Methods

Oligonucleotide synthesis and annealing. n-miR-29b and C.elegans miR-67 were obtained as annealed duplexes from Dharmaco (GE Healthcare) while bio-miR-29b (Integrated DNA Technologies, USA) and CNV\textsuperscript{K} containing oligonucleotides (Nihon Techno Service, Japan) were obtained as HPLC purified dried RNA pellets of sense and antisense strands (see Supplementary Table 6 for oligonucleotide sequences). All strands contained 5\(^\prime\) terminal phosphate and where present, biotin was attached to the 3\(^\prime\) OH group of the sense strand via a TEG (tetra-ethylene-glycol, 15 atom) spacer arm. Sense and antisense strands were annealed by mixing equimolar amounts of each oligonucleotide with oligo annealing buffer (20mM potassium acetate, 6mM HEPES pH 7.4, 400uM magnesium acetate). This resultant solution was incubated in boiling water that was left to cool down slowly at room temperature. Annealed duplexes were diluted before storing as aliquots. Successful annealing of oligos was confirmed by running annealed duplexes alongside single stranded sense and antisense oligos on a 2\% agarose gel.

MTT assay. Exponentially growing NIH3T3 cells (7000 cells/well) were reverse transfected in 96-well plates with varying doses of miR-29b \textsuperscript{CNV\textsuperscript{K}} oligos using either Lipofectamine 2000 (Life Technologies) or Lipofectamine RNAiMAX (Life Technologies). After 24 or 48 hours post-transfection, cells were treated with 1mg/ml MTT (Sigma) for 4 hours at 37\(^\circ\)C. Following removal of MTT reagent cells were treated with 100\% isopropanol for 40 minutes and the spectroscopic absorbance of coloured extract was measured at 570 nm. Empty wells with medium only served as blank. Cells treated with growth medium only were called ‘Untreated’ and represented healthy growing cells. ‘Reagent only’ cells were treated with either of the two lipid-based transfection reagents and were used to assess the reduction in cell viability caused by transfection reagents alone in the absence of any \textsuperscript{CNV\textsuperscript{K}} oligonucleotide. The fold cell viability was calculated relative to the untreated control cells.

Cell growth and transfection. HeLa and NIH3T3 cells were maintained in DMEM media supplemented with 10\% (v/v) FBS and 2mM L-glutamine. All cells were obtained from American Type Culture Collection and cultured at 37\(^\circ\)C and 5\% (v/v) CO\textsubscript{2}. All cells were tested negative for mycoplasm contamination (MycoAlert Mycoplasma Detection Kit, Lonza). One million cells were reverse transfected with equimolar amounts of \textsuperscript{CNV\textsuperscript{K}}-miR-29b and \textsuperscript{CNV\textsuperscript{K}}-scram oligos (Nihon Techno Service, Japan) per well in a 6-well plate using Lipofectamine RNAiMax reagent (Life Technologies) according to the manufacturer's instructions. Unless mentioned otherwise, mimics were transfected at a final concentration of 10nM. Cells were topped up with 2 ml of complete growth media 24 hours after transfection, and were allowed to recover for an additional 8 hours before harvesting. To prepare deep sequencing libraries from pulldown RNA approximately 9\times10\textsuperscript{6} cells were transfected with \textsuperscript{CNV\textsuperscript{K}} mimics. For qRT-PCR analysis of target genes 6\times10\textsuperscript{6} cells were transfected, while 4\times10\textsuperscript{5} cells were transfected for analysis of \textsuperscript{CNV\textsuperscript{K}} oligo associated proteins by immunoblotting.

