Identification of a pan-cancer oncogenic microRNA superfamily anchored by a central core seed motif

Mark P. Hamilton  
*Baylor College of Medicine*

Kimal Rajapakshe  
*Baylor College of Medicine*

Sean M. Hartig  
*Baylor College of Medicine*

Boris Reva  
*Memorial Sloan Kettering Cancer Center*

Michael D. McLellan  
*Washington University School of Medicine in St. Louis*

*See next page for additional authors*

Follow this and additional works at: https://digitalcommons.wustl.edu/open_access_pubs

Please let us know how this document benefits you.

**Recommended Citation**

https://digitalcommons.wustl.edu/open_access_pubs/4318

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact vanam@wustl.edu.
Authors

This open access publication is available at Digital Commons@Becker: https://digitalcommons.wustl.edu/open_access_pubs/4318
Identification of a pan-cancer oncogenic microRNA superfamily anchored by a central core seed motif

Mark P. Hamilton1, Kimal Rajapakshe1, Sean M. Hartig1, Boris Reva2, Michael D. McLellan3, Cyriac Kandoth3, Li Ding3,4,5, Travis I. Zack6, Preethi H. Gunaratne7,8, David A. Wheeler8, Cristian Coarfa1 & Sean E. McGuire1,9

MicroRNAs modulate tumorigenesis through suppression of specific genes. As many tumour types rely on overlapping oncogenic pathways, a core set of microRNAs may exist, which consistently drives or suppresses tumorigenesis in many cancer types. Here we integrate The Cancer Genome Atlas (TCGA) pan-cancer data set with a microRNA target atlas composed of publicly available Argonaute Crosslinking Immunoprecipitation (AGO-CLIP) data to identify pan-tumour microRNA drivers of cancer. Through this analysis, we show a pan-cancer, coregulated oncogenic microRNA ‘superfamily’ consisting of the miR-17, miR-19, miR-93, miR-455 and miR-210 seed families, which cotargets critical tumour suppressors via a central GUGC core motif. We subsequently define mutations in microRNA target sites using the AGO-CLIP microRNA target atlas and TCGA exome-sequencing data. These combined analyses identify pan-cancer oncogenic cotargeting of the phosphoinositide 3-kinase, TGFβ and p53 pathways by the miR-17-19-130 superfamily members.

1Department of Molecular and Cellular Biology, Baylor College of Medicine, 1 Baylor Plaza Houston M822, Houston, Texas 77030, USA. 2Computational Biology Center, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. 3The Genome Institute, Washington University, St Louis, Missouri 63108, USA. 4Department of Genetics, Washington University, St Louis, Missouri 63110, USA. 5Siteman Cancer Center, Washington University, St Louis, Missouri 63110, USA. 6The Eli and Edythe L Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02142, USA. 7Department of Biology and Biochemistry, University of Houston, 4800 Calhoun, Houston 77204, Texas, USA. 8The Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA. 9Division of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. Correspondence and requests for materials should be addressed to S.E.M. (email: sean.mcguire@bcm.edu).

DOI: 10.1038/ncomms3730
MicroRNAs are single-stranded RNA molecules (~22 nucleotides) that repress messenger RNA translation and promote mRNA degradation. MicroRNAs are critical regulators of oncogenesis and their regulation of cancer cell signalling is complex. Global microRNA expression is often repressed in cancer. However, some microRNAs are oncogenic, exhibiting amplified expression in many tumour types.

Facilitated by Argonouate proteins, microRNAs bind target mRNAs in the RNA-induced silencing complex. MicroRNA target regulation is canonically mediated by nucleotides 2–8 on the 3′-end of the microRNA strand, termed the microRNA seed. A minimum of six consecutive nucleotides is required to pair the microRNA with its target mRNA. This minimal binding requirement allows a given microRNA to potentially bind tens, hundreds or thousands of mRNA targets.

One difficulty in determining the functions of microRNAs in tumours is the wide array of potential genes that any microRNA might regulate. Established microRNA target prediction algorithms are based on inference, relying on evolutionary conservation of 3′-untranslated region (UTR) sequences complementary to the microRNA seed and biochemical binding context to determine putative microRNA binding sites. Although these algorithms are useful for predicting microRNA targets, especially within the 3′-UTRs, they are not experimental demonstrations of microRNA–target interactions and are often less able to accurately predict microRNA binding within protein-coding regions and non-coding RNAs (ncRNAs) because of reliance on site-specific conservation.

Argonouate Crosslinking Immunoprecipitation (AGO-CLIP) data sets experimentally identify microRNA–target interactions in a genome-wide manner through purification of Argonouate-protein-associated RNAs, which include bound microRNAs and their respective targets. In this study, to explore the microRNA regulatory landscape across the TCGA Pan-Cancer project, which includes data from breast adenocarcinoma (BRCA), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), uterine corpus endometrioid carcinoma (UCEC), and acute myeloid leukemia (LAML), we compiled all publicly available human AGO-CLIP data into a single unified atlas and ranked individual microRNA target sites by total occurrences across data sets. We integrated this substantial atlas of microRNA target sites with TCGA pan-cancer microRNA, mRNA, copy number variation (CNV) and exome-sequencing data sets to discover common microRNA regulatory architecture across tumour types. Finally, we developed an algorithm, miSNP, to infer somatic mutations in these regulatory binding sites. Our analysis represents integration of a new resource, the AGO-CLIP atlas, and TCGA data, creating a method by which we were able to understand microRNA regulatory architectures across multiple tumour types. Collectively, this study identified a pan-cancer oncogenic microRNA (oncomiR) network that cotargets multiple potent tumour suppressors (TS) through a common core seed motif.

