The Power Decoder simulator for the evaluation of pooled shRNA screen performance

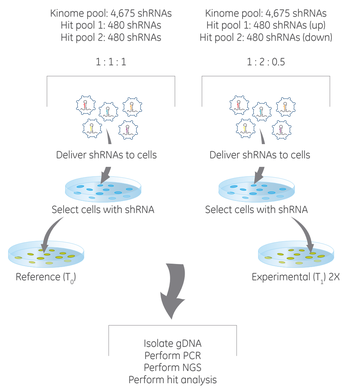
Jesse Stombaugh, Abel Licon, Žaklina Strezoska, Joshua Stahl, Sarah Bael Anderson, Michael Banos, Anja van Brabant Smith, Amanda Birmingham, Annaleen Vermeulen
Dharmacon, now part of GE Healthcare, 2650 Crescent Drive, Suite #100, Lafayette, CO 80026, USA

Abstract

RNA interference (RNAi) screening using pooled, short hairpin (sh)RNA is a powerful, high-throughput tool for determining the biological relevance of genes for a phenotype. Assessing an shRNA pooled screen's performance is difficult in practice; one can estimate the performance only by using reproducibility as a proxy for power or by employing a large number of validated positive and negative controls. Here, we develop an open-source software tool, the Power Decoder simulator, for generating shRNA pooled screening experiments *in silico* that can be used to estimate a screen's statistical power. Using the negative binomial distribution, it models both the relative abundance of multiple shRNAs within a single screening replicate and the biological noise between replicates for each individual shRNA. We demonstrate that this simulator can successfully model the data from an actual laboratory experiment. We then use it to evaluate the effects of biological replicates and sequencing counts on the performance of a pooled screen, without the necessity of gathering additional data. The Power Decoder simulator is written in R and Python, and is available for download under the GNU General Public License v3.0.

Results

Pooled screening experiments with engineered depletion and enrichment of shRNAs



Summary of screens

Experiment	Enrichment/ Depletion Magnitude	Average Independent Integrations per shRNA	Fold Representation in PCR Amplification
Screen 100_2x	2	100	100
Screen 100_4x	4	100	100
Screen 500_1.5x	1.5	500	500
Screen 500_2x	2	500	500

Materials and methods

HEK293T cells were transduced with three different shRNA pools that were combined in one lentiviral pool at the level of high titer lentiviral particles: Decode Pooled Human GIPZ Kinase Library (Dharmacon, Cat #RHS6078) enrichment and depletion sets.

gDNA was isolated from T_0 and T_1 samples and NGS libraries were prepared using Decode Indexing, PCR, and Sequencing primer kit (Dharmacon, Cat #PRM6178) following the Manufacturer's instructions and run on Illumina HiSeq 2000 (1x50 base reads, an average of 54 million reads per lane were obtained).

NGS reads were aligned using Bowtie (v0.12.7). Any shRNA with more than 50 perfect alignments was considered present in the NGS experiment. The differential abundance analysis was performed using DESeq (v1.10.1), which is an R (v2.15.3) package, part of the Bioconductor (v2.11.0) framework. DESeq uses a model based on the negative binomial distribution to estimate the significance of the fold change. It also applies the Benjamini-Hochberg multiple test correction to the reported p-values. Hits were classified as any clone that had a multiple-test corrected p-value of 0.05 or lower.

Actual differential enrichment and depletion of shRNAs in engineered screens

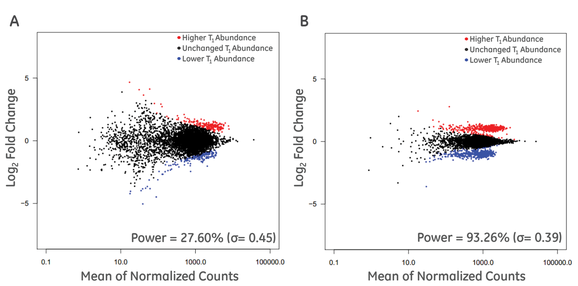


Figure 1. Differential enrichment and depletion of shRNAs in engineered screens. MA plots of representative examples of normalized data from experimental shRNA pooled screens with engineered two-fold enrichment and depletion of shRNAs where transductions were performed at **(A)** 100 and **(B)** 500 independent shRNA integrations on average. The shRNAs with significantly ($p^* \leq 0.05$) higher and lower abundance in T_1 in the NGS count data are in red and blue, respectively. Power values listed are mean +/- standard deviation over 30 normalizations.

