Title
An introduction to statistical issues in High throughput screens

Authors
Chiara Sabatti
Koppany Visnyei
Harley Kornblum

Publication Date
2011-10-25

An introduction to statistical issues in High throughput screens

Chiara Sabatti[1], Koppany Visnyei[2], Harley Kornblum[2,3]

1 Department of Human Genetics and Department of Statistics, UCLA, Los Angeles, CA.

2 Department of Pharmacology, UCLA, Los Angeles, CA

3 Semel Institute for Neuroscience and Human Behavior, Department of Psychiatry and Department of Pediatrics, UCLA, Los Angeles, CA

UCLA Statistics Department Preprint # 532

January 2008

Abstract

We describe the nature and goals of high-throughput screening experiments, focusing on the challenges they present from the view-point of statistical analysis. We suggest graphical displays to facilitate quality control. We describe sources of systematic variation and methods to correct for it. We consider the problem of ranking compounds with respect to their effects on one cell type and we suggest a couple of procedures, depending on the available number of replicates. Finally, we explore the use of a hierarchical framework for hypothesis testing, to study the effects of compounds in multiple cell lines.

1 Introduction

High-throughput screening (HTS) has been used to assay chemical compound libraries, in order to identify molecule candidates that exhibit a specific biological or biochemical effect on a target cell type. Traditionally, HTS have been mainly carried out in pharmaceutical and biotech industries, as one of the initial steps towards drug discovery. In the last few years, however, a number of academic and public research insti-tutions have established their own screening facilities, as these need notbe narrowly focused on drug discovery, but generate exciting opportunities for basic science. The biology community has now grown accustomed to large scale datasets that provide comprehensive surveys of genomes, gene expression, splice variants, etc. Datasets generated with HTS fit well within this paradigm. For example, through HTS it is possible to study the effects of RNA interference [5], and gene expression-based HTS has been used to generate the "connectivity map" resource that can be used to find connections among small molecules, physiological processes, diseases and drugs [9].

UCLA has one HTS core, the Molecular Screening Shared Resource [12] which has automated equipment and data collection that make it possible to test about 60,000 molecules on a system of interest in about a day. All the data analyzed in this paper has been generated by the UCLA MSSR.

As the academic community starts taking more and more advantage of the HTS technology, a number of challenges develop. For example, while this is a flexible technology, that can be used with multiple purposes, the design of experiments needs to be adapted to the different settings. Or, differently from the lengthy and secretive process of drug discovery, where only final outcomes are known, the transparency and publication requirements of basic research make it necessary to be able to quantify at each stage in the process which molecules appear to have an effect. This translates in a need for statistical quantification of the observed effects. Indeed, the research community has started to recognize that a greater involvement of statisticians in the analysis of HTS screen data would be beneficial [13]. This need is only going to grow as these new datasets are going to be used by scientists in conjunction with other data, attempting to understand the complex functioning of biological systems.

While few statisticians have taken up the challenges presented by HTS data (see, for example, [10, 11]), the vast majority of applied statisticians with interest in biological systems, still ignores the nature of these experiments. This is a reality bound to change, given the potential offered by HTS and their likely integration with other biological data sources. With the present paper we would like to familiarize the statistics community at large with HTS and outline some of the statistical challenges it presents. We begin with a description of the experimental procedure and its goals.

2 Highthroughput screens

The goal of HTS is the identification of a subset of small molecules (compounds) in a given library with an effect on a cell type of interest. The library could be compiled in a variety of ways, with compounds preselected (or not) with respect to the specific biological questions. While HTS can be used with a variety of compound libraries (such as natural compound libraries, small inhibitory RNA libraries, etc.), we will focus on screens based on chemical libraries of small molecules. The

read-out of the screen can be the expression of a certain protein (such as fluorescent proteins that correlate with the expression of a certain gene, for example), or even highly complex features like cell morphology, that can be analyzed with the help of automated microscopy and an appropriate software evaluating the cell shape. However, the most common read-out is cell number or cell viability. This can be measured indirectly, mostly using different commercially available kits that convert the number of live cells into readable signals, like fluorescence or luminescence. In this latter case, in a given library, one would typically find some compounds that do not modify the growth rate of the cells, some that increase it and some that decrease the growth rate, or even kill the cells. Roughly speaking, a HTS consists in exposing cells to a large collection of compounds, incubate them, and measure their growth levels after an appropriate period of time. This is carried out in a highly automated fashion, taking advantage of progress in biotechnology.

A key piece of HTS equipment is a plate: a small container, usually made of plastic, that features a grid of small, open divots called wells (see Figure **??**). Commonly both 94 well plates and 384 well plates are used. All the data analyzed in this paper was obtained using 384 well plates, with 16 rows and 24 columns.

Cells are dispensed (using an automated cell dispenser) as single cell solutions into the plates at a specific concentration, and in uniform media. The automated cell dispenser used in the experiments we analyze here has 8 pins, each filling up two rows in the plate. Rows are filled in the order described in Figure 1.

After dispensing the cells, the plates are being exposed to the compounds, which are stored in a corresponding 384 well plates. Through an automated system, 384 pins are lowered into the compound plate and then lowered into the wells of the plate containing the cells. Multiple plates are typically necessary to query all the compounds of interest. The 384 pins are being washed in DMSO, ethanol, methanol and blow-dried between each plate. Each pin is calibrated to transfer exactly 0.5?L of dissolved compound into a well.

A key piece of HTS equipment is a plate: a small container, usually made of plastic, that

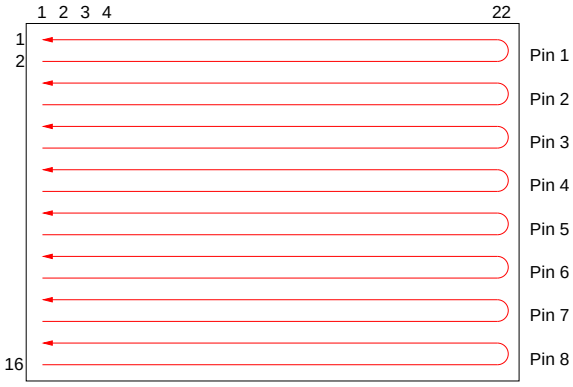


Figure 1: Schematic representation of the order with which cells are dispensed into wells by 8 pins. Each pin fills two rows of wells.

features a grid of small, open divots called wells (see Figure ??). Commonly both 94 well plates and 384 well plates are used. All the data analyzed in this paper was obtained using 384 well plates, with 16 rows and 24 columns.

Finally, one of a number of different methods is used to obtain a numerical read-out of the density of live cell in each well. For example (and as in the case of the data analyzed here), ATPLite (Perkin Elmer, USA) can be added to the plates, which are further incubated and evaluated with an automatic luminescence-reader. ATPLite reacts with ATP present in live cells, and the emitted luminescence is directly proportional with the density of live cell in each well. These intensity reads represent the final out come of the screen. It should be mentioned, that methods establishing cell densities based on concentrations of cell metabolites, will be influenced not only by the shear number of the live cells in a well, but also by the metabolic activity of those cells. Hence, results of such read-outs should be interpreted accordingly.

Typically, multiple rounds of screens are performed for each problem. Initial screens query vast number of compounds (tens of thousands) and tend not to include any replicate, mostly out of cost reasons. Secondary and tertiary screens focus on hits of initial experiments and investigate further their effects, possibly considering different compound concentrations, cell types, incubation

times and replicate experiments. Scientic questions of interest, then, are: how do we identify promising candidates in the initial screen? When is the evidence from screen sufcient to establish the functional effect of a compound?

In this paper we present some of the statistical issues one encounters when trying to address such questions. Rather then taking a very general viewpoint, we focus on the challenges encountered in the analysis of two scientific experiments, carried out at the UCLA screening facility by members of the Kornblum's research group. The overall goal of these experiments is to identify compounds that reduce or increase the grow of cancer cells (studied with two cancer cell lines, called '107' and '1600'), possibly leaving unaltered control (non-tumor) cells (investigated via cell line '293T'). We present the analysis of data from two screens conducted within this research program.

The initial screen explored the effect of over 30,000 compounds on cell line 107. Each compound was evaluated only once, requiring a total of 94 plates. The follow-up screen focused on the ≈ 1000 most promising candidates from the first screen. These were re-screened, in duplicates, in the original cell type and two additional others (1600 and 293T).

Both screens used a 384 well plates. In the primary screen, columns 1 and 24 of the plate were left empty (as it is common practice to avoid evaporation effects), and wells in columns 2 and 23 were plated with cells that were exposed to the dissolvent of the chemical molecules only, namely DMSO, (but not to actual chemical compounds). These DMSO treated wells were serving as controls. The secondary screen used three plates (plate types A, B, C for a total of 786 compounds) to assess putative "killer" compounds and one plate to access putative "enhancer" compounds (plate type D, with a total of 185 compounds). Again, columns 1 and 24 were left empty. Rows 1 and 16, and columns 2 and 23 were used for controls. In addition, control wells were distributed on the entire plate, as illustrated in Figure 2. Table 1 provides a summary description of which compounds were screened for which cell line in the 24 plates that constituted our follow-up study.

