



# Introduction to Microbiome Studies

Rob Beiko  
Dalhousie University



Creative Commons

This page is available in the following languages:

Afrikaans Български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto Castellano Castellano (AR) Español (CL) Español (Ecuador) Castellano (MX) Castellano (PE) Euskara Suomi/kieläidän français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu Nederlands Norsk Sesotho sa Leboa polski Português română slovenščina jezik српски (latinica) Sotho svenska 中文 简体 (台灣) isiZulu

 creative  
commons

Attribution-Share Alike 2.5 Canada

You are free:

 to Share — to copy, distribute and transmit the work

 to Remix — to adapt the work





Under the following conditions:

 **Attribution.** You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

 **Share Alike.** If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

[Disclaimer](#)

Your fair dealing and other rights are in no way affected by the above.  
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:  
[English](#) [French](#)

[Learn how to distribute your work using this licence](#)

**The Microbiome**

**The Microbiome**

**The Microbiome**

---

## Things we would like to understand about the microbiome

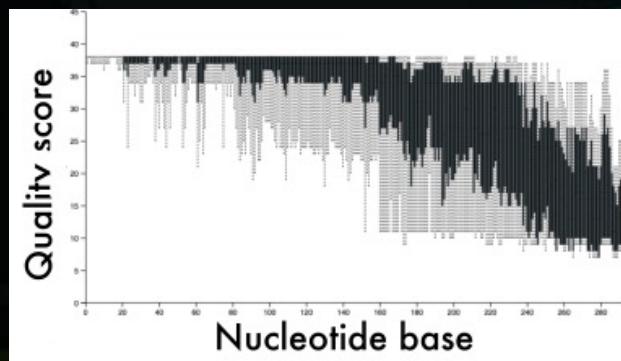
1. Who is there
2. What they are doing
  - o To each other
  - o To their environment
3. How they will respond
  - o To each other
  - o To their environment



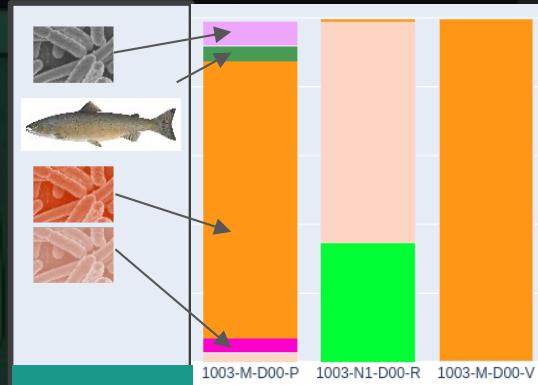
# What we want



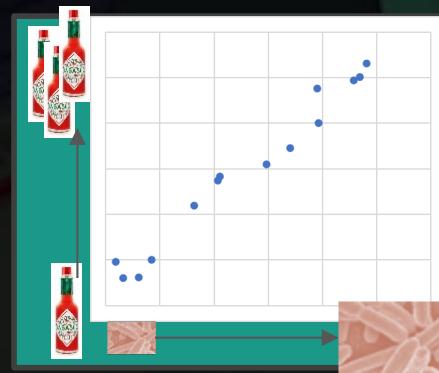
# What we get



short, messy raw data



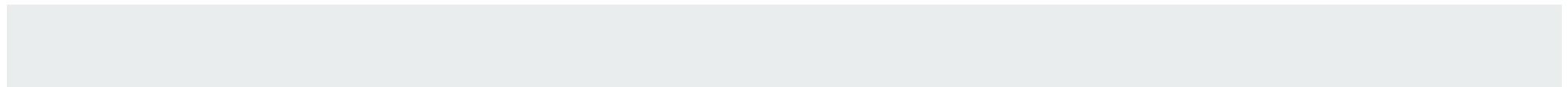
taxonomic incoherence



spurious associations



mediocre assembly



---

## The way forward (is like any other good science)

- Frame appropriate questions
- Understand how your **data** relate to the underlying **question**
  - Your data are often a **proxy** for what you really want to know
- Understand the **limitations** of what your data can tell you
- Choose appropriate **methods** and understand *their* **limitations**
- Assess the **stability** and **robustness** of your results, where possible
- (Try to) Avoid breathless **overinterpretation** of results



