Identification of intronic microRNAs altered in breast cancer through microarray host gene expression analysis

Giovanni d'Ario

European School Of Molecular Medicine

IFOM-IEO-Campus

Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

Academic year 2009-2010

Supervisor : Pier Paolo Di Fiore

Added Co-Supervisor : Fabrizio Bianchi

External Co-Supervisor : Carlos Caldas

Internal Co-Supervisor : Francesca Ciccarelli

Abstract

MicroRNA (miRNAs) are endogenous non-coding RNAs of $\sim$ 22 nucleotides in length that function as post-transcriptional regulators of gene and protein expression through degradation or translation inhibition of the target messenger RNAs. MiRNAs show altered expression profiles in several human pathologies, including cancer. They can act as tumour suppressors or as oncogenes, depending on the characteristics of their target genes.

More than half of the mammalian miRNAs, including several of the miRNAs implicated in breast cancer, are localized within the introns of protein-coding genes, often organized in clusters, and usually transcribed together with their host gene. It is therefore possible, at least in principle, to identify novel intronic cancer-regulated miRNAs by examining the expression profile of their host genes by means of microarrays. For this purpose, we analyzed the regulation of 253 miRNA host genes in five large breast cancer microarray data sets comprising more than 950 samples, examining their association with different clinical and pathological parameters such as tumour grade, estrogen and progesterone receptor status, p53 status, survival, and occurrence of relapse or metastasis.

We found that *MCM7* and *SMC4* were the most frequently and significantly overexpressed genes in high grade tumours. These genes contain two well known cancer-associated miRNA clusters: *miR25-93-106b* and *miR-15b/16-2* respectively. In addition we identified six other miRNA host genes that were significantly downregulated in high grade tumours in all the data sets. Much less evidence is available in the literature about

the involvement in cancer of the miRNAs contained in these genes (*i.e.*, *miR-218-1*, *miR-342*, *miR-483*, *miR-548f-2*, *miR-1245* and *miR-1266*).

We measured the expression of the selected miRNAs by Real Time PCR on an independent cohort of 36 formalin-fixed paraffin-embedded (FFPE) samples, and we observed reduced expressions level of such miRNAs in high grade tumours. In particular, we found *miR-342-3p*, *miR-342-5p*, *miR-483-3p*, and *miR-483-5p* to be the most significantly downregulated miRNAs. These miRNAs were also found to correlate with bad prognosis in grade 2 tumours. Finally we provided initial evidence that increased expression of *miR-342-5p*, but not *miR-342-3p*, induces apoptosis in the highly metastatic MDA-MB-231 breast cancer cell line.

Acknowledgements

First of all I would like to thank Prof. Pier Paolo Di Fiore for giving me the opportunity to work in his group, and for allowing me to complete this project despite all the adversities.

I am indebted to Fabrizio Bianchi for giving me the possibility to work on a fascinating project, and for compensating my somewhat over-pessimistic physicist-like attitude towards biological data.

I would also like to thank my internal co-supervisor Francesca Ciccarelli, and my external co-supervisor Prof. Carlos Caldas for suggestions and support.

My gratitude goes to Francesco Nicassio and Matteo Marzi for their help, suggestions, and especially patience in the many occasions when they explained me the details of the experimental procedures and the data organization. Similarly I would like to thank Stefano Marchesi, Chiara Tordonato, Carmela Mazzoccoli, Marisa Oppizzi, and Simona Monterisi for the clarity and depth of their explanations during our meetings.

I cannot count the times that James Reid helped me with the R programming language, the LaTeX document typesetting system, and so many other things that it would take too long to mention them all. Lara Lusa has been my statistical lighthouse. Before meeting her, I would have never imagined that statistics could be fun. Exactly one year ago David Cairns joined the institute, and I cannot help thanking him for the help, the good suggestions, and the good time spent together. With Alessandro Brozzi I have shared so many academic, intellectual, and above all human experiences, that I simply don't know what to say. These people,

James, Lara, Alessandro and David, have been and are to me, above anything else, precious friends. I would also like to acknowledge the essential role played by the "B4 Unit", where some of the best (and the worst) ideas have been conceived and developed.

I would like to thank Manuela Vecchi, for her help, from a scientific and a human point of view. I am deeply indebted to her.

I also would like to thank Stefano Confalonieri, Giovanni Mazzarol, Maria Capra, Micaela Quarto, Giovanna Jodice, Chiara Luise and all the molecular pathology unit for their work and their help.

I would like to acknowledge Prof. Giuseppe Viale for providing the breast cancer samples we used to validate our findings.

Huge thanks to Rosalind Gunby for her invaluable help in turning an awful collection of badly written sentences into something readable, and for showing me that, in order to make something clear, it is often better to subtract, rather than to add.

To all the bioinformaticians, past and present, with whom I have shared ideas, troubles, rants, laughs and lunches, thank you.

I have a debt of gratitude with Lorenzo, John, Ivan, and all the security staff, past and present, for their kindness and for putting me in a good mood at the beginning and at the end of every working day simply by chatting a bit about life's small things. I would also like to thank Claudia, Amelia, and all the staff of the Bar (Pino, where are you?) for the same reason.

Meeting so many people from so many parts of the world, is one of the things I am most grateful to IFOM for. It made me richer inside, and some of these people have become some of my dearest friends. I will not even try to make a list of their names, because I would certainly leave someone out. However, I would like to acknowledge most sincerely all of them. They have been, and are, the most important part of this

experience.

Finally the biggest thanks go to my mother, who supported and encouraged me even when she was the one who needed to be encouraged and supported, especially in the days when I was not with her while she needed my presence. She will never read these words, but I simply would like to tell her "thank you for *everything*". Simply this.

Acronyms

siRNA Short interfering RNA.

miRNA MicroRNA.

piRNA Piwi-interacting RNA.

dsRNA Double stranded RNA.

RNAi RNA interference.

RISC RNA-induced silencing complex.

TU Transcriptional unit.

Pol II (III) RNA polymerase II (III).

pri-miRNA primary microRNA.

pre-miRNA precursor microRNA.

DGCR8 DiGeorge syndrome critical region protein.

RIIID RNase III domain.

dsRBD Double-stranded RNA-binding domain.

ssRNA Single stranded RNA.

SILAC Stable isotope labeling with amino acids in cell culture.

ORF Open reading frame.

CML Chronyc Myeloid Leukemia.

CLL Chronyc Lymphocytic Leukemia.

BCL2 B cell leukaemia/lymphoma 2.

ER Oestrogen Receptor.

PgR Progesterone Receptor.

IHC Immunohistochemistry.

PPV Positive Predictive Value.

EGFR Epidermal Growth Factor Receptor.

GEO Gene Expression Omnibus.

SV40 Simian vacuolating virus 40.

MGI Mouse Genome Informatics.

RLE Relative Log Expression.

NUSE Normalized Unscaled Standard Error.

IQR Interquartile range.

snRNA Small nuclear RNA.

TD Terminally differentiated.

CtBP C-terminal binding protein.

TRRAP transformation/transcription domainassociated protein.

FDR False discovery rate.

MAPK mitogen-activated protein kinase.

FFPE Formalin-fixed paraffin embedded.

IGF2 insulin-like growth factor 2.

TSS Transcription start site.

TFBS Transcription factor binding sites.

GGI gene expression grade index.

FACS fluorescence-activated cell sorting.

Contents

List of Figures

List of Tables

1

Introduction

1.1 microRNAs

1.1.1 Historical Background

In 1993, in a much cited article, Victor Ambros and colleagues, Rosalind Lee and Rhonda Feinbaum, discovered that *lin-4*, a gene known to control the timing of larval development in *C. elegans*, produces two small RNAs instead of coding for a protein (Lee et al., 1993). One of the two RNAs was found to be approximately 22 nucleotide (nt) in length, while the other was about 60 nt. The longer one was predicted to fold into a stem-loop, which in turn was proposed to be the precursor of the shorter one. The Ambros and Ruvkun labs than noticed that these *lin-4* RNAs had antisense complementarity to multiple sites in the $3'$ UTR of the *lin-14* gene (Lee et al., 1993; Wightman et al., 1993). This complementarity interval fell in a region previously proposed to mediate the repression of *lin-14* by the *lin-4* gene product (Ambros, 1989). The Ruvkun lab then demonstrated that *lin-4* mediated regulation of *lin-14*, through these 3' UTR sites, substantially reduces the amount of LIN-14 protein, without noticeably changing *lin-14* mRNA levels. Together these findings supported a model in which the *lin-4* RNAs pair to the *lin-14* $3'$ UTR to specify translational repression of *lin-14*. This repression forms part of a regulatory pathway that triggers the transition from cell divisions of the first larval stage to those of the second stage (Lee et al., 1993; Wightman et al., 1993).

1. INTRODUCTION

Lin-4 is today regarded as the first example of an abundant class of small, regulatory, non-coding RNAs called microRNAs (or miRNAs, for short), although the word microRNA appeared only in 2001 (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). Surprisingly, it took seven years after the discovery of the *lin-4* RNA to find another member of this class, during which time interest in this family of small RNAs approached zero. One of the reasons for this long period of oblivion, was the fact that, despite extensive efforts to screen more distantly related nematodes for *lin-4*, this RNA remained the only example of a small, non-coding, regulatory RNA (Ambros, 2008). Thus this phenomenon was believed to be a peculiarity of the developmental timing mechanism of *C. elegans*. On the other hand, a number of findings between 1993 and 2000 provided hints for a greater generality of the *lin-4* – *lin-14* mechanism of action. It was found, for example, that *lin-4* regulates not only *lin-14*, but also *lin-28* through complementary elements in its 3′ UTR (Moss et al., 1997). This discovery was suggestive of a potential evolutionary flexibility for this type of antisense interaction.

In 2000 the Ruvkun lab identified a second small RNA in *C. elegans*, encoded by the *let-7* gene (Reinhart et al., 2000), another gene controlling developmental timing in the worm. Like the *lin-4* RNA, the 21 nt *let-7* RNA is generated from a double-stranded hairpin precursor, and it controls the production of yet another developmental timing regulatory molecule, *lin-41*. Furthermore, the *let-7* RNA was found to work through imprecise antisense base-pairing with 3′ UTR sequences of its target gene.

Another surprising finding was the observation that in most animals, except *C. elegans*, *lin-4* and *let-7* are clustered closely together in the genome and are apparently transcriptionally co-regulated (Bashirullah, 2003; Sempere et al., 2003; Sokol et al., 2008). In the following years many examples of miRNA clusters have been found. There is still little understanding of why certain miRNAs occur in clusters in the genome, often transcriptionally co-expressed, and often from polycistronic primary transcripts.

Another fundamental piece in the miRNA puzzle was added when Gary Ruvkun

and co-workers showed that the *let-7* RNA is perfectly conserved across a wide range of animal phyla (Pasquinelli et al., 2000). This finding indicated that *let-7* and *lin-4* were members of an evolutionarily ancient class of regulatory molecules, thus suggesting that the existence of other RNA with the same characteristics, not only probable, but necessary. Three laboratories, Ambros's, Tuschl's and Bartel's, undertook the search for new miRNAs almost simultaneously, and their findings were published on the same issue of *Science* (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001).

The fact that the *let-7* RNA sequence is perfectly conserved across a vast evolutionary distance added new layers of complexity to the already fairly sophisticated emerging picture. At that time, there was no reason to expect that a 22 nt sequence would be conserved to such a large extent.

1.1.2 MicroRNA biogenesis and mechanisms of actions

1.1.2.1 MicroRNAs and short interfering RNAs share much of the molecular machinery

MiRNAs are a remarkable example of a more general category of small ($\sim 20-30$ nt) non-coding RNAs that regulate genes and genomes. This regulation can occur at different levels, either separately or simultaneously, including at the level of transcription, translation, chromatin structure, chromosome segregation, RNA processing and RNA stability. The central theme is that the small RNAs serve as specificity factors that directly bind effector proteins to target nucleic acid molecules via base-paring interactions. The core component of the effector machinery is always a member of the Argonaute protein super-family (Carthew and Sontheimer, 2009).

Many classes of small RNAs have emerged in recent years, the peculiarities of their origins, structures, associated effector proteins and biological roles have led to the identification of three main categories: short interfering RNAs (siRNAs), miRNAs and piwi-interacting RNAs (piRNAs). These RNAs are known to be present in eukaryotes only, although the Argonaute proteins that function in eukaryotic silencing can also be found in scattered bacterial and archaeal species. The first two

categories, siRNAs and miRNAs, are the most broadly distributed, while piRNAs are primarily found in animals and are less well understood (Carthew and Sontheimer, 2009).

Five years after the discovery of the first miRNA, *lin-4*, by Ambros and coworkers in 1993, Andrew Fire, Craig Mello and colleagues reported that exogenous double-stranded RNA (dsRNA) specifically silences genes through a mechanism called RNA interference (RNAi) (Fire et al., 1998). For this discovery they were awarded the Nobel Prize in Physiology and Medicine in 2006. In 1999, silencing in plants was shown to be accompanied by the appearance of $\sim 20-25$ nt RNAs that matched the sequence of the silencing trigger, and very shortly thereafter, the direct conversion of dsRNAs into $\sim 21-23$ nt siRNAs was documented. By 2001 the two categories of small RNAs had become well established: miRNAs as regulators of endogenous genes, and siRNAs, as defenders of genome integrity in response to foreign or invasive nucleic acids such as viruses, transposons and transgenes (Carthew and Sontheimer, 2009). Single stranded forms of both miRNAs and siRNAs were found to associate with effector assemblies known as RNA-induced silencing complexes (RISCs) (Hammond et al., 2000). In both cases the identities of the genes to be silenced are specified by the small RNA component, as explained in greater detail in the next subsection.

MiRNAs and siRNAs appeared to differ in two major aspects: Firstly, miRNAs were viewed as endogenous, whereas siRNAs were thought to be essentially exogenous in origin, deriving from a viral, transposon or transgene trigger. Second, miRNAs appeared to be processed from stem-loop precursors with incomplete double-stranded character, whereas siRNAs were found to be excised from long, fully complementary double stranded RNAs (dsRNAs). However, it was soon clear that the similarities between miRNAs and siRNAs significantly exceeded the differences. Indeed these small RNAs are similar in size, sequence-specific inhibitory functions and, most importantly, in their dependence on the same two families of proteins: Dicer enzymes to excise them from their precursors, and Ago proteins, to support their silencing function.

1.1.2.2 Genomic localization of miRNA genes

MiRNAs have been identified in a wide range of organisms, ranging from simple multicellular ones, such as poriferans (sponges) and cnidarians (starlet sea anemone), to homo sapiens. Animal miRNAs appear to have evolved separately from those in plants because their sequences, precursor structure and biogenesis mechanisms are distinct from those in plants (Chapman and Carrington, 2007; Millar and Waterhouse, 2005).

Many mammalian miRNA genes have multiple isoforms, or more appropriately paralogs, that are probably the result of gene duplications. For an example, in the human genome there are 12 distinct loci for the *let-7* family miRNAs. These paralogs often have identical sequences in the nt positions 2–7 relative to the 5' end of the miRNA. Since these nucleotides, called the *seed* region, are essential for the regulatory function of the miRNAs, the paralogs are thought to act redundantly.

As noted above, a large proportion of the known miRNAs, approximately 50%, are found in close proximity (< 10 kb) to other miRNAs (Yu et al., 2006). These *clusters* are transcribed from a single polycistronic transcriptional unit (TU), although there are cases where individual miRNAs are derived from separate promoters (Lee et al., 2002; Monteys et al., 2010).

Approximately 50% of miRNA loci are located in the intronic region of non-coding transcripts, whereas 10% are found in the exonic region of non-coding transcriptional units (Griffiths-Jones, 2007; Saini et al., 2007). Examples of the four possible genomic localizations of miRNAs are shown in Fig. 1.1.

1.1.2.3 From primary to precursor miRNAs: the Drosha enzyme

The first step in the creation of an active, mature miRNA is the synthesis of the stem-loop *primary* transcript (pri-miRNA). The transcription of most pri-miRNAs is mediated by RNA polymerase II (Pol II) (Cai et al., 2004; Lee et al., 2004), although a small group of miRNAs that are associated with Alu repeats can be transcribed by Pol III (Borchert et al., 2006). These transcripts are typically several kilobases long. While still in the nucleus, they undergo a first cleavage at the stem of

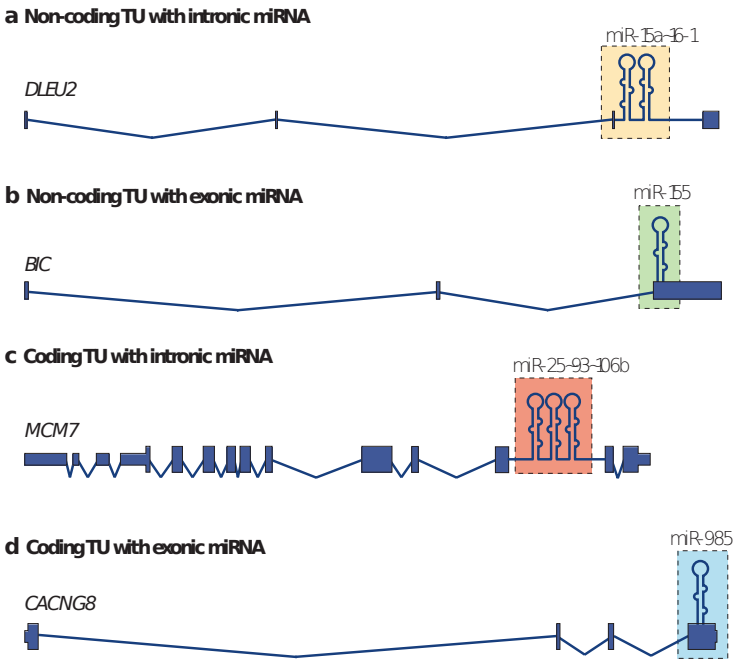


Figure 1.1: **Genomic localization of the miRNA loci - a**) Intronic mirRNA in a non-coding transcript, as is the case of *miR-15a 16-1* cluster in the non coding RNA gene DLEU2. **b**) Exonic miRNAs in non-coding transcripts, as is *miR-155* in the RNA gene BIC. **c**) Intronic miRNAs in protein-coding transcripts, as is the *miR-25 93 106b* cluster, located in the intron of the DNA replication licensing factor MCM7 transcript. **d**) exonic miRNA in protein-coding transcripts. The *miR-935* hairpin is localized in the last exon of CACNG8. Adapted from (Kim et al., 2009).

the hairpin structure by the RNase III-type protein Drosha. The result of this cleavage is a much shorter hairpin RNA, approximately 70 nt long called the *precursor* miRNA (pre-miRNA). Drosha alone is not able to achieve this cleavage, but needs to be supported by a cofactor: the DiGeorge syndrome critical region (DGCR8) protein in mammals Pasha in *D. melanogaster* and *C. elegans* (Denli et al., 2004; Gregory et al., 2004; Han et al., 2004; Landthaler et al., 2004). Mouse embryonic stem (ES) cells that are deficient in the *Dgcr8* gene fail to produce miRNAs and manifest defects in proliferation and differentiation (Wang et al., 2007). Typically a metazoan pri-miRNA consists of a stem of 33 base pairs (bp), a terminal loop and two flanking single stranded RNA (ssRNA) strands. DGCR8 assists Drosha to cleave the substrate 11 bp away from the ssRNA-dsRNA junction (Han et al., 2006; Zeng and Cullen, 2005). Interestingly, Drosha has been found to negatively regulate also its own cofactor, DGCR8, in that it cleaves the hairpins in the second exon of the DGCR8 mRNA (Han et al., 2009).

The domain compositions of Drosha, DGCR8 and Dicer (which will be described in the next subsection) are shown in Fig. 1.2. Drosha is a 130-160 kDa, nuclear protein, and together with DGCR8 forms a large ($\sim$ 650 kDa in humans) complex known as the *Microprocessor* complex. Drosha has two tandem RNase III domains (RIIIDs) and a double-stranded RNA-binding domain (dsRBD) (Fig. 1.2). The two RIIIDs interact with each other to make an intramolecular cut in the 3' strand of the dsRNA.

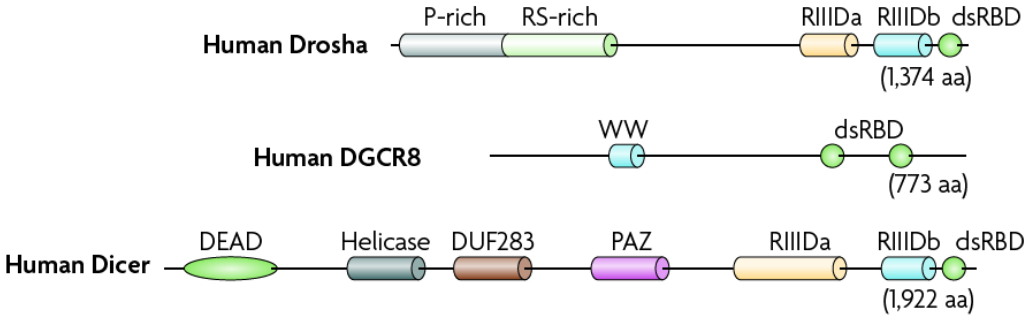


Figure 1.2: **Domain structure of Drosha** - Both Drosha and Dicer are RNase III proteins, characterized by the presence of two tandem RNase III domains (RIIIDs) and a double-stranded RNA-binding domain (dsRBD). Adapted from (Kim et al., 2009).

Recently, another class of miRNA-like RNAs embedded in short introns in flies and mammals has been discovered. Their biogenesis differs from miRNAs in that it is independent of Drosha. This special class of RNAs is sometimes referred to as *mirtrons* (Okamura et al., 2007; Ruby et al., 2007).

After being synthesized in the nucleus, the pre-miRNAs are exported to the cytoplasm (Kim, 2004). This process is mediated by exportin 5, a member of the nuclear transport receptor family. Exportin 5 recognizes the 14 bp double stranded RNA stem along with a short 3' overhang. Initially, this protein was known as a minor export factor for tRNAs, but it later emerged that miRNAs are its main cargoes (Bohnsack et al., 2004; Lund et al., 2004).

1.1.2.4 From pre-miRNAs to mature miRNAs: Dicer and the cytoplasmic processing

Pre-miRNAs are processed to mature miRNAs by the endoribonuclease called Dicer (Meister and Tuschl, 2004; Tomari and Zamore, 2005). Dicer enzymes belongs to a RNase III protein family characterized by several domains in a specific order going from the amino-to-carboxy terminus: a DEXD/H ATPase domain, a DUF283 domain, a PAZ domain, two tandem RNase III domains, and a dsRNA-binding domain (see Fig. 1.3). Some organisms such as mammals and nematodes, have only a single Dicer that controls the biogenesis of both miRNAs and siRNAs, whereas other organisms have more than one. For instance, *Drosophila* expresses two distinct Dicers and *Arabidopsis* produces four. In general, organisms with multiple Dicers exhibit a higher degree of specialization, as exemplified by *Drosophila*, where Dicer-1 is required for miRNA biogenesis and Dicer-2 is mostly devoted to the siRNA pathway (Tomari and Zamore, 2005). Biochemical, genetic and structural studies have converged on a model in which the PAZ and RNase III domains are essential for excising siRNAs preferentially from the ends of dsRNA molecules (Macrae et al., 2006; Zhang et al., 2004). PAZ domains are specialized in binding RNA ends, especially duplex ends with short ($\sim$ 2 nt) 3' overhangs. When this domain binds to an RNA end, the dsRNA substrate extends approximately two helical turns along

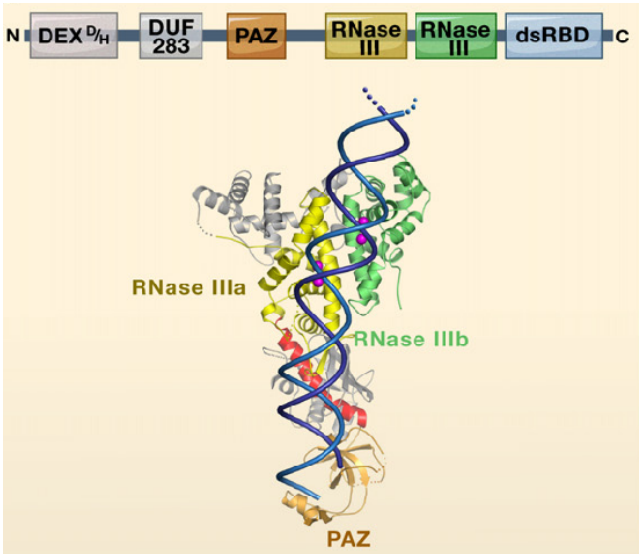


Figure 1.3: **Domain composition and structure of Dicer proteins** - The typical domain organization of Dicer proteins is shown at the top. Dicer enzymes cleave dsRNAs through the action of two RNase III domains. The dsRNA ends associate with the PAZ domain. The RNase III domains then cleave the $\sim 20-30$ nt miRNA/siRNA duplex from the precursor. Adapted from (Carthew and Sontheimer, 2009).

the surface of Dicer before reaching its processing centre. This centre is located in a cleft of an intramolecular dimer involving the RNase III domains. Each of the two RNase III active sites cleaves one of the two strands. This model pertains equally to pre-miRNA stem-loop substrates and to long, perfectly base-paired dsRNAs.

Human Dicer is also known to interact with two closely related proteins, TRBP (TAR RNA-binding protein, also known as TARBP2) and PACT (also known as PRKRA), although their roles in miRNA processing have not been completely elucidated yet (Chendrimada et al., 2005; Haase et al., 2005).

1.1.2.5 miRNA mediated silencing: the Argonaute protein family

Following Dicer cleavage, the resulting 22 nt RNA duplex is loaded onto a protein of the Argounaute superfamily, thereby generating the effector RNA induced silencing complex (RISC). The Argonaute superfamily can be divided into three separate subgroups: the Piwi clade that binds piRNAs, the Ago clade that associates with miRNAs and siRNAs, and a third clade that has only been described thus far in nematodes (Yigit et al., 2006). All gene-regulatory phenomena involving 20–30 nt

RNAs are thought to require one or more Argonaute proteins, and these proteins are the essential building blocks of various forms of RISC. The double-stranded products of Dicer enter into a RISC assembly pathway that involves duplex unwinding, culminating in the stable association of only one of the two strands with the Ago effector protein (Meister and Tuschl, 2004; Tomari and Zamore, 2005). This *guide* strand directs target recognition by Watson-Crick base pairing, whereas the other strand of the original duplex (sometimes called the *passenger* strand) is discarded. Argonaute proteins are characterized by the presence of four domains: the PAZ domain (present also in Dicer enzymes), the PIWI domain that is unique to the Argonaute superfamily, and the N and Mid domains. The arrangement of the domains and the structure of a typical Ago protein is shown in Fig. 1.4. Many aspects of Argonaute function have been elucidated by crystallographic studies (Parker et al., 2005; Song et al., 2004; Yuan et al., 2005). As shown in Fig. 1.4 the protein is bi-lobed, with

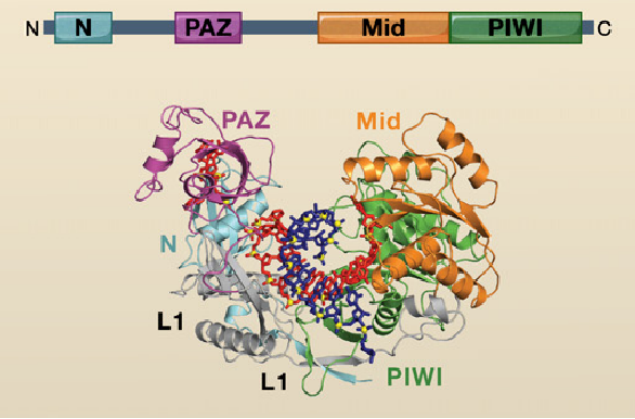


Figure 1.4: **Argonaute proteins** - The canonical arrangement of Ago domains is given at the top. Below the structure of the *Thermus thermophilus* Ago protein, with a bound DNA guide strand, is shown. Adapted from (Carthew and Sontheimer, 2009).

one lobe consisting of the PAZ domain and the other of the PIWI domain flanked by the N and Mid domains. The PAZ domain binds the 3' terminus of the guide strand. The other end of the guide strand engages a binding pocket in the Mid domain, and the remainder of the guide strand tracks along a positively charged surface to which each of the domains contributes. Guide strand nt 2–6, known as the *seed* region, are exposed and are available for base pairing, while the passenger strand is degraded.

Studies on siRNA duplexes indicate that the relative thermodynamic stability of the two ends of the duplex determines which strand is to be selected (Khvorova et al., 2003; Schwarz et al., 2003). Because strand selection is often not a stringent process, some hairpins produce miRNAs from both strands at comparable frequencies.

Although some species such as *Schizosaccharomyces pombe* express only a single Argonaute protein, most contain multiple Argonaute genes. Drosophila, for example, has five, humans eight and nematodes twenty-seven paralogues. In humans four of the eight proteins are from the Ago clade, and associate with both siRNAs and miRNAs (Meister and Tuschl, 2004; Tomari and Zamore, 2005), but little difference has been reported thus far in the populations of small RNAs that they bind, so the degree of functional specialization in mammals remains unclear (Carthew and Sontheimer, 2009).

1.1.3 How miRNAs regulate their targets

In the last few years much has been understood in terms of how miRNAs act on their targets; however, there is still a remarkable disagreement on some pivotal aspects. Two main questions arise when considering the way a miRNA influences the expression of its targets: 1) how does a miRNA select its targets? 2) how does it alter their expression?

1.1.3.1 The identification of miRNA targets

Concerning the first question, a significant difference is observed between plants and animals. In the former, targets can be efficiently detected simply by searching for extensive complementarity between the miRNA and 3' UTR sequences, as shown in (Rhoades et al., 2002). In plants, regulation of transcripts mainly occurs by means of slicing rather than translational repression. According to (Brodersen and Voinnet, 2009) however, the possibility that larger numbers of mRNAs are targeted for translational inhibition by means of imperfect matching has never been tested, and cannot be excluded *a priori*. In animals the situation is far more complicated. Extensive complementarity between miRNA and 3' UTR sequences exists, but is

very rare, and the largest part of validated targets interact with their corresponding miRNA by imperfect base-pairing. The initial effort to reliably identify miRNA targets in animals heavily relied on computational tools, and a number of different methods were developed. The list of targets produced by these methods, however, showed little overlap, suggesting a huge sensitivity of the results to the methodological differences (Bartel, 2009). One of the main problems is that, for a given mature miRNA, there is a large number of 3' UTR fragments that would achieve an identical score in terms of sequence alignment to the miRNA (Lewis et al., 2003).

It is nowadays widely accepted that one of the turning points in miRNA target identification was the inclusion of preferential evolutionary conservation to distinguish a true target from the multitude of equally matching 3' UTR fragments. Today there is increasing agreement about three major aspects of the miRNA-target interaction, all of which will be discussed in greater detail in the next sections:

1. The importance of the *seed* region, i.e. the region in the miRNA 5' centered around nucleotides 2-7

2. The importance of the conservation of the seed region across different species.

3. The observation that highly conserved miRNAs have many conserved targets.

The relevance of the so-called seed region can be summarized in the "seed rule" (Brodersen and Voinnet, 2009), that states that Watson-Crick base-pairing to the 5' miRNA nucleotides 2-7 is necessary for the regulatory effect of a miRNA on its targets (Lewis et al., 2005). This rule is based on the evidence that the seed appears to be the only contiguous region of miRNAs that, according bioinformatics analyses, is evolutionarily conserved to an extent significantly greater than expected by chance (Lewis et al., 2003). Other evidence of the fundamental role played by the seed region is provided by the fact that mutating nucleotides within the seed-pairing region of a validated target, abolishes or significantly impairs the miRNA-induced regulation (Brennecke et al., 2005; Doench and Sharp, 2004). The seed region also turned out to be the most over-represented motif among targets that act as responders at a

post-translational level as showed by Stable Isotope Labeling with Amino acids in Cell culture (SILAC) (Baek et al., 2008; Selbach et al., 2008).

Despite the wide agreement on the central role of the seed region, some researchers propose that there might be a large set of "non-seed" targets with biological relevance, although such sites would be more difficult to identify by computational approaches (Brodersen and Voinnet, 2009). Some non-seed targets have been identified in *C. elegans* (remarkably back in 1996, when neither the term miRNA or seed were in use, (Ha et al., 1996)), in *Drosophila* (Easow et al., 2007) and in human cells (Ørom et al., 2008; Stern-Ginossar et al., 2007). Despite these discoveries, the frequency of this non-seed regulation mechanism is still unknown and, currently, the seed-based mechanism is widely regarded as the major one.

Some target prediction methods, in addition to the evolutionarily conserved complementarity of the target's 3'UTR and the miRNA seed, take other factors into account. One of these is the type of nucleotide opposite to the first miRNA nucleotide. Some of these (e. g. the TargetScan prediction algorithm (Friedman et al., 2009; Grimson et al., 2007; Lewis et al., 2005), rewards an A across from position 1 (Bartel, 2009), whereas other algorithms do not treat such a case differently from others. This difference can find a justification in the increasing evidence that a non-Watson-Crick pairing at nucleotide 1 is somewhat favoured as confirmed both by site-conservation analyses, and by array and proteomics data (Baek et al., 2008; Nielsen et al., 2007).

Another possibility, although less frequent, is the presence of an additional perfect match in the eighth position. The summary of the configurations that are today regarded as *canonical* is shown in Fig. 1.5.

1.1.3.2 Supplementary and compensatory sites

It is a commonly thought that additional pairing to the remainder of the miRNA supplements the seed pairing. To date, however, there is no experimental demonstration of the validity of this assumption (Doench and Sharp, 2004), and an extensive analysis of mammalian 3' UTRs showed that the majority of sites have no more

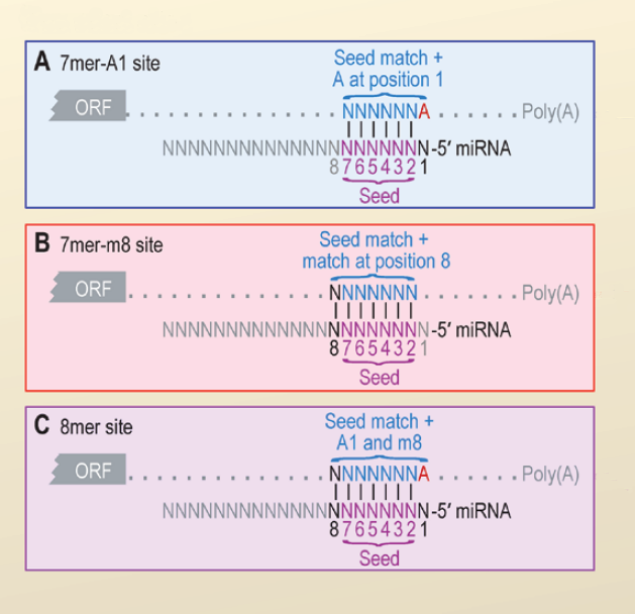


Figure 1.5: **The three *canonical* 7-8 nt seed-matched sites.** - **A**. The 7mer, optionally with an additional match when there is an A in the first position (7mer-A1). **B**. The 7mer with an additional match at position 8 (7mer-m8). **C**. The 8mer site, characterized by the presence of an eighth matching nt. Adapted from (Bartel, 2009).

3'-supplementary pairing than expected by chance (Bartel, 2009; Brennecke et al., 2005). One complication in the identification of functional and effective supplementary pairing regions is their very large number. However, an array-based screening has identified a type of 3'-supplementary site that is associated with a sufficient number of sites to be screened by means of microarray data sets (Grimson et al., 2007). This site is centered on the miRNA nt 13–16 and is predicted to be highly sensitive to bulges, mismatches or wobbles, preferring at least 3-4 contiguous uninterrupted pairs (Fig. 1.6, top). These sites are predicted with greater specificity, but their efficacy in terms of target downregulation seems to be only slightly superior to that of the canonical sites.

Sometimes the presence of additional pairing to the miRNA 3' UTR can represent a form of compensation for a defective pairing to the miRNA seed. The presence of one single mismatch in the seed can require a substantial perfect match (at least 9 nt) starting from a region comprised between nt 13 and 17 (Fig. 1.6, bottom). These compensatory sites are, however, far less conserved through evolution, most

probably due to their greater structural complexity. According to Bartel (Bartel, 2009) they represent about 1% of the preferentially conserved sites in mammals.

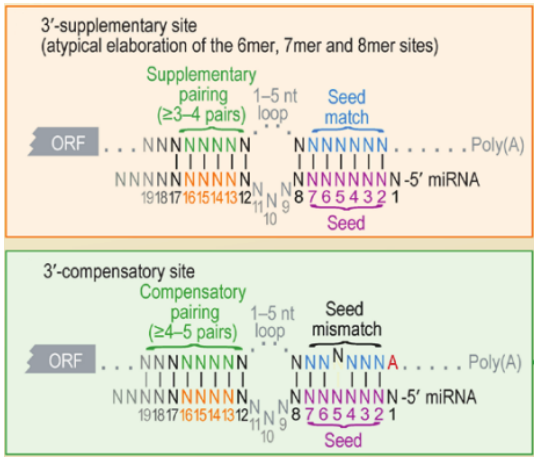


Figure 1.6: **Compensatory and supplementary sites - Top**: 3'-supplementary sites usually have Watson-Crick pairing centering around miRNA nt 13-16 to supplement a 6-8 nt site. **Bottom**: 3'-compensatory sites also have Watson-Crick pairing around nt 13-16, but are required to compensate for a seed mismatch, thereby creating a functional site. Adapted from (Bartel, 2009).

1.1.3.3 Bartel's mechanistic model

Bartel and co-workers have developed a mechanistic model in which the contribution of each miRNA region is contextualized according to the available experimental evidence (Bartel, 2004). According to this model (Fig. 1.7), the RISC complex presents the 5' nt 2-8 to the target mRNA to favour pairing (Fig. 1.7 **A**). These nucleotides are pre-organized in an A-form helix to improve affinity for the target. Thermodynamical and topological considerations show that both longer or shorter stretches of miRNA nucleotides would have a lower affinity and specificity (Bartel, 2009). According to the model, the miRNA undergoes an important conformational accommodation to facilitate Watson-Crick pairing. Moreover the miRNA is predicted to be bound along its entire length to the Argonaute protein to prevent its backbone being accessed by cellular RNase (Fig. 1.7 **A**). Once nucleation occurs at the seed, the Argonaute protein progressively loosens its grip on the central and 3' regions of the miRNA (Fig. 1.7 **C**). These miRNA regions are then free to wrap around the mRNA, thus forming a miRNA-mRNA duplex consisting of two helical

turns. After this rearrangement, the Argonaute protein rebinds to the central and 3' regions of the miRNA thus strengthening the linkage to the miRNA-mRNA duplex (Fig. 1.7 **D**), and allowing its cleavage.

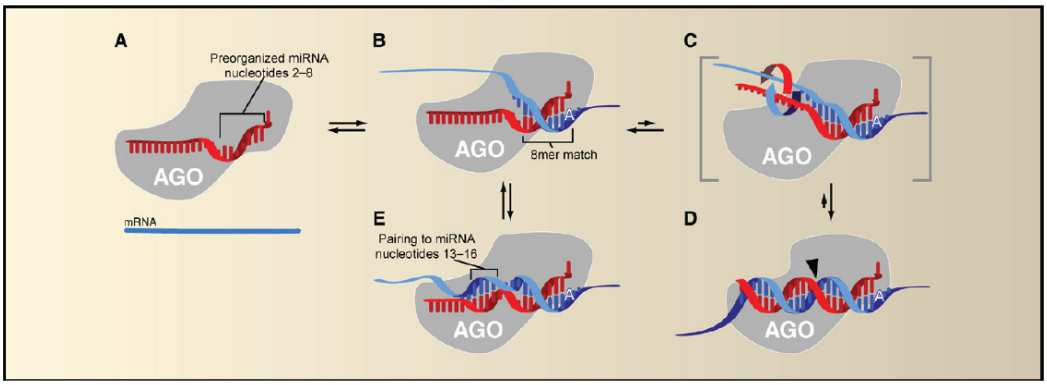


Figure 1.7: **A model for miRNAs regulatory mechanism** - **A** The Ago protein carries the miRNA guide strand (in red). The target mRNA is shown in blue. **B** The 8mer matching occurs in the binding pocket. **C - D** Following the target recognition, a large conformational change occurs. The active site is now in position to be cleaved. **E** In case of 3' supplementary pairing nucleotides 13-16 are also matched. This prevents the miRNA-mRNA duplex from further wrapping around each other. Adapted from (Bartel, 2009).

1.1.3.4 Many targets have a poorly conserved complementary seed sequence

The experimental evidence about the importance of the seed sequence and its conservation does not exclude the existence of fully effective, but poorly conserved target sites. From a purely computational point of view, poorly conserved target sites exceed well conserved sites by about ten to one (Bartel, 2009). The obvious question is what fraction of this large number of mRNAs is composed by true miRNA targets. Heterologous reporter assays show that a large number of non-conserved sites in RNA can be functional (Farh et al., 2005). Analyzing simultaneously miRNA and mRNA expression profiles, it appears that these non-conserved targets are primarily found in tissues where the cognate miRNA is absent (Farh et al., 2005). This phenomenon of mutual avoidance, where mRNAs "try" not to co-localize with their cognate miRNAs, is sometimes referred to as *selective avoidance*. One possible explanation of this phenomenon is that, over the course of evolution, sites for miRNAs

that are absent in the cells where the mRNA is expressed, can accumulate without consequence, whereas sites for miRNAs that are highly expressed in the same cell where the mRNA functions, would impart a selective disadvantage, and thus fail to be fixed in the population (Farh et al., 2005; Stark et al., 2005). However, because so many mRNAs have non-conserved 7 nt sites for each miRNA, the minority of mRNAs that are co-expressed with the miRNA will still constitute a large number, possibly exceeding the conserved sites (Farh et al., 2005; Krützfeldt et al., 2005).

