Clinical Cancer Research

Discovery and Validation of Circulating Biomarkers of Colorectal Adenoma by High-Depth Small RNA Sequencing

Brian S. Roberts[1], Andrew A. Hardigan[1,2], Dianna E. Moore[1], Ryne C. Ramaker[1,2], Angela L. Jones[1], Meredith B. Fitz-Gerald[3], Gregory M. Cooper[1], C. Mel Wilcox[4], Robert P. Kimberly[3,5], and Richard M. Myers[1]

Abstract

Purpose: Colorectal cancer is the third most common cancer worldwide, causing approximately 700,000 deaths each year. The majority of colorectal cancers begin as adenomas. Definitive screening for colorectal adenomas is currently accomplished through colonoscopy but, owing largely to costs and invasiveness, is typically limited to patient groups at higher risk by virtue of age or family history. We sought to determine if blood-based small RNA markers could detect colorectal adenoma.

Experimental Design: We applied high-depth small RNA sequencing to plasma from a large ($n = 189$) cohort of patients, balanced for age, sex, and ancestry. Our analytical methodology allowed for the detection of both microRNAs and other small RNA species. We replicated sequencing results by qPCR on plasma samples from an independent cohort ($n = 140$).

Results: We found several small RNA species with significant associations to colorectal adenoma, including both microRNAs and non-microRNA small RNAs. These associations were robust to correction for patient covariates, including age. Among the adenoma-associated small RNAs, two, a miR-335-5p isoform and an un-annotated small RNA, were validated by qPCR in an independent cohort. A classifier trained on measures of these two RNAs in the discovery cohort yields an AUC of 0.755 (0.775 with age) for adenoma detection in the independent cohort. This classifier accurately detects adenomas in patients under 50 and is robust to sex or ancestry.

Conclusions: Circulating small RNAs (including but not limited to miRNAs) discovered by sequencing and validated by qPCR identify patients with colorectal adenomas effectively. *Clin Cancer Res; 24(9); 2092–9. ©2018 AACR.*

Introduction

Colorectal cancer is the third most common cancer with a global annual incidence of 1.4 million new cases and approximately 700,000 deaths (1). In the United States, colorectal cancer incidence and mortality rates have been declining for decades due to increased screening in patients over 50 years old and to improvements in treatments (2, 3). However, colorectal cancer incidence has been rising in younger age groups in whom screening is not currently performed. Detection of early stages of disease has significant impact on colorectal cancer patient outcomes, as stage at diagnosis is inversely correlated with survival (4).

Canonical colorectal cancer pathogenesis (accounting for 85% of cases) proceeds along an adenoma to adenocarcinoma sequence wherein the successive accumulation of driver mutations in tumor suppressor genes (e.g., *p53* and *APC*) or oncogenes (e.g., *KRAS*) leads to aberrant colonic epithelial growth and adenoma formation in the colon (5). Further development of gross chromosomal instability leads to colorectal cancer invasion and metastasis (6). Because the development of colorectal cancer along the adenoma to adenocarcinoma path can take as long as 10 years, there is a large window for diagnostic and therapeutic intervention (7).

Current guidelines recommend screening by colonoscopy for all patients after the age of 50 and earlier for high-risk populations (7). Colonoscopy has several advantages to other screening methods, such as the ability to remove identified adenomas concurrently, but the procedure is invasive, costly, and requires specialized medical expertise to perform. Less invasive screening tests have been developed, such as the fecal occult blood test (FOBT), fecal immunochemical test (FIT), and more recent stool-based testing for DNA changes associated with adenoma-adenocarcinoma progression in combination with FIT. While stool-based screening tests have fewer risks than colonoscopy, they are still relatively costly and have limited colorectal cancer prediction accuracy, with follow-up colonoscopy required for positive results (4). The development of a noninvasive, sensitive, and specific blood-based test for colorectal cancer screening that could be implemented in a standard blood panel and more

[1]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama. [2]Department of Genetics, University of Alabama at Birmingham, Birmingham, Alabama. [3]Center for Clinical and Translational Science, University of Alabama at Birmingham, Birmingham, Alabama. [4]Department of Medicine, Division of Gastroenterology and Hepatology, University of Alabama at Birmingham, Birmingham, Alabama. [5]Department of Medicine, Division of Clinical Immunology and Rheumatology, University of Alabama at Birmingham, Birmingham, Alabama.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Corresponding Author: Richard M. Myers, HudsonAlpha Institute for Biotechnology, 601 Genome Way, Huntsville, AL 35806. Phone: 256-327-0431; Fax: 256-327-0978; E-mail: rmyers@hudsonalpha.org

doi: 10.1158/1078-0432.CCR-17-1960

Translational Relevance

Earlier detection of colorectal cancer and precancerous colorectal adenoma by colonoscopy improves patient outcomes. However, colonoscopy is expensive, invasive, and often restricted to older patients or those with known risk factors. Fecal detection methods have shown the promise of noninvasive molecular diagnostics for colorectal cancer screening, but have some significant limitations. Blood-based testing for colorectal adenoma would offer numerous desirable characteristics, including a simple clinical implementation and likely higher corresponding patient compliance. By large-scale sequencing, we found circulating small RNA signatures of colorectal adenoma in patient blood samples. We replicated these findings by using qPCR, a clinically compatible assay, in an independent cohort. Furthermore, we found two small RNAs that significantly predict colorectal adenoma regardless of sex, ancestry, or age, including in younger patients (<50 years old).

effectively prioritize patients for follow-up colonoscopy has the potential to decrease the economic and clinical burden of current screening methods and improve patient outcomes.

