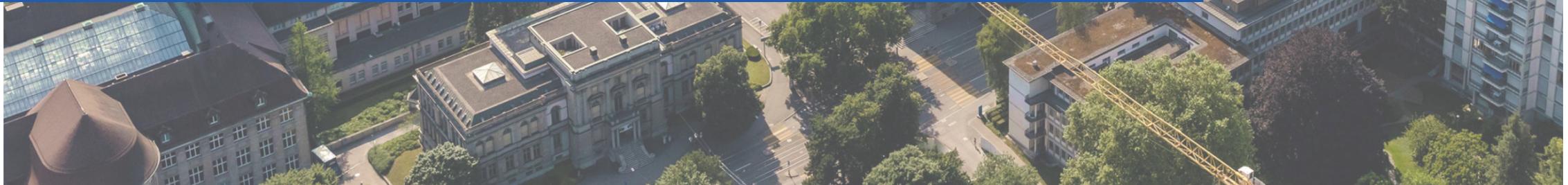
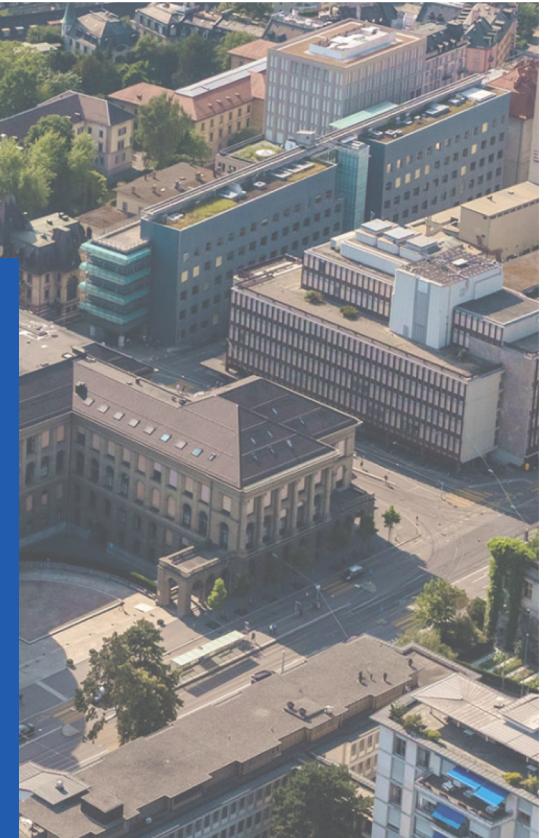


MISHMASH: Microbiome Sequence and Metadata Availability Standards

Lina Kim

January 2026

NCCR Winter School





RURAL

URBAN

SUBURBAN



RURAL

URBAN

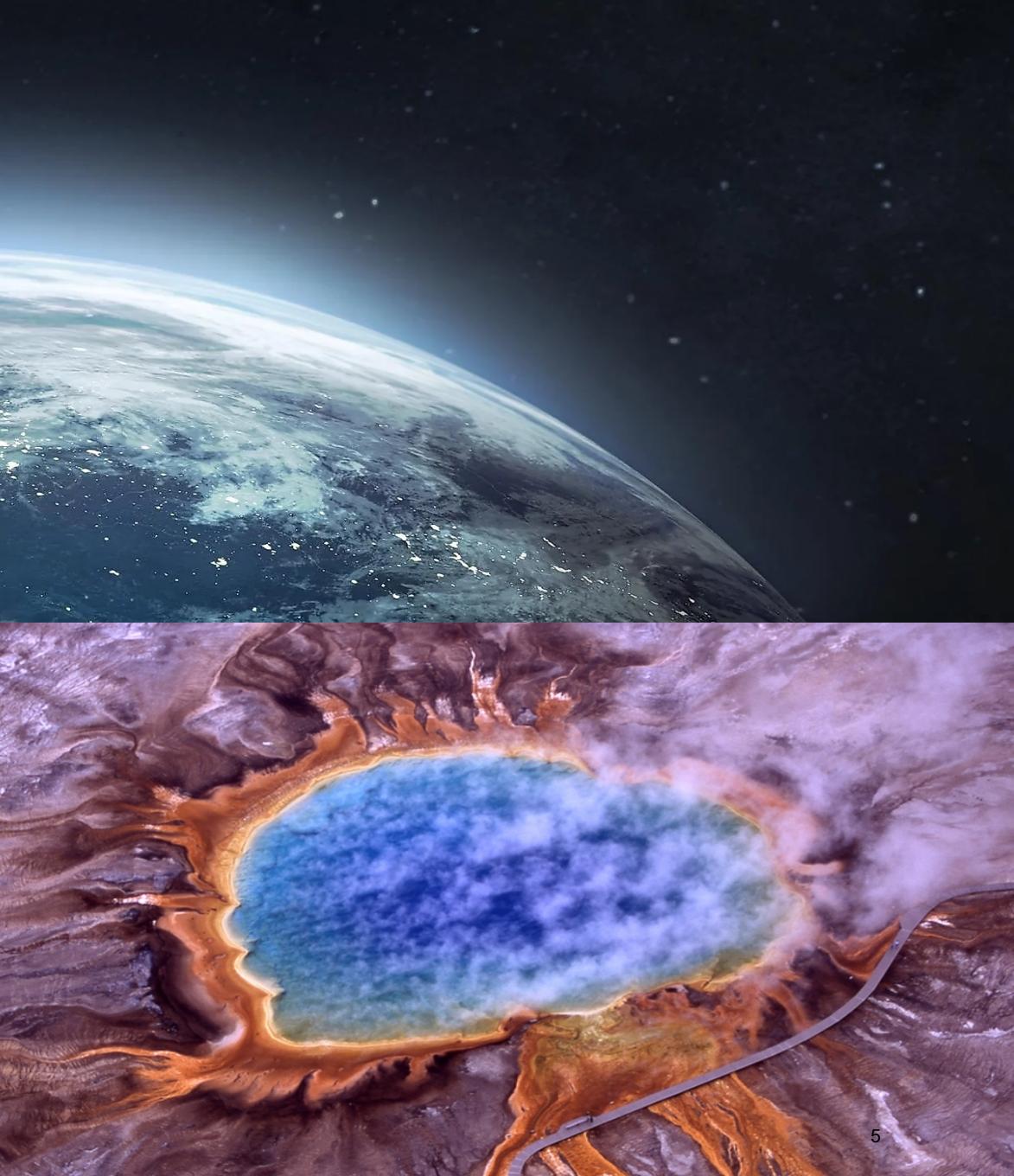
SUBURBAN



RURAL

URBAN

SUBURBAN



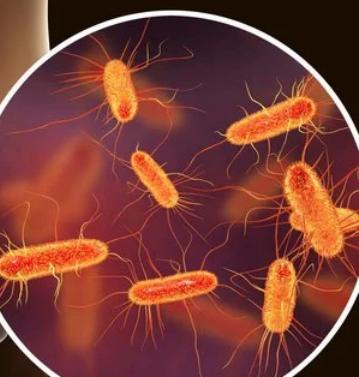
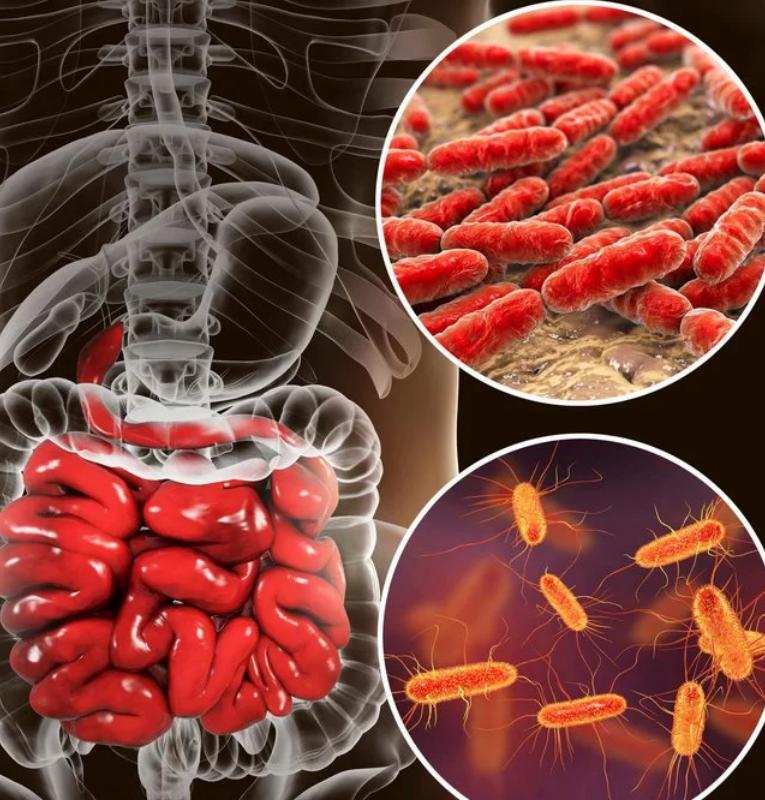
RURAL