RNA and protein pulldown using \textsuperscript{CNV\textsuperscript{K}} oligos. NIH3T3, HeLa cells were transfected with 10nM annealed oligonucleotides in 6-well format using Lipofectamine RNAiMax reagent (Life Technologies). 24 hours post-transfection, cells were washed with 2 ml complete growth media and allowed to recover for 8 hours. For ‘lysed’ crosslinking protocol, cells were rinsed in ice-cold PBS twice, harvested by scraping and lysed in hypotonic lysis buffer (10 mM KCl, 1.5 mM Mg\textsubscript{Cl\textsubscript{2}}, 10 mM Tris-Cl pH 7.5, 5 mM DTT, 0.5% NP-40, 60 U/ml SUPERase\textsuperscript{In} (Ambion) and 1 \times Complete Mini protease inhibitor cocktail (Roche)\textsuperscript{31}. Cell debris was cleared by centrifugation (12,000 g at 4\(^\circ\)C for 2 minutes). The cleared lysate was irradiated as droplets on ice-cold aluminium block for 10 minutes under a 20W, 365 nm UV lamp at a distance of ~10cm from the light source, and 1/20\(^{th}\) of its volume was kept as Input. The remaining irradiated supernatant was transferred to a clean tube to which NaCl was added to a final concentration of 1M. Dynabeads MyOne Streptavidin C1 (12 \mu l beads/million cells; Life Technologies) were pre-blocked with 1 \mu g/\mu l bovine serum albumin (Sigma) and 1 \mu g/\mu l yeast tRNA (Sigma), and incubated with the supernatant for 30 minutes at room temperature. Beads were then washed with hypotonic lysis buffer supplemented with 1M NaCl before extracting associated RNAs from pulldown and Input samples using Trizol LS (Life Technologies) (supplemented with 17ug/ml Glycobule (Ambion) and 10mM MgCl\textsubscript{2}) as per manufacturer’s instructions.
For ‘intact’ crosslinking protocol, cells were rinsed with ice-cold PBS once, 500ul PBS was added per well and living cells were irradiated on ice as above. Crosslinked cells were lysed in hypotonic lysis buffer and cell debris was cleared by centrifugation, as described above. 1/20th volume of the supernatant was kept as Input to which Trizol LS (Life Technologies) (supplemented with 17ug/ml Glycoblue (Ambion) and 10mM MgCl₂) was added and RNA extraction was carried out as per manufacturer’s instructions. The remaining supernatant was treated for biotin pulldown using Dynabeads MyOne Streptavidin C1 followed by RNA extraction as described above. RNA extraction, reverse transcription and qRT-PCR were performed as described below and one-way ANOVA followed by Tukey’s post-hoc test for multiple testing correction was used to test statistical significance amongst the three groups of cells transfected with \textsuperscript{CNV}K-miR-29b, n-miR-29b and \textsuperscript{CNV}K-scram oligos respectively. For statistical analysis, data were logarithmically transformed to ensure lognormal distribution and Brown–Forsythe test was used to confirm homogeneity of variance.

For isolation of proteins associated with \textsuperscript{CNV}K oligos, cell transfection, crosslinking and pulldown were performed as described above. For eluting proteins associated with \textsuperscript{CNV}K oligos from pulldown and input samples, the magnetic beads were treated twice with 0.1% SDS and incubated at 95°C/5min before collecting the eluate. The collected eluate was mixed with 4× Laemmli sample buffer (Bio-rad) before immunoblotting.

**Nuclear cytoplasmic fractionation.** Cytoplasmic and nuclear RNA was isolated from NIH3T3 cells as described previously\textsuperscript{4}. Briefly, the cells were trypsinized and harvested by centrifugation. The cell pellet was washed twice with ice cold PBS before incubating in buffer A (10 mM Tris-Cl pH 8, 140 mM NaCl, 1.5 mM MgCl₂, 0.5% NP-40) on ice for 5 min. After centrifugation at 1000g for 3 min, the cytoplasmic fraction was contained in the supernatant while the nuclear fraction was in the pellet. The cytoplasmic fraction was collected and spun again to remove any contaminating nuclear pellet. The nuclear pellet was washed twice in buffer A, and finally with buffer A containing 1% Tween-40 and 0.5% deoxycholic acid.

RNA from cytoplasmic supernatant and nuclear pellet was isolated using Trizol LS, as per manufacturer’s instructions. RNA quality and yield was assessed on an Agilent 2100 Bioanalyzer using the RNA 6000 Nano Chip kit (Agilent). Cell equivalent amounts of RNA from nuclear and cytoplasmic fractions were used for all analyses, unless stated otherwise. RIPA lysis buffer (50 mM Tris-Cl pH 7.4, 150 mM NaCl, 1 mM EDTA, 1% Triton-X 100, 1% deoxycholic acid, 0.1% SDS) containing 1× complete mini protease inhibitor cocktail (Roche) and 1× phosphatase inhibitor (1 mM sodium vanadate, 1 mM sodium pyrophosphate, 1 mM sodium molybdate, 10 mM sodium fluoride) was used to extract proteins from the nuclear pellet. Protein concentrations in cytoplasmic supernatant and nuclear lysate were determined using BCA protein assay (Pierce), according to the manufacturer’s protocol. 10 μg of protein from nuclear and cytoplasmic lysate was used for immunoblotting.

**RNA extraction, reverse transcription and qRT-PCR.** Total RNA was extracted from transfected cells using mirVana miRNA extraction kit (Life Technologies). Concentrations of RNA samples were measured at 260 nm using a ND1000 Nanodrop (Biolab). RNA samples were DNase treated using TURBO DNase (Ambion) and purified using RNeasy MinElute cleanup kit (Qiagen), to retain small RNAs according to manufacturer’s instructions. cDNA was synthesized from 500ng of extracted RNA using random hexamers at a concentration of 250ng per 5μg of RNA with SuperScript III Reverse Transcriptase kit (Invitrogen) as per manufacturer’s instructions. The synthesized cDNA was diluted 4 times in RNase-free water and subsequently used for qRT-PCR to quantify mouse miR-29b target genes Col3a1, Col1a1, Col5a3, Adam12 and Dnmt3a (see Supplementary Table 6 for primer sequences). Supplementary Figure 12 demonstrates primer binding sites for NIH3T3 qRT-PCR positive and negative control genes. Transcript levels for specific genes were quantified relative to expression of the housekeeping gene Gapdh and data were expressed relative to Gapdh using the ΔCt method. qRT-PCR was carried out using the SYBR Green PCR Master Mix (Applied Biosystems) in 20ul reactions, using 2ul of diluted cDNA, 4ul of primers (2uM), 4ul water and 10ul 2X master mix per reaction. For mature miRNA quantification, 5ng of RNA was used for cDNA synthesis. Reverse transcription and qRT-PCR was carried out using Taqman small RNA assays kit (Life Technologies) as per the instructions of the manufacturer. MicroRNA expression was calculated relative to
U6 snRNA expression. All reactions were carried out in triplicates in 96 well plates on the ViiA7 system (Applied Biosystems).

