Supporting Figures for: Randomized lasso links microbial taxa with aquatic functional groups inferred from flow cytometry

Peter Rubbens[1*], Marian L. Schmidt[2,3*], Ruben Props[4], Bopaiah A. Biddanda[5], Nico Boon[4], Willem Waegeman[1], and Vincent J. Denef[2]

[1] KERMIT, Department of Data Analysis and Mathematical Modelling, Ghent University, Coupure Links 653, B-9000, Gent, Belgium
[2] Department of Ecology and Evolutionary Biology, University of Michigan, 1105 North University Ave., Ann Arbor, MI 48109, USA
[3] Current address: Department of Integrative Biology, University of Texas at Austin, 2506 Speedway, Austin, Texas 78712, USA
[4] CMET, Center for Microbial Ecology and Technology, Ghent University, Coupure Links 653, B-9000, Gent, Belgium
[5] Annis Water Resources Institute, Grand Valley State University, 740 West Shoreline Drive, Muskegon, MI 49441, USA

* These authors contributed equally.

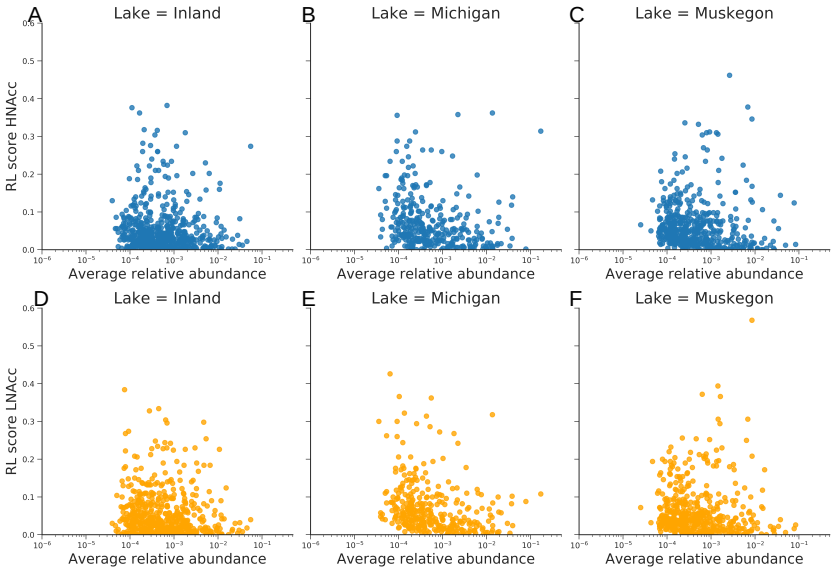


Fig. S1. Scatter plot of RL score versus the average relative abundance of every OTU for HNAcc (blue points, **A**, **B** and **C**) and LNAcc (orange points, **D**, **E** and **F**) for each lake system: Inland (**A** and **D**), Michigan (**B** and **E**) and Muskegon (**C** and **F**).

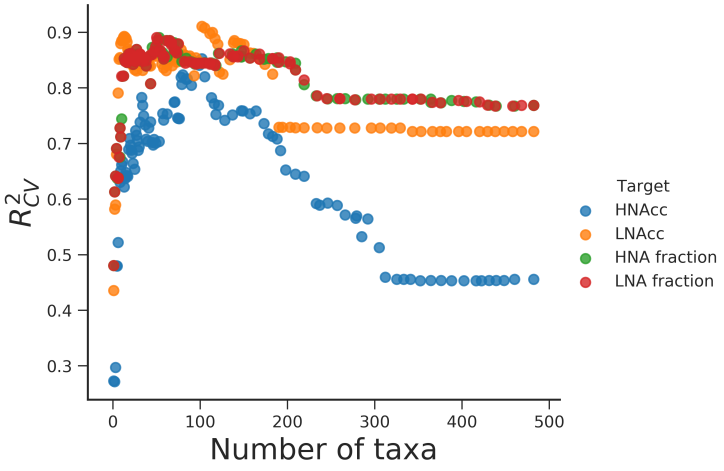


Fig. S2. Comparison of predictions of HNAcc and LNAcc versus relative fractions. This was done for lake Muskegon at the OTU level, expressed in terms of R^2_{CV}. The subset of taxonomic variables was iteratively reduced using a recursive variable elimination strategy, based on the RL score. Lowest-scored variables were removed at every step, after which the base model (i.e., the Lasso) was used to model and predict cell counts or fractions. Predictions for HNA and LNA fractions overlap (red and green dots).

