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-130, 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.
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 Argonaute proteins, microRNAs bind target mRNAs in the RNA-induced silencing complex. MicroRNA target regulation is canonically mediated by nucleotides 2–8 on the 5′-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.

Argonaute Crosslinking ImmunoPrecipitation (AGO-CLIP) data sets experimentally identify microRNA–target interactions in a genome-wide manner through purification of Argonaute-protein-associated RNAs, which include bound microRNAs and their respective targets. 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, glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), colon and rectal carcinoma (COAD, READ), bladder urothelial carcinoma (BLCA), kidney renal clear cell carcinoma (KIRC), ovarian serous cystadenocarcinoma (OV), uterine corpus endometrial carcinoma (UCEC), and acute myeloid leukemia (LAML), 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 targets 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,b, 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 Argonaute protein (Fig. 2a). AGO Photoactivatable-Ribonucleoside-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).17,18,20–24 This included 11 AGO-PAR-CLIP libraries and 3 unmodified AGO-CLIP libraries (also called Argonaute High-Throughput Sequencing of RNA Isolated by CLIP (AGO-HITS-CLIP)).16 The AGO-CLIP atlas allowed 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 (Fig. 2d).

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 Argonaute 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 Argonaute 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 >50% 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 ($R^2 = 0.86$, $P = 0.01$, $r_{100%} = 0.90$, $r_{90%} = 0.86$, $r_{80%} = 0.80$, $r_{70%} = 0.70$, $r_{60%} = 0.60$, $r_{50%} = 0.50$, $r_{40%} = 0.40$, $r_{30%} = 0.30$, $r_{20%} = 0.20$, $r_{10%} = 0.10$, $r_{0%} = 0.00$).

**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 oncogenes (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 $\geq 3$ occurrences (corresponding to a significant peak based on random permutation) without considering TargetScan; using TargetScan-only binding sites (without considering AGO-CLIP data); and, finally, combining AGO-CLIP-defined target sites with $\geq 3$ occurrences, or $\geq 1$ 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).

Table 1. Log2 expression change in tumour suppressor (TS) and oncogene (OC) targets for pan-cancer oncomiRs and miR suppressors

| Gene     | BRCA        | HNSC        | HCC         | KIRC        | MSKCC       | UCEC        |
|----------|-------------|-------------|-------------|-------------|-------------|-------------|
| hsa-mir-183 | 2.55        | 2.77        | 1.48        | 0.01        | 2.39        | 2.26        | 4.63        |
| hsa-mir-182 | 1.88        | 2.03        | 1.02        | 0.47        | 2.03        | 1.61        | 4.03        |
| hsa-mir-96  | 2.67        | 2.93        | 2.18        | 2.15        | 2.43        | 2.29        | 4.09        |
| hsa-mir-210 | 3.00        | 3.27        | 3.41        | 3.41        | 3.01        | 4.17        | 3.09        |
| hsa-mir-425 | 1.45        | 0.96        | 0.46        | 2.54        | 1.99        | 1.32        | 2.84        |
| hsa-mir-136b| 2.64        | 1.69        | 1.42        | 1.35        | 1.27        | 1.79        | 1.08        |
| hsa-mir-15a | 2.67        | 1.42        | 1.29        | 1.25        | 0.97        | 2.66        | 2.38        |
| hsa-mir-33 | 1.34        | 0.10        | 1.17        | 1.31        | 0.67        | 4.13        | 1.77        |
| hsa-mir-17  | 1.60        | 1.14        | 1.02        | 0.72        | 2.63        | 0.58        | 1.54        |
| hsa-mir-106a| 1.55        | 1.21        | 1.44        | 0.93        | 0.89        | 1.55        | 1.71        |
| hsa-mir-23b | 1.00        | 0.99        | 1.09        | 0.93        | 1.07        | 1.71        | 2.13        |
| hsa-mir-33b | 3.89        | 3.22        | 2.73        | 1.67        | 2.85        | 3.10        | 2.76        |
| hsa-mir-192 | 2.08        | 1.50        | 1.05        | 2.00        | 2.00        | 2.00        | 1.54        |
| hsa-mir-142 | 1.60        | 0.96        | 0.18        | 1.13        | 0.66        | 1.13        | 1.27        |
| hsa-mir-31a | 1.46        | 1.04        | 1.05        | 1.26        | 1.44        | 1.44        | 1.38        |
| hsa-mir-97a | 1.52        | 0.86        | 0.86        | 1.02        | 1.04        | 0.66        | 0.86        |
| hsa-mir-95 | 3.33        | 2.95        | 1.05        | 1.84        | 2.55        | 2.66        | 1.29        |
| hsa-mir-590 | 1.44        | 1.27        | 1.15        | 0.76        | 1.26        | 1.45        | 0.97        |

