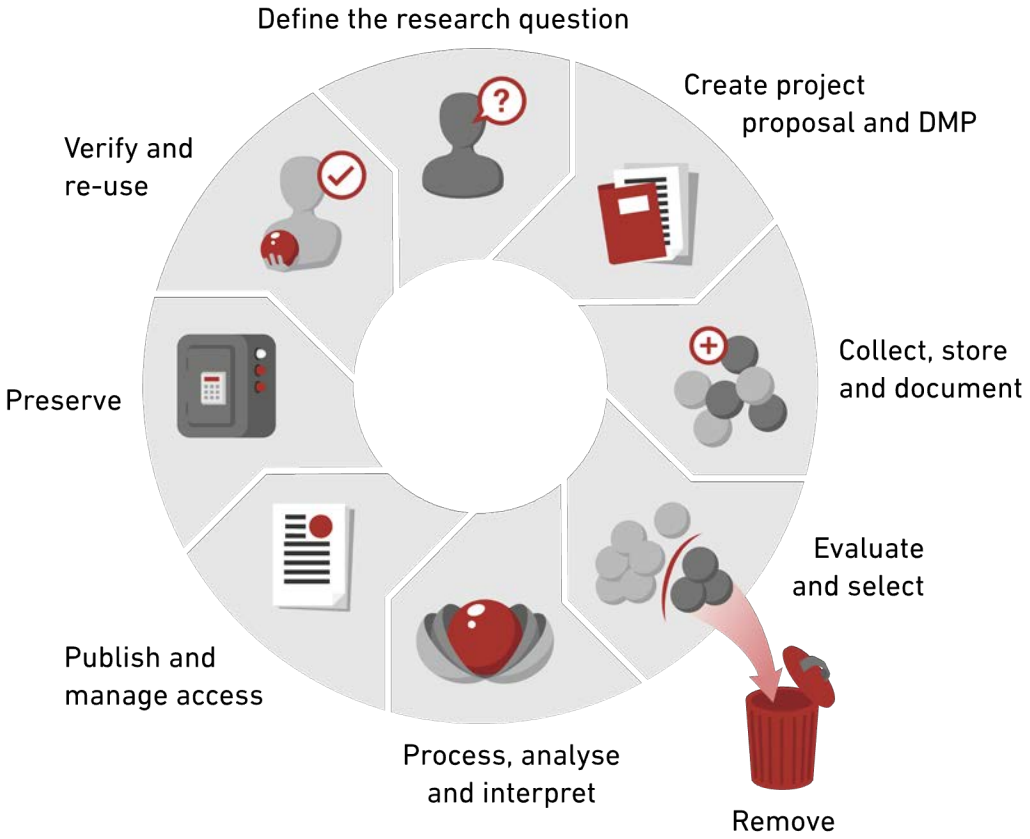
ETH zürich

FAIR practices for microbiome data

Lina Kim and Alan Pacheco

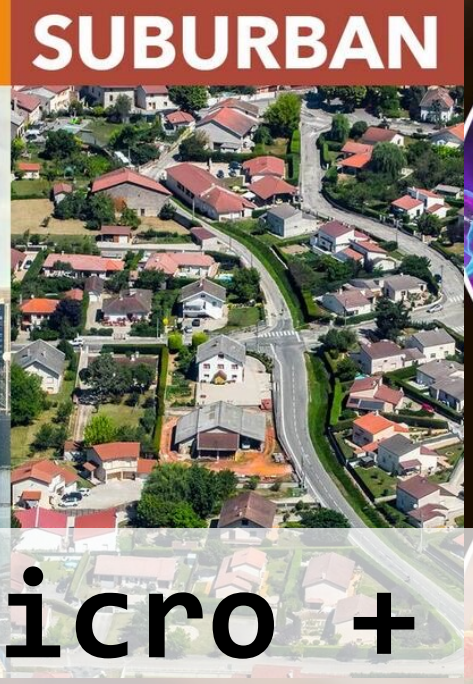
8 May 2023

How do microbiome data fit ORD frameworks?



ETH Library: Research Data Life Cycle

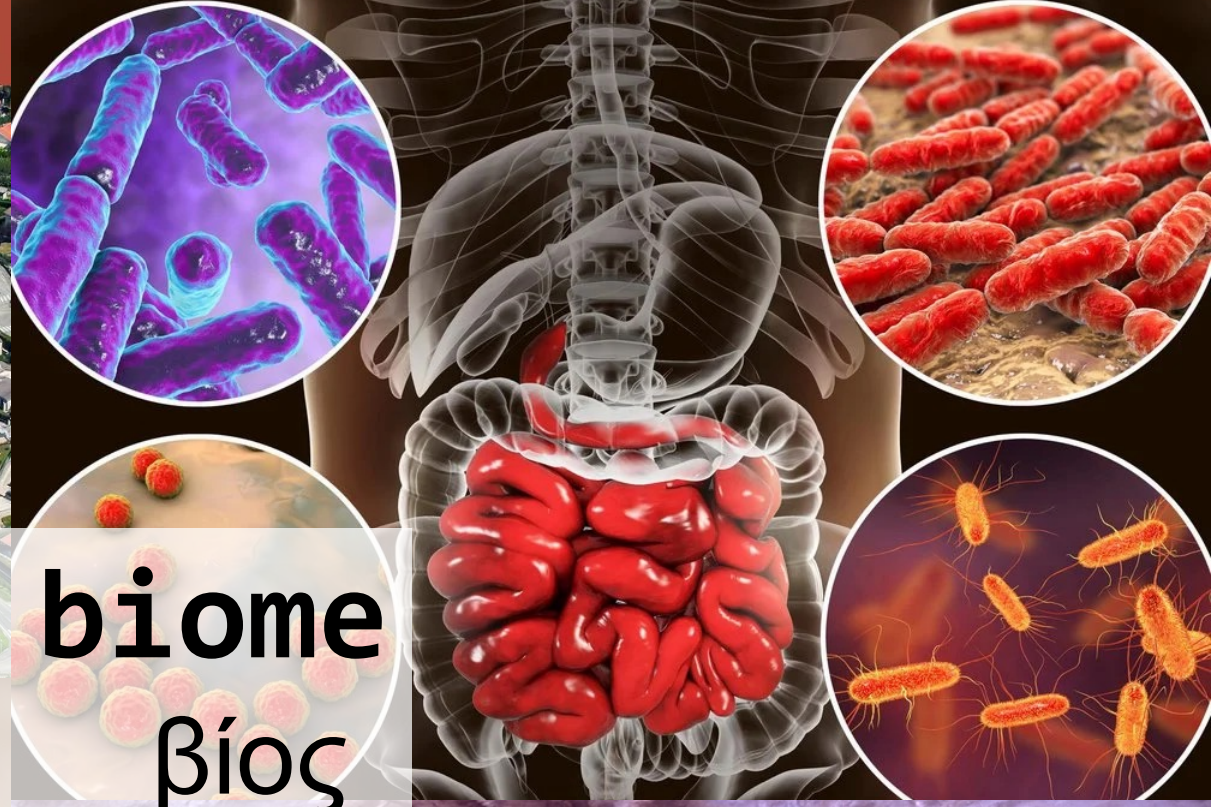
1. **Research context**
   - Goals
   - Methods
   - Data types
2. **ORD challenges and needs**
   - Significance
   - Solutions



