An Assessment of Database-Validated microRNA Target Genes in Normal Colonic Mucosa: Implications for Pathway Analysis

Cancer Informatics
Volume 16: 1–11
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1176935117716405

Martha L Slattery[1], Jennifer S Herrick[1], John R Stevens[2], Roger K Wolff[1] and Lila E Mullany[1]

[1]Department of Internal Medicine, The University of Utah, Salt Lake City, UT, USA. [2]Department of Mathematics & Statistics, Utah State University, Logan, UT, USA.

ABSTRACT

BACKGROUND: Determination of functional pathways regulated by microRNAs (miRNAs), while an essential step in developing therapeutics, is challenging. Some miRNAs have been studied extensively; others have limited information. In this study, we focus on 254 miRNAs previously identified as being associated with colorectal cancer and their database-identified validated target genes.

METHODS: We use RNA-Seq data to evaluate messenger RNA (mRNA) expression for 157 subjects who also had miRNA expression data. In the replication phase of the study, we replicated associations between 254 miRNAs associated with colorectal cancer and mRNA expression of database-identified target genes in normal colonic mucosa. In the discovery phase of the study, we evaluated expression of 18 miRNAs (those with 20 or fewer database-identified target genes along with miR-21-5p, miR-215-5p, and miR-124-3p which have more than 500 database-identified target genes) with expression of 17 434 mRNAs to identify new targets in colon tissue. Seed region matches between miRNA and newly identified targeted mRNA were used to help determine direct miRNA-mRNA associations.

RESULTS: From the replication of the 121 miRNAs that had at least 1 database-identified target gene using mRNA expression methods, 97.9% were expressed in normal colonic mucosa. Of the 8622 target miRNA-mRNA associations identified in the database, 2658 (30.2%) were associated with gene expression in normal colonic mucosa after adjusting for multiple comparisons. Of the 133 miRNAs with database-identified target genes by non-mRNA expression methods, 97.2% were expressed in normal colonic mucosa. After adjustment for multiple comparisons, 2416 miRNA-mRNA associations remained significant (19.8%). Results from the discovery phase based on detailed examination of 18 miRNAs identified more than 80 000 miRNA-mRNA associations that had not previously linked to the miRNA. Of these miRNA-mRNA associations, 15.6% and 14.8% had seed matches for CRCh38 and CRCh37, respectively.

CONCLUSIONS: Our data suggest that miRNA target gene databases are incomplete; pathways derived from these databases have similar deficiencies. Although we know a lot about several miRNAs, little is known about other miRNAs in terms of their targeted genes. We encourage others to use their data to continue to further identify and validate miRNA-targeted genes.

KEYWORDS: Colon cancer, miRNA, functionality, bias, RNA-Seq

RECEIVED: March 2, 2017. **ACCEPTED:** May 28, 2017.

PEER REVIEW: Three peer reviewers contributed to the peer review report. Reviewers' reports totaled 885 words, excluding any confidential comments to the academic editor.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by NCI grants CA163683 and CA48998 from the National Cancer Institute at the National Institutes of Health.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Martha L Slattery, Department of Internal Medicine, The University of Utah, 383 Colorow, Salt Lake City, UT 84108, USA.
Email: Marty.Slattery@hsc.utah.edu

Background

MicroRNAs (miRNAs) regulate gene expression by either repressing translation of messenger RNAs (mRNAs) or by inducing mRNA degradation by complementarity binding of target sequences.[1] Given their role in gene expression, it is not surprising that miRNAs have been shown to be associated with many physiological processes and have been linked to numerous diseases.[2–10] As such, miRNAs have the potential of being valuable biomarkers for both early disease detection and prognosis as well as serving as targets for therapeutic purposes. An understanding of the functionality of miRNAs and how they relate to various biological pathways is essential to move the field of miRNA research from associations with diseases and conditions to therapeutic and interventional tools.

Determination of functional pathways regulated by miRNAs, while an essential step in developing therapeutics centering on miRNA expression, is challenging. Functional pathway analysis is largely dependent on databases such as miRTarBase which have identified miRNA target genes that have been validated by a variety of methods.[11] MicroRNA target interactions (MTIs) incorporated into the database have varying degrees of supporting evidence. The minority of MTIs incorporated into the database have strong evidence and include reporter assays and Western blot methods. Most of the MTIs in the database come from what is considered less stringent methods of target gene identification such as microarray and next-generation sequencing methods, one of which is cross-linking and immunoprecipitation sequencing or cross-linking immunoprecipitation (CLIP)-Seq. Although quantitative polymerase chain reaction (qPCR), microarray, RNA-Seq, and Northern blot methods detect MTI associations through gene expression, other experiments such as reporter assay and Western blot studies measure expression levels of proteins. Comparisons of miRNA expression and mRNA expression patterns have been used as methods to identify target genes.[11–13]

Although the study of miRNAs as they relate to disease is a rapidly expanding field of research, databases that contain information on miRNA-validated targeted genes are dependent on what is known in the literature.

It is not surprising that some miRNAs have been studied extensively, whereas others have limited information. The extent to which information bias regarding miRNA-targeted genes influences our ability to accurately infer functional pathways associated with miRNAs is unexplored. In this study, we examine 254 miRNAs that we have previously identified as being associated with colorectal cancer, colorectal cancer survival, or the microsatellite instability (MSI) tumor phenotype[2,3,14,15] with mRNA expression. We compare expression levels between mRNA for previously identified target genes and the miRNA expression level. We analyze separately associations for those target genes previously identified by mRNA expression methods versus those identified by other methods. For a smaller subset of miRNAs with 20 or fewer reported target genes and for 3 miRNAs with hundreds of database-identified target genes, we compare miRNA expression level with all mRNA expression of protein-coding genes in colon tissue. We further compare seed region for those newly discovered mRNAs associated with miRNAs to help identify target genes that may be directly influenced by miRNAs. Figure 1 depicts the study flow.

