BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 2 2012, pages 261–268
doi:10.1093/bioinformatics/btr648

Data and text mining

Advance Access publication November 26, 2011

A novel specific edge effect correction method for RNA interference screenings

Jean-Philippe Carralot[1,*,†,‡], Arnaud Ogier[2], Annette Boese[3], Auguste Genovesio[4], Priscille Brodin[1,5], Peter Sommer[3] and Thierry Dorval[6,*,†]

[1]Biology of Intracellular Pathogens, Inserm Avenir Team, [2]Cellular Differentiation, [3]Cell Biology of Retroviruses, [4]Image Mining, Institut Pasteur Korea, Seongnam-si, Korea, [5]Chemical Genomics of Intracellular Mycobacteria, Inserm U1019, CNRS UMR8204, Institut Pasteur of Lille, France and [6]Functional Morphometry, Institut Pasteur Korea, Seongnam-si, Korea

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: High-throughput screening (HTS) is an important method in drug discovery in which the activities of a large number of candidate chemicals or genetic materials are rapidly evaluated. Data are usually obtained by measurements on samples in microwell plates and are often subjected to artefacts that can bias the result selection. We report here a novel edge effect correction algorithm suitable for RNA interference (RNAi) screening, because its normalization does not rely on the entire dataset and takes into account the specificities of such a screening process. The proposed method is able to estimate the edge effects for each assay plate individually using the data from a single control column based on diffusion model, and thus targeting a specific but recurrent well-known HTS artefact. This method was first developed and validated using control plates and was then applied to the correction of experimental data generated during a genome-wide siRNA screen aimed at studying HIV–host interactions. The proposed algorithm was able to correct the edge effect biasing the control data and thus improve assay quality and, consequently, the hit-selection step.

Contact: dorvalt@ip-korea.org; jean-philippe.carralot@roche.com

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

Received on December 10, 2010; revised on October 27, 2011; accepted on November 10, 2011

1 INTRODUCTION

In high-throughput screening (HTS) campaign, although most repetitive errors can be controlled, some biases, such as edge effects (also called border effects), which appear after a long incubation period, cannot easily be corrected due to well-to-well discrepancies inherent in the spatial structure of each plate. In these cases, a number of post-screening correction methods have been developed to normalize the data and are currently applied to small molecule-based screens. Recently, the emergence of genome-wide RNA interference (RNAi) HTS has raised experimental complications such as the requirement for more incubation time than compound-based screens. However, because of the specificity of the small interfering RNA (siRNA) library format and the diversity of phenotypes induced after siRNA treatment, the classical methods developed for post-processing in compound screens cannot be efficiently implemented in RNAi screens. Within this framework, the relevance of the positive results, or 'hits', is highly linked to the consistency of each test within the complete screening campaign (Shun *et al.*, 2011). Thus, every experimental parameter is a potential source of variation and can lower the accuracy of the entire screen. Consequently, the entire process must be designed to maximize the similarity between each individual experiment. In such a context, each spatial or time-dependent dissimilarity can also be considered to be a possible source of response variability. The most widely used platform for biological experiments in HTS is the microtitre plate (hereinafter referred to simply as 'plate'); it is basically a rectangular rigid support containing an array of wells, each containing a single experiment. This type of support displays intrinsic heterogeneity in its spatial design: each well has a unique physical location, making it different from the others in terms of its immediate neighbours. Thus, the surrounding properties of a given well can have consequences for its response (usually a cellular phenotype) in the biological experiment; this is especially true for wells that are located at the border of the plate. Therefore, this locational variability can cause a differential in response across the plate. In practice, location-dependent variations in observed phenotypes can be due to many physical causes such as temperature and evaporation differences, inhomogeneous cell or agent dispensing, cross-contamination or plate-reader edge effects (Hser, 2006; Makarenkov *et al.*, 2007).

In many applications, these factors are either neglected or prevented by the use of specific countermeasures (Lundholt *et al.*, 2003). Nevertheless, because of the expense and technology involved in a screening campaign, such preventive measures are not always feasible or applicable. In these cases, post-correction of the data is mandatory.

In compound-based screening, methods for correcting spatial patterns of variation in plate responses have been widely studied and described in the literature (Brideau *et al.*, 2003; Dragiev *et al.*, 2011; Heyse, 2002; Kevorkov and Makarenkov, 2005; Makarenkov *et al.*, 2006, 2007; Malo *et al.*, 2006). These methodologies rely

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first and last authors should be regarded as joint Authors.
‡Present address: F. Hoffmann-La Roche Ltd, Assay development and HTS, 4070 Basel, Switzerland.

on several main assumptions that are usually justified in a standard compound-based screening:

- the difference between a positive and a negative response is obvious (in an uncorrupted case); and
- the positive responses are sparse and rare.

In some cases, it can also be assumed that the artefact occurs on a large number of plates. In this case, the correction process is well established. Indeed, considering hits to be impulsive noise, a simple plate data average directly describes the trend and consequently leads to an accurate evaluation of the trend. A simple correction process allows for a practical correction of the biased data.

If the hypothesis of a reproducible artefact throughout the plate set is not fulfilled, it is still possible to apply a plate-to-plate correction algorithm, such as those reported in (Birmingham *et al.*, 2009; Brideau *et al.*, 2003; Heyse, 2002; Kevorkov and Makarenkov, 2005; Makarenkov *et al.*, 2006, 2007; Malo *et al.*, 2006). These algorithms can be classified into two different methods: global and local.

