

¹Karoline Faust*, ²J. Fah Sathirapongsasuti*, ¹Gipsi Lima-Mendez, ³Jacques Izard, ²Nicola Segata, ⁴Dirk Gevers, ²Curtis Huttenhower* and ¹Jeroen Raes*

* These authors contributed equally.

¹Bioinformatics and (Eco-)Systems Biology, VIB-VUB, Pleinlaan 2, B-1050, Brussels, Belgium

²Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, USA

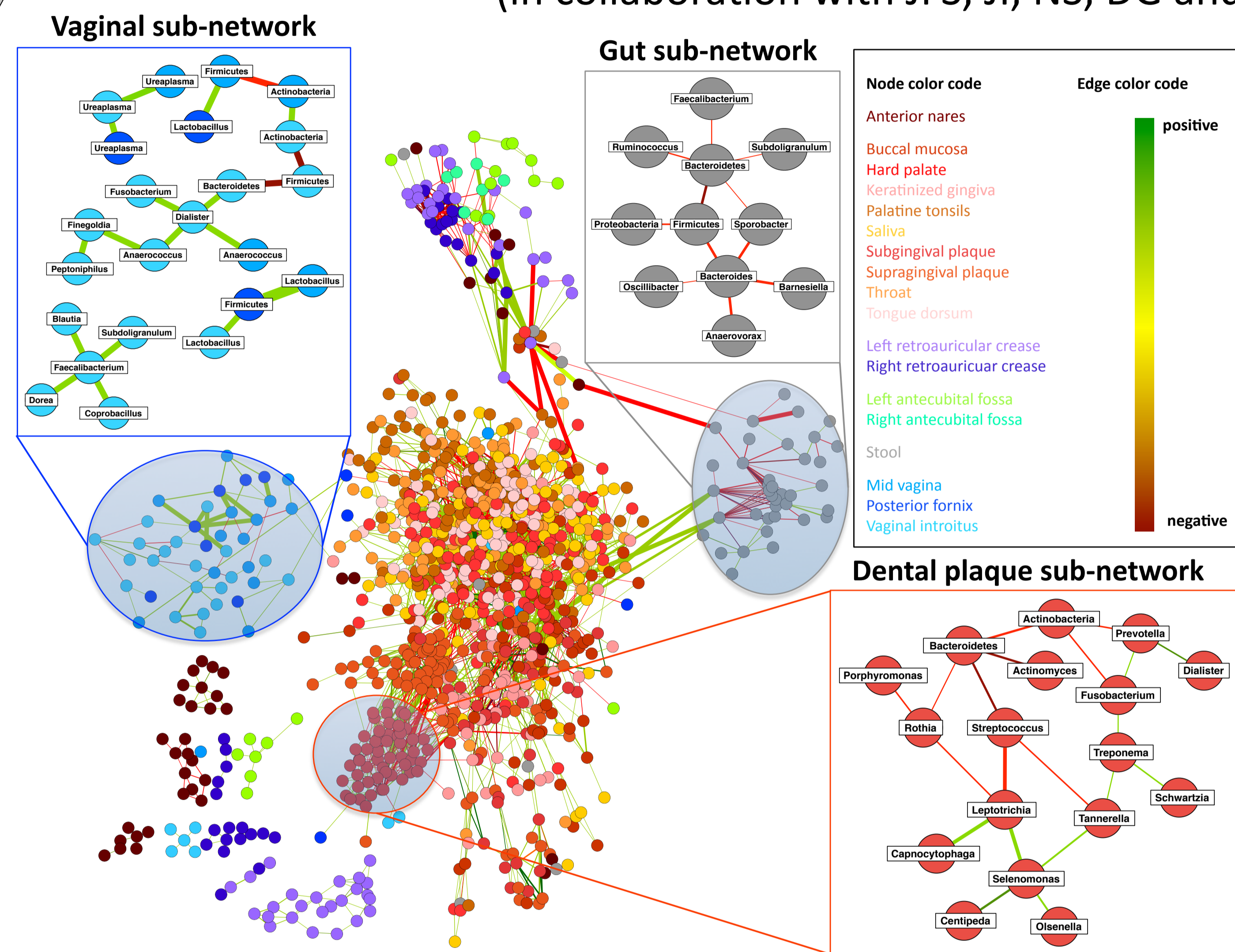
³Department of Molecular Genetics, The Forsyth Institute, 245 First Street, Cambridge, USA

