Koninklijke Nederlandse Vereniging voor Microbiologie

You must choose, but choose wisely:

Model-based approaches for microbial community analysis and data integration

Marcio F. A. Leite

M.Leite@nioo.knaw.nl

KURAMAE GROUP
MICROBIAL COMMUNITY ECOLOGY & ENVIRONMENTAL GENOMICS

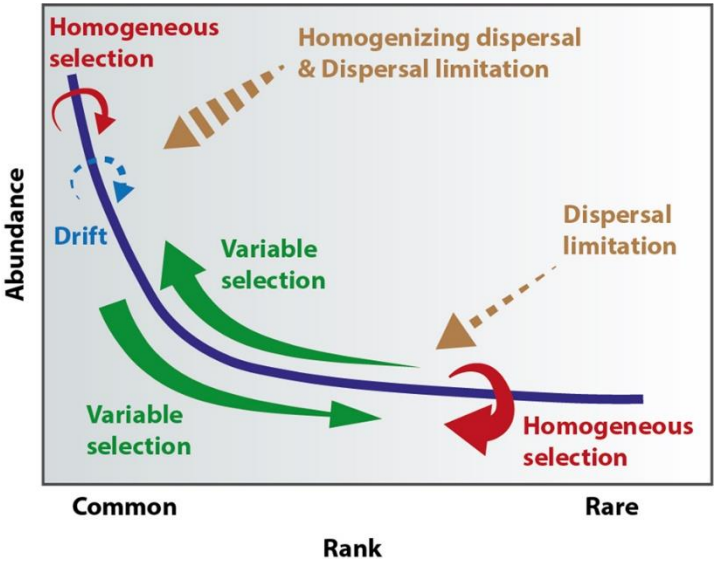
NIOO NEDERLANDS INSTITUUT VOOR ECOLOGIE
NETHERLANDS INSTITUTE OF ECOLOGY

DEPARTMENT Microbial Ecology

Feb 3, 2023

DMG symposium
"Practical applications of metagenomics towards typing"

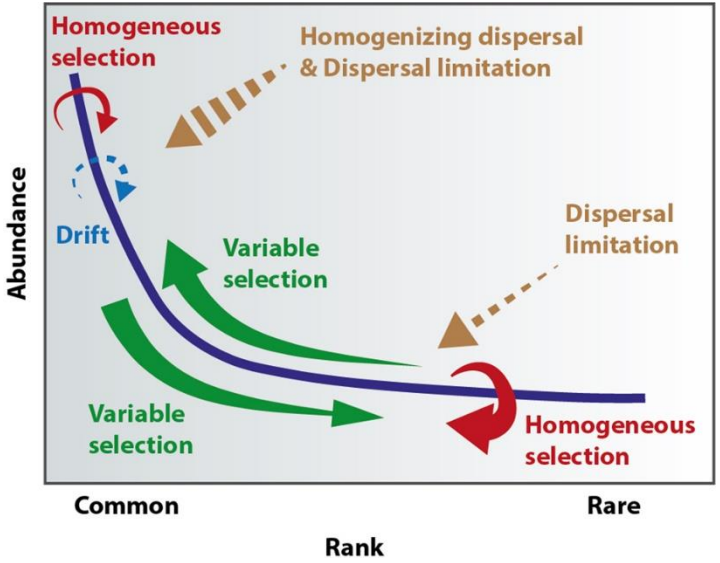
What we want/hope to know from metagenomics sequencing (MGS) data?



Jia, Dini-Andreote, Salles. *ISME COMMUN*. **2**, 96 (2022).

What we want to know from sequencing data?

- Community composition
- Diversity
- Differential abundance
- Co-occurrence



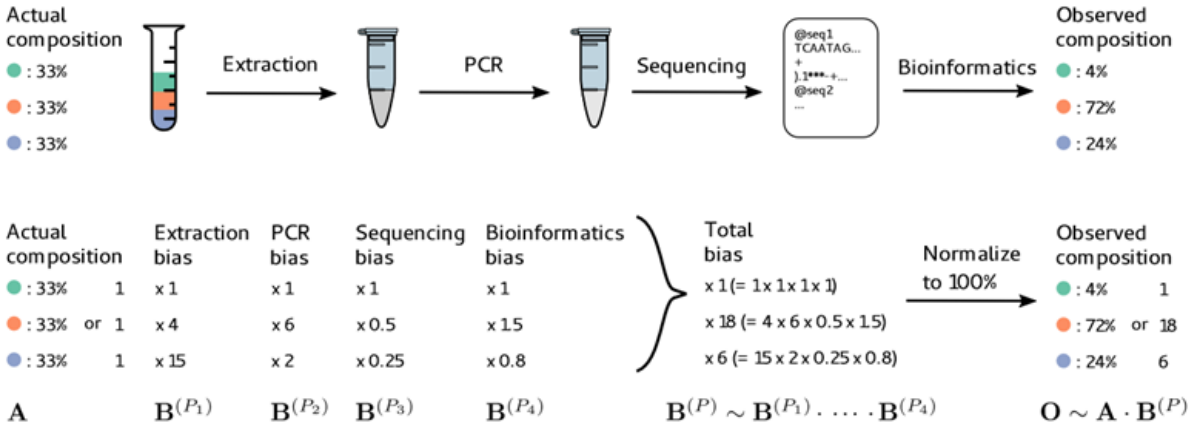
Jia, Dini-Andreote, Salles. *ISME COMMUN.* **2**, 96 (2022).

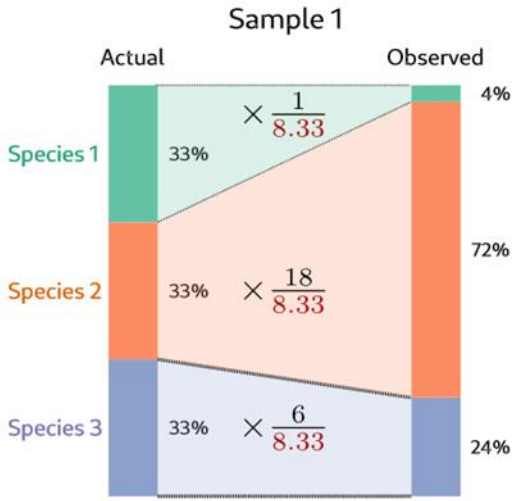

We need dissimilarity metrics
Lots of them!