In the global methodology, the background is fitted with a simple parametric function (usually a low-order polynomial), with the function considered to yield accurate bias modelling (Kevorkov and Makarenkov, 2005). With such an approach, the relevance of the model is crucial, and the function parameters should therefore be carefully chosen. Indeed, on one hand, too simple a function is unable to accurately model the spatial bias, leading to an inadequate correction, whereas on the other, a too-complex function can induce an overfit of the signal, raising the risk of the loss of true positive events. However, such approaches are well suited for smooth and well-defined spatial bias such as row, column and bowl-shaped effects as described in (Zhang, 2011).

With the local approach, the 2D signal defined by the plate spatial responses is denoised using limited-size kernel filtering, such as the non-linear median method (Brideau *et al.*, 2003; Malo *et al.*, 2006). In this situation, the sparsity of hits and the high signal-to-noise ratio of the positive versus negative controls are of primary importance. Indeed, this denoising methodology is based on the assumption of impulsive noise corruption and can thus generate artefacts when this assumption is not realistic.

Recently, in addition to traditional compound-based screening, commercially available genome-wide libraries of chemically synthesized siRNAs have enabled the emergence of RNAi HTS. RNAi using siRNA has become the gold standard method for loss-of-function studies in a variety of organisms. Compared to compound screening, one difference in siRNA-based screening is the extended incubation. Typically, the cells are first transfected with siRNA using a lipofectant compound and then incubated for 48–96 h before the phenotypic assay. Thus, when compared to small-compound screening, these two to four additional days of incubation required for an efficient knockdown of protein expression are likely to exacerbate experimental bias and cause additional artefacts. Indeed, long incubation periods combined with low-volume wells create ideal conditions for inconsistency.

The normalization methods developed for compound screening mentioned above are not applicable to siRNA-based screening, as the basic hypotheses on which these methods rely on cannot be verified (Birmingham *et al.*, 2009; Zhang, 2008). First, commercial siRNA libraries are clustered in most cases, and hits should therefore not be randomly distributed. Indeed, siRNA libraries are organized into sublibraries (e.g. kinases, phosphatases and GPCR), and siRNA experiments targeting the same pathway or the different subunits of a given heterodimer will have similar well contents in the same vicinity. Hence, positive events should appear in a clustered fashion, rejecting the main hypothesis that hits are rare, isolated events. Secondly, as opposed to compound screens, where most compound have no effect on cells, therefore the distribution of their readout can be roughly considered as Gaussian with positive event being 'outliers' (Malo *et al.*, 2006; Shun *et al.*, 2011), the results from siRNA screens are generally more widely distributed. Such a wide range of responses can be explained by several hypotheses described in Supplementary Material 1. Recent correction approaches such as B-score perform well on general cases, but can display errors when dealing with edge effects in a noisy environment, strong bias or hit clusters (Supplementary Figs S8–S17). Considering the new experimental conditions required by siRNA screening (e.g. increased incubation periods and clustered libraries) and the specificities of the RNAi-induced responses (e.g. wide phenotypic scales and possible patterns of responses), we introduce here a new method enabling the correction of edge effects occurring in RNAi screens. For each individual plate, the algorithm first extracts and evaluates the spatial bias from a control column and extrapolates the artefact model to the remainder of the plate.

In (Zhang, 2011), the authors identified a set of spatial artefacts corrupting their siRNA screenings and classified them in three categories: column-, row- and bowl-shaped effects. The last one displays similarities with what we defined here as an edge effect. However, if a second-order polynomial can usually model efficiently the effect; such approach was unable to correct properly our edge effect. The widely used B-score correction also generated similar artefacts as shown in Supplementary Figs S8 and S9. Our goal here is not to propose a generic approach for correcting unknown systematic error corrupting experimental data in high-throughput screening, but to target a specific and recurrent type of artefact known as edge effect. Thus, for our method to be effective, the user has to identify and assess the error type by performing control plate analysis prior or within the screening campaign. This article's purpose is also to show that a relevant prior on the phenomena can provide not only a better understanding of the error, but also a relevant correction by making the method adapted to one specific case. However, we proved that if no bias is present, our correction scheme does not induce modification of the original data.

We developed and validated our proposed method on control plates and applied the data-correction protocol to a genome-wide siRNA screen studying the interactions between human cells and the human immunodeficiency virus (HIV).

2 MATERIALS AND METHODS

2.1 Border-effect characterization

During screening development, the robustness and reproducibility of the assays were addressed. The accuracy and homogeneity of each automated process was verified to exclude any artefact possibly caused by automatic liquid handling. To ascertain the inter-well and intra- and inter-day reproducibility of the assay, two sets of five control plates were evaluated with a 1 day interval under exactly the same conditions as in the subsequent HTS (Fig. 1a and Supplementary Material 2).

As shown in Figure 2, this study revealed a dramatic edge effect, as the HIV infection burden was increased on the border of the 384-well plates.

