Citation: Clinical and Translational Gastroenterology (2016) 7, e152; doi:10.1038/ctg.2016.11
© 2016 the American College of Gastroenterology All rights reserved 2155-384X/16
www.nature.com/ctg

Expression Profiles of miRNA Subsets Distinguish Human Colorectal Carcinoma and Normal Colonic Mucosa

Daniel F. Pellatt, MStat[1], John R. Stevens, PhD[2], Roger K. Wolff, PhD[1], Lila E. Mullany, MS[1], Jennifer S. Herrick, MS[1], Wade Samowitz, MD[3] and Martha L. Slattery, PhD[1]

OBJECTIVES: MicroRNAs (miRNAs) are small, non-protein-coding RNA molecules that are commonly dysregulated in colorectal tumors. The objective of this study was to identify smaller subsets of highly predictive miRNAs.
METHODS: Data come from population-based studies of colorectal cancer conducted in Utah and the Kaiser Permanente Medical Care Program. Tissue samples were available for 1,953 individuals, of which 1,894 had carcinoma tissue and 1,599 had normal mucosa available for statistical analysis. Agilent Human miRNA Microarray V.19.0 was used to generate miRNA expression profiles; validation of expression levels was carried out using quantitative PCR. We used random forest analysis and verified findings with logistic modeling in separate data sets. Important microRNAs are identified and bioinformatics tools are used to identify target genes and related biological pathways.
RESULTS: We identified 16 miRNAs for colon and 17 miRNAs for rectal carcinoma that appear to differentiate between carcinoma and normal mucosa; of these, 12 were important for both colon and rectal cancer, hsa-miR-663b, hsa-miR-4539, hsa-miR-17-5p, hsa-miR-20a-5p, hsa-miR-21-5p, hsa-miR-4506, hsa-miR-92a-3p, hsa-miR-93-5p, hsa-miR-145-5p, hsa-miR-3651, hsa-miR-378a-3p, and hsa-miR-378i. Estimated misclassification rates were low at 4.83% and 2.5% among colon and rectal observations, respectively. Among independent observations, logistic modeling reinforced the importance of these miRNAs, finding the primary principal components of their variation statistically significant ($P<0.001$ among both colon and rectal observations) and again producing low misclassification rates. Repeating our analysis without those miRNAs initially identified as important identified other important miRNAs; however, misclassification rates increased and distinctions between remaining miRNAs in terms of classification importance were reduced.
CONCLUSIONS: Our data support the hypothesis that while many miRNAs are dysregulated between carcinoma and normal mucosa, smaller subsets of these miRNAs are useful and informative in discriminating between these tissues.

Clinical and Translational Gastroenterology (2016) **7,** e152; doi:10.1038/ctg.2016.11; published online 10 March 2016
Subject Category: Colon/Small Bowel

INTRODUCTION

MicroRNA (miRNA or miR) are noncoding RNA molecules that alter gene activity through both the translation reduction and the decay of mRNA.[1] They regulate key cellular processes including division and differentiation,[2] and altered miRNA expression has been associated with several diseases including cancer.[3] Dysregulated miRNAs associated with cancer, including colorectal cancer (CRC) specifically, have been discussed as potential diagnostic tools[4,5] and have generated treatment ideas.[6] Previous studies[7–9] have identified several miRNAs with differential expression between carcinoma and non-tumor tissue among individuals with CRC. Many of these studies had relatively small sample sizes and identified and focused on modest lists of differentially expressed miRNAs. Our previous analysis[10] with a comparatively large sample size indicated that over 86% of miRNAs expressed in >80% of the population were differentially expressed between carcinoma tissue and normal mucosa. As more researchers use miRNA arrays that analyze thousands of miRNAs rather than targeted miRNAs, identification of key miRNAs in the carcinogenic process becomes more challenging. Given the extent of dysregulated miRNAs in CRC, it is desirable to identify important subsets of miRNAs that distinguish these tumors from non-tumor tissue.

Here, we focus on distinctions between the miRNA expression profiles of carcinoma tissue and those of adjacent normal mucosa in individuals with CRC. Our aim is to identify smaller groups of miRNAs with expression profiles that are highly predictive of carcinoma vs. normal tissue. We use random forest analysis to identify important miRNAs for classifying tissue as either carcinoma or normal colonic tissue. Random forests are competitive classifiers in a variety of scenarios[11,12] and have been proffered as microarray analysis tools because of their ability to classify using many input variables compared with the number of observations while simultaneously providing a useful, cross-validation-based, measure of input variable importance, the out-of-bag (OOB) error rate.[13–15] Our analysis centers around tissue

[1]Department of Internal Medicine, University of Utah, Salt Lake City, Utah, USA; [2]Department of Mathematics and Statistics, Utah State University, Logan Utah, USA and [3]Department of Pathology, University of Utah, Salt Lake City, Utah, USA
Correspondence: Daniel F. Pellatt, MStat, Department of Internal Medicine, University of Utah, 383 Colorow, Salt Lake City, Utah 84108, USA. E-mail: Daniel.Pellatt@hsc.utah.edu
Received 17 November 2015; accepted 3 February 2016

Table 1 Samples after pair selection

Primary data set	Tissue			Secondary data set	Tissue		
	Normal mucosa	Carcinoma	Total		Normal mucosa	Carcinoma	Total
Colon							
Male	220	304	524	Male	47	62	109
Female	193	276	469	Female	43	48	91
Total	413	580	993	Total	90	110	200
Rectal							
Male	126	191	317	Male	48	68	116
Female	93	150	243	Female	42	42	84
Total	219	341	560	Total	90	110	200

collected from 1,953 individuals with CRC and considers colon and rectal carcinomas separately. Independent verification was performed using logistic modeling. For additional insight into the cellular processes associated with our findings, bioinformatics tools were used to identify target genes and related biological pathways associated with specific miRNAs identified.

METHODS

Study participants. Study participants came from two population-based case–control studies that included all incident colon and rectal cancers verified through tumor registries and between 30 and 79 years of age who resided along the Wasatch Front in Utah or were members of the Kaiser Permanente Medical Care Program (KPMCP) in Northern California as described previously.[16,17] Cases were obtained via a rapid-reporting system; for those cases who did not participate in the interview portion of the study, deidentified tissue was obtained via the tumor registry along with tumor characteristics. Thus, this study includes tissue obtained from 97% of all Utah cases diagnosed and for 85% of all Kaiser Permanente Medical Care Program study participants.[18] The study was approved by the University of Utah Institutional Review Board Protocol numbers IRB_00002335 and IRB_00055877. All study participants provided informed consent.

miRNA processing. After extracting RNA from formalin-fixed, paraffin-embedded tissue, the Agilent Human miRNA Microarray V.19.0 (Agilent Technologies, Santa Clara, CA) was used to obtain miRNA expression results for 2,006 unique human miRNAs.[10] Samples that failed initial quality control parameters established by Agilent that included tests for excessive background fluorescence, excessive variation among probe sequence replicates on the array, and measures of the total gene signal on the array to assess low signal were repeated a second time. After quality control, there were 1,953 individuals contributing carcinoma and/or normal mucosa tissue observations among whom 1,894 individuals contributed carcinoma tissue and 1,599 contributed normal mucosa. Normal mucosa was taken from the same site adjacent to the index carcinoma.