Noncoding small RNAs, including microRNAs (miRNA) and others, are readily found in circulation within extracellular vesicles or bound to RNA-binding proteins (8, 9). These small RNAs, in particular miRNAs, have been extensively studied because of their diverse roles in regulating cancer-related cellular processes, such as differentiation and cell-cycle exit (10, 11). Aberrant expression of miRNAs has been implicated in a variety of malignancies (12). Furthermore, advances in sequencing technologies have allowed for both the discovery of new small RNAs and evaluation of them as potential biomarkers for cancer diagnosis and progression.

Previous studies have evaluated circulating small RNA associations with colorectal cancer and colorectal adenoma, and identified several possible miRNA biomarkers for diagnosis and monitoring of disease progression (13–15). Yamada and colleagues describe a study that uses qPCR to detect circulating miRNAs in a sizable cohort of patients that yielded effective classifiers of both colorectal cancer and colorectal adenoma (16). In a different study, Vychytilova-Faltejskova and colleagues used low-coverage sequencing in pooled samples to identify circulating miRNA colorectal cancer biomarker candidates and validated them using qPCR in a large cohort (17). They demonstrated good classifier performance on colorectal cancer and found significant miRNA level changes for adenoma samples in some cases. All of these reports establish the potential for circulating small RNA in the detection of colorectal cancer and adenoma. However, they are limited in their consideration of only miRNA amongst all small RNAs, potentially confounded by covariates like patient age, and, in cases in which sequencing was used, relatively low-depth coverage per patient.

Here, we present results of a relatively larger study, in terms of sample size and sequencing depth, in which we identified circulating small RNA biomarkers of colorectal adenoma. Our approach is not limited to miRNAs, and we found that many non-miRNA small RNAs have significant associations. Highly optimized sample processing and library generation combined with rigorous data processing and statistical analysis led us to findings that reproduce with technically orthogonal assays and in independent cohorts. Our selection of large, balanced, and diverse cohorts allowed us to identify colorectal adenoma biomarkers that are independent of age and sex effects and permitted us to discover intriguing associations with self-reported ancestry. Lastly, while preliminary, our data suggest that the identified small RNA biomarkers have utility in adenoma detection in patients below age 50, a group for whom colonoscopies are not routinely offered.

Materials and Methods

Sample collection and processing

Study subjects were enrolled from patients undergoing routine colonoscopies at a clinic at the University of Alabama Birmingham (UAB) Medical Center in Birmingham, Alabama. Subjects were selected from those undergoing colorectal polyp screening; those being examined for other reasons (IBD, Crohn's Disease, others) were excluded. Written informed consent was obtained for each patient in accordance with ethical principles embodied in the Belmont Report. Prior to the colonoscopy, approximately 10 mL of blood was obtained in an EDTA-treated Vacutainer. Within 30 minutes of collection, the sample was centrifuged at 2,200 × g for 10 minutes. The plasma supernatant was isolated, yielding approximately 4 mL plasma, and immediately frozen at −80°C until further processing. For each patient, colonoscopy findings, relevant medical history, and demographic information, including self-reported ancestry, were recorded. The entire protocol was reviewed and approved by the AAHRPP-accredited UAB IRB (UAB IRB# X130327016).

RNA isolation and sequencing library generation

Plasma samples were thawed on ice. Immediately upon thawing, 1.1 mL of plasma was centrifuged at 3,000 × g for 15 minutes at 4°C to pellet debris. One mL of the supernatant was used as input for the Plasma/Serum Circulating and Exosomal RNA Purification Kit (Slurry Format; Norgen Biotek) and then concentrated using the RNA Clean-up and Concentration Micro Kit (Norgen Biotek). Manufacturer instructions were followed exactly for both kits. The purified RNA was stored at −80°C until library generation.

Small-RNA-sequencing libraries were prepared exactly as previously described (18). Briefly, adaptors were ligated, the product converted to cDNA, and amplified with 15 cycles of PCR. A blocking oligonucleotide was used which drastically reduced the level of hsa-miR-16-5p in the sequencing library, enhancing the overall complexity. The PCR product was subjected to electrophoresis under extremely denaturing conditions and fragments collected corresponding to an insert size of 15 to 30 base pairs. After purification, the library was sequenced on an Illumina HiSeq 2000. For every sample, replicate libraries were prepared from separate plasma aliquots isolated from the same blood draw. Replicate libraries were made and sequenced in separate batches to minimize the effects of technical variation on the results. All raw fastq files and a count table for all samples are publicly available from GEO (GEO Accession number: GSE110381).

Quantitative PCR of small RNAs

We used the Universal cDNA Synthesis Kit II (Exiqon) to prepare the samples for quantitative PCR (qPCR). The kit uses a poly-adenylase to tail the RNA and reverse transcription to convert to cDNA. For each sample, 4 μL of isolated RNA was added to each 10 μL reaction. The product was diluted 1:20 with 10 mmol/L TRIS buffer containing 0.05% Tween 20. Four microliters of the diluted cDNA was combined with 5 μL of 2× PowerSYBR Master Mix (Applied Biosystems) and 1 μL of assay specific primer mix (Exiqon). All samples were run in triplicate on a QuantStudio 6 Flex system (Applied Biosystems) using default settings. Threshold cycle values (C_t) were called using the manufacturer's software.

Sequencing data processing

Details of our read processing analysis and justification of parameters are provided in the Supplementary Methods, along with a set of scripts capable of precisely reproducing the results and figures in this article. A flow chart of our read processing pipeline is presented in Supplementary Fig. S1. We trimmed adaptor sequences from the raw reads and then aligned to the human genome reference assembly (GRCh37/hg19), with only perfect matches allowed. Reads aligning to multiple locations were counted in proportion to the total number of valid alignment locations. We made these choices by finding conditions that optimized the reproducibility of replicate libraries. For example, Supplementary Fig. S2 demonstrates that allowing one mismatch in the whole genome alignments reduces replicate reproducibility when compared with reads aligned with no mismatches. Also, as comparator reads were aligned to pre-miRNA sequences only. Reads aligned to the whole genome with no mismatches show similar reproducibility to the pre-miRNA aligned reads, while those with one mismatch do not (see Supplementary Methods for more details).

