NamesforLife
*Bringing meaning to life ...*

# Taxonomic inference vs. Ground truth

George M. Garrity and Charles T. Parker
Department of Microbiology and Molecular Genetics,
Michigan State University and NamesforLife, LLC,
East Lansing, MI USA

KMB 2018
45th Annual Meeting & International Symposium
The Korean Society for Microbiology & Biotechnology

## Ground Truth

## Reproducibility

## Standards

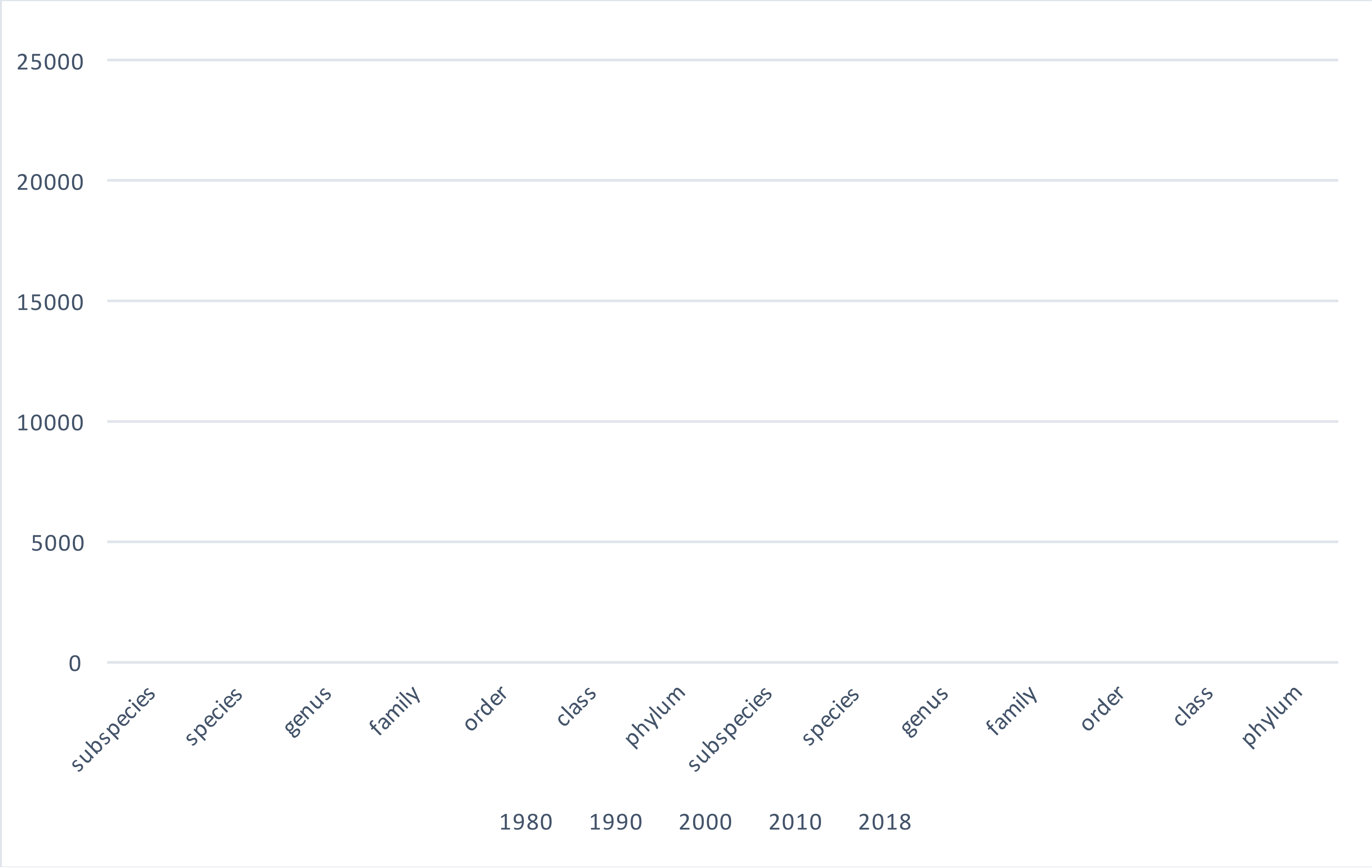
***History does not repeat itself, but it rhymes.***

***Mark Twain***

- Core activities
  - Characterization (description)
  - Classification
  - Identification

- Core resources
  - Literature and databases*
  - Reference materials
  - Tools
    - Hardware and software
    - SOPs and workflows

# It started with a simple question



Would reannotation of taxonomic reference files be useful?