The enrichment of tRNA-Lysine in the cytosol and snoRA19 in the nucleus was measured by qPCR to validate the purity of nuclear and cytoplasmic fractions. 1µg of Total RNA and cell equivalent amounts of nuclear and cytoplasmic RNA was treated with DNase I (Invitrogen) and 500ng of it was primed using random hexamers at a concentration of 250ng per 5µg of RNA. cDNA was prepared using SuperScript III Reverse Transcriptase kit (Invitrogen) as per manufacturer's instructions. Due to the lack of an invariant internal control, we performed absolute quantification for the above RNAs. Briefly, the qPCR targets were PCR amplified for 35 cycles and purified from a 2% agarose gel using the Wizard SV Gel and PCR Clean-Up System kit (Promega). The gel purified amplicons were cloned into pGEMT Easy vector (Promega) and after sequence verification at the Australian Genome Research Facility (AGR) PCR amplicons were diluted to 1ng/ul and used as templates for a 1/10 standard curve serial dilution series. qRT-PCR was carried out in triplicate reactions in 96 well plates on the ViiA7 system (Applied Biosystems), each plate included RT negative and No template controls. Data was analysed using Applied Biosystems Sequence Detection Software v1.2.3.

**Nuclear-cytoplasmic miRNA target pulldown using CNVK oligos.** For isolation of miR-29b associated transcripts in nucleus and cytoplasm, 9×10⁶ NIH3T3 cells were transfected with 10nM annealed oligonucleotides in 6-well format using Lipofectamine RNAiMax reagent (Life Technologies). 24 hours post-transfection, cells were tapped up with 2 ml complete growth media and allowed to recover for 8 hours. Cells were rinsed with ice-cold PBS once, 500ul PBS was added per well and live cells were irradiated with 365nm UV light for 10 minutes under a 20W, 365 nm UV lamp at a distance of ~10cm from the light source. Nuclear-cytoplasmic fractionation was carried out as described above, except using a slightly modified buffer A recipe (10 mM Tris-Cl pH 8, 140 mM NaCl, 1.5 mM MgCl₂, 0.5% NP-40 supplemented with 5 mM DTT and 1× Complete Mini protease inhibitor cocktail (Roche). To the nuclear pellet 130 ul of sonication buffer (0.1% SDS, 1 mM EDTA, 10 mM Tris-Cl pH 8, 100 mM NaCl) was added and nuclei were lysed using a focused-ultrasonicator (Covaris S2, SonolAB Single) using 5 pulses of 60 seconds each with the following operating conditions: duty cycle: 2%; intensity: 3; cycles/burst: 200; bath temperature: 4°C; mode: frequency sweeping; degassing mode: continuous; volume: 130 ul in Covaris microTUBE, AFA fiber with Snap-Cap. The volume of sonicated nuclear lysate and cytoplasmic supernatant was made up to 1ml with modified buffer (0.1% SDS, 1 mM EDTA, 10 mM Tris-Cl pH 8, 150 mM NaCl, 0.1% Tween-20). Membranes were then blocked with 1 µg/µl bovine serum albumin (Sigma) and 1 µg/µl yeast tRNA (Sigma), and incubated with salt-adjusted, nuclear-cytoplasmic lysates for 30 minutes at room temperature. Beads were then washed with hypotonic lysis buffer supplemented with 1M NaCl before extracting associated RNAs from pulldown and Input samples using Trizol LS (Life Technologies) (supplemented with 17ug/ml Glycoblue (Ambion) and 10mM MgCl₂) as per manufacturer's instructions. Isolated RNA was resuspended in water and RNA quality and yield was assessed on Agilent 2100 Bioanalyzer using the RNA 6000 Pico Chip kit (Agilent).

