BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 17 2013, pages 2218–2220
doi:10.1093/bioinformatics/btt358

Data and text mining

Advance Access publication June 28, 2013

SbacHTS: Spatial background noise correction for High-Throughput RNAi Screening

Rui Zhong[1], Min Soo Kim[1], Michael A. White[2,3], Yang Xie[1,2] and Guanghua Xiao[1,*]

[1]Quantitative Biomedical Research Center, Department of Clinical Science, [2]Harold C. Simmons Comprehensive Cancer Center and [3]Department of Cell Biology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: High-throughput cell-based phenotypic screening has become an increasingly important technology for discovering new drug targets and assigning gene functions. Such experiments use hundreds of 96-well or 384-well plates, to cover whole-genome RNAi collections and/or chemical compound files, and often collect measurements that are sensitive to spatial background noise whose patterns can vary across individual plates. Correcting these position effects can substantially improve measurement accuracy and screening success.

Result: We developed SbacHTS (Spatial background noise correction for High-Throughput RNAi Screening) software for visualization, estimation and correction of spatial background noise in high-throughput RNAi screens. SbacHTS is supported on the Galaxy open-source framework with a user-friendly open access web interface. We find that SbacHTS software can effectively detect and correct spatial background noise, increase signal to noise ratio and enhance statistical detection power in high-throughput RNAi screening experiments.

Availability: http://www.galaxy.qbrc.org/

Contact: Guanghua.Xiao@UTSouthwestern.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 21, 2013; revised on May 15, 2013; accepted on June 18, 2013

1 INTRODUCTION

RNAi is a process in which gene expression is inhibited by small RNA molecules such as small interfering RNAs (siRNAs) and short hairpin RNAs. High-throughput RNAi screening is a breakthrough technology for functional genomics and for drug target discovery (Orvedahl *et al.*, 2011; Whitehurst *et al.*, 2007). A frequently used screening platform uses 96-well or 384-well microtiter plates, on which each well contains siRNA oligos designed to target a specific gene. A number of methods and tools are available for data normalization (Dragiev *et al.*, 2012), visualization (Zhang and Zhang, 2013) and analysis of such experiments (Ogier and Dorval, 2012).

A challenge confronting the collection of precision measurements during High-Throughput Screening (HTS) is that the required experimental steps, such as procedures including cell culture, transfection, reagent delivery, incubation and HTS-plate scanning, may introduce spatially correlated background noise varying across experiments, batches and plates (Birmingham *et al.*, 2009; Carralot *et al.*, 2011; Malo *et al.*, 2006). Ignoring position effects results in low signal to noise ratios and reduces sensitivity. Recently, edge effects have been considered in normalization of high-throughput RNAi screening data (Carralot *et al.*, 2011); however, the only existing approach to address the global spatial background noise across a plate is to use *B*-score statistics (Malo *et al.*, 2006) to adjust row and column effects using analysis of variance. This method can be an effective approach for simple row and column effects. In our exploratory analysis of the high-throughput RNAi screening data, complex spatial patterns have been observed for most of the plates. Thus, simplified background correction approaches may often lead to overcorrection for some well plates and undercorrection for others. To help address this limitation, we sought advanced statistical models to enhance accuracy of quantification and correction of complex spatial background noise in high-throughput RNAi screening experiments.

Kriging interpolation (Banerjee *et al.*, 2003) is a well-established and widely used statistical model to fit spatial patterns in the observed data. In this study, we adapted a Kriging model to quantify and correct spatially correlated background noise in high-throughput RNAi screening data and developed a user-friendly software package, SbacHTS (Spatial background noise correction for High-Throughput RNAi Screening), together with intuitive data visualization and quality assessment tools, for open-source implementation of the Kriging correction. We find that SbacHTS software can effectively detect and correct spatial background noise, increase signal to noise ratio and enhance statistical detection power for RNAi screening experiments.

