nature | **methods**

Statistical methods for analysis of high-throughput RNA interference screens

Amanda Birmingham, Laura M Selfors, Thorsten Forster, David Wrobel, Caleb J Kennedy, Emma Shanks, Javier Santoyo-Lopez, Dara J Dunican, Aideen Long, Dermot Kelleher, Queta Smith, Roderick L Beijersbergen, Peter Ghazal & Caroline E Shamu

Supplementary figures and text:

Supplementary Table 1 Summary of Assay Parameter Differences in Cell-Based Assays for siRNA and Small Molecule Screens

Supplementary Table 2 Comparison of Statistical Methods

Supplementary Table 1: Summary of Assay Parameter Differences in Cell-Based Assays for siRNA and Small-Molecule Screens

Assay Parameter	*Small-Molecule Assays*	*siRNA Assays*
Median Signal to Background (S/B; Note 1)	5.6 (n=14)	2.9 (n=18)
Median Coefficient of Variation (CV; Note 2)	13.4% (n=21)	26.5% (n=25)
Normality of Data (Note 3)	5% Normally Distributed (n=21)	26% Normally Distributed (n=25)
Z′-factor	Generally > 0.5	Generally < 0.5

Data sets from optimized siRNA screens and cell-based small molecule screens carried out at the ICCB-Longwood Screening Facility at Harvard Medical School were compared. The number of screens analyzed for each metric is provided (*n*). For each screen, data from five independent, non-consecutive 384-well screening plates (for the most part screened in duplicate) were analyzed. For siRNA screens, siRNA reagents were used as the positive and negative (non-targeting siRNA) control conditions, typically with 6 positive and 4 negative controls per plate. For small molecule screens, only mammalian cell-based assays were considered. Positive controls were small molecules that mimicked the desired hit phenotype (typically 16 positive controls per plate); negative control conditions were no compound added (typically 16 negative controls per plate). For S/B and CV, the median of all plate values for each assay type is provided.

Note 1: The average ratio of positive control signal to average negative control signal was calculated for each screening plate (with the inverse value taken if positive control value was less than background).
Note 2: Median and median absolute deviation (MAD) of experimental well values were used as a measure of CV.
Note 3: Normality of the distribution of measured values across different reagents was tested using lillietest function (MATLAB) on all non-control wells. (Note that this should not be confused with normality of the data from replicate wells containing the same reagent.)

Nature Methods: doi: 10.1038/nmeth.1351

Supplementary Table 2: Comparison of Statistical Methods

Normalization Methods			
% of control mean	$\% = \frac{\text{sample signal} * 100}{\text{mean of control}}$	• easy to calculate • easy to interpret • biologists are accustomed to this type of normalization	• very sensitive to outliers • need many replicates of controls • does not incorporate information on control variation • does not adjust for positional effects within plates
% of sample median	$\% = \frac{\text{sample signal} * 100}{\text{median of samples}}$	• easy to calculate • easy to interpret • not very sensitive to outliers • biologists are accustomed to this type of normalization • does not need many replicates of controls, because use samples as '*de facto*' negative controls	• does not incorporate information on sample variation • does not adjust for positional effects within plates
z-score	$z = \frac{\text{sample value} - \text{sample mean}}{\text{sample standard deviation}}$	• easy to calculate • incorporates information on sample variation • easy to use results in hit identification : typically use threshold of z >=2 or 3 • does not need many replicates of controls because use samples as '*de facto*' negative controls	• can be sensitive to outliers • plate-based z scores can be skewed if hits are unevenly distributed on plates (common) • does not adjust for positional effects within plates
robust z-score	$z = \frac{\text{sample value} - \text{sample median}}{\text{sample median absolute deviation}}$	• easy to calculate • incorporates information on sample variation • not very sensitive to outliers • does not need many replicates of controls because use samples as '*de facto*' negative controls	• does not adjust for positional effects within plates • some biologists may not be accustomed this type of normalization
B-score	$B = \frac{\text{sample residual}}{\text{sample median absolute deviation}}$ where: sample residual = sample value – plate, row, column, and possibly well corrections as necessary	• not sensitive to outliers • incorporates information on sample variation • does not need many replicates of controls because use samples as '*de facto*' negative controls • adjusts for positional effects within plates	• difficult to calculate • B scores can be skewed if real hits are heavily weighted in particular rows or columns • not intuitive for many biologists
Quality Metrics	**Formulae**	**Advantages**	**Disadvantages**
Z-factor	$Z = 1 - \frac{(3*SD \text{ of sample} + 3*SD \text{ of control})}{\lvert\text{sample mean} - \text{control mean}\rvert}$	• easy to calculate • takes signal dynamic range and data variation into account	• used extensively for small molecule assays; may need to adjust thresholds for siRNA screens
Z′-factor	$Z' = 1 - \frac{(3*SD \text{ of h.v. control} + 3*SD \text{ of l.v. control})}{\lvert\text{h.v. control mean} - \text{l.v. control mean}\rvert}$	• easy to calculate • takes dynamic range and data variation into account.	• can get "good" Z′-factor using very strong controls, but this may not be representative of positives from screen • used extensively for small molecule assays; may need to adjust thresholds for siRNA screens
SSMD	$SSMD = \frac{\text{h.v. control mean} - \text{l.v. control mean}}{\sqrt{(SD^2 \text{ of h.v. control} + SD^2 \text{ of l.v. control})}}$	• easy to calculate • takes dynamic range and data variation into account	• not yet as widely recognized as Z/Z′-factor

