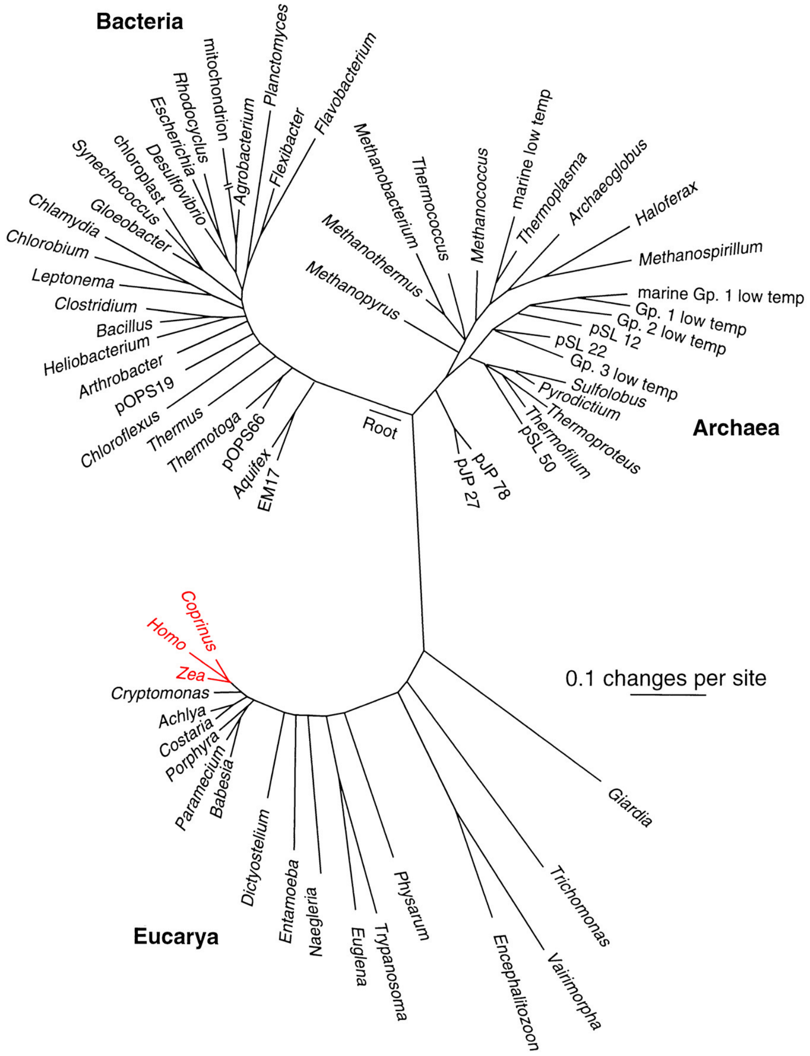
Introduction to QIIME on the IPython Notebook

Daniel McDonald-(🐦@mcdonadt)
Yoshiki Vázquez-Baeza-(🐦@yosmark)

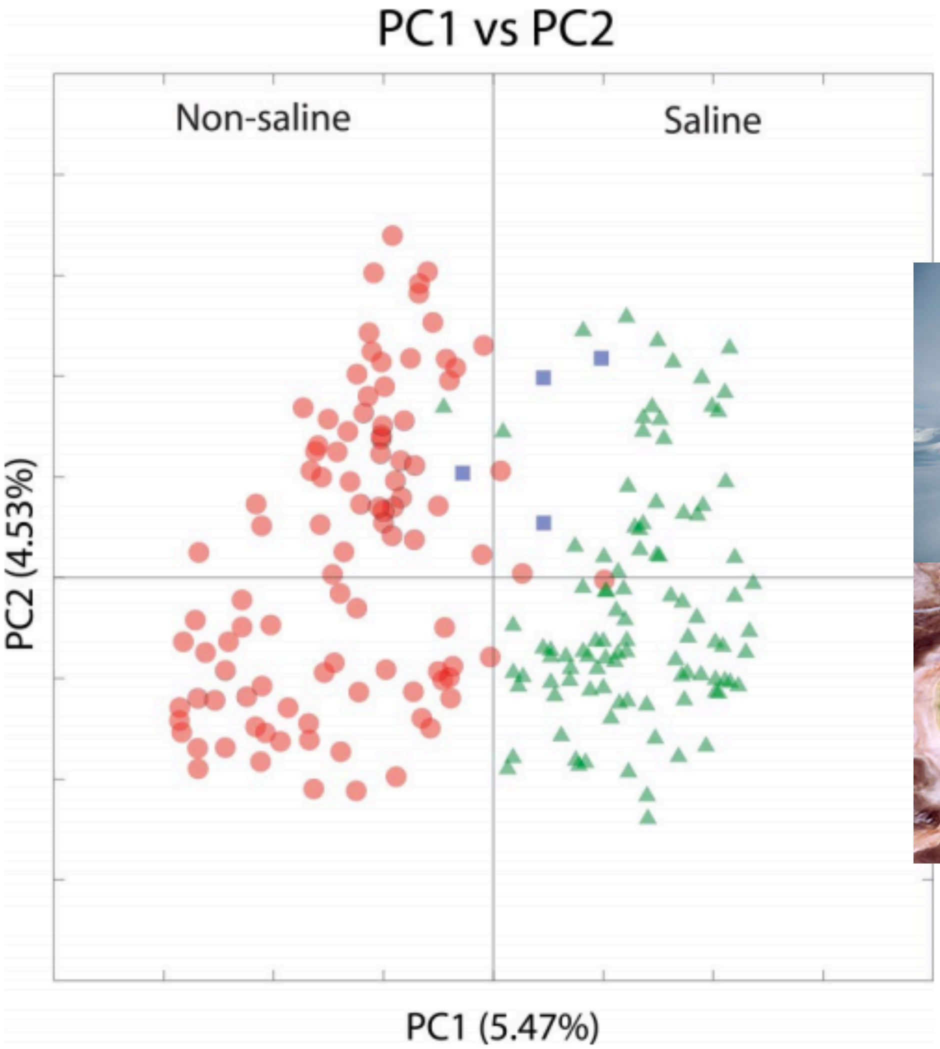
DO NOT START THE EC2 INSTANCES … WE WILL DO THIS IN A DIFFERENT WAY TODAY!!!!

A microbe dominated world



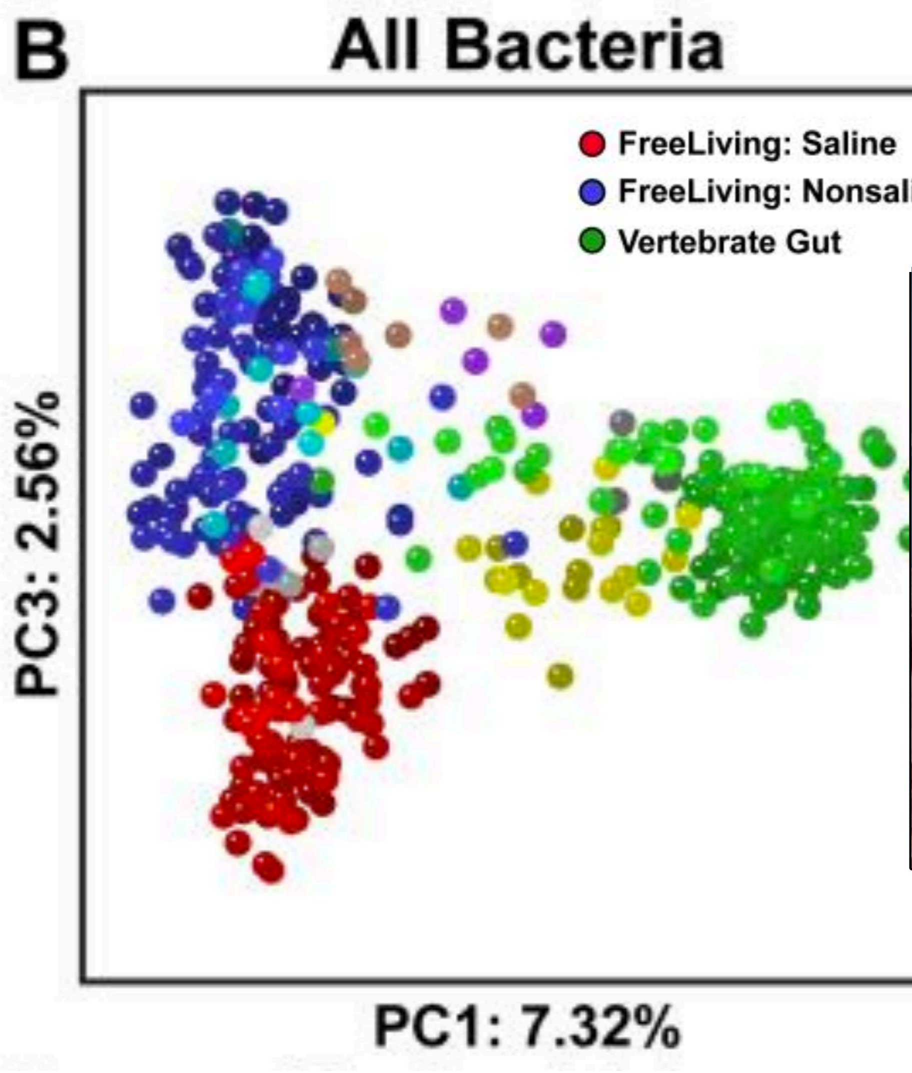
The universal nature of biochemistry. Pace NR.
Proc Natl Acad Sci U S A. 2001 Jan 30;98(3):805-8.

A microbe dominated world



Lozupone and Knight PNAS. 2007 Jul 3;104(27):11436-40

A microbe dominated world



Ley et al Nat Rev Microbiol. Oct 2008; 6(10): 776–788

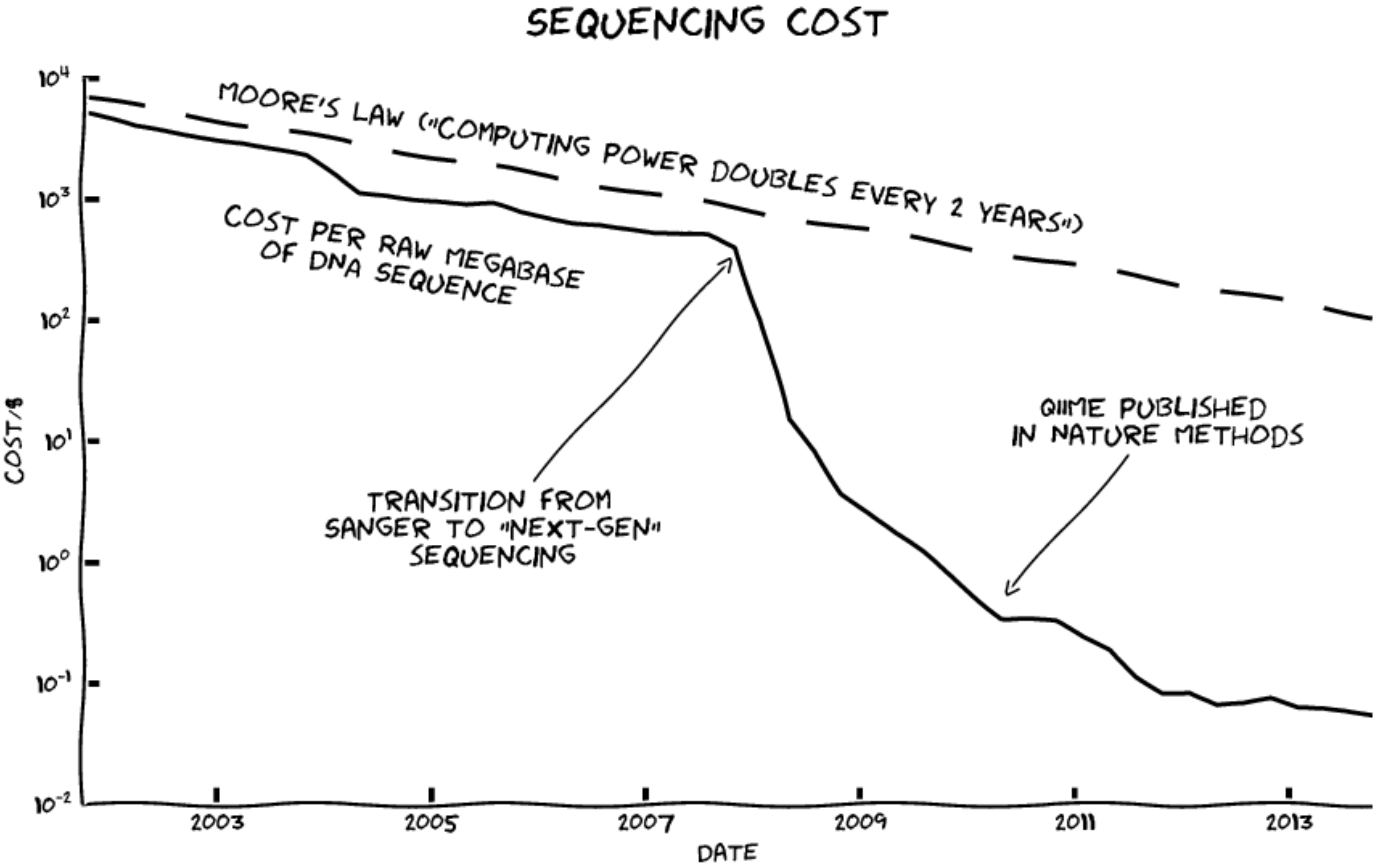
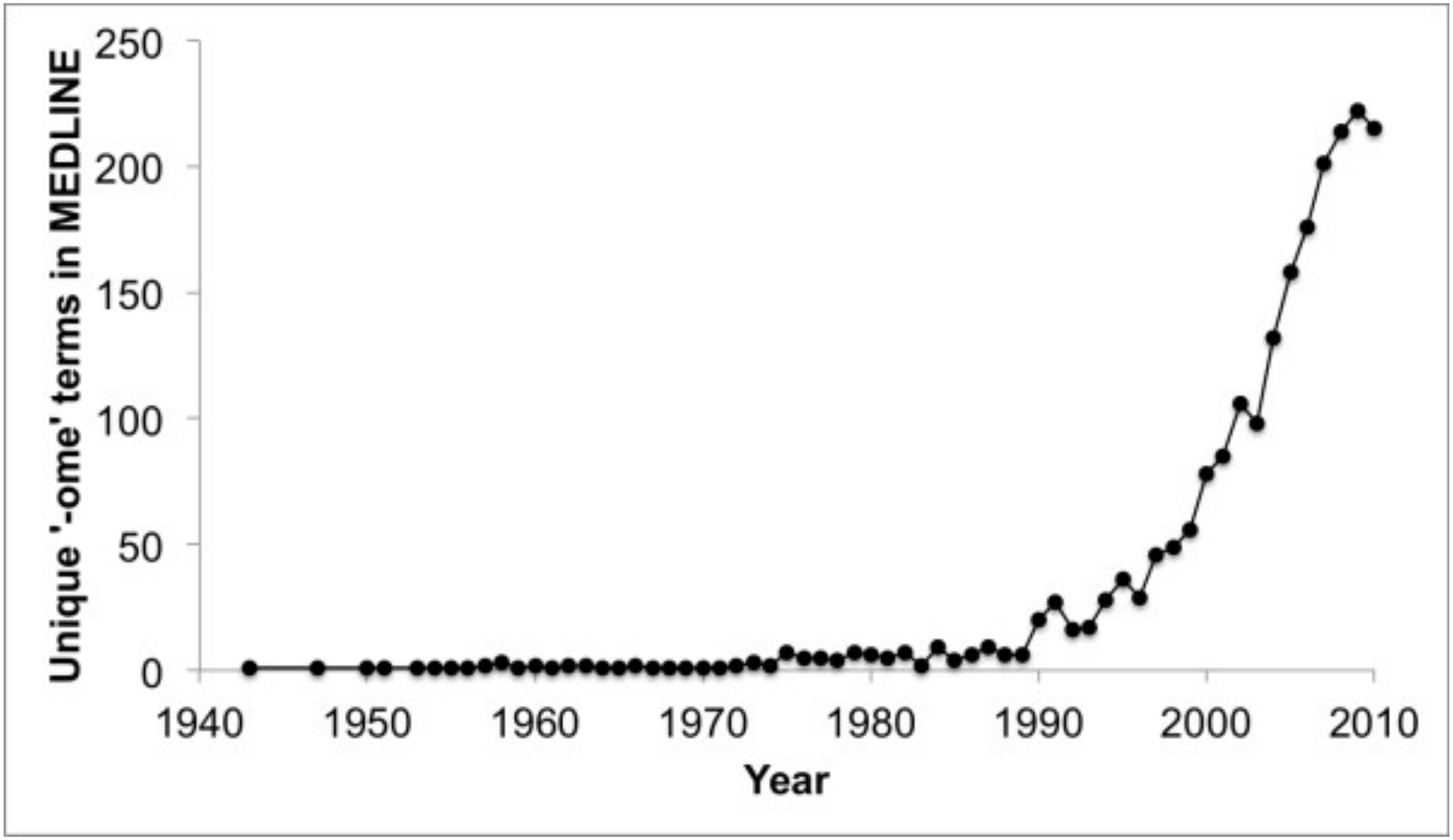


Image credit: Yoshiki Vazquez Baeza and Jorge Cañardo (Github: @ElDeveloper, @Jorge-C)

…an explosion of -omics

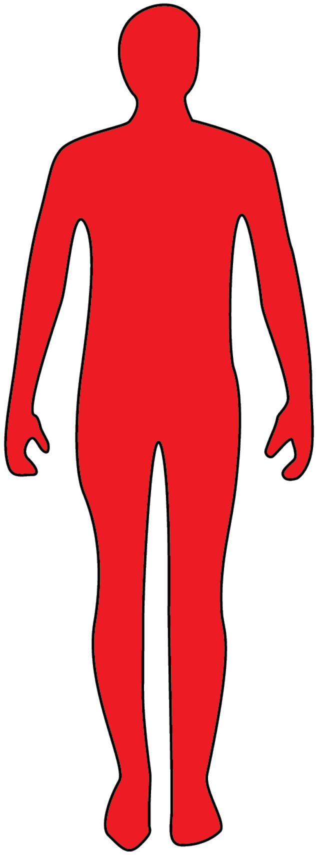


McDonald, Clemente, et al. Gigascience 2012 1:7

How “human” are we, really?

Human

10 trillion human cells
20,000 human genes



Hamady, 2009, PhD thesis

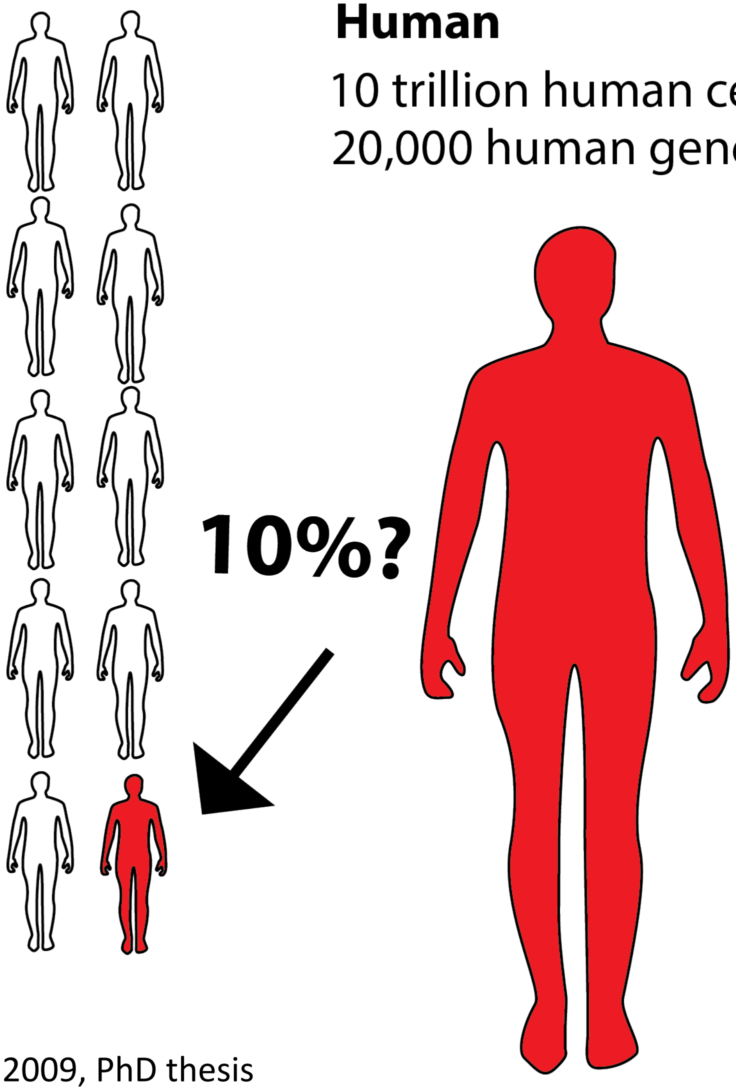
How “human” are we, really?

Microbiota

100 trillion microbial cells

Human

10 trillion human cells
20,000 human genes



Hamady, 2009, PhD thesis

http://www.g4tv.com/attackoftheshow/blog/post/717271/velociraptor-riding-a-shark/

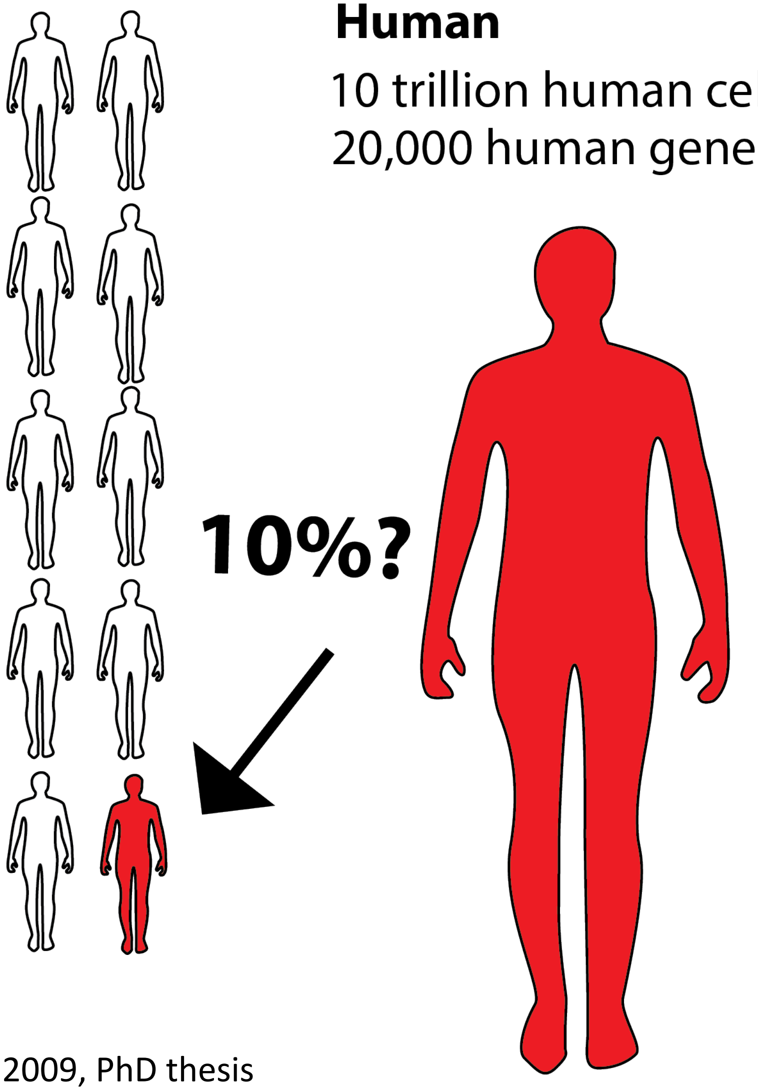
How "human" are we, really?

Microbiota

100 trillion microbial cells

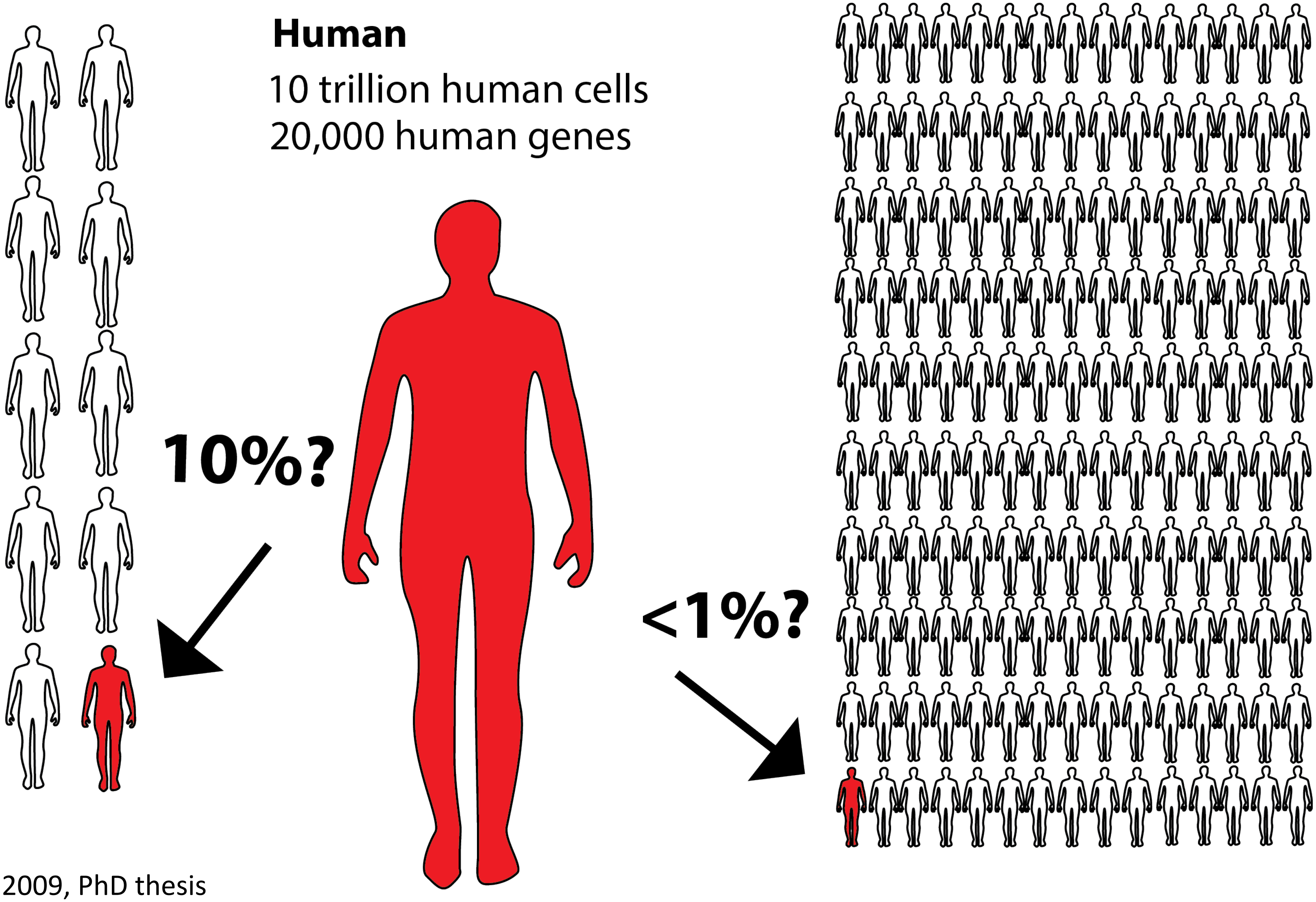
Human

10 trillion human cells
20,000 human genes



Hamady, 2009, PhD thesis

How “human” are we, really?



Microbiota
100 trillion microbial cells

Human
10 trillion human cells
20,000 human genes

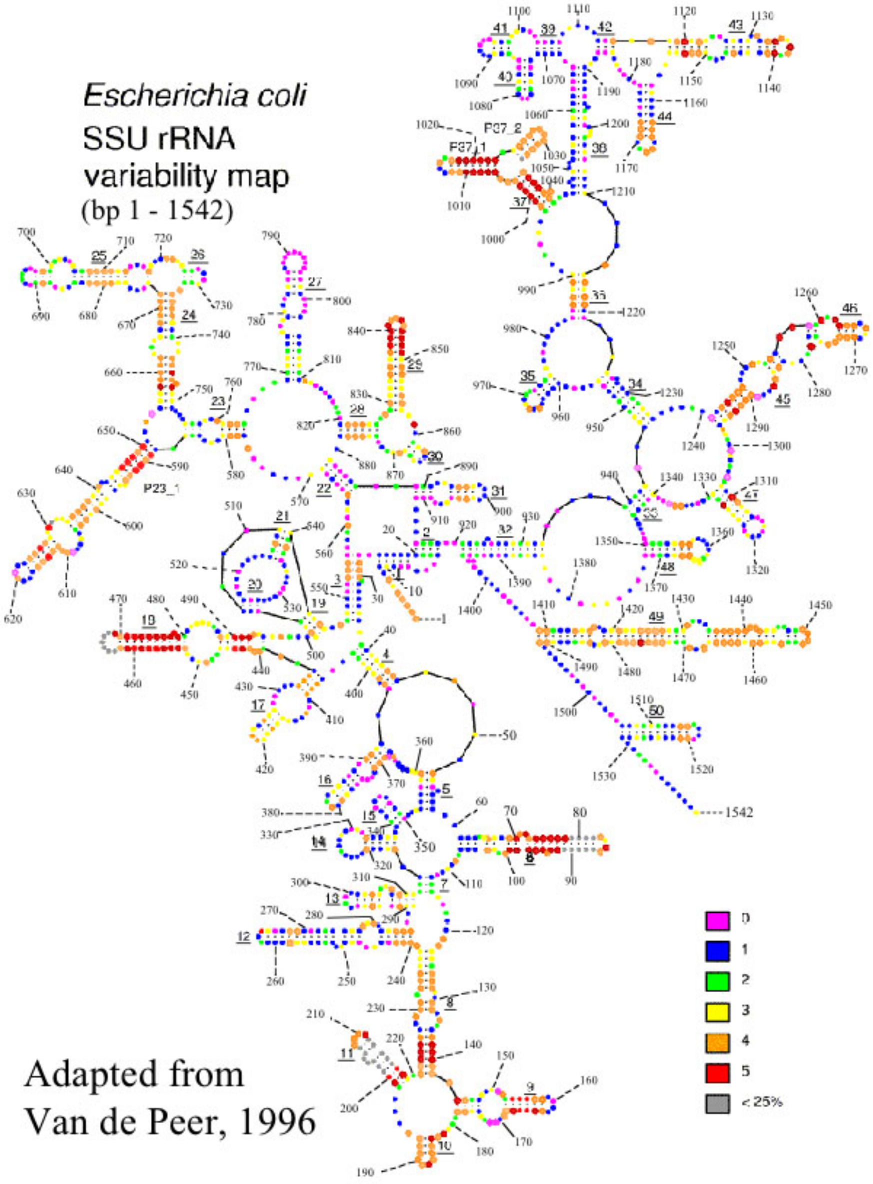
Microbiome
3,000,000 microbial genes

10%?

<1%?

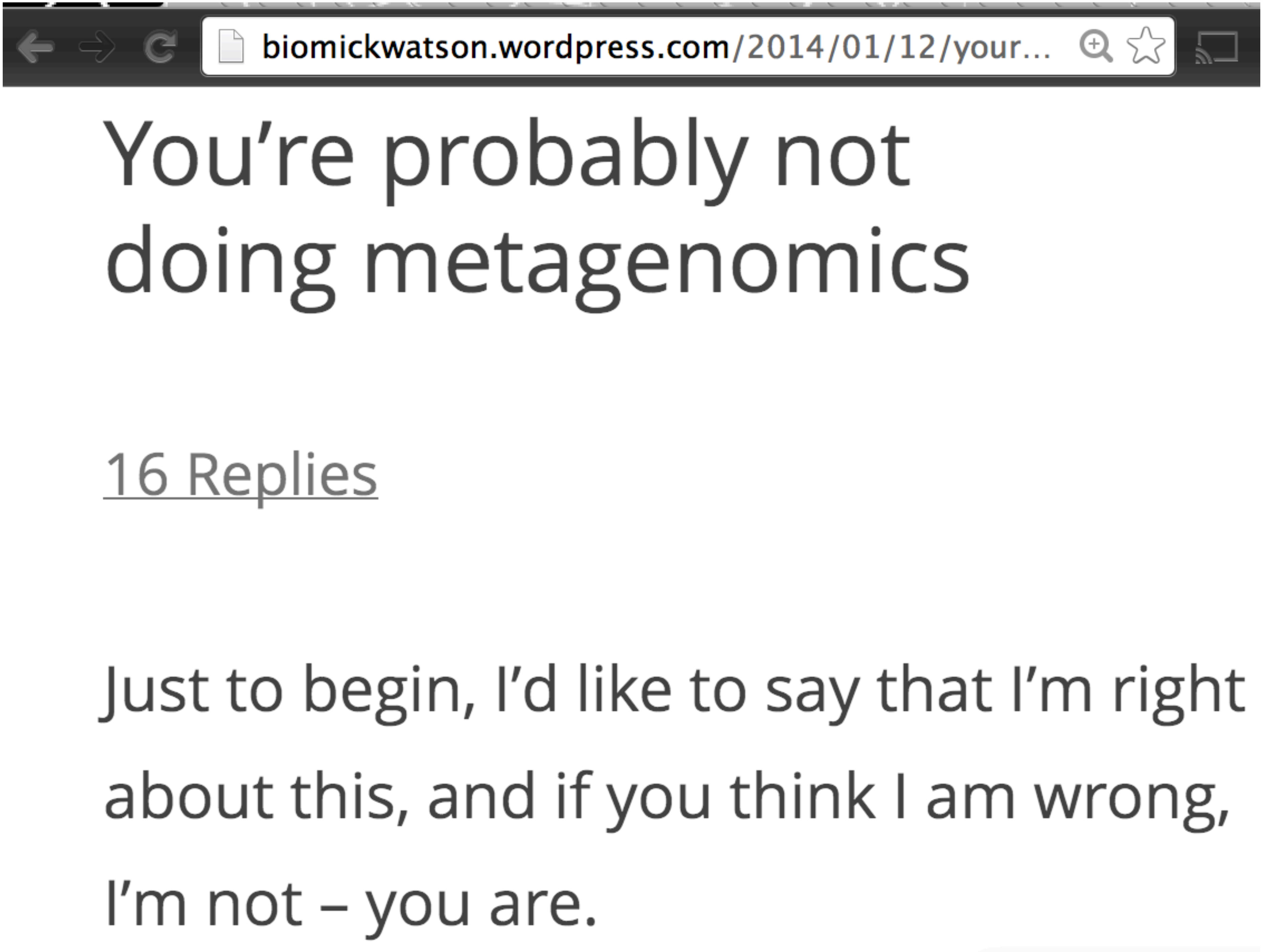
Hamady, 2009, PhD thesis

How do we assay this diversity?



Escherichia coli
SSU rRNA
variability map
(bp 1 - 1542)

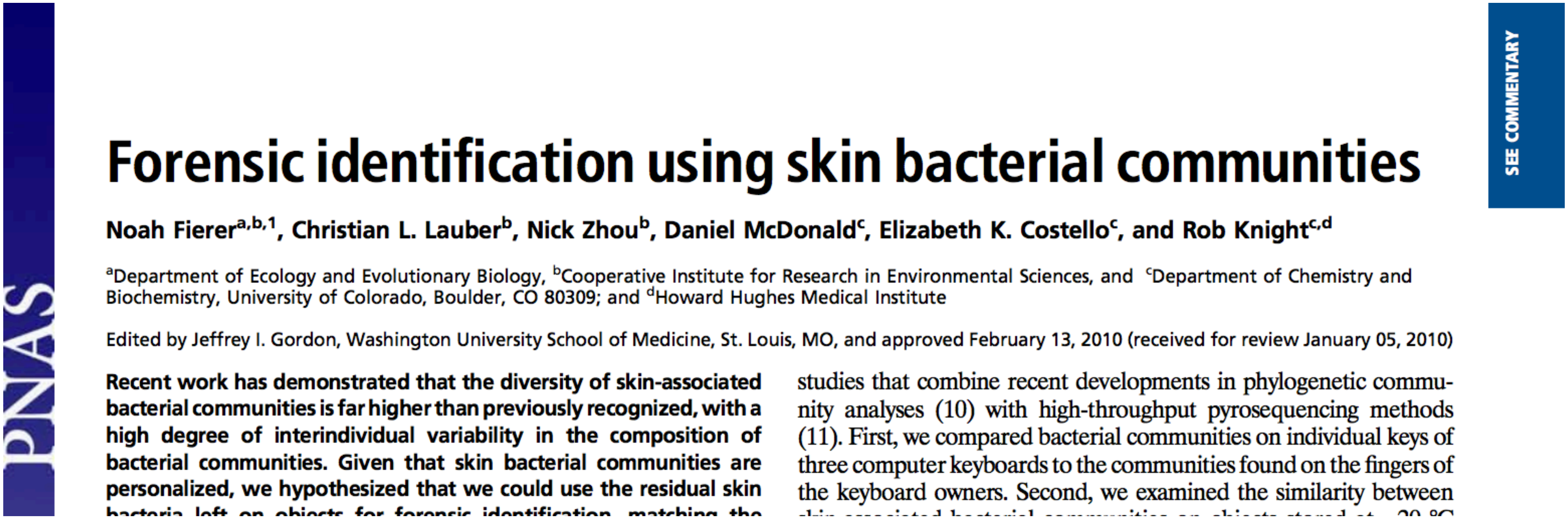
Adapted from
Van de Peer, 1996



You're probably not doing metagenomics

16 Replies

Just to begin, I'd like to say that I'm right about this, and if you think I am wrong, I'm not – you are.



Forensic identification using skin bacterial communities

Noah Fierer[a,b,1], Christian L. Lauber[b], Nick Zhou[b], Daniel McDonald[c], Elizabeth K. Costello[c], and Rob Knight[c,d]

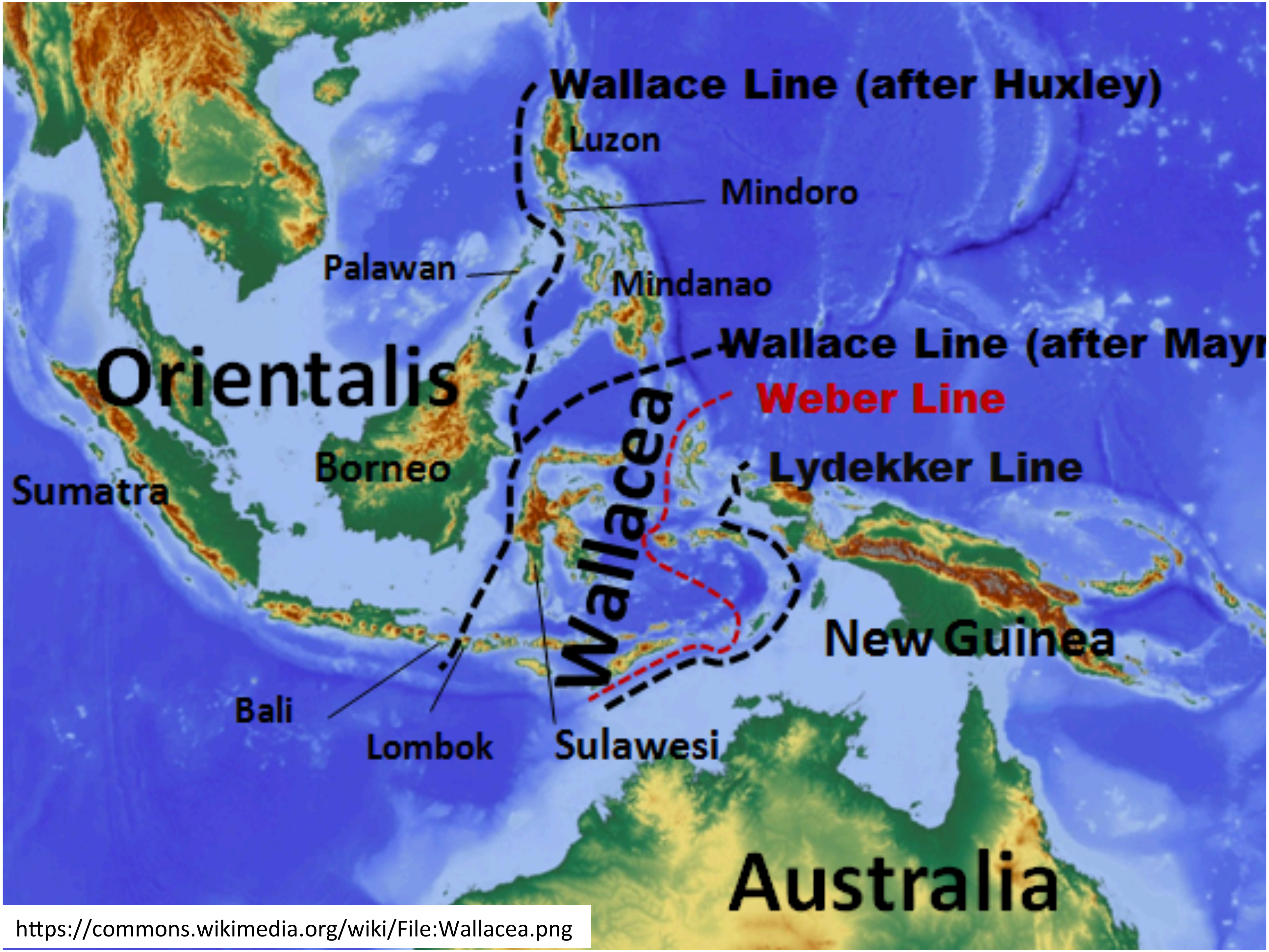
[a]Department of Ecology and Evolutionary Biology, [b]Cooperative Institute for Research in Environmental Sciences, and [c]Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO 80309; and [d]Howard Hughes Medical Institute

Edited by Jeffrey I. Gordon, Washington University School of Medicine, St. Louis, MO, and approved February 13, 2010 (received for review January 05, 2010)

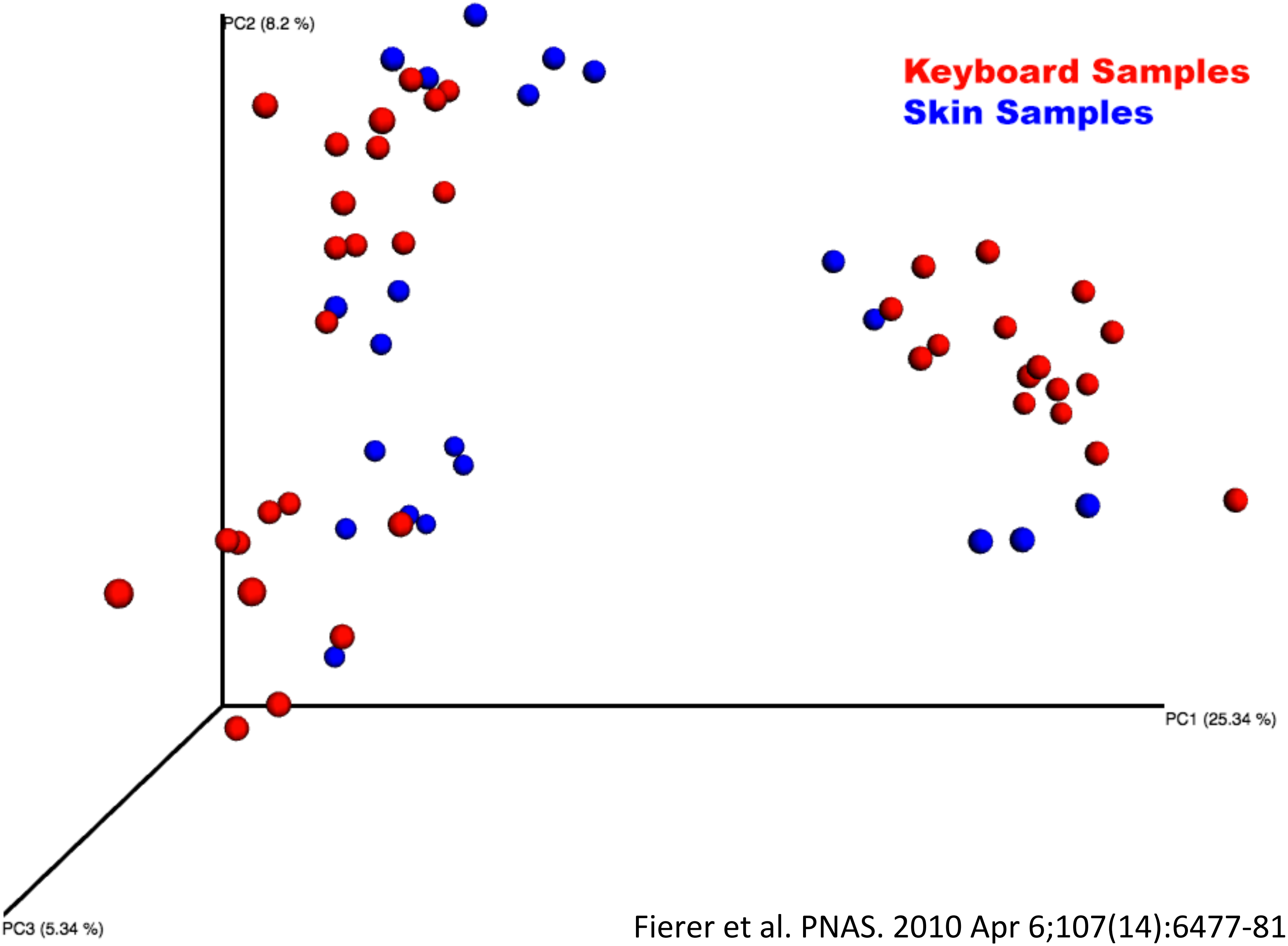
Recent work has demonstrated that the diversity of skin-associated bacterial communities is far higher than previously recognized, with a high degree of interindividual variability in the composition of bacterial communities. Given that skin bacterial communities are personalized, we hypothesized that we could use the residual skin bacteria left on objects for forensic identification, matching the

studies that combine recent developments in phylogenetic community analyses (10) with high-throughput pyrosequencing methods (11). First, we compared bacterial communities on individual keys of three computer keyboards to the communities found on the fingers of the keyboard owners. Second, we examined the similarity between

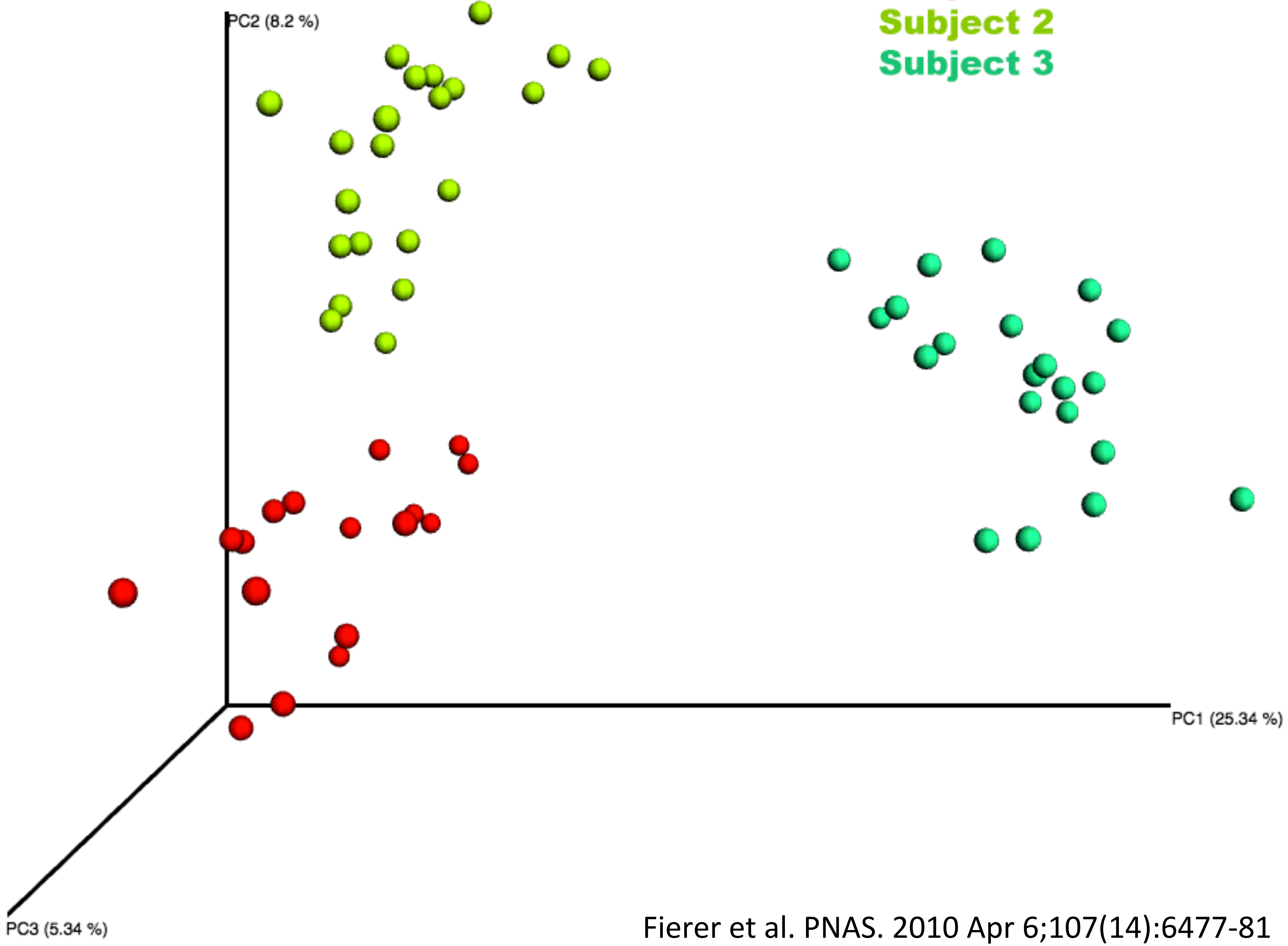
Photo credit: Steve Miller, CIRES



https://commons.wikimedia.org/wiki/File:Wallacea.png



Fierer et al. PNAS. 2010 Apr 6;107(14):6477-81



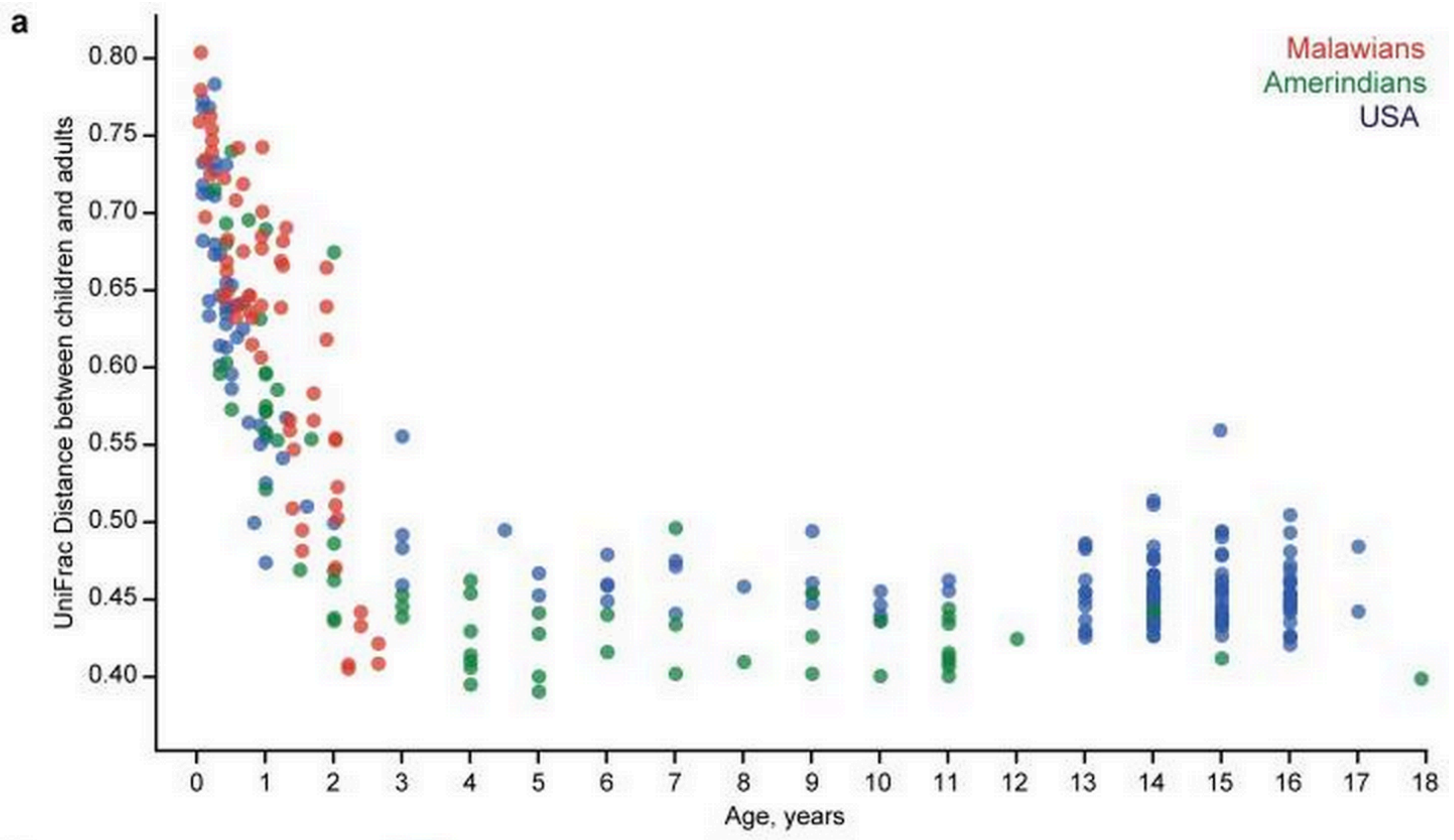
Fierer et al. PNAS. 2010 Apr 6;107(14):6477-81