1.1.3.5 The UTR context is important for miRNA mediated regulation

The conservation or the complementarity of miRNA target sites in the 3' UTR of mRNAs are not sufficient to completely explain of the mechanisms underpinning miRNA mediated repression. A clear demonstration of this is provided by the fact that identical miRNA target sites in two different mRNAs can lead to repression in one case and have no effect in the other (Brennecke et al., 2005; Farh et al., 2005). The position of such target sites inside the UTR of the target, as well as the composition of the site's flanking regions, can have a huge impact on the efficacy of the miRNA machinery to regulate the mRNA. According to Bartel (Bartel, 2009) the following features enhance the regulatory effect of a miRNA on its target: 1) the target site is located at least 15 nt from the stop codon; 2) the target site is not close to the the center of long UTRs; 3) the region surrounding the target site is rich in AU nucleotides; 4) there are multiple target sites recognized by the same miRNA or by co-expressed miRNAs acting cooperatively to repress same target gene (Grimson et al., 2007).

Both computational and experimental evidence show that targeting can occur not only in the 3' UTR of the mRNA, but also in the 5' UTR and in the open reading frames (ORFs) (Baek et al., 2008; Farh et al., 2005; Grimson et al., 2007). Targeting in ORFs is less frequent than in 3' UTRs, but is still, however, much more frequent than in 5' UTR. One possible explanation for this phenomenon could be the interference of the translational machinery with the RISC complex (Bartel, 2004).

Many mRNAs have multiple target sites in their 3' UTRs that can be recognized by the same or by different miRNAs. This observation raises the question of how multiple miRNAs act on the same targets. It has been proposed that miRNAs can act either independently or in a coordinated way; however evidence appears to favour the former hypothesis, as the response of mRNAs having multiple target sites is almost identical to the response we would expect if each site contributed independently. The net effect of this independent repression would be multiplicative and not additive (Grimson et al., 2007; Nielsen et al., 2007). Even so, however, there are exceptions to this mode of action, as shown by Grimson (Grimson et al., 2007): two target sites located at 8-40 nt from each other, led to an enhancement in repression that is far superior to the repression expected if the two acted independently.

Figure 1.8 summarizes, in a qualitative way, the mean efficacy of the miRNA target sites described so far in terms of destabilization of target mRNA after overexpression of a miRNA. The table refers to the effect at the mRNA level although, as we shall describe later, the effect at the translational level only might be equally relevant.

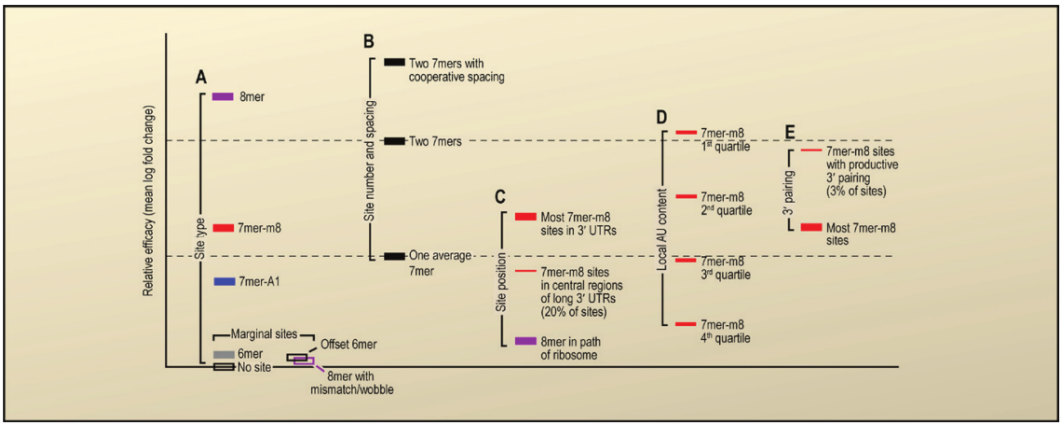


Figure 1.8: **Relative efficacy of miRNA target sites.** - The efficacy is defined as the mean downregulation (on a logarithmic scale) of target mRNAs possessing the indicated sites. The various typologies of sites have been classified in the categories labeled **A**-**E**. Adapted from (Bartel, 2009).

1.1.3.6 Does miRNA-mediated regulation occur at the transcriptional or translational level?

Initially, miRNAs were thought to regulate their targets by repressing protein output with marginal or no effect on the mRNA. This belief was initially motivated by the observation that miRNA *lin-4* downregulates *lin-14* without appreciably affecting the abundance of its mRNA (Lee et al., 1993; Olsen and Ambros, 1999; Wightman et al., 1993). Two questions arose: 1) to what extent, if any , is the targets mRNA abundance altered? 2) does translational repression occur during the initiation of translation, or at post-initiation steps?

In the last three years, a number of studies have cast some light on the first issue, although already in 2005 Lim and collaborators (Lim et al., 2005) showed, using microarrays, that the effect of miRNA overexpression was largely detectable at the mRNA level. They transfected two mature miRNAs, *miR-124* and *miR-1* into HeLa cells and observed that, in each case, about 100 mRNAs were downregulated 12 hours after transfection. The presence of the corresponding seed motifs was significantly enriched in the 3' UTR of these mRNAs. In addition, Selbach and collaborators in Rajewsky's lab (Selbach et al., 2008) exploited a modified version of the SILAC technology to quantify the change in protein levels upon *miR-223* deletion in mouse neutrophils. They also measured the levels of the corresponding transcripts and found a strong positive correlation between the two. Despite evidence for some translational-only repression, all proteins up-regulated by more than 50% derived from transcripts that were detectably increased upon *miR-223* deletion. Similarly, Baek and collaborators in Bartel's lab, observed a strong correlation between protein and mRNA levels upon *miR-1* overexpression (Baek et al., 2008).

Hendrickson and colleagues (Hendrickson et al., 2009) used a different technique: they focused on *miR-124* and identified its targets by looking at the mRNAs that were recruited to Argonaute proteins in response to ectopic *miR-124* expression in the human embryonic kidney HEK293T cell line. In order to quantify the effects of the miRNA at the protein level, the authors considered two parameters: the *ribosome occupancy* and the *ribosome density*. The former is defined as the fraction of

mRNA species bound by at least one ribosome and, presumably, undergoing translation. The latter refers only to those mRNAs that have at least one bound ribosome and is the average number of ribosomes bound every 100 bases of coding sequence. To reliably identify the real targets of *miR-124* the authors lysed the cells transfected with the miRNA and isolated the Argounaute-associated mRNAs by means of immunoprecipitation. This procedure was performed also on mock-transfected cells and the mRNAs specifically recruited to the Ago proteins by *miR-124* were identified by Significance Analysis of Microarrays (SAM) (Tusher et al., 2001). This study concluded that *miR-124* reduces both translation and transcription of its targets over a wide range of values and, most importantly, that approximately 75% of the variation at the protein level can be explained by changes in the abundance of the corresponding mRNAs. Also in this case, however, the change in protein levels was smaller than that at the mRNA level, with an average decrease of protein abundance around 12%, compared with a corresponding decrease in mRNA abundance of 34%. Therefore, despite the initial reports describing miRNA-regulated repression as a purely translational effect, there is accumulating evidence that the effect of miRNAs on mRNA is relevant and detectable.

1.1.3.7 The regulation of most intragenic miRNAs mirrors that of their host genes

The fact that a large fraction (almost half) of mammalian miRNAs are contained within the boundaries of transcriptional units suggests that the regulation of the pri-miRNA should replicate that of the host gene. The reality, however, is far more complex. It has been long held that, while intergenic miRNAs have their own promoter, intronic ones were under the transcriptional control of the host gene's promoter. This picture has changed with the recent publication of a number of works showing that a significant fraction of intronic miRNAs have a host gene-independent promoter (Corcoran et al., 2009; Ozsolak et al., 2008; Wang et al., 2009). Monteys and collaborators (Monteys et al., 2010) performed an extensive genomic analysis of intronic miRNAs, and they found that approximately 35% of them have upstream

regulatory elements consistent with promoter function.

Despite this evidence, the majority of intronic miRNAs have been found to follow the pattern of regulation of their host genes, as shown by Baskerville and Bartel (Baskerville and Bartel, 2005). To analyze the global expression of miRNAs in human tissues they considered a panel of 175 human miRNAs across 24 different organs. They observed that intronic miRNAs are usually coordinately expressed with their host gene mRNA, implying that they generally derive from a common transcript. In two cases they noticed a negative correlation between the expression of the mature miRNA and that of the host gene. This was, for example, the case of *miR-26a* and the *CTDSLP* gene. Since this miRNA does not show any evidence of a host gene-independent promoter, there must be another explanation for this finding. In fact, there are two distinct pri-miRNAs, *hsa-mir-26a-1* and *hsa-mir-26a-2*, that can produce the same mature miRNA, *miR-26a*. These pri-miRNAs are localized in the *CTDSPL* (3p21.3) and *CTDSPL2* (12q13-q15) genes respectively. *CTDSPL2* is usually more highly expressed than *CTDSPL* and the expression of the two genes was found to be negatively correlated. Thus *miR-26a* can be highly expressed even when the *CTDSLP* gene is downregulated, as it is derived from the alternative pri-miRNA, *hsa-mir-26a-2*. The authors also found that proximal pairs of miRNAs are usually co-expressed, and noticed that an abrupt drop in the correlation between pairs of expressed miRNAs occurs at a mutual distance of 50kb.

When there is no sign of host-gene independent promoters, there are essentially two scenarios that can lead to contradictory results when comparing the expression patterns of a miRNA and its host gene. The first case is when the same mature miRNA is produced by two (or more, although we will assume just two to simplify the example) primary transcripts, both of which are intragenic, but located in different genes, as it is the case of the aforementioned *miR-206a*. The second case is when the same mature miRNA is produced by two primary transcripts, one of which is intragenic, while the other is intergenic. In this latter case, one should expect to observe no correlation between the expression of the mature miRNA and the host gene if only the intragenic miRNA was expressed. An example of this class is

provided by *hsa-miR-24*, which is produced by the *hsa-mir-24-1* and *hsa-mir-24-2* primary transcripts. The first is localized in the *C9orf3* gene, on chromosome 9, while the second is an intragenic miRNA localized on chromosome 19. Finally, as described previously, maturation of a miRNA is accomplished in several steps and is mediated by different nuclear and cytosolic enzymes. Thus, the expression of a mature miRNA does not necessarily correlate with the expression of its pri-miRNA and host gene.

1.2 MiRNAs and cancer

The first direct link between cancer and miRNAs was observed by Carlo Croce and collaborators while attempting to clone a tumour suppressor gene, located in the chromosomal region 13q14, which is frequently lost in chronic lymphocytic leukemia (CLL) (Adrian et al., 2002). They observed that none of the protein coding genes were altered, but detected a translocation and, in one case, a very small deletion, both localized exactly in this region. Within the boundaries of the deleted region, they found two miRNA genes, *hsa-mir-15a* and *hsa-mir-16-1*. The translocation breakpoints cut precisely the precursor of these two miRNAs. The importance of these two miRNAs in connection with CLL was reinforced by the finding that they are lost in approximately 70% of the patients, suggesting that such a loss might be an early event.

1.2.1 MiRNAs can act as tumour suppressors or as oncogenes

The loss of *miR-15a* and *miR-16-1* in CLL led to the natural hypothesis that they might act as tumour suppressors. Since this discovery, many other miRNAs have been shown to be lost in a wide range of tumours. For example, the members of the *let-7* family, which consists of 12 miRNAs with very similar, and in some cases identical, mature sequences are located in chromosomal regions that are deleted in multiple tumours, including breast, lung, ovarian and cervical cancer (Croce, 2009). A direct causal role of the members of the *let-7* family in cancer, however, was only

established in 2005, when Johnson and collaborators demonstrated that a loss of *let-7* family members determines a constitutive *RAS* overexpression, an oncogene known to contribute to the pathogenesis of several types of tumours Johnson et al. (2005). Still in 2005, Cimmino *et al.* found that *miR-15a* and *miR-16b* target another well known oncogene, the B cell leukemia/lymphoma 2 (BCL2) gene, which has an anti-apoptotic function and its overexpression is a major cause of follicular lymphoma (Cimmino et al., 2005). Since the vast majority of miRNAs have an inhibitory effect on their targets, an oncogene-targeting miRNA acts in turn as a tumour suppressor. Importantly miRNAs, like tumour suppressor genes, can be silenced by epigenetic mechanisms such as methylation. This has been observed for *miR-15a*, *miR-16-1*, *miR-29* and in members of the *let-7* family (Calin and Croce, 2006).

Deletion or downregulation, however, are not the only mechanisms through which miRNAs exert their oncogenic potential. In fact, miRNAs have also been found to be overexpressed in human tumours, thus acting as oncogenes. However, defining a particular miRNA as either a tumour suppressor or as an oncogene, is often incorrect, since their function strongly depends on the cellular context. For example, *miR-221* and *miR-222* behave as tumour suppressors in erythroblastic leukemia targeting the oncogene *KIT*. In contrast, in a number of solid tumours they behave as oncogenes, targeting no less than four tumour suppressors (Felli et al., 2005; Pineau et al., 2010). The first miRNAs that were found to behave as oncogenes were *miR-155* and the cluster *miR-17-92*. The latter is a group of 6 miRNAs belonging to four families based on their seed sequences (see Fig. 1.9), located in a region of 800 bp in the non-protein coding gene *C13orf25* at 13q31.3 (He et al., 2005). Both the sequences and the organization of the members of this cluster are highly conserved in all vertebrates. Despite being under the transcriptional control of the same promoter, not all the members of the cluster have necessarily the same expression profiles. The mRNA splicing regulator hnRNP, for example, exclusively stimulates maturation of *miR-18a*, hence acting as a chaperone for recognition and cropping of this specific miRNA, without affecting the other members of the cluster (Guil and Caceres, 2007). Two cluster paralogs of miR-17-92 exist in mammals, presumably due to

gene duplication events (Mendell, 2008). One is found on chromosome 7, within the third intron of the *MCM7* protein coding gene, and contains 3 miRNAs, *miR-106b*, *miR-93* and *miR-25*. The other cluster is found in chromosome X and contains 6 miRNAs (Fig. 1.9). According to Mendell *et al.* this cluster is undetectable or expressed at trace levels in all the experimental settings that have been explored so far (Mendell, 2008). The members of the *miR-17-92* cluster represent probably

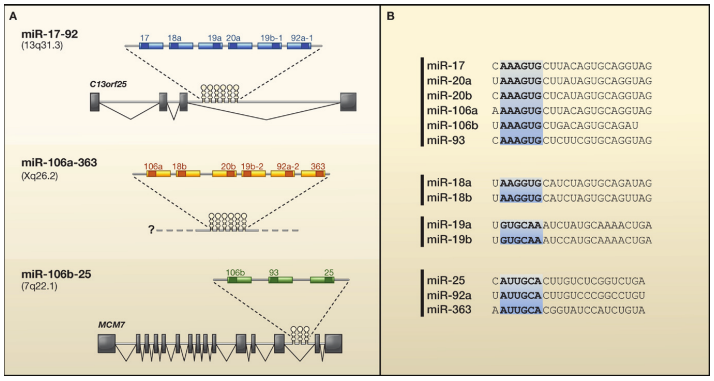


Figure 1.9: **Genomic organization and sequences of the *miR-17-92* cluster and its paralogs** - **A**: Genomic organization of the human *miR-17-92* cluster and its paralogs. **B**: MiRNAs belonging to the *miR-17-92* cluster are organized in four families on the basis of their seed sequence and, consequently, of their putative targets. Adapted from (Mendell, 2008).

the best studied example of oncogenic miRNAs. Several studies have experimentally demonstrated that the proapoptotic gene *BCL2L11/BIM* is a direct target of *miR-17-92* (Koralov et al., 2008; Petrocca et al., 2008b)). Moreover the *c-Myc* oncogene, which is a known, potent inducer of tumour angiogenesis, was shown not only to induce expression of the *miR-17-92* cluster, but also to exert its angiogenic activity, at least in part, by the downstream activation of the cluster itself (Dews et al., 2006). Finally, *miR-17-92* targets directly *E2F1*, *E2F2* and *E2F3*, members of the *E2F* family of transcription factors, which are critical regulators of the cell cycle and apoptosis (Mendell, 2008; Ventura et al., 2008). Interestingly, both E2F1 and E2F3 can directly activate transcription of these miRNAs, giving rise to a negative

feedback loop (O'Donnell et al., 2005; Petrocca et al., 2008b). In addition to the members of the *miR-17-92* cluster, several other miRNAs have been found to target important tumour suppressors, such as *miR-21* which targets *PTEN*, a tumour suppressor playing a fundamental role in preventing cells from growing and dividing too rapidly (Garofalo et al., 2009).

1.2.2 Breast cancer

According to the United States Cancer Statistics (data refer to the 1999-2006 period), breast cancer is the second cause of death among women in the western world after lung cancer (Harris et al., 2007). Despite several advances in the field, the molecular pathogenesis of this disease is still not completely understood. Human breast tumours have been classified into 18 categories, based on the histological features of the primary tumour at the time of diagnosis (Fattaneh Tavassoli, 2003). Lesion size, nuclear grade, mitotic index, necrosis, and cellular organization are some of the features that are commonly used to classify human breast tumours.

Many breast cancer molecular markers have been proposed, but the most reliable ones, according to the guidelines of the American Society of Clinical Oncology, are the estrogen receptor (ER), the progesterone receptor (PgR) and the ErbB2 status (Harris et al., 2007). These clinical parameters are considered the most valuable in identifying patients that would benefit from endocrine forms of therapy. ER negative patients, for instance, do not benefit from endocrine interventions, and have usually a poor prognosis (Harris et al., 2007). There are currently two hypotheses trying to explain the association of the ER with breast cancer. The first assumes that binding of estrogens to the ER stimulates proliferation of mammary cells. This increase in cell division and DNA synthesis elevates the risk for replication errors and, as a consequence, the acquisition of detrimental mutations. The second hypothesis states that estrogen metabolism leads to the production of genotoxic by-products that could directly damage DNA (Deroo and Korach, 2006).

ErbB2 is one of the four members of the *ErbB* family of receptor tyrosine kinases, which also includes the epidermal growth factor receptor (EGFR), *ErbB3* and

ErbB4. *ErbB2* and EGFR are implicated in the development of a number of human cancers, and several alterations in their genes have been detected in tumours, including gene amplification, mutations leading to the activation of the kinase domain, and in-frame deletions (Hynes and MacDonald, 2009). Breast tumours with *ERbB2* gene amplification and receptor overexpression are often labelled as ErbB2+.

The standard method to assess ER, PgR and ErbB2 status is immunohistochemistry (IHC). Despite being routinely used, this method has been shown to have some drawbacks, including a modest positive predictive value (PPV)[1] ranging from 30% to 60% (Bonneterre et al., 2000; Mouridsen et al., 2001), and a large false negative rate (30% - 60%). The interpretation of staining patterns can be a further cause of variability, being subjective to the individual pathologist or the threshold setting of the image analysis system in use (Rhodes, 2003).

Recently, with the advent of high-throughput, gene-expression profiling methods, a variety of novel biomarkers have been proposed and, more importantly, several genetic signatures have been described that can accurately identify different breast cancer molecular subtypes with distinct pathological behaviors (Sotiriou and Pusztai, 2009). These findings have not only offered alternative methods to refine the diagnosis of breast cancer, but have also opened new and more ambitious possibilities for an improved therapeutic intervention. The seminal work in this direction was Charles Perou's study (Perou et al., 2000), where the analysis of gene expression patterns suggested that at least four major molecular subtypes (or "molecular portraits", to use the definition provided by the authors) of breast cancer exist: luminal-like, basal-like, normal-like, and HER-2 positive. Luminal-like tumours are mostly ER positive, while basal-like tumours are essentially triple negative, *i.e.*, negative for ER, PgR, and ErbB2. Luminal tumours derive their name from the fact that they express high amounts of luminal cytokeratins and genetic markers of luminal epithelial cells of normal breast tissue (Rakha et al., 2007). Instead, some, but not all basal-like breast cancers, display high expression of "basal" cytokeratins such as CK5, and a variety of growth factors, including the epidermal and insulin

[1]The positive predictive value of a test is the proportion of patients with positive test results who are correctly diagnosed.

growth factors (Sørlie et al., 2003; Sotiriou et al., 2003). A second study suggested that further molecular subgroups exist, splitting the luminal-like tumours in luminal-A, luminal-B and luminal-C (Sørlie et al., 2001). In a third study the basal-like, normal-like, HER-2 positive and two luminal-like categories were observed (Sørlie et al., 2003).

Despite this apparent reproducibility, these categories are not immune to criticism. Different computational methods have been applied with the purpose of validating the number of subtypes and establishing their robustness. McShane, for example, proposed a method to assess subgroup stability that, when applied to the original data, showed a very low reproducibility of the luminal subgroups (McShane et al., 2002). Similarly Kapp, using a clustering-based validation method (see (Kapp and Tibshirani, 2007) for details), was able to confirm the existence of the basal-like, luminal-B and ErbB2 positive groups, while the normal-like and the luminal-A groups could not be reliably reproduced (Kapp et al., 2006).

There are several possible explanations for these controversial results. The lack of a standardized procedure for subtype definition is, however, one of the most probable causes. It is today widely accepted that a still unknown number of molecular breast cancer subclasses exist, differing in prognosis and in response to chemotherapy. It is apparent that the large overlap between the basal and luminal classes on the one hand, and the ER negative and positive on the other, is a major confounding factor. Other, more technical reasons, concerning the methodology used to identify the subclasses, also exist. A thorough discussion can be found in (Pusztai et al., 2006).

1.2.3 MiRNAs and breast cancer

Several miRNAs have been found to be deregulated in breast cancer, displaying the complex scenario of regulation patterns that we have previously illustrated (see Section 1.2.2. The locus containing the *miR-17-92* cluster, for example, is known to undergo loss of heterozygosity in breast cancer as well as in other types of cancer (Mendell, 2008). In addition, it has been found to downregulate the AIB1 (amplified

in breast cancer 1) protein, which acts as an oncogene and is commonly amplified in breast, ovarian and pancreatic cancer (Mertens-Talcott et al., 2007). Similarly, the 11q23-24 region that contains *miR-125b*, is frequently deleted in breast cancer, as well as ovarian and lung (Negrini et al., 1995). *Mir-21*, an oncogenic miRNA, is associated with increased invasion and metastatic potential in breast tumours (Yan et al., 2008), while *mir-335* is markedly downregulated in lung metastatic breast cancer cells.

Intriguingly some miRNAs have been found to be strongly associated to the molecular subtype of breast cancer. In bead-based flow cytometric miRNA expression profiling of 93 primary tumours, a number of miRNAs were found to be differentially expressed between the luminal A, luminal B, basal-like, ErbB2 positive, and normal-like subtypes (Blenkiron et al., 2007). Moreover, numerous publications show a strict correlation between miRNA levels and ER, PgR or ErbB2 status. Lowery *et al.* for example, identified a miRNA signature that accurately predicts ER, PgR and ErbB2 status. They demonstrated that *miR-342*, *miR-299*, *miR-217*, *miR-190*, *miR-135b* and *miR-218* are most strongly associated to ER status, while *miR-520g*, *miR-377*, *miR527-518a* and *miR-520f-520c* are predictive of PgR status. Finally *miR-520d*, *miR-181c*, *miR-302c*, *miR-376b*, and *miR-30e* constitute the ErbB2 signature (Lowery et al., 2009).

MiRNAs have also been implicated in breast cancer progression. A large body of evidence indicates that estrogen receptor-α-negative (ERα^-) breast tumours, which are more aggressive and less responsive to hormonal therapy, originate from ER-α-positive (ERα^+) tumours through different molecular pathways. For example, estrogen withdrawal (Santen et al., 2002), hypoxia (Stoner et al., 2002), overexpression of ErbB2 which results in hyperactivation of mitogen-activated protein kinase (MAPK) (Oh et al., 2001), and DNA methylation at the ERα promoter (Yan et al., 2001) have all been proposed to account for this transition. However, a very recent study hsa shown that the overexpression of a miRNA cluster containing *miR-221* and *miR-222* suppresses the ERα protein in ERα^+ cells, thus turning them into ERα^- (Di Leva et al., 2010). In the same study ERα was found to negatively reg-

ulate the expression of *miR-221* and *miR-222* by binding to their promoter, thus giving rise to a negative feedback loop .

1.2.4 Rationale of the project

Taken all together, these data highlight the importance of miRNAs in regulating gene expression in normal and pathological conditions. However, a comprehensive picture of the expression profiles of miRNAs in cancer and their relative correlation with different molecular and pathological tumour characteristics is still lacking. Therefore, we decided to perform a high-throughput transcriptional study of intronic miRNAs by analyzing the expression profiles of their host genes in publicly available Affymetrix breast cancer data sets.

As previously described, more than half of the human miRNAs are contained within the boundaries of transcriptional units. Although computational methods predict that roughly one third of these intragenic miRNAs might have an independent promoter, this is most probably an overestimation, since only a very small fraction of these predictions has so far been validated (Monteys et al., 2010). As a consequence, we should expect no less than two thirds of the intragenic miRNAs to be under the transcriptional control of their host gene promoter, hence it should be possible to predict miRNA expression by analyzing to the microarray expression profile of its host gene.

Microarrays represent the first and most diffused tool to simultaneously measure the expression of thousands of transcripts. Since their appearance, almost fifteen years ago, they microarrays been used to investigate the regulation of genes in hundreds of species. In Humans they have been exploited to explore the genomic aspects of many diseases and of almost every type of cancer. A number of publicly accessible microarray data repositories exist, the largest being the Gene Expression Omnibus (GEO) (Barrett et al., 2005). As of July 2010, GEO contains 450,535 entries, mostly but not only, from microarray experiments.

Breast cancer was one of the first types of cancers to be analyzed by means of microarrays. The pioneering work of van't Veer and collaborators (van 't Veer et al.,

2002) opened the search for groups of genes whose expression is coherently regulated between two or more conditions. These "gene signatures", as they are often called, have two main fields of application: the first aims to identify subgroups of tumours, as is the case of the molecular subtypes previously described; the second tries to identify novel markers for predicting of clinical outcome, response to therapy etc. Hundreds of such predictive signatures have been published, and in some cases they have represented the basis for the development of diagnostic tests that are today in routine clinical use. One such case is the MammaPrint molecular diagnostic test, which is used to predict the risk of metastasis in breast cancer patients with primary tumours less than 5 cm in diameter, no sign of lymph node metastasis, and less than 61 years of age (Buyse et al., 2006; Mook et al., 2009; van 't Veer et al., 2002).

In conclusion, the large fraction of intragenic miRNAs that are supposed to follow the regulation profile of their host gene, together with the huge amount of freely accessible breast cancer microarray data sets, offer an unprecedented opportunity to search for novel miRNAs involved in breast cancer, and the subversion of their related pathways. Results from our screening hold the potential to identify novel miRNA-based genetic signatures which can be used as diagnostic/prognostic tools for breast cancer. The miRNA signatures may offer an advantage over mRNA based signatures due to their exceptional stability even in formalin-fixed paraffin-embedded (FFPE) tissues (Li et al., 2007). At the same time, the identification of miRNAs with a role in breast cancer progression should increase our understanding of the genetics of metastasis, and allow us to explore novel miRNA-based epigenetic mechanisms subverted in human neoplasia.

2

Materials & methods

2.1 E1A and SV40 experiments

2.1.1 Affymetrix microarray preparation

Total RNA was extracted using TRIZOL (Invitrogen), and RNA integrity was checked with the 2100 Bioanalyzer (Agilent). Next, 5 μl containing of total RNA (1 μg/μl) were retrotranscribed in double stranded cDNA using T7-Oligo(dT) primers by SuperScript II kit (Invitrogen), as follows:

1. samples were incubated at 70°C for 10 minutes, then placed on ice
2. 3.5 μl of FS mix were added and samples were placed at 37°C for 2 minutes
3. 0.5 μl of SSII enzyme were added, then samples were incubated at 42°C for 1 hour
4. 65 μl of SS mix were added, then samples were placed at 16°C for 2 hours
5. 1 μl of T4 polymerase was added, then samples were placed at 16°C for 5 minutes
6. 5 μl of EDTA were added to stop reaction.

Double stranded cDNA was then purified by precipitation by adding 2 μl of glicogen, 80 μl of ammonium acetate and 400 μl of 100% ethanol, and incubated

at −80°C for 20 minutes. Samples were centrifuged at 10.000 rpm at 4°C for 30 minutes, and pellets washed by ice cold 70% ethanol (400 μl) followed by a 10 minutes centrifugation at 10.000 rpm at 4°C. Dry pellets were then resuspended in 1.5 μl of DEPC treated water. In vitro anti-sense RNA transcription was performed through an Eberwines modified in vitro transcription reaction (MEGAscript, Ambion) using labeled rNTP (Enzo® BioArray^{TM} HighYield^{TM} RNA Transcript Labeling Kit, ENZO Biolabs). In particular, we added 14.5 μl of rNTPs mix, 2 μl of T7 polymerase and 2 μl of reaction buffer to the 1.5 μl of purified cDNA, and the reaction mix was incubated at 37°C for 6 hours. Labeled cRNA were then fragmented (30-200 base fragments), checked by agarose gel, and hybridized on Affymetrix expression arrays (i.e. E1A experiments on a MOE 430 Plus 2 array, and SV40 experiment on a HG-U133A 2.0 array) using Affymetrix standard protocols.

2.1.2 E1A experiment

Infection of murine terminally differentiated myotubes with E1A is explained elsewhere (see (Nicassio et al., 2005)). Two independent experiments were performed on two different days (biological replicates). RNA was extracted from E1A infected and from control cells 24 and 36 hours after infection ("early" and "late" time points). RNA from each extraction was used to hybridize two Affymetrix Gene Chip Mouse Genome 430-2 (Affymetrix) arrays (technical replication). Microarrays were normalized using the RMA algorithm (Irizarry et al., 2003) implemented in the `affy` package from the Bioconductor suite of software libraries for the R programming language (Gentleman et al., 2004; R Development Core Team, 2010). All normalizations were performed using the default settings of the software.

Information relative to mouse intragenic miRNAs, associated host genes, and mature miRNA sequences was retrieved from the miRBase database (www.mirbase.org), release 13.0. The Entrez identifiers for the murine miRNA host genes were obtained from the `org.Mm.eg.db` Bioconductor annotation package. Probe sets mapping to miRNA host genes were identified by retrieving the probe set Entrez IDs from the `mouse4302.db` Bioconductor annotation package and successively matching these

IDs to the host genes Entrez ID.

Differentially expressed Probe sets upon E1A infection were identified in the early and the late time points by means of the `limma` Bioconductor package (Smyth, 2004, 2005). All the tests were two-sided. P-values were adjusted according to the Benjamini-Hochberg correction for multiple hypothesis testing (Benjamini and Hochberg, 1995). Probe sets were declared significant if the p-value was less than 0.1.

RT-PCR data were obtained from the TaqMan Human MicroRNA Array V.1.0 microfluidic card (Applied Biosystems). For each card, the median of the C_T values below 33 cycles was calculated, and TaqMan probes were normalized to this median C_T value. Relative expression for each mature miRNA was expressed as the ratio relative to the control cells.

We identified the human TaqMan probes that exactly matched a murine intronic mature miRNA by aligning the sequences of the TaqMan probes to the sequences of mouse mature miRNAs. Sequence alignment was performed using the `pairwiseAlignment` function, included in the `Biostrings` library from the Bioconductor suite. The default values of penalization for gap openings (-10) and gap extension (-4) were used.

We measured the strength and the significance of the agreement between TaqMan probes and the probe sets associated to miRNA host genes. The strength was quantified using the Spearman correlation coefficient. The significance of the agreement was measured by first transforming the Spearman's correlation coefficient into a new variable θ as shown in Eq. 2.1.

$$\theta = \sqrt{\frac{n-2}{1-r_s^2}} \tag{2.1}$$

For large samples ($n > 30$), θ has a t distribution with $n-2$ degrees of freedom, thus the null hypothesis of zero correlation can be tested with a t-test (Altman, 1990).

2.1.3 The SV40 experiment

The pBABE-neo vector containing the SV40-largeT antigen was transfected in amphotrophic Phoenix cells using calcium phosphate transfection reagent. 48 hours after transfection, viral supernatant was collected and used to infect MCF10A target cells. 48 hours after infection, cells were collected and total RNA was extracted.

Two Affymetrix HG-U133A2 arrays were hybridized with RNA extracted from control cells (*i.e.*, infected with "empty" retroviruses). and two arrays with RNA extracted from SV40 infected cells. The four arrays were normalized with the RMA algorithm. The same mRNA was used to hybridize two TaqMan Human MicroRNA Array V.1.0 cards, one for the control cells, and one for the SV40 infected cells.

Information on human intronic miRNAs, associated host genes and mature miRNA sequences was retrieved from miRBase 13.0. The Entrez IDs of the miRNA host genes was extracted from the `org.Hs.eg.db` Bioconductor annotation package. Probe sets mapping to miRNA host genes were identified by retrieving probe sets Entrez IDs by means of the Bioconductor `hgu133a2.db` Bioconductor annotation package, and matching such Entrez IDs to those of the miRNA host genes.

Similarly to the E1A experiment, differentially expressed probe sets were identified by means of the `limma` package, and probe sets were declared statistically significant if their Benjamini-Hochberg adjusted p-value was less than 0.1. The strength and statistical significance of the association of the TaqMan probes with the Affymetrix probe sets was measured, by the Spearman's correlation coefficient and by the t-test after transforming the correlation coefficient as shown in Eq 2.1.

2.2 The breast cancer microarray data sets

Breast cancer microarray data sets and the associated clinical information were downloaded from the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) data base. The accession numbers for the selected data sets are shown in Table 3.14. Information relative to human intronic miRNAs and associated host genes was retrieved from miRBase 14.0. All data sets were based on the Affymetrix HG-U133A

chip. We retrieved the CEL files of each data set from GEO and checked for the presence of defective arrays by means of the quality control procedure described below.

2.2.1 The quality control procedure

We have developed a semi-automated quality control procedure aimed to identify flawed arrays. The procedure is based on two statistical measures: the Relative Log Expression (RLE) and the Normalized Unscaled Standard Error (NUSE) (Bolstad, 2004; Gentleman et al., 2004).

For each data set we performed the following steps:

1. We computed the median value and the IQR of both the NUSE and the RLE statistics for each array. This gave four values for each chip: M_{NUSE}, M_{RLE}, IQR_{NUSE} and IQR_{RLE}.

2. We compared each of these values with a corresponding cutoff value. If any value exceeded the cutoff, the chip was tagged as "dubious".

3. We calculated the IQRs of M_{NUSE}, M_{RLE}, IQR_{NUSE} and IQR_{RLE} across the arrays. This gave four values: $IQR_{M_{NUSE}}$, $IQR_{M_{RLE}}$, $IQR_{IQR_{NUSE}}$ and $IQR_{IQR_{RLE}}$.

4. If any of these IQR values was greater than $q_3 + 1.5IQR$ or less than $q_1 - 1.5IQR$, where q_1 and q_3 are the first and the third quartile of the distribution, we considered such values as outliers and flagged the corresponding array as "rejected".

5. We made diagnostic plots for both the dubious and the rejected arrays for a successive visual analysis.

Cutoff values were chosen heuristically, with a preference for overestimating the number of poor quality chips rather than failing to identify a compromised chip. The four cutoff values are shown in Table 2.1. Arrays identified as defective were

M_{NUSE}	M_{RLE}	IQR_{NUSE}	IQR_{RLE}
1.10	0.2	0.10	1.0

Table 2.1: The chosen cutoff values for the four quantities defined in step 1 of the quality control procedure.

removed from the data set before normalization. Each data set was normalized separately using the RMA algorithm with default settings. The Entrez IDs of miRNA host genes was retrieved by means of the `org.Hs.eg.db` Bioconductor annotation package. Probe sets mapping to miRNA host genes were identified by retrieving the probe set Entrez IDs from the `hgu133a.db` annotation package and successively matching these IDs to the host genes Entrez ID. Probe sets were filtered for signal intensity using the Bioconductor `genefilter` package. Only the probe sets that had a normalized signal greater than 150 (7.2 on the log_2 scale) in at least 10% of the samples were retained for further analysis. Significantly regulated probe sets between the states of any given clinical parameter were identified by means of the `limma` package. Probe sets with a Benjamini-Hochberg-adjusted p-value less than 0.05 were declared significantly regulated.

2.2.2 Permutation test

We performed a permutation test to determine whether the percentage of significantly regulated miRNA host genes found in a data set would also have resulted if we had considered genes that do not contain miRNAs. For each data set:

1. We removed all miRNA-associated probe sets from the data set.

2. We randomly selected n probe sets where n is the number of miRNA-associated probe sets ($n = 465$).

3. We filtered this data set with respect to probe set intensity (intensity > 150 in at least 10% of the samples.

4. We determined the number of probe sets that were significantly regulated between G3 vs. G1 and ER+ vs. ER- tumours.

5. Repeated steps 2–4, 999 times

An empirical p-value was calculated as the fraction of simulations yielding a larger list of significantly regulated probe sets than the list obtained in the original analysis.

2.3 The cell lines and the FFPE samples

2.3.1 miRNA extraction and Real Time PCR analysis.

Total RNA was extracted from cell lines using the TRIZOL reagent (Invitrogen) and from FFPE archive breast tumors using the RNAeasy FFPE kit (QIAGEN). Next, 0.5 μg of total RNA was retrotranscribed using the miSCRIPT reverse transcription kit (QIAGEN). A total of 5 ng of cDNA was used as a template for real time PCR (LightCycler 480, Roche) in a 25μl reaction by using the miSCRIPT primer assay (QIAGEN) and miScript SYBR Grees PCR kit (QIAGEN). The following PCR protocol was used: i) hold, 15 minutes at 95°C; ii) 40 cycles, 15 seconds at 95 °C, 30 seconds at 55°C, 30 seconds at 70°C; iii) (Melting curve) 15 seconds at 95°C, 30 seconds at 55°C then ramp to 95°C (0.11°C per second).

Six miRNAs were screened on a panel of commercial cell lines for validation. The selected miRNAs were *miR-342*, *miR-483*, *miR-548f-2*, *miR-218*, *miR-1245*, *miR-1266*. Primers were ordered for both the 3p and 5p strands of *miR-218*, *miR-342*, and *miR-483*. The normal breast HMEC and MCF10A cell lines, and the breast tumour MB-231, MB-361, MCF7, BT474 cell lines. Expression was normalized to the housekeeping small nuclear RNA U6. The expression levels of *mir-218/218**, *miR-342-3p/5p*, *miR-483-3p/5p* and *miR-1266* was screened on a collection of 36 FFPE tumour samples provided by the European Institute of Oncology, Milan. Expression was normalized to the housekeeping small nuclear RNA U5A.

The statistical significance of miRNA regulation with respect to tumour grade (G3 vs. G1) and ER status was assessed by Mann-Whitney tests, performed with the R statistical environment.

2.4 Reclassification of G2 samples

Hierarchical clustering of expression data was performed using the `heatmap.2` function from the `gplot` library. Data were row-standardized by subtracting the mean and dividing by the standard deviation. We used euclidean distance as a measure of dissimilarity and the "average linkage" method to build the cluster hierarchy.

Survival analysis was performed using the functions included in the `survival` R library. Survival was estimated by Kaplan-Meier estimation procedure, and difference in survival between groups was tested by logrank test.

2.5 MiRNA overexpression

MDA-MB-231 breast cancer cells (ATCC) were transfected with 20nM miScript miRNA Mimics (QIAGEN) using Lipofectamine 2000 transfection reagents (Invitrogen). After 72 hours post-transfection, total RNA was extracted using TRIZOL and real Time PCR performed (as previously described) to check for the expression of *miR-342-3p* and *miR-342-342-5p* mimics. AllStars Negative Control siRNA (QIAGEN) was used as negative control siRNA. This siRNA has no homology to any known mammalian gene. FACS analysis of transfected cells was used to monitor cell cycle distribution and apoptosis (TUNEL assay). Briefly, one million cells of every condition were resuspended in formaldehyde 2% and incubated on ice for 20 minutes. Cells were then washed in PBS 1% BSA (Bovine Serum Albumin, SIGMA), fixed in 3:4 ethanol 100% and incubated on ice for 30 minutes. Next, cells were washed again in PBS BSA 1%, resuspended in 50 μl of TdT solution (in Situ Cell Death Detection Kit, Fluorescein, ROCHE) and incubated 1 hour at 37°C. Lastly, cells were washed in PBS 1% BSA and stained in a 500ul PBS solution of Propidium Iodide (2.5 μg/ml) plus RNAse (6.25 μg/ml), and incubated overnight at 4°C before FACS analysis.

3

Results

3.1 Introduction

The overall aim of this thesis is the identification of novel onco-miRNAs relevant to breast cancer through the analysis of microarray gene expression data. A necessary first step in our analysis was to determine whether microarray data could indeed be used to reliably infer the expression of intronic miRNAs by measuring the expression of the relative host genes. This first part of our analysis is referred to as the "proof of principle". We used two experimental model systems that are relevant to cancer for the proof of principle. The first model is based on terminally differentiated (TD) murine myotube cells induced to re-enter the cell-cycle by infection with the early region 1 A (E1A) viral oncogene, an early gene product of tumorigenic adenoviruses (Crescenzi et al., 1995; Kirshenbaum and Schneider, 1995; Tiainen et al., 1996) (see Sec. 3.2). The second model is based on transformation of the non tumorigenic human breast cell line MCF10A by infection with the Simian vacuolating virus 40 (SV40) (Hahn et al., 1999; Van Der Haegen and Shay, 1993) (see Sec. 3.3). In both experimental systems, we measured the regulation of the miRNA host genes by Affymetrix and compared it with the regulation of the relative mature miRNAs, measured by real time-PCR (RT-PCR). Following the "proof of principle" experiments, we proceeded to the second part of our analysis, *i.e.,* the identification of cancer regulated miRNAs through the profiling of more than 900 microarrays from

five distinct breast cancer data sets (see Sec. 3.4).

