# Improved Statistical Methods for Hit Selection in High-Throughput Screening

Christine Brideau, Bert Gunter, Bill Pikounis and Andy Liaw

*J Biomol Screen* 2003 8: 634
DOI: 10.1177/1087057103258285

# Improved Statistical Methods for Hit Selection in High-Throughput Screening

CHRISTINE BRIDEAU,[1] BERT GUNTER,[2] BILL PIKOUNIS,[2] and ANDY LIAW[2]

High-throughput screening (HTS) plays a central role in modern drug discovery, allowing the rapid screening of large compound collections against a variety of putative drug targets. HTS is an industrial-scale process, relying on sophisticated automation, control, and state-of-the art detection technologies to organize, test, and measure hundreds of thousands to millions of compounds in nano- to microliter volumes. Despite this high technology, hit selection for HTS is still typically done using simple data analysis and basic statistical methods. The authors discuss in this article some shortcomings of these methods and present alternatives based on modern methods of statistical data analysis. Most important, they describe and show numerous real examples from the biologist-friendly StatServer® HTS application (SHS), a custom-developed software tool built on the commercially available S-PLUS® and StatServer® statistical analysis and server software. This system remotely processes HTS data using powerful and sophisticated statistical methodology but insulates users from the technical details by outputting results in a variety of readily interpretable graphs and tables. (*Journal of Biomolecular Screening* 2003:634-647)

**Key words:** high-throughput screening, hits, statistics, screening

## INTRODUCTION

HIGH-THROUGHPUT SCREENING (HTS) is a manufacturing process. The input is samples to be measured and “reagents” (possibly including membranes, whole cells, or other biological entities as well as chemicals) with which to measure them, and the output is numbers. As with any manufacturing process, the output varies. However, this variation is deliberate, caused by the actual variability among the measured samples. In reality, of course, this is never completely true. Some of the variability in the results is due to systematic variation in the measurement process, and some is due to unsystematic “noise,” which can be thought of as “random” influences.

It is important to clearly distinguish and compensate for systematic assay variability, but only more testing (or better quality control[1]) can reduce the effect of assay noise. Figure 1 illustrates how the methods described in this article can effectively compensate for various systematic errors that frequently occur in assays. In this case, a systematic error was detected in a colorimetric immunoassay. The detector revealed a distinct positional effect in alternate rows of a 384-well plate. Although this detection pattern was observed during assay validation, the problem could not be completely resolved as the detector was performing according to the manufacturer’s specifications. Throughout the entire campaign of more than 1000 plates, values situated in row A were on average 14% lower than those situated in row P. In some other situations, the systematic variability can occur across a series of plates even though none exist within each plate. For an assay in which the enzyme was not stable over the 20-h automated experiment, a reduction in the assay window is observed in Figure 2.

As these examples show, systematic effects that are not appropriately adjusted for can bias results. Specifically, on average, they raise or lower (a subset of) the results by a fixed amount. This increases false-positive (falsely identified “hits”) and false-negative (true “hits” that are not identified) rates. This should be contrasted with noise variability that randomly raises or lowers all measured values from their “true” levels. However, it is also the case that the level of this random noise, which is often summarized by the assay variance or coefficient of variation (%CV), can change systematically. This can again lead to higher false positives and negative rates and therefore also needs to be accounted for in the hit selection procedures.

The importance of identifying systematic assay biases and changes in the assay noise level is that these can be effectively adjusted for in the analysis to reduce or eliminate their negative ef-

[1]Department of Biochemistry and Molecular Biology, Merck Frosst Centre for Therapeutic Research, Kirkland, Quebec, Canada.

[2]Biometrics Research Department, Merck Research Laboratories, Rahway, NJ.

Received Feb 9, 2003, and in revised form Jun 9, 2003. Accepted for publication Jul 4, 2003.

Journal of Biomolecular Screening 8(6); 2003
DOI: 10.1177/1087057103258285
Published by Sage Publications in association with The Society for Biomolecular Screening

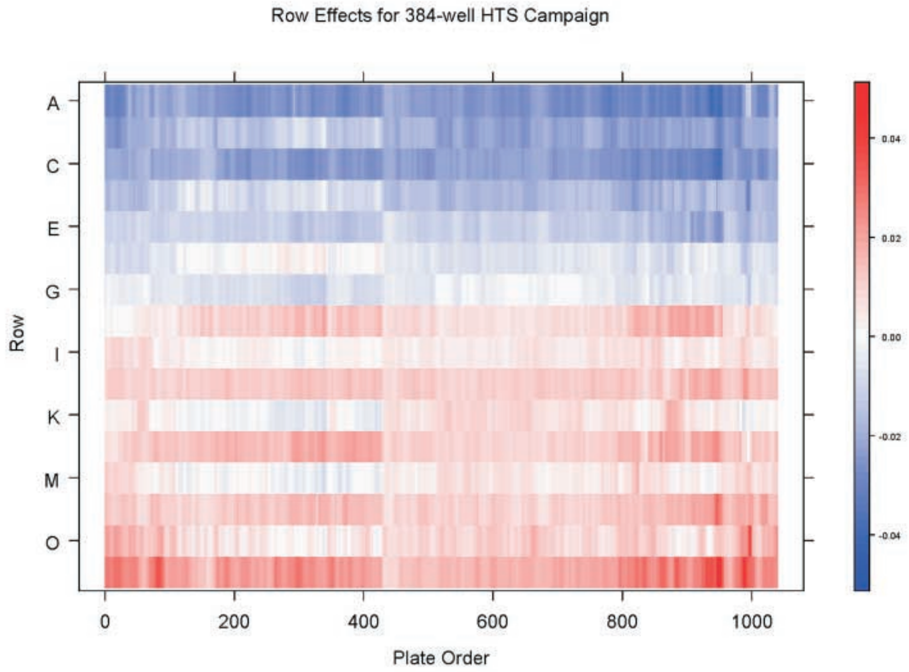


**FIG. 1.** Heat map showing the presence of a strong positional effect in rows A and P from 1183 × 384-well assay plates. Each cell in the heat map encodes with a color value the estimated row effect for the plate sequence index on the *X*-axis and row on the *Y*-axis. A systematic difference from the top to the bottom of the plate is clearly seen, as well as an alternating row-to-row bias.

fects on hit identification. Specifically, fewer false positives and negatives will result.

### *What is wrong with existing controls-based methods?*

The standard approach for dealing with systematic sources of HTS variability is to put fixed controls on each plate (chip, gel, or other experimental unit) and normalize the data to the controls. For example, in an assay in which a high value of the measured output (fluorescence, absorption, radioactivity) occurs when a biochemical reaction is not inhibited and a low result occurs when it is, one might use a vehicle (no compound) as a high control and a known potent inhibitor as a low control. Assuming several samples of each are present in each experimental unit, one typically averages the controls to get a single high (H) and single low (L) control value. Then, for each sample value, $x_i$ in the unit, the normalized percent inhibition is calculated as

$$\frac{H - x_i}{H - L} \times 100. \qquad (1)$$

Note that this does adjust for additive unit-to-unit shifts: If the values on a plate are all shifted up by 200, on average, then the 200 would cancel out in both numerator and denominator and the percent inhibition would be unaffected by the shift.

Nevertheless, there are some important potential problems with this traditional approach.