Methods

Study participants were part of a colon cancer case-control study that included incident first primary adenocarcinoma of the colon who were diagnosed between 30 and 79 years of age and resided in Utah or were members of the Kaiser Permanente Medical Care Program (KPMCP) in Northern California. Participants were non-Hispanic white, Hispanic, or African American.[16] Local Surveillance, Epidemiology, and End Results tumor registries verified all cases that were diagnosed between October 1991 and September 1994. Detailed study methods have been described.[3] Participants signed informed consent prior to release of confidential data. The Institutional Review Boards of the University of Utah and the KPMCP approved the study.

RNA Processing

Formalin-fixed paraffin-embedded tissue was used to extract RNA. Normal mucosae adjacent to the carcinoma tissue and matched carcinoma tissue were used to make RNA. Total RNA was extracted, isolated, and purified using the RecoverAll Total Nucleic Acid Isolation Kit (Ambion, Austin, Texas); RNA yields were determined using a NanoDrop spectrophotometer.

MicroRNA

The Agilent Human miRNA Microarray V19.0 was used. The microarray contains probes for 2006 unique human miRNAs as described previously. Data were required to pass stringent quality control (QC) 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. If samples failed to meet quality standards for any of these parameters, the sample was relabeled, hybridized to arrays, and rescanned. If a sample failed QC assessment a second time, the sample was deemed to be of poor quality and was excluded from analysis. Our previous analysis has shown that the repeatability associated with this microarray was extremely high ($r = 0.98$),[3] and that comparison of miRNA expression levels obtained from the Agilent microarray with those obtained from qPCR had an agreement of 100% in terms of directionality of findings and that the fold change calculated for the miRNA expression difference between carcinoma and normal colonic mucosa was almost identical.[2] Of the 2006 unique human miRNAs assessed on the Agilent microarray, 1226 were expressed in colon carcinoma tissue and 1179 in normal colon mucosa.

To normalize differences in miRNA expression that could be attributed to the array, amount of RNA, location on array, or factors that could erroneously influence miRNA expression levels, total gene signal was normalized by multiplying each sample by a scaling factor,[17] which was the median of the 75th percentiles of all the samples divided by the individual 75th percentile of each sample.

RNA-Seq Sequencing Library Preparation and Data Processing

Total RNA was run on 197 carcinoma and normal mucosa pairs; 157 of these passed QC. These samples were taken from the study subjects used for miRNA analysis and were extracted, isolated, and purified in the same manner as previously described.[18] RNA library construction was done with the Illumina TruSeq Stranded Total RNA Sample Preparation Kit with Ribo-Zero. The samples were then fragmented and primed for complementary DNA (cDNA) synthesis, adapters were then ligated onto the cDNA, and the resulting samples were then amplified using polymerase chain reaction (PCR); the amplified library was then purified using Agencourt AMPure XP beads. A more detailed description of the methods can be found in our previous work.[19]

Illumina TruSeq v3 single read flow cell and a 50-cycle single-read sequence run were performed on an Illumina HiSeq instrument. Reads were aligned to a sequence database containing the human genome (build GRCh37/hg19, February 2009 from genome.ucsc.edu). Python and a pysam module were used to calculate counts for each exon and untranslated region (UTR) of the genes using a list of gene coordinates

