

Brassicaceae microbiome response to insect herbivory: summarizing random forest, network analysis, and differential abundance



Pedro Beschoren da Costa¹, Marcela Aragón Gómez¹, Zulema Carracedo

Lorenzo¹, Karen Kloth¹, Erik Poelman¹, Marcel Dicke¹

Wageningen University and Research - Entomology Department, Wageningen, Netherlands



Introduction

Plant-microbe-insect interactions are very complex, but can still be exploited for sustainable agriculture. Data from microbial community sequencing is also complex, despite the numerous analytical tools available. While access to machine learning or network construction algorithms to analyze such data is not a limiting factor, finding a comprehensive pattern in the analysis output can still be very challenging. Here we apply different analytical methods to microbial communities of Brassicaceae plants under herbivory stress, and then summarize the multi-method outputs into a single figure.

Results

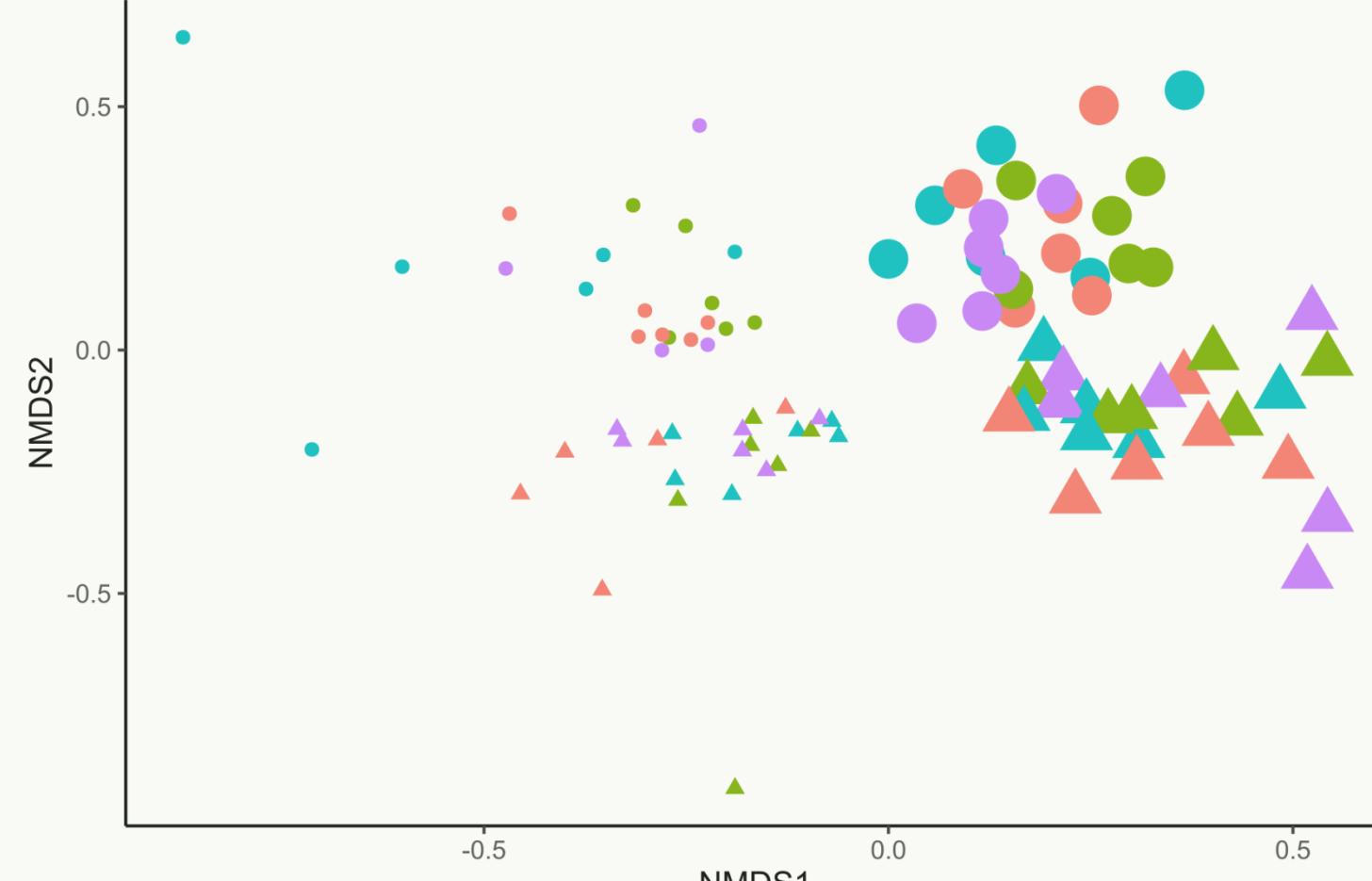


Figure 2: NMDS ordination of microbial communities. Data was normalized with Cumulative Sum Scaling. There was a very clear effect of plant species and sample type (Endosphere or rhizosphere) but treatment effects were subtle. We proceed the analysis by partitioning the data in 4 sets, so that plant species and sample type effects do not mask treatment effects.

Material and methods

We exposed *Arabidopsis thaliana* and *Brassica oleracea* to Methyl Jasmonate, a mimic of herbivory stress (Figure 1). We then performed 16S sequencing of plants roots and rhizospheric soil (figure 2). The microbiome dataset was partitioned and processed in R with Boruta for random forest, SPIEC-EASY for network analysis, and Deseq2 for differential abundance. This generated a list of 346 ASVs that were “tagged” as “important” by these different methods. A Fisher test was then used to compare the proportions of every taxonomic group in this list against the proportions of the same taxonomic group in the rest of the dataset. The output from the Fisher tests is summarized in a heat tree.