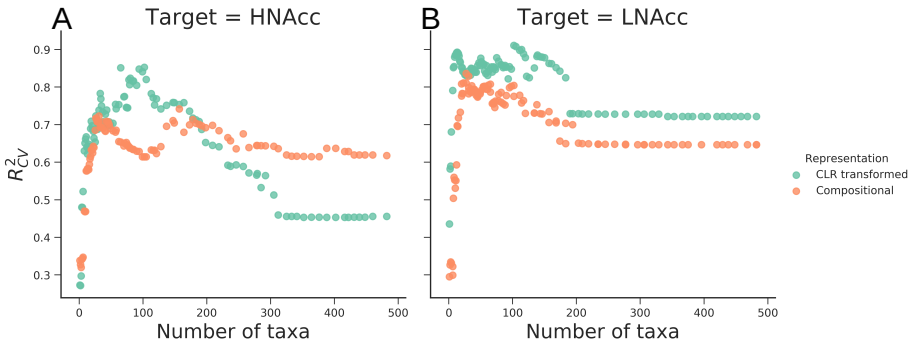


Fig. S3. Prediction of HNAcc (**A**) and LNAcc (**B**) for lake Muskegon at the OTU level, expressed in terms of R^2_{CV} using relative abundances (compositional) and CLR transformed (CLR transformed). The subset of taxonomic variables was iteratively reduced using a recursive variable elimination strategy, based on the RL score. Lowest-scored variables were removed at every step, after which the base model (i.e., the Lasso) was used to model and predict HNAcc and LNAcc.

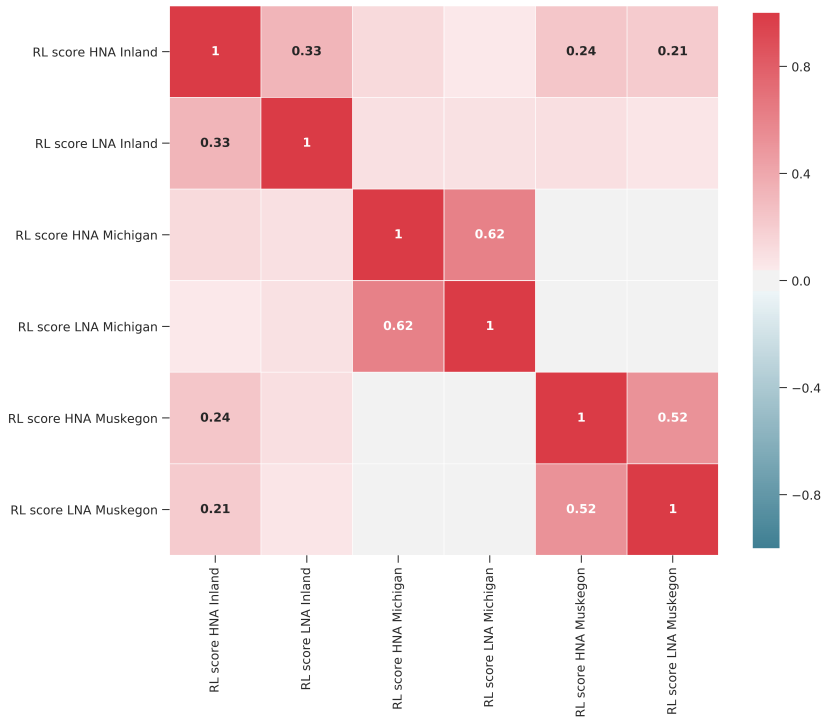
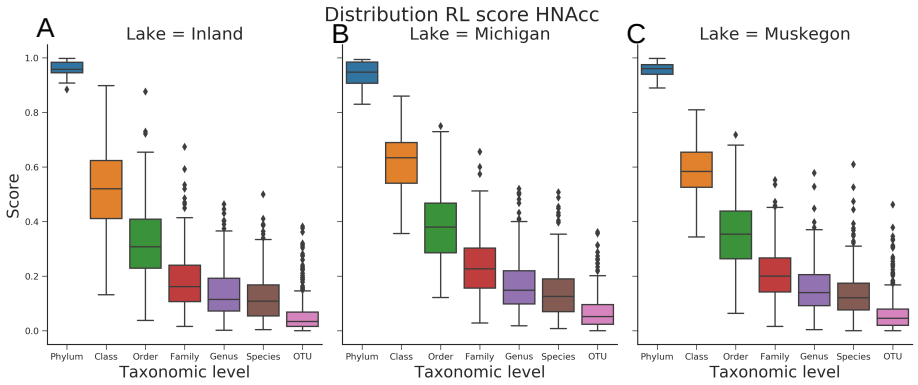


Fig. S4. Pearson correlations between RL scores assigned to OTUs in function of HNAcc and LNAcc between lake systems. Only those OTUs were considered that were present in all lake systems, which were 190 in total. Values are bolded if $P < 0.05$.



