Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Subject Section

Blind estimation and correction of microarray batch effect

Sudhir Varma[1,*]

[1]HiThru Analytics, Princeton NJ

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Batch effect (BE) has been the primary bottleneck for large-scale integration of data from multiple experiments. Current BE correction methods either need known batch identities (*ComBat*) or remove unknown true biological differences (*SVA*). Almost all of them also need to be re-computed for each new dataset. Even though the effects of technical differences on measured expression have been published, there are no BE correction algorithms that predict potential batch effects.

Results: We show that the overall BE can be decomposed into distinct Batch Effect Signatures (BES) that capture the possible directions of perturbation in the measured gene expression caused by BE. We select a Reference set of samples designed to eliminate biological differences to estimate the BES. We introduce an algorithm Batch Effect Signature Correction (*BESC*) that uses the BES calculated on the Reference set to efficiently predict and remove BE on two independent Validation sets. The correction is *blind*, i.e. without recomputing the parameters on the Validation sets and *single sample*, i.e. each sample corrected independently of each other.

Availability: R Package *besc* available from http://www.explainbio.com/besc

Contact: sudhir.varma@hithru.com

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

Introduction

Batch effect (BE) has been the primary bottleneck for the large-scale integration of data from multiple experiments. BE, defined as the systematic biases between microarray data generated by different labs at different times or under different experimental conditions (Scherer; Fare et al., 2003), can act as a confounding variable in statistical tests and usually has a stronger effect on the measured expression than the biological phenotype under study (Leek et al., 2010).

Unknown or unrecorded experimental or biological differences can add a systematic difference between putative replicates within or between two batches. Thus, we use the term *batch effect* as a general term for any heterogeneity due to experimental factors between samples that are putative experimental replicates. The heterogeneity can extend to different samples within the same sample collection, i.e. considering only average difference between two collections will likely underestimate the BE.

In practice, it has proven difficult to separate heterogeneity due to technical differences from that due to unknown biological differences. The usual approach in batch correction (Leek and Storey, 2007; Johnson et al., 2007; Marron et al., 2007) is to protect the known covariates and remove all remaining heterogeneity. Biological differences such as sex and genotype can be clinically important but they will be removed if they are not part of the protected covariates. Conversely, if the study design is unbalanced, the statistical significance of the association of gene expression with the protected covariates can be inflated beyond what one would expect by just a reduction in noise (Nygaard et al., 2015).

Additionally, current batch correction methods are intended to be used each time a new composite dataset is created. It is known that specific differences in sample condition, experimental technique (Scherer; Li *et al.*, 2014) and environmental conditions (Fare et al., 2003) can affect the measured gene expression in predictable directions irrespective of the sample type. However there has been little systematic effort to estimate how many of those common effects are shared between datasets or to compute dataset-independent batch correction parameters that can be used for "blind" prediction of BE.

In this article, we show that a large proportion of the BE in Affymetrix U133 Plus2 array data can be captured by a relatively small set of signatures, defined as the directions in which the measured expression has been perturbed by batch effects. We estimate *batch effect signatures* (BES) in the form of orthogonal components from a large dataset of reference samples. We develop an algorithm for computing the BES using the Reference dataset such that the BES are unlikely to model known or unknown biological differences. We introduce a novel batch-correction method called Batch Effect Signature Correction (*BESC*) that uses the batch effect signatures for blind prediction and correction of BE in new samples and compare the performance to *SVA*.

Methods

Batch effect signature

The measured expression for a set of samples can depend on one or more known biological factors (such as cell line name, tissue of origin or disease status), unknown or unmodeled biological factors (age, sex, genotype) and unknown or unmodeled experimental factors (such as microarray batch, FFPE vs. fresh and experimental technique). Following (Leek and Storey, 2007) we can model the expression of a sample as

$$x_{ij}=\mu_i+f_i(y_j)+\sum_{l=1}^{L}\gamma_{il}g_{lj}+e_{ij} \quad (1)$$

Where x_{ij} is the measured expression of gene i (out of m genes) for sample j (out of n samples), μ_i is the overall mean expression of gene i, $f_i(y_i)$ is a (possibly non-linear) function that models the dependence of the expression on known factor y_i, $\sum_{l=1}^{L}\gamma_{il}g_{lj}$ is the dependence of the expression on L (possibly non-linear) functions g_{lj} of the unknown biological or experimental factors and e_{ij} is uncorrelated noise.

