Statistical methods for analysis of high-throughput RNA interference screens

Amanda Birmingham[1,2], Laura M Selfors[1,3], Thorsten Forster[1,4], David Wrobel[1,3], Caleb J Kennedy[1,5], Emma Shanks[1,6], Javier Santoyo-Lopez[1,4], Dara J Dunican[1,7], Aideen Long[1,7], Dermot Kelleher[1,7], Queta Smith[1,2], Roderick L Beijersbergen[1,8], Peter Ghazal[1,4] & Caroline E Shamu[1,3,5]

RNA interference (RNAi) has become a powerful technique for reverse genetics and drug discovery, and in both of these areas large-scale high-throughput RNAi screens are commonly performed. The statistical techniques used to analyze these screens are frequently borrowed directly from small-molecule screening; however, small-molecule and RNAi data characteristics differ in meaningful ways. We examine the similarities and differences between RNAi and small-molecule screens, highlighting particular characteristics of RNAi screen data that must be addressed during analysis. Additionally, we provide guidance on selection of analysis techniques in the context of a sample workflow.

In recent years, large-scale RNA interference (RNAi) libraries that are designed to target complete genomes have been produced for multiple organisms. When arrayed in 96- or 384-well microplates, with each well containing reagents that are, in theory, directed to only one gene, genome-scale RNAi screens are possible using high-throughput screening (HTS) technologies that have been developed for small-molecule screening[1]. Here we present a sample workflow for analyzing assay-endpoint readouts from high-throughput RNAi screening data with guidance on choice and application of relevant statistical methods. Most of the work discussed in this review focuses on RNAi screens using synthetic small interfering RNA (siRNA), but the methods are also applicable to RNAi screens performed in microplates with other silencing reagents, including short hairpin RNA (shRNA), endoribonuclease-prepared siRNA (esiRNAs) and double-stranded RNA.

Comparison of RNAi and small-molecule screens

In spite of their dependence on techniques developed for small-molecule screens, many groups have observed anecdotally that there are nontrivial differences between RNAi and small-molecule HTS data. To more fully characterize such possible differences, we compared existing datasets from more than 19 optimized siRNA screens and more than 13 mammalian cell–based small-molecule screens carried out at the ICCB-Longwood Screening Facility at Harvard Medical School (**Supplementary Table 1**). Lilliefors tests on the distribution of measured values across plates containing different siRNA or small-molecule reagents showed that more siRNA screening data were normally distributed. In addition, we investigated three simple measures of assay robustness: signal-to-background ratio, coefficient of variation, and Z' factor (see below). The median signal-to-background ratio, comparing plate mean positive control signal to plate mean negative control, was about twofold lower for the siRNA screens compared to the cell-based small-molecule screens. The median coefficient of variation for siRNA assays was twice that for the small-molecule assays (26.5% versus 13.4%, respectively). Consistent with the decreased signal-to-background ratio and increased variability of siRNA screens, the Z' factors from siRNA screens tended to be low (frequently between 0.0 and 0.5). Thus, overall we

[1]The RNAi Global Initiative, Lafayette, Colorado, USA. [2]Thermo Fisher Scientific, Lafayette, Colorado, USA. [3]ICCB-Longwood Screening Facility, Harvard Medical School, Boston, Massachusetts, USA. [4]Division of Pathway Medicine and Centre for Systems Biology, The University of Edinburgh, Edinburgh, UK. [5]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. [6]Drug Discovery Unit, Division of Biological Chemistry and Drug Discovery, College of Life Sciences, University of Dundee, Dundee, Scotland. [7]Institute of Molecular Medicine, Trinity Centre for Health Sciences, Trinity College, Dublin, Ireland. [8]Division of Molecular Carcinogenesis and National Cancer Institute Robotics and Screening Center, The Netherlands Cancer Institute, Amsterdam, The Netherlands. Correspondence should be addressed to A.B. (amanda.birmingham@thermofisher.com).

PUBLISHED ONLINE 30 JULY 2009; DOI:10.1038/NMETH.1351

found that the siRNA screens were less robust by these standard measures than small-molecule screens. Several plausible reasons for these differences exist.

Library design. In theory, most siRNA reagents in a genome-scale library should have an expressed cellular target and knockdown of each target may have physiological effects; RNAi reagents thus have a higher intrinsic probability for impacting the overall network biology. In contrast, most compounds in small-molecule screening libraries have no effect on cell pathways because they lack cellular targets and are *de facto* negative controls.

Transfection. Even under well-controlled conditions, the transfection process, especially transfection efficiency, is a major source of variability for siRNA screens. Transfection also causes cell stress and can affect cell viability, which may have variable and indirect phenotypic effects in cellular assays.

