Chapter 4

Statistics and Decision Making in High-Throughput Screening

Isabel Coma, Jesus Herranz, and Julio Martin

Abstract

Screening is about making decisions on the modulating activity of one particular compound on a biological system. When a compound testing experiment is repeated under the same conditions or as close to the same conditions as possible, the observed results are never exactly the same, and there is an apparent random and uncontrolled source of variability in the system under study. Nevertheless, randomness is not haphazard. In this context, we can see statistics as the science of decision making under uncertainty. Thus, the usage of statistical tools in the analysis of screening experiments is the right approach to the interpretation of screening data, with the aim of making them meaningful and converting them into valuable information that supports sound decision making.

In the HTS workflow, there are at least three key stages where key decisions have to be made based on experimental data: (1) assay development (i.e. how to assess whether our assay is good enough to be put into screening production for the identification of modulators of the target of interest), (2) HTS campaign process (i.e. monitoring that screening process is performing at the expected quality and assessing possible patterned signs of experimental response that may adversely bias and mislead hit identification) and (3) data analysis of primary HTS data (i.e. flagging which compounds are giving a positive response in the assay, namely hit identification).

In this chapter we will focus on how some statistical tools can help to cope with these three aspects. Assessment of assay quality is reviewed in other chapters, so in **Section** 1 we will briefly make some further considerations. **Section** 2 will review statistical process control, **Section** 3 will cover methodologies for detecting and dealing with HTS patterns and **Section** 4 will describe approaches for statistically guided selection of hits in HTS.

Key words: Automated SDI computational tool, Pattern recognition, Quality assurance, Quality control, Quantitative structure activity relationship, Standard deviation of inactives, Statistical process control, Screening quality control, *ultra*-high-throughput screening.

1. Introduction

Experimentation is naturally and inevitably subject to variability. When an experiment is repeated under the same conditions or as close to the same conditions as possible, the observed results are

W.P. Janzen, P. Bernasconi (eds.), *High Throughput Screening, Methods and Protocols, Second Edition, vol. 565*

DOI 10.1007/978-1-60327-258-2_4, Springerprotocols.com

never exactly the same, and there is an apparent random and uncontrolled source of variability in the system under study. This fluctuation in responses from one experiment to another is known as *experimental variation* or *experimental error*. If something is certain for screeners it is that certainty is false. Nevertheless, "randomness isn't haphazard; it often displays an underlying order that can be quantified, and thus used to advantage" (Charles Annis *(1)*). In this context, we can see statistics as the science of decision making under uncertainty. Screening is about making decisions on how active or inactive one particular compound is against a biological system. Thus, the usage of statistical tools in the analysis of screening experiments is the rational way of correctly interpreting screening data, qualifying them with meaning and hence converting them into valuable information that supports sound decision making.

Repeated observations that differ because of experimental error often vary on a central value in a roughly symmetrical distribution in which small deviations occur more frequently than large ones. Amongst the numerous distributions described in the literature, the *Gaussian* or the *normal* distribution is the simplest and the one that plays the most important role in statistics for experimental sciences. The normal distribution for the frequency of a measure is a symmetrical curve centred around the *mean* and tailing off towards zero in both directions in a shape defined by the experimental error or the spread of data, which is estimated by the *standard deviation*. Experimental data of one particular compound tested in one biological assay protocol follow a normal distribution. When measures do not fit this model, it is an indication of either extreme errors, patterned heterogeneity or anisotropy of the testing unit in the experiments (e.g. non-random variability owing to different state of the compound, the biology and the equipment) or an asymmetrical scale in the measured variable. Furthermore, according to the central limit theorem, the distribution of an average tends to be normal, even when the distribution from which the average is computed is decidedly non-normal *(2)*. Hence, the distribution of any phenomenon under study does not have to be normal because its average will be, and it will have the same mean as the parent distribution.

In the HTS workflow, there are three primary stages where key decisions have to be made based on experimental data:

1. **Assay development**: How good is the assay in terms of reliability, reproducibility and sensitivity for detecting the kind of modulating compounds we are pursuing? Which assay protocol is better? If the assay is not good enough, we will have to invest additional resources to adjust it. If it is, we will put the assay in production, with the consequent implications of resources and data generation. Assessment of assay quality is reviewed in other chapters *(3)*. In **Section 3** we will briefly make some further considerations.

2. **HTS campaign process**: Is the screening process performing at the expected quality? Do all quality indicators stay within the range of acceptance? Are we maintaining the same quality and range of variability across all compounds tested, i.e. all wells within the plate, all plates within a daily run and all runs within the campaign. If the HTS experiment is analysed on the fly, and the assessment is that the process is not behaving properly, we may consider pausing it, invalidating the data generated and troubleshooting the process before resuming production activity. If the analysis is run retrospectively, we will try to identify patterned signs of experimental responses that are not randomly distributed in terms of time (temporal patterns) or space (spatial patterns). Depending on the strength of the pattern and the knowledge of the rule quantifying the underlying order, we may consider correcting or rejecting the data. **Section 3** will review statistical process control, and **Section 4** will cover methodologies for detecting and dealing with HTS patterns.
3. **Data analysis of primary HTS data**: Which compounds are giving a positive response in the assay? Those deemed as *positive* or hits will deserve additional investment and attention in further experiments. On the other hand, those graded as *negative* will be abandoned, probably forever. **Section 5** will describe approaches for statistically guided selection of hits in HTS.

In this chapter we will focus on how some statistical tools can help to cope with these three aspects, which are more closely related to primary HTS. Needless to say they do not comprehensively cover all the plethora of statistical issues that can be found by a screening scientist, such as design of experiments in assay optimisation (i.e. how to reduce the number of experiments while improving interpretation and decision making) and assessment of quality and reliability of dose–response screens in QSAR (i.e. how precisely and accurately can one model QSAR from screens run in different experiments over time, with different platforms or laboratories or by different scientists? Do all assays correlate or agree?). *See* Refs. *(4–6)* for a review of these subjects.

2. Assessment of Assay Quality

A review on statistical evaluation of assay quality is included in an earlier chapter *(3)*, where the reader can find a description of the most widely used parameters. Some of these and other alternative

ones have been critically reviewed elsewhere *(4, 7, 8)*. As a rule of thumb, parameters that do not contain any information regarding data variation are less appropriate in evaluating assay quality. Since the publication of Z-prime parameter *(9)*, this has become the cornerstone of assay plate quality control and assay performance adopted by screening scientists.

$$Z' = 1 - 3 \times \left(SD_{signal} + SD_{background}\right) / \left|M_{signal} - M_{background}\right|$$

where SD is the standard deviation, *M* is the mean, signal is the control of response for an active assay (100% assay activity), and background is the control of response for an inactive assay (0% assay activity).

The attractiveness of Z-prime lies in the following: (1) it combines the two key quantitative features of an assay, namely signal amplitude (i.e. distance from the mean of signal controls to the mean of background controls) and variability (i.e. standard deviation of data); (2) it is dimensionless, so it can be used as a direct comparator across assays regardless of assay modality; (3) ease of calculation; and (4) it is a relative indicator of the power of the assay to discriminate active and inactive compounds, so it directly correlates with the minimum statistically significant threshold of activity in order for a compound to be deemed *hit* (i.e. hit cut-off).

Nevertheless, there are several limitations to the use of Z-prime as a single quality indicator in the assessment of assay quality and performance:

(1) Since it is asymptotic, it is scarcely sensitive to changes in assays with high values close to 1. That limits the comparison of assays and protocols as well as the process control of screens in production. In these cases, we would recommend the use of other supplementary ndicators, such as signal to background (though this does not account for data variability) and signal to noise *(3)*.

(2) It deals with only control samples and not compound samples. Since controls in microtitre plates are usually arrayed in a small number of fixed positions, Z-prime does not account for the variability of other positions of the plate. In other words, by using Z-prime, we are blind to errors or source of variability beyond control wells. One possibility is interspersing plates fully filled with "signal" or "background" controls in separate plates or randomly distributed in the same plate. Alternatively, in these cases it is recommended to use Z parameter (note: Z is calculated as Z-prime but signal being referred to actual samples or compounds instead of controls) or other quality indicator looking at samples. In this chapter we will describe tools to identify sequential or spatial patterns in an automated manner.

(3) By itself, Z-prime is not an indication of the robustness of the assay in the sense of reproducibility of the assay performance across days, scientists, equipments, labs, reagent batches, compounds, etc. Actually, poor correlation has been found between Z-prime and confirmation rate for real screening campaigns *(10)*. For this purpose, evolution of Z-prime and other quality indicators for signal window and variability can be monitored. Even so, high Z-prime values do not guarantee a reproducible percentage of response at single shot or XC_{50} values in dose–response experiments. Thus, more adequate statistical tools have been developed for the proper assessment of reproducibility and repeatability of single shot (e.g. precision radius and accompanying statistical parameters *(9)*) or correlation and agreement of dose–response screens (e.g. B-score and R-score *(2, 11–13)*) or correlation and agreement of dose–response screens (e.g. MSR, minimum significant ratio *(4, 6)*).

a. The precision radius statistical analysis is based on the premise that HTS assays contain two distinct sample populations, i.e. the population of inactive samples and the population of outlier or active samples. The former is modelled by normal distribution and highlights the assay noise, whereas the latter is modelled by an extreme value distribution function (a sigmoid-like Gumbel distribution) and highlights the detection power of the assay. Both populations overlap and show convergence in distribution. Precision radius statistics can also be employed for the setting of hit cut-off thresholds, as will be described in another section.

b. MSR (minimum significant ratio) is described as the smallest potency ratio between any pair of compounds that is real, i.e. statistically significant. It is calculated as follows:

$$\mathrm{MSR} = 10^{2\sqrt{2}\times \mathrm{SD}}$$

where SD is the standard deviation of the replicate XC_{50} results from a test–retest experiment or historical QC data.

A last consideration is that we should bear in mind that the ultimate goal of HTS is to identify novel compounds. The higher the sensitivity of the assay to pick up modulators of weak potency, the better the screen. The quality of the assay must be subordinated to this aim. In this regard, optimising quality indicators of an assay based on configuring the assay in a way that is insensitive to small fluctuations in the levels of the response that is measured (e.g. high conversion rate of substrates in enzymatic assays *(14)*) might lead to a very robust but useless screen.

3. Statistical Process Control

High-throughput screening (HTS) has undergone critical changes in the past decade. These changes have covered all aspects of the HTS business from compound management to the production and evaluation of hits *(12, 15)*. The business is immersed in the so-called "ultra-high-throughput" (uHTS) era. Nowadays compound libraries typically surpass 1.5 million compounds available for diversity screening. Biochemical and cellular assays (screens) are carried out in high-density plates containing reactions from a few micro- to nanolitres per well. Assay plates are processed at a large scale by robots or workstation platforms. The quantity and speed of data production have increased the benchmark values of the 1990s by more than 20-fold. A typical day of HTS operation provides more than 100,000 data points. Such volumes of data need to be properly managed, stored and analysed.

As this in-depth change in "industrial" data production has settled down, an important requirement has emerged more strongly than ever. "Cost-efficient" management of the HTS processes is vitally demanded. The new uHTS systems need to minimise waste, rework, cycle time as well as the likelihood that poor-quality data may be passed on to the customers. Organisations have tried to solve the problem by seeking and adapting traditional quality strategies including quality control and quality assurance methods. The result is the "screening quality" culture in which "screening quality control" forms the core.

The culture of **statistical process control (SPC)** was pioneered by Walter A. Shewhart and W. Edwards Deming. With its emphasis on early detection and prevention of problems, SPC has a distinct advantage over quality methods, such as inspection, that apply resources to detecting and correcting problems in the end product or service. Events or occurrences are the result of an input leading to a certain output, following a process. The fundamental assumption from which theory of quality control is explained stands on the differentiation of causes of variation for any system or process. In general, any quality control system will use a variety of tools to detect and minimise assignable variability to a given process. It includes many procedures such as preventive maintenance, instrument function checks and validation tests *(16)*.

Statistical quality control is a term used to describe the aspects of a control system in which statistics are applied to determine whether observed measurements fall within the range expected due to random variation of the process. Much of SPC power lies in the ability to monitor both process centre and its variation around that centre. By collecting data from samples at various points within the process, variations in the process that may affect the quality of the end product or service can be detected and corrected *(16)*.

