BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 6 2012, pages 882–883
doi:10.1093/bioinformatics/bts034

Gene expression

Advance Access publication January 17, 2012

The sva package for removing batch effects and other unwanted variation in high-throughput experiments

Jeffrey T. Leek[1,*], W. Evan Johnson[2], Hilary S. Parker[1], Andrew E. Jaffe[1,3] and John D. Storey[4]

[1]Department of Biostatistics, JHU Bloomberg School of Public Health, Baltimore, MD, [2]Division of Computational Biomedicine, Boston University, Boston, MA, [3]Department of Epidemiology, JHU Bloomberg School of Public Health, Baltimore, MD and [4]Lewis-Sigler Institute, Department of Molecular Biology, Princeton University, Princeton, NJ, USA

Associate Editor: Janet Kelso

ABSTRACT

Summary: Heterogeneity and latent variables are now widely recognized as major sources of bias and variability in high-throughput experiments. The most well-known source of latent variation in genomic experiments are batch effects—when samples are processed on different days, in different groups or by different people. However, there are also a large number of other variables that may have a major impact on high-throughput measurements. Here we describe the `sva` package for identifying, estimating and removing unwanted sources of variation in high-throughput experiments. The `sva` package supports surrogate variable estimation with the `sva` function, direct adjustment for known batch effects with the `ComBat` function and adjustment for batch and latent variables in prediction problems with the `fsva` function.

Availability: The `R` package `sva` is freely available from http://www.bioconductor.org.

Contact: jleek@jhsph.edu

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

Received on November 7, 2011; revised on January 12, 2012; accepted on January 13, 2012

1 INTRODUCTION

High-throughput data are now commonly used in molecular biology to (i) identify genomic features associated with outcomes and (ii) build signatures for prediction. These goals are complicated by the presence of latent variables or unwanted heterogeneity in the high-throughput data. Batch effects are the most widely recognized potential latent variable in genomic experiments. The impact of batch effects can be severe, potentially completely compromising biological results (Leek *et al.*, 2010). Furthermore, batch effects are not the only potential source of latent variation that may compromise the statistical or biological validity of a study (Leek and Storey, 2007).

Here we introduce the `sva` package for identifying and removing batch effects and other unwanted sources of variation. The `sva` package contains methods for removing artifacts both by: (i) identifying and estimating surrogate variables for unknown sources of variation in high-throughput experiments (Leek and Storey, 2007, 2008) and (ii) directly removing known batch effects using ComBat (Johnson *et al.*, 2007). Removing batch effects and using surrogate variables have been shown to reduce dependence, stabilize error rate estimates and improve reproducibility (Leek and Storey, 2007, 2008; Leek *et al.*, 2010). Finally, the `sva` package includes the only publicly available function, `fsva`, for identifying and removing latent variables in genomic/epigenomic prediction problems.

*To whom correspondence should be addressed.

2 USING THE sva PACKAGE

2.1 Data format

The data are formatted as a matrix, with features (transcripts, genes, proteins) in rows and samples in the columns. Two model matrices must be created with the `model.matrix` function—the 'null model' and the 'full model'. The null model consists of the known variables and covariates that must be included as adjustment variables. The full model includes all the variables in the null model, as well as the variable of interest. The variable of interest is the outcome/phenotype being predicted or associated with the high-throughput data.

2.2 The sva function for estimating and removing surrogate variables

The `sva` function is part of a two step process that first estimates surrogate variables and then removes them in a differential expression analysis. Surrogate variables can be estimated by applying the `sva` function to the high-dimensional data matrix (`dat`), with arguments for the full model matrix (`mod`) and the null model matrix (`mod0`). The output of the `sva` function are the surrogate variables themselves. They can be included in the model matrix and null model matrix and then passed, along with the data matrix, to the `f.pvalue` function in the `sva` package to calculate parametric F-test P-values adjusted for surrogate variables.

2.3 The ComBat function for removing batch effects

The `ComBat` function adjusts for known batches using an empirical Bayesian framework (Johnson *et al.*, 2007). The `ComBat` function is again applied to the high-dimensional data matrix, passing the full model matrix created without any known batch variables.

Batch variables are passed as a separate argument (batch) to the function. The output is a set of corrected measurements, where batch effects have been removed. Standard analysis techniques can be applied to this corrected data, or the sva function can be applied to remove potentially unwanted sources of variation.