To minimize differences that could be attributed to the array, amount of RNA, location on array, or other factors that could erroneously influence expression, total gene signal was normalized by multiplying expression values of each sample by a scaling factor (median of the 75th percentiles of all the samples divided by the individual 75th percentile of each sample), stratified by carcinoma site. We refer to miRNAs using standard nomenclature used in the miRBase database.[19]

Statistical methods. After quality control, observations consisted of 1,193 individuals with colon tissue (carcinoma, normal mucosa, or both) and 760 with rectal tissue; the colon observations were considered separately from the rectal observations. Within these groups, we focused on miRNAs that were expressed in most individuals, dropping any with zero measured expression in more than 10% of individuals. This amounted to considering expression values for 522 miRNAs among colon observations and 545 miRNAs among rectal observations for analysis. For the colon study, paired samples of both carcinoma and normal mucosa tissues were available for 955 individuals, whereas only one tissue type was available for 238 individuals. Among the rectal study, these numbers were 585 and 175, respectively. As our data contained both paired and non-paired observations and because the models used are not designed for paired observations, for each individual with both carcinoma and normal mucosa tissue, one of these tissue types was randomly kept for model fitting, whereas the other was withheld for subsequent validation. Within both the colon and rectal observations, 200 individuals were randomly selected and set aside to form independent secondary data sets. Table 1 contains a summary of the non-withheld observations in the primary and secondary colon and rectal data sets; a similar table for the withheld pairs is available in the online Supplementary Table S1 online.

All statistical analysis was performed in R.[20] To identify miRNAs of particular importance in distinguishing normal mucosa from carcinoma tissue, random forests were fit to the primary colon and rectal data sets using the R package “randomForest”.[21,22] These models were fit to classify tissue type, i.e., normal mucosa vs. carcinoma tissue, using miRNA expression values, sex, and age category as explanatory variables. Age categories were $\leqslant 50$, 51–60, 61–70, and 71 or more years of age at diagnosis. Proximal vs. distal subsite also was included as an explanatory variable in the random forest modeling associated with the colon data set. Classification error rates and the rankings of variable importance scores did not meaningfully change when adjusting various model parameters beyond the default settings of the “randomForest” package. Aside from setting the number of trees equal to 10,000, the default settings were used. K-means clustering with $k=2$ was applied to the resulting importance scores associated with the explanatory variables; this was performed on importance scores resulting from the colon and rectal models separately. Those explanatory variables that clustered toward the greatest mean importance measure, all of which were miRNAs, were selected as being of particular importance. Supplementary 1 (S1 Text) provides additional information on the Random Forest procedure used.

Random forests lack some of the parametric structure and related statistical tests associated with more traditional modeling techniques. To verify our findings, we fit logistic models to the secondary colon and rectal data sets ($N=200$ in each). Normal mucosa vs. carcinoma tissue was the outcome modeled and only the miRNAs identified in the respective primary data sets as important were considered for modeling. For each model, rather than including all miRNAs designated as important as explanatory variables, dimension was reduced via principal component analysis (PCA) performed separately for both the secondary colon and rectal data sets on the mean centered and sample standard deviation scaled miRNA expressions. The first five principal component scores were considered as potential explanatory variables for inclusion in the logistic models. The Akaike information criterion (AIC) was used to select which components were included in the final models. The logistic models yielded P-values associated with the coefficients attached to the included principal components (alternative hypothesis: coefficient is non-zero). The logistic models were considered further as classification models. Tissue was classified according to which type the model estimated to be of greater probability; leave-one-out cross-validation was used to measure the associated classification accuracy of the logistic models.

To further assess the accuracy of the random forest and logistic models, we considered the withheld pairs. Among the primary and secondary data sets for both colon and rectal tissue, the models fit to the non-withheld pairs were used to predict the tissue types of the withheld observations and misclassification rates were observed. In the secondary data sets where logistic models were applied, the PCA loadings computed from the non-withheld secondary data sets were used to compute PCA scores for the withheld observations so that no further model fitting was applied to the withheld observations.

Bioinformatics analysis. To determine target genes and related biological pathways associated with the identified miRNAs that classified tissue type, miRNA targets were generated using DIANA-TarBase V.7.0 (http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=tarbase/index),[23] a repository of validated miRNA targets. Target genes were then used as input to the Database for Annotation, Visualization, and Integrated Discovery (DAVID) V.6.7[24,25] (Enriched Gene Ontology (GO))[26] terms for biological processes were then pulled using DAVID’s Functional Annotation Tool. Biological process terms were selected as significant using the criterion false discovery rate <0.05. DAVID uses the Kyoto Encyclopedia of Genes and Genomes (KEGG).[27]

Quantitative PCR for validation of Agilent platform expression measurements. We compared expression measurements from the Agilent platform with quantitative PCR (qPCR) measurements to validate data characteristics across platforms for several miRNAs identified as important. One hundred and eighty samples, representing 45 normal mucosa/carcinoma pairs from individuals with colon tumors and 45 paired samples from individuals with rectal tumors, were selected for qPCR measurement. cDNA was generated for 11 specific miRNAs using 10 ng of total RNA in a multiplex reaction using the TaqMan MicroRNA Reverse Transcription Kit (Life Technologies, Carlsbad, CA) and TaqMan assay-specific primers (all assays were purchased from Life Technologies, Carlsbad, CA). A multiplexed 12-cycle preamplification step was performed according to the manufacturer’s recommendation. The preamplified material was then diluted 1:16 and individual TaqMan assays were performed using 40 cycles and collecting real-time data on an ABI 7900HT. Data were evaluated using Life Technologies Quant Studio Flex 12K software (Thermo Fisher Scientific, Waltham, MA) to determine the number of cycles required for each sample to cross a common threshold. Ten of these miRNAs selected for comparison were identified as important among colon observations, eight of which also were identified as important among rectal observations. One miRNA, miR--1183, not identified as important in the random forest analysis was selected for inclusion to serve as a housekeeping gene to control for individual-specific variation. This specific miRNA was not differentially expressed between carcinoma and normal mucosa in our data and had high levels of expression in both tissues in almost all samples. Expression measurements were normalized using the $2^{-\Delta C_T}$ method.[28,29] To evaluate similarity between the measurements of the two platforms, Agilent expression measurements were contrasted with qPCR expression measurements among the individuals with colon tumors separately from those with rectal tumors. For each group, among individuals for whom both qPCR and Agilent expressions were available ($N=45$), each miRNA identified as important for which qPCR data was available was considered individually. For each such miRNA, correlation between Agilent and qPCR expressions was computed among both carcinoma and normal mucosa samples. Agilent platform fold changes between carcinoma tissue and normal colonic mucosa also were compared with fold change measurements from qPCR.