ARTICLE

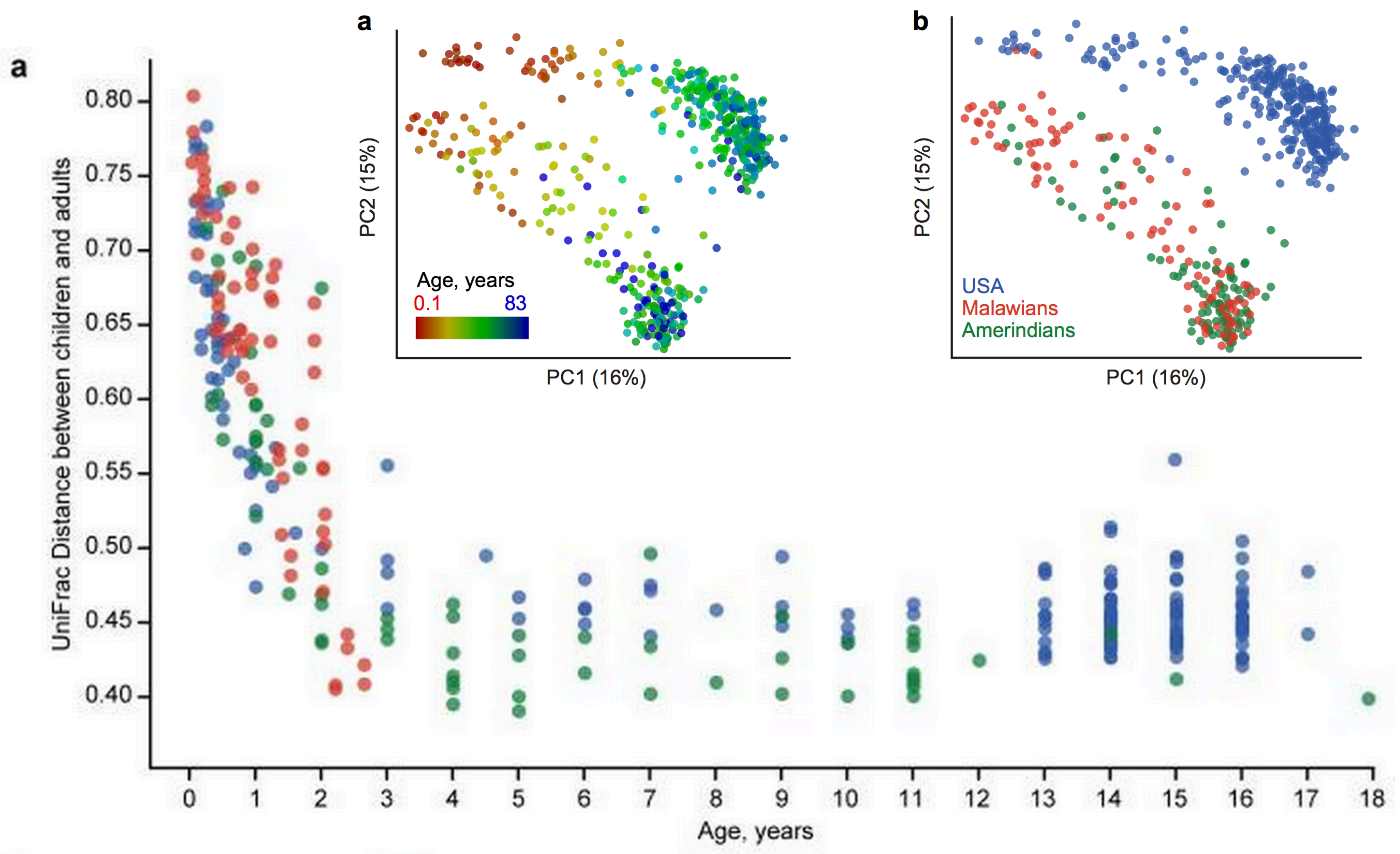
doi:10.1038/nature11053

Human gut microbiome viewed across age and geography

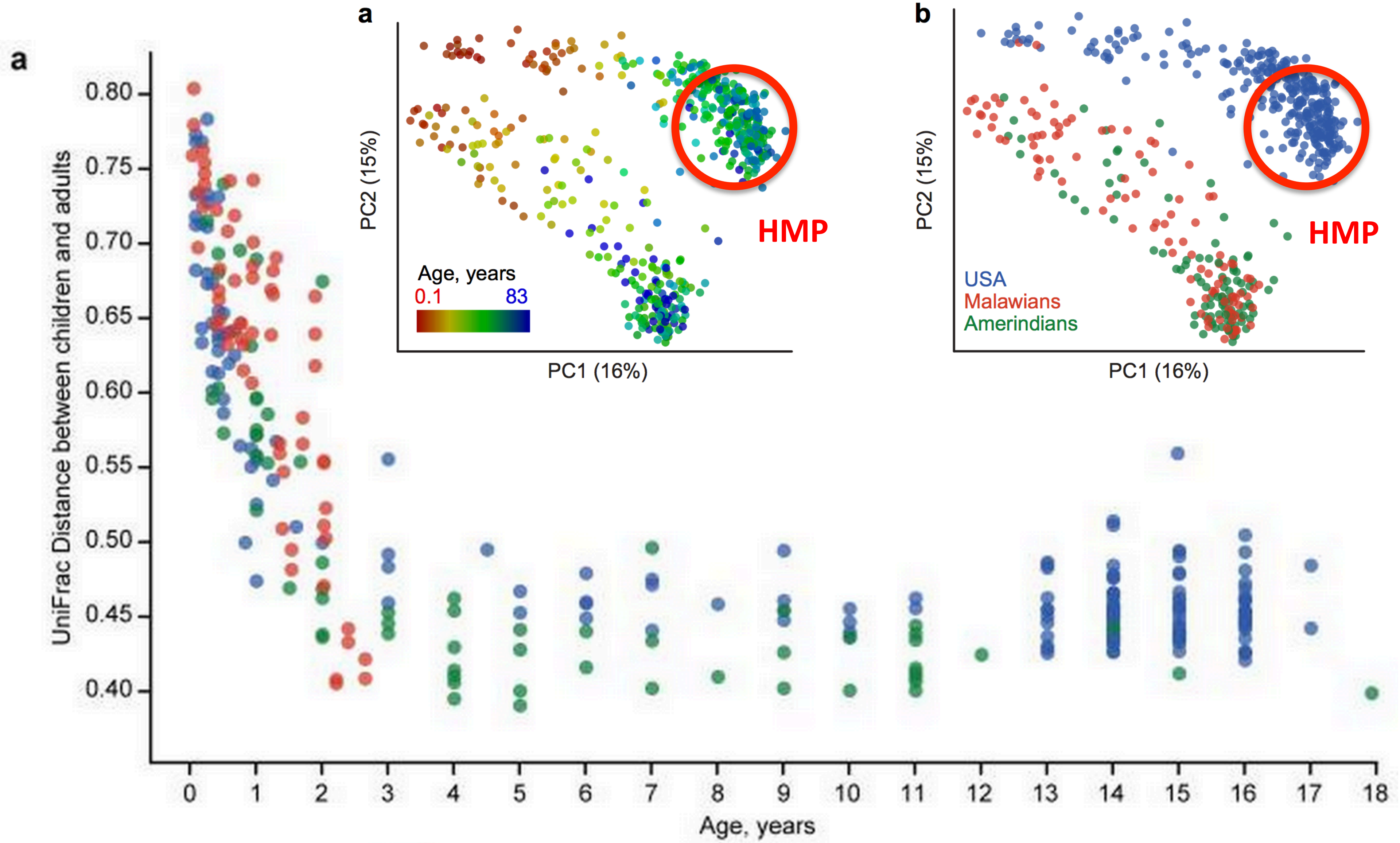
Tanya Yatsunenko[1], Federico E. Rey[1], Mark J. Manary[2,3], Indi Trehan[2,4], Maria Gloria Dominguez-Bello[5], Monica Contreras[6], Magda Magris[7], Glida Hidalgo[7], Robert N. Baldassano[8], Andrey P. Anokhin[9], Andrew C. Heath[9], Barbara Warner[2], Jens Reeder[10], Justin Kuczynski[10], J. Gregory Caporaso[11], Catherine A. Lozupone[10], Christian Lauber[10], Jose Carlos Clemente[10], Dan Knights[10], Rob Knight[10,12] & Jeffrey I. Gordon[1]



Yatsunenko et al. Nature. 2012 May 9;486(7402):222-7

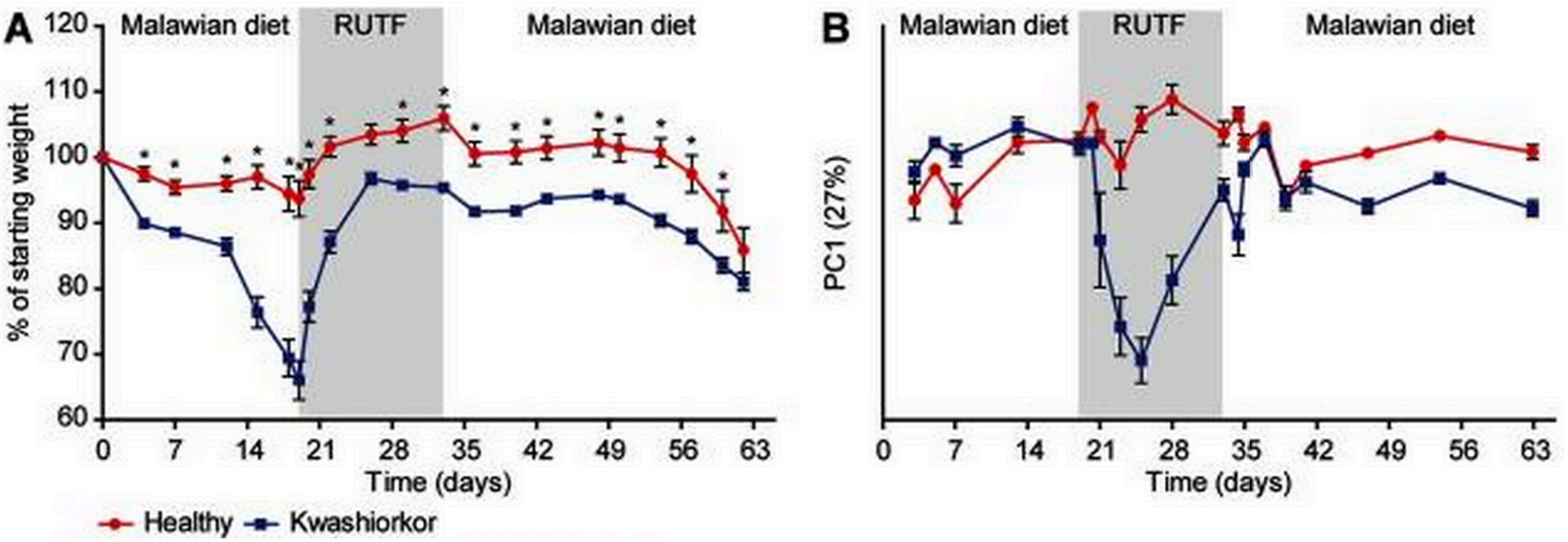


Yatsunenko et al. Nature. 2012 May 9;486(7402):222-7

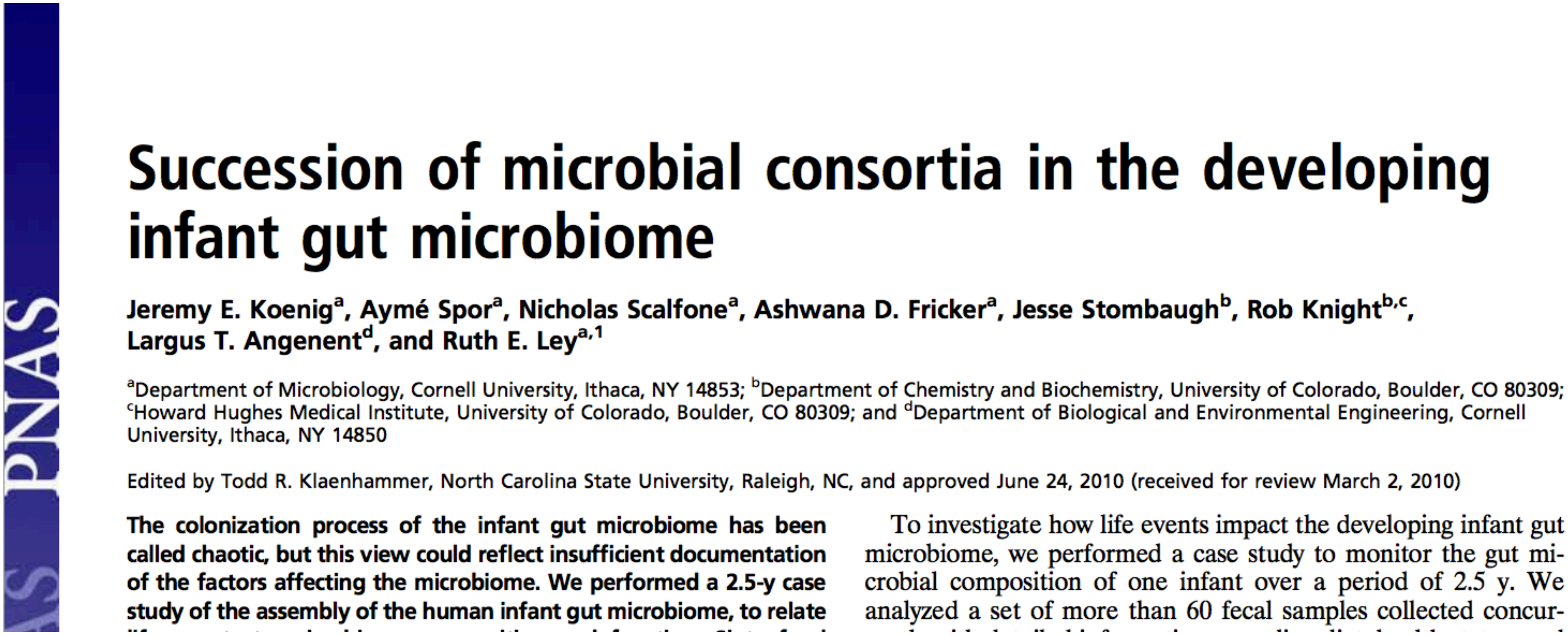


Yatsunenko et al. Nature. 2012 May 9;486(7402):222-7

Gut Microbiomes of Malawian Twin Pairs Discordant for Kwashiorkor
Michelle I. Smith *et al.*
Science **339**, 548 (2013);
DOI: 10.1126/science.1229000



Smith et al. Science. 2013 Feb 1;339(6119):548-54



Succession of microbial consortia in the developing infant gut microbiome

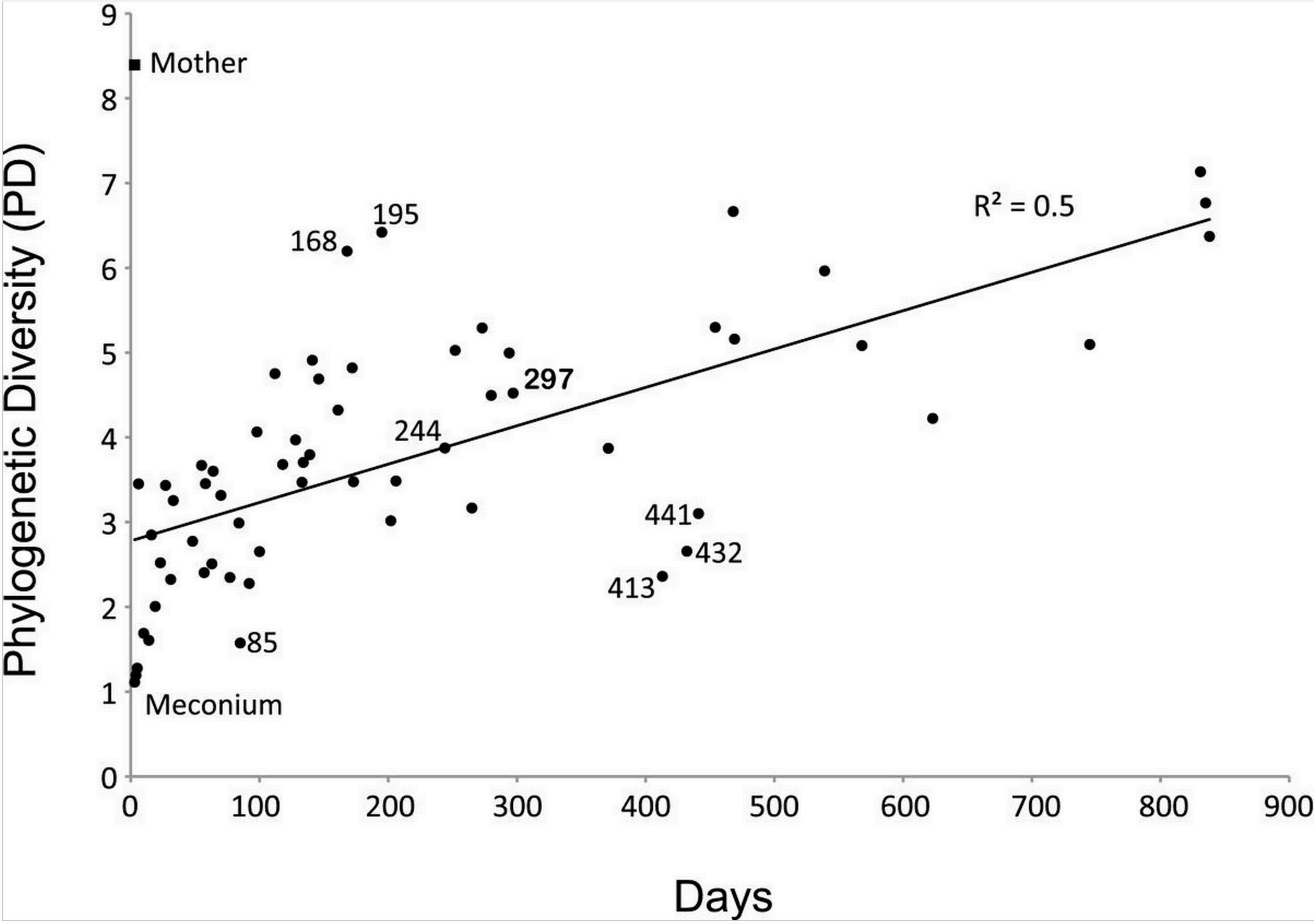
Jeremy E. Koenig[a], Aymé Spor[a], Nicholas Scalfone[a], Ashwana D. Fricker[a], Jesse Stombaugh[b], Rob Knight[b,c], Largus T. Angenent[d], and Ruth E. Ley[a,1]

[a]Department of Microbiology, Cornell University, Ithaca, NY 14853; [b]Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO 80309; [c]Howard Hughes Medical Institute, University of Colorado, Boulder, CO 80309; and [d]Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY 14850

Edited by Todd R. Klaenhammer, North Carolina State University, Raleigh, NC, and approved June 24, 2010 (received for review March 2, 2010)

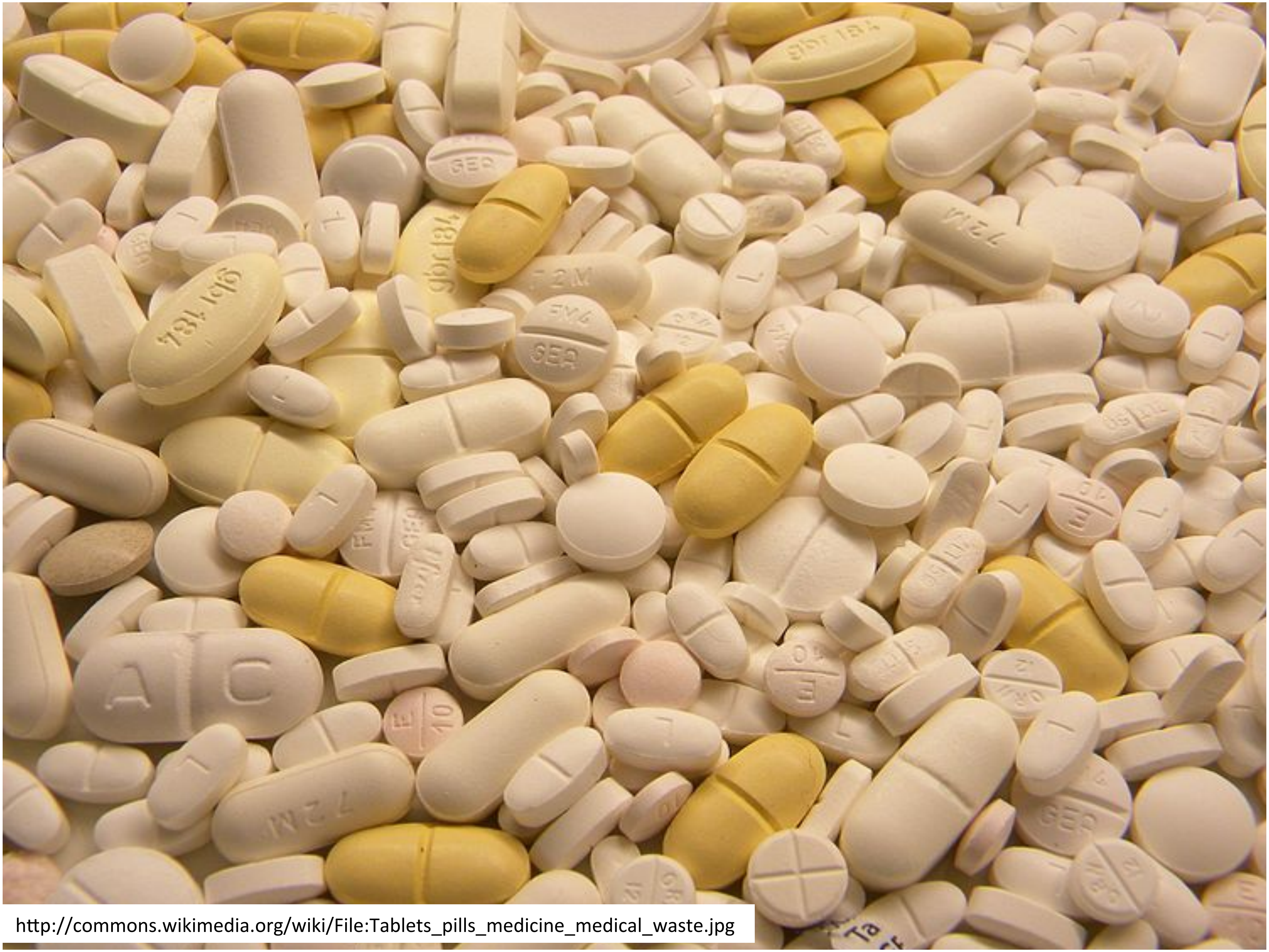
The colonization process of the infant gut microbiome has been called chaotic, but this view could reflect insufficient documentation of the factors affecting the microbiome. We performed a 2.5-y case study of the assembly of the human infant gut microbiome, to relate

To investigate how life events impact the developing infant gut microbiome, we performed a case study to monitor the gut microbial composition of one infant over a period of 2.5 y. We analyzed a set of more than 60 fecal samples collected concur-



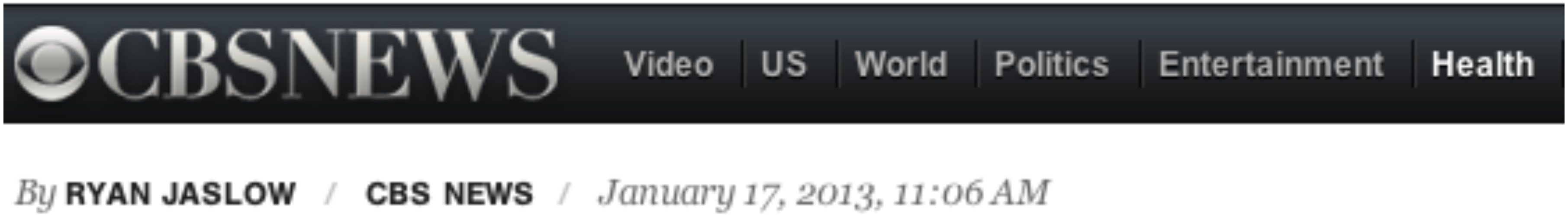
Koeing et al. Proc Natl Acad Sci U S A. 2011 Mar 15;108

Video fun time!


http://commons.wikimedia.org/wiki/File:Tablets_pills_medicine_medical_waste.jpg

By **RYAN JASLOW** / **CBS NEWS** / *January 17, 2013, 11:06 AM*

Fecal transplants beat antibiotics for curing diarrhea caused by C. difficile

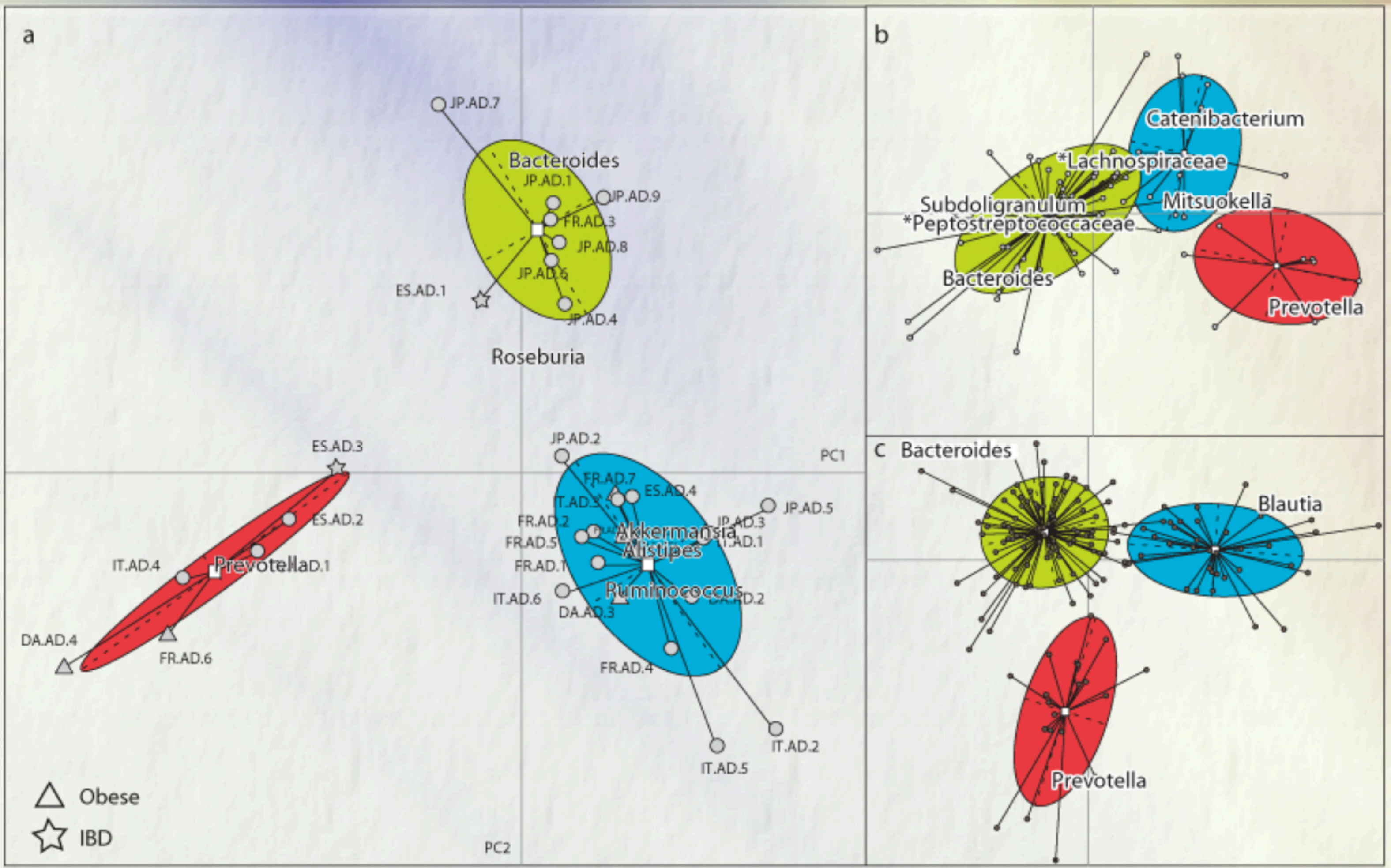


By **RYAN JASLOW** / **CBS NEWS** / *January 17, 2013, 11:06 AM*

Fecal transplants beat antibiotics for curing diarrhea caused by C. difficile

Video fun time!

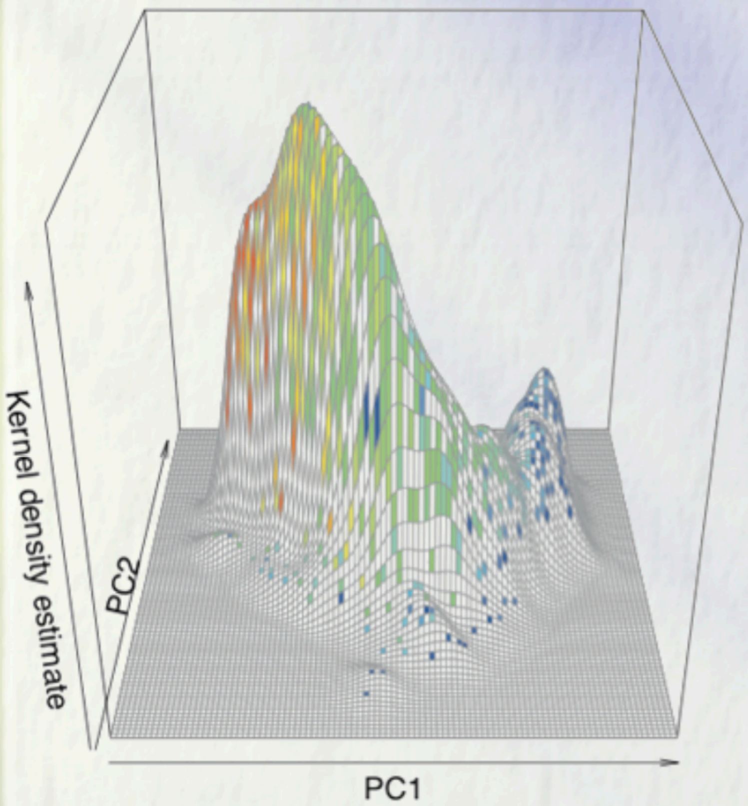
Are there enterotypes in the HMP data?



Arumugam et al. 2011 Nature

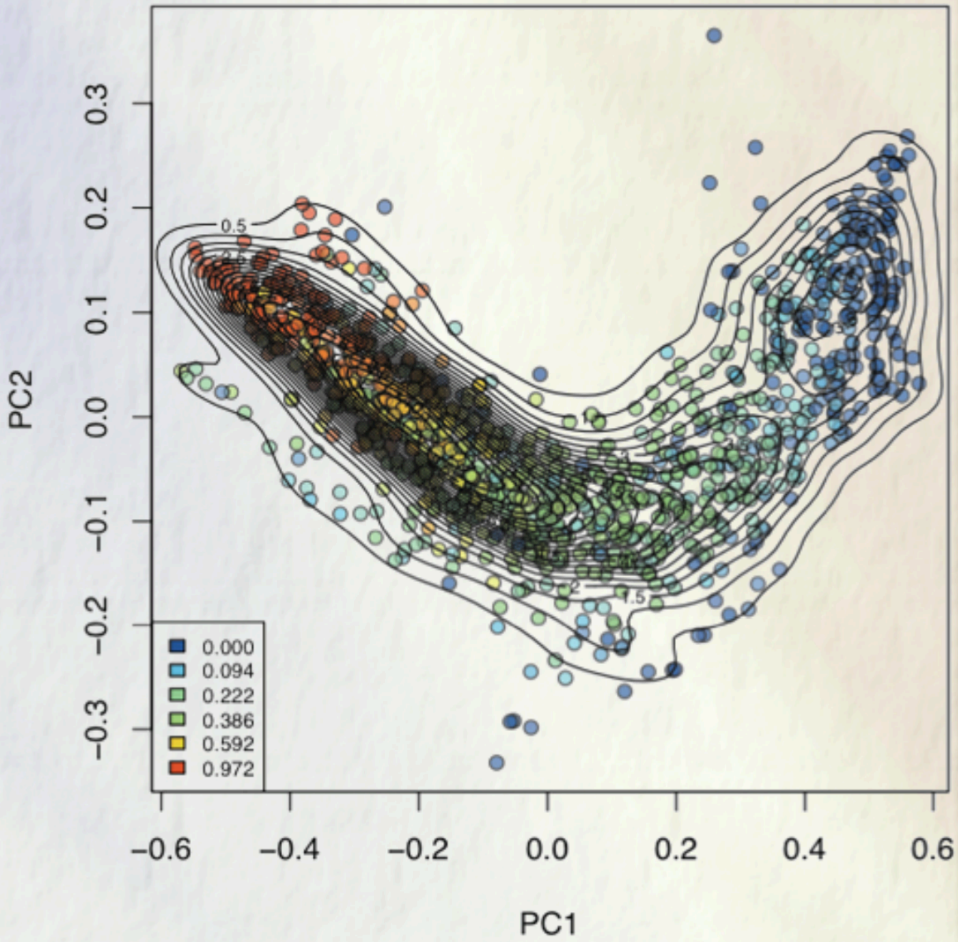
Slide adapted with permission from Rob Knight

No support for enterotypes in the HMP + community data...



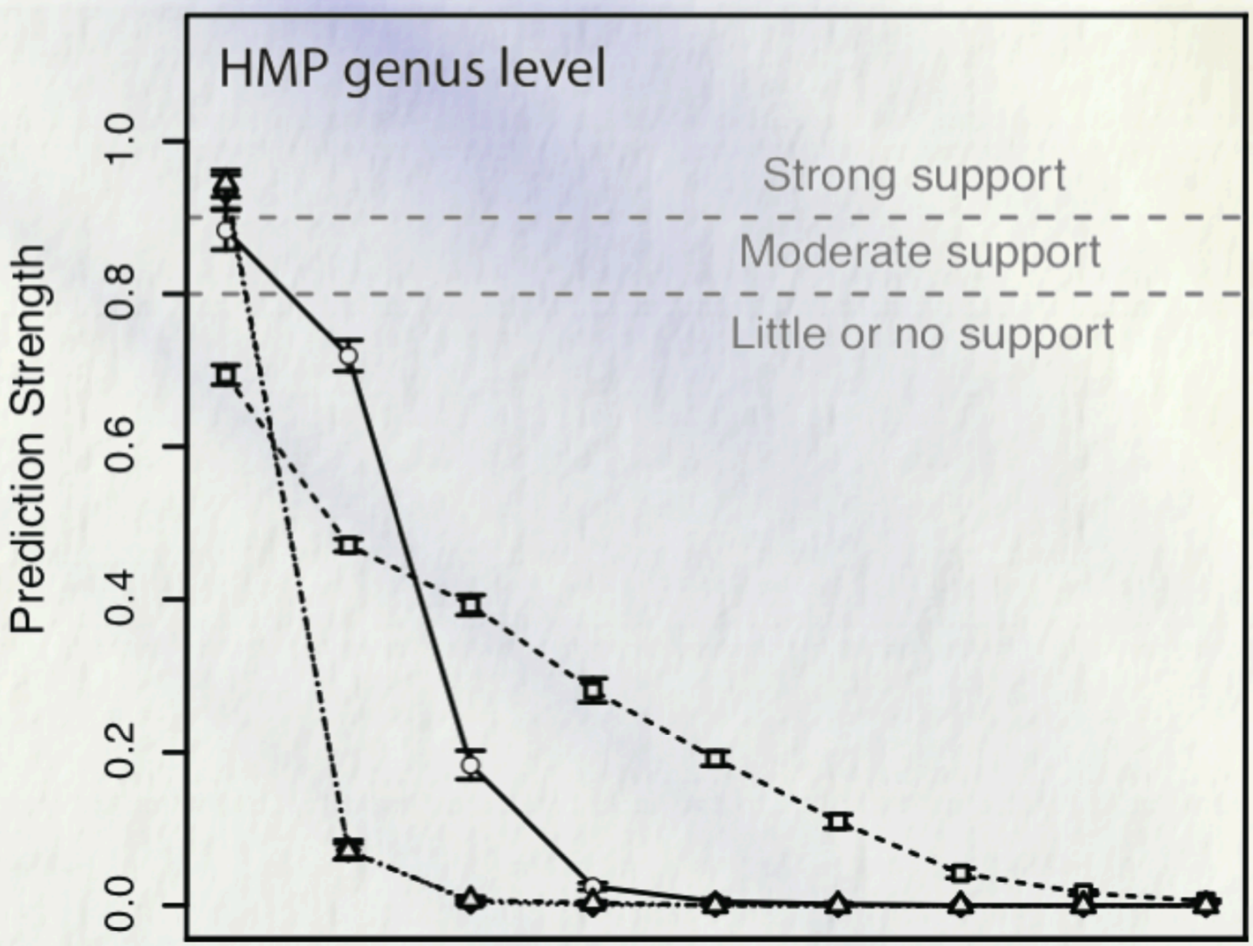
Koren et al. PLoS Com Biol 2013;9(1)

~~Koren et al. in revision PLoS Comp Bio~~



Slide adapted with permission from Rob Knight

...clustering weak, gradients explain better (ditto other sites)

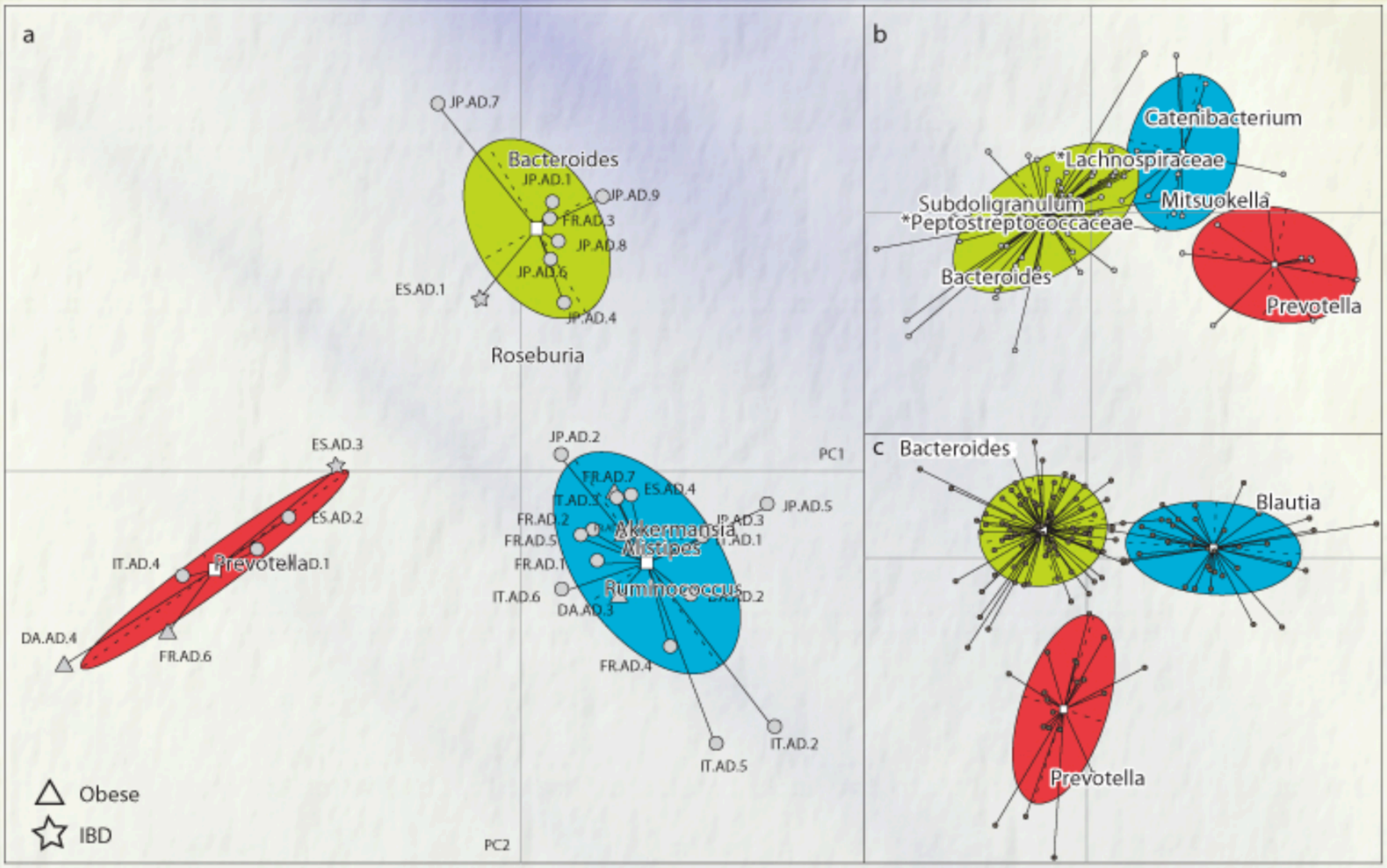


~~Koren et al. in revision PLoS Comp Bio~~

Koren et al. PLoS Com Biol 2013;9(1)

Slide adapted with permission from Rob Knight

So let's re-examine the original evidence for enterotypes

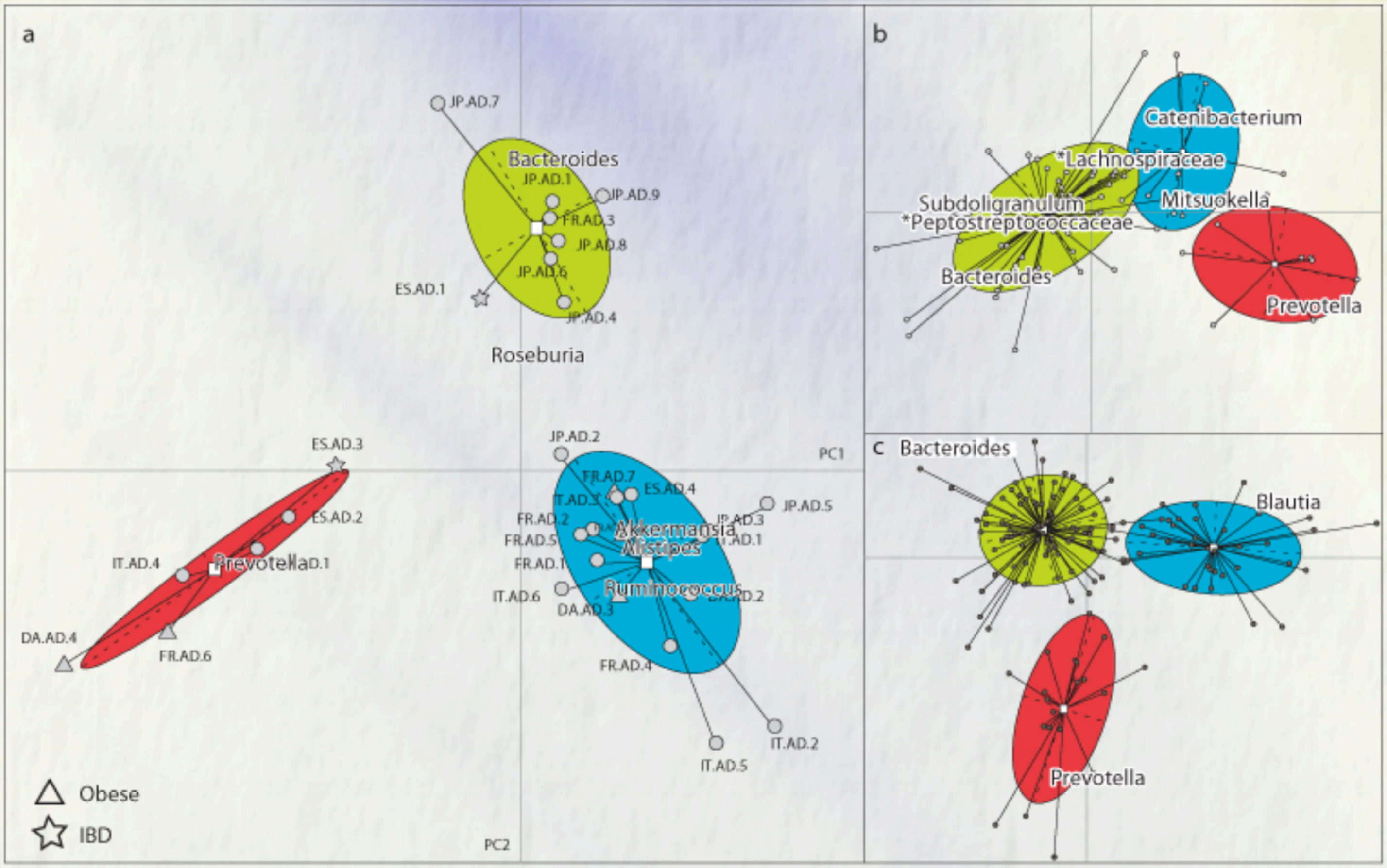


Koren et al. PLoS Com Biol 2013;9(1)

~~Koren et al. in revision PLoS Comp Bio~~

Slide adapted with permission from Rob Knight

So let's re-examine the original evidence for enterotypes

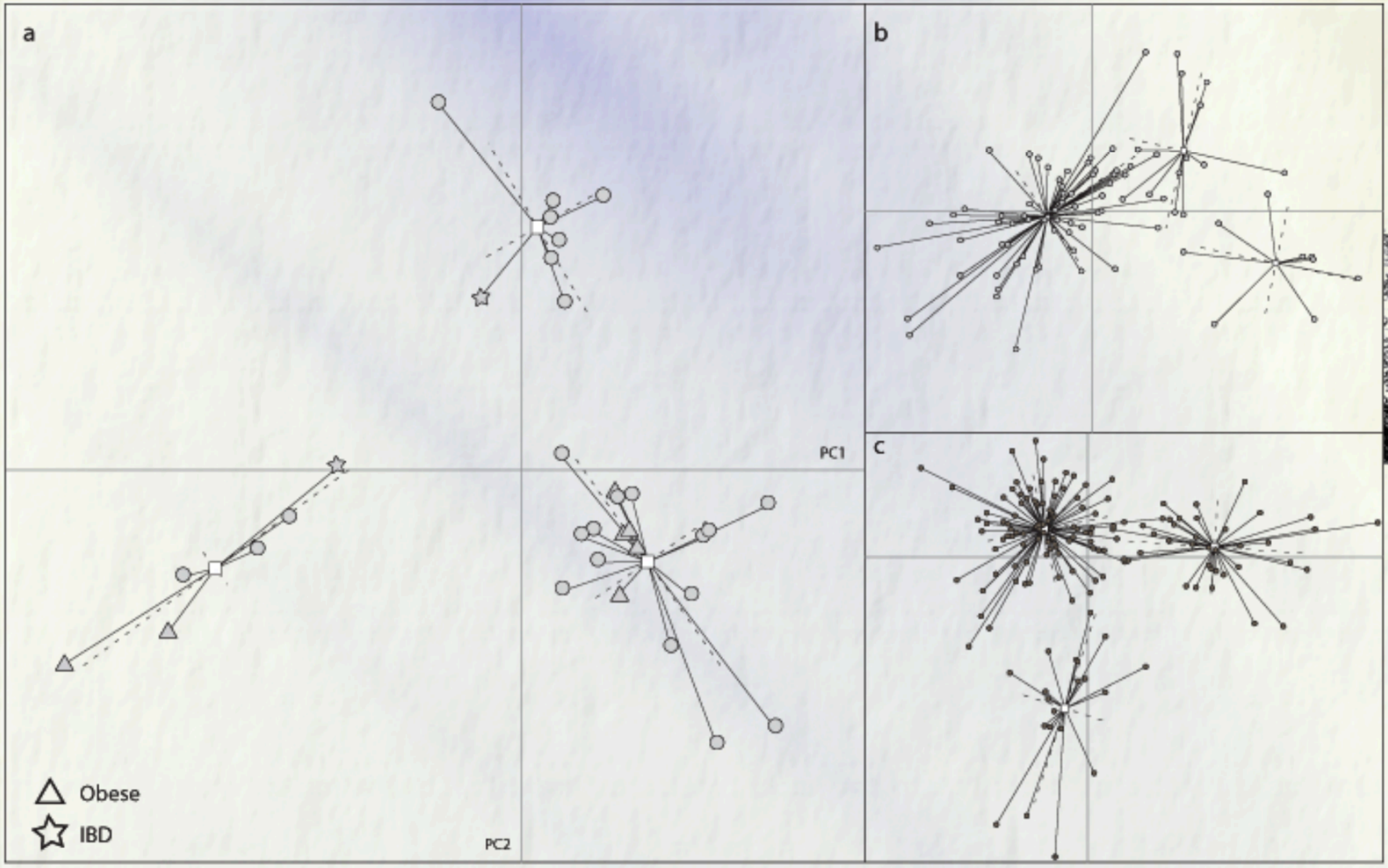


Koren et al. PLoS Com Biol 2013;9(1)

~~Koren et al. in revision PLoS Comp Bio~~

Slide adapted with permission from Rob Knight

So let's re-examine the original evidence for enterotypes

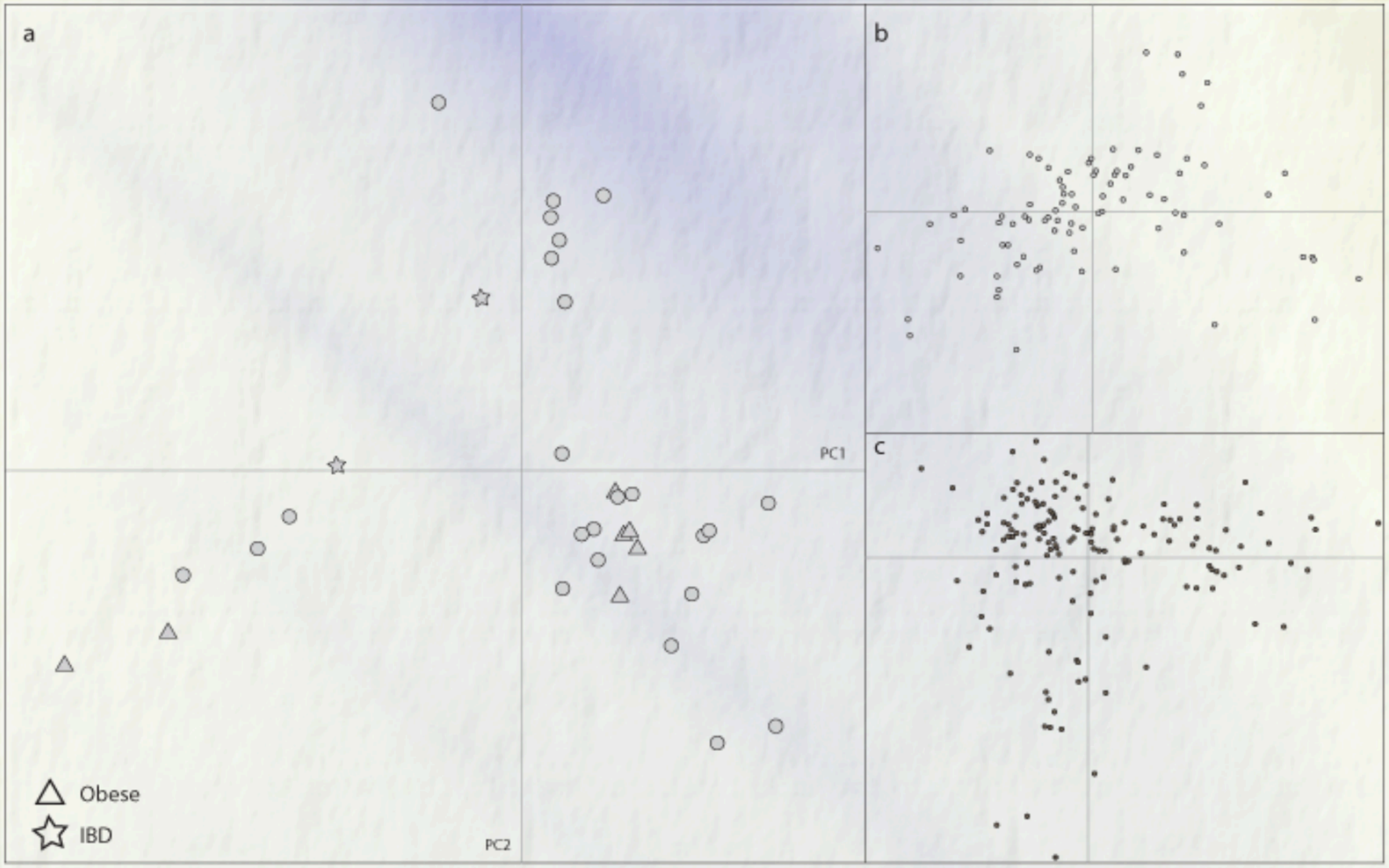


Koren et al. PLoS Com Biol 2013;9(1)

~~Koren et al. in revision PLoS Comp Bio~~

Slide adapted with permission from Rob Knight

So let's re-examine the original evidence for enterotypes

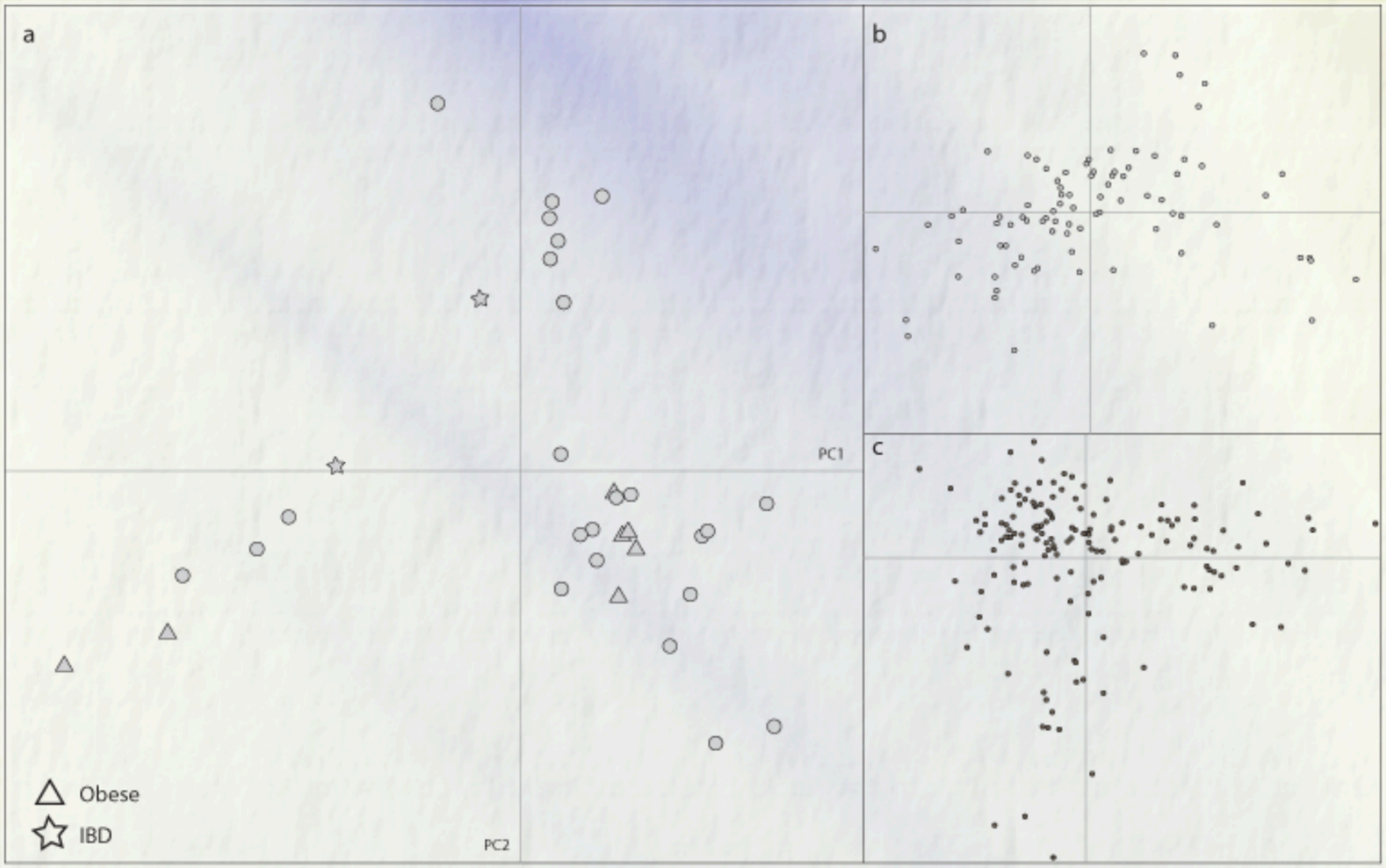


Koren et al. PLoS Com Biol 2013;9(1)

~~Koren et al. in revision PLoS Comp Bio~~

Slide adapted with permission from Rob Knight

So let's re-examine the original evidence for enterotypes



Koren et al. PLoS Com Biol 2013;9(1)

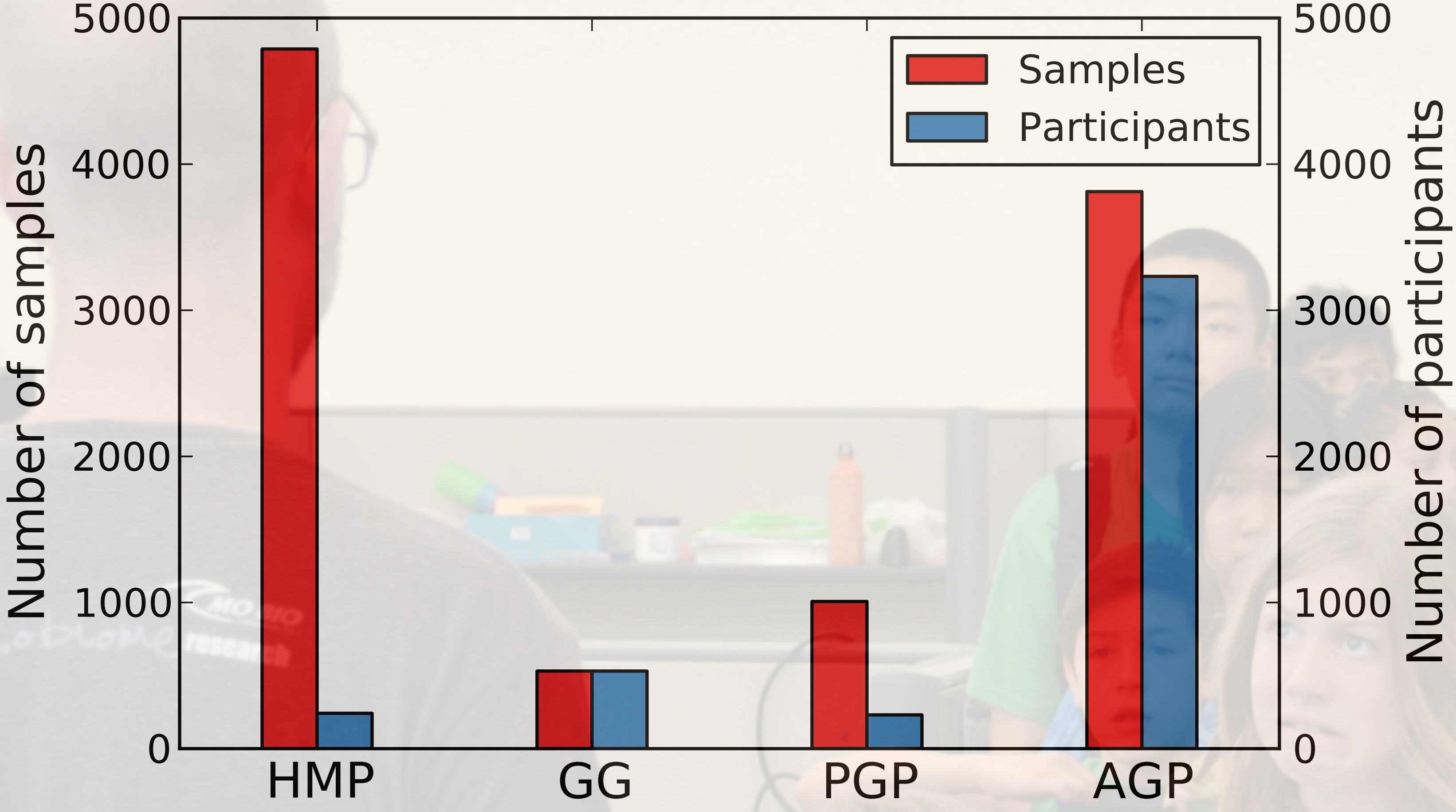
~~Koren et al. in revision PLoS Comp Bio~~

