Normalization Method Comparison

Nate Olson

2017-12-14

Normalization methods compared ; 1. TMM - weighted trim mean of M-values (EdgeR),
1. RLE - relative log expression (EdgeR),
1. CSS - cummulative sum scaling (metagenomeSeq),
1. TSS - total sum scaling, proportions,
1. UQ - upper quartile scaling.

Globally the observed and expected relative abundance values are consistent across individual and normalization methods excluding the RLE normalization method and E01JH0038 (Fig. 1 and 2). RLE assumes that feature relative abundance is constant across samples, the individual specific effect might be due to variation in how similar samples from the individuals are to the dataset as a whole. Deviation from the expected value for E01JH0038 varied by normalization method with TSS normalized counts being the most consistent with the expected values and UQ the least consistent. The negative binomial relative abundance estimates were used to inferr the theta used to calculate the expected relative abundance values. The negative binomial relative abundance estimates were most consistent with the TSS normalized counts potentially biasing the TSS. For TMM and UQ the E01JH0038 was an outlier relative to the other indivduals.

The agreement between log fold-change estimates and the expected values varied by both individual and normalization method (Fig. 3 and 4). The log fold-change values were calculated using normalized counts averaged across PCR replicates. RLE normalized counts were excluded from the log fold-change error analysis as only 3 features had log fold-change estimates between pre- and post-exposure samples greater than the threshold (> 5) used to define pre-specific and pre-dominant features. The slope of the linear model fit to the observed and expected log fold-change estimates for the RLE and TMM normalization method varied between individuals with negative slopes for E01JH0004, E01JH0016, and E01JH0017 (Fig. 3). When EdgeR was used to calculate the log fold-change estimates for the pipeline comparison, the default normalization method was used and the linear model slope was positive for all individuals.

