# Supplementary Notes

## MicroRNA Expression Profiles Classify Human Cancers

Jun Lu, Gad Getz, Eric Miska, Ezequiel Alvarez-Saavedra, Justin Lamb, David Peck, Alejandro Sweet-Cordero, Benjamin L. Ebert, Raymond H. Mak, Adolfo A. Ferrando, James R. Downing, Tyler Jacks, H. Robert Horvitz and Todd R. Golub

## *Contents*

## *Online Information*

Additional information about the paper and a frequently-asked-questions (FAQ) page are available at http://www.broad.mit.edu/cancer/pub/miGCM .

## *Supplementary Figure Legends*

**Supplementary Figure 1** Schematics of target preparation and bead detection of miRNAs. (Left panel) 18 to 26-nucleotide (nt) small RNAs were purified by denaturing PAGE (polyacrylamide gel electrophoresis) from total RNAs extracted from tissues or cells. Small RNAs underwent two steps of adaptor ligation utilizing both the 5'-phosphate and 3'-hydroxl groups, each followed by a denaturing purification. Ligation products were reverse-transcribed (RT) and PCR amplified using a common set of primers, with biotinylation on the sense primer. (Right panel) Denatured targets were hybridized to beads coupled with capture probes for miRNAs. After binding to streptavidin-phycoerythrin (SAPE), the beads went through a flow cytometer that has two lasers and is capable of detecting both the bead identity and fluorescence intensity on each bead.

**Supplementary Figure 2** Specificity and accuracy of the bead-based miRNA detection platform, probe similarity (for Fig. 1). Eleven synthetic oligonucleotides corresponding to human *let-7* family of miRNAs or mutants were PCR-labelled. Each of the labelled targets was split and hybridized separately on the bead platform and on a glass microarray. The synthetic targets are indicated on the horizontal axis, and the capture probes are indicated on the vertical axis. The similarity of the capture probes are measured by the differences in nucleotides (nt) and indicated by shades of blue.

**Supplementary Figure 3** Noise and linearity of bead detection of miRNAs. **(a)** The noise of target preparation and bead detection was analyzed. Multiple analyses of the same RNA samples were performed. Expression data were $\log_2$-transformed after thresholding at 1 to avoid negative numbers. The standard deviation (std) of each miRNA was plotted against the mean of that miRNA. Data were generated from independent labeling reactions and detections of five replicates of MCF-7, four replicates of PC-3, three replicates of HEL, three replicates of TF-1 and three replicates of 293 cell RNAs. Note that most miRNAs have a standard deviation below 0.75 when their mean is above 5 (in $\log_2$ scale). **(b)** Linearity of target preparation and bead detection. miRNAs were labeled and profiled from HEL cell total RNA with different starting amounts (10 μg, 5 μg, 2 μg and 0.5 μg, respectively). Data are averages of duplicate determinations, measured in median fluorescence intensity (MFI). Each line connects the readings of one miRNA with different amounts of starting material.

**Supplementary Figure 4** Unsupervised analysis of miRNA expression data. miRNA profiling data of 218 samples covering multiple tissues and cancers were filtered, and centred and normalized for each feature. The data were then subjected to hierarchical clustering on both the samples (horizontally oriented) and the features (vertically oriented, with probe names on the left), with average-linkage and Pearson correlation as a similarity measure. Sample names (staggered) are indicated on the top and miRNA names on the left. Tissue types and malignancy status (MAL; N for normal, T for tumor and TCL for tumor cell line) are represented by colored bars. Samples that belong to the epithelial origin (EP) or derived from the gastrointestinal tract (GI) are also annotated below the dendrogram. STOM: stomach; PAN: pancreas; KID: kidney; PROST: prostate; UT: uterus; MESO: mesothelioma; BRST (breast); FCC: follicular lymphoma; MF: mycosis fungoides; COLON: colon; LVR: liver; BLDR: bladder; OVARY: ovary; Lung: lung; MELA: melanoma; BRAIN: brain; TALL: T-cell ALL; BALL: B-cell ALL; LBL: diffused large-B cell lymphoma; AML: acute myelogenous leukaemia.

**Supplementary Figure 5** Comparison of miRNA expression levels of poorly differentiated and more-differentiated tumors. Poorly differentiated tumors (PD) with primary origins from colon, ovary, lung, breast (BRST) or lymphnode (LBL) were compared to more-differentiated tumors (non-PD) of the corresponding tissue types in the miGCM collection. After filtering out non-detectible miRNAs, the remaining 173 features were centered and normalized for each tissue type separately to a mean of 0 and a standard deviation of 1. A heatmap of the data is shown. Samples with the same tissue type and PD status were sorted according to total miRNA expression readings, with higher expressing samples on the left. Features were sorted according to the variance-thresholded t-test score.

**Supplementary Figure 6** Hierarchical clustering analyses of miRNA data and mRNA data. For 89 epithelial samples that had successful expression data of both miRNAs and mRNAs, hierarchical clustering was performed using average linkage and correlation similarity, after gene filtering. Filtering of miRNA data eliminates genes that do not have expression values above a miminum threshold in any sample (see Supplementary Methods for details). Three different filtering methods were used for mRNA data. The first method (mRNA filt-1) uses the same criteria as used for miRNA data, resulting in 14546 genes. The second method (mRNA filt-2) employed a variation filter as described [1], and resulted in 6621 genes. The third method (mRNA filt-3) focused on transcription factors that passed the above variation filter, ending with 220 genes. Samples of gastrointestinal tract (GI) or non-GI origins are indicated. Tissue type (TT) and malignancy status (MAL) for normal (N) or tumor (T) samples are also indicated. Note that the GI-derived samples largely cluster together in the space of miRNA expression, but not by mRNA expression. Abbreviations: PAN: pancreas; KID: kidney; PROST: prostate; UT: uterus; MESO: mesothelioma; BRST: breast; COLON: colon; BLDR: bladder; OVARY: ovary; Lung: lung; MELA: melanoma.

**Supplementary Figure 7** *In vitro* erythroid differentiation. Purified CD34$^+$ cells from human umbilical cord blood were induced to differentiate along the erythroid lineage. **(a)** Total cell counts were determined every two days. Data are averages of cell counts from a triplicate experiment and error bars represent standard deviations. **(b)** Markers of erythroid differentiation, CD71 and Glycophorin A (GlyA), were determined using flow cytometry. Percentages of cells with negative (-), low, or positive (+) marker staining are plotted. **(c)** miRNA expression profiles of differentiating erythrocytes were determined on days (d) indicated after induction. Data were log$_2$-transformed, averaged among successfully profiled same-day samples and normalized to a mean of 0 and a standard deviation of 1 for each miRNA. Data were then filtered to eliminate miRNAs that do not have expression values higher than a minimum cut-off (7.25 on $\log_2$ scale) in any sample. A heatmap of miRNA expression is shown, with red color indicating higher expression and blue for lower expression. Data shown are from a representative differentiation experiment of two performed.

**Supplementary Figure 8** Comparing miRNA expression levels with an mRNA signature of proliferation. A consensus set of mRNA transcripts that positively correlate with proliferation rate was assembled based on published data (see Supplementary Data). Data for miRNA and mRNA expression in lung and breast (BRST) were centered and normalized for each gene, bringing the mean to 0 and the standard deviation to 1. The mean expression of mRNAs correlated with proliferation (on the horizontal axis) was plotted against the mean expression of miRNA markers for tumor/normal distinction (on the vertical axis). Normal samples, poorly differentiated (diff.) tumors and more differentiated tumors are represented by round, triangle and square dots, respectively. Note that the mRNA proliferation signature distinguishes normal samples from tumors, reflecting faster proliferation rates in cancer specimens; however, it does not distinguish between poorly differentiated tumors and more differentiated tumors, even though the miRNA expression levels in the latter two categories are different.

## *Supplementary Methods*

### Cell culture

HEL, TF-1, PC-3, MCF-7, HL-60, SKMEL-5, 293 and K562 cells were obtained from the American Type Culture Collection (ATCC, Manassas, VA), and cultured according to ATCC instructions. All T-cell ALL cell lines were cultured in RPMI medium supplemented with 10% fetal bovine serum. CCRF-CEM and LOUCY cells were obtained from ATCC. ALL-SIL, HPB-ALL, PEER, TALL1, P12-ICHIKAWA cells were obtained from the German Collection of Microorganisms and Cell Cultures (DSMZ, Braunschweig, Germany). SUPT11 cells were a kind gift of Dr. Michael Cleary at Stanford University.

Umbilical cord blood was obtained under an IRB approved protocol from the Brigham and Women's Hospital. Light-density mononuclear cells were separated by Ficoll-Hypaque centrifugation, and CD34$^+$ cells (85-90% purity) were enriched using Midi-MACS columns (Miltenyi Biotec, Auburn, CA). Erythroid differentiation of the CD34$^+$ cells was induced in two stages in liquid culture [2]. For the first seven days, cells were cultured in Serum Free Expansion Medium (SFEM, Stem Cell Technologies, Tukwila, WA) supplemented with penicillin/streptomycin, glutamine, 100 ng/mL stem cell factor (SCF), 10 ng/mL interleukin-3 (IL-3), 1 μM dexamethasone (Sigma), 40 μg/ml lipids (Sigma), and 3 IU/ml erythropoietin (Epo). After 7 days, cells were cultured in the same medium without dexamethasone and supplemented with 10 IU/ml Epo. For flow cytometry analyses, approximately 1 to 5 x $10^5$ cells were labeled with a phycoerythrin-conjugated antibody against glycophorin-A (CD235a, Clone GA-R2, BD-Pharmingen, San Jose, CA) and a FITC-conjugated antibody against CD71 (Clone M-A712, BD-Pharmingen). Flow cytometry analyses were performed using a FACScan flow cytometer (Becton Dickinson).

