Bioinformatics, 32(13), 2016, 1959–1965
doi: 10.1093/bioinformatics/btw092
Advance Access Publication Date: 28 February 2016
Original Paper

OXFORD

Gene expression

Detection and removal of spatial bias in multiwell assays

Alexander Lachmann[1,2,3,†], Federico M. Giorgi[2,3,4,†], Mariano J. Alvarez[2] and Andrea Califano[1,2,3,5,6,7,*]

[1]Department of Biomedical Informatics (DBMI), [2]Department of Systems Biology, [3]Center for Computational Biology and Bioinformatics (C2B2), Columbia University, New York, NY, USA, [4]Scuola Superiore Sant'Anna, Pisa, Italy, [5]Department of Biochemistry and Molecular Biophysics, [6]Institute for Cancer Genetics and [7]Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
Associate Editor: Jonathan Wren

Received on August 6, 2015; revised on January 8, 2016; accepted on February 14, 2016

Abstract

Motivation: Multiplex readout assays are now increasingly being performed using microfluidic automation in multiwell format. For instance, the Library of Integrated Network-based Cellular Signatures (LINCS) has produced gene expression measurements for tens of thousands of distinct cell perturbations using a 384-well plate format. This dataset is by far the largest 384-well gene expression measurement assay ever performed. We investigated the gene expression profiles of a million samples from the LINCS dataset and found that the vast majority (96%) of the tested plates were affected by a significant 2D spatial bias.
Results: Using a novel algorithm combining spatial autocorrelation detection and principal component analysis, we could remove most of the spatial bias from the LINCS dataset and show in parallel a dramatic improvement of similarity between biological replicates assayed in different plates. The proposed methodology is fully general and can be applied to any highly multiplexed assay performed in multiwell format.
Contact: ac2248@columbia.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A revolution has occurred in the field of Quantitative Transcriptomics in the last 15 years, fueled by a significant drop in the cost of technologies for multiple gene expression profiling (Montgomery and Dermitzakis, 2011). Such technologies, namely gene microarrays (Hertzberg and Pope, 2000) and RNASeq (Wang *et al.*, 2009), are now standard procedures in biological and medical research, and they are becoming increasingly more popular as robust research and diagnostic tools in all biological fields, spanning from crop research (Hansey *et al.*, 2012) to drug discovery (Jenkins and Ma'ayan, 2013) and from microbiology (Westermann *et al.*, 2012) to personalized medicine (Derks and Diosdado, 2015). Recently, both gene expression assays and other multiplexed readout assays have been implemented in a multiwell format (microplate), lending themselves to significant microfluidic automation and scale-up. This type of experiments lend themselves to detection and normalization of a variety of potential bias sources, including effects representing 2D spatial bias introduced by the specific microfluidics apparatus. For instance, multiplex gene expression has been recently used as a reporter for large-scale cell perturbation assays, allowing for the quantitative characterization of the corresponding perturbagens as inducers of specific transcriptional responses at the molecular level (molecular phenotypes). This new class of studies, named gene expression high throughput screening (GE-HTS), promise to have significant impact on biomedical research (Ma'ayan *et al.*, 2014), and have already shown their potential to accelerate drug discovery for

human diseases, such as for leukemia (Stegmaier *et al.*, 2004) and muscle atrophy (Kunkel *et al.*, 2012). Detecting and correcting any systematic bias generated by such large-scale assays is obviously of paramount importance to maximize the value of these studies. Principal component analysis for bias removal has been applied successfully to increase the ability to increase power in eQTL identification (Fehrmann *et al.*, 2015). Spatial biases in the probe intensity levels, commonly observed on oligonucleotide arrays, are removed by normalization techniques such as GC-RMA (Wu *et al.*, 2004) or LOESS normalization (Smyth and Speed, 2003). The largest dataset of this kind to date, has been generated by the Library of Integrated Network-based Cellular Signatures (LINCS) effort (Vempati *et al.*, 2014). It represents a direct extension of the Connectivity Map (CMAP), the first large-scale GE-HTS study (Lamb *et al.*, 2006). LINCS adopted an innovative approach to measuring gene expression at an ultra-low cost, based on medium-throughput profiling of 978 Landmark (L1000) genes across a consistent number of perturbations and cell lines (Duan *et al.*, 2014). The rest of the transcriptome (20 000 genes) is then inferred by a mathematical model built on top of thousands of gene expression measurements from GEO (Barrett *et al.*, 2013). The LINCS dataset collects roughly a million experiments where different cell types (primary and transformed human cell lines) are chemically (by small molecule compounds) and genetically (by shRNA knock-down or cDNA over-expression) perturbed (Liu *et al.*, 2015). The LINCS dataset was conducted on 384-well (24 × 16 format) microplates and is so far the biggest microplate-based experiment ever produced. As such, it offers an unprecedented opportunity to both detect and normalize bias introduced by the experimental setup.

2 Methods

2.1 LINCS dataset

The normalized LINCS dataset was obtained from the LINCS consortium (http://support.lincscloud.org). Specifically, the dataset analyzed in the current manuscript is composed of a total of 1894 microplates of which 963 are perturbed by targeted gene Knock-Downs (KD), 747 by Chemical/Pharmaceutical Compounds (CPC), and 184 by gene Over-Expression (OE). In total, the dataset we analyzed is composed of eight different cell lines (PC3, MCF7, VCAP, HT29, HA1E, A375, HCC515, and A549). The total number of different experiments included in the analysis was comprised of 685 612 gene expression profiles. Following the L1000 procedure (Duan *et al.*, 2014), a total of 978 genes have a direct measurement (landmark genes) while for 22 000 more genes the transcript amounts were inferred.

