Bioinformatics 2007

An efficient method for the detection and elimination of systematic error in high-throughput screening

Vladimir Makarenkov[1*], Pablo Zentilli[1], Dmytro Kevorkov[1], Andrei Gagarin[1], Nathalie Malo[2,3] and Robert Nadon[2,4]

[1]Departement d'informatique, Université du Québec à Montreal, C.P.8888, s. Centre Ville, Montreal, QC, Canada, H3C 3P8

[2]McGill University and Genome Quebec Innovation Centre, 740 Dr. Penfield Ave., Montreal, QC, Canada, H3A 1A4

[3]McGill University, Department of Epidemiology, Biostatistics, and Occupational Health, 1020 Pine Av. West, Montreal, QC, Canada, H3A 1A4

[4]McGill University, Department of Human Genetics, 1205 Dr. Penfield Ave., N5/13, Montreal, QC, Canada, H3A 1B1

ABSTRACT

Motivation: High-throughput screening (HTS) is an early-stage process in drug discovery which allows thousands of chemical compounds to be tested in a single study. We report a method for correcting HTS data prior to the hit selection process (i.e., selection of active compounds). The proposed correction minimizes the impact of systematic errors which may affect the hit selection in HTS. The introduced method, called a *well correction*, proceeds by correcting the distribution of measurements within wells of a given HTS assay. We use simulated and experimental data to illustrate the advantages of the new method compared to other widely-used methods of data correction and hit selection in HTS.

1 INTRODUCTION

High-throughput screening (HTS) is a modern technology used for the identification of pharmacologically active compounds (i.e., hits). In screening laboratories, testing more than 100,000 compounds a day has become routine. Automated mass screening for pharmacologically active compounds is now widely distributed. It serves for the identification of chemical compounds as starting points for optimization (primary screening), for the determination of activity, specificity, physiological and toxicological properties of large libraries (secondary screening), and for the verification of structure-activity hypotheses in focused libraries (tertiary screening) (Heyse 2002). The lack of standardized data validation and quality assurance processes has been recognised as one of the major hurdles for successful implementing high-throughput experimental technologies (Kaul 2005). Therefore, automated quality assessment and data correction systems need to be applied to biochemical data in order to recognize and eliminate experimental artefacts that might confound with important biological or chemical effects. The description of several methods for quality control and correction of HTS data can be found in Zhang et al. (1999 and 2000), Heyse (2002), Heuer et al. (2003), Brideau et al. (2003), Gunter et al. (2003), Kevorkov and Makarenkov (2005), Makarenkov et al. (2006), Malo et al. (2006), and Gagarin et al. (2006a).

Various sources of systematic errors can affect experimental HTS data, and thus introduce a bias into the hit selection process (Heuer et al. 2003), including:

- Systematic errors caused by ageing, reagent evaporation or cell decay which can be recognized as smooth trends in the plate means/medians.
- Errors in liquid handling and malfunction of pipettes which can generate localized deviations from expected values.
- Variation in incubation time, time drift in measuring different wells or different plates, and reader effects which can be recognized as smooth attenuations of measurements over an assay.

Random errors produce noise that cause minor variation of the hit distribution surface. Systematic errors generate repeatable local artifacts and smooth global drifts, which become more noticeable when computing a hit distribution surface (Kevorkov and Makarenkov 2005). Often systematic errors create border, row or columns effects, resulting in the measurements in certain rows or columns that are systematically over or underestimated (Brideau et al. 2003). This paper introduces a new method allowing one to mini-

*To whom correspondence should be addressed.
Email: makarenkov.vladimir(at)uqam.ca

mize the impact of systematic error on the hit selection process. We propose to examine the hit distribution of raw data and fit the data variation within each well to correct the data at hand. The comparison of the new method to other data correction techniques used in HTS is described in the Simulations section. The latter section is followed by an application example, where we carried out the identification of active compounds in the raw and well-corrected HTS assay generated at McMaster University.

2 MATERIALS AND METHODS

Experimental data

In this paper we examine an experimental dataset generated at the HTS Laboratory of McMaster University. This test assay was proposed as a benchmark for the McMaster Data mining and docking competition (the competition website: http://hts.mcmaster.ca/Downloads/82BFBEB4-F2A4-4934-B6A8-804CAD8E25A0.html; see also Elowe et al. 2005). It consists of a screen of compounds that inhibits the *Escherichia coli* dihydrofolate reductase. Each compound was screened twice: two copies of 625 plates were run through the screening machines. This gives 1250 plates in total, each having wells arranged in 8 rows and 12 columns (the columns 1 and 12 containing controls were not considered in this study). The assay conditions reported in Elowe et al. (2005) were the following: Assays were carried out at 25° C and performed in duplicate. Each 200 μL reaction mixture contained 40 μM NADPH, 30 μM DHF, 5 nM DHFR, 50 mM Tris (pH 7.5), 0.01% (w/v) Triton and 10 mM β-mercaptoethanol. Test compounds from the screening library were added to the reaction before initiation by enzyme and at a final concentration of 10 μM. The Supplementary Materials section contains more detail on the screening method and plate layout (Figure 1sm*) for this assay.

Data pre-processing and correction in HTS

The analysis of experimental HTS data requires pre-processing to ensure the statistical meaningfulness and accuracy of the data analysis and the hit selection. Ideally, inactive samples should have similar mean values and generate a constant surface. In a real case, however, random errors produce random noise. For a large number of plates, the noise residuals should compensate each other in the computation of mean values. Systematic repeatable artifacts become more visible as the number of plates increases (Kevorkov and Makarenkov 2005). The following steps can be carried out to pre-process experimental HTS data:

A. Hit and outlier elimination (optional). This elimination can be carried out in order to reduce the influence of hits and outliers on the plates' means and standard deviations. It can be particularly important when analyzing screens with few (< 100) plates.

B. Within-plate normalization of all samples, which can be done including or excluding hits and outliers, using the *Z score transformation* (i.e., zero mean and unit variance standardization, Equation 1) or the *Control normalization* (Equation 2) can be carried out. Such transformations should be applied to analyze together experimental HTS data generated under different testing conditions. In case of *Z score,* the following formula is used:

$$x_i^{\wedge} = \frac{x_i - \bar{x}}{SD}, \qquad (1)$$

where x_i - measured value at well i, $x_i^{\wedge}$ - normalized output value at well i, $\bar{x}$ - mean value, and SD - standard deviation.

The *control normalization* (i.e., normalized percent inhibition) is based on the following formula:

$$x_i^{\wedge} = \frac{H - x_i}{H - L} * 100\%, \qquad (2)$$

where x_i - measured value at well i, H – mean of high controls, L - mean of low controls, and $x_i^{\wedge}$ - evaluated percentage at well i.

The additivity of experimental data is a necessary property that should hold prior to the application of some statistical procedures. Plate means and standard deviations vary substantially from plate to plate. In order to compare and analyze together experimental data from various plates and data tested under different conditions, all measurements should be normalized.

C. Data correction of all samples. This step can be conducted using the median polish procedure (Tukey 1977), the background correction procedure (Kevorkov and Makarenkov 2005), or the well correction method discussed in this article. The B score transformation procedure (Brideau et al. 2003; Malo et al. 2006), additionally correcting for row and column biases, can also be carried out. The comparison of the data correction techniques is presented in the Simulation study section.

Also, background plates (i.e., control plates) can be inserted throughout a screen. Background plates are separate plates containing only control wells and no screening compounds. They are particularly useful for calculating the

* Supplementary materials

background levels of an assay and help determine whether an assay has sufficient signal which can be reliably detected (Fassina 2006). Such additional plates enables one to create background signatures that lead to plate-based correction factors on, per well, per row or per column, basis for all other assay plates. The use of background plates gets around the main assumption made for the well correction procedure: when examined across plates, wells should not systematically contain compounds from the same family (see the description of the well correction method below). The main inconvenience of this method is that it does not take into account possible errors that might occur in the plates processed between two background plates.

Note that for certain methods, the corrected data can be easily denormalized in order to obtain a dataset scaled as the original one. Analysis of the hit distribution surface of the corrected data can then be carried out using the χ-square contingency test (see the results in the section 3.2).

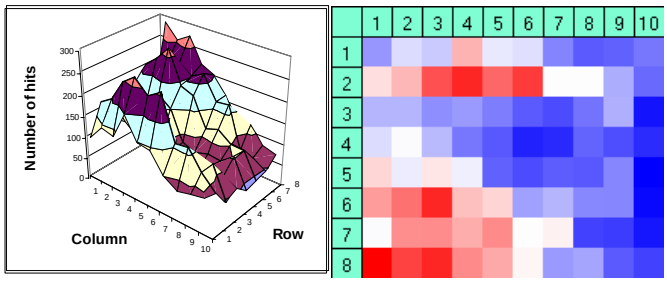
Hit selection process

In the HTS workflow, the bias correction process is followed by hit selection (i.e., inference of active compounds). The selection of hits in the corrected data is often done using a pre-selected threshold (e.g., $\bar{x} - 3SD$, in case of an inhibition assay).

Hit selection is a process that should not only consider statistical treatment of HTS data. It should also be used in conjunction with the structure-activity-relationships (SAR) observed using the corrected HTS data (Gedeck and Willett 2001). In SAR, the basic assumption for all molecule based hypotheses is that similar molecules have similar activities. The quality of hits is improved if SAR is taken into consideration. For instance, the likelihood that an identified hit is an artifact grows if a large number of highly related compounds was confirmed as inactive in the corrected data.

Analysis of hit distribution

The presence of systematic errors in an assay can be detected through the analysis of its hit distribution surface (Kevorkov and Makarenkov 2005). This surface can be computed by estimating the number of selected hits within each well location. In the case of randomly distributed compounds, hits should be distributed evenly over the well locations. In the example presented in Figure 1, we considered the normalized McMaster assay (see Elowe et al. 2005 and Zolli-Juran et al. 2003) comprising 1250 plates arranged in 8 rows and 10 columns (the control columns were not considered). For each well, we estimated the number of experimental values that deviated from the plate means by more than $\bar{x} - SD$ (i.e., number of values that are lower than $\bar{x} - SD$ at each well across plates).



(a) (b)

Figure 1. Hit distribution surface for the McMaster data (1250 plates). Values deviating from the plate means for more than one *SD* were taken into account during the computation. (a) Well positional effects in 3D are shown; (b) Well, row and column positional effects are shown.

As the common strategy utilizes the $\bar{x} - 3SD$ threshold for the hit selection, the considered data comprise all hits as well as all values close to them. The substantial variation of the measurements shown in Figure 1 illustrates the presence of systematic errors in this assay (see the first two columns in Table 5sm for the results of the χ-square contingency test conducted on this surface). The detailed analysis of the McMaster hit distributions obtained for different thresholds is presented in the section 3.2.

Compared methods

The five following methods were compared in this study. Methods 1 and 2 do not involve any data correction, whereas Methods 3 to 5 proceed by the correction of systematic error before hit selection.

Method 1. Classical hit selection using the value $\bar{x} - c * SD$ as a hit selection threshold, where the mean value $\bar{x}$ and standard deviation *SD* are *computed separately for each plate,* and *c* is a preliminary chosen constant. All values lower than or equal to the threshold value are considered as hits.

This method should be applied cautiously when samples are not randomly assigned to plates. Specifically, Method 1 can lead to increased rates of false positives and false negatives when analyzing plates with compounds belonging to the same family (e.g., in case of an inhibition assay, a high concentration of small hit values in a plate can lead to their transformation into false negative hits, whereas a high con-

centration of large values can transform some of the lowest non hit measurements into false positive hits).

Method 2. Classical hit selection using the value $\bar{x} - c * SD$ as a hit selection threshold, where the mean value $\bar{x}$ and standard deviation *SD* are computed *over all assay values*, and *c* is a preliminary chosen constant. This method can be chosen when we are certain that all plates of the given assay were processed under the "same testing conditions".

