Bioinformatics Advance Access published August 18, 2016

Gene Expression

BatchQC: Interactive software for evaluating sample and batch effects in genomic data

Solaiappan Manimaran[1,2], Heather Marie Selby[3], Kwame Okrah[4], Claire Ruberman[5], Jeffrey T. Leek[5], John Quackenbush[6,7], Benjamin Haibe-Kains[8,9,10], Hector Corrada Bravo[11] and W. Evan Johnson[1,2,3,*]

[1]Department of Biostatistics, Boston University, Boston, MA, [2]Division of Computational Biomedicine, Boston University School of Medicine, Boston, MA, [3]Bioinformatics Program, Boston University, Boston, MA, [4]gRED Oncology Biostatistics, Genentech, South San Francisco, CA, [5]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, [6]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, [7]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, [8]Departments of Medical Biophysics and Computer Science, University of Toronto, [9]Princess Margaret Cancer Centre, University Health Network, [10]Ontario Institute of Cancer Research, Toronto, Ontario, Canada, [11]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD

*To whom correspondence should be addressed.

Associate Editor: Dr. Inanc Birol

Abstract

Sequencing and microarray samples often are collected or processed in multiple batches or at different times. This often produces technical biases that can lead to incorrect results in the downstream analysis. There are several existing batch adjustment tools for '-omics' data, but they do not indicate a priori whether adjustment needs to be conducted or how correction should be applied. We present a software pipeline, BatchQC, which addresses these issues using interactive visualizations and statistics that evaluate the impact of batch effects in a genomic dataset. BatchQC can also apply existing adjustment tools and allow users to evaluate their benefits interactively. We used the BatchQC pipeline on both simulated and real data to demonstrate the effectiveness of this software toolkit.

Availability: BatchQC is available through Bioconductor and: https://github.com/mani2012/BatchQC.

Contact: wej@bu.edu

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

1 Introduction

Non-biological variation, including batch effects, is commonly present across multiple batches of '-omics' data of the same type (e.g., microarray, RNA-seq, DNA methylation, proteomics) and can be caused by technical factors such as differences in profiling platform, lab protocol, experimenter, processing procedures or reagent batch. These data can be affected by technical variation attributable to both observed and unobserved factors (Leek JT *et al.*, 2010). Batch effects can act as a confounder and produce technical biases that lead to incorrect downstream analyses (Akey,J.M. *et al.*, 2007, Lambert,C.G. *et al.*, 2012). Many effective methods have been developed to filter technical heterogeneity and batch effects from high-throughput biological data (Johnson,W.E *et al.*, 2007, Leek JT *et al.*, 2012, Gagnon-Bartsch *et al.*, 2013); however, it is often unclear which method should be applied when combining particular sets of experimental data. In some cases, significant batch correction is needed, whereas in other cases no correction is required. Thus, a complete evaluation of each case is necessary before devising an appropriate correction strategy. This process requires preliminary analyses including data visualization, hierarchical clustering, principle components analysis (PCA), and significance testing. These analyses entail considerable effort, and must be repeated for every set of experimental data. BatchQC streamlines batch evaluation by providing interactive diagnostics, visualizations, and statistical analyses to explore the extent to which batch variation impacts the data. BatchQC diagnostics guide the user in determining whether batch adjustment needs to be conducted, and how it should be applied before downstream analysis. BatchQC interactively applies multiple batch effect approaches to the data, and the user can see the benefits of each method.

© The Author(s) 2016. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

2 Methods

BatchQC is a Shiny App (http://shiny.rstudio.com/) R-package (R ver. 3.3+). The output is organized into multiple tabs, each featuring a part of the batch effect analysis of the data. The package was developed for flexible future development: tabs can be modified, added, or removed as needs change or as new approaches are developed. We highlight many of the interactive features that are provided by BatchQC:

Summary and Sample Diagnostics: BatchQC provides diagnostic and measures for the impacts of study design, batch and treatment effects, and for individual samples. These include tabular summary of the samples from each condition and batch, and estimates of the level of confounding present in the study design between batch and condition. BatchQC also provides summary statistics and figures for the percentage of the variation explained by the batch and condition variables for each gene, and a p-value analysis that tests statistical significance of both the batch and condition effects across genes (summarizing their distributions using boxplots, histograms, and tables). Moreover, BatchQC includes visualizations for between sample correlations to help identify outlying samples and batches.

Visualization and Differential Expression: BatchQC provides an interactive box plot of user-inputted genomic values (read count, probe intensity, etc.) for each sample, with options to sort and color the samples by condition or batch, which enables the user to visualize differences across conditions and batches (Figure 1, top). BatchQC also provides heatmap plots for gene-level values and a sample-level circular dendrogram that clusters the samples using the choice of several different agglomeration measures. BatchQC also provides gene-level significance tests.

PCA and Shape Analysis: BatchQC conducts PCA on the dataset and produces an interactive plot for displaying the user's choice of any two components at the same time, with the points colored by the choice of condition or batch (Figure 1, bottom). BatchQC also provides a summary table that associates the percentage of variation of each PCA component explained by batch and condition, and tests for the statistical significance of these effects. In addition, BatchQC conducts a distributional shape analysis, namely the skewness and kurtosis, to evaluate batch effects in higher moments in the data (Okrah,K *et al.*, 2015).