To conclude this section, let us introduce some notation. In the rest of the paper, we will use

	107	293T	1600
Plate Type A (293 'killers')	1, 2	9, 10	17, 18
Plate Type B (293 'killers')	3, 4	11, 12	19, 20
Plate Type C (200 'killers')	5, 6	13, 14	21, 22
Plate Type D (185 'enhancers')	7, 8	15, 16	23, 24

Table 1: Design of the secondary screen. The compounds were plated in four plate types, with plate types A, B, C containing putative killers and plate type D putative enhancers. Each of the plate types was screened twice against each of the three cell lines, leading to 24 plates, identified with number 1-24 in the table. Plates were prepared in three different times, corresponding to the three different colors of the plate numbers.

symbol y_i to indicate the i-th intensity read measuring the effect of compound k_i on cell line ℓ_i, and obtained from a well in row r_i and column c_i, on plate p_i, and batch b_i. A special class of compounds in represented by controls: we use the letter N (neutral) to indicate baseline controls and E and S to indicate positive (known to enhance growth) and negative controls (known to suppress cell growth). Finally, we precise that we are actually going to work with the logarithm of the raw intensity reads: this is standard practice, corresponds to the biological understanding of changes in cell growth, and leads to a more symmetric distribution of errors.

3 Quality control issues

As every experimental device, HTS are subject to variability in the quality of the obtained data. Perhaps the first aspect on which the existing statistical literature on HTS has focused is the need of assessing the quality of an assay. Zhang *et al.* [22], in one of the first papers on the subject, introduced what has become known as Z factor (and Z') to evaluate the discriminatory power of

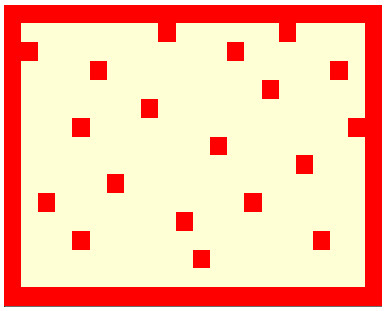


Figure 2: A possible 16×24 plate design with uniformly spaced controls. Each square corresponds to a well. Well devoted to controls are indicated with darker color.

a particular screen. At an even more basic level, the first step in the analysis of HTS data has to consists in the evaluation of the success of the experiment corresponding to each plate. While the automation and standardization of the process has increased, it is still important to consider that some plates may not lead to valuable signals. This is especially true in secondary screenings, where smaller cell amounts are used, more hands-on intervention is required, and there is generally a higher variability across plates. Figure 4,for example, illustrates one example of 'failed' plate in the secondary screen we analyzed.

We are interested in the values y_i from one plate ($p_i = p$), and hence reflecting one cell line $\ell_i = \ell$ and obviously one batch $b_i = b$. To evaluate the presence of biologically meaningful signal in the overall reads from plate p, one relies especially on the signal y_i relative to compounds of known effect. For example, the mentioned Z' has the following form:

$$Z' = 1 - \frac{(3\bar{\sigma}_E + 3\bar{\sigma}_S)}{|\bar{y}_E - \bar{y}_S|},$$

where $\bar{y}$ and $\bar{\sigma}$ indicate the sample averages and standard deviations for y_i such that $k_i = E$ and $k_i = S$.

In the screens we analyzed, only neutral controls were used, making the use of such statistics impractical. However, information can be gathered by the comparison of the reads from neutral controls $\{i : k_i = N\} := \mathcal{N}$ and the rest of the data $\{i : k_i \neq N\} := \mathcal{E}$. We can assume that

each $y_i = \mu_i + z_i$, with μ_i representing the effect of compound k_i on the given cell line and z_i the experimental error. We will devote the next section to describe in detail the systematic components of this variability and how it can be corrected. For the time being, we assume that $E(z_i) = 0$. Let's also focus on the difference between each individual μ_i and the neutral value μ_N: $\mu_i = \mu_N + \theta_i$.

In a primary screen, when a large number of compounds of unknown effect are studied, one can assume that the average $\sum_{i\in\mathcal{E}} \theta_i \approx 0$: a large number of compounds will have no effect, and there may be enhancers as well as killers. Hence, comparing the sample averages $\bar{y}_\mathcal{E}$ and $\bar{y}_\mathcal{N}$ would not be very informative. Instead, one can compare the $\sum_{i\in\mathcal{N}} y_i^2/|\mathcal{N}|$ with $\sum_{i\in\mathcal{E}} y_i^2/|\mathcal{E}|$: as long as some θ_i are different from zero and the error variance is (on average) the same across control and experimental spots, $\sum_{i\in\mathcal{E}} y_i^2/|\mathcal{E}|$ is larger in expectation. Plates for which these two values appear close are likely to be unsuccessful experiments. Rather then attempting to define a precise cut-off value that would rest on a series of unjustified hypothesis at this stage, we suggest to use graphical means of comparison. In particular, we find that comparing the spread of the distributions of $\{y_i, i \in \mathcal{N}\}$ and $\{y_i, i \in \mathcal{E}\}$ with box plots can be particularly effective (see Figure 3).

In secondary screens, where compounds have been selected, one expects a large fraction of the θ_i to be different from 0 and actually with the same sign (given that one is interested in studying enhancers or killers alone). It is reasonable, then, that $\sum_{i\in\mathcal{E}} \theta_i < 0$ (or > 0) and it becomes meaningful to compare $\bar{y}_\mathcal{E}$ and $\bar{y}_\mathcal{N}$. Again, we believe that graphical means as box plot offer an attractive instrument. Figure 4 gives one example of such comparison. As in Table 1, plate type C queries 200 compounds that appeared to be killers, and plate type D 185 compounds that appeared to be growth enhancers in the primary screen. There are 184 and 199 control wells in plate types C and D. In plate 5, the putative killers on C are re-screened against the original cell line 107: we can clearly see that on average experimental compounds have lower intensities then control wells. Analogously, on plate 7, putative enhancers on D are re-screened against cell line 107: this time, however, there is no appreciable difference between the distribution of control and experimental wells. One might interpret this result as a consequence of inappropriate choice of 'enhancers'; an

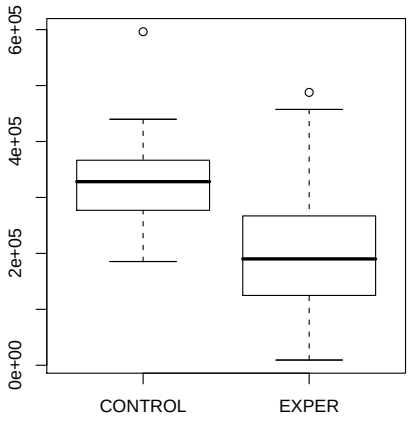


Figure 3: Different spread in control and experimental wells. The box plots are based on measurements from the first plate of our primary screen, which contained 320 experimental compounds and 32 control wells. The spread of the experimental distribution is clearly larger then the control one. Note that control also appear to have higher intensity than experimental compounds. However, this is an artifact due to the position occupied by controls in the plates (see section 4).

analysis of the results on plate 23, however, suggests another interpretation. In plate 23, the same compounds are screened against cell line 1600, and here a clear increase over the control levels is visible. It seems likely, hence, that the reaction on plate 7 did not work properly, leading to low resolution power.

As a summary, it is important to inspect the success of each experiment by comparing the distribution of controls wells and compound wells in a plate. Resorting to graphical displays like the one presented in Figures 4 and 3 is very helpful. We also recommend using positive and negative controls whenever possible.

4 Detection and correction of systematic effects

In the previous section, where the focus was on the aggregate behavior from a plate, we have assumed that $E(\sum_i z_i) \approx 0$. While this was reasonable for those purposes, it is important to recognize that, in general, $E(z_i) \neq 0$. There are a number of systematic effects that we were able to observe in our experiments, and, indeed, the most recent statistical literature on HTS has focused on their detection and correction (see for example [4, 8, 10, 11]). Following is a list of observed systematic sources of variability.