D

R^2_{NCV}	Phylum	Order	Genus	OTU
HNAcc	0.47	0.58	0.46	0.49
LNAcc	0.57	0.67	0.62	0.65

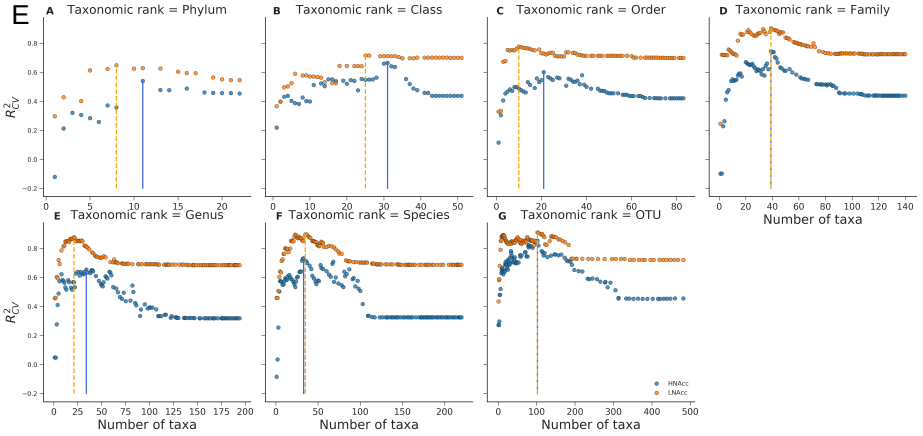


Fig. S5. Distribution of the RL score for all lake systems (**A**: Inland, **B**: Michigan and **C**: Muskegon) and all taxonomic levels in function of HNAcc. **D**: R^2_{NCV} values for HNAcc in Lake Muskegon at the Phylum, Order, Genus and OTU-level. **E**: Evaluation of HNA cell counts (HNAcc) and LNA cell counts (LNAcc) predictions using the Lasso at all taxonomic levels for the Muskegon lake system, expressed in terms of R^2_{CV}, using different subsets of taxonomic variables. Subsets were determined by iteratively eliminating the lowest-ranked taxonomic variables based on the RL score.

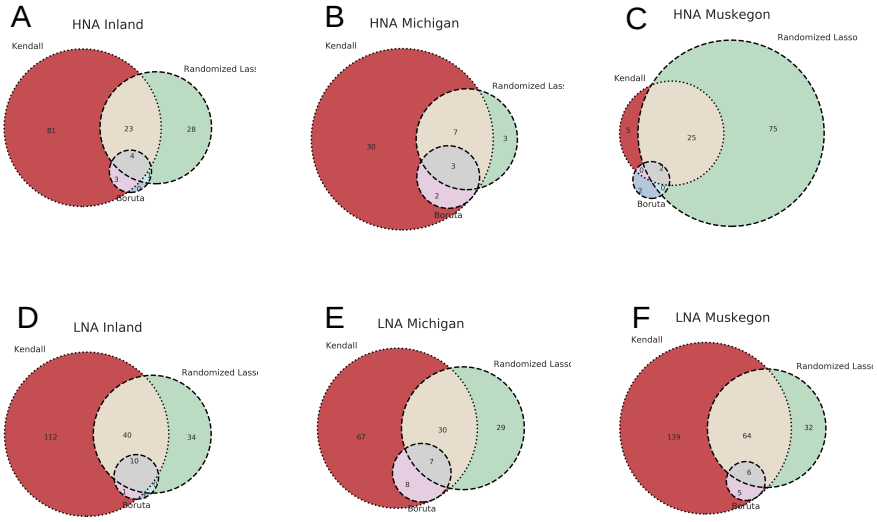


Fig. S6. Venn diagrams for selected OTUs according to the Kendall rank correlation coefficient, RL and Boruta algorithm. OTUs are selected for **A**: HNAcc, Inland; **B**: HNAcc, Michigan; **C**: HNAcc, Muskegon; **D**: LNAcc, Inland; **E**: LNAcc, Michigan; **F**: LNAcc, Muskegon.