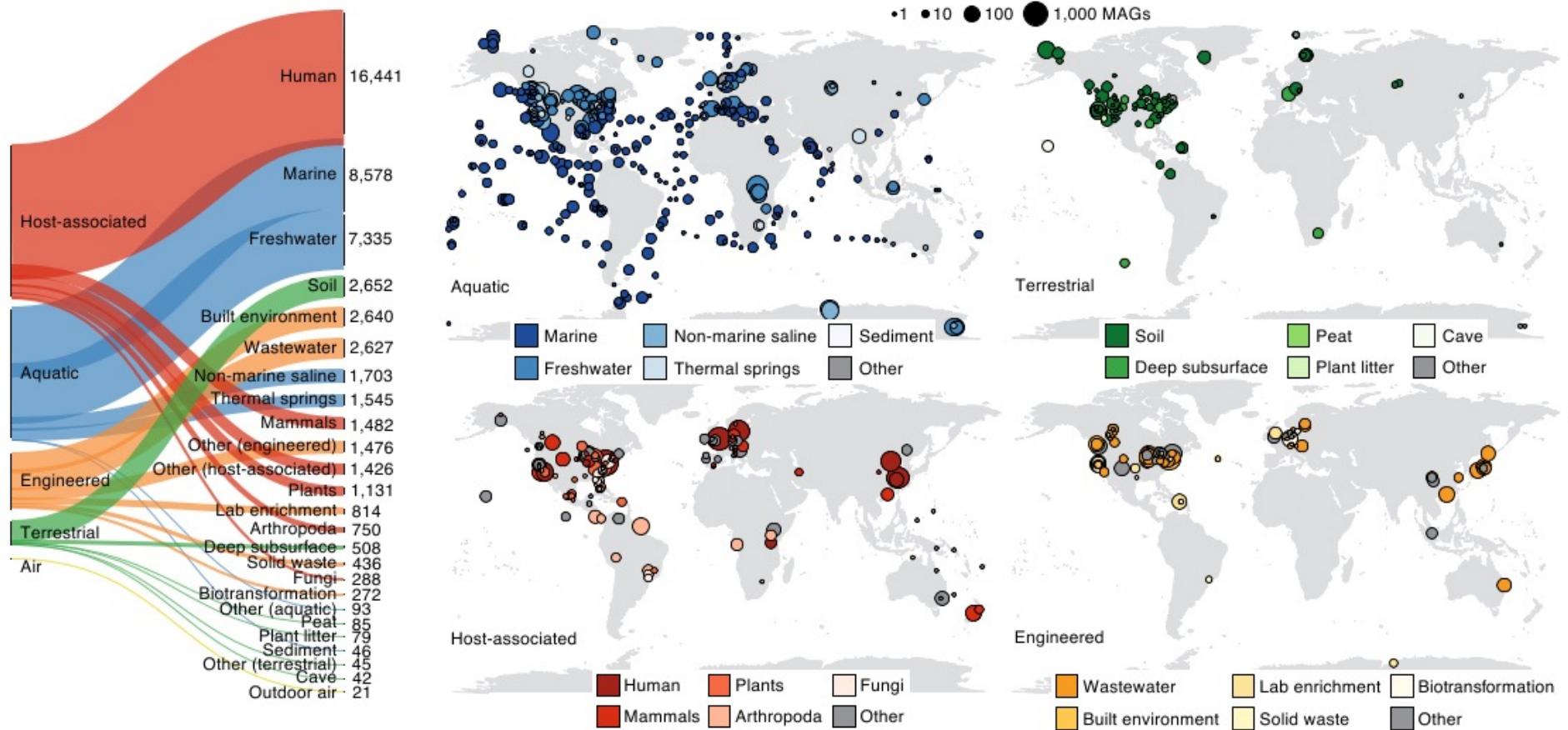
## **Why do we care about the microbiome?**

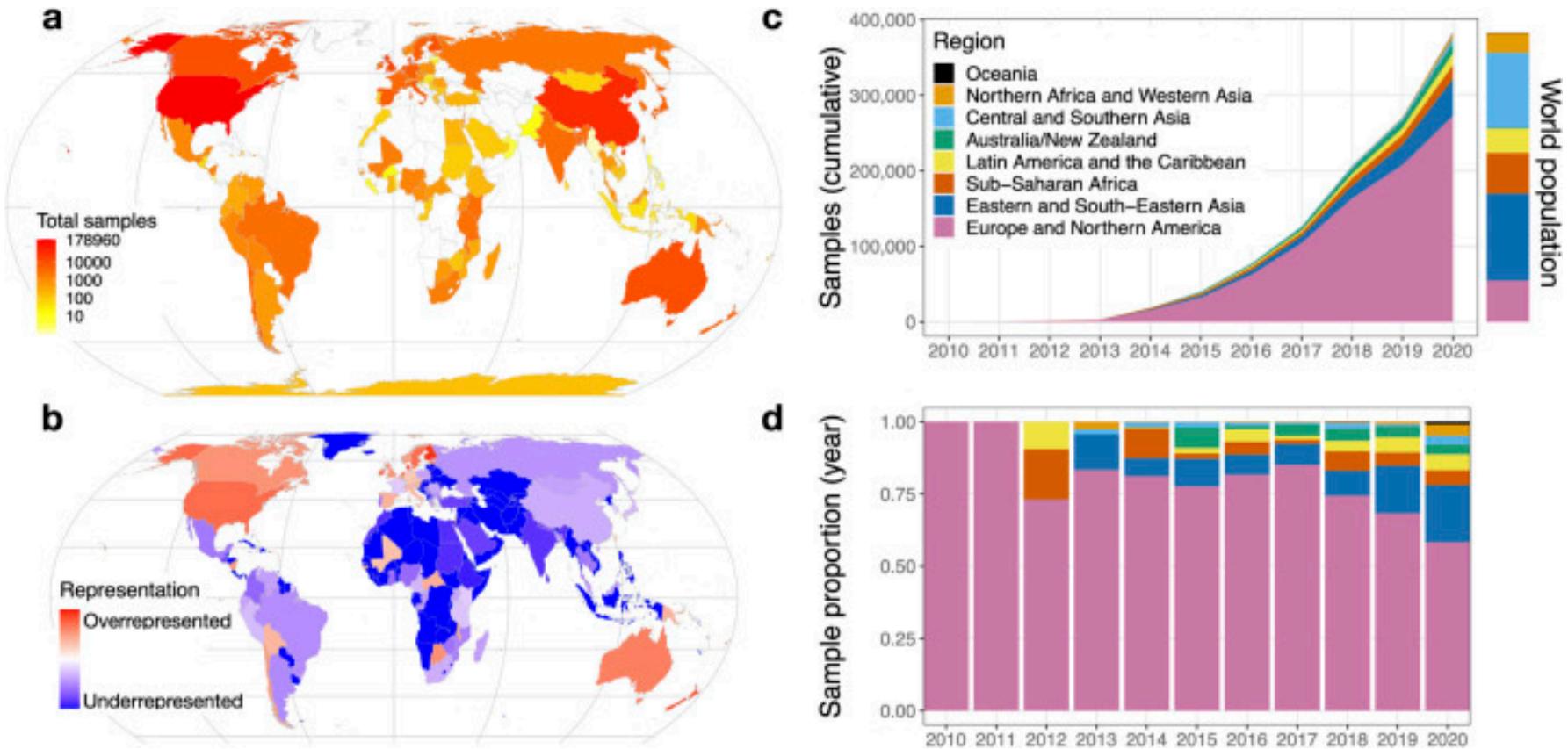
- You tell me!

