BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 19 2014, pages 2757–2763
doi:10.1093/bioinformatics/btu375

Gene expression

Advance Access publication June 6, 2014

Preserving biological heterogeneity with a permuted surrogate variable analysis for genomics batch correction

Hilary S. Parker[1], Jeffrey T. Leek[1], Alexander V. Favorov[2,3,4], Michael Considine[2], Xiaoxin Xia[5], Sameer Chavan[6], Christine H. Chung[2] and Elana J. Fertig[2,*]

[1]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, [2]Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD 21205, USA, [3]Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow 119333, Russia, [4]Research Institute for Genetics and Selection of Industrial Microorganisms "GosNIIGenetika", Moscow 117545, Russia, [5]Department of Statistics and Biostatistics, Rutgers University, NJ 08854, USA and [6]Division of Allergy & Clinical Immunology, Department of Medicine, Johns Hopkins University, Baltimore, MD 21224, USA

Associate Editor: Ziv Bar–Joseph

ABSTRACT

Motivation: Sample source, procurement process and other technical variations introduce batch effects into genomics data. Algorithms to remove these artifacts enhance differences between known biological covariates, but also carry potential concern of removing intragroup biological heterogeneity and thus any personalized genomic signatures. As a result, accurate identification of novel subtypes from batch-corrected genomics data is challenging using standard algorithms designed to remove batch effects for class comparison analyses. Nor can batch effects be corrected reliably in future applications of genomics-based clinical tests, in which the biological groups are by definition unknown a priori.

Results: Therefore, we assess the extent to which various batch correction algorithms remove true biological heterogeneity. We also introduce an algorithm, permuted-SVA (pSVA), using a new statistical model that is blind to biological covariates to correct for technical artifacts while retaining biological heterogeneity in genomic data. This algorithm facilitated accurate subtype identification in head and neck cancer from gene expression data in both formalin-fixed and frozen samples. When applied to predict Human Papillomavirus (HPV) status, pSVA improved cross-study validation even if the sample batches were highly confounded with HPV status in the training set.

Availability and implementation: All analyses were performed using R version 2.15.0. The code and data used to generate the results of this manuscript is available from https://sourceforge.net/projects/psva.

Contact: ejfertig@jhmi.edu

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

Received on December 18, 2013; revised on April 7, 2014; accepted on June 2, 2014

1 INTRODUCTION

High-throughput genome-scale data can yield powerful predictors of sample phenotype. When applied to disease such as human cancers, the resulting classifiers may predict patient prognosis or therapeutic strategy to select appropriate personalized treatment. Nonetheless, technical artifacts from sample collection and processing introduce biologically irrelevant signal into high-throughput data, referred to as 'batch effects'. Even modest changes to experimental design, such as the microarray scan date or technician, can adversely impact the accuracy of genomic analyses (Leek *et al.*, 2010; Luo *et al.*, 2010). Several batch correction algorithms can remove these artifacts statistically, provided that pertinent biological groups are known and well represented in each batch (Gagnon-Bartsch and Speed, 2012; Johnson *et al.*, 2007; Leek and Storey, 2007; Sun *et al.*, 2011).

Current batch correction algorithms remove signal unrelated to the 'protected' biological groups, thereby enhancing group differences unrelated to batch. For example, ComBat corrects genomic data using an empirical Bayes framework to fit parameters in a linear model containing both batch and biological covariates (Johnson *et al.*, 2007). Similarly, Surrogate Variable Analysis (SVA; Leek and Storey, 2007) uses an iterative procedure to estimate the effect of batch covariates spanning a space not associated with biological covariates. Although essential for class comparison, these corrections sometimes unintentionally eliminate true biological heterogeneity not encoded in the modeled biological covariates. As a result, novel patterns cannot be identified from such corrected data, limiting inference of either dynamics from time course genomics data or of novel disease subtypes. The heterogeneity of human tumors makes the removal of sample heterogeneity particularly challenging in cancer genomics. In this field, biological heterogeneity is essential to realize the central goal of identifying personalized genomic signatures to select optimal treatments for each cancer patient. Worse still, the need to prioritize collection of formalin-fixed paraffin-embedded (FFPE) samples for pathology may introduce an imbalance and significant technical artifacts in samples available for such classification (Viljoen and Blackburn, 2013).

This study provides a comprehensive assessment of the extent to which batch correction techniques, including notably SVA and ComBat, remove such biological heterogeneity when correcting for technical artifacts. We also propose a new algorithm, permuted-SVA (pSVA), to counteract such overcorrection of genomics data. This algorithm incorporates an innovative statistical model in the computational framework of SVA (Leek and

*To whom correspondence should be addressed.

Storey, 2007) to iteratively refine sample clusters after modeling the effect of known batches. This algorithm removes technical artifacts in replicate samples, while retaining biological heterogeneity in samples. As a result, correcting genomics data with pSVA improves subtype identification and class discovery in batch-affected gene expression data from head and neck squamous cell carcinoma (HNSCC) tumors.