**Immunoblotting.** For analysis of proteins associated with CNVK oligos and to analyze the purity of subcellular fractions by immunoblotting, the proteins were separated on a pre-cast 4-20% gradient SDS-PAGE gel (Bio-rad). The separated proteins were transferred onto methanol-activated polyvinylidene difluoride (PVDF) membranes and were blocked with 5% Skim milk (5% BSA was used for ab32381, Abcam) in 1X TBST buffer (50 mM Tris-Cl pH 7.6, 150 mM NaCl, 0.1% Tween-20). Membranes were then probed for specific antigens using antibodies against AGO2 (ab57113 for HeLa (Abcam); ab32381 for NIH3T3 (Abcam), GAPDH (2275-PC-100, Trevigen), Lamin A/C (4777S, Cell Signalling), H3 Histone (9715S, Cell Signalling) or Calnexin (ab140818, Abcam). Horseradish peroxidase (HRP) conjugated α-Rabbit or α-mouse antibody (Cell Signalling) was used as secondary antibody. Amersham ECL Western Blotting Detection Reagents (GE healthcare) was used for immunodetection. See Supplementary Figure 11 for full-length digital images of immunoblots.
**Luciferase assays.** To select 88 human 3'UTRs for Switchgear LightSwitch luciferase assay validation (Switchgear Genomics, CA), 827 CNV-identified human targets were binned into eight sets, from lowest to highest enrichment based on their PE value as follows: 4.15 to 5.41 (104 genes), 5.41 to 6.31 (103 genes), 6.31 to 7.38 (103 genes), 7.38 to 8.69 (104 genes), 8.69 to 10.7 (103 genes), 10.7 to 14.2 (103 genes), 14.2 to 19.5 (103 genes), 19.5 to 74.3 (104 genes). Each bin contained roughly the same number of genes and 11 genes were selected from each bin for luciferase assay validation. 35 of the 88 genes were conserved between human and mouse. A subset of previously known targets was chosen as a positive control for the luciferase assays, while an additional seven pLightSwitch reporters were used as negative controls, determined by a lack of predicted miR-29b binding sites. This negative control set included the empty vector, four constructs from non-miR-29b targeted genes (ACTB, GAPDH, LDHA, PPIA), and two controls containing a random, non-geric, non-conserved sequence in place of the 3'UTR. An optimized synthetic target consisting of sequence repeats that are fully complementary to miR-29b was used as a positive control. 96-well plates were seeded with 10,000 HeLa cells per well 18-24 hours before transfection to achieve 80% confluence at the time of transfection. Each transfection included 0.3µl of DharmaFECT DUO transfection reagent, 100ng of 3'UTR reporter and sufficient mimic or non-targeting control miRNA to yield a final concentration of 50nM in a total volume of 100µl/well. Each construct was transfected in triplicate separately with either the mimic (hsa-miR-29b) or the non-targeting control. Cells were incubated at 37°C for 24 hours post-transfection before being harvested. 100µl of LightSwitch Luciferase Assay Reagent (Active Motif) were added to each well, plates were incubated at room temperature for 30 minutes and finally read on a SpectraMax L luminometer. 3'UTR negative controls were used to normalize luminescence values, as per standard Switchgear normalization procedures.

Switchgear-provided UTR sequences were mapped to the genome with Lastz and intersected with TargetScan predictions, revealing that 34/88 had no overlap with TargetScan predicted binding sites. Switchgear constructs include 3'UTRs based on RefSeq annotations, while we utilised GENCODE annotations for RNA seq mapping and enrichment analysis. Site type for a particular gene was defined as the highest confidence type of TargetScan predicted site in the UTR of that gene, i.e. if a UTR had a 7-mer-A1 and 8-mer site it was annotated as “8-mer”. Target site efficacy was ordered as follows, i.e. 8mer > 7mer-m8 > 7mer-A1 > 6mer. Pearson's Chi-squared tests were carried out using the chisq.test function of the R stats library. Visualisation (Supplementary Fig. 5) was performed using the ggplot2 library.

For luciferase assay data analysis, a normalisation factor (NF) was calculated for the miR-29b mimic (mim) and control (NTC) experiments, according to the following formulas:

\[
NF_{\text{mim}} = \left( \frac{ACTB_{\text{mim}}}{ACTB_{\text{NTC}} + ACTB_{\text{mim}}} \right) \cdot \left( \frac{GAPDH_{\text{mim}}}{GAPDH_{\text{NTC}} + GAPDH_{\text{mim}}} \right) \cdot \left( \frac{PPIA_{\text{mim}}}{PPIA_{\text{NTC}} + PPIA_{\text{mim}}} \right) \cdot \left( \frac{LDHA_{\text{mim}}}{LDHA_{\text{NTC}} + LDHA_{\text{mim}}} \right) \cdot \left( \frac{Empty_{\text{mim}}}{Empty_{\text{NTC}} + Empty_{\text{mim}}} \right)
\]