Reads were then coalesced into features by merging overlapping reads. Because this resulted in millions of features, many of which were very lowly abundant or appeared in a small subset of samples, we reduced the feature number by both eliminating features that did not appear in at least 10 samples and by requiring a minimal mean read depth of 1 read. Read counts (or fractional counts for ambiguously aligned reads) for each retained feature and sample were then calculated.

Small RNA features were annotated overlapping their genomic locations with various candidate regions, including miRNAs, tRNAs, rRNAs, Y RNAs, snoRNAs, and others. Overlap enrichment and empirical P values were estimated by comparing actual feature overlap to overlaps with 1,000 randomly generated feature sets matched for sequence content and other characteristics (see Supplementary Methods for details).

Statistical analysis

Detailed description of the statistical analysis is provided in Supplementary Methods along with a comprehensive set of scripts. Briefly, read counts from replicate libraries were summed after the higher total read-depth replicate was downsampled to match the lower one ensuring that each replicate contributed equally. We converted the read counts from the passed features into counts for fragments as actually sequenced. For example, every isomiR (differential 5′ or 3′ ends) of a given miRNA was analyzed as a separate entity, and the equivalent done for non-miRNA small RNAs. Note that because perfect alignments were used, variants arising from nontemplated base addition, or other mechanisms that lead to mismatches between the RNA and genomic DNA, were not considered in this analysis.

We used regression via the R package DESeq2 (19) in a "leave one out" approach taking into account patient covariates (age, sex, self-reported ancestry) along with adenoma status to find significant associations. Also, DESeq2 was used to normalize and transform the data into variance-stabilized data (VSD) and used in subsequent analyses. To build a multivariate classifier, we used LASSO regression as implemented in the R package glmnet (20). LASSO attempts to balance model size and potential overfitting against classier performance while yielding a closed-form model (21). Prior to LASSO regression, we employed a linear regression to age, sex, and self-identified ancestry on VSD. We used the residuals from this regression as inputs to LASSO. Receiver operating characteristic (ROC) curves with confidence intervals were calculated using the R package pROC (22).

We normalized C_t values from qPCR analyses to a median C_t of four chosen small RNA species that included miR-26a-5p, miR-139-5p, miR-10a-5p, and a Y RNA fragment (Supplementary Table S1). Based on sequencing data, we chose these species for their low standard deviation across samples, lack of significant association with any patient covariate, and their broad range of expression levels (spanning a 100-fold range).

Results

Identification of diverse small RNA species by sequencing

We generated an average of approximately 11 million aligned small RNA-sequencing reads from each of 432 libraries. In some cases, we generated more than two replicate libraries for each sample when quality issues were detected. After selecting the best two replicates for each sample and applying a minimum total count threshold per library, 378 libraries from 189 samples remained for further analysis (Fig. 1A). The median total read depth (combining the two replicates) for each retained sample was 19,361,058 reads with a standard deviation of 8,397,223.

We found 13,927 small RNA species that had at least 10 counts in one or more samples; there were 125,806 unique fragment mappings among these 13,927 species. Of these, we found unique mappings to 685 of the 935 pre-miRNAs in defined in the UCSC Browser and 13,097 non-miRNA small RNA species. Consistent with our treatment of the RNA preparations with DNase, we found that sequencing reads are derived from only one strand of the genome for the vast majority of the non-miRNA small RNAs we measured, strongly suggesting they are genuine RNAs rather than DNA contaminants (Fig. 1B).

To describe the small RNA species we found, we compared them with RNA types previously reported to be found in human plasma or serum (9). Consistent with these reports, we found large and significant enrichments ($P < 0.001$) for miRNA, ribosomal RNA (rRNA), transfer RNA (tRNA), small nucleolar RNA (snoRNA), and Y RNA, as well as a smaller but still significant enrichment for Piwi-interacting RNA (piRNA; Fig. 1C). We also found small but significant enrichments for any transcript defined by Ensembl (GRCh37, version 87). Additionally, we compared our small RNA features with CLIP-seq binding site mappings in K562 cells obtained from the ENCODE Consortium (Fig. 1D). We found highly significant enrichments ($P < 0.001$) for all 89 of the RNA-binding proteins assayed in this set, consistent with previous reports of stabilization of small RNAs in circulation by RNA-binding proteins (23, 24).

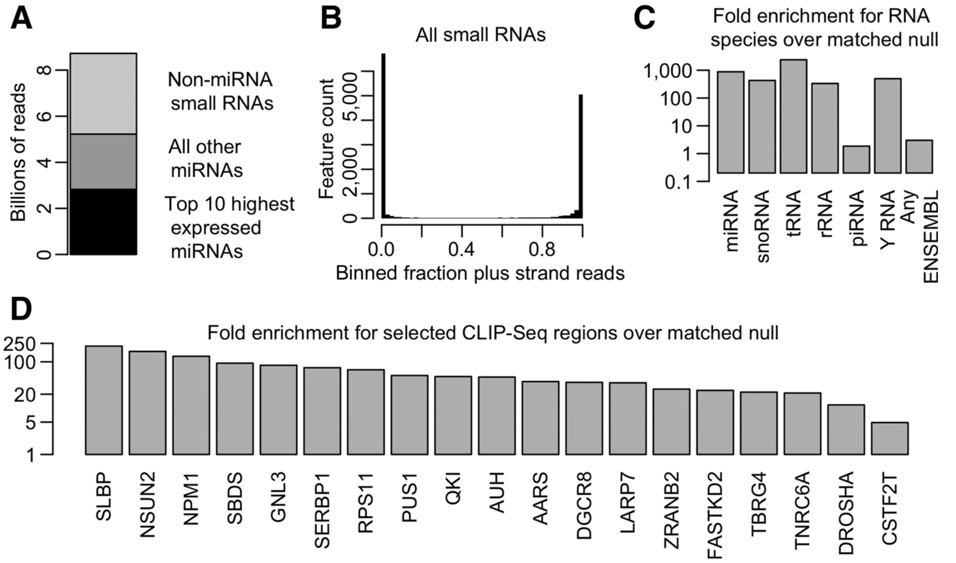

Figure 1.

