Supplemental Figures

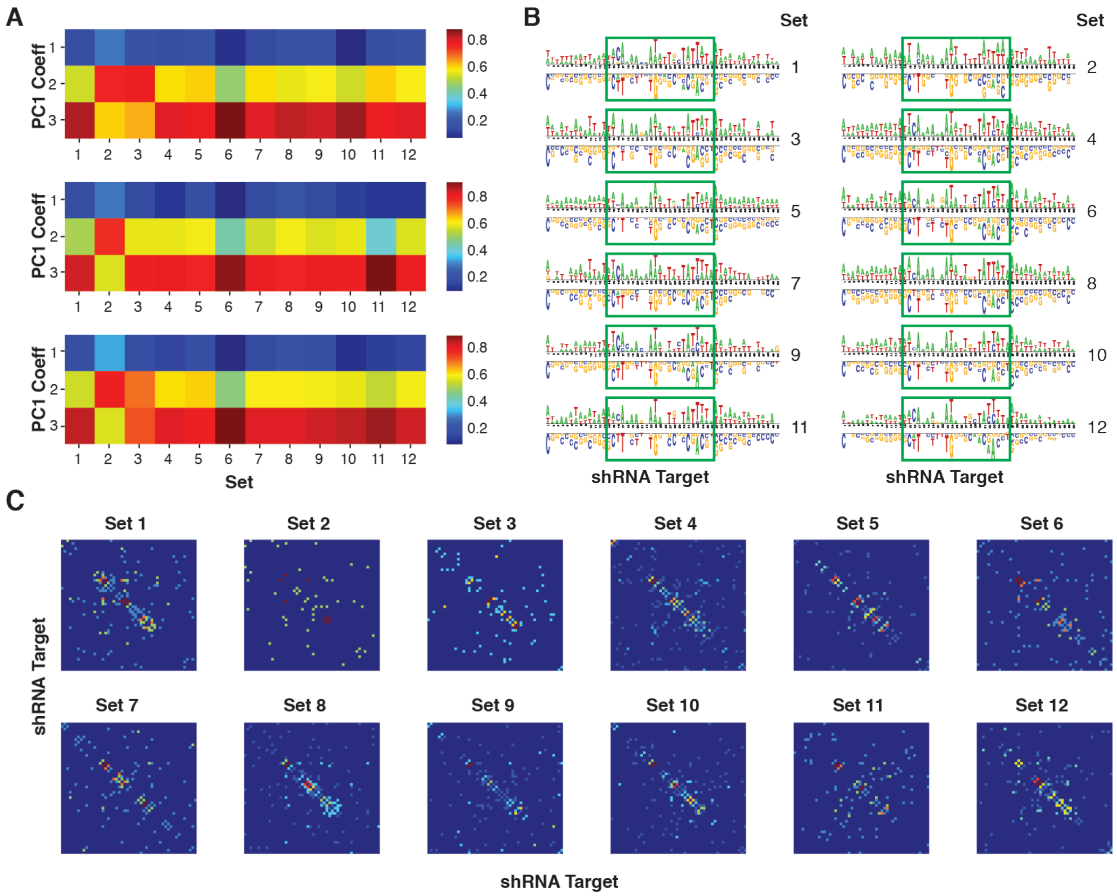

Figure S1, related to Figure 1: A) Heatmap representation of the coefficients used to extract the first principal components of the matrices described in Figure 1A. Coefficients 1, 2 and 3 represent the contribution that each of on-dox sorts 1, 2 and 3 made in defining the first principal component of the matrices. Biological replicates are shown in three different plots. Results from each of 12 separately processed shRNA sets are displayed. **B)** Significantly enriched (top) and depleted (bottom) nucleotides within potent shRNAs (with respect to weak shRNAs). Results from each of 12 separately processed shRNA sets are displayed. **C)** Heatmap representations of the predictive capacity (with respect to shRNA potency) of each pair of positions within the target region. Heatmap cells are colored colors to represent the number of nucleotide combinations that were significantly predictive, as calculated with via linear regression (p-value <0.05) at each position-pair. Results from each of 12 separately processed shRNA sets are displayed.