\[
NF_{\text{NTC}} = \left( \frac{ACTB_{\text{NTC}}}{ACTB_{\text{NTC}} + ACTB_{\text{mim}}} \right) \cdot \left( \frac{GAPDH_{\text{NTC}}}{GAPDH_{\text{NTC}} + GAPDH_{\text{mim}}} \right) \cdot \left( \frac{PPIA_{\text{NTC}}}{PPIA_{\text{NTC}} + PPIA_{\text{mim}}} \right) \cdot \left( \frac{LDHA_{\text{NTC}}}{LDHA_{\text{NTC}} + LDHA_{\text{mim}}} \right) \cdot \left( \frac{Empty_{\text{NTC}}}{Empty_{\text{NTC}} + Empty_{\text{mim}}} \right)
\]
Where $\text{gene}_{\text{aus}}$ and $\text{gene}_{\text{src}}$ are the average luminosity recorded in the three replicate miR-29b mimic and non-targeting control experiments, respectively. The luminosity of each interrogated 3' UTR reporter was then divided by the corresponding normalisation factor, and a Welch's t-test with a Benjamini-Hochberg false discovery rate of 0.05 was used to identify significant differences between miR-29b mimic and non-targeting control normalised luminosity.

**Northern blot hybridization.** To assess if cells equally take up both $\text{CNV}\text{-miR-29b}$ and $\text{CNV}\text{-scram}$ oligonucleotides, 6 million NIH3T3 cells were transfected with $\text{CNV}\text{-K}$ oligos. Crosslinking, pulldown and RNA isolation was performed as described above. 10μg of input RNA was used for each blot and entire yield of pulldown RNA was used. tRNA-Lys was used to serve as an internal control. To assess if transfected $\text{CNV}\text{-K}$ oligos are imported into the nucleus, 9 million NIH3T3 cells were transfected with 10nM $\text{CNV}\text{-K}$ oligos and nuclear, cytoplasmic pulldown RNA isolated as described above without 365 nm UV crosslinking. Entire yield of pulldown RNA was used for Northern blot hybridization. 4μg of total RNA isolated from respective transfected cells was used as positive control. To demonstrate nuclear localization of miR-29b, cell equivalent amounts of NIH3T3 nuclear and cytoplasmic RNA was used for every blot. 10μg of Total RNA was loaded per blot to serve as an internal control.

An equal volume of 2X formamide loading dye was mixed with RNA and samples denatured at 95°C for 5 min. 15% polyacrylamide/8M urea/0.25X TBE gel was used to detect miRNAs. RNA was transferred onto a nylon membrane (Hybond-NX, Amersham Biosciences) in a semi dry fashion using 0.25X TBE. The RNA was chemically cross-linked for 2h at 55°C using 1-Ethyl-3-[3-dimethylaminopropyl]carbodiimide hydrochloride (EDC). The blots were hybridized with radiolabelled probes and the membrane exposed to storage phosphor screen for at least 3 hours. Signal was detected using Typhoon 8600 variable mode imager (Amersham Biosciences).

**Splinted ligation.** Was performed as previously described. Briefly, 2μg RNA was mixed with 1pmol of radiolabelled ligation oligo in the presence of 1pmol of bridge oligonucleotide in a 10μl reaction. The reaction was incubated at 94°C for 1min, 65°C for 2min and 37°C for 10 min. The RNA and ligation oligo were ligated using Ligate–IT Rapid ligation kit (USB) and the mixture incubated at 30°C for 1hour. Finally, 1μl of Shrimp Alkaline Phosphatase (USB) was added per reaction and incubated at 37°C for 15mins. All incubations were carried out in a thermacycler PCR machine. Samples were denatured in an equal volume of 2X formamide loading dye at 95°C for 3min and loaded onto 15% PAGE 8M urea/1X GTG buffer. The bridge oligos were synthesized with three-carbon spacers (SpC3) on either end, while the ligation oligo had one spacer at its 3' end (Supplementary Table 6). The carbon spacers serve as unligatable groups thus preventing the formation of oligo concatamers. All oligos were synthesized by IDT.

**Library preparation and sequencing.** RNA samples were DNase treated using TURBO DNase (Ambion) and purified using RNaseasy MinElute cleanup kit (Qiagen) according to manufacturer's instructions. The DNA was quantified using Agilent RNA 6000 pico kit (Agilent Technologies) and a mix of oligodT and random hexamers primers were used to prepare sequencing libraries from 5ng RNA using TotalScript RNA-seq kit (Epicentre), as per manufacturer’s instructions with a few modifications. Modifications in the library preparation protocol include the use AxyPrep magnetic beads (AxyPrep Mag PCR Clean-up Kit 50 mL, Axygen) instead of AMPure XP magnetic beads throughout the protocol; 2X Phusion High Fidelity PCR Master mix was replaced with a cocktail of 5x Phusion High Fidelity PCR Master Mix HF Buffer (NEB), 200uM dNTPs and 0.5 units Phusion High-Fidelity DNA Polymerase (NEB), finally a 1:1 ratio of magnetic beads to gap-filled ampiclon was used in the final clean-up step of the protocol. Libraries were sequenced on an Illumina HiSeq 2500 platform (QCMG, Australia) to obtain 100bp paired-end reads.