⁴Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, USA

Contact: karoline.faust@vib-vub.be

Metagenomic sequencing projects are accumulating abundance data for microbial organisms in a wide variety of environments, including the human body. In macro-ecological data sets, non-random patterns of species distributions are thought to reflect ecological relationships. For instance, "checkerboard" patterns of mutual exclusion between two taxa are interpreted as competitive relationships [1]. The analysis of microbial abundance data revealed similar non-random patterns for microorganisms [2]. To assess the significance of these relationships, scores of distribution patterns are compared to random expectation. Significant relationships are then combined to form a network. Here, we present microbial networks inferred with different approaches in a variety of metagenomic data sets.

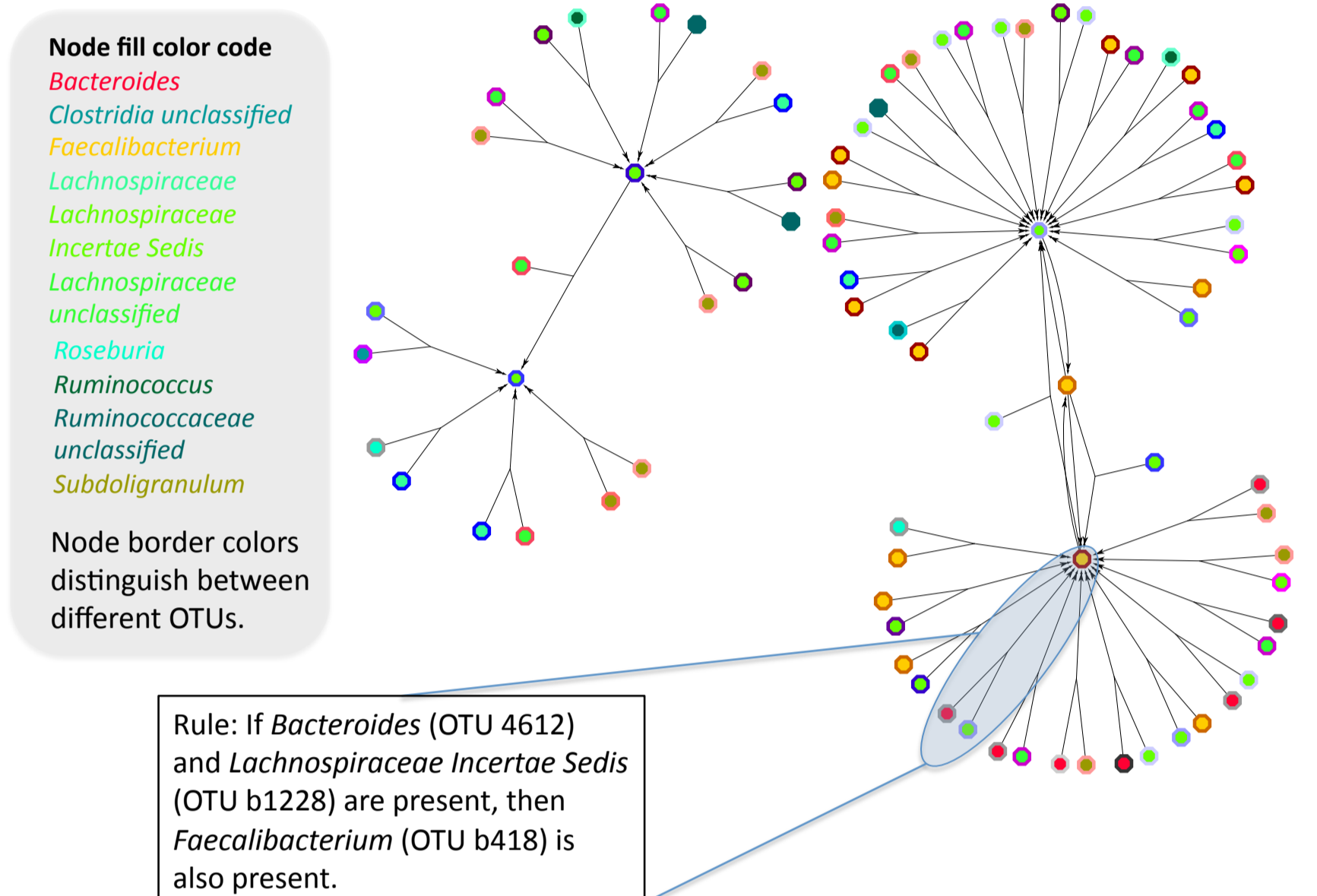
Network of microbial relationships inferred from Human Microbiome Project 16S data (in collaboration with JFS, JI, NS, DG and CH)



