# A Brief Introduction to Microbial Taxonomic Assignment

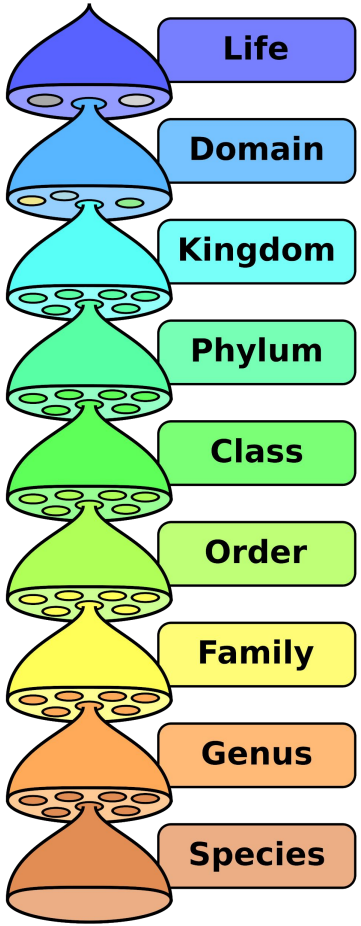
M$^3$ Workshop
January 10, 2019

# Outline

1. A brief background on taxonomy and phylogeny
2. How (microbial) taxonomy is maintained
3. Assigning an unknown sequence to a taxon
    a. One slide on OTUs/ASVs
    b. Methods and challenges
    c. Databases
4. Outline for the workshop

# What is taxonomy?

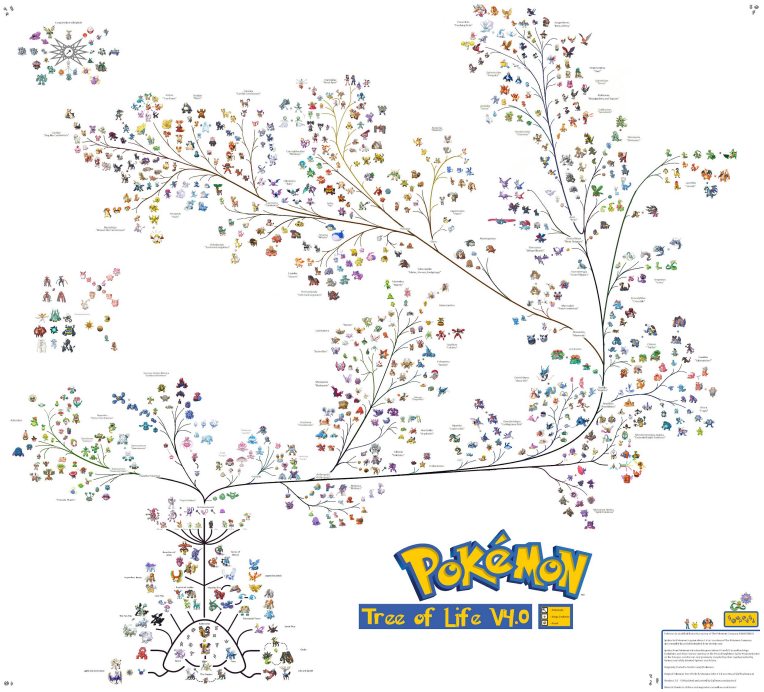
- Classification of organisms, typically arranged in hierarchical ranks (e.g. kingdom, phylum, class, order, genus, species)
- For meaningful community comparisons, taxonomic names must be consistent
- “Official” Taxonomic Names
  - Bergey’s Taxonomic Outline - manual of taxonomic names for bacteria
  - List of Prokaryotic Names with Standing in the Nomenclature
  - New species (genomically distinct, phenotypically distinguishable, deposited in two strain collections) are published in the International Journal of Systematic and Evolutionary Microbiology (IJSEM)




https://en.wikipedia.org/wiki/Taxonomic_rank

# My introduction to taxonomy

# My introduction to taxonomy

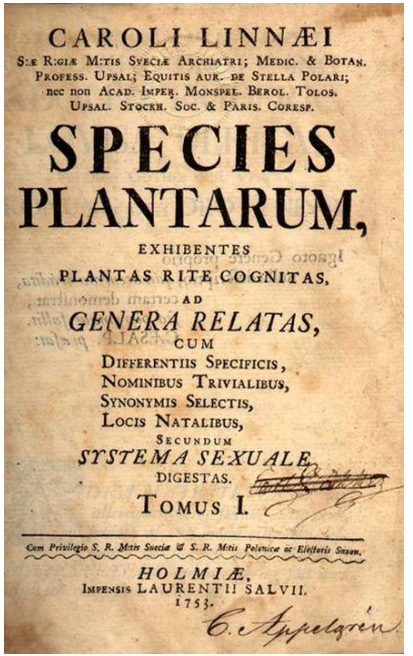



https://en.wikipedia.org/wiki/Game_Boy_Color
https://www.youtube.com/watch?v=Mhgnrq-fG5M
https://connect.unity.com/p/pokeadventure-pokemon-game-demo
https://bulbapedia.bulbagarden.net/wiki/Evolution
https://i.redd.it/yx1uxuw1cl3y.png

Pokemon evolution ≠ Real evolution
Pokemon taxonomy == A kind of taxonomy

# An actual history of taxonomy

- Taxonomy is likely as old as language
- Botanical taxonomy was of particular interest
  - Knowing and communicating the names of edible and poisonous plants to share acquired experiences
- Two works of Carl Linnaeus, a Swedish botanist, are regarded as the starting points of modern botanical and zoological taxonomy
  - *Species Plantarum*, 1753
  - The 10th edition of *Systema Naturae*, 1758
- Charles Darwin: evolutionary theory, 1858
- Ernst Haeckel and August Eichler: started constructing evolutionary trees using phenetics (anatomy, biochemistry, etc)
  - Coined the term ‘phylogeny’


CAROLI LINNÆI
S:Æ R:GIÆ M:TIS SVECIÆ ARCHIATRI; MEDIC. & BOTAN. PROFESS. UPSAL; EQUITIS AUR. DE STELLA POLARI; nec non ACAD. IMPER. MONSPEL. BEROL. TOLOS. UPSAL. STOCKH. SOC. & PARIS. CORESP.

SPECIES PLANTARUM,

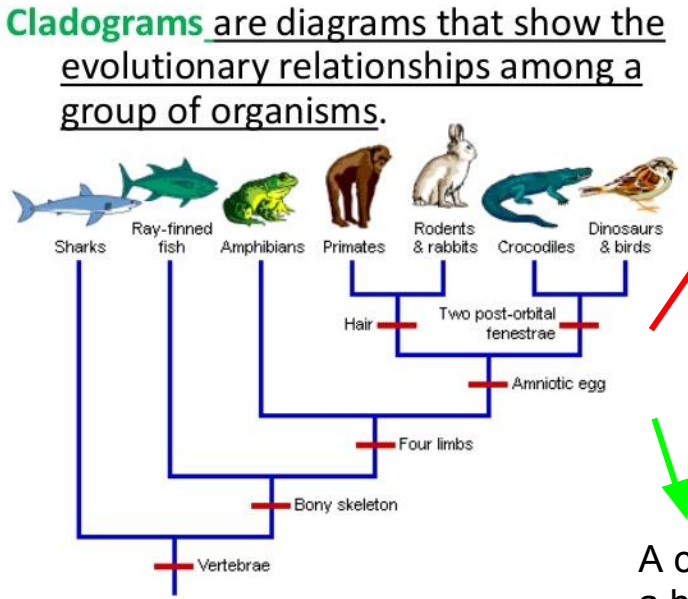
EXHIBENTES PLANTAS RITE COGNITAS, AD GENERA RELATAS, CUM DIFFERENTIIS SPECIFICIS, NOMINIBUS TRIVIALIBUS, SYNONYMIS SELECTIS, LOCIS NATALIBUS, SECUNDUM SYSTEMA SEXUALE DIGESTAS.

