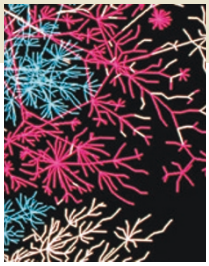

Statistical practice in high-throughput screening data analysis

Nathalie Malo[1,2], James A Hanley[2], Sonia Cerquozzi[1], Jerry Pelletier[3] & Robert Nadon[1,4]

High-throughput screening is an early critical step in drug discovery. Its aim is to screen a large number of diverse chemical compounds to identify candidate 'hits' rapidly and accurately. Few statistical tools are currently available, however, to detect quality hits with a high degree of confidence. We examine statistical aspects of data preprocessing and hit identification for primary screens. We focus on concerns related to positional effects of wells within plates, choice of hit threshold and the importance of minimizing false-positive and false-negative rates. We argue that replicate measurements are needed to verify assumptions of current methods and to suggest data analysis strategies when assumptions are not met. The integration of replicates with robust statistical methods in primary screens will facilitate the discovery of reliable hits, ultimately improving the sensitivity and specificity of the screening process.

High-throughput screening (HTS) is the backbone of drug discovery within the pharmaceutical industry. Over the past decade it has also made its way into academic settings. The combination of robotic methods, parallel processing and miniaturization of biological assays has dramatically increased throughput. The potential to increase the hit discovery rate has been offset, however, by increased research costs. Despite the current popularity of HTS and major improvements in processing, the new drug approval rate has declined significantly[1].

Developers are attempting to counter this inefficiency by various means, including developing biotech-pharmaceutical alliances and changing their internal organizational structures by merging multiple disciplines associated with lead generation and validation[2]. Likewise, HTS programs are being integrated within academic settings where alternative targets and diseases of lesser commercial value can be explored[3]. At the root, the challenge is to find the next marketable drug while simultaneously maximizing the number of screened targets and compounds, minimizing costs per well and optimizing the lead generation and validation process.

Two kinds of inference or decision error can occur at the primary screen step: 'false positives' and 'false negatives'—it is unclear if current inefficiencies are due mostly to the generation of too many false positives, too many false negatives or both. We advance the view that improving hit specificity and sensitivity cannot be met by technological and organizational improvements alone and that improvements in data analysis methods are needed to fulfill the promise of HTS.

[1]McGill University and Genome Quebec Innovation Centre, 740 avenue du Docteur Penfield, Montreal, Quebec, Canada, H3A 1A4. [2]McGill University Department of Epidemiology, Biostatistics, and Occupational Health, 1020 Pine Avenue West, Montreal, Quebec, Canada, H3A 1A4. [3]McGill University Department of Biochemistry, 3655 Promenade Sir William Osler, Montreal, Quebec, Canada, H3A 1A4. [4]McGill University Department of Human Genetics, 1205 avenue du Docteur Penfield N5/13, Montreal, Quebec, Canada, H3A 1B1. Correspondence should be addressed to R.N. (robert.nadon@mcgill.ca).

Published online 7 February 2006; doi:10.1038/nbt1186

HTS is a large-scale process (**Fig. 1**) that screens many thousands of chemical compounds in order to identify potential lead candidates rapidly and accurately. Whereas the plating format and number of compounds per plate can vary, typically just a single measurement of each compound's activity is obtained in an initial primary screen. The

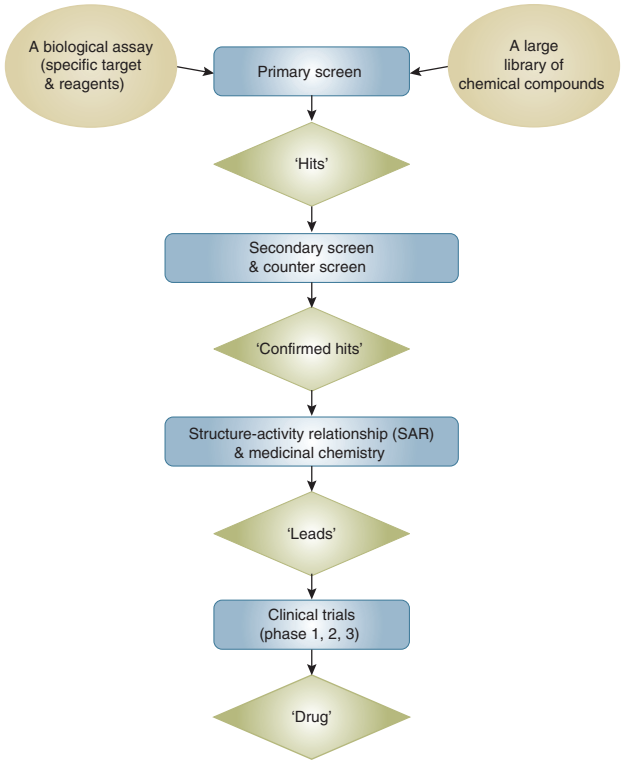

Figure 1 From HTS process to eventual drug development.