RESULTS

The primary colon and rectal data sets consisted of 52.8% and 56.6% men, respectively, whereas 54.5% and 58.0% of individuals among the secondary colon and rectal data sets were men. Of the individuals in the primary colon data set, 50.5% had proximal tumors, whereas 49.5% had distal tumors and the same percentages were observed among the secondary colon data set. Considering age at diagnosis, among individuals in the primary colon data set 9.3% were

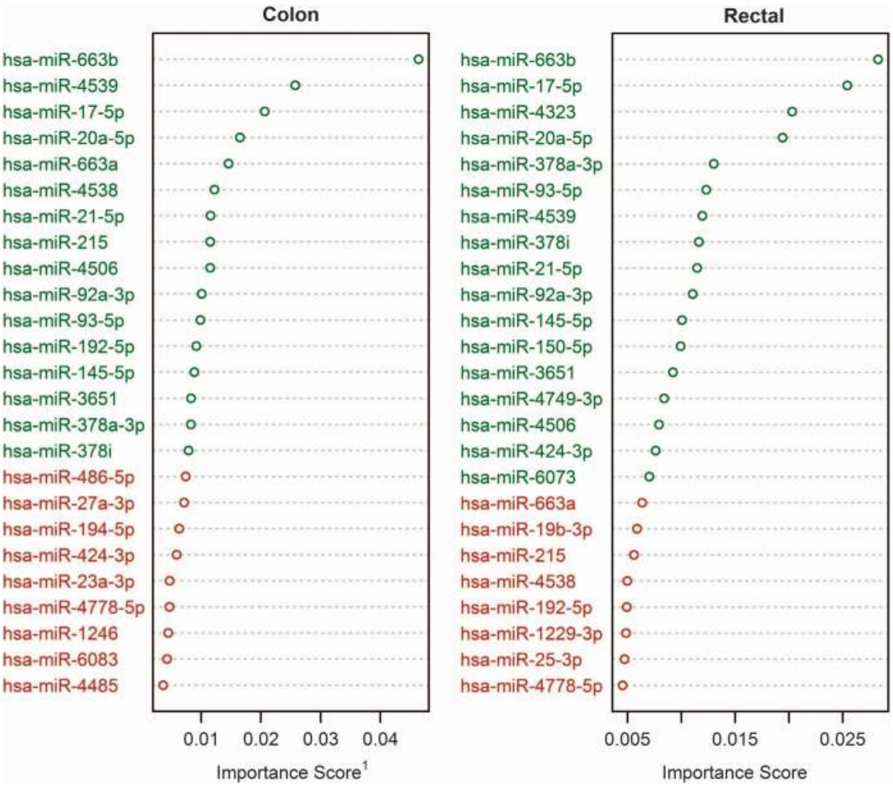

Figure 1 Greatest importance scores. Green indicates those microRNAs (miRNAs) selected as most important via *k*-means clustering. [1]Mean decrease in tree accuracy from permutation.

50 years of age or less, 19.6% were between 51 and 60 years, 36.5% were between 61 and 70 years, and 34.6% were 71 years or older; in the secondary colon data set, these percentages were 11.0%, 12.5%, 39.5%, and 37.0%, respectively. Among individuals in the primary rectal data set, 16.8% were 50 years of age or less, 22.3% were between 51 and 60 years, 33.8% were between 61 and 70 years, and 27.1% were 71 years or older; these percentages among individuals in the secondary rectal data set were 16.5%, 22%, 38%, and 23.5%, respectively.

The random forest models' importance measurements paired with *k*-means analysis ($k=2$) identified 16 miRNAs as most important in discriminating between carcinoma and normal mucosa for colon cancer; 17 miRNAs were identified for rectal cancer. Of these miRNAs, 12, hsa-miR-663b, hsa-miR-4539, hsa-miR-17-5p, hsa-miR-20a-5p, hsa-miR-21-5p, hsa-miR-4506, hsa-miR-92a-3p, hsa-miR-93-5p, hsa-miR-145-5p, hsa-miR-3651, hsa-miR-378a-3p, and hsa-miR-378i, were identified as important discriminators for both colon and rectal cancer. We identified hsa-miR-663a, hsa-miR4538, hsa-miR-215, and hsa-miR-192-5p uniquely in association with colon cancer; hsa-miR-4323, hsa-miR-150-5p, hsa-miR-4749-3p, hsa-miR-424-3p, and hsa-miR-6073 were identified uniquely in association with rectal cancer. Figure 1 displays the 25 miRNAs with the highest variable importance scores in both the colon and rectal data sets. Green text indicates those grouped by *k*-means clustering as most important, whereas red text indicates membership in the cluster of miRNAs with lower importance scores, not all of which are shown. Sex, age category, and colonic site (i.e., proximal vs. distal colon) received variable importance scores among the closest to zero; as such, the random forest analysis does not provide evidence of their interaction with the identified miRNAs in terms of classifying tissue type. Summary information regarding expression values from the Agilent platform for miRNAs among the clusters of the greatest mean importance in the colon and rectal data sets are presented in Table 2.

The OOB error (misclassification) estimate for the random forest model fit to the primary colon data set was $<5\%$ as was the misclassification rate among the associated paired observations withheld from the random forest modeling. Among the primary rectal data set, the estimated error rates were lower, with an OOB error estimate of 2.50% associated with the random forest model fit to the primary data set and an error rate of 3.51% among the pairs withheld from modeling. Figure 2 illustrates some findings of the random forest models. It contrasts the expression values of the top three miRNAs, as determined by importance score rank, against one another among individuals in the primary colon and rectal data sets and is color-coded by tissue type.

In the secondary colon data set, the logistic model included the first three principal components as explanatory variables. These accounted for 76.41% of the sample variance among the important miRNAs. For each of these components, the *P*-value associated with the model coefficient was <0.001. The estimated misclassification rate was 6.50% among the secondary colon observations and 5.10% among the associated pairs withheld from modeling. Considering the secondary rectal data set, the first, third, and four principal

