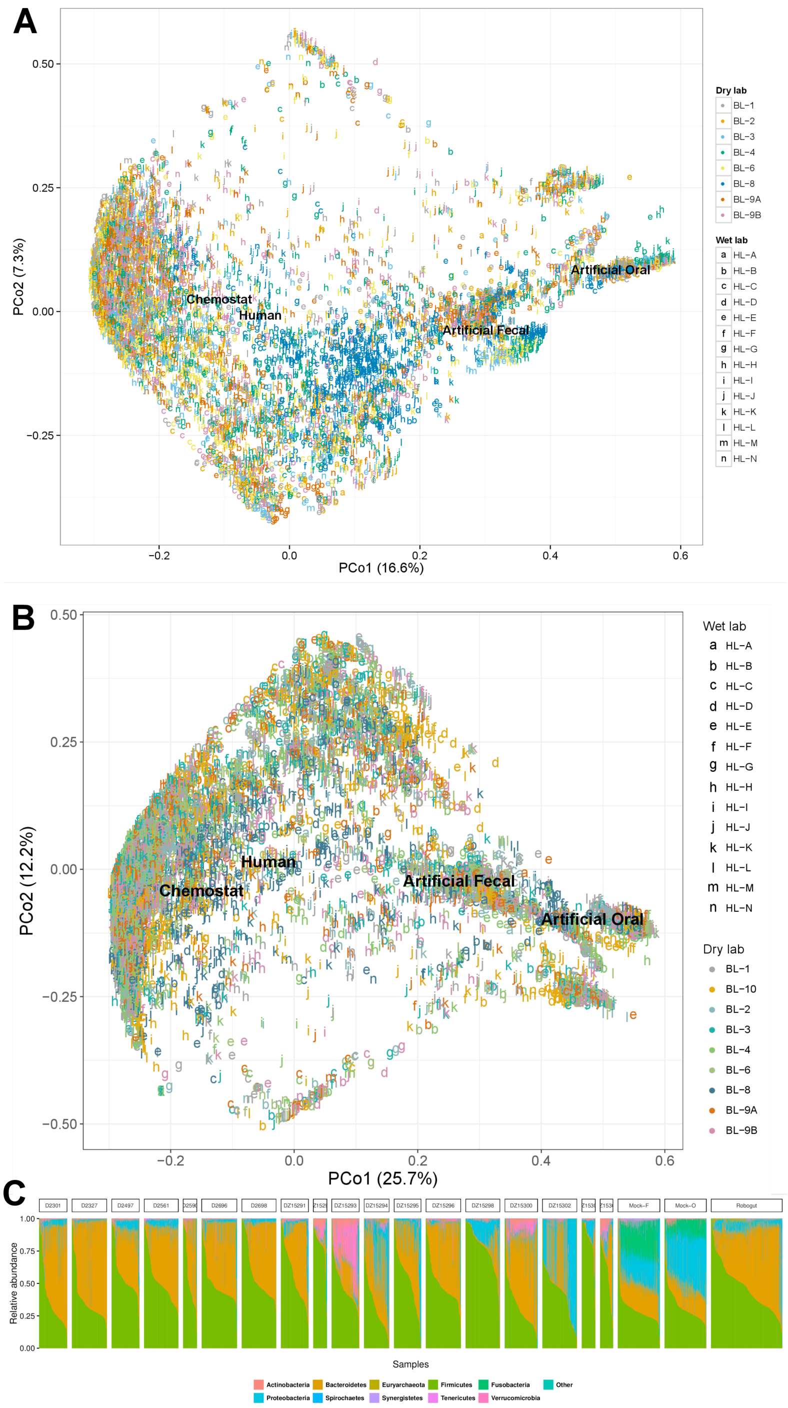
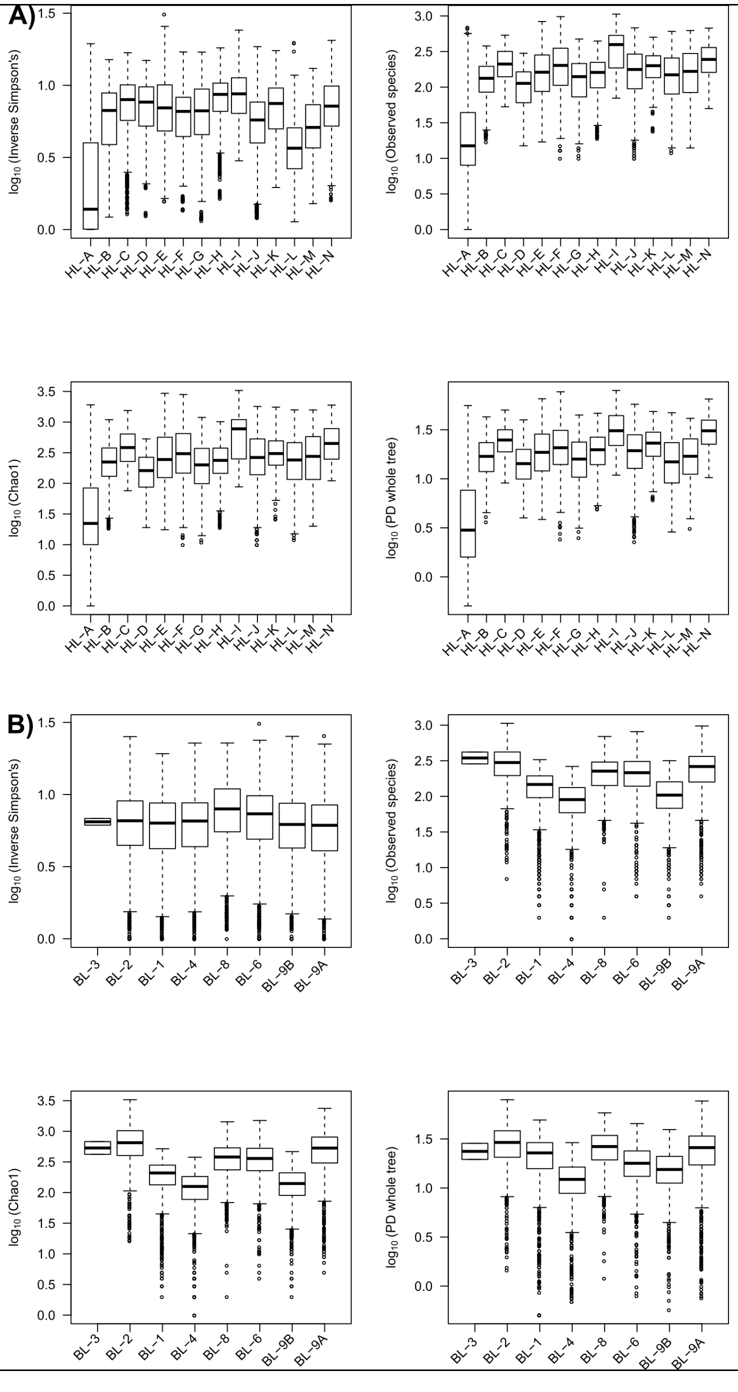


