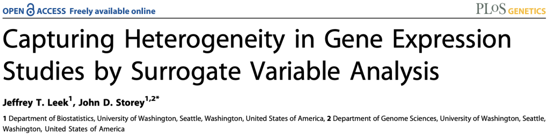
BIOST 434 - Batch Correction with pSVA

Zhining Sui

Department of Biostatistics and Computational Biology, University of Rochester

March 6, 2024

SVA (Leek and Storey, 2007):



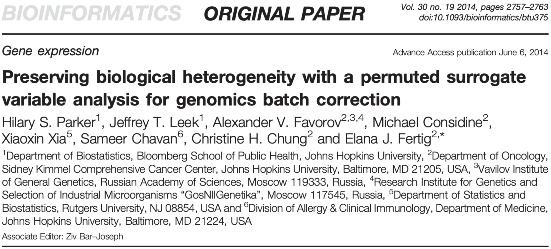
OPEN ACCESS Freely available online PLoS GENETICS

Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis

Jeffrey T. Leek[1], John D. Storey[1,2*]

1 Department of Biostatistics, University of Washington, Seattle, Washington, United States of America, 2 Department of Genome Sciences, University of Washington, Seattle, Washington, United States of America

pSVA (Parker et al., 2014):



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

Overview

1 Introduction

2 Algorithms
- Surrogate Variable Analysis (SVA) (Leek and Storey, 2007)
- Permuted-SVA (pSVA) (Parker et al., 2014)

3 Results

Introduction

Differential expression studies characterize transcriptional variation with respect to primary variables by fitting a model relating the expression for each gene to the primary variables.

Two general sources of expression variation:

1. Expression variation introduced by primary variable (may or may not exist).
2. Expression variation introduced by noise:
   - Unmodeled factors: measured variables, but are not included in the statistical model (statistically intractable to determine which to include in the model or too complicated to model).
   - Unmeasured factors: not measured in the study.

     } Batch effect
   - Gene-specific noise: random fluctuations in gene expression independently from gene to gene.

Introduction: Expression Heterogeneity

Expression heterogeneity (EH): patterns of variation due to any unmodeled factor.

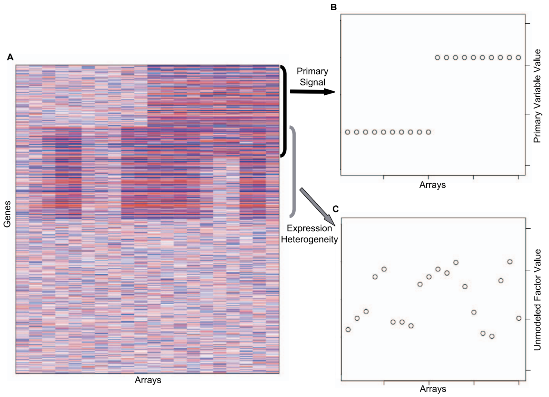


Figure 2. Example of Expression Heterogeneity

(A) A heatmap of a simulated microarray study consisting of 1,000 genes measured on 20 arrays.
(B) Genes 1–300 in this simulated study are differentially expressed between two hypothetical treatment groups; here the two groups are shown as an indicator variable for each array.
(C) Genes 201–500 in each simulated study are affected by an independent factor that causes EH. This factor is distinct from, but possibly correlated with, the group variable. Here, the factor is shown as a quantitative variable, but it could also be an indicator variable or some linear or nonlinear function of the covariates.

Signal dependence
(may or may not exist)

Noise dependence

Introduction: Addressing Noise Dependence

Consider m genes, n samples ($n < m$), 1 primary variable:

$\boldsymbol{X}_{m\times n} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_m)^T$: normalized expression matrix.
$\boldsymbol{y} = (y_1, ..., y_n)^T$: primary variable of interest.

The expression for gene i on sample j:

$$x_{ij} = \mu_i + f_i(y_j) + \epsilon_{ij}, \;\; E[\epsilon_{ij}] = 0$$

where $f_i(y_j)$ gives the relationship between gene i and the primary variable. $f(\cdot)$ is almost always an additive or linear model in high-dimensional data analysis.

Any structure among the $\boldsymbol{\epsilon}_{m\times n} = \{\epsilon_{ij}\}$ constitutes **noise dependence** caused by unmeasured or unmodeled factors.

Introduction: Addressing Noise Dependence

We want to study the dependence in $\epsilon_{m\times n}$.

A direct approach is to estimate the linear dependence between the genes using the covariance matrix, $\Sigma_{m\times m}$.

However, in a gene expression experiment, the number of genes, m, is on the order of thousands, giving millions of elements, $\frac{(m+1)m}{2}$, to estimate in the covariance matrix. Also, the number of samples is usually no more than several hundred.

Another approach is to decompose the variability in $\epsilon_{m\times n}$ into a dependent and an independent component, explained by unmodeled factors and gene-specific noise, respectively.