micro +

μικρός

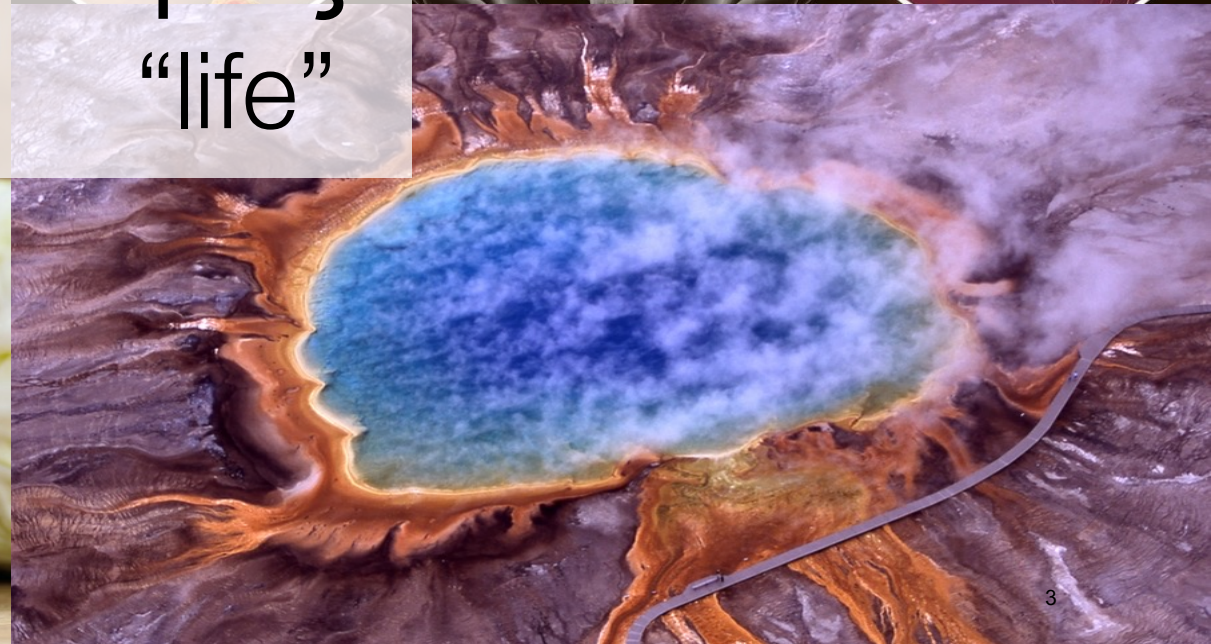
“small”



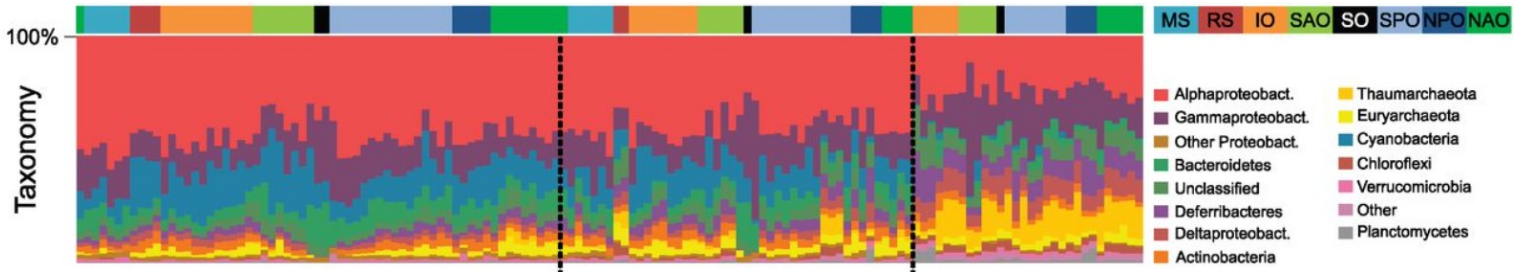
biome

βίος

“life”

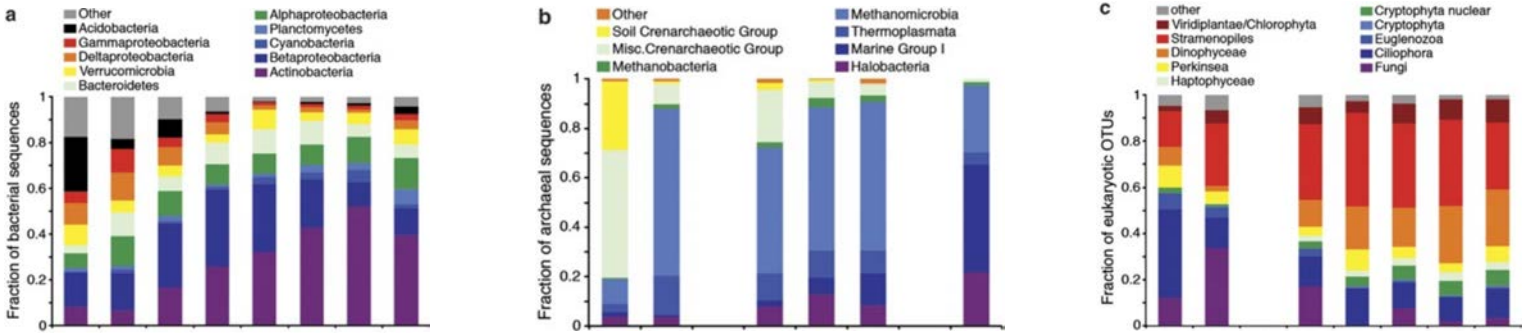


Consortia investigate many different types of microbiomes

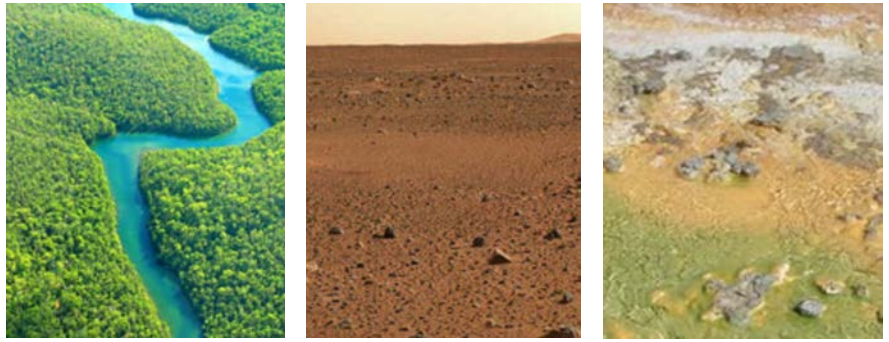


Sunagawa *et al.*, 2015

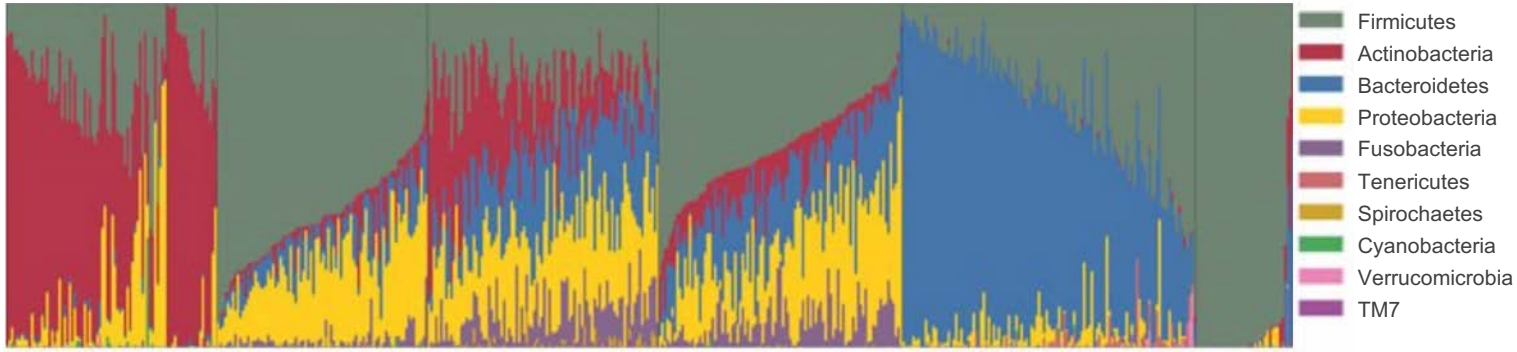
Tara Oceans



Crump *et al.*, 2012



Earth Microbiome Project



The Human Microbiome Project Consortium, 2012

Human Microbiome Project, Nature

Microbiome research: Potential solutions to many world problems

BioSpace

The Brave New World of Microbiome-based Therapies

The Brave New World of Microbiome-based Therapies. Published: Feb 11, 2020 By Mark Terry. Body Surrounded by Microorganisms. The microbiome is the ...

The Boston Globe

The Future of Food: Agriculture's extremely tiny saviors | The Future of Food

It's an attempt to capitalize on a phenomenon that scientists are still trying to fully understand: that the growth of food crops and other plants is heavily influenced ...