Summary of actual screen data analysis

Experiment	Power (%)		Specificity (%)		False Positive Rate (%)		False Negative Rate (%)	
	Average	σ	Average	σ	Average	σ	Average	σ
Screen 100_2x	27.60	0.45	98.95	0.01	1.05	0.01	72.40	0.45
Screen 100_4x	76.10	0.55	98.68	0.01	1.32	0.01	23.90	0.55
Screen 500_1.5x	58.35	0.92	98.95	0.02	1.05	0.02	41.65	0.92
Screen 500_2x	93.26	0.39	98.60	0.01	1.40	0.01	6.74	0.39

Power Decoder: Building a Statistical Model

Simulate pooled screening NGS data by modeling:

1. The number of shRNA alignments in T_0 and T_1 to determine relative shRNA abundance
2. The biological noise between replicates.

Modeling T_0 Alignment Counts

Use negative binomial distribution for simulating alignment data. Use R's **fitdistr** function, determine the mean **(μ)** and dispersion parameters from the normalized means of the actual biological replicates from T_0 data in Screen 100 and 500.

$$T_0^* = NB\ [\mu = (T_0^t),\ dispersion = f\ (\mu)]$$

T_0^*: simulated alignment counts in T_0

T_0^t: actual alignment counts in T_0

NB: sampling using the negative binomial mode and the mean-dispersion relationship function

Modeled NGS alignment data is similar to actual experimental NGS alignment data

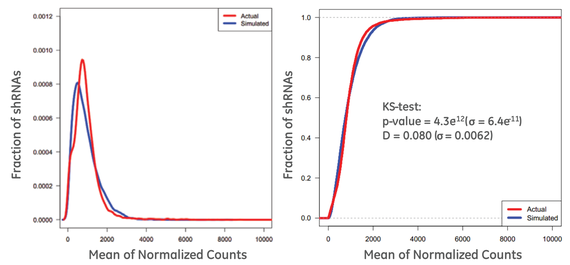


Figure 2. Modeled NGS screen data compared to actual experimental NGS screen data. Kernel density estimate plots for the distributions of NGS counts for representative examples of normalized actual (red) and simulated (blue) T_0 data generated by fitting parameters to the negative binomial distribution for Screen 500_2x (left). Cumulative distributions of the same actual and simulated T_0 count distributions for Screen 500_2x (right).

The mean-dispersion relationship of actual biological replicate noise for shRNAs in NGS data can be used to model biological noise

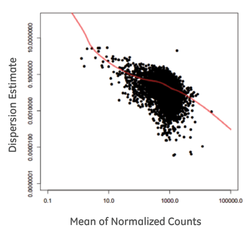


Figure 3. The mean-dispersion relationship of actual biological replicate counts for shRNAs in NGS data. Mean-dispersion relationship for representative examples of normalized experimental data from Screen 500_2x. The dispersion parameter for each individual shRNA graphed as a function of mean number of counts for that shRNA. The red line is a regression line fitted through the points using the DESeq software, and can be used to model biological noise.

Modeling T_1 Alignment Counts

Randomly select 20% of the shRNAs from modeled T_0 data and apply fold change. To produce simulated alignment (T_1^*), replicates can be sampled using the mean-dispersion function from Figure 3.

$$T_1^* = NB\ [\mu = (T_0^t * fold\_change),\ dispersion = f\ (\mu)]$$

T_1^*: simulated alignment counts in T_1

Modeling Biological Replicates

Knowing the relationship between the mean and the dispersion parameter of the negative binomial distribution, count data can be sampled to produce any number of replicates:

$$f(\mu_i) = dispersionParameter_\mu$$

Where μ_i is the mean for an individual shRNA as determined from sampling from the count distribution model.

