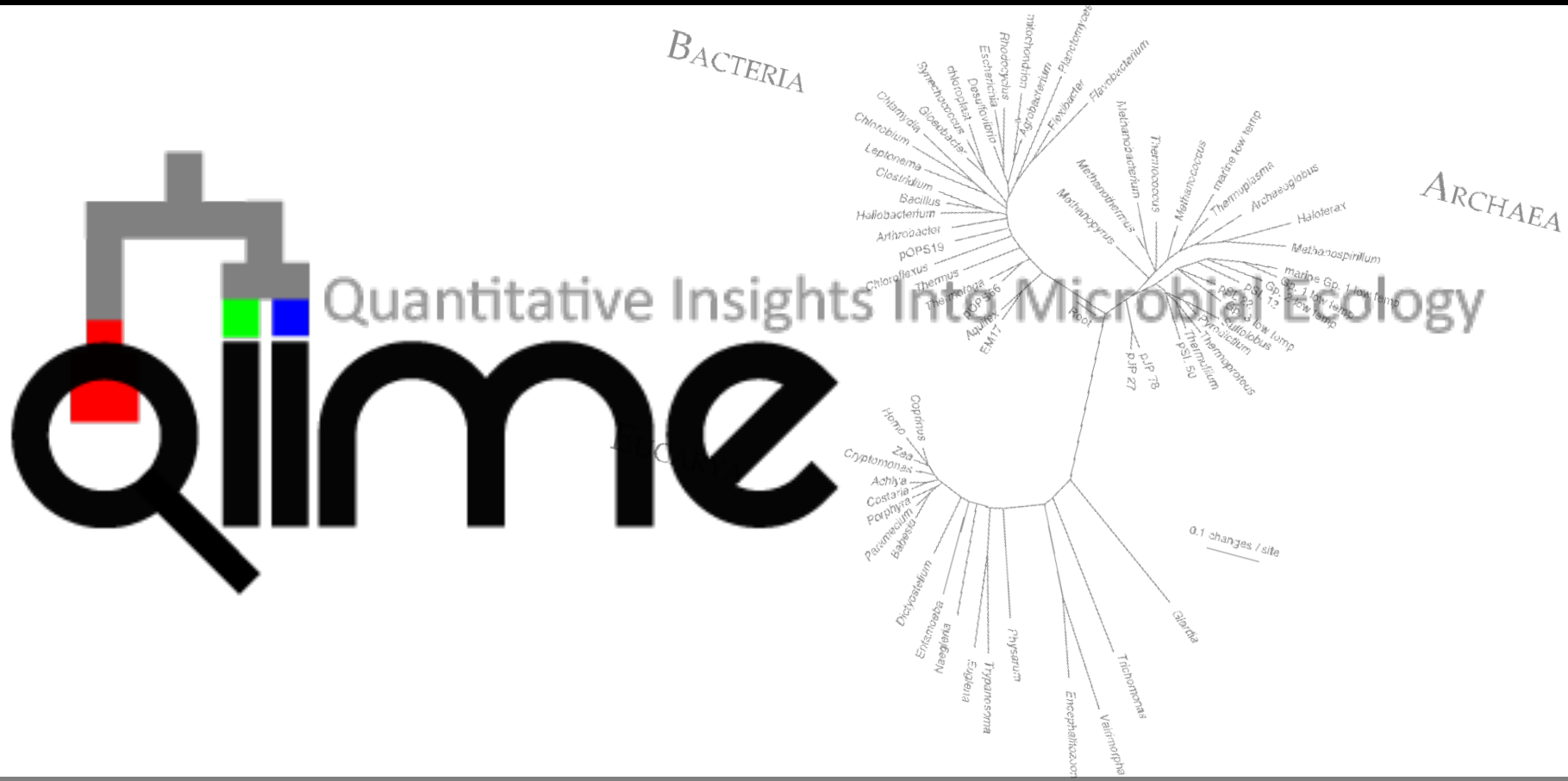
Microbial Communities Profiling via QIIME



KNIGHTLAB UCSD

Searching for significant OTUs

Which features (OTUs) of your data are most different between sample classes?

Corn

Soy

	Sample_1	Sample_2	Sample_3	Sample_4	Sample_5	Sample_6
OTU_1	100	150	1000	250	275	600
OTU_2	345	297	611	35	14	0

Searching for significant OTUs

SampleA

SampleB

SampleC

Searching for significant OTUs

SampleA
OTUX – ¼

SampleA
OTUY – ⅓

SampleB
OTUX – ½

SampleB
OTUY – ⅓

SampleC
OTUX – ¼

SampleC
OTUY – ⅓

Searching for significant OTUs

SampleA
OTUX – ¼

SampleA
OTUY – ⅓

SampleB
OTUX – ½

SampleB
OTUY – ⅓

SampleC
OTUX – ¼

SampleC
OTUY – ⅓

✓

✗

Parametric vs. non parametric

- Parametric: Assume that the data follows certain distribution, generally a normal distribution (ANOVA and t-test; Q-Q plot)

- However, there are other specific ones:
  - Kruskal-Wallis: The samples come from similar shape distribution (mean and variance might be different; Kolmogorov–Smirnov test)

What do we have in qiime for this?

- group_significance.py
  - Parametric
    - g-test (goodness-of-fit test)
    - ANOVA (one-way analysis of variance)
    - T-test
  - Non parametric
    - Kruskal-Wallis* (non-parametric ANOVA)
    - Mann-Whitney-U (non-parametric t-test)
    - Bootstrap Mann-Whitney-U
    - Bootstrap T-test

Give me the tests!

- G-test: the graphical example we saw. Originally developed for single value experiments
- ANOVA: test differences in means
- T-test: ANOVA for 2 groups
- Kruskal-Wallis: Non parametric ANOVA
- Mann-Whitney-U: Kruskal-Wallis for 2 groups
- Bootstrap: Randomizes labels and performs the given test n times and the p-value is = better_or_equal_test_statistic/ random_tests

Distance-based tests between groups

- Implementation of tests available in the R packages 'vegan' and 'ape' and Primer-E
- compare_categories.py
  - permutational
    - adonis
    - anosim
    - mrpp
    - permanova
    - permdisp
    - dbrda
  - bioenv
  - morans_i