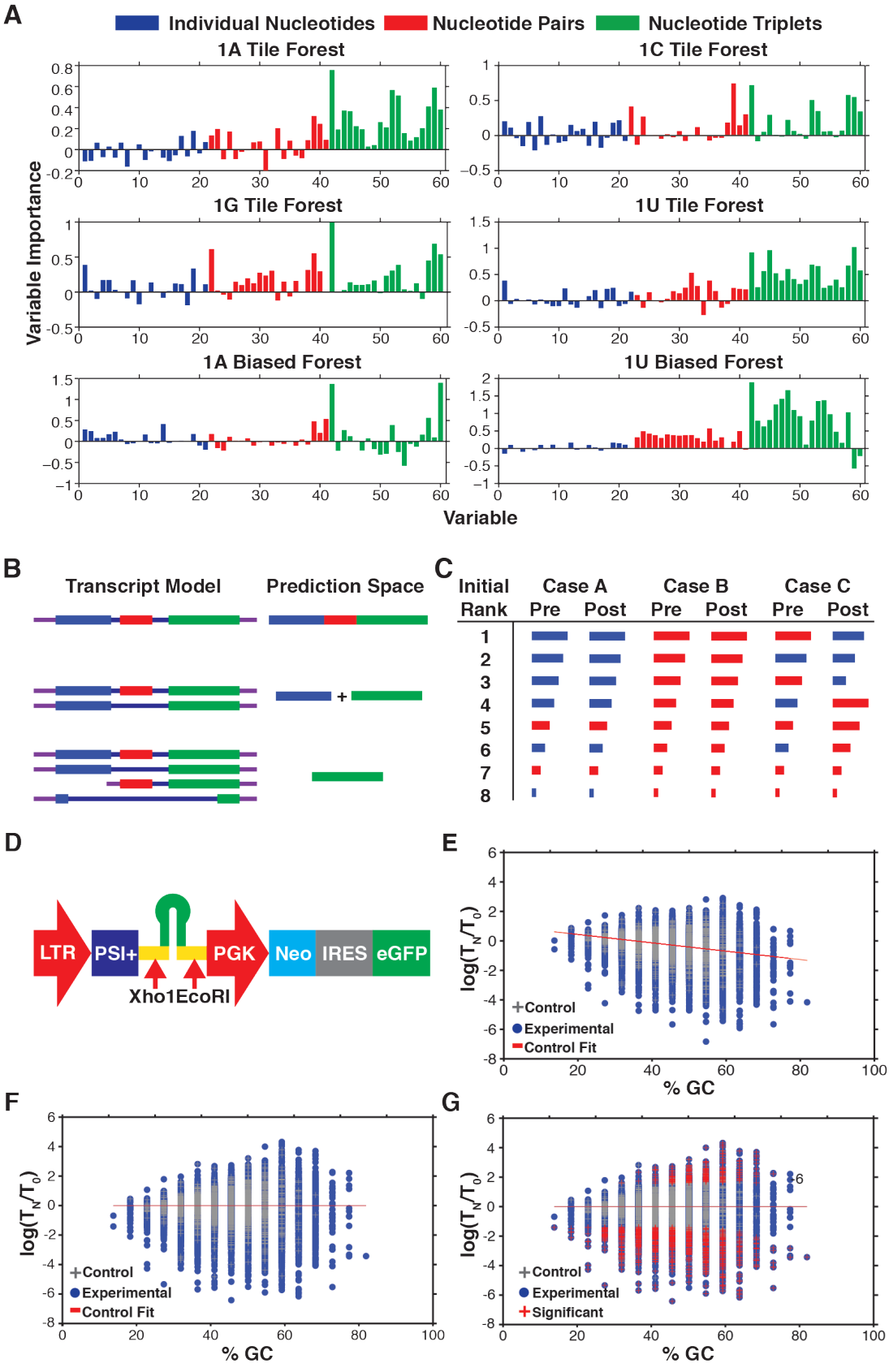
A
Individual Nucleotides
Nucleotide Pairs
Nucleotide Triplets
1A Tile Forest
1C Tile Forest
1G Tile Forest
1U Tile Forest
1A Biased Forest
1U Biased Forest
Variable Importance
Variable
B
Transcript Model
Prediction Space
C
Initial Rank
Case A
Case B
Case C
Pre
Post
D
LTR
PSI+
PGK
Neo
IRES
eGFP
Xho1EcoRI
E
log(T_N/T_0)
Control
Experimental
Control Fit
% GC
F
Control
Experimental
Control Fit
% GC
G
Control
Experimental
Significant
% GC

