Prediction of ultra-potent shRNAs with a sequential classification algorithm

Raphael Pelossof[1,*], Lauren Fairchild[1,2*], Chun-Hao Huang[3,4], Christian Widmer[1], Vipin T. Sreedharan[1], Nishi Sinha[5], Dan-Yu Lai[5], Yuanzhe Guan[5], Prem K. Premsrirut[5], Darjus F. Tschaharganeh[3], Thomas Hoffmann[6], Vishal Thapar[3], Qing Xiang[7], Ralph J. Garippa[7], Gunnar Rätsch[1], Johannes Zuber[6], Scott W. Lowe[4,8], Christina S. Leslie[1,#] and Christof Fellmann[5,9,#]

[1]Computational Biology Program, Memorial Sloan Kettering Cancer Center, New York, New York, USA.

[2]Tri-Institutional Training Program in Computational Biology and Medicine, New York, New York, USA.

[3]Memorial Sloan Kettering Cancer Center, New York, New York, USA.

[4]Cell and Developmental Biology Program, Weill Graduate School of Medical Sciences, Cornell University, New York, New York, USA.

[5]Mirimus Inc., 500 Sunnyside Blvd., Woodbury, New York, USA.

[6]Research Institute of Molecular Pathology, Vienna Biocenter, Vienna, Austria.

[7]RNAi Core, Memorial Sloan Kettering Cancer Center, New York, New York, USA.

[8]Howard Hughes Medical Institute and Memorial Sloan Kettering Cancer Center, New York, New York, USA.

[9]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, California, USA.

[*]These authors contributed equally to this work.

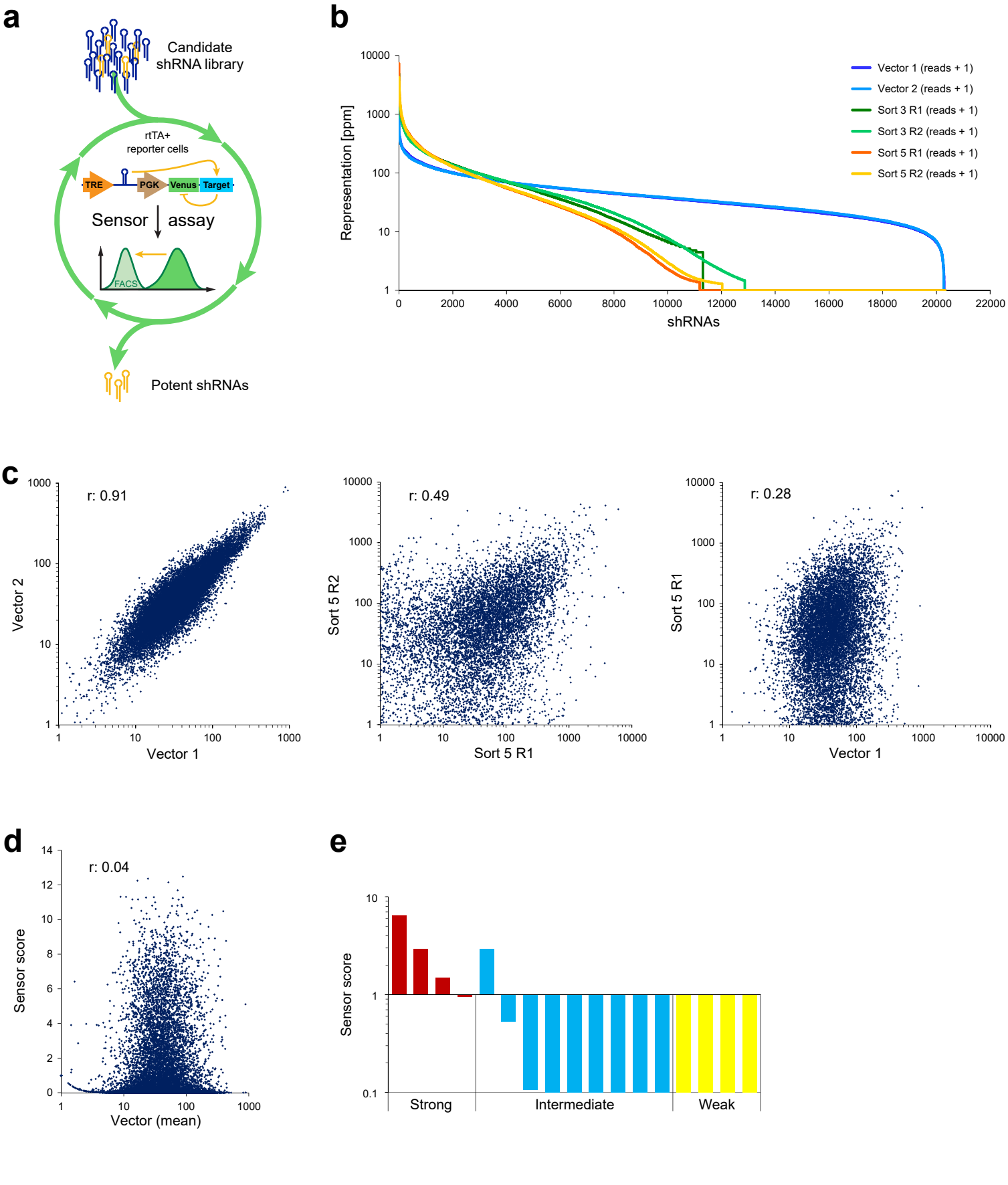
[#]Correspondence should be addressed to C.S.L. (cleslie@cbio.mskcc.org) or C.F. (fellmann@berkeley.edu).

Supplementary Information

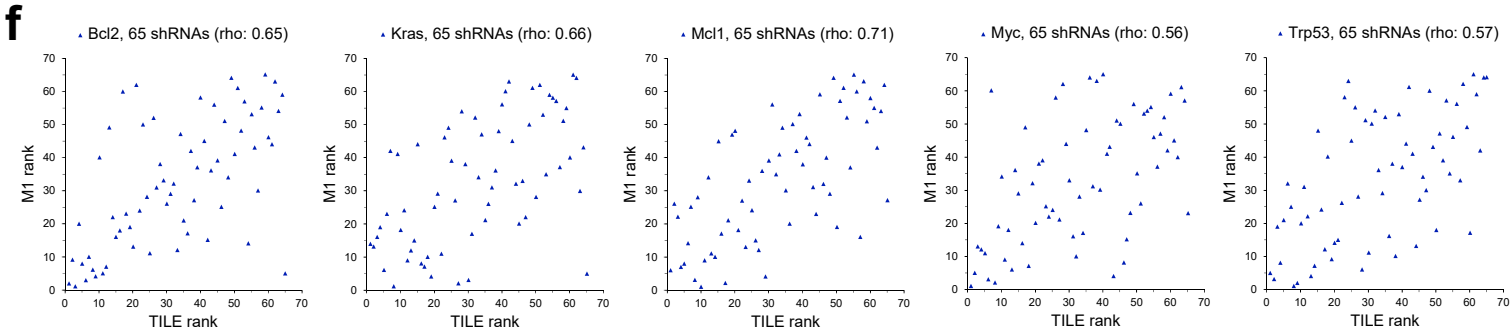
Supplementary Figures 1-6.