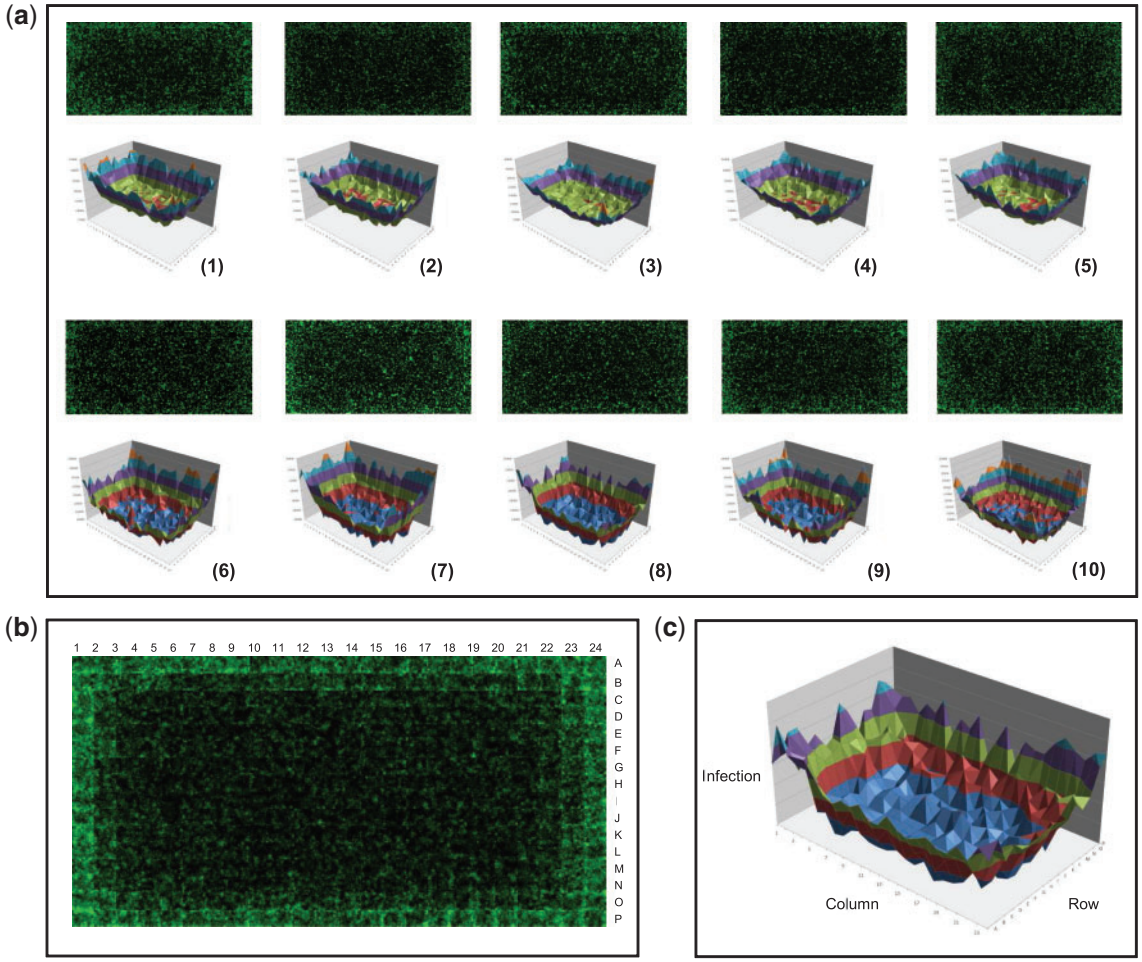

Fig. 1. Spatial bias visualisation. (**a**) Overview of the edge effects on the five control plates evaluated on Day 1 (top row) and on Day 2 (bottom row). Panels (**b**) and (**c**) display a synthetic reconstruction of a 384-well control plate with the confocal images of HIV-infected cells. For each well, a randomly chosen picture acquired within the considered well is displayed, giving an overview of the biological nature of the bias. The confocal images were acquired on Day 5 after infection. Cells overexpress the EGFP reporter protein when infected (green).

Despite the complications of plate-to-plate and overall day-to-day variation in the HIV infection, the bias that was observed for the edge wells was reproducible and seemed to follow a regular trend. Inaccurate cell dispensing and variation in knockdown efficacy were ruled out as explanations for this bias. Indeed, on Day 4 after transfection and before HIV infection, the cell number (measured by cell coverage) and silencing level (monitoring the EGFP fluorescence of cells transfected with EGFP and non-targeting siRNAs) were found to be reproducible and homogeneous compared with wells located at the centre and the edges of 10 control plates (Fig. 1a).

The data displayed a systematic strong spatial bias, corrupting the fluorescence intensities (see Figs 1 and 2 and Supplementary Fig. S5). Indeed, unlike a compound-based screening, an siRNA-based screening presents a large variety of results. Thus, it is inappropriate to consider the results as a very large number of negative responses containing rare and sparse hit events. Consequently, the data-correction algorithm should rely mainly on the dedicated positive controls, and the correction must be extrapolated over the entire plate.

To characterize the artefact, we used the 10 control plates and created a synthetic image resulting from a well-by-well average of the pictures acquired for each well. The result is displayed in Figure 1c, as is the corresponding 3D graph representing the average intensity level for each well (Fig. 1b). This analysis shows that this systematic bias was located on the borders of the plate and seems to be caused by the plate geometry itself. The intensity of the bias corresponding to each well appears to be related to the distance of the well from the plate border.

Consequently, we refocused our study on a more specific aspect of background modelling. Indeed, in the HTS context, as the liquid volume in each well is relatively low (particularly in 384-well plates), it is clearly sensitive to evaporation and hydration by atmospheric moisture (Ramadan *et al.*, 2007). Such processes have a direct impact on the quality of the results (Fig. 1) and should be taken into account during either the screening optimization procedure (Lundholt *et al.*, 2003) or the data analysis step (Birmingham *et al.*, 2009; Zhang, 2008).