TSS
TMM
DESEq2
ZIP
CSS
ANCOM-BC
ANCOM
CLR
ALDEx2
ZINB
corncob

Data essence precedes data existence: unique challenges of microbiome data

MGS bias + Compositionality + sparsity + high variability (overdispersion)

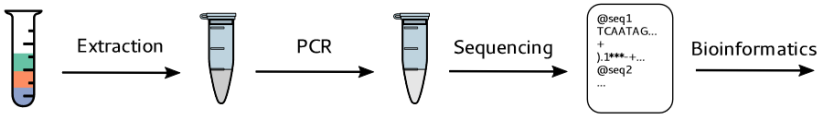


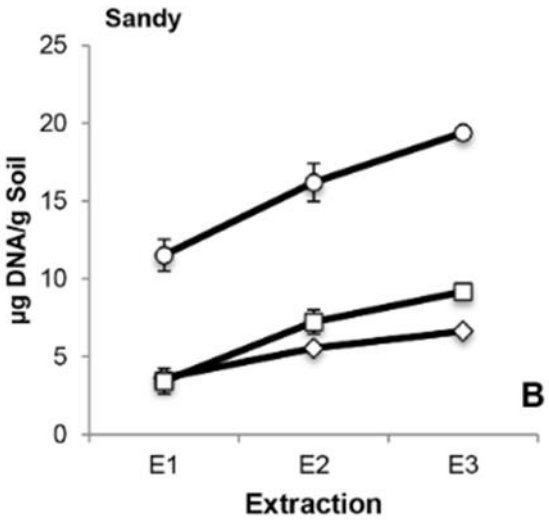


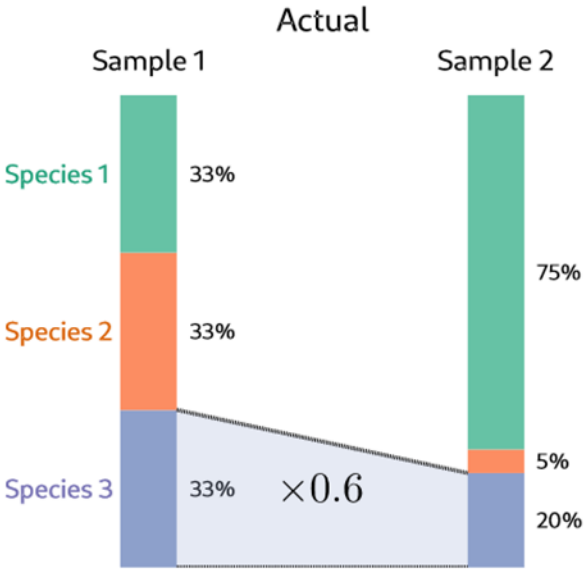
McLaren, Willis, Callahan. *Elife*. 2019.

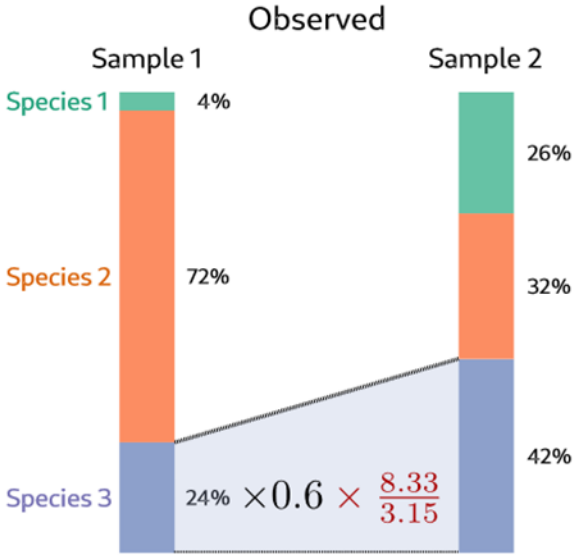
McLaren et al. *biorxiv*. 2022.

MGS bias









Dimitrov et al. *PeerJ*. 2017.

McLaren et al. *biorxiv*. 2022.

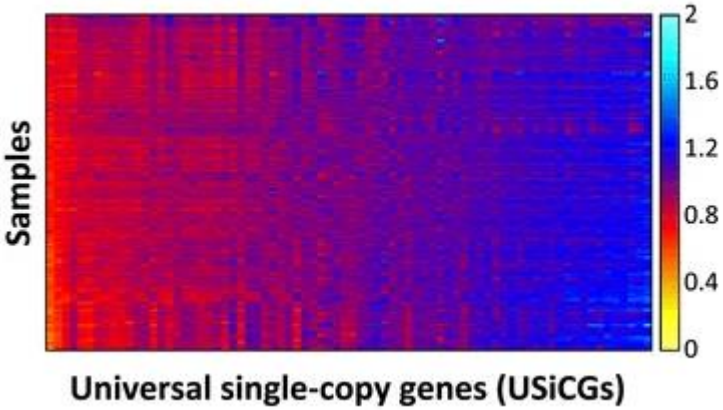
Diversity

DivNet

α-Diversity
+
β-Diversity

Willis & Martin. *Biostatistics*. 2022.