2.2 Bias detection

In order to detect 2D spatial bias, we formulated the problem through a spatial autocorrelation framework. Spatial autocorrelation is a 2D space measure to assess how variables at neighboring locations in 2D space co-vary (Dale and Fortin, 2002). The most common way to measure spatial correlation is by calculating Moran's I (Moran, 1950), which measures the correlation of signal amongst spatial neighbors (Equation 1):

$$I(X,T) = \frac{N}{\sum_i \sum_j t_{ij}} \frac{\sum_i \sum_j t_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2} \quad (1)$$

where N is the number of samples in an array (e.g. a microplate), each X is a variable (gene), $\bar{X}$ is the mean of X, and T_{ij} (plate row i and plate column j) is the topology encoded as a weight matrix (i.e. which wells are neighbors). T_{ij} is a symmetric $N \times N$ matrix with $t_{ij} \in \{0, 1\}$. t_{ij} is 1 if well i is adjacent to well j and 0 otherwise. Like Pearson's correlation, the values for Moran's I can range from 1 (where the values of the variables are completely correlated to the values of their neighbors) to −1 (a checkerboard-like scenario, where every sample is inversely correlated to its neighbors).

2.3 Bias removal

The main assumption in the spatial bias removal is the redundancy of biases across multiple genes. With redundancy we specifically mean that a bias affects all genes with to a certain degree. The observed gene expression $\hat{e}$ for a gene is composed of the true biological signal e and additive bias components. For a set of genes size G and well (i, j) on a given plate with a set of M biases $\psi_m(i,j)$ of Ψ and the corresponding coefficient matrix $C \in G \times M$, the observed gene expression can be noted as:

$$\hat{e}(g,i,j) = e(g,i,j) + \sum_{m=1}^{M} c_{gm} \times \psi_m(i,j) \quad (2)$$

$e(g,i,j)$ is the true biological gene expression in well (i, j) when all biases are removed from the observed gene expression. $\psi_m(i,j) \in \Psi$ is a function, representing a bias and Ψ is the set of all biases that are defined over the coordinates of the plates. In our approach we specifically search for the bias functions in Ψ that result in significant spatial autocorrelation. The coefficients $c_{gm} \in C$ are defining the impact of the bias n on gene g. The coefficient c_{gm} stays constant for gene g and bias m is not depended on well coordinate (i, j). A larger absolute coefficient represents a stronger effect of the bias on the gene expression. In return the coefficient indicates how much it contributes to the bias with respect to the other genes. For any of the bias functions we assume that multiple genes are contributing, resulting in non-zero coefficients for this bias. This redundancy is a key attribute of the biases, enabling the proposed solution of the problem.

To account for spatial biases, we first decompose the expression data into the major orthogonal components of variance by Singular Value Decomposition (SVD) (Golub and Reinsch, 1970; Alter *et al.*, 2000):

$$X = UDV^t \quad (3)$$

where X is the G-gene $\times N$-sample matrix ($N = I \times J$, I rows and J columns of the plate) of expression values, D is a diagonal matrix with the positive singular values $d_1 > d_2 > \ldots > d_r > 0$ on its diagonal and 0 for all other entries, with $r = min(G; N)$, and U and V are the orthogonal left- and right-singular vectors, which represent the basis, eigensamples × genes and eigengenes × samples square matrices, respectively. We compute the principal components on of X^t using the R function *prcomp*, resulting in PCs that are linear combinations of the genes in X. The output is $C = UD \in G \times r$, containing the scaled eigenvectors and a square $N \times r$ matrix V containing the principal component values. The columns of C contain the coefficients describing the linear combination of genes for the corresponding PC. The PC values in V are the expression values of the eigengenes. The original gene expression can be back transformed through $X = CV^t$.

Due to the redundancy assumption we expect to find eigengenes that describe bias functions of Ψ. Each column of V represents a PC and the rows represent samples/wells of the plate. As such we calculate the Moran's I spatial autocorrelation for each column of V by mapping its values to the corresponding plate coordinate. If a principal component shows a significant spatial autocorrelation the PC represents a function of Ψ as described above.

The autocorrelation score is transformed into a weight determining how much the principal component should be removed from the data. The weight function is defined as $W(x,\sigma,\beta) = 1 - 1/(1 + e^{-\beta(x-\sigma)})$ for autocorrelation score x, β defining the climb of W and σ for the function offset. σ defines in what range the phase transition from 0 to 1 occurs and β how quickly (Supplementary Fig. 1). For the LINCS data we choose σ that maximizes the average increase in RS for the three separate datasets KD, OE and CPC (see Supplementary Methods, Section 6).

From V we compute vector $\vec{w}$ of length m by converting the autocorrelation scores x to weights by the weight function $W(x,\sigma,\beta) = \vec{w}$. The higher the autocorrelation of the column in V the lower the weight applied to it. The function is sigmoidal and ranges from 0 to 1. Unbiased PCs will receive a weight of about 1. We compute the down-weighted PC value matrix V^* from V by $V^* = V \times diag(\vec{w})$, with matrix $diag(\vec{w})$ of size $m \times m$ with vector $\vec{w}$ on the diagonal and 0 otherwise. We can build a new gene expression matrix $X^* = CV^{*t}$ with matrix V^* from which the biased principal components are removed. The procedure is shown in Supplementary Algorithm 2 and 3.