## Glass-slide detection of miRNAs

Glass slide microarrays were spotted oligonucleotide arrays and hybridized as described previously [3]. Briefly, 5'-amino-modified oligonucleotide probes (the same ones as used on the bead platform) were printed onto amide-binding slides (CodeLink, Amersham Biosciences). Printing and hybridization were done following the slides manufacturer's protocols with the following modifications: oligonucleotide concentration for printing was 20 µM in 150 mM sodium phosphate, pH 8.5. Printing was done on a MicroGrid TAS II arrayer (BioRobotics) at 50% humidity. Labeled PCR product was resuspended in hybridization buffer (5X SSC, 0.1% SDS, 0.1 mg/ml salmon sperm DNA) and hybridized at 50°C for 10 hours. Microarray slides were scanned using an arrayWoRx$^e$ biochip reader (Applied Precision) and primary data were analyzed using the Digital Genome System suite (Molecularware).

## Northern blot analysis

Northern blot analyses were carried out as described [4]. Total RNAs from cell lines were loaded at 10 µg per lane. Blots were detected with DNA probes complementary for human miR-20, miR-181a, miR-15a, miR-16, miR-17-5p, miR-221, let-7a, and miR-21.

## Quantitative RT-PCR

Reverse transcription (RT) reactions were carried out on 50 to 200 ng total RNA in 10 μl reaction volumes, using the TaqMan reverse transcription kit (Applied Biosystems, Foster City, CA) and random hexamers, following the manufacturer's protocol. RT products were diluted 5-fold in water and assayed using TaqMan Gene Expression Assays (Applied Biosystems) in triplicates, on an ABI PRISM 7900HT real-time PCR machine. Efficiency of PCR amplification was determined by 5 two-fold-serial-diluted samples from HL-60 cDNA. The TaqMan Gene Expression Assays used are listed in the parentheses. (Dicer1: Hs00998566_m1; Ago2/EIF2C2: Hs00293044_m1; Drosha/RNase3L: Hs00203008_m1; DGCR8: Hs00256062_m1; and eukaryotic 18S rRNA endogenous control)

## Data preprocessing and quality control

To eliminate bead-specific background, the reading of every bead for every sample was first processed by subtracting the average readings of that particular bead in the two-embedded mock-PCR samples in each plate. As stated in the Methods, every sample was assayed in three wells. Each of the three wells contained 94 probes (19 common probes and 75 unique ones). Out of the 19 common probes are the two pre-labeling controls and the two post-labeling controls. Quality control was performed as part of the preprocessing by requiring that the reading from each control probe exceeds some minimal probe-specific threshold. These thresholds were determined by identifying a natural lower cutoff, i.e. a dip, in the distribution of each control probe. The cutoff values were chosen based on a set of samples in a pilot study. The lower post-control should be greater than 500 and the higher post-control must exceed 2450. The lower and higher pre-controls should exceed 1400 and 2000 respectively (after well-to-well scaling). In this study, about 70% of the samples passed the quality control. Note that the above specifications were used on version 1 of the platform. A similar preprocessing was performed on version 2 of the platform.

Preprocessing was done in four steps: (i) well-to-well scaling – the reading from each well were scaled such that the total of the two post-labeling controls, in that well, became 4500 (a median value based on a pilot study); (ii) sample scaling – the normalized readings were scaled such that total of the 6 pre-labeling controls in each sample reached 27,000 (a median value based on a pilot study); (iii) thresholding at 32 (see Supplementary Data); and (iv) $\log_2$ transformation. All control probes, as well as a probe (EAM296) which had a high background in the absence of any prepared target, were removed before any further analysis. After eliminating these probes, 217 (255 for version 2 of the platform) features were left and these were used throughout the analysis.

## Hierarchical clustering

miRNA expression data first underwent filtering. The purpose of this filtering is to remove features which have no detectable expression and thus are uninformative but may introduce noise to the clustering. A miRNA was regarded as “not expressed” or “not detectible”, if in none of the samples, that particular miRNA has an expression value above a minimal cutoff. We applied a cutoff of 7.25 (after data were $\log_2$-transformed). This cutoff value was determined based on noise analyses of target preparation and bead detection (see Supplementary Data Section and Supplementary Fig. 3a). In that experiment, the majority of features had a standard deviation below 0.75 when their mean was over 5 in $\log_2$-transformed data. Thus we used a cutoff of 3 standard deviations above the minimal expression level (5+3x0.75=7.25). Any feature that is not expressed under this criterion was filtered out before clustering. Data were then centered and normalized for each feature, bringing the mean to 0 and the standard deviation to 1. This equalizes the contributions of all features. For hierarchical clustering, we used Pearson correlation as a similarity measure, and used the average-linkage algorithm [5] for both the samples and the features.

## *k*-Nearest Neighbor (*k*NN) prediction

After feature filtration (described in the hierarchical clustering), marker selection was performed on 187 features. The variance-thresholded t-test score was used as a measure to score features. A minimal standard deviation of 0.75 was applied. Markers were searched among the filtered miRNAs. Nominal P-value was calculated for each feature, by permuting the class labels of the samples. In order to select features that best distinguish tumors from normal samples on all tissue types, i.e. taking into account the confounding tissue-type phenotype, restricted permutations were performed [6]. In restricted permutations, one shuffles the tumor/normal labels only within each tissue type to get the distribution under the desired null hypothesis. To achieve accurate estimates for the p-values, 400 times the number of features (400x187=74,800) of iterations were performed. To correct for multiple-hypotheses testing, markers were selected requiring the Bonferroni-corrected P-values to be less than 0.05. *k*NN prediction was performed using the *k*NN module in the GenePattern software, with *k*=3 and a Euclidean distance measure (GenePattern at http://www.broad.mit.edu/cancer/software/genepattern/index.html).

## Probabilistic Neural Network (PNN) prediction

A two-class PNN [7] prediction was calculated based on the following class posterior probability:

$$P(c\,|\,\mathbf{x}) = \frac{P(\mathbf{x}\,|\,c)P(c)}{\sum_{c'} P(\mathbf{x}\,|\,c')P(c')} = \frac{\dfrac{P(c)}{n_c}\sum_{i:\bar{y}_i \in c} \exp\left(-\mathrm{D}(\mathbf{x},\mathbf{y}_i)^2 / 2\sigma^2\right)}{\sum_{c'}\left[\dfrac{P(c')}{n_{c'}}\sum_{i:\bar{y}_i \in c'} \exp\left(-\mathrm{D}(\mathbf{x},\mathbf{y}_i)^2 / 2\sigma^2\right)\right]},$$

where **x** is the predicted sample and *c* is the class for which the posterior probability is calculated. The training set samples are $\mathbf{y}_i$, $n_c$ is the number of samples of class *c* in the training set, and D($\mathbf{x}$,$\mathbf{y}_i$) is the distance between the predicted sample and training sample *i*. In our case, the sum in the denominator (of c') is over two class values, since we predict a sample either to belong or not to belong to a specific tissue-type. Note that the first step is derived using Bayes rule which allows to incorporate a prior probability for each class, $P(c)$. We used a uniform prior over all 11 tissue-types which translated to 1/11 for being in a certain type and 10/11 for not being in that type. We did not use the tissue-type frequencies in the training set since they likely do not represent the frequencies of different tumors in the general population.

Multi-class prediction using PNN was achieved by breaking down the question into multiple one vs. the rest (OVR) predictions. To perform PNN OVR two-class classification, we built a model based on the training set. This model has two parameters: the number of features used, and σ (the standard deviation of the Gaussian kernel which is used to calculate the contribution of each training sample to the classification). The optimal parameters (for each OVR classifier) were selected using a leave-one-out cross-validation procedure from all possible parameter-pairs in which the number of features ranges from 2 to 30 in steps of 2 and σ takes the values from 1 to 4 times the median nearest neighbor distance, in steps of 0.5 (a total number of 105 combinations). The best model was determined by (i) the fewest number of leave-one-out errors on the training set, which include both false-positive and false-negative errors with the same weight, and (ii) among all conditions with the same error rate, the parameters that gave rise to the maximal mean log-likelihood of the training set were selected. The mean log-likelihood is defined as $L[\{\mathbf{x}_i\};M] = \frac{1}{\#\text{of training examples}} \sum_i \log(P_M(\mathrm{c}_\mathrm{i} \mid \mathbf{x}_i))$ where $c_\mathrm{i}$ is the true class of sample $\mathbf{x}_\mathrm{i}$ and the probability is evaluated using the model *M*. The top n features were selected using the variance-thresholded t-test score in a balanced manner; n/2 features with the top positive scores and n/2 features with most negative scores. The cosine distance measure was used; $\mathrm{D}(\mathbf{x},\mathbf{y}_i)=1-\mathrm{cosine}(\mathbf{x},\mathbf{y}_i)$.

## P-value calculation for the number of correct classifications

A Binomial distribution was used to calculate the probability to obtain at least the number of correct classifications (on the test set) as we observed. Assuming a random classifier would predict the tissue-type randomly with a uniform distribution over the 11 possible outcomes, the probability of a correct classification is 1/11. This is applicable to the PNN prediction, in which the background frequency of each tissue type was assumed to be 1/11. The p-value is, therefore, the tail of the Binomial distribution from the observed number of correct classifications, *s*, to the total number of samples in the test set, *n*:

$\text{P - value} = \sum_{t=s}^{n} \binom{n}{t} p^{t} (1-p)^{n-t}$ where p is one over the number of tissue-types (1/11, in our case) and *t* is the number of correct classification which goes from the observed number, *s*, to the maximum of possible correct samples *n*.

## *Supplementary Data*

### Development of a bead-based miRNA profiling platform

Compared with glass-based microarrays, bead-based profiling solutions have the advantages of higher sample throughput and liquid phase hybridization kinetics, while having the disadvantage of lower feature throughput. For the genomic analysis of miRNA expression, this disadvantage is negligible because of the relative small number of identified miRNAs. Since new miRNAs are still being discovered, the flexibility and ease of these "liquid chips" to introduce new features is of particular value.

