MOLECULAR ONCOLOGY 9 (2015) 473–487

available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/molonc

Technical Note

Mining cancer gene expression databases for latent information on intronic microRNAs

Simona Monterisi[a], Giovanni D'Ario[a,1], Elisa Dama[a,b], Nicole Rotmensz[b], Stefano Confalonieri[a,c], Chiara Tordonato[a], Flavia Troglio[a], Giovanni Bertalot[a], Patrick Maisonneuve[b], Giuseppe Viale[d,e], Francesco Nicassio[a,c,f], Manuela Vecchi[a,c], Pier Paolo Di Fiore[a,c,g,**,2], Fabrizio Bianchi[a,*,2]

[a]Molecular Medicine Program, Department of Experimental Oncology, European Institute of Oncology, Milan, Italy
[b]Division of Epidemiology and Biostatistics, European Institute of Oncology, Milan, Italy
[c]IFOM, The FIRC Institute for Molecular Oncology Foundation, Milan, Italy
[d]Division of Pathology, European Institute of Oncology, Milan, Italy
[e]School of Medicine, University of Milan, Milan, Italy
[f]Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia (IIT), Milan, Italy
[g]Department of Scienze della Salute, University of Milan, Milan, Italy

ARTICLE INFO

Article history:
Received 3 June 2014
Received in revised form
23 September 2014
Accepted 2 October 2014
Available online 15 October 2014

Keywords:
MicroRNA
Cancer
Gene expression
Breast cancer

ABSTRACT

Around 50% of all human microRNAs reside within introns of coding genes and are usually co-transcribed. Gene expression datasets, therefore, should contain a wealth of miRNA-relevant latent information, exploitable for many basic and translational research aims. The present study was undertaken to investigate this possibility. We developed an *in silico* approach to identify intronic-miRNAs relevant to breast cancer, using public gene expression datasets. This led to the identification of a miRNA signature for aggressive breast cancer, and to the characterization of novel roles of selected miRNAs in cancer-related biological phenotypes. Unexpectedly, in a number of cases, expression regulation of the intronic-miRNA was more relevant than the expression of their host gene. These results provide a proof of principle for the validity of our intronic miRNA mining strategy, which we envision can be applied not only to cancer research, but also to other biological and biomedical fields.

* *Corresponding author*. European Institute of Oncology, Department of Experimental Oncology, Via Ripamonti, 435, 20141 Milan, Italy. Tel.: +39 02 94375173; fax: +39 02 94379305.
** *Corresponding author*. European Institute of Oncology, Department of Experimental Oncology, Via Ripamonti, 435, 20141 Milan, Italy. Tel.: +39 02 57489832; fax: +39 02 94375991.
E-mail addresses: pierpaolo.difiore@ieo.eu (P.P. Di Fiore), fabrizio.bianchi@ieo.eu (F. Bianchi).
[1] Present address: Swiss Institute of Bioinformatics, Lausanne, Switzerland.
[2] These two authors are equal last authors.
http://dx.doi.org/10.1016/j.molonc.2014.10.001

Abbreviations

FFPE	formalin-fixed paraffin-embedded
PCR	polymerase chain reaction
FDR	false discovery rate
qRT-PCR	quantitative reverse transcriptase PCR
GEO	gene expression omnibus
ER	estrogen receptor
HER2 (ErbB2)	human epidermal growth factor receptor 2
IHC	immunohistochemistry

1. Introduction

MicroRNAs are small, non-coding RNA molecules (18–22 nucleotides in length) that function as endogenous triggers of the mRNA interference pathway and are involved in the regulation of pleiotropic biological functions (Krol et al., 2010; Yendamuri and Kratzke, 2011). Almost 50% of all human miRNA genes are located within introns of host genes, with which they usually share transcriptional regulation (Baskerville and Bartel, 2005; Griffiths-Jones, 2007; He et al., 2012; Monteys et al., 2010; Ozsolak et al., 2008; Rainer et al., 2009; Rodriguez et al., 2004; Saini et al., 2007). In principle, this property could be exploited to predict the expression of intronic miRNAs (int-miRNAs) through the analysis of the expression of their host genes (miRNA host genes, miR-HG). Similar approaches have already been successfully employed to identify miRNA target genes, to predict miRNA tissue expression, and to characterize miRNA/miR-HG feedback loops (Lutter et al., 2010; Radfar et al., 2011; Wang et al., 2009).

The major potential stemming from the mode of regulation of int-miRNAs is, however, untapped. In recent years, enormous effort has been dedicated to the profiling of various physiological and pathological conditions at the transcriptomic (mRNA) level. While almost every field of biology and biomedicine has been explored through this approach, cancer biology is arguably the field in which the highest investment has been made, with the dual objective of: i) obtaining a global view of cancer processes by systems-based analysis (Basso et al., 2005; Minn et al., 2005; Sweet-Cordero et al., 2005); ii) identifying biomarkers for improved management of cancer patients (Ivshina et al., 2006; Sorlie et al., 2001; Sotiriou et al., 2003; van 't Veer et al., 2002). As a result, thousands of human tumors have been profiled and the datasets made publicly available, frequently associated with high quality clinical information. In our view, these datasets are amenable to mining "latent" information on int-miRNA expression.

There is growing interest in miRNAs, both as potential cancer determinants and biomarkers (Calin and Croce, 2006). From a general perspective, miRNA profiling might be advantageous over mRNA profiling, since the complexity of miRNome is at least 20-fold lower than that of a reference transcriptome (if one makes the somewhat rough comparison of ~1000 miRNAs *vs.* ~ 20,000 genes). This means that sufficient statistical power can be reached with a much lower number of analyzed samples. This is particular relevant to studies, such as those involving human pathological samples, in which genetic variability represents a relevant confounding factor.

Thus, the explicit goal of this study was to exploit cancer datasets, in particular, breast cancer datasets, to provide a proof of principle that meta-analysis of miR-HG expression profiles can accurately identify int-miRNAs that are relevant to cancer, both in terms of their potential utility as biomarkers and their role in breast cancer cell biology.

2. Material and methods

2.1. Patient selection criteria

Written informed consent for research use of biological samples was obtained from all patients. Patients underwent surgery at the European Institute of Oncology between 1998 and 2010. Only tumor samples with a cellularity >70% were included in the study.

2.2. Affymetrix microarray analysis

Retroviral infection of the MCF10A cell line with the SV40-large T antigen was performed using a pBABE-neo retroviral vector. After 48 h of infection, cells were collected and total RNA extracted using the RNeasy Mini Kit (QIAGEN). RNA quality was controlled using the 2100 Bioanalyzer (Agilent). Total RNA (5 μg) was then retrotranscribed into double stranded cDNA using SuperScript® Double-Stranded cDNA Synthesis Kit (Invitrogen).

In vitro anti-sense RNA transcription was performed through an Eberwine's modified *in vitro* transcription reaction (MEGAscript, Ambion) using labeled rNTP (Enzo® BioArray™ HighYield™ RNA Transcript Labeling Kit, ENZO Biolabs). Briefly, we added 14.5 μl of rNTPs mix, 2 μl of T7 polymerase and 2 μl of reaction buffer to 1.5 μl of purified cDNA, and incubated the reaction mix at 37 °C for 6 h. Labeled cRNA was then fragmented (30-200 base fragments), checked by agarose gel, and hybridized on Human Genome U133A 2.0 Arrays in duplicate for each condition (i.e., MCF10A SV40-large T, and MCF10A pBABE-empty).

Data were normalized using the Robust Multi-array Average (RMA) method. Information on human int-miRNAs, associated host genes and mature miRNA sequences was retrieved from miRBase v13.0. 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 using the Bioconductor *hgu133a2.db* annotation package. Differentially expressed probe sets were identified using the limma Bioconductor package. *P*-values were adjusted using the Benjamini–Hochberg correction.

2.3. Bioinformatics analysis of external Affymetrix datasets

Breast cancer microarray datasets and associated clinical information were downloaded from the Gene Expression Omnibus database (GEO, http://www.ncbi.nlm.nih.gov/geo/). The accession numbers of the datasets used are GSE1456, GSE2990, GSE7390, and GSE4922. All datasets were based on

the GeneChip® Human Genome U133A to avoid batch bias effects during the analysis.

We applied a quality control procedure on CEL files to identify flawed arrays using Relative Log Expression (RLE) values and Normalized Unscaled Standard Error (NUSE) values (Bolstad et al., 2004; Gentleman et al., 2004). For each array, we computed the median value and the interquartile range (IQR) of both the NUSE and RLE statistics. We then calculated the IQRs across the arrays for each dataset. Arrays were rejected if IQR values were $>$q3 + 1.5IQRs or $<$ q1 − 1.5IQRs, where q1 and q3 are the first and third quartile, respectively.

This filtering resulted in the exclusion of 47 arrays in the GSE4922 dataset, 9 arrays in the GSE1456 dataset, 9 arrays in the GSE7390 dataset, and 5 arrays in the GSE2990 dataset. In addition, in the GSE2990 dataset we considered only the "Uppsala" cohort of patients because of the low signal intensity distribution of several arrays of the "Oxford" cohort compared to the "Uppsala" cohort, which determined a batch bias effect (Figure S1). Data were normalized using the RMA method.

Information relative to human int-miRNAs and associated host genes was retrieved from miRBase (www.mirbase.org, release 13.0). Probe sets were filtered for signal intensity using the Bioconductor genefilter package. Only probe sets that had a normalized signal greater than 150 (7.2 on the log2 scale) in at least 10% of the samples were retained for further analysis. Differentially expressed probe sets were identified using the limma Bioconductor package. All *P*-values were adjusted using the Benjamini–Hochberg correction.

Monte Carlo simulation was performed for each dataset in the following manner: 1) all miRNA-associated probe sets were excluded from the dataset; 2) we randomly selected *n* probe sets, where *n* is the number of miRNA-associated probe sets ($n = 422$); 3) the number of probe sets significantly regulated between G3 *vs.* G1 and/or ER + *vs.* ER-tumors were annotated; 4) we repeated steps '2' and '3' 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. Expression dataset Breast subtype analysis was performed using the TCGA breast cancer (October 2012 release, 599 patients) downloaded from the TCGA web data portal (https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp). Data were gene centered on relative medians and log2 transformed before clustering analysis.

Pathway analysis was performed using the online available webtool Ingenuity Pathway Analysis (IPA) (http://www.ingenuity.com/). Predicted and experimentally validated miRNA target gene sets were obtained using miRWalk database (Dweep et al., 2011). MicroRNA target prediction was performed using four different methods: miRanda, miRDB, miRWalk and Targetscan. Genes predicted in 4 out of 4 methods were retained for subsequent IPA analysis.