1. This does not in any way address systematic positional variability, such as edge effects, within the experimental units. In fact, it may exacerbate them. Let us suppose, for example, that all the controls are at an edge in a plate (this is frequently done for convenience). Suppose that there is an edge effect on the plate so that their values are, on average, 200 lower than the rest of the plate. The offset again cancels in the denominator of equation (1), but it does not do so in the numerator. As a result, the values of percent inhibition would be shifted lower, on average, by $\frac{200}{H-L} \times 100\%$. Thus, active compounds would tend to be missed because the percent inhibition would appear to be lower than it actually is. Note that the situation could become even more complicated if the low and high controls were on different edges with different biases for each.
2. Controls may add their own individual biases. For example, if the potency of the low control inhibitor shifted higher at some point, then the denominator in equation (1) would be smaller and the inhibitions would be inflated for those samples measured after the shift compared to those before. Of course, this could be monitored (and adjusted for) by properly tracking low controls, but this may not be done or may simply be left to individual subjective judgment based on "eyeballing" the data, a haphazard and unscientific procedure.
3. Variability among the controls is not accounted for. For example, cost and convenience aside, would better results be obtained from using 1 low and high control per experimental unit or 10 per unit? Quite clearly, the greater the number of controls, the better the precision of the results is concerned. But how much better? Should 1 control or 5 or 25 be used? In practice, these questions are usually

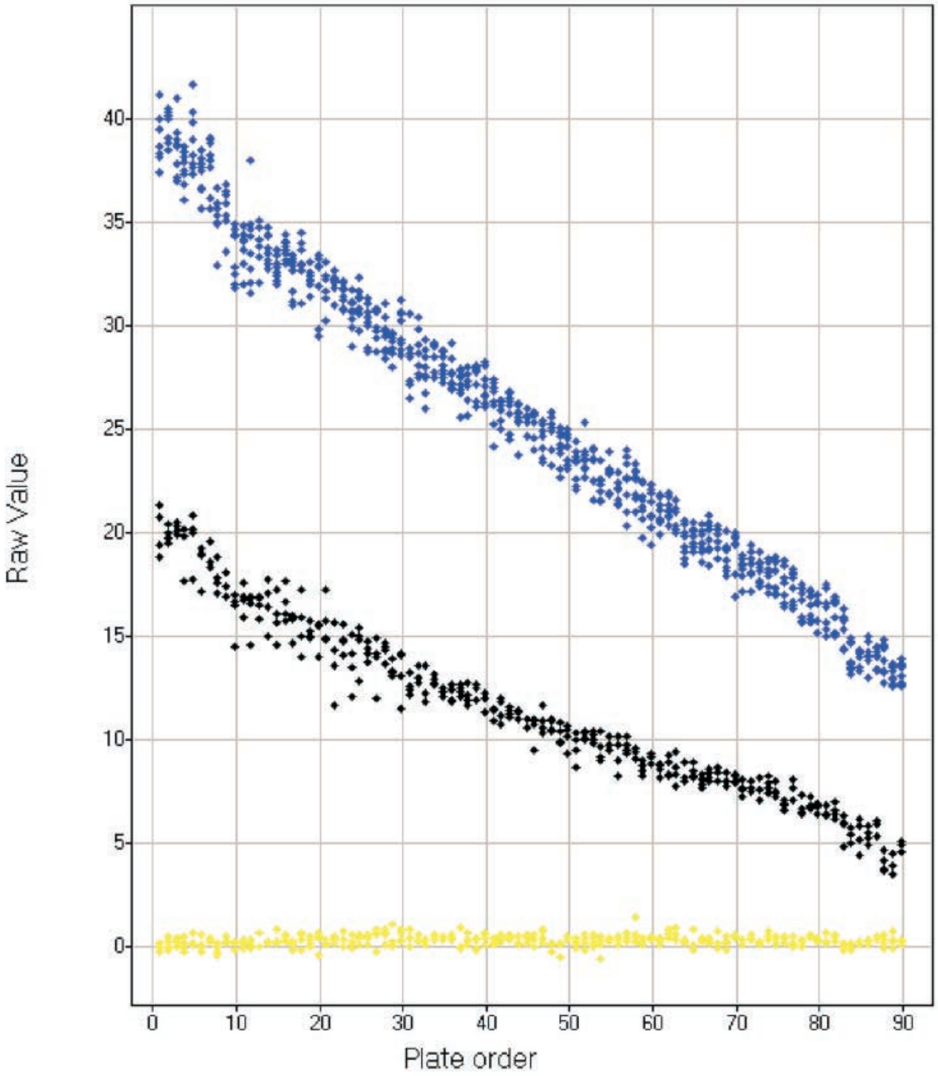


**FIG. 2.** Results of an enzymatic assay run over a 20-h period at room temperature. Highest enzyme activity controls (blue), background activity controls (yellow), and reference compound at the $IC_{50}$ concentration (black). A clear and systematic decay in enzyme activity is shown.

answered on the basis of cost, throughput, assay format, logistics, automation capabilities, and so forth. However, without informing the decisions by an understanding of the impact of variability among the controls, the assay may experience excessive variability that translates into higher false-positive and false-negative rates.

4. Finally, there is the perennial issue of what to do about "outliers" or aberrant values due to measurement anomalies. As is well known, individual wild values can considerably distort averages (and other standard statistics such as standard deviations, curve fits, etc.). This will, in turn, throw off the calculated inhibitions. Manually checking all controls to filter out outliers, even with the aid of visualization tools, can be laborious and prone to subjectivity in deciding what is or is not an outlier. There are various "automatic" schemes for looking for outliers, but these can also be problematic and ad hoc, especially when relatively few controls per unit are used.

With all these potential problems, why is traditional controls-based normalization still so widely used? We think there are several reasons. First of all, most of the time it works reasonably well. If positional effects are not large and the controls are monitored for shifts and outliers, then it is often adequate. Indeed, it is the only way we know of to deal with certain issues, for instance, individual sample units with too many potent compounds, which we shall discuss shortly. Second, it is the traditional way that biologists think about biological activity and much more natural than some of the statistical scores that are discussed below. Finally, it is simply the case that many biologists are unaware of some of the more sophisticated statistical alternatives that are available. Indeed, this is part of what motivated us to write this article. Nevertheless, the preceding discussion argues for the development of alternatives, and we shall describe several below. However, we have found no universal completely satisfactory way to score results from screening and therefore advocate the use of at least 2 together to do the job. We shall try to make it clear in our discussion why we have come to this view.

## MATERIALS AND METHODS

It is usual to provide a section on materials and methods in reports such as this. However, in this article, we are describing data analysis procedures and software that are applicable to all assays, so materials and methods that refer to a specific assay are not pertinent. Rather, we discuss the statistical methods. We shall emphasize the statistical concepts in the main body of the article, deferring more technical details to Appendix B. However, even there, we deliberately avoid complete mathematical rigor, as we feel this is inappropriate for this journal. The references can be consulted for the full details.

To illustrate these statistical methods, we have chosen results from an atypical campaign in which the assay performance changed over time[1] (Figs. 3-8).

## STATISTICAL METHODS

### *Non-controls-based normalization*

If controls are the source of these potential difficulties, then the solution is not to use them. At first glance, this may seem to throw us right back in the predicament of systematic variability: Without controls to normalize for systematic plate-to-plate (unit-to-unit, in general) variability, how can the effect of such systematic changes be avoided? But the answer is remarkably simple: The samples, themselves, are their own controls.

This may seem to be a contradiction, but the nature of screening is that the overwhelming majority of samples screened are not active. Specifically, they act like a vehicle control. Hence, if the sample average shifts from plate to plate by a significant amount, then this is almost certainly due to a shift in the measurement process and not in the samples that are being measured. The exception to this principle is that, on occasion, the vagaries of the sample collection may result in a large number of truly active compounds within a single assay plate. This might occur because the compounds were synthesized as variants around a common structure, as might happen in combinatorial chemistry, for example. In any case, this is why it is necessary to continue to use controls-based methodology in addition to the new methods proposed here: It is the only way to check for and deal with such clustering of active compounds in an experimental unit. We describe how such different measures can be used together in what follows.

We propose here 3 alternative scoring methods that use the samples to be tested as their own controls and thus avoid the problems delineated in the previous section. Each has strengths and weaknesses that are discussed in detail.

### *Median-based activities*

This method can be used only for inhibition assays in which the vehicle samples produce high raw values (e.g., counts, absorbance) while potent hits produce low raw values. This is because adding noise or bias to a relatively small denominator (compared to the numerator) results in large meaningless changes to the value of the ratio, as would be the case in screening for compounds that stimulate a response above basal levels. Recall that the median of *n* data values is the middle value when the data are ordered from low to high (the average of the 2 middle values if *n* is even). Like the mean, it is a measure of the center of the data. Unlike the mean, however, it is largely unaffected by outliers, as they affect the median only through their order in the data, not their actual value. This is important because, in HTS, it is actually the outliers in which we are most interested. The reason is that most of the data will be inactive, vehicle-like, and therefore should be measured as high values, whereas potent samples will have much lower values, thereby appearing as outliers from the bulk of the data. Hence, if we wish to clearly distinguish active outliers from a measure of the center of the inactive bulk of the data, one would prefer that the measure not be pulled away from the bulk by the outliers. The median fits this bill, but the average does not.

