Multivariate Cutoff Level Analysis (MultiCoLA) of Large Community Datasets

Angélique Gobet, Christopher Quince, Alban Ramette

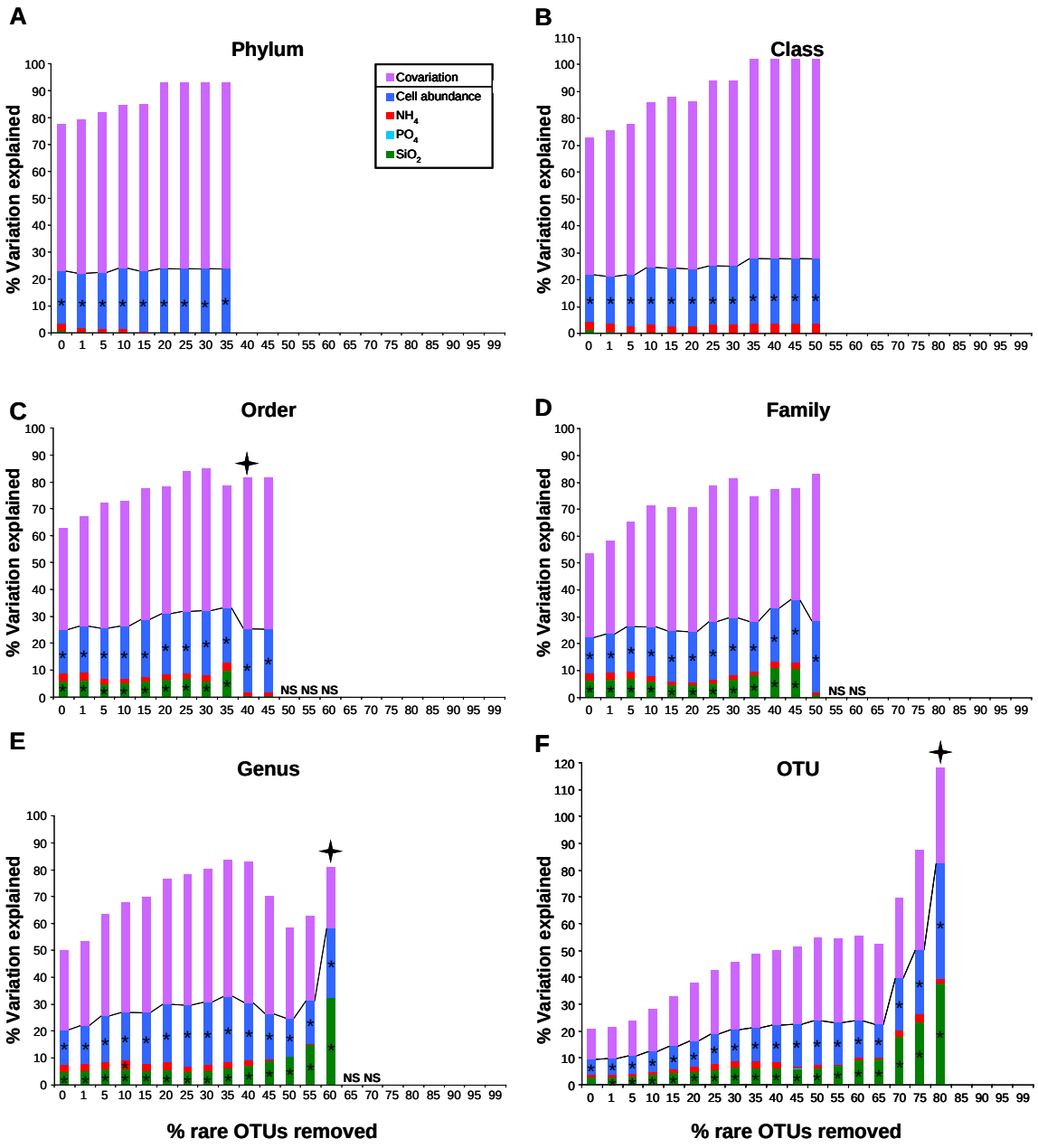
Supplementary Figures:

Supplementary Figure 1. MultiCoLA profiles of biological variation with the dataset-based cutoff approach.

