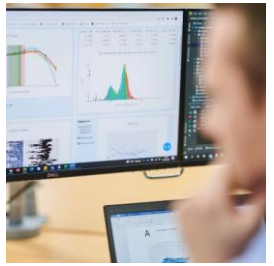
Addressing metagenomic data compositionality and confounding factors in clinical studies for the safety assessment of human microbiome perturbations

Aline Métris

SERS
Safety, Environmental & Regulatory Science

Unilever

Introduction : Assessing the safety of human microbiome perturbations from New Generation Sequencing (NGS) data


MICROBIOLOGY SOCIETY Discover our portfolio Our resources About us

MICROBIAL GENOMICS

Volume 11, Issue 5

Open Access

Assessing the safety of microbiome perturbations

Aline Metris[1], Alan W. Walker[2], Alicia Showering[3], Andrea Doolan[4], Andrew J. McBain[5], Antonis Ampatzoglou[1], Barry Murphy[6], Catherine O'Neill[7], Colette Shortt[8], Elizabeth M. Darby[9], Gary Aldis[10], Greg G. Hillebrand[11], Helen L. Brown[12], Hilary P. Browne[13,14], Jay P. Tiesman[15], Joy Leng[16], Leo Lahti[17], Nicholas S. Jakubovics[18], Oliver Hasselwander[19], Robert D. Finn[20], Silvia Klamert[1], Tamas Korcsmaros[21,22,23] and Lindsay J. Hall[9,21]

View Affiliations

Published: 15 May 2025 https://doi.org/10.1099/mgen.0.001405

- NGS data from clinical studies have shown taxonomic associations between health and disease
- A "healthy" microbiome is relative to the host general and local health status, body site, age, lifestyle, environmental factors etc.
- A "healthy" microbiome is not universally defined because of **confounding factors** & bias between studies due to extraction methods & bioinformatics analysis

<u>Some NGS data analysis challenges</u>:

- Multivariate
- Sparsity
- Heteroscedasticity
- **Compositionality**

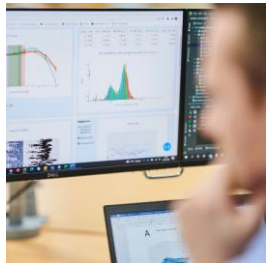
=> To understand the safety of intervention, experimental design (part 1) & data analysis (part 2) of clinical studies need to be optimised.

Unilever

Part 1

Optimising experimental design

SERS
Safety, Environmental
& Regulatory Science



Unilever

Clinical study design: targeted longitudinal studies

Randomized controlled trials with enough power are costly so for smaller studies, consider the following:

⇒ Target population (geographical, age, lifestyle, vulnerability …)

⇒ Cross-over designs
- Each participant serves as their control - minimize the people variability effect
- Time series - as microbiome resilience linked to health

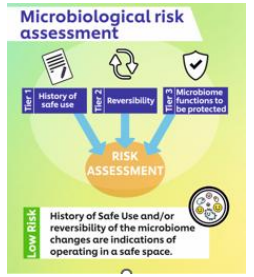
⇒ Intervention: realistic dose, exposure (site, frequency) & comparison with a control

⇒ Sampling, extraction, measurement methods (e.g. 16S rRNA region) & bioinformatics adapted to body site/question (e.g. 16S reference database)

⇒ Additional measurements to NGS data: quantitative counts (qPCR, flow cytometry), other types of data (e.g. –omics to look at function) and adequate host (e.g. cytokines) & environment (e.g. pH, moisture) metadata/measurements.