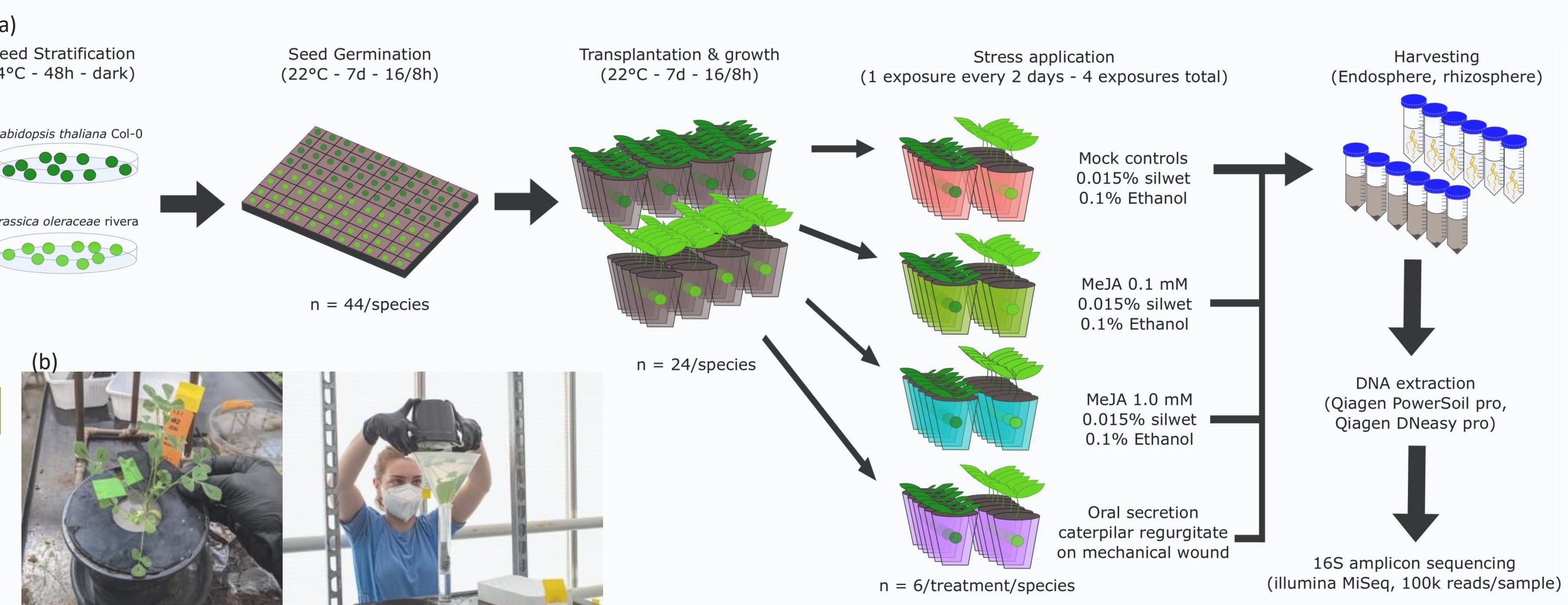


Figure 1: Scheme of the experimental design (a); representation of the MeJA dipping process (b).

