A probability-based approach for the analysis of large-scale RNAi screens

Renate König, Chih-yuan Chiang, Buu P Tu, S Frank Yan, Paul D DeJesus, Angelica Romero, Tobias Bergauer, Anthony Orth, Ute Krueger, Yingyao Zhou & Sumit K Chanda

Supplementary figures and text:

Supplementary Figure 1: Large-scale siRNA library screening and analysis.

Supplementary Figure 2: Confirmation and validation of siRNA activities.

Supplementary Table 1: Original screen data for screen A.

Supplementary Table 2: Original screen data for screen B.

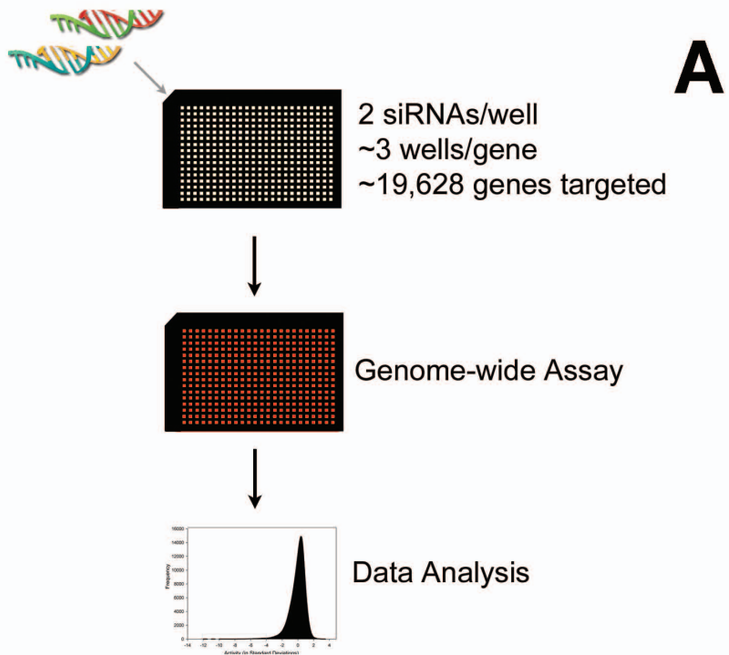
Supplementary Table 3: Scrambled control sequences.

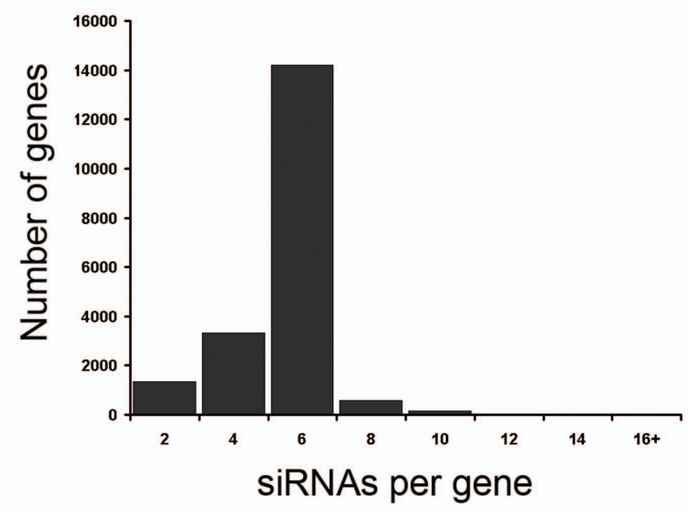
Supplementary Table 4: Reconfirmation screen data for screen A.