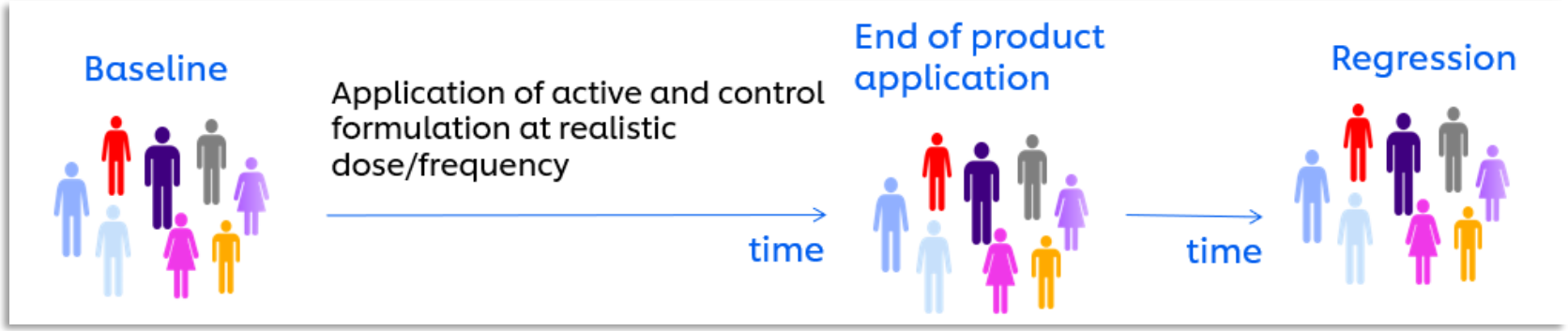
Unilever

Example of safety assessment approach: the reversibility of change for beauty and personal care products



https://doi.org/10.1016/j.mran.2021.100188

- Including a **control/placebo** to define significant change (on the same person where possible)
- Including qPCR for **quantitative** representation of the microbiome

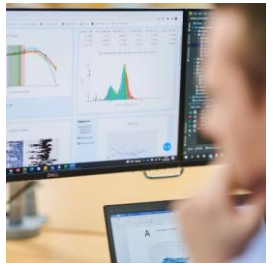


The microbiome returning to its initial state after a period of application and regression is evidence of low risk – relative Risk Assessment.

Unilever

Part 2

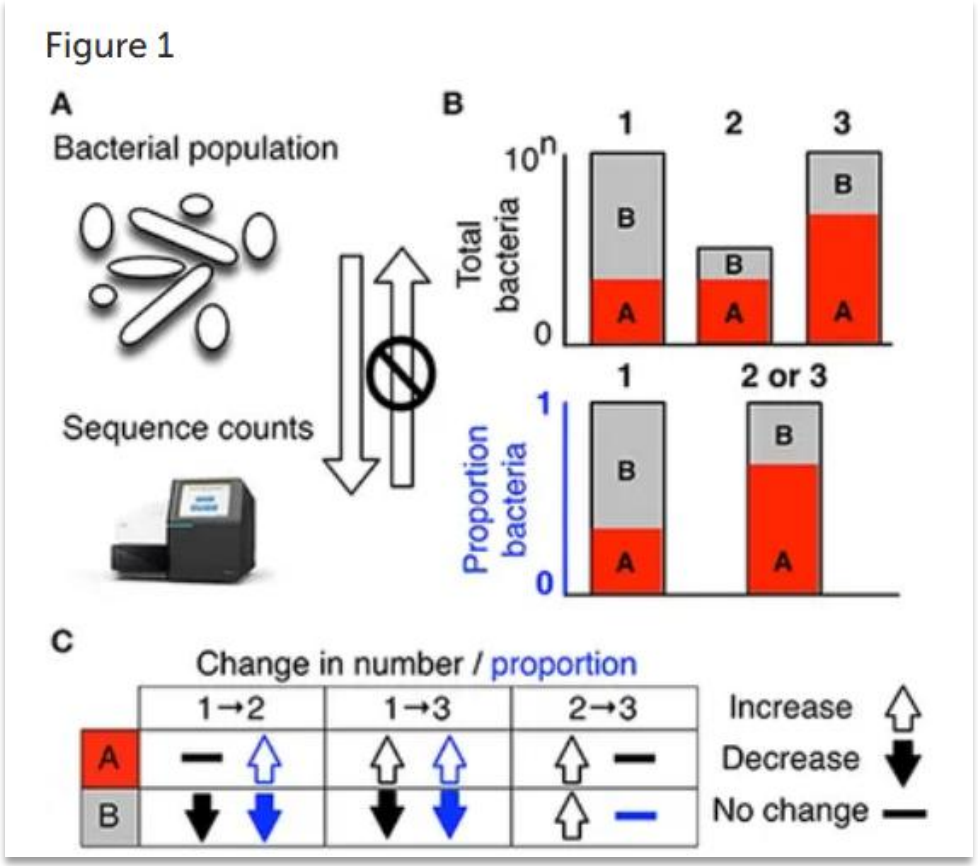
Optimising data analysis



SERS
Safety, Environmental
& Regulatory Science

Unilever

NGS data compositionality



Microbiome Datasets Are Compositional: And This Is Not Optional

Gregory B. Gloor[1*] Jean M. Macklaim[1] Vera Pawlowsky-Glahn[2] Juan J. Egozcue[3]

[1] Department of Biochemistry, University of Western Ontario, London, ON, Canada
[2] Departments of Computer Science, Applied Mathematics, and Statistics, Universitat de Girona, Girona, Spain
[3] Department of Applied Mathematics, Universitat Politècnica de Catalunya, Barcelona, Spain