Method 2 should be applied cautiously when the plates have been tested either over numerous days and/or by different machines (robots, readers, etc). Experience suggests that the "same conditions" requirement may be violated under these circumstances. Moreover, in at least some circumstances, the variability of the various compounds does not appear to be constant but rather follows an inverse gamma distribution (Malo et al. 2006). In the latter case, the pooled variance of Method 2 provides only one component of an individual compound's variance – the other component would be provided by the compound's specific variance as estimated by replicates. However, if the differences among the compound variances are very small, then Method 2 is expected to do well.

Method 3. *Median polish* procedure (Tukey 1977) can be used to remove the impact of systematic error. Median polish (Equation 3) works by alternately removing the row and column medians, and continues until the proportional reduction in the sum of absolute residuals is less than a fixed value ε or until a fixed number of iterations has been carried out. 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 as follows:

$$r_{ijp} = x_{ijp} - \hat{x}_{ijp} = x_{ijp} - (\hat{\mu} + \hat{R}_{ip} + \hat{C}_{jp}). \qquad (3)$$

The residual is defined as the difference between the observed result (x_{ijp}) and the fitted value ($\hat{x}_{ijp}$), which is 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*($\hat{C}_{jp}$). Thus, the matrix of residuals **R** replaces the original matrix in the further computations. In our simulations, Method 2 was applied to the values of the matrix **R** in order to select hits.

Method 4. *B score* (Equation 4) normalization procedure (Brideau et al. 2003) is designed to remove plate row/column biases in HTS (Malo et al. 2006). 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. In addition, for each plate *p*, the adjusted median absolute deviation (MAD_p) is obtained from the r_{ijp}'s. The B score is calculated as follows:

$$B\ score = \frac{r_{ijp}}{(1.4826 * MAD_p)}, \qquad (4)$$

where $MAD_p = median\{|r_{ijp} - median(r_{ijp})|\}$. The raw MAD used in the B score calculation is rescaled by the multiplicative constant of 1.4826. To select hits, this computation was followed by Method 2, applied to the **B score** matrix. Here we considered the version of B score presented in Malo et al. (2006); the latter version of the method does not include the smoothness parameter used by Brideau et al. (2003). The main assumption that must be met to apply Methods 3 and 4 is that the compounds should be randomly distributed within each plate. Any systematic row or column placement of compounds within a plate will bias the results given by these two methods.

Method 5. *Well correction* procedure described below followed by Method 2.

The first two methods are the classical hit selection strategies not involving any correction of systematic bias, whereas the last three methods combine a data preprocessing procedure with the hit selection by Method 2. The results of the median polish and B score methods were generated using the S-PLUS package (S-PLUS manual 2006).

Well correction procedure

To be able to apply the new correction procedure to experimental datasets, the following assumptions about HTS data should be made: screened samples can be divided into active and inactive; the majority of the screened samples are inactive; values of the active samples differ substantially from the inactive ones; and systematic error causes a repeatable influence on the measurements within wells across plates. Also, wells, across plates, should not systematically contain compound samples belonging to the same family. However, it does not require the randomization of samples within plates, which seems to be a much more frequent situation in the real HTS campaigns. We studied the chemical structure of compounds within each well of the McMaster dataset and have not found any systematic pattern in the compound distribution. Usually, each well location contains a large number of samples across plates (e.g., 1250 samples for the McMaster assay), and small systematic compound placements are very unlikely to bias the data.

The Z score normalization produces a modified dataset in which the values within each plate are zero-mean centered, whereas standard deviation and variance are equal to unity. Once the data are plate-normalized, we propose to analyze the values within each well measured across all assay plates. If no systematic error is present in the dataset, the *distribution of measurements within wells* should be also close to a zero-mean centered one with a standard deviation close to unity. The *well correction method* consists of two main steps:

1. *Least-squares approximation of the data carried out separately for each well of the assay.*
2. *Z score normalization of the data within each well location of the assay.*

The real distribution of values can differ substantially from the ideal one. The example presented in Figure 2sm features the measurements obtained for the well located in column 1 and row 8 of the McMaster data (see Elowe et al. 2005). The mean of the observed values is -0.37. Such a deviation suggests the presence of systematic error in this well location. Experimental values for a specific well location can also have ascending or descending trends. The well correction procedure first discovers these trends using the linear least-squares approximation; note that the fitting by a polynomial of a higher degree can be also carried out instead of the linear approximation. Thus, the obtained trend (e.g., a straight line $y = ax + b$ in case of the linear fitting, where x denotes the plate number and y denotes the plate-normalized measurement) is subtracted from or added to the original measurements bringing the mean value of this well to zero. Because the optimal parameters are sought for each well location of the assay independently, well correction has more fitting parameters than B score and Median Polish. For the analysis of large industrial assays, more sophisticated functions (e. g., higher degree polynomials or spline functions) can be also used. Alternatively, an assay can be divided into intervals and a particular trend function characterizing each interval can be determined through approximation.

Second, the well normalization using the Z score normalization (Equation 1) of the well measurements is carried out *independently for each well location*. Then, we can select hits in the corrected data and reexamine the hit distribution surface.

3 RESULTS AND DISCUSSION

Simulation study

To demonstrate the effectiveness of the well correction procedure we first carried out simulations with random data. Specifically, we considered three types of random symmetrically distributed data: standard normal, heavy tailed (positive kurtosis), and light tailed (negative kurtosis) distributions. As with the McMaster data, our datasets consisted of 1250-plate assays, each plate comprising wells arranged in 8 rows and 10 columns. First, three random null datasets (i.e., without hits) were generated. The hit selection procedure was carried out on these data and the false positive hit rates reported in Table 1 were found for the five different hit selection methods presented above.

Hit selection thresholds equal to $\bar{x} - 3SD$ for the standard normal, to $\bar{x} - 2.042SD$ for the light tailed, and to $\bar{x} - 3.420SD$ for the heavy tailed data, were considered. The hit selection thresholds for the light and heavy tailed data were chosen to have approximately the same hit percentage found by the hit selection method based on the assay parameters (Method 2) for the three raw datasets (~ 0.14% of hits; i.e., 140 hits). Since the simulated data did not contain any hits, the hits identified by the methods were false positives by definition. Note that Method 1 was very sensitive to the data distribution; it found the lowest number of false positives in the case of the heavy tailed (83 hits) and standard normal distributions (104 hits), but 642 false positive hits in case of the light tailed data. The median polish and B score methods were unstable, yielding the most false positives (2685 and 2676 for the light tailed data, and 288 and 361 for the heavy tailed data, respectively). The most stable results were obtained by Method 2 and the well correction procedure. As expected, the largest percentage of the false positive hits was found in the light tailed data. For each type of random data we then generated and added to plates k percent of hits, where k was consequently taking values 0.5, 1, 1.5, 2, 2.5, and 3%, whose locations and values were chosen arbitrarily; the probability of each well in each plate to contain a hit was k percent. One thousand replicates of data of each distribution and for each hit percentage were generated. The values of hits were assumed to have a standard normal distribution with the parameters $N(\bar{x} - 5SD, SD)$ for the standard normal, $N(\bar{x} - 4.9SD, SD)$ for the light tailed, and $N(\bar{x} - 5.9SD, SD)$ for the heavy

tailed distributions, respectively, where $\bar{x}$ is the mean value and *SD* is the standard deviation of the observed plate.

Table 1. False positive hit rate for the five pre-processing methods. Random data without hits having standard normal, heavy and light tailed distributions were considered.

Distributions \ Methods	Standard Normal	Light Tailed	Heavy Tailed
Threshold	$\bar{x} - 3SD$	$\bar{x} - 2.042SD$	$\bar{x} - 3.420SD$
Method 1	0.104 %	0.642 %	0.083 %
Method 2	0.140 %	0.138 %	0.137 %
Method 3	0.385 %	2.685 %	0.288 %
Method 4	0.538 %	2.676 %	0.361 %
Method 5	0.138 %	0.292 %	0.121 %

Second, row and column biases were generated as follows. For each row and each column of a given random assay we generated a systematic error that was identical for all assay plates. We also added a small random error to all assay measurements. Therefore, the error-perturbed value, x'_{ijp}, of the measurement in row *i* and column *j* on the *p*-th plate was obtained as follows:

$$x'_{ijp} = x_{ijp} + s_i + s_j + rand_{ijp}, \qquad (5)$$

where x_{ijp} is the observed result in well *ij* of plate *p*, s_i is the systematic error present in row *i*, s_j is the systematic error present in column *j*, and $rand_{ijp}$ is the random error in well *ij* of plate *p* (see also Table 1sm, case a). The variables s_i and s_j in Equation 5 had a standard normal distribution with parameters $N(0, c)$, where the variable *c* was consequently taking the values 0, 0.6*SD*, 1.2*SD*, 1.8*SD*, 2.4*SD*, and 3*SD*. For each value of the variable *c*, a different set of assays was generated and tested. For all values of the parameter *c*, the random error *rand* was always distributed according to a standard normal low with parameters $N(0, 0.6SD)$.

We carried out the five pre-processing methods described above, choosing as hits the measurements with the values lower than $\bar{x} - 3SD$, $\bar{x} - 2.04SD$, and $\bar{x} - 3.42SD$, for the standard normal, light tailed, and heavy tailed data, respectively. Statistics describing the impact of systematic error on the hit selection process are reported in Tables 2sm, 3sm, and 4sm. Specifically, the hit detection rate as well as the false positive and the false negative rates were computed during the simulations. The results in Tables 2sm to 4sm are indicated for the data with 1% of added hits whereas the systematic error varied from 0 to 3.0*SD*. The hit detection rates for the five competing methods corresponding to the systematic error of 1.2*SD* are depicted in Figures 2 and 3sm-4sm. As the tables and graphics suggest, the well correction procedure showed the most stable behavior regardless the data distribution, level of systematic bias, and hit percentage. In all situations, well correction, median polish and B score methods removed systematic error regardless of amount of bias (Figures 2 and 3sm-4sm, cases a and c). However, the well correction procedure generally outperformed the Median Polish and B score methods. The two latter methods were able to remove systematic trend and return the correct residuals in the easiest cases, but they often converged to a local instead of a global minimum when the data structure was rather fuzzy. We can also observe that Method 2 based on the assay parameters was more precise than Method 1 using the plates' parameters. Consequently, when the testing conditions are similar for all plates of the given assay, treating all assay measurements as a whole batch should be preferred to the plate-by-plate analysis. On the other hand, the usage of the well correction procedure can be advocated for any type of data regardless of hit percentage. Compared to the four other methods, the well correction procedure was particularly accurate as to the false negative rate (Tables 2sm to 4sm). At the same time, when the well correction was applied to the data free of noise, it did not have any negative influence to the false positive rate (Table 1).

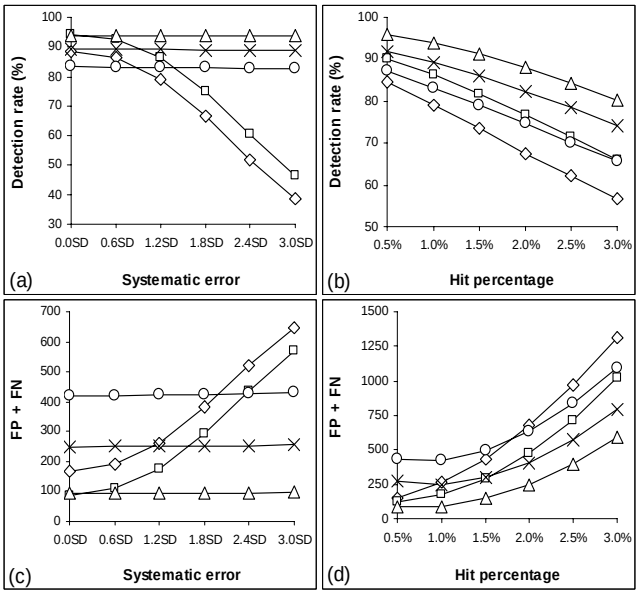


Figure 2 (see also Table 2sm). True hit rate and total of the number of false positive and false negative hits for the noisy standard normal data *with systematic error stemming from row x column interactions which are constant across plates*. The results were obtained with the methods using plates' parameters (i.e., Method 1, ◊), assay parameters (i.e., Method 2, Ä), median polish (x), B score (○), and well correction (Δ). The abscissa axis indicates the noise factor (a and c - with fixed hit percentage of 1%) and the percentage of added hits (b and d - with fixed error rate of 1.2SD).