The main assumption of the proposed method is that the evaporation process can be modelled over time by a diffusion model; i.e. the readout of each well will vary following this diffusion law across the entire plate. The evaporation begins at the plate borders, creating this 'edge effect', evolving continuously and smoothly over space and time. Consequently, our aim was to first define the correct set of parameters governing the diffusion-state model corresponding to the experimental data observed on each plate and then apply a relevant correction. However, libraries provided by vendors in 384 wells plate format do not let the user the choice of the control wells,

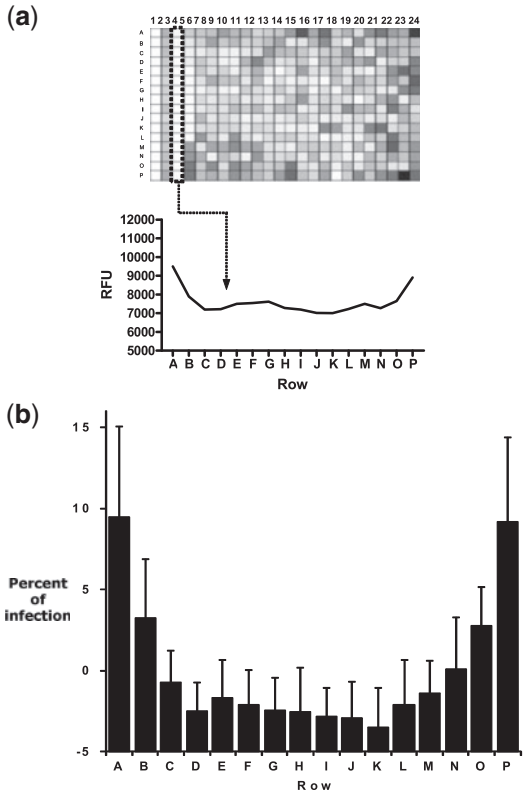

Fig. 2. Edge effects in a genome-wide siRNA screening. (**a**) Characterization of the spatial bias within the control column for plate performed at Day 1 of the screening campaign. (**b**) Average percentage of HIV infection across the control column of cells transfected with RAB9-targeting siRNA (Column 3). The results are the means ± SD for the indicated wells of the 68 plates, covering one representative genome-wide screening repeat.

which are usually located at the first and second columns near the plate edges. Even if the second column is still robust for evaluating the effect in various conditions, the relevance of our approach obviously increases with the number of controls. We propose here a correction identifying the model parameter from one column and apply the correction to the entire plate, but in case of adapted data (i.e. acceptable signal to noise ratio and no hit cluster), the model evaluation can be performed on the entire plate.

2.2 Diffusion model

2.2.1 Plate and signal modelling Assuming a multiplicative bias, the values associated with each well are given by:

$$z'_n(i,j) = (b_n(i,j) * z_n(i,j)) + \epsilon(i,j), \forall (i,j) \in \Gamma, \tag{1}$$

where (i,j) represents the spatial coordinates of the well over the plate, n is the plate index, $z_n(i,j)$ is the original fluorescence signal, $b_n(i,j)$ is the multiplicative bias and $\epsilon(i,j)$ is the standard additive noise, usually defined as Gaussian noise with low variance. Γ defines the dispensed set of well coordinates on a plate of dimension $M \times N$ ($\Gamma : [1,M] \times [1,N]$ for a rectangular, fully dispensed plate). In this specific application, we do not consider the bias to be correlated from plate to plate, and thus, we do not consider the plate index n as a relevant parameter for the surface (background) estimation. Consequently, and to simplify the notation, the plate index n is omitted hereafter. Nevertheless, if the operator can highlight a plate-to-plate repeatability in the bias, the bias estimation can be improved by averaging the signal over a set of plates and thus increasing the robustness of the proposed algorithm. Finally, assuming a multiplicative bias and a good signal-to-noise ratio for the additive noise [i.e. $\epsilon(i,j) \approx 0$], a fitting procedure for the background $b(i,y)$ should allow us to retrieve the unbiased signal $z(i,j)$ using Equation (1).

2.2.2 Diffusion equation To model the edge effects, we choose a linear version of the diffusion equation, also known as the heat equation. This specific 2D diffusion process follows the parabolic differential equation given by

$$\frac{\partial \tilde{b}(i,j,t)}{\partial t} = c \Delta \tilde{b}(i,j,t), \tag{2}$$

where $\tilde{b}(i,j,t)$ is the estimated spatiotemporal diffusion field over the plate (i.e. the estimated bias), c is the diffusion coefficient (set at unity for convenient notation) and Δ is the Laplacian operator. The partial differential equation (PDE) solutions are also defined by the initial conditions $\tilde{b}(i,j,t=0)$. Under the assumption of a bias related to plate geometry, for a fully and equally dispensed plate, these boundary conditions can be written as follows:

$$\begin{cases} \tilde{b}(i,j,t) = U_1, \ \forall (i,j) \in \mathbb{Z}^2 \setminus \Gamma \\ \tilde{b}(i,j,t=0) = U_0, \ \forall (i,j) \in \Gamma \end{cases}, \tag{3}$$

where U_0 and U_1 are positive parameters. These initial conditions can be physically seen as follows:

- at the initial time $t=0$ of the dispensing, there is no edge effect on the plate; and
- the effect strength is driven by a physical difference between the inside (Γ) and the outside ($\mathbb{Z}^2 \setminus \Gamma$) of the plate.

Note that Equation (3) can be spatially adapted with regard to the initial well dispensing and thus can take into account different spatial configurations of the wells (i.e. partial or heterogeneous dispensing). We then assume that plate background response fluorescence values can be accurately modelled by the function $\tilde{b}(i,j;\mathbf{P})$ with $\mathbf{P} = (t_{opt}, U_0, U_1)^T$. This triplet should be estimated using conventional optimization schemes. To properly model the bias, we propose to split the optimization framework into two consecutive operations: the first obtains the bias profile related to t_{opt}, and the second shifts and scales the data to properly fit the physical values acquired during the screen. In the following, we first describe the optimization process for the first parameter t_{opt}. For this application, we simulate a discrete version of the diffusion process based on the heat Equation (2), with the boundary conditions defined in Equation (3).

