

# A Computational Algorithm to Predict shRNA Potency

Simon R.V. Knott,<sup>1,3</sup> Ashley R. Maceli,<sup>1,3</sup> Nicolas Erard,<sup>1,3</sup> Kenneth Chang,<sup>1,3</sup> Krista Marran,<sup>1</sup> Xin Zhou,<sup>1</sup> Assaf Gordon,<sup>1</sup> Osama El Demerdash,<sup>1</sup> Elvin Wagenblast,<sup>1</sup> Sun Kim,<sup>1</sup> Christof Fellmann,<sup>1,4</sup> and Gregory J. Hannon<sup>1,2,\*</sup>

<sup>1</sup>Watson School of Biological Sciences, Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA

<sup>2</sup>Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK

<sup>3</sup>Co-first author

<sup>4</sup>Present address: Mirimus, Inc., 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA

\*Correspondence: [hannon@cshl.edu](mailto:hannon@cshl.edu)

<http://dx.doi.org/10.1016/j.molcel.2014.10.025>

## SUMMARY

The strength of conclusions drawn from RNAi-based studies is heavily influenced by the quality of tools used to elicit knockdown. Prior studies have developed algorithms to design siRNAs. However, to date, no established method has emerged to identify effective shRNAs, which have lower intracellular abundance than transfected siRNAs and undergo additional processing steps. We recently developed a multiplexed assay for identifying potent shRNAs and used this method to generate ~250,000 shRNA efficacy data points. Using these data, we developed shERWOOD, an algorithm capable of predicting, for any shRNA, the likelihood that it will elicit potent target knockdown. Combined with additional shRNA design strategies, shERWOOD allows the *ab initio* identification of potent shRNAs that specifically target the majority of each gene's multiple transcripts. We validated the performance of our shRNA designs using several orthogonal strategies and constructed genome-wide collections of shRNAs for humans and mice based on our approach.

## INTRODUCTION

The discovery of RNAi promised a new era in which the power of genetics could be applied to model organisms for which large-scale studies of gene function were previously inconvenient or impossible (Berns et al., 2004; Brummelkamp et al., 2002; Chuang and Meyerowitz, 2000; Fire et al., 1998; Gupta et al., 2004; Hannon, 2002; Kamath et al., 2003; Kambris et al., 2006; Paddison et al., 2004; Sánchez Alvarado and Newmark, 1999; Svoboda et al., 2000; Timmons and Fire, 1998; Tuschl et al., 1999; Zender et al., 2008). It quickly became clear that implementing RNAi, especially on a genome-wide scale, could be challenging. This was particularly true for applications in mammalian cells in which discrete sequences, in the form of small interfering RNAs (siRNAs) or short hairpin RNAs (shRNAs),

were used as silencing triggers (Brummelkamp et al., 2002; Elbashir et al., 2001; Paddison et al., 2002). The overall degree of knockdown achieved was found to vary tremendously depending on the precise sequence of the small RNA that is loaded into the RNAi effector complex (RISC) (Chiu and Rana, 2002; Khvorova et al., 2003; Schwarz et al., 2003). However, the nature of sequence and structural motifs that favor RISC loading and high turnover target cleavage has yet to be fully revealed (Ameres and Zamore, 2013).

Early studies aimed at optimizing RNAi in mammals used endogenous microRNAs as a guide for the design of effective artificial RNAi triggers (Khvorova et al., 2003; Reynolds et al., 2004; Schwarz et al., 2003; Ui-Tei et al., 2004; Zeng and Cullen, 2003). Canonical microRNAs are processed by a two-step nucleolytic mechanism (Seitz and Zamore, 2006). The initial cleavage of the primary microRNA (miRNA) transcript in the nucleus by the microprocessor yields a short, often imperfect hairpin loop, the pre-miRNA (Denli et al., 2004; Lee et al., 2003). This is exported to the cytoplasm, where a second cleavage by Dicer and its associated cofactors yields a short duplex of ~19–20 nucleotides with two nucleotide 3' overhangs (Bernstein et al., 2001; Grishok et al., 2001; Hutvagner et al., 2001; Ketting et al., 2001; Lund et al., 2004; Yi et al., 2003). This duplex serves as a substrate for preferential loading of one strand into Argonaute proteins in the context of RISC (Hammond et al., 2001; Hutvagner and Zamore, 2002; Khvorova et al., 2003; Martinez et al., 2002; Schwarz et al., 2003).

An examination of the sequences of endogenous miRNAs indicated that thermodynamic asymmetry between the two ends of the short duplex was a strong predictor of which strand would be accepted by Argonaute as the "guide" (Khvorova et al., 2003; Schwarz et al., 2003). Applying this insight to artificial triggers, initially in the form of siRNAs, validated the generality of this observation, and thermodynamic asymmetry became a key guiding principle of both siRNA and shRNA design (Reynolds et al., 2004; Silva et al., 2005). Subsequent studies of the structure of the Ago-small RNA complex have also indicated a sequence preference for a 5' terminal U that fits into a binding pocket in the mid-domain of the Argonaute protein (Seitz et al., 2008; Wang et al., 2008).

In many ways, siRNAs gain entry into RISC in mammals by simulating the end product of the two-step miRNA processing

pathway. shRNAs, which mimic either the primary miRNA or pre-miRNA, must be processed nucleolytically prior to RISC loading (Brummelkamp et al., 2002; Cullen, 2006; Paddison et al., 2002). Therefore, shRNAs are likely subject to additional constraints that lead to efficient recognition by Drosha and Dicer. We do not yet understand the selection rules for effective flux through the miRNA biogenesis pathway and, therefore, cannot predict *ab initio* which transcripts will produce small RNAs. However, studies of Drosha in particular have implicated patterns of conservation and base pairing in the basal stem, regions adjacent to the Drosha cleavage site, as determinants of efficient pre-miRNA cleavage (Auyeung et al., 2013; Chen et al., 2004; Han et al., 2006; Seitz and Zamore, 2006). Elements within the hairpin loop have also been shown to have an impact on both Drosha efficiency and its site preference (Han et al., 2006; Zhang and Zeng, 2010).

Several attempts have been made to extract predictive rules for the design of effective small RNAs from endpoint silencing data. The first serious attempt applied artificial neural networks to a set of ~2,000 paired data points, associating the sequence of siRNA guides with a corresponding knockdown measurement (established using fluorescent reporters) (Huesken et al., 2005). Experience in the field supported the effectiveness of BIOPREDSi. However, access to the algorithm eventually became impossible. The same data set was subsequently used to produce a second algorithm, Designer of Small Interfering RNA (DSIR), which included additional input variables (the frequency of each nucleotide, each 2-mer, and each 3-mer within the guide) (Vert et al., 2006). To accommodate this large number of parameters, linear modeling was performed using Lasso regression (a form of linear regression that iteratively decreases the use of nonpredictive variables in the linear model) (Tibshirani, 1996).

siRNA design algorithms could be applied for the design of shRNAs, and these did inform the design of genome-wide shRNA collections (Berns et al., 2004; Paddison et al., 2004). However, the prognostic power of siRNA design algorithms is compromised for shRNA design. shRNAs, expressed from RNA polII or polIII promoters, reach lower intracellular concentrations than transfected, synthetic siRNAs (Berns et al., 2004; Paddison et al., 2004). Moreover, shRNAs have additional constraints for effective processing. Therefore, it was imperative that shRNA-specific algorithms be developed.

The generation of accurate siRNA design algorithms was only made possible with the creation of large training data sets. So far, a corresponding shRNA data set has been lacking. Recently, we developed a “sensor” method that allows for the parallel assessment of shRNA potencies on a massive scale (Fellmann et al., 2011). Using the sensor approach, we interrogated ~250,000 shRNAs for their effectiveness in the reporter setting. We used this data set to train a machine learning algorithm for potent shRNA prediction. We tested this algorithm, which we termed, shERWOOD, both at the level of individual shRNAs and at the level of optimized shRNA mini libraries. We demonstrated that, by applying computational shRNA selection in combination with a set of target selection heuristics and an optimized micro-RNA scaffold, we are able to create highly potent shRNAs. We built upon this result to design and construct next-generation shRNA

libraries targeting the constitutive exomes of mice and humans. Predictions for other organisms and custom shRNA designs are also made available via a web-based version of shERWOOD.

## RESULTS

### Neighboring Positions of the Target Sequence Are Predictive of shRNA Strength

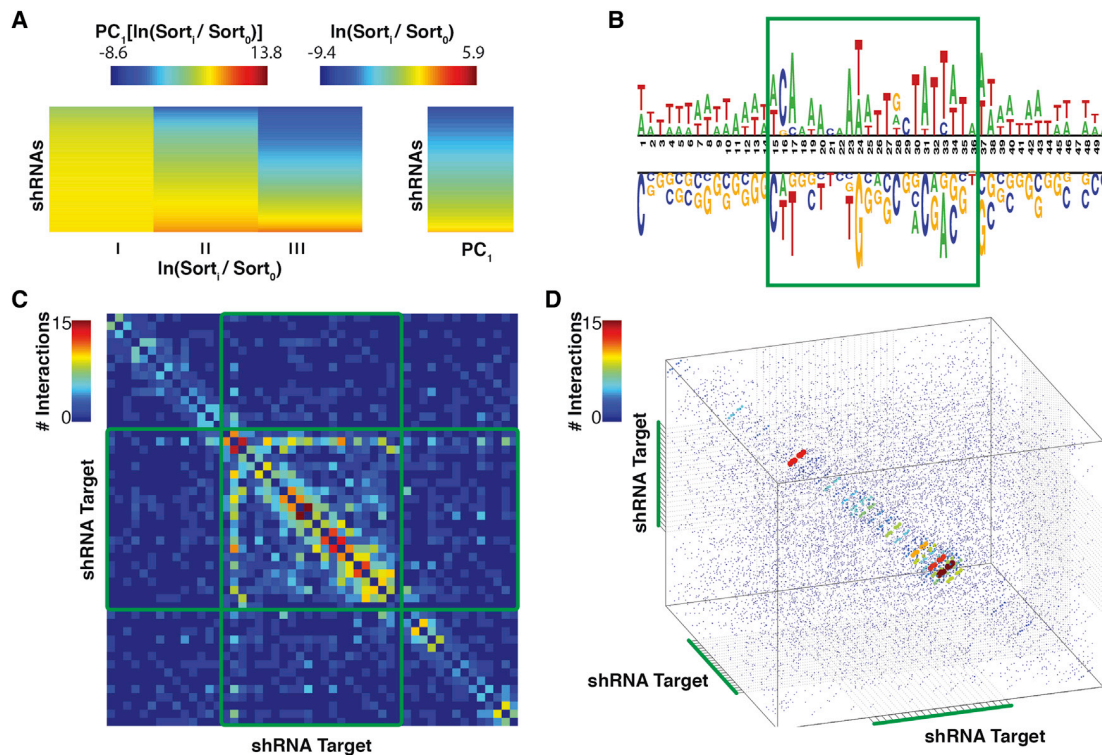
As a prelude to creating an shRNA design algorithm, we first developed a large-scale sensor data set in which shRNA potency was measured and associated with sequence information. To perform the assay, we synthesized 12 sets of ~25,000 constructs that included a doxycycline-inducible shRNA and a GFP-tagged shRNA target sequence located downstream of a constitutive promoter (Fellmann et al., 2011). Libraries were packaged and infected (at single copy) into a reporter cell line. In the absence of doxycycline, GFP was detectable in each cell. However, in the presence of doxycycline, the shRNAs became expressed, and the resultant GFP signal was reduced in a manner proportional to shRNA potency. Using fluorescence-activated cell sorting, cells with low GFP levels, in the presence of drug, were gathered and analyzed via next generation sequencing (NGS) to determine which shRNAs became enriched (i.e., which shRNAs have high potency). Operating iterative cycles of this assay has been shown to identify extremely potent constructs (Fellmann et al., 2011).