URBAN



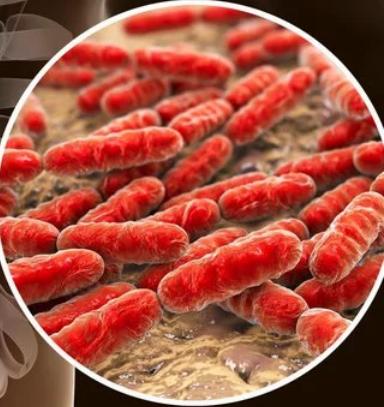
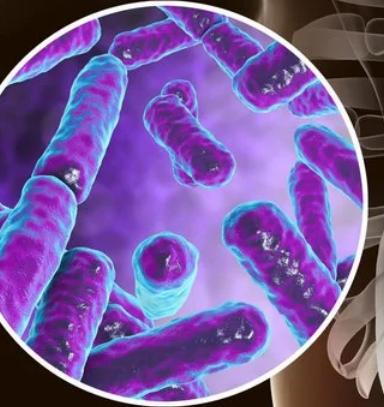
SUBURBAN



RURAL

URBAN

SUBURBAN

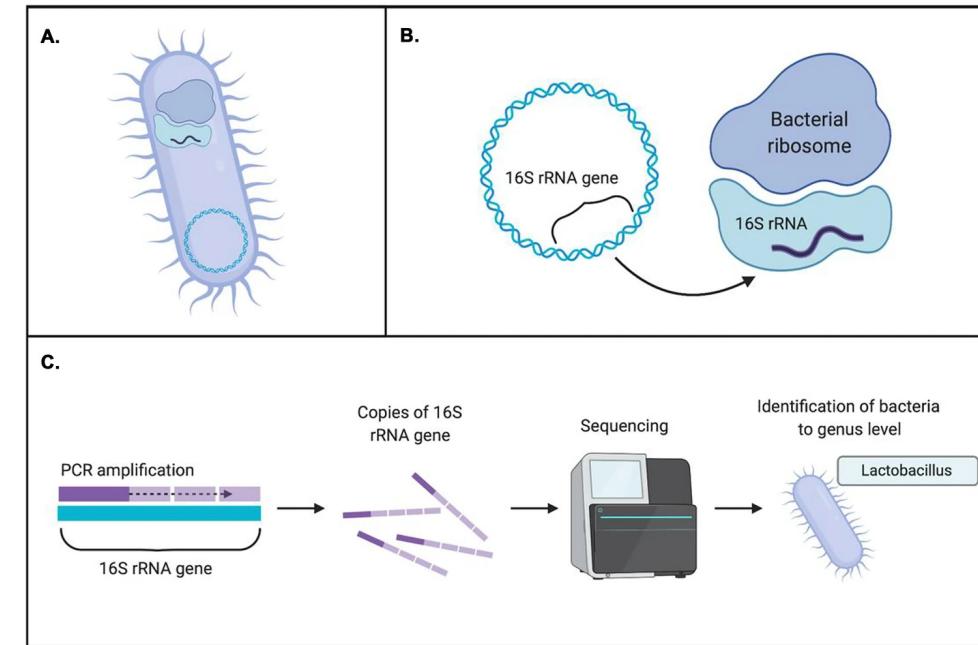
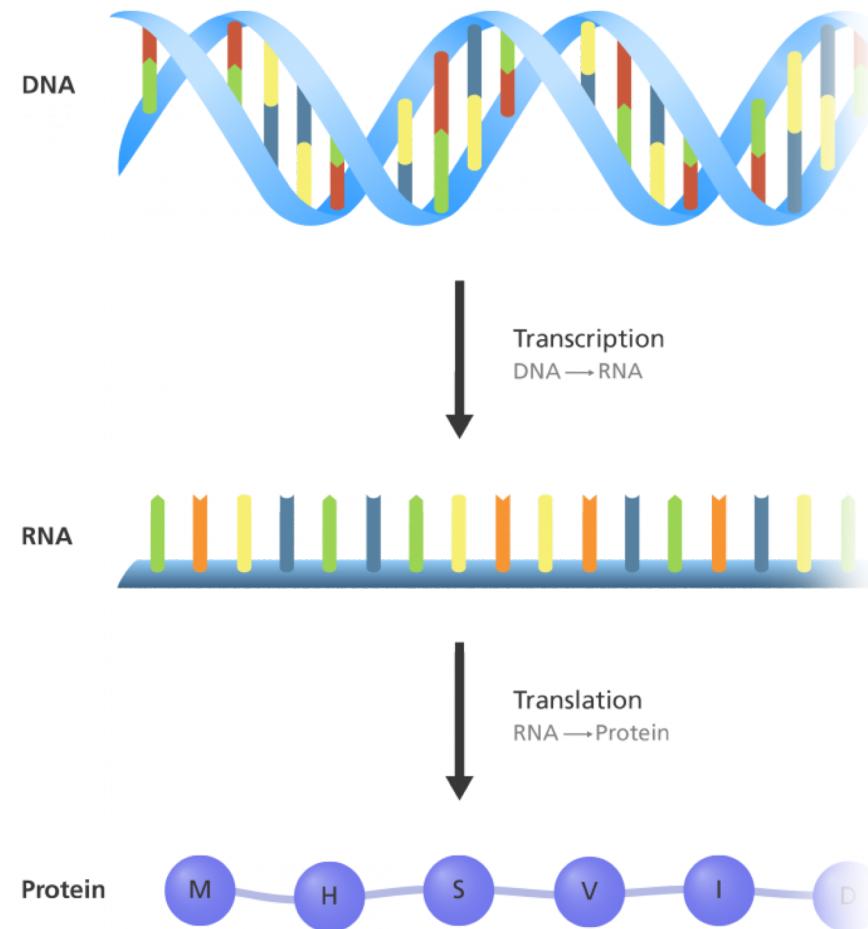


micro +
μικρός
“small”

biome
βίος
“life”

