miRNA sequence and expression analysis in breast tumors

MicroRNA sequence and expression analysis in breast tumors by deep sequencing

Thalia A. Farazi1*, Hugo M. Horlings2,3*, Jelle ten Hoeve4, Aleksandra Mihailovic1, Hans Halfwerk2,3, Pavel Morozov1, Miguel Brown1, Markus Hafner1, Fabien Reyal3, Marieke van Kouwenhove3, Bas Kreike3,5, Daoud Sie3,6, Volker Hovestadt1, Lodewyk Wessels4##, Marc J. van de Vijver2,3##, and Thomas Tuschi1##

1Howard Hughes Medical Institute, Laboratory of RNA Molecular Biology, The Rockefeller University, New York, NY 10065, USA, 2Academic Medical Center, Department of Pathology, Meibergdreef 9, 1105AZ, Amsterdam, Netherlands, 3Division of Experimental Therapy, Department of Molecular Biology, 4Department of Bioinformatics and Statistics, 5Division of Radiation Oncology, and 6Central Microarray Facility, Netherlands Cancer Institute, Plesmanlaan 121, 1066CX, Amsterdam, Netherlands

#corresponding authors (ttuschi@mail.rockefeller.edu, tel:212-327-7651, fax:212-327-7652; m.j.vandevijver@amc.uva.nl, tel:+31-20-5666646, fax:+31-20-5669523; lwessels@nki.nl, tel:+31-20-5127987, fax:+31-61-4413040)

*both authors contributed equally to the work
miRNA sequence and expression analysis in breast tumors

Running Title:

miRNA sequence and expression analysis in breast tumors

Keywords:

Breast carcinomas, triple negative, HER2-positive, miRNA, deep-sequencing

Financial support:

Supported by grants from the Dutch Cancer Society (NKB2002-2575) and the NIH (1RC1CA145442).

Conflict of interest:

T. T. is cofounder and scientific advisor to Alnylam Pharmaceuticals and scientific advisor to Regulus Therapeutics.
miRNA sequence and expression analysis in breast tumors

Abstract

MicroRNAs (miRNAs) regulate many genes critical for tumorigenesis. We profiled miRNAs from 11 normal breast tissues, 17 non-invasive, 151 invasive breast carcinomas, and 6 cell lines by in-house-developed barcoded Solexa-sequencing. miRNAs were organized in genomic clusters representing promoter-controlled miRNA expression and sequence families representing seed-sequence-dependent miRNA-target regulation. Unsupervised clustering of samples by miRNA sequence families best reflected the clustering based on mRNA expression available for this sample set. Clustering and comparative analysis of miRNA read frequencies showed that normal breast samples were separated from most non-invasive ductal carcinoma in situ and invasive carcinomas by increased miR-21 (the most abundant miRNA in carcinomas) and multiple decreased miRNA families (including miR-98/let-7), with most miRNA changes apparent already in the non-invasive carcinomas. In addition, patients that went on to develop metastasis demonstrated increased expression of mir-423, and triple negative breast carcinomas were most distinct from other tumor subtypes due to up-regulation of the mir-17~92 cluster. However, absolute miRNA levels between normal breast and carcinomas did not reveal any significant differences. We also discovered two polymorphic nucleotide variations among the more abundant miRNAs miR-181a (T19G) and miR-185 (T16G), but we did not identify nucleotide variations expected for classical tumor suppressor function associated with miRNAs. The differentiation of tumor subtypes and prediction of metastasis based on miRNA levels is statistically possible, but is not driven by deregulation of abundant miRNAs, implicating far fewer miRNAs in tumorigenic processes than previously suggested.
miRNA sequence and expression analysis in breast tumors

Introduction

Breast cancer is a heterogeneous disease involving various oncogenic pathways and/or genetic alterations. Although prognostic mRNA expression signatures have been defined for some invasive breast carcinomas (1, 2), the underlying pathways regulating breast cancer aggressiveness remain poorly understood.

miRNAs are 19-23 nucleotide (nt) RNAs that regulate gene expression (3, 4, 5). miRNAs have been suggested to act as oncogenes (or tumor suppressors) contributing to distinct tumor characteristics (6, 7, 8, 9). Some miRNAs are located in genomic regions exhibiting copy number alterations (10), and miRNA levels can be dramatically deregulated in tumors, e.g. up-regulation of mir-21 (reviewed in 11), mir-17–92 cluster (reviewed in 12), mir-26a (13), and down-regulation of let-7 family members (reviewed in 14). miRNAs control key processes in tumorigenesis, such as tumor initiation (13), metastasis (15, 16), inflammation (17), and differentiation (18). Components of the miRNA biogenesis pathway have been implicated in tumorigenesis, suggesting that miRNAs are less abundant in tumors (19, 20). For example, DICER1 was found to function as a haplo-insufficient tumor suppressor (20, 21); 27% of various tumors were suggested to have a hemizygous deletion of the gene that encodes DICER1 (20). Furthermore, miRNA profiling in a variety of tissues and cancers has not only identified cell-type-specific and cancer-deregulated miRNAs (19, 22), but also profiles that correlate with prognosis (23).

