NSF RCN QIIME Workshop
3-4 October 2014

www.qiime.org
http://bit.ly/acadia-rcn-qiime

Greg Caporaso
Jai Rideout
www.caporasolab.us
www.applied-bioinformatics.org

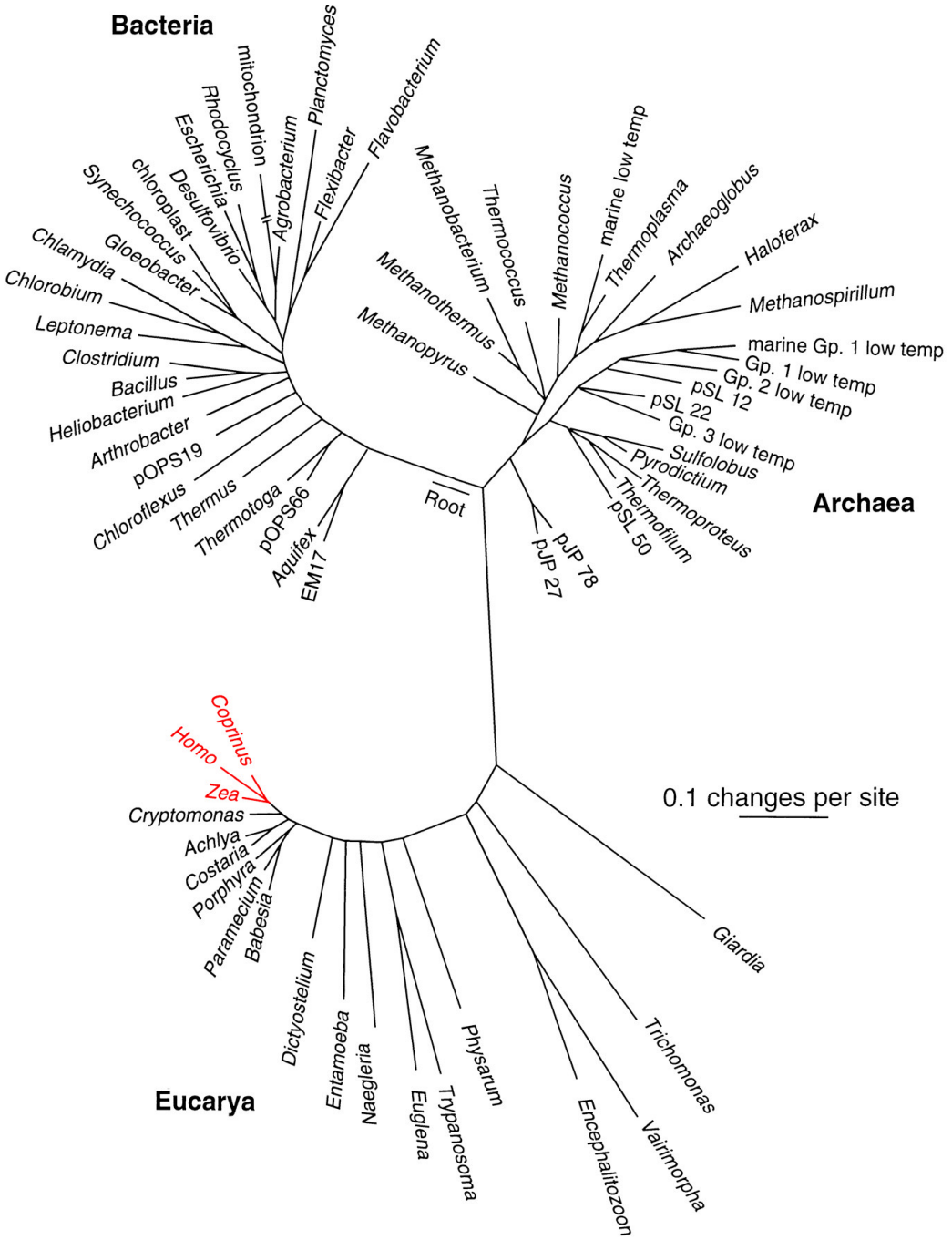
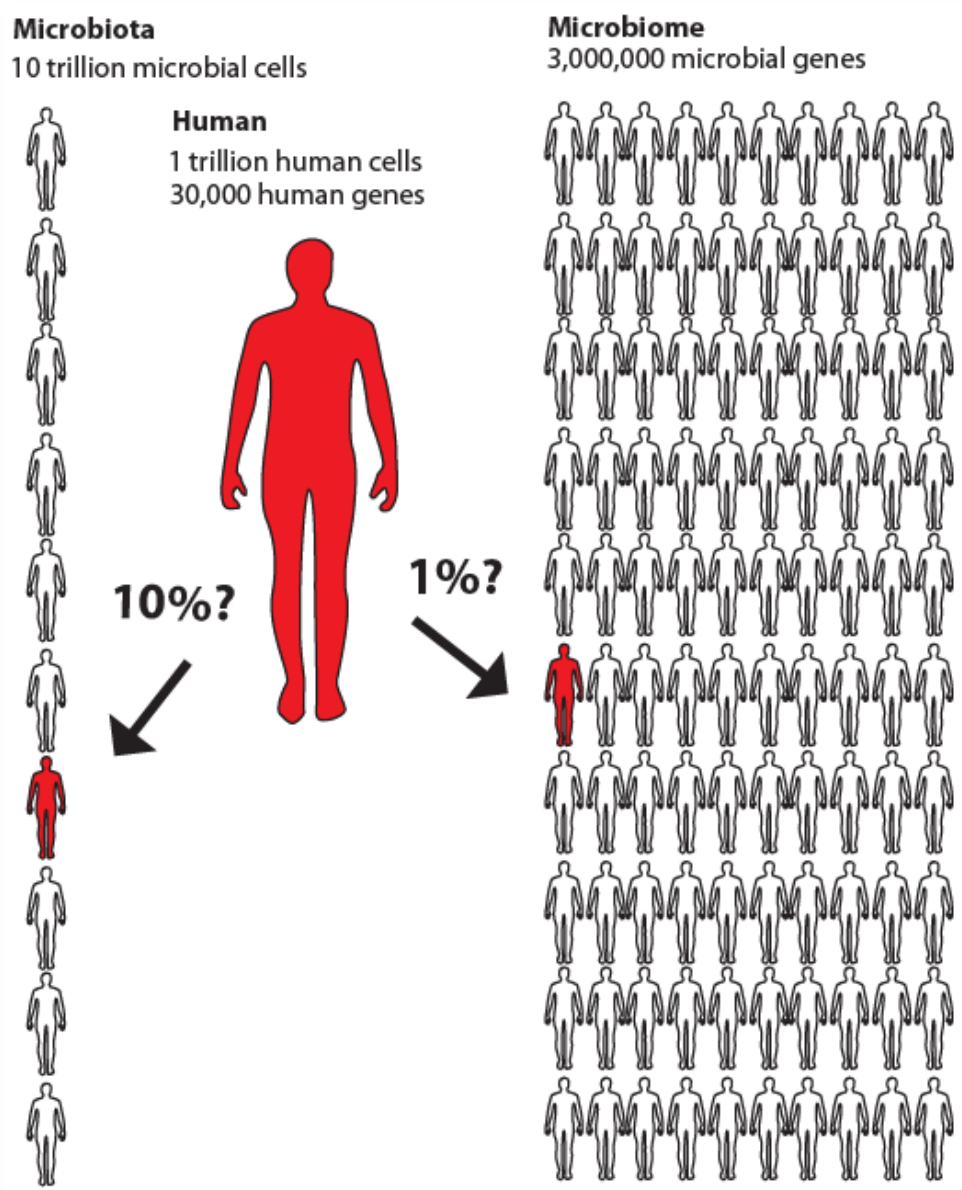
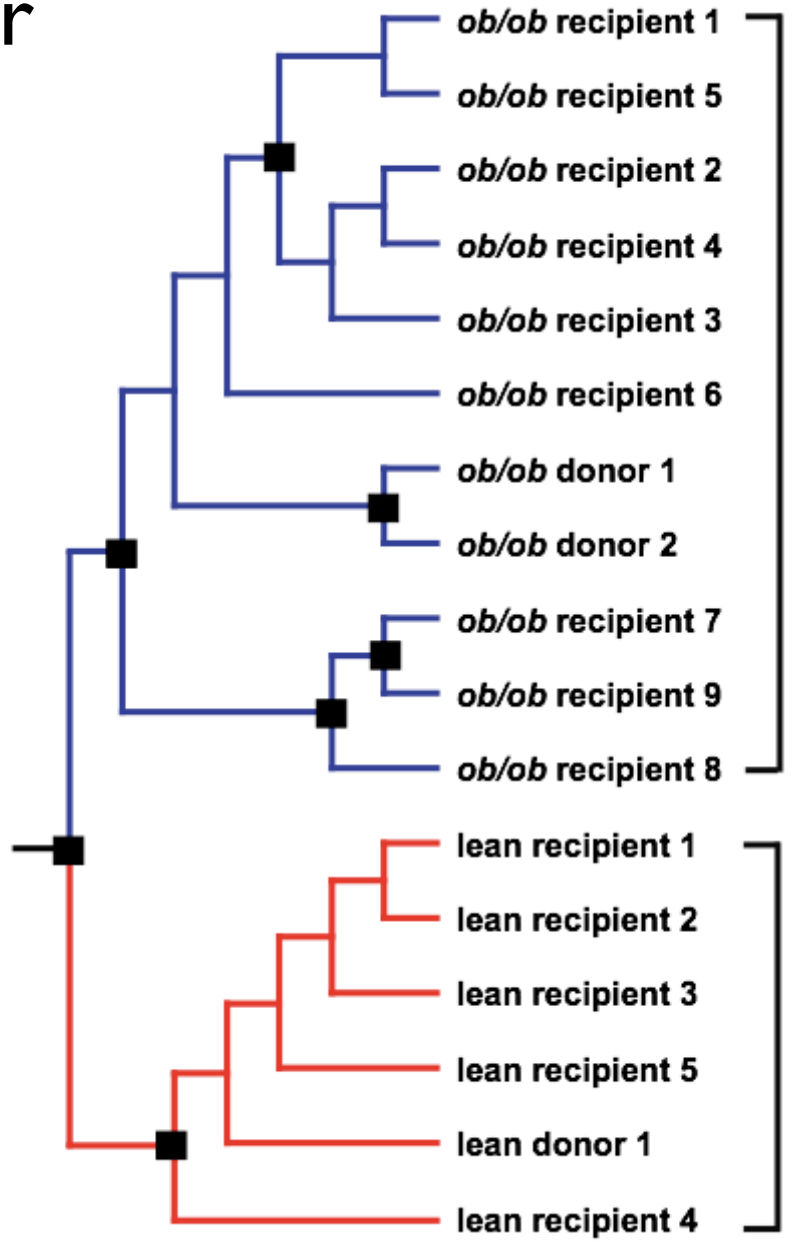