For small RNA deep sequencing analysis nuclear and cytoplasmic RNA was isolated from NIH3T3 cells using the fractionation protocol described above. These samples were used to construct two small RNA libraries using Illumina TruSeq Small RNA kit (Illumina) as per the manufacturer's instructions. The libraries were sequenced over an Illumina HiSeq 2000 platform (QCMG, Australia) to obtain 50bp single-end reads. The small RNA reads were trimmed, adaptors removed, and identical reads collapsed to a single read using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) commands "fastx_clipper -a
TGGAATTCTC -l 16 -i <input.fastq> -o <clipped.fastq> -Q33" and “fastx_collapser -i <clipped.fastq> -o <clipped.collapsed.fasta> -Q33”. The final reads were mapped against the UCSC mouse reference genome (mm10) using bowtie 2 with settings: "--local -k 100 -f -L 10 -N 1 -i L,1,0". The resulting SAM file was converted to BAM format using samtools, and sorted by both coordinate and read name using picard. Reads in name-sorted BAM overlapping annotated miRNA genes were counted using a custom Python script. All scripts are available from the authors by request.

**Read mapping and quantification.** STAR version 2.3.0 was used to map reads to the genome (mm10 for mouse and hg19 for human). GENCODE annotations (version M2 for mouse and 19 for human) were used to construct the splice junction database and as a reference for the count tables. Gene-level count tables for were obtained using the HTSeq framework in union mode. Count tables were filtered to discard lowly expressed genes, defined as those with fewer than a total of 30 tags mapping across all of the input datasets. A pseudocount of one was added to each observation of each gene, and counts were then normalised to the number of uniquely mapped reads in each of the sequencing libraries. A pulldown enrichment (PE) value was determined for every set of four experiments:

\[
P_{E} = \frac{29b_{\text{pulldown}} - \text{average}(29b_{\text{input}}, \text{scrambled}_{\text{input}})}{\text{scrambled}_{\text{pulldown}}}
\]

Genes were considered enriched if the PE of all replicates was greater than 4.

**Quality control and exploratory data analysis.** DESeq2 was used for exploratory data analysis. Hierarchical clustering, correlation and overlap analysis of mouse and human datasets (Supplementary Fig. 4) revealed good concordance between similar samples in the repetitions of the experimental series, and between genes considered to be miR-29b targets as a result of having a PE>4. Few differences were observed between samples prepared using the intact and lysed protocols.

**Gene ontology analysis.** Gene ontology, pathway, transcription factor binding site and miRNA target site enrichment analysis were carried out using the ToppGene server with a 0.05 Benjamini-Hochberg false discovery rate and requiring a minimum of 2 features counted in the test set. For mouse genes, the MGI name was used as input; this resulted in 522 unique gene names queried. For human genes, 763 genes had HGNC gene names and were present in the ToppGene database. In order to incorporate ontology annotations into the list of genes tested using SwitchGear luciferase assays in Supplementary Figure 5, the gene list was submitted to ToppGene, and genes that were part of the “Cell cycle”, “DNA Damage Response”, “Chromatin”, “Extracellular matrix and collagens” (which represents the union of “Extracellular matrix” and “Collagens” terms), and “Occluding junctions” ontologies were annotated as such (Supplementary Table 3).

**Comparison of CNVK identified genes with previously reported miR-29b targets.** The list of CNVK-identified genes was compared to a list of previously reported miR-29b targets (Supplementary Table 1), taking into consideration levels of expression of these genes in the cell lines used. This enabled us to take into account the fact that the CNVK approach cannot identify a gene to be a target of miR-29b if that gene is not expressed. Of the 69 known mouse targets (Supplementary Table 4), 55 were tested for differential enrichment since they passed the greater than 30 counts across all NIH3T3 input datasets threshold. Similarly, 83/111 known human targets (Supplementary Table 4) were interrogated.

**Comparison with predicted miR-29b targets.** TargetScan predictions were downloaded from the UCSC Genome Browser Table Browser (“TargetScan miRNA Regulatory Sites” track; Supplementary Table 7). Predictions were subdivided into four categories: those containing a conserved 8-mer site (“8mer”), those with a context score greater than 80 (“GT8ctx”), those that have both a conserved 8-mer site and a context score greater that 80 (“high-confidence”) and all sites (“total”).