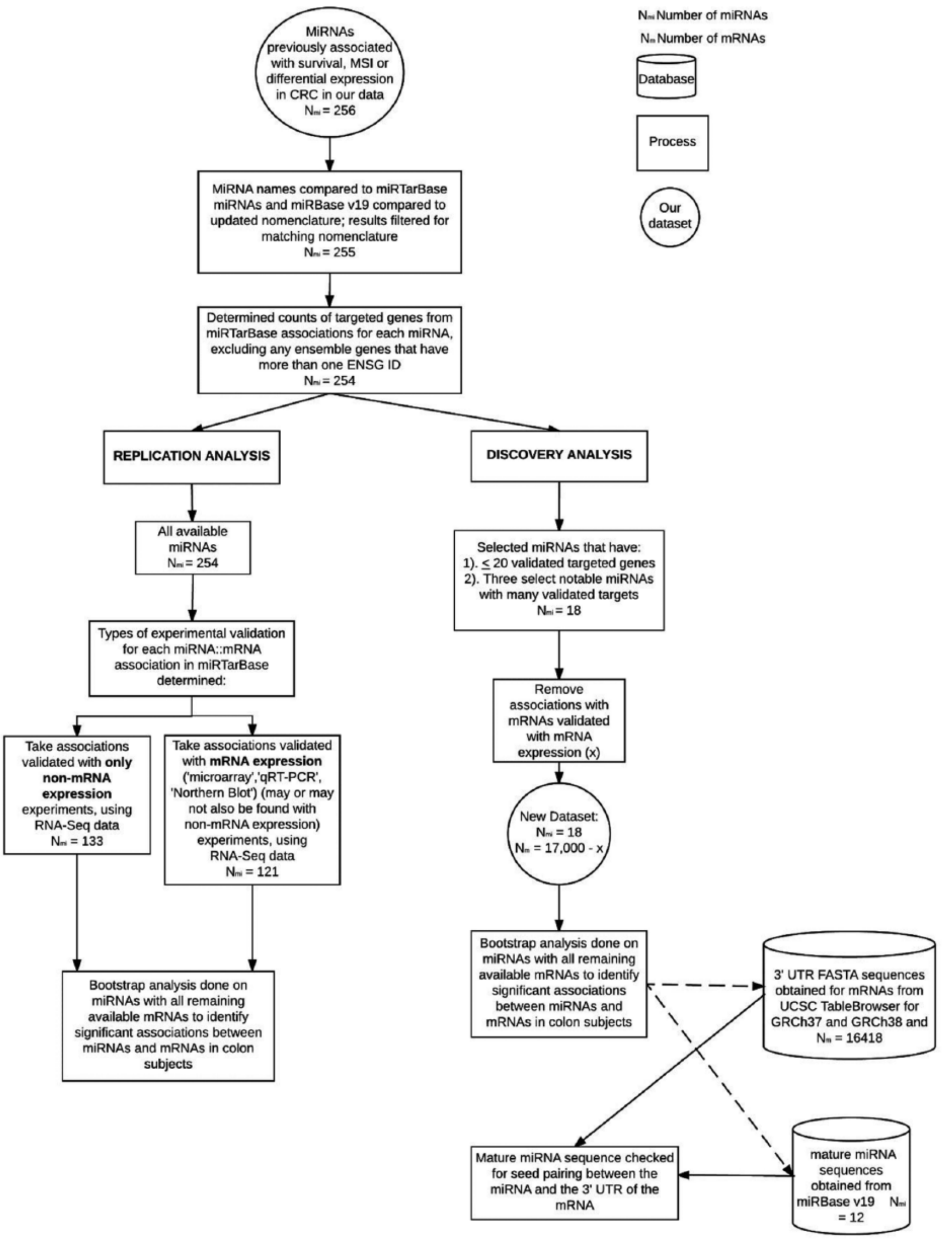
MiRNAs previously associated with survival, MSI or differential expression in CRC in our data N_{mi} = 256
MiRNA names compared to miRTarBase miRNAs and miRBase v19 compared to updated nomenclature; results filtered for matching nomenclature N_{mi} = 255
Determined counts of targeted genes from miRTarBase associations for each miRNA, excluding any ensemble genes that have more than one ENSG ID N_{mi} = 254
REPLICATION ANALYSIS
DISCOVERY ANALYSIS
All available miRNAs N_{mi} = 254
Types of experimental validation for each miRNA::mRNA association in miRTarBase determined:
Take associations validated with only non-mRNA expression experiments, using RNA-Seq data N_{mi} = 133
Take associations validated with mRNA expression ('microarray','qRT-PCR', 'Northern Blot') (may or may not also be found with non-mRNA expression) experiments, using RNA-Seq data N_{mi} = 121
Bootstrap analysis done on miRNAs with all remaining available mRNAs to identify significant associations between miRNAs and mRNAs in colon subjects
Selected miRNAs that have: 1). ≤ 20 validated targeted genes 2). Three select notable miRNAs with many validated targets N_{mi} = 18
Remove associations with mRNAs validated with mRNA expression (x)
New Dataset: N_{mi} = 18 N_m = 17,000 - x
Bootstrap analysis done on miRNAs with all remaining available mRNAs to identify significant associations between miRNAs and mRNAs in colon subjects
3' UTR FASTA sequences obtained for mRNAs from UCSC TableBrowser for GRCh37 and GRCh38 and N_m = 16418
Mature miRNA sequence checked for seed pairing between the miRNA and the 3' UTR of the mRNA
mature miRNA sequences obtained from miRBase v19 N_{mi} = 12
N_{mi} Number of miRNAs
N_m Number of mRNAs
Database
Process
Our dataset

Figure 1. Study flow.

obtained from http://genome.ucsc.edu. Total gene counts were determined. We dropped features that were not expressed in our data or for which the expression was missing for most of the samples.[19]

Bioinformatics Analysis

Our assessment of miRNAs focused on 254 miRNAs that we have previously reported as being associated with either differences in tumor and normal mucosa expression with a fold change of at least 1.5, survival, or with MSI tumor phenotype.[2,3,14] We identified experimentally validated target gene these miRNAs using miRNA-mRNA pairs from miRTarBase v6.0.[11] As our miRNA names are those from Agilent v19 (corresponding to miRBase v19) and miRTarBase v6.0 includes newer associations, we determined the new nomenclature for any miRNAs whose names did not match using archived miRBase (http://www.mirbase.org)[20] "miRNA.diff.zip" files.

To determine associations between miRNA-mRNA pairs in colon tissue and to help estimate the extent of completeness of existing databases, we conducted both a replication and discovery analysis.

For the replication component of the study, we determined how many and which target genes were identified using gene expression methods, namely, "microarray," RNA-Seq, "qRT-PCR," and "Northern blot" experiments (we refer to gene expression methods as "mRNA-methods") in miRTarBase and whether the miRNA-mRNA pairs could be replicated in normal colon tissue. Target genes for miRNAs identified by methods other than gene expression (which we refer to as "non-mRNA methods") were analyzed separately to determine how many of these could be identified by RNA-Seq, an mRNA expression method. Our hypothesis is that those target genes identified by mRNA methods should correlate more positively to our comparison of miRNA with target genes using RNA-Seq data.

For the discovery component of the study, we focused on miRNAs that had 20 or fewer validated targets in miRTarBase as well as 3 commonly validated miRNAs (hsa-miR-21-5p, hsa-miR-124-3p, and hsa-miR-215). We analyzed these miRNAs with mRNA expression generated by RNA-Seq in our data set excluding those validated target genes that had previously been identified in miRTarBase (see section "Statistical Methods"). We further analyzed miRNAs and targeted mRNAs for seed region matches, and we analyzed the mRNA 3′ UTR FASTA as well as the seed region sequence of the associated miRNA to determine seed region pairings between miRNA and mRNA. MicroRNA seed regions were calculated as described in our previous work,[21] and we calculated and included seeds of 6, 7, and 8 nucleotides in length. Our hypothesis is that a seed match would increase the likelihood that identified genes associated with a specific miRNA were more likely to have a direct association given a higher propensity for binding. As miRTarBase uses findings from many different investigations spanning across years and alignments, we used FASTA sequences generated from both GRCh37 and GRCh38 *Homo sapiens* alignments, using UCSC Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables).[22] We downloaded FASTA sequences that matched our Ensembl IDs and had a consensus coding sequences available. Analysis was done using scripts in R 3.2.3 and in perl 5.018002.