**Supplementary Figure 1**

**MBQC-base beta-diversity, major protocol variables, and taxonomic profiles.**

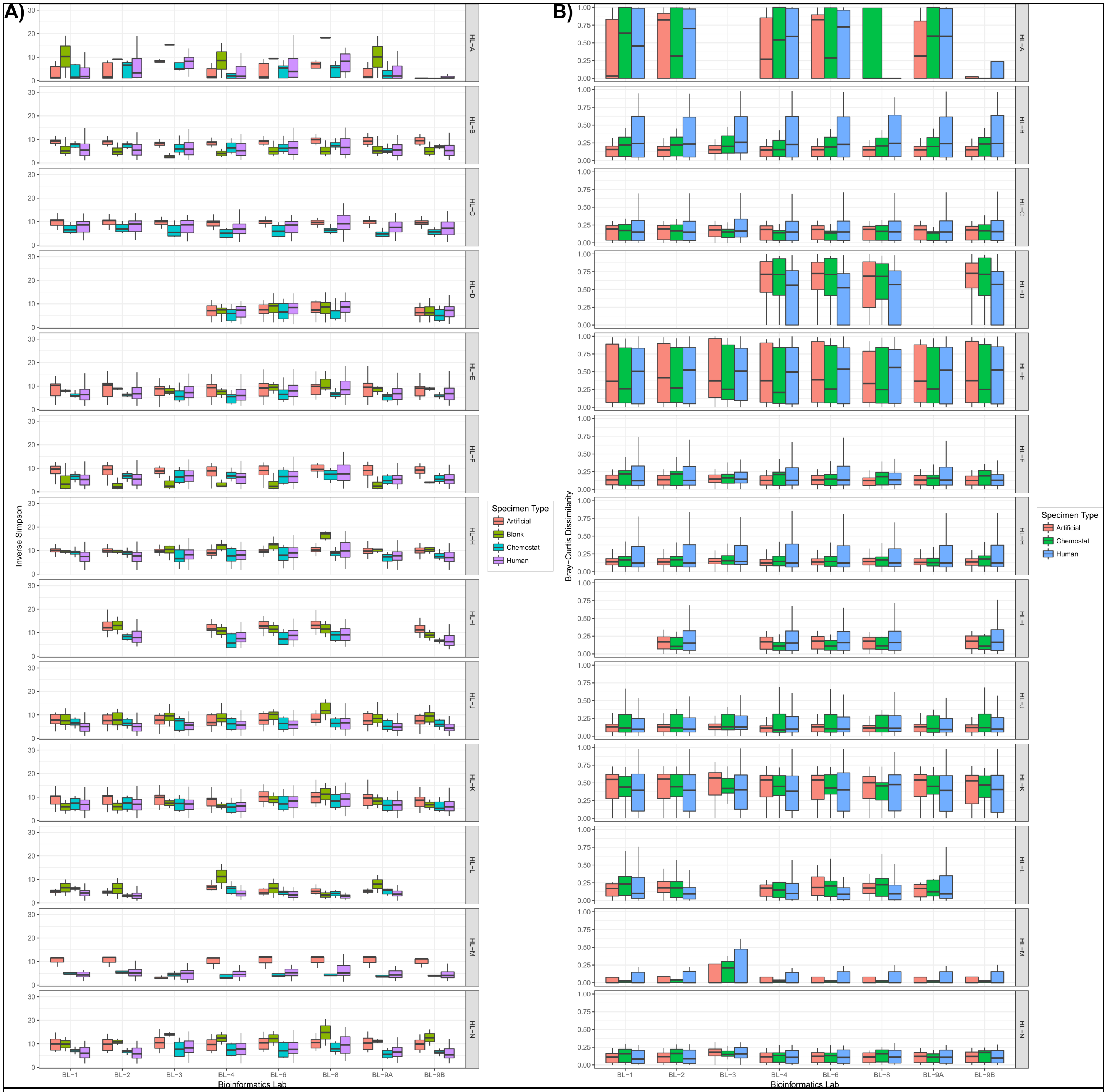
**A)** Multidimensional scaling of MBQC sample Bray-Curtis dissimilarities (see **Fig. 1**). Labels indicate centroids of the indicated sample types. **B)** As panel A, but also including post-hoc mothur-processed samples (BL-10, see Methods). Systematic taxonomic shifts from this protocol are present (see **Supplementary Dataset 6** but not of sufficient effect size to appear on the first two ordination axes. **C)** Proportions of 10 bacterial phyla that were detected with a minimum relative abundance of 0.01% in at least 10% of the 16,554 samples that were subjected to integrated analysis.



**Supplementary Figure 2**

**Within-sample alpha diversity, stratified by handling and bioinformatics lab.**

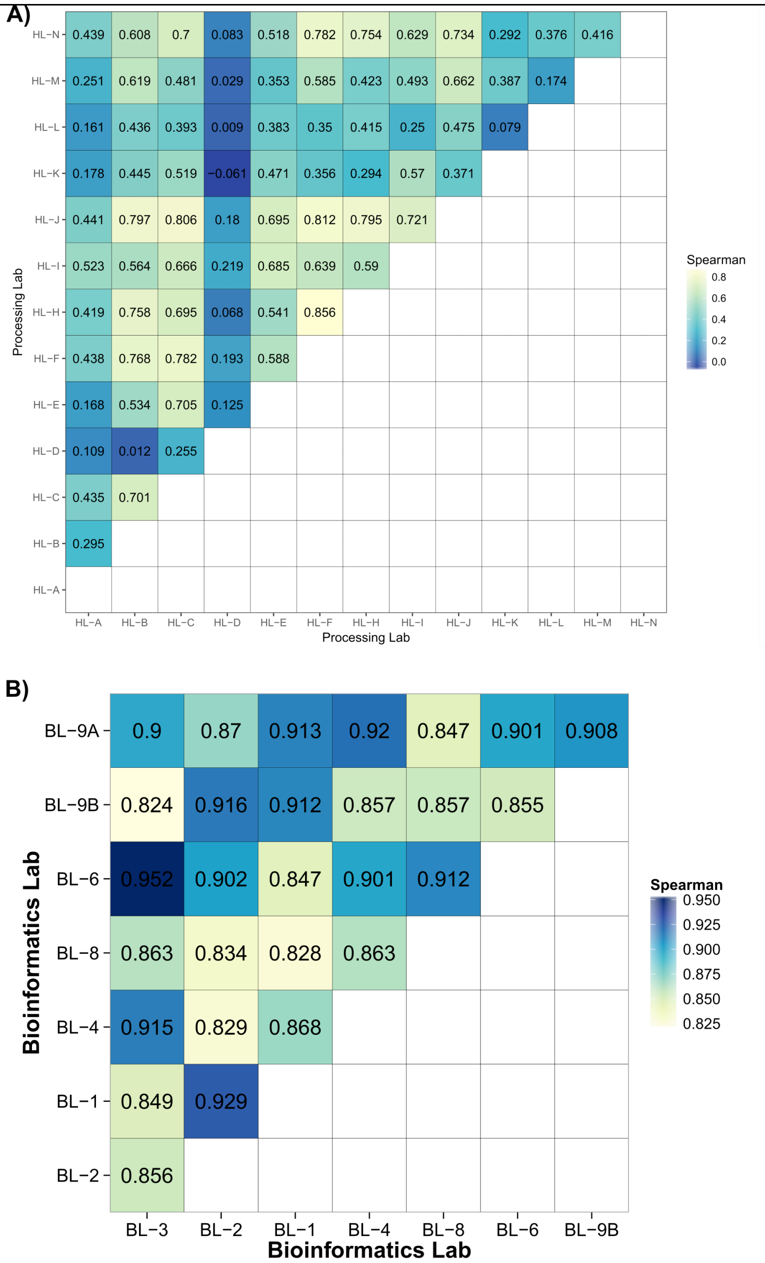
Four different alpha diversity measures (inverse Simpson, observed species richness, Chao1, and phylogenetic diversity) across all samples processed by each **A)** handling and **B)** bioinformatics lab. All diversity measures, whether qualitative (OS, Chao1, PD) or quantitative (IS) and whether taxonomic (IS, OS, Chao1) or phylogenetic (PD) correlate closely, with large but consistent differences induced by distinct handling protocol choices.



## Supplementary Figure 3

**Within-sample alpha and beta diversities, stratified by handling and bioinformatics lab.**

Distributions of **A)** within-sample Inverse Simpson alpha diversity and **B)** between-sample Bray-Curtis beta diversity, stratified by sample type (total n=2,033 for artificial communities, n=11,991 for human-derived samples, and n=1,725 for chemostat samples) handling and bioinformatics lab. As in other, higher-level summary statistics, effect sizes in decreasing order are (on average) biospecimen type, biospecimen, handling laboratory, and bioinformatics laboratory, with no bioinformatics protocols (regardless of read length or trimming) evident outliers from this trend. Outlier values outside 1.5 times the interquartile range are omitted for clarity.