2.4. RNA isolation and RT-PCR

Total RNA was extracted from cell lines using the TRIZOL reagent (Invitrogen) or from FFPE archival breast tumor samples (with a tumor cellularity $>$70%) using the RNAeasy FFPE kit (QIAGEN). RNA was quantified by Nanodrop (Agilent Technologies).

miRNA expression profiles of MCF10A cells were obtained using the TaqMan® Low Density Array microRNA Signature Panel (v1.0; Applied Biosystems) and reactions were carried out on an Applied Biosystems 7900HT thermocycler using the manufacturer's recommended cycling conditions.

miRNA expression profiles of FFPE archival breast tumor samples or of MDA-MB-231 and MDA-MB-361 cells were obtained using miScript Primer Assays and the miScript SYBR Green PCR Kit (Qiagen). Total RNA (400 ng) was reverse transcribed using the miScript Reverse Transcriptase kit (Qiagen) according to manufacturer's instructions. Briefly, the two-step protocol involves reverse transcription of miRNA to cDNA using miRNA-specific primers followed by qRT-PCR. Reactions were run in duplicate using 5 ng of cDNA as template in 20 μl final reaction volume. All probes were normalized to U5A for FFPE archival breast tumor samples or to *SNORD25* for breast cancer cell lines, as an internal control. Amplification reactions were performed with LightCycler 480 (ROCHE) using the manufacturer's recommended cycling conditions. Relative expression ratios of miRNAs were obtained using the 2^{-ddCT} method.

2.5. Cell lines and infection

The MDA-MB-231 and MDA-MB-361 breast cancer cell lines were grown in Dulbecco's Modified Eagle Medium supplemented with 10% fetal bovine serum, 2 mM L-glutamine, 100 U/ml penicillin, and 100 U/ml streptomycin at 37 °C in a humidified incubator with 5% CO_2. Mature miRNAs were overexpressed using GFP Lenti-miR™ vectors obtained from Systems Biosciences (SBI, Mountain View, CA, USA). The pCDH-CMV-MCS-EF1-copGFP was used as a control vector (SBI, Mountain View, CA, USA). GFP positive cells were sorted with a BD Influx™ cell sorter (BD Bioscience, San Jose, CA, USA). Host genes overexpression was achieved by Lipofectamine® 2000 transfection (LifeTechnologies) of pCMV-Sport6-*MYO5C*, pCMV-Sport6-*EVL* and pDEST26-*IGF2* vectors (derived from the original pENTR221-*IGF2* vector) provided by Life Science (Source BioScience, Nottingham, UK). Gene expression was verified by qRT-PCR using QuantiTect Primer Assay (Qiagen), or, for *MYO5C*, using custom primers: *MYO5C* − Forward, GAATCTCTGCCTCCACTTCG; *MYO5C* − Reverse, GATAGCTGAGAGCCGTGAGG. miR-HG expression was normalized to GUSB expression as an internal control.

2.6. Cell proliferation and colony forming assays

For proliferation assays, MDA-MB-231 and MDA-MB-361 cells were seeded in triplicate in 6-well plates (BD Falcon™ 353046) at 4×10^4 and 25×10^4 cells/well, respectively. Bio-Rad TC10 automated cell counter was used to count cells every 24 h for 4 days. The first measurement was taken at 24 h for MDA-MB-231 or at 72 h for MDA-MB-361 after seeding cells. Colony forming assays were performed by seeding 5000 cells/type in 10-cm plates (BD Falcon™ 353003) and then incubating plates for 10 days. Colony formation was visualized by staining for 5 min at RT with crystal violet (1% w/v in 35% EtOH, Santa Cruz).

2.7. Wound-healing assay

Time-lapse video microscopy was performed as described previously (Palamidessi et al., 2008; White et al., 2007) with slight modifications. Confluent monolayers of MDA-MB-231 or MDA-MB-361 cells in 12-well plates (BD Falcon™ 353043) were wounded with a plastic pipette tip to induce migration into the wound. Cells were placed on the stage of an inverted motorized microscope (Leica AF600) in a cage incubator (Okolab) at 37 °C and 0% CO_2 for time-lapse video microscopy. Phase-contrast images were collected every 20 min over a 12-h period. Videos were generated using the ImageJ software for image analysis. Cell trajectories were determined using the MTrackJ plugin of ImageJ (Meijering et al., 2012). The distance covered by each cell and the migration speed were extracted from the track plots. Fifteen cells from 3 independent experiments were analyzed for each condition, and data are expressed in micrometers as mean ± σ.

2.8. Statistical analysis

The significance of the overlap between gene lists was based on the hypergeometric distribution (Fury et al., 2006). The extension to the case of more than two overlapping lists was based on marginalization of joint probability and chain rule of probability.

RT-PCR gene expression analysis and relative statistical analyses were performed using JMP 10.0 64-bit edition (SAS Institute Inc.). Statistical analyses were performed on log2 data using parametric tests (t-test, ANOVA). Cluster 3.0 for Mac OS X (http://bonsai.hgc.jp/~mdehoon/software/cluster/) and Java Treeview (http://jtreeview.sourceforge.net) were used for the hierarchical clustering analysis. Uncentered correlation and centroid clustering methods were used on log2 median centered data. The multivariate model to predict risk of having an aggressive tumor subtype was built using diagonal linear discriminant analysis (DLDA) with BRB ArrayTools (http://linus.nci.nih.gov/BRB-ArrayTools.html). Briefly, the model assigned a risk index to every patient and classified them as high- or low-risk of having an aggressive tumor subtype based on linear combination of gene expression values weighted by coefficients calculated during training of the classifier. The critical cutoff value to predict high-/low-risk was identified by the receiver operating characteristics curve analysis (ROC) using JMP 10.0 software, and it was set at −1.23. In the training set, twenty-eight out of the 29 LuA tumors (~97%) were predicted as 'low risk', which is consistent with the reported low metastatic behavior of LuA breast tumors (Voduc et al., 2010). In contrast, 43 out of the 66 patients with the more aggressive LuB tumor subtype (~65%) were predicted as 'high risk' ($P < 0.0001$). Overall, the test displayed an accuracy, sensitivity and specificity of ~75%, ~65% and ~97%, respectively, in the training set, and it was independent from other risk factors such as nodal status, tumor size (pT), and HER2 positivity (Table S1).

Multivariate nominal logistic regression of miRNA risk class was performed using JMP 10.0 software. The Odds ratio of miRNA high-risk class was adjusted for ErbB2 (HER2), nodal status and tumor size (pT).

Table 1 – Overlapping miR-HGs in the different analyses. Overlapping miR-HGs (highlighted in grey scale) in G3 *vs.* G1 and ER+ *vs.* ER- analyses ($P < 0.05$, Benjamini–Hochberg correction). Dataset, Gene Expression Omnibus accession numbers of the Affymetrix datasets. N, total number of unique miR-HGs found regulated in the relative dataset. %, percentage of miR-HGs regulated out of the total miR-HGs mapped on the array ($N = 243$). N overlapping, number of miR-HGs commonly found regulated in highlighted datasets (light/dark grey areas). *P*, significance of the overlaps.

G3/G1

Dataset	N	% (N/243)	N overlapping		
GSE4922	110	45	53 P<0.0001	47 P<0.0001	25 P<0.0001
GSE1456	56	23			
GSE2990	78	32			
GSE7390	35	14			

ER+/ER-

Dataset	N	% (N/243)	N overlapping	
GSE4922	97	40	27 P<0.0001	17 P<0.0001
GSE2990	29	12		
GSE7390	43	18		

3. Results

3.1. An in silico approach to extract information on the regulation of int-miRNAs from microarray gene expression (mRNA) datasets

Several studies have indicated that there is a good correlation between the expression of miR-HGs and their respective int-miRNAs (Baskerville and Bartel, 2005; Wang et al., 2009; Wang and Li, 2009). Therefore, the first step in our strategy was the development of an *in silico* approach to predict int-miRNA expression by means of microarray-based analysis of miR-HG expression profiles. To this end, we used a controllable and syngeneic model system, the non-transformed breast cell line MCF10A infected with the Simian virus 40 (SV40) large T antigen, which causes cell transformation and alters the expression of numerous genes (Carbone et al., 1997; De Luca et al., 1997; Girardi et al., 1962). We compared the mRNA and miRNA expression profiles of these MCF10A-SV40 cells to those of mock-infected MCF10A cells (see Methods). miR-HGs were identified by mapping the genomic position of 706 known human miRNA precursors (pri-miRNAs) to the genomic coordinates of the entire human genome. This resulted in the identification of 317 pri-miRNAs located within the introns of 269 unique miR-HGs (Table S2). Affymetrix gene expression analysis revealed 43 miR-HGs differentially expressed in MCF10A-SV40 *vs.* control cells (FDR<10%, Table S3), which contain 51 pri-miRNAs in their introns, corresponding to 84 mature int-miRNAs. We validated the analysis, by directly measuring the expression of 47 of these miRNAs, in the same samples, by qRT-PCR (using available TaqMan assays). Of the 47 int-miRNAs, 38 were detectable by qRT-PCR, and of these 31 were congruently regulated with their miR-