Statistical Methods

Our final analysis consisted of 157 subjects with high-quality mRNA expression data and high-quality miRNA data. After excluding from the analysis any non–protein-coding mRNAs and those with fewer than 0.5 reads on average across all samples, we were able to examine 17 434 mRNAs that had unique Ensembl IDs. We used the log base 2–transformed RPKM (reads per kilobase per million) normal colonic mucosa mRNA expression data.

We examined the association between the mRNAs and the candidate miRNAs by fitting a linear model and adjusting for age at diagnosis, study center, and sex. *P* values were generated using the bootstrap method by creating a distribution of 10 000 F statistics derived by resampling the residuals from the null hypothesis model of no association between the mRNAs and miRNAs using the boot package in R. Associations were considered significant if the false discovery rate (FDR)–adjusted *P* values were less than .05.[23]

Table 1. Description of the study population.

		NO. (%)
Age	Mean (SD)	65.1 (10.2)
Sex	Male	88 (56.1)
	Female	69 (44.0)
AJCC stage	1	36 (23.1)
	2	51 (32.7)
	3	50 (32.1)
	4	19 (12.2)
Tumor instability	MSS	128 (81.5)
	MSI	29 (18.5)
TP53	Nonmutated	90 (57.3)
	Mutated	67 (42.7)
KRAS	Nonmutated	113 (72.0)
	Mutated	44 (28.0)
CIMP	Low	114 (72.6)
	High	43 (27.4)

Abbreviations: AJCC, American Joint Committee on Cancer; CIMP, CpG island methylator phenotype; MSI, microsatellite instability; MSS, microsatellite stable.

Results

Most of the study population was men, and the average age of study participants was 65.1 years (Table 1). The major molecular tumor phenotypes for the colon cancer study participants were 18.5% MSI, 42.7% with a mutated *TP53* gene, 28% with a mutation in *KRAS* gene, and 27.4% with a CpG island methylator phenotype high tumor. The distribution of targeted genes per miRNA shows that 117 of the 254 miRNAs we were able to evaluate had fewer than 100 target genes, 60 miRNAs had 100 to 200 identified target genes, 51 miRNAs had between 200 and 500 targeted genes, and 26 miRNAs had more than 500 database-identified target genes (Figure 2).

Replication phase

Of the 121 miRNAs that had at least 1 targeted gene identified by mRNA gene expression methods, 97.9% were expressed in normal colonic mucosa (Table 2 shows 50 of the 121 miRNAs; Supplemental Table 1 shows all 121 miRNAs). Of the 8622 target miRNA-mRNA associations identified in the database, 2658 (30.2%) were associated with gene expression in normal colonic mucosa after adjusting for multiple comparisons; prior to adjustment for multiple comparisons, 42.9% of database-identified associations were detected. For these

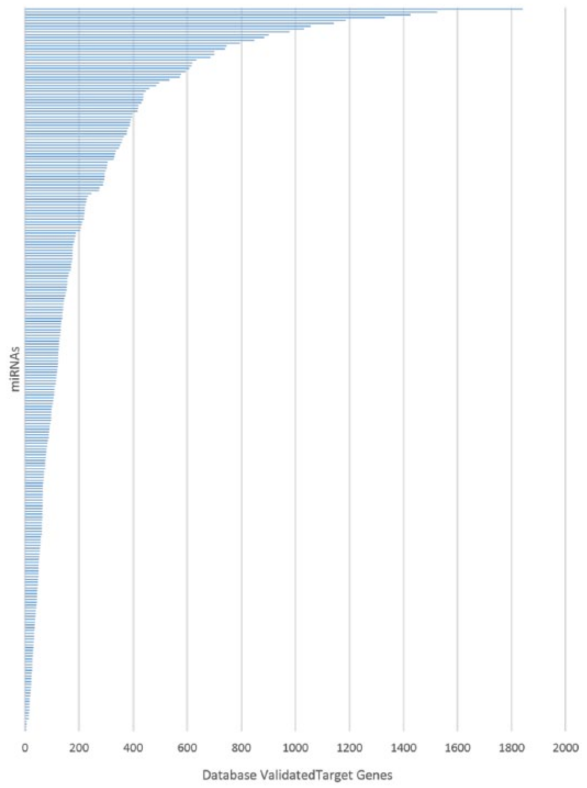

Figure 2. Distribution of database-validated target genes by individual microRNAs (miRNAs).

miRNAs that had at least 1 targeted gene identified by a gene expression method, 31 399 targeted genes were database validated by non-mRNA methods; of these, 98.3% were expressed in normal colonic mucosa. Of the database-identified target genes identified by non-mRNA expression methods, 48.0% were associated with the targeted gene in our data prior to adjustment for multiple comparisons and 37.6% were significant when an FDR of 0.05 was applied.

Evaluation of the 133 miRNAs with database-identified target genes only by non-mRNA expression methods showed that 11 850 of the 12 191 target genes were expressed in normal colonic mucosa (97.2%) (Table 3 shows 50 of the miRNAs; Supplemental Table 2 shows all 133 miRNAs). Of those expressed in normal colonic mucosa, 3770 (30.9%) miRNA-mRNA were associated in normal colonic mucosa using RNA-Seq data prior to adjustment for multiple comparisons. After adjustment for multiple comparisons, 2416 miRNA-mRNA associations remained significant (19.8%); this compares with 30.2% of miRNA:mRNA associations being detected when the original detection was a gene expression method.

Discovery phase

Examination of those miRNAs that had fewer than 20 database-identified target genes along with miR-21-5p, miR-215-5p,

Table 2. Comparison of miRNA-validated targeted genes with mRNA expression to colon RNA-Seq data.

