M.C Ricardo Gomez Reyes
Autonomous University of Baja California

Exploratory data analysis for 16S-metagenomic (V3 amplicon) libraries sequencing with illumina Miniseq from Gut microbiota of shrimp (*Litopenaeus vannamei*) across foregut, midgut, and intestine.

Compositional taxonomic summary

A set of 609 Amplicon Sequences Variants (henceforth, features) were checked from its lowest to highest taxonomic rank in order to determine the level of resolution (Figure 1). Here is defined resolution as the last taxonomic level (Linnaean system), where the features were classified with confidence (the astringency method depends on the classifier algorithm used). Redundant terms as "_unclassified | Unclassified", "Undetermined", "sp." as well as "Unknown" were homogenized as missing data in order to measure the taxonomic resolution. (Those abbreviations usually came from either the classifier algorithm or the data-base used as reference).

The features can be described in a good way until the Family taxonomic rank (85.8% of the features are within this level). Moreover, we found a positive relationship (spearman correlation of 0.944) between features and abundant percentages. Therefore, we decide to use as *bonafide* the family level for downstream analysis.

M.C Ricardo Gomez Reyes
Autonomous University of Baja California

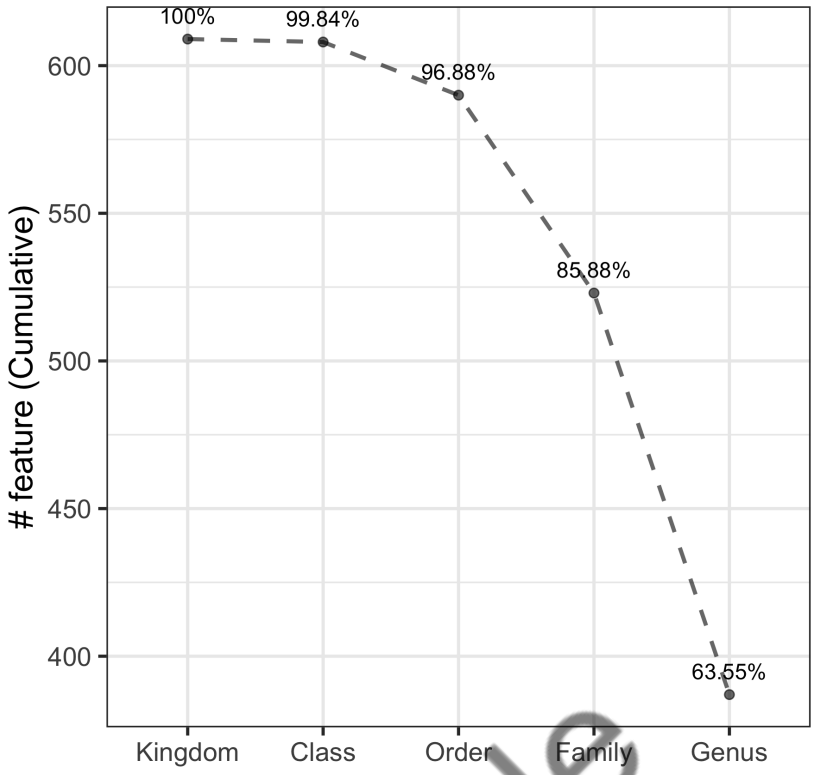

Figure 1. Percent of features that move out each taxonomic level. After Family level, the percent of features drop to 63.5 %. At the genus level 386 from 609 features are discarded.

After term cleaning, a set of 14 Phyla were obtained, here, *Proteobacteria* was quantified as the highest group of bacteria present in the data, representing 68% of the abundance, followed by *Actinobacteria* (20 %), *Verrucomicrobia* (6%) and *Firmicutes* (3 %). The phyla was divided into 116 Families. Here a value of *divergence* was calculated (1-(1/length(unique(Family)). This score can be understood as the fraction of features in a group (Families) that diverge from his highest Ancestor (Phylum). For at least *Verrucomicrobia, Planctomycetes, Actinobacteria, Bacteroidetes, Firmicutes* and *Proteobacteria* we found a positive relationship (Pearson = 0.75) between the divergence of groups and families observed (figure below).

