show positive regulation by the miRNA. Therefore, many of the predicted activating interactions are possibly false findings. They could be caused by miRNA expression mirroring the involvement of other regulatory factors not included in the data. To correct for this, the model could easily be restricted to only negative interactions (using a non-positive prior for β). This has previously been reported to increase prediction performance [77].

Previous studies have shown that microRNA signatures correlate with different breast cancer subtypes [11]. This suggests that using pooled datasets of various tumors, such as the data used in this study, is likely to miss subtype-specific miRNA effects, unless this is accounted for in the model. This could be achieved in the proposed method by constructing a hierarchical model that includes tumor-subtype data.

Another way to improve the proposed model would be to include sequence-based target information, as most published methods do. This could be achieved with indicator variables, a weighting scheme, or more elaborately by including sequence-based data within the hierarchical-shrinkage prior to impose less regularization on putative target pairs. However, as the authors of GenMir noted, including sequence features did not result in a significant improvement of their method [49].[9]

The proposed model does not account for the fact that microRNAs have several, even hundreds, of target transcripts [36]. Therefore, the regulatory effect of a single miRNA is most likely spread across several genes. In combination with transcripts having several regulating miRNAs, this many-to-many nature of microRNA regulation ultimately calls for computational methods that model the whole regulatory network at once, such as regression models with multivariate targets. This, however, becomes a much more difficult problem than multivariate linear regression.

[9]It should be noted, that GenMir uses sequence-based predictions as a preliminary filter step. Therefore, it is perhaps not surprising that including the same type of data within the model does not produce substantial improvement.

Aure et al [4] used lasso regression for a similar analysis of the same dataset. They used a multi-step process, where only miRNAs deemed significant in a univariate regression model were used as input in multivariate lasso regression. This approach is flawed in the sense, that it loses some of the power of multivariate models to identify singly weak but combinatorially strong effects, as univariate modeling is used as a filtering step. It also effectively uses the same data twice, causing bias, and introduces a multiple hypothesis testing problem. Therefore, a multivariate approach (such as the one presented here or earlier ones with slight modification) would likely be preferable.

In conclusion, the work in this thesis shows that the proposed method of projection predictive variable selection is applicable to microRNA target prediction. However, further refinements to the model are warranted to improve performance. In the presented form, compared to a simpler alternative, the method offered only a limited advantage from a modeling perspective, and no apparent advantage from a biological perspective, but incurred a large computational burden. The choice of parameters α and γ, which define the threshold for model size, proved nontrivial. The values chosen had a large impact on the sizes of resulting models and, therefore, a data-driven approach for optimizing the parameter values would perhaps be useful.

Future prospects

The recent development of CLIP-seq and similar methods has made high-throughput experimental microRNA target discovery possible, partially replacing the need for computational target prediction. Nonetheless, experimental (particularly high-throughput) methods are not immune to error, and gene regulation is vastly complex with many unconventional regulatory mechanisms having been discovered. Integrative computational approaches beyond correlation – combining several types of data – will, thus, remain important in the future. Possibilities for integrating data include incorporating copy number variation, other regulatory RNAs, transcription factors, other protein-level regulatory factors such as phosphorylation, and epigenetic mechanisms into models.

The elucidation of complex regulatory networks using network-level modeling is becoming feasible with modern experimental and computational methods. Employing this approach will be essential, as it has the ability to better capture the true nature of gene regulation and cellular biology.

Many aspects of microRNA biology and function still remain unknown. Uncovering miRNA function offers interesting possibilities in diagnostics and treatment of disease, and will further our understanding of the complexities of molecular cell biology. Therefore, microRNAs remain an exciting avenue of research.

References

[1] Agarwal, V., Bell, G. W., Nam, J. W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, 4.

[2] Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, 7(1):55–65.

[3] Ambros, V. (2004). The functions of animal microRNAs. *Nature*, 431(7006):350–355.

[4] Aure, M. R., Jernström, S., Krohn, M., Vollan, H. K., Due, E. U., Rødland, E., Kåresen, R., Ram, P., Lu, Y., Mills, G. B., Sahlberg, K. K., Børresen-Dale, A. L., Lingjærde, O. C., and Kristensen, V. N. (2015). Integrated analysis reveals microRNA networks coordinately expressed with key proteins in breast cancer. *Genome Med.*, 7(1):21.

[5] Azim, H. A., Michiels, S., Zagouri, F., Delaloge, S., Filipits, M., Namer, M., Neven, P., Symmans, W. F., Thompson, A., Andre, F., Loi, S., and Swanton, C. (2013). Utility of prognostic genomic tests in breast cancer practice: The IMPAKT 2012 Working Group Consensus Statement. *Ann. Oncol.*, 24(3):647–654.

[6] Baek, D., Villen, J., Shin, C., Camargo, F. D., Gygi, S. P., and Bartel, D. P. (2008). The impact of microRNAs on protein output. *Nature*, 455(7209):64–71.

[7] Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297.

[8] Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–233.

[9] Bernstein, E., Caudy, A. A., Hammond, S. M., and Hannon, G. J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409(6818):363–366.

[10] Betel, D., Wilson, M., Gabow, A., Marks, D. S., and Sander, C. (2008). The microRNA.org resource: targets and expression. *Nucleic Acids Res.*, 36(Database issue):D149–153.

[11] Blenkiron, C., Goldstein, L. D., Thorne, N. P., Spiteri, I., Chin, S. F., Dunning, M. J., Barbosa-Morais, N. L., Teschendorff, A. E., Green, A. R., Ellis, I. O., Tavare, S., Caldas, C., and Miska, E. A. (2007). MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biol.*, 8(10):R214.

[12] Calin, G. A. and Croce, C. M. (2006). MicroRNA signatures in human cancers. *Nat. Rev. Cancer*, 6(11):857–866.

[13] Charboneau, L., Tory, H., Scott, H., Chen, T., Winters, M., Petricoin, E. F., Liotta, L. A., and Paweletz, C. P. (2002). Utility of reverse phase protein arrays: applications to signalling pathways and human body arrays. *Brief Funct Genomic Proteomic*, 1(3):305–315.

[14] Cheloufi, S., Dos Santos, C. O., Chong, M. M., and Hannon, G. J. (2010). A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. *Nature*, 465(7298):584–589.

[15] Chen, W., Zhou, S., Mao, L., Zhang, H., Sun, D., Zhang, J., Li, J., and Tang, J. H. (2016). Crosstalk between TGF-β signaling and miRNAs in breast cancer metastasis. *Tumour Biol.*

[16] Chendrimada, T. P., Gregory, R. I., Kumaraswamy, E., Norman, J., Cooch, N., Nishikura, K., and Shiekhattar, R. (2005). TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature*, 436(7051):740–744.

[17] Chou, C. H., Chang, N. W., Shrestha, S., Hsu, S. D., Lin, Y. L., Lee, W. H., Yang, C. D., Hong, H. C., Wei, T. Y., Tu, S. J., Tsai, T. R., Ho, S. Y., Jian, T. Y., Wu, H. Y., Chen, P. R., Lin, N. C., Huang, H. T., Yang, T. L., Pai, C. Y., Tai, C. S., Chen, W. L., Huang, C. Y., Liu, C. C., Weng, S. L., Liao, K. W., Hsu, W. L., and Huang, H. D. (2016). miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.*, 44(D1):D239–247.

[18] Chugh, P. and Dittmer, D. P. (2012). Potential pitfalls in microRNA profiling. *Wiley Interdiscip Rev RNA*, 3(5):601–616.

[19] Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M. F., Kellis, M., Lindblad-Toh, K., and Lander, E. S. (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U.S.A.*, 104(49):19428–19433.

[20] Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

[21] Czech, B., Zhou, R., Erlich, Y., Brennecke, J., Binari, R., Villalta, C., Gordon, A., Perrimon, N., and Hannon, G. J. (2009). Hierarchical rules for Argonaute loading in Drosophila. *Mol. Cell*, 36(3):445–456.

[22] Davis, S. and Meltzer, P. (2007). Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 14:1846–1847.

[23] de Rinaldis, E., Gazinska, P., Mera, A., Modrusan, Z., Fedorowicz, G. M., Burford, B., Gillett, C., Marra, P., Grigoriadis, A., Dornan, D., Holmberg, L., Pinder, S., and Tutt, A. (2013). Integrated genomic analysis of triple-negative breast cancers reveals novel microRNAs associated with clinical and molecular phenotypes and sheds light on the pathways they control. *BMC Genomics*, 14:643.

[24] Du, T. and Zamore, P. D. (2005). microPrimer: the biogenesis and function of microRNA. *Development*, 132(21):4645–4652.

[25] Dupuis, J. A. and Robert, C. P. (2003). Variable selection in qualitative models via an entropic explanatory power. *Journal of Statistical Planning and Inference*, 111(1):77–94.