Description of small RNA species in the sequencing study. **A,** The total number of aligned reads used in the analysis are split into three categories. **B,** This histogram shows the fraction of reads for a given small RNA feature that were derived from the plus strand. For the vast majority of small RNA species, nearly all of the reads come from one strand only. **C,** The height of the bars represents the fold enrichment of the small RNA features found in our study for the given RNA species annotation on the *x*-axis. The fold enrichment is calculated compared with the mean overlap of 1,000 matched null sets of equal size to the small RNA feature set (Supplementary Methods). **D,** Bars show the fold enrichment of the small RNAs for selected RNA-binding protein CLIP-Seq binding site mappings.

Patient covariates have large effects

The 189 patients in the sequencing discovery cohort were balanced for sex, age, and presence or absence of adenomas (Supplementary Table S2). Forty percent of the patients in this cohort self-identified as African American, with the remainder self-identifying as of European ancestry; other self-reported identifications were not included in the analysis due to insufficient numbers. We sought to find small RNA species that significantly associate with ancestry, sex, age, and the detection of colorectal adenomas during colonoscopy (Fig. 2) using a "leave one out" regression strategy (Supplementary Methods). The *P* values plotted in Fig. 2 represent the significance of the variation associated with the given independent variable that is independent of the other considered covariates. A detailed account of the results of the differential expression analysis is presented in Supplementary Table S3.

We found numerous small RNA species with significant associations to the patients' age ($n = 40$ at $P < 1e{-4}$), sex ($n = 123$ at $P < 1e{-4}$), and self-identified ($n = 248$ at $P < 1e{-4}$) ancestry, after filtering for low counts (DESeq2 base mean >10) and condensing fragments that map to multiple loci but have the same sequence. Many of the most significant ancestry-associated small RNA species are derived from a single locus at chr14:101,250,000–101,550,000. This locus, commonly referred to as the *DLK1-DIO3* locus, is known to be imprinted in humans, with the snoRNAs and miRNAs contained therein exclusively expressed from maternal chromosomes (25). Fifty-two pre-miRNA loci are annotated within the locus, and we detected the expression of 81 distinct mature miRNAs (including 3′ isomiRs) from 31 of these pre-miRNAs above a DESeq2 base mean of 100 (our threshold for reliable differential expression measurement, see Supplemental Methods). All 81 expressed miRNAs correlate with ancestry (*P* values range from 1.1e−2 to 2.7e−9), and all of them are regulated in the same direction to nearly equal degrees. This coregulation is consistent with evidence that small RNAs in the *DLK1-DIO3* locus are transcribed as a single-polycistronic transcript and then processed to each of the smaller species (26, 27). As a validation, we measured a representative of the cluster, hsa-miR-382-5p, in an independent patient cohort by qPCR and found significant association with ancestry ($P = 2.6e{-5}$) in the same direction (Supplementary Fig. S4).

Validation of colorectal adenoma associated small RNAs

We found many small RNA species significantly associated with the presence of colorectal adenomas upon colonoscopy ($n = 225$ at $P < 1e{-4}$; Fig. 2). Because patient age, sex, and ancestry were included in the regression, the small RNA species we found are not

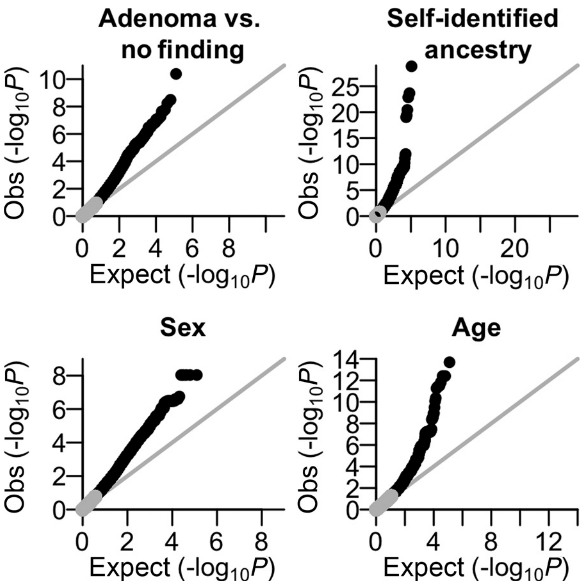

Figure 2.

Significance of the association of small RNA species with patient characteristics. Quantile-Quantile plots of $-\log_{10} P$ values from DESeq2 regressions versus those from a uniform distribution are presented for adenoma status, ancestry, sex, and age, respectively. The *P* values represent the significance between a model that includes all independent variables versus a model with the independent variable of interest left out. Black dots represent those $-\log_{10}$ *P* values that deviate from the null expectation by greater than 1.2 fold.

merely a proxy for risk associated with these covariates. We found a slight but significant enrichment for lower association P values in miRNAs versus non-miRNAs [Kolmogorov-Smirnov (KS) test $P = 1.40e-5$, Supplementary Fig. S5]. Nevertheless, we found highly significant associations in both classes.

