

logFC Assessment

Nate Olson

2017-10-24

NOTE Need reasoning for excluding slope > 4

Summary statement about the overall agreement between log fold-change estimates and expected values.

The slope and R^2 values for linear models of the estimated and expected log fold-change for individual features, all titration comparison, were used to characterize the log fold-change bias and variance across pipelines. Features with with slope > 4 were excluded from the log fold-change bias and variance analysis. A error metric of $1 - slope$ was used, where 0 is the desired value (i.e. log fold-change estimate = log fold-change expected), negative and positive values indicate the log-fold change was consistently under and over estimated. The linear model R^2 value was used to characterize the feature-level log fold-change variance as it indicates how consistent the relationship between the log fold-change estimates and expected values are across titration comparisons. Similar to the relative abundance assessment we used a mixed-effects models to take into account differences in individuals when comparing the log fold-change error rates between pipeline. The log fold-change bias metric was not significantly different between pipelines (F = 4.75, 0.43, p = 0.03, 0.65, DADA2 0.55, Mothur 0.44, QIIME 0.55, 1B). The log fold-change variance metric was not significantly different between pipelines (F = 77.88, 0.02, p = 0, 0.98, DADA2 0.39, Mothur 0.39, QIIME 0.4 1C).

TODO Incorporate analysis into text

While some of the poor agreement can be attributed to low abundance (noisy data), or smaller differences in the log fold-change between pre- and post-exposure samples. An additional mixed-effects model was used to determine feature characteristics that are correlated with logFC error rate. Increased estimated logFC and logCPM were significantly related to lower error rates. Analysis of the log fold-change error results indicated a feature specific performance effect.

Table 1: Number of pre-specific and pre-dominant features by individual and pipeline

Individual	Type	dada2	mothur	qiime	unclustered
E01JH0004	dominant	7	11	8	14
E01JH0004	specific	47	11	10	32
E01JH0011	dominant	3	7	6	11
E01JH0011	specific	38	14	11	24
E01JH0016	dominant	4	5	0	7
E01JH0016	specific	84	44	16	65

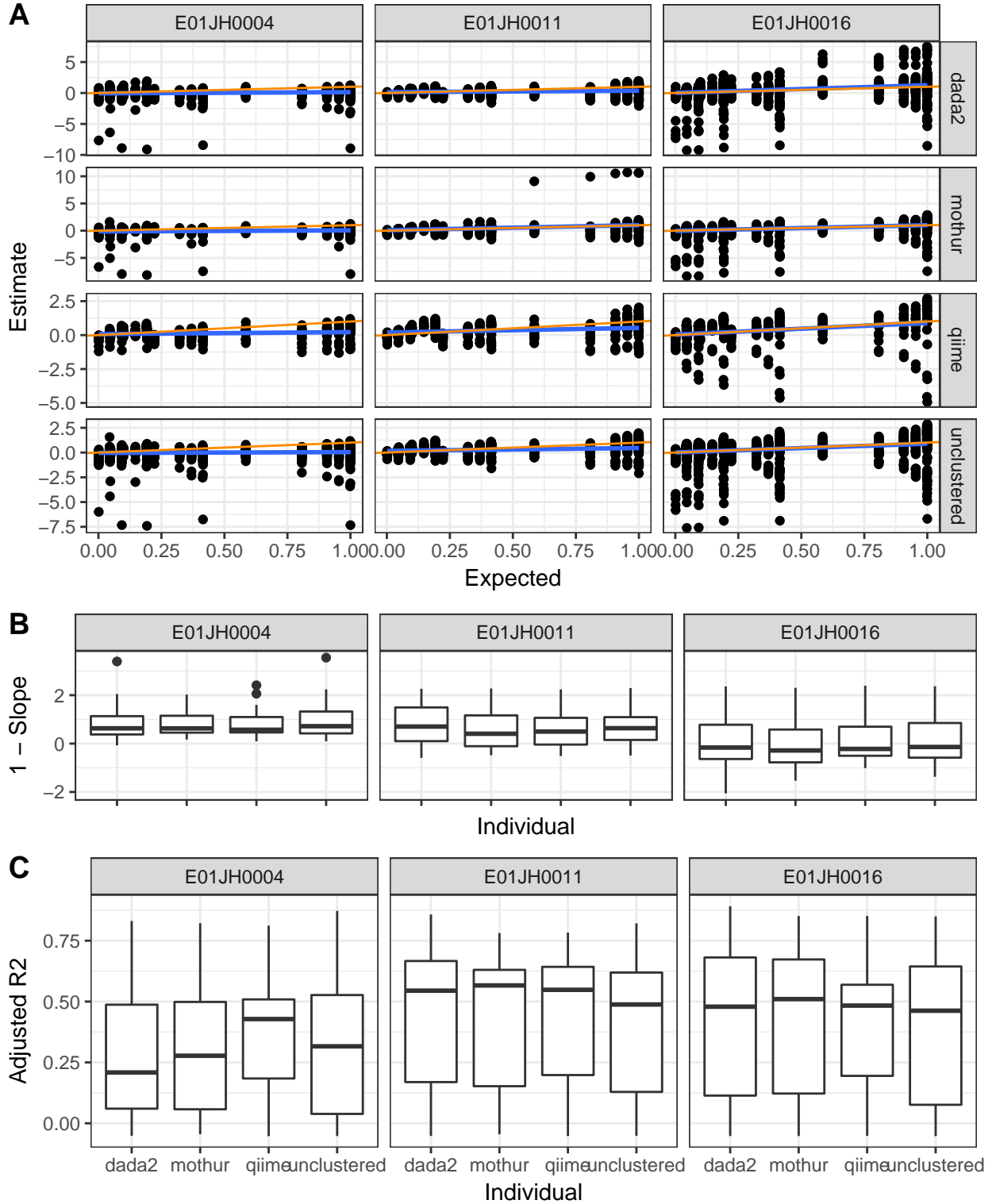


Figure 1: Relationship between the observed and expected logFC for pre-specific and pre-dominant features by pipeline and individual for all titration pair comparisons. Orange line indicates expected 1-to-1 relationship between the estimated and expected logFC. Blue line is a linear model was fit to the data and grey area is the models uncertainty estimate. Distribution of feature-level log-fold change error by individual and pipeline as the (B) slope and (C) R2 for a linear model fit to the estimated and expected log fold-change values