Kinetics and mechanism of RNAi action. Whereas small molecules rarely affect the actual abundance of the target protein itself, RNAi reagents reduce—and in some cases nearly eliminate—the target gene product in the cell[2]. Because of this mechanistic difference, RNAi reagents generally require 48–72 hours for maximal effect, whereas small molecules can directly affect their protein targets within hours. This increased time between cell plating and assay endpoints leads to greater impact of cell culture and environmental variation on phenotypes and more assay variability in RNAi screens. In addition, the 'window' of maximal RNAi effect likely varies for each gene, but typically a single endpoint is chosen for a screening assay[2]. Assaying too early for a given gene produces a false negative, whereas assaying too late may lead to false positives because of downstream effects, thus contributing to weaker, more variable phenotypes in RNAi screens.

Number and quality of controls. RNAi positive controls generally have weaker effects than small-molecule positive controls and, owing to variations in transfection efficiencies, may exhibit more inter-well variability than small-molecule controls. Furthermore, although small-molecule screens use vehicle-only wells as negative controls, no such universal negative control exists for siRNA screens because RNAi screens usually involve complex delivery vehicles that alone can have biological effects, and even 'nontargeting' siRNA controls may exhibit off-target effects in some cell lines[3]. Sequence-specific off-target effects of RNAi[4–6] resulting from partial sequence complementarity between siRNA and mRNA complicate interpretation of RNAi experiments. The mechanisms by which the off-target effects occur are not yet completely understood, and thus they cannot be fully eliminated experimentally.

As many judgment calls are required in the selection of HTS data analysis strategies, RNAi researchers should be aware of these differences to make appropriate analysis choices, as discussed below.

RNAi screening analysis workflow

Data analysis takes much longer than most screening groups have anticipated—sometimes longer than the wet-lab screen itself[7,8]. Below we present a sample workflow for analyzing RNAi HTS data from individual assays with quantitative output. We provide guidance on choice and application of relevant statistical methods (also summarized in **Table 1** and in **Supplementary Table 2**). Researchers should keep in mind that, in most situations, there is no single 'correct' analysis method for any dataset. Successful data analysis depends critically on careful experimental design and assay development before the primary screen. Plate layout and placement of control reagents is important because RNAi screens are particularly susceptible to edge effects or drifting values across a plate; several suggested layouts have been published[9–11]. It is advisable to randomize reagent plating location to avoid systematic effects of grouped similar reagents (although