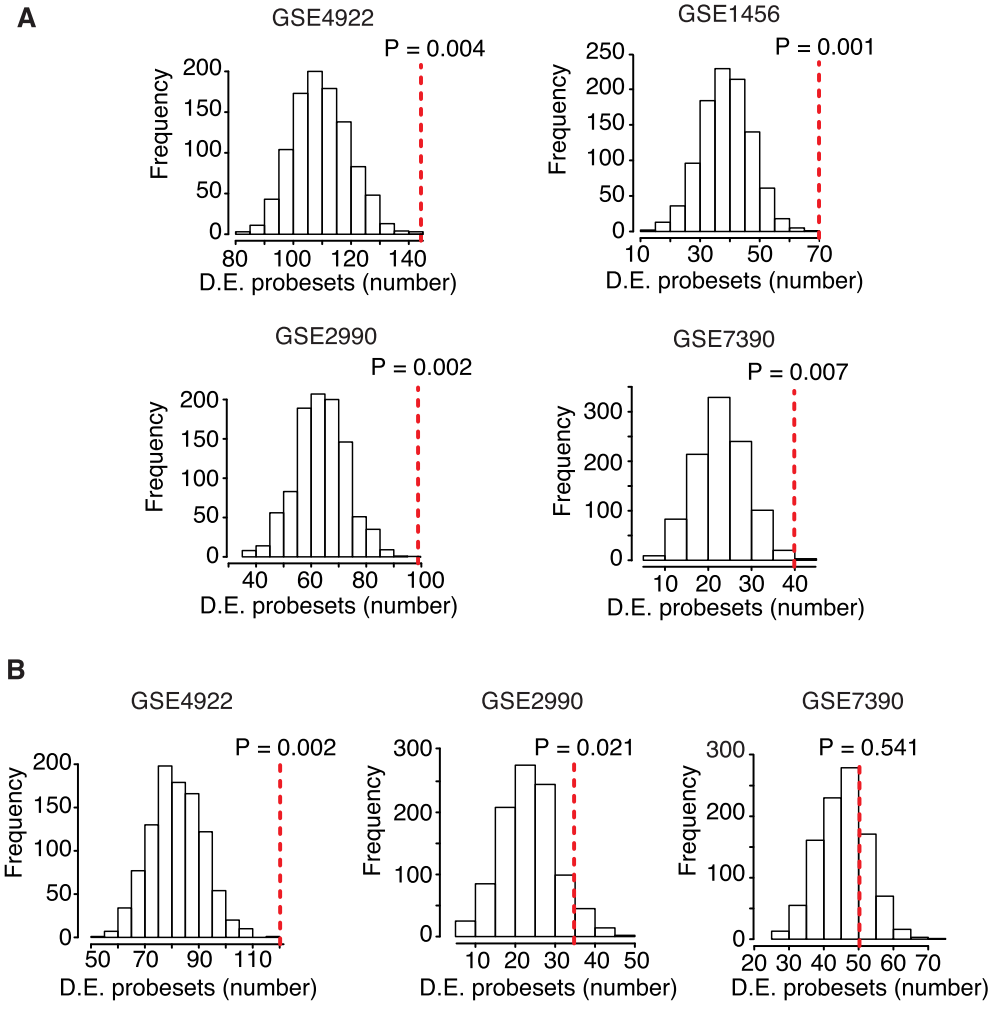


Figure 1 – Global gene expression profiles of miR-HGs in breast cancer. A. Results of the Monte Carlo simulation for the G3 *vs.* G1 breast cancer Affymetrix meta-analysis. For this simulation, we used 1000 lists of randomly selected genes that were of the same size as the original list of miR-HGs, and which contained genes that, to date, have not been associated with any int-miRNA ("non-host genes"). Histograms represent the distribution of the random lists (1000 random lists of 422 probe sets) in the indicated GEO datasets. *X*-axes: number of significantly regulated probe sets (D.E., differentially expressed probe sets) found in the random lists ($P < 0.05$; Benjamini–Hochberg correction). *Y*-axes: frequency of random lists. Red dashed lines indicate the position, within the distributions, of the miR-HG list (422 probe sets) identified through our analysis. *P*-values indicate the fraction of random lists having an equal or larger number of significantly regulated probe sets compared to the miR-HG list. B. Results of the Monte Carlo simulation for the ER+ *vs.* ER-breast cancer Affymetrix analysis. Results are presented as in (A).

HGs (Table S4). The positive correlation between the expression of miR-HGs and int-miRNAs was significant when analyzed both qualitatively (congruent direction of regulation; $P = 0.02$, Figure S2A) and quantitatively (expression ratios; $P = 0.0003$, $R = 0.4$, Figure S2B). In contrast, the 74 int-miRNAs (for which TaqMan assays were available) within introns of unregulated miR-HGs (FDR>10%), did not show significant co-regulation with their cognate miR-HGs ($P = 0.2$ and $P = 0.5$; Figure S2C and D), possibly due to a different post-transcriptional regulation of mRNA and miRNA cognate transcripts.

In summary, these data indicate that our bioinformatics approach can be used to infer int-miRNA regulation through host gene expression profile analysis.

3.2. In silico *prediction of miR-HGs differentially regulated in breast cancer*

We next analyzed the expression of miR-HGs in breast cancer through a meta-analysis of expression datasets of 666 patients, from four independent studies with clinical and pathological information, and raw data available through the GEO database (Table S5). We mapped 243 unique miR-HGs (containing 264 int-miRNAs) whose expression data were present in the datasets (Table S6). Through the comparison of tumors with different clinical and pathological parameters, we identified a significant number of differentially expressed miR-HGs ($P < 0.05$, Benjamini–Hochberg correction; Table S7A and B), especially in the comparisons between poorly differentiated (G3) and well-differentiated (G1) tumors, or between estrogen receptor positive (ER+) and estrogen receptor negative (ER−) tumors. In the first instance (G3 *vs.* G1), 14–45% of all miR-HGs (depending on the considered dataset, Table 1) were differentially regulated; in the second case (ER + *vs.* ER−) the differential expression of miR-HGs was 12–40% (Table 1).

It has been reported that the differences in the transcriptomic profiles of different types of breast cancer (i.e. G3 *vs.* G1, or ER−*vs.* ER+) are so vast that regulation of a set of genes of interest might simply reflect large-scale transcriptional changes (Ivshina et al., 2006; Sotiriou et al., 2003, 2006), and even that a significant number of randomly chosen "signatures" may have prognostic value (Venet et al., 2011). Thus, to determine whether the large fraction of miR-HGs differentially regulated in breast cancer was not the mere result of large-scale transcriptional changes, we performed a Monte Carlo simulation using 1000 random lists of non-host genes (which represent more than 95% of the entire genome). In G3 *vs.* G1 tumors, there was a significant enrichment of differentially regulated miR-HGs, with respect to non-host genes ($P < 0.01$ in all datasets; Figure 1A). A similar enrichment was also observed, although to a lesser extent, when ER+ tumors were compared with ER− (Figure 1B).

The above result argues that miR-HGs are preferentially and selectively regulated among different types of breast cancers. While this concept will be further discussed later, it is of note that there is a high degree of overlap between the G3/G1 and the ER+/ER− lists of miR-HGs regulated with same trend, in independent datasets (Table 1 and Tables S7C and D).

Table 2 — Significantly regulated miR-HGs in G3 versus G1 breast tumors and in the different tumor subtypes.

Probeset	miRBase	miRNA	Gene symbol	GEO-1 FC	GEO-2 FC	GEO-3 FC	GEO-4 FC	GEO-1 P	GEO-2 P	GEO-3 P	GEO-4 P	TCGA subtype P (ANOVA)
201664_at	MI0000115/MI0000438	*16-2/15b*	*SMC4*	1.94	1.57	1.91	1.74	<0.001	<0.001	<0.001	0.001	–
201663_s_at	MI0000115/MI0000438	*16-2/15b*	*SMC4*	1.71	1.60	1.56	1.88	<0.001	<0.001	0.002	<0.001	<0.001
210983_s_at	MI0000095	*106b/25/93*	*MCM7*	1.57	1.68	1.52	1.57	<0.001	<0.001	0.001	<0.001	<0.001
201852_x_at	MI0006380	*1245*	*COL3A1*	−1.40	−1.53	−1.70	−1.56	0.005	0.012	0.001	0.016	<0.001
209897_s_at	MI0000294	*218-1*	*SLIT2*	−1.49	−1.51	−1.57	−1.57	<0.001	0.001	<0.001	0.003	<0.001
218966_at	MI0006403	*1266*	*MYO5C*	−1.61	−1.37	−1.63	−1.54	<0.001	0.007	<0.001	0.004	<0.001
217838_s_at	MI0000805	*342*	*EVL*	−2.45	−1.88	−2.25	−1.95	<0.001	<0.001	<0.001	<0.001	<0.001
202409_at	MI0002467	*483*	*IGF2*	−2.50	−1.72	−2.65	−2.31	<0.001	0.012	<0.001	<0.001	<0.001
214053_at	MI0006375	*548f-2*	*ERBB4*	−2.79	−2.43	−2.66	−2.38	<0.001	<0.001	<0.001	<0.001	<0.001

The miR-HGs significantly regulated (fold-change > 1.5 or < −1.5, in at least 3 out of the 4 datasets) between G3 and G1 breast tumors in the four Affymetrix GEO datasets are shown. The significance by ANOVA analysis of gene expression regulation among different tumor subtypes is also reported in a fifth independent breast tumor dataset (TCGA dataset (2012)), in which molecular subtyping using the PAM50-model was available (Agilent array). miRNA genes are indicated together with their relative miRBase accession number. Fold change (FC) and P-values (P) are reported. GEO-1, GSE4922 dataset; GEO-2, GSE1456 dataset; GEO-3, GSE2034 dataset; GEO-4, GSE7390 dataset.

3.3. *From miR-HGs to int-miRNAs: in silico predictions of relevance to cancer*

We next attempted to translate the information on miR-HGs into the corresponding information concerning their hosted int-miRNAs. Since this step was preparatory to the actual validation and analysis of biological relevance in "real" tumors, we concentrated on the miR-HGs differentially expressed between G3 and G1 tumors, which displayed the most consistent and significant regulation among datasets (Figure 1). To eliminate candidates whose fluctuations might be due to technical or biological variability, we applied a stringent threshold and selected only those miR-HGs that displayed a fold-change of at least 1.5 (positive or negative) in at least 3 out of the 4 datasets. This yielded a list of 8 candidate miR-HGs, of which 2 were upregulated and 6 downregulated (Table 2).

The two upregulated miR-HGs were *SMC4* and *MCM7* that are involved in DNA synthesis, mitosis and DNA repair (Hagstrom and Meyer, 2003; Lei and Tye, 2001), and contain two miRNA clusters in their introns: the *miR-15b~16-2* and the *miR-25~106b* cluster, respectively. Both miRNA families have been described as being relevant to cancer (Bonci et al., 2008; Poliseno et al., 2010). Conversely, the six downregulated miR-HGs contain nine miRNAs whose relevance to cancer has not been investigated in detail: *miR-548f-2*, *miR-1245*, *miR-218/218**, *miR-342-3p/-5p*, *miR-483-3p/-5p* and *miR-1266* (Grady et al., 2008; Song et al., 2012; Soon et al., 2009; Tie et al., 2010; Veronese et al., 2010).

To gain initial insights into the potential relevance of these int-miRNAs to breast cancer, we performed pathway analyses of their predicted and validated target genes (see Methods). We observed a statistically significant enrichment in cancer-relevant genes among the predicted targets, which was confirmed also among the experimentally validated target genes (Figure 2A and B).