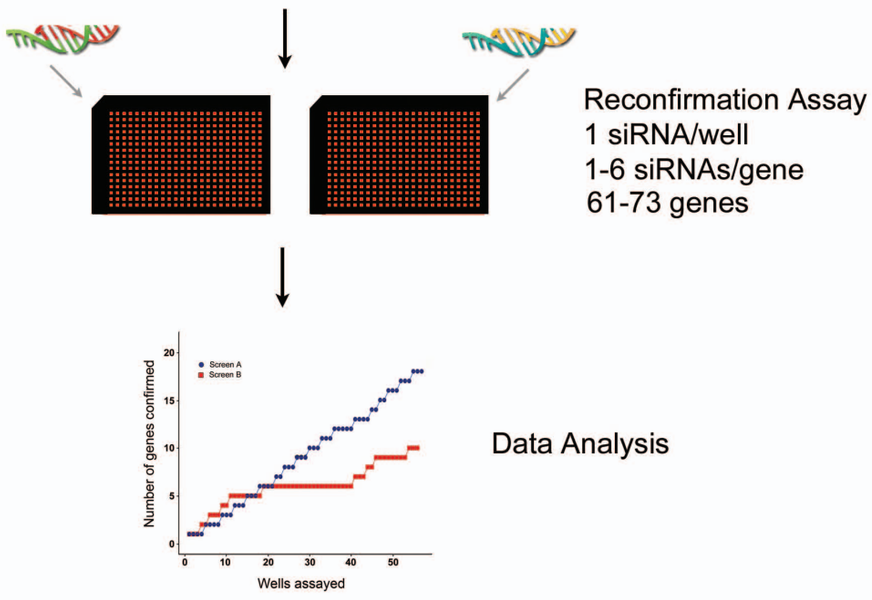
Supplementary Table 5: Reconfirmation screen data for screen B.