2.4 Reproducibility score

The LINCS data is organized in batches of replicate plates. There are 2–5 plates in each batch in which the cell line, plate layout and measurement time are identical. The LINCS data has no replicates on the same plate, instead the whole plate is replicated. The reproducibility score (RS) is a measure as to how similar replicates are across the plates of a batch compared to all other perturbations. Each well of a batch with N plates can be defined by $i \in \{A,\ldots,P\}$ and $j \in \{1,\ldots,24\}$ and plate index $p,k \in \{1,\ldots,N\}$. X is the z-score normalized gene expression matrix of all samples in the batch. $Xijk$ is the sample on plate k at well (i, j). We define the unnormalized replicate score (RS*) as follows:

$$RS^*(X_{ijk}) = \sum_{p \neq k}^{N} cor(X_{ijk}, X_{ijp})/(N-1) \quad (4)$$

To normalize RS^* we build a null model for each individual sample of the batch. For this we select one sample from each of the other plates and re-compute the average correlation. We calculate a set of 1000 random RS values and fit a normal distribution $N(\mu,\sigma^2)$. The normalized reproducibility score is calculated by $RS = RS^*/\sigma$. We define $\Delta RS(X,X^*) = RS(X^*) - RS(X)$ for uncorrected gene expression X and corrected gene expression X^*. Positive ΔRS mean increased and negative decreased in reproducibility.

2.5 Removal of random principal components

We developed a naive method or removing random PC from the data. For this, we apply the bias correction on all plates of all three datasets (KD, OE and CPC). The naive method calculates the principal component decomposition $X = UDV^t$ for each plate analog to our method. The singular values on the diagonal of D indicate the amount of variance encoded in the PCs. We iteratively set a diagonal entry of D to zero until we removed the same amount of variance from the plate as the bias removal resulting in D^*. As not the exact amount of variance can be removed unless we remove the same PCs we allow an error of $\pm\ \epsilon$, for $\epsilon = 0.005$ of the total variance. Setting entries of D to zero is identical as setting a column of V to zero. $X^R = UD^*V^t$ is the residual gene expression and we can compute the reproducibility scores for all batches for X^* and X^R with $[var(X^*) - var(X^R)]/var(X) = +/-\epsilon$.

2.6 Gene set enrichment analysis

We performed Gene Set Enrichment Analysis (GSEA, (Subramanian *et al.*, 2005)) between the values of each PC heavily affected by bias in the LINCS dataset (weight ≥ 0.5) and all Gene Ontology (GO) Biological Processes (GeneOntologyConsortium, 2013), as provided by the MSigDB database (Liberzon *et al.*, 2011). P-values were corrected according to the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). The normalized enrichment scores of the GO terms were integrated using the Stouffer procedure (Stouffer *et al.*, 1949), and the 50 terms showing the highest integrated enrichment across the dataset are shown in Figure 4.

3 Results

3.1 Detection of spatial bias in gene expression

Consistent spatial bias affecting gene expression profiles was identified in 1825 out of 1894 (96%) of the tested microplates of the LINCS dataset, as evident from observing landmark gene expression profiles on a display structure that recreates the original microplate sample positioning (Fig. 1). We could detect apparent effects on gene expression that were dependent on the sample location on the microplate, which supports the hypothesis that cells in the microplate suffered from a non-controlled gradient of external perturbation. Since the visual inspection of gene expression profiles offered such a striking display of 2D bias, we measured its effects across the dataset in terms of spatial autocorrelation. The observed autocorrelation between samples in the LINCS dataset was significantly higher than expected (Wilcoxon test $P < 2.2 \times 10^{-16}$), indicating that the location of a sample on a microplate partially determines the detected transcript abundance of landmark genes (Fig. 2). This is evident in all the three experiment subsets (KD, OE, CPC). We then applied Principal Component Analysis (PCA) on the entire dataset and decomposed the gene expression into $X = UDV^t$. We calculate the autocorrelation for the columns of V analogue to the gene spatial autocorrelation. We detected that most plates have Principal Components (PCs) with significant autocorrelation. 2445 PCs were detected with autocorrelation ≥ 0.5 (See Supplementary Methods for further details).

3.2 Bias correction improves reproducibility score

We applied the bias correction on all KD, OE and CPC plates and calculated RS before and after correction. Strict σ values can almost entirely negate the bias from the transformed matrix (Supplementary Fig. 2); however, we noticed that this has a negative impact on the experimental quality of the corrected data in terms of agreement between biological replicates (Supplementary Fig. 3). The KD subset achieved the best results with a more stringent correction ($\sigma = 0.15$) than the other two subsets ($\sigma = 0.35$). We then applied these σ values to each LINCS subset and used them to correct the entire data matrix to obtain a consistent reduction in spatial bias (Fig. 2). Globally, the dataset presented an average of 3.58 significantly biased PCs per plate with a weight ≥ 0.5 (i.e. removal of more than 50% of their contribution to gene expression). In terms of total variance, the autocorrelation bias accounted for 22% of the gene expression variability observed in the dataset. Removing spatial bias allowed us to significantly improve the agreement between replicated conditions between plates. Our procedure significantly improved the capability of clustering biological replicates located on different microplates (ΔRS mostly positive, Wilcoxon Test $p = 10^{-100}$). We observed the highest ΔRS in the KD subset (Fig. 3). Similar results were observed in the CPC