Slide adapted with permission from Rob Knight

Gilbert et al. BMC Biol. 2014 Aug 22;12(1):69

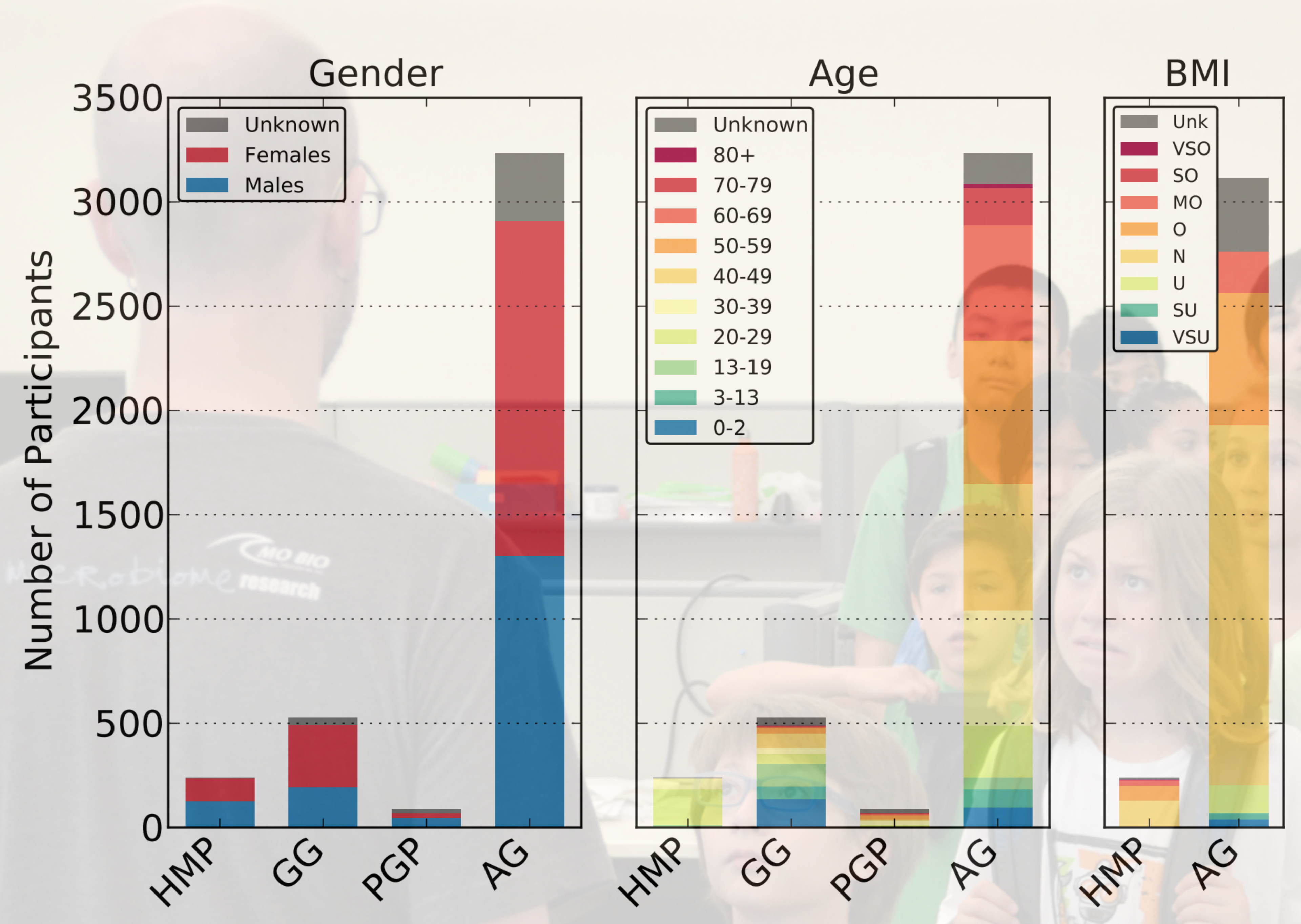

american
gut


MO BIO
microbiome research

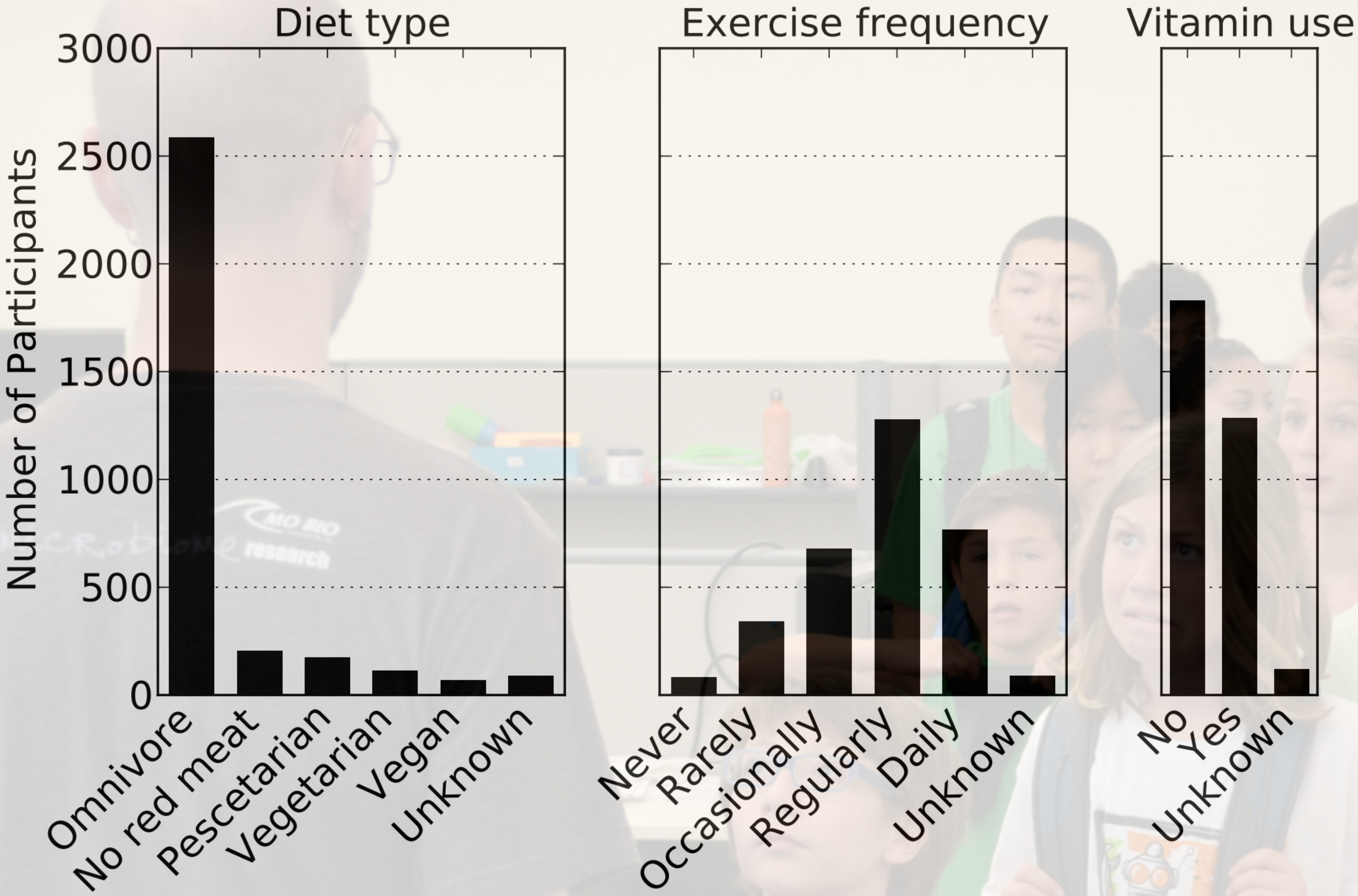


- HMP: Human Microbiome Project (2012, Nature 486: 202-214)
- GG: Global Gut (2012, Nature 486: 222-228)
- PGP: Personal Genome Project (unpublished)
- AGP: American Gut Project (unpublished)

http://microbio.me/americangut/img/mod1_main.pdf

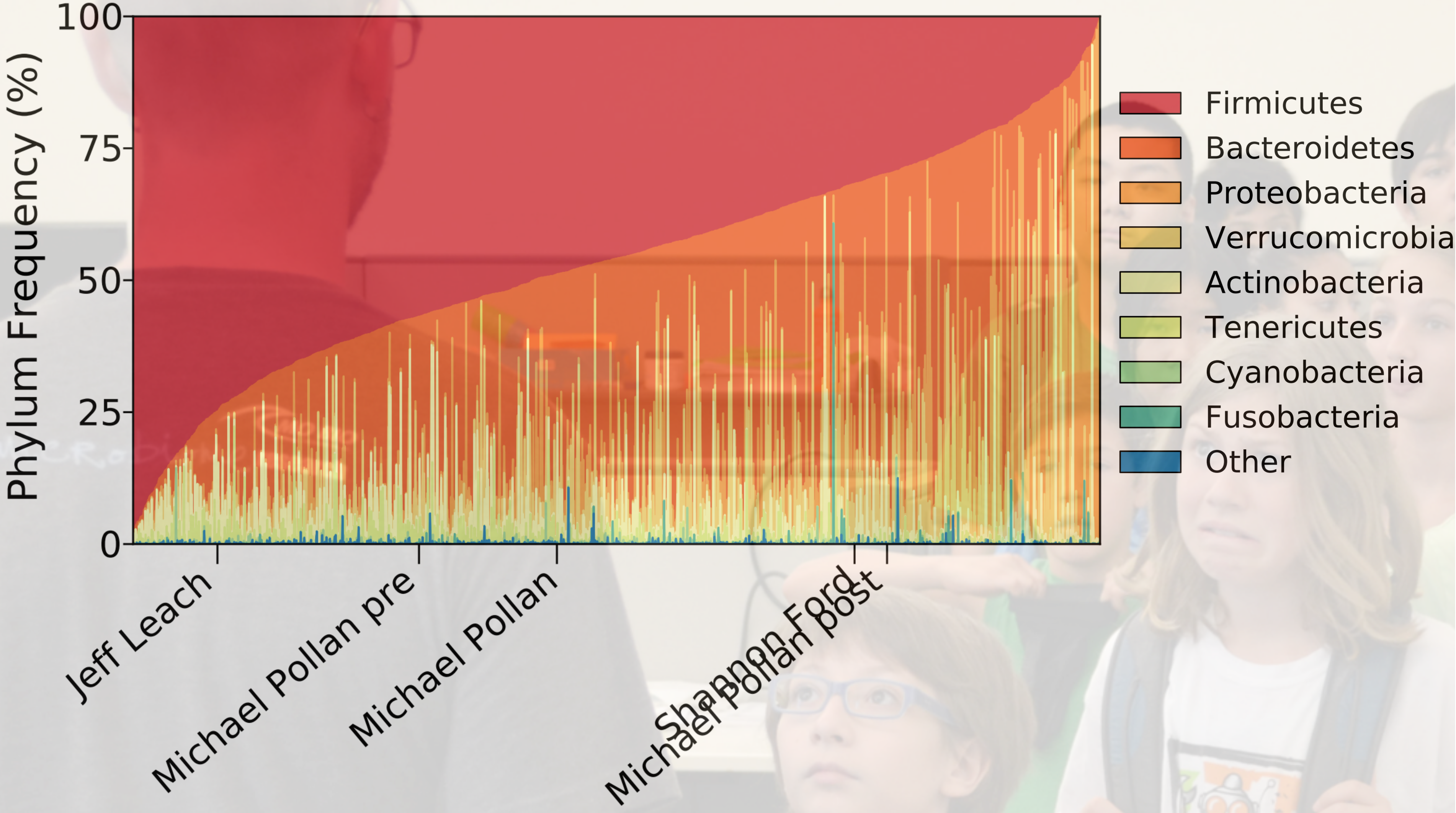


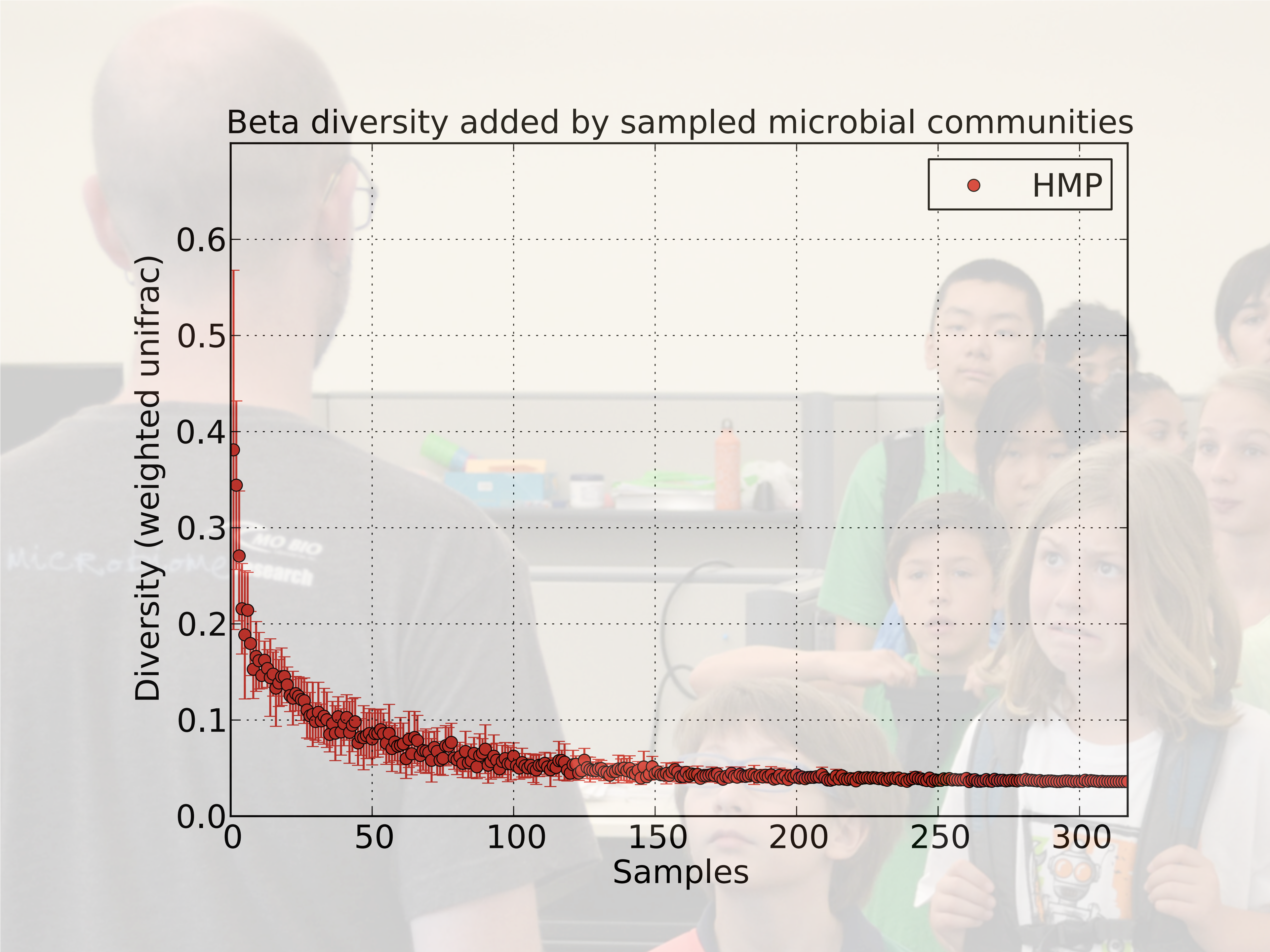
http://microbio.me/americangut/img/mod1_main.pdf

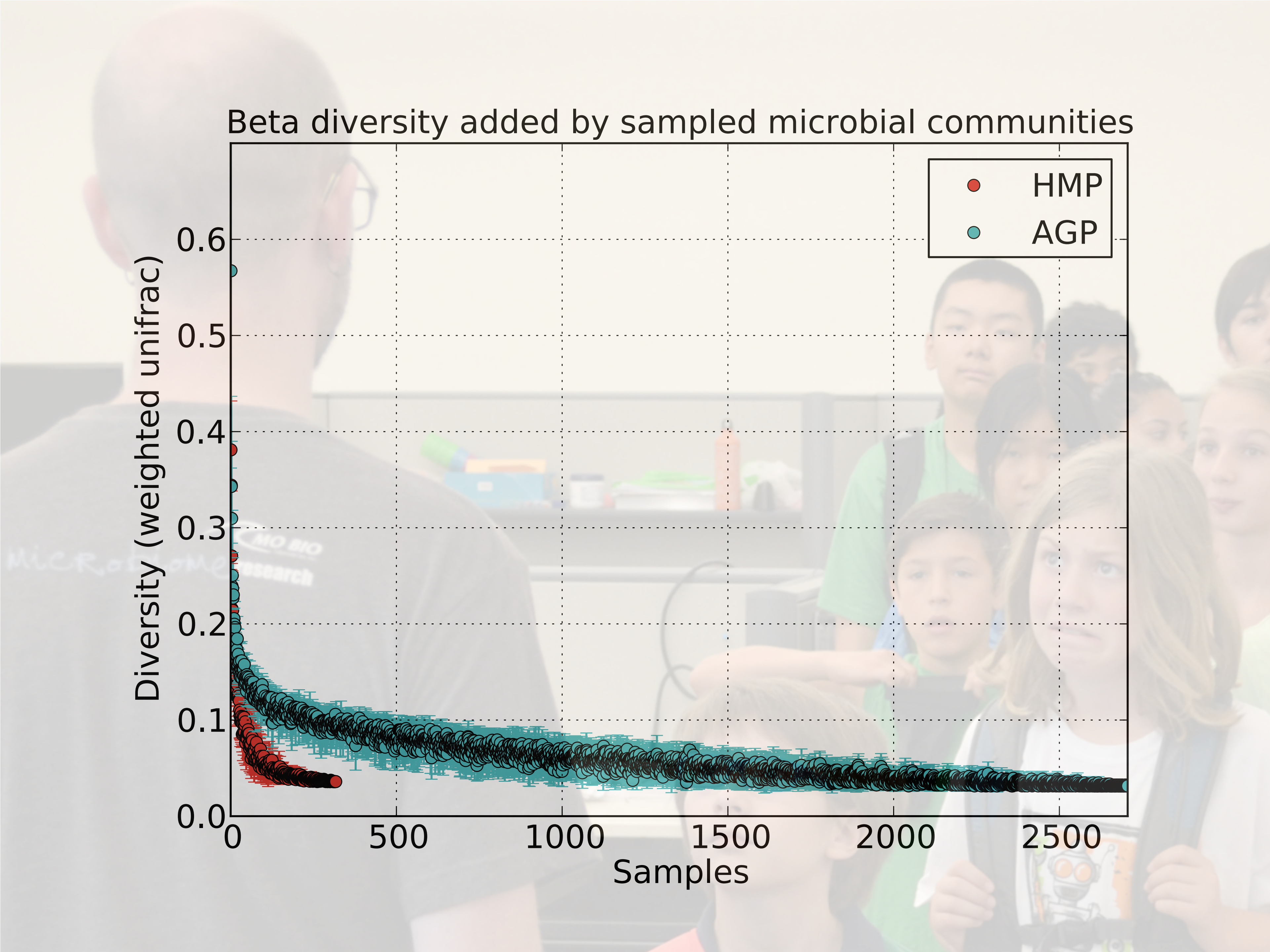


http://microbio.me/americangut/img/mod1_main.pdf

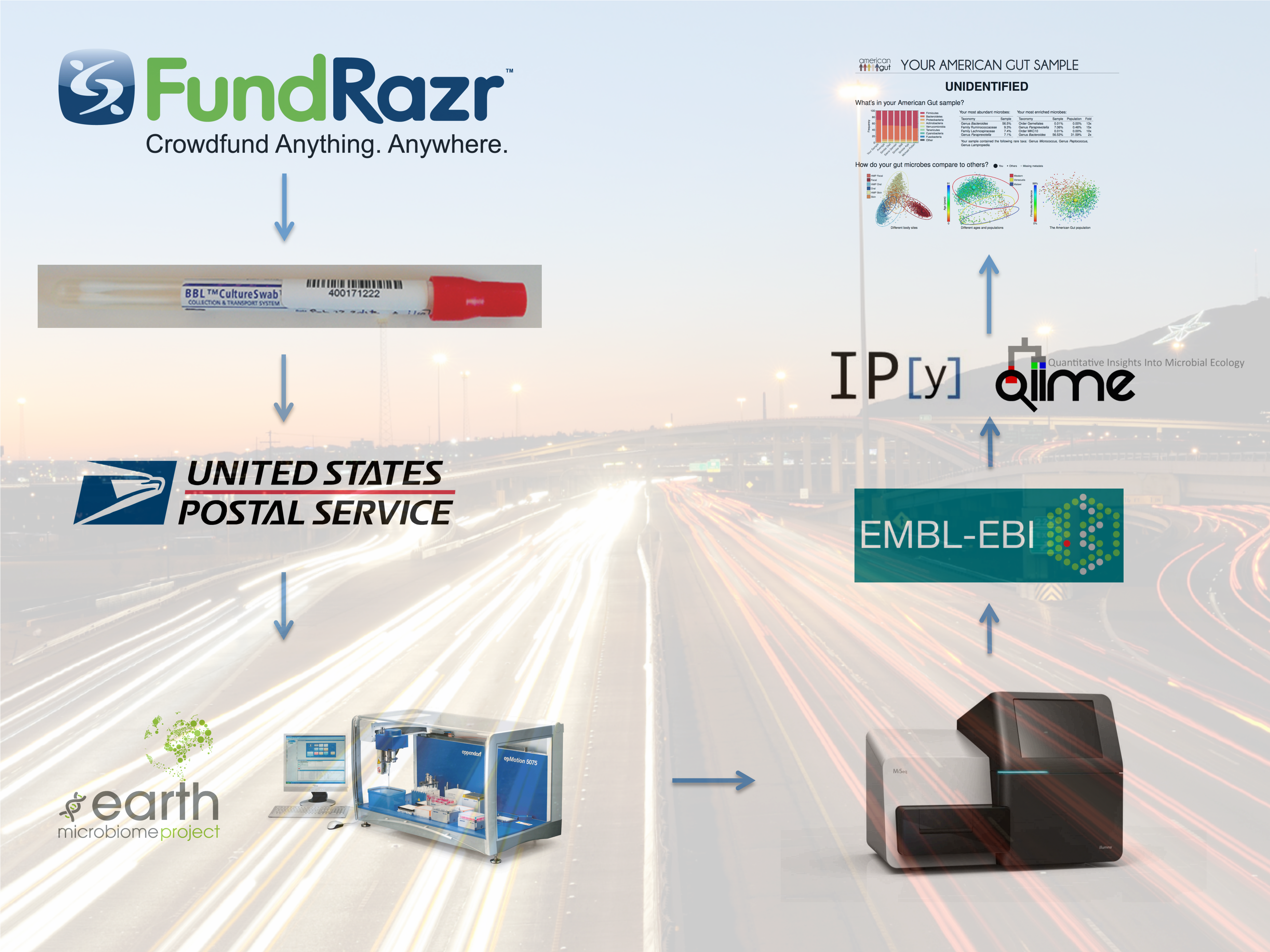
Fecal


Phylum Frequency (%)
100
75
50
25
0
Jeff Leach
Michael Pollan pre
Michael Pollan
Shannon Ford
Michael Pollan post
Firmicutes
Bacteroidetes
Proteobacteria
Verrucomicrobia
Actinobacteria
Tenericutes
Cyanobacteria
Fusobacteria
Other


Beta diversity added by sampled microbial communities
HMP
Diversity (weighted unifrac)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Samples
0
50
100
150
200
250
300


Beta diversity added by sampled microbial communities
HMP
AGP
Diversity (weighted unifrac)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Samples
0
500
1000
1500
2000
2500


22B
EXIT
16 FT 1 IN



american gut YOUR AMERICAN GUT SAMPLE

UNIDENTIFIED

What's in your American Gut sample?

Your most abundant microbes:

Taxonomy	Sample
Genus *Bacteroides*	56.5%
Family Ruminococcaceae	9.3%
Family Lachnospiraceae	7.4%
Genus *Paraprevotella*	7.1%

Your most enriched microbes:

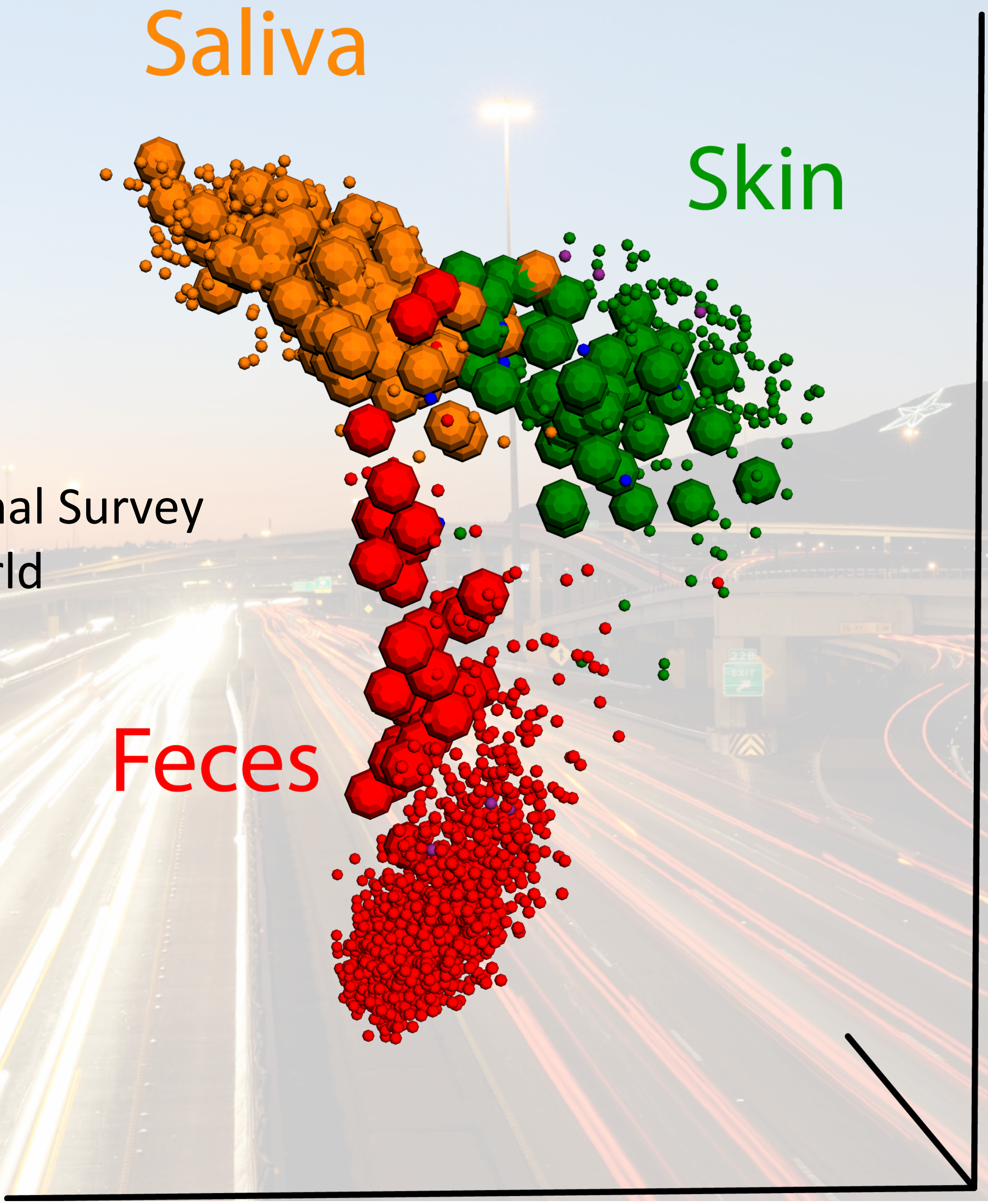
Taxonomy	Sample	Population	Fold
Order Gemellales	0.01%	0.00%	13x
Genus *Paraprevotella*	7.06%	0.46%	15x
Order MKC10	0.01%	0.00%	10x
Genus *Bacteroides*	56.53%	31.59%	2x

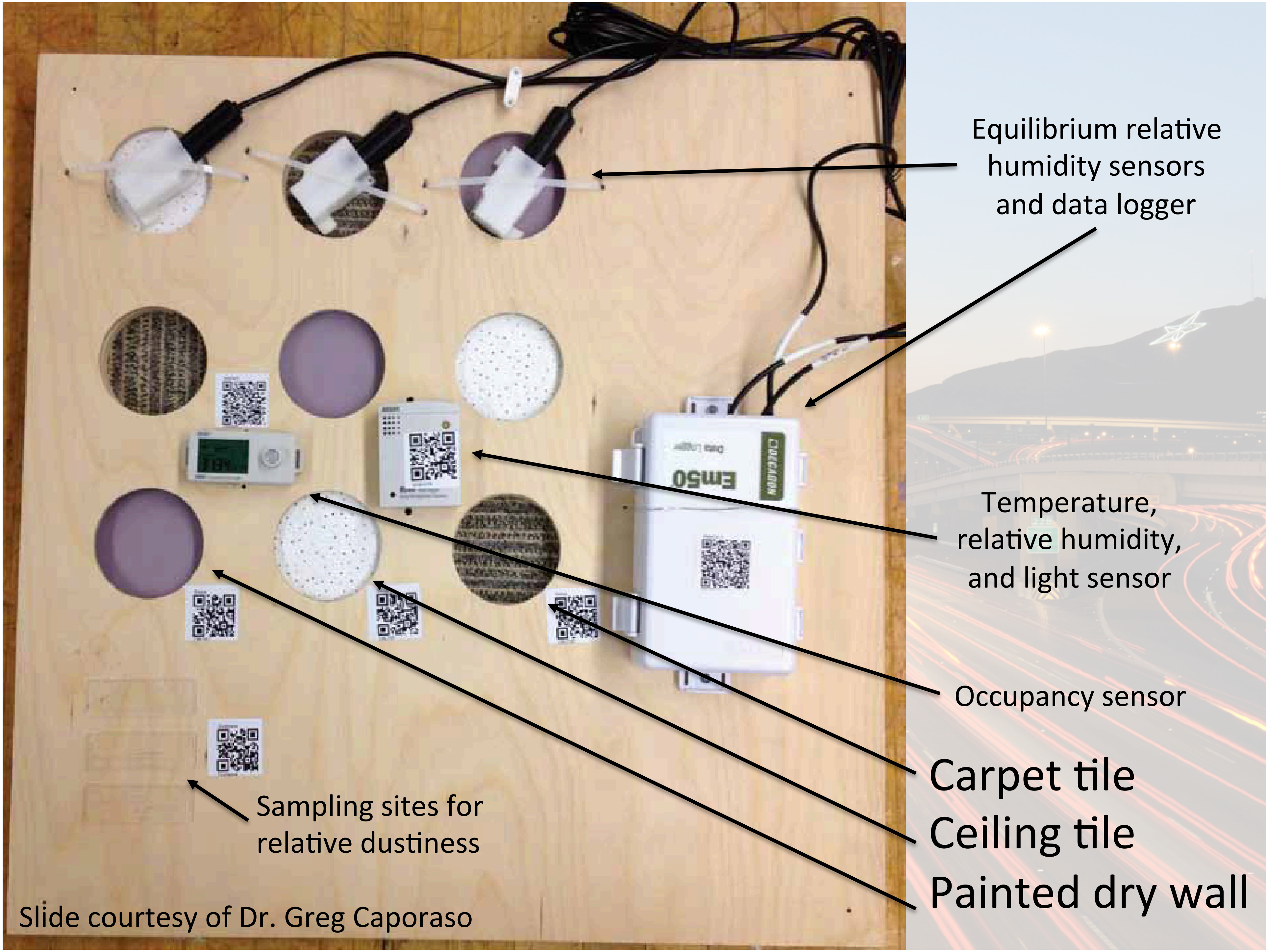
Your sample contained the following rare taxa: Genus *Micrococcus*, Genus *Peptococcus*, Genus *Lampropedia*.

How do your gut microbes compare to others?

ICU Microbiome pilot

- International Nutritional Survey
- 5 sites around the world
- Multiple time points
- Rich patient metadata





Equilibrium relative humidity sensors and data logger

Temperature, relative humidity, and light sensor

Occupancy sensor

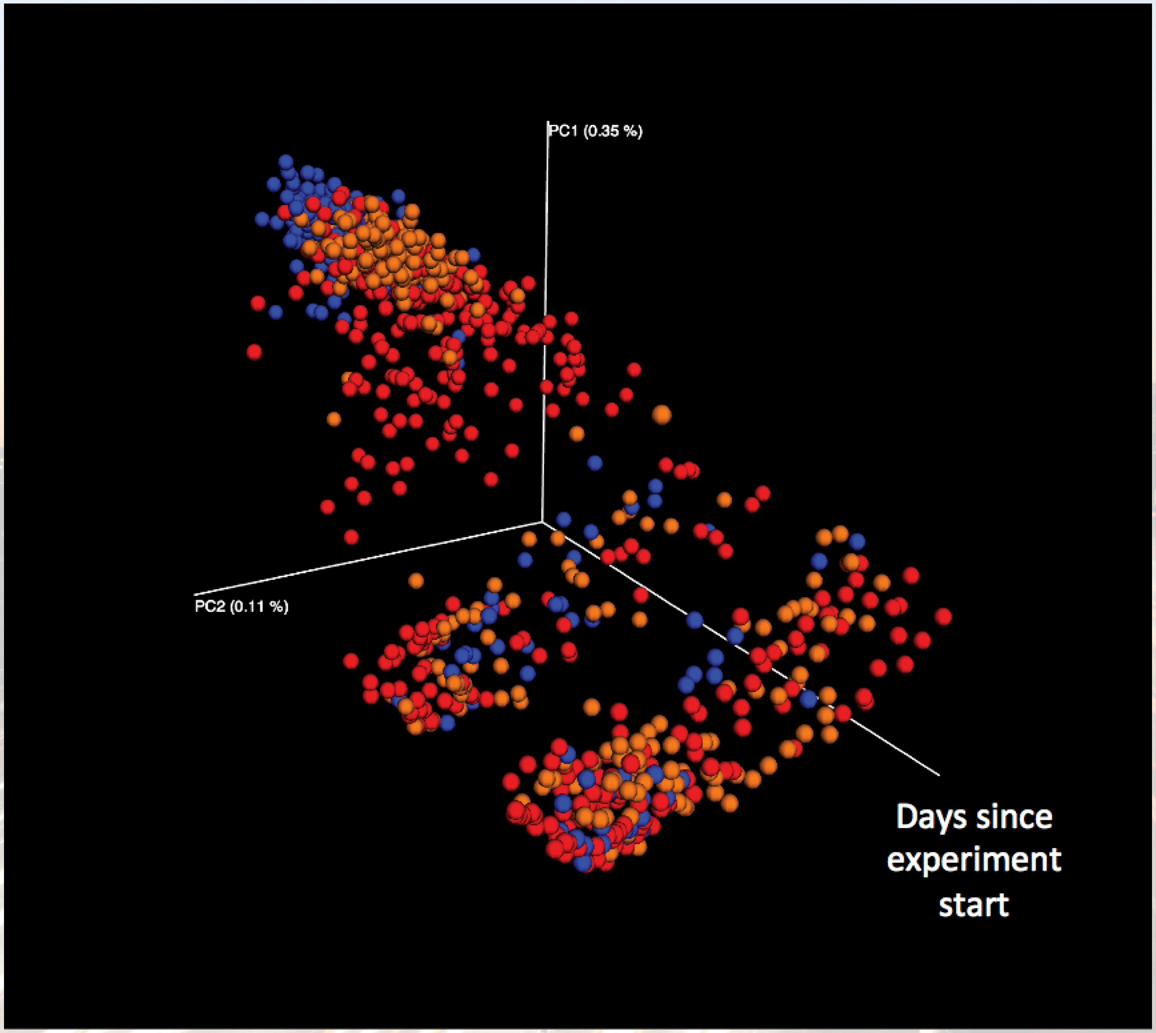
Carpet tile

Ceiling tile

Painted dry wall

Sampling sites for relative dustiness

Slide courtesy of Dr. Greg Caporaso

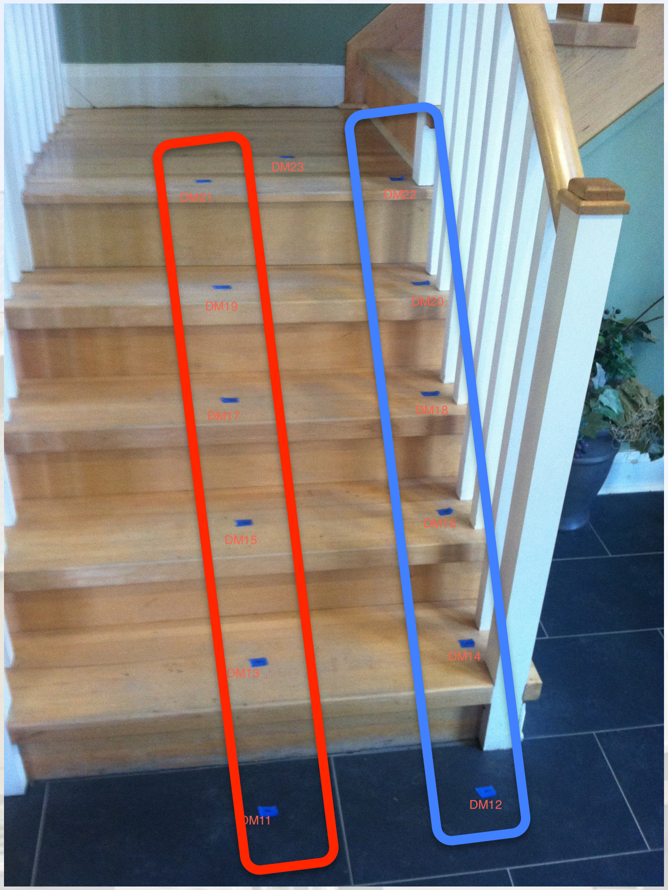


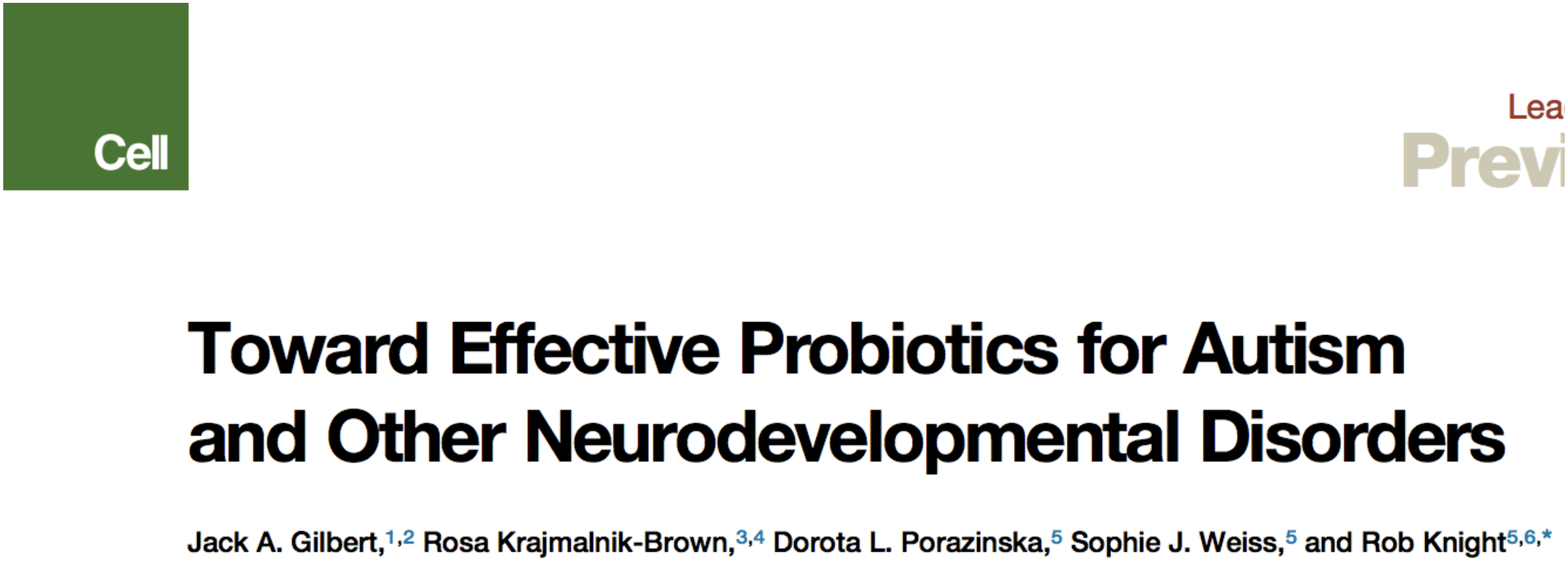
Flagstaff begins to look different, but “returns” to the other cities.

There doesn’t appear to be a city-specific composition effect.

🟦 San Diego
🟥 Flagstaff
🟧 Toronto

Slide courtesy of Dr. Greg Caporaso


DM23
DM21
DM22
DM19
DM20
DM17
DM18
DM15
DM16
DM13
DM14
DM11
DM12


Cell

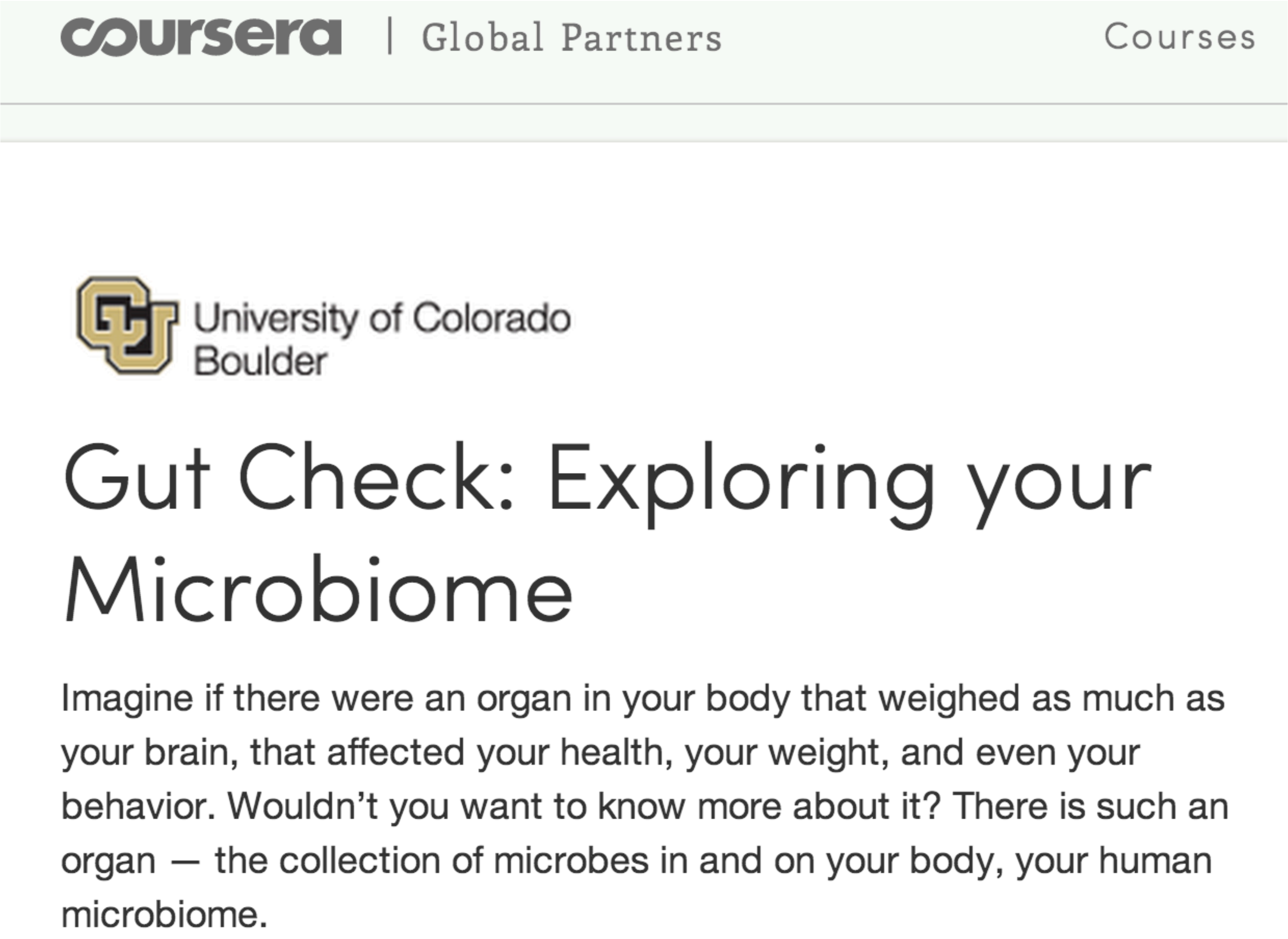
Lea

Previ

Toward Effective Probiotics for Autism and Other Neurodevelopmental Disorders

Jack A. Gilbert,[1,2] Rosa Krajmalnik-Brown,[3,4] Dorota L. Porazinska,[5] Sophie J. Weiss,[5] and Rob Knight[5,6,*]

Autism spectrum cohort of the American Gut
- 269 families to receive samples
- Conjunction with ASD-clinicians
- Donations for ASD-affected and neurotypical siblings covered by Jack Gilbert



University of Colorado Boulder

Gut Check: Exploring your Microbiome

Imagine if there were an organ in your body that weighed as much as your brain, that affected your health, your weight, and even your behavior. Wouldn't you want to know more about it? There is such an organ — the collection of microbes in and on your body, your human microbiome.

Everything Comes Down to Poo

"Galileo saw outer space through his telescope, and we want to see the inner space of your gut through modern genetics" -Rob Dunn

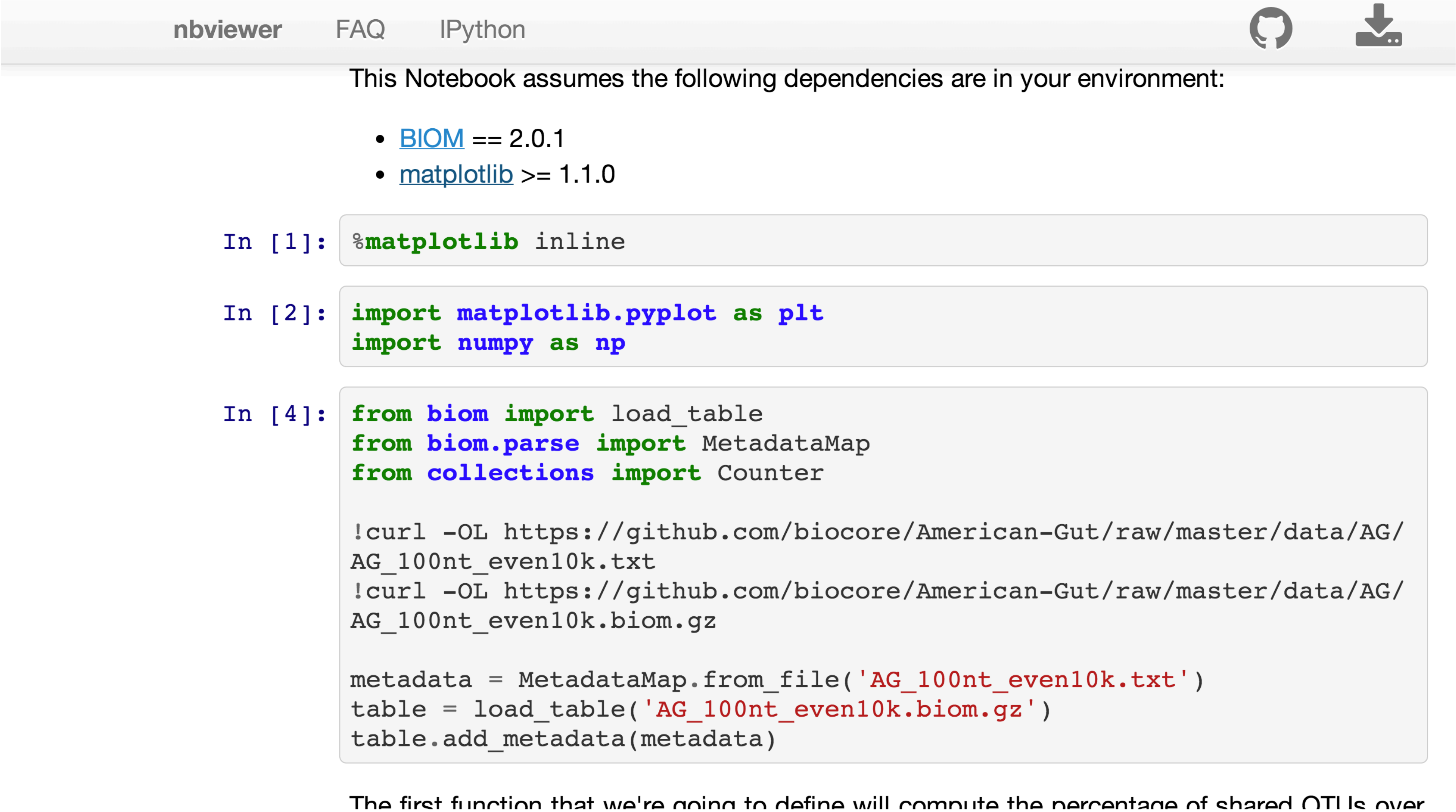
OPEN DATA

OPEN ACCESS

*

open source TM

- Analyses are executable, reproducible, and BSD-licensed



This Notebook assumes the following dependencies are in your environment:

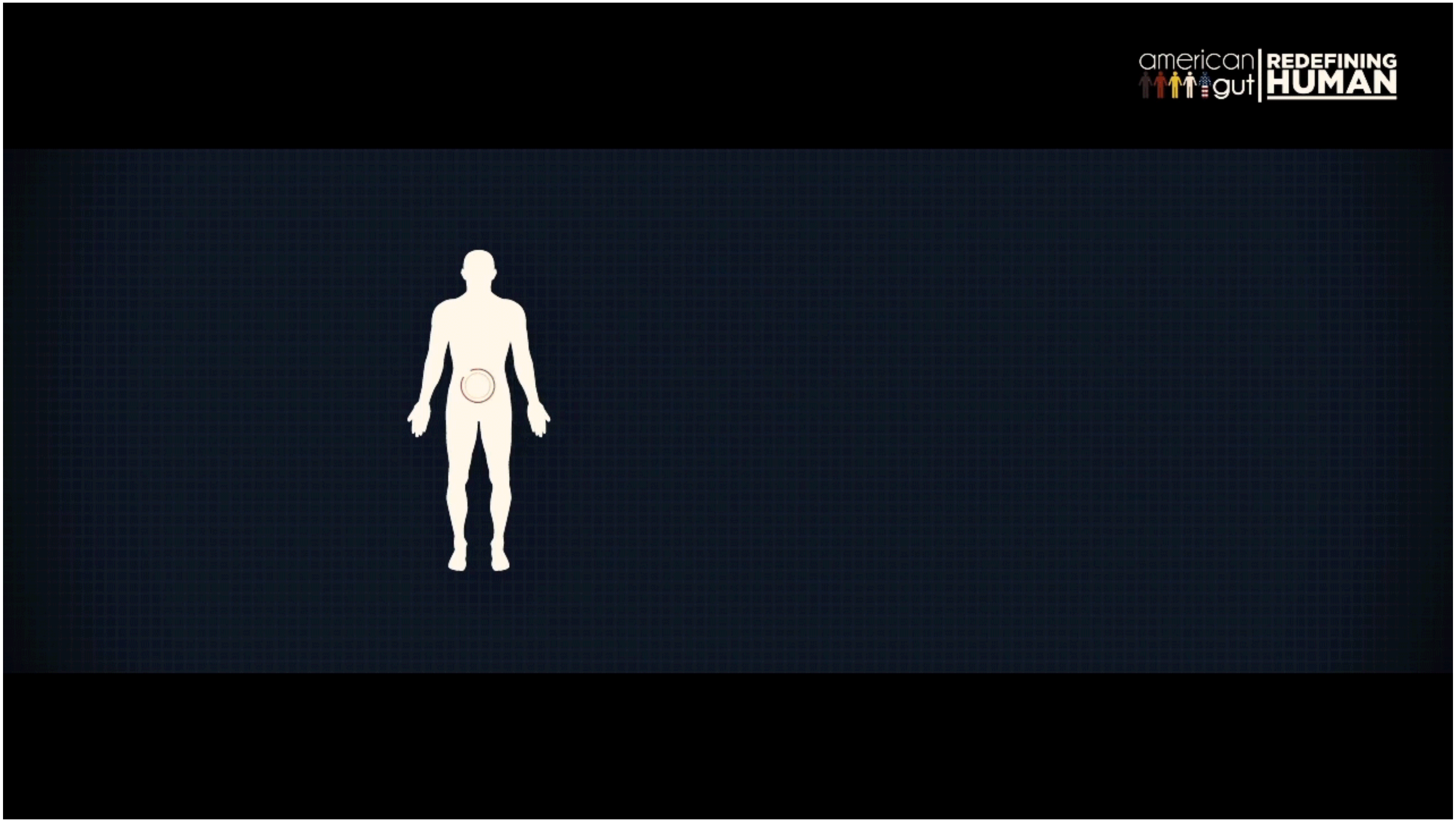
- BIOM == 2.0.1
- matplotlib >= 1.1.0

```
In [1]: %matplotlib inline
```

```
In [2]: import matplotlib.pyplot as plt
        import numpy as np
```

```
In [4]: from biom import load_table
        from biom.parse import MetadataMap
        from collections import Counter

        !curl -OL https://github.com/biocore/American-Gut/raw/master/data/AG/
        AG_100nt_even10k.txt
        !curl -OL https://github.com/biocore/American-Gut/raw/master/data/AG/
        AG_100nt_even10k.biom.gz

        metadata = MetadataMap.from_file('AG_100nt_even10k.txt')
        table = load_table('AG_100nt_even10k.biom.gz')
        table.add_metadata(metadata)
```

The first function that we're going to define will compute the percentage of shared OTUs over



Adapted from http://redefininghuman.com
Please see https://fundrazr.com/campaigns/aoaif/

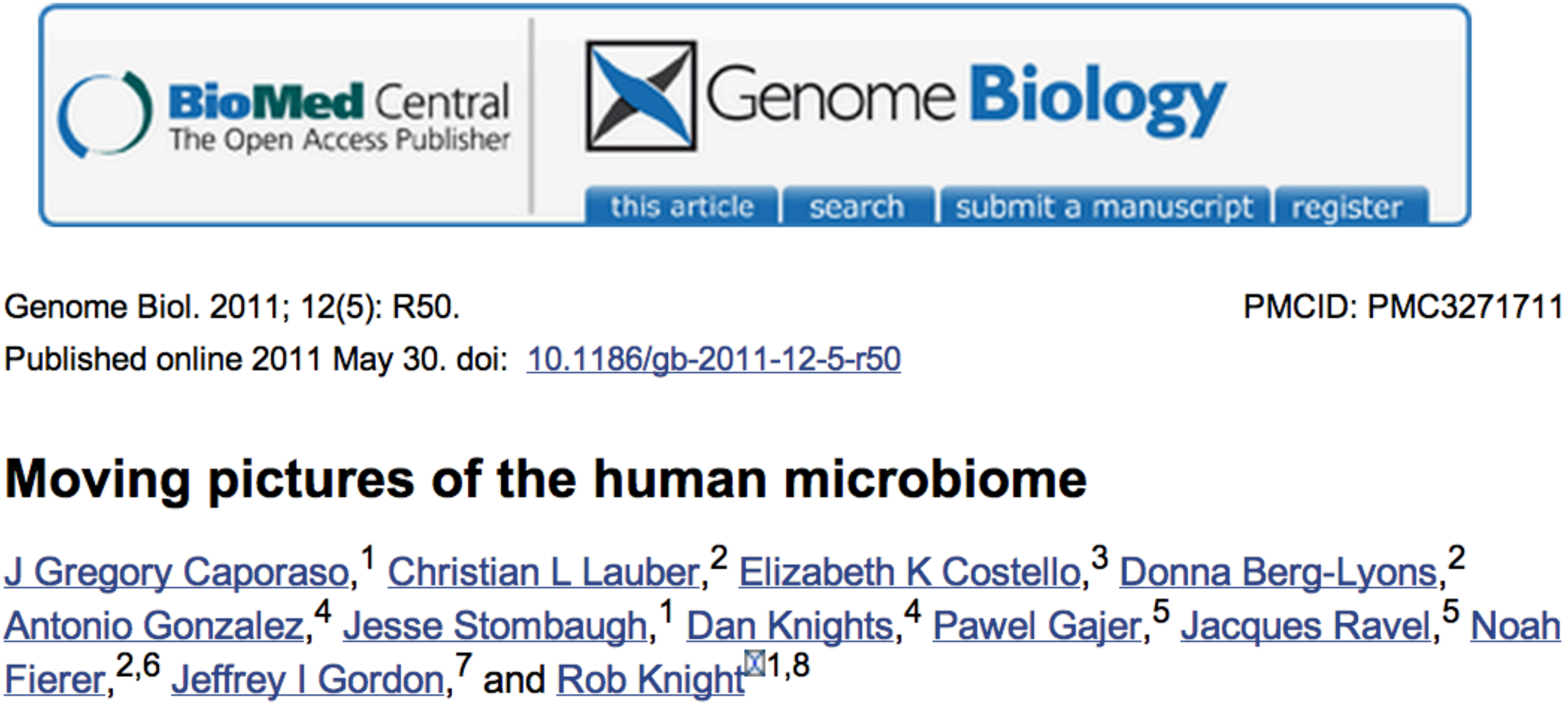
Running QIIME

- Native installation on Mac OS X or Linux
  - From laptops to compute clusters with thousands of cores
  - qiime-deploy
- Ubuntu Virtual Box
- Cloud-based installations

amazon
web services™
EC2

IP[y]: IPython
Interactive Computing

The IPython Notebook



Genome Biol. 2011; 12(5): R50. PMCID: PMC3271711

doi: 10.1186/gb-2011-12-5-r50

Moving pictures of the human microbiome

J Gregory Caporaso,[1] Christian L Lauber,[2] Elizabeth K Costello,[3] Donna Berg-Lyons,[2] Antonio Gonzalez,[4] Jesse Stombaugh,[1] Dan Knights,[4] Pawel Gajer,[5] Jacques Ravel,[5] Noah Fierer,[2,6] Jeffrey I Gordon,[7] and Rob Knight[1,8]

http://www.ncbi.nlm.nih.gov/pubmed/21624126

Moving Pictures of the Human Microbiome

- Two subjects sampled daily, one for six months, one for 18 months
- Four body sites: tongue, palm of left hand, palm of right hand, and gut (via fecal swabs).