Equation 1 indicates that the overall space of measured gene expressions can be separated into that spanned by the biological variation $f_i(y_i)$ and that spanned by systematic batch effects $\sum_{l=1}^{L}\gamma_{il}g_{lj}$. The purpose of BE correction is to estimate and remove variation in the BE space while retaining the biological differences. *Data-specific* batch correction methods (such as *ComBat* and *SVA*) assume that the BE space is unique to each composite dataset and has to be recalculated for each time a new composite dataset is created. However, we show that orthogonal basis derived from the BE space of a reference dataset can be used to estimate and remove the variation in the BE space for other test datasets (i.e. *blind* batch correction).

Fitting the functions f_i and global means μ_i using linear or non-linear regression, we can express Equation 1 in terms of the residuals of the fit

$$r_{ij}=x_{ij}-\mu_i-f_i(y_j)=\sum_{l=1}^{L}\gamma_{il}g_{lj} \quad (2)$$

The aim of *SVA* is to find a set of $K \le L$ orthogonal vectors (*surrogate variables*) that span the same linear space as g_l

$$\sum_{k=1}^{K}\lambda_{ik}h_{kj}=\sum_{l=1}^{L}\gamma_{il}g_{lj} \quad (3)$$

Each surrogate variable $h_k=[h_{k1},h_{k2},\ldots,h_{kn}]^T$ is a vector of length equal to the number of samples and together they can be used to model heterogeneity from unknown sources in any future statistical analysis on that dataset.

To motivate our approach, note that λ_{ki} is the difference in the expression of the i^{th} gene for each unit change in the k^{th} surrogate variable. Thus, the vector $\lambda_k=[\lambda_{k1},\lambda_{k2},\ldots,\lambda_{km}]$ is a *signature* that quantifies the dependence of the expression of all genes on the k^{th} surrogate variable. For example, if the surrogate variable captures the effect of a particular type of sample preparation, the *signature* is the differential expression between samples that use that preparation method *vs.* samples that do not. We call the set of K vectors λ_k the *Batch Effect Signatures* (BES). Each vector is a zero-mean unit vector (i.e. length one) that is as long as the number of probes on the array. Considering BE to be a separation in the probe expression space, between samples of the same type but different batch, each BES points in the direction of

For a new set of samples, we would not expect the values of the surrogate variables to be the same, but the expression difference (i.e. signature λ_k) between samples that use the preparation method *vs.* samples that do not would likely remain the same. Given a dataset of reference samples that is large and diverse enough, we can compute the signatures of the various known or unknown factors that contribute to BE. Those signatures can then be used to estimate and remove BEs from new samples. We call this approach the *Batch Effect Signature Correction* (*BESC*).

Difference between *SVA* and *BESC*

There are some differences between the *SVA* and *BESC* formulations. Writing Equation (2) in matrix form

$$R=DU \quad (4)$$

where R is the $m\times n$ matrix of residuals (n samples, m genes), D is the $m\times K$ matrix of batch effect signatures (element at i^{th} row and k^{th} column λ_{ik} and U is the $K\times n$ matrix of surrogate variables (element at k^{th} row and j^{th} column h_{kj}. *SVA* constrains the rows of U to be orthogonal and estimates the matrix U using singular value decomposition (SVD) on the residual matrix R. On the other hand, we constrain the columns of D to be orthogonal and estimate D by taking the Principal Components (PCA) of the transpose of the residual matrix, R^T. Note that there is no substantial difference between using SVD or PCA for decomposing R (except computational stability).