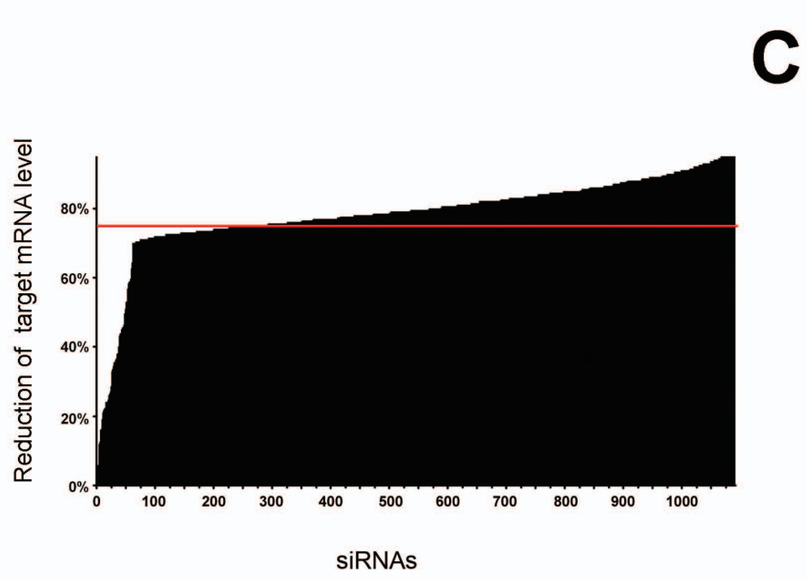
Supplementary Methods

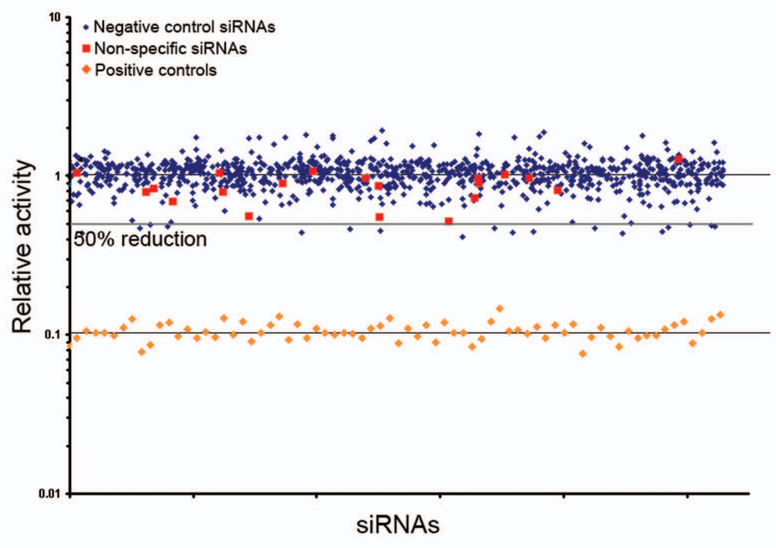
Supplementary Figure 1

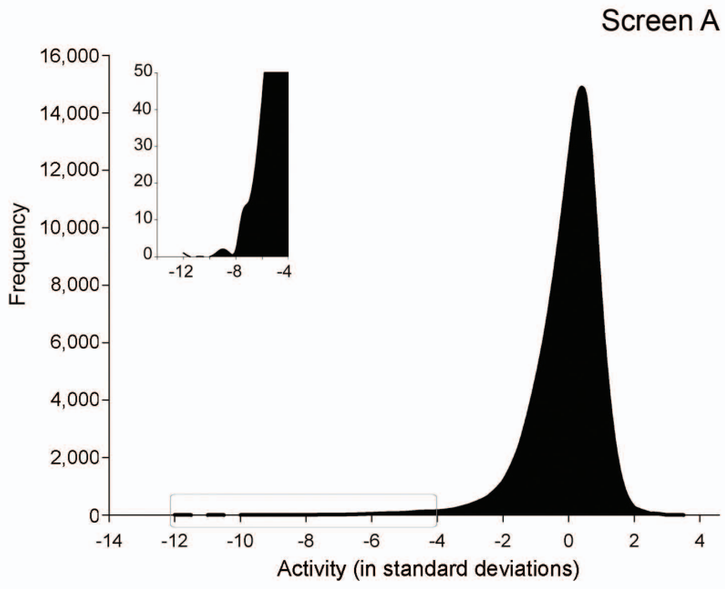


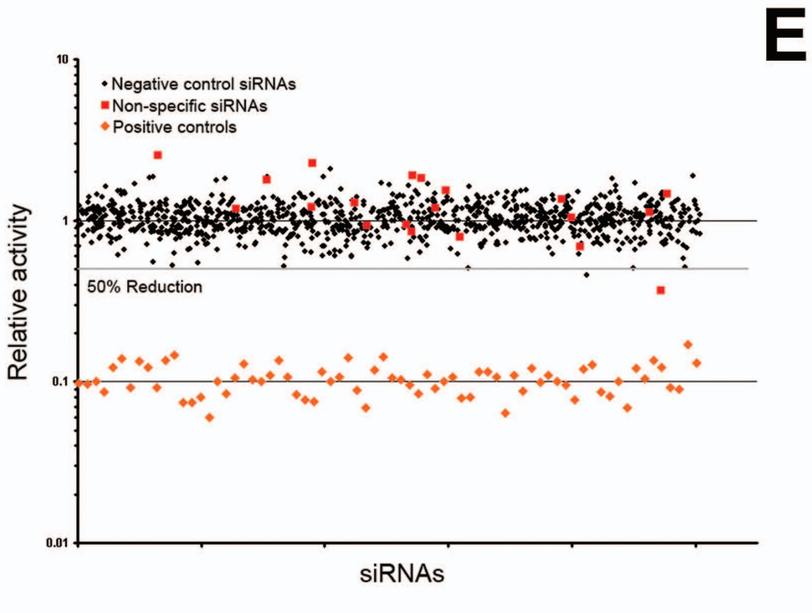


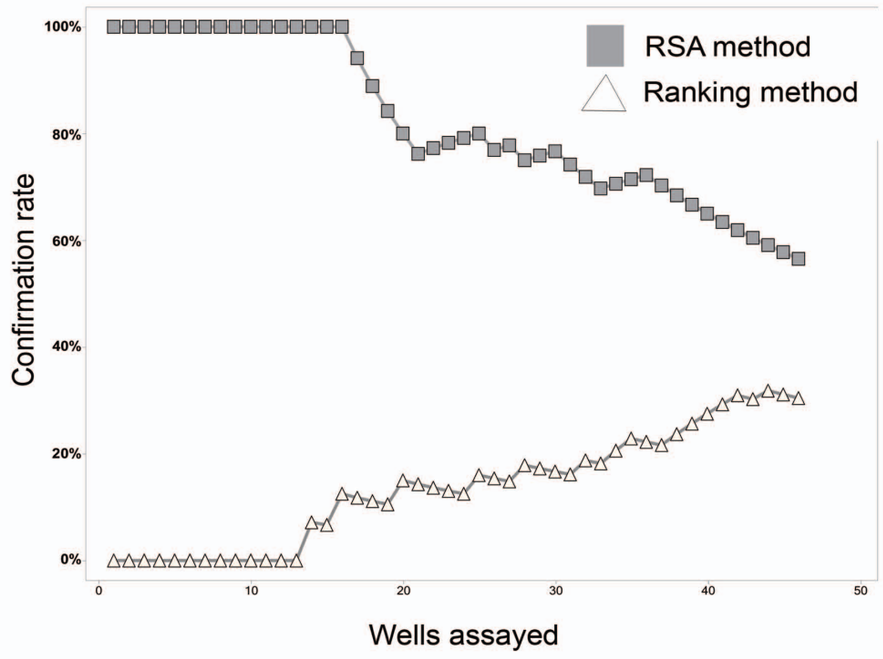












Supplementary Figure 1: Large-scale siRNA library screening and analysis.