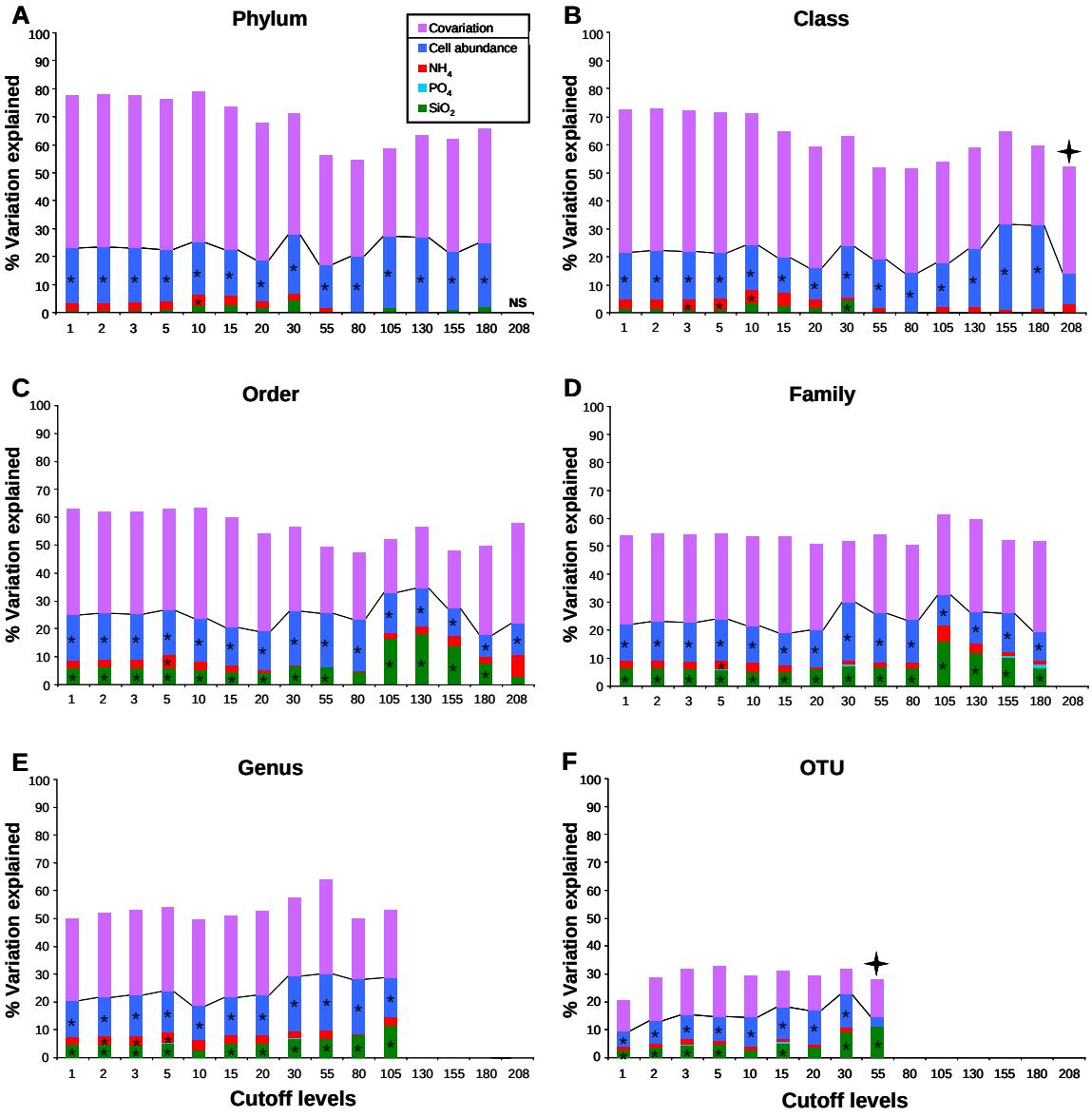
Supplementary Figure 2. MultiCoLA profiles of biological variation with the sample-based cutoff approach.