- The average read-out value varies across plates [plate effect]
- In cell number based assays, within one plate, the intensity reads for wells closer to the edges tend to be higher. More generally, the position of a well in a plate influences the associated intensity value [location effect]
- During the compound spotting process, technical artifacts may result in higher/lower amounts of the queried compounds or otherwise perturb the environment of a well, so that its reads are consistently higher or lower [well effect]

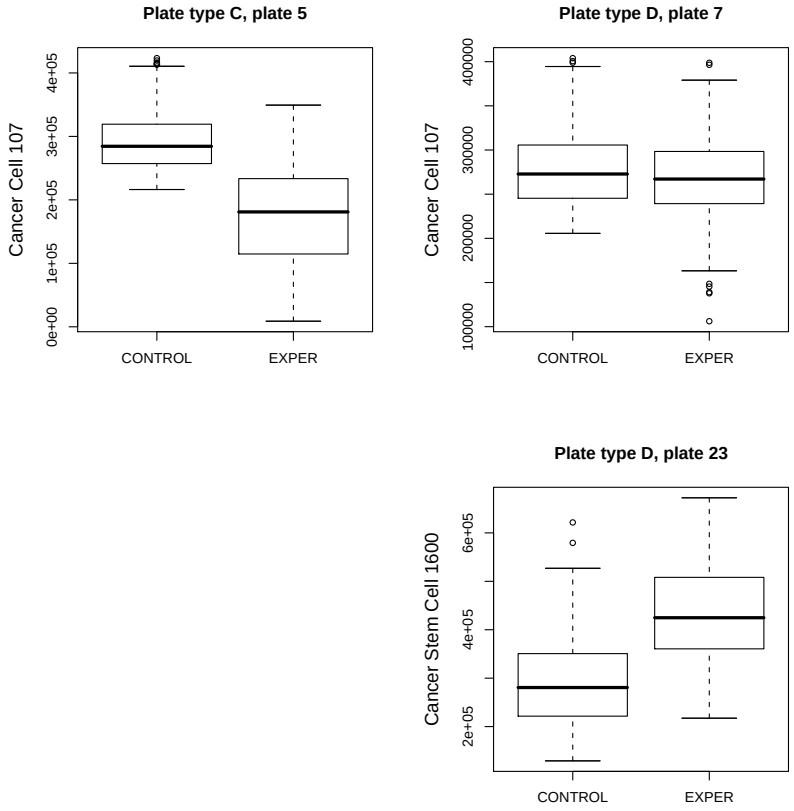


Figure 4: Visualizing the signal amounts in different experiments. Box plots in the left column refer to one experiment with plate type C, while box plots in the right column refer to two experiments with plate type D. The first row reports results obtained with cell line 107 and the bottom row results obtained with cell line 1600. The specific plate number is indicated for each graph.

- Cells cultures plated at different times may have different viability and responsiveness [batch effect]
- The pins used by the cell dispenser may exibit different behavior due to technical artifacts [pin effect]
- During a session of automated plating, the accuracy of any of the instrument my degrade, originating a time trend across plates [order effect]
- The quantity of cells available in the flasks may influence the amount spotted in different wells on the plate, depending on the order with which wells get filled [dispenser effect]

The goal of HTS is to learn about the effect of queried compounds on the cell type of interest. In an attempt to do so, it is necessary to estimate the systematic components of error described and remove them. With this goal in mind, we assume that

$$y_i = \mu(k_i, \ell_i) + s(r_i, c_i, p_i) + \epsilon_i. \tag{1}$$

The function $s(r_i, c_i, p_i)$ can capture plate, batch, pin, location, order and dispenser effects. Generally speaking, we can estimate some form of $s(r_i, c_i, p_i)$ from y, assuming that the compound-specific effects $\mu(k_i, \ell_i)$ do not have trends that correlate with the systematic errrors. The form of this function that we should estimate, depends on the type of effects that are more immediately noticeable and on the nature of the data at hand, specifically on which replicates are available. For example, if we restrict our attention to data coming from one plate only, $s(r_i, c_i, p_i)$ can be a function of r_i and c_i, capturing location and possibly dispenser effects, but well effects will be confounded with the main effects of interest. This is the approach adopted in some of the first contributions in this area. For example, in [8] the following model is fitted for each plate:

$$s(r_i, c_i, p_i) = m_{p_i} + \text{poly}_{p_i}(r_i, c_i, 2),$$

where $\text{poly}_{p_i}(r_i, c_i, d)$ indicates a polynomial of degree 2 in r_i and c_i, and $\hat{s}$ is estimated using a least squares approach. In [4], instead, expressing concerns for outliers, the authors use a median polish technique to fit an ANOVA model with row and column effects:

$$s(r_i, c_i, p_i) = m_{p_i} + R_i + C_i,$$

where R_i and C_i represent factors for the row and columns values.

To estimate well effects [10, 11] one needs to consider data from multiple plates and assume that some component of the function $s(r_i, c_i)$ is constant across plates. A two-stage approach is adopted in [11] where well and location effects are estimated in two successive steps.

Our approach to the estimation of the systematic error component in (1) varied in the case of primary versus secondary screen. In the primary screen, we observed a very marked location effect, as well as plate and well effects. This is clearly apparent from Figures 5 and 6 that present data from all the 94 plates of our initial screen. Intensity values y_i for each well in each plate are plotted against their corresponding row r_i (column c_i), stratified by their column (row). It is easy to see that, for each column, the intensity values decrease in correspondence of the middle rows and, for each row, intensity values decrease in the central columns. In addition to this rather smooth trend, there appear to be some well specific effects. The bottom of Figure 6, for example, gives a zoomed version of the plot of intensity values per each row, in column 11 across all 94 plates. It is easy to see how some wells have comparatively lower intensity then neighboring ones. Figures 5 and 6 at least partially show how location effects are fairly constants across plates. Indeed, we were unable to detect significant batch or order effects. Our controls were, unfortunately, all positioned in the same location (first and last columns), so that control means were confounded with the higher intensity values observed along the edges of the plates. On the basis of this observation, we decided not to use control intensity values to fit a smooth trend in the plates. Given that we had data one 94 plates, we were able to corrected both for well effects a smoother trends, simply fitting

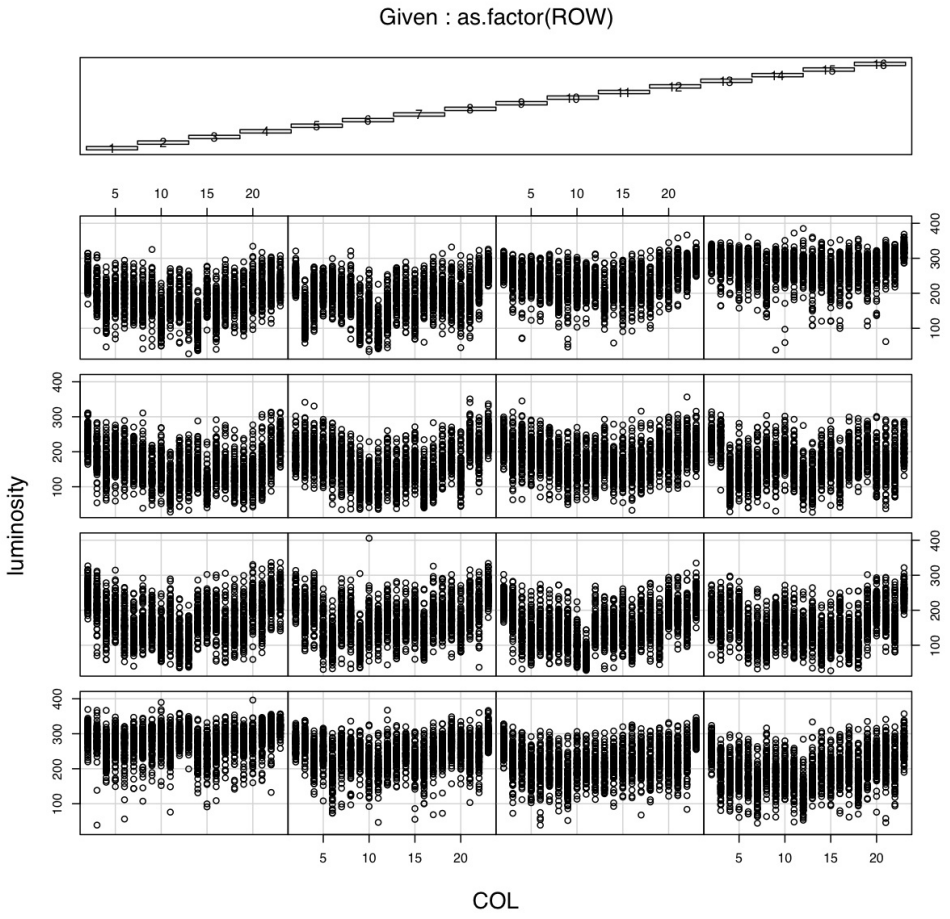
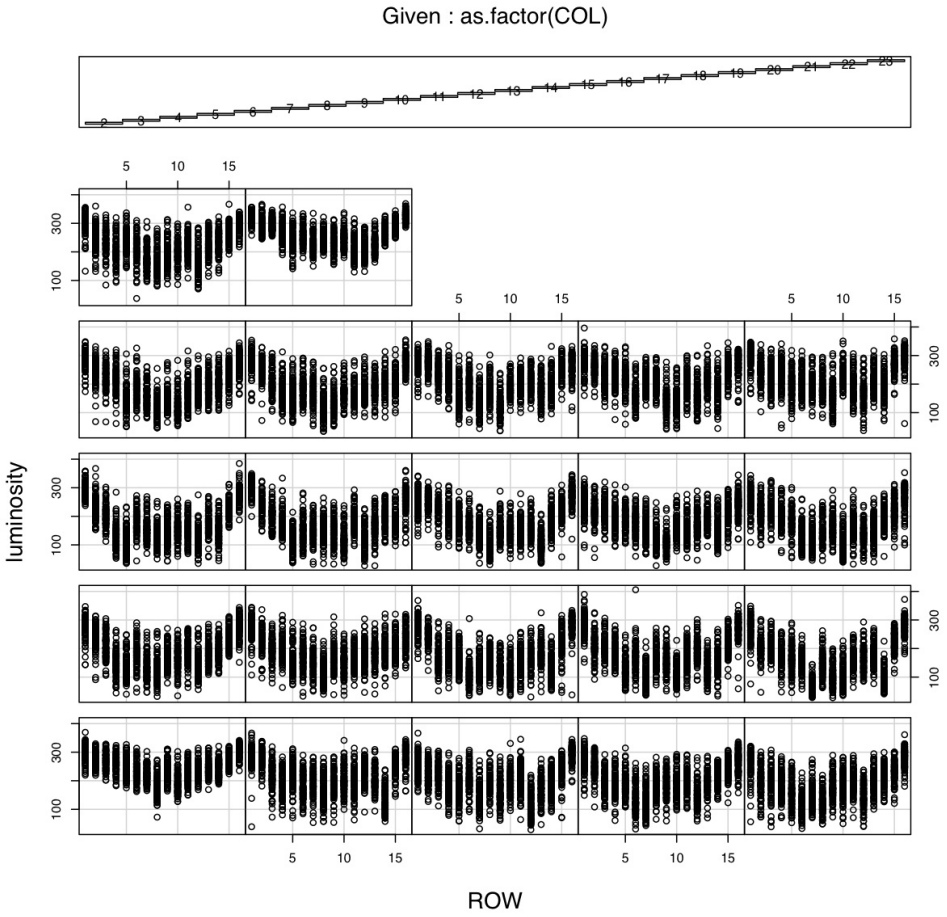