2 METHODS

2.1 pSVA

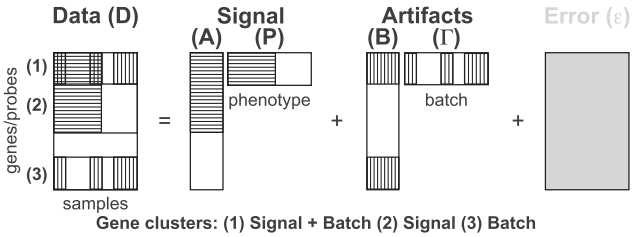
Batch correction procedures typically model a response composed of two components: biological signal and experimental artifacts. For example, SVA models a matrix of gene expression data **D** of g genes (rows) and s samples (columns) as

$$\mathbf{D} \sim \mathbf{AP} + \mathbf{B\Gamma} + \varepsilon, \quad (1)$$

where **P** is a model matrix describing pertinent biological covariates, **A** the matrix of coefficients relating each gene to the covariates in **P**, **Γ** the model matrix corresponding to unmodeled factors (typically batch covariates), **B** the matrix of coefficients relating each gene to the factors in **Γ** and ε unbiased random noise. SVA assumes that the factors in **Γ** are unknown. Therefore, rather than estimating these factors directly, SVA instead estimates their net effect on gene expression (i.e. **BΓ**) using an iterative procedure that assumes these factors span a space inferred from genes not associated with the biological covariates. This procedure assumes that the gene expression data is a linear combination of mean-zero random noise and known biological covariates (**AP**), which must be 'protected' during the estimation of **B**. By basing correction in inferred sets of genes not associated with known biological covariates enables SVA to estimate technical artifacts that are not necessarily orthogonal to biological covariates, in contrast to standard applications of PCA.

SVA's above-described deconvolution of experimental artifacts from biological signal is based upon independent assessment of biologically relevant groups, not information content in the data. However, in many cases true sources of biological heterogeneity (e.g. disease subtypes, time course response) are unknown a priori, whereas technical covariates (e.g. data source) are known. In these cases, SVA's optimization procedure will by definition model the true sources of biological heterogeneity as technical artifacts and remove them from the data when not contained in the model covariates (**P**). However, we note that the PCA procedure used at the core of SVA is frequently used as a clustering method to model subtypes in genomic data. Based upon this observation, pSVA reverses the standard application of SVA to model biological heterogeneity as those features estimated from genes not associated with known technical covariates in the model matrix **Γ** (Fig. 1). Specifically, pSVA inputs a model matrix of technical covariates (**Γ**) and then uses the iterative procedure in SVA to estimate the net effect of factors spanning from genes not associated with these technical artifacts. Namely, it identifies coefficients Ψ corresponding to an orthonormal set of vectors **Θ** that span the same space as the unknown biological covariates (columns of the model matrix **P**), so that $\Psi\mathbf{\Theta} = \mathbf{AP}$. Finally, pSVA retains this estimate for the product **AP** as the batch corrected denoised data for all subsequent analyses. Thus, the iterative use of PCA to estimate orthogonal covariates to batch in pSVA is analogous to iteratively refining estimates of batch with unknown sample subtypes. This algorithm is also naturally suited for class prediction or time course genomics because it never requires a priori knowledge of sample phenotypes unknown in future test sets. The new pSVA algorithm was implemented in an R function, provided on https://sourceforge.net/projects/psva.

Results with pSVA batch correction were compared with implementations of SVA in the SVA package (Leek *et al*., 2012). It was also compared with this package's implementation of ComBat, which also fits the model in Equation (1) from estimates of both **P** and **Γ** with an empirical Bayes procedure, as depicted in Figure 1 (Johnson *et al*., 2007). A combination of SVA and ComBat was also considered, for which SVA was applied to the data matrix resulting from ComBat correction.



Gene clusters: (1) Signal + Batch (2) Signal (3) Batch

	SVA	pSVA	ComBat
Observations	D & phenotype covariates (P)	D & batch covariates (Γ)	D & P & Γ
Desired estimate	Span of artifacts (BΓ)	Span of signal (AP)	Empirical Bayes fit to model in eq (1)
Estimation sample groups by	Inferring gene cluster (3)	Inferring gene cluster (2)	

Fig. 1. Comparison of SVA, pSVA and ComBat. Illustration of the model batch-affected genomics data from distinct phenotypes modeled in Equation (1) (top). As described in the bottom table, SVA algorithm assumes that phenotype is known. SVA then uses an iterative algorithm to find those genes unaffected by batch, and thereby fit the depicted model. pSVA follows a similar procedure, in cases where the batch is assumed known a priori. In contrast, ComBat uses an empirical Bayes algorithm to model the effects of known biological and batch covariates