We next wished to extract which sequence characteristics were most predictive of shRNA efficacy. This subset of characteristics could then be employed as inputs during machine learning. We first developed a method to consolidate the different sensor data points into a single value for each shRNA (see [Supplemental Information](#) available online). These accurately capture the enrichment pattern of individual iterations of the sensor in one single value, therefore allowing downstream machine learning to proceed more easily (Figure 1A). Analysis of the coefficients used to consolidate the sensor data shows that information from the final sensor iteration contributes the most to the final potency value. However, information from the second iteration is also included (Figure S1A).

To distinguish discretely between strong and weak shRNAs, we applied an empirical Bayes moderated t test to the shRNA potency measurements extracted from two biological replicates (Smyth, 2004). Strong and weak shRNAs were those that were enriched or depleted, respectively, with a false discovery rate (FDR) < 0.05.

To test individual nucleotide positions for their predictive capacity, we compared, at each position in the target sequence, each nucleotide's enrichment and or depletion levels in the potent compared with the weak shRNAs (Figure 1B; Figure S1B; binomial test, FDR < 0.05; Vacic et al., 2006). In general, low GC content is predictive of high efficacy, with the exception of the third nucleotide inside the guide target, which shows a strong selection for cytosine. Also of note is a lack of enrichment for thymidine at the 22<sup>nd</sup> position of the guide target (corresponding to the first position of the guide). This arose because our input data sets were derived from shRNAs preselected by DSIR.

We next tested whether any pairs of positions had predictive capacity for shRNA strength beyond what was expected based



### Figure 1. Identification of Sequence Characteristics Predictive of shRNA Efficacy

(A) shRNA score determination via sensor NGS data. On the left is a heatmap representation of normalized shRNA read counts for each on-dox sensor sort. The right panel represents shRNA potencies, calculated by extracting the first principal component of the left panel matrix. (B) A nucleotide logo representing enriched (top) and depleted (bottom) nucleotides ( $p < 0.05$ ) in potent shRNAs. (C) A heatmap demonstrating the predictive capacity (with respect to shRNA potency) of each pair of positions within the target region. Heatmap cells are colored to represent the number of nucleotide combinations that were significantly predictive ( $p < 0.05$ ) at each position-pair. (D) The predictive capacity of each triplet of positions within the target region. Data point colors and sizes represent the number of nucleotide triplets that were significantly predictive ( $p < 0.05$ ) at each position triplet.

on their individual predictive power. To calculate a measurement for each position pair, we applied linear regression to identify synergistic predictive capacity ( $p$  value  $< 0.05$ ; [Supplemental Experimental Procedures](#)). Following this, each position pair was assigned a value equal to the sum of nucleotide combinations that were predictive of shRNA potency when assessed at the two positions (Figure 1C; Figure S1C). For a given position within the target, the most predictive partner is the neighboring nucleotide. An exception to this trend is observed in the positions corresponding to the shRNA guide seed, where predictive position pairs are also observed in nucleotides separated by up to four bases.

Finally, we wished to determine whether triplets of positions showed a similar trend to that observed in the pairwise analysis. For this, we performed a modified version of the linear regression tests described above, where triplets instead of pairs of nucleotides were assessed for synergistic predictive capacity. As with the pairwise analysis, neighboring triplets of positions within the target show strong predictive power compared with triplets of nonneighboring positions (Figure 1D). Furthermore, the distance between predictive triplets is also extended slightly in the guide seed region of the shRNA.

### A Sensor-Based Computational Algorithm to Predict shRNA Efficacy

Because sequence-based characteristics correlated with shRNA efficiency, we sought to apply machine learning to the sensor-derived efficacy measurements. The goal was to develop a computational algorithm that would predict, for any target sequence, the potency of a corresponding shRNA. We reasoned that the best machine learning tool to apply to this task was random forest regression analysis (Breiman, 2001). The reasons for this decision were two-fold. First, there is no decrease in the accuracy of random forests when the number of input variables is large. Second, the architecture of the algorithm takes into account increases in accuracy that can be achieved by analyzing combinations of input variables.

Our training data set was of two distinct types. One comprised an unbiased set of shRNAs that tiled every nucleotide of nine genes (Fellmann et al., 2011). A second comprised a larger set of shRNAs preselected by the DSIR algorithm (described above). We therefore chose to separate data corresponding to each input class and to train separate forests. We also chose to separate data based on the 5' nucleotide of the guide. This was done for two reasons. First, previous studies, supported

by structural insights, had suggested that the 5' nucleotide of the guide was a prominent determinant of small RNA potency (Fellmann et al., 2011; Frank et al., 2010; Khvorova et al., 2003; Reynolds et al., 2004). Therefore, training forests individually for shRNAs initiating with each base focused the prediction process on additional determinants. Moreover, the DSIR-based predictions were already heavily biased toward U and A at the 5' position. In fact, the bias was so strong that we did not have sufficient data to train 5'C and 5'G forests for these data sets. This meant that, in the first pass, we trained six independent modules.

In each module, input data were composed of individual base information as well as all neighboring pairs of bases throughout the guide sequence. In addition, the set of triplet position/nucleotide combinations found to be predictive, as assessed by linear regression, were also included (Figure 1D). After training each of the modules, we sought to determine which input variables were relied most heavily upon. For each module, each variable was permuted across observations, and the resultant reduction in predictive capacity was recorded at each regression tree. The resultant changes were then averaged across trees, and that mean was normalized by their standard deviation. The triplet variables were heavily relied upon (Figure S2A), particularly the triplet corresponding to shRNA guide positions 2–4.

To consolidate these modules, a second-tier random forest was trained using the first-tier outputs, the corresponding shRNA guide base information, and a set of thermodynamic properties extracted from each shRNA (e.g., enthalpy, entropy). We named the compiled algorithm shERWOOD.

To test the prognostic power of shERWOOD, we took advantage of the unbiased nature of the tiled shRNA sensor data. For each of the nine genes represented, we independently trained a shERWOOD algorithm without the data corresponding to that gene. We could then test shERWOOD performance against experimental data in a manner that was not skewed by the use of those data for training. We saw an overall Pearson correlation of 0.72 between experimentally derived potency measurements and computational predictions (Figure 2A). For comparison, DSIR achieves a correlation of 0.4, and a prior shRNA prediction algorithm trained on a subset of the sensor data used in this study achieves 0.56 (Matveeva et al., 2012; Vert et al., 2006). This indicates that shERWOOD achieves a roughly 180% increase in performance over currently existing siRNA prediction algorithms and a 126% increase in efficacy over existing shRNA-specific prediction algorithms.

We supplemented shERWOOD with additional heuristics to maximize the probability of successfully reducing protein levels in most cell and tissue types. The complex nature of alternative splicing patterns provided a strong motivation for directing shRNAs against constitutive exons. We therefore developed a strategy that iteratively searches for regions within a gene that are shared by at least 80% of transcripts (Supplemental Experimental Procedures). This algorithm also tests whether high-potency shRNAs have the potential to cosuppress paralogous genes. Considered together, these strategies have the potential to maximize the probability of biologically meaningful results from studies using shRNAs.