Nodes in the predicted network represent phylotypes colored by body site. Edge thickness increases with the significance of the relationship.

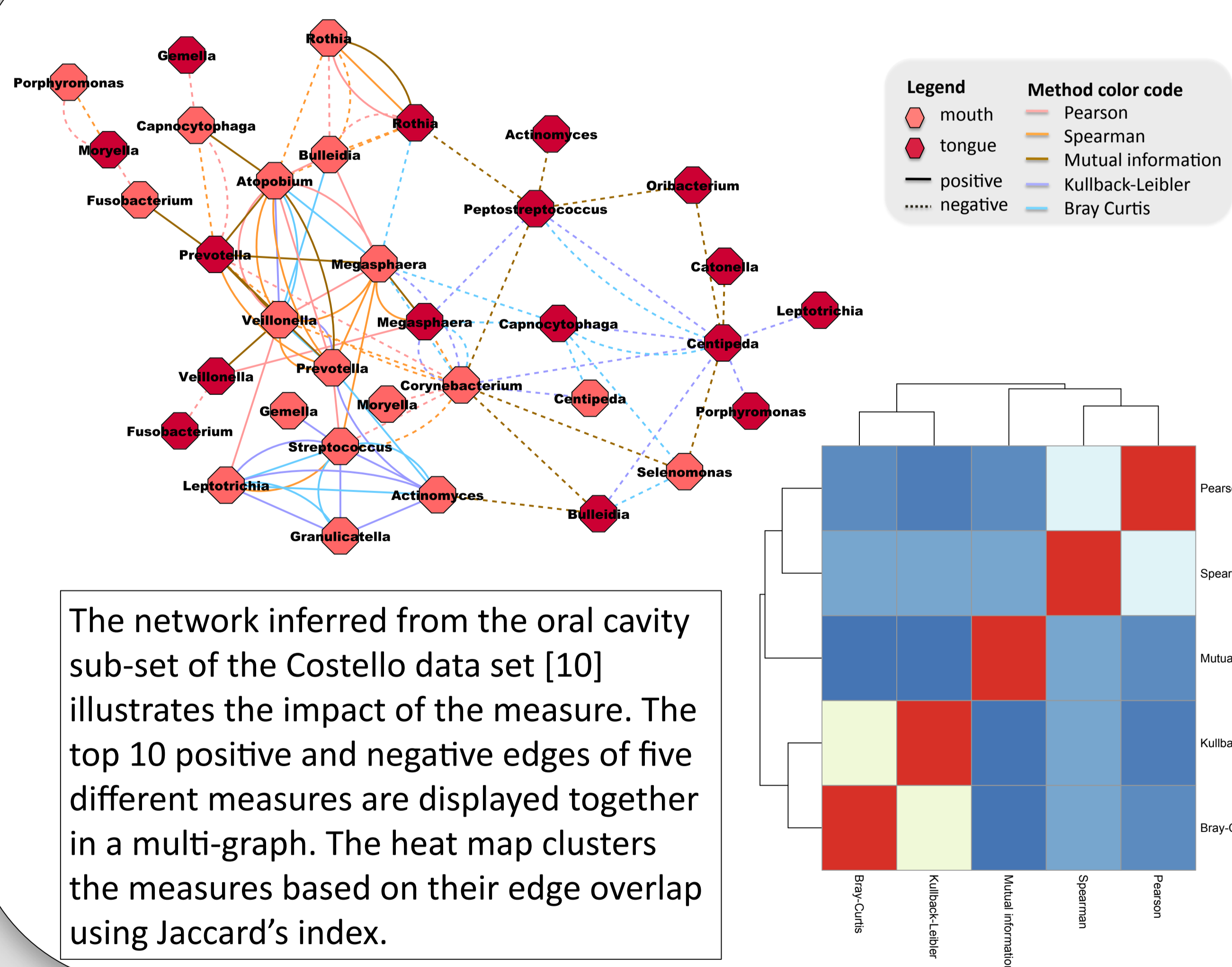
We predict microbial relationships from the HMP 16S data [3] by combining sparse regression with an ensemble of similarity measures using Fisher's method. We assessed the significance of relationships with a novel renormalization-based permutation procedure that we observed in simulations to reduce the impact of compositional bias [4] introduced by normalization. Known alternative community configurations in the vagina [5] and successional stages in the dental plaque [6] were detected as exclusive relationships between members of different communities. In the gut, we found negative relationships between two of the three enterotype drivers [7] (Bacteroides and Ruminococcus).

