Batch effect correction methods for NASA GeneLab transcriptomic datasets

Lauren M. Sanders[1,2], Hamed Chok[3], Finsam Samson[4], Ana Uriarte Acuna[2,5], San-huei Lai Polo[2,5], Valery Boyko[2,6], Yi-Chun Chen[2,5], Marie Dinh[2,7], Samrawit Gebre[2], Jonathan M. Galazka[2], Sylvain V. Costes[2], **Amanda M. Saravia-Butler**[2,5]

[1]Blue Marble Space Institute of Science, NASA Ames, [2]NASA Ames Research Center (ARC); Moffett Field, CA, 94035, USA, [3]GeneLab Multi-Omics Analysis Working Group, [4]Stanford University; Stanford, CA, 94305, USA, [5]KBR, Space Biosciences Division, NASA ARC, Moffett Field, CA 94035, USA, [6]The Bionetics Corporation, NASA ARC, Moffett Field, CA 94035, USA, [7]Logyx, LLC, Mountain View, CA 94043, USA

Abstract

RNA sequencing (RNAseq) data from space biology experiments promise to yield invaluable insights into the effects of spaceflight on terrestrial biology. However, sample numbers from each study are low due to limited crew availability, hardware, and space. To increase statistical power, spaceflight RNAseq datasets from different missions are often aggregated together. However, this can introduce technical variation or "batch effects", often due to differences in sample handling, sample processing, and sequencing platforms. Several computational methods have been developed to correct for technical batch effects, thereby reducing their impact on true biological signals.

In this study, we combined 7 mouse liver RNAseq datasets from NASA GeneLab (part of the NASA Open Science Data Repository) to evaluate several common batch effect correction methods (ComBat and ComBat-seq from the *sva* R package, and Median Polish, Empirical Bayes, and ANOVA from the *MBatch* R package). We quantitatively evaluated the ability of these methods to correct for technical batch variables in space biology RNA-seq data using the following criteria: BatchQC, principal component analysis, dispersion separability criterion, log fold change correlation, and differential gene expression analysis. Each batch variable / correction method combination was then assessed using a custom scoring approach to identify the optimal correction method for the combined dataset, by geometrically probing the space of all allowable scoring functions to yield an aggregate volume-based scoring measure.

Finally, we describe the way in which the GeneLab multi-study analysis and visualization portal will allow users to examine the presence or absence of batch effects using multiple metrics. If the user chooses to perform batch effect correction, the scoring approach described here can be implemented to identify the optimal correction method to use for their specific combined dataset prior to analysis.

Background

OSDR Mouse Liver RNAseq Datasets:

- OSD-47, OSD-48, OSD-137, OSD-168, OSD-173, OSD-242, OSD-245
- FLT and GC samples from each dataset were combined and evaluated

Two Primary Sources of Technical Variation Identified:

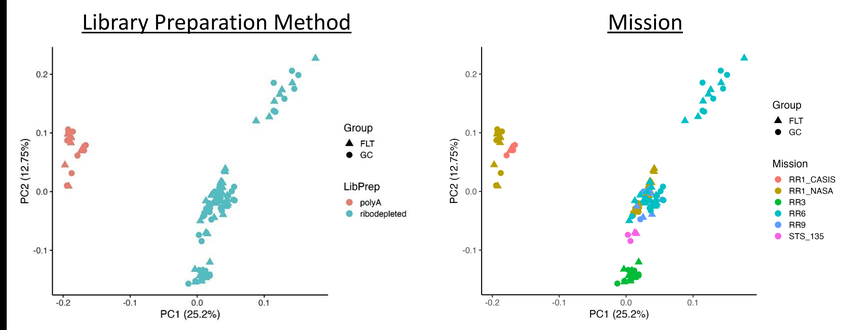

Figure 1. Principal Component Analysis (PCA) of the combined dataset. PCA plots with shapes representing either spaceflown (FLT) or ground control (GC) groups and samples colored by library preparation (LibPrep) method (*left*) or by mission (*right*).

Methods

Batch Effect Correction Methods Evaluated:

Tool	Algorithm	Sources of Batch Effect Corrected
ComBat	Empirical Bayes	Library Preparation Mission
ComBat-seq	Negative Binomial	
MBatch	Empirical Bayes	
	ANOVA	
	Median Polish	

Evaluation Criteria:

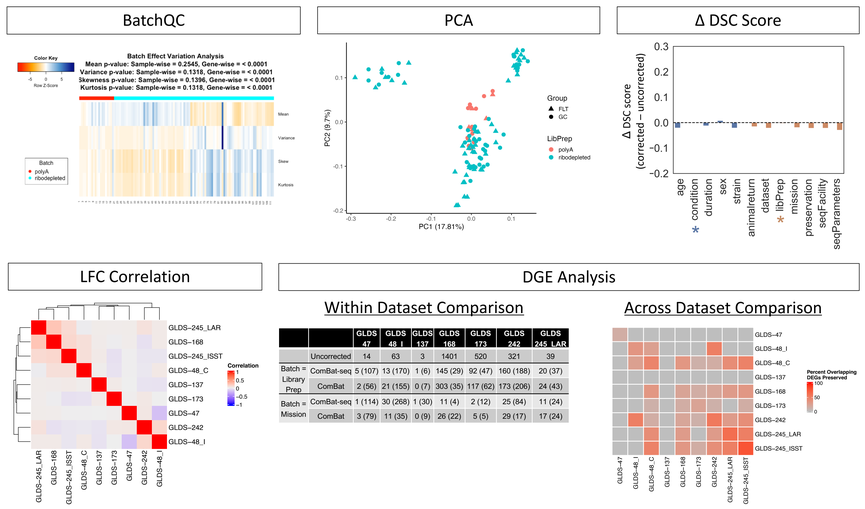

Figure 2: Evaluation criteria used to score each batch variable / correction method pair. Each evaluation criteria was measured pre- and post-correction then compared to generate an improvement score. *BatchQC* was used to measure skew and kurtosis. *PCA* was used to evaluate sample clustering. *Delta dispersion separability score (DSC)* was used to calculate the ratio of dispersion between groups versus within groups, samples were grouped based on technical (orange) or biological (blue) variables. *Log(2) fold change (LFC)* of gene expression in spaceflight versus ground control samples were used to compare how gene expression changes correlate across different datasets. *Differential gene expression (DGE) analysis* was used to determine the DEGs (adj. $p < 0.05$ and $|LFC| > 1$) in each dataset, then the DEGs were compared within each dataset and across pairwise datasets.

Custom Scoring Categorization Scheme:

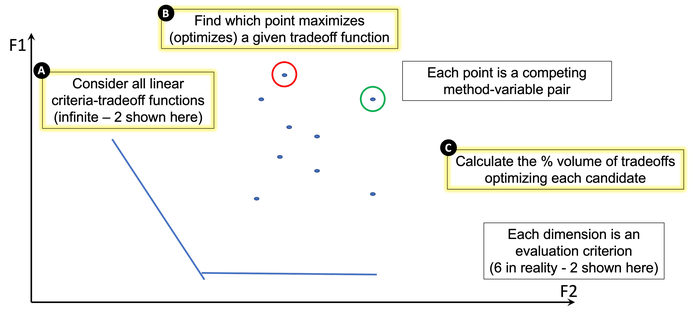

Figure 3: A geometry-based categorization scheme for ranking candidates against performance measures. Dimensions F1 and F2 each represent one of the 6 evaluation criteria. Each point represents a competing batch variable / correction method pair. Each pair's performance is dependent on the criterion. The percent volume of criteria tradeoffs optimized by each candidate pair is calculated geometrically and used to rank all candidate pairs; the underlying geometric approach also yields a quantification of the contribution from each evaluation criterion.

Results

Method	Correction variable	% Volume Assigned
ComBat	Library preparation	34.69%
ComBat-seq	Library preparation	19.31%
ComBat	Mission	18.58%
MBatch Median Polish	Mission	13.06%
MBatch Empirical Bayes	Library preparation	8.17%
ComBat-seq	Mission	3.41%
MBatch ANOVA	Library preparation	1.95%
MBatch Empirical Bayes	Mission	0.79%
MBatch Median Polish	Library preparation	0.00%
MBatch ANOVA	Mission	0.00%

Figure 4. Scoring categorization scheme results for all batch variable / correction method pairs. The table reports the final ranking of the batch variable / correction method pairs based on the percent volume assigned to each after applying the scoring categorization scheme.

Future Work

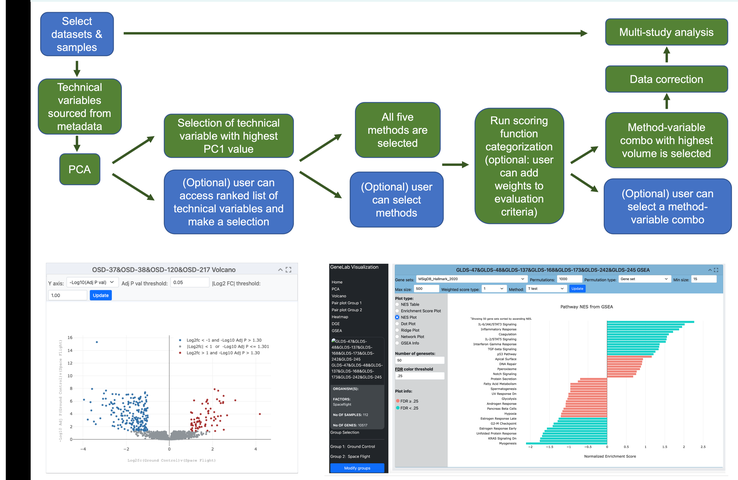

Figure 5: Overview of the GeneLab user portal for multi-study data analysis with batch effect correction and subsequent visualization. *Top:* Flow diagram of multi-study analysis. Blue indicates user choices, while green indicates automated actions. *Bottom:* Example of data visualizations (*left:* volcano plot, *right:* pathway normalized enrichment score (NES) plot from gene set enrichment analysis (GSEA)) using multi-study analysis data. The batch effect correction pipeline will be publicly available to users in GeneLab's upcoming multi-study analysis and visualization portal.

Acknowledgements

We thank the GeneLab Analysis Working Group members for their suggestions and feedback; all NASA GeneLab members for generating, hosting, and maintaining the datasets used in this study.

Funding: This work was funded by the NASA Space Biology Program within the NASA Science Mission Directorate's (SMD) Biological and Physical Sciences (BPS) Division.

Publication: Sanders LM, Chok H, Samson F, Acuna AU, Lai Polo S-H, Boyko V, Chen Y-C, Dinh M, Gebre S, Galazka JM, Costes SV, and Saravia-Butler AM (2023). Batch effect correction methods for NASA GeneLab transcriptomic datasets. *Frontiers in Astronomy and Space Sciences*. doi.org/10.3389/fspas.2023.1200132.