Caporaso JG et al. (2011) Moving pictures of the human microbiome. Genome biology 12: R50.

Moving Pictures of the Human Microbiome

- Investigate the relative temporal variability of body sites.
- Is there a temporal core microbiome?

Moving Pictures of the Human Microbiome: QIIME tutorial

- A **small** subset of the full data set to facilitate short run time: ~0.1% of the full sequence collection.
- Sequenced across six Illumina GAIIx lanes, with a subset of the samples also sequenced on 454.

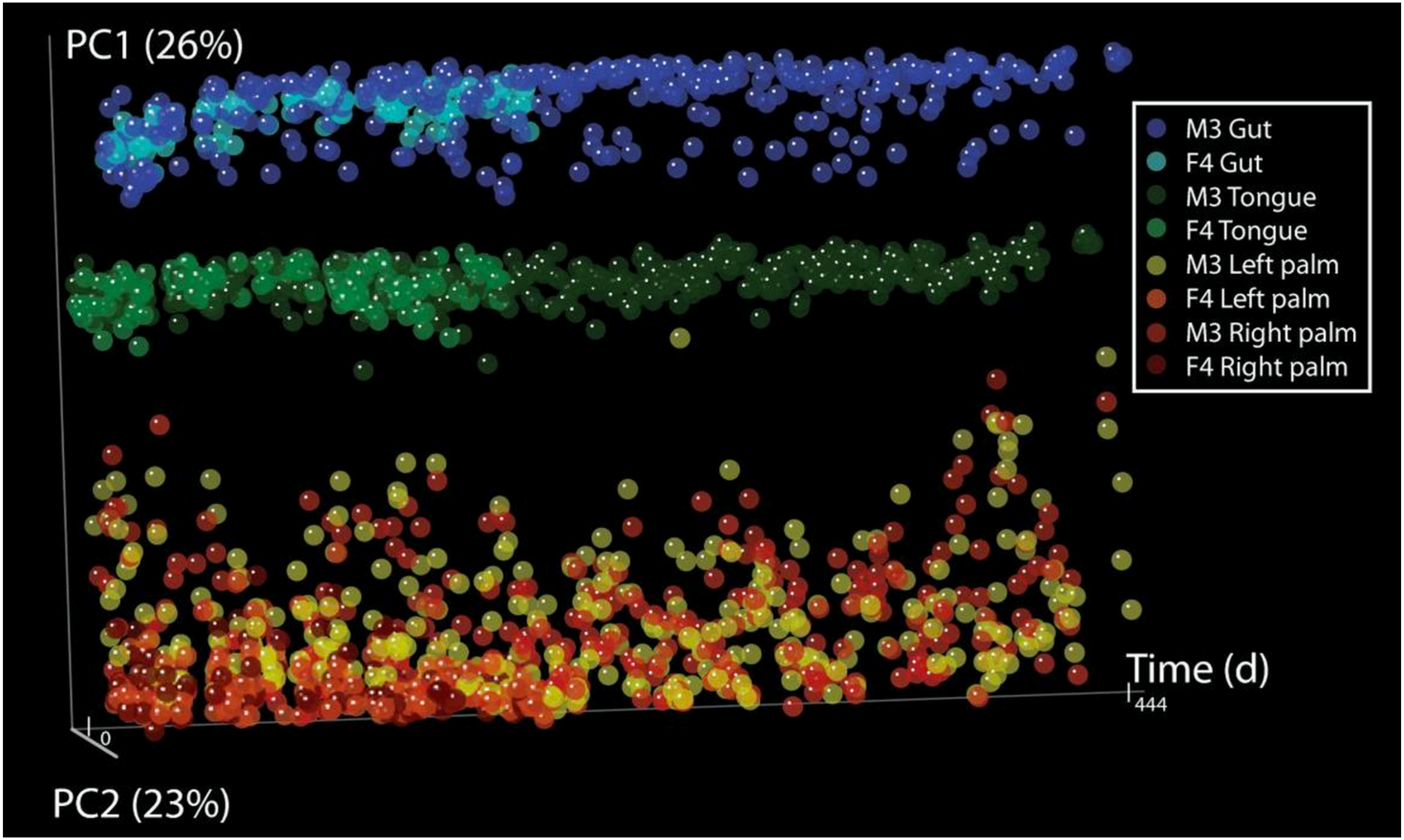
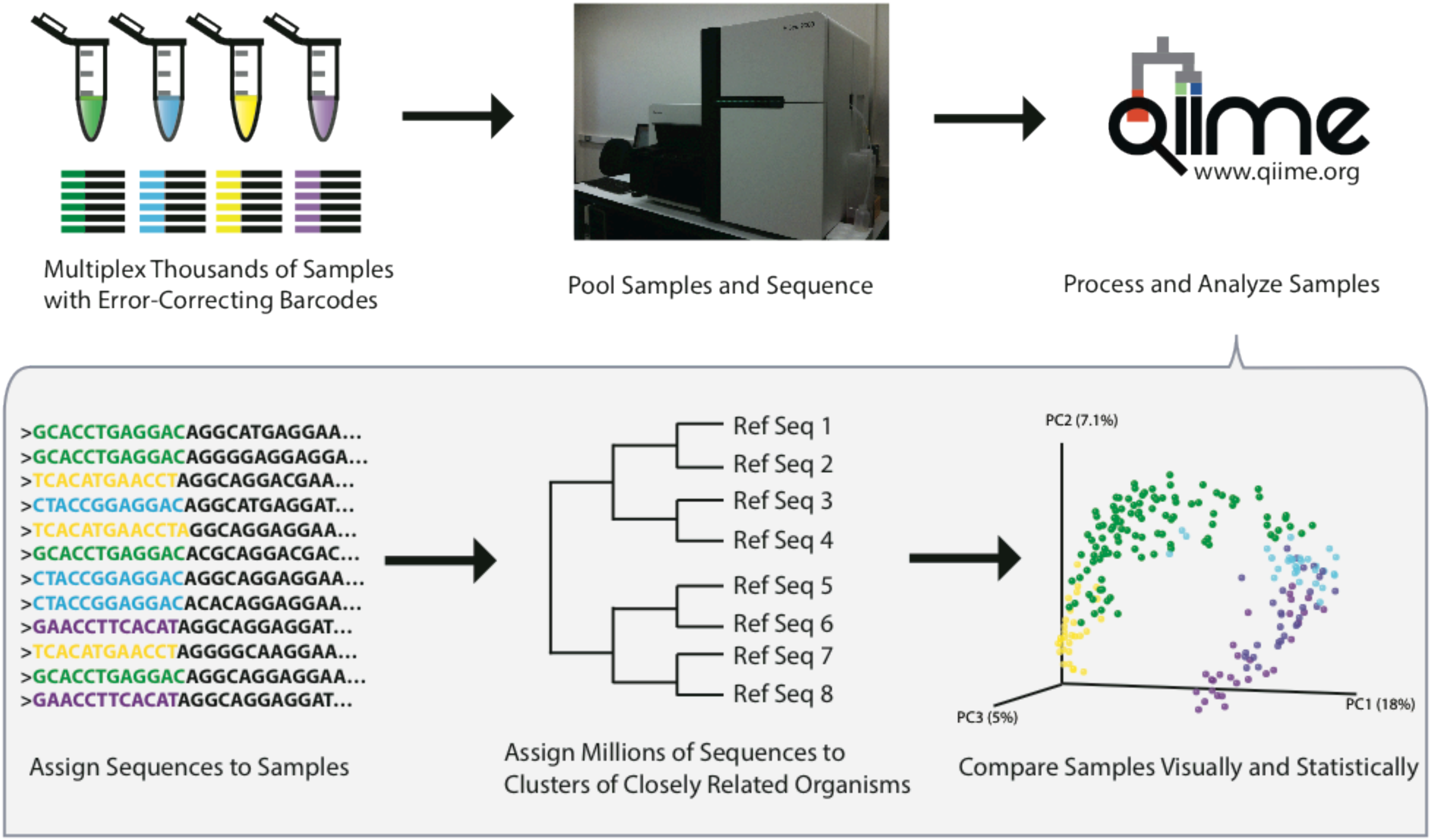


Figure 1, http://www.ncbi.nlm.nih.gov/pubmed/21624126



Multiplex Thousands of Samples
with Error-Correcting Barcodes

Pool Samples and Sequence

Process and Analyze Samples

Assign Sequences to Samples

Assign Millions of Sequences to
Clusters of Closely Related Organisms

Compare Samples Visually and Statistically

Metcalf, Jessica (2014): Overview of data generation, processing and analysis using QIIME. figshare.
http://dx.doi.org/10.6084/m9.figshare.902219

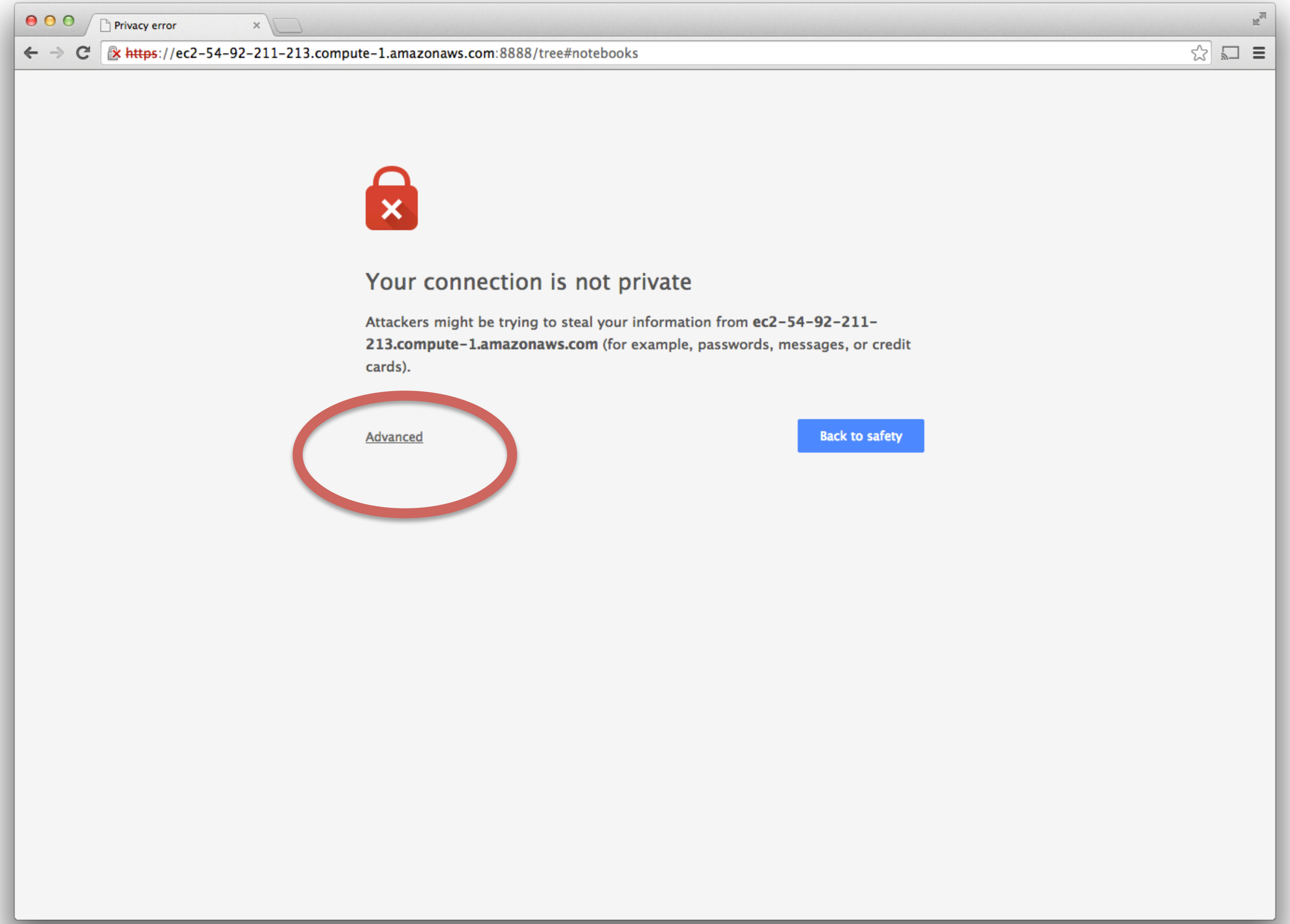
Tutorial

1) Go to the URL of your table:
   http://tinyurl.com/tableXXXX-notebook

Password: qiime2014~

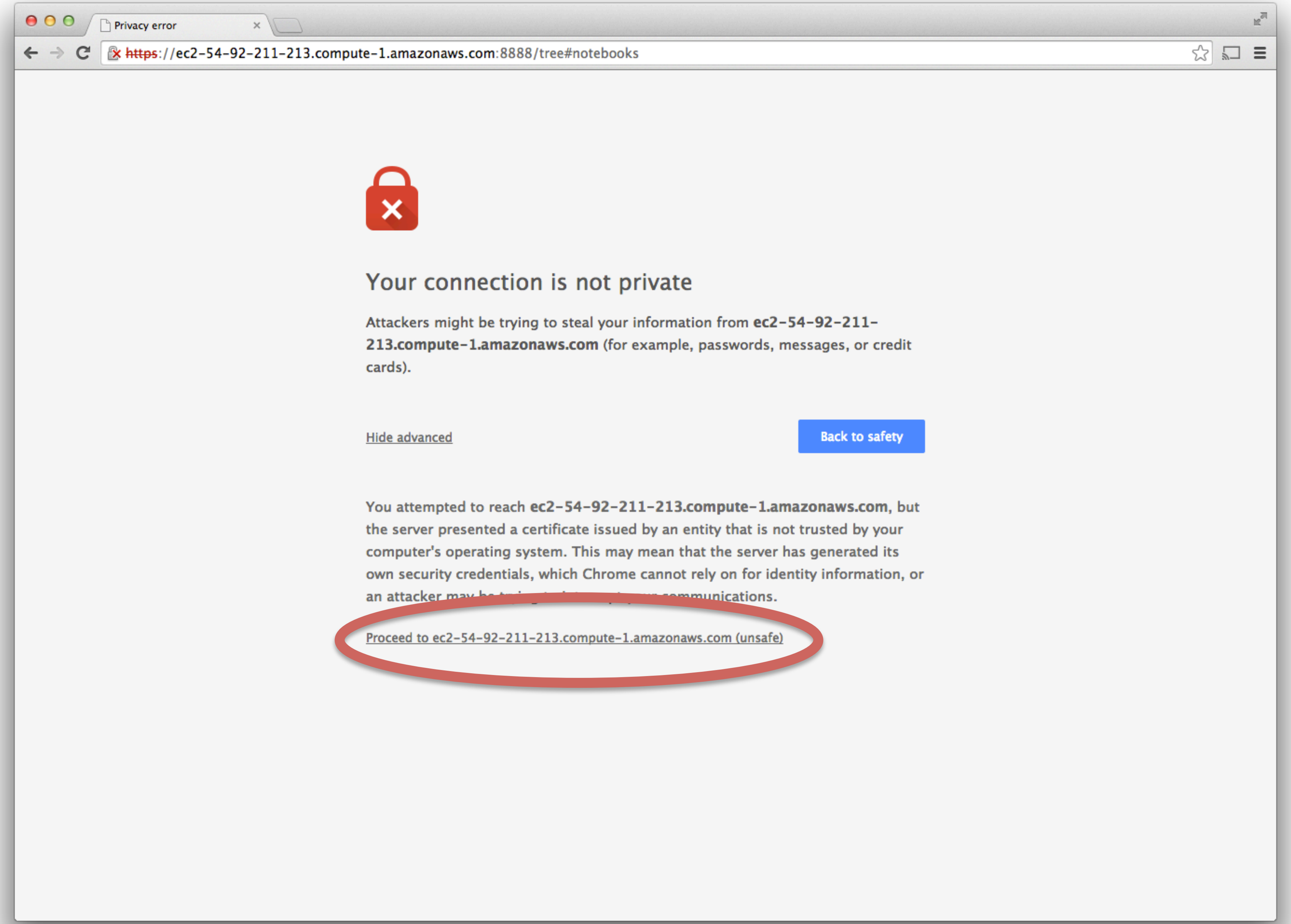
http://tinyurl.com/table1-notebook
http://tinyurl.com/table2-notebook
http://tinyurl.com/table3-notebook
http://tinyurl.com/table4-notebook
http://tinyurl.com/table5-notebook
http://tinyurl.com/table6-notebook
http://tinyurl.com/table7-notebook
http://tinyurl.com/table8-notebook
http://tinyurl.com/table9-notebook
http://tinyurl.com/table10-notebook

Replace the XXXX with the number of your table



Your connection is not private

Attackers might be trying to steal your information from **ec2-54-92-211-213.compute-1.amazonaws.com** (for example, passwords, messages, or credit cards).



Your connection is not private

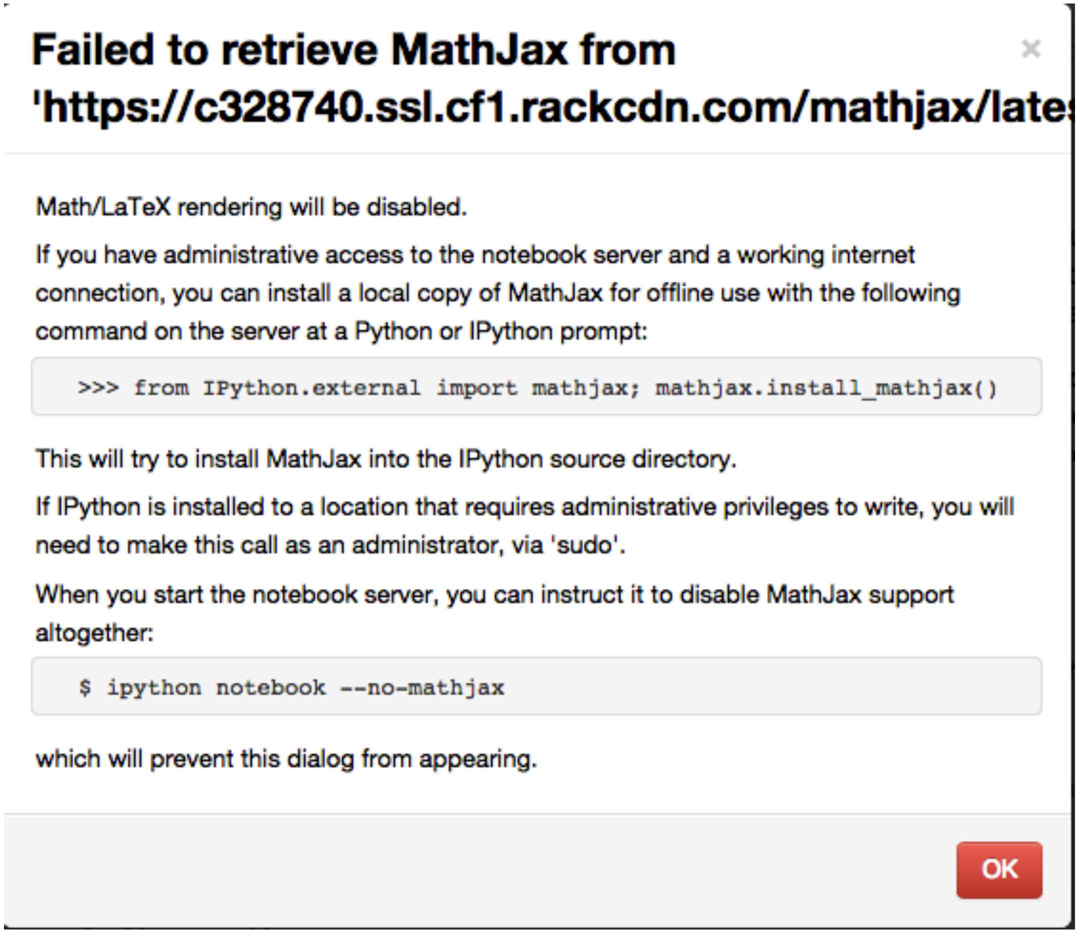
Attackers might be trying to steal your information from **ec2-54-92-211-213.compute-1.amazonaws.com** (for example, passwords, messages, or credit cards).

You attempted to reach **ec2-54-92-211-213.compute-1.amazonaws.com**, but the server presented a certificate issued by an entity that is not trusted by your computer's operating system. This may mean that the server has generated its own security credentials, which Chrome cannot rely on for identity information, or an attacker may be trying to intercept your communications.

Proceed to ec2-54-92-211-213.compute-1.amazonaws.com (unsafe)

Tutorial

2) Click on the notebook named tutorial_master.



Failed to retrieve MathJax from 'https://c328740.ssl.cf1.rackcdn.com/mathjax/late

Math/LaTeX rendering will be disabled.

If you have administrative access to the notebook server and a working internet connection, you can install a local copy of MathJax for offline use with the following command on the server at a Python or IPython prompt:

```
>>> from IPython.external import mathjax; mathjax.install_mathjax()
```

This will try to install MathJax into the IPython source directory.

If IPython is installed to a location that requires administrative privileges to write, you will need to make this call as an administrator, via 'sudo'.

When you start the notebook server, you can instruct it to disable MathJax support altogether:

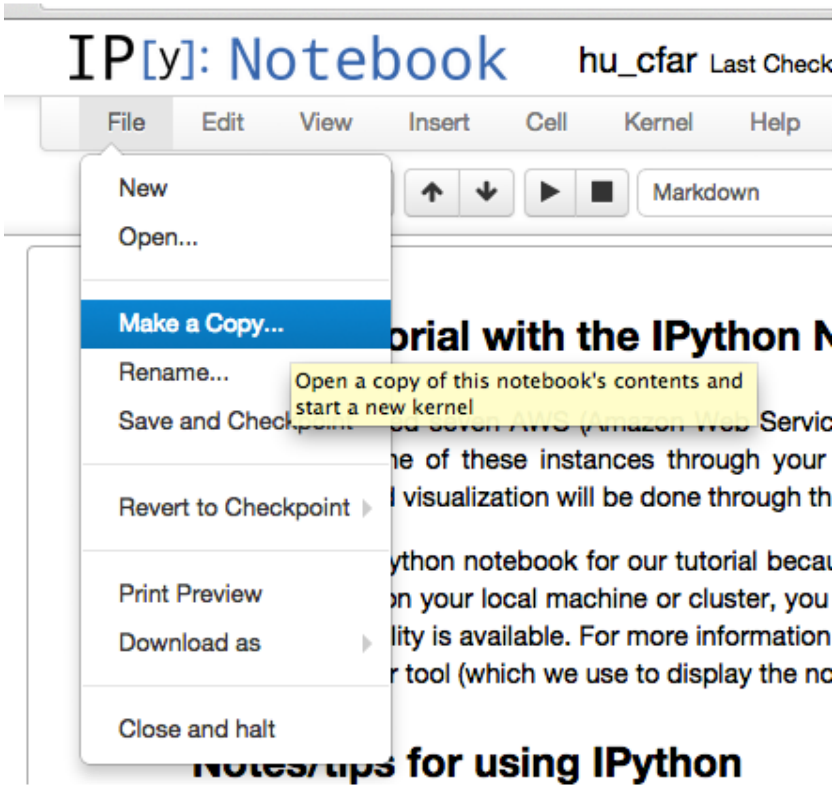
```
$ ipython notebook --no-mathjax
```

which will prevent this dialog from appearing.

Click OK

Tutorial

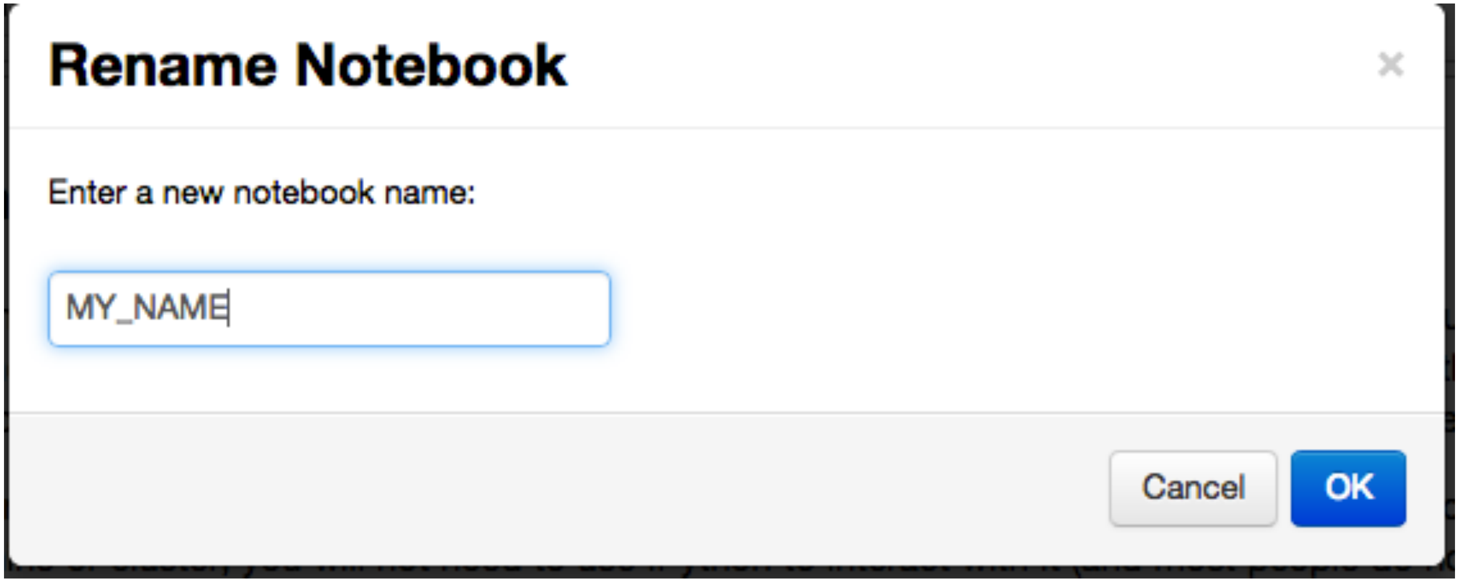
3) File>Make a Copy



Open a copy of this notebook's contents and start a new kernel

Tutorial

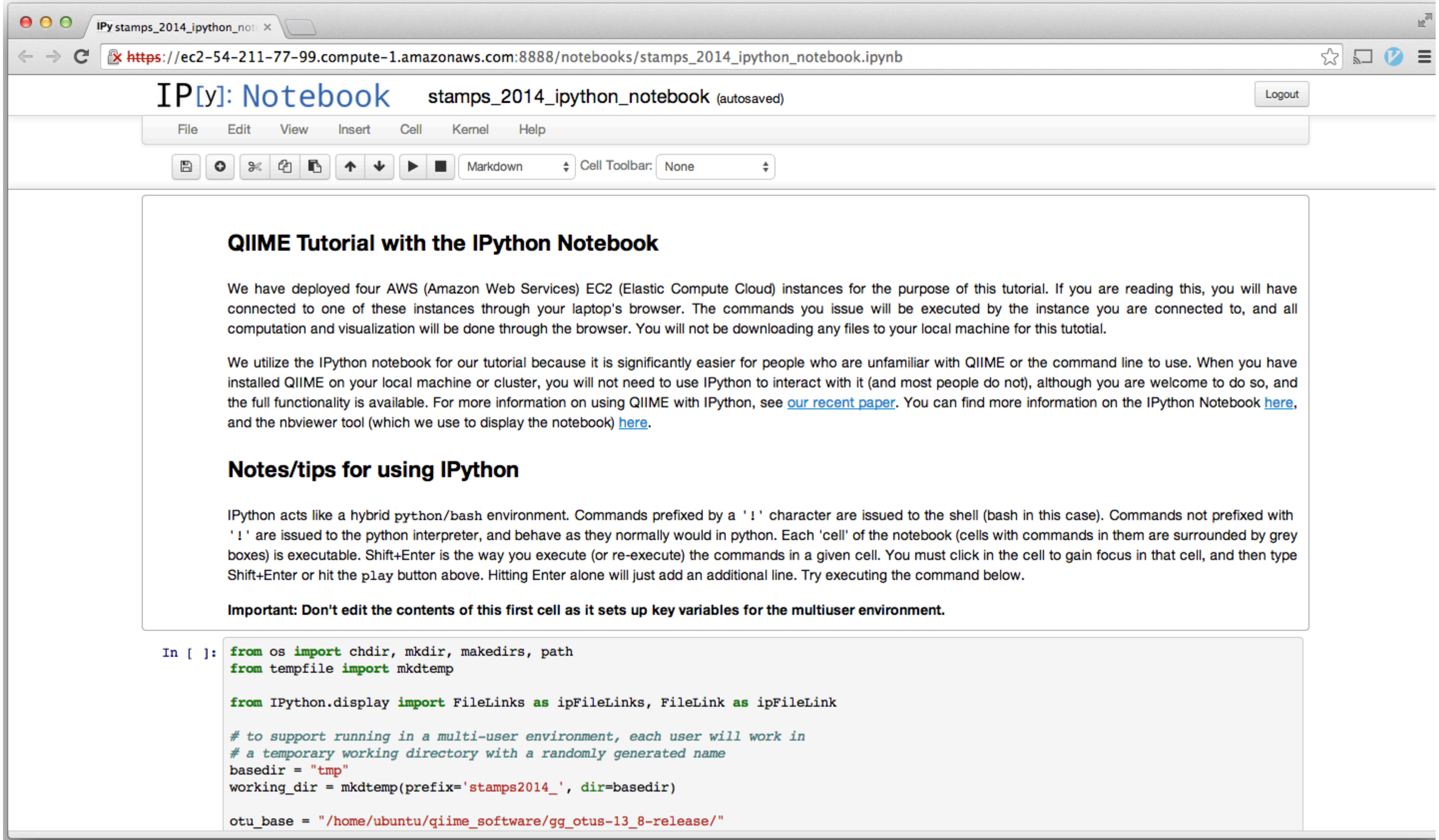
3) Rename to your name



Rename Notebook

Enter a new notebook name:

MY_NAME



QIIME Tutorial with the IPython Notebook

We have deployed four AWS (Amazon Web Services) EC2 (Elastic Compute Cloud) instances for the purpose of this tutorial. If you are reading this, you will have connected to one of these instances through your laptop's browser. The commands you issue will be executed by the instance you are connected to, and all computation and visualization will be done through the browser. You will not be downloading any files to your local machine for this tutotial.

We utilize the IPython notebook for our tutorial because it is significantly easier for people who are unfamiliar with QIIME or the command line to use. When you have installed QIIME on your local machine or cluster, you will not need to use IPython to interact with it (and most people do not), although you are welcome to do so, and the full functionality is available. For more information on using QIIME with IPython, see our recent paper. You can find more information on the IPython Notebook here, and the nbviewer tool (which we use to display the notebook) here.

Notes/tips for using IPython

IPython acts like a hybrid `python/bash` environment. Commands prefixed by a `'!'` character are issued to the shell (bash in this case). Commands not prefixed with `'!'` are issued to the python interpreter, and behave as they normally would in python. Each 'cell' of the notebook (cells with commands in them are surrounded by grey boxes) is executable. Shift+Enter is the way you execute (or re-execute) the commands in a given cell. You must click in the cell to gain focus in that cell, and then type Shift+Enter or hit the `play` button above. Hitting Enter alone will just add an additional line. Try executing the command below.

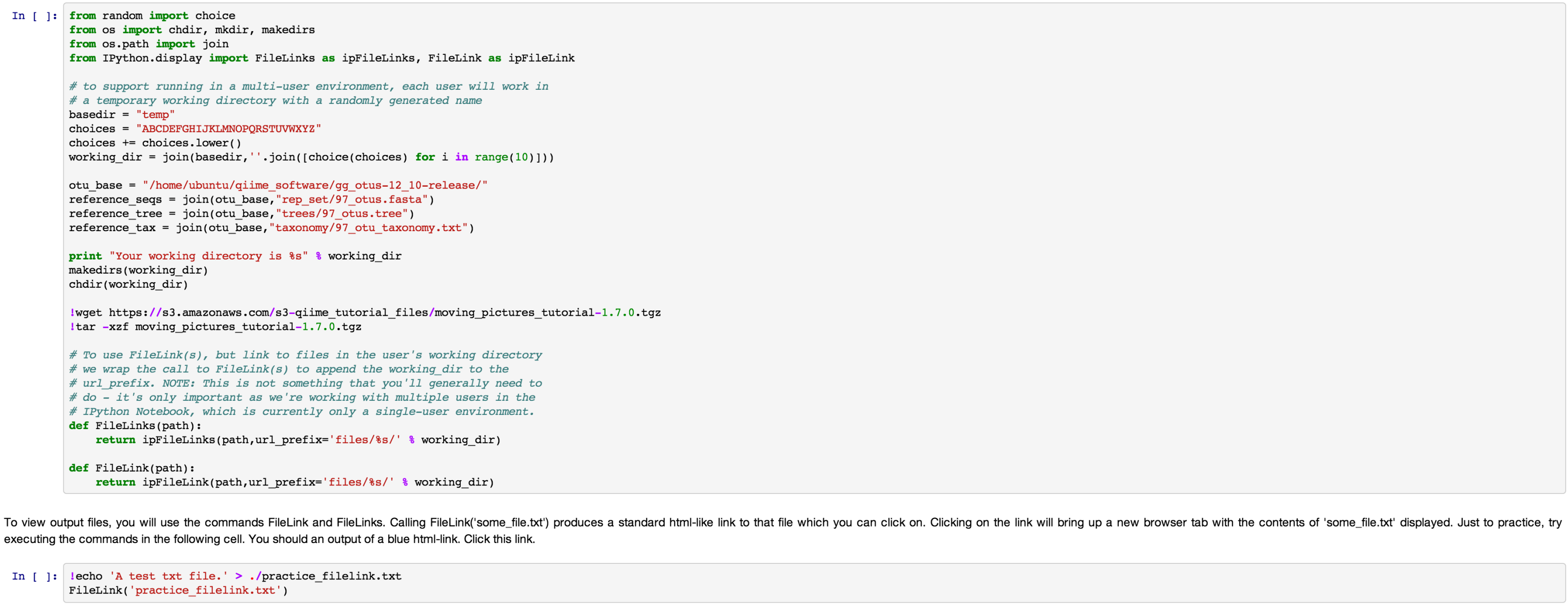
Important: Don't edit the contents of this first cell as it sets up key variables for the multiuser environment.

```
from os import chdir, mkdir, makedirs, path
from tempfile import mkdtemp

from IPython.display import FileLinks as ipFileLinks, FileLink as ipFileLink

# to support running in a multi-user environment, each user will work in
# a temporary working directory with a randomly generated name
basedir = "tmp"
working_dir = mkdtemp(prefix='stamps2014_', dir=basedir)

otu_base = "/home/ubuntu/qiime_software/gg_otus-13_8-release/"
```

IPython reference

IPython acts like a hybrid python/bash environment.

The way we interact with the IPython notebook is through the ‘cells’



In [ ]:

```
from random import choice
from os import chdir, mkdir, makedirs
from os.path import join
from IPython.display import FileLinks as ipFileLinks, FileLink as ipFileLink

# to support running in a multi-user environment, each user will work in
# a temporary working directory with a randomly generated name
basedir = "temp"
choices = "ABCDEFGHIJKLMNOPQRSTUVWXYZ"
choices += choices.lower()
working_dir = join(basedir,''.join([choice(choices) for i in range(10)]))

otu_base = "/home/ubuntu/qiime_software/gg_otus-12_10-release/"
reference_seqs = join(otu_base,"rep_set/97_otus.fasta")
reference_tree = join(otu_base,"trees/97_otus.tree")
reference_tax = join(otu_base,"taxonomy/97_otu_taxonomy.txt")

print "Your working directory is %s" % working_dir
makedirs(working_dir)
chdir(working_dir)

!wget https://s3.amazonaws.com/s3-qiime_tutorial_files/moving_pictures_tutorial-1.7.0.tgz
!tar -xzf moving_pictures_tutorial-1.7.0.tgz

# To use FileLink(s), but link to files in the user's working directory
# we wrap the call to FileLink(s) to append the working_dir to the
# url_prefix. NOTE: This is not something that you'll generally need to
# do - it's only important as we're working with multiple users in the
# IPython Notebook, which is currently only a single-user environment.
def FileLinks(path):
    return ipFileLinks(path,url_prefix='files/%s/' % working_dir)

def FileLink(path):
    return ipFileLink(path,url_prefix='files/%s/' % working_dir)
```

To view output files, you will use the commands FileLink and FileLinks. Calling FileLink('some_file.txt') produces a standard html-like link to that file which you can click on. Clicking on the link will bring up a new browser tab with the contents of 'some_file.txt' displayed. Just to practice, try executing the commands in the following cell. You should an output of a blue html-link. Click this link.

In [ ]:

```
!echo 'A test txt file.' > ./practice_filelink.txt
FileLink('practice_filelink.txt')
```

IPython reference

IPython acts like a hybrid python/bash environment.

The way we interact with the IPython notebook is through the ‘cells’

```
from random import choice
from os import chdir, mkdir, makedirs
from os.path import join
from IPython.display import FileLinks as ipFileLinks, FileLink as ipFileLink

# to support running in a multi-user environment, each user will work in
# a temporary working directory with a randomly generated name
basedir = "temp"
choices = "ABCDEFGHIJKLMNOPQRSTUVWXYZ"
choices += choices.lower()
working_dir = join(basedir,''.join([choice(choices) for i in range(10)]))

otu_base = "/home/ubuntu/qiime_software/gg_otus-12_10-release/"
reference_seqs = join(otu_base,"rep_set/97_otus.fasta")
reference_tree = join(otu_base,"trees/97_otus.tree")
reference_tax = join(otu_base,"taxonomy/97_otu_taxonomy.txt")

print "Your working directory is %s" % working_dir
makedirs(working_dir)
chdir(working_dir)

!wget https://s3.amazonaws.com/s3-qiime_tutorial_files/moving_pictures_tutorial-1.7.0.tgz
!tar -xzf moving_pictures_tutorial-1.7.0.tgz

# To use FileLink(s), but link to files in the user's working directory
# we wrap the call to FileLink(s) to append the working_dir to the
# url_prefix. NOTE: This is not something that you'll generally need to
# do - it's only important as we're working with multiple users in the
# IPython Notebook, which is currently only a single-user environment.
def FileLinks(path):
    return ipFileLinks(path,url_prefix='files/%s/' % working_dir)

def FileLink(path):
    return ipFileLink(path,url_prefix='files/%s/' % working_dir)
```

To view output files, you will use the commands FileLink and FileLinks. Calling FileLink('some_file.txt') produces a standard html-like link to that file which you can click on. Clicking on the link will bring up a new browser tab with the contents of 'some_file.txt' displayed. Just to practice, try executing the commands in the following cell. You should an output of a blue html-link. Click this link.

```
!echo 'A test txt file.' > ./practice_filelink.txt
FileLink('practice_filelink.txt')
```

IPython reference

IPython acts like a hybrid python/bash environment.

The way we interact with the IPython notebook is through the ‘cells’

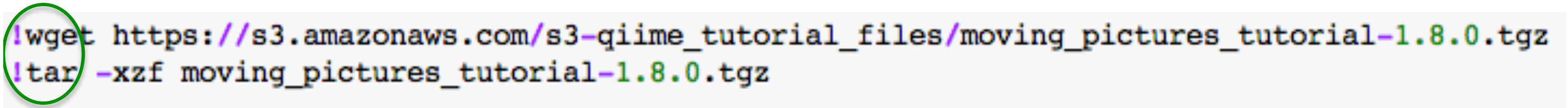
```
In [ ]: from random import choice
        from os import chdir, mkdir, makedirs
        from os.path import join
        from IPython.display import FileLinks as ipFileLinks, FileLink as ipFileLink

        # to support running in a multi-user environment, each user will work in
        # a temporary working directory with a randomly generated name
        basedir = "temp"
        choices = "ABCDEFGHIJKLMNOPQRSTUVWXYZ"
        choices += choices.lower()
        working_dir = join(basedir,''.join([choice(choices) for i in range(10)]))

        otu_base = "/home/ubuntu/qiime_software/gg_otus-12_10-release/"
        reference_seqs = join(otu_base,"rep_set/97_otus.fasta")
        reference_tree = join(otu_base,"trees/97_otus.tree")
        reference_tax = join(otu_base,"taxonomy/97_otu_taxonomy.txt")

        print "Your working directory is %s" % working_dir
        makedirs(working_dir)
        chdir(working_dir)

        !wget https://s3.amazonaws.com/s3-qiime_tutorial_files/moving_pictures_tutorial-1.7.0.tgz
        !tar -xzf moving_pictures_tutorial-1.7.0.tgz

        # To use FileLink(s), but link to files in the user's working directory
        # we wrap the call to FileLink(s) to append the working_dir to the
        # url_prefix. NOTE: This is not something that you'll generally need to
        # do - it's only important as we're working with multiple users in the
        # IPython Notebook, which is currently only a single-user environment.
        def FileLinks(path):
            return ipFileLinks(path,url_prefix='files/%s/' % working_dir)

        def FileLink(path):
            return ipFileLink(path,url_prefix='files/%s/' % working_dir)
```

To view output files, you will use the commands FileLink and FileLinks. Calling FileLink('some_file.txt') produces a standard html-like link to that file which you can click on. Clicking on the link will bring up a new browser tab with the contents of 'some_file.txt' displayed. Just to practice, try

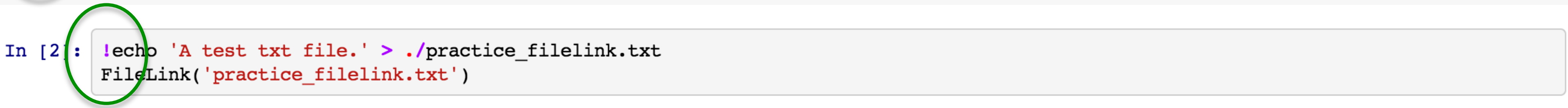
```
In [ ]: !echo 'A test txt file.' > ./practice_filelink.txt
        FileLink('practice_filelink.txt')
```

IPython reference

Commands prefixed by a '!' character are issued to the shell (just like what your terminal runs).

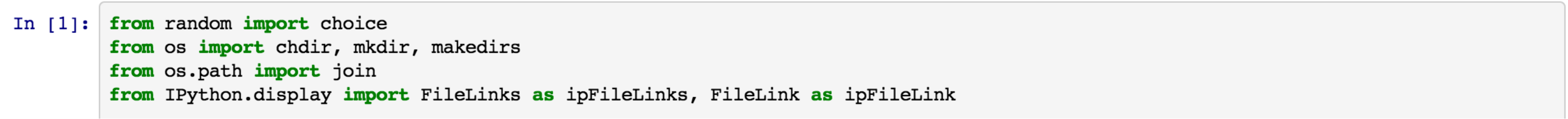


```
!wget https://s3.amazonaws.com/s3-qiime_tutorial_files/moving_pictures_tutorial-1.8.0.tgz
!tar -xzf moving_pictures_tutorial-1.8.0.tgz
```



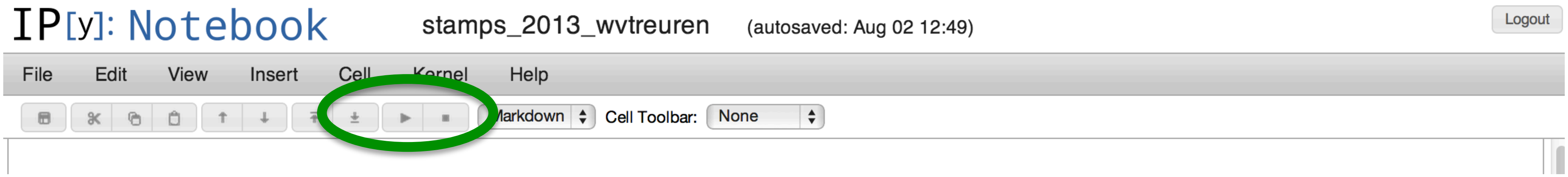
```
In [2]: !echo 'A test txt file.' > ./practice_filelink.txt
        FileLink('practice_filelink.txt')
```

Commands not prefixed with '!' are issued to python, and behave as they normally would in python.



```
In [1]: from random import choice
        from os import chdir, mkdir, makedirs
        from os.path import join
        from IPython.display import FileLinks as ipFileLinks, FileLink as ipFileLink
```

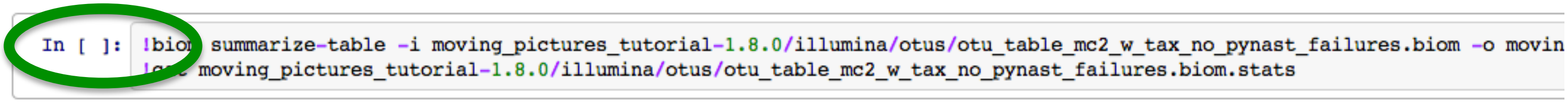
Each 'cell' of the notebook is executable. Shift+Enter (or the play button) is the way you execute (or re-execute) the commands in a given cell. You must click in the cell to gain focus in that cell, and then type Shift+Enter or hit the play button



IPython reference

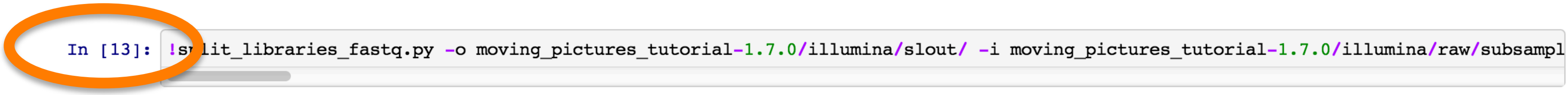
Each executable has a prefix that shows you its status (if it has been run, if it hasn't been run, or if its still running)

Hasn't been run



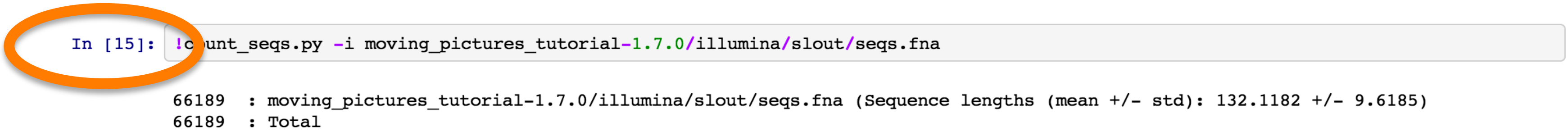
```
In [ ]: !bion summarize-table -i moving_pictures_tutorial-1.8.0/illumina/otus/otu_table_mc2_w_tax_no_pynast_failures.biom -o movin
        !cat moving_pictures_tutorial-1.8.0/illumina/otus/otu_table_mc2_w_tax_no_pynast_failures.biom.stats
```

Has been run



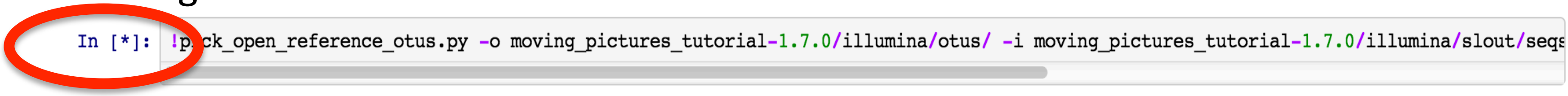
```
In [13]: !split_libraries_fastq.py -o moving_pictures_tutorial-1.7.0/illumina/slout/ -i moving_pictures_tutorial-1.7.0/illumina/raw/subsampl
```

```
In [14]: FileLinks('moving_pictures_tutorial-1.7.0/illumina/slout/')

Out[14]: moving_pictures_tutorial-1.7.0/illumina/slout/
           histograms.txt
           split_library_log.txt
           seqs.fna
```

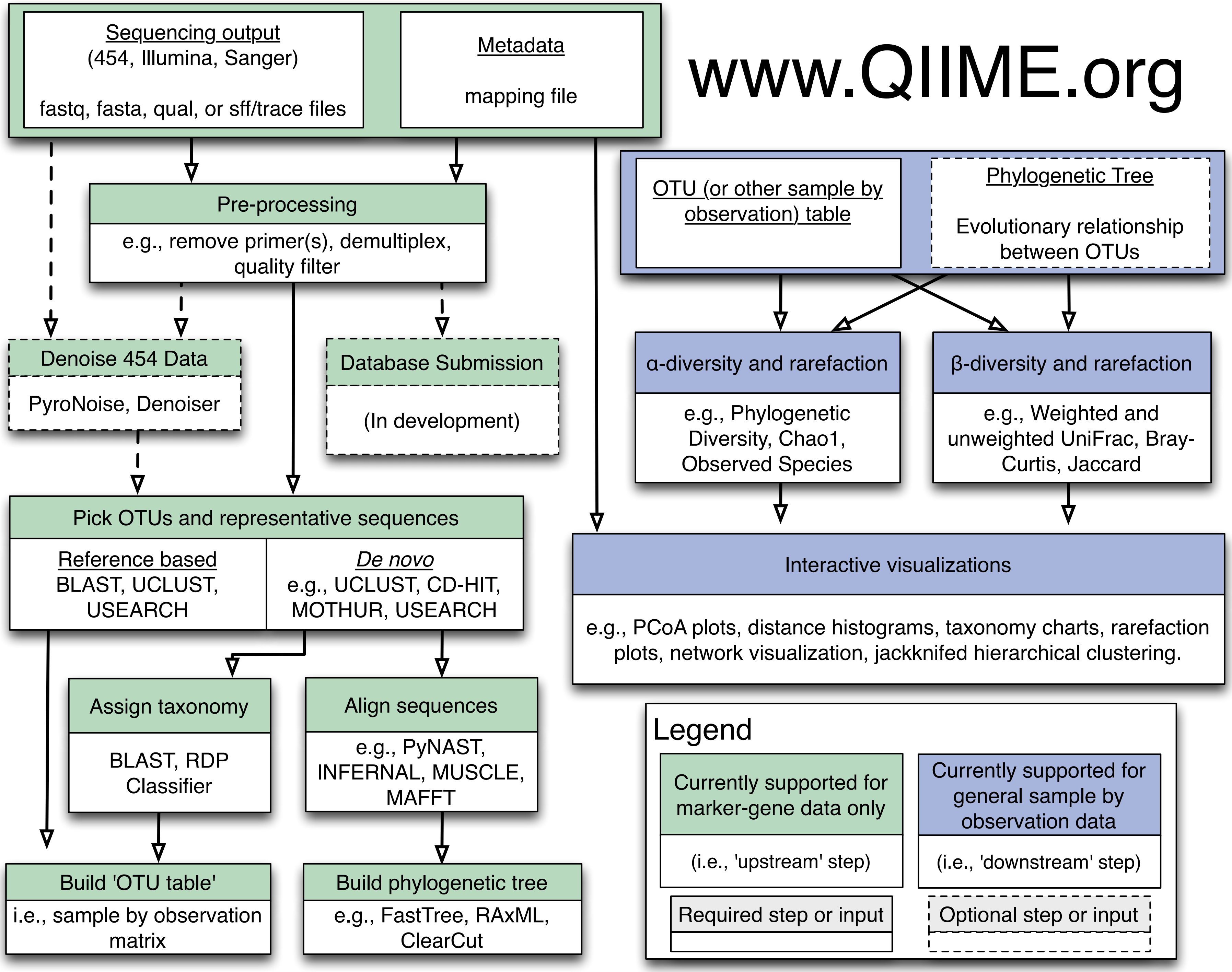


```
In [15]: !count_seqs.py -i moving_pictures_tutorial-1.7.0/illumina/slout/seqs.fna

         66189  : moving_pictures_tutorial-1.7.0/illumina/slout/seqs.fna (Sequence lengths (mean +/- std): 132.1182 +/- 9.6185)
         66189  : Total
```

Still running



```
In [*]: !pick_open_reference_otus.py -o moving_pictures_tutorial-1.7.0/illumina/otus/ -i moving_pictures_tutorial-1.7.0/illumina/slout/seqs
```



Sequencing output
(454, Illumina, Sanger)
fastq, fasta, qual, or sff/trace files

Metadata
mapping file

www.QIIME.org

Pre-processing
e.g., remove primer(s), demultiplex, quality filter

Denoise 454 Data
PyroNoise, Denoiser

Database Submission
(In development)

Pick OTUs and representative sequences

Reference based
BLAST, UCLUST, USEARCH

De novo
e.g., UCLUST, CD-HIT, MOTHUR, USEARCH

Assign taxonomy
BLAST, RDP Classifier

Align sequences
e.g., PyNAST, INFERNAL, MUSCLE, MAFFT

Build 'OTU table'
i.e., sample by observation matrix

Build phylogenetic tree
e.g., FastTree, RAxML, ClearCut

OTU (or other sample by observation) table

Phylogenetic Tree
Evolutionary relationship between OTUs

α-diversity and rarefaction
e.g., Phylogenetic Diversity, Chao1, Observed Species

β-diversity and rarefaction
e.g., Weighted and unweighted UniFrac, Bray-Curtis, Jaccard

Interactive visualizations
e.g., PCoA plots, distance histograms, taxonomy charts, rarefaction plots, network visualization, jackknifed hierarchical clustering.

