Experimental Design and Statistical Methods for Improved Hit Detection in High-Throughput Screening

Nathalie Malo, James A. Hanley, Graeme Carlile, Jing Liu, Jerry Pelletier, David Thomas and Robert Nadon

J Biomol Screen 2010 15: 990
DOI: 10.1177/1087057110377497

Experimental Design and Statistical Methods for Improved Hit Detection in High-Throughput Screening

NATHALIE MALO,[1,2] JAMES A. HANLEY,[2] GRAEME CARLILE,[3] JING LIU,[3] JERRY PELLETIER,[3] DAVID THOMAS,[3] and ROBERT NADON[1,4]

Identification of active compounds in high-throughput screening (HTS) contexts can be substantially improved by applying classical experimental design and statistical inference principles to all phases of HTS studies. The authors present both experimental and simulated data to illustrate how true-positive rates can be maximized without increasing false-positive rates by the following analytical process. First, the use of robust data preprocessing methods reduces unwanted variation by removing row, column, and plate biases. Second, replicate measurements allow estimation of the magnitude of the remaining random error and the use of formal statistical models to benchmark putative hits relative to what is expected by chance. Receiver Operating Characteristic (ROC) analyses revealed superior power for data preprocessed by a trimmed-mean polish method combined with the RVM *t*-test, particularly for small- to moderate-sized biological hits. (*Journal of Biomolecular Screening* 2010:990-1000)

Key words: experimental design, statistical tests, hit detection, replication, bias correction

IDENTIFICATION OF ACTIVE COMPOUNDS IN high-throughput screening (HTS) contexts can be substantially improved by applying classical experimental design and statistical inference principles to all phases of HTS studies. Good experimental design at the data acquisition phase serves 2 broad purposes: (1) improves internal validity by reducing the possibility that observed effects have been caused by confounding factors and (2) minimizes unwanted variation in activity measurements stemming from human, biological, and equipment sources. Statistical methods at the data preprocessing (normalization) phase can further reduce unwanted variation, which cannot be controlled procedurally. At the inference phase, the magnitude of the remaining random error, inherent in any biological system, can be estimated by replicate measurements and taken into consideration when deciding which of the putative hits are sufficiently reliable to warrant follow-up. The information from the random error observed in a particular screen can also be used to estimate anticipated false-negative rates for future similar studies.

Although the advantages of statistical procedures for HTS were described more than a decade ago,[1] statistical treatment of HTS data is only now becoming more common as researchers search for ways to reduce false positives and false negatives. Various methods proposed to characterize the quality of screens[2-4] include: estimating of false-positive and false-negative rates as part of assay quality assessments,[5] removing bias within and across plates,[6-8] improving hit/nonhit ranking,[9] conceptualizing assay validation within a statistical framework,[10] and obtaining random error estimates for use in statistical tests to identify hits.[11,12] There remains a need, however, for elucidation of experimental design and data analysis principles tailored to HTS applications.

We extend our previous arguments[12] that popular methods for bias correction and inference are deficient and offer alternatives. We propose an analytical protocol that combines randomization, replication, and efficient statistical methods, allowing examination of model assumptions, calculation of *p*-values to benchmark putative hits, control of false-positive rates at levels specified by the experimenter, and an increase in true-positive rates for HTS applications.

Specifically, we use 4 data sets to compare various approaches for addressing unwanted variation in HTS measurements: experimental data obtained initially without randomization procedures (as is the case with most screening studies), a repeat of the experiment but with randomization procedures, a control experiment, and a dilution series. We counter the fundamental misunderstanding among many researchers that high correlations between

[1]McGill University and Genome Quebec Innovation Centre, Montreal, Quebec, Canada.
[2]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada.
[3]Department of Biochemistry, McGill University, Montreal, Quebec, Canada.
[4]Department of Human Genetics, McGill University, Montreal, Quebec, Canada.

Received Jan 21, 2010, and in revised form May 27, 2010; Accepted for publication May 29, 2010.

Supplementary material for this article is available on the *Journal of Biomolecular Screening* Web site at http://jbx.sagepub.com/supplemental.

Journal of Biomolecular Screening 15(8); 2010
DOI: 10.1177/1087057110377497

replicate plates are desirable; we demonstrate why this typically reflects bias in the measurements rather than good biological reproducibility. We also show that normalization based on a trimmed-mean polish (TP) provides the desirable statistical characteristics of bias correction and measurement independence and performs better than Z-scores or B-scores. Finally, we show that TP scores, when combined with the RVM *t*-test approach, provide variance and *p*-value distributions that agree with theoretical expectations.

METHODS

Experiment A: Nonrandomized immunofluorescent screen

Some 1120 chemical compounds were tested to determine if they correct the trafficking defect of the phenylalanine deletion mutant form of cystic fibrosis transmembrane conductance regulator (CFTR) protein ΔF508 (see supplementary information methods online). Fourteen 96-well plates were run in duplicate. Including incubation time, the screen was run in 4 days. Plates were processed in sets of 5, followed immediately by a duplicate set processed in the same sequence. Compounds that correct the mutant protein trafficking defect are detected by an increase in fluorescence (arbitrary units)—large measured values are more likely to be regarded as biologically valid hits.

Experiment B: Randomized immunofluorescent screen

This screen was the same as experiment A except for 2 aspects: processing order was randomized for all steps in the protocol, and replicates were obtained in 3 independent runs (i.e., blocks).

Experiment C: Measurement experiment

An inactive compound from experiment A was tested in all of the 80 middle wells of six 96-well plates. Plate processing order was randomized for all steps.

Experiment D: Dilution series in vitro translation assay

This experiment uses a different assay and target than the cystic fibrosis screen. A known protein synthesis inhibitor was arrayed within each of 6 replicated plates in 10 concentrations (0.0098, 0.0195, 0.039, 0.078, 0.1563, 0.2344, 0.3125, 0.4687, 0.625, and 1.25 μM). Four replicates of each of the 10 concentrations and 24 negative controls (DMSO) were randomly located in the 64 middle wells of 96-well plates. Positive controls (anisomycin at 50 μM) and negative controls (DMSO) were placed in alternating wells on the 1st, 2nd, 11th, and 12th columns. Firefly and renilla luciferase activity measurements were obtained for each well; low measured values corresponded to hits.

To circumvent the unrealistically high proportion (40/64) of true hits within each plate, we generated random samples from the data to mimic hit proportions that might be expected from a standard primary screen. Removing potential row and column biases with the TP score normalization method was deemed inappropriate for these data because differences among the rows and columns reflected biological differences in addition to any potential biases. Let $i = 1,\ldots, I$ rows; $j = 1,\ldots, J$ columns; and $p = 1,\ldots, P$ plates. Accordingly, the data were normalized as follows:

$$\frac{x_{ijp} - \tilde{x}_p}{MAD_p} \tag{1}$$

where x_{ijp} is the compound measurement corresponding to the well located in row i, column j, and plate p; $\tilde{x}_p$ and MAD_p are, respectively, the median and the median absolute deviation of all measurements within plate p.

