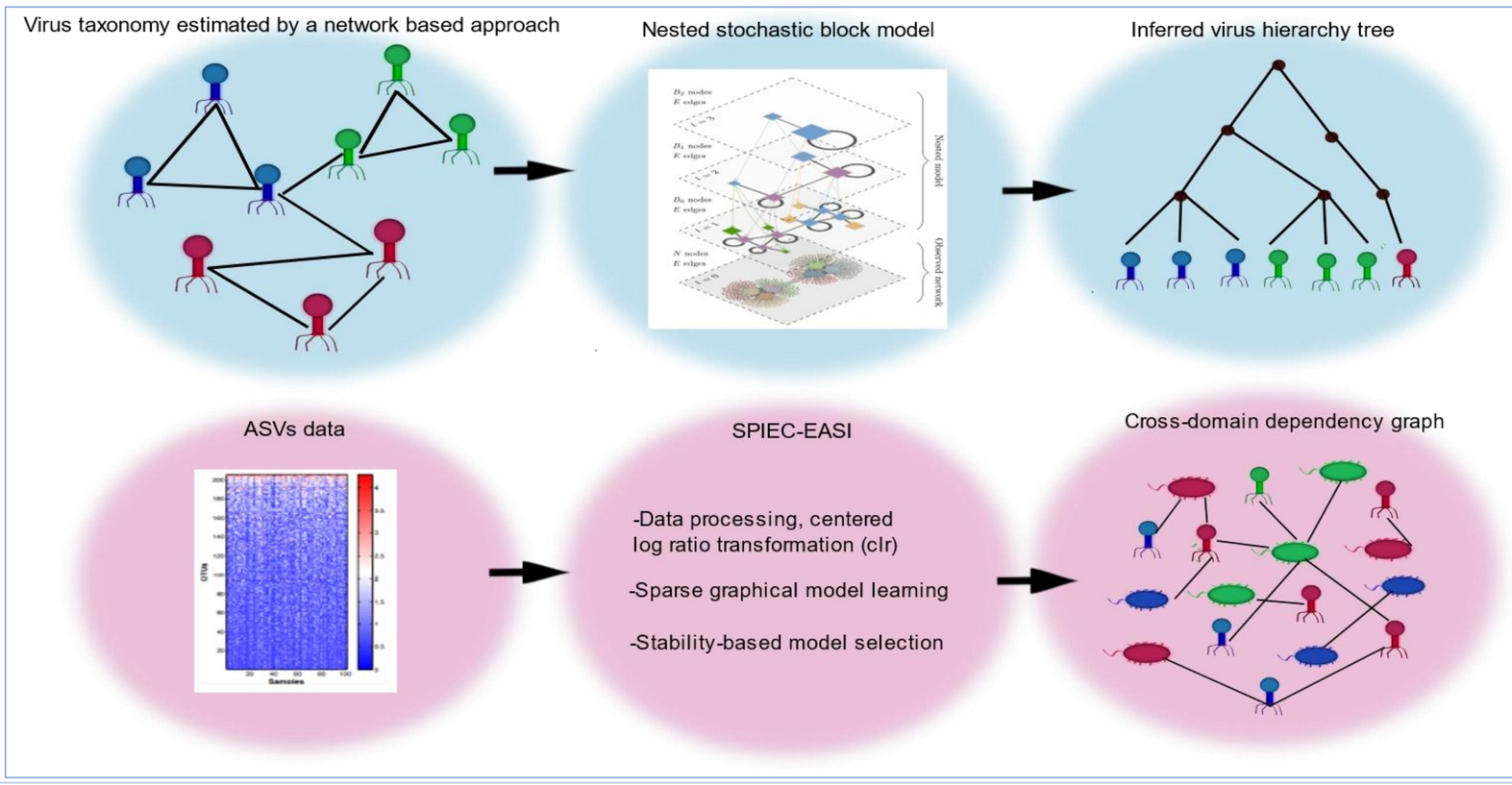
Learning hierarchical phage-bacteria associations from high-throughput sequencing data

Daniele Pugno[1,2], Jinlong Ru[2,4], Jinling Xue[2,4],Li Deng[2,4], Christian L. Müller[1,2,3]