TOMUS I.

Cum Privilegio S. R. Mtis Suecia & S. R. Mtis Poloniæ ac Electoris Saxon.

HOLMIÆ, IMPENSIS LAURENTII SALVII. 1753.

# Phylogeny




Taxonomy

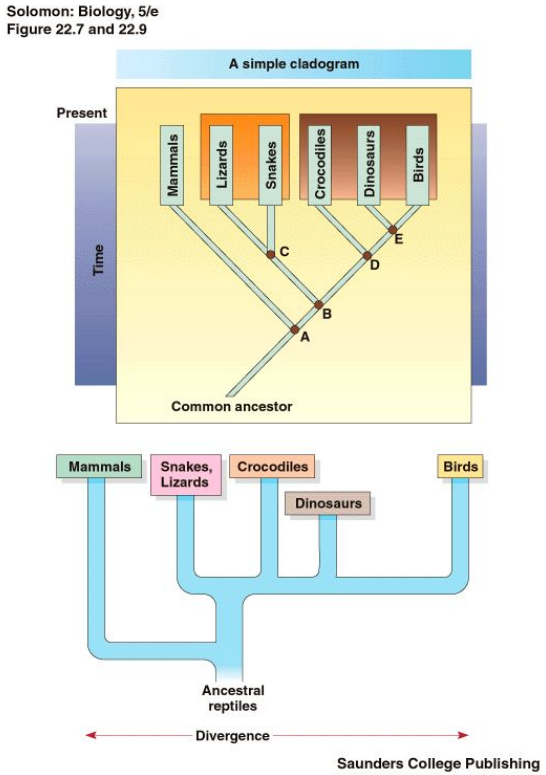
A cladogram, representing a hypothesis of the evolution of vertebrates

- Evolutionary development and diversification of organisms
- Phylogenetic trees do not use taxonomic information but are generated directly from sequencing data via neighbor-joining, maximum parsimony, maximum likelihood, or other methods
- Difficult to make good phylogenetic trees!

https://www.slideshare.net/stubeck/cladograms-73805705

# What is phylogeny?

- A phylogeny is a visual description of how organisms might be evolutionarily related through common ancestors
- Phylogenies are often built off of sequence alignments and different models of evolution that have different assumptions
  - Neighbor joining: assumes the distance between sequences is an estimate of evolution
  - Parsimony: build the phylogeny with the fewest evolutionary events that can explain the relationship
  - Maximum likelihood: assign probabilities based on mutation models



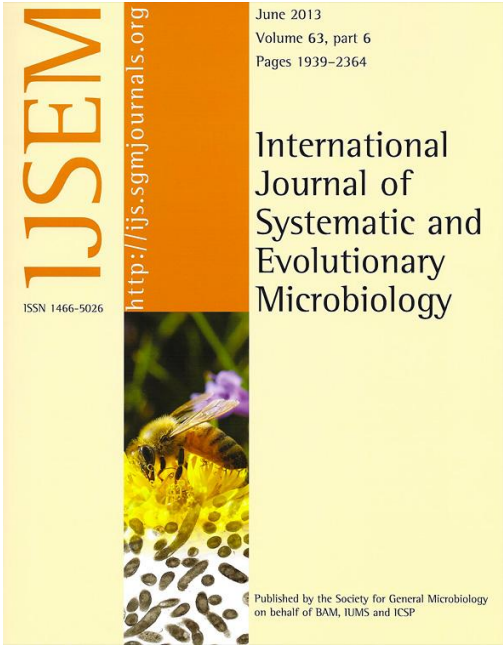

http://www.zo.utexas.edu/faculty/sjasper/images/SO22_07.gif

# Outline

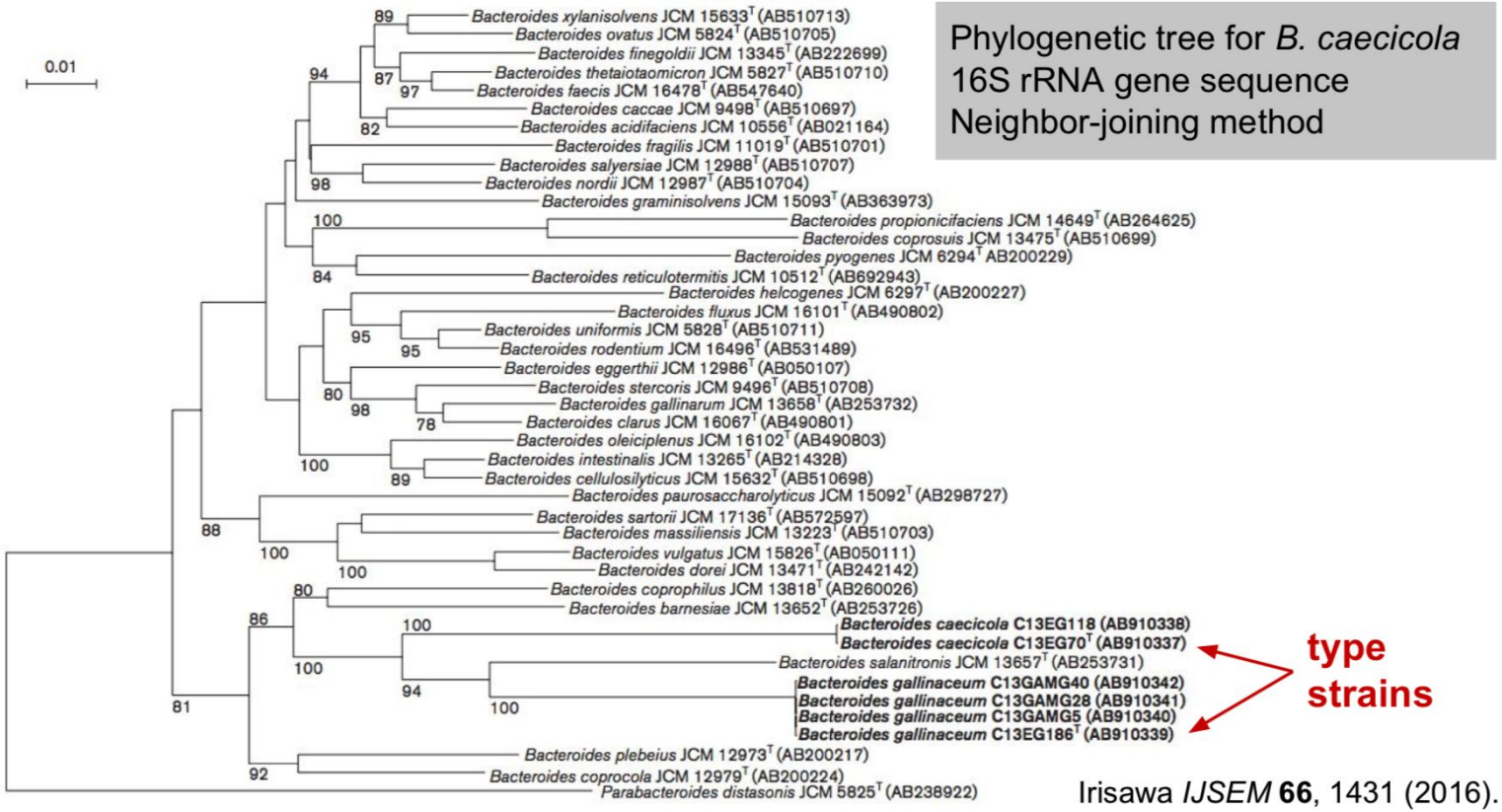
1. A brief background on taxonomy and phylogeny
2. How (microbial) taxonomy is maintained
3. Assigning an unknown sequence to a taxon
    a. One slide on OTUs/ASVs
    b. Methods and challenges
    c. Databases
4. Outline for the workshop

