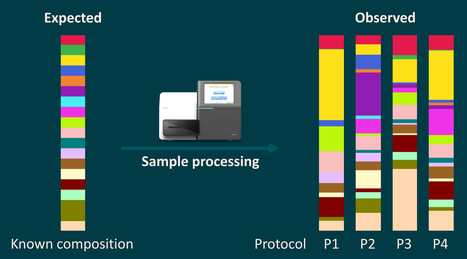
Meta-analysis of positive controls and laboratory metainformation in microbiome data

Luise Rauer (luise.rauer[at]tum.de)[1,2,3], Tanja Gentz[1], Rajiv Karbhal[1], Claudia Hülpüsch[1,2,3], Matthias Reiger[1,2,3,4], Claudia Traidl-Hoffmann[1,2,3,4,5], Christian L. Müller[6,7,8], Avidan U. Neumann[1,2,3,4]

1. Environmental Medicine, Faculty of Medicine, University of Augsburg, Augsburg, Germany, 2. Chair of Environmental Medicine, Technical University Munich, Munich, Germany, 3. Institute of Environmental Medicine, Helmholtz Center Munich, Augsburg, Germany, 4. CK-CARE Centre for Allergy Research and Education, Davos, Switzerland, 5. Institute for Food & Health (ZIEL), Technical University Munich, Freising, Germany, 6. Center for Computational Mathematics, Flatiron Institute, New York, NY, USA, 7. Department of Statistics, LMU Munich, Munich, Germany, 8. Institute of Computational Biology, Helmholtz Center Munich, Neuherberg, Germany

Introduction

Human microbiome research has revolutionized our understanding of the **microbiome's contribution to human health, and diseases** such as obesity, Inflammatory Bowel Disease, or Atopic Dermatitis.
The huge variety of available methods for generating microbiome data leads to distinct errors and biases depending on the chosen laboratory method, which limits the comparability and clinical application of microbiome data.
These **protocol-specific errors and biases can be quantified by mock samples, i.e. positive controls** with known species composition that are processed along with biological samples.



Aim

We aim to build a **database of published microbiome studies** that used standardized, commercially available mock communities as positive controls. We then collect the **studies' laboratory metadata** to quantify the impact of different laboratory methods on microbiome data.

Pilot study

Search strategy & exclusion criteria

Company	Mock name	[mock name]	"[mock name]"	[mock name] [company]	"[mock name]" [company]
ATCC	MSA-1000	116	116	17	17
	MSA-1001	21	21	14	14
	MSA-1002	53	53	49	49
	MSA-1003	37	37	32	32
	MSA-2002*	482	482	32	32
	MSA-2003	384	384	33	33
BEI Resources	HM-280	176	176	9	9
	HM-281	162	162	10	10
	HM-782D	140	140	137	137
	HM-783D	75	75	74	74
Zymo-BIOMICS	D6300	10,400	741	134	134
	D6305	231	124	77	77
	D6310	192	36	7	7
	D6311	200	74	9	9
	D6322	103	34	5	5
	D6331	130	38	13	13

After identifying companies that provide commercially available mock communities, we performed a systematic literature search in Google Scholar to find scientific papers that use these mock communities. More specific results were found when the company's name was added to the search, instead of the name of the mock alone.
The n=32 articles mentioning MSA-2002 by ATCC (highlighted by asterisk) were then screened. Studies were excluded for the following reasons:

- Duplicate (n=1),
- **Full text not available (n=3),**
- Content not relevant (n=13),
- Sequencing technology out of scope (n=4),
- **Raw sequencing data not deposited (n=4),**

leading to only n=7 studies included for collection of laboratory metadata.