We next sought to find a minimal set of small RNAs with maximal ability to classify patients with adenoma from those without by employing LASSO binomial regression (21). We restricted the input to the regression to those small RNA species that we also predicted to yield agreement sequencing and qPCR. We generated these predictions by evaluating, for each small RNA, the likelihood of cross-reaction between closely related fragments with similar 5′ and 3′ ends (i.e., isomiRs of the same mature form) and the effects of a cross-reaction due to dissimilarity between the expression values of the fragments (Supplementary Methods). From the LASSO regression results, we selected the top 10 weighted small RNA species for further validation (Supplementary Table S1).

To validate the association of this set of 10 small RNAs with colorectal adenoma status, we measured their expression by qPCR in an independent validation cohort of 140 patient samples that had not been included in the discovery cohort from our sequencing study. Similar to the discovery cohort, the patient characteristics of this validation cohort are largely balanced for sex, ancestry, age, and colorectal adenoma status (Supplementary Table S2). We investigated the association of the expression values of these 10 small RNA species as measured by sequencing in the original discovery cohort and their performance as measured by qPCR on the separate validation cohort. For three of the small RNA species, qPCR failed due to low C_t values (>33) and poor melting curve behavior (Supplementary Table S1). For two of the remaining seven, we found consistent and significant fold-changes between the sequencing discovery cohort and the qPCR validation cohort for two (Fig. 3). The other five's association with adenoma was not significant (Supplementary Fig. S6). The reasons for the lack of agreement may be technical or true biological variation. Of the two small RNA species in agreement, the first maps to two locations, chr16:89206352-89206369 and chr20:60808046-60808063. The chromosome 16 mapping is within an intron of *ACSF3*, and the chromosome 20 mapping is within an exon of a hypothetical transcript, AK126744. The second is an isoform of hsa-miR-335-5p, one base pair shorter than the canonical form at the 3′ end.

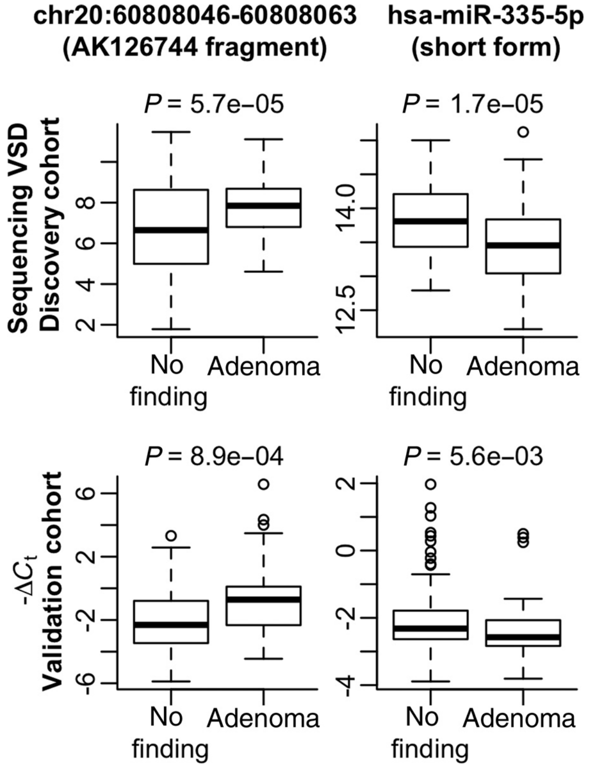

Figure 3.

Validation of adenoma-associated small RNAs in an independent cohort. The top row shows boxplots of two validated adenoma-associated small RNAs from sequencing VSD in the discovery cohort. The bottom row shows boxplots of the same two small RNAs from qPCR $-\Delta C_t$ values in the validation cohort. Linear regression P values between the two categories are presented above the plots.

Small RNA-based classifiers predict colorectal adenoma and replicate in an independent cohort

We created a classifier using binomial regression from the two small RNA species that validated. As inputs, we used *z*-scored sequencing VSD from the entire discovery cohort and applied this model's intercepts and coefficients to subsequent predictions, ensuring no backward flow of information. When applied to sequencing data in the discovery cohort, this classifier performs well without age (AUC = 0.710, $P = 2.81e-7$) and more accurately when age is added (AUC = 0.781, $P = 1.27e-11$, Fig. 4, left). At a threshold chosen to optimize sensitivity and specificity (event probability = 0.51), the classifier yields sensitivity of 69.1%, specificity of 72.6%, positive predictive value (PPV) of 71.4%, and negative predictive value of (NPV) of 70.4% applied to sequencing data with age in the discovery cohort. When applied to qPCR data from the validation cohort (to which it was not trained) the performance is even better (AUC = 0.755, $P = 1.11e-7$) also improving with age inclusion (AUC = 0.775, $P = 1.08e-8$) (Fig. 4, right). Again at the chosen threshold, the classifier with age yields a sensitivity of 71.1%, a specificity of 70.3%, PPV of 74.3%, and NPV of 68.2% in the validation cohort. The classifier outperforms age alone in both cohorts (discovery cohort age only AUC = 0.678, validation cohort age only AUC = 0.682). The choice of normalizing species in circulating small RNA qPCR can be difficult (28). We explored whether our choice to normalize to four small RNA species based on sequencing data (Supplementary Table S1) could affect the performance of the classifier. We normalized to all four species and all subsets of three, two, and one, and found the performance of the classifier was robust to the choice of normalizers (Supplementary Fig. S7). Furthermore, the classifier (expression data only) performed slightly better in African Americans (AUC = 0.787) than in individuals of European ancestry (AUC = 0.744) and equally well in males (AUC = 0.759) and females (AUC = 0.752).