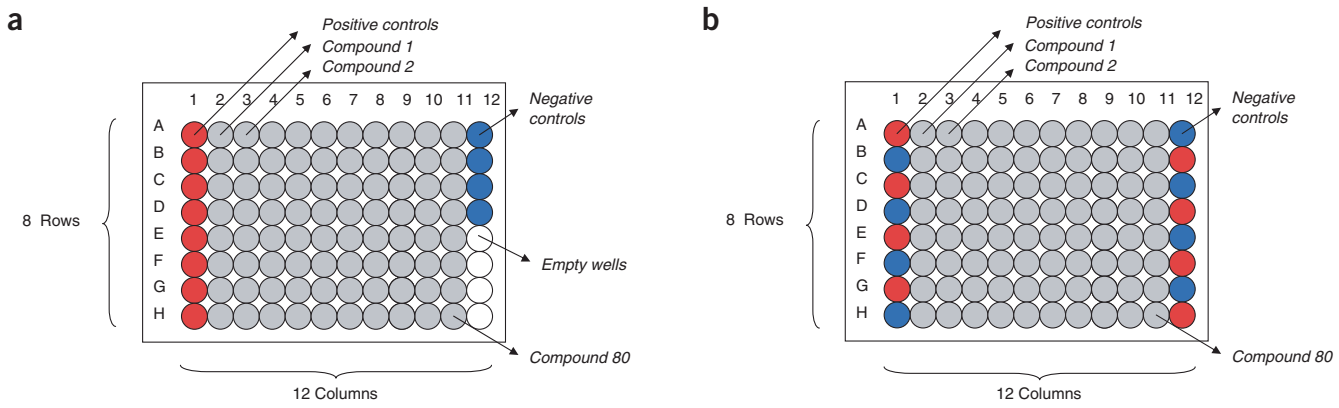

Figure 2 Typical location of controls on a 96-well plate. In a primary screen, the designed biological assay is performed by using a robot to add the target of interest and specific reagents to each well, which already contain a different compound or control. After incubation or other required manipulations, an activity measurement is obtained for every well by automated plate reading. These raw data represent the activity measurement of each compound or control against a specified target. The measurement units and the scales depend on the design of the biological assay, the target of interest and the specific reader or imager that is used. (**a**) Generally, in a compound library, 80 different compounds (gray circles) are stored in the middle of a 96-well plate and wells on the first and last columns are left empty. Often in a high-throughput screen, eight positive controls (red circles) are placed in column 1 and four negative controls (blue circles) are placed in column 12. The other four wells (white circles) in column 12 remain empty and are not used. (**b**) Ideally, controls should be located randomly among the 96 wells of each plate. Only the first and the last columns are typically available for controls, since compounds (gray circles) are stored in the 80 middle wells. Despite this limitation, edge-related bias can be minimized by alternating the eight positive controls (red circles) and the eight negative controls (blue circles) in the available wells, such that they appear equally on each of the eight rows and each of the two available columns.

automated process allows the testing of several hundred plates over a period of weeks. Compounds identified for follow-up (labeled 'hits') are evaluated for biological relevance by a counter screen and confirmed as bona fide hits by a secondary screen.

Secondary screens test many fewer compounds (e.g., the 1% most active compounds from the primary screen[4]) and typically use at least duplicate measurements. Paradoxically, compounds with the highest measured activity levels on a primary screen will on average be less extreme on a secondary screen because of a statistical artifact known as 'regression toward the mean'[5,6]. Accordingly, marginal hits on the first run may fail to validate on the second run merely because of random measurement error, although the size of the statistical artifact can be minimized by improving measurement precision (e.g., by obtaining replicate measurements). Confirmed hits with an established biological activity according to a structure-activity relationship (SAR) series and medicinal chemistry are termed 'leads' that can develop into drug candidates for clinical testing.

Box 1 Formulae for normalization

Percent of control. A qualitative measure of test compound activity defined as

$$POC = \frac{x_i}{\bar{c}} \times 100$$

where x_i is the raw measurement on the i^{th} compound and $\bar{c}$ is the mean of the measurements on the positive controls in an antagonist assay.

Normalized percent inhibition. Another normalization method using controls:

$$NPI = \frac{\bar{c}_+ - x_i}{\bar{c}_+ - \bar{c}_-}$$

where x_i is the raw measurement on the i^{th} compound, $\bar{c}+$ and $\bar{c}-$ are the means of the measurements on the positive and negative controls, respectively, in an antagonist assay.

Z score. A simple and widely known normalizing method calculated as

$$Z = \frac{x_i - \bar{x}}{s_x}$$

where x_i is the raw measurement on the i^{th} compound, $\bar{x}$ and s_x are the mean and the standard deviation, respectively, of all measurements within the plate.

B score[9]. The residual (r_{ijp}) of the measurement for row i and column j on the p^{th} plate is obtained by fitting a two-way median polish and is defined below as

$$r_{ijp} = y_{ijp} - \hat{y}_{ijp} = y_{ijp} - (\hat{\mu} + \hat{R}_{ip} + \hat{C}_{jp})$$

The residual is defined as the difference between the observed result (y_{ijp}) and the fitted value ($\hat{y}_{ijp}$, defined as the estimated average of the plate ($\hat{\mu}_p$) + estimated systematic measurement offset for row i on plate p ($\hat{R}_{ip}$) + estimated systematic measurement column offset for column j on plate p ($\hat{C}_{jp}$)). For each plate p, the adjusted median absolute deviation (MAD_p) is obtained from the r_{ijp}'s (MAD_p). The B score is calculated as follows:

$$Bscore = \frac{r_{ijp}}{MAD_p}$$

Median absolute deviation (MAD). A robust estimate of spread of the r_{ijp}'s values:

$$median\{|r_{ijp} - median(r_{ijp})|\}$$

Inferential errors can be caused by 'noise' due to technical or procedural factors, including assay formats, poor pipette delivery, robotic failures and unintended differences in compound concentrations due to evaporation of solvent, either from the compound collection or during the assay set-up. Errors of unknown origin may also develop over the course of the entire screen. Their adverse effects can often be minimized by quality control procedures, although statistical corrections may also be needed to mitigate the effects of uncontrolled variation (see "HTS data processing" section). Other factors that are less amenable to procedural quality control but that can nonetheless add extraneous variation include potency differences across compounds, and systematic across-plate and within-plate column or row biases (e.g., edge effects).

Differences in variability can also create inequalities among the compounds. The measured activity of low variability compounds will almost always be close to their true levels. Thus, even when measured in singlet, hits are more easily discovered and false hits more easily avoided with these compounds. By contrast, the measured activity levels of highly variable compounds may differ considerably from their true values. It is correspondingly more difficult to discover hits and to avoid false positives.

Once technical and procedural efficiencies have been optimized, the only way to minimize variability further is to obtain estimates of activity levels by taking averages (e.g., mean, median) across replicate measurements. Activity estimates based on repeated measurements are less variable than estimates based on single measurements. Replicate measurements also provide direct estimates of variability, which can be used to estimate the probability of detecting true hits (power analysis), facilitating cost/benefit analyses. Moreover, replicates reduce the number of false negatives without increasing the number of false positives (see "Use of replicates" section).

We review current data preprocessing and hit identification methods for primary screening. We discuss their use and limitations, problems with the constant error assumption, the influence of hit threshold on false-positive and false-negative rates, and factors that can degrade assay sensitivity and specificity. We also discuss the advantages of replicates and make recommendations for the statistical analysis of HTS.