**Mouse targets.** Predictions were intersected with ensembl transcripts using the UCSC Genome Browser Table Browser functionality. BiomaRt was used to retrieve ENSMG gene identifiers for the ENSMT
transcripts, and as most of these were represented in GENCODEM2 the corresponding ENSMUSG was used. Identifiers for those not present were retrieved using biomaRt of the ensembl archive version May 2012, and all but three ENSMUSG gene names were present in the current assembly. Those three genes were manually queried and current identifiers recovered: ENSMUSG00000079828 (not in primary assembly, BLAST of the peptide sequence against the “nr” database showed 100% identity with ENSMUSG00000093930), ENSMUSG00000091454 (merged into ENSMUSG00000095677 record as of Ensembl 68), and ENSMUSG00000091645 (only present in release 67, no peptide product; visualising this gene in UCSC shows overlap with Zfp3611, which now has the identifier). A total of 743 TargetScan miR-29b sites were annotated in UCSC, and these overlapped with 1486 ensembl transcripts, corresponding to 664 ensembl mouse genes (ENSMG) [13 sites had no overlap with an ensembl transcript]; of these 664, 509 were expressed above the 30+ tags threshold across all four input datasets and hence considered for potential enrichment. 226 8-mer miR-29b sites were identified, which overlapped with 449 ensembl transcripts, corresponding to 203 ensembl mouse genes (ENSMG) [6 sites had no overlap with an ensembl transcript]; of these 203, 148 were expressed above the 30+ tags threshold across all four input datasets and hence considered for potential enrichment. 263 TargetScan miR-29b sites had a context score >= 80, which overlapped with 516 ensembl transcripts, corresponding to 234 ensembl mouse genes (ENSMG) [4 sites had no overlap with an ensembl transcript]; of these 234, 167 were expressed above the 30+ tags threshold across all four input datasets and hence considered for potential enrichment. 170 TargetScan miR-29b sites had a context score >= 80 and an 8-mer sequence, which corresponded to 153 ensembl mouse genes (ENSMG); of these, 106 were expressed above the 30+ tags threshold across all four input datasets and hence considered for potential enrichment.

**Human targets.** Predictions were intersected with GENCODE 19 Comprehensive transcripts using the UCSC Genome Browser Table Browser functionality. BiomaRt was used to retrieve ENSG gene identifiers for the ENST transcripts. A total of 948 miR-29b sites were identified, which overlapped with 2235 transcripts, corresponding to 888 genes (ENSG) [4 sites had no overlap with an annotated transcript]; of these 888, 669 were expressed above the 30+ tags threshold across all three input datasets and hence considered for potential enrichment. 272 8-mer miR-29b sites were identified, which overlapped with 623 transcripts, corresponding to 268 human genes (ENSG) [1 site had no overlap with an annotated transcript]; of these 268, 202 were expressed above the 30+ tags threshold across all three input datasets and hence considered for potential enrichment. 311 TargetScan miR-29b sites had a context score >= 80, which overlapped with 657 transcripts, corresponding to 299 human genes (ENSG) [1 site had no overlap with an annotated transcript]; of these 299, 218 were expressed above the 30+ tags threshold across all three input datasets and hence considered for potential enrichment. 186 TargetScan miR-29b sites had a context score >= 80 and an 8-mer sequence, which corresponded to 203 human genes (ENSG); of these, 106 were expressed above the 30+ tags threshold across all four input datasets and hence considered for potential enrichment.

**Permutation testing.** A total of 567 (mouse) and 827 (human) miR-29b target hits were identified based on the PE metric filtering described above. We used the following random permutation model to assess whether the high degree of overlap with TargetScan predictions occurred by chance alone. We generated 10000 random sets of 567 and 827 genes, selecting from those expressed (>30 reads) in the input mouse and human datasets, respectively. We then counted how many of these genes were represented in the TargetScan predictions, with a context score > 80, predicted to have an 8-mer binding site or both of these. A p-value for each comparison was calculated by using the R statmod44 library using the approach described by Phipson45, which considers the number of times the number of overlaps between a random set and the TargetScan predictions exceeded that of our CNV-K-detected target set. A total of 1 (mouse) and 10 (human) genes were characterised as enriched using a PE metric for the scrambled oligonucleotide, none of which were predicted to be miR-29b targets by TargetScan.

**Target conservation analysis.** BiomaRt was used to identify human orthologs of mouse genes based on the cross-species mappings for GENCODE 19 and GENCODE M2. 542/567 genes identified as miR-29b targets in NIH3T3 cells had annotated human orthologs, of which 483 were expressed at levels of greater than 30 tags in HeLa cells across all of the input datasets and hence tested for pulldown enrichment. Of the
827 identified human hits, 656 were expressed above the minimum tag threshold in mouse (total of at least 30 counts across all eight input datasets) and hence tested for enrichment. The VennDiagram and ggplot2 R package was used to visualise the results.

METHOD REFERENCES

SUPPLEMENTARY NOTE 1 - Rationale for sequencing library synthesis approach

The \textsuperscript{CNV}K-miR-29b streptavidin pulldown from 10\textsuperscript{7} cells resulted in the purification of very small quantities of RNA (~5ng), less that the 100ng minimum required for the Illumina TruSeq stranded Total RNA kit. In addition to this we were interested in capturing both polyadenylated and non-polyadenylated target transcripts and to use random hexamer priming for library preparation. Low RNA input could result in significantly reduced library complexity.

To address this limitation, we compared sequencing results for a TruSeq library made from 1ug of K562 RNA\textsuperscript{T} with two libraries made from 5ng of RNA with the Epicenter TotalScript kit using either random hexamer (RH) or a 1:1 mix of random hexamer and oligo-dT (RHdT) primers. Unlike the TruSeq protocols, which use RiboZero to reduce the number of rRNA molecules that will be sequenced in the library, the TotalScript kit does not include a ribodepletion step.