To numerically solve this partial differential, we use the classical approach based on finite differences by substituting the partial derivatives with truncated Taylor series approximations. Then, Equation (2), using the finite forward difference in time and the central difference in space (FTCS), can be written in a 2D discrete manner as follows:

$$\begin{aligned} \frac{\tilde{b}(i,j,t+k) - \tilde{b}(i,j,t)}{k} = & \frac{\tilde{b}(i+h,j,t) - 2\tilde{b}(i,j,t) + \tilde{b}(i-h,j,t)}{h^4} \\ & + \frac{\tilde{b}(i,j+h,t) - 2\tilde{b}(i,j,t) + \tilde{b}(i,j-h,t)}{h^4}, \end{aligned} \tag{4}$$

where k is the time step, and h is the spatial step (which here are horizontally and vertically identical). In our application, it is important to note that the minimum value of h is limited by the horizontal and vertical number of wells.

The time evolution of the bias generated by the model at location $(i,j) \in \Gamma$ can be written as follows:

$$\begin{aligned} \tilde{b}(i,j,t+k) = & \left(1 - 4\tfrac{k}{h^4}\right)\tilde{b}(i,j,t) + \\ & \tfrac{k}{h^4}\big(\tilde{b}(i-h,j,t) + \tilde{b}(i+h,j,t) \\ & + \tilde{b}(i,j-h,t) + \tilde{b}(i,j+h,t)\big). \end{aligned} \tag{5}$$

The solution of such a partial differential equation is known to be stable and convergent if the Courant–Friedrichs–Lewy condition (CFL condition) is respected (Courant *et al.*, 1967). This condition is particularly critical in

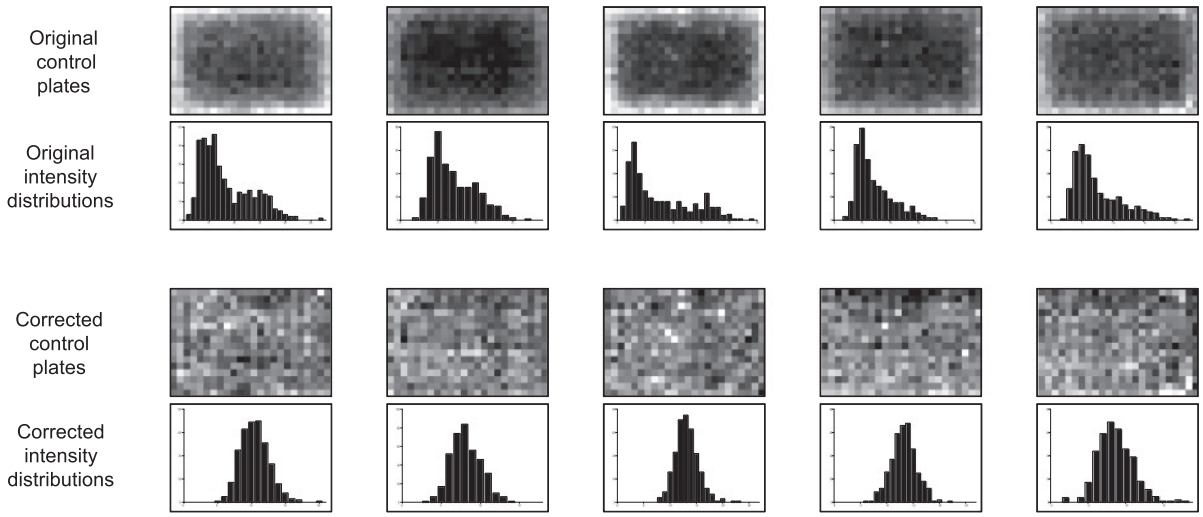

Fig. 3. Model validations. The figure displayed here is a validation of the correction process over the experimental control plates (Fig. 1a). The first two rows display five plates from the control set with their associated intensity histograms. The following two rows display the exact same plates after correction. Visually, the improvement is clear, as the correction removed the edge effect. Moreover, the distributions changed from heavy-tailed (or even bimodal) to standard Gaussian distributions.

our case, where the spatial resolution is low due to the large area of the wells compared with total area of the plate, which limits the Γ tessellation resolution.

Specifically, the iterative process described by Equation (5) is achieved by applying a discrete Laplacian operator over the image (see Supplementary Movie 1 for an example of plate-based diffusion process) and adding it to the previous solution, generating a set of 2D solutions over time [i.e. $\tilde{b}(i,j,t), \forall t \in \mathbb{R}^+$]. Among these solutions, one should correspond to the diffusion state of the considered plate $z'(i,j)$ (Supplementary Fig. S1). To find it, we must define a metric for evaluating the similarity between the model $\tilde{b}(i,j,t)$ and the considered plate $z'(i,j)$ (Fig. 1).

As the solutions of Equation (2) are related to the initial border value U_0, we consider them to be true as multiplicative and additive coefficients that are defined by the well-response values (here, fluorescence quantification). This assumption eliminates many non-normalized distances, such as the standard Euclidean distance. For the purposes of our application, we decided to use the standardized Euclidean distance ρ as a cost function. The optimization scheme can thus be written as follows:

$$\underset{t_{opt} \in \mathbb{R}^{+*}}{\operatorname{argmax}} \rho(z', \tilde{b}). \tag{6}$$

Thanks to the low dimensionality of the model, the optimization process can be achieved using a brute-force approach, avoiding local *maxima* results.