Figure S2, related to Figure 2: A) Variable importance in each first tier module of the shERWOOD algorithm. Each bar represents either the importance of individual nucleotide composition (blue), nucleotide pair composition (red) or nucleotide triplet composition (green) at a different position in the shRNA guide. The left most bars of each class represent the nucleotide, pair or triplet beginning at the second position of the guide. The right most bars represent those ending at the 22nd position of the guide. **B)** Example extractions of target regions for a single transcript gene (top) and two multiple transcript genes (middle and bottom). For the middle gene, a target region (composed of >250 bp present in >80% of transcripts) was identified on the first algorithm iteration. For the bottom gene, a second algorithm iteration was required, where the smallest transcript was not considered. **C)** Example shRNA off-target algorithm implementation. In case A, all rank 1-4 are non-multimappers, so no shuffling occurs. In case B, all rank 1-8 shRNAs are multimappers (indicting that the gene is a paralogue), so no shuffling occurs. In case C, some but not all rank 1-4 and rank 5-8 shRNAs are multimappers and shuffling occurs to select a set of 4 shRNAs that include the highest scoring non-multimappers. **D)** Schematic of the retroviral vector employed in the validation RNAi screens. **E)** Plot of shRNA log-fold changes with respect to shRNA-guide GC-content. The red line represents a one-dimensional polynomial fit to the control shRNA population data-points. **F)** Plot of shRNA log-fold changes with respect to shRNA-guide GC-content after the polynomial fit described above was subtracted from all data-points. **G)** shRNA hits calling by an Empirical-Bayes Moderated T-Test (FDR < 0.05).

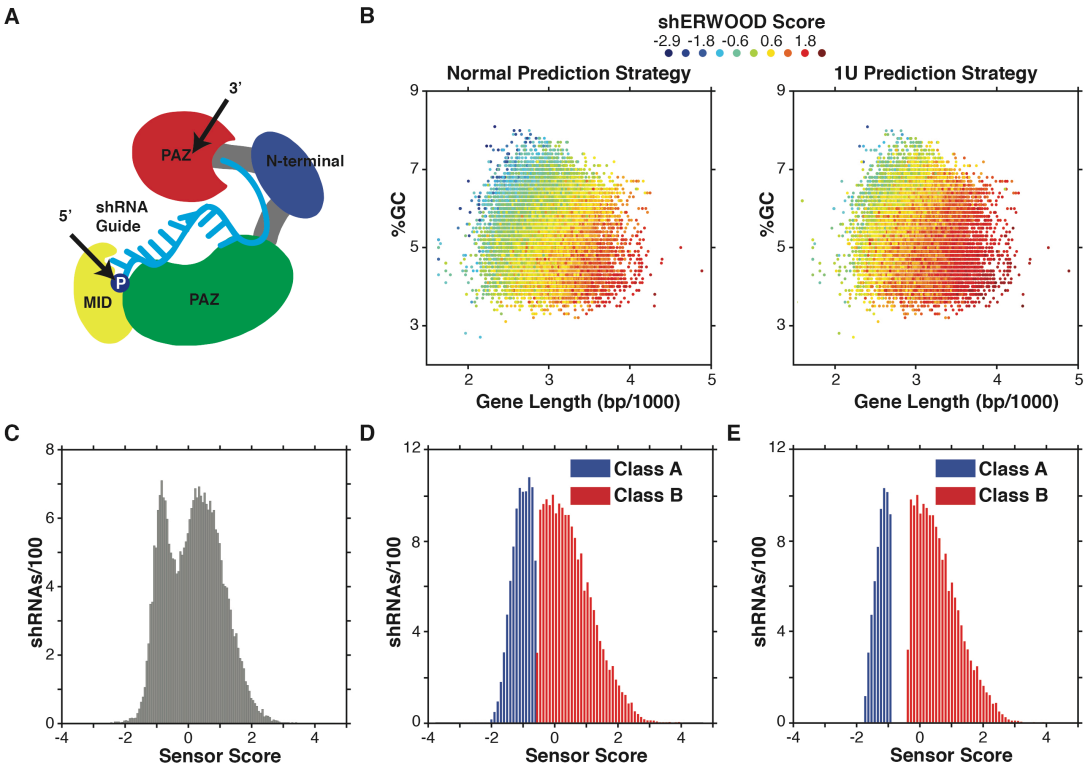

Figure S3, related to Figure 3: **A)** Schematic representation of an shRNA guide loaded into argonaute. **B)** shRNA predicted potencies (data-point colors) for all human genes with-respect to gene length (x-axis) and %GC content (y-axis), as predicted under the normal (left) and 1U (right) strategies. **C)** Histogram of ~26,000 emitted values from a mixed-gaussian model fitted to the shERWOOD-1U selected shRNA sensor measurements. **D)** Clustering of the shERWOOD-1U selected shRNA sensor measurements by the mixed-gaussian model into poor (blue) and potent (red) shRNA classes. **E)** Histogram of sensor-scores for shRNAs selected to train the 1U-classifier algorithm. Blue bars represent the weak shRNAs and red bars represent the potent shRNAs. The training set was selected by applying a 70% confidence cutoff to the clustering data described in D.

