TAXONOMY
STAMPS 2016

Robert Edgar

Independent scientist
robert@drive5.com
www.drive5.com

Let's dive into the hornets' nest

"Bacterial taxonomy is a hornets' nest that no one, really, wants to get into."

Referee #1, UTAX paper

Authoritative classifications

- Assume prokaryotic "species" meaningful
- Starting point for automated classification
  - Database of sequences + taxonomy annotations
- Bacteria & Archaea
  - ~10k sequenced isolated strains

Authoritative classifications

- Classified prokaryotes
  - ~12k named species
  - ~2,300 genera
  - Tiny fraction of total
- RDP Classifier training set v14 (RDP14)
  - 10k full-length 16S sequences
  - classified to genus but not species
  - ~2k genera
  - Best approximation I know of for authoritative db
  - Named isolate set with species names, no longer supported?
  - No 16S database documents "gold standard" subset AFAIK

Large databases

- SSU sequences + taxonomy annotations
- Greengenes
  - 1.3M 16S sequences
  - Obsolete? Last updated May 2013, secondgenome.com
- SILVA
  - 1.8M 16S sequences
  - ~100k genera
  - 98% not named
  - Small fraction of extant species / strains (billions?)

Length & phylogenetic "resolution"

- Full-length sequences can identify species
  - If ~100% identical to known sequence
- 97% "rule" not reliable
  - Paralogs in one species can be as low as 89%
  - Different species can be >97%
- Short tags (V4) cannot resolve species
  - Different species may have identical V4 sequences
  - Genus resolution good, but not perfect
  - 10% of genera in RDP14 have same V4 as another genus
  - Even if only one 100% id hit, could be novel species

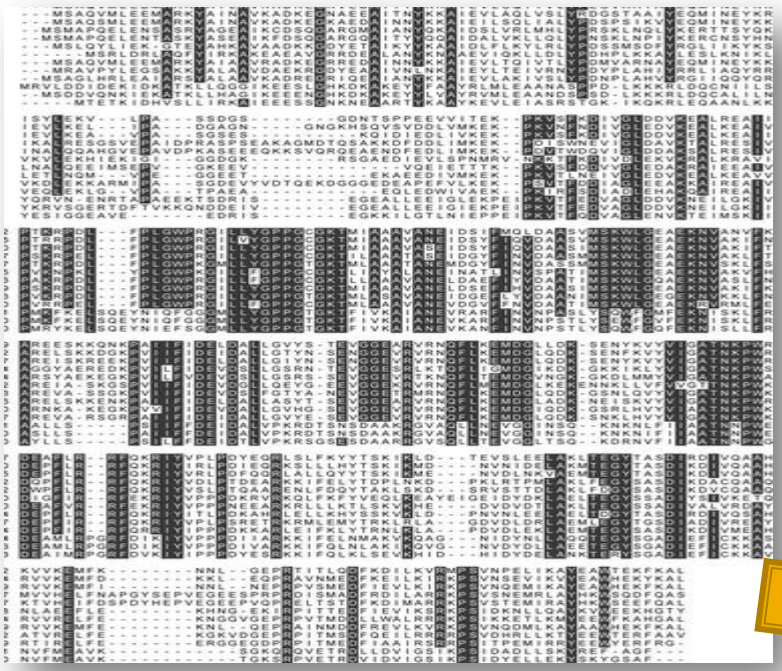
RDP, SILVA & GG taxonomies

- Different nomenclatures
  - RDP: Based on Bergey's
  - GG: Based on NCBI
  - SILVA: Based on LSPN
- Conflicts between sequence & taxonomy
- Example: *Escherichia* and *Shigella*
  - Sequence shows that these genera not monophyletic
  - GG: leaves genus & species blank
  - SILVA: new genus *Escherichia-Shigella*
  - RDP: new genus *Escherichia/Shigella*

SILVA & GG "taxonomies"

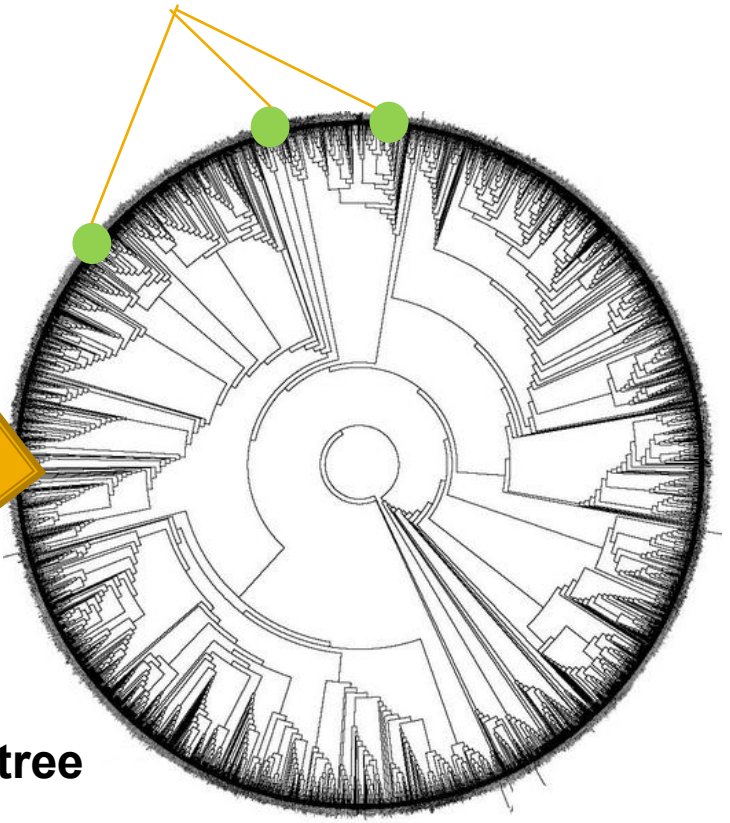
- Large majority are environmental
- Known only from sequence
- Taxonomy annotations are **predictions**!
- Manual + automated methods
- Error rates…?