[26] Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*, 4(8):1184–1191.

[27] Dweep, H. and Gretz, N. (2015). miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nat. Methods*, 12(8):697.

[28] Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30(1):207–210.

[29] Eiring, A. M., Harb, J. G., Neviani, P., Garton, C., Oaks, J. J., Spizzo, R., Liu, S., Schwind, S., Santhanam, R., Hickey, C. J., Becker, H., Chandler, J. C., Andino, R., Cortes, J., Hokland, P., Huettner, C. S., Bhatia, R., Roy, D. C., Liebhaber, S. A., Caligiuri, M. A., Marcucci, G., Garzon, R., Croce, C. M., Calin, G. A., and Perrotti, D. (2010). miR-328 functions as an RNA decoy to modulate hnRNP E2 regulation of mRNA translation in leukemic blasts. *Cell*, 140(5):652–665.

[30] Emmadi, R., Canestrari, E., Arbieva, Z. H., Mu, W., Dai, Y., Frasor, J., and Wiley, E. (2015). Correlative Analysis of miRNA Expression and Oncotype Dx Recurrence Score in Estrogen Receptor Positive Breast Carcinomas. *PLoS ONE*, 10(12):e0145346.

[31] Engelmann, J. C. and Spang, R. (2012). A least angle regression model for the prediction of canonical and non-canonical miRNA-mRNA interactions. *PLoS ONE*, 7(7):e40634.

[32] Eulalio, A., Behm-Ansmant, I., Schweizer, D., and Izaurralde, E. (2007). P-body formation is a consequence, not the cause, of RNA-mediated gene silencing. *Mol. Cell. Biol.*, 27(11):3970–3981.

[33] Fabian, M. R., Sundermeier, T. R., and Sonenberg, N. (2010). Understanding how miRNAs post-transcriptionally regulate gene expression. *Prog. Mol. Subcell. Biol.*, 50:1–20.

[34] Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., and Bray, F. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer*, 136(5):E359–386.

[35] Filipowicz, W., Bhattacharyya, S. N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.*, 9(2):102–114.

[36] Friedman, R. C., Farh, K. K., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, 19(1):92–105.

[37] Gao, L. B., Bai, P., Pan, X. M., Jia, J., Li, L. J., Liang, W. B., Tang, M., Zhang, L. S., Wei, Y. G., and Zhang, L. (2011). The association between two polymorphisms in pre-miRNAs and breast cancer risk: a meta-analysis. *Breast Cancer Res. Treat.*, 125(2):571–574.

[38] Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

[39] Goldhirsch, A., Wood, W. C., Gelber, R. D., Coates, A. S., Thurlimann, B., and Senn, H. J. (2007). Progress and promise: highlights of the international expert consensus on the primary therapy of early breast cancer 2007. *Ann. Oncol.*, 18(7):1133–1144.

[40] Gregory, R. I., Chendrimada, T. P., Cooch, N., and Shiekhattar, R. (2005). Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. *Cell*, 123(4):631–640.

[41] Gregory, R. I., Yan, K. P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., and Shiekhattar, R. (2004). The Microprocessor complex mediates the genesis of microRNAs. *Nature*, 432(7014):235–240.

[42] Griffiths-Jones, S. (2004). The microRNA Registry. *Nucleic Acids Res.*, 32(Database issue):D109–111.

[43] Grundhoff, A. and Sullivan, C. S. (2011). Virus-encoded microRNAs. *Virology*, 411(2):325–343.

[44] Guo, H., Ingolia, N. T., Weissman, J. S., and Bartel, D. P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308):835–840.

[45] Ha, M. and Kim, V. N. (2014). Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.*, 15(8):509–524.

[46] Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 100(1):57–70.

[47] Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674.

[48] Huang, J. C., Babak, T., Corson, T. W., Chua, G., Khan, S., Gallie, B. L., Hughes, T. R., Blencowe, B. J., Frey, B. J., and Morris, Q. D. (2007). Using expression profiling data to identify human microRNA targets. *Nat. Methods*, 4(12):1045–1049.

[49] Huang, J. C., Frey, B. J., and Morris, Q. D. (2008). Comparing sequence and expression for predicting microRNA targets using GenMiR3. *Pac. Symp. Biocomput.*, pages 52–63.

[50] Huang, Y., Zou, Q., Wang, S. P., Tang, S. M., Zhang, G. Z., and Shen, X. J. (2011). The discovery approaches and detection methods of microRNAs. *Mol. Biol. Rep.*, 38(6):4125–4135.

[51] Hunt, E. A., Broyles, D., Head, T., and Deo, S. K. (2015). MicroRNA Detection: Current Technology and Research Strategies. *Annu. Rev. Anal. Chem.*, 8:217–237.

[52] Ibanez-Ventoso, C., Vora, M., and Driscoll, M. (2008). Sequence relationships among C. elegans, D. melanogaster and human microRNAs highlight the extensive conservation of microRNAs in biology. *PLoS ONE*, 3(7):e2818.

[53] Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., and Liu, Y. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.*, 37(Database issue):98–104.

[54] Jones-Rhoades, M. W., Bartel, D. P., and Bartel, B. (2006). MicroRNAS and their regulatory roles in plants. *Annu. Rev. Plant. Biol.*, 57:19–53.

[55] Karhu, K. (2009). Exploring the effect of different microrna target prediction techniques. Master's thesis, Helsinki University of Technology.

[56] Kim, V. N. (2005). MicroRNA biogenesis: coordinated cropping and dicing. *Nat. Rev. Mol. Cell Biol.*, 6(5):376–385.

[57] Kim, V. N., Han, J., and Siomi, M. C. (2009). Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell Biol.*, 10(2):126–139.

[58] Knudson, A. G. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. U.S.A.*, 68(4):820–823.

[59] Kozomara, A. and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, 42(Database issue):68–73.

[60] Krek, A., Grun, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M., and Rajewsky, N. (2005). Combinatorial microRNA target predictions. *Nat. Genet.*, 37(5):495–500.

[61] Krol, J., Loedige, I., and Filipowicz, W. (2010). The widespread regulation of microRNA biogenesis, function and decay. *Nat. Rev. Genet.*, 11(9):597–610.

[62] Le, T. D., Zhang, J., Liu, L., and Li, J. (2015). Ensemble methods for miRNA target prediction from expression data. *PLoS ONE*, 10(6):e0131627.

[63] Lee, C. T., Risom, T., and Strauss, W. M. (2007). Evolutionary conservation of microRNA regulatory circuits: an examination of microRNA gene complexity and conserved microRNA-target interactions through metazoan phylogeny. *DNA Cell Biol.*, 26(4):209–218.

[64] Lee, L. W., Zhang, S., Etheridge, A., Ma, L., Martin, D., Galas, D., and Wang, K. (2010). Complexity of the microRNA repertoire revealed by next-generation sequencing. *RNA*, 16(11):2170–2180.

[65] Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854.

[66] Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., and Kim, V. N. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature*, 425(6956):415–419.

[67] Lee, Y., Kim, M., Han, J., Yeom, K. H., Lee, S., Baek, S. H., and Kim, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.*, 23(20):4051–4060.

[68] Lin, S. and Gregory, R. I. (2015). MicroRNA biogenesis pathways in cancer. *Nat. Rev. Cancer*, 15(6):321–333.

[69] Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B. L., Mak, R. H., Ferrando, A. A., Downing, J. R., Jacks, T., Horvitz, H. R., and Golub, T. R. (2005). MicroRNA expression profiles classify human cancers. *Nature*, 435(7043):834–838.

[70] Lund, E., Guttinger, S., Calado, A., Dahlberg, J. E., and Kutay, U. (2004). Nuclear export of microRNA precursors. *Science*, 303(5654):95–98.

[71] Mannsperger, H. A., Gade, S., Henjes, F., Beissbarth, T., and Korf, U. (2010). RPPanalyzer: Analysis of reverse-phase protein array data. *Bioinformatics*, 26(17):2202–2203.

[72] Melchor, L. and Benitez, J. (2013). The complex genetic landscape of familial breast cancer. *Hum. Genet.*, 132(8):845–863.

[73] Melo, S. A. and Esteller, M. (2011). Dysregulation of microRNAs in cancer: playing with fire. *FEBS Lett.*, 585(13):2087–2099.

[74] Miranda, K. C., Huynh, T., Tay, Y., Ang, Y. S., Tam, W. L., Thomson, A. M., Lim, B., and Rigoutsos, I. (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*, 126(6):1203–1217.

[75] miRBase (2014). miRBase: the microRNA database. [Online; accessed 17-February-2016].

[76] Muniategui, A., Nogales-Cadenas, R., Vazquez, M., Aranguren, X. L., Agirre, X., Luttun, A., Prosper, F., Pascual-Montano, A., and Rubio, A. (2012). Quantification of miRNA-mRNA interactions. *PLoS ONE*, 7(2):e30766.