# usearch guided reannotation

**Manual review of fasta taxonomic annotations**

- Adjustment of taxonomic depth (seven levels)
- Convert annotations to usearch format
- Removal of eukaryote, plastid and cyanobacterial squences

**Reassignment of sequence identity**

- Classification of relabeled sequences using usearch sintax function
  - NamesforLife type strain database as reference (April 2018 release)
  - Default cutoff value (pseudo-bootstrap of 80)
  - Sequence identified at all seven levels - reassign
  - Sequence identified at 5 or six levels – reassign if mean score > 80
  - Sequence identified at 4 or less levels – retain original identity
- Correct reassigned names
  - Comparison of species level identity to NamesforLife nomenclature

**Results**

- Eliminate virtually all taxa with multiple parents found in source files
- Increase number of correctly identified (high scoring) 16S sequences

**However,**

- None of the reference databases cover >82.8% of the validly published bacteria and archaea

# The microbiome experiment

**Hypothesis** – are  OTU - OTU and  OTU - taxon name consistent and meaningful when different reference taxonomies are applied?

**Input data**

eight diverse Illumina 16S (v4) metagenome samples

**Software**

mothur version 1.39, variation of Schloss' MiSeq SOP

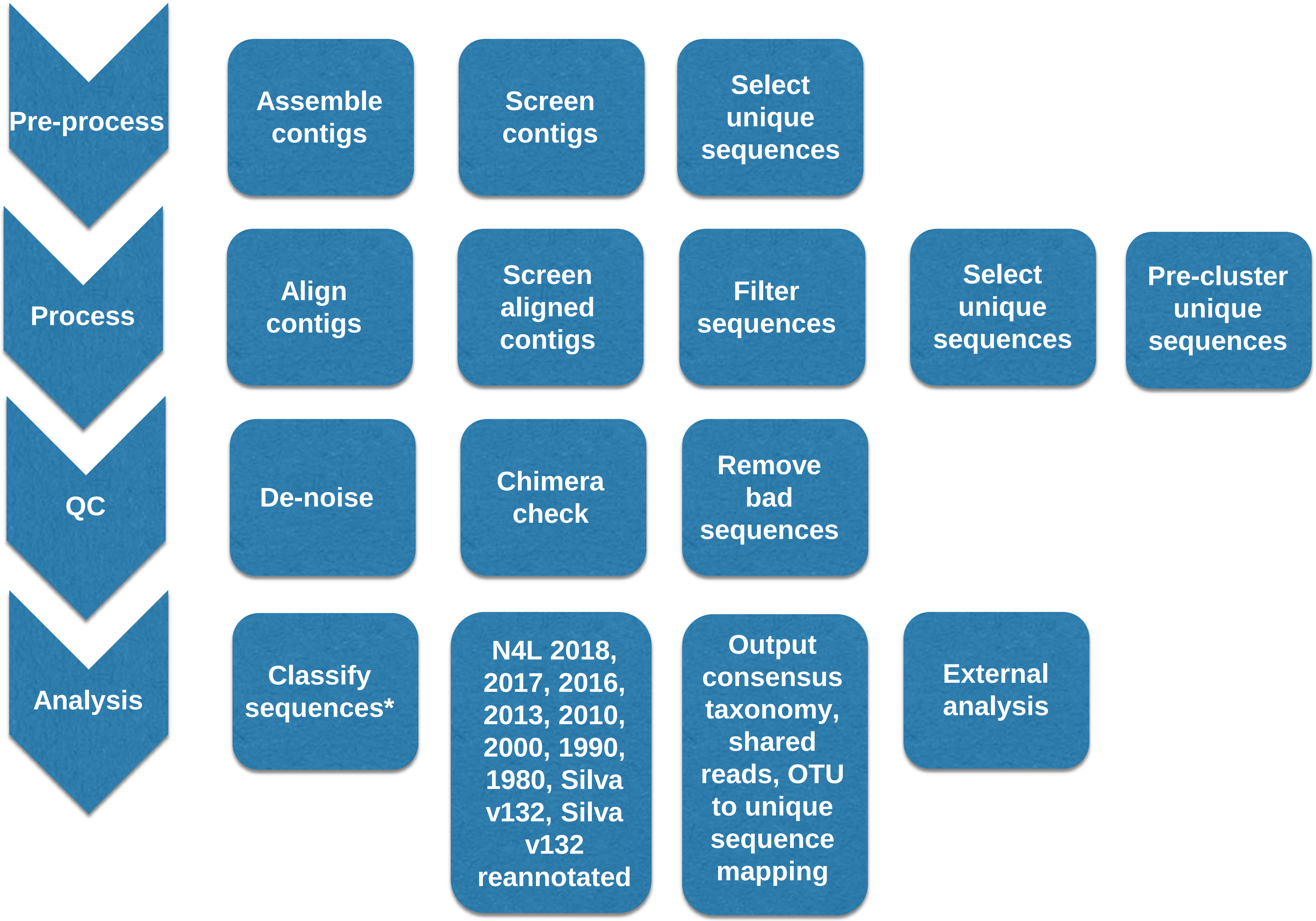
**Hardware**

Mac Pro 3.7 GHz Quad Core Intel Xeon E5, 32GB RAM/SSD

**Reference taxonomies**

NamesforLife type strain database (May 17, 2018 release)
Silva nr_v.132 (trimmed, using both original annotation and reannotated)