3.1. Design and Implementation of Statistical Quality Control Systems for uHTS Operation: The Complexity of Pseudo-Manufacturing Processes

One of the principal difficulties of implementing quality control practices in uHTS processes lies in their complexity. uHTS processes include multiple parameters potentially acting on the quality of the final output (data). For instance, an anomaly during any HTS run may be due to mechanical reasons (e.g. liquid handling, clogged tips, temperature gradients and plate storage conditions) or biological reasons (e.g. wrong enzyme concentration and faulty batch of cells). The anomaly may be in an isolated plate or may be more pervasive, indicative of a more serious systematic pattern affecting a run.

On the other hand, every screen or production line may have unique quality matters related to particular aspects (e.g. molecular mechanism, stability of reagents, additives, cofactors and tool compounds). Besides these factors, the variety of assay formats currently available requires that every line needs specialised people who are familiar with these requirements. In this context, the term controller can be used to refer to the scientist or the team of scientists in charge of analysis and problem solving. Screen controllers must have an in-depth knowledge not only of biochemical aspects of the screen but also of assay technologies, automation, statistical and data handling areas.

Another point of difficulty lies in the cultural challenges that people working on this field need to absorb. The way to implement typical "manufacturing" values such as supply chain working or 6σ assumptions has to be adapted for HTS teams that normally belong to discovery research organisations. There are underlying cultural barriers to excessive metrics, steps, measurements and to the intense focus on reducing variability that is seen to water down the discovery process. Striking the right balance between the application of statistical quality control and the unencumbered research is a key issue.

Moreover, there are other factors to consider when adapting statistical quality control to uHTS processes, for instance, reliability of instrumentation and design of the HTS production lines. Reliability can be defined as the probability that a device will perform its intended function during a specified period of time under stated conditions *(17)*. In general, company suppliers of uHTS equipment tend to put more emphasis on innovation and enhanced functionality than on device reliability. This situation adds another level of complexity to the HTS statistical quality control systems. Production lines comprise a plethora of instruments with different levels of reliability and automation.

Setting appropriate error limits is actually a general principle in quality management *(16)*. Cost–benefit implications of type I errors (probability of a false error rejection) or type II errors (probability of not rejecting an error) must be optimised when setting quality limits in order to guarantee the overall quality of the

product at the minimum cost. A balanced mixture of absolute and flexible quality limits is required in order to cover the variety of screens, assay formats and production currently available. It is also important to ensure enough room for controllers to make final decisions about passing or failing results. At the same time there must exist a global frame of quality business rules in order to assure standards of data quality to customers.

Finally, the ability to monitor lines in real time vs. offline and the ability to monitor quality of production units (plates) vs. production rows are key elements to consider in efficient quality control systems.

3.2. Efficient Quality Control Systems in uHTS

Several quality HTS quality control systems have been described in recent years.

Gunter et al. *(12)* described a custom-developed software application for HTS quality control called *StatServer® HTS System* built up from the commercially available S-PLUS® and StatServer®. The system comprises a series of charts for the analysis of shifts and trends in data. Different tools of analysis are described like the graphical analysis of plate centre (i.e. median raw data for samples) vs. plate sequence, high- and low-control averages (i.e. raw values) vs. plate order and robust plate coefficient of variation (i.e. by using a smoothing statistical algorithm) vs. plate. There was a second group of QC plots developed in order to address positional effects like box plots of row/column effects and plate maps with appropriate colour scales. The entire software interaction process includes manual uploading of data and offline analysis of results. Flexibility for plotting and downloading was also provided.

The system described by Padmanabha et al. *(15)* implies a more holistic approach to HTS quality control. They emphasised the importance of quality in all processes around HTS including compound management. Their quality system depicts results in real time through a variety of control charts for controls (e.g. raw data values and Z′) and samples (e.g. mean of raw plate values). Several other analyses are carried out, such as well average of activity across plates, in order to track anomalous trends. Once the screen is completed, the data undergo a process of manual cleanup prior to hit selection. They finally point out the importance of real-time data analysis and processing in order to quickly mine HTS results to build models that rank or select compounds for further testing and early value to the HTS results.

Gribbon et al. *(10)* nicely illustrated the value of screen monitoring beyond Z′ and variability of controls. They proposed to explore errors in whole assay plates and across them as temporal or spatial plate effects. Emphasis was focused on visualisation of large data sets in two-dimensional plate-based formats (plate maps) in order to recognise trends on plates and within the runs. A deeper

review about pattern recognition and correction in HTS can be found in the next section. Certain assay methodologies, such as fluorescence polarisation and TR-FRET, provide ratiometric read-outs that can eventually be used for in-well QC. They also described a general workflow to give a global coherence to the elements of QC that they proposed.

The wave of quality control systems for HTS has been empowering tools for analysis and user's interface capabilities. In the next section we are describing the quality control system developed in GlaxoSmithKline (*Screening Quality Control*, i.e. *SQC software*). It comprises basic elements for quality control and provides new and differentiated key features with regard to existing systems including quality limits, business rules and a broad spectra of analytical tools covering full relevant aspects to monitor uHTS processes.

Figure 4.1 describes the SQC framework design and the principal components that comprise the system. Both intra-plate and inter-plate data are the system inputs. There are analytical tools and business rule algorithms that are applied automatically in order to provide corresponding intra- and inter-plate qualifiers or system outputs. The system operates in two temporal sessions including different algorithms for analysis: online and offline sessions.

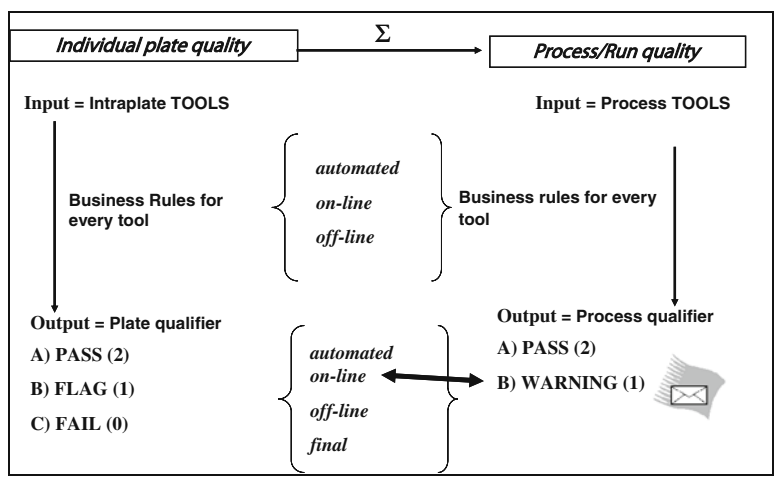


Fig. 4.1. Screening quality control software (SQC) framework design.

Online QC attempts to identify unusual plate behaviour based on the likelihood of unusual occurrences or on expected behaviour of plate parameters. The expected behaviour of parameters is initially captured during the HTS validation stage *(3, 18)*. The system provides online dynamic calculations in order to correct potential drift throughout the life of the screen. The online QC process is fully automated, with no intervention from the

controller unless a problem is identified and the controller is notified by e-mail alert. For that to happen successfully, business rules have been defined accordingly.

In some cases the delay between plate preparation and final reading makes online QC impossible, but in any event the same tools and additional ones made available after the run should be applied to ensure quality HTS post-run (offline QC).

Analytical tools have been designed to evaluate different aspects of screening experiment quality and are used to define the business rules. Such business rules trigger qualifiers for individual plates (intra-plate: pass, flag or fail) and group of plates in a run (inter-plate: ok or process out-of-control).

The qualifiers generated during online QC are saved and fed into the post-run analysis. During the post-run analysis, the controller will have the opportunity to review the information generated. He/she will validate the production line every day and will look for any factor affecting data quality, as in the case of process problems not caught in real time. The controller will finally fail or pass plates according to general guidelines, screen specific limits and context of the run in question. The original qualifiers themselves cannot be removed from the record. This enables business rule sensitivity to be reviewed across screens.

The qualifiers also provide easily accessible QC metrics for process improvement data mining. They serve as a basis for subsequent planning of quality assurance (QA) guidelines focused on increasing productivity by reducing the rate of errors and rework.

In order to avoid manual intervention by controllers to identify and remove random outliers within the replicated controls and standards, algorithms based on robust statistics [*see* **Section 5** and *(19)*] are employed to calculate mean and standard deviation for each set of controls and standards.

The graphical users interface (GUI) of the software provides tabular and graphical information. The tabular reports are updated in real time including information about individual plates. They also provide information on process status, along with a description of the process flags. Graphical information includes a display of all the charts being used, also updated in real time. The data generated in every tool with the limits used in every rule are also available for graphical and tabular display on demand. There are secondary reports, accessed by controller on demand, including plate maps and particular rule details. *See* **Fig. 4.2** for details of the user's interface.

3.2.1. Intra-Plate QC

Intra-plate QC parameters are calculated for each plate in the assay using raw and normalised (i.e. percent inhibition or activation) data for controls, samples and standard compounds.

Several intra-plate QC tools are assessed so that plates are classified accordingly as *passed*, *failed* or just *flagged*.

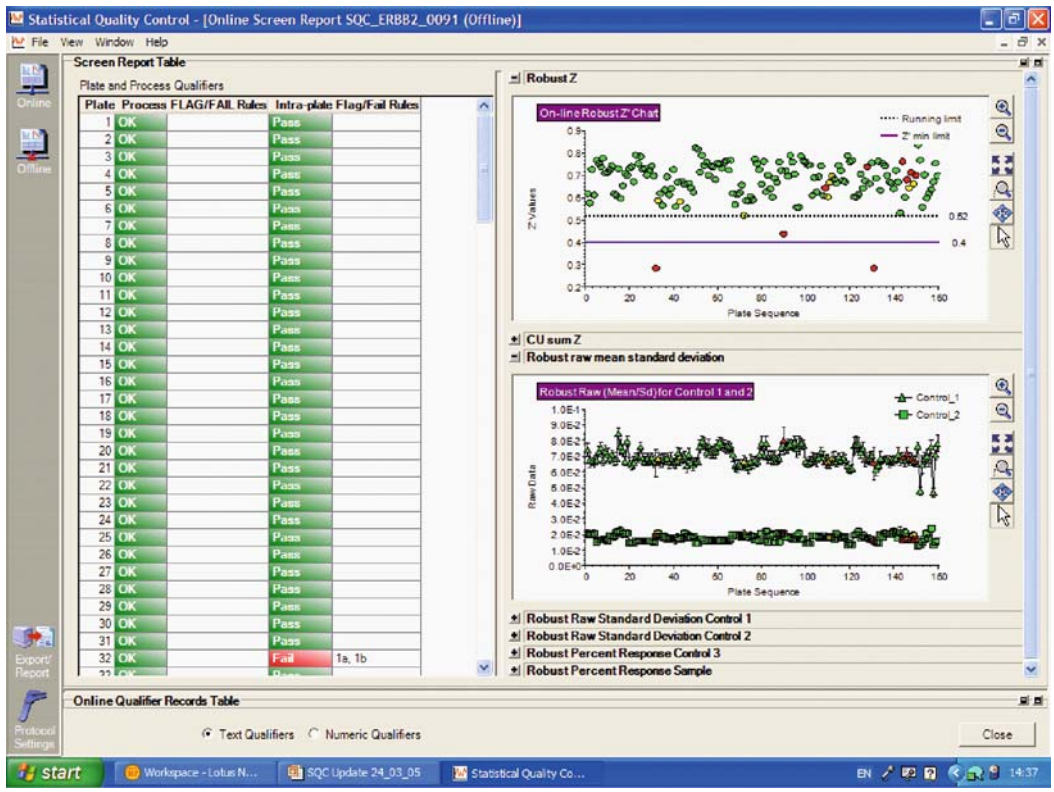


Fig. 4.2. **Screening quality control software (SQC) graphical user's interface**. Screenshot of tabular and graphical information.

Z′ is a measure of the separation between the distribution tails of the low and high controls in an assay *(9)*. An adequate value of Z′ is necessary to ensure that the assay has the ability to distinguish between inactive controls and compounds with a certain degree of biological activity.

Our goal in GSK is to run assays with average Z′ greater than 0.6. If after reasonable optimisation trials an assay does not reach this average, we can accept $Z' < 0.6$.

There are algorithms to determine various Z′ limits including Z′ absolute minimum and running Z′ at 2 and 3 σ levels for both online and offline sessions.

The location of control outliers as detected by the robust statistical algorithms is recorded whenever the bias is greater than 10% from the average.