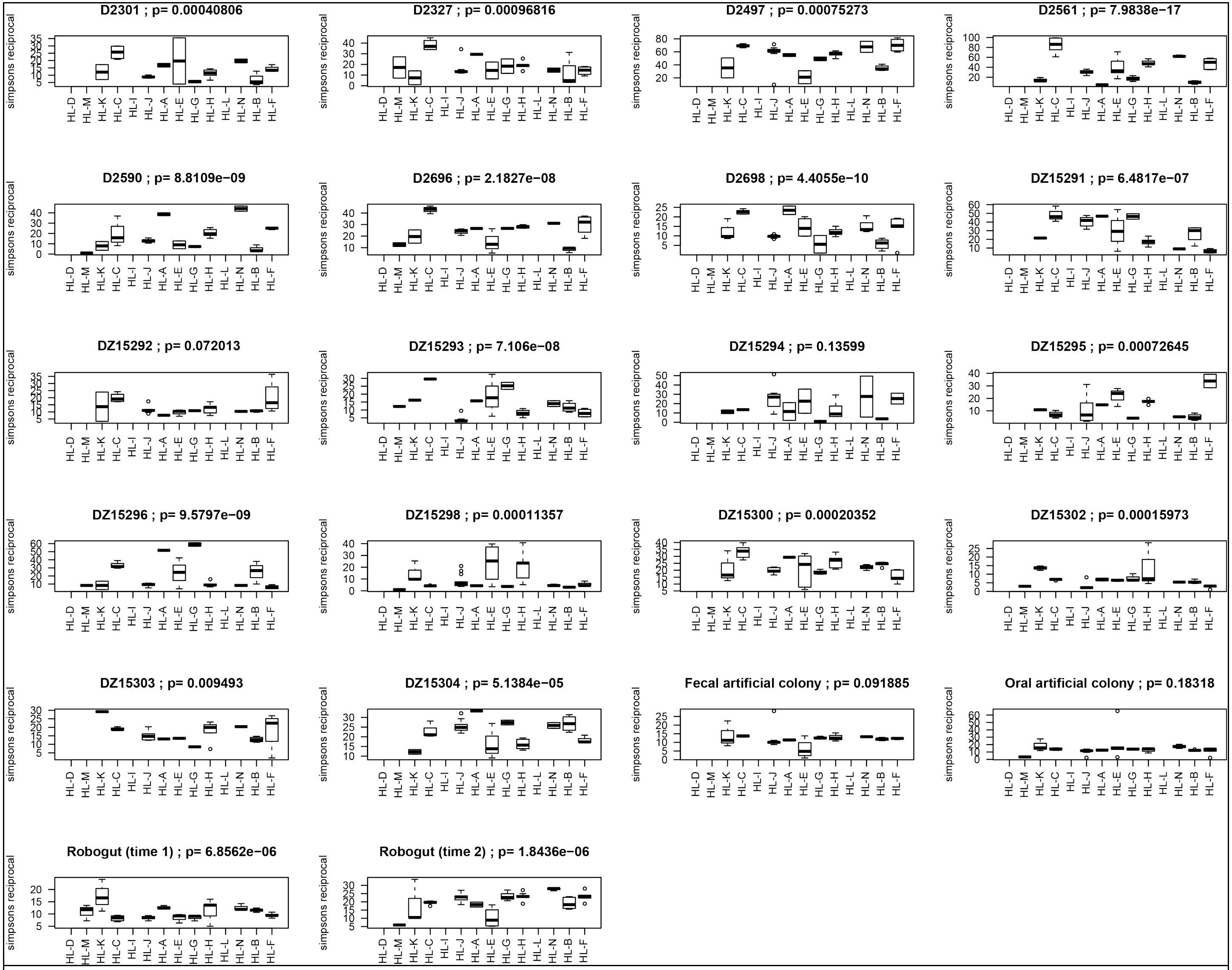
| | HL−A | HL−B | HL−C | HL−D | HL−E | HL−F | HL−H | HL−I | HL−J | HL−K | HL−L | HL−M | HL−N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HL−N | 0.439 | 0.608 | 0.7 | 0.083 | 0.518 | 0.782 | 0.754 | 0.629 | 0.734 | 0.292 | 0.376 | 0.416 | |
| HL−M | 0.251 | 0.619 | 0.481 | 0.029 | 0.353 | 0.585 | 0.423 | 0.493 | 0.662 | 0.387 | 0.174 | | |
| HL−L | 0.161 | 0.436 | 0.393 | 0.009 | 0.383 | 0.35 | 0.415 | 0.25 | 0.475 | 0.079 | | | |
| HL−K | 0.178 | 0.445 | 0.519 | −0.061 | 0.471 | 0.356 | 0.294 | 0.57 | 0.371 | | | | |
| HL−J | 0.441 | 0.797 | 0.806 | 0.18 | 0.695 | 0.812 | 0.795 | 0.721 | | | | | |
| HL−I | 0.523 | 0.564 | 0.666 | 0.219 | 0.685 | 0.639 | 0.59 | | | | | | |
| HL−H | 0.419 | 0.758 | 0.695 | 0.068 | 0.541 | 0.856 | | | | | | | |
| HL−F | 0.438 | 0.768 | 0.782 | 0.193 | 0.588 | | | | | | | | |
| HL−E | 0.168 | 0.534 | 0.705 | 0.125 | | | | | | | | | |
| HL−D | 0.109 | 0.012 | 0.255 | | | | | | | | | | |
| HL−C | 0.435 | 0.701 | | | | | | | | | | | |
| HL−B | 0.295 | | | | | | | | | | | | |
| HL−A | | | | | | | | | | | | | |

| | BL−3 | BL−2 | BL−1 | BL−4 | BL−8 | BL−6 | BL−9B |
|---|---|---|---|---|---|---|---|
| BL−9A | 0.9 | 0.87 | 0.913 | 0.92 | 0.847 | 0.901 | 0.908 |
| BL−9B | 0.824 | 0.916 | 0.912 | 0.857 | 0.857 | 0.855 | |
| BL−6 | 0.952 | 0.902 | 0.847 | 0.901 | 0.912 | | |
| BL−8 | 0.863 | 0.834 | 0.828 | 0.863 | | | |
| BL−4 | 0.915 | 0.829 | 0.868 | | | | |
| BL−1 | 0.849 | 0.929 | | | | | |
| BL−2 | 0.856 | | | | | | |

**Supplementary Figure 4**

**Correlations of alpha diversities for samples as processed by different handling and bioinformatics labs.**

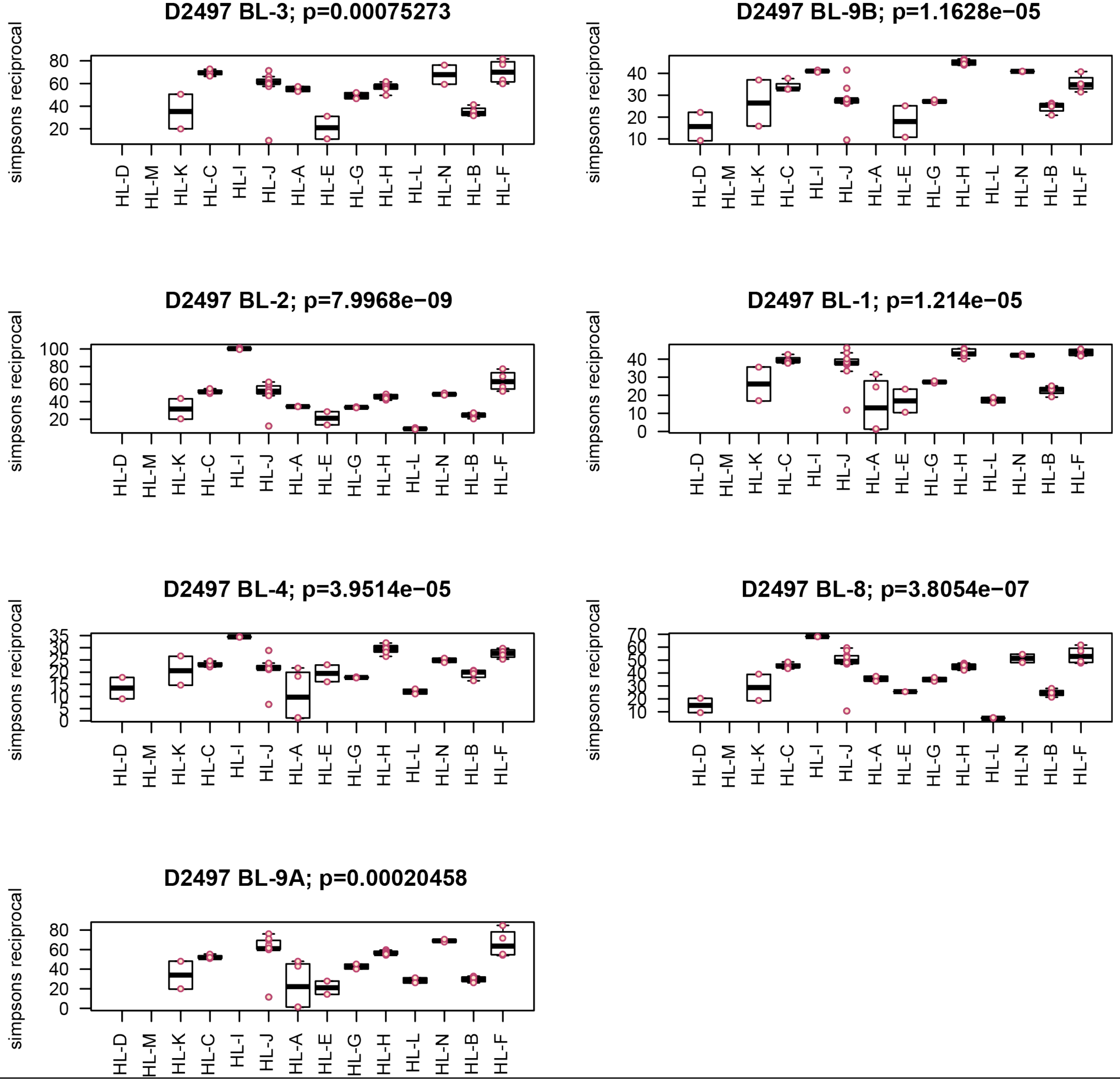
Each tile represents a Spearman rank correlation coefficient between pairwise comparisons of all log10 transformed Inverse Simpson index estimates for the overlapping subsets of 20,708 samples that survived quality control for each pair of **A)** handling and **B)** bioinformatics laboratories. High correlation in OTU Inverse Simpson estimates, which accounts for richness and evenness, across labs implies robustness (or consistent bias) in microbial community *in silico* reconstruction protocols across laboratories. Correlations in diversity are lower among handling than bioinformatics labs but generally highly significantly positive; exceptions include potential external (e.g. HL-A) and within-batch (e.g. HL-D) contaminants (see **Supplementary Fig. 14**).