Santa Barbara Independent

A Microbe That Might Eat Gasoline

Scientists are examining if the newly found bacteria could perhaps clean up harmful toxins from the oceans. Among them is Kenneth Cullings, ...

Jun 6, 2020

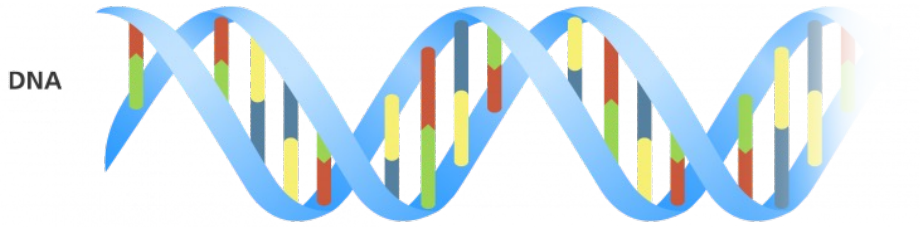
nature

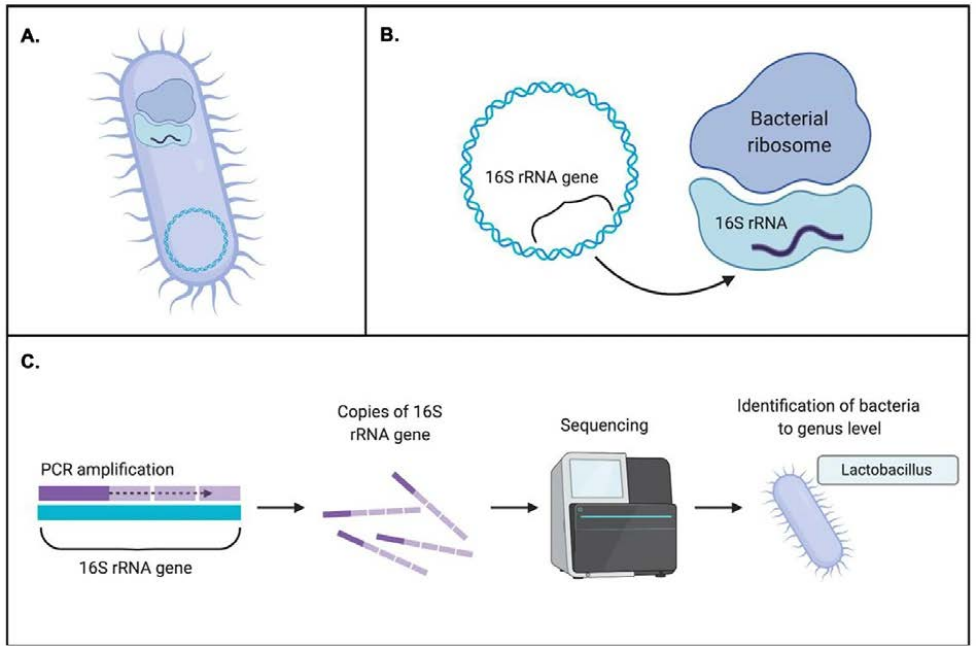
Jan 29, 2020 • 12:00 EST

Fighting Cancer with Microbes

Targeting the microbiome could hold the key to combating a range of malignant diseases.

16S sequencing reveals bacterial composition

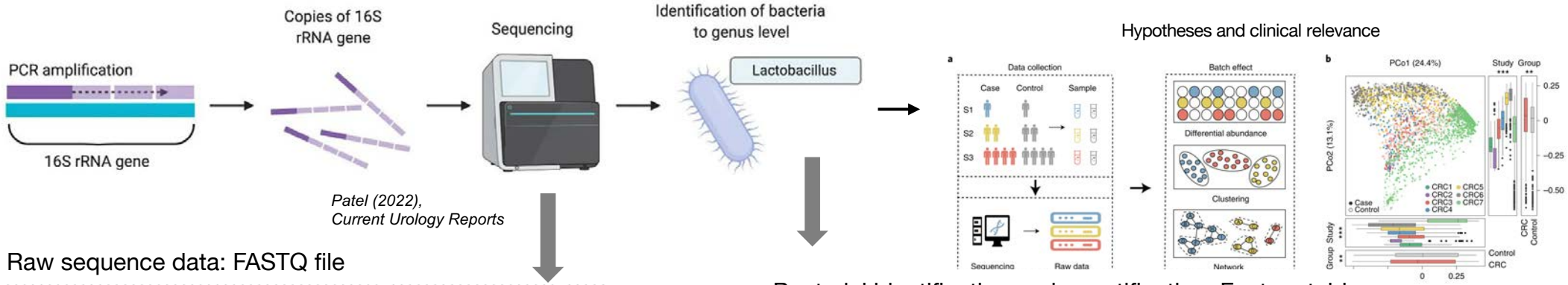




Patel (2022),
Current Urology Reports

16S data come in various forms

Data available for upload



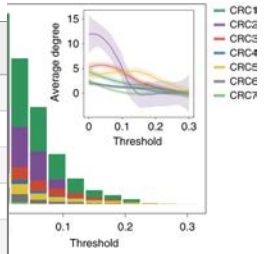
Patel (2022), Current Urology Reports

Raw sequence data: FASTQ file

```
@M00477:2390:000000000-BVV3Y:1:2106:21296:3777 1:N:0:GTAGAG
AGCGTAGACGGCATGGCAAGCCAGATGTGAAAGCCCGGGGCTCAACCCCGGGACTGCATTTGGAACTGTCAG
GCTAGAGTGTCGGAGAGGAAAGCGGAATTCCTAGTGTAGCGGTGAAATGCGTAGATATTAGGAGGAACACCA
GTGGCGAAGGCGGCTTTCTGGACGATGACTGACGTTGAGGCTCGAAAGCGTGGGGAGCAAACAGGATTAGAT
+
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@M00477:2390:000000000-BVV3Y:1:2106:6796:3942 1:N:0:GTAGAG
TGCGTAGGTGGTTTCTTAAGTCAGAGGTGAAAGGCTACGGCTCAACCGTAGTAAGCCTTTGAAACTGGGAAA
CTTGAGTGCAGGAGAGGAGAGTGGAATTCCTAGTGTAGCGGTGAAATGCGTAGATATTAGGAGGAACACCAG
TTGCGAAGGCGGCTCTCTGGACTGTAACTGACACTGAGGCACGAAAGCGTGGGGAGCAAACAGGATTAGATA
+
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGE
FGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGCGGGGGGGGGGGGGGGGGGGGGGGGGA
@M00477:2390:000000000-BVV3Y:1:2106:24737:4004 1:N:0:GTAGAG
AGCGCAGGCGGATTGGCAAGTTGGGAGTGAAATCTATGGGCTCAACCCATAAATTGCTTTCAAAACTGTCAG
TCTTGAGTGGTGTAGAGGTAGGCAGAATTCCCGGTGTAGCGGTGGAATGCGTAGATATCGGGAGGAACACCA
GTGGCGAAGGCGGCCTACTGGGCACTAACTGACGCTGAGGCTCGAACGCATGGGTAGCAAACAGGATTAGAT
+
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGC
FGGGGGGGFGGFFGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGF
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGG*8CCGGGCFFGGGGGGGGGGGGFF7
```