**There are lots of data out there...  
...sort of.**



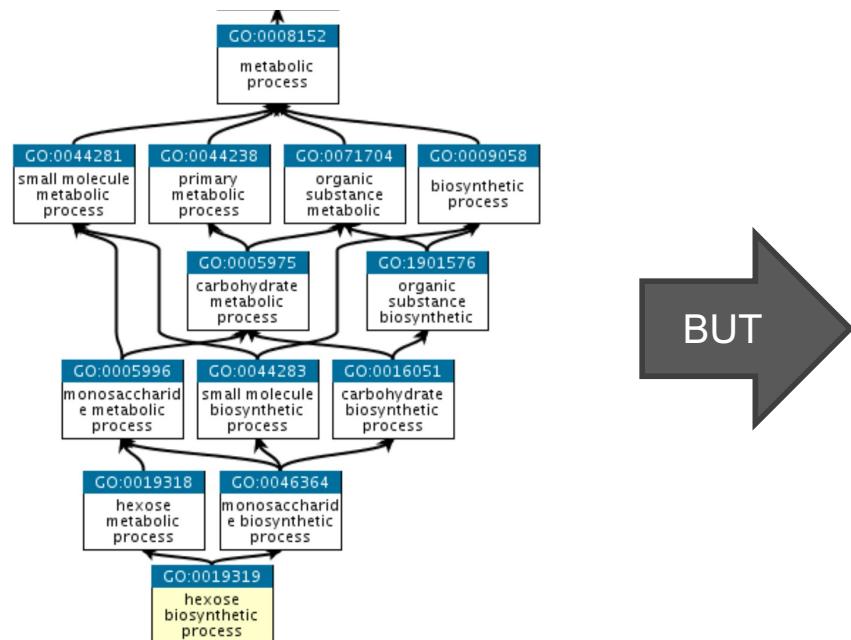
# Distribution of metagenome-assembled genomes





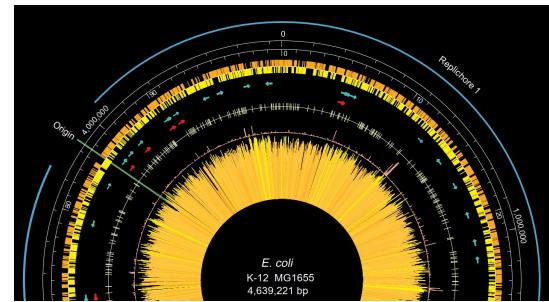
# Protein function

Gene Ontology: 43,303 terms  
(2022-11-03)



BUT

*Escherichia coli* K-12 MG1655



Blattner et al. (1997) *Science*

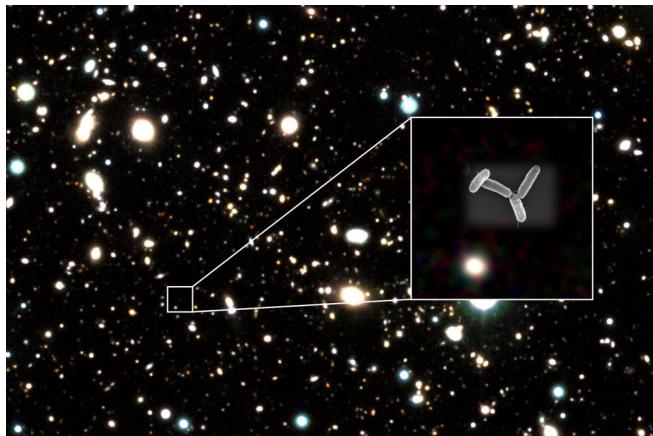
Dec 1, 2022:

4298 annotated proteins  
687 “putative”  
278 “domain-containing”  
+ high-level “XXX family”

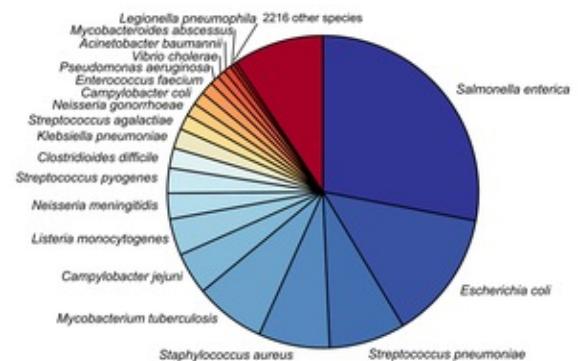
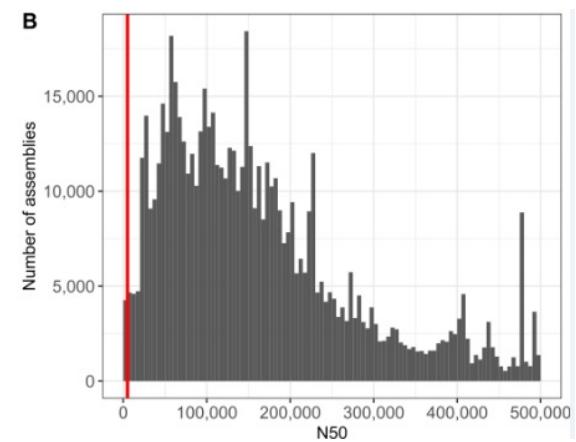
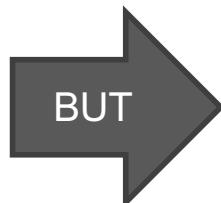
Critical Assessment of Functional Annotation (CAFA): Zhou et al. (2019) *Genome Biol*

## Reference genomes

639,981 high-quality sequenced genomes!