Biological replicate noise generated by the model is comparable to actual biological replicate noise

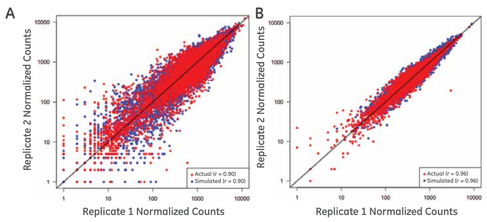


Figure 4. Simulated replicate noise compared to actual biological replicate noise. Comparison of replicate noise of representative examples of normalized actual (red) and simulated (blue) data for **(A)** Screen 100_2x and **(B)** Screen 500_2x. For each graph, the number of counts for each shRNA in biological replicate 1 is plotted as a function of the number of counts for the corresponding shRNA in biological replicate 2.

Differential enrichment and depletion of shRNAs in simulator generated screens

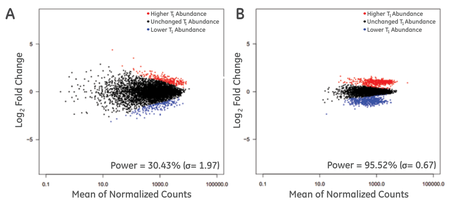


Figure 5. Differential enrichment and depletion of shRNAs in simulated screens. MA plots for two representative examples of simulated data from shRNA pooled screens with *in silico* two-fold enrichment and depletion of shRNAs based on **(A)** Screen 100_2x and **(B)** Screen 500_2x. The shRNAs with significantly ($p^* \leq 0.05$) higher and lower abundance in T_1 in the simulated NGS count data are in red and blue, respectively. Power values listed are mean +/- standard deviation over 900 simulations.

Comparison of actual and simulated screen data analysis

Experiment	Number of Replicates	Actual Power (%)		Actual Specificity (%)		Simulated Power (%)		Simulated Specificity (%)	
		Average	σ	Average	σ	Average	σ	Average	σ
Screen 100_2x	3	27.60	0.45	98.95	0.01	30.43	1.97	99.66	0.1
Screen 100_4x	3	76.10	0.55	98.68	0.01	84.37	1.25	99.15	0.15
Screen 500_1.5x	3	58.35	0.92	98.95	0.02	63.19	1.75	99.33	0.14
Screen 500_2x	3	93.26	0.39	98.60	0.01	95.52	0.67	99.00	0.16

Using the Power Decoder, screen sensitivity can be improved with additional biological replicates

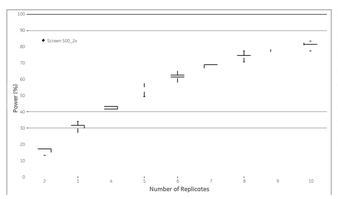


Figure 6. Power as a function of replicate number. Box plots represent powers derived from DESeq analysis of 900 simulated NGS experiments of Screen 100_2x per replicate level. For comparison, the actual power of the Screen 500_2x using two biological replicates is also plotted.

Conclusion

These investigations demonstrate how the Power Decoder simulator can help scientists plan future screens and easily investigate the likely effects of various experimental factors *in silico*, saving both time and money. Data from existing screens can also be analyzed retrospectively to evaluate their power and thus estimate the completeness of their resulting hit lists. Additionally, the Power Decoder simulator can be used to streamline optimization of novel pooled screening technologies such as gene knockout screens employing the new CRISPR (clustered regularly interspaced short palindrome repeats)-associated nuclease Cas9. The ability to do fast, easy, accurate power analyses before screening will enable researchers to perform adequately powered experiments, thereby delivering reliable answers to crucial biological questions.

gelifesciences.com/dharmacon

Dharmacon™
part of GE Healthcare