SVA Model Leek and Storey (2007)

Consider L biologically meaningful unmodeled factors:

Decomposition of the random noise:

$$\epsilon_{ij} = \sum_{\ell=1}^{L} \gamma_{\ell i} g_{\ell j} + \epsilon_{ij}^{*},$$

where $g_{\ell j}$ is an arbitrarily complicated function of the ℓth unmodeled factor of sample j and $\gamma_{\ell i}$ is a gene-specific coefficient for the ℓth unmodeled factor.

Here, the linear term $\sum_{\ell=1}^{L} \gamma_{\ell i} g_{\ell j}$ is the dependent variation across genes due to unmodeled factors and ϵ_{ij}^{*} is the independent true gene-specific noise. All of the noise dependence across features is quantified by the term $\sum_{\ell=1}^{L} \gamma_{\ell i} g_{\ell j}$.

SVA: Estimating the unmodeled variables

Now, the expression for gene i on sample j is

$$x_{ij} = \mu_i + f_i(y_j) + \sum_{\ell=1}^{L} \gamma_{\ell i} g_{\ell j} + \epsilon^*_{ij}$$

It is not possible to directly estimate $\boldsymbol{g}_\ell = (g_{\ell 1}, ..., g_{\ell n})$ for $\ell = 1, 2, .., L$ unmodeled factors. One intuitive idea is to find a set of statistically tractable **surrogate variables** that gives the same overall effects on the gene expression as the unmodeled factors.

Key observation: For any set of vectors $\boldsymbol{g}_\ell$, it is possible to identify mutually orthogonal vectors $\boldsymbol{h}_k = (h_{k1}, ..., h_{kn})$, $k = 1, ..., K$ $(K \leq L)$ that spans the same linear space as $\boldsymbol{g}_\ell$, i.e.,

$$\sum_{\ell=1}^{L} \gamma_{\ell i} g_{\ell j} = \sum_{k=1}^{K} \lambda_{ki} h_{kj}.$$

SVA: Estimating the unmodeled variables

Now, the expression for gene *i* on sample *j* is

$$x_{ij} = \mu_i + f_i(y_j) + \sum_{k=1}^{K} \lambda_{ki} h_{kj} + \epsilon^*_{ij}.$$

Goal: identify and estimate the surrogate variables $\boldsymbol{h}_k$ based on consistent patterns of expression variation.

Requirements for surrogate variables:

- represent signal due to sources other than the primary variable.
- allow for potential overlap with the primary variable.

Intuition: Directly estimate surrogate variables that explain a large proportion of the consistent variation across features from the data.

SVA: Estimating the surrogate variables

Singular Value Decomposition *Any real $m \times n$ matrix $\boldsymbol{X}$ has a decomposition $\boldsymbol{X} = \boldsymbol{UDV}^T$, where the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary with dimensions $m \times m$ and $n \times n$, respectively. The columns of $\boldsymbol{U}$ are called the left singular vectors of $\boldsymbol{X}$ and the columns of $\boldsymbol{V}$ are called the right singular vectors of $\boldsymbol{X}$. The matrix $\boldsymbol{D}$ is diagonal with non-negative elements. The number of non-zero diagonal elements of $\boldsymbol{D}$ is equal to the rank of $\boldsymbol{X}$.*

SVD is designed to partition variation in rows and columns of the matrix $\boldsymbol{X}$ into linear components that represent maximal variation.

SVA: Estimating the surrogate variables

Alter et al. (2000) showed that the right singular vectors, or eigengenes, represent trends that account for a large proportion of the variation in the expression matrix.

The surrogate variable estimates are based on the **right non-zero singular vectors**.

The SVD is applied rather than some other matrix decomposition because the SVD is both easy to compute and is optimal in the sense of providing the least squares solution among all bilinear fits to the data.

SVA Algorithm

Two parts of the algorithm:

- Detection of unmodeled factors (estimating the number of surrogate variables).
- Construction of surrogate variables.

SVA Algorithm 1: Estimating the number of surrogate variables

SVA uses a permutation algorithm (Buja and Eyuboglu, 1992) to estimate the number of significant right singular vectors. "Significant" means that the eigengene represents a greater proportion of variation than expected by chance.

Test statistic T_k for eigengene k: the proportion of variation explained by the kth right singular vector.

The observed matrix is assumed to be obtained from a random sample of n subjects drawn from a large population: it is a realization of random variables x_{ij} where the random vectors of gene expression for each sample $(x_{1j}, ..., x_{mj})$ are independently and identically distributed for samples $j = 1, ..., n$.

SVA Algorithm 1: Estimating the number of surrogate variables

Null hypothesis: there exists no dependence between gene expressions in each sample (i.e., rows are independently distributed).

Under this null hypothesis, the variables in the matrix are independently distributed, and the variables within row i share a common marginal distribution F_i, which may be different from row to row.

