# microRNA associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer

Francesca M. Buffa, Carme Camps, Laura Winchester, Cameron E. Snell, Harriet E. Gee, Helen Sheldon, Marian Taylor, Adrian L. Harris, Jiannis Ragoussis

# Supplementary Information

## Patient characteristics

Individual samples from a retrospective series of 219 patients with early primary breast cancer (BC), who received surgery in Oxford between 1989 and 1993, were considered. Ethical approval for analysis of samples and notes was obtained from the local research ethics committee. This cohort has been described previously (1); demographics, treatment and assays details are provided in Table S1. Complete information was available for ER status, nodal status, histology and treatment. Tumor grade was available for 194 out of 219 cases and HER2 status was available for 129 patients. Patients received surgery followed by adjuvant chemotherapy, adjuvant hormone therapy, both of these therapies, or no adjuvant treatment. Tamoxifen was used as endocrine therapy for 5 years in all $ER^+$ BCs. Patients who were <50 years of age, with lymph node positive tumors, or $ER^-$ and/or >3 cm in diameter, received adjuvant cyclophosphamide, methotrexate, and 5-fluorouracil (CMF). Patients >50 years of age with $ER^-$, lymph node–positive tumors also received CMF for six cycles, at a thrice weekly intravenous regimen. Complete 10 year follow-up was available in all but 3 cases. Clinical endpoints considered were distant relapse-free survival (DRFS) and recurrence-free survival (RFS) as defined by the STEEP guidelines (2). Clinical covariates considered in Cox analyses were: age, tumor size, ER status, Grade (from low to high), nodal status (binary, nodes involved=1, no=0) and number of nodes involved, Tamoxifen treatment and chemotherapy treatment (Yes=1, No=0). As this is not a treatment-naive group, the factors identified by the Cox analysis, could be not solely prognostic but predictive of a specific treatment. However, these patients received different treatments, and treatment was

included as covariate in the Cox analysis to identify factors whose significance was independent of treatment.

Eight published expression signatures derived from large-scale analysis of cancer datasets were chosen as surrogate markers of critical processes in cancer biology and introduced in Cox analyses; these are described below ("Mapping of gene expression signatures and calculation of summary expression score"). The HER2 signaling signature was used in Cox analyses both as continuous and binary variable; mRNA levels were used to classify patients into HER2 positive and negative using a Bayesian clustering method (TwoStep clustering, SPSS 15.0). This classification showed >98% agreement with HER2 immunohistochemistry (IHC) staining which was available on a subset (N=129) of patients confirming published clinical studies (3). The same method was applied to identify a triple negative receptor (TNR) group using mRNA levels of ER, HER and PRg together with ER status as measured by IHC. Patients classified as negative by both gene expression of the three markers and ER IHC were selected as TNR. All TNR patients for which HER2 IHC was available were confirmed HER2 negative by IHC.

## RNA extraction, mRNA and microRNA microarrays

Total RNA was isolated from fresh frozen tumor samples by the Trizol method (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. Frozen samples were visually assessed and a large portion representative of the whole tumor was taken. No selection of tumor cells was applied, as there is an increasing number of studies providing evidence that signaling from the stroma has an important role in cancer progression. However, when we classified the samples based on mRNA expression using published methods (Table S1), no tumor samples were assigned to the "Normal" group that has been previously indicated to reflect samples with very high levels of normal tissue contamination. This suggests that the level of contamination in the present series is relatively contained.

RNA was extracted from cells using the miRVana miRNA Isolation Kit (Ambion). RNA was extracted and purified from liquid nitrogen–frozen breast tumor samples or normal breast tissue for the controls using Tri-reagent (Sigma-Aldrich) and ethanol precipitation. Only tumor samples with good quality of RNA were considered for further analysis. Specifically, in both cases RNA quality and abundance were determined after extraction using an Agilent 2100 Bioanalyzer (Agilent Technologies) and a Nanodrop ND-1000 spectrophotometer

(Nanodrop Technologies), respectively. A RIN number greater than 6 was considered as good quality sample.