3.2 The proof of principle - E1A experiment

The E1A viral oncogene is able to force re-entry into the cell cycle of terminally differentiated (TD) mouse myotubes derived from C2C12 myoblasts upon serum deprivation (Tiainen et al., 1996). The action of E1A has been largely linked to its interference with the growth suppression function of the pRb-family proteins (Bandara and La Thangue, 1991; Whyte et al., 1988). The, the deletion of pRb alone, however, is not sufficient to induce cell cycle re-entry of TD cells (Camarda et al., 2004; Huh et al., 2004). Thus, additional E1A-induced pathways must also be involved, most probably the E1A-regulated pocket protein/pRb-independent pathways, including those relying on CycAE/CDK2, C-terminal binding protein (CtBP), transformation/transcription domain-associated protein (TRRAP), p400, PCAF–associated factor (PCAF) and other chromatin remodeling activities (Alevizopoulos et al., 1998; Deleu et al., 2001; Faha et al., 1993; Fuchs et al., 2001; Ghosh and Harter, 2003; Reid et al., 1998; Shi et al., 2003). Considering the number and complexity of the molecular mechanisms required for cell cycle re-entry of TD cells, we expected miRNAs to play a relevant role in this process. We thus measured the change in expression of the mature intronic-miRNAs upon E1A infection, and compared it to the regulation of their host genes in the same experimental conditions, to assess their correlation and, at the same time, to possibly identify E1A regulated intronic miRNAs.

The gene expression profiling was based on Affymetrix MOUSE430-2 murine microarrays. mRNA was extracted 24 and 36 hours after E1A infection. The purpose of this double extraction was to identify genes that are regulated at an early stage of infection, mainly by pocket protein/pRb dependent pathways, and those regulated at a late phase by pocket protein/pRb-independent pathways (Nicassio et al., 2005). We used eight microarrays: four for the E1A infected samples, and four for controls infected with a deleted mutant of E1A that completely abrogates its ac-

tivity (Tiainen et al., 1996). We performed two independent mRNA extractions on different days (biological replicates) and used each of these replicates to hybridize two microarrays (technical replication). Thus, the technical replicates were nested within the biological replicates (see Fig. 3.1). At the time of this analysis, no high

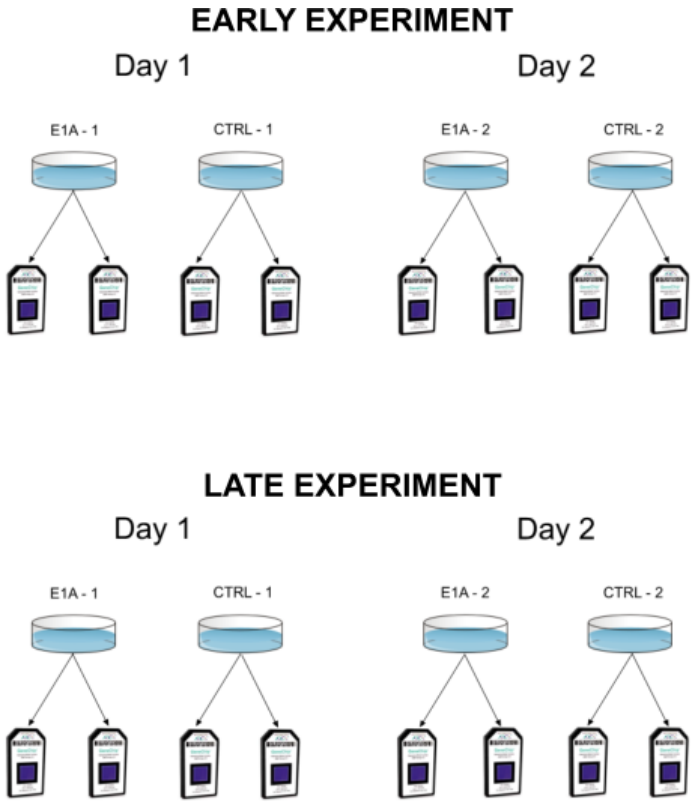


Figure 3.1: **Design of the E1A experiment**. We extracted the RNA from the control (i.e., an E1A deleted mutant) and the E1A infected cells on two different days, thus having two independent biological replicates of the two experimental condition. Each of these RNAs was used to hybridize two microarrays (technical replication). The same scheme was used both for the "early" and the "late" experiments.

throughput RT-PCR platform was available to investigate the expression of mature miRNAs in mouse. For this reason the mature miRNA profiling was based on the Applied Biosystems TaqMan Human MicroRNA Array V. 1.0 which contains 365 probes designed to target an equal number of mature human sequences. The probes on the card were based on the mature sequences stored in the miRBase miRNA database (www.mirbase.org) release 10.0. Since these probes refer to human mature sequences, it was necessary to identify which mature miRNAs were conserved between mouse and human. In the RT-PCR experiment one card was used for the

	All	Intragenic	Transcript Names
Mus Musculus	547	254	244

Table 3.1: Total number of pri-miRNAs, intragenic miRNAs, and associated transcript names in Mus Musculus according to miRBase release 13.0.

E1A transfected cells, and another for the control cells. To determine which miRNAs were regulated, we considered the fold change of expression between the two experimental conditions (control vs. E1A infected cells).

To compare the Affymetrix probe sets with the TaqMan PCR probes we:

1. Identified the intronic miRNAs, their host genes, and the mature miRNAs they produce in mouse.

2. Identified which of these mature miRNAs were conserved in human, and could therefore be analyzed by a human TaqMan microfluidic card.

3. Determined which Affymetrix probe sets could be associated with the murine miRNA host genes.

4. Identified potentially problematic cases (see below).

5. Analyzed the microarray experiment, identified all the statistically significant probe sets associated with an intronic miRNA host gene, and compared their regulation to that of the corresponding mature miRNA.

We used miRBase (release 13.0) as a source of information for our analysis. In miRBase, miRNAs are classified either as intragenic or as associated with a transcript name, usually (in the case of murine miRNAs) a Mouse Genome Informatics (MGI) symbol. In some cases however, the transcript name refers to other types of identifiers, such as NCBI Nucleotide loci, MGI nucleotide/probe clones etc. The total number of mouse miRNAs and the number of the intragenic mouse miRNAs, are shown in Table 3.1.

We retrieved the NCBI Entrez Gene identifier (Entrez ID) of the host genes of the intragenic miRNA. This step was necessary for the successive mapping of miRNA host genes to Affymetrix probe sets. To retrieve the Entrez IDs, we took

pri-miRNAs	mature miRNAs	Entrez Ids	Affy probes	Taqman probes
66	63	58	156	63

Table 3.2: Number of Pri-miRNAs, mature miRNAs, Entrez Ids, Affymetrix probe sets and TaqMan probes that could be reliably used for the comparison of the miRNA host genes with the corresponding mature sequences

advantage of the R statistical environment (R Development Core Team, 2010), and of the Bioconductor suite of bioinformatics packages (Gentleman et al., 2004). More specifically we used the `org.Mm.eg.db` annotation package, which provides the mapping of the Entrez IDs to a large number of biological identifiers, including the MGI symbols and their aliases. We found that 194 of the 244 available transcript names had an associated Entrez Gene ID.

To map murine mature sequences to human TaqMan probes we assumed that even a single nucleotide mismatch could compromise the specificity of the probe. We, therefore, applied a stringent sequence alignment procedure to identify the TaqMan probes that perfectly matched the mature murine sequences, or contained them completely without mismatches. For this purpose we used the functions available in the `Biostrings` library, from the Bioconductor suite. From this alignment, we found that 128 human TaqMan probes could reliably be associated with the same number of murine mature miRNAs and to 150 miRNA host genes[1]. Sixty-six of these 150 miRNA host genes could be mapped to 156 Affymetrix probe sets. For this mapping we used the `mouse4302.db` annotation package from Bioconductor, which contains the mapping from the probe set IDs of the Mouse-420-2 chip to several other identifiers, including the Entrez ID. The summary of this mapping step is shown in Table 3.2.

While the majority of mature miRNAs are the product of a single precursor (we refer to this category of miRNAs as *singletons*), some mature miRNAs have more than one precursor. Table 3.3 shows the number of mature miRNAs ordered by the number of precursors from which they are derived. For the mature miRNAs that have more than one precursor, and are thus encoded by different miRNA genes, we

[1]We have more miRNA genes than mature sequences because different miRNA genes can produce the same mature miRNA.

No. of precursors	1	2	3	6
Frequency	522	40	7	1

Table 3.3: The same mature sequence can be produced by more than one precursor. Here we report the frequency of mature miRNAs having a given number of precursors. In the vast majority of cases one precursor produces one mature sequence (*singletons).*

	mature miRNAs	Singletons	all intergenic	all intragenic	inter and intragenic
All miRNAs	570	522	16	12	20
selected miRNAs	63	48	0	4	11

Table 3.4: Classification of the murine mature miRNAs stored in miRBase release 13.0 with respect to the typology of the precursors.

can have three situations: 1) all the precursors are intergenic, 2) all the precursors are intragenic, 3) some of the precursors are intergenic, while others are intragenic. The first case does not represent a problem, since we are not able to infer the regulation of such miRNAs with gene expression microarrays. Both the second and the third cases can however be problematic, since the information obtained from the Affymetrix probe sets would be only partial, and therefore potentially misleading. The 547 murine precursors sequences contained in miRBase 13.0 correspond to 570 mature miRNAs[1], which can be classified as shown in Table 3.4.

Restricting our attention to miRNAs that can be simultaneously mapped to an Affymetrix probe set and to a TaqMan probe, we found that 48 mature miRNAs were intragenic singletons, 4 had intragenic only precursors, and 11 had inter and intragenic precursors.

3.2.0.1 Analysis Of The Microarray Data

Before identifying the differentially expressed probe sets we performed a quality control procedure on the Affymetrix chips ensure that no artifacts or technical problems were present. We used the `affyPLM` library from the Bioconductor suite (Bolstad, 2004), which provides a false color visualization of the surface of chips, making scratches or other spatial defects clearly visible. From the quality control procedure

[1]There are more mature miRNAs than precursors because one precursor can give rise to two distinct miRNAs, depending on the selected strand

Time point	Probe sets	Gene Symbols	Precursors	Mature miRNAs
Early	53	30	36	33
Late	46	31	38	36

Table 3.5: Number of probe sets and corresponding gene symbols, precursors and mature miRNAs that were differentially expressed upon E1A transfection at the early and late time point.

no particular problems emerged. We therefore proceeded to normalize the arrays with the Robust Multichip Average algorithm (Irizarry et al., 2003). The Affymetrix Mouse-430-2 array contains 45,101 probe sets, of which 37,316 could be unambiguously mapped to 20,448 gene symbols. We restricted the data set to the 156 probe sets that could be associated with a TaqMan probe and examined their regulation at the early and late time points separately. We used the `limma` Bioconductor package (Smyth, 2004, 2005) to identify the differentially expressed probe sets. The relatively complex design of the experiment, with technical replicates nested into biological replicates, made the use of a t-test highly inefficient. The `limma` package, instead, is well suited to this kind of experimental designs. We adjusted the p-values with the Benjamini-Hochberg procedure to control the false discovery rate (FDR)[1] (Benjamini and Hochberg, 1995) and retained only the probe sets with an adjusted p-value smaller than 0.1. We opted for this relatively mild cutoff to compensate for the limited statistical power due to the small sample size. We identified 53 regulated probe sets at the "early" time point, corresponding to 30 gene symbols, 36 precursors, and 33 mature miRNAs. At the "late" time point, we identified 46 regulated probe sets corresponding to 31 gene symbols, 38 precursors[2] and 36 mature miRNAs. These numbers are summarized in Table 3.5.

We considered the Affymetrix probe set/TaqMan probe pairs separately, instead of taking the average or the median of the probe sets referring to the same gene, as is common practice in the field. This practice can be misleading, as illustrated by the following example: the estrogen receptor 1 (*ESR1*) gene is mapped by nine probe

[1] Since these values relate to FDRs rather than rejection probabilities, they are sometimes called q-values.

[2] The fact that there are more precursors than host genes is due to the fact that some miRNA genes are clustered within the same host gene.

sets in the Affymetrix HG-U133A array. By definition, ER+ samples have higher levels of *ESR1* than ER- samples. If we plot the boxplots of the expression levels of these nine probe sets, stratifying by ER status, the difference between the probe sets clearly emerges (Fig. 3.2). Only the `205225_at` probe set clearly discriminates the ER+ from the ER- samples, while such discrimination is much weaker, or absent, in the other probe sets. Thus, averaging these probe sets would make this difference undetectable. This example explains why an increasing number of software tools select only the most regulated probe set for each gene, discarding the others, when two or more probe sets match the same gene. This is, for example, the default setting in the Gene Set Enrichment Analysis (GSEA) software package (Subramanian et al., 2005).

We compared the regulation of the miRNA host genes, measured by Affymetrix, and of their corresponding mature miRNAs, measured by TaqMan. The results of these comparisons are shown in Figs. 3.3 and 3.4 for the early and late time points respectively. These plots show all the Affymetrix probe set/TaqMan probe pairs, including the problematic cases; however, for the quantitative comparison of the two technologies only the singleton miRNAs were considered.

At the early time point, 31 of the 42 statistically significant probe sets (i.e., 74%) had the same direction of regulation as the corresponding TaqMan probes (Spearman correlation coefficient $r_s = 0.70$, p-value 6×10^{-9})[1]. At the late time point 34 of the 39 significant probe sets (i.e., 87%) were regulated coherently with the corresponding TaqMan probes (correlation coefficient $r_s = 0.62$, p-value 2×10^{-6}).

Affymetrix probe set/TaqMan probe pairs displaying an opposite direction of regulation are referred to as the "disagreeing" cases. At the early time point there were 6 disagreeing cases (Table 3.6), while at the late time point there were 3 disagreeing cases (Table 3.7). For more than half of these cases the C_T values were greater than 30 in both the E1A transfected and control samples. Thus, the estimation of the log fold change is unreliable for these samples.

The miRNAs *let-7g* and *mir-27b*, however, have C_T values less than 28, therefore,

[1]The p-value were calculated based on a correlation-test, where the null hypothesis was that the correlation coefficient equaled zero.

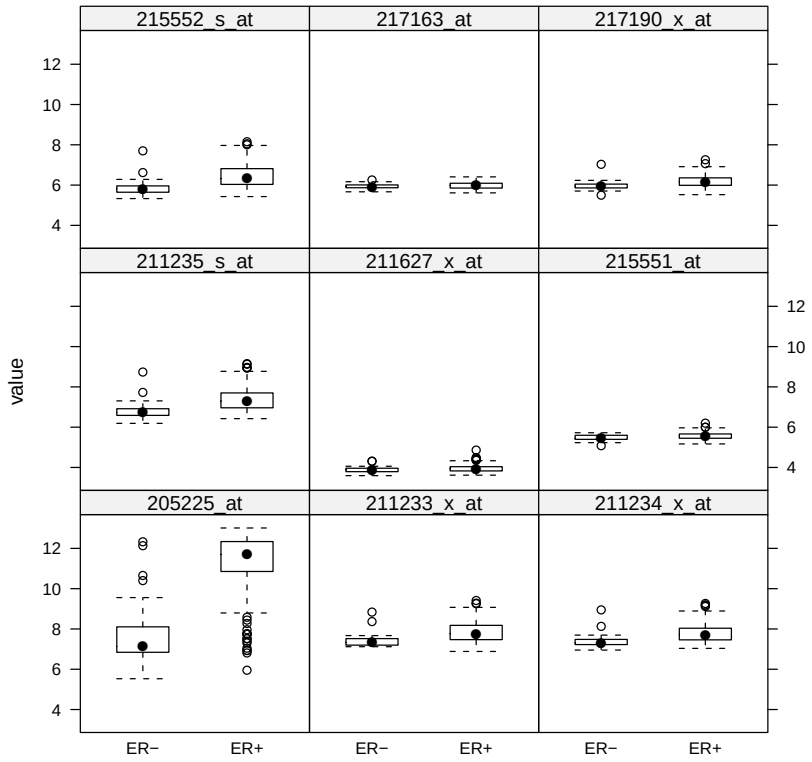


Figure 3.2: Boxplot of the expression values of the nine probe sets mapping the ESR1 gene, stratified by ER status. The probe set in the lower left corner, 205225_at, clearly discriminates the two groups, while the other probe sets display little regulation.

transcript_name	mature miRNA	PCR log-FC	Affy log-FC	Ct e1a_24h	Ct ctrl
Tln2	hsa-miR-190	0.11	-0.20	33.79	33.90
2610318N02Rik	hsa-miR-130b	-0.17	0.63	29.71	29.54
2010111I01Rik	hsa-miR-27b	0.25	-0.66	27.09	27.33
Nfyc	hsa-miR-30e-5p	-0.18	0.34	30.95	30.76
Wdr82	hsa-let-7g	-0.31	0.22	27.97	27.66
3110082I17Rik	hsa-miR-339	-0.66	0.35	31.02	30.37

Table 3.6: Disagreeing cases at the early time point. The host gene (transcript name), mature miRNA ID, TaqMan (PCR) and Affymetrix (Affy)log fold change (log-FC), and adjusted C_T values for the E1A-infected (E1A-24h) and control (ctrl) samples, are reported for the disagreeing cases.

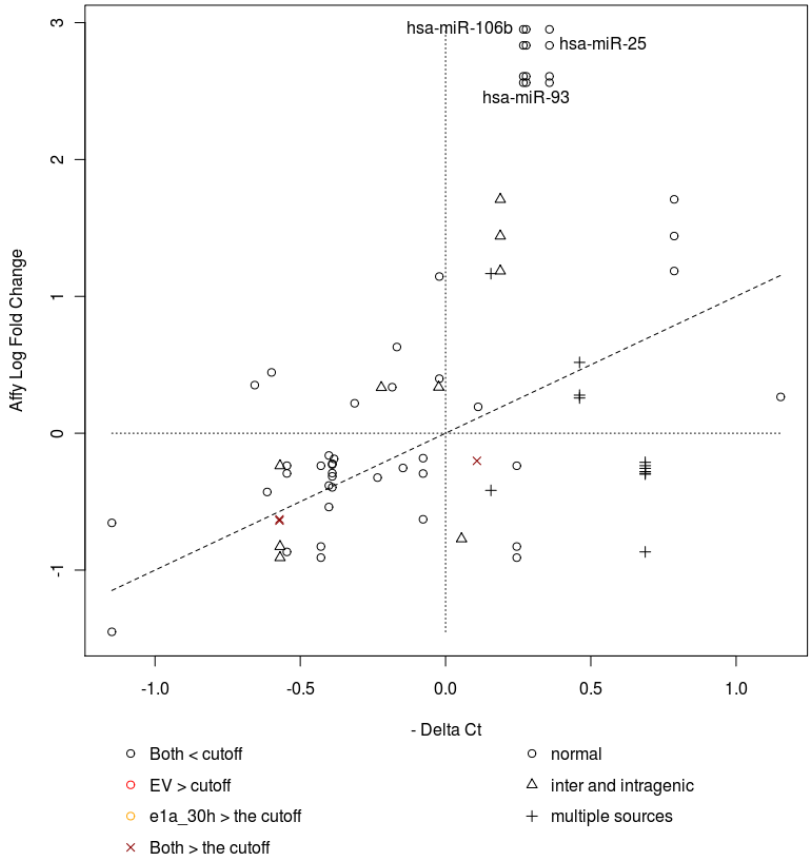


Figure 3.3: **Comparison of the regulation of the mature miRNAs and of their host genes at the early time point.** - The log-2 fold change of the mature miRNAs, as measured by TaqMan is shown on the x-axis. The log-2 fold change of the corresponding host genes, as measured by Affymetrix, is shown on the y-axis. More than one probe set can be associated with the same mature miRNA and vice versa. Triangles indicate mature miRNAs that have both intergenic and intragenic precursors. Vertical crosses refer to mature miRNAs having multiple precursors that are all intragenic, although located within different genes (multiple sources). Red circles indicate cases in which the C_T of the control cells was above 33 cycles (meaning that the results were not reliable). Orange circles refer to a similar problem in the E1A transfected cells. Finally red crosses indicate that the C_T of both the infected and the control cells were above 33 cycles. The dashed line represents the ideal case of a perfect agreement between Affymetrix and TaqMan data.

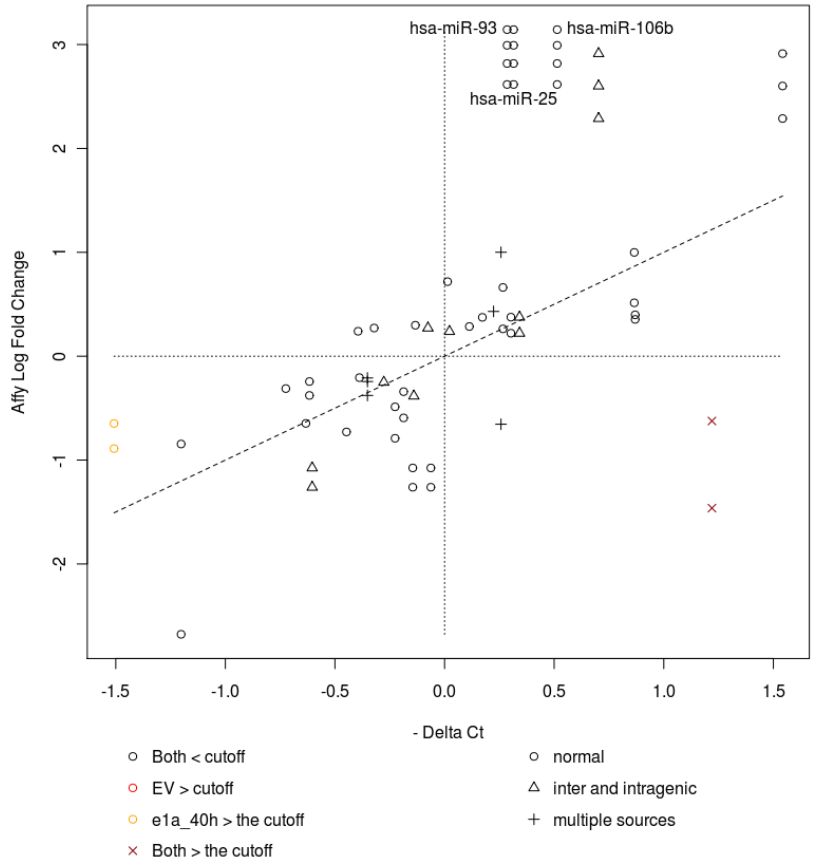


Figure 3.4: **Comparison of the fold change of the mature miRNAs and of the host genes at the late time point** - Same plot as in Fig. 3.3 although for the late time point.

the estimation of the log fold change should be more reliable for these miRNAs. A possible explanation for the discrepancy in results could be the presence of upstream regulatory elements that control the expression of miRNAs independently from the host gene promoter. Indeed, Monteys and collaborators have demonstrated that roughly one third of human intronic miRNAs, including *let-7g* and *miR-27b* have upstream regulatory elements possessing features consistent with promoter function (Monteys et al., 2010). It is therefore plausible that in mouse these miRNAs might be regulated in a host gene independent fashion. In addition Newman and co-workers showed that some RNA binding proteins, such as LIN-28, can inhibit the processing of the *let-7g* precursor, thus preventing the accumulation of its mature form (Newman et al., 2008).

At both the early and late time points, several probe set/TaqMan pairs showed the same direction of regulation, although the degree of regulation was strikingly different. This was the case for *miR-106b*, *miR-25* and *miR-93* (Figs 3.3 and 3.4) which are members of a cluster contained in the MCM7 gene. The four probe sets that mapped this gene were 5 to 8 fold up-regulated at both time points (2.5 to 3 fold on the log scale), while the up-regulation of the three TaqMan probes matching these miRNAs ranged from 1.2 to 1.5 fold (0.25 to 0.5 on the log scale). Similarly to *let-7g* and *miR-27b*, regulatory elements with features compatible with promoter functions have been found upstream of the *miR-25-93-106b* cluster, suggesting that the it might be regulated independently of its host gene (Monteys et al., 2010).

In this analysis we only considered the probe sets that displayed statistical significance. When, however, we considered all probe set/TaqMan probe pairs, without filtering for significance, we observed no correlation between the Affymetrix and

transcript_name	mature miRNA	PCR log-FC	Affy log-FC	Ct e1a_36h	Ct ctrl
Fgf13	hsa-miR-504	1.22	-1.04	34.12	35.34
Nfyc	hsa-miR-30e-5p	-0.32	0.27	31.08	30.76
Wdr82	hsa-let-7g	-0.13	0.30	27.79	27.66

Table 3.7: Disagreeing cases at the E1A late time point. The host gene (Transcript name), mature miRNA ID, TaqMan (PCR) and Affymetrix (Affy) log fold change (log-FC), and adjusted C_T values for the E1A infected (E1A 36h) and control (ctrl) samples, are reported for the disagreeing.

Total	Intragenic	Intergenic
706	407	299

Table 3.8: Total number of intragenic and intergenic human miRNAs in release 13.0 of miRBase.

the TaqMan measurements (Fig. 3.5). This result demonstrates the importance of filtering for significance.

3.3 The proof of principle - SV40 experiment

SV40 is a polyomavirus that is found in both monkeys and humans. It was first identified in 1960 in cultures of rhesus monkey kidney cells that were being used to produce polio vaccine. We used SV40 to transform the non-tumorigenic breast cancer cell line, MCF10A, and performed gene expression profiling of both SV40-transformed and control cells. The microarray experiment consisted of four Affymetrix HG-U133A2 arrays, each containing 22,277 probe sets corresponding to 12,690 gene symbols. Two arrays were hybridized with RNA extracted from SV40 infected cells, and two with the RNA from the control cells (*i.e.*, the empty retroviral vector). In the RT-PCR TaqMan experiment, we used one microfluidic card for the infected cells, and one card for the control cells. The card was the same as the one used in the E1A experiment (TaqMan Human MicroRNA Array V. 1.0). We followed the same procedure as described for the E1A experiment to analyze data from the SV40 experiment, with the exception of step 2, *i.e.*, the mapping from murine to human miRNAs. This step was not necessary because we used the human cell line, MCF10A, in the SV40 experiment.

Release 13.0 of miRBase contains 706 human pri-miRNAs, corresponding to 703 mature miRNAs. The number of inter and intragenic miRNAs in human is shown in Table 3.8.

The 407 intragenic miRNAs were associated with 386 transcript names, 317 of which had a unique Entrez ID. We aligned TaqMan probe sequences to the mature miRNA sequences contained in miRBase 13.0 to remove possible erroneously designed probes, and found that 268 of the 365 probes on the card reliably matched a

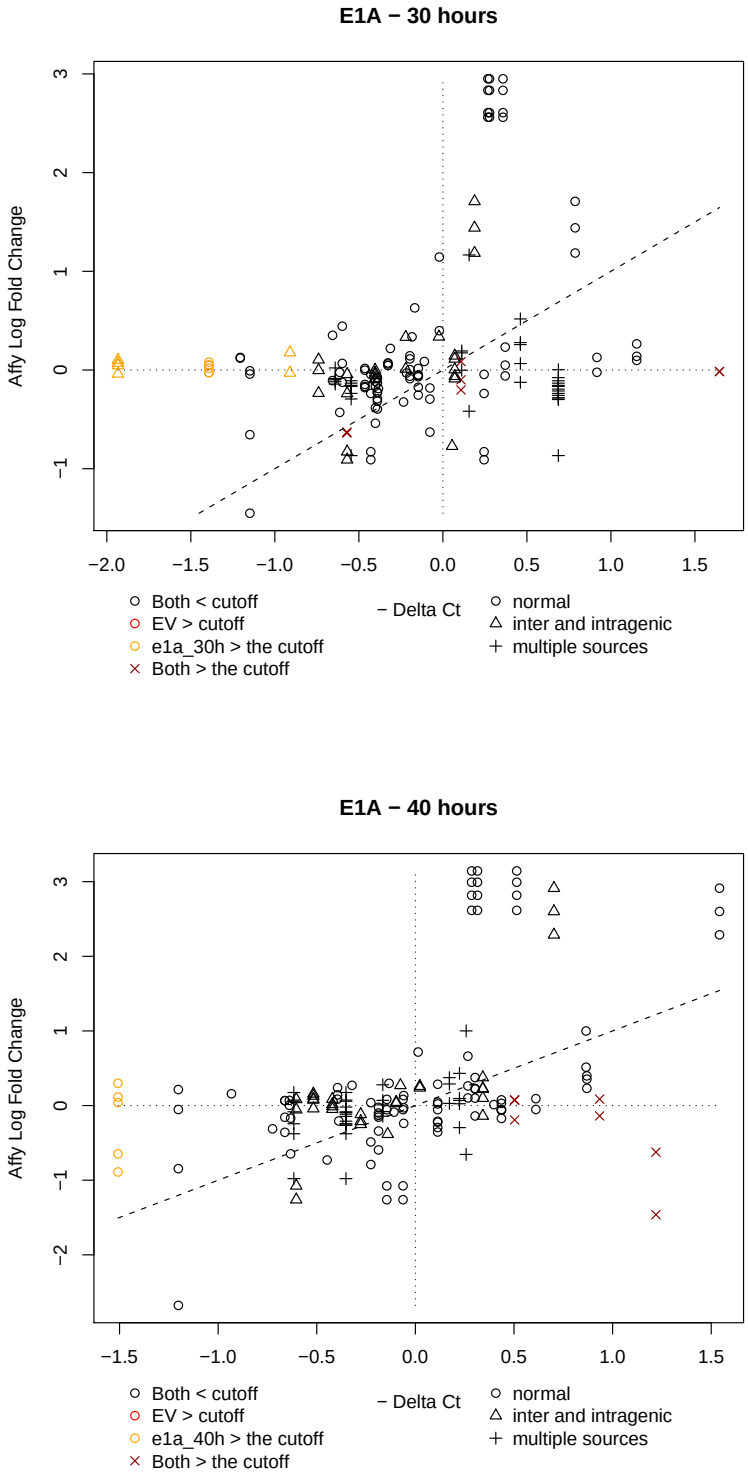


Figure 3.5: **Comparison of the log fold-changes of all the probe sets, irrespective of their statistical significance, in the E1A experiment**. *Top*: same plot as in fig. 3.3, although without filtering the probe sets for statistical significance. *Bottom*: analogous plot for the late time point.

N. of precursors	1	2	3	4	5	8
Frequency	632	59	7	3	1	1

Table 3.9: Distribution of the number of human mature miRNAs associated with a particular number of precursors.

mature sequence.

As in the E1A experiment we determined the number of mature miRNAs falling into the different precursor categories and identified the problematic cases (Table 3.10).

For the Affymetrix data, we used the `hgu133a2.db` annotation package from Bioconductor to retrieve the probe sets from the Entrez IDs and found 436 probe sets mapping to the 317 Entrez IDs. We identified 102 mature miRNAs that could be associated with both a TaqMan probe and at least one Affymetrix probe set (Table 3.11).

As in the previous analysis, we used the `limma` package to identify significantly regulated probe sets using a cutoff adjusted p-value of less than 0.1 (Benjamini-Hochberg adjustment). We identified 56 significantly regulated probe sets, corresponding to 42 gene symbols and 54 intragenic miRNAs.

Excluding the problematic cases, we found that 21 of the 29 (72%) statistically significant probe sets probe/TaqMan pairs had the same direction of regulation (Fig. 3.6). The Spearman correlation coefficient was positive ($r_s = 0.44$), and significantly different from zero ($p = 0.008$). We observed 8 disagreeing probe set/TaqMan probe pairs (Table 3.12).

For 6 of the disagreeing cases, the C_T values were greater than 30 in both the control and infected samples, indicating that these results are unreliable. In contrast, *hsa-miR-149* and *hsa-miR-328* displayed low C_T values and, therefore, the results should be reliable. The presence of a putative promoter region upstream of the *hsa-*

singletons	all_intergenic	all_intragenic	inter_and_intragenic
632	30	22	19

Table 3.10: Distribution of human mature miRNAs in the different precursor categories.

pri-miRNAs	mature miRNAs	Entrez Ids	Affy probes	Taqman probes
111	102	89	154	102

Table 3.11: Number of Pri-miRNAs, mature miRNAs, Entrez Ids, Affymetrix probe sets and TaqMan probes that could be reliably used for the comparison of the miRNA host genes and the mature sequences

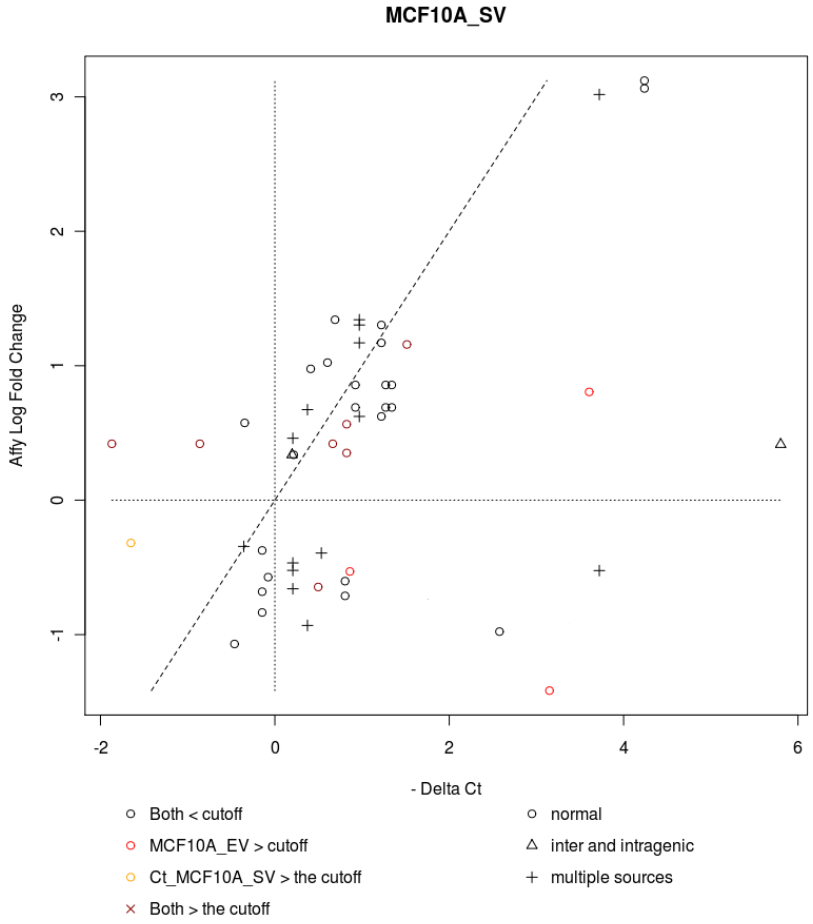


Figure 3.6: **Results of the SV40 experiment** - Comparison of the log fold changes of the mature miRNAs and of the host genes of their precursors in the SV40 experiment.

transcript_name	assay_name	PCR log-FC	Affy log-FC	Ct SV40	Ct ctrl
SREBF2	hsa-miR-33	0.86	-0.53	32.42	33.28
GPC1	hsa-miR-149	0.81	-0.66	25.43	26.24
DNM3	hsa-miR-214	-0.35	0.57	32.36	32.02
LARP7	hsa-miR-302c	-0.86	0.42	36.24	35.38
LARP7	hsa-miR-367	-1.87	0.42	36.60	34.73
KIAA1199	hsa-miR-549	3.15	-1.42	32.86	36.01
ELMO3	hsa-miR-328	2.58	-0.98	25.90	28.48
DENND1A	hsa-miR-601	0.50	-0.65	34.38	34.88

Table 3.12: Transcript name, mature miRNA, PCR and Affymetrix log fold change, adjusted C_T values in the infected and in the control cases for the disagreeing case in the SV40 experiment.

miR-149 host gene, GPC1, might explain the apparent discrepancy in the results for this miRNA (Monteys et al., 2010).

3.3.0.2 Proof of principle - conclusions

The main goal of the proof of principle experiments was to determine whether the regulation of intronic miRNAs could be inferred from Affymetrix profiling experiments assessing gene expression levels. Overall we found a good correlation between the expression of miRNA host genes, assessed by Affymetrix, and mature miRNA expression assessed by RT-PCR. In terms of direction of regulation, the agreement between the Affymetrix and PCR results ranged from 72% (in the SV40 experiment) to 87% (in the E1A late experiment). In terms of the Spearman correlation coefficient, the agreement ranged from a minimum of $r_s = 0.44$ (SV40 experiment) to a maximum of $r_s = 0.77$ (E1A, early time point). In some cases the two technologies, despite being in agreement regarding the direction of regulation, were remarkably different in terms of quantification of the effect. This was not surprising considering the different sensitivities of the two methods, and the small sample size of the Affymetrix experiments.

3.4 The breast cancer data sets

The second step of the present research project was the identification of novel oncomiRNAs in breast cancer. Breast cancer microarray data sets were taken from Gene

Expression Omnibus (GEO) database (Barrett et al., 2005) which is the largest repository of publicly available gene expression data. The database contains more than 450,000 entries, most of which come from microarray experiments, although Next Generation Sequencing and high throughput RT-PCR data are also available. We used four criteria to select microarray data for our analysis. First, data sets had to have a large sample size in order to achieve high statistical power. Second, data sets had to be accompanied with adequate clinical and pathological information. Third, data sets needed to be based on the same microarray platform, to avoid the introduction of additional non-biological variability. Fourth, the raw needed to be available. The raw data were necessary because we wanted to perform an accurate quality control of the samples in order to reduce the noise introduced by poor quality arrays. In the E1A and SV40 experiments, the small number of arrays allowed for a visual inspection of every chip. To analyze hundreds of arrays, however, it was necessary to develop a quality control pipeline that allowed us to automatically identify potentially compromised arrays, which could then be visually inspected.

3.4.1 The quality control procedure

Microarray experiments can be influenced by two types of problems: those affecting single arrays, and those conditioning a significant portion of the data set (*batch effects*). In the first case it usually suffices to remove the defective array. In the second case, however, no easy solution can be provided, and additional analyses need to be performed. Affymetrix arrays are designed to be robust against local effects, but a large spatial defect such as the one shown in Fig. 3.7 can compromise the array's reliability. Two statistical measures have been introduced by Benjamin Bolstad, which allow the identification of this kind of defect, the Relative Log Expression (RLE) and the Normalized Unscaled Standard Error (NUSE) (Gentleman et al., 2004). In the RLE approach, the log scale estimates of the expression $\hat{\theta}_{gi}$ of each gene g on each array i are used to compute the median expression m_g of each gene across arrays. The relative expression is then defined as $M_{g1} = \hat{\theta}_{gi} - m_g$. These values are used to draw a boxplot for each array. Normally, we expect that

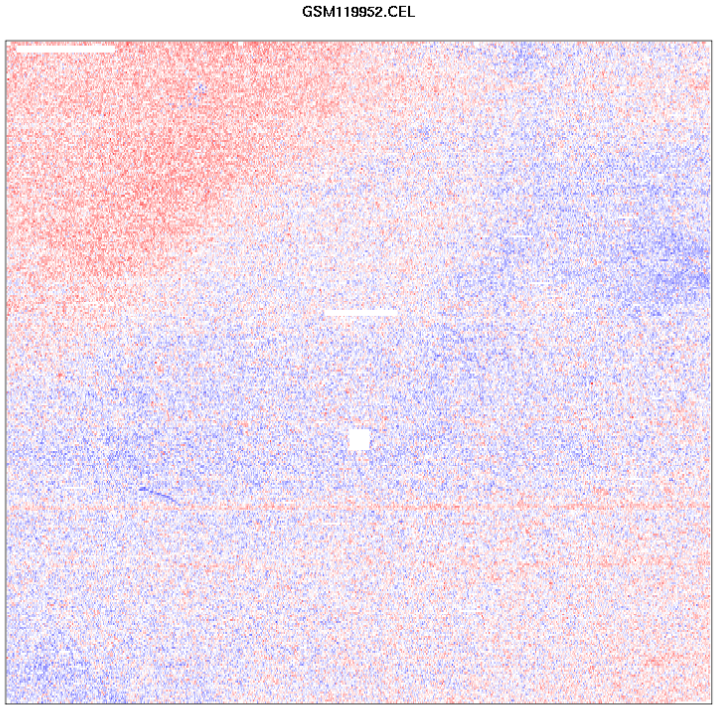


Figure 3.7: **An example of an extensive spatial defect on the surface of an Affymetrix microarray** - By design Affymetrix arrays are robust against small defects, scratches or dust. The `affyPLM` Bioconductor library allows the false colour visualization of the surface of the array. A large region of inhomogeneity, such as the region shown in red in the top left hand corner of the array, represents an artefact which renders the array unusable.

the majority of genes are not differentially expressed, therefore the boxplots should be centered around $M = 0$ and have small spread[1]. The NUSE approach, instead, uses an estimate of the standard error of each gene on each array obtained from the `affyPLM` library previously described. To account for the fact that variability differs considerably between genes, the standard error is standardized, so that the median value across arrays is 1 for each gene. The NUSE values are calculated using Eq. 3.1.

$$\text{NUSE}(\hat{\theta}_{gi}) = \frac{\text{SE}(\hat{\theta}_{gi})}{\text{med}_i(\text{SE}(\hat{\theta}_{gi}))} \tag{3.1}$$

Low quality arrays have boxplots that are significantly shifted from 1 or that have a large spread (*i.e.*, a larger IQR).

Since our analysis involves a large number of arrays, we automated the computation of the NUSE and RLE statistics. For each data set we performed the following steps:

1. We computed the median value and the IQR of both the NUSE and the RLE statistics for each array. This gave four values for each chip: M_{NUSE}, M_{RLE}, IQR_{NUSE} and IQR_{RLE}.

2. We compared each of these values with a corresponding cutoff value (see Table 3.13). If any value exceeded the cutoff, the chip was tagged as "dubious".

3. We then generated the distributions of M_{NUSE}, M_{RLE}, IQR_{NUSE} and IQR_{RLE} across the arrays and calculated its IQR. This gave four values: $IQR_{M_{NUSE}}$, $IQR_{M_{RLE}}$, $IQR_{IQR_{NUSE}}$ and $IQR_{IQR_{RLE}}$.