Figure 3 | Determination of pan-cancer microRNAs and their targets. (a) List of broadly conserved pan-cancer oncomiRs and miR suppressors reveals microRNAs undergoing consistent expression changes across the pan-cancer data set. (b) A dual nomination strategy uses AGO-CLIP microRNA target definitions to associate pan-cancer microRNAs to respective TS and OC targets to define relevant pan-tumour microRNA-target relationships. (c) TS target versus oncomiR target enrichments for pan-cancer oncomiRs (blue bars) and miR suppressors (red bars) across the top 100–3,000 (~10% of genes) TS and OCs. This graph represents the enrichments and significance levels of those enrichments of both pan-cancer oncomiRs and miR suppressors. Individual bars in the graph represent the log2-based per cent TS over per cent OC targets. Positive bars demonstrate average total enrichment for TS. Negative bars demonstrate average enrichment for OCs. An asterisk denotes significant enrichments. Red box highlights enrichment of pan-cancer oncomiRs with their top 250 targets used in subsequent analysis. Student’s t-test, *P < 0.05, **P < 0.005. N-values for enrichment reflect the total number of pan-cancer oncomiRs (n = 22) and pan-cancer miR suppressors (n = 25).

AGO-CLIP-defined targets within the 3’-UTR but not called by TargetScan (results are summarized in Supplementary Table S3).

TARGETSCAN may produce bias towards TS, because the top 3,000 OCs have 31% more AGO-CLIP clusters binding them than the top 3,000 OCs. This observation may reflect overexpression of these OCs in the rapidly growing or are oncogenic, the relative lack of AGO-CLIP clusters on the TS may reflect the cellular context of the 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...
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 oncomiR1 (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-17 targets, 79.3% of mir-130/301ab targets, 39.2% of mir-17/106a/93 targets, 42.3% of mir-18a targets, 75.7% mir-455 targets and 62.5% of mir-210 targets (Fig. 4c).

The entire mir-17-19-130-93-455-18-210 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 TGFB2, across all tumour types. Figure 5a demonstrates correlations with the top four highest-ranked TS (PTEN, TGFB2, TGFB2 and SMAD4) that are targeted by all three seed families. PTEN, TGFB2 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 correlations heat map for the microRNA–target pairs 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, TGFB2 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 TGFB2 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 oncogenic phosphoinositide 3-kinase pathway33. TGFB2 and SMAD4 are tumour-suppressive components of the transforming growth factor-β (TGFβ) pathway34. ZBTB4 is described as a mediator of the p53 response35. 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 GUGC 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 TS for mir-17, -19, -130, -210, -18 and -455 families demonstrates TS coregulating relationships with 39.2% of mir-17/106a/93 targets, 70.2% of mir-19 targets, 79.3% of mir-130/301ab targets, 42.3% of mir-18a targets, 75.7% mir-455 targets and 62.5% of mir-210 targets having predicted coregulating with at least one other superfamily member. The mir-93 family has largely the same predicted target spectrum as that of the mir-93 family. (d) miR-17-19-130 superfamily members are strongly correlated across pan-cancer tumours, demonstrating potent coregulation of these microRNAs (P < 1E-200, Student’s t-test). MicroRNAs are colour-coded based on co-localization to the same genomic cluster. microRNA–microRNA correlate P = 0.796.
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 superfamily 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 cotargeting 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 5nM 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.

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 variations27,39. This process, however, remains imperfect. Many missense mutations are not characterized as contributing significant functional impact on a gene. Further, large percentages of coding region mutations are silent. Finally, the number of mutations outside the coding region of genes
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).

Regulation of the selected site, silencing luciferase expression by demonstrating strong evidence of microRNA binding and miSNP analysis (Supplementary Data 10). Four of six tested mutations in coding sequences to capture the diversity of the a novel mechanism of tumour escape from microRNA regulation.