mRNA expression was measured using Illumina Human RefSeq-8 arrays (illumina inc., San Diego, CA, USA). Illumina platforms were chosen as they performed well in large comparative studies (4, 5).
RNA was amplified using Ambion Illumina Amplification Kit (Catalog #I1755). 850ng of amplified RNA product was hybridized to Illumina Sentrix Beadchip 8x1 GAP REFSEQ2 using single chamber hybridization cartridges. Washing, staining and scanning were carried out as specified in the Illumina Whole Genome Expression Manual v.1. Average signal was background subtracted with local background subtraction (BeadStudio), quantile normalized in Bioconductor (www.bioconductor.org) and logged (base 2).
microRNA expression was measured using Illumina miRNA arrays v.1. 200ng of total RNA was polyadenylated and converted to biotinylated cDNA, which was attached to a solid phase and hybridized with a pool of microRNA-specific oligonucleotides (MSO). Universal PCR amplification was performed, creating fluorescently labeled products identifiable by MSO unique sequence. These were hybridized on the array; signals were detected and quantified using Illumina scanner and BeadStudio. Data were quantile normalized and logged (base 2). miRBase (http://www.mirbase.org/) human microRNAs were considered.

## Real-time reverse transcription PCR

Although Illumina arrays have been shown to be accurate and reproducible by a recent large comparative study (5), we validated the expression measurement obtained with the arrays by performing real-time PCR and this represents an alternative technique to measure expression which has been generally used to validate results from array data.
microRNA expression was assessed by real-time PCR with TaqMan MicroRNA assay protocol (Applied Biosystems) using 5 ng total RNA per gene (for full details and methods see (1)). Expression values were normalised to the geometrical mean of RNU43, RNU44 and RNU48 using the ΔΔCT methods as implemented in the SLqPCR Bioconductor package (http://www.bioconductor.org/packages/2.2/bioc/html/SLqPCR.html). Furthermore, due to lack of consensus on normalization of RT-PCR experiments measuring microRNA expression, we have also carried out analyses by using non normalized RT-PCR CT values.

## Meta-analysis criteria and GEO breast cancer datasets

Literature was reviewed and studies to be included in the meta-analysis were selected and assessed for suitability. Specifically, NCBI Gene Expression Omnibus, GEO (http://www.ncbi.nlm.nih.gov/geo/) was searched for gene expression studies in breast cancer using platforms where most of the microRNAs precursors and targets could be matched. GEO was chosen as it is a comprehensive and MIAME complaint database, and raw data are often published, in particular for Affymetrix arrays, so that data can be processed uniformly. Thus, studies with Affymetrix U133A, U133 B or plus2, were considered (www.affymetrix.com). These were chose because Affymetrix U133 arrays very good agreement with Illumina arrays in the MAQC comparative study (4). Other necessary requirements were that the studies were published in peer-reviewed journals, clinical outcome was published and available for analysis, and they did not have cases in common. For one dataset which was overlapping with the dataset used in this study, only the non-overlapping set of samples were considered, thus samples not included in the present cohort of 207 patients (see Table S2 for details). Datasets that passed these criteria are summarized in Table S2.

Raw gene expression data from GEO (http://www.ncbi.nlm.nih.gov/sites/geo/) datasets were retrieved and processed as described previously (6). Briefly, processing was performed using 'simpleaffy' (7); the 'gcrma' function was used to estimate expression values, data were quantile normalised and logged (base2). Affymetrix annotation was used; Ensemble IDs were used as common identifiers for target analysis. Matching of Affymetrix and Illumina data was done as described in the Section "Mapping of gene expression signatures and calculation of summary expression score". Survival Cox analyses for these datasets were performed as described previously (6).

## Mapping of published signatures and calculation of Signature Summary Scores

Eight published expression signatures derived from large-scale analysis of cancer datasets were chosen as surrogate markers of critical processes in cancer biology. These were: proliferation, ERS1 and HER2 signaling (for these three signatures, genes in common between two studies were used (8, 9)), stem cell (10), invasion, immune-response, apoptosis (9) and hypoxia (6).