HTS data processing

A well-defined and highly sensitive test system requires both quality control and accurate measurements. Within-plate reference controls are typically used for these purposes. Controls help to identify plate-to-plate variability and establish assay background levels. Normalization of raw data removes systematic plate-to-plate variation, making measurements comparable across plates. Systematic errors decrease the validity of results by either over- or underestimating true values. These biases can affect all measurements equally or can depend on factors such as well location, liquid dispensing and signal intensity. Although recent improvements in automation can minimize bias, and thereby provide more reproducible results, equipment malfunctions can nonetheless introduce systematic errors, which must be corrected at the data processing and analysis stages.

Measured compound activity is a function of at least two factors: the compound's true activity and random error (see also "Use of replicates" section). Symbolically, one simple additive model might be $Y_{ijp} = \mu_{ijp} + \varepsilon_{ijp}$ where Y_{ijp} is the observed raw measurement obtained from the well located on row i and column j on the p^{th} plate, μ_{ijp} is the 'true' activity and ε_{ijp} is the effect of all sources of error. Assuming no bias, the ε_{ijk}'s are assumed to have zero mean and a specified probability distribution (e.g., normal). Another simple model is $Y_{ijp} = \mu_{ijp} + R_{ip} + C_{jp} + \varepsilon_{ijp}$ where R and C represent plate-specific row and column artifacts, respectively, and ε_{ijp} represents remaining sources of error. (This latter model is assumed by the median polish procedure described below.) Specifying models explicitly in this manner has the advantage of suggesting how sensitivity and specificity gains can be achieved most cost effectively.

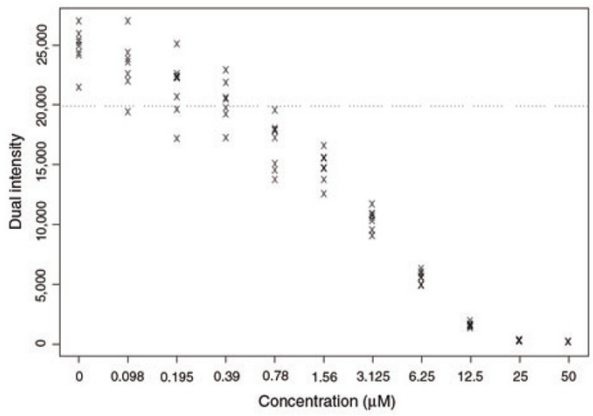

Figure 3 Titration series in a translation assay. These results from an anisomycin titration in a *Renilla* luciferase translation assay show that variability differs across the various concentrations. A hit may be defined as any activity measurement that is at least three standard deviations away from the mean of the control measurements. This corresponds to a dual intensity value of 19,894 (dotted line). All of the measurements for concentrations ≥0.78 are hits (all of the values are below the dotted line). There were six false positives, however, for the three lowest nonnull concentrations.

Current practice. Because of the manner in which compound collections are plated, controls are typically placed contiguously on the outer columns (**Fig. 2**). Unfortunately, a systematic outer column effect affects all of the measurements on the plate because they are adjusted relative to these controls. For example, edge effects may lower (or increase) detection levels on average along the edge compared to the

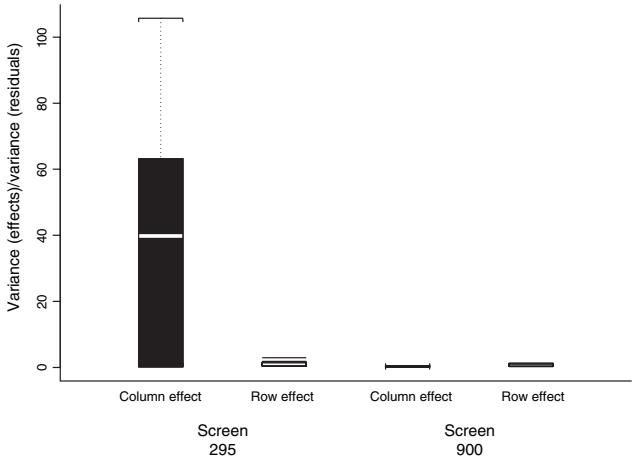

Figure 4 Presence of edge effects in a high-throughput screen. Data from two different screens (http://chembank.broad.harvard.edu/screens) with duplicate measurements across plates are presented. Tukey's two-way median polish was applied to each plate to obtain estimates of row and column effects and of residuals (that is, compound measurements after the polish procedure removed any row and column effects). For each plate, variances of the 16 row effects and of the 24 column effects were divided by the variance of the 384 residuals. Box plots of these variance ratios illustrate the presence of a column effect for screen 295.

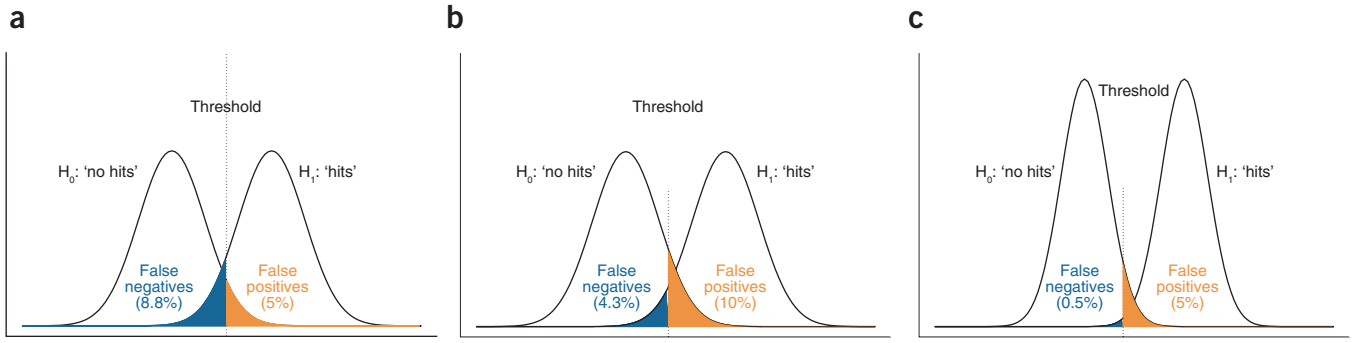

Figure 5 Replicates, false-positive and false-negative rates. In hypothesis testing a false-positive rate (type I error) is the probability of rejecting the null hypothesis (H_0) given that this hypothesis is true. The false-negative rate (type II error) is the probability of failing to reject the null hypothesis (H_0) given that the alternative hypothesis (H_1) is true. (**a**) Given a fixed threshold value, the false-negative and false-positive rates are represented by the blue and the orange areas under the curve, respectively. (**b**) Decreasing the threshold value results in an increase in the false-positive rate and a decrease in the false-negative rate. The opposite would be true if the threshold value were increased. (**c**) The benefit of multiple measurements (replicates) is illustrated. The use of replicates reduces data variability, which is reflected in the narrowed data distributions. Consequently, the false-negative rate is minimized whereas the false-positive rate remains fixed.