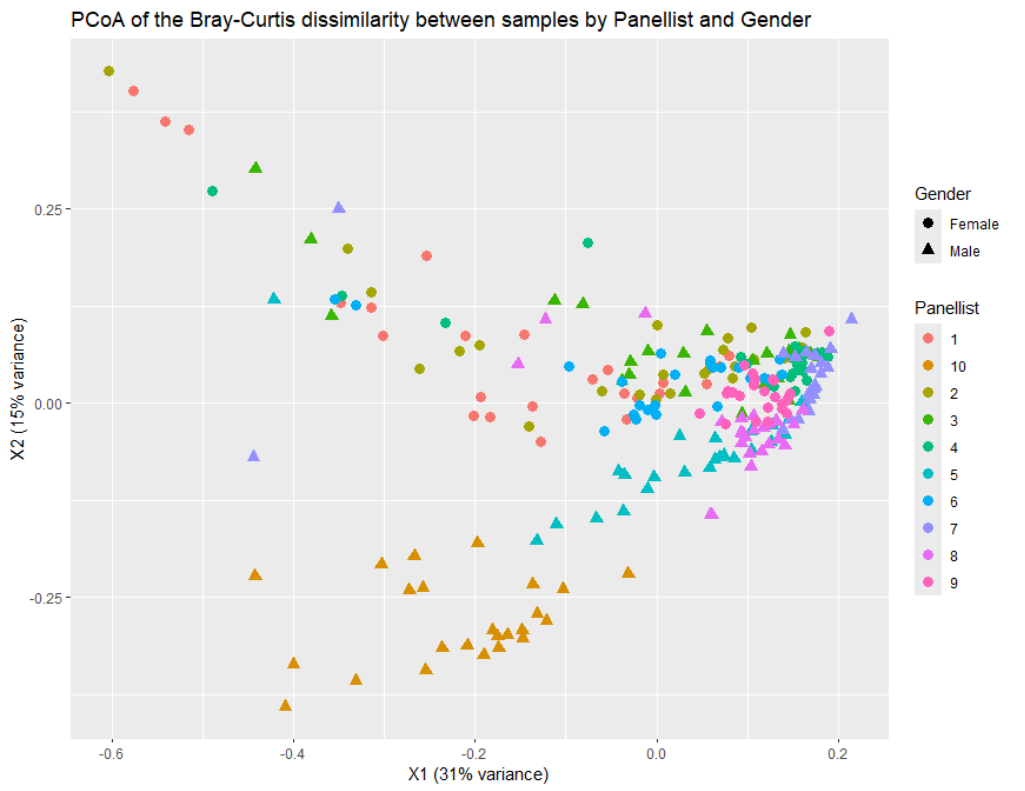
https://doi.org/10.3389%2Ffmicb.2017.02224

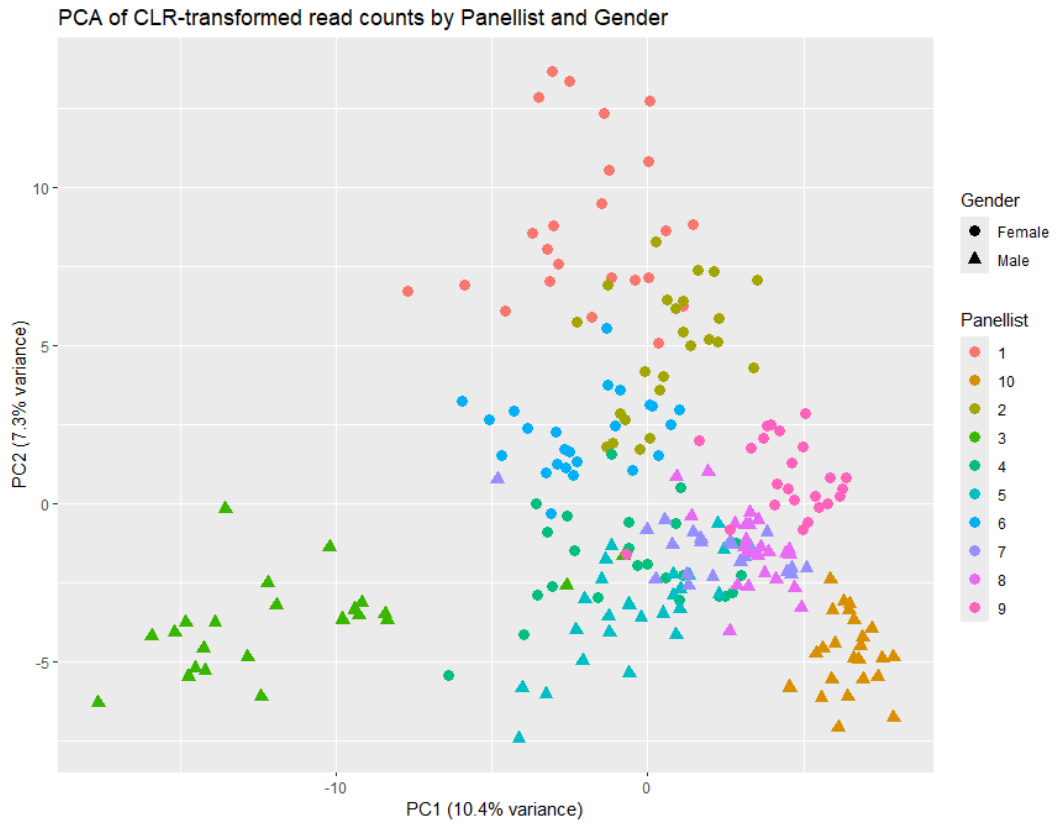
Figure 1.
When sequencing we lose information on total count (A).