Affymetrix probes IDs for these signatures were retrieved from the original publications and mapped to the Illumina gene expression arrays used in this study using NCBI RefSeq13 database. Perfect matches with the same Entrez gene ID were considered. Affymetrix probe sets were defined as matching if at least 80% of the probes were perfect matches; where multiple probe/probesets matched a transcript the one closest to the 3' end of the transcript was selected. Probes for which a match was not found were filtered out. When published Affymetrix datasets were considered for the analysis, the published probesets were used. For each signature, a summary signature expression score was computed as described in the original publications. The signature summary scores were then ranked and normalized between 0 (low) and 1 (high); this ranked score was introduced in the linear regression analysis and Cox analysis.

## Penalized regression

An expression profiling analysis approach that includes also clinic-pathological covariates allows one to identify microRNAs associated with clinical outcome (Section A below), or a specific clinical covariate (Section B), independently from other microRNAs and clinical covariates. Thus, it limits identification of spurious associations due to inherent heterogeneity of tumor populations. How this approach relate to using univariate approaches, such as significance analysis of microarrays (SAM), has been discussed previously (11-13). Briefly, in univariate analysis a different model is built for each predictor variable ignoring the dependence between predictors and requiring multiple test correction for variable selection; in contrast, in a multiple covariate analysis all predictors are tested simultaneously in the same regression model.

Gene expression datasets are characterized by a much greater number of covariates than samples and a high structure (i.e. some of the covariates are highly correlated). Thus, in order to prevent over-fit, a regression approach should be chosen that performs efficient variable selection and regularization (i.e. encourages a grouping effect where highly correlated predictors tend to stay in or out of the model together). Here, we used L1- and L2-penalized linear regression (11, 12), a method that performs least-square minimization whilst enforcing a constraint on a combination term including the sum of the absolute values of the regression coefficients and the sum of their squares (for details see (11)). This approach is particularly powerful with respect to other linear regression methods when the size of the study (i.e. the number of cases) is small with respect to the number of covariates, and when some of the covariates are highly correlated (11).

In standard stepwise selection procedures, the model size (i.e. the number of selected covariates) depends on the choice of the p-value required for selection; on the contrary, in penalized regression the model size depends on the penalty applied to the least-squares which is estimated using k-fold cross-validation (see (11) for more details). Briefly, in k-fold cross-validation the data are split in k equally sized parts (or folds); k-1 folds are pooled and constitute the training set. The model is trained on this set and a prediction is made for the kth fold, the validation set. The error on this prediction is estimated and the procedure is repeated iteratively until all the k folds have been used for training and prediction. The prediction errors from the k datasets are combined and the cross-validation error estimate is obtained for each model size; the model size (or penalty) that results in the lowest cross-validation error is selected. We implemented cross-validation as described previously (11). We also used iterative leave-one-out to test stability, bias and performance of the cross-validated models as done in previous marker study by our group (14). Specifically, at each step one case (i.e. a single patient) is left out from the analysis, a cross-validated model is selected as described above using the remaining cases and prediction is made for the left-out case.

The specific steps for the cross-validation and leave-one-out process were:

1) leave one case out of the analysis
2) for the rest of the cases perform cross-validation (see also Sweave files for details of implementation):
   2.1) select a grid of values for L2-parameter (0, 0.01, 0.1, 1, 10, 100),
   2.2) tune for each value the L1-parameter by cross-validation,
   2.3) select the L2-parameter value which gave the smallest cross-validation error
3) fit the model with the L1- and L2-penalization selected by cross-validation and estimate covariate effects
4) use the cross-validated model to predict status of the left out case

The above steps were implemented in R (http://cran.r-project.org) and the glmpath 0.94 package (http://cran.r-project.org/web/packages/glmpath) was used for penalized regression, with a combined L1- and L2-penatly as previously described (12). Sweave was used for Automatic Generation of Reports (see http://www.stat.uni-uenchen.de/~leisch/Sweave).