compare_categories.py

compare_categories.py

-i <distance matrix file path>

-m <mapping file path>

-c <category in the mapping file>

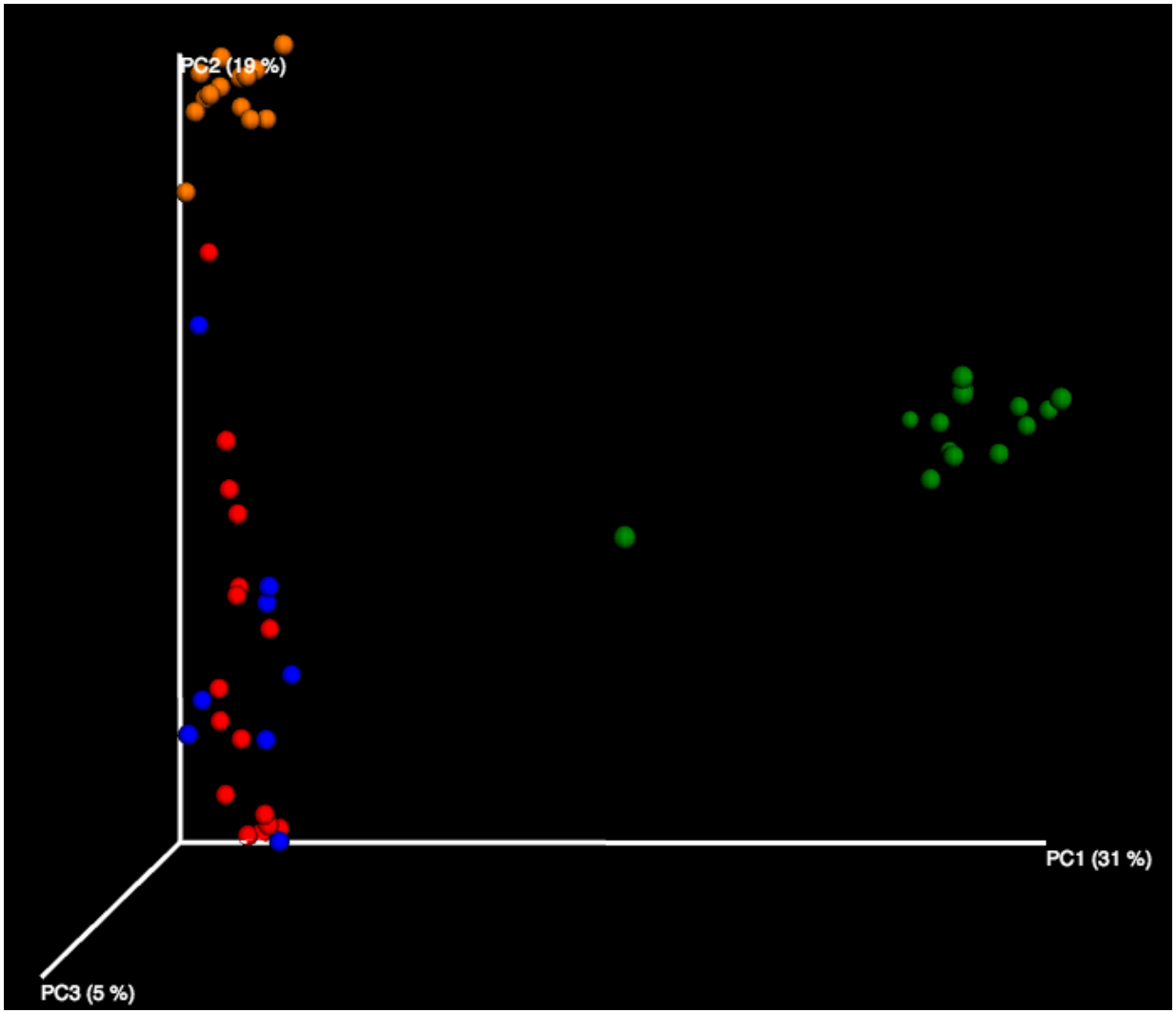
-o <output directory>

--method <statistical method>

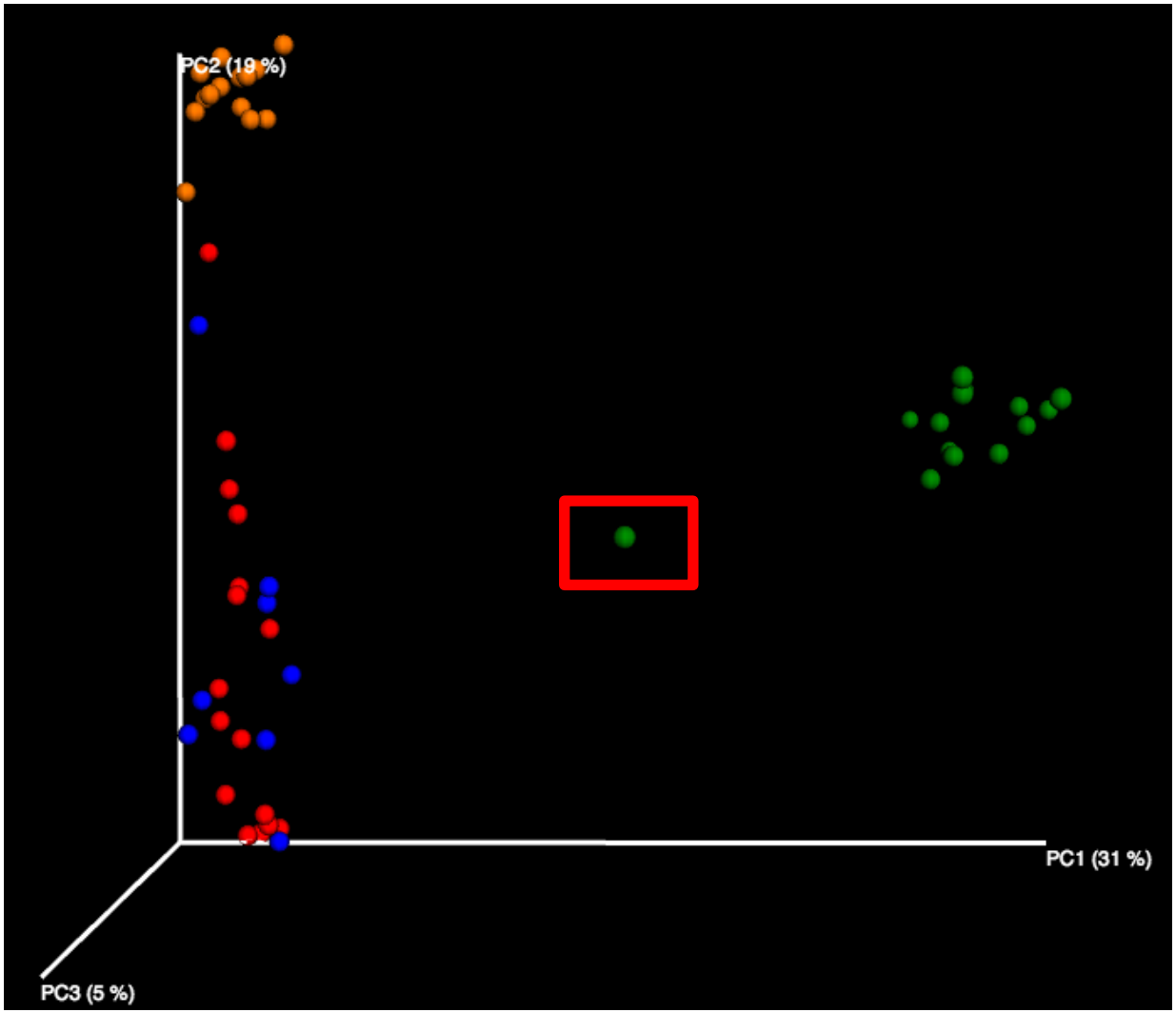
Optional:

-n <number of permutations>

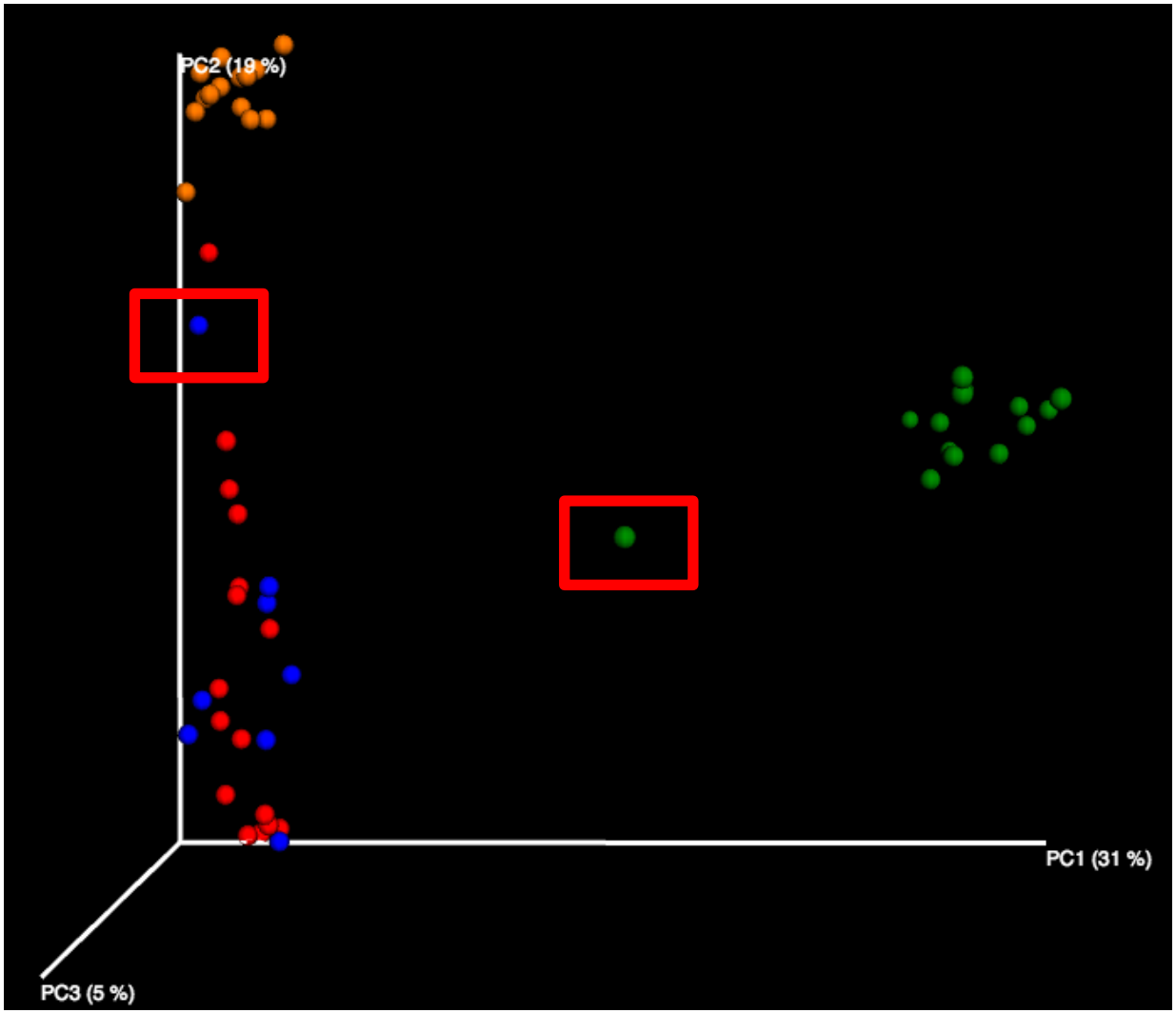
Should my sample really be there?



Should my sample really be there?



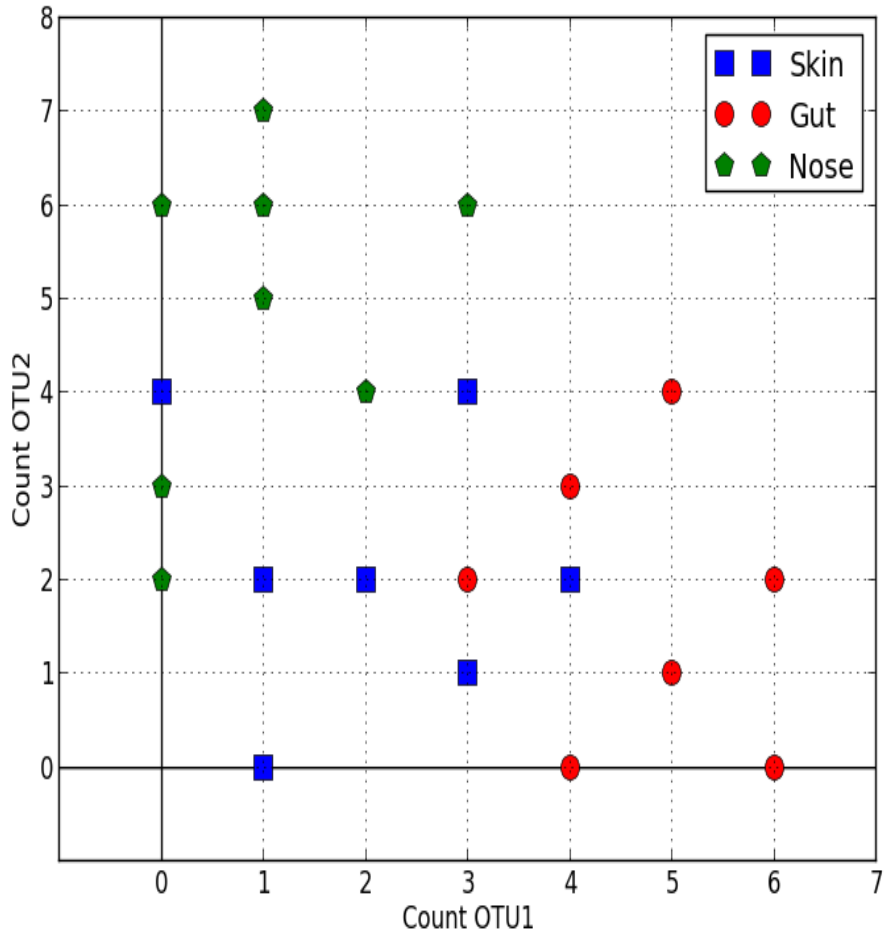
Should my sample really be there?



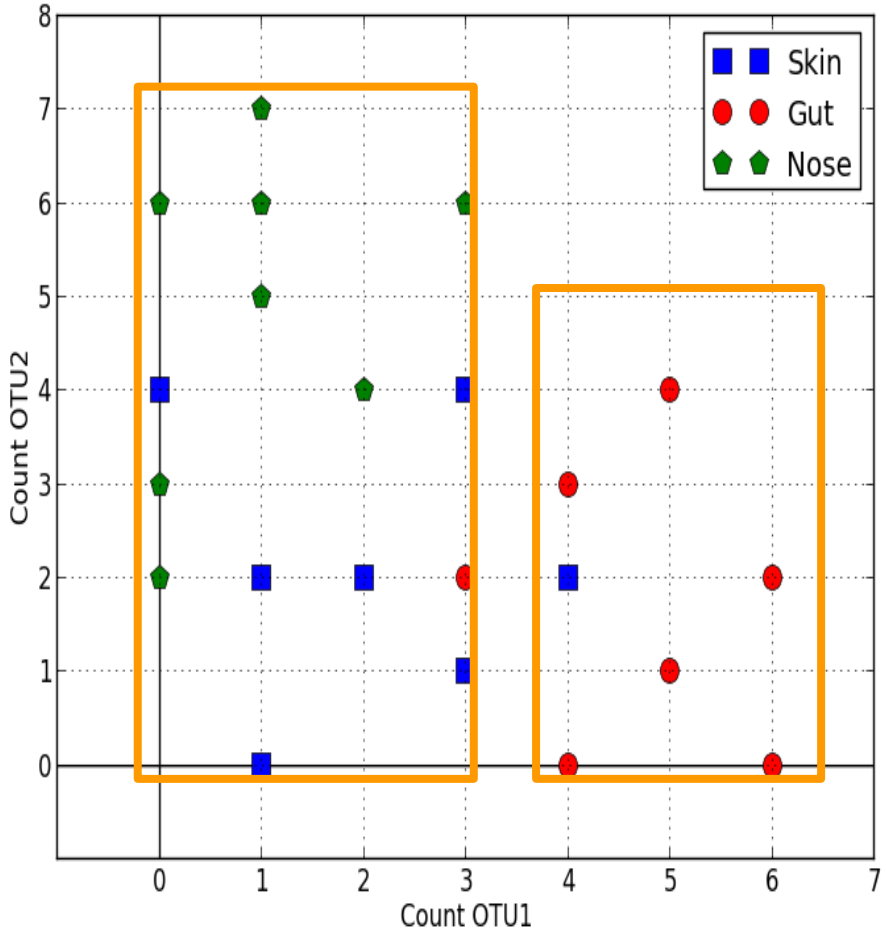
Supervised learning
random forests

- Can you tell your sample classes apart?
  - One of many machine learning strategies available
- What does it tell you?
  - Whether or not you samples are separated by a group of features
  - What the inter and intra-class variation looks like