GG & SILVA predictions



huge multiple alignment…

Named isolates



…huge tree

GG & SILVA predictions

- Perfect alignment impossible
  - Very hard to align across many phyla
  - May not be possible / meaningful in hypervariable regions
  - Especially GG
  - NAST *designed* to introduce mis-alignents!
- Perfect tree prediction impossible
- Must be errors
  - Plausibly could be many

Taxonomy annotation errors

"Mathematics is the art of giving the same name to different things"

Henri Poincaré

"Taxonomy should not give the same name to different things"

Robert Edgar

Taxonomy annotation errors

- Common name
  - Identical name found in all systems (GG, SILVA & RDP)
  - Most names are common
- Pair of databases
- Choose a rank, e.g. genus
- **Identical** sequences with **common** names
- If disagree, one annotation is **wrong**

GG & SILVA errors

GG-QIIME vs. SILVA-mothur

Rank	Common Names	Same Name	Different Name
Phylum	29098	28616 (98.3%)	481 (1.7%)
Class	24476	21592 (88.2%)	1201 (4.9%)
Order	21919	17121 (78.1%)	2804 (12.8%)
Family	15805	13141 (83.1%)	1428 (9.0%)
Genus	7735	5352 (69.2%)	1868 (24.1%)

Combined error rate:

24% genus
9% family
2% phylum

Disagreement implies error in one or both dbs.

Probably just one

Resolve by comparing with RDP14

GG-QIIME and RDP14

Rank	Common Names	Same Name	Different Name
Phylum	477	475 (99.6%)	2 (0.4%)
Class	1761	1678 (95.3%)	27 (1.5%)
Order	1786	1583 (88.6%)	79 (4.4%)
Family	1545	1423 (92.1%)	78 (5.0%)
Genus	1404	1253 (89.2%)	151 (10.8%)

SILVA-mothur and RDP14

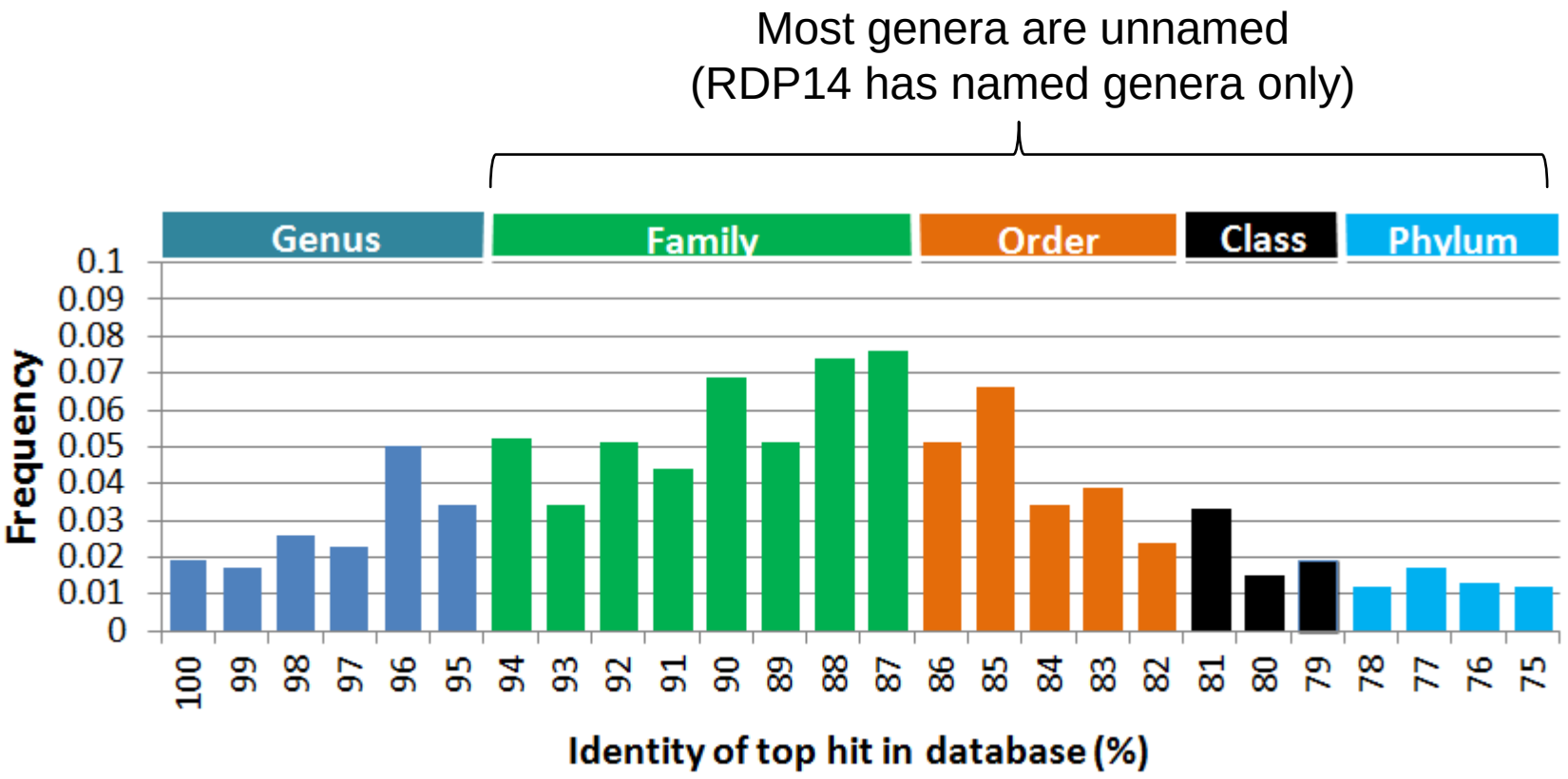
Rank	Common Names	Same Name	Different Name
Phylum	1030	1028 (99.8%)	2 (0.2%)
Class	4324	4299 (99.4%)	17 (0.4%)
Order	3359	3148 (93.7%)	57 (1.7%)
Family	4291	4070 (94.8%)	141 (3.3%)
Genus	4510	4386 (97.3%)	124 (2.7%)

GG ~3 – 4x more dis-agreements with RDP

Implies GG error rate at least ~3x > SILVA

SILVA ~6% genus errors
GG ~**18%** genus errors!