**Results**

**Global microRNA expression patterns in normal and tumour tissue.** The TCGA pan-cancer data set represents the single largest compilation of microRNA-sequencing data in cancer produced to date. Global analysis of microRNA expression patterns in 4,186 tumours and 334 normal tissue samples revealed the top 30 microRNAs constitute, on average, ~90% of all microRNA expression across heterogeneous normal tissues. The same 30 microRNAs likewise comprise 80–90% of microRNA expression in tumours (Fig. 1a, Supplementary Tables S1 and S2).

**miR-143** is the single, most highly expressed microRNA in normal tissue, and **miR-21** is the most highly expressed microRNA in cancer (Fig. 1b). MicroRNA expression patterns undergo global population changes between cancer and normal, primarily due to increased **miR-21** expression (from 6.9 to 19% of all microRNA detected) and decreased **miR-143** expression (from 33 to 11.2% of detectable microRNA) across tumour types.

**AGO-CLIP atlas identifies global microRNA binding events.** AGO-CLIP technology employs ultraviolet crosslinking of RNA to protein followed by immunoprecipitation to determine RNA species bound to the Argonouate protein (Fig. 2a). AGO Photoactivatable-Ribonucleotide-Enhanced CLIP (AGO-PAR-CLIP) includes an added step where nucleotide analogues such as 4-thiouridine are introduced before crosslinking. These nucleotide analogues, when crosslinked, undergo T–C transitions during the reverse-transcription step of the AGO-CLIP experiment, allowing more confident visualization of RNA–protein interaction (Fig. 2b).

We began by generating a large atlas of microRNA binding sites by compiling publicly available AGO-CLIP data (Supplementary Data 1). This included 11 AGO-PAR-CLIP libraries and 3 unmodified AGO-CLIP libraries (also called Argonouate High-Throughput Sequencing of RNA Isolated by CLIP (AGO-HITS-CLIP)). The AGO-CLIP atlas allows us to integrate experimentally defined microRNA–target interactions with TCGA data to create the most accurate prediction of microRNA binding patterns across TCGA cancers. The AGO-CLIP seed atlas consists of 124,000 microRNA target clusters that subsequently infer over 300,000 putative seed motifs within those clusters. Individual seed sites were used as genomic anchors to tabulate recurrent definition of a given seed across 14 AGO-CLIP data sets (Fig. 2c, Supplementary Data 1). Clusters were then randomly permuted across the genome to determine an exact binomial probability of cluster occurrence at a given seed complement. False discovery rates (FDRs) for each seed-complementary target were calculated from their probability of recurrence (Supplementary Data 1–3). We found that ≥3 occurrences of an AGO-CLIP peak on a given target site corresponded to a significant event relative to a random distribution of clusters (q < 0.05 based on binomial P-value, Supplementary Data 3). AGO-CLIP defined that cluster localization by mRNA region is largely consistent with previous reports, with 60% of clusters mapping to the 3′-UTR, 24.7% of clusters mapping to the coding region, 8.2% mapping to the 5′-UTR and 7% mapping to ncRNAs.

**DICER1, MDM2** and the long ncRNA (lncRNA) **Xist** are among the top ten, most frequently targeted genes in the atlas, suggesting that the microRNA functional roles may include autoregulation, apoptotic sensitization through TP53 and lncRNA function (Supplementary Data 2). Importantly, traditional target prediction algorithms do not predict the high-frequency interactions on both MDM2 and Xist, demonstrating the strength of the unbiased AGO-CLIP platform. We also found numerous interactions between the Argonouate proteins and lncRNAs, small nucleolar RNAs and transfer RNAs in the AGO-CLIP atlas. These findings are consistent with recent evidence, suggesting that ncRNAs are microRNA targets or Argonouate binding substrates. Discovery of these interactions supports a...
A growing consensus that microRNA function extends beyond the regulation of protein-coding genes.

Analysis of TCGA microRNA expression data. To define which microRNAs consistently change relative to matched normal tissue, we performed significance testing on tumour versus normal microRNA expression levels across samples. Inspection of the TCGA microRNA data set revealed that significance testing (Fisher’s exact test) between tumour and normal samples on the raw reads per million values generated by high-level processed TCGA data produced significantly more increased microRNAs than decreased microRNAs (Supplementary Fig. S1A). The reason for this differential significance level between increased and decreased microRNAs is due to loss of highly expressed microRNAs in tumour samples, especially loss of miR-143, which accounts for 35–70% of microRNA expression in normal tissue. miR-143 expression often decreases by 450% in tumours (Fig. 1b). As microRNA expression in sequenced samples is expressed as a population value (reads per million microRNAs mapped), and because total RNA between tumour and normal samples used in sequencing experiments is constant, the loss of miR-143 leads to reciprocal gains in the proportion of other microRNAs. This relationship is demonstrated by the strong association of the absolute number of significantly increased microRNAs and the loss of miR-143 ($R^2 = 0.86, P = 0.01$).