Therefore, if one calculates a pseudo percent inhibition as

$$100 \times \left(1 - \frac{x_i}{median(all\, x_i{}' s)}\right),$$

this value will be close to zero for the bulk of inactive samples and near 100 for potent samples with low measured values. A background value (e.g., the average of several completely empty wells) may first be subtracted from the data if desired, although this usually does not make much difference.

Note that this calculation adjusts for plate-to-plate changes differently than the standard inhibition calculation. In that calculation, either additive or multiplicative plate-to-plate differences are adjusted for, but the median calculation works only for multiplicative changes. What this means is illustrated by the following example. If the median raw value is 1000, and a particular well has a value of 100, then its median-based percent inhibition is 90%. Suppose now that all the values on a plate increase by an average of 100. Then, as we saw, the usual percent inhibition would not change, but the median-based inhibition would now be 100 × (1 – 200/1100) = 82%. This effect could be quite drastic. If results increased by 1000, the median-based inhibition would be 45%; if they decreased by 100, it would be 100%!

On the other hand, if all the values were increased by 10% = multiplied by 110%, both the numerator and denominator would be multiplied by 1.1 and thus the value of the fraction would not change. So the median-based calculation assumes that systematic plate-to-plate changes occur multiplicatively, not additively. This is often reasonable, so this approach often adjusts properly in practice. Nevertheless, it is important to make the underlying assumptions explicit so that these issues are clearly highlighted.

Although avoiding the problems of the controls, this statistical potency score still does not deal with positional effects. They will continue to bias results in the same manner as before. Also, as already mentioned, if more than half the samples in the plate are actually active, then the median will be close to 0 and completely ridiculous negative values would result. One would then need to use the median of just the highest values, which sometimes can be difficult to unequivocally identify.

The great advantage of the median-based activity is that it gives results in a form that biologists and chemists can easily interpret while avoiding many of the problems with controls. For this reason, there are some who find this to be a useful way to characterize HTS results and define hit criteria. However, our view is that this interpretability comes at a price that we prefer to avoid. Instead, we prefer to rely primarily on statistical scoring procedures for these purposes. Two such approaches are presented next.

### *Z scores*

Statistical scoring methods depend on a completely different principle to normalize results and select hits. No measure of intrinsic potency is calculated. Rather, such methods are based on the observation that potent compounds are outliers compared to the bulk of nonpotent results. Hence, one can use statistical and graphical methods to identify the outliers as the hits among the statistical distribution of normalized scores

Any relative scoring and outlier detection procedure suffers from a key limitation compared to intrinsic potency measures: It identifies only "different"-looking results, which may or may not be potent enough to be of real biological interest. Consider, for example, an assay that runs for 4 weeks. Suppose that each week, the 100 best results were picked for confirmatory testing. Such a scheme is often necessary for logistical reasons in large-scale testing operations (choosing the 100 best can be preferable to using arbitrary statistical cutoffs based on, often risky, mathematical assumptions). Then, in the extreme, it could be the case that the following occurs. During the first 3 weeks of testing, there are no potent compounds, so that the 300 compounds sent off for testing are just the randomly best-appearing 300 from a distribution of nonpotent compounds. Suppose now that there are 400 (or more) truly potent compounds the 4th week. Then, by choosing only the best 100, 300 true hits have been lost. All 400 of these would have been better than any in the first 300.

Obviously, this is a caricature, but it illustrates the issues. As we shall discuss further in the next section, using both an intrinsic potency measure (controls-based inhibition) along with a relative potency score overcomes such difficulties.

The simplest and most widely known statistical scoring method is the Z score. Computed on a plate-by-plate basis, the Z score for the raw value of $x_i$ is simply defined as

$$z_i = \frac{x_i - \bar{x}}{S_x},$$

where $x$ is the mean of all the sample values (no controls) on a plate, and $S$ is the standard deviation of these values. If the values were approximately normally distributed, then cutoffs of about ±3 or greater might be appropriate. However, in practice, the normal distribution tends to be too well behaved, and longer tailed or even skew distributions may occur. Because modeling such situations (especially in an automated way) can be difficult, we prefer just to rank the results and choose a fixed proportion of the most extreme. We also supplement such rankings with plots of the results and the controls-based calculations, as will be discussed in the next section.

How do Z scores handle the kinds of issues that we have discussed? Obviously, they ignore controls and so are not subject to any of their problems. Unlike median-based activities, Z scores properly handle additive or multiplicative offsets from one plate (experimental unit) to another. For additive offsets, the offset is removed by the subtraction in the numerator; the standard deviation does not change if the mean is shifted. For multiplicative offsets, both the numerator and denominator are multiplied by the same amount, and so the score does not change.

Z scores, like the B score method to be described next, also sensibly handle any plate-to-plate changes in the assay noise or sample variability. These would show up as changes in the variability of the results. For example, suppose one had 2 plates, both with the same mean, but one with results varying from 1000 to 3000 and another with results mostly between 1900 and 2100. Again, under the assumption that most results in a plate are from nonpotent samples, a value of 1500 on the first plate would not look unusual and would have a Z score of perhaps –1. In the 2nd plate, a value of 1500 would stand out with a Z score of perhaps –5 or less. This is clearly what one would want to happen. If the controls behaved similarly, so that the assay window was proportional to the variability, controls-based inhibitions would also adjust properly. However, if this were not the case, for example, if the assay became noisier but the window did not change, then inhibitions would remain the same, on average, thus leading to an increased false-positive rate. Median-based inhibitions would exhibit similar behavior: The median of the nonpotent compounds corresponds to the high control. Thus, the ability of statistical scoring methods to react appropriately to systematic changes in variability in the assay may be an advantage over controls-based activity measures.

There are 2 obvious drawbacks to Z scores and a 3rd that is more subtle. Obviously, they still fail to deal with positional effects. Second, their performance is compromised by outliers, which are the results of greatest interest. To demonstrate, suppose, which nonpotent results have high values and that there are several hits on a plate, that is, low values. This will lower the average, $\bar{x}$, and raise the standard deviation, $S_x$, which has the effect of moving the Z score closer to 0, thus making the hits look less like outliers. Therefore, Z scores may tend to miss more marginal hits or may miss even clear hits if there are enough on a plate to enlarge $S_x$ (the standard deviation is more affected than the mean).

A 3rd, more subtle effect is that none of the methods discussed thus far fully utilize the information that may be present in the data. As an example to illustrate, suppose that plate-to-plate centers and controls are not changing over a range of plates. Then it would be sensible and statistically desirable to combine the data over all the plates in the range to compute means, medians, standard deviations, or other statistics of interest. The advantage is that the calculations used become more efficient because less of the underlying assay measurement variability is transmitted to the estimates (scores or inhibitions). Therefore, the estimated values will be closer to the true values, thus reducing both false positives and false negatives. Although it is difficult to say how large an issue this is in practice for any given assay, it is clearly something that is desirable if it can be accomplished.

We have developed a new statistical scoring method for HTS named B scores (for "better" scores) that is designed to deal with all these issues and thereby provide improved performance over Z scores. However, if there are no positional effects in an assay and changes occur in a plate, then Z scores will be slightly more efficient. Z scores are also simple, calculable by any spreadsheet or database software, whereas B scores require powerful statistical software. However, our experience is that they are worth the effort, providing an effective, non-controls-based methodology to deal with positional effects when they occur while retaining near optimal performance when they do not. For these reasons, we rely on the combined use of B scores and controls-based inhibitions, using median-based inhibitions and Z scores only for backup and examination.

### *B scores*

Like Z scores, B scores are a relative potency score based on the raw sample values. In broad outline, they are similar to Z scores in that they are the ratio of an adjusted raw value in the numerator to a measure of variability in the denominator. However, both the adjustment and measure of variability are more extensive:

- The numerator adjustment accounts for both row and column effects* (positional effects) as well as plate-to-plate changes in the mean. Moreover, it is computed in such a way that it is resistant to outliers. Finally, the computation also attempts to pool results from plates that are "near" in the run order if such pooling appears to be appropriate. This helps stabilize and improve the efficiency of the results.
- The denominator is a resistant measure of the residual variability in a plate after the row and column effects are fitted. Moreover, it, too, is a pooled estimate that tries to appropriately utilize information from nearby (in sequence) plates.