The joint null distribution of the matrix is invariant under permutation within the rows (i.e., each of the $n!^m$ permuted matrix is equally likely).

SVA Algorithm 1: Estimating the number of surrogate variables

By randomly permuting each individual row of the matrix, any structure across features is eliminated. So, the null distribution of the test statistic conditional on the observed data can be obtained by calculating all possible values of the test statistic under all permutations of the observed data.

However, in the actual computation, a full enumeration of all permutations is infeasible, therefore we use Monte Carlo simulations in which an affordable number B of row permutations is sampled with replacement.

Count how often permutation statistics exceed the observed ones.

SVA Algorithm 1: Estimating the number of surrogate variables

Algorithm to detect unmodeled factors:

1. Fit the model $x_{ij} = \mu_i + f_i(y_j) + \epsilon_{ij}$ to get $\hat{\mu}_i$ and $\hat{f}_i$. Calculate the residual matrix $\boldsymbol{R}_{m\times n}$ whose the (i,j)-th element is $r_{ij} = x_{ij} - \hat{\mu}_i - \hat{f}_i(y_i)$.
2. Calculate SVD: $\boldsymbol{R} = \boldsymbol{UDV}^T$.
3. Denote the ℓth eigenvalue as d_ℓ for $\ell = 1, ..., n$. The observed statistic for the kth eigengene is:

$$T_k = \frac{d_k^2}{\sum_{\ell=1}^{n-df} d_\ell^2}.$$

4. Permute each row of $\boldsymbol{R}$ independently to form $\boldsymbol{R}^* = \{r_{ij}^*\}$.

SVA Algorithm 1: Estimating the number of surrogate variables

Algorithm to detect unmodeled factors:

5. Fit the model $r_{ij}^* = \mu_i^* + f_i^*(y_j) + \epsilon_{ij}^*$. Calculate the $m \times n$ model-subtracted null residual matrix $\boldsymbol{R}_0$ whose the (i, j)-th element is $r_{ij}^0 = r_{ij}^* - \hat{\mu}_i^* - \hat{f}_i^*(y_i)$.
6. Calculate the SVD: $\boldsymbol{R}_0 = \boldsymbol{U}_0 \boldsymbol{D}_0 \boldsymbol{V}_0^T$.
7. Form a null statistic for eigengene k:

$$T_k^0 = \frac{d_{0k}^2}{\sum_{\ell=1}^{n-df} d_{0\ell}^2}.$$

SVA Algorithm 1: Estimating the number of surrogate variables

Algorithm to detect unmodeled factors:

8. Repeat steps 4-7 for B times to obtain null statistics T_k^{0b}.
9. Compute the p-value for eigengene k as

$$p_k = \frac{\#\{T_k^{0b} \geq T_k; b = 1, ..., B\}}{B}.$$

10. Eigengene k is a significant signature of residual expression heterogeneity if $p_k \leq \alpha$.

SVA Algorithm 2: Estimating surrogate variables

The surrogate variables are estimated based on the right non-zero singular vectors of ...?

Requirements for surrogate variables:

1. represent signal due to sources other than the primary variable.
2. allow for potential overlap with the primary variable.

SVA Algorithm 2: Estimating surrogate variables

The surrogate variables are estimated based on the right non-zero singular vectors of ...?

Requirements for surrogate variables:

1. represent signal due to sources other than the primary variable.
2. allow for potential overlap with the primary variable.

Run SVD on the original expression matrix?

SVA Algorithm 2: Estimating surrogate variables

The surrogate variables are estimated based on the right non-zero singular vectors of ...?

Requirements for surrogate variables:

1. represent signal due to sources other than the primary variable.
2. allow for potential overlap with the primary variable.

Run SVD on the original expression matrix?
⇒ Violates Requirement 1: Surrogate variable estimates are biased towards the primary variable signal.

SVA Algorithm 2: Estimating surrogate variables

The surrogate variables are estimated based on the right non-zero singular vectors of ...?

Requirements for surrogate variables:

1. represent signal due to sources other than the primary variable.
2. allow for potential overlap with the primary variable.

Run SVD after regressing out the primary variable (i.e., on the residual gene expression matrix)?

SVA Algorithm 2: Estimating surrogate variables

The surrogate variables are estimated based on the right non-zero singular vectors of ...?

Requirements for surrogate variables:

1. represent signal due to sources other than the primary variable.
2. allow for potential overlap with the primary variable.

Run SVD after regressing out the primary variable (i.e., on the residual gene expression matrix)?
$\Rightarrow$ Violates Requirement 2: Surrogate variables estimates are orthogonal to the primary variables.

SVA Algorithm 2: Estimating surrogate variables

The surrogate variables are estimated based on the right non-zero singular vectors of **a carefully defined subset of the expression matrix.**

Combining the two previous approaches:

1. Identify the subset of genes that are only affected by the unmodeled variables. (Identify eigengenes of the residual expression matrix and identify subsets of genes that are most highly associated with each residual eigengenes).
2. Estimate common patterns in the original expression of this subset of genes using SVD.