Image credit: Norman Pace



Microbial Ecology of the Gastrointestinal Tract
Annual Review of Microbiology **31**: 107–33. Savage, D. C. (1977).

Do differences in our microbiota matter?



Peter J. Turnbaugh et al., Nature 2006
An obesity-associated gut microbiome with increased capacity for energy harvest

Microbes rarely live or act alone.

Microbes rarely live or act alone.

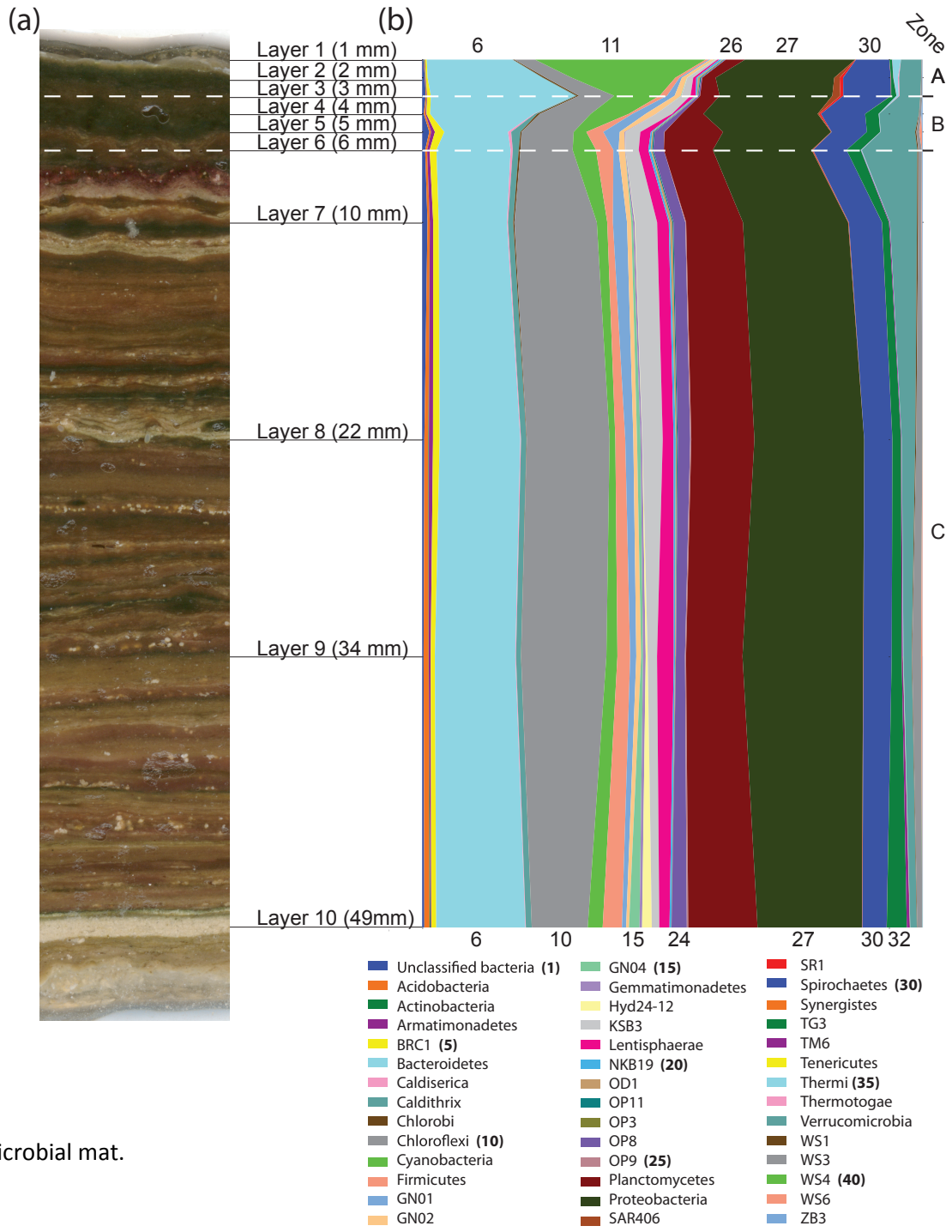


Image source:
Phylogenetic stratigraphy in the Guerrero Negro hypersaline microbial mat.
Harris, Caporaso *et al.* (2012)
International Society for Microbial Ecology Journal

