Statistical Contributions to Data Analysis for High-Throughput Screening of Chemical Compounds

Nathalie Malo

Doctor of Philosophy

Department of Epidemiology, Biostatistics, and Occupational Health

McGill University

Montreal,Quebec

June 2006

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Your file Votre référence
ISBN: 978-0-494-27817-8
Our file Notre référence
ISBN: 978-0-494-27817-8

PREFACE

Contributions of Authors

This thesis is the beginning of an important collaboration between statisticians and life scientists working in high-throughput screening (HTS) of chemical compounds. When Dr. Nadon agreed to become my supervisor, he introduced me to Dr. Jerry Pelletier and Dr. David Thomas, professors in the Department of Biochemistry. They, at their turn, introduced me to the HTS process and their respective laboratories. Two years ago, my supervisors, Dr. Hanley, Dr. Nadon, and I were just new to this field.

When I started the literature review, I realized the urgent need for statisticians to get involved in this field. Although a large amount of data is generated daily by the new automated technology, only a few basic statistical methods are currently used. Dr. Nadon and I determined statistical questions that could be answered in this research project and designed a first small measurement experiment. I spent days in the HTS laboratory, in order to observe the entire process, ask questions to the technicians, and thus, get a better understanding of the origins of the data. I did the statistical analysis of the data, and I wrote the first review paper. During this work, looking to develop new methods, I came across another research question of statistical interest, which became the topic of the second paper. I was also responsible for designing the simulation study, and for programming the diverse methods.

Dr. Nadon had responsibility for day to day supervision. He offered his expertise in microarrays and suggested some references to me. He gave me advice with respect

to the research questions and the data analysis. Dr. Hanley offered his statistical expertise by giving me recommendations with regard to the statistical methods and the simulation study. He helped me to determine the objectives of the thesis and with the policies related to the department. I also wrote the two other papers and the thesis. Both Dr. Hanley and Dr. Nadon offered support during the editing of each of the three manuscripts and the thesis.

Statement of Originality

The doctoral thesis consists of three manuscripts. In the first manuscript, I critically examine the current practice in HTS data analysis, and provide statistical recommendations. To my knowledge, it is the first statistical review in that field. The second manuscript is aimed at a statistical audience. It evaluates the performance of various robust methods for handling replicates in two-way layouts. I compared software and asked authorities in that field, but there was no consensus on which methods should be used. Finally, the third manuscript is an application of the statistical methods to HTS data from both real screens and in-house experiments. Similar methods have been used in microarrays, but I adapted them in order to minimize potential biases specific to HTS data.

Notes to the reader

Since this work is the result of multi-disciplinary collaborations, I have included sufficient statistical material so that most of the thesis can be understood by statisticians, epidemiologists and life scientists. In addition, I described the HTS process,

covered some background material in the introduction, and defined technical terms in the glossary. The thesis does not included a separate literature review, since it is the essence of the first manuscript. Finally, I used 'we' in the writing of the three manuscript chapters, and 'I' throughout the other chapters.

ACKNOWLEDGEMENTS

Six years ago, when I finished my master's degree, I remember, I did not think I would do a PhD. But today I am here, writing these last sentences of my thesis. What did happen meanwhile? I just met amazing persons who transmitted me their knowledge and their passion for science. I think about John Raelson and Majid Belouchi who introduced me to statistical genetics; and Nicholas Schork who guided me in this new field and convinced me to pursue my study at the doctoral level.

The last years have certainly been trying as well as enriched. I first thank my two supervisors, Dr. James Hanley and Dr. Robert Nadon, without whom this thesis would not have been possible. I thank you for your financial and especially your day to day support. I appreciate your confidence which allowed me to follow my own schedule, to work from home, and thus, to be successful. I want to thank Dr. Jerry Pelletier and Dr. David Thomas for introducing me to their laboratories, and also, Jing Liu, Jany Lapointe, and Graeme Carlile for performing specific designed experiments.

In my thoughts, I cannot forget all my new colleagues and friends from both McGill University and the Genome Center Innovation Center, above all, Marie-Ève Beauchamp, my first 'amie de fille'. I thank you for your organization, your listening and your judicious advice on science and on life. I also wish to thank Sonia Cerquozzi for initiating my understanding of biochemistry, Mathieu Miron for his invaluable help with biology and Adobe Illustrator, and Carl Murie for his programming support.

During these years, in addition to my PhD, I also had two great opportunities: traveling and playing music. This would certainly not have been possible without the motivation from very important persons to me. Sébastien Maître made me rediscover my piano, the scene, as well as jazz music; the grade from the Faculty of Music that appears on my transcript is testament to this! Chu accompanied me in Morroco, and there I first met Sergi who has been here for me, whatever the size of the ocean or the number of countries that there is between us. I have learned so much from you about simplicity, honesty, life, love and traveling.

Les derniers mais non les moindres, ma famille. Une fois de plus, je remercie mes parents de toujours être avec moi peu importe les décisions que je prends au cours de ma vie et qui parfois vous semblent complètement folles! Aussi surprise et heureuse que je puisse l'être, je remercie particulièrement ma petite soeur et nouvelle confidente, Mélissa, qui a mis au monde mes deux adorables filleuls : Océane et Nolan, pour qui je serai toujours là. Je remercie aussi Nancy, une cousine auparavant méconnue, qui a partagé mon quotidien durant ces dernières années. Alain et Sylvain, les jumeaux, les Mailloux, mes cousins, mes frères, mes colocs, peu importe la nomenclature, je n'oublierai jamais tous les instants passés avec vous. Bref, merci à tous pour votre sourire et vos encouragements.

ABSTRACT

High-throughput Screening (HTS) is a relatively new process which allows several thousand chemical compounds to be tested rapidly in order to identify their potential as drug candidates. Despite increasing numbers of promising candidates, however, the numbers of new compounds that ultimately reach the market have declined. One way to improve upon this situation is to develop efficient and accurate data processing and statistical testing methods tailored for HTS. Human, biological or mechanical errors may develop across the several days it takes to run the entire screen and cause unwanted variation or "noise". Consequently, HTS data need to be preprocessed in order to reduce the effect of systematic errors. Robust statistical methods for outlier detection can then be applied to identify the most promising compounds. Current practice typically uses only single measurements, which negates the use of standard statistical methods and forces scientists to rely on strong untested assumptions and on arbitrary choices of significance thresholds.

The broad objectives of this research are to develop and evaluate robust and reliable statistical methods for both data preprocessing and statistical inference. This thesis is divided into three papers. The first manuscript is a critical review of the current practices in HTS data analysis. It includes several recommendations for improving sensitivity and specificity of screens. The second manuscript compares the performance of different robust preprocessing methods applied to replicated two-way data with respect to detection of outlying cells. The third manuscript evaluates

some of the statistical methods described in the first manuscript with respect to their performance when applied to several empirical data sets.

ABRÉGÉ

Le criblage à haut débit est un nouveau processus permettant de tester plusieurs milliers de composés chimiques rapidement dans le but d'identifier des candidats potentiels pour le développement de nouveaux médicaments. Malgré le nombre croissant de candidats prometteurs, le nombre de nouveaux composés qui atteignent le marché à toutefois diminué. Une façon d'amélioré cette situation est de développer des méthodes efficaces et précises pour le traitement de ces données ainsi que pour l'inférence statistique. Des erreurs humaines, biologiques et mécaniques peuvent survenir durant les semaines requises pour procéder à un dépistage complet et ainsi causer du "bruit", soit de la variation non désirée. D'où l'importance de traiter les données afin de réduire l'effet d'erreurs systématiques. Des méthodes robustes pour la détection de valeurs aberrantes peuvent ensuite être utilisées pour identifier les composés les plus prometteurs. En pratique, une seule mesure est obtenue pour chaque composé et cette absence de mesures répliquées empêche l'utilisation de méthodes statistiques habituelles et oblige les scientifiques à baser leur analyses sur de fortes hypothèses non vérifiées et sur des choix arbitraires de seuils de signification.

Les objectifs principaux de cette recherche consistent en le développement et l'évaluation de méthodes statistiques robustes et fiables pour le traitement des données et l'inférence statistique. Cette thèse est divisée en trois articles. Le premier manuscrit est une revue critique des pratiques courantes pour l'analyse de données

provenant de criblage à haut débit de composés chimiques. Plusieurs recommandations quant à l'amélioration de la sensitivité et de la spécificité des dépistages sont également incluses. Dans le deuxième manuscrit, je compare la performance de diverses méthodes robustes pour le traitement de tableaux de données répliquées lors de la détection de cellules aberrantes. Dans le troisième manuscrit, j'évalue différentes méthodes statistiques, décrites dans le premier article, lorsque appliquées à plusieurs jeux de données empiriques.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

1.1 Rationale

High-throughput screening (HTS) is a large-scale process that is the first critical step in drug discovery. A **collection** of chemical **compounds** is tested against a specified therapeutic **target** in order to identify potential drug candidates rapidly and accurately. The scientific challenge is to test a very large number of compounds against a number of targets while minimizing the research costs. This process was made possible in part by the recent integration of new automated technology that works with very small volumes.

In a single experimental run, over a period of weeks, thousands of compounds are tested in hundreds of plates, each containing a two-way array of wells. Typically, 80 different compounds are stored on a single 96-well plate that contains 8 rows and 12 columns. The first and the last columns are left empty for future use of **controls**. Raw data have no units, since activity values, generally obtained by luminescence or fluorescence, are measured relative to each other and depend on the technology used, the **assay** format, etc.

The purpose in analyzing the large amount of data points generated daily is to find the small unknown proportion (maybe 1%) of "outliers", i.e. chemical compounds with an extreme activity level (labeled "**hits**") that may later be developed

into drugs. The focus on outliers is for the opposite purpose than that is in traditional analyses up to now. Traditionally, outliers are undesirable, since they arise from errors in measurements and thus, are usually removed before performing any statistical analysis. In HTS, the outliers (hits) are of interest by themselves and statistical analyses are performed specifically to *identify* them and retain them for further testing and commercial potential. The non-outliers are discarded.

Paradoxically to the increasing number of tested compounds, only a single measurement of each compound's activity is obtained in an initial **primary screen**. From a statistical point of view, each tested compound may be thought of as an individual experiment with a $n = 1$. Despite the improvement of the HTS process and the reduction of the research cost, fewer new drugs enter the market. Part of this may be because of the considerable noise, since the activity of each compound is determined on the basis of $n = 1$ value. One obvious improvement would be through replication and averaging. The absence of replicate measurements is mostly due to cost and time issues. Screeners need to be convinced of the benefit of replicates. Without replicates, the use of standard statistical methods is negated and scientists are forced to rely on strong untested assumptions. Replicates are also needed to verify assumptions of current methods and to suggest data analysis strategies when assumptions are not met.

In statistical analyses, compounds are typically assumed to have been randomly located in the wells of a plate, but the presence of row, column or well effects have sometimes been observed. For example, edge effects may be caused by evaporation at the edges and a better focus when reading the middle wells of a plate. Plates

containing more wells of smaller size (e.g. 16 rows $\times$ 24 columns = 384 wells) are starting to be used, but as the volume decreases, the effect of potential sources of error increases.

In order to introduce some of the issues, we present data from real **screens** performed on 384-well plates. Figure 1–1 shows the distributions of the raw data from two publicly-available screens with duplicate measurements (http://chembank.broad.harvard.edu/screens) for a yeast peptide inhibition assay (top half of figure) and a DNA synthesis assay (bottom half of figure). Although the distribution of data from the second screen is closer to Gaussian, the first one has three modes. How can we analyze such data? Which compounds may be deemed as hits? Should we first transform the data?

For the same two datasets, with the values now plotted in well order (row1column1, ..., row1-column24, row2-column1, ..., row2-column24, ..., row16-column1, ..., row16-column24), Figure 1–2 shows two types of variation. First, since the points belonging to a same plate are linked, from one curve to another one can observe plate-to-plate variability. In the yeast peptide inhibition assay, we notice that half of the plates have a higher signal in comparison to the other half for both duplicate measurements. I do not have enough information on the provenance of these data to explain why the two streams don't overlap, but this shift can be caused by several reasons such as a difference in environmental conditions, if the screen have been performed in two different days, or by the use of different batches of **reagents** and solutions. Second, each curve shows a similar 'zigzag' pattern which corresponds to within plate variability, more specifically to column effects, since higher values correspond to the first

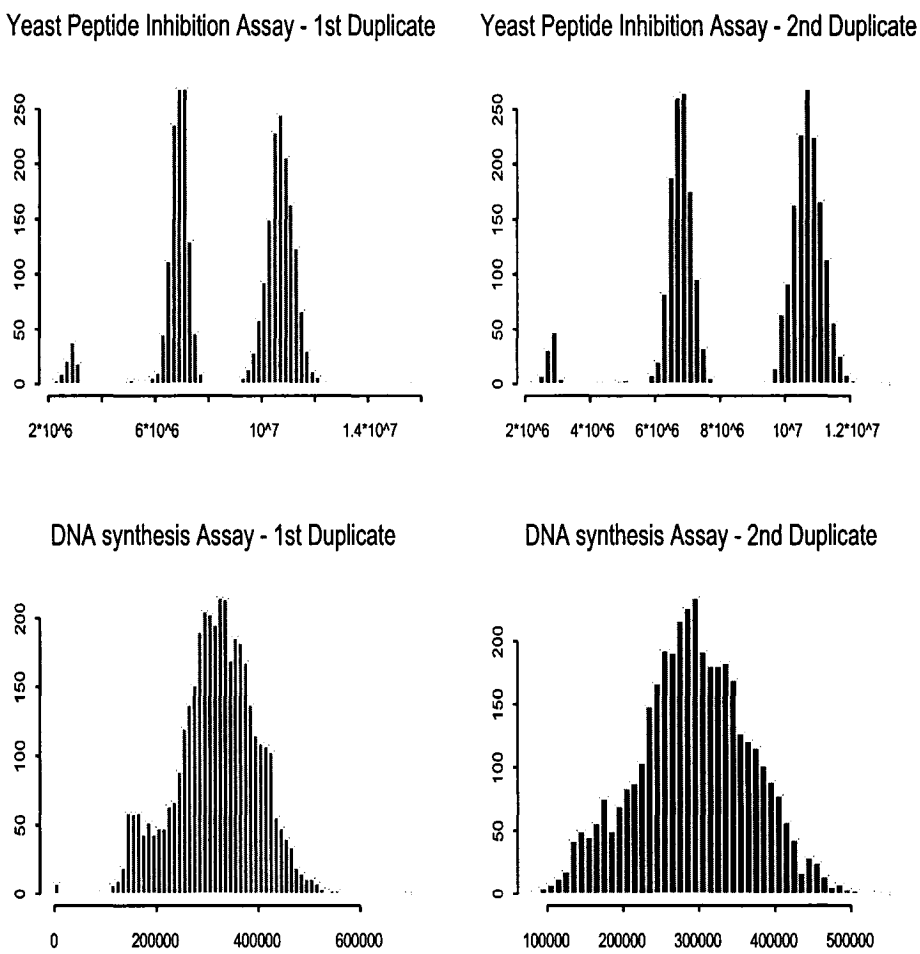

Figure 1–1: Histograms of raw data for two publicy-available screens (see text for details).

columns and lower values to the last columns. In addition, again in the yeast peptide inhibition assay, for the plates with higher signal, an important effect is observed on the last row since these wells have lower signal that all the others (right bottom

of the figure). This may have been caused by some procedural factor such as poor pipetting delivery or evaporation.

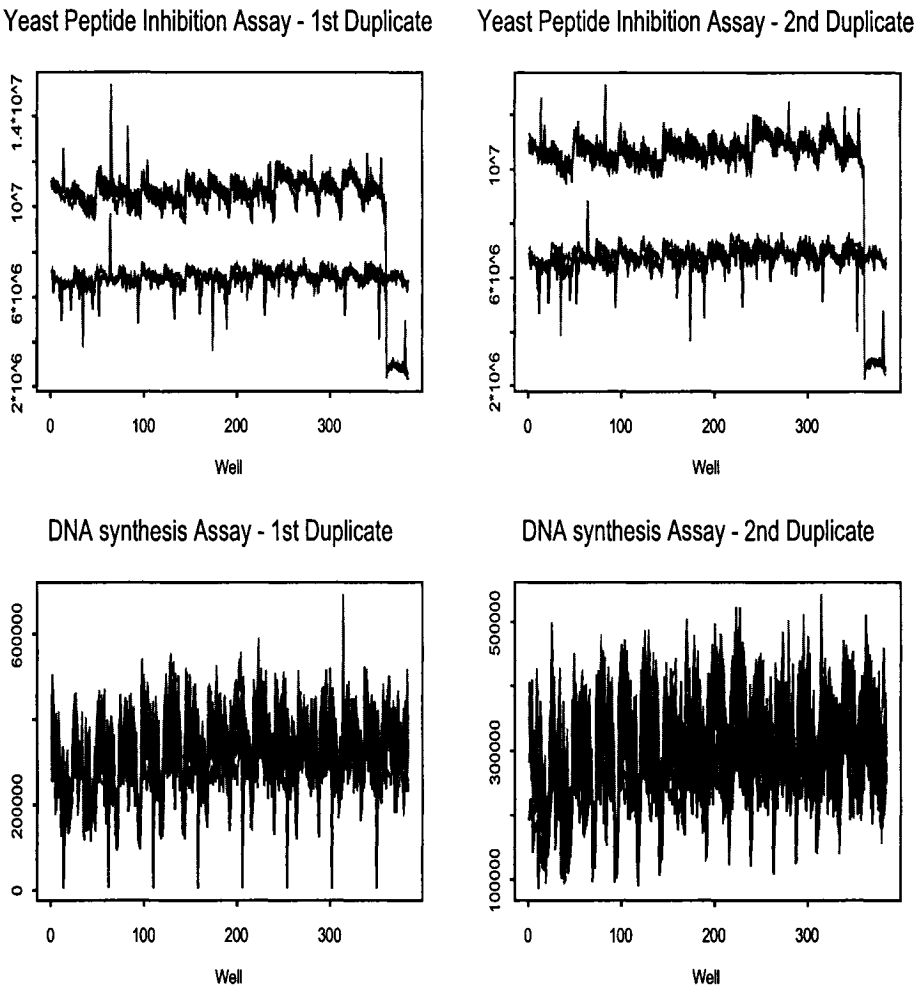

Figure 1–2: Raw Data vs well locations ordered by row (i.e. row1column1, row1-column2, ..., row1-column24, row2-column1, ..., row2-column24, ..., row16-column1, ..., row16-column24).

In the HTS literature, the possibility of positional effects that may occur in a real screen has never been fully assessed. Are the measurement errors systematic or random? Can we control for them in the laboratory process? Or must we do so at the data-analysis stage, and to what extent can we? I started my research by designing a small measurement experiment. The idea was simple: to test the *same* compound in each well in several plates, i.e. to repeat the exact *same experiment everywhere.* To minimize biases that may occur from procedural and technical factors, we randomized the plate processing order at every step of the protocol. Needless to say, this was very laborious for the life scientists! The observed variation in the values from different wells on the same and on different plates, even though the values are generated by the same compound, allowed me to observe the presence of errors in the measurements.

After a few days spent in the laboratory proceeding to the experiment, I realized that there are several potential sources of errors that may create noise. Unfortunately, some screeners tend to believe that results from automated technology do not contain any distortions. Robots are obviously faster and more reliable than manual work by humans; however, they are not infallible and may introduce their own biases (e.g. mechanical failures, differences in plate manufacturing etc.). Sources of errors may be biological, human and mechanical, and most of the time are of unknown origin, so they cannot be controlled during the HTS process itself. Thus, *preprocessing* of the raw data is required before any inference is done. By preprocessing I mean an efficient "normalization" of the data in order to reduce the effects of systematic errors.

At the same time, I started to consult the literature and ask screeners about the methods they currently use to analyze HTS data. I first noticed a lack of preprocessing methods. Most of the time, only plate-to-plate variation was corrected for by using biological controls. Because compounds are placed in the middle wells of a plate, controls have to be placed on the edges and thus, they may introduce their own biases. Brideau et al. [2] have recently introduced a method to remove row and column biases that do not use the controls, but it is not yet integrated into practice. As for inference, methods vary among laboratories; moreover, the choice of a significance threshold is totally arbitrary. When I heard comments like "don't worry about false positives unless the rate is very high"; "strive for highest possible quality and don't worry too much about the one that got away"; and "adjust the hit threshold until you have the number of hits you want", I realized that screeners are unaware of the importance of false positives and false negatives, and of how they can be affected by the methods used for preprocessing and by the criteria used for decision making.

The statistical community has also been slow to respond to the new inferential challenges posed by HTS data. Tukey's median polish [3], developed almost 30 years ago, would seem to be a natural tool. However, it was developed as an informal and general tool for analyzing data in a two-way layout. It has been used for several different purposes. In some instances, the focus is on obtaining an additive model for the data, with the necessary examination of outliers as a secondary/subordinate objective. Sometimes the focus is on interaction patterns. In some applications, the primary focus has been on detecting outliers as items of scientific interest, rather than

on identifying those of a nuisance nature. Even then, the median polish method did not allow for the estimate of the size of the outliers to be accompanied by standard errors or other such measures of statistical stability. Part of this lack of a precision measure may have stemmed from the fact that median polish has typically been applied to data with just one value per cell. Moreover, median polish was developed as an "exploratory data analysis" (EDA) tool, and it was not envisaged that it could be automated for HTS data. The refinement of methods for outlier detection seems to have been neglected for the next 25 years. It is only just recently that Terbeck and Davies [4, 5] have developed new robust methods to detect outliers in two-way data. However, there has been no formal evaluation of these newer methods and no comparisons with the earlier methods for outlier detection.

Moreover, no one has investigated how best to extend those older and newer methods to two-way data with replicates. Both the median polish and the more recent statistical methods are designed to work with a single observation per cell. I could not find any guidelines on how *replicates* should be analyzed. Consequently, there is an important need for new efficient statistical tools that may handle replicate measurements in order to improve HTS data analysis.

1.2 Objectives

The broad objectives of my thesis are to develop and evaluate new and efficient statistical methods for both data preprocessing and statistical inference for HTS data. The purpose of these tools is to better identify high quality hits with a high

degree of confidence, and to be able to do so in a semi-automatic mode in order to handle the increasing volume of raw HTS data being generated.

Since interest is on outlier detection, my first focus is on robustness. The use of statistics that are robust to the presence of outliers, and thus, can at the same time identify outliers, will give more reliable results than classical methods that are influenced by extreme values. For example, the use of a median instead of an arithmetic mean in the statistical analysis is a first step towards robustness.

My second focus is to justify and promote the use of replicate measurements in HTS practice. Although it may be expensive, I believe that the use of replicates in both preprocessing and inference will help to minimize false positive and false negative rates, and thereby increase the sensitivity and specificity of screens. Consequently, to demonstrate these benefits, I wish to provide proper statistical methods that allow replicates to be used in HTS data analyzes.

My thesis is divided in three parts, with results that are of interest to both life scientists and statisticians; each part is presented as a separate manuscript. I begin with a critical review of the current practices in HTS data analysis. This first manuscript also includes several recommendations to improve sensitivity and specificity of screens. The second manuscript is mostly of statistical interest. It compares different robust preprocessing methods to deal with replicated two-way data. The last manuscript is an application of the statistical methods described in the first manuscript to data from both real screen and in-house laboratory experiments.