2.2 Gene expression data, normalization and subtype identification

We formulated training data from our long-term retrospective profiling HNSCC primary tumors using Affymetrix hgu133plus2.0 in GSE3292 (Slebos *et al*., 2006), GSE10300 (Cohen *et al*., 2009) and GSE53355 (accession numbers for each sample in Supplementary Table S1). The tumors were collected from HNSCC patients under an Institutional Review Board (IRB) approval at Vanderbilt University. Samples selected were either from frozen or FFPE tumors, results for some of which have been previously published (Chung *et al*., 2010, 2011; Gilbert *et al*., 2012; Slebos *et al*., 2006). Briefly, the RNA samples were isolated from either frozen or FFPE tumors and amplified using four different commercially available RNA isolation kits over 4 years. To ensure sample balance with batch, we retained only those frozen samples that used WT-Ovation RNA Amplification System V1 (called Ovation 1), Ovation RNA Amplification System V2 (called Ovation 2), WT-Ovation FFPE-beta RNA Amplification System (called FFPE_beta) or WT-Ovation FFPE RNA Amplification System called (FFPE), from NuGEN Technologies, Inc., San Carlos, CA. Cross-study validation was performed relative to a single batch of frozen samples measured with the hgu133plus2.0 array available in GSE6791 (Pyeon *et al*., 2007) (Table 1).

All arrays were normalized with fRMA (McCall *et al*., 2010). Gene-level estimates were obtained by selecting the probe with maximum fold change across all samples and patient-specific estimates by average expression values for all replicate samples. Batch correction was applied considering sample procurement and RNA amplification kit as dominant batches. Batch correction with ComBat, SVA and their combination included a model for **AP** that protected for the known Human Papillomavirus (HPV) status of samples. HNSCC subtypes were defined from batch-corrected data using the classifier from Walter *et al.* (2013), which distinguishes HPV-positive HNSCC as the 'atypical' subtype.

2.3 Genomic prediction of HPV status

HPV status of the tumors was determined by PCR as previously described (Slebos *et al*., 2006). We developed a genomic classifier of

Table 1. Number of samples of each HPV status across batches

Test set	HPV-negative	Unique HPV-negative	HPV-positive	Unique HPV-positive
Frozen Nugen Ovation	24	24	10	10
FFPE Nugen FFPE	13	12	4	4
Frozen Nugen FFPE	15	14	14	13
GSE6791 Pyeon *et al.* (2007)	26	26	16	16

Notes: Columns labeled with 'unique' count the number of samples from distinct tumors in that batch, excluding replicate samples from the same tumor within each batch.

HPV status using the PAM algorithm, implemented in the CRAN package pamr (Tibshirani *et al.*, 2002). Cross-batch validation considered data from Pyeon *et al.* (2007), GSE6791, as an additional batch to each of the three batches distinguished by sample processing in our dataset. Training was then performed on three of the four batches and testing on data from the remaining batch, excluding any replicate samples from tumors also measured in the training set. The biological groups of the test set are by definition unknown in such class prediction problems. Therefore, we only applied ComBat, SVA and the combination of SVA and ComBat to the training data. We applied pSVA to batch correct both test and training data, training the batch correction in the latter algorithm on the same data used in the classifier. Although such correction of the test set may bias our results toward pSVA, we note that its exclusion of a priori knowledge of biological groups makes it the only algorithm for which we may perform such correction. Estimates of classes in the test set were obtained at the individual tumor level by voting across replicate samples, breaking ties with an assignment of the more probable and worse prognosis HPV-negative class.

The effects of varying degrees of confounding were simulated. Specifically, training sets were selected by setting a proportion p and selecting $10\,p$ HPV-positive samples from GSE6791 (Pyeon *et al.*, 2007) and $20(1-p)$ HPV-negative samples from our Frozen Nugen Ovation batch. An additional $10(1-p)$ HPV-positive samples were selected from our Frozen Nugen Ovation batch and $20\,p$ HPV-negative from GSE6791 (Pyeon *et al.*, 2007), for a total of 10 HPV-positive and 20 HPV-negative samples. Thus, batch and HPV status are perfectly confounded when p is zero or one, and perfectly balanced when p is 0.5. A total of 100 training sets were selected from independent samples randomly for values of p between 0 and 1 at intervals of 0.1, for a total of 1100 simulations. Testing was performed on independent frozen samples in our study processed with the Nugen FFPE amplification kit.

3 RESULTS

3.1 Sample procurement and RNA isolation drive dominant gene expression signals

Our HNSCC dataset contained a total of 80 samples, with sample size and distribution of HPV status and sample processing shown in Table 1. The counts in rows 1 to 3 include replicate samples from the same tumor spread along processing techniques (Supplementary Table S1). Excluding these replicate samples leaves a total of 39 unique HPV-negative and 22 unique HPV-positive samples. The distribution of expression values was more variable in FFPE samples than frozen samples (Supplementary Fig. S2a). Moreover, RNA expressed at high levels showed the greatest variability in the FFPE samples, consistent with anticipated RNA degradation.