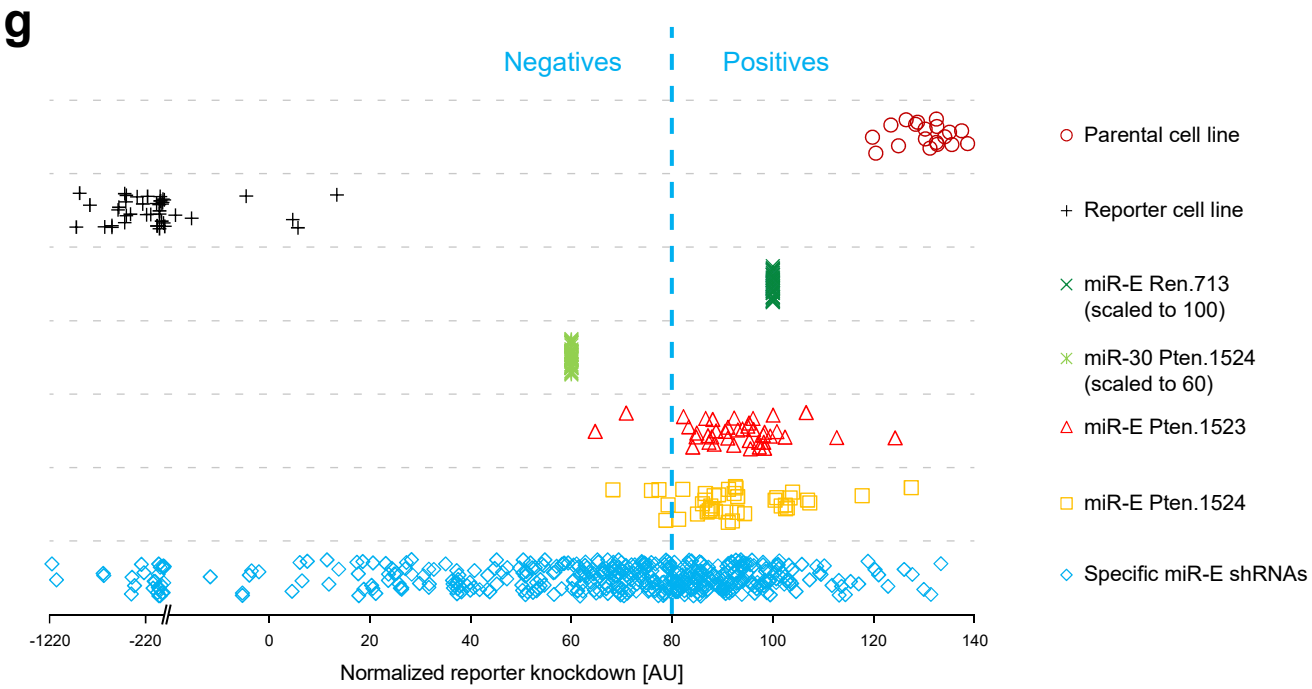
Supplementary Tables 1-3.

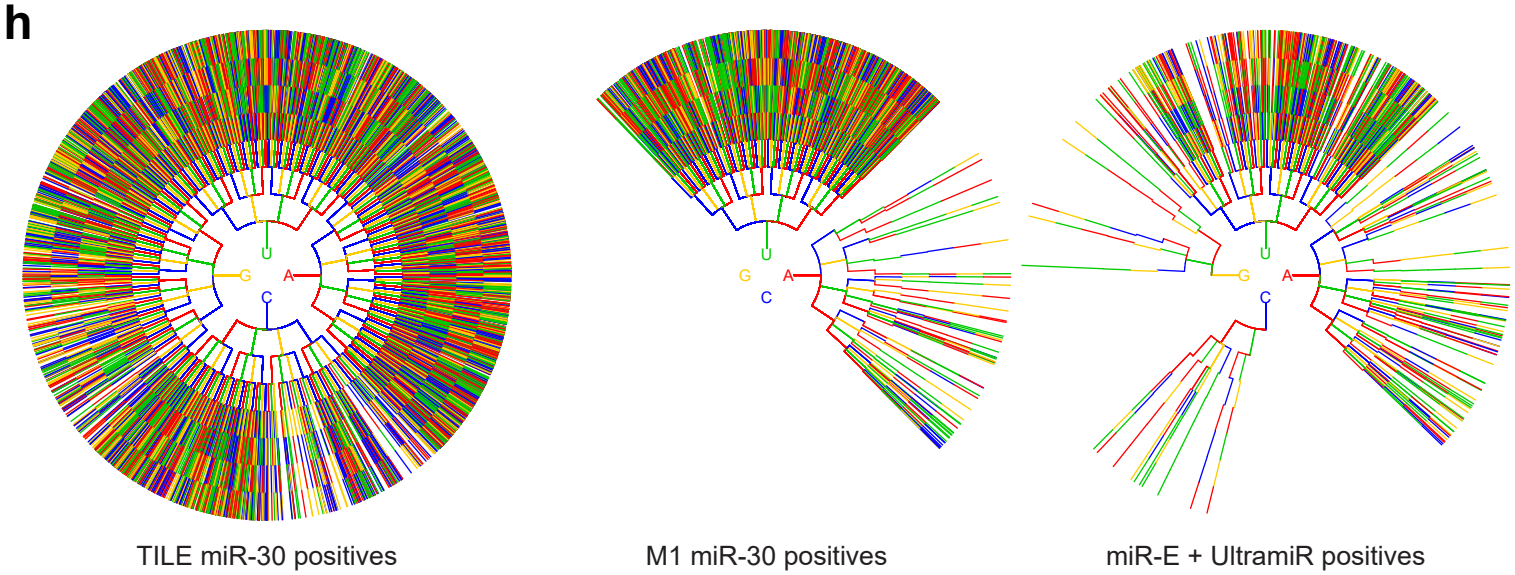
Supplementary Figure S1



Supplementary Figure S1







Supplementary Figure S1 Dataset generation.

(a-f) Generation of the M1 (miR-30, 20,400 shRNAs) Sensor assay dataset (**Sup Table S2, Methods**).

(a) Schematic of our previously published Sensor assay that enables large-scale functional assessment of shRNA potency (**Methods**).