4. If any of these IQR values were greater than $q_3 + 1.5IQR$ or less than $q_1 - 1.5IQR$, where q_1 and q_3 are the first and the third quartile of the distribution, we considered such values as outliers and flagged the corresponding array as "rejected"[2].

[1] Henceforth, we will use the interquartile range (IQR) as a measure of spread, which is defined as the difference between the third and the first quartile of the values under consideration.

[2] This definition of an outlier is the one adopted by most statistical software when plotting boxplots.

M_{NUSE}	M_{RLE}	IQR_{NUSE}	IQR_{RLE}
1.10	0.2	0.10	1.0

Table 3.13: The chosen cutoff values for the four quantities defined in step 1 of the quality control procedure.

5. We made diagnostic plots for both the dubious and the rejected arrays for a successive visual analysis.

The four cutoff values were selected using a trial and error approach, with a preference for overestimating the number of poor quality chips rather than failing to identify a compromised chip.

3.4.2 The chosen data sets

The data sets used in our screening are summarized in Table 3.14. These data sets were chosen according to the criteria outlined in Section 3.4. Additional details on the data sets are provided in the following sections.

Data Set	Year	Samples (filtered)	Reference	GEO acc.
Ivshina	2006	289 (242)	(Ivshina et al., 2006)	GSE4922
Pawitan	2005	159 (150)	(Pawitan et al., 2005)	GSE1456
Sotiriou	2006	189 (85)	(Sotiriou et al., 2006)	GSE2990
TRANSBIG	2007	198 (189)	(Desmedt et al., 2007)	GSE7390
Wang	2005	286 (286)	(Wang et al., 2005)	GSE2034

Table 3.14: The data sets used in the breast cancer microarray screening. The year of publication, number of samples before and after the filtering (in parenthesis), the relative publication and the GEO accession number are reported.

3.4.2.1 The Ivshina data set

The Ivshina data set is comprised of samples from two institutes, referred to as the Uppsala and Singapore cohorts (Ivshina et al., 2006). The data set retrieved from the GEO database (accession no. GSE4922) included data on 289 patients, 40 of which came from the Singapore cohort and 249 from the Uppsala cohort. RNA from these patients was used to hybridize Affymetrix HG-U133A and HG-U133B

microarrays, for a total of 578 arrays. We only analyzed the 289 arrays based on HG-U133A because the HG-U133B arrays were available for some but not all of the selected data sets. Samples from Singapore were also excluded, due to the lack of clinical information, leaving an initial sample size of 249 samples.

We performed the quality control procedure on the 249 samples and, as a result, 14 arrays were flagged as "dubious" and 3 as "rejected". After visual inspection, we removed 7 of these flagged arrays from the data set, and proceeded to analyze the remaining 242 samples. The arrays were then normalized using the RMA algorithm with default parameters. The same normalization procedure was used for all the other data sets.

The available clinical parameters for the Ivshina data set were tumour grade, event, disease free survival, ER status, lymph node (LN) status, p53 status, age and tumour size. The definition of "event" was rather broad: "any type of recurrence (local, regional or distant) or death from breast cancer".

We mapped 465 Affymetrix probe sets to 253 miRNA host genes, which were in turn associated with 286 pri-miRNAs, representing 40% of the 721 human pri-miRNAs contained in miRBase, release 14. In this, and in the following analyses, we filtered out probe sets with signals less than 150 (linear scale) in at least 10% of the samples. After this filtering step we had 278 probe sets. We used the `geneFilter` package for the filtering, and the `limma` package for the successive analysis. Both packages are part of the Bioconductor suite. Probe sets having a Benjamini-Hochberg adjusted p-value less than 0.05 were considered as statistically significant. We performed a series of comparisons to determine the number of significantly regulated probe sets between different clinical states. We found 146 significantly regulated probe sets between grade 1 (G1) and G3 (G3) tumours, 122 between grade 3 and grade 2 (G2), and 45 between G2 and G1 tumours (Table 3.15).

Table 3.15 suggests that G1 tumours are genetically more different from G3 tumours than from G2 tumours, since the number of regulated probe sets in the first comparison approximately three times larger than in the second comparison. The probe sets that were regulated in the various comparisons were largely overlapping.

Comparison (N. of observations)	Probe sets	Host genes	Pri-miRNAs
G3 (55) vs. G1 (66)	146	108	110
G3 (55) vs. G2 (121)	122	95	106
G2 (121) vs. G1 (66)	45	39	44

Table 3.15: Number of regulated probe sets, host genes and pri-miRNAs across the possible comparisons concerning the tumour grade in the Ivshina data set.

For example 106 probe sets were common to both the G3 vs. G1 list (146 probe sets) and the G3 vs. G2 (122 probe sets) list. We then determined the number of significantly regulated probe sets with respect to ER status, LN status, P53 status, and the occurrence of an event (Table 3.16).

Comparison (N. of observations)	Probe sets	Host genes	Pri-miRNAs
ER+ (204) vs. ER- (34)	127	99	111
LN+ (78) vs. LN- (155)	23	18	22
wild type p53 (182) vs. mutated p53 (58)	124	94	100
Event (85) vs. no event (157)	6	5	5

Table 3.16: Identification of significantly regulated probe sets in the Ivshina data set. The type of comparison and the number of significantly regulated probe sets, host genes and pri-miRNAs are reported.

Interestingly, the number of significantly regulated probe sets in the G3 vs. G1, ER+ vs. ER-, p53 wild type vs. p53 mutated comparisons were similar. Therefore we determined the extent of overlap between these lists of probe sets.

From Table 3.17 it is quite clear that a large overlap exists, presumably due to the high degree of correlation between these clinical parameters. Some caution is therefore necessary in the interpretation of these data, since we cannot confidently distinguish the source of the regulation of the probe sets. To separate the individual contributions of the different clinical parameters, we re-analyzed the data stratifying

Comparison	Common probe sets
(G3 vs. G1) $\cap$ (ER+ vs. ER-)	98
(G3 vs. G1) $\cap$ (p53 wt vs. p53 mut)	114
(ER+ vs. ER-) $\cap$ (p53 wt vs. p53 mut)	97

Table 3.17: Number of overlapping probe sets in the G3 vs. G1, ER+ vs. ER- and p53 mutated vs. p53 wild type lists from the Ivshina data set. The $\cap$ sign indicates intersection.

with respect to the ER status (Tables 3.18 and 3.19).

Considering samples restricted for ER status, we observed very few regulated probe sets in the ER- samples, which can be easily explained by the small sample size. There were in fact only 34 ER- samples compared with 204 ER+ samples (the ER status was missing in 4 samples). We also found that 93% of the significantly regulated probe sets in the G3 vs. G1 comparison, were also included in the unrestricted data set list. Similarly all probe sets in the G3 vs. G2 list in the ER+ data set were also included in the corresponding unrestricted list. We performed a similar analysis stratifying with respect to the p53 status. No significant probe sets were found in the mutated p53 subset for any of the above comparisons, while 90% of the significant probe sets in the G3 vs. G1, and 88% of those in the G3 vs. G2 comparisons, restricted to the wild type p53 restricted set, were contained in the corresponding unrestricted list. These results suggest that tumour grade is the main cause of probe set regulation.

3.4.2.2 The Pawitan data set

The Pawitan data set consists of 159 breast cancer patients operated at the Karolinska Hospital from January 1994 to December 1996 (Pawitan et al., 2005). The clinical data includes information on: tumour grade, occurrence and time until breast cancer relapse, death from any cause, time until death, and death specifically due to breast cancer.

ER Positive samples			
Comparison (N. of observations)	Probe sets	Host genes	Pri-miRNAs
G3 (34) vs. G1 (61)	122	96	99
G3 (34) vs. G2 (109)	73	60	68
G2 (102) vs. G1 (61)	0	0	0
Event (72) vs. no event (132)	6	6	6
LN+ (66) vs. LN- (132)	0	0	0
p53 wt (163) vs. p53 mutated (39)	86	67	77

Table 3.18: Identification of significantly regulated probe sets in the subset of ER+ samples from the Ivshina data set. Thy type of comparison and the number of significantly regulated probe sets, host genes and pri-miRNAs are reported.

ER Negative samples			
Comparison (N. of observations)	Probe sets	Host genes	Pri-miRNAs
G3 (21) vs. G1 (2)	3	2	2
G3 (21) vs. G2 (11)	0	0	0
G2 (11) vs. G1 (2)	0	0	0
Event (12) vs. no event (22)	0	0	0
LN+ (12) vs. LN- (20)	0	0	0
p53 wt (15) vs. p53 mutated (19)	0	0	0

Table 3.19: Identification of significantly regulated probe sets in the subset of ER-samples from the Ivshina data set. Thy type of comparison and the number of significantly regulated probe sets, host genes and pri-miRNAs are reported

Comparison (N. of observations)	Probe sets	Host gene	Pri-miRNAs
G3 (58) vs. G1 (27)	69	53	63
G3 (58) vs. G2 (54)	87	66	74
G2 (54) vs. G1 (27)	0	0	0
Relapse (38) vs. no relapse (112)	56	49	56
Death from any reason (36) vs. no death (114)	20	17	18
Death from breast cancer (27) vs. no death (123)	43	38	41

Table 3.20: Identification of significantly regulated probe sets in the Pawitan data set. The type of comparison and the number of significantly regulated probe sets, host genes and pri-miRNAs are reported.

We applied our quality control procedure to the data set: 10 arrays were flagged as "dubious" and one as "rejected". Following the visual inspection of these arrays we excluded 9 arrays from our analysis. The remaining 150 samples, were then normalized using the RMA method and filtered to remove low signal probe sets leaving 234 probe sets. We then performed a series of comparisons to determine the number of significantly regulated probe sets between different clinical parameters (Table 3.20).

We found more regulated probe sets in the G3 vs. G2 comparison (87) than in the G3 vs. G1 comparison (69, see Table A.4 off the appendix for the complete list). This finding can be explained by the fact that the G2 group is twice as large as the G1 group. Thus the statistical power of the G3 vs. G2. comparison is greater than that of the G3 vs. G1 comparison. Moreover, we identified 55 *i.e.*, 82% of the G3 vs. G1 list, that were common to both lists. We also identified probe sets that were

Origin (label)	Samples	Treatment
Uppsala (KIT)	24	Yes
Oxford (OXFT)	40	Yes
Uppsala (KIU)	64	No
Oxford (OXFU)	61	No

Table 3.21: Partition of the samples in the Sotiriou data set with respect to the institute of the origin and the treatment.

significantly associated with relapse (see Table A.5) and death (see Tables A.6 and A.7); however, the fold change in expression was low for these probe sets.

3.4.2.3 The Sotiriou data set

The Sotiriou data set (GEO database accession GSE2990) contains information on 189 patients with primary operable invasive breast cancer. The frozen tumour specimens were obtained from two institutes: the John Radcliffe Hospital (Oxford, UK) and the Uppsala University Hospital (Uppsala, Sweden). RNA samples from Oxford were processed at the Jules Bordet Institute in Brussels, Belgium (Sotiriou et al., 2006). For the Uppsala samples, RNA was extracted at the Karolinska Institute and processed at the Genome Institute of Singapore. Some of the patients were treated with tamoxifen while others were not. Table 3.21 shows the partition of the samples with respect to the institute of origin and treatment. We performed the quality control procedure on the Sotiriou data set, which led to the exclusion of 5 arrays, leaving 184 for the successive analysis.

The fact that the samples were from different institutes and had undergone different manipulations was a reason of concern. We plotted the boxplots of the raw signal of the 184 arrays using different colours to indicate the institute/treatment combinations (Fig. 3.8). A strong batch effect was clearly visible: the intensity distributions in the OXFT group of arrays were significantly shifted to the low signal end compared with the other three groups. Normalization, at least to some extent, should correct this kind of problem, so we proceeded to check whether the batch effect was relevant even after normalizing the data set. We normalized the arrays using RMA and we removed all of low signal probe sets. Additionally, we removed

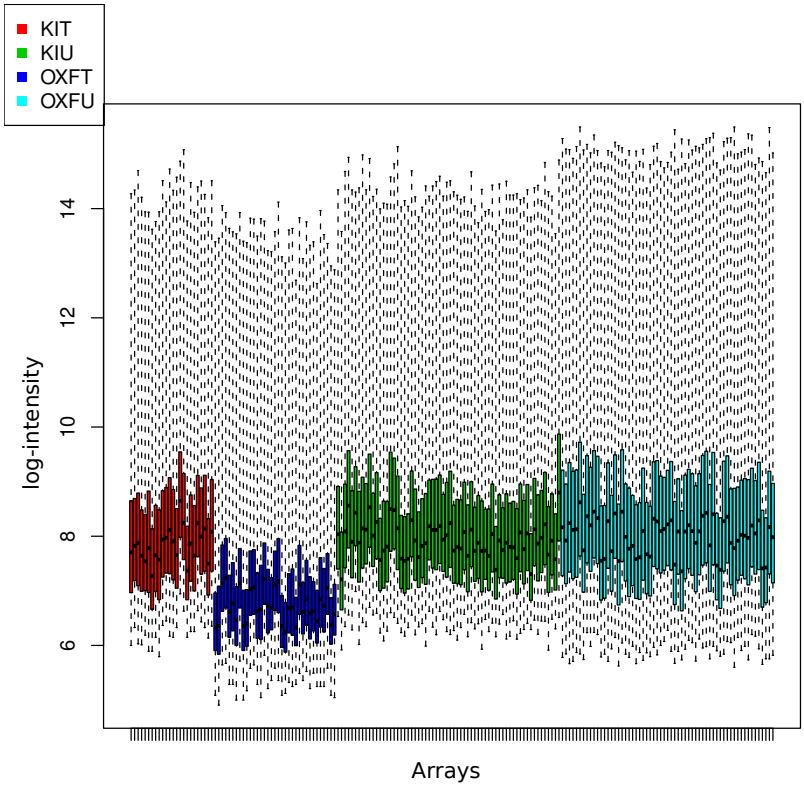


Figure 3.8: **Batch effect in the Sotiriou data set.** - Boxplot showing the log-intensity of each array at the probe level (i.e.,, before normalizing). Different colours indicate the four origin/treatment possibilities. A clear batch effect is visible, in that the arrays in the OXFT group have systematically a lower signal compared with the other groups.

all probe sets that did not have an IQR of the log-scale expression level greater than 0.8. The purpose of this second filtering step was to remove the "flat" probe sets, thus keeping only those that showed some variability. This filtering, although not particularly stringent, left only 1677 of the initial 22,283 probe sets. We then performed a probe-set level ANOVA adjusting the p-values with the Benjamini-Hochberg method, and found that 1256 of the 1677 probe sets (i.e., 75%) were significantly associated with the institute of origin using a cutoff of 5%. We selected the 100 most significant probe sets and performed a hierarchical clustering to visually assess differences between the groups (Fig. 3.9). The difference between the Oxford and the Uppsala data sets was still present: the Uppsala and the Oxford samples form two perfectly separated groups. Within these two clusters we observed that the OXFT and the OXFU groups were well separated, while there was no marked difference between the KIT and the KIU samples. We, therefore, concluded that it was inappropriate to treat the data as a single data set. Thus, we analyzed only the 85 arrays from Uppsala group that had passed the initial quality control.

After restricting the data set to the 465 probe sets associated with miRNA host genes, we filtered them to remove probe sets having weak intensities, leaving 280 probe sets. We then compared the expression of these probe sets according to different clinical parameters as summarized in Table 3.22. We found 101 probe sets significantly regulated in G3 vs. G1, 13 in G3 vs. G2 and 42 in ER positive vs. ER-. Again we observed a large overlap between the lists of significantly regulated probe sets: the 13 probe sets regulated in G3 vs. G2 are included in the 101 regulated probe sets in G3 vs. G1 while 32 of the 42 probe sets (*i.e.*, 76%) regulated in ER+ vs. ER- are contained in the G3 vs. G1 probe set list.

We then restricted our analysis to the ER+ samples (Table 3.23). The overlap between regulated probe sets in the G3 vs. G1 and the G3 vs. G2 comparisons was very high: 82 of the 88 (i.e., 93%) regulated probe sets in G3 vs. G1 restricted to the ER+ samples also appear in the same comparison performed on the whole data set (Table 3.22). Similarly, all the 13 probe sets in the unrestricted G3 vs. G2 list (Table 3.22) were contained in the corresponding list from the ER+ subset

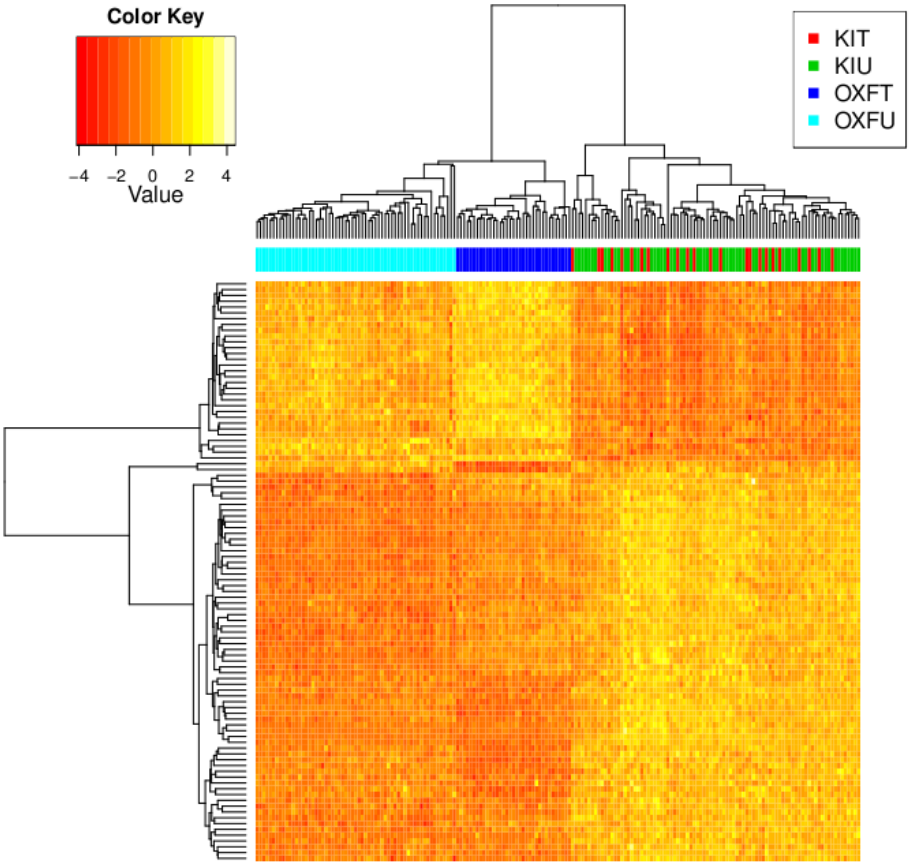


Figure 3.9: **Hierarchical clustering of the 184 arrays from the Sotiriou data set that passed the quality control procedure.** - Clustering was performed using the 100 probe sets that were most significantly associated with the institute of origin. We used the euclidean distance as a dissimilarity measure and the complete linkage method to group clusters. The colour key indicates the log-2 expression ratios.

Comparison (N. of observations)	Probe sets	Host genes	Pri-miRNAs
G3 (20) vs. G1 (36)	101	78	84
G3 (20) vs. G2 (26)	13	12	14
G2 (26) vs. G1 (36)	0	0	0
ER+ (74) vs. ER- (10)	42	42	37
LN+ (14) vs. LN- (67)	0	0	0
Recurrence (27) vs. no recurrence (56)	0	0	0
Distant metastasis (16) vs. no metastasis (66)	0	0	0

Table 3.22: Identification of significantly regulated probe sets in the Sotiriou-Uppsala data set. The type of comparison and the number of significantly regulated probe sets, host genes and pri-miRNAs are reported.

Comparison	Probe sets	Host genes	Pri-miRNAs
G3 (16) vs. G1 (34)	88	72	78
G3 (16) vs. G2 (21)	44	34	41

Table 3.23: Identification of significantly regulated probe sets in the Sotiriou-Uppsala data set restricted to the ER+ samples. The type of comparison and the number of significantly regulated probe sets, host genes and pri-miRNAs are reported.

(Table 3.23). The fact that only 10 samples were ER- justifies the fact that we could not find any significantly regulated probe set for any comparison in that subset.

We also stratified patients with respect to the tamoxifen treatment. Table 3.24 shows the distribution of tumour grade and ER status between the treated and untreated groups. We then identified significantly regulated probe sets between G3 vs. G1, G3 vs. G2 and ER+ vs. ER- samples for the two groups (Table 3.25) The regulated probe sets in the untreated group largely overlapped with the regulated probe sets in the corresponding comparisons performed on the complete data set (Table 3.22). In fact 90% of the G3 vs. G1 regulated probe sets the untreated group were also identified in the same comparison performed on the complete data set. There were more regulated probe sets (33) for the G3 vs. G2 comparison in the untreated group than in the complete data set (13 probe sets). This was

Tamoxifen treatment	G1	G2	G3	ER+	ER-
No	26	16	10	10	51
Yes	10	0	10	0	23

Table 3.24: Distribution of patients according to tumour grade, ER status and tamoxifen treatment in the Sotiriou Uppsala data set.

	Tamoxifen untreated			Tamoxifen treated		
Comparison	probe sets	host genes	pri-miRNAs	probe sets	host genes	pri-miRNAs
G3 vs. G1	86	66	73	5	5	5
G3 vs. G2	33	27	33	0	0	0
ER+ vs. ER-	42	33	37	0	0	0

Table 3.25: Identification of significantly regulated probe sets in the Sotiriou-Uppsala data set stratified by treatment. The type of comparison and the number of significantly regulated probe set, host genes and pri-miRNAs are reported for the tamoxifen untreated and treated groups..

presumably a consequence of the fact that the untreated group included the ER- tumours, i.e., the most aggressive tumours, whereas the G3 tumours in the tamoxifen group should, as a consequence of the treatment, be less aggressive. Moreover, 11 of the 13 probe sets of the latter list were also included in the former list. The complete list of significantly regulated probe sets for the G3 vs. G1 and the ER+ vs. ER- comparisons can be found in Tables A.8 and Table A.9 respectively.

3.4.2.4 The TRANSBIG data set

TRANSBIG is a consortium that was launched in 2004 to promote international collaboration in translational research (Buyse et al., 2006). According to the TRANSBIG website it comprises 40 institutions in 22 countries. The data set stored in the GEO data base with accession ID GSE7390 contains 198 arrays. A complete description of the data set can be found in (Buyse et al., 2006) and (Desmedt et al., 2007). The frozen tumour samples originated from five different institutes, but the pathological evaluation of ER status and tumour grade was performed in a single institute (European Institute of Oncology, Milan, Italy). Similarly, RNA was extracted in a single institute (Netherlands Cancer Institute, Amsterdam) with the exception of one subset of arrays (those from the Centre René Huguenin, Saint Cloud). The institutes of origin were: Institut Gustave Roussy, Villejuif, France (IGR); Karolinska Institute, Stockholm and Uppsala University Hospital, Uppsala, Sweden (KAR); Centre Rene Huguenin, Saint-Cloud, France (RH); Guys Hospital, London, United Kingdom (GUY); John Radcliffe Hospital, Oxford, United Kingdom (JRH). The distribution of the samples across the institutes is shown in Table 3.26.

GUY	IGR	JRH	KAR	RH
36	50	24	51	37

Table 3.26: Distribution of the patients across the five institutes composing the TRANSBIG data set.

We applied our quality control procedure, which led to the exclusion of 9 arrays from the data set, leaving 189 for the successive analysis. We also checked for the presence of batch effects as described for the Sotiriou data set by plotting an array-wise boxplot using the raw data (Fig. 3.10). No obvious batch effects were observed.

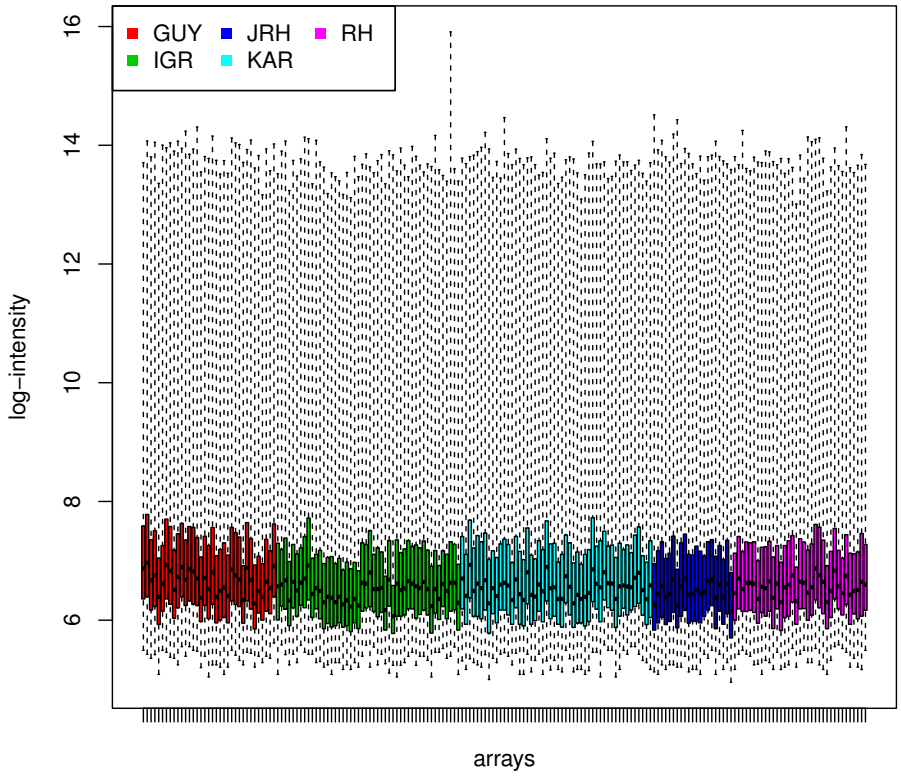


Figure 3.10: **Boxplot of the raw data of the TRANSBIG data set** - Different colours indicate the different hospitals that provided patient samples.

The available clinical data for the TRANSBIG data set includes information on tumour grade, ER status, death from any cause, occurrence of relapse, occurrence of distant metastasis. All patients were lymph node negative. Only 103 probe sets

Comparison (N. of observations)	Probe sets	Host genes	Pri-miRNAs
G3 (80) vs. G1 (27)	39	33	38
G3 (80) vs. G2 (80)	54	42	47
G2 (80) vs. G1 (27)	0	0	0
ER+ (128) vs. ER- (61)	52	43	47
Recurrence (89) vs. no recurrence (100)	0	0	0
Distant metastasis (62) vs. no metastasis (127)	0	0	0
Death from any cause (56) vs. no death (133)	0	0	0

Table 3.27: Identification of significantly regulated probe sets in the TRANSBIG data set. The type of comparison and the number of significantly regulated probe sets, host genes and pri-miRNAs are reported.

Comparison	Length 1^{st} list	Length 2^{nd} list	Common
(G3 vs. G1) $\cap$ (G3 vs. G2)	39	54	35
(G3 vs. G1) $\cap$ (ER+. vs. ER-.)	39	52	31
(G3 vs. G2) $\cap$ (ER+. vs. ER-.)	54	52	43

Table 3.28: Overlap of the lists of significantly regulated probe sets in the tumour grade and ER status comparisons in TRANSBIG data set.

of the initial 465 passed the probe set-signal intensity filtering. We determined the number of significantly regulated probe sets between different types of samples based on clinical parameters (Table 3.27)

As with the Sotiriou data set, we observed more regulated probe sets in G3 vs. G2 comparison than in G3 vs. G1 comparison, which is likely due to the larger number of G2 samples. Again, there was a large overlap between the two lists: 35 of the 39 probe sets (i.e., 90%) in the G3 vs. G1 probe set list, also appeared in the G3 vs. G2 list. There was also a large overlap between the ER status list and the tumour grade lists, as shown in Table 3.28.

We then considered ER+ and ER- patients separately. In our filtered data set there were 61 ER- and 128 ER+ samples. In the ER- subset we identified only one significantly regulated probe set in G3 vs. G1. In the ER+ subset there was a large overlap between the lists of regulated probe sets: 8 of the 12 probe sets (i.e 67%) in the G3 vs. G1 list were also present in the G3 vs. G2 list (Table 3.29). Moreover, 11 of the 12 significantly regulated probe sets in the G3 vs. G1 comparison in the ER+ subset (i.e., 92%) were included in the corresponding list of significantly regulated probe sets in the complete (Table 3.27) data set. This observation suggests that

Comparison	Probe sets	Genes	Pri-miRNAs
G3 (35) vs. G1 (26)	12	12	13
G3 (35) vs. G2 (65)	19	15	18

Table 3.29: Identification of significantly regulated probe sets in the ER+ subset of the TRANSBIG data set. The type of comparison and the number of significantly regulated probe sets, host genes and pri-miRNAs are reported.

stratifying with respect to the ER status does not add any additional information. The complete list of significantly regulated probe sets, the log-fold change, the for the G3 vs. G1 and the ER+ vs. ER- comparisons, can be found in Tables A.10 and A.11 respectively.

3.4.2.5 The Wang data set

The Wang data set consists of 286 samples from LN- patients. This data set was not initially included in our repository, due to some technical issues. First, the raw data were not available. Thus no quality control procedure was possible. Second, the clinical information was limited to the ER status, the LN status (negative for all samples), the occurrence of relapse and the time to relapse or to the last follow up, and the occurrence of brain relapse. Third, the arrays had been normalized with Affymetrix MAS5 algorithm, while we used RMA to normalize all the other data sets. Nevertheless, we decided to include this data set in our analysis for the following reasons: First, the large sample size should minimize any negative effects due to the presence of a small number of defective arrays. Second, the addition of this data set allowed us to have four data sets with information on ER status (not available in the Pawitan data set) and four with information on tumour grade (not available in the Wang set). Third, we re-normalized our data sets with the MAS5 method, repeated the analyses, and compared the results with those obtained with the RMA normalized data. We found that the lists of significantly regulated probe sets were almost identical with similar p-values and fold changes. For example we show the log fold change of significantly regulated the probe sets in the G3 vs. G1 comparison for the MAS and the RMA normalized data from the Wang data set were in good agreement (Fig. 3.11).

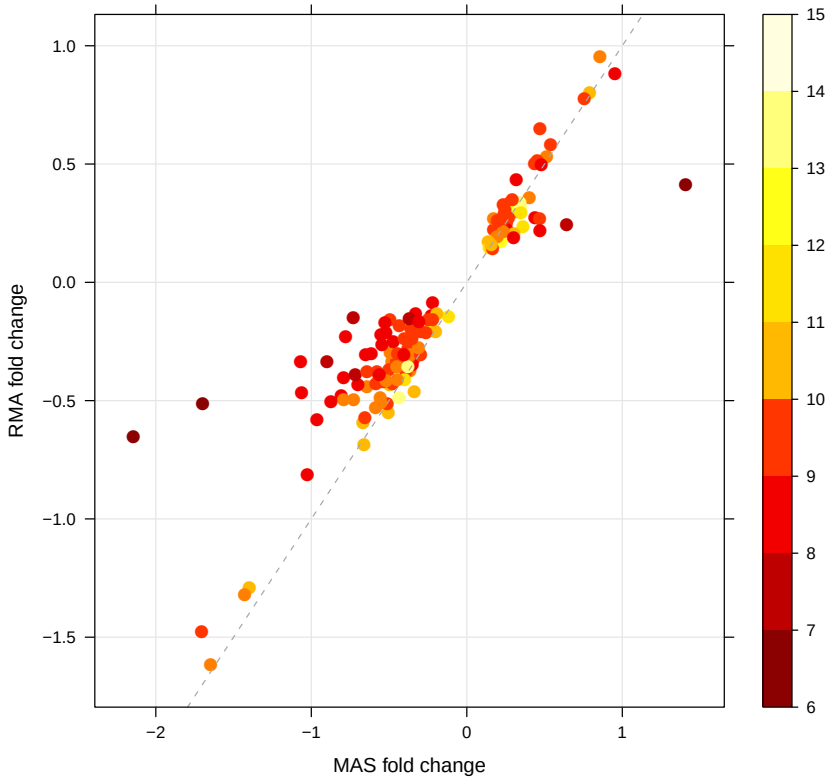


Figure 3.11: **Comparison of the results obtained with MAS and RMA normalized data from the Wang data set.** - Scatterplot of the log fold changes in the G3 vs. G1 comparison for the probe sets that were significantly regulated both in the MAS (x-axis) and the RMA (y-axis) normalized data. The colors represent the log-intensity of the probe sets. The points most deviating from perfect agreement (grey dashed line) have very low average intensities.

Comparison (N. of observations)	Probe sets	Gene symbols	miRNAs
ER+ (209) vs. ER- (77)	162	125	145
Relapse (107) vs. No relapse (179)	6	5	7
Brain relapse (10) vs. No relapse (276)	7	6	9

Table 3.30: Identification of significantly regulated probe sets in the Wang data set. The type of comparison and the number of significantly regulated probe sets, host genes and pri-miRNAs are reported

We filtered the probe sets for signal intensity: leaving 372 probe sets for our analysis. The number of significantly regulated probe sets for the different comparisons is shown in Table 3.30. The data set contains 209 ER+ and 77 ER- samples. We stratified the analysis with respect to the ER status, and found only 4 significantly regulated probe sets in Relapse vs. No relapse and 3 in Brain relapse vs. No relapse. The complete list of significantly regulated probe sets for all comparisons can be found in (Tables A.12, A.13, and A.14).

3.4.3 Permutation test on results

From the above analyses we made the following conclusions:

1. The majority of significantly regulated probe sets/miRNA host genes were found in the tumour grade (G3 vs. G1, G3 vs. G2) and ER status (ER+ vs. ER-) comparisons.

2. The G3 vs. G1 and G3 vs. G2 lists of regulated probe sets overlapped extensively.

3. Stratifying the tumour grade comparisons with respect to the ER status was not informative due to the small number of ER- samples.

4. In the comparisons of other clinical parameters (*e.g.*, LN status, event occurrence), we identified few significantly regulated probe sets, which displayed smaller fold changes in expression and/or higher p-values than the tumour grade/ER status comparisons.

Based on these conclusions, we focused our attention on the tumour grade and ER status comparisons. A summary of the G3 vs. G1 and ER+ vs. ER- comparison

	G3 vs. G1			ER+ vs. ER-		
Data set	probe sets	host genes	pri-miRNAs	probe sets	host genes	pri-miRNAs
Ivshina	146	108	110	127	109	111
Pawitan	69	53	63	NA	NA	NA
Sotiriou	101	78	84	42	42	37
TRANSBIG	39	33	38	52	43	47
Wang	NA	NA	NA	162	125	145

Table 3.31: Summary of the number of significantly regulated probe sets, host genes, and pri-miRNAs in each of the five data sets for the tumour grade and ER status analyses.

for the five data sets is shown in Table 3.31. Overall we observed that a large percentage (ranging from 29% to 55%) of the tested probe sets were significantly regulated in the tumour grade and ER status comparisons.

To determine whether a similar percentage would have resulted if we had considered genes that do not contain miRNAs, we performed the following simulation on each data set:

1. We removed all miRNA-associated probe sets from the data set.

2. We randomly selected n probe sets where n is the number of miRNA-associated probe sets (465 in our case).

3. We filtered this data set with respect to probe set intensity (intensity > 150 in at least 10% of the samples.

4. We determined the number of probe sets that were significantly regulated between G3 vs. G1 and ER+ vs. ER- tumours.

5. Repeated steps 2–4, 999 times

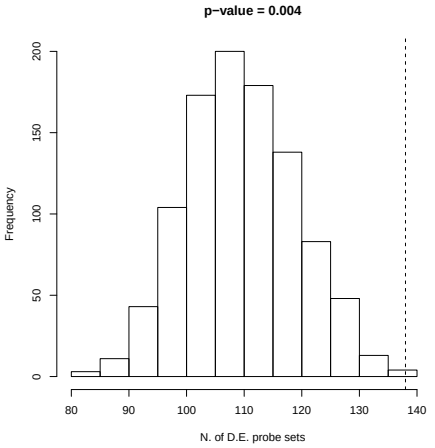
For each data set, and for each comparison, we thus had an empirical distribution of the expected number of significantly regulated probe sets. Each distribution contained 1000 values, 999 from the simulations and one deriving from our analysis. From these distributions, we then determined the number of simulations that had a regulated probe set list as large or larger than the original list. The empirical p-value is given by the ratio of this number to the total number of simulations. For

the G3 vs. G1 comparison, we observed that the lists identified through our analyses were significantly larger than what we should expect according to the distribution (Fig. 3.12). Thus the difference in expression of miRNA host genes between low and high grade tumours is greater than for non-host genes. This results suggests that intronic miRNAs might be involved in cancer progression and the switch to a more aggressive phenotype (e.g. enhanced proliferation, invasion and resistance to apoptosis)

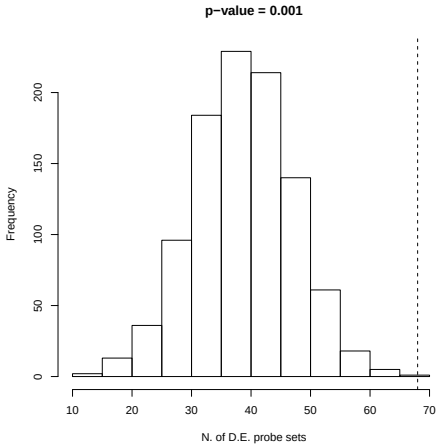
For ER status comparison the the situation is less clear. For three of the data sets, Ivshina, Sotiriou and Wang, the lists identified through our analyses were significantly larger than the majority of simulated lists (Fig. 3.13) In contrast, for the TRANSBIG data set, the empirical p-value was non-significant. On the basis of this simulation, we considered as prime candidates for onco-miRNAs those associated with the probe sets significantly regulated between G1 and G3 tumours.

3.4.4 Identification of candidate miRNAs involved in breast cancer

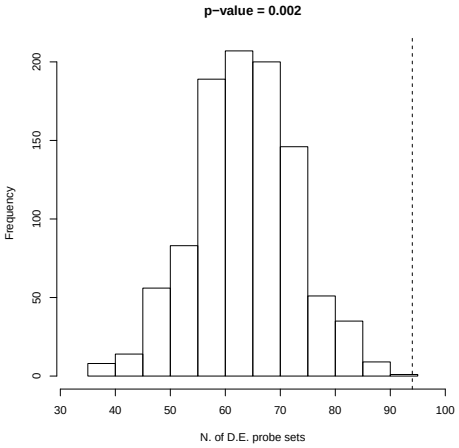
We observed that many probe sets were significantly regulated, with similar fold changes in expression, in both the tumour grade and ER status comparison in all data sets. These probe sets, therefore, represented a good starting point for the identification of novel onco-miRNAs involved in breast cancer (see Tables A.17 and A.18 of the Appendix for the complete list). From the list of commonly regulated probe sets/miRNAs (Table A.17), we selected those that were regulated by at least 1.5 fold in either direction in all data sets. This resulted in a list of 11 probe sets specific for 8 genes containing a total of 11 miRNAs, 6 of which were downregulated and 5 up-regulated in G3 tumours (Table A.19). The up-regulated miRNAs were already known to be involved in cancer (see Sec. 1.2), while few reports had been published on the downregulated miRNAs. Thus we focused our search for novel onco-miRNAs on these downregulated miRNAs. The up-regulated miRNAs belong to the *miR-25-93-106b*, and *miR-15b/16-2* clusters, contained in the MCM7 and SMC4 genes, respectively. The *miR-25-93-106b* cluster has been implicated in hepatocellular carcinoma (Li et al., 2009), gastric cancer (Petrocca et al., 2008b)



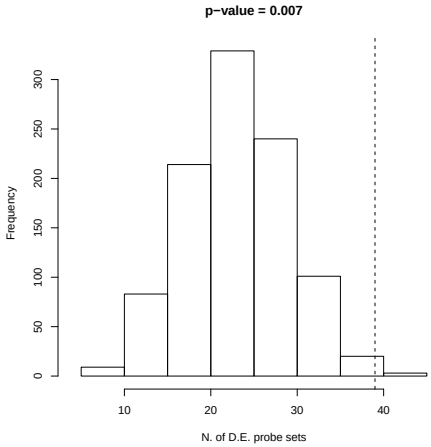
(a) *The Ivshina data set.*



(b) *The Pawitan data set.*

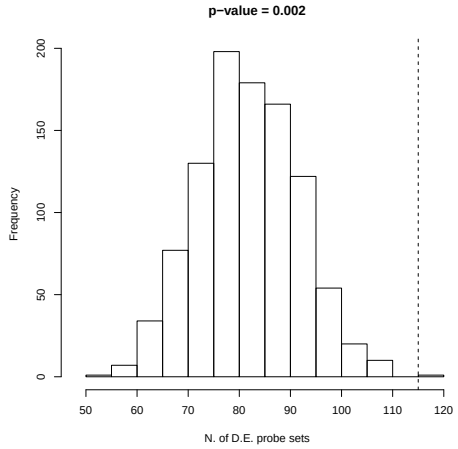


(c) *The Sotiriou data set.*

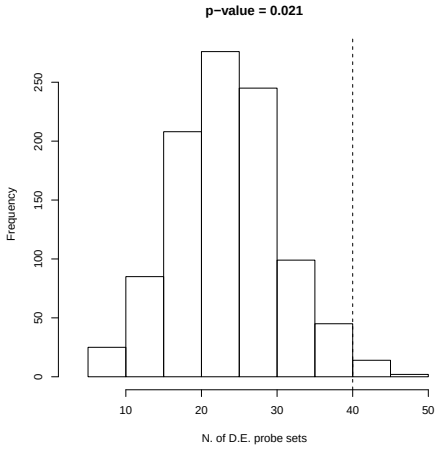


(d) *TRANSBIG data set.*

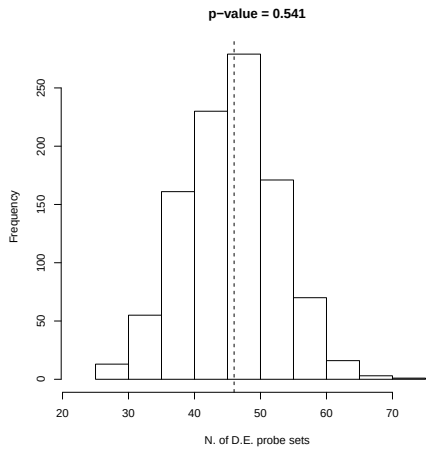
Figure 3.12: Results of the simulations for the G3 vs. G1 comparisons. Distribution histograms for the indicated data sets. The dashed lines indicate the positions within the distributions of the probe set lists determined through our analyses. The p-values indicate the fraction of lists with an equal or larger number of significantly regulated probe sets compared with the ones from our analyses.