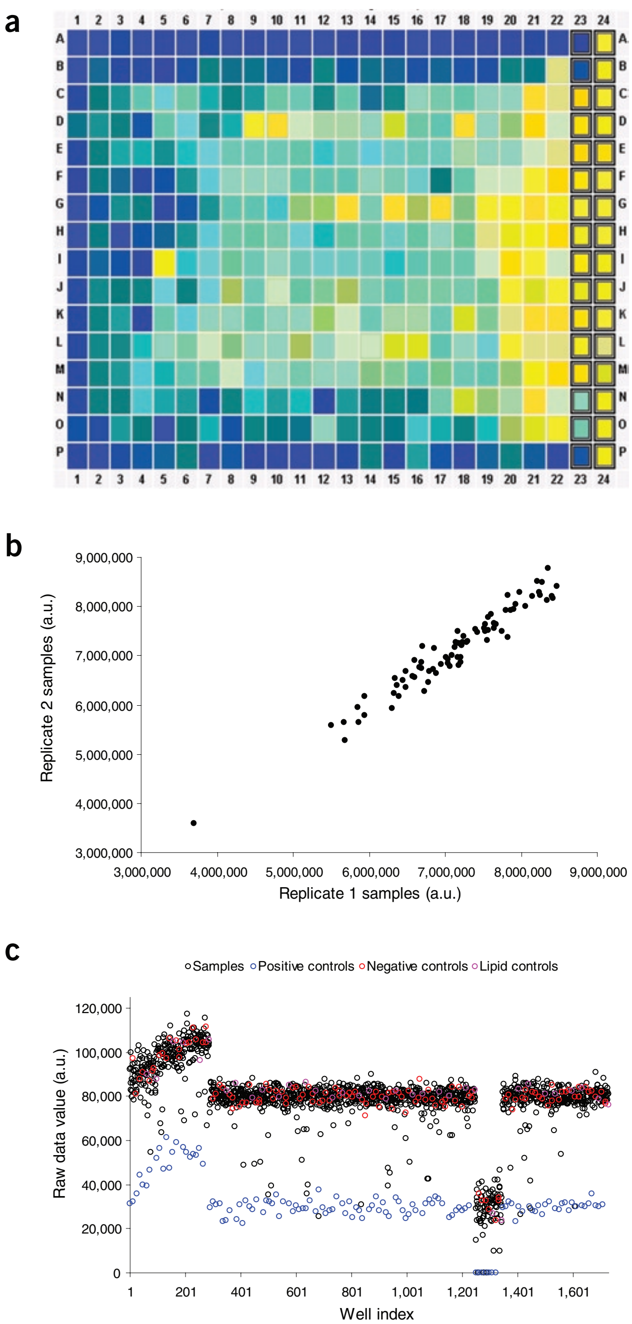

Figure 1 | Methods for visualizing data outputs of a screen. (**a**) Heat map of raw values from a hypothetical 384-well plate with strong edge effects. (**b**) Replicate correlation plots of raw values from two hypothetical replicates of the same master plate showing good agreement and suggesting overall good reproducibility. (**c**) Plate-well scatter plot of raw values from a hypothetical screen of eighteen 96-well plates including evidence of drift in the first three plates and a plate with unusually low values near the end of the screen.

BOX 1 CONTROL-BASED VERSUS SAMPLE-BASED NORMALIZATION

Some normalization techniques compare individual experimental sample values to aggregated values of controls, whereas others use the samples themselves as *de facto* negative controls (and some though not all techniques can accommodate either approach). The use of samples as *de facto* negative controls on an assay plane can provide more accurate measurements because there are usually more experimental samples than controls on an assay plate from which to gather data. Additionally, some assays lack good negative controls that work effectively across all plates. In such cases, the best available choice is to use the majority of sample wells as a negative reference. However, this approach is based on the assumption that most samples display no biological effect in the assay being analyzed; this assumption needs to be explicitly assessed by RNAi screeners for their particular assay, especially when using screening libraries in which reagents targeting structurally or functionally related genes are plated together or when using a very general assay such as cell viability in which many RNAi reagents may actually have a biological effect. Clearly, this assumption will also be violated in the case of confirmatory screens in which many primary screening positive reagents are evaluated on the same assay plate. In these cases, screeners should consider including a large number of each type of control[22] on each assay plate and using control-based statistics instead.

The use of samples as *de facto* negative controls is especially problematic for nonrobust normalization methods that use statistical measures (such as mean and standard deviation) that are strongly sensitive to outliers in the data. Such methods may be strongly affected in RNAi screens containing either large numbers of hits or very strong hits (which will constitute 'outliers' in the sample data).

not all researchers find this technically feasible). Additionally, siRNA screens of whole genomes can be carried out in <30,000 wells and thus can be performed routinely in duplicate or higher replicate numbers to decrease both false positive and false negative rates[12].

Step 1: data triage

To ensure a successful high-throughput screen, data should be examined while the screen is in progress to allow the identification of potential problems as they occur. Plate visualization is one of the most effective techniques for uncovering undesirable patterns that might indicate technical problems (for example, extreme edge effects or a clogged plate filler manifold). Library plates that have much higher hit rates than others can also be uncovered by data visualization. Such an observation would suggest that targeting reagents are nonrandomly plated in the library and that sample-based normalization strategies are inappropriate (**Box 1**). Visualization formats developed for small-molecule screens[13], including heat maps and plate-well scatter plots to display overall screen performance as well as replicate correlation plots to visualize overall reproducibility (**Fig. 1**), are equally helpful for RNAi screens. Calculation of various quality metrics (discussed below) on plates as they are screened can also help ensure the production of usable data.

Step 2: normalization

Normalization is a process intended to remove systematic errors from the data and to allow comparison and combination of data from different plates in the screen. Normalization is generally performed per plate, but is occasionally performed per experiment when there is a known confounding per-plate bias such as a large number of true hits on a particular plate (which may occur when plating of library reagents has been nonrandom—for example, when screening a transcription factor siRNA library for genes involved in stem cell differentiation).

Many normalization methods have been developed (see ref. 9 for an overview). An important issue in this area is the use of on-plate controls versus samples for normalizing (**Box 1**). Most RNAi screeners will find that one of the following methods is suitable to their data and analysis capacities and will probably receive acceptable results with any reasonable choice. As evidence of this, Wiles *et al.*[14] investigated the effects of seven normalization techniques (six of which had not previously been applied to RNAi screens) on validation rates of hits selected for secondary screening from a large-scale *Drosophila* RNAi screen and concluded that no single method excelled.

Fraction or percent of control. One common approach, often favored by biologists for its easy calculation and interpretability, is division of each sample value by the mean of the control of interest (either negative or positive). This method requires a large number of controls to provide adequate estimation of their mean. Additionally, this method is adequate only for tightly distributed and 'well-behaved' data as it incorporates no information on the variation of the sample measurements and is sensitive to outliers in the controls.