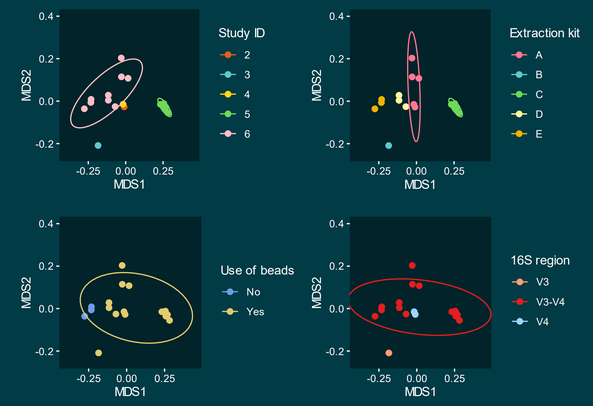
Metadata collection

Study ID	1	2	3	4	5	6	7
Mock lot number							
Amount of input cells / dilution							
Cryoprotectant							
Storage buffer							●
Storage temperature							
Extraction kit	●	●	●	●	●	●	●
Use of beads	●	●	●	●	●	●	●
Bead size and material	●	●		●	●	●	●
Lysis conditions	●	●		●	●	●	●
DNA quantification	●		●		●	●	●
16S region	●	●	●	●	●	●	●
Primer sequence	●	●	●	●	●	●	●
Number of PCR cycles	●				●		●
PCR temperatures		●		●	●		●
Polymerase		●	●	●	●		●
PCR product quantification		●		●	●		●
Equimolar sample pooling	●	●		●	●		●
Sequencing platform	●	●	●	●	●	●	●
PhiX spike-in	●						●
Sequencing kit	●	●	●	●	●		●
Bidirectional sequencing length	●	●	●	●	●	●	●
Basecalling software							
Sum (out of 23)	13	13	9	13	15	9	17

The n=7 studies using MSA-2002 **provided on average 13 (median) out of 23 required pieces of laboratory metadata**, indicated by dots in the table. Study ID 7 specified the most pieces of laboratory metadata, but this study is a best-practice protocol for improved sample processing. The laboratory metainformation is particularly missing in beginning of sample processing (upper quarter of the table). Underlined metadata were further investigated in the bias evaluation.

Bias evaluation

After further excluding studies with ID 1 and 7, which had their raw sequencing data deposited in a non-reusable format, n=5 studies with a total of 17 samples were bioinformatically processed.
The microbiome sequencing results of the mock community MSA-2002, as generated by the different laboratory methods chosen per study, show **substantial variation between individual studies** (A), as indicated by non-metric multidimensional scaling of Bray-Curtis dissimilarities between samples.
Among the chosen laboratory metadata for further bias evaluation, the choice of extraction kit seems to lead to the largest variation in results (B), compared to the use of beads (C) or 16S region (D).



References

Study ID 1: Martí et al. Cell Rep Med. 2021. doi: 10.1016/j.xcrm.2021.100206.
Study ID 2: Glendinning et al. Poult Sci. 2022. doi: 10.1016/j.psj.2021.101624.
Study ID 3: Pollock et al. Anim Microbiome. 2021. doi: 10.1186/s42523-021-00144-x.
Study ID 4: Glendinning et al. Anim Microbiome. 2019. doi: 10.1186/s42523-019-0017-z.
Study ID 5: Porcellato et al. Sci Rep. 2020. doi: 10.1038/s41598-020-77054-6.
Study ID 6: Dumont-Leblond et al. Commun Biol. 2021. doi: 10.1038/s42003-021-01690-5.
Study ID 7: Martí et al. STAR Protoc. 2021. doi: 10.1016/j.xpro.2021.100652.

Conclusions and next steps

- **Novel, powerful approach** to quantify biases in microbiome research.
- **Open access** to literature and raw data required to fully leverage the approach's power.
- **Specific reusable format** of deposited data required for state-of-the-art bioinformatic sample processing.
- Microbiome research needs **standardized reporting guidelines** for laboratory methods.
- Can the **search strategy** (search terms or search engine) be improved?
- Which laboratory metadata are the most important, and need to be available for re-analysis in our **database**?
- Inconsistent description and scattered distribution of metadata across the papers' methods sections require a **paper scraping algorithm**?
- **Expansion of meta-analysis approach** to other mock communities.