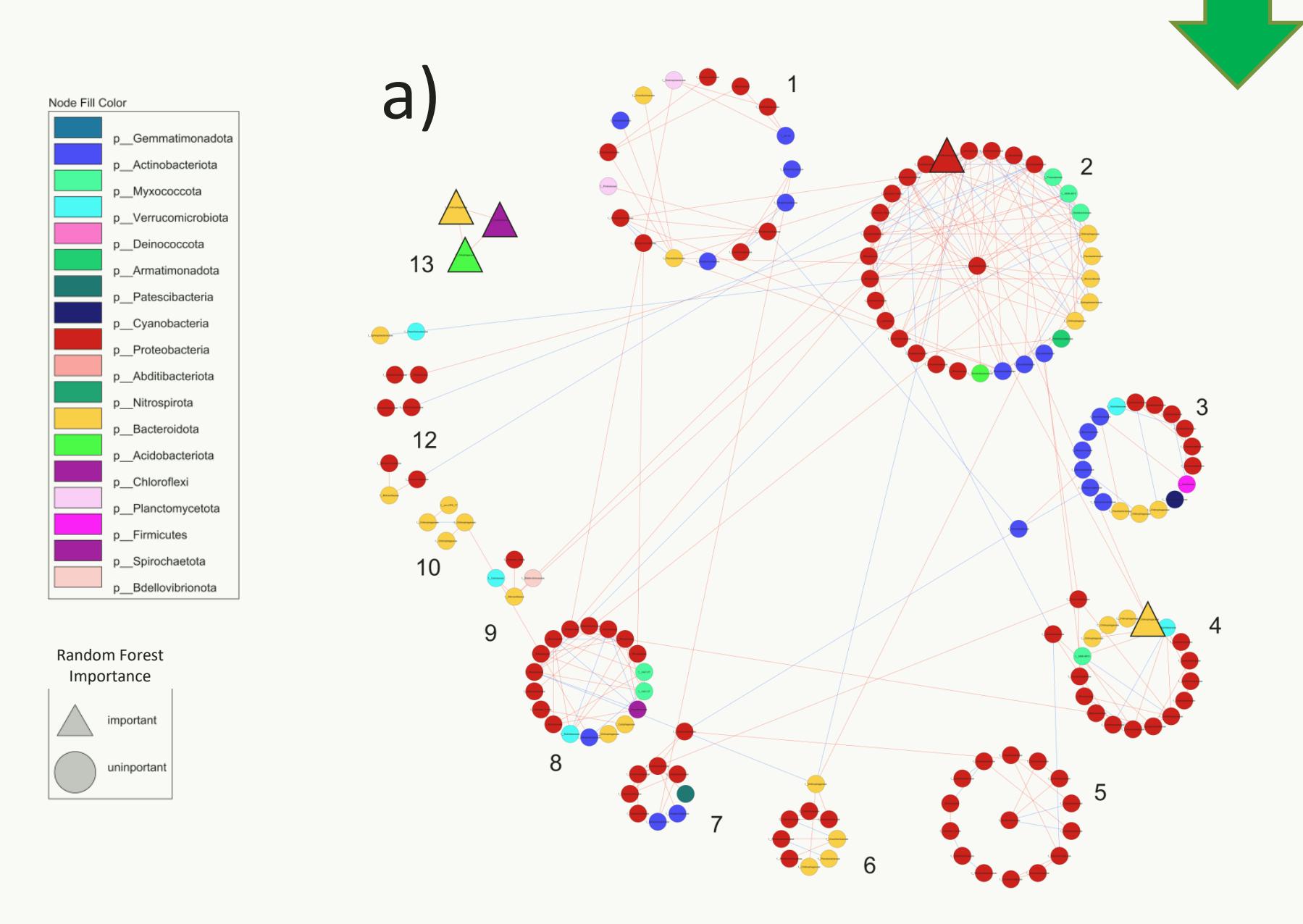


Figure 3: Co-variance networks of *A. thaliana* endophytic (a) and rhizospheric (b) communities; *B. oleracea* endophytic (c) and rhizospheric (d) communities. Groups of nodes represent modules. Red edges represent positive co-variance, and blue edges represent negative co-variance. Nodes in the center of modules are module hubs ($Z_i > 2.5$), nodes outside modules are module connectors ($p > 0.62$). Keystone taxa (significantly high z-transformed degree, betweenness centrality, and closeness centrality) have thicker edges. Triangles represent ASVs important in random forest, diamonds represent keystone taxa.