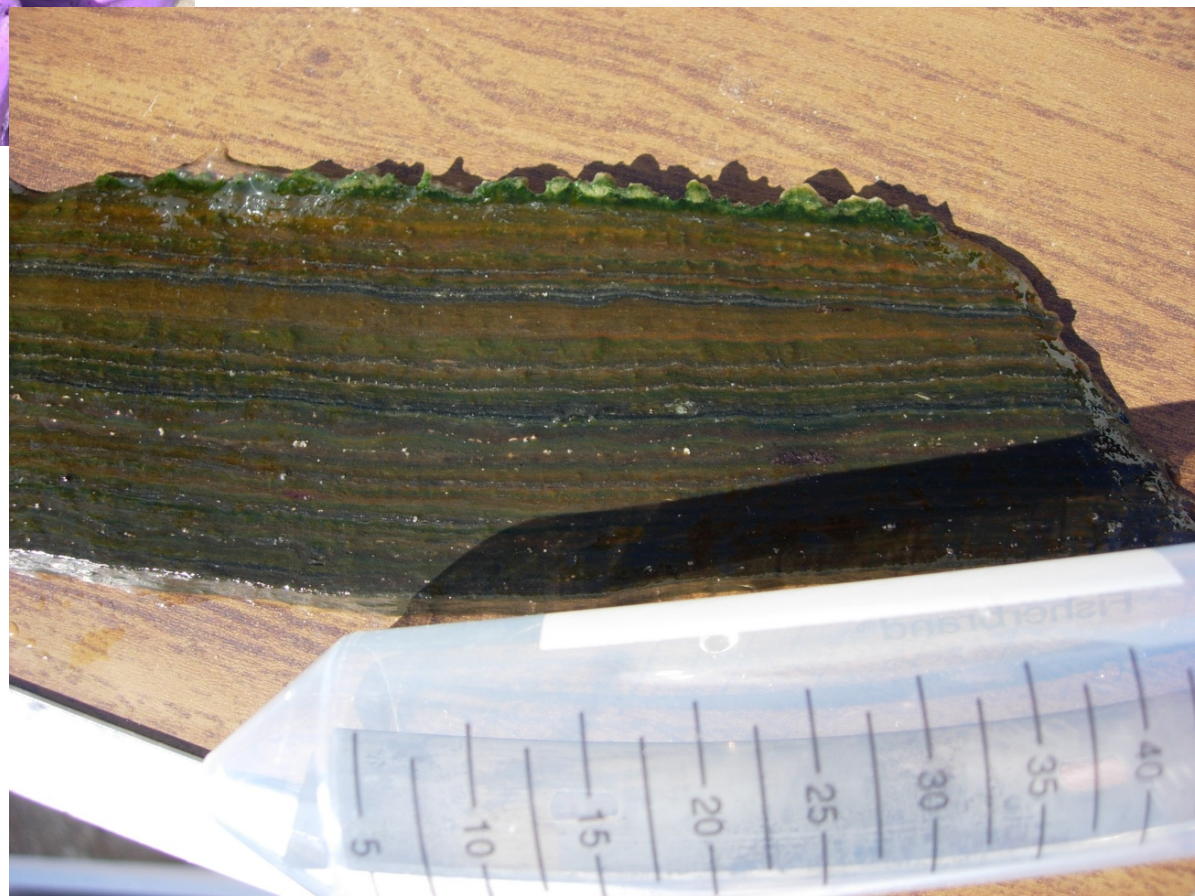
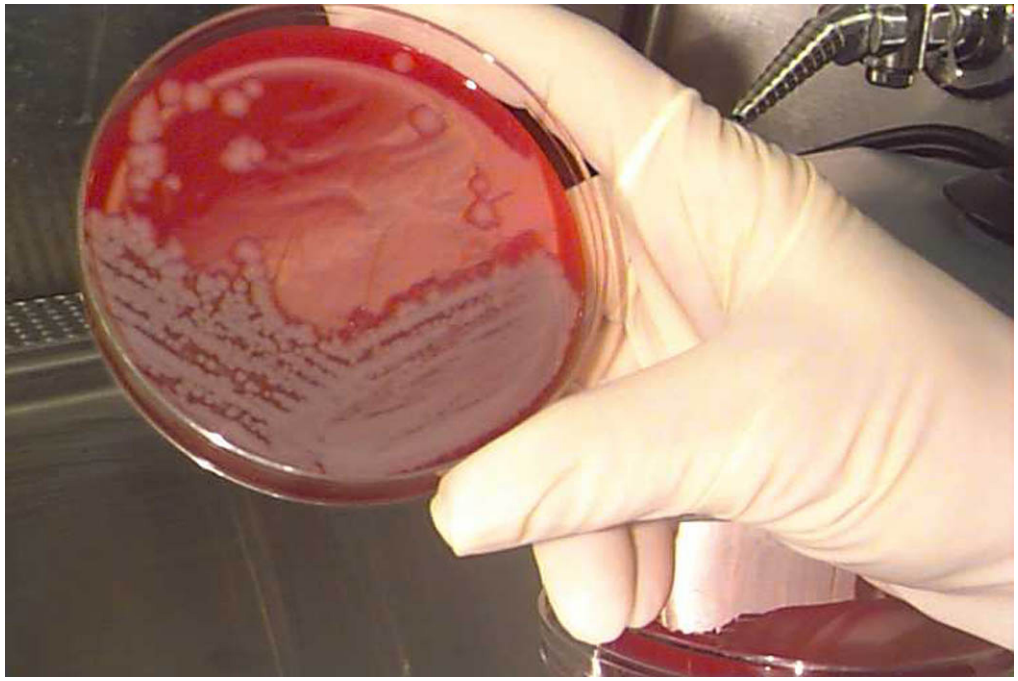


Photo credit: John Spear

Culturing microbes is hard

Back of the envelope calculation: less than 13% of bacterial species* have a representative that has been grown in culture.

Many recent advances are based on *culture-independent* approaches for studying microbial communities.



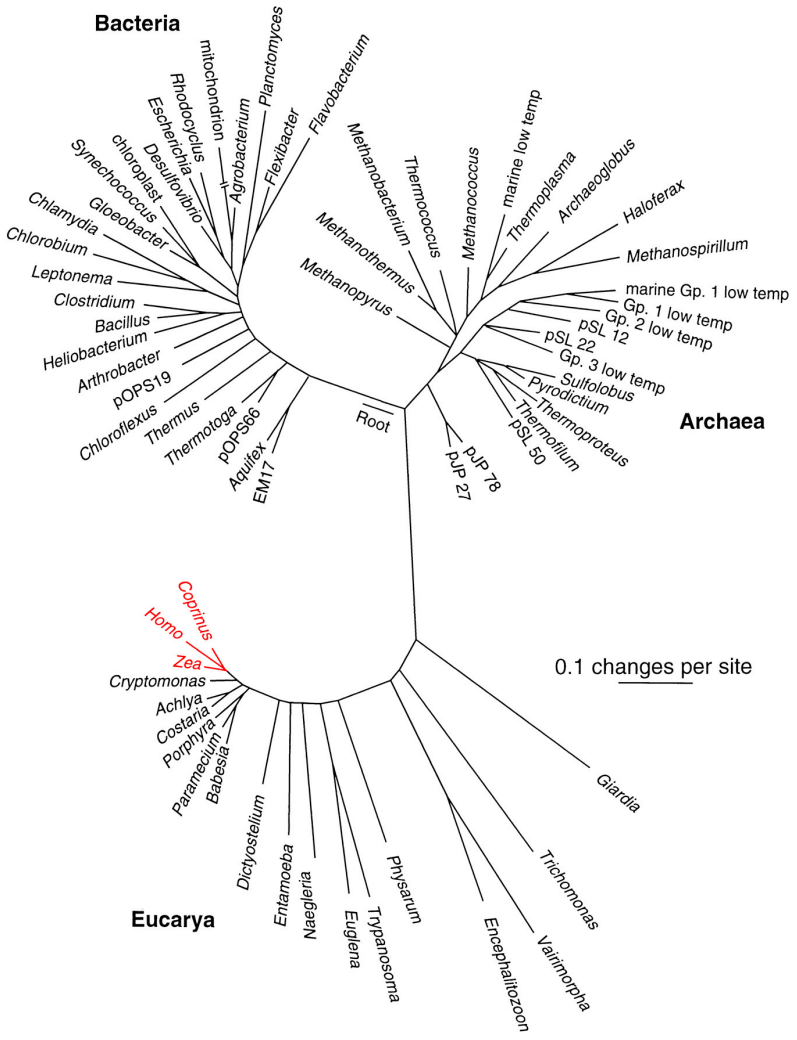
Bacillus anthracis in culture

* Defined as 97% OTUs in the Greengenes 13_5 reference database.

Culture-independent investigation of microbial communities

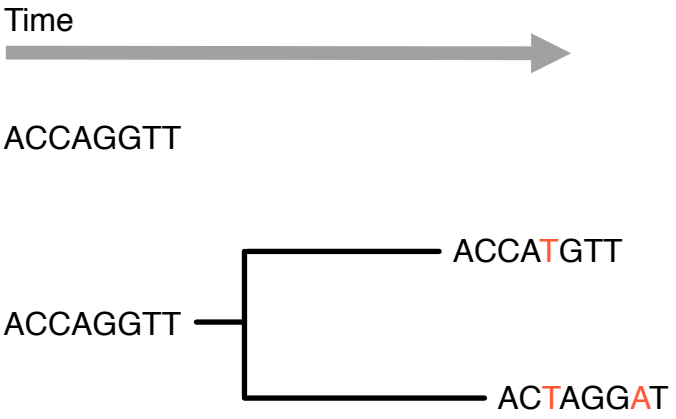
All cellular life has a shared evolutionary history, and some genes are shared by all organisms.

The sequence of those genes can be used as a *genetic fingerprint* for different organisms.