Features A and B in samples 2 and 3 appear with the same relative abundances, even though the counts in the environment are different (B).

And hence no difference between samples 2 and 3 in relative terms (C) while for the host, it is the concentration it is exposed to that matters.

Consequence of compositionality –representation with PCoA vs. PCA





In a study with 10 panellists of volar and dorsal skin samples treated by ethanol, the largest source of variability is panellists (ANOSIM R=0.41, Significance: 0.0002) followed by gender. The PCA representation renders the grouping more clearly than PCoA.

Transitory Shifts in Skin Microbiota Composition and Reductions in Bacterial Load and Psoriasin following Ethanol Perturbation

Charis R. Saville,[a] Aline Metris,[b] Gavin J. Humphreys,[a] Catherine O'Neill,[c] Paul Barrett,[b] Judith Fernandez-Piquer,[b] Andrew J. McBain[a]

Consequence of compositionality – Differential Abundance (DA)

LETTER

doi:10.1038/nature24460

Quantitative microbiome profiling links gut community variation to microbial load

Doris Vandeputte[1,2,3]*, Gunter Kathagen[1,2]*, Kevin D'hoe[1,2,3]*, Sara Vieira-Silva[1,2]*, Mireia Valles-Colomer[1,2], João Sabino[4], Jun Wang[1,2], Raul Y. Tito[1,2,3], Lindsey De Commer[1], Youssef Darzi[1,2], Séverine Vermeire[4], Gwen Falony[1,2]§ & Jeroen Raes[1,2]§

https://www.nature.com/articles/nature24460

- Stools samples analysed with flow cytometry for cell counts.
- Cells counts did not correlate with sequencing depth but biological process like transit time.
- Analysis of the differential abundance leads to different results when looking at quantitative vs. relative abundance with Crohn's disease (CD) .