2.4 fsva for prediction

For genomic prediction, datasets are generally composed of a training set and a test set. For each sample in the training set, the outcome/class is known, but latent sources of variability are unknown. For the samples in the test set, neither the outcome/class nor the latent sources of variability are known. When applying genomic predictors, individual samples must be corrected. But most functions for batch correction and surrogate variable estimation have been developed in the context of population studies. 'Frozen' surrogate variable analysis can be used to remove latent variation in the training and test sets, as well as individual samples obtained in future studies, similar to the recently developed normalization procedures (McCall *et al.*, 2010).

The arguments that must be passed to fsva are a database of measurements from the training set (dbdat), the model matrix for the training set (mod), the sva object obtained from running sva on the training set and optionally the data from the test set (newdat). The fsva function returns corrected training data (db) and corrected test data (new). If new samples are obtained, they can be adjusted for surrogate variables by including them in the newdat data matrix while leaving all other arguments the same. To illustrate this method, we applied the fsva function to a previously published study of gene expression in bladder cancer. Adjustment with fsva led to increased accuracy and improved clustering of samples in the test set (Supplemental Materials).

3 DISCUSSION

We have introduced the sva package, including the popular ComBat function for removing batch and other unmeasured or unmodeled sources of variation. We have also introduced the first function for removing batch effects in genomic prediction problems. The sva package is freely available from the Bioconductor website and is compatible with widely used differential expression software such as limma (Smyth, 2004).

3.1 Surrogate variables versus direct adjustment

The goal of sva is to remove all unwanted sources of variation while protecting the contrasts due to the primary variables specified in the function call. This leads to the identification of features that are consistently different between groups, removing all common sources of latent variation.

In some cases, latent variables may be important sources of biological variability. If the goal of the analysis is to identify heterogeneity in one or more subgroups, the sva function may not be appropriate. For example, suppose it is expected that cancer samples represent two distinct, but unknown subgroups of biological interest. If these subgroups have a large impact on expression, then one or more of the estimated surrogate variables may be highly correlated with the subgroup (Teschendorff *et al.*, 2011). This is true regardless of whether the surrogate variables are estimated with principal components, singular vectors (Leek and Storey, 2007, 2008) or independent components (Teschendorff *et al.*, 2011). However, removing surrogate variables that are correlated with the phenotype of interest may lead to inconsistent and anti-conservatively biased significance analysis, specially if unknown latent variables are correlated with the phenotype of interest (Leek and Storey, 2007). Thus, whether exclusion of surrogate variables improves inference or not is an open unsolved problem.

In contrast, direct adjustment only removes the effect of known batch variables. Batch effects are the best-known source of latent variation in genomic experiments (Leek *et al.*, 2010). However, there are many variables that may have a substantial impact on genomic measurements, from environmental variables (Gibson, 2008) to genetic variation (Brem *et al.*, 2002; Schadt *et al.*, 2003). These variables may be the focus of the study being performed. But there are many studies that focus on identifying the association between genomic measurements and specific outcomes or phenotypes. In these studies, genetic and environmental variables are often unmeasured or unmodeled. If ignored, these biological variables may act in the same way that batch effects act by obscuring signal, reducing power and biasing biological conclusions (Leek and Storey, 2007).

As a rule of thumb, when there are a large number of known or unknown potential confounders, surrogate variable adjustment may be more appropriate. Alternatively, when one or more biological groups is known to be heterogeneous, and there are known batch variables, direct adjustment may be more appropriate.

ACKNOWLEDGEMENT

We would like to thank Rafa Irizarry and the Feinberg Lab for helpful comments and feedback on the sva package.

Funding: National Institutes of Health grants: (RR021967 and R01 HG002913).

Conflict of Interest: none declared.

REFERENCES

Brem,R.B. *et al*. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Gibson,G. (2008) The environmental contribution to gene expression profiles. *Nat. Rev. Genet.*, **9**, 575–581.

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

Leek,J. and Storey,J. (2007) Capturing heterogeneity in gene expression studies by 'surrogate variable analysis'. *PLoS Genet.* **3**, e161.

Leek,J. and Storey,J. (2008) A general framework for multiple testing dependence. *Proc. Natl Acad. Sci.* **105**, 18718–18723.

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.

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

Schadt,E.E. *et al.* (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**, 297–302.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article3.

Teschendorff,A.E. *et al.* (2011) Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics*, **27**, 1496–1505.