Harikane et al. (2022) *ApJ*



Blackwell et al. (2021) *PLoS Biol*

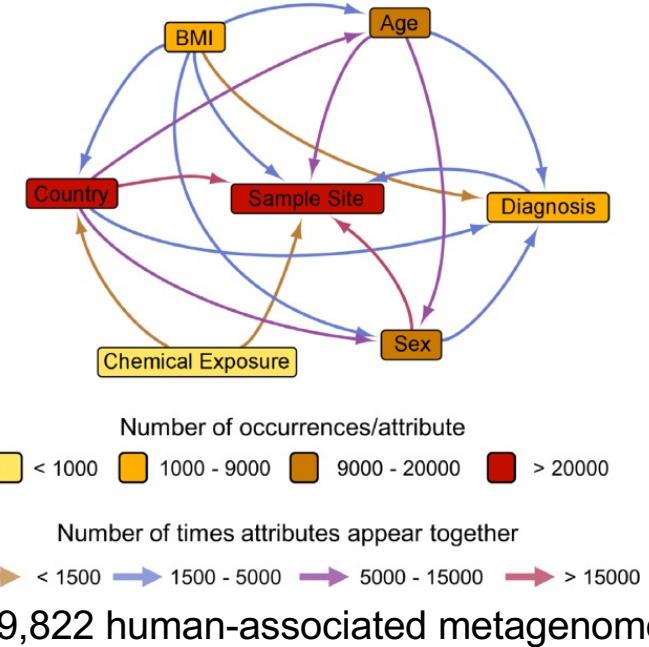
# Metadata

## PERSPECTIVE

**Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIXS) specifications**

Slots			
MixS	Name	Cardinality and Range	Description
Reference			
Introduction			
	MixS ID	Name	Cardinality and Range
	0000065	x_16s_recover	0..1 xsd:boolean
	0000066	x_16s_recover_software	0..1 xsd:string
	0000048	adapters	0..1 recommended xsd:string
	0000094	alt	0..1 recommended xsd:string
	0000059	annot	0..1 xsd:string
	0000057	assembly_name	0..1 recommended xsd:string
	0000056	assembly_qual	1..1 xsd:string
			Can a 16S gene be recovered from the submitted SAG or MAG?
			Tools used for 16S RNA gene extraction
			Adapters provide priming sequences for both amplification and sequencing of t...
			Altitude is a term used to identify heights of objects such as airplanes, spa...
			Tool used for annotation, or for cases where annotation was provided by a com...
			Name/version of the assembly provided by the submitter that is used in the ge...
			The assembly quality category is based on sets of criteria outlined for each ...

BUT



Kasmanas et al. (2021) *Nucleic Acids Res*

<https://genomicsstandardsconsortium.github.io/mixs/Mimag/>

# Prokaryotic taxonomy has an...interesting history

INTERNATIONAL JOURNAL OF SYSTEMATIC BACTERIOLOGY, July 1988, p. 321–325  
0020-7713/88/030321-05\$02.00/0  
Copyright © 1988, International Union of Microbiological Societies

Vol. 38, No. 3

## *Proteobacteria* classis nov., a Name for the Phylogenetic Taxon That Includes the “Purple Bacteria and Their Relatives”

E. STACKEBRANDT,<sup>1</sup> R. G. E. MURRAY,<sup>2\*</sup> AND H. G. TRÜPER<sup>3</sup>

*Lehrstuhl für Allgemeine Mikrobiologie, Biologiezentrum, Christian-Albrechts Universität, 2300 Kiel, Federal Republic of Germany<sup>1</sup>; Department of Microbiology and Immunology, University of Western Ontario, London, Ontario, Canada N6A 5C1<sup>2</sup>; and Institut für Mikrobiologie, Universität Bonn, 5300 Bonn 1, Federal Republic of Germany<sup>3</sup>*

*Proteobacteria* classis nov. is suggested as the name for a new higher taxon to circumscribe the  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\delta$  groups that are included among the phylogenetic relatives of the purple photosynthetic bacteria and as a suitable collective name for reference to that group. The group names (alpha, etc.) remain as vernacular terms at the level of subclass pending further studies and nomenclatural proposals.

“*Proteus*, a Greek god of the sea, capable of assuming many different shapes; [...] *bakterion*, a small rod; *Proteobacteria* protean **group of bacteria of diverse properties** despite a common ancestry”

## Valid publication of the names of forty-two phyla of prokaryotes

Aharon Oren<sup>1,\*</sup> and George M. Garrity<sup>2,\*</sup>

### Abstract

After the International Committee on Systematics of Prokaryotes (ICSP) had voted to include the rank of phylum in the rules of the International Code of Nomenclature of Prokaryotes (ICNP), and following publication of the decision in the IJSEM, we here present names and formal descriptions of 42 phyla to effect valid publication of their names, based on genera as the nomenclatural types.

### **BDELLOVIBRIONOTA PHYL. NOV.**

(Bdello.vi.bri.o.no'ta. N.L. masc. n. *Bdellovibrio*, type genus of the phylum; -ota, ending to denote a phylum; N.L. pl. neut. n. *Bdellovibrionota*, the *Bdellovibrio* phylum)

The properties of the taxon are as described by Waite *et al.*, 2020 [14].