**Analysis** – Principle Coordinate Analysis
Non-metric Multidimensional Scaling
Test for significance - Kolmogorov – Smirnov



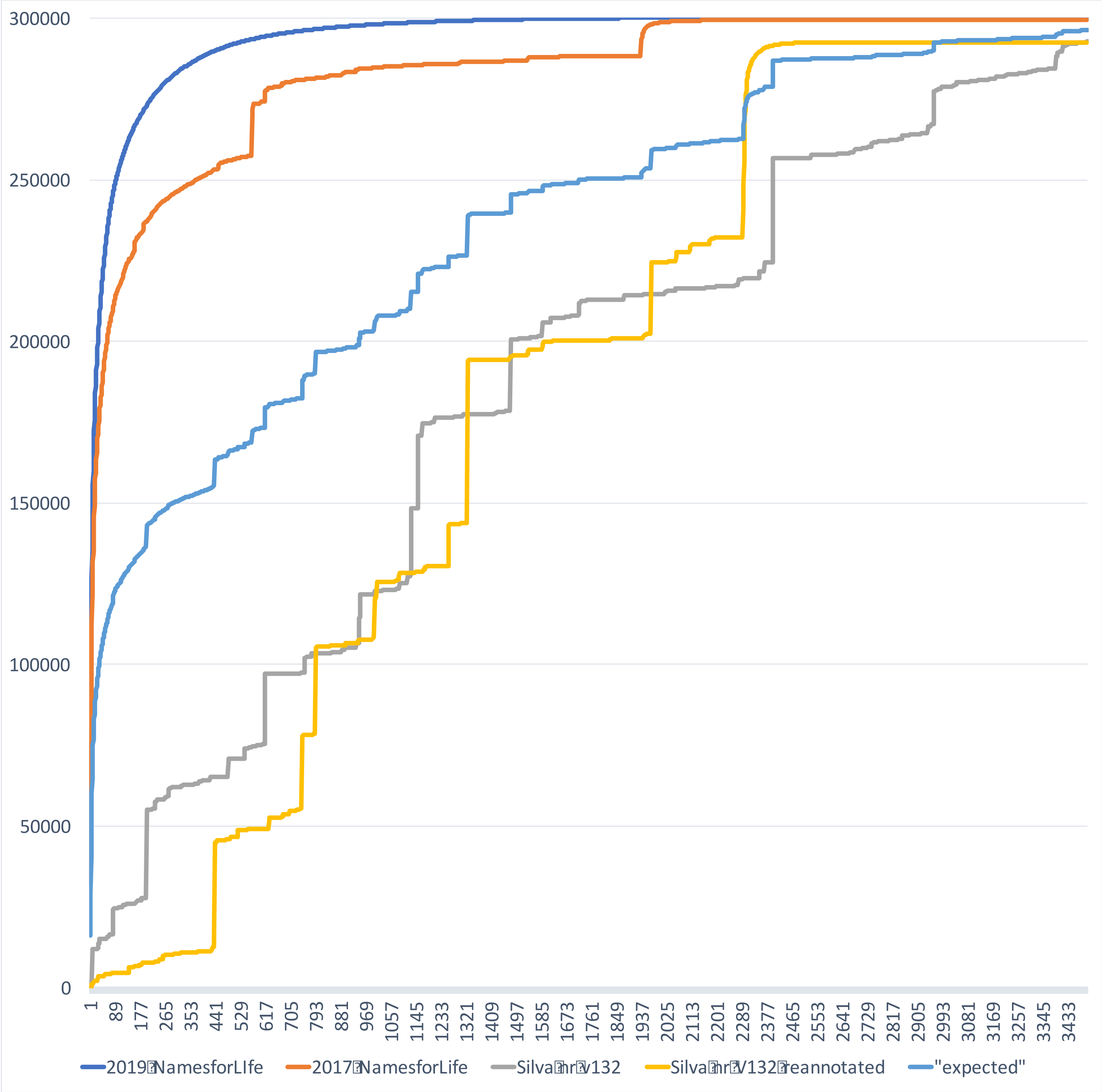
**Pre-process:** Assemble contigs → Screen contigs → Select unique sequences

**Process:** Align contigs → Screen aligned contigs → Filter sequences → Select unique sequences → Pre-cluster unique sequences