The B score method with this type of normally distributed constant variance data did not perform well, although the median for the hits was separated from the median for the non-hits to a greater extent than in the well correction method which, in turn, was less separated from the non hits than raw data (see Figure 8sm). This effect was offset, however, by the increased variance for both the non hits and the hits. The B score method improved accuracy somewhat but at the high cost of a large decrease in precision. Accordingly, B score method should not be used unless there is evidence of row or column effects.

We also constructed the ROC curves for the five methods under study. ROC curves provide a graphical representation of the relationship between the true positive and false positive prediction rate of a model. There are many advantages to this approach, including that thresholds do not need to be determined in advance.

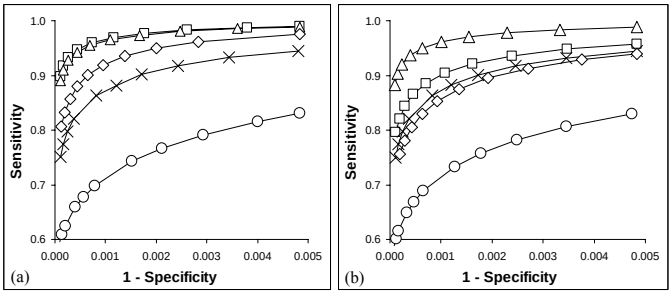


Figure 3. ROC curves for the noisy standard normal data *with systematic error stemming from row x column interactions which are constant across plates*. The results were obtained using: Z score (i.e. Method 1, ◊), raw data (i.e. Method 2, Ä), median polish (x), B score method (○), and the well correction procedure (Δ). The graph (a) corresponds to the case: 1% of added hits and no systematic error; the graph (b) corresponds to the case: 1% of added hits and systematic error of 1.2*SD*.

The *y*-axis corresponds to the *sensitivity* of the model, i.e. how well the model is able to predict true positives (real cleavages); the *y*-coordinates are calculated as follows:

$$Y = \frac{TP}{(TP + FN)}, \qquad (6)$$

where *TP* is the number of true positives and *FN* is the number of false negatives. The *x*-axis corresponds to 1-*specificity*, i.e. the ability of the model to identify true negatives. An increase in specificity (i.e. a decrease along the *x*-axis) results in an increase in sensitivity. The *x*-coordinates are calculated as follows:

$$X = 1 - \frac{TN}{(TN + FP)}, \qquad (7)$$

where TP is the number of true negatives and FP is the number of false positives. The greater the sensitivity at high specificity values (i.e. high y-axis values at low x-axis values) the better the model. A numerical measure of the accuracy of the model can be obtained from the area under the curve, where an area of 1.0 signifies near perfect accuracy, while an area of less than 0.5 indicates that the model is worse than just random. Figure 3 illustrates the ROC curves associated with the five methods compared in this article. The curves were obtained for the standard normal data with 1% of added hits without (a) and with (b) systematic error. The ROC curves confirm the conclusions that can be made while observing the methods' performances reported in Table 2sm and depicted in Figure 2. The well correction procedure and Method 2 provide the best results for data free of systematic bias (Figure 3a), whereas median polish and B-score methods fail to recover correct hits in this situation. After the addition of systematic noise (Figure 3b) well correction procedure outperformed the four other methods, whereas the performances of Methods 1 and 2, not assuming any correction of systematic bias, decreased compared to the case of the error free data.

Finally, for the noisy standard normal data only, we carried out simulations with 4 additional types of error. The data generation diagram for these simulations is presented in Table 1sm (see cases b to e). The following additional error conditions were considered:

b) Systematic error with column effects across plates (Figure 5sm).

c) Systematic error varying from well to well (no row x column interactions involved) across plates (Figure 6sm).

d) Systematic error stemming from row x column interactions and changing from plate to plate (Figure 4).

e) Random error only varying from well to well and from plate to plate (Figure 7sm).

These situations account for the most realistic scenarios, although it is of course not possible to represent all contingencies. For these additional datasets, the well correction procedure was generally more accurate than the four other methods. The only case when the B score method outperformed the well correction procedure was the case where systematic error stemmed from row x column interactions which were changing from plate to plate (i.e., systematic error was not constant across plates, see Figure 4) and this error was sufficiently large (1.2*SD* and more for the true hit

rate, and 2.3*SD* and more for the sum of false positives and false negatives).

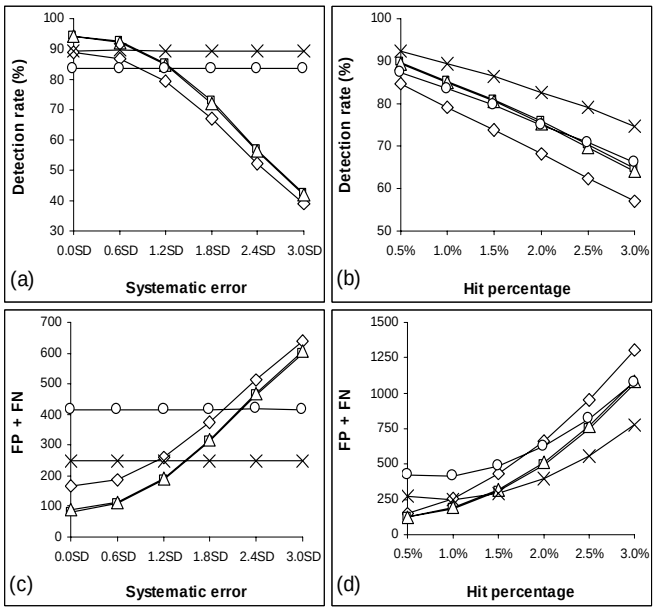


Figure 4. True hit rate and total of the number of false positive and false negative hits for the noisy standard normal data *with systematic error stemming from row x column interactions which are varying across plates*. The same 5 methods as in Figure 2 above are presented.

Well correction of the McMaster data

We fitted the McMaster assay data to a Gaussian distribution (see Figure 9sm). The raw values were first plate normalized using Z scores. The experimental distribution was evaluated by counting the number of measurements in $0.1SD$ intervals in the range $(\bar{x}-5SD\,;\bar{x}+5SD)$. The Gaussian distribution was modeled using the parameters of the experimental one. The hit selection area $(\bar{x}-5SD\,;\bar{x}-3SD)$ is shown in the upper right corner of Figure 9sm.

One popular hit selection method in high-throughput screening proceeds by fixing a constant threshold (usually, $\bar{x}-3SD$) for all considered wells. For an inhibition assay, all measurements that are lower than this threshold are identified as hits. The procedure assumes that the measurement distributions in each well have the same shapes and properties. We verified this assumption while examining the cumulative distribution functions at wells of the McMaster assay (80 functions for 8x10 plates). These functions have a broad band of shapes. These differences can be due to systematic biases and can have a substantial impact on the hit selection procedure.

After the plate normalization by Z scores, the values in each plate are zero-mean centered, and their standard deviation and variance are equal to unity. However, the values in wells, measured across all plates, can have different standard deviations and variances.

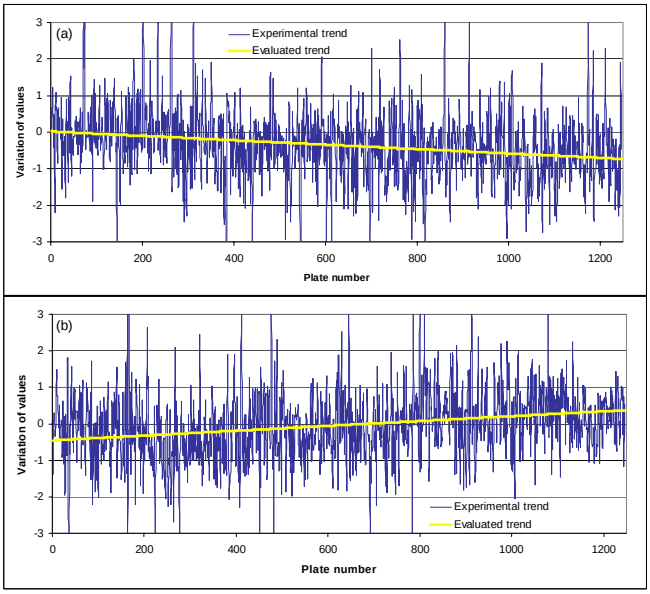


Figure 5. Variation of the plate-normalized measurements in two different wells along 1250 plates of the McMaster assay; descending (a) and ascending (b) trends are highlighted.

The example in Figure 2sm shows that the mean value of the normalized measurements from the well located in column 1 and row 8 (McMaster assay) is -0.37. This deviation is likely to be caused by systematic error. Furthermore, Figure 5 shows that the variation of values within a well can have descending (Figure 5a) and ascending (Figure 5b) trends. Thus, to identify hits in the McMaster assay we applied the classical hit selection algorithm based on the assay mean and standard deviation (Method 2) and the well correction procedure (Method 5). The well correction algorithm first evaluates trends using a least-squares approximation. These trends are removed from the experimental data. Then, the algorithm normalizes the modified assay values within each well separately using Z scores. We examined and compared the hit distribution surfaces obtained using Methods 2 and 5 for different hit selection thresholds. The hit distribution surfaces for the McMaster raw and well-corrected datasets are shown in Figures 6 and 10sm. Figure 6 presents the hit distributions for the following hit selection thresholds ($\bar{x}-SD$, $\bar{x}-1.5SD$, and $\bar{x}-2SD$) and Figure 10sm presents the hit distribution surfaces for the thresholds ($\bar{x}-2.5SD$, $\bar{x}-3SD$, and $\bar{x}-3.5SD$). Each value on the graphic depicts the number of hits found at the associated well. Figures 6 and 10sm suggest that the well correction procedure allows one to attenuate the impact of systematic

bias. The improvements in the hit distribution surfaces are more evident for the bigger hit selection thresholds (Figure 6). For all obtained hit distributions, we also carried out a χ-square contingency test (with the parameter α of 0.01). In our case, the null hypothesis (H_0) assumes that the observed hit distribution is a constant surface. The results of this test are reported in Tables 5sm and 6sm.

Figure 6 (a; raw data) and (b; well-corrected data) depicts the hit distributions for the $\bar{x} - SD$ threshold. The null hypothesis was rejected in both cases (see Table 5sm). However, the well-corrected dataset demonstrated a substantial improvement of the χ-square statistic compared to the raw data. The χ-square value decreased from 2377.4 to 204.6 (with critical value equal to 111.1), i.e., this value for the corrected data is about 11.6 times lower compared to the raw ones. Figure 6 (c; raw data) and (d; well-corrected data) depicts the hit distributions for the $\bar{x} - 1.5SD$ threshold. After the well correction, the χ-square coefficient decreased from 1258.4 to 173.6; i.e., it is about 7 times lower for the well-corrected data compared to the raw ones, but it was still larger than the χ-square critical value (111.1). Figure 6 (e; raw data) and (f; well-corrected data) shows the hit distribution surfaces for the $\bar{x} - 2SD$ threshold. The χ-square contingency test failed to reject the null hypothesis (H_0) for the corrected data (χ-square value of 74.8) and rejected it for the raw data (χ-square value of 438.6). Figure 10sm illustrates the hit distribution surfaces obtained for the raw and well-corrected data for commonly used hit selection thresholds. In this study, the following thresholds were employed to identify hits in the raw and corrected McMaster data: ($\bar{x} - 2.5SD$, $\bar{x} - 3SD$, and $\bar{x} - 3.5SD$). The null hypothesis, postulating that the hit distribution corresponds to a constant surface, was not rejected for the well-corrected data in case of all three considered hit selection thresholds (Table 6sm). For the raw data, the null hypothesis was rejected for all considered thresholds, except $\bar{x} - 3.5SD$, for which the values of the χ-square coefficient for the raw and corrected data were close (106.2 and 105.3, respectively). This is certainly due to the decrease in the number of hits when lowering the hit selection threshold. The mean numbers of hits per well for the well-corrected dataset were usually slightly lower than for the raw data (Tables 5sm and 6sm). Tables 7sm and 8sm report the hit numbers per well in the raw and well-corrected McMaster data computed for the $\bar{x} - 3SD$ threshold. Even though for the corrected dataset the null hypothesis was rejected for the thresholds $\bar{x} - SD$ and $\bar{x} - 1.5SD$, the well correction procedure led to an important reduction of the χ-square statistic.