SVA is not strictly a batch correction algorithm. The surrogate variables also capture heterogeneity due to unknown biological differences since they are calculated without reference to the batch. Other batch correction methods such as ComBat (Johnson *et al.*, 2007) require known batch assignment, i.e. which samples belongs to which batch. As *SVA* does not require that information, it is the closest comparable algorithm to *BESC*.

Selection of reference samples

To ensure that the BES computed from a reference dataset does not capture unknown biological covariates of the samples (e.g. sex, genotype), we use a dataset of 2020 cell line samples. Fitting the linear model in Equation (1) using the cell line name as the known covariate captures all known and unknown biological differences between the samples. The same cell line, under untreated conditions, should give similar expression profiles between replicates. Although it is possible for growth conditions (e.g. passage number, growth medium) to affect the expression for a cell line, it can be argued that cell lines that have been grown from a standardized population of cells are one of the most replicable biological samples. Any differences in expression measured in two batches can be treated as mostly arising from BEs and experimental noise.

Correcting BE using BES

Given a set of BES, the batch effect in a new set of samples is computed by fitting a linear model of the BES to the expressions of each sample, i.e. we find weights a_k that minimize the squared residuals

$$\left\| x_{new} - \sum_{k=1}^{K} a_k S_k \right\|^2 \quad (5)$$

Then $\sum_{k=1}^{K} a_k S_k$ is the estimate of the BEs and the residuals $x_{corr} = x_{new} - \sum_{k=1}^{K} a_k S_k$ is the corrected expression for the sample.

BES is thus a "blind" estimate of the BE; we do not recalculate BE parameters on the set of new samples. That enables *single sample correction* of new samples, i.e. without having to re-compute the correction for all the samples whenever new samples are added.

BE score

To summarize the BE in a set of samples we developed a BE score; the Distance Ratio Score (DRS). Consider a set of samples from one or more sample types hybridized in multiple batches. For a sample of a certain sample type, we take the log of the ratio of the distance to the closest sample belonging to same batch but a different sample type to the distance to the closest sample belonging to a different batch but the same sample type.

$$DRS = \frac{1}{n}\sum_{i=1}^{n} log_2\left(\frac{d(x_i, x_{i,sb/dt})}{d(x_i, x_{i,db/st})}\right) \quad (6)$$

Where $d()$ is any measure of dissimilarity between two samples, x_i is the i^{th} sample, $x_{i,sb/dt}$ is the closest sample from the same batch of a different sample type and $x_{i,db/st}$ is the closest sample from a different batch of the same sample type as x_i. Intuitively, the DRS is high if samples of the same type cluster together irrespective of batch since the denominator will be small compared to the numerator. Conversely if most of the samples cluster according to batch rather than sample type, the DRS will be small. For our analysis, we used the Euclidean distance as the dissimilarity measure.

Cross-validation of Batch Effect Signatures

The question remains whether BES calculated on one dataset are general enough to be predictive of the BE in another dataset. To investigate that, we selected and annotated a Reference Set of 2020 cell line samples (242 cell lines from 348 collections) on the Affymetrix U133 Plus 2.0 platform from the Gene Expression Omnibus (Supplementary Table 1).

We did 5-fold cross-validation, i.e. created 5 random splits of the samples in the Reference Set into training and testing sets (approximately 80% samples for training and 20% samples for testing) ensuring that all samples from an individual collection are either all in the testing set or in the training set (Supplementary Table 1).

For each train/test split, we used the training set to compute the residuals after fitting a model to the expression with the known sample type as the covariate. Then we computed the principal components of the transpose of the residual matrix. The eigenvectors of the covariance matrix (i.e. the principal components) were used as putative BES. We used varying numbers of eigenvectors with the highest eigenvalues to correct the samples in the testing set and computed a Distance Ratio Score (DRS) of the test set samples for each number of BES.