Figure 5: Systematic column effects. Each of the displays corresponds to one of the 16 rows in the plate (ordered from 1 to 16, from left to right, top to bottom). Within each display, the intensity values for all wells in a given row across all plates are plotted against their column number. Recall that column range from 2 to 23, and columns 2 and 23 are occupied by controls.

(a)



(b)

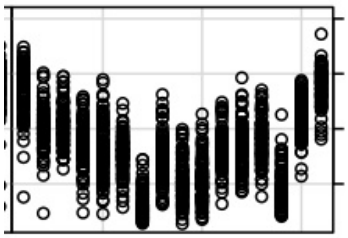


Figure 6: Systematic row effects. (a) Each of the displays corresponds to one of the 22 columns in the plate (ordered from 2 to 23, from left to right, top to bottom). Within each display, the intensity values for all wells in a given column across all plates are plotted against their row number. (b) Enlarged version of the display containing data from column 11.

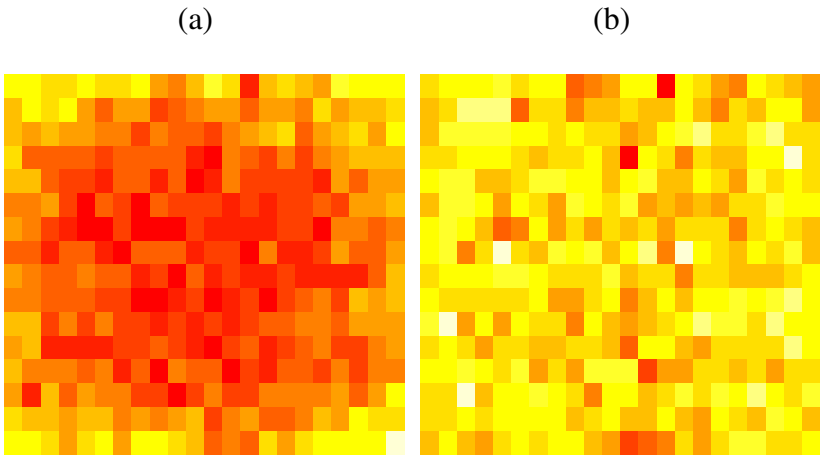


Figure 7: Removal of location effects. Heat-map representation of the intensity values from one plate in the primary screen before (a) and after (b) the removal of location effects. Darker colors correspond to lower values.

an ANOVA model with a plate effect and an interaction term for the column and row factors:

$$s(r_i, c_i, p_i) = m_{p_i} + R_i + C_i + R_i \times C_i.$$

Since outliers were not a substantial problem in the dataset, there was no need to resort to median polish fitting. Figure 7 compares the intensity values of one of the plates prior to and after correction. While we will report elsewhere on the scientific results of our experiment, it is important to note that accounting for these location effects lead to a prioritization of the compounds with higher reproducibility in secondary screens.

The noted variation in intensities has very important implications for plate design. Traditionally, wells in the first and last column are reserved for controls, and this was indeed the design of our primary screen. However, when the background luminosity values vary across locations in a plate, such design is unsatisfactory. For our secondary screen, we designed a plate where controls were distributed uniformly (as in Figure 2). This was particularly important because the screened compounds were selected for their apparent effect. At the same time, we investigated the possible

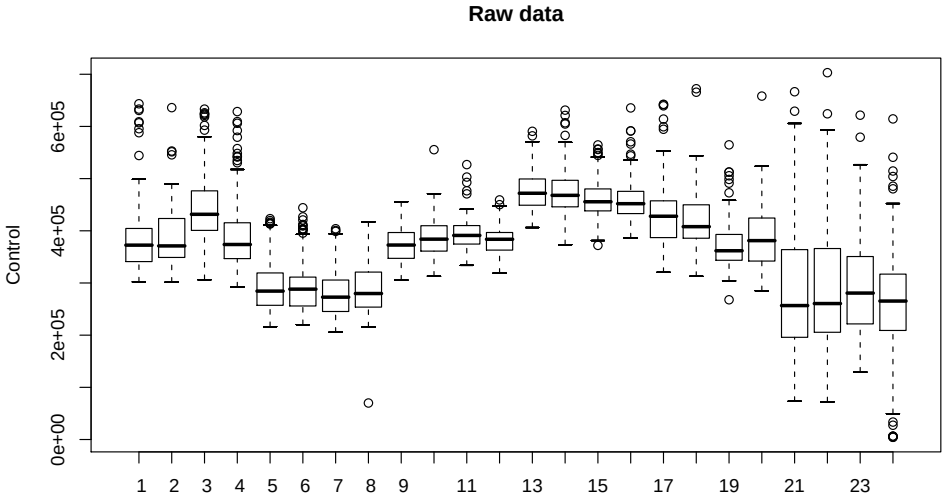


Figure 8: Distributions of control intensities. The 24 plates that constitute the secondary screen are indicated with numbers 1-24, corresponding to those given in Table 1. Corresponding box plots indicate the distribution of intensities of control wells in each of the plates.

biological reason leading to a such pronounced location effect, which has been often attributed to evaporation [].

Turning now to the analysis of the secondary screen, we notice much less prominent location effects (displays like the ones in Figures 6 and 5 did not reveal clear patterns). Instead, a batch effect was evident, with the variability of one batch much higher then the remaining, and average control values significantly different across experimental groups, as it is illustrated in Figure 8. We were able to detect the presence of other systematic variability. Figure 9 displays the correlation values between control well in the same position in each of the 24 plates. It is clearly possible to identify structure in this correlation matrix: replicate experiments tend to have higher correlations, and "blocks" formed by 4 experiments can be identified. These blocks correspond again to the experimental groups identifiable in Figure 8 and hint that location effects are present, and possibly

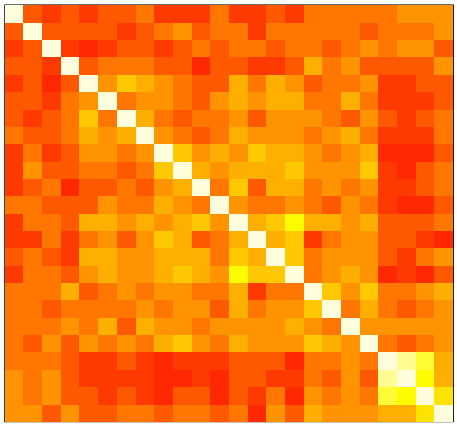


Figure 9: Correlation between control intensities. Heat map display of the correlation matrix of intensities for control wells in the 24 plates. Darker color corresponds to lower value. Each plate is treated as a variable and wells that occupy the same position are treated as same observation.

variable. Note that he last four experiments (21-24) appear in Figure 9 to have higher correlation then the others. One possible explanation for this, is the stronger presence of spatial effects in this last batch, which can be at least partially explained in terms of pin/dispenser effect. Too few cells were prepared for dispensing in plates 21-24, resulting in lower intensities in the well filled last (see Figure 10). This is just one example of the fact that pin effects related to cell dispensing are possible, and they should be taken into account when correcting for systematic errors.