Keeping assay sensitivity constant along screen rows and across screen runs is a key metric to monitor in uHTS. The standard is defined as a stable compound (available in sufficient quantity) that helps monitor the biological relevance of reagents. When used for QC purposes, it is added at XC50 concentration. The mean of

normalised response for standards is monitored by a Shewhart control chart with limits (±3 SD) based on the estimates of mean and SD from HTS validation runs.

Sample means of normalised data are also monitored to identify factors affecting only the samples. Although any plate with an unusual sample mean will be flagged, a single unusual plate may not be a cause for concern. Therefore, action is recommended only after persistent behaviour is observed. Limits are defined by mean and SD from HTS validation runs.

Plates with a high hit rate are also denoted by the system. Although hits may be clustered in certain plates, most plates should not give exceptional hit rates. After every run, cut-off is calculated as mean samples + 3 SD samples (run). The business rule for high hit rate plates is based on hit rate values from screen validation.

Plates with unusually high variability of samples (e.g. tail of negative or positive results) are analysed and flagged offline. Threshold is set at a certain level of significance according to a normal distribution.

See **Fig. 4.3** for details on the software display for the record of samples in the run distribution tails. When the proportion of samples in one tail of the distribution run consistently appears over limit, this algorithm is very powerful to address systematic errors in plate areas. Please refer to the next section on systematic error detection in online mode and to the next chapter for a deeper analysis of patterns in HTS.

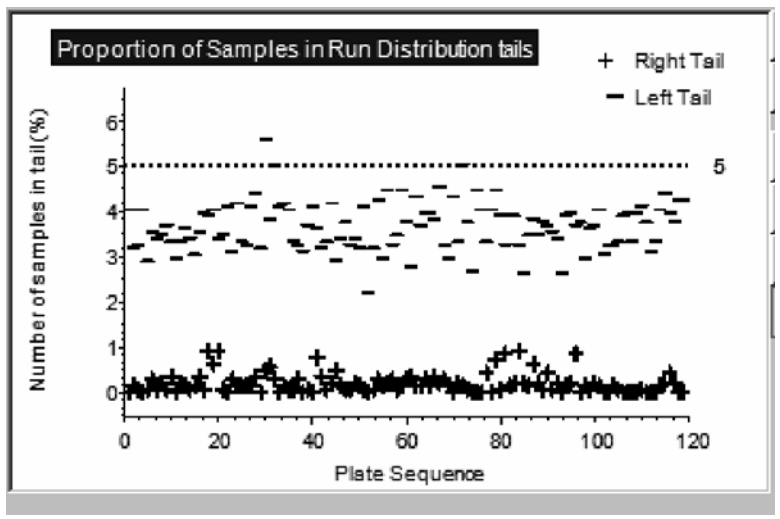


Fig. 4.3. **Graphical display for proportion of samples in run distribution tails (offline)**. The example shows a real case where there is a bias to the left tail of the distribution run for most of the plates in a process line. Although many plates fall below limit, the graph warns about a potential pattern in assay plates (negative values). This fact must be investigated by the inspection of individual plate maps and the incidence of other tools as systematic error.

3.2.2. Inter-Plate QC (Process)

Random data failure is normal in HTS experiments. Such plates are flagged as "failed" during the data acquisition process and may be subsequently retested. Failed plate data are excluded from all subsequent analyses.

Repeated, non-random occurrences of plate failure, on the other hand, indicate a systematic problem with the process, which could be the result of a biological, mechanical, software or human error.

This component of the system makes use of tools that seek patterns and trends within runs and alert for possible systematic errors in the run resulting in a process flag.

Unlike intra-plate QC, inter-plate QC does not result in decisions to pass or fail a plate. Instead, it generates warning flags to inform the controller that something unusual is happening in the run. Several inter-plate QC tools are defined.

Trend of robust Z′ per plate is monitored by a CuSum chart to detect consistent decreasing trends in the Z′ values. A "process out-of-control" message appears if the drift in Z′ falls below the limit of the CuSum control chart. Upon investigation of the cause of the problem, the controller may (i) identify and correct the problem (the CuSum will be reset as if it were the beginning of the run), (ii) be unable to identify or correct the problem and continue or (iii) be unable to identify or correct the problem and stop.

The online software keeps track of flag or fail qualifiers on individual plates and will trigger alarms according to the number and magnitude of errors detected. For instance, four consecutive "fail" or "flag" plates or five plates within any segment of ten plates trigger a "process out-of-control" online alarm.

Systematic errors detected in real time can result in substantial savings by avoiding waste of time and reagents. As a standard practice, patterns identified as systematic errors will trigger process flags and alert the controller of potential problems with the screen that must be investigated.

Three types of systematic errors are detected and flagged:

- Well level: Wells are flagged when they consistently give the highest or lowest responses that are significantly different from the rest of the wells.
- Row level: A row is flagged if it consistently gives the highest or lowest average value that is also significantly different from the other row averages.
- Column level: A column is flagged if it consistently gives the highest or lowest average value that is also significantly different from the other column averages.

See **Fig. 4.4** for details on the software display for systematic error tool. This is a key quality indicator in production lines. Thanks to this algorithm, controllers can explore the magnitude and possible causes of any systematic error in real time.

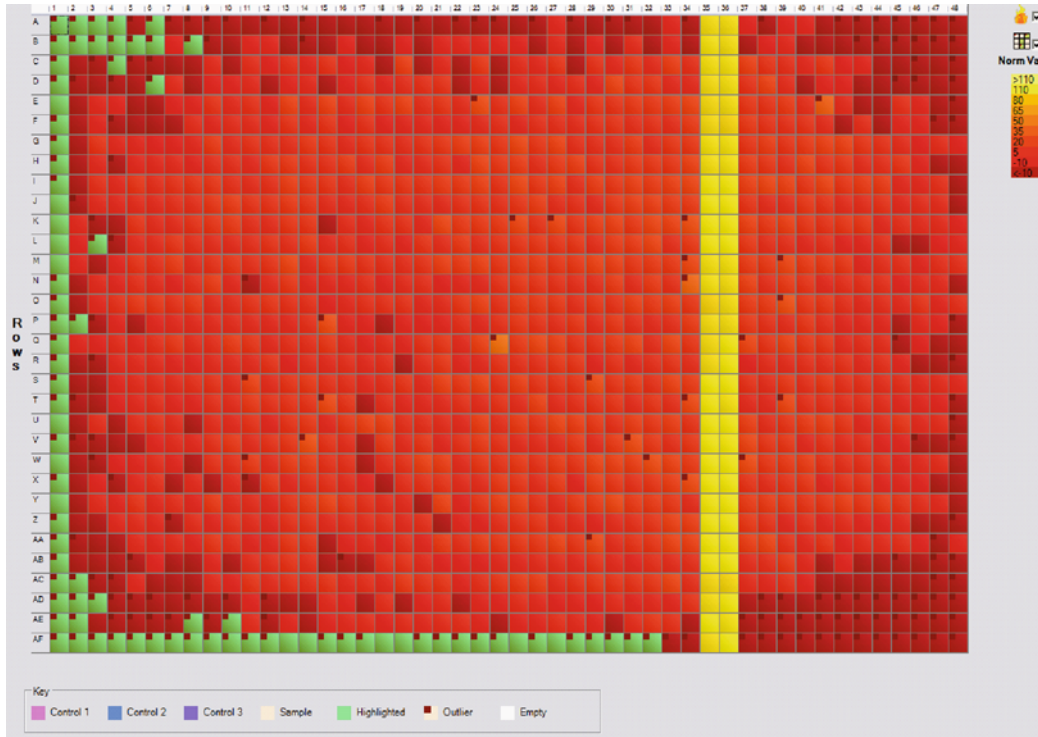


Fig. 4.4. **SQC display of plate maps and systematic error wells**. Wells are coloured according to actual activity. Systematic error wells are depicted *highlighted*.

False positives and false negatives in screen rows are also estimated from control populations. Signal and background control wells are defined as false positives or false negatives if they exhibit a normalised effect greater or lower than a threshold. This gives an idea of how many inactive samples appear to be active and how many active samples appear to be inactive in every run. Runs are flagged according to different criteria.

Observation of false-positive and -negative rates in production lines is another key quality indicator of the actual performance of the HTS processes. This tool is particularly powerful when it is related to run hit rate.

3.3. Summary

HTS quality control systems need special working frameworks. We have described key elements to consider in efficient uHTS quality control systems. The screening quality control system described contains all the requirements that we consider are of key importance in statistical process controlling for HTS. It monitors a very wide range of relevant statistical aspects, which have been adapted from classical SPC to HTS, including tendencies and systematic pipetting errors. Thanks to the development of a core set of business rules, the software automatically audits and grades screening

plates and runs. The system also allows controllers to input particular limits to certain rules according to actual screen performance. By measuring online quality levels, the controller has the opportunity of pausing the run and quickly correcting an obvious problem or stopping the run altogether in an effort to save reagents and/or compounds until the problem is solved. Controllers have the ability to make final decisions on pass/fail criteria according to the current cost–benefit requirements of particular production lines.

4. Detecting and Dealing with Patterns in HTS

When designing an HTS experiment, we aim to attain identical experimental conditions in every well, with compound identity or its concentration as single variables. As discussed earlier, we expect experimental errors but if our quest for uniformity were successful, errors should be exclusively of random nature. However, we often see spatial and temporal systematic errors, because homogenous experimental conditions are very difficult to maintain throughout the plate, and also in the course of time. A full HTS campaign usually takes several days or weeks. Often, the activity data of the compounds increase or decrease at the edges of the plates. This is normally referred to as systematic errors or spatial patterns. At other times, we observe temporal systematic errors when the errors are repeated in the same position in consecutive plates. A visual inspection of plate data can help to detect these patterns if they are present, but we need objective measures to evaluate the importance of these patterns and to deal with them. From a statistical point of view, these two kinds of systematic errors are treated as different problems, and different statistical techniques need to be applied in order to detect each one *(10, 12)*.

4.1. Spatial Patterns

HTS experiments use automation that can contribute to the occurrence of spatial patterns in the HTS and also to the difficulty in maintaining constant experimental conditions in the course of time. The reasons why spatial patterns arise can vary greatly, and although it is not the purpose of this chapter to examine these reasons, we can put forward some of the most frequent reasons: evaporation, liquid handling errors, gradual loss of reagent's activity, temperature gradients, reader calibrations, physical handling of plates, freezing, centrifugation, etc. (**Fig. 4.5**).

The presence of patterns has consequences in the selection of active compounds. Some active compounds can be misclassified as inactive (the so-called false negatives) and some of the inactive compounds can be misclassified as active (the so-called false

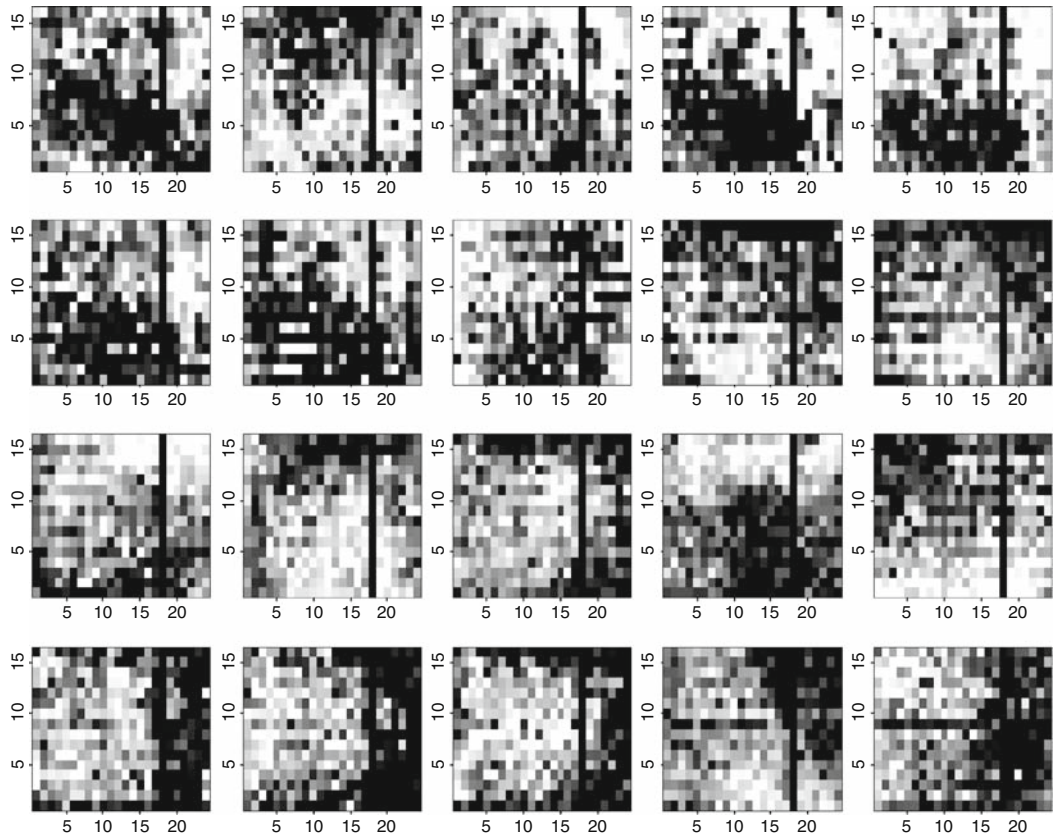


Fig. 4.5. **Different examples of spatial patterns found in consecutive plates in a HTS screen**. Grey colour scale range from –20% activity (*white*) to 30% activity (*black*).

positives). When the HTS campaign is complete, and we add up the total number of hits per row and column, we can assess how the presence of spatial patterns is affecting the hit selection. In these situations, the number of hits in the rows or columns located around the edges may be higher or lower than expected. This usually happens gradually, depending on the distance from the edges *(10, 12)*.