2 METHODS

SbacHTS software uses Kriging interpolation to fit spatial noise patterns and correct noise in high-throughput screening data. For each individual plate, at well s, let $Y_s = X_s + \varepsilon_s$ be the observed intensity (e.g. cell viability readout from the well), which includes both signal (X_s) and spatially correlated background noise (ε_s). We assume X_s is from a normal distribution $X_s \sim N(\mu_s, \sigma_s^2)$, where μ_s is the mean effect of siRNA in well s, and ε_s is from a multivariate Gaussian distribution $\varepsilon_s \sim \mathbb{N}(\vec{0}, \Sigma)$, where $\Sigma = \sigma^2\rho(\phi, d_{i,j}) + \tau^2 I$. Here, $\rho(\phi, d_{i,j})$ is a function of the distance between well plates i and j, $d_{i,j} = |s_i - s_j|$, with parameters ϕ, and $\tau^2 I$ models the independent white Gaussian noise. Using this model, we can estimate the spatial background noiseε_s and the signal X_s at each location s from the observed value Y_s. The software exports the estimated signals as a data matrix. For details about statistical inference and parameters estimation, reader can refer to Banerjee *et al.* (2003).

*To whom correspondence should be addressed.

3 RESULTS

The SbacHTS software (Fig. 1a) is developed with an intuitive web interface to facilitate broad implementation in the screening community. We used a real screening dataset (Whitehurst *et al.*, 2007) to evaluate the performance of SbacHTS software (Fig. 1b–e). This experiment used a commercial whole-genome arrayed siRNA library with 267 96-well plates to compare siRNA effects between paclitaxel-treated (experimental group) lung cancer cells and vehicle-treated (control group) lung cancer cells in triplicate.

3.1 Spatial noise pattern visualization

SbacHTS displays the spatial background noise pattern across each plate (Fig. 1b) and the corresponding fitted spatial background noise (Fig. 1c). This allows for fast and intuitive user appreciation of the position effects and spatial distribution of the calculated background noise across each plate for each screen. In addition, SbacHTS software exports observed intensity display too.

3.2 Improvement of coefficients of variation and statistical detection power

In the example dataset, the genome-wide siRNA screening experiments were carried out in triplicate. The coefficients of variation can, therefore, be calculated for each siRNA pool. Importantly, Figure 1d demonstrates that spatial background correction decreases the coefficients of variation, indicating improvement of the signal to noise ratio in the experiments. For example, the 90th percentile of coefficients of variation decreases from 0.044 in the original data to 0.039 after spatial correction. As a result, for 90% of genes, the statistical power to detect 10% changes increases from 0.88 to 0.94 when the Type I error rate is 5%. The goal of this screening experiment was to identify siRNAs, which significantly decrease cell viability in the experimental group compared with the control group (referred as hits). For each well, the cell viabilities were compared between the experimental group and the control group, using Student's *t*-test, and a *P*-value was calculated. A Beta Uniform model (Pounds and Morris, 2003) was used to determine the false discovery rate (FDR). After applying spatial background correction, we identified 867 hits with an FDR of <0.05, whereas we only identified 101 hits from the original data (Supplementary Figure S1) using the same criterion. Therefore, we successfully reduced the false-negative rate when controlling the same FDR (Supplementary Figure S2).

3.3 Visualization of batch effects

SbacHTS software provides a function to summarize the measurements (such as observed readouts or *z* scores) of each plate using a box-plot, grouping the box-plots by relevant variables (e.g. processing dates or batches). This function allows users to detect systematic bias or batch effects caused by uncontrolled experimental factors (Figure 1e).

4 IMPLEMENTATION

SbacHTS software is implemented in R and available as a Galaxy (Giardine *et al.*, 2010) tool with a user-friendly web interface. A detailed user manual and demonstration data are available at http://www.galaxy.qbrc.org/.

Although a genome-wide siRNA screen was used to demonstrate SbacHTS, the software is also applicable to other high-throughput screening experiments with similar workflow, such as chemical compound screening. Our results show that SbacHTS can correct spatial background noise, increase signal to noise ratio and improve hit identification from high-throughput screening experiments. In addition, SbacHTS is computationally efficient, requiring <5 minutes to process the 267-plate genome-wide data.