Supplementary Figure 3. MultiCoLA profiles of biological variation with the dataset-based and sample-based approaches on PyroNoise corrected data.

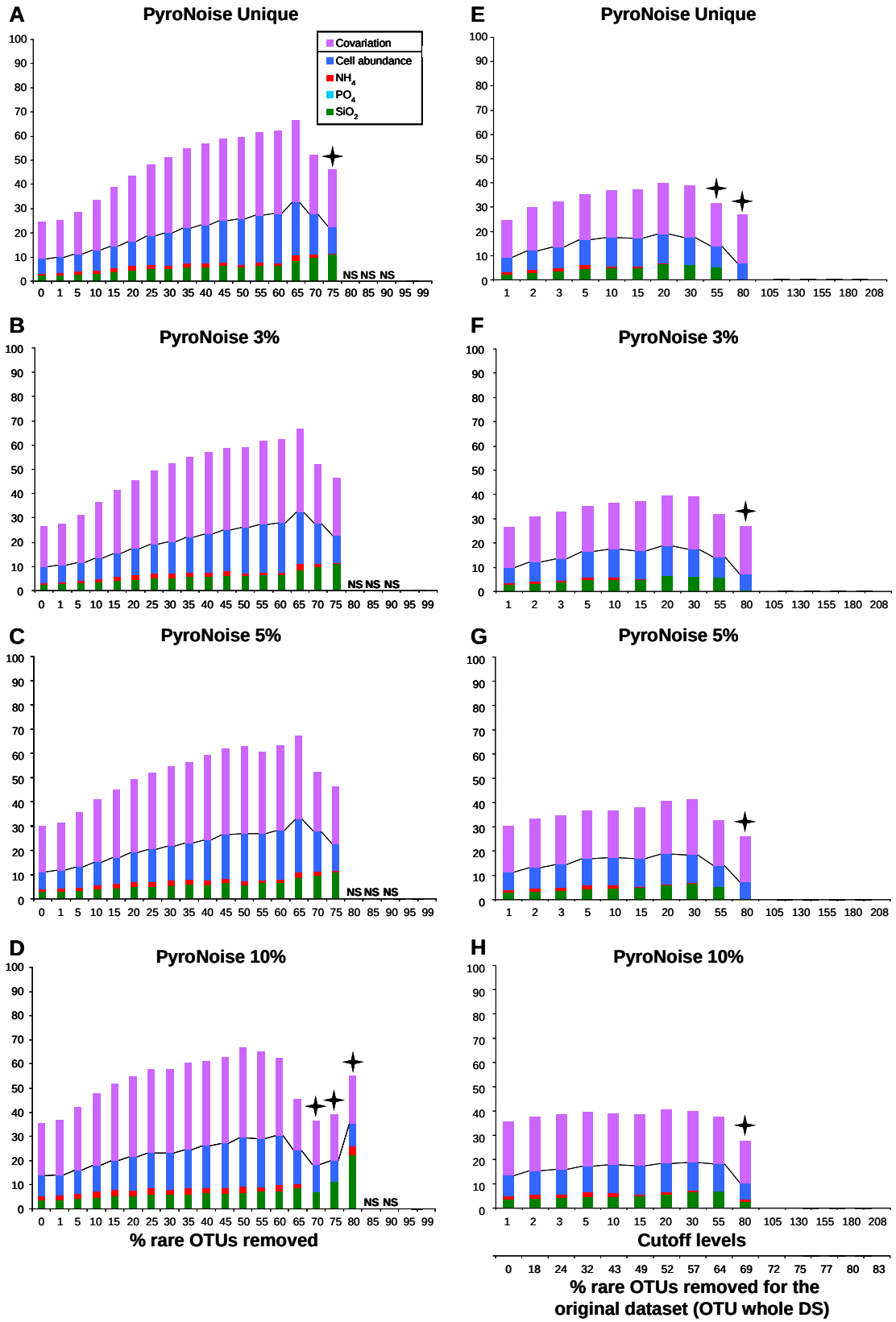
Supplementary Figure 4. MultiCoLA profiles based on the dataset- (A, B, C) and sample-based (D, E, F) cutoff approaches only retaining the rare OTUs in each truncated dataset. (**A**, **D**) Abundance of rare OTUs in each truncated dataset at the phylum, class, order, family, genus and OTU levels. A black solid line indicates comparisons at the OTU level for the dataset with a complete annotation and a black dashed line indicates the OTU level for the whole dataset (OTU whole DS). (**B**, **E**) Non-parametric Spearman correlations comparing the deviation in complete data structure between the original matrix and truncated matrices. (**C**, **F**) Comparison of most important axes of extracted variation between the original and truncated datasets. Lacking points are due to sample loss by applying a given cutoff to the original dataset. In the panels **D**, **E**, **F**, the upper x-axis corresponds to cutoff levels defined as a function of the sample-based approach, and the lower x-axis represents the corresponding proportion of removed sequences in the OTU dataset (all OTUs). This enables the comparison of the dataset-based approach with the sample-based approach. ODS, original dataset.