Supplementary Figure 1a: Execution of original and reconfirmation screens**.** A genome-wide library comprising 107,734 synthetic siRNAs targeting 19,628 genes was spotted in an arrayed format such that each well contains two unique and identifiable siRNAs per gene (in average 3 wells/gene or 6 siRNAs/gene). Using this library matrix two different biological screen assays were performed and independently screened in either duplicates or triplicates. The results of each screen were analyzed by two different ranking methods. Approximately 55 top wells determined by each analysis methodology and 10 non-active wells were picked from each screen and the corresponding siRNAs were individually rearrayed in duplicate such as each well contained 1 siRNA. The reconfirmation assays were run in duplicates and comprised 61or 73 genes with 1 to 6 siRNAs/gene, screen A or B, respectively.

Supplementary Figure 1b: Redundancy of the siRNA library**.** The distribution of the number of siRNAs/gene is shown for the RNAi library used to execute the screens described in this study.

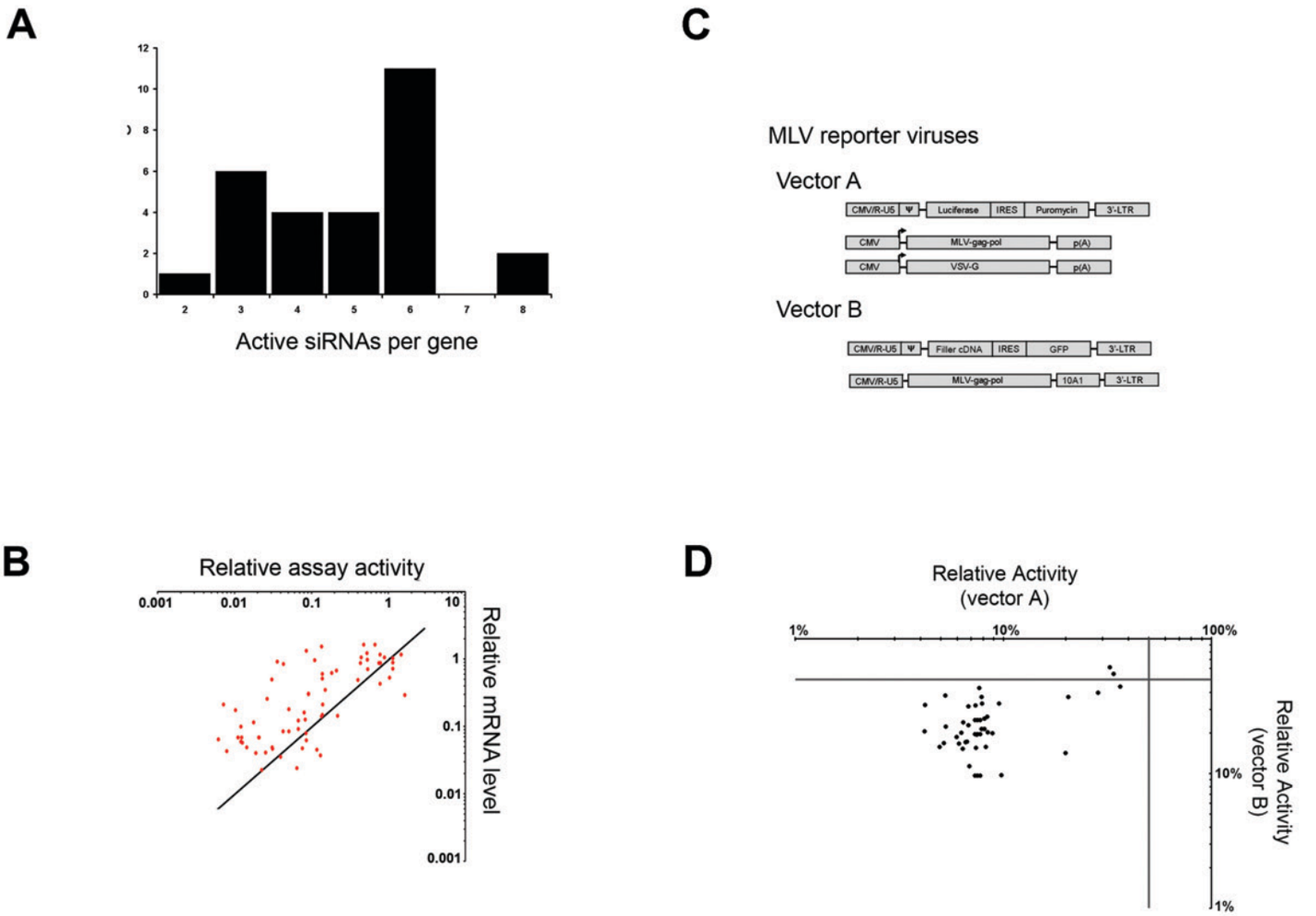
Supplemental Figure 1c: Efficiency of the siRNA library**.**1091 siRNAs targeting 548 genes were individually transfected into Hela, MCF-7, HEK 293, HELA S3, HEPG2,or A549 cells (dependent upon the presence of mRNA expression in the cell line), and quantitative RT-PCR was conducted to determine level of knockdown. Values were normalized to an internal control and plotted as a percentage of expression in comparison to levels seen in cells transfected with an siRNA scrambled control. 75% suppression is indicated by the red line.