The experimental conditions of each HTS are very different, and in consequence, the patterns generated are also different. But as an HTS campaign progresses, these conditions can change. Therefore, a very flexible and robust statistical method is necessary, which is able to analyse the great variety of spatial patterns that can be found in the hundreds or thousands of plates run in a whole HTS campaign. Also, in most of the cases, HTS data analysis methods are run automatically, and we have only a general perspective of the effects of applying a data correction method, because we cannot analyse all the plates generated in an HTS campaign in detail.

Although visual inspection has proven to be a simple way of recognising spatial and temporal patterns, there exist software packages that can better cope with pattern recognition and

correction for complex and massive data sets. Some of these are public domain [e.g. HTS Corrector *(20, 21)*] or commercially available (e.g. Paratek and Genedata), whereas others are developed by companies for their internal use [e.g. StatServer HTS *(12, 13)*].

Pattern recognition methods are reliable only in plates where the majority of the compounds are inactive, and the few active compounds present are randomly scattered.

4.2. Description of a Spatial Pattern Recognition Method

There are different statistical approaches for spatial pattern recognition and treatment in HTS plates. The most common techniques use discrete Fourier transform *(22)*, Bayesian statistics, median polish algorithm *(13, 23)*, etc. We propose an approach based on exploratory data analysis (EDA). The steps followed are the following:

- Detection of patterns
- Evaluation of the importance of these patterns
- Correction of the HTS data, if necessary

A pattern recognition method has to be very flexible and robust. It has to be capable of analysing a great variety of HTS data. The variety of biological conditions, experimental designs, readers, previous calculations with the data, etc. lead to different kinds of data, and we have to take all of them into consideration. Also, we need a robust method, which is not sensitive to outliers. An outlier is an observation that is a long way from the majority of the values in the data set. In HTS data, the presence of outliers is quite common, due to data errors and especially due to the presence of active compounds, with values that are far off the majority of inactive compounds.

Experience shows that patterns in each plate are different in strength and shape. Likewise, patterns may appear and disappear along the HTS campaign. Hence, we propose an intra-plate methodology where the analysis unit is the plate and where we do not use information from previous or subsequent plates.

4.2.1. Detection of Patterns

We can consider HTS data as values in a three-dimensional space, where the rows and columns in the plate represent the x- and y-axis and the activity value of each sample is represented along the z-axis. The main aim of the pattern recognition method is to describe the surface that best fits this data set. **Figures 4.6** and **4.7** show examples of plates with spatial patterns. We assume that the majority of assayed compounds are inactive. If the values are randomly scattered on the plate and no spatial systematic errors are present, we can expect the data to be randomly situated below or above the plane $z = 0$, and then, when we fit a surface to this data set, the surface will

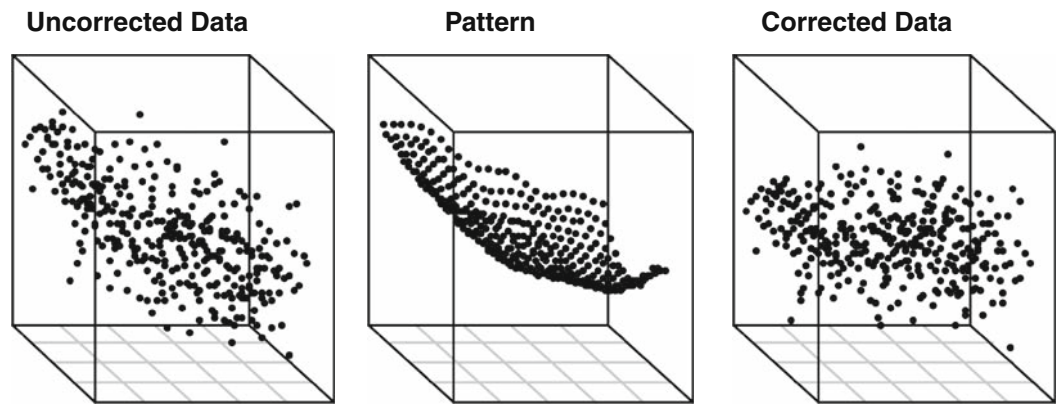


Fig. 4.6. **Example of a robust surface fitted by the pattern recognition method**. Three graphs are shown in the figure: uncorrected data (*left*), pattern found (*centre*) and corrected data (*right*).

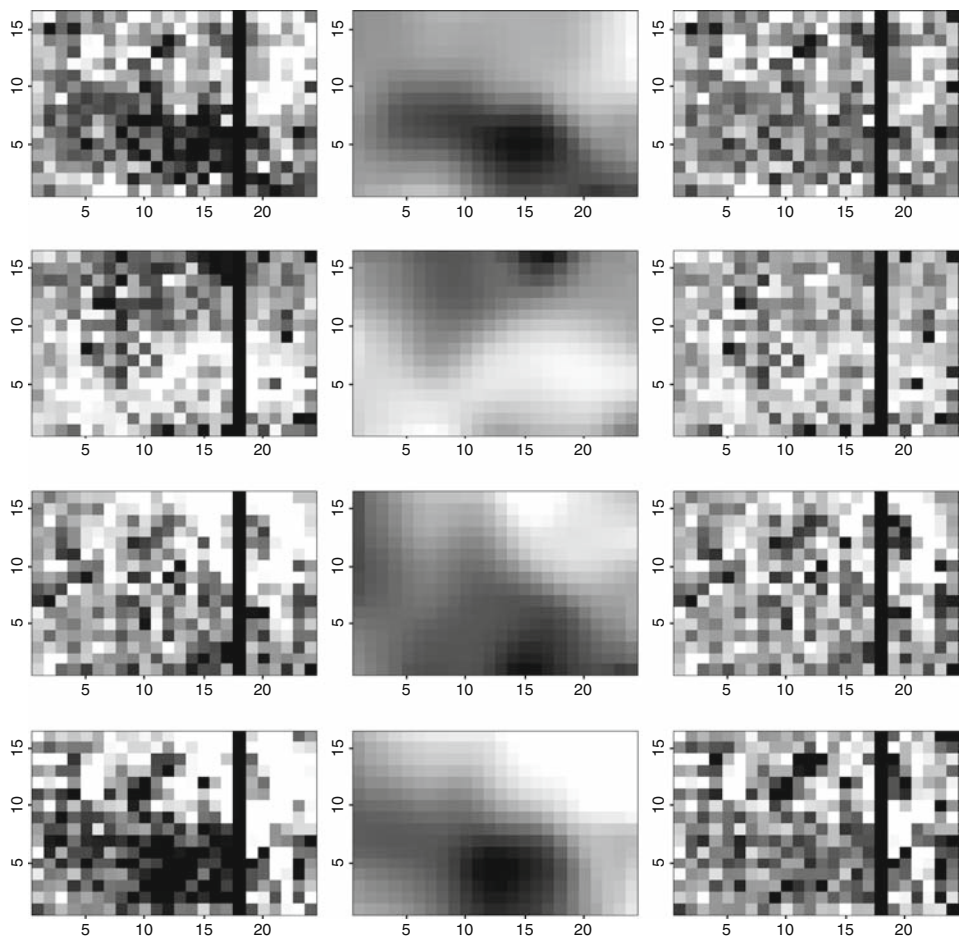


Fig. 4.7. **Examples of the correction done by the pattern recognition method in four plates**. Three columns are shown in the figure: uncorrected data (*left*), pattern found (*centre*) and corrected data (*right*). Grey colour scale range from –20% activity (*white*) to 30% activity (*black*).

be precisely on the plane z = 0. We could say that our data set has one spatial pattern when the surface fitted is not a plane. Later we will discuss how to assess whether the fitted surface is relevant for the original data.

The method to detect the pattern is based on a two-dimensional extension of the repeated running median procedure defined by Tukey for smoothing data sequences *(24, 25)*. Therefore, the method is a nonlinear smoothing method. The technique is descriptive and does not imply a theoretical model for the data. It is an exploratory data analysis and can also be considered as a robust local regression model. Its versatility means that it can be used in very different data sets such as those generated in HTS, and it is sufficiently flexible for a much wider range of situations found in HTS. The robustness of the method comes from the use of the medians to estimate the surface.

One of the best features of the running median procedure in analysing data series is that it detects the increment or the decrement parts of the data series extremely well. In the two-dimensional extension of the algorithm, this feature means that it is very easy to describe the areas of the data where they are increasing or decreasing, i.e. the gradients of the patterns we want to detect. Consequently, this method will adjust patterns well if they occur gradually, which is usually the case when they are generated by changes in temperature, centrifugation, reader calibration, etc.

The most basic smoother for data sequences is the "running median of 3", called the "3" smoother *(24)*. For example, for a sequence of data $\{x_t : t = 1, ..., T\}$, the "3" smoother replaces x_t with the median of $\{x_{t-1}, x_t, x_{t+1}\}$. Repeated smoothing or "resmoothing" (R) uses the output of a smoothing operation as input to the same smoothing operation and is repeated until no change occurs.

The basic idea of the two-dimensional running median extension for smoothing surfaces is to substitute each data point by the median of the surrounding data.

Suppose $\{x_{rc}; r = 1, ..., R, c = 1, ...C\}$ are the activity data of the compounds in an HTS plate with R rows and C columns. Formally, the "9R" smoother is defined as a repetitive algorithm with the following iterations:

- Iteration 1: Each x_{rc} is replaced by the median of the nine surrounding data

$$x_{rc}^1 = \underset{\substack{i=r-1,r+1 \\ j=c-1,c+1}}{median}\{x_{ij}\} \ \textit{where} \ r = 1, ..., R \ \textit{and} \ c = 1, ..., C$$

- Iteration 2: Each data are replaced by the median of the nine surrounding data obtained in iteration 1

$$x_{rc}^2 = \underset{\substack{i=r-1,r+1 \\ j=c-1,c+1}}{median}\{x_{ij}^1\} \ \textit{where} \ r = 1, ..., R \ \textit{and} \ c = 1, ..., C$$

- The algorithm stops when $x_{rc}^p = x_{rc}^{p-1} \ \forall r, c$.

The use of the median in this iterative method produces a non-continuous surface, like a "staircase", and to smooth this effect, as a last step, a smoother is applied in the form of a weighted mean with the 25 surrounding data, similar to the Hanning (after von Hann) defined by Tukey *(24)*. Each x^p_{rc} value calculated in the last iteration is replaced by

$$y_{rc} = \underset{\substack{i=r-2,r+2\\ j=c-2,c+2}}{average}\{w_{ij} \cdot x^p_{ij}\} \;\; where \sum_{\substack{i=r-2,r+2\\ ej=c-2,c+2}} w_{ij} = 1$$

This smoother is called 9RH.

The fact that this is an iterative method makes it particularly appropriate to fit gradients that gradually increase or decrease, because the peaks present in the surface disappear when the median is calculated in each step.

4.2.2. Variance Explained by Patterns: How to Evaluate Patterns

Visual inspection is the intuitive method for evaluating whether or not HTS data present spatial patterns. In today's HTS environment with thousands of plates analysed at once, this visual analysis has been aided by programs such as Spotfire. However, evaluating results by means of visual inspection can be subjective, and in fact we could easily get distracted by little patterns, without practical impact, that are present in the majority of the HTS.