In the results section we report various statistics for our analysis (see e.g. Fig. S2). The main ones are:-

- the effect estimate for each covariate in the cross-validated model, in each leave-one-out iteration (**c)**, its full distribution and related summary statistics such as and average effect

over all leave-one-out iterations (**C**). This is the coefficient estimate for each covariate in the cross-validated model and indicates the magnitude of its effect:- c=0 no effect in the cross-validated model (i.e. no association with the dependent variable, e.g. *x* in Eq. SM1 below), c<0 negative effect, or c>0 positive effect. Relative to other regression methods, penalized regression tends to shrink the coefficient estimates, c. Specifically, a pure L2 penalty tends to result in all coefficients to be small but non-zero, whilst a pure L1 penalty tends to result in many coefficients shrunk exactly to zero and a few other coefficients with relatively little shrinkage. A combined L1- and L2-penality, as applied here, gives a result in between, with fewer regression coefficients shrunk to zero than in a pure L1 setting, and a greater shrinkage of the other coefficients (for more details see (11)). The exact amount of shrinkage is determined by the L1 and L2 parameters, and is optimized using cross-validation as described above.

- the fraction of leave-one-out iterations in which the variable is selected in the cross-validated model (**F**). This measures the stability (or consistence) of the models in the leave-one-out process. It varies between 0 and 1. Specifically, 0 is the worst scenario and indicates that simply leaving one case out from the analysis causes the variable not to be selected in the cross-validated model generated using all other cases. At the other end of the spectrum, 1 is the most consistent scenario and indicates that no matter which sample is taken out of the analysis the variable is always selected in the cross-validated model. In this study, variables with F>0.95 (thus selected in the cross-validated model in at least 95% of the leave-one-out iterations) were selected for further analysis unless otherwise stated. For a general schema of how the approach was applied to this study see Figure S1; below are the details for Cox and linear association analyses respectively.

## Cox regression using penalization

The above described penalized regression approach was applied to the Cox survival analysis. Specifically, a Cox model assumes that the log of the hazard for a given outcome (e.g. DRFS, in our case) is a linear model of the clinical covariates and that the ratio of the hazards for any two observations (i.e. cases) is independent of time. The baseline hazard in a Cox model is left unspecified and the model can be estimate using partial likelihood methods. The model is typically reduced using stepwise regression methods and a Cox model is obtained including only the significant covariates. The Cox models in our analyses were optimized, and reduced models were obtained, by using the penalized regression approach with cross-validation and

leave-one-out as described above. The function *coxpath* in the glmpath 0.94 package (12) was used.

The analysis was performed in two steps.

Firstly, microRNAs whose expression was correlated with DRFS independently from other microRNAs were selected. This was done using a Cox model where the log of the hazard (*log h*) of the clinical endpoint (DRFS in this study) was expressed as a linear function of the expression levels of all *N* microRNAs:

$$\log h = c_0 + \sum_{n=1}^{N} c_n \cdot e_n \qquad \text{[Eq. SM1]}$$

where the coefficients in the model represent the log Hazard Ratio (log HR) of survival (DRFS in this study) for the N microRNAs. A positive coefficient (corresponding to HR > 1) indicates that high values of the covariate are associated with a higher likelihood of bad prognosis, vice versa for negative coefficients (corresponding to HR < 1). L-1 and L2-penalized regression was used to selected prognostic microRNAs in this model as described in the previous section. Pre-screening was used to make the computation more manageable (11); specifically, microRNAs associated with DRFS in univariate analysis ($p<0.005$ in univariate Cox model where microRNA expression was ranked and normalized between 0 and 1, this was done to overcome the effect of outliers) were selected each time that a model was cross-validated (i.e. in each leave-one-out training set). In the Cox linear regression all covariates were standardized to have unit variance. microRNAs that were consistently selected in the cross-validated models (ie. selected in at least 95% of the leave-one-out iterations) were further analyzed.

Second step of the analysis considered a Cox analysis accounting for microRNA expression, clinical covariates and key breast cancer biological processes. Only microRNA that passed selection in the first step were considered. The log of the hazard (*log h*) of the clinical endpoint (DRFS) was expressed as a linear function of the *ith* microRNA expression, $e_i$, the *Np* gene signatures of biological processes ($x_{1,\ldots.}\ x_n$) and *Nc* clinical covariates ($y_1,\ldots.y_n$):

$$\log h(e,\bar{x},\bar{y})_i = a_i + d_i \cdot e_i + \sum_{n=1}^{N_p} b_i^n \cdot x_n + \sum_{n=1}^{N_c} c_i^n \cdot y_n \qquad \text{[Eq. SM2]}$$

where the coefficients in the model represent the log Hazard Ratio (log HR) of DRFS for the respective covariate. Clinical covariate and gene signature of key biological processes in breast cancer described above were considered. L-1 and L2- penalized regression was used to selected prognostic microRNAs in this model as described in the previous section.

microRNAs were considered as consistently independently associated with DRFS when they were consistently selected in both the first and second above described steps.