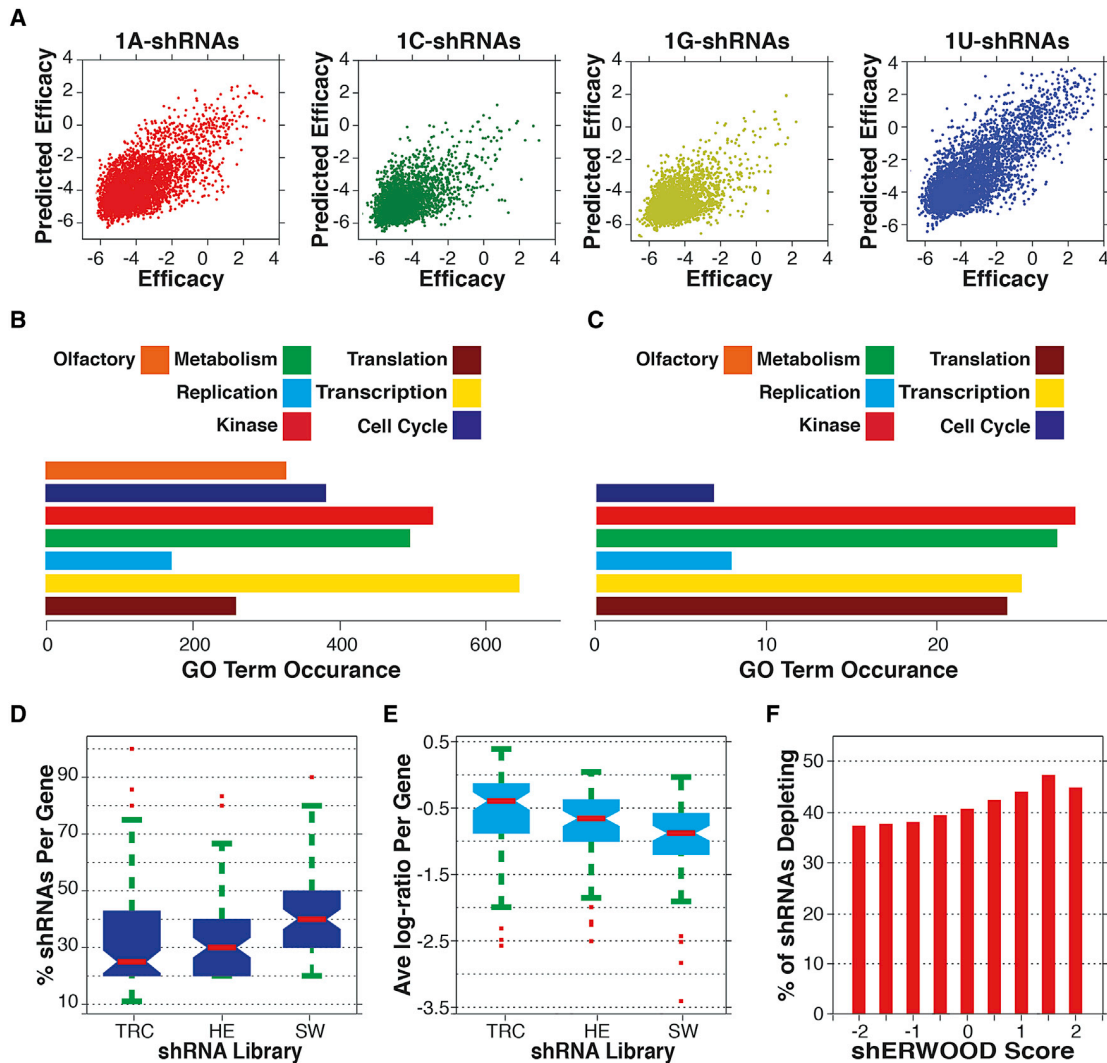
### Benchmarking shERWOOD

To assess the performance of the shERWOOD algorithm, we felt that it was necessary to test a large number of shRNAs for their biological effects because one can find anecdotal evidence for excellent performance for nearly any algorithm or strategy. We therefore chose ~2,200 genes based on their enrichment in gene ontology (GO) categories likely to impact the growth and survival of cells in culture (Figure 2B). As controls, particularly for the likelihood of off-target effects, we included 400 olfactory receptor genes. Olfactory receptors are expressed only in olfactory neurons, and even then, they display allelic choice so that only one paralog is expressed per cell. Therefore, shRNAs targeting olfactory receptors are highly unlikely to have relevant, on-target biological effects in any cell line screened *in vitro*. To benchmark the performance of shERWOOD, we compared a focused mini library predicted with this algorithm to two widely used genome-wide collections, namely The RNAi Consortium (TRC) collection distributed by Sigma-Genosys and the so-called Hanon-Elledge V3 library distributed presently by GE Dharmacon (K.C., unpublished data). To produce the shERWOOD-based library and a deeper simulation of the V3 library, we used either shERWOOD or DSIR to predict their top 10 scoring shRNAs for our test genes. The sequences of TRC shRNAs are listed on a public web portal, and we selected all listed shRNAs for each gene. In the case of TRC shRNAs, it was necessary to adapt them to a 22-base pair stem for placement into the miR-30 context.

For each test library, we synthesized 27,000 oligonucleotides in solid phase on microarrays (Cleary et al., 2004). These were cleaved, amplified, and cloned directly into a miR-30 scaffold within a murine stem cell virus (MSCV)-based retroviral vector without sequence validation. In this arrangement, the primary shRNA was transcribed from the long terminal repeat (LTR) promoter, whereas GFP and Neomycin resistance were expressed separately as a bicistronic transcription unit from the phosphoglycerate kinase promoter (PGK) (Figure S2D). Pilot sequencing showed that each library was of similar quality and representation.

Each library was infected separately into the pancreatic ductal adenocarcinoma cell line A385. Two days after infection, cells were collected for a reference time point, and, after ~12 doublings, cells were again harvested for a final time point (Supplemental Experimental Procedures). shRNA representation was determined following amplification of hairpin inserts from genomic DNA (Sims et al., 2011), and, after processing, shRNA read counts were compared between the initial and final time points (Supplemental Experimental Procedures; Figures S2E–S2G).

To enable direct comparisons between libraries, we censored the shERWOOD- and DSIR-based libraries on a per gene basis to contain the same number of hairpins as were available in the TRC library, keeping those with the best algorithmic scores. We then selected the consensus set of “essential” genes, accepting only those where at least two hairpins in each library passed the statistical threshold (FDR < 0.1). As expected, the resulting set of genes that were important for the growth and survival of A385 was depleted of olfactory receptor shRNAs



**Figure 2. Construction and Validation of an shRNA-Specific Predictive Algorithm**

(A) Consolidated cross-validation of predictions versus sensor scores for all shRNAs in the [Fellmann et al. \(2011\)](#) data set (shRNAs are separated by the guide 5' nucleotide).

(B) GO term instances associated with the targeted gene set selected for shRNA validation screens.

(C) GO term instances associated with genes for which at least two hairpins were significantly depleted in each of the TRC, Hannon-Elledge (HE), and shERWOOD (SW) validation screens.

(D) The percentage of shRNAs targeting consensus-essential genes that were depleted in each of the TRC, HE, and shERWOOD shRNA screens. The plot was made with the Matlab Boxplot function using default parameters. The edges of the box are the 25<sup>th</sup> and 75<sup>th</sup> percentiles. The error bars extend to the values  $q_3 + w(q_3 - q_1)$  and  $q_1 - w(q_3 - q_1)$ , where  $w$  is 1.5 and  $q_1$  and  $q_3$  are the 25<sup>th</sup> and 75<sup>th</sup> percentiles.

(E) Average log-fold change for shRNAs targeting consensus-essential genes (per gene) for each of the TRC, EH, and shERWOOD validation screens. The plot was made with the Matlab Boxplot function using default parameters. The edges of the box are the 25<sup>th</sup> and 75<sup>th</sup> percentiles. The error bars extend to the values  $q_3 + w(q_3 - q_1)$  and  $q_1 - w(q_3 - q_1)$ , where  $w$  is 1.5 and  $q_1$  and  $q_3$  are the 25<sup>th</sup> and 75<sup>th</sup> percentiles.

(F) The percentage of shRNAs corresponding to consensus-essential genes that, for any given shERWOOD score, were depleted in the shERWOOD validation screen.

(Figure 2C). In contrast, the set of consensus-essential genes was enriched for GO terms associated with translation.

To benchmark shRNA selection strategies against each other, we determined the percentage of shRNAs in each mini library that scored for each consensus essential gene. For the TRC library, 24% of shRNAs achieved significant depletion, whereas

31% of DSIR-predicted sequences and 40% of shERWOOD-based hairpins scored (Figure 2D). We also considered performance from the perspective of median log-fold depletion. For the TRC collection, the average log-fold change was  $-0.4$ . For DSIR, this rose to  $-0.62$ , and it increased further to  $-0.78$  for shERWOOD shRNAs (Figure 2E). We note that this type of

analysis slightly favors the library with the weakest overall shRNAs because it will be this collection that sets entry criteria for the consensus-essential gene set.

To assess whether shERWOOD scores were a proxy for shRNA potency, we examined the relationship between the shERWOOD score and the probability of being significantly depleted for each consensus-essential gene. For this, we analyzed all ten shERWOOD predictions using a sliding scale of shERWOOD score cutoffs (Figure 2F). As an example, considering shRNAs with a score greater than 0.5, the likelihood that an shRNA will be depleted if it targets one of our consensus essential genes is 42%. Again, this underestimates the information content of shERWOOD scores because, in the cumulative plot shown, the minimum number of scoring hairpins for a given gene, irrespective of scores, is two (i.e., 20%).

### Structure-Guided Insights Expand the shRNA Prediction Space

Regardless of the accuracy of predictive models, we sometimes found it difficult to identify potent shRNAs because of search space restrictions imposed by sequence constraints (e.g., GC content), gene length, or the complexity of alternative splicing patterns. We therefore sought ways to expand the sequence space to which we could apply the shERWOOD approach. Analysis of miRNA seed sequences as well as other data have suggested that the first base of the small RNA guide does not pair with its target (Lai, 2002; Lewis et al., 2005; Yuan et al., 2006). Structural studies have supported this hypothesis by showing that the first base of the guide is tightly bound within a pocket in the mid-domain of Ago proteins (Figure S3A; Elkayam et al., 2012; Frank et al., 2010; Nakanishi et al., 2012; Wang et al., 2008). Because the first base of the guide is a strong contributor to shRNA efficacy, we reasoned that we could expand the range of possible effective shRNAs by simply changing the first base of all potential guides to a U, promoting their binding to RISC, and, theoretically, not altering target site choice. We will henceforth refer to this as the 1U strategy. A simulated construction of a human genome-wide shRNA library demonstrates that, when this strategy is implemented, predicted shRNA potencies increase dramatically, particularly for short GC-rich genes (Figure S3B).

To test the 1U strategy in a high-throughput manner, we constructed a sensor library where the top 15 shRNAs targeting a set of ~2,000 “druggable” genes were predicted using the 1U strategy. The constructs were designed so that the shRNAs contained the 1U conversion and the target sites contained the endogenous base. shRNA potencies were extracted as described in Figures 1 and 3A. The distribution indicates that ~50% of the shRNAs were strong or very strong (knockdown efficiency > 75%) based on the scores of control shRNAs that were assayed in parallel. When shRNAs were separated into native and artificial 1U sets and the score distributions were plotted, we were surprised to see a significant reduction in the efficacy of the nonnative 1U shRNAs (Figure 3B; Wilcoxon rank-sum test,  $p$  value < 0.01). This was strongly suggesting that RISC interacts not only with the 1U of the guide but also with the first base of the target site.

We therefore stratified 1U shRNAs into four sets based on their endogenous 5' nucleotide (Figure 3C). This analysis indicated

that only a subset of shRNAs performs well when a 1U switch is made (based on the bimodal distributions for endogenous 1A, 1C, and 1G shRNAs) but that the subset that does perform well is predicted to be quite efficacious by the sensor assay. This bimodal distribution is not observed for shERWOOD-selected endogenous 1U shRNAs, and we see that the majority of this shRNA class are efficient.