Although no apparent differences were observed in the distribution of frozen samples, the complexity in such study designs often introduces significant batch effects after apparent normalization. Hierarchical clustering confirmed that the RNA amplification kit was the dominant cause of expression differences between samples instead of any clinical attribute of the tumor samples (Supplementary Fig. S2b). We observed no significant batch effects between samples processed with Nugen Ovation 1 or Ovation 2 or samples processed with Nugen FFPE or FFPE_beta kits. Therefore, we label both Ovation 1 and Ovation 2 amplification kits as Nugen Ovation and labeled FFPE and FFPE_beta as Nugen FFPE.

3.2 Removing technical artifacts with pSVA preserves sample heterogeneity

We applied pSVA, ComBat, SVA and a combination of SVA and ComBat to data from our retrospective genomics study of HNSCC (Chung *et al.*, 2010, 2011; Gilbert *et al.*, 2012; Slebos *et al.*, 2006) to remove the apparent artifacts from amplification kit and procurement observed in Supplementary Figure S2. The applications of SVA and ComBat both model HPV status as the known biological covariate. Each of these techniques successfully mixes sample-processing groups in hierarchical clusters (Supplementary Fig. S3) and makes the distribution of expression levels more similar between FFPE and frozen samples (Supplementary Fig. S4a). Despite this correction, each batch correction algorithm preserves differential expression of the established HPV biomarker in HNSCC, p16 (CDKN2A) (Robinson *et al.*, 2010; Smeets *et al.*, 2007) (Supplementary Fig. S4b).

We also compared the correlation of gene expression profiles between samples corrected with each batch correction technique to assess the relative similarity between methods (Supplementary Fig. S5). Samples corrected with pSVA had most variable correlation coefficients with samples in the uncorrected data relative to any other batch correction technique. Applying ComBat alone yielded more similar gene expression profiles to pSVA than expression resulting after SVA correction alone or in combination with ComBat. Data corrected with pSVA was highly correlated to data obtained from applying SVA with known batches, suggesting that noise estimates for ε in Equation (1) are truly unbiased. Similarly high correlations were observed for data corrected with SVA, including both HPV status and batch in the model, suggesting that pSVA is retaining pertinent heterogeneity without any knowledge of biological groups.

To further validate the success of these batch correction algorithms, we compared the correlation of expression profiles in replicate samples (Fig. 2a). Prior to batch correction, there was substantial variation in expression profiles between replicate samples. Each batch correction algorithm increased the correlation between replicate samples, with greatest improvement observed when combining SVA and ComBat. However, only pSVA preserves the high heterogeneity between non-replicate samples reflective of tumor heterogeneity (Fig. 2b).

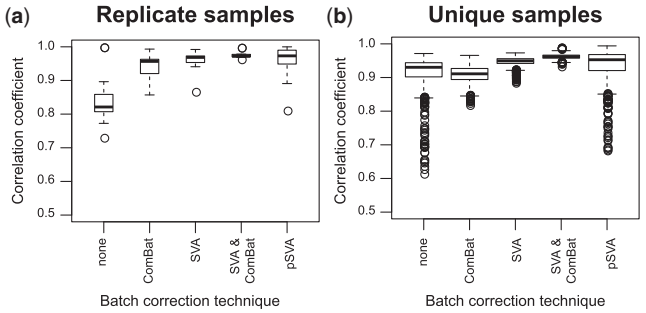


Fig. 2. Intra- versus intersample heterogeneity after batch correction. (**a**) Distribution of the Pearson correlation coefficient between replicate samples from the same tumor after each of the indicated batch correction algorithms is performed. Specifically, we computed Pearson correlation coefficients between all combinations of replicate samples from the same tumor. The boxplot shows the distribution of these correlation coefficients for all tumors with replicate samples, listed in Supplementary Table S1. (**b**) Distribution of the Pearson correlation coefficient between samples from unique tumors for each batch correction algorithm. Correlation coefficients are computed between all pairs of samples using the same processing technique. Correlations coefficients include those from replicate samples of the same tumor correlated to all other tumors, but exclude correlation coefficients between replicate samples plotted in (**a**)

