Prediction of potent shRNAs with a sequential classification algorithm

Raphael Pelossof1,12, Lauren Fairchild1,2,12, Chun-Hao Huang3,4, Christian Widmer1,5, Vipin T Sreedharan1, Nishi Sinha6, Dan-Yu Lai6, Yuanzhe Guan6, Prem K Prem sirut6, Darjus F Tschaharganeh3, Thomas Hoffmann7, Vishal Thapar3, Qing Xiang6, Ralph J Garippa8, Gunnar Rätsch1,9, Johannes Zuber7, Scott W Lowe3,4,10, Christina S Leslie1 & Christof Fellmann6,11

We present SplashRNA, a sequential classifier to predict potent microRNA-based short hairpin RNAs (shRNAs). Trained on published and novel data sets, SplashRNA outperforms previous algorithms and reliably predicts the most efficient shRNAs for a given gene. Combined with an optimized miR-E backbone, >90% of high-scoring SplashRNA predictions trigger >85% protein knockdown when expressed from a single genomic integration. SplashRNA can significantly improve the accuracy of loss-of-function genetics studies and facilitates the generation of compact shRNA libraries.

Experimental RNA interference (RNAi) acts by providing exogenous sources of double-stranded RNA that mimic endogenous triggers and enable reversible, transcript-specific gene knockdown1. Whereas short interfering RNAs (siRNAs) allow for rapid gene knockdown, they are not suitable for many long-term and in vivo studies due to their transient nature. Stem-loop shRNAs can be used as a continuous source of RNAi triggers when expressed from suitable vectors, but suffer from various technical limitations including inaccurate processing2 and off-target effects through saturation of the endogenous microRNA machinery3–5. State-of-the-art microRNA-based shRNA vectors can overcome these limitations by providing a natural substrate of the RNAi pathway that is accurately and efficiently processed6–9, resulting in minimal or no off-target effects when expressed from a single genomic integration (single-copy)10,11. Still, our limited understanding of RNAi processing requirements and the lack of robust algorithms for the design of microRNA-based shRNAs with high potency and low off-target activity has hampered the utility of RNAi tools.

To understand the sequence requirements of potent RNAi and identify efficient microRNA-based shRNAs for any gene, we previously developed a functional high-throughput “Sensor” assay that enables biological assessment of tens of thousands of shRNAs in parallel (Supplementary Fig. 1a)10. We used this assay to generate focused and genome-wide shRNA libraries11,12. Furthermore, to increase the potency of all shRNAs, especially when expressed at single-copy, we established miR-E7, an optimized microRNA backbone that boosts processing efficiency7,13 and leads to stronger target knockdown when compared to standard miR-30 designs3.

To build an accurate miR-E shRNA predictor, we developed SplashRNA, a sequential learning algorithm combining two support vector machine (SVM) classifiers trained on judiciously integrated data sets (Supplementary Table 1). SplashRNA models the sequential advances in shRNA technology to enable efficient learning on unbiased and biased data (Fig. 1a,b). To train the algorithm, we generated a large-scale miR-30 data set (referred to as M1; Supplementary Fig. 1b–f) and a miR-E data set (referred to as miR-E; Supplementary Fig. 1g) using our RNAi Sensor and reporter assays, respectively (Supplementary Table 2)7,10. We also used the previously published TILE10 and UltramiR12 sets. TILE is unbiased as it was generated by complete tiling of nine genes. By contrast, M1, miR-E and UltramiR are based on preselected input libraries showing biased coverage of potent shRNAs (Supplementary Fig. 1h). Yet, together these data sets comprehensively sample the distributions of features of non-functional and functional shRNAs. Effective integration of all sets is thus crucial for efficient miR-E shRNA prediction.

Combining diverse data sets presents a machine-learning challenge. Our approach of using a sequential classifier stems from classification strategies used in face detection14–15, where a first classifier evaluates simple face-like features to reject obvious non-faces and a second classifier evaluates refined features on retained potential faces. Similarly, SplashRNA contains a sequence of two SVM classifiers trained on miR-30 and miR-E data. The miR-30 classifier evaluates shRNA sequence features to reject obvious non-functional shRNAs, whereas the miR-E classifier evaluates refined sequence features for retained, potentially potent shRNAs (Fig. 1b and Supplementary Fig. 2a). Each classifier is composed of a combination of k-mer feature representations16,17. To capture AU content and position-specific k-mer features10, we represented an shRNA as a sum of a spectrum kernel on sequence positions 1–15, a spectrum kernel on sequence positions 16–22 and a weighted degree kernel on the entire sequence (Supplementary Fig. 2b). We found that this kernel combination yields the best performance (Supplementary Fig. 2c,d).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.
outperformed the miR-30-based shERWOOD algorithm on the UltramiR set (Supplementary Fig. 4e), compared to its published maximum performance\textsuperscript{12}. Additionally, SplashRNA consistently showed the highest predictive performance on independent data sets when benchmarked against DSIR and two other shRNA prediction tools, sequence score\textsuperscript{19} (seqScore) and miR_Scan\textsuperscript{20}.

We also observed the high performance of SplashRNA in two large-scale biological RNAi screens\textsuperscript{19,21}, run with shRNAs functionally equivalent to miR-E (Supplementary Fig. 4g)\textsuperscript{22}, which tested ~25 preselected shRNAs per gene. In both cases, SplashRNA was able to retrospectively predict which shRNAs were potent and thus were enriched or depleted in the positive or negative selection screen, respectively. SplashRNA achieved the most significant difference in potency between its top five and bottom five predictions per targeted gene and was the only algorithm to reach significance in both screens (P < 0.01, one-sided Wilcoxon rank sum test). Top SplashRNA predictions also showed equally good or better accuracy compared to larger sets of preselected shRNAs when tested on a subset of the negative-selection screen that included only a previously established set of ‘gold-standard’ essential genes\textsuperscript{11,22}. The top ten SplashRNA predictions identified true positives significantly better than the bottom ten (P < 0.001, empirical permutation test), minimizing off-target hit identification (Fig. 2c).