In these chapters, for a better understanding of the biochemistry part, relevant technical terms are identify in bold and are defined in a glossary (appendix). Finally, I give a general conclusion, and discuss the potential impact of this research.

Preamble to Manuscript I

The first objective of my thesis is to better understand what is the HTS process and the currently practiced methods for data preprocessing and hit identification, while keeping in mind my statistical knowledge. Few statisticians are involved in HTS. Being one of them, I wish to convince screeners to be conscious of, to appreciate, and to deal with statistical issues that are present in actual HTS procedures.

Consequently, this manuscript is a critical look at the diverse statistical and non-statistical tools used to analyze the large amount of HTS data generated daily. In this manuscript, I also go further and recommended statistical methods that may be used to improve both preprocessing of, and inference from, HTS data. However, the presentation is mostly at a theoretical level and I restrict my attention to the two publicly-available data sets presented in the introduction.

This manuscript has been published in the February 2006 issue of Nature Biotechnology under the computational biology section. The reprint can be found in appendix. The references are included in the global thesis bibliography.

CHAPTER 2
Manuscript I - Statistical Practice in High-Throughput Screening Data Analysis.

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

[1] McGill University and Genome Quebec Innovation Centre, 740 avenue du Docteur Penfield, Montreal, Quebec, Canada, H3A 1A4

[2] McGill University Department of Epidemiology, Biostatistics, and Occupational Health, 1020 Pine Avenue West, Montreal, Quebec, Canada, H3A 1A4

[3] McGill University Department of Biochemistry, 3655 Promenade Sir William Osler, Montreal, Quebec, Canada, H3A 1A4

[4] McGill University Department of Human Genetics, 1205 avenue du Docteur Penfield N5/13, Montreal, Quebec, Canada, H3A 1B1

Abstract

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

2.1 Introduction

High-throughput screening (HTS) is the backbone of drug discovery within the pharmaceutical industry. Over the past decade it has also made its way into academic settings. The combination of robotic methods, parallel processing, and miniaturization of biological assays has dramatically increased throughput. The potential to increase the hit discovery rate has been offset, however, by increased research costs. Despite the current popularity of HTS and major improvements in processing, the new drug approval rate has declined significantly [6].

Developers are attempting to counter this inefficiency by various means, including developing biotech-pharmaceutical alliances and changing their internal organizational structures by merging multiple disciplines associated with **lead** generation and validation [7]. Likewise, HTS programs are being integrated within academic settings where alternative targets and diseases of lesser commercial value can be explored [8]. At the root, the challenge is to find the next marketable drug while simultaneously maximizing the number of screened targets and compounds, minimizing costs per well, and optimizing the lead generation and validation process.

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

HTS is a large-scale process (Figure 2–1) that screens many thousands of chemical compounds in order to identify potential lead candidates rapidly and accurately. Whereas the plating format and number of compounds per plate can vary, typically just a single measurement of each compound's activity is obtained in an initial primary screen. The automated process allows the testing of several hundred plates over a period of weeks. Compounds identified for follow-up (labeled "hits") are evaluated for biological relevance by a **counter screen** and confirmed as bona fide hits by a **secondary screen**.

Secondary screens test many fewer compounds (e.g. the 1% most active compounds from the primary screen, [9]) and typically use at least duplicate measurements. Paradoxically, compounds with the highest measured activity levels on a primary screen will on average be less extreme on a secondary screen because of a statistical artifact known as "regression toward the mean" [10, 11]. Accordingly, marginal hits on the first run may fail to validate on the second run merely because of random measurement error, although the size of the statistical artifact can be minimized by improving measurement precision (e.g. by obtaining replicate measurements). Confirmed hits with an established biological activity according to a structure-activity relationship (SAR) series and medicinal chemistry are termed "leads" that can develop into drug candidates for clinical testing.

Inferential errors can be caused by "noise" due to technical or procedural factors, including assay formats, poor pipette delivery, robotic failures and unintended differences in compound concentrations due to evaporation of solvent, either from the compound collection or during the assay set-up. Errors of unknown origin may

also develop over the course of the entire screen. Their adverse effects can often be minimized by quality control procedures, although statistical corrections may also be needed to mitigate the effects of uncontrolled variation (see "HTS Data Processing" section). Other factors which are less amenable to procedural quality control but which can nonetheless add extraneous variation include potency differences across compounds, and systematic across-plate and within-plate column or row biases (e.g. edge effects).

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

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

We review current data pre-processing and hit identification methods for primary screening. We discuss their use and limitations, problems with the constant

error assumption, the influence of hit threshold on false positive and false negative rates, and factors that can degrade assay sensitivity and specificity. We also discuss the advantages of replicates and make recommendations for the statistical analysis of HTS.

2.2 HTS Data Processing

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

Measured compound activity is a function of at least two factors: the compound's true activity and random error (see also "Use of Replicates" section). Symbolically, one simple additive model might be $Y_{ijp} = \mu_p + \epsilon_{ijp}$ where Y_{ijp} is the observed raw measurement obtained from the well located on row i and column j on the p^{th} plate, μ_p is the "true" activity and ϵ_{ijp} is the effect of all sources of error. Assuming no bias, the ϵ_{ijp} are assumed to have zero mean and a specified probability

distribution (e.g. normal). Another simple model is $Y_{ijp} = \mu_p + R_{ip} + C_{jp} + \epsilon_{ijp}$ where R and C represent plate-specific row and column artifacts, respectively, and ϵ_{ijp} represents remaining sources of error. (This latter model is assumed by the median polish procedure described below). Specifying models explicitly in this manner has the advantage of suggesting how sensitivity and specificity gains can be achieved most cost-effectively.

2.2.1 Current Practice

Because of the manner in which compound collections are plated, controls are typically placed contiguously on the outer columns. For example, Figure 2–2 shows the typical location of compounds and controls in a 96-well plate. Unfortunately, a systematic outer column effect affects all of the measurements on the plate because they are adjusted relative to these controls. For example, edge effects may lower (or increase) detection levels on average along the edge compared to the remainder of the plate. Consequently, background correction will be lower (or higher) if controls are located on this edge, causing compound activities to appear higher (or lower) than their true states. Worse still, the edge effects may be present in some plates but not others (see "Recommendations" section below). Cell-based biological controls are especially problematic because of variable growth patterns [12]; cell clumping or evaporation within different areas of the plate can lead to different growth conditions and ultimately to position-related bias. Regardless of cause, positional effects increase the rate of false positives and false negatives.

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

The three methods described above implicitly assume a random error distribution that is common to all measurements within a single plate, although without replicates this assumption cannot be verified directly. Positive and negative controls may exhibit differences in variability, however, raising questions about the constant error assumption. Differences in variability among compounds are also likely inasmuch as inactive compounds are similar to negative, and active compounds are similar to positive controls [13]. For example, Figure 2–3 shows results from a titration series of a protein translation assay in which variability among replicates differs across the various concentrations. In general, non-constant variances among the compounds of interest may be due to differences in luminescence, reactivity, or solubility. The serious errors of inference that can arise from incorrectly assuming one distribution

even when departures from it are minimal, have been cogently described by Tukey [14].

Another potential difficulty is that these three methods rely on non-robust statistics. Means and standard deviations are greatly influenced by statistical outliers, which in the context of HTS are putative hits. In statistical terms, the mean and the standard deviation have low breakdown points, in contrast to more resistant location and scale estimators (e.g. median, Tukey biweight, median absolute deviation (MAD)). One recent proposal circumvents these issues by adopting a more robust data analysis procedure.

The B score [15] is a robust analog of the Z score which uses an index of dispersion that is more resistant to the presence of outliers and more robust to differences in the measurement error distributions of the compounds (Box 2) . A two-way median polish is first computed to account for row and column effects of the plate. The resulting residuals within each plate are then divided by their MAD to standardize for plate-to-plate variability. The B score has three advantages: it is non-parametric (i.e., makes minimal distributional assumptions), it minimizes measurement bias due to positional effects and is resistant to statistical outliers.

2.2.2 Recommendations

In the absence of compelling reasons to the contrary, we prefer normalizing the data without using controls. Specifically, we prefer the B score method, especially if row or column biases are suspected. Evidence of these biases can be obtained by examining the variability of the row and column effects estimated by the median polish

procedure relative to the residual compound measurements. To illustrate, we reanalyzed two publicly-available screening data sets with duplicate measurements for a yeast peptide inhibition assay and a DNA synthesis assay (http://chembank.broad.harvard.edu/screens; Screen numbers 295 and 900, respectively). Figure 2–4 shows a strong and variable column effect for Screen 295. Moreover, as we demonstrate in the "Use of Replicates" section, the variability of B scores may more adequately reflect actual random error conditions. This in turn facilitates the decision process because the compound measurements can be benchmarked against theoretical error distributions.

If researchers were to use the Z score method, we would advise they use robust versions in order to minimize the undesirable influence of outlier compounds (i.e. "hits"). For example, in a "jackknife" Z score method, $\bar{x}$ and s_x (third equation in Box 2) are calculated excluding the compound of interest (x value in the equation); accordingly, s_x differs for each individual compound. Alternatively, in a "robust" Z score method, $\bar{x}$ and s_x are replaced by more robust measures (e.g. median and MAD, respectively).

Controls, if necessary for a specific assay, should be used carefully. Ideally, they should be located randomly within plates, thereby minimizing row or column biases. Current compound collection formats, however, do not lend themselves to randomization. Potential positional effects can nonetheless be minimized by varying the location of controls within plates in a systematic manner. One way consists of alternating well-locations for the positive and negative controls along the available edges of the library (Fig. 2–2a) . Thus, positive and negative controls will appear

equally in each row and in each column and may minimize edge-related bias. For example, in a 96-well plate, an order effect may produce different biases among the different columns. In such a case, the alternating method (Fig. 2–2b) will be more efficient than current practice consisting of 8 positive controls on the first column and 4 negative controls on the last column (Fig. 2–2a) .

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

2.3 Statistical Inference: Threshold for Hit Identification

Regardless of library design strategy (rational or combinatorial), statistical methods offer the means to characterize quality of screens and of hits within a probabilistic framework. Quality can be defined as the ability of the screening process to accurately identify compounds that can be developed into potential leads [16]. A statistical approach to these issues has a number of advantages, including objectivity, reproducibility, and ease of comparison across screens.

Once data have been pre-processed with quality control checks and normalization procedures, the next critical step is to decide which compounds should be processed in a secondary screen. Currently, this inferential process is not well defined statistically: procedures for hit identification are based on informal rules of thumb rather than on probabilistic judgments of error rates. In academic settings and in smaller companies, informal rules may also be based on particular laboratory constraints such as capacity limitations. Although it is generally appreciated that lowering hit-threshold increases false positive rates while lowering false negative rates, statistical models can better quantify the balance between specificity and sensitivity by assigning probabilities to the two types of inferential errors (Fig. 2–5)

.

2.3.1 Current Practice

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

Hits can also be identified as a percentage of the compounds that generate the highest measured activity (e.g. top 1%, [9]). From an optimization perspective, this method is arbitrary and suffers from the absence of a probability model. Without prior consideration of the true number of active compounds, one cannot optimize

the percentage of primary screen compounds to be screened a second time. If the number of identified potential hits is dictated by the capacity for secondary screening, specificity and sensitivity may vary widely across screens. Consequently, the quality of the results from screen to screen within a laboratory will depend on the extent to which threshold choice reflects the actual number of true active compounds in the various screens.

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

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

As with the pre-processing methods described earlier, the threshold methods described above assume a common magnitude of random error for all measurements and rely on non-robust statistics, which may lead to inferential errors in hit detection. Hit detection depends jointly on compound concentration, potency, and variability.

Potency will differ across compounds within a screen, as will actual concentrations due to uncontrolled factors such as solvent evaporation and compound solubility. The easiest hits to detect will be compounds with high relative potencies and concentrations and low variability. The titration series in Figure 2–3 illustrates this issue. Singlet measurement false positives for the three lowest non-null concentrations were eliminated when activity measurements were based on means across the eight replicate measurements per concentration. Methods which estimate random error without assuming constant error are described in "Use of Replicates: Recommendations" below.

2.3.2 Recommendations

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

Verifying data handling assumptions and contrasting various approaches in formal methodological studies are important steps in determining the most cost effective procedures. Additivity assumptions, for example, can readily be verified from a simple graphical procedure once the data have been pre-processed by the median polish procedure [17]. This same procedure provides a simple method for determining the appropriate data transformation (e.g. log) which will produce additive measurements.

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

2.4 Use of Replicates

Random error reflects inevitable uncertainties in all scientific measurements. This "noise" unpredictably raises or lowers measurements relative to their true values. Potential sources of random error include biological, instrument, and human-related influences. Random error accumulates as a collection of several minimal differences across assays, such as voltage variation, liquid dispensing differences, as well as reagent or sample preparation and handling [16]. Compound-related problems involving chemical properties and activity (e.g. stability, solubility, auto-fluorescence and degradation) also affect measurement precision.

Precision can be increased by obtaining replicates and by minimizing extraneous variation due to sample handling and processing. Random error estimates, which are central to statistical inference, are typically obtained from replicate measurements of the same attribute or process. Having empirical estimates of variability allows one to use statistical power analysis to control the false negative rate while maintaining a fixed false positive rate (Fig. 2–5) . We anticipate that obtaining replicates for at least some compounds in primary screens will become more routine.

2.4.1 Current Practice

Compounds in primary screens are typically measured only once because of time and cost issues, although the use of duplicate measurements has been recognized for secondary screens and is beginning to be recommended for primary screens (http://iccb.med.harvard.edu/screening/guidelines.htm). Absent replicates, strong assumptions must be made in order to estimate random error. For example, Buxser and Vroegop [18] describe an approach in which the variability among replicated

control measurements is used to estimate variability of the unreplicated compound measurements. Alternatively, random error can be estimated from the variability across single measurements of all compounds on a plate, assuming that all compounds are inactive and that they all have the same random error; early approaches to gene expression microarray analysis adopted a similar approach for estimating error from single measurements [19]. Single measurement methods have ultimately proven inadequate [20], however, and it is now standard practice to obtain at least three replicates per measurement in recognition that replicates offer advantages which outweigh short term cost considerations [21, 22].

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

in compound pooling, but may arise if pooled compounds have opposite biological effects of similar size [7].

2.4.2 Recommendations

Replicates offer the twin advantages of greater precision for activity measurements and the means to estimate variability associated with the measurements. Compared with the uncertainty of a single measurement, the imprecision (standard error) of a mean is reduced by $100 \times (1 - 1/\sqrt{n})\%$ where n refers to the number of replicates. Having two replicates reduces imprecision by 29%; having three replicates reduces it by a further 13% while having four replicates reduces it an additional 8% (i.e. to 50% of the imprecision associated with a single measurement). Thus, replicates make minimally and moderately active compounds easier to detect.

Replicates may appear in wells on the same or on different plates. Although within-plate variation (due, for example, to plate composition and handling) will typically be smaller, across-plate replication is preferred because it represents a more realistic estimate of variation necessary for generalizing results beyond the immediate sample. In general, it is important to obtain estimates of total variability of any measurement process, what has been called "genuine replication" [23].

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

caveats notwithstanding, minimal replication can be used to examine the reasonableness of current assumptions and to potentially improve overall screen sensitivity and specificity.

We illustrate the importance of pre-processing, the need to check assumptions regarding error distributions and the other options available when assumptions are not met, by performing additional analyses on the Figure 2–4 data. If the errors associated with the normalized compound measurements from these screens were normally distributed with constant variance across compounds, the sample variances based on the duplicate measurements would follow a $\chi^2_{(1)}$ distribution (Box 3). Figure 2–6 illustrates the lack of fit, however, between the theoretical and the observed variance distributions for these data, indicating that the normality/constant variance combined assumption is not tenable after pre-processing by either the B score or the Z score procedures.

Alternatively, one can assume that the error associated with compound measurements is normally distributed but with unequal variances distributed across the compounds according to an inverse gamma distribution . An Empirical Bayes approach using this model has been used successfully for analysis of microarray data with minimal replication [20, 24, 25]. Figure 2–7 shows that the error variances of the data sets from Figure 2–6 fit an inverse gamma distribution for both data sets for the B scores and for one of the data sets for the Z scores. An important advantage of this variance distribution pattern is that standard statistical tests of compound activity can be constructed using a weighted average of the compound-specific variances estimated from replicated measurements and the overall estimate obtained

from the variance distribution; when only a random subsample of the compounds has been replicated, the latter variance estimate can be applied to compounds measured only in singlet from the same screen (Box 4) . In either case, the more similar the compound-specific variances are to each other, the more reliable the overall variance estimate will be. This in turn will provide more degrees of freedom and more power for the statistical tests. Figure 2–7 also illustrates the value of correcting for row and column effects. In the presence of column or row biases (screen 295), B scores more closely approximated the theoretical inverse gamma distribution than the corresponding Z scores, although in their absence (Screen 900) the B score method produced a slightly poorer fit.

As more extensively replicated data sets become available, other data analytic approaches can be examined and optimized. For example, although we found no evidence of a relationship between signal intensity and replicate variability in the two data sets we examined, such a relationship has been used in the microarray context in combination with the inverse gamma variance distribution assumption [26]; this type of relationship may provide additional useful information for estimating random error associated with replicate and singlet measurements. Similarly, if various classes of compounds are thought to differ in terms of variability, random subsets of the various classes may produce more accurate estimates of variability when examined separately. Another approach which may show promise is to model the distribution of activity measurements as a mixture of two distributions (inactive and active compounds) [18] . In short, the principle of "borrowing strength" from information available from the data in total can provide useful information that would normally

only be obtained from large numbers of replicates.

2.5 Conclusions

Statistics currently serve a limited role in HTS. One use is to correlate chemical properties with activity levels at the screen development stage to provide information for compound selection and for property modification to enhance chemical activity [27, 28]. Once the screen has been run, data mining software packages are increasingly being used for quality control. Notwithstanding these advances in data analysis, HTS continues to lack universal procedures for processing and extracting knowledge from screens [29]. We discuss four broad conclusions below that we believe are warranted at this early stage of development for the statistical modeling of HTS data.

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

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

The assumption of a common error variance across compounds implicit to many current hit identification approaches is incorrect at least some of the time. At a minimum, the assumption should be routinely verified by replicating some of the

compounds and checked against theoretically-derived distributions. When the assumption of constant error is untenable, the Empirical Bayes approach to estimating random error offers an attractive alternative. It provides an amalgam of the specific within-compound variations (if measured in replicate) and the error estimate derived from the distribution of the within-compound variances, with the latter alone providing the "best" estimate when a particular compound has not been replicated. This combination of sources of information is a compromise between using only the within-compound (and thus highly variable) error estimates and the average but unrealistic (and thus falsely precise) pooled error estimate that would be appropriate under a common error model.

The limitations of standard statistical approaches with minimal replication can be partially offset by "borrowing strength" from the large number of available measurements (compounds). We have provided one example of this principle by using the distribution of sample error variances to obtain error estimates for individual compounds.

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

potentially costly false negatives.

2.6 Boxes

2.6.1 Box 1 : Formulae for Normalization

Percent of Control: A qualitative measure of test compound activity defined as:

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

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

Normalized Percent Inhibition: Another normalization method using controls:

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

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

Z score: A simple and widely know normalizing method calculated as:

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

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

B score [14]: The residual (r_{ijp}) of the measurement for row i and column j on the p^{th} plate is obtained by fitting a two-way median polish and is defined below:

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

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

For each plate p, the adjusted median absolute deviation (MAD_p) is obtained from the r_{ijp}'s. The B score is calculated as follows:

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

Median Absolute Deviation (MAD): A robust estimate of spread of the r_{ijp}'s values:

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

2.6.2 Box 2: Examining the Distribution of Sample Variances

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

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

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

For each compound, consider the model:

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

where $k = 1, 2, ..., K$ replicates and it is assumed that:

$$\epsilon_k \sim N(0, \sigma^2).$$

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

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

The a and b parameters are assumed to be constant across compounds and can be estimated from the data from all compounds by fitting an F-distribution to the sample variances s^2:

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

Wright and Simon's [12] procedure for estimating the a and b parameters was used to generate the data shown in Figure 2–7.

2.6.3 Box 3: Test Statistics for Hit Detection with Replicates

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

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

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

"Modified" one-sample *t*-test: After estimation of the a and b parameters by fitting an inverse gamma distribution to the set of variances across replicates for each compound (see Box 3), a variation of the previous standard t-test is:

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

where

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

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

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

of the compound-specific and the "typical" variances for the B scores, and similarly 1:1.11 and 1:4.12 for the Z scores.

For an unreplicated compound, so that $K - 1 = 0$, $\tilde{s}^2$ is simply the typical value, estimated by the quantity $(ab)^{-1}$ with $2a$ degrees of freedom (for example approximately 6 for the B scores), which is a compromise between zero degrees of freedom associated with single measurements and $number of compounds - 1$ degrees of freedom (i.e., 2687 and 3839 degrees of freedom, respectively for screen 295 and 900) associated with a common error model.

2.7 Figures

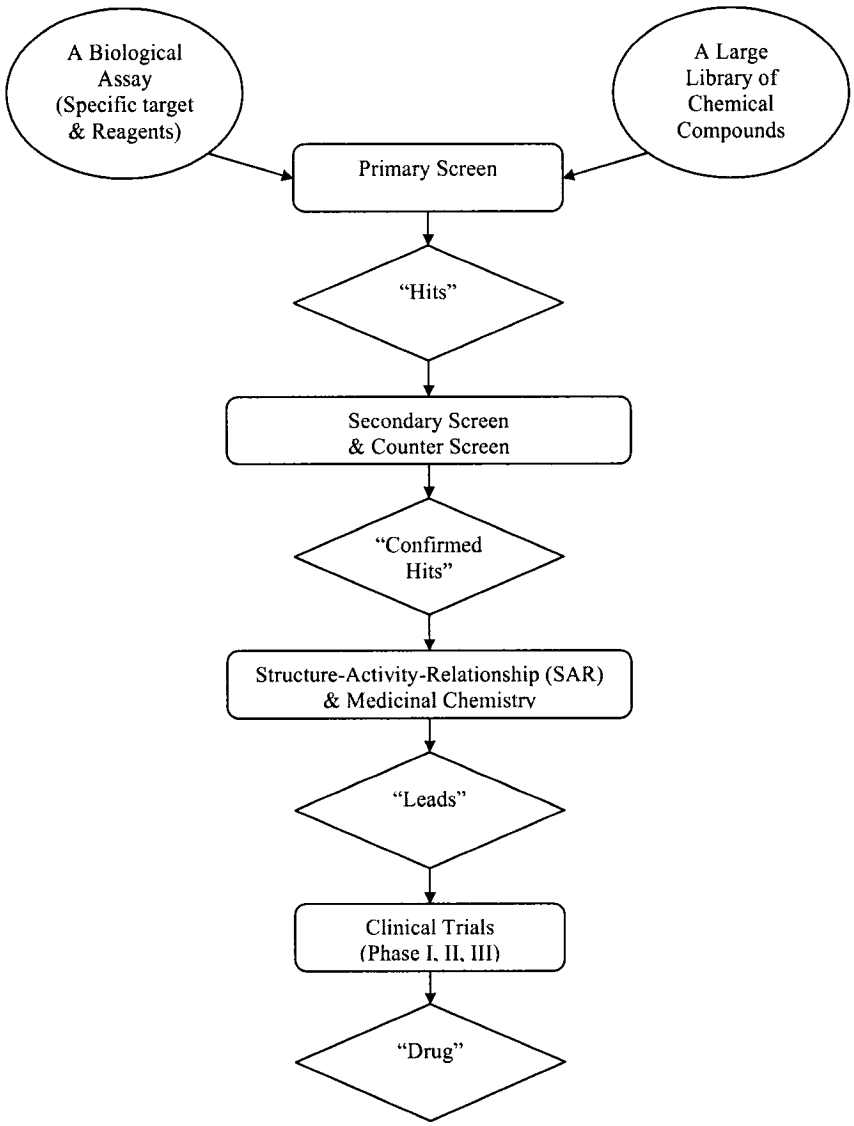

Figure 2–1: From HTS process to eventual drug development.