Fraction or percent of samples. In another common normalization approach, the mean of the samples on the plate may be substituted for the mean of negative controls in the percent-of-control calculation, which reduces the need for large numbers of controls but may exacerbate the issue of nonrobustness (**Box 1**). A more robust variation on this method that is adequate for many screens is division of each sample value by the median value for all samples on the plate, although this method still cannot incorporate information on the degree of variation in the sample data.

***z* score and robust *z* score.** *z* score (the number of standard deviations from the mean) is frequently used to normalize data in a way that provides explicit information on the strength of each siRNA relative to the rest of the sample distribution. An advantage of *z* score is its incorporation of information on the variation in sample measurements, but this method also depends on the use of samples as *de facto* negative controls. Because *z* score is sensitive to outliers, a variation known as the robust *z* score, which substitutes the outlier-insensitive median and median absolute deviation for mean and standard deviation in the *z*-score calculation, is generally considered preferable for RNAi screens.

***B* score.** The above techniques, as most normalization methods, can account for cross-plate systematic effects but not for within-plate systematic effects. If data triage has identified the presence of within-plate systematic effects, the B score normalization[15] may be applied to remove row, column or well effects by iterative application of the Tukey median polish algorithm. B score is relatively robust to outliers; however, because its calculation examines the entire plate to identify systematic effects, it essentially uses the samples as negative controls.

The first three approaches can be easily calculated using standard spreadsheet formulas, but the B score cannot because it is based on an iterative algorithm. However, a version of the B score is available in the cellHTS2 package[16] of the open-source BioConductor bioinformatics software.

Step 3: calculation of quality metrics

Before proceeding to intensive analysis, a screener must ensure that the resulting data meet the minimum standards of quality to permit legitimate conclusions.

***Z′* or *Z* factor**. By far the most common quality metrics reported for both RNAi screens and small-molecule screens are Z and Z' factor[17]. Z' factor is often used during assay optimization because it is based on controls, whereas Z factor may be used during screening to assess performance of the screen on actual samples. Note that the Z' and Z factor should not be confused with the z score discussed above.

$$Z' \text{ factor} = 1 - (3\sigma_{hc} + 3\sigma_{lc}) / |\mu_{hc} - \mu_{lc}|$$
$$Z \text{ factor} = 1 - (3\sigma_{s} + 3\sigma_{c}) / |\mu_{s} - \mu_{c}|$$

where μ indicates mean, σ indicates standard deviation, "hc" indicates the high-value control, "lc" indicates the low-value control, "s" indicates sample value, and "c" indicates negative control. The range of both measures is negative infinity to 1, with > 0.5 as a very good assay, > 0 an acceptable assay and < 0 an unacceptable assay.

A potential issue in using the Z' factor as a measure of assay resolution is that it is possible to generate a high Z' factor using a very strong positive control, which may not realistically represent more moderate screening positives. This issue is of special concern for RNAi screens, in which weak effects might be biologically meaningful and in which the signal-to-background ratio can be of lower magnitude than in small-molecule screens (**Supplementary Table 1**). Thus, researchers are advised whenever possible to use positive controls that are similar in strength to the hits they anticipate finding. It may also be necessary to adjust Z'-factor quality guidelines for RNAi screens; we have found that assays with Z' factors of zero or greater have been successful in identifying validated hits when we screened library plates in duplicate or triplicate.

Strictly standardized mean difference. Researchers desiring a more statistically rigorous means to address the limitations of moderate controls are advised to adopt the strictly standardized mean difference (SSMD). SSMD, which was developed for use with RNAi screening, is the ratio between the difference of the means and the standard deviation of the difference between two populations, 1 and 2 (such as positive and negative controls)[18]. For independent populations

$$\text{SSMD} = (\mu_1 - \mu_2) / \sqrt{(\sigma_1^2 + \sigma_2^2)}$$

Acceptable values for SSMD depend on the strength of the positive controls used, but SSMD has been shown to be an accurate, less conservative indicator of quality than Z' or Z factor, as demonstrated by examples in which signal-to-noise ratio and Z-factor quality metrics produce misleading results but SSMD correctly 'passes' the data in question[18]. Indeed, we have incorporated the use of SSMD into quality assessment workflows and found it more informative in the context of RNAi screening than the Z' factor in some cases. In one case, samples were assayed 20 hours after known maximal response to identify long-term repressors; this extended timeline caused increased variability in high controls that decreased the Z' factor below standard acceptable thresholds for the majority of assay plates. SSMD, in contrast, accurately captured the clear difference between the high and low populations.