Supplementary Figure 1. MultiCoLA profiles of biological variation with the dataset-based cutoff approach.. Partitioning of the biological variation at the (**A**) phylum, (**B**) class, (**C**) order, (**D**) family, (**E**) genus and (**F**) OTU levels for the dataset with a complete annotation, into the respective effects of environmental factors (nutrients and cell abundance). Negative values, unexplained variation and non-significant models are not shown. SiO_2, silicate; PO_4, phosphate; NH_4, ammonium; Covariation of any of the 4 environmental factors is represented under the same category. A star indicates a significant effect of the pure factors (P<5%), whereas "NS" indicates non-significant models. A cross indicates non-significant Bonferroni corrected models. Absence of data (lacking bar) is due to sample loss by applying a given cutoff to the original dataset.

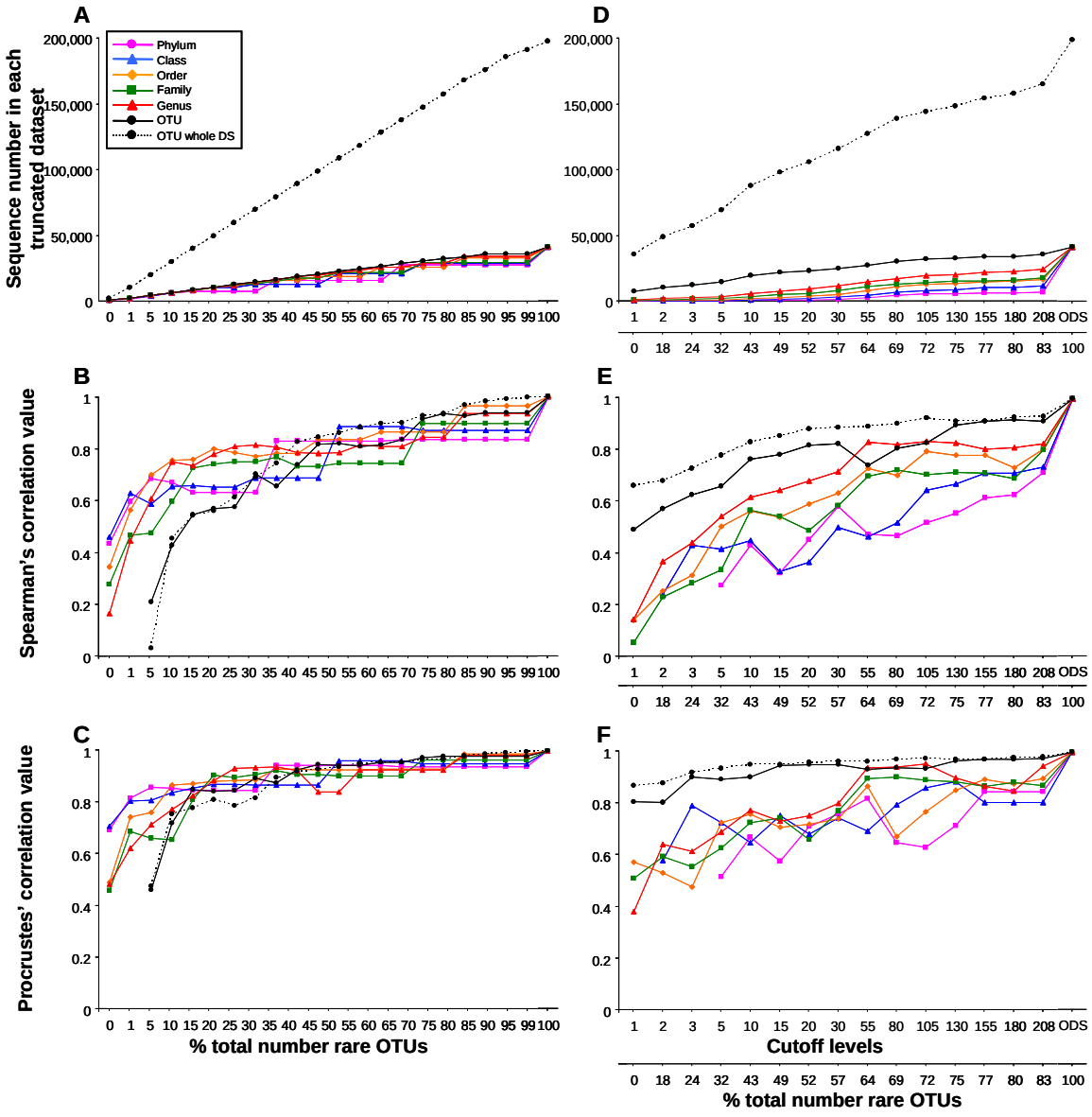


Supplementary Figure 2. MultiCoLA profiles of biological variation with the sample-based cutoff approach. See Supplementary Figure 1 for details.



Supplementary Figure 3. MultiCoLA profiles of biological variation with the dataset-based (A, B, C, D) and sample-based (E, F, G, H) approaches for PyroNoise

corrected data. For the sample-based approach panel **H**, the upper x-axis corresponds to cutoff levels defined as a function of the sample-based approach (as for panels **E**, **F**, **G**), and the lower x-axis represents the corresponding proportion of removed