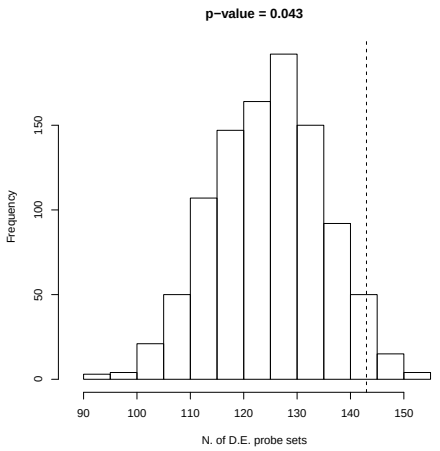
(a) *the Ivshina data set.*



(b) *The Sotiriou data set.*



(c) *TRANSBIG data set.*



(d) *Wang's data set.*

Figure 3.13: Results of the simulation for the ER status comparison on the indicated data sets. Figure as described in Fig 3.12.

and renal cell carcinoma (Slaby et al., 2010). This cluster also interferes with cell cycle arrest and apoptosis when over-expressed in gastrointestinal and other cancer cells, by modulating TGFβ signaling (Petrocca et al., 2008a). It has also recently been demonstrated that the *miR-25-93-106b* cluster targets the tumour suppressor PTEN and, simultaneously, cooperates with its host gene MCM7, to induce cellular transformation both in vitro and in vivo. Indeed, concomitant expression of the *miR-25-93-106b* cluster and MCM7 triggers prostatic intraepithelial neoplasia in transgenic mice (Poliseno et al., 2010). The *miR-15b/16-2* cluster has also been implicated in cancer, in particular malignant melanoma (Satzger et al., 2010). In addition it induces cell cycle progression by targeting cyclins in glioma cells (Xia et al., 2009), and apoptosis by targeting BCL2 in chronic lymphocytic leukemia (Cimmino et al., 2005). In contrast with the up-regulated miRNAs, there are few reports in the literature of an involvement of the down-regulated miRNAs (*i.e.*, *miR-218-1*, *miR-342*, *miR-483*, *miR-548f-2*, *miR-1245*, and *miR-1266*) in cancer. *Mir-218-1* has been shown to be downregulated in metastatic cancer together with its host gene, *SLIT3* (Tie et al., 2010). *MiR-342* has been associated with acute promyelocytic leukemia (De Marchis et al., 2009; Garzon et al., 2007) and colorectal cancer (Grady et al., 2008). Moreover, in breast cancer, the downregulation/expression of this miRNA correlates with ER status (Lowery et al., 2009) and tamoxifen resistance (Miller et al., 2008) but, to our knowledge, no direct correlation with tumor grade has been reported. For *miR-483*, 2 recent articles published after we performed our analysis, have linked this miRNA to cancer. In particular, *miR-483* has been described as a potential prognostic predictor in adrenocortical cancer (Soon et al., 2009) and the mature form of this miRNA has been shown to be over-expressed in Wilm's tumours and in a fraction of liver cancers (Veronese et al., 2010). Finally, to date, there are no publications describing the role *miR-548f-2*, *miR-1245*, and *miR-1266* either in cancer or in any other biological context.

3.4.5 Validation of results

We first examined the expression of the downregulated miRNAs in normal and tumour breast cell lines by RT-PCR. This screening allowed us to verify the expression of these miRNAs in mammary cells, as well as to optimize the RT-PCR protocol, for the subsequent analysis on tumour specimens. Two normal (HMEC and MCF10A), and four tumour (MB-231, MB-361, MCF7, BT474) cell lines were screened. For *miR-218*, *miR-342* and *miR-483* we examined the expression of both the 3p and 5p[1] versions. MiRNAs with a C_T values equal to 40 cycles were considered as "not amplified" (i.e., undetectable). The results of the RT-PCR are shown in Fig. 3.14: the top panel shows the raw C_T values of the target miRNAs and of the corresponding housekeeping RNA, while the bottom panel shows the absolute ΔC_Ts values calculated with respect to the housekeeping RNA. In this latter plot, the higher the ΔC_T value, the *less* expressed the miRNA. From these results we made the following observations 1) *miR-342-3p* and *miR-342-5p* were both strongly expressed in all the cell lines; 2) *miR-483-3p* and, in particular, *miR-483-5p* were expressed at low levels in the majority of cell lines, the exceptions being MB-361 and MCF7; 3) *miR-1266* is very weakly expressed; 4) *miR-218* and *miR-218** were very weakly expressed or undetectable in all cell lines; 5) the results for *miR-1245* were untrustworthy due to technical problems (i.e., the amplification curves were not sigmoidal and some technical replicates failed); 6) *miR-548f-2* was never detected.

We next examined the expression of the miRNAs that were detected in the cell line screening (*i.e.*, *miR-218/218**, *miR-342-3p/5p*, *miR-483-3p/5p* and *miR-1266*) in a collection of thirty-six formalin-fixed paraffin-embedded (FFPE) tumour samples provided by the European Institute of Oncology, Milan, Italy. The samples consisted of sixteen G1 and twenty G3 tumours. To reduce the risk of stromal contamination we used a needle to extract cylinders of tumour tissue instead of slicing sections of the samples. Two cylinders were collected from each sample, and RNA was extracted from these cylinders for the successive RT-PCR analysis using a specific kit for total RNA extraction from paraffin embedded tissues (see Sec. 2.3.).

[1] An asterisk, as in the case of *miR-218**, indicates the strand opposite to the predominant one.

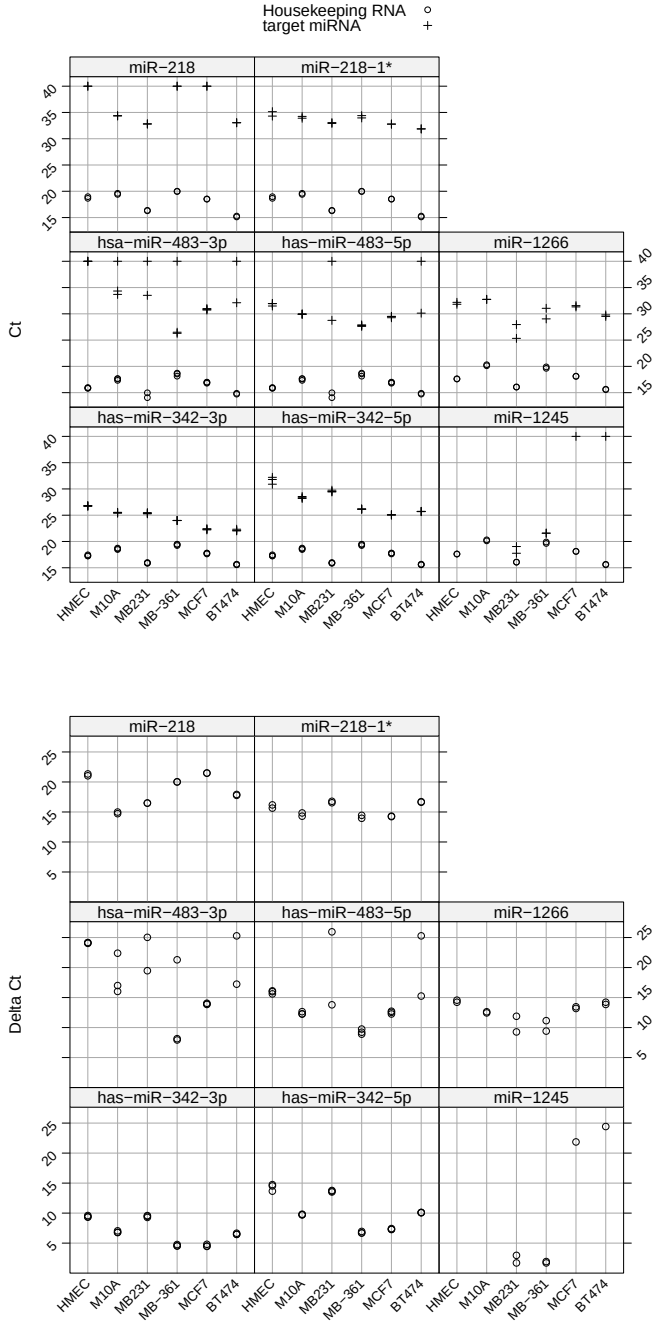


Figure 3.14: RT-PCR analysis of the expression of selected miRNAs in breast cell lines. **Top**: raw C_T values of the selected miRNAs (target miRNA) and of the corresponding housekeeping RNA (U6 snRNA). A C_T value = 40 indicates no amplification. **Bottom**: absolute value of the ΔC_T of the selected miRNAs with respect to the housekeeping RNA. The higher the ΔC_T value, the *lower* the expression of the miRNA. The multiple points on each column refer to the different experimental replicates.

miRNA ID	p-value
miR-1266	0.0028
miR-218	0.058
miR-218*	0.1
miR-342-3p	9.1×10^{-06}
miR-342-5p	0.0019
miR-483-3p	0.0019
miR-483-5p	0.0022

Table 3.32: Results of the Mann-Whitney test on the ΔC_Ts with respect to the tumour grade.

MiRNAs were then selectively retro-transcribed into cDNA. We obtained a good yield from this procedure ($\sim 2\mu g$ of RNA on average).

We observed that *miR-342-3p* and *miR-342-5p* were strongly expressed in breast tumours, particularly in the G1 tumours (Fig. 3.15). In contrast *miR-218*, *miR-218** and *miR-1266* were not expressed in any of the samples (Fig. 3.15). For *miR-483-3p* we detected a moderate level of expression in a fraction of G1 tumours, while its expression was lower in G3 tumours (Fig. 3.15). Finally, we discarded the results of *miR-1245* due to the unreliability of the assay, as evidenced in the cell line screening.

For *miR-218* and *miR-342-5p*, influential values (outliers) can clearly be seen from the boxplots (Fig. 3.15, top panel). Thus, we applied a non-parametric statistical tests, the Mann-Whitney test on each miRNA to compare the ΔC_T values of G3 vs. G1 tumours Table 3.32. We observed that the expression of the selected miRNAs, excluding *miR-218* and *miR-218**, significantly separated G1 from G3 patients, confirming the results from the breast cancer data sets analysis. Down-regulation of these miRNAs in high grade breast cancers indicates a possible role in tumour progression, and suggests they might act as tumour suppressors, a possibility that we are currently investigating.

We then performed a hierarchical clustering of the ΔC_T values of the thirty-six samples with respect to the seven selected miRNAs to assess their discriminating power. As before, the higher the ΔC_T value, the *less* expressed the miRNA. The resulting dendrogram separated G1 and G3 tumours, although the separation did not appear robust (Fig. 3.16). Next we considered only *miR-342-3p/5p* and *miR-*

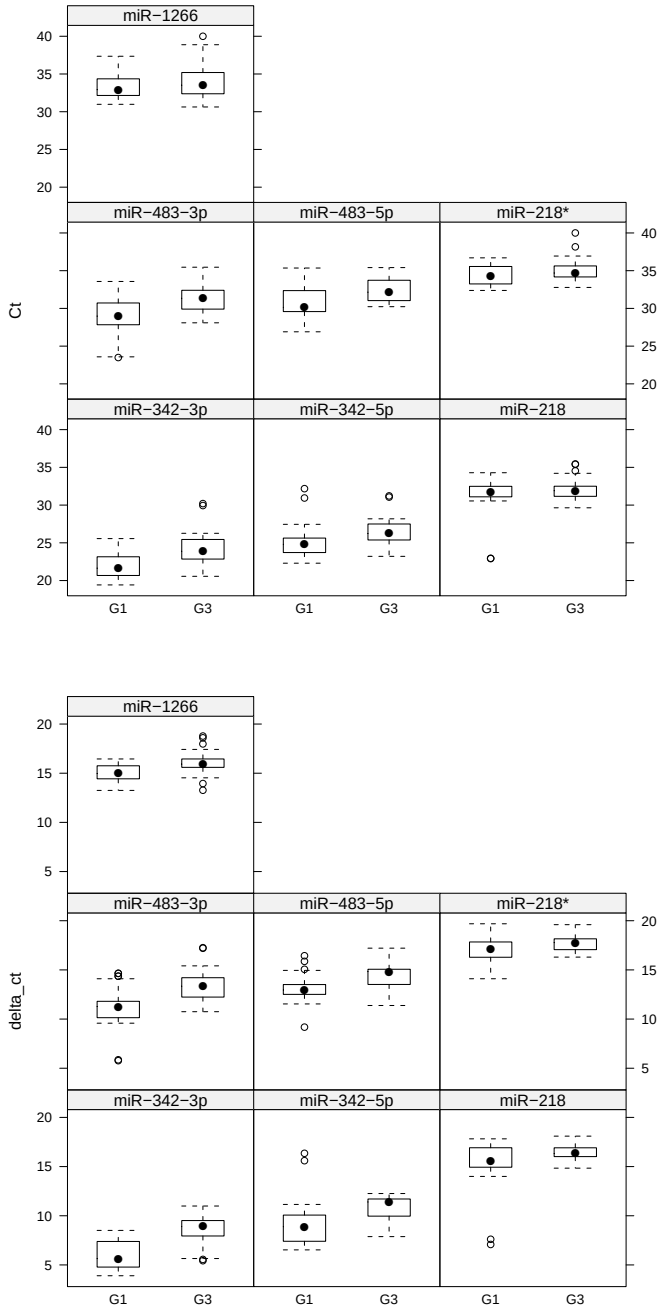


Figure 3.15: RT-PCR analysis of the expression of indicated miRNAs in breast cancer specimens stratified by tumour grade. **Top**: Boxplots of the raw C_T values of the selected miRNAs in G1 and G3 tumours. **Bottom**: Boxplots of the ΔC_T values of the selected miRNAs with respect to the housekeeping RNA, U5A snRNA. The RT-PCR of each tumour specimen was performed in duplicate.

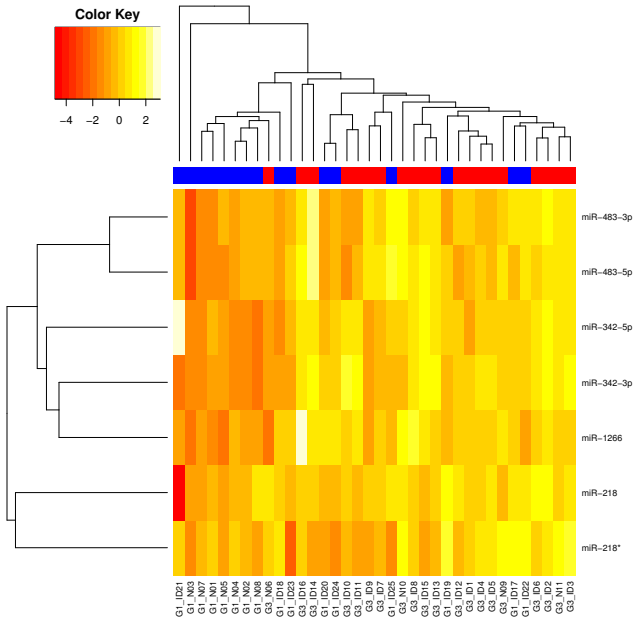


Figure 3.16: Hierarchical clustering of the ΔC_T values of thirty-six tumour samples (blue: G1, red: G3) with respect to all the seven selected miRNAs. The higher the ΔC_T value, the *less* expressed the miRNA. Expression values of each miRNA have been standardized by subtracting the mean and dividing by the standard deviation. The color key refers to the standardized ΔC_T values. Dissimilarity measure: euclidean distance. Linkage method: average linkage.

483-3p/5p and repeated the clustering procedure, and the resulting dendrogram showed a more robust separation of the G1 and G3 tumours. (Fig. 3.17).

We then examined the regulation of the selected miRNAs with respect to ER status (Fig. 3.18). We observed a marked downregulation *miR-342-3p* and *miR-342-5p* in ER+ compared with ER-, while little of no difference in expression was observed for the other miRNAs.

The downregulation of *miR-342-3p/5p* was statistically significant, as was the small downregulation of *miR-1266*, although much less so (Table 3.33).

3.4.6 Reclassification of G2 tumours

Finally we investigated whether the probe sets mapping the *miR-342* and *miR-483* host genes, *EVL* and *INS-IGF2* could divide G2 tumours,in the breast cancer data sets, for prognosis. Classification of G2 tumours was achieved by hierarchical clustering. We excluded the Sotiriou data set because of the relatively small number (26)

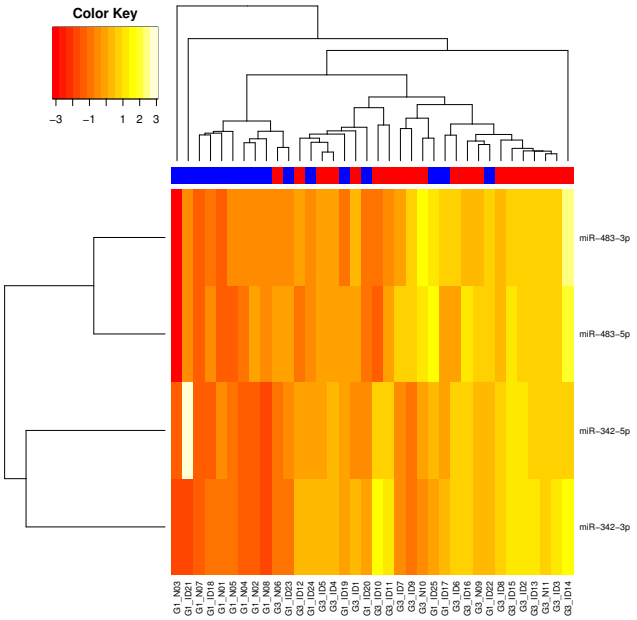
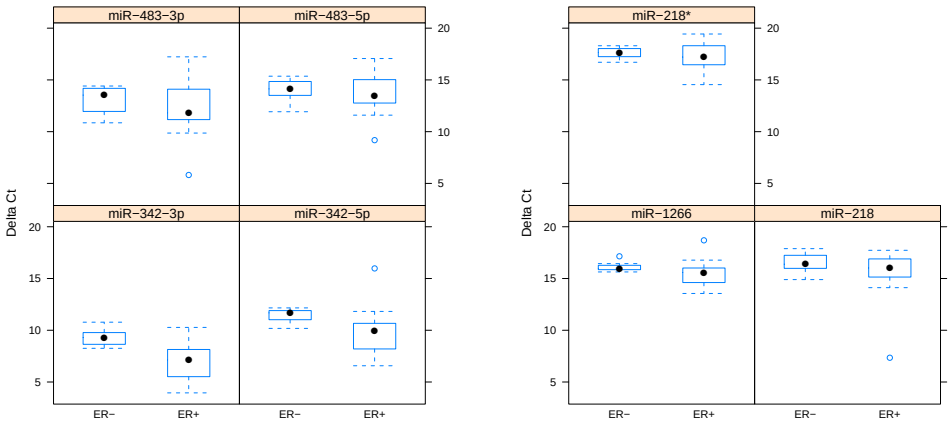


Figure 3.17: Hierarchical clustering of the ΔC_T values of the same samples, but restricted to *miR-342-3p/5p* and *miR-483-3p/5p*. As in Fig. 3.16, The higher the ΔC_T value, the *less* expressed the miRNA. Expression values of each miRNA have been standardized by subtracting the mean and dividing by the standard deviation. The color key refers to the standardized ΔC_T values.



(a) *ER status in miR-342 and miR-483.* (b) *ER status in miR-218 and miR-1266.*

Figure 3.18: **Regulation of the selected miRNAs with respect to ER status.** Boxplots of the ΔC_T values between ER+ and ER- in *miR-342-3p/5p*, *miR-483-3p/5p*, *miR-218/218** and *miR-1266*.

miRNA ID	p-value
miR-1266	0.0089
miR-218	0.21
miR-218*	0.32
miR-342-3p	6.63×10^{-06}
miR-342-5p	2.19×10^{-05}
miR-483-3p	0.14
miR-483-5p	0.64

Table 3.33: Results of the Mann-Whitney test on the ΔC_T values with respect to ER status.

of G2 patients. We also excluded the TRANSBIG data set because of an anomaly in the relationship between tumour grade and recurrence-free survival. In fact, a survival analysis on patients in the TRANSBIG data set after stratifying by tumour grade, revealed that G3 patients had a longer survival compared with G2 patients (Fig. 3.19). This result contrasts the literature (see for example Soerjomataram et al. (2008)) and with our results from the other data sets.

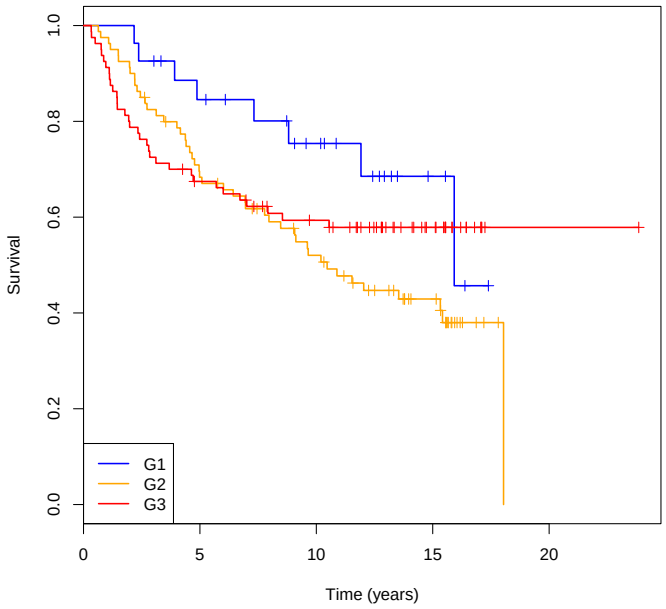


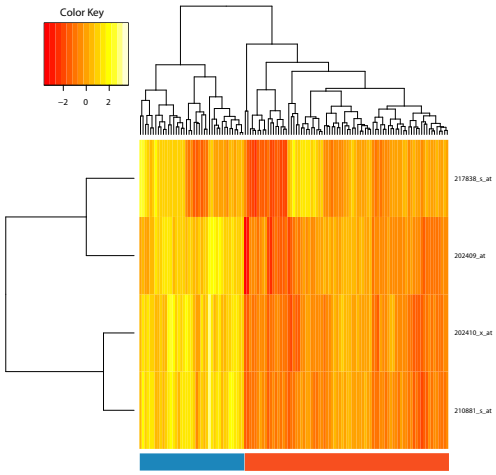
Figure 3.19: Kaplan Meier curves for the G1 (in blue), G2 (in orange) and G3 (in red) patients in TRANSBIG data set.

We, therefore considered only the Ivshina and the Pawitan data sets, which in-

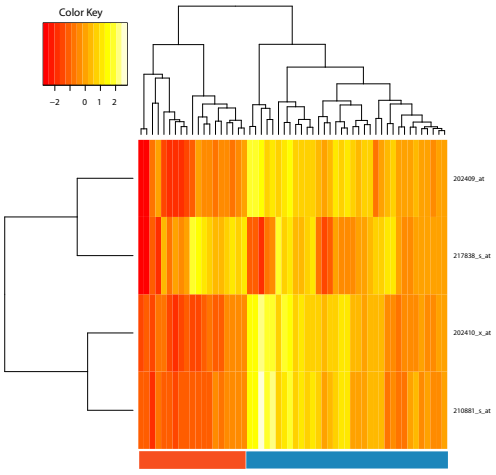
cluded, after the quality control procedure, 121 and 54 G2 samples respectively. We performed a hierarchical clustering with respect to the expression levels of four probe sets, three associated with *INS-IGF2* and one with *EVL* (Fig. 3.20, top panels). Patients were clearly separated into two groups in both data sets (Fig. 3.20, top panels). We then plotted Kaplan-Meier survival curves for the two data sets to compare recurrence-free survival in such groups, and performed a log-rank test to assess whether the observed differences were statistically significant (Fig. 3.20, lower panels). The p-values were significant for the Pawitan, but not for the Ivshina data set. We then repeated the same steps after removing the single probe set that mapped to the *miR-342* host gene, *EVL*. Again G2 patients were separated into two groups in both data sets, but this time we observed a statistically significant difference in survival in both the Pawitan and the Ivshina data sets. Thus the *miR-483* host gene, *INS-IGF2*, appears to discriminate high risk from low risk G2 patients. Since we observed a high correlation between the expression of *INS-IGF2* and *miR-483*, it is plausible that the latter might serve as a useful tool for the prognosis of breast cancer.

3.4.7 Over-expression of *miR-342-3p* and *miR-342-5p*

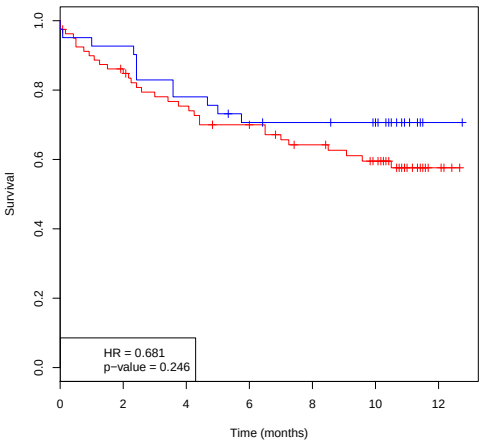
While validating the expression of *miR-342-3p* and *miR-342-5p* in a panel of commercial breast cancer cell lines, we observed that these mature miRNAs were expressed at low levels in the highly metastatic MDA-MB231 cell line (Fig. 3.14). We thus decided to investigate the biological effects of *miR-342-3p* and *miR-342-5p* by over-expressing them in MDA-MB231 cells. Over-expression was obtained by transfecting MDA-MB231 cells with synthetic miRNAs (see Methods). As a control, we transfected cells with a scrambled synthetic miRNA (CTR in Fig. 3.22). The concentration of *miR-342-3p* and *miR-342-5p* were increased by 860 and 12,000 fold respectively upon transfection with the synthetic miRNAs. We compared the effects on apoptosis of these transfections by means of fluorescence-activated cell sorting (FACS). We observed a marked increase in apoptosis in the cells transfected with *miR-342-5p* while we did not score any effect on the cells transfected with *miR-*



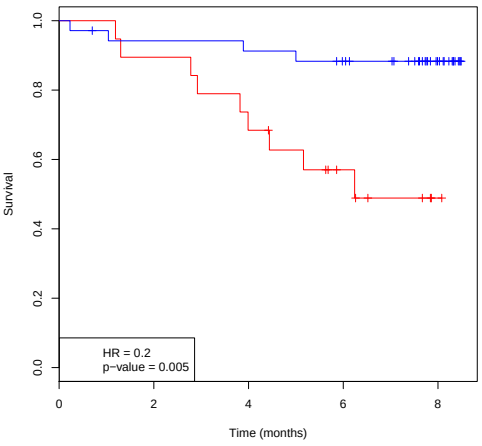
(a) *Hierarchical clustering on the Ivshina data.*



(b) *Hierarchical clustering on the Pawitan data.*

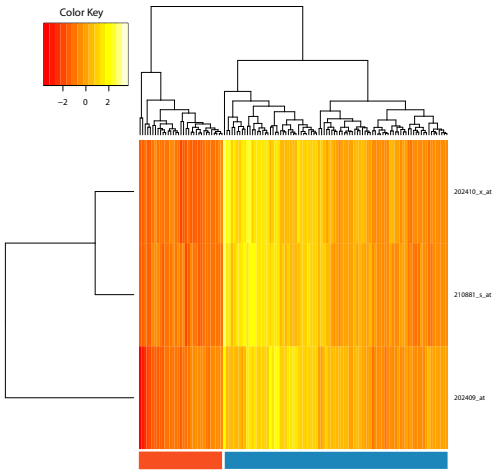


(c) *Kaplan Meier curves for the Ivshina data.*

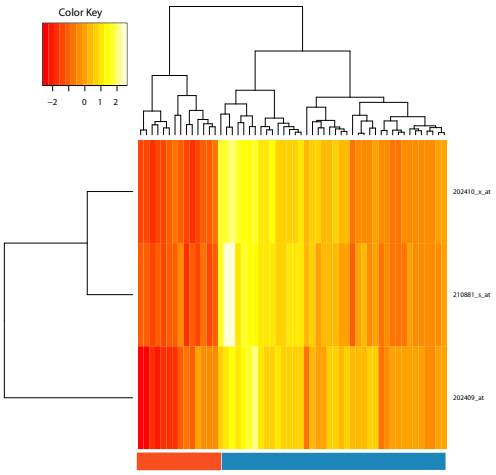


(d) *Kaplan Meier curves for the Pawitan data.*

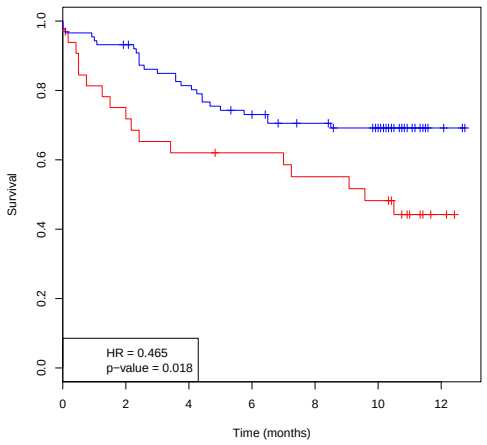
Figure 3.20: **Hierarchical clustering and Kaplan-Meier curves for the Ivshina and the Pawitan data after clustering with respect to *miR-342* and *miR-483*. Top:** G2 samples were clustered with respect to the expression levels of the probe sets matching the *miR-342-3p/5p* and *miR-483-3p/5p* host genes, *EVL* and *INS-IGF2*. The colour key refers to the standardized log expression levels as measured by Affymetrix. Here, the higher the value, the higher the log expression level. Hierarchical clustering identified two groups, indicated in red and blue. **Bottom:** survival curves of two groups identified by hierarchical clustering. The plots report also the hazard ratio (HR) and the log-rank test p-value. Results were significant for the Pawitan but not the Ivshina data set. The colour key refers to the standardized log fold changes measured by Affymetrix.



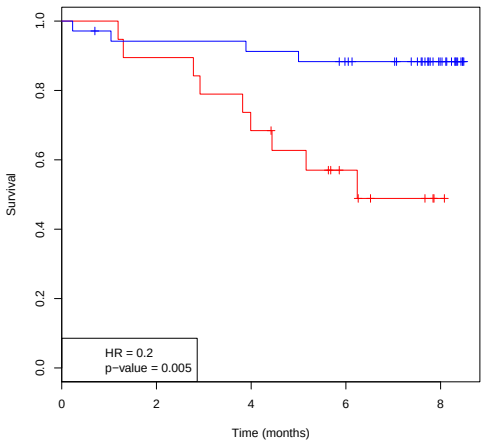
(a) *Hierarchical clustering on the Ivshina data.*



(b) *Hierarchical clustering on the Pawitan data.*



(c) *Kaplan Meier curves for the Ivshina data.*



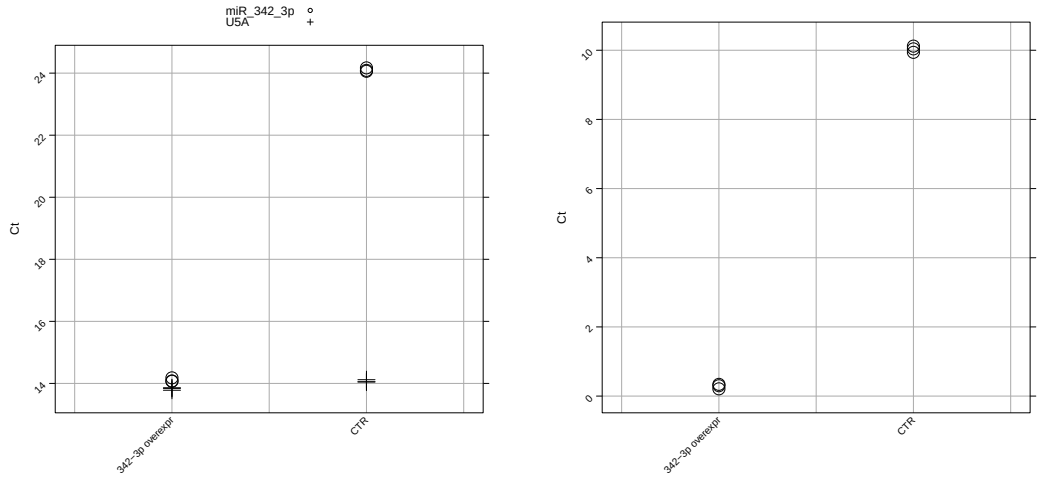
(d) *Kaplan Meier curves for the Pawitan data.*

Figure 3.21: **Hierarchical clustering and Kaplan-Meier curves for the Ivshina and the Pawitan data after clustering with respect to *miR-483* only.** Same as Fig. 3.20, only restricted to the three probe sets mapping to the *miR-483* host gene, *INS-IGF2*. The coulour key refers to the standardized log expression values as measured by Affymetrix. The higher the value, the higher the log expression value.

Control	*miR-342-3p*	*miR-342-5p*
26.1	24.1	48.8

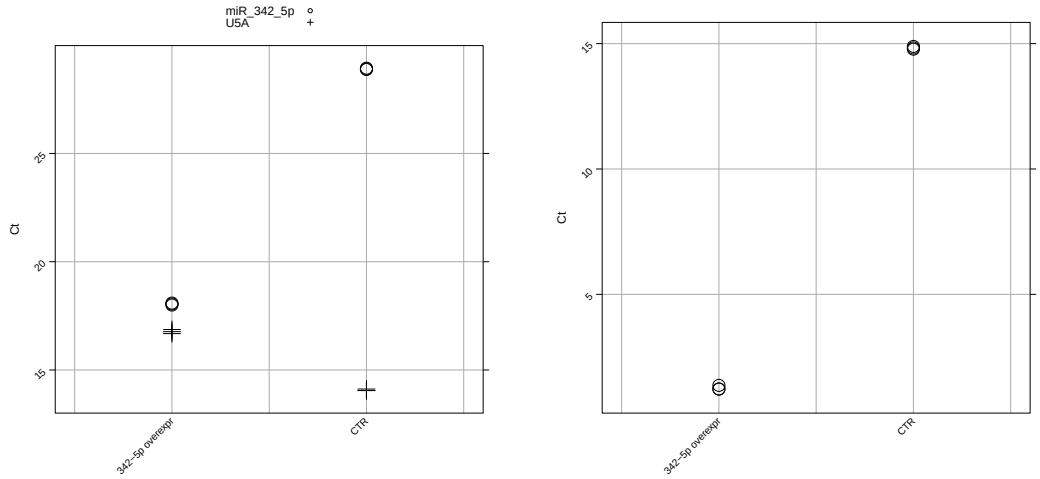
Table 3.34: Percentage of apoptotic cells in the control experiment and in the cells transfected with *miR-342-3p* and *miR-342-5p*.

342-3p (Fig. 3.23 and Table 3.34). This result suggests a role for *miR-342-5p* in cell viability. In line with these preliminary observations, Grady and co-workers have reported that over-expression of *miR-342* in the HT29 colon carcinoma cell line induces apoptosis (Grady et al., 2008).



(a) C_T *values of miR-342-3p*

(b) ΔC_T *values of miR-342-3p.*



(c) C_T *values of miR-342-5p.*

(d) ΔC_T *values of miR-342-5p.*

Figure 3.22: **Overexpression of *miR-342-3p* and *miR-342-5p***. Raw C_T values and ΔC_T values for *miR-324-3p* (top) and *miR-342-5p* (bottom) in cells where *miR-342-3p/5p* were over-expressed, and in cells transfected with a scrambled miRNA (CTR).

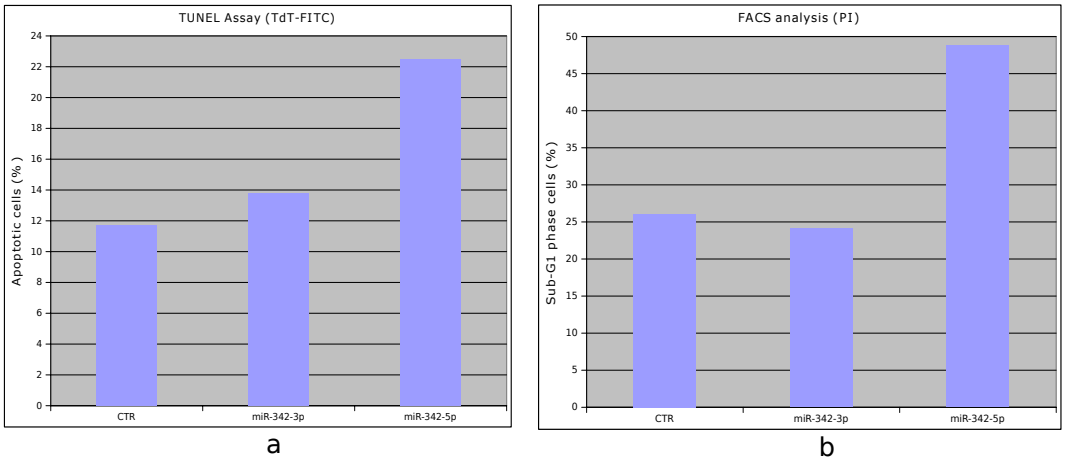


Figure 3.23: **FACS analysis of MDA-MB231 cells transfected with *miR-342-3p/5p***. **a**: TUNEL assay analysis. Y-axis, percentage of apoptotic cells relative to parental cells. X-axis, cells transfected with a negative control (CTR), cells transfected with *miR-342-3p* and with *miR-342-5p*. **b**: propidium iodide (PI) analysis of cell cycle distribution. Y-axis, percentage of cells in Sub-G1 phase (apoptotic) relative to parental cells. X-axis, cells transfected with a negative control (CTR), cells transfected with *miR-342-3p* and with *miR-342-5p*.

4

Discussion

4.1 The E1A and the SV-40 experiments

Intronic miRNAs represent a large fraction of the total number of miRNAs in mammals. As new miRNAs are discovered, the relevance of intragenic miRNAs is becoming more apparent. In miRBase 13.0, for example, the fraction of human intronic miRNAs was 58%; this proportion rose to 60% in release 14.0 and to 65% in release 15.0 (April 2010). MiRNA genes are frequently located in fragile regions of the genome as illustrated in Fig. 4.1 (Calin et al., 2004; Croce, 2009), and some intragenic miRNAs have been found to be associated with cancer, such as the *mir-17-92* and *mir-15b/16-2* miRNA clusters, often through complex pathways and feedback loops (see Sec. 1.2). However, despite the large number of publications describing the involvement of intragenic miRNAs in tumorigenesis and cancer progression, it is likely that as yet unidentified miRNAs exist that may play an influential role in cancer. This project represents, to the best of our knowledge, the first attempt to identify novel intronic onco-miRNAs implicated in breast cancer, by examining the expression patterns of miRNA host genes in a large collection of publicly available breast cancer microarray data sets.

We based our search for novel intragenic onco-miRNAs on the assumption that the regulation of intronic miRNAs closely resembles that of their host genes. To test whether this assumption was reasonable, we compared the expression of mature

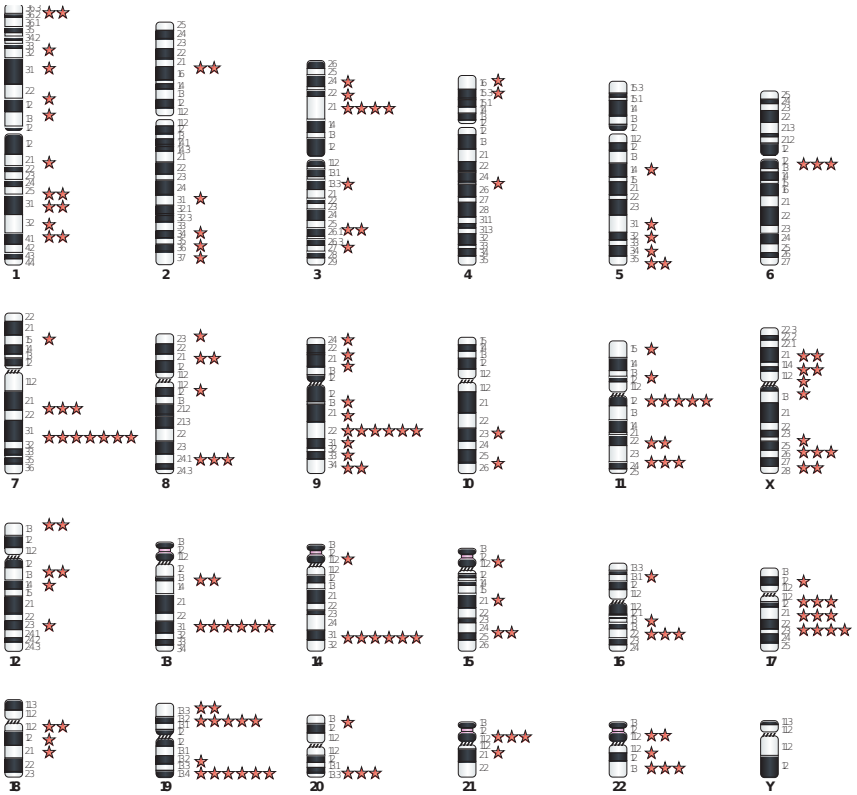


Figure 4.1: MiRNA genes are frequently located in chromosome regions that are involved in rearrangements in human cancers. Here the 24 human chromosomes are shown, with stars indicating the locations of miRNA genes found to be implicated in cancer. Adapted from (Croce, 2009).

miRNAs and of their host genes in two model systems that have been extensively studied in our group: i) the re-entry of terminally differentiated murine myotubes into the cell cycle, upon infection with the adenoviral E1A protein (Sec. 3.2); ii) the MCF10A breast cell line transformed with SV40 (Sec 3.3). For both models, we performed a microarray screening to profile the expression of mRNA genes and a high-throughput PCR screening to profile the expression of mature miRNAs.