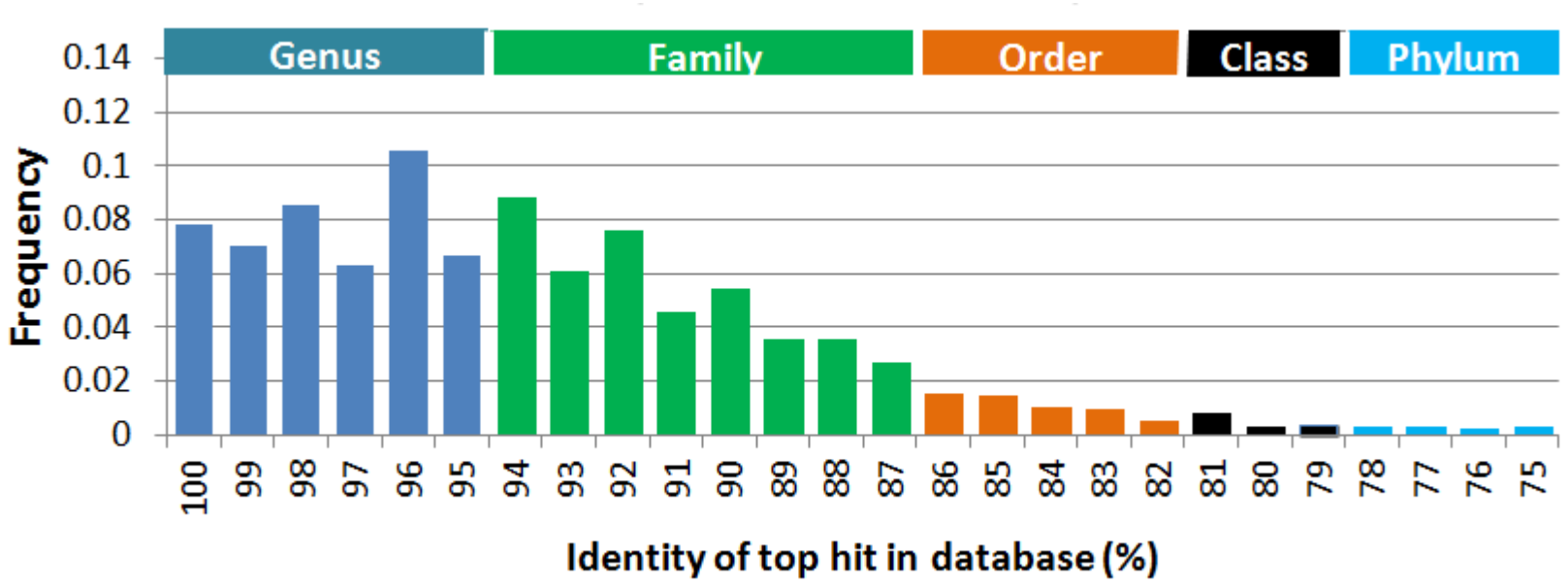
Soil OTUs vs. RDP14



V4 region unless otherwise stated

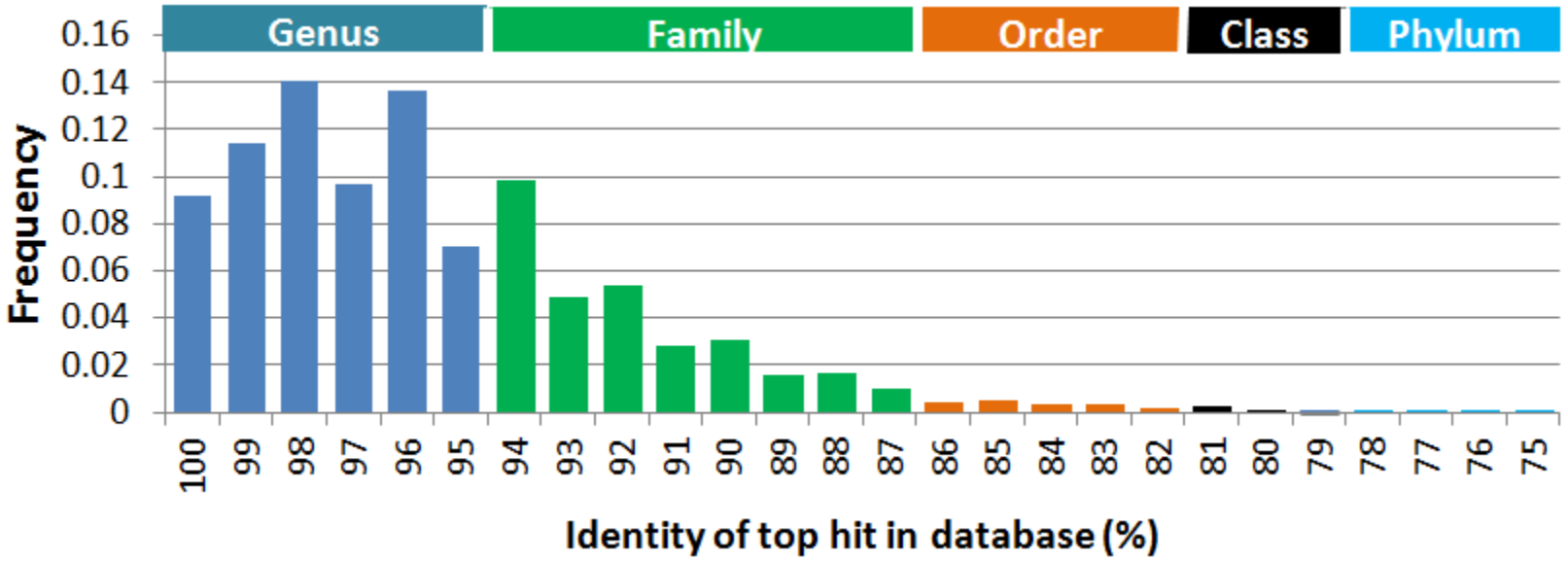
Soil OTUs vs. SILVA-mothur

Better coverage than you might expect for 20x bigger db
but still many genera not in db
~6% of named genera wrong