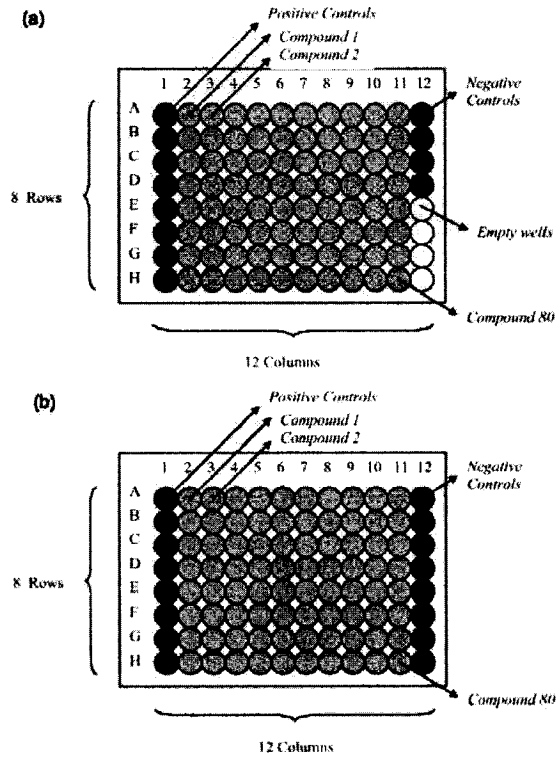

Figure 2–2: Typical location of controls on a 96-well plate. In a primary screen, the designed biological assay is performed by using a robot to add the target of interest and specific reagents to each well, which already contain a different compound or control. After incubation or other required manipulations, an activity measurement is obtained for every well by automated plate reading. These raw data represent the activity measurement of each compound or control against a specified target. The measurement units and the scales depend on the design of the biological assay, the target of interest and the specific reader or imager that is used. *(a)* Generally, in a compound library, 80 different compounds are stored in the middle of a 96-well plate and wells on the first and last columns are left empty. Often in a high-throughput screen, eight positive controls are placed in column 1 and four negative controls are placed in column 12. The others four wells in column 12 remain empty and are not used. *(b)* Ideally, controls should be located randomly among the 96 wells of each plate. Only the first and the last columns are typically available for controls, since compounds are stored in the 80 middle wells. Despite this limitation, edge-related bias can be minimized by alternating the 8 positive controls and the 8 negative controls in the available wells, such that they appear equally on each of the 8 rows and each of the 2 available columns.

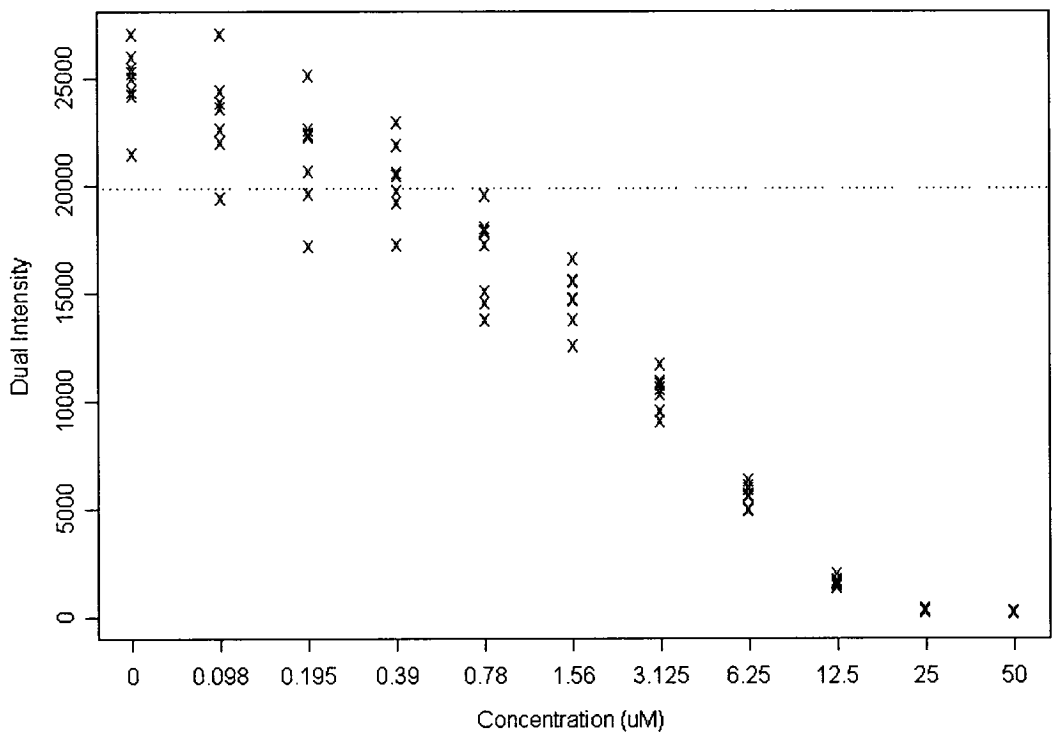

Figure 2–3: Titration series in a translation assay. These results from an Anisomycin titration in a Renilla luciferase translation assay show that variability differs across the various concentrations. A hit may be defined as any activity measurement that is at least 3 standard deviations away from the mean of the control measurements. This corresponds to a dual intensity value of 19894 (dashed line). All of the measurements for concentrations greater than or equal to 0.78 are hits (all of the values are below the dashed line). There were six false positives, however, for the three lowest non-null concentrations.

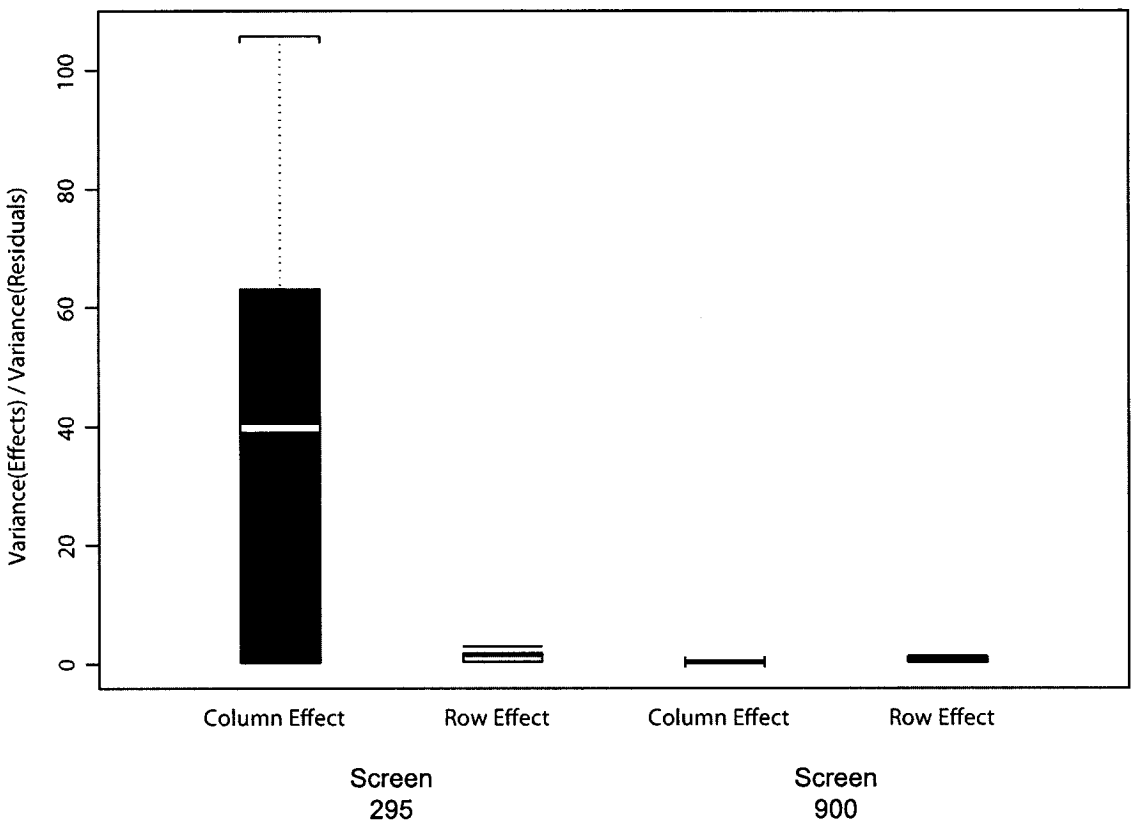

Figure 2–4: Presence of edge effects in a high-throughput screen. Data from two different screens (http://chembank.broad.harvard.edu/screens) with duplicate measurements across plates are presented. Tukey's two-way median polish was applied to each plate in order to obtain estimates of row and column effects and of residuals (i.e. compound measurements after the polish procedure removed any row and column effects). For each plate, variances of the 16 row effects and of the 24 column effects were divided by the variance of the 384 residuals. Boxplots of these variance ratios illustrate the presence of a column effect for Screen No. 295.

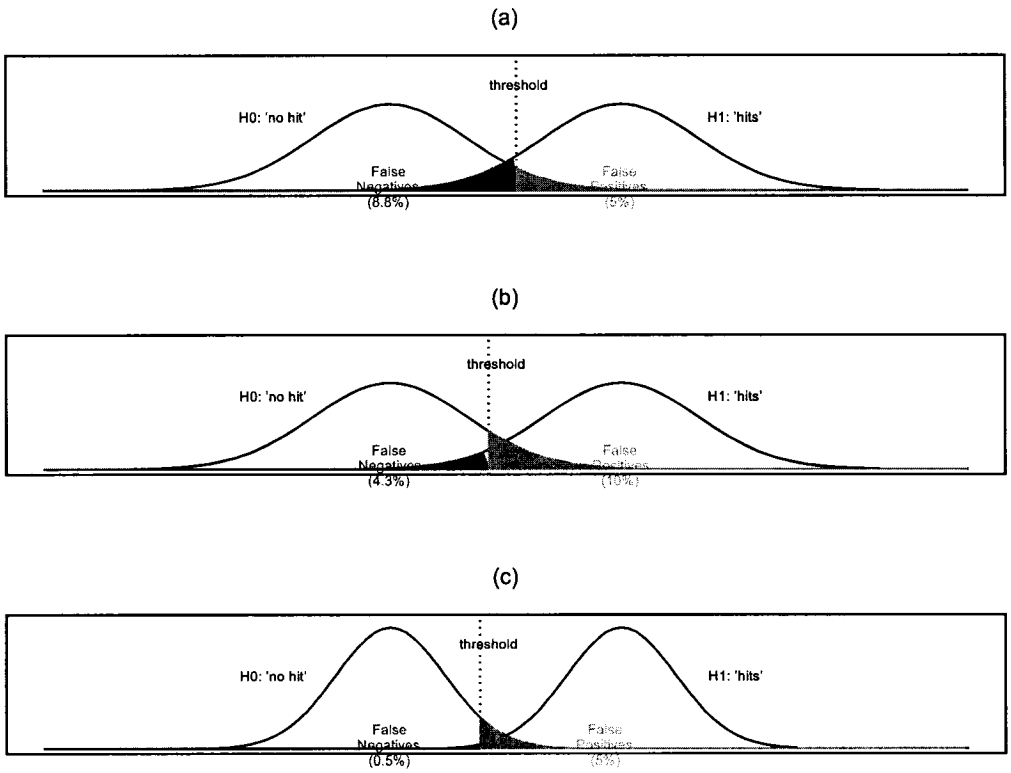

Figure 2–5: Replicates, false positive and false negative rates. In hypothesis testing a false positive rate (Type I error) is the probability of rejecting the null hypothesis (H0) given that this hypothesis is true. The false negative rate (Type II error) is the probability of failing to reject the null hypothesis (H0) given that the alternative hypothesis (H1) is true. *(a)* Given a fixed threshold value, the false negative and false positive rates are represented by the blue and the red areas under the curve, respectively. *(b)* Decreasing the threshold value results in an increase in the false positive rate and a decrease in the false negative rate. The opposite would be true if the threshold value were increased. *(c)* The benefit of multiple measurements (replicates) is illustrated. The use of replicates reduces data variability which is reflected in the narrowed data distributions. Consequently, the false negative rate is minimized while the false positive rate remains fixed.

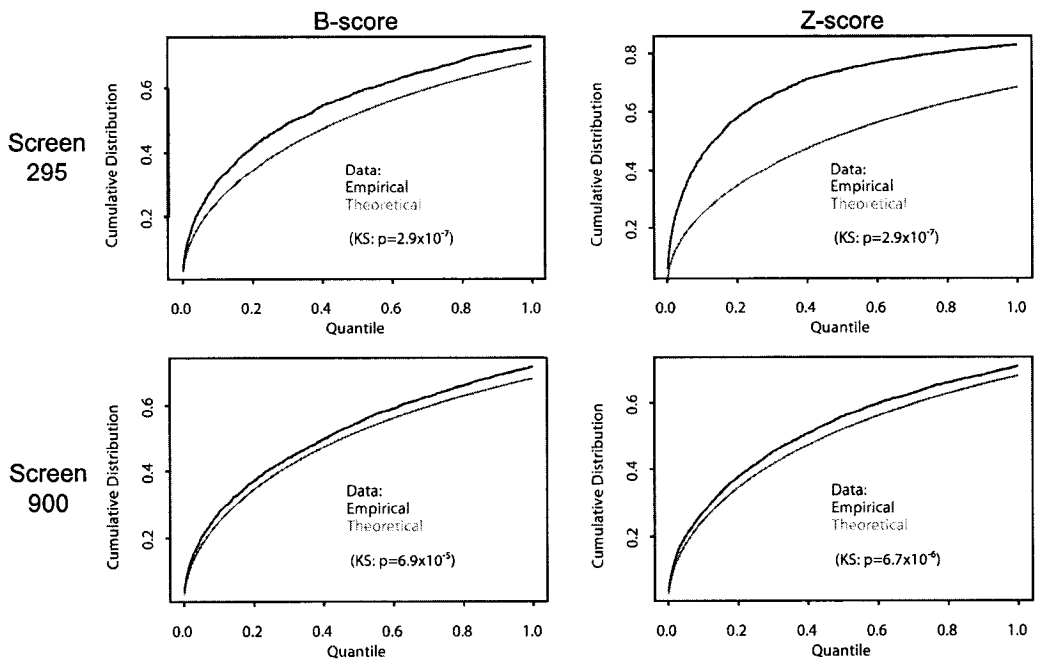

Figure 2–6: Verification of the assumptions of normally distributed data with constant variance among compounds. Empirical values correspond to a function of the sample variances. Under the assumption of a constant variance among compounds, the overall variance might be estimated by the mean of the sample variances. Each sample variance (obtained from the duplicate measurements) is divided by the overall variance estimate and the ratio should follow a chi-square distribution with 1 degree of freedom (Box 3). Results of the Kolmogorov-Smirnov (KS) test of differences between the theoretical and the empirical distributions are shown.

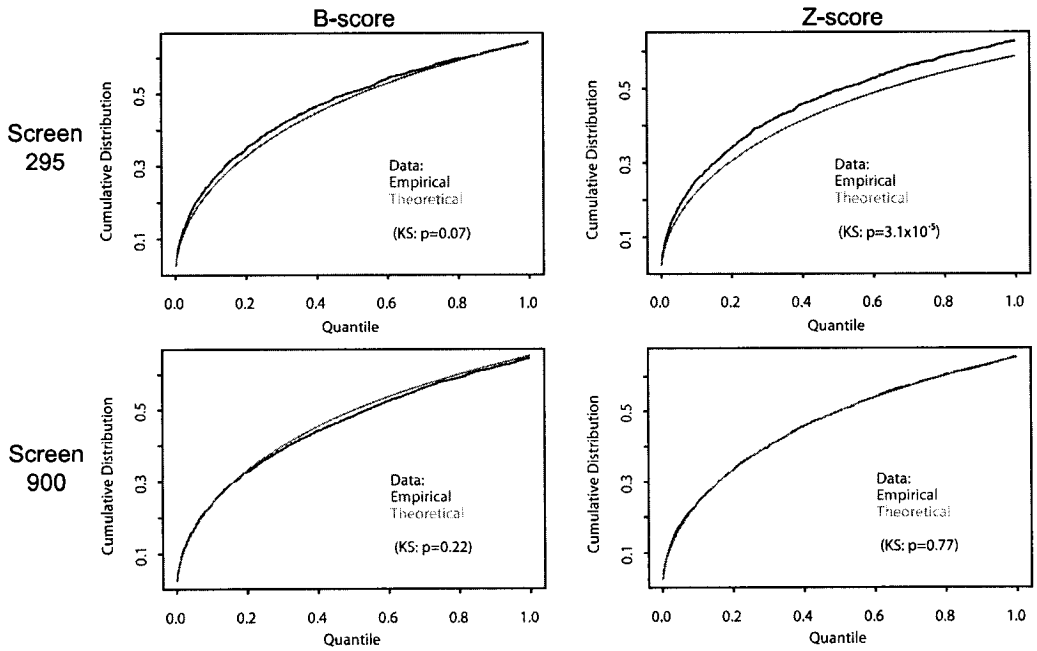

Figure 2–7: Verification of the assumption that the within-compound variances follow an inverse gamma distribution. Empirical values correspond to a function of the sample variances. Under the assumption of normally distributed data, each sample variance (obtained from the duplicate measurements) is multiplied by the estimated a and b parameters of the inverse gamma distribution and the result should follow an F distribution with 1 and 2a degrees of freedom (Box 3). Results of the Kolmogorov-Smirnov (KS) test of differences between the theoretical and the empirical distributions are shown.

Preamble to Manuscript II

Current HTS practice does not include replicate measurements. However, replicates are being increasingly appreciated since they (i) allow aggregated measurements, thus reducing the variability of these averages on which decisions are made, and (ii) offer the advantage of improving both sensitivity and specificity of screens.

During the course of my work on the previous manuscript, I realized that as currently practiced, robust methods cannot handle, or it is not clear how they should handle, replicates. For example, a two-way median polish [3] can clearly be applied to repeated measurements. However, the statistical literature does not provide detailed technical guidance on the appropriate algorithm to use.

The purpose of this second manuscript is to evaluate and compare, via simulation studies, the performance of different robust preprocessing methods when applied to replicated two-way data with respect to detection of outlying cells.

This manuscript will be submitted to JASA. The references are included in the global thesis bibliography.

CHAPTER 3
Manuscript II - Robust Efficient Identification of Outlying Cells in a Two-Way Layout with Replicates.

Nathalie Malo[1,2], James A. Hanley[2], and Robert Nadon[1,3]

[1] McGill University and Genome Quebec Innovation Centre, 740 avenue du Docteur Penfield, Montreal, Quebec, Canada, H3A 1A4

[2] McGill University Department of Epidemiology, Biostatistics, and Occupational Health, 1020 Pine Avenue West, Montreal, Quebec, Canada, H3A 1A2

[3] McGill University Department of Human Genetics, 1205 avenue du Docteur Penfield N5/13, Montreal, Quebec, Canada, H3A 1B1

Abstract

The increasing amount of two-way data and the recent movement towards using multiple measurements (i.e., replicates) in diverse research applications have lead to renewed interest in robust methods for detecting outlying cells. Residuals from such robust methods are not affected by the "leakage" produced by those from a least squares fit. Thus, outlying cells are easier to detect. However, the statistical literature provides no technical guidance on how older and newer algorithms should be modified or adapted to handle multiple observations per cell. We compare the performance of four preprocessing options with respect to detection of outlying cells, which are defined by four inferential rules.

Tukey's median polish is a preprocessing method which was introduced in the 1970s as a general statistical tool. The residuals are those obtained in a classical two-way ANOVA model, but using medians rather than means. To overcome the potential lack of uniqueness of the L^1 solution, Terbeck and Davies [4], and Davies [5] have developed robust methods based on M-estimators. Using these older and newer methods, we consider four different options for obtaining residuals in replicated two-way data. The options are: median polish applied to individual values, median polish applied to the cell medians, and two methods from Davies [5] also applied to the cell medians. We adopt four arbitrary inferential rules to define outlying cells. ROC curves are used to compare tests, while effectively maintaining a constant test 'size'. A median polish applied to *individual values* perform best in detecting an *single* outlying cell. This method is also the most accurate of the four when applied to a real dataset. In contrast, in the presence of *several* outlying cells containing extreme

signals, preprocessing methods applied to *cell medians* have the best performance. Median polish also offers the advantage of being easier to understand and faster to compute than Davies [5] methods. We recommend the use of median polish applied to individual values, especially when interest is on detecting outlying cells with a small effect size.

3.1 Introduction

Technological advances in several scientific fields have led to very large data sets with a two-way (row/column) structure. Row and column effects may be of direct scientific interest, or merely a nuisance to be dealt with. Irrespective of the primary focus, there is a need for automated robust methods: the large number of two-way tables, and the resulting volume of data preclude a detailed table by table examination. Areas of interest include estimating additive main effects (rows and columns) and detecting non-additive effects (individual outlying cells or interaction patterns).

Tukey's median polish [3] is an early example of an exploratory statistical tool which has been used to examine two-way data structures. However, both its use, and theoretical study of its properties, subsequently declined. More recently, however, it has been revived for data-intensive applications such as geostatistics [30, 31], microarrays [32], and high-throughput screening (HTS) of chemical compounds [2].

Although the objectives of the contemporary use of median polish may differ across applications, the algorithm is mathematically and computationally the same. For example, in geostatistics, spatial data are obtained for irregularly distributed sampling locations, which results in a two-way array with missing values. The goal is to predict the phenomenon under study at unobserved locations using the correlation between neighboring observations. Median polish is used to provide robust and accurate estimates of spatial trends. Applying kriging methods to the residuals from the median polish is a powerful way for spatial estimation and prediction, since it eliminates biases caused by spatial trends. Median polish is also commonly

applied to Affymetrix microarray data in which each gene's expression is estimated by numerous gene-specific probes. Here, the scientific focus is on obtaining gene expression estimates after removing microarray (columns) and probe-specific (rows) biases on a gene-by-gene basis. In HTS, several thousand chemical compounds are tested in a single experimental run involving hundreds of plates. Each plate contains a two-way array of wells (e.g. 8 rows by 12 columns); often the first and last columns are used for positive and negative controls so that each 96-well plate can accommodate 80 compounds. Median polish can be used to minimize processing biases which can create artifactual row and column effects within plates. The median polish preprocessing is a necessary step prior to the primary focus, which is to identify large residuals (outlying values) which represent active compounds that may later be developed into a drug. Median polish use in these data-intensive applications has raised several issues that were not even considered in its original applications.