Batch Adjustment: BatchQC can interactively adjust the data for batch effects using ComBat or Surrogate Variable Analysis (SVA). BatchQC determines the number of surrogate variables to identify in the given data set, estimates the surrogate variables, and performs the batch adjustment. After batch adjustment, all of the previous diagnostics can be viewed for the raw or adjusted data, and the user can interactively toggle between the two and evaluate the impact of batch adjustment.

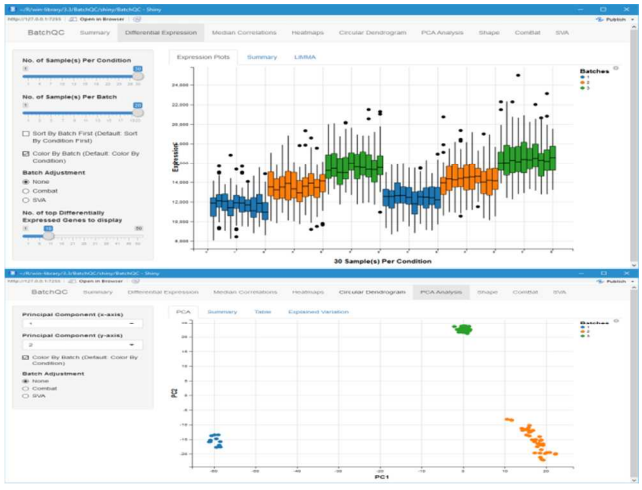

Figure 1: Examples from the BatchQC interface. (top) Boxplots from the simulated dataset showing clear distributional differences between batches. (bottom) The first two PCA components from the signature dataset shows strong batch effects.

3 Data Examples and Comparisons

The supplementary materials and package vignettes illustrate the functionalities of BatchQC using multiple simulated and real datasets that are included with the package. The need for batch adjustment in the first simulated dataset (details in Supplement) can be noted in the boxplots (Figure 1, top), variation analysis, PCA and the clustered heatmap. After adjustment, the boxplots and variation analysis showed no remaining evidence of batch effects, and the samples clustered strongly by condition and not by batch as was previously the case in heatmap and PCA. Interestingly, in the second simulated dataset, PCA did not reveal the batch differences, while the boxplot and heatmaps did show batch differences, indicating the need for multiple diagnostic measures in each data scenario.

As a real-data example, we applied BatchQC to sequencing data captured from human mammary epithelial cells after activating key growth pathway genes (BatchQC examples; GEO accession GSE73628). The data consists of three batches and ten different conditions corresponding to control and activation of nine different pathways (details in Supplement). The PCA plots show a batch effect (Figure 1, bottom). Similarly, the circular dendrogram and median correlation plots reveal clustering based on batch status. Batch adjustment using either ComBat or SVA diffuses the batch clusters in the PCA and circular dendrogram. These examples demonstrate the utility of all features of the BatchQC package.

4 Conclusions

BatchQC is beneficial for the analyses of '-omic' data and is particularly advantageous in studies where samples are collected in multiple batches or at different times. BatchQC streamlines batch preprocessing and evaluation by providing interactive diagnostics, visualizations, and statistical analyses to explore whether batch adjustment needs to be conducted and how correction should be applied (see Supplementary Materials for recommendations on when batch adjustment should be applied). Moreover, BatchQC interactively applies multiple batch effect approaches to the data, and the user can quickly see the benefits of each method interactively. BatchQC is the first software tool to integrate batch diagnostics and correction methods. BatchQC is available as a Shiny user interface that makes the application easy to use and can be easily installed on any computer with standard R installation (R ver. 3.3+) and pandoc (1.12.0+).

Acknowledgements

This research was supported by funds from the NIH (R01 HG005692, R01 ES025002), the NSF (DGE 0654108). Conflict of Interest: none.

References

Akey,J.M., Biswas,S., Leek,J.T. and Storey,J.D. (2007) On the design and analysis of gene expression studies in human populations. Nat. Genet., 39, 807–808.

Gagnon-Bartsch J.A, Jacob L., Speed T.P., Removing Unwanted Variation from High Dimensional Data with Negative Controls, 2013

Johnson,W.E., Li,C. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods, Biostatistics (2007), 8, 1, pp. 118–127.

Lambert,C.G. and Black,L.J. (2012) Learning from our GWAS mistakes: from experimental design to scientific method. Biostatistics, 13, 195–203.

Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012 Mar 15;28(6):882-3.

Leek J.T, Scharpf R.B, Bravo H.C, Simcha D., Langmead B., Johnson W.E., Geman D., Baggerly K., and Irizarry R.A., Tackling the widespread and critical impact of batch effects in high-throughput data, Nature Reviews, Genetics, Volume 11, October 2010, p733-739

Okrah,K., Bravo,H.C, (2015) Shape analysis of high-throughput transcriptomics experiment data, Biostatistics (2015), pp. 1–14

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016