For each of 100 simulation runs, we randomly sampled (with replacement) 1120 normalized measurements from the empirical data set (14 plates × 80 values per plate). Some 1064 "nonhits" were sampled from the 144 negative control measurements (6 plates × 24 values per plate). Four consecutive concentrations were chosen. For each concentration, 14 hits were sampled from the 24 concentration-specific measurements (6 plates × 4 replicate values per plate), yielding a rate of true hits of 5% within each simulation run. We repeated this simulation for 3 different sets of concentrations (i.e., the 4 highest, the 4 lowest, and the 4 in the middle). Hits were identified according to various statistical criteria, and false-positive/false-negative rates were calculated (see Inferential Statistics section below).

Preprocessing statistics

We compared 3 non-control-based normalization methods. First, the Z-score method:

$$Z_{ijp} = \frac{x_{ijp} - \bar{x}_p}{s_p} \tag{2}$$

where x_{ijp} is the compound measurement corresponding to the well located in row i, column j, and plate p; $\bar{x}_p$ and s_p are, respectively, the mean and the standard deviation of all measurements within plate p.

Second, for the B-score, the residual (r_{ijp}) of the measurement for row i and column j on the pth plate is obtained by fitting a 2-way median polish[13] and is defined below as

$$r_{ijp} = y_{ijp} - \check{y}_{ijp} = y_{ijp} - (\check{\mu}_p + \check{R}_{ip} + \check{C}_{jp}). \tag{3}$$

The residual is defined as the difference between the observed result (y_{ijp}) and the fitted value ($\hat{y}_{ijp}$), defined as the estimated average of the plate ($\hat{\mu}_p$) + estimated systematic measurement offset for row i on plate p ($\hat{R}_{ip}$) + estimated systematic

measurement column offset for column j on plate p ($\hat{C}_{jp}$). The median polish is an iterative algorithm that alternates row and column operations. Considering the rows first, for each row, the row median is subtracted from every element in that row. For each column, the median of the revised numbers is then subtracted from every element in that column. This continues until all medians are 0 or reach some predefined minimal difference from 0. For each plate p, MAD_p is the median absolute deviation of all residuals within the plate (r_{ijp}). The B-score, without the smoothing function,[6] is calculated as follows:

$$B_{ijp} = \frac{r_{ijp}}{MAD_p}. \quad (4)$$

Third, the TP(10) score method:

$$TP(10)_{ijp} = \frac{r_{ijp}^{(10)}}{MAD_p}, \quad (5)$$

where $r_{ijp}^{(10)}$ are the residuals obtained by a 2-way polish[13] using the S-Plus (TIBCO Spotfire, Somerville, MA) 2-way function with trim = 0.10.

All 3 methods rescale measurements so that they are comparable across plates; in addition, the B-score and the TP score correct for row and column effects and are resistant to outliers.[6]

Inferential statistics

The p-values to decide which compounds should be deemed as hits were defined using statistical tests on K replicates. For each compound measurement, a standard 1-sample t-test with $K - 1$ degrees of freedom was calculated as

$$t = \frac{\bar{x}_k - \text{constant}}{s_k \sqrt{1/K}} \quad (6)$$

where $\bar{x}_k$ and s_k are the arithmetic mean and the standard deviation, respectively, of the K replicated normalized measurements; the constant was taken to be zero. The ratio is then evaluated against a t-distribution with $K - 1$ degrees of freedom for estimation of associated p-values. Because of cost and time issues, the number of replicates is usually very small. As such, this test relies on imprecise estimates of variance and has corresponding low sensitivity (high false-negative rates).

The RVM 1-sample t-test provides a compromise between the low sensitivity of the 1-sample t-test and the strong common error assumption of a 1-sample z-test. Compound-specific variances are assumed to follow an inverse-gamma distribution with parameters a and b as follows[14,15]:

$$\sigma^{-2} \sim G(x; a, b) \equiv \frac{x^{a-1} \exp(-x/b)}{\Gamma(a) b^a}. \quad (7)$$

We estimated the a and b parameters by fitting the sample variances to an F-distribution according to the maximum likelihood method described by Wright and Simon[15]:

$$(ab) s_k^2 \sim F_{(k-1),2a}. \quad (8)$$

The RVM t-test is calculated as

$$\tilde{t} = \frac{\bar{x}_k - \text{constant}}{\tilde{s}\sqrt{1/K}} \quad (9)$$

where $\tilde{s}^2 = \frac{(K-1)s_k^2 + 2a(ab)^{-1}}{(K-1) + 2a}$ and where $\bar{x}_k$ and s_k^2 are the arithmetic mean and the variance, respectively, of the K replicated measurements, and the parameters a and b are estimated from the data from all compounds by fitting an F-distribution to the sample variances s^2.

$\tilde{t}$ follows a t-distribution with $K - 1 + 2a$ degrees of freedom. Note that variance ($\tilde{s}^2$) is estimated by a weighted average of the compound-specific variances s_k^2 and an estimate $(ab)^{-1}$ of the "typical" error variance underlying the error distributions of different compounds, with weights equal to $(K - 1)$ and $2a$, respectively.[12] This leads to an increase of $2a$ degrees of freedom over the standard t-test.

RESULTS

Examination of raw data

Figure 1 shows histograms and line plots of column effects of raw data for 2 immunofluorescent screens (experiments A and B; see Methods). Under the usual assumptions of unbiased measurements and few hits, the majority of the measured values would be symmetrically distributed around a central null value. **Figure 1a** shows, however, that the distribution of the first replicate set in the nonrandomized screen contains 2 modes. Moreover, **Figure 1b** shows evidence of systematic error between replicate sets, columns, and differences in the pattern of column bias across replicate sets (see Suppl. Fig. S1 for plate-by-plate column effect plots for the 2 immunofluorescent screens). The distributions of the randomized screen data are more in line with expectations (**Fig. 1c**), although the 3 sources of systematic error remained (**Fig. 1d**).