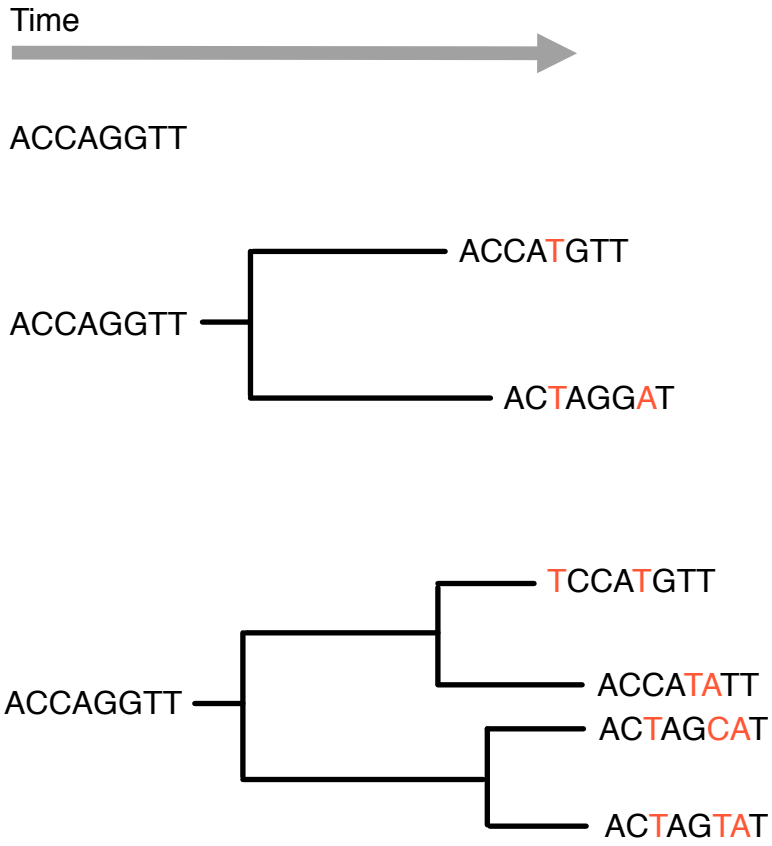


ACCAGGTT

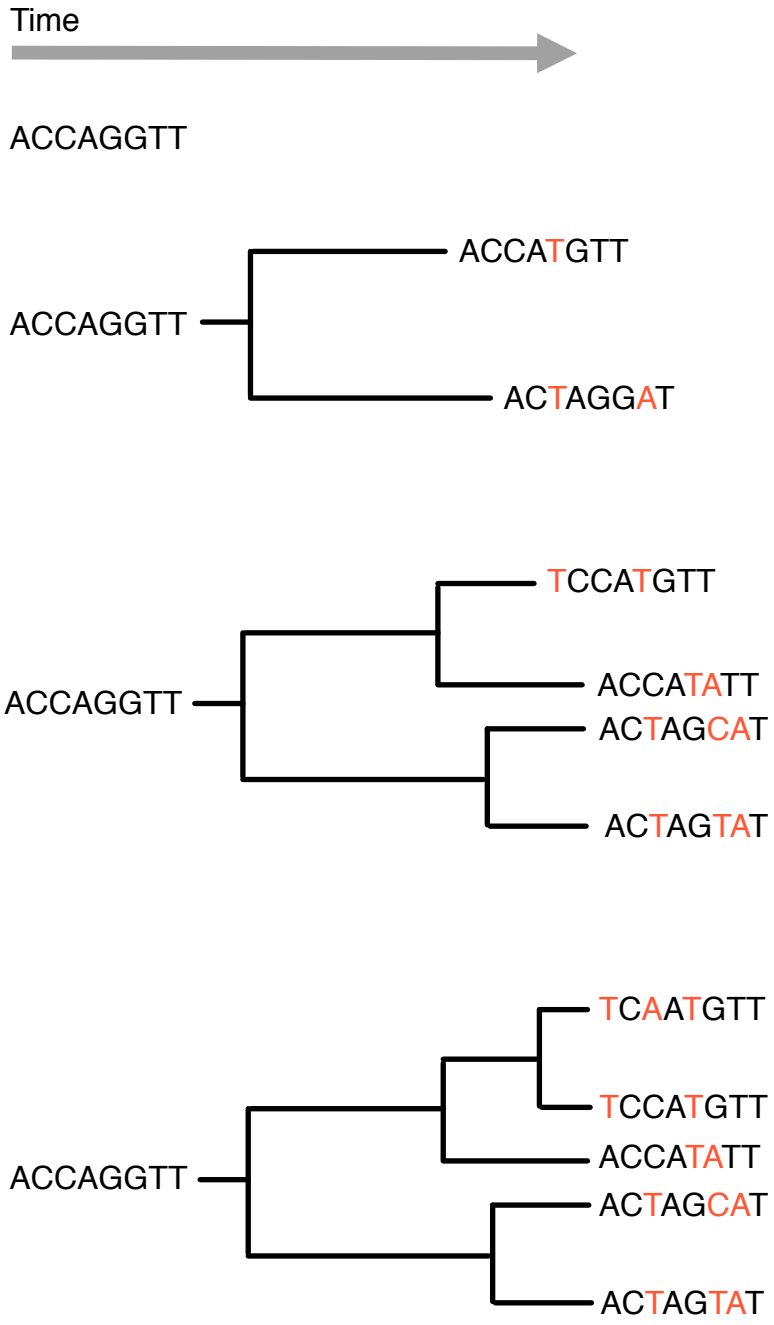
The random accumulation of *mutations* (changes to gene sequences over evolutionary time) gives us information for identifying and comparing organisms.



The random accumulation of *mutations* (changes to gene sequences over evolutionary time) gives us information for identifying and comparing organisms.

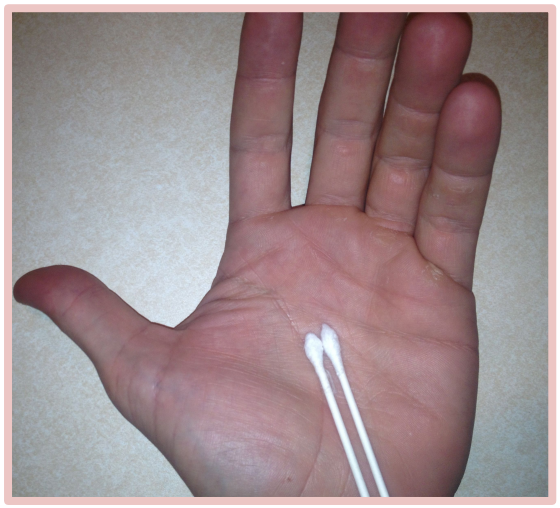


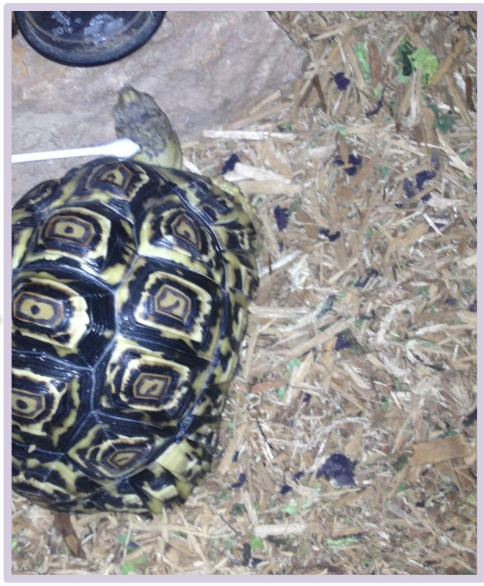
The random accumulation of *mutations* (changes to gene sequences over evolutionary time) gives us information for identifying and comparing organisms.

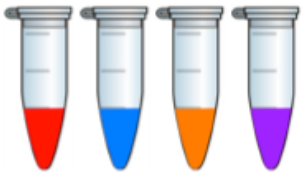


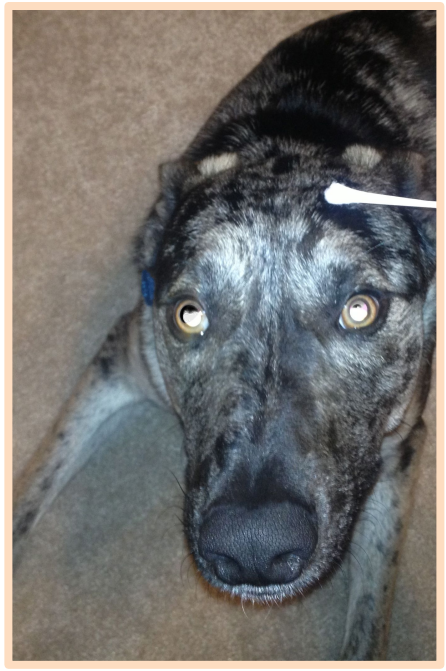
The random accumulation of *mutations* (changes to gene sequences over evolutionary time) gives us information for identifying and comparing organisms.

Collect samples









Extract DNA

(you can do this at home!)

Image source and instructions:
http://learn.genetics.utah.edu/content/labs/extraction/howto/

Isolate the *small subunit ribosomal RNA* gene to "fingerprint" different microbial organisms.

Why this gene?

- It's ubiquitous.
- Contains regions that identical across organisms, and regions that are variable across organisms.