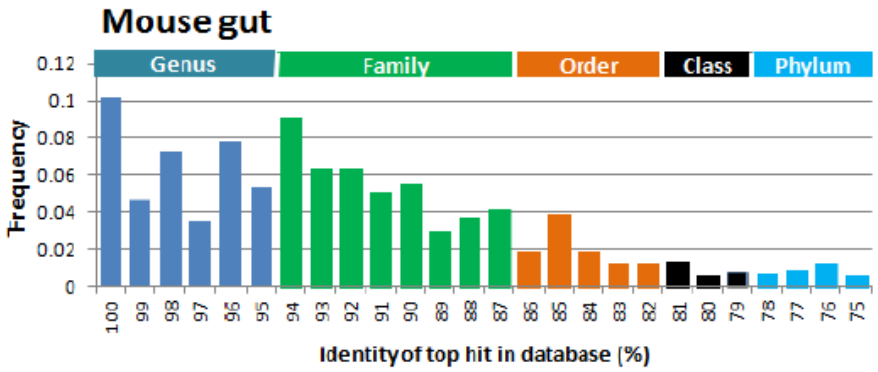


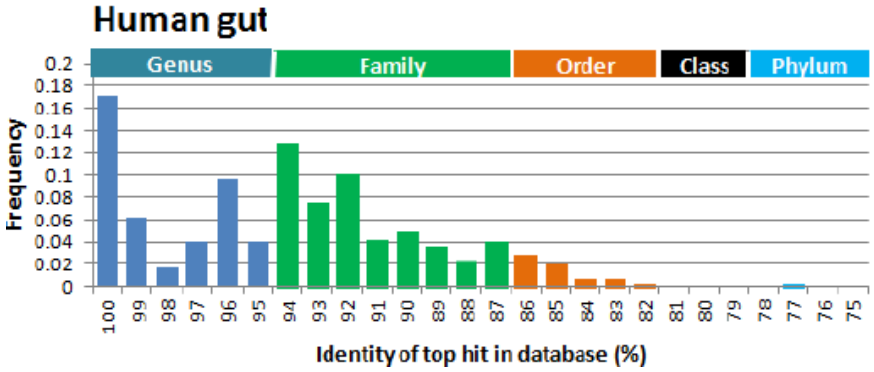
Soil OTUs vs. GG-QIIME

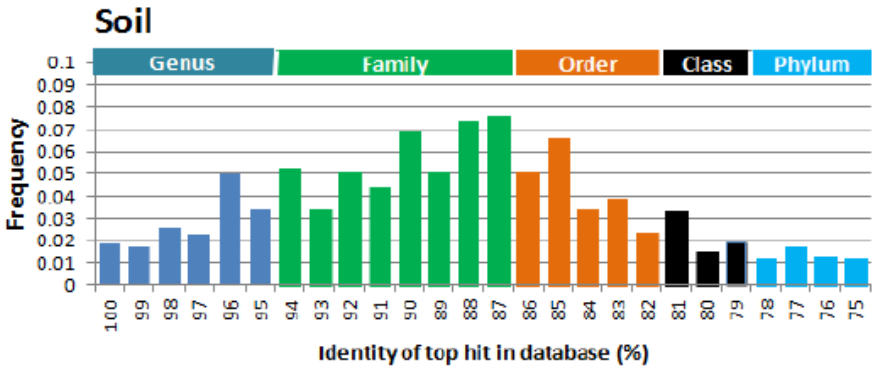
Many genera are novel
Better coverage than RDP or SILVA-mothur
but **18%** named genera wrong!