3.4. *Validation of predicted breast cancer-regulated int-miRNAs by qRT-PCR*

To validate the results of the *in silico* analysis, we analyzed the eleven identified int-miRNAs by qRT-PCR in a cohort of 36 FFPE archival G1 and G3 breast cancers (Table S8A). *miR548f-2* and *miR-1245* were undetectable in all samples and were thus excluded from further analyses. The hierarchical clustering analysis of the tumor samples, based on the expression of the remaining nine int-miRNAs, displayed a clear separation at the first tree branching between G3 and G1 tumors (70% and 75% of G3 and G1 tumors, respectively, clustering at the first branch; $P = 0.006$, likelihood-ratio test; Figure 3A). This result confirmed a distinct pattern of expression of the int-miRNA signature in high-*vs.* low-grade breast tumors, as predicted by the miR-HG expression analysis. Individually, *miR-342-3p/5p*, *miR-483-3p/5p* and *miR-*

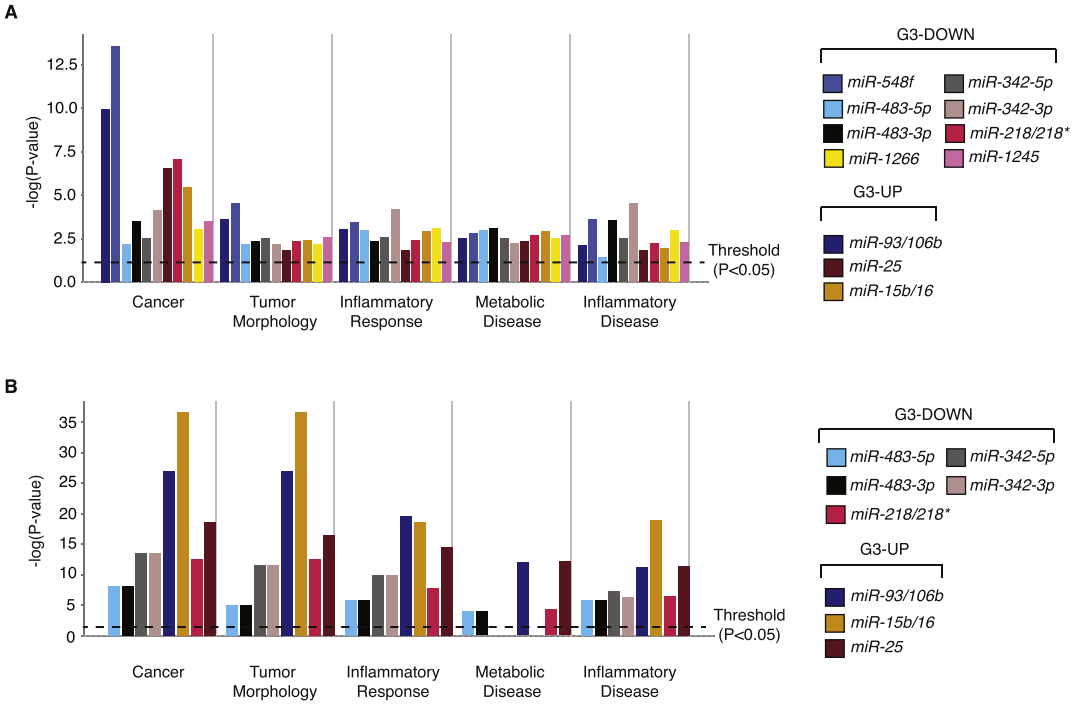
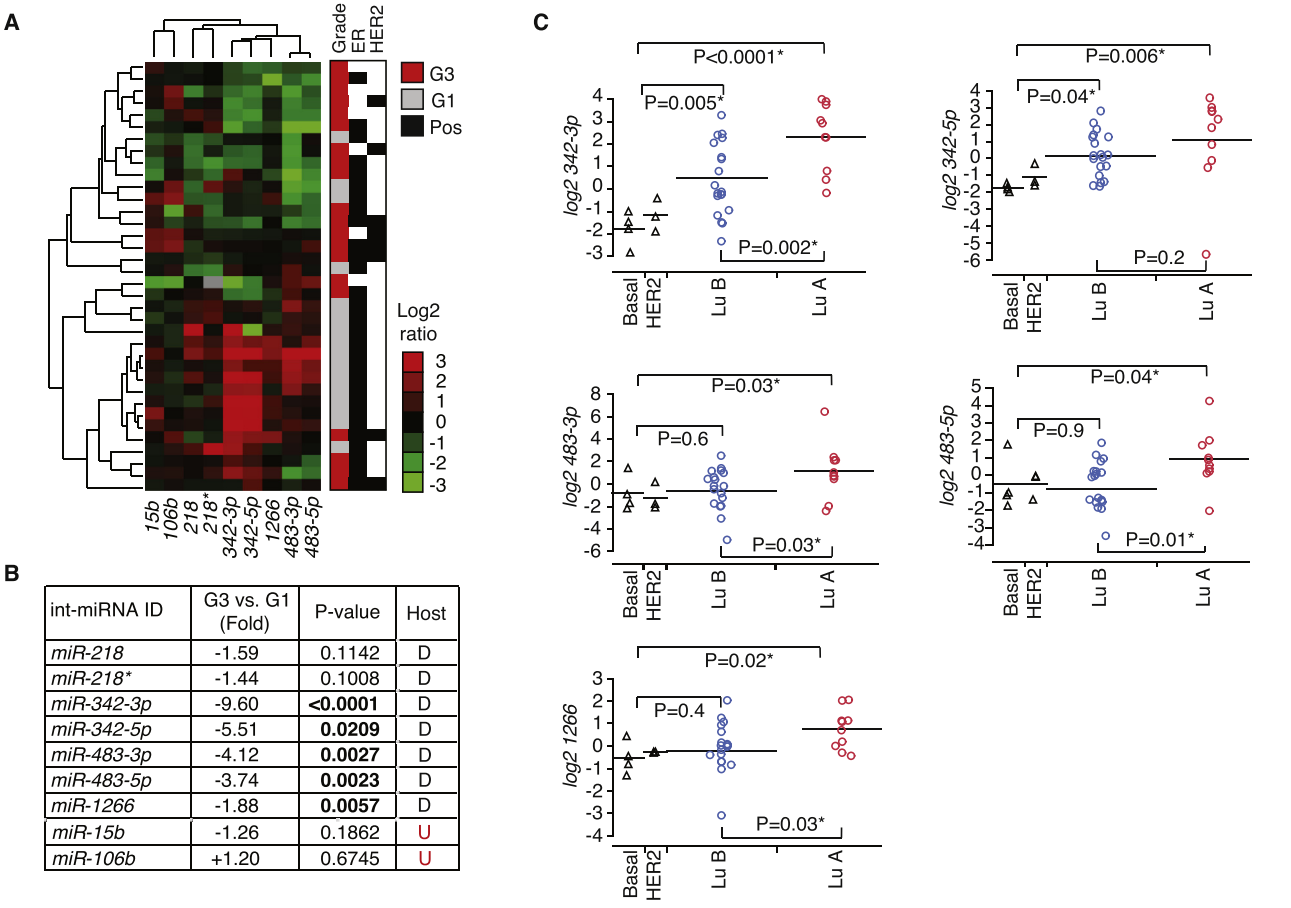


Figure 2 – Ingenuity Pathway Analysis of predicted and experimentally validated int-miRNA target genes. A. Bio-functions analysis of predicted target genes. B. Bio-functions analysis of experimentally validated target genes (see also Table S12). No validated targets were available for *miR-548f*, *miR-1245* and *miR-1266*. *Y*-axis, *P*-value (-log10; Fisher's Exact test) of the enrichment of int-miRNA targets in the biomolecular functions (displayed on the *X*-axis) annotated in the Ingenuity Bio-functions database. G3-DOWN, target genes of int-miRNAs located in host genes that are downregulated in G3 breast tumors. G3-UP, target genes of int-miRNAs located in host genes that are upregulated in G3 breast tumors. *P*-value (Fisher's exact test) cutoff was set at 0.05 (Threshold).



int-miRNA ID	G3 vs. G1 (Fold)	P-value	Host
miR-218	-1.59	0.1142	D
*miR-218**	-1.44	0.1008	D
miR-342-3p	-9.60	**<0.0001**	D
miR-342-5p	-5.51	**0.0209**	D
miR-483-3p	-4.12	**0.0027**	D
miR-483-5p	-3.74	**0.0023**	D
miR-1266	-1.88	**0.0057**	D
miR-15b	-1.26	0.1862	U
miR-106b	+1.20	0.6745	U

Figure 3 – Validation of the int-miRNA signature in the 36-patient cohort of G1 and G3 breast tumors by qRT-PCR. A. Hierarchical clustering of tumors based on the expression of selected int-miRNAs. *miR-15b* and *miR-106b* were used as representatives of their respective co-transcribed miRNA clusters, while *miR-548f-2* and *miR-1245* were undetectable by qRT-PCR. Columns represent log2 ratios of expression of each miRNA (median centered); rows represent tumor samples. Colored bars indicate the class of each patient. The color code, on the right, shows the characteristics of each patient: red, grade 3 tumor (G3); grey, grade 1 tumors (G1); ER, estrogen receptor (black = positive, white = negative); HER2, ErbB2 receptor (black = positive, white = negative). B. Differences in int-miRNA expression between G3 and G1 tumors. G3 *vs.* G1 (Fold): fold-change difference in expression; *P*-value calculated by Student's *t*-test; Host: Affymetrix host gene expression change in G3 *vs.* G1 tumors (D: downregulated, U: upregulated). C. Expression ratios (Log2) of int-miRNAs in breast tumor subtypes defined by ER/PgR, HER2 and Ki67 status. The tumor subtypes, shown on the *x*-axes, were identified as follows: basal = ER−, PgR- and HER2-; HER2 = ER−, PgR- and HER2+; LuB = ER+ and/or PgR+, Ki67 ≥ 14%; LuA = ER+ and/or PgR+, Ki67 < 14%. *P*-values were calculated using the Student's *t*-test. Asterisks, statistically significant *P*-values.

1266 were strongly downregulated in G3 breast tumors ($P < 0.05$; Figure 3B). *miR-218/218** were also downregulated, but not significantly ($P > 0.05$; Figure 3B). Similarly, the regulation of *miR-15b* and *miR-106b* (used as representatives of the *miR-15b∼16-2* and the *miR-25∼106b* naturally co-transcribed clusters) was not significant ($P > 0.05$; Figure 3B).