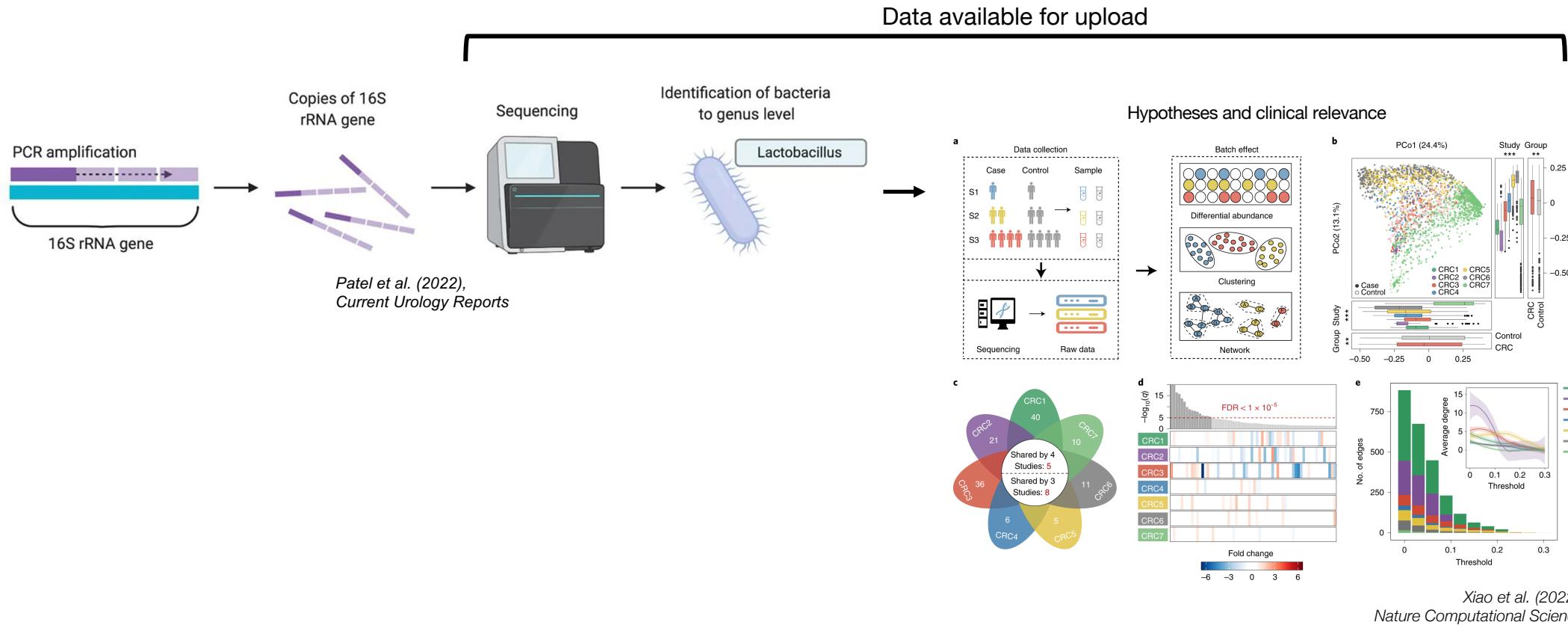


Sequencing reveals bacterial composition and function

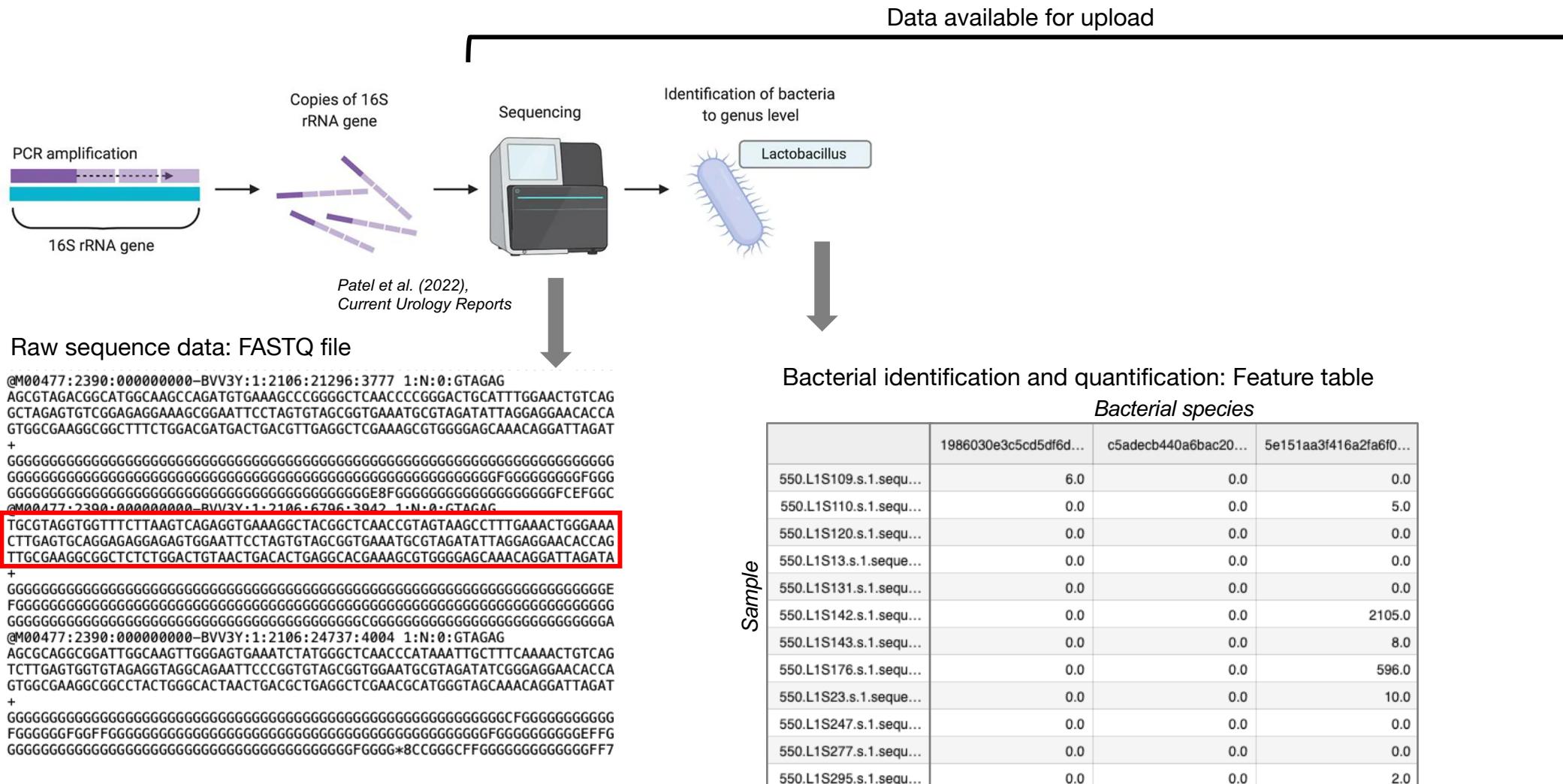


*Patel et al. (2022),
Current Urology Reports*

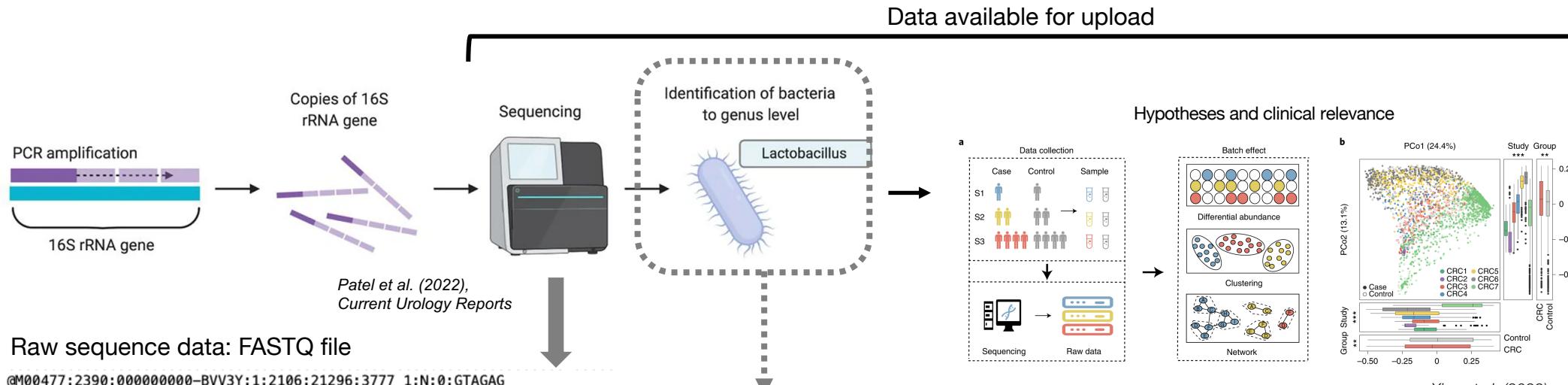
Sequencing data come in various forms



Sequencing data come in various forms

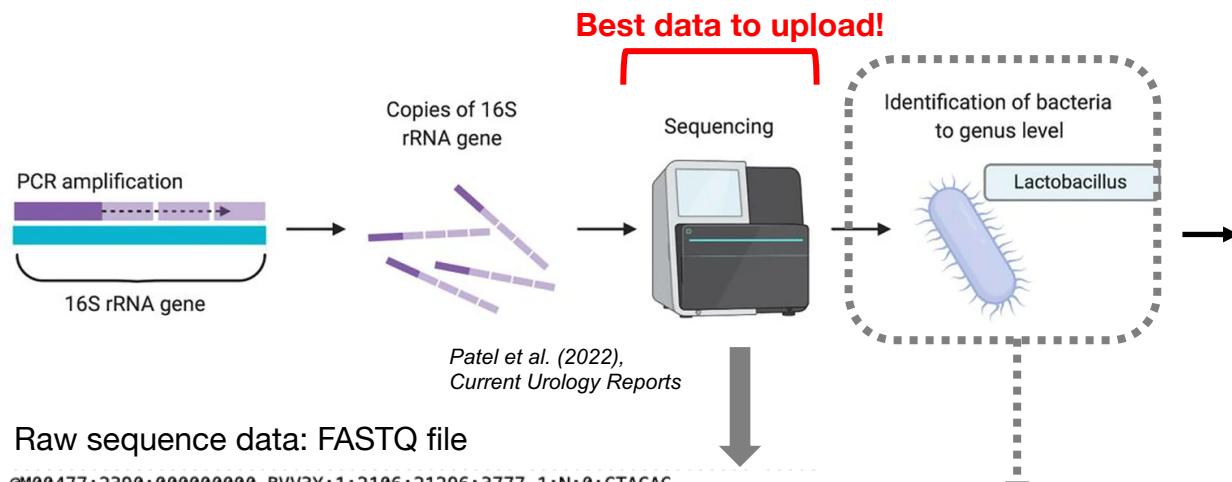


Sequencing data come in various forms

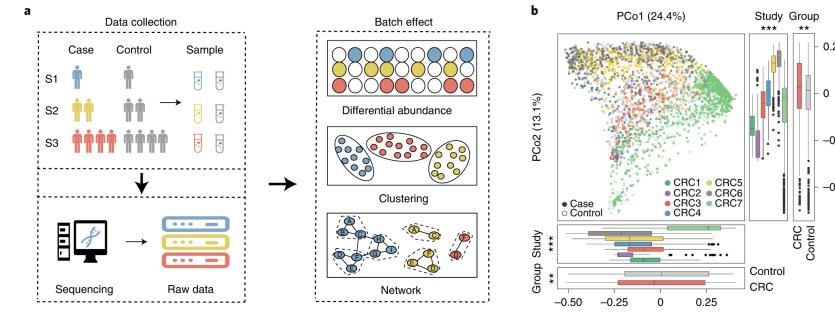


This process has many diverse steps, with many different options for tools and parameters!