(b) Library complexity over Sensor assay sort cycles. Shown are normalized read numbers (parts per million, ppm) in both duplicates for each shRNA represented within the initial libraries (Vector) and the pools after the indicated sorts (Sort 3, 5).

(c) Correlation of reads per shRNA between the two replicates before sorting (left panel), after Sort 5 (middle panel) and between the initial and endpoint population (right panel; shown for one representative replicate). r, Pearson correlation coefficient.

(d) Correlation of Sensor score and reads per shRNA in the vector libraries, showing that the score is independent of the initial shRNA representation. r, Pearson correlation coefficient.

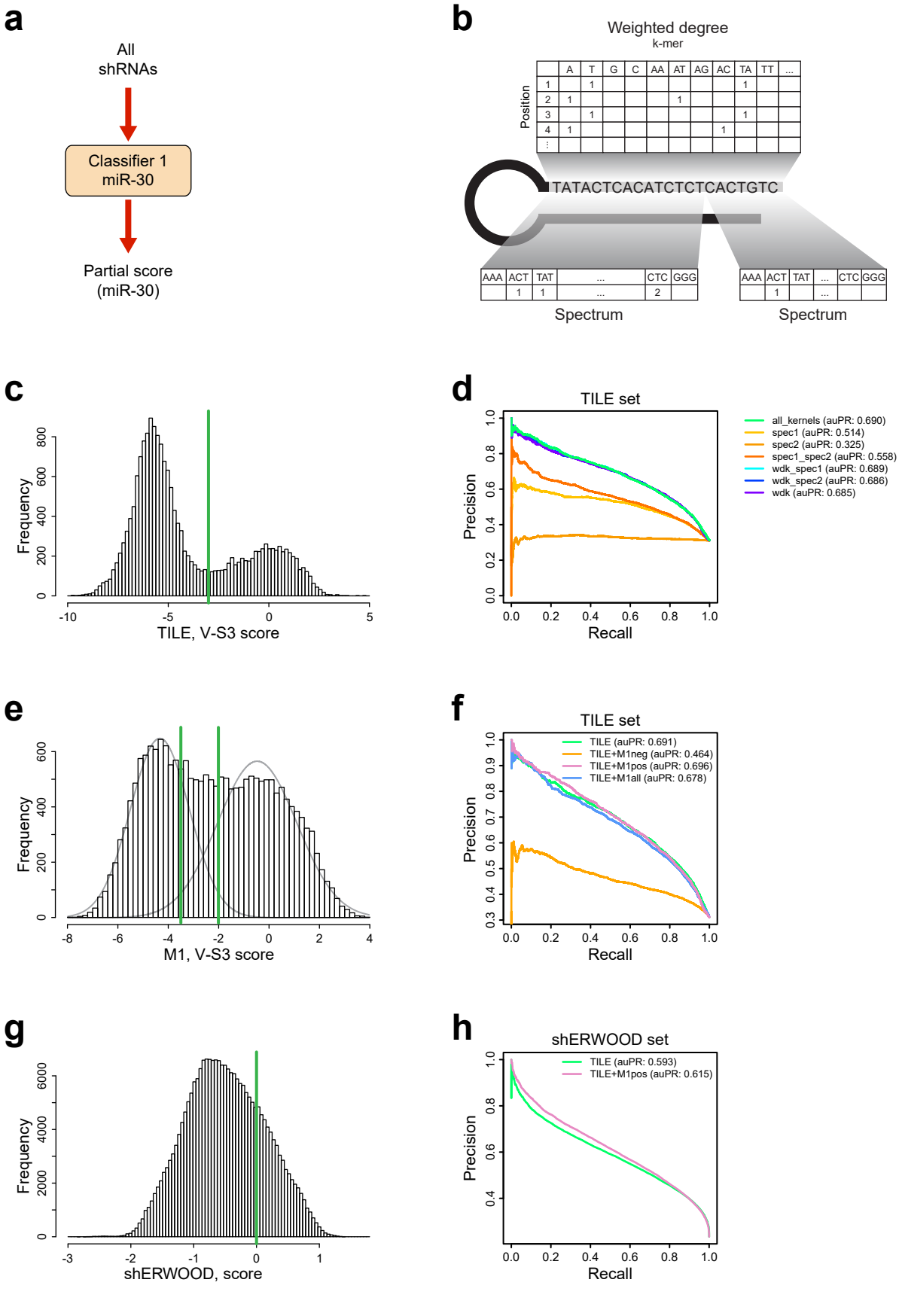
(e) Enrichment or depletion of 17 control shRNAs after Sort 5. All controls have been used in previous Sensor assays (e.g. TILE, mRas + hRAS) and are classified into a strong, intermediate and weak class according to their knockdown potency assessed by immunoblotting.

(f) Rank correlation of 325 performance control shRNAs. 65 shRNAs per gene targeting mouse *Bcl2*, *Kras*, *Mcl1*, *Myc* and *Trp53* that had previously been tested as part of the TILE dataset were chosen as supplemental controls to assess Sensor assay performance for weak, intermediate and strong shRNAs. The individual shRNA ranks between TILE and M1 were highly correlated (325 shRNAs, Spearman rank correlation coefficient rho: 0.63; gene-specific correlation coefficients are also reported), even though the TILE and M1 datasets were generated several years apart, using mostly different equipment, reagents and operators.

(g) Generation of the miR-E reporter assay dataset (**Sup Table S2, Methods**). Normalized reporter knockdown values of miR-E shRNAs assessed one-by-one in an RNAi reporter assay. The shRNAs were tested in 42 individual batches, each including several control shRNAs for data scaling (miR-E Ren.713, miR-30 Pten.1524) and quality control (miR-E Pten.1523, miR-E Pten.1524). Background fluorescence of the parental chicken cell line (ERC) and maximal fluorescence of the batch-specific reporter cell line (ERC cells expressing the shRNA target reporter) were also measured. All shRNAs were grouped into either a positive or negative class. A threshold value of 80 was chosen as a cutoff, based on the performance of miR-30 Pten.1524 and miR-E Ren.713.

(h) Nucleotide representation of positive shRNAs from the indicated datasets. Shown are the nucleotides one to eight of the guide strand (starting in the center), including the entire seed region. Unbiased TILE (miR-30) set, showing a diversified nucleotide composition (left panel). Preselected M1 (miR-30, DSIR + Sensor rules selected) set, showing a biased nucleotide representation (middle panel). Preselected miR-E + UltramiR set, showing a different nucleotide bias due to the altered shRNA backbone. More shRNAs starting with a C were found to be potent (compared to TILE, p-value = 0.002, Fisher's exact test), indicating less restrictive sequence requirements when using the miR-E backbone.

Supplementary Figure S2



Supplementary Figure S2 Kernel selection and data integration.