Receiver operating characteristic curves. The receiver operating characteristic (ROC) curve[19] has been used as a quality metric in microarray transcriptomics[20,21], and as it provides a quick and intuitive understanding of dynamic ranges in data given positive and negative controls, we here highlight it for use in RNAi screening. ROC curves plot sensitivity versus (1 – specificity) (see **Supplementary Table 2** for definitions). The area under the ROC curve can be used as a quality metric: an area of 1 represents a perfect predictor and area of 0.5 represents performance as bad as random chance. One advantage of using ROC curves is that multiple thresholds for defining positives and the resulting trade-offs between sensitivity and specificity can easily be investigated by plotting multiple ROC curves. This method has been used to compare validation performance of hits generated from differently normalized RNAi data[14].

Step 4: hit identification

The identification of 'hits' or 'screening positives' is the goal of any primary RNAi screen, and yet remains a point of considerable contention in data analysis. Hit identification is, essentially, the process of deciding which sample values differ meaningfully from those of the negative controls. Although some screeners simply select a discrete number of top-scoring samples as screening positives (often as determined by follow-up capacity), many hit identification techniques are available. The selected hit list forms the basis for subsequent validation screens or investigations.

To reduce the risk of false positives, many practitioners recommend screening multiple reagents targeting the same gene of interest and selecting hits based on the combined results[6]; generally, genes are chosen as hits when a majority of tested reagents are screening positives, although the 'redundant siRNA activity' technique described below offers a more rigorous approach to combining the results of multiple reagents. False positives may also be limited by combining information from multiple screening outputs[22], an approach that has become particularly viable with the advent of high-content screening. Although the techniques of multiparametric analysis are complex and beyond the scope of this work, a useful overview is available in ref. 23. Such approaches may identify real hits that have high variability in a single readout metric[22].

Below we discuss the features of both small molecule–derived techniques (mean ± threshold (k) standard deviation, median ± k median absolute deviation and multiple t-tests) and RNAi techniques (quartile-based selection, SSMD for hit identification, redundant siRNA activity, rank product and Bayesian models).

Mean ± *k* standard deviation. This approach, which involves selecting a standard deviation threshold (k) of the normalized data relative to the mean and identifying positives as samples that surpass this threshold, is by far the most frequently used hit identification technique in RNAi screening literature (for example, refs. 24,25). It is often used with z-score normalization, but is sometimes used on data that has been normalized by other approaches (such as the B score). This method is particularly appropriate for normally distributed data because the standard deviation from the mean links to an estimate of the probability that hit values are significantly different than the distribution of values for *de facto* negative controls. Another advantage is that this method is very easy to calculate and implement; however, it is not robust to outliers. Thus, especially for data in which outliers appear frequently, the application of the commonly used 3 standard deviation cut-off with this approach tends to miss weak hits, whereas lowering the standard deviation threshold to capture such hits may unacceptably increase the rate of false positives[26,27].

Median ± *k* median absolute deviation. An improvement on the mean ± k standard deviation approach is median ± k median absolute deviation (for example, ref. 28). This method is robust to outliers and has been shown to identify weak hits in RNAi data more effectively than mean ± k standard deviation while still capturing the strong hits and controlling false positives[27]. It has also been shown to generate fewer false negatives than mean ± k standard deviation when applied to a nonnormal data distribution[27] and is very easy to calculate and implement (although it sacrifices the former method's easy link to a probability distribution). For these reasons, Chung *et al.*[27] recommend it as the first-choice approach for hit selection in RNAi screens.

Multiple t-tests. For certain assays, such as those comparing RNAi treatment in the presence of drug versus RNAi treatment alone, it may be appropriate to assess the difference in means between replicates for each condition with multiple t-tests (for example, ref. 29). This approach is simple to implement and understand, but it requires three or more replicates of each condition and assumes normality of the replicate data. In addition, it is imperative to apply multiple-comparison corrections to the resulting P-values of each individual test if a high false positive rate cannot be tolerated[30], and results of such t-tests are sensitive to outliers[31].

Table 1 | Statistical analysis methods for hit identification

Strategy	Formula	Advantages	Disadvantages
Mean ± k standard deviation (s.d.)	Hit with increased activity is any sample whose value is ≥ sample mean + k s.d. Hit with decreased activity is any sample whose value is ≤ sample mean – k s.d.	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 ± k median absolute deviation (MAD)	Hit with increased activity is any sample whose value is ≥ sample median + k MAD Hit with decreased activity is any sample whose value is ≤ sample median – k MAD	Easy to calculate Can identify weaker hits Not very sensitive to outliers	Not easily linked to hit P-values
Multiple t-tests	Hit is any reagent for which t-test between samples at two conditions is less than a threshold (usually P = 0.05 or P = 0.01)	Simple 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 is any sample whose value is > third quartile value + c interquartile range, where c is a threshold constant Hit with decreased activity is any sample whose value is < first quartile value – c interquartile range	Easy to calculate Can identify weaker hits Not sensitive to outliers Good for nonsymmetrical data distributions	Limited additional power over median ± 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	Allows control of both false positive and false negative rate Not dependent on sample size 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	Can identify weaker hits Not sensitive to outliers Provides hit P-values May reduce false positives owing 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	Provides hit P-values Can identify weaker hits Not sensitive to outliers	Difficult to calculate Requires many replicates
Bayesian	Appropriate equations depend on whether negative-control or activation-inhibition-negative control model is applied	Provides hit P-values Not sensitive to outliers 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