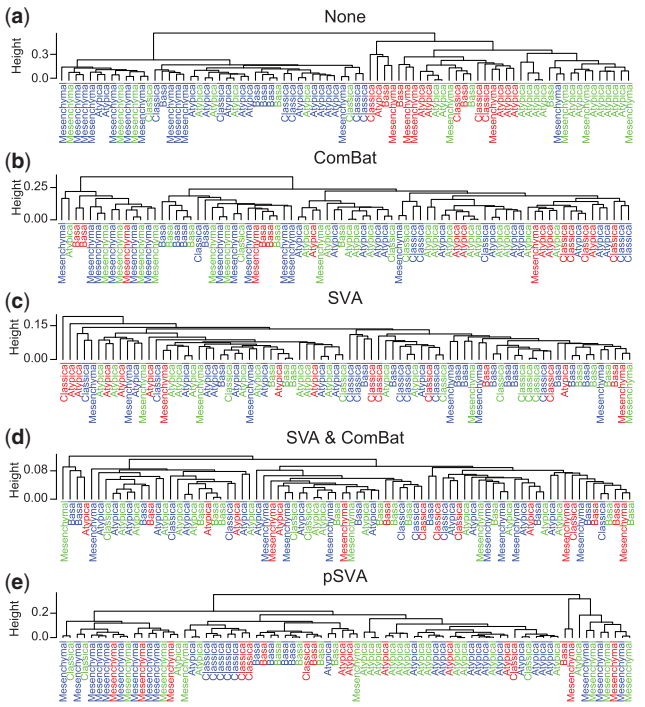


Fig. 3. Subtype identification from batch corrected data. Hierarchical clustering of batch-corrected genomic data from probes with maximum intratumor heterogeneity. Samples are colored according to the processing technique by which they were collected (blue for frozen samples processed with Ovation amplification kits, red for FFPE samples processed with FFPE amplification kits and green for frozen samples processed with FFPE amplification kits) and labeled according to the subtype (i.e. basal, mesenchymal, atypical and classical) inferred by applying the Walter *et al*. (2013) classifier to the batch-corrected data. Clustering is performed on data (a) without batch correction, (b) with ComBat, (c) with SVA, (d) with SVA combined with ComBat, and (e) with pSVA

3.3 pSVA preserved validated clinical subtypes in HNSCC

The technical artifacts in gene expression data most significantly impacted clustering (Supplementary Fig. S1), potentially confounding subtype identification algorithms. To explore the preservation of subtypes after batch correction, we applied hierarchical clustering (Fig. 3). In each dataset, we selected probes with larger average between sample variation than average between replicate variation to mitigate the impact of technical artifacts on inferred clusters, as described by Chung *et al*. (2004). We then compared the relationship of inferred clusters with batch and to established HNSCC subtypes inferred with the Walter *et al*. (2013) classifier. In Chung *et al*. (2004) four subtypes of HNSCC were described based on gene expression with distinct molecular characteristics: Group 1 with high Epidermal Growth Factor Receptor (EGFR) and its ligand expressing tumors, Group 2 with characteristics of epithelial-to-mesenchymal transition, Group 3 with normal mucosal epithelium-like expression profile and Group 4 with an upregulation of xenobiotic metabolism seen in heavy smokers. Subsequently, these four Groups were termed: Group 1—Basal, Group 2—Mesenchymal, Group 3—Atypical and Group 4—Classical. Additional analysis indicated Atypical tumors included the majority of the HPV-positive HNSCC.

Despite the probe selection, batch dominated clusters identified from the raw data (Fig. 3a), with 61% of merged samples being in the batch. Nonetheless, within each sample groups closely matched documented HNSCC subtypes, with 79% of merged groups being assigned to the same sample. Applying SVA alone (Fig. 3c) or in combination with ComBat (Fig. 3d) best removed the relationship between batch and clusters, with 19 and 8% of samples merged from the same batch, respectively. Consistent with the protected groups in its model, applying SVA alone or in combination with ComBat retained only a cluster containing the Atypical group, characterized by a predominance of HPV-positive samples. However, both of these techniques also removed the relationship between other HNSCC subtypes, with only 46% of merged samples being of the same subtype for SVA and 44% for the combination of SVA and ComBat.

On the other hand, ComBat (Fig. 3b) and pSVA (Fig. 3e) removed the association of clusters with batch with 34% or merged samples from the same batch, and also preserved sample subtypes. pSVA had higher similarity of subtypes in merged samples than ComBat (78% versus 72%). Notably, pSVA preserves these groups without encoding HPV status, which protects for the separation of 'atypical' HNSCC. Furthermore, the apparent clustering of 'classical' and 'atypical' subtypes in pSVA is consistent with recent suggestion of subtypes within HPV-positive (Atypical) HNSCC (Walter *et al*., 2013).

Despite the apparent differences in inferred clusters, HNSCC subtypes inferred from pSVA closely matched those inferred in the non–batch-corrected data (9% discrepancy from pSVA) or ComBat-corrected data (8% discrepancy). Applying subtype classification to SVA or SVA and ComBat-corrected data

Table 2. Number of samples of each HPV status across batches

Test set	None (%)	ComBat (%)	SVA (%)	SVA & ComBat (%)	pSVA (%)	HPV (%)
Frozen Nugen Ovation	88	91	91	94	76	71
FFPE Nugen FFPE	69	75	75	88	81	75
Frozen Nugen FFPE	74	63	74	81	70	52
GSE6791	86	83	81	88	86	62

Notes: Columns labeled with 'unique' count the number of samples from distinct tumors in that batch, tumors with replicate samples in other batches.

yielded subtype inference with 50 and 63% discrepancy from subtypes inferred in the pSVA-corrected data, respectively.