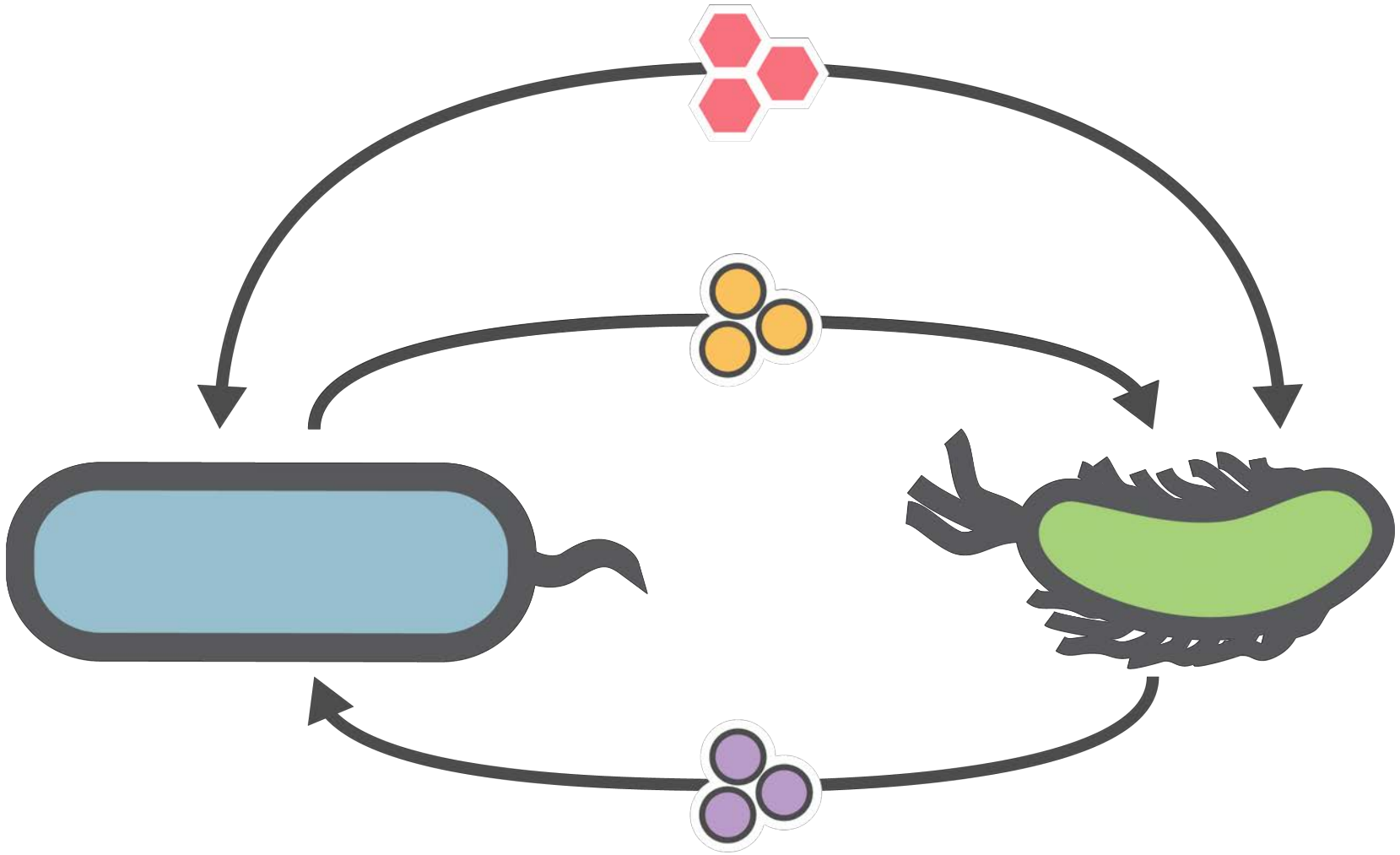
Bacterial identification and quantification: Feature table

	1986030e3c5cd5df6d...	c5adecb440a6bac20...	5e151aa3f416a2fa6f0...
550.L1S109.s.1.sequ...	6.0	0.0	0.0
550.L1S110.s.1.sequ...	0.0	0.0	5.0
550.L1S120.s.1.sequ...	0.0	0.0	0.0
550.L1S13.s.1.seque...	0.0	0.0	0.0
550.L1S131.s.1.sequ...	0.0	0.0	0.0
550.L1S142.s.1.sequ...	0.0	0.0	2105.0
550.L1S143.s.1.sequ...	0.0	0.0	8.0
550.L1S176.s.1.sequ...	0.0	0.0	596.0
550.L1S23.s.1.seque...	0.0	0.0	10.0
550.L1S247.s.1.sequ...	0.0	0.0	0.0
550.L1S277.s.1.sequ...	0.0	0.0	0.0
550.L1S295.s.1.sequ...	0.0	0.0	2.0



Xiao (2022), ...e Computational Science

What are they doing?

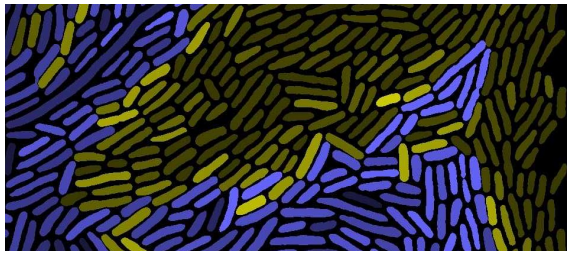


What are they doing?

Kishore *et al.*, 2020.

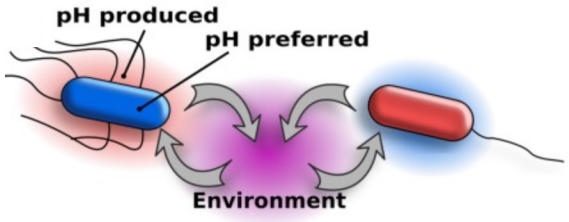
Interactions are highly context-dependent

