Mock Communities

STAMPS 2016

Robert Edgar

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

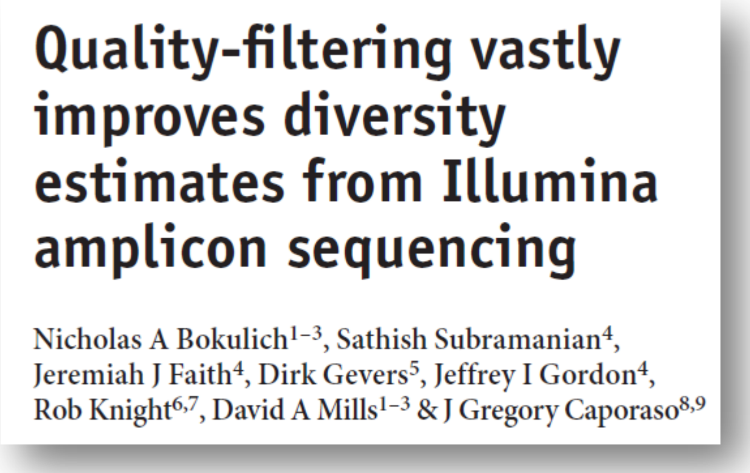
Mock communties

- Artificial sample, mix of known strains
- Typically 10 - 80 strains
    - HMP widely used for 16S, has 21
- *Even*: equal concentrations
    - cells, genome mass or 16S mass -- quite different!
- *Staggered*: range of abundances
- *Extreme*: species >97% identical
    - Validate denoising

Alice, mock turtle and griffyn

Mock case study #1

- MiSeq 2x250 V4
- Mock samples only
  - 2 Even, 2 Staggered



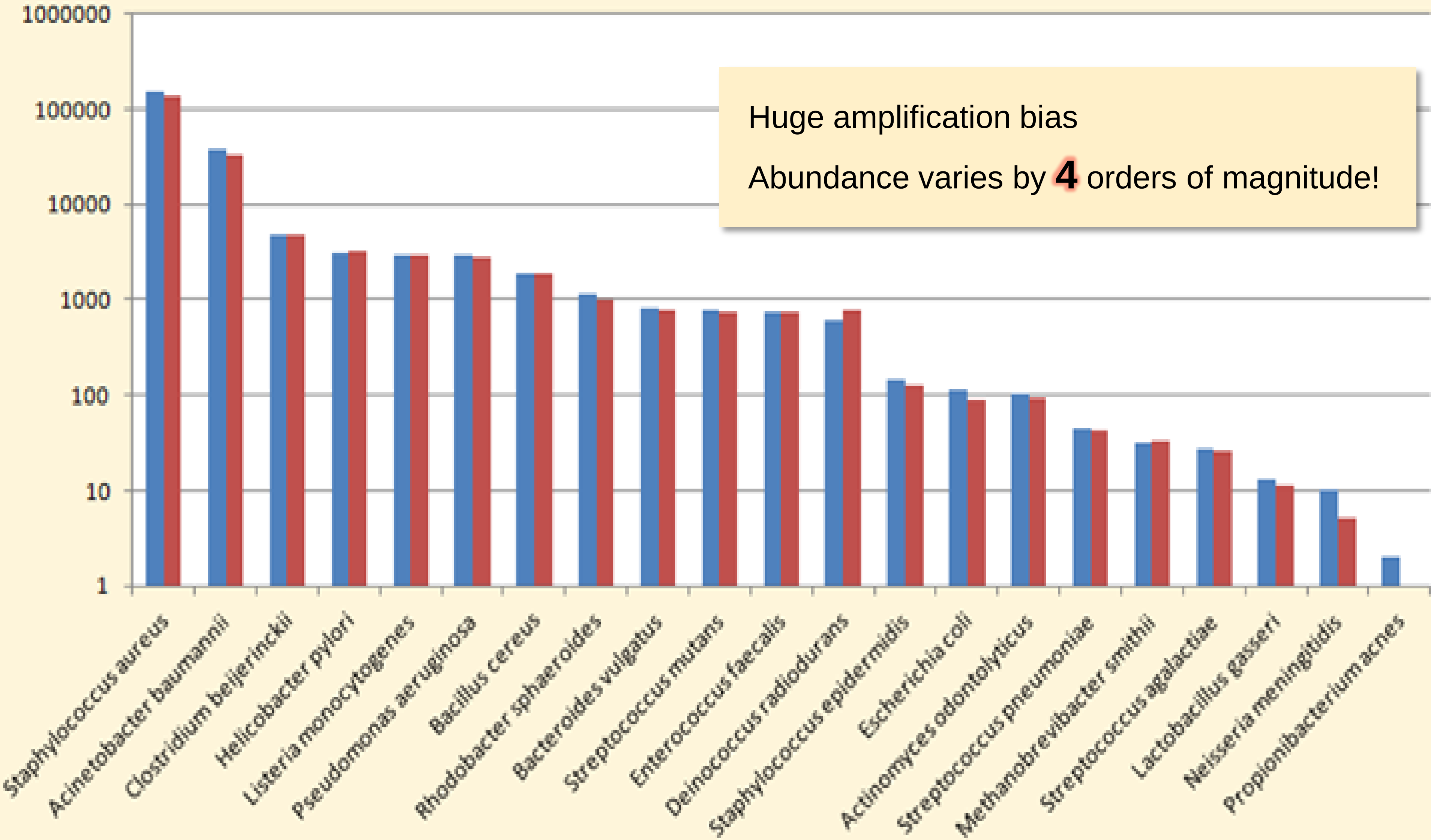
Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing

Nicholas A Bokulich[1–3], Sathish Subramanian[4], Jeremiah J Faith[4], Dirk Gevers[5], Jeffrey I Gordon[4], Rob Knight[6,7], David A Mills[1–3] & J Gregory Caporaso[8,9]

Contaminants

- At least 26 contaminant species
  - <97% to mock and not chimeric and 100% match to SILVA
  - Cluster at 97% to minimize double-counting
  - More contaminants than designed!
- Abundances **93, 73**, 6, 5, 5, … reads
  - Higher than some mock species
  - Mostly singletons

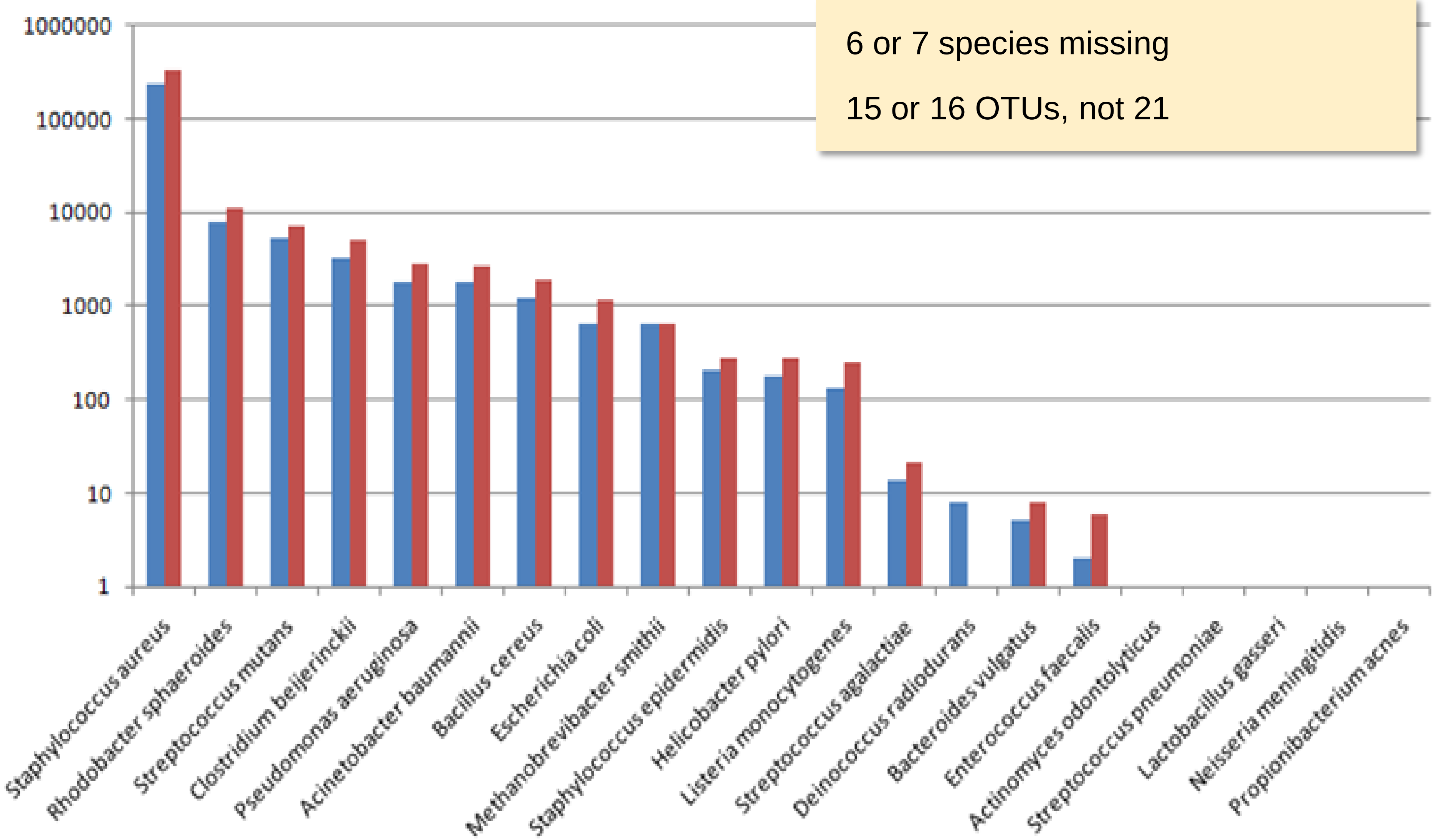
"Even" read abundances



Huge amplification bias

Abundance varies by **4** orders of magnitude!