[77] Muniategui, A., Pey, J., Planes, F. J., and Rubio, A. (2013). Joint analysis of miRNA and mRNA expression data. *Brief. Bioinformatics*, 14(3):263–278.

[78] Oesterreich, S. and Davidson, N. E. (2013). The search for ESR1 mutations in breast cancer. *Nat. Genet.*, 45(12):1415–1416.

[79] Ovaska, K., Laakso, M., Haapa-Paananen, S., Louhimo, R., Chen, P., Aittomäki, V., Valo, E., Ñúnez-Fontarnau, J., Rantanen, V., Karinen, S., Nousiainen, K., Lahesmaa-Korpinen, A. M., Miettinen, M., Saarinen, L., Kohonen, P., Wu, J., Westermarck, J., and Hautaniemi, S. (2010). Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Med*, 2(9):65.

[80] Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., and Bernard, P. S. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.*, 27(8):1160–1167.

[81] Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., Hayward, D. C., Ball, E. E., Degnan, B., Muller, P., Spring, J., Srinivasan, A., Fishman, M., Finnerty, J., Corbo, J., Levine, M., Leahy, P., Davidson, E., and Ruvkun, G. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–89.

[82] Payne, S. H. (2015). The utility of protein and mRNA correlation. *Trends Biochem. Sci.*, 40(1):1–3.

[83] Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A. L., Brown, P. O., and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752.

[84] Piironen, J. and Vehtari, A. (2015). Projection predictive variable selection using stan+r. *arXiv:1508.02502 [stat.ME]*.

[85] Piironen, J. and Vehtari, A. (2016). Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, pages 1–25.

[86] Place, R. F., Li, L. C., Pookot, D., Noonan, E. J., and Dahiya, R. (2008). MicroRNA-373 induces expression of genes with complementary promoter sequences. *Proc. Natl. Acad. Sci. U.S.A.*, 105(5):1608–1613.

[87] R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[88] Rubin, D. B. (1981). The bayesian bootstrap. *Ann. Statist.*, 9(1):130–134.

[89] Ruby, J. G., Jan, C. H., and Bartel, D. P. (2007). Intronic microRNA precursors that bypass Drosha processing. *Nature*, 448(7149):83–86.

[90] Saetrom, O., Snøve, O., and Saetrom, P. (2005). Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. *RNA*, 11(7):995–1003.

[91] Sah, S., McCall, M. N., Eveleigh, D., Wilson, M., and Irizarry, R. A. (2010). Performance evaluation of commercial mirna expression array platforms. *BMC Research Notes*, 3(1):1–6.

[92] Sales, G., Coppe, A., Bisognin, A., Biasiolo, M., Bortoluzzi, S., and Romualdi, C. (2010). MAGIA, a web-based tool for miRNA and Genes Integrated Analysis. *Nucleic Acids Res.*, 38(Web Server issue):W352–359.

[93] Slade, I., Bacchelli, C., Davies, H., Murray, A., Abbaszadeh, F., Hanks, S., Barfoot, R., Burke, A., Chisholm, J., Hewitt, M., Jenkinson, H., King, D., Morland, B., Pizer, B., Prescott, K., Saggar, A., Side, L., Traunecker, H., Vaidya, S., Ward, P., Futreal, P. A., Vujanic, G., Nicholson, A. G., Sebire, N., Turnbull, C., Priest, J. R., Pritchard-Jones, K., Houlston, R., Stiller, C., Stratton, M. R., Douglas, J., and Rahman, N. (2011). DICER1 syndrome: clarifying the diagnosis, clinical features and management implications of a pleiotropic tumour predisposition syndrome. *J. Med. Genet.*, 48(4):273–278.

[94] Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P. E., and Børresen-Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U.S.A.*, 98(19):10869–10874.

[95] Stan Development Team (2016). RStan: the R interface to Stan.

[96] Stingo, F. C., Chen, Y. A., Vannucci, M., Barrier, M., and Mirkes, P. E. (2010). A Bayesian graphical modeling approach to microRNA regulatory network inference. *Ann. Appl. Stat.*, 4(4):2024–2048.

[97] Strachan, T. and Read, A. (2011). *Human Molecular Genetics, 4th edition*. TGarland Science, Taylor & Francis Group.

[98] Suomen Syöpärekisteri (2016). http://www.cancer.fi/syoparekisteri. [Online; accessed 31-July-2016].

[99] Tavassoli, F. and Devilee, P., editors (2003). *Pathology and Genetics of Tumours of the Breast and Female Genital Organs*. World Health Organization Classification of Tumours. IARC Press: Lyon.

[100] Tchatchou, S., Jung, A., Hemminki, K., Sutter, C., Wappenschmidt, B., Bugert, P., Weber, B. H., Niederacher, D., Arnold, N., Varon-Mateeva, R., Ditsch, N., Meindl, A., Schmutzler, R. K., Bartram, C. R., and Burwinkel, B. (2009). A variant affecting a putative miRNA target site in estrogen receptor (ESR) 1 is associated with breast cancer risk in premenopausal women. *Carcinogenesis*, 30(1):59–64.

[101] Thomson, D. W., Bracken, C. P., and Goodall, G. J. (2011). Experimental strategies for microRNA target identification. *Nucleic Acids Res.*, 39(16):6845–6853.

[102] Tibes, R., Qiu, Y., Lu, Y., Hennessy, B., Andreeff, M., Mills, G. B., and Kornblau, S. M. (2006). Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Mol. Cancer Ther.*, 5(10):2512–2521.

[103] van der Pas, S. L., Kleijn, B. J. K., and van der Vaart, A. W. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electron. J. Statist.*, 8(2):2585–2618.

[104] van Iterson, M., Bervoets, S., de Meijer, E. J., Buermans, H. P., 't Hoen, P. A., Menezes, R. X., and Boer, J. M. (2013). Integrated analysis of microRNA and mRNA expression: adding biological significance to microRNA target predictions. *Nucleic Acids Res.*, 41(15):e146.

[105] van Rooij, E. and Kauppinen, S. (2014). Development of microRNA therapeutics is coming of age. *EMBO Mol. Med.*, 6(7):851–864.

[106] van Schooneveld, E., Wildiers, H., Vergote, I., Vermeulen, P. B., Dirix, L. Y., and Van Laere, S. J. (2015). Dysregulation of microRNAs in breast cancer and their potential role as prognostic and predictive biomarkers in patient management. *Breast Cancer Res.*, 17:21.

[107] VanGuilder, H. D., Vrana, K. E., and Freeman, W. M. (2008). Twenty-five years of quantitative PCR for gene expression analysis. *BioTechniques*, 44(5):619–626.

[108] Varmus, H. (1988). Retroviruses. *Science*, 240(4858):1427–1435.

[109] Vasudevan, S., Tong, Y., and Steitz, J. A. (2007). Switching from repression to activation: microRNAs can up-regulate translation. *Science*, 318(5858):1931–1934.

[110] Vehtari, A. and Ojanen, J. (2012). A survey of bayesian predictive methods for model assessment, selection and comparison. *Statist. Surv.*, 6:142–228.

[111] Vlachos, I. S., Paraskevopoulou, M. D., Karagkouni, D., Georgakilas, G., Vergoulis, T., Kanellos, I., Anastasopoulos, I. L., Maniou, S., Karathanou, K., Kalfakakou, D., Fevgas, A., Dalamagas, T., and Hatzigeorgiou, A. G. (2015).

DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res.*, 43(Database issue):D153–159.

[112] Vogel, C. and Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.*, 13(4):227–232.

[113] Wang, X. and El Naqa, I. M. (2008). Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*, 24(3):325–332.

[114] Weigelt, B. and Reis-Filho, J. S. (2009). Histological and molecular types of breast cancer: is there a unifying taxonomy? *Nat. Rev. Clin. Oncol.*, 6(12):718–730.

[115] Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75(5):855–862.

[116] Yan, M., Huang, H. Y., Wang, T., Wan, Y., Cui, S. D., Liu, Z. Z., and Fan, Q. X. (2012). Dysregulated expression of dicer and drosha in breast cancer. *Pathol. Oncol. Res.*, 18(2):343–348.

[117] Yu, F., Yao, H., Zhu, P., Zhang, X., Pan, Q., Gong, C., Huang, Y., Hu, X., Su, F., Lieberman, J., and Song, E. (2007). let-7 regulates self renewal and tumorigenicity of breast cancer cells. *Cell*, 131(6):1109–1123.

[118] Yue, D., Liu, H., and Huang, Y. (2009). Survey of Computational Algorithms for MicroRNA Target Prediction. *Curr. Genomics*, 10(7):478–492.