# New bacterial species are published in IJSEM

- Species are defined by a type strain.
- To be a new species, a newly isolated strain must be sufficiently different from an existing type strain
- Requirements:
  - Genomically distinct
  - Distinguishable by phenotype
  - Deposited at two strain collections
- Genus, family, and class level taxa are also named in this system
- bacterio.net has info on named bacterial species, including links to papers and 16S sequences.
- SILVA maintains aligned 16S sequences for named species.




https://pennchopmicrobiome.chop.edu/sites/default/files/field/images/files/taxonomy-bittinger-talk.pdf




Phylogenetic tree for *B. caecicola*
16S rRNA gene sequence
Neighbor-joining method

Irisawa *IJSEM* **66**, 1431 (2016).

https://pennchopmicrobiome.chop.edu/sites/default/files/field/images/files/taxonomy-bittinger-talk.pdf

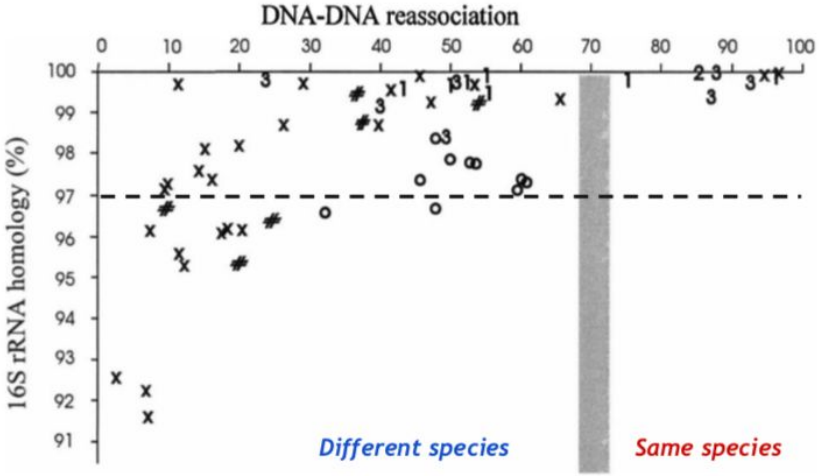
# 16S rRNA gene similarity is sensitive but not selective for new species identification

“16S rRNA gene sequences alone do not describe a species, but may provide the first indication that a novel species has been isolated (less than 97% gene sequence similarity).

Where 16S rRNA gene sequence similarity values are more than 97% (over full pairwise comparisons), other methods such as DNA-DNA hybridization or analysis of gene sequences with a greater resolution must be used.

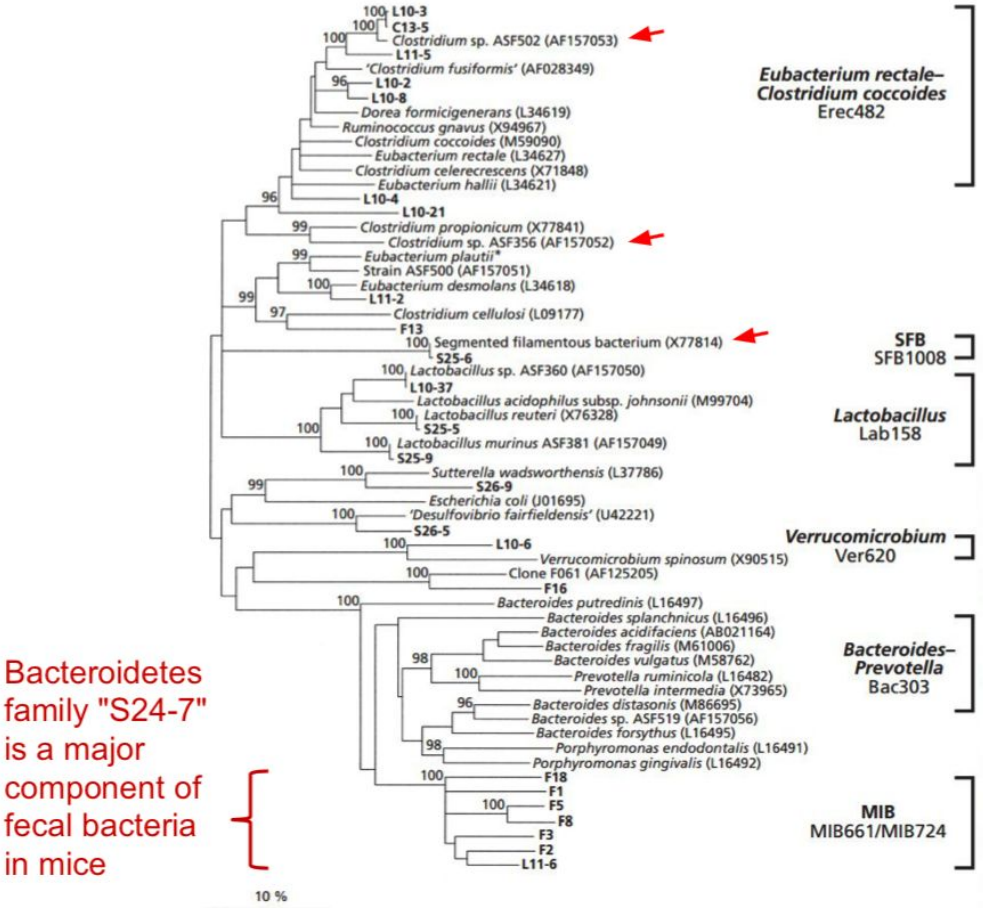
These methods must be correlated with the characterization based on phenotypic tests.”

Tindall *IJSEM* **60**, 249 (2010)




Stackebrandt *IJSEM* **44**, 846 (1994)

https://pennchopmicrobiome.chop.edu/sites/default/files/field/images/files/taxonomy-bittinger-talk.pdf




# Many bacteria are unnamed

- Clostridium sp. ASF502 and ASF356 are part of Altered Schaedler Flora used in research for 40 years
- Segmented filamentous bacteria (SFB) were the topic of high-profile research
- Mouse intestinal bacteria (MIB) may comprise up to 80% of bacteria in feces of lab mice
- "*Candidatus*" is a component of the taxonomic name for a bacterium that cannot be maintained in a culture collection.