Since 2005, when miRNA deregulation was described in breast tumors (24), over 400 studies have been published regarding changes in miRNA levels, their regulation and role in breast cancer (reviewed in 25). Most of these studies were conducted for a
miRNA sequence and expression analysis in breast tumors

specified subset of miRNAs using microarrays or RT-PCR, and did not identify miRNA abundance or sequence variation. Studies investigating the role of specific miRNAs in metastasis were conducted in cell lines and animal models, and were supported by small patient cohorts. For example, over-expression/knockdown of several miRNAs, including miR-10b, miR-9, miR-31, miR-126 and miR-335, was shown to play a role in metastasis (15, 16, 26, 27). However, correlation between these miRNAs and metastasis was not identified in large clinical studies (28, 29).

Here, we surveyed by barcoded Solexa-sequencing 185 breast specimens, including 11 normal breast, 17 ductal carcinoma in situ (DCIS), 151 invasive carcinomas (including 126 invasive ductal carcinomas (IDC)) and 6 ductal cell lines and obtained expression levels based on sequence read number. This approach not only provided the fold-differences, but also characterized miRNA abundance and nucleotide variation, accommodating a large number of clinical samples to address inter-patient variation in a cost-effective manner. We performed unsupervised clustering to assess if miRNAs organized in genomic clusters or sequence families could resolve breast cancer types, as classified by immunohistochemistry (IHC) i.e. estrogen receptor (ER), progesterone receptor (PR) and HER2 status, and molecular subtypes based on mRNA profiles (30).

By comparing HER2-positive non-invasive (DCIS) and invasive breast carcinomas we identified miRNAs with altered levels in tumor invasion. By comparing breast tumor types with different IHC characteristics we identified miRNAs unique to triple negative breast carcinomas (TNBC), lacking expression for ER, PR and HER2. 97 out of 151 invasive breast carcinoma patients had clinical follow-up records available, allowing us to evaluate the role of miRNAs as prognostic markers.
Materials and Methods

Patient clinical and tumor histo-pathological characteristics

Samples were obtained from patients treated at the Netherlands Cancer Institute between 1985 and 2006, comprising DCIS (n=17) and invasive carcinoma (n=151) patients with non-metastatic disease at diagnosis. Eleven normal breast specimens were obtained from mammoplasty patients treated at the Institut Curie. The medical ethical committee of the Netherlands Cancer Institute approved this study. For detailed clinical and pathological information see Suppl. Table S1. Clinical follow-up data were available for 48 TNBC and 49 HER2-positive patients. Therapy for these patients consisted of breast conserving surgery or modified radical mastectomy, with or without adjuvant therapy (chemotherapy, radiation or hormonal therapy). 22 out of 49 HER2-positive and 28 out of 48 TNBC patients received anthracycline-based, cyclophosphamide, methotrexate and fluorouracil chemotherapy, hormonal treatment or a combination modality. The MCF7, HCC38, MCF10A and BT-474 cell lines were purchased from ATCC. ZR-751 was a gift from John Hilkens and MDA-MB134 from Marleen Kok.

RNA isolation

Fresh-frozen (frozen immediately after surgery and stored at -80°C) normal and tumor tissue was collected. RNA was isolated from approximately thirty 30-µm cryosections corresponding to approximately 20 mg of tissue, using the first and last section to assess tumor content; only samples containing ≥50% tumor were further characterized. The tissues were homogenized in Trizol (Invitrogen, Carlsbad, CA) using a Polytron
miRNA sequence and expression analysis in breast tumors

instrument (polytron®, PT, MR2100, Kinematica AG, Luzern) for 1 min and total RNA was isolated by a modified Trizol protocol (Supplementary Methods). Quality of isolated RNA was assessed on a 1% agarose gel based on the relative abundance of 18S and 28S subunits of ribosomal RNA.

**Small RNA sequencing and bioinformatics analysis**

We used a barcoded small RNA sequencing approach (described in (31) and summarized in Supplementary Methods). We employed 21 different barcodes obtaining 1.8-26.5 mio reads per sequence run (18-87% reads with extractable barcodes) (Suppl. Table S2). We selected reads with an insert of 16-25 nt. Adapter sequences were extracted from sequence reads, using the following criteria: 4-nt minimum overlap of 3’ adapter, or 5-nt minimum 3’ overlap of adapter with one mismatch excluding insertions and deletions in the first nucleotide of the adapter past the barcode; barcodes were assigned without allowing any mismatches. On average 80% of the extracted reads represented prototypical miRNAs (based on our annotation database – see below), 0.005% viral miRNAs and 0.2% piRNAs (based on NCBI definitions). The samples with lower miRNA content had higher percentage of rRNA, likely reflecting sample quality. The miRNA expression profiles were submitted to the Gene Expression Omnibus (GEO) with accession number GSE28884.