HIST1H3B

0

the seed-complementary region of the binding site. This selection highly conserved and the relative location of the mutation within coding region or 3′-UTR, whether TargetScan called the site as strongly reduced, but did not ablate, variably reduced, but did not ablate, and demonstrate the ability to tier the probable functional impact only because we utilize the AGO-CLIP technology that we are able to define these microRNA target-site mutations at all. In addition, KIRC exome sequencing has a small number of 3′-UTR mutations annotated and we incorporated 36 COAD WGS samples that contain full 3′-UTR SNV annotations.

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 microRNA binding sites with corresponding somatic mutations demonstrated strong evidence of microRNA binding and regulation of the selected site, silencing luciferase expression by >40% in each case and strongly suggesting our analysis identifies functional microRNA binding sites (Fig. 6c).

For the four target sites with validated target repression, we reproduced the endogenous tumour mutation in the 3′-UTR–luciferase construct. The binding site SNVs corresponding to the miR-17 seed on the SKIL 3′-UTR and the miR-9 seeds on the HIST1H3B 3′-UTR were able to ablate microRNA binding. Mutations complementary to the Let-7 seed on ANP32E and miR-142 seed on FAM114A1 variably reduced, but did not ablate, the repressive ability of the microRNA on the luciferase reporter (Fig. 6c). The ability of a mutation to ablate microRNA binding was directly related to the relative position of the mutation within the region complementary to the microRNA seed. Mutations in the first or last nucleotide of the seed complement had a reduced ability to ablate binding relative to mutations near the centre of the seed complement (Fig. 6d). These observations are consistent with established concepts of microRNA binding and demonstrate the ability to tier the probable functional impact of microRNA binding site mutations based on its location within the seed-complementary region of the mRNA target.

One microRNA target site mutation validated in our study was a deletion of a miR-17 seed family binding site in the SKI-like OC (SKIL/SnoN) 3′-UTR. SKIL is a known OC and was ranked in the top 7% of pan-cancer OCs in our pan-cancer mRNA driver index due to expression gain and copy number amplification (Supplementary Data 6). SKIL oncogenic function is known to occur through direct repression of the TGFβ signalling pathway, and part of the TGFβ signalling pathway activation involves targeting and degradation of the SKIL protein41–43.

Targeting of the SKIL 3′-UTR by the miR-17 seed family reveals potential antiautoregulatory feedback that can attenuate silencing of the TGFβ pathway by the miR-17-19-130 superfamily. Mutation of the miR-17 seed-family binding site on the SKIL 3′-UTR may represent a mechanism to allow escape from this feedback regulation, allowing unregulated SKIL expression while simultaneously enhancing the oncogenicity of the miR-17 seed family and creating enhanced suppression of the TGFβ pathway (Fig. 7). The miR-17 seed-family binding site on the SKIL 3′-UTR is identified in 8 out of 14 AGO-CLIP data sets, indicating that the site itself is highly active endogenously. Despite this strong evidence of microRNA binding in the AGO-CLIP atlas, TargetScan, Pictar and MiRanda, motif calling algorithms do not nominate SKIL as a potential target of the miR-17 family, again highlighting the value of unbiased genome-wide binding assays as a useful method of determining active microRNA seeds.