Supplementary Figure 1d,e: Performance of negative and non-specific control siRNAs. (**d** and **e**) Negative controls were designed by scrambling ~ 6,000 siRNA sequence (with the 1st base G), and subsequently eliminating those with less than 3-4 mismatches to human and mouse mRNA sequences. Then, the remaining sequences were aligned to the human genome sequence, and sequences with less than 2 mismatches were removed. Finally, siRNAs with stretches of 4 nucleotides (i.e. CCCC, GGGG, etc) or runs of more than 5 of any 2 nucleotides (i.e. 6 C+G's. 6 A+T's or 6 A + G's, etc.) were removed. Of the 86 sequences that were remained from this analysis, 42 sequences were synthesized and used as negative controls for confirmation screens, as well as 3 commercially available controls (negative control GL2-Luciferase (Dharmacon), All-Star Negative Control (Qiagen) and Negative Control siRNA (Qiagen). The distribution of their activities (~24 replicates) after normalization to the median of negative controls for both Screen A (**d**) and Screen B (**e**) are shown (blue and black, respectively). Additionally, the activities of positive control siRNAs (targeting GL3 luciferase, Dharmacon) are shown (yellow). Finally, 10 non-active siRNA wells from Screen A and Screen B were picked and assayed for reconfirmation activity (red). Analysis of these wells showed that only 1 of 20 wells had activity which reduced signal <50%, suggesting false-negatives are not a major issue in large-scale siRNA analysis (5 $\pm$ 5%).

Supplementary Figure 1f. Analysis of Genome-wide siRNA Data. The frequency distribution of siRNA activities for Screen A (top panel) and Screen B (bottom panel) are plotted as a function of standard deviation from the screen median. Enlargement of the distribution "tail" (boxed, insets) shows Screen A contained a large number and more active outliers than Screen B.

Supplementary Figure 1g: Analysis of Screen C. Reconfirmation rates of wells based upon their original scores using either activity-based or probability-based rankings for Screen C are plotted from left to right.

Supplementary Figure 2



Supplementary Figure 2a,b: Activities of siRNAs targeting confirmed genes. (**a**) The distribution of phenotypically active siRNAs per confirmed gene in Screens A & B is shown. (**b**) Quantitative RT-PCR was performed to analyze the effects of 70 siRNAs targeting 10 confirmed genes from Screens A & B. Relative transcript levels after RNAi transfections was plotted against screen activity, confirming a correlation between RNAi silencing efficiency and phenotypic activity. All values were normalized to negative control values arbitrarily set to 1.

Supplementary Figure 2c,d: Reconfirmation rates in a secondary assay. (**c**) For additional confirmation analysis, we utilized a distinct set of viral packaging plasmids to generate virus carrying a GFP marker and psedutyped with an amphotropic envelopic (10A1; vector b). We used this virus to measure siRNA activities which were originally identified using a VSVG-pseudtyped virus modified to encode luciferase (vector a). (**d**) 45/47 siRNAs inhibited both viruses 50% or greater (gray lines). Each data point represents the average of 4 measurements by either FACS (vector b) or luminescence (vector a) readouts.

In the version of this supplementary file originally posted online, the equations were incorrect. The error has been corrected in this file.