A Table of model properties

Table A1: Properties of fitted models for all 105 genes. A missing value for N miRNAs indicates a projected model was not found (i.e. the stopping criterion was not satisfied before reaching 200 covariates), a zero indicates no miRNA variables were chosen. R^2 was not computed for models with no miRNA variables. Only the significant miRNAs chosen are listed for compactness of display.
R^2_{gene}: R^2 for gene-only model
$R^2_{\perp}$: R^2 for projected model obtained with PPVS
$\Delta\bar{R}^2$: Difference of adjusted R^2 of projected model versus gene-only model
*: gene expression variable is significant (95% credible interval)
N_{miRNA}: number of miRNA variables in projected model (number of significant miRNA variables, 95% credible interval)

Gene	R^2_{gene}	$R^2_{\perp}$	$\Delta\bar{R}^2$	N_{miRNA}	**Significant miRNAs**
ACACA	0.462*	0.524*	0.062	2 (2)	miR-30a, miR-370
AKT1	0.111*	0.174*	0.064	2 (2)	miR-449a, miR-342-5p
AKT2	0.001	0.230	0.195	14 (4)	miR-342-5p, miR-449a, miR-96, miR-146b-5p
AKT3	0.000	0.251	0.216	15 (4)	miR-342-5p, miR-449a, miR-96, miR-146b-5p
ANXA1	0.380*	0.425*	0.047	1 (1)	miR-765
AR	0.713*			0	
BAK1	0.119*	0.228*	0.110	2 (2)	miR-29c, miR-505*
BAX	0.130*	0.405*	0.228	23 (4)	miR-557, miR-659, miR-142-3p, miR-199a-3p
BCL2	0.728*			0	
BCL2L1	0.088*	0.354*	0.123	53 (1)	miR-622
BCL2L11	0.176*	0.318*	0.143	2 (2)	miR-29c, miR-34a
BECN1	0.005				
BID	0.015*	0.134	0.122	1 (1)	miR-1246
BIRC2	0.018*	0.243*	0.226	2 (2)	miR-1246, miR-425
BRAF	0.094*	0.456*	0.352	8 (8)	miR-638, miR-125b, miR-1321, miR-107, miR-765, miR-148a, miR-135b, miR-505*
CASP8	0.014	0.111*	0.091	4 (4)	miR-99a, miR-148a, miR-631, miR-126*
CAV1	0.284*	0.432*	0.147	3 (3)	miR-551b, miR-24
CCNB1	0.638*	0.714*	0.076	2 (2)	miR-199a-5p, miR-30a
CCND1	0.334*	0.497*	0.153	8 (7)	miR-936, miR-181c, miR-622, miR-493*, miR-19a, miR-126, miR-9*
CCNE1	0.544*			0	
CDH1	0.449*			0	
CDH2	0.006			0	
CDH3	0.165*	0.584*	0.398	17 (13)	miR-155, miR-10b*, miR-502-5p, miR-224, miR-489, miR-148a, miR-195, miR-197, miR-361-5p, miR-650, miR-150, miR-501-5p, miR-582-5p
CDK1	0.037*				
CDKN1B	0.418*	0.491*	0.075	1 (1)	miR-195
CHEK1	0.019*	0.168	0.122	11 (0)	
CHEK2	0.512*			0	
CLDN7	0.234*	0.363*	0.125	4 (4)	miR-29c, miR-200c, miR-30b, miR-150
COL6A1	0.000	0.325	0.321	4 (4)	miR-125b, miR-638, miR-210, miR-24
CTNNA1	0.081*	0.113*	0.033	2 (1)	miR-125a-3p
CTNNB1	0.002	0.339	0.287	22 (5)	miR-711, miR-10a, miR-31*, miR-16, miR-28-5p

Continued on next page...

Gene	$R^2_{\mathbf{gene}}$	$R^2_{\perp}$	$\Delta\bar{R}^2$	$\mathrm{N}_{\mathbf{miRNA}}$	**Significant miRNAs**
DIABLO	0.089*	0.483*	0.336	31 (6)	miR-378, miR-339-3p, miR-762, miR-15b, miR-144*, miR-582-5p
DVL3	0.068*	0.377*	0.283	14 (9)	miR-24, miR-498, miR-140-3p, miR-223, miR-29c, miR-21, miR-432, miR-662, miR-204
EEF2	0.009	0.330	0.292	14 (3)	miR-106b, miR-196b, miR-29c*
EEF2K	0.310*			0	
EGFR	0.148*	0.356*	0.189	10 (7)	miR-181d, miR-181b, miR-1182, miR-495, miR-30c, miR-126, miR-183*
EIF4E	0.100*	0.252*	0.050	36 (0)	
EIF4EBP1	0.495*			0	
ERBB2	0.729*			0	
ERBB3	0.204*	0.334*	0.128	3 (3)	miR-199a-5p, miR-451, miR-484
ERCC1	0.004			0	
ERRFI1	0.012				
ESR1	0.825*			0	
FN1	0.495*			0	
FOXO3	0.094*	0.225*	0.128	3 (3)	miR-140-3p, miR-631, miR-197
GAB2	0.597*			0	
GATA3	0.708*			0	
GSK3A	0.106*	0.473*	0.319	26 (5)	miR-21, miR-29a, miR-20a, miR-100, miR-92a
GSK3B	0.064*	0.524*	0.275	81 (0)	
IGF1R	0.637*			0	
IGFBP2	0.553*			0	
INPP4B	0.754*			0	
IRS1	0.397*	0.441*	0.046	1 (1)	miR-93
KDR	0.000	0.281	0.272	6 (6)	miR-150, miR-663, miR-495, miR-24-1*, miR-363, miR-140-3p
KIT	0.581*			0	
KRAS	0.037*	0.335	0.264	16 (2)	miR-96, miR-21
MAP2K1	0.006	0.141	0.138	1 (1)	miR-21
MAPK14	0.017*	0.346*	0.301	14 (9)	miR-145, miR-92a, miR-181c, miR-142-3p, miR-425, miR-339-3p, miR-342-5p, miR-18b, miR-1226*
MAPK9	0.193*	0.251*	0.061	1 (1)	miR-342-5p
MAPT	0.020*	0.335	0.270	20 (3)	miR-30c, miR-132, miR-17*
MET	0.031*	0.085*	0.054	2 (2)	miR-125b, miR-139-5p
MRE11A	0.012	0.290	0.052	70 (0)	
MSH2	0.333*			0	
MSH6	0.340*	0.628*	0.265	19 (10)	miR-125b, miR-324-5p, miR-25, miR-195*, miR-451, miR-154, miR-551b, miR-26b, miR-513a-5p
MYC	0.025*	0.101*	0.076	2 (2)	miR-150*, miR-24
NCOA3	0.132*			0	
NF2	0.132*	0.321*	0.187	3 (3)	miR-638, miR-125b, miR-22
NOTCH1	0.216*	0.279*	0.064	2 (2)	miR-199b-5p, miR-502-5p
NOTCH3	0.150*	0.401*	0.245	5 (5)	miR-125b, miR-27b, miR-193a-5p, miR-32, miR-139-5p
PARK7	0.288*	0.428*	0.140	2 (2)	miR-93, miR-29c
PCNA	0.229*	0.283*	0.057	1 (1)	miR-199a-5p
PECAM1	0.054*	0.169*	-0.028	43 (0)	
PGR	0.819*			0	
PIK3CA	0.051*			0	

Continued on next page...