SVA Algorithm 2: Estimating surrogate variables

This approach combines the advantages of the two previous methods.

First, the signatures of the unmodeled factors are identified in the residuals, so the selection of the subset of genes for each surrogate variable is not affected by signal from the primary variable. Each subset is enriched for genes showing strong association with the corresponding surrogate variable. The maximal source of variation within that subset is then likely to be the surrogate variable of interest.

By calculating the SVD on the original expression data for that subset of genes, this approach also allows for correlation with the primary variable.

SVA Algorithm 2: Estimating surrogate variables

Algorithm to construct surrogate variables:

1. Fit the model $x_{ij} = \mu_i + f_i(y_j) + \epsilon_{ij}$ and calculate the residual matrix $\boldsymbol{R}_{m \times n}$.
2. Calculate SVD: $\boldsymbol{R} = \boldsymbol{UDV}^T$. Let $\boldsymbol{\epsilon}_k = (\epsilon_{k1}, ..., \epsilon_{kn})^T$ be the column of $\boldsymbol{V}$. These residual eigengenes represent the orthogonal residual expression heterogeneity signals independent of the signal due to the primary variable.
3. Run previous algorithm to get the number of significant eigengenes, $\hat{K}$.

SVA Algorithm 2: Estimating surrogate variables

Algorithm to construct surrogate variables:
For each significant eigengene $\epsilon_k, k = 1, ..., \hat{K}$.

4. Regress ϵ_k on the $\boldsymbol{x}_i$ $(i = 1, ..., m)$. Calculate the p-value for an association between the residual eigengene and each gene's expression.
5. Let π_0 denote the proportion of genes with expression not truly associated with ϵ_k. Estimate $\hat{\pi}_0$ (Storey and Tibshirani, 2003) and the number of genes associated with the eigengene by $\hat{m}_1 = \lfloor(1 - \hat{\pi}_0) \times m\rfloor$. Let $s_1, ..., s_{\hat{m}_1}$ be the indices of the genes with $\hat{m}_1$ smallest p-values from the test.

SVA Algorithm 2: Estimating surrogate variables

Algorithm to construct surrogate variables:
For each significant eigengene $\epsilon_k, k = 1, ..., \hat{K}$.

6. Form the $\hat{m}_1 \times n$ reduced expression matrix $\boldsymbol{X}_r = (\boldsymbol{x}_{s_1}, ..., \boldsymbol{x}_{s_{\hat{m}_1}})^T$. Calculate the right singular vectors of $\boldsymbol{X}_r$, and denote these by ϵ_j^r for $j = 1, ..., n$.
7. Let $j^* = \text{argmax}_{1 \leq j \leq n} \textbf{cor}(\epsilon_k, \epsilon_j^r)$ and set $\hat{\boldsymbol{h}}_k = \epsilon_{j^*}^r$.

After surrogate variables are estimated, they are treated as fixed independent variables in the model.

SVA Summary

- SVA developed an iterative procedure to account for noise dependence in high throughput experiments.
- The framework is based on the idea of decomposing the dependence between features into a set of common shared factors that can be estimated from the data directly.
- Estimating the fixed values of unmodeled factors rather than population covariance Σ reduces the computational cost.
- Instead of estimating the specific unmodeled factors, SVA estimates their linear combination, i.e., the net effect.
- The surrogate variables are calculated based on carefully selected subsets of genes in the original expression matrix that correspond to patterns observed in a residual expression matrix where the main effects of the primary variables have been removed.

Problem in SVA

SVA can remove unmodeled factors (e.g., technical artifacts) statistically, provided that primary variables (e.g. pertinent biological groups) are known and well-represented.

However, in many cases, true sources of biological heterogeneity (e.g. disease subtypes) are unknown, whereas the unmodeled factors are known.

Then, SVA will eliminate true (intragroup) biological heterogeneity since it is not encoded in the model as the primary variable.

So, novel patterns cannot be identified, limiting inference of either dynamics from time course genomics data or of novel disease subtypes or of any personalized genomic signatures.

Permuted-SVA (pSVA) Algorithm

Model in SVA:

$$\boldsymbol{D} \sim \boldsymbol{AP} + \boldsymbol{B\Gamma} + \epsilon$$

Notations:

$\boldsymbol{D}_{g \times s}$ = Gene expression matrix for g genes and s samples.
$\boldsymbol{P}$ = Matrix describing pertinent biological covariates.
$\boldsymbol{\Gamma}$ = Matrix corresponding to unmodeled factors (typically batch covariates).

pSVA reverses the standard application of SVA: It models biological heterogeneity as those surrogate variables estimated from genes that are not associated with known technical covariates in the model matrix $\boldsymbol{\Gamma}$.