We developed a bead-based miRNA profiling platform, as detailed in the Methods section. Version 1 of this platform (used for most samples in this study) covers 164 human, 185 mouse, and 174 rat miRNAs, according to Rfam 5.0 miRNA registry database [8, 9] (http://www.sanger.ac.uk/Software/Rfam/mirna/index.shtml). Version 2 of this platform (used for the acute lymphoblastic leukemia study and the erythroid differentiation study) covers additional 24 human, 13 mouse and 2 rat miRNAs (refer to Supplementary Table 1 for details).

This profling platform is compatible in theory with any miRNA labeling method that labels the sense strand. For our study, we followed one described by Miska *et al.* [3] that labels mature miRNAs through adaptor ligation, reverse-transcription and PCR amplification. We reasoned that the amplification step will allow future use of these labeled materials, which were from precious clinical samples. Defined amounts of synthetic artificial miRNAs were added into each sample of total RNAs as pre-labeling controls. This allows us to normalize the profiling data according to the starting amount of total RNA, using readings from capture probes for these synthetic miRNAs (see Methods for details). This contrasts the use of total feature intensity to normalize the readings of different samples; the hidden assumption of the latter is that the total miRNA expression is the same in all samples, which may not be true considering the small known number of miRNAs.

We analyzed the variation caused by labeling and detection using repetitive assays of the same RNA samples of a few cell lines originated from different tissues; these cell lines have different miRNA profiles. We plotted the standard deviation of each probe versus its means, after the data were $\log_2$-transformed (Supplementary Fig. 3a). The variations are large for low means, and decrease and stabilize with increasing means. For most measured features with mean above 5 (32 before $\log_2$-transformation), the standard deviation is below 0.75. This value of mean provides a good cutoff for a lower threshold of the data, which was thus used in this study.

We compared the data from expression profiles and northern blots on a panel of 7 cell lines; the same quantities of the same starting total RNAs were used for both analyses. We picked eight miRNAs that are expressed in any of these cell lines and that show differential expression according to the expression profiles, and probed them with northern blots. All eight display good concordance between the two assays (Fig. 1c), indicating that our profiling platform has good accuracy.

We next examined the linearity of profiling (both labeling and detection) by measuring a series of starting materials, covering 0.5 µg to 10 µg of total RNAs from HEL cells. Most miRNAs report good linearity up to 3500 median fluorescence intensity readings (after normalization with pre-labeling-controls, Supplementary Fig. 3b). Taken together with the threshold level of 32, the profiling method has roughly 100-fold of dynamic range.

One common issue that affects hybridization-based analyses for miRNAs is the specificity of detection, since many miRNAs are closely-related on the sequence level. To assess the specificity of detection, we synthesized oligonucleotides corresponding to the reverse-transcription products of adaptor-ligated miRNAs, in this case the human *let-7* family of miRNAs and a few artificial mutants. The sequences for these oligonucleotides, as well as the alignment of human *let-7* miRNAs and mutant sequences, are listed below. They were then labeled through PCR using the same primer sets. This provides a collection of sequence-pairs that differ by one, two, or a few nucleotides (Supplementary Fig. 2 and the alignment below). Results are presented in the main text and in Fig. 1a,b.

**Alignment of Human *let-7* miRNAs and Mutant Sequences**

| | |
|---|---|
| UGAGGUAGUAGUUUGUACAGU | hsa-let-7g |
| UGAGGUAGUACUUUCUACAGUUA | let-7-mut1 |
| UGAGGUAGUAGGUUGUAUGGUU | hsa-let-7c |
| UGAGGUACUAGCUUGUAUGGUU | let-7-mut2 |
| UGAGGUAGUAGGUUGUGUGGUU | hsa-let-7b |
| UGAGGUACUAGCUUGUGUGGUU | let-7-mut3 |
| UGAGGUAGUAGGUUGUAUAGUU | hsa-let-7a |
| UGAGGUAGGAGGUUGUAUAGU | hsa-let-7e |
| AGAGGUAGUAGGUUGCAUAGU | hsa-let-7d |
| UGAGGUAGUAGAUUGUAUAGUU | hsa-let-7f |
| UGAGGUAGUAGUUUGUGCU | hsa-let-7i |

**Table: Oligonucleotide Sequences for Detection Specificity Experiment**

| miRNA or Mutant Name | Oligonucleotide Sequence (5’ to 3’) |
|---|---|
| hsa-let-7g | CTGGAATTCGCGGTTAAAACTGTACAAACTACTACCTCATTTAGTGAGGAATTCCGT |

| | |
|---|---|
| let-7-mut1 | CTGGAATTCGCGGTTAAATAACTGTAGAAAGTACTACCTCATTTAGTGAGGAATTCCGT |
| hsa-let-7c | CTGGAATTCGCGGTTAAAAACCATACAACCTACTACCTCATTTAGTGAGGAATTCCGT |
| let-7-mut2 | CTGGAATTCGCGGTTAAAAACCATACAAGCTAGTACCTCATTTAGTGAGGAATTCCGT |
| hsa-let-7b | CTGGAATTCGCGGTTAAAAACCACACAACCTACTACCTCATTTAGTGAGGAATTCCGT |
| let-7-mut3 | CTGGAATTCGCGGTTAAAAACCACACAAGCTAGTACCTCATTTAGTGAGGAATTCCGT |
| hsa-let-7a | CTGGAATTCGCGGTTAAAAACTATACAACCTACTACCTCATTTAGTGAGGAATTCCGT |
| hsa-let-7e | CTGGAATTCGCGGTTAAAACTATACAACCTCCTACCTCATTTAGTGAGGAATTCCGT |
| hsa-let-7d | CTGGAATTCGCGGTTAAAACTATGCAACCTACTACCTCTTTTAGTGAGGAATTCCGT |
| hsa-let-7f | CTGGAATTCGCGGTTAAAAACTATACAATCTACTACCTCATTTAGTGAGGAATTCCGT |
| hsa-let-7i | CTGGAATTCGCGGTTAAAAGCACAAACTACTACCTCATTTAGTGAGGAATTCCGT |

## Hierarchical clustering of multiple cancer and normal samples

We applied this miRNA profiling platform for 140 human cancer specimens, 46 normal human tissues, and various cell lines. The collection of samples covers more than ten tissues and cancer types. This collection was referred to as miGCM (for miRNA Global Cancer Map). We first examined the miRNA expression profiles to see whether we can detect previously reported tissue-restricted expression of miRNAs. Indeed, we observed tissue-restricted expression patterns. For example, miR-122a, a reported liver-specific miRNA [10], is exclusively expressed in the liver samples, whereas miR-124a, a brain-specific miRNA [10], is abundantly expressed in the brain samples.

We performed hierarchical clustering on this data set, as described in the Methods. Hierarchical clustering is an unsupervised analysis tool that captures internal relationship between the samples. It organizes the samples (or features) into a tree structure (a dendrogram) according to the similarity between the samples (or the features). Close pairs of samples (ones with similar expression profiles) will generally be connected in the dendrogram at an earlier phase, while samples with larger distances (with less similar expression profiles) will be connected at a later phase (details can be found in reference [11]). The detailed result of hierarchical clustering on both the samples and features using correlation metrics is presented in Fig. 2a and Supplementary Fig. 4.

## Comparison of miRNA and mRNA clustering in regard to GI samples

After finding that the gastrointestinal tract samples were clustered together (main text and Fig. 2a), we asked whether or not this structure is similarly displayed by clustering in the mRNA space. We took 89 epithelial samples that have both successful mRNA and miRNA profiling data, and subjected them to hierarchical clustering. Both data underwent identical gene filtering, i.e. a lower threshold filter to eliminate genes that do not have expression values over 7.25 (on $\log_2$ scale) in any sample, and underwent the same clustering procedure. This gene filtering resulted in 195 miRNAs and 14546 mRNAs. Data were presented in the main text, Fig. 2c and Supplementary Fig. 6. Results show that the mRNA clustering does not recover the coherence of GI samples, as identified in the miRNA expression space. Of note, the exact outcome of hierarchical clustering is dependent on the collection of samples present for analysis. Consequently, the cluster of the GI samples in miRNA clustering in Fig. 2c is slightly different from that of Fig. 2a, since the latter comprises of many more samples.

In order to test whether the lack of coherence of GI samples in the mRNA clustering is sensitive to the choice of genes that were used to represent each sample, we tested two additional gene filtering methods. First, we used a variation filter as was performed in Ramaswamy et al. [1] (lower threshold of 20, upper threshold of 16000, the maximum value is at least 5 fold greater than the minimum value, and the maximum value is more than 500 greater than the minimum value), which yielded 6621 genes. Second, we examined only transcription factors, a set of gene regulators as are miRNAs. We took the genes that passed the above variation filter and that are also annotated with transcription factor activity in the Gene Ontology (www.geneontology.org, GO:0003700). This resulted in 220 transcription factors as listed in the table below. Similar to the minimum-expression filter on the mRNA data, these two gene selection methods yielded clustering by tissue types to a certain degree. However, none recovered the gut coherence (Supplementary Fig. 6). This indicated either that the miRNA space contains some different information from the mRNA space or that in the mRNA space, the gut signal is masked by other signals or noise. Importantly, a set of transcription factors did not mimic miRNAs in this test, suggesting the difference is not solely due to the gene regulator nature of miRNAs.