Identified small RNA classifier has utility in younger patients

Because we enrolled all patients undergoing colorectal adenoma screening, our patient cohort had a small number of patients younger than 50 years of age, as expected based on colonoscopy screening guidelines. Nevertheless, 28 of the patients in the discovery cohort were younger than age 50 (6 adenoma, 22 no finding, median age 43) and 19 patients in the validation cohort

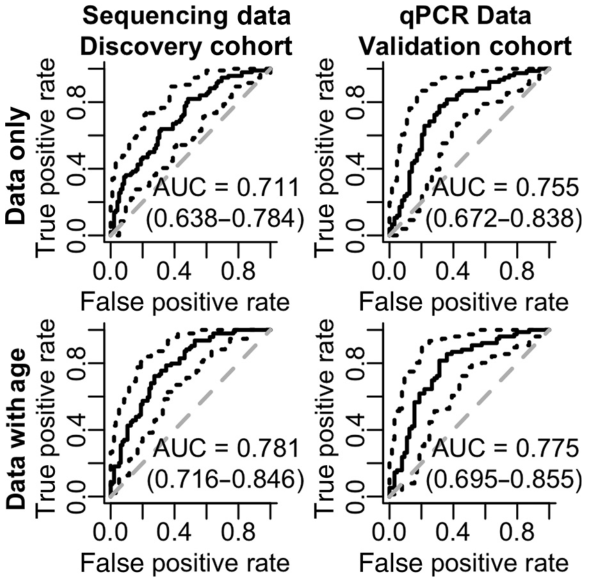

Figure 4.

Performance of a small RNA expression classifier on two cohorts. Each panel shows an ROC curve based on the classifier's ability to distinguish adenoma status in patients. The black line represents the mean of ROC curves from 2,000 randomly sampled subsets while the dashed lines represent 95% confidence intervals. The top row shows the performance of the classifier based on small RNA expression measurements only while the bottom shows the performance with age also included in the classifier.

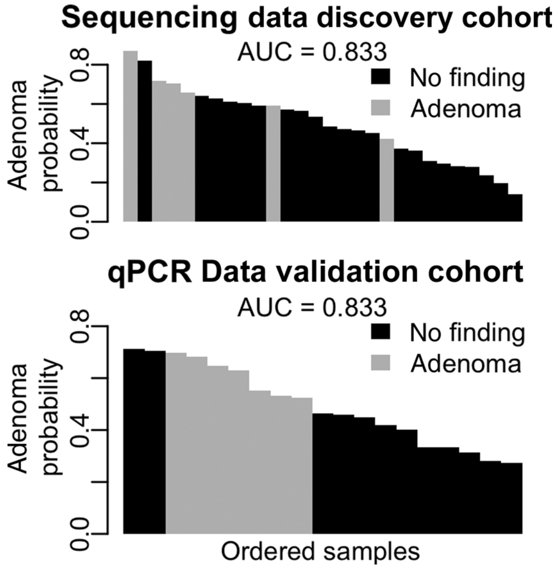

Figure 5.

Prediction of adenoma status in younger patients. Bar heights represent the predicted probability of adenoma in patients less than 50 years old. Bar colors represent the true status of the patients. Sequencing data from the discovery cohort is used for the top panel and qPCR data in the validation cohort is used in the bottom panel.

(7 adenoma, 12 no finding, median age 37). On sequencing data in the discovery cohort, we found the classifier (trained on the full cohort) has good performance on the 28 "young" patient samples (AUC = 0.833, $P = 5.9e{-}3$, Fig. 5, top). Also, when applied to qPCR data from the validation cohort, the classifier also performs well on the 19 "young" patients (AUC = 0.833, $P = 8.5e{-}3$, Fig. 5, bottom).

Discussion

Our study identified several small RNA species with significant associations to colorectal adenoma status. Key features of our method include isolation of the plasma within 30 minutes of the blood draw, preparation and sequencing of replicate libraries from each patient plasma sample, blocking of the highly abundant hsa-miR-16-5p from the libraries, high-depth sequencing of each sample, alignment with parameter choices empirically justified based on replicate reproducibility, and rigorous statistical analyses. Furthermore, the selection of a cohort balanced for sex, ancestry, and age allowed not only the discovery of significant small RNA associations with these characteristics and concomitant biological validation of the methodology, but also rigorous control of the effects of these covariates on colorectal adenoma status.

Our approach to alignment and feature definition found small RNAs that are heavily enriched for miRNA, tRNA, Y RNA, snoRNA, and rRNA annotations, consistent with previous work by others (9). These features were overwhelmingly driven from strand-biased reads, indicating transcription, rather than genomic DNA contamination, as a likely source. Furthermore, we found small RNAs mapping to genomic locations with significant enrichments for a variety of non-Argonaute RNA-binding protein binding sites, which is, to our knowledge, a novel finding. Intriguingly, this binding suggests that, like Argonaute (23), these proteins may be stabilizing small RNAs in circulation. However, it should be noted that these enrichments are based on uncurated CLIP-seq data, which may contain spurious peaks. Nevertheless, we found that for the most enriched RNA-binding protein, SLBP, the top 5 sites most uniquely bound by SLBP (having few overlapping non-SLBP binding sites) were all histone mRNA fragments, consistent with SLBP's known function (29).

Of the small RNA species selected for qPCR testing, we found two that validate in an independent cohort. The first maps to two genomic locations, an intron of *ACSF3* (Acyl-CoA synthetase family member 3) and an exon of hypothetical transcript AK126744. No function has been attributed to AK126744, and while *ACSF3* dysfunction is causal for combined malonic and methylmalonic aciduria (30), its possible relationship colorectal adenoma is unclear. However, both regions are annotated as transcribed, giving validity to the existence of an RNA product with this sequence.