Table 2 Important miRNA summary statistics

	Mean-adjusted counts				Mean-adjusted counts		
miRNA	**Normal mucosa**	**Carcinoma tissue**	**Fold change**[a]	**miRNA**	**Normal mucosa**	**Carcinoma tissue**	**Fold change**
Colon data set				*Rectal data set*			
Hsa-miR-663b	34.40	68.54	1.99	Hsa-miR-663b	30.79	61.96	2.01
Hsa-miR-4539	83.22	53.67	0.64	Hsa-miR-17-5p	14.18	53.07	3.74
Hsa-miR-17-5p	15.07	49.57	3.29	Hsa-miR-4323	14.39	8.51	0.59
Hsa-miR-20a-5p	16.32	57.34	3.51	Hsa-miR-20a-5p	15.07	60.65	4.02
Hsa-miR-663a	276.96	416.95	1.51	Hsa-miR-378a-3p	152.32	113.25	0.74
Hsa-miR-4538	188.66	138.38	0.73	Hsa-miR-93-5p	13.18	35.72	2.71
Hsa-miR-21-5p	145.44	389.67	2.68	Hsa-miR-4539	67.90	44.73	0.66
Hsa-miR-215	66.77	39.18	0.59	Hsa-miR-378i	75.13	57.02	0.76
Hsa-miR-4506	6.62	10.06	1.52	Hsa-miR-21-5p	127.02	381.56	3.00
Hsa-miR-92a-3p	39.52	99.46	2.52	Hsa-miR-92a-3p	44.28	115.80	2.61
Hsa-miR-93-5p	13.39	34.34	2.56	Hsa-miR-145-5p	270.28	131.80	0.49
Hsa-miR-192-5p	128.10	81.90	0.64	Hsa-miR-150-5p	38.50	12.87	0.33
Hsa-miR-145-5p	210.80	125.59	0.60	Hsa-miR-3651	27.47	59.73	2.17
Hsa-miR-3651	23.57	52.27	2.22	Hsa-miR-4749-3p	14.90	8.83	0.59
Hsa-miR-378a-3p	152.31	113.77	0.75	Hsa-miR-4506	5.42	8.88	1.64
Hsa-miR-378i	75.85	57.65	0.76	Hsa-miR-424-3p	23.45	36.87	1.57
—	—	—	—	Hsa-miR-6073	7.14	4.43	0.62

miRNA, microRNA.
[a] Calculated as mean normal mucosa/mean carcinoma.

components were included in the logistic model. These accounted for 59.88% of the sample variance among important rectal miRNAs. The *P*-values corresponding to the coefficients associated with these components within the fitted logistic model were all <0.05. The misclassification rate among the secondary rectal data set was 3.50%; the corresponding misclassification rate among the associated withheld pairs was 3.16%. Confusion matrices associated with the random forest and logistic models for both the colon and rectal data sets are presented in Table 3. Summary information regarding the PCA and logistic modeling results is found in Table 4. Additionally, among the secondary data sets we tested for differences in PCA score profiles across sex, tumor site, and tumor stage categories, considering tumor observations separately from non-tumor observations. We did not observe evidence of interactions between the PCA scores and these categorical variables; a summary of these results is included in the Supplementary Table S2.

Regarding the miRNAs identified as most important in distinguishing between tissue type among the colon and rectal data sets, Table 5 contains bioinformatics analysis results including the number of target mRNAs identified, and associated enriched biological processes and pathways. Fairly conservative identification methods were used (i.e., verified mRNA targets only combined with false discovery rate <0.05); important miRNAs for which target mRNAs were not identified were excluded from presentation in Table 5. Of the miRNAs associated with colon tissue only, hsa-miR-663a was associated with four enriched biological processes, and hsa-miR-92-5p was associated with nine enriched biological processes and eight enriched KEGG pathways. Three miRNAs associated with both colon and rectal tissue targeted mRNAs enriched for multiple processes and KEGG pathways. Hsa-miR-17-5p was associated with six enriched biological processes and seven enriched KEGG pathways, and hsa-miR-20a-5p and hsa-miR-21-5p were both associated with the same nine enriched processes and eight enriched KEGG pathways. Hsa-miR-378a-3p was identified among both the colon and rectal data sets and was associated with just one significantly enriched KEGG pathway, “pathways in cancer”, whereas hsa-miR-145-5p, also identified among both data sets, targeted genes that were not significantly enriched for any pathways but were enriched for five biological processes. None of the miRNAs associated with rectal tissue only had any targeted genes enriched for either biological processes or KEGG pathways. Of the gene (mRNA) targets identified in association with the important miRNAs, 26 contributed to the identification of significantly enriched KEGG pathways, five of which were targeted by multiple miRNAs. Regarding the enriched biological processes, 100 target genes contributed to the identification of these processes, 11 of which were targeted by multiple miRNAs identified as important. The genes associated with KEGG pathways and biological processes as well as the miRNAs for which they are verified targets are included in the Supplementary Table S3.

Table 6 contains results comparing the Agilent platform measurements with those from qPCR among several miRNA identified as important. qPCR miRNA expression was measured for hsa-miR-663b, hsa-miR-17-5p, hsa-miR-20a-5p, hsa-miR-93-5p, hsa-miR-21-5p, hsa-miR-92a-3p, hsa-miR-3651, hsa-miR-4506, hsa-miR-4538, hsa-miR-215, and hsa-miR-1183. All fold changes between carcinoma tissue and normal colonic mucosa followed the same pattern among qPCR measurements and Agilent measurements, i.e., each miRNA with a fold change of <1 when measured by the Agilent platform also had a fold change of <1 when measured via qPCR and *vice versa* for miRNA with fold changes >1. Additionally, most of the miRNA expressions displayed a high level of correlation between Agilent and qPCR measurements; 75% of the correlations computed were >0.5. We believe this demonstrates a high level of agreement between the Agilent platform expression measurements and those from qPCR and

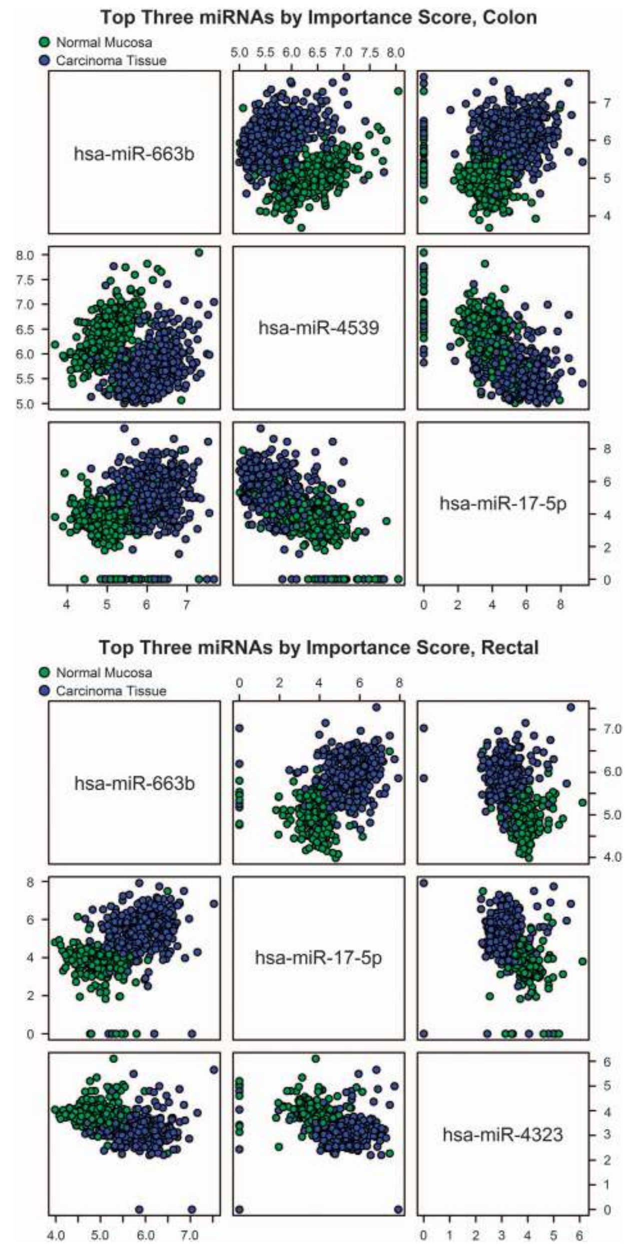

Figure 2 Scatterplots of microRNA (miRNA) expression values among primary colon and rectal data sets. Axes measure log 2(1+expression value). Observations were randomly ordered to avoid any data-order artifacts in this visualization.

provides a degree of validation with regard to the platform used in generating our data, the Agilent platform.