Setting a threshold of ≥10 reads per miRNA for the pool of all 49,479,978 miRNA sequence reads (179 patient samples and 6 cell lines), we identified a total of 888 mature and miRNA star sequences (representing the two strands obtained after miRNA
miRNA sequence and expression analysis in breast tumors

precursor processing). Resetting the threshold to 5,000 reads, we identified 231 miRNA and miRNA star species together constituting 99% of all miRNA reads.

To assess if experimental variables affected miRNA profiles, we profiled 54 samples in replicate and performed Spearman correlation yielding a median correlation coefficient of 0.92 using the top expressed miRNAs (represented with >5,000 sequence reads across all samples) (Suppl. Table S3).

miRNA genomic clusters were redefined taking into consideration EST evidence and levels of miRNA expression from our data (similarly to 22). Typically, the greatest genomic distance between clustered miRNAs was 5 kb. Sequence families were defined based on seed sequence similarity (position 2-8) allowing only one transition in these positions, as well as 3’ end similarity (position 9 through 3’ end) allowing up to 50% mismatches with additional manual curation (Suppl. Table S4).

Northern Blot analysis

5 μg of total RNA was used for Northern blot analysis as previously described (22). Equal loading of the lanes was confirmed by ethidium bromide staining of the tRNA band. Synthetic standards containing 2, 10, and 50 fmol oligoribonucleotide corresponding to each miRNA were run on each gel and used as references to quantify miRNA levels. After each experiment, blots were stripped in 0.1% SSC at 85 °C for 5 min, and re-probed up to 4 times, the final time for U6 snRNA as control for loading and degradation. The spot intensities were quantified using ImageGauge software in units of photo-stimulated luminescence and corrected for background intensity.
miRNA sequence and expression analysis in breast tumors

**Statistical analysis**

Kaplan-Meier analysis was conducted for the top 5 significantly deregulated miRNAs ($\alpha \leq 0.001$ as multiple testing correction for 20 comparisons) in patients who developed metastasis. For each miRNA tested, the patients were split into two groups at the median. Difference between the two Kaplan-Meier curves was assessed with the Log-Rank test.

**Results**

*Small RNA cDNA library preparation and miRNA abundance analysis*

We isolated total RNA from 11 normal breast samples, 17 DCIS and 151 invasive breast carcinomas (Suppl. Table S1). IHC typing of the breast carcinomas showed that 71 lacked ER, PR and HER2 expression (TNBC), while 97 were positive for a single or multiple of the three markers. 23 TNBCs represented special histological types: metaplastic (9), atypical medullary (8), adenoid cystic (2), and apocrine carcinomas (4). In addition, we isolated RNA from 6 ductal breast cell lines (MCF7, MCF10A, HCC38, BT474, MDA-MB134, and ZR-751).

We processed all samples by 3’-adapter barcoded small RNA cDNA library Solexa-sequencing (31). A total of 61,319,767 barcoded sequence reads were extracted yielding a median of 134,022 reads per sample (range: 7,403-1,608,855) (Suppl. Table S2); these reads were mapped to the genome allowing one mismatch/insertion/deletion and then to our non-coding RNA and miRNA databases allowing up to two mismatches or one insertion/deletion (22). We constructed a miRNA read database from >1,000 human samples sequenced in the Tuschl laboratory, defined prototypical miRNAs (557 precursors, corresponding to 1112 mature and star sequences, miR-451 and miR-618.
miRNA sequence and expression analysis in breast tumors

being the only miRNAs without a star sequence). We added 269 not yet reported star sequences, ignored putative miRNAs from miRBase for which we did not obtain read evidence and renamed specific miRNAs according to the read ratio between mature/star sequences (Suppl. Table S4).

Previous reports suggested that miRNAs were less abundant in tumor compared to normal samples (19, 20, 32). For a representative set of 31 samples (5 normals, 4 DCIS, 18 IDC and 4 cell lines), we determined the absolute amount of miRNA per µg of total RNA. We added a cocktail of an equimolar amount of 10 synthetic 22-nt 5’ phosphorylated RNAs distinct from human sequences per µg of total RNA, on average representing 18% of the total reads (range 7-57%) (31). We were not able to detect significant changes in miRNA content between normal, disease tissue or cell lines, assuming that calibrator RNAs and miRNAs were cloned with similar average efficiency. Normal breast contained an average of 16±4, DCIS 14±4, IDC 15±5 and cell lines 9±5 fmol miRNA/µg of total RNA. We did not account for tumor cell type heterogeneity, but tumor samples were selected to comprise ≥50% tumor cells. Consistent with these results, examination of mRNA levels in the same samples for miRNA pathway components or other factors implicated in miRNA biogenesis (reviewed in 9) did not suggest globally differential miRNA processing in normal breast and carcinomas (Suppl. Figure S1). Based on calculations in MCF7 cells, each tumor cell contained 145,000 miRNA molecules, illustrating that miRNAs expressed at 1% of the total miRNA content (see below) in each cell would represent 1,500 copies per cell. Considering that miRNAs regulate many mRNAs, each represented by many transcripts, lowly abundant miRNAs would be insufficient to confer measurable target mRNA regulation, unless they act like...
miRNA sequence and expression analysis in breast tumors

siRNAs on nearly fully complementary target mRNAs. Quantitative Western blot for EIF2C2/AGO2 protein, the main component of the miRNA effector complex, in MCF7 demonstrated the presence of approximately 42,000 copies per cell (Suppl. Figure S1). Assuming similar abundance for the often co-expressed EIF2C/AGO members, the number of effector complexes matches the miRNA copy number.