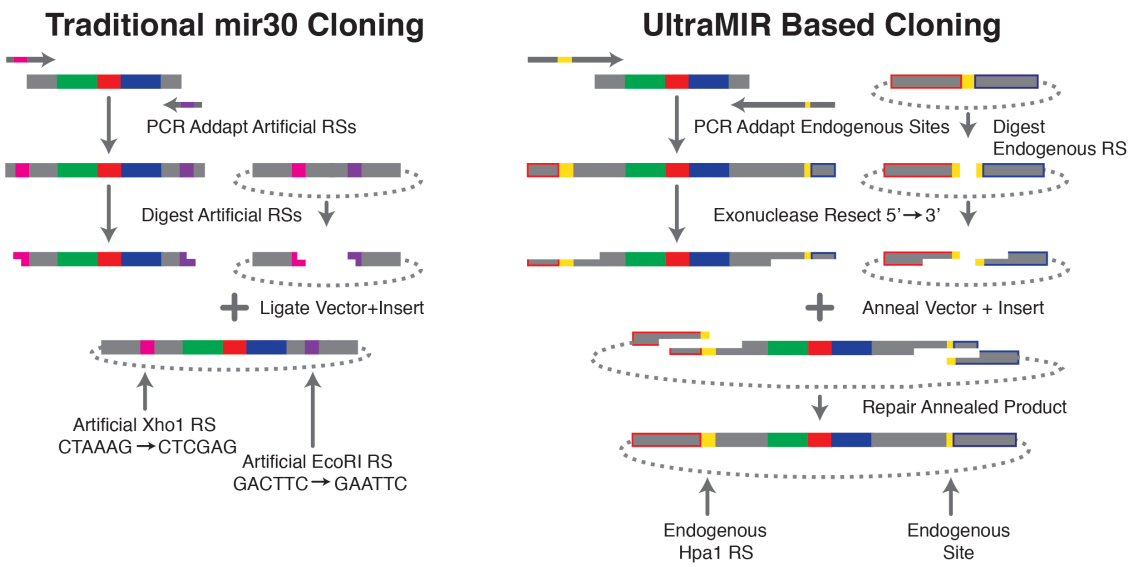

Figure S4, related to Figure 4: Schematic representation of the cloning schemes for traditional miR30 and ultramiR shRNA scaffolds.

Supplemental Methods

Extraction of shRNA efficacy from Sensor Data

To define a potency measurement for each shRNA, a matrix was constructed wherein rows correspond to shRNAs and columns represents the enrichment level of each shRNA at each iteration of the sensor (log_2 fold change with respect to the initially infected shRNA population). Columns of the matrix were then mean centered. Following this, principal component coefficients were extracted using the singular value decomposition (SVD) algorithm. Scores for each shRNA were then extracted by multiplying their sort values in the mean centered matrix with the first column of coefficient matrix (this corresponds to the first principle component loadings).

Linear Regression Analysis of Position Pairs

For a given position-pair, for each combination of nucleotides, a binary matrix was developed that represented, for each shRNA (rows in the matrix), whether the first nucleotide was present at the first position (first column), whether the second nucleotide was present at the second position (second column) and whether both nucleotides were present at both positions (third column). For example, when assessing the combination of Adenine at position 1 and Guanine at position 2, for each shRNA, if Adenine is located at position 1 then the first column assigned a value of 1 (zero if not). If the shRNA contains a Guanine at position II, the second column is assigned a value of 1 (zero if not). Finally, if Adenine is present at position 1 and Guanine at position 1, then column 3 is assigned a value of 1 (zero if not).

For each position-pair/nucleotide-combination, linear regression was applied to develop two models. In the first model only the first two columns of the coded matrix were included as inputs, whereas in the second model all columns were included. Following this, the sum-squared errors (SSEs) of the two models were compared via a rank sum test, and if the second model showed an increase in predictive capacity (p-value <0.05), the corresponding position pair score was incremented by one. The final predictive value of each position pair is the total number of nucleotide combinations that were found to be predictive at those positions (minimum of zero maximum of 16).

Linear Regression Analysis of Position Triplets

For a given position-triplet, for each triplet of nucleotides, a binary matrix was developed in a manner similar to described above for position pairs. However in these matrices there is a column representing each individual nucleotides presence at its corresponding position, each pairwise-nucleotide combination's presence at their corresponding positions and the triplet of nucleotide's presence at the corresponding positions (for a total of 7 columns).