Sequencing data come in various forms



Hypotheses and clinical relevance



Xiao et al. (2022), Nature Computational Science

Raw sequence data: FASTQ file

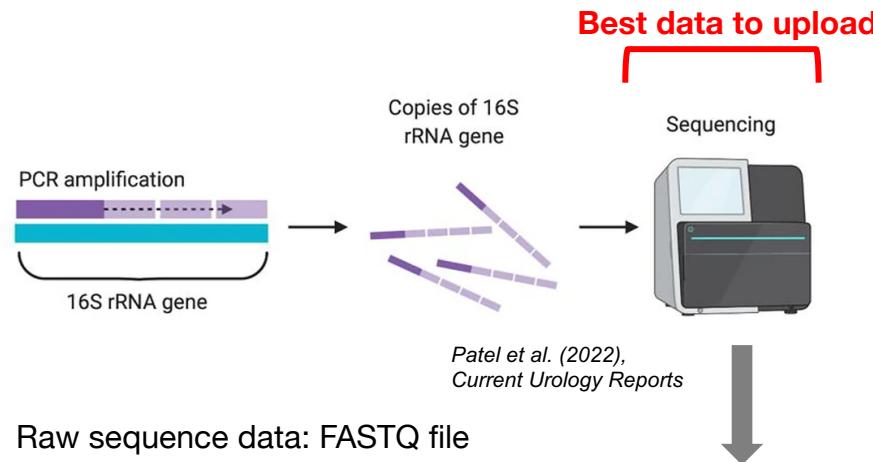
```

@M00477:2390:00000000-BVV3Y:1:2106:21296:3777 1:N:0:GTAGAG
AGCGTAGACGGCATGGCAAGCCAGATGTGAAAGCCCAGGGCTAACCCCGGGACTGCATTGAACTGTCA
GCTAGAGTGTGGAGAGGAAGCGGAATTCTAGTGTAGCGGTGAAATCGTAGATATTAGGAGGAACACCA
GTGGCGAAGGCCGTTCTGGACATGACTGACCTTGAGGCTGAAAGCGTGGGGAGCAAACAGGATTAGAT
+
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@M00477:2390:00000000-BVV3Y:1:2106:6706:3042 1:N:0:GTAGAG
TGGCTAGGTGTTCTTAAGTCAGAGGTAAAGGCTACGGCTAACCGTAGTAAGCTTGAACACTGGAAA
CTTGAGTGCAGGAGAGGAGAGTGGAAATCCTAGTGTAGCGGTGAAATCGTAGATATTAGGAGGAACACCA
TTGCGAAGGCCGCTCTGACTGAACTGACACTGAGGACGAAAGCGTGGGGAGCAAACAGGATTAGATA
+
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
FGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@M00477:2390:00000000-BVV3Y:1:2106:24737:4004 1:N:0:GTAGAG
AGCCAGCGGATTGGCAAGTTGGAGTGAATCTATGGGCTAACCCATAATTGCTTCAAACACTGTCA
TCTTGAATGGTAGAGGTAGGAGAATTCCGGTGTAGCGGTGAAATCGTAGATATCGGGAGGAACACCA
GTGGCGAAGGCCGCTACTGGCACTAAGTGCAGCTGAGGCTGAAACCGTAGCAAACAGGATTAGAT
+
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
FGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG

```

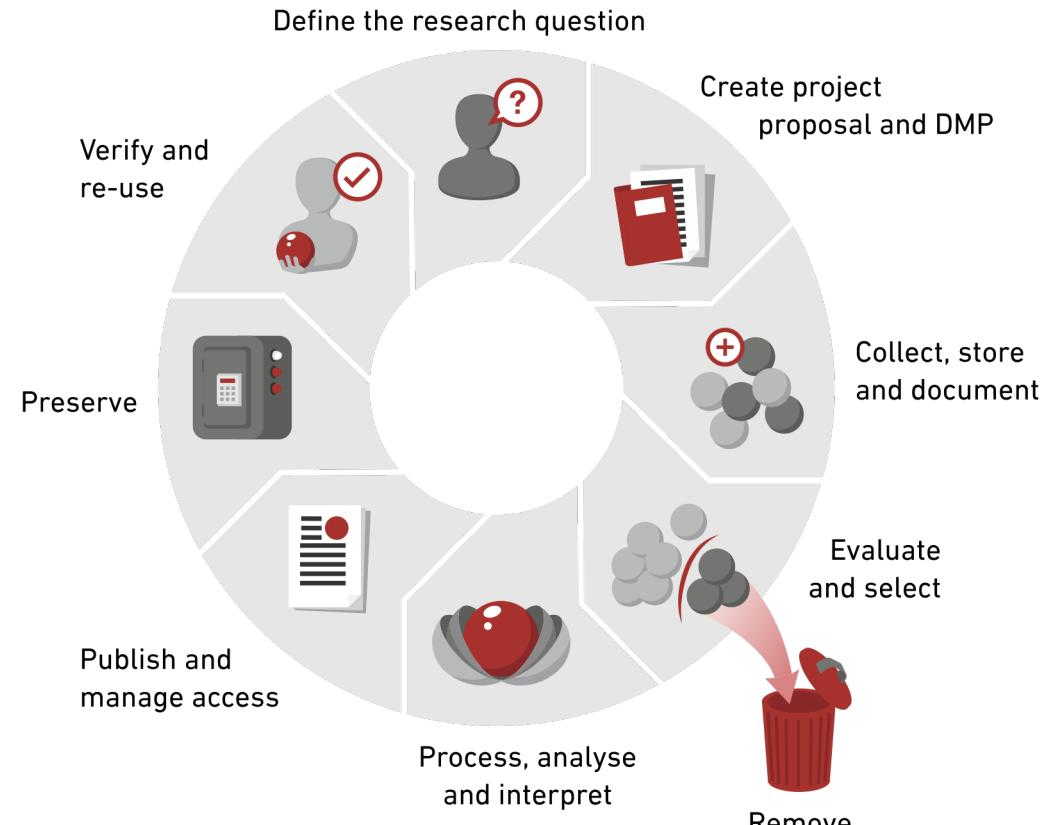
This process has many diverse steps, with many different options for tools and parameters!

Sequencing data come in various forms



Raw sequence data: FASTQ file

```
@M00477:2390:00000000-BV3Y:1:2106:21296:3777 1:N:0:GTAGAG
AGCGTAGACGGCATGCCAACCCGGGCTCAACCCGGGACTGCATTGAACTGTCA
GCTAGAGTGTGGAGAGGAAGCGGAATTCTAGTGTAGCGGTGAAATCGTAGATATTAGGAGGAACCCA
GTGGCGAAGGCCTTCTGGACATGACTGACCTTGAGGCTCAAAGCGTGGGAGCAAACAGGATTAGAT
+
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@M00477:2390:00000000-BV3Y:1:2106:6704:3042 1:N:0:GTAGAG
TGGCTAGGTGTTCTTAAGTCAGAGGTAAAGGCTACGGCTAACCGTAGTAAGCCTTGAACACTGGAAA
CTTGAGTGCAGGAGAGGAGTGGAAATTCTAGTGTAGCGGTGAAATCGTAGATATTAGGAGGAACACCA
TTGCGAAGGCCTCTGACTGACACTGAGGACGAAAGCGTGGGAGCAAACAGGATTAGATA
+
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
FGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@M00477:2390:00000000-BV3Y:1:2106:24737:4004 1:N:0:GTAGAG
AGCCAGCGGATTGCCAAGTTGGAGTGAATCTATGGGCTAACCCATAATTGCTTCAAACACTGTCA
TCTTGAAGTGTAGAGGTAGGAGAATTCCCGGTGAGCGGTGAAATCGTAGATATCGGGAGGAACACCA
GTGGCGAAGGCCTACTGGCACTGACCTGAGGCTGAGGCTGAAACAGGATTAGAT
+
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
FGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
```



Evaluating published microbiome data

- Data Availability Statements (DAS) required by journals
 - “Data available upon reasonable request”

Evaluating published microbiome data

- Data Availability Statements (DAS) required by journals
 - “Data available upon reasonable request”

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.

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>).

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)**

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.

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>).