Legend

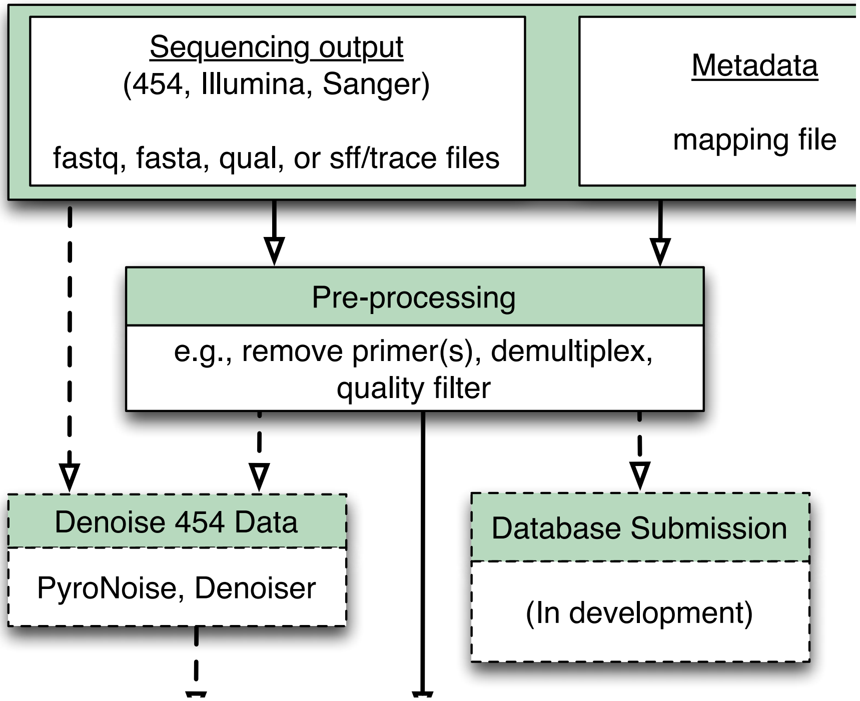
Currently supported for marker-gene data only
(i.e., 'upstream' step)

Currently supported for general sample by observation data
(i.e., 'downstream' step)

Required step or input

Optional step or input

Samples to sequences



<u>Sequencing output</u>
(454, Illumina, Sanger)

fastq, fasta, qual, or sff/trace files

<u>Metadata</u>

mapping file

Pre-processing

e.g., remove primer(s), demultiplex, quality filter

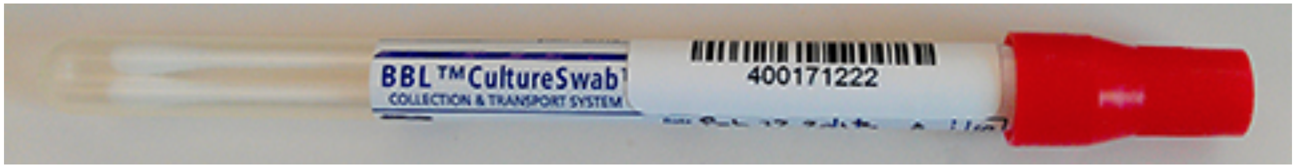
Denoise 454 Data

PyroNoise, Denoiser

Database Submission

(In development)

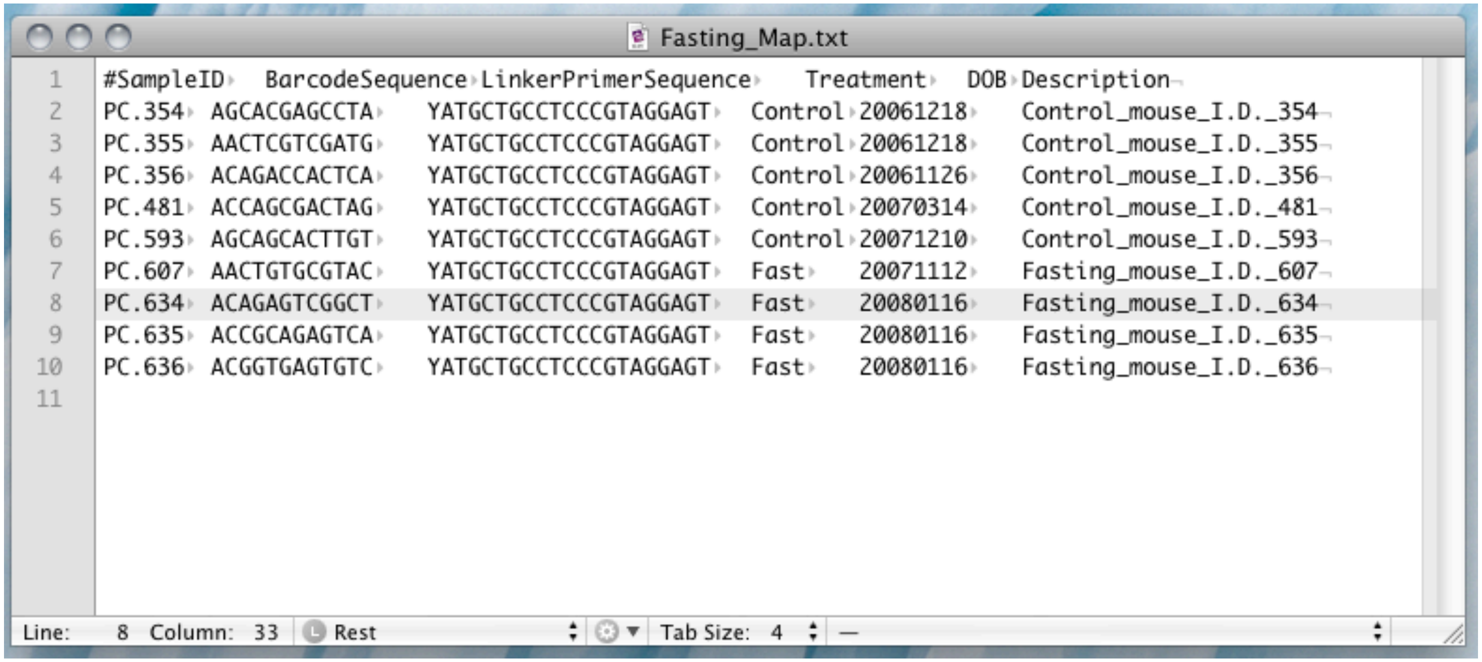
Relates samples to variables



SampleID	Sex	SampleType	Plotting your doom?
400171222	Male	Poop	Naturally

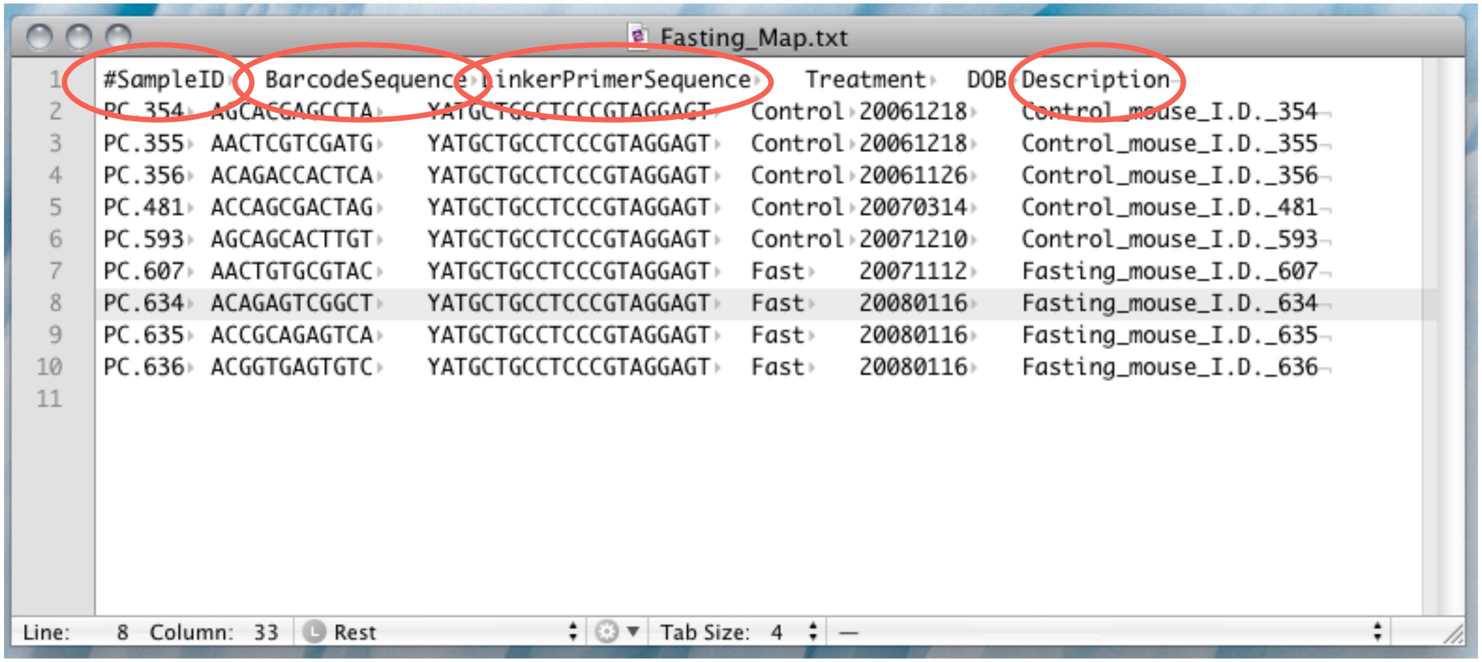
Cat image: http://www.heatherbrookes.org/hb-blog/?p=2341

Mapping file



```
#SampleID	BarcodeSequence	LinkerPrimerSequence	Treatment	DOB	Description
PC.354	AGCACGAGCCTA	YATGCTGCCTCCCGTAGGAGT	Control	20061218	Control_mouse_I.D._354
PC.355	AACTCGTCGATG	YATGCTGCCTCCCGTAGGAGT	Control	20061218	Control_mouse_I.D._355
PC.356	ACAGACCACTCA	YATGCTGCCTCCCGTAGGAGT	Control	20061126	Control_mouse_I.D._356
PC.481	ACCAGCGACTAG	YATGCTGCCTCCCGTAGGAGT	Control	20070314	Control_mouse_I.D._481
PC.593	AGCAGCACTTGT	YATGCTGCCTCCCGTAGGAGT	Control	20071210	Control_mouse_I.D._593
PC.607	AACTGTGCGTAC	YATGCTGCCTCCCGTAGGAGT	Fast	20071112	Fasting_mouse_I.D._607
PC.634	ACAGAGTCGGCT	YATGCTGCCTCCCGTAGGAGT	Fast	20080116	Fasting_mouse_I.D._634
PC.635	ACCGCAGAGTCA	YATGCTGCCTCCCGTAGGAGT	Fast	20080116	Fasting_mouse_I.D._635
PC.636	ACGGTGAGTGTC	YATGCTGCCTCCCGTAGGAGT	Fast	20080116	Fasting_mouse_I.D._636
```

Mapping file: always run `validate_mapping_file.py`



```
#SampleID	BarcodeSequence	LinkerPrimerSequence	Treatment	DOB	Description
PC.354	AGCACGAGCCTA	YATGCTGCCTCCCGTAGGAGT	Control	20061218	Control_mouse_I.D._354
PC.355	AACTCGTCGATG	YATGCTGCCTCCCGTAGGAGT	Control	20061218	Control_mouse_I.D._355
PC.356	ACAGACCACTCA	YATGCTGCCTCCCGTAGGAGT	Control	20061126	Control_mouse_I.D._356
PC.481	ACCAGCGACTAG	YATGCTGCCTCCCGTAGGAGT	Control	20070314	Control_mouse_I.D._481
PC.593	AGCAGCACTTGT	YATGCTGCCTCCCGTAGGAGT	Control	20071210	Control_mouse_I.D._593
PC.607	AACTGTGCGTAC	YATGCTGCCTCCCGTAGGAGT	Fast	20071112	Fasting_mouse_I.D._607
PC.634	ACAGAGTCGGCT	YATGCTGCCTCCCGTAGGAGT	Fast	20080116	Fasting_mouse_I.D._634
PC.635	ACCGCAGAGTCA	YATGCTGCCTCCCGTAGGAGT	Fast	20080116	Fasting_mouse_I.D._635
PC.636	ACGGTGAGTGTC	YATGCTGCCTCCCGTAGGAGT	Fast	20080116	Fasting_mouse_I.D._636
```

◯ = required field

Check for errors

- Work through the “Check our mapping file for errors” section, and execute the command
- Try validating the “bad” mapping file

Missing mapping file

From: XXXX@gmail.com
To: jose.clemente@gmail.com
Subject: mapping file

Hi,

I am doing microbial analysis with QIIME. I got the SFF files back from the sequencing center, but I lost the mapping file. Can you help me please?

Thanks,
XXXX

HT to Dr. Clemente for sharing this e-mail

Missing mapping file

From: XXXX@gmail.com
To: jose.clemente@gmail.com
Subject: mapping file

Hi,

I am doing microbial a ysis h QIIM I got t SFF files back fr the sequencing center, but I lost the m ping Can u help e please?

Thanks,
XXXX

NO

HT to Dr. Clemente for sharing this e-mail

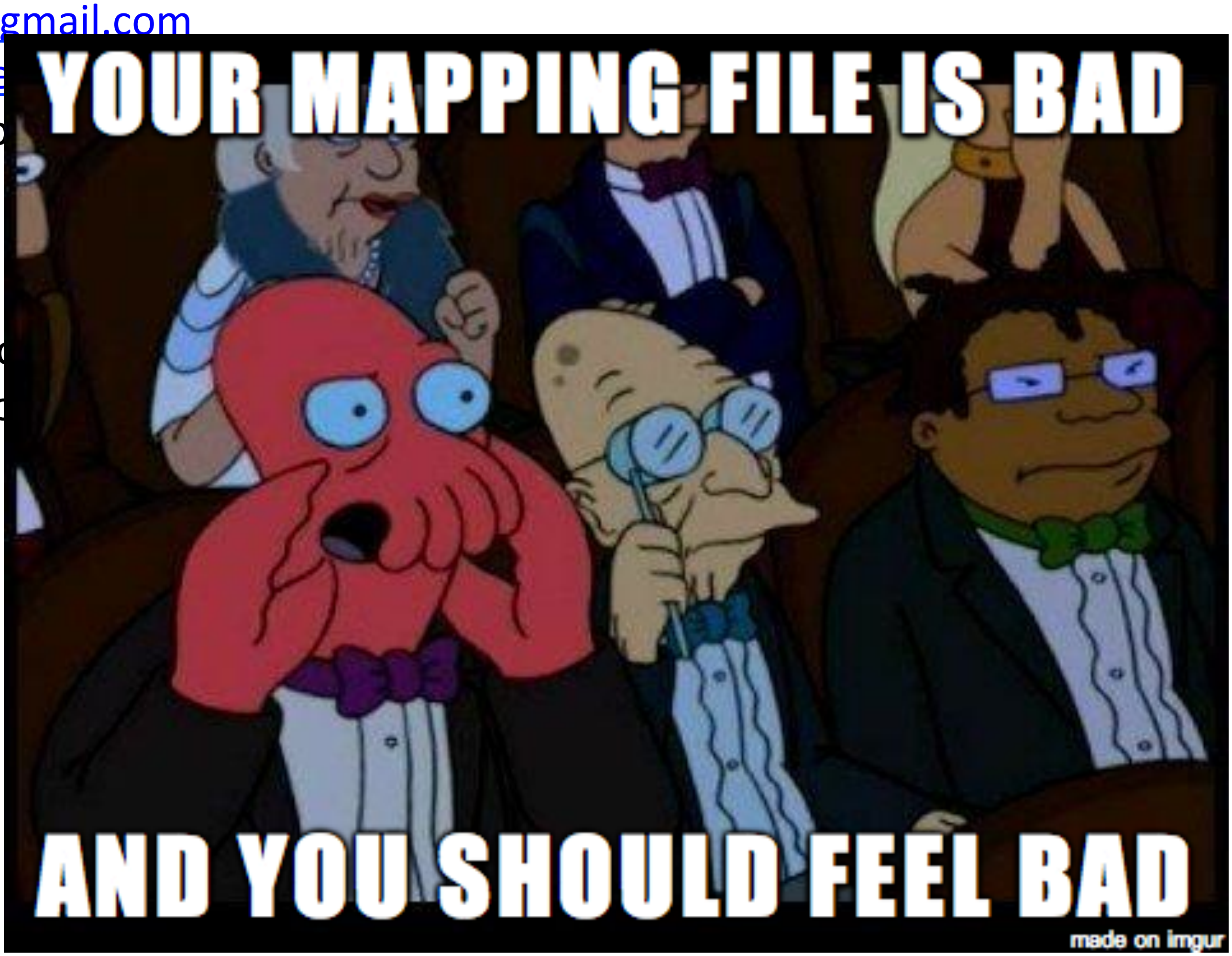
Missing mapping file

From: XXXX@gmail.com
To: jose.cleme
Subject: mapp

Hi,

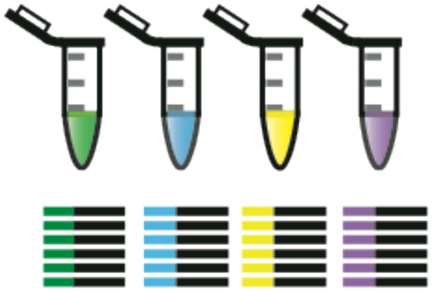
I am doing mic ncing center, but I lo

Thanks,
XXXX



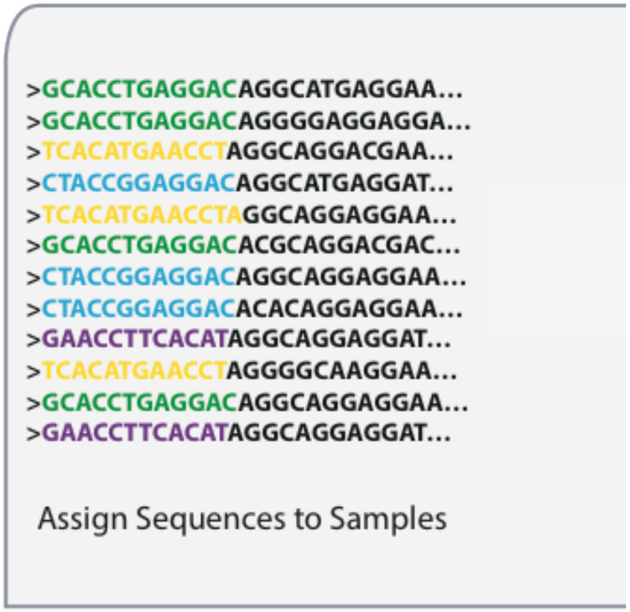
HT to Dr. Clemente for sharing this e-mail

Split libraries



Multiplex Thousands of Samples
with Error-Correcting Barcodes

- Perform quality filtering
- Demultiplex sequences



```
>GCACCTGAGGACAGGCATGAGGAA...
>GCACCTGAGGACAGGGGAGGAGGA...
>TCACATGAACCTAGGCAGGACGAA...
>CTACCGGAGGACAGGCATGAGGAT...
>TCACATGAACCTAGGCAGGAGGAA...
>GCACCTGAGGACACGCAGGACGAC...
>CTACCGGAGGACAGGCAGGAGGAA...
>CTACCGGAGGACACACAGGAGGAA...
>GAACCTTCACATAGGCAGGAGGAT...
>TCACATGAACCTAGGGGCAAGGAA...
>GCACCTGAGGACAGGCAGGAGGAA...
>GAACCTTCACATAGGCAGGAGGAT...
```

Assign Sequences to Samples

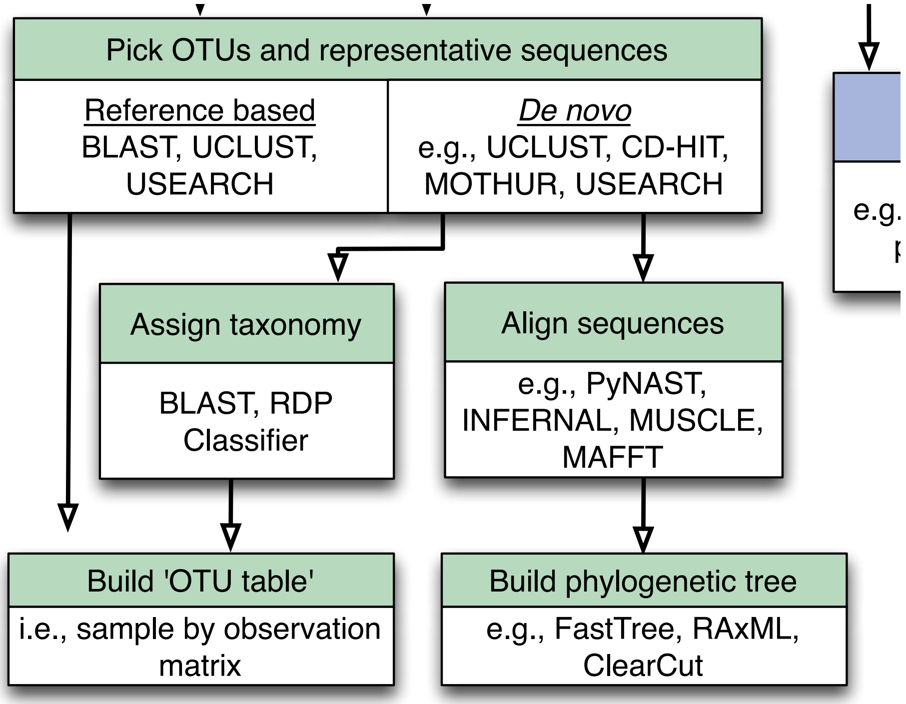
http://www.earthmicrobiome.org/emp-standard-protocols/16s/

Demultiplex your data

- Read through and execute the “Demultiplexing and quality filtering sequences” section

Sequences to OTUs and Phylogeny

And coming soon:
SortMeRNA



Pick OTUs and representative sequences

Reference based
BLAST, UCLUST,
USEARCH

De novo
e.g., UCLUST, CD-HIT,
MOTHUR, USEARCH

Assign taxonomy

BLAST, RDP
Classifier

Align sequences

e.g., PyNAST,
INFERNAL, MUSCLE,
MAFFT

Build 'OTU table'

i.e., sample by observation
matrix

Build phylogenetic tree

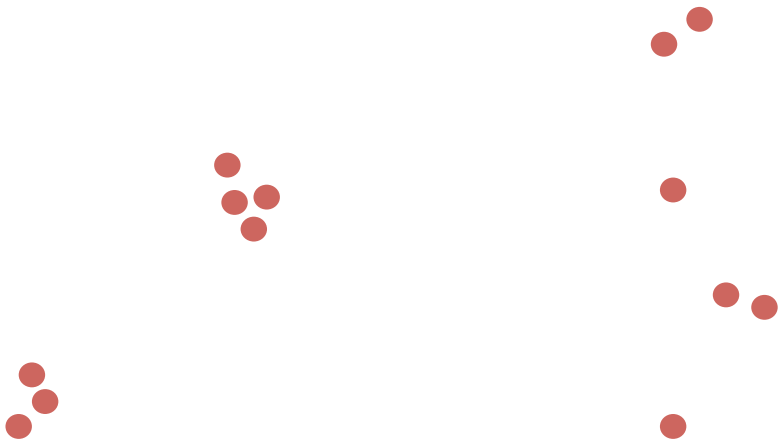
e.g., FastTree, RAxML,
ClearCut

e.g.

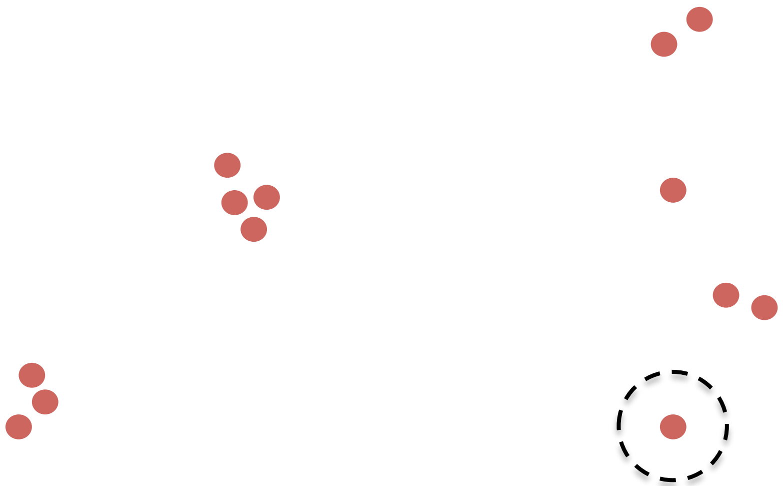
Pick OTUs

- Execute the cell that begins with:
  - `!pick_open_reference_otus.py –o mov…`

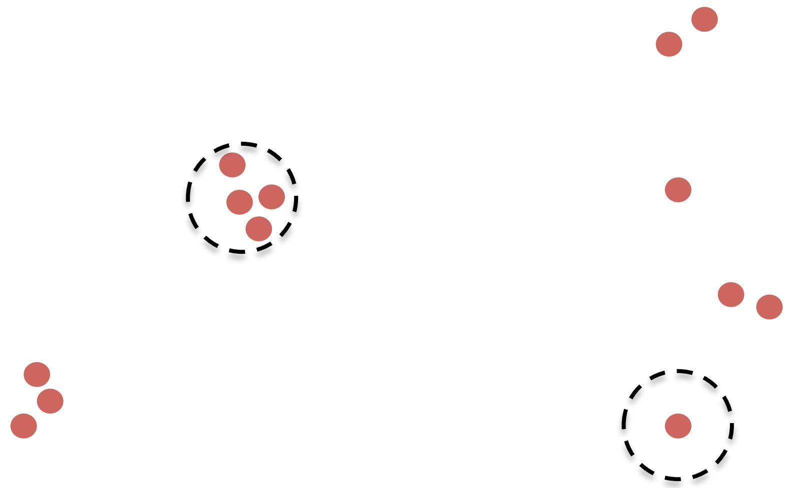
Understanding OTU picking



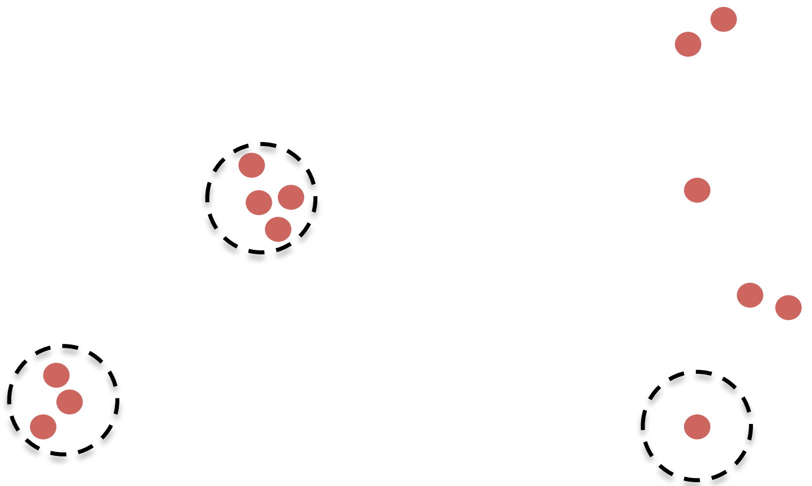
Understanding OTU picking



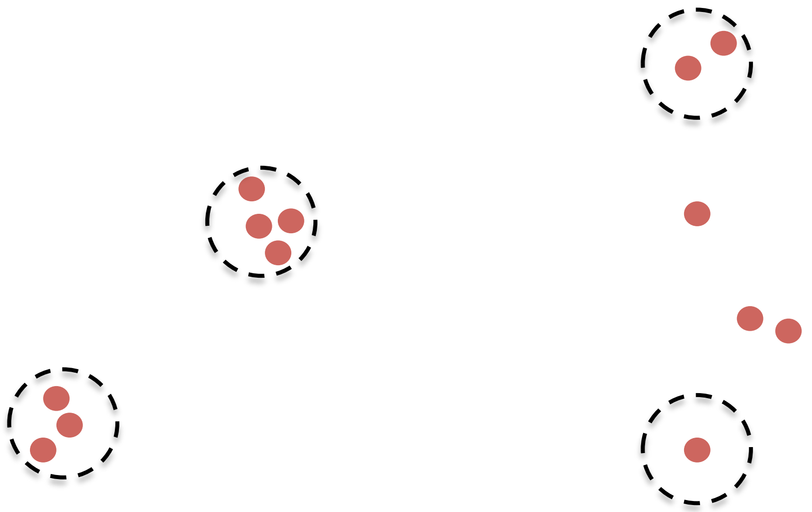
Understanding OTU picking



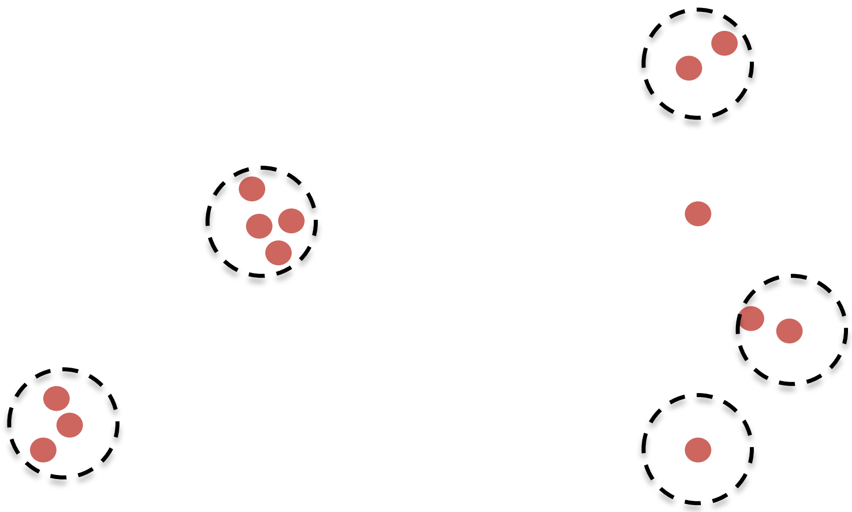
Understanding OTU picking



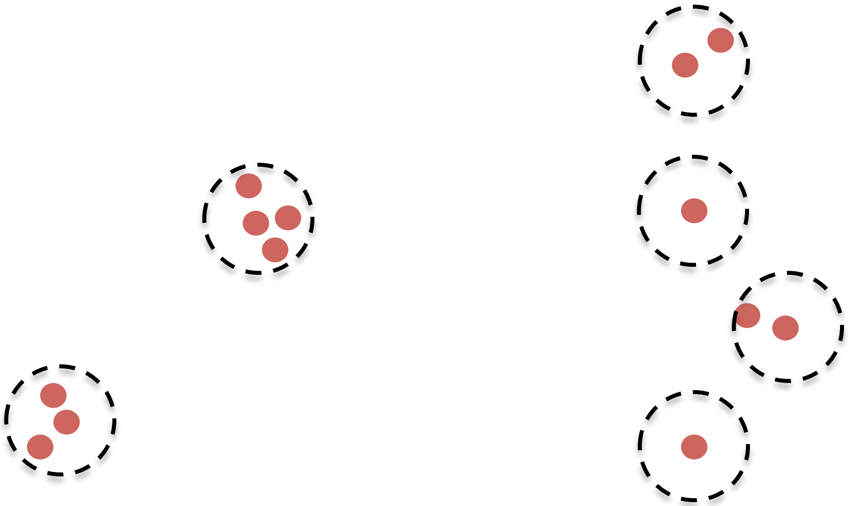
Understanding OTU picking



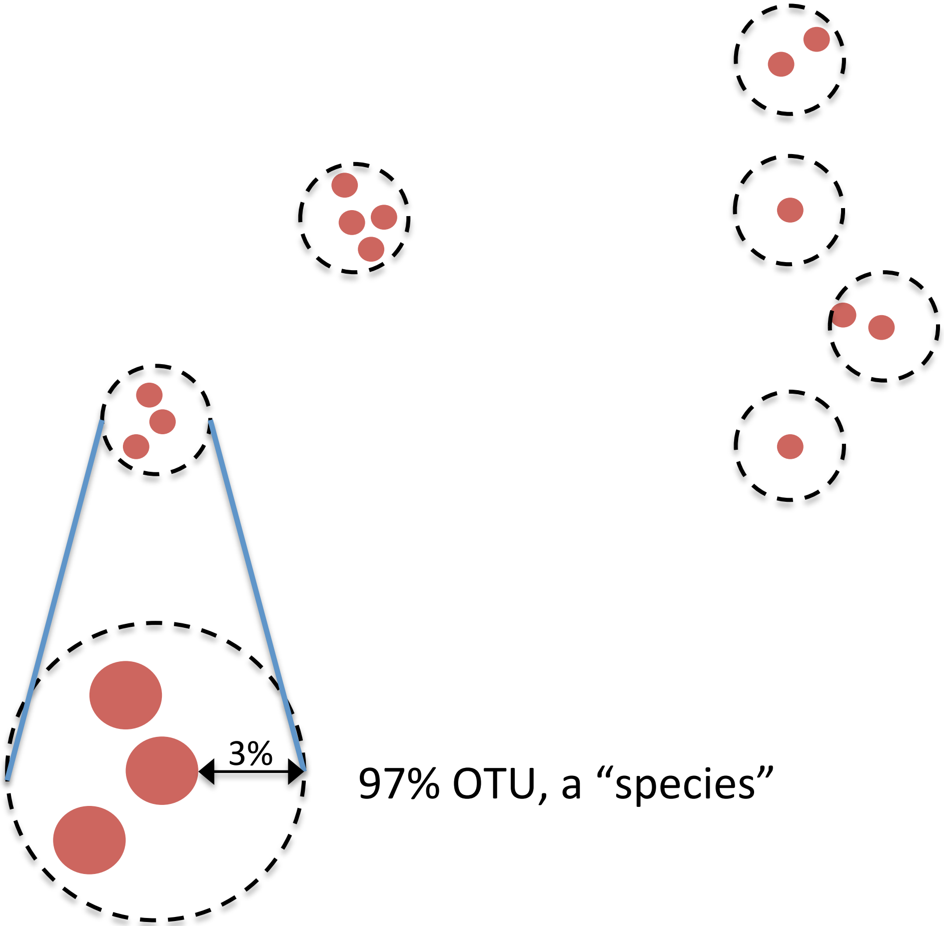
Understanding OTU picking



Understanding OTU picking

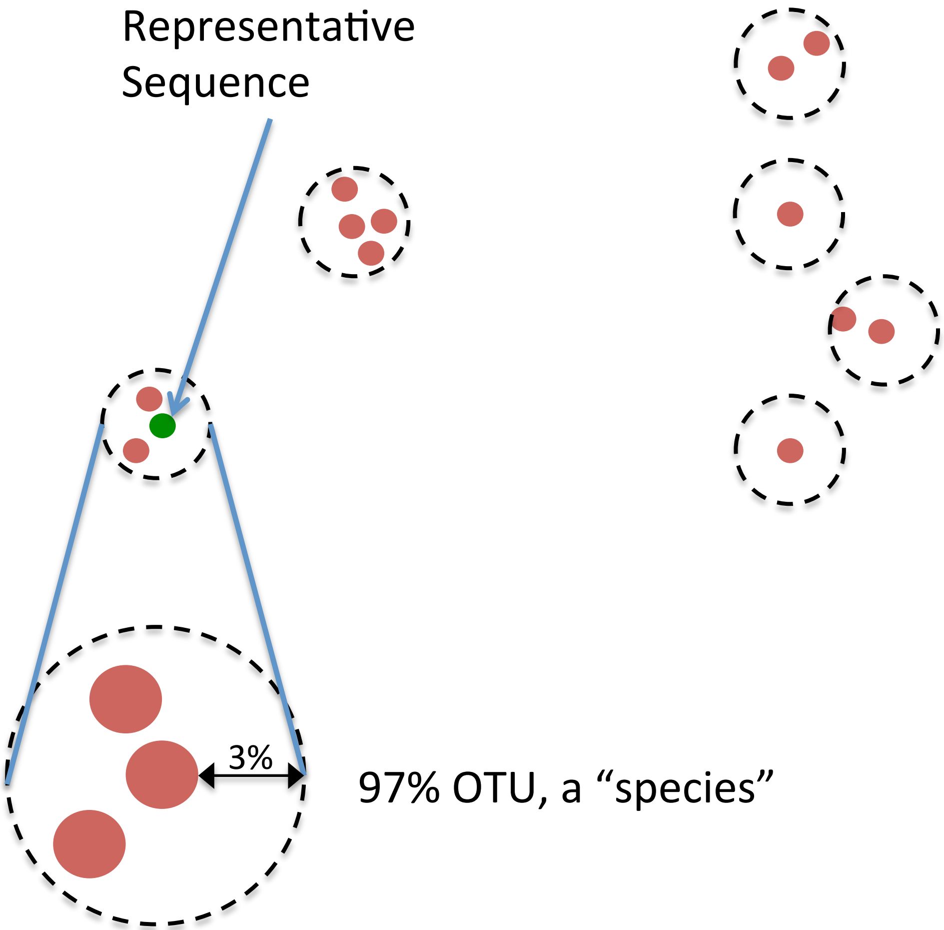


Understanding OTU picking

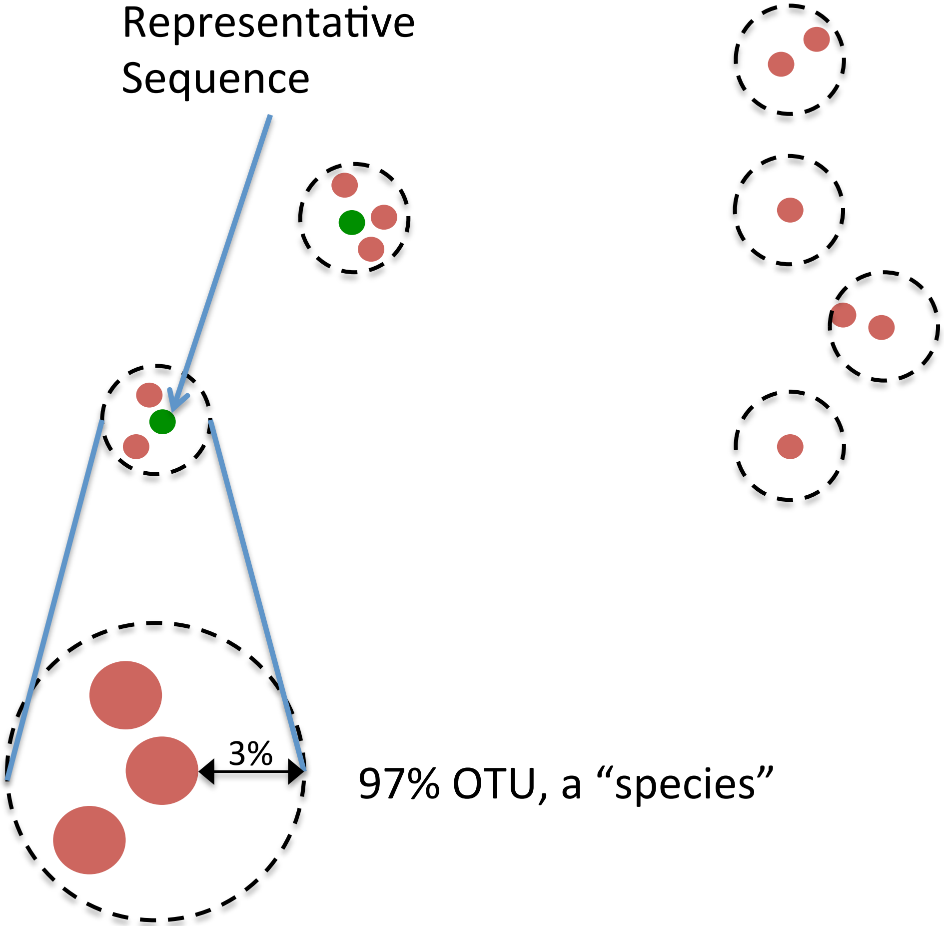


97% OTU, a "species"

Understanding OTU picking



Understanding OTU picking

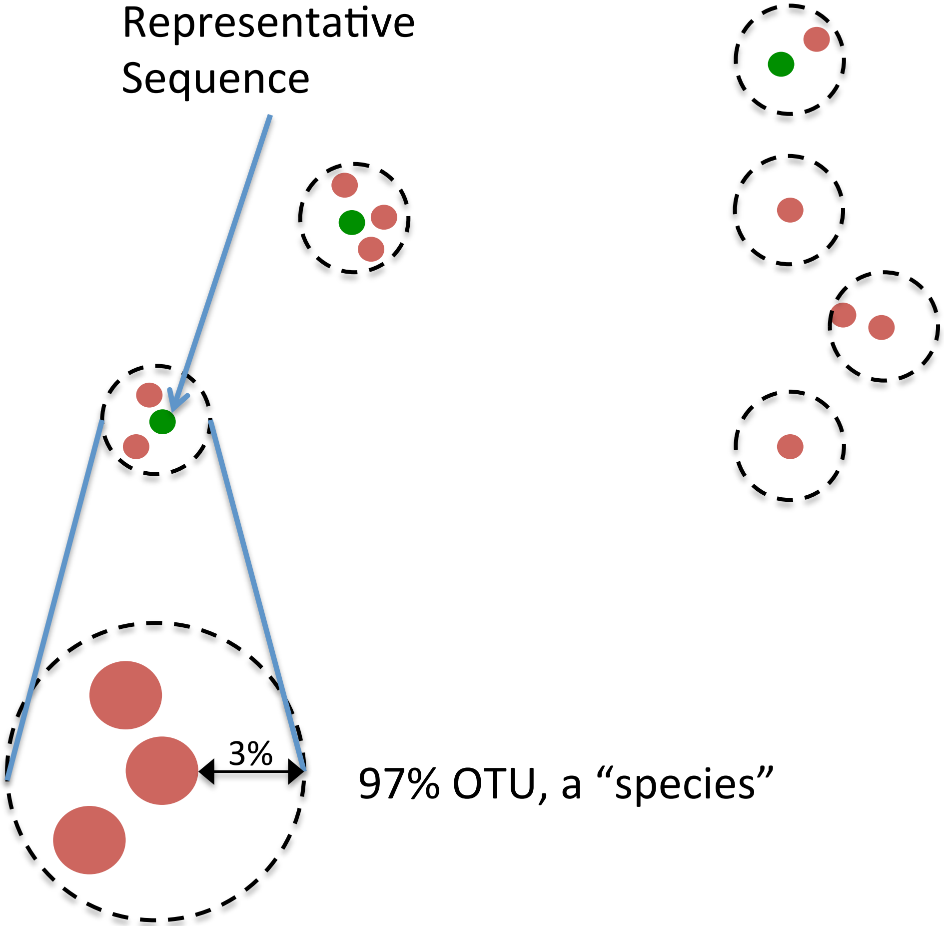


Representative Sequence

3%

97% OTU, a “species”

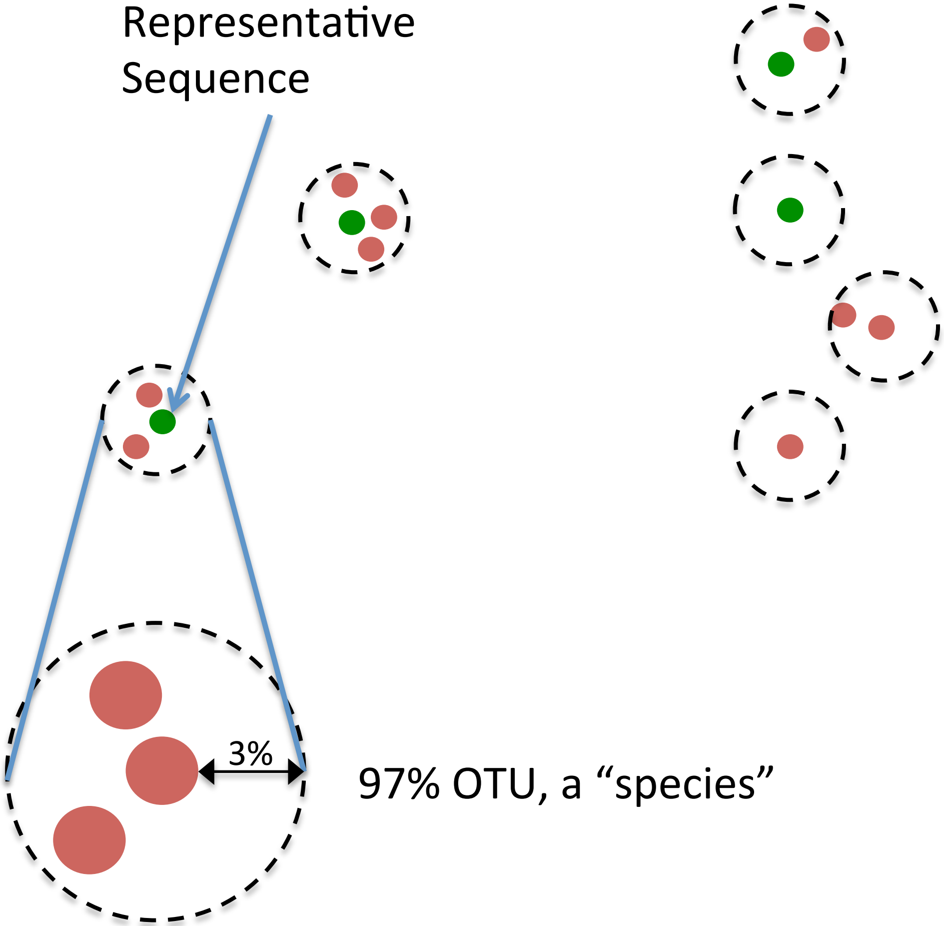
Understanding OTU picking



Representative Sequence

97% OTU, a “species”

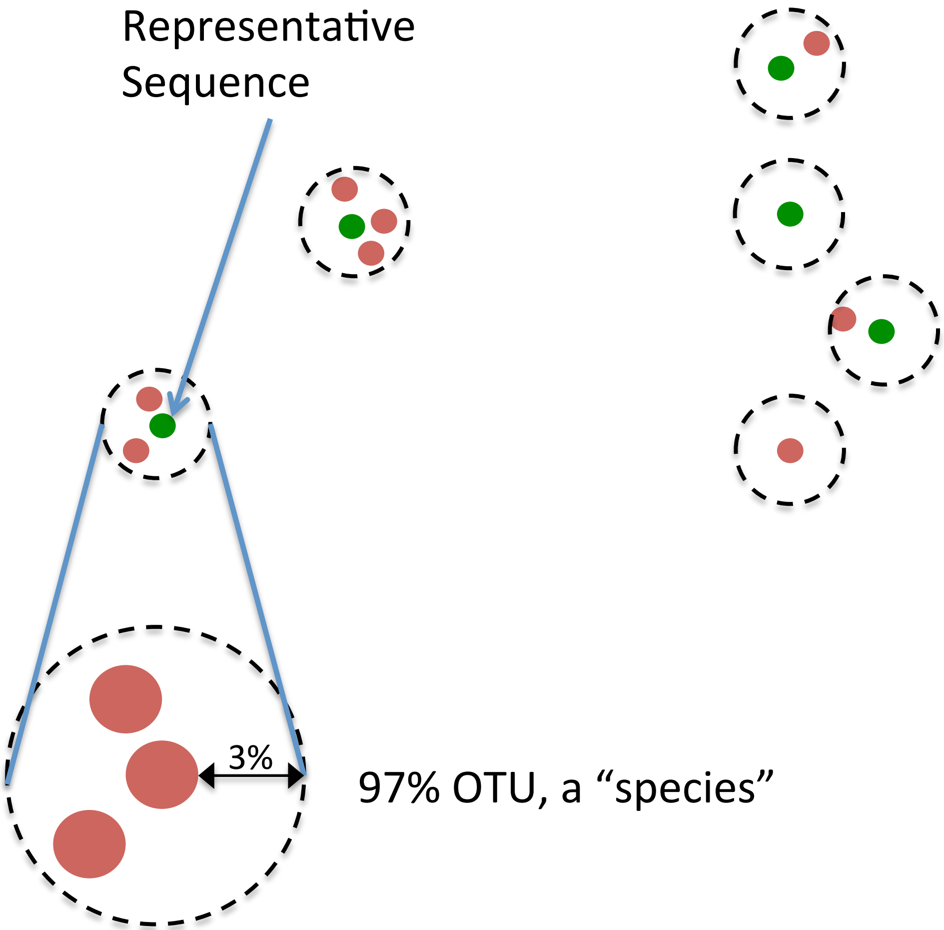
Understanding OTU picking



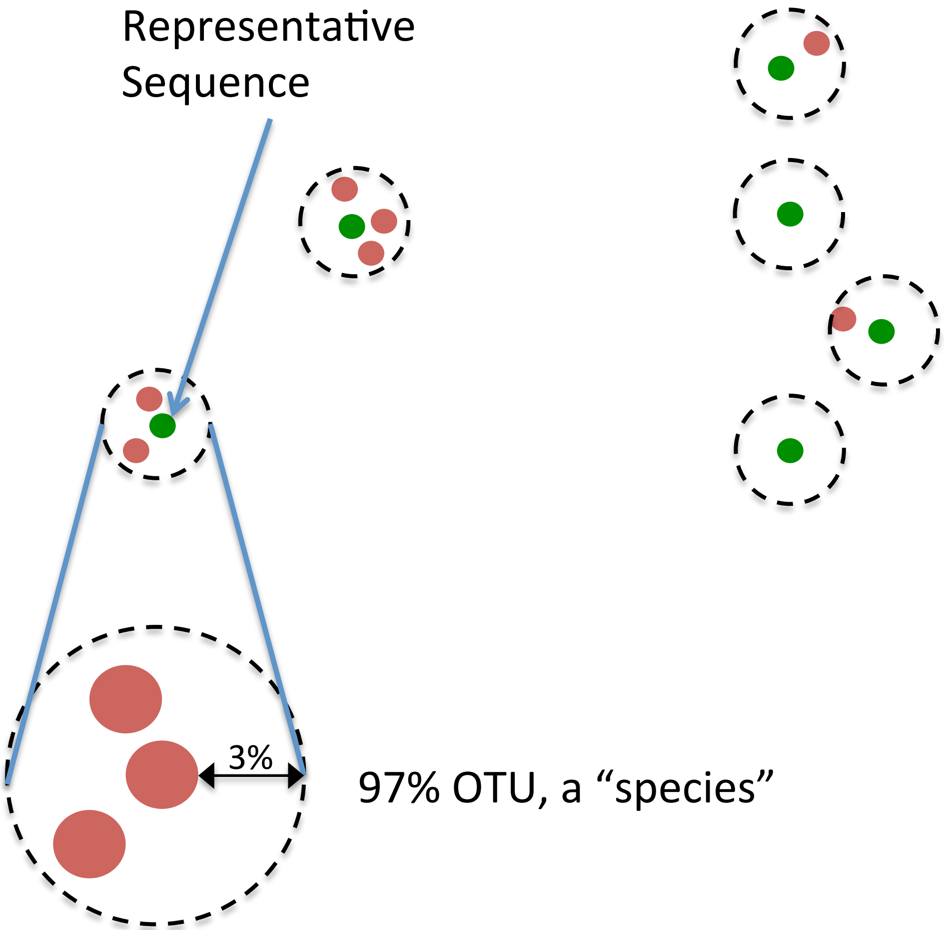
Representative Sequence

97% OTU, a “species”

Understanding OTU picking



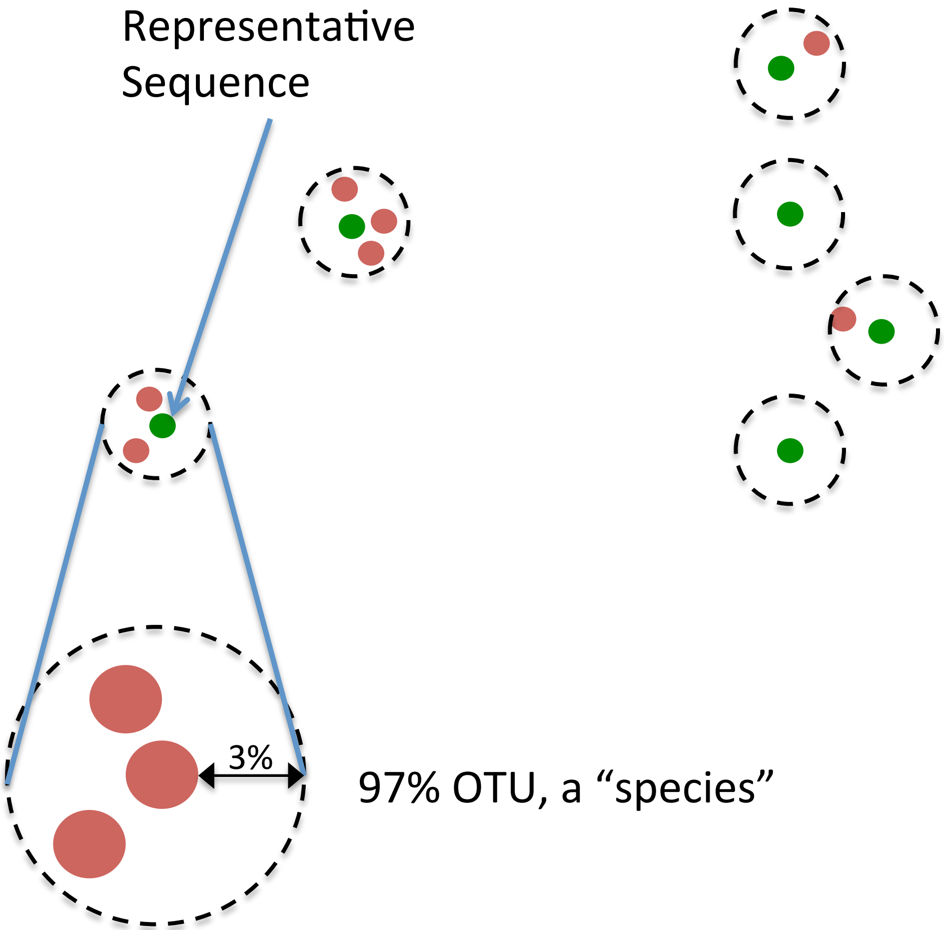
Understanding OTU picking



Representative Sequence

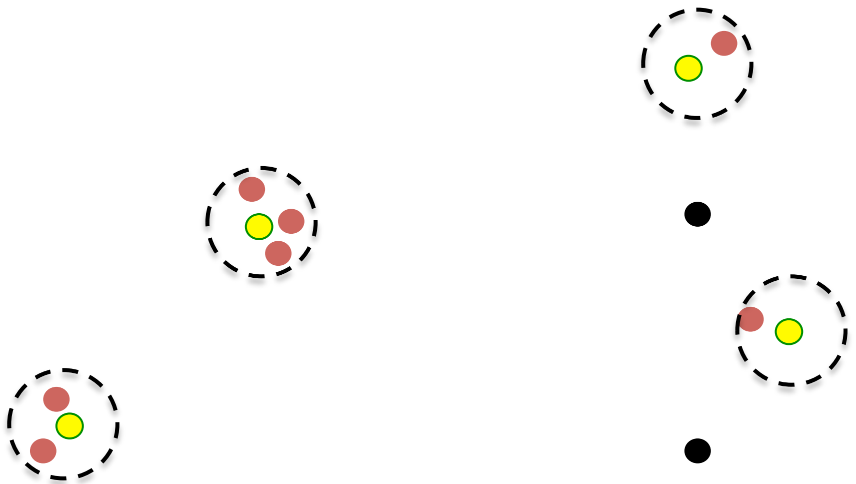
97% OTU, a “species”

