Predicting the HMA-LMA Status in Marine Sponges by Machine Learning

Moitinho-Silva, L.[1,2]; Steinert, G.[3]; Nielsen, S.[1,2]; Hardoim, C.C.P.[4]; Wu Y-C[5,6]; McCormack, G.P.[7]; López-Legentil, S.[8]; Marchant, R.[9]; Webster, N.[10,11]; Thomas, T.[1,2] and Hentschel, U[5,6]

[1]Centre for Marine Bio-Innovation, UNSW, Sydney, Australia; [2]School of Biological, Earth and Environmental Sciences, UNSW, Sydney, Australia; [3]Laboratory of Microbiology, WUR, Wageningen, NL; [4]Laboratório de Interação Hospedeiro-Microbiota, UNESP-CLP, São Paulo, Brazil; [5]RD3 Marine Microbiology, GEOMAR Helmholtz Centre for Ocean Research Kiel; [6]Christian-Albrechts Univ., Kiel, Germany; [7]Zoology, Ryan Institute, School of Natural Sciences, National Univ. of Ireland, Galway, Ireland; [8]Dep. of Biology and Marine Biology, and Center for Marine Science, UNCW, Wilmington, NC, USA; [9]Centre for Translational Data Science, School of Information Technologies, Univ. of Sydney, Sydney, Australia; [10]Australian Institute of Marine Science, Townsville, QLD, Australia; [11]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, Univ. of Queensland, St. Lucia, QLD, Australia

The dichotomy between high microbial abundance (HMA) and low microbial abundance (LMA) sponges has been observed in sponge-microbe symbiosis, although the extent of this pattern remains poorly unknown. We characterized the differences between the microbiomes of HMA (n = 19) and LMA (n = 17) sponges (575 specimens) present in the Sponge Microbiome Project. HMA sponges were associated with richer and more diverse microbiomes than LMA sponges, as indicated by the comparison of alpha diversity metrics. Microbial community structures differed between HMA and LMA sponges considering Operational Taxonomic Units (OTU) abundances and across microbial taxonomic levels, from phylum to species. The largest proportion of microbiome variation was explained by the host identity. Several phyla, classes, and OTUs were found differentially abundant in either group, which were considered "HMA indicators" and "LMA indicators." Machine learning algorithms (classifiers) were trained to predict the HMA-LMA status of sponges. Among nine different classifiers, higher performances were achieved by Random Forest trained with phylum and class abundances. Random Forest with optimized parameters predicted the HMA-LMA status of additional 135 sponge species (1,232 specimens) without a priori knowledge. These sponges were grouped in four clusters, from which the largest two were composed of species consistently predicted as HMA (n = 44) and LMA (n = 74). In summary, our analyses shown distinct features of the microbial communities associated with HMA and LMA sponges. The prediction of the HMA-LMA status based on the microbiome profiles of sponges demonstrates the application of machine learning to explore patterns of host-associated microbial communities.