DISCUSSION

The miRNAs identified appear viable for distinguishing between carcinoma and normal colonic mucosa at the molecular level. Estimated misclassification rates were low and we identified 16 miRNAs of particular importance in discriminating between colon carcinoma tissue and normal mucosa; similarly, 17 miRNAs were identified as particularly important with regard to rectal carcinoma classification. Among independent observations, logistic modeling coupled with PCA verified findings via parsimonious predictive models of carcinoma vs. normal mucosa using only the miRNAs identified via random forest analysis and similar classification accuracy was observed. This study helps describe the landscape of miRNAs as they relate to CRC. It is hoped that the ability to narrow focus to key molecular differences between carcinoma and adjacent normal mucosa will aid future clinical research, from screening tools to targeted therapeutic modalities.

Bioinformatics tools identified several miRNA targets, enriched biological processes, and pathways associated with miRNAs identified as important among the colon and rectal data sets. One of the most common threads in the pathways identified with these miRNAs was angiogenesis. This suggests that these miRNAs have the potential to contribute to the metastatic potential of tumors. Additionally, several of the miRNAs identified as important have been identified in the previous research. Hsa-miR-21 identified in this study as important for both colon and rectal study has been studied extensively with colon cancer.[9,30–34] Hsa-miR-663b was shown to be upregulated in bladder cancer plasma, as such has been proposed as a biomarker in clinical bladder cancer detection,[35] and it has also been seen to be involved in cell proliferation, migration, apoptosis, and regulation of MAP/ERK (mitogen-activated protein/extracellular signal-regulated kinase) signaling in a study of CRC cell lines.[36] Both hsa-miR-21-5p and hsa-miR-17-5p were seen to be significantly dysregulated in a CRC study by Kara *et al*.[37] Additionally, higher hsa-miR-17-5p expression was correlated with drug resistance and metastasis in CRC patients in a 2014 study by Fang *et al*.[38] MiRNA hsa-miR-4323 was correlated with tumor relapse in patients with small-cell esophageal carcinoma.[39] Other miRNAs identified as important, which have been previously reported as associated with CRC, include miR-20a,[31,32,40] miR-92a,[41] miR-192-5p,[42] miR-145,[40,43–45] miR-93,[46] and miR-150.[47]

Previous analysis of our data indicated that a large percentage of miRNAs exhibit dysregulation between carcinoma tissue and normal mucosa.[10] To consider if we could achieve similar misclassification rates with alternative subsets of miRNAs, we removed those identified as important from consideration and repeated our analysis. Obtaining new subsets of important miRNAs for colon and rectal in the same manner as before, we again removed these secondary findings from consideration and repeated the analysis a third time. Considering the colon data set, the second round of analysis identified 27 miRNAs as most important and resulted in an OOB estimated error rate for the new random forest model among the primary data set of 7.78% and a leave-one-out estimated error rate for the newly fitted logistic model of 10.50% among the secondary data set. The third round of analysis using the colon data set identified 48 miRNAs as most important, whereas the OOB error estimate among the primary data set using random forest modeling was 11.48% and the leave-one-out estimate associated with logistic modeling in the secondary data set was 12.66%. Considering the rectal data set, the second round of analysis identified 20 miRNAs as most important; the OOB error estimate associated with random forest modeling was 4.29%, whereas the

Table 3 Confusion matrices associated with random forest and logistic classification models

Random forest model				Error rate[a]	Logistic model				Error rate[a]
Colon data set									
Primary data set					Secondary data set				
		Predicted					Predicted		
		Normal mucosa	Carcinoma				Normal mucosa	Carcinoma	
Actual	Normal mucosa	384	29		Actual	Normal mucosa	83	7	
	Carcinoma	19	561			Carcinoma	6	104	
				4.83%					6.50%
Pairs withheld from model fitting					Pairs withheld from model fitting				
		Predicted					Predicted		
		Normal mucosa	Carcinoma				Normal mucosa	Carcinoma	
Actual	Normal mucosa	389	27		Actual	Normal mucosa	73	5	
	Carcinoma	10	372			Carcinoma	3	76	
				4.64%					5.10%
Rectal data set									
Primary data set					Secondary data set				
		Predicted					Predicted		
		Normal mucosa	Carcinoma				Normal mucosa	Carcinoma	
Actual	Normal mucosa	213	6		Actual	Normal mucosa	87	3	
	Carcinoma	8	333			Carcinoma	4	106	
				2.50%					3.50%
Pairs withheld from model fitting					Pairs withheld from model fitting				
		Predicted					Predicted		
		Normal mucosa	Carcinoma				Normal mucosa	Carcinoma	
Actual	Normal mucosa	211	11		Actual	Normal mucosa	69	2	
	Carcinoma	4	201			Carcinoma	3	84	
				3.51%					3.16%

OOB, out-of-bag.

[a]Error rates for primary data sets are OOB estimates; error rates for secondary data sets are leave-one-out estimates.