(a) Schematic of the first support vector machine (SVM) classifier that serves to eliminate non-functional sequences and prioritize shRNAs that are likely to be potent.

(b) Schematic of the kernel representation used by SplashRNA. A weighted degree kernel is calculated across the entire guide sequence, while two spectrum kernels are calculated across nucleotides 1-15 and 16-22, respectively.

(c) TILE score distribution (**Methods**). We set a potency threshold separating the negative from the positive class at the minimal point between the two modes of the distribution (green line, for thresholds see **Sup Table S1**).

(d) Testing of multiple kernel combinations in a leave-one-gene-out nested cross-validation setting on the TILE dataset found that the combination of a weighted degree kernel over positions 1-22 and two spectrum kernels at positions 1-15 and 16-22 (allKernels) yields the best performance. Spec1 is a spectrum kernel over positions 1-15. Spec2 is a spectrum kernel over positions 16-22. Spec1_spec2 is a combination of spec1 and spec2. Wdk is a weighted degree kernel over positions 1-22. Wdk_spec1 is a combination of wdk and spec1. Wdk_spec2 is a combination of wdk and spec2. All_kernels is a combination of wdk, spec1 and spec2.

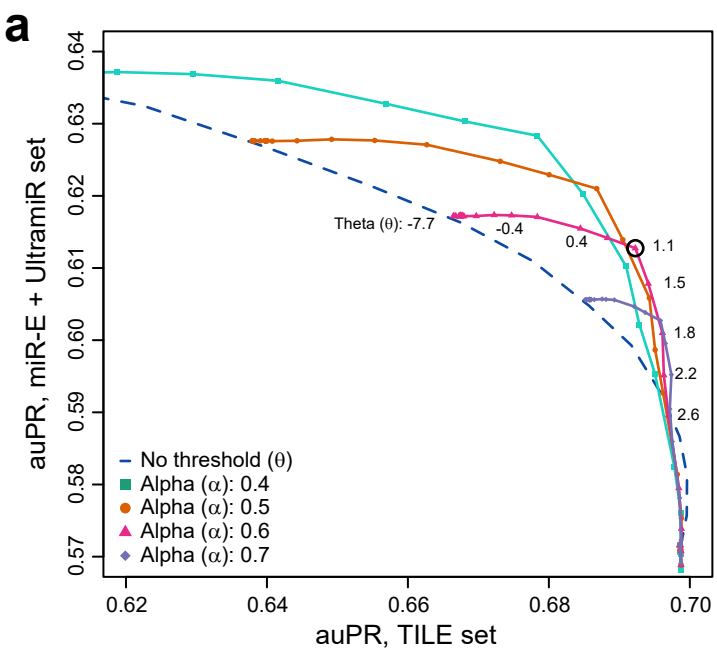
(e) M1 score distribution (**Sup Table S1, Methods**). Cutoffs (green lines) were calculated by fitting Gaussian distributions to the modes and setting thresholds at 5% FPR and 5% FNR.

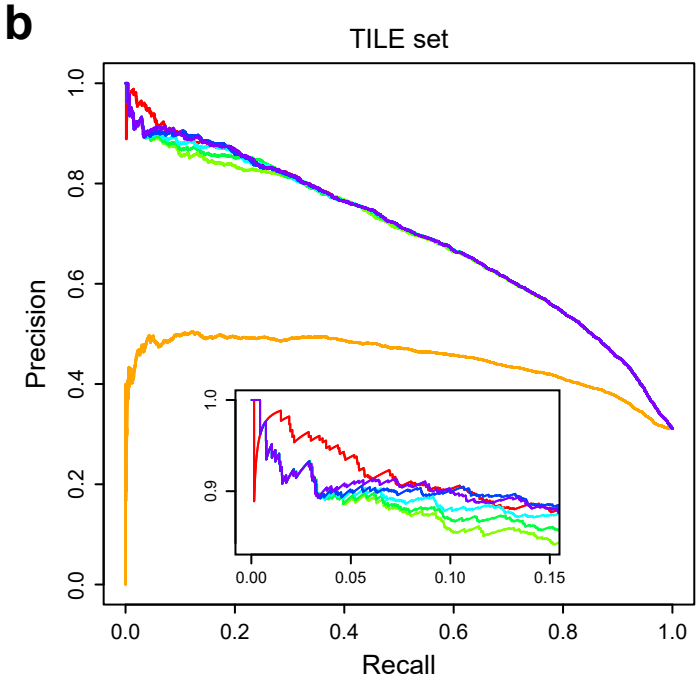
(f) Incorporation of M1 positives, negatives or both into the TILE training set was tested in a nested leave-one-gene-out cross-validation setting. Inclusion of M1 negatives deteriorated performance on the TILE dataset, whereas inclusion of the M1 positives alone improved performance. Note: TILE+M1pos = $\text{Splash}_{\text{miR-30}}$, the miR-30 classifier.

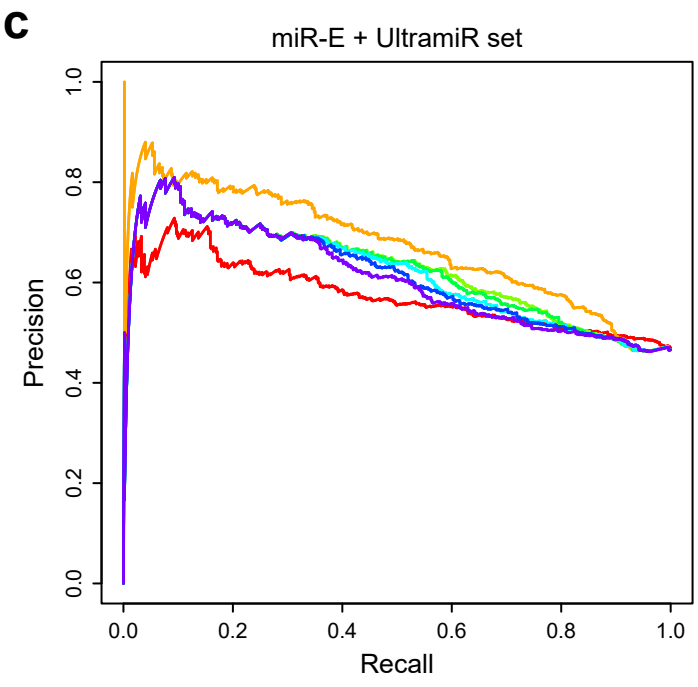
(g) Score distribution for the shERWOOD miR-30 set (**Sup Table S1, Methods**). We set the threshold at an arbitrary cutoff of zero (green line).

(h) Incorporation of M1 positives into the TILE training set improved performance on the external shERWOOD dataset. Note: TILE+M1pos = $\text{Splash}_{\text{miR-30}}$, the miR-30 classifier.