Figure 1 | The landscape of microRNA expression in the TCGA pan-cancer data set. (a) Thirty microRNAs constitute 90% of microRNA expression across all normal tissues. (b) Global microRNA expression change occurring in tumours is due principally to loss of miR-143 expression and gain of miR-21 expression. MicroRNAs represented in columns from bottom to top listed left to right by row legend.
Supplementary Fig. S1B). We corrected for these composition differences using upper quartile and trimmed median of M-values to normalize the data set 26. These methods are designed to compensate for differences between tissues (for example, comparing the liver and the kidney) and are thus useful when comparing tumour versus normal tissue microRNA values, because they compensate for artefact generated by large expression changes in the most prevalent microRNAs (Supplementary Fig. S1A–D; Supplementary Data 4 contains detailed significance calculations for microRNAs in all tumour types).

Definition of microRNA–target interactions. We next defined pan-cancer oncomiRs and miR suppressors (tumour-suppressing microRNAs) based on consistent expression changes across cancer types. Pan-cancer oncomiRs were defined by significant expression gain (q < 0.05, Fisher’s exact test) in at least six out of seven pan-cancer tumour types containing tumour versus normal microRNA-sequencing data. Pan-cancer miR suppressors were similarly defined by significant expression loss in at least six of seven tumour types (Fig. 3a; Supplementary Data 5 contains detailed pan-cancer microRNA selection data). To ensure we were observing target interactions with highly expressed microRNAs that had many conserved target sites in the 3’-UTR, we focused on the dominant arms of the 87 broadly conserved microRNA families with an Argonaute-bound read group corresponding to the microRNA in at least 3 of the 14 AGO-CLIP data sets in our analysis.

We examined interactions between putative pan-cancer oncomiRs or miR suppressors, and their driver targets, based on the assumption that pan-cancer oncomiRs are enriched for TS targets and pan-cancer miR suppressors are enriched for oncogenic targets. We then performed integrative analysis of pan-cancer TS and oncopgenes (OCs) by using available pan-cancer data, including exome-sequencing single-nucleotide variation (SNV) scores (MuSiC27,28 and MSKCC29 algorithms), CNV analysis (GISTIC30,31 algorithm) and mRNA expression changes as building blocks for integrated gene nomination (Fig. 3b). We generated a continuous scale of relevant pan-cancer genes that describes putative TS as increasingly negative values and putative OCs as increasingly positive values, based on SNV, CNV and mRNA expression changes across TCGA tumour types (Supplementary Data 6 and Methods).

We tested four methods for calling microRNA–target interactions including the following: using all AGO-CLIP-defined binding sites without considering site conservation (for example, TargetScan); using only AGO-CLIP-defined sites with ≥3 occurrences and a TargetScan call. We found that combining AGO-CLIP and TargetScan results (final method) was the only method that produced enrichments of TS targets for pan-cancer oncomiRs and OC targets for pan-cancer miR suppressors (Fig. 3c, Supplementary Fig. S2A–D).

To gain insight into the differing enrichments, we explored the target spectrums of AGO-CLIP-defined target sites and TargetScan-defined target sites for the selected pan-cancer microRNAs. We found that on average 25.6% of TargetScan targets are also nominated by AGO-CLIP. Reciprocally, TargetScan also nominates 31.47% of AGO-CLIP targets. In total, 74.4% of all TargetScan targets were not called by AGO-CLIP, and TargetScan did not call 68.53% of AGO-CLIP targets. In the case of AGO-CLIP, 34.61% of all targets were outside the 3’-UTR (coding region or 5’-UTR), leaving another 34.39% of all
AGO-CLIP-defined targets within the 3′-UTR but not called by TargetScan (results are summarized in Supplementary Table S3).

Calling interactions using the AGO-CLIP atlas alone produced bias towards enrichment of OCs (Supplementary Fig. S2A,B), whereas TargetScan alone produced bias towards TS (Supplementary Fig. S2C). The AGO-CLIP data set produces slight bias towards OCs, because the top 3,000 OCs have 31% more AGO-CLIP clusters binding them than the top 3,000 TS. This observation may reflect overexpression of these OCs in the rapidly growing or oncogenic, the relative lack of AGO-CLIP experiments wherein these genes could have reduced expression owing to culturing conditions, rather than representing a general phenomenon.

Ultimately, determining microRNA–target interactions using AGO-CLIP-defined target sites with ≥3 occurrences, or ≥1 occurrences, and a TargetScan call was the only method to generate expected enrichments in TS targets for oncomiRs and OC targets for miR suppressors. This method had the added utility of combining target site conservation with genomic experimental validation. Discovering expected enrichments when combining TargetScan and AGO-CLIP values may suggest that combined microRNA target calling yields improved accuracy over a single method by reducing the false negatives and false positives inherent in each technology. As such, we chose to define a microRNA–target interaction as ≥3 AGO-CLIP-defined occurrences, and a microRNA–target interaction as ≥1 AGO-CLIP-defined occurrences.
occurrences, or ≥ 1 occurrences, and a TargetScan for subsequent analysis of tumour-driving microRNA–target interactions.