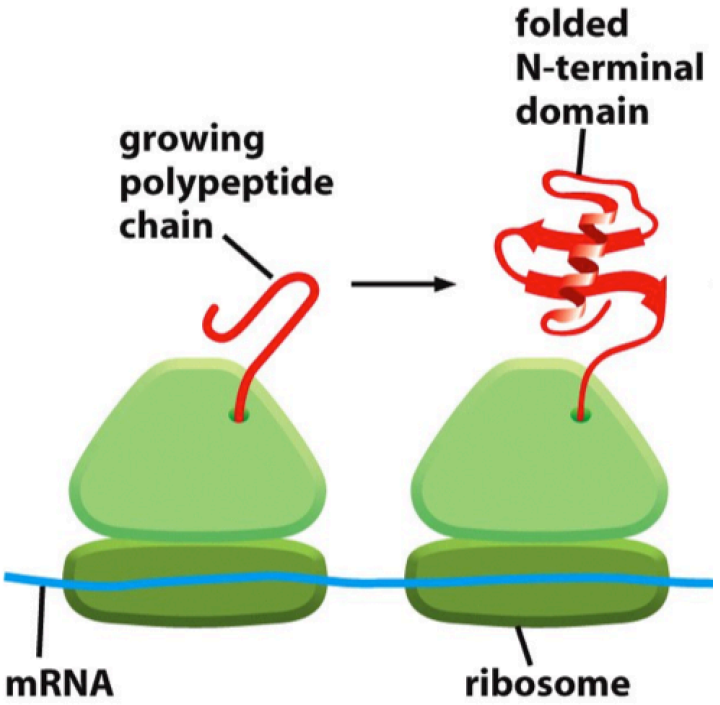
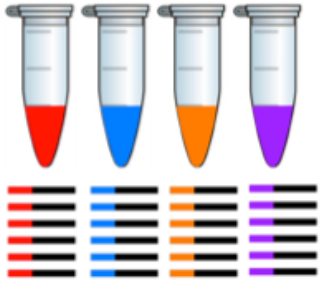


Figure 6-84 *Molecular Biology of the Cell* (© Garland Science 2008)

Sequence the rRNA from all samples on a “high-throughput” DNA sequencer

Per-sample rRNA



Pool samples and sequence

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

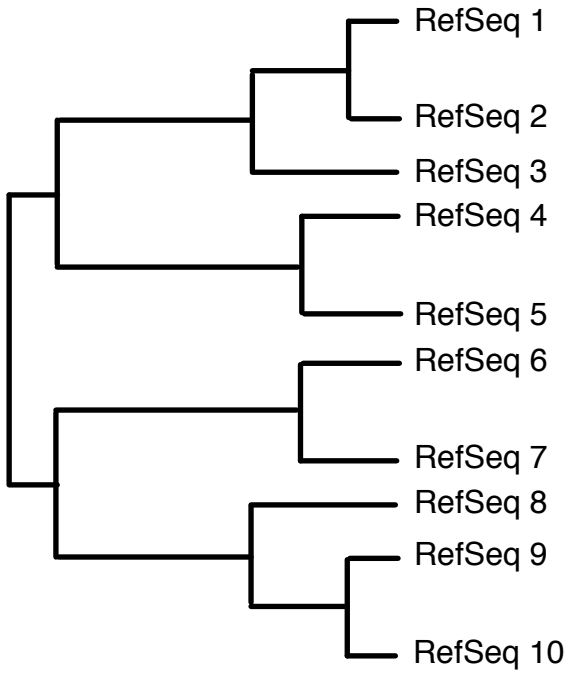
Micah Hamady, et al., Nature Methods, 2008.
Error-correcting barcodes for pyrosequencing hundreds of samples in multiplex.

Which microbial organisms are represented by the rRNA gene sequences in each sample?

rRNA reference database

```
>PC.634_1 FLP3FBN01ELBSX
CTGGGCCGTGTCTCAGTCCCAATGTGGCCGTTTACCCTCTCAGGCCGG
CTACGCATCATCGCCTTGGTGGGCCGTTACCTCACCAACTAGCTAATG
CGCCGCAGGTCCATCCATGTTCACGCCTTGATGGGCGCTTTAATATAC
TGAGCATGCGCTCTGTATACCTATCCGGTTTTAGCTACCGTTTCCAGC
AGTTATCCCGGACACATGGGCTAGG
>PC.634_2 FLP3FBN01EG8AX
TTGGACCGTGTCTCAGTTCCAATGTGGGGGCCTTCCTCTCAGAACCCC
TATCCATCGAAGGCTTGGTGGGCCGTTACCCCGCCAACAACCTAATGG
AACGCATCCCCATCGATGACCGAAGTTCTTTAATAGTTCTACCATGCG
GAAGAACTATGCCATCGGGTATTAATCTTTCTTTCGAAAGGCTATCCC
CGAGTCATCGGCAGGTTGGATACGTGTTACTCACCCGTGCGCCGGT
>PC.354_3 FLP3FBN01EEWKD
TTGGGCCGTGTCTCAGTCCCAATGTGGCCGATCAGTCTCTTAACTCGG
CTATGCATCATTGCCTTGGTAAGCCGTTACCTTACCAACTAGCTAATG
CACCGCAGGTCCATCCAAGAGTGATAGCAGAACCATCTTTCAAACTCT
AGACATGCGTCTAGTGTTGTTATCCGGTATTAGCATCTGTTTCCAGGT
GTTATCCCAGTCTCTTGGG
```

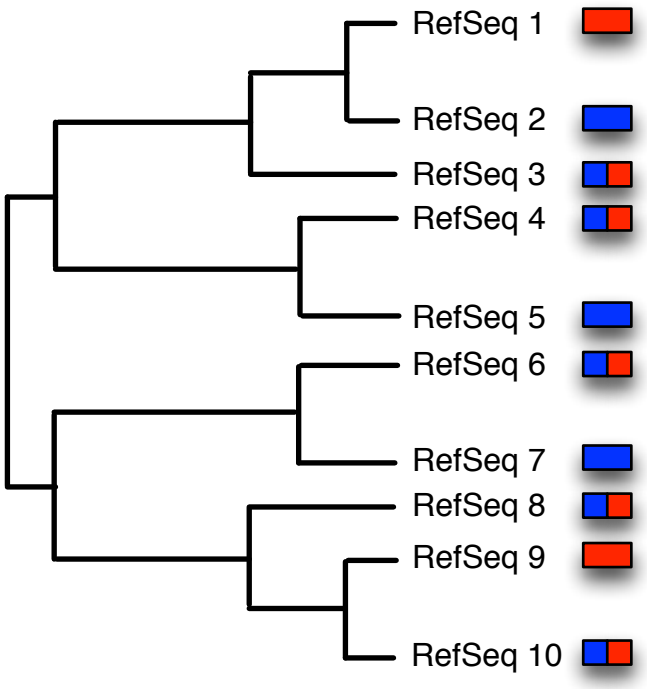
Search against reference sequences



Which microbial organisms are represented by the rRNA gene sequences in each sample?

```
>PC.634_1 FLP3FBN01ELBSX
CTGGGCCGTGTCTCAGTCCCAATGTGGCCGTTTACCCTCTCAGGCCGG
CTACGCATCATCGCCTTGGTGGGCCGTTACCTCACCAACTAGCTAATG
CGCCGCAGGTCCATCCATGTTCACGCCTTGATGGGCGCTTTAATATAC
TGAGCATGCGCTCTGTATACCTATCCGGTTTTAGCTACCGTTTCCAGC
AGTTATCCCGGACACATGGGCTAGG
>PC.634_2 FLP3FBN01EG8AX
TTGGACCGTGTCTCAGTTCCAATGTGGGGGCCTTCCTCTCAGAACCCC
TATCCATCGAAGGCTTGGTGGGCCGTTACCCCGCCAACAACCTAATGG
AACGCATCCCCATCGATGACCGAAGTTCTTTAATAGTTCTACCATGCG
GAAGAACTATGCCATCGGGTATTAATCTTTCTTTCGAAAGGCTATCCC
CGAGTCATCGGCAGGTTGGATACGTGTTACTCACCCGTGCGCCGGT
>PC.354_3 FLP3FBN01EEWKD
TTGGGCCGTGTCTCAGTCCCAATGTGGCCGATCAGTCTCTTAACTCGG
CTATGCATCATTGCCTTGGTAAGCCGTTACCTTACCAACTAGCTAATG
CACCGCAGGTCCATCCAAGAGTGATAGCAGAACCATCTTTCAAACTCT
AGACATGCGTCTAGTGTTGTTATCCGGTATTAGCATCTGTTTCCAGGT
GTTATCCCAGTCTCTTGGG
```

Search against reference sequences

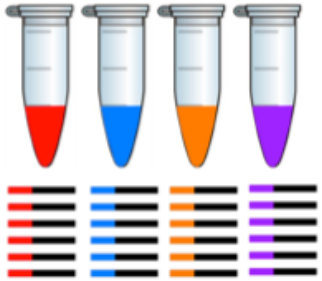


Comparing microbial communities

Who is there?

How many “species” are there?