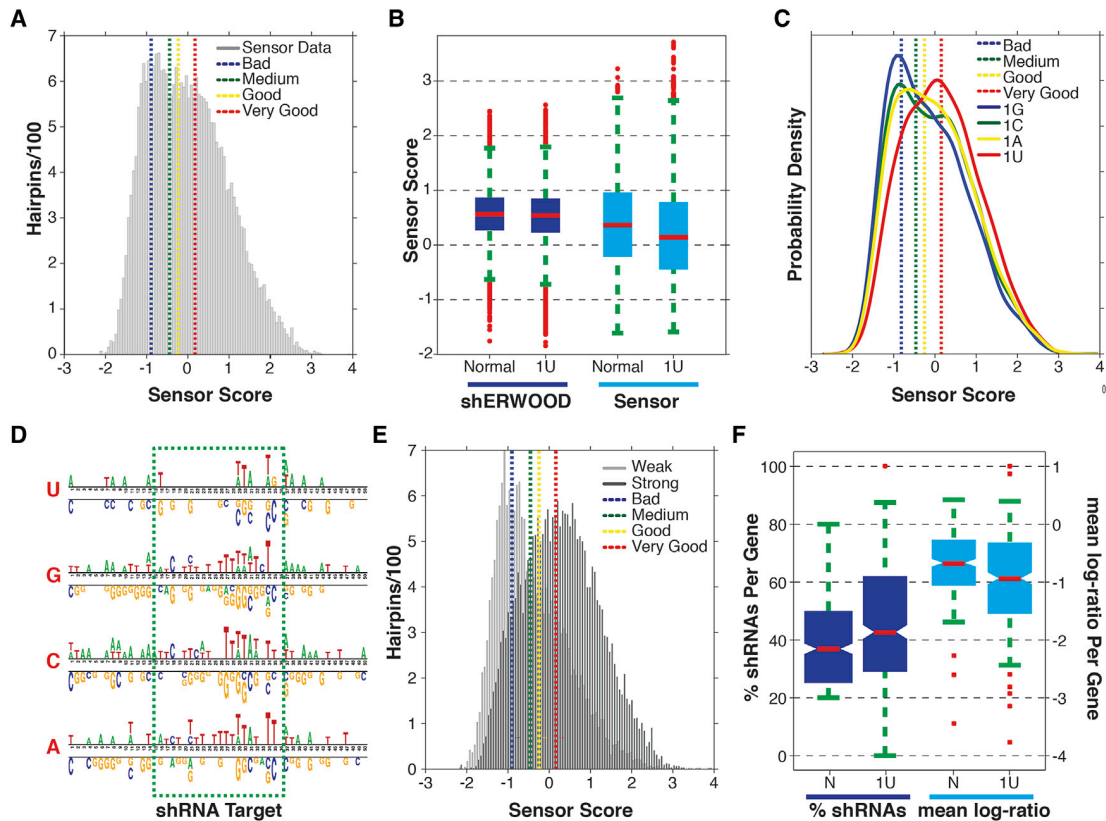
Given these results, we sought to determine whether we could predict sequences for which a 1U conversion would result in a highly effective shRNA. We fit a Gaussian mixture model to the sensor scores (Figure S3C) and applied this model to assign shRNAs into one of the two resultant populations (Figure S3D). Following clustering, we applied a binomial test separately for shRNAs where the endogenous base was 1A, 1C, 1G, and 1U to determine whether any nucleotides were enriched/depleted in the strong shRNAs with respect to weak shRNAs. All sets show a strong enrichment for U in the target region corresponding to the shRNA guide positions 3, 7, and 8 (Figure 3D). There is also a strong selection for Cs in the target region—corresponding position 19 of the endogenous 1A, 1C, and 1G shRNA guides.

These results prompted us to develop a computational algorithm that could both select the strongest endogenous 1U shRNAs and identify which endogenous 1C, 1G, and 1A shRNAs were likely to yield potent 1U-converted molecules. Data points for which the mixed Gaussian clustering resulted in less than a 70% confidence group assignment were censored (Figure S3E). We trained a random forest using the 22 nucleotides of the endogenous base as well as all neighboring pairs of nucleotides as input and the corresponding 1U conversion sensor scores as output. The algorithm was able to achieve 80% specificity while maintaining 50% sensitivity. Notably, we were able to increase the specificity to 85% through the supplemental application of previously reported rules for shRNA selection (Figure 3E; Fellmann et al., 2011; Matveeva et al., 2012).

To validate this addition to the shERWOOD algorithm, we performed an shRNA screen as described above, in which shRNAs were selected with the 1U strategy with or without applying the additional filter. We also applied this variant of the algorithm to the shRNA screen described in Figure 2. We found that, when additional filters were applied to the 1U strategy, shRNAs targeting our set of consensus-essential genes showed a significantly higher percentage of depleted shRNAs per gene (Wilcoxon rank-sum test,  $p$  < 0.01) and a stronger mean depletion, as measured by log ratio (Wilcoxon rank-sum test,  $p$  < 0.01; Figure 3F).

### A Variant miRNA Scaffold Increases shRNA Potency

Recently completed studies of evolutionarily conserved determinants of Drosha processing raised the possibility that the placement of the EcoRI site in the standard miR-30 scaffold might have reduced the efficiency of pre-miRNA cleavage (Auyeung et al., 2013). Others have reported that alternatively positioning the EcoRI site within the scaffold increases small RNA levels, presumably by improving biogenesis. This led to overall more potent knockdown (Fellmann et al., 2013). We therefore chose to create shRNAs by Gibson assembly, removing restriction sites altogether from the shRNA scaffold (Figure S4). We felt that this was the surest way to avoid any unanticipated effects of altering processing signals. We termed this scaffold ultramiR.



**Figure 3. Structure-Guided Maximization of shRNA Prediction Space**

(A) Histogram of sensor scores for the top 15 shRNAs as identified by the shERWOOD-1U strategy, targeting ~2000 druggable genes. Overlaid are the mean sensor scores for control shRNAs representing poor, medium, potent, and very potent shRNAs (with mean knockdown efficiencies of 25%, 50%, 75%, and >90%, respectively).

(B) The distribution of shERWOOD-1U prediction scores for shRNAs where endogenous 1U shRNAs are separated from endogenous non-1U shRNAs. Sensor scores for endogenous 1U and non-1U shRNAs are displayed on the left. The plot was made with the Matlab Boxplot function using default parameters. The edges of the box are the 25<sup>th</sup> and 75<sup>th</sup> percentiles. The error bars extend to the values  $q3 + w(q3 - q1)$  and  $q1 - w(q3 - q1)$ , where  $w$  is 1.5 and  $q1$  and  $q3$  are the 25<sup>th</sup> and 75<sup>th</sup> percentiles.

(C) Distribution of sensor scores for shERWOOD-1U-selected shRNAs, separated by endogenous guide 5' nucleotides.

(D) A nucleotide logo representing enriched (top) and depleted (bottom) nucleotides ( $p < 0.05$ ) in potent shERWOOD-1U-selected shRNAs (separated by endogenous guide 5' nucleotides).

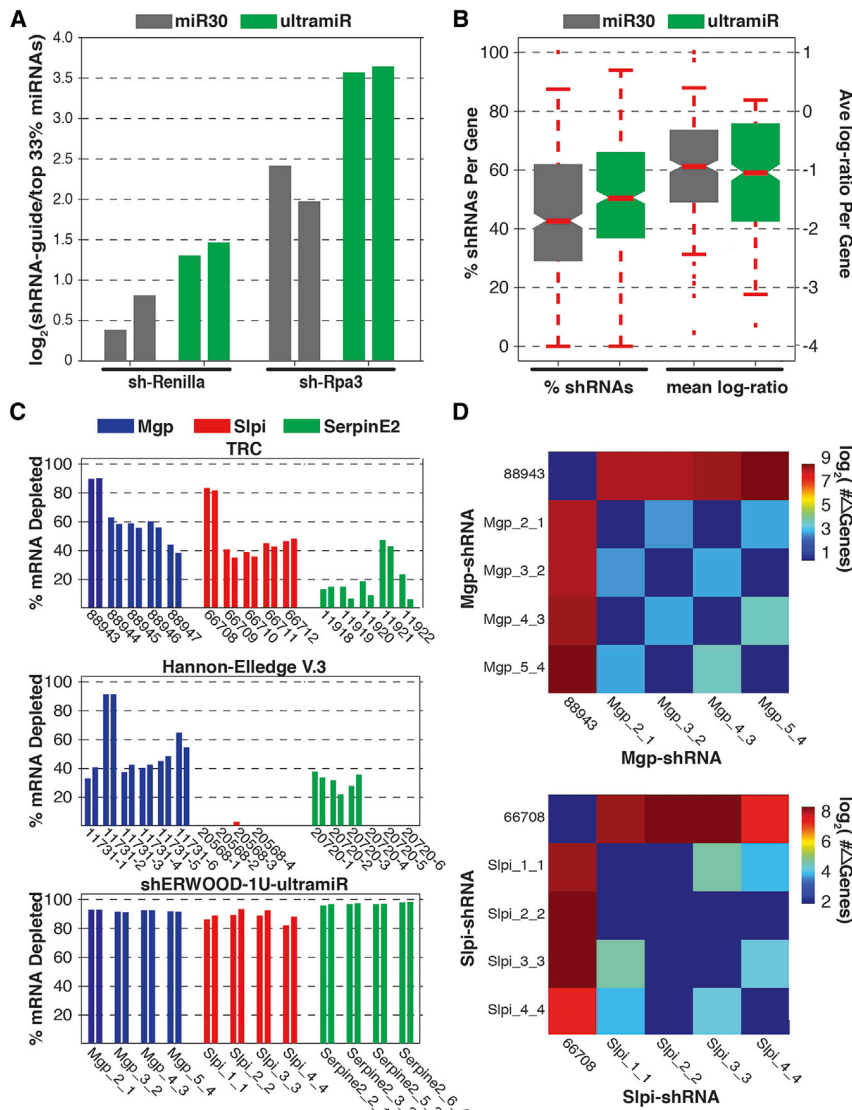
(E) The distribution of sensor scores for shRNAs classified as weak and strong by a random forest classifier trained on the shERWOOD-1U sensor data.

(F) The distributions of the percentage of shERWOOD- and shERWOOD-1U-selected shRNAs targeting consensus-essential genes that were depleted in validation screens (left). In addition, normalized log-fold changes of shRNAs, identified under each selection scheme, are displayed (right). The plot was made with the Matlab Boxplot function using default parameters. The edges of the box are the 25<sup>th</sup> and 75<sup>th</sup> percentiles. The error bars extend to the values  $q3 + w(q3 - q1)$  and  $q1 - w(q3 - q1)$ , where  $w$  is 1.5 and  $q1$  and  $q3$  are the 25<sup>th</sup> and 75<sup>th</sup> percentiles.