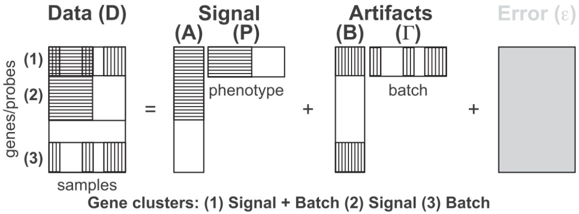
Permuted-SVA (pSVA) Algorithm

$$\boldsymbol{D} \sim \boldsymbol{AP} + \boldsymbol{B\Gamma} + \epsilon$$

pSVA inputs a model matrix of technical covariates ($\boldsymbol{\Gamma}$) and then uses the iterative procedure in SVA to estimate the net effect of factors spanning from genes not associated with these technical artifacts.

It identifies coefficients Ψ corresponding to an orthonormal set of vectors $\boldsymbol{\Theta}$ that span the same space as the unknown biological covariates, so that $\Psi\boldsymbol{\Theta} = \boldsymbol{AP}$.

Permuted-SVA (pSVA) Algorithm



	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

pSVA Applied to Human Expression Study

1. **Gene Expression Data**: 80 samples from retrospective profiling of head and neck squamous cell carcinoma (HNSCC) primary tumors over four years.
2. **Primary variable**: Human Papillomavirus (HPV) status (39 unique HPV-negative, 22 unique HPV-positive).
3. **Dominant batches**:
   - Sample procurement: Frozen tumors; FFPE tumors.
   - RNA amplification kit: Nugen FFPE (Nugen FFPE or FFPE_beta); Nugen Ovation (Nugen Ovation_1 or Ovation_2).
4. **Hidden biological heterogeneity**: HNSCC tumor subtypes.
   - Group 1 - Basal, Group 2 - Mesenchymal, Group 3 - Atypical and Group 4 - Classical (distinguishing HPV-positive HNSCC as the ‘atypical’ subtype).

Result 3.1. Sample procurement and RNA isolation drive dominant gene expression signals

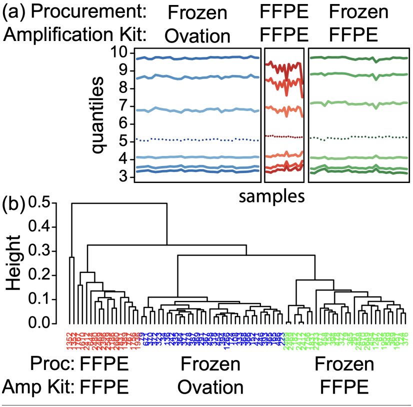


Figure: Supplementary Fig. S2.

Result 3.2. Removing technical artifacts with pSVA preserves sample heterogeneity

Method:
Remove technical artifacts from the amplification kit and procurement using pSVA, ComBat, SVA, and a combination of SVA and ComBat. Both SVA and ComBat model HPV status as the known biological covariate.

Result 3.2. Removing technical artifacts with pSVA preserves sample heterogeneity

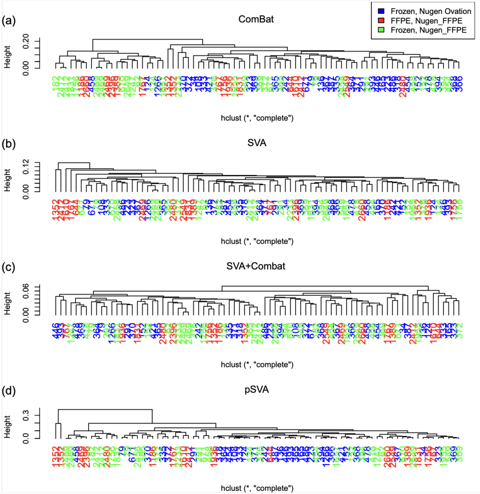


Figure: Supplementary Fig. S3.

Each algorithm successfully mixes sample-processing groups in hierarchical clusters.

Result 3.2. Removing technical artifacts with pSVA preserves sample heterogeneity

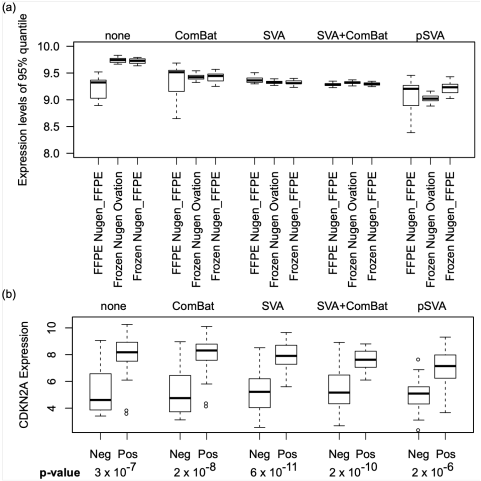


Figure: Supplementary Fig. S4.