Understanding OTU picking

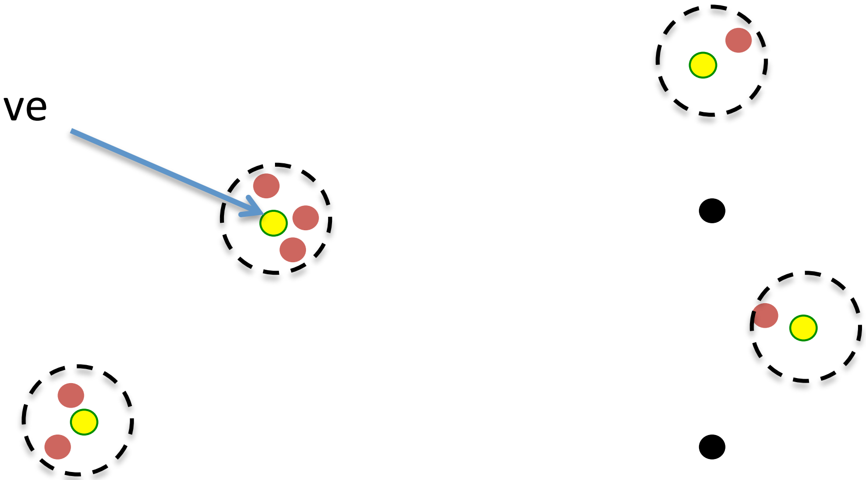


De novo
OTU picking

Understanding OTU picking

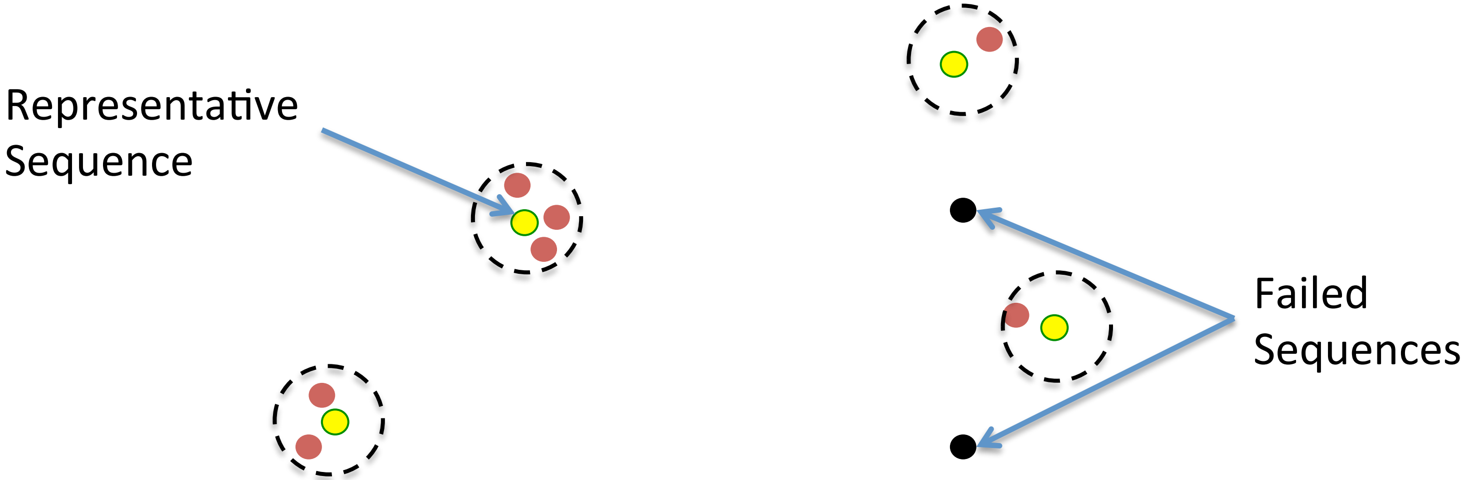


Understanding OTU picking



Representative Sequence

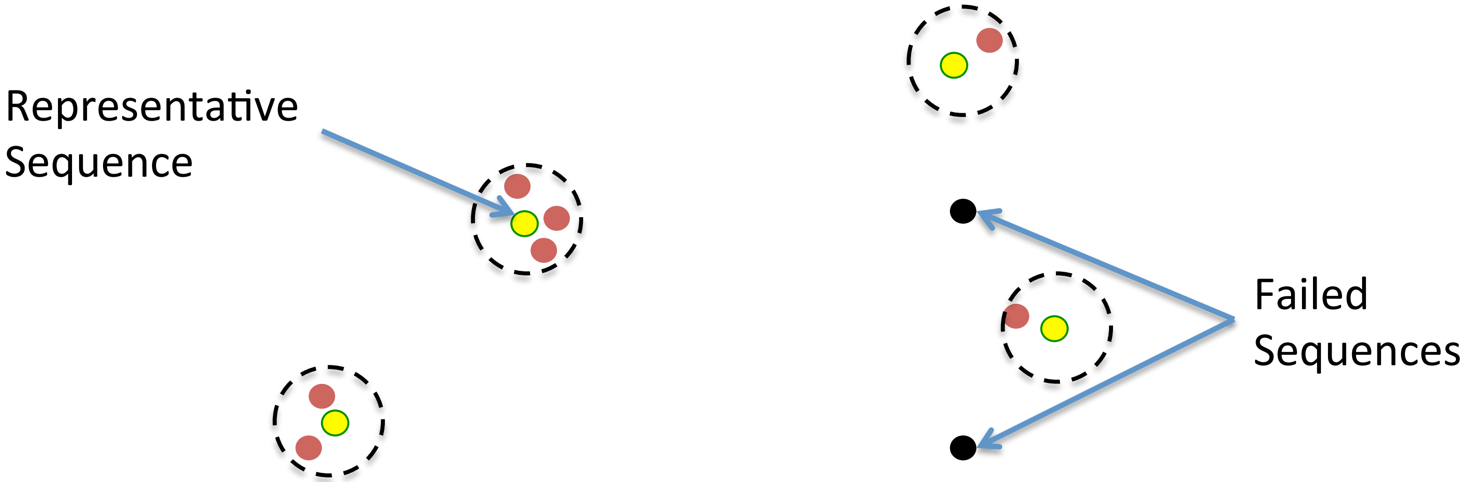
Understanding OTU picking



Representative Sequence

Failed Sequences

Understanding OTU picking

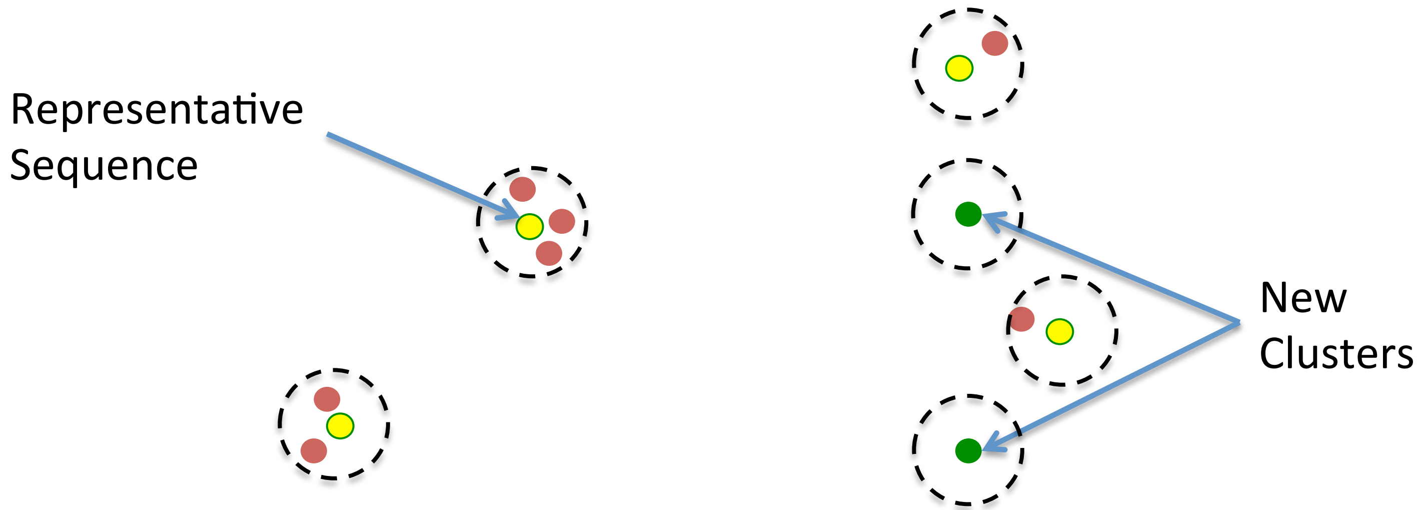


Representative Sequence

Failed Sequences

Closed reference OTU picking

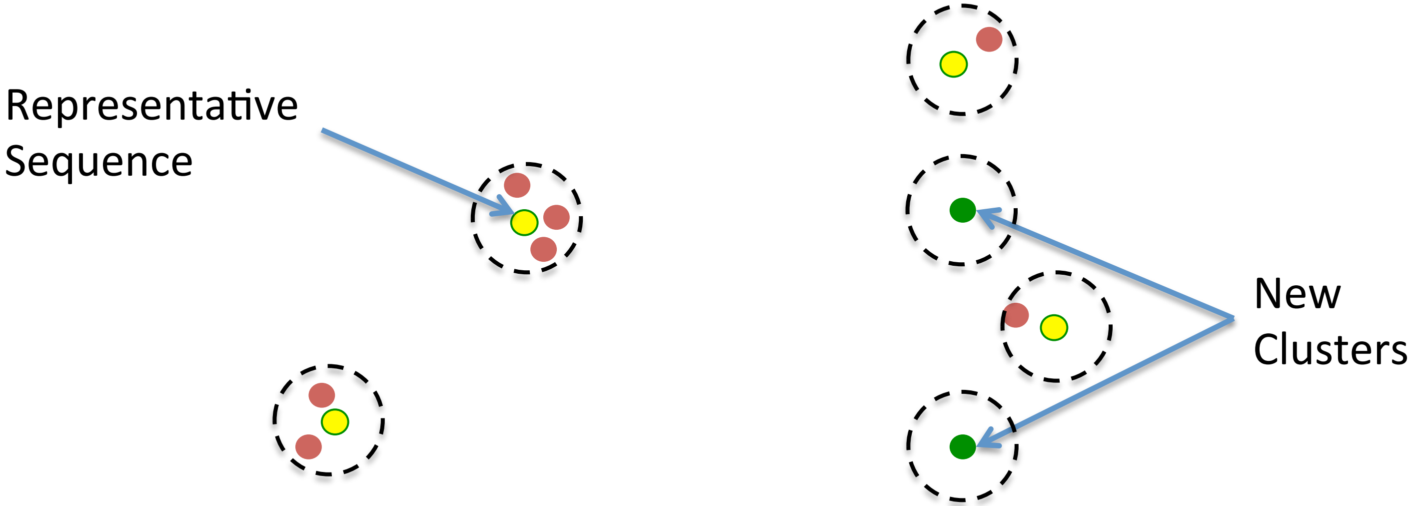
Understanding OTU picking



Representative Sequence

New Clusters

Understanding OTU picking



Representative Sequence

New Clusters

Open reference OTU picking

Understanding OTU picking

- De novo
  - Single study
  - Poor characterization of existing taxa
  - Good characterization of novel taxa
- Closed-reference
  - Combine studies
  - Good characterization of existing taxa
  - Poor characterization of novel taxa
- Open-reference (QIIME default)
  - Single study
  - Good characterization of existing taxa
  - Good characterization of novel taxa

In depth discussion here in Rideout et al 2014 (https://peerj.com/articles/545/)

Pick open reference OTUs

- Workflow script, performs all steps through building an OTU table
  - Determine the OTU clusters
  - Pick the representative sequence for each OTU cluster
  - Align the sequences to a template or other reference alignment
  - Assign taxonomy if performing a de novo OTU picking
  - Remove non-phylogenetically informative positions
  - Construct a phylogeny from an alignment
  - Constructs the actual OTU table object

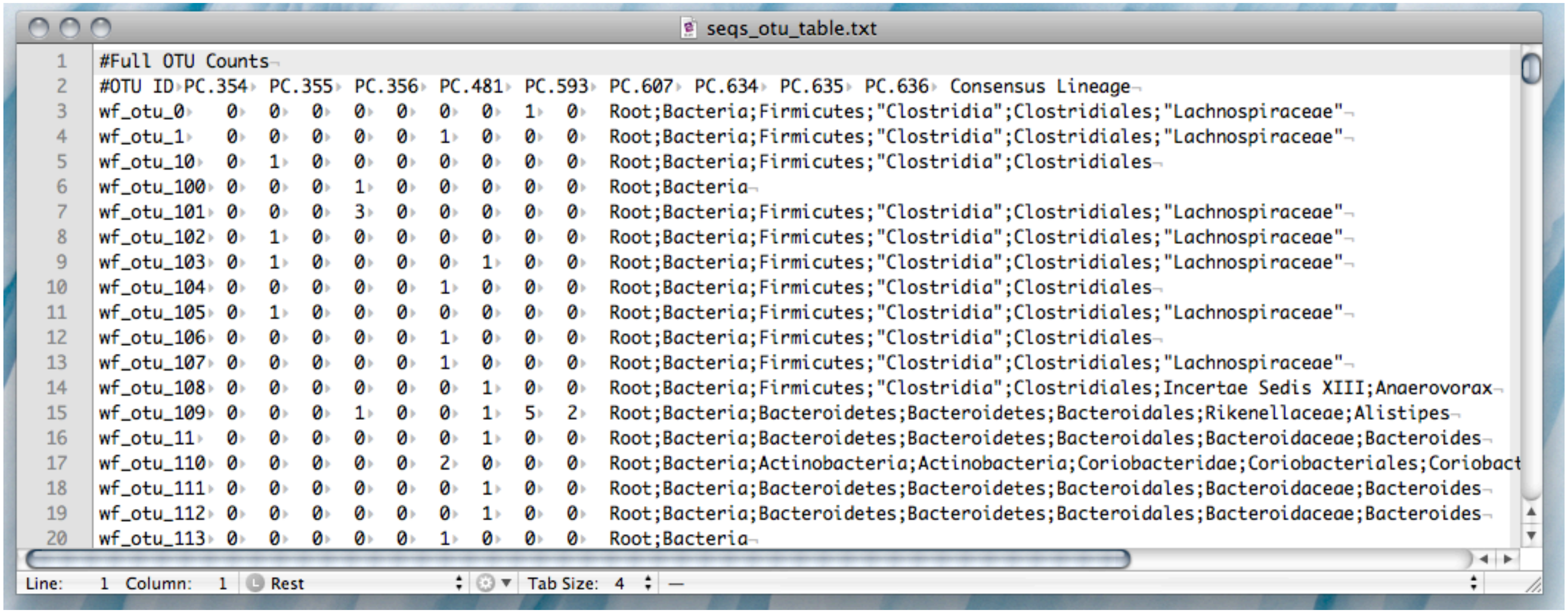
Pick OTUs

- Read through the rest of the text in the “OTU Picking…” section
- Once the “pick_open_reference_otus.py” cell has completed, execute the remaining cells in the section
- Read through and execute the “Create a single mapping file…” section

OTU table

(classic format)

sample x OTU matrix

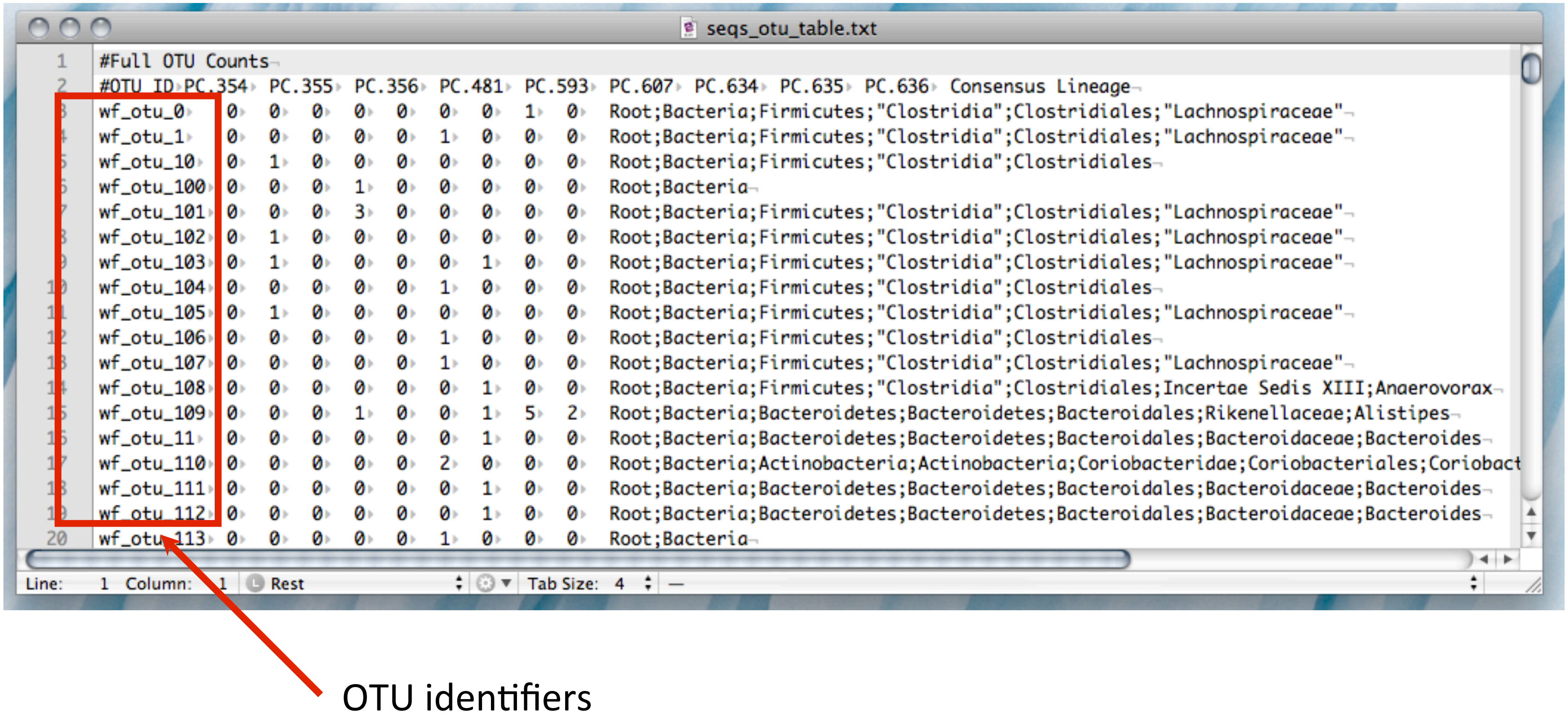


```
#Full OTU Counts
#OTU ID	PC.354	PC.355	PC.356	PC.481	PC.593	PC.607	PC.634	PC.635	PC.636	Consensus Lineage
wf_otu_0	0	0	0	0	0	0	0	1	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_1	0	0	0	0	0	1	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_10	0	1	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales
wf_otu_100	0	0	0	1	0	0	0	0	0	Root;Bacteria
wf_otu_101	0	0	0	3	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_102	0	1	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_103	0	1	0	0	0	0	1	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_104	0	0	0	0	0	1	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales
wf_otu_105	0	1	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_106	0	0	0	0	0	1	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales
wf_otu_107	0	0	0	0	0	1	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_108	0	0	0	0	0	0	1	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;Incertae Sedis XIII;Anaerovorax
wf_otu_109	0	0	0	1	0	0	1	5	2	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Rikenellaceae;Alistipes
wf_otu_11	0	0	0	0	0	0	1	0	0	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides
wf_otu_110	0	0	0	0	0	2	0	0	0	Root;Bacteria;Actinobacteria;Actinobacteria;Coriobacteridae;Coriobacteriales;Coriobact
wf_otu_111	0	0	0	0	0	0	1	0	0	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides
wf_otu_112	0	0	0	0	0	0	1	0	0	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides
wf_otu_113	0	0	0	0	0	1	0	0	0	Root;Bacteria
```

OTU table

(classic format)

sample x OTU matrix



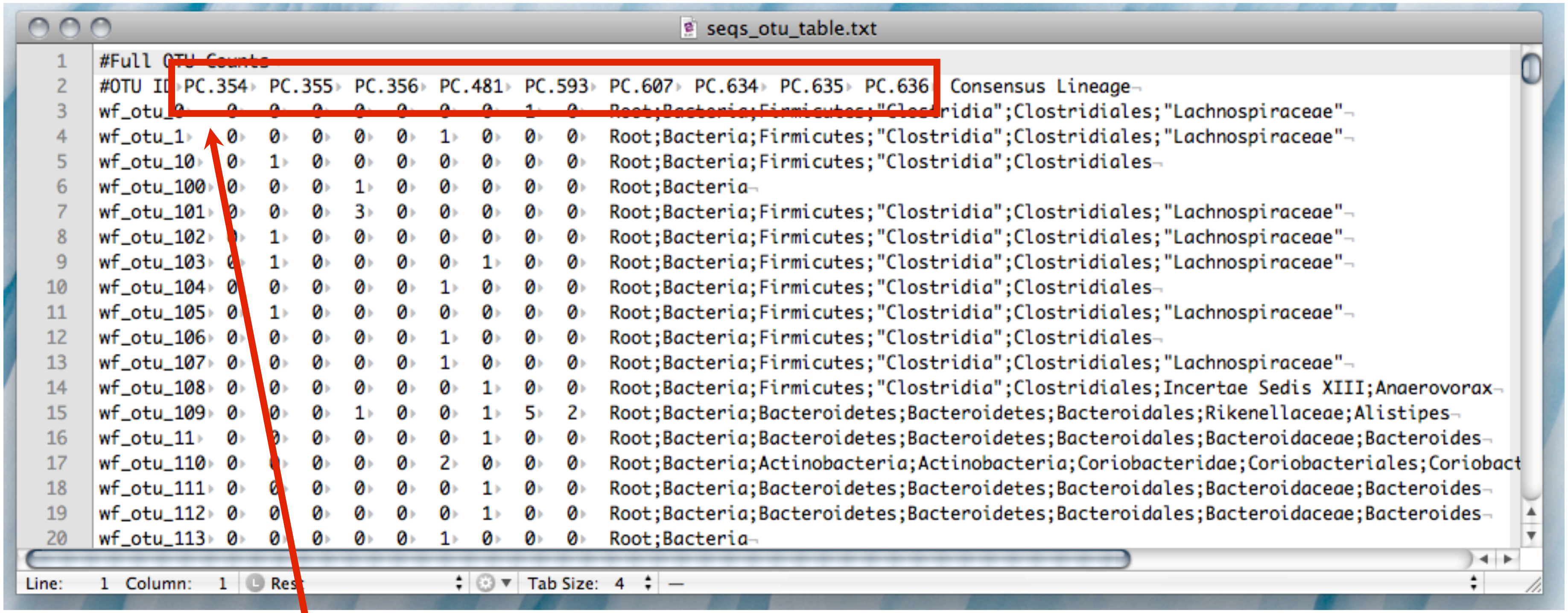
```
#Full OTU Counts
#OTU ID	PC.354	PC.355	PC.356	PC.481	PC.593	PC.607	PC.634	PC.635	PC.636	Consensus Lineage
wf_otu_0	0	0	0	0	0	0	0	1	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_1	0	0	0	0	0	1	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_10	0	1	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales
wf_otu_100	0	0	0	1	0	0	0	0	0	Root;Bacteria
wf_otu_101	0	0	0	3	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_102	0	1	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_103	0	1	0	0	0	0	1	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_104	0	0	0	0	0	1	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales
wf_otu_105	0	1	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_106	0	0	0	0	0	1	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales
wf_otu_107	0	0	0	0	0	1	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_108	0	0	0	0	0	0	1	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;Incertae Sedis XIII;Anaerovorax
wf_otu_109	0	0	0	1	0	0	1	5	2	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Rikenellaceae;Alistipes
wf_otu_11	0	0	0	0	0	0	1	0	0	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides
wf_otu_110	0	0	0	0	0	2	0	0	0	Root;Bacteria;Actinobacteria;Actinobacteria;Coriobacteridae;Coriobacteriales;Coriobact
wf_otu_111	0	0	0	0	0	0	1	0	0	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides
wf_otu_112	0	0	0	0	0	0	1	0	0	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides
wf_otu_113	0	0	0	0	0	1	0	0	0	Root;Bacteria
```

OTU identifiers

OTU table

(classic format)

sample x OTU matrix



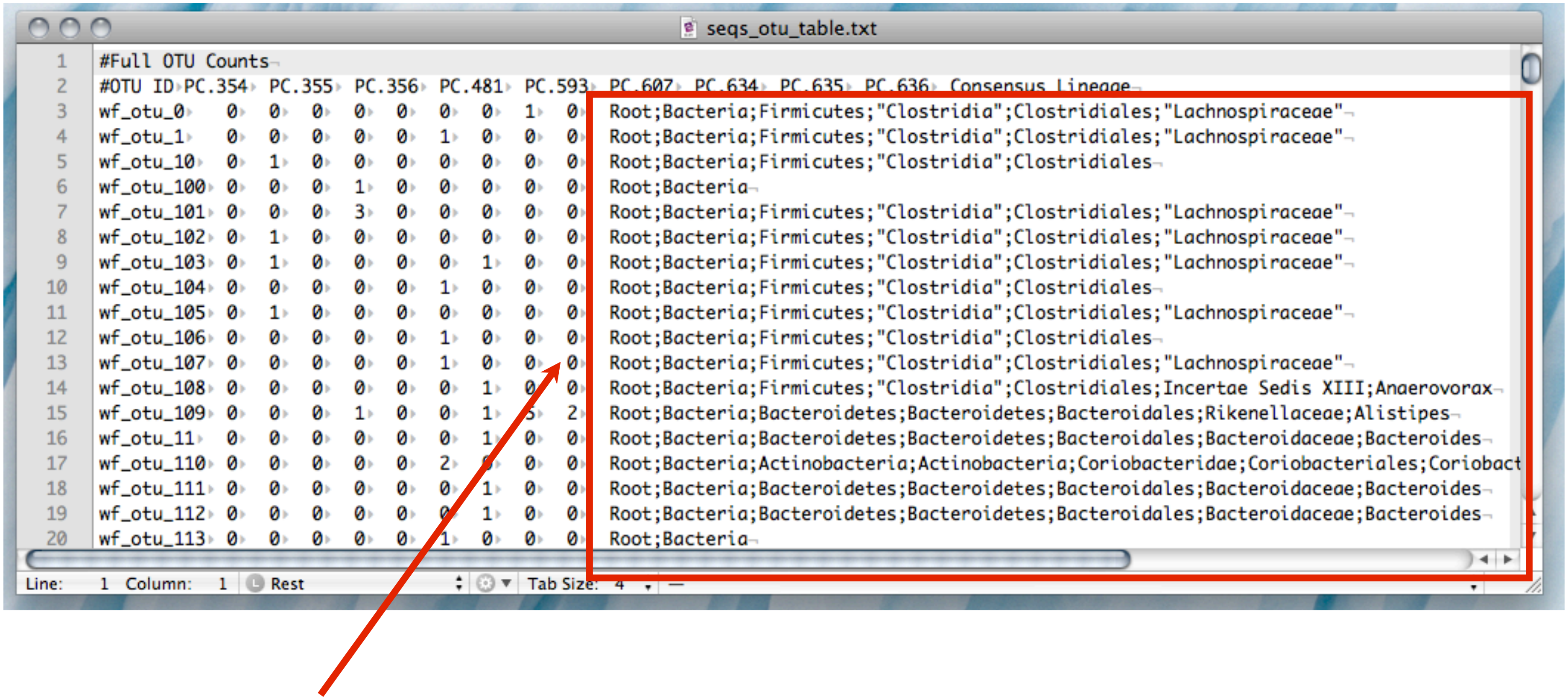
```
#Full OTU Counts
#OTU ID PC.354  PC.355  PC.356  PC.481  PC.593  PC.607  PC.634  PC.635  PC.636  Consensus Lineage
wf_otu_0        0   0   0   0   0   0   0   1   0   Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_1        0   0   0   0   0   1   0   0   0   Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_10       0   1   0   0   0   0   0   0   0   Root;Bacteria;Firmicutes;"Clostridia";Clostridiales
wf_otu_100      0   0   0   1   0   0   0   0   0   Root;Bacteria
wf_otu_101      0   0   0   3   0   0   0   0   0   Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_102      0   1   0   0   0   0   0   0   0   Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_103      0   1   0   0   0   0   1   0   0   Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_104      0   0   0   0   0   1   0   0   0   Root;Bacteria;Firmicutes;"Clostridia";Clostridiales
wf_otu_105      0   1   0   0   0   0   0   0   0   Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_106      0   0   0   0   0   1   0   0   0   Root;Bacteria;Firmicutes;"Clostridia";Clostridiales
wf_otu_107      0   0   0   0   0   1   0   0   0   Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_108      0   0   0   0   0   0   1   0   0   Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;Incertae Sedis XIII;Anaerovorax
wf_otu_109      0   0   0   1   0   0   1   5   2   Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Rikenellaceae;Alistipes
wf_otu_11       0   0   0   0   0   0   1   0   0   Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides
wf_otu_110      0   0   0   0   0   2   0   0   0   Root;Bacteria;Actinobacteria;Actinobacteria;Coriobacteridae;Coriobacteriales;Coriobact
wf_otu_111      0   0   0   0   0   0   1   0   0   Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides
wf_otu_112      0   0   0   0   0   0   1   0   0   Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides
wf_otu_113      0   0   0   0   0   1   0   0   0   Root;Bacteria
```

Sample identifiers

OTU table

(classic format)

sample x OTU matrix



```
#Full OTU Counts
#OTU ID	PC.354	PC.355	PC.356	PC.481	PC.593	PC.607	PC.634	PC.635	PC.636	Consensus Lineage
wf_otu_0	0	0	0	0	0	0	0	1	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_1	0	0	0	0	0	1	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_10	0	1	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales
wf_otu_100	0	0	0	1	0	0	0	0	0	Root;Bacteria
wf_otu_101	0	0	0	3	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_102	0	1	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_103	0	1	0	0	0	0	1	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_104	0	0	0	0	0	1	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales
wf_otu_105	0	1	0	0	0	0	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_106	0	0	0	0	0	1	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales
wf_otu_107	0	0	0	0	0	1	0	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_108	0	0	0	0	0	0	1	0	0	Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;Incertae Sedis XIII;Anaerovorax
wf_otu_109	0	0	0	1	0	0	1	5	2	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Rikenellaceae;Alistipes
wf_otu_11	0	0	0	0	0	0	1	0	0	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides
wf_otu_110	0	0	0	0	0	2	0	0	0	Root;Bacteria;Actinobacteria;Actinobacteria;Coriobacteridae;Coriobacteriales;Coriobact
wf_otu_111	0	0	0	0	0	0	1	0	0	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides
wf_otu_112	0	0	0	0	0	0	1	0	0	Root;Bacteria;Bacteroidetes;Bacteroidetes;Bacteroidales;Bacteroidaceae;Bacteroides
wf_otu_113	0	0	0	0	0	1	0	0	0	Root;Bacteria
```

Optional per OTU taxonomic information

Biological Observation Matrix

- http://biom-format.org
- Genomic Standards Consortium standard format

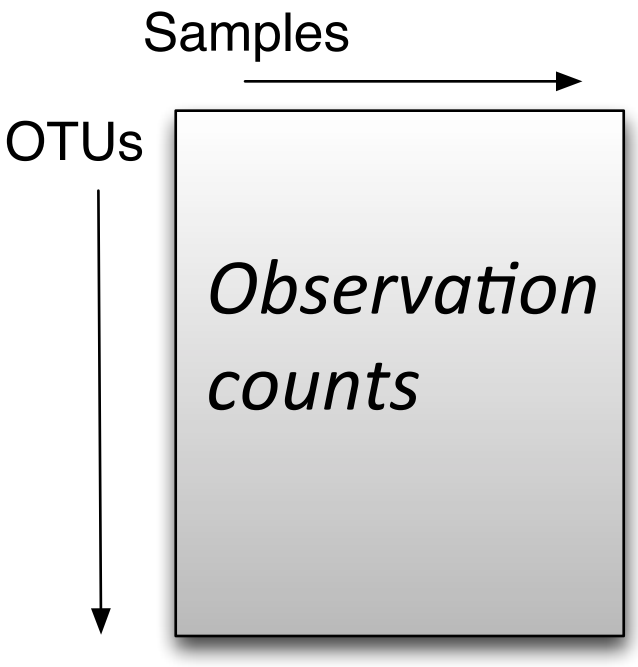
MG-RAST
metagenomics analysis server

qiime
Quantitative Insights Into Microbial Ecology

VAMPS
The Visualization and Analysis
of Microbial Population Structures

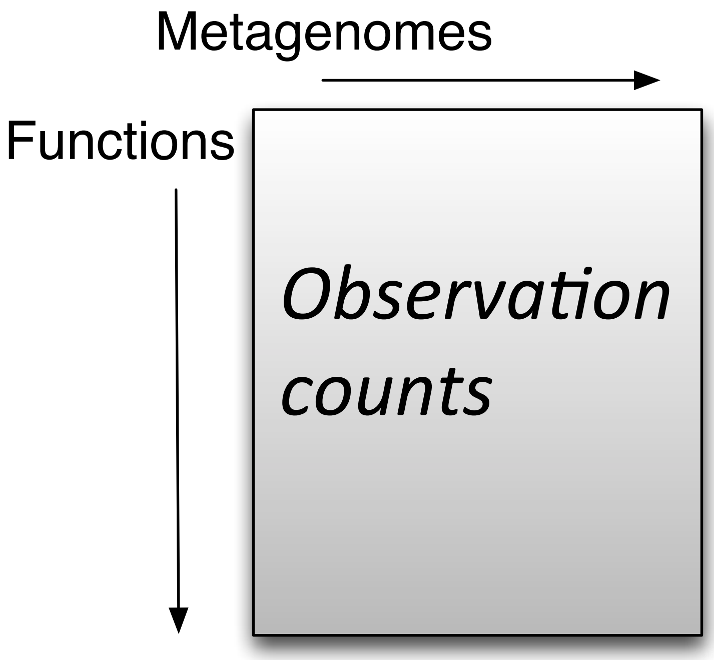
The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. McDonald, Clemente, et al. Gigascience 2012 1:7

sample x observation contingency matrix



qiime
Quantitative Insights Into Microbial Ecology

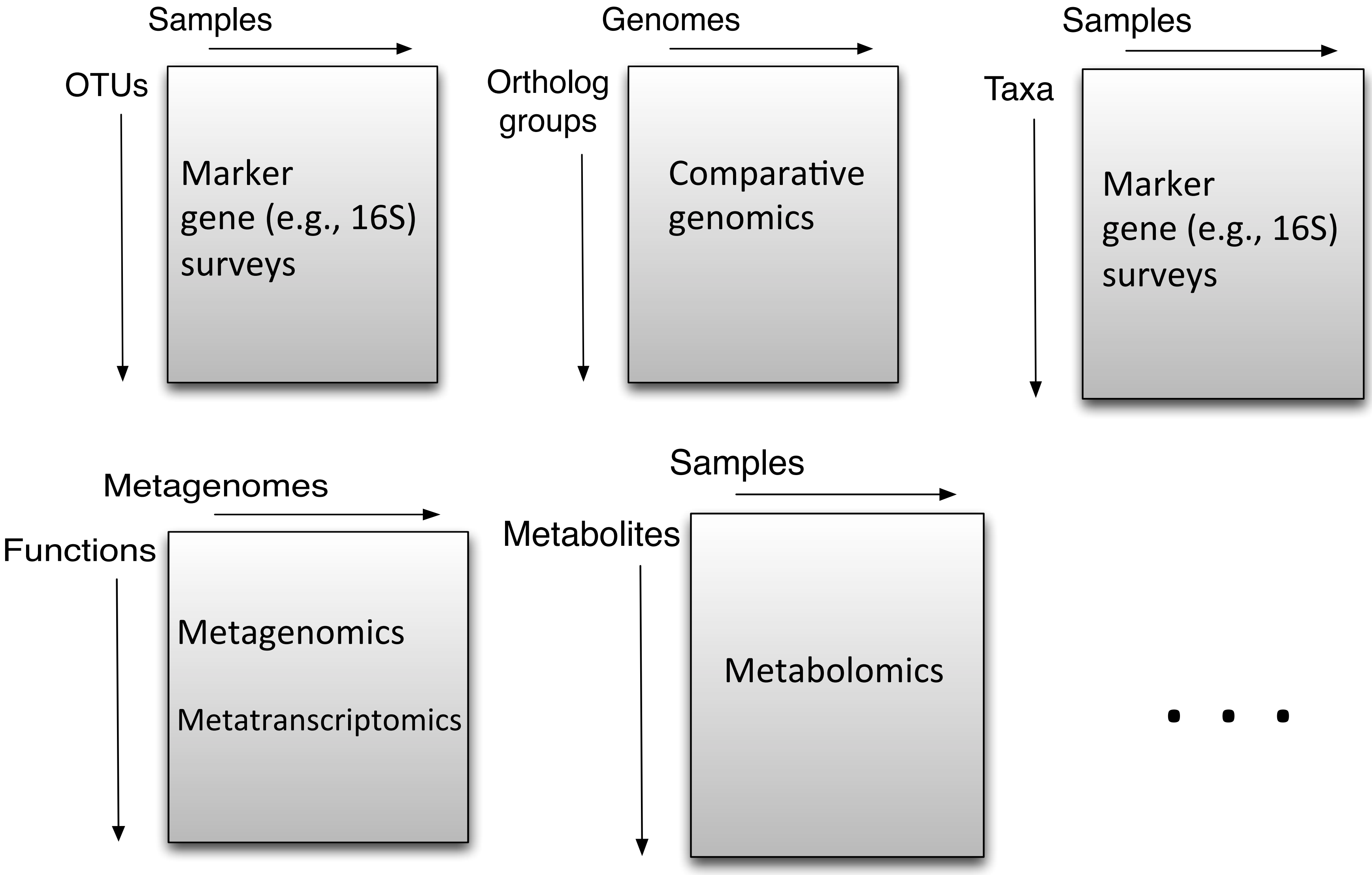
sample x observation contingency matrix



MG-RAST

metagenomics analysis server

sample x observation contingency matrix



Samples

OTUs

Marker gene (e.g., 16S) surveys

Genomes

Ortholog groups

Comparative genomics

Samples

Taxa

Marker gene (e.g., 16S) surveys

Metagenomes

Functions

Metagenomics

Metatranscriptomics

Samples

Metabolites

Metabolomics

. . .

http://biom-format.org/

Biological
Observation
Matrix

Tab Delimited Files

{ JSON }

HDF5

HDF5

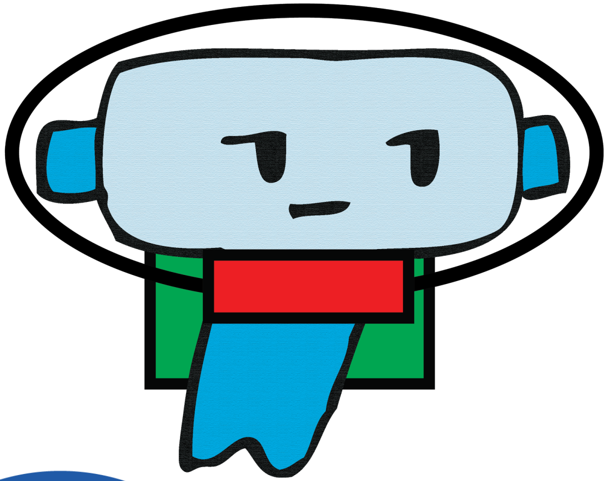
- Blazing fast!
- Space efficient.
- Widely used in scientific fields

http://home.web.cern.ch/

http://www.nasa.gov/

HDF5

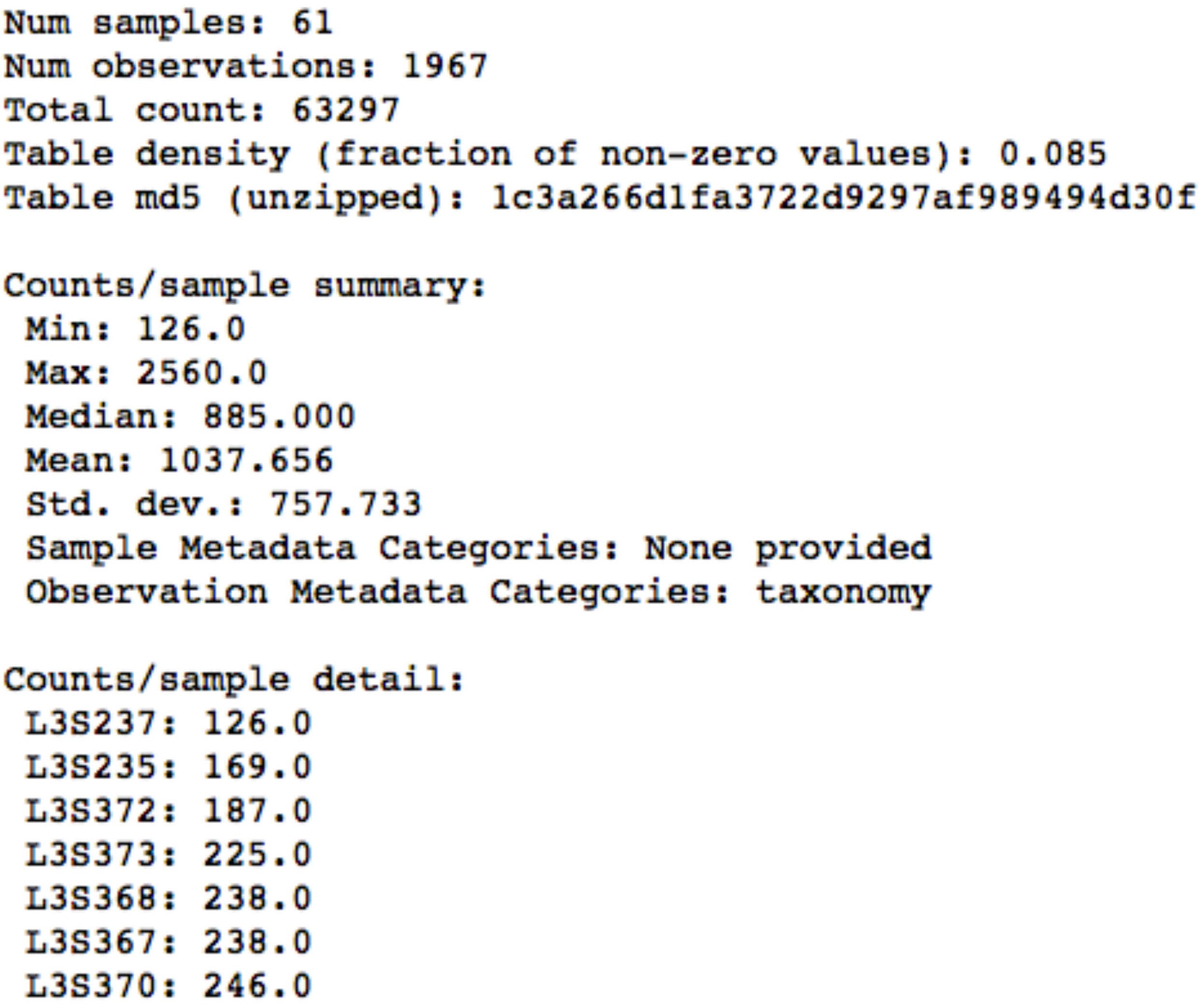
- Blazing fast!
- Space efficient.
- Widely used in scientific fields



http://home.web.cern.ch/

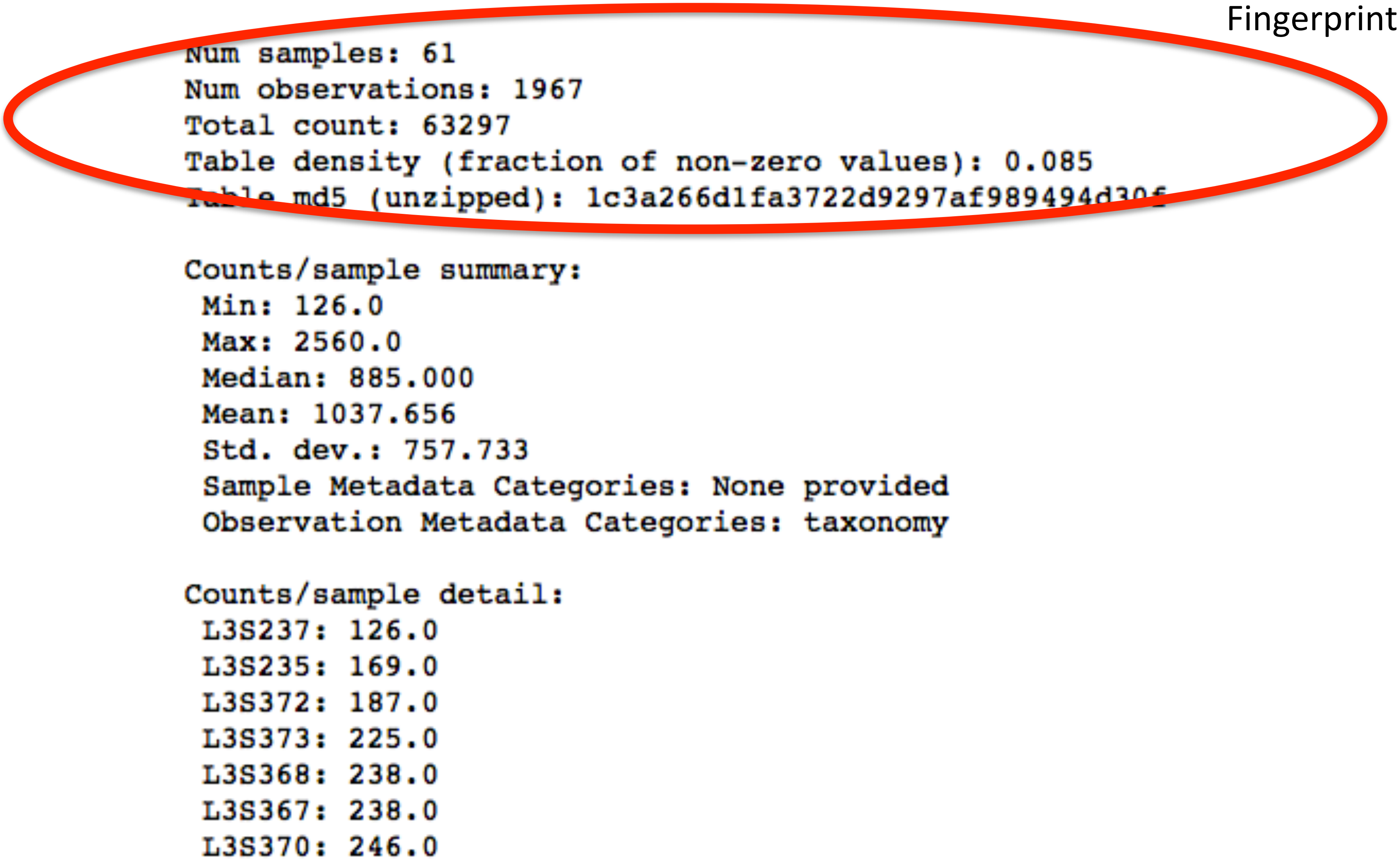
http://www.nasa.gov/

biom summarize-table



```
Num samples: 61
Num observations: 1967
Total count: 63297
Table density (fraction of non-zero values): 0.085
Table md5 (unzipped): 1c3a266d1fa3722d9297af989494d30f

Counts/sample summary:
 Min: 126.0
 Max: 2560.0
 Median: 885.000
 Mean: 1037.656
 Std. dev.: 757.733
 Sample Metadata Categories: None provided
 Observation Metadata Categories: taxonomy

Counts/sample detail:
 L3S237: 126.0
 L3S235: 169.0
 L3S372: 187.0
 L3S373: 225.0
 L3S368: 238.0
 L3S367: 238.0
 L3S370: 246.0
```

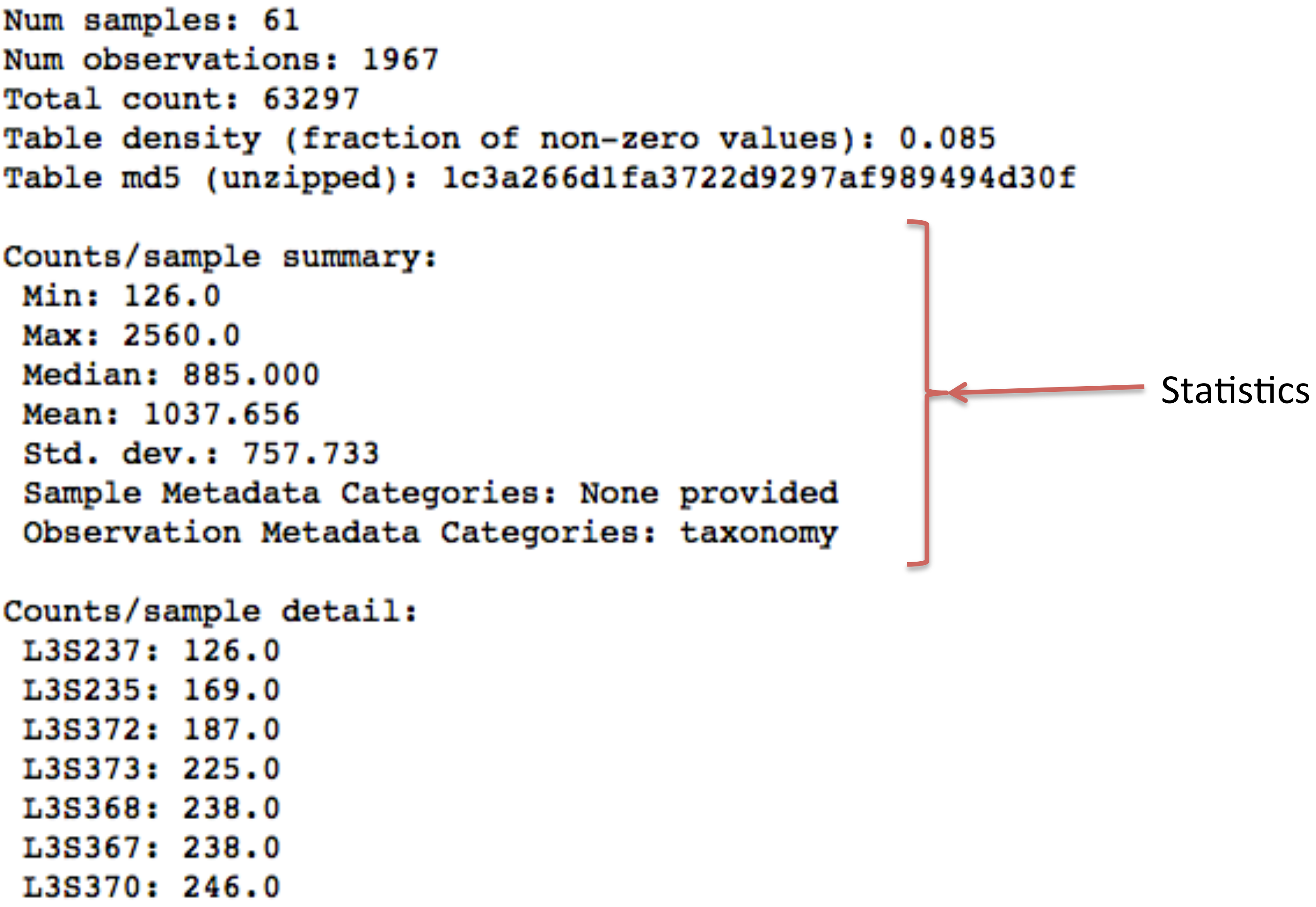
biom summarize-table



Fingerprint

```
Num samples: 61
Num observations: 1967
Total count: 63297
Table density (fraction of non-zero values): 0.085
Table md5 (unzipped): 1c3a266d1fa3722d9297af989494d30f

Counts/sample summary:
 Min: 126.0
 Max: 2560.0
 Median: 885.000
 Mean: 1037.656
 Std. dev.: 757.733
 Sample Metadata Categories: None provided
 Observation Metadata Categories: taxonomy

Counts/sample detail:
 L3S237: 126.0
 L3S235: 169.0
 L3S372: 187.0
 L3S373: 225.0
 L3S368: 238.0
 L3S367: 238.0
 L3S370: 246.0
```

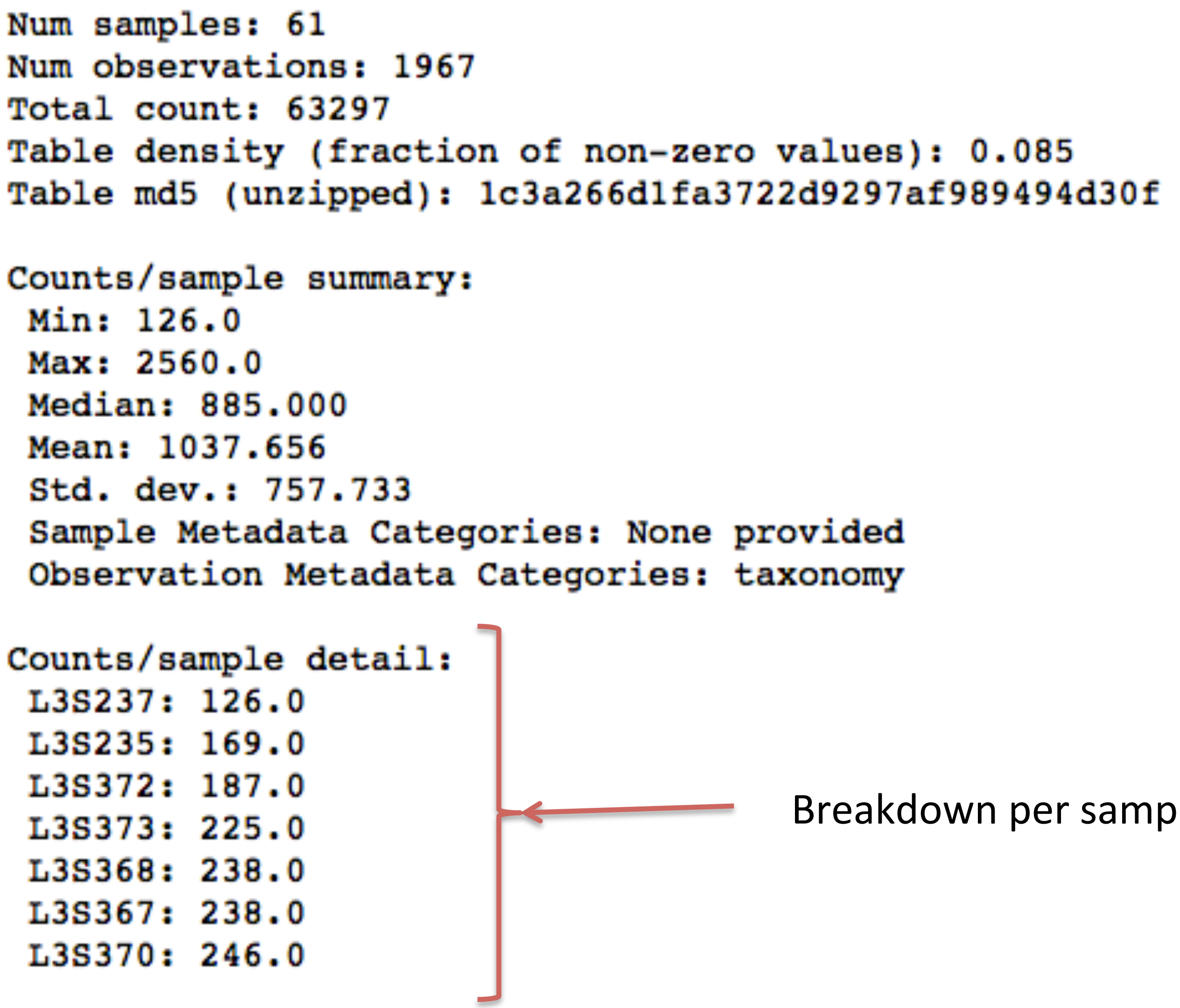
biom summarize-table



```
Num samples: 61
Num observations: 1967
Total count: 63297
Table density (fraction of non-zero values): 0.085
Table md5 (unzipped): 1c3a266d1fa3722d9297af989494d30f

Counts/sample summary:
 Min: 126.0
 Max: 2560.0
 Median: 885.000
 Mean: 1037.656
 Std. dev.: 757.733
 Sample Metadata Categories: None provided
 Observation Metadata Categories: taxonomy

Counts/sample detail:
 L3S237: 126.0
 L3S235: 169.0
 L3S372: 187.0
 L3S373: 225.0
 L3S368: 238.0
 L3S367: 238.0
 L3S370: 246.0
```