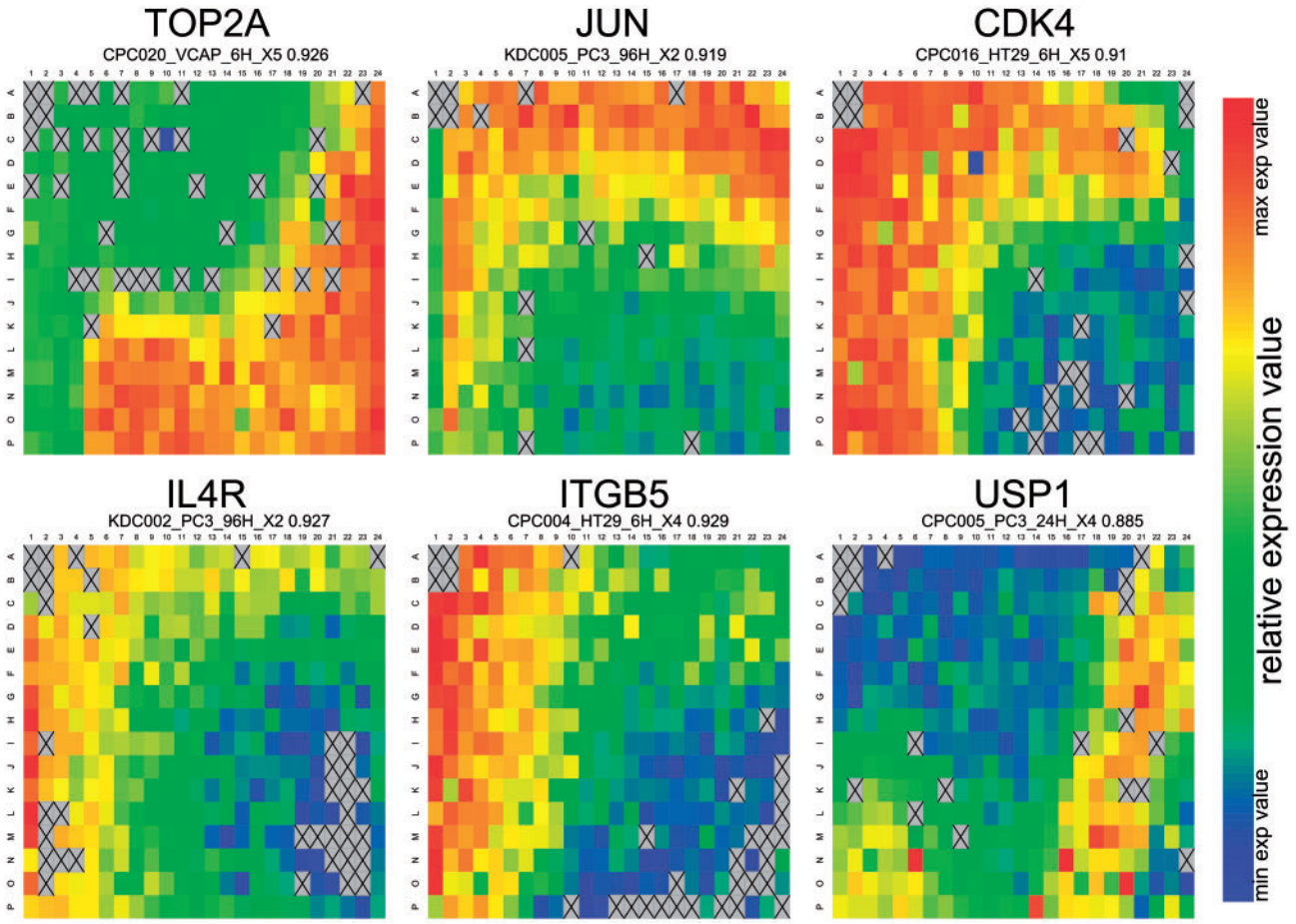

Fig. 1. Spatial bias affecting gene expression levels. The heatmaps show the most affected genes in each of the six LINCS plates shown. The columns of the plates are labeled from 1 to 24 and the rows from A to P. Plates contain 384 unique perturbations and no replicates. The color scale reflects the differential gene expression (Duan *et al.*, 2014). Grey wells marked with X are experiments for which no gene expression was available

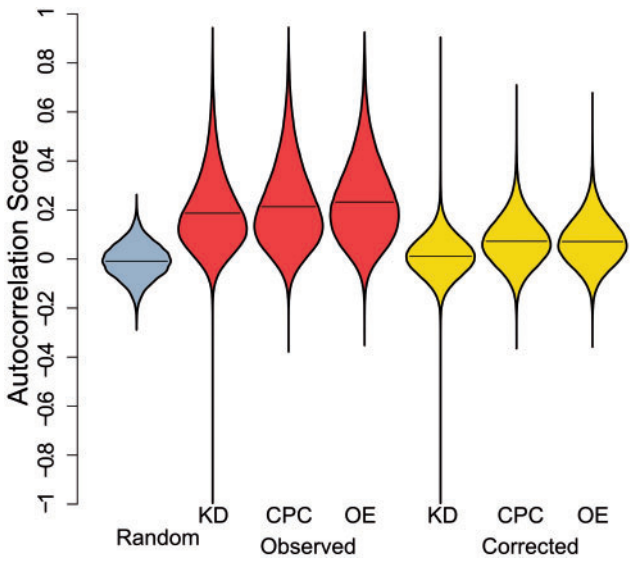

Fig. 2. Systematic spatial bias affecting the LINCS dataset. The violin plots show the probability distribution for the autocorrelation score (Moran's I) of a bias-free spatial location on a 24 × 16 microplate (Random, shown in blue); for each the three uncorrected LINCS subsets (KD, CPC and OE, shown in red); and for their corresponding corrected versions (shown in yellow). LINCS subsets were corrected using $\sigma = 0.15$ for KD experiments, and $\sigma = 0.35$ for CPC and OE experiments