remainder of the plate. Consequently, background correction will be lower (or higher) if controls are located on this edge, causing compound activities to appear higher (or lower) than their true states. Worse still, the edge effects may be present in some plates but not others (see "Recommendations" section below). Cell-based biological controls are especially problematic because of variable growth patterns[7]; cell clumping or evaporation within different areas of the plate can lead to different growth conditions and ultimately to position-related bias. Regardless of cause, positional effects increase the rate of false positives and false negatives.

'Percent of control' is one preprocessing method that attempts to correct for plate-to-plate variability by normalizing compound measurements relative to controls. Raw measurements for each compound, for example, can be divided by the average of within-plate controls. 'Normalized percent inhibition' is another control-based method in which the difference between the compound measurement and the mean of the positive controls is divided by the difference between the means of the measurements on the positive and the negative controls. The 'Z score' method excludes control measurements altogether under the assumption that most compounds are inactive and can serve as controls; compound measurements are rescaled relative to within-plate variation by subtracting the average of the plate values and dividing the difference by the standard deviation estimated from all measurements of the plate (see **Box 1**).

The three methods described above implicitly assume a random error distribution that is common to all measurements within a single plate, although without replicates this assumption cannot be verified directly. Positive and negative controls may exhibit differences in variability, however, raising questions about the constant error assumption. Differences in variability among compounds are also likely inasmuch as inactive compounds are similar to negative controls, and active compounds are similar to positive controls[8]. For example, **Figure 3** shows results from a titration series of a protein translation assay in which variability among replicates differs across the various concentrations. In general, nonconstant variances among the compounds of interest may be due to differences in luminescence, reactivity or solubility. The serious errors of inference that can arise from incorrectly assuming one distribution even when departures from it are minimal have been cogently described by Tukey[9].

Another potential difficulty is that these three methods rely on nonrobust statistics. Means and standard deviations are greatly influenced by statistical outliers, which in the context of HTS are putative hits. In

Box 2 Examining the distribution of sample variances

Under the assumption of normally distributed errors with mean μ and variance σ^2, the statistic

$$\frac{(K-1)s^2}{\sigma^2}$$

is distributed as a χ^2 with $K-1$ degrees of freedom where s^2 is the sample variance for each of the K replicated compound measurements.

For each compound, consider the model:

$$y_k = x_k' \beta + \varepsilon_k$$

where $k = 1,\ldots,K$ replicates and it is assumed that:

$$\varepsilon_k \sim N(0,\sigma^2)$$

A standard Bayesian choice for a prior distribution of the variances is an inverse gamma with unknown parameters a and b:

The a and b parameters are assumed to be constant across

$$\sigma^{-2} \sim G(a,b) \equiv \frac{x^{a-1}\exp(-x/b)}{\Gamma(a)b^a}$$

compounds and can be estimated from the data from all compounds by fitting an F-distribution to the sample variances s^2:

Wright and Simon's[18] procedure for estimating the a and b

$$ab(s^2) \sim F_{(K-1),2a}$$

parameters was used to generate the data shown in **Figure 7**.

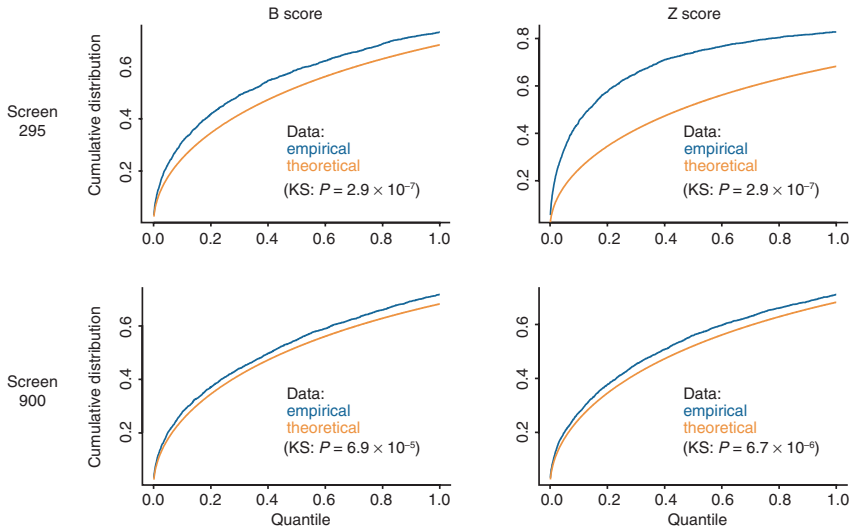

Figure 6 Verification of the assumptions of normally distributed data with constant variance among compounds. Empirical values correspond to a function of the sample variances. Under the assumption of a constant variance among compounds, the overall variance might be estimated by the mean of the sample variances. Each sample variance (obtained from the duplicate measurements) is multiplied by (K – 1) and divided by the overall variance estimate and the ratios should follow a chi-square distribution with 1 degree of freedom (**Box 2**). Results of the Kolmogorov-Smirnov (KS) test of differences between the theoretical and the empirical distributions are shown.

statistical terms, the mean and the standard deviation have low breakdown points, in contrast to more resistant location and scale estimators (e.g., median, Tukey biweight, median absolute deviation). One recent proposal circumvents these issues by adopting a more robust data analysis procedure.