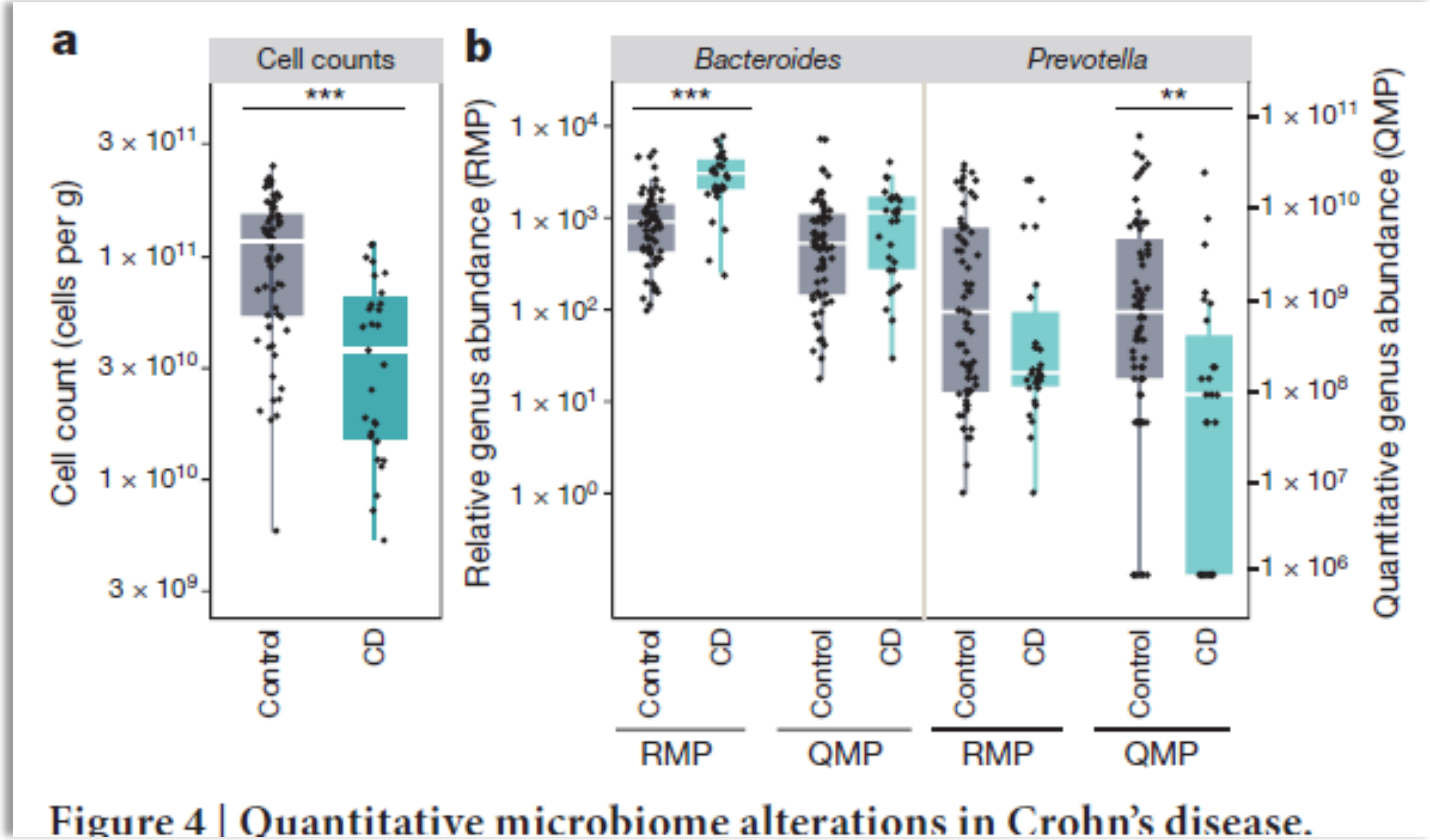


Figure 4 | Quantitative microbiome alterations in Crohn's disease.