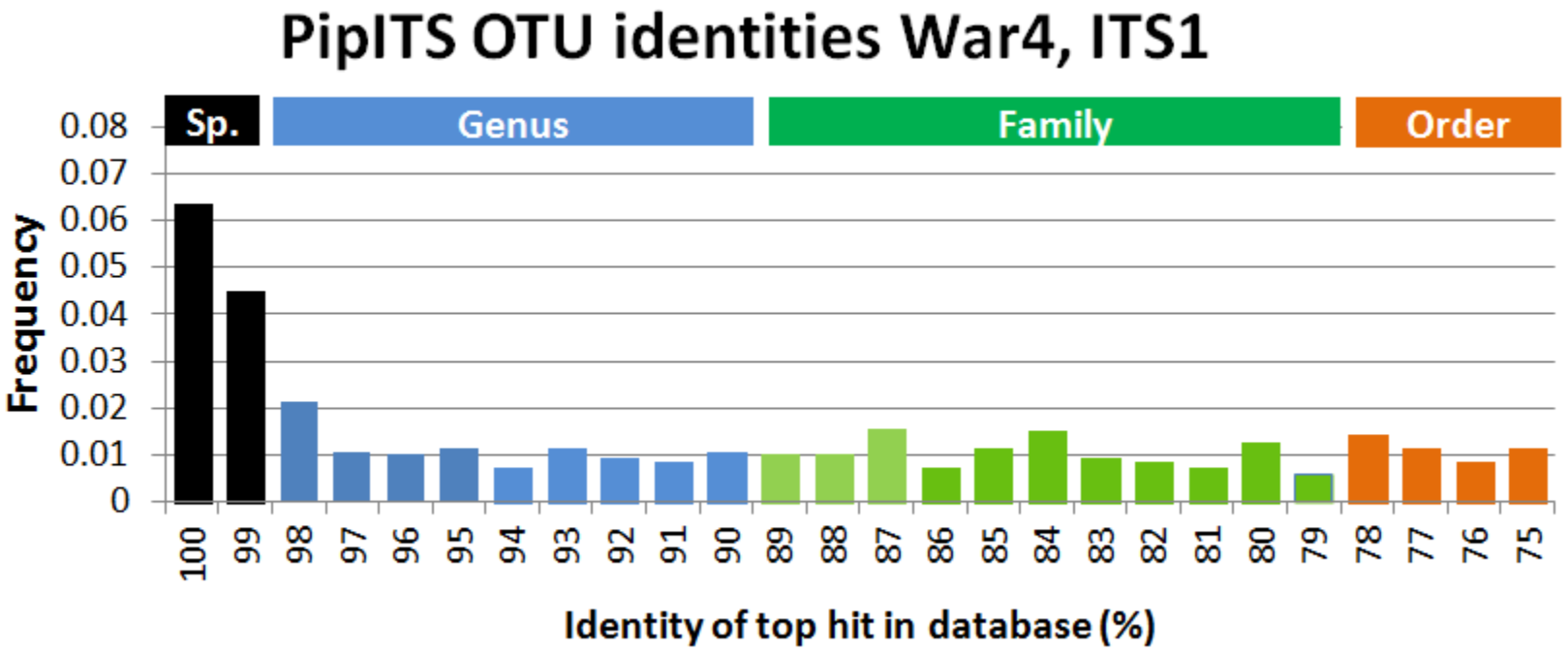
RDP14 vs. soil and gut OTUs







Similar story with fungal ITS



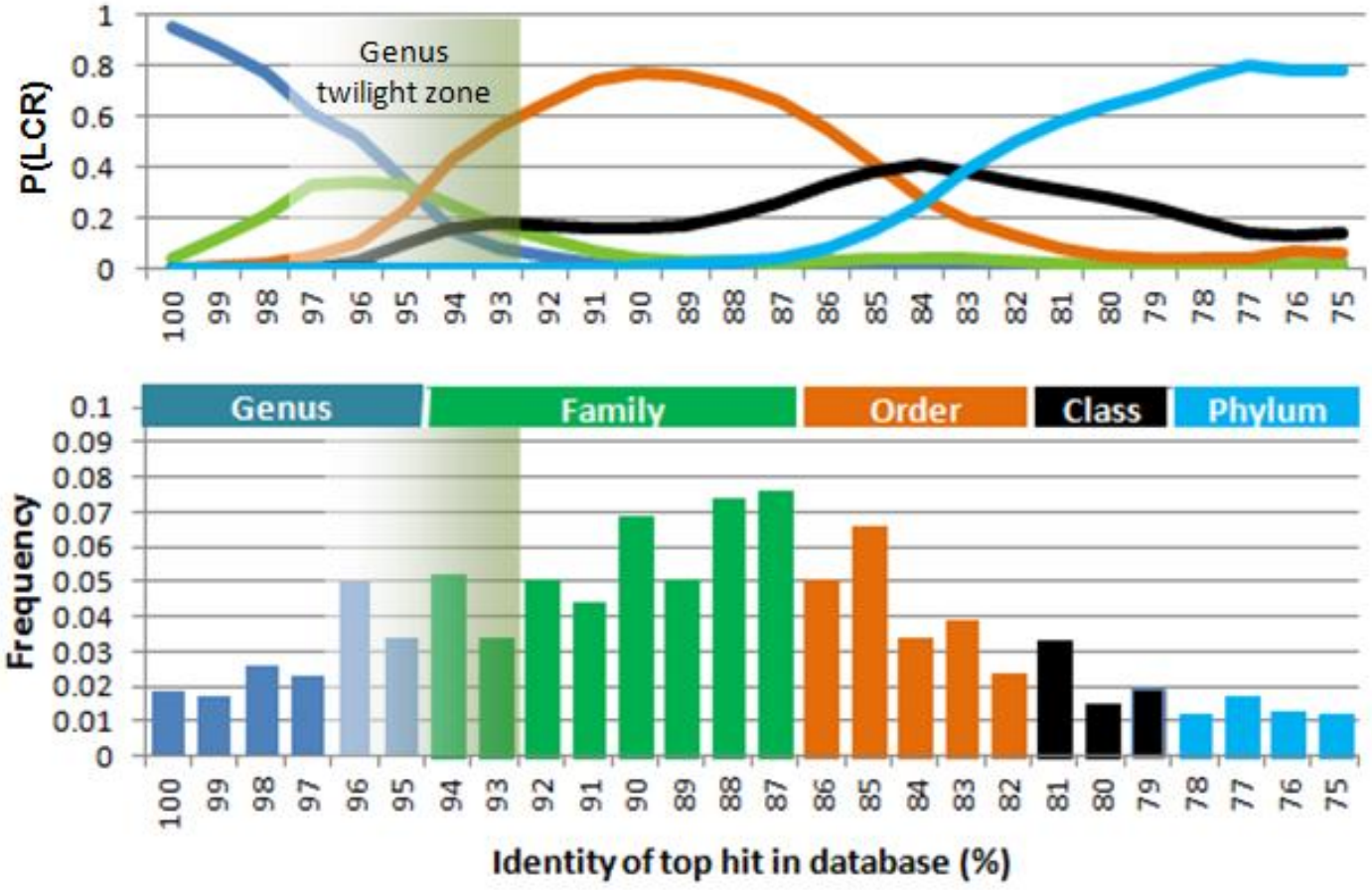
Lowest Common Rank

- Reference data is sparse
- Top database hit has 90% id
- Does it have same genus, family…?
  - What is Lowest Common Rank (LCR)
- **Easy** to find top hit(s)
  - All algorithms find the top hit(s), more or less
- **Hard** to predict LCR
  - This is the real challenge for taxonomy prediction

Twilight zone

- Half of genera have only one sequence
- Impossible to find genus-specific features
- Top hit 95% identity
  - Same genus?
  - Hard / impossible to predict
  - Must choose between FPs and FNs
  - Algorithm should indicate confidence
- ~95% is genus "twilight zone"
  - Similar to 20% a.a. identity for protein homology