"Staggered" read abundances

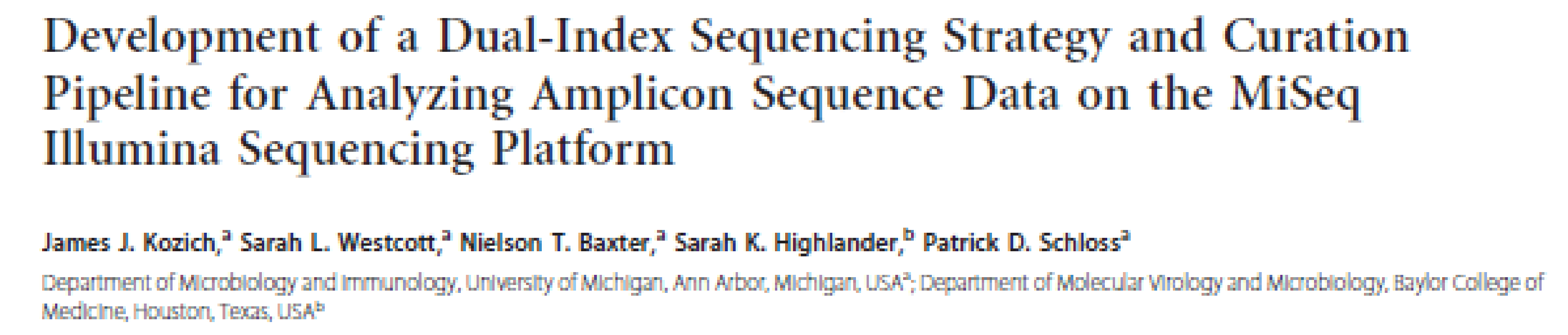


6 or 7 species missing

15 or 16 OTUs, not 21