Quartile-based selection. Researchers who determine that their data distribution is not symmetrical may wish to use the quartile-based hit identification method. This approach sets upper and lower hit selection thresholds based on number of interquartile ranges above or below the first and third quartiles of the data. Like median ± k median absolute deviation, the quartile method has been shown to identify both strong and weak hits while controlling false positives[26]. Although this method is easy to calculate, it has not been generally implemented in the RNAi screening community, perhaps because of its modest improvement over the more common median ± k median absolute deviation approach on approximately normal data. Additionally, in quartile-based selection, as in many other robust methods, the rankings produced are not easily translatable into P-values.

SSMD for hit identification. The SSMD metric discussed earlier can be used for hit identification by screeners concerned with controlling the rate at which siRNAs that have real large or moderate effects fail to be identified as screening positives as well as the rate at which siRNAs that should be considered negative are identified as screening positives[32]. Formulas are available[33,34] for calculating the SSMD limits for hit selection based on the desired false positive rate, false negative rate or both; although these require many negative controls (>50), follow-up work[31] suggested SSMD cut-offs for screens without large numbers of negative samples, such as confirmatory screens.

Although the SSMD metric has linear relationship to z score when only one replicate per siRNA is measured in a screen, these statistically based guidelines may make SSMD more meaningfully interpretable to researchers. Currently SSMD-based hit identification is not calculated by standard analysis packages and is not trivial to implement from scratch.

Redundant siRNA activity. The redundant siRNA activity (RSA) analysis method[35] is appropriate for researchers seeking to integrate information about multiple RNAi reagents tested for each gene. RSA ranks silencing reagents according to experimental effect and assigns a P-value to all reagents for a single gene based on whether the reagents for that gene are distributed significantly higher in the rankings than would be expected by chance. Because of its use of chance performance as a basis for statistical calculations, RSA is able to provide P-values for gene hits without sacrificing robustness.

'Positive' reagents identified by this method have been found to have higher rates of reconfirmation than those identified with conventional methods, with discrepancies attributable to low reproducibility of orphan individual siRNAs with high activities[35]. Although RSA is not currently included in common analysis software packages, its developers have made available implementations in C# (for Windows), R and Perl (see http://carrier.gnf.org/publications/RSA/).

Rank product. Screeners intending to perform screens in biological replicate and seeking a robust hit identification approach that provides estimated P-values may also wish to consider the rank-product method, originally developed for use with microarray data[36]. The premise of the rank-product approach is that a consistent hit should be highly ranked in each independent biological replicate set. The rank-product statistic for each sample across all independent sets estimates this consistency; it can then be translated into a measure of statistical significance by comparing the observed rank product statistic to a rank product statistic obtained from a large number of simulated datasets (providing the statistic expected by chance).

This approach provides P-values for potential hits without requiring the assumption of an underlying probability distribution, but does require substantial computation and several replicates per screen to work. Although similar to RSA in its comparison of true data rankings to those produced by chance, it does not depend on the use of multiple different RNAi reagents per gene. A rank-product implementation suited for use with RNAi screening data has recently been made available as part of the RNAither package[37] in the BioConductor open-source bioinformatics software.

Bayesian models. Screeners with appropriate computational resources who seek explicit estimated probabilities that a given siRNA has no effect, an inhibition effect or an activation effect (rather than the single score produced by other methods) may wish to use a Bayesian approach described recently[38]. Bayesian statistics use Bayes' theorem to calculate the probability that a particular hypothesis is true given the observed evidence and offer a means to update these probabilities when additional evidence is collected. Zhang *et al.*[38] identify three hypotheses of interest (an siRNA has no effect, an siRNA has an activation effect or an siRNA has an inhibition effect) and develop two models to describe the posterior probability that each of these hypotheses are true for a particular sample given the evidence of this sample's observed value. The first, and simpler, model is based on using only the negative controls to describe the posterior distribution of the true mean value for the sample given the observed data value. The second, more complex model describes a posterior distribution that assumes the availability of data from both positive-inhibition and positive-activation controls as well as negative controls. Both models also provide the means to calculate the false discovery rate associated with any given hit threshold, but are usable only on screens without replicates.