The B score[10] is a robust analog of the Z score which uses an index of dispersion that is more resistant to the presence of outliers and more robust to differences in the measurement error distributions of the compounds (**Box 1**). A two-way median polish is first computed to account for row and column effects of the plate. The resulting residuals within each plate are then divided by their median absolute deviation to standardize for plate-to-plate variability. The B score has three advantages: it is nonparametric (that is, makes minimal distributional assumptions), it minimizes measurement bias due to positional effects and is resistant to statistical outliers.

Recommendations. In the absence of compelling reasons to the contrary, we prefer normalizing the data without using controls. Specifically, we prefer the B score method, especially if row or column biases are suspected. Evidence of these biases can be obtained by examining the variability of the row and column effects estimated by the median polish procedure relative to the residual compound measurements. To illustrate, we reanalyzed two publicly available screening data sets with duplicate measurements for a yeast peptide inhibition assay and a DNA synthesis assay (http://chembank.broad.harvard.edu/screens; screen numbers 295 and 900, respectively). **Figure 4** shows a strong and variable column effect for screen 295. Moreover, as we demonstrate in the "Use of replicates" section, the variability of B scores may more adequately reflect actual random error conditions. This in turn facilitates the decision process because the compound measurements can be benchmarked against theoretical error distributions.

If researchers were to use the Z score method, we would advise they use robust versions to minimize the undesirable influence of outlier compounds (that is, 'hits'). For example, in a 'jackknife' Z score method, $\bar{x}$ and s_x (third equation in **Box 1**) are calculated excluding the compound of interest (x value in the equation); accordingly, s_x differs for each individual compound. Alternatively, in a 'robust' Z score method, $\bar{x}$ and s_x are replaced by more robust measures (e.g., median and median absolute deviation, respectively).

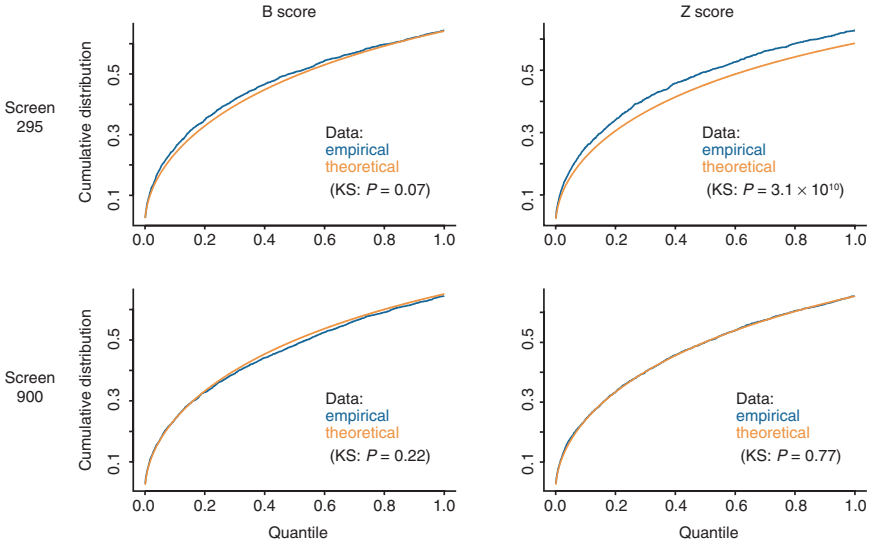

Figure 7 Verification of the assumption that the within-compound variances follow an inverse gamma distribution. Empirical values correspond to a function of the sample variances. Under the assumption of normally distributed data, each sample variance (obtained from the duplicate measurements) is multiplied by the estimated a and b parameters of the inverse gamma distribution and the result should follow an F distribution with 1 and 2a degrees of freedom (**Box 2**). Results of the Kolmogorov-Smirnov (KS) test of differences between the theoretical and the empirical distributions are shown.

Box 3 Test statistics for hit detection with replicates

One sample *t*-test: With K replicates, for each compound a Student *t* statistic is

$$t = \frac{\bar{x} - cons}{s\sqrt{1/K}}$$

where $\bar{x}$ and s are the arithmetic mean and the standard deviation, respectively, of the K replicate measurements, *cons* is a constant typically equal to zero. t follows a t-distribution with $K - 1$ degrees of freedom.

'Modified' one-sample *t*-test: After estimation of the a and b parameters by fitting an inverse gamma distribution to the set of variances across replicates for each compound (see **Box 2**), a variation of the previous standard t-test is:

$$\tilde{t} = \frac{\bar{x} - cons}{\tilde{s}\sqrt{1/K}}$$

where

$$\tilde{s}^2 = \frac{(K-1)s^2 + 2a(ab)^{-1}}{(K-1) + 2a}$$

where $\bar{x}$ and s^2 are the arithmetic mean and the variance, respectively, of the K replicate measurements. The degrees of freedom for the test are now $K - 1 + 2a$, an increase of $2a$ over the standard t-test.

$\tilde{s}^2$ can be viewed as a weighted average of the observed compound-specific variance s^2 and an estimate $(ab)^{-1}$ of the 'typical' error variance underlying the error distributions of different compounds. The weights are $(K - 1)$ and $2a$, respectively. A very large value of a is equivalent to assuming a common variance across all compounds and to simply averaging all of the observed variances, thereby virtually ignoring compound-specific variances. Smaller values of a imply that the underlying variances across compounds are heterogeneous and that the observed compound-specific variances be 'trusted' more. In **Figure 7**, the values of a for screens 295 and 900 were 2.84 and 3.64, respectively for the B scores, and 1.11 and 4.12, respectively, for the Z scores. Accordingly, the estimates were 1:2.84 and 1:3.64 amalgams of the compound-specific and the 'typical' variances for the B scores, and similarly 1:1.11 and 1:4.12 for the Z scores.

For an unreplicated compound, so that $K - 1 = 0$, $\tilde{s}^2$ is simply the typical value, estimated by the quantity $(ab)^{-1}$ with $2a$ degrees of freedom (e.g., ~6 d.f. for the B scores), which is a compromise between zero degrees of freedom associated with single measurements and 'number of compounds – 1' degrees of freedom (that is, 2,687 and 3,839 degrees of freedom, respectively for screen 295 and 900) associated with a common error model.