The second validated small RNA is miRNA miR-335-5p. We chose to validate an isoform that is one base shorter on the 3′ end than the canonical form because it showed maximal effect size. However, we find that both the canonical form and an isoform two bases shorter on the 3′ end also have significant association with adenoma in the same direction, although another isoform one base longer on the 3′ end does not (Supplementary Table S3). The role of miR-335-5p in a variety of cancers has been well documented. Although its functions in cancer can be diverse, in many cancers and subtypes it demonstrates decreased expression in tumors and exhibits tumor suppressor activity (31–33), notably in colorectal cancer (34). Also, consistent with our observations, Tsikitis and colleagues find miR-335-5p levels decreased in colorectal adenoma tissue samples when compared with normal adjacent mucosa or hyperplastic polyps, and, in particular, serrated adenomas (35). These independent studies strongly corroborate our observations of lower expression of miR-335-5p in the plasma of patients with adenoma.

We also demonstrated that a classifier derived from just two small RNA expression measurements has utility in distinguishing patients with colorectal adenomas from those without (Fig. 4). This classifier (AUCs of 0.710 and 0.755 on discovery and replication cohorts) was more predictive of adenoma status than age (AUCs of 0.678 and 0.682 on discovery and replication cohorts, respectively), which is a predominant factor in screening recommendations. Furthermore, adding age to classifiers built from the small RNA measurements improves their performance above what either yields separately, indicating they contain independent information. This is important because many small RNAs significantly associate with age (Fig. 2), and failure to control for this could have resulted in a classifier that was merely a surrogate for age.

The classifier is robust to patient sex and self-identified ancestry, indicating adequate analytical control for these covariates as well. Furthermore, we found that the classifier also performs well on the subset of patients under the age of 50, albeit on small sample numbers (Fig. 5). We do not have information as to why these young patients were undergoing colonoscopy, and cannot rule out that they could represent a higher risk population, potentially confounding our results. Confirmation and refinement of this result using a larger study is warranted, as assays capable of identifying younger patients not currently screened by colonoscopy with colorectal adenoma could be highly impactful.

In summary, we demonstrate the discovery of technically and biologically reproducible small RNA associations with adenoma and potentially early-stage cancer from unbiased large-scale sequencing of blood samples. Moreover, these findings may extend to younger patients, who currently do not receive routine colorectal cancer screening. Given that sequencing costs are likely to continue to decrease, methods for cell-free nucleic acid isolation are becoming clinically feasible, and our findings strongly confirm a subset of the sequencing results, we propose that small RNA sequencing of biofluids may have clinically relevant utility for colorectal cancer screening in the near future.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: B.S. Roberts, A.A. Hardigan, R.P. Kimberly, R.M. Myers
Development of methodology: B.S. Roberts, A.A. Hardigan, R.P. Kimberly, R.M. Myers
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): B.S. Roberts, A.A. Hardigan, D.E. Moore, A.L. Jones, M.B Fitz-Gerald, C.M. Wilcox, R.P. Kimberly, R.M. Myers
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): B.S. Roberts, A.A. Hardigan, R.C. Ramaker, G.M. Cooper, C.M. Wilcox, R.P. Kimberly, R.M. Myers
Writing, review, and/or revision of the manuscript: B.S. Roberts, A.A. Hardigan, R.C. Ramaker, G.M. Cooper, C.M. Wilcox, R.M. Myers, R.P. Kimberly
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): D.E. Moore, A.L. Jones, R.P. Kimberly, R.M. Myers
Study supervision: B.S. Roberts, M.B Fitz-Gerald, R.P. Kimberly, R.M. Myers

Acknowledgments

We thank the patients who participated in this study, the members of the University of Alabama at Birmingham Clinical Research Support Program who aided the collection and management of the plasma samples, the Center for Clinical and Translational Science Laboratory at the University of Alabama at Birmingham for sample processing, and the Genomic Services Laboratory, led by Dr. Shawn Levy, at the HudsonAlpha Institute for Biotechnology for provided sequencing services and data processing.

This study was supported by funding from the Center for Clinical and Translational Science at the University of Alabama in Birmingham, NCATS UL1 TR0001417 (M. Fitz-Gerald, C. Wilcox, and R. Kimberly), the State of Alabama, HudsonAlpha Institute (B. Roberts, D. Moore, and R. Myers), the UAB Medical Scientist Training Program, NIGMS 5T32GM008361 (A. Hardigan and R. Ramaker), and the contribution of an anonymous private donor.

Received July 7, 2017; revised January 2, 2018; accepted February 21, 2018; published first February 28, 2018.