**Table: 220 mRNA genes with transcription factor activity annotation**

| Chip | Probe Set ID | Gene Title |
|---|---|---|
| Hu6800 | AB000468_at | ring finger protein 4 |
| Hu6800 | D43642_at | transcription factor-like 1 |
| Hu6800 | D83784_at | pleiomorphic adenoma gene-like 2 |
| Hu6800 | D86479_at | AE binding protein 1 |
| Hu6800 | D87673_at | heat shock transcription factor 4 |
| Hu6800 | J03161_at | serum response factor (c-fos serum response element-binding transcription factor) |
| Hu6800 | J03827_at | nuclease sensitive element binding protein 1 |
| Hu6800 | L02785_at | solute carrier family 26, member 3 |
| Hu6800 | L11672_at | zinc finger protein 91 (HPF7, HTF10) |
| Hu6800 | L11672_r_at | zinc finger protein 91 (HPF7, HTF10) |
| Hu6800 | L13203_at | forkhead box I1 |
| Hu6800 | L13740_at | nuclear receptor subfamily 4, group A, member 1 |
| Hu6800 | L17131_rna1_at | high mobility group AT-hook 1 |
| Hu6800 | L20298_at | core-binding factor, beta subunit |
| Hu6800 | L22342_at | SP110 nuclear body protein |
| Hu6800 | L22454_at | nuclear respiratory factor 1 |
| Hu6800 | L40904_at | peroxisome proliferative activated receptor, gamma |

| | | |
|---|---|---|
| Hu6800 | M14328_s_at | enolase 1, (alpha) |
| Hu6800 | M16938_s_at | homeo box C6 |
| Hu6800 | M19720_rna1_at | v-myc myelocytomatosis viral oncogene homolog 1, lung carcinoma derived (avian) |
| Hu6800 | M23263_at | androgen receptor (dihydrotestosterone receptor; testicular feminization; spinal and bulbar muscular atrophy; Kennedy disease) |
| Hu6800 | M24900_at | thyroid hormone receptor, alpha (erythroblastic leukemia viral (v-erb-a) oncogene homolog, avian) /// nuclear receptor subfamily 1, group D, member 1 |
| Hu6800 | M25269_at | ELK1, member of ETS oncogene family |
| Hu6800 | M31627_at | X-box binding protein 1 |
| Hu6800 | M36542_s_at | POU domain, class 2, transcription factor 2 |
| Hu6800 | M38258_at | retinoic acid receptor, gamma |
| Hu6800 | M64673_at | heat shock transcription factor 1 |
| Hu6800 | M65214_s_at | transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47) |
| Hu6800 | M68891_at | GATA binding protein 2 |
| Hu6800 | M76732_s_at | msh homeo box homolog 1 (Drosophila) |
| Hu6800 | M77698_at | YY1 transcription factor |
| Hu6800 | M79462_at | promyelocytic leukemia |
| Hu6800 | M79463_s_at | promyelocytic leukemia |
| Hu6800 | M93650_at | paired box gene 6 (aniridia, keratitis) |
| Hu6800 | M95929_at | sideroflexin 3 |
| Hu6800 | M97676_at | msh homeo box homolog 1 (Drosophila) |
| Hu6800 | M97935_s_at | signal transducer and activator of transcription 1, 91kDa |
| Hu6800 | M97936_at | signal transducer and activator of transcription 1, 91kDa |
| Hu6800 | M99701_at | transcription elongation factor A (SII)-like 1 |
| Hu6800 | S81264_s_at | T-box 2 |
| Hu6800 | U00968_at | sterol regulatory element binding transcription factor 1 |
| Hu6800 | U11861_at | maternal G10 transcript |
| Hu6800 | U18018_at | ets variant gene 4 (E1A enhancer binding protein, E1AF) |
| Hu6800 | U20734_s_at | jun B proto-oncogene |
| Hu6800 | U28687_at | zinc finger protein 157 (HZF22) |
| Hu6800 | U29175_at | SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4 |
| Hu6800 | U35048_at | transforming growth factor beta 1 induced transcript 4 |
| Hu6800 | U36922_at | forkhead box O1A (rhabdomyosarcoma) |
| Hu6800 | U39840_at | forkhead box A1 |
| Hu6800 | U44755_at | small nuclear RNA activating complex, polypeptide 2, 45kDa |
| Hu6800 | U51003_s_at | distal-less homeo box 2 |
| Hu6800 | U51127_at | interferon regulatory factor 5 |
| Hu6800 | U53830_at | interferon regulatory factor 7 |
| Hu6800 | U58681_at | neurogenic differentiation 2 |
| Hu6800 | U63842_at | neurogenin 1 |
| Hu6800 | U69126_s_at | KH-type splicing regulatory protein (FUSE binding protein 2) |
| Hu6800 | U72649_at | BTG family, member 2 |
| Hu6800 | U73843_at | E74-like factor 3 (ets domain transcription factor, epithelial- |

| | | |
|---|---|---|
| | | specific ) |
| Hu6800 | U76388_at | nuclear receptor subfamily 5, group A, member 1 |
| Hu6800 | U81599_at | homeo box B13 |
| Hu6800 | U81600_at | paired related homeobox 2 |
| Hu6800 | U82759_at | homeo box A9 |
| Hu6800 | U85193_at | nuclear factor I/B |
| Hu6800 | U85658_at | transcription factor AP-2 gamma (activating enhancer binding protein 2 gamma) |
| Hu6800 | U95040_at | tripartite motif-containing 28 |
| Hu6800 | X03635_at | estrogen receptor 1 |
| Hu6800 | X06614_at | retinoic acid receptor, alpha |
| Hu6800 | X12794_at | nuclear receptor subfamily 2, group F, member 6 |
| Hu6800 | X13293_at | v-myb myeloblastosis viral oncogene homolog (avian)-like 2 |
| Hu6800 | X13810_s_at | POU domain, class 2, transcription factor 2 |
| Hu6800 | X16316_at | vav 1 oncogene |
| Hu6800 | X16665_at | homeo box B2 |
| Hu6800 | X16706_at | FOS-like antigen 2 |
| Hu6800 | X17360_rna1_at | homeo box D4 |
| Hu6800 | X17651_at | myogenin (myogenic factor 4) |
| Hu6800 | X51345_at | jun B proto-oncogene |
| Hu6800 | X52541_at | early growth response 1 |
| Hu6800 | X55005_rna1_at | thyroid hormone receptor, alpha (erythroblastic leukemia viral (v-erb-a) oncogene homolog, avian) |
| Hu6800 | X55037_s_at | GATA binding protein 3 |
| Hu6800 | X56681_s_at | jun D proto-oncogene |
| Hu6800 | X58072_at | GATA binding protein 3 |
| Hu6800 | X60003_s_at | cAMP responsive element binding protein 1 |
| Hu6800 | X61755_rna1_s_at | homeo box C5 |
| Hu6800 | X65463_at | retinoid X receptor, beta |
| Hu6800 | X66079_at | Spi-B transcription factor (Spi-1/PU.1 related) |
| Hu6800 | X68688_rna1_s_at | zinc finger protein 11b (KOX 2) /// zinc finger protein 33a (KOX 31) |
| Hu6800 | X69699_at | paired box gene 8 |
| Hu6800 | X70683_at | SRY (sex determining region Y)-box 4 |
| Hu6800 | X72632_s_at | thyroid hormone receptor, alpha (erythroblastic leukemia viral (v-erb-a) oncogene homolog, avian) /// nuclear receptor subfamily 1, group D, member 1 |
| Hu6800 | X78992_at | zinc finger protein 36, C3H type-like 2 |
| Hu6800 | X85786_at | regulatory factor X, 5 (influences HLA class II expression) |
| Hu6800 | X90824_s_at | upstream transcription factor 2, c-fos interacting |
| Hu6800 | X93996_rna1_at | myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 7 |
| Hu6800 | X96401_at | MAX binding protein |
| Hu6800 | X96506_s_at | DR1-associated protein 1 (negative cofactor 2 alpha) |
| Hu6800 | X99101_at | estrogen receptor 2 (ER beta) |
| Hu6800 | Y08976_at | FEV (ETS oncogene family) |
| Hu6800 | Z11899_s_at | POU domain, class 5, transcription factor 1 |
| Hu6800 | Z17240_at | high-mobility group box 2 |
| Hu6800 | Z22951_rna1_s_at | --- |
| Hu6800 | Z49825_s_at | hepatocyte nuclear factor 4, alpha |

| | | |
|---|---|---|
| Hu6800 | Z50781_at | delta sleep inducing peptide, immunoreactor |
| Hu6800 | Z56281_at | interferon regulatory factor 3 |
| Hu35KsubA | AA010750_at | LAG1 longevity assurance homolog 2 (S. cerevisiae) |
| Hu35KsubA | AA036900_at | FOS-like antigen 2 |
| Hu35KsubA | AA091017_at | nuclear factor of activated T-cells 5, tonicity-responsive |
| Hu35KsubA | AA099501_at | p66 alpha |
| Hu35KsubA | AA127183_s_at | serologically defined colon cancer antigen 33 |
| Hu35KsubA | AA157520_at | signal transducer and activator of transcription 5B |
| Hu35KsubA | AA287840_at | Runt-related transcription factor 2 |
| Hu35KsubA | AA328684_at | SLC2A4 regulator |
| Hu35KsubA | AA347664_at | lymphoid enhancer-binding factor 1 |
| Hu35KsubA | AA355201_at | SRY (sex determining region Y)-box 4 |
| Hu35KsubA | AA418098_at | cAMP responsive element binding protein-like 2 |
| Hu35KsubA | AA424381_s_at | Forkhead box C1 |
| Hu35KsubA | AA431268_at | --- |
| Hu35KsubA | AA436315_at | forkhead box O3A |
| Hu35KsubA | AA456687_at | nuclear factor I/A |
| Hu35KsubA | AA459542_s_at | regulatory factor X-associated ankyrin-containing protein |
| Hu35KsubA | AA489299_at | transcriptional adaptor 3 (NGG1 homolog, yeast)-like |
| Hu35KsubA | AA504413_at | Solute carrier family 25, member 29 |
| Hu35KsubA | AB002302_at | myeloid/lymphoid or mixed-lineage leukemia 4 |
| Hu35KsubA | AB002305_at | aryl-hydrocarbon receptor nuclear translocator 2 |
| Hu35KsubA | AB004066_at | basic helix-loop-helix domain containing, class B, 2 |
| Hu35KsubA | C02099_s_at | methionine sulfoxide reductase B2 |
| Hu35KsubA | D45333_at | prefoldin 1 |
| Hu35KsubA | D61676_at | Pre-B-cell leukemia transcription factor 1 |
| Hu35KsubA | D82636_at | CCR4-NOT transcription complex, subunit 7 |
| Hu35KsubA | H45647_at | hairy/enhancer-of-split related with YRPW motif 1 |
| Hu35KsubA | IKAROS_at | zinc finger protein, subfamily 1A, 1 (Ikaros) |
| Hu35KsubA | L07592_at | peroxisome proliferative activated receptor, delta |
| Hu35KsubA | L13203_at | forkhead box I1 |
| Hu35KsubA | L16794_s_at | MADS box transcription enhancer factor 2, polypeptide D (myocyte enhancer factor 2D) |
| Hu35KsubA | L40904_at | peroxisome proliferative activated receptor, gamma |
| Hu35KsubA | L41067_at | nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 3 |
| Hu35KsubA | M23263_at | androgen receptor (dihydrotestosterone receptor; testicular feminization; spinal and bulbar muscular atrophy; Kennedy disease) |
| Hu35KsubA | M62626_s_at | T-cell leukemia, homeobox 1 |
| Hu35KsubA | M79462_at | promyelocytic leukemia |
| Hu35KsubA | M92299_s_at | homeo box B5 |
| Hu35KsubA | M93650_at | paired box gene 6 (aniridia, keratitis) |
| Hu35KsubA | M96577_s_at | E2F transcription factor 1 |
| Hu35KsubA | M97676_at | msh homeo box homolog 1 (Drosophila) |
| Hu35KsubA | N32724_at | high-mobility group 20B |
| Hu35KsubA | N83192_at | KIAA0669 gene product |
| Hu35KsubA | RC_AA029288_at | zinc finger protein 83 (HPF1) |
| Hu35KsubA | RC_AA040699_at | ELK3, ETS-domain protein (SRF accessory protein 2) |