# Some bacteria are named inconsistently

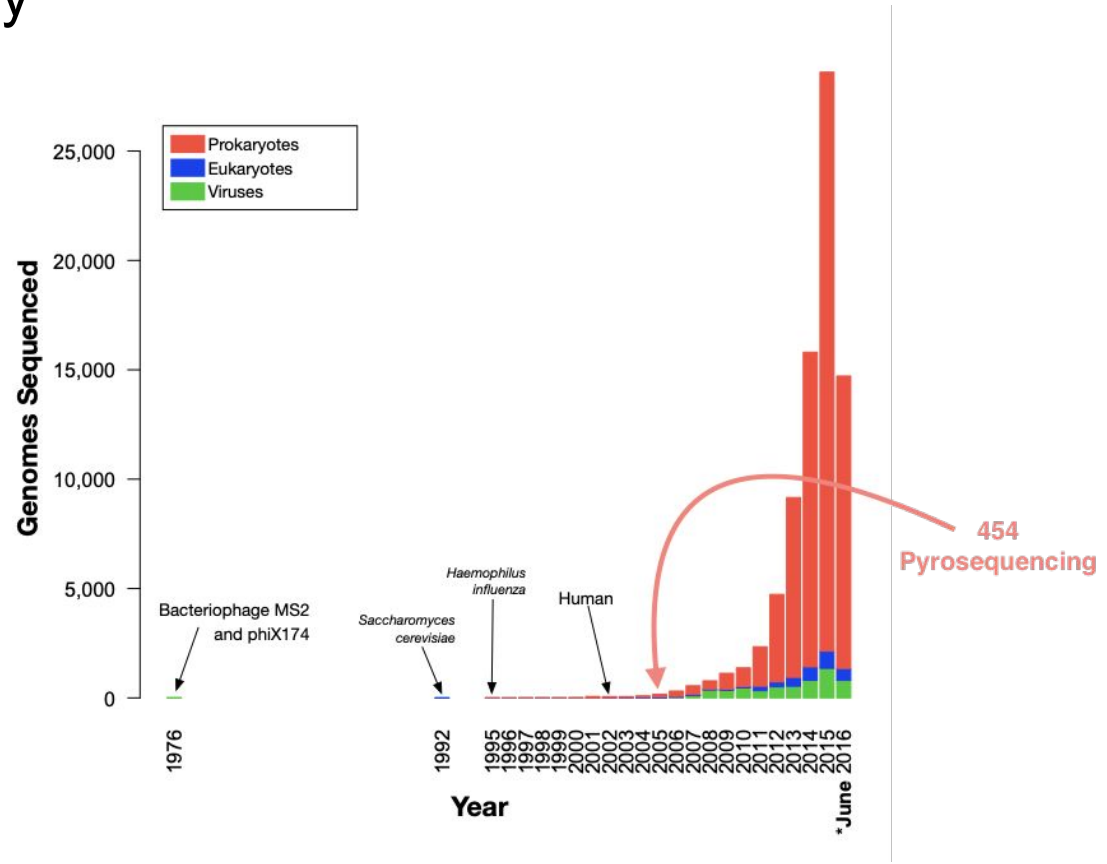
*Eubacterium*, *Clostridium*, *Ruminococcus*

https://pennchopmicrobiome.chop.edu/sites/default/files/field/images/files/taxonomy-bittinger-talk.pdf

Salzman *Microbiology* **148**, 3651 (2002)

# Genomes sequenced annually

- Perhaps an overwhelming number of genomes being sequenced annually
  - Number of genomes sequenced annually doubles just about every year
- Add to this the growing *number* of “metagenome assembled genomes” (MAGs) sequenced and assembled directly from environmental samples
- Our inability to obtain most bacteria in pure culture continues to prevent the development of a robust bacterial taxonomy




Produced from NCBI GenBank files June 2016 - DJ Nasko

# A new taxonomy - GTDB

- If you really want to go down the rabbit hole...

RESOURCE

nature biotechnology

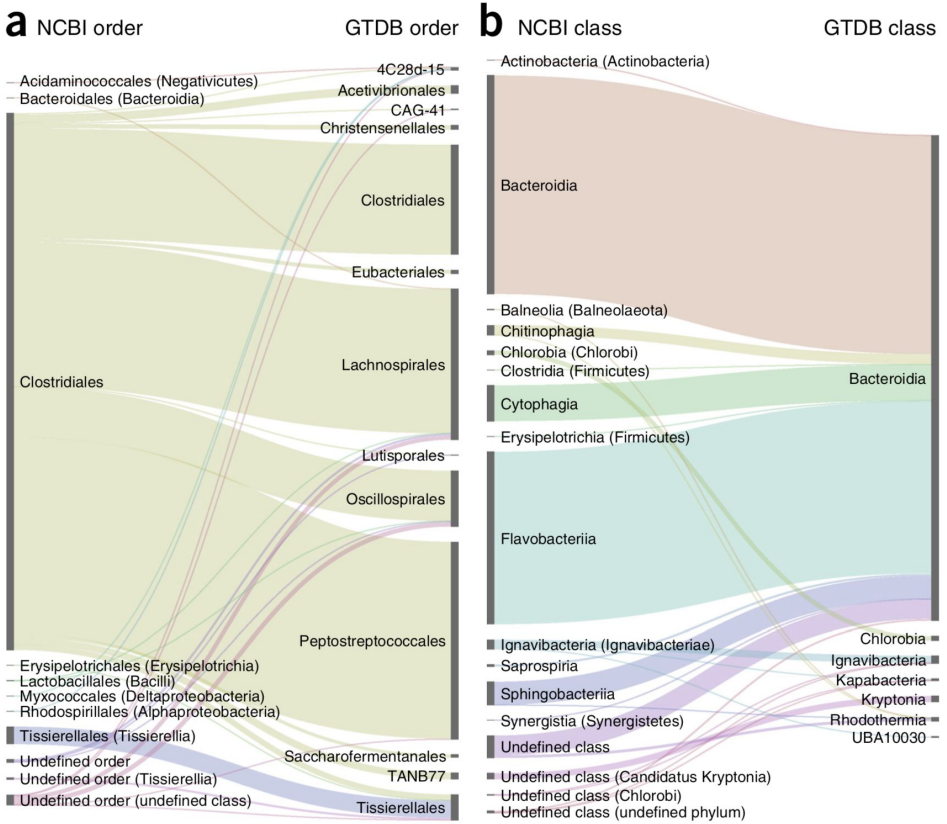
A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life

Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil & Philip Hugenholtz

**Taxonomy is an organizing principle of biology and is ideally based on evolutionary relationships among organisms. Development of a robust bacterial taxonomy has been hindered by an inability to obtain most bacteria in pure culture and, to a lesser extent, by the historical use of phenotypes to guide classification. Culture-independent sequencing technologies have matured sufficiently that a comprehensive genome-based taxonomy is now possible. We used a concatenated protein phylogeny as the basis for a bacterial taxonomy that conservatively removes polyphyletic groups and normalizes taxonomic ranks on the basis of relative evolutionary divergence. Under this approach, 58% of the 94,759 genomes comprising the Genome Taxonomy Database had changes to their existing taxonomy. This result includes the description of 99 phyla, including six major monophyletic units from the subdivision of the Proteobacteria, and amalgamation of the Candidate Phyla Radiation into a single phylum. Our taxonomy should enable improved classification of uncultured bacteria and provide a sound basis for ecological and evolutionary studies.**

The rapid expansion of sequenced bacterial and archaeal genomes in the past decade has enabled the construction of genome-based phylogenies[1–3] suitable for defining taxonomy. A robust taxonomy is needed to accurately describe microbial diversity, to interpret metagenomic data and to provide a common language for communicating scientific results[4]. Sequence-based phylogenetic trees provide a framework for the development of a taxonomy that takes into account both evolutionary relationships and differing rates of evolution. Current microbial taxonomies such as those provided by NCBI[5], SILVA[6], RDP[7], Greengenes[8] and EzTaxon[3] are often inconsistent with evolutionary relationships, because many taxa circumscribe polyphyletic groupings. This inconsistency is partly attributable to historical phenetic. A proposal to standardize taxonomic ranks by using 16S rRNA sequence identity thresholds has identified a high degree of discordance between these thresholds and the SILVA taxonomy[11].