To estimate systematic effects in the secondary screen, we divided the 24 plates in the 6 experimental groups clearly identifiable from Figure 8 and comprising the plates relative to one cell line and one experimental batch only (see table 1). For each of this experimental groups, we estimated a systematic component of the form:

$$s(r_i, c_i, p_i) = m_{p_i} + \text{poly}_{p_i}(r_i, 2) + \text{poly}_{p_i}(c_i, 2) + \text{pin}(r_i) + \text{order}(c_i),$$

where the component $\text{pin}(r_i)$ describes which pin in the cell dispenser filled in the row in question and $\text{order}(c_i)$ describes at what stage the well was filled with cells. Given the limited number of observations from multiple plates, we decided against estimating a well effect. The pin and or-

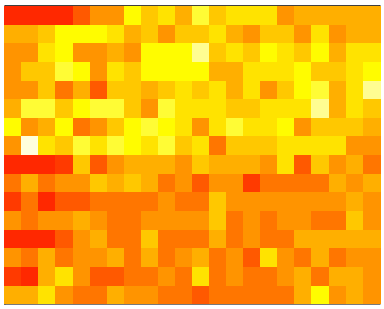


Figure 10: Pin/Order effects. Heat map displays of the intensity values obtained for plate 24. Comparing this display with the diagram in Figure 1, it is possible to see how the last wells to be filled display lower intensity values.

der effects were actually included in the model only for the last experimental group. To estimate $s(r_i, c_i, p_i)$ we used a weighted least square approach on y_i, with higher weight given to intensities of control wells. Recall that in the secondary screen, our controls are spotted uniformly across the plates (see figure 2): we can rely on their values to estimate the parametric form of $s(r_i, c_i, p_i)$. Additionally, experimental compounds have been preselected because they appear to have an effect on cell viability: while the relative strengths of their effects should not be confounded with their location on the plate, their overall mean effect can be expected not to be zero. Because of this reason, the systematic component is estimated using not the original y_i intensities, but the difference between their value and the average intensity value of experimental (or control) compounds in the plate of interest. To clarify this, let us introduce a little notation: with $\bar{y}_{p_i\mathcal{N}}$ we indicate the mean of the y_j values for which $p_j = p_i$ and $k_j = N$, and with $\bar{y}_{p_i\mathcal{E}}$ the mean of the y_j values for which $p_j = p_i$ and $k_j \neq N$. Then, for the purpose of estimating $s(r_i, c_i, p_i)$ with a weighted least square approach, we use $y_i - \bar{y}_{p_i\mathcal{E}}$ for wells containing experimental compounds and $y_i - \bar{y}_{p_i\mathcal{N}}$ for well containing controls.

5 Ranking compounds of interest and tests of significance

We have devoted the previous section to the estimation of systematic effects. In what follows we assume that systematic effects are subtracted from observed intensities, and that we are working with residuals $r_i = y_i - \hat{s}(p_i, r_i, c_i)$. The remaining challenge is that of identification of "interesting" compounds. This assumes different connotations in the context of primary or secondary screens.

In primary screens, typically, only one cell type is studied, and each compound is tested only in one well. The goal is to identify those compounds for which θ_i is substantially different from 0. With one observation per compound, there is really little room for statistical testing. Researchers are looking substantially for a ranking of the compounds and some guidance of when there appear to be no substantial difference between experimental compounds and controls. For this purpose, it is important to keep in mind that different plates can exhibit different dynamic ranges. Typically, $\hat{s}(p_i, r_i, c_i)$ contains an estimate of a mean plate effect, but it may still be appropriate to to correct for different variances across plates. In this context we suggest to rank compounds with respect to the following statistics:

$$t_i = \frac{r_i - \bar{r}_{p_i\mathcal{N}}}{\sigma_{p_i\mathcal{N}}}, \tag{2}$$

where $\sigma^2_{p_i\mathcal{N}} = \sum_{j:p_j=p_i,k_j=N}(r_j - \bar{r}_{p_i\mathcal{N}})^2/|\mathcal{N}|$. This is clearly inspired from classical t-statistics. In absence of replicates for the intensities associated with any experimental compounds, one can gather some information on the experimental variability looking at the variance of controls wells in each plate. Note that because of the removal of systematic effects, the average value of control spots in a plate $\bar{r}_{p_i\mathcal{N}}$ may be practically zero. Moreover, given that the number of control wells per plate is much larger then 1, we have decided to omit consideration of the variability of $\bar{r}_{p_i\mathcal{N}}$ in the denominator of the t_i statistics.

If we assume that intensities for each well have gaussian distributions with identical variances and mean that depend on the compounds, t_i has a gaussian distribution (recall that the number

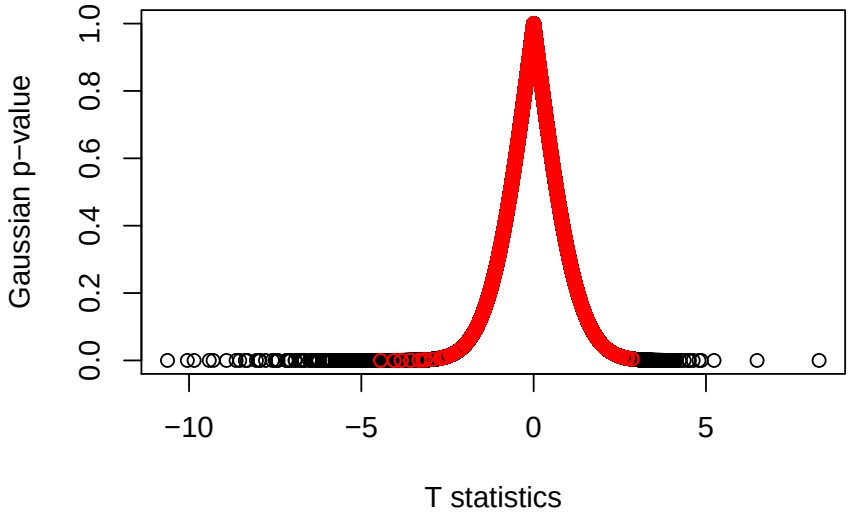


Figure 11: Values of the statistics t_i (2) for all the wells in the first plate of the primary screen. The value of the statistics is on the x-axis, while the p-value of a two-sided gaussian test based on t_i is on the y-axis. Points corresponding to control wells are colored in red.

of control cases is fairly high, so that the difference between t and gaussian distribution become trascurable). Therefore, one could use quantiles of the gaussian distribution to define thresholds of significance. However, we are hardly in the position to justify such assumptions. A more reasonable guideline can be obtained comparing the values of the t_i statistics for control wells with the values of these statistics for experimental wells: the compounds that lead to a t_i statistics smaller in absolute value then the statistic of one of the control wells do not provide strong evidence for an effect (see Figure 11).

Let us now consider the context of secondary screens, or more generically, screens where replicates for experimental compounds are available on different plates. (Of course, it is also possible that plates are designed so as to contain replicates of the same compounds, but this was not the case in the data set we have analyzed so we do not consider such possibility here. It is, however, quite easy to generalize some of the suggestions below). In our secondary screen, the effect of each compound on each cell line was queried with two replicate plates, where the compound occupies the same well. We start focusing on the analysis of the compound effects on the original cell line, 107. The two replicate plates showed a strong correlation, as illustrated in Figure 12, and as to be expected given that they were always part of the same experimental batch. Recall that each compound is queried only in one plate type, so that if we restrict our attention to the observations i for which the compound $k_i = \kappa$, we will obtain two observations, corresponding to the same well in two plates. We will indicate the plates that contain compound κ with π_κ.

Because we have at least one replicate, non parametric tests have bigger resolution power and we have the potential to consider compound (or well) specific variance. Obviously, relying on two observations only to estimate a variance is rather unsatisfactory. At the same time, one has to recall that control wells, for which we still have a large number of replicates, are a good source of information on variability. We considered two approaches: on the one hand, we consider Wicoxon test statistics, and on the other hand, we construct t-like statistics, using a variety of methods to pull information across wells to estimate the compounds variances.