Association rule mining in presence/absence data



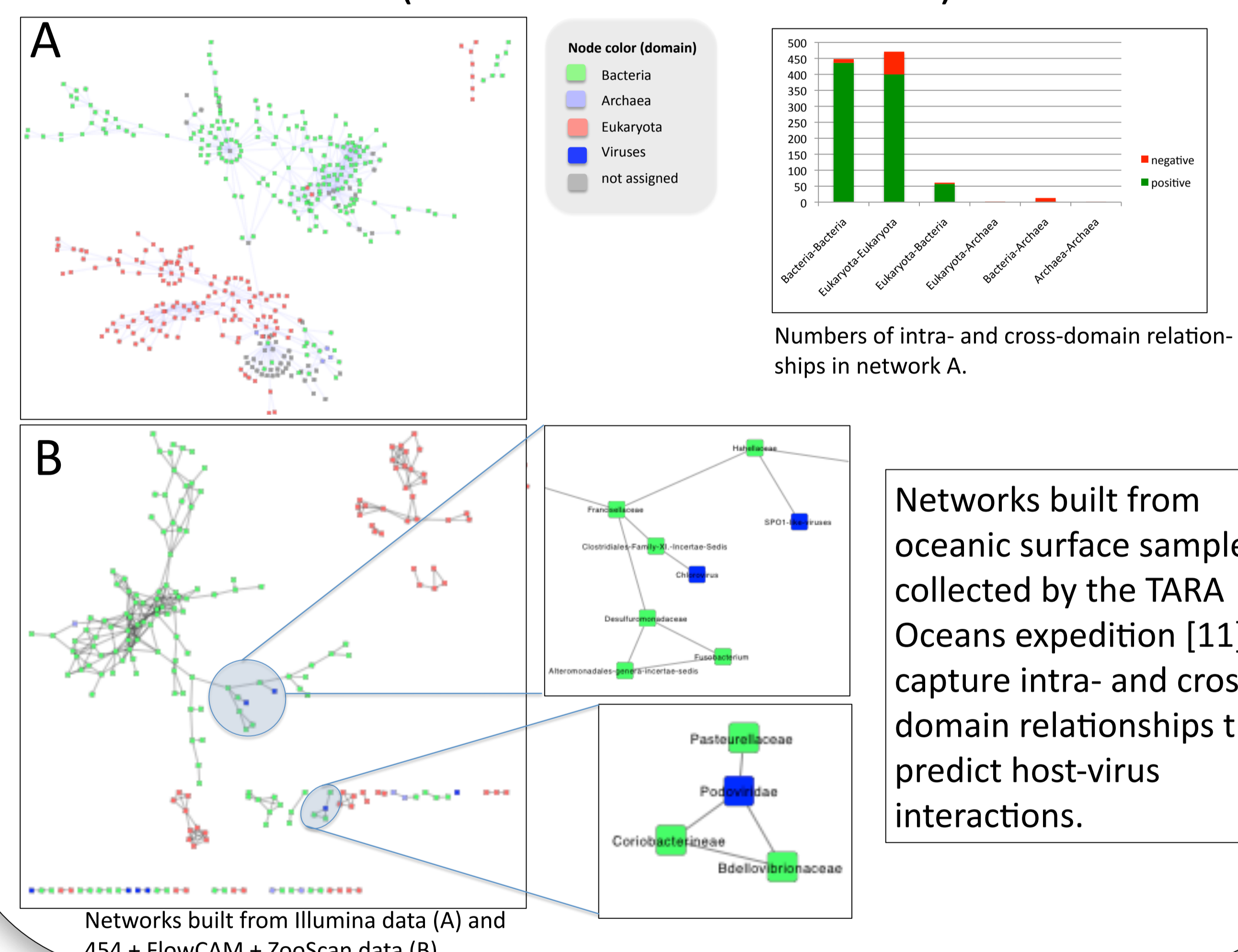
Association rule mining is a technique that detects significant logical rules in large data sets, for instance: "If A is present and B is absent, C is present." The network summarizes rules mined in a global OTU presence/absence data set [8] with the apriori algorithm [9]. Each node in the network represents an OTU, whereas each (hyper-)edge corresponds to a rule. For ease of interpretation, the same OTU (same node fill and border color) may occur multiple times.