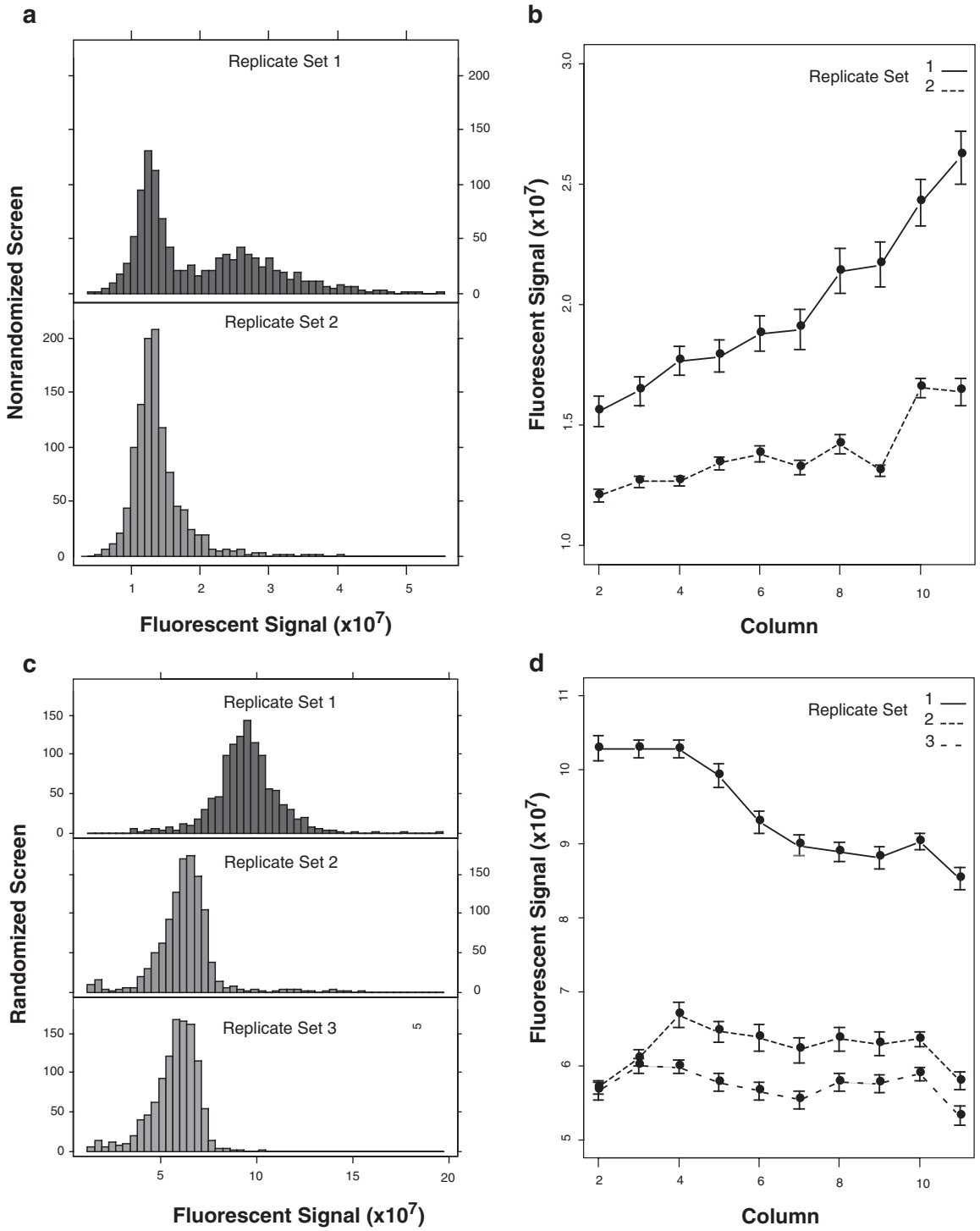

FIG. 1. Graphical display of raw data for each replicate set of immunofluorescent screens. (**a**) Histograms of raw data for the nonrandomized (experiment A) screen show large variability, especially in the first replicate set. The first distribution contains 2 modes and a very long tail on the right (i.e., more large values than the usual expected proportion of hits). The distribution of the second replicate set is closer to expectation with 1 mode and smaller asymmetry on the right end. (**b**) Plot of average measurements against column number shows that column effects are present, especially for the right-most columns. (**c**) Histograms of raw data for the randomized screen (experiment B) again show different patterns. The data are improved, however, in that all 3 distributions are unimodal. (**d**) Column effects were not removed by randomization. Error bars for column effects represent standard errors.

Data preprocessing

Unwanted variation in the measurements that cannot be controlled procedurally may nonetheless be minimized by appropriate normalization of the data (see Methods section for more details on the procedures). **Figure 2** shows that 10% trimmed-mean polish (TP(10)) scores corrected the distributional asymmetry (**Fig. 2a,c**) and column effects (**Fig. 2b,d**) observed in **Figure 1**, in contrast to the popular Z-score normalization method (see Suppl. Fig. S2). Similar results were obtained for the less pronounced row effects (data not shown).

The advantages of TP(10) scores are illustrated further in **Figure 3** by analysis of additional data in which the same compound was tested in every well in the same concentration across all plates (experiment C: measurement experiment; see Methods). As such, in the absence of systematic bias, the same signal plus random noise was expected for all wells of every plate. Consequently, measured values should be uncorrelated with their counterparts in the same locations on other plates and should show no autocorrelations within the series of measurements. **Figure 3a**, however, shows that the raw data were positively correlated across plates, indicating the presence of procedurally induced location-specific biases. **Figure 3b** shows that the TP(10) scores greatly minimized the bias, producing the expected null correlation (scatter plots between plates for Z-scores generate results identical to the raw data because they are simply rescaled raw scores). Similarly, the autocorrelation plots across all 6 replicate plates in **Figure 3c** show substantial correlations among putatively independent measurements for the raw data. The correlation at lag 1 indicates that wells in immediate proximity to each other (down each successive column) are highly correlated ($r = 0.55$). Successive lags indicate correlations between each well and the nth succeeding well (lag n). A pattern was observed that repeated at every eighth lag. Within columns, the correlations decreased until the middle row and then began increasing, reaching a maximum at the well in the first row of the adjacent column (e.g., the highest correlation of $r = 0.67$ was observed for lag 8, which corresponds to immediately adjacent wells across columns). Although Z-scores provided some degree of correction (**Fig. 3d**), TP(10) scores again provided the best correction (**Fig. 3e**), reducing the autocorrelations at the various lags to near-zero values.

Because Z-scores retain any row and column biases, they can generate nonnormally distributed data of various types even if the true underlying biological signals are normally distributed. Although B-scores correct these biases, they can transform normally distributed data into long-tailed distributions, producing excess numbers of false positives.[16] In agreement with these previous findings, both the Z-score and the B-score data produced nonnormally distributed data for the experiment C data (**Fig. 4a,b**). TP(10) scores, by contrast, generated approximately normally distributed data with only a few outliers (**Fig. 4c**).

Hit detection

One major advantage of replicates is that formal statistical models can be used to benchmark putative hits relative to what is expected by chance under the statistical model. **Figure 5** illustrates our investigation of the assumptions of our model as applied to the nonrandomized experiment A and to the randomized experiment B data (see Methods for a detailed description of the tests).