MIRNA	MRNA DATABASE–IDENTIFIED TARGET GENES				NON-MRNA DATABASE–IDENTIFIED TARGET GENES			
	TOTAL, N	EXPRESSED IN COLON TISSUE, N	VALIDATED BY RNA-SEQ $P_{UNADJ} < .05$	VALIDATED BY RNA-SEQ FDR < 0.05, N	TOTAL, N	EXPRESSED IN COLON TISSUE, N	VALIDATED BY RNA-SEQ $P_{UNADJ} < .05$, N	VALIDATED BY RNA-SEQ FDR < 0.05, N
Total targets	8808	8622	3782	2658	31 399	30 855	15 080	11 795
hsa-let-7a-5p	37	36	17	10	528	522	341	257
hsa-let-7e-5p	12	12	4	1	519	511	165	87
hsa-let-7f-5p	20	20	11	9	325	318	208	173
hsa-let-7g-5p	17	17	11	8	274	269	198	163
hsa-let-7i-5p	5	4	3	3	278	273	242	226
hsa-miR-1-3p	419	411	38	9	424	421	28	7
hsa-miR-10a-5p	11	11	6	5	408	400	229	179
hsa-miR-15a-5p	79	78	37	14	573	566	179	56
hsa-miR-16-5p	108	108	96	89	1339	1320	1159	1077
hsa-miR-17-5p	40	40	27	22	1050	1034	772	661
hsa-miR-19b-3p	12	12	3	1	658	650	96	19
hsa-miR-20a-5p	33	33	23	16	953	939	674	537
hsa-miR-20b-5p	11	11	6	5	838	824	617	488

(Continued)

Table 2. (Continued)

MIRNA	MRNA DATABASE–IDENTIFIED TARGET GENES				NON-MRNA DATABASE–IDENTIFIED TARGET GENES			
	TOTAL, N	EXPRESSED IN COLON TISSUE, N	VALIDATED BY RNA-SEQ $P_{UNADJ} < .05$	VALIDATED BY RNA-SEQ FDR < 0.05, N	TOTAL, N	EXPRESSED IN COLON TISSUE, N	VALIDATED BY RNA-SEQ $P_{UNADJ} < .05$, N	VALIDATED BY RNA-SEQ FDR < 0.05, N
hsa-miR-21-3p	3	3	3	2	62	61	38	35
hsa-miR-21-5p	438	434	365	319	120	118	93	79
hsa-miR-23a-3p	23	21	16	13	184	181	146	122
hsa-miR-23b-3p	13	12	5	3	282	278	155	102
hsa-miR-24-3p	259	255	217	194	509	492	415	397
hsa-miR-25-3p	16	16	12	12	401	394	337	312
hsa-miR-26a-5p	26	25	19	17	376	374	298	261
hsa-miR-26b-5p	1465	1383	267	91	267	266	61	22
hsa-miR-27a-3p	38	37	30	23	336	328	250	202
hsa-miR-27b-3p	25	24	11	6	345	336	185	135
hsa-miR-28-5p	2	2	0	0	69	68	56	50
hsa-miR-29a-3p	64	64	47	40	146	140	107	97
hsa-miR-29b-3p	68	68	24	12	150	144	68	50
hsa-miR-29c-3p	40	40	10	6	172	167	53	34
hsa-miR-30a-5p	15	15	1	1	658	653	55	15
hsa-miR-30b-5p	16	16	5	1	368	366	114	53
hsa-miR-30c-5p	25	24	4	3	454	451	23	11
hsa-miR-30d-5p	5	5	3	2	337	335	275	243
hsa-miR-30e-5p	7	7	0	0	314	313	32	8
hsa-miR-31-5p	37	36	1	0	132	127	2	0
hsa-miR-34a-5p	67	63	50	43	511	491	452	425
hsa-miR-92a-3p	34	34	31	30	1222	1205	1144	1124
hsa-miR-92b-3p	2	2	0	0	603	593	65	38
hsa-miR-93-5p	28	28	13	12	1114	1099	868	753
hsa-miR-98-5p	470	461	225	111	235	230	103	56
hsa-miR-99a-5p	10	10	3	1	115	114	40	19
hsa-miR-99b-5p	3	3	2	1	42	42	17	7
hsa-miR-100-5p	23	23	8	8	209	207	115	65
hsa-miR-101-3p	23	23	0	0	300	294	1	0
hsa-miR-106b-5p	79	75	14	1	934	920	205	58
hsa-miR-124-3p	1232	1209	1058	918	123	116	105	93
hsa-miR-125b-5p	70	70	50	38	305	303	250	218
hsa-miR-127-3p	7	7	3	2	13	13	5	4
hsa-miR-129-5p	12	12	11	9	353	344	291	267
hsa-miR-130a-3p	20	20	11	9	334	325	157	83
hsa-miR-130b-3p	9	9	0	0	507	497	5	1

Abbreviations: FDR, false discovery rate; mRNA, messenger RNA; miRNA, microRNA.

Table 3. Comparison of miRNA targets for 50 miRNAs with non-mRNA expression methods was used to RNA-Seq in colon tissue.

MIRNA	TOTAL, N	EXPRESSED IN COLON TISSUE, N	VALIDATED BY RNA-SEQ $P_{UNADJ} < .05$, N	VALIDATED BY RNA-SEQ FDR < 0.05, N
Total target genes	12 191	11 850	3770	2416
hsa-miR-30c-1-3p	421	410	20	6
hsa-miR-139-3p	70	66	1	0
hsa-miR-145-3p	45	44	5	4
hsa-miR-151a-3p	73	71	46	40
hsa-miR-151b	24	24	6	2
hsa-miR-192-3p	121	117	13	5
hsa-miR-204-3p	94	90	1	0
hsa-miR-324-5p	276	272	121	45
hsa-miR-361-3p	121	118	81	73
hsa-miR-378b	27	26	0	0
hsa-miR-378d	27	26	1	0
hsa-miR-378g	57	57	3	1
hsa-miR-378i	27	26	1	0
hsa-miR-425-3p	38	36	22	16
hsa-miR-455-3p	426	416	6	2
hsa-miR-466	157	155	17	9
hsa-miR-501-3p	62	58	4	3
hsa-miR-513c-3p	172	169	124	80
hsa-miR-532-3p	200	196	12	4
hsa-miR-550a-3-5p	81	77	12	4
hsa-miR-550b-2-5p	86	82	0	0
hsa-miR-583	133	129	16	8
hsa-miR-623	212	201	1	0
hsa-miR-652-3p	130	129	95	61
hsa-miR-654-5p	89	84	74	73
hsa-miR-659-5p	23	23	17	13
hsa-miR-662	16	16	0	0
hsa-miR-664a-3p	84	79	0	0
hsa-miR-664a-5p	108	105	13	6
hsa-miR-664b-3p	138	135	99	70
hsa-miR-664b-5p	13	13	8	6
hsa-miR-769-3p	145	140	129	119
hsa-miR-877-5p	210	207	158	118
hsa-miR-892b	54	54	7	2
hsa-miR-934	17	17	1	1