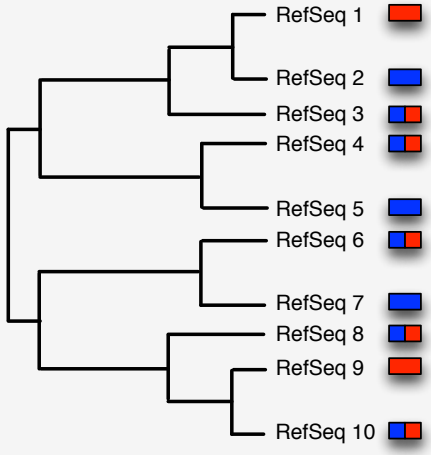
How similar are pairs of samples?



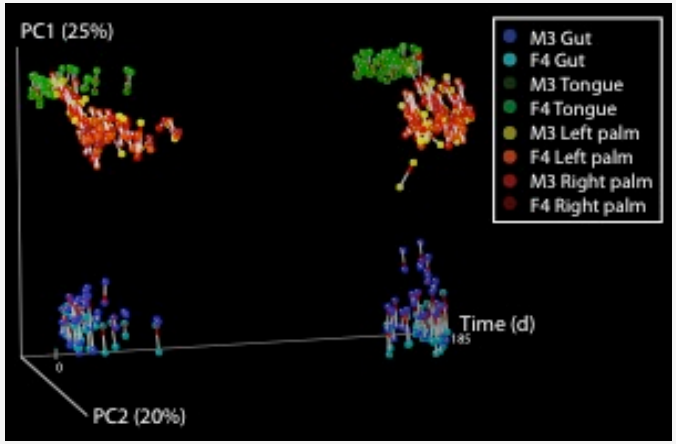
qiime
www.qiime.org

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

Assign reads to samples



Assign millions of sequences from thousands of samples to reference



Compare samples statistically and visually

A few of the 1754 articles citing QIIME (Google Scholar, 27 Aug 2014)

Is It Time for a Metagenomic Basis of Therapeutics?

Henry J. Haiser and Peter J. Turnbaugh*

The trillions of microbes associated with the human body are a key part of a comprehensive view of pharmacology. A mechanistic understanding of how the gut microbiota directly and indirectly affects drug metabolism is beginning to emerge.

Science (2012)

Gastrointestinal Microbiome Signatures of Pediatric Patients With Irritable Bowel Syndrome

Delphine M. Saulnier*, ‡, §, Kevin Riehle‖, ¶, Toni–Ann Mistretta*, ‡, Maria–Alejandra Diaz*, ‡, Debasmita Mandal#, Sabeen Raza*, ‡, Erica M. Weidler**, ‡‡, Xiang Qin§§, Cristian Coarfa‖, ¶, Aleksandar Milosavljevic‖, ¶, Joseph F. Petrosino§§, ‖‖, ¶¶, Sarah Highlander§§, ‖‖, Richard Gibbs§§, Susan V. Lynch#, Robert J. Shulman**, ‡‡, James Versalovic*, ‡, ‖, **, ‖‖,

Gastroenterology (2011)

Hydrocarbon-Degrading Bacteria and the Bacterial Community Response in Gulf of Mexico Beach Sands Impacted by the Deepwater Horizon Oil Spill

Joel E. Kostka[1,*], Om Prakash[1], Will A. Overholt[1], Stefan J. Green[1,3], Gina Freyer[1,2], Andy Canion[1], Jonathan Delgardio[1], Nikita Norton[1], Terry C. Hazen[4] and Markus Huettel[1]

Applied and Environmental Microbiology (2011)

Obesity (2012); **20** 4, 738–747. doi:10.1038/oby.2011.111

Responses of Gut Microbiota to Diet Composition and Weight Loss in Lean and Obese Mice

Yann Ravussin[1], Omry Koren[2], Ayme Spor[2], Charles LeDuc[1], Roee Gutman[1,5], Jesse Stombaugh[3], Rob Knight[3,4], Ruth E. Ley[2] and Rudolph L. Leibel[1]

The Microbiology of Malting and Brewing

Nicholas A. Bokulich[a,b] and Charles W. Bamforth[a]

Microbiology and Molecular Biology Reviews (2013)

Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice

Andrew L. Goodman[1], George Kallstrom, Jeremiah J. Faith, Alejandro Reyes, Aimee Moore, Gautam Dantas, and Jeffrey I. Gordon[2]

+ Author Affiliations

PNAS (2011)

Review

The Impact of the Gut Microbiota on Human Health: An Integrative View

Jose C. Clemente[1], Luke K. Ursell[1], Laura Wegener Parfrey[1], Rob Knight[1, 2],

Cell (2012)

Microbial diversity associated with four functional groups of benthic reef algae and the reef-building coral *Montastraea annularis*

Katie L. Barott[1,*], Beltran Rodriguez-Brito[1], Jan Janouškovec[2], Kristen L. Marhaver[3], Jennifer E. Smith[3], Patrick Keeling[2], Forest L. Rohwer[1]

Issue

Environmental Microbiology
Volume 13, Issue 5, pages 1192–1204, May 2011

Article first published online: 27 JAN 2011
DOI: 10.1111/j.1462-2920.2010.02419.x

OPEN ACCESS Freely available online

PLoS one

Fecal Microbiota in Premature Infants Prior to Necrotizing Enterocolitis

Volker Mai[1*], Christopher Michael Young[2], Maria Ukhanova[1], Xiaoyu Wang[1], Yijun Sun[1], George Casella[4], Douglas Theriaque[5], Nan Li[2], Renu Sharma[3], Mark Hudak[3], Josef Neu[2]

Current Concepts of the Intestinal Microbiota and the Pathogenesis of Infection

Leslie H. Wardwell · Curtis Huttenhower · Wendy S. Garrett

Current Infectious Disease Reports (2011)

Composition of Human Skin Microbiota Affects Attractiveness to Malaria Mosquitoes

Niels O. Verhulst[1*], Yu Tong Qiu[1], Hans Beijleveld[1,2], Chris Maliepaard[3], Dan Knights[4], Stefan Schulz[5], Donna Berg-Lyons[6], Christian L. Lauber[6], Willem Verduijn[7], Geert W. Haasnoot[7], Roland Mumm[8,9], Harro J. Bouwmeester[10], Frans H. J. Claas[7], Marcel Dicke[1], Joop J. A. van Loon[1], Willem Takken[1], Rob Knight[4,11], Renate C. Smallegange[1]

PLoS One (2011)

Learning QIIME

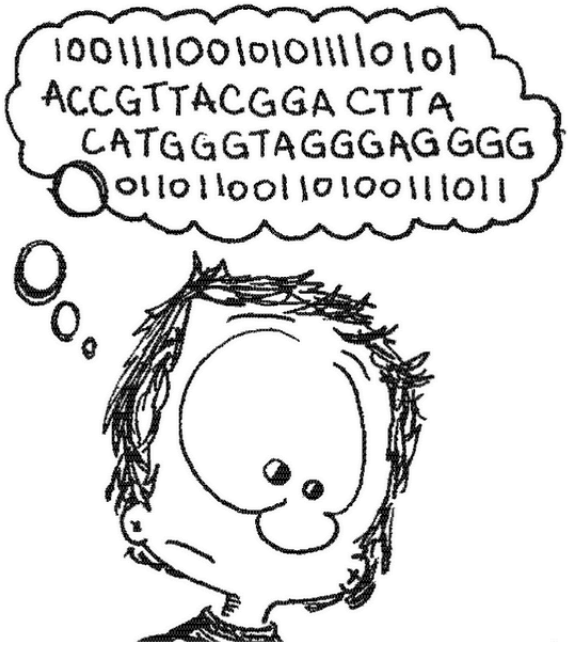
- Start with the tutorials
  - http://qiime.org/tutorials/index.html
- Call any script with –h to get help or see the script usage pages
  - http://qiime.org/documentation/script_index.html
- Ask questions on the QIIME Forum
  - http://forum.qiime.org
- Report bugs on the issue tracker
  - http://github.com/qiime/qiime/issues

An Introduction To Applied Bioinformatics

Interactive lessons in bioinformatics.

View the Project on GitHub
gregcaporaso/An-Introduction-To-Applied-Bioinformatics

Download ZIP File | Download TAR Ball | View On GitHub



Bioinformatics, as I see it, is the application of the tools of computer science (things like programming languages, algorithms, and databases) to address biological problems (for example, inferring the evolutionary relationship between a group of organisms based on fragments of their genomes, or understanding if or how the community of microorganisms that live in my gut changes if I modify my diet). Bioinformatics is a rapidly growing field, largely in response to the vast increase in the quantity of data that biologists now grapple with. Students from varied disciplines (e.g., biology, computer science, statistics, and biochemistry) and stages of their educational careers (undergraduate, graduate, or postdoctoral) are becoming interested in bioinformatics.