Because the details of the algorithms that do this are complex, we defer a detailed description to Appendix B. However, we describe a couple of aspects here, just to suggest the flavor of how it is done.

### *Numerator adjustment*

Assuming only systematic and assay noise effects, one can model the values in each plate as being the sum of the overall plate center, row, and column effects. This is a standard 2-way layout, and is expressed algebraically as

$$y_{ijp} = \mu_p + \rho_i + \gamma_j + \varepsilon_{ijp}, \tag{2}$$

where $y_{ijp}$ is the value in the $i$th row and $j$th column of the $p$th plate. Here, $\mu_p$ is the plate center, $\rho_i$ is the $i$th row effect, and $\gamma_j$ is the $j$th column effect. $\varepsilon_{ijp}$ is the random noise of the assay on the plate. The standard textbook approach to fitting such a model is to use 2-way ANOVA, a procedure based on fitting row, column, and plate means. However, means are sensitive to outliers, so the B score algorithm uses an iterative "median polish" procedure invented by John Tukey,[2,3] based on medians instead. It has the additional property of easily dealing with missing wells, a ubiquitous problem in HTS.

Further steps of the algorithm then smooth the fitted effects over nearby plates, as described in Appendix B. The adjusted numerator of the B Score for $y_{ijp}$ is then $y_{ijp} - (\hat{\mu}_p + \hat{\rho}_i + \hat{\gamma}_j)$, where the hats over the Greek letters are the estimates from the median polish. This adjusted value is clearly an estimate of $\varepsilon_{ijp}$, the random noise error associated with this value. The Z score numerator adjustment subtracts only $\hat{\mu}_p = \bar{x}$ from the measured value, so it fails to account for the possible row and column effects. By adding this additional complexity to the model (and using resistant techniques to fit it), B scores adjust for positional effects. Of course, if there are no such effects present, then their estimates will be near 0 and the B scores and Z scores will be similar.

### *Denominator*

The denominator of both Z and B scores is an estimate of the amount of spread of the random noise, assuming no hits are in the plate. As we have seen, $S_x$, the Z score estimate, is biased by the presence of hits in a plate. How can one obtain an estimate that would not be so biased? It turns out that this has been a much-studied issue in modern statistics, and many such resistant scale estimates are available. We describe a simple one, "mad" or median absolute deviation, to give the flavor.

After the row, column, and plate adjusted values are computed for all the wells in a plate, these values should look like random noise if there are no hits present. In fact, they should look like random noise centered at 0 because the row/column model first subtracts any systematic positional effect, including the plate center, from each value. Hence, the absolute values of the $\varepsilon_{ijp}$s, which are sometimes called the absolute deviations, are a measure of noise

*These methods are for plate-based assays. Other assay formats would require different positional effects modeling methods.

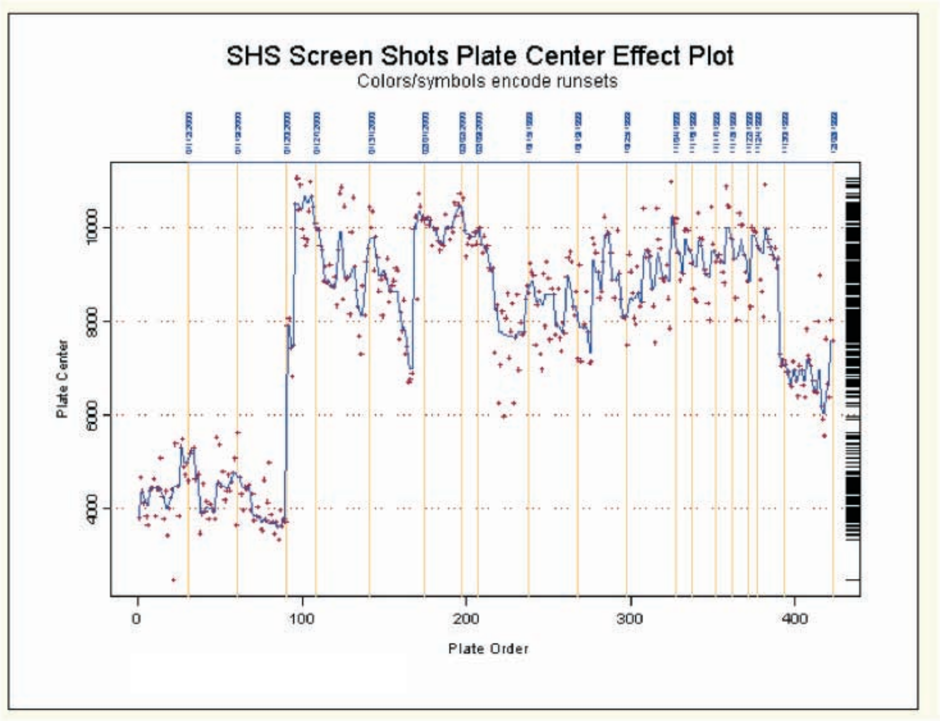


**FIG. 3.** Plot of plate centers versus plate sequence. The centers are the estimated median (estimated $\mu_p$ from the robust fitting algorithm) of the test samples on each plate (controls are ignored). An estimated trend curve is superimposed to help the eye determine the overall behavior. This curve is constructed by an algorithm that attempts to ignore isolated outliers and track step changes as shown in the region of plate 95.

distribution variability. The more variable the noise, the more spread out the $\varepsilon_{ijp}$s and hence the larger their absolute values tend to be. Hence, the median of the absolute deviations (mad) is a measure of the noise spread. In fact, it can be shown that for a Gaussian (normal) distribution, the theoretical mad is close to 0.6745 × the theoretical standard deviation, $\sigma$. For this reason, 1/0.6745 = 1.43 × sample mad is a widely used resistant estimate of the scale (standard deviation) of a distribution.

What is important to note here is that the mad will not be affected by a few hits in the plate. Such hits would show up as a few large absolute deviations (because the adjusted values themselves use a resistant adjustment), which would not much affect the median.

Perceptive readers may note a problem. If there are C columns and R rows in the plate, then it should be the case that more than C/2 hits in any one row or more than R/2 hits in any one column would throw that column's or row's numerator adjustment off, as these adjustments are based on the median of the values in the row or column. This is correct, and these are considerably fewer than the R × C/2 hits that would throw the median off when row and column positional adjustments are not made. However, the smoothing procedure alluded to earlier compensates for this potential problem. If a row or column median on a particular plate is far different from that of plates that occur in sequence before and after, then the smoothing procedure will replace the aberrant value by a value more typical of those in the nearby plates. Thus, the final adjustment for a row or column with "too many" hits in a given plate will be correct.

Details of this and other issues are taken up in Appendix B, but this should give those who do not wish to get involved at the level of technical detail at least some idea of how and why the B score algorithm works.

However, no one scoring method is uniformly perfect. There are circumstances in which B scores can fail to identify a hit whereas other methods will. In HTS, avoiding such false negatives is important; therefore, we have developed a strategy that uses B scores and controls-based activity together (with Z scores and median activities as supplementary information) to reduce such occurrences. In the next section, we describe this strategy using screen shots from StatServer® HTS application (SHS).[1]

## RESULTS

To illustrate the following results, data from a high-throughput campaign of 423 × 96-well plates are used as an example (Figs. 3-8). The data were obtained from a receptor-ligand binding assay using the scintillation proximity assay (SPA) technology. In this example, a sudden increase in raw values after the first 100 assay plates (Fig. 3) was observed for all samples and controls. The effect could not be assigned to any known cause.