We used two different statistical measures to quantify the agreement between the two technologies: the Spearman correlation coefficient, and the fraction of agreeing Affymetrix probe set/TaqMan probe pairs. We opted for the non-parametric version of the correlation coefficient because of its greater robustness to the presence of outliers (Sprent and Smeeton, 2007). In the E1A experiment, we found that the Spearman correlation between the Affymetrix and the TaqMan probes was positive, at both the early and late time points ($r_s = 0.70$ and $r_s = 0.62$ respectively) and significantly different from zero ($p = 6 \times 10^{-9}$ and $p = 2 \times 10^{-6}$ respectively). We observed a fraction of agreement equal to 0.74 at the early time point, and to 0.87 at the late time point. In the SV40 experiment the Spearman correlation coefficient was $r_s = 0.44$, with $p = 0.008$, and the fraction of agreement was 0.72. The mean correlation coefficient in our experiments was $\bar{r} = 0.59$. These results are in good agreement with those obtained by Baskerville and co-workers (Baskerville and Bartel, 2005). These authors used miRNA microarrays to measure the expression levels of 175 human miRNAs across twenty-four different organs and compared the profiles of the intronic miRNAs to the expression patterns of the respective host genes, measured by an Affymetrix screening (Baskerville and Bartel, 2005). In most cases they found that the correlation between the miRNA and the host gene expression was highly positive (mean correlation coefficient $\bar{r} = 0.55$), and concluded that most intronic miRNAs are processed from the same primary transcripts as their host genes. These authors also observed a negative correlation between *miR-26-a* and its host gene, *CTDSPL*, and explained this anti-correlation by the fact that the mature *hsa-miR-26a* miRNA is produced by two distinct precursors, *hsa-mir-26a-1* and *hsa-mir-26a-2*, the latter being located in the *CTDSP2* gene. They noticed that the two

Probe set	miRNA ID	Gene symbol	Affy log FC	PCR log FC	Ct SV40	Ct control
201904 s at	hsa-mir-26a-1	CTDSPL	-0.47	0.21	21.84	21.64
201905 s at	hsa-mir-26a-1	CTDSPL	-0.66	0.21	21.84	21.64
201906 s at	hsa-mir-26a-1	CTDSPL	-0.52	0.21	21.84	21.64
208735 s at	hsa-mir-26a-2	CTDSP2	0.46	0.21	21.84	21.64

Table 4.1: Probe set and miRNA IDs, miRNA host gene symbol, Affymetrix and TaqMan logarithmic fold changes and C_T values of the two precursors of the *hsa-miR-26a* mature miRNA.

host genes are often anti-correlated, *CTDSP2* being expressed at much higher levels than *CTDSPL*. We compared the regulation of the two genes with the regulation of the mature *hsa-miR-26a* miRNA in the SV40 experiment which, like Baskerville's experiment, was based on human cell lines. We observed that *hsa-miR-26a* had an opposite direction of regulation compared with the expression of *CTDSPL*, while it showed the same direction of regulation as *CTDSP2*, thus confirming Baskerville's findings (Table 4.1).

In contrast with our results, a recent study by Sikand *et al.* reported little correlation between the expression of intronic miRNAs and their host genes (Sikand et al., 2009). The authors compared the expression patterns of the members of the *miR-17-92*, *miR-106b-25* and *miR-23b-24* miRNA clusters with the regulation of their host genes, *C13orf25*, *MCM7* and *C9orf3*, in androgen-sensitive and androgen-refractory prostate cancer cell lines (Sikand et al., 2009). In the first two cases no statistically significant correlation was found between the expression levels of the mature miRNAs and the respective host genes. The expression of *C9orf3*, on the other hand, was positively correlated with the expression of *miR-23b* and *miR-27b*, but no correlation was observed with *miR-24*.

To compare our results with the observations reported by Sikand and co-workers, we analyzed the expression of the members of the *miR-25-93-106b* and the *miR-23b-24-27b* clusters, and of their host genes, *MCM7* and *C9orf3*, respectively. We could not examine the regulation of *C13Orf25*, the gene containing the *miR-17-92* cluster, because no probe set mapped to it on the Affymetrix HG-U133A2 platform. In contrast with Sikand *et al.*, we observed a good agreement between the expression of all the members of the clusters and their host gene (Table 4.2).

Probe set	miRNA ID	Gene symbol	Affy log FC	PCR log FC	Ct SV40	Ct control
208795_s_at	hsa-miR-106b	MCM7	0.86	1.27	21.68	22.95
210983_s_at	hsa-miR-106b	MCM7	0.69	1.27	21.68	22.95
208795_s_at	hsa-miR-25	MCM7	0.86	1.34	22.52	23.87
210983_s_at	hsa-miR-25	MCM7	0.69	1.34	22.52	23.87
208795_s_at	hsa-miR-93	MCM7	0.86	0.92	20.57	21.50
210983_s_at	hsa-miR-93	MCM7	0.69	0.92	20.57	21.50
212848_s_at	hsa-miR-23b	C9orf3	-0.16	-0.48	27.43	26.95
212848_s_at	hsa-miR-24	C9orf3	-0.16	-0.82	19.84	19.02
212848_s_at	hsa-miR-27b	C9orf3	-0.16	-0.48	24.59	24.06

Table 4.2: Probe set and miRNA IDs, miRNA host gene symbol, Affymetrix and TaqMan logarithmic fold changes and C_T values of the miRNAs analyzed by Sikand et *al.*

One possible explanation for these contrasting findings, is that the three clusters considered by Sikand have upstream motifs compatible with promoter function (Monteys et al., 2010). It is therefore probable that the three clusters can be transcribed independently of their host gene. thus making a positive correlation between the expression of mature miRNAs and their host genes a possibility, but not a necessity.

4.2 The breast cancer microarray data set analysis

To the best of our knowledge, this analysis represents the first large scale screening of intronic miRNAs in a collection of breast cancer data sets. Our analysis differs from previous studies on the role of miRNAs in breast cancer (see Sec. 1.2.3) in two main respects. First, almost all such studies compared normal with tumour samples, while our research was restricted to tumour samples only. Second, we focused specifically on the intronic miRNAs, and their association with their host gene. The latter point meant that we could take advantage of thousands of breast cancer microarray samples stored in publicly accessible databases, allowing us to achieve a sample size that would have otherwise been impossible.

We analyzed five breast cancer microarray data sets, to identify miRNA host genes that were differentially regulated with respect to the available clinical parameters. In all the data sets we analyzed, we found that the G3 vs. G1 and ER+

vs. ER- comparisons were the most enriched in significantly regulated probe sets. Moreover, the lists of significantly regulated probe sets in the tumour grade and ER status analyses were largely overlapping. This finding reflects the strong association between tumour grade and ER status in the three data sets where both parameters were available, *i.e.*, the Ivshina, Sotiriou and TRANSBIG data sets. We performed a Fisher test to assess the significance of this association, and found that it was highly significant in the Ivshina and TRANSBIG data set ($p = 1.7 \times 10^{-7}$ and $p = 2.8 \times 10^{-9}$) respectively. In contrast, in the Sotiriou data set the association was less significant ($p = 0.058$) due to the very small number of ER- patients (Table. 4.3).

	Ivshina			Sotiriou			TRANSBIG		
ER status	G1	G2	G3	G1	G2	G3	G1	G2	G3
ER-	2	11	21	1	5	4	1	15	45
ER+	61	109	34	34	21	16	26	65	35

Table 4.3: Distribution of tumour grade stratified by ER status in the Ivshina, Sotiriou and the TRANSBIG data set.

The association between tumour grade and ER status was likely to introduce a strong confounding effect. We, therefore, considered the ER+ and ER- data sets separately, and observed that more than 90% (93% in the Ivshina and Sotiriou data sets, and 92% in the TRANSBIG data set) of the significantly regulated probe sets in the G3 vs. G1 comparison, restricted to the ER+ tumours subset, were also included in the corresponding list in the whole data set. This suggested that tumour grade was the dominant cause of regulation of the significantly modulated probe sets.

We then assessed whether the large number of significantly regulated probe sets were specific for miRNA host genes or if it was a general effect due to the profound genetic differences characterizing G1 vs. G3 and ER+ vs. ER- tumours. The simulation described in Section 3.4.3 showed that miRNA host genes were enriched in the significantly regulated probe sets compared with non-host genes in the tumour grade comparison. This observation suggests that the molecular differences between high and low grade tumours (*e.g.*, in terms of cell differentiation and proliferation)

might affect miRNAs on a more global scale, than non-host genes,thus indicating a possible role for intronic miRNAs in cancer progression.

4.3 Identification of candidate novel onco-miRNAs

Our microarray breast cancer screening led to the identification of 27 probe sets, associated with 23 miRNAs (Table A.17). Each of these probe sets was regulated between G3 and G1 tumours in a statistically significant manner, and with the same direction of regulation, in all four data sets containing the tumour grade information. From this list of probe sets, we selected those having an average regulation of at least 1.5 fold in either direction. This resulted in a list of 11 probe sets corresponding to an equal number of intronic miRNAs (Tab. A.19). Of these 11 probe sets, five were upregulated while six were downregulated in G3 tumours. The upregulated probe sets mapped the *MCM7* and *SMC4* genes, which host the *hsa-miR-25-93-106b* and *hsa-miR-15b/16-2* miRNA clusters respectively. the *MCM7* gene is a well known marker of proliferation and, in agreement with our findings, has been found upregulated in several types of cancer (Fujioka et al., 2009; Giaginis et al., 2010; Li et al., 2005; Nishihara et al., 2008). The *hsa-miR-25-9-106b* miRNA cluster is a paralog of the *miR-17-92* cluster, which has also been found overexpressed in a number of human malignancies, as described in Section 3.4.4.

The *SMC4* gene is essential for chromosome assembly and segregation (Losada and Hirano, 2005), but we are not aware of any study describing the role of *SMC4* in breast cancer or any other type of human malignancy. The *mir-15b/16-2* cluster is the paralog of the *miR-15a/16-1* cluster, known to behave as a tumour suppressor in a number of malignancies (see Sec. 3.4.4). However, the *mir-15b/16-2* cluster has been found to be upregulated in melanoma cell lines and in melanoma tissue samples compared to melanocytes and melanocytic nevi respectively (Satzger et al., 2010). In the WM1205 melanoma cell line, characterized by high *miR-15b* expression, downregulation of *miR-15b* resulted in decreased tumour cell proliferation and increased apoptosis (Satzger et al., 2010). Therefore, despite the large homology of

Gene symbol	miRNA ID
SLIT2	*miR-218-1*
EVL	*miR-342*
INS-IGF2	*miR-483*
ErbB4	*miR-548f-2*
COL3A1	*miR-1245*
MYO5C	*miR-1266*

Table 4.4: Gene symbol and miRNA ID of the miRNAs selected for further validation

the two clusters, *miR-15b* exhibits oncogenic features that have not been reported in *miR-15a*.

We focused our attention on the six miRNA host genes that were found consistently downregulated in high grade tumours, since little was known about the hosted miRNAs and their implication in breast cancer. Table 4.4 shows such genes and the miRNAs they host. The *SLIT2* gene, host to *hsa-miR-218-1*, is frequently inactivated in breast and lung cancer (Dallol et al., 2002). Contrary to our findings, high levels of the *EVL* gene, host to *hsa-mir-342*, have been observed in breast cancer (Hu et al., 2008). However, epigenetic silencing of both *EVL* and *miR-342* has been observed in colorectal cancer (Grady et al., 2008). The *mir-483* host gene, *INS-IGF2*[1], belongs to the insulin-like growth factor (IGF) family of growth-regulating hormones that are relevant not only in the growth of normal mammary glands, but also in the development and progression of breast cancer (Espelund et al., 2008; Mu et al., 2009; Yu and Rohan, 2000). The *mir-548f-2* host gene, *ErbB4*, belongs to the EGFR subfamily of receptor tyrosine kinases. ErbB receptors regulate multiple cellular responses, including cell proliferation, survival, migration and differentiation (Hynes and Lane, 2005). In breast cancer, increased expression of *ErbB4* has been associated with low tumour grade and reduced cancer cell proliferation indices, in agreement with our findings (Bièche et al., 2003; Pawlowski et al., 2000). To our knowledge, no studies concerning the *miR-1245* host gene, *COL3A1*, and its connection to breast or any other type of cancer have been published. Finally we could

[1]More precisely, *miR-483* is located in a locus that includes two alternatively spliced readthrough transcript variants which align to the *INS* gene in the 5' region and to the *IGF2* gene in the 3' region.

not find any publication describing either *hsa-mir-1266*, or its host gene, *MYO5C*.

4.4 The validation of our findings

We screened the expression of the selected miRNAs in a panel of commercial breast cell lines and breast tumour samples. We found that:

1. *Mir-342-3p* and *miR-342-5p* were both strongly expressed in the breast cell lines as well as in the tumour samples, and significantly downregulated in G3 compared to G1 tumours ($p = 9.1 \times 10^{-6}$ for *miR-342-3p* and $p = 0.0019$ for *miR-342-5p*).

2. *MiR-483-3p* and *miR-483-5p* were expressed at low levels in the cell lines. These miRNAs were detectable at moderate levels in G1 tumours, and at low levels in G3 tumours, the difference being statistically significant ($p = 0.0019$ for *miR-483-3p* and $p = 0.0022$ for *miR-483-5p*).

3. *Mir-218*, *miR-218** and *miR-1266* were expressed at very low levels in either the cell lines and the breast tumour samples.

4. *Mir-1245* was discarded because of technical problems with the corresponding PCR probe.

Interestingly, we could not detect *miR-548f-2* in either the commercial breast cell lines, or the tissue samples derived from breast cancer patients. This finding is in striking contrast with the strong average signal of the *hsa-mir-548f-2* host gene, *ErbB4*, we measured in the microarray experiment. A possible explanation for this discrepancy is the fact that miRNAs can be under the transcriptional control of a promoter which is different from that of the host gene, as observed by Monteys *et al.* (Monteys et al., 2010). In this study the authors examined the regions surrounding 253 intronic miRNAs, searching for DNA features commonly found in promoters as CpG islands, transcription start sites (TSS), conserved transcription factor binding sites (TFBS), poly(A) signals, and EST data (Fujita and Iba, 2008; Saini et al., 2007). They observed that approximately 34% of intronic miRNAs have upstream

regulatory regions consistent with promoter function. Moreover, for a number of intronic miRNAs, including *hsa-miR-548f-2*, the authors cloned the regions encompassing the miRNAs and and their upstream Pol II or Pol III sequences into a promoter-less plasmid, and confirmed that miRNA expression occurred independently of host gene transcription. The authors also ranked these 253 miRNAs by the number of DNA features compatible with the presence of host-gene independent promoters. We noted that other miRNAs, among those we had chosen for validation, were included in the list of miRNAs possessing promoter-compatible DNA features: *hsa-miR-483* and *hsa-miR-218-1* possess a TSS as well as TFBS, but none of them have been experimentally tested. On the contrary, no promoter-compatible DNA feature was identified in the region surrounding *hsa-miR-342*. Finally, *hsa-miR-1266* was not in Montey's list.

4.5 Re-classification of G2 tumours

We found that *miR-342-3p/5p* and *miR-483-3p/5p* were significantly regulated between high and low grade tumours. This observation suggests a possible use of such miRNAs for grade classification purposes. Tumour grade is a semi-quantitative measure that summarizes three distinct morphological features: percentage of tubule formation, degree of nuclear pleomorphism, and mitotic count. Concordance between institutions on tumour grade evaluation is less than 80% (Robbins et al., 1995). Moreover, 30-60% of tumours are classifed as G2, which is uninformative for clinical decision making because it is associated with an intermediate risk of recurrence.

Sotiriou and collaborators have performed a primary tumour gene expression profiling on sixty-four breast cancers to investigate the molecular basis of histologic grade (Sotiriou et al., 2006). All the tumours were ER+ and none was classified as G2. The authors identified a set of 97 genes that were significantly regulated between G1 and G3 tumours, and used this gene signature to develop a score called the gene expression grade index (GGI). The GGI was then used to re-classify an independent set of G2 tumours. They found that a high GGI value was more

significantly associated with a higher risk of recurrence than a low gene expression grade index.

We considered the four probe sets mapping to the *miR-342-3p/5p* and *miR-483-3p/5p* host genes, *EVL* and *INS-IGF2* and performed an unsupervised hierarchical clustering to assess whether these miRNAs could divide G2 tumours into survival-related subgroups. We applied this procedure to the Ivshina and Pawitan data set, and we found that we could indeed identify two subgroups, and that these were significantly associated with survival in the Pawitan data set, but that such association was not significant in the Ivshina data set (Fig. 3.20). If, however, we considered only the three probe sets mapping to *INS-IGF2*, the patients were separated into two groups displaying a significant association with survival both in the Ivshina and the Pawitan data set (Fig. 3.21). Therefore, the *miR-342* and *miR-483* host genes appear to be associated to tumour grade in different ways. The former is more significantly regulated between G3 and G1, but displays little association with survival, whereas the latter discriminates a high-risk and a low-risk groups within G2 patients. Our validation on G1 and G3 breast cancer samples confirmed that *INS-IGF2* and *miR-483-3p/5p* have highly correlated patterns of regulation. We therefore expect to observe a similar separation of G2 tumours also at a miRNA level.

4.6 Ongoing work and future plans

To determine the biological relevance of the identified intronic miRNAs in breast cancer, we recently started a series of experiments based on the overexpression of synthetic miRNAs. We thus overexpressed *miR-342-3p* and *miR-342-5p* in the metastatic breast cancer cell line MDA-MB231, by transfecting it with the relative synthetic miRNAs. We chose the MDA-MB231 cell line because we found *miR-342-3p/5p* expressed at low levels in these cells compared to other commercial breast cell lines (Fig. 3.14). We found that the overexpression of *miR-342-5p*, but not that of *miR-342-3p*, induces apoptosis in MDA-MB231, therefore suggesting a role for

miR-342 in regulating cell viability (Fig. 3.23). In line with this preliminary observation, Grady *et al.* (Grady et al., 2008) observed that overexpression of *miR-342* in HT29 colon carcinoma cell line induces apoptosis. Further experiments will be needed to characterize better the role of *miR-342* and of *miR-483* in breast cancer progression. In addition, we are collecting G2 tumour samples provided by the European Institute of Oncology, to investigate the capacity of *miR-483-3p/5p* and of *miR-342-3p/5p* to discriminate patients with high and low risk of recurrence, thus confirming what we observed in silico (Sec. 3.4.6).

Appendix A

Appendix

ID	logFC	adj.P.Val	mirna_id	gene_symbol
201664_at	0.95	1e-14	hsa-mir-16-2	SMC4
201664_at	0.95	1e-14	hsa-mir-15b	SMC4
210365_at	-0.81	1.5e-12	hsa-mir-802	RUNX1
214053_at	-1.48	6.6e-12	hsa-mir-548f-2	ERBB4
200710_at	-0.59	9.6e-12	hsa-mir-324	ACADVL
217838_s_at	-1.29	9.6e-12	hsa-mir-342	EVL
203130_s_at	-1.62	2.1e-11	hsa-mir-1978	KIF5C
202409_at	-1.32	2.3e-11	hsa-mir-483	IGF2
219396_s_at	-0.48	2.3e-11	hsa-mir-631	NEIL1
208795_s_at	0.58	5.4e-11	hsa-mir-93	MCM7
208795_s_at	0.58	5.4e-11	hsa-mir-25	MCM7
208795_s_at	0.58	5.4e-11	hsa-mir-106b	MCM7
207783_x_at	-0.16	6.5e-11	hsa-let-7f-2	HUWE1
207783_x_at	-0.16	6.5e-11	hsa-mir-98	HUWE1
201839_s_at	0.80	9e-11	hsa-mir-559	TACSTD1
200875_s_at	0.53	1.1e-10	hsa-mir-1292	NOL5A
201663_s_at	0.78	1.6e-10	hsa-mir-16-2	SMC4
201663_s_at	0.78	1.6e-10	hsa-mir-15b	SMC4
217844_at	-0.29	2e-10	hsa-mir-26b	CTDSP1
210983_s_at	0.65	2.5e-10	hsa-mir-25	MCM7
210983_s_at	0.65	2.5e-10	hsa-mir-93	MCM7
210983_s_at	0.65	2.5e-10	hsa-mir-106b	MCM7
217892_s_at	-0.55	2.4e-09	hsa-mir-1293	LIMA1
221580_s_at	0.51	4.6e-09	hsa-mir-1304	TAF1D
218966_at	-0.69	4.6e-09	hsa-mir-1266	MYO5C
209897_s_at	-0.57	4.8e-09	hsa-mir-218-1	SLIT2
218782_s_at	0.88	4.8e-09	hsa-mir-548d-1	ATAD2
202754_at	0.34	4.9e-09	hsa-mir-128-1	R3HDM1
204134_at	-0.47	2.2e-08	hsa-mir-139	PDE2A
202410_x_at	-0.65	5.5e-08	hsa-mir-483	IGF2
216515_x_at	0.24	7.5e-08	hsa-mir-1244	PTMA
210881_s_at	-0.51	1e-07	hsa-mir-483	IGF2
209219_at	0.30	1.3e-07	hsa-mir-1236	RDBP
218131_s_at	0.36	1.3e-07	hsa-mir-640	GATAD2A
217094_s_at	0.27	2.1e-07	hsa-mir-644	ITCH
201904_s_at	-0.41	2.3e-07	hsa-mir-26a-1	CTDSPL
201906_s_at	-0.44	3.5e-07	hsa-mir-26a-1	CTDSPL
203594_at	0.50	4.1e-07	hsa-mir-553	RTCD1
212156_at	-0.19	4.7e-07	hsa-mir-627	VPS39
209744_x_at	0.25	4.7e-07	hsa-mir-644	ITCH
216384_x_at	0.31	5.2e-07	hsa-mir-1244	PTMA
203812_at	-0.30	5.5e-07	hsa-mir-585	SLIT3
203812_at	-0.30	5.5e-07	hsa-mir-218-2	SLIT3
221221_s_at	-0.31	5.5e-07	hsa-mir-874	KLHL3
211921_x_at	0.34	5.5e-07	hsa-mir-1244	PTMA
202561_at	-0.51	6e-07	hsa-mir-597	TNKS
200772_x_at	0.32	6.3e-07	hsa-mir-1244	PTMA
209360_s_at	-0.43	9.7e-07	hsa-mir-802	RUNX1
212770_at	-0.33	1.3e-06	hsa-mir-629	TLE3
200785_s_at	-0.43	1.9e-06	hsa-mir-1228	LRP1

Table A.1: Affymetrix ID, log fold change, adjusted p-value, pre-miRNA ID, and host gene symbol for the 50 most regulated probe sets in the G3 vs. G1 comparison in the Ivshina data set.

ID	logFC	adj.P.Val	mirna_id	gene_symbol
212209_at	0.55	1.7e-09	hsa-mir-620	MED13L
208795_s_at	-0.55	5.7e-09	hsa-mir-25	MCM7
208795_s_at	-0.55	5.7e-09	hsa-mir-93	MCM7
208795_s_at	-0.55	5.7e-09	hsa-mir-106b	MCM7
218966_at	0.73	5.7e-09	hsa-mir-1266	MYO5C
217838_s_at	1.17	5.7e-09	hsa-mir-342	EVL
218131_s_at	-0.40	1.1e-08	hsa-mir-640	GATAD2A
214053_at	1.30	1.1e-08	hsa-mir-548f-2	ERBB4
212770_at	0.40	1.7e-08	hsa-mir-629	TLE3
203988_s_at	0.66	3.8e-08	hsa-mir-625	FUT8
202409_at	1.10	2.5e-07	hsa-mir-483	IGF2
217892_s_at	0.50	3.7e-07	hsa-mir-1293	LIMA1
212256_at	0.67	1e-06	hsa-mir-1294	GALNT10
210983_s_at	-0.54	1.1e-06	hsa-mir-106b	MCM7
210983_s_at	-0.54	1.1e-06	hsa-mir-25	MCM7
210983_s_at	-0.54	1.1e-06	hsa-mir-93	MCM7
221580_s_at	-0.45	1.3e-06	hsa-mir-1304	TAF1D
212208_at	0.47	3e-06	hsa-mir-620	MED13L
209360_s_at	0.44	3.8e-06	hsa-mir-802	RUNX1
202756_s_at	0.46	5.2e-06	hsa-mir-149	GPC1
202754_at	-0.28	5.2e-06	hsa-mir-128-1	R3HDM1
35666_at	0.43	6.1e-06	hsa-mir-566	SEMA3F
221934_s_at	0.36	1.3e-05	hsa-mir-425	DALRD3
221934_s_at	0.36	1.3e-05	hsa-mir-191	DALRD3
204398_s_at	0.39	1.8e-05	hsa-mir-330	EML2
212207_at	0.36	1.8e-05	hsa-mir-620	MED13L
217844_at	0.21	2.5e-05	hsa-mir-26b	CTDSP1
212156_at	0.17	3.4e-05	hsa-mir-627	VPS39
203130_s_at	1.07	4.3e-05	hsa-mir-1978	KIF5C
200875_s_at	-0.37	5.6e-05	hsa-mir-1292	NOL5A
216515_x_at	-0.19	6.2e-05	hsa-mir-1244	PTMA
209730_at	0.28	7.7e-05	hsa-mir-566	SEMA3F
200710_at	0.38	0.0001	hsa-mir-324	ACADVL
219155_at	-0.32	0.00012	hsa-mir-548d-2	PITPNC1
221958_s_at	0.58	0.00014	hsa-mir-1262	GPR177
201839_s_at	-0.52	0.00014	hsa-mir-559	TACSTD1
200773_x_at	-0.17	0.00014	hsa-mir-1244	PTMA
210365_at	0.47	0.00015	hsa-mir-802	RUNX1
201664_at	-0.51	0.00022	hsa-mir-15b	SMC4
201664_at	-0.51	0.00022	hsa-mir-16-2	SMC4
210881_s_at	0.38	0.00028	hsa-mir-483	IGF2
201663_s_at	-0.49	0.00032	hsa-mir-16-2	SMC4
201663_s_at	-0.49	0.00032	hsa-mir-15b	SMC4
218457_s_at	-0.26	0.00033	hsa-mir-1301	DNMT3A
203266_s_at	0.44	0.0004	hsa-mir-744	MAP2K4
217988_at	-0.31	0.0004	hsa-mir-1201	CCNB1IP1
202308_at	0.52	0.00043	hsa-mir-33b	SREBF1
217726_at	0.24	0.00043	hsa-mir-148b	COPZ1
202410_x_at	0.46	0.00043	hsa-mir-483	IGF2
208963_x_at	-0.38	0.00043	hsa-mir-1908	FADS1

Table A.2: Affymetrix ID, log fold change, adjusted p-value, pre-miRNA ID, and host gene symbol for the 50 most regulated probe sets in the ER positive vs. ER negative comparison in the Ivshina data set.

ID	logFC	adj.P.Val	mirna_id	gene_symbol
202754_at	0.31	6.2e-11	hsa-mir-128-1	R3HDM1
218966_at	-0.65	6.2e-11	hsa-mir-1266	MYO5C
202409_at	-1.05	3.8e-10	hsa-mir-483	IGF2
214053_at	-1.14	4.8e-10	hsa-mir-548f-2	ERBB4
216515_x_at	0.22	1.6e-09	hsa-mir-1244	PTMA
203130_s_at	-1.23	1.6e-09	hsa-mir-1978	KIF5C
201664_at	0.63	1.6e-09	hsa-mir-16-2	SMC4
201664_at	0.63	1.6e-09	hsa-mir-15b	SMC4
200710_at	-0.44	2.1e-09	hsa-mir-324	ACADVL
201663_s_at	0.61	2.6e-09	hsa-mir-16-2	SMC4
201663_s_at	0.61	2.6e-09	hsa-mir-15b	SMC4
218131_s_at	0.33	6.2e-09	hsa-mir-640	GATAD2A
217838_s_at	-0.92	9.2e-09	hsa-mir-342	EVL
217844_at	-0.22	1.4e-08	hsa-mir-26b	CTDSP1
208795_s_at	0.43	1.4e-08	hsa-mir-93	MCM7
208795_s_at	0.43	1.4e-08	hsa-mir-106b	MCM7
208795_s_at	0.43	1.4e-08	hsa-mir-25	MCM7
207783_x_at	-0.12	1.5e-08	hsa-let-7f-2	HUWE1
207783_x_at	-0.12	1.5e-08	hsa-mir-98	HUWE1
203594_at	0.45	6.1e-08	hsa-mir-553	RTCD1
210983_s_at	0.47	6.1e-08	hsa-mir-106b	MCM7
210983_s_at	0.47	6.1e-08	hsa-mir-93	MCM7
210983_s_at	0.47	6.1e-08	hsa-mir-25	MCM7
219396_s_at	-0.32	1.2e-07	hsa-mir-631	NEIL1
221580_s_at	0.39	1.3e-07	hsa-mir-1304	TAF1D
217892_s_at	-0.40	3.1e-07	hsa-mir-1293	LIMA1
209360_s_at	-0.37	6.1e-07	hsa-mir-802	RUNX1
209897_s_at	-0.42	6.2e-07	hsa-mir-218-1	SLIT2
218213_s_at	0.28	8.8e-07	hsa-mir-611	C11orf10
209744_x_at	0.20	1.1e-06	hsa-mir-644	ITCH
212209_at	-0.35	1.2e-06	hsa-mir-620	MED13L
200773_x_at	0.17	1.4e-06	hsa-mir-1244	PTMA
210881_s_at	-0.39	1.9e-06	hsa-mir-483	IGF2
221221_s_at	-0.25	2.2e-06	hsa-mir-874	KLHL3
202410_x_at	-0.48	2.4e-06	hsa-mir-483	IGF2
218782_s_at	0.60	2.6e-06	hsa-mir-548d-1	ATAD2
201881_s_at	0.21	3.1e-06	hsa-mir-630	ARIH1
203265_s_at	-0.38	3.1e-06	hsa-mir-744	MAP2K4
200772_x_at	0.25	3.2e-06	hsa-mir-1244	PTMA
210365_at	-0.45	3.6e-06	hsa-mir-802	RUNX1
200045_at	0.19	4.1e-06	hsa-mir-877	ABCF1
203266_s_at	-0.44	4.5e-06	hsa-mir-744	MAP2K4
211921_x_at	0.26	6.1e-06	hsa-mir-1244	PTMA
212256_at	-0.48	1e-05	hsa-mir-1294	GALNT10
216384_x_at	0.23	1.1e-05	hsa-mir-1244	PTMA
220296_at	-0.35	1.3e-05	hsa-mir-1294	GALNT10
204398_s_at	-0.31	1.3e-05	hsa-mir-330	EML2
221511_x_at	-0.29	1.3e-05	hsa-mir-628	CCPG1
214151_s_at	-0.22	1.7e-05	hsa-mir-628	CCPG1
201116_s_at	-0.50	1.7e-05	hsa-mir-1979	CPE

Table A.3: Affymetrix ID, log fold change, adjusted p-value, pre-miRNA ID, and host gene symbol for the 50 most regulated probe sets in the mutated vs. wild type p53 comparison in the Ivshina data set.

ID	logFC	adj.P.Val	mirna_id	gene_symbol
208795_s_at	0.71	1.5e-06	hsa-mir-93	MCM7
208795_s_at	0.71	1.5e-06	hsa-mir-106b	MCM7
208795_s_at	0.71	1.5e-06	hsa-mir-25	MCM7
210983_s_at	0.75	2.4e-05	hsa-mir-106b	MCM7
210983_s_at	0.75	2.4e-05	hsa-mir-25	MCM7
210983_s_at	0.75	2.4e-05	hsa-mir-93	MCM7
214053_at	-1.28	6.5e-05	hsa-mir-548f-2	ERBB4
221511_x_at	-0.41	6.5e-05	hsa-mir-628	CCPG1
201663_s_at	0.68	6.5e-05	hsa-mir-16-2	SMC4
201663_s_at	0.68	6.5e-05	hsa-mir-15b	SMC4
212770_at	-0.42	8.4e-05	hsa-mir-629	TLE3
209360_s_at	-0.55	0.00021	hsa-mir-802	RUNX1
217838_s_at	-0.91	0.00025	hsa-mir-342	EVL
217844_at	-0.26	0.00027	hsa-mir-26b	CTDSP1
203130_s_at	-1.18	0.00027	hsa-mir-1978	KIF5C
201664_at	0.65	0.00053	hsa-mir-15b	SMC4
201664_at	0.65	0.00053	hsa-mir-16-2	SMC4
217892_s_at	-0.44	0.00056	hsa-mir-1293	LIMA1
203594_at	0.54	0.00065	hsa-mir-553	RTCD1
212256_at	-0.68	0.0007	hsa-mir-1294	GALNT10
209485_s_at	-0.81	0.00075	hsa-mir-320c-2	OSBPL1A
201935_s_at	-0.48	0.00086	hsa-mir-1256	EIF4G3
218213_s_at	0.31	0.0012	hsa-mir-611	C11orf10
208158_s_at	-0.46	0.0013	hsa-mir-320c-2	OSBPL1A
207357_s_at	-0.46	0.0013	hsa-mir-1294	GALNT10
209897_s_at	-0.59	0.0013	hsa-mir-218-1	SLIT2
201906_s_at	-0.42	0.0019	hsa-mir-26a-1	CTDSPL
204906_at	-0.26	0.0019	hsa-mir-1913	RPS6KA2
201622_at	0.22	0.0024	hsa-mir-593	SND1
214151_s_at	-0.30	0.0033	hsa-mir-628	CCPG1
209219_at	0.28	0.0037	hsa-mir-1236	RDBP
202754_at	0.27	0.0041	hsa-mir-128-1	R3HDM1
212912_at	-0.46	0.0065	hsa-mir-1913	RPS6KA2
202561_at	-0.45	0.0065	hsa-mir-597	TNKS
212556_at	0.41	0.0071	hsa-mir-937	SCRIB
212785_s_at	-0.24	0.0071	hsa-mir-302c	LARP7
212785_s_at	-0.24	0.0071	hsa-mir-367	LARP7
212785_s_at	-0.24	0.0071	hsa-mir-302d	LARP7
212785_s_at	-0.24	0.0071	hsa-mir-302b	LARP7
212785_s_at	-0.24	0.0071	hsa-mir-302a	LARP7
200785_s_at	-0.29	0.0071	hsa-mir-1228	LRP1
218966_at	-0.45	0.0071	hsa-mir-1266	MYO5C
215076_s_at	-0.52	0.0079	hsa-mir-1245	COL3A1
221958_s_at	-0.64	0.0081	hsa-mir-1262	GPR177
211161_s_at	-0.55	0.0083	hsa-mir-1245	COL3A1
200875_s_at	0.35	0.011	hsa-mir-1292	NOL5A
201852_x_at	-0.61	0.013	hsa-mir-1245	COL3A1
202777_at	-0.30	0.013	hsa-mir-548e	SHOC2
200045_at	0.18	0.013	hsa-mir-877	ABCF1
33132_at	0.32	0.013	hsa-mir-939	CPSF1

Table A.4: Affymetrix ID, log fold change, adjusted p-value, pre-miRNA ID, and host gene symbol for the 50 most regulated probe sets in the G3 vs. G1 comparison in the Pawitan data set.

ID	logFC	adj.P.Val	mirna_id	gene_symbol
201906_s_at	-0.36	0.0054	hsa-mir-26a-1	CTDSPL
218290_at	0.19	0.0054	hsa-mir-1227	PLEKHJ1
212674_s_at	0.21	0.0054	hsa-mir-1226	DHX30
210235_s_at	0.42	0.0054	hsa-mir-548k	PPFIA1
209485_s_at	-0.61	0.0054	hsa-mir-320c-2	OSBPL1A
208310_s_at	0.35	0.0054	hsa-mir-198	FSTL1
221763_at	-0.40	0.0054	hsa-mir-1296	JMJD1C
221763_at	-0.40	0.0054	hsa-mir-1296	RP11-10C13.2
212912_at	-0.44	0.0054	hsa-mir-1913	RPS6KA2
202409_at	-0.79	0.0073	hsa-mir-483	IGF2
204235_s_at	-0.53	0.008	hsa-mir-561	GULP1
208158_s_at	-0.34	0.0081	hsa-mir-320c-2	OSBPL1A
202561_at	-0.37	0.0099	hsa-mir-597	TNKS
203812_at	-0.21	0.0099	hsa-mir-585	SLIT3
203812_at	-0.21	0.0099	hsa-mir-218-2	SLIT3
214053_at	-0.74	0.0099	hsa-mir-548f-2	ERBB4
201879_at	-0.31	0.0099	hsa-mir-630	ARIH1
217844_at	-0.17	0.011	hsa-mir-26b	CTDSP1
221958_s_at	-0.54	0.011	hsa-mir-1262	GPR177
201935_s_at	-0.32	0.012	hsa-mir-1256	EIF4G3
204355_at	0.21	0.012	hsa-mir-1226	DHX30
202374_s_at	0.22	0.012	hsa-mir-664	RAB3GAP2
209897_s_at	-0.43	0.012	hsa-mir-218-1	SLIT2
202410_x_at	-0.48	0.012	hsa-mir-483	IGF2
209863_s_at	-0.50	0.012	hsa-mir-944	TP63
206621_s_at	0.14	0.013	hsa-mir-590	WBSCR1
201663_s_at	0.38	0.013	hsa-mir-16-2	SMC4
201663_s_at	0.38	0.013	hsa-mir-15b	SMC4
202565_s_at	-0.35	0.013	hsa-mir-604	SVIL
202565_s_at	-0.35	0.013	hsa-mir-938	SVIL
210983_s_at	0.40	0.013	hsa-mir-106b	MCM7
210983_s_at	0.40	0.013	hsa-mir-93	MCM7
210983_s_at	0.40	0.013	hsa-mir-25	MCM7
212208_at	-0.33	0.013	hsa-mir-620	MED13L
209177_at	0.25	0.013	hsa-mir-191	C3orf60
200753_x_at	0.25	0.013	hsa-mir-636	SFRS2
215116_s_at	-0.34	0.013	hsa-mir-199b	DNM1
221221_s_at	-0.15	0.014	hsa-mir-874	KLHL3
202328_s_at	-0.17	0.015	hsa-mir-1225	PKD1
203594_at	0.33	0.016	hsa-mir-553	RTCD1
209744_x_at	0.17	0.017	hsa-mir-644	ITCH
220189_s_at	0.21	0.018	hsa-mir-1229	MGAT4B
203130_s_at	-0.65	0.018	hsa-mir-1978	KIF5C
201117_s_at	-0.54	0.018	hsa-mir-1979	CPE
201117_s_at	-0.54	0.018	hsa-mir-578	CPE
203445_s_at	-0.17	0.02	hsa-mir-26a-2	CTDSP2
200912_s_at	-0.20	0.021	hsa-mir-1248	EIF4A2
212701_at	-0.12	0.023	hsa-mir-190	TLN2
219411_at	0.32	0.023	hsa-mir-328	ELMO3
212414_s_at	-0.35	0.023	hsa-mir-766	SEPT6

Table A.5: Affymetrix ID, log fold change, adjusted p-value, pre-miRNA ID, and host gene symbol for the 50 most regulated probe sets in the relapse vs. no relapse comparison in the Pawitan data set.

ID	logFC	adj.P.Val	mirna_id	gene_symbol
210235_s_at	0.44	0.016	hsa-mir-548k	PPFIA1
209897_s_at	-0.50	0.016	hsa-mir-218-1	SLIT2
214053_at	-0.80	0.016	hsa-mir-548f-2	ERBB4
218290_at	0.18	0.016	hsa-mir-1227	PLEKHJ1
219411_at	0.40	0.016	hsa-mir-328	ELMO3
217838_s_at	-0.61	0.023	hsa-mir-342	EVL
217844_at	-0.17	0.023	hsa-mir-26b	CTDSP1
204235_s_at	-0.50	0.027	hsa-mir-561	GULP1
209485_s_at	-0.54	0.027	hsa-mir-320c-2	OSBPL1A
212701_at	-0.13	0.031	hsa-mir-190	TLN2
212414_s_at	-0.39	0.034	hsa-mir-766	SEPT6
202066_at	0.38	0.034	hsa-mir-548k	PPFIA1
201663_s_at	0.38	0.034	hsa-mir-15b	SMC4
201663_s_at	0.38	0.034	hsa-mir-16-2	SMC4
221221_s_at	-0.15	0.036	hsa-mir-874	KLHL3
209863_s_at	-0.48	0.036	hsa-mir-944	TP63
212912_at	-0.37	0.036	hsa-mir-1913	RPS6KA2
212703_at	-0.15	0.037	hsa-mir-190	TLN2
201664_at	0.38	0.047	hsa-mir-16-2	SMC4
201664_at	0.38	0.047	hsa-mir-15b	SMC4
220189_s_at	0.20	0.049	hsa-mir-1229	MGAT4B
203594_at	0.32	0.049	hsa-mir-553	RTCD1

Table A.6: Affymetrix ID, log fold change, adjusted p-value, pre-miRNA ID, and host gene symbol for all the 20 regulated probe sets in the comparison between patients who died from any cause and those who did not in the Pawitan data set.