To test ultramiR performance, we inserted two shRNAs, targeting luciferase or mouse RPA3, into the standard scaffold and into ultramiR. These constructs were packaged and infected in duplicate (multiplicity of infection [MOI] < 0.3) into human embryonic kidney 293T (HEK293T) cells and the modified DF1 reporter line used for the sensor screen, respectively (Fellmann et al., 2011). Following selection for singly infected cells, we analyzed the levels of mature shRNAs by small RNA sequencing (Malone et al., 2012). shRNA guide counts were normalized across libraries by determining their log-fold enrichment relative to the 66<sup>th</sup> quantile of endogenous microRNA levels. A comparison of the normalized shRNA values indicated that, when shRNAs were placed into the ultramiR scaffold, mature small

RNA levels were increased significantly relative to levels observed using the standard miR-30 scaffold (Figure 4A). Notably, the performance of ultramiR and the previously described alternate scaffold, miR-E, were indistinguishable (data not shown).

To provide a more rigorous test of ultramiR performance, we created a variant of the shERWOOD-selected 1U strategy shRNA library and compared its performance to that of the same library in the standard scaffold. Considering the consensus-essential gene set, over half of all shRNAs in the library were depleted significantly (Figure 4B). This substantial improvement (from 42% to 51%, Wilcoxon rank-sum test,  $p < 0.01$ ) was accompanied by a greater degree of mean log-fold



**Figure 4. Validation of an Alternative Mir Scaffold**

(A) Relative abundances of processed guide sequences for two shRNAs (as determined via small RNA cloning and NGS analysis) when cloned into traditional miR30 and ultramiR scaffolds. Values represent the log-fold enrichment of shRNA guides with respect to sequences corresponding to the ten most abundant microRNAs.

(B) Distributions of the percentage of shERWOOD-1U-selected shRNAs targeting consensus-essential genes that were depleted in validation screens when shRNAs were placed into miR30 and ultramiR scaffolds. Log-fold changes for the same constructs are displayed on the left. The plot was made with the Matlab Boxplot function using default parameters. The edges of the box are the 25<sup>th</sup> and 75<sup>th</sup> percentiles. The error bars extend to the values  $q3 + w(q3 - q1)$  and  $q1 - w(q3 - q1)$ , where  $w$  is 1.5 and  $q1$  and  $q3$  are the 25<sup>th</sup> and 75<sup>th</sup> percentiles.

(C) Knockdown efficiencies for shRNAs targeting the mouse genes Mgp, Slpi, and Serpine2. shRNAs assessed were those contained within the TRC collection, those initially designed for the Hannon-Elledge V.3 library, and those designed using the current strategies. The TRC and Hannon-Elledge V.3 shRNAs are housed within each library's lentiviral vectors, whereas the shERWOOD-1U-selected shRNAs are housed within an ultramiR scaffold in a retroviral vector. UltramiR is constitutively expressed from the LTR.

(D) The number of differentially expressed genes (>2-fold change and FDR < 0.05) identified through pairwise comparisons of the cell lines corresponding to Mgp and Slpi knockdown by the shERWOOD-1U-selected shRNAs and the TRC shRNAs 88943 and 66708.

depletion for each construct (from  $-0.95$  to  $-1.05$ , Wilcoxon rank-sum test,  $p < 0.01$ ).

We also tested a limited number of individual shRNAs for their potency by measuring reductions in target mRNA levels. We selected the four shRNAs with the highest shERWOOD scores for mouse Mgp, Serpine2, and Slpi. These were cloned into an MSCV-based ultramiR vector in which hygromycin resistance and mCherry were also expressed as a bicistronic transcript from the PGK promoter. We also chose to compare these shRNAs to those developed using previous library construction strategies. For this, we obtained the current TRC (five shRNAs per gene) and V.3 Hannon-Elledge (six shRNAs per gene) library constructs targeting these genes. For the Hannon-Elledge library, because there were not four precloned shRNAs for each gene, we assembled the remaining shRNAs that were designed as part of that library but never constructed. We failed to clone two constructs (both targeting Slpi) after multiple attempts, meaning that only four V3 constructs were tested for that

gene. Mouse 4T1 cells were infected at single copy, and knockdown was tested following selection of infected cells.

The TRC library is carried within a vector lacking a fluorescent marker. We therefore calibrated infection levels to achieve single copy by comparison with parallel infections and selections with V3 constructs. The knockdown efficiency of each shRNA was assessed by comparing transcript levels (via quantitative PCR) to those in cells infected with corresponding empty vectors. The TRC shRNAs showed modest knockdown in most cases, with only two shRNAs showing more than 80% of transcript reduction (88943 and 66708, Figure 4C). The Hannon-Elledge V.3 shRNAs produced relatively modest levels of knockdown. In comparison the majority of shRNAs designed using the strategies outlined here reduced target mRNA levels by over 80%, with most reducing target mRNA levels by more than 90% (Figure 4D). Considered together, our data indicate that the combined use of shERWOOD and the ultramiR scaffold consistently produces highly potent shRNAs.

To assess the specificity of shRNA knockdown, we performed RNA sequencing (RNA-seq) on all cell lines expressing

shERWOOD-ultramiR shRNAs targeting *Slpi* and *Mgp* and the two cell lines harboring TRC constructs 88943 and 66708, which target *Mgp* and *Slpi*, respectively. Even in the absence of off-target effects, the silencing of a gene through RNAi will likely elicit biological effects that result in changes in the abundance of other mRNAs. Unlike so-called “off-target” effects, phenotypic effects that emanate from on-target silencing should be consistent for all efficacious shRNAs. Therefore, by comparing the expression profiles of cells harboring different shRNAs corresponding to a single gene, one should be able to infer the scope of off-target effects for each construct. The shRNAs that show the greatest propensity to off-targets will be those that create expression profiles most dissimilar to the mean profile.

When either *Mgp* or *Slpi* were silenced using the strategies outlined here, the expression profiles in the resultant lines were found to be highly similar. Less than 25 genes were altered in their expression (DESeq, fold change > 2 and FDR < 0.05) between any pair of corresponding lines. However, when these were compared with lines that had *Mgp* or *Slpi* silenced using potent TRC constructs, a significant difference in expression profiles was observed. Over 500 genes are altered in the line where *Mgp* has been silenced using the TRC constructs, and approximately 250 are altered in the line expressing the TRC *Slpi*-shRNA (Figure 4D).

These results could reflect our current strategies for reducing off-targeting or our use of a microRNA-based scaffold. Recently, others have observed strong phenotypic changes, related to microRNA dysregulation, when U6 driven stem-loop shRNAs were expressed in cells where the target gene had been deleted (Baek et al., 2014). In contrast, when these same shRNAs were expressed from a microRNA scaffold, the phenotype was not observed. Overall, the aforementioned analysis indicates that shRNAs produced using the strategies outlined in this report, when expressed in an ultramiR scaffold, show strong knock-down capacity and limited off-target effects.

## DISCUSSION

The application of RNAi in mammalian cells promised a revolution in understanding gene function and in the discovery and validation of therapeutic targets. Although the impact of RNAi has been enormous, there has also been substantial frustration in attempts to fully realize the potential of this technology. Many different sequences often need to be tested to obtain one that potently suppresses expression, a problem that is particularly acute with shRNAs expressed from single-copy transgenes. This, and the resulting variability in the quality of publicly available genome-wide shRNA collections, has caused consternation, particularly when very similar shRNA screens carried out by different investigators yield largely nonoverlapping results (Babij et al., 2011; Luo et al., 2009; Scholl et al., 2009). We tried to address problems with current shRNA technologies by optimizing target sequence choice and small RNA production.

We leveraged our prior development of a high-throughput assay for testing shRNA potency to develop a computational algorithm capable of accurately predicting the outcome of the sensor screen and, in turn, predicting potentially potent shRNAs. Through iterative cycles of training and refinement we produced

a tool that permits highly efficacious shRNAs to be generated for nearly any gene.

We validated the performance of our approach and benchmarked it against current tools using nonsequence verified, focused shRNA libraries. Based on our analyses, we can now generate shRNA libraries where nearly 60% of all hairpins targeting essential genes are strongly depleted in multiplexed screens. This means that, for any library containing, on average, four hairpins per gene, most bona fide hits will be identified by multiple hairpins, greatly reducing the probability of false-positive calls. Because our libraries were used in their raw forms, we feel that this is a lower boundary of performance because sequence-validated and arrayed collections will not contain a mixture of shRNA variants generated by synthesis and PCR errors.

Given the promise of our approach, we have undertaken the construction of fourth- and fifth-generation sequence-verified shRNA libraries targeting the mouse and human genomes. The fourth generation toolkit takes advantage of shERWOOD in a canonical miR-30 scaffold and currently comprises over 75,000 shRNAs targeting human genes and 40,000 shRNAs targeting mouse genes. The fifth-generation toolkit places shERWOOD shRNAs in the ultramiR scaffold and is presently ~50% complete.

We have predicted shERWOOD shRNAs targeting constitutive exons of annotated human, mouse, and rat protein coding genes, and these are available via a web portal (<http://sherwood.cshl.edu:8080/sherwood/>). We have additionally made shERWOOD available as a web-based tool for custom shRNA prediction, for example for the design of shRNAs for other model organisms or for specific mRNA isoforms or non-coding RNAs.

Overall, we feel that the combination of improvements to shRNA technologies described herein creates a next-generation RNAi toolkit that will produce more reliable outcomes for investigators, whether applied on a gene-by-gene basis or in the context of unbiased, genome-wide screens.

## EXPERIMENTAL PROCEDURES

### Cell Lines

The sensor algorithm was performed using Eco-rtTA-chicken (ERC) cells (derived from DF-1 chicken embryonic fibroblasts) (Fellmann et al., 2011). All shRNA screens were performed in the pancreatic adenocarcinoma cell line A385 (Cui et al., 2012). Small RNA analysis for RPA2 shRNAs was performed in the ERC cell line (Fellmann et al., 2011) and in HEK293T cells for the *Renilla* shRNAs. Individual shRNA knockdown experiments were performed in the 4T1 murine mammary cancer cell line (Dexter et al., 1978).