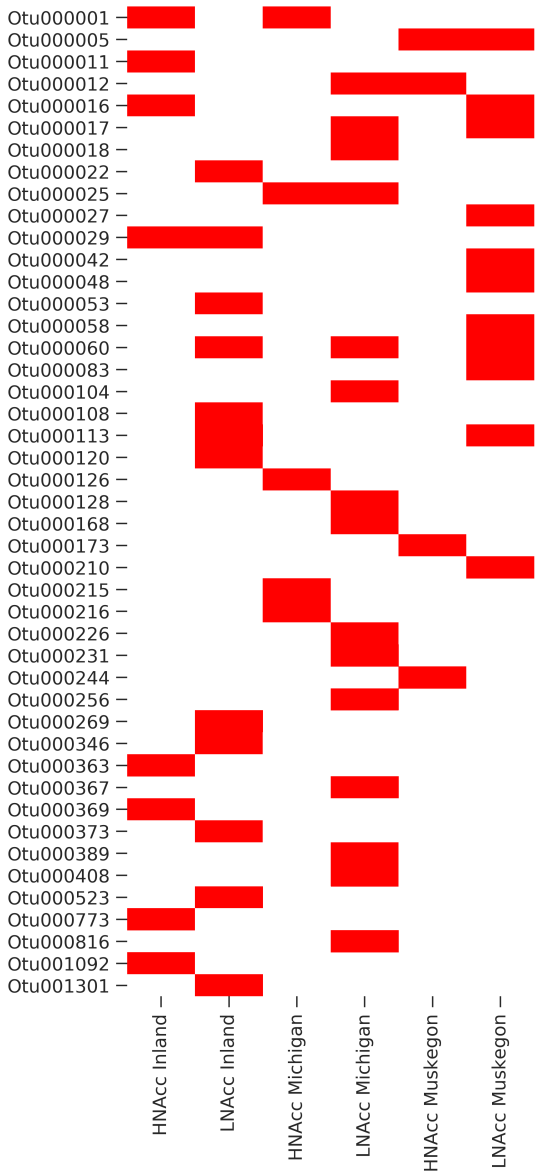


Fig. S7. Selected OTUs (in red) according to the Boruta algorithm for each lake system and functional group

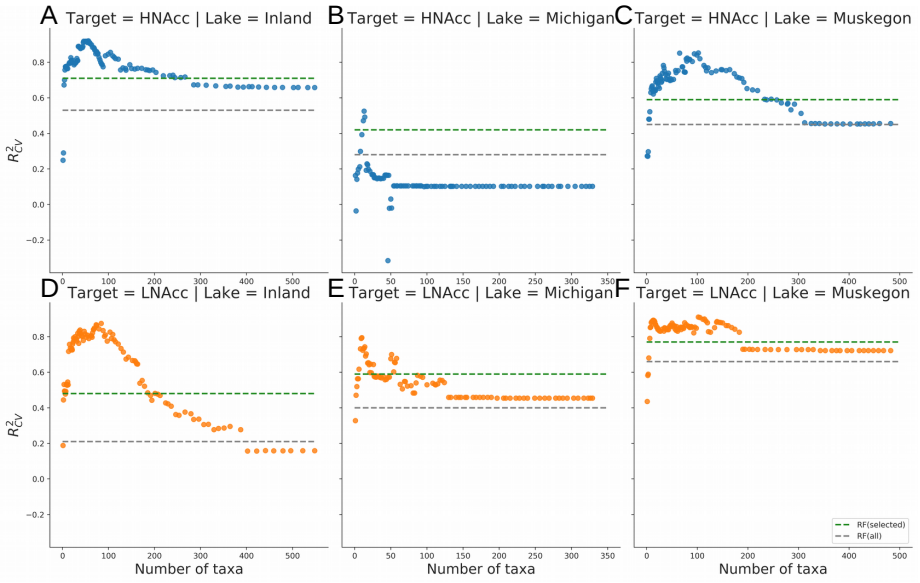


Fig. S8. Comparison of Random Forest predictions using all OTUs (grey dashed line) or selected OTUs (green dashed line) using the Boruta algorithm. This is compared with predictions using the Lasso and RL score at different thresholds, for HNAcc (blue points, **A**, **B** and **C**) and LNAcc (orange points, D, E and F) for each lake system: Inland (**A** and **D**), Michigan (**B** and **E**) and Muskegon (**C** and **F**). Performance is expressed in terms of R^2_{CV}.

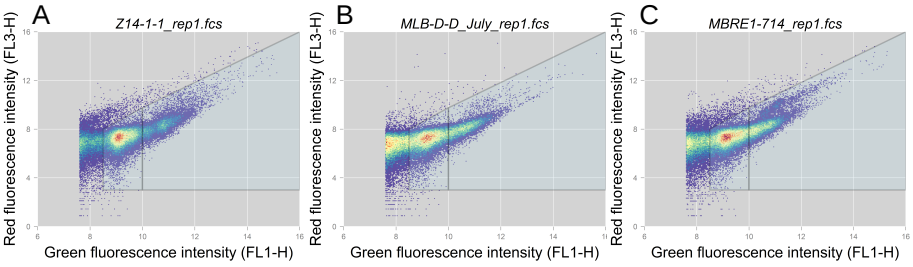


Fig. S9. Examples of the gating strategy for the three lake systems. The same two gates are applied to all samples to determine HNAcc and LNAcc. The gating strategy is performed in the arcsinh(x) transformed bivariate space of the FL1-H and FL3-H channel, following guidelines of Prest et al., 2013. **A**: Inland, **B**: Michigan, **C**: Muskegon.