The method for detecting patterns described earlier provides the grounds for putting forward a more objective way of evaluating the found pattern. We believe that part of the variability in the original data is due to the pattern, because on the edges or in the centre, we find data that are greater or smaller than expected, thus increasing variance. After fitting the surface that describes the pattern, we can use it to evaluate how much of the original variability is due to the data and how much is due to the pattern.

We define a measure of the strength of the patterns, called **variance explained by the patterns (VEP)**, as the ratio between variability of the found pattern and variability of the uncorrected data (**Fig. 4.8**). This is similar to the R-square, the coefficient of determination used in linear regression. To estimate variability, we use a robust variance estimator, $c \cdot \text{MAD}^2$, where c is a constant and MAD is the median of absolute deviations. Formally

$$\text{VEP} = \frac{\text{variance(pattern)}}{\text{variance(uncorrected)}} = \frac{\text{MAD}^2_{\text{patt}}}{\text{MAD}^2_{\text{uncorr}}}$$

A low VEP means that the found pattern does not contribute significantly to the variability of the original data, and therefore, the pattern is small, and we do not need to correct the data. If there is no pattern, the fitted surface adjusted by the method is near to the plane

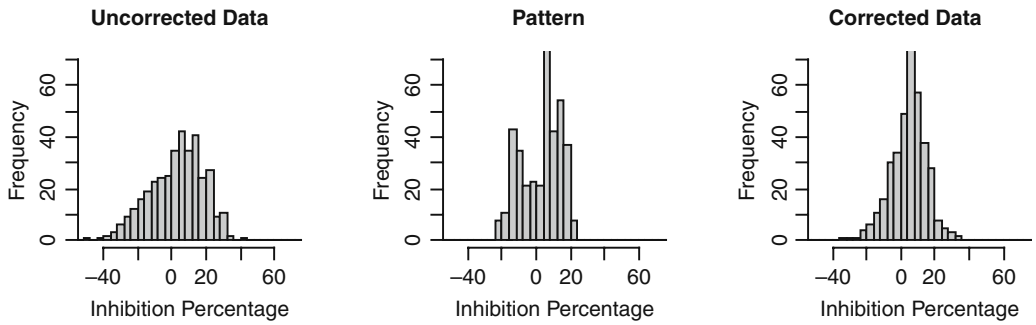


Fig. 4.8. **Variance explained by the patterns (VEP)**. Part of the original variance (*left histogram*) is due to the pattern found (*centre histogram*) and after correcting the data, this variability decreases (*right histogram*).

$z = 0$, which has null variability, and therefore, we would obtain VEP = 0. A high VEP means the pattern found contributes significantly to the variability of data, and the data must therefore be corrected.

Our experience has led us to define limits for VEP:

- VEP > 0.3 strong patterns
- VEP of 0.2–0.3 medium patterns
- VEP < 0.2 smooth patterns, or lack of pattern

4.2.3. Screening Data Correction

In some cases, after detecting the patterns, we can correct the experimental conditions to make some of these patterns disappear or decrease in intensity. However, this is not always possible and sometimes we need to correct the original data, in order to make the best decisions about which compounds are active.

The basic idea of correction is to take the distance to the surface as a new value of compound activity. In practice, this means adding or subtracting a quantity to the original data, depending on whether the well is positioned in a high or low area of the surface.

$$\text{Corr}_{rc} = \text{Uncorr}_{rc} + \text{Weight}_{rc} \cdot (\text{median}(\text{pattern}) - \text{Pattern}_{rc}) \forall r, c$$

where r and c are row and column identifiers, respectively.

This weight function will take values near to 1 for all the low activity values and near to 0 for high activity. The use of a weight function compensates for the lack of linearity in the response.

4.3. Temporal Patterns

The use of automation in HTS is the main cause of systematic errors across time, associated with a certain position in several consecutive plates. For example, these problems relate to compound dispensing, obstructed pipettes, contaminated wells, etc. The presence of temporal systematic errors principally affects false-positive findings, because inactive compounds are misclassified as active compounds.

The statistical techniques that we use to detect this kind of systematic error are different from those used to detect the spatial

patterns, but the mathematical foundation is similar, since we use smoothers based on medians as defined by Tukey *(24)*, which can estimate the tendency of data coming from a temporal series.

This inter-plate analysis is based on studying the data of each position on the plate, and each well, as an independent temporal series. In other words, in an HTS run of p plates with n samples each, we study n temporal series with p point each. The idea is to adjust a very robust temporal series to find strong trends in one of these series, describing what happens in each position of the plate across time.

Systematic error level (SEL) can be defined by applying a robust smoother. Normally, we use estimators such as the 11RH or the 15RH *(24)*, which are stronger than those advised by Tukey (4253H, 3RSSH), since our objective is to find a strong trend that is maintained in several consecutive plates and thus identify it as a systematic error.

If a position on a plate does not present a systematic error, and assuming that the majority of the compounds are inactive, we can hope that the data will be randomly distributed below or above 0% of the activity. In this case, when we fit to a robust smoother, the trend of the series (i.e. SEL) should be close to 0%. **Figure 4.9** shows different examples of the evolution of the activity in a well across time. **Figure 4.9A** shows a high systematic error, where all values are around 100% of activity in a sequence of 100 consecutive plates. We imagine that this will be the case in the majority of the wells. However, **Fig. 4.9B** shows a position of the plate where the values are around 20% of activity, and this could be classified as a low systematic error. Finally, **Fig. 4.9C** shows a position of the plate without systematic error, during 100 consecutive plates, when the values are randomly below or above 0% activity, and the fitted trend is near to 0% in all cases.

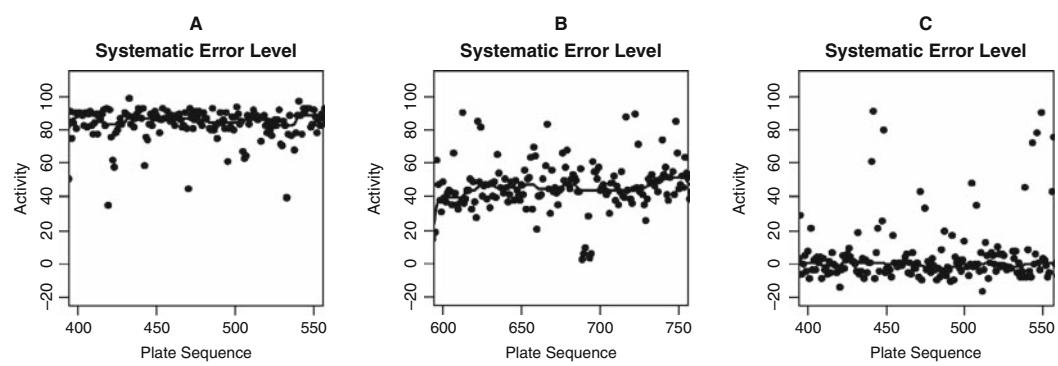


Fig. 4.9. **Evolution of the response values in a well along a set of consecutive plates, and the systematic error level (SEL)**. (**A**) A well with a high systematic error (SEL ≈ 100%). (**B**) A well with a medium systematic error (SEL ≈ 45%). (**C**) A well without systematic error (SEL ≈ 0%). *Dots* correspond to experimental activity. *Curve line* is the calculated activity upon fitting of the temporal sequence to the SEL robust smoother.

After calculating the SEL for all the plate positions, the data analysis is focused on detecting errors in the set of hits.

4.4. Summary and Conclusions

The presence of spatial and temporal patterns in HTS data can be a significant problem in some assays. Exploratory techniques based on repeated running median has been found to be an effective tool for detecting and correcting these systematic errors. The method has been found to be very flexible and robust for dealing with them in all HTS where they have been used. We have validated the method, and it shows a great capacity to detect false positives and recover false negatives.

5. Statistically Guided Selection of Hits in HTS

Besides a sound scientific rationale, an appropriate compound collection and a high-quality execution, the process of data mining is a key to success of every screening campaign. At the end of the day, HTS is a number game. In order to make the screening valuable, the vast amount of data and information that is gathered in any HTS blitz has to be conveniently processed, interpreted and transformed into real and meaningful information and knowledge. We assume here that in primary HTS every compound is tested at the same concentration and just once. Screening in replicates *(11)* or at different compound concentrations, such as quantitative HTS (qHTS) described by Inglese et al. *(26)*, can certainly help reduce the identification as hits of compounds associated with assay errors. It should be noted though that some claims of diminishing the burden in false negatives and false positives are based on generic assumptions, such as fixed activity threshold or 3 and 6 SD of the mean of actives being commonly used cut-offs. As will be reviewed below, the field has moved away from simple cut-offs, and high-quality assays are routinely run. Therefore, some of these claims should be downplayed.

The decision point relates to which compounds from a single shot test will be pushed forward as positives in the screen. Although the selection of hits can be guided **biologically** (e.g. potency and profiling of positives through secondary assays for specificity, selectivity or enablers) and **chemically** (e.g. clustering into representative chemotypes and deselecting intractable structures), the first question is merely **statistical**, i.e. where to set the threshold of activity that best segregates true positives from true negatives? On the other hand, the screening scientist struggles with a **logistical** constraint: the number of samples selected cannot surpass a limit dictated by the maximum number of samples that can be reasonably (that is in a timely and resource-efficient manner) prepared and tested through the subsequent assays.

It is not uncommon for primary positives in HTS simply to be selected on the basis of potency above a particular cut-off of activity that accommodates logistics. The ultimate consequence is that some putatively weak, but still valuable, true hits may have to be abandoned. However, the assignment of potency from a single shot experiment is rather risky. First, reliability of the activity value depends much upon the quality of the assay and the distribution of activity of the sample population tested. In other words, the same threshold of activity does not have the same reliability meaning in all assays. Second, the actual concentration of the compound in the assay might significantly differ from the nominal value *(27)*, so apparently weak compounds may turn out to be potent if they were actually tested as a trace. In all, a hit selection process that minimises the rates of false positives and negatives, regardless of their level of activity, would eventually optimise the use of limited screening resources.

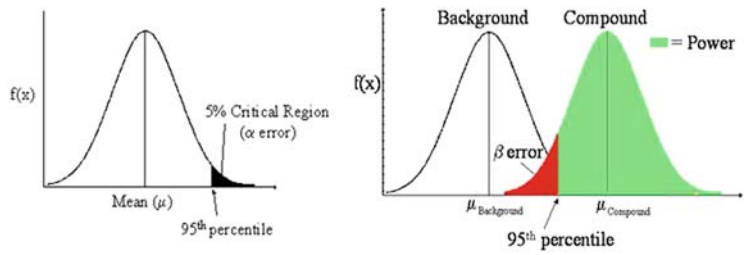
False positives are annoying. False negatives are unacceptable, because they are usually abandoned forever. Highly potent positives that eventually turn out to be false are worthless, disappointing and can negatively bias the selection. On the other hand, true weak positives are valid for SAR and provide a starting point in a hit-to-lead chemical programme. Our preference is giving the highest chances of picking up weak hits, accepting the risk of progressing with them false positives that will be unveiled in subsequent stages.

The process of hit identification involves three steps: (1) validation, removal and adjustment of screening data (*see* **Section 4**), (2) ranking compounds by activity and (3) setting a meaningful threshold to declare positive compounds. Below, we will describe statistical approaches to address the two last steps.

5.1. What Is a Statistical Cut-Off?

Currently, the most commonly used method for hit selection in HTS experiments is statistical significance (or *p*-value) for testing no mean difference and in particular the $mean \pm k$ SD method and its variants, where SD is the standard deviation of the negative reference (i.e. background controls or inactive samples) and k is a multiplying scalar. Alternatively, methods of clustering have been described on the basis of finding two statistically significant clusters of samples, namely active and inactive samples *(28)*. The statistical significance approach addresses the question of controlling the rate of false positives, also known as type I error or α. The value of k is chosen so that the false-positive rate (i.e. α) can be kept below a certain level. The higher the value of k, the more stringent the cut-off we set to lower the rate of false positives, but the higher the rate of false negatives (also known as type II error or β), and vice versa. Hence, it is a hard challenge to make both rates low at the same time, and there is always a trade-off between minimising the two types of error.