For each position-triplet/nucleotide-combination, linear regression was applied to develop two models. In the first model only the first six columns of the coded matrix were included as inputs, whereas in the second model all columns were included. Following this, the sum-squared errors (SSEs) of the two models were compared via a rank sum test, and if the second model showed an increase in predictive capacity (p-value <0.05), the corresponding position-triplet score was increased by one. The final predictive value of each position triplet is the total number of nucleotide combinations that were found to be predictive at those positions (minimum of zero maximum of 64).

A Heuristic for Maximizing the Number of Transcripts Targeted Per Gene

We've developed a set of heuristics for selecting target regions that ensures the majority of transcripts are targeted, while maintaining sufficient predictive space for the identification of potent shRNAs. For a target gene, we search for genomic regions (including splice sites) that are represented in at least 80% of transcripts. If these areas's lengths sum to at least 250 bases, we select these as the target regions for the gene. If, however, there is no such set of regions, we iteratively remove the smallest isoform from consideration and search for a set of sequences that are shared by at least 80% of the remaining transcripts (whose summed length is greater than 250 bp). This process continues iteratively until a set of regions is identified, or only a single transcript remains. In the later case, shRNAs are predicted for each individual transcript (Figure S2B). The removal of short transcripts as a step in the heuristic is sub-optimal, however this step is necessary to maintain a search space that allows for potent shRNA selection.

A Heuristic for Minimizing the Number of Off-Target Effects

We've developed a strategy that minimizes off-target effects, and takes into account the fact that paralogues, with nearly identical sequences, likely share function and, thus should be targeted in parallel in large genetic screens (with the assumption that the particular paralogue whose targeting results in the phenotype of interest can be identified during validation experiments).

The algorithm was designed with the goal of constructing of a genome-wide library harboring four shRNAs per gene. For each gene, the top eight shRNAs within the target regions, as defined above, are assigned a rank of 1-8 based on shERWOOD scores. Following this, shRNAs are mapped to the transcriptome using the bowtie algorithm, allowing up to three mismatches outside of the shRNA-seed, and classified as a non-multi-mapper or multi-mapper (Langmead et al., 2009). If the ranks 1-4 shRNAs are all non-multi-mappers, they are selected for the library. If the ranks 1-8 shRNAs are all multi-mappers, they are assumed to target a set of paralogues, and they are selected for the library. If some (but not all) of the ranks 1-4 shRNAs are multi-mappers, and some of the ranks 5-8 shRNAs are non-multi-mappers, the algorithm selects a delivery set, equal to the ranks 1-4 shRNAs, and then iterates as follows: replace the lowest rank multi-mapping shRNA in the

delivery set with the highest ranking non-multi-mapping shRNA outside of the delivery set. Continue until no non-multi-mappers exist outside of the delivery set (Figure S2C).

Analysis of small RNA Processing

All small RNA libraries were constructed using Illumina's sRNA cloning kit. Libraries were sequenced on an Illumina MiSeq. Following sequencing, reads were aligned to an bowtie index containing all endogenous microRNA guide sequences as well as sequences corresponding to the shRNA being studied using the bowtie algorithm (allowing three mismatches). Ilumina adapter sequences were appended to each sequence during the construction of the index. Reads were then normalized between libraries as their $\log_2$ fold difference to the 66th quantile of the count distribution of the endogenous microRNAs.

RNAseq Library Construction and Analysis

Total RNA was purified and DNAse treated using the Qiagen RNeasy Mini Kit. RNA integrity (RNA Integrity score > 9) and quantity was measured on an Agilent Bioanalyzer (RNA Nano kit). The NuGEN Ovation RNA-Seq V2 protocol was carried out on 100 ng of total RNA. cDNA was fragmented using the Covaris LE220 sonicator according to the manufacturer's instruction to yield a target fragment size of 200 bp. The fragmented cDNA was subsequently processed using the NuGEN Ovation Ultralow DR Multiplex System.

Each sample was sequenced on the Illumina HiSeq-2.0 platform, generating 76 nt single-end (SE) reads. Reads were aligned to the mm10 genome using the Bowtie-2 alignment tool under default parameters (Langmead and Salzberg, 2012). Mapped reads were then assigned to genes using HTSeq-count (using the latest version of RefSeq.gtf file for gene coordinates)(Anders et al., 2014). Resultant counts were then normalized and compared using DESeq (Anders and Huber, 2010). For a gene to be considered over-expressed it had to show an at least 2-fold change with FDR < 0.05.

Supplemental References

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome biology *11*, R106.

Anders, S., Pyl, P.T., and Huber, W. (2014). HTSeq; A Python framework to work with high-throughput sequencing data.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods *9*, 357-359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology *10*, R25.