sequences in the OTU dataset (all OTUs). This enables the comparison of the sample-based with dataset-based approach. Each panel consists of PyroNoise-corrected datasets whose sequences were clustered at various sequence dissimilarity levels (0-10%). See Supplementary Figure 1 for further details.



Supplementary Figure 4. MultiCoLA profiles based on the dataset- (A, B, C) and sample-based (D, E, F) cutoff approaches only retaining the rare OTUs in each truncated dataset. (**A**, **D**) Abundance of rare OTUs in each truncated dataset at the phylum, class, order, family, genus and OTU levels. A black solid line indicates comparisons at the OTU level for the dataset with a complete annotation and a black dashed line indicates the OTU level for the whole dataset (OTU whole DS). (**B**, **E**) Non-parametric Spearman correlations comparing the deviation in complete data structure between the original matrix and truncated matrices. (**C**, **F**) Comparison of most important axes of extracted variation between the original and truncated datasets. Lacking points are due to sample loss by applying a given cutoff to the original dataset. In the panels **D**, **E**, **F**, the upper x-axis corresponds to cutoff levels defined as a function of the sample-based approach, and the lower x-axis represents the corresponding proportion of removed sequences in the OTU dataset (all OTUs). This enables the comparison of the dataset-based approach with the sample-based approach. ODS, original dataset.