Blind estimation and correction of microarray batch effect
Sudhir Varma1,* sudhir.varma@hithru.com
1HiThru Analytics, Princeton NJ 08540
*To whom correspondence should be addressed.

Abstract
Microarray 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 have the potential to overcorrect, by removing true but unknown biological differences (SVA). Even though the effects of technical differences on measured expression have been published, there are no BE correction algorithms that take the approach of predicting technical effects from parameters computed from a fixed reference sample set. We show that a set of signatures, each of which is a vector the length of the number of probes, calculated on a Reference set of microarray samples can predict much of the batch effect in other Validation sets. We present a rationale of selecting a Reference set of samples designed to estimate technical differences without removing biological differences. Putting both together, we introduce the Batch Effect Signature Correction (BESC) algorithm that uses the BES calculated on the Reference set to efficiently predict and remove BE. Using two independent Validation sets, we show that BESC is capable of removing batch effect without removing unknown but true biological differences. Much of the variations due to batch effect is shared between different microarray datasets. That shared information can be used to predict signatures (i.e. directions of perturbation) due to batch effect in new datasets. The correction is blind (without needing to re-compute the parameters on new samples to be corrected), single sample, (each sample is corrected independently of each other) and conservative (only those perturbations known to be likely to be due to technical differences are removed ensuring that unknown but important biological differences are maintained). Those three characteristics make it ideal for high-throughput correction of samples for a microarray data repository. An R Package besc implementing the algorithm is available from http://explainbio.com.

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 [1,2], 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 [3].

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 [4–6] 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 [7,8]. 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 [9].

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 [1,10] and environmental conditions [2] 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 take another approach: instead of estimating and removing all differences between two batches, we only remove those differences that are known to be associated with technical variations. 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 Reference dataset of 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.

Materials and methods

Batch effect signature

The measured expression for a set of samples can depend on one or more biological factors (such as cell line name, tissue of origin or disease status) and unknown or unmodeled experimental factors (such as microarray batch, FFPE vs. fresh samples, and experimental technique). Following [4] we can model the expression of a sample as a linear combination of known biological covariates, unknown batch effect and noise

\[ x_{ij} = \mu_i + f_i(y_i) + \sum_{l=1}^{L} \gamma_{il} g_{lj} + e_{ij} \]  

(1)

Where \( x_{ij} \) is the measured expression of gene \( i \) (out of \( m \) genes) for sample \( j \) (out of \( n \) samples), \( \mu_i \) 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 biological factors \( y_i \). The batch effect is modeled as \( \sum_{l=1}^{L} \gamma_{il} g_{lj} \), the perturbation of the measured expression by \( L \) experimental factors. \( g_{lj} \) is the effect of the \( l^{th} \) experimental factor on the \( i^{th} \) gene 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 Com-
**Predicting microarray batch effect**

Bat and SVA) assume that the BE space is unique to each composite dataset and has to be recalculated each time a new composite dataset is created. However, we show that an 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 $\mu_i$ using linear or non-linear regression, we can express Equation 1 in terms of the residuals of the fit

$$ r_{ij} = x_{ij} - \mu_i - f_i(y_j) = \sum_{l=1}^{L} \gamma_l g_{lj} $$  \hspace{1cm} (2)$$

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

$$ \sum_{k=1}^{K} \lambda_{ki} h_{kj} = \sum_{l=1}^{L} \gamma_l g_{lj} $$  \hspace{1cm} (3)$$

Each surrogate variable $h_k = [h_{k1}, h_{k2}, ..., h_{kn}]^T$ 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 $\lambda_{ki}$ is the difference in the expression of the $i^{th}$ gene for each unit change in the $k^{th}$ surrogate variable. Thus, the vector $\lambda_k = [\lambda_{1k}, \lambda_{2k}, ..., \lambda_{mk}]$ 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 $\lambda_k$ 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 batch effects to be perturbations in various directions between samples of the same type but different batch, each perturbation can be estimated as a weighted sum of the BES.

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 $\lambda_k$) between samples that differ on a certain experimental factor (e.g. amplified using a certain RNA amplification protocol vs. un-amplified) would be more likely to 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

$$ R = DU $$  \hspace{1cm} (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 ($\lambda_{ik}$ the element at $i^{th}$ row and $k^{th}$ column) and $U$ is the $K \times n$ matrix of surrogate variables ($h_{kj}$ the element at $k^{th}$ row and $j^{th}$ column). SVA constrains the rows of $U$ to be orthogonal and estimates the matrix $U$ using singular value decomposition (SVD) on the residual
matrix $R$ (which also constrains the columns of $D$ to be orthogonal). On the other hand, we con-
strain 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 be-
tween using SVD or PCA for decomposing $R$ (except computational stability).