### Vectors

All RNAi screens and small RNA cloning experiments were performed with an MSCV-based retroviral vector harboring a bicistronic transcript (eGFP-IRES-Neomycin) downstream of the PGK promoter (Figure S2D). Single-target knockdown experiments for shERWOOD-ultramiR shRNAs were performed with a similar vector, where Neomycin is replaced with Hygromycin, and enhanced GFP is replaced with mCHERRY. Single-target knockdown experiments for the Hannon-Elledge V3 and TRC shRNAs were performed with the GIPZ and pLKO.1 vectors, respectively (GE Dharmacon).

### shRNA Library Construction

To ensure high-complexity end products, all shRNA libraries were amplified from raw chip material using 16 separate reactions with 22 PCR cycles. For

each reaction, 1  $\mu$ l of 100  $\mu$ M chip material was used. All transformations were performed with Invitrogen's MegaX DH10B T1 electrocompetent cells using a Bio-Rad Gene Pulser Xcell and Bio-Rad Gene Pulser 1 mm cuvettes for electroporation. For each library, a minimum of 25 M successfully transformed cells were obtained.

#### shRNA Library Screening

shRNA libraries were packaged using the Platinum-A retrovirus packaging cell (Cell Biolabs). Cells were cotransfected with glycoprotein G of the vesicular stomatitis virus and siRNAs targeting the shRNA processing protein Pasha (QIAGEN). Viral infections were performed at an MOI of 0.3 to ensure a maximum of one shRNA infection per cell. shRNA representation in the infected cell population was maintained at a minimum of 1,000 infected cells per shRNA on each passage. All screens were performed in triplicate. Two days after infection, cells were collected for a reference time point, and, after  $\sim$ 12 doublings, cells were again harvested for a final time point. Neomycin selection began after the initial time point and continued throughout the screens.

#### shRNA Library Processing and Analysis

Following cell harvests, DNA was extracted with the QIAGEN QIAamp DNA Blood Maxi kit. For each sample, shRNA molecules were extracted from genomic DNA in 96 separate 25-cycle PCR reactions where 2  $\mu$ g of input DNA was included in each reaction. Following this initial PCR, Illumina adapters were added via PCR, and samples were processed on the Illumina Hi-Seq-2.0 platform (read depth was maintained at  $\sim$ 1,000 short reads per shRNA). Following sequencing, shRNA counts were extracted with the bowtie algorithm (allowing zero mismatches) and normalized by their total counts. Log-fold changes demonstrated a GC bias in the control shRNA population (Figure S2E). To remove this bias, a 1<sup>o</sup> polynomial was fit to each screen replicate's log-fold change versus GC content data, and this curve was then subtracted from each data point (Figure S2F). Following this, values were further normalized so that the control population had a population variance of one. shRNAs were classified as depleted with an FDR cutoff of 0.1 using an empirical Bayes moderated test (Figure S2G; Smyth, 2004).

For further details, see the [Supplemental Experimental Procedures](#).

#### ACCESSION NUMBERS

All raw and processed data are available through the National Center for Biotechnology Information under the accession number GSE62189.

#### SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures and four figures and can be found with this article online at <http://dx.doi.org/10.1016/j.molcel.2014.10.025>.

#### AUTHOR CONTRIBUTIONS

S.R.V.K. and G.J.H. designed the experiments and wrote the manuscript. S.R.V.K. designed the algorithm. A.M. performed all shRNA screens. A.M. and X.Z. performed the 1U sensor experiments. N.E. performed all small RNA cloning. S.R.V.K., A.M., and N.E. constructed the sequence-verified libraries. K.C. and K.M. performed the DSIR-sensor experiments. S.R.V.K. and A.G. developed and implemented the exon inclusion and off-target minimization strategies. A.G. and O.E.D. designed the shERWOOD website. E.W. and S.K. performed the individual knockdown experiments. S.R.V.K. and S.K. performed the RNA-seq experiments.

#### ACKNOWLEDGMENTS

This work was supported by the Howard Hughes Medical Institute, by grants from the NIH (to G.J.H.), and by a gift from Kathryn W. Davis. S.R.V.K. is supported by a fellowship from The Hope Funds for Cancer Research. E.W. is supported by a fellowship from the Boehringer Ingelheim Foundation. We thank all members of the Hannon laboratory for help with initial library construction. We

would also like to thank Blake Simmons (Transomics Technologies) for help with constructing sequence-verified versions of the libraries. Finally, we thank Vasily Vagin for helpful comments on this manuscript. S.R.V.K. and G.H. are associated with Transomic Technologies, who have commercialized libraries constructed using the shERWOOD and ultramiR design strategies.

Received: April 3, 2014

Revised: September 11, 2014

Accepted: October 23, 2014

Published: November 26, 2014

#### REFERENCES

- Ameres, S.L., and Zamore, P.D. (2013). Diversifying microRNA sequence and function. *Nat. Rev. Mol. Cell Biol.* **14**, 475–488.
- Auyeung, V.C., Ulitsky, I., McGeary, S.E., and Bartel, D.P. (2013). Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. *Cell* **152**, 844–858.
- Babij, C., Zhang, Y., Kurzeja, R.J., Munzli, A., Shehabeldin, A., Fernando, M., Quon, K., Kassner, P.D., Ruefli-Brasse, A.A., Watson, V.J., et al. (2011). STK33 kinase activity is nonessential in KRAS-dependent cancer cells. *Cancer Res.* **71**, 5818–5826.
- Baek, S.T., Kerjan, G., Bielas, S.L., Lee, J.E., Fenstermaker, A.G., Novarino, G., and Gleeson, J.G. (2014). Off-target effect of doublecortin family shRNA on neuronal migration associated with endogenous microRNA dysregulation. *Neuron* **82**, 1255–1262.
- Berns, K., Hijmans, E.M., Mullenders, J., Brummelkamp, T.R., Velds, A., Heimerikx, M., Kerkhoven, R.M., Madiredjo, M., Nijkamp, W., Weigelt, B., et al. (2004). A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428**, 431–437.
- Bernstein, E., Caudy, A.A., Hammond, S.M., and Hannon, G.J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**, 363–366.
- Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5–32.
- Brummelkamp, T.R., Bernards, R., and Agami, R. (2002). A system for stable expression of short interfering RNAs in mammalian cells. *Science* **296**, 550–553.
- Chen, C.Z., Li, L., Lodish, H.F., and Bartel, D.P. (2004). MicroRNAs modulate hematopoietic lineage differentiation. *Science* **303**, 83–86.
- Chiu, Y.L., and Rana, T.M. (2002). RNAi in human cells: basic structural and functional features of small interfering RNA. *Mol. Cell* **10**, 549–561.
- Chuang, C.F., and Meyerowitz, E.M. (2000). Specific and heritable genetic interference by double-stranded RNA in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **97**, 4985–4990.
- Cleary, M.A., Kilian, K., Wang, Y., Bradshaw, J., Cavet, G., Ge, W., Kulkarni, A., Paddison, P.J., Chang, K., Sheth, N., et al. (2004). Production of complex nucleic acid libraries using highly parallel in situ oligonucleotide synthesis. *Nat. Methods* **1**, 241–248.
- Cui, Y., Brosnan, J.A., Blackford, A.L., Sur, S., Hruban, R.H., Kinzler, K.W., Vogelstein, B., Maitra, A., Diaz, L.A., Jr., Iacobuzio-Donahue, C.A., et al. (2012). Genetically defined subsets of human pancreatic cancer show unique in vitro chemosensitivity. *Clinical cancer research* **18**, 6519–6530.
- Cullen, B.R. (2006). Induction of stable RNA interference in mammalian cells. *Gene Ther.* **13**, 503–508.
- Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F., and Hannon, G.J. (2004). Processing of primary microRNAs by the Microprocessor complex. *Nature* **432**, 231–235.
- Dexter, D.L., Kowalski, H.M., Blazar, B.A., Fligiel, Z., Vogel, R., and Heppner, G.H. (1978). Heterogeneity of tumor cells from a single mouse mammary tumor. *Cancer Res.* **38**, 3174–3181.
- Elbashir, S.M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494–498.