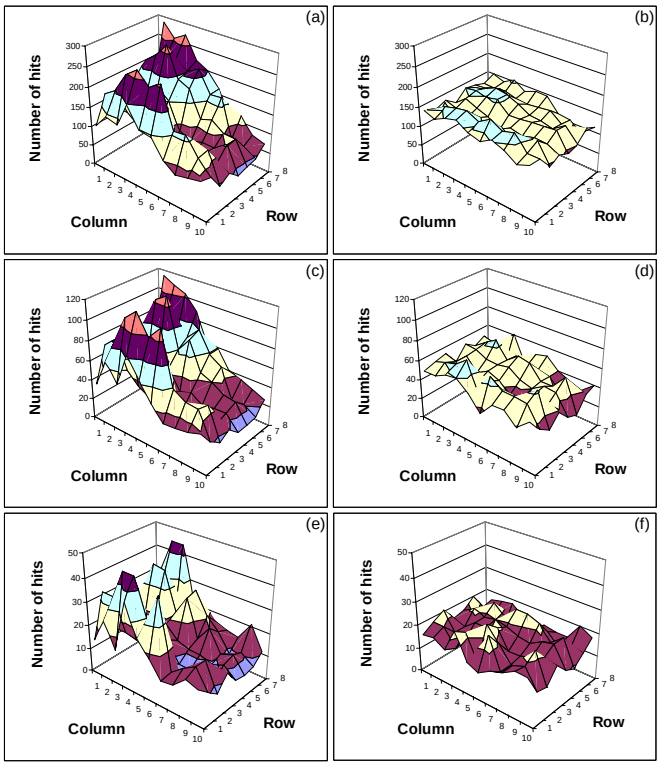


Figure 6. Hit distributions for the raw (a, c, and e) and well-corrected (b, d, and f) McMaster data obtained for the thresholds $\bar{x} - SD$ (a and b), $\bar{x} - 1.5SD$ (c and d), and $\bar{x} - 2SD$ (e and f).

We also analyzed the list of active compounds from the Test Library of the McMaster dataset. The samples in the original dataset were identified as Consensus hits by the organizers of the McMaster HTS competition if both of their replicate measurements were lower or equal to 75% with respect to the reference controls. Only 42 of 50 000 different tested compounds indicated by their MAC-IDs in Table 9sm satisfied this property. Our experiments showed that the selection of these 42 replicate compounds can be reached by carrying out Method 1 on the non-normalized data (hit selection by plates, where the values lower than $\bar{x} - c * SD$ are identified as hits) with the standard deviation coefficient c = 2.29. Among the 42 consensus hits identified in such a manner, the competition organizers also identified 14 compounds having well behaved dose-response curves (they are highlighted in Table 9sm). We also performed the analysis of the original data set using the well correction procedure which was carried out prior to the selection of hits (Method 5 with hit selection carried out by plate). All other parameters were identical to those of Method 1. With these parameters the application of Method 5 led to the

identification of 40 hits. Among these hits, 31 were those found by Method 1 (see Table 9sm), and 9 hits were new. Note that the proportion of compounds with well behaved dose-response curves was better for the well-corrected data (~42%; this percentage does not include the 9 new compounds for which the follow up tests were not conducted) than for the raw data (~33%). However, a more detailed study comparing the dose-response behaviour of the hits identified by all the five competing methods was not possible because the dose-response follow up information is not available for the 9 hits found by Method 5. To conduct a comprehensive study of the five methods compared in this manuscript an experimental dose-response follow-up of the hits obtained using each of these methods is certainly necessary.

It would be quite practical to have the benchmark datasets for which all the results, including the confirmed hits, and testing conditions are known. We think that the scientists from the HTS Laboratory of McMaster University who organized the HTS Data Mining and Docking competition are on the way of doing it: after the announcement of the competition results a special issue of Journal of Bimolecular Screening was dedicated to the analysis of the test data set (see Elowe et al. 2005 and the articles in the same issue).

CONCLUSION

We described a method that can be used to refine the analysis of experimental HTS data by eliminating systematic biases from them prior to the hit selection procedure. The proposed method, called a *well correction,* rectifies the distribution of assay measurements by normalizing data within each considered well across all assay plates. Simulations were carried out with standard normal, heavy and light tailed random datasets. They suggest that the well correction procedure is a robust method that should be used prior to the hit selection process. Well correction generally outperformed the Median polish and B score methods as well as the classical hit selection procedure. In the situations when neither hits nor systematic errors were present in the data, the well correction method showed the performance similar to the traditional method of hit selection. The well correction method also compares advantageously (see Gagarin et al. 2006b) to the background correction procedure (Kevokov and Makarenkov 2005). In the future, it would be interesting to examine how robust the methods are to violations of normality of HTS data.

We also examined an experimental assay generated at the HTS Laboratory of McMaster University. The analysis of the hit distribution of the raw McMaster dataset showed the presence of systematic errors. The McMaster data were processed using different hit selection thresholds varying from $\bar{x} - SD$ to $\bar{x} - 3.5SD$. Note that for all considered thresholds, except $\bar{x} - 3.5SD$, for the raw McMaster data, the χ-square contingency test rejected the null hypothesis, postulating that the hit distribution surface is a constant. The analysis of the well-corrected datasets showed that the new method considerably smoothes the hit distribution surfaces for the $\bar{x} - SD$ and $\bar{x} - 1.5SD$ thresholds. When applied to the well-corrected dataset, the χ-square contingency test failed to reject the null hypothesis for the thresholds $\bar{x} - 2SD$ to $\bar{x} - 3.5SD$. Furthermore, the simulation study also confirmed that Method 2 based on the assay parameters was more accurate than Method 1 based on the plates' parameters. Therefore, in case of identical testing conditions for all plates of the given assay, all assay measurements should be treated as a single batch.

The HTS Corrector software (Makarenkov et al. 2006, http://www.labunix.uqam.ca/~makarenv/hts.html), including the methods for data pre-processing and correction of systematic error, was developed. HTS Corrector includes all data correction methods discussed is this article (Well correction, B score, and Median polish) as well as different methods of hit selection (e.g., Methods 1 and 2 compared in this study). Note that for large industrial assays, a procedure allowing one to divide assays into homogeneous sub-assays has been included in the program. HTS Corrector first establishes a user-defined distance measure between plates and carries out a k-means partitioning algorithm (MacQueen 1967; Legendre and Legendre 1998) to form *k* homogeneous sub-assays.

ACKNOWLEDGEMENTS

We thank Genome Quebec for funding this project. We also thank two anonymous referees for their helpful comments.

REFERENCES

Brideau,C., Gunter,B., Pikounis,W., Pajni,N. and Liaw A. (2003) Improved statistical methods for hit selection in HTS. *J. Biomol. Screen.*, **8**, 634-647.

Elowe,N.H., Blanchard,J.E. Cechetto,J.D. and Brown,E.D. (2005) Experimental Screening of Dihydrofolate Reductase Yields a "Test Set" of 50,000 Small Molecules for a Computational Data-Mining and Docking Competition, *J. Biomol. Screen.,* **10**, 653 - 657.

Fassina,G. (2006) High-Throughput Screening of Combinatorial Libraries. Survey of Applications of Combinatorial Technologies. *Training Course Presentation*, URL: http://www.ics.trieste.it/Documents/Downloads/df3983.pdf.

Gagarin,A., Makarenkov,V. and Zentilli,P. (2006a) Clustering techniques to improve the hit selection in HTS, *J. Biomol. Screen.,* **11**, 903-914.

Gagarin,A., Kevorkov,D. and Makarenkov,V. (2006b) Comparison of two methods for detecting and correcting systematic error in HTS data. In *Data Science and Classification*, V. Batagelj, H.H. Bock, A. Ferligoj, and A. Ziberna (Eds.), IFCS 2006, Studies in Classification, Data Analysis, and Knowledge Organization, Springer Verlag, 241-249.

Gedeck,P. and Willett, P. (2001) Visual and computational analysis of structure-activity relationships in high-throughput screening data. *Curr. Opin. Chem. Biol.* 5, 389 – 395.

Gunter,B., Brideau,C., Pikounis,B., Pajni,N. and Liaw,A. (2003) Statistical and graphical methods for quality control determination of HTS data. *J. Biomol. Screen.*, **8**, 624-633.

Heuer,C., Haenel,T., Prause,B. (2003) A novel approach for quality control and correction of HTS data based on artificial intelligence. *Pharmaceutical discovery & development report*. PharmaVentures.

Heyse,S. (2002) Comprehensive analysis of high-throughput screening data. In *Proc. of SPIE* 2002, **4626**, 535-547.

Kaul,A (2005) The impact of sophisticated data analysis on the drug discovery process. *Business briefing: future drug discovery 2005*.

Legendre,P. and Legendre,L. (1998) *Numerical ecology*. 2nd English ed. Amsterdam: Elsevier Science BV, 739-746.

Kevorkov,D. and Makarenkov,V. (2005) Statistical analysis of systematic errors in HTS. *J. Biomol. Screen.*, **10**, 557-567.

MacQueen,J. (1967) Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 1, Statistics.* L. M. Le Cam and J. Neyman (Eds.), University of California Press.

Makarenkov,V., Kevorkov,D., Zentilli, P., Gagarin,A., Malo,N. and Nadon,R. (2006) HTS-Corrector: new application for statistical analysis and correction of experimental data, *Bioinformatics*, **22**, 1408-1409.

Malo,N, Hanley,J.A, Cerquozzi,S. Pelletier,J. and Nadon R (2006) Statistical practice in high-throughput screening data analysis. *Nature Biotechnol.*, **24**, 167-175.

S-PLUS 6. (2006). S-plus programmer's guide, Insightful, URL : http://www.insightful.com/support/doc_splus_win.asp.

Tukey,J.W. (1977) *Exploratory Data Analysis.* Cambridge, MA: Addison-Wesley.

Zhang,J.H, Chung,T.D.Y. and Oldenburg,K.R. (1999) A simple statistic parameter for use in evaluation and validation of HTS assays. *J. Biomol. Screen.*, **4**, 67-73.

Zhang,J.H, Chung,T.D.Y and Oldenburg,K.R. (2000) Confirmation of primary active substances from HTS of chemical and biological populations: a statistical approach and practical considerations. *J. Comb. Chem.*, **2**, 258-265.

Zolli-Juran,M., Cechetto,J.D., Hartlen,R., Daigle,D.M and Brown,E.D. (2003) HTS identifies novel inhibitors of *Escherichia coli* dihydrofolate reductase that are competitive with dihydrofolate. *Bioorg. Med. Chem. Lett.*, **13**, 2493-2496.

Supplementary Materials

Screening Method

The high throughput screen of *E. coli* dihydrofolate reductase (DHFR) against 50,000 small molecules from ChemBridge Corporation was considered. The screen was performed at the McMaster High Throughput Screening Laboratory in duplicate in 96-well plates using the Beckman-Coulter Integrated Robotic System.