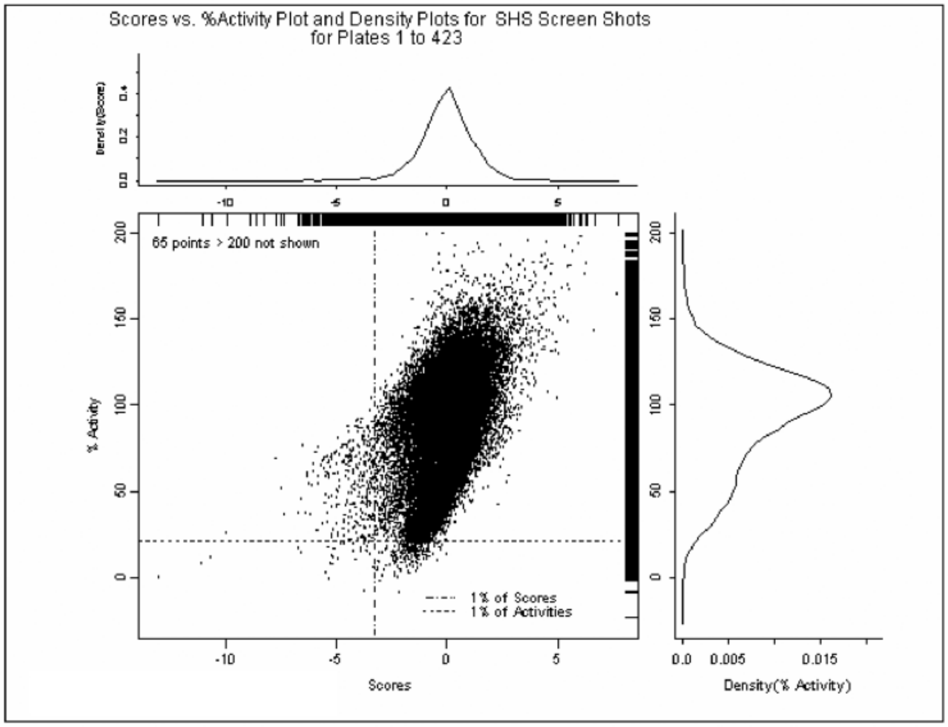


**FIG. 4.** Data from 34,000 samples are represented in a scatterplot of scores versus activity. For each sample, the controls-based activity is plotted on the *y*-axis and the corresponding B score on the *x*-axis. For each measure, marginal density plots (a kind of smoothed histogram) are plotted in the corresponding margins. A "rug" plot in which each value is indicated by a short line segment is also shown for each axis.

Raw data from the assay may be assembled and submitted to SHS in one of several different ways: as text data, as an Excel® file, or as a database query. The data may be the results of the entire assay or any portion thereof, for example, a day's or week's worth to be examined for quality control purposes. Once submitted, the user may restrict output to any contiguous (in sequence) subset of plates. This functions as a kind of noninteractive zoom capability to examine results at any desired level of detail. Although the various graphs and data displays are static, the system allows sufficient flexibility and ease of use for users to investigate what the data have to say at the needed level of detail. In addition, as will be shown, the ability to combine many graphs in a regular array on a single page in so-called "trellis displays"* is a very effective way to visualize assay performance over many plates.

*Hit summaries: plots and tables*

The initial set of graphs and tables provides a summary of the overall assay performance, identifying possible hits and highlighting issues for further investigation. In these results, a hit by any criterion has been defined as any result that is among the best 1% by the criterion. This threshold is arbitrary and user specifiable, but we have found that a 1% default generally works well when initiating the hit selection process.

The first display, a scores-versus-activities plot for approximately 34,000 samples in the data set, is shown in Figure 4. It is a scatterplot (also called a crossplot or xyplot) of the controls-based activities on the *y*-axis versus their corresponding B scores on the *x*-axis. The plot is enhanced in several different ways to provide more information.

- To avoid even more of the plot condensing into a black blob and to speed up the plotting, the graph "bins" the points to reduce the number of points plotted without removing key detail.
- Extreme outliers that would distort the plot and lose detail are omitted and annotated with a note: "65 points > 200 not shown."
- "Rug" plots on the axes give the marginal distribution of the activities and B scores. A short line is drawn for each data value.
- Nonparametric density plots in the margins of the graph plot the marginal distributions.
- The user-specified hit cutoffs by each criterion are indicated by horizontal and vertical lines. They are correctly drawn in the right position whether potent values are low or high (in this particular assay, hits corresponded to low activities and therefore high inhibition).

Ideally, one would want the 2 scoring methods to track each other closely, which would mean that the scatterplot would fall close to a 45° line and most hits would fall in the lower left quadrant of agreement. Clearly this is not the case for this assay, which

*A term coined by William Cleveland of Lucent Research. See the references for details.

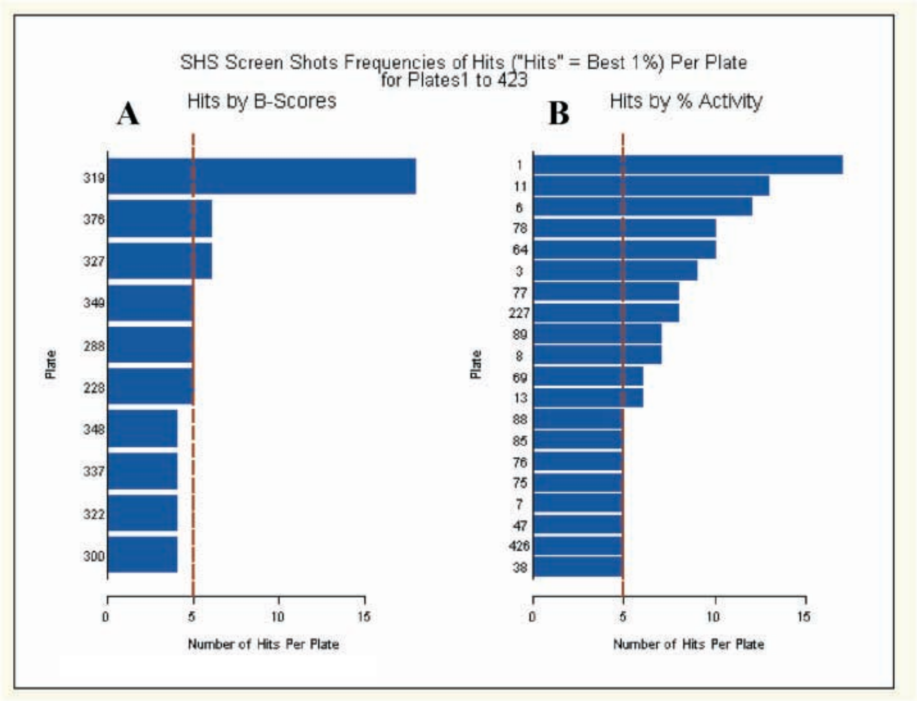


**FIG. 5.** Pareto plot of the top 1% hits by plate number. The plates are ranked by the number of hits selected by B score **(A)** and by percent activity **(B)** present on the plate. The vertical limits on each plot represent the number of hits per plate that would occur purely by chance (based on the binomial distribution). The plates with just a few hits are not shown. This plot focuses attention on plates with an unusually large number of hits by one or the other criteria, indicating problems that need to be addressed.

is already an important diagnostic: Why did the B scores not agree with the controls-based results? Further plots follow that will shed light on this question.

The next part of the output (not shown) provides the hits and nonhits lists as a series of tables that include details of the data including the percent activity and B scores for each sample. These tables can be scrolled and examined in the browser window or downloaded and viewed as a local file (with Excel, for example). They can be used to identify specific samples for confirmatory testing and/or to gain insight into the reasons for the discrepancies among the scores that may be diagnostic of assay problems.

A sequence of hit frequency plots follows and provides an extensive visualization of where the hits occurred (in the run sequence) and in which rows and columns on the plates they occurred for both B scores and controls-based activities.

Figure 5 illustrates barplots of the hits by plate sequence number, listed in rank order of the hits (this is known as a Pareto plot). The vertical lines are statistical limits (based on the binomial distribution) of what would occur purely by chance if hits had the same probability of appearing in each plate. Note that using the B score criterion, only 3 plates were above this limit (only one, 319, very much so) whereas there were many plates significantly above it for percent activities. Specifically, using B scores, hits were, indeed, spread out fairly uniformly among the plates, whereas using the activity criterion, they were concentrated on relatively few plates. Note that the barplots show only a subset of plates that got the most hits. The many plates with just a few (where "few" is defined as statistically "random") hits are not shown, as this does not provide any useful information.

Even more interesting, the controls-based hits occur almost entirely among the first 100 plates or so, whereas for B scores, they are more evenly spread out but mostly not among the first 100 plates. This is a very clear indication of something going on in the assay that requires investigation.