Gene	$R^2_{\mathbf{gene}}$	$R^2_{\perp}$	$\Delta\bar{R}^2$	$\mathrm{N}_{\mathbf{miRNA}}$	**Significant miRNAs**
PIK3R1	0.077*	0.377*	0.295	5 (5)	miR-142-3p, miR-342-3p, miR-501-5p, miR-145*, miR-92a
PRKAA1	0.119*	0.240*	0.121	2 (2)	miR-199a-3p, miR-342-3p
PRKCA	0.368*			0	
PTCH1	0.029*	0.167*	0.127	6 (2)	miR-145, miR-934
PTEN	0.029*	0.468*	0.370	35 (5)	miR-204, miR-498, miR-324-3p, miR-30d, miR-22
PTGS2	0.065*	0.114*	0.052	1 (1)	miR-1246
PTK2	0.021*	0.273*	0.242	6 (5)	miR-1246, miR-150, miR-21, miR-200c, miR-32
PXN	0.133*	0.431*	0.140	63 (0)	
RAB25	0.031*				
RAD50	0.076*	0.106*	0.033	1 (1)	miR-139-5p
RAD51	0.011	0.205	0.195	2 (2)	miR-1246, miR-181c
RAF1	0.121*	0.446*	0.306	12 (9)	miR-125b, miR-1260, miR-498, miR-449a, miR-130a, miR-30b, miR-28-5p, miR-374a, miR-195*
RB1	0.006	0.233*	0.222	4 (4)	miR-145, miR-1260, miR-451, miR-195*
RPS6KB1	0.516*	0.617*	0.102	2 (2)	miR-497, miR-106b
SMAD1	0.359*	0.417*	0.060	1 (1)	miR-106b
SMAD3	0.354*	0.475*	0.112	7 (7)	miR-7, miR-451, miR-181d, miR-29b, miR-365, miR-1225-5p, miR-1299
SMAD4	0.000			0	
SNAI1	0.001			0	
SRC	0.308*			0	
STAT5A	0.292*	0.438*	0.147	2 (2)	miR-155, miR-324-5p
STMN1	0.000	0.364	-0.028	109 (0)	
SYK	0.227*	0.714*	0.462	26 (20)	miR-125b, miR-155, miR-324-5p, miR-195, miR-449a, miR-711, miR-762, miR-940, miR-22, miR-615-3p, miR-663, miR-377*, miR-598, miR-126, miR-29b, miR-1228*, miR-144, miR-204, miR-601, miR-30d
TP53	0.017*	0.186*	0.161	5 (5)	miR-301b, miR-29c*, miR-551b, miR-19a, miR-874
TP53BP1	0.260*	0.357*	0.099	1 (1)	miR-96
TSC2	0.065*	0.318*	0.225	13 (3)	miR-1915, miR-30b, miR-92a
VASP	0.151*	0.193*	0.045	1 (1)	miR-29c*
XIAP	0.089*	0.224*	0.136	2 (2)	miR-638, miR-99a
XRCC1	0.130*			0	
YAP1	0.164*	0.351*	0.183	4 (4)	miR-106b, miR-486-5p, miR-28-5p, miR-595
YBX1	0.016*	0.329*	0.314	2 (2)	miR-125b, miR-96
YWHAE	0.001	0.145	0.138	4 (4)	miR-21, miR-1260, miR-365*, miR-204

B Model size distributions

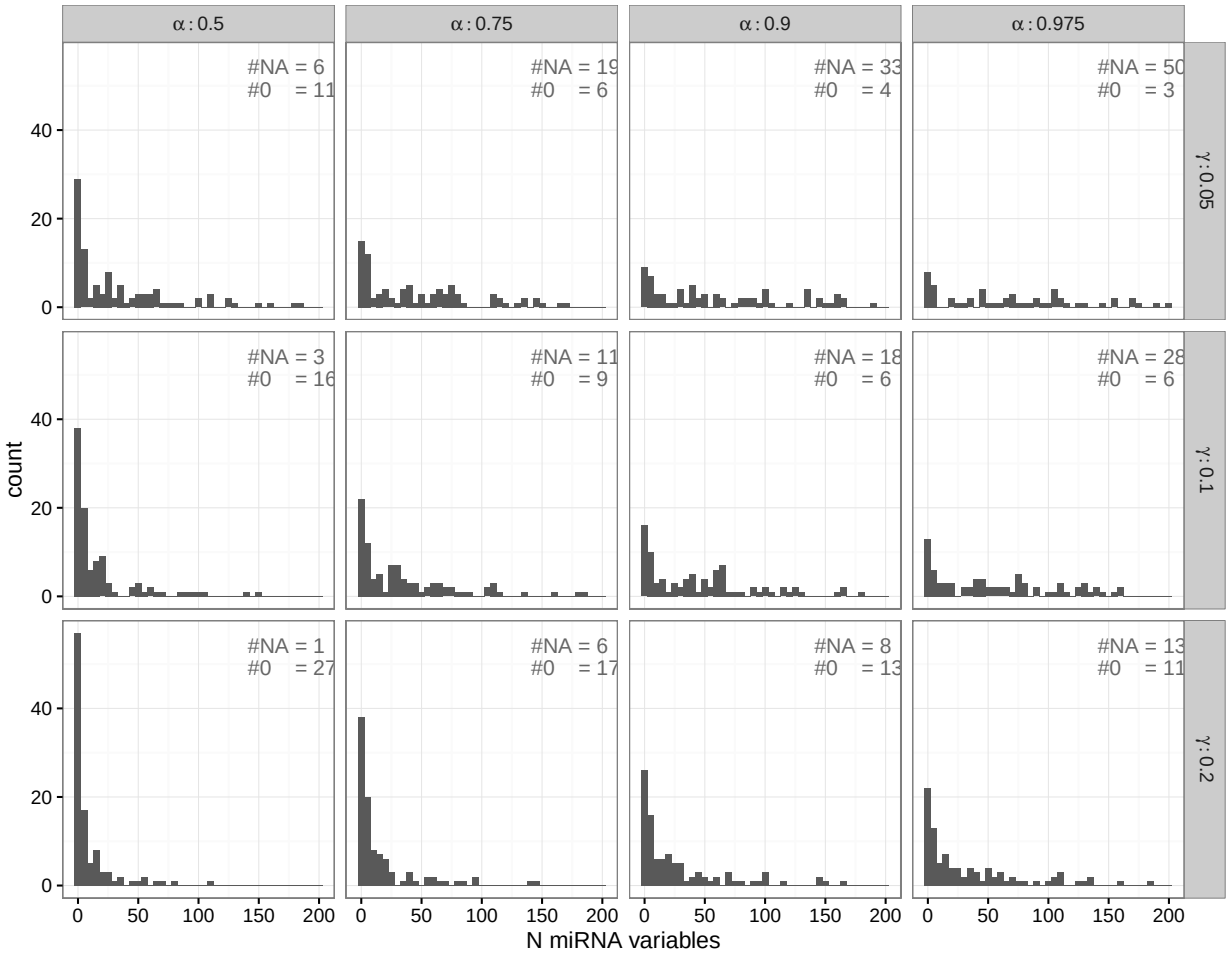
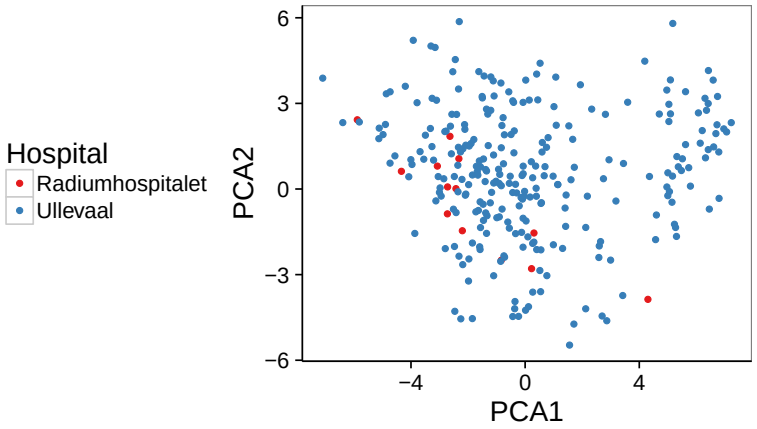
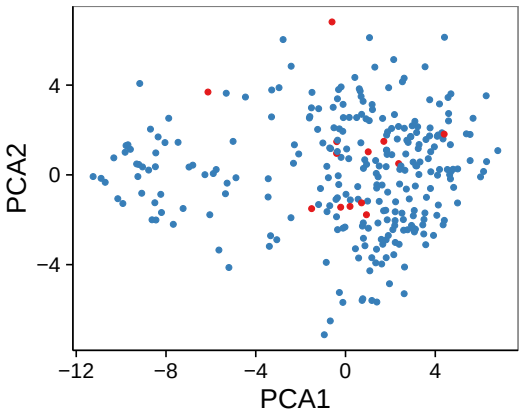


Figure B1: Distributions of chosen model sizes for differrent values of model-size parameters α and γ. $\#0$ refers to the number of models with no covariates (i.e. not even the mRNA covariate was chosen) and #NA to the number of models where the model-size criterion was not met. The parameter values had a large impact on the final model sizes. Strict values (α close to one and small γ) generated very large models and had the effect of the size-criterion not being met for many genes (larger #NA).