Supplementary Methods

High-throughput siRNA screens

A 384-well plate-based assay was optimized to identify siRNAs that influence infection of human 293T cells by either a VSV-G pseudotyped MLV vector encoding luciferase (Screen A) or by a VSV-G pseudotyped HIV-1 reporter virus encoding luciferase (Screen B). The optimal ratio of the effect of the positive control siRNA (GL3) to that of the negative control siRNA (GL2) was obtained with a 1:10 dilution of the MLV virus stock and with 10 ng pf p24 of the HIV virus stock. A third assay (Screen C) was optimized to identify siRNAs that influence cell toxicity.

Two genome-wide libraries comprising 107,734 synthetic siRNAs in total were arrayed such that each well contains two unique and identifiable siRNAs per gene, and, on average, 3 wells/gene or 6 siRNAs/gene. This collection is comprised of a previously described library[1,2] (~2 siRNAs/gene) and a commercially available set (HP GenomeWide siRNA, Qiagen; ~4siRNAs/gene). 7 ng of each siRNA (in total 14 ng per well) were spotted on 384-well plates using a Minitrak liquid transfer instrument. Each plate contained also positive (GL3-Luciferase, Dharmacon) and negative (GL2-Luciferase, Dharmacon) control siRNAs (Screen A and B) or positive control RPS27a (AAGCUGGAAGAUGGACGUACU) for Screen C. The library matrix is introduced into mammalian cells through a high throughput transfection process[3,4]. These steps, and all subsequent procedures in the screening process, are performed through the use of proprietary robotic automation at GNF to ensure robustness and reproducibility of screening data. Briefly, a Staeubli robotic arm was used to catalog each plate by scanning unique barcodes on each of 384 well plates. Then, each plate was moved from a room temperature incubator to a bottle valve dispenser which deposits 10 µL of Optimem (Invitrogen) for dilution of the siRNAs. siRNA transfection reagent (Lipofectamine 2000, 45-80 nl/well) diluted in Optimem was then dispensed at 10 ul per well. After a 30-minute incubation, the transfection was completed by returning each plate to the bottle valve to dispense 20 µL of cells (3000 - 5000 293T cells/well) in serum containing medium. Each plate was then moved to a 37ºC and 5% CO_2 incubator. For Screen A and B, after 48 hours, MLV or HIV virus, respectively, was added at 10 ul/well and the cells were incubated for another 24 hours. The appropriate detection reagent (Bright-glo, Promega) for Screen A and B, or Cell-titer-glo (Promega) for Screen C, was added in equal volume to each well using the bottle valve dispenser, and the relative luminescence for each well was analyzed on a 384-well plate reader (Viewlux). For Screen C, the cells were incubated between 48-72 hours. Each screen was executed at least twice. For secondary screen with MLV-GFP virus, FACS analysis was used to establish virus infection efficiency.

Expression vectors and reporter viruses

For Screen A, we used the MLV-based retroviral vector pVGIP3, a derivative of pBabe[5], to generate a Moloney-based virus. pVGIP3-luciferase was generated by replacing

EGFP with GL3-luciferase by cloning into the unique EcoRI and NotI sites from pVGIP3. Then, Moloney MLV vector supernatant was generated by cotransfection of 293T cells with 12 ug of plasmid DNA consisting of a mixture of pcVSV-G, pCMVgp[6] and pVGIP3-luciferase in a 1:2:2 ratio. Two days later, retroviral supernatant was filtered and aliquoted. Similar methodologies were used to generate amphotropic envelope pseudotyped GFP virus used in the secondary assay.

For Screen B, we used the pNL43-Luc-E^-R^+ (HIV-1 Wild-type Δenv, encoding firefly luciferase GL3) vector to generate lentiviral supernatant[7]. Specifically, single-cycle HIV-1 luciferase reporter virus was generated by cotransfection of 293T cells with 8 ug of pNL43-luc- E^-R^+ and 4 ug of pcVSV-G. Two days posttransfection, virus was harvested, filtered through 0.45 um filter, aliquoted and quantified by p24 enzyme-linked immunosorbent assay (ELISA) from Beckman Coulter.

