Impact of tree choice in metagenomics differential abundance studies

Antoine Bichat[1, 2, 3], Christophe Ambroise[2], Jonathan Plassais[1], and Mahendra Mariadassou[3]

[1]Entérome, Paris, France
[2]LaMME, Université Evry Val d'Essonne, Evry, France
[3]INRA - MaIAGE, Université Paris Saclay, Jouy-en-Josas, France

We consider the problem of incorporating evolutionary information (e.g. taxonomic or phylogenic trees) in the context of metagenomics differential analysis. Recent results published in the literature propose different ways to leverage the tree structure to increase the detection rate of differentially abundant taxa. Here, we propose instead to use a different hierachical structure, in the form of a correlation-based tree, as it may capture the structure of the data better than the phylogeny. We first show that the correlation tree and the phylogeny are significantly different before turning to the impact of tree choice on detection rates. Using synthetic data, we show that the tree does have an impact: smoothing p-values according to the phylogeny leads to equal or inferior rates as smoothing according to the correlation tree. However, both trees are outperformed by the classical, non hierachical, Benjamini-Hochberg (BH) procedure in terms of detection rates. Other procedures may use the hierachical structure with profit but do not control the False Discovery Rate (FDR) a priori and remain inferior to a classical Benjamini-Hochberg procedure with the same nominal FDR. On real datasets, no hierarchical procedure had significantly higher detection rate that BH. Although intuition advocates the use of a hierachical structure, be it the phylogeny or the correlation tree, to increase the detection rate in microbiome studies, current hierachical procedures are still inferior to non hierachical ones and effective procedures remain to be invented.