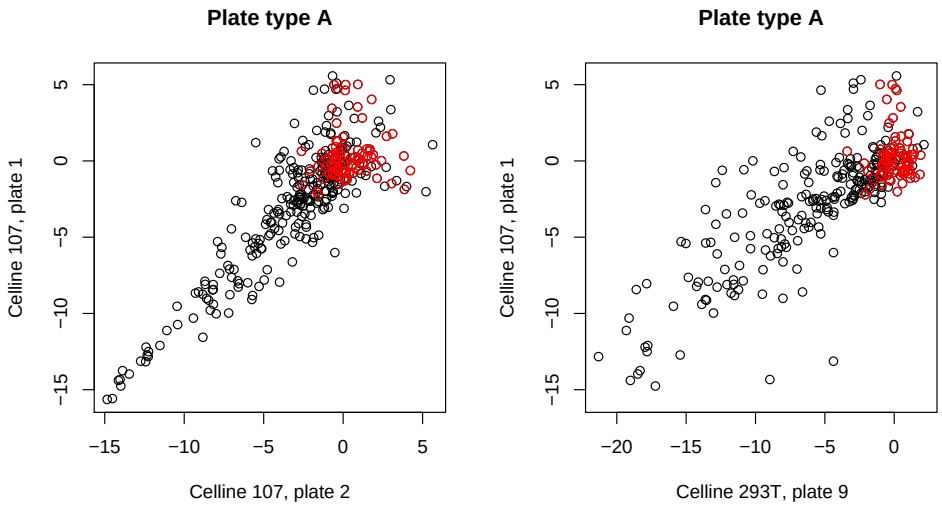


Figure 12: Signal correlation across replicate experiments

For each compound κ, we consider two samples:

$$\mathcal{E}_\kappa = \{r_i, i : k_i = \kappa\} \qquad \mathcal{N}_\kappa = \{r_i, i : k_i = N \quad \& \quad p_i \in \pi_\kappa\},$$

corresponding to observations from the wells containing κ ($\mathcal{E}_\kappa$) and control wells from the corresponding plates ($\mathcal{N}_\kappa$). Let us indicate with W_κ the Wilcoxon rank-sum statistics to test the hypothesis that the two samples $\mathcal{E}_\kappa$ and $\mathcal{N}_\kappa$ come from the same distribution. We can then rank the different compounds on the basis of their W_κ value. Given the non-parametric nature of this test, one can actually rely on their p-values for meaningful evaluations.

One may still be interested in constructing t-like statistics, however, in order to better discriminate between compounds that appear to have some effect. With this goal in mind, we consider different estimators for compound-specific variances. Let d_κ the mean pairwise square difference between intensity associated with compound κ in plates in π_κ: in our dataset, with two replicates, this is simply $d_\kappa = (r_{i,k_i=\kappa,p_i=\pi_\kappa(1)} - r_{i,k_i=\kappa,p_i=\pi_\kappa(2)})^2$. Let m_κ the mean of the intensity values used to define d_κ. Then, $E(d_\kappa) = 2\sigma_\kappa^2$ if σ_κ^2 is the variance of observations for compound κ. With a slight abuse of notation, let ν index all the positions in a plate occupied by control spots, and

treat each such well as if testing for a different compound. Then $s^2_{\mathcal{N}_\kappa} = \frac{\sum d^2}{2|\mathcal{N}|}$ provides an estimate of the control variance. To obtain an estimate of the compound specific σ_κ we can consider linear combinations of d_κ and $s^2_{\mathcal{N}_\kappa}$, as well as linear combinations of d_j, over a set of compounds j that we determine to be similar to κ. Specifically, we considered two strategies. We can group compounds on the basis of their d_κ values, and use all the d_j values in a group to estimate a common variance (s^*_κ). Or we can assume that the variance of the compounds is related to their mean level (s^m_κ). Consider the entire collections of d_j associated to one plate type ($j = 1, \ldots, n$) and let $d_{(j)}$ indicate its ordered values. We divide the n values d_j in R groups of equal size, corresponding to the R quantiles. Let $\ell_\kappa \in \{0, n/R, 2n/R, 3n/R, \ldots, n\}$ be such that $d_{(\ell_\kappa)} < d_\kappa < d_{(\ell_\kappa + n/R)}$. We can then obtain:

$$s^*_\kappa = \frac{\sum_{\ell_\kappa < j < \ell_\kappa + n/R} d^2_{(j)}}{2n/R}$$

We can instead define s^m_κ as the fitted value for d_κ of a spline regression of $\{d_j/2\}$ on the corresponding $\{m_j\}$ values. We can then consider estimators of σ^2_κ that are linear combinations of d_κ,$s^2_{\mathcal{N}_\kappa}$, s^*_κ and s^m_κ. Note that typically one can justify the use of estimators of this form using an empirical Bayes prospective (see for example similar approaches adopted in the analysis of gene expression array data [20, 1, 17]). We used two version of these test statistics:$t^m_\kappa = 0.5 s^2_{\mathcal{N}_\kappa} + 0.5 s^m_\kappa$ and $t^*_\kappa = 0.5 s^2_{\mathcal{N}_\kappa} + 0.5 s^*_\kappa$. Figure 13 reports the scatterplots of p-values of the W_κ statistics versus the values of t^*_κ: it is possible to note the consistent variation of the Wilcoxon p-value with the value of the t-statistics, which suggests that t^*_κ provides a reliable instrument for ranking. Due to the small number of observations in the experimental sample, there are a large number of ties among the small p-values for W_κ; the associated t^*_κ statistics allow further discrimination between the different compounds. Indeed, it is biologically interesting to pursue compounds that have sizeable effect on the cell line growth and it is important to be able to separate these from the compounds that may have a significant but small effect.

Throughout this section we have focused on ranking compounds rather then performing a

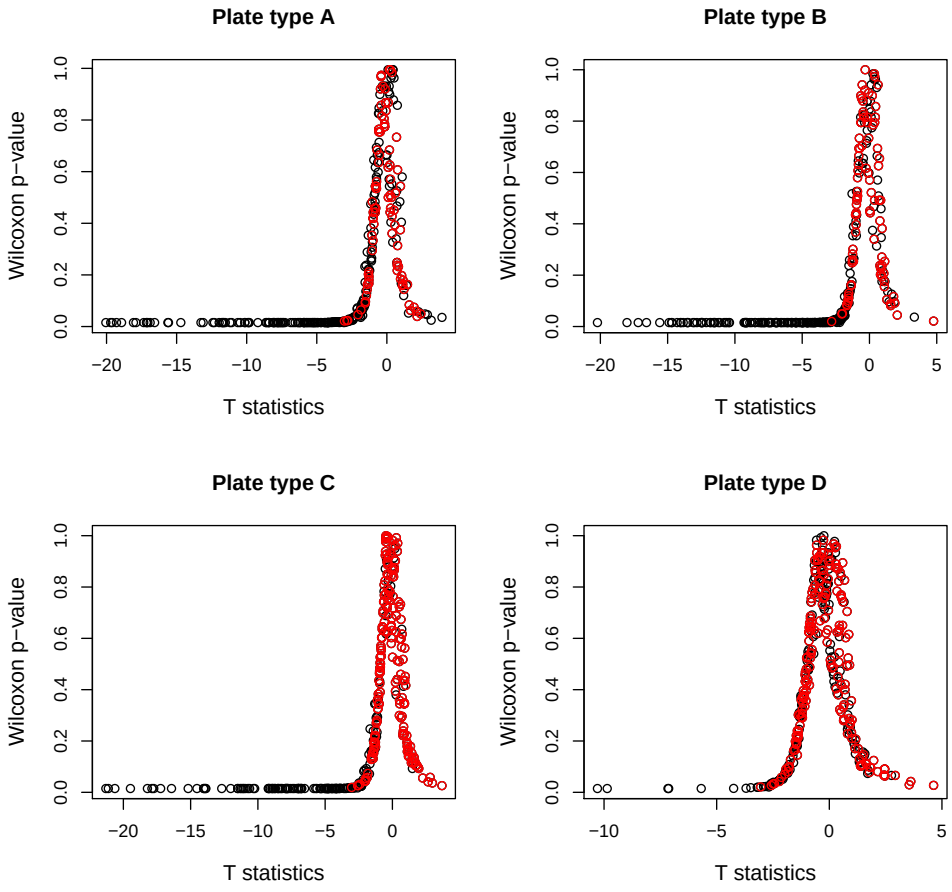


Figure 13: Wilcoxon tests p-values and t statistics for the effects of compounds on cell line 107. Each display focuses on the compounds tested in one of the four plate types. Two side p-values for the W_κ Wilcoxon rank-sum test are plotted against the value of t^*_κ statistics. Points corresponding to controls are indicated in red.

test of hypothesis on their effect. This is both to conform to the actual scientific practice and because, especially in primary screens, we do not actually have enough information to conduct valid test of hypothesis. At the same time, displays as the one presented in Figure 11 provide the researcher with some guidance on what compounds actually appear to have a significantly different effect than controls. Obviously one can approach this problem in a more formal manner and explicitely consider the multiple testing involved. This may be appropriate in secondary screens when replicates are available. To avoid repetitions, we will describe possible strategies with this respect in following sections.