Since the expression of the miR-HGs was also differentially regulated in different breast cancer molecular subtypes (Table 2, Figure S3), we compared the expression levels of the significantly regulated int-miRNAs in the breast tumor subtypes in the 36-patient cohort. The different molecular subtypes – luminal A (LuA), Luminal B (LuB), basal, and HER2 – were identified using ER, progesterone receptor (PgR), HER2 and Ki67 immunohistochemistry markers (Blows et al., 2010; Cheang et al., 2009; Nielsen et al., 2004). Overall, the int-miRNAs showed significant downregulation in the most aggressive molecular subtypes (LuB, basal, and HER2) with respect to the less aggressive LuA subtype (Figure 3C). Similar results were obtained using an external dataset (Enerly et al., 2011) with matched miR-HG and int-miRNA expression profile (Figure S4). Importantly, we also observed a general downregulation of these int-miRNAs, and of their relative miR-HGs, in breast tumors when compared to normal breast epithelium (Figure S5). This latter finding suggests that the loss of expression of the analyzed int-miRNAs/miR-HGs might be directly involved in the transformation process.

3.5. *Detection of aggressive G2 breast cancers using the int-miRNA signature*

The sum of the previous results suggested an association between the downregulation of four int-miRNAs – *miR-342-3p*, *miR-483-3p*, *miR-483-5p* and *miR-1266* – and features of aggressiveness in breast cancer. We directly tested this possibility by taking advantage of a category of breast tumors with a moderate degree of differentiation (G2 tumors). The reason for doing so was dual. On the one hand, G2 tumors were not considered in our previous analyses, thus circumventing the risk of overfitting the data because of the selection of candidate int-miRNAs. On the other, G2 tumors represent a heterogeneous category, composed of tumors with varying degrees of aggressiveness (Gnant et al., 2011; Ivshina et al., 2006; Rakha et al., 2010; Sotiriou et al., 2006).

An independent cohort of 95 G2 tumors was profiled for *miR-483-3p/5p*, *miR-342-3p* and *miR-1266* expression (Table S8B). Hierarchical clustering analysis of int-miRNA expression profiles of this cohort, alone or together with those of the previously described 36-tumor G1/G3 cohort, revealed two main clusters, characterized by opposite regulation of the four int-miRNAs (Figure 4A). The cohort of G2 breast cancer patients was almost equally distributed between the two clusters. Interestingly, luminal A (less aggressive) and luminal B (more aggressive) G2 cancers co-segregated significantly with G1 and G3 tumors, respectively (Figure 4A), as also confirmed by contingency analysis ($P < 0.0001$; Figure 4B).

Three of the four int-miRNAs, *miR-483-3p*, *miR-483-5p* and *miR-1266*, were significantly downregulated in G2-LuB *vs.* G2-LuA tumors ($P < 0.0001$; Figure 4C), suggesting that they might identify more aggressive subtypes (i.e. Luminal B) even in G2 tumors. To investigate this possibility, we built a multivariate model based on these three miRNAs using the 95-patient cohort as the training set (see Materials and Methods). The model was then validated in an additional independent cohort

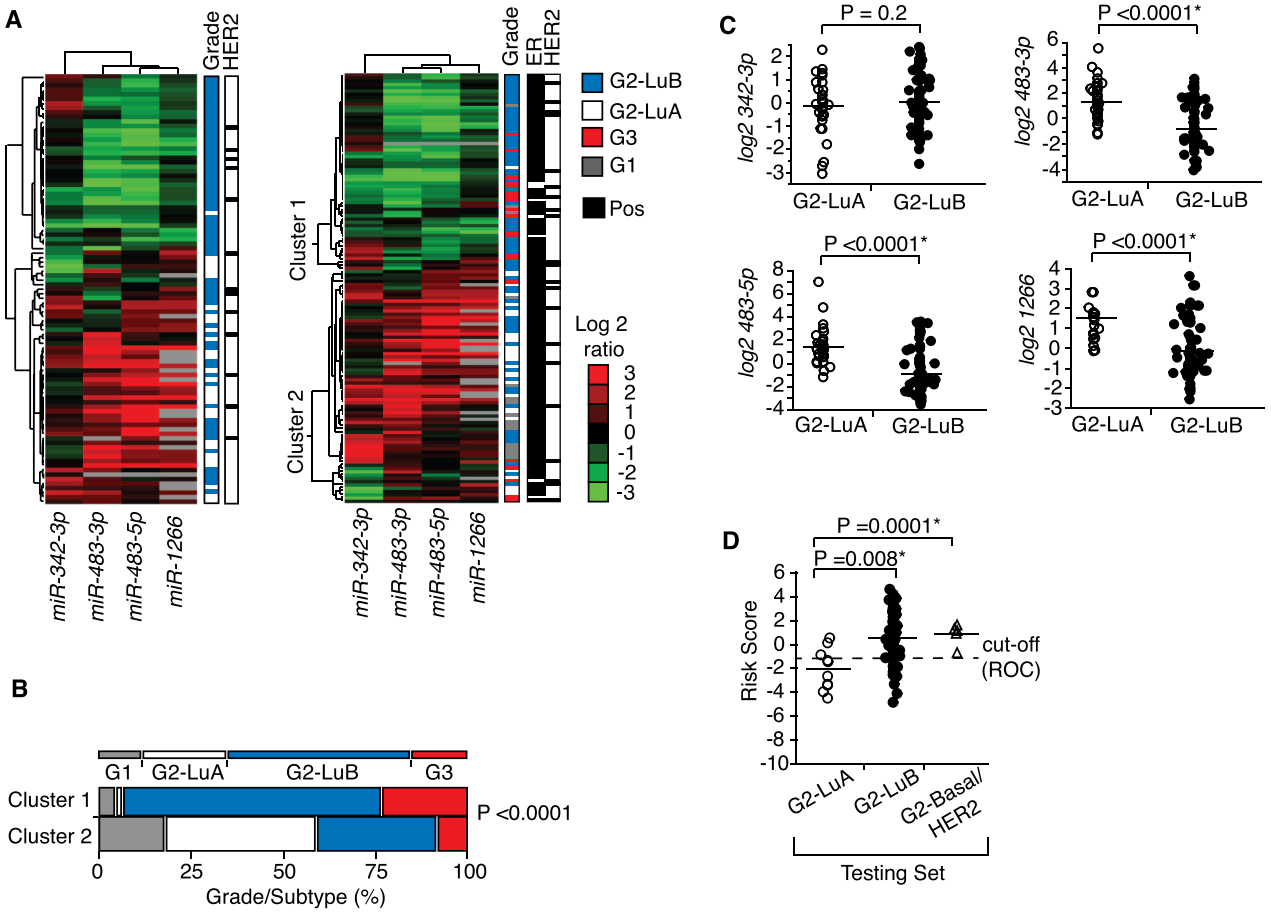


Figure 4 – Molecularly "aggressive" G2 breast cancers are identified by an int-miRNA signature. A. The cohort of 95 G2 tumors was analyzed by qRT-PCR for the expression of *miR-342-3p*, *miR-483* and *miR-1266*. Results, alone or together with those obtained in G3:G1 cohort (36 patients), were subjected to hierarchical clustering analysis. Columns represent log2 ratios of expression of int-miRNAs (median centered); rows represent tumor samples. The color code, on the right, indicates tumor grade/subtype: G3, red; G1, grey; G2-LuB, light blue; G2-LuA, white; ER, (black = positive); HER2, (black = positive). B. Contingency analysis of tumor distribution in Clusters 1 and 2. Color codes as in (A). *P*-value, likelihood-ratio test. C. Expression analysis of int-miRNAs in G2-luminal A (G2-LuA) and G2-luminal B (G2-LuB) tumors. *P*-values, Student's *t*-test. D. Performance of the int-miRNA signature composed of *miR-483-3p*, *miR-483-5p* and *miR-1266* in the additional independent cohort (Testing Set) of 90 G2 breast tumors. *Y*-axes: risk scores of the model. *X*-axes: breast tumor subtypes. Dashed line: decision cutoff used to classify patients in the high-*vs*. low-risk category, determined by nominal logistic regression and ROC analysis. *P*-values were calculated by the Student's *t*-test. Asterisks, statistically significant *P*-values.