SVA alone or in combination with ComBat successfully makes the distribution of expression levels more similar between FFPE and frozen samples.

Each algorithm preserves the differential expression of the established HPV biomarker in HNSCC, p16 (CDKN2A).

Result 3.2. Removing technical artifacts with pSVA preserves sample heterogeneity

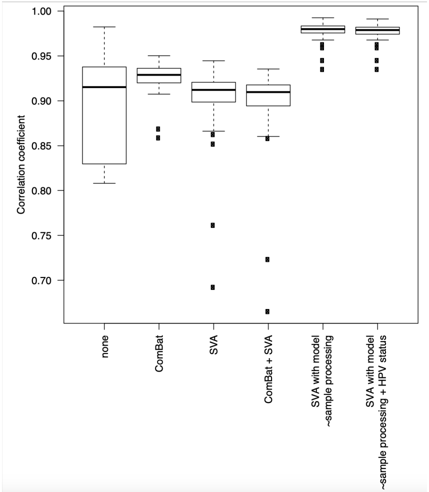


Figure: Supplementary Fig. S5.

Samples corrected with pSVA had the 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.

Result 3.2. Removing technical artifacts with pSVA preserves sample heterogeneity

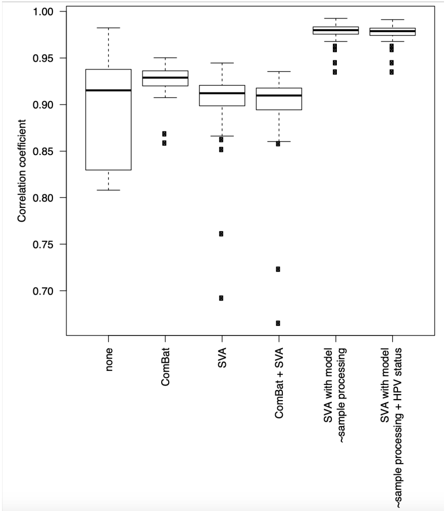


Figure: Supplementary Fig. S5.

Data corrected with pSVA was highly correlated to data corrected with SVA with both HPV status and batch in the model, suggesting that pSVA is retaining pertinent heterogeneity without any knowledge of biological groups.

Result 3.2. Removing technical artifacts with pSVA preserves sample heterogeneity

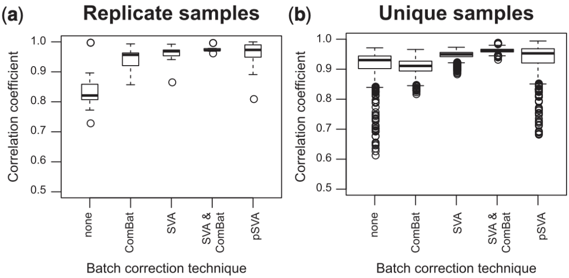


Figure: Fig. 2.

Before batch correction, there was substantial variation in expression profiles between replicate samples.

Each batch correction algorithm increased the correlation between replicate samples, with the greatest improvement observed when combining SVA and ComBat.

Result 3.2. Removing technical artifacts with pSVA preserves sample heterogeneity

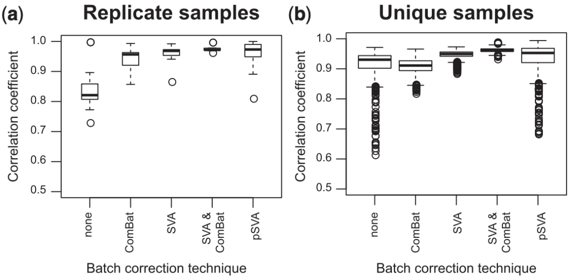


Figure: Fig. 2.

Only pSVA preserves the high heterogeneity between non-replicate samples reflective of tumor heterogeneity.

Result 3.3. pSVA preserved validated clinical subtypes in HNSCC

Method:
Applied hierarchical clustering.
Compared the relationship of inferred clusters with batch and to the established HNSCC subtypes.

Result 3.3. pSVA preserved validated clinical subtypes in HNSCC

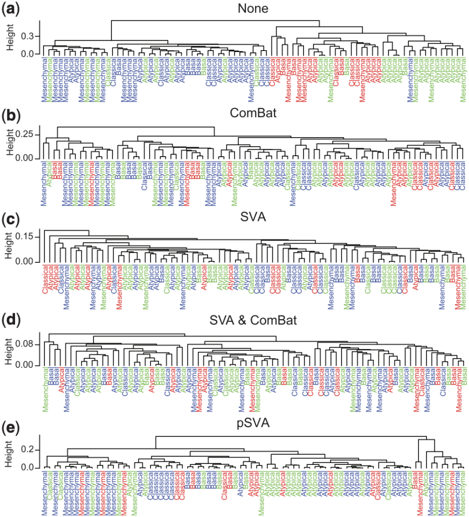


Figure: Fig. 3.