subset (Supplementary Fig. 4) and in the OE subset (Supplementary Fig. 5), however with weaker effects, possibly due to these parts of the dataset generally being less responsive to perturbations (Supplementary Fig. 6). To test this hypothesis we compared the strength of KD and OE perturbation with ΔRS. CPC samples do not have a direct measure of perturbation effect. For knockdown samples we show that the fold change of target genes negatively correlates with RS while in OE the correlation is positive (Supplementary Materials). We used signature strength (absolute sum of z-score values of a sample) and correlated it with ΔRS. We could show a correlation between signature strength and ΔRS in all three datasets (Supplementary Materials). In total, 85% of the samples showing significant reproducibility before spatial bias removal (see Supplementary Materials) had improved RS when compared to the original expression matrix. In CPC, the average improvement is not as drastic as in the other two subsets with individual percentages of improved replicate scores at 89% for KD, 81% for OE, and 68% for CPC.

Removing spatially biased variance, as proposed, thus clearly improves reproducibility; however, it could be argued that any method that reduces data variance may produce similar results. Therefore, we show that removal of random principal components has no positive systematic effect on the reproducibility score. Specifically, we applied a naive method based on the removal of non-statistically significant principal components from the dataset. The autocorrelation based

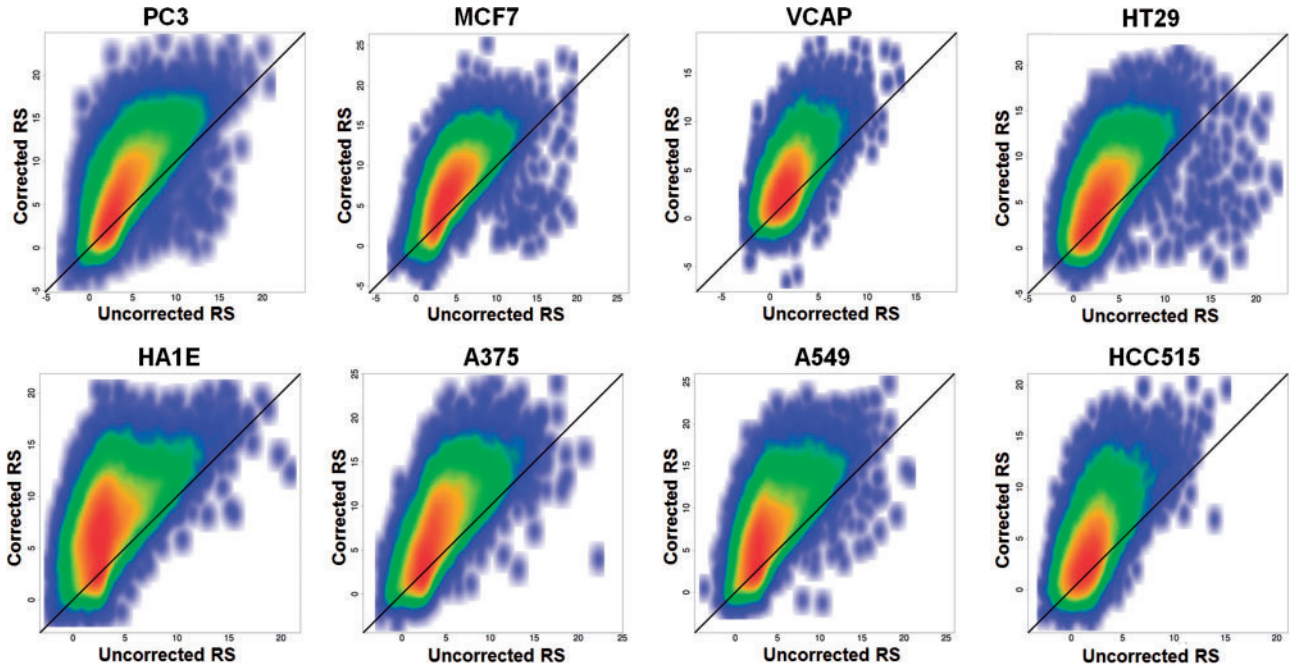

Fig. 3. Effect of spatial bias correction on the reproducibility score (RS). Scatter plot showing RS values before (*x*-axis) and after (*y*-axis) spatial bias correction for the KD subset in 8 cell lines. The colors indicate a Gaussian kernel estimation for the joint probability density, from blue (low density) to red (high density). The diagonal line indicates the points where the correction had no effect in terms of RS

Table 1. Top 10 biased biological processes (mean autocorrelation) across all plates in LINCS

Process name	avg AC
Cation homeostasis	0.287
Cellular cation homeostasis	0.287
Chemical homeostasis	0.285
Nuclear organization and biogenesis	0.284
Apoptotic nuclear changes	0.284
Ion homeostasis	0.279
Cellular homeostasis	0.279
Energy derivation by oxidation of organic compounds	0.278
Coenzyme metabolic process	0.277
Homeostatic process	0.273

method significantly outperformed removal of equivalent amounts of variance from gene expression data using principal components, when tested for overall replicate agreement in the three datasets (KD, OE and CPC) (Supplementary Fig. 7, Supplementary Algorithm 4). The left three panels show the distribution of fractions # positive ΔRS/# ΔRS over all batches. In OE and CPC the random removal of PCs resulted in and average decrease of RS in all batches (fraction $\leq$ 0.5). In KD most batches also show a decrease in reproducibility after correction, with only few plates showing a fraction of $\geq$ 0.5.