| | | |
|---|---|---|
| Hu35KsubA | RC_AA045545_at | glucocorticoid modulatory element binding protein 2 |
| Hu35KsubA | RC_AA055932_at | TAF5-like RNA polymerase II, p300/CBP-associated factor (PCAF)-associated factor, 65kDa |
| Hu35KsubA | RC_AA065094_at | trinucleotide repeat containing 4 |
| Hu35KsubA | RC_AA069549_at | zinc finger protein 37a (KOX 21) |
| Hu35KsubA | RC_AA114866_s_at | homeo box A11 |
| Hu35KsubA | RC_AA121121_at | Huntingtin interacting protein 2 |
| Hu35KsubA | RC_AA135095_at | high-mobility group 20B |
| Hu35KsubA | RC_AA136474_at | Meis1, myeloid ecotropic viral integration site 1 homolog 2 (mouse) |
| Hu35KsubA | RC_AA150205_at | Kruppel-like factor 7 (ubiquitous) |
| Hu35KsubA | RC_AA156112_at | Krueppel-related zinc finger protein |
| Hu35KsubA | RC_AA156359_at | TAR DNA binding protein |
| Hu35KsubA | RC_AA156792_at | hairy/enhancer-of-split related with YRPW motif-like |
| Hu35KsubA | RC_AA235980_at | transcription factor EB |
| Hu35KsubA | RC_AA252161_at | p66 alpha |
| Hu35KsubA | RC_AA253429_at | zinc finger protein 175 |
| Hu35KsubA | RC_AA256678_at | CCR4-NOT transcription complex, subunit 7 |
| Hu35KsubA | RC_AA256680_at | Nuclear factor I/B |
| Hu35KsubA | RC_AA280130_at | checkpoint suppressor 1 |
| Hu35KsubA | RC_AA284143_at | arginine-glutamic acid dipeptide (RE) repeats |
| Hu35KsubA | RC_AA286809_at | upstream binding protein 1 (LBP-1a) |
| Hu35KsubA | RC_AA292717_at | forkhead box P1 |
| Hu35KsubA | RC_AA347288_at | growth arrest-specific 7 |
| Hu35KsubA | RC_AA379087_s_at | apoptosis antagonizing transcription factor |
| Hu35KsubA | RC_AA393876_s_at | nuclear receptor subfamily 2, group F, member 2 |
| Hu35KsubA | RC_AA419547_at | E74-like factor 5 (ets domain transcription factor) |
| Hu35KsubA | RC_AA421050_at | zinc finger protein 444 |
| Hu35KsubA | RC_AA425309_at | Nuclear factor I/B |
| Hu35KsubA | RC_AA428024_at | ubinuclein 1 |
| Hu35KsubA | RC_AA430032_at | pituitary tumor-transforming 1 |
| Hu35KsubA | RC_AA431399_at | arginine-glutamic acid dipeptide (RE) repeats |
| Hu35KsubA | RC_AA436608_at | SATB family member 2 |
| Hu35KsubA | RC_AA443090_s_at | interferon regulatory factor 7 |
| Hu35KsubA | RC_AA443962_at | MYST histone acetyltransferase 2 |
| Hu35KsubA | RC_AA452256_at | zinc finger protein 265 |
| Hu35KsubA | RC_AA456289_at | nuclear factor I/A |
| Hu35KsubA | RC_AA456677_at | zinc finger protein, subfamily 1A, 4 (Eos) |
| Hu35KsubA | RC_AA464251_at | LOC440448 |
| Hu35KsubA | RC_AA476720_at | nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 1 |
| Hu35KsubA | RC_AA478590_at | forkhead box O3A |
| Hu35KsubA | RC_AA478596_at | zinc fingers and homeoboxes 2 |
| Hu35KsubA | RC_AA504110_at | v-ets erythroblastosis virus E26 oncogene homolog 1 (avian) |
| Hu35KsubA | RC_AA504144_at | CAMP responsive element binding protein 1 |
| Hu35KsubA | RC_AA504147_s_at | Solute carrier family 25, member 29 |
| Hu35KsubA | RC_AA609017_s_at | forkhead box O1A (rhabdomyosarcoma) |
| Hu35KsubA | RC_AA621179_at | forkhead box J2 |

| | | |
|---|---|---|
| Hu35KsubA | RC_AA621680_at | Kruppel-like factor 4 (gut) |
| Hu35KsubA | RC_D59299_i_at | myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 10 |
| Hu35KsubA | U09366_at | zinc finger protein 133 (clone pHZ-13) |
| Hu35KsubA | U17163_at | ets variant gene 1 |
| Hu35KsubA | U28687_at | zinc finger protein 157 (HZF22) |
| Hu35KsubA | U33749_s_at | thyroid transcription factor 1 |
| Hu35KsubA | U53831_s_at | interferon regulatory factor 7 |
| Hu35KsubA | U62392_at | zinc finger protein 193 |
| Hu35KsubA | U63824_at | TEA domain family member 4 |
| Hu35KsubA | U76388_at | nuclear receptor subfamily 5, group A, member 1 |
| Hu35KsubA | U81600_at | paired related homeobox 2 |
| Hu35KsubA | U85707_at | Meis1, myeloid ecotropic viral integration site 1 homolog (mouse) |
| Hu35KsubA | U88047_at | AT rich interactive domain 3A (BRIGHT- like) |
| Hu35KsubA | U89995_at | forkhead box E1 (thyroid transcription factor 2) |
| Hu35KsubA | W20276_f_at | CG9886-like |
| Hu35KsubA | W26259_at | forkhead box O3A |
| Hu35KsubA | W55861_at | Myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila) |
| Hu35KsubA | W67850_s_at | TGFB-induced factor 2 (TALE family homeobox) |
| Hu35KsubA | X13403_s_at | POU domain, class 2, transcription factor 1 |
| Hu35KsubA | X16666_s_at | homeo box B1 |
| Hu35KsubA | X52402_s_at | homeo box C5 |
| Hu35KsubA | X52560_s_at | CCAAT/enhancer binding protein (C/EBP), beta |
| Hu35KsubA | X58431_rna2_s_at | homeo box B6 |
| Hu35KsubA | X68688_rna1_s_at | zinc finger protein 11b (KOX 2) /// zinc finger protein 33a (KOX 31) |
| Hu35KsubA | X70683_at | SRY (sex determining region Y)-box 4 |
| Hu35KsubA | X99101_at | estrogen receptor 2 (ER beta) |
| Hu35KsubA | X99350_rna1_at | forkhead box J1 |
| Hu35KsubA | Y10746_at | methyl-CpG binding domain protein 1 |
| Hu35KsubA | Z14077_s_at | YY1 transcription factor |

## Normal/tumor classifier and *k*NN prediction of mouse lung samples

In order to build a classifier of normal samples vs. tumor samples based on the miGCM collection, we first picked tissues that have enough normal and tumor samples (at least 3 in each class). The following list summarizes the tissues for this analysis.

**Table: Number of Training Samples Used to Build the Normal/Tumor Classifier**

| Tissue | Number of Normal | Number of Tumor |
|---|---|---|
| Colon | 5 | 10 |

| | | |
|---|---|---|
| Kidney | 3 | 5 |
| Prostate | 8 | 6 |
| Uterus | 9 | 10 |
| Lung | 4 | 6 |
| Breast | 3 | 6 |

*k*NN [11] is a predicting algorithm that learns from a training data set (in this case, the above samples from the miGCM data set) and predicts samples in a test data set (in this case, the mouse lung sample set). A set of markers (features that best distinguishes two classes of samples, in this case, normal vs. tumor) was selected using the training data set. Distances between the samples were measured in the space of the selected markers. Prediction is performed, one test sample at a time, by: (i), identifying the *k* nearest samples (neighbors) of the test sample among the training data set; and (ii) assigning the test sample to the majority class of these *k* samples.

We first selected markers that best differentiate the normal and tumor samples (see Supplementary Methods) out of the 187 features that passed the filter (which was applied on the training set alone). This generated a list of 131 markers that each has a p-value <0.05 after Bonferroni correction; 129/131 markers are over-expressed in normal samples, whereas 2/131 are over-expressed in the tumor samples. The following table lists these markers.