Raw data:

Batch dominated clusters identified.

Result 3.3. pSVA preserved validated clinical subtypes in HNSCC

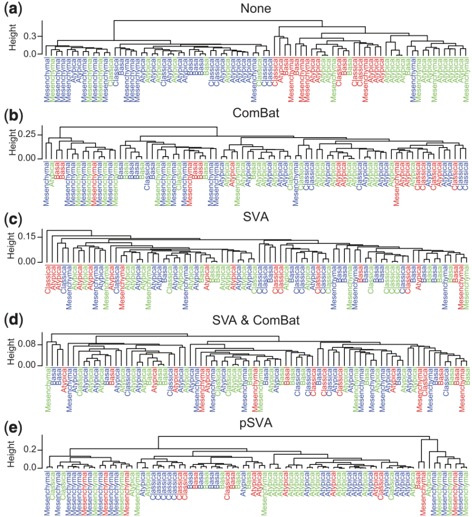


Figure: Fig. 3.

SVA alone or in combination with ComBat:

Best removed the relationship between batch and clusters.

Retained only a cluster containing the Atypical group, characterized by a predominance of HPV-positive samples.

Removed the relationship between other HNSCC subtypes.

Result 3.3. pSVA preserved validated clinical subtypes in HNSCC

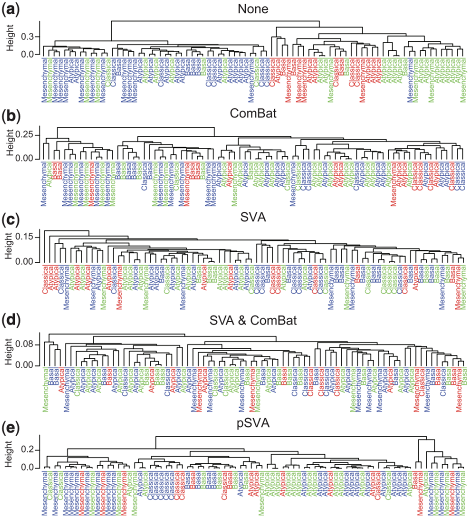


Figure: Fig. 3.

ComBat or pSVA:

Removed the association of clusters with batch and also preserved sample subtypes.

Subtypes inferred from pSVA closely matched those inferred in the non–batch-corrected data or ComBat-corrected data.

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

Method:

Developed a genomic classifier of HPV status using the PAM algorithm.

Considered data from GSE6791 as an additional batch to each of the three batches in the dataset.

Trained the PAM classifier with three of the four batches and tested it on data from the remaining batch, excluding any replicate samples from tumors also measured in the training set.

Applied ComBat, SVA, and the combination of SVA and ComBat to the training data.

Applied pSVA to batch correct both test and training data.

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

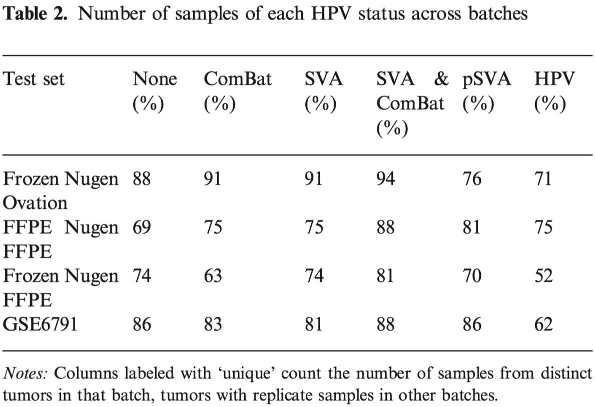


Table 2. Number of samples of each HPV status across batches

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.

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.

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

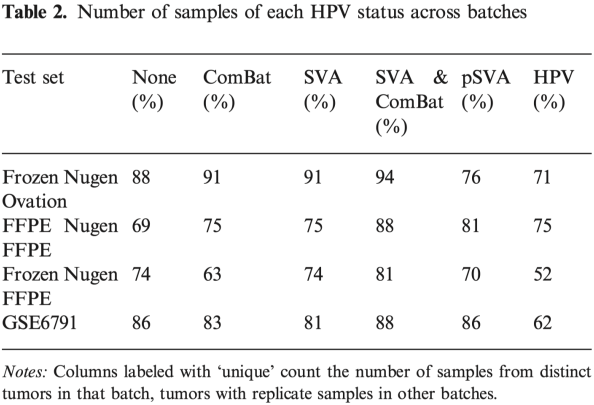


Table 2. Number of samples of each HPV status across batches

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.

In all cases, the combination of SVA and ComBat most accurately predicted HPV status.