Strictly speaking, SVA is not 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 [5] require known batch assign-
ment, i.e. which samples belongs to which batch. A modification to SVA, permuted-SVA (or
pSVA) [8] has been proposed to prevent the algorithm from removing unknown biological differ-
ences. However, pSVA has the same limitation as ComBat, i.e. it is applicable only when the
technical covariates that contribute to batch effect are known. As SVA does not require that in-
formation, it is the closest comparable algorithm to BESC and was selected for comparison to
BESC in this article.

Selecting samples for the Reference set

The selection of the Reference set is crucial to ensure that the BES computed from it does not
capture unknown biological covariates of the samples (e.g. sex, genotype). We selected and an-
notated a Reference set of 2020 cell line samples (242 unique cell lines) on the Affymetrix U133
Plus 2.0 platform, selected from 348 collections from the Gene Expression Omnibus (S1 Table).
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 standard-
ized population of cells are one of the most replicable biological samples. Any differences in ex-
pression 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

$$
\left\| x_{new} - \sum_{k=1}^{K} a_k \lambda_k \right\|^2
$$

(5)

Where $x_{new}$ is the vector of expressions for a new sample to be corrected and $\lambda_k$ is the $k^{th}$ BES.
$\sum_{k=1}^{K} a_k \lambda_k$ is the estimate of the BEs and the residuals $x_{corr} = x_{new} - \sum_{k=1}^{K} a_k \lambda_k$ is the corrected
expression for the sample. Since the BES $\lambda_k$ are zero sum orthonormal vectors, it is simple to
show that $a_k = \lambda_k^T x_{new}$ where the superscript T indicates the transpose, i.e. the weights for each
BES is the scalar product of the BES and the expression vector.

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.
Predicting microarray batch effect

Quantifying batch effect

Several methods have been proposed in literature to visualize or quantify the batch effect in a dataset [11]. Visualization methods include clustering (dendrograms) and principal component analysis (PCA). However, they are not suited for very large number of samples, as is the case for the datasets in this paper. Principal Variance Component Analysis (PVCA) [1] has been proposed as a quantitative measure. PVCA fits a linear mixed model to the principal components of the data using the known biological covariates and batch identities as random effects. The variance due to the different covariates and batch is summed up across the principal components and offers a way to compare the primary sources of variance. For the two Validation sets, we used PVCA to compute the contribution of variance due to sample-type as a measure of the batch effect. The higher the contribution, the lower the batch effect.

However, PVCA has the disadvantage that the variances due to batch effect in different test datasets cannot be averaged together (as is required for cross-validation). So, we developed another batch effect measure, the Distance Ratio Score (DRS) that is intuitive, provides one value for each sample and can be averaged across various test sets. 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 of a different sample type to the distance to the closest sample belonging to a different batch but the same sample type.

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

(6)

Where \(d()\) is any measure of dissimilarity between two samples, \(x_i\) is the \(i^{th}\) sample, \(x_{i,dt}\) is the closest sample 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-validating 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 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 (S1 Table).

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 DRS 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) – 878 samples of primary healthy tissue (532 blood, 228 colon and 118 lung samples from 41 collections) on the U133 Plus 2.0 platform from GEO (S2 Table). We predicted the sex of each sample using 5 chromosome-Y genes (Supplementary Methods). Samples were selected to create a balanced dataset with respect to collection, organ and sex.

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 (S3 Table). We have compiled the reported Micro-Satellite Stability/Instability (MSS/MSI) status for 513 out of the 2565 colon cancer samples (154 MSI, 359 MSS). Out of those 513, we selected a balanced set of 503 samples with known MSI/MSS status to test for differential expression between MSI and MSS.

Permutation p-value for DRS

To test whether the improvement in the DRS is statistically significant, we performed a permutation test of the DRS 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 DRS obtained was compared to the true DRS at different numbers of BES. The procedure was repeated 100 times and the null DRSs used to compute the p-value for the true DRS using a Student’s t-test at each number of BES.

Comparison to SVA

We compared the DRS 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 [7,8]. 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. The computation of statistical significance for Validation set 1 was done using organ source, sex and batch as covariates and for Validation set 2 was done using batch and MSI/MSS status as covariates.
Results

**BES computed on training data is predictive of batch effect in test data**

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

**BES computed on Reference set corrects batch effect in Validation sets**

Fig 2a shows the DRS on Validation set 1 using various numbers of BES calculated on the Reference set. The Fig compares the performance of the BESC method to the SVA method using the same number of surrogate variables as the number of BES. The BESC method shows improvement in the DRS when up-to 15 BES are used for correction. Using more than 15 signatures begins to decrease the effectiveness. Fig 2b shows similar results using the variance contributed to the sample type (in this case, the organ of origin of the samples). The variance increases for the sample type (with a corresponding decreasing in variance due to batch effect; not shown) as the number of BES/number of surrogate variables increases. The results for SVA show a steady improvement in batch effect with at each level. The performance for BESC reaches a maximum when 15 BES are used.