Figure 6 tallies the hits by which row or column they appeared in under both hit criteria. The format for this assay was the usual 8 row × 12 column 96-well plate; columns 1 and 12 were used for controls, so only the middle 10 columns contained samples. One would not expect any row or column to have preferentially more hits than any other. Specifically, one would expect the hits to occur randomly and evenly among the different rows and columns, which is what one sees for the B scores. However, using activities to define hits, it appears that row 8 has a scarcity, perhaps indicating a possible edge effect. Similar column hit plots are, of course, also provided but are omitted here.

Figures 7 and 8 are "maps" to provide more detail on the previous plots. In each plot, the plate sequence (1 to 423) is plotted on the horizontal axis, while the vertical axis plots the row number, 1 to 8, for all rows containing samples. The corresponding column hit maps are omitted. Each position in the map corresponds to a particular row in a particular plate. A blue scale bar is drawn on the map in that position according to the legend on the right to indicate

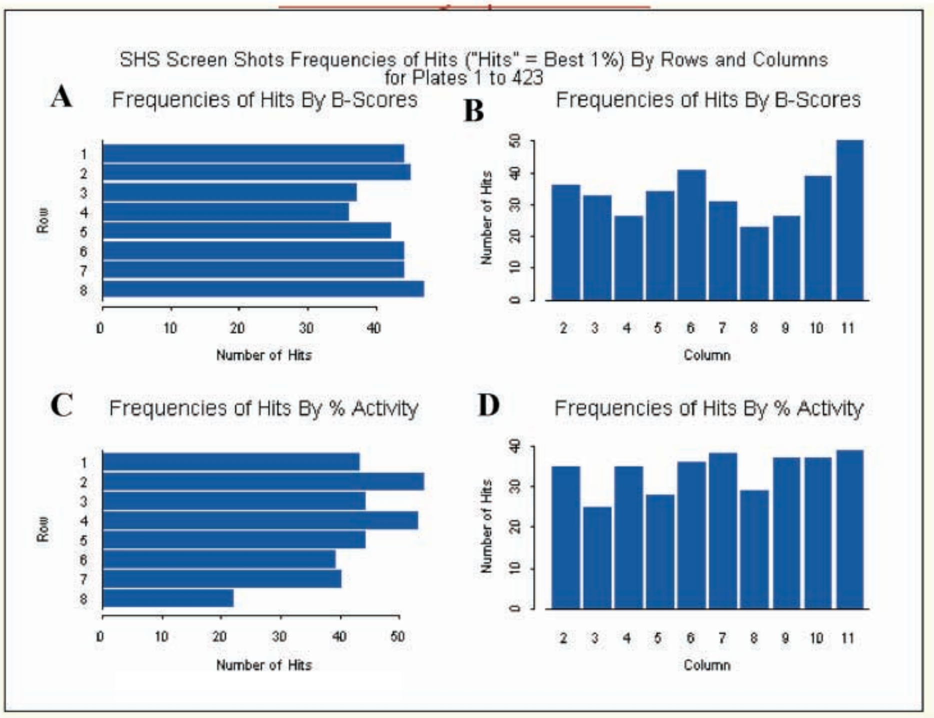


**FIG. 6.** Barplot of the total number of hits selected from B scores by row **(A)** and by column **(B)** compared to total number of hits selected from percent activity by row **(C)** and by column **(D)**. This set of plots helps visualize the effect, if any, positional biases have on hit determination.

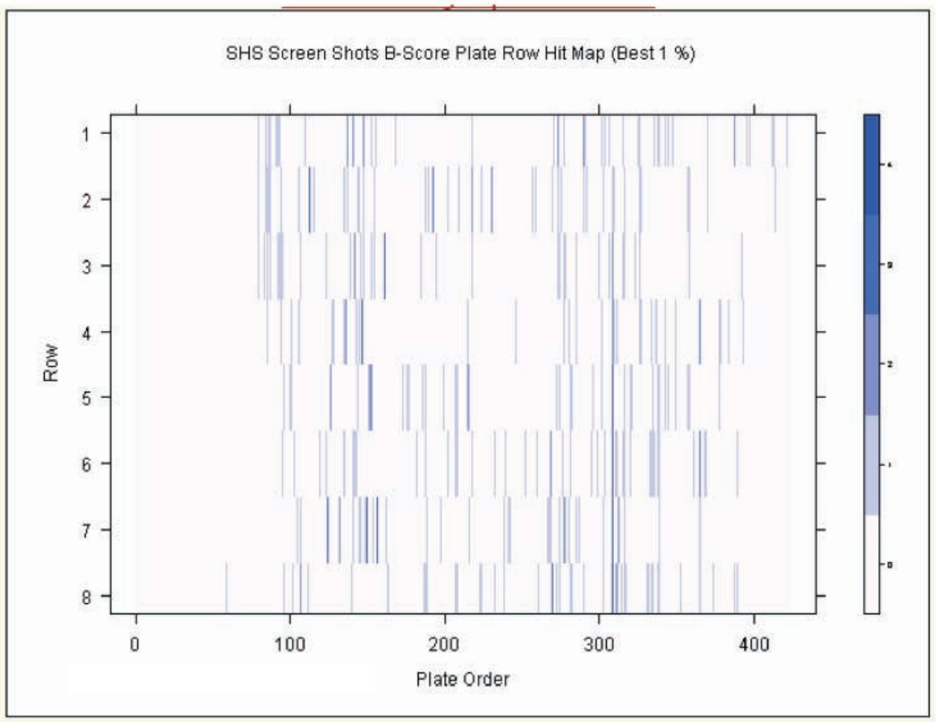


**FIG. 7.** Map of hits (best 1%) selected from the B scores by row. The *x*-axis indicates the assay plates in the sequence they were screened and the *y*-axis indicates the row number. Similar to a heat map, each (thin) bar encodes the number of hits in the corresponding plate and row position by saturation on a blue scale (legend on the right).

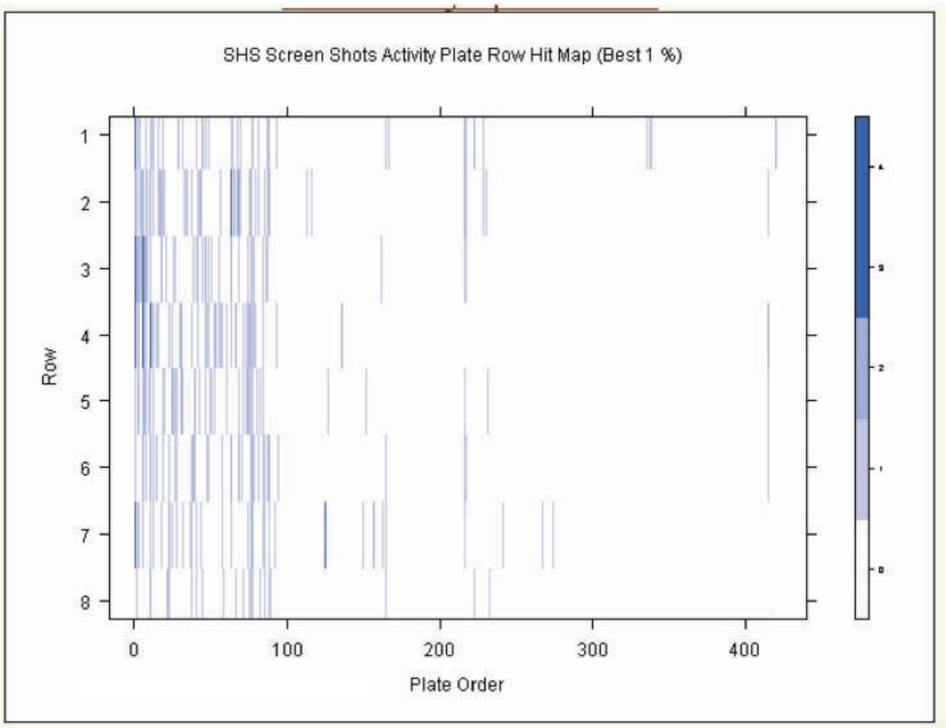
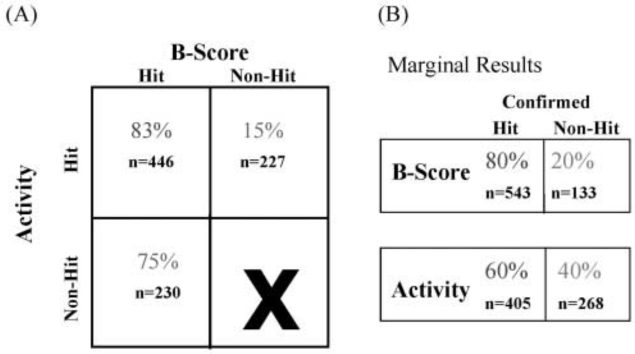


**FIG. 8.** Map of hits (best 1%) selected from controls-based activity by row. Similar to the map in Figure 7, except that it displays hits selected by the controls-based activity criterion. The dramatic difference between the 2 plots shows the impact of positional effects on the hit selection. This demonstrates the value of using a criterion that attempts to compensate for positional effects to supplement standard hit selection procedures.

the number of hits, if any, that were in that row on that plate. The maps clarify and extend the previous summaries. They make it is strikingly clear that B score hits occurred almost exclusively after the first 100 plates, whereas controls-based hits occurred almost exclusively before.