Discussion

Combinatorial 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-SKIL 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’46 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 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 NHI sequence read archive (SRR codes: SRR048973, SRR048975, SRR048976, SRR048977, SRR048978, SRR048979, SRR048980, SRR048981, SRR359787, SRR189786, SRR189787, SRR189784, SRR189785, SRR189783, SRR080362, SRR080363, SRR080352, SRR030533, SRR030534, SRR030535, SRR030539, SRR03060, SRR030631, SRR030356, SRR030357, SRR433334, SRR433336, SRR433335, SRR392689, SRR392688, SRR392687, SRR926865; 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, on 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), HeLa (1/14 experiments), BCBL1 (1/14 experiments), U2OS (1/14 experiments), HeLa (1/14 experiments), BCBL1 (1/14 experiments), LCL-BAC (1/14 experiments), LCL-BAC-EBV-infected (2/14 experiments) and EFSD (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.17 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 ribonuclease antago-AGO-CLIP data sets allow a determination of physical interlinkage between protein–RNA pairs through a mismatch repair defect initialized at crosslinked nucleoside 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 for processing these PAR-CLIP data is provided.48 This 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. Cluster underpinning analysis for complements of microRNA seeds to infer the identity of the microRNA binding partner. PAR-CLIP Bowtie files were processed through the PARalyzer algorithm to generate clusters and seeds using established parameters.17 AGO-HITS-CLIP Bowtie files were also processed through PARalyzer and grouped to generate AGO-HITS-CLIP 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 a novel target site identification. Meta-analyses concerning the location of clusters and their density on various segments of the transcriptome (3′-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 seed-site recurrence across all data sets. Using this strategy, 99% of 123,752 PAR-CLIP clusters mapping to the UCSC known gene transcriptome 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 complements may be inferred by single clustered seed sites and these seeds often overlap a single site. These sites often highlight putative microRNA super-seed targets, which are readily apparent in AGO-CLIP data. The identity of the actual binding partner may be one of the complementary microRNAs, all of them, or may represent a form of non-canonical binding that is not currently considered in our motif analysis.20,21 Following target identification, the seed complements of each target were grouped for recurrence across the 11 PAR-CLIP data sets. To this, 3 HITS-CLIP read groups were intersected to combine data from 14 total AGO-CLIP sources. PAR-CLIP clusters and HITS-CLIP groups were then permuted across the genome 20 times using BedTools49 and analysis was performed to determine the likelihood of a given target being recurrently identified by chance across randomization. A FDR was assigned based on binomial P-values established from the actual measured probability of a seed-complementary site recurring at random based on random distribution of the target sites across the transcriptome. We determined target site recurrence of three or more corresponded to a Q-value <0.05.

**TCGA data acquisition.** All TCGA data, except for microRNA expression, were downloaded from the Synapse archive as part of the TCGA Pan-Cancer project and correspond to TCGA pan-cancer whitelist files (originally downloaded 23 January 2013). The TCGA Pan-Cancer project consists of 12 available tumour types that include COAD, READ, LUAD, LUSC, BLCA, BRCA, GBM, UCSC, KIRC, LAML, OV and HNSC. Some components of available data are currently incomplete, such as missing normal microRNA-sequencing samples for LAML, ovarian carcinoma, GBM, COAD and READ.

MicroRNA data were compiled individually from the TCGA data portal to analyse a range of factors including expression, mutation data and survival. MicroRNA data were processed directly from the TCGA data portal isoform files for all whitelist tumours as of 20 November 2012. Multiple reads from an individual isoform were collapsed into a single read count; the reads per million microRNAs mapped data form was used, which establishes each microRNA read count as a processing of the total microRNA population. MicroRNA-sequencing data used in this study are summarized in Supplementary Table S1. MicroRNA data underwent upper-quartile normalization50 using the edgeR software package51, which produced the best overall normalization results compared with reads per million or trimmed median of M-values normalizaion, followed by determination of significant differences between tumour and normal samples using a Fisher’s exact test with Bonferroni correction to determine FDRs. As part of the Pan-Cancer project, some processed data were available for second-line analysis. These data included pan-cancer CNV data processed through the ABSOLUTE-GISTIC52 pipeline and mutation data processed through the MSKCC driver analytical pipeline, which were incorporated into driver gene nominations. A list of data IDs in Synapse is provided in Supplementary Table S4.

**Pan-cancer oncomiR and miR-suppressor selection.** Our goal in nominating pan-cancer oncomiRs and miR suppressors was to determine microRNAs that change consistently in the same direction across most cancers. Thus, we used a stringent (q <0.05, Fisher’s exact test with Bonferroni correction for multiple testing) threshold comparing tumour versus normal microRNA expression, and required pan-cancer oncomiRs to have increased expression in six out of seven tumour types with available tumour versus normal data. Alternately, pan-cancer miR suppressors were required to have decreased expression in six out of seven tumour types. Finally, when determining microRNAs to analyse for microRNA–target interactions, we additionally filtered for dominant isoforms in browsed microRNA families and microRNA families that peak identified the 14 AGO-CLIP data sets. This helped limit false positives in the microRNA–target analysis by ensuring the interactions we observed consisted of conserved microRNA seeds derived from microRNAs expressed in the AGO-CLIP cell lines.