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)**
- Project MiShMASh: **Microbiome Sequence and Metadata Availability Standards**



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

Aim 1: Tier availability for microbiome sequence analysis

For each paper, assign a badge (bronze, silver, gold) based on sequencing data availability:

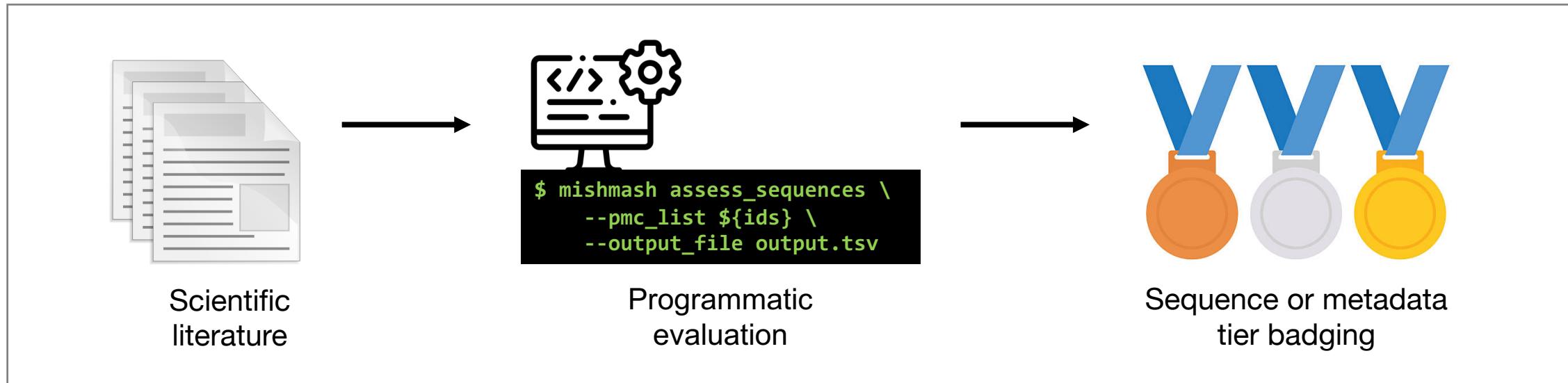
Requirement Type	Bronze	Silver	Gold
Data accessibility	Sequence data are downloadable and not paywalled	Sequence data are downloadable and not paywalled	Sequence data are downloadable and not paywalled
Data accessibility	Accession numbers provided	Accession numbers provided	Accession numbers provided
Data accessibility	Raw sequences given as BCL or FASTQ	Raw sequences given as BCL or FASTQ	Raw sequences given as BCL or FASTQ
Sample processing metadata		Sequencing method provided: 16S or metagenomics	Sequencing method provided: 16S or metagenomics
Sample processing metadata		PCR primer sequences	PCR primer sequences
Data accessibility			If data are not public, DAS clarifies access requirements
Data accessibility			Database is publicly accessible, with no login and/or institutional affiliation e.g. SRA or Figshare
Data processing metadata			Code provided

Each badge mandates all requirements from the previous tier!

Note that “No badge” is also an assessment option.

Aim 2: Automated evaluation of standard adherence

Software easily validates data accessibility statements to ensure compliance

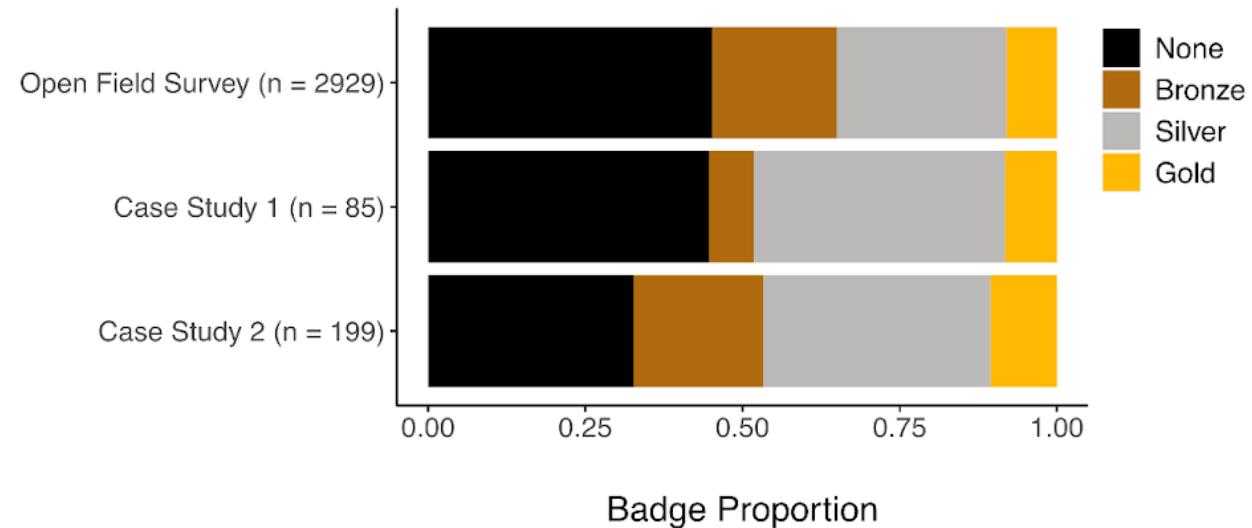
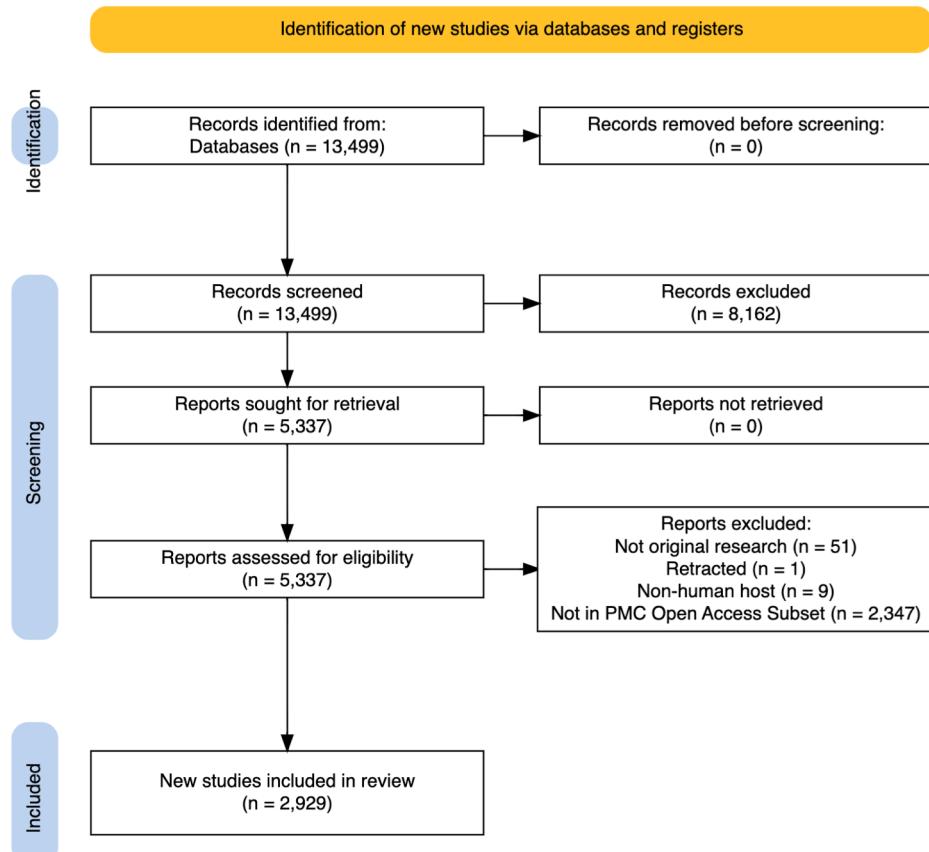


Programmatic tool encourages:

- Researchers to self-assess data availability in own publications
- Users to evaluate meta-analysis data prior to engagement
- Journal editors or grant agencies to verify data FAIRness in stakeholder publications