Overall, both protocols strongly correlated (Spearman > 0.9) with the standard input dataset when comparing expression levels of GENCODE17 annotated genes (Supplementary Figure SN1). Mapping statistics, however, for the three sequenced libraries revealed that the low-input protocols were characterised by lower mapping rates than the standard input (8-13\% vs 80\%), and preferentially detected transcripts expressed at higher levels in K562 cells. The values for proportion of ribosomal transcripts in mapped data indicate that, as expected, using a 1:1 ratio of RH:oligo-dT reduces rRNA detection (18.99\% of reads in RH vs 6.61\% RHdT)).

The comparison of the K562 datasets demonstrated that the TotalScript 5ng protocol generated a representative library, especially for highly expressed transcripts, and that using RHdT priming enabled a more complex library than RH alone. For all \textsuperscript{CNV}K experiments described, libraries were made with a mix of random hexamer and oligo-dT priming from 5ng of RNA using the Epicenter TotalScript RNA kit.
Supplementary Figure SN1: Comparison of standard Illumina TruSeq stranded Total RNA kit with low-input TotalScript random hexamer and random hexamer:oligo-dT primed libraries. The left (RH) and centre (RHdT) panels show scatterplots of normalised to library depth read counts in the corresponding 5ng low-input (x -axis) and 1ug input (y -axis) datasets; both RH and RHdT counts correlate well with the standard input, with Spearman rank-based correlation coefficients of > 0.9. The right panel demonstrates that both low-input protocols preferentially detect genes that are more highly expressed in K562 cells.
SUPPLEMENTARY NOTE 2 – Sylamer analysis of CNVK identified target genes.

We used Sylamer\(^2\) to investigate k-mer overrepresentation in the 3' UTR of human and mouse genes identified as likely targets (PE > 4) of miR-29b using CNVK pulldown. BiomaRt was used to carry out gene:transcript mapping for target genes, and all annotated 3'UTRs in the pull down dataset were considered. Basic and Comprehensive Gencode M11 and Gencode 26 annotations were used for mouse and human, respectively. The seed sequence for miR-29 family is as follows:

\[
\begin{align*}
\text{hsa, mmu, mo-miR-29a} & : \text{UAAGCACC}_{\text{GGUGUGGU}} \\
\text{hsa, mmu, mo-miR-29b} & : \text{UAAGCACC}_{\text{UGUGUGGU}} \\
\text{hsa, mmu, mo-miR-29c} & : \text{UAAGCACC}_{\text{UGUGUGGU}}
\end{align*}
\]

Therefore, in theory we’d expect the complementary sequence \text{AATGGTGCT} (seed shown in bold) in miR-29b interacting mRNA molecules. Results for enriched k-mers of length 6 – 8 are presented below. These results show that through k-mer analysis alone we are able to recover the miR-29b complementary sequence \textit{de novo} in both the human and mouse datasets.
Overlapping the most significant of these with the theoretical sequence described above:

miR-29b comp seq:   AATGGTGCTA  (seed shown in bold)
Seq1:                TGGTGC
Seq2:                GGTGCT

Sorted sequences

log10(enrichment P-value)
Overlapping the most significant of these with the theoretical sequence described above:

miR-29b comp seq: AATGGTGCTA (seed shown in bold)
Seq1: TGGTGCT
Seq2: GGTGCTA
Seq3: TTGGTGC
Seq4: ATGGTGC
Seq5: GGTGCTT
Seq6: GGGAGGC #does not align
Overlapping the most significant of these with the theoretical sequence described above:

miR-29b comp seq: AATGGTGC(TA (seed shown in bold)
Seq1: TGGTGCTA
Seq2: TTGGTGCT
Seq3: ATGGTGCT
Seq4: TGGTGCTT
Seq5: AATGGTGC
Seq6: ATTTGGTGC
Similar results are observed for mouse:

**Mouse 6-mer**

**Sylamer landscape using words of length: 6**

Overlapping the most significant of these with the theoretical sequence described above:

miR-29b comp seq: **AATG**TGCTA (seed shown in bold)

Seq1: **T**GGTGC

Seq2: **G**GTGCT
Mouse 7-mer

Sylamer landscape using words of length: 7

miR-29b comp seq: AATGGTGCTA (seed shown in bold)
Seq1: TGGTGCT
Seq2: GGTGCTA
Seq3: ATGGTGC
Seq4: TTGGTGC
Seq5: GGTGCTT
Overlapping the most significant of these with the theoretical sequence described above:

miR-29b comp seq: AATGGTGCTA (seed shown in bold)

Seq1: TGGTGCTA
Seq2: ATGGTGCT
Seq3: TTGGTGCT
Seq4: TGTTGCTT
Seq5: TGTTGCTG
Seq6: GGTGCTAT
SUPPLEMENTARY NOTE REFERENCES