Robust shRNA prediction starts with the selection of the right transcript region. Analyses of unbiased TILDE data showed that efficient shRNAs are more prevalent in 3′ UTRs compared to coding sequences and 5′ UTRs (Supplementary Fig. 5a), likely due to the shared high AU content (Supplementary Fig. 5b–d)\textsuperscript{28}. Whereas 3′ UTRs often present ample design space because of their lengths, when validating top predictions in mouse fibroblasts, many shRNAs targeting the distal end of \textit{Pten} resulted in minimal or no protein knockdown (Supplementary Fig. 5e and Supplementary Table 2). Inspection of the \textit{Pten} mRNA (NCBI, NM_000895) revealed that all these shRNAs target regions past alternative cleavage and polyadenylation (ApA) signals, which lead to shorter transcript variants\textsuperscript{22} lacking the respective target sites (Supplementary Fig. 5f). Hence, to eliminate ApA as a source of non-functional shRNAs, we used ApA atlases\textsuperscript{25,26} to annotate the human and mouse reference transcriptomes (NCBI) and discard 3′ UTR portions that may be absent due to ApA. Similarly, we report predictions only on the intersection of all transcript variants for each gene and filter multi-matching sequences.

Testing an extensive set of individual \textit{de novo} predictions targeting \textit{Pten}, \textit{Bap1}, \textit{Pbrm1}, \textit{Rela}, \textit{Bcl211}, \textit{Axin1}, \textit{NF2} and \textit{Cd9} (Supplementary Table 2) under single-copy conditions\textsuperscript{2} by conventional western blot analysis (Fig. 2d,e and Supplementary Fig. 6a–f) or flow-cytometry-based immunofluorescence of surface proteins (Supplementary Fig. 6g), we found that protein knockdown levels were very high: 91% of predictions (41/45) with a SplashRNA score of >1 showed >85% protein knockdown (Supplementary Fig. 6h). Even in the case of human \textit{NF2}, a gene with nine annotated transcript variants that share only 198 nucleotides (excluding the 5′ UTR, Supplementary Fig. 6i), the top eight SplashRNA predictions triggered 77–96% (median 89%) protein suppression under single-copy conditions (Supplementary Fig. 6j). Additionally, \textit{Cd9} knockdown analyses in mouse fibroblasts showed that SplashRNA clearly outperforms DSIR in \textit{de novo} prediction and achieves near knockout levels comparable to CRISPR–Cas9 (Supplementary Fig. 6g). Potent microRNA-based shRNAs have an equally low chance of off-target effects as non-functional sequences when expressed at single-copy\textsuperscript{11}.

Extrapolating beyond the tested shRNAs, we calculated the proportion of genes for which SplashRNA would find at least five shRNAs above a given threshold (Fig. 2f). After shortening of transcripts due to ApA and considering only the intersection of all transcript variants
per gene, we found that 87% of mouse genes and 81% of human genes have at least five shRNAs with SplashRNA scores above 1, corresponding to an 80% probability (e.g., four out of five shRNAs) of more than 85% knockdown at single-copy (Supplementary Fig. 6i).

Building on our Sensor assay and the optimized miR-E backbone, here we have established a robust algorithm to predict ultra-potent microRNA-based shRNAs targeting nearly any gene. SplashRNA is able to accurately predict the potency of independently validated and novel shRNAs and outperforms existing algorithms. Our sequential predictor approach facilitates the integration of biased and unbiased data sets and can serve as a blueprint for other prediction problems. An open source implementation of SplashRNA is accessible at http://splashrna.msckcc.org. Mouse and human genome-wide predictions are also provided separately (Supplementary Table 3).

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J.A. Doudna, G.J. Hannon, L.E. Dow and S.N. Floor for continuous support and valuable discussions. We gratefully acknowledge assistance and support from A. Banito, V. Sridhar, L. Faletti, C.C. Chen and S. Tian. C.F. was supported in part by a K99/R00 Pathway to Independence Award (K99GM118909) from the National Institutes of Health (NIH), National Institute of General Medical Sciences (NIGMS). C.F. is a founder of Mirimus Inc., a company that develops RNAi-based reagents and transgenic mice. This work was also supported in part by grant CA113106 (S.W.L.) and CRISPR-P (CA180936). S.W.L. is a founder and member of the scientific advisory board of Mirimus Inc., the Geoffrey Beene Chair of Cancer Biology at MSKCC and an investigator of the Howard Hughes Medical Institute. J.Z. is a member of the scientific advisory board, and P.K.P. is a founder and employee of Mirimus Inc. C.S.L. was supported in part by NIH/NCI U01 grants HG007033 and HG007893 and NCI U01 grant CA164190. A375 cells were a kind gift from Neal Rosen, MSKCC.

AUTHOR CONTRIBUTIONS

R.P., L.F., C.S.L. and C.F. conceived and designed the study, and developed the data integration framework. R.P., L.F., and C.W. built the algorithm, and carried out the model training and computational validation. C.-H.H., N.S.,
D.-Y.L., Y.G., P.K.P., D.F.T., T.H., J.Z., S.W.L. and C.F. generated the biological
data sets and validated knockdown potency. R.P., L.F., C.W. and V.T.S. built the
web page. V.T. and G.R. assisted with study design and advised on algorithmic
development. Q.X. and R.J.G. helped with validation of predictions.
R.P., L.F., C.-H.H., T.H., J.Z., S.W.L., C.S.L. and C.F. analyzed data and
wrote the manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare competing financial interests: details are available in the online
version of the paper.