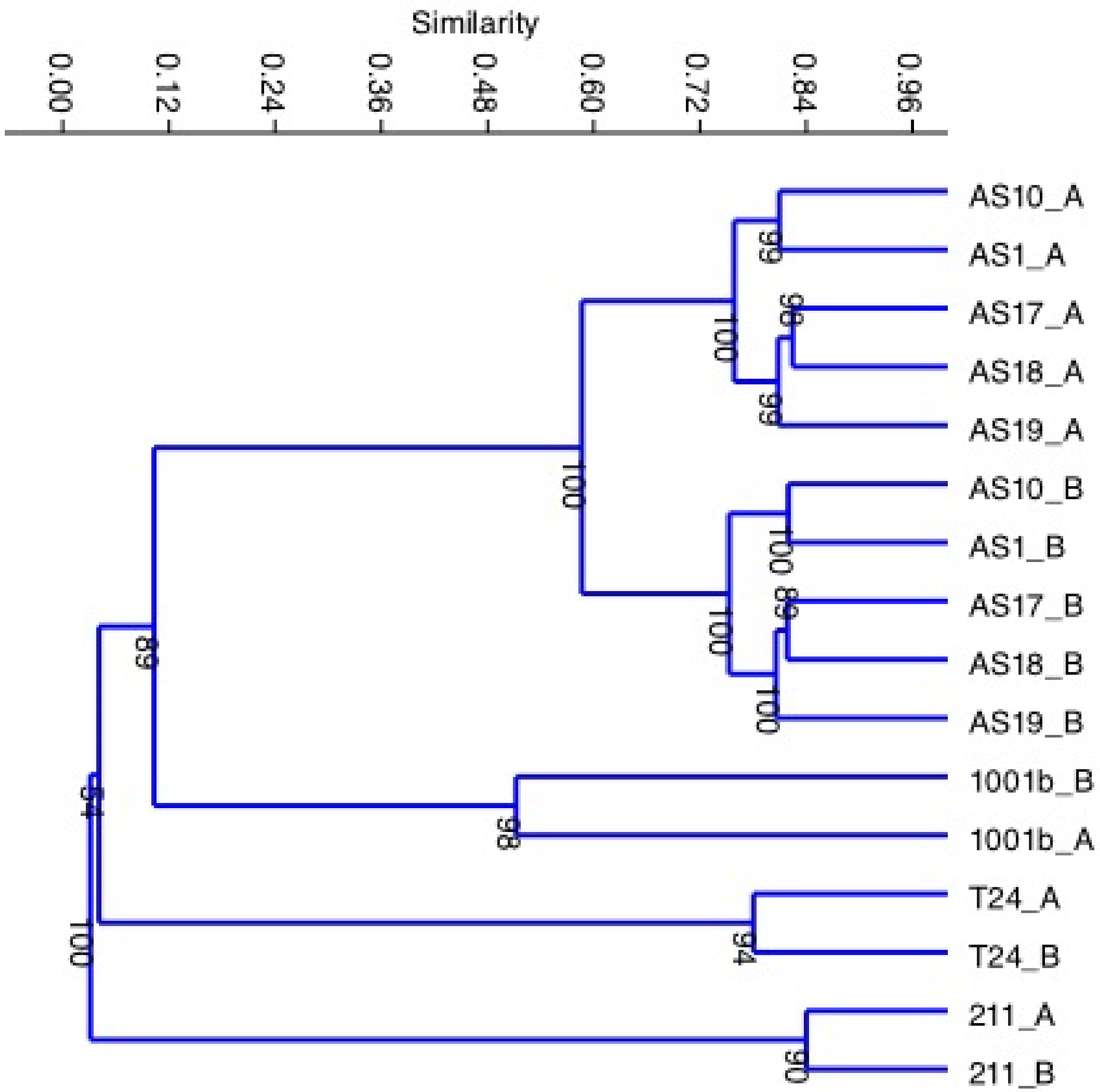
**QC:** De-noise → Chimera check → Remove bad sequences

**Analysis:** Classify sequences* → N4L 2018, 2017, 2016, 2013, 2010, 2000, 1990, 1980, Silva v132, Silva v132 reannotated → Output consensus taxonomy, shared reads, OTU to unique sequence mapping → External analysis