Once this parameter t_{opt} is obtained, we can consider the bias profile to be defined. Until this point, the physical values in the heat equation were normalized by simple practical boundary conditions (i.e. $U_1 = 1$ and $U_0 = 0$). To complete the process, we must estimate the last two parameters, U_1 and U_0, based on the dynamics of the plate-fluorescence responses. Basically, we must estimate the shift and stretch values of the modelled profile based on physical values. This is performed by fitting the model $\tilde{b}$ to the data z' using a L_2 distance function. This second optimization scheme can be written as follows:

$$\underset{U_0, U_1 \in \mathbb{R}^+}{\operatorname{argmax}} \left\| \tilde{b}_{t_{opt.}}(U_0, U_1) - z' \right\|_2 \tag{7}$$

This step is achieved by using a conjugated gradient algorithm, leading finally to the triplet $\mathbf{P} = (t_{opt}, U_0, U_1)^T$ characterizing the diffusion profile, i.e. the solution of the equation under the physical constraints dictated by the system.

The correction process is concluded by dividing the original signal z by the estimated bias $\tilde{b}$, leading to the corrected plate $\tilde{z}(i,j), (i,j) \in \Gamma$ (Supplementary Fig. S1). The results obtained within this framework tend to confirm that the assumptions used to model the bias found in these control plates using a diffusion equation are relevant. Figure 3 displays five control plates before and after correction with their associated readout values (i.e. intensities). A first visual inspection of the plates confirms the elimination of the bias, leading to a more evenly distributed intensity readout. Moreover, the central limit theorem states that the resulting distribution of the intensity average should approach the normal distribution, as the readouts are assumed to have independent and identical distributions. The original plates tend to display heavy-tailed distributions (at times even seeming bimodal), whereas the corrected plates present a normal intensity distribution. We can thus conclude that the correction process removed the spatial bias, leading to the expected probability distribution for the results. As our algorithm targets the edges effects, a test validating the presence of this artefact is required to provide an accurate correction. Such test can be performed either on fully dispensed plates prior or within the screen campaign. A relevant and regular way to assess the well-funded correction can be done by applying the correction algorithm to the defined plates followed by a normality test to the corrected readouts. Either a visual inspection of the bell-shaped data distribution or a graphical tool such as a normal probability plot can be performed. As the second option does not require a large amount of readouts, the linearity of the probability plot can also be assessed on the control column of each individual plate of the screen. Finally a pseudo code summarizing the correction method is provided to the user in Supplementary Material 4.

2.3 Model extrapolation: from control wells to entire plate

The method proposed in the previous section is based on a standard, fully dispensed plate defined by $\Gamma : [1, M] \times [1, N]$. In a screening context, those control plates are regularly acquired in order to assess the reproducibility of the readouts within a large set of experiments. These plates can be used to validate and extract the parameters of the diffusion model. However, our modelling allows us to restrain the domain Γ to a subset of wells $\Omega \subset \Gamma$ designated during assay development. Indeed, as explained in Section 1 and Supplementary Material 1, the readout values extracted from an siRNA

screen cannot be used for the direct modelling of the artefact: we must rely on a set of control wells that are specifically chosen for this purpose. Our method is not constrained to any specific pattern for the control (Supplementary Fig. S4), and the user can even choose an unconnected set of wells, as was proposed in (Zhang, 2008). In our case, each plate contains a negative control that was designed as a column at a specific and recurrent position on every plate.

The number and spatial configuration of the wells required to estimate the diffusion background is difficult to assess as it is related to the quality of the control experiments, their positions and their numbers. To evaluate the robustness of our methodology, allowing for an accurate extrapolation of the model from the controls to the complete plate, we simulated a set of experiments representing the standard conditions of an siRNA screening (Supplementary Fig. S6). Based on the diffusion model presented in Section 2.2.2, we generated three different artificial plates corresponding to three different diffusion times (i.e. incubation times). The background was then estimated using (i) the full plate or (ii) only the third column. Finally, the same experiment was performed using the same set of plates but corrupted by additional Gaussian noise, as defined in Equation (1) (Supplementary Fig. S6). The noise varies up to a peak signal-to-noise ratio (PSNR) of 19.48 dB, representing fairly poor plate conditions.

It appears that even under the worst signal-to-noise conditions, the use of only the third column in modelling the diffusion process is sufficient to provide accurate results similar to those provided by the processing of the full plate (Supplementary Fig. S3 for a more complete study of the column position impact on the reconstruction quality).

3 RESULTS AND DISCUSSION

3.1 HIV genome-wide siRNA screening: a case study

To identify host factors involved in the interactions with the HIV, we performed an RNAi screening of human cells transfected with a genome-spanning siRNA library and infected with HIV-1. The assay is described in more details in Supplementary Material 3.

Based on the means $(\bar{z}_p, \bar{z}_n)$ and standard deviations (SD_p, SD_n) of, respectively, the positive and negative controls, a Z' factor (Zhang *et al.*, 1999) was calculated for each plate n based on the following formula:

$$Z'_n = 1 - 3\frac{SD_p + SD_n}{|\bar{z}_p - \bar{z}_n|}. \quad (8)$$

A clear spatial artefact affected the infection burden, as increased fluorescence values of the reporter protein were obtained in the edge wells compared to those located at the centre of the plate (Fig. 4a). Intra- and inter-day reproducibility of the artefact was studied and revealed that increased infection values were consistently obtained in the edge wells (Fig. 5b). However, both the mean infection burden and the dynamic range of the variation between the lowest and highest values varied within each screening day and, more noticeably, between screening days.