Supplementary Figure S3







Splash$_{miR-30}$ (auPR: 0.699)
Splash$_{miR-E}$ (auPR: 0.449)
Threshold (θ): 0.5 (auPR: 0.689)
Threshold (θ): 0.7 (auPR: 0.693)
Threshold (θ): 0.9 (auPR: 0.695)
Threshold (θ): 1.1 (auPR: 0.696) = SplashRNA
Threshold (θ): 1.3 (auPR: 0.697)

Splash$_{miR-30}$ (auPR: 0.572)
Splash$_{miR-E}$ (auPR: 0.670)
Threshold (θ): 0.5 (auPR: 0.623)
Threshold (θ): 0.7 (auPR: 0.620)
Threshold (θ): 0.9 (auPR: 0.614)
Threshold (θ): 1.1 (auPR: 0.611) = SplashRNA
Threshold (θ): 1.3 (auPR: 0.604)

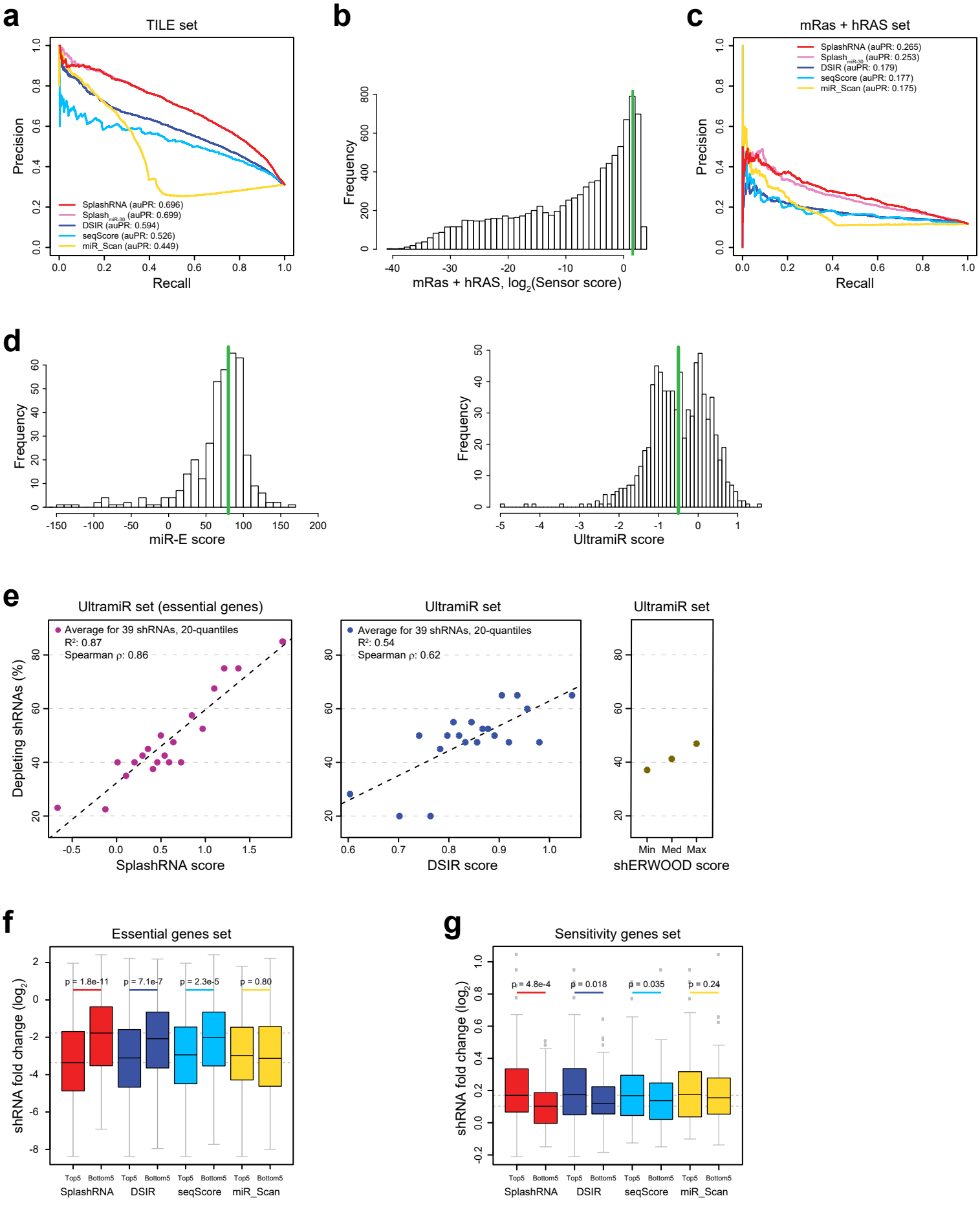
Supplementary Figure S3 Calibration of the sequential SVM classifier SplashRNA.

(a) Precision-recall trade-off between the two classifiers $Splash_{miR\text{-}30}$ and $Splash_{miR\text{-}E}$. Selection of alpha and theta hyperparameters leads to varied performance (area under the precision-recall curve, auPR) on the TILE miR-30 (x-axis) and miR-E + UltramiR (y-axis) sets. Each line represents a setting of alpha; points on the line represent distinct theta values. The circle indicates the alpha and theta choices for the final sequential classifier (SplashRNA, alpha = 0.6, theta = 1.1). The dotted line represents the performance of the convex linear classifier without a threshold at every alpha. Note that the performance of a sequential classifier equals or exceeds that of a linear combination since one can set the threshold to a large enough value such that all examples are evaluated by both classifiers.

(b) Performance on the TILE set, varying the value for theta with alpha set to 0.6. The insert shows a zoom in of the first 15% of the precision-recall.

(c) Performance on the miR-E + UltramiR set, varying the value for theta with alpha set to 0.6.

Supplementary Figure S4



Supplementary Figure S4 Prediction performance of SplashRNA.

(a) Precision-recall curves on the TILE dataset, comparing leave-one-gene-out nested cross-validation predictions from SplashRNA (auPR: 0.696) and $Splash_{miR\text{-}30}$ (auPR: 0.699) against the alternative prediction tools DSIR (auPR: 0.594), seqScore (auPR: 0.526) and miR_Scan (auPR: 0.449).

(b) Score distribution of the mRas + hRAS set (DSIR + Sensor rules selected). The green line indicates the threshold (**Methods, Sup Table S1**).

(c) Prediction performance comparison of the indicated algorithms on the external mRas + hRAS Sensor dataset (**Sup Table S1**). SplashRNA outperformed the other algorithms.