(Continued)

Table 3. (Continued)

MIRNA	TOTAL, N	EXPRESSED IN COLON TISSUE, N	VALIDATED BY RNA-SEQ $P_{UNADJ} < .05$, N	VALIDATED BY RNA-SEQ FDR < 0.05, N
hsa-miR-939-5p	125	120	27	6
hsa-miR-1183	77	75	1	0
hsa-miR-1203	24	23	3	0
hsa-miR-1207-3p	67	63	25	10
hsa-miR-1225-5p	32	30	24	14
hsa-miR-1228-5p	26	26	9	0
hsa-miR-1229-5p	42	40	18	7
hsa-miR-1233-5p	123	122	4	2
hsa-miR-1271-5p	67	66	4	0
hsa-miR-1288-3p	25	25	10	6
hsa-miR-1305	176	173	84	57
hsa-miR-1913	112	110	0	0
hsa-miR-1973	5	5	4	3
hsa-miR-2117	50	48	20	5
hsa-miR-2392	106	104	19	5

Abbreviations: FDR, false discovery rate; mRNA, messenger RNA; miRNA, microRNA.

and miR-124-3p which all have more than 500 database-identified target genes with all genes expressed in normal colonic mucosa gave us an indication as to how the databases compared at both ends of the mRNA target gene spectrum. We showed a large number of miRNA-mRNA associations for that had not previously linked to the miRNA (Table 4). Of the more than 80 000 miRNA-mRNA associations we detected using RNA-Seq data, 15.6% and 14.8% had seed matches for CRCh38 and CRCh37, respectively, supporting the hypothesis that seed matches would increase the likelihood of a direct association given the higher propensity for binding.

Discussion

Evaluation of 254 miRNAs previously associated with colon cancer, MSI tumor phenotype, or with survival after diagnosis with colorectal cancer showed a great deal of variability in the number of targeted genes for each miRNA in miRTarBase. A major finding is the documentation of the incompleteness of the target genes for many miRNAs. Although our study was limited to colon cancer and could be considered a partial database with respect to miRTarBase, we identified numerous genes whose expression was associated with miRNA expression in colon tissue that were not identified in miRTarBase by similar gene expression methods. We believe that this variability is indicative of the extent to which miRNAs have been studied than actual differences in the number of targeted genes by specific miRNAs. However, the implications for determining functionality and pathways associated with genes targeted by miRNAs resulting from this discrepancy are many; pathways are driven by those miRNAs which have been researched the most. Second, although our partial database is restricted to only colon tissue and gene expression, it points out other limitations. First, our inability to identify target genes incorporated in the database from similar gene expression studies suggests that tissue specificity is important in determine disease-specific pathways. Pathways that encumber non–site-specific genes could obviously be irrelevant for the disease and tissue of interest. In short, although miRNA-mRNA associations may exist in some tissues, they may not be relevant in other tissue types. Our data suggest that 2% to 3% of targeted genes are not expressed in colon tissue and of those miRNA-mRNA associations reported in databases that are expressed in colon tissue, and less than 50% of targeted genes previously identified with mRNA validation methods could be validated with RNA-Seq data in our colon cancer samples. Lack of specificity can limit the accuracy of projected pathways generated by existing databases.

The variability in the number of targeted genes identified in the existing database with the 254 miRNAs that we evaluated is considerable (Figure 2). Although some miRNAs such as miR-21-5p had more than 500 previously identified targeted genes, 15 miRNAs had fewer than 20 targeted genes identified.

Table 4. miRNA-mRNA associations in colon tissue using RNA-Seq data and matching seed region.

MIRNA	DATABASE-VALIDATED TARGET GENES, N	VALIDATED TARGET GENES EXPRESSING IN COLON TISSUE, N	DATABASE-VALIDATED TARGET GENES BY MRNA EXPRESSION METHODS, N	MIRNA-MRNA ASSOCIATIONS IDENTIFIED IN COLON TISSUE (P_{UNADJ} < .05), N	MIRNA-MRNA ASSOCIATIONS IDENTIFIED IN COLON TISSUE (FDR < 0.05), N	NEW MIRNA-MRNA ASSOCIATIONS IDENTIFIED (FDR < 0.01) USING CRCH 38 TO IDENTIFY SEED MATCH, N	NEW MIRNA-MRNA ASSOCIATIONS IDENTIFIED (FDR < 0.01) USING CRCH37 TO IDENTIFY SEED MATCH
hsa-miR-3677-3p	4	4	0	13 363	13 059	1039	966
hsa-miR-6068	4	4	0	9805	6834	486	450
hsa-miR-1973	5	5	0	6918	1891	231	216
hsa-miR-3181	5	4	0	11 983	11 268	765	688
hsa-miR-4315	14	14	0	728	0	N/A	N/A
hsa-miR-664b-5p	13	13	0	11 803	10 472	1876	1757
hsa-miR-4730	15	15	0	8796	4884	564	527
hsa-miR-3621	16	16	0	10 100	7869	657	589
hsa-miR-662	16	16	0	461	0	N/A	N/A
hsa-miR-6717-5p	15	15	0	5982	1300	58	56
hsa-miR-4681	17	17	0	123	0	N/A	N/A
hsa-miR-572	17	17	1	6178	0	N/A	N/A
hsa-miR-934	17	17	0	785	0	N/A	N/A
hsa-miR-127-3p	20	20	7	6908	844	41	36
hsa-miR-4787-5p	20	20	0	10 883	9616	1142	1068
hsa-miR-21-5p	558	552	434	8715	6646	1851	1754
hsa-miR-215-5p	713	707	670	2664	0	N/A	N/A
hsa-miR-124-3p	1355	1325	1209	13 370	12 860	4961	4815

Abbreviations: FDR, false discovery rate; mRNA, messenger RNA; miRNA, microRNA.