Hierarchical mixed-effects models for random and confounding factors

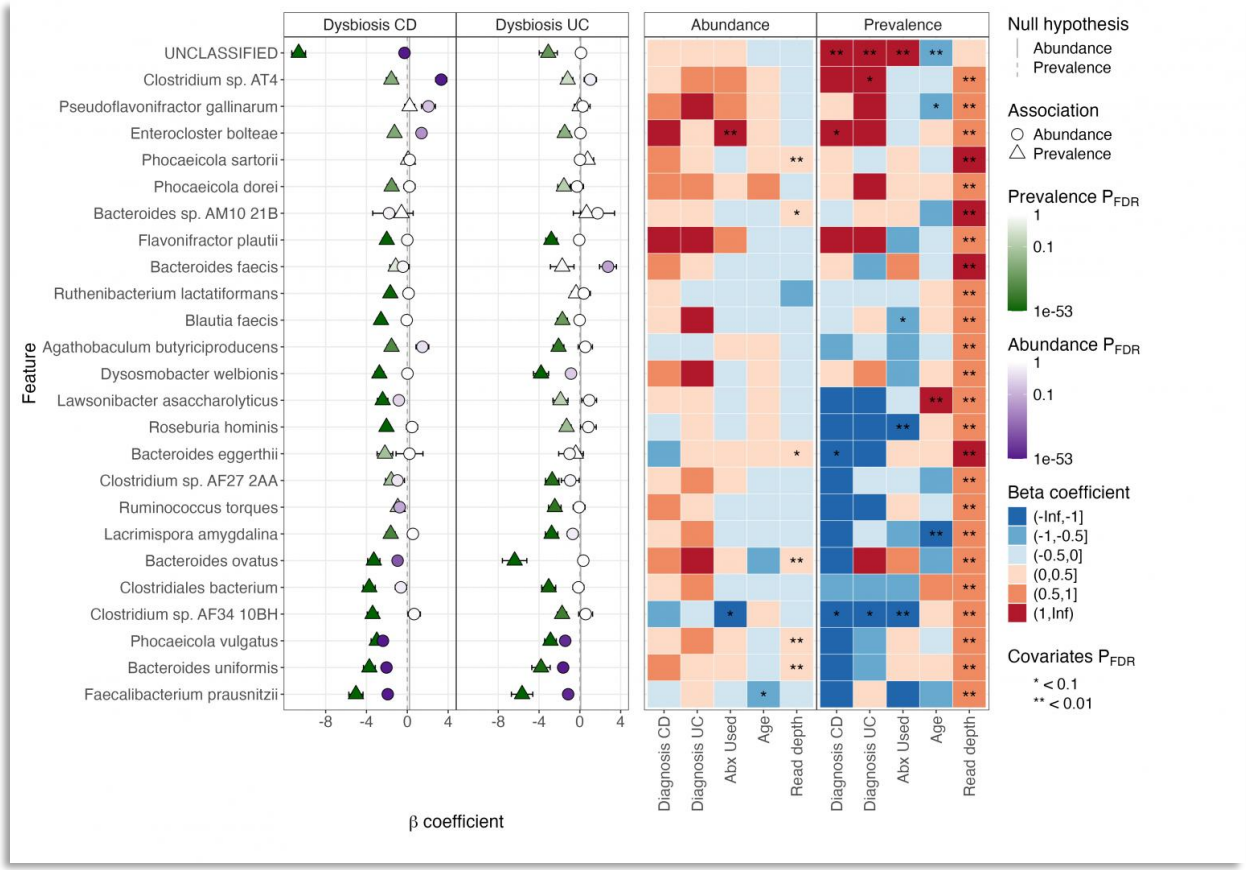
Hierarchical mixed-effects models

$$y_{ij} = \alpha_{[j]} + \beta \times x_i + \epsilon_i$$

Random effects
e.g. people (j)

Fixed effects e.g. time
or intervention (i)

https://huttenhower.sph.harvard.edu/maaslin3/



For example, in Maaslin3, abundance and prevalence are regressed separately, there is an option to separate fixed from random effects (equation based on lme4 R library) and scaling options.
CD Crohn's disease, UC Ulcerative Colitis

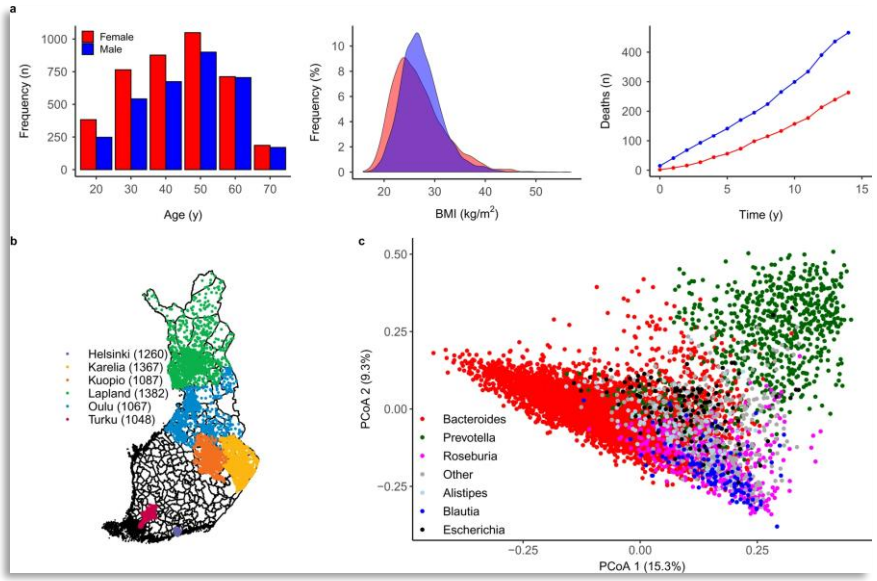
The potential of big data & metadata for better predictions with ML methods