3.3 Biased biological processes

To test whether the bias in gene expression affects enrichment analysis we apply GSEA (Subramanian *et al.*, 2005) on the individual samples of a plate and calculate the autocorrelation of z-scores of each biological process in MsigDB. After correction the autocorrelation of enrichment scores is reduced in all three datasets. We calculate enrichment scores over the landmark gene signatures for all plates for 323 biological processes from MsigDB with at least 5 landmark genes and infer the autocorrelation score. Similar to the

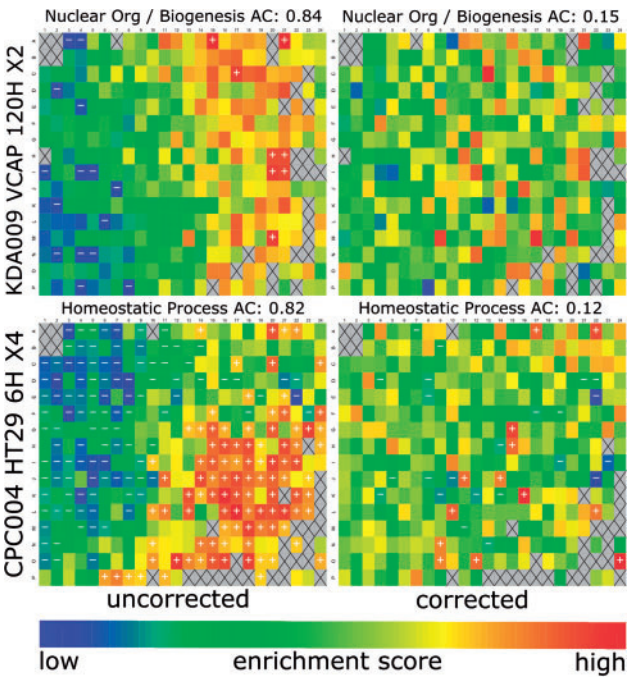

Fig. 4. Enrichment of two commonly biased biological processes for two example plates. + marked wells indicate significant up-regulation and - significant down-regulation. The left shows enrichment z-scores for uncorrected and the right bias corrected plates. AC is the autocorrelation score for the enrichment scores. Wells marked with X are experiments for which gene expression was not available

biases observed in individual genes biological processes are biased by plate location (Supplementary Fig. 8). Processes sorted by average autocorrelation score are shown in Table 1 (for a full list see Supplementary Table 3). The processes are related to homeostasis and apoptosis suggesting that cells in different plate locations

encounter varying growth conditions resulting in gene expression changes. As shown in Figure 4 the location of the sample on the plate affects enrichment analysis. Significant down and upregulation in uncorrected gene expression is linked to well coordinate. After correction the biases are resolved. Negatively and positively enriched samples are not spatially separated and there are generally fewer significant enrichments observed.

Finally, in order to get additional insights about the source of the spatial bias, we performed the enrichment of functional gene sets on the PCs significantly associated to the spatial bias (with a correction weight ≥ 0.5) using Gene Set Enrichment Analysis. We noted a few recurring Gene Ontology Biological Processes (GeneOntology Consortium, 2013) that were significantly associated to the biased components across the entire dataset (Fig. 5). Specifically we use the columns of the coefficient matrix C from the singular value decomposition as signatures for the enrichment analysis (see Supplementary Methods). Although not extraordinarily significant (with adjusted p-values reaching a minimum value 10^{-7}), the enriched components seem to circumscribe the source of bias into two major groups inversely correlated to each other and characterized by alterations in (i) cell proliferation and (ii) ion channel activity. Furthermore, similar PCs could be observed across distinct microplates and cell lines, indicating that spatial bias affects similar biological processes across different plates and cell lines.

4 Discussion

The overarching achievement of transcriptomics is the ability to obtain large quantities of highly reliable gene expression data representing multiple cellular states. Thus, as dataset size increases with microfluidic automation, computational approaches for data processing are becoming more and more important. This study shows that 2D spatial bias, likely associated with gradient driven differential conditions affecting the microfluidic setup, is a significant source of error that can be effectively detected and corrected. Previous studies have dealt with similar problems in contexts other than that of gene expression measurement; however, these studies simply evaluated the issue without proposing a solution (Harrison and Hammock, 1988) or proposed a solution that involved changing the design layout of the microplates (Liang *et al.*, 2013). In other cases, the systematic bias was observed in high-throughput screening technologies with a reduced set of experimental variables (Caraus *et al.*, 2015). No method to our knowledge addresses the spatial bias problem in the context of multivariate assay scenarios, like the LINCS dataset.