Controls, if necessary for a specific assay, should be used carefully. Ideally, they should be located randomly within plates, thereby minimizing row or column biases. Current compound collection formats, however, do not lend themselves to randomization. Potential positional effects can nonetheless be minimized by varying the location of controls within plates in a systematic manner. One way consists of alternating well locations for the positive and negative controls along the available edges of the plate (**Fig. 2**). Thus, positive and negative controls will appear equally in each row and in each column and may minimize edge-related bias. For example, in a 96-well plate, an order effect may produce different biases among the different columns. The current practice consisting of eight positive controls on the first column and four negative controls on the last column (**Fig. 2a**) is less efficient than the alternating method (**Fig. 2b**).

If controls are used to normalize compound intensities, it is important to obtain as accurate and precise measurements as possible: any inaccuracies and random measurement errors will lower the accuracy and precision of the normalized values through error propagation. One way to improve precision is to obtain a relatively large number of control measurements (see the "Use of replicates: recommendations" section). Another way is to delete outliers among the controls before normalizing. Identifying measurement outliers among controls is more straightforward than among the compounds of interest because the control measurements are replicates of the same measurement process and should have similar values.

Statistical inference and hit identification thresholds

Regardless of library design strategy (rational or combinatorial), statistical methods offer the means to characterize quality of screens and of hits within a probabilistic framework. Quality can be defined as the ability of the screening process to accurately identify compounds that can be developed into potential leads[11]. A statistical approach to these issues has a number of advantages, including objectivity, reproducibility and ease of comparison across screens.

Once data have been preprocessed with quality control checks and normalization procedures, the next critical step is to decide which compounds should be processed in a secondary screen. Currently, this inferential process is not well defined statistically: procedures for hit identification are based on informal 'rules of thumb' rather than on probabilistic judgments of error rates. In academic settings and in smaller companies, informal rules may also be based on particular laboratory constraints such as capacity limitations. Although it is generally appreciated that lowering the hit-threshold increases false-positive rates while lowering false-negative rates, statistical models can better quantify the balance between specificity and sensitivity by assigning probabilities to the two types of inferential errors (**Fig. 5**).

Current practice. One way to identify hits is to plot raw or preprocessed measurements against compound identity (that is, plot each activity measurement on the y-axis and the well identity 1,2,... 96 on the x-axis) for each plate separately. Compounds whose measured activity deviates from the bulk of the activity measurements are identified as hits. Although this subjective 'eyeball' method may be adequate for identifying highly active compounds, potentially important compounds of low or intermediate potency are difficult to identify reliably and may be missed.

Hits can also be identified as a percentage of the compounds that generate the highest measured activity (e.g., top 1%[4]). From an optimization perspective, this method is arbitrary and suffers from the absence of a probability model. Without prior consideration of the true number of active compounds, one cannot optimize the percentage of primary screen compounds to be screened a second time. If the number of identified potential hits is dictated by the capacity for secondary screening, specificity and sensitivity may vary widely across screens. Consequently, the quality of the results from screen to screen within a laboratory will depend on the extent to which threshold choice reflects the actual number of true active compounds in the various screens.

Compounds whose activity exceeds a fixed 'percent of control' threshold may also be considered as hits. For example, in an agonist assay any compound with an activity measurement that is at least twice the average of the measurements on the negative controls is deemed a hit.

Alternatively, the hit threshold may be defined as a number of standard deviations (typically 3) beyond the mean of the raw or processed data. However, hits (outliers) may cause the distribution of the compound measurements to be skewed. Such a phenomenon may be observed when performing a fluorescent-based assay and when a large number of compounds in the collection are fluorescent. Statistically, imagine the observations as arising from a mixture of two populations with different means (e.g., nonactive compound measurements centered around one mean and active compound measurements around a different mean—likely with different standard deviations also).

As with the preprocessing methods described earlier, the threshold methods described above assume a common magnitude of random error for all measurements and rely on nonrobust statistics, which may lead to inferential errors in hit detection. Hit detection depends jointly on compound concentration, potency and variability. Potency will differ across compounds within a screen, as will actual concentrations due to uncontrolled factors such as solvent evaporation and compound solubility. The easiest hits to detect will be compounds with high relative potencies and concentrations and low variability (**Fig. 3**). Singlet-measurement false positives for the three lowest nonnull concentrations were eliminated when activity measurements were based on means across the eight replicate measurements per concentration. Methods that estimate random error without assuming constant error are described in "Use of replicates: recommendations" below.

Recommendations. One view about false negatives is that little can be done about them and so it is best to adopt a forward-looking perspective and to pursue the hits one does have. We contend, however, that it is important to quantify potential false-negative rates before deciding whether or not they are negligible in a particular screen. If 0.1% of a million compounds to be screened are truly active, a low false-negative rate of 2% represents 20 potential candidates lost. With synthetic compound collections, the potential loss may be lessened because they are made from a set number of basic scaffolds. Thus, in practice, missing an active compound may not matter if related compounds are detected. When screening natural products or extracts, however, truly unique chemical entities will go undetected. Although it is difficult to assign a monetary value to these lost candidates, decisions to not follow-up will typically not be revisited and as such represent irretrievable financial losses.

Verifying data handling assumptions and contrasting various approaches in formal methodological studies are important steps in determining the most cost-effective procedures. Additivity assumptions, for example, can readily be verified from a simple graphical procedure once the data have been preprocessed by the median polish procedure[12]. This same procedure provides a simple method for determining the appropriate data transformation (e.g., log), which will produce additive measurements.

These various steps are necessary for quantifying many aspects of the decision-making process in HTS. Currently, many important go/no-go decisions are based on perceived necessity (e.g., affordability, capacity), subjective perception and past experience. These considerations must enter into any decision process. Statistical modeling of the type we are encouraging enhances rather than replaces this process. Although we believe that currently practiced methods are often insufficiently sensitive to detect hits that arise from potentially important but marginally active compounds, attempts to improve sensitivity must be balanced against specificity and demonstrate cost effectiveness. One way to quantify this balance is to obtain estimates of random error from replicate measurements and to conduct statistical power analysis. Judicious use of replicates will improve sensitivity to minimally active but pharmacologically important compounds that go undetected otherwise.