Such an observation emphasizes the need for a correction method that is able to evaluate and correct for the bias within each plate individually and can also adapt to the dynamic range of both the infection and the artefact. Our normalization algorithm was adapted for automated correction and applied for the correction of the 68 assay plates in the screening. The correction method was very efficient in eradicating the edge effect, as homogeneous infection values were obtained throughout the columns containing positive and negative controls (Fig. 6A and data not shown). As a consequence, the assay window of detection was significantly improved. The average Z' factor was increased significantly

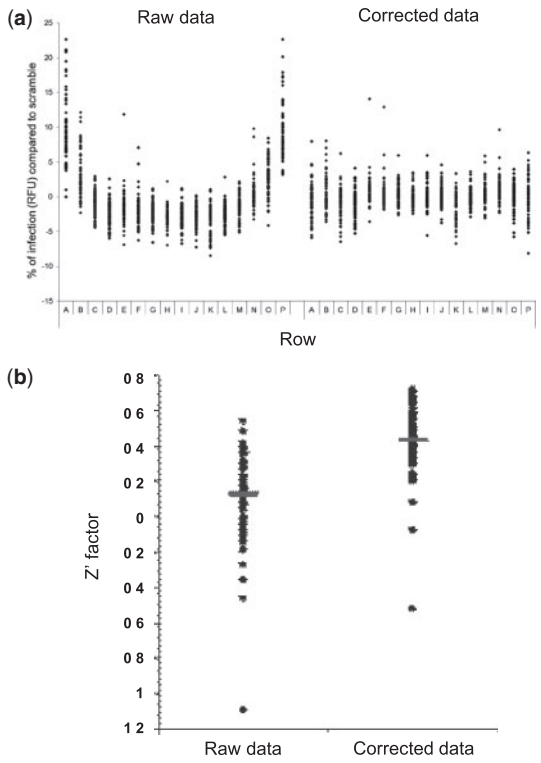

Fig. 4. Data correction eradicates the spatial bias and improves assay quality. (**a**) 3D representation of the infection values across Column 3 for the 68 assay plates of the genome-wide screening before and after correction. (**b**) Comparison of the Z' factor of each of the 68 plates before and after correction.

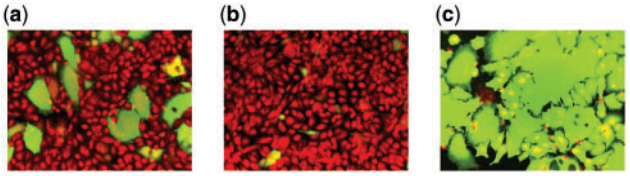

Fig. 5. Typical RNAi-induced phenotypes of HIV-infected cells. HeLa P4 LTR-EGFP 2B4 cells transfected with non-targeting siRNA (**a**) and with the siRNAs decreasing (**b**) or increasing (**c**) the infection burden. The confocal images were acquired on Day 5 after infection. The cells were counterstained with syto60 dye (red) and overexpressed EGFP reporter protein when infected (green).

($P < 0.0001$), from 0.12 (ranging from −1.09 to 0.54 with a median value of 0.17) to 0.43 (range −0.52 to 0.72, median value 0.47) when comparing the raw and corrected data. Notably, three aberrant plates that suffered from inaccurate cell dispensing and hence displayed very low Z' factors (values −1.1, −0.47 and −0.36) were not validated after correction (Z' factors values −0.52, −0.08 and −0.08). This illustrates the adaptability and the mildness of our algorithm, which prevents any overcorrection issues.

The standard score (or z-score) for each well was determined and used to select genes that were potentially promoting or restricting HIV infection. A z-score transformation for a specific readout $z_n(i,j)$

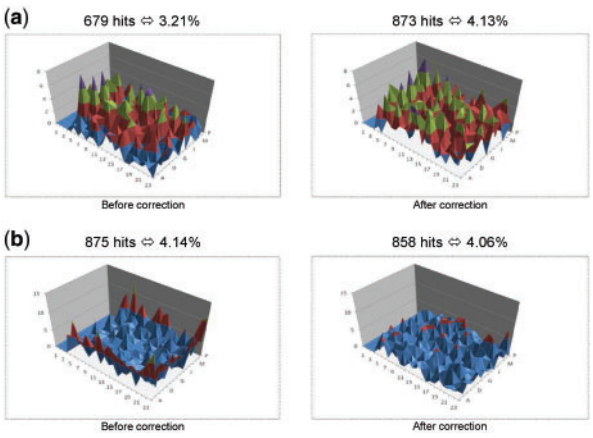

Fig. 6. Data correction improves the spatial distribution of hits. The spatial distribution of the selected hits decreasing (**a**) or increasing (**b**) HIV infection burden before (left images) and after (right images) data correction. In (a), the spatial bias led to underestimation of the number of hits located at the boundaries of the plates; they were overestimated in (b). The correction process tended to equalize the hit distribution over the plate, without preference for any area.