As with our previous findings,[12] the distributions of sample variances did not conform to the constant variance assumption (chi-square distribution) implicitly assumed by the Z-score statistic applied to unreplicated data across compounds or across replicates for each compound for both the nonrandomized and randomized screens (**Fig. 5a,e**), suggesting that the larger number of observed hits with the latter test likely reflects an unduly high false-positive rate (data not shown). The variances matched the theoretically expected inverse-gamma distribution under the RVM model for both the nonrandomized (**Fig. 5b**) and the randomized screens (**Fig. 5f**). The relative lack of smoothness for the nonrandomized screen fit and the small p-value reflect the fewer number of replicates used for calculating the sample variances (2 vs. 3). Both the standard and the RVM 1-sample t-tests failed to generate the uniform 2-tailed p-value distributions within the higher p-value ranges for the nonrandomized screens, which would justify the use of the tests (**Fig. 5c,d**). By contrast, the randomized screen data produced the expected p-value distributions for both statistical tests, although with greater power for the RVM test (**Fig. 5g,h**). In this context, the standard t-test suffers from a lack of degrees of freedom due to the small number of replicates. Rank ordering of the 2 t-statistics is the same, but the quantiles are different because the activity measurements are divided by different estimates of the standard error.

Other considerations

In the 2 immunofluorescent screens (experiments A and B), statistical hits were expected in both directions, and accordingly, we examined 2-tailed p-value distributions as a check of assumptions. Decrease of fluorescent signal may arise from a number of different causes. A compound may be toxic, remain in the cell during the experiment, and have the ability to quench the fluorescence of the tag on the secondary antibody or bind to the cystic fibrosis transmembrane regulator close to the location of the 3HA tag and mask the antibody binding site from the antibody detection. For the biological purposes of the studies, however, the interest lies in the activity measurements that correspond to high positive TP(10) score values (increase in fluorescence). Accordingly, it is appropriate to estimate 1-tailed p-values for hit detection, with the understanding that effects in the opposite (negative) direction will be ignored, no matter how large the effects might be.

Outliers among replicates threaten the validity of results obtained from statistical tests based on means (such as the ones

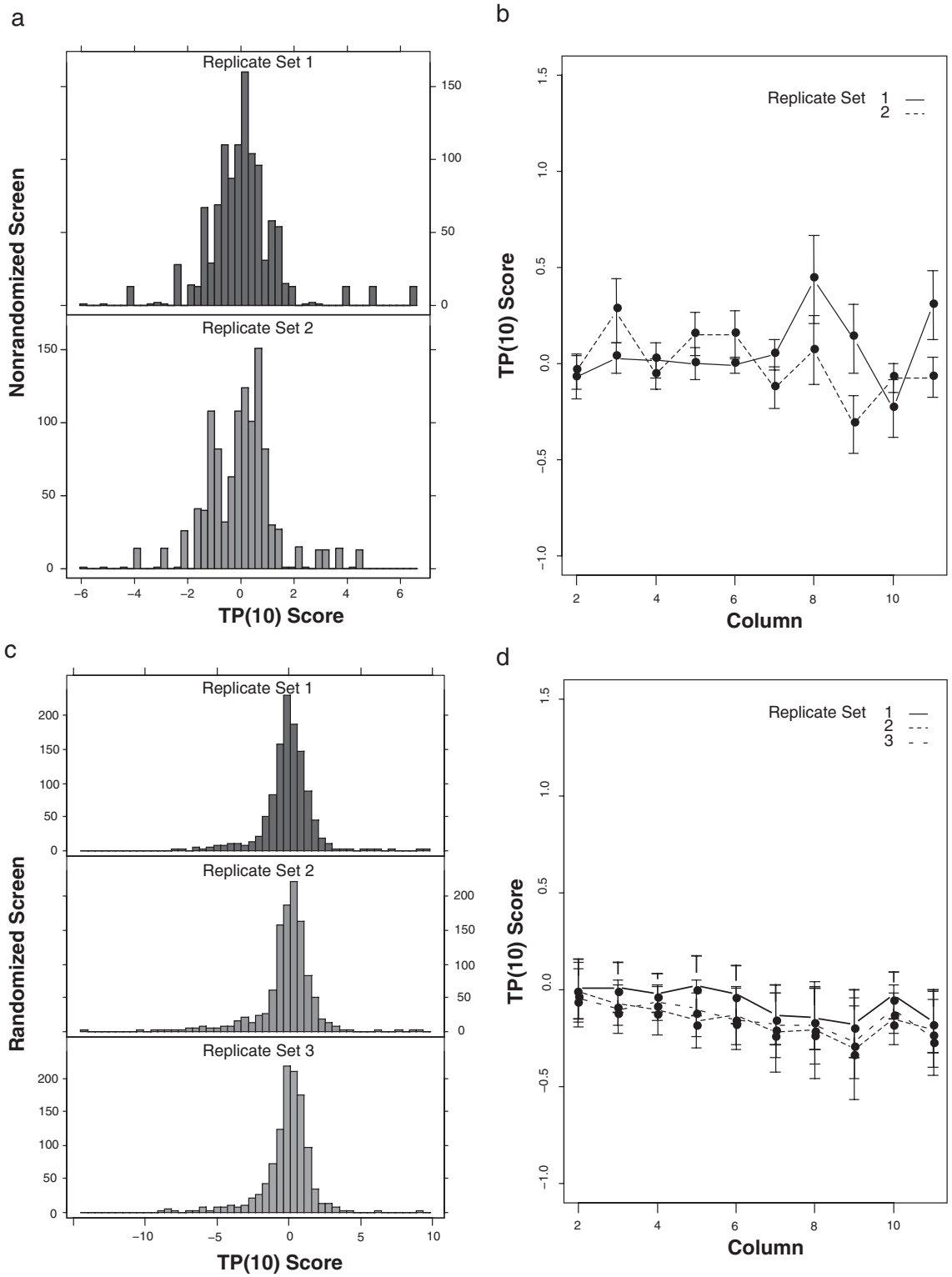

FIG. 2. Graphical display of preprocessed data using the TP(10) score method. (**a**) Histograms of TP(10) scores for the nonrandomized screen (experiment A) show that the distributional asymmetries have been corrected. (**b**) Plot of average measurements against column number shows that the TP(10) scores corrected plate and lessened column effects for the nonrandomized screen. (**c**) Histograms of TP(10) scores for the randomized screen (experiment B) show distributions that are more similar across replicate sets than for the raw data. (**d**) The TP(10) scores corrected both plate and column effects for the randomized screen. Error bars for column effects represent standard errors.