(d) Score distributions of the miR-E and UltramiR datasets. For the miR-E set, the threshold was set to 80 (green line, **Methods**). The UltramiR set represents the distribution of log depletion scores of shRNAs tested in a cell-viability screen (**Sup Table S1**).

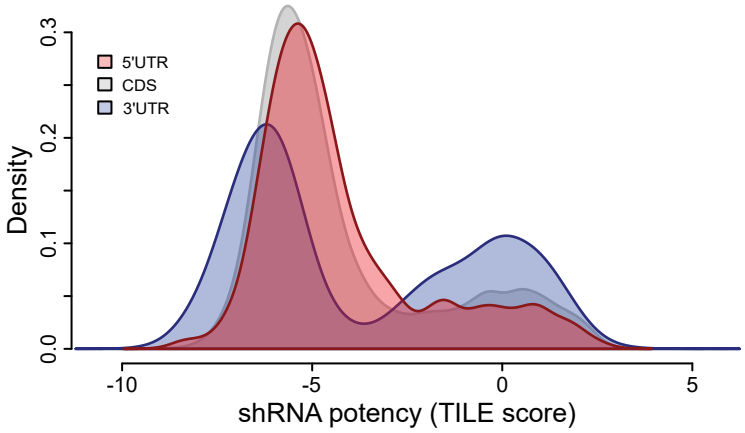
(e) SplashRNA and DSIR based re-ranking of shERWOOD selected UltramiR shRNAs targeting essential genes that were tested in a cell viability screen. X-axis: mean SplashRNA or DSIR score for equally sized groups (purple and blue dots, 20 groups) of 39 shRNAs each. Y-axis: Percent of shRNAs in each group that were potent (**Methods**). SplashRNA and DSIR were compared against the published minimum (Min), median (Med) and maximum (Max) shERWOOD algorithm performance on the same dataset (green-brown dots).

(f) Retrospective potency prediction of shRNAs from a large-scale essential genes RNAi screen. The biological screen used 20-25 miR-E-like shRNAs per gene to identify essential genes. shRNA potency was quantified by assessing their log fold changes (**Methods**). For each of the top 50 essential genes, all tested algorithms selected their top and bottom five sequences by prediction score. Log fold changes for all selected shRNA across the 50 genes were compared. SplashRNA achieved the most significant discrimination between top and bottom predictions ($p = 1.8e\text{-}11$, one-sided Wilcoxon rank sum test). seqScore ($p = 2.3e\text{-}5$) was used to generate the initial library of approximately 25 shRNAs per gene.

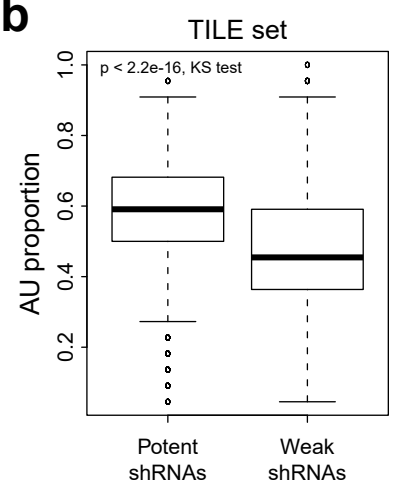
(g) Retrospective potency prediction of shRNAs from a large-scale toxin resistance and sensitivity RNAi screen. The biological screen used 25 miR-E-like shRNAs per gene to identify resistance and sensitivity genes. shRNA potency was quantified by assessing their log fold changes (**Methods**). For each of the top 20 sensitivity genes, all tested algorithms selected their top and bottom five sequences by prediction score. Log fold changes for all selected shRNA across the 20 genes were compared. SplashRNA was the only algorithm to achieve significant discrimination between the top and bottom predictions at $p < 0.01$ ($p = 4.8e\text{-}4$, one-sided Wilcoxon rank sum test). Of note, SplashRNA also outperformed the other algorithms when selecting smaller or larger numbers of top sensitivity genes from the biological screen (data not shown). seqScore was used to generate the initial library of approximately 25 shRNAs per gene.