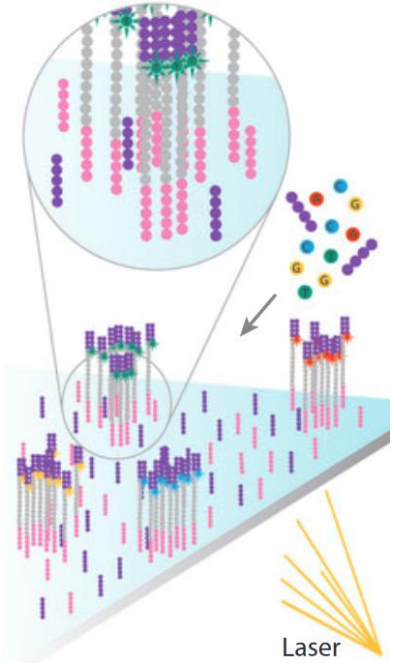
MUSiCC



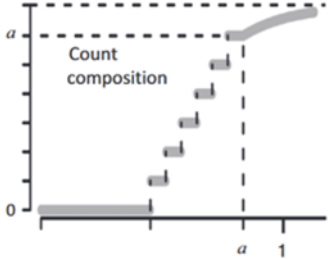
Manor & Borenstein. *Genome Biology*. 2015.

Compositionality

Sequencing is not counting.

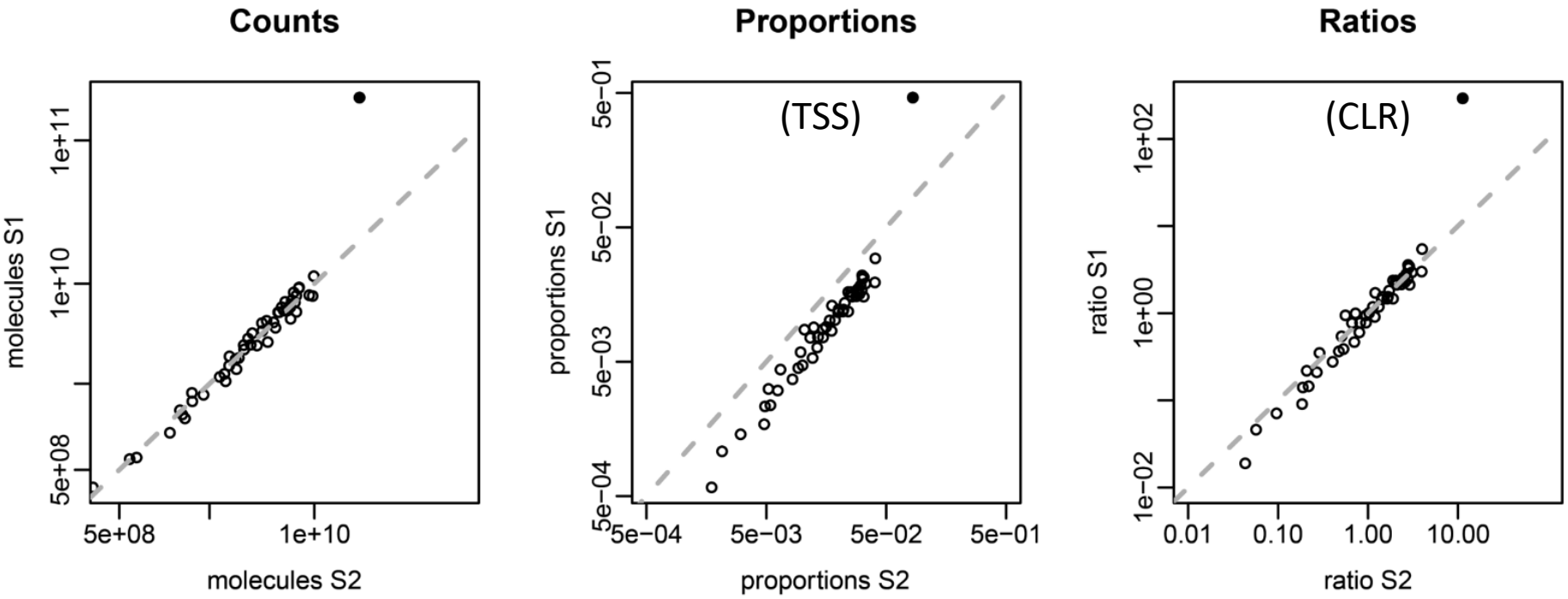


Mardis. *Annu. Rev. Genom. Human Genet*. 2008.



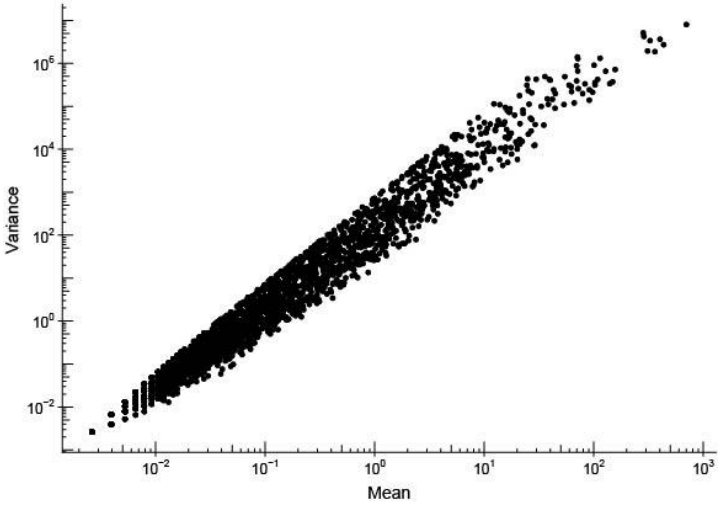
Compositionality

Relative abundance (proportions) is prone to errors.