**T5 and QC definitions.** This analysis utilized three external data sets generated for the purpose of pan-cancer analysis by the TCGA: MuSiC, MSKCC driver target analysis and GISTIC. The strength of the MuSiC algorithm, developed at Washington University28, is its ability to quality-control SNV samples, eliminate outliers (such as certain hypermutated samples) and derive P-values to determine significantly mutated genes versus the background mutation rate. We thus utilize MuSiC P-values and mutation frequencies in our analysis. Specific mutations are likely to either activate or inactivate genes, and definition of these mutation sites in a single gene can ultimately define that gene as T5 or T. The MSKCC driver algorithm defines SNVs that we are able to use to stratify mutated genes as either OCs or TS based on a functional impact score that weighs the probable impact of mutation at a specific amino acid residue. Finally, GISTIC20, developed at the Broad Institute, is an algorithm that controls for sample quality and low-amplitude copy number shifts in CNV. We processed GISTIC data to define CNV log ratios and set CNV thresholds. Our mRNA q-values are performed in house and set at stringent, common threshold controls (q <0.05, Student’s t-test with Bonferroni correction).

A ranking system was developed, which integrates TCGA mRNA, CNV and mutation data. T5 threshold comparing tumour versus normal microRNA expression, and gene expression changes as gene mutations as three orthogonal methods of identifying T5 and OCs across tumours. This method generated a continuous ranked list for every gene, based on consistent changes across tumours, ranging from more negative (T5) to more positive (OCs).

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
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 whitest 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, MSKKC 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 MSKKC driver analysis as a TS $(+1)$ or an OG $(+1)$. Genes nominated as both TS and OG are negated. Truncating mutations based on MSKKC 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 values were scaled to roughly equal weight to place equal emphasis on the three orthogonal technologies used in the analysis. The scoring equation is described below:

$$\text{Final score } = \left( \frac{\text{mRNA increases}}{\text{mRNA decreases}} \right) + \left( \frac{\text{CNV amplification}}{\text{CNV deletion}} \right) + \left( \frac{\text{mutation frequency across all tumours}}{ \pm 1 \, \text{MSKKC drivers} } \right)$$

(1)

There are six tumours with tumour versus normal mRNA (BLCA, BRCA, HNSC, KIRC, LUAD and LUSC). All tumours contain CNV data and mutation data. Mutation frequency must be significant ($q < 0.005$, Fisher’s exact test) to be considered at all. The MSKKC mutation analysis assigns overall scoring. Mutation score is highly dominant in several genes such as $\text{TGFBR2}$, $\text{SMAD4}$ and $\text{PTEN}$ in $30\%$ of samples, $\pm 0.3$ for each sample. For each whitelist $P_{\text{total targets per microRNA}}$ was developed in Python utilizing the Numpy and Scipy modules. In the current study, 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 $\pm 3$ occurrences corresponding to a non-random event.

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 microRNA driven 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:

$$x(TS \text{ targets per microRNA}/\text{total targets per microRNA})$$

versus

$$x(OC \text{ targets per microRNA}/\text{total targets per microRNA})$$

(2)

$\text{Co-targeting representation was performed with the Venn Diagram package in R.}$

MicroRNA pan-cancer correlations. Two sets of correlations were used in this study. The first was microRNA to microRNA correlation for miR-17-19-130 family members identified as pan-cancer oncomiRs. These correlations consist of a simple microRNA-to-microRNA Pearson’s $R^2$ value. To generate a null distribution, 100 mature, dominant isoform microRNAs were randomly selected and 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 MSKKC driver analysis as a TS $(+1)$ or an OG $(+1)$. Genes nominated as both TS and OG are negated. Truncating mutations based on MSKKC 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 values were scaled to roughly equal weight to place equal emphasis on the three orthogonal technologies used in the analysis. The scoring equation is described below:

$$\text{Final score } = \left( \frac{\text{mRNA increases}}{\text{mRNA decreases}} \right) + \left( \frac{\text{CNV amplification}}{\text{CNV deletion}} \right) + \left( \frac{\text{mutation frequency across all tumours}}{ \pm 1 \, \text{MSKKC drivers} } \right)$$

(1)

There are six tumours with tumour versus normal mRNA (BLCA, BRCA, HNSC, KIRC, LUAD and LUSC). All tumours contain CNV data and mutation data. Mutation frequency must be significant ($q < 0.005$, Fisher’s exact test) to be considered at all. The MSKKC mutation analysis assigns overall scoring. Mutation score is highly dominant in several genes such as $\text{TGFBR2}$, $\text{SMAD4}$ and $\text{PTEN}$ in $30\%$ of samples, $\pm 0.3$ for each sample. For each whitelist $P_{\text{total targets per microRNA}}$ was developed in Python utilizing the Numpy and Scipy modules. In the current study, 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.