A strength of this approach is that it incorporates both plate-wide and experiment-wide information as well as (depending on the model used) information from both negative controls and the assumed *de facto* negative samples. Comparing several hit identification approaches, Zhang *et al.*[38] found the simpler Bayesian model to perform best, followed by plate-wise median ± k median absolute deviation. Unfortunately, although the published Bayesian models show great promise, they have not yet been incorporated into commonly available analysis software and are not trivial to implement. Until software applications implementing Bayesian modeling are available, the plate-wise median ± k median absolute deviation approach may be the best alternative.

Conclusions and future directions

High-throughput RNAi offers a wealth of exciting new data but also provides new challenges in the statistical interpretation of these data. When considering appropriate statistical techniques for quality control and hit selection in HTS, it is important to keep in mind that statistical analysis methods developed in industry for small-molecule screening and therapeutic lead identification are optimized for a very low false positive rate. RNAi screens, however, appear inherently more variable and often produce a broader range of biological hits. Thus, researchers applying small-molecule techniques to RNAi data should have a clear understanding of the implicit assumptions and limitations, and possibly apply altered thresholds to account for common RNAi-specific data features. Several methods for RNAi HTS data analysis have been developed specifically with these issues in mind.

The statistical methods outlined here, while certainly not comprehensive, offer effective, practical tools for RNAi HTS analysis. However, additional work is needed; many analysis techniques continue to favor conservative limits on false positive rates, which artificially reduce the number of weak and/or variable yet biologically important hits found in screens. In addition, many researchers are investigating the use of microscopy-based high-content assays, which may eventually offer even more statistical power by providing multiple readouts and finer characterization of phenotypic results. However, statistical techniques and software for these assays are still in their infancy; no doubt this area will see considerable development effort in the coming years.

Finally, it should be emphasized that rigorous secondary screening is necessary to validate primary screen findings[6]. Even the best statistics cannot establish the biological importance of genes identified as primary hits.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

Work at ICCB-Longwood was funded by US National Institutes of Health grants CA078048, AI067751 and AI057159. Funding to P.G. was provided by the Wellcome Trust and the Biotechnology and Biological Sciences Research Council (BBSRC). The Centre for Systems Biology at Edinburgh is a Centre for Integrative Systems Biology (CISB) funded by BBSRC and the Engineering and Physical Sciences Research Council (EPSRC), reference BB/D019621/1. Work at the University of Edinburgh was supported by edikt2 (Scottish Funding Council grant HR04019; http://www.edikt.org/). R.L.B. is supported by Shering-Plough and TIPharma. D.J.D., A.L. and D.K. are supported by Marie Curie MTKD-CT-2005-029798 (European Union FP6).