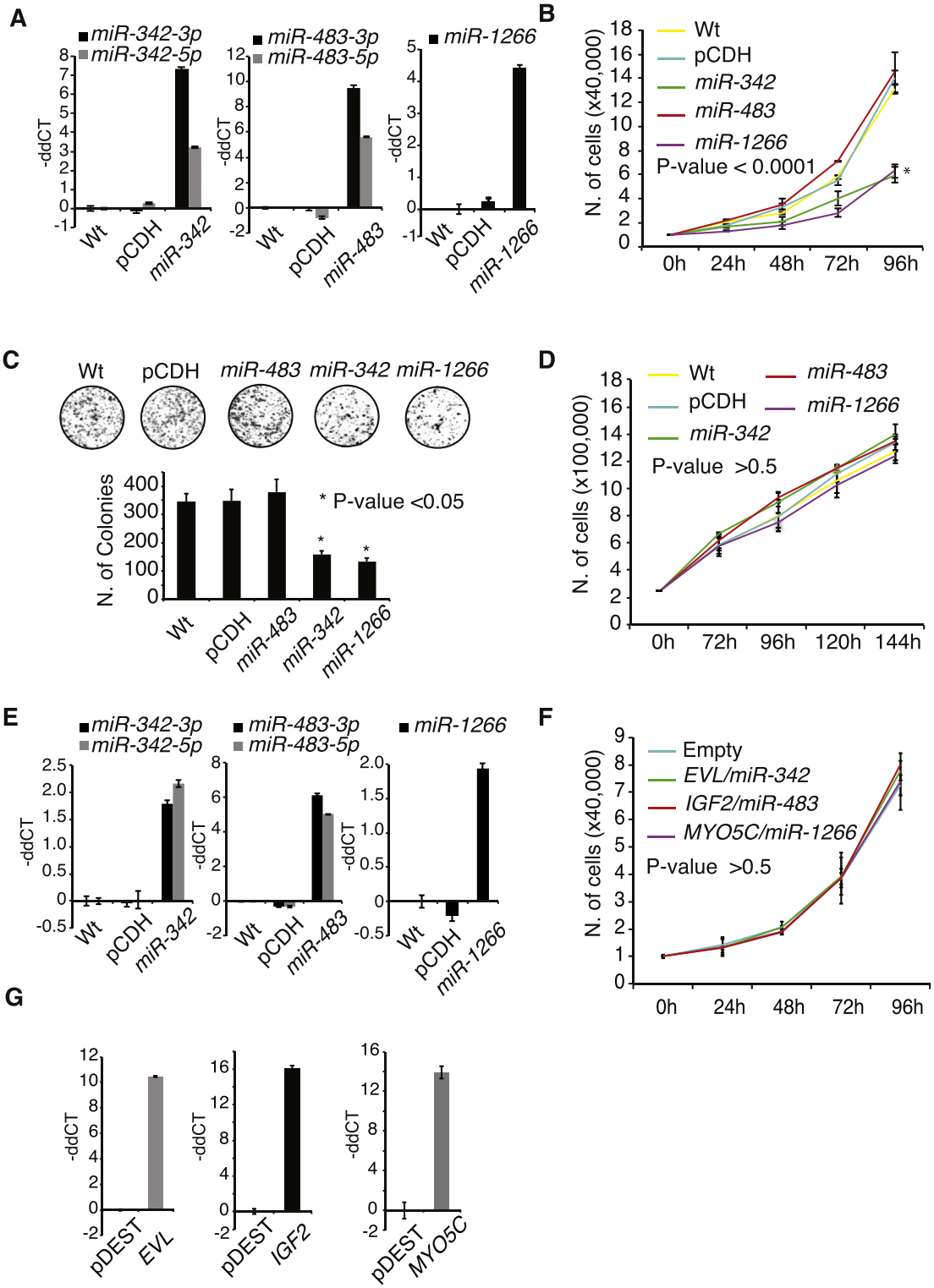


Figure 5 – Effects of overexpression of *miR-1266*, *miR-342-3p/5p* and *miR-483-3p/5p* and of relative host genes in breast cancer cell lines. Cells were infected with lentiviral vectors expressing precursors of *miR-342*, *miR-483* and *miR-1266*, or transfected with vectors expressing full-length *EVL*, *IGF2* and *MYO5C* miR-HGs. A. qRT-PCR analysis of int-miRNA expression in MDA-MB-231 cells. miRNA levels are reported as the Log2 normalized ratio of expression (-ddCT) relative to wild-type (non-infected) cells. pCDH, the empty vector pCDH-CMV-MCS-EF1-GFP was used as control. B. Cell proliferation assay with MDA-MB-231 cells. Data represent the mean ± σ from three independent experiments ($n = 3$). *P*-value, two-way ANOVA test relative to control cells (i.e., WT or pCDH). The asterisk indicates statistical significance. C. Colony forming assay with MDA-MB-231 cells. Images (top) represent colonies formed ten days after seeding. The bar graph (bottom) displays the mean ± σ from two independent experiments ($n = 2$). *, statistically significant *P*-value ($P < 0.05$) relative to control cells (WT or pCDH). D. Cell proliferation assay with MDA-MB-361 cells. Data represent the mean ± σ from three independent experiments. No significant differences

of 90 patients with G2 node-negative breast cancer, in which all subtypes were represented (Table S9). In this cohort, the model correctly identified all 5 basal/HER2 tumors and 60 of 75 LuB tumors (80%) as 'high risk', and 7 of 10 LuA tumors (70%) as 'low-risk' (Figure 4D). Thus, the risk model detected aggressive tumor subtypes with an accuracy, sensitivity and specificity of 80%, 81% and 70%, respectively.

In conclusion, we have provided a proof of principle that the detection of int-miRNAs, through data mining of miR-HGs in published cancer expression datasets, represents a viable strategy for the identification *in silico* of miRNAs of potential cancer relevance.

3.6. *Increased expression of miR-342 and miR-1266, but not of their host genes, impairs breast cancer cell proliferation and migration*

We investigated whether the regulated expression of the loci corresponding to *miR-342-3p/5p*, *miR-483-3p/5p* and *miR-1266* has an impact on the biology of breast cancer cells. Hierarchical clustering of the *miR-342*, *miR-483* and *miR-1266* miR-HG expression profiles (i.e. *EVL*, *IGF2* and *MYO5C*, respectively) in a panel of 51 breast cancer cell lines, for which expression data are publicly available [Table S10 (Neve et al., 2006)], revealed two main clusters, enriched in cell lines of the basal (Cluster 1, 20 out of 22, $P < 0.0001$) or luminal (Cluster 2, 23 out of 29, $P < 0.0001$) subtypes (Figures S6A and B). We selected as a model system, the MDA-MB-231 cell line that displayed the lowest median expression levels of the *IGF2*, *EVL2* and *MYO5C* miR-HGs (Figure S6C). As a control, we selected the MDA-MB-361 cell line, which displayed an opposite, and quantitatively comparable, regulation of the same miR-HGs (Figure S6C). By qRT-PCR analysis, we confirmed that the expression of *miR-342-3p/5p*, *miR-483-3p/5p* and *miR-1266* was indeed congruent with the expression of their host genes in both of the selected cell lines (Figure S6D).

We selected two relevant cancer phenotypes, proliferation and migration, to test the impact of restoration of high levels of expression of *miR-342*, *miR-483* and *miR-1266* in MDA-MB-231 cells, using MDA-MB-361 as a specificity control. We also tested the effects of overexpression of the corresponding miR-HGs (*IGF2*, *EVL2* and *MYO5C*). The lentiviral-mediated expression of *miR-342* and *miR-1266* (Figure 5A), caused a significant reduction in the proliferation rate and colony forming ability of MDA-MB-231, but not MDA-MB-361, cells (Figure 5B–E). Similarly, *miR-1266* expression significantly impaired cell migration of MDA-MB-231, but not MDA-MB-361, cells (Figure 6A–C). Conversely, overexpression of the miR-HGs did not affect proliferation or migration of MDA-MB-231 cells (Figure 5F–G, and Figure 6D). Finally, since negative self-regulation of miR-HGs has been reported in the literature (Bosia et al., 2012), we analyzed the expression of miR-HGs upon overexpression of their cognate int-miRNAs: no significant changes were observed (Figure S7).

4. Discussion

Here, we describe an approach to exploit an intrinsic characteristic of miRNAs, i.e. that ~50% of their genes reside within introns of protein-coding genes and share their regulation (He et al., 2012; Monteys et al., 2010; Ozsolak et al., 2008; Rodriguez et al., 2004). We reasoned that the wealth of publicly available microarray (mRNA) expression datasets might contain "encrypted" miRNA-related information that could be exploited for the discovery of biologically relevant miRNAs simply through meta-analysis.

In designing a proof-of-principle validation, we concentrated on breast cancer for which several high-quality, independent, transcriptome datasets are publicly available. We did so with multiple intents: i) to verify whether we could identify differentially expressed int-miRNAs simply by extracting the latent information in published gene expression (mRNA) datasets, ii) to verify whether int-miRNA signatures can be identified that would allow patient stratification, iii) to identify int-miRNAs whose involvement in biological processes, most notably cancer, was not previously known, iv) to investigate whether, at least in some cases, the dysregulations emerging from gene expression profiling might be more informative if viewed from the point of view of the hosted int-miRNA, rather than of the hosting gene.

Our efforts were successful on all accounts. Firstly, we were able to identify several miR-HGs, and their corresponding int-miRNAs, that are differentially expressed in various breast cancer subtypes. Importantly, several of the int-miRNAs that we identified have recently been found to be regulated in high-throughput miRNA expression profilings of independent cohorts of breast cancer patients [Table S11; (Aure et al., 2013; Blenkiron et al., 2007; Dvinge et al., 2013; Volinia et al., 2012)]. These data further support the effectiveness of our *in silico* approach to predict cancer-regulated int-miRNAs.

Secondly, we were able to identify an int-miRNA cancer signature – composed of *miR-342*, *miR-483* and *miR-1266* – that successfully stratified G2 cancers according to their molecular subtype, and therefore, according to their aggressiveness. We are not claiming a direct, even prospective, clinical utility of the identified signature or of the related risk model. Clearly, further studies are needed in this direction, aimed at, for example, comparing our int-miRNA cancer signature with existing stratification tools, such as molecular subtyping. However, our data demonstrate that the int-miRNA-related information "hidden" in transcriptomic profiles can be

were found by two-way ANOVA test relative to control cells ($P > 0.5$; WT or pCDH). E. qRT-PCR analysis of int-miRNA expression in MDA-MB-361 cells. miRNA levels are reported as the Log2 normalized ratio of expression (-ddCT) relative to wild-type (non-infected) cells. pCDH, the empty vector pCDH-CMV-MCS-EF1-GFP was used as control. F. Cell proliferation assay in MDA-MB-231 cells overexpressing the indicated miR-HGs. Data represent the mean ± σ from three independent experiments ($n = 3$). No significant differences were found by two-way ANOVA test ($P > 0.5$) relative to control cells (transfected with a pDEST26 empty vector). G. qRT-PCR analysis of miR-HG expression in MDA-MB-231 cells. *EVL*, *IGF2* or *MYO5C* host gene expression is reported as the Log2 normalized ratio of expression (-ddCT) relative to control (empty vector transfected) cells. Empty, the empty vector pDEST26 was used as a control.

