Batch effects in omics data

Session Intro

The session looks at a major issue that underlies many omics datasets, viz. batch effects. Batch effects are technical biases that may confound analysis of omics data. They are very complex and effective mitigation is highly context dependent. Do they affect identification of discriminating/causal factors when we analyze patient datasets? Do prediction models (constructed on training datasets) work well on future patients? How do you mitigate batch effects?

Session Plan

Part I, What batch effects are and how they affect biomedical data analysis and model building.

Suggested readings:

- Leek et al., "Tackling the widespread and critical impact of batch effects in high-throughput data", *Nat. Rev. Genet.*, 11(10):733-739, 2010

Presentation team #3: CHENG YI, MD SALMAN SHAMIL, WANG JINGTAN

Part II, How batch effects can be measured. How do you know they are big enough to worry over?

Suggested readings:

- kBET (Buttner et al., "A test metric for assessing single-cell RNA-seq batch corrections", *Nat. Methods*, 16:43-49, 2019)
- LISI (Korsunsky et al., "Fast, sensitive, and accurate integration of single-cell data with Harmony", *Nat. Methods*, 16:1289-1296, 2019)
- gPCA (Reese et al., "A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis", Bioinformatics, 29(22):2877-2883, 2013)
- PCA side-by-side boxplot (Goh & Wong, "Protein complex-based analysis is resistant to the obfuscating consequences of batch effects – a case study in clinical proteomics", *BMC Genomics*, 18:142, 2017)

Note: It is not necessary to present the papers above. Just focus on the parts about batch effect measurements.

Presentation team #8: HE YINGZHI, LI XIANG, ZHAO ZITONG

Part III, Normalization methods and batch effect-correction methods. What are these and what are their important differences?

Suggested readings:

- Common normalization methods such as linear scaling, quantile normalization, z-score transformation, and specialized methods such as GFS (Belorkar & Wong, "GFS: Fuzzy preprocessing for effective gene expression analysis", *BMC Bioinformatics*, 17(Suppl 17):540, 2016)
- Some popular batch effect-correction methods are ComBat (Johnson et al., "Adjusting batch effects in microarray expression data using empirical Bayes methods", *Biostatistics*, 8:118-127, 2007), Harman (Otyam et al., "Risk-conscious correction of batch effects: Maximising information extraction from high-throughput genomics datasets", *BMC Bioinformatics*, 17:332, 2016), SVA (Leek & Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis", *PLoS Genet*, 3:1724-1735, 2007), and Batch mean centering (Sim et al., "The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets – improving meta-analysis and prediction of prognosis", *BMC Med. Genomics*, 1:42, 2008)

Note: No need to present all papers in details. Just focus on key concepts.

Presentation team #6: DIBYADIP CHATTERJEE, GAO TIANYU, LIU NIAN

Part IV, How should a normalization method be applied when there are multiple classes and batches?

Suggested readings:

- Zhao et al., "How to do quantile normalization correctly for gene expression data analysis", *Scientific Reports*, 10:15534, 2020

Project: The paper above provides some ideas/scenarios on how to apply quantile normalization. The results can be sharpened. The discussion can also be extended to cover other normalization approaches, e.g. z-score transformation. The presentation team should identify some other studies (or even make your own study) and present/discuss these as well.

Presentation team #9: TANG KAIWEN, NGUYEN HA LINH

Part V, How do normalization methods interact with batch effects and batch effect-correction methods

Suggested readings:

- Zhou et al., “Examining the practical limits of batch effect-correction algorithms: When should you care about batch effects?”, *J Genet. Genomics*, 46:433-443, 2019.

Project: This paper compares normalization methods and batch effect-correction methods. It mainly considered applying normalization methods to whole datasets and before batch-effect correction. What if normalization was applied in a class-specific manner? What if batch-effect correction was done before normalization? The presentation team should identify some other studies (or even make your own study) and present/discuss these as well.

Presentation team #7: FU GUOJI, LUO YANG, PRABOWO DJONATAN

Part VI, If a dataset has lots of missing values and also batch effects, what happens and what can/should you do?

Suggested readings:

- Some missing value-imputation methods (imputation by global mean, same-batch mean, nearest neighbours, etc.)
- Voss et al., “HarmonizR enables data harmonization across independent proteomic datasets with appropriate handling of missing values”, *Nat. Comm.,* 13: 3523, 2022
- Sun & Goh, “Why batch sensitization is important for missing value imputation”, https://doi.org/10.21203/rs.3.rs-1328989/v1

Key questions: Are these imputation methods sensitive to batch effects? Do missing value imputation, normalization and batch-effect correction confound each other?

Presentation team #4: DUAN KEYU, LU XINYANG, STEFAN PUTRA LIONAR

Project for the class

Can you think of a better way than the methods in Part II for measuring and quantifying batch effects?

No need a fully worked out method. Just provide an outline of your idea, along with an explanation of why you think it will work. Best keep to 1-2 pages, though it is fine if you have done a lot of work and want to show more.

Submission date: 31 October 2023