Statistics

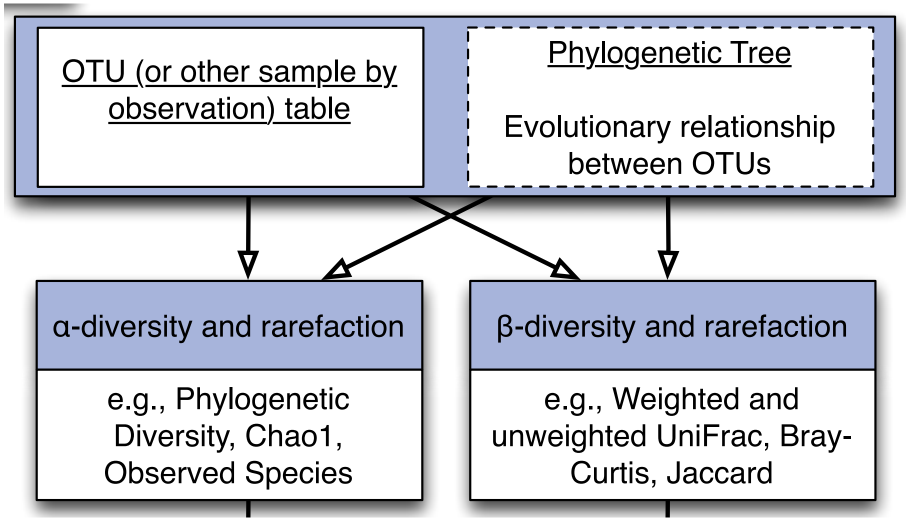
biom summarize-table



```
Num samples: 61
Num observations: 1967
Total count: 63297
Table density (fraction of non-zero values): 0.085
Table md5 (unzipped): 1c3a266d1fa3722d9297af989494d30f

Counts/sample summary:
 Min: 126.0
 Max: 2560.0
 Median: 885.000
 Mean: 1037.656
 Std. dev.: 757.733
 Sample Metadata Categories: None provided
 Observation Metadata Categories: taxonomy

Counts/sample detail:
 L3S237: 126.0
 L3S235: 169.0
 L3S372: 187.0
 L3S373: 225.0
 L3S368: 238.0
 L3S367: 238.0
 L3S370: 246.0
```

Breakdown per sample

Computing alpha and beta diversity



OTU (or other sample by observation) table

Phylogenetic Tree

Evolutionary relationship between OTUs

α-diversity and rarefaction

e.g., Phylogenetic Diversity, Chao1, Observed Species

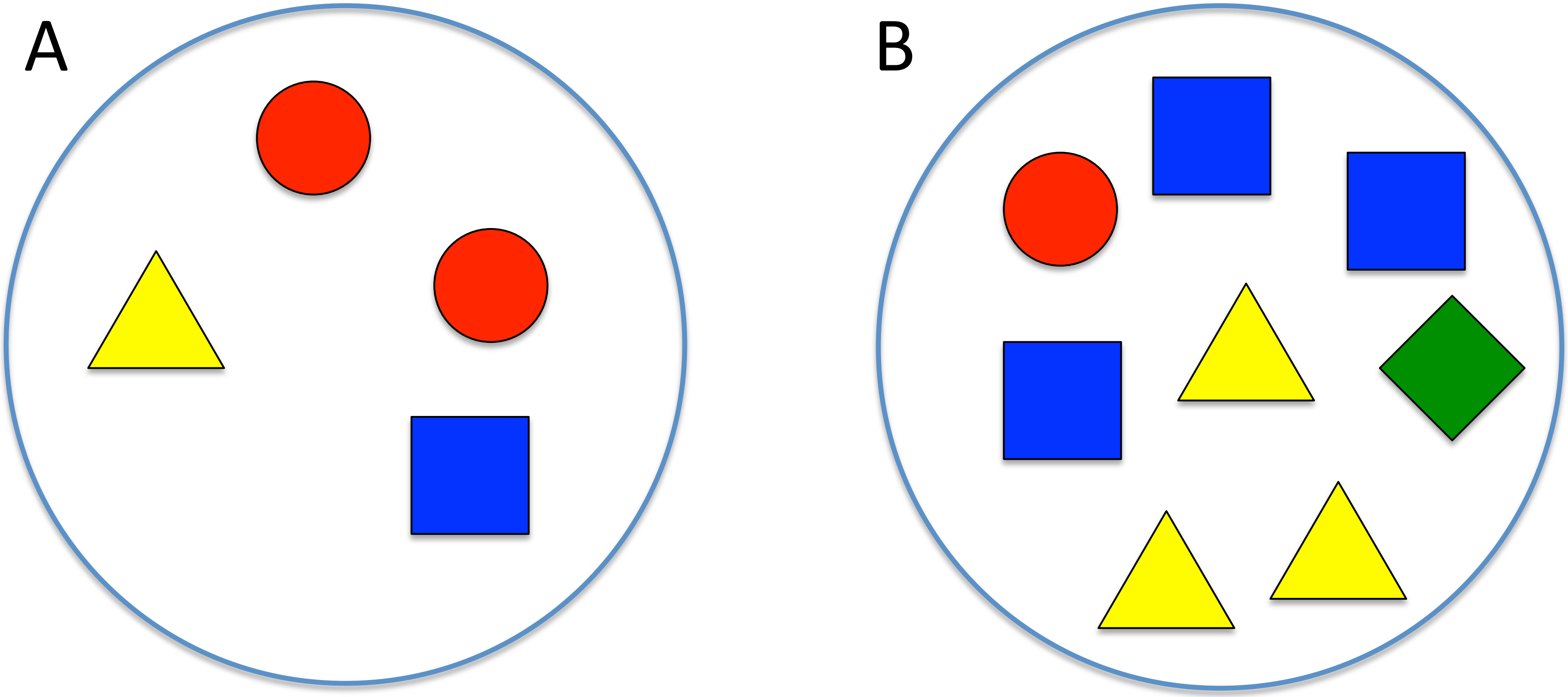
β-diversity and rarefaction

e.g., Weighted and unweighted UniFrac, Bray-Curtis, Jaccard

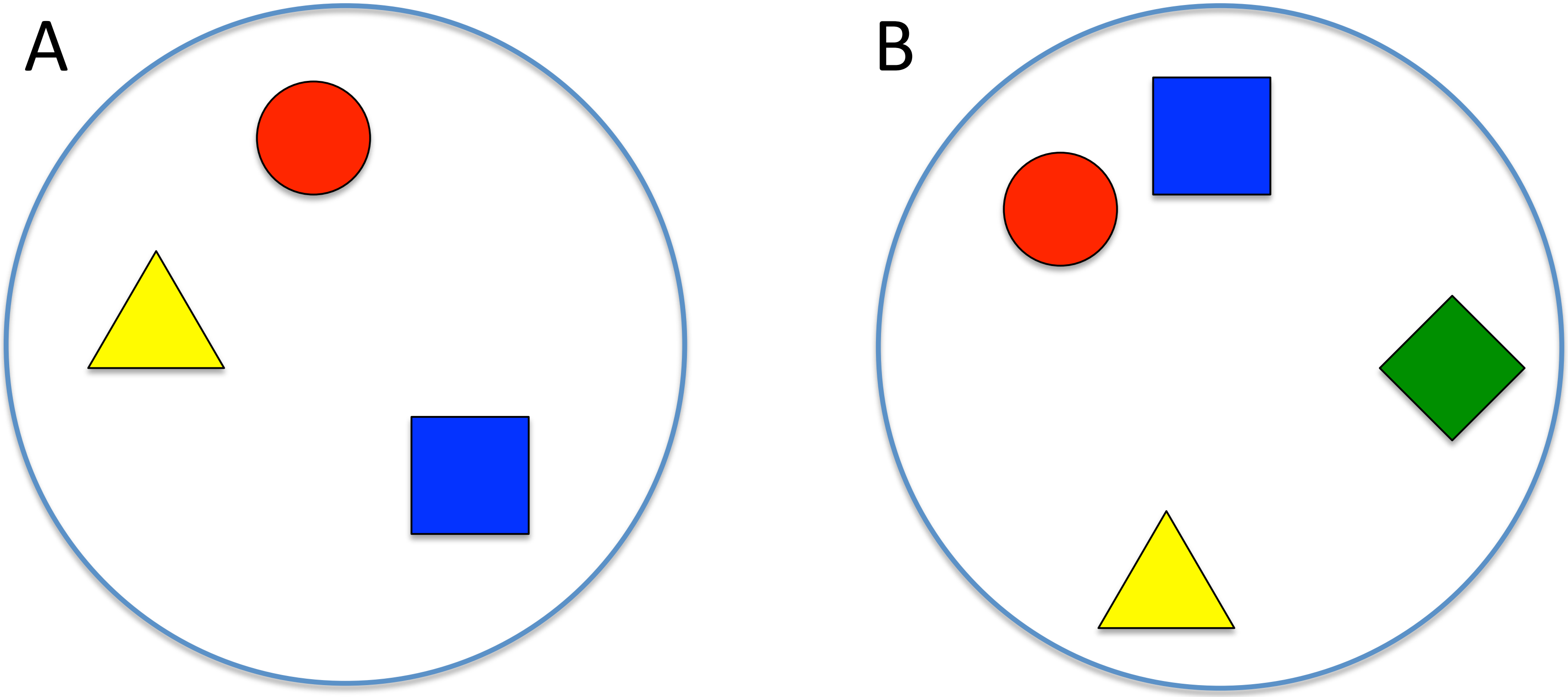
Alpha and beta diversity

- Alpha is within a sample
  - E.g., how many species are in a sample
- Beta is between samples
  - E.g., how similar are two samples
- Lots of ways to calculate these

Alpha diversity



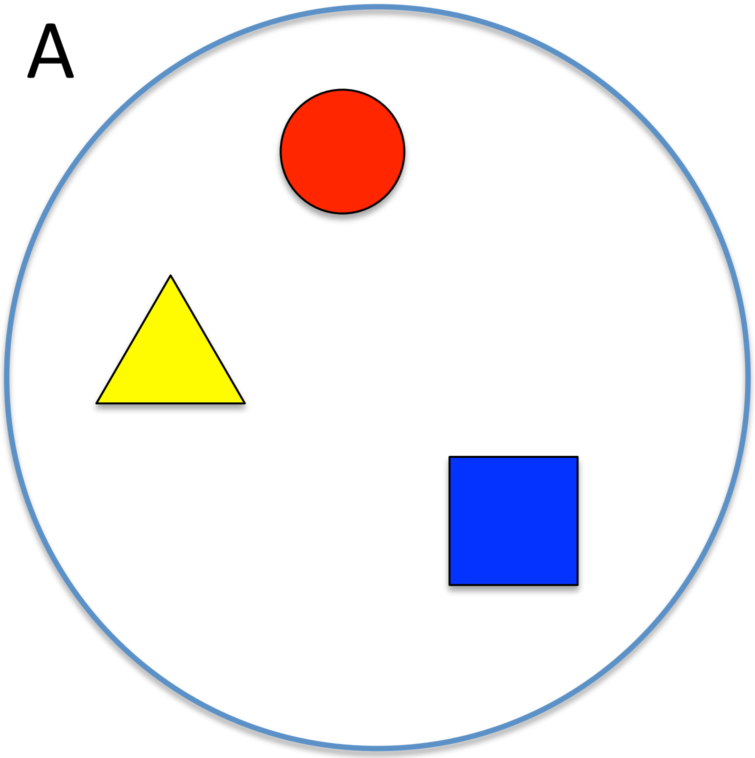
Alpha diversity



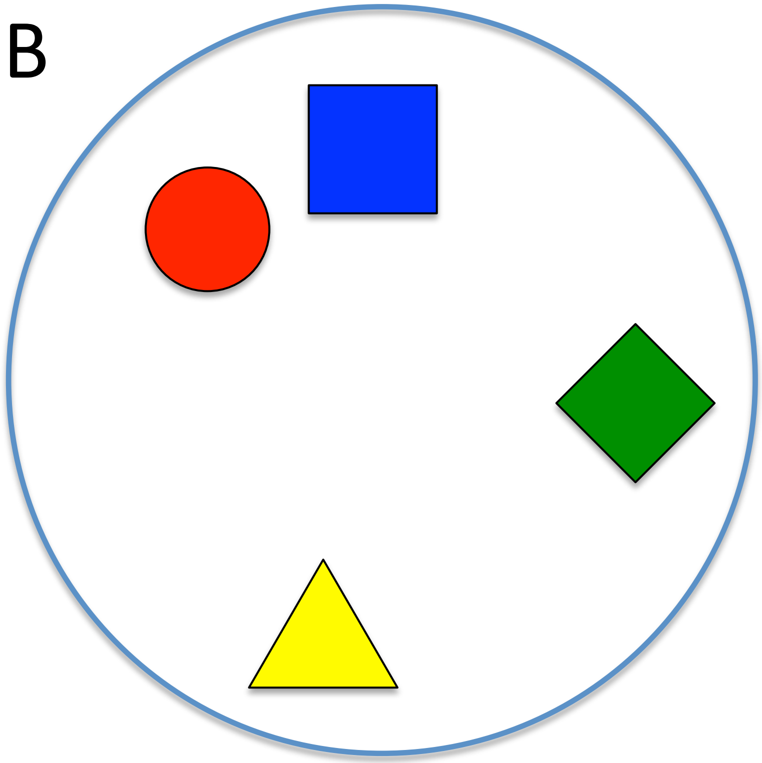
Alpha diversity = 3

Alpha diversity = 4

Alpha diversity



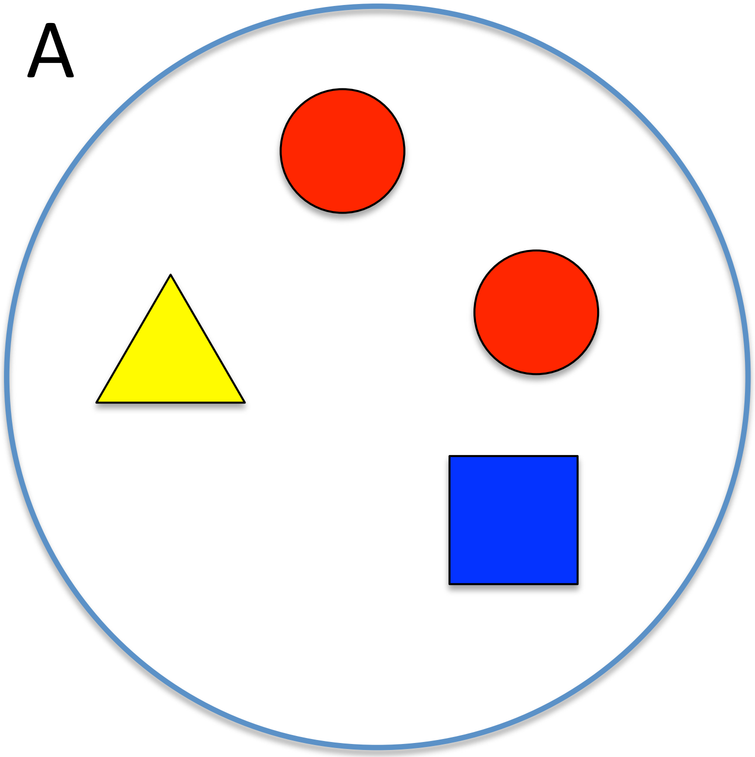
Alpha diversity = 3



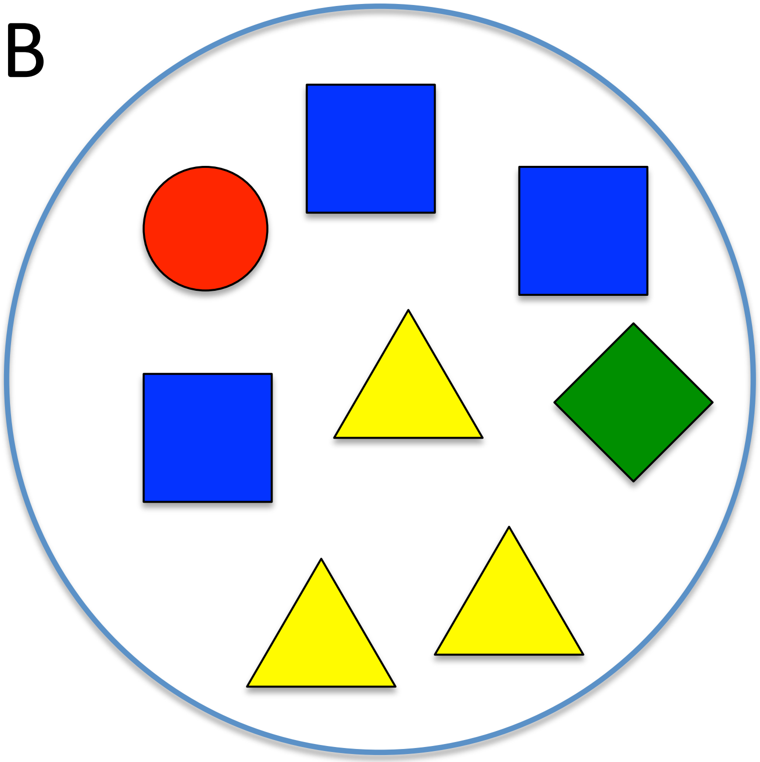
Alpha diversity = 4

Sample B is more diverse than sample A

Alpha diversity



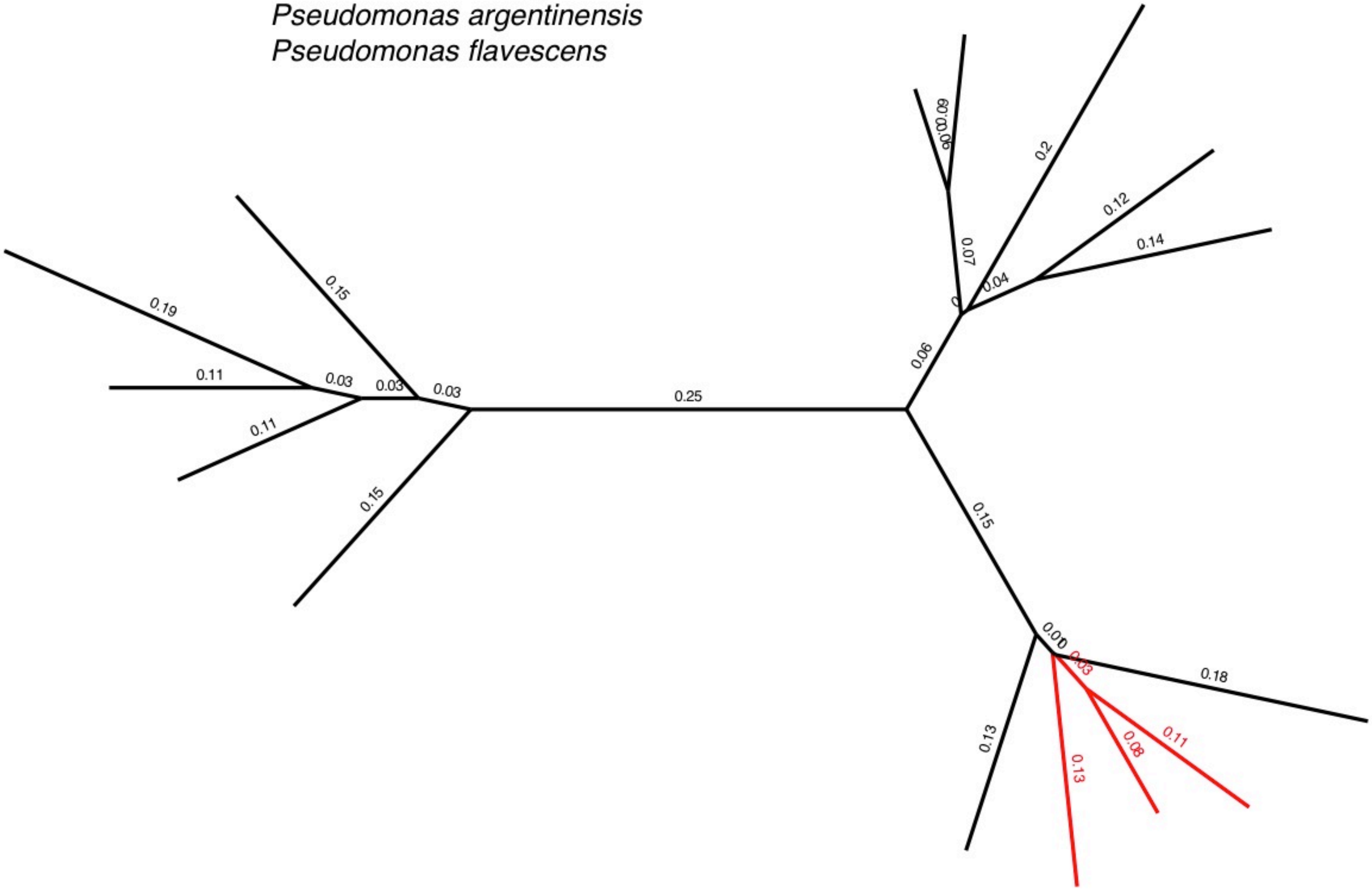
Alpha diversity = 3
Total sequences = 4



Alpha diversity = 4
Total sequences = 8

Sample A

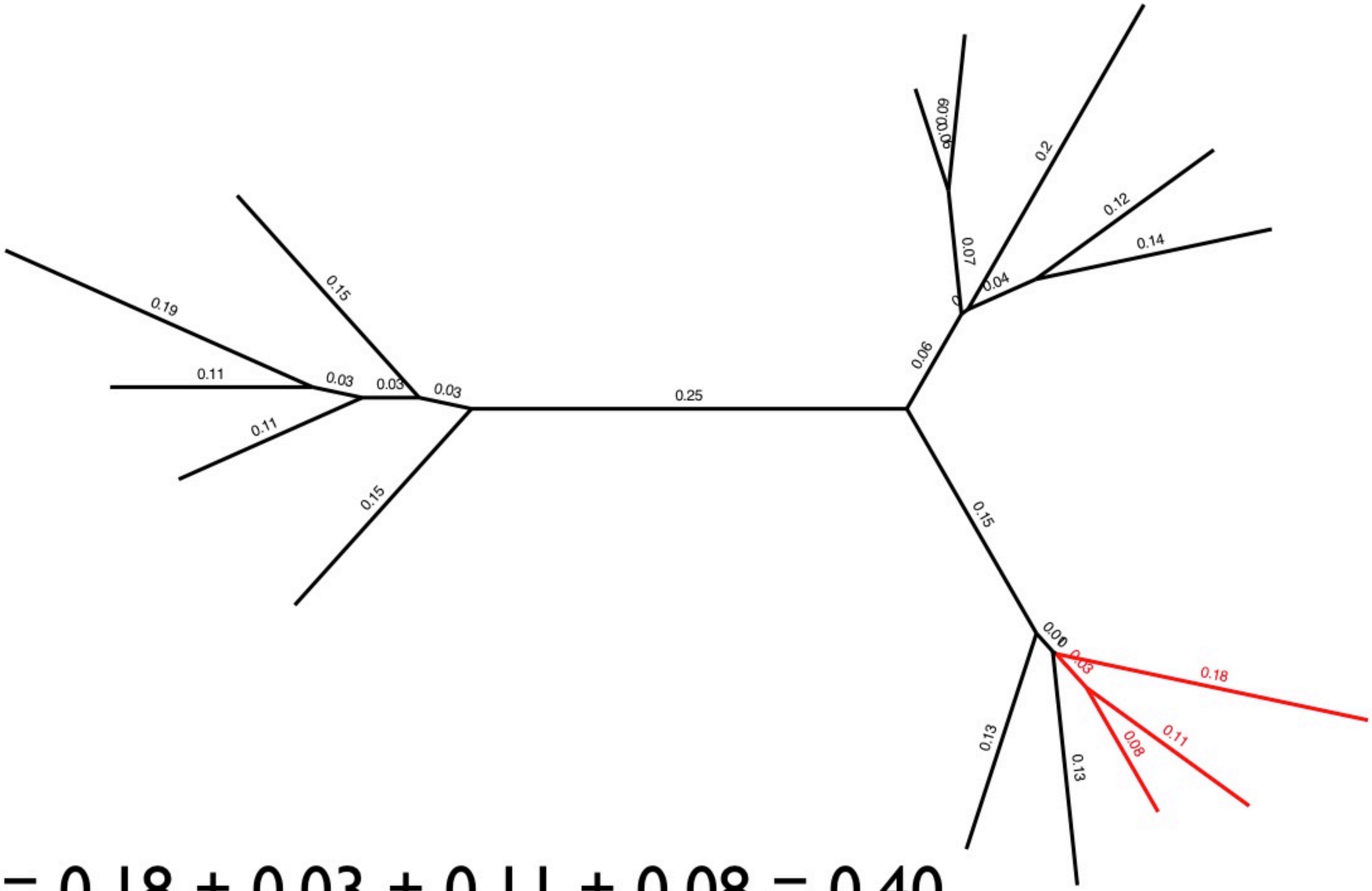
Pseudomonas aeruginosa
Pseudomonas argentinensis
Pseudomonas flavescens



PD = 0.13 + 0.03 + 0.11 + 0.08 = 0.35

Sample B

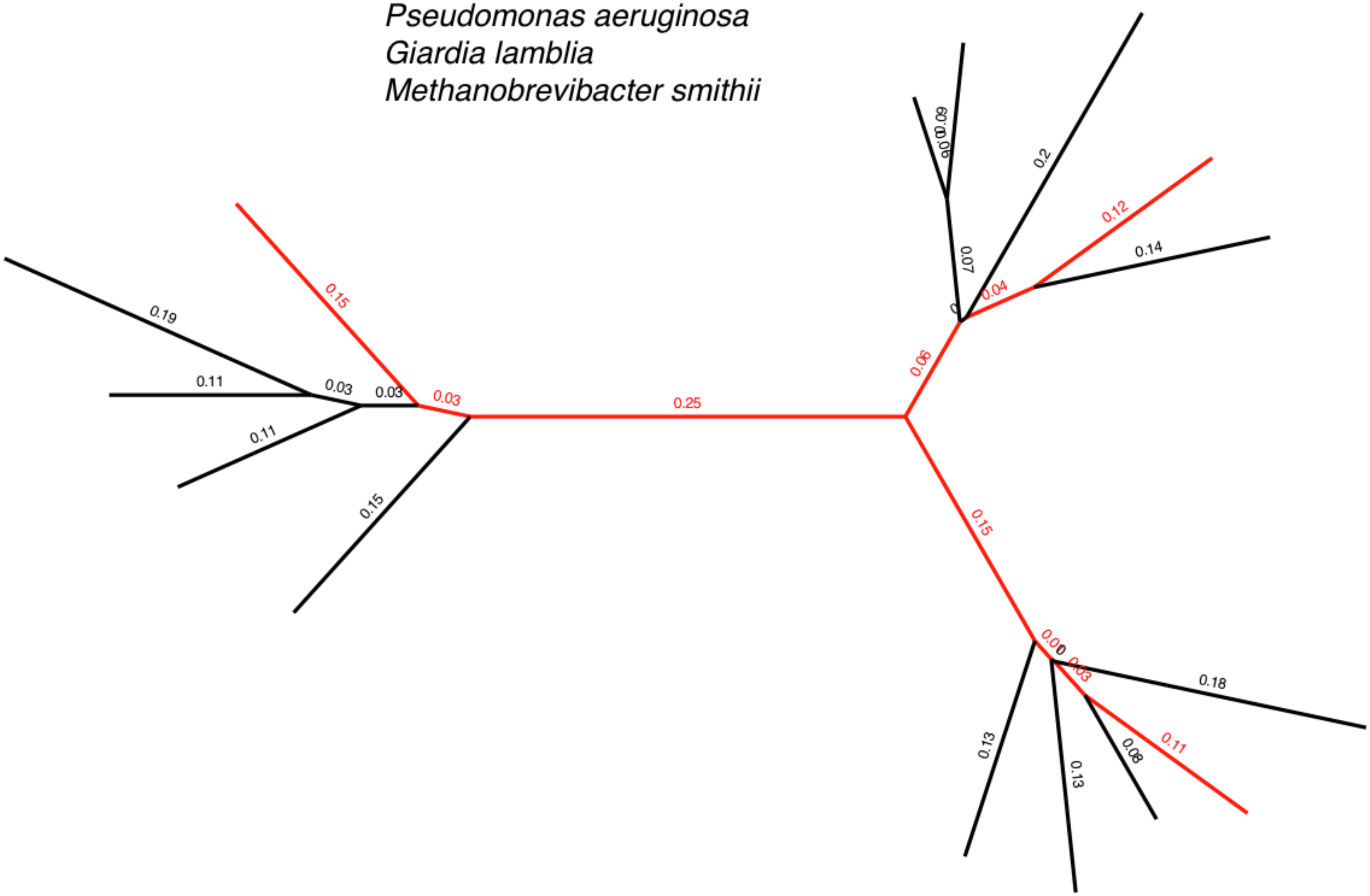
Pseudomonas aeruginosa
Pseudomonas argentinensis
Escherichia coli



PD = 0.18 + 0.03 + 0.11 + 0.08 = 0.40

Sample C

Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii



PD = 0.15+0.03+0.25+0.06+0.04+0.12+0.15+0.01+0.03+0.11= 0.95

Phylogenetic Diversity (PD)

Sample A

Pseudomonas aeruginosa
Pseudomonas argentinensis
Pseudomonas flavescens

PD = 0.35

Sample B

Pseudomonas aeruginosa
Pseudomonas argentinensis
Escherichia coli

PD = 0.40

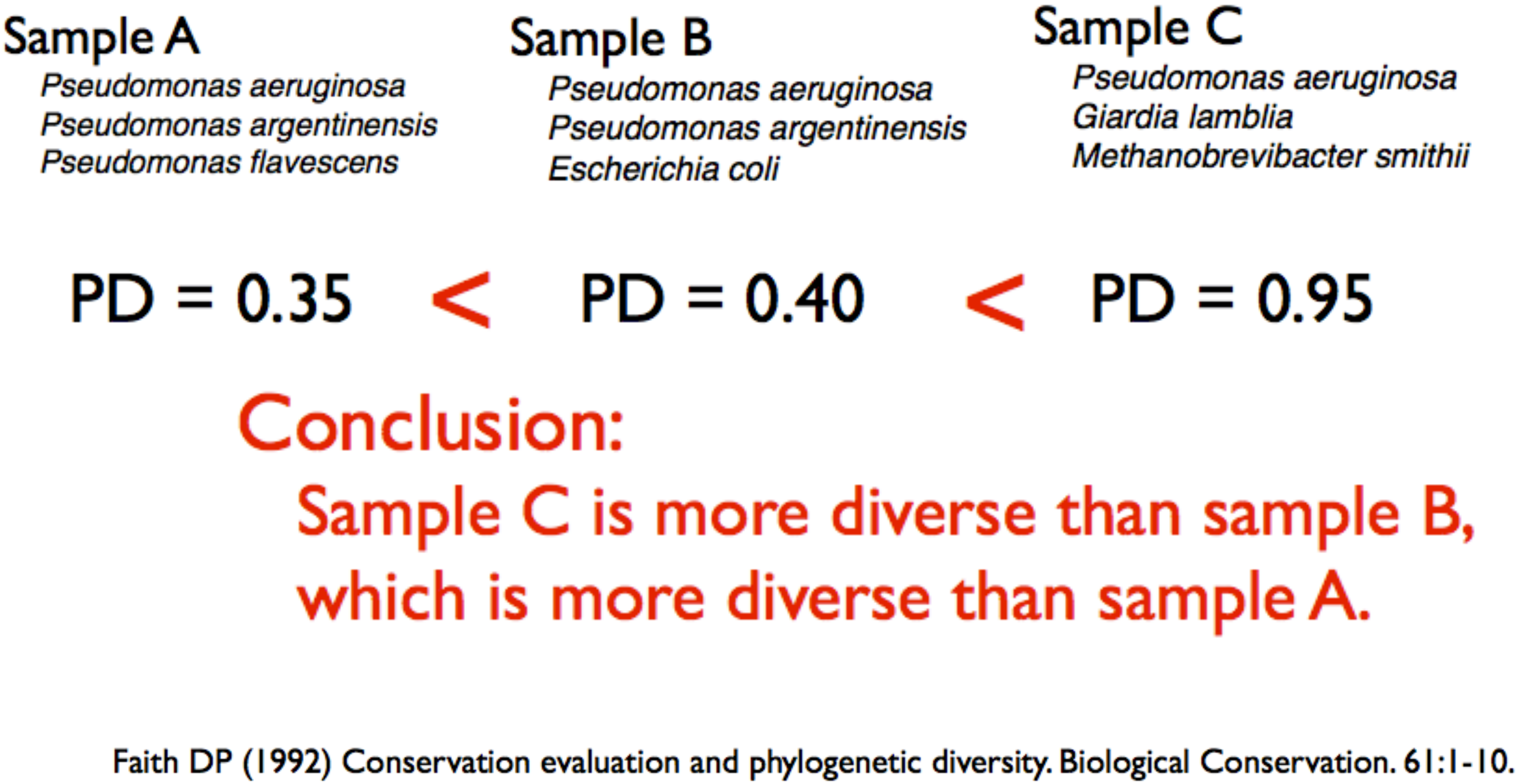
Sample C

Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii

PD = 0.95

Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biological Conservation. 61:1-10.

Phylogenetic Diversity (PD)



Sample A

Pseudomonas aeruginosa
Pseudomonas argentinensis
Pseudomonas flavescens

Sample B

Pseudomonas aeruginosa
Pseudomonas argentinensis
Escherichia coli

Sample C

Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii

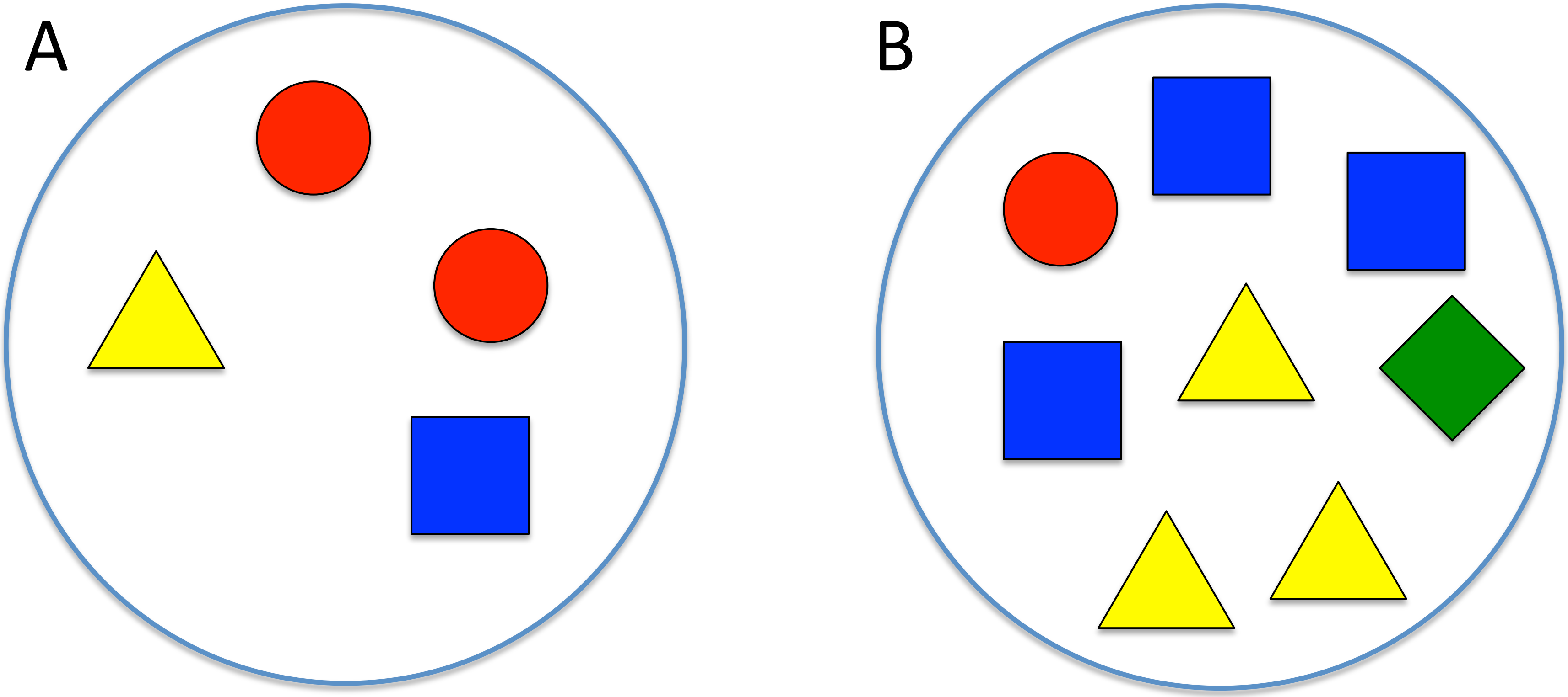
$PD = 0.35 < PD = 0.40 < PD = 0.95$

Conclusion:

Sample C is more diverse than sample B, which is more diverse than sample A.

Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biological Conservation. 61:1-10.