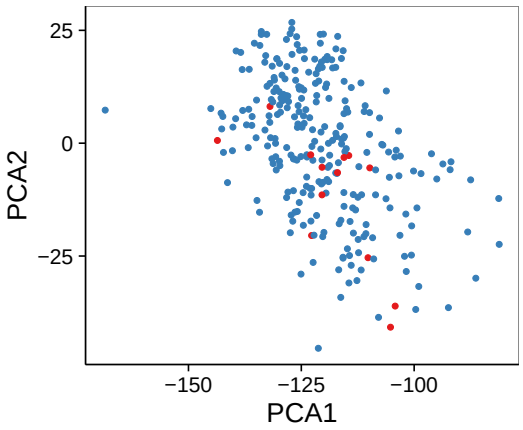
C Quality control plots



(a) protein

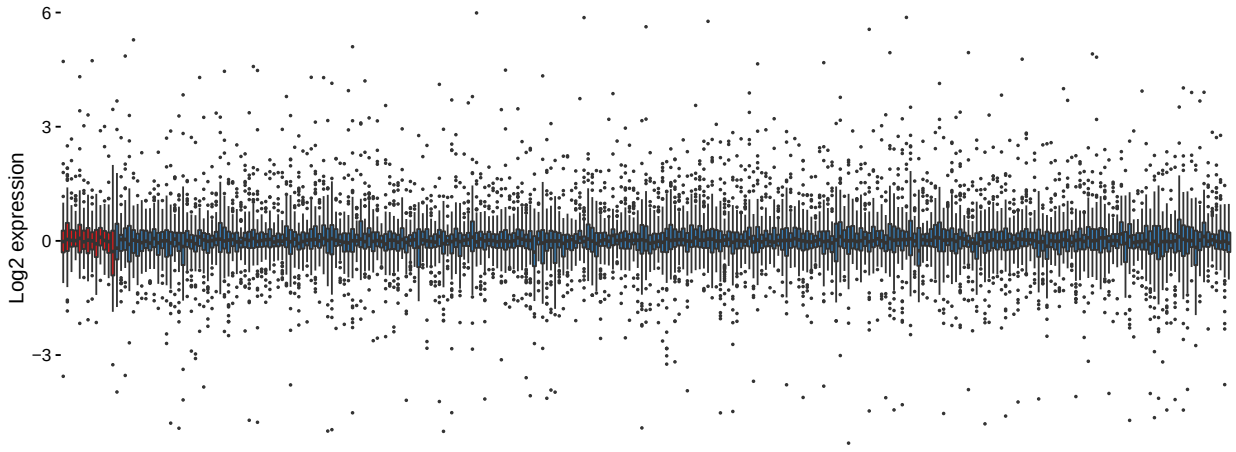


(b) mRNA

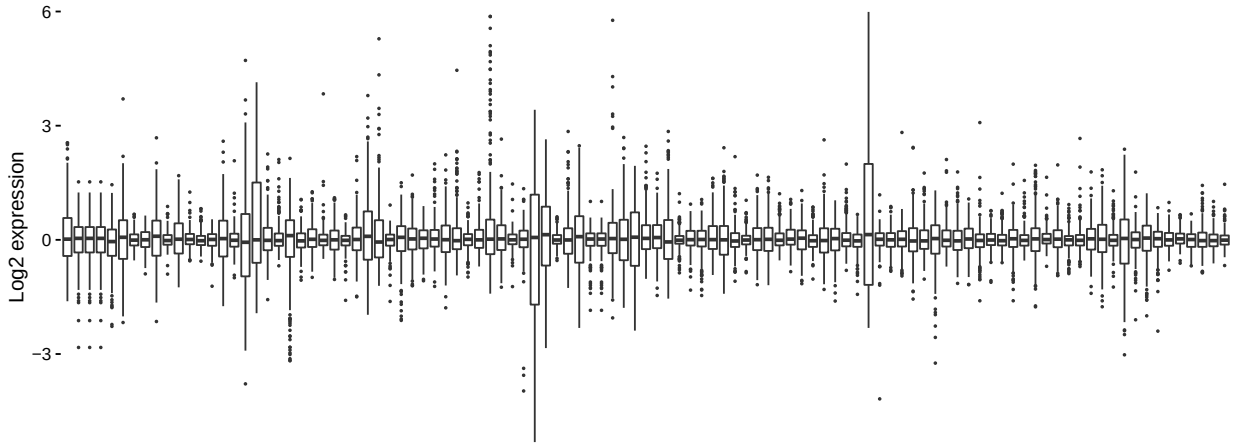


(c) miRNA

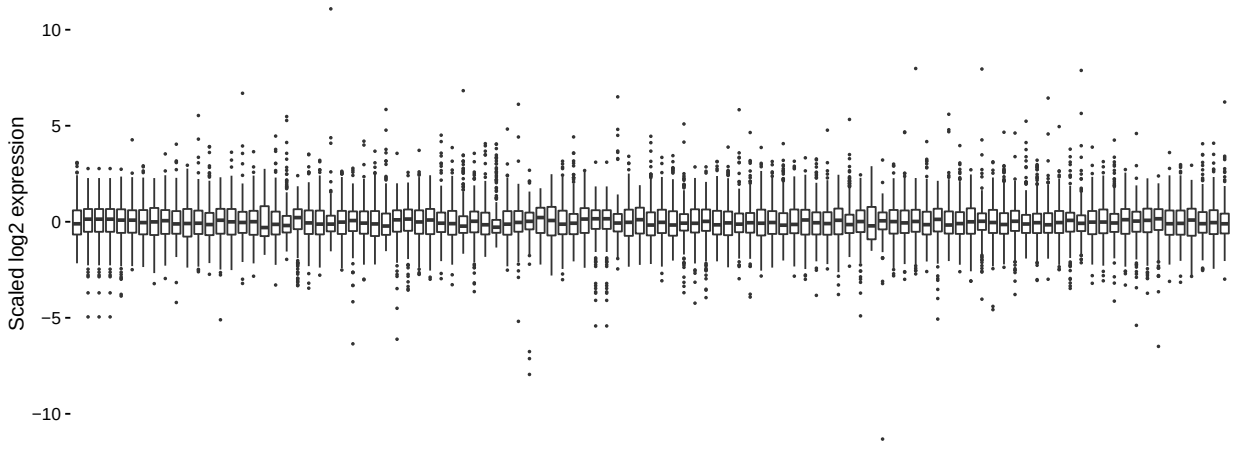
Figure C1: Scatter plots of first two principal components for each tumor sample (or microarray) computed from (a) protein expression, (b) mRNA expression, and (c) miRNA expression data. The point color corresponds to the hospital where each sample was handled. The samples from different hospitals were not distinguishable using the principal components, which suggests that no significant hospital batch effect was present.



(a) Protein expression by tumor samples



(b) Protein variables



(c) Scaled protein variables ($\mu_y = 0, \sigma_y = 1$)

Figure C2: Distribution of protein expression, grouped by (a) tumor samples, (b) protein variables (i.e. microarrays), and (c) protein variables (and scaled to zero mean and unit variance). The fill color in (a) corresponds to the hospital where the sample was collected. It is evident in (c), that some of the protein variables have significant outliers (especially the two values beyond 10 and -10).