An alternative to the median polish, especially for the identification of outlying cells, has been proposed recently by Terbeck and Davies [4], and Davies [5]. Their methods, based on M-estimators, circumvent the potential lack of uniqueness of median polish and guarantee scale invariance.

Both the median polish and the Davies' methods were designed to work with a single observation per cell. In two-way biomedical data of the type described above, there is a movement towards obtaining replicate measurements. After initial reluctance, their benefit is now more widely recognized in microarrays [33]. The same appreciation of the advantages of replicated measurements is beginning to be recognized in HTS applications [34]. For technical reasons, entire plates are replicated, so

that replicated compounds are located on the same well of different plates. However, the statistical literature provides no technical guidance on how median polish should be modified or adapted to handle multiple observations per cell, even though the literature does explain how median polish can be applied to a three way layout [35]. Terbeck and Davies [4] and Davies [5] also restricted their attention to the simple case of one observation per cell. Applications of their methods to two-way data with replicates have not been investigated.

The focus of this paper is to compare the performance of four different options for dealing with replicates, when the ultimate task, after pre-processing, is to detect outlying cells in a two-way layout. The paper proceeds as follows. In section 2, we introduce the notation and describe four preprocessing procedures for handling replicates, namely two ways of adapting median polish, and one way of handling replicates using two methods from Davies [5]. In section 3, we describe a simulation study to compare performance of (i) the four preprocessing procedures and (ii) four inferential rules applied to the preprocessed data for defining outlying cells. ROC curves are used to compare performance. Results are presented in section 4. The four preprocessing options are applied to a real dataset in section 5.

3.2 Background

When the data consist of a single observation per cell ($K = 1$) in a two-way table, the usual focus is on the standard additive model

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

where $1 \leq i \leq I$ and $1 \leq j \leq J$. The row effects (α_i), column effects (β_j), and the grand mean (μ) are typically estimated by minimizing the sum of squared residuals

$$SSR = \sum_{i=1}^{I}\sum_{j=1}^{J}(\hat{\epsilon}_{ij})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)^2.$$

The analytic solution satisfying this L^2 criterion is given in the balanced case by

$$\hat{\mu} = \frac{1}{IJ}\sum_i\sum_j y_{ij}; \quad \hat{\alpha}_i = \frac{1}{J}\sum_j(y_{ij} - \hat{\mu}); \quad \hat{\beta}_j = \frac{1}{I}\sum_i(y_{ij} - \hat{\mu}).$$

It is helpful to note for the discussion of the median polish below that the residuals produced by this 'analysis by means' or 'mean polish' equal those obtained by a three step process: (i) find the grand mean and subtract it from all observations; (ii) subtract the mean of each resulting row from all cell values in the respective row; (iii) subtract the mean of each new column from all values in that new column.

The least-squares method, however, has poor resistance to outliers. Because of the restriction $\sum_i \sum_j \hat{\epsilon}_{ij} = 0$, residuals in all cells and estimates of row and column effects will be heavily influenced by the presence of an extreme value in one cell. One important consequence of this "leakage" problem is that the outlying cell will be less distinct after the polish, as illustrated by a simple example [36, 4]. Consider a 3×3 table with zero values in 8 of the cells, and a value of 9 in a single outlying cell. Table 3–1 shows that the L^2 criterion applied to these data leads to nonzero residuals in all cells, and that the residual in the outlying cell is now only 6 units higher than those in each of the others. By contrast, after the median polish, 8 of the residuals remain at zero, and the value in the outlying cell remains at a distance of 9 from these values, and thus, is more readily detected.

3.2.1 Robust Preprocessing Methods for Two-Way Data

In contrast to the analysis by means, robust methods protect the fit from being distorted by extreme values and yield better estimates of the main effects, especially for contaminated or long-tailed data. We examine different robust approaches.

Tukey's Median Polish.

By analogy with the ANOVA model, Tukey's median polish is another simple method of fitting the standard additive model to the data, but using medians instead of means. The procedure operates iteratively and starts as follows:

1. Estimate the row effects by calculating the median of each row ($\tilde{\alpha}_i = median[y_{i1}, ..., y_{iJ}]$);
2. Estimate the residuals by subtracting each row median from all observations in the corresponding row ($\tilde{\epsilon}_{ij} = y_{ij} - \tilde{y}_i, \forall j$);
3. Estimate the common value by taking the median of all row medians ($\tilde{\mu} = median[\tilde{y}_1, ..., \tilde{y}_I]$);
4. Subtract the common value from each row-median ($\tilde{\alpha}_i = \tilde{y}_i - \tilde{\mu}$);
5. Repeat all previous steps on columns of residuals $\tilde{\epsilon}_{ij}$, rather than rows, to estimate the column effects ($\tilde{\beta}_j = median[\tilde{\epsilon}_{1j}, ..., \tilde{\epsilon}_{Ij}]$.

The polishing is repeatedly applied to rows and columns of residuals alternatively until all row and column medians are zero or until no further improvement is obtained. Beginning the iteration with columns instead of rows will not necessarily yield the same fit; however, the differences are typically small [36].

After the main effects have been removed by a robust method, residuals ($\hat{\epsilon}_{ij}$) that originate from outlying cells are larger, because the median has a high breakdown point of 50%, and thus, these cells are easier to identify. For example, applying

the aforementioned median polish procedure to the data from table 3–1, the single outlier does not contribute to the estimation of the row, column and main effects. The residualized cell values remain at zero. As a result, the residuals correspond exactly to the original data where the first cell is clearly an outlier.

Median polish has several uses: to identify row and column structure; to check if there is evidence of an interaction, i.e., if the model is non-additive ; to detect global non additive patterns (e.g. increasing row effects in first column, decreasing effects in last column); to identify cell-specific abberations (e.g. data generally conform to additive model but residual in one cell is suspiciously high or low relative to the others). After fitting the standard additive model using median polish, a large residual appearing in one cell may come from an outlying cell, and thus, should be of particular attention [36].

Unfortunately, the median polish fit does not always coincide with the corresponding L^1 solution [36, 37], i.e., it does not always minimize the sum of absolute residuals

$$SAR = \sum_{i=1}^{I}\sum_{j=1}^{J}|\tilde{\epsilon}_{ij}| = \sum_{i=1}^{I}\sum_{j=1}^{J}|Y_{ij} - \tilde{\mu} - \tilde{\alpha}_i - \tilde{\beta}_j|.$$

Median polish can therefore only be thought of as an approximate least-absolute-deviations method of fitting.

Davies' Methods.

To overcome the potential lack of uniqueness of the L^1 solution, Terbeck and Davies [4] proposed methods based on M-estimators. The 'M' stands for 'maximum-likelihood-like'. These estimators were introduced by Huber [38] and Hampel *et al.* [39]. M-estimators can be thought of as a generalization of maximum likelihood

estimation in which the function to be maximized has been modified. Under the assumption of a Gaussian model for errors, maximizing the likelihood is equivalent to minimizing the SSR. Let ρ be a function of the residuals. The least-squares method minimizes SSR, where $\rho(x) = x^2$, which is unstable in the presence of outliers. In contrast, M-estimators minimize

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\rho(\hat{\epsilon}_{ij}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\rho(Y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)$$

where ρ is a symmetric, positive-definite function of the residuals. If $\rho(x) = |x|$, then the previous sum corresponds to the SAR and a minimum solution always exists, but in general may not be unique. However, if ρ is chosen to be a strictly convex function, then the solution is always unique.

M-estimators are invariant to the scale if they minimize

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\rho(\frac{\hat{\epsilon}_{ij}}{s}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\rho(\frac{Y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j}{s})$$

where s is a robust estimator of scale (e.g. $1.483 \times$ median absolute deviation (MAD)). Consequently, neither mean nor median polish is invariant to scale because the variance of the residuals is not taken into account in the L^2 and L^1 criteria. Note that M-estimators can be calculated iteratively using reweighted least-squares. Davies [5] has argued that Terbeck and Davies [4] external estimator of scale is complicated and unsatisfactory; consequently, he has developed simpler and computationally more stable methods also based on M-estimators. Davies' [5] two methods are based on the same strictly convex function ρ of the residuals. The first method,

'*Davies robust method*', iterates over the set of L^1 solutions, and uses the corresponding estimator of scale (see equation in Appendix). Davies [5] claims that "interactions and outliers can be more reliably identified by the residuals from appropriate re-descending M-estimators than from L^1 residuals" or those obtained by the previous robust method. Consequently, his main method, '*Davies reweighted method*', is based on re-descending M-estimators and overcomes the potential problem of multiple solutions corresponding to different local minima related to such estimators. Briefly, the second method minimizes a weighted function of the residuals obtained from the first method and uses an estimator of scale which is asymptotically Fisher consistent for normal errors (see Appendix for details).

3.2.2 Multiple Observations per Cell

All of the above methods refer to a single observation per cell. In a two-way layout with $K_{ij} > 1$ replicate measurements in cell ij, the observations y_{ijk} represent the replicated values for each combination of the $(I \times J)$ levels of the two factors. The standard additive model becomes

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}.$$

With $K_{ij} > 1$ replicates in some or all cells, it is possible to fit an additive model with interaction (γ_{ij}) terms

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}.$$

It is possible to distinguish between interaction (γ_{ij}) and unstructured noise (ϵ_{ijk}) only if one has replicated data (i.e. $K > 1$). Typically, the fitting of such a model

focuses on means and uses the L^2 criterion to determine the 'best' fit. But as previously demonstrated, the least-squares fit is not resistant to the presence of outliers and suffers from the 'leakage' problem.

To circumvent these problems, a limited number of robust methods have been developed for detecting interactions in two-way data with replicates. Among others, Brown and Mood[40] and Hettmansperger and Elmore [41] have introduced a median and a rank test, respectively. Although these robust tests allow for the detection of interaction patterns, they do not allow one to *identify specific outlying cells.*

At this point, it is important to distinguish between interaction and outlying cells. Interaction is a more general term. Interaction may be conceptualized as a *pattern* of residuals after row and column effects have been removed. Our interest is in detection of individual outlying cells irrespective of the residualized structure pattern. We are not concerned with the interaction pattern, but in the special case of outlying cells.

There is a lack of statistical research literature about the performance of robust methods to detect outlying cells when there are replicate observations in each cell. To fill this gap, we examine the performance of various combinations of preprocessing robust methods and inferential rules when applied to replicated two-way tables. Since Davies [5] methods are defined for a single observation per cell, the only way to handle replicates is to replace the K_{ij} observations in each cell by their median. Until now, attention has been restricted to the specific case of $K = 1$ and the performance of these methods for the general case of $K \geq 2$ has not yet been investigated.

3.2.3 Median Polish With Replicates

Since its introduction, applications of the median polish procedure have been primarily to two-way (or higher) layouts with $K = 1$ observation per cell. In contrast to newer methods, two ways to handle the case of $K > 1$ suggest themselves with Tukey's median polish.

Median Polish Applied to Table of Medians.

One way is to reduce the replicates in a cell to a single value by taking the median of all K observations, i.e. to calculate

$$\tilde{Y}_{ij} = median\{Y_{ij1}, ..., Y_{ijK_{ij}}\}.$$

The usual two-way median polish can then be applied to the resulting table $(I \times J)$ of cell medians ($\tilde{y}_{ij}$'s). For example, in a step where one polishes the rows, the median of the i^{th} row is obtained by calculating the median of the cell medians in the row

$$\tilde{Y}_i = median\{\tilde{Y}_{i1}, ..., \tilde{Y}_{iJ}\}.$$

At the end of the entire procedure, one residual ($\hat{\epsilon}_{ij}$) is obtained for each cell of the two-way table, whatever the number of replicated observations per cell.

Median Polish Applied to Individual Values

A second natural method is to use all individual replicated measurements in a row (column) when estimating row (column) medians during the polishing. To polish the rows, the median for the i^{th} row is obtained by calculating the median of all $K_{i\cdot} = \sum_{j=1}^{J} K_{ij}$ observations in this row

$$\tilde{Y}_i = median\{Y_{i11}, ..., Y_{i1K_{i1}}, Y_{i21}, ..., Y_{i2K_{i2}}, ..., Y_{iJ1}, ..., Y_{iJK_{iJ}}\}.$$

This way of polishing creates a total of $K.. = \sum_{i=1}^{I} \sum_{j=1}^{J} K_{ij}$ residuals ($\hat{\epsilon}_{ijk}$), i.e., *one for each individual original observation.*

Lack of Consensus

There are no guidelines in the statistical literature as to which of the two previous algorithms is preferable. Few software packages include procedures to perform median polish and each of the two major ones that do forces the end user to treat replicates differently. In S-PLUS, the 'twoway' function limits its input to a two-way table with a single measurement in each cell. Consequently, the only way to handle multiple measurements is to first calculate the median of each cell and than use the 'two-way' function to apply the median polish to the table of medians (as in *Median Polish Applied to Table of Medians* subsection). Minitab allows for multiple measurements per cell, but its documentation does not specify the algorithm used. By calculating a small example both manually and with Minitab, we determined that all individual replicated values are used to estimate row and column effects (as in *Median Polish Applied to Individual Values* subsection).

Nor is there any consensus among prominent researchers who have developed and refined robust data analysis methods. We asked four authorities which of the above two median polish methods should be used to handle replicated measurements. Their answers, shown in table 3–2, illustrate that there is no consensus.

3.3 Methods

Four preprocessing options are investigated:

1. Median Polish Applied to Individual Values

2. Median Polish Applied to Cell Medians
3. Davies Robust Method Applied to Cell Medians
4. Davies Reweighted Method Applied to Cell Medians

In addition, for the two first procedures, we examined the effect of starting the iterative median polish with columns rather than rows. For the two last procedures, we used the code provided by Davies [5]. All procedures were run in R 1.6.2 under Linux.

A simulation study was performed to compare the four preprocessing procedures with respect to their performance in detecting outlying cells in a two-way layout. The study included single and multiple outlying cells, increasing number of observations (both table size and number of replicated observations per cell), outlying values of various sizes, the absence or presence of row/column effects, the four options for obtaining residuals and four different rules for defining outliers. For simplicity, we considered only square tables with equal numbers of rows and columns ($I = J$), balanced designs with equal number of observations per cell ($K_{ij} = K, \forall i, j$), Gaussian errors, and no missing values.

3.3.1 Amount of Data

Two factors are considered:

1. Table Size: 5×5 or 10×10;
2. Number of Replicates: $K = 1$, 2, 3, 5 or 10;

For each of the 20 combinations of table size, number of replicates, outlier size, and column effect, we simulated 1000 data sets. Values in 'null' cells were drawn from a $N(0, 1)$ distribution. The replicated values in the single outlying cell or the

multiple outlying cells were drawn from a $N(\mu, 1)$ distribution where μ was either 1 or 2.

3.3.2 Patterns of Cells

We studied two situations. Table 3–3 illustrates each of the following patterns.

Single Outlying Cell

Values in one selected cell are generated as outlier values. In addition, for this specific pattern, a column effect was created by adding a constant (2) to each of the values in the column that contained the outlying cell (see Figure 3–1).

Multiple Outlying Cell

Terbeck and Davies [4] and Davies [5] have restricted their attention to patterns they call "unconditionally identifiable" that were also the main object of their study. Here we focus on two corollaries (2.7 and 2.8 in Terbeck and Davies [4]). Corollary 2.7 states that an interaction pattern in which fewer than 50% of the rows and fewer than 50% of the columns contain interactions, is unconditionally identifiable. On the other hand, by corollary 2.8, an interaction pattern in which fewer than 25% of the cells in each row and in each column are outlying, is also unconditionally identifiable. Thus, the case of a single outlying cell satisfies the conditions in both corollaries. Consequently, we decided to test three other patterns that do not satisfy either one or the two corollaries.

The second pattern satisfies the conditions in Corollary 2.8, but does not satisfy those in Corollary 2.7. Terbeck and Davies [4] mentioned that "Tukey's median polish can be shown to detect all interaction patterns described by Corollary 2.7, but it does not detect all those described by corollary 2.8". We investigate the most extreme

case where all rows and columns will each contains a maximum number (smaller than 25%) of outlying cells. Consequently, each row and each column contains one outlying cell in a 5×5 table, and two outlying cells in a 10×10 table. That is, 20% of the cells in each table (distributed evenly among all rows and columns) were outlying cells.

The third pattern corresponds to the opposite case, i.e., it satisfies the condition in Corollary 2.7, but does not satisfy those in Corollary 2.8. We used a neighboring group of outlying cells, while maintaining the number of outlying cells per row or column under the median breakdown point of 50%. Thus, four cells forming a 2×2 cluster in a corner of a 5×5 table, and 16 cells forming a 4×4 cluster in a corner of a 10×10 table, contain outlying values. Consequently, each table contains 16% of outlying cells located in the same corner.

The fourth pattern is a compromise between the two previous patterns of several outlying cells. However, it does not satisfy any of the two corollaries. We decided to maintain a similar percentage of outlying cells. Thus, for a 10×10 table, the majority (60%) of the rows and columns contain three outlying cells. The 18 outlying cells are distributed in two clusters of nine outlying cells located in two 3×3 tables in opposite corners of the 10×10 table.

3.3.3 Inferential Rules for Defining an Outlying Cell

After each simulated dataset had been preprocessed by each of the four options, interest is on cell residuals. For the median polish applied to individual values, since one residual is obtained for each original observation, a 'cell residual' was defined

as the mean of all residuals of the cell. For the three other methods, no additional aggregation was required.

We adopted the following four arbitrary statistical rules to define a cell as an outlier:

1. *SDs away from the Mean:* A cell in which the cell residual is more than x standard deviations away from the mean of all cell residuals;
2. *Jackknife SDs away from the Mean:* As in previous rule, but where the cell residual in the candidate cell is removed from the mean and standard deviation calculations;
3. *MADs away from the Median:* A cell in which the cell residual is more than x (rescaled) Median Absolute Deviations away from the median of all cell residuals;
4. *IQR away from Q1 or Q3:* Any observation that is more than x times the Inter Quartile Range away from the 1st (Q1) or the 3rd (Q3) quartile of all cell residuals.

3.3.4 Comparing Performance of Combinations of Preprocessing Options and Inferential Rules

We wished to apply each of the four inferential rules to each resulting table of cell residuals. Since power may be higher at the expense of higher type I error, we needed to compare sensitivity (power, i.e., 1-probability of a type II error) at a common specificity (i.e., 1-α, 1-probability of a Type I error). Otherwise, the sensitivity of one rule could artificially be higher than that of another because of its larger α level. Since we were unable to choose a priori the threshold 'x' which would result in a certain type I error, we employed an ROC analysis to control for different

α levels. We varied the threshold 'x' from 0.1 to 3.5 in steps of 0.1, for a total of 35 data points for each ROC curve.

To illustrate, consider residuals from a 5×5 table with one outlying cell, and one specific rule, e.g. 'x SDs away from the mean'. For the first value of x, we calculated the corresponding cutoff according to the specific rule. Using this cutoff, we calculated sensitivity and specificity. For each table, sensitivity is either 1/1 or 0/1 depending on whether the absolute value of the cell residual in the true outlying cell exceeds the cutoff. Similarly, specificity is the proportion of the remaining 24 cell residuals that are correctly classified as non-outliers. We recalculated sensitivity and specificity for each value of x. Then, we repeated the calculation for each of the 1000 datasets preprocessed by the same option and obtained an average sensitivity and an average specificity across the 1000 values for each value of x.

3.4 Results

First, we examined differences when the median polish algorithm starting the iteration with rows rather with columns, but results were the same no matter which was used first (data not shown), so we will consider only iterations starting with rows.

The results using the four different rules for defining an outlying cell were all equivalent (data not shown). Consequently, we will fix on the first rule 'standard deviations away from the mean' that is simple, intuitive and well known.

Figure 3–2 shows ROC curves comparing the four different options of obtaining residuals when trying to detect a single outlying cell. As expected, the number of

replicates (K) has a pronounced effect on performance. For $K = 1$ or $K = 2$ (panel a), there is a slight advantage for the Davies reweighted and the median polish applied to cell medians methods, especially for the 10×10 tables. However, for $K \geq 3$ (panels b, c, and d), it becomes clear that median polish applied to individual values performs best when trying to detect an outlying cell in two-way data with replicates.

When there are several outlying cells, results may differ. For the first pattern, where 20% of outlying cells are evenly distributed, the conclusions remain the same, but the differences are smaller (Figure 3–3). For the third and fourth patterns, when the outlying cells are grouped in one (Figure 3–4) or two (Figure 3–5) corners of the table, for outlying values of small size ($\mu = 1$), the conclusions still hold but the differences are again smaller. However, for outlying values of bigger size ($\mu = 2$), performance of the median polish applied to individual values decreases and becomes poorer in comparison to the other methods, as the number of replicates increases. We noticed that specificity remain the same for all methods whatever the number of replicates. Thus, preprocessing methods applied to cell medians performed best because, for fixed specificity, sensitivity increases when the number of replicates increases, in contrast with the median polish applied to individual values for which the sensitivity remains the same. Also, there is a slight advantage for median polish applied to cell medians in comparison to the two Davies methods also applied to cell medians.

Figures 3–6 and 3–7 show that increasing the number of observations increases the sensitivity for a fixed specificity for the median polish applied to individual values

method. Importantly, the gain is greater when increasing the number of replicates than the size of the table because of the stability of the estimate with more replicates. For example, with three replicates per cell, the gain in performance from a 5×5 table (i.e. 75 data points) to a 10×10 table (i.e. 300 data points) is small in comparison to having four time more data points. In contrast, for a 10×10 table, increasing from two replicates (i.e. 200 data points) to three replicates (i.e. 300 data points) represents a larger gain for only 1.5 times more data points. Also, power to be able to detect outlying cells is influenced by both the number of replicates and the size of the outlying values. Since the latter is divided by $\sqrt{1/K}$ in power calculations, large effect with few replicates end up with the same ratio as of small effect with more replicates.

Finally, in the case of a single outlying cell per table, all four preprocessing options handle column effects equally well; sensitivity and specificity are the same whenever a column effect is present or absent (data not shown).

3.5 Example

The example is taken from Hahn *et al.* [1]. The data measure fighting behavior of pairs of mice after maturation, and can be found in Scheirer *et al.* [42]. Aggression was measured by seconds of tail rattling per seconds of fighting. The data are represented in a 2×3 table with 7 replicates (pairs of mice) per cell. The first factor corresponds to different environmental conditions (0 and 1) and the second factor to brain weight (small, medium, and large). Hettmansperger and Elmore [41] have also analyzed this data set. They present boxplots of cell data with 85% confidence

intervals. The application of their test allows detection of the presence of interaction. However, from the raw data, it appears that the 'environment 0, small brain group' is an outlying cell (Table 3–4).

Table 3–5 shows residuals obtained when applying the four preprocessing options for obtaining residuals. As in the simulation study, residuals in each cell for the median polish applied to individual values have been aggregated by taking their mean in order to define cell residuals. Since the environment factor has only two levels, interpretation may be facilitated by examining the difference between the two environmental conditions [43]. Differences are presented in Table 3–6. In all cases, the largest residual appears correctly in the small brain group. However, as confirmed by the simulation study, the outlying cell can be identified with greater power when looking at the residuals obtained from the median polish applied to individual values.