Naturally, it would be useful to know why this occurred. Using the quality control methods and graphs discussed in Gunter and others,[1] we show that there was a systematic bias in the assay controls prior to plate 100: high controls (corresponding to nonhits) averaged about 6000, whereas the actual sample values averaged about 4000 (Fig. 3). This bias essentially disappeared after plate 100. Recalling how activity scores are calculated, it is clear that this bias between plate and high control mean levels translates into lower scores for the first 100 plates. This effect was also heightened by a change in the assay control window, the difference being about 4700 (6000–1300) for the first 100 plates and 6500 (9000-2500) thereafter. As this is the denominator of the activity score, the smaller denominator also produced lower scores for the early plates. Finally, the variability of the sample measurements, measured as CV, was also much greater prior to plate 100. The CV decreased from about 25% to 30% just before plate 100 to 5% to 10% just after. This means that not only were the activity scores lower on average before plate 100, but they also varied more around that average, and hence there were more extremely low scores, that is, hits.



(A)

| | | B-Score | |
|---|---|---|---|
| | | Hit | Non-Hit |
| Activity | Hit | 83%<br>n=446 | 15%<br>n=227 |
| | Non-Hit | 75%<br>n=230 | X |

(B)

Marginal Results

| | Confirmed | |
|---|---|---|
| | Hit | Non-Hit |
| B-Score | 80%<br>n=543 | 20%<br>n=133 |
| Activity | 60%<br>n=405 | 40%<br>n=268 |

**FIG 9.** Results from high-throughput screening campaign with a constant positional effect (as shown in Fig. 1) where the top 0.2% hits were selected by either the B score or activity (>3SD) criteria. All hits were subsequently retested in the same assay, and an $IC_{50}$ was determined for each. A hit was confirmed active if the $IC_{50}$ was below 10 μM, the initial screening concentration. **(A)** Confirmation results for each criteria shown by hit or nonhit quadrants. **(B)** Confirmation results for hits selected by B score only or by activity (>3SD) only.

Conversely, the B scores were unaffected by the bias between controls and samples because controls were not used to calculate them. Moreover, the large change in CV had exactly the opposite effect, as sample results were much noisier prior to plate 100; thus, fewer individual values stood out as extreme among them. Hence, there were fewer B score hits (especially after adjustment for the positional effects that were present) prior to plate 100 than subsequently.

Which scoring system is correct? Clearly, this depends on whether the shifts were caused more by a reduction in assay noise or a reduction in assay sensitivity (these are often linked together). What is important is that the use of alternative statistical scoring methods, adjustment for positional effects, and related quality control methods exposed problems in this example assay to provide greater insight as to their possible causes. The various plots and summaries pinpointed when the problems occurred and demonstrated their distorting effect on hit selection. Moreover, we have also shown that when positional biases are not present, the overall strategy we advocate still performs well and similarly to standard hit selection metrics.

## DISCUSSION

One of the challenges in HTS is to clearly distinguish active outliers from the inactive bulk of the data. The conventional selection criterion is to choose an arbitrary percent activity (or inhibition) threshold such as 50%. The use of 3 standard deviations (3SDs) from the mean of all samples is another common method. However, this may mean that too many compounds are selected for further investigation and may cause a bottleneck in secondary assays. If too few are chosen, then a potential lead may be missed. For practical reasons, the number of hits is often chosen by how many compounds can be tested in secondary assays. Choosing a small percentage (say 0.5%) of active compounds may reduce the number of hits, but it may also lead to an increased false-negative rate.

All HTS campaigns are prone to systematic errors. Individually, none of the conventional methods for hit selection will do an adequate job of identifying the real hits when such errors are present. Detecting systematic error is essential during assay validation and optimization. However, there are instances in which this error is difficult to eliminate, as was the case with the example assay discussed in the results. This is especially true for cell-based assays in which edge effects are common. Occasionally, the systematic error is unpredictable (such as a blocked pipette tip) and may be overlooked by the analyst during quality control validation.

In most instances, the most active samples will be selected as hits. The challenge is to select the marginally active hits. It has been our experience that the activity of a compound tested as a singlet at a fixed concentration will not predict the real intrinsic potency (i.e., $IC_{50}$) of the compound. Therefore, marginal hits may have very good true potency but may be masked during the HTS campaign for several reasons (compound instability, random or systematic errors during the assay). Consider the example (Fig. 1) in which the mean inhibition values of samples situated in row A are less than those in row P. Simply choosing a threshold of less than 50% inhibition would have led to missed actives in row A and more false positives in row P. The same holds true if the 3SD method were used for selection. Using the B score algorithm in addition to the activity threshold, true hits were identified in row A. In fact, when these hits were subsequently tested again, we found many of the initial nonhits (using the 3SD cutoff) in row A were actually potent hits. One of these actives was identified as a potential lead compound for development because it was shown to be potent (200 nM $IC_{50}$) and cell permeable. After a thorough examination of hits chosen by either the 3SD or B score criteria (0.2% hit rate), we observed that 80% of the B score hits confirmed activity in follow-up experiments in contrast to 60% for the 3SD hits (Fig. 9). More important, of the hits that met the 3SD criteria and not B score, only 15% confirmed as true hits in contrast to 75% for hits chosen by B score with the low activity.

The chosen threshold level is always arbitrary and may depend on the target and the quality of the assay results. Combining a robust statistical method such as the B score with the controls-based percent activity, median score, or Z score helps to select the most appropriate threshold. In assays for which the results are inherently variable (cell-based assays, for example), the number of active compounds may not correlate as well with the B score. Therefore, choosing the top active samples by both B score and controls-based activity will increase the probability that true hits are selected (Fig. 9). The correlation between the 2 measures will also indicate the presence of positional effects or systematic errors. For example, consider a compound library of 500,000 samples in which the top 0.2% is chosen by percent activity and by B score. Then 1000 compounds would have been selected if there were a perfect correlation. However, if systematic errors existed during the HTS campaign, then fewer samples would fall into the active quadrants and perhaps only 0.15% of the compounds would be considered hits simultaneously by both criteria.

There is a general limitation to the B score algorithm where it will fail to identify hits. This occurs when compounds are not randomly distributed across plates. Compound libraries that are synthesized around a common core may result in many active hits. If these compounds were all tested on the same assay plate, they would be ignored by the B score algorithm because the low raw values would be normalized to each other. The high active hit plates can be easily identified using the SHS by viewing the hit barplots (Fig. 5). Typically, the active plate will show a large number of hits on the hits-by-activity plot, but the same plate will be absent from the B score plot. The analyst can then examine these plates individually to select the appropriate compounds. This is usually rare when screening in higher density plates such as 384 and 1536 because most of the compound library plates were generated in 96-well plates.

In summary, we have developed a method to help identify active compounds from HTS data. The SHS software allows the analyst to view the entire HTS campaign results within a series of plots and helps identify the active compounds. The analysis is rapid and helps identify systematic errors. It is important to view HTS data through many visualization methods because no single graph will be able to represent an overall picture of the assay performance. Similarly, it is also important to use more than one calculation method to identify hits. A robust statistical model such as the B score adds value to the other traditional calculation methods.

## ACKNOWLEDGMENT

We thank our colleagues Dr. Vladimir Svetnik and Jerry Ferentinos for their support and valuable contributions to the development and validation of the methodology.