## Association Analyses using penalized linear regression

Association of expression of prognostic microRNAs with clinico-pathological factors and gene signatures of key biological processes in breast cancer was analyzed using penalized linear regression analysis. Specifically, for each of these microRNAs a model was built where the expression of the $i^{th}$ microRNA was considered as a function of the expression of the *Nc* clinical variables and the *Np* gene signatures of biological processes:

$$e^i(x,y) = a^i + \sum_{n=1}^{Np} b_n^i \cdot x_n + \sum_{n=1}^{Nc} c_n^i \cdot y_n \qquad \text{[Eq. SM3]}$$

covariates in the model were the *Np* signature scores ($x_{1,\ldots}\ x_n$) and the values ($y_1, \ldots y_n$) of the *Nc* clinical covariates in the patient population:- age (decades), tumour size (cm), ER positivity, Grade, Number of Nodes Involved, Tamoxifen treatment, Chemotherapy.
L1- and L2-penalized regression analysis was used as described in the previous sessions to selected covariate associated with the microRNAs. Predictors were standardized to have unit variance.

## Comprehensive list of *in-silico* predicted targets

Several algorithms have been suggested for target prediction. However, results from these algorithms present significant discrepancies, and accurate target prediction is still challenging (15). Discrepancies are due to differences in implementation, different requirements for site conservation across species and different hypotheses regarding the microRNA action on its target genes. Several different microRNAs algorithms have been compared in various studies (15). Algorithms used in this study were: TargetScan v5.1 (16); Pictar(17); MiRanda v5 (18); DianaLab microT v3.0 (19), miRTarget2 (20) and miBridge (21). For microT v3.0 a low stringency score threshold of 2 was used (19). miRTarget2 was used as implemented in miRDB (22); this algorithm includes also predictions from TargetScan, Pictar and Miranda and selection of targets using microarray expression data. Ensemble IDs were used as common identifiers. Below we list schematically the main characteristics and differences between algorithms used in this study.

**Characteristics:**

- **Sequence**
  - Perfect seed match rule: **TargetScan**

- Preference for perfect seed match: **Pictar**
- Empirically determined binding rules: **DianaLab**
- Dynamic programming alignment score cutoff: **miBridge, miRanda**
- Seed 5' and/or 3' flank requirements: **TargetScan, miRTarget2 (miRDB)**

❖ **Thermodynamics**

- ΔG calculations using traditional RNA folding programs: **DianaLab**, **miRanda, miRTarget2 (miRDB)**
- ΔG calculations: short nucleic acid hybridization programs: **miBridge, Pictar**

❖ **Conservation**

- Among human, rodent, and chicken: **miRTarget2 (miRDB)**
- Among human, rodent, and dog: **miBridge**
- Among human, chimp, rodent, and dog: **TargetScan, DianaLab, miRanda, Pictar**
- Residing in an 'island' of conservation: **TargetScan, DianaLab**

In addition, miBridge predicts microRNA targets containing simultaneous 5'- and 3'-UTR interaction sites.

In this study, we build on previous results and in particular on the conclusion common to multiple studies that using the union of the predictions from these algorithms rather than the intersection provides a more promising strategy to select true targets (15, 23). Thus, the union of the predicted targets from all algorithms was used in the analyses as follows:

1) Map microRNAs to the databases relative to the different algorithms
2) If microRNAs doesn't have a match in database then either introduce as missing or if possible use target prediction for custom seeds (available in TargetScan)
3) For each database compute a ranked strength of prediction. Specifically, this was a continuous rank (from 0=lowest algorithm score to 1=top algorithm score) or binary score (0=not predicted by algorithm, 1=predicted) depending on the "conservativeness" of the algorithm; the first was used for MiRanda and DianaLab, the second for others.
4) Calculate prediction meta-score. Specifically, the overall prediction score was calculated for each transcript as average of the ranks. This score was used as a weight in the target analyses (see next two sections), but not as selection criteria: the union of the predictions from all algorithms was always considered.