References

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-tieulent J, Jemal A. Global Cancer Statistics, 2012. CA Cancer J Clin 2015;65:87–108.
2. Siegel RL, Fedewa SA, Anderson WF, Miller KD, Ma J, Rosenberg PS, et al. Colorectal cancer incidence patterns in the United States, 1974–2013. J Natl Cancer Inst 2017;109:27–32.
3. Siegel RL, Miller KD, Fedewa SA, Ahnen DJ, Meester RGS, Barzi A, et al. Colorectal Cancer Statistics, 2017. CA Cancer J Clin 2017;67:177–93.
4. van Lanschot MCJ, Carvalho B, Coupé VMH, van Engeland M, Dekker E, Meijer GA. Molecular stool testing as an alternative for surveillance colonoscopy: a cross-sectional cohort study. BMC Cancer 2017;17: 116.
5. Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, et al. Genetic alterations during colorectal-tumor development. N Engl J Med 1988;319:1557–62.
6. Dienstmann R, Vermeulen L, Guinney J, Kopetz S, Tejpar S, Tabernero J. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. Nat Rev Cancer 2017;17:79–92.
7. Winawer SJ, Fletcher RH, Miller L, Godlee F, Stolar MH, Mulrow CD, et al. Colorectal cancer screening: clinical guidelines and rationale. Gastroenterology 1997;112:594–642.
8. Dhahbi JM, Spindler SR, Atamna H, Yamakawa A, Boffelli D, Mote P, et al. 5′ tRNA halves are present as abundant complexes in serum, concentrated in blood cells, and modulated by aging and calorie restriction. BMC Genomics 2013;14:298.
9. Yeri A, Courtright A, Reiman R, Carlson E, Beecroft T, Janss A, et al. Total extracellular small RNA profiles from plasma, saliva, and urine of healthy subjects. Sci Rep 2017;7:44061.
10. Ivey KN, Srivastava D. MicroRNAs as regulators of differentiation and cell fate decisions. Cell Stem Cell 2010;7:36–41.
11. Carleton M, Cleary MA, Linsley PS. MicroRNAs and cell cycle regulation. Cell Cycle 2007;6:2127–32.
12. Lujambio A, Lowe SW. The microcosmos of cancer. Nature 2012;482: 347–55.
13. Uratani R, Toiyama Y, Kitajima T, Kawamura M, Hiro J, Kobayashi M, et al. Diagnostic potential of cell-free and exosomal MicroRNAs in the identification of patients with high-risk colorectal adenomas. PLoS One 2016;11:1–16.
14. Verma AM, Patel M, Aslam MI, Jameson J, Pringle JH, Wurm P, et al. Circulating plasma microRNAs as a screening method for detection of colorectal adenomas. Lancet 2015;385:S100.

15. Ho GYF, Jung HJ, Schoen RE, Wang T, Lin J, Williams Z, et al. Differential expression of circulating microRNAs according to severity of colorectal neoplasia. Transl Res 2015;166:225–32.
16. Yamada A, Horimatsu T, Okugawa Y, Nishida N, Honjo H, Ida H, et al. Serum MIR-21, MIR-29a, and MIR-125b are promising biomarkers for the early detection of colorectal neoplasia. Clin Cancer Res 2015; 21:4234–42.
17. Vychytilova-Faltejskova P, Radova L, Sachlova M, Kosarova Z, Slaba K, Fabian P, et al. Serum-based microRNA signatures in early diagnosis and prognosis prediction of colon cancer. Carcinogenesis 2016;37:941–50.
18. Roberts BS, Hardigan AA, Kirby MK, Fitz-Gerald MB, Wilcox CM, Kimberly RP, et al. Blocking of targeted microRNAs from next-generation sequencing libraries. Nucleic Acids Res 2015;43:e145.
19. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:1–34.
20. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33:1–22.
21. Tibshirani R. Regression selection and shrinkage via the lasso. J R Stat Soc 1996;58:267–88.
22. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77.
23. Arroyo JD, Chevillet JR, Kroh EM, Ruf IK, Pritchard CC, Gibson DF, et al. Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. Proc Natl Acad Sci U S A 2011; 108:5003–8.
24. Wang K, Zhang S, Weber J, Baxter D, Galas DJ. Export of microRNAs and microRNA-protective protein by mammalian cells. Nucleic Acids Res 2010;38:7248–59.
25. da Rocha ST, Edwards CA, Ito M, Ogata T, Ferguson-Smith AC. Genomic imprinting at the mammalian Dlk1-Dio3 domain. Trends Genet 2008; 24:306–16.
26. Takada S, Tevendale M, Baker J, Georgiades P, Campbell E, Freeman T, et al. Delta-like and gtl2 are reciprocally expressed, differentially methylated linked imprinted genes on mouse chromosome 12. Curr Biol 2000; 10:1135–8.
27. Tierling S, Dalbert S, Schoppenhorst S, Tsai C-E, Oliger S, Ferguson-Smith AC, et al. High-resolution map and imprinting analysis of the Gtl2–Dnchc1 domain on mouse chromosome 12. Genomics 2006;87:225–35.
28. Schwarzenbach H, da Silva AM, Calin G, Pantel K. Data normalization strategies for MicroRNA quantification. Clin Chem 2015;61:1333–42.
29. Zhang M, Lam TT, Tonelli M, Marzluff WF, Thapar R. Interaction of the histone mRNA hairpin with stem-loop binding protein (SLBP) and regulation of the SLBP-RNA complex by phosphorylation and proline isomerization. Biochemistry 2012;51:3215–31.
30. Sloan JL, Johnston JJ, Manoli I, Chandler RJ, Krause C, Carrillo-Carrasco N, et al. Exome sequencing identifies ACSF3 as a cause of combined malonic and methylmalonic aciduria. Nat Genet 2011;43:883–6.
31. Kawaguchi T, Yan L, Qi Q, Peng X, Gabriel EM, Young J, et al. Overexpression of suppressive microRNAs, miR-30a and miR-200c are associated with improved survival of breast cancer patients. Sci Rep 2017;7:15945.
32. Sandoval-Borquez A, Polakovicova I, Carrasco-Veliz N, Lobos-Gonzalez L, Riquelme I, Carrasco-Avino G, et al. MicroRNA-335-5p is a potential suppressor of metastasis and invasion in gastric cancer. Clin Epigenetics 2017;9:114.
33. Luo LJ, Wang DD, Wang J, Yang F, Tang JH. Diverse roles of miR-335 in development and progression of cancers. Tumor Biol 2016;15399:410.
34. Sun Z, Zhang Z, Liu Z, Qiu B, Liu K, Dong G. MicroRNA-335 inhibits invasion and metastasis of colorectal cancer by targeting ZEB2. Med Oncol 2014;31:982.
35. Tsikitis VL, Potter A, Mori M, Buckmeier JA, Preece CR, Harrington CA, et al. MicroRNA signatures of colonic polyps on screening and histology. Cancer Prev Res 2016;9:942–9.