6 Comparing the effects of compounds across different cell lines

In secondary screens one is often interested in studying the effect of selected compounds on multiple cell types. This is, for example, the design of the secondary screen described in Table 1, where compounds that in a primary screen were judged interesting for cell line 107 are additionally screened on cell lines 293T and 1600.

There are multiple questions that one can try to address when looking at the effects of compounds on multiple cell lines. For example, one can ask if if the compounds that have an effect on the primary cell line also have an effect on the other surveyed ones, or if the compounds appear to have an effect at all, in any of the cell lines, or if compounds have differential effects on cell lines. The first question (do compounds that have an effect on primary cell line have also an effect of other cell lines?) can be addressed using the same test statistics we described in the previous section. This is an interesting example of nested test of hypothesis: we will discuss this aspect later, and start by considering, instead, what statistics can be used to investigate the other questions.

To evaluate if a compound has an effect overall or if it has differential effects on cell lines, one needs to compare measurements obtained by experimental plates that show substantially different intensity distributions. This is a situation somewhat analogous to the case of gene expression

arrays, where the need to normalize across arrays has been long noted and addressed (see for example [3]). The correction for systematic effects that we have described in previous section was done within experimental groups that employed the same cell lines, so that it does not address this particular problem. The techniques of quantile-quantile normalization often adopted in gene expression array studies cannot be immediately adapted to secondary screens. These are generally conducted on compounds that are expected to have an effect on one cell line at least and one has to evaluate carefully the assumption that the overall distribution of intensities should be the same across plates containing different cell lines. The availability of a number of control wells in each of the plates comes to the researchers aid: these can be used to calibrate signal intensities and dynamic range across plates.

With this caveat in mind, we describe a few statistics that may aid the researcher in the desired comparisons. Firstly we present a statistics to evaluate the overall effect of a compound. Let us introduce some notation. For a given compound κ and a cell line λ, let π_κ^λ be the set of plates that measure the effect of κ on λ, let $\mathcal{E}_\kappa^\lambda = \{r_i, k_i = \kappa, \ell_i = \lambda\}$ and let $\mathcal{N}_\kappa^\lambda = \{r_i, i : k_i = N, \ell_i = \lambda \quad \& \quad p_i \in \pi_\kappa^\lambda\}$, be the collection of control values on the relevant plates. Assume that F_κ^λ is the distribution from which the observations in $\mathcal{E}_\kappa^\lambda$ were independently generated and let G_κ^λ the analogous distribution for $\mathcal{N}_\kappa^\lambda$. For a given compound κ, we want to test if $F_\kappa^\lambda = G_\kappa^\lambda$ for all cell lines λ. Formally, let $H_0^{\kappa\lambda} := F_\kappa^\lambda = G_\kappa^\lambda$; we are interested in testing $H_0^\kappa = \cap_\lambda H_0^{\kappa\lambda}$. Note that we are allowing the distributions of controls as well as compounds to be different across cell lines: the null hypothesis consists in the equality between control and compound distributions within each cell line. Let us justify this null hypothesis choice. Typically, the different cell lines studied are closely related and one expects a priori that a generic compound would have the same type of effect in all of them. Of course, compounds with differential effects exist and are particularly interesting. Clearly, efforts should be devoted to their identification. However, given the large number of screened compounds, it is important to first single out those that appear to have an effect at all. As one proceeds from secondary to tertiaries and following screens, the focus of the investigation

may shift, but we are here considering a still rather broad and exploratory series of experiments. In this setting, testing H_o^κ rather then separately exploring each $H_0^{\kappa\lambda}$ offers multiple advantages. Firstly, we are able to use a larger number of observations. If a compound has similar effects across cell lines, this translates in an increase our power to detect modest effects. Secondly, reducing the number of tests, and the severity of multiple comparisons, further results in an increase of power.

To test the null H_0^κ we suggest using the van Elteren test [14], which is the weighted average of Wilcoxon rank sum tests for each $H_0^{\kappa\lambda}$:

$$\overline{W}_\kappa = \sum_\lambda \frac{W_\kappa^\lambda}{|\mathcal{E}_\kappa^\lambda| + |\mathcal{N}_\kappa^\lambda| + 1}.$$

In this stratified test, compounds and controls intensities are compared within a cell line group, thereby avoiding the need of "normalization" previously described. The sample sizes in our experiment may not allow one to rely on asymptotic approximation for the distribution of $\overline{W}_\kappa$; evaluating this statistics also on each control well provides an empirical estimator of its distribution under the null hypothesis in our data set which can be used successfully.

We now turn to the problem of identifying compounds with differential effects across cell lines. Firstly, we want to stress that we found that our dataset (with only two replicates for each cell line and a substantial variability of signal across cell lines) was underpowered to identify such compounds. Following is a description of a couple of statistics that appeared reasonable and whose efficacy we are planning to explore in more depth in future datasets.

Again, to take into account the considerable difference in signal distributions across cell lines, we suggest that the comparisons relies on statistics calculated for each cell line separately. Two are natural candidates: the Wilcoxon and t-type statistics that we described in the previous section. Compounds for which these statistics have most divergent values across cell lines are the one that are more likely to have differential effects. For each cell line pair, we hence suggest ranking

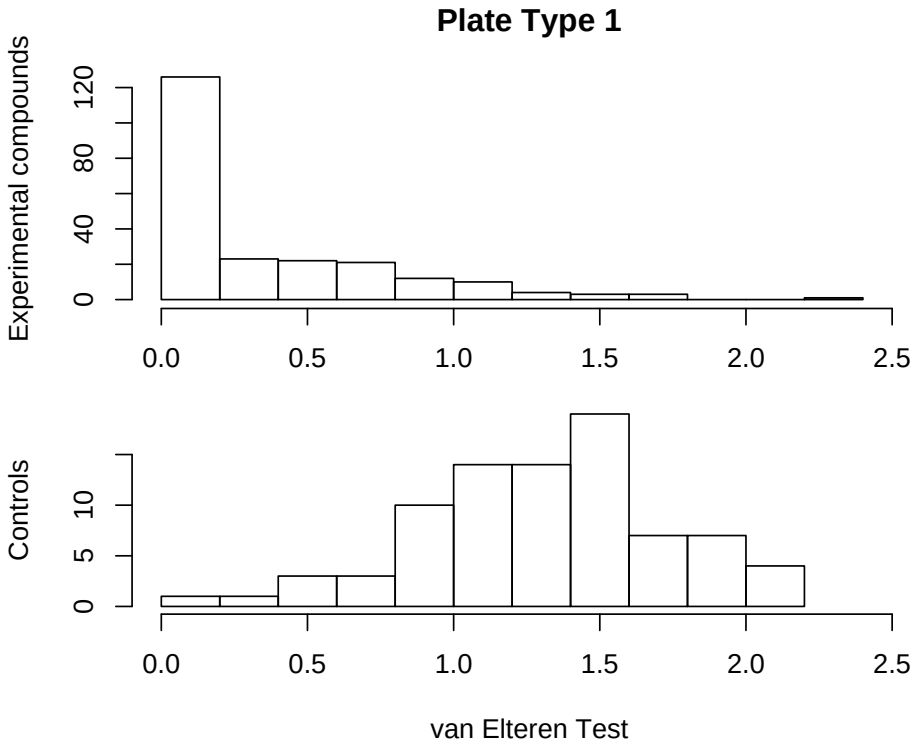


Figure 14: Distribution of the van Elteren test statistics for compounds queried in plate type A. A separate histogram is provided for experimental compounds and control.

compounds on the basis of one of the followings

$$\begin{aligned} Wd_\kappa &= W_\kappa^\lambda - W_\kappa^\delta \\ td_\kappa &= t_\kappa^\lambda - t_\kappa^\delta. \end{aligned}$$

Again, applying this test statistics to control spots provides needed benchmark values.

7 Nested tests of hypothesis

While in the previous sections we have often used a testing framework to clarify what scientific hypothesis we were interested in studying and we have proposed a number of statistics, we have been careful to suggest that these would be primarily used for ranking compounds and providing benchmark values. That is, we have not aspired to declare any result statistically significant. There are multiple reasons for this. Firstly, the nature of some of the described experiments is purely exploratory and a ranking of compounds is precisely what the researchers are aiming for. Secondly, in many cases our information on the experimental mechanism is insufficient to specify the form of the null distribution. Thirdly, a great many comparisons are carried out and if one wants to meaningfully adopt a test of hypothesis framework, one has to take into account this issue, overlooked this far.