Fig 3a shows the DRS on Validation set 2 for varying number of BES used in the correction. For that dataset too, the DRS reaches a maximum plateau at 15 BES. Using more than 15 BES does not significantly change the DRS. Fig 3b shows similar results using the PVCA-computed variance contribution due to sample-type (i.e. disease status- colon cancer vs. normal in this case). In both cases, the SVA corrected data shows superior performance in removing batch effect.

**Comparison to SVA**

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

However, Fig 2c and 3c 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” 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.

Fig 3c 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 differ-
ences between the MSI and MSS samples. On the other hand, the BES corrected data retains most of the differential expression.

Figs 2c and 3c 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 (Fig 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 not possible to know how much of that residual structure is due to batch effect compared to unknown but true biological differences. We show that SVA is very likely to remove unknown but important biological information in our Validation sets (e.g. sex in the case of normal samples and MSI vs. MSS differences in the case of colon cancer). We show that BESC is to 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.
Predicting microarray batch effect

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 re-compute the vectors on that set. Other BE correction methods are platform agnostic.

Conclusions

This paper describes a novel finding that batch effect (BE) perturbs measured gene expression in predictable directions which we call Batch Effect Signatures (BES). That characteristic can be used to compute possible directions of the perturbations in a Reference dataset which can used to predict the BE in an independent Validation set. Selecting the Reference set to contain only known cell-lines ensures that all (known or unknown) biological differences are fixed by specifying the cell-line name. That ensures that the BES capture differences due only to technical differences between the batches.

We show that the BES calculated on the Reference set efficiently removes batch effect in two Validation sets, as measured by PVCA and the Distance Ratio Score (DRS), a novel measure of batch effect. Compared to SVA, our algorithm does not remove all possible differences between samples of the same type in different batches, but we show that SVA also over-corrects by removing unknown but true biological differences.

An R Package besc implementing the algorithm is available from http://www.explainbio.com. All data used are public data available from GEO (Gene Expression Omnibus https://www.ncbi.nlm.nih.gov/geo/). GEO accession numbers for all samples used in the analyses are in S1-S3 Tables.

Acknowledgements

I gratefully acknowledge the valuable feedback on this paper given by Vinodh Rajpakse and Augustin Luna.

References

1. Scherer A. Batch Effects and Noise in Microarray Experiments: Sources and Solutions [Internet]. [cited 2016 Nov 29]. Available from: http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470741384.html

2. Fare TL, Coffey EM, Dai H, He YD, Kessler DA, Kilian KA, et al. Effects of Atmospheric Ozone on Microarray Data Quality. Anal. Chem. 2003;75:4672–5.

3. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. 2010;11:733–9.

4. Leek JT, Storey JD. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. PLoS Genet. 2007;3:e161.

5. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.
6. Marron JS, Todd MJ, Ahn J. Distance-Weighted Discrimination. J. Am. Stat. Assoc. 2007;102:1267–71.

7. Jaffe AE, Hyde T, Kleinman J, Weinberg DR, Chenoweth JG, McKay RD, et al. Practical impacts of genomic data “cleaning” on biological discovery using surrogate variable analysis. BMC Bioinformatics. 2015;16:372.

8. Parker HS, Leek JT, Favorov AV, Considine M, Xia X, Chavan S, et al. Preserving biological heterogeneity with a permuted surrogate variable analysis for genomics batch correction. Bioinformatics. 2014;30:2757–63.

9. Nygaard V, Rødland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics. 2015;kxv027.

10. Li S, Łabaj PP, Zumbo P, Sykacek P, Shi W, Shi L, et al. Detecting and correcting systematic variation in large-scale RNA sequencing data. Nat. Biotechnol. 2014;32:888–95.

11. Lazar C, Meganck S, Timinau J, Steenhoff D, Coletta A, Molter C, et al. Batch effect removal methods for microarray gene expression data integration: a survey. Brief. Bioinform. 2012;
Fig 1: Cross-validated performance on Reference set
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
Fig 2: Performance on Validation set 1
a) DRS for the Validation Set 1 using SVA, BES and permuted null BES
b) Contribution of variance due to organ-type, calculated using PVCA
c) Number of genes differentially expressed between male and female samples at various levels of correction by BES and SVA.
Fig 3: Performance on Validation set 2

a) DRS for the Validation Set 2 using SVA, BES and permuted null BES

b) Contribution of variance due to disease status, calculated using PVCA

c) Number of genes differentially expressed between MSI and MSS samples at various levels of correction by BES and SVA.