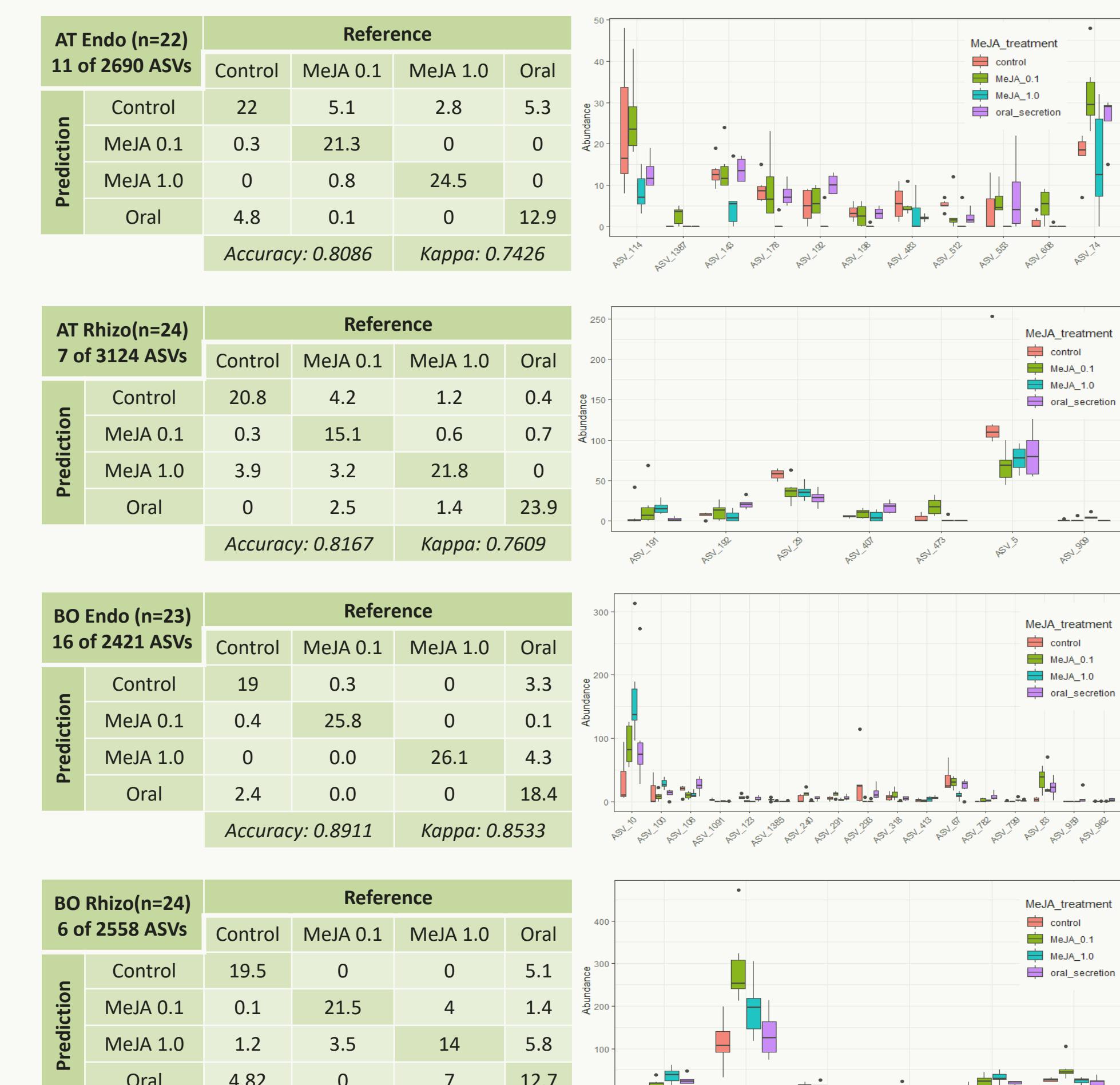


Figure 4: Random forest output showing the confusion matrix (table to the left) and the abundance of ASVs selected as predictors for stress treatment in each data partition (right). AT = *Arabidopsis thaliana*, BO = *Brassica oleracea*, Endo = endosphere samples, Rhizo = rhizosphere samples.