Type genus: *Bdellovibrio* Stolp and Starr 1963 (Approved Lists 1980).

### **CAMPYLOBACTEROTA PHYL. NOV.**

(Cam.py.lo.bac.te.ro'ta. N.L. masc. n. *Campylobacter*, type genus of the phylum; -ota, ending to denote a phylum; N.L. pl. neut. n. *Campylobacterota*, the *Campylobacter* phylum)

The properties of the taxon are as described by Waite *et al.*, 2018 [17]. Replacement of the illegitimate name: *Epsilonbacterota* Waite *et al.*, 2017 [18], which is an earlier synonym for *Campylobacterota* Waite *et al.* 2018, but is illegitimate as it was based on the illegitimate class *Campylobacteria* Waite *et al.* 2017, which is a later homotypic synonym of *Epsilonproteobacteria* Garrity 2006.

Type genus: *Campylobacter* Sebald and Véron 1963 (Approved Lists 1980).

### **PSEUDOMONADOTA CORRIG. PHYL. NOV.**

(Pseu.do.mo.na.do'ta. N.L. fem. n. *Pseudomonas*, type genus of the phylum; -ota, ending to denote a phylum; N.L. pl. neut. n. *Pseudomonadota*, the *Pseudomonas* phylum)

The properties of the taxon are as described by Garrity *et al.*, 2005 [36]. Correction of the effectively published synonym: *Proteobacteria* (sic) Garrity *et al.* 2005.

Type genus: *Pseudomonas* Orla Jensen 1921 (Approved Lists 1980).

INTESTINAL FLORA IN NEW-BORN INFANTS  
WITH A DESCRIPTION OF A NEW PATHOGENIC ANAEROBE,  
BACILLUS DIFFICILIS

IVAN C. HALL, PH.D.  
AND  
ELIZABETH O'TOOLE  
DENVER

*"Bacillus difficile"* (1935)

Published in final edited form as:  
*Environ Microbiol.* 2013 October; 15(10): 2631–2641. doi:10.1111/1462-2920.12173.

A genomic update on clostridial phylogeny: Gram-negative  
spore-formers and other misplaced clostridia

Natalya Yutin and Michael Y. Galperin\*

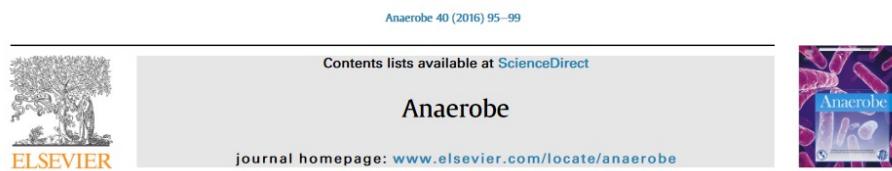
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

cellulolytic clostridia that belong to the family *Ruminococcaceae*. As a tentative solution to resolve the current taxonomical problems, we propose assigning 78 validly described *Clostridium* species that clearly fall outside the family *Clostridiaceae* to six new genera: *Pectoclostridium*, *Lachnoclostridium*, *Ruminiclostridium*, *Erysipelatoclostridium*, *Gottschalkia*, and *Tyzzera*. This work reaffirms that 16S rRNA and ribosomal protein sequences are better indicators of evolutionary proximity than phenotypic traits, even such key ones as the structure of the cell envelope and Gram-staining pattern.

*"Pectoclostridium difficile"* (2013)

Prévot AR. Études de systématique bactérienne. IV. Critique de la conception actuelle du genre *Clostridium*. *Annales de l'Institut Pasteur (Paris)* 1938; **61**:72-91.

*"Clostridium difficile"* (1938)



*Clostridium difficile*

Reclassification of *Clostridium difficile* as *Clostridioides difficile* (Hall and O'Toole 1938) Prévot 1938

Paul A. Lawson <sup>a,\*</sup>, Diane M. Citron <sup>b</sup>, Kerin L. Tyrrell <sup>b</sup>, Sydney M. Finegold <sup>c, d, e</sup>

*"Clostridioides difficile"* (2016)



Contents lists available at ScienceDirect

## Systematic and Applied Microbiology

journal homepage: [www.elsevier.com/locate/syapm](http://www.elsevier.com/locate/syapm)



### Development of the SeqCode: A proposed nomenclatural code for uncultivated prokaryotes with DNA sequences as type



William B. Whitman <sup>a,\*</sup>, Maria Chuvochina <sup>b</sup>, Brian P. Hedlund <sup>c</sup>, Philip Hugenholtz <sup>b</sup>, Konstantinos T. Konstantinidis <sup>d</sup>, Alison E. Murray <sup>e</sup>, Marike Palmer <sup>c</sup>, Donovan H. Parks <sup>b</sup>, Alexander J. Probst <sup>f</sup>, Anna-Louise Reysenbach <sup>g</sup>, Luis M. Rodriguez-R <sup>h</sup>, Ramon Rossello-Mora <sup>i</sup>, Iain Sutcliffe <sup>j</sup>, Stephanus N. Venter <sup>k</sup>

“...a new code of nomenclature, the Code of Nomenclature of Prokaryotes Described from Sequence Data (SeqCode), has been developed over the last two years to allow **naming of Archaea and Bacteria using DNA sequences as the nomenclatural types.**”

# The Plan





## IMPACCT Bioinformatics Workshop Schedule

ARTS 150, 853 Rue Sherbrooke Ouest, McGill University, Montréal  
Dec 6-7, 2022

### Tuesday – Dec 6, 2022

09:00 – 09:10     Welcome and Student Introduction

#### Module 1

09:10 – 09:40     Lecture: Introduction to Microbiome Studies  
Instructor: Dr. Rob Beiko