I teach bioinformatics at the undergraduate and graduate levels at Northern Arizona University. This repository contains some of the materials that I've developed in these courses, and represents an initial attempt to organize these materials in a standalone way. In some cases, I'm just linking out to other materials for now.

http://applied-bioinformatics.org

Key QIIME files

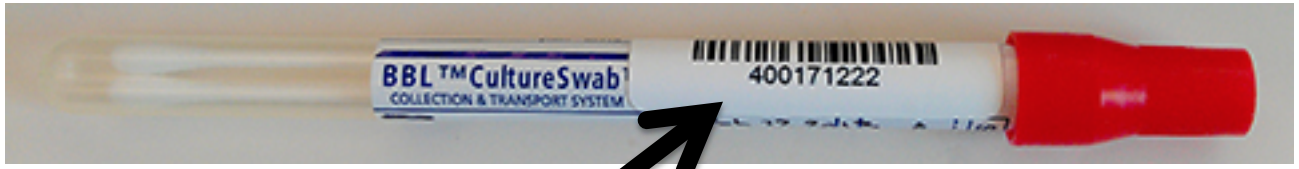
Metadata mapping file: per sample meta-data, user-defined (tab-separated text; we recommend using Google Docs)

Sequence files (in fasta, fastq, or sff format)

OTU table: sample x OTU matrix, central to downstream analyses (in BIOM format)

Phylogenetic tree (if applicable; in newick format)

Mapping file relates samples to variables



moving-pictures-qiime-tutorial_merged-map

File Edit View Insert Format Data Tools Help Last edit was made 2 minutes ago by gregcaporaso

gregcaporaso@gmail.com

Comments Share

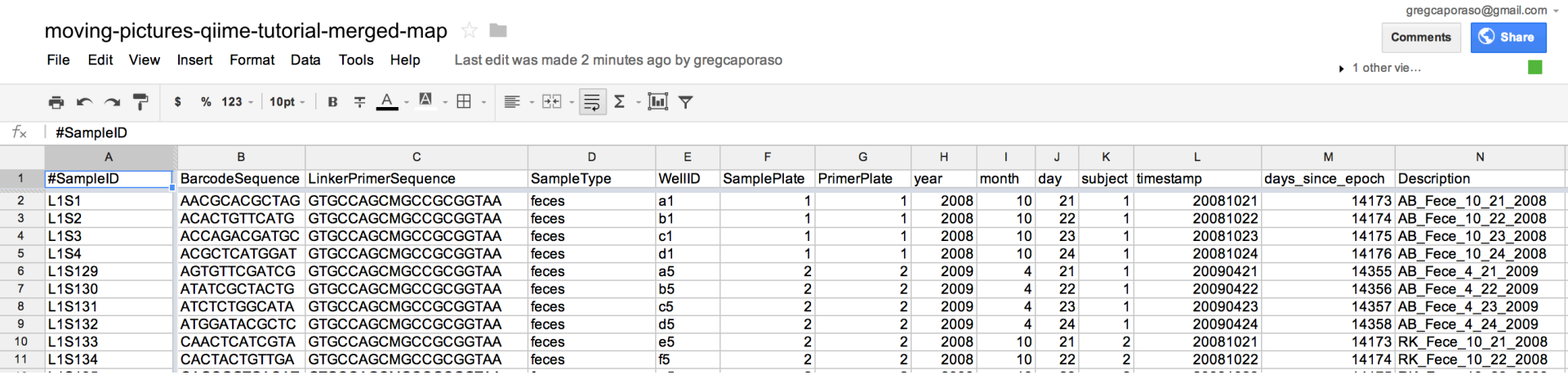
1 other vie…

10pt

#SampleID

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	#SampleID	BarcodeSequence	LinkerPrimerSequence	SampleType	WellID	SamplePlate	PrimerPlate	year	month	day	subject	timestamp	days_since_epoch	Description
2	L1S1	AACGCACGCTAG	GTGCCAGCMGCCGCGGTAA	feces	a1	1	1	2008	10	21	1	20081021	14173	AB_Fece_10_21_2008
3	L1S2	ACACTGTTCATG	GTGCCAGCMGCCGCGGTAA	feces	b1	1	1	2008	10	22	1	20081022	14174	AB_Fece_10_22_2008
4	L1S3	ACCAGACGATGC	GTGCCAGCMGCCGCGGTAA	feces	c1	1	1	2008	10	23	1	20081023	14175	AB_Fece_10_23_2008
5	L1S4	ACGCTCATGGAT	GTGCCAGCMGCCGCGGTAA	feces	d1	1	1	2008	10	24	1	20081024	14176	AB_Fece_10_24_2008
6	L1S129	AGTGTTCGATCG	GTGCCAGCMGCCGCGGTAA	feces	a5	2	2	2009	4	21	1	20090421	14355	AB_Fece_4_21_2009
7	L1S130	ATATCGCTACTG	GTGCCAGCMGCCGCGGTAA	feces	b5	2	2	2009	4	22	1	20090422	14356	AB_Fece_4_22_2009
8	L1S131	ATCTCTGGCATA	GTGCCAGCMGCCGCGGTAA	feces	c5	2	2	2009	4	23	1	20090423	14357	AB_Fece_4_23_2009
9	L1S132	ATGGATACGCTC	GTGCCAGCMGCCGCGGTAA	feces	d5	2	2	2009	4	24	1	20090424	14358	AB_Fece_4_24_2009
10	L1S133	CAACTCATCGTA	GTGCCAGCMGCCGCGGTAA	feces	e5	2	2	2008	10	21	2	20081021	14173	RK_Fece_10_21_2008
11	L1S134	CACTACTGTTGA	GTGCCAGCMGCCGCGGTAA	feces	f5	2	2	2008	10	22	2	20081022	14174	RK_Fece_10_22_2008

Metadata mapping file – create these in Google Docs (see http://bit.ly/mphm-qiime)



moving-pictures-qiime-tutorial-merged-map

gregcaporaso@gmail.com

Comments Share

File Edit View Insert Format Data Tools Help Last edit was made 2 minutes ago by gregcaporaso

