Removal of unwanted variation in differential expression analysis of single-cell transcriptome sequencing data

Sofia Prieto Leon[1] *sofia.prietoleon@uhasselt.be*
Koen Van den Berge [2] *KVande14@its.jnj.com*
Ewoud De Troyer [2] *edetroye@its.jnj.com*
Olivier Thas [1,3,4] *olivier.thas@uhasselt.be*
[1]Data Science Institute and I-BioStat, Hasselt University, Belgium
[2] Statistics and Decision Sciences, Janssen Research & Development, Beerse, Belgium
[3] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium
[4] National Institute for Applied Statistics Research Australia (NIASRA),
University of Wollongong, New South Wales, Australia

Abstract

Analyzing high-dimensional biological data poses significant challenges, particularly in the presence of unwanted variation introduced by external factors that can obscure biological signals. In this work, we delve into the application of Remove Unwanted Variation (RUV) methods to address these challenges, focusing on pseudo-bulk gene expression data derived from single-cell RNA sequencing (scRNA-seq) experiments.

Initially, we review current RUV methodologies for bulk analysis, drawing insights from work by Risso et al. (2014), Gagnon-Bartsch and Speed (2012), Gagnon-Bartsch et al. (2013), and Molania et al. (2019). These references provide a solid foundation for understanding the principles and practical implementations of RUV, ranging from microarray to modern RNA-seq and nanostring technologies.

Subsequently, we explore the application of RUV methods to pseudo-bulk data, where gene expression profiles are aggregated from scRNA-seq data at a cell type-subject level. By leveraging insights from bulk RNA-sequencing analyses, we compare three strategies for integrating RUV techniques into pseudo-bulk differential expression analysis, thereby improving downstream analyses' robustness and biological interpretability.

Finally, an illustrative example utilizing single-cell RNA-seq data from Perez et al. (2022) study on lupus provides a practical demonstration of the discussed methodologies. With this work, we intend to showcase the utility of RUV methods in uncovering cell type-specific gene expression associations in complex diseases, clarify the strengths and limitations of each method, and highlight scenarios where a particular approach outperforms others.

References

Gagnon-Bartsch, J. A., Jacob, L., and Speed, T. P. (2013). Removing unwanted variation from high dimensional data with negative controls. *Berkeley: Tech Reports from Dep Stat Univ California*, pages 1–112.

Gagnon-Bartsch, J. A. and Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552.

Molania, R., Gagnon-Bartsch, J. A., Dobrovic, A., and Speed, T. P. (2019). A new normalization for nanostring ncounter gene expression data. *Nucleic Acids Research*, 47(12):6073–6083.

Perez, R. K., Gordon, M. G., Subramaniam, M., Kim, M. C., Hartoularos, G. C., Targ, S., Sun, Y., Ogorodnikov, A., Bueno, R., Lu, A., et al. (2022). Single-cell rna-seq reveals cell type–specific molecular and genetic associations to lupus. *Science*, 376(6589):eabf1970.

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology*, 32(9):896–902.