Reconfirmation screens

Approximately 50 top wells determined by each analysis methodology and 10 non-active wells were picked from each screen and the corresponding siRNAs were individually rearrayed in duplicate such as each well contained 1 siRNA (7 ng). 43 negative controls were designed by scrambling (see Supplemental Figure 3) and added to each plate in quadruplicates in addition to 3 commercially available controls (negative control GL2-Luciferase (Dharmacon), All-Star Negative Control (Qiagen) and Negative Control siRNA (Qiagen) and the respective positive controls. Each plate was measured twice in duplicate runs. The results were analyzed such as an siRNA is considered to be confirmed only if the median signal of all four readings is below 0.5 (50% reduction). The original well was considered to be reconfirmed if at least one of the 2 siRNAs showed this activity, and a gene was considered a true positive, only if there are at least two independent siRNAs confirmed based on this criterion.

Real Time PCR

Total RNA was extracted by using RNeasy 96 Kit according to the manufacture's instruction (Qiagen). RNA sample was reverse transcribed using the QuantiTect reverse transcription kit (Qiagen). PCR products were detected using the fluorescent dye SYBR green (Applied Biosystems). Formation of a unique DNA product was confirmed by verifying that products had a single melting temperature. Fluorescence-monitored PCR values were normalized.

Statistical Analyses

Data Normalization

Each screen was run in duplicates or triplicates. We only consider duplicates here to simplify the method description. Each well was assigned reading raw1 and raw2 from the two runs of a screen, and they are normalized into rawNorm1' and rawNorm2' by dividing with the median value of their plate readings, respectively. In the plate median calculation, all data wells were considered for the original screens, and only negative control wells were considered for the confirmation screens. On a given plate *K*, a weighting factor W_K was then calculated as:

$$W_K = |\mu_p - \mu_n|/(\sigma_p - \sigma_n), \quad (1)$$

where μ_p and μ_n denotes the mean of positive and negative control signals, σ_p and σ_n denotes their standard deviations. W_K, conceptually equivalent to Z-factor, represents the quality of a plate. *Score*' was calculated as the weighted average of rawNorm1' and rawNorm2':

$$Score' = (W_{k1} \cdot rawNorm1' + W_{k2} \cdot rawNorm2')/(W_{k1} + W_{k2}). \quad (2)$$

To further reduce among-plate signal variations, we scaled rawNorm1', rawNorm2' and *Score*' based on the positive control signals on each plate according to the following function:

$$f(x) = \exp(\ln(x)/\ln(\mu_p) \cdot \ln(0.1)) \text{ for } x<1 \text{ and}$$

$$f(x) = x \text{ for } x \geq 1. \quad (3)$$

The above scaling function normalized all positive control signals to 0.1, while keeping the plate median at 1.0. As the result, inhibition effect of siRNAs are more comparable across plates. rawNorm1, rawNorm2, and Score are the final normalized readings we reported for replicates and the weighted average. The Score value is used by the two hit picking mythologies in this study.

Ranking Method

Activity values were normalized to respective plate medians and the weighted average of replicates were sorted from most potent to least potent. Top hits were selected based upon this ranking methodology.

Redundant siRNA Activity Analysis (Probability-based Method)

All wells (N) are sorted, as in the Ranking Method, based upon their activities. A well i is assigned rank r_i in the sorted list. We then set two arbitrarily defined activity thresholds A_{min} and A_{max}, so that a well that is more active than A_{min} is guaranteed to be considered as an active well and a well that is less active than A_{max} is guaranteed to be considered as a negative well. In these screens, we define A_{min} as 0.2 and A_{max} as 0.8, when the screen median is 1.0. Thus, well 1, 2, …, a have activities better than A_{min} and are automatically assigned as active wells. Conversely, well b, …, n have activities worse than A_{max} and are automatically assigned as negative wells.

a. We then calculated the chance of having a wells from gene g, if one randomly select r_a wells from the library as:

$$P_a = P(N, n, r_a, r = a). \quad (4)$$

The function P is called the cumulated hypergeometric distribution function, and is mathematically written as

$$P(N,n,m,r) = \sum_{i=r}^{\min(m,n)} \frac{\binom{n}{i}\binom{N-n}{m-i}}{\binom{N}{m}} \quad (5)$$

b. and also calculated the chance of having a+1 wells from gene g, if one randomly select r_{a+1} wells from the library as:

$$P_{a+1} = P(N, n, r_{a+1}, r = a+1) \quad (6)$$

c. This was repeated for a+2, …, b wells to obtain P_{a+2}, …, P_b.

d. We then assigned the minimum of $\{P_a, \ldots, P_b\}$ as the p-value for all n wells. Thus, if P_k is the minimum value, then well 1, 2, …, a, …, k are called active wells and well k+1, …, b, … n are called negative wells.