Luciferase assays. The PTEN 3′-UTR–luciferase reporter constructs were purchased from Addgene (Cambridge, MA)26,27. All other constructs were purchased directly from Switchgear Genomics (Palo Alto, CA) and generated using the 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 β-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 β-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

1. Selbach, M. et al. Widespread changes in protein synthesis induced by microRNAs. Nature 455, 58–63 (2008).
2. Mukherji, S. et al. MicroRNAs can generate thresholds in target gene expression. Nat. Genet. 43, 854–859 (2011).
3. Guo, H., Ingolia, N. T., Weissman, J. S. & Bartel, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature 466, 835–840 (2010).
4. Lu, J. et al. MicroRNA expression profiles classify human cancers. Nature 435, 834–838 (2005).
5. Martello, G. et al. A microRNA targeting dicer for metastasis control. Cell 141, 1195–1207 (2010).
6. Kumar, M. S., Lu, J., Mercer, K. L., Golub, T. R. & Jacks, T. Impaired microRNA processing enhances cellular transformation and tumorigenesis. Nat. Genet. 39, 673–677 (2007).
7. Volinia, S. et al. Reprogramming of miRNA networks in cancer and leukemia. Genome Res. 20, 589–599 (2010).
8. Darido, C. et al. Targeting of the tumor suppressor GRHL3 by a miR-21-dependent proto-oncogenic network results in PTEN loss and tumorigenesis. Cancer Cell 20, 635–645 (2011).
9. Olive, V. et al. miR-19 is a key oncogenic component of mir-17-92. Genes Dev. 23, 2839–2849 (2009).
10. Conkrite, K. et al. miR-17–92 cooperates with RB pathway mutations to promote retinoblastoma. Genes Dev. 25, 1734–1745 (2011).
11. Le从此开始, D. P. MicroRNAs: target recognition and regulatory functions. Cell 136, 215–233 (2009).
12. Wang, Y. et al. Structure of an argonaut silencing complex with a seed-containing guide DNA and target RNA duplex. Nature 456, 921–926 (2008).
13. Friedman, R. C., Farh, K. K., Burge, C. B. & Bartel, D. P. Most mammalian microRNAs act at conserved targets of microRNAs. Genome Res. 19, 92–105 (2009).
14. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120, 15–20 (2005).
15. Grimson, A. et al. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol. Cell 27, 91–105 (2007).
16. Chi, S. W., Zhang, J. B., Mele, A. & Darnell, R. B. Argonaute HITs-CLIP decodes microRNA-mRNA interaction maps. Nature 460, 479–486 (2009).
17. Hafner, M. et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 141, 129–141 (2010).
18. Hafner, M., Lianoglou, S., Tuschl, T. & Betel, D. Genome-wide identification of miRNA targets by PAR-CLIP. Methods 58, 94–105 (2012).
19. The Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. 45, 1113–1120 (2013).
20. Skalsky, R. L. et al. The viral and cellular microRNA targeted in lymphoblastoid cell lines. PLoS Pathog. 8, e1002484 (2012).
21. Lipchina, I. et al. Genome-wide identification of microRNA targets in human ES cells reveals a role for miR-302 in modulating BMP response. Genes Dev. 25, 2173–2186 (2011).
22. Gottwein, E. et al. Viral microRNA targetome of KSHV-infected primary effusion lymphoma cell lines. Cell Host Microbe 10, 515–526 (2011).
23. Kishore, S. et al. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nat. Methods 8, 559–564 (2011).
24. Haecker, I. et al. AGO-CLIP expands understanding of Kapoosi sarcoma-associated herpesvirus miRNA function in primary effusion lymphomas. PLoS Pathog. 8, e1002884 (2012).
25. Helwak, A., Kudla, G., Dudnackova, T. & Tollervey, D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell 153, 654–665 (2013).
26. O’Donnell, K. A., Wentzel, E. A., Zeller, K. L., Dang, C. V. & Mendell, J. T. c-Myc-regulated microRNAs modulate E2F1 expression. Nature 435, 839–843 (2005).
27. Dees, N. D. et al. MuSiC: identifying mutational significance in cancer genomes. Genome Res. 22, 1589–1598 (2012).
28. Kandoth, C. et al. Mutational landscape and significance across 12 major cancer types. Cancer Discov. 3, 282–304 (2013).
29. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 39, e118 (2011).
30. Mermel, C. H. et al. GIST2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 12, R41 (2011).