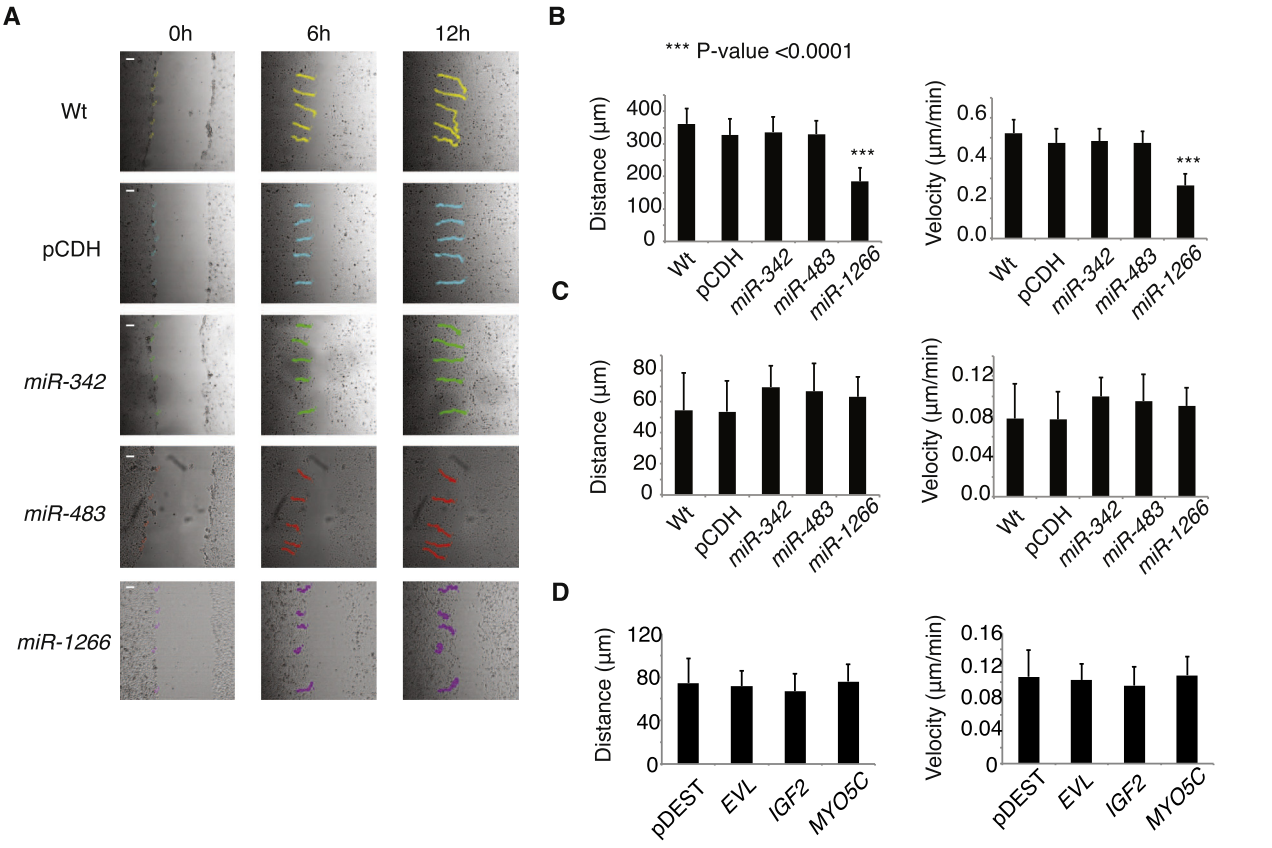


Figure 6 – Effects of overexpression of *miR-1266*, *miR-342-3p/5p* and *miR-483-3p/5p* on the migration of breast cancer cell lines. A. Monolayers of infected or wild-type (WT) MDA-MB-231 cells were scratch-wounded, as shown in the images on the left, and monitored by time-lapse video microscopy. Representative images were taken from movies at 0, 6 and 12 h. Colored lines show tracks of 5 representative cells. White Bar, 30 µm. B. Quantitation of the experiment shown in (A). Mean distance covered (left) and velocity (right) are shown. Data represent the mean ± σ from 15 individually tracked cells from 3 independent experiments. *P*-values were calculated using Welch's *t*-test analysis. ***, $P < 0.0001$ relative to control cells (Wt or pCDH). C. Quantitation of migration of MDA-MB-361 cells overexpressing *miR-1266*, *miR-342-3p/5p* and *miR-483-3p/5p*, relative to control cells (Wt or pCDH). Mean distance covered (left) and velocity (right) are shown. Data represent the mean ± σ from 15 individually tracked cells from 3 independent experiments. *P*-values were calculated using Welch's *t*-test analysis and were not significant ($P > 0.5$). D. Quantitation of migration of MDA-MB-231 cells overexpressing *EVL*, *IGF* and *MYO5C*, relative to control cells (pDEST). Mean distance covered (left) and velocity (right) are shown. Data represent the mean ± σ from 15 individually tracked cells from 3 independent experiments. *P*-values were calculated using Welch's *t*-test analysis and were not significant ($P > 0.5$).

successfully extracted *in silico* and used to identify miRNAs of potential clinical utility.

From a biological viewpoint, our approach led to the identification of two miRNAs, *miR-342* and *miR-1266*, that are involved in cancer-related phenotypes. The fact that overexpression of these miRNAs inhibited cancer-relevant phenotypes specifically in cells that display low expression of these miRNAs, but not in cells that express normal levels, argues that the downmodulation of *miR-342* and *miR-1266* might have a causal role in determining the aggressiveness of some breast cancers. This latter notion is supported by our findings that *miR-342* and *miR-1266* are differentially expressed between different breast cancer subtypes, and also underexpressed in some tumor tissues with respect to the normal breast epithelium. Of note, while there is some evidence in the literature indicating an involvement of *miR-342* in cancer (Dvinge et al., 2013; Veronese et al., 2010), there have been no reports, prior to this study, of a role of *miR-1266* in cancer (Ichihara et al., 2012).

Lastly, in the case of the loci encoding *miR-342/EVL* and *miR1266/MYO5C*, we report the intriguing observation that restoration of the expression of the int-miRNA, but not that of the host gene, inhibits cancer-relevant phenotypes. Thus, adding the "int-miRNA perspective" to the analysis and validation of gene expression studies is likely to increase the likelihood of identifying significant biological mechanisms involved in cancer.

5. Conclusions

In summary, we have developed an approach for the identification of biologically relevant int-miRNAs that has the potential to accelerate miRNA research by bypassing the lengthy and costly phase of initial screenings, and substituting them with meta-analysis of miR-HGs in publicly available expression datasets. The approach apparently performs well both in the holistic-oriented field of signature identification, which

finds primary application in projects of clinical interest, and in the more traditional field of "gene hunting" to guide high-resolution studies. While we have applied the methodology to the breast cancer setting to obtain a proof of principle of its utility, we envision applications in several fields of biology and medicine for which high quality gene expression datasets are available.

Acknowledgments

We are indebted with the Molecular Pathology Unit and the Clinical Biomarker Lab (Francesca De Santis, Stefania Pirroni) and the Imaging Facility at the Molecular Medicine Program of IEO; Pascale R. Romano and Rosalind Gunby for critically editing the manuscript. We also thank the IEO Biobank and Biomolecular Resource Infrastructure (IBBRI) and the Division of Pathology at IEO. This work was supported by grants to PPDF from Associazione Italiana per la Ricerca sul Cancro (AIRC, IG 10349 and 14404 and MCO 10.000 to PPDF), MIUR (the Italian Ministry of University and Scientific Research), the Italian Ministry of Health; from European Research Council (Mammastem Project), and from the Monzino Foundation.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.molonc.2014.10.001.

REFERENCES

Aure, M.R., Leivonen, S.K., Fleischer, T., Zhu, Q., Overgaard, J., Alsner, J., Tramm, T., Louhimo, R., Alnaes, G.I., Perala, M., Busato, F., Touleimat, N., Tost, J., Borresen-Dale, A.L., Hautaniemi, S., Troyanskaya, O.G., Lingjaerde, O.C., Sahlberg, K.K., Kristensen, V.N., 2013. Individual and combined effects of DNA methylation and copy number alterations on miRNA expression in breast tumors. Genome Biol. 14, R126.

Baskerville, S., Bartel, D.P., 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA 11, 241–247.

Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A., 2005. Reverse engineering of regulatory networks in human B cells. Nat. Genet. 37, 382–390.

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

Blows, F.M., Driver, K.E., Schmidt, M.K., Broeks, A., van Leeuwen, F.E., Wesseling, J., Cheang, M.C., Gelmon, K., Nielsen, T.O., Blomqvist, C., Heikkila, P., Heikkinen, T., Nevanlinna, H., Akslen, L.A., Begin, L.R., Foulkes, W.D., Couch, F.J., Wang, X., Cafourek, V., Olson, J.E., Baglietto, L., Giles, G.G., Severi, G., McLean, C.A., Southey, M.C., Rakha, E., Green, A.R., Ellis, I.O., Sherman, M.E., Lissowska, J., Anderson, W.F., Cox, A., Cross, S.S., Reed, M.W., Provenzano, E., Dawson, S.J., Dunning, A.M., Humphreys, M., Easton, D.F., Garcia-Closas, M., Caldas, C., Pharoah, P.D., Huntsman, D., 2010. Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. Plos Med. 7, e1000279.

Bolstad, B.M., Collin, F., Simpson, K.M., Irizarry, R.A., Speed, T.P., 2004. Experimental design and low-level analysis of microarray data. Int. Rev. Neurobiol. 60, 25–58.

Bonci, D., Coppola, V., Musumeci, M., Addario, A., Giuffrida, R., Memeo, L., D'Urso, L., Pagliuca, A., Biffoni, M., Labbaye, C., Bartucci, M., Muto, G., Peschle, C., De Maria, R., 2008. The miR-15a-miR-16-1 cluster controls prostate cancer by targeting multiple oncogenic activities. Nat. Med. 14, 1271–1277.

Bosia, C., Osella, M., Baroudi, M.E., Cora, D., Caselle, M., 2012. Gene autoregulation via intronic microRNAs and its functions. BMC Syst. Biol. 6, 131.

Calin, G.A., Croce, C.M., 2006. MicroRNA signatures in human cancers. Nat. Rev. Cancer 6, 857–866.

Carbone, M., Rizzo, P., Grimley, P.M., Procopio, A., Mew, D.J., Shridhar, V., de Bartolomeis, A., Esposito, V., Giuliano, M.T., Steinberg, S.M., Levine, A.S., Giordano, A., Pass, H.I., 1997. Simian virus-40 large-T antigen binds p53 in human mesotheliomas. Nat. Med. 3, 908–912.

Cheang, M.C., Chia, S.K., Voduc, D., Gao, D., Leung, S., Snider, J., Watson, M., Davies, S., Bernard, P.S., Parker, J.S., Perou, C.M., Ellis, M.J., Nielsen, T.O., 2009. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. J. Natl. Cancer Inst. 101, 736–750.

Chen, D.T., Nasir, A., Culhane, A., Venkataramu, C., Fulp, W., Rubio, R., Wang, T., Agrawal, D., McCarthy, S.M., Gruidl, M., Bloom, G., Anderson, T., White, J., Quackenbush, J., Yeatman, T., 2010. Proliferative genes dominate malignancy-risk gene signature in histologically-normal breast tissue. Breast Cancer Res. Treat. 119, 335–346.

De Luca, A., Baldi, A., Esposito, V., Howard, C.M., Bagella, L., Rizzo, P., Caputi, M., Pass, H.I., Giordano, G.G., Baldi, F., Carbone, M., Giordano, A., 1997. The retinoblastoma gene family pRb/p105, p107, pRb2/p130 and simian virus-40 large T-antigen in human mesotheliomas. Nat. Med. 3, 913–916.