3.4 Batch correcting training data enhanced cross-batch and cross-study prediction accuracy

We extended our study to test prediction accuracy of the PAM classifier when including the independent dataset from Pyeon *et al*. (2007), GSE6791, containing additional frozen samples to assess the prediction accuracy in future samples subject to new batch effects. Specifically, we tested the accuracy of HPV status predicted for each batch from PAM classifiers trained on the remaining three independent batches (Table 2). Even without batch correction, the prediction accuracy was higher than the median cross-validation accuracy for the independent dataset, GSE6791 (86%) and the frozen samples processed with the Nugen Ovation Amplification kit (88%) but degraded for the samples processed with the Nugen FFPE Amplification kit (69% for FFPE samples and 74% for frozen samples).

We applied pSVA to batch correct the entire dataset and SVA and/or ComBat only to the training data. Batch correction was most essential to improving the prediction accuracy of the FFPE samples above that of a naive classifier that assigns all samples to be HPV-negative. In all cases, the combination of SVA and ComBat most accurately predicted HPV status. All other techniques had mixed accuracy depending upon the batch. Notably, applying pSVA to batch correct the test set improved the accuracy most for FFPE samples, which have the most technical artifacts (Supplementary Fig. S2) and class imbalance (Table 2).

3.5 pSVA batch correction stabilized prediction accuracy in classifiers trained on samples with high confounding between batch and HPV status

In the previous examples, each training and test set contain samples with widely different proportions of batch and HPV status, potentially introducing confounding which altered results of prediction accuracy. To control for this effect, we generated training sets selected from GSE6791 (Pyeon *et al*., 2007) and Frozen Nugen Ovation samples that were representative of the distribution of HPV-positive and -negative HNSCC in these batches (Table 1) at various levels of confounding between batch and

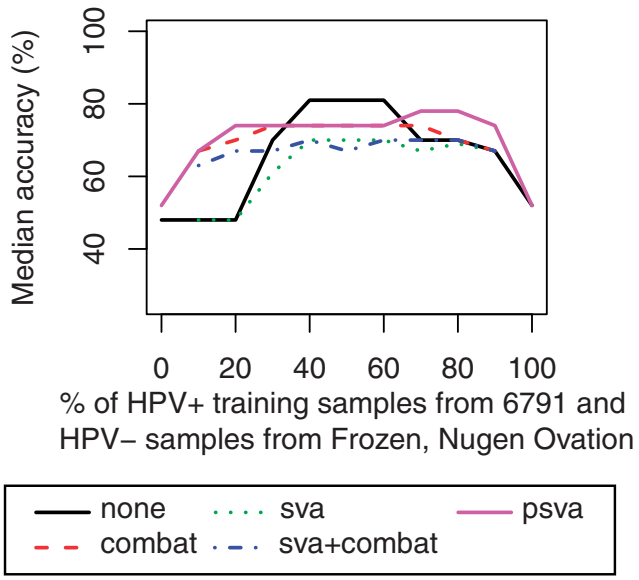


Fig. 4. Classification accuracy with confounded data. Median accuracy of classifiers of HPV-status trained on subsets of batch-corrected data simulated at varying levels of confounding, tested on independent frozen samples processed with the Nugen_FFPE amplification kit

biology. We tested the accuracy of the resulting PAM classifiers of HPV status before and after batch correction on the well-balanced Frozen Nugen FFPE samples.

We observed that prediction accuracy depended significantly on confounding, with PAM yielding median prediction accuracy below that for the naïve classifier that classifies everything as HPV-negative when 8 of 10 HPV-positive samples were selected from GSE6791 (Pyeon *et al*., 2007) and 16 of 20 HPV-negative from Frozen Nugen Ovation samples (Fig. 4). However, simply balancing the HPV status with batch increased the median prediction accuracy to that observed with both SVA and ComBat when developing the HPV classifier on the entire dataset excluding Frozen Nugen FFPE samples (Table 2). Both ComBat and pSVA stabilized the median prediction accuracy (Fig. 4) and variability in prediction accuracy across levels of confounding (Supplementary Fig. S6). Although the combination of ComBat and SVA stabilized median prediction accuracy, the accuracy was lower and more variable than that observed for ComBat or pSVA. Notably, pSVA improved median prediction accuracy above ComBat at high levels of confounding and was the only well-defined algorithm at 100% confounding, improving median prediction accuracy to levels observed for the naïve classifier. SVA consistently degraded prediction accuracy below that observed without batch correction. Moreover, no bias correction techniques yielded classifiers that matched the high median accuracy observed through simply having a balanced study design.