Identification of a pan-cancer oncomiR network. We observed that the strongest overall microRNA–target enrichment was for oncomiRs targeting the top 250 ranking TS (Fig. 3c, red box). Many of these interactions involved cotargeting of multiple microRNAs on the same TS. As many top TS targets were cotargeted by the pan-cancer oncomiRs, we analysed seed sequences of pan-cancer oncomiRs to determine whether any core sequences were common to the individual oncomiR seeds. Intriguingly, 10 of 22 (45.4%, Fig. 4a) pan-cancer oncomiRs in our analysis share similar sequence homology in their seed region that aligns around a central GUGC motif defining a microRNA seed ‘superfamily’. GUGC motifs occur in 36 of 187 (ref. 13) microRNAs from broadly conserved seed families (5.19%), enriching seed families with a GUGC motif in their seed region 8.7-fold among pan-cancer oncomiRs compared with all broadly conserved microRNAs (P = 0.008, Wilcoxon rank-sum test, Fig. 4a). None of the 25 pan-tumour miR suppressors identified in our analysis contain a GUGC motif, making this motif significantly depleted among the identified miR suppressors (P = 0.016, Wilcoxon rank-sum test). This motif is also enriched when including all microRNAs meeting our significance threshold, and not just the microRNAs from broadly conserved families (P = 5E – 10, Wilcoxon rank-sum test; the GUGC motif is present in 4.6% of all dominant-arm miRbase microRNAs and 26% of all microRNAs significantly increased in 6/7 TCGA tumours).

Several superfamily microRNAs (miR-17/106a, miR-210, miR-130b/301ab and miR-93/105) derive from the same seed family (Fig. 4a). The miR-93/105 and miR-17/106a seed families have virtually the same seeds, leading to highly similar predicted target spectrums. miR-17, miR-106a, miR-18a and miR-19a are part of the well-described miR-17–92 oncogenic cluster, also known as oncomiR-1 (refs 9,10,32). MicroRNA seed similarity in the pan-cancer oncomiRs led us to hypothesize that these microRNAs may undergo coordinate regulation to mutually cotarget and suppress critical TS.

To test this hypothesis, we defined the target spectrum of the pan-cancer microRNA superfamily (Fig. 4b,c). We observed oncogenic microRNA superfamily cotargeting of high-ranking pan-cancer TS such as SMAD4, ZBTB4 and TGFBR2, often at a single complementary seed-target site we termed a microRNA ‘super-seed’ target, where multiple families of microRNAs bind and regulate a specific 3’-UTR (Fig. 4b). In most superfamily oncomiRs, the majority of targets predicted in the top 3,000 TS are shared with at least one other superfamily member, including 70.2% of miR-19 targets, 79.3% of miR-130/301ab targets, 39.2% of miR-17/106a/93 targets, 42.3% of miR-18a targets, 75.7% of miR-455 targets and 62.5% of miR-210 targets (Fig. 4c).

The entire miR-17-19-130-93-455-210 superfamily of pan-tumour oncomiRs identified in this analysis forms three separate super-seed target sites; one consisting of the miR-17, miR-19 and miR-130 families, one consisting of the miR-18 and miR-19 families, and one consisting of the miR-19 and miR-455 families (Fig. 4b). The miR-17, miR-19 and miR-130 seed families exhibited the majority of total TS cotargeting on high-ranking TS in our data set. We thus focused further studies on tumour regulation by this subset of oncomiRs.

To test possible coregulation of the miR-17, miR-19 and miR-130 families, we correlated the expression levels of family members across TCGA tumours and found strong positive correlation of these microRNAs (average miR-17-19-130 family member microRNA–microRNA correlate across 11 tumour types, R² = 0.33, P < 1E – 200 versus null distribution, Student’s t-test, Fig. 4d). These data suggest that the miR-17-19-130 superfamily members undergo coordinate regulation in tumours to mediate silencing of TS genes in a synergistic manner.

To demonstrate the potential for cotargeting of TS by the miR-17-19-130 superfamily members, we determined pan-cancer correlates for all microRNA–target interactions in the top 250-ranked TS. Pan-cancer correlation of high-ranking TS targeted by the miR-17-19-130 superfamily revealed strong negative correlation of the superfamily with many pan-cancer TS, including PTEN, ZBTB4 and TGFBR2, across all tumour types. Figure 5a demonstrates correlations with the top four highest-ranked TS (PTEN, TGFBR2, TGFBR2 and SMAD4) that are targeted by all three seed families. PTEN, TGFBR2 and ZBTB4, all significantly negatively correlated with the miR-17-19-130 family members versus a null distribution of random microRNA–mRNA correlates (P < 1E – 15 for each, Student’s t-test). SMAD4 positively correlates with the superfamily in BLCA (P < 1E – 10, Student’s t-test), but otherwise shows no significant correlation, potentially suggesting a role for the microRNAs in translational repression of this target. Full correlate heat map for the microRNA–target interactions in the top 250 TS versus all pan-cancer oncomiRs is provided in Supplementary Data 7, a complementary heat map for targets of pan-cancer miR suppressors paired with the top 250 OCs is contained in Supplementary Data 8.

Next, we determined the ability of the miR-17-19-130 family to suppress translation of the top cotargeted TS, PTEN, ZBTB4, TGFBR2 and SMAD4 using 3’-UTR–luciferase fusions. We used miR-17, -19a and -130b as representative members of each seed family. We found cosuppressive capacity by pan-cancer oncomiRs on all four pan-cancer TS (Fig. 5b,c). In the case of ZBTB4, PTEN and SMAD4, all miR-17-19-130 superfamily members were able to bind to the 3’-UTR and significantly repress luciferase activity. miR-19a did not significantly suppress the TGFBR2 3’-UTR (Fig. 5b).

The SMAD4 gene contains a single miR-17-19-130 super-seed site that is highly conserved and few potential compensatory sites able to bind the miR-17, -19 or -130 seeds. As such, we deleted the central six nucleotides of the SMAD4 super-seed and measured strong ablation of each microRNA seed family’s ability to bind and regulate the SMAD4 3’-UTR (Fig. 5c). This finding illustrates the ability of a single 3’-UTR binding site to undergo coregulation by multiple microRNA families at a microRNA ‘super-seed’ target site.

