NoiseMaker: simulated screens for statistical assessment
Phoenix Kwan and Amanda Birmingham*
Thermo Fisher Scientific, 2650 Crescent Drive, Lafayette, CO 80026, USA
Associate Editor: John Quackenbush

ABSTRACT
Summary: High-throughput screening (HTS) is a common technique for both drug discovery and basic research, but researchers often struggle with how best to derive hits from HTS data. While a wide range of hit identification techniques exist, little information is available about their sensitivity and specificity, especially in comparison to each other. To address this, we have developed the open-source NoiseMaker software tool for generation of realistically noisy virtual screens. By applying potential hit identification methods to NoiseMaker-simulated data and determining how many of the pre-defined true hits are recovered (as well as how many known non-hits are misidentified as hits), one can draw conclusions about the likely performance of these techniques on real data containing unknown true hits. Such simulations apply to a range of screens, such as those using small molecules, siRNAs, shRNAs, miRNA mimics or inhibitors, or gene over-expression; we demonstrate this utility by using it to explain apparently conflicting reports about the performance of the B score hit identification method.

Availability and implementation: NoiseMaker is written in C++, an ECMA and ISO standard language with compilers for multiple operating systems. Source code, a Windows installer and complete unit tests are available at http://sourceforge.net/projects/noisemaker.

Full documentation and support are provided via an extensive help file and tool-tips, and the developers welcome user suggestions.

Contact: amanda.birmingham@thermofisher.com

Received on May 14, 2010; revised on July 23, 2010; accepted on August 5, 2010

1 INTRODUCTION

Data analysis and hit identification are points of confusion for many screeners (Birmingham et al., 2009). Those asking questions such as “Which method identifies the most "true hits" for my particular screen circumstances?” or “What will the false positive rate of my chosen method be?” are frequently stymied, since answering these requires them to know the identity of the real hits. However, developing a list of the anticipated real biological hits for any given assay is extremely challenging and is likely to be both noisy and incomplete, especially for medium- to weak-strength effects.

The difficulty in assessing the performance of hit identification methods can be avoided by moving to in silico-based strategies. In the computational environment, one can generate a virtual screen containing defined true hits at known locations, and then perturb these true values with varying degrees and types of noise (both systematically biased and random) to simulate the variation inherent in biological screens; statistical techniques can then be evaluated based on their ability to identify known true positives and true negatives. These evaluations will be valid to the extent that the in silico hit distributions and types of noise are congruent with those of the real system. This approach offers both speed and flexibility, providing the opportunity to profile a method’s performance in many different realistic screening scenarios as well as the ability to simulate whole screens within minutes.

To enable such in silico testing, we have developed the NoiseMaker tool for generating simulated high-throughput screening datasets. A NoiseMaker user selects a realistic scenario for his or her simulated screen, including a range of hit properties as well as noise characteristics, derived from previous screens or assay development work (Supplementary Appendix 1); the software then randomly assigns “true hits” conforming to this scenario and generates noisy replicates of the screen. The user applies potential analysis approaches to this noisy data, using the known true hits to calculate metrics of interest (such as sensitivity, specificity or positive predictive value), and selects the most effective method.
shaped (non-linear) positional biases. Address non-Gaussian and multiplicative noise, as well as bowl-additive noise and linear positional effects. Future development will contain the input true values and one column of noisy values for range) using optional floor and/or ceiling values. The output file introduces the intended systematic effects. Noise can also be limited to a specific range (such as that simulating an instrument's detection well values away from their initial values by approximately the amount assigned to the Noise definition's mean change value, and the row noise combined. This simulates the convergence of realistic outcomes, from evaporation of reagents in edge wells to a blocked dispensing tip at a single well position.

Noise definitions are additive; e.g. if one is specified for the entire screen, one for Plate 2 and one for Row 5, then all values in Row 5 of Plate 2 will be permuted with the screen noise, the plate noise, and the row noise combined. This simulates the convergence of disparate systematic effects in real screens. All Noise definitions model noise as a Gaussian perturbation of the true values. They adjust the well values away from their initial values by approximately the amount assigned to the Noise definition’s mean change value, with the exact amount of adjustment being randomly chosen from a Gaussian distribution centered on the mean change value and with the specified standard deviation (SD) value. This ensures that each Noise definition produces realistically noisy adjustments even as it introduces the intended systematic effects. Noise can also be limited to a specific range (such as that simulating an instrument’s detection range) using optional floor and/or ceiling values. The output file contains the input true values and one column of noisy values for each simulated replicate.

Currently NoiseMaker is limited to Gaussian noise distributions, additive noise and linear positional effects. Future development will address non-Gaussian and multiplicative noise, as well as bowl-shaped (non-linear) positional biases.

3 SAMPLE APPLICATION

The B score (Brideau et al., 2003) is a normalization and hit identification method employing Tukey’s median polish, and has been proposed for use in screens displaying within-plate positional effects such as row and/or column biases. However, Makarenkov et al. (2007) have reported that it failed to recover correct hits in a scenario with ‘noisy standard normal data with systematic error stemming from row × column interactions which are constant across plates’. To address this apparent inconsistency in the literature and demonstrate how NoiseMaker can be applied in evaluating the performance of statistical techniques, we evaluated the B score in scenarios with different types of row and column positional effects: varying size of SD (Group A), varying size of mean change (Group B) and varying size of both mean change and SD (Group C).

After data sets with appropriate noise were created by NoiseMaker, we calculated B scores for all wells and identified the wells whose scores were in the top 1% as positives. We found that the true positive rates of datasets in Group A decrease and the false positive rates slightly increase as SD of the row and column noise increases (Supplementary Appendix II). However, the true positive rates and false positive rates of datasets in Group B remain steady regardless of the amount of mean change of the row and column noise, while the true positive rates and false positive rates of data sets in Group C behave similarly to those in Group A. These results suggest that B score is an appropriate choice for correction of systemic influences that primarily affect mean rather than variance. Notably, Makarenkov’s work examined simulated data with varying SDs, which is consistent with this finding.

4 CONCLUSION

NoiseMaker is simulation software for creating realistic, virtual high-throughput screens that can be used to evaluate hit identification methods and quality criteria. We establish its power by using it to clarify the utility of the B score under various screening conditions. This tool will be useful for broader comparisons of available hit identification methods, and is freely available for download and use by others interested in modeling screens in silico.

ACKNOWLEDGEMENTS

The authors thank Gabor Bakos of Trinity College, Dublin, and members of the RNAi Global Initiative for suggestions and helpful discussions, and Kevin Sullivan and John Quinn for code-review.

Conflict of Interest: none declared.

REFERENCES