- Elkayam, E., Kuhn, C.D., Tocilj, A., Haase, A.D., Greene, E.M., Hannon, G.J., and Joshua-Tor, L. (2012). The structure of human argonaute-2 in complex with miR-20a. *Cell* **150**, 100–110.
- Fellmann, C., Zuber, J., McGjunkin, K., Chang, K., Malone, C.D., Dickins, R.A., Xu, Q., Hengartner, M.O., Elledge, S.J., Hannon, G.J., and Lowe, S.W. (2011). Functional identification of optimized RNAi triggers using a massively parallel sensor assay. *Mol. Cell* **41**, 733–746.
- Fellmann, C., Hoffmann, T., Sridhar, V., Hopfgartner, B., Muhar, M., Roth, M., Lai, D.Y., Barbosa, I.A., Kwon, J.S., Guan, Y., et al. (2013). An optimized microRNA backbone for effective single-copy RNAi. *Cell Reports* **5**, 1704–1713.
- Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811.
- Frank, F., Sonenberg, N., and Nagar, B. (2010). Structural basis for 5'-nucleotide base-specific recognition of guide RNA by human AGO2. *Nature* **465**, 818–822.
- Grishok, A., Pasquinelli, A.E., Conte, D., Li, N., Parrish, S., Ha, I., Bailly, D.L., Fire, A., Ruvkun, G., and Mello, C.C. (2001). Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell* **106**, 23–34.
- Gupta, S., Schoer, R.A., Egan, J.E., Hannon, G.J., and Mittal, V. (2004). Inducible, reversible, and stable RNA interference in mammalian cells. *Proc. Natl. Acad. Sci. USA* **101**, 1927–1932.
- Hammond, S.M., Boettcher, S., Caudy, A.A., Kobayashi, R., and Hannon, G.J. (2001). Argonaute2, a link between genetic and biochemical analyses of RNAi. *Science* **293**, 1146–1150.
- Han, J., Lee, Y., Yeom, K.H., Nam, J.W., Heo, I., Rhee, J.K., Sohn, S.Y., Cho, Y., Zhang, B.T., and Kim, V.N. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* **125**, 887–901.
- Hannon, G.J. (2002). RNA interference. *Nature* **418**, 244–251.
- Huesken, D., Lange, J., Mickanin, C., Weiler, J., Asselbergs, F., Warner, J., Meloon, B., Engel, S., Rosenberg, A., Cohen, D., et al. (2005). Design of a genome-wide siRNA library using an artificial neural network. *Nat. Biotechnol.* **23**, 995–1001.
- Hutvagner, G., and Zamore, P.D. (2002). A microRNA in a multiple-turnover RNAi enzyme complex. *Science* **297**, 2056–2060.
- Hutvagner, G., McLachlan, J., Pasquinelli, A.E., Bálint, E., Tuschl, T., and Zamore, P.D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* **293**, 834–838.
- Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., et al. (2003). Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237.
- Kambris, Z., Brun, S., Jang, I.H., Nam, H.J., Romeo, Y., Takahashi, K., Lee, W.J., Ueda, R., and Lemaitre, B. (2006). *Drosophila* immunity: a large-scale in vivo RNAi screen identifies five serine proteases required for Toll activation. *Current biology: CB* **16**, 808–813.
- Ketting, R.F., Fischer, S.E., Bernstein, E., Sijen, T., Hannon, G.J., and Plasterk, R.H. (2001). Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. *Genes Dev.* **15**, 2654–2659.
- Khvorova, A., Reynolds, A., and Jayasena, S.D. (2003). Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115**, 209–216.
- Lai, E.C. (2002). Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat. Genet.* **30**, 363–364.
- Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S., and Kim, V.N. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**, 415–419.
- Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20.
- Lund, E., Güttinger, S., Calado, A., Dahlberg, J.E., and Kutay, U. (2004). Nuclear export of microRNA precursors. *Science* **303**, 95–98.
- Luo, J., Emanuele, M.J., Li, D., Creighton, C.J., Schlabach, M.R., Westbrook, T.F., Wong, K.K., and Elledge, S.J. (2009). A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. *Cell* **137**, 835–848.
- Malone, C., Brennecke, J., Czech, B., Aravin, A., and Hannon, G.J. (2012). Preparation of small RNA libraries for high-throughput sequencing. *Cold Spring Harbor protocols* **2012**, 1067–1077.
- Martinez, J., Patkaniowska, A., Urlaub, H., Lührmann, R., and Tuschl, T. (2002). Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell* **110**, 563–574.
- Matveeva, O.V., Nazipova, N.N., Ogurtsov, A.Y., and Shabalina, S.A. (2012). Optimized models for design of efficient miR30-based shRNAs. *Front. Genet.* **3**, 163.
- Nakanishi, K., Weinberg, D.E., Bartel, D.P., and Patel, D.J. (2012). Structure of yeast Argonaute with guide RNA. *Nature* **486**, 368–374.
- Paddison, P.J., Caudy, A.A., Bernstein, E., Hannon, G.J., and Conklin, D.S. (2002). Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. *Genes Dev.* **16**, 948–958.
- Paddison, P.J., Silva, J.M., Conklin, D.S., Schlabach, M., Li, M., Aruleba, S., Balija, V., O'Shaughnessy, A., Gnoj, L., Scobie, K., et al. (2004). A resource for large-scale RNA-interference-based screens in mammals. *Nature* **428**, 427–431.
- Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W.S., and Khvorova, A. (2004). Rational siRNA design for RNA interference. *Nat. Biotechnol.* **22**, 326–330.
- Sánchez Alvarado, A., and Newmark, P.A. (1999). Double-stranded RNA specifically disrupts gene expression during planarian regeneration. *Proc. Natl. Acad. Sci. USA* **96**, 5049–5054.
- Scholl, C., Fröhling, S., Dunn, I.F., Schinzel, A.C., Barbie, D.A., Kim, S.Y., Silver, S.J., Tamayo, P., Wadlow, R.C., Ramaswamy, S., et al. (2009). Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells. *Cell* **137**, 821–834.
- Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115**, 199–208.
- Seitz, H., and Zamore, P.D. (2006). Rethinking the microprocessor. *Cell* **125**, 827–829.
- Seitz, H., Ghildiyal, M., and Zamore, P.D. (2008). Argonaute loading improves the 5' precision of both MicroRNAs and their miRNA\* strands in flies. *Current biology: CB* **18**, 147–151.
- Silva, J.M., Li, M.Z., Chang, K., Ge, W., Golding, M.C., Ricles, R.J., Siolas, D., Hu, G., Paddison, P.J., Schlabach, M.R., et al. (2005). Second-generation shRNA libraries covering the mouse and human genomes. *Nat. Genet.* **37**, 1281–1288.
- Sims, D., Mendes-Pereira, A.M., Frankum, J., Burgess, D., Cerone, M.A., Lombardelli, C., Mitsopoulos, C., Hakas, J., Murugaesu, N., Isacke, C.M., et al. (2011). High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing. *Genome Biol.* **12**, R104.
- Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* **3**, Article3.
- Svoboda, P., Stein, P., Hayashi, H., and Schultz, R.M. (2000). Selective reduction of dormant maternal mRNAs in mouse oocytes by RNA interference. *Development* **127**, 4147–4156.
- Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. A Stat. Soc.* **58**, 267–288.
- Timmons, L., and Fire, A. (1998). Specific interference by ingested dsRNA. *Nature* **395**, 854.

- Tuschl, T., Zamore, P.D., Lehmann, R., Bartel, D.P., and Sharp, P.A. (1999). Targeted mRNA degradation by double-stranded RNA in vitro. *Genes Dev.* *13*, 3191–3197.
- Ui-Tei, K., Naito, Y., Takahashi, F., Haraguchi, T., Ohki-Hamazaki, H., Juni, A., Ueda, R., and Saigo, K. (2004). Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. *Nucleic Acids Res.* *32*, 936–948.
- Vacic, V., Iakoucheva, L.M., and Radivojac, P. (2006). Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* *22*, 1536–1537.
- Vert, J.P., Foveau, N., Lajaunie, C., and Vandenbrouck, Y. (2006). An accurate and interpretable model for siRNA efficacy prediction. *BMC Bioinformatics* *7*, 520.
- Wang, Y., Sheng, G., Juranek, S., Tuschl, T., and Patel, D.J. (2008). Structure of the guide-strand-containing argonaute silencing complex. *Nature* *456*, 209–213.
- Yi, R., Qin, Y., Macara, I.G., and Cullen, B.R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev.* *17*, 3011–3016.
- Yuan, Y.R., Pei, Y., Chen, H.Y., Tuschl, T., and Patel, D.J. (2006). A potential protein-RNA recognition event along the RISC-loading pathway from the structure of *A. aeolicus* Argonaute with externally bound siRNA. *Structure* (London, England: 1993) *14*, 1557–1565.
- Zender, L., Xue, W., Zuber, J., Semighini, C.P., Krasnitz, A., Ma, B., Zender, P., Kubicka, S., Luk, J.M., Schirmacher, P., et al. (2008). An oncogenomics-based in vivo RNAi screen identifies tumor suppressors in liver cancer. *Cell* *135*, 852–864.
- Zeng, Y., and Cullen, B.R. (2003). Sequence requirements for micro RNA processing and function in human cells. *RNA* *9*, 112–123.
- Zhang, X., and Zeng, Y. (2010). The terminal loop region controls microRNA processing by Drosha and Dicer. *Nucleic Acids Res.* *38*, 7689–7697.