PTEN is a conserved TS that regulates the oncoprotein phosphoinositide 3-kinase pathway33. TGFBR2 and SMAD4 are tumour-suppressive components of the transforming growth factor-β (TGFβ) pathway34,35. ZBTB4 is described as a mediator of the p53 response36. Thus, the sum of this analysis suggests that the pan-cancer oncomiR superfamily consisting of miR-17, miR-19 and miR-130 seed families coordinate target multiple critical tumour-suppressing pathways across tumour types (modelled in Fig. 7). We focused our analysis on the highest-ranking TS targets defined in an unbiased pan-cancer analysis of microRNA–target interactions in this study. Many of the described interactions have been defined previously9,36–38. However, this study defines these pathway targets as significant across multiple tumour contexts, based on an unbiased estimation of microRNA–target interactions in the largest single data set of human tumours produced to date. Hundreds of potential, novel interactions between these microRNAs and other targets are defined in Supplementary Data 7.

The AGO-CLIP atlas reveals mutations in microRNA targets. To define additional novel mechanisms of microRNA regulation in tumours, we next integrated the AGO-CLIP data set with TCGA mutation data to identify somatic SNVs in microRNA
Figure 4 | Pan-cancer oncomiRs are grouped into a coregulated superfamily of pan-oncomiRs. (a) Broadly conserved pan-cancer oncomiRs (45.4%) share a central GUCC seed sequence homology, grouping them into a larger oncomiR ‘superfamily’ consisting of miR-17, -19, -130, -210, -18 and -455 seeds. (b) MicroRNA superfamilies often target complementary ‘super seeds’. (c) Overview of target predictions for the top 3,000 highest-ranked T5 for miR-17, -19, -130, -210 and -455 families demonstrates TS cotargeting relationships with 39.2% of miR-17/106a/93 targets, 70.2% of miR-19 targets, 79.3% of miR-130/310ab targets, 42.3% of miR-18a targets, 75.7% miR-455 targets and 62.5% of miR-210 targets having predicted cotargeting with at least one other superfamily member. The miR-93 family has largely the same predicted target spectrum as that of the miR-17 family. miR-19 has few AGO-CLIP-defined targets and was thus not included in the Venn. miR-93 is grouped with the miR-17/93/106a super-seed target superfamily. The miR-19a super-seed target superfamily is grouped with the miR-93/106a/130a/310/19a ‘super-seed target’.
target sites across tumours. SNV analysis commonly involves identification of relevant coding-region somatic mutations through predicted functional impacts of amino acid changes associated with the nucleotide variations\textsuperscript{27,39}. This process, however, remains imperfect. Many missense mutations are not characterized as contributing significant functional impact on a gene.

Figure 5 | The miR-17-19-130 pan-cancer oncomiR superfamily binds and suppresses potent pan-cancer suppressor genes. (a) miR-17-19-130 microRNA–mRNA target correlations across tumours reveals strong negative correlation between superfamily members and their top ranked TS targets TGFBR2, PTEN and ZBTB4, but not SMAD4. (b) The miR-17-19-130 superfamily is able to coordinately bind and suppress expression of TS 3′-UTR-luciferase reporter constructs, indicating powerful interaction potential. (c) Superfamily coticatgering on the SMAD4 3′-UTR occurs at a novel microRNA super-seed locus where multiple microRNA seed families can bind, allowing for potential binding of more than individual microRNAs. Mix, an equimolar mixture of miR-17, -19a and -130b to demonstrate the co-repressive capacity of the oncomiR superfamily as it would exist in the cellular context. *P<0.05, **P<0.005, ***P<0.0005. Student’s t-test. Luciferase assays were performed twice at 5 nM mimic and twice at 10 nM in quadruplicate. Results were combined for final analysis (n = 16). Error bars are s.e.m. microRNA–mRNA correlate n-values are as follows: BLCA = 95, BRCA = 794, COAD = 177, HNSC = 301, KIRC = 466, LAML = 173, LUAD = 313, LUSC = 193, ovarian carcinoma (OV) = 225, READ = 65 and uterine corpus endometrioid carcinoma (UCEC) = 320. These numbers reflect the total number of TCGA tumour samples that are characterized with both mRNA and microRNA sequencing.
in regulatory regions (5′-UTR and 3′-UTR) outnumbers coding-region mutations. In our analysis of COAD whole-genome sequencing (WGS) samples, 60% of mutations in coding mRNAs are in the 3′-UTR, highlighting the potential importance of mutations in these regions (Fig. 6a).

As disruption of microRNA target sites complementary to the microRNA seed region directly interferes with microRNA binding40, any mutation in the microRNA target site complementary to the microRNA’s seed is likely to attenuate microRNA control of that site. As such, analysing mutations intersecting with microRNA seed-complementary sites has the potential to greatly expand the search for relevant cancer mutations by imbuing silent mutations and 3′-UTR mutations with functional significance.