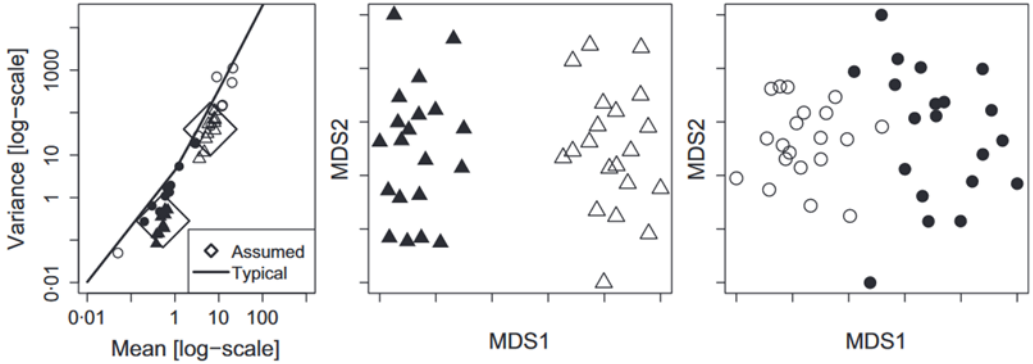


Gloor & Reid. *Can. J. Microbiol*. 2016.

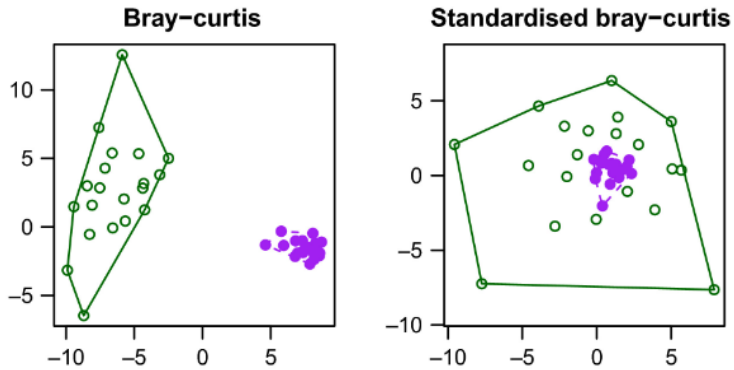
Sparsity + Overdispersion



Leite & Kuramae. *SBB.* 2020.



Warton et al. *Methods Ecol. Evol.* 2012.

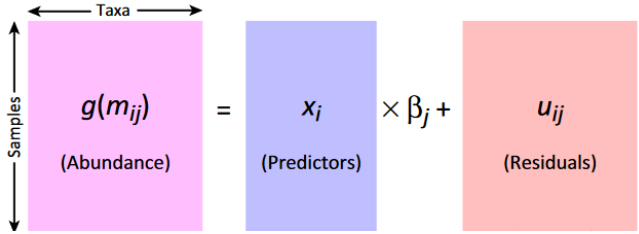


Warton & Hui. *Methods Ecol. Evol.* 2017.

What we need… (the Plato's method)

i. Investigate microbial community under the **compositional constraint**;
ii. Determine the influence of **sequencing bias**;
iii. Control data **sparsity and overdispersion**;
iv. Account for the influence of **rare taxa**;
v. Integrate additional information (e.g., traits, environmental covariates, experimental design);
vi. Disentangle biotic interaction from environmental response.

So far… Model-based approaches

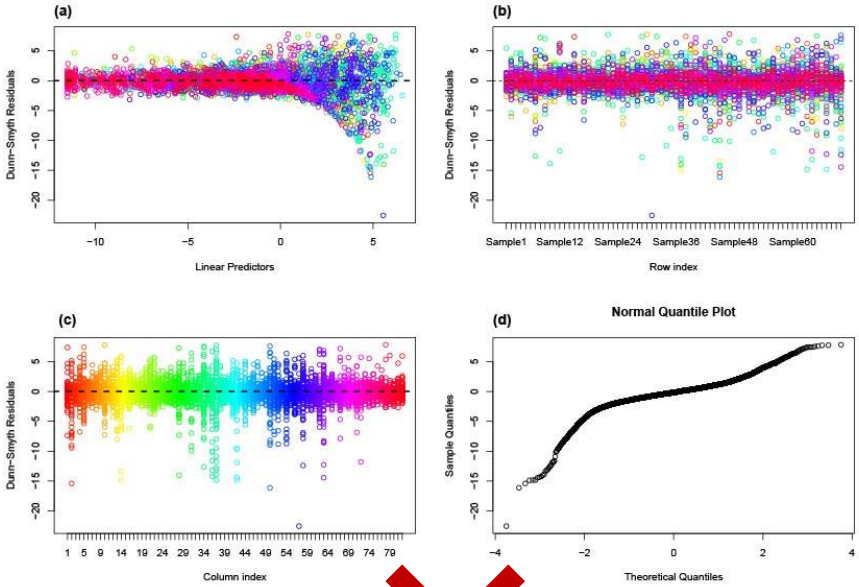


Warton et al. *TREE*. 2015.

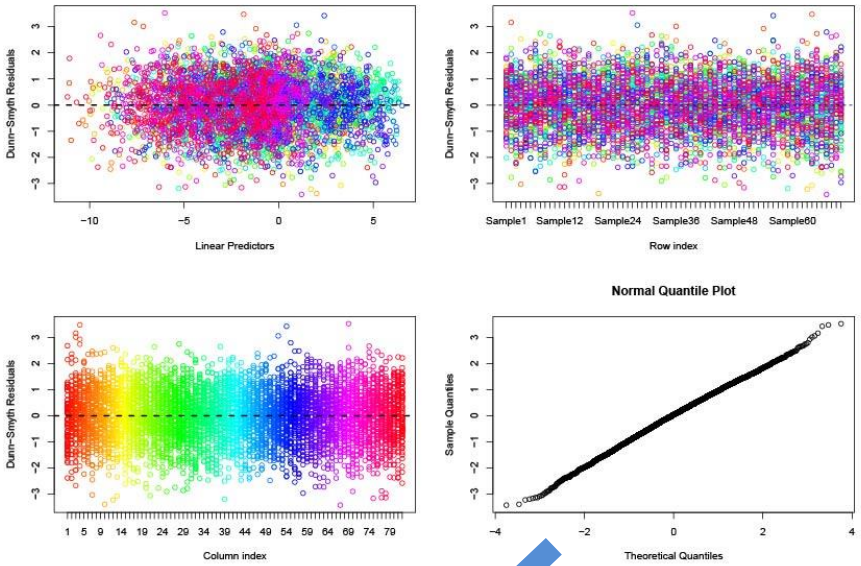
- ☑ Compositional constraint;
- ☑ MGS bias;
- ☑ Overdispersion via mean-variance relationship;
- ☑ Rare taxa;
- ☑ Integrate additional information (e.g, traits, environmental covariates, experimental design);
- ☑ Disentangling interaction.
- ☑ Bonus: Unobserved/Latent variables