[1]Ludwig-Maximilians-Universität München, Germany; [2]Helmholtz Zentrum München, Neuherberg, Germany [3] Center for Computational Mathematics, Flatiron Institute, New York, USA; [4]Technische Universität München, Germany

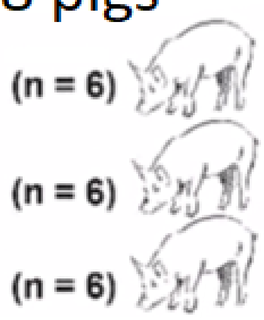
Abstract

The gut microbiome, i.e., the collection of microorganisms and viruses populating the gut, is widely believed to have a considerable impact on human health. The viral component of the human gut microbiome is dominated by bacteriophages (viruses infecting bacteria), which are known to play crucial roles in shaping microbial composition and function. To uncover the potential impact of bacteriophages on bacterial communities in the gut, we develop novel statistical workflows using tools from high-dimensional statistics and network science to infer phage-bacteria associations from high-throughput sequencing data. The core of our workflow uses sparse graphical model estimation to robust partial correlations among the microbial taxa from amplicon and metagenomics abundance data [1]. In addition, we use gene-sharing networks derived from shared protein clusters between viral genomes [2] to learn virus similarity and hierarchy. Here, we use nested stochastic block models (nSBM) [3], a generative model able to detect the hierarchical organization of networks across multiple scales. Combining these data-driven hierarchies with taxonomic and phylogenetic information from bacteria and fungi enables a principled multi-resolution grouping of virus-bacteria-fungi associations. We illustrate our proposed workflow on a large pig gut microbiome dataset where bacterial, fungal, and viral sequencing data are available and present promising initial findings regarding high-level phage-bacteria associations in the pig gut.