Space



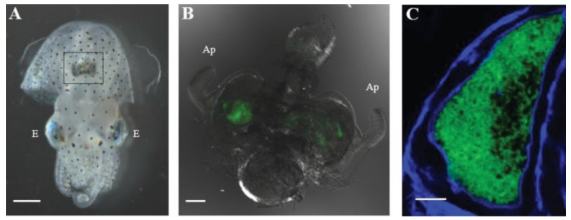
Dal Co *et al.*, 2020

pH



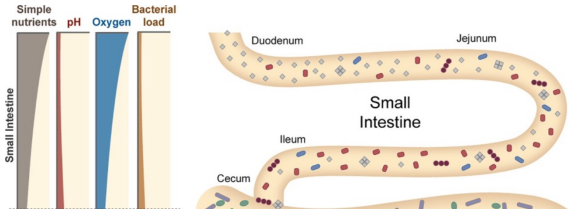
Ratzke and Gore, 2018

Time



Verma and Miyashiro, 2013

Nutrients

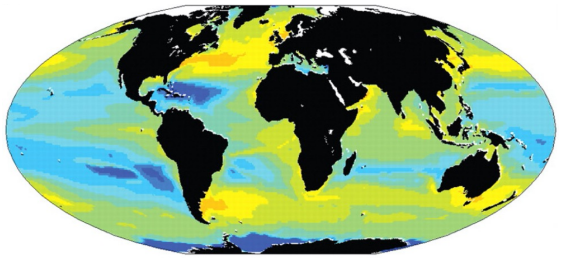


Pereira and Berry, 2017

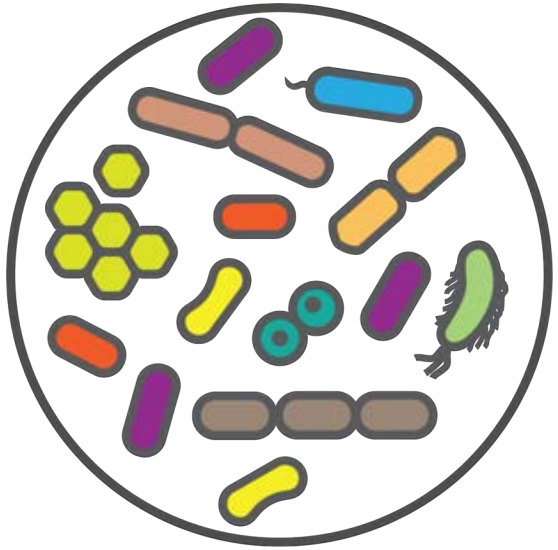
Cost

Adapted from Pacheco *et al.*, 2019

Temperature



Follows *et al.*, 2007



Community composition + function

Toward systematic comparisons of microbial interactions

Daniel Segrè
Boston University

Charlie Pauvert
Uniklinik RWTH Aachen

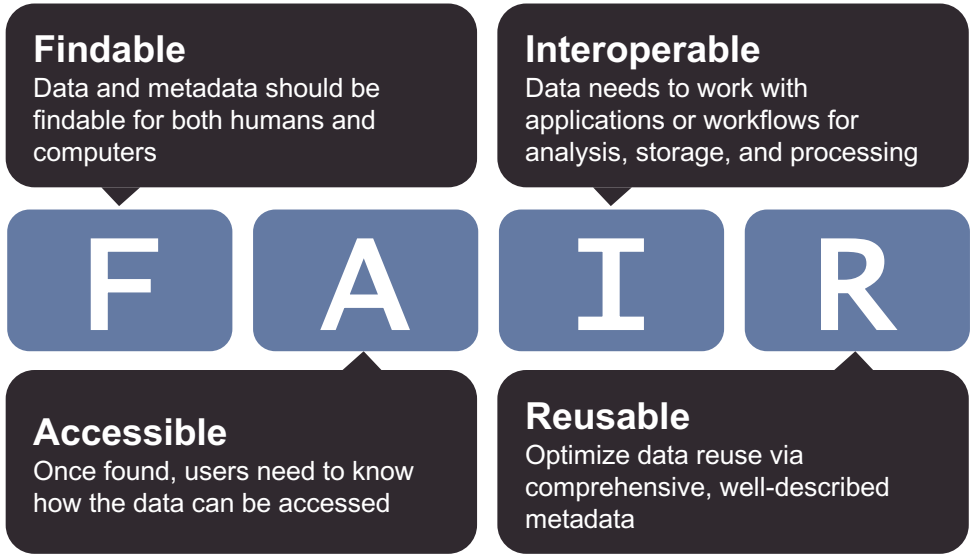
Dileep Kishore
Boston University

Are certain kinds of interactions more prevalent among certain species?

Which microbial network motifs are common across biomes?

Do all strains of my species of interest engage in similar microbial relationships?

Adapted from Pacheco, Pauvert, *et al.*, 2022



Adapted from Wilkinson, *et al.*, 2016

nature

Published: 26 September 2002

Microarray standards at last

nature biotechnology

Correspondence | Published: 02 March 2020

MEMOTE for standardized genome-scale metabolic model testing

nature biotechnology

Published: 06 May 2011

Minimum information about a marker gene se (MIMARKS) and minimum information about sequence (MIxS) specifications

Bioinformatics, 36(24), 2020, 5712–5718
doi: 10.1093/bioinformatics/btaa622
Advance Access Publication Date: 8 July 2020
Letter to the Editor

OXFORD

Databases and ontologies

The Minimum Information about a Molecular Interaction usal STatement (MI2CAST)

an microbiome research:

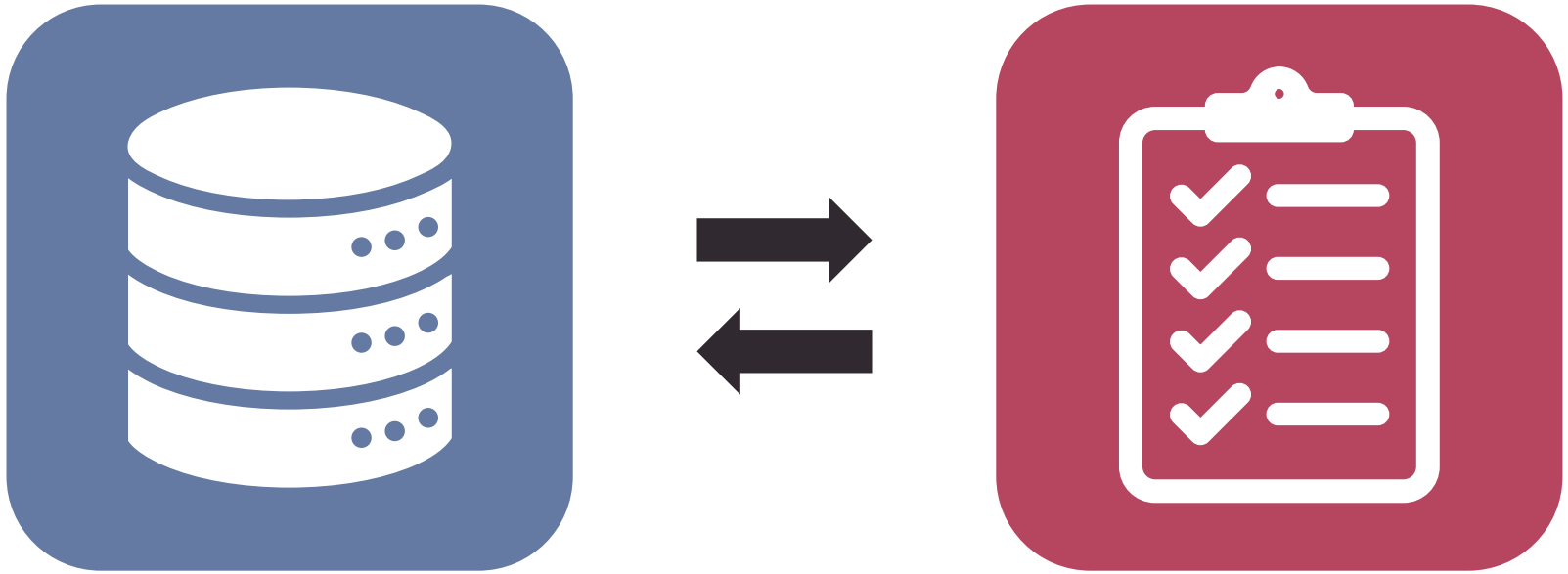
nature biotechnology

BIOINFORMATICS

Vol. 19 no. 4 2003, pages 524–531
DOI: 10.1093/bioinformatics/btg015

The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models

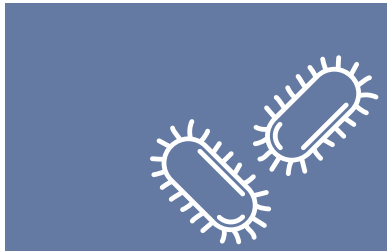
Toward FAIR representations of microbial interactions



Open web catalog (FA)

Minimal metadata requirements (IR)

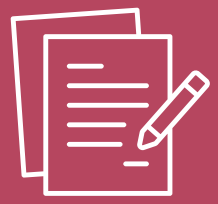
A minimal set of metadata requirements



Microbial entities

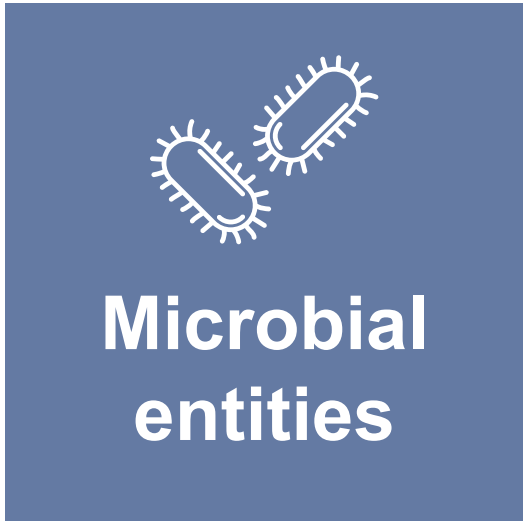
Inference methods

Interaction context



Interaction attributes

A minimal set of metadata requirements



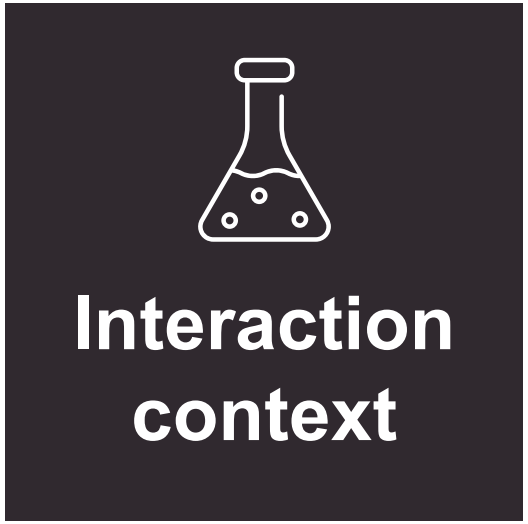
- **Names of participants**
- Identifiers (**taxonomic** and *sequence*)
- Whether it is from a commercial or academic collection or directly isolated

A minimal set of metadata requirements

- **Is the evidence experimental, computational or both?**
- **Reference** (DOI or URL)
- *Type of method*:
  - Simulation
  - Microscopy
  - Cultivation
  - Sample (sequencing)