Assumptions first

Poisson Distribution



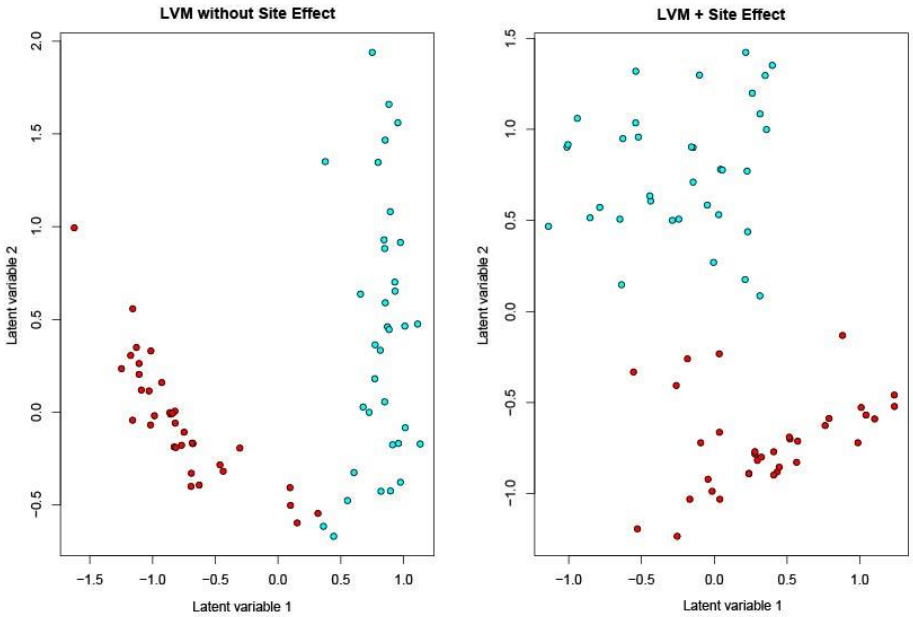
Negative Binomial Distribution



Leite & Kuramae. *SBB*. 2020.

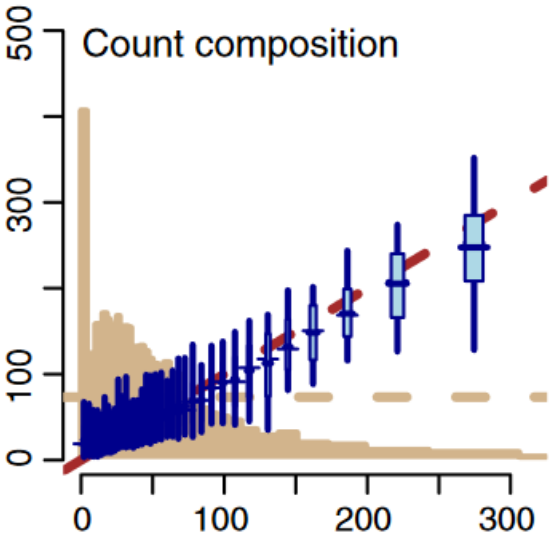
Assumptions first

Compositionality via site effects



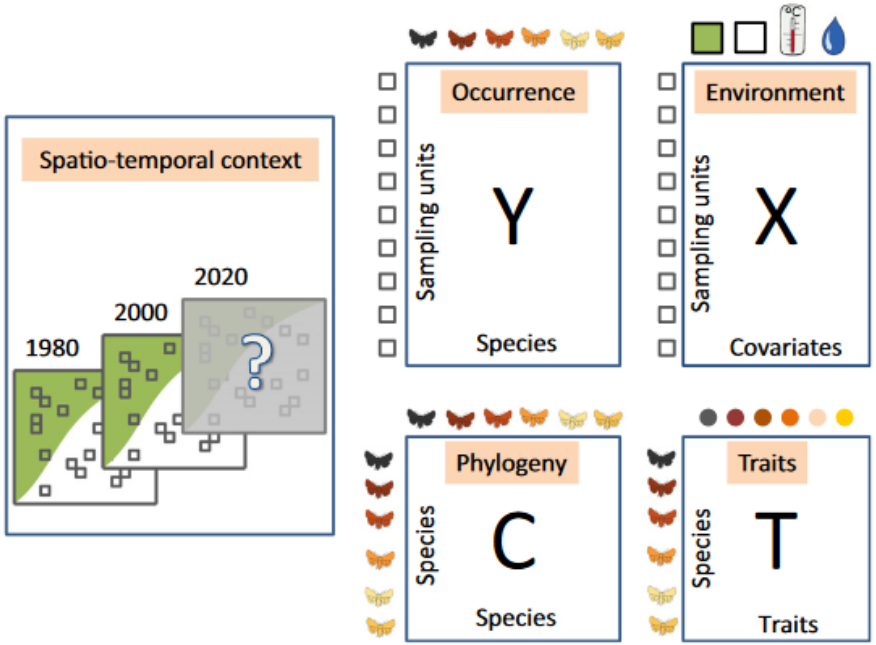
Leite & Kuramae. *SBB.* 2020.