09:40 – 10:00     Lab: Introduction to AWS  
Instructor: Zhibin Lu

10:00 – 10:30     AM Break

#### Module 2

10:30 – 11:00     Lecture: QIIME2 from Sequence to ASV Table  
Instructor: Dr. Rob Beiko

11:00 – 12:30     Lab: QIIME2 from Sequence to ASV Table  
TA: Diana Haider

12:30 – 13:15     Lunch Break

#### Special Topic

13:15 – 14:00     Lecture: Experimental Design, Sample Collection & Storage  
Instructor: Dr. Corinne Maurice

14:00 – 14:15     Q&A: Experimental Design, Sample Collection & Storage  
TA: Michael Shamash

#### Module 3

14:15 – 15:00     Lecture: Statistics & Data Visualization  
Instructor: Dr. Rob Beiko

15:00 – 16:00     Lab: Statistics & Data Visualization  
TA: Diana Haider

16:00 – 16:15     PM Break

#### Module 3 (continued)

16:15 – 17:00     Lab: Statistics & Data Visualization  
TA: Diana Haider

17:00 – 19:00     Cocktail Hour (McGill Faculty Club, 3450 McTavish Street)



## Learning Outcomes – Day 1 (Lecture)

- You will be able to:
  - **Part 1**
    - **Understand** what the microbiome is
    - **Despair** of our ability to characterize the microbiome
  - **Part 2**
    - **Understand** the main strengths and weaknesses of marker gene-based approaches
    - **Read** and **interpret** the contents of sequence files
    - **Describe** the process of sequence clustering
  - **Part 3**
    - **Interpret** the results of analyses including:
      - Taxonomic summaries
      - Diversity analysis
      - Differential abundance



## **Learning Outcomes – Day 1 (Lab)**

- By the end of the tutorial, you will be able to:
  - Conduct and end-to-end microbiome analysis using QIIME2
  - Know the main tools available to conduct statistical and diversity analysis

The screenshot shows the QIIME 2 website at [qiime2.org](https://qiime2.org). The header includes a navigation bar with links for Home, Library, Docs, Forum, Workshops, and View. A star icon for bookmarking is also present. The main content features the QIIME 2 logo with a stylized DNA helix graphic. A descriptive text block states: "QIIME 2™ is a next-generation microbiome bioinformatics platform that is extensible, free, open source, and community developed." Below this are three buttons: "Code of Conduct »" (blue), "Citing QIIME 2 »" (green), and "Learn more »" (blue). The page then displays four key features with corresponding icons: 1) "Automatically track your analyses with decentralized data provenance — no more guesswork on what commands were run!" (represented by a network graph icon). 2) "Interactively explore your data with beautiful visualizations that provide new perspectives." (represented by a bar chart with a magnifying glass icon). 3) "Easily share results with your team, even those members without QIIME 2 installed." (represented by two people sharing a document icon). 4) "Plugin-based system — your favorite microbiome methods all in one place." (represented by a puzzle piece icon).

qiime2.org

Home Library Docs Forum Workshops View

# qiime2

QIIME 2™ is a next-generation microbiome bioinformatics platform that is extensible, free, open source, and community developed.

[Code of Conduct »](#) [Citing QIIME 2 »](#) [Learn more »](#)

Automatically track your analyses with decentralized data provenance — no more guesswork on what commands were run!

Interactively explore your data with beautiful visualizations that provide new perspectives.

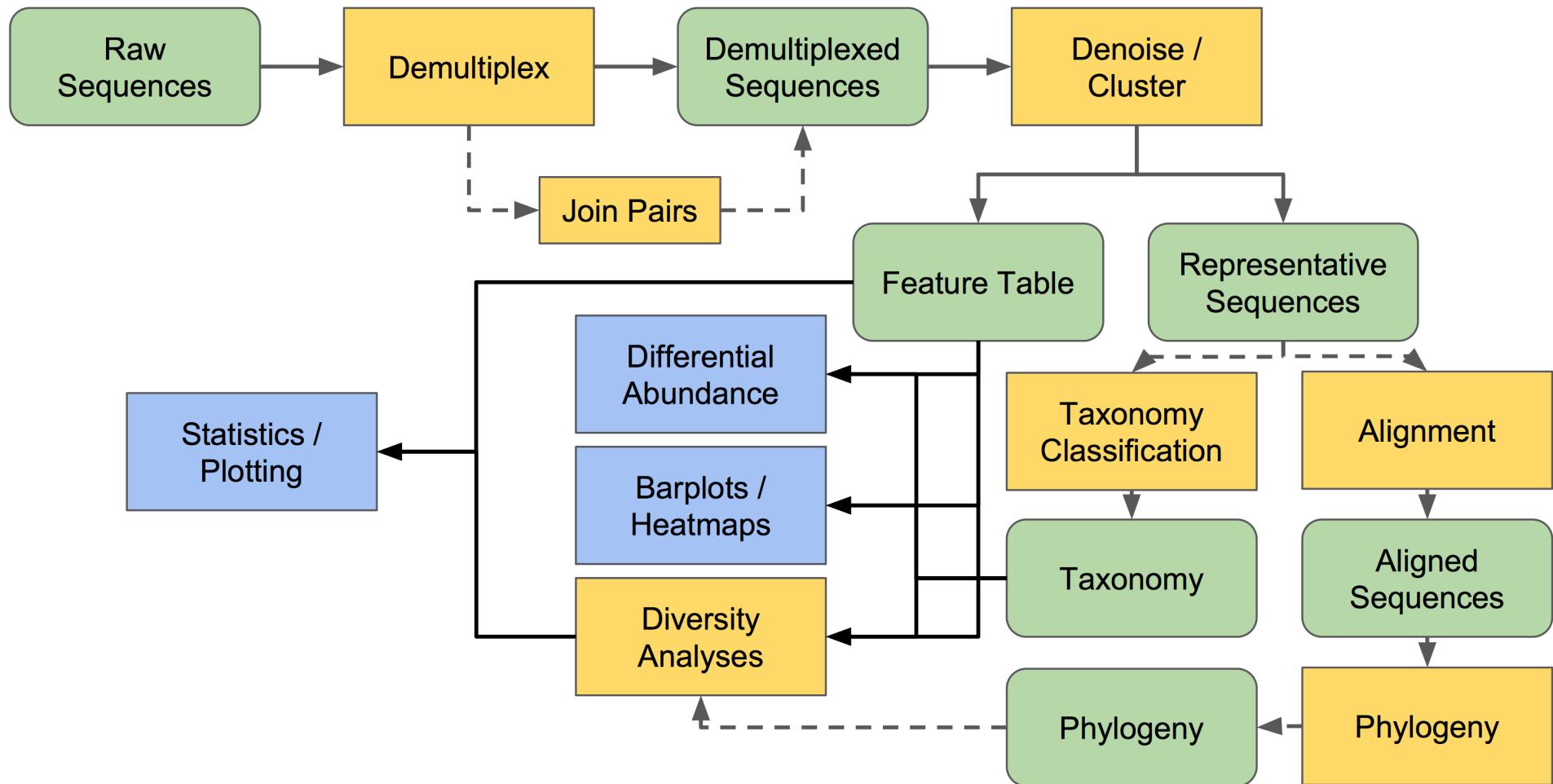
Easily share results with your team, even those members without QIIME 2 installed.