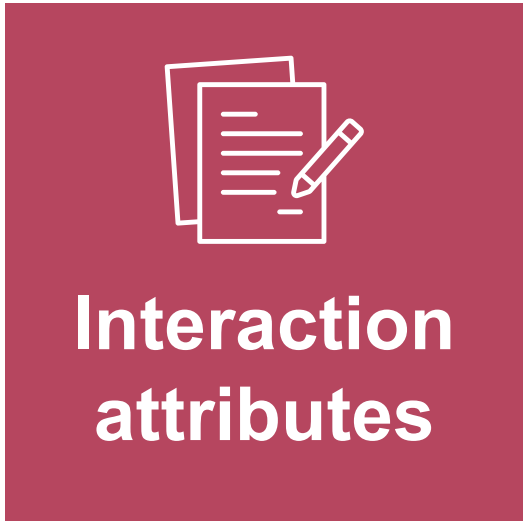
A minimal set of metadata requirements

- *Which biome?*
- Medium/pH/Carbon- or nitrogen-source
- Compounds involved

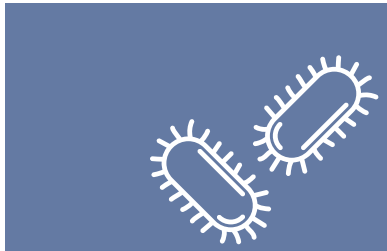


A minimal set of metadata requirements

- *How are participants affected? 0 / 1 / -1*
- Ecological outcome? Co-occurrence, competition
- What are the dependencies? Time, contact, etc.



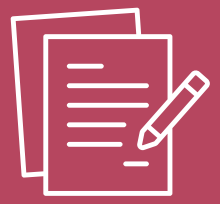
A minimal set of metadata requirements



Microbial entities

Inference methods

Interaction context



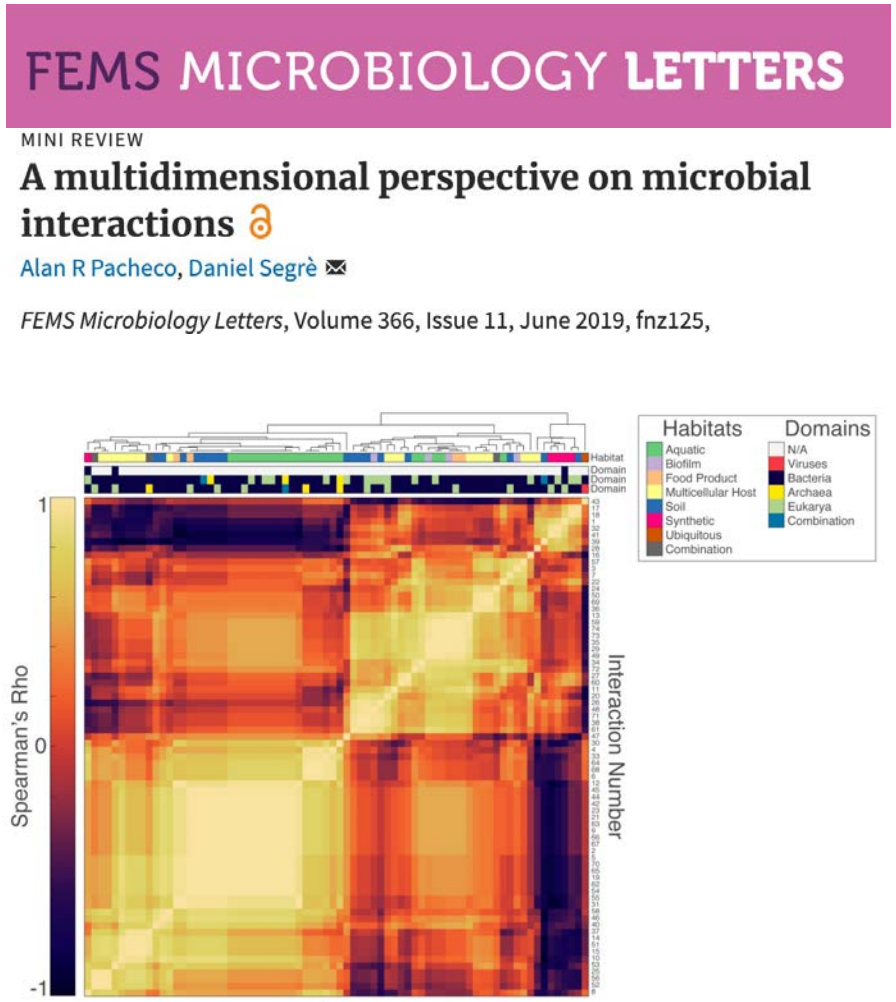
Interaction attributes

A minimal set of metadata requirements

Metadata	Level	Description
A. Which microbial entities are involved?		
participants	M	Comma-separated list of the microbial entities' names, with descriptions of any genetic manipulations performed.
tax_id	M	Comma-separated list of the matching identifiers from the NCBI Taxonomy at the relevant taxonomic level. (e.g., NCBI:txid1043002, NCBI:txid411903). Novel taxa lacking identifiers are denoted by N/A[b].
sequence_id	R	Comma-separated list of the accessions to the matching sequence data (e.g., genome, marker gene sequence). Taxa from presequencing era articles could be denoted by N/A.
env_origin	X	Term from the Environmental Ontology indicating from which biome the microbial entities originate (e.g., soil [ENVO:00001998]).
source_collection	X	Comma-separated list of the source of the participants engaging in this interaction: isolation, commercial collection, academic collection
B. How was the interaction uncovered?		
evidence_type	M	Type of evidence used to determine the interaction using the Evidence and Conclusion Ontology. At least one of the following broader terms are required: exptl [ECO:0000006], computational [ECO:0007672], or both [ECO:0007661].
method_type	R	One or several of the following types of methods used to determine the interaction:
		· Simulation-based (e.g., generalized Lotka-Volterra model, genome-scale metabolic model)
		· Microscopy-based (e.g., co-localization with fluorescent markers, assisted motility)
		· Cultivation-based (e.g., continuous co-culture in bioreactor, co-plating on solid media)
		· Sample-based (e.g., co-occurrences drawn from analyses of abundances obtained from *in situ* or *in vivo* sampling).

Mini-case study of interaction comparisons

- Catalog of 74 interactions from literature
- Each interaction represented as a numerical ‘barcode’ defined by each attribute
- Standardized method of comparing attributes across studies and biomes
- Stepping stone for larger network-driven analyses with more data

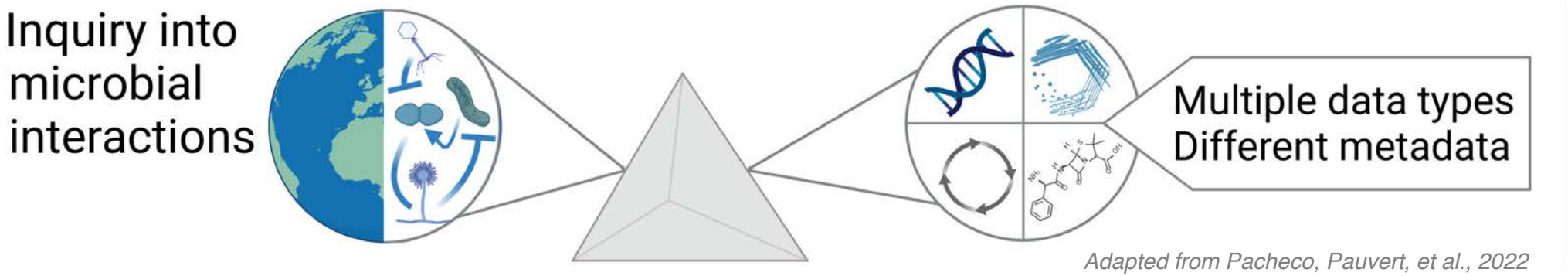


Toward community-driven implementations of FAIR interaction data

+

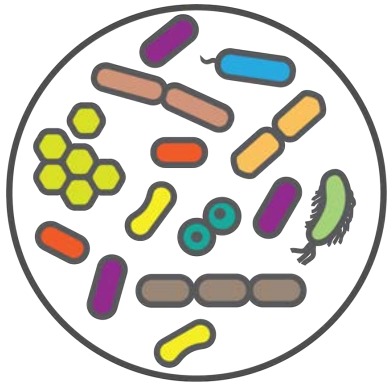
"Minimal information for a microbial interaction"