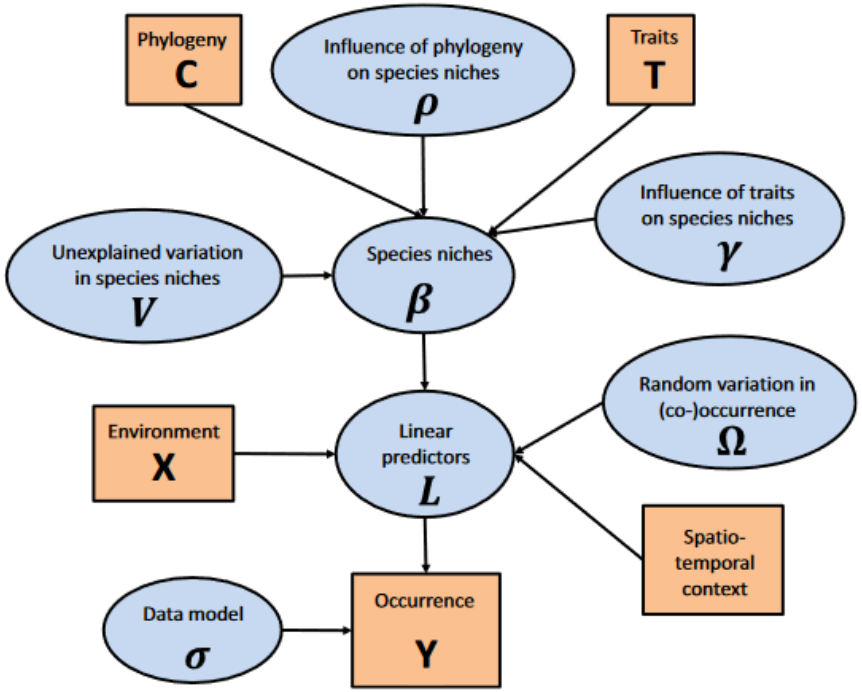
Composition scale + Sampling effort



Clark et al. *Ecol. Monogr.* 2017.

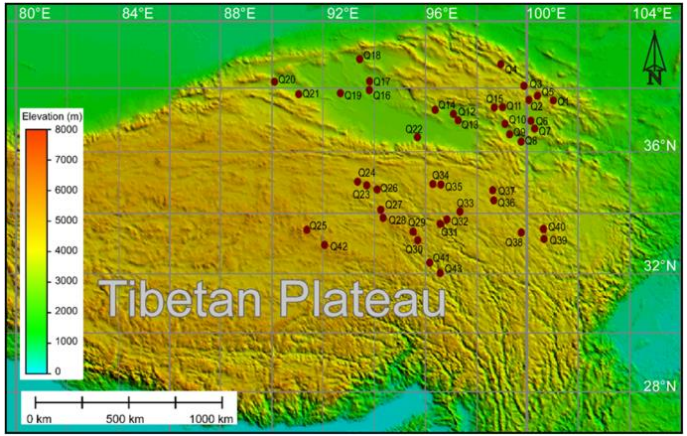
Joint and hierarchical models

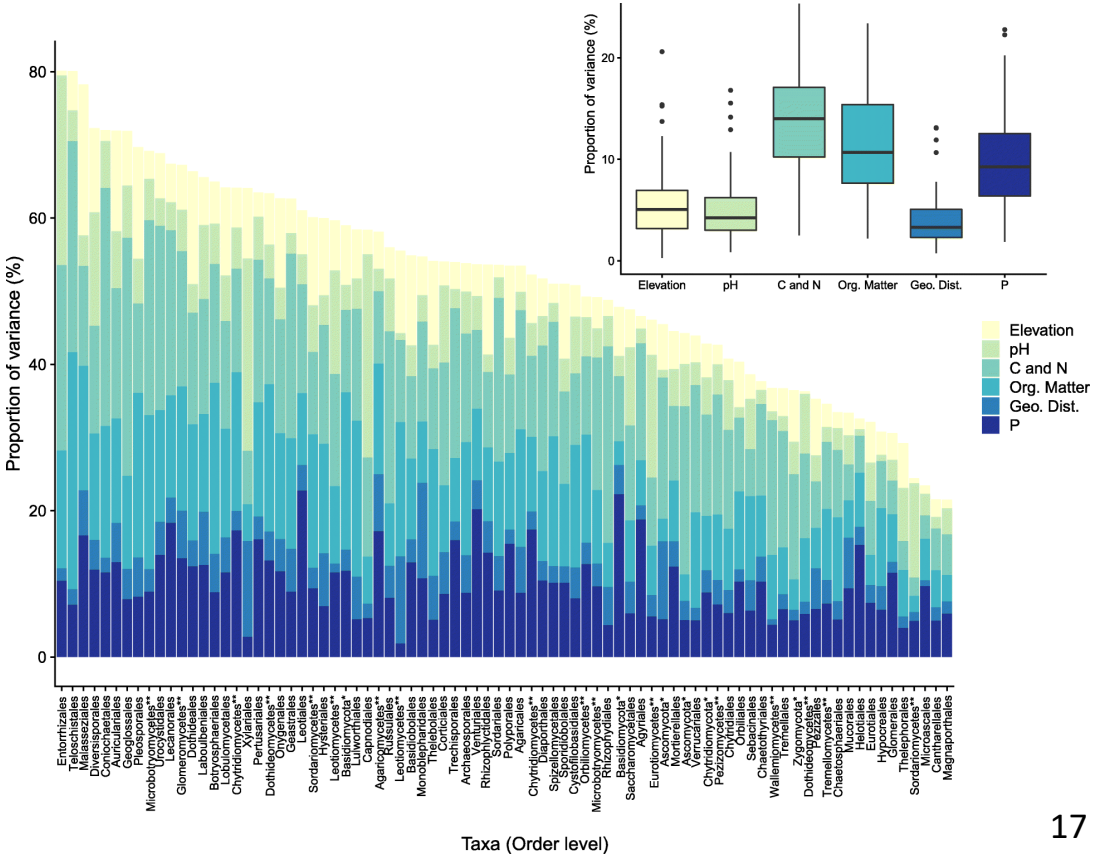




Ovaskainen et al. *Ecol.* 2017.