Reprints and permissions information is available online at http://www.nature.com/
reprints/index.html.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims
in published maps and institutional affiliations.

1. Fellmann, C. & Lowe, S.W. Nat. Cell Biol. 16, 10–18 (2014).
2. Guda, S. et al. Mol. Ther. 23, 1465–1474 (2015).
3. Grimm, D. et al. Nature 441, 537–541 (2006).
4. McBride, J.L. et al. Proc. Natl. Acad. Sci. USA 105, 5868–5873 (2008).
5. Baek, S.T. et al. Neuron 82, 1255–1262 (2014).
6. Zuber, J. et al. Nat. Biotechnol. 29, 79–83 (2011).
7. Fellmann, C. et al. Cell Rep. 5, 1704–1713 (2013).
8. Gu, S. et al. Cell 151, 900–911 (2012).
9. Watanabe, C., Cuellar, T.L. & Haley, B. RNA Biol. 13, 25–33 (2016).
10. Fellmann, C. et al. Mol. Cell 41, 733–746 (2011).
11. Yuan, T.L. et al. Cancer Discov. 4, 1182–1197 (2014).
12. Knott, S.R.V. et al. Mol. Cell 56, 796–807 (2014).
13. Auyeung, V.C.C., Ullitsky, I., McGearry, S.E.E. & Bartel, D.P.P. Cell 152, 844–858
   (2013).
14. Viola, P. & Jones, M. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.
   1, 511–518 (2001).
15. Pelossof, R. Learning with Stochastic Focus of Attention PhD thesis, (Columbia
   Univ. 2011).
16. Leslie, C., Eskin, E. & Noble, W.S. Pac. Symp. Biocomput. 575, 564–575 (2002).
17. Sonnenburg, S., Rätsch, G. & Rieck, K. Large scale learning with string kernels.
   Large-scale Kernel Machines. (eds. Bottou, L., Chapelle, O., DeCoste, D. & Weston, J.)
   73–104 (MIT Press, Cambridge, MA 2007).
18. Vert, J.P., Fouveau, N., Lajaunie, C. & Vandenbergrouck, Y. BMC Bioinformatics 7, 520
   (2006).
19. Kampmann, M. et al. Proc. Natl. Acad. Sci. USA 112, E3384–E3391 (2015).
20. Matveeva, O.V., Nazipova, N.N., Ogurtsov, A.Y. & Shabalina, S.A. Front. Genet. 3,
   163 (2012).
21. Morgeis, D.W., Deans, R.M., Li, A. & Bassik, M.C. Nat. Biotechnol. 34, 634–636
   (2016).
22. Kampmann, M., Bassik, M.C. & Weissman, J.S. Proc. Natl. Acad. Sci. USA 110, E2317–E2326
   (2013).
23. Hart, T., Brown, K.R., Siriwattaram, F., Rottapel, R. & Moftak, J. Mol. Syst. Biol. 10,
   733 (2014).
24. Spies, N., Burge, C.B. & Bartel, D.P. Genome Res. 23, 2078–2090 (2013).
25. Derti, A. et al. Genome Res. 22, 1173–1183 (2012).
26. Iwagao, H., Iwagaki, K., Yang, J., Chiu, E., Loyer, J.-H. & Yeo, C.T. Cell 151, 900–911 (2012).
27. Wang, J., Lee, J.H., Puy, M., Yau, C.S., Mayr, C. Genes Dev. 27, 2380–2396 (2013).
ONLINE METHODS

MicroRNA-based shRNAs and minimization of off-target effects. Though RNAi triggers can be expressed as simple stem-loop shRNAs from RNA polymerase III (Pol-III) promoters in mammalian cells, such strategies can lead to off-target effects associated with high shRNA expression levels, likely due to saturation of the endogenous microRNA machinery. Many Pol-III-based systems also suffer from inaccurate processing of precursor molecules, yielding undesired mature small RNAs. In contrast, use of microRNA-embedded shRNAs expressed from RNA polymerase II (Pol-II) promoters results in accurate processing and can alleviate the toxic side effects, especially when used at single genomic integration (single-copy). Notably, highly potent mir-30-based shRNAs expressed at single-copy show the same low levels or absence of off-target effects as analogous weak and non-functional sequences. Hence, to develop an improved shRNA prediction algorithm, we focused on the optimized miR-E system that is based on the endogenous human Mir30A.

Here, to determine the extent of sequence-based off-target effects we applied the GESS algorithm to shRNAs validated by immunoblotting, and to previously reported Sensor assay and gene expression microarray results. GESS analyzes genome-wide enrichment of seed sequence matches. We tested whether potent shRNAs do not have more off-target effects than their weaker counterparts and if these targets have common sequences.