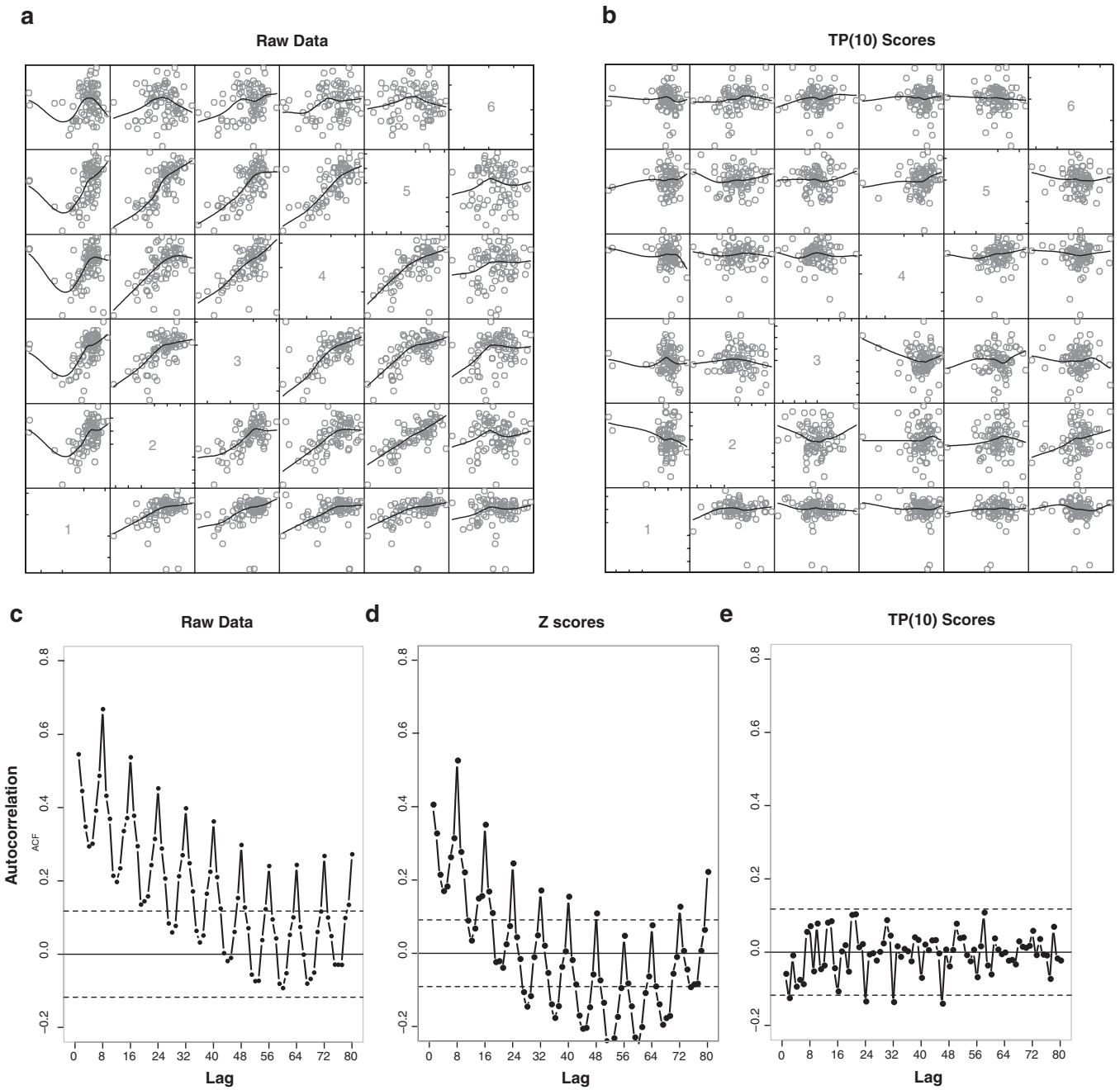

FIG. 3. Pairwise scatter plots between plates for both raw and preprocessed data, for pairwise comparison of plates, from the measurement experiment (experiment C) in which the same compound was tested in all wells of 6 plates. (**a**) Because of procedurally induced bias, raw measurements across plates were correlated. (**b**) The TP(10) score method reduced across-plate correlations to near zero. High autocorrelations across wells within plates were observed for (**c**) raw data and (**d**) Z-scores. (**e**) TP(10) scores reduced the autocorrelations to near-zero values.

employed here). Outliers are difficult to detect, however, when there are few replicates. One method to circumvent this problem in the current context is to investigate whether any of the replicate variances (rather than the individual fluorescent values) may be considered outliers. The advantage is that outlier variances are more readily detected because there are many variances distributed according to a known distribution under the RVM model used here. The idea is that compounds with individual replicate fluorescent outliers should have unusually large variances. The *F*-distribution (**Fig. 5b,f**) can be used as the reference probability