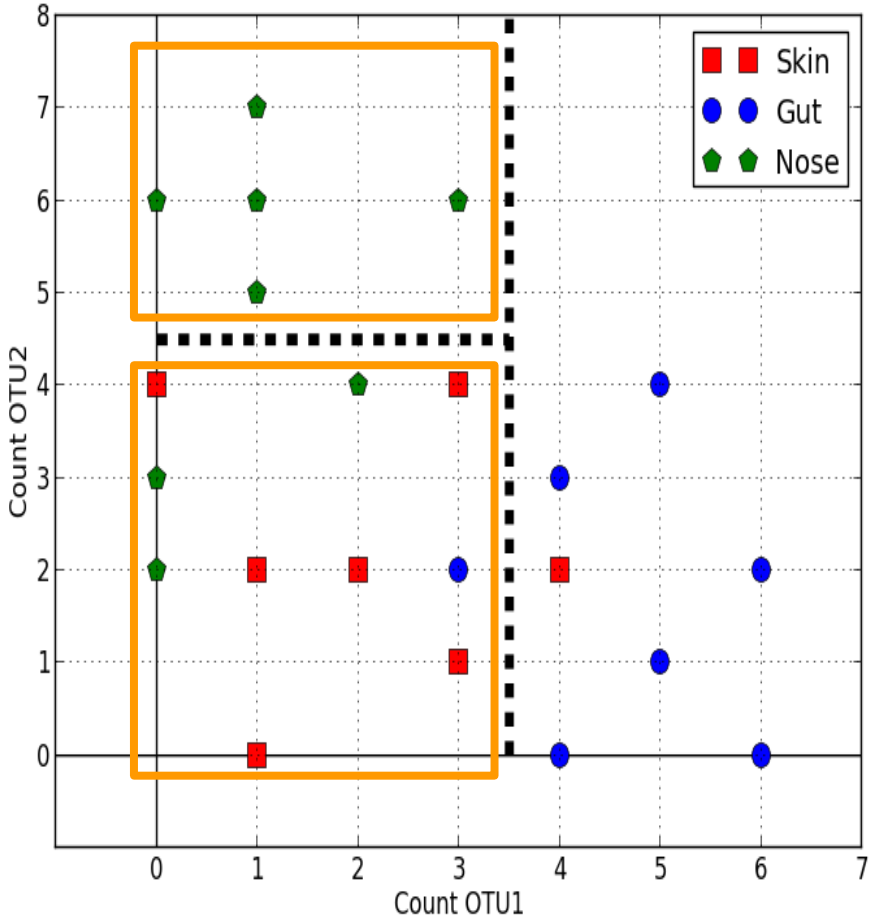
How does it work?

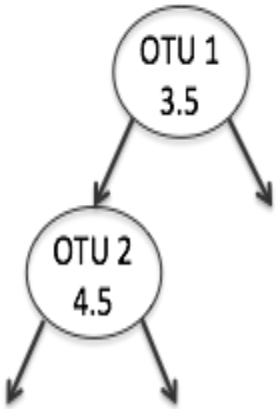


How does it work?

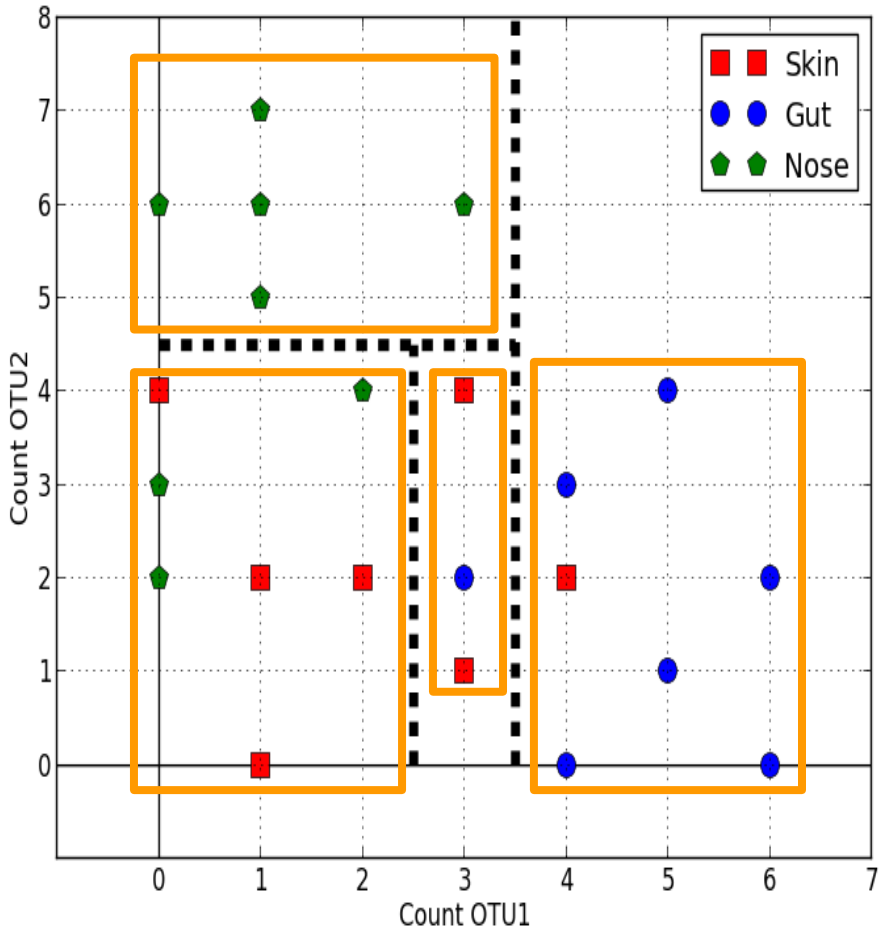


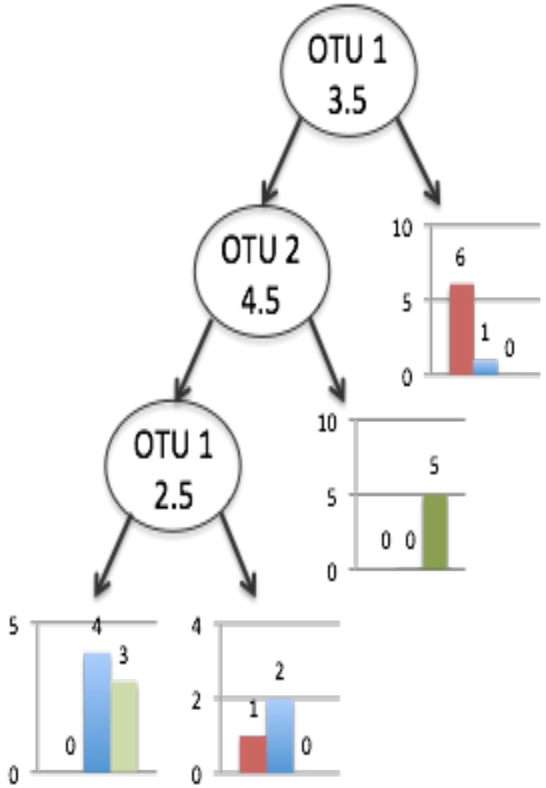
How does it work?



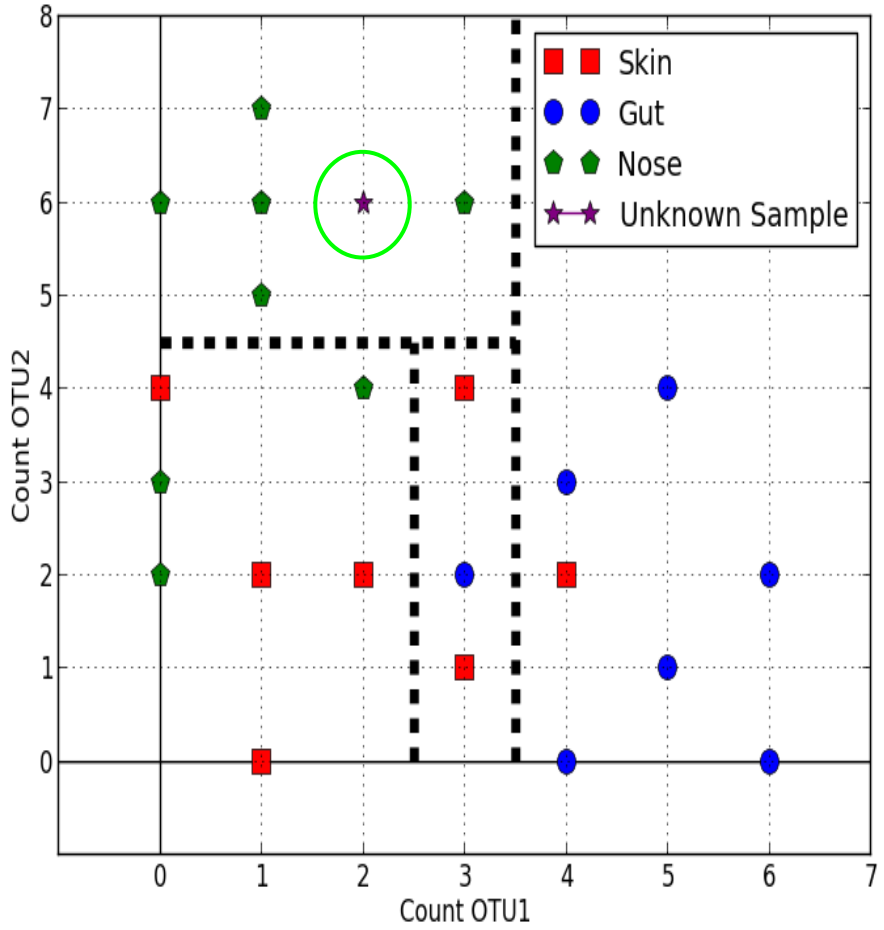


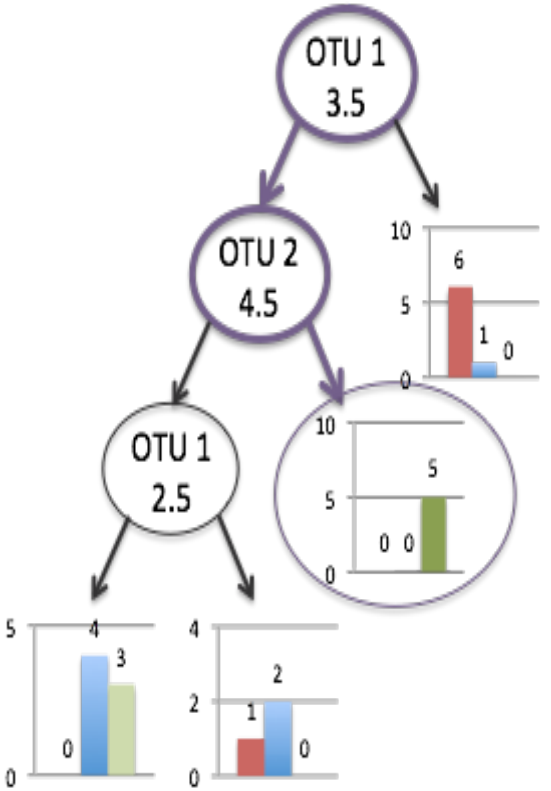
How does it work?