Supplementary Figure S5

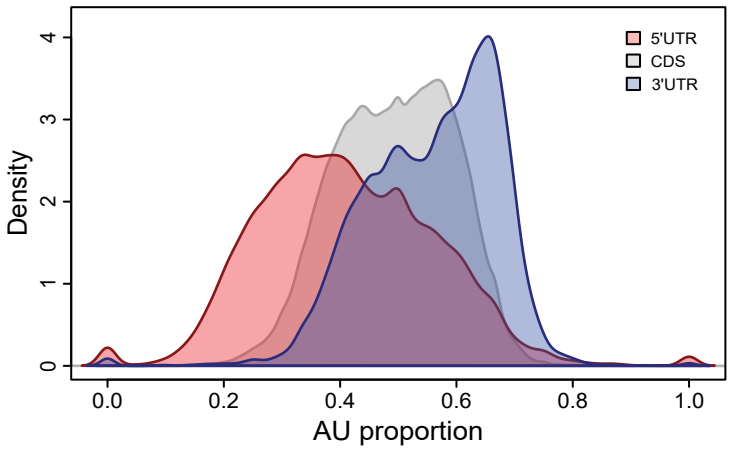
a



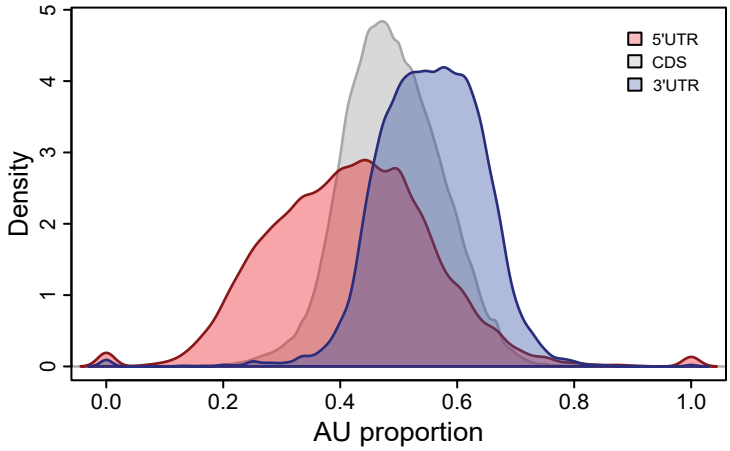
b



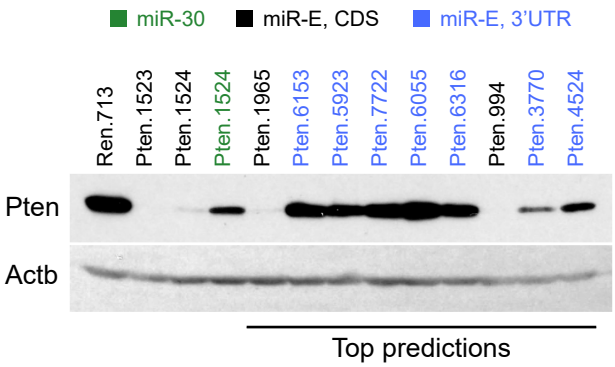
c



d



e



f

shRNA	Chr	Start (mm9)	End (mm9)	KD	ApA (mm9)	2P-Seq	Comment
Pten.994	19	32850552	32850573	+++			
Pten.1523	19	32889925	32889946	+++			
Pten.1524	19	32889926	32889947	+++			
Pten.1965	19	32894404	32894425	+++			
					32894814	184.0	Major APA
Pten.3770	19	32896209	32896230	++			
Pten.4524	19	32896963	32896984	+			
					32897818	47.0	Minor APA
Pten.5923	19	32898362	32898383	-			
Pten.6055	19	32898494	32898515	-			
Pten.6153	19	32898592	32898613	-			
Pten.6316	19	32898755	32898776	-			
Pten.7722	19	32900161	32900182	-			
					32900648	49.5	Poly(A)
Consensus poly(A)	19	NA	32900649	NA			

Supplementary Figure S5 Transcript selection.

(a) Distribution of shRNA potency in functionally distinct transcript regions. Shown is the potency distribution of shRNAs in the unbiased TILE dataset that target the 5'UTR, CDS or 3'UTR. Since these shRNAs were evaluated using the Sensor assay, their targets are not subject to alternative cleavage and polyadenylation (APA) and/or splicing events.

(b) A/U content of potent and weak miR-30 shRNAs from the unbiased TILE set. Potent shRNAs tend to have a higher proportion of A/U nucleotides ($p < 2.2e-16$, two-sided Kolmogorov-Smirnov test).

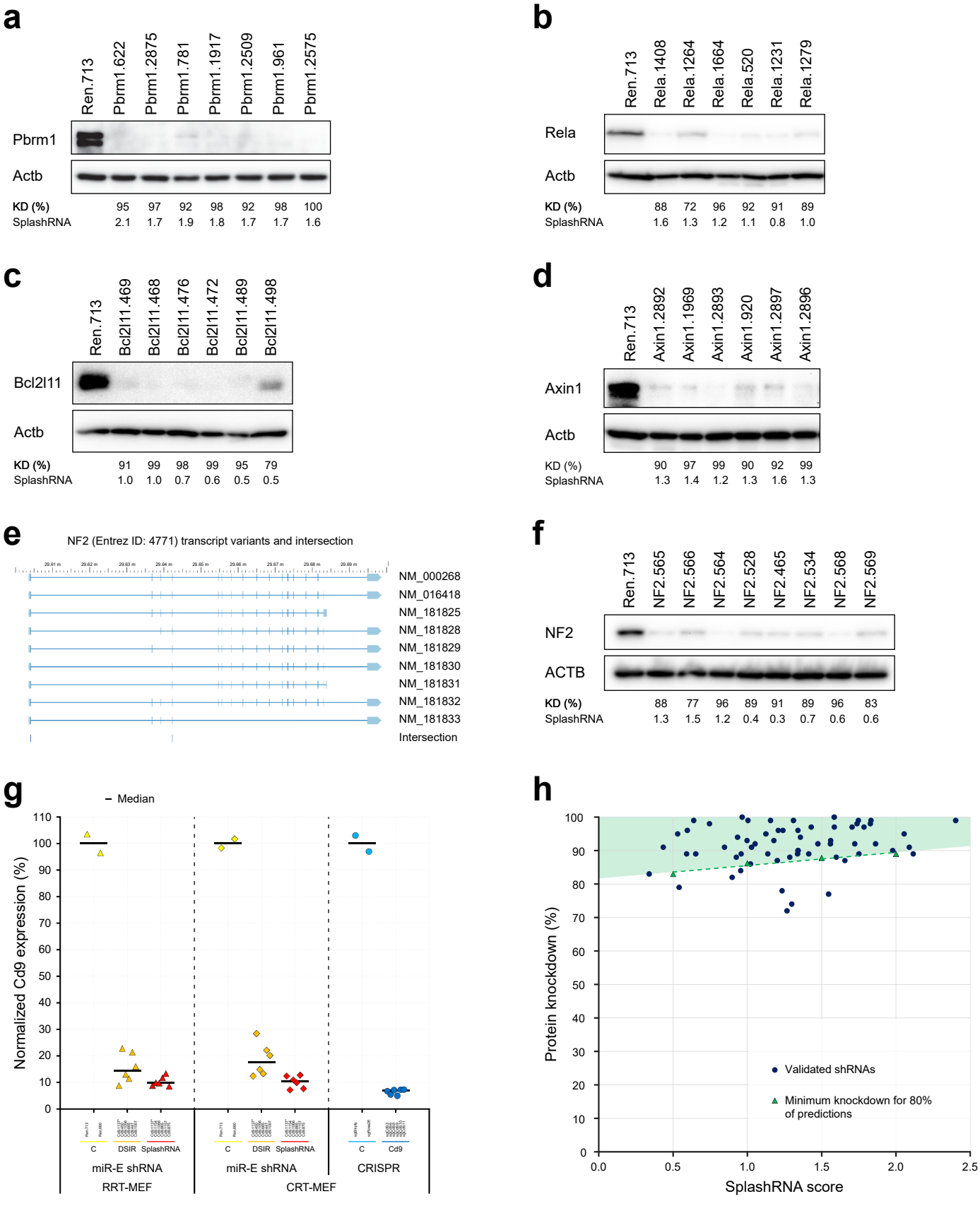
(c) A/U content of functionally distinct transcript regions in the human genome. Shown are the A/U densities in 5'UTR, CDS and 3'UTR.

(d) A/U content in mouse transcripts.

(e) Alternative cleavage and polyadenylation (ApA) prevents potent shRNAs from inhibiting their putative target gene. Immunoblotting of Pten in NIH/3T3s transduced at single-copy with LEPG expressing the indicated shRNAs. Nine top predictions targeting the CDS or the 3'UTR after early ApA sites were compared alongside controls for their ability to suppress mouse *Pten*. Actb was used as loading control.

(f) Comparison of knockdown efficiency and annotation of ApA sites. Shown are potent Pten shRNA predictions and their position (start, end) on the mouse genome (mm9). KD indicates a qualitative degree of the knockdown observed in immunoblotting analyses of NIH/3T3s (**e**). ApA indicates previously published positions on the mouse genome (mm9) of ApA sites identified in NIH/3T3 and mouse ES cells by 3P-seq. 2P-Seq shows the quantification of transcript expression levels measured by 2P-Seq. All shRNAs and ApA sites are ordered according to their position along the mouse genome.