Taxonomic signatures of cause-specific mortality risk in human gut microbiome

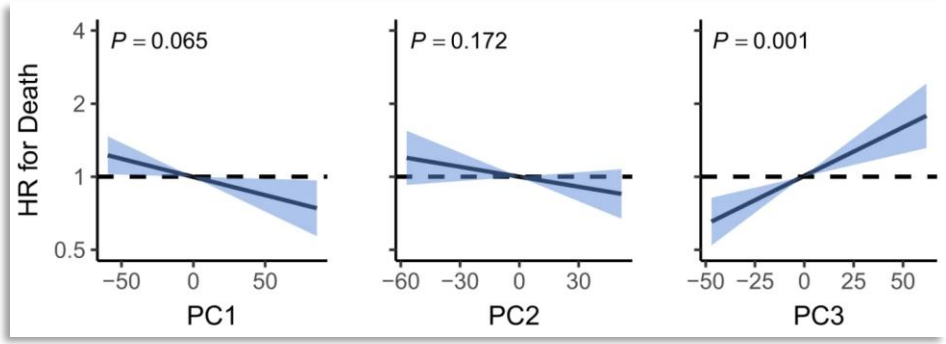
Aaro Salosensaari, Ville Laitinen, Aki S. Havulinna, Guillaume Meric, Susan Cheng, Markus Perola, Liisa Valsta, Georg Alfthan, Michael Inouye, Jeramie D. Watrous, Tao Long, Rodolfo A. Salido, Karenina Sanders, Caitriona Brennan, Gregory C. Humphrey, Jon G. Sanders, Mohit Jain, Pekka Jousilahti, Veikko Salomaa, Rob Knight, Leo Lahti & Teemu Niiranen

Nature Communications **12**, Article number: 2671 (2021) | Cite this article

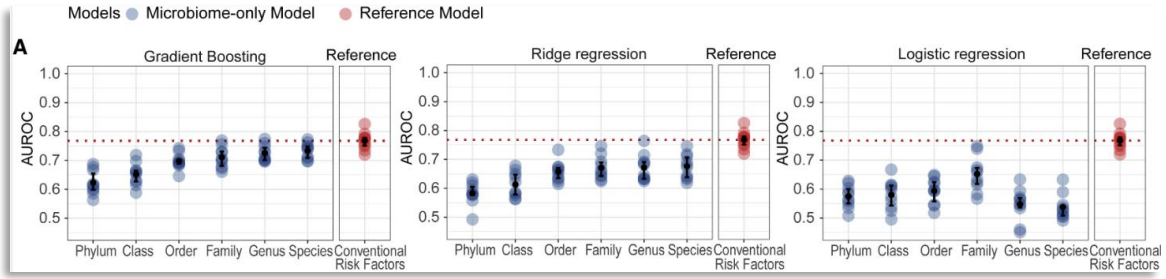
11k Accesses | **16** Citations | **359** Altmetric | Metrics



Finnish faecal sample collection with diet, lifestyle and linked to health data.



The hazard ratio (HR, all causes) for death correlates with the third coordinate of beta diversity (PC3 driven by species of the Enterobacteriaceae family, based on centred-log-ratios of abundance).



- Microbiome predictive of risk score for incident liver disease
- Gradient boosting outperforms ridge & logistic regression