The effect of this disparity is evident in current pathway tools. For instance, when hsa-miR-21-5p, which had 558 validated targets, is entered into miRPath[24] (http://snf-515788.vm.okeanos.grnet.gr), an updated miRNA pathway tool, 34 pathways associated with an FDR correction applied and when using TarBase-validated target genes. When hsa-miR-3677-3p, which had 4 validated targets, is entered using the same parameters, 1 significant pathway is returned. Development of pathways associated with dysregulated miRNAs is thus heavily influenced by those miRNAs that have many genes previously identified and is only minimally represented by many miRNAs that have been examined in less detail and are shown to be associated with diseases. Pathways used to determine functionality of miRNAs or to identify therapeutic targets are incomplete and are dominated by a subset of the total miRNAs associated.

Tissue specificity of miRNA expression, while a concern, appears to have a less impact in terms of pathway identification because 97% to 98% of mRNAs were expressed in our targeted colon tissue. However, a greater concern is that although miRNAs and mRNAs may be expressed in the tissue, they may not have the same regulatory impact. We saw that less than half of the mRNAs previously identified by gene expression methods were actually associated with their identified target in our data. This lack of reproducibility could be from tissue specificity. However, other reasons for the lack of association could also exist. Because we were looking at gene expression from RNA-Seq, we adjusted for more comparisons than perhaps some of the previously reported studies. It should be kept in mind that the microarray and next-generation sequencing methods are considered less reliable and our lack of confirmation of some of these associations could stem from the level of confidence in data such as these.

Validation of miRNA target genes includes methods that have varying degrees of their strength of evidence as well as what they are validating. Western blot and reporter assays that detect protein levels along with qPCR which measures mRNA expression levels have been considered stronger evidence than microarrays or next-generation sequencing techniques such as RNA-Seq or CLIP-Seq.[11] Methods that can validate protein expression and are target-specific miRNAs to determine whether changing miRNA alters mRNA expression level or protein expression are ideal.[25–27] However, these techniques do not lend themselves to more widespread exploration of targeted genes. Microarrays and RNA-Seq, both methods that can more broadly curate gene expression, detect mRNA expression levels and do not evaluate protein expression. Because miRNAs function through posttranscriptional regulation to effect protein expression, these techniques could fail to identify associations with target genes that might exist. MicroRNAs can also affect mRNA expression through degradation by partial complementarity binding of target sequences. Target gene validation techniques, such as RNA-Seq, would detect variation in mRNA that could be correlated with miRNA expression. It has been stated that RNA-Seq provides "an alternative to microarray gene expression analysis allowing a deeper analysis to provide a larger list of inferring miRNA targets in comparable over-expression studies."[28]

Although we have shown that some miRNAs are associated with thousands of mRNAs, we have attempted to identify mRNAs that are more directly associated with the miRNA of interest. It is most likely that many mRNAs are not directly associated with the miRNA of interest, but rather dysregulation is seen secondary to other directly targeted genes.[27] To identify direction interactions, we used seed matches between the 5′ region of the mature miRNA from nucleotides 2 to 8, or the seed region, and the 3′ UTR of the mRNA.[21] A seed region match suggests a more functional binding between the miRNA and the targeted gene. It also has been suggested by Thomson et al[28] that seed matches can provide a mechanism of enriching for direct miRNA targets over indirect or secondary effects. Evaluation of seed region matches between significant miRNA-mRNA associations showed that there were a considerable number of previously unreported target genes (Table 4). This would suggest that many of the associations are indirectly related, but that the databases are incomplete for most miRNAs when it comes to identifying targeted genes and subsequent pathways constructed from those genes.

Another important consideration when interpreting miRNA and their targeted pathways is the complexity of miRNAs and their associations with gene expression. MicroRNAs regulate hundreds if not thousands of genes, and individual genes are regulated by many miRNAs. Furthermore, it is unlikely that miRNA-mRNA associations apply to all tissues. The lack of specificity in these associations adds complexity to constructing pathways because it is difficult to know which miRNA-mRNA–targeted pairs are most relevant for the condition being studied.

The study has several strengths. First is our depth of data that includes individuals with both miRNA and RNA-Seq data. This allows us to attempt colon cancer–specific functionality assessment. In addition, using RNA-Seq data to evaluate mRNA expression, we are able to consider many more genes expressed in colon tissue than other methods that are more labor-intensive and tissue-intensive. This allows us to undertake a broader discovery of new miRNA-mRNA associations. However, there are limitations in that associations were detected by non-mRNA methods that may imply translational miRNA mechanisms that would not be detected when looking at gene expression. Availability of protein data would enhance our analysis but is unavailable. We adjusted for multiple comparisons and through this adjustment could have missed previously identified associations. Although we have applied rigid QC and compared subsets of both our miRNA and mRNA data to qPCR with good results,[2,29] there is potential for inaccuracies in these techniques.

There are several other considerations when constructing functional pathways associated with miRNAs. Our data suggest that miRNA-targeted gene databases are incomplete; pathways derived from these databases have similar deficiencies. Although we know a lot about several miRNAs, we know very little about other miRNAs in terms of what genes they target. It appears that for most miRNAs, the information is incomplete in terms of validated targeted genes and that tissue-specific associations exist.

Conclusions

Existing databases of miRNA-targeted genes have limitations both in terms of coverage for specific miRNAs and tissue-specific miRNA-mRNA associations. We encourage others to use their data to continue to further identify and validate miRNA-targeted genes to improve the likelihood that research conducted on miRNAs will help translate to improve medical care.