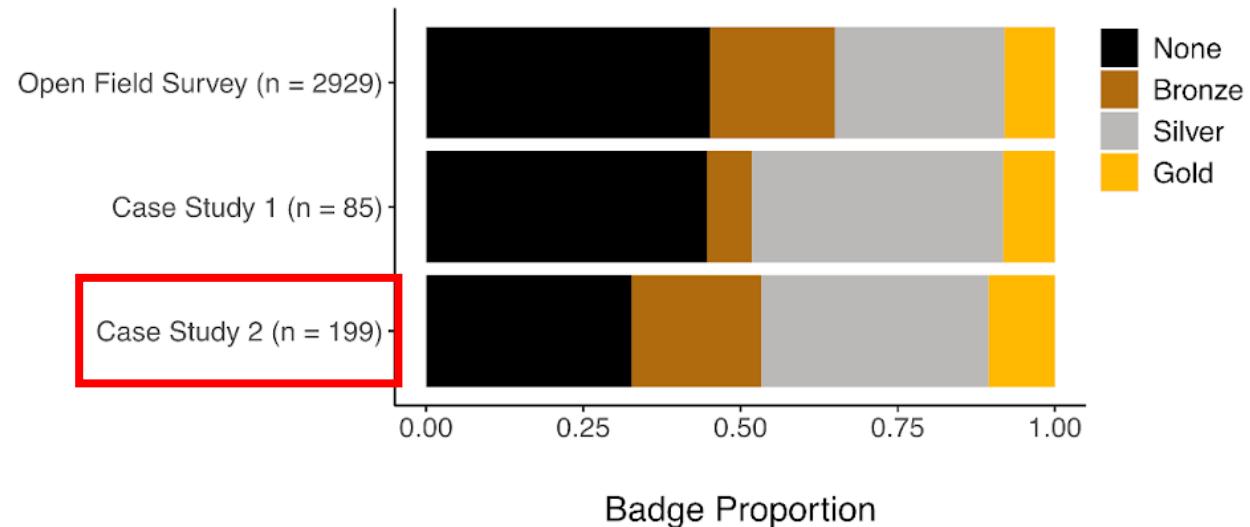
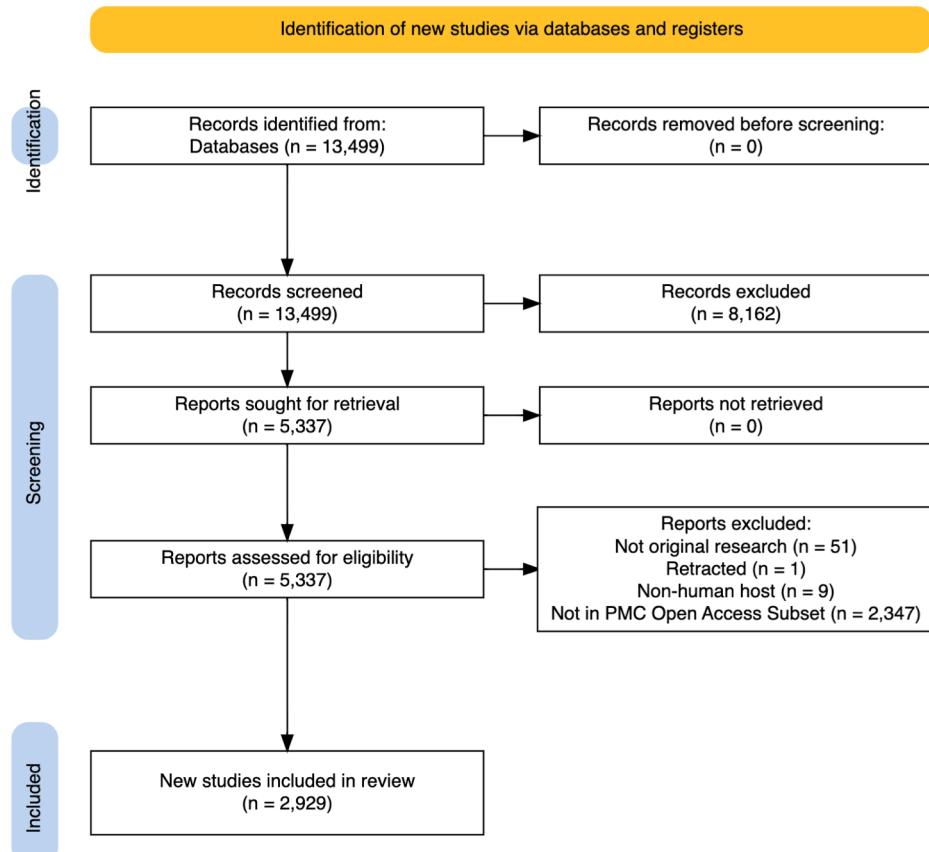
Aim 2: Automated evaluation of standard adherence

Open field survey of 2,929 human gut microbiome papers over 20 years reveals shockingly high proportion of papers with no accessible sequence data!



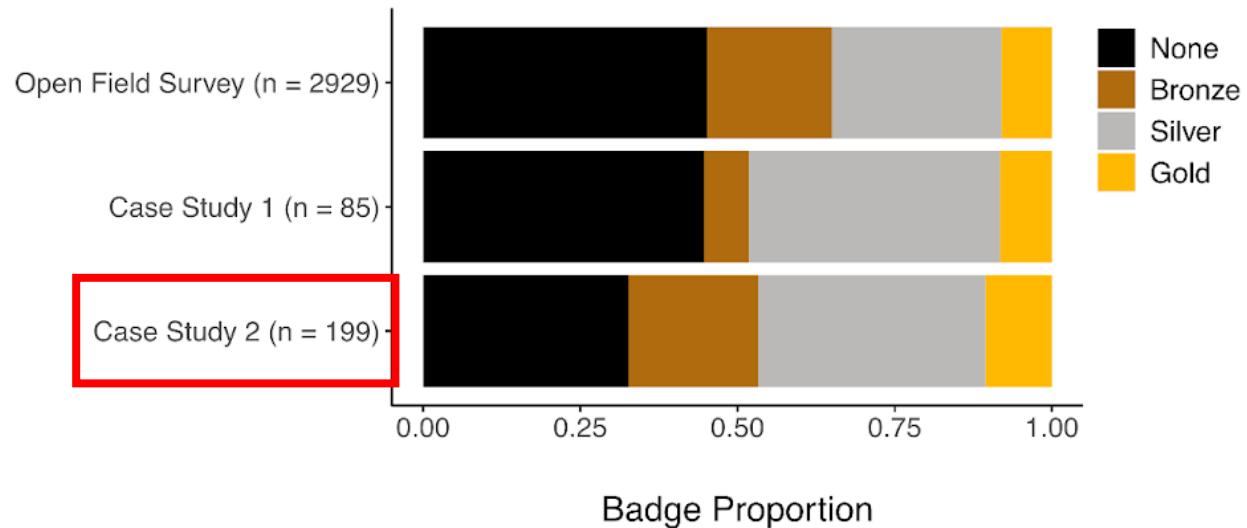
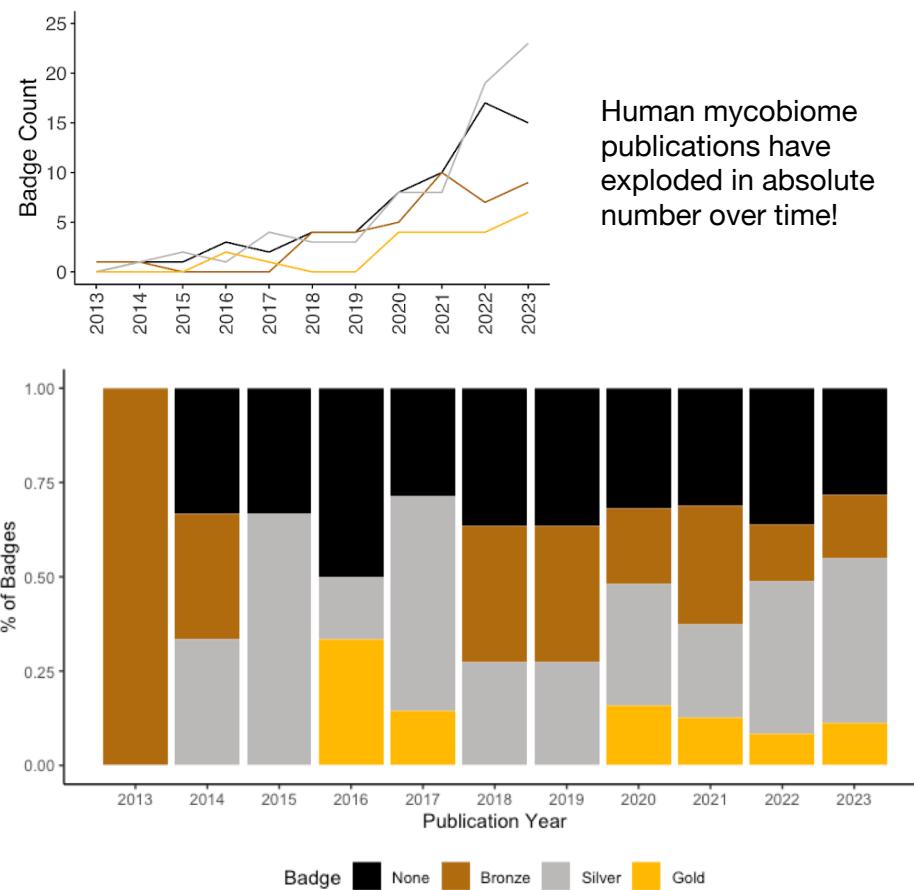
Aim 2: Automated evaluation of standard adherence

Open field survey of 2,929 human gut microbiome papers over 20 years reveals shockingly high proportion of papers with no accessible sequence data!