Current microbial taxonomies based on 16S rRNA gene relationships[3,6–8] have several limitations, including low phylogenetic resolution at the highest and lowest taxonomic ranks[14], missing diversity as a result of primer mismatches[15] and PCR-produced chimeric sequences that can corrupt tree topologies by drawing together disparate groups[16]. Trees inferred from the concatenation of single-copy vertically inherited proteins provide higher resolution than those obtained from a single phylogenetic-marker gene[17–19] and are increasingly representative of microbial diversity,



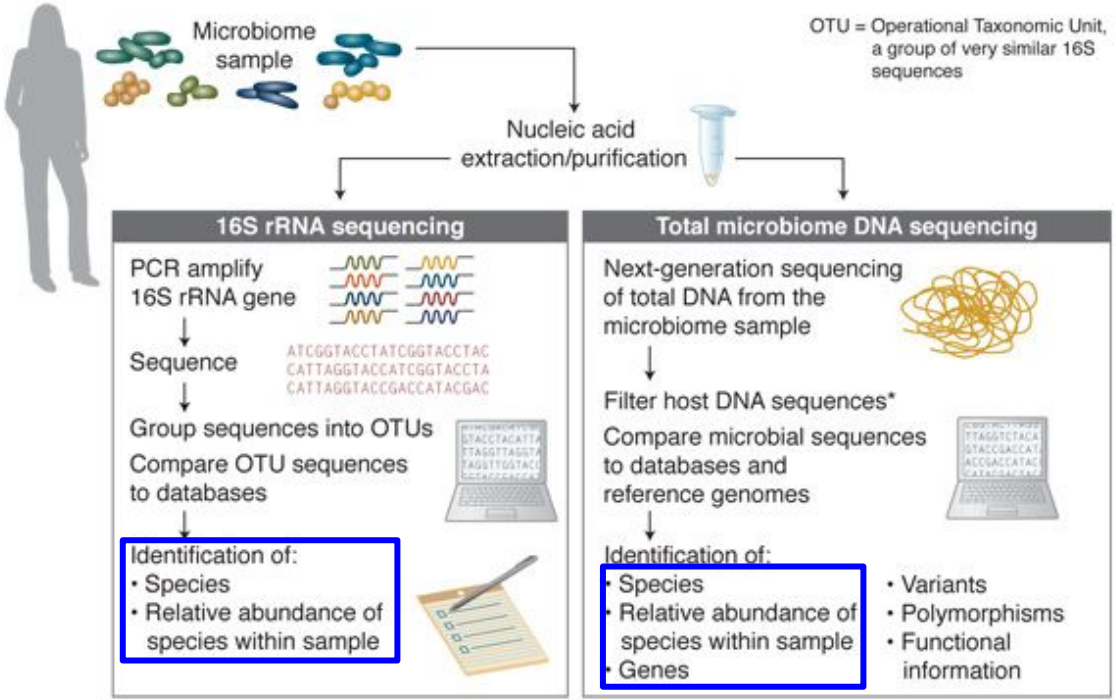

Parks, Donovan H., et al. "A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life." Nature biotechnology (2018).

# Outline

1. A brief background on taxonomy and phylogeny
2. How (microbial) taxonomy is maintained
3. Assigning an unknown sequence to a taxon
    a. One slide on OTUs/ASVs
    b. Methods and challenges
    c. Databases
4. Outline for the workshop

# What is in my microbial community sample?

- Assuming we have a good phylogeny from the tree of life
- Want to know what taxa are in our sample(s)
- How abundant are they?




https://www.neb.com/tools-and-resources/feature-articles/addressing-challenges-in-microbiome-dna-analysis

# Goals of Taxonomic Assignment

- Taxonomic Identification
  - Specifically, *what taxa are in my sample*?
- Abundance Estimation
  - *How many* of each taxa are in my sample (in relative terms)?
- Approach/assumptions depends on the ultimate goals of the project
  - What are you planning on doing with the labels?

# One slide on OTUs and ASVs

**Operational Taxonomic Units (OTUs) are used as a numerical stand-in for species**
**Amplicon Sequence Variants (ASVs) are also used as a numerical stand-in for species**

Coined by Sneath and Sokal in *Principles of Numerical Taxonomy*, 1963

OTUs are clusters of DNA sequences that are assumed to have species-level similarity.
(*ASVs are newer and use more elaborate means to group sequences*)
Why make clusters?

97% similarity threshold used in 16S sequencing taken from Stackebrandt and Goedel *Int. J. Syst. Bacteriol.* **44**, 846 (1994):

"At sequence homology values below about 97.5% it is unlikely that two organisms have more than 60 to 70% DNA similarity and hence they are related at the species level."

# One slide on OTUs and ASVs

Grouping sequences into OTUs and ASVs is one way to organize your data, **but you still need to assign these groups to taxa**

The process of assigning sequences to a taxon is what we'll be covering today, not OTU clustering or ASV grouping.

# Challenges to Taxonomic Assignment

- **Reference database coverage and completeness**-- what if a sequence is not in a database? What can we say about the data?
- **Resolution of different methods**-- e.g. certain hypervariable regions of the 16S gene cannot distinguish between *E. coli* and *Shigella spp*.
- **Computational bottlenecks**-- trade-offs between speed and accuracy
- **Computational reproducibility**-- use of different software/db versions, etc negatively impact ability to reproduce results
- **No ground truth**-- how do we evaluate the results?

# Reference Database Considerations

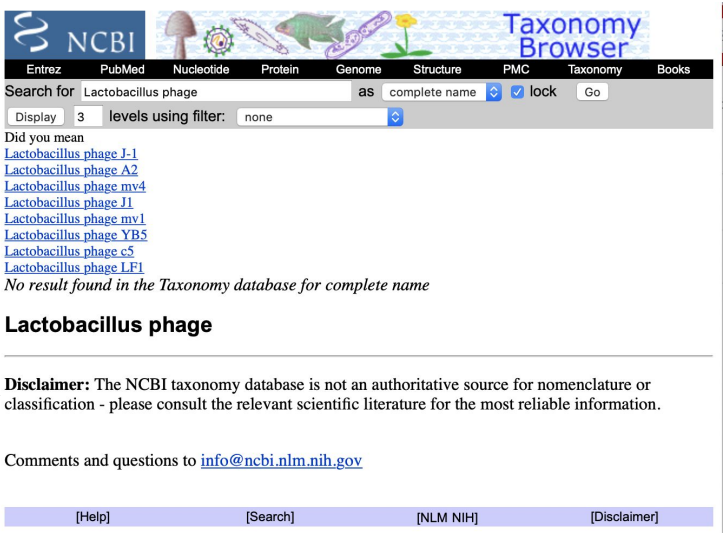
1. **Size-** *Does it contain reference sequences similar to your data? Is it dense or sparse?*
2. **Taxonomy-** *Are the references classified to genus or species level?*
3. **Quality-** *Are there chimeras or low-quality sequences?*

from Sue Huse

# Common Taxonomic Systems for Bacteria

- RDP (https://rdp.cme.msu.edu/)
- Greengenes (http://greengenes.secondgenome.com/)
- SILVA (https://www.arb-silva.de/)
- NCBI (https://www.ncbi.nlm.nih.gov/taxonomy)

Note: NCBI is not an “authoritative source for nomenclature or classification”