is based on the following formula:

$$z_n^*(i,j) = \frac{z_n(i,j) - \bar{z}}{SD}, \tag{9}$$

where $z_n^*(i,j)$ represents the normalized output value, $\bar{z}$ and SD, respectively, the readout mean and SD values over the complete screen. A negative z-score corresponds to the decrease or the blockage of HIV infection as compared with non-targeting control (Fig. 5a), suggesting that the targeted gene has a decreasing effect on an HIV burden (Fig. 5b). Pools were classified as host-dependency factor (HDF) positive if they decreased in both screens of EGFP fluorescence by at least 3 SD from the negative controls. In contrast, a positive z-score correlates with an increased HIV burden, suggesting that the corresponding gene is a host cell factor involved in the control of viral infection (Fig. 5c). Pools were classified as host-restriction factor (HRF) positive if they increased in both screens of EGFP fluorescence by at least 3 SD from the negative controls. As observed for the control columns (Figs 2 and 4a), the entire assays suffered from a dramatic spatial artefact, as infection burden was increased at the border of the plates (Fig. 6). Consequently, the selection of hits was strongly biased by this trend. Indeed, hits classified as HDF because they decreased the infection as compared to controls were preferentially located in the centres of the plates. Conversely, hits increasing the infection compared with the controls and classified as HRF occurred at the borders of the plates (Fig. 6). However, after plate correction, the selected hits were much more evenly distributed throughout the assay plate for both HDF and HRF populations (see right Fig. 6a and b, respectively). In addition, to select a comparable number of HRF hits, much more stringent selection criteria were needed to be applied after the correction, showing the overall assay detection window improvement.

4 CONCLUSIONS AND MODEL IMPROVEMENT

Due to the extended incubation period required for the knockdown procedure, RNAi-based HTS are likely subject to edge effects. Moreover, due to the wide range of phenotypes, these screens raise new challenges for the post-processing data normalization. We developed a novel algorithm that is able to predict the edges-related spatial bias for each plate individually using the data from a single control column. The methodology is based on a physical modelling of the bias, providing for its robust estimation. Supplementary Figures 7 assesses the robustness of our approach on unbiased data. Moreover, when the bias is associated to a diffusion process, the cost function displays a clear global minimum. This can be used to reject the diffusion model as the source of bias (Supplementary Fig. S2). Our correction was applied to the data from a genome-wide siRNA screen and was shown to significantly increase the overall assay quality and to improve the hit-selection process.

The constant development of RNAi screens and the emergence of experiments requiring extended incubation periods (e.g. stem cell differentiation) emphasize the need for a method of correcting recurring edge effects. Moreover, the trend towards assay miniaturization (for instance, plates containing 1536 samples) can also increase the evaporation impact. However, with such plates, and regarding the experiments, the cross-contamination artefact should be taken into account in the correction model.

In addition, such a correction method should be able to circumvent any bias that encountered in each library and/or biological process and be adaptable to each plate individually. Finally, our correction method could be made more robust in cases where the user has prior knowledge of a redundancy in the bias over a set of plates and can then average the control areas over those plates, further reducing the experimental noise.

ACKNOWLEDGMENTS

We gratefully acknowledge MEST Korea, Gyonggido and KISTI.

Funding: PB and JPC have received financial support from the Korea Research Foundation (Grant K204EA000001-07E0100-00100, K204EA000001-08E0100-00100, K204EA000001-09E0100-00100, K20802001409-09B1300-03700) and INSERM-Avenir. TD was partially supported by a grant from MEST 2011-0019430. We wish to thank Thierry Christophe, Chang Bok Lee and Natalie DeWitt for fruitful discussions and technical support.

Conflict of Interest: none declared.

REFERENCES

Birmingham,A. *et al.* (2009) Statistical methods for analysis of highthroughput rna interference screens. *Nat. Methods*, **6**, 569–575.

Brideau,C. *et al.* (2003) Improved statistical methods for hit selection in high-throughput screening. *J. Biomol. Screen.*, **8**, 634–647.

Courant,R. *et al.* (1967) On the partial difference equations of mathematical physics. *IBM J. Res. Development*, **11**, 215–234.

Dragiev,P. *et al.* (2011) Systematic error detection in experimental high-throughput screening. *BMC Bioinformatics*, **12**, 25.

Heyse,S. (2002) Comprehensive analysis of high-throughput screening data. In Bornhop,D.J. (eds) *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, vol. 4626 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*. SPIE, pp. 535–547.

Hüser,D.J. (ed.) (2006) *High-Throughput Screening in Drug Discovery*. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim.

Kevorkov,D. and Makarenkov,V. (2005) Statistical analysis of systematic errors in high-throughput screening. *J. Biomol. Screen.*, **10**, 557–567.

Lundholt,B.K. *et al.* (2003) A simple technique for reducing edge effect in cell-based assays. *J. Biomol. Screen.*, **8**, 566–570.

Makarenkov,V. *et al.* (2006) Hts-corrector: software for the statistical analysis and correction of experimental high-throughput screening data. *Bioinformatics*, **22**, 1408–1409.

Makarenkov,V. *et al.* (2007) An efficient method for the detection and elimination of systematic error in high-throughput screening. *Bioinformatics*, **23**, 1648–1657.

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

Ramadan,N. *et al.* (2007) Design and implementation of high-throughput RNAi screens in cultured Drosophila cells. *Nat. Protoc.*, **2**, 2245–2264.

Shun,T.Y. *et al.* (2011) Identifying actives from hts data sets: practical approaches for the selection of an appropriate hts data-processing method and quality control review. *J. Biomol. Screen.*, **16**, 1–14.

Zhang,X.D. (2008) Novel analytic criteria and effective plate designs for quality control in genome-scale RNAi screens. *J. Biomol. Screen.*, **13**, 363–377.

Zhang,X.D. (2011) *Optimal High-Throughput Screening: Practical Experimental Design and Data Analysis for Genome-Scale RNAi Research*. Cambridge University Press, New York, NY, USA.

Zhang,J. *et al.* (1999) A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screen.*, **4**, 67–73.