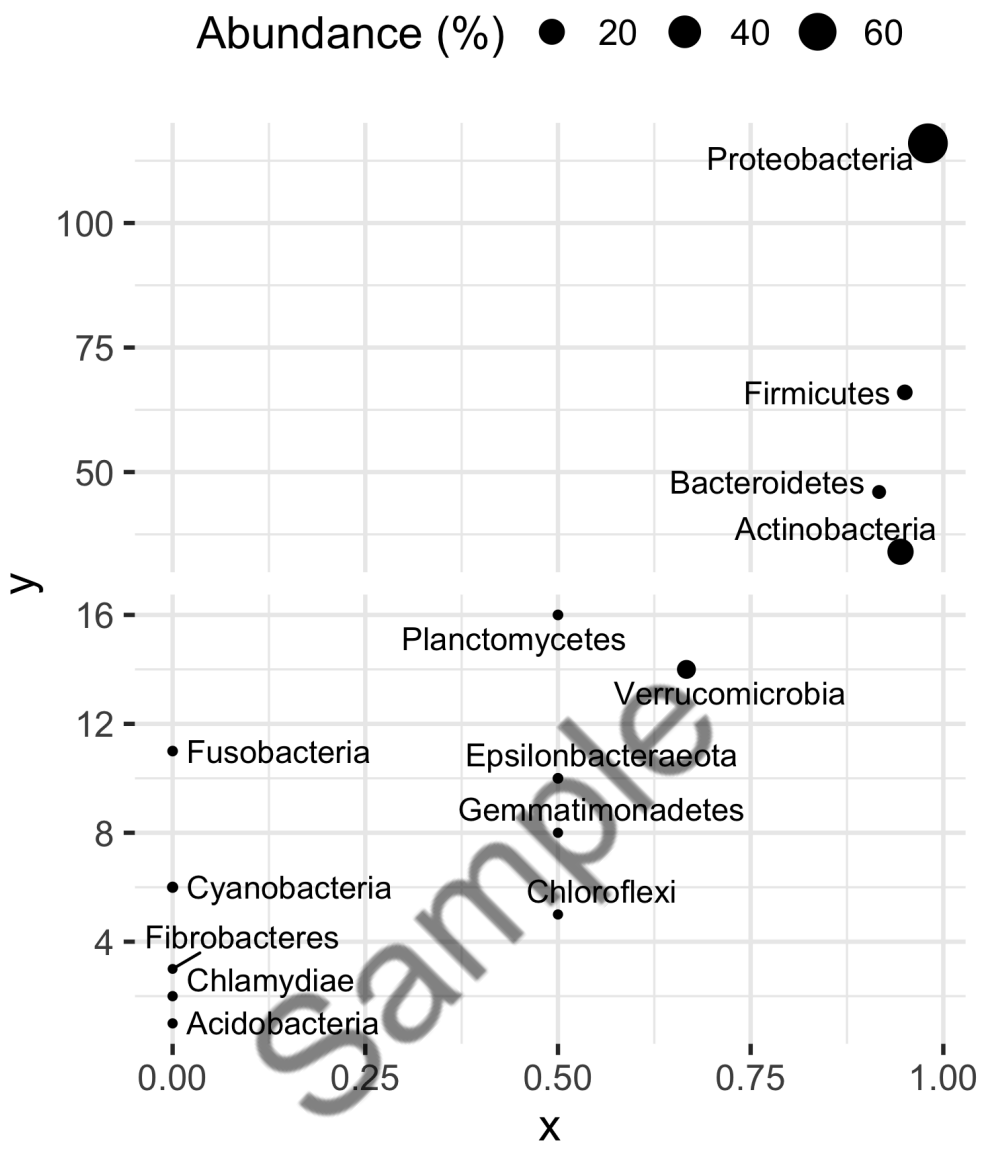

Figure 2. The cumulative number of families (y-axis) against his divergence score (x-axis). A cumulative sum of 13 families are Low abundance (Relative Ab < 1 %). These groups include a set of single families with divergence <= 0.5.

Low abundance filtering

The previous filtration steps all depended upon the taxonomic annotations. Ignoring taxonomies, we used prevalence filtering as a form of unsupervised filtration. As the description in figure 2, *Proteobacteria* have the highest spread in the Prevalences, followed by *Actinobacteria, Verrucomicrobia, Firmicutes, Bacteroidetes,* and *Cyanobacteria* (Figure 4). A total set of 577 features were retained after this filtration step (180,879 (0.996 %) absolute abundance).

M.C Ricardo Gomez Reyes
Autonomous University of Baja California

Phylum	cor	mean prevalence	total abundance
Proteobacteria	0.644	4.08	123669 (68%)
Actinobacteria	0.743	2.86	37093 (20 %)
Verrucomicrobia	0.793	10.2	11093 (6%)
Firmicutes	0.614	1.5	6435 (3.5)
Bacteroidetes	0.832	2.41	2273 (1.2%)
Cyanobacteria	0.983	2.12	316 (0.2 %)
Planctomycetes	0.901	2.77	176
Patescibacteria	0.968	3.5	121
Epsilonbacteraeota	0.99	2.33	102
Fusobacteria	-0.90	1.67	76
Chlamydiae	-	2	25
Acidobacteria	-	1	16
Chloroflexi	-	1	16
Gemmatimonadetes	-	1.5	16
Dependentiae	-	2	2
Fibrobacteres	-	1	2

Table 1. Prevalence of features at the Phylum level based on the 1% of the total abundance as a threshold to set the low taxa (horizontal blue line in the table). Percent of total abundance in parenthesis. Total reads = 181,475 (100%). Total reads after filtering = 180,879 (99.6 %).