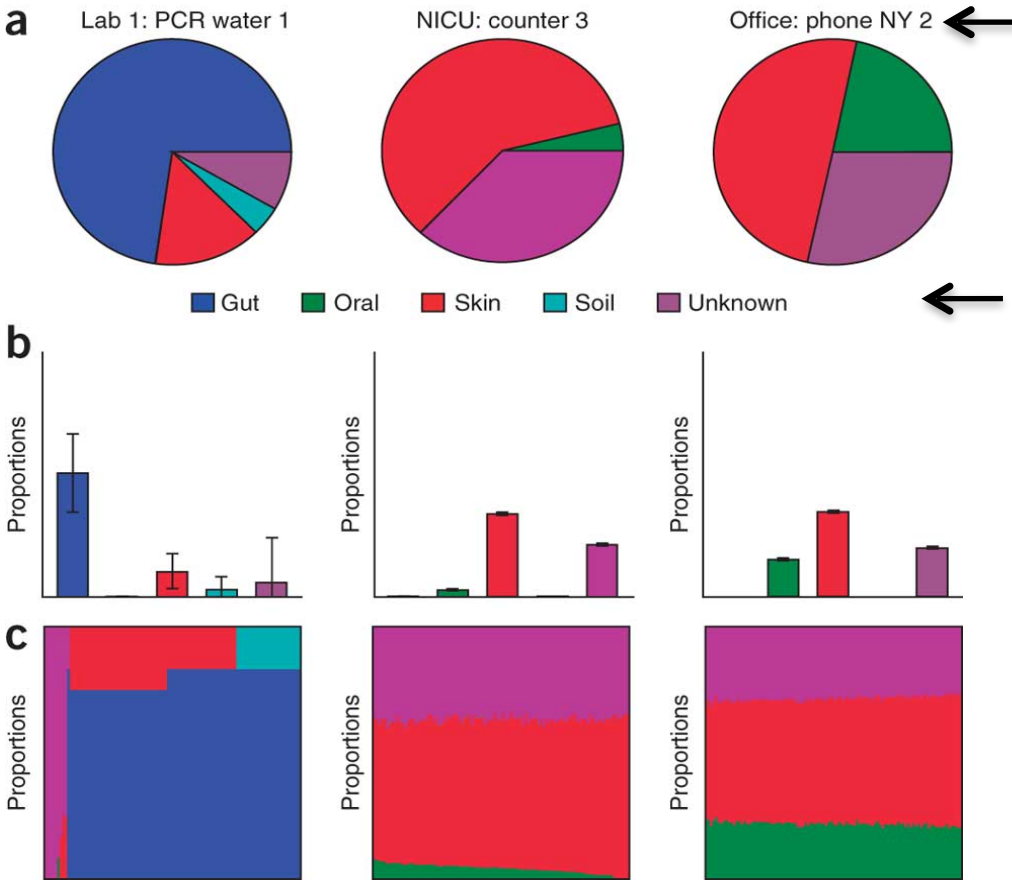


How does it work?





SourceTracker (Bayesian approach)



in a set of input samples

estimate the proportion that come from possible source environments

Bayesian community-wide culture-independent microbial source tracking
Knights et al. 2011 NATURE METHODS

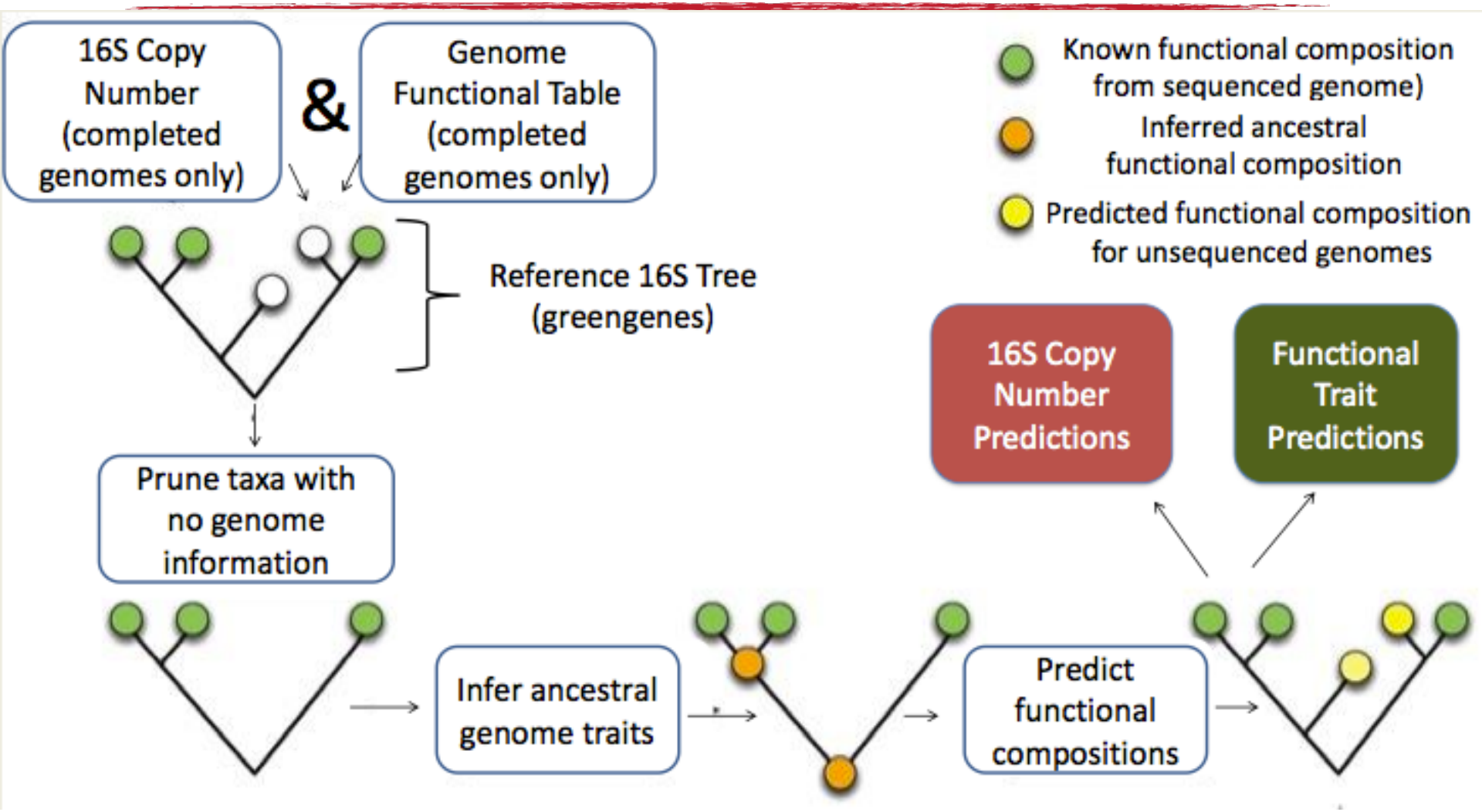
Running these in Qiime

- http://qiime.org/tutorials/running_supervised_learning.html
- http://qiime.org/tutorials/source_tracking.html

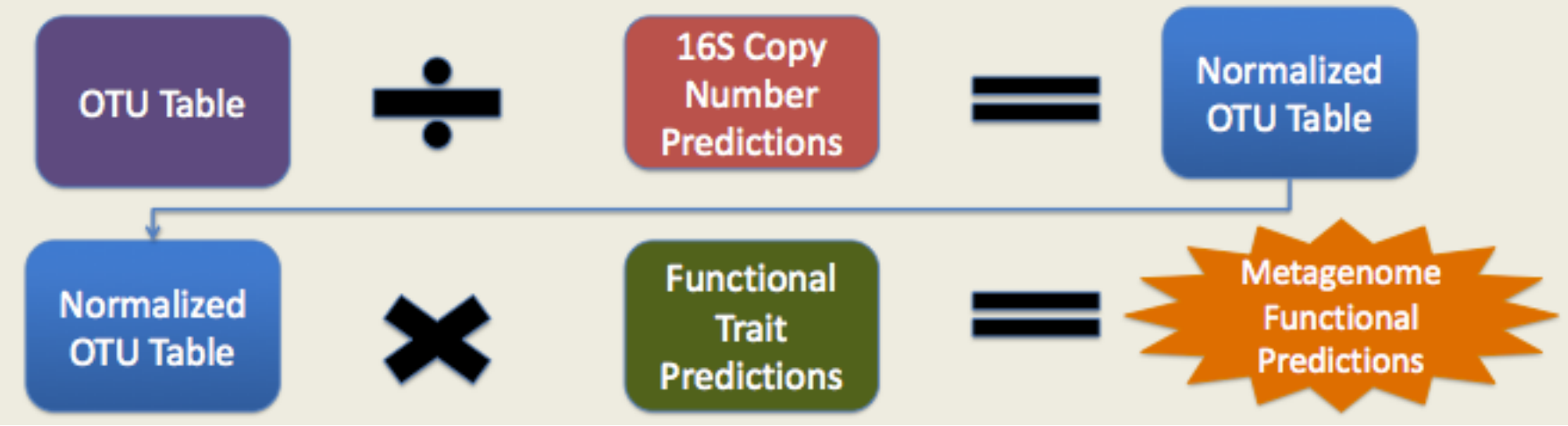
Metagenomic prediction

- Phylogenetic Investigation of Communities by Reconstruction of Unobserved States
  - PICRUSt
- Infer abundance profiles
  - KEGG, PFAM, etc
- Collaboration between Beiko, Knight and Huttenhower labs