Mock case study #2

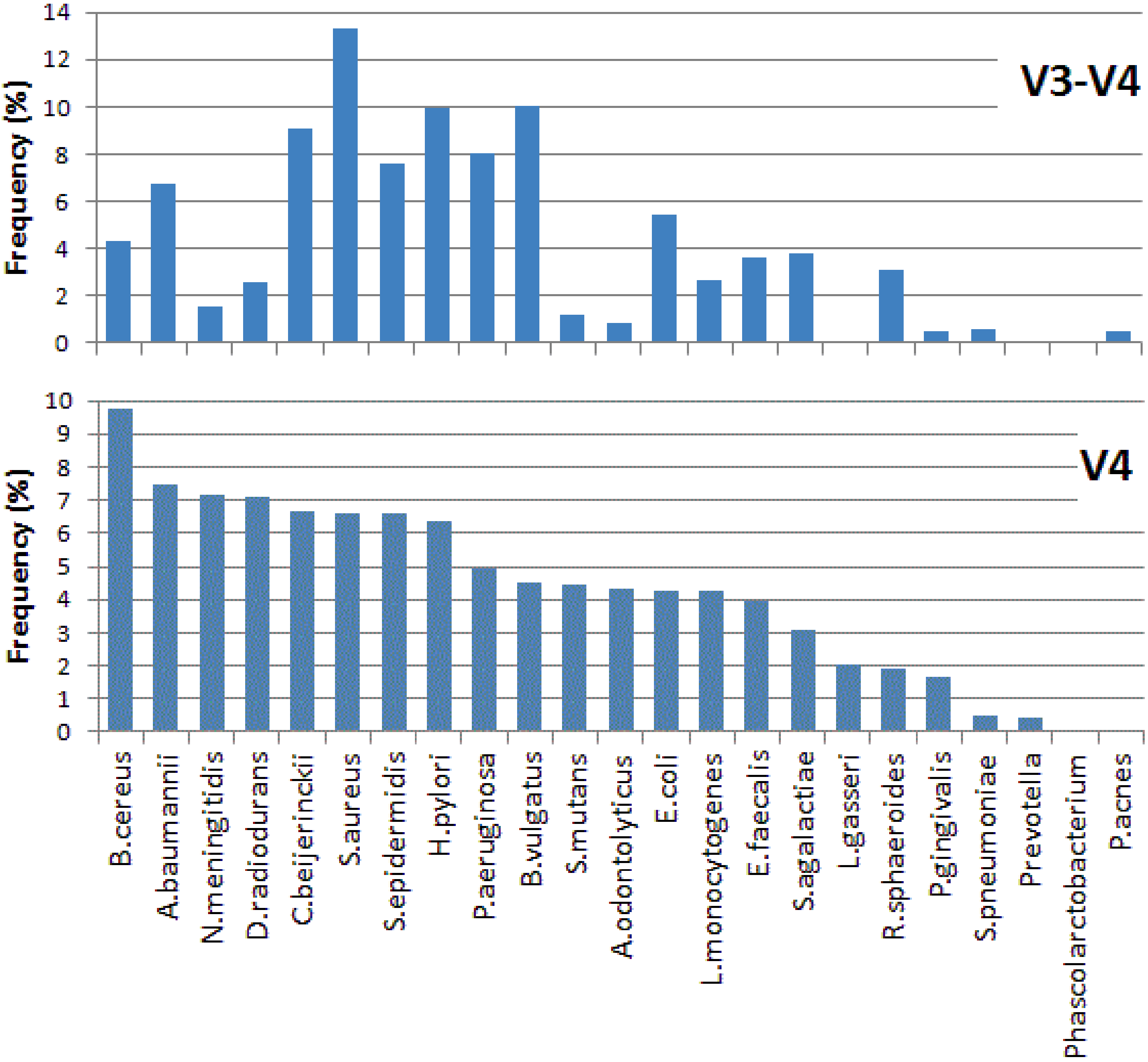
- MiSeq 2x250 V4
- Mock, Soil, Human gut, Mouse gut



Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform

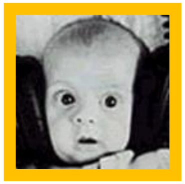
James J. Kozich,[a] Sarah L. Westcott,[a] Nielson T. Baxter,[a] Sarah K. Highlander,[b] Patrick D. Schloss[a]

Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan, USA[a]; Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, Texas, USA[b]


V3-V4
V4
Frequency (%)
Frequency (%)
0
2
4
6
8
10
12
14
0
1
2
3
4
5
6
7
8
9
10
B.cereus
A.baumannii
N.meningitidis
D.radiodurans
C.beijerinckii
S.aureus
S.epidermidis
H.pylori
P.aeruginosa
B.vulgatus
S.mutans
A.odontolyticus
E.coli
L.monocytogenes
E.faecalis
S.agalactiae
L.gasseri
R.sphaeroides
P.gingivalis
S.pneumoniae
Prevotella
Phascolarctobacterium
P.acnes

Re-analysis with UPARSE

- **2,350** OTUs in the mock samples
- Expected about 21… oops
- What the heck are they?
  - read errors, chimeras…??



Green -- mock L.monocytogenes.3
Orange -- other Faecalibacterium

OTU	Mock	Soil	Human	Mouse	%id	Refseq
OTU_6	730490(13.69%)	587(0.01%)	562(0.01%)	897(0.01%)	100%	S.aureus.2
OTU_8	522856(9.80%)	1668(0.02%)	388(0.01%)	3(0.00%)	100%	B.cereus.9
OTU_11	405378(7.60%)	345(0.01%)	307(0.00%)	37(0.00%)	100%	A.baumannii.4
OTU_13	388965(7.29%)	331(0.00%)	318(0.00%)	0(0.00%)	100%	B.vulgatus.5
OTU_14	385642(7.23%)	324(0.00%)	280(0.00%)	0(0.00%)	100%	D.radiodurans.3
OTU_15	360000(6.75%)	581(0.01%)	329(0.00%)	2(0.00%)	100%	C.beijerinckii.10
OTU_16	344151(6.45%)	220(0.00%)	278(0.00%)	1(0.00%)	100%	H.pylori.2
OTU_17	267306(5.01%)	638(0.01%)	217(0.00%)	10(0.00%)	100%	P.aeruginosa.3
OTU_12	242303(4.54%)	142(0.00%)	430082(6.00%)	2528(0.04%)	100%	B.vulgatus.5
OTU_19	240857(4.51%)	181(0.00%)	176(0.00%)	47(0.00%)	100%	S.mutans.4
OTU_20	232806(4.36%)	130(0.00%)	258(0.00%)	1(0.00%)	100%	A.odontolyticus.2
OTU_18	230458(4.32%)	7223(0.11%)	880(0.01%)	30944(0.44%)	100%	E.coli.5
OTU_21	229180(4.29%)	207(0.00%)	192(0.00%)	2(0.00%)	100%	L.monocytogenes.3
OTU_23	215401(4.04%)	173(0.00%)	171(0.00%)	18(0.00%)	100%	E.faecalis.2
OTU_26	167749(3.14%)	69(0.00%)	117(0.00%)	0(0.00%)	100%	S.agalactiae.4
OTU_30	108607(2.04%)	32(0.00%)	78(0.00%)	0(0.00%)	100%	L.gasseri.4
OTU_29	104428(1.96%)	247(0.00%)	69(0.00%)	0(0.00%)	100%	R.sphaeroides.4
OTU_31	89937(1.69%)	60(0.00%)	71(0.00%)	0(0.00%)	100%	P.gingivalis.3
OTU_66	27287(0.51%)	80(0.00%)	74(0.00%)	259(0.00%)	100%	S.pneumoniae.2
OTU_1	16093(0.30%)	127(0.00%)	3244970(45.26%)	13110(0.19%)	99%	AB064923|S000768314 Provotella
OTU_7752	6894(0.13%)	39(0.00%)	1431022(19.96%)	6920(0.10%)	98%	AB064923|S000768314 Provotella