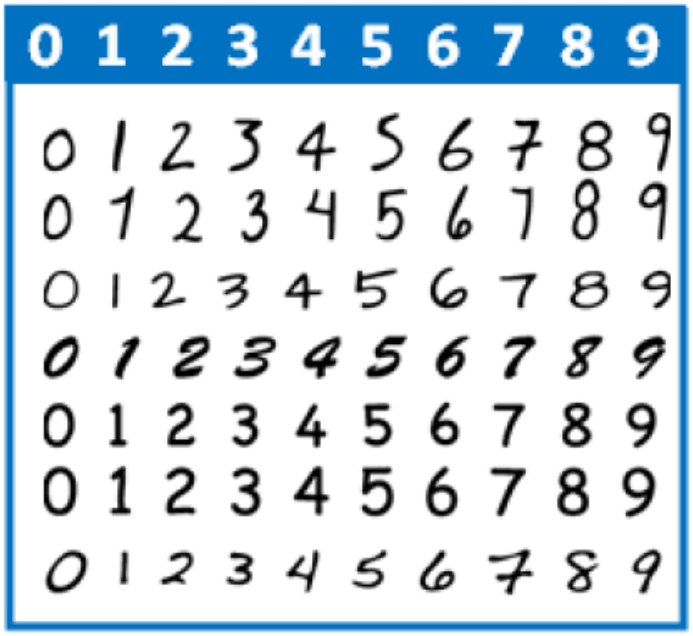
Twilight zone

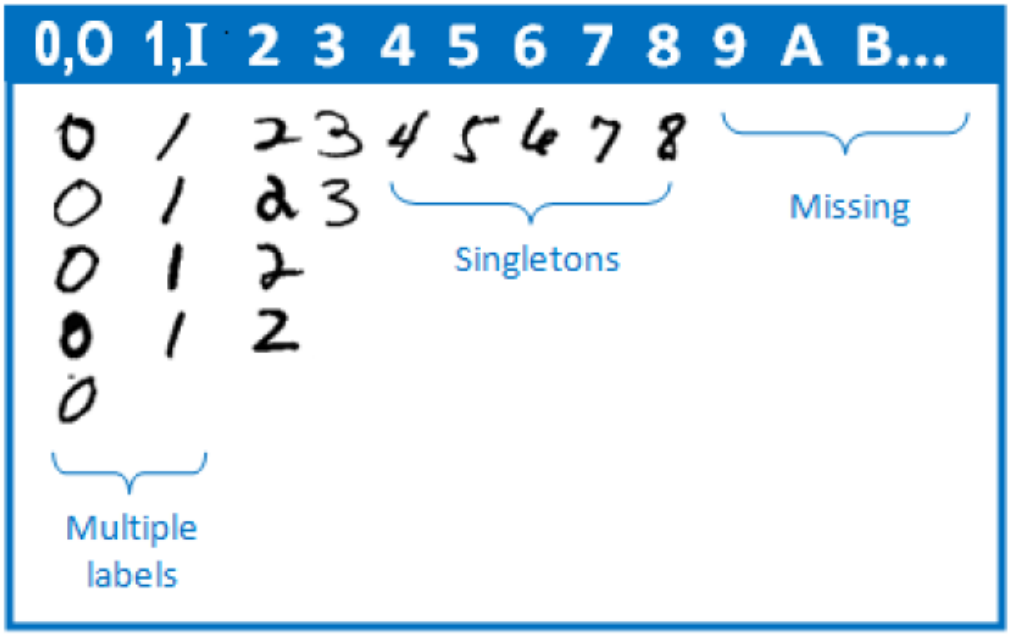


Prediction algorithms

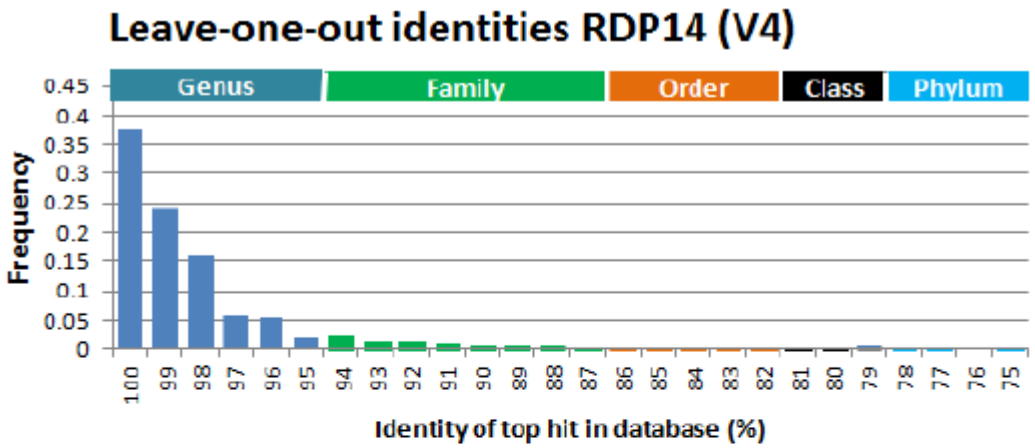
Method	Confidence	Published / Documented?	Description
RDP Classifier	Bootstrap	Yes	Naive "Bayesian" 8-mers
UTAX	*P*-value (EPQ)	Yes	8-mer distances to top hit & nearest neighbors at each rank
GAST	No	Yes	*Ad-hoc* top hit consensus
mothur-knn	No	No	?
QIIME-uclust	No	No	?
QIIME-blast	No	No	? (blast does not predict taxonomy)
QIIME-sortmerna	No	No	?
QIIME-RDP	Yes	Yes	RDP at 50% bootstrap (!)