Metagenomic prediction



Metagenomic prediction



PICRUSt

- http://picrust.github.io/picrust/

PICRUSt

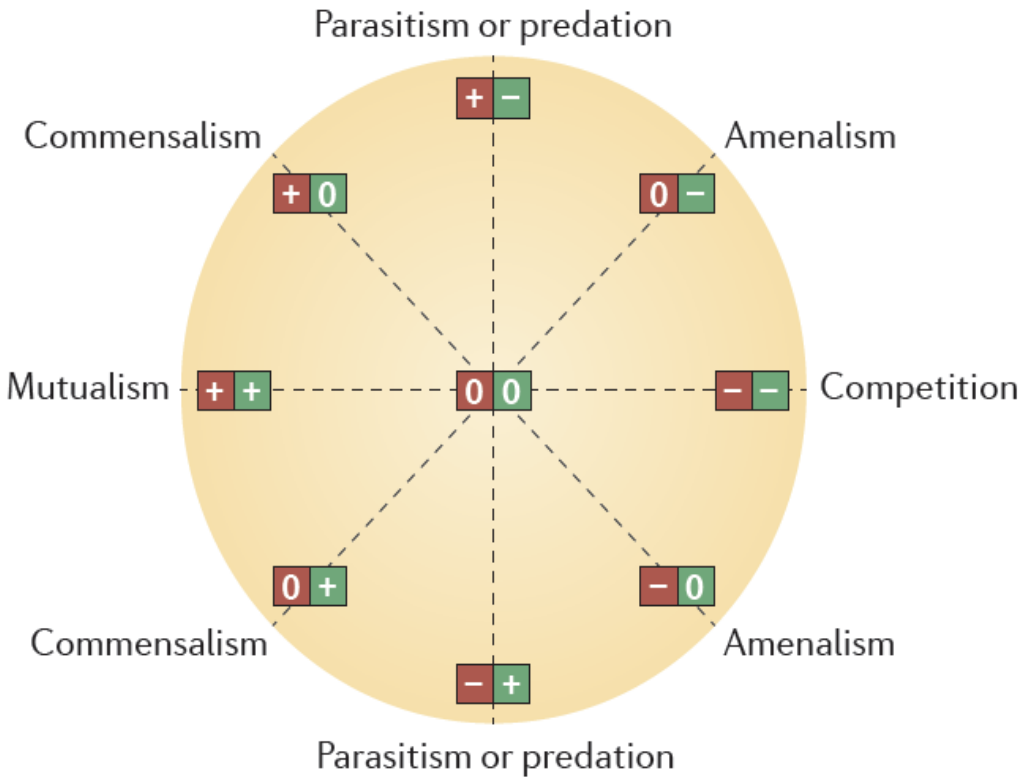
- http://picrust.github.io/picrust/

  Nearest Sequenced Taxon Index (NSTI): how good the predictions are, the lower the better.

  Really well characterized human body sites you get 0.03 ± 0.02 s.d.

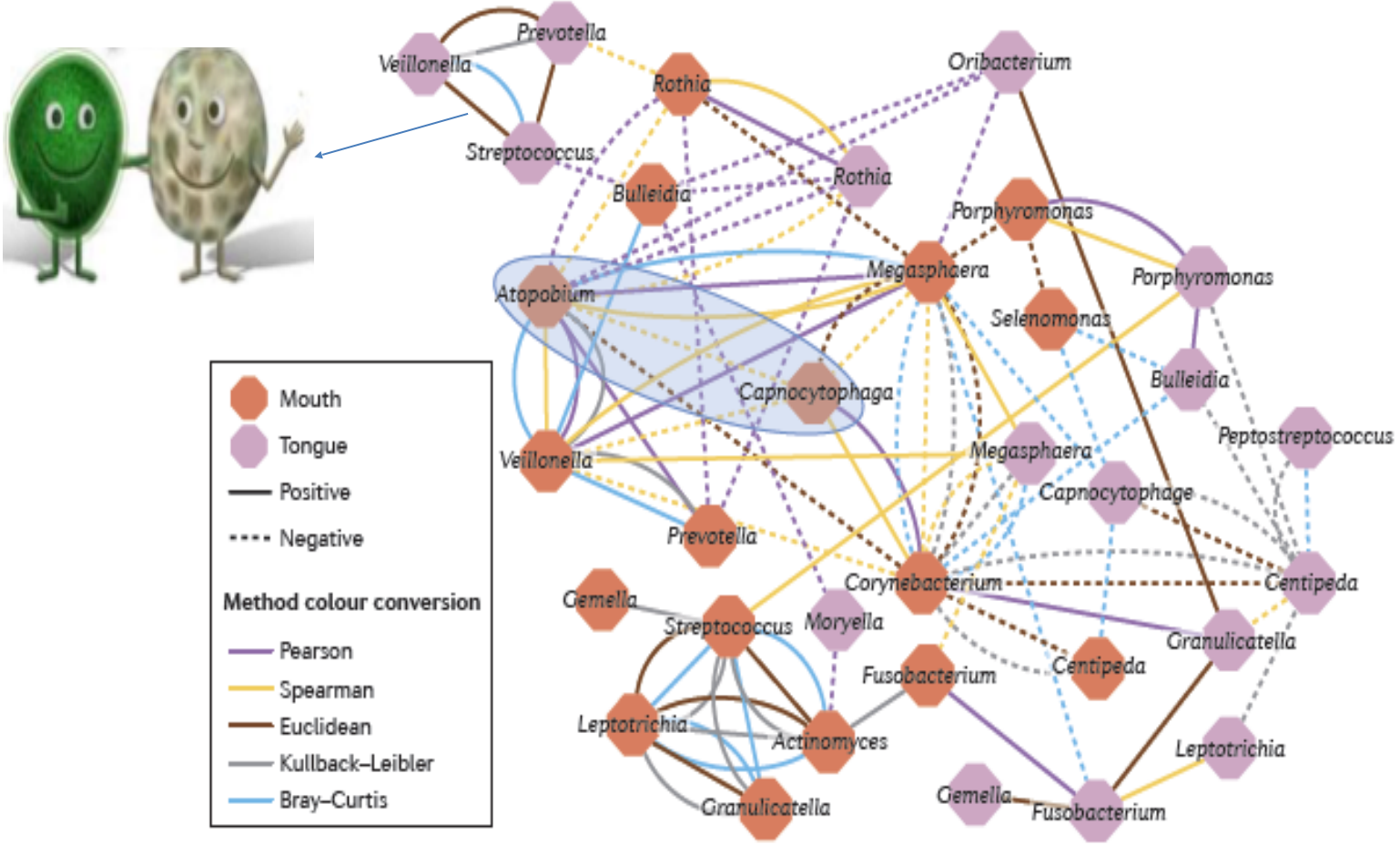
  Hypersaline mat microbiome: 0.23 ± 0.07 s.d.

Co-occurrence



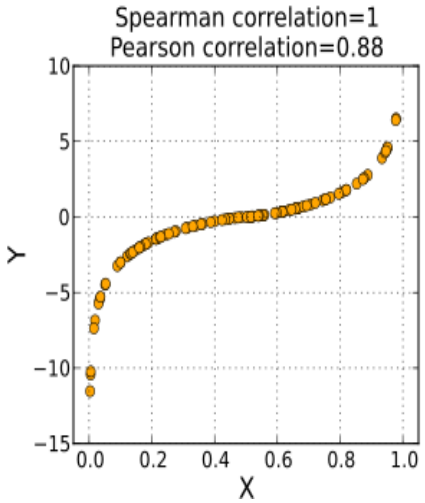
Faust, K., Raes, J., Microbial Interactions from networks to models. *Nature Reviews Microbiology* 10(8):538-50

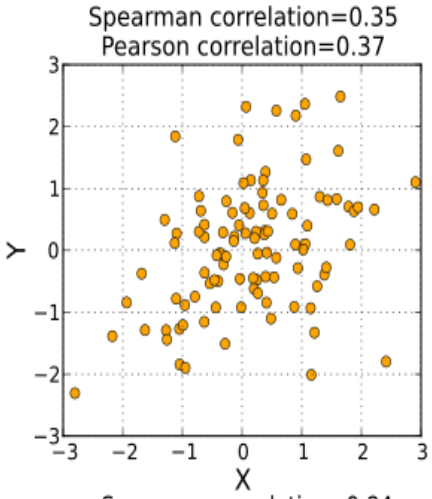
Network structure changes depending on the method (definition of co-occurrence)

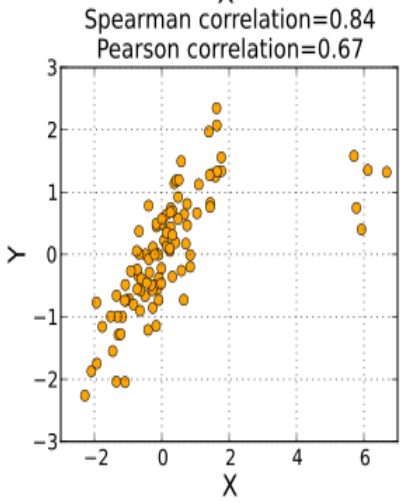


Faust, K., Raes, J., Microbial Interactions from networks to models. *Nature Reviews Microbiology* 10(8):538-50

Co-occurrence: evaluate most popular network methods (Naïve vs. Toolkits)







Naïve correlations:

- Pearson correlation
- Spearman correlation
- Bray Curtis

Toolkits:

- CoNet: Correlation Networks
- SparCC: Sparse Correlations for Compositional Data
- RMT: Random Matrix Theory
- LSA: Local Similarity Analysis
- MIC: Mutual Information Coefficient

Faust, K., et al. (2012): Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS Computational Biology* 8(7): e1002606

Friedman, J., Alm, E. (2012): Inferring Correlation Networks from Genomic Survey Data. *PLoS Computational Biology* 8(9): e1002687.

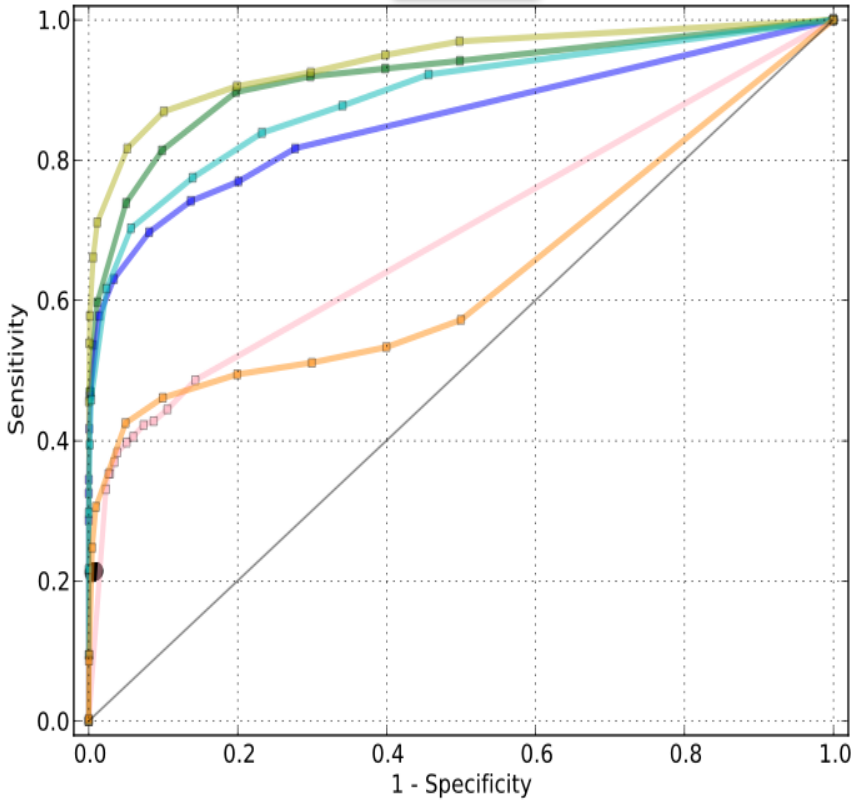
Deng, Y. et al. (2012): Molecular Ecological Network Analysis. *BMC Bioinformatics* (13)113 doi:10.1186/1471-2105-13-113