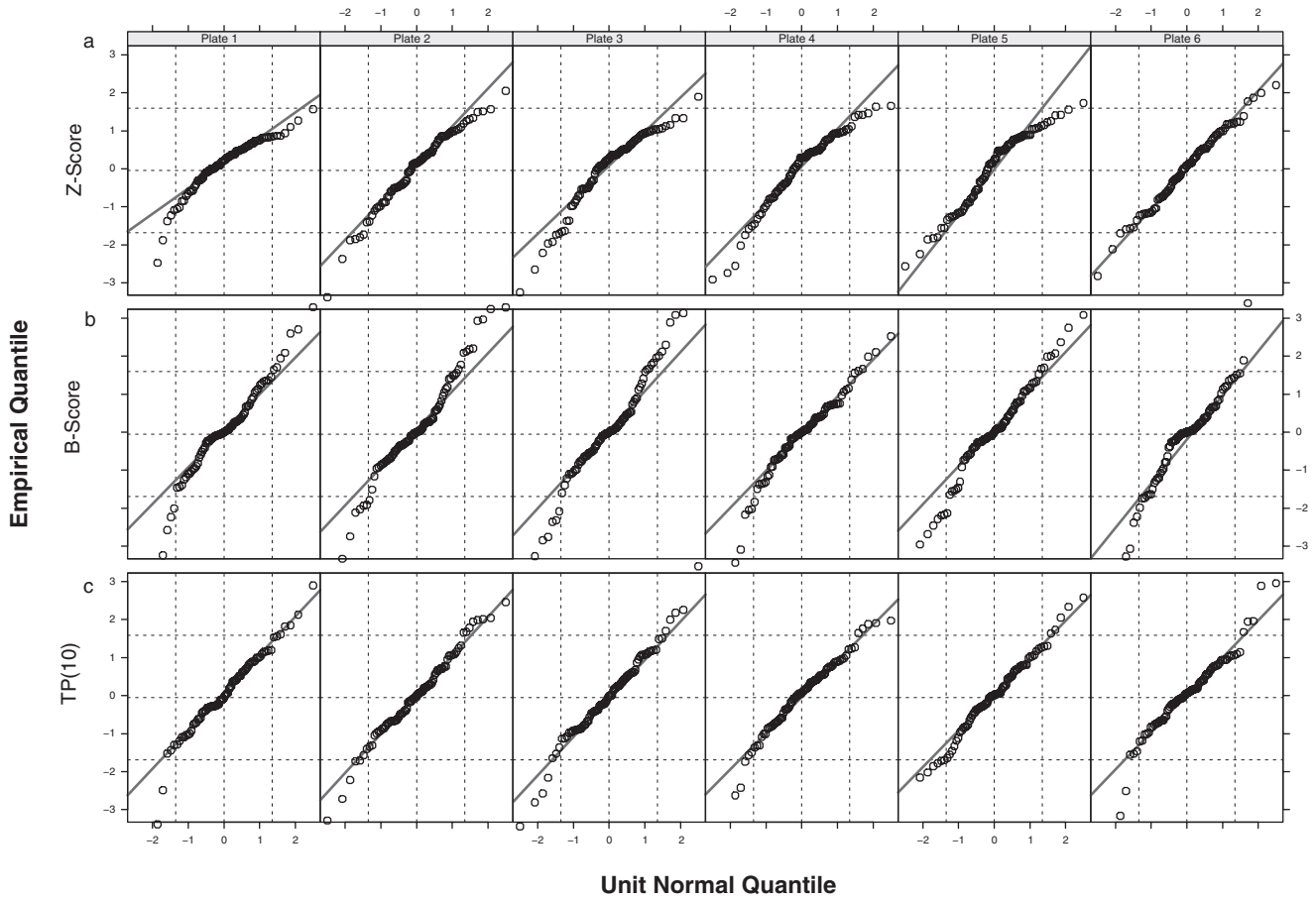

FIG. 4. Normal QQ plots for the measurement experiment (experiment C) data show pronounced deviations from normality for (**a**) Z-scores and (**b**) B-scores. (**c**) TP(10) scores, by contrast, show near-normal distributions with few outliers.

model. One way to evaluate if there are any outlier variances is to estimate the number of "rescaled" variances (i.e., the observed variances multiplied by a and b, the estimated parameters of the inverse-gamma distribution) that exceed a predefined false discovery rate (FDR) threshold with an F-distribution with $K - 1$ and $2a$ degrees of freedom as reference (where K is the number of replicates). For the randomized screen (experiment B) data, we found no evidence of outlier variances (and hence no obvious fluorescent outliers) using the positive false-discovery rate procedure (pFDR)[17] with a q threshold of 0.10.

Finally, interpretation of individual p-values needs to be understood within the multiple testing context. For example, 5% of the compounds are expected to have p-values ≤ 0.05 merely by chance. For the randomized screen, 9% of the individual p-values were ≤ 0.05, suggesting that hits are present (**Fig. 5h**). Indeed, 58 (5%) of the compounds were identified as hits using a pFDR q = 0.12 threshold. Larger q values could also have been used to reduce the false-negative rate (with the consequential increase in false positives).

Empirical demonstration of statistical power

We performed a dilution series experiment (experiment D: dilution series in vitro translation assay; see Methods) in which various concentrations of an active compound were randomly assigned well positions on a 96-well plate. **Figure 6** presents Receiver Operating Characteristic (ROC) curves that compare the performance of 3 statistical tests based on random samples generated from the data. The RVM t-test performed best, generating the fewest false negatives at fixed false-positive levels. **Figure 6** also shows that false negatives were reduced by increasing the number of replicates, especially for low-concentration hits.

DISCUSSION

We provide experimental design principles and statistical methods to improve hit detection in HTS. Our results illustrate that sensitivity and specificity of screens can be increased by (1) robust preprocessing of raw data that removes row and