3.6 Discussion

Since one cannot know in advance the numbers and the location of outlying cells, in a general manner, we recommend the use of the median polish applied to individual values. In most cases, this method offers higher performance when trying to detect either one or several outlying cells. The three other preprocessing methods applied to cell medians perform better when there are several grouped outlying cells and when the size of the outlying values is high in comparison to the size of the other values. However, such outlying cells are easier to detect than outlying cells of lower value. Thus, in these cases, median polish applied to individual values is not the best method, but still performs well.

As expected, the number of observations (i.e. table size and especially number of replicates), has a big effect on performance. Most of the time, with only $K = 3$ or 5, there is an substantial increase in power.

Results from our simulation study did not show any advantages for Davies [5] methods applied to cell medians. Median polish also applied to cell medians performed as well and sometimes even slightly better. In addition, median polish has the advantages of being easier to understand and faster to calculate. However, there might be some interaction patterns that are 'unconditionally identifiable' but that do not satisfy either of the two Corollaries [4]. Note that Corollary 2.7 was also the definition studied by Daniel [44]

For example in HTS, the last two patterns of cells may occur if compounds are not randomly located among the wells of the plates, and if analogue compounds are located in a same corner. Thus, if one of these compounds is active, the others have higher chances to be also active since they share similar chemical properties.

In the future, additional simulation studies could be done to assess the performance when the errors are not Gaussian. Also, using median polish with repeated measurements would also help in developing tests of significance of fitted row and column effects, since there is currently no theory on this issue.

3.7 Appendix

Davies [5] preprocessing methods are based on a class of M-estimators. To guarantee uniqueness, the following strictly convex function is used

$$\rho_\lambda(x) = \frac{x^2}{\lambda + |x|}$$

where $\lambda > 0$, and satisfies $sup_x|\rho_\lambda(x) - |x|| = \lambda$.

Also, the two methods are invariant to the scale, since they both minimize

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\rho_\lambda(\frac{\epsilon_{ij}}{s}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\rho_\lambda(\frac{y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j}{s})$$

where s is an estimator of scale as defined below.

3.7.1 Davies Robust Method

A unique L^1 solution $(\hat{\mu}^0, \hat{\alpha}^0, \hat{\beta}^0)$ is obtained by minimizing, over the set of L^1 solutions,

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\rho_\lambda(\frac{y_{ij} - \hat{\mu}^0 - \hat{\alpha}_i^0 - \hat{\beta}_j^0}{s^0})$$

for some specified value λ_0 of the parameter λ, and where the corresponding estimator of scale is

$$s^0 = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}|y_{ij} - \hat{\mu}^0 - \hat{\alpha}_i^0 - \hat{\beta}_j^0|$$

Davies [5] uses $\lambda_0 = 0.1$ as default value in the calculations arguing that it is a reasonable choice according to simulations.

3.7.2 Davies Reweighted Method

An external estimator of scale, which is asymptotically Fisher consistent for normal errors, is given by

$$s = u_{(\frac{IJ-K(IJ)}{z(\alpha)})}$$

where $u_{(1)} \leq u_{(2)} \leq ... \leq u_{(IJ)}$ are the ordered absolute residuals ($u_{ij} = |r_{ij}| = |y_{ij} - \hat{\mu}^0 - \hat{\alpha}^0 - \hat{\beta}^0|$) from the previous solution; $K(IJ) = min\{(J - [\frac{J-1}{2}])[\frac{I-2}{2}], (I - [\frac{I-1}{2}])[\frac{J-2}{2}]\} + [\frac{I-1}{2}][\frac{J-1}{2}]$ is the maximal number of interactions in an unconditionally identifiable interaction pattern; $\alpha = \frac{IJ-I-J+1-K(I,J)/2}{IJ-I-J+1}$, and $z(\alpha)$ denotes the α quantile of the standard normal distribution.

Using this resulting estimator of scale s and residuals r_{ij} from the previous robust method, the procedure consists in minimizing

$$\sum_{i=1}^{I}\sum_{j=1}^{J} w(\frac{r_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j}{s})$$

where $w(x) = 1 - (1 - (\frac{x}{c})^2)^3$ is a weight function, and c is a tuning constant set to $c = 3.5$ (default value in code).

Table 3–1: Hypothetical data to illustrate 'leakage'

Two-way Data			Least-Squared Residuals			Median Polish Residuals		
9	0	0	4	-2	-2	9	0	0
0	0	0	-2	1	1	0	0	0
0	0	0	-2	1	1	0	0	0

Table 3–2: Responses to questions on how to handle replicates in median polish

Authority	Preference		Comments
	Median Polish Applied to		
	Individual Values	Cell Medians	
1		√	"I believe you should first aggregate the replicates in each cell."
2	√		"My intuition is that you can take the median of all numbers in a row or column."
3	√	√	"Approach used for 3-way tables would appear to suggest using medians over all the replicates. But working with cell medians would certainly give more resistance to outliers locally."
4	√	√	"One natural question is whether the number of observations is the same in all the cells of the two-way table."

3.8 Tables and Figures

Table 3–3: Patterns of outlying cells for 10 × 10 tables as used in the simulation study

1^{st} Pattern: 1 Outlying Cell										2^{nd} Pattern: 20% Outlying Cells									
μ (+2)	0	0	0	0	0	0	0	0	0	μ	μ	0	0	0	0	0	0	0	0
0 (+2)	0	0	0	0	0	0	0	0	0	μ	μ	0	0	0	0	0	0	0	0
0 (+2)	0	0	0	0	0	0	0	0	0	0	0	μ	μ	0	0	0	0	0	0
0 (+2)	0	0	0	0	0	0	0	0	0	0	0	μ	μ	0	0	0	0	0	0
0 (+2)	0	0	0	0	0	0	0	0	0	0	0	0	0	μ	μ	0	0	0	0
0 (+2)	0	0	0	0	0	0	0	0	0	0	0	0	0	μ	μ	0	0	0	0
0 (+2)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	μ	μ	0	0
0 (+2)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	μ	μ	0	0
0 (+2)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	μ	μ
0 (+2)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	μ	μ
3^{rd} Pattern: 16% Outlying Cells										4^{th} Pattern: 18% Outlying Cells									
μ	μ	μ	μ	0	0	0	0	0	0	μ	μ	μ	0	0	0	0	0	0	0
μ	μ	μ	μ	0	0	0	0	0	0	μ	μ	μ	0	0	0	0	0	0	0
μ	μ	μ	μ	0	0	0	0	0	0	μ	μ	μ	0	0	0	0	0	0	0
μ	μ	μ	μ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	μ	μ	μ
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	μ	μ	μ
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	μ	μ	μ

Table 3–4: Cell medians of the original data from Hahn *et al.* [1] (see text for details)

	Brain Weight		
	Small	Medium	Large
Environment 0	3.50	0.98	0.37
Environment 1	1.35	1.27	0.82

Table 3–5: Cell residuals for data from Table 3–4

	Brain Weight			Environment
	Small	Medium	Large	
Median Polish Applied to Individual Values	2.50	-0.41	-0.25	0
	-0.54	0.44	1.01	1
Median Polish Applied to Table of Cell Medians	1.22	0.00	-0.08	0
	-1.22	0.00	0.08	1
L^1 Solution Applied to Table of Cell Medians	1.22	0.00	-0.16	0
	-1.22	0.00	0.00	1
M Functional Applied to Table of Cell Medians	0.84	-0.38	-0.46	0
	-0.84	0.38	0.46	1

Table 3–6: Cell residual differences between the two environmental conditions (environment 0 advantage) for data from Table 3–4

	Brain Weight		
	Small	Medium	Large
Median Polish Applied to Individual Values	3.04	-0.85	-1.26
Median Polish Applied to Table of Cell Medians	2.44	0.00	-0.16
L^1 Solution Applied to Table of Cell Medians	2.44	0.00	-0.16
M Functional Applied to Table of Cell Medians	1.68	-0.76	-0.92

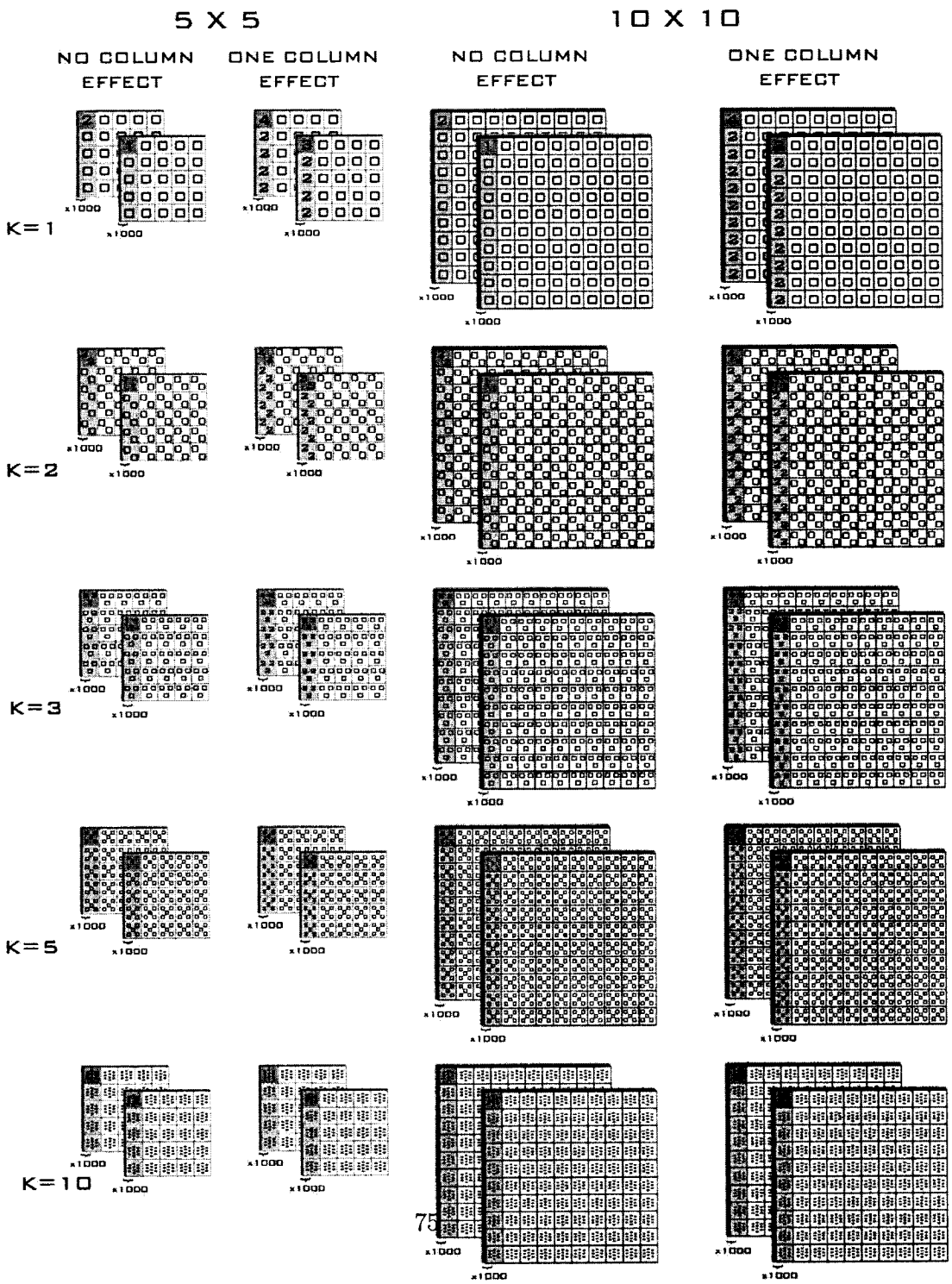

Figure 3–1: Overview of simulation study design

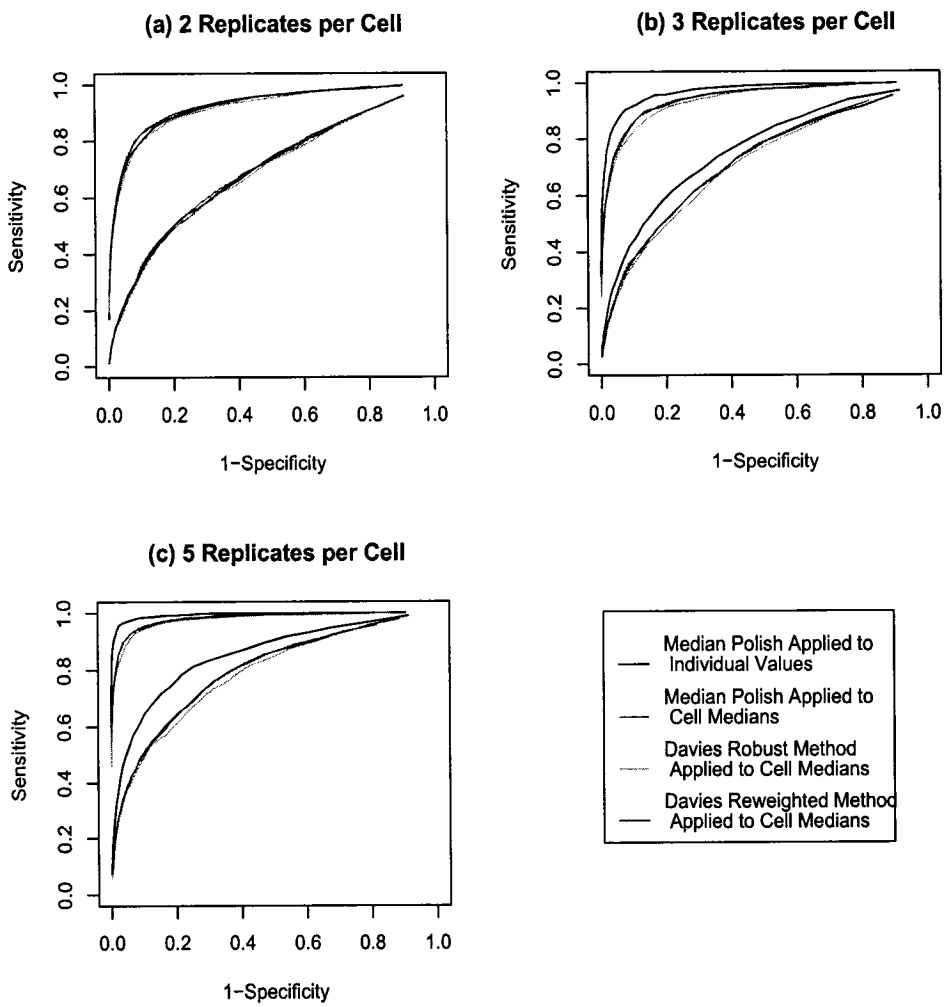

Figure 3–2: ROC curves to compare the performance of four options of obtaining residuals. Each ROC curve represents the performance of one of the 4 preprocessing options when trying to detect a single outlying cell (Table 3–3, pattern 1). Lower curves correspond to standard normal dataset with an outlying cell with a low 'signal' ($\mu = 1$, plain lines) while upper curves correspond to standard normal dataset with an outlying cell with a higher signal ($\mu = 2$, dashed lines). In each case, an effect (value of 2) was added to the observations in the column containing the outlying cell. The 'standard deviations away from the mean' rule is used to define a cell residual as an outlying cell. Panels (a), (b), and (c) are for 10×10 tables with respectively 2, 3, and 5 replicates per cell.

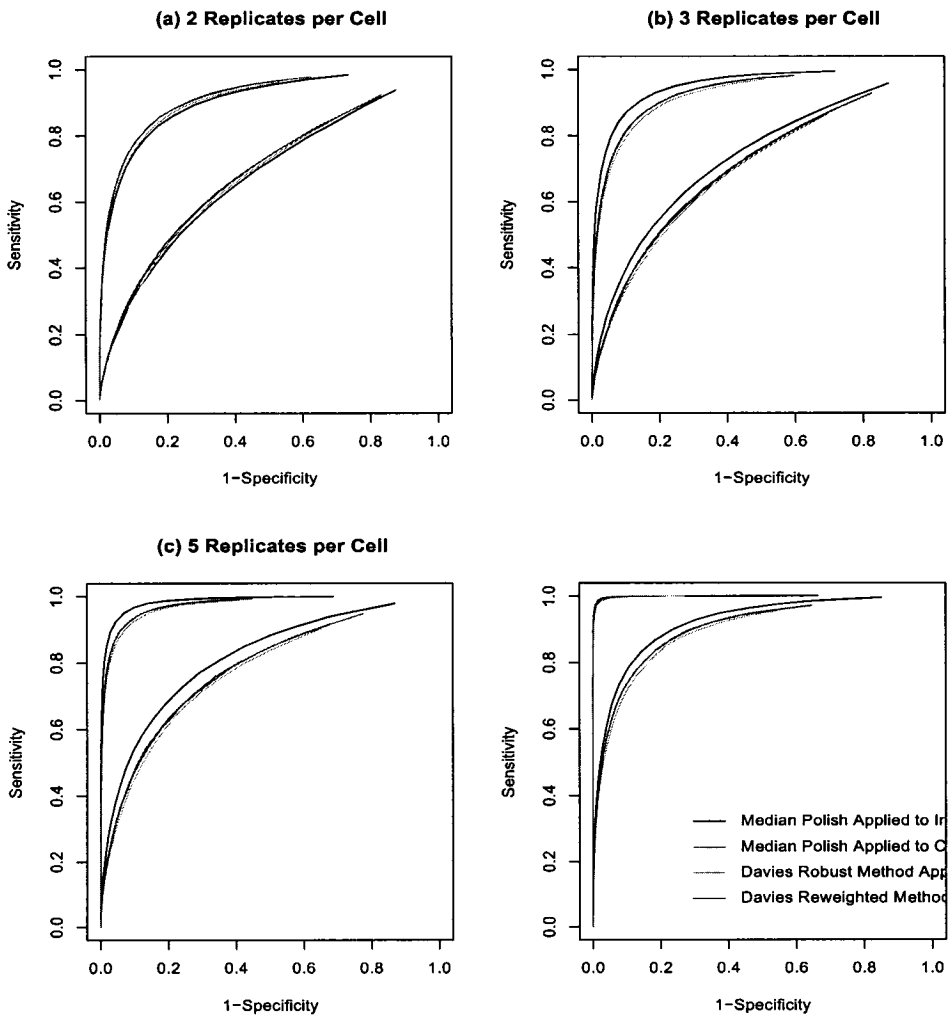

Figure 3–3: ROC curves to compare the performance of four options of obtaining residuals. Each ROC curve represents the performance of one of the 4 preprocessing options when trying to detect several (20%) outlying cells evenly distributed in each row and column (Table 3–3, pattern 2). Lower curves correspond to standard normal dataset with outlying cells of low signal ($\mu = 1$, plain lines) while upper curves correspond to standard normal dataset with outlying cells of higher signal ($\mu = 2$, dashed lines). The 'standard deviations away from the mean' rule is used to define a cell residual as an outlying cell. Panels (a), (b), and (c) are for 10×10 tables with respectively 2, 3, and 5 replicates per cell.

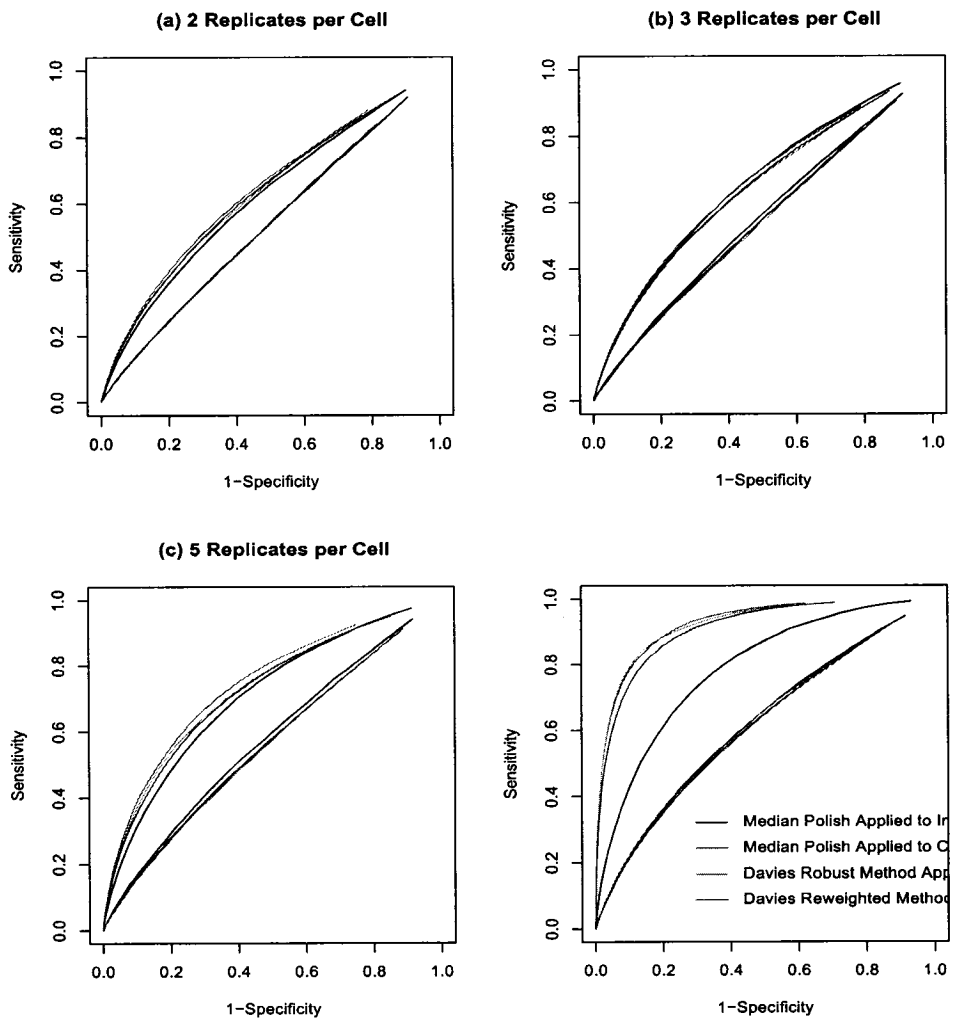

Figure 3–4: ROC curves to compare the performance of four options of obtaining residuals. Each ROC curve represents the performance of one of the 4 preprocessing options when trying to detect several (16%) outlying cells all located in a same corner of the table (Table 3–3, pattern 3). Lower curves correspond to standard normal dataset with outlying cells of low signal ($\mu = 1$, plain lines) while upper curves correspond to standard normal dataset with outlying cells of higher signal ($\mu = 2$, dashed lines). The 'standard deviations away from the mean' rule is used to define a cell residual as an outlying cell. Panels (a), (b), and (c) are for 10×10 tables with respectively 2, 3, and 5 replicates per cell.