Use of replicates

Random error reflects inevitable uncertainties in all scientific measurements. This noise unpredictably raises or lowers measurements relative to their true values. Potential sources of random error include biological, instrument and human-related influences. Random error accumulates as a collection of several minimal differences across assays, such as voltage variation, liquid dispensing differences, as well as reagent or sample preparation and handling[11]. Compound-related problems involving chemical properties and activity (e.g., stability, solubility, autofluorescence and degradation) also affect measurement precision.

Precision can be increased by obtaining replicates and by minimizing extraneous variation due to sample handling and processing. Random error estimates, which are central to statistical inference, are typically obtained from replicate measurements of the same attribute or process. Having empirical estimates of variability allows one to use statistical power analysis to control the false-negative rate while maintaining a fixed false-positive rate (**Fig. 5**). We anticipate that obtaining replicates for at least some compounds in primary screens will become more routine.

Current practice. Compounds in primary screens are typically measured only once because of time and cost issues, although the use of duplicate measurements has been recognized for secondary screens and is beginning to be recommended for primary screens (http://iccb.med.harvard.edu/screening/guidelines.htm). Absent replicates, strong assumptions must be made to estimate random error. For example, Buxser and Vroegop[13] describe an approach in which the variability among replicated control measurements is used to estimate variability of the unreplicated compound measurements. Alternatively, random error can be estimated from the variability across single measurements of all compounds on a plate, assuming that all compounds are inactive and that they all have the same random error; early approaches to gene expression microarray analysis adopted a similar approach for estimating error from single measurements[14]. Single measurement methods have ultimately proven inadequate[15], however, and it is now standard practice to obtain at least three replicates per measurement in recognition that replicates offer advantages that outweigh short-term cost considerations[16,17].

Ideal replicates are those measurements that are repeated for the same compound under the same experimental conditions. For this reason and because they underestimate total random error, multiple rereadings of the same plate are not recommended as replicates, except as a check for possible extraneous variation due to the reading process itself. Similarly, structurally similar compounds (analogs) are not recommended as replicates, despite the fact that they may show comparable activity. Nor should measurements of the same compounds under different experimental circumstances (e.g., primary versus secondary screen) be used as replicates because they may be influenced by different extraneous factors (e.g., differences among reagents, batches of compounds and time effects). Finally, pooling compounds in various combinations within individual wells offers timesaving advantages but cannot be considered replication in the usual sense. For example, false positives are more likely to arise when weakly interacting compounds are pooled in the same well or when true active compounds within a row increase. By contrast, false negatives are less common in compound pooling, but may arise if pooled compounds have opposite biological effects of similar size[2].

Recommendations. Replicates offer the twin advantages of greater precision for activity measurements and the means to estimate variability associated with the measurements. Compared with the uncertainty of a single measurement, the imprecision (standard error) of a mean is reduced by

$$100 \times (1 - 1/\sqrt{n})\ \%$$

where n refers to the number of replicates. Having two replicates reduces imprecision by 29%; having three replicates reduces it by a further 13% while having four replicates reduces it an additional 8% (that is, to 50% of the imprecision associated with a single measurement). Thus, replicates make minimally and moderately active compounds easier to detect.

Replicates may appear in wells on the same or on different plates. Although within-plate variation (due, for example, to plate composition and handling) will typically be smaller, across-plate replication is preferred because it represents a more realistic estimate of variation necessary for generalizing results beyond the immediate sample. In general, it is important to obtain estimates of total variability of any measurement process, what has been called 'genuine replication'[18].

We have argued that much of current practice makes strong assumptions about the data (e.g., same magnitude of random error associated with all measurements), which if incorrect can increase both the false-positive and the false-negative rates. Without large-scale studies with replicated measurements, these assumptions and the advantages of more complex statistical modeling approaches are difficult to verify. Moreover, it is unlikely that one approach will be optimal for all screens. These caveats notwithstanding, minimal replication can be used to examine the reasonableness of current assumptions and to potentially improve overall screen sensitivity and specificity.

We illustrate the importance of preprocessing, the need to check assumptions regarding error distributions and the other options available when assumptions are not met, by performing additional analyses on the **Figure 4** data. If the errors associated with the normalized compound measurements from these screens were normally distributed with constant variance across compounds, the sample variances based on the duplicate measurements would follow a χ^2 (1) distribution (**Box 2**). **Figure 6** illustrates the lack of fit, however, between the theoretical and the observed variance distributions for these data, indicating that the normality/constant variance combined assumption is not tenable after preprocessing by either the B score or the Z score procedures.

Alternatively, one can assume that the error associated with compound measurements is normally distributed but with unequal variances distributed across the compounds according to an inverse gamma distribution (**Box 2**). An empirical Bayes approach using this model has been used successfully for analysis of microarray data with minimal replication[15,19,20]. **Figure 7** shows that the error variances of the data sets from **Figure 6** fit an inverse gamma distribution for both data sets for the B scores and for one of the data sets for the Z scores. An important advantage of this variance distribution pattern is that standard statistical tests of compound activity can be constructed using a weighted average of the compound-specific variances estimated from replicated measurements and the overall estimate obtained from the variance distribution; when only a random subsample of the compounds has been replicated, the latter variance estimate can be applied to compounds measured only in singlet from the same screen (**Box 3**). In either case, the more similar the compound-specific variances are to each other, the more reliable the overall variance estimate will be. This in turn will provide more degrees of freedom and more power for the statistical tests. **Figure 7** also illustrates the value of correcting for row and column effects. In the presence of column or row biases (screen 295), B scores more closely approximated the theoretical inverse gamma distribution than the corresponding Z scores, although in their absence (screen 900) the B score method produced a slightly poorer fit.

As more extensively replicated data sets become available, other data-analytic approaches can be examined and optimized. For example, although we found no evidence of a relationship between signal intensity and replicate variability in the two data sets we examined, such a relationship has been used in the microarray context in combination with the inverse gamma variance distribution assumption[21]; this type of relationship may provide additional useful information for estimating random error associated with replicate and singlet measurements. Similarly, if various classes of compounds are thought to differ in terms of variability, random subsets of the various classes may produce more accurate estimates of variability when examined separately. Another approach that may show promise is to model the distribution of activity measurements as a mixture of two distributions (inactive and active compounds)[13]. In short, the principle of 'borrowing strength' from information available from the data in total can provide useful information that would normally be obtained only from large numbers of replicates.