Table 4 Logistic models and related PCA results

	Intercept	Colon				Intercept	Rectal		
		Comp. 1[a]	Comp. 2	Comp. 3			Comp. 1[a]	Comp. 3	Comp. 4
Logistic model									
Est. coefficient	1.19	– 1.40	– 1.09	– 1.95	Est. coefficient	3.07	– 2.90	– 1.32	0.95
S.e.	0.46	0.27	0.28	0.48	S.e.	0.83	0.54	0.46	0.43
P-value	0.01	< 0.001	< 0.001	< 0.001	*P*-value	< 0.001	< 0.001	< 0.01	0.03
Corresponding PCA results[b]									
% of sample var.		38.24	24.86	13.30	% of Sample var.		44.66	8.67	6.56
Included miRNA		**Loadings**			**Included miRNA**		**Loadings**		
Hsa-miR-663b		– 0.08	– 0.07	– 0.62	Hsa-miR-663b		– 0.23	– 0.36	– 0.30
Hsa-miR-4539		0.26	0.25	– 0.23	Hsa-miR-17-5p		– 0.31	– 0.08	0.07
Hsa-miR-17-5p		– 0.38	0.09	– 0.09	Hsa-miR-4323		0.24	– 0.40	0.30
Hsa-miR-20a-5p		– 0.37	0.10	– 0.08	Hsa-miR-20a-5p		– 0.30	– 0.08	0.10
Hsa-miR-663a		0.04	0.01	– 0.62	Hsa-miR-378a-3p		0.20	– 0.12	– 0.39
Hsa-miR-4538		0.26	0.26	– 0.16	Hsa-miR-93-5p		– 0.30	– 0.04	– 0.01
Hsa-miR-21-5p		– 0.33	0.14	0.00	Hsa-miR-4539		0.28	– 0.30	– 0.01
Hsa-miR-215		– 0.08	0.42	0.18	Hsa-miR-378i		0.19	– 0.12	– 0.39
Hsa-miR-4506		– 0.20	– 0.22	0.04	Hsa-miR-21-5p		– 0.26	0.12	0.35
Hsa-miR-92a-3p		– 0.33	0.13	– 0.12	Hsa-miR-92a-3p		– 0.29	– 0.12	0.13
Hsa-miR-93-5p		– 0.36	0.11	– 0.02	Hsa-miR-145-5p		0.10	0.29	0.07
Hsa-miR-192-5p		– 0.10	0.42	0.17	Hsa-miR-150-5p		0.15	0.21	– 0.14
Hsa-miR-145-5p		– 0.02	0.26	0.18	Hsa-miR-3651		– 0.26	– 0.15	0.09
Hsa-miR-3651		– 0.33	0.11	– 0.12	Hsa-miR-4749-3p		0.25	– 0.30	0.31
Hsa-miR-378a-3p		0.17	0.41	– 0.08	Hsa-miR-4506		– 0.24	0.27	– 0.39
Hsa-miR-378i		0.17	0.39	– 0.09	Hsa-miR-424-3p		– 0.24	– 0.22	–0.05
—		—	—	—	Hsa-miR-6073		0.15	0.42	0.27

AIC, Akaike information criterion; miRNA, microRNA; PCA, principal component analysis.

[a]The components listed are those included in the model; model selection was based on AIC statistics.

[b]Loadings and % of sample variance refer to the component listed in the corresponding column.

Table 5 miRNA contribution to enriched biological processes and pathways[a]

miRNA	No. of targets[b]	Enriched biological processes[c]	Enriched pathways[c]
Rectal and colon tissue			
Hsa-miR-17-5p	5	Phosphate metabolic process, angiogenesis, intracellular signaling cascade, blood vessel and vasculature development, interphase of mitotic cell cycle, blood vessel morphogenesis	Melanoma, chronic myeloid leukemia, small-cell lung cancer, glioma, prostate cancer, pathways in cancer, p53 signaling pathway
Hsa-miR-20a-5p	8	Phosphate metabolic process, angiogenesis, intracellular signaling cascade, blood vessel and vasculature development, positive regulation of angiogenesis, negative regulation of cell differentiation, interphase of mitotic cell cycle, blood vessel morphogenesis, positive regulation of transcription from RNA polymerase II promoter	Melanoma, pancreatic cancer, chronic myeloid leukemia, small-cell lung cancer, glioma, prostate cancer, pathways in cancer, p53 signaling pathway
Hsa-miR-378a-3p	1		Pathways in cancer
Hsa-miR-21-5p	289	Phosphate metabolic process, angiogenesis, intracellular signaling cascade, blood vessel and vasculature development, positive regulation of angiogenesis, negative regulation of cell differentiation, interphase of mitotic cell cycle, blood vessel morphogenesis, positive regulation of transcription from RNA polymerase II promoter	Melanoma, pancreatic cancer, chronic myeloid leukemia, small-cell lung cancer, glioma, prostate cancer, pathways in cancer, p53 signaling pathway
Hsa-miR-92a-3p	3		
Hsa-miR-145-5p	10	Phosphate metabolic process, intracellular signaling cascade, interphase of mitotic cell cycle, negative regulation of cell differentiation, positive regulation of transcription from RNA polymerase II promoter	
Hsa-miR-3651	24		
Colon tissue only			
Hsa-miR-663a	1	Blood vessel and vasculature development, blood vessel morphogenesis, positive regulation of transcription from RNA polymerase II promoter	
Hsa-miR-92-5p	346	Phosphate metabolic process, angiogenesis, intracellular signaling cascade, blood vessel and vasculature development, positive regulation of angiogenesis, negative regulation of cell differentiation, interphase of mitotic cell cycle, blood vessel morphogenesis, positive regulation of transcription from RNA polymerase II promoter	Melanoma, pancreatic cancer, chronic myeloid leukemia, small-cell lung cancer, glioma, prostate cancer, pathways in cancer
Rectal tissue only			
Hsa-miR-150-5p	1		

FDR, false discovery rate; miRNA, microRNA.

[a] All enriched processes and pathways were identified using DAVID (https://david.ncifcrf.gov/home.jsp).

[b] Targets are from TarBase, using filters "Homo sapiens".

[c] Processes and pathways selected using <0.05 FDR.

leave-one-out estimated error rate associated with the logistic model was 4.68%. The third round of analysis considering rectal observations identified 35 miRNAs as important and the random forest model had an OOB error estimate of 6.25%, whereas the leave-one-out estimated error associated with the logistic model was 8.43%. That is, we found several disjoint subsets of important miRNAs from which decent misclassification rates could be achieved. However, we found that as miRNAs previously identified as important were removed from consideration, model classification suffered and the differences in the importance rank between miRNAs ranked highest and those ranked lower diminished. In general, several subsets of miRNAs were capable of discriminating between carcinoma tissue and normal mucosa based on expression; however, as those miRNAs initially identified as most important were discarded from analysis, a greater number of miRNAs were required and lower model accuracy was observed. The miRNAs identified for the colon and rectal data sets during the secondary and tertiary analysis runs as well as the associated confusion matrices are included in the online Supplementary Tables S4–S7.

Our study has several strengths including a large sample size that enabled us to reproduce findings using alternative models and methods among observations independent of those considered in the primary analysis. Additionally, qPCR allowed us to validate data characteristics with an alternative expression measurement platform for several miRNAs identified as important. However, all of our analysis restricted attention to frequently expressed miRNAs, and other less frequently expressed miRNAs could also be important for subsets of the population. Although we used an Agilent microarray platform and validated data characteristics among several key miRNAs using qPCR, other platforms and methods of assessment could have produced additional and/or alternative results in terms of which miRNA expressions are deemed important, as concordance between results from different expression measurement platforms can vary.[48] We used normal mucosa adjacent to the tumor. While this was the only option to obtain non-tumor colonic tissue, it has been shown that adjacent tissue may also have genetic alterations.[49] However, we were able to identify unique miRNA patterns between the normal colonic mucosa and carcinoma tissue.