Xia, LC. et al. (2013): Efficient Statistical Significance Approximation for Local Similarity Analysis of High-Throughput Time Series Data. *Bioinformatics* 29(2): 230-237

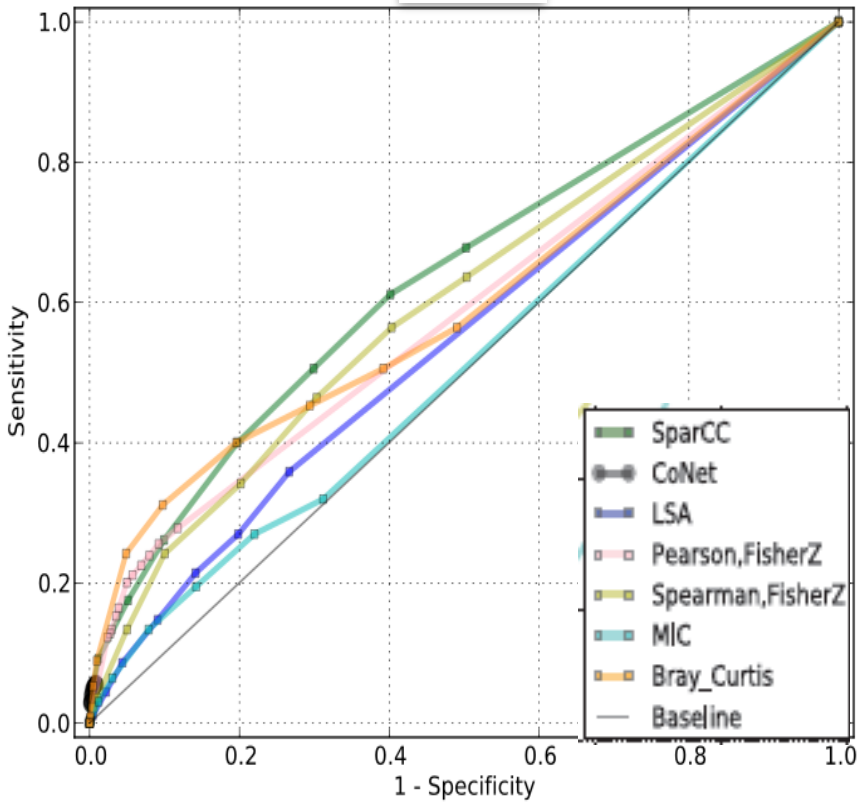
Reshef, DN. Et al. (2011): Detecting Novel Associations in Large Data Sets. *Science. 334:1518-1524*