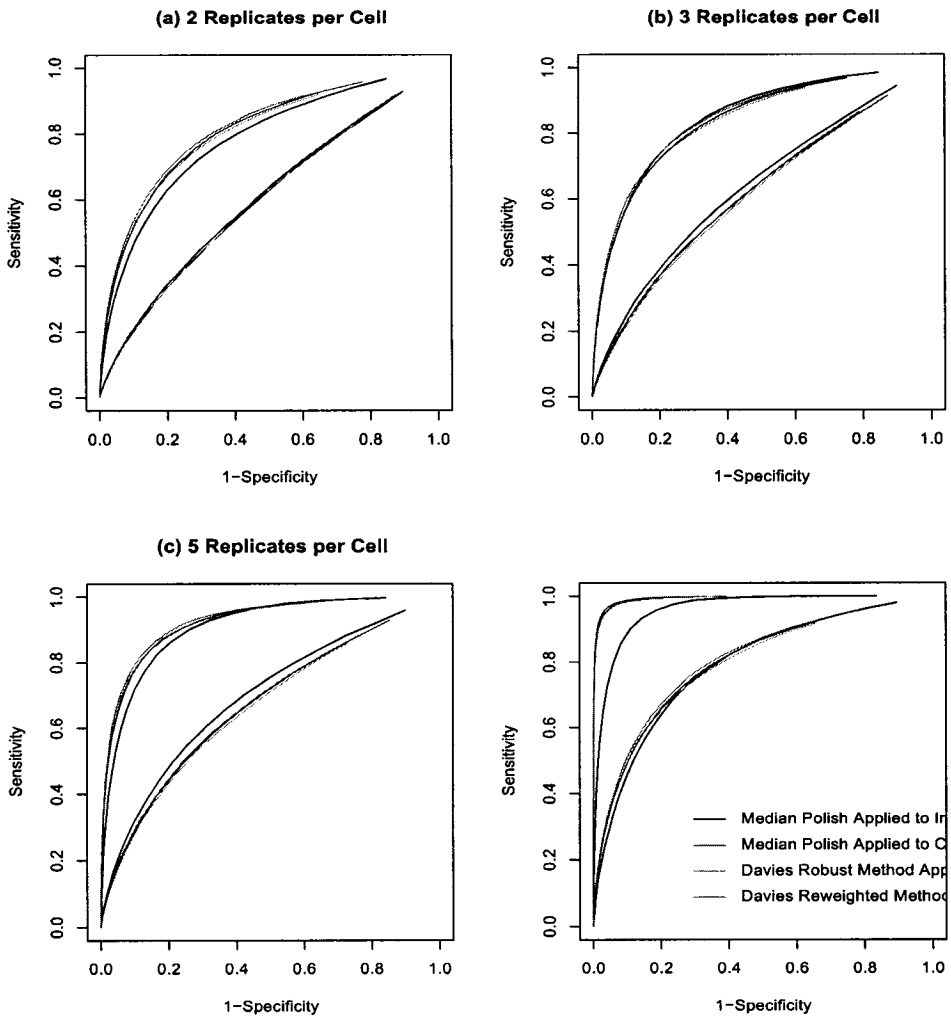

Figure 3–5: ROC curves to compare the performance of four options of obtaining residuals. Each ROC curve represents the performance of one of the 4 preprocessing options when trying to detect several (18%) outlying cells located in two opposite corners of the table (Table 3–3, pattern 4). Lower curves correspond to standard normal dataset with outlying cells of low signal ($\mu = 1$, plain lines) while upper curves correspond to standard normal dataset with outlying cells of higher signal ($\mu = 2$, dashed lines). The 'standard deviations away from the mean' rule is used to define a cell residual as an outlying cell. Panels (a), (b), and (c) are for 10×10 tables with respectively 2, 3, and 5 replicates per cell.

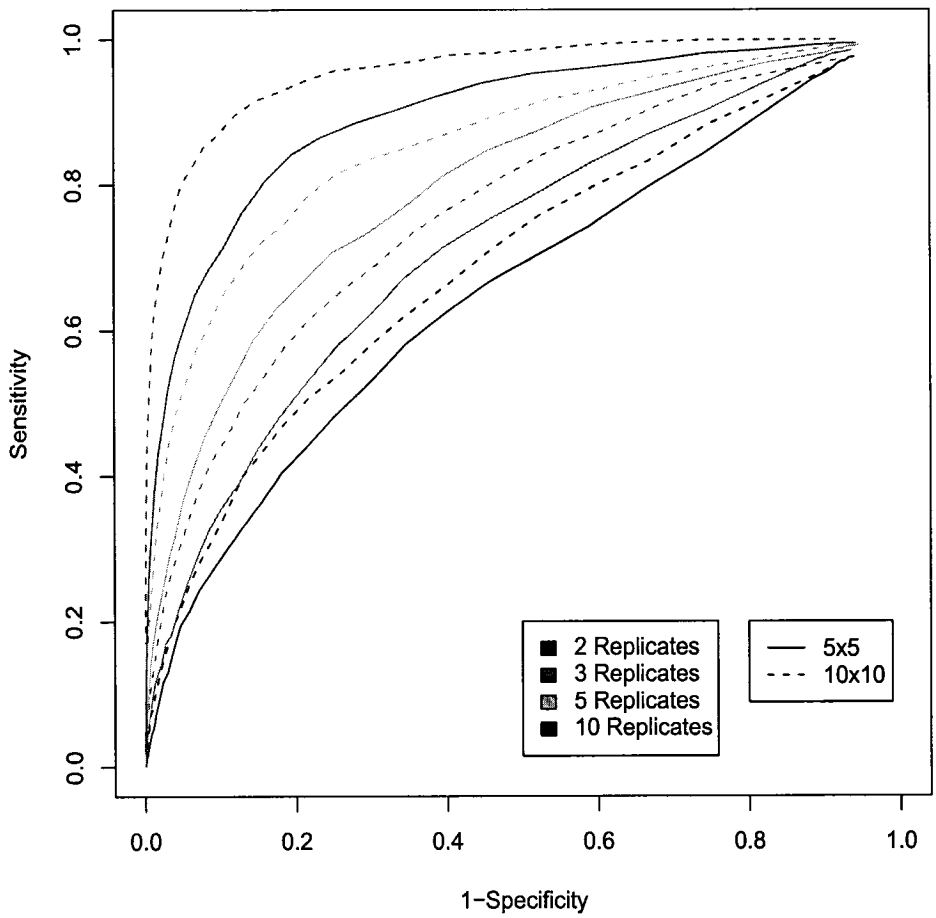

Figure 3–6: ROC curves to illustrate the effect of increasing number of observations. All ROC Curves represent results obtained when median polish is applied to individual values and when the 'standard deviation away from the mean' rule is used to define an outlying cell. Data are standard normal with an outlying cell of low size ($\mu = 1$) and a column effect in the corresponding row ($\mu = 1$). The number of replicates varies for both 5×5 table (plain lines) and 10×10 tables (dashed lines).

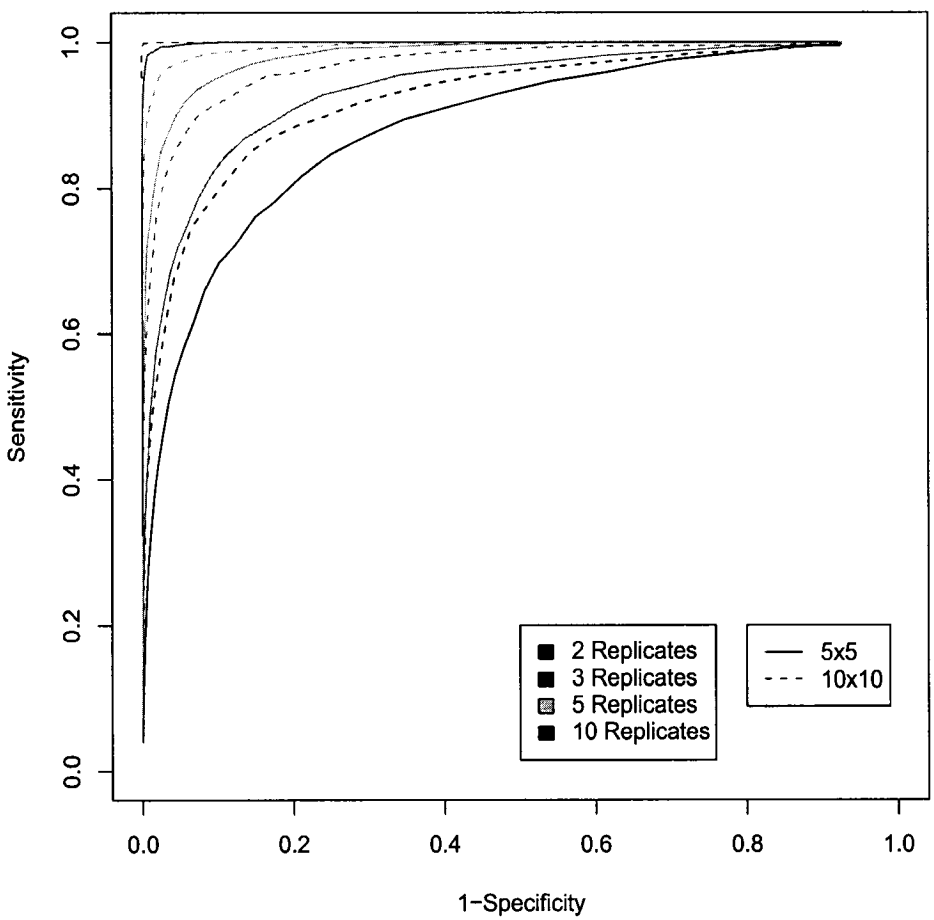

Figure 3–7: ROC curves to illustrate the effect of increasing number of observations. All ROC Curves represent results obtained when median polish is applied to individual values and when the 'standard deviation away from the mean' rule is used to define an outlying cell. Data are standard normal with an outlying cell of higher size ($\mu = 2$) and a column effect in the corresponding row ($\mu = 1$). The number of replicates varies for both 5×5 table (plain lines) and 10×10 tables (dashed lines).

Preamble to Manuscript III

The last chapter of my thesis reveals crucial applications of my research to HTS data. In the two previous manuscripts, I was mainly theoretical. In the first chapter, I reviewed both preprocessing and inferential methods, and made statistical suggestions to improve HTS data analysis. In the second, I performed a simulation study to compare the power of different statistical approaches to detect outlying cells in a replicated two-way dataset. Here, in the third chapter, the focus is on real-life applications.

The numerous HTS datasets generated are always examined with the goal of finding hits, and variation in primary screen has never been investigated systematically because of cost issues. Also, I don't want users to become overly optimistic and to expect statistical tools to do as well in real applications as they do in simulations. Looking at simulated data is not sufficient since we don't know where the hits are in real data and we cannot expect perfection.

In this manuscript, based on empirical datasets and data from real screens, I give a statistical view on the presence of unwanted variation; I provide designed procedures to optimally generate replicated HTS data; and I recommend steps, methods and guidance for statistical analysis of these data. Briefly, I demonstrate the benefits of (i) reducing unwanted variation, (ii) obtaining replicates, and (iii) using robust

efficient statistics to improve sensitivity and specificity of screens, and thus, hit detection.

This manuscript will be submitted to Nature Biotechnology. The references are included in the global thesis bibliography.

CHAPTER 4
Manuscript III - 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]

[1] McGill University and Genome Quebec Innovation Centre, 740 avenue du Docteur Penfield, Montreal, Quebec, Canada, H3A 1A4

[2] McGill University Department of Epidemiology, Biostatistics, and Occupational Health, 1020 Pine Avenue West, Montreal, Quebec, Canada, H3A 1A4

[3] McGill University Department of Biochemistry, 3655 Promenade Sir William Osler, Montreal, Quebec, Canada, H3A 1A4

[4] McGill University Department of Human Genetics, 1205 avenue du Docteur Penfield N5/13, Montreal, Quebec, Canada, H3A 1B1

Abstract

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. We make several procedural and statistical recommendations to increase sensitivity and specificity of screens. First, randomization of plate processing order at every step improves accuracy in activity measurements by minimizing unwanted variation stemming from human, biological and equipment errors. Second, the use of robust data preprocessing methods, such as the B-score normalization method, can further reduce unwanted variation by removing row, column and plate biases, which would otherwise potentially increase both false positives and false negatives. Third, replicate measurements allow estimation of the magnitude of the remaining random error and the use of formal statistical models, such as by an Empirical Bayes t-test to benchmark putative hits relative to what is expected by chance. Thus, all these approaches together increase confidence in hit identification.

4.1 Introduction

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 two broad purposes: it facilitates data interpretation by reducing the possibility that observed effects have been caused by confounding factors and it minimizes unwanted variation in activity measurements stemming from human, biological and equipment errors. Statistical methods at the data preprocessing (normalization) phase can further reduce unwanted variation which would otherwise potentially increase both false positives and false negatives. 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 this screen can also be used to estimate anticipated false negative rates for similar future studies.

Although the advantages of statistical procedures for HTS analysis were described a decade ago [45], statistical articles are only now becoming more common as researchers search for ways to improve the sensitivity and specificity of their screens. Various methods have been proposed to characterize the quality of screens [16, 46], to remove bias within and across plates [2, 47] and to obtain random error estimates for use in statistical tests to identify hits [18, 34].

We presented a data analysis strategy in a recent review of preprocessing and inferential methods for HTS [34]. For preprocessing, we argued in favor of non-control

based normalization methods and specifically recommended the B-score procedure [2]. We argued further that replicate measurements are needed to minimize variability, to verify method assumptions, and to suggest alternative data analysis strategies when assumptions are not met. Specifically, we demonstrated that aggregating the random error estimate for an individual compound with an estimate obtained across all compounds can provide a more precise estimate of random error. The described Empirical Bayes approach provides an effective framework for verifying model assumptions, estimating false positive rates, and reducing false negative rates.

Here we extend those arguments and demonstrate the utility of the approaches with a replicated primary and secondary screen and with two control experiments. We illustrate the advantages of randomization in the screening setup to minimize unwanted variation. We show that the B-score method provides the desirable statistical characteristics of bias correction and measurement independence. Finally, we show that B-scores when combined with Empirical Bayes t-test approach provide variance and p-value distributions which agree with theoretical expectations. The combination of randomization, replication B-score normalization, and the Empirical Bayes t-test should improve both specificity and sensitivity for HTS applications.

4.2 Results

4.2.1 Examination of raw data.

Figure 4–1 shows histograms and line plots of column effects of raw data for the Imunofluorescent screens (see Methods section for description of these datasets). Under the usual assumptions of unbiased measurements and few hits, one should

expect that the majority of the measured values will be symmetrically distributed about a central null value. Panel a shows, however, that the distribution of the first replicate set in the non-randomized screen contains two modes. Moreover, Panel b shows evidence of column bias (the right-most columns within the plates have higher signals on average; see Supplementary Results online for plate-by-plate column effect plots). Panels c-d show that variability and the column effects in the randomized screen have been reduced.

4.2.2 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 4–2 show that relative to raw data, Z-scores were more symmetrically distributed and reduced the column effects. Figure 4–3 shows that B-scores, however, provided the best adjustment for distributional asymmetry and column effects. Similar results were obtained for the less- pronounced row effects (data not shown). Moreover, overlap among the 100 largest/smallest values (averaged across replicates) between the two screens was higher for B-scores (14%/65%) than for either raw data (8%/43%) or Z-scores (7%/54%). Thus, although it is always best to avoid unwanted variation with procedural solutions, these results suggest that the B-score method provides a degree of reproducibility even in the presence of substantial procedurally-induced bias.

The advantages of B-scores are illustrated further in Figure 4–4 by analysis of additional data from an in-vitro translation assay experiment in which the same compound was tested in every well in the same concentration across all plates (see

Methods for description). 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 4–4a, however, shows that the raw data are positively correlated, indicating the presence of procedurally-induced location-specific biases. Figure 4–4b shows that the B-scores greatly minimized the bias, producing the expected null correlation (Scatterplots between plates for Z-scores generate results identical to the raw data because on a plate-by-plate basis they are simply rescaled raw scores and as such generate identical scatterplots). Similarly, the autocorrelation plots across all six replicate plates in Figure 4–4c show substantial correlations between putatively independent measurements for the raw data. The correlation at lag 1 indicates that wells in immediate proximity to each other down columns and up the next column (column 2: well at row 1 with well at row 2, well at row 2 with well at row 3, well at row 8 with well at row 1/column 3 well at row 7/column 10 with well at row 8/column 10) are highly correlated (r = 0.55). Successive lags indicate correlations between each well and the nth succeeding well (lag n). A pattern was observed which repeated at every 8^{th} lag. The closer wells were to each other within columns, the higher the correlation (e.g. the lag 1 correlation is higher than that for lag 2). A similar pattern was observed across columns (e.g. the highest correlation of r = 0.68 was observed for lag 8, which corresponds to immediately adjacent wells across columns). Although Z-scores provided some degree of correction (Figure 4–4b), B-scores again provided

the best correction (Figure 4–4c), reducing the autocorrelations at the various lags to near zero values.

4.2.3 Hit detection.

A major advantage of having replicates is the use of formal statistical models to benchmark presumed hits relative to what is expected by chance under the statistical model being used for data analysis. Figure 4–5 illustrates our investigation of the assumptions of our model, a required step before use of any statistical test (see Methods section for a detailed description of the tests).

The Empirical Bayes t-test produced the theoretically expected inverse gamma distribution for the non-randomized and randomized screens for replicate variances (Figures 4–5 a, d) and a uniform distribution for null p-values (Figures 4–5 c, f), increasing confidence in the validity of the results. The standard one-sample t-test generated fewer hits (small p-values) and a non-uniform null 2-tailed p-value distribution (Figures 4–5 b, e), indicating that the test is inappropriate for the data. In this context, the standard t-test suffers from a lack of degrees of freedom due to the small number of replicates and it may be more vulnerable to any non-normality within each location. Rank ordering of the two t-statistics is the same, but the quantiles are different because the activity measurements are divided by different estimates of the standard error. Finally, as we found previously with other data sets,(Malo et al., 2006), results from the one sample z-test were also not valid for these data. The common variance assumption was grossly violated, suggesting that the larger number of observed hits likely reflect an unduly high false positive rate (data no shown).

4.2.4 Other Considerations.

Statistical hypotheses may be investigated as 1 or 2-tailed tests. The former are used when the direction of the effect is predicted; the latter are used when effects of interest may be in either direction. In the two imunofluorescent screens, statistical hits were expected in both directions and accordingly we examined 2-tailed p-value distributions as a check of assumptions. For the biological purposes of the studies, however, the interest lies in the activity measurements which correspond to high positive B-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. Decrease of fluorescent signal may arise from a number of different causes. A compound may be toxic and cause 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.

Outliers among replicates threaten the validity of results obtained from statistical tests based on means (such as the ones 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 Empirical Bayes model used here. The idea is that compounds with replicate fluorescent

outliers should have unusually large variances. The F-distribution (Figures 4–5 b and f) can be used as the reference probability model. At a fixed alpha level, any 'rescaled' variance (i.e. the observed variance multiplied by a and b, the estimated parameters of the inverse-gamma distribution) that is greater than the quantile of a F-distribution with K-1 and 2a degrees of freedom is deemed an outlier. For the 'randomized screen', at alpha=0.001 we found no outliers, and at alpha=0.01 we found 7 outliers. Since 1120 compounds were tested, these numbers are smaller than the expected numbers, and consequently there are no obvious variance outliers (and hence no obvious fluorescent outliers) in the randomized screen.

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 (Figure 4–5h). Notwithstanding, we were unable to identify individual hits using the false-discovery rate procedure, which provides sensitive p-value adjustment in multiple testing contexts (FDR [48]) procedure, despite allowing a relatively high FDR of 0.25 (see Methods section). This apparent contradiction can be explained as follows. The lowest 1-tailed p-value was 0.003, a not unusually small p-value under the null hypothesis, given that there were 1120 compounds in the screen. That is, although there were many more small p-values than expected, none were so small as to merit individual attention. This in turn suggests that any true hits are likely to have small effect sizes (i.e. low intensities and/or high variability). This does not present insurmountable problems in

the current context because unlike for other high-throughput technologies (e.g. microarrays), secondary screens can be performed at medium throughput at relatively low cost. Accordingly, the net can be cast widely (likely generating large numbers of false positives) so as to minimize the number of false negatives.

4.2.5 Empirical demonstration of statistical power.

We performed a 'dilution series experiment' (see Methods section) in which various concentrations of a true active compound were randomly assigned well positions on a 96-well plate. Figures 4–6 presents ROC curves which compare the performance of three statistical tests based on random samples generated from the data. The Empirical Bayes t-test performed best, generating the fewest false negatives at fixed false positive levels. Figure 4–6 also shows that false negatives are reduced by increasing the number of replicates, especially for low concentration hits.

4.3 Discussion

We make several procedural and statistical recommendations to improve HTS hit detection.

For unavoidable sources of variation, randomization and blocking of processing steps provide the means to make valid assessments of compounds' activity levels by minimizing the effects of potential confounds such as processing order.

Exploratory graphics [49] of raw and preprocessed data allow assessment of measurement adequacy before performing further statistical analysis. Looking at the data distribution provide the means to check for gross errors in the measurements. Plots of plate and row/column medians can highlight a frequent source of

bias which can be minimized by robust preprocessing methods such as the B-score [2]. Autocorrelation plots can provide checks for measurement independence.

Finally, we show how replicates increase sensitivity of screens. With replicates, the significance threshold for hit identification can be based on p-values offering the advantage of understanding the probability of what should be expected by chance. Assumptions must be verified to ensure that one uses the appropriate test. Triplicate measurements offer several advantages over duplicates. With triplicates, undesirable outlier measurements (e.g.. an extreme value due to a procedural error) can be deleted or corrected before further statistical analysis. Triplicates also produce a non-trivial increase in power. For the t-statistic, one additional replicate provides the largest gains when sample sizes are small. For example, the critical t value threshold for identifying a hit with a one-sample t-test with two replicates is 12.7 whereas the threshold for three replicates is reduced to 4.3. Lesser gains are observed for four and five replicates (thresholds of 2.57 and 2.28) Additional degrees of freedom can be achieved with the Empirical Bayes t-test [34, 20, 24] which acts as a proxy for adding replicates.

Ultimately, biological validation will provide definitive evidence on the merits of various analytical approaches. How best to validate findings from high-throughput technologies is an unresolved philosophical question [50]. For example, a compound may be statistically deemed validated if it is significant in both tests or if the two p-values are not significantly different. It is left to the field to operationally define

validation and to decide on the methods that should be used for statistical confirmation of validation.

4.4 Methods

4.4.1 Data Sources

Immunofluorescent screen (non-randomized).

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 $\Delta F508$. Fourteen 96-well plates were run in duplicate. Including incubation time, the screen was run in four days. Plates were processed in sets of five, 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.

Immunofluorescent screen (randomized).

This screen was the same as the previous non-randomized screen except for two aspects: processing order was randomized for all steps in the protocol and replicates were obtained in three independent runs (i.e. blocks).

Measurement experiment.

An inactive compound from cystic fibrosis immunofluorescent assay screen described above was tested in all of the 80 middle wells of six 96-well plates. Plate processing order was randomized for all steps.

Dilution series in-vitro translation assay.