## Global expression analysis of microRNA predicted targets

While microRNAs main function is to inhibit translation they also promote mRNA degradation, and although the effect on each individual transcript might be small, evidence for significant global down-regulation of target mRNAs has been reported (24, 25). We therefore investigated our dataset for relationships between the expression of hypoxia-associated microRNAs, that of their target transcripts, and hypoxia itself.

All transcripts on the gene expression arrays (Illumina in this study, or Affymetrix in published studies considered) were ranked by their correlation with microRNA expression. The cumulative Relative Risk (RR) of observing a given number of targets whose expression is inversely correlated with that of the targeting microRNA *i*, was calculated at each correlation level, *r*:

$$RR_i(r) = \frac{\sum_{j=1}^{Nt} \chi\left(p_i^j > 0\right) \cdot \chi\left(r_i^j - r \le 0\right)}{\sum_{j=1}^{Nt} \chi\left(r_i^j - r \le 0\right)} \Bigg/ \frac{\sum_{j=1}^{Nt} \chi\left(p_i^j > 0\right) \cdot \chi\left(r_i^j - r \ge 0\right)}{\sum_{j=1}^{Nt} \chi\left(r_i^j - r \ge 0\right)} \qquad \text{[Eq. SM4]}$$

where $N_t$ is the number of transcripts in the array, $p_i^j$ is the target prediction score (see previous section) for transcript *j* relative to microRNA *i*, $r^j_i$ is the correlation between microRNA *i* and transcript *j* expression levels, $\chi$ is the indicator function: $\chi$(true)=1 and $\chi$(false)=0.

A schematic representation of the method is also provided in Figure S3B, bottom panel.

To estimate RR significance, a random simulation was carried out. Specifically, set of transcripts of the size of the predicted target lists were randomly sampled, RR was calculated and this was repeated 100 times. A hypergeometric test was also performed to test the probability of obtaining by chance the observed, or greater, number of targets inversely correlated with the microRNA (with correlation $\ge r$) for any given correlation *r*. The alternative hypothesis is enrichment, so this is a one-sided test with critical region right.

As the next step, we extend this to analysis of clinical covariates. If a microRNA is associated with a given clinical covariate/biological process/clinical outcome, and if a global inverse relationship between microRNA and target expression can be detected (i.e. RR>1 in Eq. SM4), we would expect to detect association between the targets and the clinical covariate/biological process/clinical outcome. Specifically, for each microRNA, *i*, up-regulated in tumors showing activation of a given biological process, *k*, (or poor prognosis) we estimated the RR of observing a given number of targets whose expression is inversely correlated with the status of the biological process (or outcome, that is high values associated with good prognosis), at each correlation level *r* (or log Hazard Ratio, HR, level):

$$RR_i^k(r) = \frac{\sum_{j=1}^{Nt} \chi\left(p_i^j > 0\right) \cdot \chi\left(r_k^j - r \le 0\right)}{\sum_{j=1}^{Nt} \chi\left(r_k^j - r \le 0\right)} \Bigg/ \frac{\sum_{j=1}^{Nt} \chi\left(p_i^j > 0\right) \cdot \chi\left(r_k^j - r \ge 0\right)}{\sum_{j=1}^{Nt} \chi\left(r_k^j - r \ge 0\right)} \qquad \text{[Eq. SM5]}$$

where $r_k^j$ is the correlation between the summary signature score $k$ and expression of transcript $j$ (or the log HR for transcript $j$), and the other symbols are as defined above. On the contrary, for microRNAs downregulated in tumors showing activation of a given biological process (or good prognosis), the opposite relation was expected, thus the reciprocal RR was estimated:

$$RR_i^k(r) = \frac{\sum_{j=1}^{Nt} \chi\left(p_i^j > 0\right) \cdot \chi\left(r_k^j - r \ge 0\right)}{\sum_{j=1}^{Nt} \chi\left(r_k^j - r \ge 0\right)} \Bigg/ \frac{\sum_{j=1}^{Nt} \chi\left(p_i^j < 0\right) \cdot \chi\left(r_k^j - r \le 0\right)}{\sum_{j=1}^{Nt} \chi\left(r_k^j - r \le 0\right)} \qquad \text{[Eq. SM6]}$$