The statistical parameters Z and Z' (Zhang *et al*. 1999) for the screen were 0.57 and 0.72 respectively. These values were comparable to those calculated for the previously reported screen of DHFR against a 50,000 small molecule library from Maybridge plc (Zolli-Juran *et al.* 2003 and Elowe *et al.* 2005).

The following information about the screening procedure can be found in the McMaster HTS laboratory report available on the competition web site: http://hts.mcmaster.ca/Downloads/82BFBEB4-F2A4-4934-B6A8-804CAD8E25A0.html.

Once an assay plate was transferred to the SpectraMax Plus, it was shaken for 5 seconds and each well was read at 340 nm every 15 seconds for 5 minutes (without shaking between reads). All raw data were transferred directly to Activity Base for analysis. Three different controls, High, Low, and Reference, were used in the screen as outlined in Figure 1sm below. For each of these controls, library compounds were excluded from the assay reaction and replaced with: (i) High controls: 2 µL DMSO; (ii) Low controls: 2 µL of 150 µM trimethoprim in DMSO; Reference controls: 2 µL of 1.2 µM trimethoprim in DMSO.

- Enzymatic activity was calculated in Activity Base by the slope of the 7 data points between 20-130 seconds (inclusive) of the 5 minute kinetic read.
- Percent residual activity was calculated using a variant of Formula (2):

$$x_i^{\circ} = \frac{x_i - L}{H - L} * 100\%,$$

where x_i - measured value at well i, H – mean of high controls, L - mean of low controls, and x_i° - evaluated percentage at well i.

The detailed description of the hit selection procedure can be found in the McMaster procedure report.

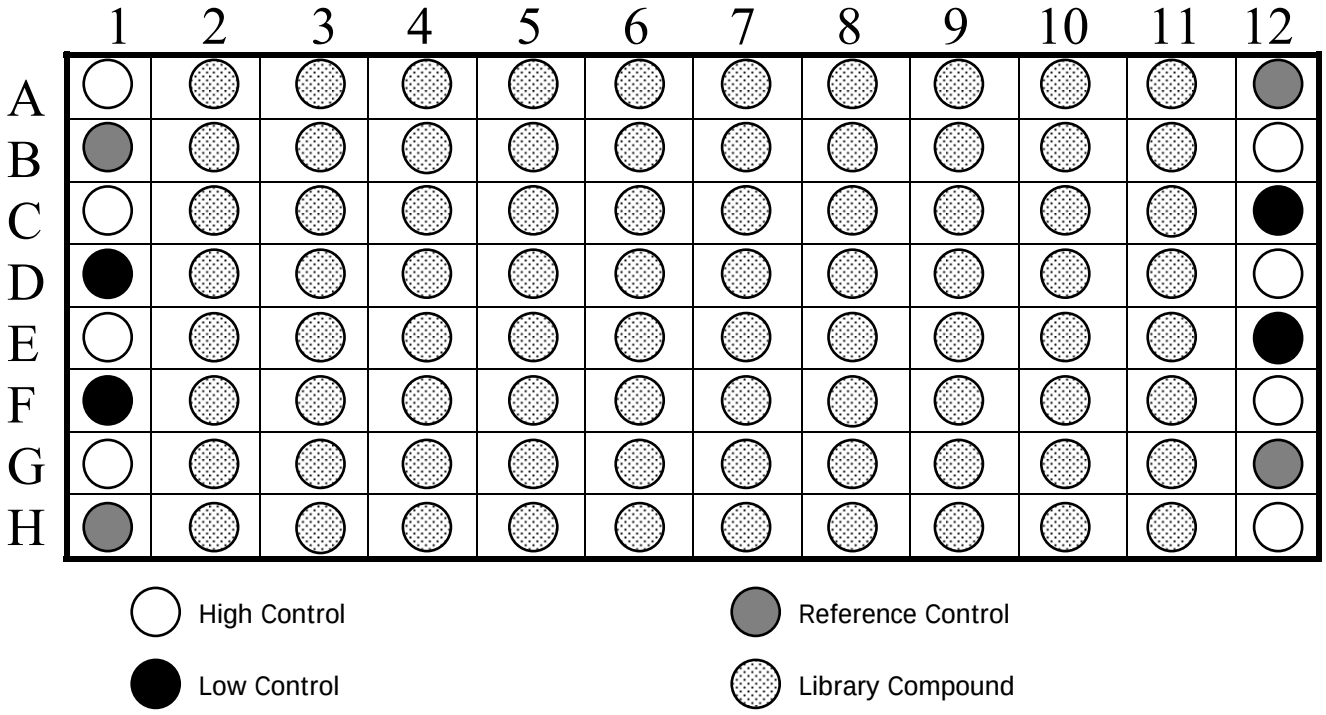


Figure 1sm. Plate layout of the McMaster test assay.

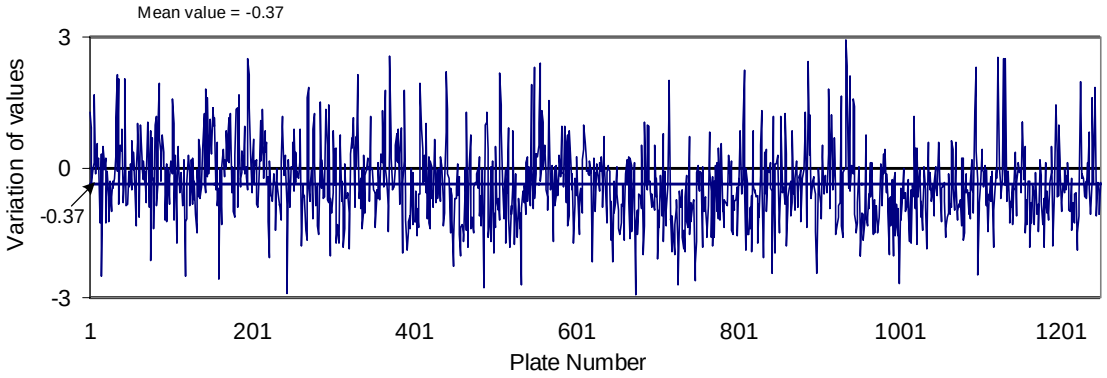


Figure 2sm. Variation of plate normalized values across different plates for the well located in column 1 and row 8 of the McMaster dataset (measured over 1250 plates).

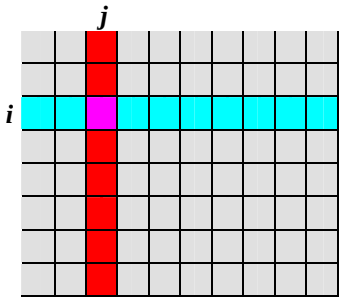
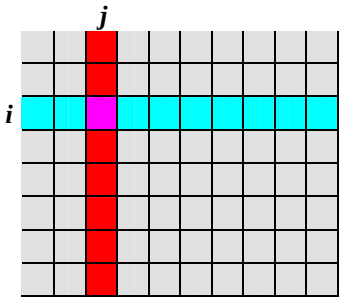
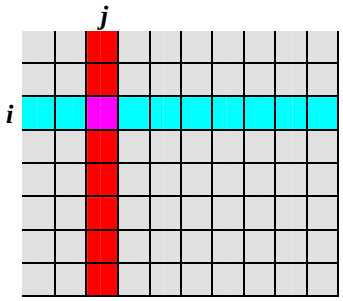
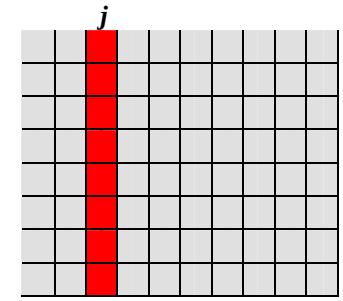
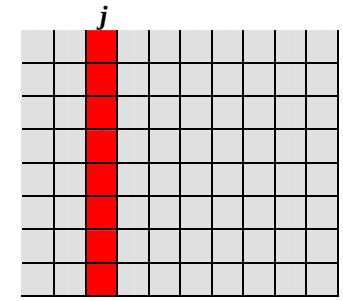
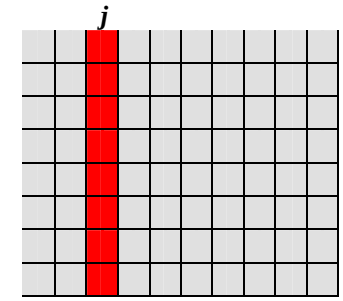
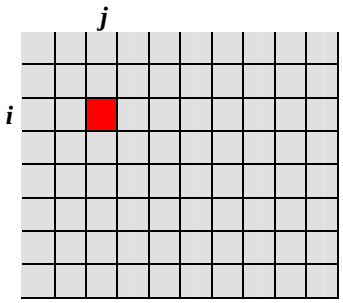
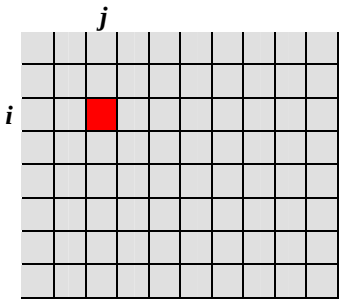
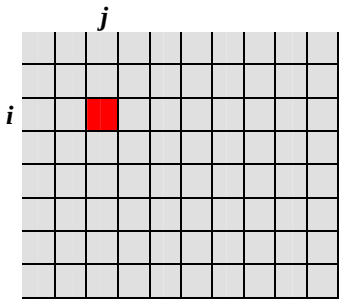
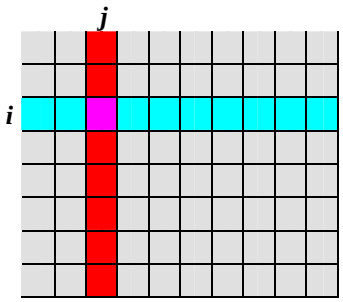
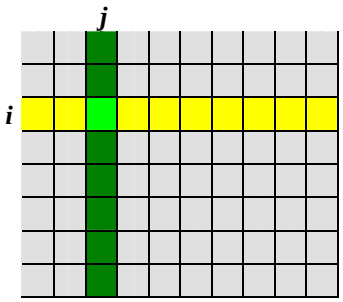
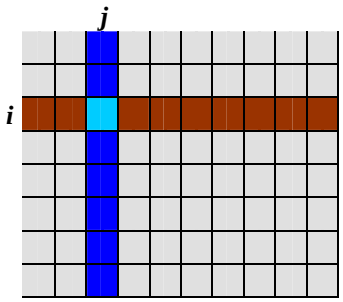
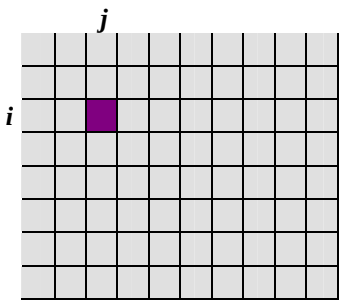
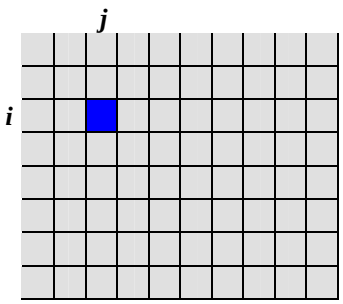
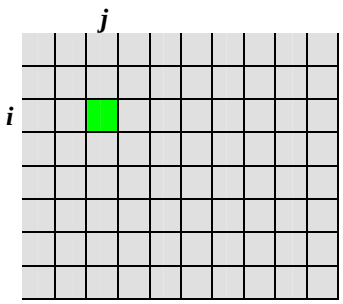
	Plate p	Plate $p + 1$	Plate $p + 2$
(a - Figure 2) $x'_{ijp} = x_{ijp} + s_i + s_j + rand_{ijp}$, $1 \le i \le 8; 1 \le j \le 10; 1 \le p \le 1250$			
(b - Figure 5sm) $x'_{ijp} = x_{ijp} + s_j + rand_{ijp}$, $1 \le i \le 8; 1 \le j \le 10; 1 \le p \le 1250$			
(c - Figure 6sm) $x'_{ijp} = x_{ijp} + s_{ij} + rand_{ijp}$, $1 \le i \le 8; 1 \le j \le 10; 1 \le p \le 1250$			
(d - Figure 4) $x'_{ijp} = x_{ijp} + s_{ip} + s_{jp} + rand_{ijp}$, $1 \le i \le 8; 1 \le j \le 10; 1 \le p \le 1250$			
(e - Figure 7sm) $x'_{ijp} = x_{ijp} + Rand_{ijp}$, $1 \le i \le 8; 1 \le j \le 10; 1 \le p \le 1250$			

Table 1sm: Schematic diagram of a set of "plates" for each of the 5 manipulations used to simulate systematic (or random) error. (a-d) Colored locations represent typical bias + random error effects; (e) Colored locations represent typical random error effects. Each row and each column (cases a, b and d) as well as each well (cases c and e) of each plate (gray colored locations) were also affected by this kind of systematic and random errors.