# Specialized Databases

- HOMD- Human Oral Microbiome Database (http://www.homd.org)
- OSU CORE for oral (http://microbiome.osu.edu/)
- UNITE- Fungal ITS (http://unite.ut.ee/)
- Virus-Host Database (https://www.genome.jp/virushostdb/)
- EuPathDB - Eukaryotic pathogens (https://eupathdb.org/eupathdb/)

from Sue Huse

# Challenges to Taxonomic Assignment

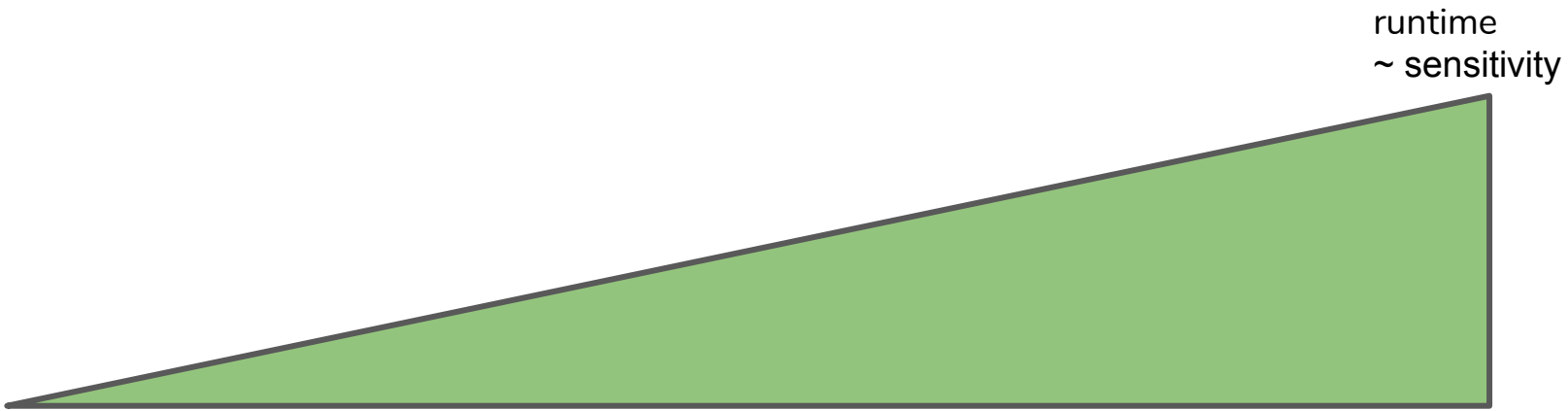
- **Reference database coverage and completeness**-- what if a sequence is not in a database? What can we say about the data?
- **Resolution of different methods**-- e.g. certain hypervariable regions of the 16S gene cannot distinguish between *E. coli* and *Shigella spp.*, read-based versus gene-based versus contig-based assignment
- **Computational bottlenecks**-- trade-offs between speed and accuracy
- **Computational reproducibility**-- use of different software/db versions, etc negatively impact ability to reproduce results
- **No ground truth**-- how do we evaluate the results?


Cross-stitch by Dr. Jennifer Glass
https://twitter.com/methanojen/status/1033749220396814336

# Challenges to Taxonomic Assignment

- **Reference database coverage and completeness**-- what if a sequence is not in a database? What can we say about the data?
- **Resolution of different methods**-- ie. certain hypervariable regions of the 16S gene cannot distinguish between *E. coli* and *Shigella spp.*, read-based versus gene-based versus contig-based assignment
- **Computational bottlenecks**-- trade-offs between speed and accuracy
- **Computational reproducibility**-- use of different software/db versions, etc negatively impact ability to reproduce results
- **No ground truth**-- how do we evaluate the results?

# Landscape of Taxonomic Approaches




**k-mer counting**
kraken

**k-mer stat model**
RDP

**Sequence Alignment**
blast
usearch
diamond

**Phylogenetic**
pplacer
TIPP

Assume taxonomically related sequences have similar k-mer frequencies

Assume taxonomically related sequences are similar

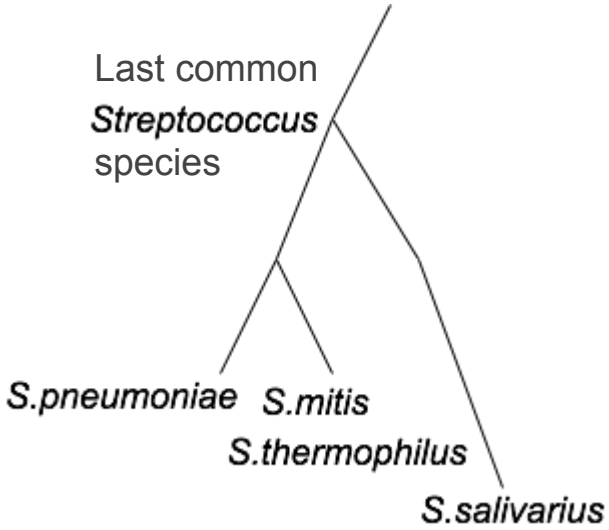
Simulate evolution

# Challenges to Taxonomic Assignment

- **Reference database coverage and completeness**-- what if a sequence is not in a database? What can we say about the data?
- **Resolution of different methods**-- ie. certain hypervariable regions of the 16S gene cannot distinguish between *E. coli* and *Shigella spp.*, read-based versus gene-based versus contig-based assignment
- **Computational bottlenecks**-- trade-offs between speed and accuracy
- **Ambiguous assignments**
- **Computational reproducibility**-- use of different software/db versions, etc negatively impact ability to reproduce results
- **No ground truth**-- how do we evaluate the results?

# Dealing with Ambiguities

- Common solution is to use the most recent common ancestor

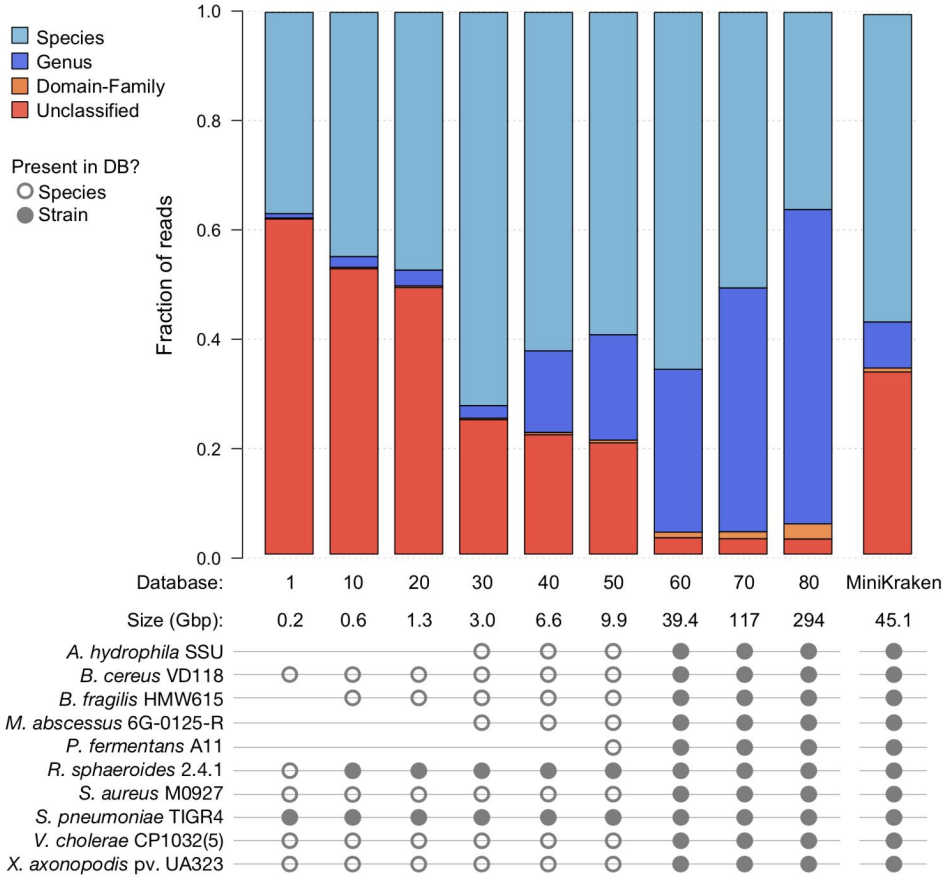



| Common Name | Human | Common Chimpanzee | Grey Wolf | Tiger Snake | Monarch Butterfly |
|---|---|---|---|---|---|
| Domain | Eukaryota | Eukaryota | Eukaryota | Eukaryota | Eukaryota |
| Kingdom | Animalia | Animalia | Animalia | Animalia | Animalia |
| Phylum | Chordata | Chordata | Chordata | Chordata | Arthropoda |
| Class | Mammalia | Mammalia | Mammalia | Reptilia | Insecta |
| Order | Primates | Primates | Carnivora | Squamata | Lepidoptera |
| Family | Hominidae | Hominidae | Canidae | Elapidae | Nymphalidae |
| Genus | Homo | Pan | Canis | Notechis | Danaus |
| Species | Homo sapiens | Pan troglodytes | Canis lupus | Notechis scutatus | Danaus plexippus |