4 DISCUSSION

Batch correction techniques have been well established for class comparison analysis of genomic data. Nonetheless, we observed that these algorithms potentially removed true biological heterogeneity to facilitate such comparison. Therefore, we developed a

new batch correction technique, pSVA, which reversed the typical SVA (Leek *et al*., 2010) process to infer dominant sources of signal that is independent of known sources of technical artifacts. This algorithm simultaneously reduced differences in replicate samples arising from technical artifacts while maintaining differences between unique samples not observed with other batch correction techniques. Moreover, hierarchical clustering of pSVA-corrected data clearly delineated established HNSCC subtypes (Chung *et al*., 2004; Walter *et al*., 2013) from batch-affected gene expression data. The pSVA classification further divided the Atypical group into two primary clusters, consistent with the recent observation of the association of the Atypical group with HPV-positive samples (Walter *et al*., 2013) and of subtypes within HPV-positive HNSCC (Keck *et al*., 2013). Moreover, encoding some of the known subgroups delineated by HPV status in our model did not substantially alter the batch-corrected data with pSVA. Therefore, we anticipate that future application of pSVA to genomics data from other human cancers without a priori knowledge of diseases will similarly retain patient-specific heterogeneity to facilitate class discovery. For example, encoding cancer types as batch with pSVA, similar to the model employed for batch correction in the TCGA PAN-CANCER project (Ciriello *et al*., 2013), could better identify the molecularly distinct subtypes across human cancers that are required for personalized therapeutic selection. Beyond subtype identification, the pSVA algorithm may also facilitate more subtle patterns in genomic data using pattern-finding algorithms (Fertig *et al*., 2010), particularly adept for time course data. For example, pSVA may enable the detection of the complex, interacting dynamic signals in time course data without the prior knowledge of specific time groups required for batch correction in Colantuoni *et al*. (2011) and should be explored in future work.

An additional goal for genomics data is the development of classifiers to infer clinical groups in future genomics datasets that better assess prognosis or improve therapeutic selection. Implementation of such personalized genomics requires application of classification rules learned on training data to independent test data, for which phenotype is by definition unknown a priori. Because pSVA does not require knowledge of these groups, the algorithm can naturally be applied to also batch correct test data in such class discovery problems, and should be compared with other emerging techniques including fSVA (Parker *et al*., 2013). Consistent with our previous findings (Leek *et al*., 2010; Parker and Leek, 2012), we find that a well-designed and balanced training set yields classifiers with the most accurate prediction results. When training sets were large and represented subtypes well across batches, batch correcting the training data alone using SVA and ComBat yielded the most accurate classifiers. However, these algorithms overcorrected the data for smaller training sets, thereby degrading the accuracy of genomics classifiers. Moreover, pSVA was notably the only technique able to accurately predict HPV status in the presence of substantial confounding between technical artifacts and sample groups. This result suggests that pSVA may facilitate the development of genomic biomarkers by combining public datasets provided that at least one such dataset contains a balanced sample design. Moreover, the ability of pSVA to correct for batch effects between frozen and FFPE samples may facilitate the development of classifiers for human cancers, in which the need to prioritize collection of FFPE samples from tissue availability often introduces an imbalance in samples available to train genomic classifiers. Nonetheless, careful cross-validation similar to that used in Parker and Leek (2012) must be performed to assess the appropriate batch correction technique for each dataset prior to building a genomic classifier for use on independent datasets.

Taken together, our results suggest that the batch correction algorithm for genomics should be selected on the basis of the desired analysis goal. Briefly, these results suggest the use of the combination of SVA and ComBat for differential expression analysis and class discovery algorithms. However, pSVA gains an advantage for class discovery when the biological covariates represented in some of the genomic datasets are highly confounded with batch. Moreover, pSVA also yields the most consistent clusters, suggesting its use for class discovery problems. We note that alternative approaches for class comparison, including Cancer Outlier Profile Analysis (MacDonald and Ghosh, 2006), that account for biological heterogeneity may likewise benefit from use of pSVA for batch correction. Because both pSVA and ComBat require prior knowledge of batch variables, either SVA (Leek and Storey, 2007) or control probe-based techniques (Gagnon-Bartsch and Speed, 2012) are optimal for cases in which the batch information is unknown a priori. The lack of biological information in control-probes makes algorithms that use them less likely to remove biological signal, making them possible alternatives to pSVA for class discovery when batch is unknown.

The lack of technical artifacts in genes relevant to HNSCC biology was essential to the accurate subtype identification observed without batch correction. This may also explain the accuracy of cross-study and cross-batch prediction of HPV status from classifiers inferred when only the training data was batch corrected. However, we cannot assume the generality of these results in other systems (Parker *et al*., 2013). Alternative batch correction algorithms based upon control probes similarly preserve hierarchical clusters and sample heterogeneity (Gagnon-Bartsch and Speed, 2012) may be more adept at correcting for technical artifacts when making inference of biological subtypes directly from batch affected genes. However, the lack of control genes in RNA-seq or common controls across array technologies make pSVA a better candidate algorithm for preserving sample subtypes in cross-platform and RNA-seq analyses. Therefore, future extensions of pSVA are needed to account for technical artifacts when a gene or set of genes used for sample classification are themselves subject to batch effects.