Diagram caption (Table 1sm)

(a) Systematic errors stemming from *row x column interactions*: Different constant values were applied to each row and each column of the first plate. The same constants were added to the corresponding rows and columns of all other plates. (b) Systematic error stemming from *column effects*: Different constant values were applied to each column of the first plate. The same constants were added to the corresponding columns of all other plates. (c) Systematic error stemming from *well effects*: Different values were added to each well of the first plate. These values were constant for each well across all plates. (d) Systematic error stemming from *changing row x column interactions*: As in (a) but with the values of the row and column constants varying across plates. (e) Random error only (present in each well of each plate and affecting all wells differently).

The error-perturbed value, x'_{ijp}, of the measurement in row i and column j on the p-th plate was obtained using the formulas indicated in the left, where x_{ijp} is the correct measurement in well ij of plate p, s_i is the systematic error affecting row i, s_j is the systematic error affecting column j, s_{ij} is the systematic error affecting well located on the intersection of line i and column j, s_{ip} is and the systematic error affecting line i of plate p, s_{jp} is the systematic error affecting column j of plate p, $rand_{ijp}$ and $Rand_{ijp}$ are the random error affecting well ij of plate p.

The variables s_i, s_j, s_{ij}, s_{ip} and s_{jp} (see also Equation 5 in the manuscript) had a standard normal distribution with parameters $N(0, c)$, where c equals 0, $0.6SD$, $1.2SD$, $1.8SD$, $2.4SD$, or $3SD$ for the various simulation conditions and SD is the standard deviation of the variable x_{ijp} distributed according to a standard normal distribution. For all values of the parameter c, the random error *rand* (cases a to d) was always distributed according to a standard normal distribution with parameters $N(0, 0.3SD)$. The random error *Rand* (case e) was distributed according to a standard normal distribution with parameters $N(0, 1.2SD)$.

Table 2sm (see also Figure 2 in the manuscript)**.** True, false positive and false negative hit detection rates for the 5 methods used to process the random *standard normal data* with 1% of added hits. The hit detection rates were obtained by dividing the number of the true (or false positive, or false negative) hits by the total number of generated hits.

Methods \ Systematic error		0	0.6*SD*	1.2*SD*	1.8*SD*	2.4*SD*	3.0*SD*
Hit detection rate (in %)	1. $\bar{x} - 3SD$ per plate	88.65	86.37	79.12	66.52	51.90	38.74
	2. $\bar{x} - 3SD$ per assay	94.21	92.47	86.41	74.78	60.37	46.61
	3. Median polish	89.24	89.16	89.21	88.98	88.94	88.75
	4. B score	83.34	83.26	83.25	83.03	82.90	82.69
	5. Well correction	93.91	93.91	93.95	93.83	93.78	93.66
False positive rate (in %)	1. $\bar{x} - 3SD$ per plate	5.41	5.34	5.21	4.71	4.02	3.30
	2. $\bar{x} - 3SD$ per assay	2.79	3.39	3.96	4.04	3.83	3.36
	3. Median polish	14.30	14.31	14.33	14.30	14.29	14.27
	4. B score	25.38	25.39	25.49	25.57	25.53	25.57
	5. Well correction	3.13	3.12	3.16	3.19	3.20	3.25
False negative rate (in %)	1. $\bar{x} - 3SD$ per plate	11.35	13.63	20.88	33.48	48.09	61.26
	2. $\bar{x} - 3SD$ per assay	5.79	7.53	13.59	25.22	39.63	53.39
	3. Median polish	10.75	10.84	10.79	11.02	11.06	11.25
	4. B score	16.66	16.74	16.75	16.97	17.10	17.31
	5. Well correction	6.09	6.09	6.04	6.17	6.22	6.34

Table 3sm (see also Figure 3sm). True, false positive and false negative hit detection rates for the 5 methods used to process the random *heavy tailed data* with 1% of added hits. The hit detection rates were obtained by dividing the number of true (or false positive, or false negative) hits by the total number of generated hits.

Methods \ Systematic error		0	0.6*SD*	1.2*SD*	1.8*SD*	2.4*SD*	3.0*SD*
Hit detection rate (in %)	1. $\bar{x} - 3SD$ per plate	93.24	91.84	87.33	79.49	69.15	58.82
	2. $\bar{x} - 3SD$ per assay	94.67	93.45	89.37	82.02	72.00	61.89
	3. Median polish	87.93	87.91	87.80	87.77	87.64	87.50
	4. B score	84.66	84.61	84.57	84.47	84.37	84.18
	5. Well correction	94.57	94.56	94.50	94.52	94.44	94.30
False positive rate (in %)	1. $\bar{x} - 3SD$ per plate	35.84	60.93	95.94	120.02	131.36	136.64
	2. $\bar{x} - 3SD$ per assay	0.00	36.72	86.55	117.76	132.83	140.02
	3. Median polish	176.20	176.07	176.36	177.12	178.10	178.62
	4. B score	189.31	189.19	189.12	189.81	190.79	191.07
	5. Well correction	3.25	3.47	4.11	5.34	7.23	9.71
False negative rate (in %)	1. $\bar{x} - 3SD$ per plate	6.76	8.16	12.67	20.51	30.85	41.18
	2. $\bar{x} - 3SD$ per assay	5.33	6.55	10.63	17.98	28.00	38.11
	3. Median polish	12.07	12.09	12.20	12.23	12.36	12.49
	4. B score	15.34	15.39	15.43	15.53	15.63	15.82
	5. Well correction	5.43	5.44	5.50	5.48	5.56	5.70

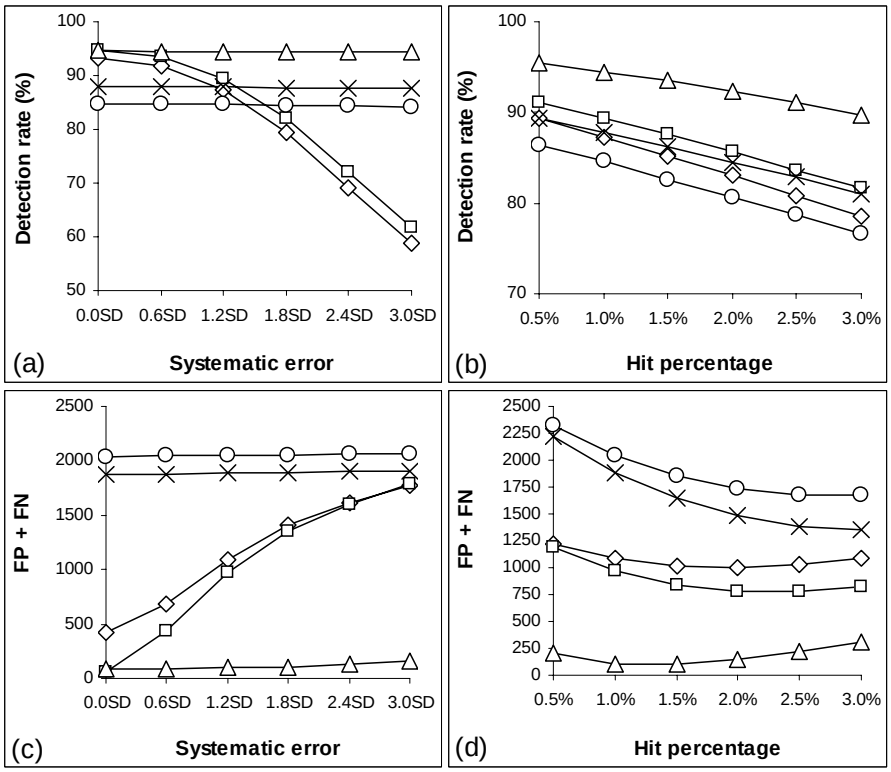


Figure 3sm (see also Table 3sm). True hit rate and total of the number of false positive and false negative hits for the noisy *heavy tailed data* obtained with the methods using plates' parameters (i.e., Method 1, ◊), assay parameters (i.e., Method 2, Ä), median polish (x), B score method (○), and the well correction procedure (Δ). The abscissa axis indicates the noise factor (a and c - with fixed hit percentage of 1%) and the percentage of added hits (b and d - with fixed error rate of 1.2*SD*).

Table 4sm (see also Figure 4sm). True, false positive, and false negative hit detection rates for the 5 methods used to process the random *light tailed data* with 1% of added hits. The hit detection rates were obtained by dividing the number of true (or false positive, or false negative) hits by the total number of generated hits.

Methods \ Systematic error		0	0.6*SD*	1.2*SD*	1.8*SD*	2.4*SD*	3.0*SD*
Hit detection rate (in %)	1. $\bar{x} - 3SD$ per plate	85.24	82.63	74.01	59.91	43.77	30.45
	2. $\bar{x} - 3SD$ per assay	94.50	92.57	85.50	72.18	55.65	40.86
	3. Median polish	90.78	90.77	90.69	90.59	90.43	90.25
	4. B score	82.88	82.90	82.81	82.67	82.50	82.33
	5. Well correction	94.04	94.01	93.96	93.91	93.78	93.62
False positive rate (in %)	1. $\bar{x} - 3SD$ per plate	3.89	3.53	2.83	2.01	1.29	0.85
	2. $\bar{x} - 3SD$ per assay	0.00	0.24	1.13	1.34	1.17	0.89
	3. Median polish	7.11	7.08	7.18	7.17	7.18	7.19
	4. B score	14.46	14.46	14.49	14.37	14.26	14.24
	5. Well correction	0.12	0.13	0.14	0.14	0.17	0.20
False negative rate (in %)	1. $\bar{x} - 3SD$ per plate	14.76	17.38	25.99	40.09	56.23	69.55
	2. $\bar{x} - 3SD$ per assay	5.50	7.43	14.50	27.82	44.35	59.14
	3. Median polish	9.22	9.23	9.31	9.40	9.57	9.75
	4. B score	17.12	17.10	17.19	17.33	17.50	17.67
	5. Well correction	5.96	5.99	6.04	6.09	6.22	6.38

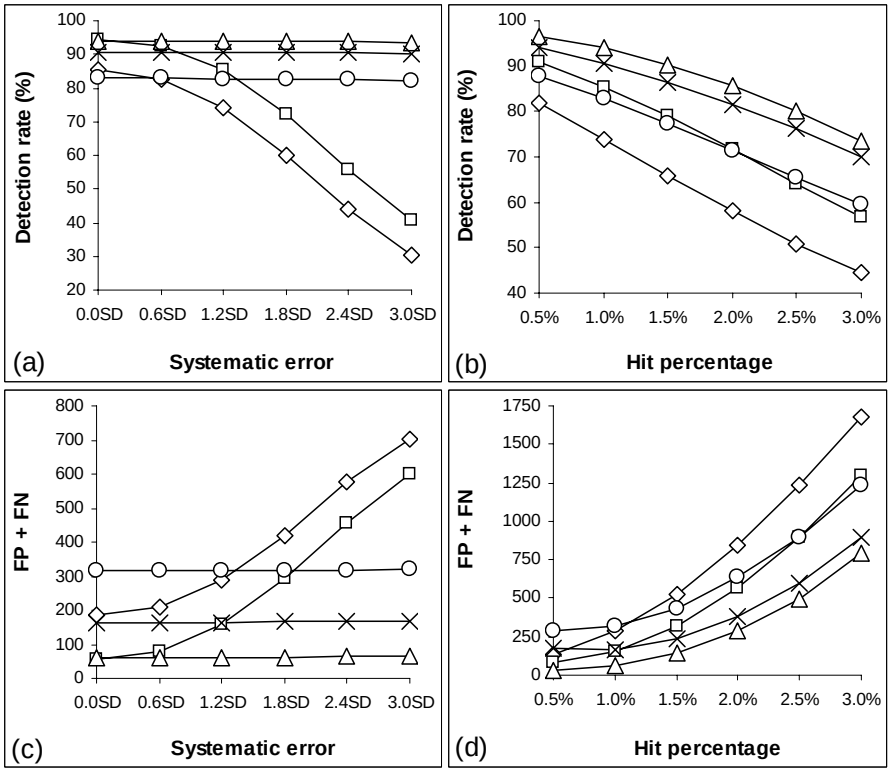


Figure 4sm (see also Table 4sm). True hit rate and total of the number of false positive and false negative hits for the noisy *light tailed data* obtained with the methods using plates' parameters (i.e., Method 1, ◊), assay parameters (i.e., Method 2, Ä), median polish (x), B score method (○), and the well correction procedure (Δ). The abscissa axis indicates the noise factor (a and c - with fixed hit percentage of 1%) and the percentage of added hits (b and d - with fixed error rate of 1.2*SD*).