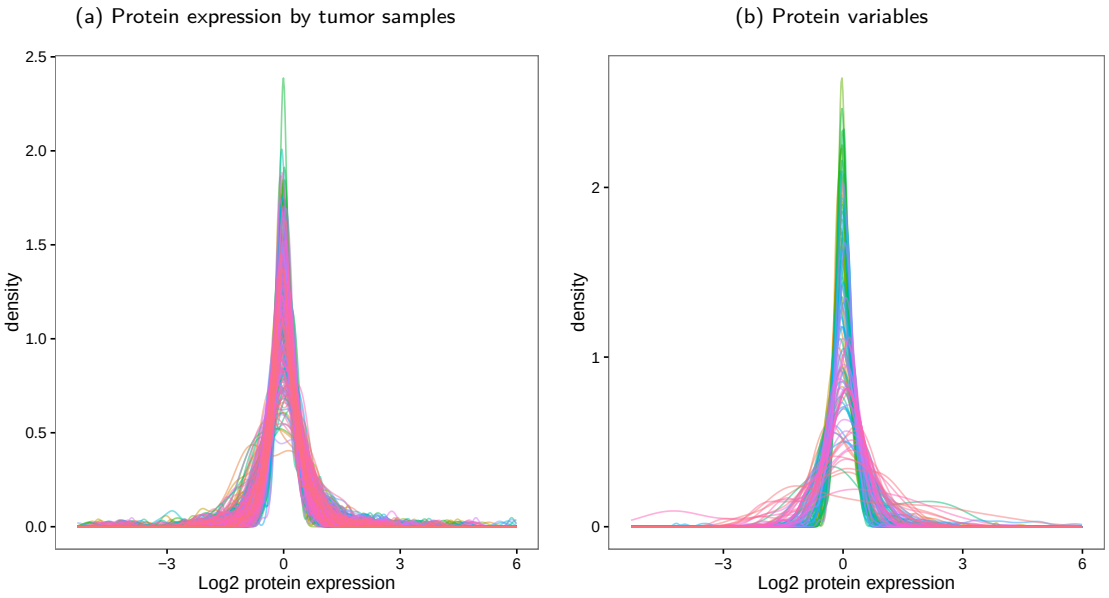
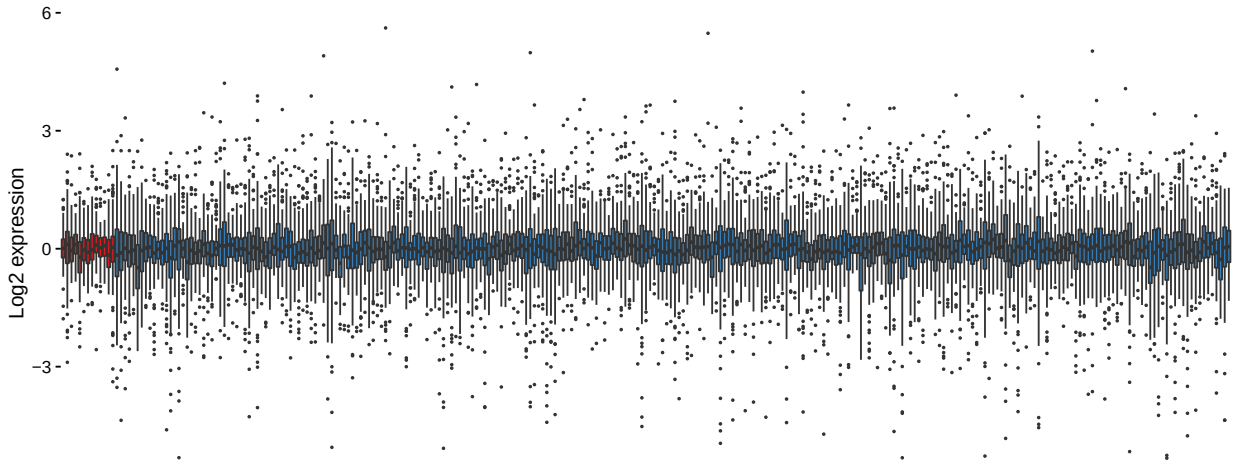
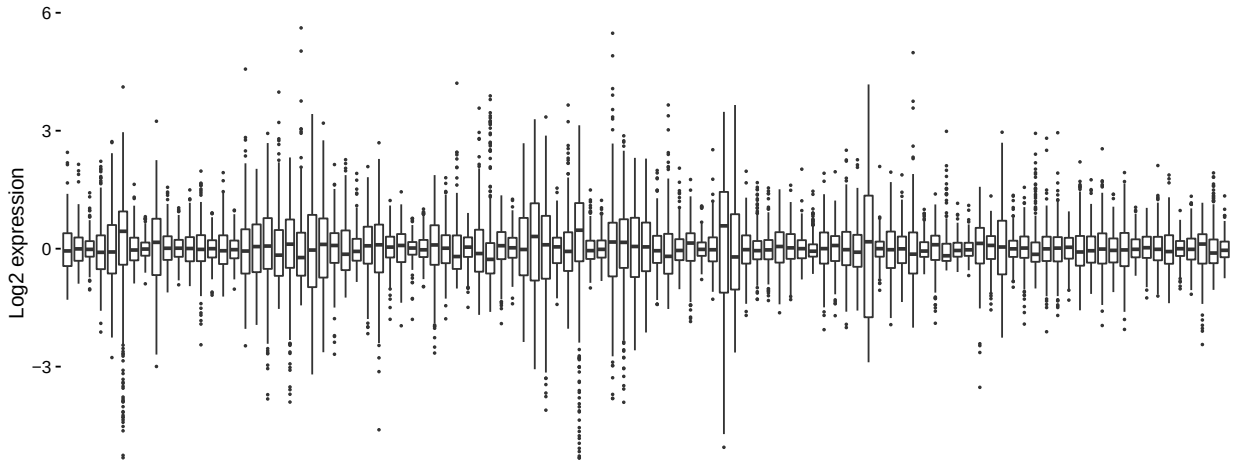


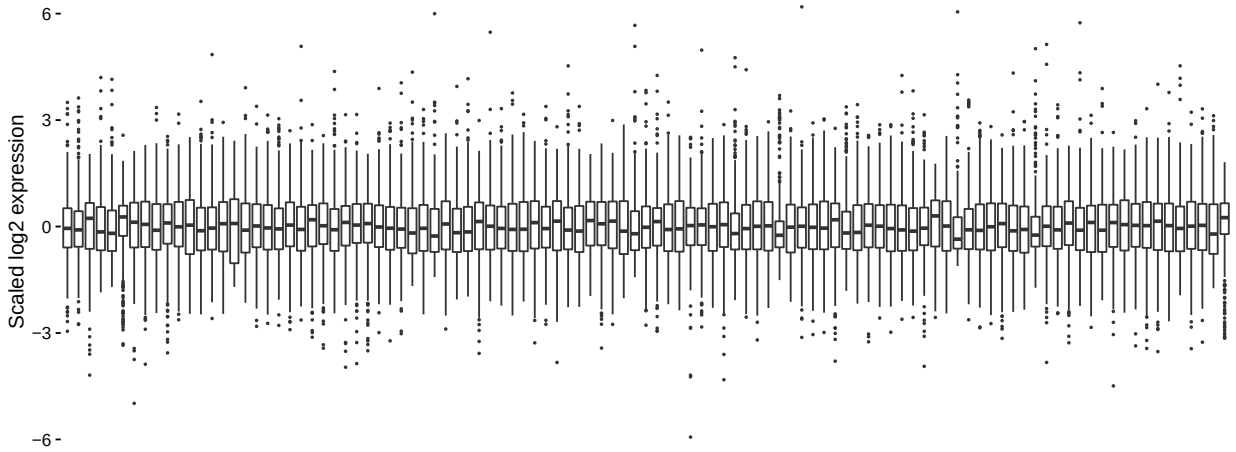
Figure C3: Density estimates of protein expression data for each tumor sample (a) and each protein variable (b). The color of the curves has no significance. Distributions were generally uniform, but long tails for several protein variables are visible in (b), suggesting the presence of significant outliers and that these variables were not approximated well by a normal ditribution.



(a) mRNA microarrays



(b) mRNA variables



(c) Scaled mRNA variables ($\mu_z = 0, \sigma_z = 1$).

Figure C4: Distribution of mRNA expression, grouped by (a) tumor samples (i.e. microarrays), (b) mRNA variables, and (c) mRNA variables (and scaled to zero mean and unit variance). The fill color in (a) corresponds to the hospital where the sample was collected. The distributions were fairly uniform and only a few significant outliers (further than 5 standard devations from the mean) were present in the mRNA variables.