**Supplementary Figure 5**

**Simpson diversity estimates for each individual sample under a single bioinformatics protocol (BL-3).**

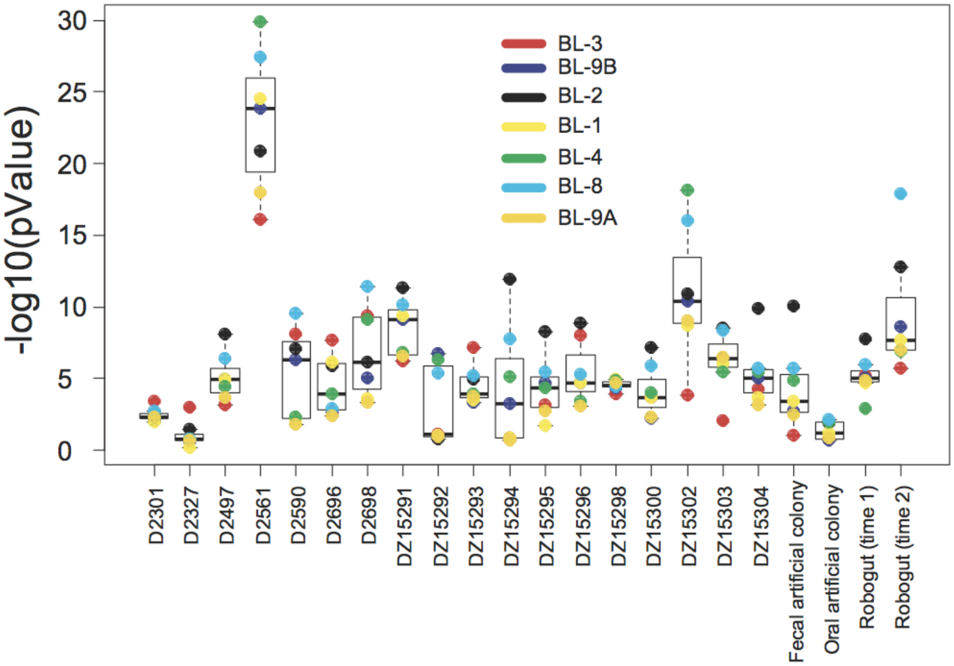
Each boxplot within each panel represents data from a distinct handling lab extraction event. A null hypothesis of no difference for each sample is evaluated by one-way ANOVA with nominal p-values shown above each panel.



**Supplementary Figure 6**

**Simpson’s diversity for a specific sample under different bioinformatics protocols.**

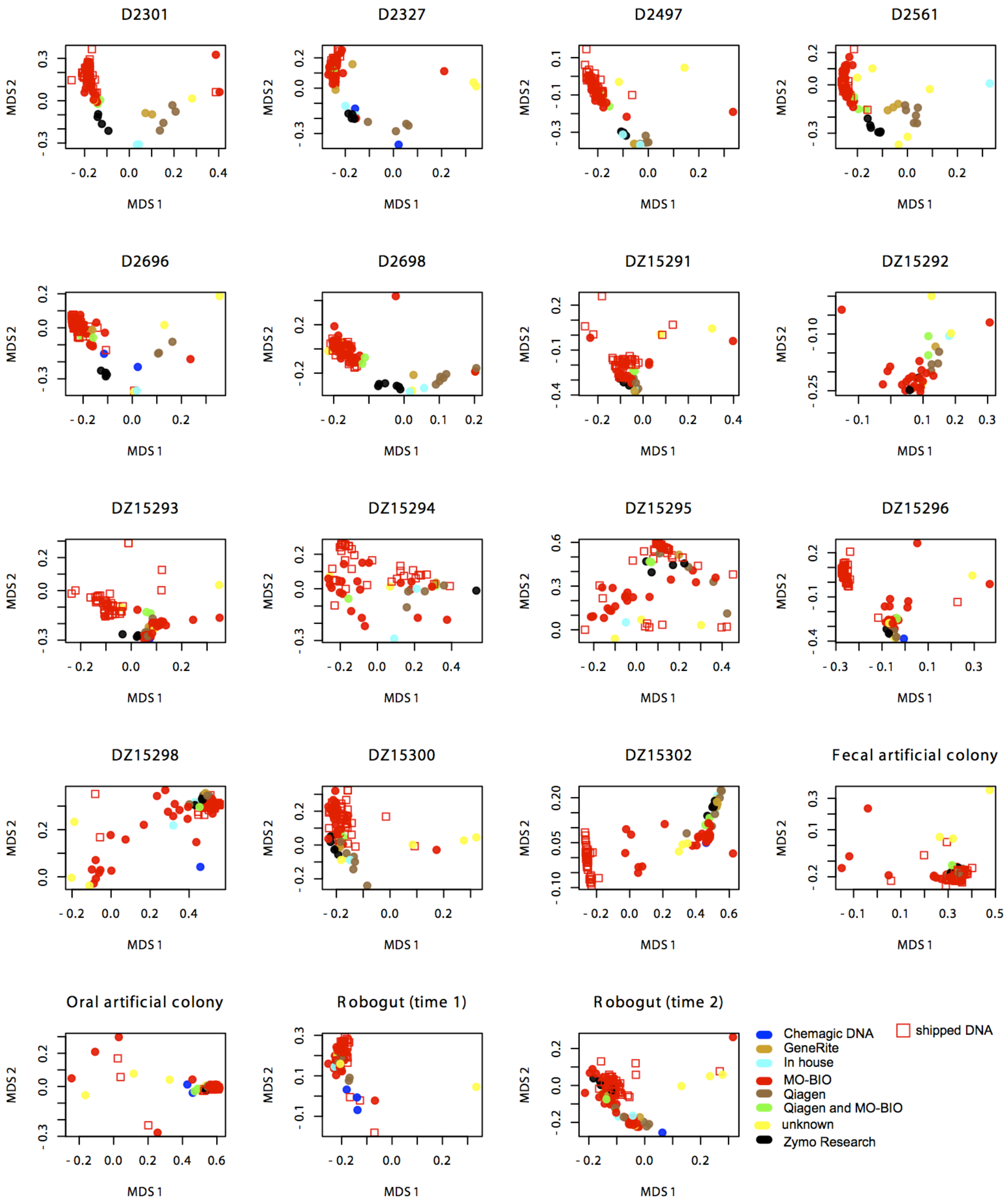
Using a single sample (D2497), the overall pattern of diversity is similar between different bioinformatics protocols, but the absolute diversity reported by each protocol varies by up to a factor of two or more. Different bioinformatics protocols when applied to the same sequences, therefore, do not produce absolute diversity estimates that are directly comparable. Direct comparisons of absolute alpha diversity, therefore, are most feasible for data processed by a single bioinformatics protocol, while relative alpha diversities can be more safely compared between protocols.