Prior to the development of the barcoded sequencing with addition of calibrator RNAs, we had performed quantitative Northern blotting for a subset of 10 miRNAs in 84 tumor samples from our collection (Suppl. Table S5). The Spearman correlation coefficients for miRNA expression based on sequence reads compared to Northern quantitation varied between 0.20 (miR-96) to 0.72 (miR-375) when comparing across all 84 samples. The absolute amount of each miRNA per µg of total RNA derived from Northern blotting was in general agreement with calibrator-assisted sequencing calculations.

**miRNA profiles in normal and tumor specimens**

miRNA profiles of a sample can be presented as relative percent (%) miRNA read frequencies (rf) by dividing miRNA read counts by total miRNA reads per library. Furthermore, miRNA profiles can be condensed either by assigning individual miRNA and miRNA star reads to their originating miRNA genomic clusters or to sequence families (denoted cluster-mir and sf-miR respectively, listing number of cluster/family members in parenthesis (Suppl. Table S4)). Either approach reduces the complexity of the data. The genomic cluster profiles represent promoter-controlled miRNA expression, while the sequence families are most informative for characterization of seed-sequence-
miRNA sequence and expression analysis in breast tumors

dependent miRNA-target regulation. For sample comparison, we required ≥5,000 total miRNA sequence reads per replicate merged library, resulting in inclusion of 179 samples (from 183 sequenced) and 6 cell lines.

We conducted unsupervised hierarchical clustering for 179 clinical samples using individual miRNAs, precursor clusters and sequence families expressed at 85% of the total miRNA reads in at least one sample (Figure 1 and Suppl. Figure S2). To enhance visualization of changes in miRNA expression, miRNA rf was transformed by standardization across each miRNA. Normal breast samples clustered together, close to a small group of ER- and HER2-positive tumors characterized by lower expression of sf-miR-21(1) and higher expression of sf-miR-22(1) compared to the remainder of tumor samples. Some DCIS samples also clustered together, breaking up groups of invasive tumor samples, suggesting that DCIS samples accumulate changes in miRNA expression early in the tumorigenic process. Invasive tumor samples positive for one or more IHC marker clustered together while TNBCs emerged as several groups distinct from the other tumors. Samples did not cluster according to other pathological and clinical characteristics.

To understand miRNA expression in the context of tissue heterogeneity, we compared the six ductal cell lines to human samples to identify miRNAs expressed in tissues, but not ductal cell lines, likely contributed by other cell types. We visualized miRNA abundance in normal breast, DCIS and IDC IHC types. To simplify the figure, we selected four representative samples from each IHC category that demonstrated miRNA expression closest to the average expression for each category, representing miRNAs expressed at 85% of the total miRNA reads in at least one sample or cell line.
miRNA sequence and expression analysis in breast tumors

(Figure 2). We identified miRNAs present at similar levels in cell lines and tumors that may be involved in tumorigenic processes, and miRNAs absent from cell lines that are likely expressed by other cell types, such as cluster-mir-126(1) (cardiovascular system), cluster-mir-143(2) (adipose tissue), cluster-mir-144(2) and cluster-mir-142 (hematopoietic system) (9, 22).

Clustering of 179 samples based on miRNAs (using 98% of miRNAs expressed in at least one sample) and clustering of 161 of these samples according to their mRNA profiles (using the same number of genes as miRNAs, selecting genes with most variance) is depicted in Figure 3. mRNA profiles better separated HER2-positive samples (p=1.90x10^{-15}), suggesting that the HER2 pathway is not related to miRNA expression changes. miRNA clustering was weakly correlated with TNBC. Three TNBC groups emerged by clustering miRNA profiles, one of which included mostly special histological types; these groups did not demonstrate distinct patient characteristics or outcome.

We used the EdgeR package (33) to identify individual mature miRNAs, miRNA genomic clusters and miRNA sequence families differentially represented between normal, DCIS and invasive tumor samples. We included miRNAs that were amongst the top 85% sequence reads in at least one sample, to limit our analysis to regulatory important miRNAs and potential biomarkers. The EdgeR algorithm compensates for the difference in sequence reads between samples, as well as the different number of samples within the categories compared.