Impact of the similarity measure



The network inferred from the oral cavity sub-set of the Costello data set [10] illustrates the impact of the measure. The top 10 positive and negative edges of five different measures are displayed together in a multi-graph. The heat map clusters the measures based on their edge overlap using Jaccard's index.

Prediction of cross-domain relationships (in collaboration with GLM)



Network inference, already widely used in genomics, is applied to more and more metagenomic data sets. However, it faces numerous challenges, such as data sparsity and biases due to normalization. In addition, the choice of the inference methodology (similarity measure, assessment of significance) has a huge impact on the results, necessitating further evaluation and benchmarking. When these issues are addressed, network inference has the potential to become a standard methodology in the analysis of metagenomic data.

References

- [1] Diamond, J.M. "Assembly of species communities" in Ecology and evolution of communities (eds. Cody, M. & Diamond, J.M.) 342-444 (Harvard University Press, 1975).
- [2] Horner-Devine, M.C. et al. "A Comparison Of Taxon Co-Occurrence Patterns For Macro- And Microorganisms" *Ecology* 88, 1345-1353 (2007).
- [3] Turnbaugh, P.J. et al. "The Human Microbiome Project" *Nature* 449, 804-810 (2007).
- [4] Aitchison, J. "A Concise Guide to Compositional Data Analysis" in CDA Workshop Girona (2003).
- [5] Ravel, J. et al. "Vaginal microbiome of reproductive-age women" *PNAS* 108, 4680-4687 (2011).
- [6] Kolenbrander, P.E. et al. "Communication among Oral Bacteria." *MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS* 66, 486-505 (2002).
- [7] Arumugam, M., Raes J. et al. "Enterotypes of the human gut microbiome" *Nature* 473, 174-180 (2011).
- [8] Chaffron, S. et al. "A global network of coexisting microbes from environmental and whole-genome sequence data." *Genome Research* 20, 947-959 (2010).
- [9] Agrawal, R., Imielinski, T. & Swami, A. in ACM SIGMOD Conference (eds. Buneman, P. & Jajodia, S.) 207-216 (ACM Press, 1993).
- [10] Costello, E.K. et al. "Bacterial Community Variation in Human Body Habitats Across Space and Time." *Science* 326, 1694-1697 (2009).
- [11] Karsenti, E. et al. "A Holistic Approach to Marine Eco-Systems Biology" *PLoS Biology* 9, e1001177 (2011).

Acknowledgement

The presenter of this work is a recipient of a travel award funded by the NIH.