ACKNOWLEDGEMENTS

We thank Vonn Walter and Neil Hayes for providing the subtype classifier and Michael F. Ochs, Ludmila Danilova, Luigi Marchioni, Robert Scharpf and Rafael Guerro-Preston for advice.

Funding: This work was funded by NIH NCI (CA141053), the Cleveland Foundation, and the NIH NCI SPORE in Head and Neck Cancer at the Johns Hopkins University to E.J.F. The project was funded in part by NIH NIDCR R01 (DE017982) to C.H.C and RBRF 13-04-40279-H, Johns Hopkins University

Framework for the Future, and the Commonwealth Foundation to A.V.F.

Conflict of Interest: none declared.

REFERENCES

Chung,C.H. *et al.* (2004) Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. *Cancer Cell*, **5**, 489–500.

Chung,C.H. *et al.* (2010) Nuclear factor-kappa b pathway and response in a phase ii trial of bortezomib and docetaxel in patients with recurrent and/or metastatic head and neck squamous cell carcinoma. *Ann. Oncol.*, **21**, 864–870.

Chung,C.H. *et al.* (2011) Insulin-like growth factor-1 receptor inhibitor, amg-479, in cetuximab-refractory head and neck squamous cell carcinoma. *Head Neck*, **33**, 1804–1808.

Ciriello,G. *et al.* (2013) Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.*, **45**, 1127–1133.

Cohen,E.E.W. *et al.* (2009) A feed-forward loop involving protein kinase calpha and micrornas regulates tumor cell cycle. *Cancer Res.*, **69**, 65–74.

Colantuoni,C. *et al.* (2011) Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature*, **478**, 519–523.

Fertig,E.J. *et al.* (2010) Cogaps: an r/c++ package to identify patterns and biological process activity in transcriptomic data. *Bioinformatics*, **26**, 2792–2793.

Gagnon-Bartsch,J.A. and Speed,T.P. (2012) Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, **13**, 539–552.

Gilbert,J. *et al.* (2012) Phase 2 trial of oxaliplatin and pemetrexed as an induction regimen in locally advanced head and neck cancer. *Cancer*, **118**, 1007–1013.

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

Keck,M.K. *et al.* (2013) Genomic profiling of kinase genes in head and neck squamous cell carcinomas to identify potentially targetable genetic aberrations in fgfr1/2, ddr2, epha2, and pik3ca. *J. Clin. Oncol.*, **31**, 365s.

Leek,J.T. and Storey,J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, **3**, 1724–1735.

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. *et al.* (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, **28**, 882–883.

Luo,J. *et al.* (2010) A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-ii microarray gene expression data. *Pharmacogenomics J.*, **10**, 278–291.

MacDonald,J.W. and Ghosh,D. (2006) Copa–cancer outlier profile analysis. *Bioinformatics*, **22**, 2950–2951.

McCall,M.N. *et al.* (2010) Frozen robust multiarray analysis (fRMA). *Biostatistics*, **11**, 242–253.

Parker,H.S. and Leek,J.T. (2012) The practical effect of batch on genomic prediction. *Stat. Appl. Genet. Mol. Biol.*, **11**, Article 10.

Parker,H.S. *et al.* (2013) Removing batch effects for prediction problems with frozen surrogate variable analysis. *arXiv*, 1301.3947.

Pyeon,D. *et al.* (2007) Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers. *Cancer Res.*, **67**, 4605–4619.

Robinson,M. *et al.* (2010) Refining the diagnosis of oropharyngeal squamous cell carcinoma using human papillomavirus testing. *Oral Oncol.*, **46**, 492–496.

Slebos,R.J.C. *et al.* (2006) Gene expression differences associated with human papillomavirus status in head and neck squamous cell carcinoma. *Clin. Cancer Res.*, **12** (3 Pt 1), 701–709.

Smeets,S.J. *et al.* (2007) A novel algorithm for reliable detection of human papillomavirus in paraffin embedded head and neck cancer specimen. *Int. J. Cancer*, **121**, 2465–2472.

Sun,Z. *et al.* (2011) Batch effect correction for genome-wide methylation data with illumina infinium platform. *BMC Med. Genomics*, **4**, 84.

Tibshirani,R. *et al.* (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

Viljoen,K.S. and Blackburn,J.M. (2013) Quality assessment and data handling methods for affymetrix gene 1.0 ST arrays with variable RNA integrity. *BMC Genomics*, **14**, 14.

Walter,V. *et al.* (2013) Molecular subtypes in head and neck cancer exhibit distinct patterns of chromosomal gain and loss of canonical cancer genes. *PLoS One*, **8**, e56823.