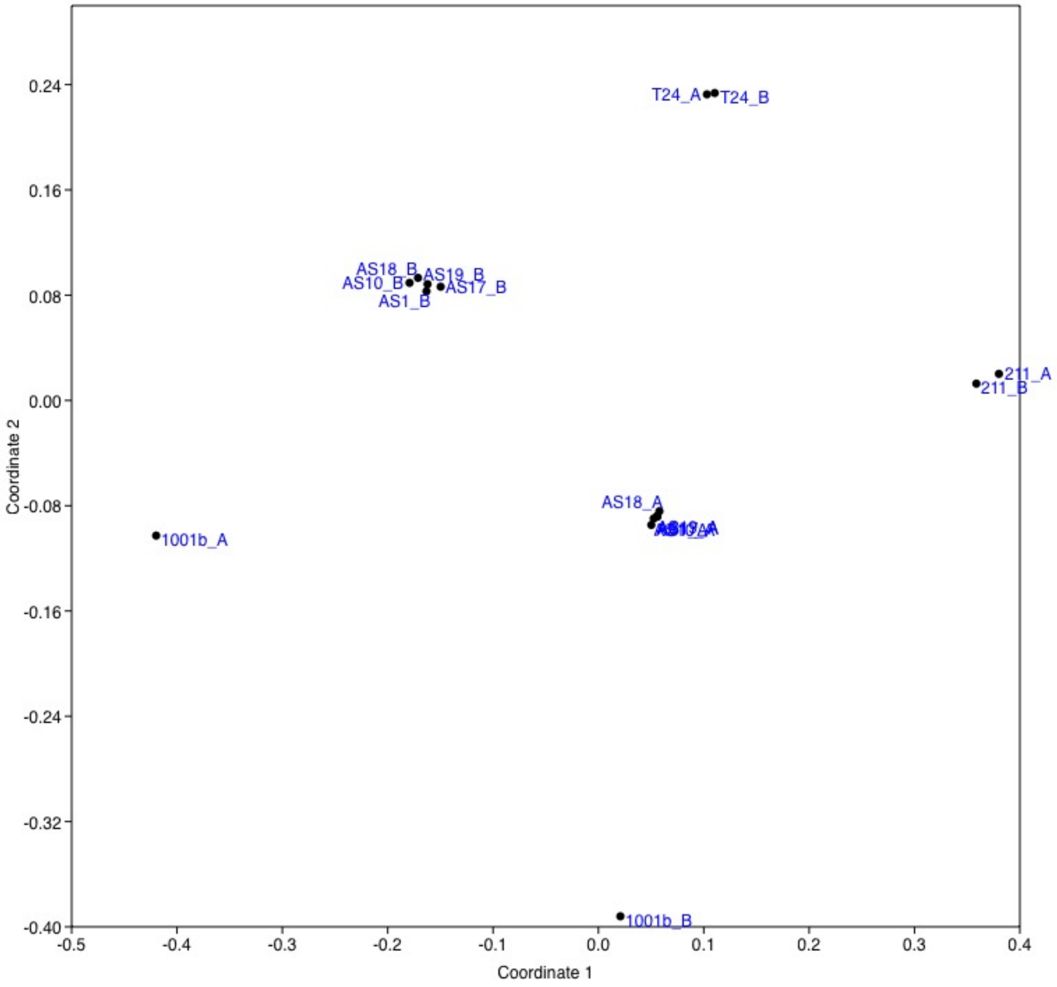
# Cumulative taxonomic abundance, combined metagenome samples compared using four taxonomies