31. Zack, T. I. et al. Pan-cancer patterns of somatic copy number alteration. Nat. Genet. 45, 1134–1140 (2013).
32. Hong, L. et al. The miR-17-92 cluster of microRNAs confers tumorigenicity by inhibiting oncogene-induced senescence. Cancer Res. 70, 8547–8557 (2010).
33. Song, M. S., Salmena, L. & Pandolfi, P. P. The functions and regulation of the TGFbeta in cancer. Nat. Rev. Mol. Cell Biol. 13, 233–245 (2012).
34. Massague, J. TGFBeta in cancer. Cell 134, 215–230 (2008).
35. Weber, A. et al. Zeb4 represses transcription of P21CIP1 and controls the cellular response to p53 activation. EMBO J. 27, 1563–1574 (2008).
36. Li, L., Shi, J. Y., Zhu, G. Q. & Shi, B. MiR-17-92 cluster regulates cell proliferation and collagen synthesis by targeting TGFβ pathway in mouse palatal mesenchymal cells. J. Cell. Biochem. 113, 1235–1244 (2012).
37. Mestdagh, P. et al. The miR-17-92 microRNA cluster regulates multiple components of the TGF-beta pathway in neuroblastoma. Mol. Cell 40, 762–773 (2010).
38. Kim, K. et al. Identification of oncogenic microRNA-17-92/ZBTB4/specificity protein axis in breast cancer. Oncogene 31, 1034–1044 (2012).
39. Banerji, S. et al. Sequence analysis of mutations and translocations across breast cancer subtypes. Nature 486, 405–409 (2012).
40. Doench, J. G. & Sharp, P. A. Specificity of microRNA target selection in translational repression. Genes Dev. 18, 504–511 (2004).
41. Solomon, E., Li, H., Duhachek Muggy, S., Syta, E. & Zolkiewska, A. The role of SnoN in transforming growth factor beta-induced expression of metalloprotease-disintegrin ADAM12. J. Biol. Chem. 285, 21969–21977 (2010).
42. Tecalo-Cruz, A. C. et al. Transforming growth factor-beta/Smad TARGET gene SKIL is negatively regulated by the transcriptional cofactor complex SNON-SMAD4. J. Biol. Chem. 287, 26764–26776 (2012).
43. Bonnie, S. et al. TGFB-β induces assembly of a Smad2–Smurf2 ubiquitin ligase complex that targets SnoN for degradation. Nat. Cell Biol. 3, 587–595 (2001).
44. Krek, A. et al. Combinatorial microRNA target predictions. Nat. Genet. 37, 495–500 (2005).
45. Betel, D., Wilson, M., Gabow, A., Marks, D. S. & Sander, C. The microRNA.org resource: targets and expression. Nucleic Acids Res. 36, D149–D153 (2008).
46. Ebert, M. S., Neillon, J. R. & Sharp, P. A. MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. Nat. Methods 4, 721–726 (2007).
47. Corcoran, D. L. et al. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. Genome Biol. 12, R79 (2011).
48. Loeh, G. B. et al. Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. Mol. Cell 48, 760–770 (2012).
49. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
50. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11, 94 (2010).
51. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140 (2010).
52. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11, R25 (2010).
53. Burkhardt, D. L. et al. Regulation of RB transcription in vivo by RB family members. Mol. Cell Biol. 30, 1729–1745 (2010).

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 IK01DK096093 (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 Mosses, David Bader and Joel Neilson for editing contributions.

Author contributions

M.P.H. contributed to study design and interpretation, construct design, wet lab experimentation, AGO-CLIP data set compilation and paper text. K.R. and C.C. helped in generation of miSNP algorithm and paper text. S.M.H. contributed to study design and paper editing. P.H.G., R.A.G. and D.A.W. contributed to study design. C.K., M.D.M. and L.D. contributed to generation of MuSiC pan-cancer analysis. B.R. contributed to generation of MSKCC pan-cancer functional driver predictions. T.Z. contributed to generation of pan-cancer ABSOLUTE-GISTIC analysis. S.E.M is the senior contributing author and helped in study design and interpretation, and paper editing.

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/

How to cite this article: Hamilton, M. P. et al. Identification of a pan-cancer oncogenic microRNA superfamily anchored by a central core seed motif. Nat. Commun. 4, 2730 doi: 10.1038/ncomms3730 (2013).

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/