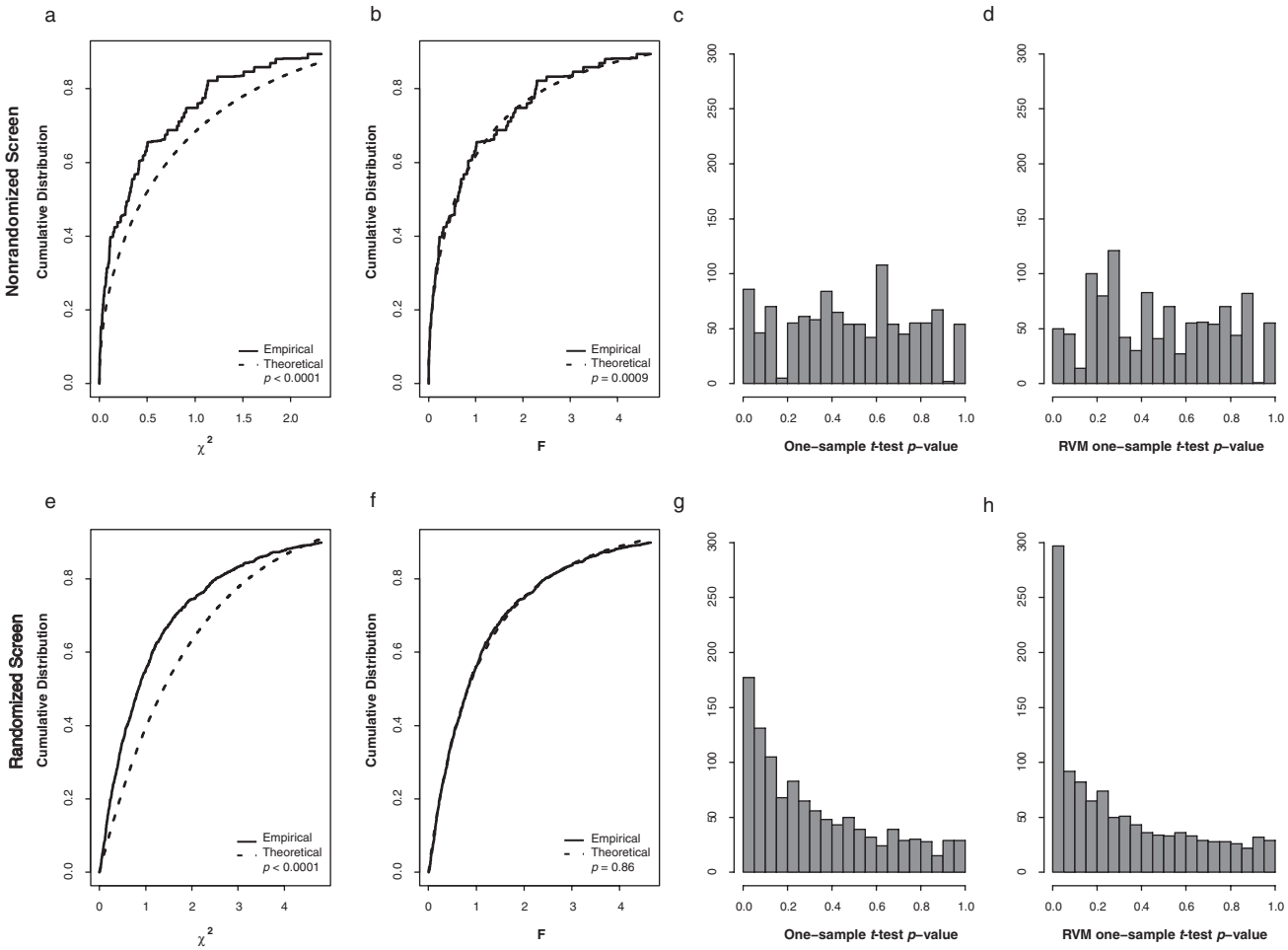

FIG. 5. Checking of assumptions for statistical testing. (**a, e**) Distributions of sample variances across replicates were inconsistent with the typical assumption of constant variance for both the nonrandomized (experiment A) and randomized (experiment B) screens. The variances matched the theoretically expected inverse-gamma distribution under the RVM model for (**b**) the nonrandomized (experiment A) and (**f**) the randomized (experiment B) screens. (**c, d**) The *p*-values in the higher range did not follow the theoretically expected uniform distribution for the nonrandomized screen data with either the standard or the RVM 1-sample *t*-tests. (**g, h**) By contrast, the randomized screen data generated the expected *p*-value distributions for both the standard and the RVM 1-sample *t*-tests, although with greater power for the RVM test.

column effects and (2) statistical inferential methods based on replicates and that borrow strength from other compounds to improve variance estimates. Gains in accuracy and reproducibility that can be achieved by such an approach are especially noteworthy for decreasing the false-negative rate among low- to moderate-sized biological hits.

For unavoidable sources of variation, we recommend randomization and blocking of processing steps to provide the means to make valid assessments of compounds' activity levels by minimizing the effects of potential confounds such as processing order. To reduce differences across blocks and thus increase precision, we adopted the rule to "block what you can and randomize what you cannot"[18(p93)] (see also Box[19]). In the immunofluorescence randomized screen, for example, we defined a block as a replicated run, and we randomized plate processing order within each run.

Various exploratory graphics[20] of raw and preprocessed data allow assessment of measurement adequacy before performing further statistical analysis. We suggest examining data distributions to check against gross measurement errors. Plots of plate and row/column means can highlight a frequent source of bias that can be minimized by robust preprocessing methods such as the TP score. Autocorrelation plots can provide checks for measurement independence. Scatter plots of replicate plates can reveal biases in the measurements. It is essential in all of these methods to compare the observed data to measurement expectations.

In Malo et al.,[12] we argued in favor of non-control-based normalization methods and specifically recommended the B-score

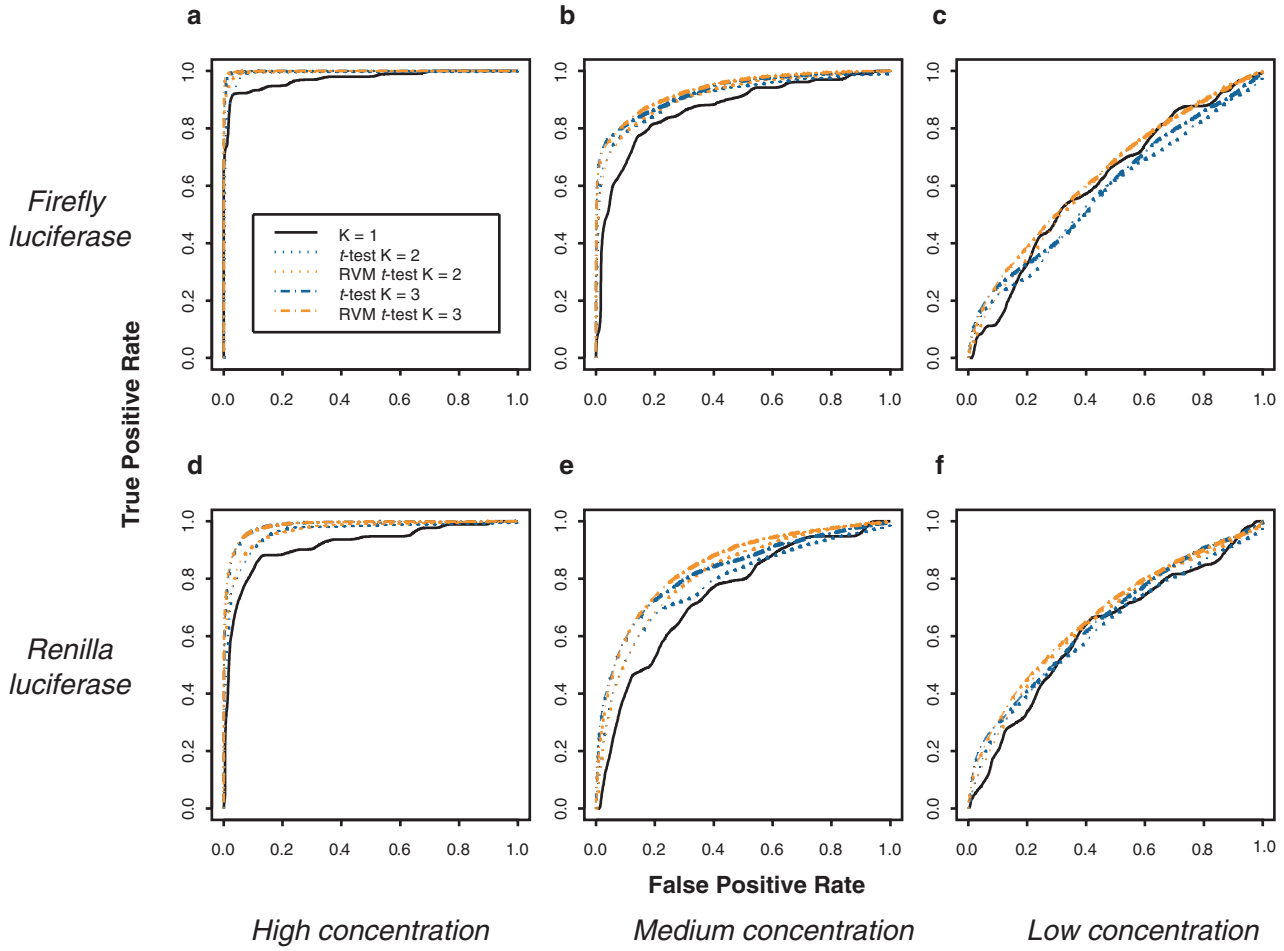

FIG. 6. Receiver operating characteristic (ROC) curves to compare power achievable with various inferential approaches and various numbers of replicates. Data were generated according to a dilution series experiment (experiment D). The black line represents the rank ordering of activity measurements in the absence of replicates. The color curves illustrate the benefit of using statistical tests based on replicates. (**a-c**) For the firefly luciferase assay, hits were easily identified by all tests at high and medium concentrations. (**d-f**) For the Renilla luciferase assay, hits were more readily identified with the RVM 1-sample *t*-test than with the standard 1-sample *t*-test at the medium concentration. All methods failed to identify hits at very low concentrations.

procedure.[6] We have noted more recently, however, that the B-score can potentially generate excessive false positives because normally distributed null data generate long-tailed B-score distributions.[16] We propose the trimmed-mean polish as an alternative to the B-score, which has good robustness and superior distributional properties with normally distributed data. In this study, we used a trim value of 10%, although higher trim values should be used if more than 10% true hits are expected within some columns or rows (to a maximum of 50%, the value used by the B-score method).