Sparsity destroys co-occurrence relationship detection

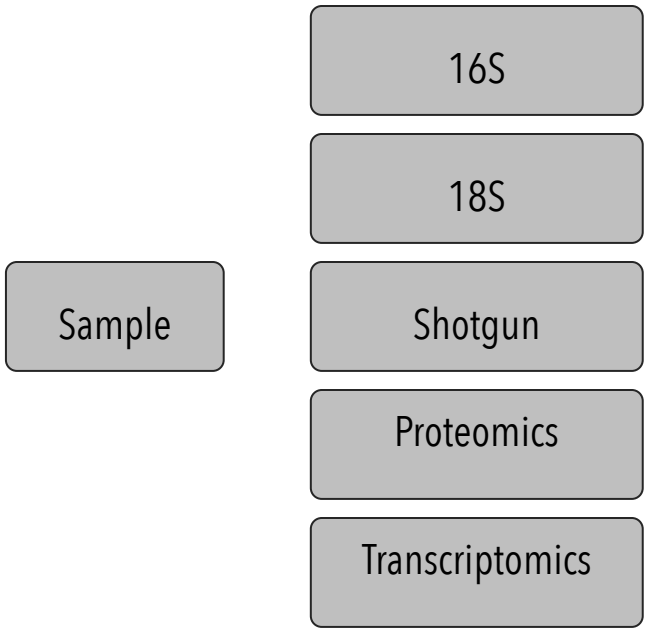
No sparsity



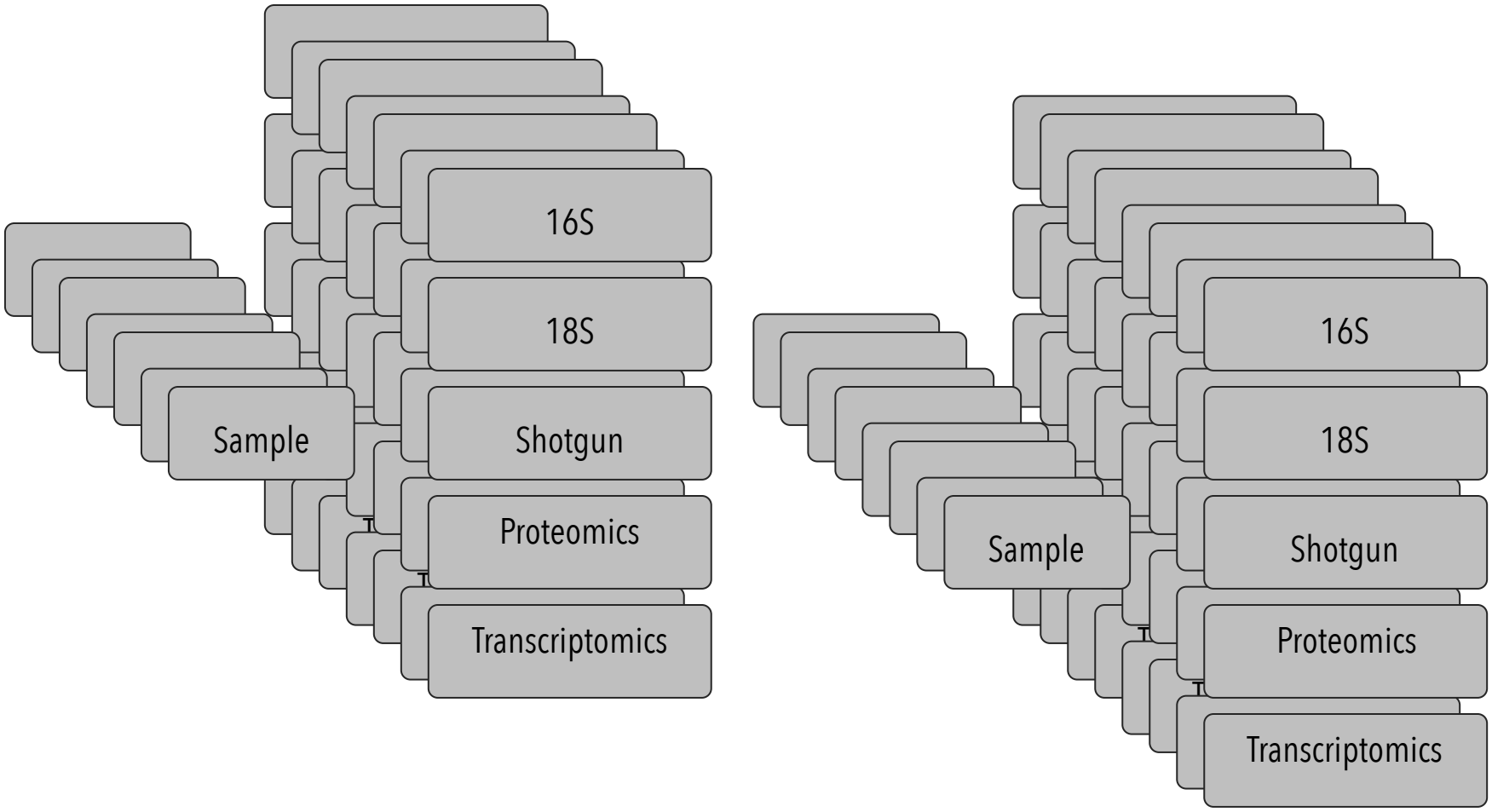
50% sparsity



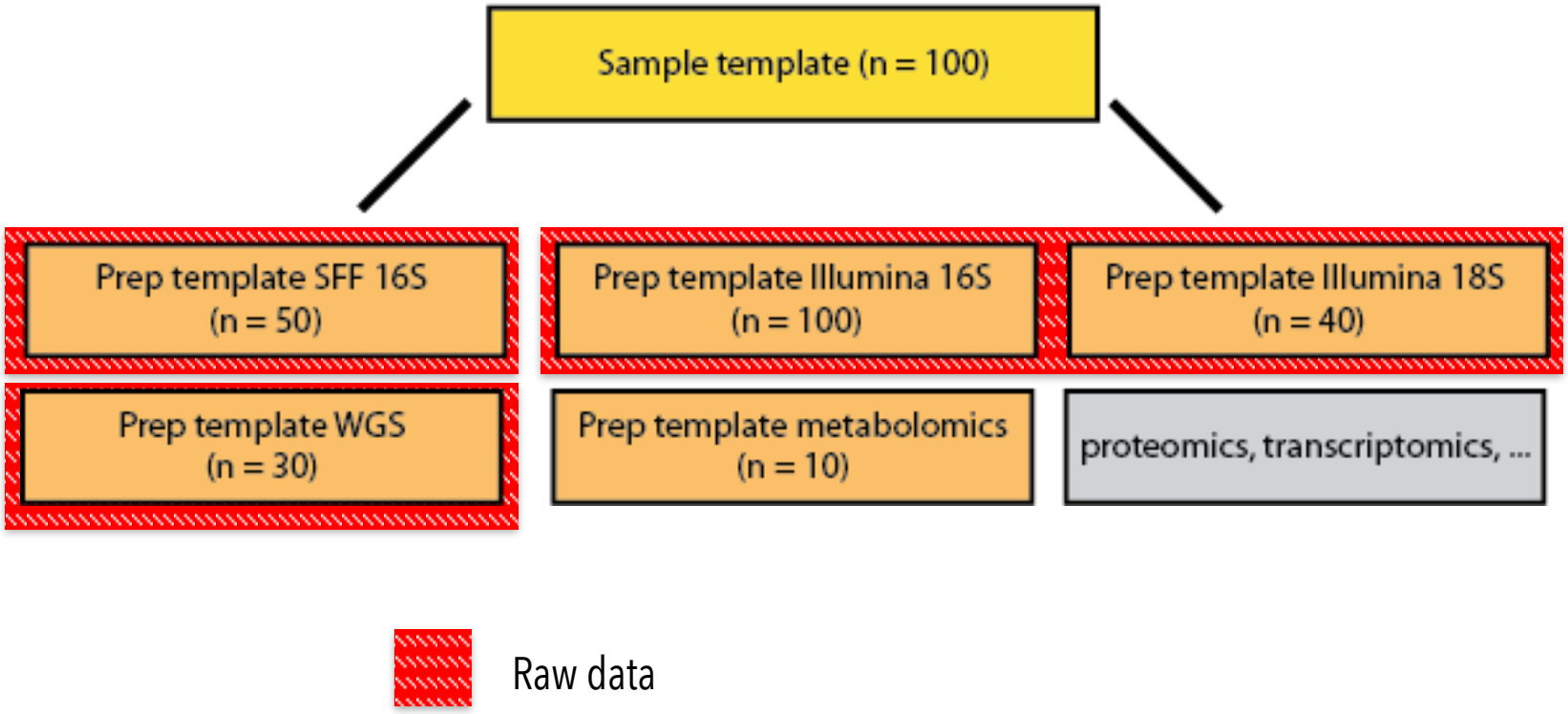
1 sample x m preps



n sample x m preps



Managing your study



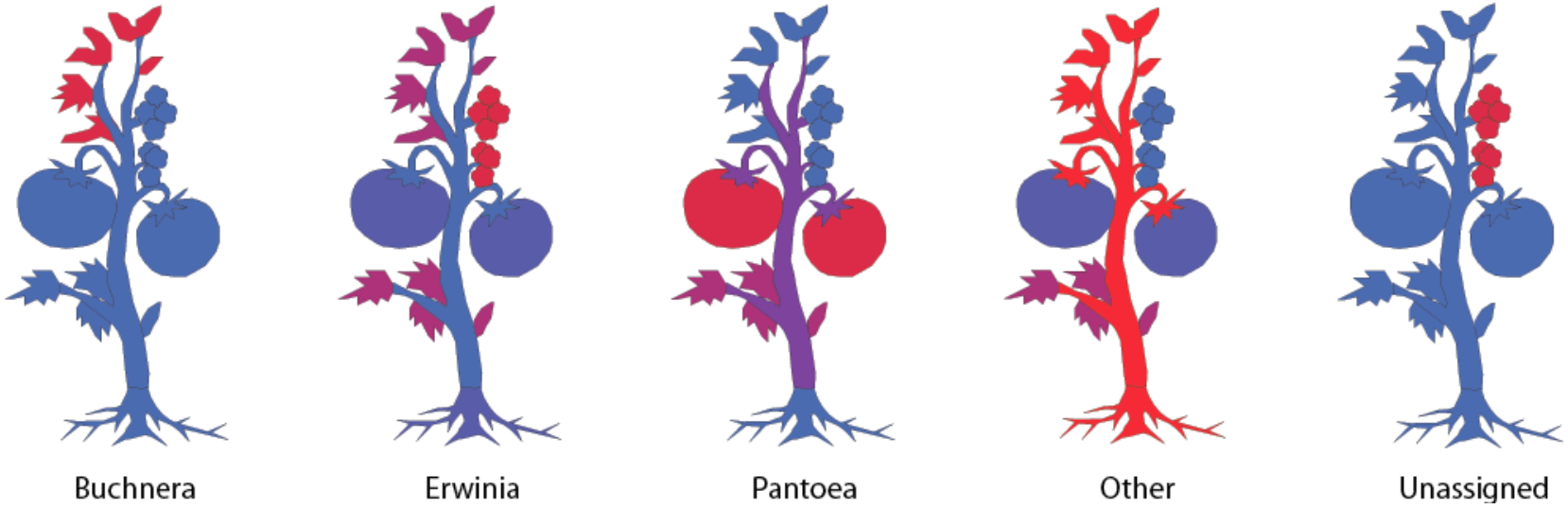
http://qiita.microbio.me

https://github.com/biocore/

Finding species level with shotgun sequencing

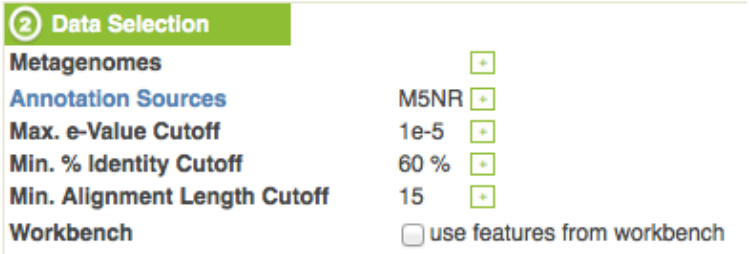
FDA dataset

- Biogeography of the tomato plant
- 16S (Qiime), 18S (Qiime), WGS (MG-RAST)

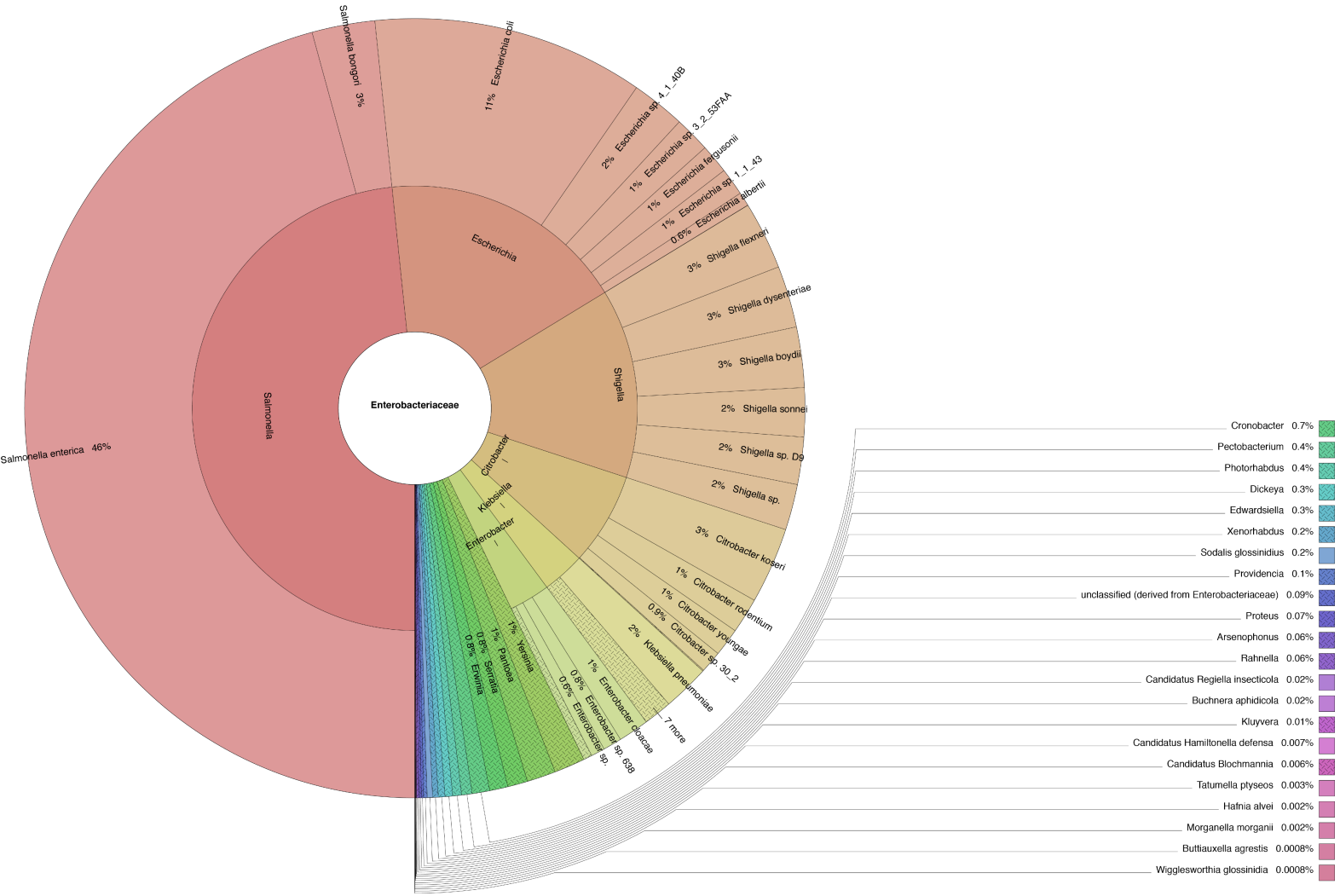


WGS analysis in MG-RAST

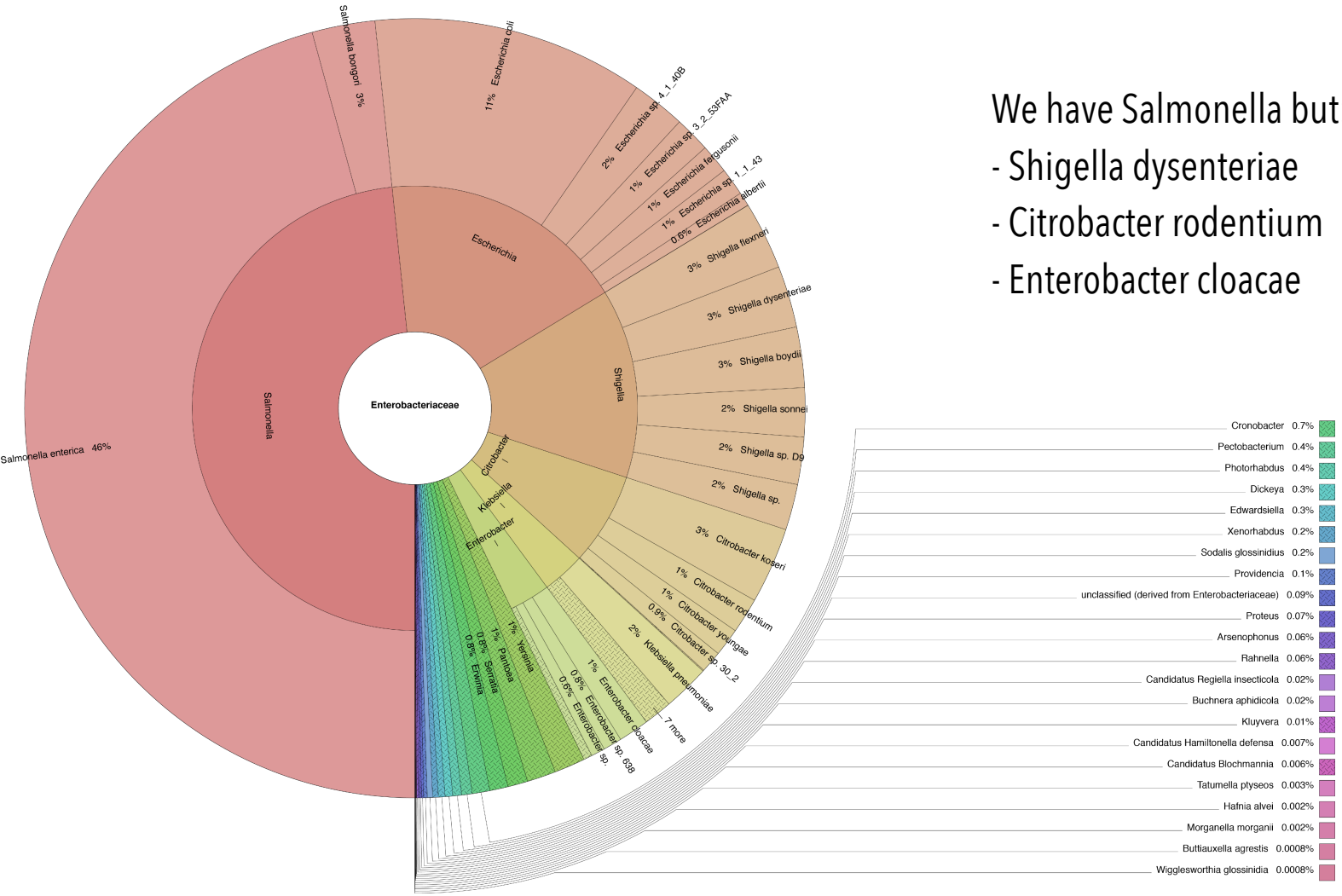
- Upload your datasets, takes some time
- Process the files, takes forever
- Analyze your data set, which default to use?