Therefore, we propose here a pure data-driven solution for multivariate gene expression datasets based on reducing the principal components of the variance most associated with spatial bias. We introduce a tunable parameter that can either be set a priori or can be assessed to optimize biological replicate agreement in a large

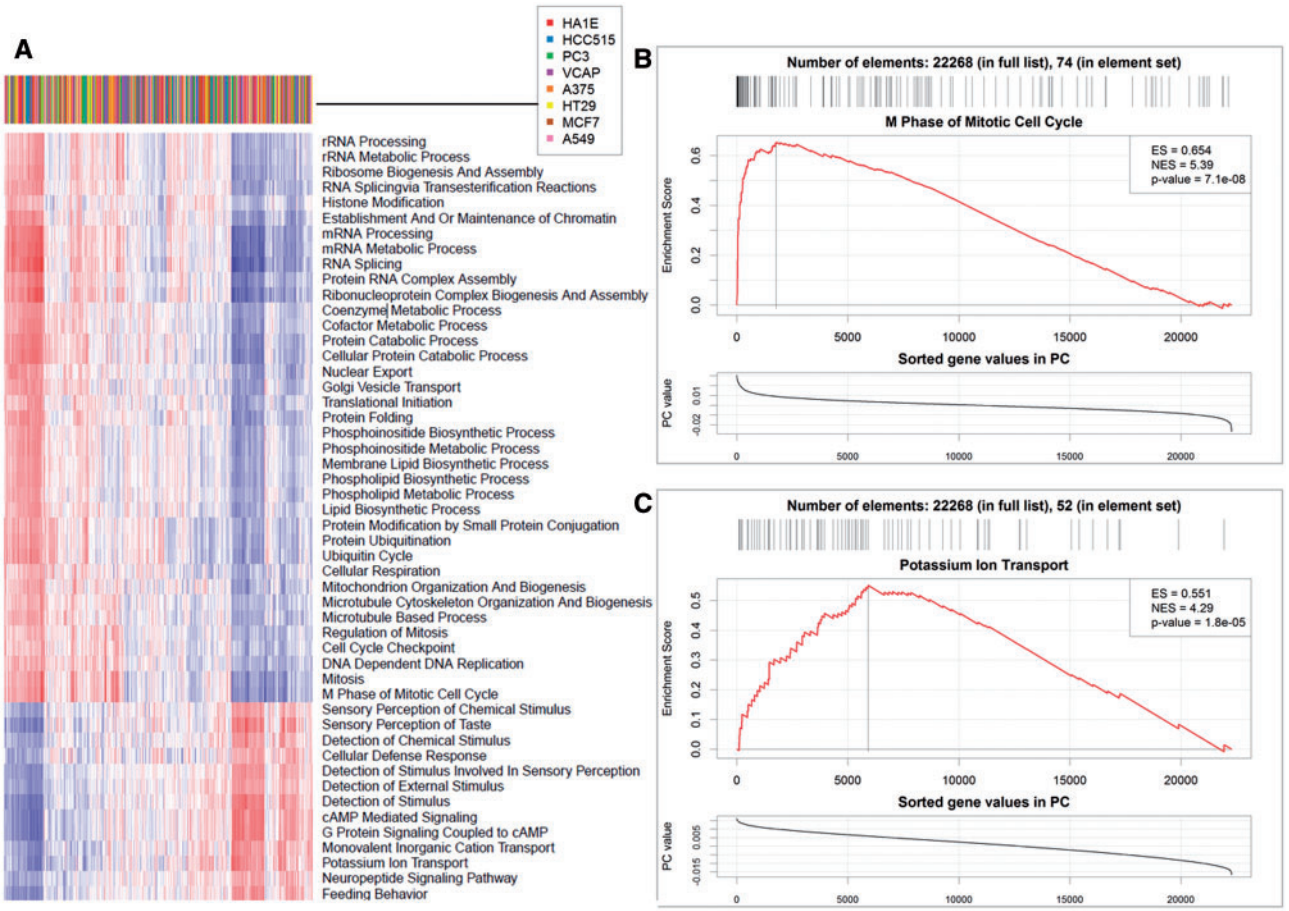

Fig. 5. Association between Gene Ontology Biological Processes (GO-BP) and the Principal Components capturing most of the spatial bias. (A) Heatmap showing the enrichment in GO-BP gene sets (rows), expressed as relative normalized enrichment score (NES), for each of the 2,445 PCs (columns) most affected by spatial bias (weight ≥ 0.5). Hierarchical clustering was performed using Euclidean distance and the Wards method. The eight cell lines included in the analysis are indicated by the color scale on top of the heatmap. (B and C) Gene Set Enrichment Analysis for the most enriched GO-BP term in the upper cluster (B) and in the lower cluster (C) in their respectively most co-segregating PCs

multiplate assay. Our solution significantly improved replicate similarity in the LINCS dataset, an effect that could not be achieved by simply removing equivalent levels of global and non-specific variance from gene expression profiles. The optimal parameters differed between specific portions of the LINCS dataset, but every subset analyzed benefited from the bias removal independently of the analyzed cell line. In fact, these results strongly support the incorporation of spatial bias analysis and removal as a critical step in the normalization of all GE-HTS assays. While spatial bias can occur both before and after cell lysis and nucleic acids extraction presence of spatial bias in biological process enrichment values on most plates and consistent enrichment of a reduced set of biological processes on the most biased principal components suggests that part of such effects are originated prior to cell lysis. Processes most strongly biased on average across all plates in KD, OE, and CPC were associated to homeostasis and apoptosis. Our analysis of the principal component coefficients shows differential cell proliferation associated exclusively with the sample positioning on the microplates, a recurring effect that seems to be independent from the microplate or the cell line affected by the bias. This difference in cell proliferation can be caused by differential growth conditions of the plated cells, perhaps due to a gradient in the evaporation of the medium partially connected to the parallel enrichment of Potassium Channel activity (Abdul *et al.*, 2002). No matter how carefully monitored and executed, any GE-HTS experiment can suffer from such operational bias issues. The proposed method allows for the detection and correction of the most significant location-specific undesired perturbation effects, regardless of the experimental stage at which they were introduced. Given that the LINCS dataset is the first of its kind, such spatial bias in multivariate gene expression data has never been observed at this scale before. Our method is similar to PCA-based noise removal techniques (Thomas *et al.*, 2002) since it operates on the Principal Component space in order to improve the measurement reliability. However, while previous methods remove the low variance components to reduce random noise (Liebermeister, 2002), our method is the first that specifically captures location-dependent effects, and decreases their contribution to the corrected dataset.