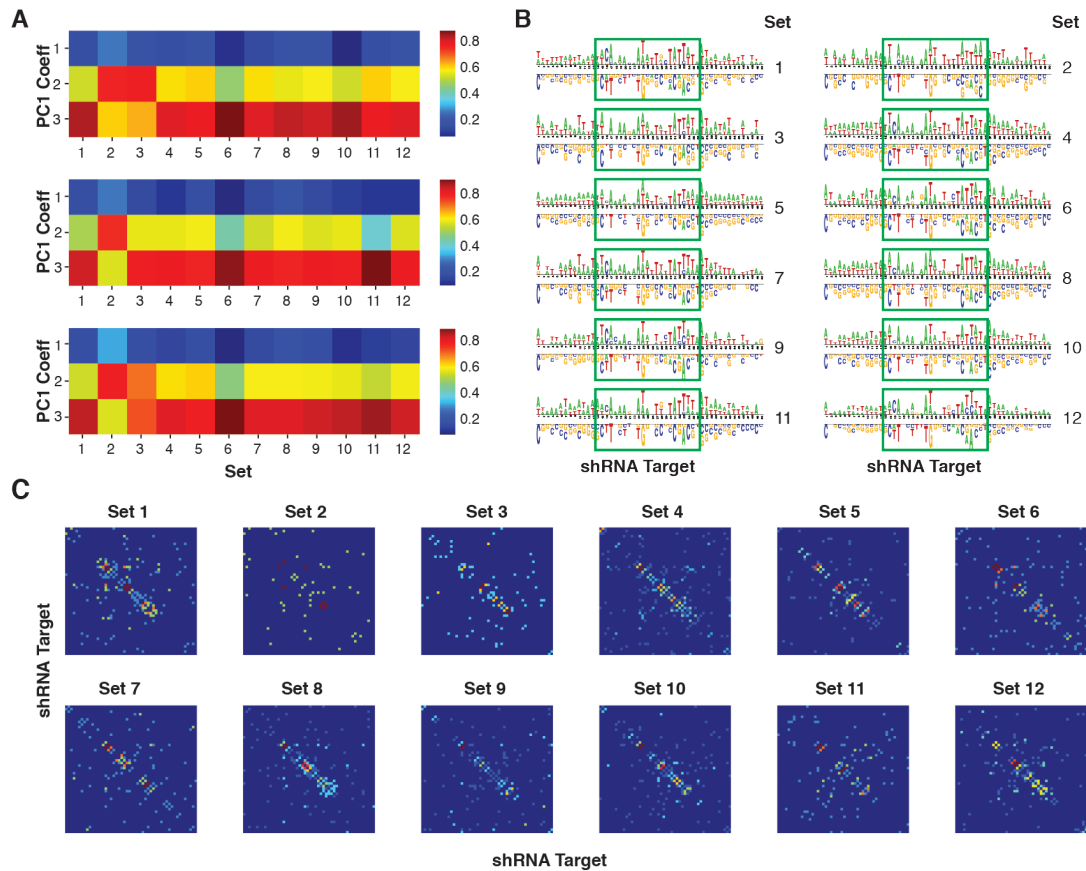
**Molecular Cell, Volume 56**

**Supplemental Information**

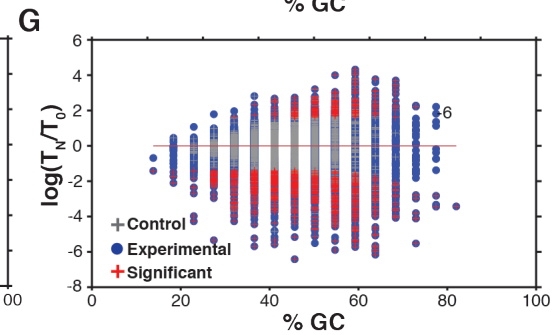
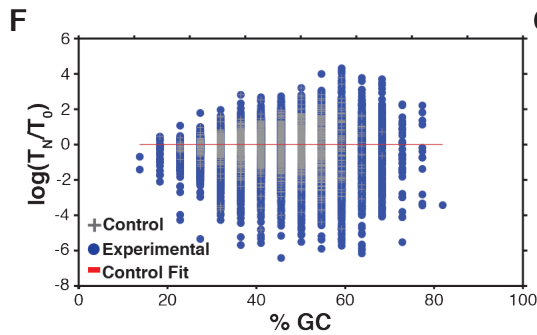
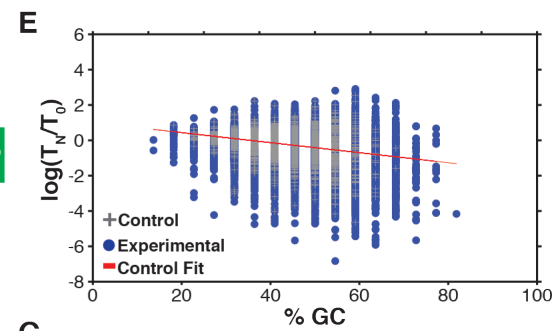
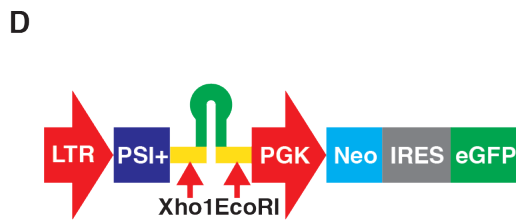
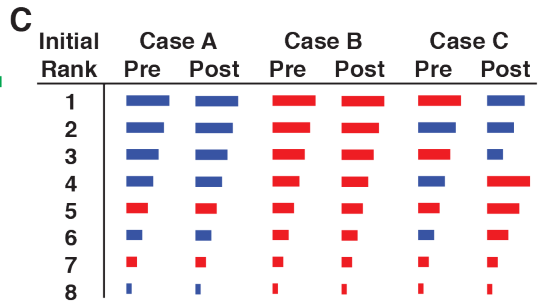
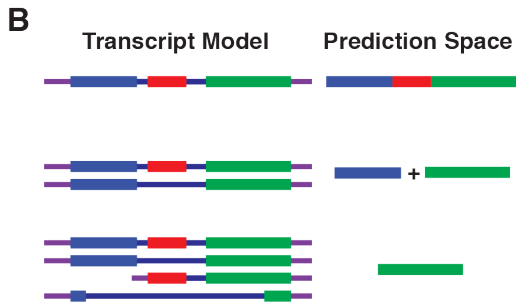
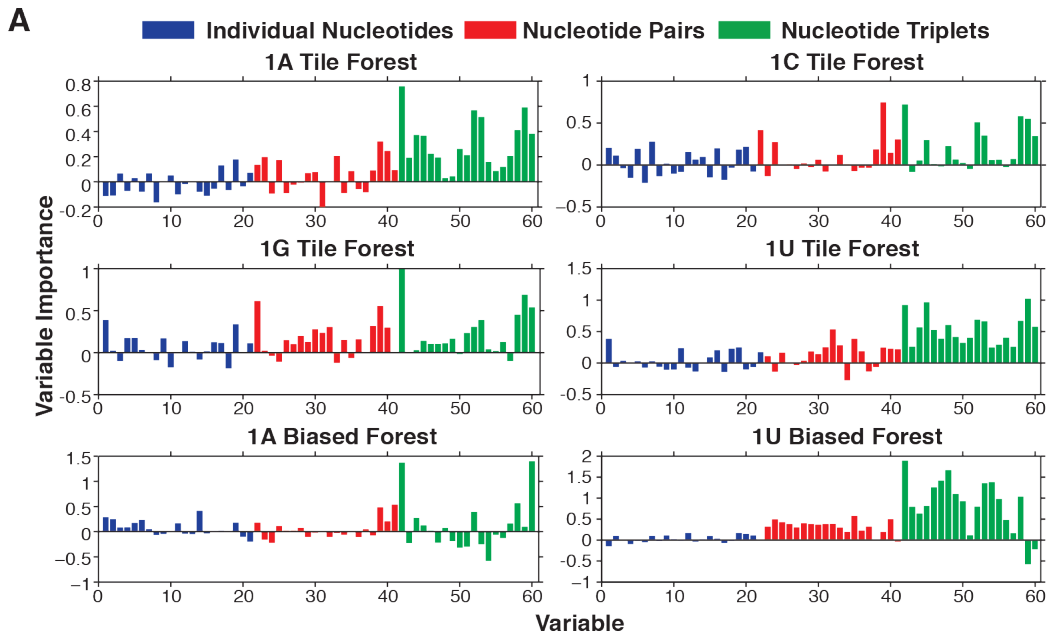
**A Computational Algorithm to Predict shRNA Potency**

Simon R.V. Knott, Ashley R. Maceli, Nicolas Erard, Kenneth Chang, Krista Marran, Xin Zhou, Assaf Gordon, Osama El Demerdash, Elvin Wagenblast, Sun Kim, Christof Fellmann, and Gregory J. Hannon

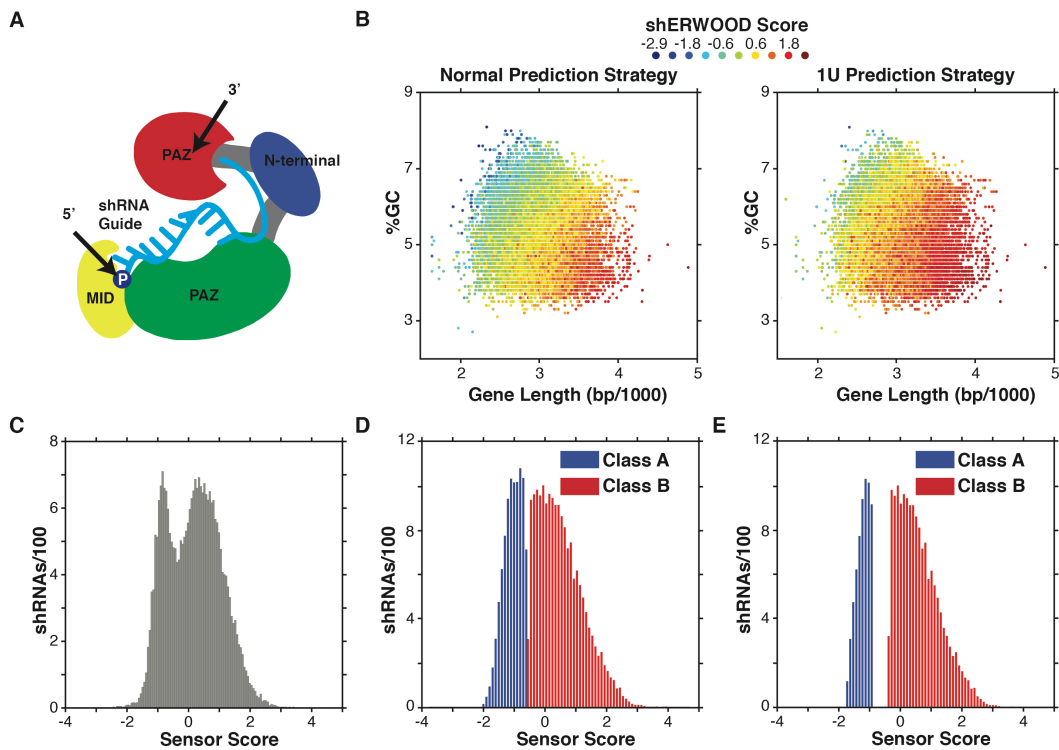
## Supplemental Figures



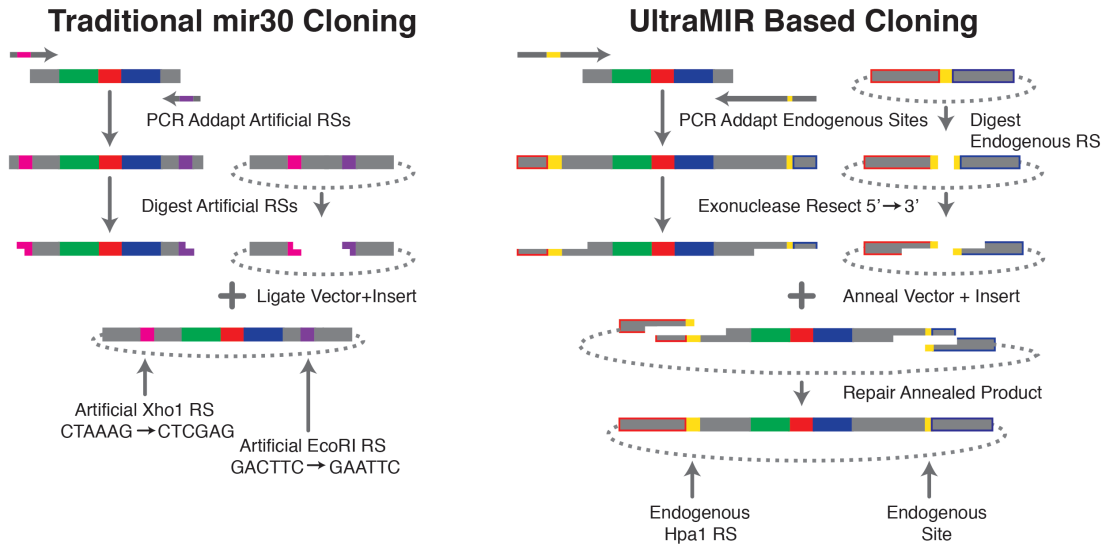
**Figure S1, related to Figure 1: A)** Heatmap representation of the coefficients used to extract the first principal components of the matrices described in Figure 1A. Coefficients 1, 2 and 3 represent the contribution that each of on-dox sorts 1, 2 and 3 made in defining the first principal component 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.



**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 22<sup>nd</sup> 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 shRNAs are non-multimappers, so no shuffling occurs. In case B, all rank 1-8 shRNAs are multimappers (indicating 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 < 0.10).



**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 represent 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 the lengths of these areas 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 a 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). Illumina 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 instructions 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 2000 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.