**Supplementary Figure 7**

**Individual samples show differences in how much diversity estimates are dependent on wet lab extraction.**

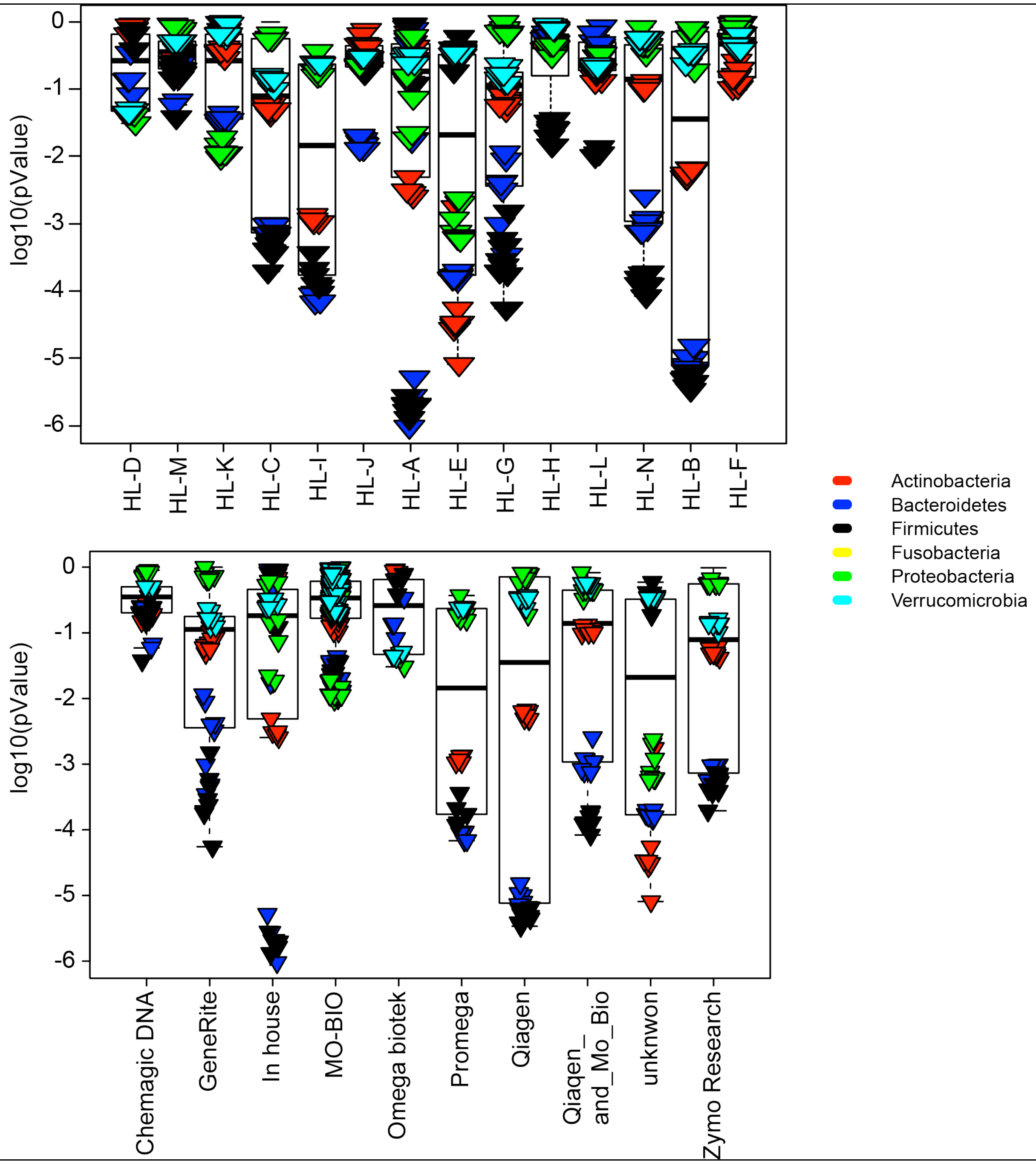
Each column represents one specimen, with the distribution of p-values across bioinformatics protocols (colors) testing whether handling lab has a significant effect on the estimation of Simpson diversity. For some samples (such as D2327), diversity estimates across different handling labs were not affected by bioinformatics protocol, while for others (such as D2561) different handling protocols produced very different absolute diversity measurements depending on bioinformatics protocol. In the case of D2561, a freeze-dried specimen, different extraction protocols produced unusually variable distributions of Bacteroidetes versus Firmicutes, and some extraction results included large proportions of likely contaminants such as *Methylobacterium*, *Staphylococcus*, and *Spirochaetes*. A smaller study only incorporating a few specimens or technical replicates using the same specimen might thus reach different conclusions based solely on which specimens happen to be included.



**Supplementary Figure 8**

**Ordination for each sample comparing replicates extracted with different kits for a single bioinformatics pipeline.**

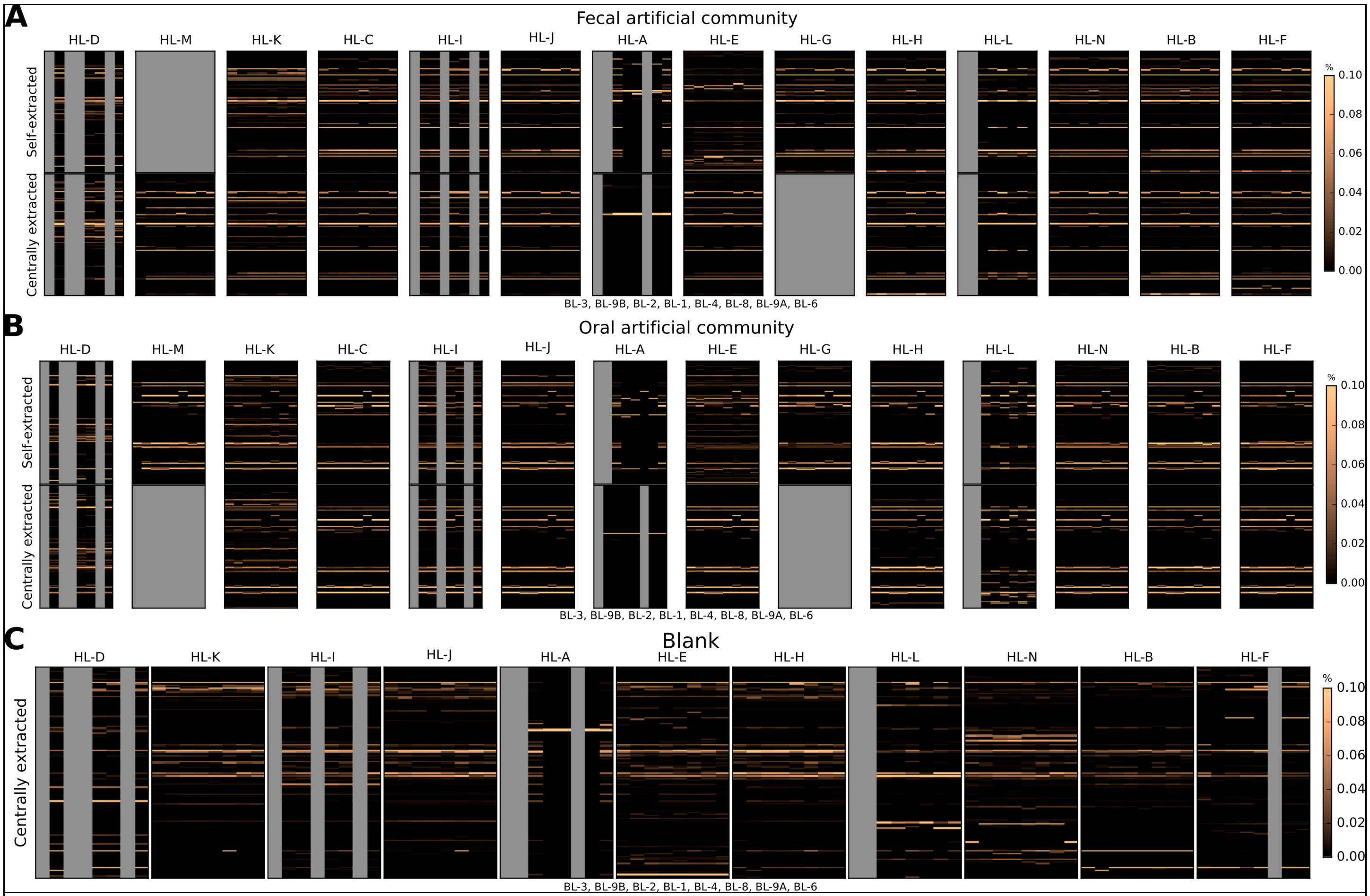
For each sample (panel), MDS ordination was performed using Bray-Curtis dissimilarity to examine how each kit (color) influenced microbial composition. In general, samples extracted centrally (open symbols) and in each lab (filled symbols) overlapped when local extraction used the same extraction kit as centralized extraction (Mo-Bio, red symbols), but this was not true for all samples (for example sample DZ15294).