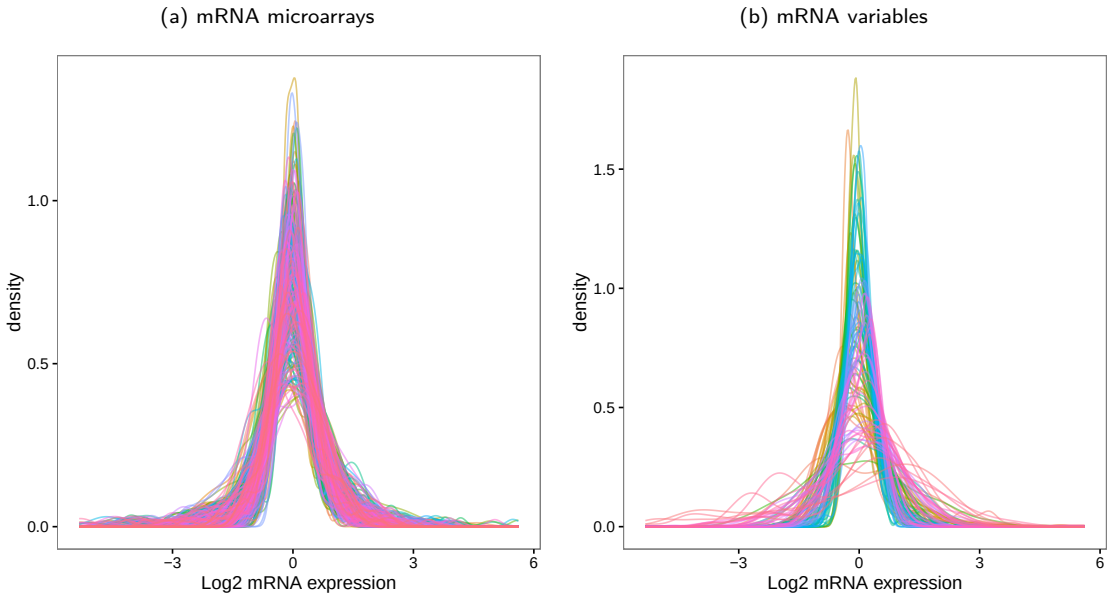
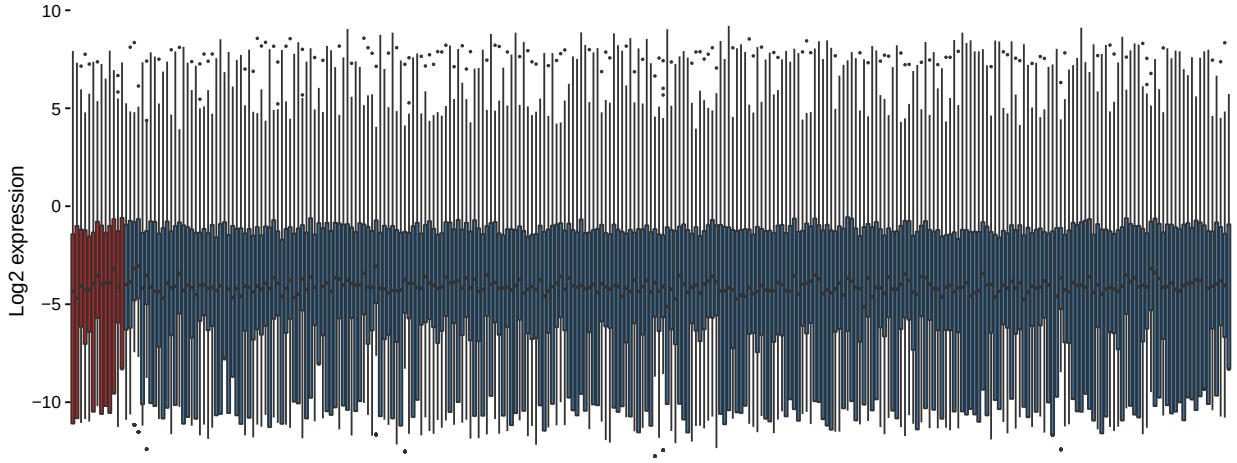
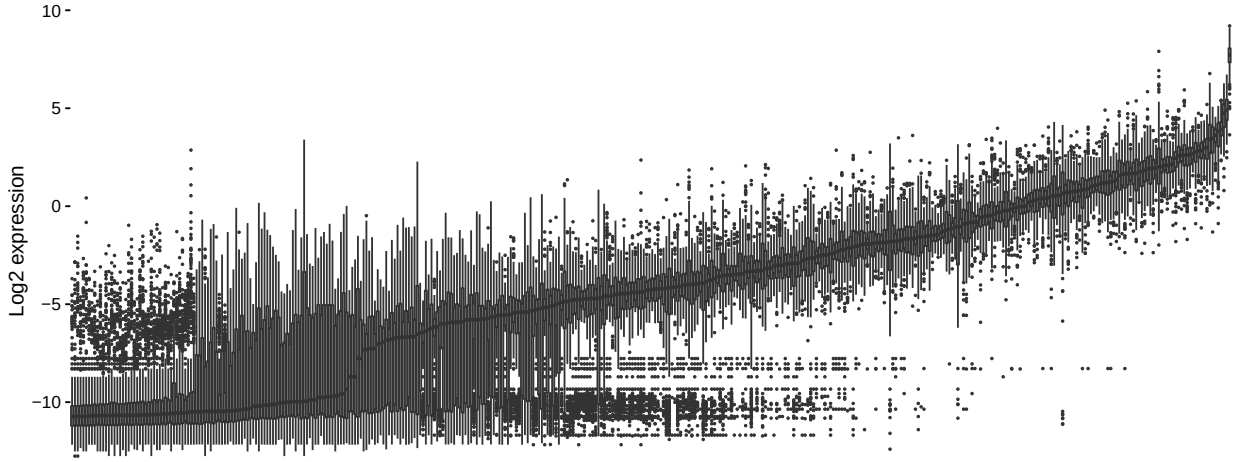


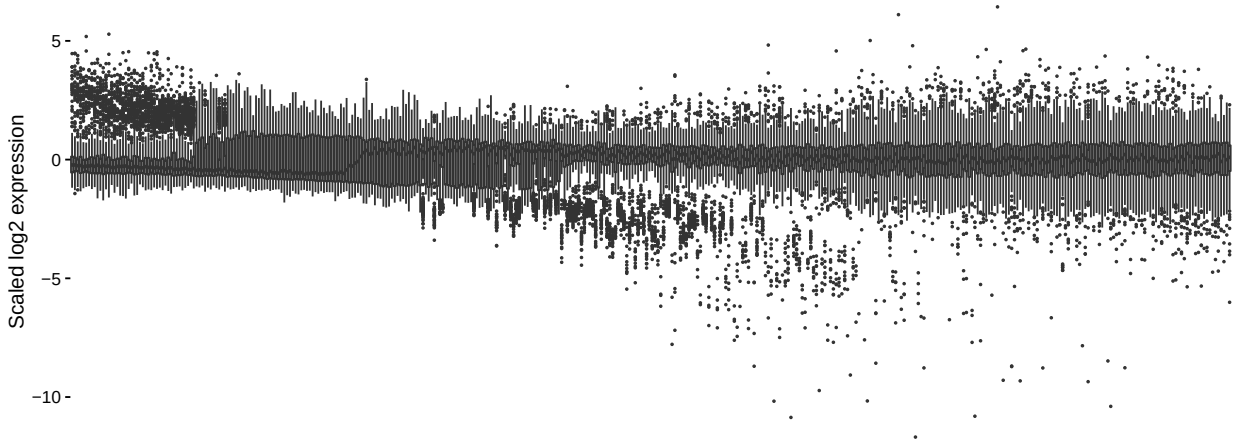
Figure C5: Density estimates of mRNA expression data for each microarray (a) and each mRNA variable (b). The color of the curves has no significance. The array distributions were uniform and quite close to normal. Some of the variable distributions had long tails as seen in (b).



(a) miRNA microarrays



(b) miRNA variables



(c) Scaled miRNA variables ($\mu_x = 0, \sigma_x = 1$)

Figure C6: Distribution of miRNA expression, grouped by (a) tumor samples (i.e. microarrays), (b) miRNA variables, and (c) miRNA variables (and scaled to $\mu_x = 0, \sigma_x = 1$). Fill in (a) corresponds to hospital. The miRNA variables in (b) have been sorted by median expression to highlight that some miRNAs had very low expression. There appeared to be a gap at around -8 (confirmed in Fig. C7). Measurements below the gap possibly corresponded to miRNAs not expressed in the samples, and thus, likely consisted mainly of background noise, Real miRNA abundances should follow a more continuous distribution. There also seemed to be some technical artifact around the gap, where some miRNA variables had exactly the same expression value. No biological phenomenon would explain this. The same order is retained in (c). The scaling did not correct for the highly skewed distributions of the most lowly expressed miRNAs.

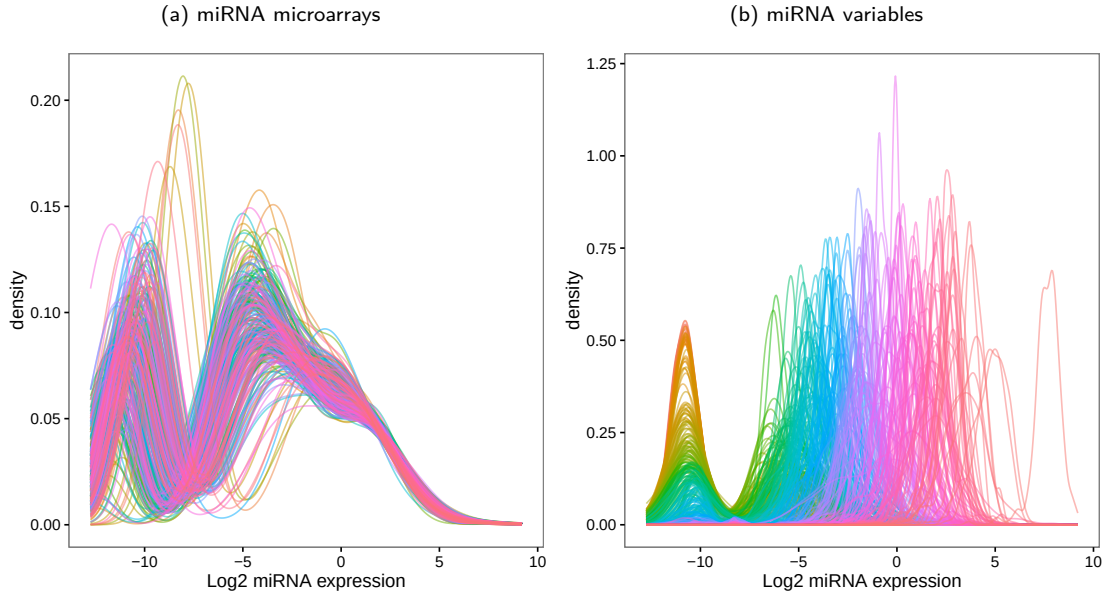


Figure C7: Density estimates of miRNA expression data for each microarray (a) and each miRNA variable (b). Color of the curves has no significance. Note the bimodal distributions in (a) and the corresponding break at around −8 and large spread in location in (b). This suggests that the miRNAs below the gap were present in such low quantities, possibly not expressed at all, that the measured expression values likely consisted mostly of background noise.