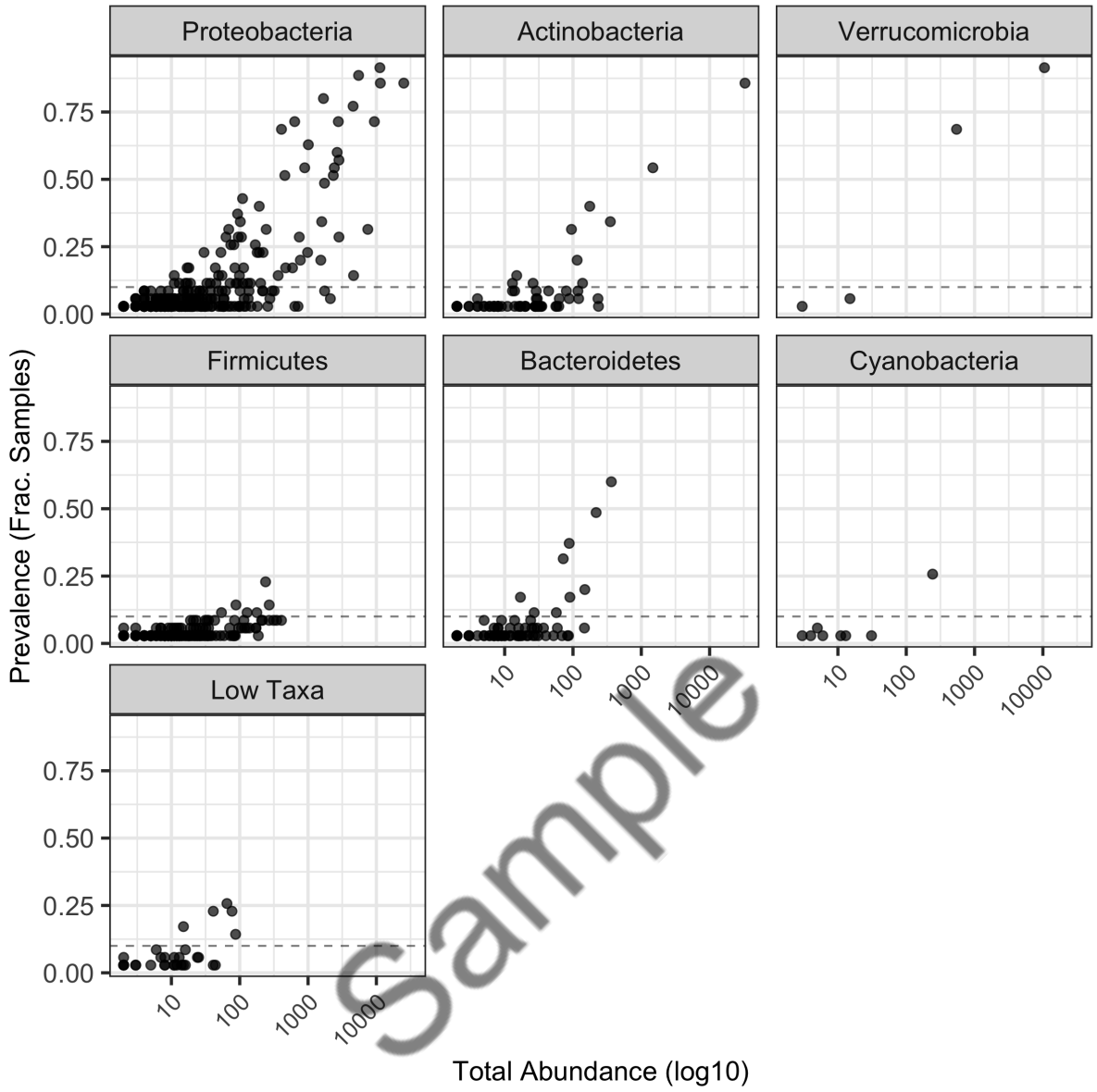

Figure 4. The Fraction of Prevalence (y - axis) versus Total abundance (x axis) shows a positive relationship: *Proteobacteria* (0.643), *Actinobacteria* (0.743), *Verrucomicrobia* (0.767), *Firmicutes* (0.614), *Bacteroidetes* (0.831), *Cyanobacteria* (0.983). According to table 1. Low taxa are the group of taxa with abundance lower than 1% of the total abundance.