Acknowledgements

We thank Kristin M. Beiswenger and Mahalaxmi Aburi for the help during manuscript preparation.

Funding

This work was supported by the LINCS grant for designing new computational tools.

Conflict of Interest: none declared.

References

Abdul,M. *et al.* (2002) Activity of potassium channel-blockers in breast cancer. *Anticancer Res.*, **23**, 3347–3351.

Alter,O. *et al.* (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci.*, **97**, 10101–10106.

Barrett,T. *et al.* (2013) Ncbi geo: archive for functional genomics data sets update. *Nucleic Acids Res.*, **41**, D991–D995.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B (Methodological)*, 289–300.

Caraus, I. *et al.* (2015) Detecting and overcoming systematic bias in high-throughput screening technologies: a comprehensive review of practical issues and methodological solutions. *Brief Bioinform.*, **16**, 974–986.

Dale,M.R. and Fortin,M.J. (2002) Spatial autocorrelation and statistical tests in ecology. *Ecoscience*, 162–167.

Derks,S. and Diosdado,B. (2015) Personalized cancer medicine: next steps in the genomic era. *Cell. Oncol.*, **38**, 1–2.

Duan,Q. *et al.* (2014) Lincs canvas browser: interactive web app to query, browse and interrogate lincs l1000 gene expression signatures. *Nucleic Acids Res.*, **42**, W449–W460.

Fehrmann,R.S. *et al.* (2015) Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.*, **47**, 115–125.

GeneOntologyConsortium (2013) Gene ontology annotations and resources. *Nucleic Acids Res.*, **41**, D530–D535.

Golub,G.H. and Reinsch,C. (1970) Singular value decomposition and least squares solutions. *Numerische Mathematik*, **14**, 403–420.

Hansey,C.N. *et al.* (2012) Maize (zea mays l.) genome diversity as revealed by rna-sequencing. *PLoS One*, **7**, e33071.

Harrison,R.O. and Hammock,B.D. (1988) Location dependent biases in automatic 96-well microplate readers. *J. Assoc. Off. Anal. Chem.*, **71**, 981–987.

Hertzberg,R.P. and Pope,A.J. (2000) High-throughput screening: new technology for the 21st century. *Curr. Opin. Chem. Biol.*, **4**, 445–451.

Jenkins,S.L. and Ma'ayan,A. (2013) Systems pharmacology meets predictive, preventive, personalized and participatory medicine. *Pharmacogenomics*, **14**, 119.

Kunkel,S.D. *et al.* (2012) Ursolic acid increases skeletal muscle and brown fat and decreases diet-induced obesity, glucose intolerance and fatty liver disease. *PloS One*, **7**, e39332.

Lamb,J. *et al.* (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Liang,Y. *et al.* (2013) Correction of microplate location effects improves performance of the thrombin generation test. *Thrombosis J.*, **11**, 12.

Liberzon,A. *et al.* (2011) Molecular signatures database (msigdb) 3.0. *Bioinformatics*, **27**, 1739–1740.

Liebermeister,W. (2002) Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, **18**, 51–60.

Liu, C. *et al.* (2015) Compound signature detection on lincs l1000 big data. *Mol. BioSyst.*, **11**, 714–722.

Ma'ayan,A. *et al.* (2014) Lean big data integration in systems biology and systems pharmacology. *Trends Pharmacol. Sci.*, **35**, 450–460.

Montgomery,S.B. and Dermitzakis,E.T. (2011) From expression qtls to personalized transcriptomics. *Nat. Rev. Genet.*, **12**, 277–282.

Moran,P.A. (1950) Notes on continuous stochastic phenomena. *Biometrika*, 17–23.

Smyth,G.K. and Speed,T. (2003) Normalization of cdna microarray data. *Methods*, **31**, 265–273.

Stegmaier,K. *et al.* (2004) Gene expression based high-throughput screening (ge-hts) and application to leukemia differentiation. *Nat. Genetics*, **36**, 257–263.

Stouffer,S.A. *et al.* (1949). The American soldier: adjustment during army life.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Thomas,C.G. *et al.* (2002) Noise reduction in bold-based fmri using component analysis. *Neuroimage*, **17**, 1521–1537.

Vempati,U.D. *et al.* (2014) Metadata standard and data exchange specifications to describe, model, and integrate complex and diverse high-throughput screening data from the library of integrated network-based cellular signatures (lincs). *J. Biomol. Screen.*, **19**, 803–816.

Wang,Z. *et al.* (2009) Rna-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genetics*, **10**, 57–63.

Westermann,A.J. *et al.* (2012) Dual rna-seq of pathogen and host. *Nat. Rev. Microbiol.*, **10**, 618–630.

Wu,Z. *et al.* (2004) A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.*, **99**, 909–917.