## APPENDIX A
## Further Annotated Statistical References

### *For robust/resistant estimators, median polish of 2-way tables, and boxplots*

A fairly accessible discussion of the ideas and methods can be found in:

Hoaglin DC, Mosteller F, Tukey JW: *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley, 1983.

A good, nontechnical discussion for social scientists that minimizes mathematics is found in:

Fox J, Long JS: *Modern Methods of Data Analysis*. Newbury Park (CA): Sage, 1990.

The original source of many of the ideas and methods requires nothing more than high school math but is still extremely challenging:

Tukey JW: *Exploratory Data Analysis*. Reading (MA): Addison-Wesley, 1977.

Robust/resistant methods have been an area of major research interest in statistics for the past 30 years, so there are thousands of papers on various aspects of the topic. One statistical text that covers many of the main ideas is:

Rousseeuw PJ, Leroy AM: *Robust Regression and Outlier Detection*. New York: John Wiley, 1987.

### *For robust smoothing*

This is another topic with a huge statistical literature. Both *Exploratory Data Analysis* and *Modern Methods of Data Analysis* contain good nontechnical discussions of one form of smoothing by repeated running medians. The B score algorithm smoothing method is a combination of ideas borrowed from running median procedures, the loess regression method discussed in the Cleveland reference below, and fitting by elemental sets, for which a good reference is:

Hawkins DM: The accuracy of elemental set approximations for regression. *J Am Stat Assoc* 1993;88:580-589.

### *For trellis and other graphical displays*

A wonderful, totally nontechnical reference on principles of graphical display of data that is suitable (and recommended!) for all biologists is:

Tufte ER: *The Visual Display of Quantitative Information*. Cheshire (CT): Graphics Press, 1983.

A slightly more mathematically oriented treatment that also includes an extensive discussion of trellis plots is:

Cleveland WS: *Visualizing Data*. Summit (NJ): Hobart, 1993.

### *For the S-Plus software system and S-language*

The obvious reference is the user manuals and guides. Two excellent books by Venables and Ripley are also available that discuss both the statistical and graphical capabilities of the software and its programming:

Ripley BD, Venables WN: *Modern Applied Statistics with S*, 4th ed. New York: Springer, 2002.

Ripley BD, Venables WN: *S Programming*. New York: Springer, 2000.

## APPENDIX B
## B Score Model and Algorithm

*Notation:* Plates are indexed by the order in which they were run, $p = 1, 2, \ldots, P$. Rows and columns are indexed by $i$ and $j$, respectively. For 96-well plates, for example, rows are indexed as $i = 1, \ldots, 8$, and columns are indexed as $j = 2, \ldots, 11$. Thus, the raw measured valued (optical density, number of counts, etc., not percentage potency) for row $i$ and column $j$ on the $p$th plate is $x_{ijp}$.

The model fit by the algorithm is

$$x_{ijp} = \mu_p + R_{ip} + C_{jp} + \text{smooth}_p(e_{ijp}) + \varepsilon_{ijp}, \text{ where } \varepsilon_{ijp} \sim (0, \sigma_p^2) \qquad (1)$$

In words, the value in the $i,j$th sample well on plate $p$ is the sum of an overall average for plate $p$ + a possible systematic measurement row offset for row $i$ on plate $p$ + a possible systematic measurement column offset for column $j$ on plate $p$ + a possible systematic interaction effect for the $ij$th well on plate $p$ + measurement noise with 0 mean and possibly different variances from plate to plate.

Here is a further explanation of the "smooth" function. The general idea is this: although there could be different positional effects on different plates (e.g., a row problem that appears suddenly due to a tip clog), we would generally expect systematic positional effects such as edge effects or a localized high region on a plate in the center or corner to be fairly consistent from plate to plate, especially for plates that are measured close together in time sequence. Therefore, results across plates ought to be combined in some sensible way to reflect this. The smooth function does this by doing a kind of local averaging at each well position longitudinally (over the plate sequence). This enables the systematic fit to capture local positional effects, including even an individual well that is consistently different perhaps due to a clogged tip, which would not be captured by the overall row/column fits.

The smoothing algorithm used here must additionally allow for step changes—most standard smoothing algorithms blur over such changes (to make them smooth!). To do this, a robust windowing procedure is used as follows:

1. Choose the ½ window width, $k$. Then, successively, for all indices, $i$, away from the series beginning and end (for which special end value procedures have to be used, as usual), form $S_{iL}$ and $S_{iR}$, the $k$ values to the left of the $i$th and the $k$ values to the right.
2. Let $m_{iL}$ = median(slopes of all pairs of values in $S_{iL}$) and similarly for $m_{iR}$. These pairs of values are known as "elemental subsets" in the statistical literature.[4] The left and right medians are outlier resistant estimates of the slopes of all possible lines fit to the pairs of values in the left and right subsets, $S_{iL}$ and $S_{iR}$.
3. For each of the $k$ indices $j = i - k, i - k + 1, \ldots, i - 1$ to the left of $i$, fit a line through the point $(j, x_j)$ with slope $m_{iL}$. Then this line has intercept $(x_j - m_{iL} \times j)$. The median of all these intercepts is then a resistant version of the usual intercept for the left subset. Do the same for the $k$ points to the right of $x_i$ to get a resistant intercept for these values. So this gives 2 resistant fitted lines (i.e., slopes and intercepts), 1 each for the left and right subsets.
4. Use the 2 fitted lines to get left and right predicted values at $i$, $\hat{x}_{iL}$, and $\hat{x}_{iR}$. Then the smoothed value at index $i$, $\hat{x}_i$, is median $(\hat{x}_{iL}, \hat{x}_i, \hat{x}_{iR})$.

An obvious issue here is the choice of $k$: the larger, the generally smoother the result will be, but the less able the curve will be to follow true changes in the data (the so-called variance/bias trade-off). We chose fixed values of $k$ for each plate format that we use on the basis of experience that generally seem to work well. More sophisticated strategies, using cross-validation, for example, could be entertained.

As the examples show, this procedure ignores outliers, tracks sudden jumps in level, and smooths out (to some extent) the random wiggles.

With this explanation in hand, the B score algorithm proceeds as follows:

Input data: Raw data from the sample wells in each plate arranged in plate run order.

Step 1: For each plate, $p$, resistantly fit a row-column additive model to the plate data using the 2-way median polish procedure built into S-Plus. This gives

$$y_{ijp} = \mu_p + R'_{ip} + C'_{jp} + e_{ijp} \text{ for each plate, } p.$$

Note that the primes on the row and column effects indicate that they then might be smoothed (along $p$) by the smoothing algorithm.

Step 2: For each $i,j$ combination, smooth (along $p$) the $e_{ijp}$s using the smoothing algorithm described above. This then gives

$$y_{ijp} = \mu_p + R'_{ip} + C'_{jp} + \text{smooth}_p(e_{ijp}) + r_{ijp}.$$

The first 4 terms represent the fit and contain the systematic positional effects, if any, whereas the $r_{ijp}$s are the residuals and represent the true measured potency—including extreme potencies of active compounds—without the distortion of the positional effects.

Step 3: For each plate $p$, obtain the Gaussian adjusted median absolute deviation (mad) of the $r_{ijp}$'s, $mad_p$ (these could also be smoothed, if desired). Then the B score for the $ijp$th well is $r_{ijp}/mad_p$.

## REFERENCES

1. Gunter B, Brideau C, Pikounis B, Pajni N, Liaw A: Statistical and graphical methods for quality control determination of high throughput screening data. *J Biomol Screen* 2003;8:624-633.
2. Tukey JW: *Exploratory Data Analysis*. Reading (MA): Addison-Wesley, 1977.
3. Emerson J, Hoaglin D: Analysis of two-way tables by medians. In Hoaglin J, Mosteller F, and Tukey J (eds.): *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley, 1983.
4. Hawkins DM: The accuracy of elemental set approximations for regression. *J Am Stat Assoc* 1993;88:580-589.

Address reprint requests to:
*Christine Brideau*
*Merck Frosst Centre for Therapeutic Research*
*Merck Frosst Canada*
*16711 Trans-Canada Highway*
*Kirkland, Quebec, Canada, H9H 3L1*

*E-mail:* christine_brideau@merck.com