NFDI4 Microbiota
Use Case GUT – „Crowd-sourcing high-quality descriptions of novel taxa“ by Dr. Thomas Hitch
How many prokaryotic species are there?


<table>
<thead>
<tr>
<th></th>
<th>Bacteria</th>
<th>Archaea</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phylum</td>
<td>127</td>
<td>19</td>
<td>146</td>
</tr>
<tr>
<td>Class</td>
<td>360</td>
<td>47</td>
<td>407</td>
</tr>
<tr>
<td>Order</td>
<td>1,163</td>
<td>116</td>
<td>1,279</td>
</tr>
<tr>
<td>Family</td>
<td>2,886</td>
<td>336</td>
<td>3,222</td>
</tr>
<tr>
<td>Genus</td>
<td>12,037</td>
<td>851</td>
<td>12,888</td>
</tr>
<tr>
<td>Species</td>
<td>45,555</td>
<td>2,339</td>
<td>47,894</td>
</tr>
</tbody>
</table>

[Image of a bar chart and a pie chart showing the distribution of taxa across different levels, with percentages indicated for each category.]
How much is left unknown?

16S rRNA gene amplicons (>15,000 mouse gut samples)

Diversity of unknown MAGs

Species clusters (UHGG)
(n = 1,332 cultured)
(n = 3,312 uncultured)

Cultured fraction of the human gut (MAGs)

Almeida et al. 2020 Nat Biotech
Cultured Uncultured

Lesker et al. 2020 Cell Rep

Species status
- Cultured (human gut)
- Cultured (other/unknown source)
- Uncultured

- Cyanobacteria (63)
- Patescibacteria (11)
- Elusimicrobiota (6)
- Ermioibacterota (4)
- Mycocibacterota (1)
- Bdellovibrioibacterota (1)
- Firmicutes G (1)
- Verrucomicrobiota (66)
- Firmicutes B (14)
- Spirochaetota (19)
- Firmicutes A (180)
- Thermoplasmatota (14)
- Actinobacteria (828)
- Bacteroidota (603)
- Desulfobacterota A (27)
- Firmicutes C (148)
- Firmicutes (488)
- Synergistota (9)
- Proteobacteria (349)
- Fusobacterota (35)
- Euryarchaeota (11)
- Halobacterota (3)
- Campylobacterota (47)
- Fibrobacterota (1)

Phyla colouring
- Firmicutes
- Proteobacteria
- Bacteroidetes
- Actinobacteria
- Verrucomicrobiota
- Deferribacteres
- Spirochaetota
- Tenericutes
- Cyanobacteria

Number of MAGs
- > 300
- < 100
- < 50
- < 10
- 1
How are novel species proposed and validated?

1. Isolate a pure culture

2. Is this isolate distinct from its closest relatives?
   - Yes: You have a novel taxon!
   - No: You have a strain of a known species

   (Branches: A, B, C)

   - A: Deposit in two national culture collections
   - B: Define how novel it is
   - C: What are its functional features?

   (Further steps: Name your taxon and define the isolate as the ‘Type material’)}
Protologger: Automated analysis and comprehensive description

7 lines of taxonomic evidence for accurate placement
4 pieces of functional information for description of predicted physiology and niche occupation
2 lines of ecological information to study prevalence across ecosystems

protologger.de

- >3,000 jobs run on the web-server
- Paper downloaded >2,000 times
- Second most cited paper in ISME comms for 2021

Hitch et al. 2021 ISME Comms.
The need for quality descriptions

Tindall. 1999 IJSB

The introduction of a ‘protopogue’

Descriptions of new taxa appear in a wide range of publications, and the requirements of the journals are diverse. This has led to various problems with locating information relating to the description of the new taxon. In some cases the name of the taxon may be included in the text, without clear reference to its properties. It is therefore, suggested that all descriptions of new names or combinations should adhere to a uniform format (c.f. the BioCode; Greuter et al., 1996). This format should be adopted for the valid publication of all names (whether as original articles, or by announcement in the ‘Validation Lists’ and reference to a previous effective publication) appearing in the IJSB starting on 1 January 2000 (or a suitable date thereafter). This would be included as part of Rule 27(2):

‘(1) The new name or new combination should be clearly stated and indicated as such (i.e. fam.nov., gen.nov., sp. nov., comb. nov., etc.).

(2) The derivation of a new name must be given.

(3) The properties of the taxon being described must be given directly after (1) and (2). This may include reference to tables or figures in the same publication, or reference to previously effectively published work.

(4) All information contained in (3) must be accessible.

(5) The type of the taxon must be designated. In the case of species or subspecies the culture collection number(s) where a subculture of the type strain has been deposited must be indicated.’