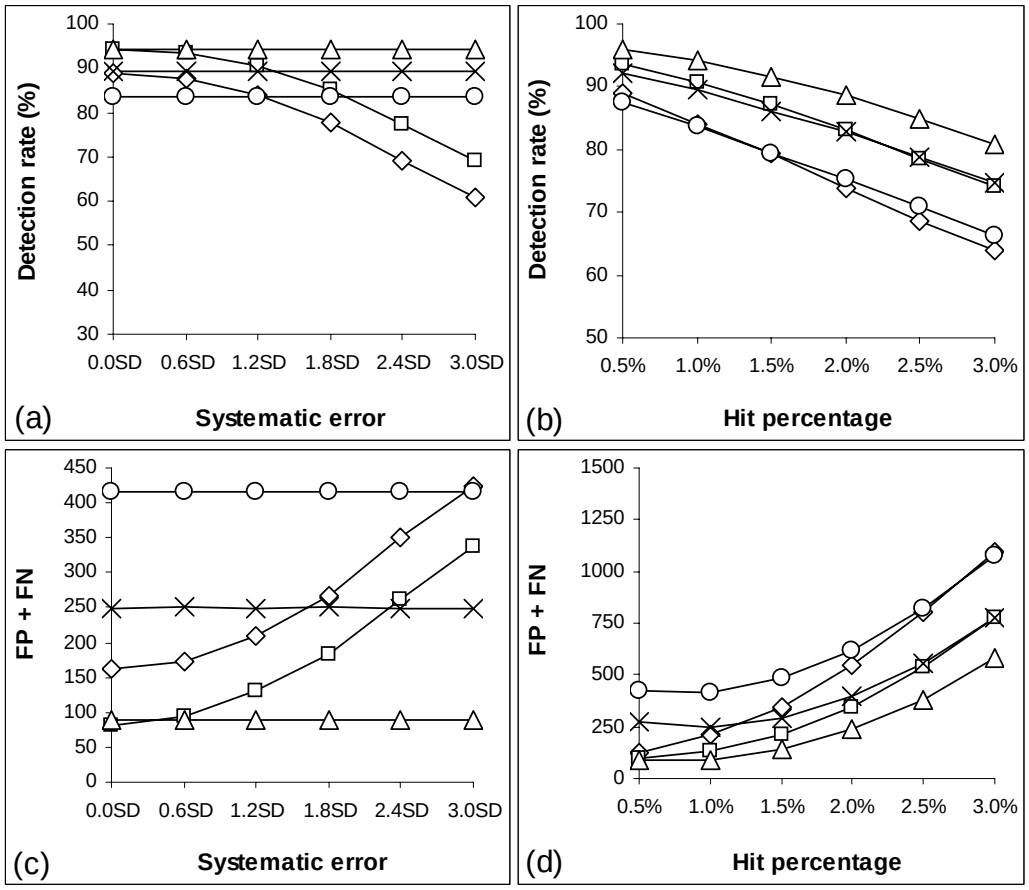


Figure 5sm. True hit rate and total of the number of false positive and false negative hits for the noisy *standard normal data with systematic error stemming from column effects only*. The results were obtained with the methods using plates' parameters (i.e., Method 1, ◊), assay parameters (i.e., Method 2, Ä), median polish (x), B score method (○), and the well correction procedure (Δ).The abscissa axis indicates the noise factor (a and c - with fixed hit percentage of 1%) and the percentage of added hits (b and d - with fixed error rate of 1.2*SD*).

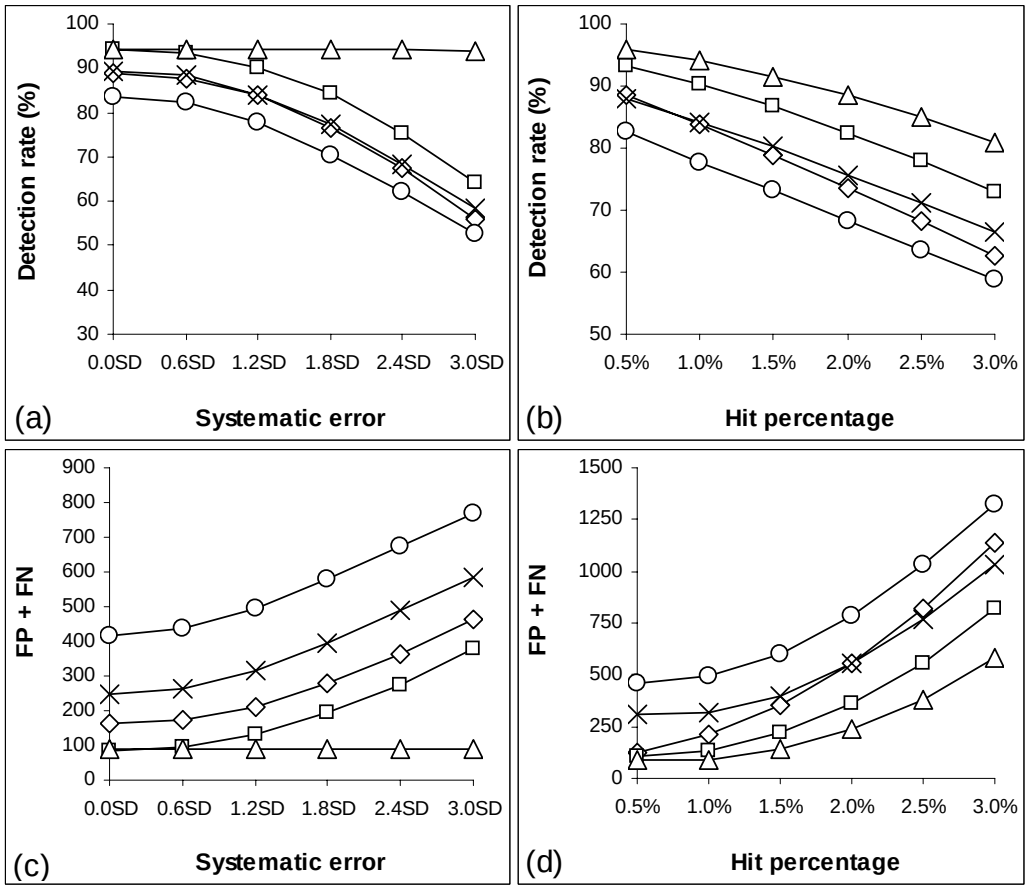


Figure 6sm. True hit rate and total of the number of false positive and false negative hits for the noisy *standard normal data with systematic error different for all wells (no row x column interactions was involved)*. The results were obtained with the methods using plates' parameters (i.e., Method 1, ◊), assay parameters (i.e., Method 2, Ä), median polish (x), B score method (○), and the well correction procedure (Δ). The abscissa axis indicates the noise factor (a and c - with fixed hit percentage of 1%) and the percentage of added hits (b and d - with fixed error rate of 1.2*SD*).

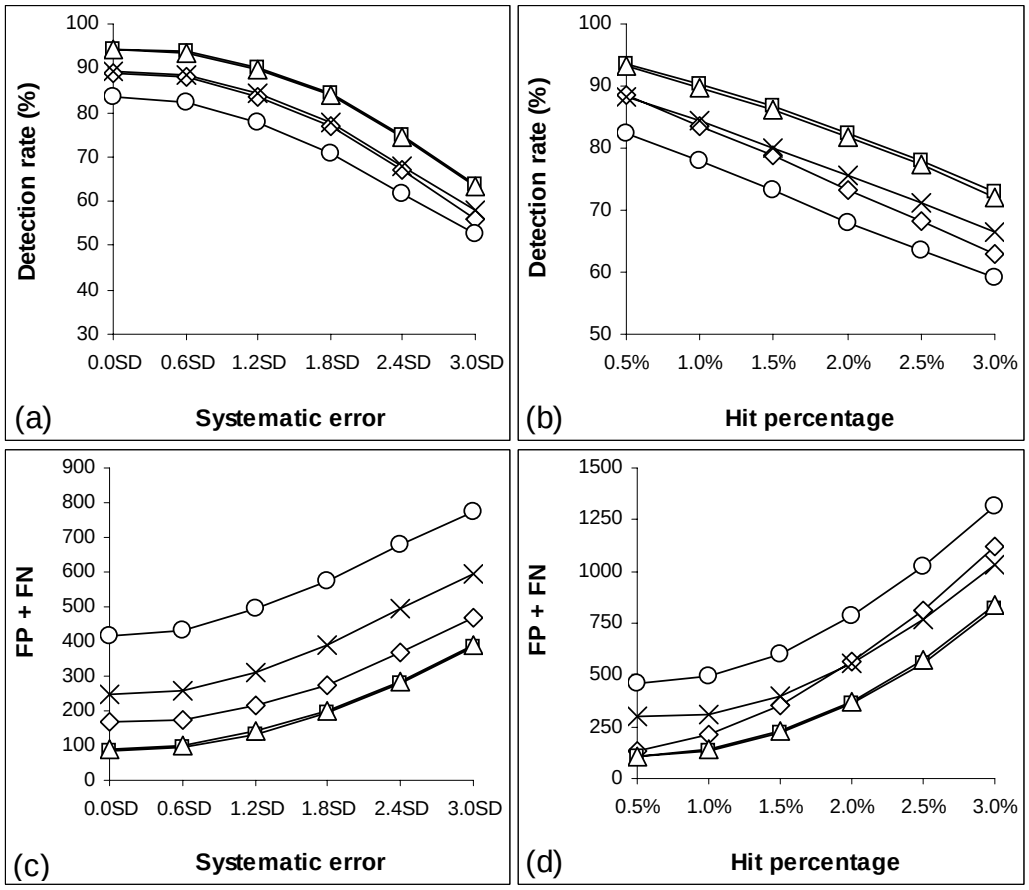


Figure 7sm. True hit rate and total of the number of false positive and false negative hits for the noisy *standard normal data with random error only stemming and varying from plate to plate*. The results were obtained with the methods using plates' parameters (i.e., Method 1, ◊), assay parameters (i.e., Method 2, Ä), median polish (x), B score method (○), and the well correction procedure (Δ). The abscissa axis indicates the noise factor (a and c - with fixed hit percentage of 1%) and the percentage of added hits (b and d - with fixed error rate of 1.2*SD*).