First, to investigate sequence-based off-target effects, we analyzed RNA expression microarray data from Trp53−/− MEF cells infected at single or high copy with one of six Trp53 shRNAs. Repetition of the published differential expression analysis found zero differentially expressed genes in the single-copy transfection setting relative to control experiments for either potent or weak shRNAs. In the high-copy transfection setting, 702 genes were upregulated and 326 genes were downregulated in the cells with potent shRNA with respect to control experiments (FDR < 0.05). Additionally, 2,437 genes were upregulated and 1,731 genes were downregulated in cells transfected with weak shRNA relative to their controls. Therefore, potent shRNAs in this setting did not induce more gene expression changes than weak shRNAs. Furthermore, both the potent and weak high-copy transfections resulted in near identical lists of differentially expressed genes: 702 genes were significantly upregulated in both lists and 324 genes were significantly downregulated in both lists. These intersections significantly overlapped (upregulated: P < 2.2 × 10−16, downregulated: P = 2.2 × 10−16, Fisher’s exact test), indicating that the main changes in gene expression are similar regardless of potency or shRNA sequence composition.

Second, we applied the GESS algorithm to our validation shRNAs that were quantified by immunoblotting to determine potential sequence-based off-target effects in our current experiments. We attributed our shRNAs to three categories based on western blot knockdown: Low (less than 80% knockdown), Mid (between 80% and 95% knockdown), High (95% knockdown or greater). For each gene and potency-level group, we ran GESS and found the genes that were potentially targeted by all the shRNAs in the group. We found no statistically significant off-targeted genes by GESS (FDR < 0.1). We also tested if the level of potency is associated with the number of potential off-target genes as measured by the number of perfect 7-mer seed matches (nucleotides 2–8). Grouping shRNAs into three groups by percent knockdown, High: >95%, Medium: 90–95%, and Low: 80–90%, and testing for a significant difference in the number of gene seed matches found no statistically significant difference between any pair of groups (P = 0.74, 0.53, and 0.73 for Low vs. Medium, Low vs. High, and Medium vs. High, respectively).

Third, we calculated all perfect 22-mer multi-mapping matches transcriptome-wide, since perfect matching of an shRNA to several genes would be highly undesirable. Consequently, we incorporated an additional feature into the SplasHRNA algorithm and web site that alerts the user if a predicted hairpin perfectly matches multiple genes in the human or mouse transcriptomes (hg38, mm10).

Sequence requirements of potent RNAi and prediction rules. The initial rules of RNAi potency contained many non-sequence elements, but later rules inferred from larger screens found that sequence-based features are more predictive and capture the other characteristics. BIOPREDi, a neural network approach, was trained on over 2,000 functionally tested siRNAs and set a new performance standard. Using the same data set, DSIR improved prediction through the use of an L1 regularized linear model with a combination of position-specific nucleotide features and mono-, di-, and tri-nucleotide counts. However, the rules governing siRNA potency differ from the ones dictating shRNA potency due to the additional biogenesis steps and siRNA-based algorithms perform relatively poorly in shRNA prediction tasks. Hence, we and others have previously used our large-scale data sets to generate miR-30-specific prediction algorithms. Yet, with a shift toward the more efficiently processed miR-E backbone, these algorithms are no longer designed for the task at hand as key sequence requirements have changed (Fig. 1a).

TILE, mRS + hRAS, and shERWOOD data sets. Over the years, a series of diverse shRNA potency data sets have been created, each having different characteristics and leveraging knowledge gained from previous studies. In the initial RNAi Sensor assay (referred to as “TILE”), we screened nearly 20,000 miR-30 based shRNAs that tailed nine mammalian genes in an unbiased manner to test all possible 22-mer sequences within these genes. This sampling strategy produces a low fraction of potent shRNAs. To reduce costs and increase the ratio of potent shRNAs, subsequent screens only assessed shRNAs that were predicted to be efficient by various in silico methods; these include the mRAS + hRAS and shERWOOD data sets. These data sets contain a higher percentage of potent shRNAs (as assessed by immunoblotting and functional RNAi screens, data not shown; Supplementary Table 1), but also represent a biased sampling of the sequence space. Additionally, the recent shift toward the use of “miR-E type” backbones that contain a 5’-DCNNC 3’ motif in their 3’-flank for improved pri-miRNA processing has further increased the fraction of efficient shRNAs and altered the overall sequence requirements for potent RNAi by relaxing constraints of Drosha processing (Supplementary Fig. 1h).

Sensor assay and M1 data set generation. A drawback of the unbiased TILE data set is that it contains few positives (potent shRNAs), with the benefit that it includes a large and comprehensive representation of negatives. Using the Sensor assay, we thus set out to establish a second large-scale miR-30 based data set containing a more comprehensive representation of positives (here referred to as M1; Supplementary Fig. 1a–f and Supplementary Table 2). The Sensor assay evaluates pools of shRNAs under conditions of single-copy genomic integration (“single-copy”) for their ability to repress a cognate target sequence placed downstream of a fluorescent reporter expressed in cis. This surrogate system showed an 85–90% specificity in identifying potent shRNAs when compared to knockdown of the corresponding endogenous genes by immunoblotting. Here, the Sensor assay was carried out as previously described, with several improvements to enhance deep-sequencing library preparation and readout accuracy. To assemble the candidate list, 60 shRNAs per gene were selected using a combination of algorithmic predictions and “Sensor rules” requiring shRNA-specific features. Specifically, to generate the M1 shRNA Sensor library, a custom oligonucleotide array (Agilent Technologies) was designed containing 20,400 185-mer sequences.