Dvinge, H., Git, A., Graf, S., Salmon-Divon, M., Curtis, C., Sottoriva, A., Zhao, Y., Hirst, M., Armisen, J., Miska, E.A., Chin, S.F., Provenzano, E., Turashvili, G., Green, A., Ellis, I., Aparicio, S., Caldas, C., 2013. The shaping and functional consequences of the microRNA landscape in breast cancer. Nature 497, 378–382.

Dweep, H., Sticht, C., Pandey, P., Gretz, N., 2011. miRWalk–database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. J. Biomed. Inform. 44, 839–847.

Enerly, E., Steinfeld, I., Kleivi, K., Leivonen, S.K., Aure, M.R., Russnes, H.G., Ronneberg, J.A., Johnsen, H., Navon, R., Rodland, E., Makela, R., Naume, B., Perala, M., Kallioniemi, O., Kristensen, V.N., Yakhini, Z., Borresen-Dale, A.L., 2011. miRNA-mRNA integrated analysis reveals roles for miRNAs in primary breast tumors. PLoS One 6, e16915.

Fury, W., Batliwalla, F., Gregersen, P.K., Li, W., 2006. Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion. Conf. Proc. IEEE Eng. Med. Biol. Soc. 1, 5531–5534.

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., Zhang, J., 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80.

Girardi, A.J., Sweet, B.H., Slotnick, V.B., Hilleman, M.R., 1962. Development of tumors in hamsters inoculated in the neonatal period with vacuolating virus, SV-40. Proc. Soc. Exp. Biol. Med. 109, 649–660.
Gnant, M., Harbeck, N., Thomssen, C., 2011. St. Gallen 2011: summary of the consensus discussion. Breast Care (Basel) 6, 136–141.
Grady, W.M., Parkin, R.K., Mitchell, P.S., Lee, J.H., Kim, Y.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., Tewari, M., 2008. Epigenetic silencing of the intronic microRNA hsa-miR-342 and its host gene EVL in colorectal cancer. Oncogene 27, 3880–3888.
Griffiths-Jones, S., 2007. Annotating noncoding RNA genes. Annu. Rev. Genomics Hum. Genet., 279–298.
Hagstrom, K.A., Meyer, B.J., 2003. Condensin and cohesin: more than chromosome compactor and glue. Nat. Rev. Genet. 4, 520–534.
He, C., Li, Z., Chen, P., Huang, H., Hurst, L.D., Chen, J., 2012. Young intragenic miRNAs are less coexpressed with host genes than old ones: implications of miRNA-host gene coevolution. Nucleic Acids Res. 40, 4002–4012.
Ichihara, A., Jinnin, M., Oyama, R., Yamane, K., Fujisawa, A., Sakai, K., Masuguchi, S., Fukushima, S., Maruo, K., Ihn, H., 2012. Increased serum levels of miR-1266 in patients with psoriasis vulgaris. Eur. J. Dermatol. 22, 68–71.
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., Miller, L.D., 2006. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer Res. 66, 10292–10301.
Krol, J., Loedige, I., Filipowicz, W., 2010. The widespread regulation of microRNA biogenesis, function and decay. Nat. Rev. Genet. 11, 597–610.
Lei, M., Tye, B.K., 2001. Initiating DNA synthesis: from recruiting to activating the MCM complex. J. Cell Sci. 114, 1447–1454.
Lutter, D., Marr, C., Krumsiek, J., Lang, E.W., Theis, F.J., 2010. Intronic microRNAs support their host genes by mediating synergistic and antagonistic regulatory effects. BMC Genomics 11.
Meijering, E., Dzyubachyk, O., Smal, I., 2012. Methods for cell and particle tracking. Methods Enzymol. 504, 183–200.
Minn, A.J., Kang, Y., Serganova, I., Gupta, G.P., Giri, D.D., Doubrovin, M., Ponomarev, V., Gerald, W.L., Blasberg, R., Massague, J., 2005. Distinct organ-specific metastatic potential of individual breast cancer cells and primary tumors. J. Clin. Invest. 115, 44–55.
Monteys, A.M., Spengler, R.M., Wan, J., Tecedor, L., Lennox, K.A., Xing, Y., Davidson, B.L., 2010. Structure and activity of putative intronic miRNA promoters. RNA 16, 495–505.
Neve, R.M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F.L., Fevr, T., Clark, L., Bayani, N., Coppe, J.P., Tong, F., Speed, T., Spellman, P.T., DeVries, S., Lapuk, A., Wang, N.J., Kuo, W.L., Stilwell, J.L., Pinkel, D., Albertson, D.G., Waldman, F.M., McCormick, F., Dickson, R.B., Johnson, M.D., Lippman, M., Ethier, S., Gazdar, A., Gray, J.W., 2006. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell 10, 515–527.
Nielsen, T.O., Hsu, F.D., Jensen, K., Cheang, M., Karaca, G., Hu, Z., Hernandez-Boussard, T., Livasy, C., Cowan, D., Dressler, L., Akslen, L.A., Ragaz, J., Gown, A.M., Gilks, C.B., van de Rijn, M., Perou, C.M., 2004. Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. Clin. Cancer Res. 10, 5367–5374.
Ozsolak, F., Poling, L.L., Wang, Z., Liu, H., Liu, X.S., Roeder, R.G., Zhang, X., Song, J.S., Fisher, D.E., 2008. Chromatin structure analyses identify miRNA promoters. Genes Dev. 22, 3172–3183.
Palamidessi, A., Frittoli, E., Garre, M., Faretta, M., Mione, M., Testa, I., Diaspro, A., Lanzetti, L., Scita, G., Di Fiore, P.P., 2008. Endocytic trafficking of Rac is required for the spatial restriction of signaling in cell migration. Cell 134, 135–147.
Poliseno, L., Salmena, L., Riccardi, L., Fornari, A., Song, M.S., Hobbs, R.M., Sportoletti, P., Varmeh, S., Egia, A., Fedele, G., Rameh, L., Loda, M., Pandolfi, 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. Sci. Signal. 3, ra29.
Radfar, M.H., Wong, W., Morris, Q., 2011. Computational prediction of intronic microRNA targets using host gene expression reveals novel regulatory mechanisms. PLoS One 6.
Rainer, J., Ploner, C., Jesacher, S., Ploner, A., Eduardoff, M., Mansha, M., Wasim, M., Panzer-Grumayer, R., Trajanoski, Z., Niederegger, H., Kofler, R., 2009. Glucocorticoid-regulated microRNAs and mirtrons in acute lymphoblastic leukemia. Leukemia 23, 746–752.
Rakha, E.A., Reis-Filho, J.S., Baehner, F., Dabbs, D.J., Decker, T., Eusebi, V., Fox, S.B., Ichihara, S., Jacquemier, J., Lakhani, S.R., Palacios, J., Richardson, A.L., Schnitt, S.J., Schmitt, F.C., Tan, P.H., Tse, G.M., Badve, S., Ellis, I.O., 2010. Breast cancer prognostic classification in the molecular era: the role of histological grade. Breast Cancer Res. 12, 207.
Rodriguez, A., Griffiths-Jones, S., Ashurst, J.L., Bradley, A., 2004. Identification of mammalian microRNA host genes and transcription units. Genome Res. 14, 1902–1910.
Saini, H.K., Griffiths-Jones, S., Enright, A.J., 2007. Genomic analysis of human microRNA transcripts. Proc. Natl. Acad. Sci. USA, 17719–17724.
Song, L., Dai, T., Xie, Y., Wang, C., Lin, C., Wu, Z., Ying, Z., Wu, J., Li, M., Li, J., 2012. Up-regulation of miR-1245 by c-myc targets BRCA2 and impairs DNA repair. J. Mol. Cell Biol. 4, 108–117.
Soon, P.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., Sidhu, S.B., 2009. miR-195 and miR-483-5p identified as predictors of poor prognosis in adrenocortical cancer. Clin. Cancer Res. 15, 7684–7692.
Sorlie, 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., Lonning, P.E., Borresen-Dale, A.L., 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. USA 98, 10869–10874.
Sotiriou, C., Neo, S.Y., McShane, L.M., Korn, E.L., Long, P.M., Jazaeri, A., Martiat, P., Fox, S.B., Harris, A.L., Liu, E.T., 2003. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc. Natl. Acad. Sci. USA 100, 10393–10398.
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., Delorenzi, M., 2006. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J. Natl. Cancer Inst. 98, 262–272.
Sweet-Cordero, A., Mukherjee, S., Subramanian, A., You, H., Roix, J.J., Ladd-Acosta, C., Mesirov, J., Golub, T.R., Jacks, T., 2005. An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. Nat. Genet. 37, 48–55.
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., Fan, D., 2010. MiR-218 inhibits invasion and metastasis of gastric cancer by targeting the Robo1 receptor. Plos Genet. 6, e1000879.
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., Friend, S.H., 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530–536.

Venet, D., Dumont, J.E., Detours, V., 2011. Most random gene expression signatures are significantly associated with breast cancer outcome. Plos Comput. Biol. 7, e1002240.

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., Negrini, M., 2010. Oncogenic role of miR-483-3p at the IGF2/483 locus. Cancer Res. 70, 3140–3149.

Voduc, K.D., Cheang, M.C., Tyldesley, S., Gelmon, K., Nielsen, T.O., Kennecke, H., 2010. Breast cancer subtypes and the risk of local and regional relapse. J. Clin. Oncol. 28, 1684–1691.

Volinia, S., Galasso, M., Sana, M.E., Wise, T.F., Palatini, J., Huebner, K., Croce, C.M., 2012. Breast cancer signatures for invasiveness and prognosis defined by deep sequencing of microRNA. Proc. Natl. Acad. Sci. USA 109, 3024–3029.

Wang, D., Lu, M., Miao, J., Li, T.T., Wang, E., Cui, Q.H., 2009. Cepred: predicting the co-expression patterns of the human intronic microRNAs with their host genes. PLoS One 4.

Wang, Y.P., Li, K.B., 2009. Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. BMC Genomics 10, 218.

White, D.P., Caswell, P.T., Norman, J.C., 2007. Alpha v beta3 and alpha5beta1 integrin recycling pathways dictate downstream Rho kinase signaling to regulate persistent cell migration. J Cell Biol. 177, 515–525.

Yendamuri, S., Kratzke, R., 2011. MicroRNA biomarkers in lung cancer: MiRacle or quagMiRe? Transl. Res. 157, 209–215.