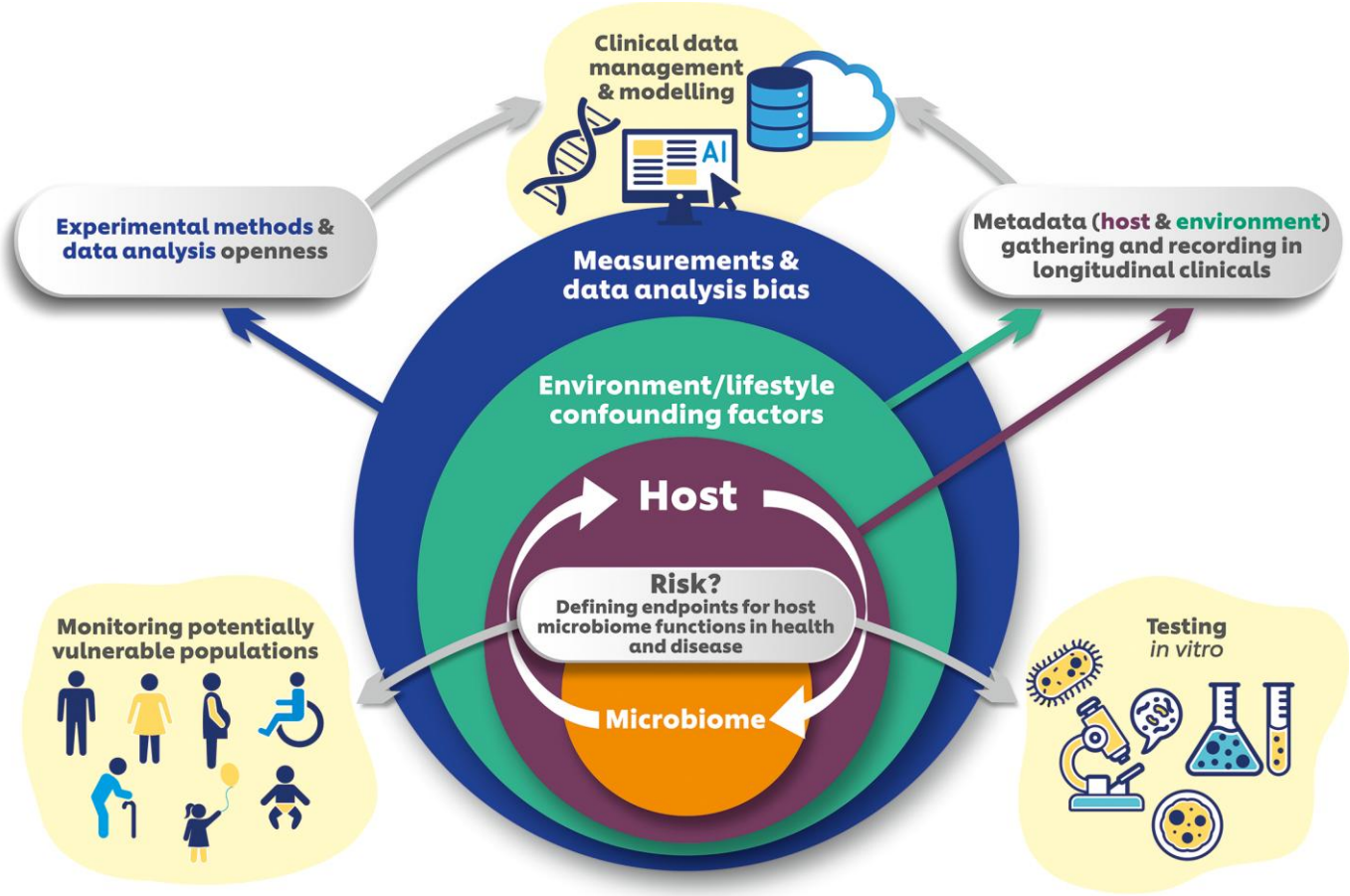
Future work: the microbiome and risk assessments

MICROBIOLOGY SOCIETY Discover our portfolio Our resources About us

MICROBIAL GENOMICS
Volume 11, Issue 5
Open Access

Assessing the safety of microbiome perturbations

Aline Metris[1], Alan W. Walker[2], Alicia Showering[3], Andrea Doolan[4], Andrew J. McBain[5],



Metris et al., 2025. **Microbiome perturbation safety assessments. Microbial Genomics.**

To characterise endpoints, need to have a transparent - data, **metadata (host, environment & methods)** and models, especially those based on ML methods.

Conclusions

At present, as we have no well-defined health endpoint, risk assessments are relative, and clinical longitudinal studies are the most informative.

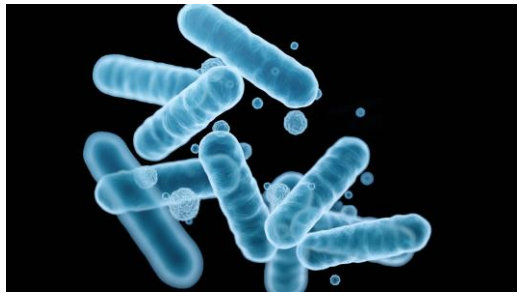
- Sequencing data are compositional => appropriate normalisation, metrics & quantitative measurements are necessary.
- Current development of scaling methods & tools based on linear mixed-effect model are promising for differential abundance allowing to consider people variability and other sources of variability.
- ML methods applied to longitudinal population cohorts has the potential to disentangle complex & overlapping relationships and identify vulnerable populations but require link to health records.

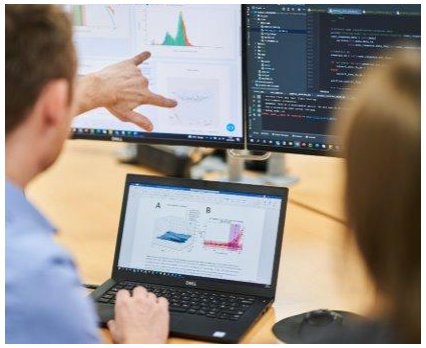
To advance risks assessments NGS data need to be linked to the host health, environmental conditions & methods (metadata).

Acknowledgements

SERS
Safety, Environmental & Regulatory Science

Pioneers & Trusted Partners in Regulatory Science, co-creating the Future for Superior, Safe & Sustainable Products





Unilever SERS microbiology team & computational team
Isaac Sharp (R coding)

In memory of

Prof. Kostas Koutsoumanis

Dr. Moira Parker

Unilever