**Table: Normal/Tumor Makers Selected On the Training Set**

| Probe | Description | Bonferroni-corrected p-value | Variance-thresholded t-test score |
|---|---|---|---|
| EAM159 | hmr_miR-130a | 0 | 10.984 |
| EAM331 | hmr_miR-30e | 0 | 10.756 |
| EAM311 | hmr_miR-101 | 0 | 10.392 |
| EAM299 | hmr_miR-195 | 0 | 9.957 |
| EAM314 | hmr_miR-126 | 0 | 9.498 |
| EAM300 | h_miR-197 | 0 | 8.762 |
| EAM181 | hmr_let-7f | 0 | 8.299 |
| EAM380 | r_miR-140* | 0 | 8.238 |

| | | | |
|---|---|---|---|
| EAM111 | hm_let-7g | 0 | 8.235 |
| EAM381 | r_miR-151* | 0 | 8.198 |
| EAM218 | hmr_miR-152 | 0 | 8.180 |
| EAM183 | hmr_let-7i | 0 | 8.098 |
| EAM253 | hmr_miR-218 | 0 | 8.077 |
| EAM155 | hmr_miR-136 | 0 | 8.058 |
| EAM192 | hmr_miR-126* | 0 | 7.991 |
| EAM222 | hm_miR-15a | 0 | 7.970 |
| EAM161 | hmr_miR-28 | 0 | 7.949 |
| EAM184 | hmr_miR-100 | 0 | 7.894 |
| EAM271 | hmr_miR-30c | 0 | 7.848 |
| EAM270 | hmr_miR-30b | 0 | 7.731 |
| EAM303 | hm_miR-199a* | 0 | 7.519 |
| EAM121 | hmr_miR-99a | 0 | 7.515 |
| EAM392 | r_miR-352 | 0 | 7.476 |
| EAM255 | hmr_miR-22 | 0 | 7.465 |
| EAM249 | hmr_miR-214 | 0 | 7.338 |
| EAM160 | hmr_miR-26b | 0 | 7.313 |
| EAM133 | hmr_miR-324-5p | 0 | 7.266 |
| EAM238 | hm_miR-1 | 0 | 7.259 |
| EAM179 | hmr_let-7d | 0 | 7.235 |
| EAM339 | hmr_miR-99b | 0 | 7.225 |
| EAM185 | hmr_miR-103 | 0 | 7.047 |
| EAM168 | hmr_let-7e | 0 | 7.034 |
| EAM200 | hmr_miR-133a | 0 | 6.959 |
| EAM278 | hmr_miR-98 | 0 | 6.952 |
| EAM333 | hmr_miR-32 | 0 | 6.951 |
| EAM291 | hmr_miR-185 | 0 | 6.910 |
| EAM187 | hmr_miR-107 | 0 | 6.879 |
| EAM263 | hmr_miR-26a | 0 | 6.818 |
| EAM261 | hmr_miR-23b | 0 | 6.814 |
| EAM371 | hmr_miR-342 | 0 | 6.743 |
| EAM330 | hmr_miR-30a-5p | 0 | 6.717 |
| EAM280 | hmr_miR-30a-3p | 0 | 6.662 |
| EAM233 | hmr_miR-196a | 0 | 6.630 |
| EAM292 | hmr_miR-186 | 0 | 6.602 |
| EAM115 | hmr_miR-16 | 0 | 6.558 |
| EAM272 | hmr_miR-30d | 0 | 6.516 |
| EAM367 | hmr_miR-338 | 0 | 6.428 |
| EAM379 | r_miR-129* | 0 | 6.323 |
| EAM193 | hmr_miR-125a | 0 | 6.222 |
| EAM273 | hmr_miR-33 | 0 | 6.209 |
| EAM223 | hmr_miR-15b | 0 | 6.148 |
| EAM105 | hmr_miR-125b | 0 | 6.111 |
| EAM385 | hmr_miR-335 | 0 | 6.011 |
| EAM237 | hmr_miR-19b | 0 | 5.981 |
| EAM320 | hm_miR-189 | 0 | 5.938 |
| EAM262 | hmr_miR-24 | 0 | 5.909 |

| | | | |
|---|---|---|---|
| EAM240 | hmr_miR-20 | 0 | 5.908 |
| EAM260 | hmr_miR-23a | 0 | 5.901 |
| EAM297 | hmr_miR-193 | 0 | 5.856 |
| EAM236 | hmr_miR-19a | 0 | 5.789 |
| EAM264 | hmr_miR-27b | 0 | 5.780 |
| EAM205 | hmr_miR-138 | 0 | 5.721 |
| EAM234 | hmr_miR-199a | 0 | 5.718 |
| EAM207 | hmr_miR-140 | 0 | 5.561 |
| EAM217 | hmr_miR-150 | 0 | 5.531 |
| EAM235 | h_miR-199b | 0 | 5.516 |
| EAM190 | hr_miR-10b | 0 | 5.511 |
| EAM282 | m_miR-199b | 0 | 5.483 |
| EAM335 | h_miR-34b | 0 | 5.315 |
| EAM288 | m_miR-10b | 0 | 5.291 |
| EAM275 | hmr_miR-34a | 0 | 5.287 |
| EAM195 | hmr_miR-128b | 0 | 5.253 |
| EAM328 | hmr_miR-301 | 0 | 5.203 |
| EAM365 | hmr_miR-331 | 0 | 5.191 |
| EAM131 | hmr_miR-92 | 0 | 5.155 |
| EAM215 | hmr_miR-148b | 0 | 5.091 |
| EAM325 | hmr_miR-27a | 0 | 5.090 |
| EAM279 | hmr_miR-29c | 0 | 5.025 |
| EAM369 | hmr_miR-340 | 0 | 4.959 |
| EAM354 | m_miR-297 | 0 | 4.953 |
| EAM119 | hmr_miR-29b | 0 | 4.937 |
| EAM210 | hmr_miR-143 | 0 | 4.908 |
| EAM361 | hmr_miR-326 | 0 | 4.790 |
| EAM324 | hmr_miR-25 | 0 | 4.764 |
| EAM226 | hmr_miR-181a | 0 | 4.742 |
| EAM343 | mr_miR-151 | 0 | 4.740 |
| EAM228 | hmr_miR-181c | 0 | 4.675 |
| EAM366 | mr_miR-337 | 0 | 4.661 |
| EAM349 | mr_miR-292-3p | 0 | 4.652 |
| EAM189 | hmr_miR-10a | 0 | 4.494 |
| EAM355 | mr_miR-298 | 0 | 4.446 |
| EAM318 | h_miR-17-3p | 0 | 4.324 |
| EAM387 | r_miR-343 | 0 | 4.140 |
| EAM363 | mr_miR-329 | 0 | 4.118 |
| EAM268 | hmr_miR-29a | 0 | 4.044 |
| EAM175 | hmr_miR-320 | 0 | 3.875 |
| EAM212 | hmr_miR-145 | 0 | 3.869 |
| EAM378 | mr_miR-7b | 0 | 3.853 |
| EAM281 | mr_miR-217 | 0 | 3.670 |
| EAM307 | m_miR-202 | 0 | 3.625 |
| EAM209 | hmr_miR-142-5p | 0 | 3.594 |
| EAM163 | hmr_miR-142-3p | 0 | 3.545 |
| EAM384 | r_miR-333 | 0 | 3.410 |
| EAM362 | hmr_miR-328 | 0 | 3.356 |

| | | | |
|---|---|---|---|
| EAM329 | hm_miR-302a | 0 | 3.348 |
| EAM368 | hmr_miR-339 | 0 | 3.007 |
| EAM351 | m_miR-293 | 0 | 2.852 |
| EAM153 | hmr_let-7a | 0 | 2.818 |
| EAM360 | mr_miR-325 | 0 | 2.753 |
| EAM145 | hmr_let-7c | 0 | 2.393 |
| EAM348 | mr_miR-291-5p | 0 | 2.092 |
| EAM298 | hmr_miR-194 | 0 | 2.068 |
| EAM250 | h_miR-215 | 0 | 1.746 |
| EAM229 | hm_miR-182 | 0.005 | -4.074 |
| EAM224 | hmr_miR-17-5p | 0.005 | 4.875 |
| EAM341 | m_miR-106a | 0.005 | 4.185 |
| EAM242 | hmr_miR-204 | 0.005 | 3.457 |
| EAM295 | hmr_miR-190 | 0.005 | 3.186 |
| EAM353 | m_miR-295 | 0.005 | 2.916 |
| EAM246 | h_miR-211 | 0.005 | 2.663 |
| EAM248 | hmr_miR-213 | 0.01 | 3.369 |
| EAM186 | h_miR-106a | 0.01 | 4.650 |
| EAM137 | hmr_miR-132 | 0.01 | 3.388 |
| EAM258 | hmr_miR-222 | 0.015 | 4.257 |
| EAM230 | hmr_miR-183 | 0.02 | -3.977 |
| EAM364 | mr_miR-330 | 0.02 | 3.982 |
| EAM206 | hmr_miR-139 | 0.02 | 3.761 |
| EAM327 | hmr_miR-299 | 0.025 | 2.353 |
| EAM232 | hmr_miR-192 | 0.04 | 1.065 |
| EAM257 | hmr_miR-221 | 0.04 | 4.321 |
| EAM216 | hm_miR-149 | 0.04 | 3.711 |

These 131 markers were used without modification to predict the 12 mouse lung samples using the *k*-nearest neighbour algorithm. Each mouse sample was predicted separately, using $\log_2$ transformed mouse and human expression data. The tumor/normal phenotype prediction of a mouse sample was based on the majority type of the *k* nearest human samples using the chosen metric in the selected feature space. Since the tumor/normal distinction was observed at the raw miRNA expression levels, we decided to use Euclidean distance to measure the distances between samples. Thus, we performed *k*NN with the Euclidean distance measure and *k*=3, resulting in 100% accuracy. The detailed prediction results are available in Supplementary Table 3. Similar classification results were obtained with other *k*NN parameters, with the exception of one mouse tumor T_MLUNG_5 (3rd column from right in Fig. 3b). This sample was occasionally classified as normal, for example, when using cosine distance measure (*k*=3). It should be pointed out that cosine distance captures less an overall shift in expression levels compared to Euclidean distance. It rather focuses on comparing the relationships among the different miRNAs So it appears that the same miRNA data capture different information with different distance metrics; Pearson correlation captures information about the lineage (as seen in clustering results), and Euclidean distance captures the normal/tumor distinction.