First, we compared normal samples to HER2-positive IDC and DCIS samples. sf-miR-21(1) and cluster-mir-142(1) members (sf-miR-142-3p(1), sf-miR-142-5p(1)) had higher abundance in IDC compared to normal breast with a p-value <0.001 (Figure 4A,
miRNA sequence and expression analysis in breast tumors

Table 1 and Table S6). Cluster-mir-98(13) members (sf-miR-125a(3), sf-miR-99a(3)), sf-miR-22(1), and cluster-mir-143(2) members (sf-miR-145(1), sf-miR-143*(1)), sf-miR-378(1), sf-miR-497(1), sf-miR-320(1), and cluster-mir-144(2) members (sf-miR-451(1), sf-miR-144(1)) were abundant miRNAs in normal breast reduced in IDC. miRNAs present both in ductal cell lines and tumors likely represent tumor down-regulated miRNAs (cluster-mir-98/let-7(13) and cluster-mir-22(1)). miRNAs absent in ductal cell lines likely reflect differences in tissue composition (i.e. adipose specific cluster-mir-143(2) is likely related to the presence of adipose tissue in the biopsies) (9). Most miRNA changes in IDC were already apparent in DCIS samples.

To identify miRNAs altered in tumor invasion, we compared HER2-positive DCIS and IDC samples. cluster-mir-142(1) members (sf-miR-142-3p(1) and sf-miR-142-5p(1)), were over-represented in HER2-positive IDC compared to DCIS. Given that cluster-mir-142(1) is hematopoietic lineage specific, this most likely reflects a change in tissue composition (22).

We then employed EdgeR to determine whether miRNA levels correlated with IHC or special histological characteristics (Figure 4B and Suppl. Table S6). Differences in miRNA representation between IHC subtypes involved less abundant miRNAs. TNBC showed the largest number of miRNA level changes. Cluster-mir-17(12) member sf-miR-19a(3) (present in ductal cell lines), sf-miR-205(1) and sf-miR-146a(2) (lower-represented in ductal cell lines) were higher expressed in TNBC compared to either HER2- and/or ER-positive tumors. sf-miR-451(1) was less abundant in TNBC. Special histological TNBCs (atypical medullary, metaplastic, adenoid cystic and apocrine...
miRNA sequence and expression analysis in breast tumors

carcinomas) were differentiated from IDC TNBCs by lowly abundant sf-miR-224(1), absent from some ductal cell lines.

Sequence-specific biases in the efficiency of cDNA library preparation can distort the representation of individual miRNAs by number of sequence read counts in a reproducible manner. We used an equimolar pool of 770 synthetic miRNAs to address possible biases in our method (Hafner, unpublished data). This analysis did not show an over-representation above the median rf >5-fold for any of the miRNAs meeting our analysis cutoff, but did show under-representation >5-fold for miR-193a, miR-193b, miR-26b, 29c and miR-30b. Among the >5-fold under-represented miRNAs was miR-31, previously implicated in metastasis, which did not meet our analysis cutoff. These biases in absolute abundance would not affect the comparisons across sample groups. Supplementary Table S7 lists EdgeR comparisons for all detected miRNAs, irrespective of abundance, including a summary of our findings for miRNAs investigated in animal models (see introduction). miR-520c and miR-9 demonstrated statistically significant up-regulation in patients that developed metastasis, consistent with their proposed pro-metastatic roles.

miRNAs as prognostic markers within TNBC and HER2 positive tumors

We evaluated the prognostic capacity of miRNAs on 48 TNBC and 49 HER2-positive patients, for which clinical follow-up information (distant metastasis free survival and overall survival) was available with a mean follow-up of 5.6 years. Lowly represented cluster-mir-423(2) members (sf-miR-423-3p(1) and sf-miR-423-5p(1)), and cluster-mir-375(1)/sf-miR-375(1) were more abundant in patients that went on to develop metastasis,
miRNA sequence and expression analysis in breast tumors

while cluster-mir-184(1)/sf-miR-184(1) was less abundant (p values ≤0.001) (Figure 4C, Table S6). Kaplan-Meier analyses supported the findings for cluster-mir-423(2) (p=0.013) and sf-miR-184(1)/cluster-mir-184(1) (p=0.041) in HER2-positive patients (Suppl. Figure 3). Univariate and multivariate Cox regression indicated that cluster-mir-423(1) is an independent predictor of outcome in the presence of other clinical parameters (tumor size, grade and lymph node status; data not shown). However, this should be interpreted with caution since these p values do not account for multiple testing.

Analysis of miRNA sequence variation in clinical specimens and cell lines

miRNAs have been proposed to act as tumor suppressors or oncogenes, yet somatic mutations in cancer patients have not been detected that support this proposal. For identification of nucleotide variations relative to the reference human genome, we required >10 varied sequence reads covering a given position per sample, with >25% variation frequency (>40X coverage). Based on analysis of deep-sequencing data from a pool of 770 synthetic miRNAs (Hafner, unpublished data), 10% or higher variation frequency guaranteed that 98% of the identified variations were not random events due to sequencing errors, but likely due to mis-mapping between miRNAs similar in sequence. We excluded the 3’ most terminal residue of all sequence reads from somatic mutational analysis if it was altered because it frequently contained 3’ untemplated nucleotide addition (22).