Testing on Validation Sets

We selected two Validation Sets to test the effectiveness of the BES calculated on the Reference Set for removing BE. All of these samples were collected from public GEO datasets. Annotations for the samples were compiled from annotations provided by the original data submitter.

Different numbers of the top BES were used to estimate and remove the BE in the Validation Sets and the DSR BE score was calculated. Note that the BES were calculated using only the Reference Set samples. The sample type (or any other covariate) of the Validation Set samples were not used during the BE correction.

Validation set 1 (Primary normal samples) – 4485 samples of primary healthy tissue (50 organs, 328 collections) on the U133 Plus 2.0 platform from GEO (Supplementary Table 2). We predicted the sex of each sample using 5 chrY genes (Supplementary Methods).

Validation set 2 (Colon cancer and normal) – 3041 samples of primary colon cancer and normal samples (476 normal colon and 2565 colon cancer samples from 60 GEO collections) on the U133 Plus 2.0 platform from GEO (Supplementary Table 3). We have compiled the reported Micro-Satellite Instability (MSI) status for 513 out of the 2565 colon cancer samples (154 MSI, 359 MSS).

Permutation p-value for DSR

To test whether the improvement in the DSR is statistically significant, we performed a permutation test of the DSR obtained with various numbers of BES. We permuted the data for each gene in the Reference Set to destroy any batch effect signatures and then computed the BES using that null dataset. Varying numbers of those null BES were used to correct the samples in the Validation Set and the DSR obtained was compared to the true DSR at different numbers of BES. The procedure was repeated 100 times and the null DSRs used to compute the p-value for the true DSR using a Student's t-test at each number of BES.

Comparison to *SVA*

We compared the DSR for the Validation set for increasing numbers of BES to that for increasing numbers of surrogate variables as estimated by *SVA*. Note that it is not a direct comparison, since *SVA* uses the Validation Set samples to compute the BE parameters. Thus, the training set/validation set approach we have taken with BES cannot be applied to *SVA*.

One disadvantage of *SVA* is that it will remove unknown biological differences (e.g. sex) from the cleaned data. To show that, we predicted the sex of each sample in the Validation set 1 using the expression of chromosome Y genes (RPS4Y1, KDM5D, USP9Y, DDX3Y, EIF1AY, see Supplementary Methods) and looked at the number of genes significantly different between male and female samples at various levels of correction for *BESC* and *SVA*. For Validation set 2, we had the reported MSI status for 513 colon cancer samples. We looked at the number of genes significantly different between MSI and MSS samples.

Results

BES computed on one training dataset is predictive of test data BEs

Figure 1 shows the 5-split cross-validated batch effect DRS on the Reference set. The plot shows the average DRS on the test set corrected using BES calculated on the training set. The DRS increases as the number of BES increases, indicating reduction of BE with increasing correction. It reaches a maximum with 30 BES used in the correction, indicating that the first 30 Batch Effect Signatures estimated on the training set capture variation due to BE in the test set. Further correction reduces the

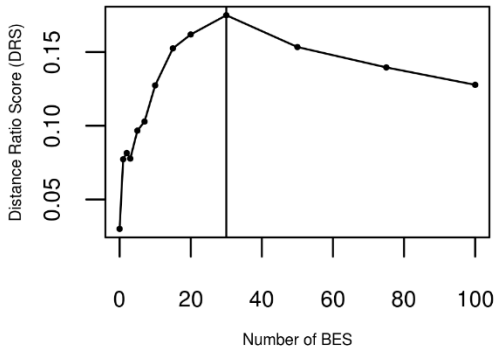

Figure 1: The cross-validated Distance Ratio Score (DRS) for the Reference set vs. the number of Batch Effect Signatures (BES) used for the correction. Higher DRS indicate lower levels of batch effect. The DRS reaches a maximum for 30 BES

DRS, indicating that the BES above 30 do not capture any information about the BE.