Table 6 Agilent platform measurements compared with qPCR measurements

Important miRNA	Fold change[a]		Correlation between platforms	
	Agilent platform	qPCR	Normal obs.	Tumor obs.
Colon				
Hsa-miR-663b	2.16	2.23	0.58	0.57
Hsa-miR-17-5p	2.75	3.05	0.88	0.83
Hsa-miR-20a-5p	2.91	3.15	0.87	0.83
Hsa-miR-4538	0.77	0.14	0.01	0.34
Hsa-miR-21-5p	2.70	3.15	0.71	0.65
Hsa-miR-215	0.55	0.29	0.59	0.39
Hsa-miR-4506	1.34	1.22	0.06	0.38
Hsa-miR-92a-3p	2.56	2.59	0.82	0.72
Hsa-miR-93-5p	2.09	2.17	0.85	0.73
Hsa-miR-3651	2.37	2.25	0.70	0.65
Rectal				
Hsa-miR-663b	2.09	2.78	-0.08	0.39
Hsa-miR-17-5p	3.13	4.22	0.78	0.91
Hsa-miR-20a-5p	3.34	4.45	0.82	0.93
Hsa-miR-93-5p	2.25	2.79	0.87	0.70
Hsa-miR-21-5p	3.09	4.07	0.55	0.95
Hsa-miR-92a-3p	2.33	2.85	0.87	0.76
Hsa-miR-3651	2.24	2.56	0.67	0.80
Hsa-miR-4506	1.46	1.11	-0.14	0.07

qPCR, quantitative PCR.
[a] Calculated as mean tumor/mean normal.

Results from this study are an important step to understanding the clinical relevance of miRNAs and their potential use for screening, early detection, and therapeutic modalities. In our previous work, we identified over 500 miRNAs that were differentially expressed between carcinoma and normal mucosa.[10] However, given the number of dysregulated miRNAs, our aim in this study was to determine if a smaller subset of miRNAs could be identified that could then be explored for their clinical relevance. Based on the analysis of all known miRNAs, we identified a small group that could accurately distinguish between normal colonic mucosa and carcinoma tissue. This group of miRNAs represents a group of miRNAs that when considered together helps to define colorectal carcinoma. Although some of these miRNAs have been studied extensively, we know little about the pathways and functions of other miRNAs in this group. Having a well-defined subset of miRNAs associated with colorectal carcinoma tissue allows for more focused studies regarding specific gene targets and subsequent pathways. While we examined KEGG pathways to help identify relevant pathways for these miRNAs, there are limitations in the current knowledgebase. Some of the miRNAs that have been studied extensively, such as miR21, have been linked to many genes and disease pathways. It is unclear which of these genes and pathways are most relevant for CRC. Other miRNAs that have not been studied extensively have been linked only to a few genes and a few pathways. Having a group of miRNAs specifically associated with CRC will guide research to better defined important pathways that will hopefully provide the basis for developing CRC-specific therapeutics.

Using a set of miRNAs to delineate a CRC-specific pathway may be especially important as new screening methods emerge, including the use of liquid biopsies. Nonspecificity of few well-studied miRNAs presents a problem for such screening test, whereas a set of miRNAs specific to CRC could prove to be useful for that purpose. The metastatic potential associated with this set of miRNAs is currently not known. However, further evaluation of the ability of this set of miRNAs to predict adenomas that have a greater potential for carcinoma development could enhance our ability to identify individuals who may require different screening guidelines. The clinical relevance of this subset of miRNAs, although promising, has yet to be fully understood. This study helps describe the landscape of miRNAs as they relate to CRC.

CONFLICT OF INTEREST

Guarantor of the article: Martha L. Slattery, PhD.
Specific author contributions: DFP conducted statistical analysis and wrote the manuscript. JRS provide overall guidance to statistical analysis. WS reviewed slides for identification of carcinoma and normal tissue. LEM coordinated bioinformatics analysis and assisted with data interpretation. RKW provided oversight to laboratory analysis. MLS obtained funding and oversaw study and helped edit final manuscript. All authors read and approved the manuscript.
Financial support: This study was supported by NCI grants CA163683, CA48998, and CA61757 from the National Cancer Institute at the National Institutes of Health.
Potential competing interests: The authors declare no conflict of interest.

Acknowledgments. The contents of this manuscript are solely the responsibility of the authors and do not necessarily represent the official view of the National Cancer Institute. We acknowledge Sandra Edwards for data oversight and Dr Caan and the staff at the Kaiser Permanente Medical Care Program for their contribution to data collection and to Dr Corcoran at Utah State University for his input into the statistical analysis. We also acknowledge Michael D. Hoffman and Erica Wolff for laboratory analysis.

Study Highlights

WHAT IS CURRENT KNOWLEDGE

✓ MiRNA are noncoding RNA molecules that alter gene activity.

✓ Many miRNAs are dysregulated in CRC tissue.

WHAT IS NEW HERE

✓ 16 miRNAs for colon and 17 miRNAs for rectal carcinoma differentiated between carcinoma and normal mucosa.

✓ 12 miRNAs were important discriminators of carcinoma and normal mucosa for colon and rectal carcinoma.

✓ MiRNA expression levels were validated with qPCR.