Sensor assay and M1 data set generation. A drawback of the unbiased TILE data set is that it contains few positives (potent shRNAs), with the benefit that it includes a large and comprehensive representation of negatives. Using the Sensor assay, we thus set out to establish a second large-scale miR-30 based data set containing a more comprehensive representation of positives (here referred to as M1; Supplementary Fig. 1a–f and Supplementary Table 2). The Sensor assay evaluates pools of shRNAs under conditions of single-copy genomic integration (“single-copy”) for their ability to repress a cognate target sequence placed downstream of a fluorescent reporter expressed in cis. This surrogate system showed an 85–90% specificity in identifying potent shRNAs when compared to knockdown of the corresponding endogenous genes by immunoblotting. Here, the Sensor assay was carried out as previously described, with several improvements to enhance deep-sequencing library preparation and readout accuracy. To assemble the candidate list, 60 shRNAs per gene were selected using a combination of algorithmic predictions and “Sensor rules” requiring shRNA-specific features. Specifically, to generate the M1 shRNA Sensor library, a custom oligonucleotide array (Agilent Technologies) was designed containing 20,400 185-mer sequences.

Sensor assay and M1 data set generation. A drawback of the unbiased TILE data set is that it contains few positives (potent shRNAs), with the benefit that it includes a large and comprehensive representation of negatives. Using the Sensor assay, we thus set out to establish a second large-scale miR-30 based data set containing a more comprehensive representation of positives (here referred to as M1; Supplementary Fig. 1a–f and Supplementary Table 2). The Sensor assay evaluates pools of shRNAs under conditions of single-copy genomic integration (“single-copy”) for their ability to repress a cognate target sequence placed downstream of a fluorescent reporter expressed in cis. This surrogate system showed an 85–90% specificity in identifying potent shRNAs when compared to knockdown of the corresponding endogenous genes by immunoblotting. Here, the Sensor assay was carried out as previously described, with several improvements to enhance deep-sequencing library preparation and readout accuracy. To assemble the candidate list, 60 shRNAs per gene were selected using a combination of algorithmic predictions and “Sensor rules” requiring shRNA-specific features. Specifically, to generate the M1 shRNA Sensor library, a custom oligonucleotide array (Agilent Technologies) was designed containing 20,400 185-mer sequences.

Sensor assay and M1 data set generation. A drawback of the unbiased TILE data set is that it contains few positives (potent shRNAs), with the benefit that it includes a large and comprehensive representation of negatives. Using the Sensor assay, we thus set out to establish a second large-scale miR-30 based data set containing a more comprehensive representation of positives (here referred to as M1; Supplementary Fig. 1a–f and Supplementary Table 2). The Sensor assay evaluates pools of shRNAs under conditions of single-copy genomic integration (“single-copy”) for their ability to repress a cognate target sequence placed downstream of a fluorescent reporter expressed in cis. This surrogate system showed an 85–90% specificity in identifying potent shRNAs when compared to knockdown of the corresponding endogenous genes by immunoblotting. Here, the Sensor assay was carried out as previously described, with several improvements to enhance deep-sequencing library preparation and readout accuracy. To assemble the candidate list, 60 shRNAs per gene were selected using a combination of algorithmic predictions and “Sensor rules” requiring shRNA-specific features. Specifically, to generate the M1 shRNA Sensor library, a custom oligonucleotide array (Agilent Technologies) was designed containing 20,400 185-mer sequences.
Xhol/EcoRI (NEB) digested pTNL backbone vector. Ligation products were MfeI-HF (NEB) digested to reduce background noise. In step 2, the 3′ miR30-PGK-Venus fragment was cloned into the EcoRI/MluI sites, followed by BamHI-HF (NEB) digestion of the resulting ligation product to further reduce background noise. During each cloning step, a representation of at least 1,000-fold the complexity of the library was maintained. All cell culture and flow cytometry procedures of the Sensor assay, to gradually enrich for the most potent shRNAs, were conducted as previously described.10,11

High-throughput sequencing based quantification of library composition and analysis of changes in shRNA representation over sort cycles were carried out as previously described10,11, with several adaptations to enhance readout precision. In contrast to previous procedures, deep-sequencing template libraries were generated by PCR amplification of shRNA guide strands including adjacent 3′ flanking regions, from vector libraries or genomic DNA, leading to longer PCR products (361 nt). The forward primer binding to the shRNA loop, HiSeq_LOOP (p7+loop, 5′-CAACAGAAGACGGCATACAGATTAG TAAGGCAACAGATGT-3′, IDT), was shortened by one nucleotide in order for each PCR to start with the same base. To enable sequencing of pooled libraries, an index primer binding site and 6 nt indices were included in the reverse primers (HiSeq_Index-p5-NS, 5′-AATCTGATACGCCGAGCGCGAG ATCTGATACGCAAGTGCCACGTCGAACCTGCACTNNNNNAC TTGTGTTACCGCACTGTTCCAGGACGAGATGT-3′, N = index, IDT). The indices used were (index, library): 5′-TTGTGTAGCGCCAAGTGCCCAGC-3′ for Vector 1, 5′-ATCTGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNAC TTGTGTTACCGCACTGTTCCAGGACGAGATGT-3′ (index, IDT) that reads reverse into the guide strand. Per library, 4 to 38 million initial sequencing reads were obtained (Illumina). For each shRNA and condition, the number of completely matching sequences was determined, normalized to the total reads per sample, and imported into a database for further analysis (Access 2007, Microsoft).