Nature Methods: doi: 10.1038/nmeth.1351

		• less conservative estimator than Z/Z′-factor • has rigorous statistical estimator for non-normal data • different thresholds available for different control strengths	

ROC	plot sensitivity vs. (1 – specificity) where:
	sensitivity = (# true-positives) / (# true-positives + # false-negatives)
	specificity = (# true-negatives) / (# true-negatives + # false-positives)
		• provides visual quality control
		• allows investigation of the effect of changing hit-threshold
		• can be quantitated as the area under the ROC curve
			• ideally, needs many replicates of multiple known positive and negative controls
			• time-consuming to review and interpret curves
			• loses some information when used only quantitatively

Hit Identification Strategies			
Mean + or - k standard deviations	Hit with increased activity = any sample whose value is >= sample mean + k standard deviations; Hit with decreased activity = any sample whose value is <= sample mean – k standard deviations	• easy to calculate • easily linked to hit p-values	• sensitive to outliers • can miss weak positives • requires multiple comparison corrections if using p-values
Median + or - k MAD	Hit with increased activity = any sample whose value is >= sample median + k MADs; Hit with decreased activity = any sample whose value is <= sample median – k MADs	• easy to calculate • can identify weaker hits • not very sensitive to outliers	• not easily linked to hit p-values
Multiple t-tests	Hit = any reagent for which t-test result between samples at two conditions < threshold (usually p = 0.05 or p = 0.01)	• easy to calculate • provides hit p-values	• requires triplicates, at minimum • sensitive to outliers • inappropriate if data is not normally distributed • requires multiple comparison corrections of p-values
Quartile-based	Hit with increased activity = any sample whose value is > Q3 + c IQR Hit with decreased activity = any sample whose value is < Q1 - c IQR (where c is a threshold constant)	• easy to calculate • can identify weaker hits • not sensitive to outliers • good for non-symmetrical data distributions	• limited additional power over median + or - k MAD for approximately normal data • not easily linked to hit p-values • not available in most analysis software
SSMD	Appropriate equations depend on whether goal is control of rates of false negatives, false positives or both; see publication	• allows control of both false positive and false negative rate • linked to rigorous probability interpretation	• not available in most analysis software • not intuitive for many biologists
RSA	Iterative ranking algorithm that cannot be reduced to a single equation; see publication	• can identify weaker hits • not sensitive to outliers • provides hit p-values • may help reduce false-positives due to off-target effects of single reagents	• difficult to calculate • may have limited utility for pool-based screens
Rank-product	Iterative ranking algorithm that cannot be reduced to a single equation; see publication	• can identify weaker hits • not sensitive to outliers • provides hit p-values	• difficult to calculate • requires many replicates
Bayesian	Appropriate equations depend on whether negative-control or activation-inhibition-negative control model is applied; see publication	• not sensitive to outliers • provides hit p-values • allows direct calculation of false discovery rate • includes both experiment-wide and plate-wide information • uses both negative controls and samples	• difficult to calculate • not intuitive for many biologists

Nature Methods: doi: 10.1038/nmeth.1351

SD, standard deviation; h.v., high-value; l.v., low-value; MAD, median absolute deviation; SSMD, Strictly Standardized Mean Difference; ROC, receiver operator characteristic; IQR, interquartile range; RSA, Redundant siRNA Activity

Nature Methods: doi: 10.1038/nmeth.1351