With replicates, the significance threshold for hit identification can be based on *p*-values offering the advantage of understanding the probability of what is expected by chance alone. The RVM approach we used here is particularly well suited for HTS applications with few replicates. As in the early days of microarrays, HTS researchers have begun to use duplicate measurements as recommended by screening centers such as the Harvard Medical School Screening Center (http://iccb.med.harvard.edu/screening/guidelines.htm).

Triplicate measurements, however, offer several advantages over duplicates. With triplicates, undesirable outlier measurements can be detected and corrected before the statistical analyses are performed to improve sensitivity and specificity. Triplicate measurements and the RVM approach may be less necessary when the primary goal of a study is to determine large effects ("low-hanging fruit"). The substantial increases in statistical power afforded by these methods gain in importance, however, when there is interest in detecting small to moderate effects (e.g., when screening small molecular fragments with the objective of identifying small fragments with modest biological activity). For the *t*-statistic, one

additional replicate provides large gains when sample sizes are small. For example, the critical *t*-value 5% false-positive threshold for identifying a hit with a 1-sample *t*-test with 2 replicates is 12.7, whereas the threshold for 3 replicates is 4.3. Lesser gains are observed for 4 and 5 replicates (thresholds of 2.57 and 2.28, respectively). Additional degrees of freedom can be achieved with the RVM *t*-test, which acts as a proxy for adding replicates.[12,14,15,21]

High-throughput biotechnologies present challenges and opportunities for statistical inference relative to the more traditional low-throughput contexts. The typically small number of replicates available contravenes the standard statistical practice of obtaining larger sample sizes. Conversely, the very large number of features (compounds, genes, etc.) present in high-throughput studies can be leveraged, for example, through empirical Bayes approaches such as the RVM model. Our results suggest that in addition to these analytical considerations, researchers need to consider randomization as an integral part of the screening process. Ultimately, consensus on the merits of specific pre-preprocessing, statistical inference, and study design methods will be decided by their respective performances in applied contexts.

ACKNOWLEDGMENTS

We thank Janie Lapointe for generating the randomized and the nonrandomized screen data. This work was supported by the "Informatics and Chemical Genomics" funding to RN under the Genome Quebec Phase II Bioinformatics Consortium program and Le Fonds Québécois de la Recherche sur la Nature et les Technologies (FQRNT) grant 119258 (Statistical Methods for High-Throughput Screening).

REFERENCES

1. Lutz MW, Menius JA, Laskody RG, Domanico PL, Goetz AS, Saussy DL, Rimele T: Statistical considerations in high throughput screening [Online]. Retrieved from http://www.netsci.org/Science/Screening/feature05.html
2. Gunter B, Brideau C, Pikounis B, Liaw A: Statistical and graphical methods for quality control determination of high-throughput screening data. *J Biomol Screen* 2003;8:624-633.
3. Zhang JH, Chung TD, Oldenburg KR: A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J Biomol Screen* 1999;4:67-73.
4. Zhang XD: A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays. *Genomics* 2007;89:552-561.
5. Sui YX, Wu ZJ: Alternative statistical parameter for high-throughput screening assay quality assessment. *J Biomol Screen* 2007;12:229-234.
6. Brideau C, Gunter B, Pikounis B, Liaw A: Improved statistical methods for hit selection in high-throughput screening. *J Biomol Screen* 2003;8:634-647.
7. Kevorkov D, Makarenkov V: Statistical analysis of systematic errors in high-throughput screening. *J Biomol Screen* 2005;10:557-567.
8. Wu ZJ, Liu DM, Sui YX: Quantitative assessment of hit detection and confirmation in single and duplicate high-throughput screenings. *J Biomol Screen* 2008;13:159-167.
9. Zhang XD, Ferrer M, Espeseth AS, Marine SD, Stec EM, Crackower MA, et al: The use of strictly standardized mean difference for hit selection in primary RNA interference high-throughput screening experiments. *J Biomol Screen* 2007;12:497-509.
10. Coma I, Clark L, Diez E, Harper G, Herranz J, Hofmann G, et al: Process validation and screen reproducibility in high-throughput screening. *J Biomol Screen* 2009;14:66-76.
11. Buxser S, Vroegop S: Calculating the probability of detection for inhibitors in enzymatic or binding reactions in high-throughput screening. *Anal Biochem* 2005;340:1-13.
12. Malo N, Hanley JA, Cerquozzi S, Pelletier J, Nadon R: Statistical practice in high-throughput screening data analysis. *Nat Biotechnol* 2006;24:167-175.
13. Tukey JW: *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977.
14. Rocke DM: Design and analysis of experiments with high throughput biological assay data. *Semin Cell Dev Biol* 2004;15:703-713.
15. Wright GW, Simon RM: A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* 2003;19:2448-2455.
16. Makarenkov V, Zentilli P, Kevorkov D, Gagarin A, Malo N, Nadon R: An efficient method for the detection and elimination of systematic error in high-throughput screening. *Bioinformatics* 2007;23:1648-1657.
17. Storey JD: A direct approach to false discovery rates. *J Roy Stat Soc Ser B (Stat Method)* 2002;64:479-498.
18. Box GEP, Hunter JS, Hunter WG: *Statistics for Experimenters: Design, Innovation, and Discovery*. Hoboken, NJ: Wiley-Interscience, 2005.
19. Box GEP: *Improving Almost Anything: Ideas and Essays*. Hoboken, NJ: Wiley-Interscience, 2006.
20. Cleveland WS: *Visualizing Data*. Murray Hill, NJ: Hobart Press, 1993.
21. Smyth G: Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004;3:Article 3.

Address correspondence to:
Robert Nadon
Department of Human Genetics, McGill University
1205 avenue du Docteur Penfield N5/13
Montreal, Quebec, Canada, H3A 1B1

E-mail: robert.nadon@mcgill.ca