Deep sequencing after two-step cloning of the shRNA-Sensor libraries showed that >99.7% of all designed vectors were successfully constructed and detected in both replicates (Supplementary Table 2). Five iterative rounds of fluorescence-activated cell sorting, with gates set to progressively select for only the most functional shRNAs, enriched for potent shRNAs (Supplementary Fig. 1b,c), as previously shown.10,11 While independent biological replicates correlated throughout the sorting procedure, correlation to the initial representation was progressively lost, showing that the assay specifically enriched potent shRNAs. The final Sensor score was uncorrelated to the initial representation (Supplementary Fig. 1d), and known controls behaved as expected and in high correlation with previous Sensor runs, even for non-functional shRNAs (Supplementary Fig. 1e,f). A Sensor score was computed as readout for shRNA potency (Supplementary Table 2). The Sensor score represents an integration of shRNA enrichment over all replicates. The Sensor score for each shRNA sequence (x) was quantified as the log fold-change of the number of read counts (rho) between third sort (S3) and its respective vector library (v), averaged over replicates (r). Thus the potency score takes the form:

\[
\text{score}(x) = \frac{1}{\sqrt{S3}} \sum_{r \in S3} \log_2(\rho(x,r,S3) + 1) - \frac{1}{\sqrt{|v|}} \sum_{r \in v} \log_2(\rho(x,r,v) + 1)
\]

To avoid potential division by 0, the counts used for the calculations were reads (parts per million, p.p.m.) + 1. To distinguish positives from negatives and integrate the data sets, we defined score cutoffs based on the score distributions for each data set. The distribution of scores for the TILE data set gives a clear separation of positive and negative shRNAs (Supplementary Fig. 2c and Supplementary Table 1). Thus we selected a threshold at the minimum score density between the two modes. The M1 set was generated by selecting shRNAs that were likely to be potent, and therefore the score distributions of the negatives and positives are less distinct. To determine the label for different shRNAs in the M1 set, we fit each mode of the distribution with a Gaussian function. Using these two Gaussians we calculated two thresholds, one at a false-positive rate of 5% and one at a false-negative rate of 5% (Supplementary Fig. 2e, Supplementary Table 1) in order to define the positive and negative examples.

Assessing the potency of an shRNA for the TILE and M1 data sets. A Sensor score was computed as readout for shRNA potency (Supplementary Table 2). The Sensor score represents an integration of shRNA enrichment over all replicates. The Sensor score for each shRNA sequence (x) was quantified as the log fold-change of the number of read counts (rho) between third sort (S3) and its respective vector library (v), averaged over replicates (r). Thus the potency score takes the form:

\[
\text{score}(x) = \frac{1}{\sqrt{S3}} \sum_{r \in S3} \log_2(\rho(x,r,S3) + 1) - \frac{1}{\sqrt{|v|}} \sum_{r \in v} \log_2(\rho(x,r,v) + 1)
\]

To avoid potential division by 0, the counts used for the calculations were reads (parts per million, p.p.m.) + 1. To distinguish positives from negatives and integrate the data sets, we defined score cutoffs based on the score distributions for each data set. The distribution of scores for the TILE data set gives a clear separation of positive and negative shRNAs (Supplementary Fig. 2c and Supplementary Table 1). Thus we selected a threshold at the minimum score density between the two modes. The M1 set was generated by selecting shRNAs that were likely to be potent, and therefore the score distributions of the negatives and positives are less distinct. To determine the label for different shRNAs in the M1 set, we fit each mode of the distribution with a Gaussian function. Using these two Gaussians we calculated two thresholds, one at a false-positive rate of 5% and one at a false-negative rate of 5% (Supplementary Fig. 2e, Supplementary Table 1) in order to define the positive and negative examples.

Assessing the potency of an shRNA for the shERWOOD data set. This data set was previously published.12

Assessing the potency of an shRNA for the miR-E data set. The score for each shRNA in the miR-E set was calculated as the relative reporter knockdown level measured by flow cytometry, normalized to the knockdown level measured for miR-E Ren.713 and miR-30 Pten.1524 in the same batch. The data were scaled independently for each batch to set miR-E Ren.713 at 100% and miR-30 Pten.1524 at 60% relative knockdown. All shRNAs above 80% were classified as positive, while all shRNAs below 80% were classified as negative (Supplementary Figs. 1g, 4d and Supplementary Table 1).

Assessing the potency of an shRNA for the UltramiR data set. The scores from the UltramiR cell viability screen were previously published (NCBI Gene Expression Omnibus, Series GSE62185)12. We limited our analysis to the shRNAs targeting 78 essential genes, as defined in the shERWOOD paper (Supplementary Table 2). UltramiR shRNAs were considered to be potent if they had a depletion score of less than –0.5 (Supplementary Fig. 4d).
Assessing the potency of an shRNA for the Essential genes data set. This data set was previously published\textsuperscript{19}. Phenotypes for each shRNA were calculated as the mean log\textsubscript{2} fold-change for the two replicates. Gene-level scores were calculated as the mean phenotype for the five shRNAs with the most negative phenotypes for each gene.

Assessing the potency of an shRNA for the Sensitivity genes data set. This data set was previously published\textsuperscript{19}. Only shRNAs appearing in both replicates were used for the analyses. Hit genes were defined as those with a reported $P$-value less than 0.05. The top sensitivity genes were those with the most positive mean phenotypes of their top five targeting shRNAs. Phenotype is defined as log\textsubscript{2} (toxin-treated/untreated).

Identifying gold-standard essential genes. The set of gold-standard essential genes and gold-standard non-essential genes was previously published\textsuperscript{23}. We reevaluated data from a published RNAi screen that used approximately 25 shRNAs per gene, or 4 sgRNAs per gene\textsuperscript{21}, to assess the efficiency of SplasRNA predictions to identify hit genes. We ranked shRNAs according to their SplasRNA score and compared the mean cell depletion values for the top scoring shRNAs against the reported gene-level cell depletion values using the reported gold-standard genes. We found that a library made from the top ten SplasRNA predictions per gene performed at least as well as the full library when identifying the gold standard genes (Supplementary Fig. 2c). Additionally, a library created by selecting the ten lowest scoring SplasRNA predictions for each gene performed statistically worse than a library created by selecting the top ten scoring shRNAs per gene ($P < 0.001$, empirical permutation test). This shows that SplasRNA allows selecting superior shRNAs, which in turn decreases off-target effects by reducing the false-discovery rate. The need for fewer shRNAs per gene also enables minimizing the complexity of RNAi libraries for multiplexed screens.