These steps were repeated for all genes in the library, and all N wells were sorted by their p-value and then by their individual activities. This ranked list was utilized to select siRNAs for confirmation studies.

Mathematical Model to Determine the Relationship between Library Redundancy and Confirmation

For a "true hit", we denote P_w as the probability of a randomly selected mixture well to give active signal beyond a certain threshold (50% in this experimental scenario), and P_0, P_1, P_2 as the probability of an active hit picked well to yield 0, 1, or 2 validated siRNAs in the confirmation test, respectively. Therefore, given a randomly selected well from a true hit, the probability of that well lead to 0, 1, or 2 confirmable siRNAs are the following:

$$Q_0 = (1 - P_w) + P_w P_0 \quad (7)$$
$$Q_1 = P_w P_1 \quad (8)$$
$$Q_2 = P_w P_2. \quad (9)$$

Considering a gene has n wells in a library, it can be confirmed as long as there are at least two validated siRNAs among all the wells. The probability of this happening is

$$P_{\text{validated}} = 1 - Q_0^{\,n} - n \cdot Q_1 Q_0^{\,n-1}, \quad (10)$$

where the second term represents none of the n wells contain validated siRNAs, and the third term represents the probability that one out of the n wells contributes a validated siRNA and the remaining n-1 wells do not contribute any validated siRNA.

Mathematical Model to Account for Varying Library Efficiencies

For a given siRNA library design, we assume an arbitrary siRNA for a true positive gene has a probability of p in producing significant biological phenotype, and there are n

siRNAs per gene. Considering each siRNA is independently measured, we can formulate the probability of having two or more siRNA that confirm to be:

$$P_{\text{validated}} = 1 - (1-p)^n - n \cdot p \cdot (1-p)^{n-1} \quad (11)$$

The table below lists the confirmation rate as a function of siRNA quality (represented by p) and library redundancy (represented by *n*).

*n**p*	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
1	0	0	0	0	0	0	0	0	0
2	0.01	0.04	0.09	0.16	0.25	0.36	0.49	0.64	0.81
3	0.028	0.104	0.216	0.352	0.5	0.648	0.784	0.896	0.972
4	0.0523	0.1808	0.3483	0.5248	0.6875	0.8208	0.9163	0.9728	0.9963
5	0.08146	0.26272	0.47178	0.66304	0.8125	0.91296	0.96922	0.99328	0.99954
6	0.114265	0.34464	0.579825	0.76672	0.890625	0.95904	0.989065	0.9984	0.999945
7	0.149694	0.423283	0.670583	0.84137	0.9375	0.981158	0.996209	0.999629	0.999994
8	0.186895	0.496684	0.744702	0.893624	0.964844	0.99148	0.99871	0.999916	0.999999
9	0.225159	0.563792	0.803997	0.929456	0.980469	0.996199	0.999567	0.999981	1
10	0.263901	0.62419	0.850692	0.953643	0.989258	0.998322	0.999856	0.999996	1

The source code for the RSA algorithm is available for download from http://carrier.gnf.org/publications/RSA in both R and Perl languages.

References for Supplementary Methods

1. Mukherji, M. et al. Genome-wide functional analysis of human cell-cycle regulators. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 14819-24 (2006).
2. Huesken, D. et al. Design of a genome-wide siRNA library using an artificial neural network. [erratum appears in Nat Biotechnol. 2005 Oct;23(10):1315]. *Nature Biotechnology* **23**, 995-1001 (2005).
3. Chanda, S. K. et al. Genome-scale functional profiling of the mammalian AP-1 signaling pathway. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 12153-8 (2003).
4. Aza-Blanc, P. et al. Identification of modulators of TRAIL-induced apoptosis via RNAi-based phenotypic screening. *Molecular Cell* **12**, 627-37 (2003).
5. Morgenstern, J. P. & Land, H. Advanced mammalian gene transfer: high titre retroviral vectors with multiple drug selection markers and a complementary helper-free packaging cell line. *Nucleic Acids Research* **18**, 3587-96 (1990).
6. Agarwal, S. et al. Isolation, characterization, and genetic complementation of a cellular mutant resistant to retroviral infection. *Proc Natl Acad Sci U S A* **103**, 15933-8 (2006).

7. Connor, R. I., Chen, B. K., Choe, S. & Landau, N. R. Vpr is required for efficient replication of human immunodeficiency virus type-1 in mononuclear phagocytes. *Virology* **206**, 935-44 (1995).