Let's look at the tomato results



Let's look at the tomato results



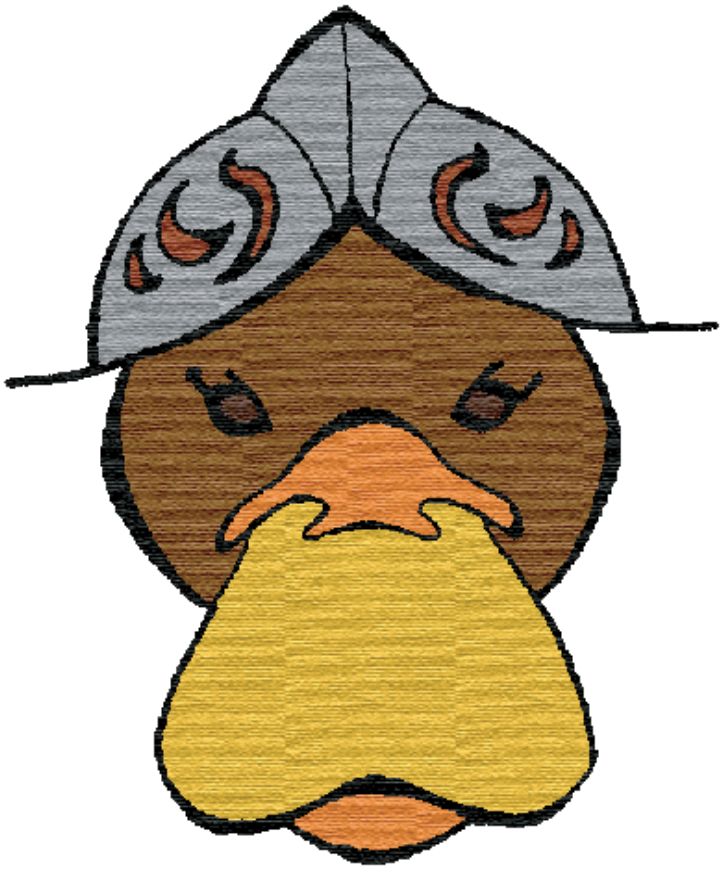
We have Salmonella but:

- Shigella dysenteriae
- Citrobacter rodentium
- Enterobacter cloacae

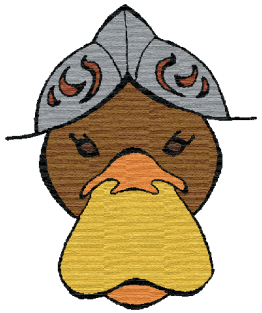
… and lots of other macro fauna

The solution
Platypus Conquistador

- Confirm the presence/absence of specific taxa



The pipeline



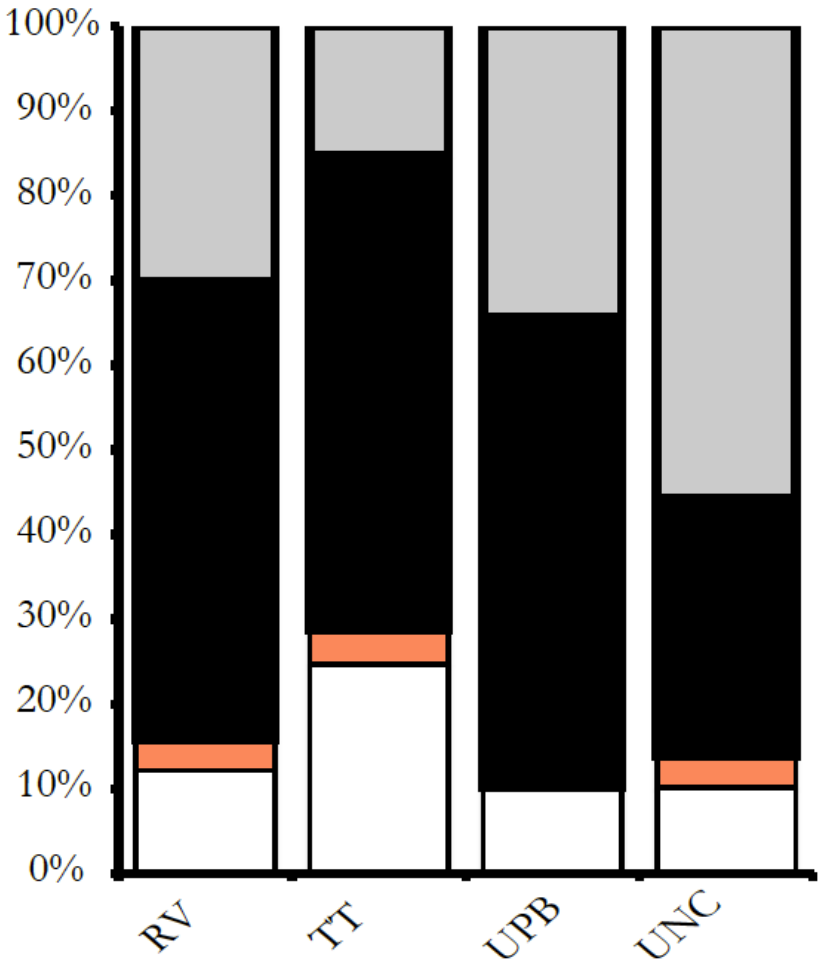
Split reference in 2 (interest and rest)

BLAST/usearch/ sortmerna your sequences against both with low parameters

Platypus to compare at different %ids and alignment lengths

img

Testing on other datasets



RV: Rappaport-Vassiliadis

TT: Tetrathionate

UPB: Universal pre-enrichment broth

UNC: uncultured

OPEN ACCESS PEER-REVIEWED
RESEARCH ARTICLE

1,249 VIEWS | 2 CITATIONS | 10 SAVES | 10 SHARES

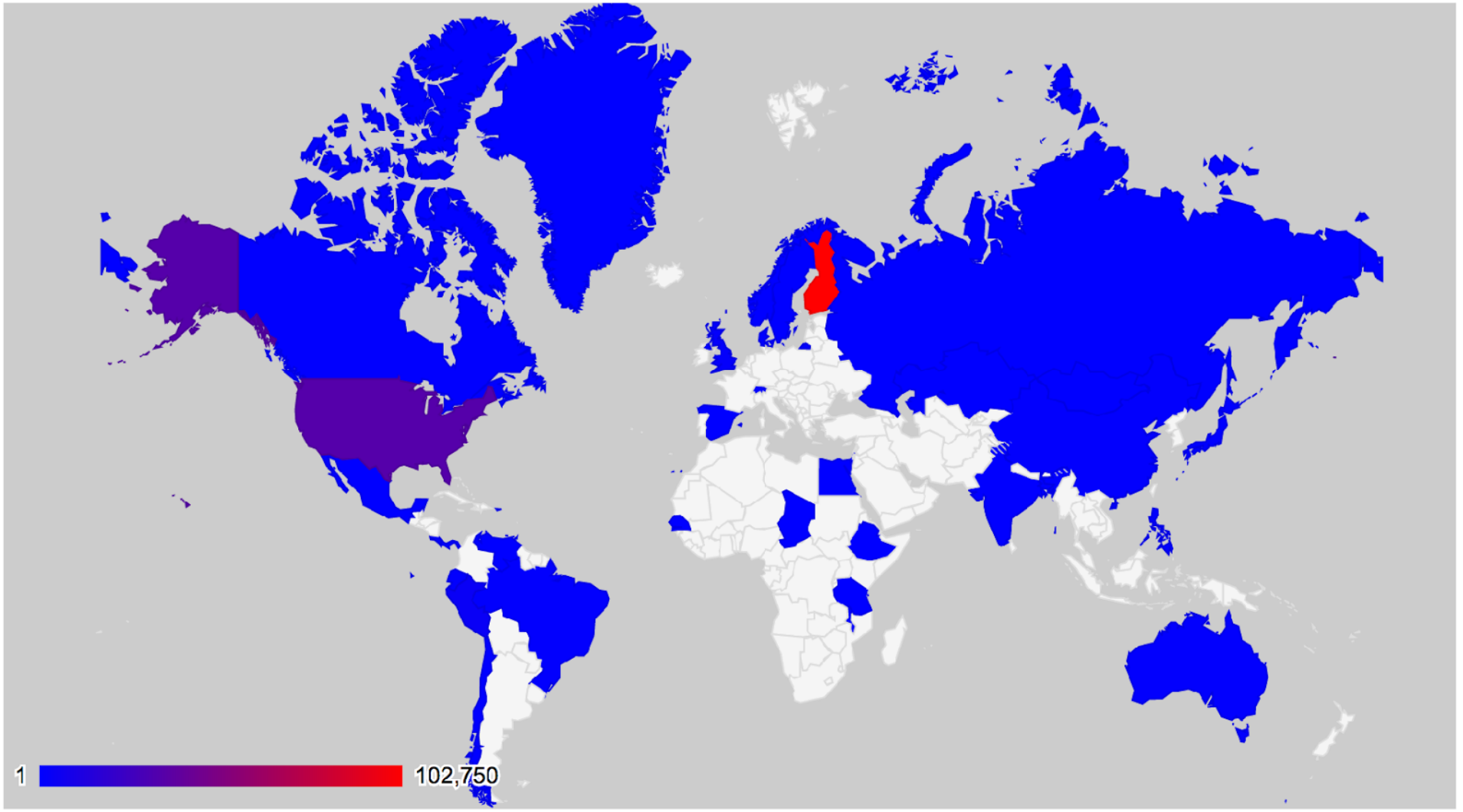
Co-Enriching Microflora Associated with Culture Based Methods to Detect *Salmonella* from Tomato Phyllosphere

Andrea R. Ottesen, Antonio Gonzalez, Rebecca Bell, Caroline Arce, Steven Rideout, Marc Allard, Peter Evans, Errol Strain, Steven Musser, Rob Knight, Eric Brown, James B. Pettengill

Published: September 09, 2013 • DOI: 10.1371/journal.pone.0073079

How pervasive?

How pervasive?
Long live the Platypus!



Extinct species

Dodo bird

Biomes:

2 marine benthic biome

11 terrestrial biome

Country:

USA

Tasmanian tiger

Biomes:

989 terrestrial biome

105 marine benthic biome

63. Prairie Division (250)

Countries:

USA, Chad, Finland, Malaysia, China

Building GreenGenes (from Daniel McDonald)

1. Get primary data
2. Align with ssu-align

   For QIIME we use PyNAST, which is a quick but not perfect solution
3. Toss out anything that doesn't align reasonably well, track what is tossed out
4. Parse records for the sequences that seem to make sense to retrieve named isolate detail, sequence stats, record info, etc.
   1. Take the named isolate subset (sequenced genome associated)
   2. Pick de novo otus at 99% on those sequences
   3. Take all non named isolates, pick at closed reference against the db formed by previous step
   4. Take all failures from, and pick de novo
   5. Combine to get a single set of OTUs

Building GreenGenes (from Daniel McDonald)

5. Mask and expand alignment such that it fits into the lane mask
6. Build and root the tree
7. Take sequences that went into tree, and realign with PyNAST

   ssu-align drops internal positions, thus not suitable for primer design
8. Dump records in chunked arb format for Phil

Thanks!

- See you tomorrow!