Looping back from interaction data to sequencing data



Adapted from Pacheco, Pauvert, et al., 2022

Community composition



Raw sequence data: FASTQ file format

```
@M00477:2390:000000000-BVV3Y:1:2106:21296:3777 1:N:0:GTAGAG
AGCGTAGACGGCATGGCAAGCCAGATGTGAAAGCCCGGGGCTCAACCCCGGGACTGCATTTGGAACTGTCAG
GCTAGAGTGTCGGAGAGGAAAGCGGAATTCCTAGTGTAGCGGTGAAATGCGTAGATATTAGGAGGAACACCA
GTGGCGAAGGCGGCTTTCTGGACGATGACTGACGTTGAGGCTCGAAAGCGTGGGGAGCAAACAGGATTAGAT
+
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGFGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGE8FGGGGGGGGGGGGGGGGGGFCEFGGC
@M00477:2390:000000000-BVV3Y:1:2106:6796:3942 1:N:0:GTAGAG
TGCGTAGGTGGTTTCTTAAGTCAGAGGTGAAAGGCTACGGCTCAACCGTAGTAAGCCTTTGAAACTGGGAAA
CTTGAGTGCAGGAGAGGAGAGTGGAATTCCTAGTGTAGCGGTGAAATGCGTAGATATTAGGAGGAACACCAG
TTGCGAAGGCGGCTCTCTGGACTGTAACTGACACTGAGGCACGAAAGCGTGGGGAGCAAACAGGATTAGATA
```

MiShMASh: Evaluating published microbiome data

- Data Availability Statements (DAS) required by journals
  - “Data available upon reasonable request”
  - **7%** of 1,800 publications provided usable data (Gabelica 2022)

- Proj

The datasets generated for this study are available on request to the corresponding author at **<REDACTED>**; procedures to also have sequencing data available in a centralized repository (ENA and/or NCBI SRA archives) are ongoing at the time of publication.

d

original article by the developers [33]. Data from this study have been stored in the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra).

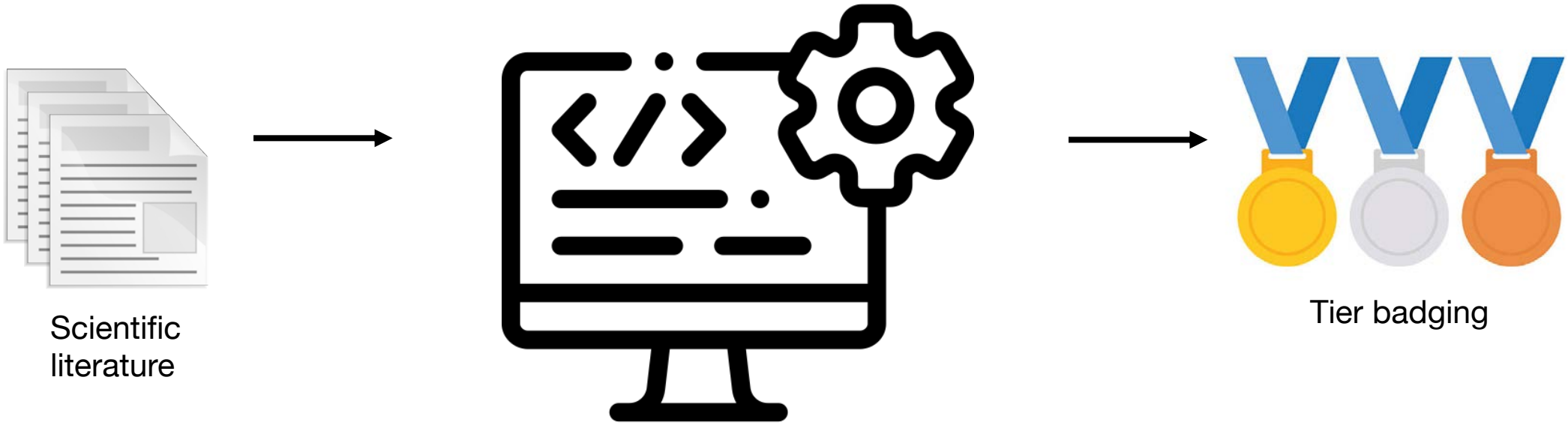
Aim 1: Tier availability for microbiome sequences

- Project MiShMASh: **Mi**crobiome **S**equence and **M**etadata **A**vailability **S**tandards

Bronze	Silver	Gold
	All requirements in Bronze	All requirements in Bronze and Silver
Database named in DAS	Database is under controlled access	Database is publicly accessible
Sequence data are given as FASTQ files or processed feature tables		Sequence data are given as FASTQ files
Primers are given as variable region names	Primers are given as variable region names and primer numbers	Primers are given as sequenced variable region names, primer numbers, and sequences

Aim 2: Automated evaluation of standard adherence

- Software easily validates data accessibility statements to ensure compliance



Scientific literature

Tier badging

- Review microbiome domain: Evaluate current literature

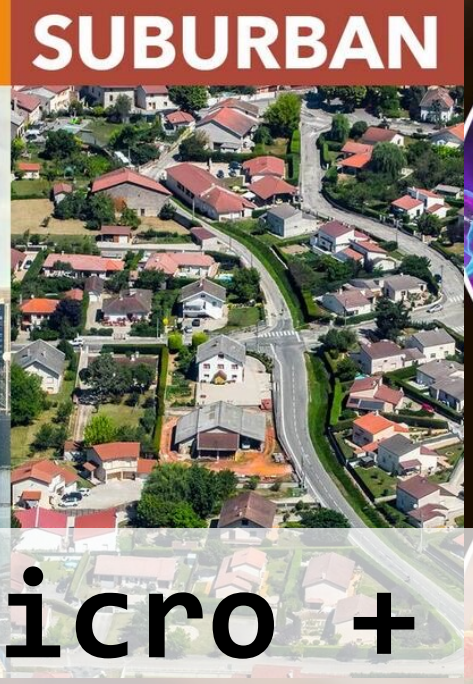
MiShMASh: Evaluating published microbiome data

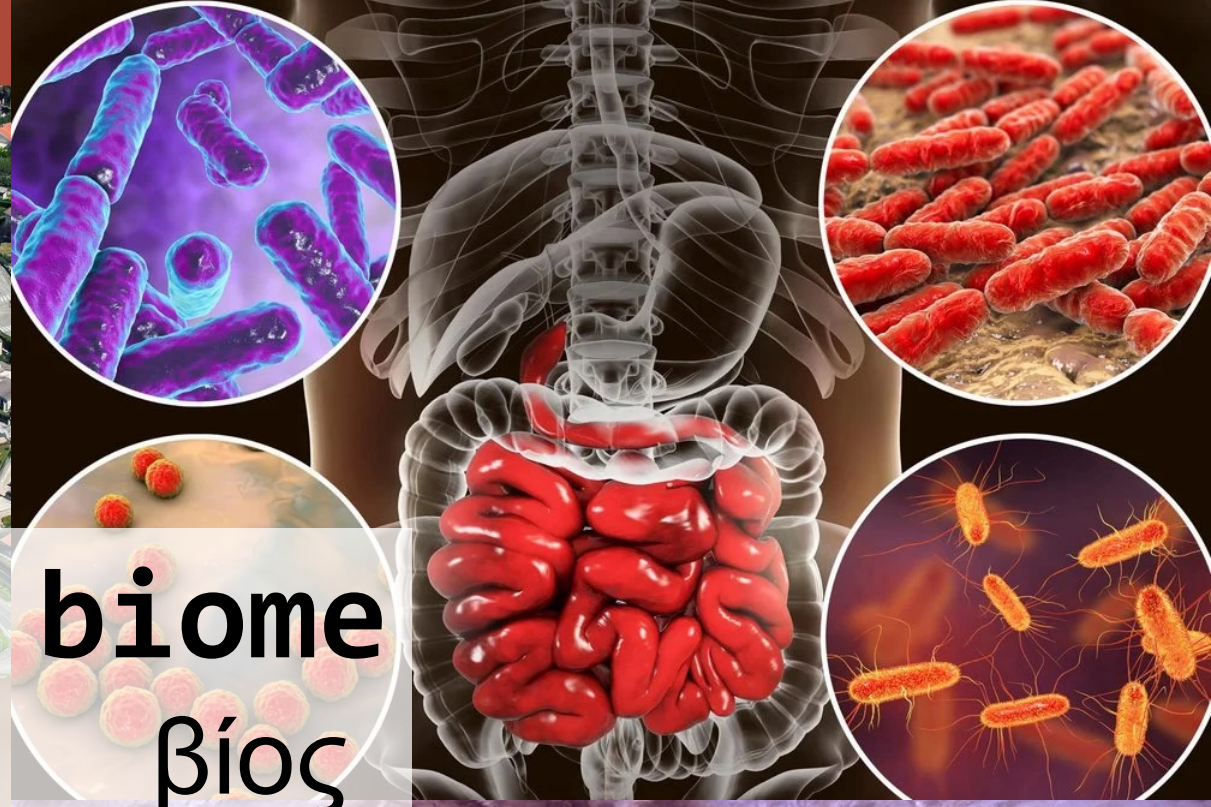
- Project MiShMASh: **Mi**crobiome **S**equence and **M**etadata **A**vailability **S**tandards