BES computed on Reference set corrects BE in Validation sets

Figure 2a shows the DRS on Validation set 1 using various numbers of BES calculated on the Reference set. The figure compares the performance of the *BESC* method to the *SVA* method using the same number of

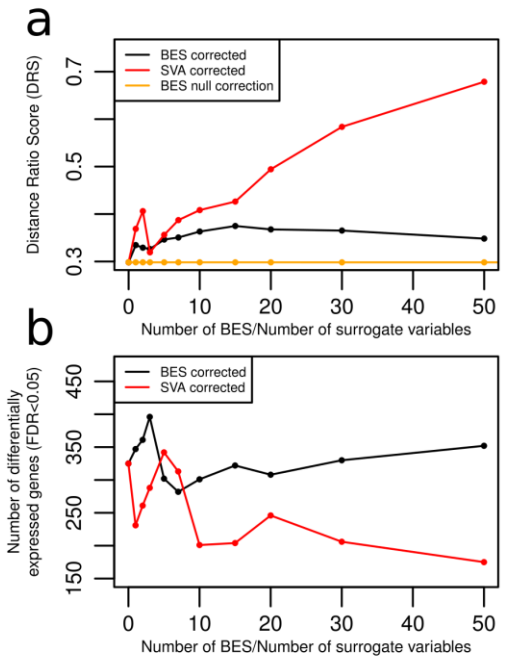

Figure 2: a) DRS for the Validation Set 1 using SVA, BES and permuted null BES b) Number of genes differentially expressed between male and female samples at various levels of correction by BES and SVA

surrogate variables as the number of BES. The *BESC* method again shows improvement in the DRS when up-to 30 BES are used for correction. Using more than 30 signatures begin to decrease the effectiveness.

Figure 3a shows the DRS on Validation set 2 for varying number of BES used in the correction. For that dataset, the DRS reaches a maximum plateau at 15 BES. Using more than 15 BES does not significantly change the DRS.

Comparison to *SVA*

SVA shows superior performance (Figure 2a) over most of the range of the x-axis (number of BES/number of surrogate variables used for correction).

However, Figure 2b and 3b illustrates the primary disadvantage of *SVA* which is the "normalizing away" of unknown but true biological differences. When sample sex is not included as one of the "protected"

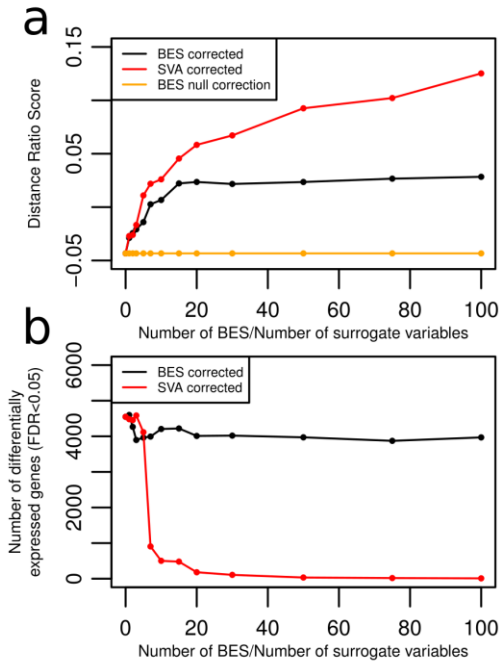

Figure 3:a) DRS for the Validation Set 2 using SVA, BES and permuted null BES b) Number of genes differentially expressed between MSI and MSS samples at various levels of correction by BES and SVA

covariates in the *SVA* algorithm, the difference between male and female samples (i.e. the number of statistically significant genes) decreases with increasing *SVA* correction. On the other hand, the difference is maintained (and slightly increased) with increasing BES correction.

Figure 3b shows the number of genes that are differentially expressed between MSI and MSS samples at a statistically significant level (FDR<0.05). There is a steep drop off of the number of genes for the *SVA*-corrected data. *SVA* correction eventually removes almost all of the differences between the MSI and MSS samples. On the other hand, the BES corrected data retains most of the differential expression.