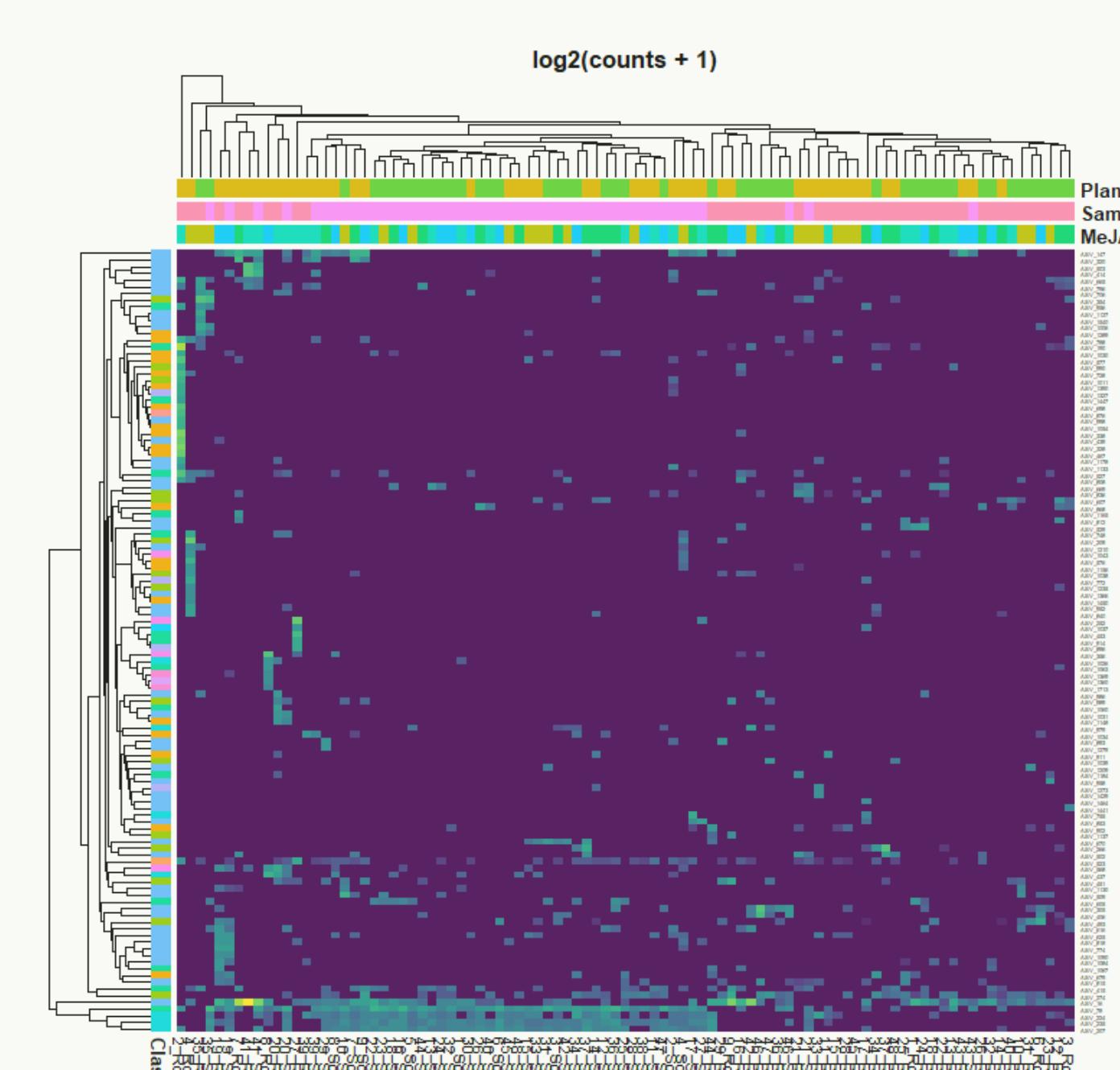


Figure 5: Biplot heatmap representing differential abundance analysis of ASVs in each treatment against their respective controls. ASVs are represented in rows, samples are represented in columns, cell colors represent log-transformed abundances.

B.oleracea soil	Important ASV	Other ASV	Total (R)
Mucilaginibacter ASV	5	57	62
ASV in other genus	48	2442	2490
Total (C)	53	2499	2557

Table 1: Example of proportions compared in fisher tests (Mucilaginibacter in rhizosphere of *B. oleracea*), used to define node color and size in figure 6.