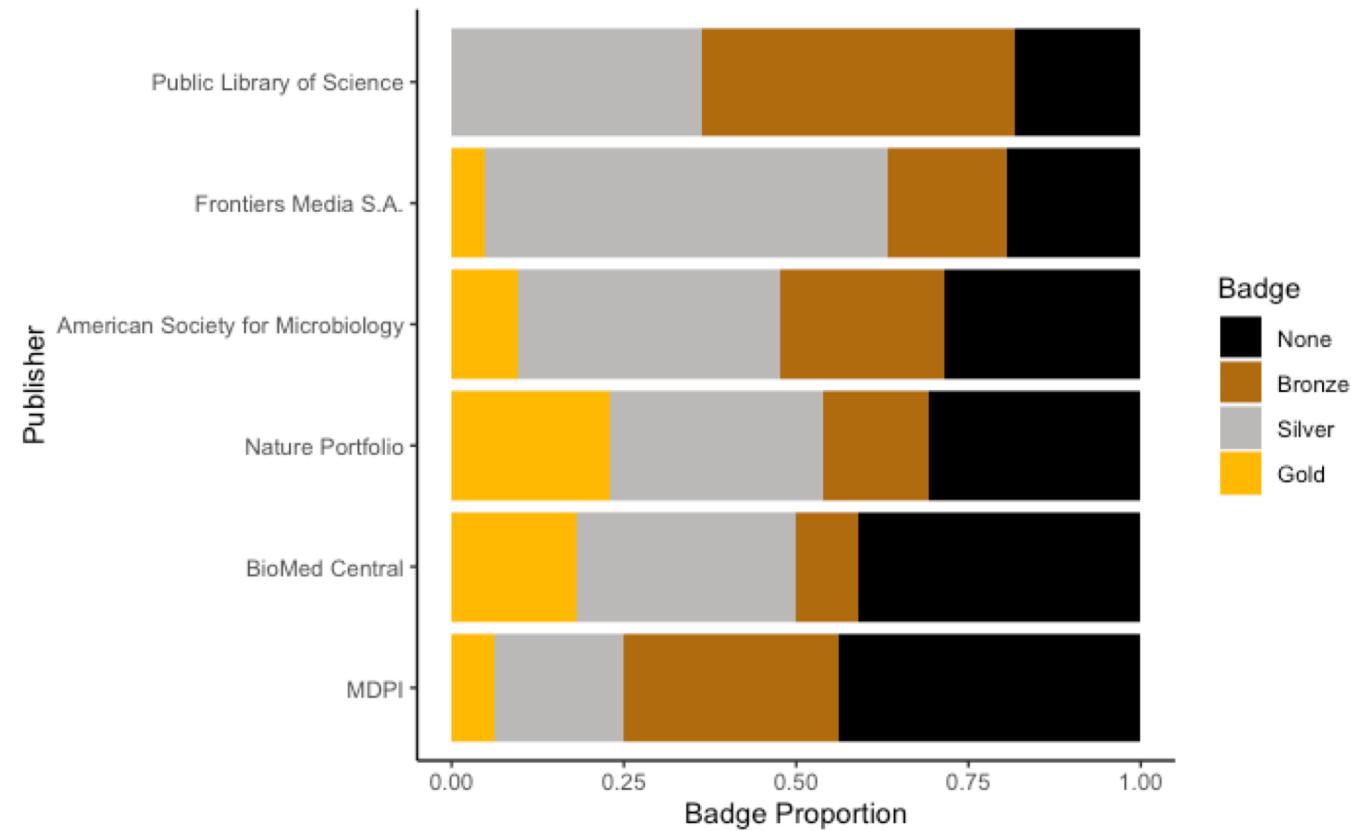
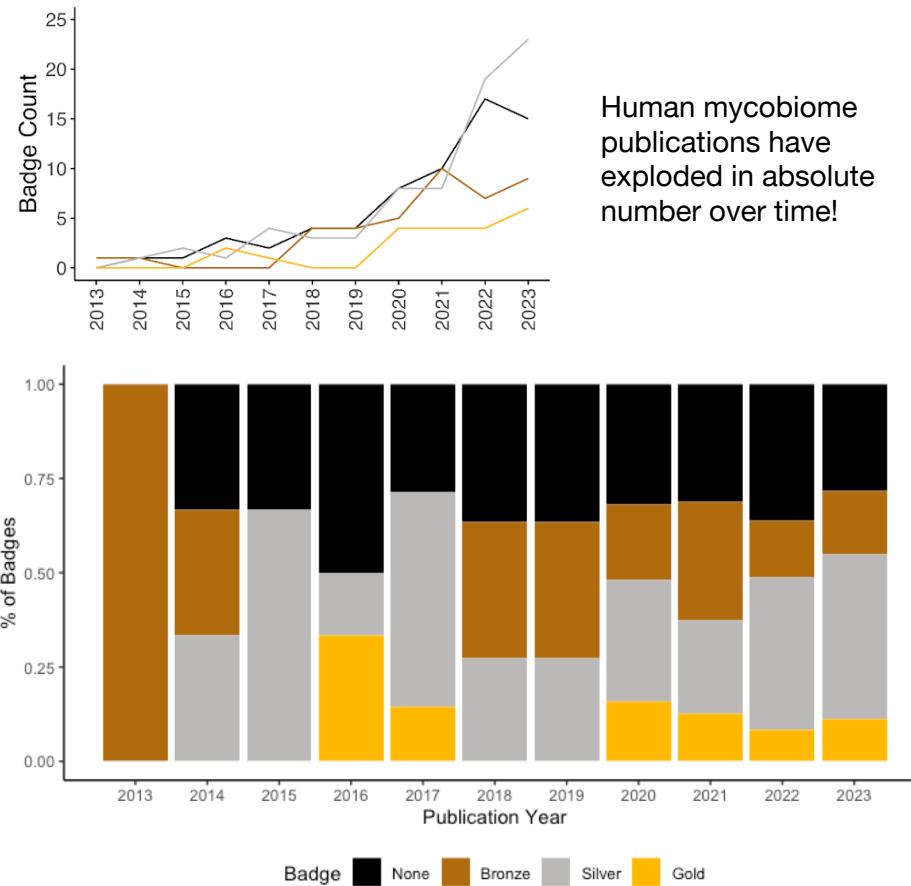
Aim 2: Automated evaluation of standard adherence

We see interesting temporal trends in data accessibility. Data from Case Study 2: A comprehensive assessment of gut mycobiome literature over past decade.



Aim 2: Automated evaluation of standard adherence

We see interesting temporal trends in data accessibility. Data from Case Study 2: A comprehensive assessment of gut mycobiome literature over past decade.



Aim 1: Tier availability for microbiome metadata analysis

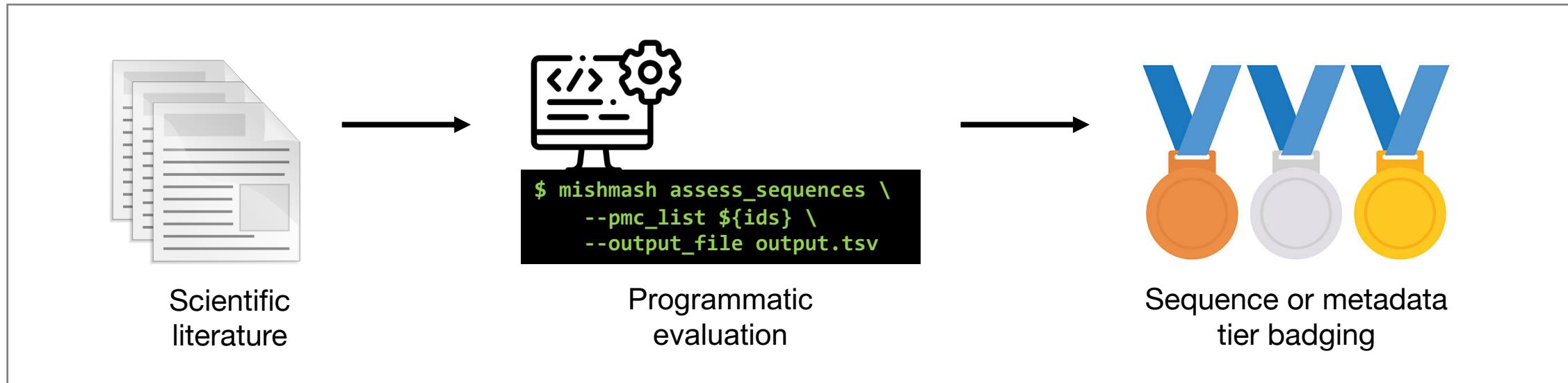
These tiers are prototype standards for human gut microbiome data, based on existing MIxS (Minimal Information about any [x] Sequence) standards.

MIxS v.6.0 Source	Bronze	Silver	Gold	MIxS requirement
MIxS core	Sample name	Sample name	Sample name	Mandatory ¹
MIxS core	Project name	Project name	Project name	Mandatory
MIxS core	Taxonomy ID of DNA sample	Taxonomy ID of DNA sample	Taxonomy ID of DNA sample	Mandatory ¹
MIxS core	Collection date	Collection date	Collection date	Mandatory ^{1,2}
MIxS core	Geographic location (latitude/longitude)	Geographic location (latitude/longitude)	Geographic location (latitude/longitude)	Mandatory ^{1,2}
MIxS core	Geographic location (country/region)	Geographic location (country/region)	Geographic location (country/region)	Mandatory ¹
MIxS core	Broad-scale environmental context	Broad-scale environmental context	Broad-scale environmental context	Mandatory
MIxS core	Local-scale environmental context	Local-scale environmental context	Local-scale environmental context	Mandatory
MIxS core	Environmental medium	Environmental medium	Environmental medium	Mandatory
MIxS core		Sequencing method	Sequencing method	Mandatory
MIxS core		Negative control type	Negative control type	Optional

MIxS core / MIMARKS checklist		Target gene	Target gene	Mandatory
MIxS core / MIMARKS checklist		PCR primers	PCR primers	Optional
MIxS human-gut package		Host subject ID	Host subject ID	Optional
MIxS human-gut package		Host disease status	Host disease status	Optional
MIxS human-gut package		Host total mass	Host total mass	Optional
MIxS human-gut package		Host sex	Host sex	Optional
MIxS core			Sample material processing	Optional
MIxS core			Nucleic acid extraction	Optional
MIxS human-gut package			Host height	Optional
MIxS human-gut package			Host diet	Optional
MIxS human-gut package			Host occupation	Optional