**Supplementary Figure 9**

**Different extraction kits produce different estimates of relative abundance.**

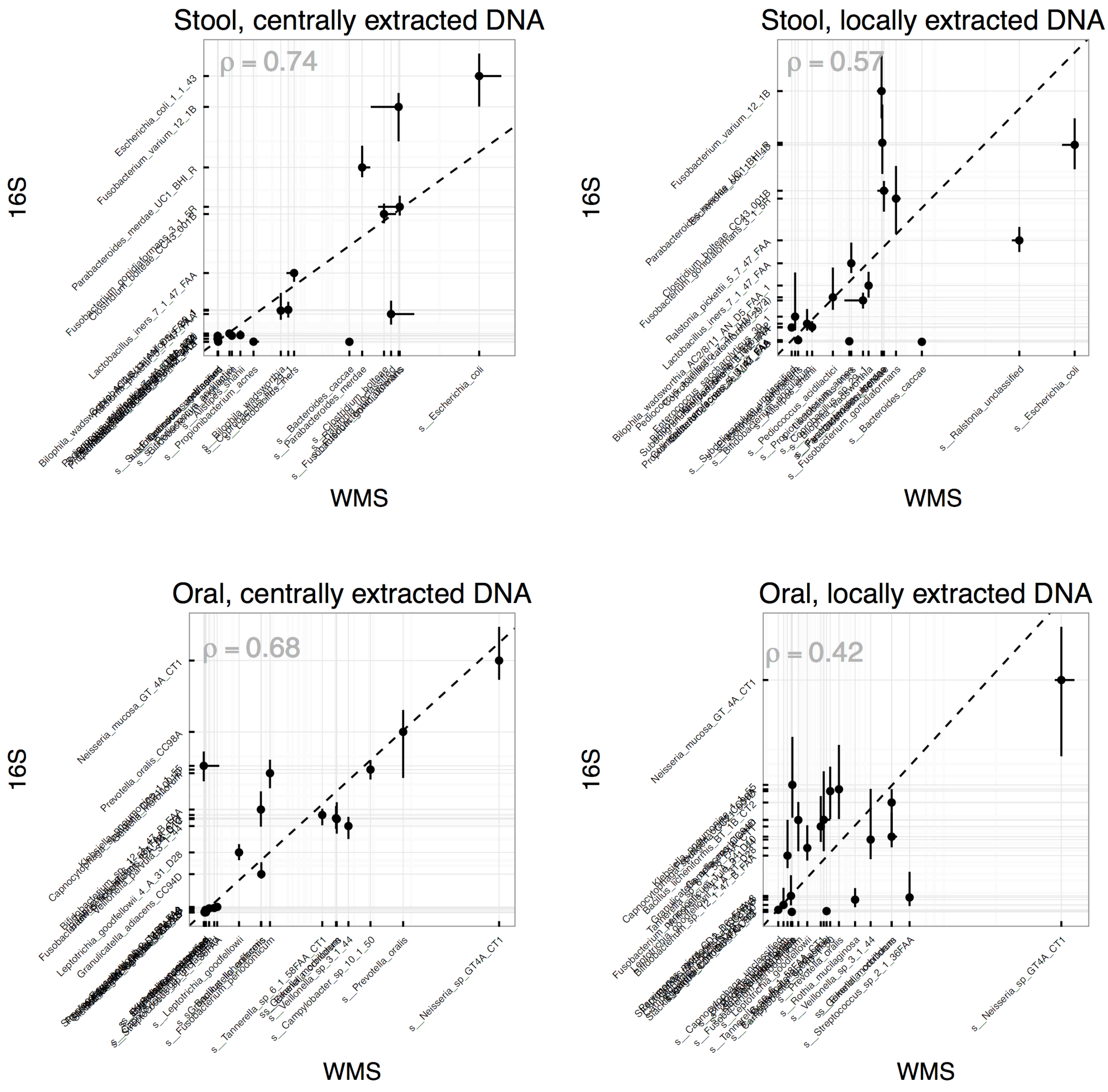
For each extraction lab (top panel) and kit manufacturer (bottom panel), log10(p-values) from a paired t-test for no difference between shipped and locally extracted DNA. Different colors represent results from different phyla as indicated in the figure legend. Bioinformatics pipelines agreed closely at the phylum level, leading to clusters of near-identical points for each color. When kits from certain manufacturers (such as MO-BIO and Omega biotek) were used by the locally extracting lab, there was good agreement with the shipped DNA (which was extracted with a MO-Bio kit). However, for other kits, there were substantial differences in the relative abundance calls at the phylum level, producing small p-values.



**Supplementary Figure 10**

**Taxonomic profiles of positive and negative control samples stratified by extraction, handling lab, and bioinformatics.**

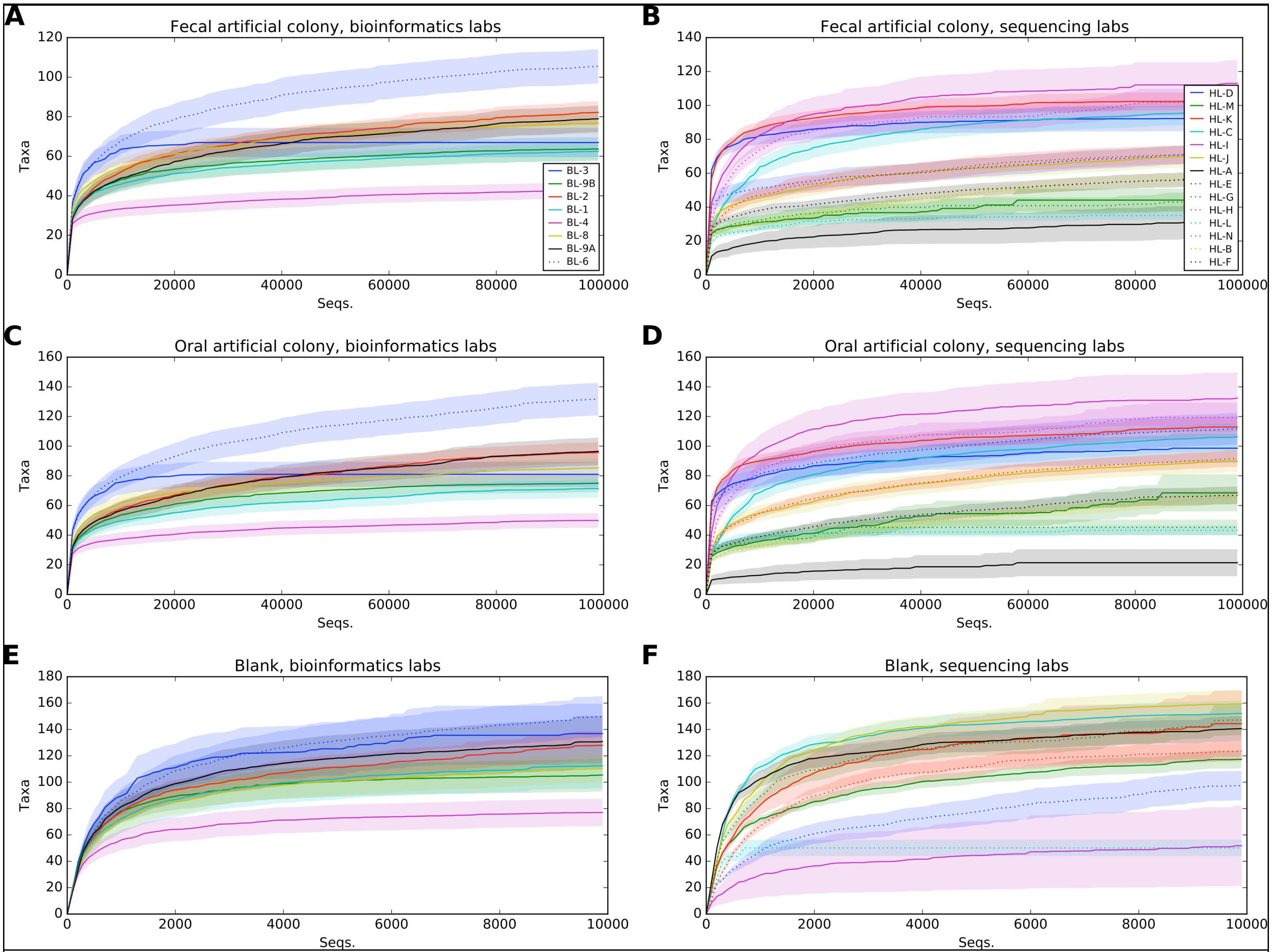
Average relative abundance profiles for **A)** fecal and **B)** oral artificial communities (see **Supplementary Table 1**) and **C)** buffer blank sample profiles stratified by extraction location (super-row), handling lab (super-column), and bioinformatics lab (column). Only taxa (rows) achieving at least 0.1% relative abundance in at least one sample are shown; combinations for which no data were provided are gray.