Acknowledgements

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. The authors acknowledge Sandra Edwards for data oversight and Dr Bette Caan and the staff at the KPMCP for their contribution to data collection, Dr Wade Samowitz for pathologic slide review and Erica Wolff and Michael Hoffman for miRNA analysis, and Brett Milash and the Bioinformatics Shared Resource of the Huntsman Cancer Institute and University of Utah for miRNA and mRNA bioinformatics data processing.

Author Contributions

MLS obtained funding, conducted research, and wrote the manuscript. LEM assisted in writing manuscript and conducted bioinformatics analysis. JSH conducted statistical analysis. JRS provided input into statistical analysis. RKW oversaw laboratory conducting miRNA and RNA-Seq analysis.

Availability of Data and Material

Data are made available in accordance with signed consent forms. Please contact author for data requests.

Ethical Approval and Consent to Participate

Participants signed informed consent prior to release of confidential data. The Institutional Review Boards of the University of Utah and the KPMCP approved the study.

REFERENCES

1. Engels BM, Hutvagner G. Principles and effects of microRNA-mediated post-transcriptional gene regulation. *Oncogene*. 2006;25:6163–6169.
2. Pellatt DF, Stevens JR, Wolff RK, et al. Expression profiles of miRNA subsets distinguish human colorectal carcinoma and normal colonic mucosa. *Clin Transl Gastroenterol*. 2016;7:e152.
3. 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;37:245–261.
4. Liu B, Ding JF, Luo J, Lu L, Yang F, Tan XD. Seven protective miRNA signatures for prognosis of cervical cancer. *Oncotarget*. 2016;7:56690–56698.
5. Zeng YB, Liang XH, Zhang GX, et al. miRNA-135a promotes hepatocellular carcinoma cell migration and invasion by targeting forkhead box O1. *Cancer Cell Int*. 2016;16:63.
6. Bandopadhyay M, Sarkar N, Datta S, et al. Hepatitis B virus X protein mediated suppression of miRNA-122 expression enhances hepatoblastoma cell proliferation through cyclin G1-p53 axis. *Infect Agent Cancer*. 2016;11:40.
7. Chakraborty C, Chin KY, Das S. miRNA-regulated cancer stem cells: understanding the property and the role of miRNA in carcinogenesis. *Tumour Biol*. 2016;37:13039–13048.
8. Lee S, Lim S, Ham O, et al. ROS-mediated bidirectional regulation of miRNA results in distinct pathologic heart conditions. *Biochem Biophys Res Commun*. 2015;465:349–355.
9. Yue J. miRNA and vascular cell movement. *Adv Drug Deliv Rev*. 2011;63:616–622.
10. Li T, Yang GM, Zhu Y, et al. Diabetes and hyperlipidemia induce dysfunction of VSMCs: contribution of the metabolic inflammation/miRNA pathway. *Am J Physiol Endocrinol Metab*. 2015;308:E257–E269.
11. Chou CH, Chang NW, Shrestha S, et al. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res*. 2016;44:D239–D247.
12. Lupini L, Bassi C, Ferracin M, et al. miR-221 affects multiple cancer pathways by modulating the level of hundreds messenger RNAs. *Front Genet*. 2013;4:64.
13. Wang YP, Li KB. Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. *BMC Genomics*. 2009;10:218.
14. Slattery ML, Herrick JS, Pellatt DF, et al. Site-specific associations between miRNA expression and survival in colorectal cancer cases. *Oncotarget*. 2016;7:60193–60205.
15. Slattery ML, Herrick JS, Mullany LE, et al. Colorectal tumor molecular phenotype and miRNA: expression profiles and prognosis. *Mod Pathol*. 2016; 29:915–927.
16. Slattery ML, Potter J, Caan B, et al. Energy balance and colon cancer—beyond physical activity. *Cancer Res*. 1997;57:75–80.
17. Agilent Technologies I. *Agilent GeneSpring User Manual*. 2013. Accessed July 16, 2015. http://genespring-support.com/files/gs_12_6/GeneSpring-manual.pdf.
18. Slattery ML, Herrick JS, Mullany LE, et al. An evaluation and replication of miRNAs with disease stage and colorectal cancer-specific mortality. *Int J Cancer*. 2015;137:428–438.
19. Slattery ML, Pellatt DF, Mullany LE, Wolff RK, Herrick JS. Gene expression in colon cancer: a focus on tumor site and molecular phenotype. *Genes Chromosomes Cancer*. 2015;54:527–541.
20. Kozomara AG-JS. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*. 2014;42:D68–D73.
21. Mullany LE, Herrick JS, Wolff RK, Slattery ML. MicroRNA seed region length impact on target messenger RNA expression and survival in colorectal cancer. *PLoS ONE*. 2016;11:e0154177.
22. Karolchik D, Hinrichs AS, Furey TS, et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. 2004;32:D493–D496.
23. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc*. 1995;57:289–300.
24. Vlachos IS, Zagganas K, Paraskevopoulou MD, et al. DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res*. 2015;43:W460–W466.
25. Kuhn DE, Martin MM, Feldman DS, Terry AV Jr, Nuovo GJ, Elton TS. Experimental validation of miRNA targets. *Methods*. 2008;44:47–54.
26. Martinez-Sanchez A, Murphy CL. MicroRNA target identification-experimental approaches. *Biology (Basel)*. 2013;2:189–205.
27. Vasudevan S. Functional validation of microRNA-target RNA interactions. *Methods*. 2012;58:126–134.
28. Thomson DW, Bracken CP, Goodall GJ. Experimental strategies for microRNA target identification. *Nucleic Acids Res*. 2011;39:6845–6853.
29. Slattery ML, Pellatt DF, Mullany LE, Wolff RK. Differential gene expression in colon tissue associated with diet, lifestyle, and related oxidative stress. *PLoS ONE*. 2015;10:e0134406.