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

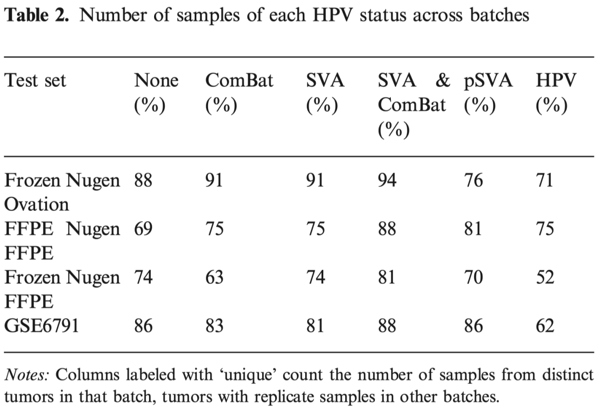


Table 2. Number of samples of each HPV status across batches

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.

Applying pSVA to batch correct the test set improved the accuracy most for FFPE samples, which have the most technical artifacts and class imbalance.

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

Method:

Training set: selected from GSE6791 (26 unique HPV-, 16 unique HPV+) and Frozen Nugen Ovation samples (24 unique HPV-, 10 unique HPV+) that were representative of the distribution of HPV-positive and -negative HNSCC in these batches at various levels of confounding between batch and biology.

Testing set: the well-balanced Frozen Nugen FFPE samples (52% HPV-positive).

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

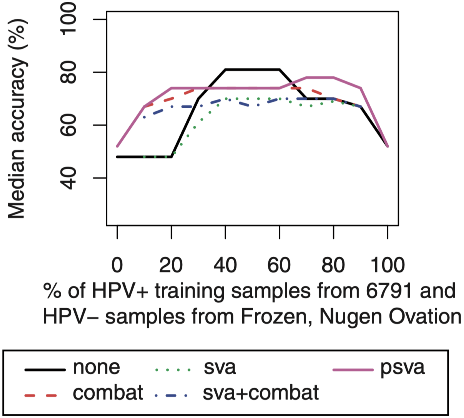


Fig. 4. Classification accuracy with confounded data. Median accuracy of classifiers of HPV-status trained on subsets of batch-corrected data simulated at varying levels of confounding, tested on independent frozen samples processed with the Nugen_FFPE amplification kit

Prediction accuracy depended significantly on confounding.

Simply balancing the HPV status with batch increased the median prediction accuracy.

No batch correction techniques yielded classifiers that matched the high median accuracy observed through the balanced study design.

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

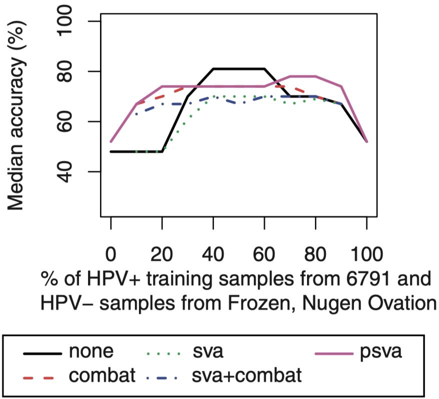


Fig. 4. Classification accuracy with confounded data. Median accuracy of classifiers of HPV-status trained on subsets of batch-corrected data simulated at varying levels of confounding, tested on independent frozen samples processed with the Nugen_FFPE amplification kit

Both ComBat and pSVA stabilized the median prediction accuracy and variability in prediction accuracy across levels of confounding.

pSVA improved median prediction accuracy above ComBat at high levels of confounding and was the only well-defined algorithm at 100% confounding.

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

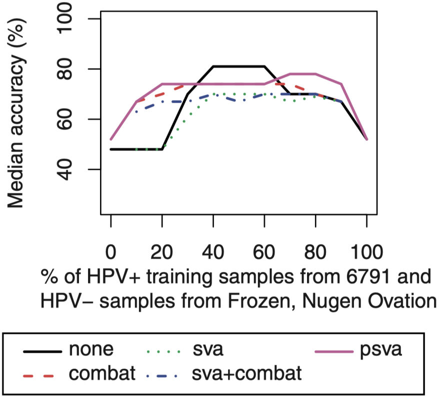


Fig. 4. Classification accuracy with confounded data. Median accuracy of classifiers of HPV-status trained on subsets of batch-corrected data simulated at varying levels of confounding, tested on independent frozen samples processed with the Nugen_FFPE amplification kit

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.

SVA consistently degraded prediction accuracy below that observed without batch correction.

References

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U. S. A.*, 97(18):10101–10106.

Buja, A. and Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behav. Res.*, 27(4):509–540.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, 3(9):1724–1735.

Parker, H. S., Leek, J. T., Favorov, A. V., Considine, M., Xia, X., Chavan, S., Chung, C. H., and Fertig, E. J. (2014). Preserving biological heterogeneity with a permuted surrogate variable analysis for genomics batch correction. *Bioinformatics*, 30(19):2757–2763.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.*, 100(16):9440–9445.