Dominant groups

In order to test differential groups of taxa, the feature table was agglomerated at the family level. Because the filtering was unsupervised upon the taxonomic annotations, the undetermined families were labeled as incomplete lineages. Here, we assumed that the features (prior to filtering steps in the bioinformatic

workflow, ej. chimeric removal, quality read filtering, low filtering features, etc.) represent a *combination* of real biological variants that remain insufficiently documented due to a lack of representatives in the reference database (Hibbet D., et al 2011; Porter et al 2018). Figure 5, shows a clear example of an incomplete lineage from the facet *Proteobacteria* (*Alphaproteobacter* Order) had a prevalence in > 75% of the samples.

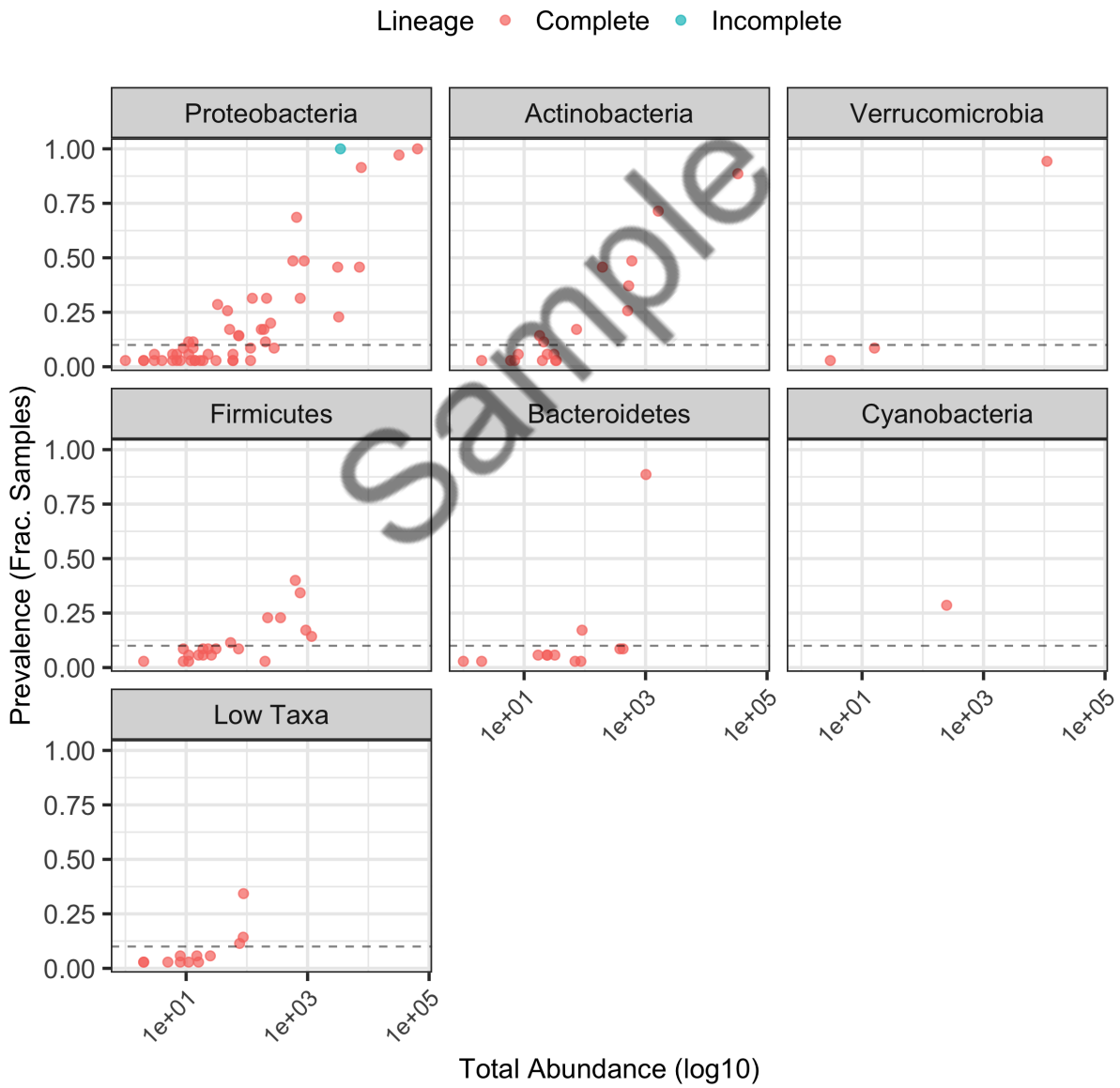

Figure 5. The Fraction of Prevalence (y - axis) versus Total abundance (x axis). The feature table was agglomerated at the family level. See table 1 to know the low taxa Phylums. Low taxa was not presented in < 35% of the samples (y-axis of the Low taxa panel) and total absolute abundance < 1%.