## Differentiation of HL-60 cells

One hypothesis for the global decrease of miRNA expression in tumors (Fig. 2a, Fig. 3a,b) is that many miRNAs are upregulated during differentiation. We examined an *in vitro* differentiation system, the differentiation of HL-60 acute myeloblastic leukemia cells. HL-60 cells differentiate with increasing neutrophil characteristics upon treatment with *all-trans* retinoic acid (ATRA) during a course of 5 days [12]. We found 59 miRNAs commonly expressed (see Supplementary Methods for the definition of “expressed”) in three independent experiments of HL-60 cells with or without ATRA treatment. A list of these 59 miRNAs is shown below. A heatmap is shown in Fig. 3c, reflecting averages of successfully profiled same condition samples. Results indicate increased expression of many miRNAs after 5 days of ATRA-induced differentiation (5d+). Since HL-60 is a cancerous cell line, this result supports the hypothesis that the global miRNA downregulation in cancer is related to differentiation. Whether or not the observed global miRNA expression change is associated with certain windows of differentiation needs further investigation.

**Table: 59 miRNAs Detected in HL-60 Cells**

| Probe | miRNA |
|---|---|
| EAM103 | Hmr_miR-124a |
| EAM111 | Hm_let-7g |
| EAM115 | Hmr_miR-16 |
| EAM119 | Hmr_miR-29b |
| EAM131 | Hmr_miR-92 |
| EAM145 | Hmr_let-7c |
| EAM270 | hmr_miR-30b |
| EAM163 | hmr_miR-142-3p |
| EAM186 | h_miR-106a |
| EAM209 | hmr_miR-142-5p |
| EAM223 | hmr_miR-15b |
| EAM224 | hmr_miR-17-5p |
| EAM226 | hmr_miR-181a |
| EAM227 | hmr_miR-181b |
| EAM236 | hmr_miR-19a |
| EAM257 | hmr_miR-221 |
| EAM258 | hmr_miR-222 |
| EAM259 | hmr_miR-223 |
| EAM273 | hmr_miR-33 |
| EAM297 | hmr_miR-193 |
| EAM282 | m_miR-199b |
| EAM279 | hmr_miR-29c |
| EAM278 | hmr_miR-98 |
| EAM272 | hmr_miR-30d |
| EAM264 | hmr_miR-27b |
| EAM263 | hmr_miR-26a |
| EAM262 | hmr_miR-24 |
| EAM261 | hmr_miR-23b |
| EAM260 | hmr_miR-23a |
| EAM244 | hmr_miR-21 |
| EAM240 | hmr_miR-20 |
| EAM237 | hmr_miR-19b |
| EAM228 | hmr_miR-181c |
| EAM222 | hm_miR-15a |
| EAM219 | hmr_miR-153 |
| EAM218 | hmr_miR-152 |
| EAM206 | hmr_miR-139 |
| EAM193 | hmr_miR-125a |
| EAM187 | hmr_miR-107 |
| EAM185 | hmr_miR-103 |
| EAM181 | hmr_let-7f |

| | |
|---|---|
| EAM179 | hmr_let-7d |
| EAM175 | hmr_miR-320 |
| EAM160 | hmr_miR-26b |
| EAM153 | hmr_let-7a |
| EAM147 | hmr_let-7b |
| EAM311 | hmr_miR-101 |
| EAM313 | hmr_miR-106b |
| EAM318 | h_miR-17-3p |
| EAM324 | hmr_miR-25 |
| EAM329 | hm_miR-302a |
| EAM331 | hmr_miR-30e |
| EAM337 | hmr_miR-93 |
| EAM341 | m_miR-106a |
| EAM352 | m_miR-294 |
| EAM364 | mr_miR-330 |
| EAM368 | hmr_miR-339 |
| EAM380 | r_miR-140* |
| EAM392 | r_miR-352 |

## Erythroid differentiation of primary hematopoietic cells *in vitro*

We profiled the expression of miRNAs during erythroid differentiation *in vitro* to ask whether the increase in miRNA expression observed in the differentiation of HL-60 cells also occurs in primary cells. The accessibility of normal hematopoietic progenitor cells and the ability to recapitulate erythropoiesis *in vitro* provide a model to study normal differentiation. We purified CD34$^+$ hematopoietic progenitor cells from umbilical cord blood. Erythroid differentiation was induced *in vitro* using a two phase liquid culture system. The state of differentiation of cultured cells was monitored every other day by evaluating expression of CD71 and glycophorin A (Gly-A) (Supplementary Fig. 7b). CD71 expression increases early in erythroid differentiation and gradually decreases in terminal erythroid differentiation. Gly-A expression increases later in erythropoiesis and remains elevated through terminal differentiation. As in HL60 cells, the expression of many miRNAs increased during differentiation (Supplementary Figure 7c). Unlike HL-

60 cells, the erythroid cells continued to proliferate at the time points when miRNA expression increased (Supplementary Figure 7a). This suggests that proliferation itself, which is often integrally linked to differentiation, cannot account completely for the increased miRNA expression during differentiation.

## Analyzing tissue samples using an mRNA proliferation signature

It is conceivable that differences in cellular proliferation, often integrally linked to differentiation, may contribute to the global miRNA signals. We asked whether the miRNA global expression differences among samples are merely a consequence of their differences in proliferation rates. To estimate the proliferation rates in tissue samples, we assembled a consensus mRNA signature of proliferation, reported to positively correlate with proliferation or mitotic index in breast tumors, lymphomas and HeLa cells [13-15]. The table below summarizes this list.

We first asked whether the mRNA proliferation signature reflects proliferation rates in our samples. Indeed, we noticed that the mean expression of these mRNAs is higher in tumors than normal tissues (Supplementary Fig. 8), reflecting faster proliferation rates in tumor samples.

Next, we examined in the tumor samples the expression of the mRNA proliferation signature. We focused on lung and breast, two tissues that we have sufficient numbers of poorly differentiated tumors and more differentiated tumors. It is important to point out that poorly differentiated tumors have globally lower miRNA expression than more differentiated tumors. However, we did not observe any difference in the mRNA proliferation signature between these two categories of samples

(Supplementary Fig. 8). This result also suggests that the global miRNA expression is unlikely to be solely dependent on proliferation rates.

**Table: mRNAs used to estimate proliferation rates**

| Chip | Probe Set ID | Gene Title |
|---|---|---|
| Hu6800 | AB003698_at | CDC7 cell division cycle 7 (S. cerevisiae) |
| Hu6800 | D00596_at | thymidylate synthetase |
| Hu6800 | D14134_at | RAD51 homolog (RecA homolog, E. coli) (S. cerevisiae) |
| Hu6800 | D21063_at | MCM2 minichromosome maintenance deficient 2, mitotin (S. cerevisiae) |
| Hu6800 | D38073_at | MCM3 minichromosome maintenance deficient 3 (S. cerevisiae) |
| Hu6800 | D38550_at | E2F transcription factor 3 |
| Hu6800 | D84557_at | MCM6 minichromosome maintenance deficient 6 (MIS5 homolog, S. pombe) (S. cerevisiae) |
| Hu6800 | J00139_s_at | dihydrofolate reductase pseudogene 1 /// dihydrofolate reductase |
| Hu6800 | J04088_at | topoisomerase (DNA) II alpha 170kDa |
| Hu6800 | J05614_at | proliferating cell nuclear antigen |
| Hu6800 | L07493_at | replication protein A3, 14kDa |
| Hu6800 | L25876_at | cyclin-dependent kinase inhibitor 3 (CDK2-associated dual specificity phosphatase) |
| Hu6800 | L32866_at | baculoviral IAP repeat-containing 5 (survivin) |
| Hu6800 | L47276_s_at | topoisomerase (DNA) II alpha 170kDa |
| Hu6800 | M15796_at | proliferating cell nuclear antigen |
| Hu6800 | M25753_at | cyclin B1 |
| Hu6800 | M34065_at | cell division cycle 25C |
| Hu6800 | M74093_at | cyclin E1 |
| Hu6800 | M87339_at | replication factor C (activator 1) 4, 37kDa |
| Hu6800 | M94362_at | lamin B2 |
| Hu6800 | S49592_s_at | E2F transcription factor 1 |
| Hu6800 | S78187_at | cell division cycle 25B |
| Hu6800 | U04810_at | trophinin associated protein (tastin) |
| Hu6800 | U05340_at | CDC20 cell division cycle 20 homolog (S. cerevisiae) |
| Hu6800 | U14518_at | centromere protein A, 17kDa |
| Hu6800 | U20979_at | chromatin assembly factor 1, subunit A (p150) |
| Hu6800 | U22398_at | cyclin-dependent kinase inhibitor 1C (p57, Kip2) |
| Hu6800 | U26727_at | cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4) |
| Hu6800 | U28386_at | karyopherin alpha 2 (RAG cohort 1, importin alpha 1) |
| Hu6800 | U30872_at | centromere protein F, 350/400ka (mitosin) |
| Hu6800 | U37022_rna1_at | cyclin-dependent kinase 4 |
| Hu6800 | U47677_at | E2F transcription factor 1 |
| Hu6800 | U56816_at | membrane-associated tyrosine- and threonine-specific cdc2-inhibitory kinase |