We identified 144 distinct nucleotide variations located within 117 mature and star miRNA sequences. 109 variations occurred in the last two positions of the
miRNA sequence and expression analysis in breast tumors

predominant mature sequence read, likely representing instances of untemplated 3’ terminal addition that were insufficiently repressed by our computational approach of not considering the 3’ most nucleotide (102 variations represented changes into A or U) (Suppl. Table S8A/B). Further targeted analysis of the 3’ end variations is included in Supplementary Table S8C. None of the remainder 35 variations were detected in abundant miRNAs (≥1% rf). These sequence variations could represent RNA editing events (including deamination, polyuridylation, polyadenylation), SNPs or somatic mutations. The most common of the 35 variations observed in the mature and star miRNA sequences were A to G, likely representing A to I RNA editing by dsRNA-specific adenosine deaminases (22, 34, 35, 36). This was further supported by a well-represented unimodal distribution of the nucleotide variation frequency for miR-376a and miR-376c, previously reported as edited in normal tissues (35). miR-625 (detected in 17 samples), miR-497* (n=9) and miR-381 (member of cluster-mir-134(41), that also includes miR-376; n=17) could represent editing events not previously reported (Suppl. Figure 4, and 5). Deamination events were not observed in cell lines.

We detected two known SNPs (SNPdb version 131) in lowly abundant miR-196a-2* and miR-146a which have been studied in the context of breast cancer risk (rs11614913 and rs2910164) (37, 38, 39, 40). rs2910164 (C5G) was detected in five carcinomas, while rs11614913 (C18T) was detected in 39 samples including one normal breast sample and three cell lines. We identified 10 nucleotide variations that are candidate new SNPs, based on the tri-modal distribution of their variation frequency. Two of these variations were identified in miRNAs present within 85% of all miRNA reads in at least one sample: miR-181a (T19G) observed in 10 carcinomas and miR-185
miRNA sequence and expression analysis in breast tumors

(T16G) observed in three carcinomas. There was no evidence for somatic mutations in miRNAs.

Discussion

Do miRNAs hold potential as diagnostic and prognostic markers in breast cancer?

Our sequencing approach using a large diverse sample collection allowed us to evaluate miRNA deregulation in the context of miRNA abundance, sequence variation and tissue heterogeneity, important elements in identifying miRNAs that would be useful prognostic and diagnostic markers, or prioritizing miRNAs for further studies. The first study on miRNA deregulation in breast cancer by Iorio et al. used microarrays to compare a variety of breast carcinomas (n=76) to normal (n=10) breast tissue (24), identifying 17 up-regulated and 12 down-regulated miRNAs in carcinomas, and miRNAs differentially expressed in ER-positive (11 miRNAs) and PR-positive (7 miRNAs) samples. Follow-up studies validated some of these miRNAs (41), suggested >30 miRNAs differentiating tumor subtypes (42, 43), and defined the cell-type-specific localization of some of these miRNAs (44).

When comparing miRNA levels between normal breast and carcinomas we identified 10 abundant miRNA sequence families (≥1% rf) showing 3- to 12-fold changes (Table 1). We confirmed up-regulation of sf-miR-21(1) and down-regulation of cluster-mir-98(13) members, sf-miR-22(1), sf-miR-145(1), sf-miR-378(1), sf-miR-497(1), sf-miR-320(1), sf-miR-451(1), and identified up-regulation of sf-miR-142-3p(1) not previously reported. Oncogenic cluster-mir-17–92 member sf-miR-19a(3) demonstrated approximately three-fold higher levels in TNBC suggesting regulatory potential, but
miRNA sequence and expression analysis in breast tumors exhibiting a modest change for a robust diagnostic marker. Mostly lowly abundant miRNAs demonstrated changes between different categories of ER/HER2 IHC groups, challenging their direct regulatory role in ER/HER2-related pathways. Moreover, when comparing the potential of overall miRNA profiles for differentiating tumor types to the potential of mRNAs profiles, miRNAs did not further clarify the assignment.

In the two large patient studies (>100 patients) only miR-210 was shown to be inversely correlated with time to metastasis, disease-free and overall survival (28, 29). Foekens et al. focused on ER-positive lymph node-negative breast tumors, but also extended their findings to TNBC, while Camps et al. focused on ER-positive patients. Our study showed higher levels of miR-210 in invasive compared to non-invasive carcinomas, but did not confirm a significant prognostic role. Other studies investigating miRNAs in breast cancer metastasis based on cell line and animal models were validated with smaller (n~20) patient sample collections (15, 16, 26, 27). The differences in miRNA levels in patients that developed metastases involve lowly abundant miRNAs, challenging to translate into prognostic markers given the detection limits of currently available experimental methods.

By comparing ductal cell lines with ductal tumors we identified miRNAs expressed in ductal versus other cell types. Changes in miRNA levels reflect either up- or down-regulation in ductal cells likely signifying associated tumor cell oncogenic pathways, or changes in tissue composition, i.e. presence of lymphocytic infiltrate. Given this lineage specificity, miRNA levels may allow estimation of cell types present in heterogeneous biopsy samples to normalize molecular array-based diagnostic tests.
miRNA sequence and expression analysis in breast tumors

Understanding miRNA involvement in oncogenic processes.