A known protein inhibitor was arrayed within each of six 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 (water) were randomly located in the 64 middle wells of 96-well plates. Positive controls (Anisomycin at 50 μM) and negative controls (water) were placed in alternating wells on the 1^{st}, 2^{nd}, 11^{th} and 12^{th} 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 which might be expected from a valid screen. Removing potential row and column biases with the B-score normalization method was deemed inappropriate for these data because differences among the rows and columns reflected biological differences as well as any potential biases due to the large number of hits of differing effect sizes. Accordingly, the data were normalized as follows:

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

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 the plate.

For each of 100 simulation runs, we randomly sampled (with replacement) 1120 normalized measurements from the empirical dataset (14 plates x 80 values per plate).

Some 1064 'non-hits' were sampled from the 240 negative control measurements (6 plates x 40 values per plate). Four consecutive concentrations were chosen. For each concentration, 14 hits were sampled from the 144 concentration-specific measurements (6 plates x 24 values per plate) yielding a rate of true hits of 5% within each simulation run. We repeat this simulation for three different sets of concentrations, i.e. the four highest, the four lowest, and the four in the medium. Hits were identified according to various statistical criteria and false positive/false negative rates were calculated (see Inferential Statistics section below).

4.4.2 Preprocessing statistics.

We compared two non-control-based normalization methods. Let $i=1,,I$ rows; $j=1,..,J$ columns; and $p=1,,P$ plates.

$$Zscore_{ijp} = \frac{x_{ijp} - \bar{x}_p}{s_p}$$

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 the plate.

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

where r_{ijp} are the residuals obtained by a two-way median polish [3] and MAD_p is the median absolute deviation of all residuals within the plate. Since we did not observe consistency in positional effects from plate-to-plate and since we randomized the plate processing order, we did not used the smooth function in our calculations of B-scores.

Both the Z-score and B-score methods rescale measurements so that they are comparable across plates; in addition, the B-score corrects for row and column effects and is resistant to outliers [2]. Because the same compound was tested in all wells, within-plate variation reflected errors in measurement only (random error and potentially bias).

4.4.3 Inferential statistics.

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

$$t = \frac{\bar{x}_K - constant}{s_K \sqrt{1/K}}$$

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 referred to 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).

To overcome this problem, a z-test was calculated for each compound using s, the square root of the average of all the compound-specific variances:

$$z = \frac{\bar{x}_K - constant}{s \sqrt{1/K}}$$

The ratio is then referred to a standard normal distribution. The z-test makes the strong assumption that the true variance is the same for all compound measurements, an assumption often not verified.

The Empirical Bayes t-test provides a compromise between the low sensitivity of the local t-test and the strong common error assumption of the z-test. Compound-specific variances are assumed to follow an inverse-gamma distribution with parameters a and b [20, 24]:

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

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. $\tilde{t}$ follows a t-distribution with *K-1+2a* degrees of freedom. Variance ($\tilde{s}^2$) is estimated by a weighted average of the compound-specific variances 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 [34]. This leads to an increase of *2a* degrees of freedom over the standard t-test.

4.4.4 False discovery rate (FDR) control.

Benjamini *et al.* [48] have proposed a method to control for the expected proportion of false positives among the positives which they called the false discovery rate (FDR).

Once a nominal p-value *P(i)* is obtained, corresponding to each compound *i=1,,m*, the compound is deemed a hit if:

$$P_{(i)} \leq 1 - \frac{(i-1)}{m}$$

This method weakly controls the familywise error rate (FWER) and is more powerful than other FWER controlling methods.

4.5 Figures

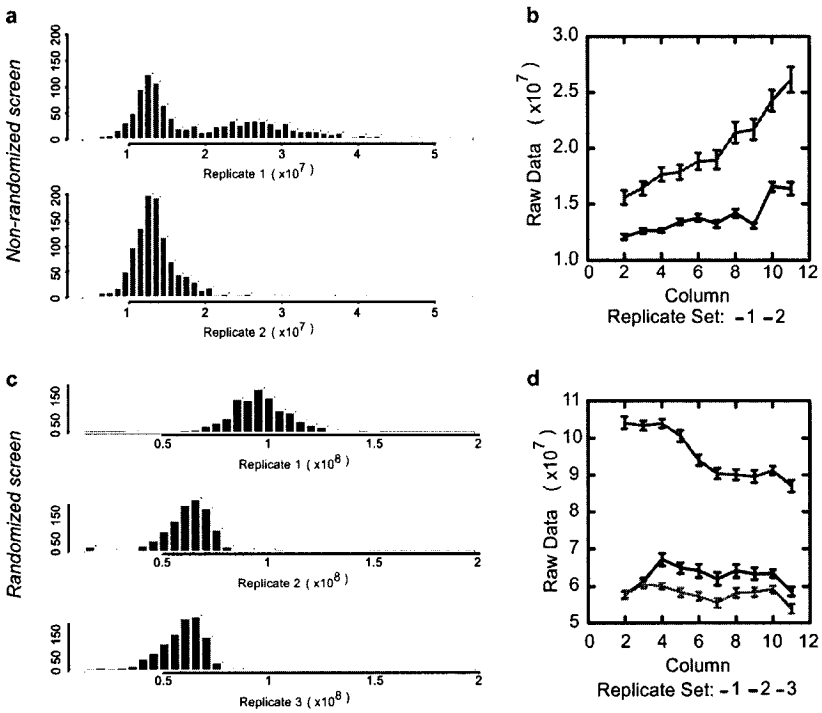

Figure 4–1: Graphical display of raw data for each replicated set of imunofluorescent screens as exploratory analysis. (a) Histograms of raw data for the non-randomized screen show large variability, especially in the first replicated set. The first distribution contains two modes and a very long tail on the right, i.e., more large values than the usual expected proportion of hits. The second distribution is closer to expectation with one 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 again show different patterns. However, distributions all three distributions are unimodal. (d) Although column effects remained, they were reduced by randomization of processing plate order.

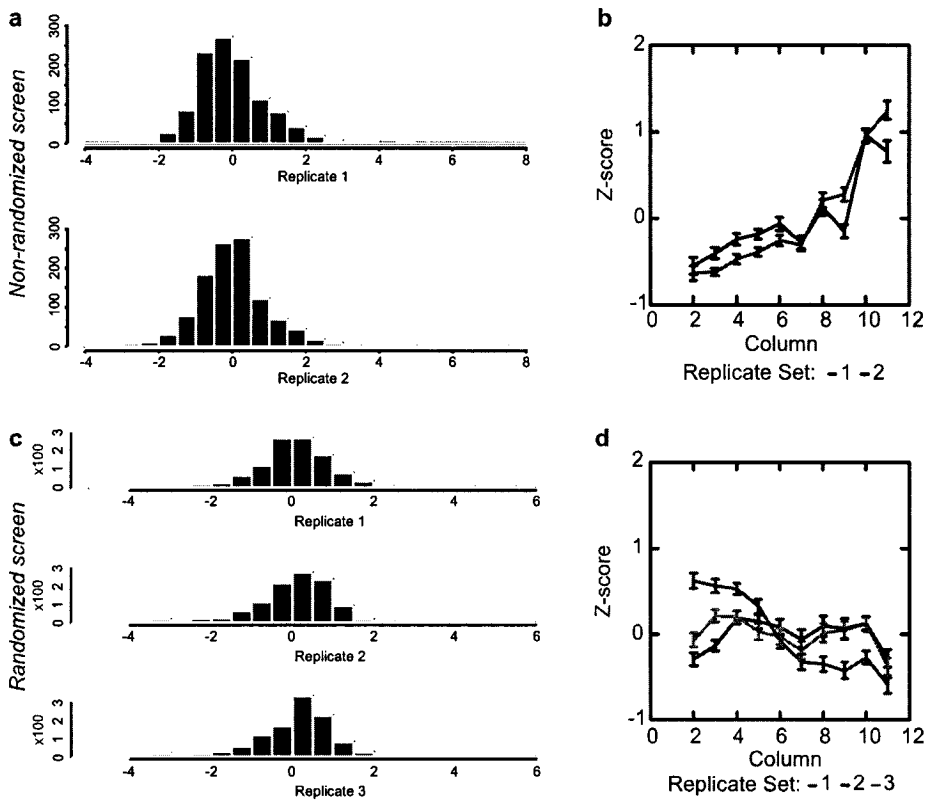

Figure 4–2: Graphical display of preprocessed data using the Z-score method. (a) Histograms of Z-scores for the non-randomized screen show less variability than the raw data. (b) Plot of average measurements against column number shows that column effects are present, especially for the right-most columns. (c) Histograms of Z-scores for the randomized screen again show more similar patterns than for the raw data. However, distributions all three distributions are unimodal. (d) The Z-scores corrects for plate effects, but not for column effects.

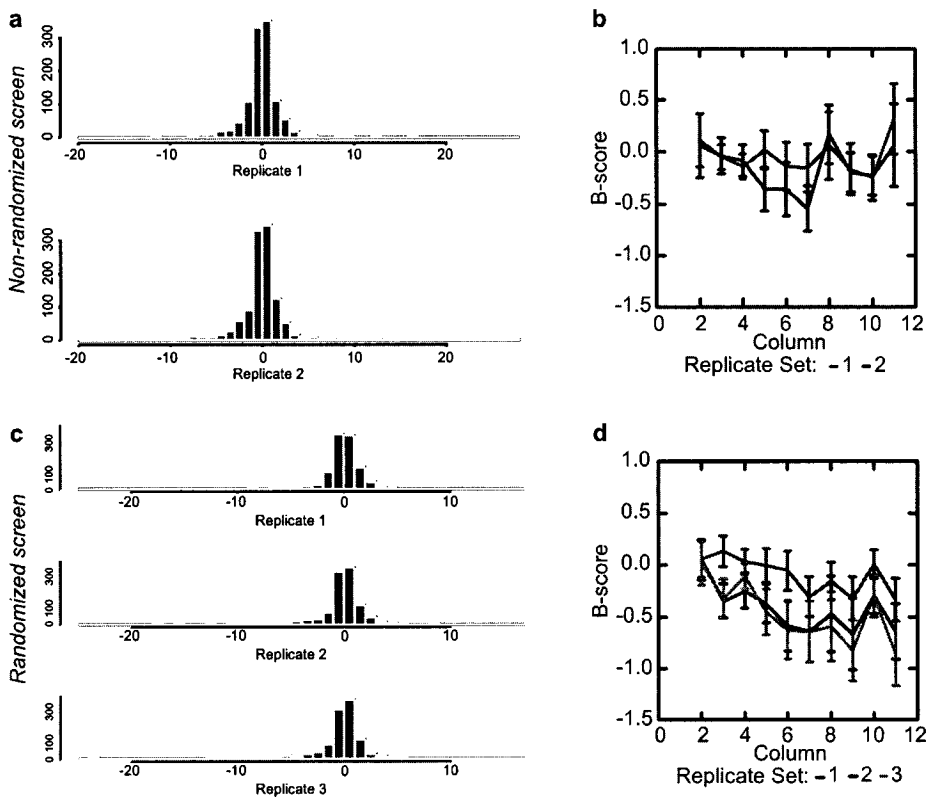

Figure 4–3: Graphical display of preprocessed data using the B-score method. (a) Histograms of B-scores for the non-randomized screen show less variability than the raw data or the Z-scores. (b) Plot of average measurements against column number shows that column effects have been removed. (c) Histograms of B-scores for the randomized screen again show more similar patterns than for the raw data or the Z-scores, and all three distributions are unimodal. (d) The B-scores corrects for plate effects as well as for row and column effects.

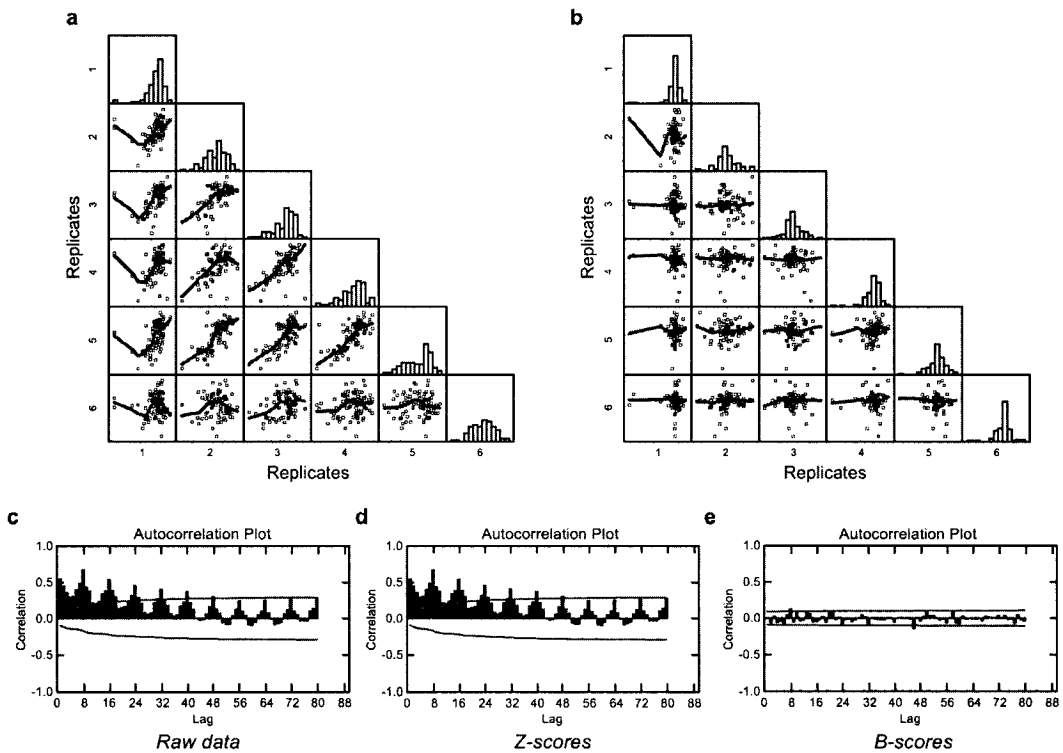

Figure 4–4: Scatter plots of raw and preprocessed data from a 'measurement experiment' in which the same compound was tested in all wells of several plates. (a) Because of procedurally-induced bias, measurements across plates were correlated. (b) The B-score method eliminates these biases, as evidenced by the lack of correlation among the replicated plates. The benefit of the B-scores normalization is also shown when looking at the autocorrelations (c,d,e), see text for details.

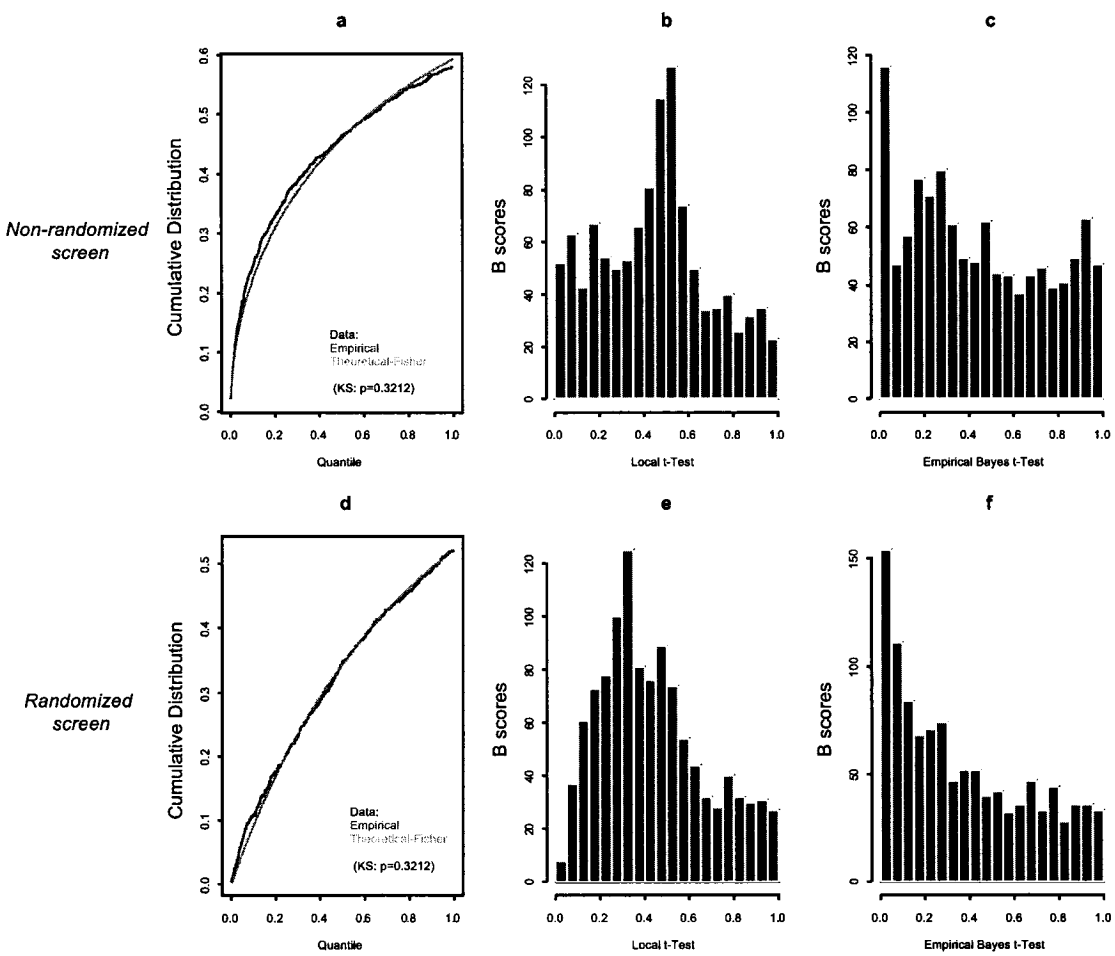

Figure 4–5: Checking of assumptions for statistical testing. (b,e) Under the null hypothesis of no hits, p-values should follow a uniform distribution, which is not the case with the 'local t-test'. However, for the Empirical Bayes t-test, we obtained a good fit under the assumption that the variances follow an inverse gamma distribution (a,d) and distribution of p-values is uniform with more low p-values, as expected in the presence of hits (c,f).

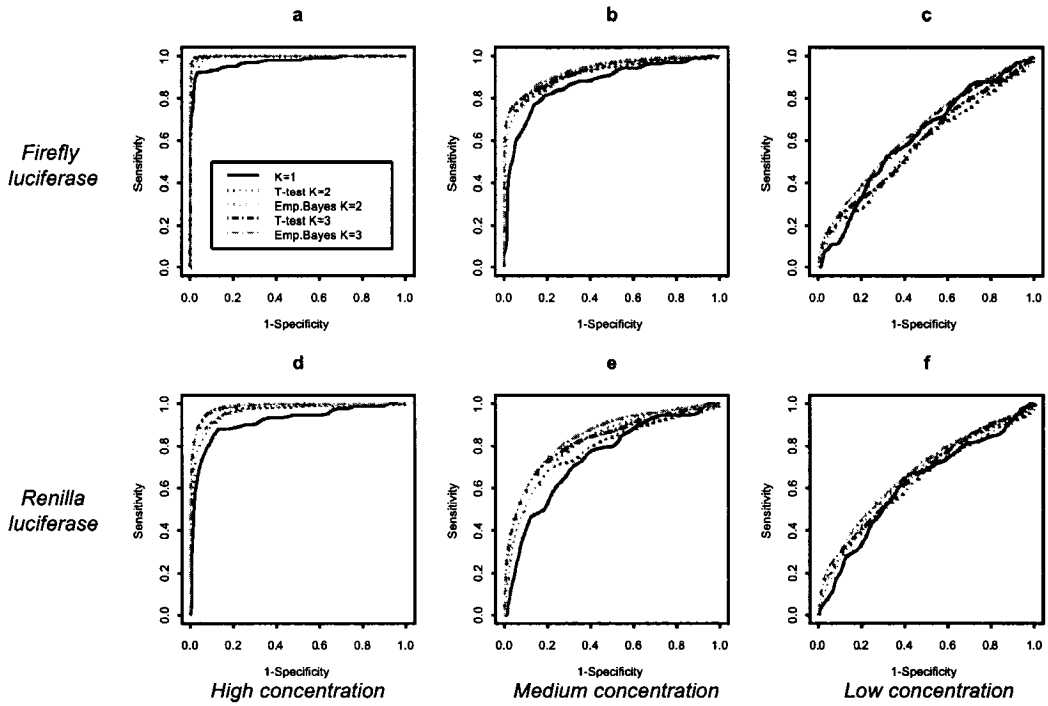

Figure 4–6: ROC curves to compare power achievable with various inferential approaches and various numbers of replicates. Data were generated according to a 'dilution series experiment'. The black line represents the rank ordering of activity measurements in the absence of replicates. The color curves illustrate the benefit of using statistical tests based on replicates. For the Firefly protein (a,b,c), hits are easily identify by all tests. For the Renilla protein (d,e,f), hits are more readily identified with the Empirical Bayes t-test (orange curves) than with the 'local t-test' (blue curves). All methods failed to identify hits at very low concentrations (c,f).

CHAPTER 5
Conclusion

At the beginning of my research three years ago, current practices relied on arbitrary and often non-statistical tools to analyze the increasing amount of HTS data generated daily. Worse, life scientists were relying on single measurements. They resisted obtaining replicates because of cost and time issues, and they did not fully appreciate the utility and benefit of statistics. They believed that automated technology and quality control were enough to produce reliable data. They were unaware of the biases that are caused by the presence of unwanted variation and of the importance of controlling false positive and false negative rates.

Consequently, the main objective of this thesis was to provide new efficient statistical methods to improve hit detection, and thus, the discovery of new drugs. However, the biggest part of my work has been to convince screeners of the benefit of statistical methods based on replicate measurements and to promote their use in HTS data analysis. These objectives have mostly been achieved by the writing of the three previously presented manuscripts, the last one containing the applications. Although each manuscript contains its own discussion, here are some overall conclusions.

I made several statistical recommendations which I divided in three different stages: experimental design, data preprocessing, and inference. When I first came into a HTS laboratory, with my statistical background and my limited knowledge

of biochemistry, everything was different from my point of view. My focus was on minimizing unwanted variation occurring during the entire screening process. For example, I proposed to prepare a big dilution and then split it in parts rather than preparing several small dilutions; to always use the same batch of compounds and reagents; to take plates from a same box; etc. Statistically, one solution was: since one cannot control for unknown potential sources or errors, one should randomize. I greatly complicated screeners'lives and they were discouraged to see me with my long list of random numbers! But my results suggest that randomization of plate processing order improves the reliability of results. However, it would be even better to perform *at the same time* the same screen twice, i.e. with and without randomization of plate processing order, in order to conclude strongly on the evidence of effects of randomization.

Second, at the preprocessing stage, I recommend the use of the B-score method [2]. This method offers the advantages of being robust and of removing row and column biases by using a two-way median polish [3]. In the first manuscript, I argued that normalization should not be based on controls, unless there is a major biological reason, since they may introduce their own biases. Although my results showed that the B scores are highly reliable, I pursued research on preprocessing methods in collaboration with bioinformaticians from UQÀM. The main idea is to correct for potential well effects, which can be though of as row and column interactions [51].

In addition, the second manuscript contain a generalization of the median polish procedure. I first believed that using median polish with replicates could improve data preprocessing. However, I finally realized that it was not the best way to

work with HTS data. Since each plate is performed separately, it is preferable to preprocess the data on a plate-by-plate basis. But the key findings of the second paper answer an open statistical question and thus, are of high statistical interest in and of themselves, and can certainly be applied to other fields.