To perform microRNA seed-target mutation analysis, we developed the miSNP algorithm to integrate AGO-CLIP data with TCGA-defined cancer mutations. miSNP intersects microRNA seed targets with mutation data and retrieves mRNA expression changes corresponding to mutations in these active sites (Fig. 6b), providing a powerful method to examine interactions between features and search for subsequent changes in mRNA associated with the interaction. Using the miSNP algorithm, we defined thousands of putative microRNA target-site mutations (Supplementary Data 9). The majority of TCGA pan-cancer SNV data derives from exome sequencing focused solely on coding-region sites. Therefore, the majority of microRNA target mutations we define from the 12 pan-cancer tumours occur specifically in the coding region. Importantly, it is...
reproduced the endogenous tumour mutation in the 3′ functional microRNA binding sites (Fig. 6c).

To demonstrate the ability of the AGO-CLIP technology and the miSNP algorithm to detect relevant microRNA target-site mutations, we selected six binding site mutations for experimental validation (Supplementary Data 10). These sites were chosen based on the number of times the specific binding site was identified in the AGO-CLIP atlas, the location of the seed in the coding region or 3′-UTR, whether TargetScan called the site as highly conserved and the relative location of the mutation within the seed-complementary region of the binding site. This selection included three mutations in 3′-UTR sequences and three mutations in coding sequences to capture the diversity of the miSNP analysis (Supplementary Data 10). Four of six tested mutations annotated and we incorporated 36 COAD WGS -UTR-3′ mutations from 36 COAD WGS samples.

Figure 7 | A model of Pan-cancer suppressor pathway regulation by miR-17-19-130 superfamily as defined by AGO-CLIP analysis. The microRNA-17-19-130 superfamily heavily targets critical TS in multiple pathways, including the TGFβ pathway, the phosphoinosside 3-kinase/AKT pathway and the PS3 pathway. Additional target site mutation analysis reveals ablation of a miR-17-mediated negative feedback loop through mutation of the miR-17 binding site on the SKIL 3′-UTR, demonstrating a novel mechanism of tumour escape from microRNA regulation.

Discussion

Combinaltorial definition of high-confidence microRNA binding sites using multiple transcriptome-wide AGO-CLIP data sets generated clear evidence of endogenous microRNA binding at specific locations on an mRNA strand. Points of microRNA interaction tested in this study included multiple microRNA binding sites, such as the miR-17-SKI-3′-UTR binding site and binding sites in coding sequences of the mRNA, that are difficult to detect through other means of microRNA target site prediction.

We found that 45% of broadly conserved pan-cancer oncomiRs share strong homology in their seed motifs. Seed similarity leads to redundant cotargeting, and therapeutic suppression of any one of these microRNAs is likely to face compensation from other members of the superfamily. The 3′-UTR–luciferase binding assays and anticorrelates of microRNA–target pairs support the possibility that these microRNAs redundantly cotarget important TS across multiple tumour types.

The ability of super-seed target sites to bind multiple members of the oncogenic superfamily make them an attractive therapeutic candidate in the future, because it may effectively act as a microRNA ‘sponge’ that can bind and titrate off multiple superfamily members to restore normal cellular regulation in cancer cells by de-repressing critical TS. This therapy may prove more effective than targeting a single oncomiR family, because it has the potential to concurrently sequester multiple oncogenic microRNA seed families to disrupt redundant oncogenic co-repression of TS.

Using the miSNP algorithm, we identified thousands of mutations in microRNA binding sites. These mutations were discovered in the exome-sequencing-defined coding-region mutations available from the Pan-Cancer project, a small number of 3′-UTR mutations available from KIRC exome sequencing and whole-genome 3′-UTR mutations from 36 COAD WGS samples. AGO-CLIP characterization of microRNA binding in additional tissue types and integration of additional 3′-UTR mutations from broader cohorts of WGS samples will improve the yield of relevant microRNA target site mutations in the future.
In conclusion, we generated a novel resource, the AGO-CLIP atlas, to integrate experimentally defined microRNA binding sites with TCGA tumour data, creating a new method and framework to understand microRNA regulation of cancer. This method addresses the difficult question of determining accurate genome-wide microRNA–target interactions. Integration of the AGO-CLIP-defined microRNA-binding data with TCGA tumour data revealed several novel insights into microRNA regulation of human tumours, including the definition of a pan-cancer oncomiR superfamily and genome-wide identification of microRNA-binding site mutations.

**Methods**

**Argonaute Crosslinking Immunoprecipitation.** AGO-CLIP sequence read archive (SRR) files corresponding to all publicly available human AGO-CLIP experiments were downloaded from the NIH sequence read archive (SRR codes: SRR048973, SRR048975, SRR048976, SRR048977, SRR048978, SRR048979, SRR048980, SRR048981, SRR359787, SRR189786, SRR189787, SRR189784, SRR189785, SRR189783, SRR08362, SRR08363, SRR08352, SRR80353, SRR80354, SRR80355, SRR80359, SRR80360, SRR80361, SRR80356, SRR80357, SRR43334, SRR433334, SRR433335, SRR92689, SRR92688, SRR92687, SRR92686, data were downloaded on 11 January 2013). Files were individually pre-processed using Fastx toolkit and cut-adapt to remove adaptor sequences and control for sequence quality. Data set quality was individually discerned using Fast QC reader. Individual sequencing runs were tiered and grouped, based first on group publishing, or cell line used, then on individual treatment and finally based on total reads. In this manner, 14 independent AGO-CLIP data sets were defined (summarized in Supplementary Data 1). Individual cell lines with correspondent AGO-CLIP data include: 293T (4/14 experiments), hESC (1/14 experiments), BCBL1 (1/14 experiments), BC-3 (1/14 experiments), BC-1 (1/14 experiments), LCL35 (1/14 experiments), LCL-BAC (1/14 experiments), LCL-BAC-EV3-infected (2/14 experiments) and EF3D (1/14 experiments). Eleven of these data sets were AGO-PAR-CLIP experiments. Three were AGO-HITS-CLIP experiments. Reads from both experiment types were mapped to hg19 using established Bowtie parameters.