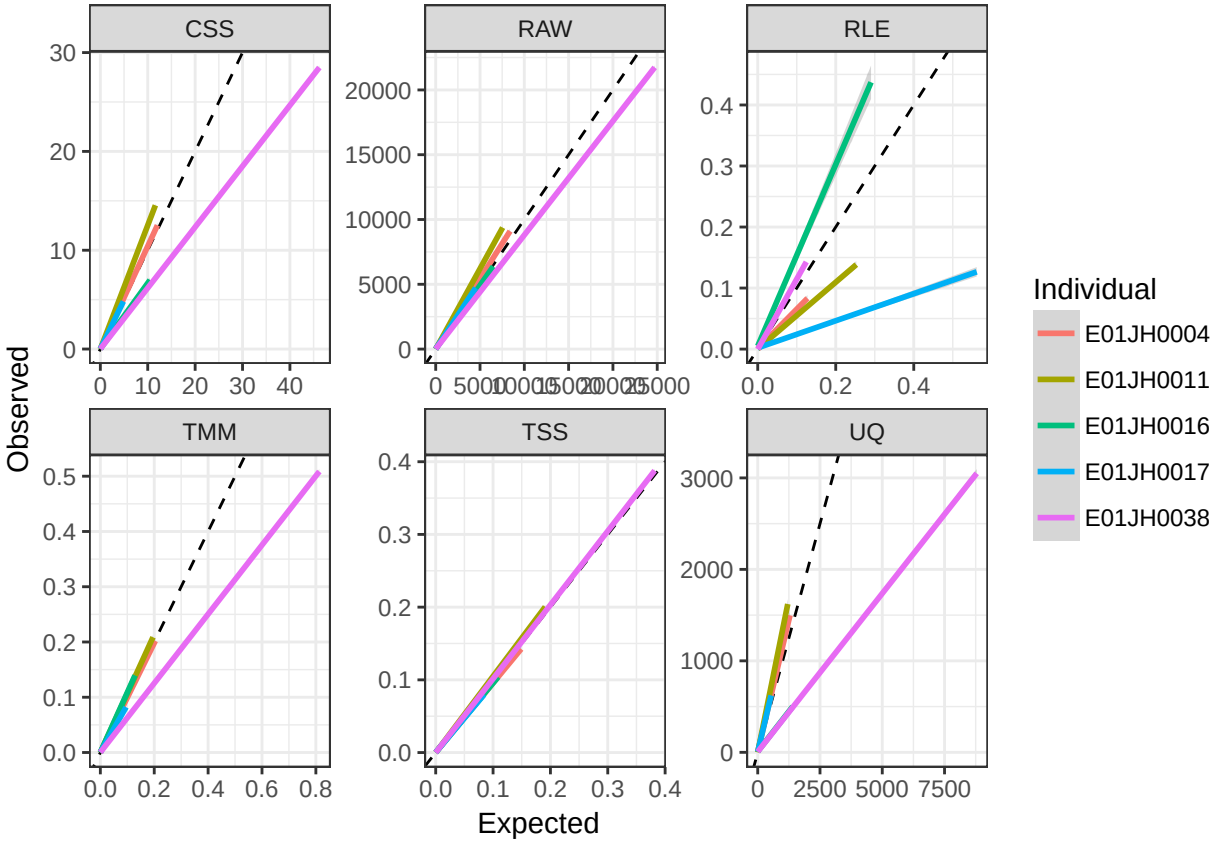

Figure 1: Linear model fit to normalized relative abundance counts averaged across PCR replicates and the expected value by normaization method and individual. The dashed line indicates the expected 1-to-1 relationship.