Taxonomy is not a textbook case

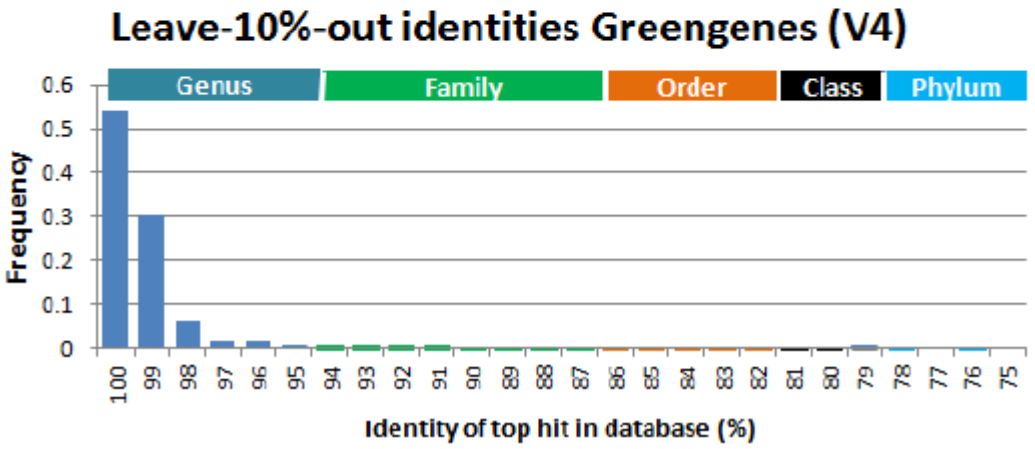




Leave-*x*-out validation



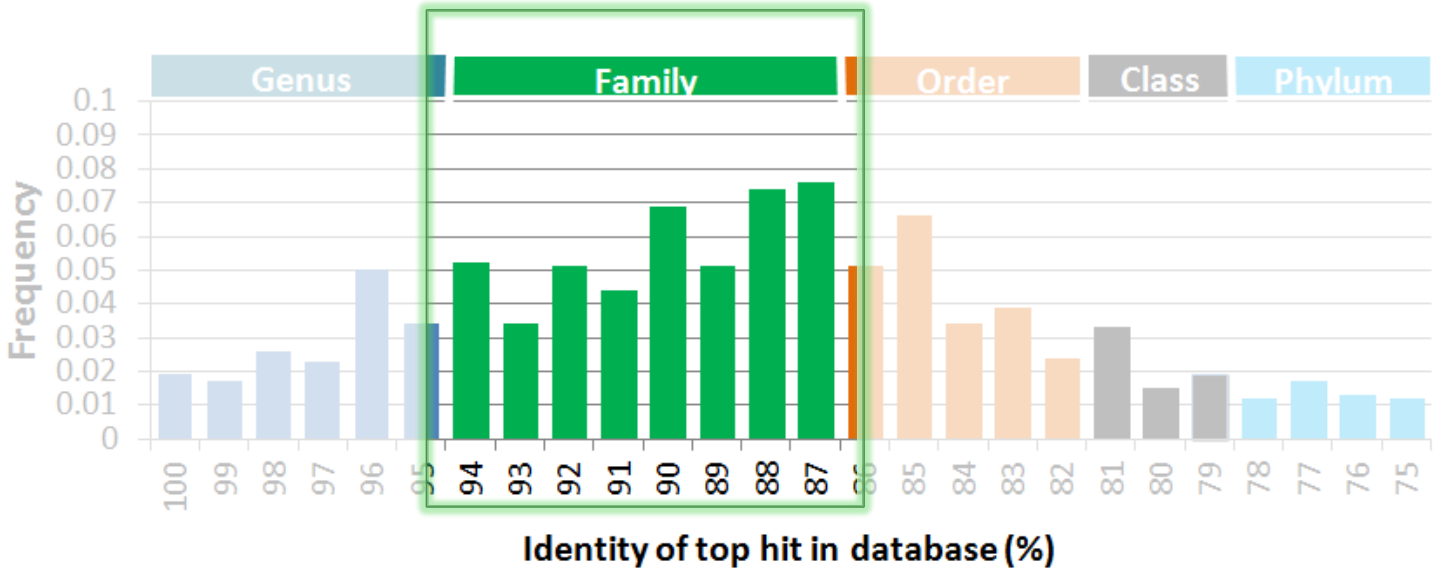
Avg. 4 seqs / genus
Most ≥98% id
Max accuracy < 100% due to singletons