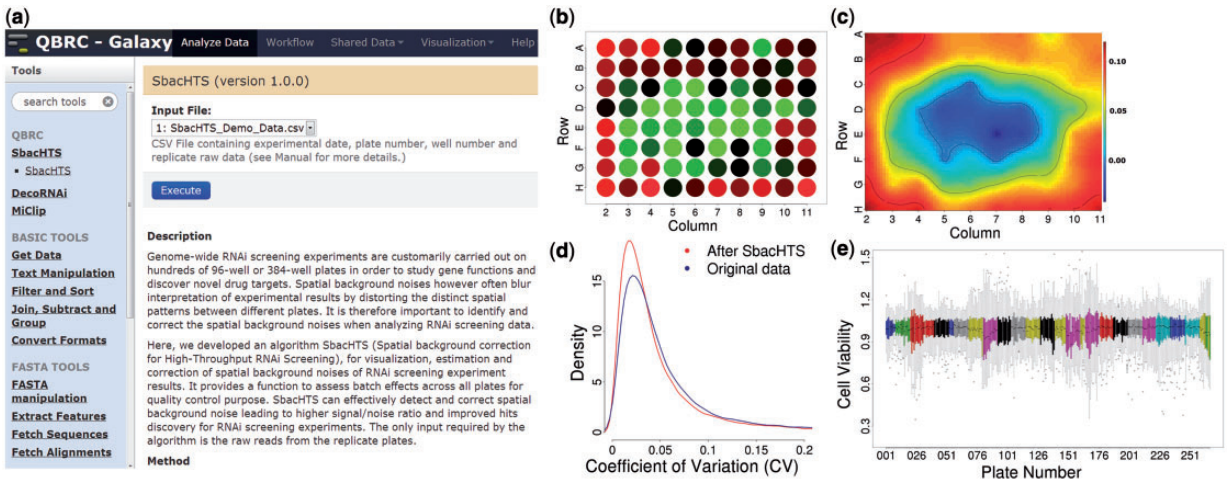

Fig. 1. (**a**) A Galaxy-based user-friendly web application of SbacHTS. (**b**) Visualization of spatial background noise pattern across plate. (**c**) Visualization of the fitted spatial background noise. (**d**) Density plots of coefficients of variation before (black line) and after (red line) spatial correction. (**e**) Visualization of batch effects using box-plots. Each box-plot summarizes the readouts from one plate, and the color indicates different experiment dates

Funding: National Science Foundation award (DMS-0907562) to G.X, National Institutes of Health grants (RO1 CA152301 to Y.X, and R33 DA027592 to G.X), the Cancer Prevention Research Institute of Texas (RP101251) to Y.X., and the Welch Foundation award (I-1414) to M.W.

Conflict of Interest: none declared.

REFERENCES

Banerjee,S. *et al.* (2003) *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC Press, Boca Raton, FL.

Birmingham,A. *et al.* (2009) Statistical methods for analysis of high-throughput RNA interference screens. *Nat. Methods*, **6**, 569–575.

Carralot,J.P. *et al.* (2011) A novel specific edge effect correction method for RNA interference screenings. *Bioinformatics*, **28**, 261–268.

Dragiev,P. *et al.* (2012) Two effective methods for correcting experimental high-throughput screening data. *Bioinformatics*, **28**, 1775–1782.

Giardine,B. *et al.* (2010) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.

Malo,N. *et al.* (2006) Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.*, **24**, 167–175.

Ogier,A. and Dorval,T. (2012) HCS-Analyzer: open source software for high-content screening data correction and analysis. *Bioinformatics*, **28**, 1945–1946.

Orvedahl,A. *et al.* (2011) Image-based genome-wide siRNA screen identifies selective autophagy factors. *Nature*, **480**, 113–117.

Pounds,S. and Morris,S.W. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, **19**, 1236–1242.

Whitehurst,A.W. *et al.* (2007) Synthetic lethal screen identification of chemosensitizer loci in cancer cells. *Nature*, **446**, 815–819.

Zhang,X.D. and Zhang,Z. (2013) displayHTS: a R package for displaying data and results from high-throughput screening experiments. *Bioinformatics*, **29**, 794–796.