A summary expression score of the predicted target signature (PTSign) was estimated by calculating in each patient sample, *pt*, the weighted average global expression of predicted targets, where the weight was set to the target prediction score $p$ (or 1, for experimentally validated targets):

$$PTsign_{pt}^{\ i} = -\left( \frac{\sum_{j=1}^{Nt} p_i^j \cdot e_{pt}^j}{\sum_{j=1}^{Nt} p_i^j} \right) \qquad \text{[Eq. SM7]}$$

where $N_t$ is the number of transcripts in the array, $p^j{}_i$ is the target prediction score (see previous section) for transcript $j$ relative to microRNA $i$, and $e^j_{pt}$ is the expression of transcript $j$ in patient *pt*.

**References**


1. Camps C, Buffa FM, Colella S, et al. hsa-miR-210 Is induced by hypoxia and is an independent prognostic factor in breast cancer. Clin Cancer Res 2008;14:1340-8.
2. Hudis CA, Barlow WE, Costantino JP, et al. Proposal for standardized definitions for efficacy end points in adjuvant breast cancer trials: the STEEP system. J Clin Oncol 2007;25:2127-32.
3. Gong Y, Yan K, Lin F, et al. Determination of oestrogen-receptor status and ERBB2 status of breast carcinoma: a gene-expression profiling study. Lancet Oncol 2007;8:203-11.
4. Shi L, Reid LH, Jones WD, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 2006;24:1151-61.
5. Git A, Dvinge H, Salmon-Divon M, et al. Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. RNA 2010;16:991-1006.

6. Buffa FM, Harris AL, West CM, Miller CJ. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. British Journal of Cancer 2010;102:428-35.
7. Wilson CL, Miller CJ. Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. Bioinformatics 2005;21:3683-5.
8. Wirapati P, Sotiriou C, Kunkel S, et al. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Res 2008;10:R65.
9. Desmedt C, Haibe-Kains B, Wirapati P, et al. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. Clin Cancer Res 2008;14:5158-65.
10. Ben-Porath I, Thomson MW, Carey VJ, et al. An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nat Genet 2008;40:499-507.
11. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Statist Soc B 2005;67:301-20.
12. Park MY, Hastie T. L1-regularization path algorithm for generalized linear models. J R Statist Soc B 2007;69:659-77.
13. Wu B. Differential gene expression detection and sample classification using penalized linear regression models. Bioinformatics 2006;22:472-6.
14. Generali D, Buffa FM, Berruti A, et al. Phosphorylated ERalpha, HIF-1alpha, and MAPK signaling as predictors of primary endocrine treatment response and resistance in patients with breast cancer. J Clin Oncol 2009;27:227-34.
15. Sethupathy P, Megraw M, Hatzigeorgiou AG. A guide through present computational approaches for the identification of mammalian microRNA targets. Nat Methods 2006;3:881-6.
16. Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 2009;19:92-105.
17. Lall S, Grun D, Krek A, et al. A genome-wide map of conserved microRNA targets in C. elegans. Curr Biol 2006;16:460-71.
18. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. Nucleic Acids Res 2008;36:D154-8.
19. Maragkakis M, Reczko M, Simossis VA, et al. DIANA-microT web server: elucidating microRNA functions through target prediction. Nucleic Acids Res 2009;37:W273-6.
20. Wang X, El Naqa IM. Prediction of both conserved and nonconserved microRNA targets in animals. Bioinformatics 2008;24:325-32.
21. Lee I, Ajay SS, Yook JI, et al. New class of microRNA targets containing simultaneous 5'-UTR and 3'-UTR interaction sites. Genome Res 2009;19:1175-83.
22. Wang X. miRDB: a microRNA target prediction and functional annotation database with a wiki interface. RNA 2008;14:1012-7.
23. Ritchie W, Flamant S, Rasko JE. Predicting microRNA targets and functions: traps for the unwary. Nat Methods 2009;6:397-8.
24. Arora A, Simpson DA. Individual mRNA expression profiles reveal the effects of specific microRNAs. Genome Biol 2008;9:R82.
25. Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. Nature 2008;455:58-63.