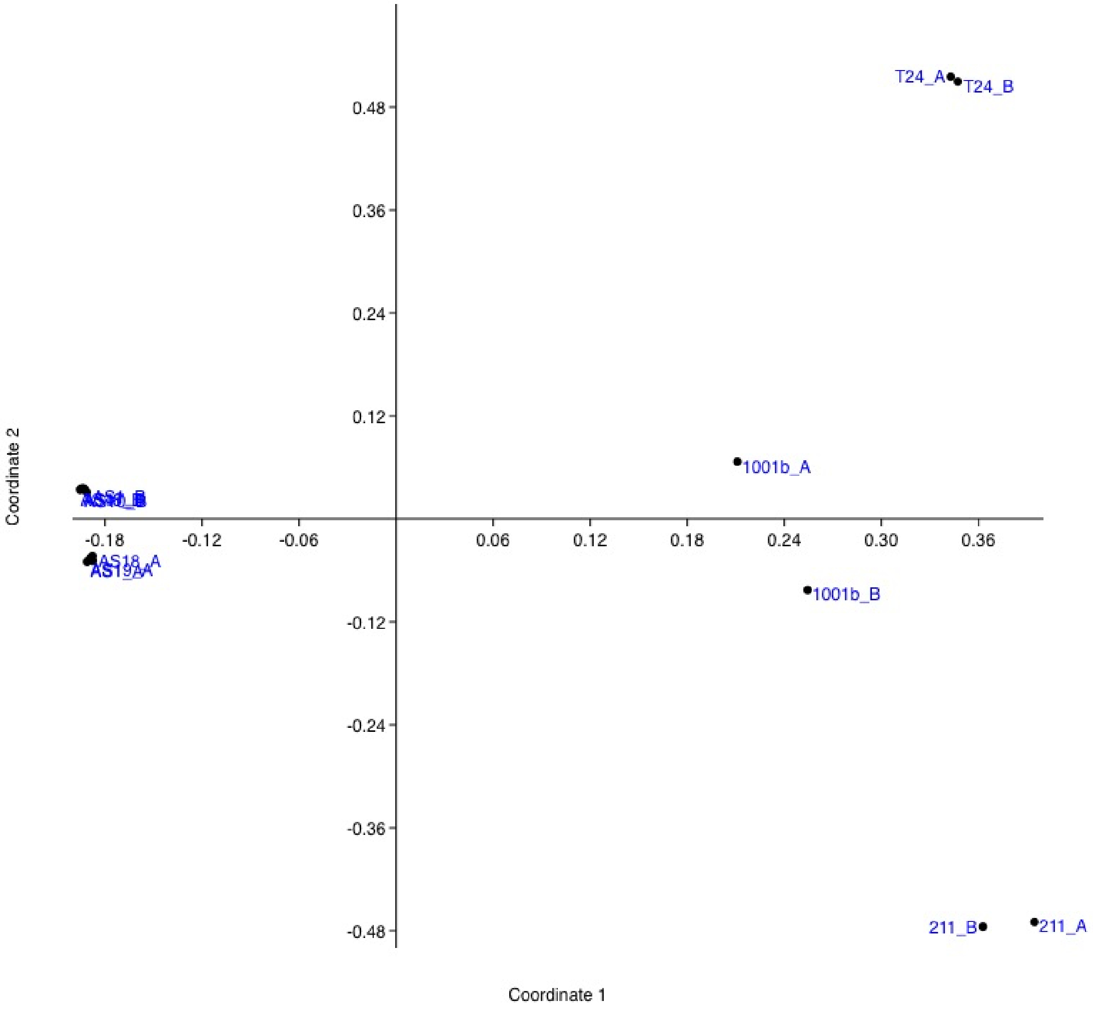


# Bray-Curtis distance, 2018 vs 2017 taxonomy, UPGMA, bootstrapped 500 iterations



# NMDS and PCoA of metagenome analysis using 2017 and 2018 taxonomies





# The analysis

**The goal** – objective way of comparing two or more taxonomies applied to the same metagenome samples

$H_0$ – no difference between taxonomies

$H_a$ – taxonomies different, comparison requires reannotation using same taxonomy

Nature of metagenome and taxonomic data – nonparametric, unbounded

The Kolmogorov-Smirnov Test (2 sample)

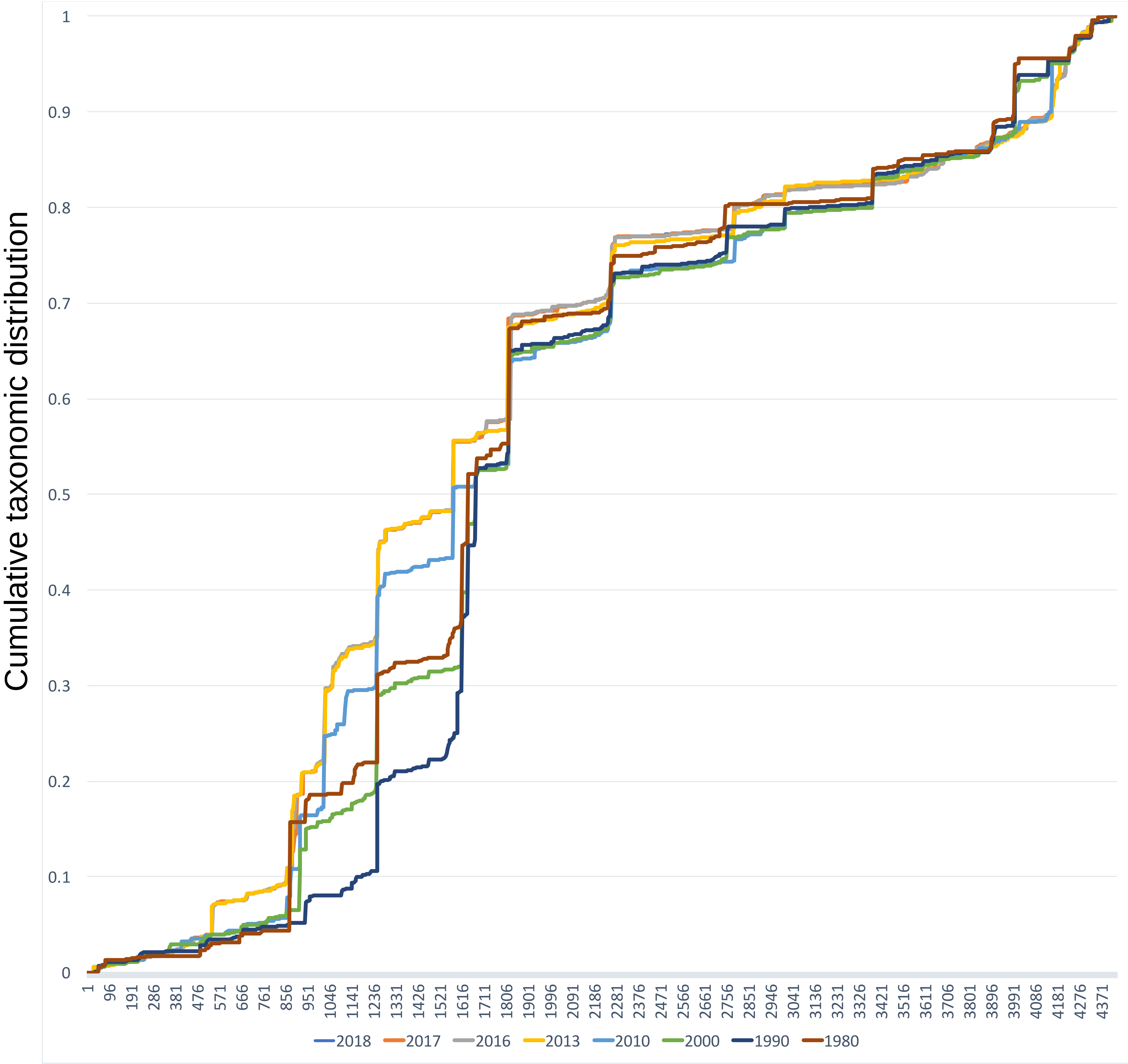
1. Test whether two samples come from the same/different distributions
2. Cumulative distribution of named reads are compared
3. KS test statistic is estimated