Figures 2b and 3b emphasize the conservative nature of *BESC*; only differences that are known to be due to BE are removed. Any unknown/unmodeled difference between the samples that is due to true biology is maintained. In contrast, *SVA* captures those differences unless they are protected, and correction with the surrogate variables removes those differences from the samples.

BES correction is statistically significant

The orange line (Figure 2a, 3a) shows the average DRS for the Validation sets corrected using the BES calculated on the permuted Reference set data. As expected, there is no significant correction of the data (i.e. the DRS does not improve from baseline). The z-score p-value of the DRS using the true Reference set (black line) is <1e-16 over the entire range of numbers of BES.

Discussion

We show that the BE parameters, i.e. the directions in multi-variate space along which BE perturbs the measure gene expression is shared between different datasets. That enables us to compute Batch Effect Signature (BES) vectors that capture the direction of perturbation on a Reference dataset and apply them to predict and remove the batch in independent Validation datasets. As far as we know, no other "blind" methods of batch correction have been published. All methods, (including *SVA*) require the correction factors to be computed on the entire sample set, needing re-calculation each time new samples are added.

Crucial to the correct operation of our algorithm is the selection of a Reference set composed of cell lines since specifying the cell line name completely fixes all known and unknown biological covariates. We argue that cell lines that have been grown from a standardized population of cells are one of the most replicable biological experiments, and any difference between the same cell line sample from different experiments is quite likely due only to technical variation. However, note that we can use any sample that has been analyzed in multiple experiments by different labs. For example, samples from The Cancer Genome Atlas (TCGA) or any other sample repository have the property that specifying the sample id completely specifies all of the known and unknown biological covariates. All that is required for inclusion in the Reference set is that the sample must be uniquely identified across multiple experiments.

We compared the *BESC* algorithm to Surrogate Variable Analysis (*SVA*). At larger number of surrogate variables, *SVA* is more effective at detecting and removing residual structure from the dataset, however it is also likely to remove unknown but important biological information (e.g. sex in the case of normal samples and MSI vs. MSS differences in the case of colon cancer). We find *BESC* to be much more conservative about retaining unknown biological differences while removing technical differences.

The characteristics of the *BESC* algorithm make it ideally suited for large-scale batch correction in microarray data repositories. *BESC* is conservative, i.e. only BE known to be likely due to technical differences is removed. We have shown that biological variation, even if unknown to the *BESC* algorithm, is preserved. *BESC* can be applied to individual samples and does not need to be recomputed as more samples are added to the repository.

One primary disadvantage of the *BESC* algorithm is that it will remove all differences between samples that are parallel to the BES vectors. It is possible that there is important biological information along those directions. However, that biological difference is confounded with the technical differences and cannot be separated out in any analysis.

Another disadvantage is that the current BES vectors are only applicable to the Affymetrix U133 Plus2 array. We used samples from the U133 Plus2 platform to calculate the BES, mainly because it is the most commonly used platform with a large number of cell line samples. Before BES can be used on another platform, we will have to compile a Reference set of cell line or other identified samples on that platform and recompute the vectors on that set. Other BE correction methods are platform agnostic.

Funding

Conflict of Interest: none declared.

References

Fare,T.L. *et al.* (2003) Effects of Atmospheric Ozone on Microarray Data Quality. *Anal. Chem.*, **75**, 4672–4675.

Johnson,W.E. *et al.* (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.

Leek,J.T. *et al.* (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, **11**, 733–739.

Leek,J.T. and Storey,J.D. (2007) Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet*, **3**, e161.

Li,S. *et al.* (2014) Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.*, **32**, 888–895.

Marron,J.S. *et al.* (2007) Distance-Weighted Discrimination. *J. Am. Stat. Assoc.*, **102**, 1267–1271.

Nygaard,V. *et al.* (2015) Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, kxv027.

Scherer,A. Batch Effects and Noise in Microarray Experiments: Sources and Solutions.