Viewing miRNAs in the context of their abundance defined miRNAs whose levels suggest regulatory functions (1% rf roughly equivalent to 1,500 copies per cell). miRNA nucleotide variations are implicated in tumorigenesis; however, evidence for their significance is limited (45). In our study, nucleotide variations were not identified in highly abundant miRNAs. We were not able to detect statistically significant differences between the occurrence of sequence variations in normal breast compared to carcinomas (Suppl. Table S8). It is important, however, to note that lowly abundant miRNAs could be highly expressed in a subpopulation of cells responsible for tumor invasion.

We noted the drastic up-regulation of miR-21 in carcinomas, recently identified as an oncogene in mouse models (46, 47). We showed that miRNA levels were comparable in normal breast and tumors, suggesting that the down-regulation of other abundant miRNAs in tumors could be a consequence of competition for processing by the RNAi machinery. Up-regulation of miR-21 may drive tumorigenesis, both through a direct effect on targets involved in repressor functions, as well as down-regulation of tumor suppressor miRNAs, such as members of the cluster-mir-98(13), as suggested from our data. In conclusion, our abundance-based view of miRNA expression in breast cancer supports a focus on oncogenic miR-21, and miR-21 targets with potential tumor suppressor functions, as promising therapeutic targets. Given that miR-21 is up-regulated in many other malignancies, identifying the tumorigenic pathways it regulates has broad implications in oncology.

Acknowledgements
miRNA sequence and expression analysis in breast tumors

We thank Scott Dewell at the Rockefeller University Genomics Center; Iddo Ben-Dov and Sean McGeary for editing the manuscript.

References

miRNA sequence and expression analysis in breast tumors


miRNA sequence and expression analysis in breast tumors


miRNA sequence and expression analysis in breast tumors


miRNA sequence and expression analysis in breast tumors


miRNA sequence and expression analysis in breast tumors


Figure and Table Legends

**Figure 1:** Unsupervised hierarchical clustering with complete linkage and Spearman correlation for patient samples conducted using the miRNA sequence families making up 85% of the sequence reads within each sample. Color histogram represents miRNA rf standardized across each miRNA.

**Figure 2:** Unsupervised hierarchical clustering with complete linkage and Spearman correlation depicting (A) the 85% top expressed mature miRNAs, (B) miRNA genomic clusters, and (C) sequence families. Every clustering includes 4 samples from each ER/HER2 IHC category, normal breast and 6 cell lines. Color histogram represents log2 miRNA abundance (rf). Examples of lineage specific miRNA genomic clusters are noted.
Figure 3: Comparison of clustering using miRNA sequence families (characterizing seed-sequence-dependent miRNA target regulation) and mRNA profiles. A. Unsupervised hierarchical clustering performed for miRNA sequence families, using the top 98% expressed families within each sample (221 families) (Spearman correlation). B. Unsupervised hierarchical clustering of 161 samples with available mRNA profiles, using the 221 genes with the highest variance (Pearson correlation).

Figure 4: Results of EdgeR comparison analysis between groups of samples. Results plotted as log$_2$ of the fold change between normal and/or tumor categories as a function of the log$_2$ of the average miRNA abundance in the two categories compared. Colored dots represent miRNAs that are significantly differentially expressed (p value<0.001). miRNAs in green are over-expressed, while miRNAs in red are under-expressed in the second category of samples within each comparison. Abundant miRNAs (log$_2$ concentration $\geq$-8) are labeled in black font, while low abundant miRNAs are labeled in grey font. A. miRNA sequence families (sf) differentially represented in tumor stages. B. miRNA sf that differentiate tumor IHC types. C. miRNA sf that predict metastasis in IDC, also separately evaluated in HER2 positive and TNBC patients.