Conclusions

Statistics currently serve a limited role in HTS. One use is to correlate chemical properties with activity levels at the screen development stage to provide information for compound selection and for property modification to enhance chemical activity[22,23]. Once the screen has been run, data mining software packages are increasingly being used for quality control. Notwithstanding these advances in data analysis, HTS continues to lack universal procedures for processing and extracting knowledge from screens[24]. We discuss four broad conclusions below that we believe are warranted at this early stage of development for the statistical modeling of HTS data.

Replicate measurements provide numerous advantages for checking measurement assumptions and improving hit/non-hit decisions. Moreover, quantification and characterization of error variances obtained from replicate measurements allow specificity and sensitivity optimization of individual screens. Given fixed costs, standard statistical power analysis can be used to reach cost-effective decisions regarding the number of plates within a screen to be replicated and the number of replicates.

Statistically adjusting measurements for row and column effects through procedures such as the median polish offers gains in inference and should be used routinely.

The assumption of a common error variance across compounds implicit to many current hit identification approaches is incorrect at least some of the time. At a minimum, the assumption should be routinely verified by replicating some of the compounds and checked against theoretically derived distributions. When the assumption of constant error is untenable, the empirical Bayes approach to estimating random error offers an attractive alternative. It provides an amalgam of the specific within-compound variations (if measured in replicate) and the error estimate derived from the distribution of the within-compound variances, with the latter alone providing the 'best' estimate when a particular compound has not been replicated. This combination of sources of information is a compromise between using only the within-compound (and thus highly variable) error estimates and the average but unrealistic (and thus falsely precise) pooled error estimate that would be appropriate under a common error model.

The limitations of standard statistical approaches with minimal replication can be partially offset by 'borrowing strength' from the large

number of available measurements (compounds). We have provided one example of this principle by using the distribution of sample error variances to obtain error estimates for individual compounds.

Advances in statistical modeling of HTS data will provide objective benchmarks against which to compare experimental results and as a consequence help to standardize the hit identification process. By improving measurement quality and by providing quantifiable false-positive/false-negative ratios, statistical modeling can improve the efficacy of nonstatistical considerations for lead development (such as counter screens to identify nonspecific interference). In this manner, the often-cited advice to identify false leads early and quickly can be strengthened while minimizing potentially costly false negatives.

ACKNOWLEDGMENTS

We thank Jing Liu and Janie Lapointe for generating the **Figure 3** data. This work was supported by the "Informatics and Chemical Genomics" funding to R.N. under the Genome Quebec Phase II Bioinformatics Consortium program.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

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

1. Dove, A. Screening for content—the evolution of high throughput. *Nat. Biotechnol.* **21**, 859–864 (2003).
2. Landro, J.A. *et al.* HTS in the new millennium: the role of pharmacology and flexibility. *J. Pharmacol. Toxicol. Methods* **44**, 273–289 (2000).
3. Stein, R.L. High-throughput screening in academia: the Harvard experience. *J. Biomol. Screen.* **8**, 615–619 (2003).
4. Nelson, R.M. & Yingling, J.D. *Introduction to High-Throughput Screening for Drug Discovery* (IBC USA Conferences, Inc., San Diego, CA, 2004).
5. Campbell, D.T. & Kenny, D.A. *A Primer on Regression Artifacts* (Guilford Press, New York, 1999).
6. Stigler, S.M. Statistics on the Table: the History of Statistical Concepts and Methods (Harvard University Press, Cambridge, MA, 1999).
7. Lundholt, B.K., Scudder, K.M. & Pagliaro, L. A simple technique for reducing edge effect in cell-based assays. *J. Biomol. Screen.* **8**, 566–570 (2003).
8. Zhang, J.H., Chung, T.D.Y. & Oldenburg, K.R. Confirmation of primary active substances from high throughput screening of chemical and biological populations: a statistical approach and practical considerations. *J. Comb. Chem.* **2**, 258–265 (2000).
9. Tukey, J.W. A survey of sampling from contaminated distributions. in *Contributions to Probability and Statistics* (ed. Olkin, I.) 448–485 (Stanford University Press, Stanford, CA, 1960).
10. 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).
11. 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).
12. Hoaglin, D.C., Mosteller, F. & Tukey, J.W. *Understanding Robust and Exploratory Data Analysis* (Wiley, New York, 1983).
13. Buxser, S. & Vroegop, S. Calculating the probability of detection for inhibitors in enzymatic or binding reactions in high-throughput screening. *Anal. Biochem.* **340**, 1–13 (2005).
14. Chen, Y., Dougherty, E.R. & Bittner, M.L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Opt.* **2**, 364–374 (1997).
15. Rocke, D.M. Design and analysis of experiments with high throughput biological assay data. *Semin. Cell Dev. Biol.* **15**, 703–713 (2004).
16. Lee, M.L., Kuo, F.C., Whitmore, G.A. & Sklar, J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. USA* **97**, 9834–9839 (2000).
17. Nadon, R. & Shoemaker, J. Statistical issues with microarrays: processing and analysis. *Trends Genet.* **18**, 265–271 (2002).
18. Box, G.E.P., Hunter, J.S. & Hunter, W.G. *Statistics for Experimenters: Design, Innovation, and Discovery*, edn. 2 (Wiley-Interscience, Hoboken, N.J., 2005).
19. Wright, G.W. & Simon, R.M. A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **19**, 2448–2455 (2003).
20. Smyth, G. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, no.1, art. 3 (2004).
21. Baldi, P. & Long, A.D. A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519 (2001).
22. Verkman, A.S. Drug discovery in academia. *Am. J. Physiol. Cell Physiol.* **286**, C465–C474 (2004).
23. Kerns, E.H. & Di, L. Pharmaceutical profiling in drug discovery. *Drug Discov. Today* **8**, 316–323 (2003).
24. Fay, N. & Ullmann, D. Leveraging process integration in early drug discovery. *Drug Discov. Today* **7**, S181–S186 (2002).