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Boutros, M. & Ahringer, J. The art and design of genetic screens: RNA interference. *Nat. Rev. Genet.* **9**, 554–566 (2008).
2. Perrimon, N. & Mathey-Prevot, B. Applications of high-throughput RNA interference screens to problems in cell and developmental biology. *Genetics* **175**, 7–16 (2007).
3. Jackson, A.L. *et al.* Expression profiling reveals off-target gene regulation by RNAi. *Nat. Biotechnol.* **21**, 635–637 (2003).
4. Birmingham, A. *et al.* 3′ UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nat. Methods* **3**, 199–204 (2006).
5. Jackson, A.L. *et al.* Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity. *RNA* **12**, 1179–1187 (2006).
6. Echeverri, C.J. *et al.* Minimizing the risk of reporting false positives in large-scale RNAi screens. *Nat. Methods* **3**, 777–779 (2006).
7. Chatterjee-Kishore, M. From genome to phenome–RNAi library screening and hit characterization using signaling pathway analysis. *Curr. Opin. Drug Discov. Devel.* **9**, 231–239 (2006).
8. Ramadan, N., Flockhart, I., Booker, M., Perrimon, N. & Mathey-Prevot, B. Design and implementation of high-throughput RNAi screens in cultured *Drosophila* cells. *Nat. Protoc.* **2**, 2245–2264 (2007).
9. Malo, N., Hanley, J.A., Cerquozzi, S., Pelletier, J. & Nadon, R. Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.* **24**, 167–175 (2006).
A comprehensive but accessible introduction to the small-molecule screening analysis techniques, which form the starting point for most RNAi screen analyses.
10. Zhang, X.H.D. Novel analytic criteria and effective plate designs for quality control in genome-scale RNAi screens. *J. Biomol. Screen.* **13**, 363–377 (2008).
This article describes the use of SSMD for quality assessment and suggests appropriate thresholds, as well as provides an in-depth discussion of plate layout issues.
11. Zhang, X.H.D. *et al.* Integrating experimental and analytic approaches to improve data quality in genome-wide RNAi screens. *J. Biomol. Screen.* **13**, 378–389 (2008).
12. Zhang, X.H.D. & Heyse, J.F. Determination of sample size in genome-scale RNAi screens. *Bioinformatics* **25**, 841–844 (2009).
13. Gunter, B., Brideau, C., Pikounis, B. & Liaw, A. Statistical and graphical methods for quality control determination of high-throughput screening data. *J. Biomol. Screen.* **8**, 624–633 (2003).
14. Wiles, A.M., Ravi, D., Bhavani, S. & Bishop, A.J. An analysis of normalization methods for *Drosophila* RNAi genomic screens and development of a robust validation scheme. *J. Biomol. Screen.* **13**, 777–784 (2008).
15. Brideau, C., Gunter, B., Pikounis, B. & Liaw, A. Improved statistical methods for hit selection in high-throughput screening. *J. Biomol. Screen.* **8**, 634–647 (2003).
16. Boutros, M., Brás, L.P. & Huber, W. Analysis of cell-based RNAi screens. *Genome Biol.* **7**, R66 (2006).
17. Zhang J.H., Chung T.D. & Oldenburg K.R. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screen. B*, **4** 67–73 (1999).
18. Zhang, X.H.D. A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays. *Genomics* **89**, 552–561 (2007).
19. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874 (2006).
20. Wagner, E.K. *et al.* Practical approaches to long oligonucleotide-based DNA microarray: lessons from herpesviruses. *Prog. Nucleic Acid Res. Mol. Biol.* **71**, 445–491 (2002).
21. Forster, T., Roy, D. & Ghazal, P. Experiments using microarray technology: limitations and standard operating procedures. *J. Endocrinol.* **178**, 195–204 (2003).
22. Stone, D.J. *et al.* High-throughput screening by RNA interference: control of two distinct types of variance. *Cell Cycle* **6**, 898–901 (2007).
23. Ainscow, E. Statistical techniques for handling high content screening data. *European Pharmaceutical Review* **5**, 30–38 (2007).
24. Bard, F. *et al.* Functional genomics reveals genes involved in protein secretion and Golgi organization. *Nature* **439**, 604–607 (2006).
25. DasGupta, R., Kaykas, A., Moon, R.T. & Perrimon, N. Functional genomic analysis of the Wnt-wingless signaling pathway. *Science* **308**, 826–833 (2005).
26. Zhang, X.H.D. *et al.* Robust statistical methods for hit selection in RNA interference high-throughput screening experiments. *Pharmacogenomics* **7**, 299–309 (2006).
27. Chung, N. *et al.* Median absolute deviation to improve hit selection for genome-scale RNAi screens. *J. Biomol. Screen.* **13**, 149–158 (2008).
This article describes experimental and computational studies validating the effectiveness of median ± *k* median absolute deviations as a hit identification approach for RNAi screens.
28. Müller, P., Kuttenkeuler, D., Gesellchen, V., Zeidler, M.P. & Boutros, M. Identification of JAK/STAT signalling components by genome-wide RNA interference. *Nature* **436**, 871–875 (2005).
29. Whitehurst, A.W. *et al.* Synthetic lethal screen identification of chemosensitizer loci in cancer cells. *Nature* **446**, 815–819 (2007).
30. Manly, K.F., Nettleton, D. & Hwang, J.T. Genomics, prior probability, and statistical tests of multiple hypotheses. *Genome Res.* **14**, 997–1001 (2004).
31. Zhang, X.H.D. Genome-wide screens for effective siRNAs through assessing the size of siRNA effects. *BMC Res. Notes* **1**, 33 (2008).
32. Zhang, X.H.D., Marine, S.D. & Ferrer, M. Error rates and powers in genome-scale RNAi screens. *J. Biomol. Screen.* **14**, 230–238 (2009).
33. Zhang, X.H.D. A new method with flexible and balanced control of false negatives and false positives for hit selection in RNA interference high-throughput screening assays. *J. Biomol. Screen.* **12**, 645–655 (2007).
34. Zhang, X.H.D. *et al.* The use of strictly standardized mean difference for hit selection in primary RNA interference high-throughput screening experiments. *J. Biomol. Screen.* **12**, 497–509 (2007).
35. König, R. *et al.* A probability-based approach for the analysis of large-scale RNAi screens. *Nat. Methods* **4**, 847–849 (2007).
36. Breitling, R., Armengaud, P., Amtmann, A. & Herzyk, P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.* **573**, 83–92 (2004).
37. Rieber, N., Knapp, B., Eils, R. & Kaderali, L. RNAither, an automated pipeline for the statistical analysis of high-throughput RNAi screens. *Bioinformatics* **25**, 678–679 (2009).
38. Zhang, X.H.D. *et al.* Hit selection with false discovery rate control in genome-scale RNAi screens. *Nucleic Acids Res.* **36**, 4667–4679 (2008).