

Evaluation of Correction Methods for NASA GeneLab Transcriptomic Datasets

Finsam Samson, Amanda Saravia-Butler

Space Biosciences Division, Ames Research Center, NASA

Abstract (long)

Conducting space biology experiments aboard the International Space Station (ISS), particularly those utilizing complex model organisms such as mice, is expensive and difficult due to limited crew availability, hardware, and space. As a result, sample numbers from these studies are low, reducing the statistical power of any one experiment. Aggregating spaceflight datasets, such as those hosted on the NASA GeneLab Data Repository, serves as a method to increase sample numbers, allowing for novel insights through bioinformatic analysis of 'omics data from merged datasets. However, aggregating datasets can introduce unwanted variation. Sources of variation may include 1) differences in sample handling, processing, and sequencing platforms between datasets (technical variation) as well as 2) differences in experimental design between datasets such as sex or age of the model organism at the time of flight. There are several tools available, which utilize different statistical methods, to correct for unwanted variation when combining different batches (or sets) of data. To determine the most robust batch correction tool to use when combining spaceflight data from different studies, GeneLab-hosted RNAseq datasets from rodent liver tissue were used to evaluate the following correction methods: ComBat and ComBat_seq from the SVA package, median polish, empirical Bayes, and ANOVA-based algorithms from the MBatch package, and negative binomial regression normalization in the DESeq2 package. These methods were evaluated using two approaches. The first (reference-based) approach, applied correction factors calculated from Universal Mouse RNA Reference (UMRR) samples, to the spaceflight and ground control samples from respective datasets in an attempt to correct for technical variation only. For the second (standard) approach, each correction method was applied to all spaceflight and ground control samples from select datasets. For each approach, after the correction was applied, differential gene expression (DGE) analysis of flight and ground control samples was performed with the combined data. The robustness of each tool was evaluated using BatchQC, to determine statistical differences between datasets before and after correction, Principal Component Analysis, to evaluate global gene expression in samples before and after correction, and by comparing DGE analysis of individual datasets and combined datasets before and after correction. The results showed that the reference-based approach introduced several additional (and likely artificial) differentially expressed genes when compared with the standard approach. As such, the standard, unmodified correction tools/algorithms were utilized. The most robust standard method(s) identified in this evaluation will be implemented in the GeneLab Visualization 2.0 platform.

Abstract (300-word max for ASGSR)

Conducting space biology experiments aboard the International Space Station, particularly those utilizing complex model organisms like mice, is expensive and difficult due to limited crew availability, hardware, and space. As a result, sample numbers from these studies are low, reducing the statistical power of any one experiment. Aggregating spaceflight datasets serves as a method to increase sample numbers, allowing for novel insights through bioinformatic analysis of 'omics data from merged datasets. However, aggregating datasets can introduce unwanted variation including 1) differences in sample handling, processing, and sequencing platforms between datasets (technical variation) as well as 2) differences in experimental design between datasets such as sex or age of the model organism used. In the present study, NASA GeneLab-hosted RNAseq datasets from rodent liver tissues were used to evaluate several statistical methods to correct for this unwanted variation through two approaches, reference-based and standard. The following correction algorithms were applied with (reference-based) and/or without (standard) considering Universal Mouse RNA Reference samples: ComBat and ComBat_seq from the SVA package, median polish, empirical Bayes, and ANOVA-based algorithms from the MBatch package, and negative binomial regression normalization in the DESeq2 package. For each approach, after the correction algorithm was applied, differential gene expression (DGE) analysis of flight and ground control samples was performed with the combined data. The robustness of each tool was evaluated using BatchQC, to determine statistical differences between datasets before and after correction, Principal Component Analysis, to evaluate global gene expression in samples before and after correction, and by comparing DGE analysis of individual datasets and combined datasets before and after correction. The results showed that the reference-based approach introduced several additional (and likely artificial) DEGs when compared with the standard approach. Thus, the most robust standard correction will be implemented in the GeneLab Visualization 2.0 platform when datasets are combined.