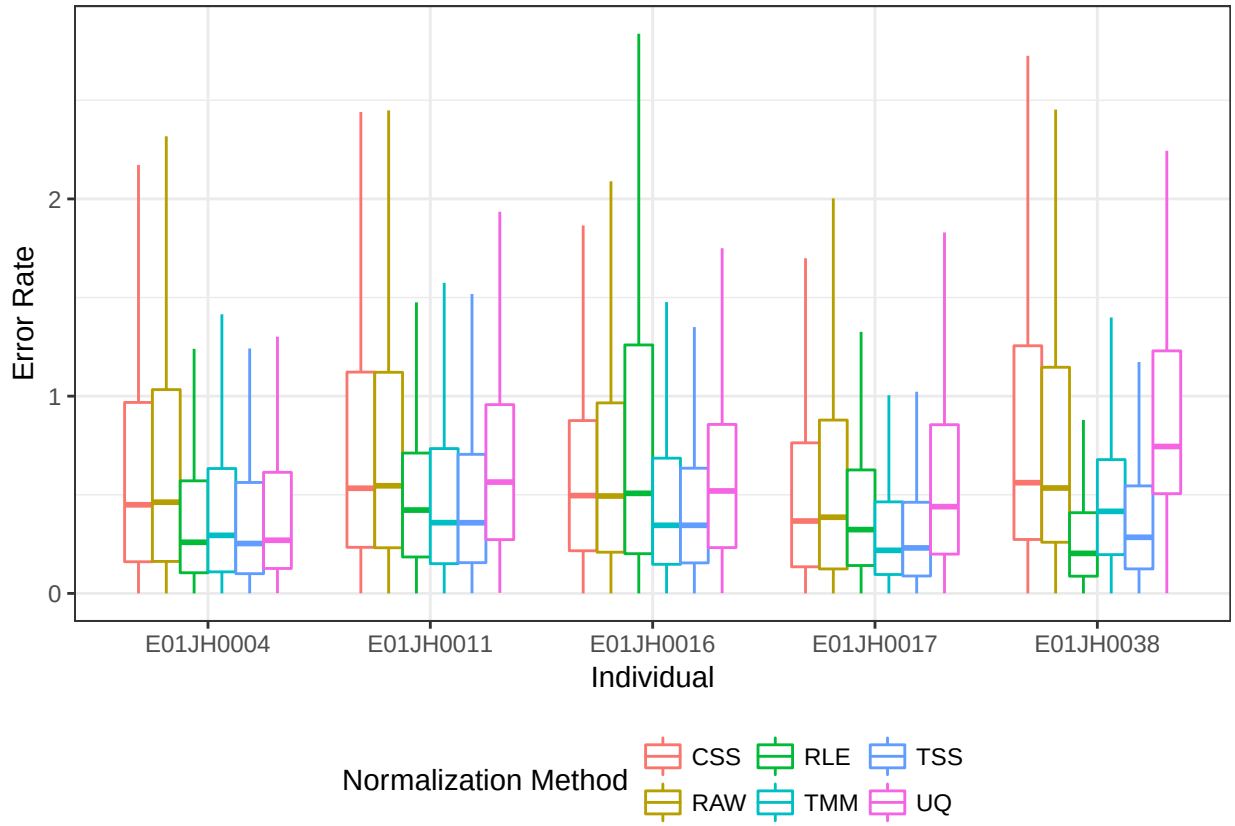

Figure 2: Error rate distribution represented as boxplots across normalization methods by individual. Boxplot outliers were excluded.

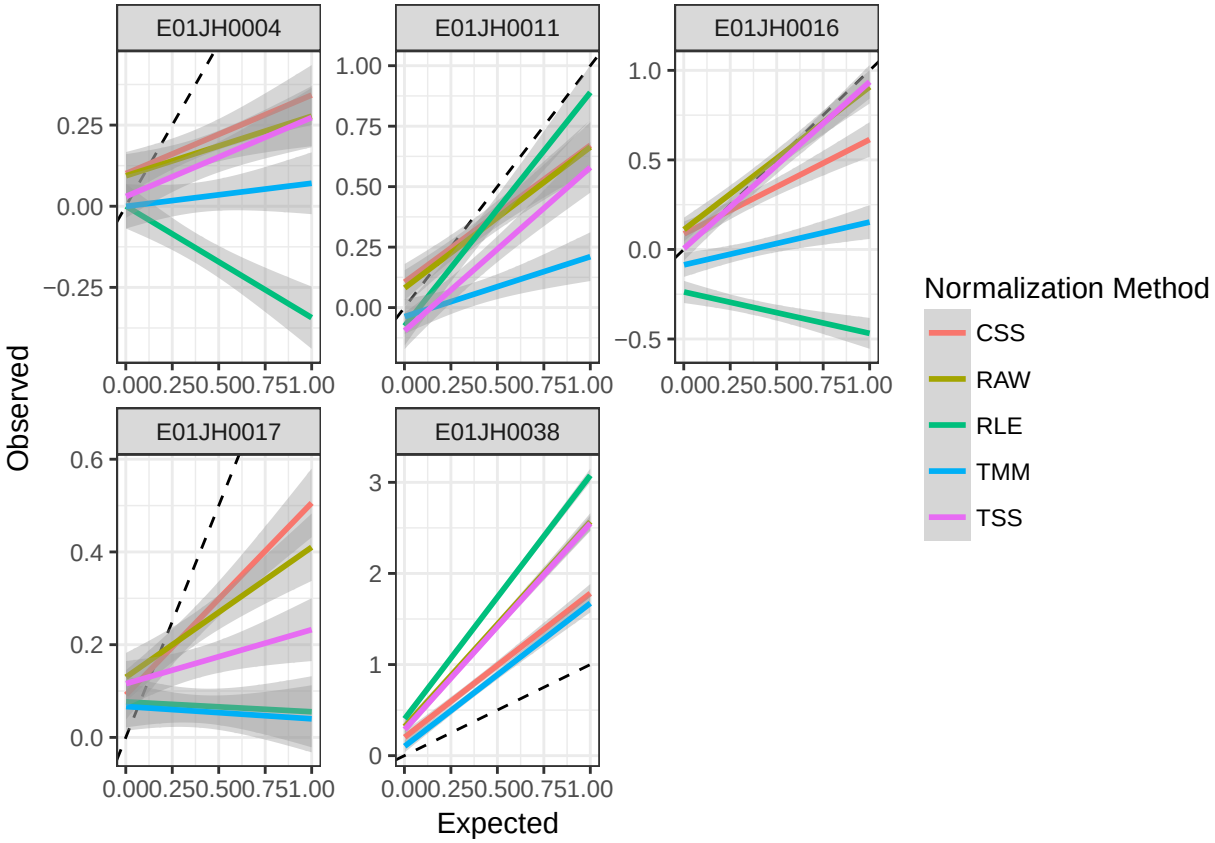

Figure 3: Relationship between observed and expected log fold-change estimates by individual and normalization method.