Leave-10%-out (Bokulich *et al.* PeerJ)
Most ≥99% id

Benchmark test

- Test with e.g. LCR = family
- Models OTUs with ~94% to 87% id with top hit

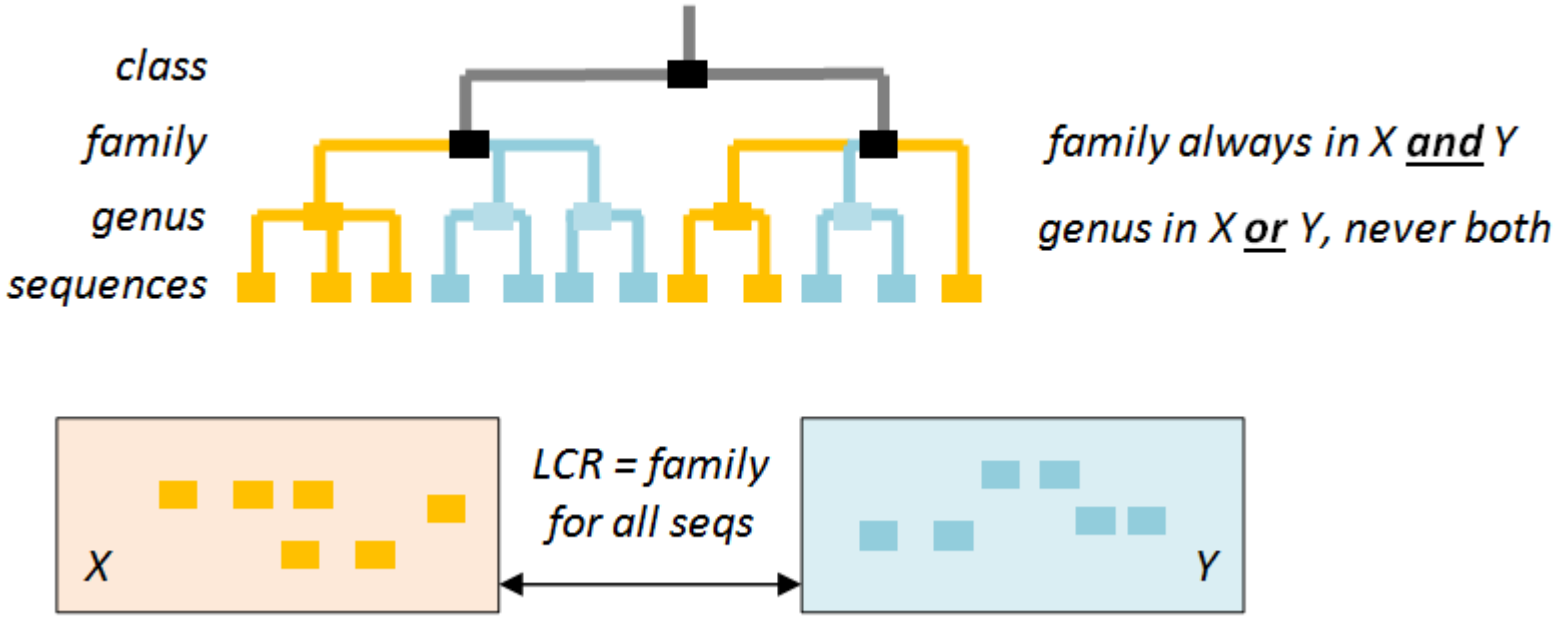


Benchmark test: "Rank split"

- Split trusted db (RDP14) into X_{rank} and Y_{rank}
- Example: LCR=family, make X_{family} and Y_{family}
- For each family, genus in X or Y (never both)
  - genus is NEVER known
  - family is ALWAYS known

Rank split construction

Method for making query - database pairs
with known LCRs from trusted ref db.

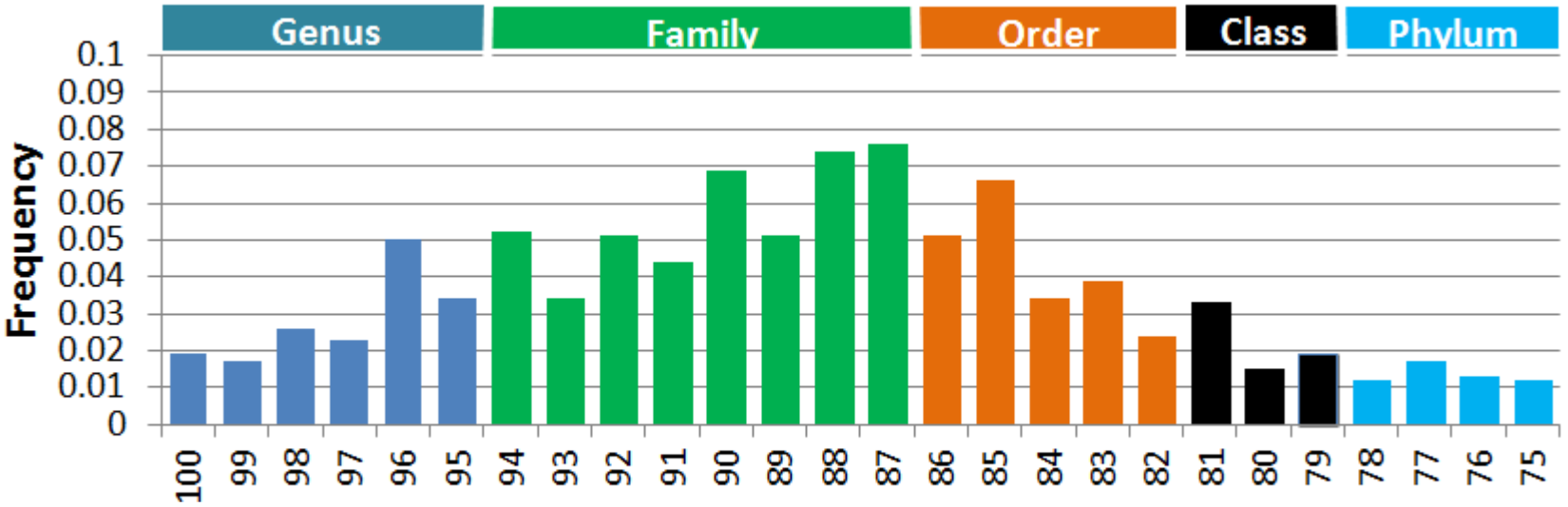


Benchmark test

- On each rank split, e.g. family
- Measure sensitivity to family & above
  - Fraction families correctly predicted (all are known)
- Mis-classifications (FPs):
  - Known but wrong, e.g. predict wrong family
- Over-classifications (FPs):
  - Novel but classified, e.g. predicted a genus name

LCR frequencies

Genus=20% Family=45% Order=20% Cl=8% Ph=7%



Estimated fraction of OTUs with LCR at each rank
"Novelty profile" of the OTUs w.r.t. reference database

LCR frequencies

- If 100% have LCR=genus
  - Leave-one-out is a good test
- If 100% have LCR=family
  - Then test with query= X_{family} and db.=Y_{family}
- Realistic test
  - Mixture of all LCRs, weighted by LCR frequencies

Results

	Soil				Mouse gut				Human gut			
	Genus		Phylum		Genus		Phylum		Genus		Phylum	
	Sens	EPQ	Sens	EPQ	Sens	EPQ	Sens	EPQ	Sens	EPQ	Sens	EPQ
UTAX (0.9)	70.9	10.5	95.7	1.2	75.1	9.6	97.2	0.8	78.9	10.3	98.5	0.3
RDP (80)	80.5	17.4	92.7	0.0	82.2	16.0	95.3	0.0	83.7	17.2	97.5	0.0
QIIME-rdp (50)	87.9	40.3	95.6	1.2	89.0	35.9	97.1	0.8	90.1	37.7	98.5	0.5
QIIME-uc	80.0	45.4	78.3	0.1	80.9	39.0	86.7	0.1	81.7	40.0	91.8	0.1
QIIME-blast	89.7	77.7	91.2	5.6	91.0	61.6	94.5	3.8	92.2	58.6	97.0	2.1
QIIME-sm	77.0	41.9	78.9	0.0	77.8	35.5	87.1	0.0	78.6	36.3	91.9	0.0
GAST	89.1	70.4	97.0	2.8	90.4	55.0	98.0	1.9	91.6	51.7	99.0	0.9
mothur-knn	34.9	4.4	91.7	0.6	35.8	3.3	93.7	0.4	36.6	3.2	95.5	0.2