**Supplementary Figure 11**

**Correlation between taxonomic profiles from whole metagenome shotgun (WMS) and 16S amplicon sequence data on stool and oral artificial communities.**

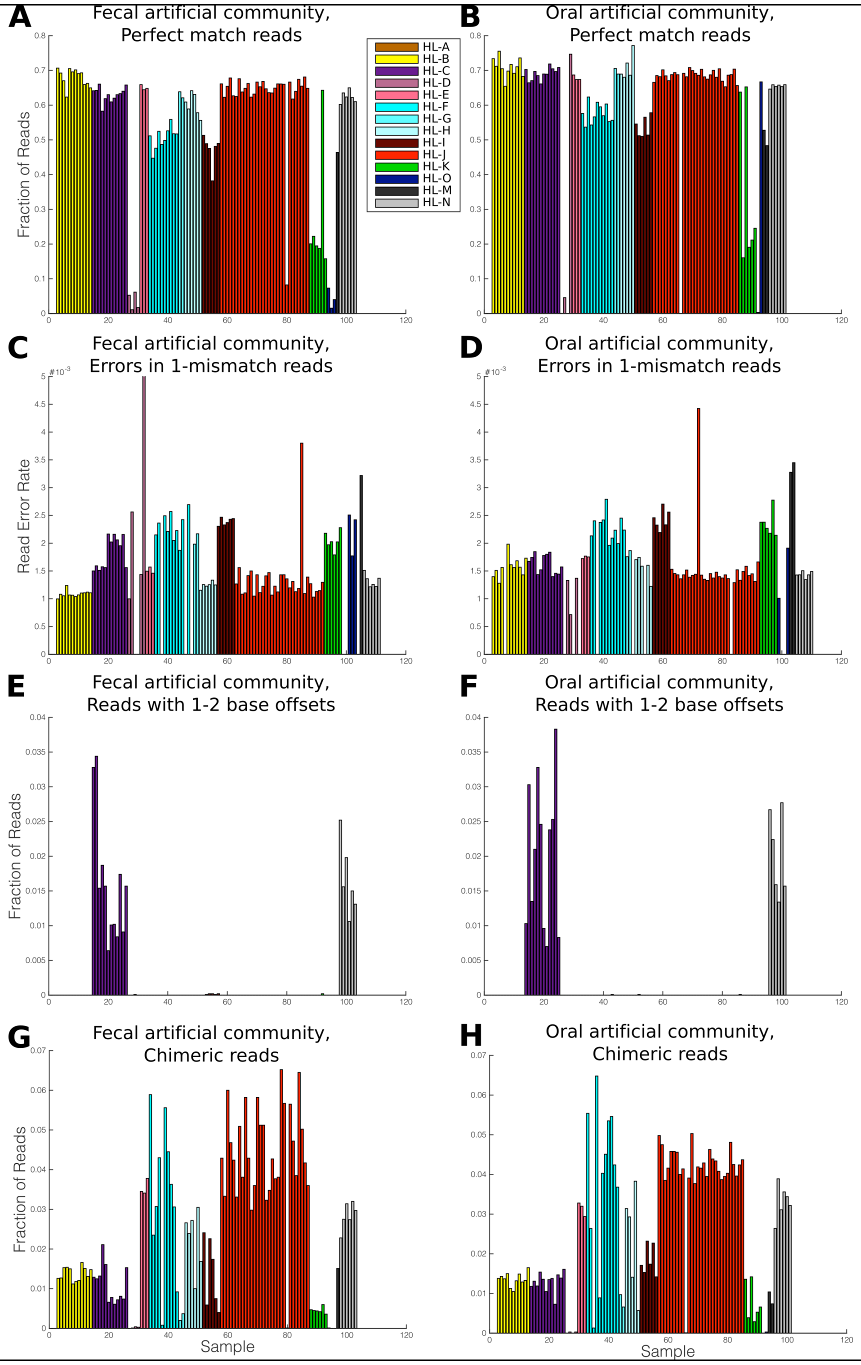
Relative abundances calculated from WMS (horizontal) and 16S rRNA gene (vertical) artificial community sequence data for 17 and 19 species, from gut and oral artificial communities, which were identifiable from both sequence sets. Each point represents the species relative abundance interquartile ranges (IQRs) for WMS and 16S; the IQRs intersect at respective median values. Spearman rho correlation coefficients are shown in the top left of each plot. The dashed diagonal line represents the diagonal. Data are summarized from 43, 36, 43, and 40 artificial community 16S amplicon samples for gut centrally and locally, and oral centrally and locally extracted DNA samples, respectively; twelve WMS samples, three in each respective group, were summarized in each subplot.



**Supplementary Figure 12**

**Alpha diversity of gut- and oral-derived artificial communities and negative control blanks as stratified by handling and bioinformatics laboratories.**