1. Develop a tier-based FAIR ORD standard for the field
2. Build software to assess adherence to standards

Impact:

- Self-assess data availability in own publications
- Evaluate meta-analysis data prior to engagement
- Promote conversation and efforts in microbiome ORD

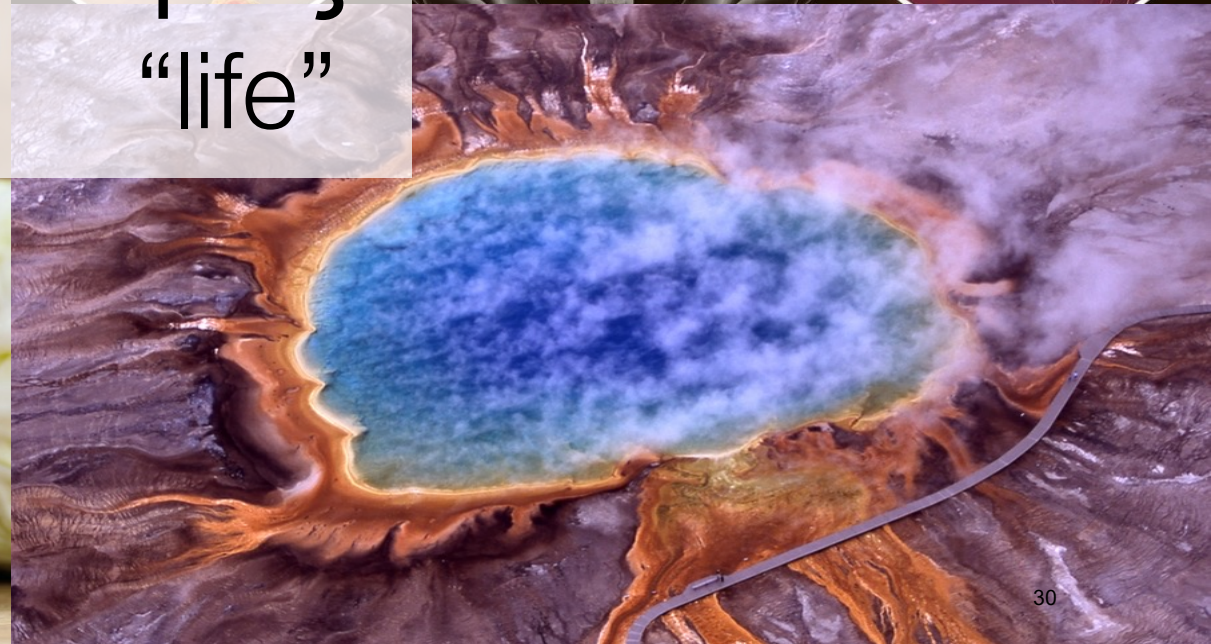




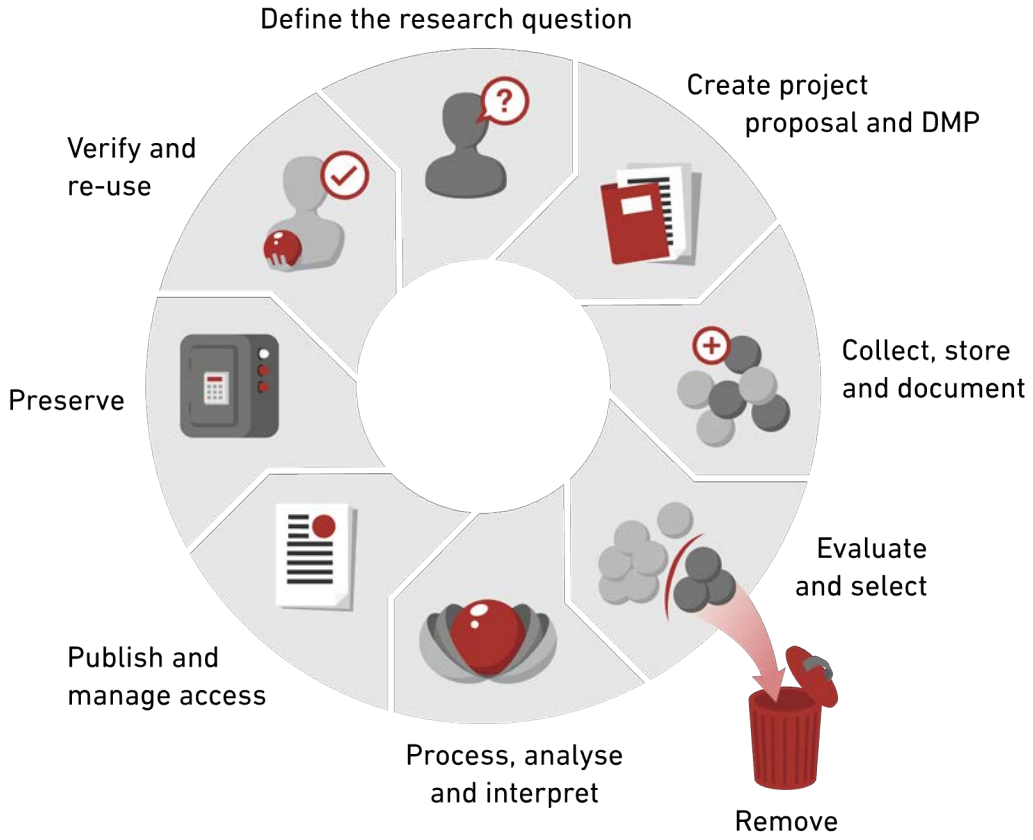
micro + biome

μικρός
“small”

βίος
“life”



How do microbiome data fit ORD frameworks?



ETH Library: Research Data Life Cycle

1. **Research context**
   - Microbiome sequence data
   - Microbial interaction data
2. **FAIR representations**
   - Microbial interaction database for systematic comparisons
   - Tier-based standards for data availability statements

Thank you!

ETH Laboratory of Food Systems Biotechnology
Dr. Anton Lavrinienko
Prof. Dr. Nicholas A. Bokulich

NCCR *Microbiomes*
Dr. Kendra Brown
Prof. Dr. Sara Mitri
Tania Miguel Trabajo

RWTH Aachen
Dr. Charlie Pauvert

Boston University
Prof. Dr. Daniel Segrè
Dr. Dileep Kishore

Figure credits

- Figures not otherwise cited within these slides are stock images or generated from open-source data.
- Slide 3
  - Rural/urban/suburban communities: https://examples.yourdictionary.com/identifying-difference-between-rural-urban-suburban
  - Earth: https://www.worldatlas.com/space/earth.html
  - Pickled spicy foods: https://med.stanford.edu/news/all-news/2021/07/fermented-food-diet-increases-microbiome-diversity-lowers-inflammation
  - Yellowstone: https://news.mit.edu/2018/mit-eaps-research-shows-how-life-survives-extreme-environments-1214
  - Gut: https://www.smithsonianmag.com/science-nature/scientists-find-possible-link-between-gut-bacteria-and-depression-180971411/
- Slide 6
  - Central dogma: https://www.yourgenome.org/wp-content/uploads/2022/04/dna_central_dogma_yourgenome.png