Both HITS-CLIP and PAR-CLIP methods utilize ultraviolet crosslinking of RNA-binding proteins to their respective RNA partners, followed by protein immunoprecipitation and high-throughput sequencing of the bound RNA. The difference between the two methods lies in the use of photoactivatable ribonucleic acid (PAR-CLIP) data sets. In PAR-CLIP, the determination of physical interlinkage between protein–RNA pairs through a mismatch repair defect initialized at crosslinked nucleotide analogues during complementary DNA synthesis, leading to T→C transitions in the generated cDNA. Of the two, the majority of human data sets are PAR-CLIP generated, and an established pipeline was used for processing these PAR-CLIP data sets. A pipeline algorithm, termed PARalyzer, uses a kernel-density algorithm centred on crosslinks to generate putative microRNA target sites. In the PARalyzer algorithm, reads are first processed into read groups based on total number of reads. T→C transitions in read groups are then used to define clusters based on the kernel-density algorithm. Clusters are used for analysis for complexes of microRNA seeds to infer the identity of the microRNA binding partner. PAR-CLIP Bowtie files were processed through the PARalyzer57 algorithm to generate clusters and seeds using established parameters.

AGO-HITS-CLIP Bowtie files were also processed through PARalyzer and groups were noted to be AGO-HITS-CLIP targets. These groups were superimposed over microRNA target sites identified by the PAR-CLIP reads. In this way, AGO-CLIP data sets were allowed to support the strength of a target site identified in the PAR-CLIP runs by contributing to site recurrence, but were not allowed to perform de novo target site identification. Meta-analyses concerning the location of clusters and their density on various segments of the transcriptome (5′-UTR, coding sequence and 5′-UTR) were performed using only the 11 AGO-PAR-CLIP libraries. A lenient seed inference strategy was used, which included all miRBase seed families. The purpose of this lenient mapping was to anchor redundant read clusters to the genome to determine site-seed coregulation. Using this strategy, 99% of 123,752 PAR-CLIP cluster mapping to the UCSC known gene transcriptional site received at least one seed inference, although likely to be at the expense of false positives in less-expressed microRNAs. From these cluster sequences, 306,733 microRNA seed-complementary sequences were inferred. Multiple seed combinations may be inferred by TCGA data available from the MusCic, MSKCC driver target and GISTIC algorithms. This system equally weighted CNV, mRNA expression change and gene mutations as three orthogonal methods of identifying TS and OCs across tumours. This method generated a continuous ranked list for every gene, based on consistent changes across tumours, ranging from more negative (TS) to more positive (OC).

For mRNA data, +1 point was given for each of seven tumours with microRNA-seq data available, in which there was a tumour versus normal expression change and gene mutations as three orthogonal methods of identifying TS and OCs across tumours. This method generated a continuous ranked list for every gene, based on consistent changes across tumours, ranging from more negative (TS) to more positive (OC).
significant increase (Student’s t-test, \( q < 0.005 \)), and \( -1 \) point was assigned for significant decreases in mRNA expression. For CNV data, GISTIC scores for HUGO-gene-annotated locus copy-number changes were used. We set an amplification/deletion threshold of 0.3 or \(-0.3\) for each sample. For each whitened tumour in which a given gene achieved amplification or in 30% of samples, \(+0.5\) or \(-0.5\) points was awarded accordingly. Finally, for mutation scores, MSKCC and MuSiC mutation analyses were both integrated. Gene mutations were only considered based on MuSiC-determined Fisher’s combined P-test FDR \( q < 0.005 \). Mutation frequency was multiplied by 100 and then by 1 or \(-1\), based on MSKCC driver analysis as a TS (\( +1 \)) or an OG (\( -1 \)). Genes nominated as both TS and OG are negated. Truncating mutations based on MSKCC truncation tabulation were assigned additional significance, and the fraction of truncations/total mutations was multiplied by \(-5\) to attribute additional negative value to any gene with high frequency of truncating mutations. All scores were then summed to generate a final pan-cancer TS versus OC score. In sum, this analysis generated a continuous negative-to-positive scale that ranked pan-cancer drivers based on consistent mRNA, CNV or mutation changes across tumours. These three orthogonal technologies used in the analysis. The scoring equation is described below:

\[
\text{Final score} = (\text{mRNA increases} - (\text{mRNA decreases} + (0.5) \text{(CNV amplification)} - (0.5) \text{(CNV deletion)} + (100) \text{mutation frequency across all tumours})
\]

\[ (\pm 1 \text{ MSKCC drivers}) - (5) \text{(truncation frequency)} \]

\[ n = 1 - 3,000; \]

\[ x(TS \text{ targets per microRNA/targets per microRNA}) \]

\[ x(\text{OC targets per microRNA/targets per microRNA}) \]

\[ \text{Targeting representation was performed with the Venn Diagram package in R.} \]