Table 1: Comparison between normal, DCIS and IDC specimens using EdgeR.
<table>
<thead>
<tr>
<th>miRNA levels in carcinomas</th>
<th>miRNA genomic cluster</th>
<th>Normal (rf)</th>
<th>DCIS (rf)</th>
<th>HER2 positive IDC (rf)</th>
<th>Normal vs DCIS fold change (p value)</th>
<th>Normal vs HER2 positive IDC fold change (p value)</th>
<th>DCIS vs HER2 positive IDC fold change (p value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Downregulated miRNAs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sf-miR-22(1)</td>
<td>cluster-mir-22(1)</td>
<td>0.120</td>
<td>0.033</td>
<td>0.042</td>
<td>3.7 (0.0006)</td>
<td>2.9 (0.0008)</td>
<td>0.8 (1)</td>
</tr>
<tr>
<td>sf-miR-125a(3)</td>
<td>cluster-mir-98(13)</td>
<td>0.088</td>
<td>0.029</td>
<td>0.013</td>
<td>3.1 (0.007)</td>
<td>6.8 (1.3E-13)</td>
<td>2.2 (0.02)</td>
</tr>
<tr>
<td>sf-miR-99a(3)</td>
<td>cluster-mir-98(13)</td>
<td>0.066</td>
<td>0.018</td>
<td>0.015</td>
<td>3.6 (0.0008)</td>
<td>4.5 (3.9E-08)</td>
<td>1.2 (1)</td>
</tr>
<tr>
<td>sf-let-7a-1(12)</td>
<td>cluster-mir-98(13)</td>
<td>0.148</td>
<td>0.076</td>
<td>0.120</td>
<td>2.0 (1)</td>
<td>1.2 (1)</td>
<td>0.6 (1)</td>
</tr>
<tr>
<td>sf-miR-451-DICER1(1)</td>
<td>cluster-mir-144(2)</td>
<td>0.053</td>
<td>0.009</td>
<td>0.004</td>
<td>5.7 (2.4E-07)</td>
<td>11.8 (8.8E-23)</td>
<td>2.1 (0.06)</td>
</tr>
<tr>
<td>sf-miR-144(1)</td>
<td>cluster-mir-144(2)</td>
<td>0.003</td>
<td>0.000</td>
<td>0.000</td>
<td>2.6 (0.067)</td>
<td>7.0 (5.67E-14)</td>
<td>2.7 (0.0005)</td>
</tr>
<tr>
<td>sf-miR-145(1)</td>
<td>cluster-mir-143(2)</td>
<td>0.046</td>
<td>0.015</td>
<td>0.009</td>
<td>3.1 (0.006)</td>
<td>4.9 (3.0E-09)</td>
<td>1.6 (1)</td>
</tr>
<tr>
<td>sf-miR-143(1)</td>
<td>cluster-mir-143(2)</td>
<td>0.042</td>
<td>0.043</td>
<td>0.022</td>
<td>1.0 (1)</td>
<td>1.9 (0.559)</td>
<td>1.9 (0.130)</td>
</tr>
<tr>
<td>sf-miR-143*1(1)</td>
<td>cluster-mir-143(2)</td>
<td>0.008</td>
<td>0.001</td>
<td>0.001</td>
<td>8.0 (4.48E-10)</td>
<td>7.9 (6.54E-16)</td>
<td>1.0 (1)</td>
</tr>
<tr>
<td>sf-miR-320-RNASEN(1)</td>
<td>cluster-mir-320-RNASEN(1)</td>
<td>0.036</td>
<td>0.006</td>
<td>0.005</td>
<td>6.5 (2.5E-08)</td>
<td>7.7 (1.2E-15)</td>
<td>1.2 (1)</td>
</tr>
<tr>
<td>sf-miR-378(1)</td>
<td>cluster-mir-378(1)</td>
<td>0.022</td>
<td>0.004</td>
<td>0.003</td>
<td>5.7 (3.3E-07)</td>
<td>8.5 (4.6E-17)</td>
<td>1.5 (1)</td>
</tr>
<tr>
<td>sf-miR-497(1)</td>
<td>cluster-mir-195(2)</td>
<td>0.011</td>
<td>0.003</td>
<td>0.003</td>
<td>4.3 (0.00004)</td>
<td>4.2 (1.7E-07)</td>
<td>1.0 (1)</td>
</tr>
<tr>
<td>sf-miR-16-1(3)</td>
<td>cluster-mir-195(2)</td>
<td>0.002</td>
<td>0.007</td>
<td>0.006</td>
<td>0.3 (0.012)</td>
<td>0.3 (0.052)</td>
<td>1.3 (1)</td>
</tr>
<tr>
<td>Upregulated miRNAs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sf-miR-21(1)</td>
<td>cluster-mir-21(1)</td>
<td>0.035</td>
<td>0.365</td>
<td>0.357</td>
<td>10.4 (2.5E-09)</td>
<td>10.2 (1.4E-09)</td>
<td>1.0 (1)</td>
</tr>
<tr>
<td>sf-miR-142-3p(1)</td>
<td>cluster-mir-142(1)</td>
<td>0.002</td>
<td>0.003</td>
<td>0.010</td>
<td>1.5 (1)</td>
<td>5.6 (0.00001)</td>
<td>3.9 (8.2E-06)</td>
</tr>
<tr>
<td>sf-miR-142-5p(1)</td>
<td>cluster-mir-142(1)</td>
<td>0.001</td>
<td>0.006</td>
<td>0.008</td>
<td>4.7 (0.0003)</td>
<td>5.9 (7.31E-06)</td>
<td>1.3 (1)</td>
</tr>
</tbody>
</table>

miRNA sf expressed at >1% rf, p value ≤0.001 for at least one comparison in bold
miRNA sf that belong to same miRNA genomic cluster irrespective of expression and p values in italics
MicroRNA sequence and expression analysis in breast tumors by deep sequencing

Thalia A Farazi, Hugo M Horlings, Jelle ten Hoeve, et al.

Cancer Res  Published OnlineFirst May 17, 2011.

Updated version  Access the most recent version of this article at: doi:10.1158/0008-5472.CAN-11-0608

Supplementary Material  Access the most recent supplemental material at: http://cancerres.aacrjournals.org/content/suppl/2011/05/17/0008-5472.CAN-11-0608.DC1

Author Manuscript  Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts  Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions  To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions  To request permission to re-use all or part of this article, use this link http://cancerres.aacrjournals.org/content/early/2011/05/13/0008-5472.CAN-11-0608. Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.