Joint and hierarchical models

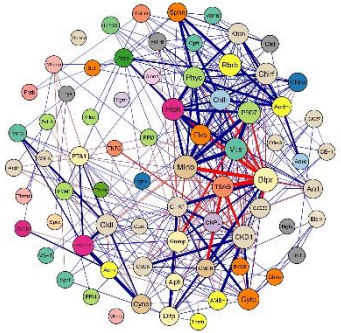




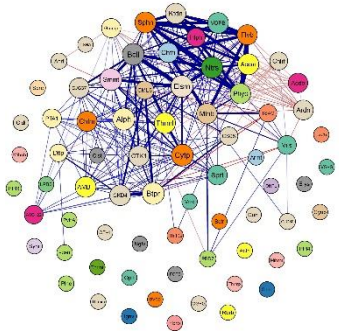
Zhou et al. *Environ. microbiome.* 2021.

Disentangling correlations

Plant-Soil driven interactions

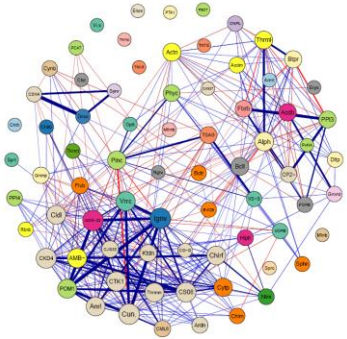


Bacteria-Bacteria interactions

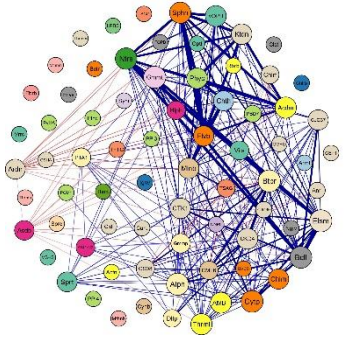


×

Plant-Soil driven interactions



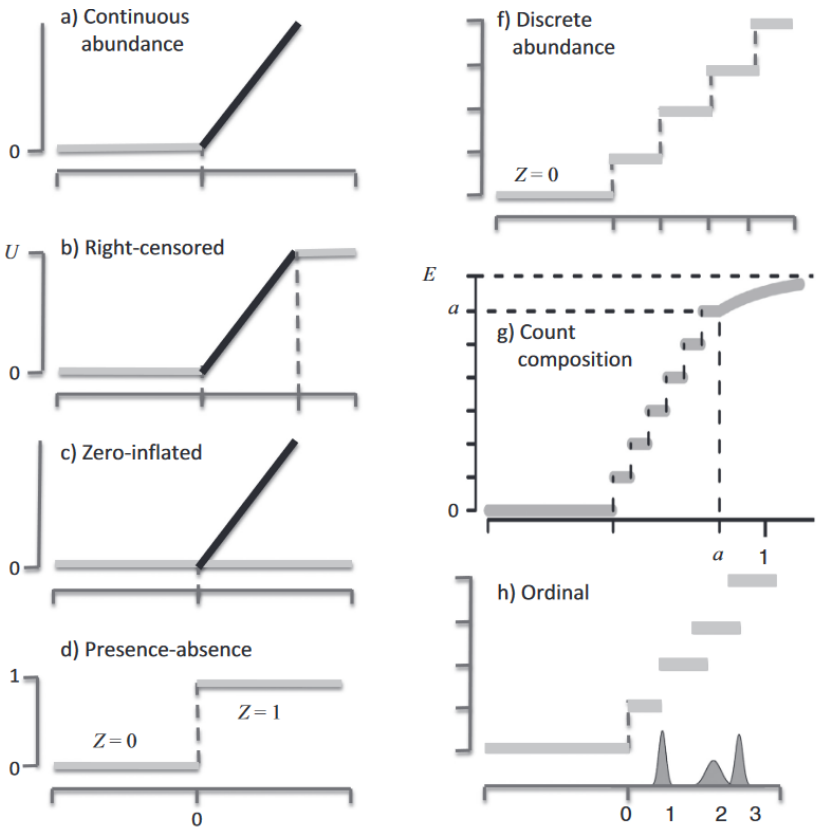
Bacteria-Bacteria interactions



Data Integration

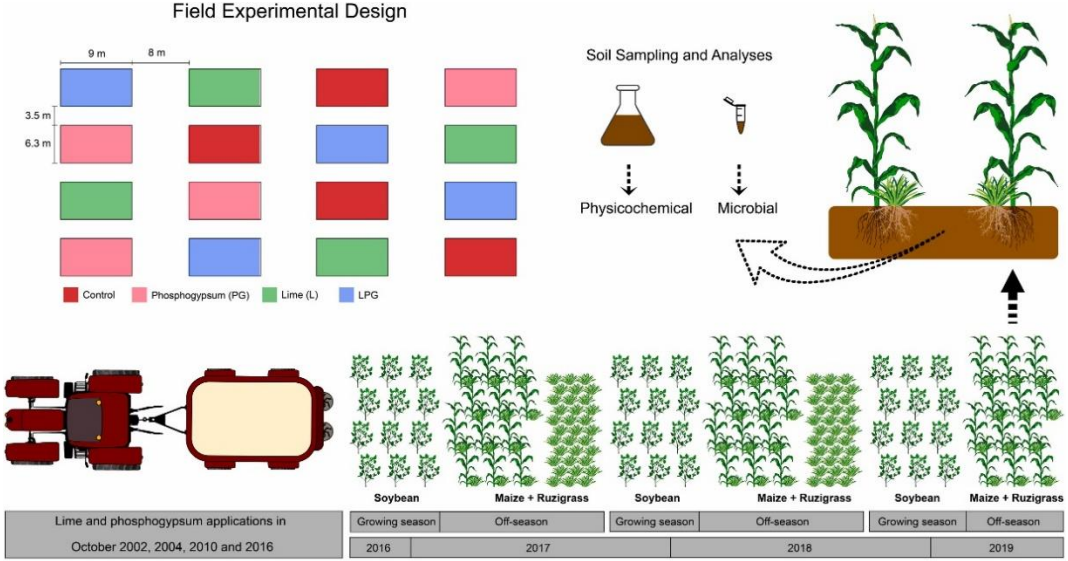
- How to integrate data that has different types?