$$\mathrm{KS} = \surd|TD_{1,n} - TD_{2,n}|_{\mathrm{max}}/\mathrm{n}$$

where

TD = cumulative taxonomic distribution measured for differences between paired taxon frequencies

n = number of unique taxa in sample


Cumulative taxonomic distribution
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
96
191
286
381
476
571
666
761
856
951
1046
1141
1236
1331
1426
1521
1616
1711
1806
1901
1996
2091
2186
2281
2376
2471
2566
2661
2756
2851
2946
3041
3136
3231
3326
3421
3516
3611
3706
3801
3896
3991
4086
4181
4276
4371
2018
2017
2016
2013
2010
2000
1990
1980

# What we found

Of the possible pairwise comparisons of taxonomic distributions for the amplicon metagenomics data, only two distributions were considered the same: 2018-2017.  All others were significantly different from one another.

Comparison of samples required reanalysis with and against the same taxonomic file to ensure that any differences between samples are due to biological or environmental factors.

Results of analyses and any assertions of novel taxa or functions may not be meaningful if an out-of-date reference taxonomy is used.

Given the rate of change, taxonomic reference files more than one year old should be reannotated prior to use.

Reproducibility – it depends
Replicability – it depends
Generalizability – it depends

# The ANI experiment

**Objective** – Reproduce previous comparative studies using $ANI_M$, $ANI_B$, $ANI_{BBH}$, and extend to $ANI_G$ and AAI

**Hypothesis** – closely related strains will have higher ANI/AAI scores. Same strain will have identical score.

$$ANI_{A\&B} = \sum(\%identity * alignment\ length)/ \sum(length\ genes\ in\ genome\ A)$$

$$AF_{A\&B} = \sum(length\ BBH\ genes)/ \sum(length\ genes\ in\ genome\ A)$$

Differences among methods regarding source/quality of genomes

- coding vs. non-coding
- Use protein coding genes only vs. relaxed approach
- Plasmid and other extrachromosomal genes removed
- Self contained vs. external calls to 3rd party software or services