Lastly, at the inferential stage, I suggested the use of an empirical Bayes t-test [20, 24, 34]. In the third manuscript, I demonstrated that the assumption of the constant variance among compounds was not satisfied and thus, that a classical z-test cannot be used. Since it is unrealistic to have a large number of replicates, a traditional t-test calculated individually for each compound will rely on few degrees of freedom. Consequently, the empirical Bayes t-test offers the advantages of estimating variance by a weighted mean of the compound-specific variance and the variance based on all compounds. The statistic is also compared to a t-distribution with more degree of freedom. My results showed the benefit of this method and that the data satisfied the assumptions.

In summary, I have provided warnings, recommendations, statistical thinking and methods, and my results have shown how to increase both the sensitivity and specificity of screens. In addition to publicly-available datasets, I had the opportunity to design my own empirical study and to perform specific experiments. Consequently, I was able to take advantage of scientific principles and to evaluate the performance of different methods.

However, since statisticians have just recently started to be involved in the HTS field, more work needs to be done in this area. First, statistical validation of hits must be defined. In microarrays, interest is on relative validation, which is easier. When a

gene is found in a RNA sample, then the same RNA sample and the same gene twice the amount are used for validation. In contrast, in HTS, interest is on absolute value. One way to validate a hit would be via a dilution series where concentrations are randomly located on plates which also contains a majority of non-active compounds. But again, one needs to define what the 'null' would be.

Second, replicates certainly improve sensitivity of screens, but their cost-benefit ratio needs to be examined. Although the higher the number of replicates, the better the estimates of the a and b parameters in the Empirical Bayes method, and the better the power. It is not clear that the increased costs are warranted by improvements in sensitivity. Somewhat differently, funds spent on new equipment and new technology to reduce variation in measurements, could also be spent to get replicates. For the same cost, one can test more compounds in single measurement or less compounds in replicates. Consequently, a very large simulation study needs to be performed to answer these questions, since the cost-benefit of replicates depends on several factors.

Third, in the first manuscript, I suggested partial replication as a compromise between the benefit and the cost of replicates. Again, all the aforemention general questions on optimization, even when the entire screen is replicated, are important here in addition to shrinkage issue. In addition, one needs to determine the number of compounds that may be get away to have the shrinkage working; i.e. the size of the subsample that must be replicated, the number of replicated plates that must be obtained, and the effect of shrinkage on the estimation of the parameters, and thus, power; etc. Several parameters needs to be considered in order to give guidelines.

Finally, this thesis is a big step towards efficient detection of high-quality hits, and thus, I believe that statistics will help increasing the number of drugs reaching the market.

Appendix A : Glossary

Agonist: A compound that binds to a receptor, enzyme, or protein and results in its activation.

Antagonist: A compound that acts to inhibit a receptor, enzyme, binding interaction, or cellular process.

Assay: An experimentally controlled biochemical or biologic system for detecting activity.

Compound: Chemical substance tested for desired molecular or cellular activity against the target in an assay or screen (e.g., clofoctol or anisomycin).

Collection: A large library, set, file, deck, bank, dispensary of chemical compounds.

Controls: A standard of comparison for screening results. Within plate controls are essential for identifying plate-to-plate variability and establishing assay background levels. Two types of controls are commonly used in early stages of HTS data analysis. Negative controls (referred to as background) represent the lowest possible measurement for the assay. Positive controls depict the maximum attainable measurement [46]. For example, in a yeast assay where a low activity measurement occurs when cell growth is inhibited by an active compound and a high activity measurement occurs when it is not, the absence of any compound might be used as a positive control and the absence of yeast as a negative control.

Counter Screen: A screen that tests the same compound library as in the primary screen, but against a related target in order to eliminate some hits seen in primary

screen.

Hit: A compound identified as having a "significant" molecular or cellular activity.

High-Throughput Screening (HTS): A process that allows the screening of several thousand chemical compounds in a period of a few weeks. The major applications of HTS are drug discovery and understanding of protein structures or biological pathways.

Lead: A hit validated by medicinal chemistry and structure-activity-relationship (SAR). A lead compound becomes a drug candidate for clinical trial.

Primary Screen: An initial high-throughput screen in which a compound library is tested against a target of interest in order to identify hits.

Reagent: A chemical or solution used to produce a desired chemical reaction (e.g., yeast or translation mix).

Screen: A large-scale assay performed using HTS automation.

Secondary Screen: A screen used for confirmation of initial hit compounds obtained from the primary screen by repeating the same assay and retesting against the same target in a new run, often done in duplicate.

Target: DNA, RNA, or protein that is involved in a disease process and is a suitable target for therapeutic compound development (e.g., rapamycin or protein synthesis) .

Appendix B : Reprint of Manuscript I

computational BIOLOGY

REVIEW

Statistical practice in high-throughput screening data analysis

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

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

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

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

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

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

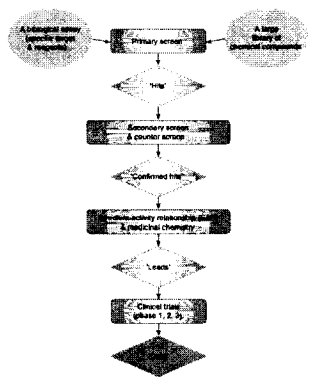

Figure 1 From HTS process to eventual drug development.

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

Published online 7 February 2006; doi:10.1038/nbt1186

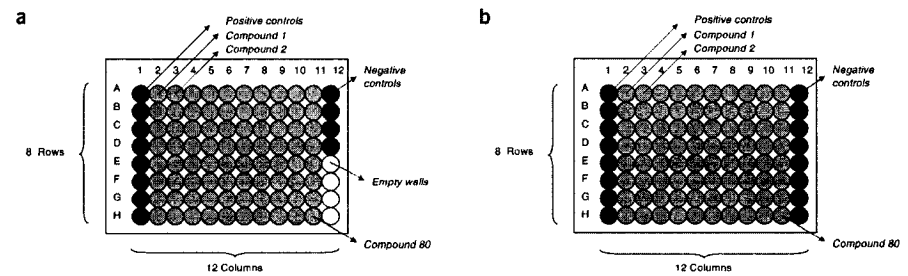

Figure 2 Typical location of controls on a 96-well plate. In a primary screen, the designed biological assay is performed by using a robot to add the target of interest and specific reagents to each well, which already contain a different compound or control. After incubation or other required manipulations,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

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

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

Box 1 Formulae for normalization

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

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

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

Normalized percent inhibition. Another normalization method using controls:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

HTS data processing

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

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

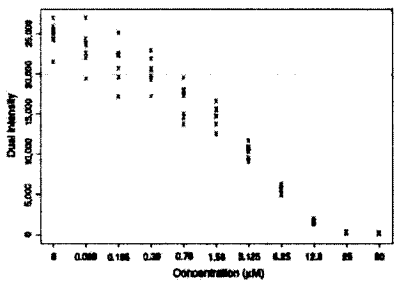

Figure 3 Titration series in a translation assay. These results from an anisomycin titration in a *Renilla* luciferase translation assay show that variability differs across the various concentrations. A hit may be defined as any activity measurement that is at least three standard deviations away from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

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

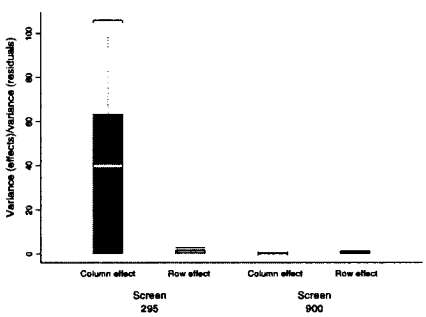

Figure 4 Presence of edge effects in a high-throughput screen. Data from two different screens (http://chembank.broad.harvard.edu/screens) with duplicate measurements across plates are presented. Tukey's two-way median polish was applied to each plate to obtain estimates of row and column effects and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

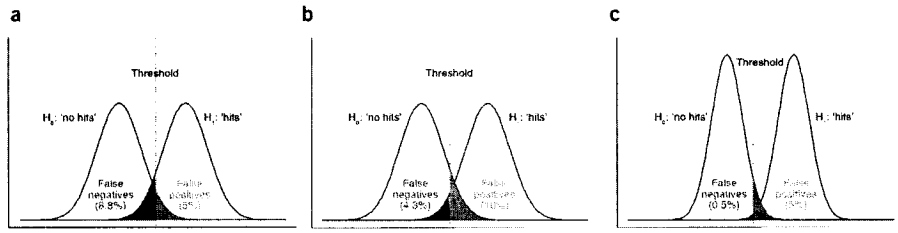

Figure 5 Replicates, false-positive and false-negative rates. In hypothesis testing a false-positive rate (type I error) is the probability of rejecting the null hypothesis (H_0) given that this hypothesis is true. The false-negative rate (type II error) is the probability of failing to reject the null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

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

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

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

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

Box 2 Examining the distribution of sample variances

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

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

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

For each compound, consider the model:

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

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

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

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

The a and b parameters are assumed to be constant across

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

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

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

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

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

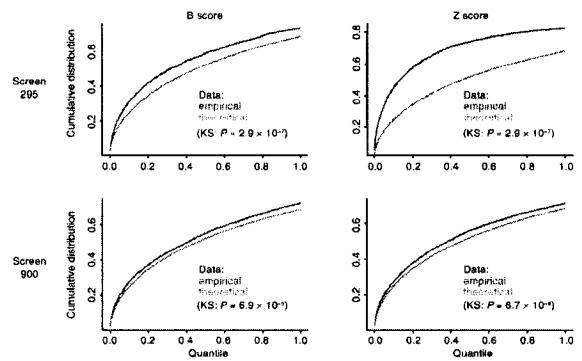

Figure 6 Verification of the assumptions of normally distributed data with constant variance among compounds. Empirical values correspond to a function of the sample variances. Under the assumption of a constant variance among compounds, the overall variance might be estimated by the mean of the sample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

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

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

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

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

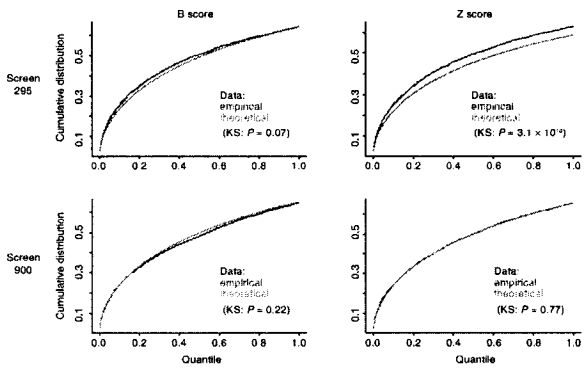

Figure 7 Verification of the assumption that the within-compound variances follow an inverse gamma distribution. Empirical values correspond to a function of the sample variances. Under the assumption of normally distributed data, each sample variance (obtained from the duplicate measurements) is multiplied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

Box 3 Test statistics for hit detection with replicates

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

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

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

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

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

where

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

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

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

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

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

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

Statistical inference and hit identification thresholds

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

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

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

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

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

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

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

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

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

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

Use of replicates

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

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

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

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

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

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

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

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

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

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

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

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

Conclusions

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

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

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

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

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

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

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

ACKNOWLEDGMENTS

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

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

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

1. Dove, A. Screening for content—the evolution of high throughput. *Nat. Biotechnol.* **21**, 859–864 (2003).
2. Landro, J.A. *et al.* HTS in the new millennium: the role of pharmacology and flexibility. *J. Pharmacol. Toxicol. Methods* **44**, 273–289 (2000).
3. Stein, R.L. High-throughput screening in academia: the Harvard experience. *J. Biomol. Screen.* **8**, 615–619 (2003).
4. Nelson, R.M. & Yingling, J.D. *Introduction to High-Throughput Screening for Drug Discovery* (IBC USA Conferences, Inc., San Diego, CA, 2004).
5. Campbell, D.T. & Kenny, D.A. *A Primer on Regression Artifacts* (Guilford Press, New York, 1999).
6. Stigler, S.M. Statistics on the Table: the History of Statistical Concepts and Methods (Harvard University Press, Cambridge, MA, 1999).
7. Lundholt, B.K., Scudder, K.M. & Pagliaro, L. A simple technique for reducing edge effect in cell-based assays. *J. Biomol. Screen.* **8**, 566–570 (2003).
8. Zhang, J.H., Chung, T.D.Y. & Oldenburg, K.R. Confirmation of primary active substances from high throughput screening of chemical and biological populations: a statistical approach and practical considerations. *J. Comb. Chem.* **2**, 258–265 (2000).
9. Tukey, J.W. A survey of sampling from contaminated distributions. in *Contributions to Probability and Statistics* (ed. Olkin, I.) 448–485 (Stanford University Press, Stanford, CA, 1960).
10. Brideau, C., Gunter, B., Pikounis, B. & Liaw, A. Improved statistical methods for hit selection in high-throughput screening. *J. Biomol. Screen.* **8**, 634–647 (2003).
11. Gunter, B., Brideau, C., Pikounis, B. & Liaw, A. Statistical and graphical methods for quality control determination of high-throughput screening data. *J. Biomol. Screen.* **8**, 624–633 (2003).
12. Hoaglin, D.C., Mosteller, F. & Tukey, J.W. *Understanding Robust and Exploratory Data Analysis* (Wiley, New York, 1983).
13. Buxser, S. & Vroegop, S. Calculating the probability of detection for inhibitors in enzymatic or binding reactions in high-throughput screening. *Anal. Biochem.* **340**, 1–13 (2005).
14. Chen, Y., Dougherty, E.R. & Bittner, M.L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Opt.* **2**, 364–374 (1997).
15. Rocke, D.M. Design and analysis of experiments with high throughput biological assay data. *Semin. Cell Dev. Biol.* **15**, 703–713 (2004).
16. Lee, M.L., Kuo, F.C., Whitmore, G.A. & Sklar, J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. USA* **97**, 9834–9839 (2000).
17. Nadon, R. & Shoemaker, J. Statistical issues with microarrays: processing and analysis. *Trends Genet.* **18**, 265–271 (2002).
18. Box, G.E.P., Hunter, J.S. & Hunter, W.G. *Statistics for Experimenters: Design, Innovation, and Discovery*, edn. 2 (Wiley-Interscience, Hoboken, N.J., 2005).
19. Wright, G.W. & Simon, R.M. A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **19**, 2448–2455 (2003).
20. Smyth, G. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, no.1, art. 3 (2004).
21. Baldi, P. & Long, A.D. A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519 (2001).
22. Verkman, A.S. Drug discovery in academia. *Am. J. Physiol. Cell Physiol.* **286**, C465–C474 (2004).
23. Kerns, E.H. & Di, L. Pharmaceutical profiling in drug discovery. *Drug Discov. Today* **8**, 316–323 (2003).
24. Fay, N. & Ullmann, D. Leveraging process integration in early drug discovery. *Drug Discov. Today* **7**, S181–S186 (2002).

References

[1] M. E. Hahn, S. B. Haber, and J. L. Fuller. Differential agonistic behavior in mice selected for brain weight. *Phys. and Behavior*, 10:759–762, 1973.

[2] C. Brideau, B. Gunter, B. Pikounis, and A. Liaw. Improved statistical methods for hit selection in high-throughput screening. *Journal of Biomolecular Screening*, 8:634–647, 2003.

[3] J. W. Tukey. *Exploratory Data Analysis.* Addison-Wesley, Reading, MA, 1977.

[4] W. Terbeck and P. L. Davies. Interactions and outliers in the two-way analysis of variance. *Annals of Statistics*, 26:1279–1305, 1998.

[5] P. L. Davies. Identifying interactions and outliers in the two-way analysis. *Unpublished*, http://wwwstat.mathematik.uni-essen.de/~davies/f2way.html, 2002.

[6] A. Dove. Screening for content–the evolution of high throughput. *Nature Biotechnology*, 21:859–64, 2003.

[7] J. A. Landro, I. C. Taylor, W. G. Stirtan, D. G. Osterman, J. Kristie, E. J. Hunnicutt, P. M. Rae, and P. M. Sweetnam. Hts in the new millennium: The role of pharmacology and flexibility. *Journal of Pharmacological and Toxicological Methods*, 44:273–89, 2000.

[8] R. L. Stein. High-throughput screening in academia: The harvard experience. *J Biomol Screen*, 8:615–9, 2003.

[9] Yingling Jeffrey D. Nelson, Richard M. Introduction to high-throughput screening for drug discovery. In IBC's TRAINING ACADEMY Courses, editor, *ScreenTech World Summit*, San Diego, Calif.,, 2004. IBS USA Conferences, Inc.

[10] Donald Thomas Campbell and David A. Kenny. *A primer on regression artifacts.* Methodology in the social sciences. Guilford Press, New York :, 1999.

[11] Stephen M. Stigler. *Statistics on the table: The history of statistical concepts and methods.* Harvard University Press, Cambridge, Mass. :, 1999.

[12] B. K. Lundholt, K. M. Scudder, and L. Pagliaro. A simple technique for reducing edge effect in cell-based assays. *Journal of Biomolecular Screening*, 8:566–570, 2003.

[13] J. H. Zhang, T. D. Y. Chung, and K. R. Oldenburg. Confirmation of primary active substances from high throughput screening of chemical and biological populations: A statistical approach and practical considerations. *Journal of Combinatorial Chemistry*, 2:258–265, 2000.

[14] J. W. Tukey. A survey of sampling from contamined distributions. In Ingram Olkin, editor, *Contributions to probability and statistics*, pages 448–485. Stanford University Press, Stanford, Calif.,, 1960.

[15] C. Brideau, B. Gunter, B. Pikounis, and A. Liaw. Improved statistical methods for hit selection in high-throughput screening. *Journal of Biomolecular Screening*, 8:634–647, 2003.

[16] B. Gunter, C. Brideau, B. Pikounis, and A. Liaw. Statistical and graphical methods for quality control determination of high-throughput screening data. *J Biomol Screen*, 8:624–33, 2003.

[17] David C. Hoaglin, Frederick Mosteller, and John Wilder Tukey. *Understanding robust and exploratory data analysis.* Wiley, New York, 1983.

[18] S. Buxser and S. Vroegop. Calculating the probability of detection for inhibitors in enzymatic or binding reactions in high-throughput screening. *Analytical Biochemistry*, 340:1–13, 2005.

[19] Dougherty E.R. Bittner M.L. Chen, Y. Ratio-based decisions and the quantitative analysis of cdna microarray images. *Journal of Biomedical Optics*, 2:364–374, 1997.

[20] David M. Rocke. Design and analysis of experiments with high throughput biological assay data. *Seminars in Cell and Developmental Biology*, 15:703–713, 2004.

[21] M. L. Lee, F. C. Kuo, G. A. Whitmore, and J. Sklar. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cdna hybridizations. *Proceedings of the National Academy of Sciences of the United States of America*, 97:9834–9, 2000.

[22] R. Nadon and J. Shoemaker. Statistical issues with microarrays: processing and analysis. *Trends Genet*, 18:265–71, 2002.

[23] George E. P. Box, J. Stuart Hunter, and William Gordon Hunter. *Statistics for experimenters: Design, innovation, and discovery.* Wiley series in probability and statistics. Wiley-Interscience, Hoboken, N.J., 2nd edition, 2005.

[24] G. W. Wright and R. M. Simon. A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*, 19:2448–2455, 2003.

[25] G. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3, 2004.

[26] P. Baldi and A. D. Long. A bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17:509–519, 2001.

[27] A. S. Verkman. Drug discovery in academia. *American Journal of Physiology and Cell Physiology*, 286:C465–74, 2004.

[28] E. H. Kerns and L. Di. Pharmaceutical profiling in drug discovery. *Drug Discovery Today*, 8:316–23, 2003.

[29] N. Fay and D. Ullmann. Leveraging process integration in early drug discovery. *Drug Discov Today*, 7:S181–6, 2002.

[30] N. Cressie. Kriging nonstationary data. *Journal of the American Statistical Association*, 81(395):625–634, 1986.

[31] O. Berke. Modified median polish kriging and its application to the wolfcamp-acquifer data. *Environmetrics*, 12:731–748, 2001.

[32] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J.Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[33] M. L. Lee, F. C. Kuo, G. A. Whitmore, and J. Sklar. Importance of replication in microarray gene expression studies: Statistical methods and evidence from

repetitive cdna hybridizations. *Proc. Nat.1 Acad. Sci.U. S. A.*, 97:9834–9839, 2000.

[34] N. Malo, J. A. Hanley, S. Cerquozzi, J. Pelletier, and R. Nadon. High-throughput screening data analysis: A critical look at statistical practice. *Nature Biotechnology*, 2006.

[35] D. C. Hoaglin, F. Mosteller, and J. W. Tukey. *Exploring Data Tables, Trends, and Shapes.* John Wiley and Sons, New York, 1985.

[36] J. D. Emerson and D. C. Hoaglin. *Analysis of Two-Way Tables by Medians, in: Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (Eds.), Understanding Robust and Exploratory Data Analysis.* John Wiley and Sons, Inc., New York, 1983.

[37] V. A. Sposito. On median polish and l_1 estimators. *Computational Statistics and Data Analysis*, 5:155–162, 1987.

[38] P. J. Huber. *Robust Statistics.* Wiley, New York, 1981.

[39] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: An Approach Based on Influence Functions.* Wiley, New York, 1986.

[40] G. W. Brown and A. M. Mood. On median tests for linear hypotheses. *Second Berkeley Symposium*, pages 159–166, 1951.

[41] T. P. Hettmansperger and R. Elmore. Tests for interaction in a two-way layout: Should they be included in a nonparametrics course? *ICOTS6*, 2002.

[42] C. J. Scheirer, W. S. Ray, and N. Hare. The analysis of ranked data derived from completely randomized factorial designs. *Biometrics*, 32:429–434, 1976.

[43] R. Rosenthal and R. L. Rosnow. *Essentials of Behavioral Research: Methods and Data Analysis.* McGraw-Hill, Inc., New York, 2nd edition, 1991.

[44] C. Daniel. Patterns in residuals in the two-way layout. *Technometrics*, 20:385–395, 1978.

[45] M.W. Lutz, J.A. Menius, T.D. Choi, R. Gooding Laskody, P.L. Domanico, A.S. Goetz, and D.L. Saussy. Experimental design for high-throughput screening. *Drug Discovery Today*, 1:277–286, 1996.

[46] J. H. Zhang, T. D. Y. Chung, and K. R. Oldenburg. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *Journal of Biomolecular Screening*, 4:67–73, 1999.

[47] D. Kevorkov and V. Makarenkov. Statistical analysis of systematic errors in high-throughput screening. *Journal of Biomolecular Screening*, 10:557–567, 2005.

[48] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:289–300, 1995.

[49] W.S. Cleveland. *Visualizing Data.* AT&T Bell Laboratories, Murray Hill, New Jersey, 1993.

[50] D.B. Allison, X. Cui, G.P. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nature Genetics*, 7:55–65, 2006.

[51] V. Makarenkov, D. Kevorkov, P. Zentilli, A. Gagarin, N. Malo, and R. Nadon. Statistical analysis of systematic errors in high-throughput screening. *Journal of Biomolecular Screening*, 10:557–567, 2005.