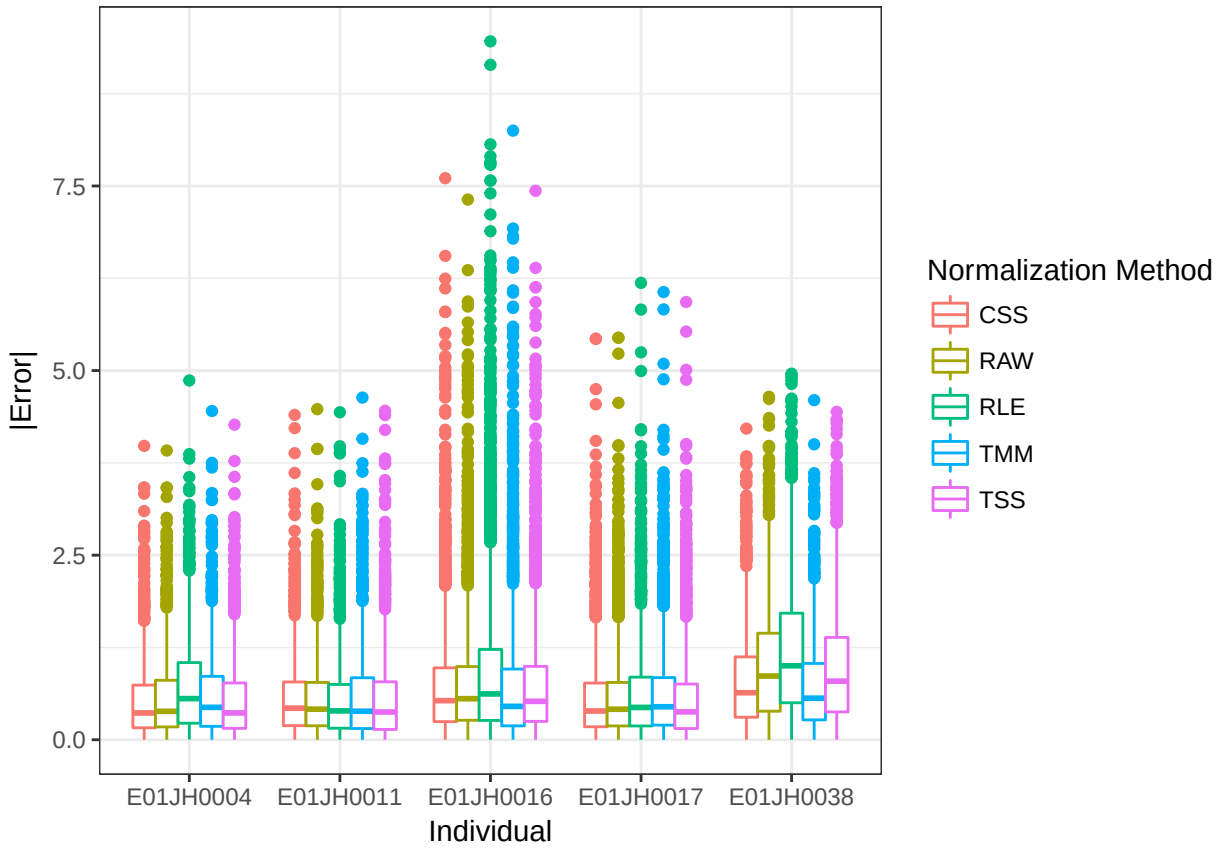

Figure 4: Comparison of log fold-change absolute error distribution across normalization methods by individual.