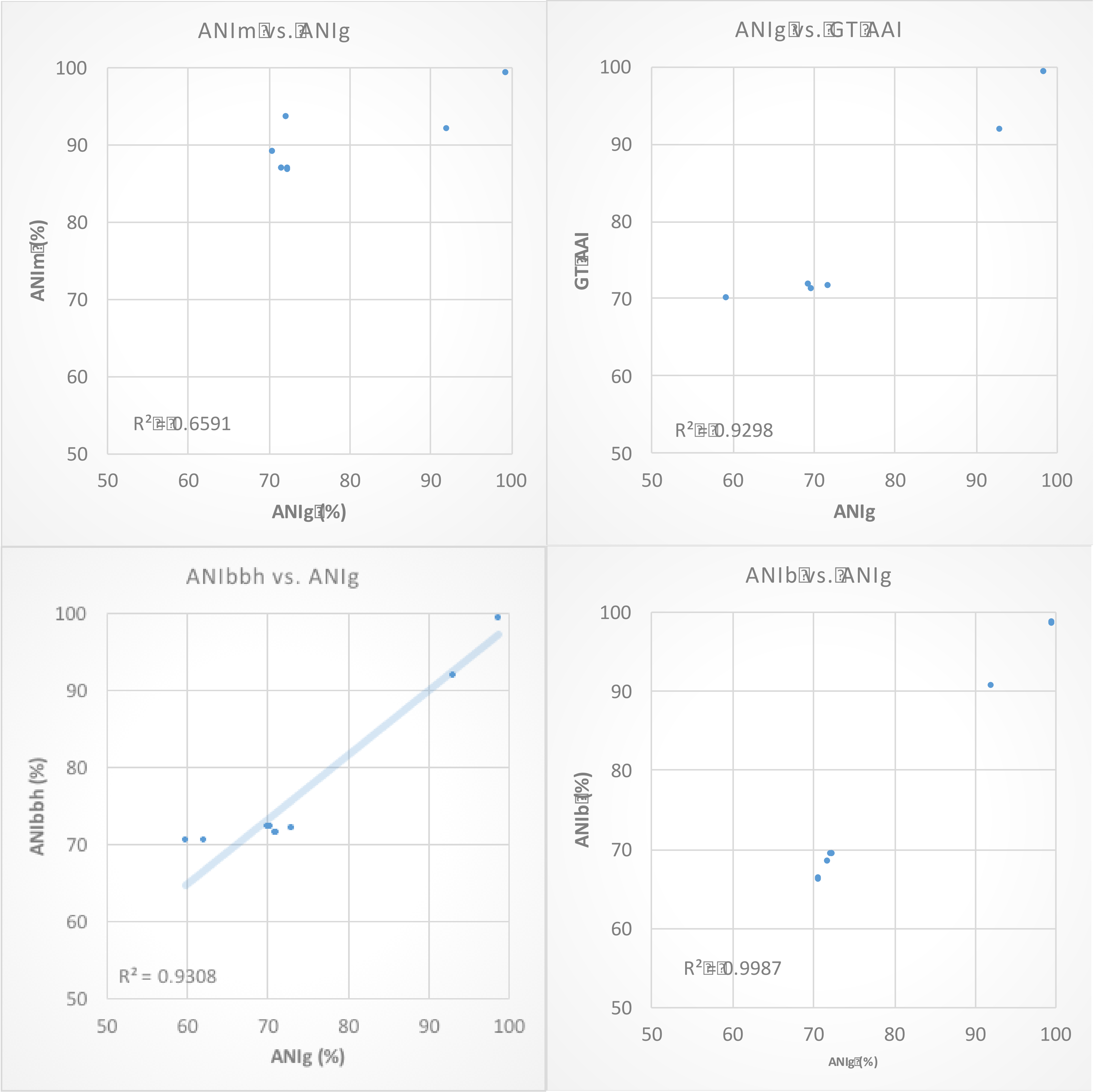
Output is % similarity (A-> B, B->A), but may not always be transitive.

- Establish thresholds for identifying species/subspecies
- At this time the method may not yet be useful for identification and classification.

**NamesforLife**

*Bringing meaning to life ...*

# Comparison of major ANI algorithms, reanalysis of Krebs reference genome collection



KMB 2018
45th Annual Meeting & International Symposium
The Korean Society for Microbiology & Biotechnology

# Findings

**Objective** – Reproduce previous comparative studies using $ANI_M$, $ANI_B$, $ANI_{BBH}$, and extend to $ANI_G$ and AAI

Problems in establishing exact input sequences

Some results were ambiguous across methods

- Idea of fixed cut-offs problematic but thresholds may be useful (e.g. MiSi)
- In ongoing studies comparing true replicate genome sequences, ANI methods rarely show identity.
- Effect of sample preparation, sequencing, assembly and annotation methods likely to prove important or significant.

Current thoughts

- Documentation of source materials and methods are frequently inadequate or lacking
- Documentation of ANI method and version, date of web service used, any methodological or analytical variations needed to correctly interpret results

Reproducibility – sometimes
Replicability – sometimes
Generalizability – not yet

# Acknowledgments

JGI

Nikos Kyrpides
Neha Varghese
Emily Eloe-Fadrosh

USC University of Southern California

Roman Barco

UAMS

Dave Ussery
Michael Robeson
Trudy Wassenar

MICHIGAN STATE UNIVERSITY

Terry Marsh
Ashley Shade
John Guittar
Keven Petersen

*NamesforLife*
*Bringing meaning to life ...*

Charles T. Parker
Nicole Osier
Vo Phan Chuong
Dorothea Taylor
Kara Mannor

Sarah Wigley
Nicole Osier
Grace Rodriguez
Amber Roberts
Danny Bakoz

muci MICHIGAN UNIVERSITIES COMMERCIALIZATION INITIATIVE

MICHIGAN ECONOMIC DEVELOPMENT CORPORATION

SIMB Society for Industrial Microbiology and Biotechnology

U.S. DEPARTMENT OF ENERGY | Office of Science

KMB The Korean Society for Microbiology and Biotechnology

This work was funded through the Small Business Technology Transfer program of the United States Department of Energy under grants number DE-FG02-07ER86321 and DE-SC0006191. Funding for business development was provided through grants and loans from the Michigan Economic Development Corporation and the Michigan Universities Commercialization Initiative. NamesforLife semantic resolution technology is covered under US Patent 7,925,444 B2. The SOSCC systems and methods is covered under US Patent 8,036,997. Semiotic fingerprinting is covered under US 8,903,824. Semantic tagging, indexing, equivalency mapping, and latent semantic analysis of molecular sequence data are the subjects of pending US and EPO patent applications.