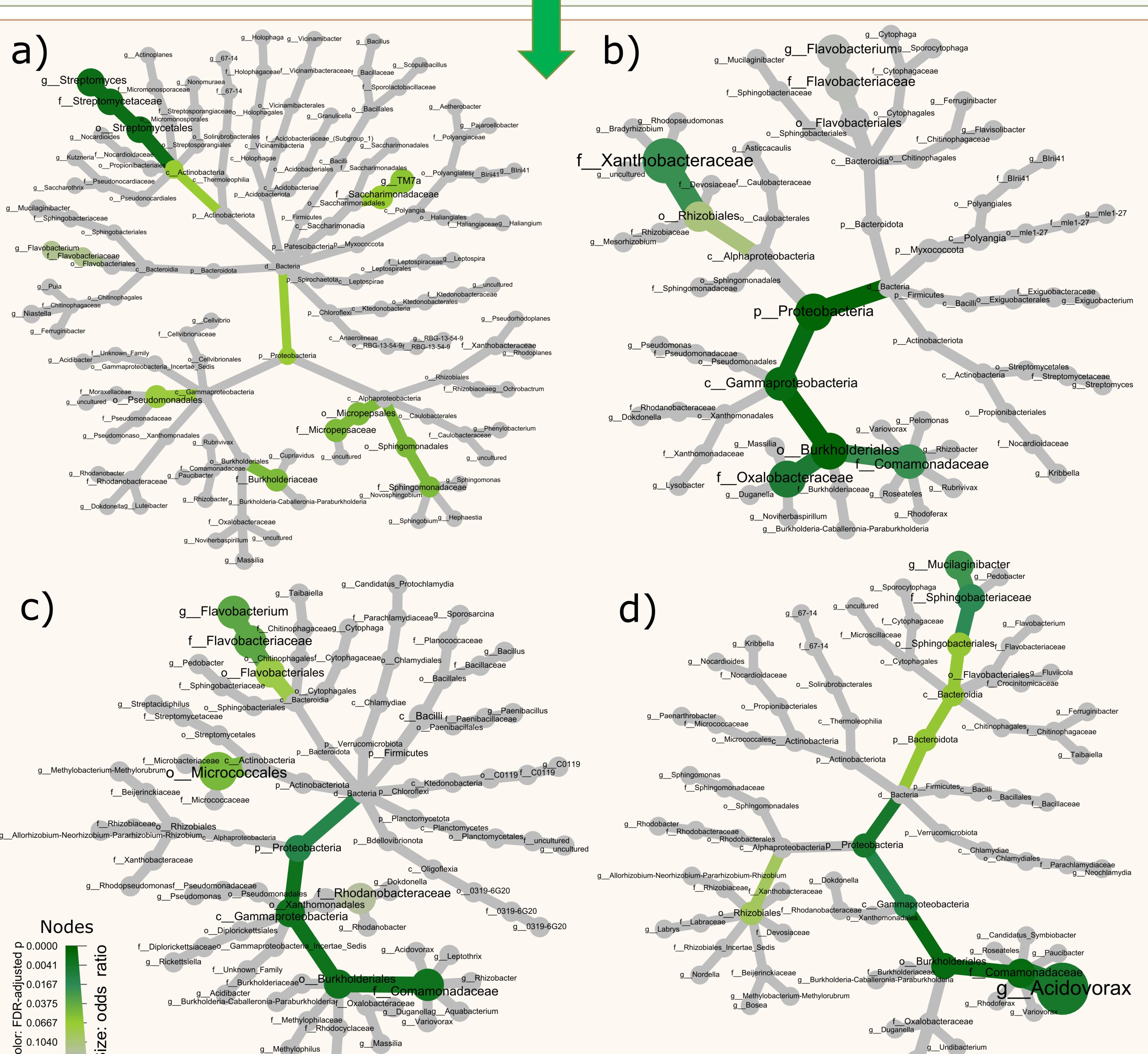


Figure 6: Summary of “important” ASVs. Proportions in the occurrence of ASVs in each taxon level are tested with Fisher test as in Table 1. An ASV is “important” if highlighted by random forest, network analysis, or differential abundance. Node colors indicate adjusted p values, node size indicates odds ratio with a maximum of: 12.9 in *A. thaliana* endosphere (a), 5.8 in *A. thaliana* rhizosphere (b); 21.3 in *B. oleracea* endosphere (c); and 6.1 in *B. oleracea* rhizosphere (d).

Conclusions

The summarization allows us to, in a glance, detect which microbial taxa are significantly more represented as “important”, having the full microbial community as a context. Noticeably, Family *Comamonadaceae* is summarized as significantly more represented in 3 out of 4 data partitions. This diverse family is not only common in plant endosphere and rhizosphere, but is also common in the gut of herbivorous insects and other organisms that survive on poor diets. It is possible that members of this family use both plant and herbivore as alternative hosts to increase its niche breadth and dispersal, similarly to *Enterobacteriaceae* that also thrive in both plant roots and mammal gut. While this analysis cannot tell if *Comamonadaceae* are beneficial or detrimental to the plant, it allows us to focus our attention for literature mining, further analysis, and new experiments in 1 out of the 1.111 taxonomic groups detected by 16S amplicon sequencing.