Match to Human + Grey Wolf: Best call *Mammalia*
Match to Tiger Snake + Monarch Butterfly: Best call *Animalia*

# Dealing with Ambiguities

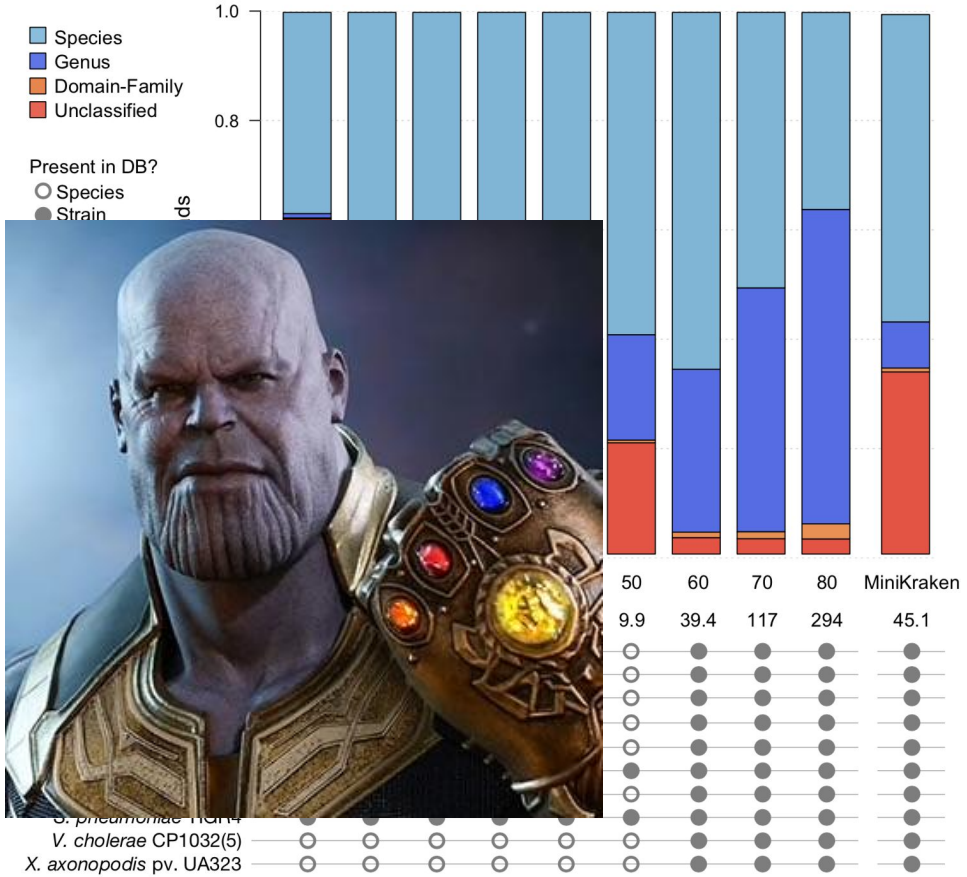
- Common solution is to use the most recent common ancestor.
- Certain genera are being sequenced more times than others.
- We have too many of the same few bugs in the database




Nasko et al. *Genome Biology* 2018

# Dealing with Ambiguities

- Common solution is to use the most recent common ancestor.
- Certain genera are being sequenced more times than others.
- We have too many of the same few bugs in the database



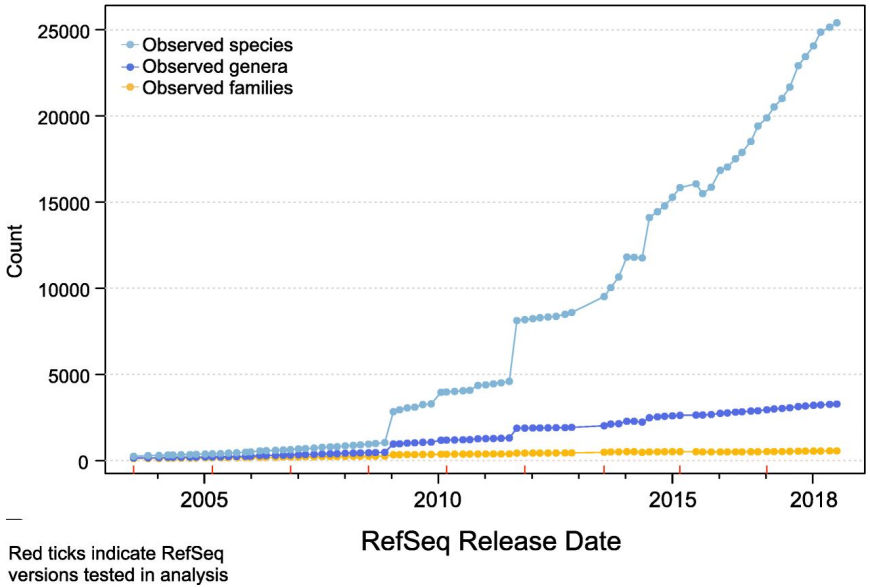

https://www.sideshowtoy.com/collectibles/marvel-thanos-hot-toys-903429/

Nasko et al. *Genome Biology* 2018

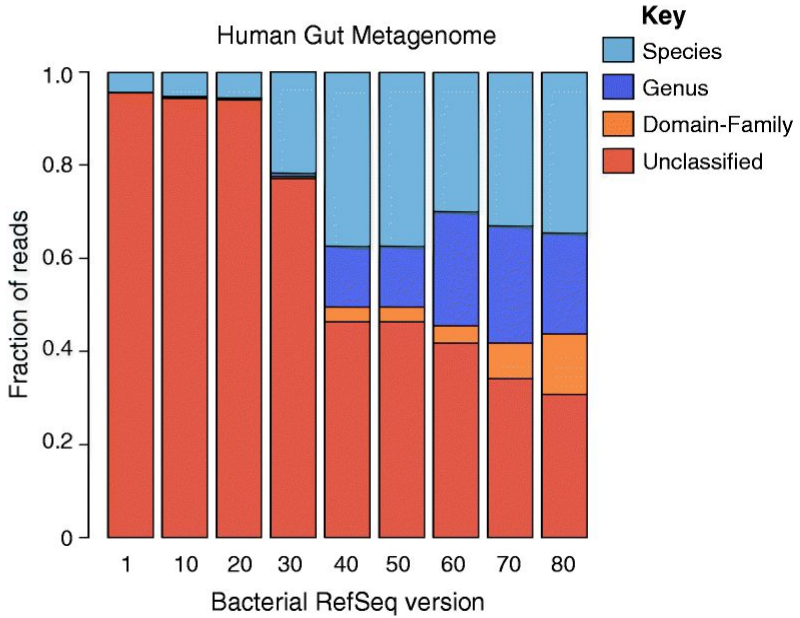
# Challenges to Taxonomic Assignment

- **Reference database coverage and completeness**-- what if a sequence is not in a database? What can we say about the data?
- **Resolution of different methods**-- ie. certain hypervariable regions of the 16S gene cannot distinguish between *E. coli* and *Shigella spp.*, read-based versus gene-based versus contig-based assignment
- **Computational bottlenecks**-- trade-offs between speed and accuracy
- **Ambiguous assignments**
- **Computational reproducibility**-- use of different software/database versions, etc negatively impact ability to reproduce results
- **No ground truth**-- how do we evaluate the results?

# RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification




Red ticks indicate RefSeq versions tested in analysis




Nasko et al. *Genome Biology* 2018

# Singularity containers

Different versions of software can produce different results.

Singularity images reduce the barriers to running the same versions of code on any platform.

We have singularity images for most programs that we'll cover today and you're welcome to copy them to your laptops and take them back to your institution.

# Challenges to Taxonomic Assignment

- **Reference database coverage and completeness**-- what if a sequence is not in a database? What can we say about the data?
- **Resolution of different methods**-- ie. certain hypervariable regions of the 16S gene cannot distinguish between *E. coli* and *Shigella spp.*, read-based versus gene-based versus contig-based assignment
- **Computational bottlenecks**-- trade-offs between speed and accuracy
- **Ambiguous assignments**
- **Computational reproducibility**-- use of different software/db versions, etc negatively impact ability to reproduce results
- **No ground truth**-- how do we evaluate the results?

# Evaluation

CAMI (Critical Assessment of Metagenome Interpretation)

Mock Communities

What do names mean after all?

Sczyrba, Alexander, et al. "Critical assessment of metagenome interpretation—a benchmark of metagenomics software." Nature methods 14.11 (2017): 1063.

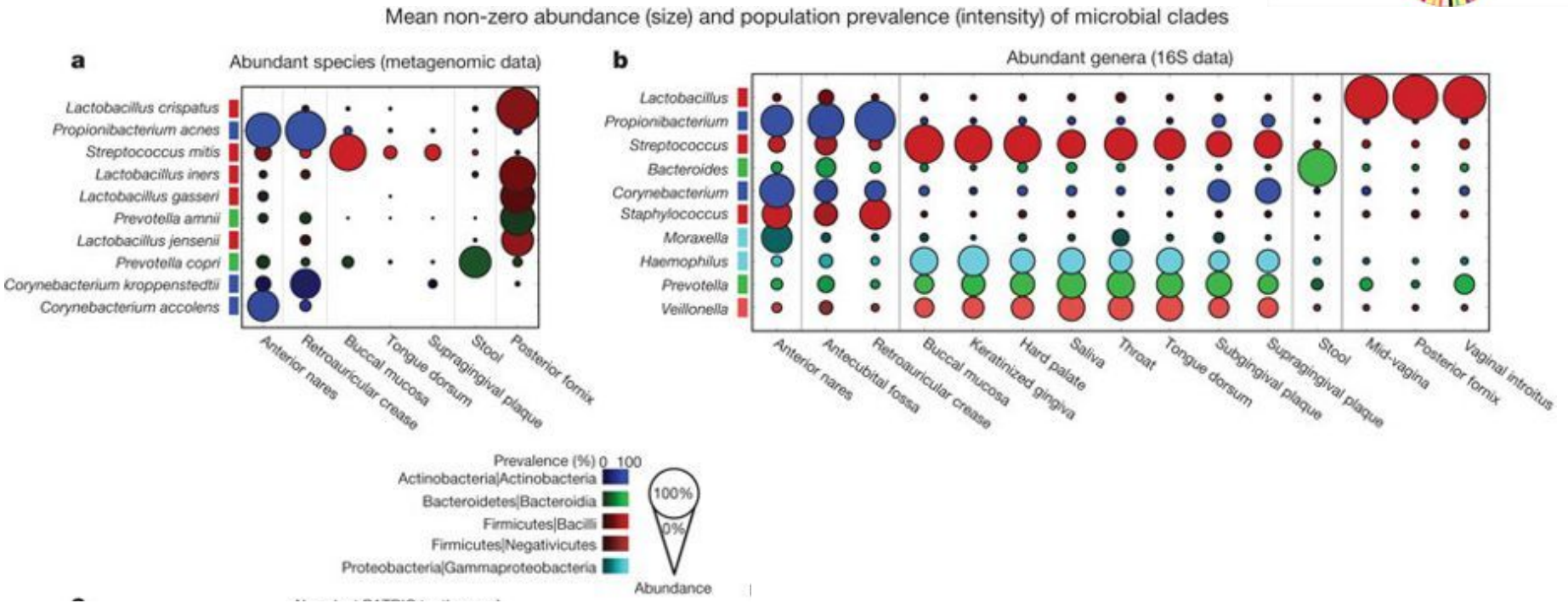
# Outline

1. A brief background on taxonomy and phylogeny
2. How (microbial) taxonomy is maintained
3. Assigning an unknown sequence to a taxon
    a. One slide on OTUs/ASVs
    b. Methods and challenges
    c. Databases
4. Outline for the workshop

# What are we going to cover today?

- **Machine Learning Approaches**-- RDP
- **Fast Metagenomic Profiling Methods**-- kraken
- **Database Searching**-- BLAST
- **Phylogenetic Methods**-- TIPP
- **How to critically interpret and evaluate taxonomic output**

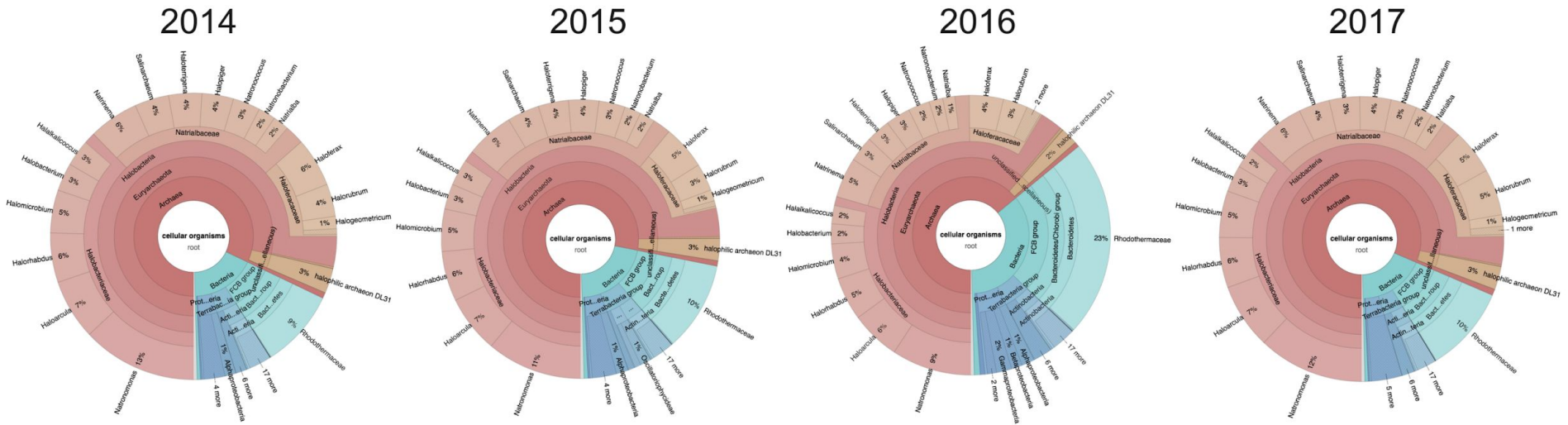
# Datasets: Human Microbiome Project




C Huttenhower *et al. Nature* **486**, 207-214 (2012) doi:10.1038/nature11234

nature

# Datasets: Halite (Salt Rock) from the Atacama Desert, Chile




Special thanks to Gherman Uritskiy and Jocelyne DiRuggiero for providing the data!
Check out the preprint: https://www.biorxiv.org/content/early/2018/10/13/442525