Gilroy et al. 2021 PeerJ

Table 1 Protologues for new Candidatus taxa identified from metagenomic analysis of chicken gut samples.

<table>
<thead>
<tr>
<th>Description of Candidatus Acetatifactor stercoripullorum sp. nov.</th>
<th>✓ Define novelty ✓ Name taxon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidatus Acetatifactor stercoripullorum (ster.co.rī.pul.lō’rum. L. neut. n. stercus dung; L. masc. n. pullus a young chicken; N.L. gen. n. stercoripullorum of the faeces of young chickens)</td>
<td></td>
</tr>
</tbody>
</table>

A bacterial species identified by metagenomic analyses. This species includes all bacteria with genomes that show ≥95% average nucleotide identity (ANI) to the type genome, which has been assigned the MAG ID CHK195-6426 and which is available via NCBI BioSample SAMN15816622. The GC content of the type genome is 48.46% and the genome length is 3.1 Mbp.

✓ State the type material

A call to arms for systematists: revitalising the purpose and practises underpinning the description of novel microbial taxa

Iain C. Sutcliffe · Martha E. Trujillo · Michael Goodfellow
Application to isolate collections

- **HBC;** human gut bacterial collection from Forster *et al.* (2019)

- From the HBC collection 64 isolates were deposited at the DSMZ

- 40 were identified to represent 34 novel taxa

- 17 novel species and 17 novel genera

- Includes; 3 high priority, 13 medium priority and 6 low priority species according to HMP

Hitch *et al.* 2021 ISME Comms.
Remaining isolate collections

![Circle charts for HBC, BIO-ML, and Hungate1000, showing numbers and color-coded categories (Known, Novel species, Novel genus).]

The need for strain deposition could call for collaboration with NFDI4biodiversity.

The case for Candidatus taxa

The List of Prokaryotic names with Standing in Nomenclature (LPSN) currently contains 2,431 “Candidatus” taxa

Lists of names of prokaryotic Candidatus taxa

Aharon Oren¹*, George M. Garrity²,³, Charles T. Parker³, Maria Chuvochina⁴ and Martha E. Trujillo⁵

INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY

Extensive microbial diversity within the chicken gut microbiome revealed by metagenomics and culture

Rachel Gilroy¹, Anuradha Ravi¹, Maria Getino², Isabella Pursley², Daniel L. Horton², Nabil-Fareed Alikhan¹, Dave Baker¹, Karim Gharbi³, Neil Hall³,⁴, Mick Watson⁵, Evelien M. Adriaenssens¹, Ebenezer Foster-Nyarko¹, Sheikh Jarju⁶, Arss Secka⁷, Martin Antonio⁶, Aharon Oren⁸, Roy R. Chaudhuri⁹, Roberto La Ragione⁵, Falk Hildebrand¹,⁵ and Mark J. Pallen¹,²,⁴

150 Candidatus genus names.

650 distinctive binomials for new Candidatus species
What is *ProtoBIOME*?

Current issues:

- There are **thousands** of microbes still to be described
- There is a **lack of standardisation** regarding what is useful
- If this isn’t addressed soon a high-volume of low-quality *Candidatus* protologues will be produced

Aims of *ProtoBIOME*:

- To provide high-quality descriptions of these novel taxa
- Help set standards in taxonomic description
- Spread these standards via community engagement
The ProtoBIOME framework

1. Select an environment to study
   - Experts in the field are needed to ensure relevant descriptions
   - Individuals who know these taxa to ensure accurate taxonomic placement

2. Genomic database
   - Identify a collection of isolates or MAGs for the selected environment
   - Extract or sequence 16S rRNA gene sequences

3. Genomic analysis via Protologger
   - Each genome/16S pair will be run through Protologger to provide data

4. Make the output available
   - Put the raw Protologger output on the website for the community

5. Engage with the community
   - Train members of the community so they can contribute to the analysis and describe taxa of interest to them
Applying the *ProtoBIOME* framework to the human gut

Almeida *et al.* 2020 *Nat Biotech*

Cultured

Uncultured

Species clusters (UHGG)

\( n = 1,332 \) cultured

\( n = 3,312 \) uncultured

16S rRNA gene sequences were extracted from 1,443 (31.3%)
CandiBIOME framework and integration within NFDI4microbiota

1. Environment specific MAG collections
   - M1.4: Connection to other NFDI’s

2. Produce taxonomic, functional and ecological outputs
   - M2.3: Workflow standards
   - M3.2: Analytical services

3. Teach the community how to understand the data and write a protologue
   - M1.3: Community outreach
   - M1.1: Training courses

4. Final curation of the crowd-sourced protologues
   - M2.1: Data and metadata standards
   - M3.6: Long-term preservation of metadata
Thank you!