Rarefaction

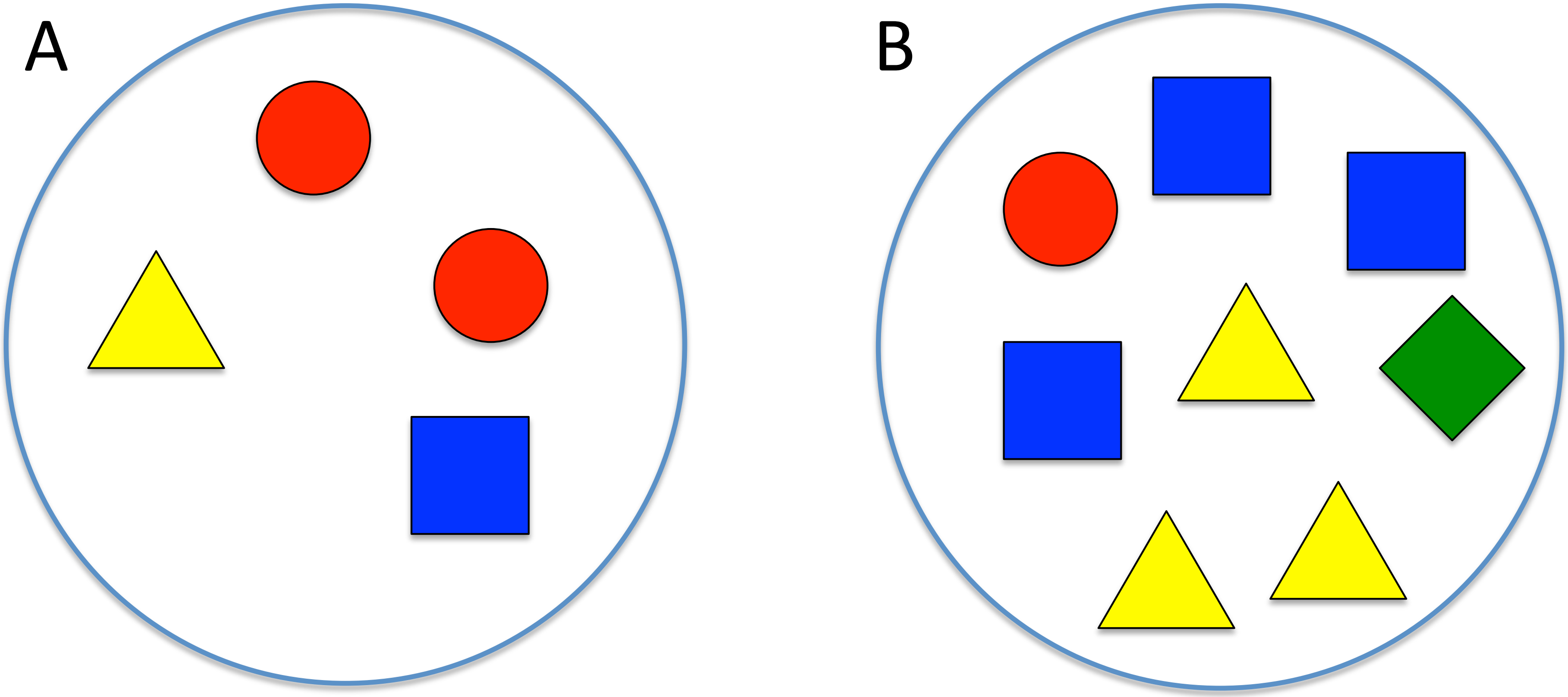


Alpha diversity = 3
Total sequences = 4

Alpha diversity = 4
Total sequences = 8

Randomly select 4 sequences from B

Rarefaction

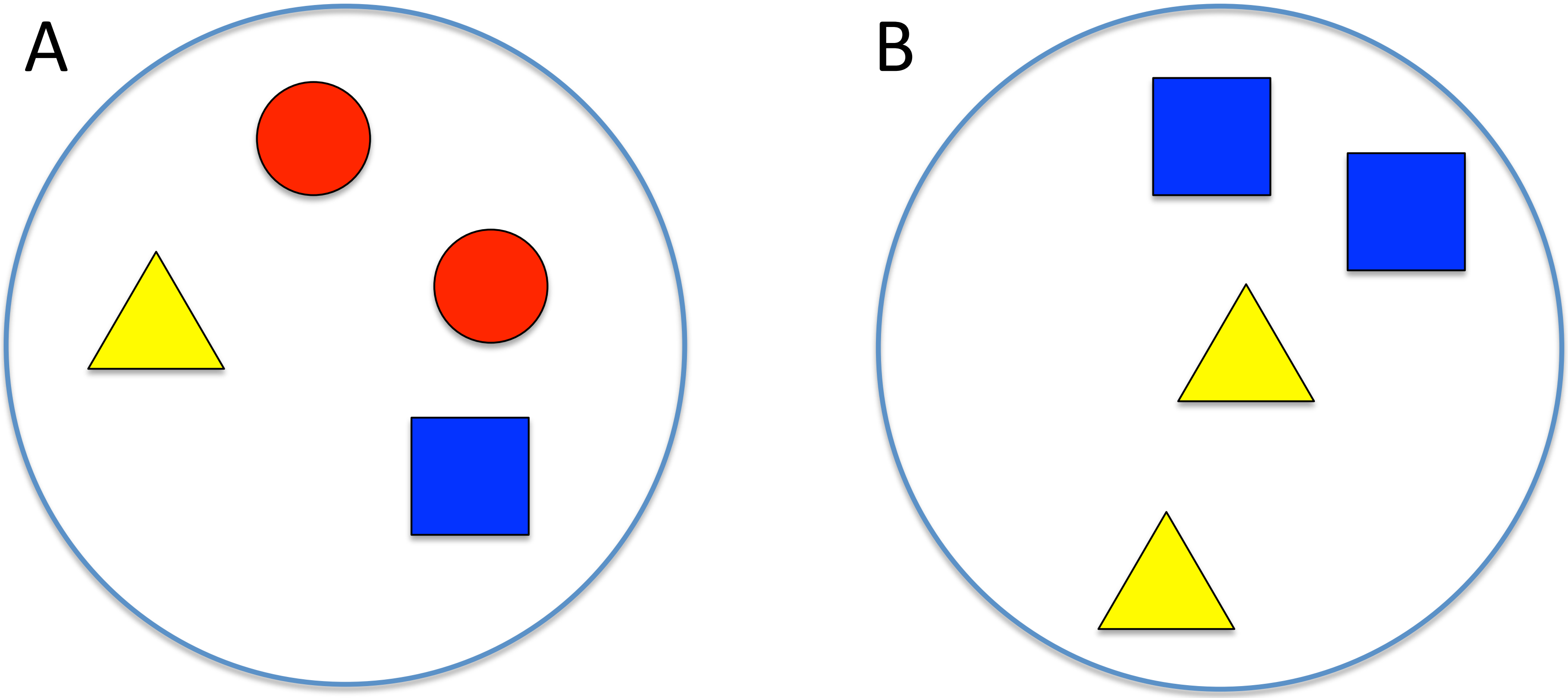


Alpha diversity = 3
Total sequences = 4

Alpha diversity = 4
Total sequences = 8

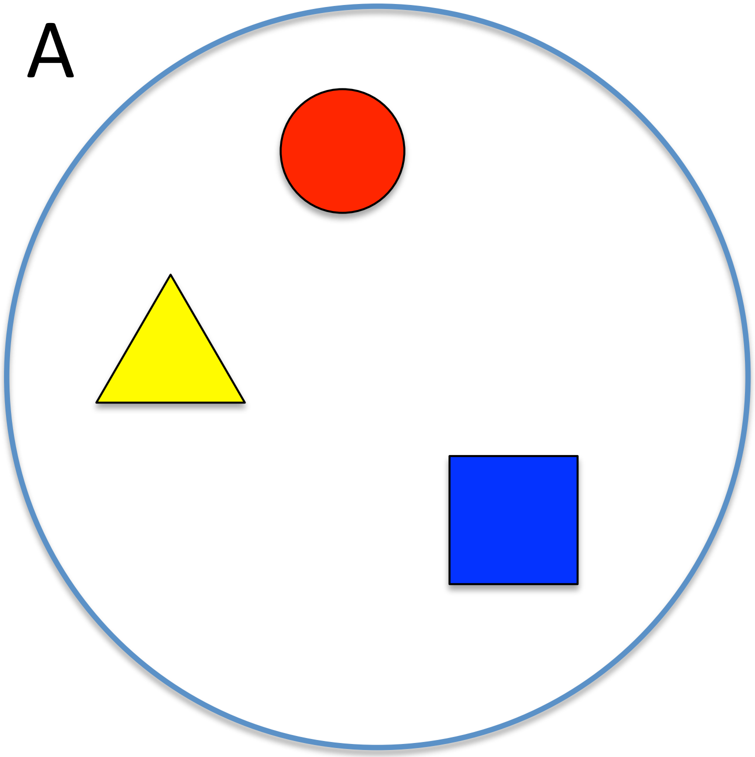
Rarefy to 4 sequences

Rarefaction



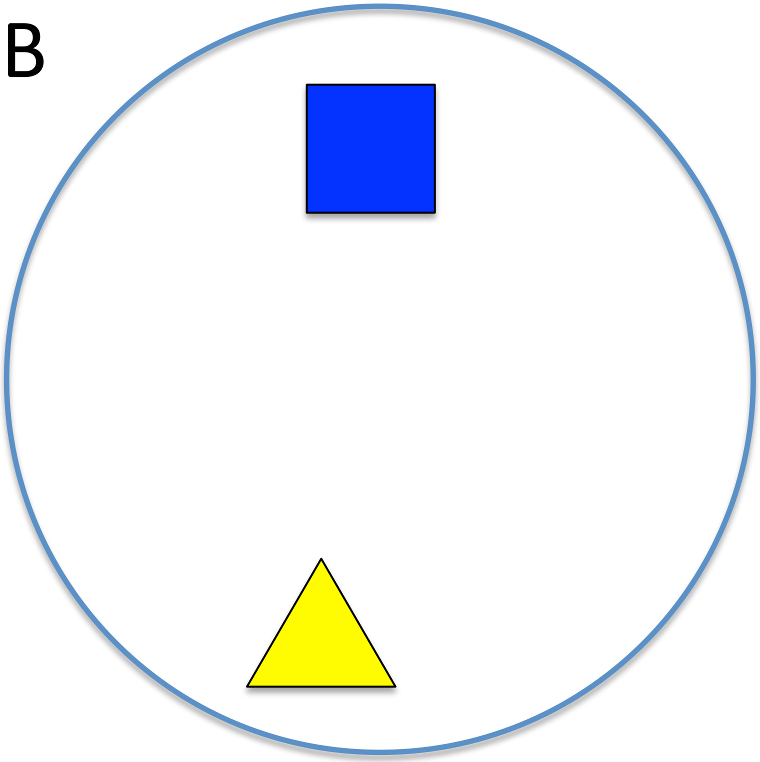
Rarefaction

A



Alpha diversity = 3

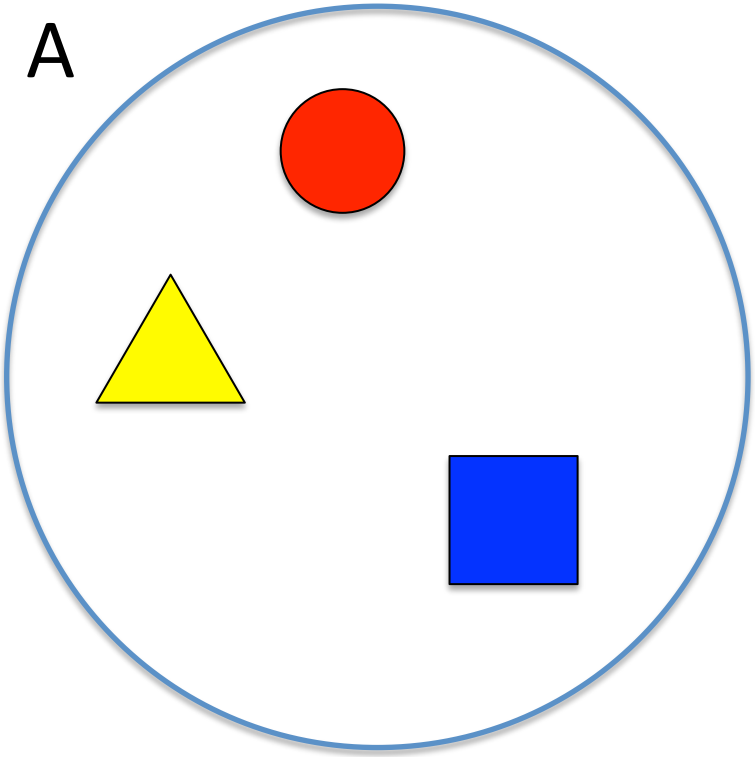
B



Alpha diversity = 2

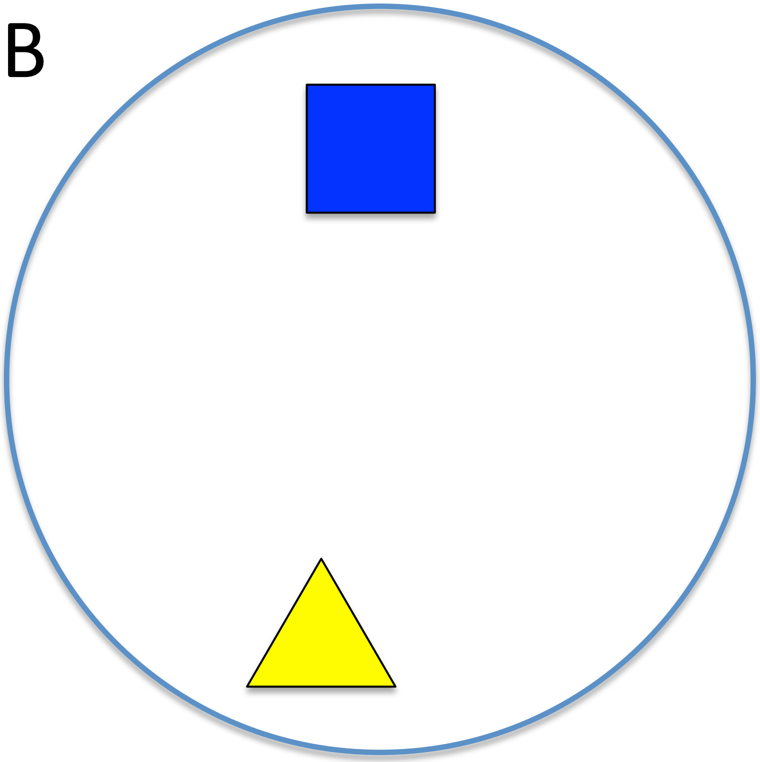
Rarefaction

A



Alpha diversity = 3

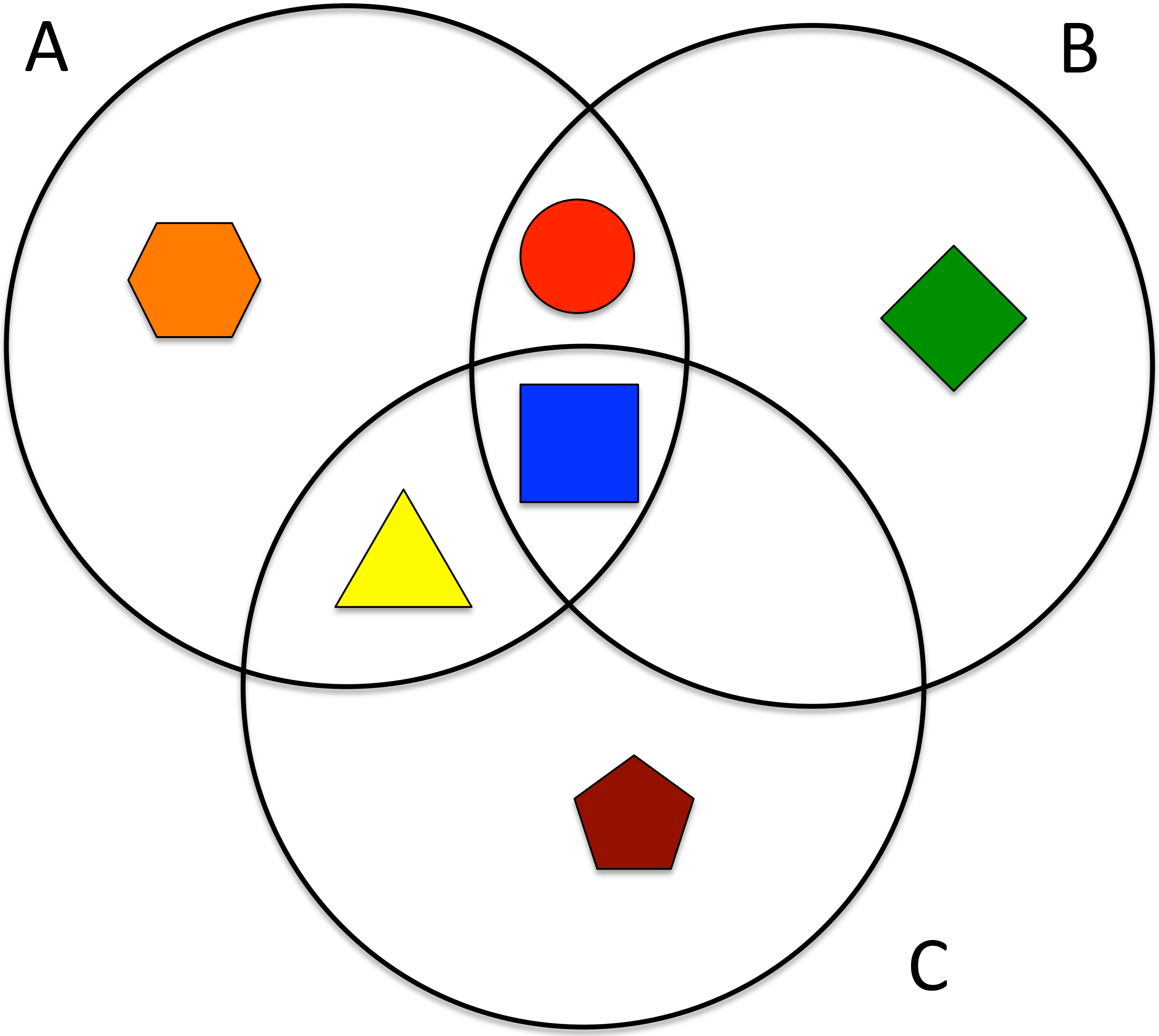
B



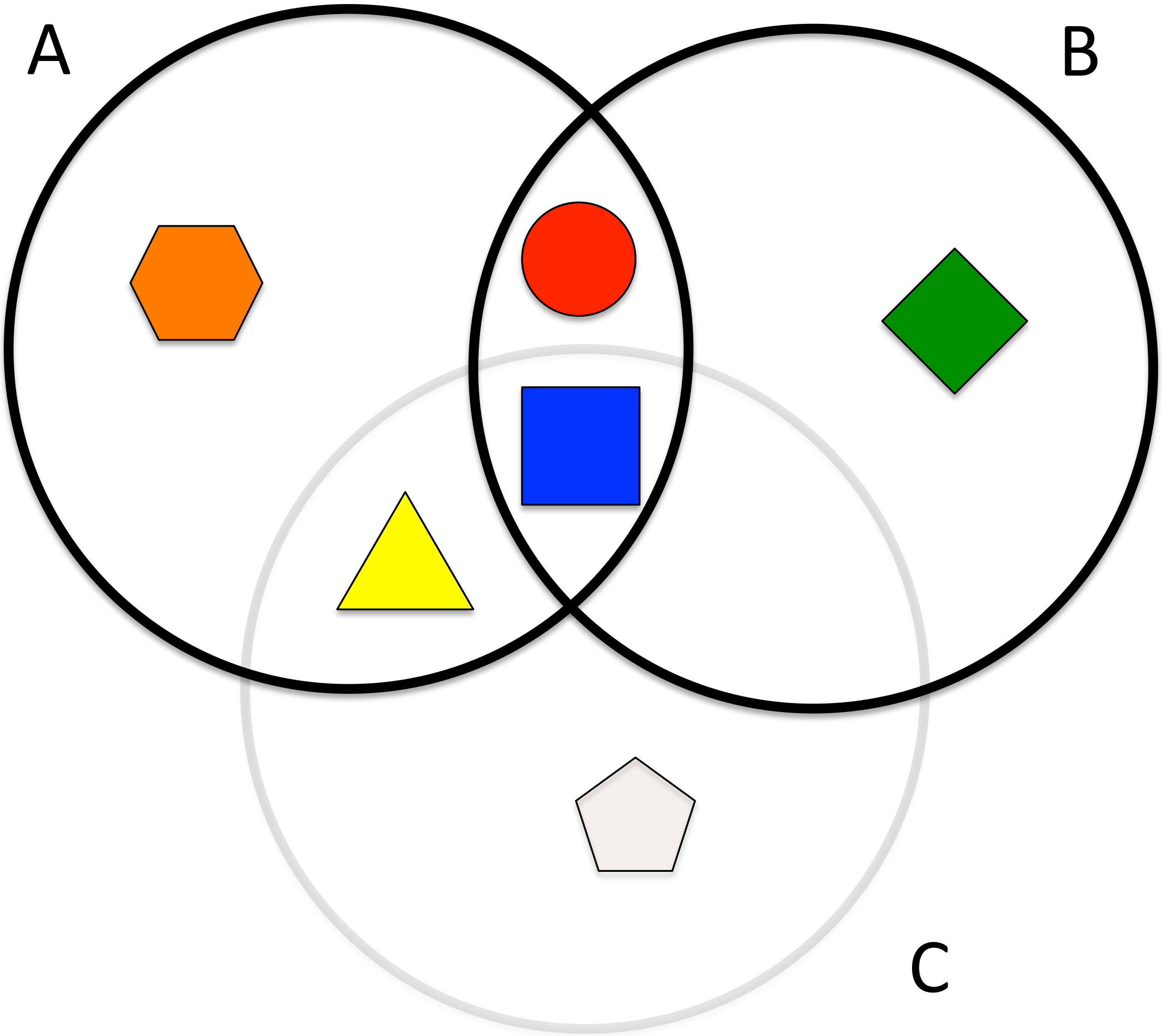
Alpha diversity = 2

Sample A is more diverse than sample B

Beta diversity

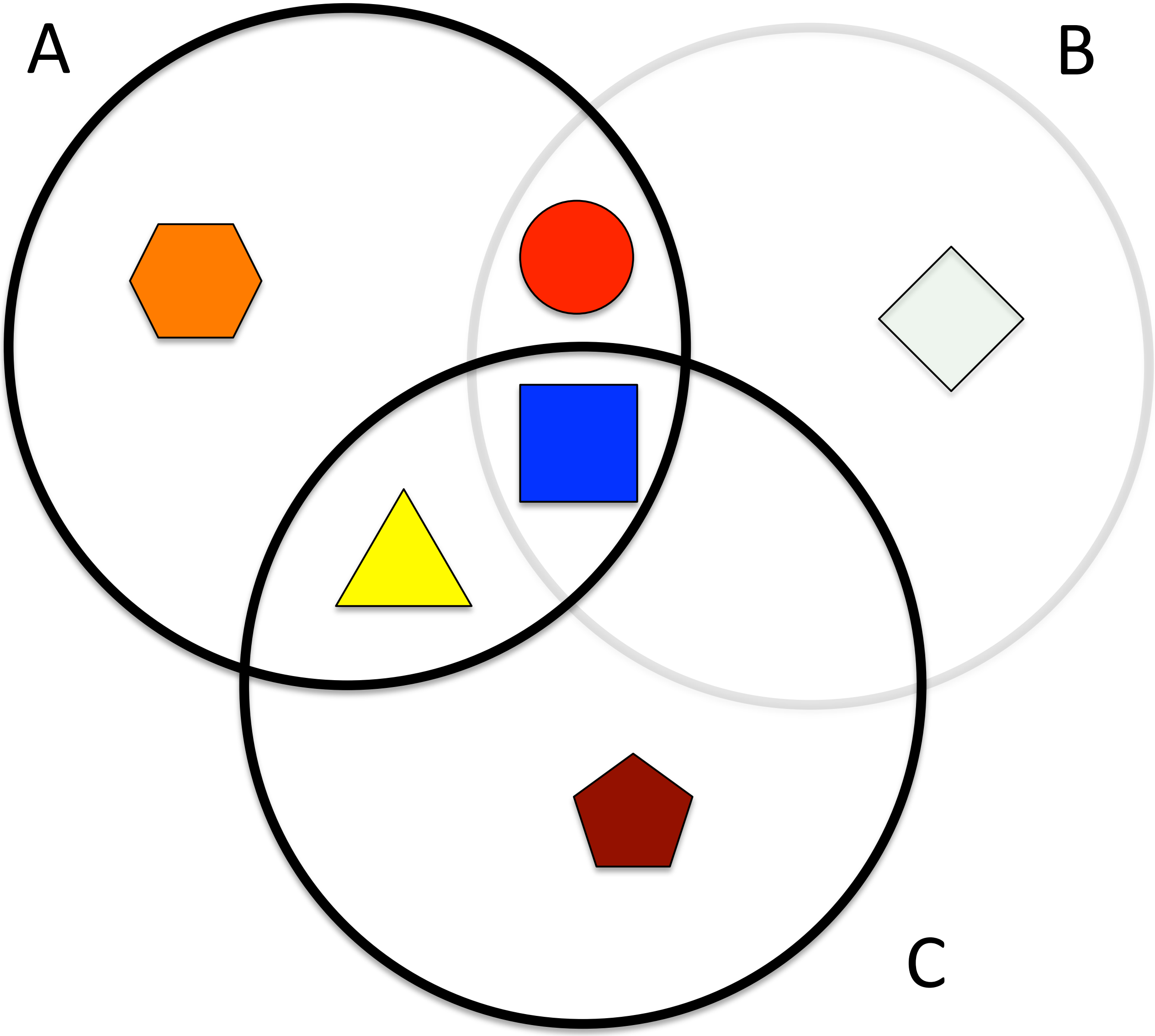


Beta diversity



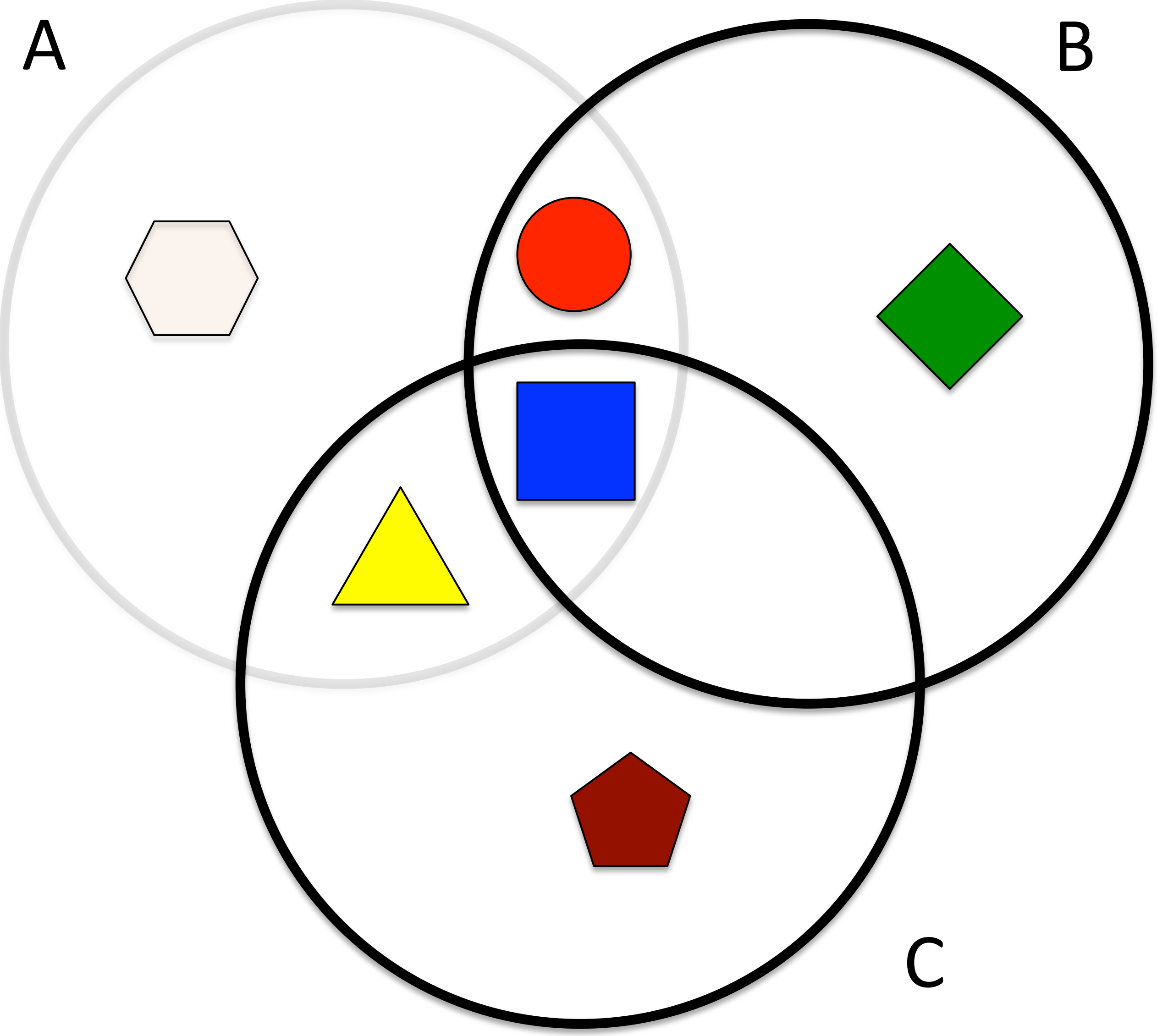
A – B = 2

Beta diversity



A x B = 2
A x C = 2

Beta diversity

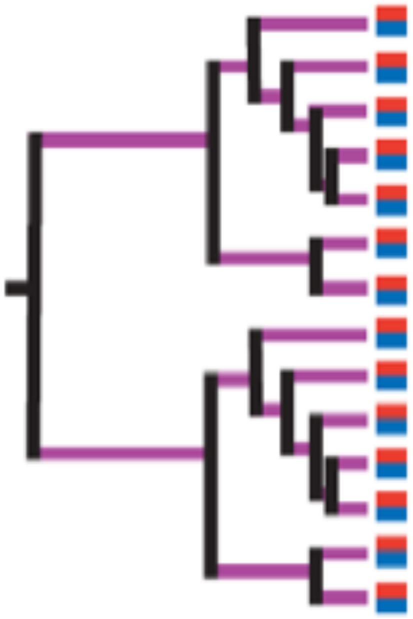


A x B = 2

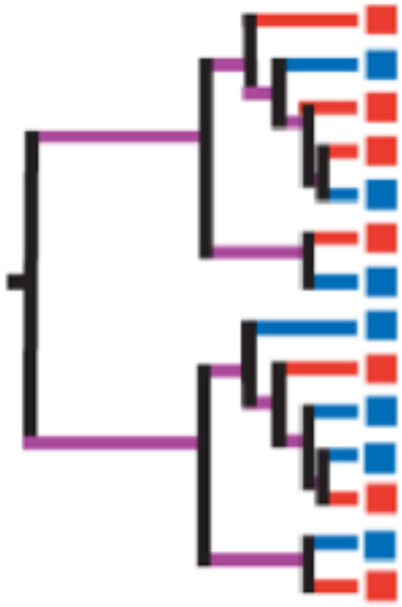
A x C = 2

B x C = 1

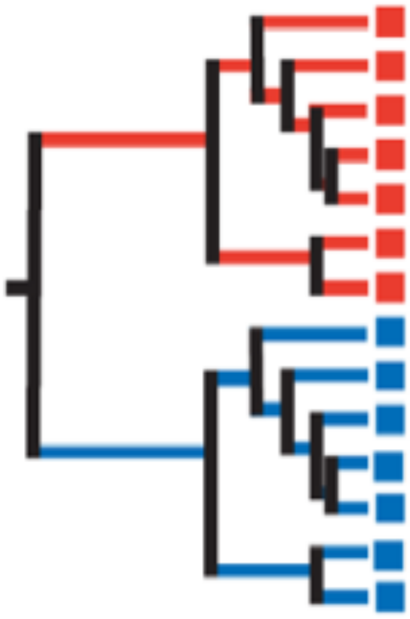
Unweighted UniFrac:
a qualitative, phylogenetic β-diversity metric



Identical communities
D = 0.0



Related communities
D ~ 0.5

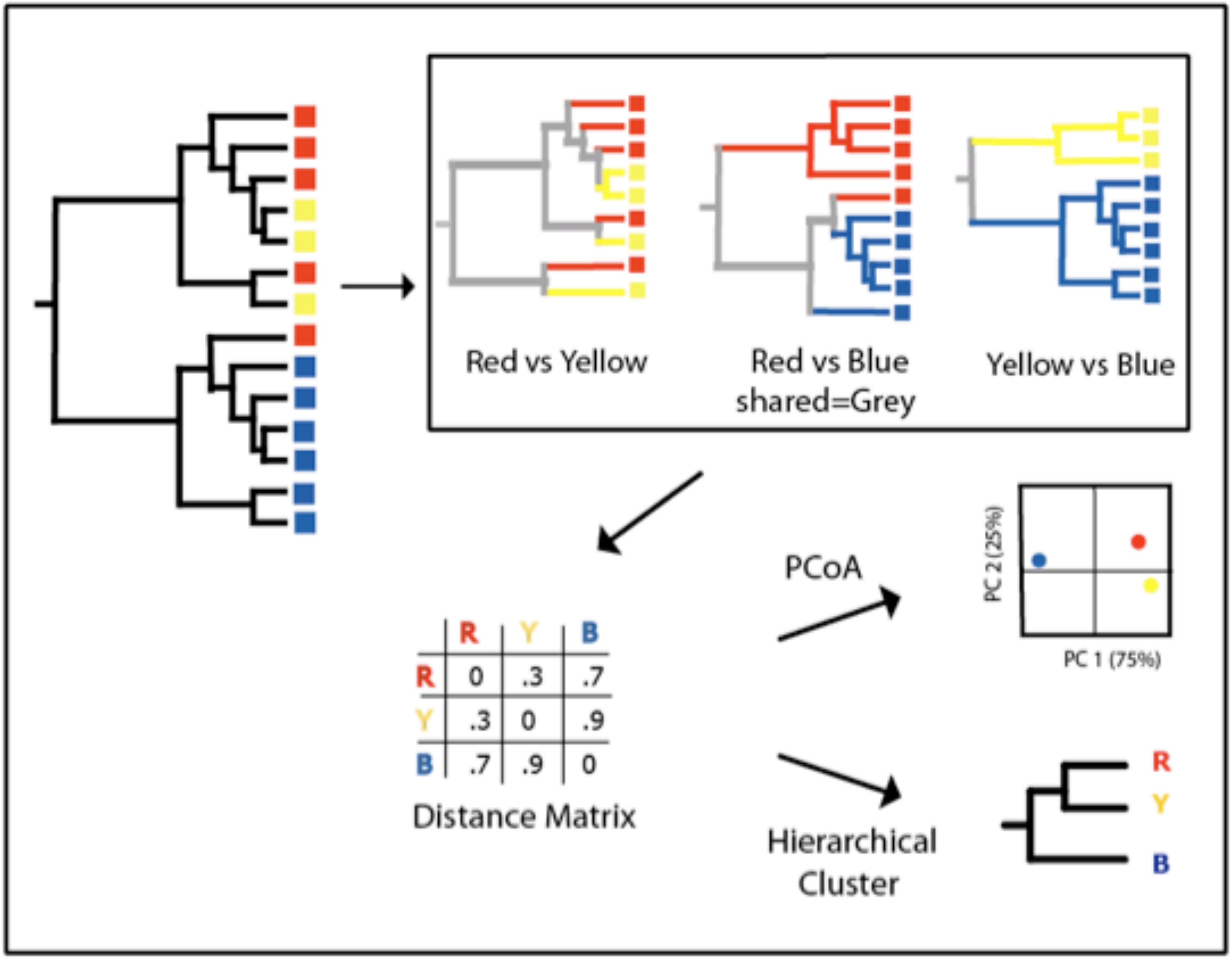


Unrelated communities
D = 1.0

Percent of observed branch length that is unique to either sample

Lozupone and Knight, 2005, Appl Environ Microbiol 71:8228

Clustering by UniFrac distance



	R	Y	B
R	0	.3	.7
Y	.3	0	.9
B	.7	.9	0

In the ancient times of ... 2012

- We used KiNG to view 3D plots in QIIME.

It's 2014!

EMPEROR

Vázquez-Baeza *et al. GigaScience* 2013, **2**:16
http://www.gigasciencejournal.com/content/2/1/16

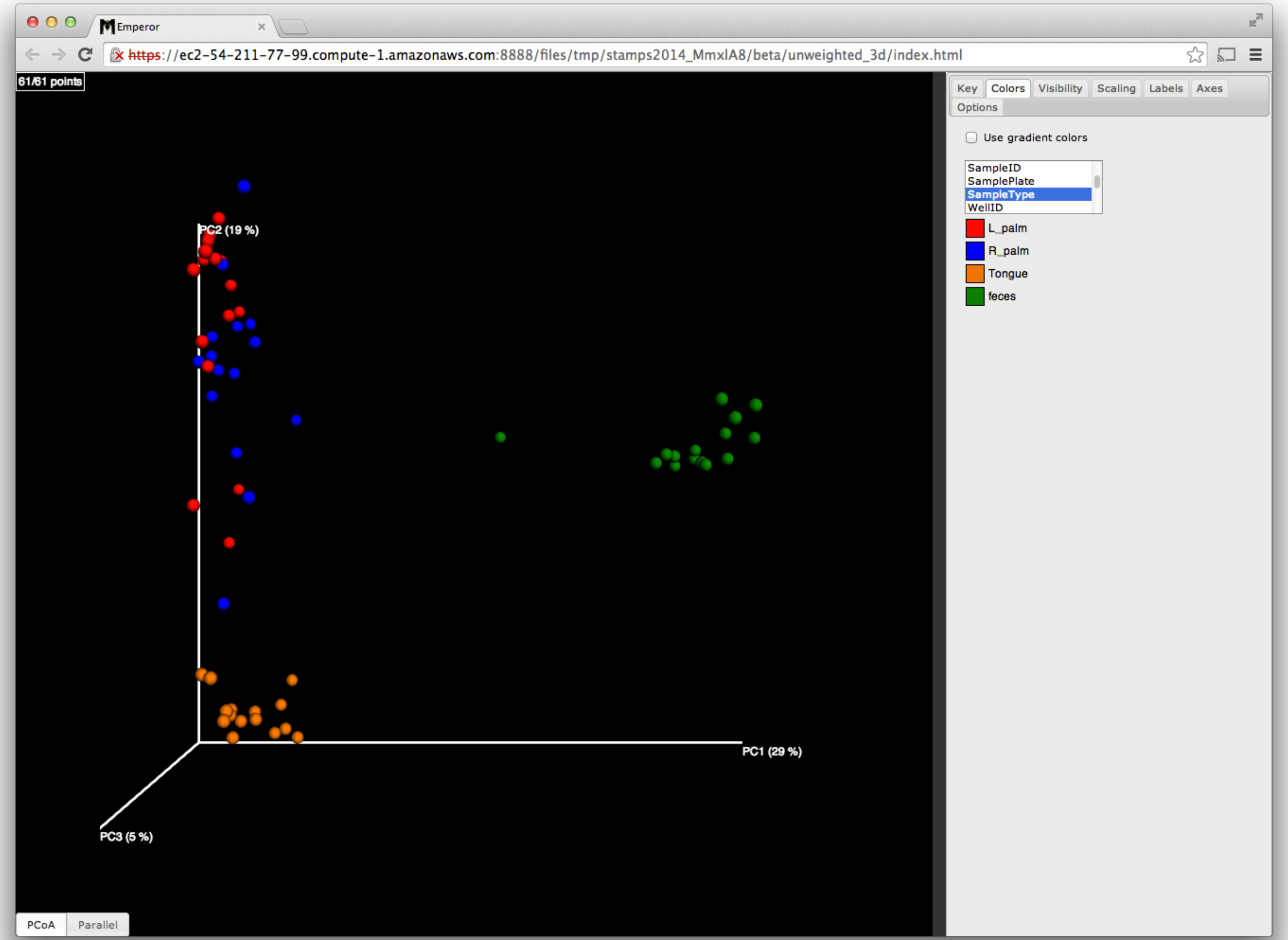
(GIGA)^n SCIENCE

TECHNICAL NOTE **Open Access**

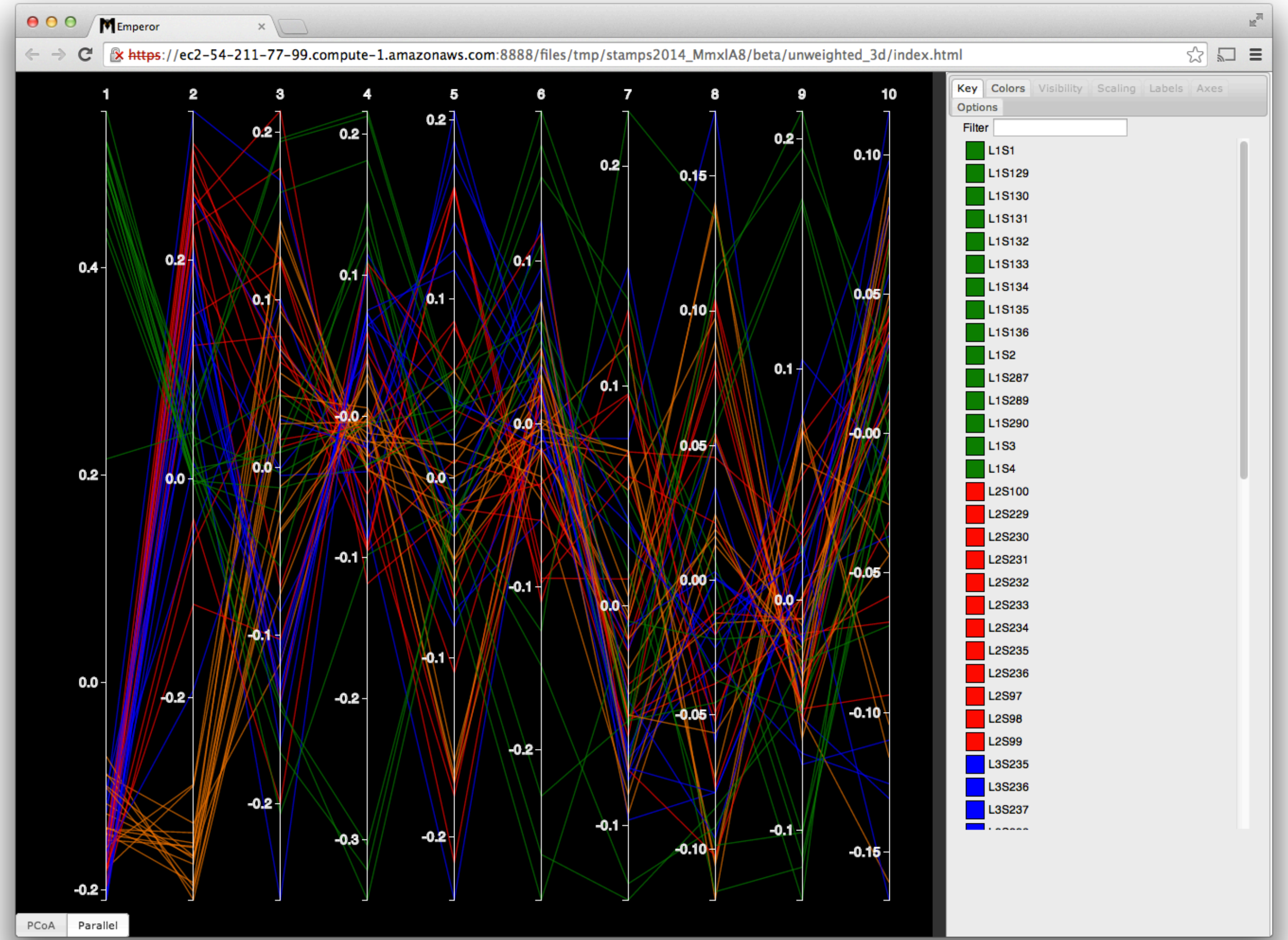
EMPeror: a tool for visualizing high-throughput microbial community data

Yoshiki Vázquez-Baeza[1], Meg Pirrung[2], Antonio Gonzalez[3] and Rob Knight[3,4,5*]

Visualizing ordination plots (3D)



Visualizing ordination plots (10D)



Parallel Plots

Overview

- Integrates with QIIME and it's workflows
- Use case-driven
- Deals well with rich metadata
- Lightweight

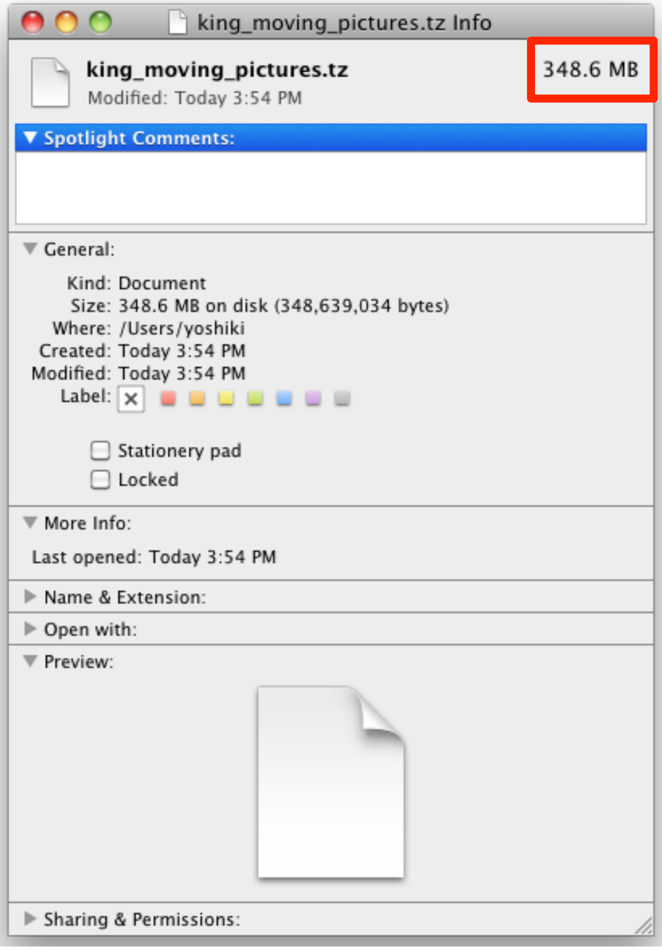
✗ Java

✓ WebGL

http://www.khronos.org/webgl/
http://www.oracle.com/

Output File Size Comparison

KiNG



king_moving_pictures.tz Info

king_moving_pictures.tz 348.6 MB

Modified: Today 3:54 PM

▼ Spotlight Comments:

▼ General:

Kind: Document
Size: 348.6 MB on disk (348,639,034 bytes)
Where: /Users/yoshiki
Created: Today 3:54 PM
Modified: Today 3:54 PM
Label: ☒

☐ Stationery pad
☐ Locked

▼ More Info:

Last opened: Today 3:54 PM

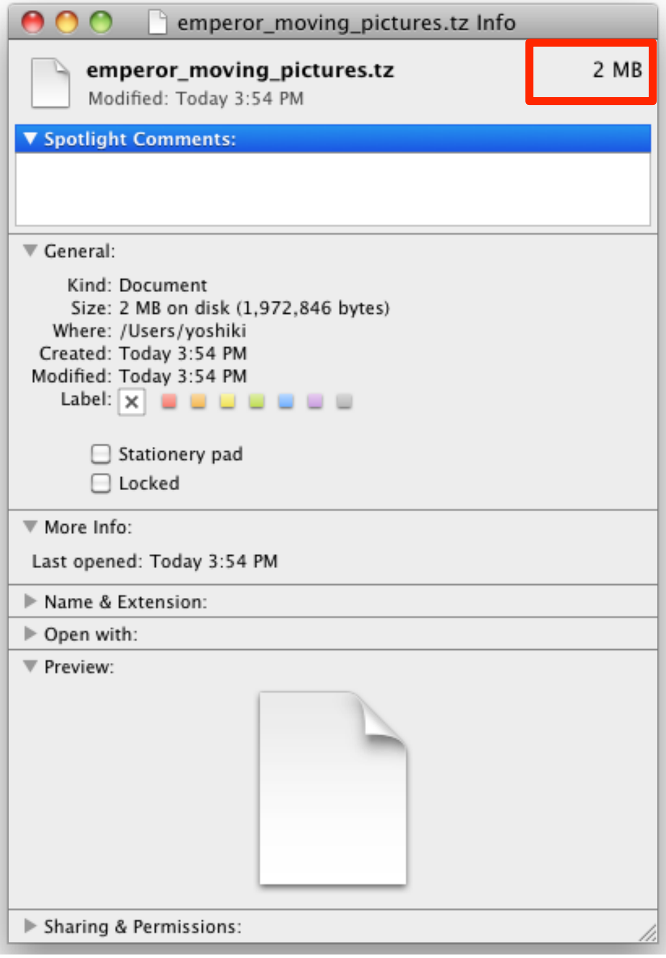
▶ Name & Extension:

▶ Open with:

▼ Preview:

▶ Sharing & Permissions:

EMPEROR



emperor_moving_pictures.tz Info

emperor_moving_pictures.tz 2 MB

Modified: Today 3:54 PM

▼ Spotlight Comments:

▼ General:

Kind: Document
Size: 2 MB on disk (1,972,846 bytes)
Where: /Users/yoshiki
Created: Today 3:54 PM
Modified: Today 3:54 PM
Label: ☒

☐ Stationery pad
☐ Locked

▼ More Info:

Last opened: Today 3:54 PM

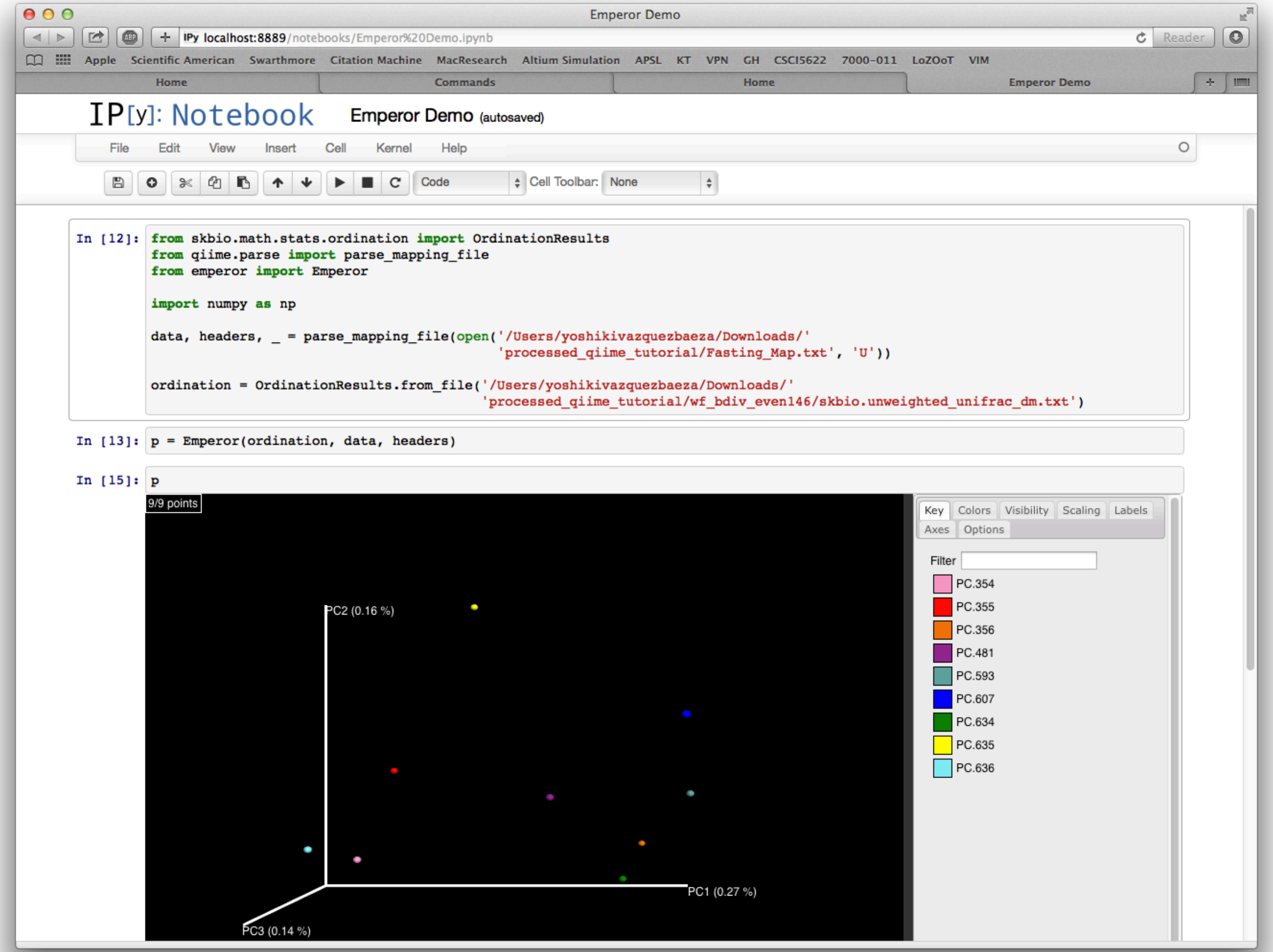
▶ Name & Extension:

▶ Open with:

▼ Preview:

▶ Sharing & Permissions:

Reproducible Science



```
In [12]: from skbio.math.stats.ordination import OrdinationResults
         from qiime.parse import parse_mapping_file
         from emperor import Emperor

         import numpy as np

         data, headers, _ = parse_mapping_file(open('/Users/yoshikivazquezbaeza/Downloads/'
                                              'processed_qiime_tutorial/Fasting_Map.txt', 'U'))

         ordination = OrdinationResults.from_file('/Users/yoshikivazquezbaeza/Downloads/'
                                             'processed_qiime_tutorial/wf_bdiv_even146/skbio.unweighted_unifrac_dm.txt')

In [13]: p = Emperor(ordination, data, headers)

In [15]: p
```

Online Resources

http://emperor.colorado.edu

- Request a feature:
    - www.github.com/biocore/emperor
- Or contact the QIIME Forum
    - http://groups.google.com/group/qiime-forum

Differential feature abundance

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

Disease

Control

	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

Differential feature abundance

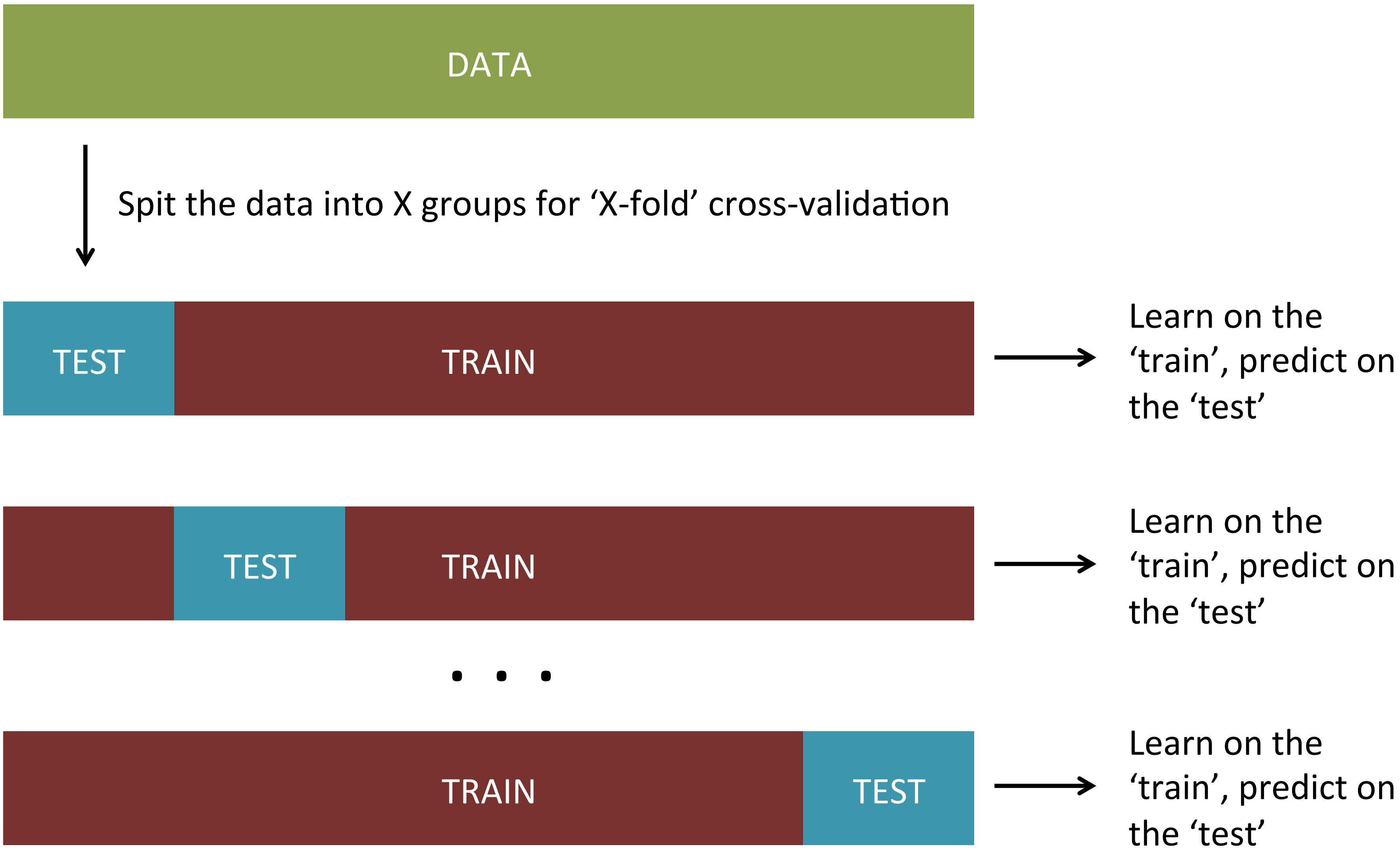
- Huge body of literature and statistical techniques available. Choose wisely.
- Not a substitute for:
  - A trained statistician
  - Proper experimental design
  - Mechanistic understanding

Differential feature abundance

- New techniques are emerging in microbial ecology (they've been used in other fields for a while elsewhere)
  - "Waste Not Want Not" McMurdie and Holmes 2014
- QIIME 1.9 (next release) will contain some new methods.

Supervised learning

- Subset of machine learning.
- We ‘supervise’ the algorithms learning. It learns from **training** data and applies what it learned to **test** data.
- **Training** data and **test** data both come from your dataset - you aren’t taking new samples or using someone else’s data.



DATA

Spit the data into X groups for ‘X-fold’ cross-validation

TEST TRAIN → Learn on the ‘train’, predict on the ‘test’

TEST TRAIN → Learn on the ‘train’, predict on the ‘test’

. . .

TRAIN TEST → Learn on the ‘train’, predict on the ‘test’

Supervised learning

- **Cross-validation** prevents **overfitting**.
- Overfitting is 'over-learning' your data – learning its idiosyncrasies rather than it's underlying pattern.

"Supervised Classification of Human Microbiota" Knights et al. 2011

Andrew Ng's Coursera Course on Machine Learning

"Elements of Statistical Learning" Hastie, Tibshirani, Friedman

scikit-learn: http://scikit-learn.org/stable/

That is all … for now

http://github.com/biocore/qiime

http://octodex.github.com/front-end-conftocat/

Modified from
http://bio.sacnas.org/uploads/Judges/we_need_you.jpg

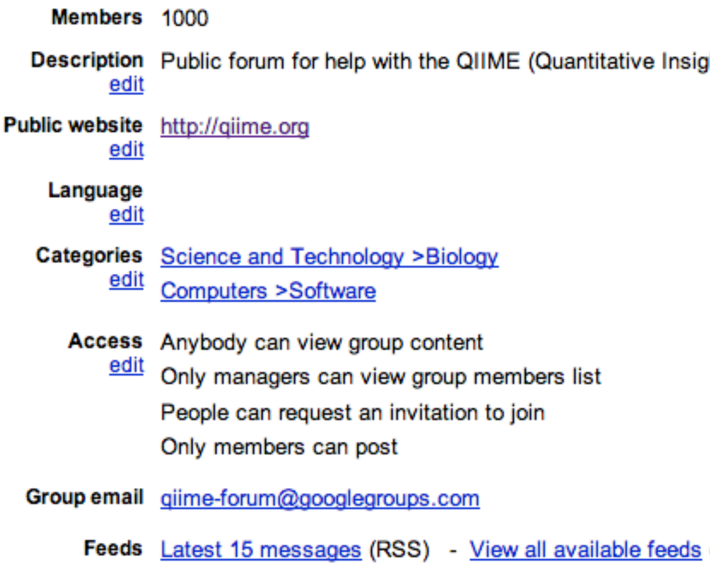
Qiime Forum

Search the forum

http://forum.qiime.org

We try to answer within 24 hours

QIIME Forum



Members 1000

Description edit
Public forum for help with the QIIME (Quantitative Insights Into Microbial Ecology) Toolkit. See http://qiime.org for more details.

Public website edit
http://qiime.org

Language edit

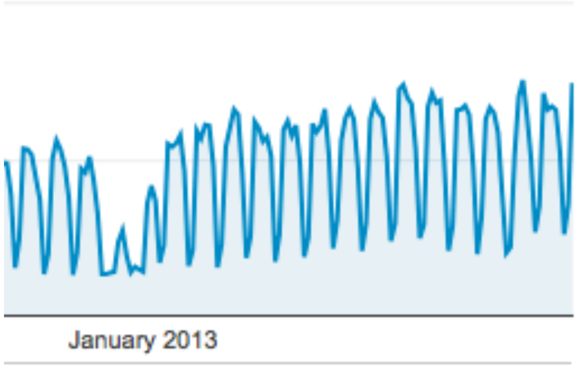
Categories edit
Science and Technology >Biology
Computers >Software

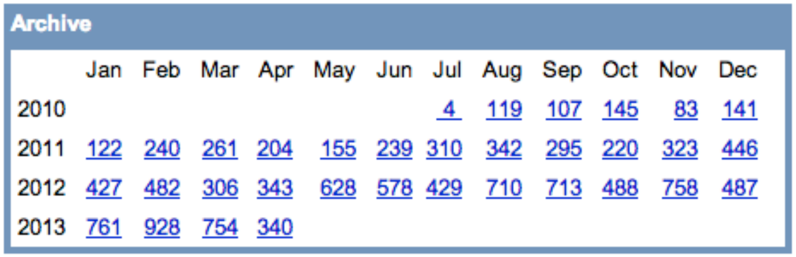
Access edit
Anybody can view group content
Only managers can view group members list
People can request an invitation to join
Only members can post

Group email qiime-forum@googlegroups.com

Feeds Latest 15 messages (RSS) - View all available feeds (RSS and Atom)

QIIME allows analysis of high-throughput community sequencing data
JG Caporaso, J Kuczynski, J Stombaugh, K Bittinger… - Nature …, 2010 - nature.com
To the Editor: High-throughput sequencing is revolutionizing microbial ecology studies. Efforts like the Human Microbiome Projects 1 and the US National Ecological Observatory Network 2 are helping us to understand the role of microbial diversity in habitats within our …





Archive

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2010							4	119	107	145	83	141
2011	122	240	261	204	155	239	310	342	295	220	323	446
2012	427	482	306	343	628	578	429	710	713	488	758	487
2013	761	928	754	340								

Acknowledgments

- Rob Knight
- Antonio Gonzalez
- Adam Robbins-Pianka
- Will Van Treuren
- Luke Ursell
- Jose Clemente
- Daniel McDonald
- Greg Caporaso
- Jackson Chen
- Nicholas Bokulich
- The Knight Laboratory
- Team of QIIME Developers
- These slides have been adapted from multiple sets of slides.

qiime

IP[y]:

KNIGHTLAB

StarCluster

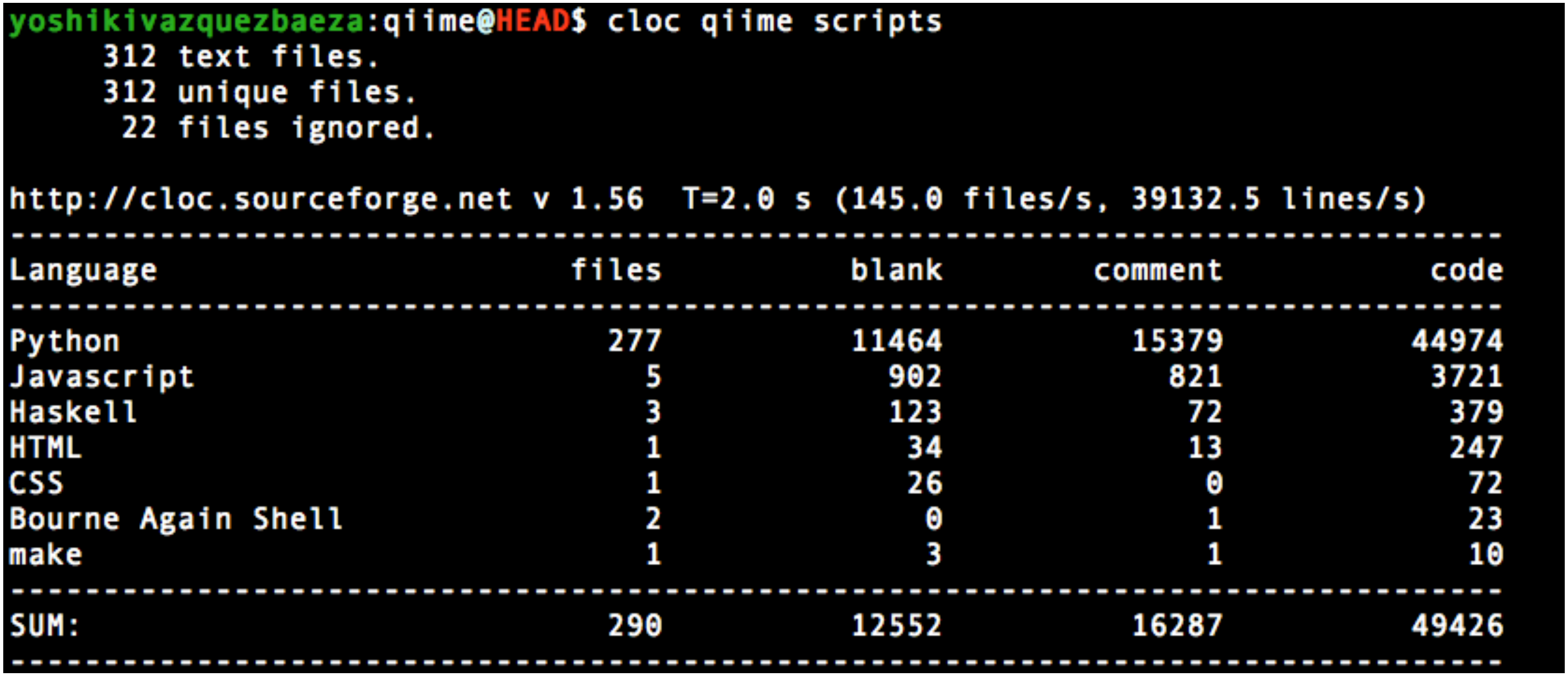
IQ BIOLOGY
INTERDISCIPLINARY QUANTITATIVE BIOLOGY

University of Colorado
Boulder

amazon web services™ EC2

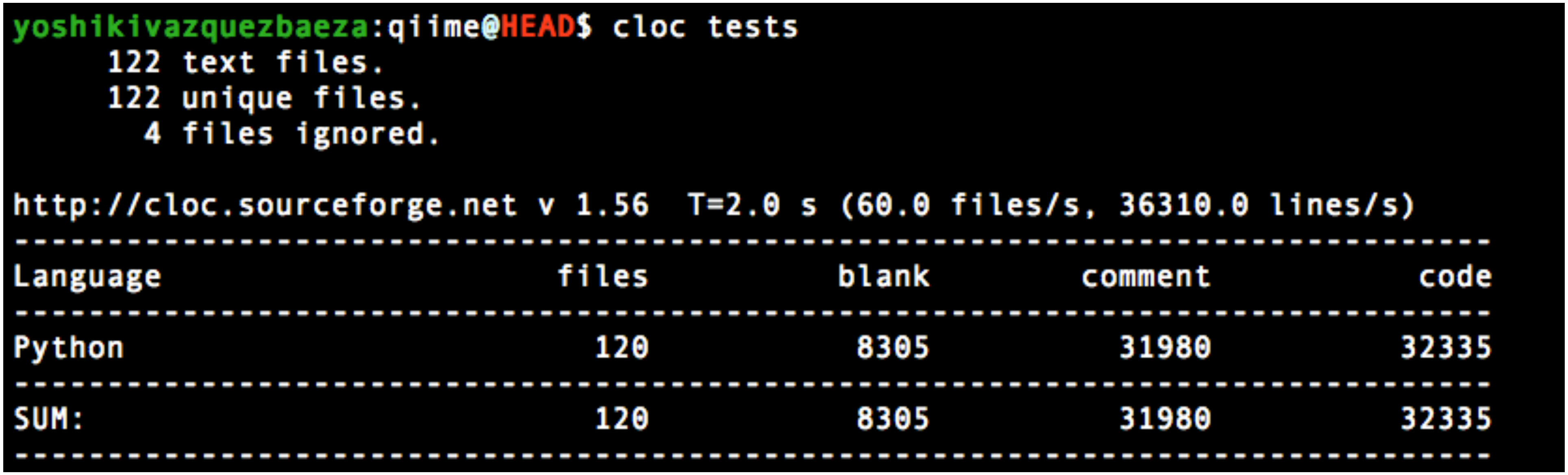
Biological
Observation
Matrix

Testing



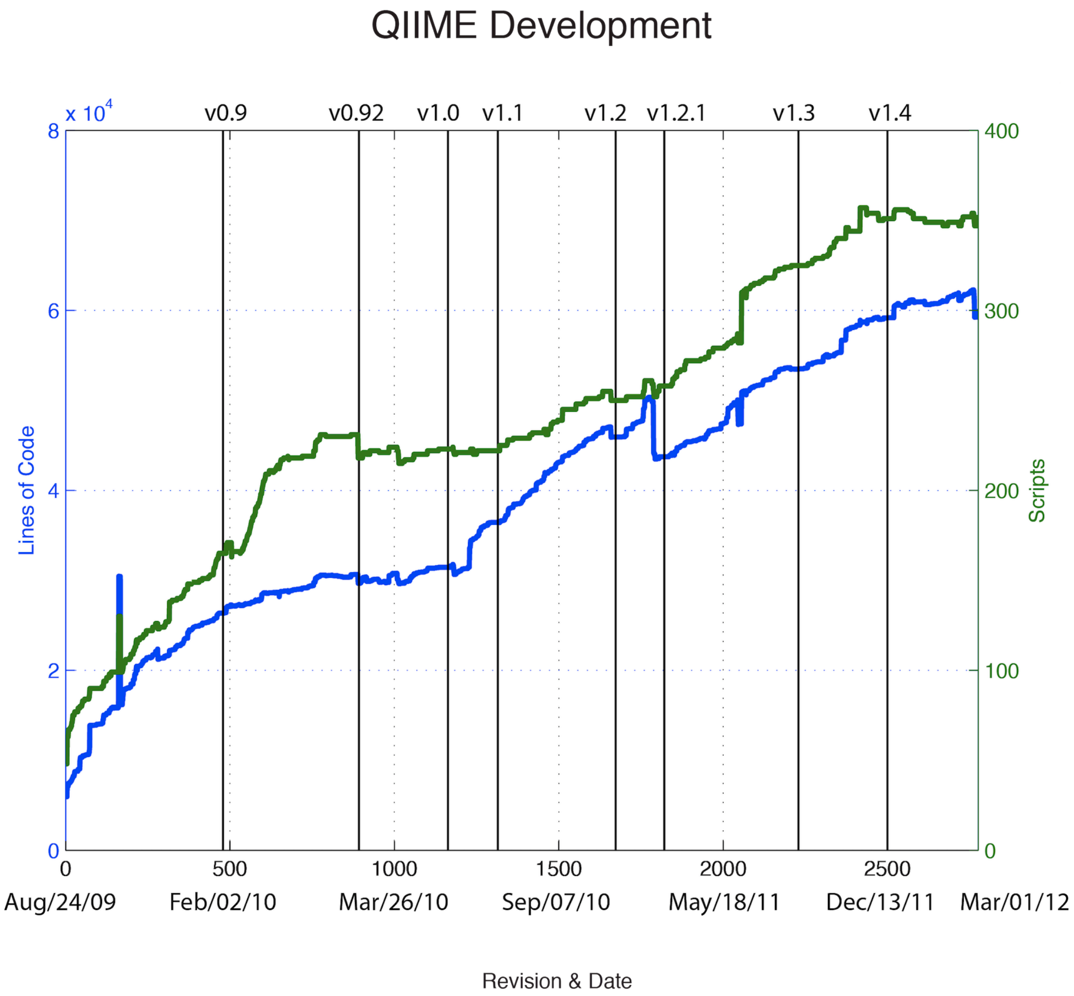
```
yoshikivazquezbaeza:qiime@HEAD$ cloc qiime scripts
     312 text files.
     312 unique files.
      22 files ignored.

http://cloc.sourceforge.net v 1.56  T=2.0 s (145.0 files/s, 39132.5 lines/s)
--------------------------------------------------------------------------------
Language                      files          blank        comment           code
--------------------------------------------------------------------------------
Python                          277          11464          15379          44974
Javascript                        5            902            821           3721
Haskell                           3            123             72            379
HTML                              1             34             13            247
CSS                               1             26              0             72
Bourne Again Shell                2              0              1             23
make                              1              3              1             10
--------------------------------------------------------------------------------
SUM:                            290          12552          16287          49426
--------------------------------------------------------------------------------
```

Testing



```
yoshikivazquezbaeza:qiime@HEAD$ cloc tests
     122 text files.
     122 unique files.
       4 files ignored.

http://cloc.sourceforge.net v 1.56  T=2.0 s (60.0 files/s, 36310.0 lines/s)
-------------------------------------------------------------------------------
Language                     files          blank        comment           code
-------------------------------------------------------------------------------
Python                         120           8305          31980          32335
-------------------------------------------------------------------------------
SUM:                           120           8305          31980          32335
-------------------------------------------------------------------------------
```

Evolution of QIIME



About UNIX

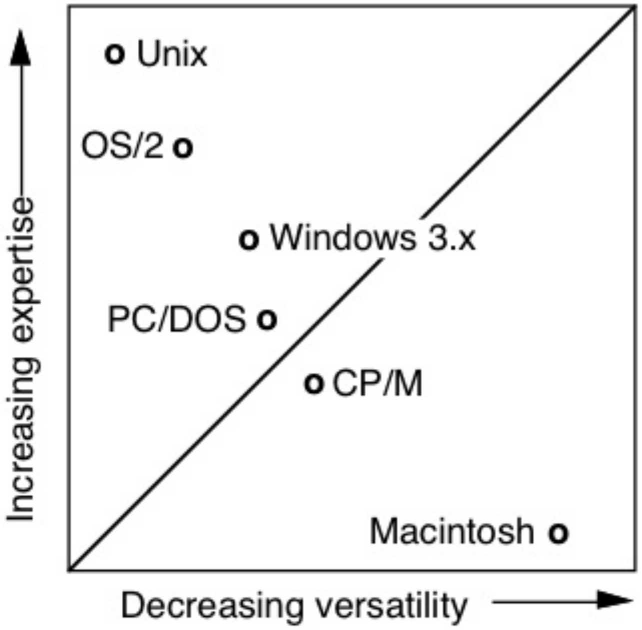


Figure 1.11 Disk operating systems–the compromise between DOS versatility and user expertise in real-time applications.

circa 1992

Taken from, Biomedical Digital Signal Processing, C-Language Examples and Laboratory Experiments for the IBM PC, Willis J. Tompkins Editor p. 18.