OTU_10	2203(0.04%)	5(0.00%)	429297(5.99%)	1275(0.02%)	100%	X72865|S000013701 Phascolarctobacterium
OTU_1159	733(0.01%)	0(0.00%)	1(0.00%)	9(0.00%)	100%	P.acnes.3
OTU_8494	657(0.01%)	1(0.00%)	0(0.00%)	0(0.00%)	97%	S.mutans.1
OTU_38	499(0.01%)	0(0.00%)	86037(1.20%)	190(0.00%)	100%	AJ413954|S000128478 Faecalibacterium
OTU_36	404(0.01%)	0(0.00%)	71575(1.00%)	465(0.01%)	100%	AY126616|S000546342 Bacteroides
OTU_22	335(0.01%)	1000(0.01%)	71623(1.00%)	203334(2.89%)	100%	AB021164|S000008023 Bacteroides
OTU_34	317(0.01%)	0(0.00%)	63445(0.88%)	122(0.00%)	100%	AB238928|S000650592 Parabacteroides
OTU_33	309(0.01%)	243032(3.64%)	3(0.00%)	480(0.01%)	100%	GG4402730 Acidobacteria-6
OTU_39	287(0.01%)	3(0.00%)	54256(0.76%)	288(0.00%)	100%	GG116083 Rikenellaceae

MiSeq cross-talk

- Spurious OTUs in mock samples
- MiSeq index read errors
- ~0.5% of reads assigned to wrong sample
- QIIME Illumina filter: discard OTUs < 0.005%
  - Bokulich *et al.* (2013) *Nat Meth*

Richness is a poor metric

- Kozich *et al.* and Bokulich *et al.* did not analyze mock OTU *sequences*
- Only the *number* of OTUs
  - a.k.a. richness or alpha diversity

Richness is a poor metric

- "Correct" nr. OTUs can be > or < nr. strains
  - More: Contaminants and cross-talk
  - More: Paralogs <97% identical
  - Less: Species missing, e.g. primer mismatches
  - Less: Strains >97% identical
- Right number for wrong reason
  - **Plus** spurious OTUs due to cross-talk, chimeras, errors
  - **Minus** missing strains due to bias, mismatches
  - Tune parameters on mock, results may not generalize
- Should **identify** and **classify** **sequences**!

Mock reference sequences

- Sequence analysis requires mock ref. db.
  - All the mock 16S sequences, and nothing but
- No such database exists, as far as I know
  - HMP mock has a ref db in circulation
    - Not published or explained
  - Has all known sequences for the species(?)
    - not just the ATCC strains
- Missing resource
  - would be useful contribution, especially for HMP
  - might be possible using finished genomes?

Mock is an essential control

- Always include a mock sample as a control
- Make OTUs from reads for all samples
- Validate mock OTUs by aligning to ref. seqs.
  - Not taxonomy prediction (RDP, UTAX) -- low resolution
- Check for chimeras, contaminants, cross-talk
- USEARCH v9 (coming soon)
  - **annotate** command compares to mock ref & SILVA
  - reports good sequences and chimeras
  - hard to implement (fake models)