While formal test of hypothesis is probably inappropriate for primary screens, it becomes more meaningful in secondary and further follow-up screens: there are both more replicates and the nature of the experiment becomes less exploratory. We now turn to consideration of how to address the problem of multiple comparisons in this context.

Firstly, we suggest that the False Discovery Rate [2] provides a meaningful measure of global error in this type of problems. The usefulness of this criterion in modern high-through put biological investigations has been underscored multiple times, in gene expression as well as genetic

studies [17, 15, 19, 18].

Secondly, we want to emphasize the importance of carefully selecting the groups of hypothesis to be combined and the order with which different hypotheses are explored. In the secondary screen we considered, both a set of potential killers and potential growth enhancers were followed up. Typically the effect of a killer compound is easier to detect then that of an enhancer: exposure to a toxic substance my easily reduce the number of cells to a small fraction, while an analogous fold increase is difficult to achieve. This translates to a different expected number of false null hypothesis in the class of enhancers and killers. Furthermore, we have noted repeatedly, how experiment conducted with different cell lines and in different batches appear to have substantially different signal strength. A recent contribution by Efron [6] underlines how in situations analogous to the described one FDR can be effectively controlled in a stratified fashion, where strata correspond to group of hypothesis that are somewhat inhomogeneous. We believe that this approach is appropriate here.

Furthermore, we believe that the statistical analysis of secondary screens, that include data from multiple cell lines, would substantially benefit from a hierarchical structure in hypothesis testing. As we argued in the last section, it makes little sense to investigate if a compound has a different effect on separate cell lines before establishing that the compound has an effect at all. Conducting a global test to start with allows one to reduce the total number of hypothesis and spend the allowable global error on the test that have some reasonable chance of leading to a discovery. An interesting recent work by Yekutieli [21] describes how one can control FDR in settings were we can identify a hierarchy of hypothesis: this approach has already been successfully applied to a genomic problem [16] and we believe would be relevant in HTS.

Figure 15 provides a graphical illustration of a possible testing plan. Compounds are grouped according to the plate type on which they were tested. This identifies a number of families of tests (three in the figure, and four in our secondary screen data): FDR should be controlled in each of these families, according to the stratified approach proposed by Efron [6]. Within each of the

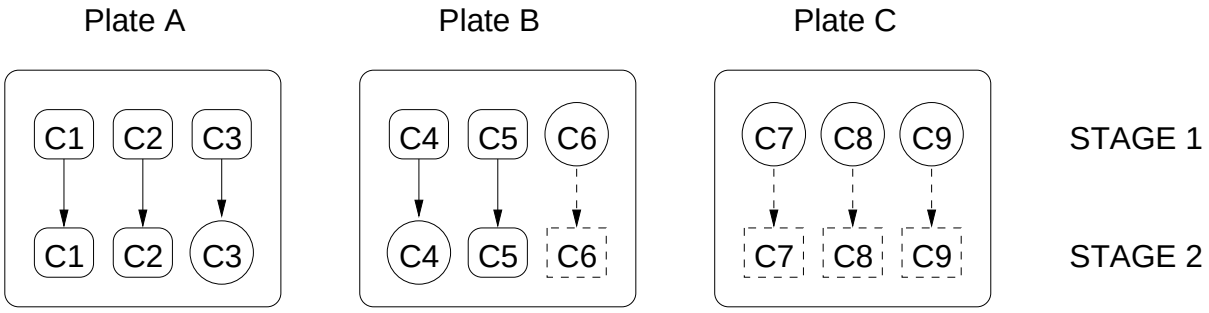


Figure 15: Hierarchical testing strategy.

families of tests, there is a hierarchical structure: initially the effect of each compound is tested on the main cell line of interest, as described in section 5. Only compounds for which the null hypothesis is rejected are then tested for differential effects in one other (or multiple other) cell lines. To control FDR in this hierarchical testing setting, one can use methodology described in [21].

8 Discussion

We provided an introduction to the statistical issues involved in the analysis of HTS. These range from the need of global quality control measures, the estimate and removal of systematic effects, the description of statistics that lead to effective ranking of the studied compounds with respect to property of interests, to the definition of an appropriate framework for the interpretation of the results of thousands of tests. For each stage of analysis, we have provided examples from real datasets, references to the existing literature, illustrated our own novel contributions, and highlighted open problems.

Aknowledgements

C.S. gratefully acknowledges support of NSF grants DMS0239427 and CCF-0326606 and NIH grants RO1 GM53275.

References

[1] Baldi, P. and A. Long (2001) A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized t-Test and Statistical Inferences of Gene Changes, *Bioinformatics*, 17: 509–519.

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

[3] Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19: 185–193.

[4] C Brideau, B Gunter, B Pikounis and A Liaw (2003) Improved Statistical Methods for Hit Selection in High-Throughput Screening, *Journal of Biomolecular Screening* 8: 634–647.

[5] Echeverri, C. and N. Perrimon (2006) High-throughput RNAi screening in cultured cells: a user's guide, *Nature Reviews Genetics* 7: 373–384.

[6] Efron,B. (2008) Simultaneous inference: when should hypothesis testing problems be combined? *Annals of applied statistics*, to appear.

[7] Gagarin A, Makarenkov V, Zentilli P. (2006) Using clustering techniques to improve hit selection in high-throughput screening, *J Biomol Screen* 11: 903–14.

[8] Kevorkov D, Makarenkov V. (2005) Statistical analysis of systematic errors in high-throughput screening, *J Biomol Screen*, 10 :557–67.

[9] Lamb J, E. Crawford, D Peck, J Modell, I Blat, M Wrobel, J Lerner, J Brunet, A Subramanian, K Ross, M Reich, H Hieronymus, G Wei, S Armstrong, S Haggarty, P Clemons, R Wei, S

Carr, E Lander, T Golub (2006) The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease, *Science* 313: 1929–1935.

[10] Makarenkov V, Kevorkov D, Zentilli P, Gagarin A, Malo N, Nadon R. (2006) HTS-Corrector: software for the statistical analysis and correction of experimental high-throughput screening data. *Bioinformatics*. 22: 1408–9.

[11] Makarenkov V, Zentilli P, Kevorkov D, Gagarin A, Malo N, Nadon R. (2007) An efficient method for the detection and elimination of systematic error in high-throughput screening. *Bioinformatics* 23: 1648–57.

[12] `http://mssr.pharmacology.ucla.edu/`

[13] N. Malo, J. Hanley, S. Cerquozzi, J. Pelletier and R. Nadon (2006) Statistical practice in high-throughput screening data analysis, *Nature Biotechnology* 24: 167–175.

[14] Puri, M. (1969) The van Elteren W test and non-null hypothesis, *Review of the International Statistical Institute*, 37: 166–175. Van Elteren P. (1960) On the combination of independent two sample tests of Wilcoxon. Bull. Inst. Int.Stat.37:351.

[15] Reiner A, D. Yekutieli, and Y. Benjamini (2003) Identifying differentially expressed genes using false discovery rate controlling procedures *Bioinformatics* 19: 368–375.

[16] Reiner A., Yekutieli D., Letwin N. E., Elmer G. I., Lee N. H., Kafkafi N., Benjamini Y. (2007) Associating quantitative behavioral traits with gene expression in the brain: searching for diamonds in the hay, *Bioinformatics* 23: 2239–2246.

[17] Sabatti, C., S. Karsten, and D. Geschwind (2002) Thresholding rules for recovering a sparse signal from microarray experiments, *Mathematical Biosciences* 176: 17–34.

[18] Sabatti, C., S. Service, and N. Freimer (2003) False discovery rates in linkage and association linkage genome screens for complex disorders, *Genetics* 164: 829–833.

[19] Storey, J. and R. Tibshirani (2003) Statistical significance for genome-wide studies, *Proceedings of the National Academy of Sciences* 100: 9440–9445.

[20] Tusher, V., R. Tibshirani and Chu (2001) Significance analysis of microarrays applied to the ionizing radiation response, *PNAS* 98: 5116–5121.

[21] Yekutieli, D. (2008) Hierarchical False Discovery Rate controlling methodology, *Journal of the American Statistical Association*, to appear.

[22] Zhang JH, Chung TD, Oldenburg KR. (1999) A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays, *J Biomol Screen*. 4: 67–73.

[23] Zhang XD, Yang XC, Chung N, Gates A, Stec E, Kunapuli P, Holder DJ, Ferrer M, Espeseth AS. (2006) Robust statistical methods for hit selection in RNA interference high-throughput screening experiments, *Pharmacogenomics* 7: 299–309.