ID	logFC	adj.P.Val	mirna_id	gene_symbol
210235_s_at	0.53	0.0035	hsa-mir-548k	PPFIA1
201906_s_at	-0.41	0.0046	hsa-mir-26a-1	CTDSPL
202561_at	-0.48	0.0058	hsa-mir-597	TNKS
209485_s_at	-0.70	0.0058	hsa-mir-320c-2	OSBPL1A
218290_at	0.21	0.0058	hsa-mir-1227	PLEKHJ1
214053_at	-0.89	0.0086	hsa-mir-548f-2	ERBB4
203594_at	0.45	0.0086	hsa-mir-553	RTCD1
202409_at	-0.89	0.0086	hsa-mir-483	IGF2
212701_at	-0.16	0.0086	hsa-mir-190	TLN2
219411_at	0.44	0.0086	hsa-mir-328	ELMO3
202410_x_at	-0.58	0.0086	hsa-mir-483	IGF2
221221_s_at	-0.19	0.0086	hsa-mir-874	KLHL3
209897_s_at	-0.52	0.0086	hsa-mir-218-1	SLIT2
217756_x_at	0.22	0.0086	hsa-mir-1282	SERF2
208158_s_at	-0.37	0.0094	hsa-mir-320c-2	OSBPL1A
221763_at	-0.42	0.01	hsa-mir-1296	RP11-10C13.2
221763_at	-0.42	0.01	hsa-mir-1296	JMJD1C
210881_s_at	-0.49	0.011	hsa-mir-483	IGF2
202066_at	0.44	0.012	hsa-mir-548k	PPFIA1
209744_x_at	0.21	0.016	hsa-mir-644	ITCH
217844_at	-0.18	0.016	hsa-mir-26b	CTDSP1
208310_s_at	0.33	0.016	hsa-mir-198	FSTL1
220189_s_at	0.25	0.016	hsa-mir-1229	MGAT4B
203812_at	-0.22	0.016	hsa-mir-218-2	SLIT3
203812_at	-0.22	0.016	hsa-mir-585	SLIT3
201663_s_at	0.42	0.016	hsa-mir-16-2	SMC4
201663_s_at	0.42	0.016	hsa-mir-15b	SMC4
202374_s_at	0.23	0.018	hsa-mir-664	RAB3GAP2
212208_at	-0.36	0.02	hsa-mir-620	MED13L
204513_s_at	-0.27	0.025	hsa-mir-1200	ELMO1
215116_s_at	-0.37	0.027	hsa-mir-199b	DNM1
212912_at	-0.39	0.032	hsa-mir-1913	RPS6KA2
202328_s_at	-0.18	0.032	hsa-mir-1225	PKD1
212414_s_at	-0.39	0.034	hsa-mir-766	SEPT6
206621_s_at	0.14	0.035	hsa-mir-590	WBSCR1
217908_s_at	0.31	0.036	hsa-mir-1255b-2	IQWD1
209863_s_at	-0.48	0.04	hsa-mir-944	TP63
202065_s_at	0.22	0.04	hsa-mir-548k	PPFIA1
203130_s_at	-0.68	0.041	hsa-mir-1978	KIF5C
209219_at	0.20	0.041	hsa-mir-1236	RDBP
209177_at	0.25	0.043	hsa-mir-191	C3orf60
210983_s_at	0.39	0.043	hsa-mir-25	MCM7
210983_s_at	0.39	0.043	hsa-mir-106b	MCM7
210983_s_at	0.39	0.043	hsa-mir-93	MCM7
217838_s_at	-0.53	0.043	hsa-mir-342	EVL
201664_at	0.39	0.044	hsa-mir-15b	SMC4
201664_at	0.39	0.044	hsa-mir-16-2	SMC4
205756_s_at	-0.26	0.048	hsa-mir-1184	F8
200753_x_at	0.24	0.048	hsa-mir-636	SFRS2

Table A.7: Affymetrix ID, log fold change, adjusted p-value, pre-miRNA ID, and host gene symbol for all the 43 regulated probe sets in the comparison between patients who died from breast cancer and those who did not in the Pawitan data set.

ID	logFC	adj.P.Val	mirna_id	gene_symbol
203130_s_at	-1.82	1.5e-06	hsa-mir-1978	KIF5C
201664_at	0.93	1.5e-06	hsa-mir-15b	SMC4
201664_at	0.93	1.5e-06	hsa-mir-16-2	SMC4
202409_at	-1.41	4.7e-06	hsa-mir-483	IGF2
214053_at	-1.41	6.2e-06	hsa-mir-548f-2	ERBB4
202754_at	0.44	7.3e-06	hsa-mir-128-1	R3HDM1
210365_at	-0.92	9.8e-06	hsa-mir-802	RUNX1
201839_s_at	0.86	1.5e-05	hsa-mir-559	TACSTD1
209897_s_at	-0.65	1.7e-05	hsa-mir-218-1	SLIT2
200785_s_at	-0.63	2.4e-05	hsa-mir-1228	LRP1
217844_at	-0.31	6.7e-05	hsa-mir-26b	CTDSP1
209360_s_at	-0.60	7.3e-05	hsa-mir-802	RUNX1
201116_s_at	-0.75	0.00012	hsa-mir-1979	CPE
201116_s_at	-0.75	0.00012	hsa-mir-578	CPE
200875_s_at	0.59	0.00017	hsa-mir-1292	NOL5A
217838_s_at	-1.17	0.00017	hsa-mir-342	EVL
202410_x_at	-0.73	0.00017	hsa-mir-483	IGF2
218966_at	-0.70	0.00017	hsa-mir-1266	MYO5C
219396_s_at	-0.48	0.00017	hsa-mir-631	NEIL1
210881_s_at	-0.58	0.00032	hsa-mir-483	IGF2
209219_at	0.34	0.00032	hsa-mir-1236	RDBP
217094_s_at	0.32	0.00032	hsa-mir-644	ITCH
221580_s_at	0.56	0.00041	hsa-mir-1304	TAF1D
201906_s_at	-0.50	0.00047	hsa-mir-26a-1	CTDSPL
200784_s_at	-0.35	0.00052	hsa-mir-1228	LRP1
208795_s_at	0.59	0.00052	hsa-mir-93	MCM7
208795_s_at	0.59	0.00052	hsa-mir-106b	MCM7
208795_s_at	0.59	0.00052	hsa-mir-25	MCM7
212733_at	0.34	0.00054	hsa-mir-922	KIAA0226
221511_x_at	-0.43	0.00056	hsa-mir-628	CCPG1
218782_s_at	0.84	0.00063	hsa-mir-548d-1	ATAD2
201852_x_at	-0.77	0.00063	hsa-mir-1245	COL3A1
208782_at	-0.63	0.00078	hsa-mir-198	FSTL1
203812_at	-0.30	0.00088	hsa-mir-218-2	SLIT3
203812_at	-0.30	0.00088	hsa-mir-585	SLIT3
210983_s_at	0.61	0.00088	hsa-mir-106b	MCM7
210983_s_at	0.61	0.00088	hsa-mir-93	MCM7
210983_s_at	0.61	0.00088	hsa-mir-25	MCM7
217892_s_at	-0.54	0.0009	hsa-mir-1293	LIMA1
218825_at	-0.31	0.00097	hsa-mir-126	EGFL7
207783_x_at	-0.15	0.00097	hsa-let-7f-2	HUWE1
207783_x_at	-0.15	0.00097	hsa-mir-98	HUWE1
202561_at	-0.50	0.0011	hsa-mir-597	TNKS
213090_s_at	0.45	0.0011	hsa-mir-1257	TAF4
221221_s_at	-0.29	0.0012	hsa-mir-874	KLHL3
200772_x_at	0.38	0.0015	hsa-mir-1244	PTMA
212701_at	-0.24	0.0015	hsa-mir-190	TLN2
201663_s_at	0.64	0.0015	hsa-mir-16-2	SMC4
201663_s_at	0.64	0.0015	hsa-mir-15b	SMC4
210794_s_at	-0.42	0.0016	hsa-mir-770	MEG3

Table A.8: Affymetrix ID, log fold change, adjusted p-value, pre-miRNA ID, and host gene symbol for the 50 most regulated probe sets in the G3 vs. G1 comparison in the Sotiriou data set.

ID	logFC	adj.P.Val	mirna_id	gene_symbol
208795_s_at	-0.99	6.7e-06	hsa-mir-93	MCM7
208795_s_at	-0.99	6.7e-06	hsa-mir-25	MCM7
208795_s_at	-0.99	6.7e-06	hsa-mir-106b	MCM7
218131_s_at	-0.67	4.7e-05	hsa-mir-640	GATAD2A
210983_s_at	-0.91	0.00036	hsa-mir-93	MCM7
210983_s_at	-0.91	0.00036	hsa-mir-25	MCM7
210983_s_at	-0.91	0.00036	hsa-mir-106b	MCM7
203988_s_at	0.81	0.00092	hsa-mir-625	FUT8
218966_at	0.84	0.001	hsa-mir-1266	MYO5C
200875_s_at	-0.70	0.0016	hsa-mir-1292	NOL5A
214053_at	1.32	0.0016	hsa-mir-548f-2	ERBB4
221580_s_at	-0.65	0.0059	hsa-mir-1304	TAF1D
202754_at	-0.39	0.006	hsa-mir-128-1	R3HDM1
217844_at	0.29	0.008	hsa-mir-26b	CTDSP1
214151_s_at	0.41	0.0083	hsa-mir-628	CCPG1
217988_at	-0.45	0.0097	hsa-mir-1201	CCNB1IP1
217726_at	0.40	0.012	hsa-mir-148b	COPZ1
217892_s_at	0.62	0.012	hsa-mir-1293	LIMA1
221511_x_at	0.47	0.012	hsa-mir-628	CCPG1
52005_at	-0.39	0.012	hsa-mir-1470	WIZ
201117_s_at	0.79	0.015	hsa-mir-1979	CPE
201117_s_at	0.79	0.015	hsa-mir-578	CPE
203827_at	0.64	0.016	hsa-mir-635	WIPI1
201856_s_at	-0.44	0.016	hsa-mir-579	ZFR
221934_s_at	0.41	0.016	hsa-mir-425	DALRD3
221934_s_at	0.41	0.016	hsa-mir-191	DALRD3
35666_at	0.45	0.016	hsa-mir-566	SEMA3F
214152_at	0.35	0.017	hsa-mir-628	CCPG1
202409_at	1.12	0.017	hsa-mir-483	IGF2
212256_at	0.77	0.017	hsa-mir-1294	GALNT10
201622_at	-0.26	0.017	hsa-mir-593	SND1
211921_x_at	-0.44	0.017	hsa-mir-1244	PTMA
208003_s_at	-0.65	0.02	hsa-mir-1538	NFAT5
217838_s_at	0.98	0.02	hsa-mir-342	EVL
200785_s_at	0.51	0.02	hsa-mir-1228	LRP1
200773_x_at	-0.25	0.022	hsa-mir-1244	PTMA
200772_x_at	-0.38	0.023	hsa-mir-1244	PTMA
202756_s_at	0.50	0.027	hsa-mir-149	GPC1
201852_x_at	0.74	0.029	hsa-mir-1245	COL3A1
201116_s_at	0.64	0.029	hsa-mir-1979	CPE
201116_s_at	0.64	0.029	hsa-mir-578	CPE
221783_at	-0.28	0.03	hsa-mir-1470	WIZ
218457_s_at	-0.40	0.034	hsa-mir-1301	DNMT3A
219396_s_at	0.37	0.035	hsa-mir-631	NEIL1
218750_at	-0.45	0.04	hsa-mir-1304	TAF1D
203130_s_at	1.16	0.047	hsa-mir-1978	KIF5C
213249_at	0.41	0.047	hsa-mir-887	FBXL7
209744_x_at	-0.25	0.049	hsa-mir-644	ITCH
216384_x_at	-0.33	0.049	hsa-mir-1244	PTMA

Table A.9: Affymetrix ID, log fold change, adjusted p-value, pre-miRNA ID, and host gene symbol for the 50 most regulated probe sets in the ER positive vs. ER negative comparison in the Sotiriou data set.

ID	logFC	adj.P.Val	mirna_id	gene_symbol
203130_s_at	-1.52	5e-05	hsa-mir-1978	KIF5C
217844_at	-0.34	8.2e-05	hsa-mir-26b	CTDSP1
201906_s_at	-0.58	8.2e-05	hsa-mir-26a-1	CTDSPL
202409_at	-1.21	8.2e-05	hsa-mir-483	IGF2
200045_at	0.37	8.2e-05	hsa-mir-877	ABCF1
208795_s_at	0.54	8.2e-05	hsa-mir-93	MCM7
208795_s_at	0.54	8.2e-05	hsa-mir-106b	MCM7
208795_s_at	0.54	8.2e-05	hsa-mir-25	MCM7
217838_s_at	-0.96	8.2e-05	hsa-mir-342	EVL
221934_s_at	-0.65	0.00012	hsa-mir-425	DALRD3
221934_s_at	-0.65	0.00012	hsa-mir-191	DALRD3
201663_s_at	0.91	0.00018	hsa-mir-15b	SMC4
201663_s_at	0.91	0.00018	hsa-mir-16-2	SMC4
210983_s_at	0.65	0.00023	hsa-mir-25	MCM7
210983_s_at	0.65	0.00023	hsa-mir-93	MCM7
210983_s_at	0.65	0.00023	hsa-mir-106b	MCM7
217726_at	-0.41	0.00031	hsa-mir-148b	COPZ1
214053_at	-1.25	0.00036	hsa-mir-548f-2	ERBB4
219561_at	-0.55	0.00036	hsa-mir-152	COPZ2
201116_s_at	-0.95	0.00044	hsa-mir-1979	CPE
201116_s_at	-0.95	0.00044	hsa-mir-578	CPE
201664_at	0.80	0.00054	hsa-mir-16-2	SMC4
201664_at	0.80	0.00054	hsa-mir-15b	SMC4
203988_s_at	-0.54	0.0013	hsa-mir-625	FUT8
217892_s_at	-0.46	0.0018	hsa-mir-1293	LIMA1
212256_at	-0.69	0.0019	hsa-mir-1294	GALNT10
201904_s_at	-0.41	0.0023	hsa-mir-26a-1	CTDSPL
202308_at	-0.60	0.0024	hsa-mir-33b	SREBF1
200710_at	-0.39	0.0025	hsa-mir-324	ACADVL
203266_s_at	-0.47	0.0028	hsa-mir-744	MAP2K4
209897_s_at	-0.65	0.0028	hsa-mir-218-1	SLIT2
218131_s_at	0.33	0.0029	hsa-mir-640	GATAD2A
214882_s_at	0.29	0.0038	hsa-mir-636	SFRS2
218966_at	-0.63	0.0043	hsa-mir-1266	MYO5C
203445_s_at	-0.26	0.0045	hsa-mir-26a-2	CTDSP2
35666_at	-0.45	0.0081	hsa-mir-566	SEMA3F
212209_at	-0.38	0.01	hsa-mir-620	MED13L
201852_x_at	-0.64	0.017	hsa-mir-1245	COL3A1
221580_s_at	0.42	0.024	hsa-mir-1304	TAF1D
211161_s_at	-0.56	0.024	hsa-mir-1245	COL3A1
200875_s_at	0.26	0.027	hsa-mir-1292	NOL5A
200754_x_at	0.21	0.027	hsa-mir-636	SFRS2
215076_s_at	-0.48	0.028	hsa-mir-1245	COL3A1
200829_x_at	0.13	0.034	hsa-mir-632	ZNF207
209360_s_at	-0.34	0.039	hsa-mir-802	RUNX1
208782_at	-0.41	0.039	hsa-mir-198	FSTL1
211921_x_at	0.22	0.048	hsa-mir-1244	PTMA

Table A.10: Affymetrix ID, log fold change, adjusted p-value, pre-miRNA ID, and host gene symbol for the 39 regulated probe sets in the G3 vs. G1 comparison in the TRANSBIG data set.

ID	logFC	adj.P.Val	mirna_id	gene_symbol
217838_s_at	1.28	7.6e-17	hsa-mir-342	EVL
214053_at	1.80	2.8e-16	hsa-mir-548f-2	ERBB4
203988_s_at	0.75	3.6e-13	hsa-mir-625	FUT8
208795_s_at	-0.63	8.5e-13	hsa-mir-93	MCM7
208795_s_at	-0.63	8.5e-13	hsa-mir-106b	MCM7
208795_s_at	-0.63	8.5e-13	hsa-mir-25	MCM7
210983_s_at	-0.73	2.7e-10	hsa-mir-93	MCM7
210983_s_at	-0.73	2.7e-10	hsa-mir-106b	MCM7
210983_s_at	-0.73	2.7e-10	hsa-mir-25	MCM7
212209_at	0.58	6e-10	hsa-mir-620	MED13L
203130_s_at	1.32	5.3e-09	hsa-mir-1978	KIF5C
217726_at	0.42	8.3e-09	hsa-mir-148b	COPZ1
35666_at	0.61	1.2e-08	hsa-mir-566	SEMA3F
212208_at	0.62	8.9e-08	hsa-mir-620	MED13L
218966_at	0.77	9.3e-08	hsa-mir-1266	MYO5C
212256_at	0.75	1.6e-07	hsa-mir-1294	GALNT10
209177_at	0.41	2.6e-07	hsa-mir-191	C3orf60
202409_at	1.04	3.1e-07	hsa-mir-483	IGF2
221934_s_at	0.55	4.8e-07	hsa-mir-191	DALRD3
221934_s_at	0.55	4.8e-07	hsa-mir-425	DALRD3
217844_at	0.27	7.3e-07	hsa-mir-26b	CTDSP1
202308_at	0.61	1.6e-06	hsa-mir-33b	SREBF1
221580_s_at	-0.55	2.6e-06	hsa-mir-1304	TAF1D
217892_s_at	0.44	3.8e-06	hsa-mir-1293	LIMA1
200045_at	-0.28	5.1e-06	hsa-mir-877	ABCF1
208336_s_at	0.38	5.1e-06	hsa-mir-639	GPSN2
201906_s_at	0.42	7.9e-06	hsa-mir-26a-1	CTDSPL
201663_s_at	-0.68	2.9e-05	hsa-mir-16-2	SMC4
201663_s_at	-0.68	2.9e-05	hsa-mir-15b	SMC4
203445_s_at	0.25	6.5e-05	hsa-mir-26a-2	CTDSP2
200710_at	0.33	0.00017	hsa-mir-324	ACADVL
203266_s_at	0.38	0.00022	hsa-mir-744	MAP2K4
219561_at	0.38	0.00037	hsa-mir-152	COPZ2
217756_x_at	0.22	0.0004	hsa-mir-1282	SERF2
217865_at	-0.23	0.00089	hsa-mir-340	RNF130
201664_at	-0.52	0.0009	hsa-mir-16-2	SMC4
201664_at	-0.52	0.0009	hsa-mir-15b	SMC4
201116_s_at	0.61	0.0009	hsa-mir-1979	CPE
201116_s_at	0.61	0.0009	hsa-mir-578	CPE
201857_at	-0.20	0.0017	hsa-mir-579	ZFR
218680_x_at	0.29	0.0024	hsa-mir-1282	HYPK
218680_x_at	0.29	0.0024	hsa-mir-1282	SERF2
200912_s_at	0.20	0.0029	hsa-mir-1248	EIF4A2
200775_s_at	0.11	0.0029	hsa-mir-7-1	HNRNPK
209360_s_at	0.32	0.0033	hsa-mir-802	RUNX1
200097_s_at	0.12	0.0037	hsa-mir-7-1	HNRNPK
203594_at	-0.30	0.0037	hsa-mir-553	RTCD1
217118_s_at	0.20	0.0048	hsa-mir-1249	C22orf9
209120_at	0.40	0.0052	hsa-mir-1469	NR2F2
201839_s_at	-0.46	0.0055	hsa-mir-559	TACSTD1

Table A.11: Affymetrix ID, log fold change, adjusted p-value, pre-miRNA ID, and host gene symbol for the 50 most regulated probe sets in the ER positive vs. ER negative comparison in TRANSBIG data set.

ID	logFC	adj.P.Val	mirna_id	gene_symbol
214053_at	1.80	8.9e-24	hsa-mir-548f-2	ERBB4
215729_s_at	-3.66	1.2e-21	hsa-mir-934	VGLL1
217838_s_at	1.33	6.9e-19	hsa-mir-342	EVL
218966_at	0.78	8.9e-19	hsa-mir-1266	MYO5C
221934_s_at	0.75	8.9e-19	hsa-mir-191	DALRD3
221934_s_at	0.75	8.9e-19	hsa-mir-425	DALRD3
203130_s_at	1.22	1.9e-15	hsa-mir-1978	KIF5C
35666_at	0.70	1.9e-14	hsa-mir-566	SEMA3F
205487_s_at	-1.72	2.9e-13	hsa-mir-934	VGLL1
213249_at	0.82	2.9e-12	hsa-mir-887	FBXL7
212207_at	0.71	8.5e-12	hsa-mir-620	MED13L
212209_at	0.63	1.1e-11	hsa-mir-620	MED13L
203999_at	1.19	4.8e-11	hsa-mir-1252	SYT1
200045_at	-0.35	1.8e-10	hsa-mir-877	ABCF1
212256_at	0.63	2.9e-10	hsa-mir-1294	GALNT10
204537_s_at	-0.94	3.1e-10	hsa-mir-452	GABRE
204537_s_at	-0.94	3.1e-10	hsa-mir-224	GABRE
218131_s_at	-0.52	1.6e-09	hsa-mir-640	GATAD2A
206794_at	0.78	3.5e-09	hsa-mir-548f-2	ERBB4
203988_s_at	0.65	4.5e-09	hsa-mir-625	FUT8
212715_s_at	-1.05	8.1e-09	hsa-mir-648	MICAL3
208795_s_at	-0.46	4.3e-08	hsa-mir-93	MCM7
208795_s_at	-0.46	4.3e-08	hsa-mir-25	MCM7
208795_s_at	-0.46	4.3e-08	hsa-mir-106b	MCM7
219474_at	0.90	4.8e-08	hsa-mir-567	C3orf52
214151_s_at	0.46	5e-08	hsa-mir-628	CCPG1
203998_s_at	1.63	5.6e-08	hsa-mir-1252	SYT1
201906_s_at	0.45	7e-08	hsa-mir-26a-1	CTDSPL
217892_s_at	0.46	8.6e-08	hsa-mir-1293	LIMA1
212208_at	0.48	2.2e-07	hsa-mir-620	MED13L
204496_at	0.46	2.7e-07	hsa-mir-624	STRN3
201663_s_at	-0.57	3.6e-07	hsa-mir-16-2	SMC4
201663_s_at	-0.57	3.6e-07	hsa-mir-15b	SMC4
220296_at	0.78	3.6e-07	hsa-mir-1294	GALNT10
212770_at	0.53	7.1e-07	hsa-mir-629	TLE3
219411_at	0.46	1.5e-06	hsa-mir-328	ELMO3
212349_at	0.37	2.2e-06	hsa-mir-1825	POFUT1
222156_x_at	0.46	3.4e-06	hsa-mir-628	CCPG1
216109_at	1.06	8.3e-06	hsa-mir-620	MED13L
221580_s_at	-0.39	1.2e-05	hsa-mir-1304	TAF1D
218433_at	0.34	1.5e-05	hsa-mir-103-1	PANK3
218433_at	0.34	1.5e-05	hsa-mir-103-1-as	PANK3
212785_s_at	0.27	1.8e-05	hsa-mir-302b	LARP7
212785_s_at	0.27	1.8e-05	hsa-mir-367	LARP7
212785_s_at	0.27	1.8e-05	hsa-mir-302c	LARP7
212785_s_at	0.27	1.8e-05	hsa-mir-302d	LARP7
212785_s_at	0.27	1.8e-05	hsa-mir-302a	LARP7
204398_s_at	0.35	2e-05	hsa-mir-330	EML2
203775_at	-0.36	2.5e-05	hsa-mir-591	SLC25A13
209744_x_at	-0.25	2.5e-05	hsa-mir-644	ITCH

Table A.12: Affymetrix ID, log fold change, adjusted p-value, pre-miRNA ID, and host gene symbol for the 50 most regulated probe sets in the ER positive vs. ER negative comparison in the Wang data set.

ID	logFC	adj.P.Val	mirna_id	gene_symbol
201664_at	0.30	0.0068	hsa-mir-16-2	SMC4
201664_at	0.30	0.0068	hsa-mir-15b	SMC4
211251_x_at	-0.25	0.016	hsa-mir-30c-1	NFYC
211251_x_at	-0.25	0.016	hsa-mir-30e	NFYC
212474_at	0.20	0.023	hsa-mir-550-2	AVL9
202216_x_at	-0.24	0.024	hsa-mir-30c-1	NFYC
202216_x_at	-0.24	0.024	hsa-mir-30e	NFYC
202066_at	0.31	0.047	hsa-mir-548k	PPFIA1
209219_at	0.20	0.049	hsa-mir-1236	RDBP

Table A.13: Affymetrix ID, log fold change, adjusted p-value, pre-miRNA ID, and host gene symbol for the 6 regulated probe sets in the Relapse vs. No relapse comparison in the Wang data set.

ID	logFC	adj.P.Val	mirna_id	gene_symbol
204022_at	0.99	0.00039	hsa-mir-140	WWP2
209905_at	2.18	0.017	hsa-mir-196b	HOXA9
214651_s_at	1.47	0.024	hsa-mir-196b	HOXA9
201857_at	0.46	0.024	hsa-mir-579	ZFR
208795_s_at	0.68	0.024	hsa-mir-25	MCM7
208795_s_at	0.68	0.024	hsa-mir-93	MCM7
208795_s_at	0.68	0.024	hsa-mir-106b	MCM7
201664_at	0.66	0.027	hsa-mir-16-2	SMC4
201664_at	0.66	0.027	hsa-mir-15b	SMC4
206832_s_at	-1.34	0.04	hsa-mir-566	SEMA3F

Table A.14: Affymetrix ID, log fold change, adjusted p-value, pre-miRNA ID, and host gene symbol for the 7 regulated probe sets in the Brain relapse vs. No relapse comparison in the Wang data set.

	Ivshina	Pawitan	Sotiriou	Transbig
hsa-mir-106b	1	1	1	1
hsa-mir-1245	-1	-1	-1	-1
hsa-mir-1266	-1	-1	-1	-1
hsa-mir-1292	1	1	1	1
hsa-mir-1293	-1	-1	-1	-1
hsa-mir-1294	-1	-1	-1	-1
hsa-mir-1304	1	1	1	1
hsa-mir-152	-1	-1	-1	-1
hsa-mir-15b	1	1	1	1
hsa-mir-16-2	1	1	1	1
hsa-mir-1978	-1	-1	-1	-1
hsa-mir-198	1	-1	-1	-1
hsa-mir-218-1	-1	-1	-1	-1
hsa-mir-25	1	1	1	1
hsa-mir-26a-1	-1	-1	-1	-1
hsa-mir-26a-2	-1	-1	-1	-1
hsa-mir-26b	-1	-1	-1	-1
hsa-mir-33b	-1	-1	-1	-1
hsa-mir-342	-1	-1	-1	-1
hsa-mir-483	-1	-1	-1	-1
hsa-mir-548f-2	-1	-1	-1	-1
hsa-mir-636	1	1	1	1
hsa-mir-640	1	1	1	1
hsa-mir-802	-1	-1	-1	-1
hsa-mir-93	1	1	1	1

Table A.15: List of the miRNAs found to be regulated betwen G3 and G1 in all the data sets where the information was available. The numbers indicate the direction of regulation: 1 stands for upregulated in G3 with respect to G1 and -1 means the opposite.

	Ivshina	Pawitan	Sotiriou	Transbig
hsa-mir-106b	-1	-1	-1	-1
hsa-mir-1244	-1	-1	-1	-1
hsa-mir-1266	1	1	1	1
hsa-mir-1293	1	1	1	1
hsa-mir-1294	1	1	1	1
hsa-mir-1304	-1	-1	-1	-1
hsa-mir-148b	1	1	1	1
hsa-mir-191	1	1	1	1
hsa-mir-1978	1	1	1	1
hsa-mir-1979	1	1	1	1
hsa-mir-25	-1	-1	-1	-1
hsa-mir-26b	1	1	1	1
hsa-mir-342	1	1	1	1
hsa-mir-425	1	1	1	1
hsa-mir-483	1	1	1	1
hsa-mir-548f-2	1	1	1	1
hsa-mir-566	1	1	1	1
hsa-mir-578	1	1	1	1
hsa-mir-593	-1	-1	-1	-1
hsa-mir-625	1	1	1	1
hsa-mir-93	-1	-1	-1	-1

Table A.16: List of the miRNAs found to be regulated betwen ER positive and ER negative in all the data sets where the information was available. The numbers indicate the direction of regulation: 1 stands for upregulated in ER positive with respect to ER negative and -1 means the opposite.

probe set	mirna id	Ivshina FC	Ivshina p-val	Pawitan FC	Pawitan p-val	Sotiriou FC	Sotiriou p-val	Transbig FC	Transbig p-val
208795_s_at	hsa-mir-106b	0.58	5.4e-11	0.71	1.5e-06	0.59	0.00052	0.54	8.2e-05
210983_s_at	hsa-mir-106b	0.65	2.5e-10	0.75	2.4e-05	0.61	0.00088	0.65	0.00023
201852_x_at	hsa-mir-1245	-0.49	0.0056	-0.61	0.013	-0.77	0.00063	-0.64	0.017
215076_s_at	hsa-mir-1245	-0.36	0.018	-0.52	0.0079	-0.53	0.0031	-0.48	0.028
218966_at	hsa-mir-1266	-0.69	4.6e-09	-0.45	0.0071	-0.70	0.00017	-0.63	0.0043
200875_s_at	hsa-mir-1292	0.53	1.2e-10	0.35	0.011	0.59	0.00017	0.26	0.027
217892_s_at	hsa-mir-1293	-0.55	2.4e-09	-0.44	0.00056	-0.54	0.0009	-0.46	0.0018
212256_at	hsa-mir-1294	-0.51	0.00011	-0.68	0.0007	-0.48	0.036	-0.69	0.0019
221580_s_at	hsa-mir-1304	0.51	4.6e-09	0.34	0.044	0.56	0.00041	0.42	0.024
219561_at	hsa-mir-152	-0.34	3.1e-05	-0.27	0.038	-0.37	0.0061	-0.55	0.00036
201663_s_at	hsa-mir-15b	0.78	1.6e-10	0.68	6.6e-05	0.64	0.0015	0.91	0.00018
201664_at	hsa-mir-15b	0.95	1e-14	0.65	0.00053	0.93	1.5e-06	0.80	0.00054
201663_s_at	hsa-mir-16-2	0.78	1.6e-10	0.68	6.6e-05	0.64	0.0015	0.91	0.00018
201664_at	hsa-mir-16-2	0.95	1e-14	0.65	0.00053	0.93	1.5e-06	0.80	0.00054
203130_s_at	hsa-mir-1978	-1.62	2.1e-11	-1.18	0.00027	-1.82	1.5e-06	-1.52	5e-05
208782_at	hsa-mir-198	-0.41	0.0022	-0.46	0.013	-0.63	0.00077	-0.41	0.039
209897_s_at	hsa-mir-218-1	-0.57	4.8e-09	-0.59	0.0013	-0.65	1.7e-05	-0.65	0.0028
208795_s_at	hsa-mir-25	0.58	5.4e-11	0.71	1.5e-06	0.59	0.00052	0.54	8.2e-05
210983_s_at	hsa-mir-25	0.65	2.5e-10	0.75	2.4e-05	0.61	0.00088	0.65	0.00023
201904_s_at	hsa-mir-26a-1	-0.41	2.3e-07	-0.25	0.037	-0.39	0.0021	-0.41	0.0023
201906_s_at	hsa-mir-26a-1	-0.44	3.5e-07	-0.42	0.0019	-0.50	0.00047	-0.58	8.2e-05
203445_s_at	hsa-mir-26a-2	-0.20	0.00017	-0.18	0.04	-0.21	0.012	-0.26	0.0045
217844_at	hsa-mir-26b	-0.29	2e-10	-0.26	0.00027	-0.31	6.7e-05	-0.34	8.2e-05
202308_at	hsa-mir-33b	-0.50	0.0004	-0.49	0.021	-0.56	0.021	-0.60	0.0024
217838_s_at	hsa-mir-342	-1.29	9.6e-12	-0.91	0.00025	-1.17	0.00017	-0.96	8.2e-05
202409_at	hsa-mir-483	-1.32	2.3e-11	-0.78	0.013	-1.41	4.7e-06	-1.21	8.2e-05
214053_at	hsa-mir-548f-2	-1.48	6.6e-12	-1.28	6.6e-05	-1.41	6.2e-06	-1.25	0.00036
200754_x_at	hsa-mir-636	0.15	0.003	0.21	0.031	0.17	0.036	0.21	0.027
214882_s_at	hsa-mir-636	0.21	0.0016	0.27	0.014	0.22	0.024	0.29	0.0038
218131_s_at	hsa-mir-640	0.36	1.3e-07	0.26	0.041	0.33	0.0072	0.33	0.0029
209360_s_at	hsa-mir-802	-0.43	9.7e-07	-0.55	0.00021	-0.60	7.3e-05	-0.34	0.038
208795_s_at	hsa-mir-93	0.58	5.4e-11	0.71	1.5e-06	0.59	0.00052	0.54	8.2e-05
210983_s_at	hsa-mir-93	0.65	2.5e-10	0.75	2.4e-05	0.61	0.00088	0.65	0.00023

Table A.17: Probe sets, miRNA IDs, fold changes and adjusted p-values of the probe sets found to be significantly regulated in G3 vs. G1, in all the data sets

probe set	mirna id	Ivshina FC	Ivshina p-val	Pawitan FC	Pawitan p-val	Sotiriou FC	Sotiriou p-val	Transbig FC	Transbig p-val
208795_s_at	hsa-mir-106b	-0.55	5.7e-09	-0.46	4.3e-08	-0.99	6.7e-06	-0.63	8.5e-13
210983_s_at	hsa-mir-106b	-0.54	1.1e-06	-0.56	0.00028	-0.91	0.00036	-0.73	2.7e-10
200772_x_at	hsa-mir-1244	-0.22	0.0016	-0.20	0.012	-0.38	0.023	-0.15	0.035
211921_x_at	hsa-mir-1244	-0.24	0.00097	-0.19	0.01	-0.44	0.017	-0.17	0.022
218966_at	hsa-mir-1266	0.73	5.7e-09	0.78	8.9e-19	0.84	0.001	0.77	9.3e-08
217892_s_at	hsa-mir-1293	0.50	3.7e-07	0.46	8.6e-08	0.62	0.012	0.44	3.8e-06
212256_at	hsa-mir-1294	0.67	1e-06	0.63	2.9e-10	0.77	0.017	0.75	1.6e-07
221580_s_at	hsa-mir-1304	-0.45	1.3e-06	-0.39	1.2e-05	-0.65	0.0059	-0.55	2.6e-06
217726_at	hsa-mir-148b	0.24	0.00043	0.32	2.8e-05	0.41	0.012	0.42	8.3e-09
221934_s_at	hsa-mir-191	0.36	1.3e-05	0.75	8.9e-19	0.41	0.016	0.55	4.8e-07
203130_s_at	hsa-mir-1978	1.07	4.3e-05	1.22	1.9e-15	1.16	0.047	1.31	5.3e-09
201116_s_at	hsa-mir-1979	0.48	0.0013	0.43	0.016	0.64	0.029	0.60	0.0009
208795_s_at	hsa-mir-25	-0.55	5.7e-09	-0.46	4.3e-08	-0.99	6.7e-06	-0.63	8.5e-13
210983_s_at	hsa-mir-25	-0.54	1.1e-06	-0.56	0.00028	-0.91	0.00036	-0.73	2.7e-10
217844_at	hsa-mir-26b	0.21	2.5e-05	0.21	6.7e-05	0.29	0.008	0.27	7.3e-07
217838_s_at	hsa-mir-342	1.17	5.7e-09	1.33	6.9e-19	0.98	0.02	1.28	7.6e-17
221934_s_at	hsa-mir-425	0.36	1.3e-05	0.75	8.9e-19	0.41	0.016	0.55	4.8e-07
202409_at	hsa-mir-483	1.10	2.5e-07	0.49	0.011	1.12	0.017	1.04	3.1e-07
214053_at	hsa-mir-548f-2	1.30	1.1e-08	1.80	8.9e-24	1.32	0.0016	1.80	2.8e-16
35666_at	hsa-mir-566	0.43	6.1e-06	0.70	1.9e-14	0.45	0.016	0.61	1.2e-08
201116_s_at	hsa-mir-578	0.48	0.0013	0.43	0.016	0.64	0.029	0.60	0.0009
201622_at	hsa-mir-593	-0.12	0.025	-0.14	0.027	-0.27	0.017	-0.16	0.0055
203988_s_at	hsa-mir-625	0.66	3.8e-08	0.65	4.5e-09	0.81	0.00092	0.75	3.6e-13
208795_s_at	hsa-mir-93	-0.55	5.7e-09	-0.46	4.3e-08	-0.99	6.7e-06	-0.63	8.5e-13
210983_s_at	hsa-mir-93	-0.54	1.1e-06	-0.56	0.00028	-0.91	0.00036	-0.73	2.7e-10

Table A.18: Probe sets, miRNA IDs, fold changes and adjusted p-values of the probe sets found to be significantly regulated in ER+ vs. ER-, in all the data sets.