Plugin-based system — your favorite microbiome methods all in one place.



## Foundations of QIIME2

- *Artifacts* (qza): intermediate data, typically produced by one action and fed into another
- *Visualizations* (qzv): machine-readable visualizations (quick and dirty: <https://view.qiime2.org/>)
- *Plugins*: a package that provides one or more steps in a pipeline (e.g., demultiplexing)
- *Data provenance*: Information about the steps that led to the present set of results
  
- NB: QIIME2 images are often lifted from the tutorial page:  
<https://docs.qiime2.org/2022.8/tutorials/overview/>



# Alternatives exist!

RESEARCH ARTICLE

## Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing

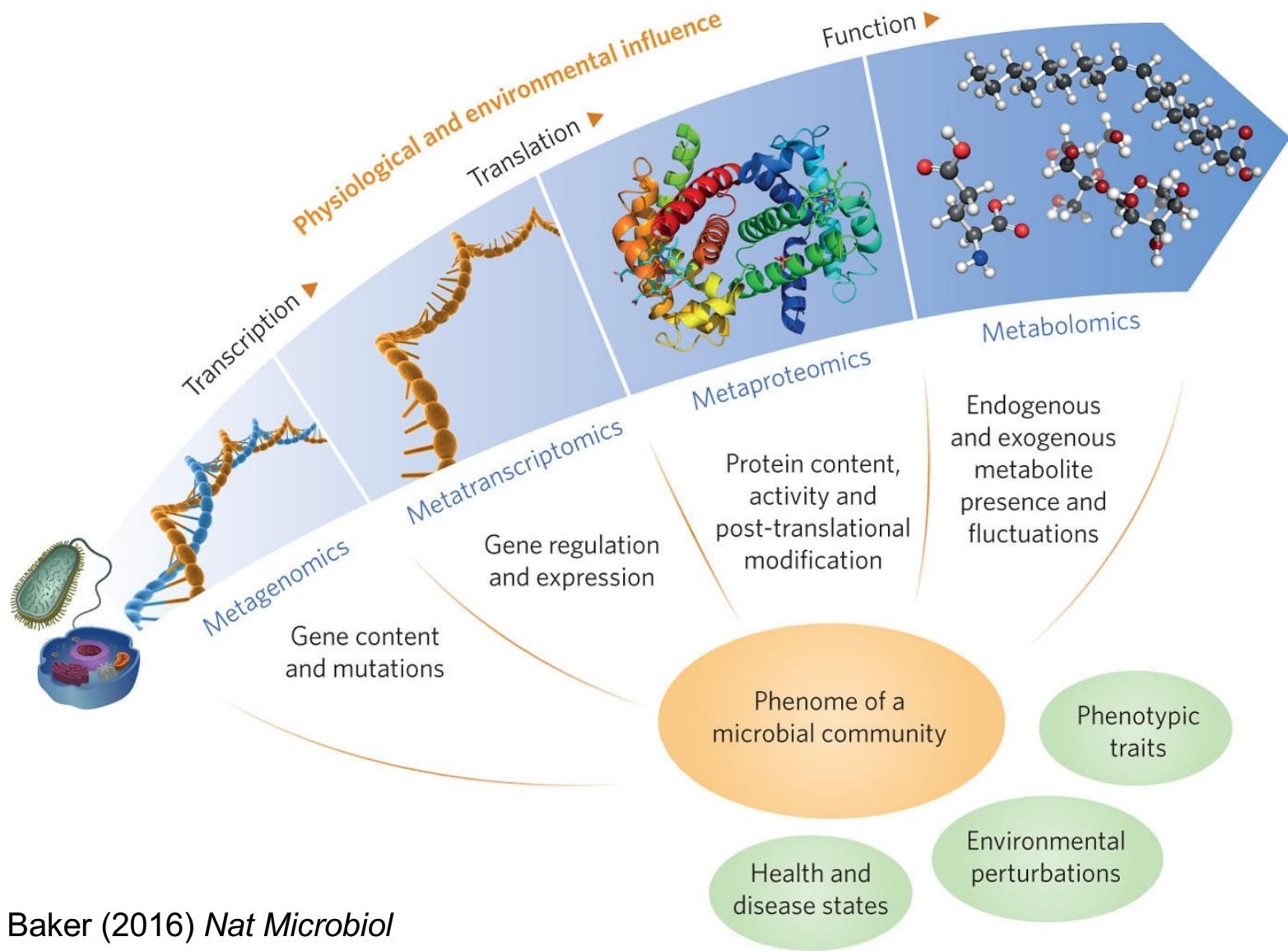
Andrei Prodan<sup>1\*</sup>, Valentina Tremaroli<sup>2</sup>, Harald Brolin<sup>2</sup>, Aeilko H. Zwinderman<sup>3</sup>, Max Nieuwdorp<sup>1</sup>, Evgeni Levin<sup>1,4</sup>