A simple way of estimating the probability of making true assignments for negative compounds at one particular threshold of compound activity is the calculation of the *power* of an assay *(19)* (**Fig. 4.10**). The power parameter is calculated as the complementary probability of the type II error (i.e. $1 - \beta$) of the assay:



α error is an estimation of false positive rate

β error is an estimation of false negative rate

$$\textbf{Power} = 1 - \beta$$

Fig. 4.10. Meaning of power of discrimination for an assay, alpha- and beta-errors.

$$\text{Power} = 1 - \beta = \Phi\left(\frac{|\mu_S - \mu_C|}{\sqrt{\frac{(\text{SD}_S^2)}{n_S} + \frac{(\text{SD}_C^2)}{n_C}}}\right) - \Phi^{-1}(\alpha)$$

where Φ is the cumulative distribution function of the standard normal distribution (i.e. $\Phi(x)$ represents the area under the standard normal distribution between minus infinity and x), n is the number of replicates, SD is the standard deviation, μ is the mean of activity in the assay, S is the population of samples and C is the population of controls or inactive samples.

Zhang et al. *(29)* and Fogel et al. *(30)* have developed predictive models of hit confirmation rates based merely on the primary HTS data obtained from compound testing in singlet. Recently, a new statistical parameter (SSMD, strictly standardised mean difference) has been introduced that also contemplates false-negative rate in an attempt to achieve a balanced control of both *(31)*. The idea of *power* is underlying SSMD, which can also be applied to quality control of HTS assays. We will focus later in this chapter on the description of methodologies that estimate the lowest threshold with statistical significance for the call of true positives. Although there may be labs where HTS is run in simultaneous multiple testing (e.g. qHTS *(11, 26)*), we will not discuss statistical approaches based on replicate testing *(2, 11)* because these data are not ordinarily available in routine HTS.

As a result of an HTS campaign, we obtain the distributions of three populations: signal controls, background controls and samples. The distribution of sample population in an HTS experiment can be modelled as a composition of one major population of inactive compounds (>95%) with a single mean and experimental variability, and a combination of many other minor populations corresponding to several active compounds (<5%) whose means of activity are spread throughout a broad range of values. Furthermore, there will be a non-normal distribution of extreme errors. The composed distribution results from the sum of all individual distributions. The statistical cut-off is defined as the lowest threshold of activity that distinguishes between active and inactive compounds with statistical significance. The statistical significance of this segregation is determined by the distance of separation between the means of the inactive samples and the weakest active population that falls apart from the noise of the screen. The fundamental question is how to estimate the noise of a screening experiment.

5.2. Permutations in the Methodological Approach

There are a number of alternatives that can be independently permutated for estimating screen noise and hence calculating a statistical cut-off:

- *Raw data or normalised values*: Although any statistical approach described can be applied to raw data from individual wells in a plate, the normalisation of data as a percentage of the response of controls in the plate is a simple way of mitigating the systematic plate-to-plate variation derived from different status of reagents or reading equipments *(2)*. Simplest normalisation formulae are percent of control or percent of response, both of them using the means of signal and background control wells. Unfortunately, a systematic error in control wells will affect all measurements on the samples of the plate. Gribbon et al. *(10)* have nicely illustrated the shortfalls of data analysis approaches solely relying on controls and measures derived from them. Although more robust normalisation estimators based on median polish have been defined in order to estimate and remove row and column effects [e.g. B-score *(2, 12, 13)* and R-score *(11)*], simple percent of response (%response) becomes a reliable parameter when preceded by pattern recognition and correction algorithms, as those described in this chapter.

$$\%\text{response} = 100 \times \frac{\left|M_{\text{signal}} - x\right|}{\left|M_{\text{signal}} - M_{\text{background}}\right|}$$

where M is the mean of the raw data for signal or background controls and x is the raw data for compound or well x.

- *Sample or background control populations*: Although the variability of the background control population is a fair and indeed the simplest way of calculating an indicator of screen variability, it has some shortcomings:
  - The inactive sample population is not centred at 0% of activity for all screens (**Fig. 4.11B**).

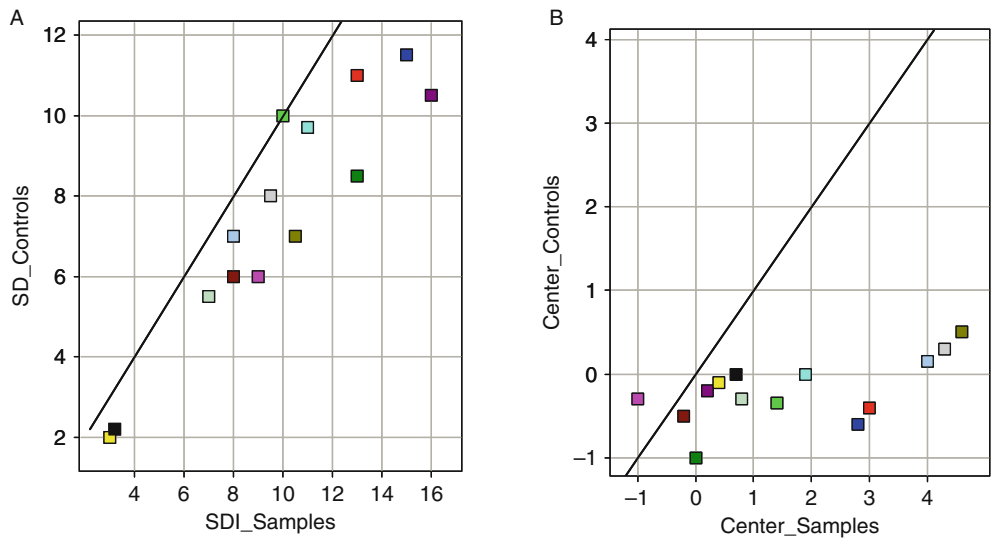


Fig. 4.11. Comparison of mean variability (**A**) and mean (**B**) of the populations of inactive samples (SDI, standard deviation of inactives) and signal controls (SD_Controls, standard deviation of controls with no compound effect) for 14 different screening campaigns.

  - The variability of the inactive samples (compounds) is often higher than that observed for the low controls (only DMSO is present). This is known as the "matrix effect" (e.g. organic load, chemical nature of the sample and procedure to dispense samples vs. controls) (**Fig. 4.11A**).
  - The number of inactive samples in an HTS campaign is much higher than the number of background controls (>95% vs. <5%, respectively), and they are randomly distributed in time and space, which results in a higher goodness of the estimation method whenever inactive samples are mostly abundant.
  - Thus, our preference it is to make use of the sample population distribution rather than the controls.
- *Size and identity of the batch of analysis*: Compounds in an HTS campaign are tested in groups or batches. Each compound is tested in a well within a plate, several plates are tested in an experimental run or independent experiment (e.g. same batch of reagents and single load in an automated platform),

and several runs are performed along the full course of the HTS campaign. In the pursuit of identifying a homogenous population of samples and controls with high significance, the run-wise analysis may be a good compromise. Ideally, screen quality should remain invariable throughout the campaign, but this does not always happen. On the other hand, plate-based variability may be misleading when some individual plates behave distinctly (e.g. spatial pattern and high number of active compounds).

- *Separation distance from screen noise (statistical significance)*: As discussed earlier, variability of experimental measures in screening can be modelled to normal distributions. Since the variability of an experimental population is estimated by its standard deviation (SD), statistical significance can be simply interpreted as distance from the mean in number of standard deviations. In this sense, 1, 2 and 3 SD accounts for approximately 68, 95 and 99% of the whole distribution. The probability of a value falling within a specified region of a distribution is equal to the area under the curve within the region, as a fraction of the total area under the curve. Based on the well-known 3σ rule, it is commonly accepted that a distance of $3 \times$ SD is the lowest statistical significance. The probability for values greater than $3 \times$ SD is 0.13% in a normal distribution, or in other words, the probability for a compound falling further than $3 \times$ SD from the mean of inactive samples of being a true positive is 99.97%. This estimate is based on the assumption that the whole population of inactive samples or controls is all randomly distributed and not subjected to any kind of patterns or extreme accidental errors. It is a good practice that the activity of compounds in single-shot screening is normalised as distance from the mean in number of SDs, as Z score does *(2)*.By doing so, the comparison of the profile of activity of a compound across different assays or targets will be more meaningful and reliable.

$$Z_{score} = \frac{x - M_{sample}}{SD_{sample}}$$

where x is the raw data on compound or well x and M_{sample} and SD_{sample} are the mean and standard deviation of the whole sample population, respectively.

- *Statistical methodologies for estimating screen noise*: The list of statistical approaches described to address the issue of screen noise estimation is so large that it cannot be comprehensively reviewed within the scope of this chapter *(32, 33)*. As discussed above, we recommend estimating screen noise through the analysis of sample distribution rather than from controls. Nevertheless, a simple approach to estimate screen noise, and

hence screen cut-off, can be done from Z-prime values. As mentioned earlier, Z-prime is a relative indicator of the power of the assay to discriminate active and inactive compounds. A rearrangement of the Z-prime equation can be derived as follows:

$$\%H_{\text{co}} = 100 \times \frac{3 \cdot SD_{\text{signal}}}{\left|M_{\text{Signal}} - M_{\text{Background}}\right|} = 100 \times \frac{(1 - Z')}{(1 + \alpha)}$$

where $\%H_{\text{co}}$ is the hit statistical cut-off as normalised percentage of activity and $\alpha = \text{SD}_{\text{background}}/\text{SD}_{\text{signal}}$. Hence, a correlation between Z-prime and a minimum statistically significant threshold of activity for a compound to be deemed *hit* (i.e. hit cut-off) can be simply estimated (**Fig. 4.12**). Note that α equals the inverse of signal-to-background ratio if we assume an identical coefficient of variation (CV = SD/M) for both types of controls. This approach can be easily used to accommodate hit thresholds of activity according to assay quality.

Below we will describe some methodologies based on the analysis of real screening compounds or samples.

5.3. Estimation of Screen Noise Based on Distribution of Sample Populations in the Screening Campaign

As discussed above, the whole population of biological activity from samples in a screening campaign does not follow a sheer normal distribution. Instead, it comprises overlapping observations from inactive samples, hits and uncorrected systematic errors, which will all contribute to skew the normality of the distribution. Since screen noise is dictated by the variability of inactive samples, which constitute the majority (>95%) of the whole population, we can conceptually approach the problem through two different routes: (1) sample population is just one single homogenous population containing outlier observations and (2) sample populations can be modelled as an overlapping sum of normal distributions from inactive samples (>95%) and hits (<5%). Despite the fact that the population of inactive samples constitutes the vast majority of the distribution, the usual estimates of the standard deviation and the mean are not accurate because they can be substantially inflated by extreme outliers, the number of which will depend on the proportion of true positives and extreme errors. The sample mean minimises the sum of squares (SS), and this is the source of its sensitivity to gross outliers as these large errors or actives inflate SS significantly. Usually, sample variance is even more sensitive to outliers than the sample mean. Therefore, more robust and adequate methods are needed in order to properly estimate screen noise from sample distribution.

5.3.1. Robust Statistics

The concept of robust statistics (sometimes also called resistant statistics) was initially developed to cope with the problem of outliers caused by errors in the measurement of an observation.