Aim 2: Automated evaluation of standard adherence

Software easily validates data accessibility statements to ensure compliance

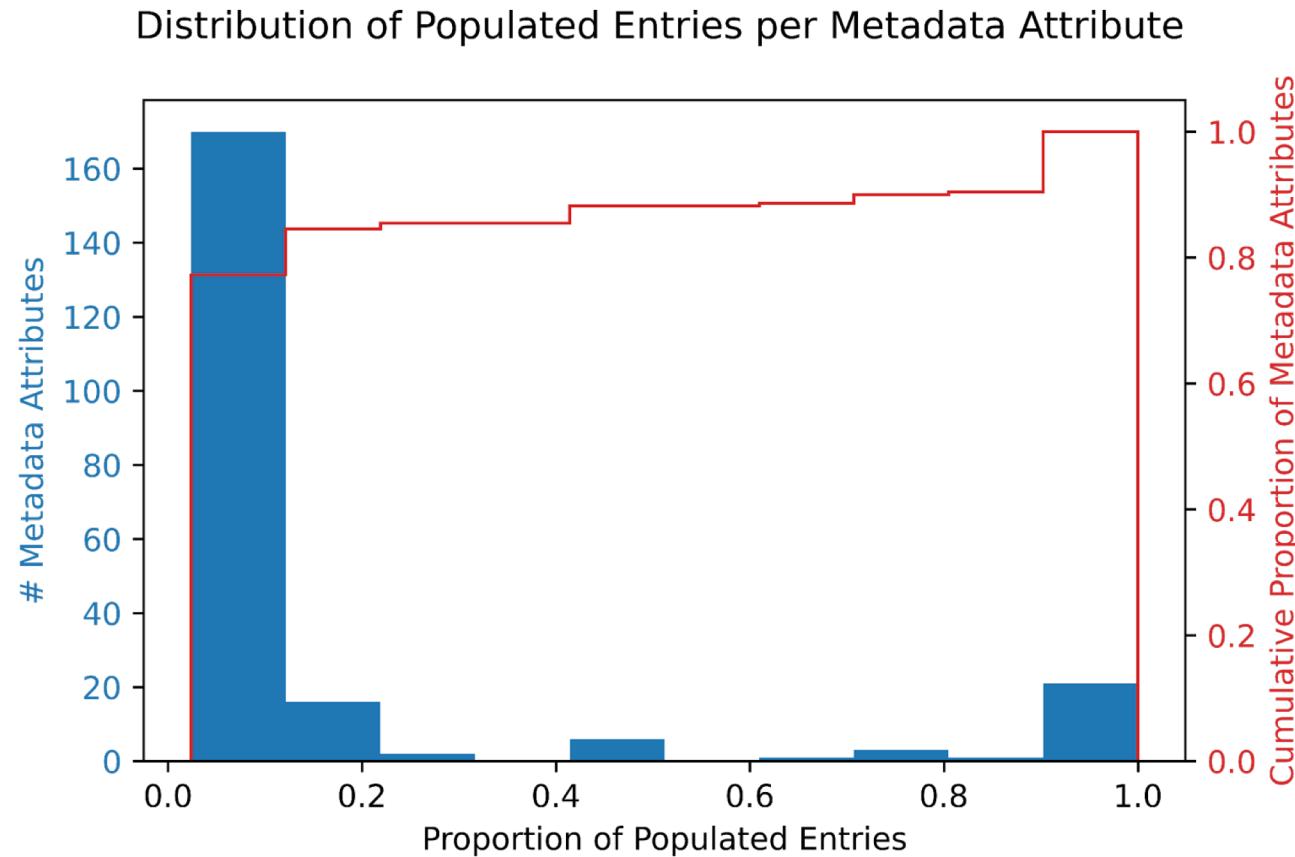


Programmatic tool encourages:

- Researchers to self-assess data availability in own publications
- Users to evaluate meta-analysis data prior to engagement
- Journal editors or grant agencies to verify data FAIRness in stakeholder publications

Aim 2: Automated evaluation of standard adherence

Metadata are not easily interoperable across studies!



Sample ID
Sample-ID
sample_id
Sample.num
sam_number
Sample accession
Sample_accession_ID

Microbiome data must meet FAIR standards

- Project MISHMASH: Microbiome Sequence and Metadata Availability Standards
- Aims:
 1. Develop a tier-based FAIR ORD standard for the field, for microbiome sequence data and metadata
 2. Build software to assess adherence to standards

Impact:

1. Self-assess data availability in own publications
2. Evaluate meta-analysis data prior to engagement
3. Promote dialogue and efforts in microbiome ORD



Bronze: Available data meet the minimum standard for data re-use.

Silver: Facilitated reproducibility; this is a balance between comprehensiveness and feasibility for most.

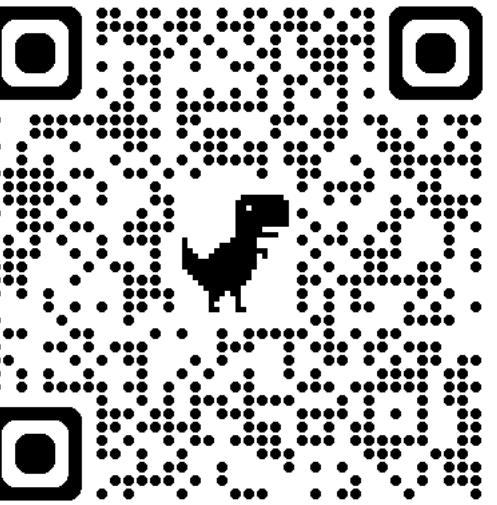
Gold: These data provide valuable information to address research questions beyond the original scope.

Thank you!

MISHMASH co-authors:

Dr. Anton Lavrinienko
Zuzana Sebechlebska
Sven Stoltenberg
Prof. Dr. Nicholas Bokulich

Dr. Alan Pacheco
Chumei Tang
Prof. Dr. Shana Sturla



Publication



ETH zürich



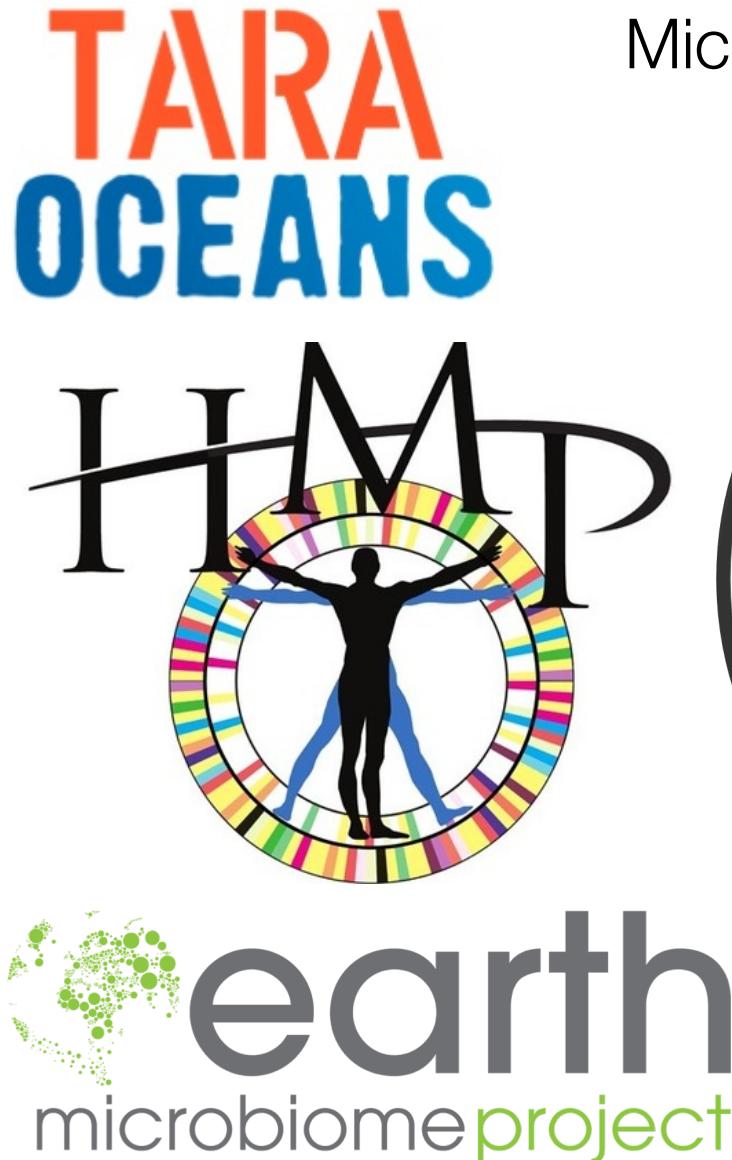
Strategic Focus Area
**Personalized Health
and Related Technologies**



MiM Microbiology
Immunology

Figure credits

- Figures not otherwise cited are stock images or generated from open-source data.
- Slide 2
 - Earth: <https://www.worldatlas.com/space/earth.html>
 - Rural/urban/suburban communities: <https://examples.yourdictionary.com/identifying-difference-between-rural-urban-suburban>
 - 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 3
 - Central dogma: https://www.yourgenome.org/wp-content/uploads/2022/04/dna_central_dogma_yourgenome.png



Microbiomes impact our health
and our environment!



MISHMASH impact requires partnership