1. Arora S, Rana R, Chhabra A *et al.* miRNA-transcription factor interactions: a combinatorial regulation of gene expression. *Mol Genet Genomics* 2013;**288**:77–87.
2. Miska EA. How microRNAs control cell division, differentiation and death. *Curr Opin Genet Dev* 2005; **15**: 563–568.
3. Iorio MV, Croce CM. MicroRNAs in cancer: small molecules with a huge impact. *J Clin Oncol* 2009; **27**: 5848–5856.
4. Dong Y, Wu WK, Wu CW *et al.* MicroRNA dysregulation in colorectal cancer: a clinical perspective. *Br J Cancer* 2011;**104**:893–898.
5. Luo X, Burwinkel B, Tao S *et al.* MicroRNA signatures: novel biomarker for colorectal cancer? *Cancer Epidemiol Biomarkers Prev* 2011;**20**:1272–1286.
6. Polytarchou C, Hommes DW, Palumbo T *et al.* MicroRNA214 is associated with progression of ulcerative colitis, and inhibition reduces development of colitis and colitis-associated cancer in mice. *Gastroenterology* 2015;**149**:981–92.e11.
7. Hamfjord J, Stangeland AM, Hughes T *et al.* Differential expression of miRNAs in colorectal cancer: comparison of paired tumor tissue and adjacent normal mucosa using high-throughput sequencing. *PLoS One* 2012;**7**:e34150.
8. Piepoli A, Tavano F, Copetti M *et al.* Mirna expression profiles identify drivers in colorectal and pancreatic cancers. *PLoS One* 2012;**7**:e33663.
9. Reid JF, Sokolova V, Zoni E *et al.* miRNA profiling in colorectal cancer highlights miR-1 involvement in MET-dependent proliferation. *Mol Cancer Res* 2012;**10**:504–515.
10. Slattery ML, Herrick JS, Pellatt DF *et al.* MicroRNA profiles in colorectal carcinomas, adenomas, and normal colonic mucosa: variations in miRNA expression and disease progression. *Carcinogenesis* 2016.
11. Svetnik V, Liaw A, Tong C *et al.* Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci* 2003;**43**:1947–1958.
12. Breiman L. Random forests. *Machine Learn* 2001; **45**: 5–32.
13. Archer KJ, Kimes RV. Empirical characterization of random forest variable importance measures. *Comput Stat Data Anal* 2008; **52**: 2249–2260.
14. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. *BMC Bioinform* 2006;**7**:1–13.
15. Cutler A, Stevens JR. Random forests for microarrays. *Methods Enzymol* 2006; **411**: 422–432.
16. Slattery ML, Potter J, Caan B *et al.* Energy balance and colon cancer—beyond physical activity. *Cancer Res* 1997;**57**:75–80.
17. Slattery ML, Caan BJ, Benson J *et al.* Energy balance and rectal cancer: an evaluation of energy intake, energy expenditure, and body mass index. *Nutr Cancer* 2003;**46**:166–171.
18. Slattery ML, Edwards SL, Palmer L *et al.* Use of archival tissue in epidemiologic studies: collection procedures and assessment of potential sources of bias. *Mutat Res* 2000;**432**: 7–14.
19. Griffiths-Jones S. miRBase: the microRNA sequence database. *Methods Mol Biol* 2006; **342**: 129–138.
20. Core Team R. *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2015. Available at: http://www.R-project.org/; last accessed on 20 January 2016.
21. Liaw A, Wiener M. Classification and regression by random forest. *R News* 2002; **2**: 18–22.
22. Breiman L, Friedman J, Stone CJ *et al. Classification and Regression Trees*. Boca Raton, FL: CRC Press, 1984.
23. Vlachos IS, Paraskevopoulou MD, Karagkouni D *et al.* DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res* 2015;**43**:D153–D159.
24. Huang, da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009; **37**: 1–13.
25. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protocols* 2008; **4**: 44–57.
26. Ashburner M, Ball CA, Blake JA *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–29.
27. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000; **28**: 27–30.
28. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2 – ΔΔCT method. *Methods* 2001; **25**: 402–408.
29. Schmittgen TD, Livak KJ. Analyzing real-time PCR data by the comparative CT method. *Nat Protocols* 2008; **3**: 1101–1108.
30. Asangani IA, Rasheed SA, Nikolova DA *et al.* MicroRNA-21 (miR-21) post-transcriptionally downregulates tumor suppressor Pdcd4 and stimulates invasion, intravasation and metastasis in colorectal cancer. *Oncogene* 2008;**27**:2128–2136.
31. Baffa R, Fassan M, Volinia S *et al.* MicroRNA expression profiling of human metastatic cancers identifies cancer gene targets. *J Pathol* 2009;**219**:214–221.
32. Schetter AJ, Leung SY, Sohn JJ *et al.* MicroRNA expression profiles associated with prognosis and therapeutic outcome in colon adenocarcinoma. *JAMA* 2008;**299**:425–436.
33. Valeri N, Gasparini P, Braconi C *et al.* MicroRNA-21 induces resistance to 5-fluorouracil by down-regulating human DNA MutS homolog 2 (hMSH2). *Proc Natl Acad Sci USA* 2010;**107**: 21098–21103.
34. Xia X, Yang B, Zhai X *et al.* Prognostic role of microRNA-21 in colorectal cancer: a meta-analysis. *PLoS One* 2013;**8**:e80426.
35. Du M, Shi D, Yuan L *et al.* Circulating miR-497 and miR-663b in plasma are potential novel biomarkers for bladder cancer. *Sci Rep* 2015;**5**:10437.
36. Ragusa M, Statello L, Maugeri M *et al.* Specific alterations of the microRNA transcriptome and global network structure in colorectal cancer after treatment with MAPK/ERK inhibitors. *J Mol Med (Berl)* 2012;**90**:1421–1438.
37. Kara M, Yumrutas O, Ozcan O *et al.* Differential expressions of cancer-associated genes and their regulatory miRNAs in colorectal carcinoma. *Gene* 2015;**567**:81–86.
38. Fang L, Li H, Wang L *et al.* MicroRNA-17-5p promotes chemotherapeutic drug resistance and tumour metastasis of colorectal cancer by repressing PTEN expression. *Oncotarget* 2014;**5**:2974–2987.
39. Okumura T, Shimada Y, Omura T *et al.* MicroRNA profiles to predict postoperative prognosis in patients with small cell carcinoma of the esophagus. *Anticancer Res* 2015;**35**:719–727.
40. Motoyama K, Inoue H, Takatsuno Y *et al.* Over- and under-expressed microRNAs in human colorectal cancer. *Int J Oncol* 2009;**34**:1069–1075.
41. Nishida N, Nagahara M, Sato T *et al.* Microarray analysis of colorectal cancer stromal tissue reveals upregulation of two oncogenic miRNA clusters. *Clin Cancer Res* 2012;**18**: 3054–3070.
42. Schee K, Lorenz S, Worren MM *et al.* Deep sequencing the microRNA transcriptome in colorectal cancer. *PLoS One* 2013;**8**:e66165.
43. Bandres E, Cubedo E, Agirre X *et al.* Identification by real-time PCR of 13 mature microRNAs differentially expressed in colorectal cancer and non-tumoral tissues. *Mol Cancer* 2006;**5**:29.
44. Izzotti A, Calin GA, Steele VE *et al.* Chemoprevention of cigarette smoke-induced alterations of microRNA expression in rat lungs. *Cancer Prev Res (Phila)* 2010;**3**:62–72.
45. Kheirelseid EA, Miller N, Chang KH *et al.* mRNA/miRNA correlations in colorectal cancer: novel mechanisms in cancer initiation and progression. *Int J Colorectal Dis* 2013;**28**: 1031–1034.
46. Nakano H, Miyazawa T, Kinoshita K *et al.* Functional screening identifies a microRNA, miR-491 that induces apoptosis by targeting Bcl-X(L) in colorectal cancer cells. *Int J Cancer* 2009;**127**:1072–1080.
47. Wu H, Neilson JR, Kumar P *et al.* miRNA profiling of naive, effector and memory CD8 T cells. *PLoS One* 2007;**2**:e1020.
48. Mestdagh P, Hartmann N, Baeriswyl L *et al.* Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. *Nat Methods* 2014;**11**:809–815.
49. Sanz-Pamplona R, Berenguer A, Cordero D *et al.* Aberrant gene expression in mucosa adjacent to tumor reveals a molecular crosstalk in colon cancer. *Mol Cancer* 2014;**13**:46.

Supplementary Information accompanies this paper on the Clinical and Translational Gastroenterology website (http://www.nature.com/ctg)