Probability-based Set Relationships

Set relationships among shared categories, e.g. taxonomy or function, between groups of samples cannot be accurately declared when analyzing datasets whose sampling cannot be fully saturated. This is the situation with metagenomic sequencing, where the richness of the assayed characteristics will be dependent on the sequencing depth recovered across the samples taken. For categories with very low abundances, there is a probability that their recovery, at any abundance, will not be possible upon re-performing the sampling and sequencing process. Thus, it is necessary to take into account, not only the abundance of categories assayed, but also the variation across samples which were assumed to be from the same group. To compute probability-based set overlap between groups, bootstrapping was performed, resampling with replacement both the samples in each group and the reads acquired, to take into account inter-sample and intra-sample variation, respectively. Across the bootstrap iterations, the number of set relationships (A only, B only, A and B, neither A nor B) that was assigned for each category between groups was accumulated. When the number of each of the set relationships was divided by the number of bootstrap iterations, each of the resultant values would then be the probability of that set relationship being declared upon re-performing the sampling and sequencing process for each category. This produced the information used in the Venn bar plots, where the length of the bar represented the probability of that set relationship being re-established upon re-performing the experiment. To further summarize the set relationships across all the categories, the set relationship with the greatest probability for each category was chosen as its most-likely assignment. This most-likely set relationship assignment was then utilized in the final Venn diagram.