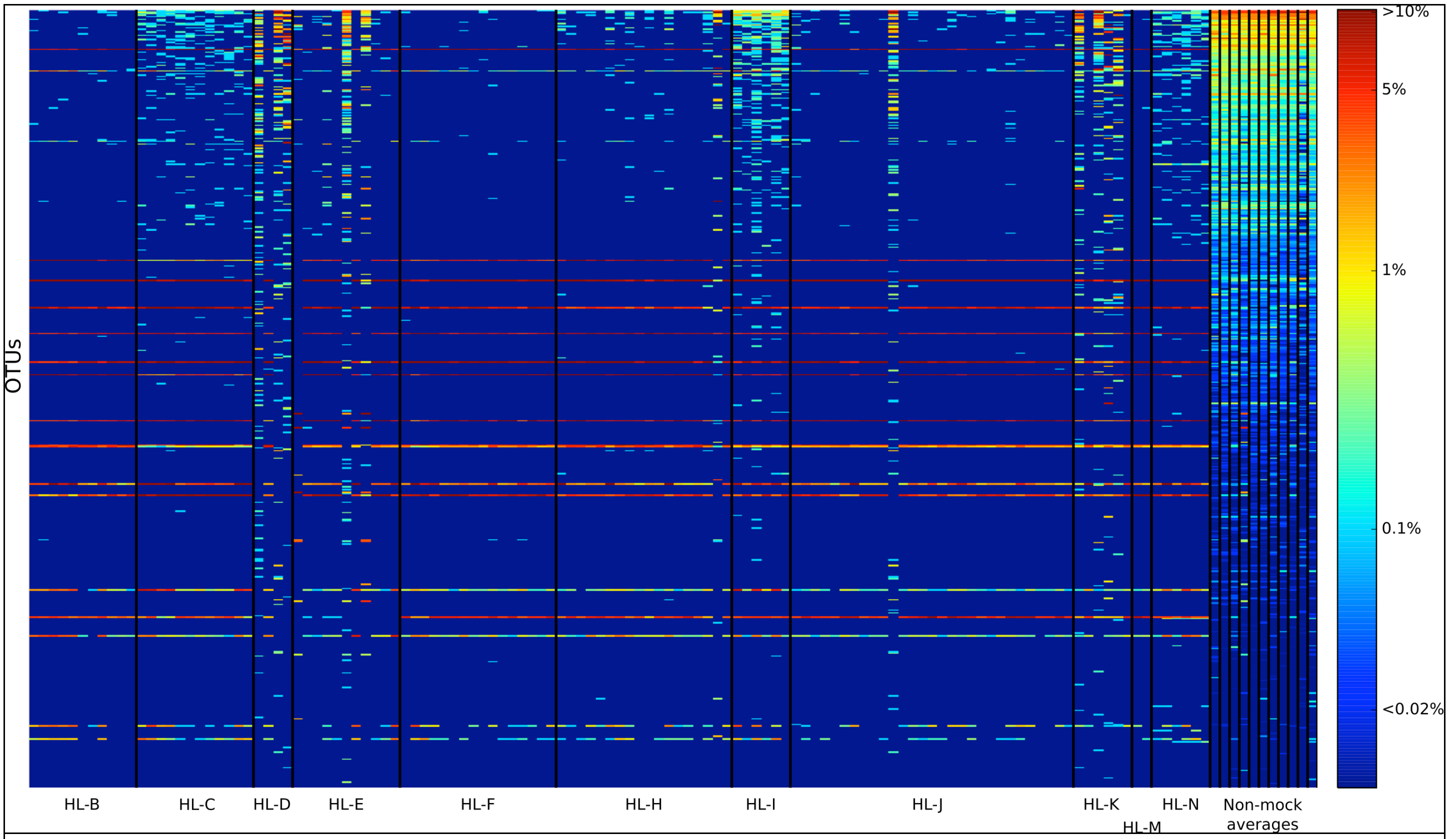
Rarefaction curves for mean number of OTUs as a function of rarefaction depth for the fecal (**A-B**) and oral (**C-D**) artificial communities and for negative control blanks (**E-F**). Means and standard error for a minimum of 5 samples are shown for different bioinformatic pipelines (average across handling labs; A, C, E) or for different handling labs (averaging across bioinformatics; B, D, F). Target values are 20 for fecal and 22 for oral artificial communities, respectively.



**Supplementary Figure 13**

**Mismatched raw reads in artificial communities account for ~30% of sequences and vary by handling laboratory.**

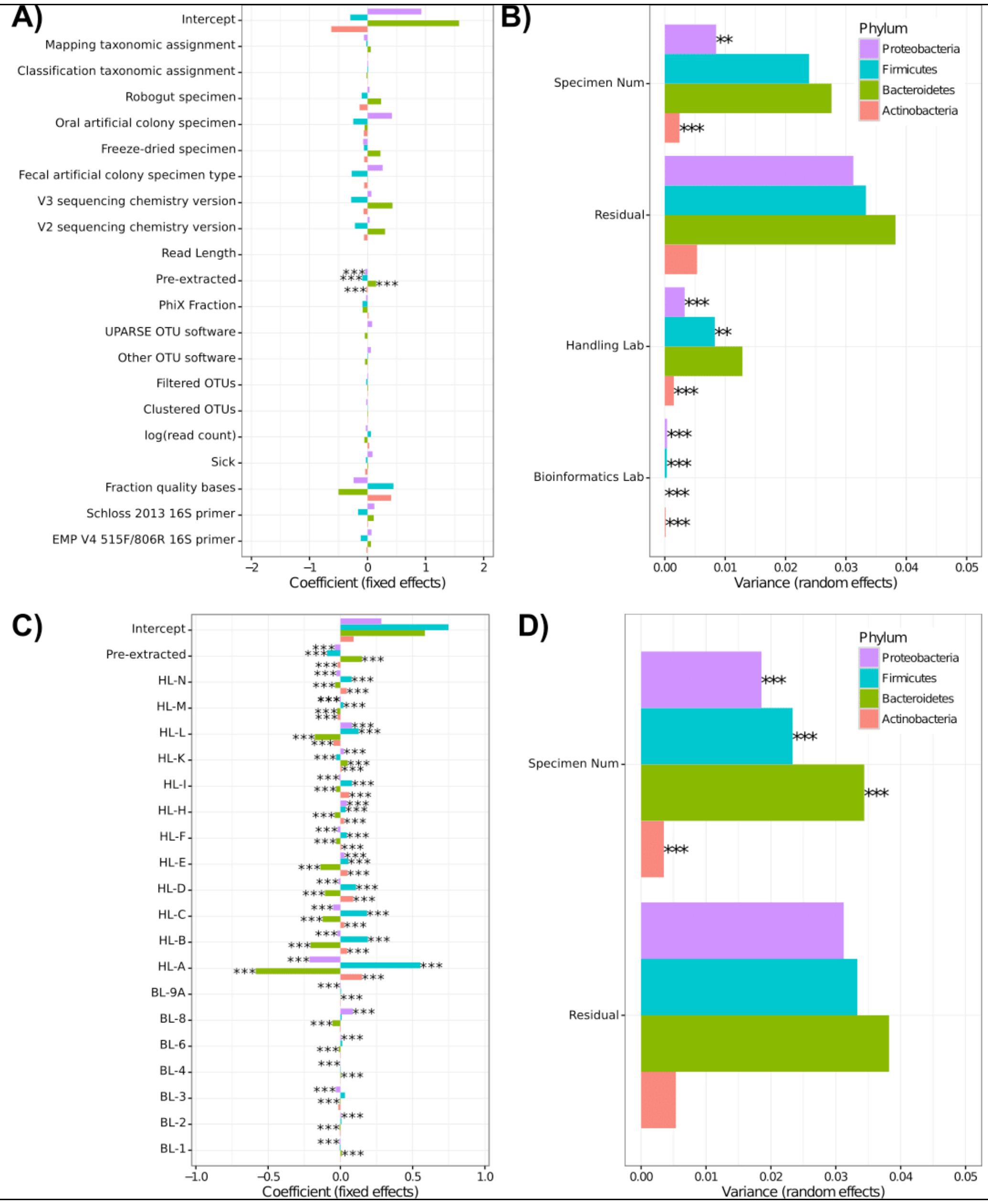
**A)** Fecal and **B)** oral artificial community fractions of reads exactly matching one of the 20 or 22 reference 16S rRNA gene sequences, respectively. Each bar represents one sample, with copy numbers varying depending on the data deposited and the number of sample sets handled. **C)** Fecal and **D)** oral per-nucleotide error rates estimated based on reads containing exactly one mismatch to reference 16S rRNA gene sequences. **E)** Fecal and **F)** oral reads identical to references but offset by either one or two nucleotides. **G)** Fecal and **H)** oral reads containing chimeric sequences from two known community members (as determined by exhaustive search of all reference pairs).



**Supplementary Figure 14**

**Reads in oral artificial communities are often apparently derived from abundant taxa in non-artificial samples.**

Rows correspond to OTUs abundant in non-artificial MBQC-base samples, columns to samples (grouped by handling lab). Averages on the right are per lab in the same order across all non-artificial samples. Only results from the BL-9B bioinformatics method are shown for simplicity, and handling lab HL-A samples are omitted due to an incongruous pattern of apparent external contamination by other organisms (see **Supplementary Fig. 10**).



**Supplementary Figure 15**

**Significance of all variables from a multivariate model of experimental and bioinformatic protocol variables.**

**A)** Magnitudes and significance levels of only fixed effects from a random effects model capturing all handling and bioinformatics lab variables for which sufficient measurements were available (see **Supplementary Table 7**). Comparably high variability in phylum-level taxonomic abundance readouts was associated with subsets of both handling and bioinformatics experimental protocol choices. **B)** Magnitudes and significance levels of only random effects from the full model. These highlight large differences induced by biological variation relative to sample handling or bioinformatics protocol choices on the resulting abundances. **C)** Random effects from a simplified model including only individual handling and bioinformatics laboratory identifiers and differences between pre- and locally-extracted samples. **D)** Fixed effects of sample handling and bioinformatics laboratories from the simplified model. These suggest that that variability in taxonomic profiling is primarily driven by sample handling protocol choices, with lesser differences induced at the phylum level by bioinformatics choices. All effects were evaluated using a likelihood ratio test with Benjamini-Hochberg-Yekutieli correction within models.