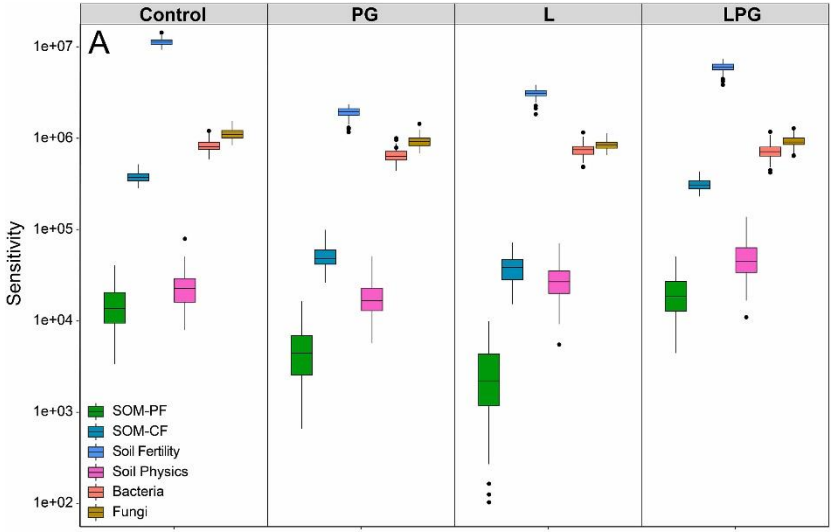
Generalized Joint Attribute Model (GJAM)



Clark et al. *Ecol. Monogr.* 2017.

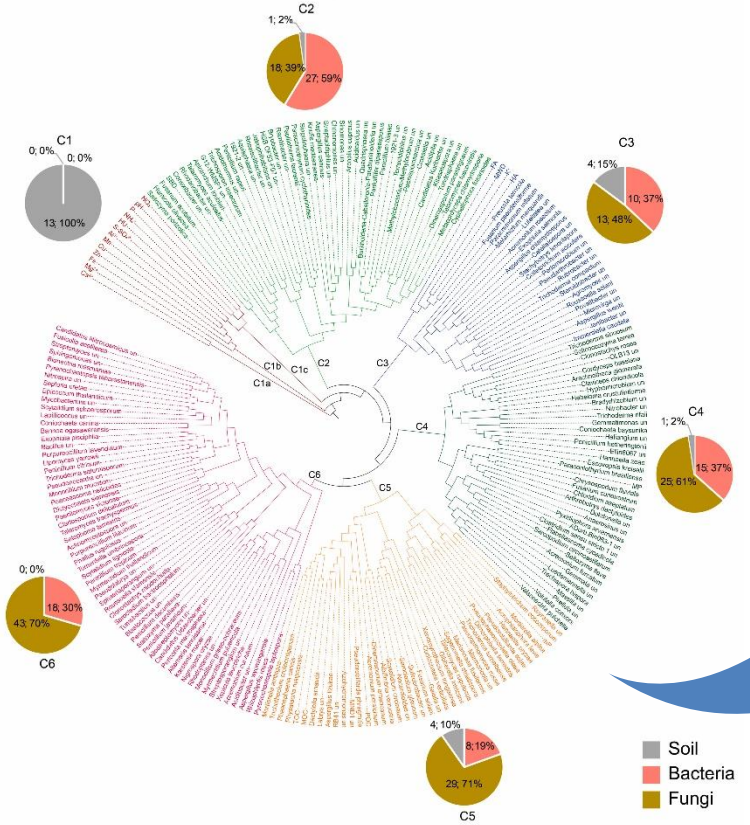
Data integration for agriculture

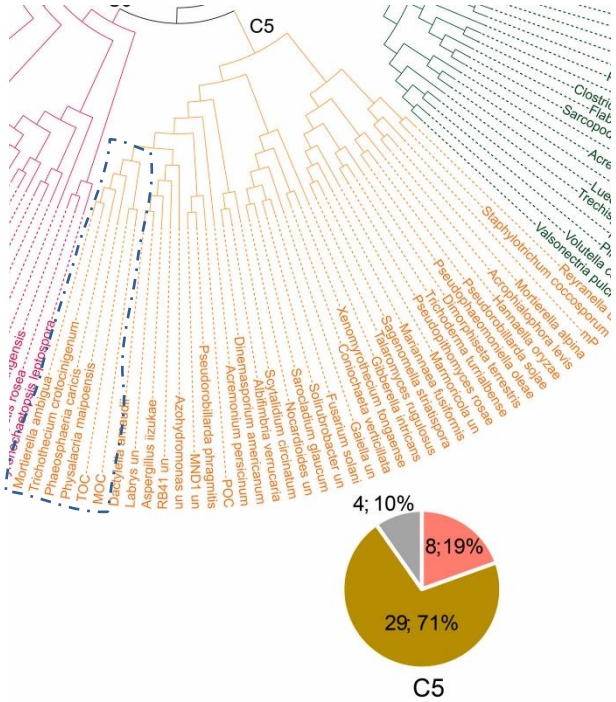




Bossolani et al. *SBB*. 2021.

Data integration for agriculture





Bossolani et al. *SBB*. 2021.

Take-home message

- When facing a large arsenal of choices: assumption checking as the first step;

- Model-based approaches can explicitly account for key statistical properties of metagenomics data.

Thank you!

@MarcioLeiteEco

Leitemfa

/DMG-symposium

M.Leite@nioo.knaw.nl

NIOO NEDERLANDS INSTITUUT VOOR ECOLOGIE
NETHERLANDS INSTITUTE OF ECOLOGY

KURAMAE GROUP