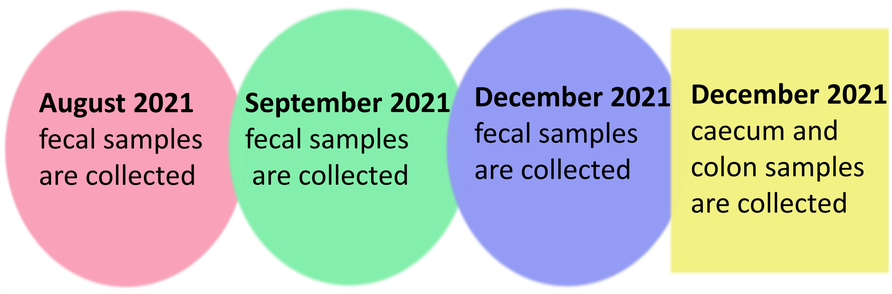


Overview

18 pigs

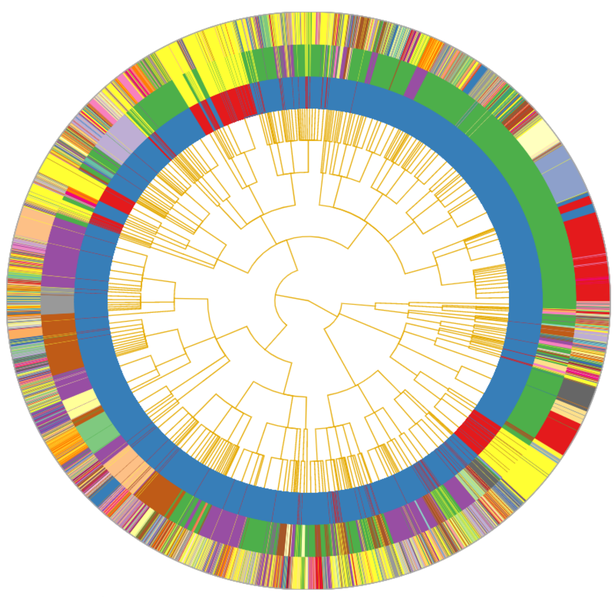


Data collection



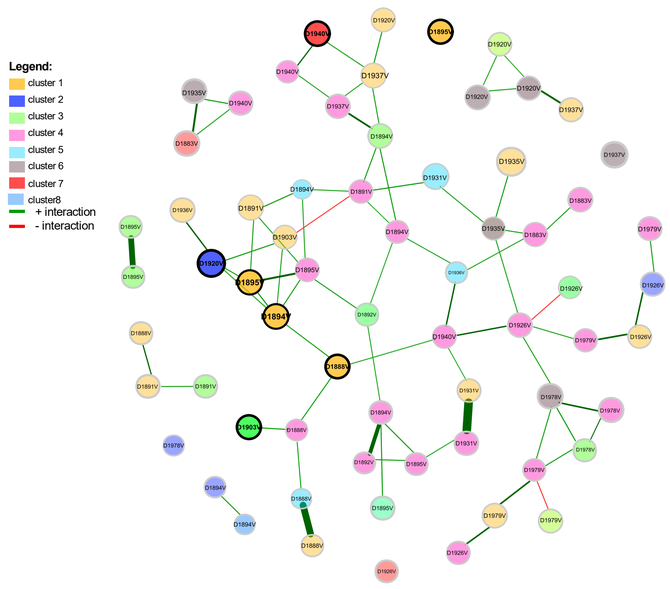
Networks

Estimated virus hierarchy



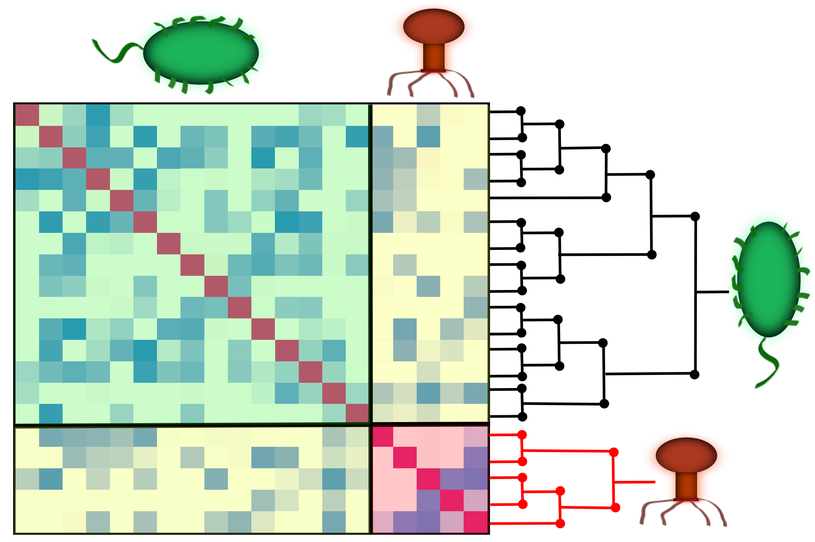
Virus hierarchy estimated from a taxonomy network approach, using nested Stochastic block model. Nested SBM, that is a generative model for graphs organized into communities. The three external layers represent if a virus contig is internal or external in the analysis. The viruses at family lavel. The viruses at genome level.

Virus-virus network



The edges capture partial correlations, that is, the correlation between two variables when controlling for all other variables included in the data set. These association are the non zero elements in the inverse covariance matrix. The network on the left represents the statistical associations between viruses. The network on the right represents statistical associations between viruses and bacteria.

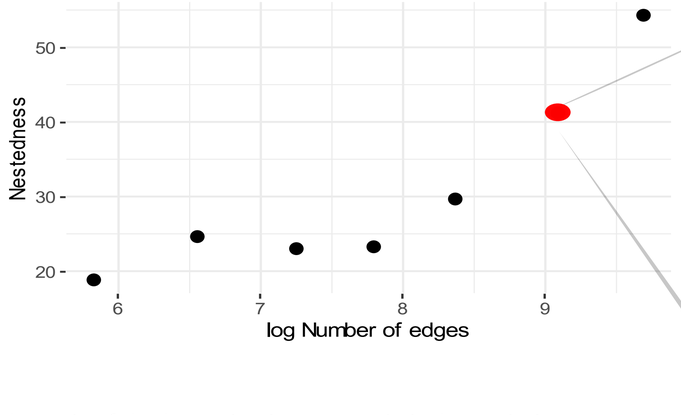
Cross-domain network and hierarchical level