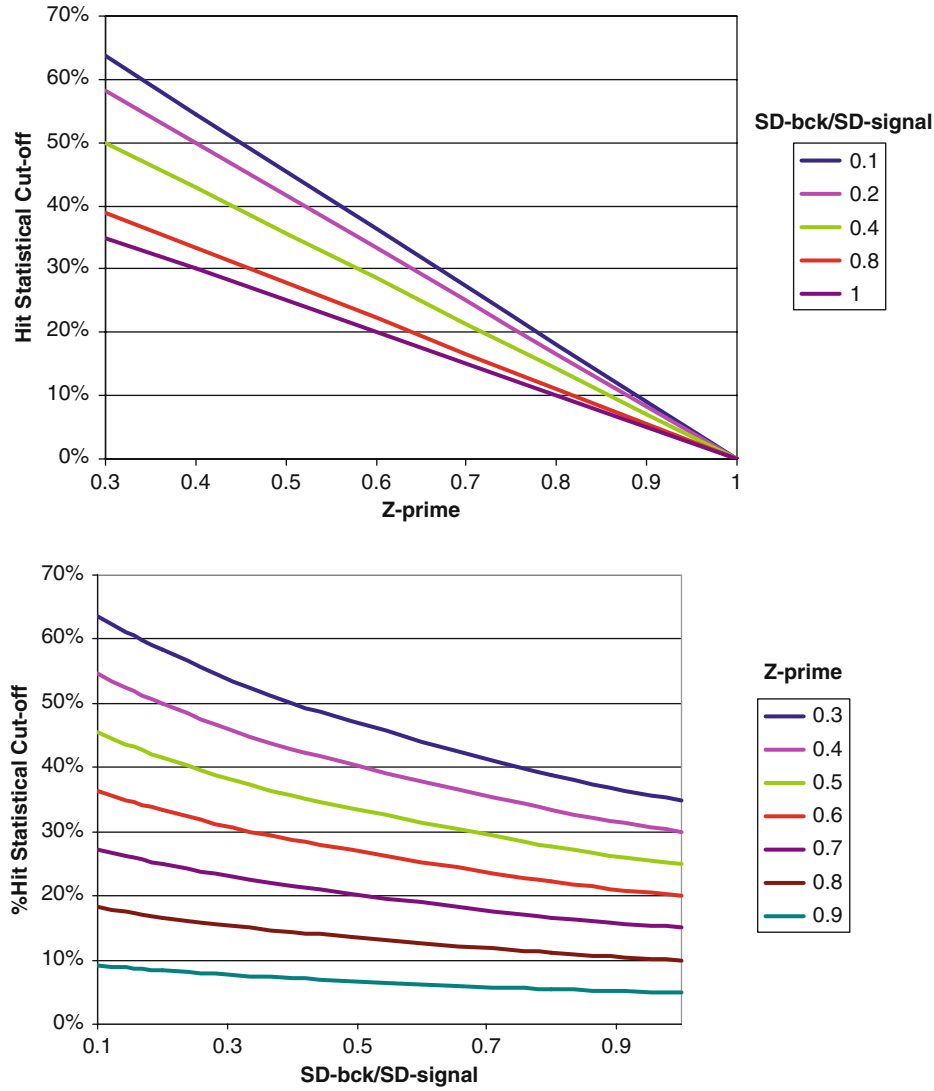


Fig. 4.12. Correlation between Z-prime and statistical cut-off based on the screen noise estimated from variability of controls. Hit statistical cut-off (%H_{co}) is calculated from the following expression derived from the Z′ definition: $\%H_{co} = 100 \times \frac{3 \cdot SD_{signal}}{|M_{Signal} - M_{Background}|} = 100 \times \frac{(1-Z')}{(1+\alpha)}$ where %H_{co} is the hit statistical cut-off as normalised percentage of activity and $\alpha = SD_{background}SD_{signal}$ for population of control samples.

Real errors do not fit the normal distribution and mask the true values of mean and standard deviation of the distribution of results. Although it has been common practice to reject outliers as errors and to delete them from the set of data, the prevailing philosophy of robust statistics has changed the emphasis from rejection to accommodation. Robust methods are as useful in assessing variability, which is even more sensitive to outliers than the mean, as they are for central tendency of the true value. The two main principles of robust statistics are the following: (1) they have to work well for heavy-tailed distributions close to the normal

and (2) they have to protect against gross errors. Both specifications are relevant for the problem of screen noise from the sample population, so they can also be applied to our advantage. Many different procedures have been described that vary in complexity of calculus and adequacy. Although *trimmed means* and *IQR* (inter-quartile range: difference between observations one quarter in from each end; note: IQR = 1.35·SD for a normal distribution) are easily calculated by hand and obey both principles of robust statistics, more sophisticated methods have been developed that require simple computation. For instance, in Ref. *(19)*, a method and its corresponding computational programme are described. It is based on an iterative minimisation of sum of squares of errors by downweighting extreme errors. The method is easy to compute and rapidly converge. When applied to screening data, it can be flexibly used for any set of data no matter the size, i.e. from a few controls within a plate to a whole set of millions of samples in a complete HTS campaign.

5.3.2. Standard Deviation of Inactives (SDI)

Below we will discuss four methodologies described for the estimation of SDI. Although they conceptually differ from the robust statistics for outliers, these approaches are also based on the calculation of robust statistical descriptors, and thus they all share methodological practices.

5.3.2.1. Normal Probability Plot (NPP)

This well-known graphical tool provides a simple approach for calculating SDI *(30, 34)*. Since the population distribution is not normal, only a central portion of it turns out to follow a straight line. The SDI is calculated from the slope of the tangent at the origin (**Fig. 4.13**).

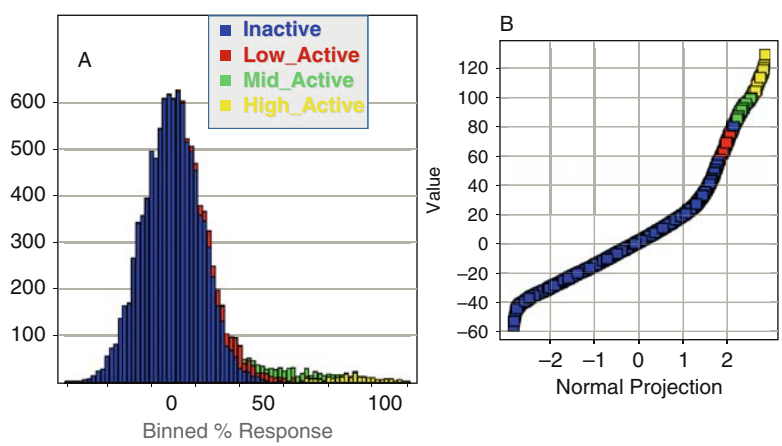


Fig. 4.13. Sample population is simulated as overlapping Gaussian distributions of inactive ($N = 10{,}000$, mean = 0, SD = 15; in *blue*), low active ($N = 500$, mean = 30, SD = 15; in *red*), mid active ($N = 300$, mean = 60, SD = 15; in *green*) and high active ($N = 200$, mean = 90, SD = 15; in *yellow*). Panel B is a normal probability plot for the whole sample population.

5.3.2.2. Empirical Cumulative Function

The SDI can also be estimated numerically through a nonlinear algorithm based on the empirical cumulative function of the activities of the whole population of samples [*see* Ref. *(30)* for a comprehensive description of mathematical formula]. The SDI can be obtained by minimising the sum of squares for the difference between the empirical cumulative function $F_n(x)$ and the standard error function $\mathrm{Erf}(x/\sigma)$:

$$\mathrm{SDI} = \arg\min \sum_{i=1}^{n} \left\{ F_n(x_i) - \mathrm{Erf}\left(\frac{x_i}{\sigma}\right) \right\}^2$$

Based on this, an estimate of the proportion of the inactive compounds is derived. Ultimately, an algorithm is set up to reach the quantitative description of the whole distribution of samples by a mixture distribution model for inactives and actives, and hence a calculated probability for a sample being a true active *(29, 30)*. From this kind of model, a predicted confirmation rate can be derived, so the number of primary positives selected for follow-up can be optimised to maximise the number of true positives without picking up too many false negatives.

5.3.2.3. Precision Radius

In an earlier section, we have discussed the basis of precision radius and its applicability in assessing assay reproducibility *(35)*. Based on the hypothesis that any HTS assay contains two distinct populations, namely inactive samples (*parent population*) and outlier/active samples (*child population*), the statistics are performed within the IQR (inter-quartile range) of the sample distribution. Therefore, to some extent, it can be considered as a derivation of robust statistics. The first step consists of establishing the centre of the assay. The bottom and top 25% of the samples are removed from the statistical analysis in order to clearly eliminate the child population of actives from the analysis of the centre. Then, a 3σ band is calculated from the IQR from which all the analysis of the centre takes place. The noise ratio of population, and hence the statistical cut-off, is contained within this band of 3 SD of its mean (namely, the centre). From the analysis of this parent population, and based on ANOVA, other statistical parameters are calculated aimed at describing the quality of the assay:

- *Precision radius*, an estimate of the expected variability in future measurements of the same sample.
- *Repeatability ratio*, percentage of the variation within the 3σ band of the assay that can be explained by variation in the measurements of the same sample. Its complementary parameter is *reproducibility ratio*, i.e. the percentage explained by variation between samples.

5.3.2.4. ASDIC (Automated SDI Computational Tool)

A new statistical tool has been developed for the automated computation of SDI from the distribution of screening samples. The method follows three steps (**Fig. 4.14**):

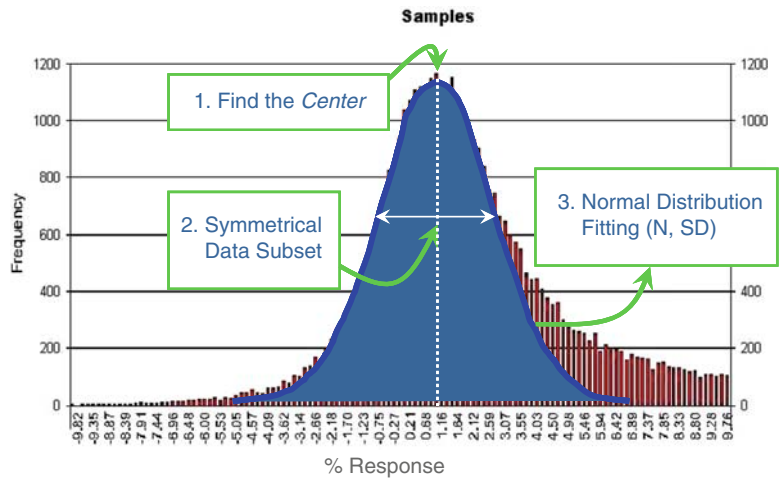


Fig. 4.14. Steps for the computation of SDI in ASDIC.

1. Establish the data centre
2. Establish the range within data that is symmetric around the centre
3. Fit to a normal distribution for this region of data, the mean of which is the data centre and standard deviation is the SDI estimator.

Step 1:

The data centre is the point where the histogram of the distribution reaches its maximum, or in other words at the highest data point density. As discussed above, the mean and the median are not robust estimators of this centre when asymmetry and outliers are significant. The method makes use of LTSq, i.e. least trimmed squares quartile. LTSq is equivalent to the mean for the subset containing the quarter of data, where these data points are mostly concentrated. Both mean and LTSq minimise the sum of squared residuals: mean, for all data set, and LTSq, for just the most populated quarter of the data set. LTSq turns out to be a more robust estimator than LTS (least trimmed square, which uses half of the population) and LMS (least median squares) for cases of high asymmetry *(36, 37)* (*see* **Appendix 1** for a description of the statistical parameters).

Step 2:

The symmetry around the data centre can be analysed adapting a methodology commonly used in exploratory data analysis for the study of symmetry in the tails of distribution. The method consists of setting a data interval centred on the LTSq value, with the same number of data points at each side of the centre. Then, the midrange for this interval (i.e. the average of the two corresponding limits of the interval) is calculated. If the midrange is equal to the LTSq estimator, the distribution symmetry is deemed

as being symmetrical in this interval. The difference is evaluated within a level of tolerance defined as a percentage of the value of the inter-quartile range (IQR), a robust estimator for the dispersion of the distribution. Successive iterations are run by broadening the interval to include a bigger number of data. It is said that there is asymmetry when the difference between the LTSq estimator and the midrange is greater than a certain tolerance parameter (e.g. $|\text{MidRange} - \text{LTSq}| > 0.02 \times \text{LTSq}$).

Step 3:

Once the data centre (i.e. LTSq) and the range of symmetry are determined, SDI is calculated as the standard deviation (SD) of the normal distribution where the mean equals the LTSq that best fit the data within the symmetrical interval. The size of the sample population (N) and the SD are iteratively varied, and the goodness of fit assessed by the chi-square test is based on the following statistics:

$$\chi^2 = \sum_{i=i}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where Q_i are the observed frequencies, E_i are the expected frequencies for a normal distribution and k is the number of intervals. The test evaluates the disagreement between observed and expected frequencies and concludes that the observed data follow the theoretical normal distribution if the sum of these weight differences is smaller than a critical point.

Figures 4.15 and **4.16** show an illustration of how ASDIC software works. ASDIC has proved to be more resistant to long tails and asymmetry, usually rendering lower SD values and hence

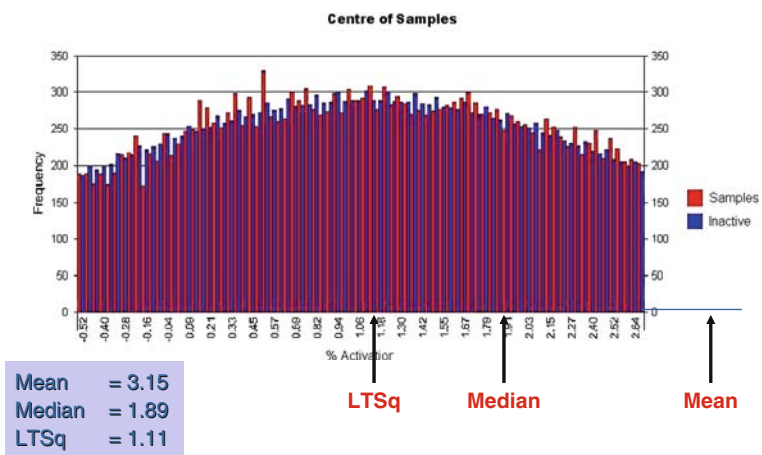


Fig. 4.15. **Estimation of centre of data by ASDIC**. LTSq (least trimmed squares quartile), median and mean are calculated for the whole population of samples. The plot depicts a zoom around the calculated LTSq. Theoretical distribution of the population of inactive samples is displayed in *blue*.