“DADA2 offered the best sensitivity, at the expense of decreased specificity compared to USEARCH-UNOISE3 and Qiime2-Deblur. USEARCH-UNOISE3 showed the best balance between resolution and specificity. OTU-level USEARCH-UPARSE and MOTHUR performed well, but with lower specificity than ASV-level pipelines. QIIME-uclust produced large number of spurious OTUs as well as inflated alpha-diversity measures and should be avoided in future studies.”

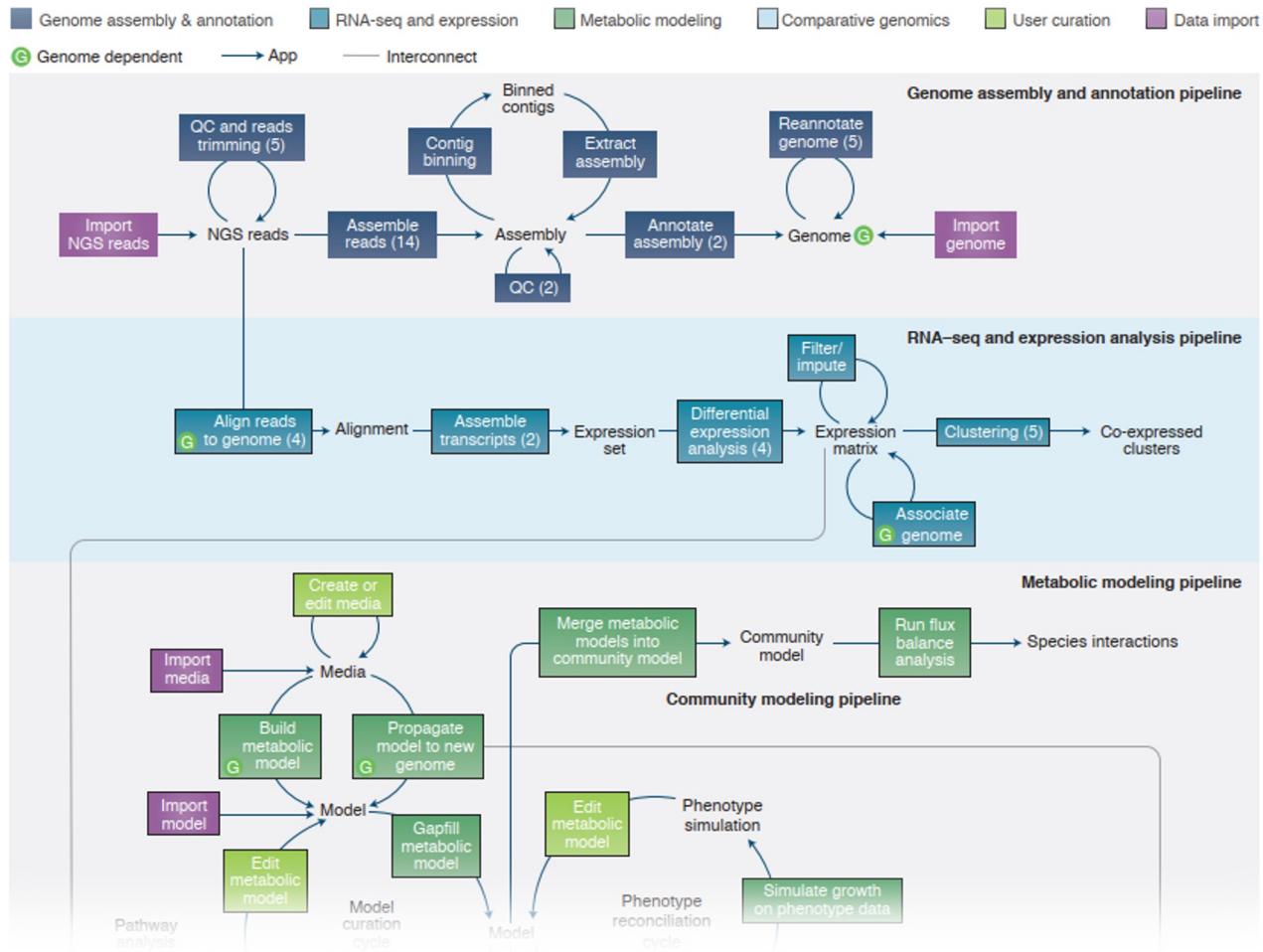
Prodan et al. (2020) *PLoS ONE*

# How we assess the microbiome

---



## DoE Knowledgebase (KBase)





End of  
Part I

30