System biology

**Link-HD: a versatile framework to explore and integrate heterogeneous microbial communities**

Laura M. Zingaretti\(^1,2,*\), Gilles Renand\(^3\), Diego P. Morgavi\(^4\) and Yulixasis Ramayo-Caldas\(^3,5\)

\(^1\)Plant and Animal Genomics, Statistical and Population Genomics Group, CSIC-IRTA-UAB-UB Consortium, Centre for Research in Agricultural Genomics (CRAG), 08193 Bellaterra, Spain, \(^2\)IAPCBA and IAPCH, UNVM, Villa María, Córdoba 5900, Argentina, \(^3\)URM Animal Genetics and Integrative Biology, GABI, INRA, AgroParisTech, Université Paris-Saclay, 78352 Jouy-en-Josas, France, \(^4\)Animal Physiology and Livestock Systems Divisions, INRA, Herbivore Research Unit, Clermont Auvergne University, Saint Genés-Champanelle 63122, France and \(^5\)Animal Breeding and Genetics Program, IRTA, 08140 Caldes de Montbui, Spain

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on June 19, 2019; revised on October 4, 2019; editorial decision on November 14, 2019; accepted on November 15, 2019

**Abstract**

**Motivation:** We present Link-HD, an approach to integrate multiple datasets. Link-HD is a generalization of ‘Structuration des Tableaux A Trois Indices de la Statistique–Analyse Conjointe de Tableaux’, a family of methods designed to integrate information from heterogeneous data. Here, we extend the classical approach to deal with broader datasets (e.g. compositional data), methods for variable selection and taxon-set enrichment analysis.

**Results:** The methodology is demonstrated by integrating rumen microbial communities from cows for which methane yield \((\text{CH}_4\text{y})\) was individually measured. Our approach reproduces the significant link between rumen microbiota structure and \text{CH}_4 emission. When analyzing the TARA’s ocean data, Link-HD replicates published results, highlighting the relevance of temperature with members of phyla Proteobacteria on the structure and functionality of this ecosystem.

**Contact:** m.lau.zingaretti@gmail.com

**Availability and implementation:** The source code, examples and a complete manual are freely available in GitHub [https://github.com/lauzingaretti/LinkHD](https://github.com/lauzingaretti/LinkHD) and in Bioconductor [https://bioconductor.org/packages/release/bioc/html/LinkHD.html](https://bioconductor.org/packages/release/bioc/html/LinkHD.html).

---

1 **Introduction**

The reduction of ‘omics’ technology costs now enables collection of data from multiple sources. This allows researchers to simultaneously study several datasets and investigate their relationship with complex traits. The integration of these heterogeneous datasets is not trivial and several statistical methods have been developed to address this challenge (Argelaguet et al., 2018; Mariette and Villa-Vialaneix, 2018; Meng et al., 2014). In particular, the amalgamation of multiple microbial ecosystems poses unique challenges as these are compositional and sparse data. MixKernel (Mariette and Villa-Vialaneix, 2018) is a well-known tool designed to integrate heterogeneous datasets including microbial communities, but no method to perform a taxonomic enrichment analysis is available. Another popular integrative approach is MOFA (Argelaguet et al., 2018), however, it is unable to deal with compositional data. Here, we present Link-HD, a tool to integrate and explore multiple microbial communities based on STATIS (Des Plantes, 1976), a family of multivariate methods to integrate multiple datasets. Link-HD generalizes STATIS with Regression Biplot (Ter Braak, 1997), clustering, differential abundance, enrichment taxonomic analysis and visualization tools. Link-HD analyzes distance tables computed from numerical, categorical, or compositional data as a generalization of multidimensional scaling (Abdi et al., 2007). Furthermore, Link-HD performs variable selection and can link the obtained common sub-space with phenotype information.

2 **Materials and methods**

Like STATIS, Link-HD aims to compare and analyze the relationships between datasets with a shared set of observations or variables. However, our package was specifically designed to integrate microbial communities and incorporate distances and transformations to deal with compositional data (Aitchison, 1982). The method is implemented in three main phases (Fig. 1).
replicates the relevant role of temperature and Proteobacteria phyla on the structure of this ecosystem, as described in Mariette and Villa-Vialaneix (2018). Finally, we show the potential of Link-HD to integrate other omics layers by using transcriptome NCI-60 cell lines. Link-HD recapitulates the reported data structure (Meng et al., 2014) and ontology analysis reveals several cancer-related pathways.

In all, our results demonstrate that Link-HD is robust in combining several heterogeneous data types. A detailed description of these case studies and the theory behind Link-HD is available at https://lazingaretti.github.io/LinkHD/ and in Bioconductor (https://bioconductor.org/packages/release/bioc/html/LinkHD.html).

4 Conclusions
We have developed an R package to integrate multiple microbial communities and other ‘omics’ layers combining a plethora of statistical methods in a fast, simple and flexible way.

Acknowledgements
The authors would like to thank Miguel Pérez-Enciso, Rayner Gonzalez-Prendes and Jordi Estelle for valuable discussions and comments on the manuscript and also to APIS-Gène for funding the MicroFicient project.

Funding
L.M.Z. is recipient of a Ph.D. grant from Ministry of Economy and Science, Spain associated with ‘Centro de Excelencia Severo Ochoa 2016–2019’ award SEV-2015-0533 to CRAG. Y.R.C. was funded by Marie Skłodowska-Curie grant (P-Sphere) agreement No 6655919 (EU).

Conflict of Interest: none declared.

References
Kittelmann, S. et al. (2014) Two different bacterial community types are linked with the low-methane emission trait in sheep. PLoS One, 9, e103171.