Classifier kernel. All SVMs were trained with the Shogun package\textsuperscript{37} using a weighted-degree kernel of order 22 and two spectrum kernels (k-mer length = 3). Each of our classifiers was constructed by the following kernel combination: ClassifierKernel = SpectrumKernel(pos1-15) + SpectrumKernel(pos16-22) + WeightedDegreeKernel(pos1-22) (Supplementary Fig. 2b,d).

Training the miR-30 classifier. When fitting the regularization parameter C for our miR-30 SVM, we used leave-one-gene-out nested cross-validation. We grouped shRNAs from the TILE miR-30 data set by target gene into outer-folds. For each outer fold, we held out shRNAs targeting one gene and optimized the parameter C on the shRNAs targeting the remaining genes through tenfold cross-validation. The M1 positive set was added to all training sets but was not used for selection of C or for validation. Performance on the outer held-out genes (Supplementary Fig. 2f). We trained our final classifier with the parameter setting C = 15 using all the TILE data and the M1 positive shRNAs. This classifier was used to predict on all other data sets.

Training the miR-E classifier. We used nested tenfold cross-validation to fit the C parameter for our miR-E SVM. We did not use leave-one-gene-out due to the lower number of shRNAs targeting each gene. The miR-E and UltramiR shRNAs were combined and split into ten outer folds. Within each of these folds, tenfold cross validation was performed to determine the optimal C parameter for that fold. Performance on the miR-E and UltramiR sets is reported on the outer held-out folds (Supplementary Fig. 3d). We trained our final classifier with the parameter setting C = 15 using all the miR-E and UltramiR data. This classifier was used to predict on all other data sets.

Calculating sequential predictor (SplasRNA) scores. The potency scores for all shRNA are first calculated using the miR-30 classifier. If the score does not exceed the threshold theta, this partial score is the final score for the shRNA. If the score does exceed the threshold, the final score is a weighted combination of the predicted scores from the miR-30 and miR-E classifiers.

$FinalScore(x) = \begin{cases} 
\alpha SVM_{miR30}(x) & \text{if } \alpha SVM_{miR30}(x) < \theta \\
\alpha SVM_{miR30}(x) + (1 - \alpha) SVM_{miRE}(x) & \text{if } \alpha SVM_{miR30}(x) \geq \theta 
\end{cases}$

Here x is the sequence of the shRNA to be evaluated, alpha (\(\alpha\)) is the mixing proportion between the two classifiers and theta (\(\theta\)) is the threshold.

Optimizing the sequential predictor. We set alpha to 0.6 and theta to 1.1 to retain good performance on miR-30 classification after analysis of the precision-recall trade-off between the miR-30 and miR-E classifiers. This performance accuracy is unattainable by a simple linear classifier \(\alpha SVM_{miR30} + (1 - \alpha) SVM_{miRE}\) (Supplementary Fig. 3a-c).

Calculation of DSIR scores. DSIR scores were calculated according to the published 21-nt linear model\textsuperscript{18,35}.

Calculation of sequence score (seqScore) scores. Scores were calculated as described in the paper\textsuperscript{19}.

Calculation of miR_Scan scores. Scores were calculated using software provided by the authors\textsuperscript{20}.

Calculation of intersections of all transcript variants per gene. Genomic regions and annotations for hg38 and mm10 were downloaded using the make-TranscriptDbFromUCSC function from the GenomicFeatures Bioconductor package\textsuperscript{38,39}. Transcript variants were grouped by gene using their Entrez ID and regions shared across all RefSeq transcript variants were calculated in R using the Bioconductor intersection function. Sequences for these intersections were then extracted using the BSgenome.Hsapiens.UCSC.hg38 and BSgenome.Musculus.UCSC.mm10 packages.

Primary data for hg38 was obtained from: Team TBD. BSgenome.Hsapiens.UCSC.hg38. Primary data for mm10 was obtained from: Team TBD. BSgenome.Musculus.UCSC.mm10: Full genome sequences for Mus musculus (UCSC version hg38). R package version 1.4.1.

Primary data for mm10 was obtained from: Team TBD. BSgenome.Musculus.UCSC.mm10: Full genome sequences for Mus musculus (UCSC version mm10). R package version 1.4.0.

Cell culture. Phoenix HEK293T viral packaging cells were grown in DMEM supplemented with 10% FBS (FBS), 100 U/ml penicillin and 100 \(\mu\)g/ml streptomycin (100-Pen-Strep). ERC chicken reporter cells were grown in DMEM supplemented with 10% FBS, 1 mM sodium pyruvate and 100-Pen-Strep, and frozen in 5% DMSO, 70% FBS and 25% culture medium. NIH/3T3 (ATCC) cells were maintained in DMEM with 10% bovine calf serum or 10% FBS (FBS) containing 100-Pen-Strep and were tested for absence of mycoplasma contamination. A375 (kind gift from Neal Rosen, MSKCC) were maintained in DMEM with 10% FBS and 100-Pen-Strep. All cell cultures were maintained in a 37 °C incubator at 5% CO\textsubscript{2}.