| | | |
|---|---|---|
| Hu6800 | U65410_at | MAD2 mitotic arrest deficient-like 1 (yeast) |
| Hu6800 | U74612_at | forkhead box M1 |
| Hu6800 | U77949_at | CDC6 cell division cycle 6 homolog (S. cerevisiae) |
| Hu6800 | X05360_at | cell division cycle 2, G1 to S and G2 to M |
| Hu6800 | X13293_at | v-myb myeloblastosis viral oncogene homolog (avian)-like 2 |
| Hu6800 | X51688_at | cyclin A2 |
| Hu6800 | X54942_at | CDC28 protein kinase regulatory subunit 2 |
| Hu6800 | X59543_at | ribonucleotide reductase M1 polypeptide |
| Hu6800 | X59618_at | ribonucleotide reductase M2 polypeptide |
| Hu6800 | X62153_s_at | MCM3 minichromosome maintenance deficient 3 (S. cerevisiae) |
| Hu6800 | X65550_at | antigen identified by monoclonal antibody Ki-67 |
| Hu6800 | X74330_at | primase, polypeptide 1, 49kDa |
| Hu6800 | X74794_at | MCM4 minichromosome maintenance deficient 4 (S. cerevisiae) |
| Hu6800 | X74795_at | MCM5 minichromosome maintenance deficient 5, cell division cycle 46 (S. cerevisiae) |
| Hu6800 | X87843_at | menage a trois 1 (CAK assembly factor) |
| Hu6800 | X89398_cds2_at | uracil-DNA glycosylase |
| Hu6800 | X95406_at | cyclin E1 |
| Hu6800 | X97795_at | RAD54-like (S. cerevisiae) |
| Hu6800 | Z15005_at | centromere protein E, 312kDa |
| Hu6800 | Z29066_s_at | NIMA (never in mitosis gene a)-related kinase 2 |
| Hu6800 | Z29077_xpt1_at | cell division cycle 25C |
| Hu6800 | Z36714_at | cyclin F |
| Hu35KsubA | AA436304_at | RAN, member RAS oncogene family |
| Hu35KsubA | AF004709_at | mitogen-activated protein kinase 13 |
| Hu35KsubA | M96577_s_at | E2F transcription factor 1 |
| Hu35KsubA | RC_AA599859_at | Cyclin B1 |
| Hu35KsubA | RC_AA620553_s_at | flap structure-specific endonuclease 1 |
| Hu35KsubA | U75285_rna1_at | baculoviral IAP repeat-containing 5 (survivin) |
| Hu35KsubA | U78310_at | pescadillo homolog 1, containing BRCT domain (zebrafish) |
| Hu35KsubA | W28391_at | proliferation-associated 2G4, 38kDa |
| Hu35KsubA | X74794_at | MCM4 minichromosome maintenance deficient 4 (S. cerevisiae) |
| Hu35KsubA | Z68092_s_at | cell division cycle 25B |

## RT-PCR analyses of genes involved in miRNA machinery

One possible mechanism of the observed global miRNA expression difference between normal samples and tumors is changes in expression levels of miRNA processing enzymes. In lung cancer, Dicer levels were reported to correlate with prognosis [16]. We decided to examine Dicer1, Drosha, DGCR8 and Argonaute 2 (Ago2), which are critical in miRNA processing [17]. Lacking probe sets representing these genes in our mRNA data, we used quantitative RT-PCR and analyzed 79 samples (32 normal samples and 47 tumors, covering 8 tissues, including colon, breast, uterus, lung, kidney, pancreas, prostate and bladder). We normalized the quantitative PCR data with 18S rRNA levels. We performed Student's t-test (two-tail, unequal variance) for normal/tumor phenotypes on all samples examined (P = 0.3 for Dicer1, P = 0.11 for Drosha, P = 0.0011 for DGCR8, P = 0.0138 for Ago2). DGCR8 and Ago2 have significant nominal p-values under the above test. However, the fold differences of DGCR8 and Ago2 are small between tumors and normal samples (tumor samples have higher mean threshold cycle (Ct) values for these two genes; the mean Ct differences between normal and tumor samples are: 0.776 for DGCR8 and 0.798 for Ago2, corresponding to 1.7-fold and 1.5-fold absolute level differences respectively, after correction for PCR amplification efficiency). Whether or not the observed weak decreases on the transcript level may account for the differences in miRNA expression needs further investigation. It is also important to note that these results do not exclude the possibility that these miRNA machinery genes are involved in regulating tumor/normal miRNA expression in certain cancer types, or are regulated on the protein and activity levels.

## Analyses of poorly differentiated tumors

We first set out to determine whether poorly differentiated tumors show a globally weaker miRNA expression than tumor samples in the miGCM collection, which represent more differentiated states. To this end, we made a comparison of poorly differentiated tumors to more differentiated tumors of the corresponding tissue types. The analysis was performed on 180 features, after the data were filtered to eliminate non-expressing miRNAs on the 55 samples which belong to tissue types that have both more-differentiated and poorly-differentiated samples (see the hierarchical clustering section in Supplementary Methods for data filtration). Supplementary Fig. 5 shows that poorly differentiated tumors indeed have globally lower miRNA expression. Out of the 180 features, 95 miRNAs display lower mean expression levels in poorly differentiated tumors ($p<0.05$ with a variance-thresholded t-test).

We used PNN for prediction of tissue origin of poorly differentiated tumors. PNN is a probability based prediction algorithm and can be considered as a smooth version of *k*NN. For a multi-class prediction, PNN avoids the ambiguity often encountered with *k*NN, when multiple training classes are equally presented in the *k* nearest neighbours of a test sample. For a two-class classification problem, PNN assigns a probability for a test sample to be classified into one of the two classes. The contribution of each training sample to the classification of a test sample is related to their distance and follows the Gaussian distribution: the closer the test sample, the larger the contribution. The probability for a test sample to belong to a certain class is the total contribution from every training sample belonging to that class, divided by the total contributions of all training samples (see Supplementary Methods for more details).

For the prediction of poorly differentiated tumors, the training sample set consists of 68 tumor samples with both miRNA and mRNA profiling data, covering 11 tissue types. The test set contains 17 poorly differentiated tumors. A table below summarizes the information on the 17 poorly differentiated tumors. To solve this multi-class prediction problem, we broke down the task into 11 two-class predictions. Each two-class prediction assigns a probability for a test sample to belong to a certain tissue-type vs. the rest of the tissue-types (one vs. the rest, OVR), for example, colon vs. non-colon. After performing OVR classifications for all 11 tissues, the one tissue-type that receives the highest probability marks the predicted tissue type. The prediction results are summarized in supplementary Table 4.

**Table: Information on Poorly Differentiated Tumor Samples**

| Sample Name | Sample of Primary or Metastatic Origin | Primary Site | Metastatic Site |
|---|---|---|---|
| PDT_BRST_1 | Primary | Breast | |
| PDT_BRST_2 | Primary | Breast | |
| PDT_BRST_3 | Primary | Breast | |
| PDT_BRST_4 | Primary | Breast | |
| PDT_BRST_5 | Metastatic | Breast | Lymph node /supraclavic |
| PDT_COLON_1 | Primary | Colon | |
| PDT_LBL_1 | Primary | Lymph node | Groin |
| PDT_LUNG_1 | Metastatic | Lung | Kidney |
| PDT_LUNG_2 | Primary | Lung | |
| PDT_LUNG_3 | Primary | Lung | |
| PDT_LUNG_4 | Primary | Lung | |
| PDT_LUNG_5 | Metastatic | Lung | Adrenal |
| PDT_LUNG_6 | Primary | Lung | |
| PDT_LUNG_7 | Primary | Lung | |
| PDT_LUNG_8 | Primary | Lung | |
| PDT_OVARY_1 | Primary | Ovary | |
| PDT_OVARY_2 | Metastatic | Ovary | Omentum |
| PDT_OVARY_3 | Primary | Ovary | |
| PDT_STOM_1 | Primary | Stomach / GE_Jct | |

## *Supplementary References*

1. Ramaswamy, S. et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 98, 15149-54 (2001).
2. Ebert, B. L. et al. An RNA interference model of RPS19 deficiency in Diamond Blackfan Anemia recapitulates defective hematopoiesis and rescue by dexamethasone: identification of dexamethasone responsive genes by microarray. Blood (2005).
3. Miska, E. A. et al. Microarray analysis of microRNA expression in the developing mammalian brain. Genome Biol 5, R68 (2004).
4. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. Science 294, 858-62 (2001).
5. Jain, A. K. & Dubes, R. C. Algorithms for clustering data. Prentice-Hall Inc., Upper Saddle River (1988).
6. Good, P. I. Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses, 2nd Ed. Springer-Verlag, New York (2000).
7. Specht, D. F. Probabilistic Neural Networks, Neural Networks. Elsevier Science Ltd., St. Louis 3, 109-118 (1990).
8. Griffiths-Jones, S. The microRNA Registry. Nucleic Acids Res 32 Database issue, D109-11 (2004).
9. Ambros, V. et al. A uniform system for microRNA annotation. RNA 9, 277-9 (2003).
10. Lagos-Quintana, M. et al. Identification of tissue-specific microRNAs from mouse. Curr Biol 12, 735-9 (2002).
11. Duda, R. O., Hart, P. E. & Stork, D. G. Pattern Classification, 2nd Ed. Wiley-Interscience, Hoboken (2000).
12. Stegmaier, K. et al. Gene expression-based high-throughput screening(GE-HTS) and application to leukemia differentiation. Nat Genet 36, 257-63 (2004).
13. Alizadeh, A. A. et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, 503-11 (2000).
14. Perou, C. M. et al. Molecular portraits of human breast tumours. Nature 406, 747-52 (2000).
15. Whitfield, M. L. et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell 13, 1977-2000 (2002).
16. Karube, Y. et al. Reduced expression of Dicer associated with poor prognosis in lung cancer patients. Cancer Sci 96, 111-5 (2005).
17. Tomari, Y. & Zamore, P. D. MicroRNA biogenesis: drosha can't cut it without a partner. Curr Biol 15, R61-4 (2005).