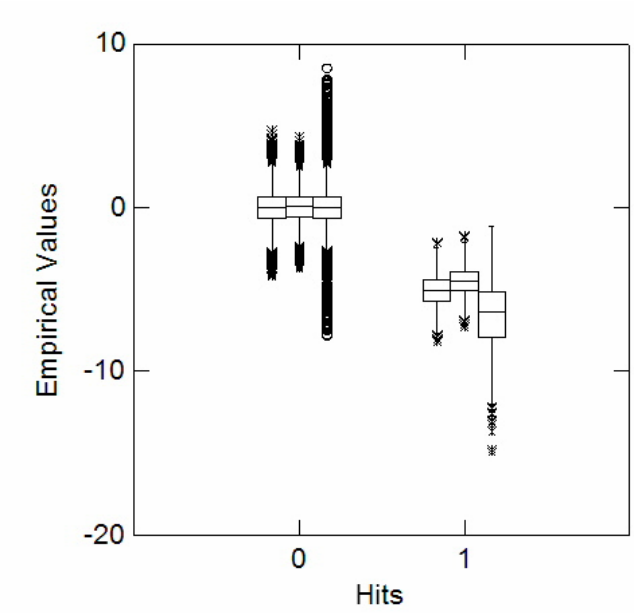


Figure 8sm. Box plots for null (Hits = 0) and "standard normal + 1% hits" > (Hits = 1) data. From left to right, empirical values are for raw (Method 2), well-corrected, and B-score data.

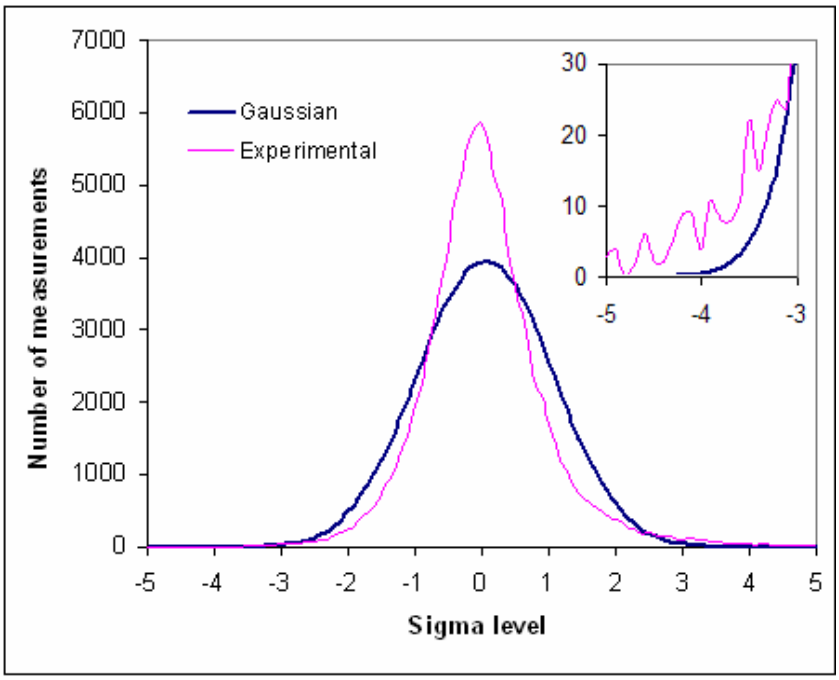


Figure 9sm. Distribution of measurements in the McMaster assay (1250 plates) and its comparison to a Gaussian distribution.

Table 5sm. Results of the χ-square contingency tests carried out for the raw and well-corrected (Well Cor.) McMaster datasets with the hit selection thresholds: $\bar{x} - SD$, $\bar{x} - 1.5SD$, and $\bar{x} - 2SD$.

$\alpha = 0.01$	$\bar{x} - SD$		$\bar{x} - 1.5SD$		$\bar{x} - 2SD$	
	Raw	Well Cor.	Raw	Well Cor.	Raw	Well Cor.
Mean number of hits per well	137.7	134.3	49.8	46.8	18.4	16.9
χ-square value	2377.4	204.6	1258.4	173.6	438.6	74.8
χ-square critical value	111.1	111.1	111.1	111.1	111.1	111.1
χ-square contingency hypothesis H_0	No	No	No	No	No	Yes
Figure 6	a	b	c	d	e	f

Table 6sm. Results of the χ-square contingency tests carried out for the raw and well-corrected (Well Cor.) McMaster datasets with the hit selection thresholds: $\bar{x} - 2.5SD$, $\bar{x} - 3SD$, and $\bar{x} - 3.5SD$.

$\alpha = 0.01$	$\bar{x} - 2.5SD$		$\bar{x} - 3SD$		$\bar{x} - 3.5SD$	
	Raw	Well Cor.	Raw	Well Cor.	Raw	Well Cor.
Mean number of hits per well	7.3	7.1	3.2	3.1	1.5	1.5
χ-square value	172.0	86.6	129.0	110.7	106.2	105.3
χ-square critical value	111.1	111.1	111.1	111.1	111.1	111.1
χ-square contingency hypothesis H_0	No	Yes	No	Yes	Yes	Yes
Figure 10sm	a	b	c	d	e	f

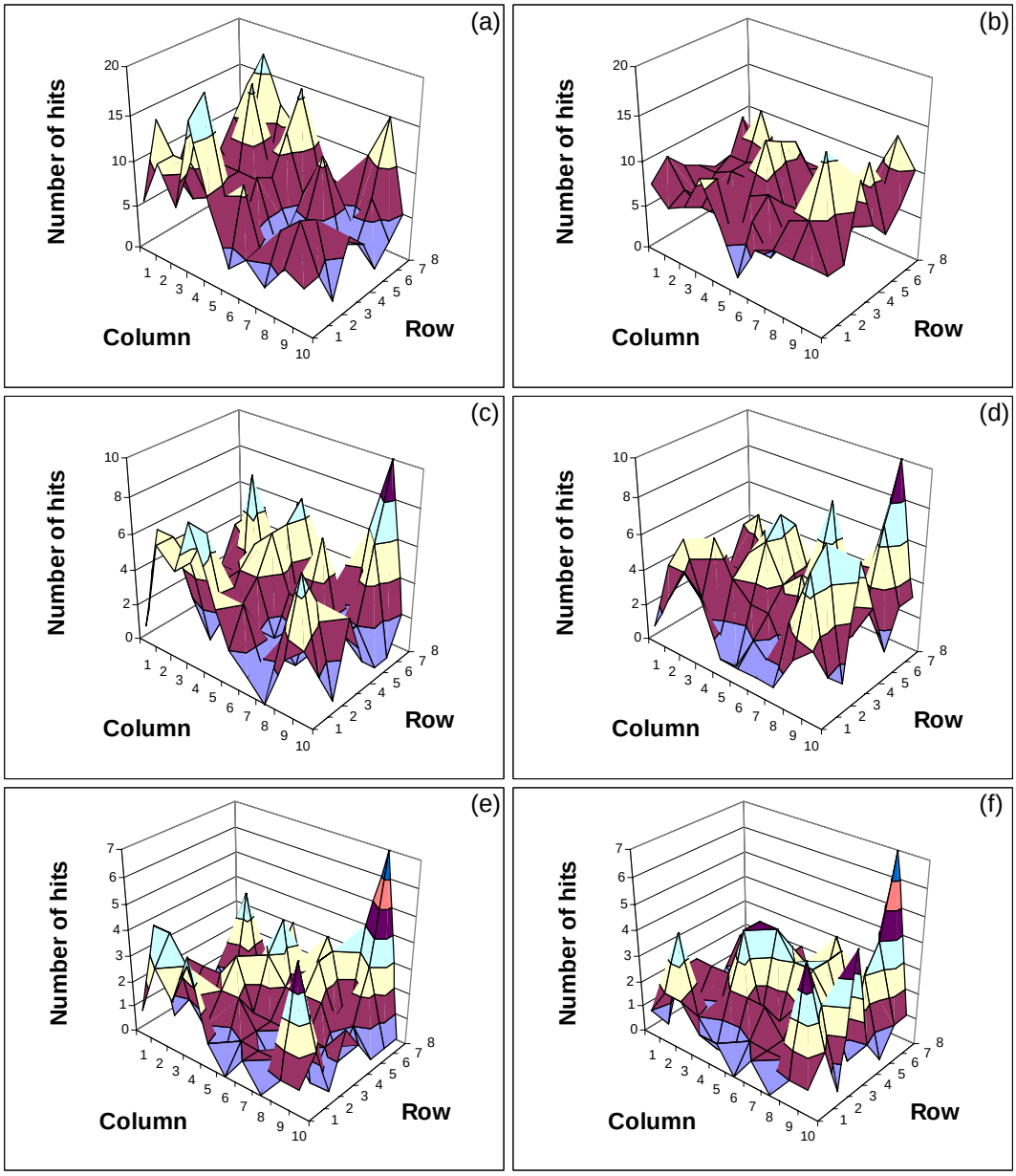


Figure 10sm. Hit distributions for the raw (a, c, and e) and well-corrected (b, d, and f) McMaster datasets obtained with the hit selection thresholds $\bar{x} - 2.5SD$ (a and b), $\bar{x} - 3SD$ (c and d), and $\bar{x} - 3.5SD$ (e and f).

Table 7sm. Hit distribution of the raw McMaster dataset computed for the $\bar{x} - 3SD$ threshold (mean value of hits per well is 3.18 and standard deviation is 2.28).

Row\Column	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	1	6	5	8	5	2	1	0	3	4
2	6	6	5	7	2	4	1	3	7	2
3	5	2	0	2	3	2	1	1	5	0
4	5	4	3	4	4	1	2	1	2	4
5	3	0	1	5	6	3	3	2	2	1
6	2	2	8	1	6	0	6	1	2	0
7	4	5	1	3	7	3	3	1	6	0
8	3	5	1	6	1	3	1	6	10	2

Table 8sm. Hit distribution of the well-corrected McMaster dataset computed for the $\bar{x} - 3SD$ threshold (mean value of hits per well is 3.10 and standard deviation is 2.08).

Row\Column	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	1	3	5	4	1	1	1	1	3	4
2	4	4	3	1	0	2	1	4	7	2
3	5	2	0	4	3	3	1	2	8	1
4	4	5	3	5	6	4	5	1	2	7
5	3	0	2	5	7	3	4	2	2	4
6	2	2	4	2	6	3	8	1	3	1
7	4	5	0	3	2	2	3	2	7	3
8	3	1	0	3	1	3	1	5	10	3

Table 9sm. Consensus hits (i.e. hits identified in both copies) obtained by both Method 1 and Method 5 (with hit selection carried out on the plate-by-plate basis). Method 1 identified as hits all compounds with the consensus residual score <=75%. Method 5 identified as hits all compounds with the consensus residual score lower than $\bar{x} - 2.29SD$ (this threshold corresponds to the 75% residual score used in Method 1).

McMaster samples identified as hits in both copies by Methods 1 and 5	
MAC-0103980	MAC-0115794
MAC-0104038	MAC-0116655
MAC-0104867	MAC-0117820
MAC-0107329	MAC-0119733
MAC-0108994	MAC-0122586
MAC-0109949	MAC-0122661
MAC-0110019	MAC-0122959
MAC-0110027	MAC-0127264
MAC-0110039	MAC-0130938
MAC-0112108	MAC-0131221
MAC-0112179	MAC-0136174
MAC-0112287	MAC-0139408
MAC-0112764	MAC-0140910
MAC-0114159	MAC-0144586
MAC-0114615	MAC-0145030
MAC-0114842	
McMaster samples identified as hits in both copies only by Method 1 (consensus of 75%) and not identified by Method 5	
MAC-0110562	
MAC-0115469	
MAC-0117240	
MAC-0128921	
MAC-0130772	
MAC-0132669	
MAC-0133856	
MAC-0140989	
MAC-0145361	
MAC-0149343	
MAC-0150029	