Retroviral transduction. Cells were transduced as previously described\textsuperscript{10}. Transduction efficiency was assessed 48 h after infection by quantification of fluorescent reporters using flow cytometry (Guava EasyCyte, Millipore). Where a specific infection rate was desired, test infections were carried out at different dilution rates and ideal infection ratios deduced. All shRNAs were assessed at single copy genomic integration (“single-copy”) by infecting target cell population at <20% of their maximal infection rate, guaranteeing <2% cells with multiple integrations\textsuperscript{10}. Transduced cell populations were usually selected 48 h after infection, using 1.0-2.0 \(\mu\)g/ml puromycin (Sigma-Aldrich) or 500–2,000 \(\mu\)g/ml G418 (Geneticin, Gibco-Invitrogen).

Immunoblotting. Cells were transduced at single-copy with the constitutive retroviral vector LEPG\textsuperscript{7} expressing the indicated miR-E shRNA constructs. NIH/3T3 or A375 cell pellets were lysed in Laemmli buffer (100 mM Tris-HCl pH 6.8, 5% glycerol, 2% SDS, 5% 2-mercaptoethanol). Equal amounts of protein were separated on SDS-polyacrylamide gels and transferred to PVDF membranes. The abundance of \(\beta\)-actin (ACTB, Afb) was monitored to ensure equal loading. Images were analyzed using the AlphaView software (proteinSimple) and quantified by ImageJ. Immunoblotting was performed using antibodies for Pten (1:1,000, Cell Signaling Technology, #9188, https://media.celsignal.com/pdf/9188.pdf), Bap1 (1:500, Bethyl Laboratories, #A302-243A, http://www.bethyl.com/product/pdf/A302-243A.pdf), Pbrm1 (1:500, Bethyl Laboratories, #A302-243A, http://www.bethyl.com/product/pdf/A302-243A.pdf).
Evaluation of shRNA and CRISPR-Cas9 based suppression of Cd9 in immortalized MEFs. miR-E shRNAs targeting murine Cd9 were designed using SplashRNA or our previous design strategy involving DSIR predictions filtered by “Sensor rules”. The six top predictions from each algorithm were cloned into RT3CEN (TRE3G-mCherry-miRE-PGK-Neo; generated based on RT3GEN7). sgRNAs were cloned into a retroviral vector (RU6sgC- pSIN.U6.sgRNA-EF1as-mCherry), which we constructed based on the pQCXIX backbone (Clontech). Parallel Tet-inducible shRNA and CRISPR-Cas9 based loss-of-function studies were performed in immortalized double-transgenic MEFs (CRT-MEFs) constitutively expressing Cas9 and rtTA-M2 from transgenic knock-in alleles at the Rosa26 loci. These MEFs were isolated from Rosa26.CAGGS-Cas9.P2A.GFP; Rosa26.rtTA-M2 (ref. 42) double-transgenic embryos (using standard protocols) and immortalized through retroviral transduction of a potent shRNA targeting Trp53 (MSCV-shTrp53.814), followed by serial passaging. Retroviral shRNA/sgRNA expression vectors were packaged using standard calcium-phosphate based transfection into Platinum-E cells (Cellbiolabs) and transduced into CRT-MEFs and RRT-MEFs under strict single-copy conditions, as previously described. Two days post-infection, shRNA expression was induced through addition of doxycycline (1 μg/ml); 6 d later cells were stained for surface Cd9 expression (Anti-mouse Cd9-APC, eBioscience, #17-0091-82). Cd9 expression was analyzed in mCherry+/shRNA-expressing cells and quantified by flow cytometry (LSR-II Fortessa, BD Biosciences). The sgRNA transduced cells were analyzed in the same way, quantifying Cd9 expression in mCherry+/sgRNA-expressing cells 8 d post-infection.

Statistical analysis. Specific statistical tests used are indicated in all cases.

Code availability. Source code that implements the main SplashRNA algorithm is provided (Supplementary Code).

Data availability. Screen data from the M1 Sensor assay and the miR-E reporter assay are provided (Supplementary Table 2). UltramiR data is also provided (Supplementary Table 2). Data from the other screens used for SplashRNA training and validation (Supplementary Table 1) have been previously published as reported.

Strehl, T., Doehle, B.P., Qin, Y., Macara, I.G. & Cullen, B.R. RNA 11, 220–226 (2005).
28. Boudreau, R.L., Martins, I. & Davidson, B.L. Mol. Ther. 17, 169–175 (2009).
29. Sigal, I.D. et al. Nature Methods 9, 363–366 (2012).
30. Khvorova, A., Reynolds, A. & Jayasena, S.D. Cell 115, 209–216 (2003).
31. Reynolds, A. et al. Nat. Biotechnol. 22, 326–330 (2004).
32. Schwarz, D.S. et al. Cell 115, 199–208 (2003).
33. Huesken, D. et al. Nat. Biotechnol. 23, 955–1001 (2005).
34. Sigoillot, F.D. et al. Nat. Methods 9, 363–366 (2012).
35. Filhol, O. et al. PLoS One 7, e48057 (2012).
36. Taxman, D.J. et al. BMC Biotechnol. 6, 7 (2006).
37. Sonnenburg, S. et al. J. Mach. Learn. Res. 11, 1799–1802 (2010).
38. Huber, W. et al. Nat. Methods 12, 115–121 (2015).
39. Lawrence, M. et al. PLoS Comput. Biol. http://dx.doi.org/10.1371/journal. pcbi.1003118 (2013).
40. Dow, L.E. et al. Nat. Protoc. 7, 374–393 (2012).
41. Platt, R.J. et al. Cell 159, 440–455 (2014).
42. Hochedlinger, K., Yamada, Y., Beard, C. & Jaenisch, R. Cell 121, 465–477 (2005).