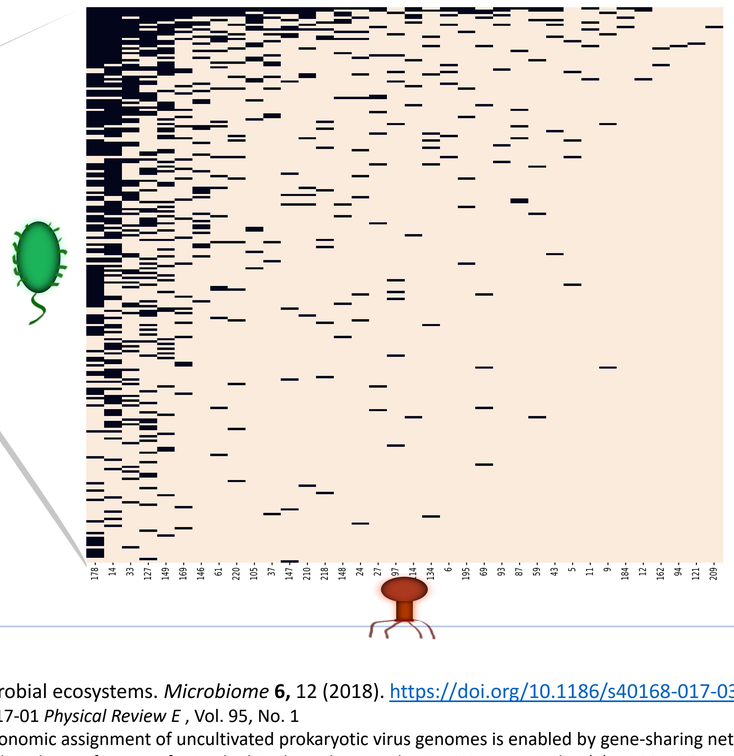


We estimate bacteria-bacteria, virus-virus, and bacteria-virus statistical associations. We observe more connections between viruses that are inferred to have a close taxonomic relationship.

Virus hierarchy and nested matrix

Nestedness vs Sparsity





Nested matrix of the bacteria-viruses stastistical association. Every row represents a different bacteria family. Every column represents a different estimated virus group. In this case nestedness is a measure from ecology.

Nested stochastic block model

$$\mathbf{P}(b|A) = \frac{P(A|\theta,b)P(\theta,b)}{P(A)}$$

- b is the set of partitions
- A is the generative model of thenetwork
- $\boldsymbol{\theta}$ is the set of parameters

Sparse graphical model

$$\widehat{\Theta} = \min_{\Theta \epsilon \boldsymbol{PD}} -\left(\log \boldsymbol{det}(\Theta) + \boldsymbol{tr}\left(\Theta\widehat{\boldsymbol{\Sigma}}\right) + \boldsymbol{\lambda}\|\Theta\|\right)$$

- PD is the space of positive matrices
- $\lambda \geq 0$ is a scalar tuning parameter
- $\widehat{\boldsymbol{\Sigma}}$ is the empirical covariance estimate
- $\|.\|$ is the L1 norm

Literature:

1. Tipton, L., Müller, C.L., Kurtz, Z.D. *et al.* Fungi stabilize connectivity in the lung and skin microbial ecosystems. *Microbiome* **6,** 12 (2018). https://doi.org/10.1186/s40168-017-0393-0
2. Peixoto, Tiago P. Nonparametric Bayesian inference of the microcanonical stochastic block model 2017-01 *Physical Review E* , Vol. 95, No. 1
3. Ho Bin Jang, Benjamin Bolduc, Olivier Zablocki, Jens H. Kuhn, Simon Roux. Nature Biotechnology. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks
4. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA (2015) Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLoS Comput Biol 11(5): e1004226. https://doi.org/10.1371/journal.pcbi.1004226

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

HELMHOLTZ MUNICH

SPP 2330
New Concepts in Prokaryotic Virus-host interactions – From Single Cells to Microbial Communities