**MicroRNA–target enrichment calculations.** To define an optimal method of determining microRNA–target interactions, we intersected pan-cancer oncomiRs and miR suppressors with pan-cancer TS versus pan-cancer OCs based on four different possible approaches that included the following: using all AGO-CLIP-defined binding sites without considering site conservation (for example, TargetScan), using only AGO-CLIP-defined sites with \( \geq 3 \) occurrences (corresponding to a significant peak based on random permutation) without considering TargetScan, TargetScan-only binding sites, and finally by combining AGO-CLIP-defined target sites with \( \geq 3 \) occurrences, or \( \geq 1 \) occurrences and a TargetScan call (Supplementary Fig. S2). Only well-conserved TargetScan calls were considered in this analysis. To calculate enrichments, the per cent of total targets per microRNA defined as TS as compared with the per cent of total number of targets defined as OCs for all genes in the top 100, 250, 500, 1,000, 1,500, 2,000, 2,500 and 3,000 TS versus OCs (see Supplementary Fig. S2 for all levels of data and Supplementary Data 6 for complete mRNA driver analysis).

The average per cent of TS versus OC targets was compared for oncomiRs and miR suppressors using Student’s t-test based on the following equation:

For \( n \) number of OC or TS ranked in the top 3,000 (Supplementary Data 6), where \( n = 1 - 3,000; \) \( x(TS \text{ targets per microRNA}/\text{targets per microRNA}) \)

\[ x(\text{OC targets per microRNA}/\text{targets per microRNA}) \]

\[ \text{Correlation} \]

\[ \text{representation was performed with the Venn Diagram package in R.} \]

**MicroRNA–mRNA binding results and RNA-seq gene expression data across multiple TCGA data sets.** miSNP performs two types of integration. First, by using only the mutation data and the AGO PAR-CLIP microRNA/mRNA binding sites, miSNP aggregates and reports at gene level both the microRNA binding sites and the mutations. Data collected for a particular gene include the individual microRNAs targeting the gene, as well as the types of mutations in the microRNA binding sites. Next, miSNP incorporates RNA-seq gene expression data to enable a user to perform a quantitative analysis of the effects of microRNAs and mutations on gene expression across a TCGA data set. The algorithm operates on a gene-by-gene basis. It first partitions the tumour samples into those that contain mutations in microRNA binding sites; the algorithm can be customized to consider only individual mutation types (i.e., only coding, only splicing) or both. Next, it reports the gene expression data for these genes and samples in a tabular format for further analysis. Finally, it identifies genes for which the expression associates with the mutation status in microRNA binding sites by comparing the gene expression distributions of tumour samples with or without common sites using a two-tailed Welch’s t-test; miSNP reports both the fold-change and the t-test P-value for each gene. miSNP was developed in Python utilizing the NumPy and Scipy modules. In the current paper, we analysed all AGO-CLIP-defined microRNA target sites for mutations to define a global perspective of possible interactions, but selected sites for validation only from interactions with \( \geq 3 \) occurrences corresponding to a non-random event.

The miSNP software package (Supplementary Software 1) is an open source (Free BSD license) and is available for community download at: www.genboree.org/miSNP.

**Luciferase assays.** The PTEN 3‘-UTR–luciferase reporter constructs were purchased from Addgene (Cambridge, MA). All other constructs were purchased directly from Switchgear Genomics (Palto Alto, CA) and generated using LightSwitch 3‘-UTR vectors. MicroRNA MirVana mimics and negative control mimics were purchased from Ambion (Grand Island, NY). Luciferase assays were performed using the LightSwitch Switchgear LightSwitch luciferase assay kit protocols. Reporter plasmids and microRNA mimics were transfected using Lipofectamine 2000 (Life Technologies, Grand Island, NY). Luciferase activity was normalized to \( \beta \)-galactosidase activity. For coding-region constructs, synthetic binding assays were generated by placing the coding region downstream of the luciferase reporter in the UTR. All binding assays were performed in confluent HEK293T cells for 24 h at 5 and 10 nM concentrations microRNA mimic and 40 ng of luciferase and \( \beta \)-galactosidase vector. Experiments were replicated twice at 5 nM and twice at 10 nM experimental concentrations in quadruplicate, and results were combined for final statistical analysis. Insert sequences novel to this study are available in Supplementary Data 11. HEK293T cells were a kind gift from Dr Weiven Long.

**References**


Acknowledgements

We gratefully acknowledge the contributions from the TCGA Research Network and its TCGA Pan-Cancer Analysis Working Group (contributing consortium members are listed in the Supplementary Note 1). The TCGA Pan-Cancer Analysis Working Group is coordinated by J.M. Stuart, C. Sander and I. Shmulevich. This work was supported by the Caroline Wiss Law Foundation (S.E.M); Dan L. Duncan Cancer Center Scholar Award (S.E.M); S.E.M. is a member of the Dan L. Duncan Cancer Center supported by the National Cancer Institute Cancer Center Support Grant P30CA125123. We acknowledge the joint participation by the Diana Hela Henry Medical Research Foundation through its direct engagement in the continuous active conduct of medical research in conjunction with Baylor College of Medicine Baylor Research Advocates for Student Scientists (BRASS) Foundation (M.P.H); the Robert and Janice McNair Foundation (M.P.H); and NIH IK1DK096093 (S.M.H) with additional funding provided by the Diabetes and Endocrinology Research Center (P30-DK079638) at Baylor College of Medicine. A special thanks to Kat Harris and the Switchgear Genomics team for fast, efficient and kind service. We thank Robb Moses, David Bader and Joel Neison for editing contributions.

Author contributions


Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/


This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/