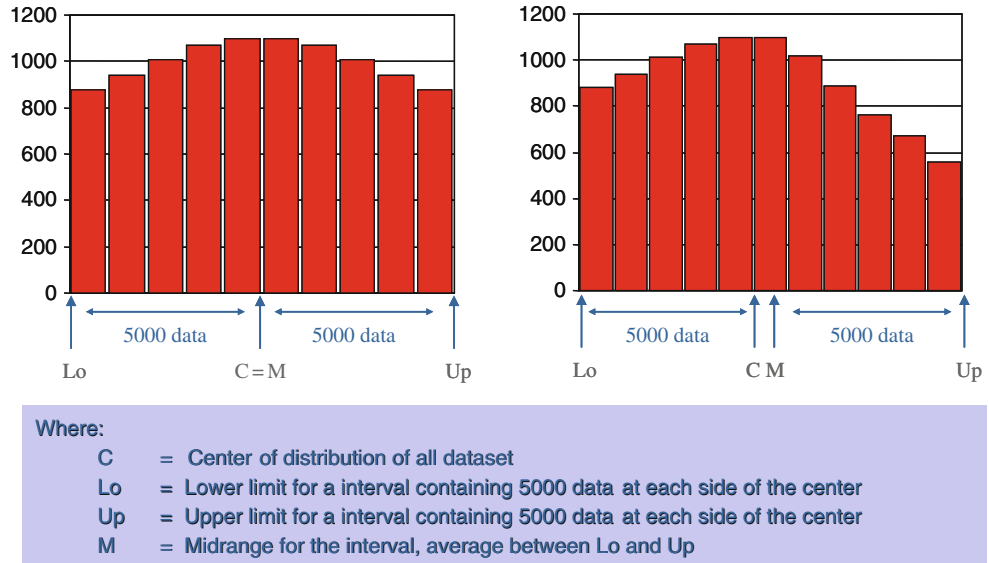


Fig. 4.16. **Setting symmetrical range by ASDIC**. **Left panel**: symmetry, the midrange for this interval equals the centre of distribution. **Right panel**: asymmetry, the midrange for this interval is greater than the centre of distribution.

lower statistical hit cut-offs. ASDIC has proved to be useful for setting statistical cut-offs of promiscuous screens (e.g. hit rates higher than 10%), circumventing spatial patterns (e.g. sample population from patterned positions can be segregated and separately analysed) and assessing screen quality as a QC indicator (e.g. inter-run comparison or comparison with SD from controls for establishing matrix effects).

Acknowledgements

The authors are greatly indebted to Ricardo Macarron, Mike Snowden, Mark Lennon, Gavin Harper, Martin Everett, Liz Clark, Glenn Hofmann, Geoff Mellor, Chris Molloy, Andy Vines, Dave Bolton and Javier Sanchez-Vicente for all the productive discussions about how to best implement statistical methodologies in the HTS process at GlaxoSmithKline. Likewise, we would like to thank many other colleagues in IT and Screening for their ideas and experimental data. SQC software has been the result of a joint collaborative effort with Tessella. We are also grateful to Robert Hertzberg, Stephen Pickett and Emilio Diez for their support in the writing of this manuscript.

Abbreviations

EDA: Exploratory Data Analysis
IQR: Inter-Quartile Range
M: Mean
MSR: Minimum Significant Ratio
PR: Pattern Recognition
QA: Quality Assurance
QC: Quality Control
QSAR: Quantitative Structure Activity Relationship
SD: Standard Deviation
SDI: Standard Deviation of Inactives
SEL: Systematic Error Level
SPC: Statistical Process Control
SQC: Screening Quality Control
uHTS: *ultra*-High-Throughput Screening
VEP: Variance Explained by the Patterns

Appendix 1: Estimation of the data centre in ASDIC

If the results of an HTS campaign are as below, we note:

n	size of the sample, number of compounds or pools
$(x_1, x_2, ..., x_n)$	activity values
$(x_{1:n}, x_{2:n}, ..., x_{n:n})$	ordered activity values
$x_{i:n}$	*i*th value in the ordered sample
$\hat{\theta}$	location estimator
$\hat{\theta}$	location estimator
$r_i = \left(x_i - \hat{\theta}\right)$	residuals
$\left(r^2\right)_{i:n}$	ordered squared residuals

– The mean is the LS (least squares) estimator, because it minimises the expression

$$\min_{\hat{\theta}} \sum_{i=1}^{n} r_i^2$$

– The LMS (least median squares) estimator minimises the expression

$$\min_{\hat{\theta}} \left(\underset{i=1,\ldots,n}{median} \left(r_i^2 \right) \right)$$

– The LTS (least trimmed squares) estimator minimises the expression

$$\min_{\hat{\theta}} \sum_{i=1}^{h} \left(r^2 \right)_{i:n}$$

where $h = [n/n22] + 1$ is the half sample size

– The LTSq (least trimmed squares quarter) estimator minimises the expression

$$\min_{\hat{\theta}} \sum_{i=1}^{q} \left(r^2 \right)_{i:n}$$

where $q = [n/4] + 1$ is the quarter sample size.

References

1. Charles Annis, Statistical Engineering. Available online at http://www.statisticalengineering.com
2. Malo N, Hanley JA, Cerquozzi S, Pelletier J, Nadon R. (2006) Statistical practice in high-throughput screening data analysis. Nat Biotechnol; 24(2): 167–175.
3. Macarron, R and Hertzberg R. Chapter 2 of this book, Design and Implementation of High Throughput Screening Assays.
4. Assay Guidance Manual Version 4.1. (2005) Eli Lilly and Company and NIH Chemical Genomics Center. Available online at http://www.ncgc.nih.gov/manual/toc.html
5. Taylor P, Stewart F, Dunnington DJ et al. (2000) Automated assay optimization with integrated statistics and smart robotics. J Biomol Screen; 5: 213–225.
6. Eastwood BJ, Farmen MW, Iversen PW, Craft TJ, Smallwood JK, Garbison KE, Delapp NW, Smith GF. (2006) The minimum significant ratio: a statistical parameter to characterize the reproducibility of potency estimates from concentration-response assays and estimation by replicate-experiment studies. J Biomol Screen; 11(3): 253–261.
7. Sittampalam GS, Iversen PW, Boadt JA, Kahl SD, Bright S, Zock JM, Janzen WP, Lister MD. (1997) Design of signal windows in high throughput screening assays for drug discovery. J Biomol Screen; 2: 159–169.
8. Iversen PW, Eastwood BJ, Sittampalam GS, Cox KL. (2006) A comparison of assay performance measures in screening assays: signal window, Z' factor, and assay variability ratio. J Biomol Screen; 11: 247–252.
9. Zhang JH, Chung TDY, Oldenburg KR. (1994) A simple statistical parameter for use in evaluation and validation of high throughput screening assays. J Biomol Screen; 4: 67–73.
10. Gribbon P, Lyons R, Laflin P, Bradley J, Chambers C, Williams BS, Keighley W. (2005) Sewing A. Evaluating real-life high-throughput screening data. J Biomol Screen; 10(2): 99–107.
11. Wu Z, Sui, Y. (2008) Quantitative assessment of hit detection and confirmation in single and duplicate high-throughput screenings. J Biomol Screen Online First; first published on January 23, 2008 as doi:10.1177/1087057107312628.
12. Gunter B, Brideau C, Pikounis B, Liaw A. (2003) Statistical and graphical methods for quality control determination of high-throughput screening data. J Biomol Screen; 8(6): 624–633.

13. Brideau C, Gunter B, Pikounis B, Liaw A. (2003) Improved statistical methods for hit selection in high-throughput screening. J Biomol Screen; 8(6): 634–647.
14. Wu G, Yuan Y, Hodge CN. (2003) Determining appropriate substrate conversion for enzymatic assays in high-throughput screening. J Biomol Screen; 8(6): 694–700.
15. Padmanabha R, Cook L, Gill J. (2005) HTS quality control and data analysis: a process to maximize information from a high-throughput screen. Comb Chem High Throughput Screen; 8(6): 521–527.
16. Westgard JO. (2001) Six Sigma Quality Design & Control. Desirable Precision and Requisite QC for Laboratory Measurement Processes. Westgard QC, Inc., Madison.
17. Enrick NL. (1985) Quality, Reliability, and Process Improvement. Industrial Press Inc, New York.
18. Coma I, Clark L, Diez E, Harper G, Herranz J, Hofmann G, Lennon M, Richmond N, Valmaseda M, Macarron R. (2009) Process validation and screen reproducibility in high-throughput screening. J Biomol Screen; 4(1): 66–76.
19. Analytical Methods Committee. Robust Statistics-How Not to Reject Outliers. (1989); Analyst 114: 1693–1697.
20. Kevorkov D, Makarenkov V. (2005) Statistical analysis of systematic errors in high-throughput screening. J Biomol Screen; 10(6): 557–567.
21. Available online at http://www.info2.uqam.ca/~makarenv/HTS/old/hts.html
22. Root DE, Kelley BP, Stockwell BR. (2003) Detecting spatial patterns in biological array experiments. J Biomol Screen; 8(4): 393–398.
23. 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(13): 1648–1657.
24. Tukey JW. (1977) Exploratory Data Analysis. Addison-Wesley, Reading, MA.
25. Hoaglin J, Mosteller F, Tukey J. (1983) Understanding Robust and Exploratory Data Analysis. John Wiley, New York.
26. Inglese J, Auld DS, Jadhav A et al. (2006) Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries. Proc Natl Acad Sci USA; 103(31): 11473–11478.
27. Popa-Burke IG, Issakova O, Arroway JD, Bernasconi P, Chen M, Coudurier L, Galasinski S, Jadhav AP, Janzen WP, Lagasca D, Liu D, Lewis RS, Mohney RP, Sepetov N, Sparkman DA, Hodge CN. (2004) Streamlined system for purifying and quantifying a diverse library of compounds and the effect of compound concentration measurements on the accurate interpretation of biological assay results. Anal Chem; 76(24): 7278–7287.
28. Gagarin A, Makarenkov V, Zentilli P. (2006) Using clustering techniques to improve hit selection in high-throughput screening. J Biomol Screen; 11(8): 903–914.
29. Zhang JH, Chung TD, Oldenburg KR. (2000) Confirmation of primary active substances from high throughput screening of chemical and biological populations: a statistical approach and practical considerations. J Comb Chem; 2(3): 258–265.
30. Fogel P, Collette P, Dupront A, Garyantes T, Guedin D. (2002) The confirmation rate of primary hits: a predictive model. J Biomol Screen; 7(3): 175–190.
31. Zhang XD. (2007) A new method with flexible and balanced control of false negatives and false positives for hit selection in RNA interference high-throughput screening assays. J Biomol Screen; 12 (5): 645–655.
32. Wu X, Sills MA, Zhang JH. (2005) Further comparison of primary hit identification by different assay technologies and effects of assay measurement variability. J Biomol Screen; 10(6): 581–589.
33. Sui Y, Wu Z. (2007) Alternative statistical parameter for high-throughput screening assay quality assessment. J Biomol Screen; 12(2): 229–234.
34. Li Z, Mehdi S, Patel I, Kawooya J, Judkins M, Zhang W, Diener K, Lozada A, Dunnington D. (2000) An ultra-high throughput screening approach for an adenine transferase using fluorescence polarization. J Biomol Screen; 5(1): 31–38.
35. Janzen W, Bernasconi P, Cheatham L, Mansky P, Popa-Burke I, Williams K, Worley J, Hodge N. (2004) Optimizing the chemical genomics process. In: Darvas F, Guttman A, Dorman F (eds) Chemical Genomics: Advances in Drug Discovery and Functional Genomics Applications. Marcel Dekker, New York.
36. Rousseeuw PJ, Leroy AM. (1987) Robust Regression and Outliers Detection. John Wiley, New York.
37. Ripley BD, Venables WN. (2000) Modern Applied Statistics with S. Springer.