Table A.19: Probe sets, miRNA IDs, fold changes and adjusted p-values of the probe sets that were significantly regulated in G3 vs. G1 in all the data sets, and had an absolute log-fold change greater than 1.5.

probe set	mirna id	Ivshina FC	Ivshina p-val	Pawitan FC	Pawitan p-val	Sotiriou FC	Sotiriou p-val	Transbig FC	Transbig p-val
208795_s_at	hsa-mir-106b	0.58	5.4e-11	0.71	1.5e-06	0.59	0.00052	0.54	8.2e-05
210983_s_at	hsa-mir-106b	0.65	2.5e-10	0.75	2.4e-05	0.61	0.00088	0.65	0.00023
201852_x_at	hsa-mir-1245	-0.49	0.0056	-0.61	0.013	-0.77	0.00063	-0.64	0.017
215076_s_at	hsa-mir-1245	-0.36	0.018	-0.52	0.0079	-0.53	0.0031	-0.48	0.028
218966_at	hsa-mir-1266	-0.69	4.6e-09	-0.45	0.0071	-0.70	0.00017	-0.63	0.0043
201663_s_at	hsa-mir-15b	0.78	1.6e-10	0.68	6.6e-05	0.64	0.0015	0.91	0.00018
201664_at	hsa-mir-15b	0.95	1e-14	0.65	0.00053	0.93	1.5e-06	0.80	0.00054
201663_s_at	hsa-mir-16-2	0.78	1.6e-10	0.68	6.6e-05	0.64	0.0015	0.91	0.00018
201664_at	hsa-mir-16-2	0.95	1e-14	0.65	0.00053	0.93	1.5e-06	0.80	0.00054
209897_s_at	hsa-mir-218-1	-0.57	4.8e-09	-0.59	0.0013	-0.65	1.7e-05	-0.65	0.0028
208795_s_at	hsa-mir-25	0.58	5.4e-11	0.71	1.5e-06	0.59	0.00052	0.54	8.2e-05
210983_s_at	hsa-mir-25	0.65	2.5e-10	0.75	2.4e-05	0.61	0.00088	0.65	0.00023
217838_s_at	hsa-mir-342	-1.29	9.6e-12	-0.91	0.00025	-1.17	0.00017	-0.96	8.2e-05
202409_at	hsa-mir-483	-1.32	2.3e-11	-0.78	0.013	-1.41	4.7e-06	-1.21	8.2e-05
214053_at	hsa-mir-548f-2	-1.48	6.6e-12	-1.28	6.6e-05	-1.41	6.2e-06	-1.25	0.00036
208795_s_at	hsa-mir-93	0.58	5.4e-11	0.71	1.5e-06	0.59	0.00052	0.54	8.2e-05
210983_s_at	hsa-mir-93	0.65	2.5e-10	0.75	2.4e-05	0.61	0.00088	0.65	0.00023

References

Adrian, G., Dan, C., Shimizu, M., Bichi, R., Zupo, S., Noch, E., Aldler, H., Rattan, S., Keating, M., Rai, K., Rassenti, L., Kipps, T., Negrini, M., Bullrich, F. and Croce, C. M. (2002). Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proceedings of the National Academy of Sciences of the United States of America *99*, 15524–15529. 22

Alevizopoulos, K., Catarin, B., Vlach, J. and Amati, B. (1998). A novel function of adenovirus E1A is required to overcome growth arrest by the CDK2 inhibitor p27(Kip1). The EMBO journal *17*, 5987–5997. 40

Altman, D. G. (1990). Practical Statistics for Medical Research (Statistics texts). 1 edition, Chapman & Hall. 33

Ambros, V. (2008). The evolution of our thinking about microRNAs. Nature medicine *14*, 1036–1040. 2

Ambros, V. R. (1989). A hierarchy of regulatory genes controls a larva-to-adult developmental switch in C. elegans. Cell *57*, 49–57. 1

Baek, D., Villén, J., Shin, C., Camargo, F. D., Gygi, S. P. and Bartel, D. P. (2008). The impact of microRNAs on protein output. Nature *455*, 64–71. 13, 17, 19

Bandara, L. R. and La Thangue, N. B. (1991). Adenovirus E1a prevents the retinoblastoma gene product from complexing with a cellular transcription factor. Nature *351*, 494–497. 40

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W.-C. C., Ledoux, P., Rudnev, D., Lash, A. E., Fujibuchi, W. and Edgar, R. (2005). NCBI GEO: mining millions of expression profiles–database and tools. Nucleic acids research *33*. 29, 56

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281–297. 15, 17

Bartel, D. P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. Cell *136*, 215–233. 12, 13, 14, 15, 16, 17, 18

Bashirullah, A. (2003). Coordinate regulation of small temporal RNAs at the onset of Drosophila metamorphosis. Developmental Biology *259*, 1–8. 2

Baskerville, S. and Bartel, D. P. (2005). Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA (New York, N.Y.) *11*, 241–247. 21, 95

REFERENCES

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) *57*, 289–300. 33, 45

Bièche, I., Onody, P., Tozlu, S., Driouch, K., Vidaud, M. and Lidereau, R. (2003). Prognostic value of ERBB family mRNA expression in breast carcinomas. International journal of cancer. Journal international du cancer *106*, 758–765. 100

Blenkiron, C., Goldstein, L., Thorne, N., Spiteri, I., Chin, S. F., Dunning, M., Morais, N. B., Teschendorff, A., Green, A., Ellis, I., Tavare, S., Caldas, C. and Miska, E. (2007). MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. Genome biology *8*, R214+. 28

Bohnsack, M. T., Czaplinski, K. and Gorlich, D. (2004). Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. RNA (New York, N.Y.) *10*, 185–191. 8

Bolstad, B. M. (2004). Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization. PhD thesis, University of California, Berkeley. 35, 44

Bonneterre, J., Thürlimann, B., Robertson, J. F., Krzakowski, M., Mauriac, L., Koralewski, P., Vergote, I., Webster, A., Steinberg, M. and von Euler, M. (2000). Anastrozole versus tamoxifen as first-line therapy for advanced breast cancer in 668 postmenopausal women: results of the Tamoxifen or Arimidex Randomized Group Efficacy and Tolerability study. Journal of clinical oncology *18*, 3748–3757. 26

Borchert, G. M., Lanier, W. and Davidson, B. L. (2006). RNA polymerase III transcribes human microRNAs. Nature structural & molecular biology *13*, 1097–1101. 5

Brennecke, J., Stark, A., Russell, R. B. and Cohen, S. M. (2005). Principles of MicroRNATarget Recognition. PLoS Biol *3*, e85+. 12, 14, 17

Brodersen, P. and Voinnet, O. (2009). Revisiting the principles of microRNA target recognition and mode of action. Nature Reviews Molecular Cell Biology *10*, 141–148. 11, 12, 13

Buyse, M., Loi, S., van't Veer, L., Viale, G., Delorenzi, M., Glas, A. M., d'Assignies, M. S. S., Bergh, J., Lidereau, R., Ellis, P., Harris, A., Bogaerts, J., Therasse, P., Floore, A., Amakrane, M., Piette, F., Rutgers, E., Sotiriou, C., Cardoso, F., Piccart, M. J. and TRANSBIG Consortium (2006). Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. Journal of the National Cancer Institute *98*, 1183–1192. 30, 69

Cai, X., Hagedorn, C. H. and Cullen, B. R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. RNA (New York, N.Y.) *10*, 1957–1966. 5

Calin, G. A. and Croce, C. M. (2006). MicroRNA signatures in human cancers. Nature reviews. Cancer *6*, 857–866. 23

Calin, G. A. A., Sevignani, C., Dumitru, C. D. D., Hyslop, T., Noch, E., Yendamuri, S., Shimizu, M., Rattan, S., Bullrich, F., Negrini, M. and Croce, C. M. (2004). Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. Proceedings of the National Academy of Sciences of the United States of America *101*, 2999–3004. 93

Camarda, G., Siepi, F., Pajalunga, D., Bernardini, C., Rossi, R., Montecucco, A., Meccia, E. and Crescenzi, M. (2004). A pRb-independent mechanism preserves the postmitotic state in terminally differentiated skeletal muscle cells. The Journal of Cell Biology *167*, 417–423. 40

Carthew, R. W. and Sontheimer, E. J. (2009). Origins and Mechanisms of miRNAs and siRNAs. Cell *136*, 642–655. 3, 4, 9, 10, 11

Chapman, E. J. and Carrington, J. C. (2007). Specialization and evolution of endogenous small RNA pathways. Nature reviews. Genetics *8*, 884–896. 5

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*, 740–744. 9

Cimmino, A., Calin, G. A. A., Fabbri, M., Iorio, M. V., Ferracin, M., Shimizu, M., Wojcik, S. E., Aqeilan, R. I., Zupo, S., Dono, M., Rassenti, L., Alder, H., Volinia, S., Liu, C.-G. G., Kipps, T. J., Negrini, M. and Croce, C. M. (2005). miR-15 and miR-16 induce apoptosis by targeting BCL2. Proceedings of the National Academy of Sciences of the United States of America *102*, 13944–13949. 23, 79

Corcoran, D. L., Pandit, K. V., Gordon, B., Bhattacharjee, A., Kaminski, N. and Benos, P. V. (2009). Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data. PloS one *4*, e5279+. 20

Crescenzi, M., Soddu, S. and Tatò, F. (1995). Mitotic cycle reactivation in terminally differentiated cells by adenovirus infection. Journal of cellular physiology *162*, 26–35. 39

Croce, C. M. (2009). Causes and consequences of microRNA dysregulation in cancer. Nature reviews. Genetics *10*, 704–714. 22, 93, 94

Dallol, A., Da Silva, N. F. F., Viacava, P., Minna, J. D., Bieche, I., Maher, E. R. and Latif, F. (2002). SLIT2, a human homologue of the Drosophila Slit2 gene, has tumor suppressor activity and is frequently inactivated in lung and breast cancers. Cancer research *62*, 5874–5880. 100

De Marchis, M. L., Ballarino, M., Salvatori, B., Puzzolo, M. C., Bozzoni, I. and Fatica, A. (2009). A new molecular network comprising PU.1, interferon regulatory factor proteins and miR-342 stimulates ATRA-mediated granulocytic differentiation of acute promyelocytic leukemia cells. Leukemia *23*, 856–862. 79

Deleu, L., Shellard, S., Alevizopoulos, K., Amati, B. and Land, H. (2001). Recruitment of TRRAP required for oncogenic transformation by E1A. Oncogene *20*, 8270–8275. 40

Denli, A. M., Tops, B. B., Plasterk, R. H., Ketting, R. F. and Hannon, G. J. (2004). Processing of primary microRNAs by the Microprocessor complex. Nature *432*, 231–235. 7

Deroo, B. J. and Korach, K. S. (2006). Estrogen receptors and human disease. The Journal of clinical investigation *116*, 561–570. 25

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., d'Assignies, M. S. S., Bergh, J., Lidereau, R., Ellis, P., Harris, A. L., Klijn, J. G., Foekens, J. A., Cardoso, F., Piccart, M. J., Buyse, M., Sotiriou, C. and TRANSBIG Consortium (2007). Strong

time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. Clinical cancer research : an official journal of the American Association for Cancer Research *13*, 3207–3214. 59, 69

Dews, M., Homayouni, A., Yu, D., Murphy, D., Sevignani, C., Wentzel, E., Furth, E. E., Lee, W. M., Enders, G. H., Mendell, J. T. and Thomas-Tikhonenko, A. (2006). Augmentation of tumor angiogenesis by a Myc-activated microRNA cluster. Nature genetics *38*, 1060–1065. 24

Di Leva, G., Gasparini, P., Piovan, C., Ngankeu, A., Garofalo, M., Taccioli, C., Iorio, M. V., Li, M., Volinia, S., Alder, H., Nakamura, T., Nuovo, G., Liu, Y., Nephew, K. P. and Croce, C. M. (2010). MicroRNA Cluster 221-222 and Estrogen Receptor alpha Interactions in Breast Cancer. J. Natl. Cancer Inst. *102*, 706–721. 28

Doench, J. G. and Sharp, P. A. (2004). Specificity of microRNA target selection in translational repression. Genes & development *18*, 504–511. 12, 13

Easow, G., Teleman, A. A. and Cohen, S. M. (2007). Isolation of microRNA targets by miRNP immunopurification. RNA (New York, N.Y.) *13*, 1198–1204. 13

Espelund, U., Cold, S., Frystyk, J., Ørskov, H. and Flyvbjerg, A. (2008). Elevated free IGF2 levels in localized, early-stage breast cancer in women. European journal of endocrinology / European Federation of Endocrine Societies *159*, 595–601. 100

Faha, B., Harlow, E. and Lees, E. (1993). The adenovirus E1A-associated kinase consists of cyclin E-p33cdk2 and cyclin A-p33cdk2. Journal of virology *67*, 2456–2465. 40

Farh, K. K.-H. K., Grimson, A., Jan, C., Lewis, B. P., Johnston, W. K., Lim, L. P., Burge, C. B. and Bartel, D. P. (2005). The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. Science (New York, N.Y.) *310*, 1817–1821. 16, 17

Fattaneh Tavassoli, P. D. (2003). Pathology and genetics of tumours of the breast and female genital organs. Oxford University Press, Lyon. 25

Felli, N., Fontana, L., Pelosi, E., Botta, R., Bonci, D., Facchiano, F., Liuzzi, F., Lulli, V., Morsilli, O., Santoro, S., Valtieri, M., Calin, G. A. A., Liu, C.-G. G., Sorrentino, A., Croce, C. M. and Peschle, C. (2005). MicroRNAs 221 and 222 inhibit normal erythropoiesis and erythroleukemic cell growth via kit receptor down-modulation. Proc Natl Acad Sci U S A *120*. 23

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E. and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806–811. 4

Friedman, R. C., Farh, K. K., Burge, C. B. and Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. Genome Research *19*, 92–105. 13

Fuchs, M., Gerber, J., Drapkin, R., Sif, S., Ikura, T., Ogryzko, V., Lane, W. S., Nakatani, Y. and Livingston, D. M. (2001). The p400 Complex Is an Essential E1A Transformation Target. Cell *106*, 297–307. 40

Fujioka, S., Shomori, K., Nishihara, K., Yamaga, K., Nosaka, K., Araki, K., Haruki, T., Taniguchi, Y., Nakamura, H. and Ito, H. (2009). Expression of minichromosome maintenance 7 (MCM7) in small lung adenocarcinomas (pT1): Prognostic implication. Lung cancer (Amsterdam, Netherlands) *65*, 223–229. 99

Fujita, S. and Iba, H. (2008). Putative promoter regions of miRNA genes involved in evolutionarily conserved regulatory systems among vertebrates. Bioinformatics *24*, 303–308. 101

Garofalo, M., Di Leva, G., Romano, G., Nuovo, G., Suh, S.-S., Ngankeu, A., Taccioli, C., Pichiorri, F., Alder, H. and Secchiero, P. (2009). miR-221&222 Regulate TRAIL Resistance and Enhance Tumorigenicity through PTEN and TIMP3 Downregulation. Cancer Cell *16*, 498–509. 25

Garzon, R., Pichiorri, F., Palumbo, T., Visentini, M., Aqeilan, R., Cimmino, A., Wang, H., Sun, H., Volinia, S., Alder, H., Calin, G. A., Liu, C.-G. G., Andreeff, M. and Croce, C. M. (2007). MicroRNA gene expression during retinoic acid-induced differentiation of human acute promyelocytic leukemia. Oncogene *26*, 4148–4157. 79

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H. and Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology *5*, R80. 32, 35, 43, 56

Ghosh, M. K. and Harter, M. L. (2003). A Viral Mechanism for Remodeling Chromatin Structure in G0 Cells. Molecular Cell *12*, 255–260. 40

Giaginis, C., Vgenopoulou, S., Vielh, P. and Theocharis, S. (2010). MCM proteins as diagnostic and prognostic tumor markers in the clinical setting. Histology and histopathology *25*, 351–370. 99

Grady, W. M., Parkin, R. K., Mitchell, P. S., Lee, J. H., Kim, Y.-H. H., Tsuchiya, K. D., Washington, M. K., Paraskeva, C., Willson, J. K., Kaz, A. M., Kroh, E. M., Allen, A., Fritz, B. R., Markowitz, S. D. and Tewari, M. (2008). Epigenetic silencing of the intronic microRNA hsa-miR-342 and its host gene EVL in colorectal cancer. Oncogene *27*, 3880–3888. 79, 90, 100, 104

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*, 235–240. 7

Griffiths-Jones, S. (2007). Annotating noncoding RNA genes. Annual review of genomics and human genetics *8*, 279–298. 5

Grimson, A., Farh, K. K., Johnston, W. K., Garrett-Engele, P., Lim, L. P. and Bartel, D. P. (2007). MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. Molecular Cell *27*, 91–105. 13, 14, 17, 18

Guil, S. and Caceres, J. F. (2007). The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. Nature Structural & Molecular Biology *14*, 591–596. 23

Ha, I., Wightman, B. and Ruvkun, G. (1996). A bulged lin-4/lin-14 RNA duplex is sufficient for Caenorhabditis elegans lin-14 temporal gradient formation. Genes & Development *10*, 3041–3050. 13

REFERENCES

Haase, A. D., Jaskiewicz, L., Zhang, H., Lainé, S., Sack, R., Gatignol, A. and Filipowicz, W. (2005). TRBP, a regulator of cellular PKR and HIV-1 virus expression, interacts with Dicer and functions in RNA silencing. EMBO reports *6*, 961–967. 9

Hahn, W. C., Counter, C. M., Lundberg, A. S., Beijersbergen, R. L., Brooks, M. W. and Weinberg, R. A. (1999). Creation of human tumour cells with defined genetic elements. Nature *400*, 464–468. 39

Hammond, S. M., Bernstein, E., Beach, D. and Hannon, G. J. (2000). An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. Nature *404*, 293–296. 4

Han, J., Lee, Y., Yeom, K.-H. H., Kim, Y.-K. K., Jin, H. and Kim, V. N. (2004). The Drosha-DGCR8 complex in primary microRNA processing. Genes & development *18*, 3016–3027. 7

Han, J., Lee, Y., Yeom, K.-H. H., Nam, J.-W. W., Heo, I., Rhee, J.-K. K., Sohn, S. Y. Y., Cho, Y., Zhang, B.-T. T. and Kim, V. N. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. Cell *125*, 887–901. 7

Han, J., Pedersen, J. S., Kwon, S. C., Belair, C. D., Kim, Y.-K. K., Yeom, K.-H. H., Yang, W.-Y. Y., Haussler, D., Blelloch, R. and Kim, V. N. (2009). Posttranscriptional crossregulation between Drosha and DGCR8. Cell *136*, 75–84. 7

Harris, L., Fritsche, H., Mennel, R., Norton, L., Ravdin, P., Taube, S., Somerfield, M. R., Hayes, D. F., Bast, R. C. and American Society of Clinical Oncology (2007). American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. Journal of clinical oncology *25*, 5287–5312. 25

He, L., Thomson, J. M., Hemann, M. T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., Cordon-Cardo, C., Lowe, S. W., Hannon, G. J. and Hammond, S. M. (2005). A microRNA polycistron as a potential human oncogene. Nature *435*, 828–833. 23

Hendrickson, D. G., Hogan, D. J., McCullough, H. L., Myers, J. W., Herschlag, D., Ferrell, J. E. and Brown, P. O. (2009). Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. PLoS biology *7*, e1000238+. 19

Hu, L.-D. D., Zou, H.-F. F., Zhan, S.-X. X. and Cao, K.-M. M. (2008). EVL (Ena/VASP-like) expression is up-regulated in human breast cancer and its relative expression level is correlated with clinical stages. Oncology reports *19*, 1015–1020. 100

Huh, M. S., Parker, M. H., Scimè, A., Parks, R. and Rudnicki, M. A. (2004). Rb is required for progression through myogenic differentiation but not maintenance of terminal differentiation. The Journal of cell biology *166*, 865–876. 40

Hynes, N. E. and Lane, H. A. (2005). ERBB receptors and cancer: the complexity of targeted inhibitors. Nature Reviews Cancer *5*, 341–354. 100

Hynes, N. E. and MacDonald, G. (2009). ErbB receptors and signaling pathways in cancer. Current opinion in cell biology *21*, 177–184. 26

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostat *4*, 249–264. 32, 45

Ivshina, A. V., George, J., Senko, O., Mow, B., Putti, T. C., Smeds, J., Lindahl, T., Pawitan, Y., Hall, P., Nordgren, H., Wong, J. E., Liu, E. T., Bergh, J., Kuznetsov, V. A. and Miller, L. D. (2006). Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer research *66*, 10292–10301. 59

Johnson, S. M., Grosshans, H., Shingara, J., Byrom, M., Jarvis, R., Cheng, A., Labourier, E., Reinert, K. L., Brown, D. and Slack, F. J. (2005). RAS Is Regulated by the let-7 MicroRNA Family. Cell *120*, 635–647. 23

Kapp, A. V., Jeffrey, S. S., Langerød, A., Børresen-Dale, A.-L. L., Han, W., Noh, D.-Y. Y., Bukholm, I. R., Nicolau, M., Brown, P. O. and Tibshirani, R. (2006). Discovery and validation of breast cancer subtypes. BMC genomics *7*, 231+. 27

Kapp, A. V. and Tibshirani, R. (2007). Are clusters found in one dataset present in another dataset? Biostatistics *8*, 9–31. 27

Khvorova, A., Reynolds, A. and Jayasena, S. D. (2003). Functional siRNAs and miRNAs Exhibit Strand Bias. Cell *115*, 209–216. 11

Kim, V. (2004). MicroRNA precursors in motion: exportin-5 mediates their nuclear export. Trends in Cell Biology *14*, 156–159. 8

Kim, V. N., Han, J. and Siomi, M. C. (2009). Biogenesis of small RNAs in animals. Nature reviews. Molecular cell biology *10*, 126–139. 6, 7

Kirshenbaum, L. A. and Schneider, M. D. (1995). Adenovirus E1A represses cardiac gene transcription and reactivates DNA synthesis in ventricular myocytes, via alternative pocket protein- and p300-binding domains. The Journal of biological chemistry *270*, 7791–7794. 39

Koralov, S., Muljo, S., Galler, G., Krek, A., Chakraborty, T., Kanellopoulou, C., Jensen, K., Cobb, B., Merkenschlager, M. and Rajewsky, N. (2008). Dicer Ablation Affects Antibody Diversity and Cell Survival in the B Lymphocyte Lineage. Cell *132*, 860–874. 24

Krützfeldt, J., Rajewsky, N., Braich, R., Rajeev, K. G., Tuschl, T., Manoharan, M. and Stoffel, M. (2005). Silencing of microRNAs in vivo with 'antagomirs'. Nature *438*, 685–689. 17

Lagos-Quintana, M., Rauhut, R., Lendeckel, W. and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. Science (New York, N.Y.) *294*, 853–858. 2, 3

Landthaler, M., Yalcin, A. and Tuschl, T. (2004). The human DiGeorge syndrome critical region gene 8 and Its D. melanogaster homolog are required for miRNA biogenesis. Current biology : CB *14*, 2162–2167. 7

Lau, N. C., Lim, L. P., Weinstein, E. G. and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science (New York, N.Y.) *294*, 858–862. 2, 3

REFERENCES

Lee, R. C. and Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. Science (New York, N.Y.) *294*, 862–864. 2, 3

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*, 843–854. 1, 19

Lee, Y., Jeon, K., Lee, J.-T. T., Kim, S. and Kim, V. N. (2002). MicroRNA maturation: stepwise processing and subcellular localization. The EMBO journal *21*, 4663–4670. 5

Lee, Y., Kim, M., Han, J., Yeom, K.-H. H., Lee, S., Baek, S. H. H. and Kim, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. The EMBO journal *23*, 4051–4060. 5

Lewis, B. P., Burge, C. B. and Bartel, D. P. (2005). Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. Cell *120*, 15–20. 12, 13

Lewis, B. P., Shih, I.-h. H., Jones-Rhoades, M. W., Bartel, D. P. and Burge, C. B. (2003). Prediction of mammalian microRNA targets. Cell *115*, 787–798. 12

Li, J., Smyth, P., Flavin, R., Cahill, S., Denning, K., Aherne, S., Guenther, S., O'Leary, J. and Sheils, O. (2007). Comparison of miRNA expression patterns using total RNA extracted from matched samples of formalin-fixed paraffin-embedded (FFPE) cells and snap frozen cells. BMC Biotechnology *7*, 36+. 30

Li, S. S., Xue, W. C., Khoo, U. S., Ngan, H. Y., Chan, K. Y., Tam, I. Y., Chiu, P. M., Ip, P. P., Tam, K. F. and Cheung, A. N. (2005). Replicative MCM7 protein as a proliferation marker in endometrial carcinoma: a tissue microarray and clinicopathological analysis. Histopathology *46*, 307–313. 99

Li, Y., Tan, W., Neo, T. W., Aung, M. O., Wasser, S., Lim, S. G. and Tan, T. M. (2009). Role of the miR-106b-25 microRNA cluster in hepatocellular carcinoma. Cancer science *100*, 1234–1242. 76

Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J., Bartel, D. P., Linsley, P. S. and Johnson, J. M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature *433*, 769–773. 19

Losada, A. and Hirano, T. (2005). Dynamic molecular linkers of the genome: the first decade of SMC proteins. Genes & development *19*, 1269–1287. 99

Lowery, A. J., Miller, N., Devaney, A., McNeill, R. E., Davoren, P. A., Lemetre, C., Benes, V., Schmidt, S., Blake, J., Ball, G. and Kerin, M. J. (2009). MicroRNA signatures predict oestrogen receptor, progesterone receptor and HER2/neu receptor status in breast cancer. Breast cancer research : BCR *11*, R27+. 28, 79

Lund, E., Güttinger, S., Calado, A., Dahlberg, J. E. and Kutay, U. (2004). Nuclear export of microRNA precursors. Science (New York, N.Y.) *303*, 95–98. 8

Macrae, I. J., Zhou, K., Li, F., Repic, A., Brooks, A. N., Cande, W. Z., Adams, P. D. and Doudna, J. A. (2006). Structural basis for double-stranded RNA processing by Dicer. Science (New York, N.Y.) *311*, 195–198. 8

McShane, L. M., Radmacher, M. D., Freidlin, B., Yu, R., Li, M.-C. and Simon, R. (2002). Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. Bioinformatics *18*, 1462–1469. 27

Meister, G. and Tuschl, T. (2004). Mechanisms of gene silencing by double-stranded RNA. Nature *431*, 343–349. 8, 10, 11

Mendell, J. T. (2008). miRiad roles for the miR-17-92 cluster in development and disease. Cell *133*, 217–222. 24, 27

Mertens-Talcott, S. U., Chintharlapalli, S., Li, X. and Safe, S. (2007). The oncogenic microRNA-27a targets genes that regulate specificity protein transcription factors and the G2-M checkpoint in MDA-MB-231 breast cancer cells. Cancer research *67*, 11001–11011. 28

Millar, A. A. and Waterhouse, P. M. (2005). Plant and animal microRNAs: similarities and differences. Functional & integrative genomics *5*, 129–135. 5

Miller, T. E., Ghoshal, K., Ramaswamy, B., Roy, S., Datta, J., Shapiro, C. L., Jacob, S. and Majumder, S. (2008). MicroRNA-221/222 confers tamoxifen resistance in breast cancer by targeting p27Kip1. The Journal of biological chemistry *283*, 29897–29903. 79

Monteys, A. M. M., Spengler, R. M., Wan, J., Tecedor, L., Lennox, K. A., Xing, Y. and Davidson, B. L. (2010). Structure and activity of putative intronic miRNA promoters. RNA (New York, N.Y.) *16*, 495–505. 5, 20, 29, 50, 55, 97, 101

Mook, S., Schmidt, M. K., Viale, G., Pruneri, G., Eekhout, I., Floore, A., Glas, A. M., Bogaerts, J., Cardoso, F., Piccart-Gebhart, M. J., Rutgers, E. T., Van't Veer, L. J. and TRANSBIG Consortium (2009). The 70-gene prognosis-signature predicts disease outcome in breast cancer patients with 1-3 positive lymph nodes in an independent validation study. Breast cancer research and treatment *116*, 295–302. 30

Moss, E. G., Lee, R. C. and Ambros, V. (1997). The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA. Cell *88*, 637–646. 2

Mouridsen, H., Gershanovich, M., Sun, Y., Pérez-Carrión, R., Boni, C., Monnier, A., Apffelstaedt, J., Smith, R., Sleeboom, H. P., Jänicke, F., Pluzanska, A., Dank, M., Becquart, D., Bapsy, P. P., Salminen, E., Snyder, R., Lassus, M., Verbeek, J. A., Staffler, B., Chaudri-Ross, H. A. and Dugan, M. (2001). Superior efficacy of letrozole versus tamoxifen as first-line therapy for postmenopausal women with advanced breast cancer: results of a phase III study of the International Letrozole Breast Cancer Group. Journal of clinical oncology *19*, 2596–2606. 26

Mu, L., Katsaros, D., Wiley, A., Lu, L., de la Longrais, I. R. A., Smith, S., Khubchandani, S., Sochirca, O., Arisio, R. and Yu, H. (2009). Peptide concentrations and mRNA expression of IGF-I, IGF-II and IGFBP-3 in breast cancer and their associations with disease characteristics. Breast cancer research and treatment *115*, 151–162. 100

Negrini, M., Rasio, D., Hampton, G. M., Sabbioni, S., Rattan, S., Carter, S. L., Rosenberg, A. L., Schwartz, G. F., Shiloh, Y. and Cavenee, W. K. (1995). Definition and refinement of chromosome 11 regions of

loss of heterozygosity in breast cancer: identification of a new region at 11q23.3. Cancer research *55*, 3003–3007. 28

Newman, M. A., Thomson, J. M. and Hammond, S. M. (2008). Lin-28 interaction with the Let-7 precursor loop mediates regulated microRNA processing. RNA (New York, N.Y.) *14*, 1539–1549. 50

Nicassio, F., Bianchi, F., Capra, M., Vecchi, M., Confalonieri, S., Bianchi, M., Pajalunga, D., Crescenzi, M., Bonapace, I. M. and Di Fiore, P. P. (2005). A cancer-specific transcriptional signature in human neoplasia. The Journal of clinical investigation *115*, 3015–3025. 32, 40

Nielsen, C. B., Shomron, N., Sandberg, R., Hornstein, E., Kitzman, J. and Burge, C. B. (2007). Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. RNA (New York, N.Y.) *13*, 1894–1910. 13, 18

Nishihara, K., Shomori, K., Fujioka, S., Tokuyasu, N., Inaba, A., Osaki, M., Ogawa, T. and Ito, H. (2008). Minichromosome maintenance protein 7 in colorectal cancer: implication of prognostic significance. International journal of oncology *33*, 245–251. 99

O'Donnell, K. A., Wentzel, E. A., Zeller, K. I., Dang, C. V. and Mendell, J. T. (2005). c-Myc-regulated microRNAs modulate E2F1 expression. Nature *435*, 839–843. 25

Oh, A. S., Lorant, L. A., Holloway, J. N., Miller, D. L., Kern, F. G. and El-Ashry, D. (2001). Hyperactivation of MAPK induces loss of ERalpha expression in breast cancer cells. Molecular endocrinology (Baltimore, Md.) *15*, 1344–1359. 28

Okamura, K., Hagen, J. W., Duan, H., Tyler, D. M. and Lai, E. C. (2007). The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. Cell *130*, 89–100. 8

Olsen, P. H. and Ambros, V. (1999). The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. Developmental biology *216*, 671–680. 19

Ørom, U. A., Nielsen, F. C. and Lund, A. H. (2008). MicroRNA-10a Binds the 5UTR of Ribosomal Protein mRNAs and Enhances Their Translation. Molecular Cell *30*, 460–471. 13

Ozsolak, F., Poling, L. L., Wang, Z., Liu, H., Liu, X. S., Roeder, R. G., Zhang, X., Song, J. S. and Fisher, D. E. (2008). Chromatin structure analyses identify miRNA promoters. Genes & Development *22*, 3172–3183. 20

Parker, J. S., Roe, S. M. and Barford, D. (2005). Structural insights into mRNA recognition from a PIWI domainsiRNA guide complex. Nature *434*, 663–666. 10

Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., Hayward, D. C., Ball, E. E., Degnan, B., Müller, 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*, 86–89. 3

Pawitan, Y., Bjöhle, J., Amler, L., Borg, A. L., Egyhazi, S., Hall, P., Han, X., Holmberg, L., Huang, F., Klaar, S., Liu, E. T., Miller, L., Nordgren, H., Ploner, A., Sandelin, K., Shaw, P. M., Smeds, J., Skoog, L., Wedrén, S. and Bergh, J. (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res *7*. 59, 62

Pawlowski, V., Révillion, F., Hebbar, M., Hornez, L. and Peyrat, J. P. (2000). Prognostic value of the type I growth factor receptors in a large series of human primary breast cancers quantified with a real-time reverse transcription-polymerase chain reaction assay. Clinical cancer research : an official journal of the American Association for Cancer Research *6*, 4217–4225. 100

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*, 747–752. 26

Petrocca, F., Vecchione, A. and Croce, C. M. (2008a). Emerging role of miR-106b-25/miR-17-92 clusters in the control of transforming growth factor beta signaling. Cancer research *68*, 8191–8194. 79

Petrocca, F., Visone, R., Onelli, M. R., Shah, M. H., Nicoloso, M. S., de Martino, I., Iliopoulos, D., Pilozzi, E., Liu, C.-G. and Negrini, M. (2008b). E2F1-Regulated MicroRNAs Impair TGF-Dependent Cell-Cycle Arrest and Apoptosis in Gastric Cancer. Cancer Cell *13*, 272–286. 24, 25, 76

Pineau, P., Volinia, S., McJunkin, K., Marchio, A., Battiston, C., Terris, B., Mazzaferro, V., Lowe, S. W., Croce, C. M. and Dejean, A. (2010). miR-221 overexpression contributes to liver tumorigenesis. Proceedings of the National Academy of Sciences of the United States of America *107*, 264–269. 23

Poliseno, L., Salmena, L., Riccardi, L., Fornari, A., Song, M. S. S., Hobbs, R. M., Sportoletti, P., Varmeh, S., Egia, A., Fedele, G., Rameh, L., Loda, M. and Pandolfi, P. P. P. (2010). Identification of the miR-106b˜25 microRNA cluster as a proto-oncogenic PTEN-targeting intron that cooperates with its host gene MCM7 in transformation. Science signaling *3*. 79

Pusztai, L., Mazouni, C., Anderson, K., Wu, Y. and Symmans, W. F. (2006). Molecular classification of breast cancer: limitations and potential. The oncologist *11*, 868–877. 27

R Development Core Team (2010). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria. ISBN 3-900051-07-0. 32, 43

Rakha, E. A., El-Sayed, M. E., Green, A. R., Paish, E. C., Powe, D. G., Gee, J., Nicholson, R. I., Lee, A. H., Robertson, J. F. and Ellis, I. O. (2007). Biologic and clinical characteristics of breast cancer with single hormone receptor positive phenotype. Journal of clinical oncology *25*, 4772–4778. 26

Reid, J. L., Bannister, A. J., Zegerman, P., Martinez-Balbas, M. A. and Kouzarides, T. (1998). E1A directly binds and regulates the P/CAF acetyltransferase. The EMBO Journal *17*, 4469–4477. 40

Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., Horvitz, H. R. and Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature *403*, 901–906. 2

REFERENCES

Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B. and Bartel, D. P. (2002). Prediction of Plant MicroRNA Targets. Cell *110*, 513–520. 11

Rhodes, A. (2003). Quality assurance in immunohistochemistry. The American journal of surgical pathology *27*. 26

Robbins, P., Pinder, S., de Klerk, N., Dawkins, H., Harvey, J., Sterrett, G., Ellis, I. and Elston, C. (1995). Histological grading of breast carcinomas: a study of interobserver agreement. Human pathology *26*, 873–879. 102

Ruby, G. G., Jan, C. H. and Bartel, D. P. (2007). Intronic microRNA precursors that bypass Drosha processing. Nature *448*, 83–86. 8

Saini, H. K., Griffiths-Jones, S. and Enright, A. J. (2007). Genomic analysis of human microRNA transcripts. Proceedings of the National Academy of Sciences *104*, 17719–17724. 5, 101

Santen, R. J., Song, R. X. X., McPherson, R., Kumar, R., Adam, L., Jeng, M.-H. H. and Yue, W. (2002). The role of mitogen-activated protein (MAP) kinase in breast cancer. The Journal of steroid biochemistry and molecular biology *80*, 239–256. 28

Satzger, I., Mattern, A., Kuettler, U., Weinspach, D., Voelker, B., Kapp, A. and Gutzmer, R. (2010). MicroRNA-15b represents an independent prognostic parameter and is correlated with tumor cell proliferation and apoptosis in malignant melanoma. International journal of cancer. Journal international du cancer *126*, 2553–2562. 79, 99

Schwarz, D. S., Hutvágner, G., Du, T., Xu, Z., Aronin, N. and Zamore, P. D. (2003). Asymmetry in the Assembly of the RNAi Enzyme Complex. Cell *115*, 199–208. 11

Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R. and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature *455*, 58–63. 13, 19

Sempere, L. F., Sokol, N. S., Dubrovsky, E. B., Berger, E. M. and Ambros, V. (2003). Temporal regulation of microRNA expression in Drosophila melanogaster mediated by hormonal signals and broad-Complex gene activity. Developmental biology *259*, 9–18. 2

Shi, Y., Sawada, J.-i., Sui, G., Affar, E. B., Whetstine, J. R., Lan, F., Ogawa, H., Po-Shan Luke, M., Nakatani, Y. and Shi, Y. (2003). Coordinated histone modifications mediated by a CtBP co-repressor complex. Nature *422*, 735–738. 40

Sikand, K., Slane, S. and Shukla, G. (2009). Intrinsic expression of host genes and intronic miRNAs in prostate carcinoma cells. Cancer Cell International *9*, 21+. 96

Slaby, O., Jancovicova, J., Lakomy, R., Svoboda, M., Poprach, A., Fabian, P., Kren, L., Michalek, J. and Vyzula, R. (2010). Expression of miRNA-106b in conventional renal cell carcinoma is a potential marker for prediction of early metastasis after nephrectomy. Journal of experimental & clinical cancer research : CR *29*. 79

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology *3*. 33, 45

Smyth, G. K. (2005). Limma: linear models for microarray data. In Bioinformatics and Computational Biology Solutions using R and Bioconductor, (Gentleman, R., Carey, V., Dudoit, S. and R. Irizarry, W. H., eds), pp. 397–420. Springer New York. 33, 45

Soerjomataram, I., Louwman, M. W., Ribot, J. G., Roukema, J. A. and Coebergh, J. W. W. (2008). An overview of prognostic factors for long-term survivors of breast cancer. Breast cancer research and treatment *107*, 309–330. 86

Sokol, N. S., Xu, P., Jan, Y.-N. N. and Ambros, V. (2008). Drosophila let-7 microRNA is required for remodeling of the neuromusculature during metamorphosis. Genes & development *22*, 1591–1596. 2

Song, J.-J., Smith, S. K., Hannon, G. J. and Joshua-Tor, L. (2004). Crystal Structure of Argonaute and Its Implications for RISC Slicer Activity. Science *305*, 1434–1437. 10

Soon, P. S. H. S., Tacon, L. J., Gill, A. J., Bambach, C. P., Sywak, M. S., Campbell, P. R., Yeh, M. W., Wong, S. G., Clifton-Bligh, R. J., Robinson, B. G. and Sidhu, S. B. (2009). miR-195 and miR-483-5p Identified as Predictors of Poor Prognosis in Adrenocortical Cancer. Clinical cancer research : an official journal of the American Association for Cancer Research *15*. 79

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., Eystein Lønning, P. and Børresen-Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences of the United States of America *98*, 10869–10874. 27

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lønning, P. E., Brown, P. O., Børresen-Dale, A.-L. and Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the National Academy of Sciences of the United States of America *100*, 8418–8423. 27

Sotiriou, C., Neo, S.-Y. Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., Fox, S. B., Harris, A. L. and Liu, E. T. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proceedings of the National Academy of Sciences of the United States of America *100*, 10393–10398. 27

Sotiriou, C. and Pusztai, L. (2009). Gene-expression signatures in breast cancer. The New England journal of medicine *360*, 790–800. 26

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Van de Vijver, M. J., Bergh, J., Piccart, M. and Delorenzi, M. (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. Journal of the National Cancer Institute *98*, 262–272. 59, 64, 102

Sprent, P. and Smeeton, N. C. (2007). Applied Nonparametric Statistical Methods, Fourth Edition. Chapman & Hall/CRC Texts in Statistical Science. 95

REFERENCES

Stark, A., Brennecke, J., Bushati, N., Russell, R. and Cohen, S. (2005). Animal MicroRNAs Confer Robustness to Gene Expression and Have a Significant Impact on 3UTR Evolution. Cell *123*, 1133–1146. 17

Stern-Ginossar, N., Elefant, N., Zimmermann, A., Wolf, D. G., Saleh, N., Biton, M., Horwitz, E., Prokocimer, Z., Prichard, M., Hahn, G., Goldman-Wohl, D., Greenfield, C., Yagel, S., Hengel, H., Altuvia, Y., Margalit, H. and Mandelboim, O. (2007). Host Immune System Gene Targeting by a Viral miRNA. Science *317*, 376–381. 13

Stoner, M., Saville, B., Wormke, M., Dean, D., Burghardt, R. and Safe, S. (2002). Hypoxia induces proteasome-dependent degradation of estrogen receptor alpha in ZR-75 breast cancer cells. Molecular endocrinology (Baltimore, Md.) *16*, 2231–2242. 28

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America *102*, 15545–15550. 46

Tiainen, M., Spitkovsky, D., Jansen-Dürr, P., Sacchi, A. and Crescenzi, M. (1996). Expression of E1A in terminally differentiated muscle cells reactivates the cell cycle and suppresses tissue-specific genes by separable mechanisms. Molecular and cellular biology *16*, 5302–5312. 39, 40, 41

Tie, J., Pan, Y., Zhao, L., Wu, K., Liu, J., Sun, S., Guo, X., Wang, B., Gang, Y., Zhang, Y., Li, Q., Qiao, T., Zhao, Q., Nie, Y. and Fan, D. (2010). MiR-218 Inhibits Invasion and Metastasis of Gastric Cancer by Targeting the Robo1 Receptor. PLoS Genet *6*, e1000879+. 79

Tomari, Y. and Zamore, P. D. (2005). Perspective: machines for RNAi. Genes & development *19*, 517–529. 8, 10, 11

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America *98*, 5116–5121. 20

Van Der Haegen, B. and Shay, J. (1993). Immortalization of human mammary epithelial cells by SV40 large T-antigen involves a two step mechanism. In Vitro Cellular & amp; Developmental Biology - Animal *29*, 180–182. 39

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature *415*, 530–536. 29, 30

Ventura, A., Young, A. G., Winslow, M. M., Lintault, L., Meissner, A., Erkeland, S. J., Newman, J., Bronson, R. T., Crowley, D., Stone, J. R., Jaenisch, R., Sharp, P. A. and Jacks, T. (2008). Targeted deletion reveals essential and overlapping functions of the miR-17 through 92 family of miRNA clusters. Cell *132*, 875–886. 24

Veronese, A., Lupini, L., Consiglio, J., Visone, R., Ferracin, M., Fornari, F., Zanesi, N., Alder, H., D'Elia, G., Gramantieri, L., Bolondi, L., Lanza, G., Querzoli, P., Angioni, A., Croce, C. M. and Negrini, M. (2010). Oncogenic role of miR-483-3p at the IGF2/483 locus. Cancer research *70*, 3140–3149. 79

Wang, X., Xuan, Z., Zhao, X., Li, Y. and Zhang, M. Q. (2009). High-resolution human core-promoter prediction with CoreBoost_HM. Genome Research *19*, 266–275. 20

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., Jatkoe, T., Berns, E. M., Atkins, D. and Foekens, J. A. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet *365*, 671–679. 59

Wang, Y., Medvid, R., Melton, C., Jaenisch, R. and Blelloch, R. (2007). DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. Nature genetics *39*, 380–385. 7

Whyte, P., Buchkovich, K. J., Horowitz, J. M., Friend, S. H., Raybuck, M., Weinberg, R. A. and Harlow, E. (1988). Association between an oncogene and an anti-oncogene: the adenovirus E1A proteins bind to the retinoblastoma gene product. Nature *334*, 124–129. 40

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*, 855–862. 1, 19

Xia, H., Qi, Y., Ng, S. S., Chen, X., Chen, S., Fang, M., Li, D., Zhao, Y., Ge, R., Li, G., Chen, Y., He, M.-L. L., Kung, H.-f. F., Lai, L. and Lin, M. C. (2009). MicroRNA-15b regulates cell cycle progression by targeting cyclins in glioma cells. Biochemical and biophysical research communications *380*, 205–210. 79

Yan, L., Yang, X. and Davidson, N. E. (2001). Role of DNA methylation and histone acetylation in steroid receptor expression in breast cancer. Journal of mammary gland biology and neoplasia *6*, 183–192. 28

Yan, L.-X. X., Huang, X.-F. F., Shao, Q., Huang, M.-Y. Y., Deng, L., Wu, Q.-L. L., Zeng, Y.-X. X. and Shao, J.-Y. Y. (2008). MicroRNA miR-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis. RNA (New York, N.Y.) *14*, 2348–2360. 28

Yigit, E., Batista, P. J., Bei, Y., Pang, K. M., Chen, C.-C. G., Tolia, N. H., Joshua-Tor, L., Mitani, S., Simard, M. J. and Mello, C. C. (2006). Analysis of the C. elegans Argonaute Family Reveals that Distinct Argonautes Act Sequentially during RNAi. Cell *127*, 747–757. 9

Yu, H. and Rohan, T. (2000). Role of the insulin-like growth factor family in cancer development and progression. Journal of the National Cancer Institute *92*, 1472–1489. 100

Yu, J., Wang, F., Yang, G., Wang, F., Ma, Y., Du, Z. and Zhang, J. (2006). Human microRNA clusters: Genomic organization and expression profile in leukemia cell lines. Biochemical and Biophysical Research Communications *349*, 59–68. 5

Yuan, Y.-R., Pei, Y., Ma, J.-B., Kuryavyi, V., Zhadina, M., Meister, G., Chen, H.-Y., Dauter, Z., Tuschl, T. and Patel, D. J. (2005). Crystal Structure of A. aeolicus Argonaute, a Site-Specific DNA-Guided

Endoribonuclease, Provides Insights into RISC-Mediated mRNA Cleavage. Molecular Cell *19*, 405–419. 10

Zeng, Y. and Cullen, B. R. (2005). Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences. The Journal of biological chemistry *280*, 27595–27603. 7

Zhang, H., Kolb, F. A., Jaskiewicz, L., Westhof, E. and Filipowicz, W. (2004). Single Processing Center Models for Human Dicer and Bacterial RNase III. Cell *118*, 57–68. 8