Supplementary Figure S6



Supplementary Figure S6 Extensive validation of *de novo* SplashRNA predictions.

(a-f) Western blot validation of *de novo* SplashRNA predictions. All shRNAs were expressed using LEPG at single-copy conditions. β-Actin (Actb, ACTB) was used for normalization.

(a) Immunoblotting of Pbrm1 in NIH/3T3s (median KD: 97%, median SplashRNA score: 1.70).

(b) Immunoblotting of Rela in NIH/3T3s (median KD: 90%, median SplashRNA score: 1.15).

(c) Immunoblotting of Bcl2l11 in NIH/3T3s (median KD: 96.5%, median SplashRNA score: 0.65).

(d) Immunoblotting of Axin1 in NIH/3T3s (median KD: 94.5%, median SplashRNA score: 1.30).

(e) Graphical depiction of the multiple human NF2 transcript variants. NF2 has nine variants with an intersection of only 198 nucleotides, excluding the 5'UTR, rendering the prediction task especially difficult due to limited sequence space.

(f) Predicting miR-E shRNAs for extremely short transcripts. Immunoblotting of NF2 in A375s transduced with the indicated shRNAs targeting all nine NF2 variants (median KD: 89%, median SplashRNA score: 0.65).

(g) Comparison of SplashRNA and DSIR predictions against CRISPR-Cas9 mediated suppression of Cd9 in mouse embryonic fibroblasts (MEFs). Shown are normalized (relative to the indicated controls) median anti-Cd9-APC fluorescence intensities of RRT-MEFs and CRT-MEFs expressing the indicated shRNAs or sgRNAs (**Methods**). The six top-scoring predictions from DSIR + Sensor rules (DSIR) or SplashRNA (ordered according to their respective scores) were compared to six sgRNA sequences (**Sup Table S2**). *, Cd9.1137 is the top prediction of both algorithms and was plotted twice for clarity. While DSIR predictions triggered Cd9 knockdown with variable efficacy, SplashRNA predictions consistently induce strong Cd9 suppression, closely approaching knockout conditions.

(h) Transfer function of SplashRNA score versus protein knockdown for all 62 *de novo* predicted shRNAs validated by immunofluorescence (**Sup Table S2**). Green triangles indicate the minimum knockdown for 80% of the predictions for a given SplashRNA score bin. Bins were defined to have a width of 0.5 with the leftmost bin starting at 0.25. For the bin centered on SplashRNA score = 1, 80% of predictions showed at least 86% protein knockdown. The expected knockdown for the top 80% of predictions (e.g. 4/5 shRNAs) increases with the SplashRNA score. Together, 91% of predictions with a SplashRNA score >1 showed more than 85% protein knockdown.

Supplementary Table S1

Dataset	Backbone	Screen type	n (shRNAs)	n-pos	n-neg	Pos threshold	Neg threshold	Score type	Use	Availability
TILE	miR-30	Sensor assay, pooled	18720	5736	12685	-3	-3	V-S3	Training, validation	Published
M1	miR-30	Sensor assay, pooled	20324	9602	10722	-2	-3.5	V-S3	Training, validation	New
mRas + hRAS	miR-30	Sensor assay, pooled	9804	1139	8665	3	3	Score**	Validation	Published
shERWOOD 250k	miR-30	Sensor assay, pooled	227673	53234	174439	0	0	Score**	Validation	Published
miR-E	miR-E	Reporter assay, one-by-one	397	170	227	80	80	Score	Training, validation	New
UltramiR	UltramiR*	Cell viability, pooled	780	378	402	-0.5	-0.5	Log fold depletion	Training, validation	Published
Essential genes, Top50 hits	Mini miR-30 with DCNNC motif*	Cell viability, pooled	1002					Log fold depletion	Validation	Published
Sensitivity genes, Top20 hits	Mini miR-30 with DCNNC motif*	Toxin resistance and sensitivity, pooled	500					Log fold enrichment	Validation	Published

Supplementary Table S1 Novel and existing shRNA potency datasets used for training and performance assessment. The total count of shRNAs in each library (or selected sub-library) is indicated (n) along with the number of positive (n-pos) and negative (n-neg) examples chosen using the indicated thresholds. The score type indicates how the read counts were integrated (**Methods**). *, These microRNA-based shRNA backbones are functionally equivalent to miR-E. **, Score from original paper. For TILE, shRNAs with 0 reads in V1 or V2 were excluded from the set.

Supplementary Table S2 Novel datasets and sequences of validated shRNAs.

Sensor-M1 dataset tab: Sequences and scores of Sensor assay evaluated shRNAs from the M1 dataset. For each of the 20,324 unique shRNAs, the name (Gene name, Entrez ID, Species), the sequence (97-mer, Oligo 185-mer) and the final Sensor assay readout (Rank, Sensor score) are indicated along with the read counts (parts per million, ppm; Vector-1, Vector-2, S3R1, S3R2, S5R1, S5R2).

miR-E dataset tab: Sequences and scores of reporter assay tested shRNAs from the miR-E set. For each shRNA, the 97-mer sequence, the normalized reporter score and the class attribution are provided (**Methods**).

UltramiR dataset tab: Sequences and scores of shRNAs from the previously published UltramiR cell viability screen. For each shRNA the name (Name, Algorithm, Gene), sequence (97-mer, Guide), score (Score, pval) and class attribution are indicated (**Methods**).

Validation shRNAs tab: Sequences of shRNAs used for immunoblotting based validation of SplashRNA. For each sequence, the gene (Gene name, Entrez ID), shRNA name and 97-mer sequence are indicated. For *de novo* predictions, the SplashRNA score and measured protein knockdown level (KD%) are also indicated.

CRISPR sgRNAs tab: Sequences of sgRNAs used for Cd9 knockout and comparison to shRNA-based Cd9 knockdown. For each sgRNA, the name, target gene, guide sequence and measured median knockout levels are indicated.

Supplementary Table S3 Genome-wide SplashRNA predictions for all human and mouse protein coding genes. The predictions were designed to target the intersection of all transcript variants per gene (NCBI), after shortening of transcripts due to ApA (**Methods**). Where no ApA annotation was available, shRNAs were designed to target the CDS only (5041 human, 5058 mouse), unless this resulted in an intersection of length 0. For the few genes where there was no intersection between all transcript variants (118 human, 71 mouse), no predictions are reported. Predictions fully targeting (22/22 nucleotides) two or more distinct sites in the transcriptome were eliminated to avoid off-targeting. Transcript variant-specific or other custom predictions can be generated using the online implementation of SplashRNA (http://splashrna.mskcc.org).

Human tab: Top SplashRNA predictions for all human protein coding genes.

Mouse tab: Top SplashRNA predictions for all mouse protein coding genes.