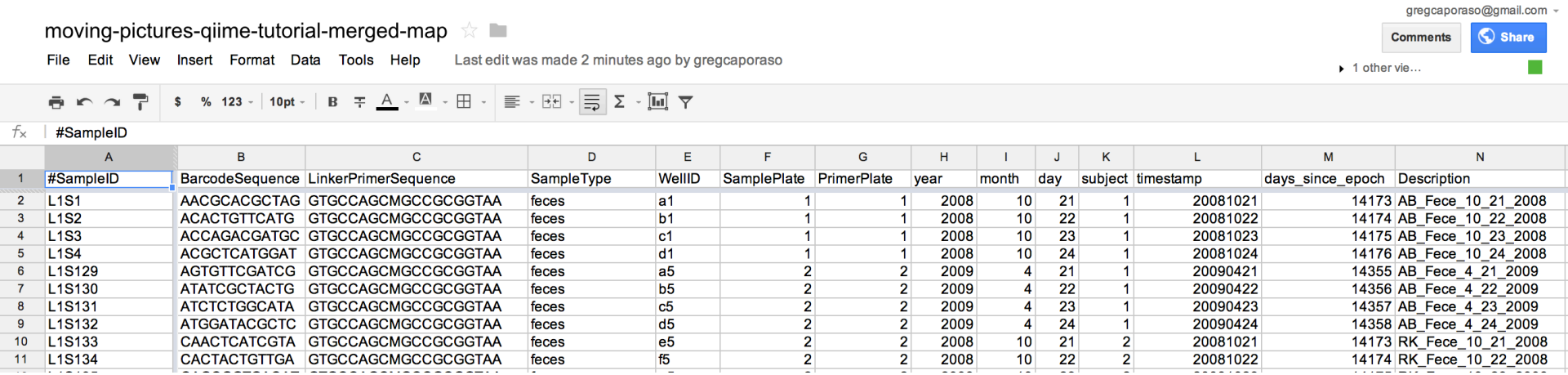
1 other vie…

$ % 123 10pt

fx #SampleID

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	#SampleID	BarcodeSequence	LinkerPrimerSequence	SampleType	WellID	SamplePlate	PrimerPlate	year	month	day	subject	timestamp	days_since_epoch	Description
2	L1S1	AACGCACGCTAG	GTGCCAGCMGCCGCGGTAA	feces	a1	1	1	2008	10	21	1	20081021	14173	AB_Fece_10_21_2008
3	L1S2	ACACTGTTCATG	GTGCCAGCMGCCGCGGTAA	feces	b1	1	1	2008	10	22	1	20081022	14174	AB_Fece_10_22_2008
4	L1S3	ACCAGACGATGC	GTGCCAGCMGCCGCGGTAA	feces	c1	1	1	2008	10	23	1	20081023	14175	AB_Fece_10_23_2008
5	L1S4	ACGCTCATGGAT	GTGCCAGCMGCCGCGGTAA	feces	d1	1	1	2008	10	24	1	20081024	14176	AB_Fece_10_24_2008
6	L1S129	AGTGTTCGATCG	GTGCCAGCMGCCGCGGTAA	feces	a5	2	2	2009	4	21	1	20090421	14355	AB_Fece_4_21_2009
7	L1S130	ATATCGCTACTG	GTGCCAGCMGCCGCGGTAA	feces	b5	2	2	2009	4	22	1	20090422	14356	AB_Fece_4_22_2009
8	L1S131	ATCTCTGGCATA	GTGCCAGCMGCCGCGGTAA	feces	c5	2	2	2009	4	23	1	20090423	14357	AB_Fece_4_23_2009
9	L1S132	ATGGATACGCTC	GTGCCAGCMGCCGCGGTAA	feces	d5	2	2	2009	4	24	1	20090424	14358	AB_Fece_4_24_2009
10	L1S133	CAACTCATCGTA	GTGCCAGCMGCCGCGGTAA	feces	e5	2	2	2008	10	21	2	20081021	14173	RK_Fece_10_21_2008
11	L1S134	CACTACTGTTGA	GTGCCAGCMGCCGCGGTAA	feces	f5	2	2	2008	10	22	2	20081022	14174	RK_Fece_10_22_2008

Metadata mapping file – create these in Google Docs (see http://bit.ly/mphm-qiime)



gregcaporaso@gmail.com

moving-pictures-qiime-tutorial-merged-map

File Edit View Insert Format Data Tools Help Last edit was made 2 minutes ago by gregcaporaso

Comments Share

1 other vie…

$ % 123 10pt

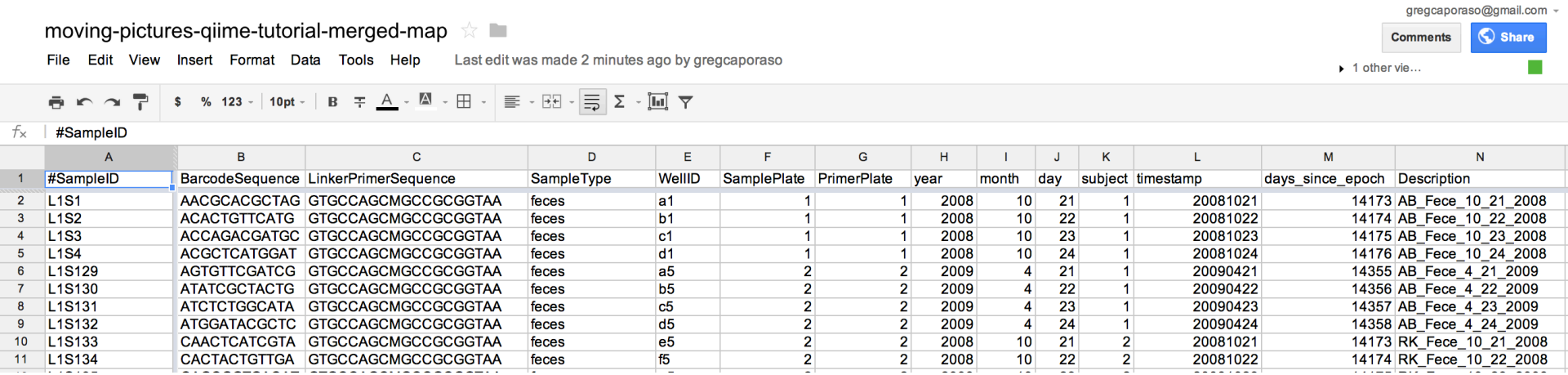
#SampleID

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	#SampleID	BarcodeSequence	LinkerPrimerSequence	SampleType	WellID	SamplePlate	PrimerPlate	year	month	day	subject	timestamp	days_since_epoch	Description
2	L1S1	AACGCACGCTAG	GTGCCAGCMGCCGCGGTAA	feces	a1	1	1	2008	10	21	1	20081021	14173	AB_Fece_10_21_2008
3	L1S2	ACACTGTTCATG	GTGCCAGCMGCCGCGGTAA	feces	b1	1	1	2008	10	22	1	20081022	14174	AB_Fece_10_22_2008
4	L1S3	ACCAGACGATGC	GTGCCAGCMGCCGCGGTAA	feces	c1	1	1	2008	10	23	1	20081023	14175	AB_Fece_10_23_2008
5	L1S4	ACGCTCATGGAT	GTGCCAGCMGCCGCGGTAA	feces	d1	1	1	2008	10	24	1	20081024	14176	AB_Fece_10_24_2008
6	L1S129	AGTGTTCGATCG	GTGCCAGCMGCCGCGGTAA	feces	a5	2	2	2009	4	21	1	20090421	14355	AB_Fece_4_21_2009
7	L1S130	ATATCGCTACTG	GTGCCAGCMGCCGCGGTAA	feces	b5	2	2	2009	4	22	1	20090422	14356	AB_Fece_4_22_2009
8	L1S131	ATCTCTGGCATA	GTGCCAGCMGCCGCGGTAA	feces	c5	2	2	2009	4	23	1	20090423	14357	AB_Fece_4_23_2009
9	L1S132	ATGGATACGCTC	GTGCCAGCMGCCGCGGTAA	feces	d5	2	2	2009	4	24	1	20090424	14358	AB_Fece_4_24_2009
10	L1S133	CAACTCATCGTA	GTGCCAGCMGCCGCGGTAA	feces	e5	2	2	2008	10	21	2	20081021	14173	RK_Fece_10_21_2008
11	L1S134	CACTACTGTTGA	GTGCCAGCMGCCGCGGTAA	feces	f5	2	2	2008	10	22	2	20081022	14174	RK_Fece_10_22_2008

Required fields for demultiplexing steps:
SampleID, BarcodeSequence, LinkerPrimerSequence, Description

Provide all of the information you have about your samples – metadata is the key to interpretation.

Metadata mapping file – create these in Google Docs (see http://bit.ly/mphm-qiime)



moving-pictures-qiime-tutorial-merged-map

File Edit View Insert Format Data Tools Help Last edit was made 2 minutes ago by gregcaporaso

gregcaporaso@gmail.com

Comments Share

1 other vie…

#SampleID

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	#SampleID	BarcodeSequence	LinkerPrimerSequence	SampleType	WellID	SamplePlate	PrimerPlate	year	month	day	subject	timestamp	days_since_epoch	Description
2	L1S1	AACGCACGCTAG	GTGCCAGCMGCCGCGGTAA	feces	a1	1	1	2008	10	21	1	20081021	14173	AB_Fece_10_21_2008
3	L1S2	ACACTGTTCATG	GTGCCAGCMGCCGCGGTAA	feces	b1	1	1	2008	10	22	1	20081022	14174	AB_Fece_10_22_2008
4	L1S3	ACCAGACGATGC	GTGCCAGCMGCCGCGGTAA	feces	c1	1	1	2008	10	23	1	20081023	14175	AB_Fece_10_23_2008
5	L1S4	ACGCTCATGGAT	GTGCCAGCMGCCGCGGTAA	feces	d1	1	1	2008	10	24	1	20081024	14176	AB_Fece_10_24_2008
6	L1S129	AGTGTTCGATCG	GTGCCAGCMGCCGCGGTAA	feces	a5	2	2	2009	4	21	1	20090421	14355	AB_Fece_4_21_2009
7	L1S130	ATATCGCTACTG	GTGCCAGCMGCCGCGGTAA	feces	b5	2	2	2009	4	22	1	20090422	14356	AB_Fece_4_22_2009
8	L1S131	ATCTCTGGCATA	GTGCCAGCMGCCGCGGTAA	feces	c5	2	2	2009	4	23	1	20090423	14357	AB_Fece_4_23_2009
9	L1S132	ATGGATACGCTC	GTGCCAGCMGCCGCGGTAA	feces	d5	2	2	2009	4	24	1	20090424	14358	AB_Fece_4_24_2009
10	L1S133	CAACTCATCGTA	GTGCCAGCMGCCGCGGTAA	feces	e5	2	2	2008	10	21	2	20081021	14173	RK_Fece_10_21_2008
11	L1S134	CACTACTGTTGA	GTGCCAGCMGCCGCGGTAA	feces	f5	2	2	2008	10	22	2	20081022	14174	RK_Fece_10_22_2008

More information on metadata:

MIMARKS/MIxS standards: http://www.nature.com/nbt/journal/v29/n5/full/nbt.1823.html

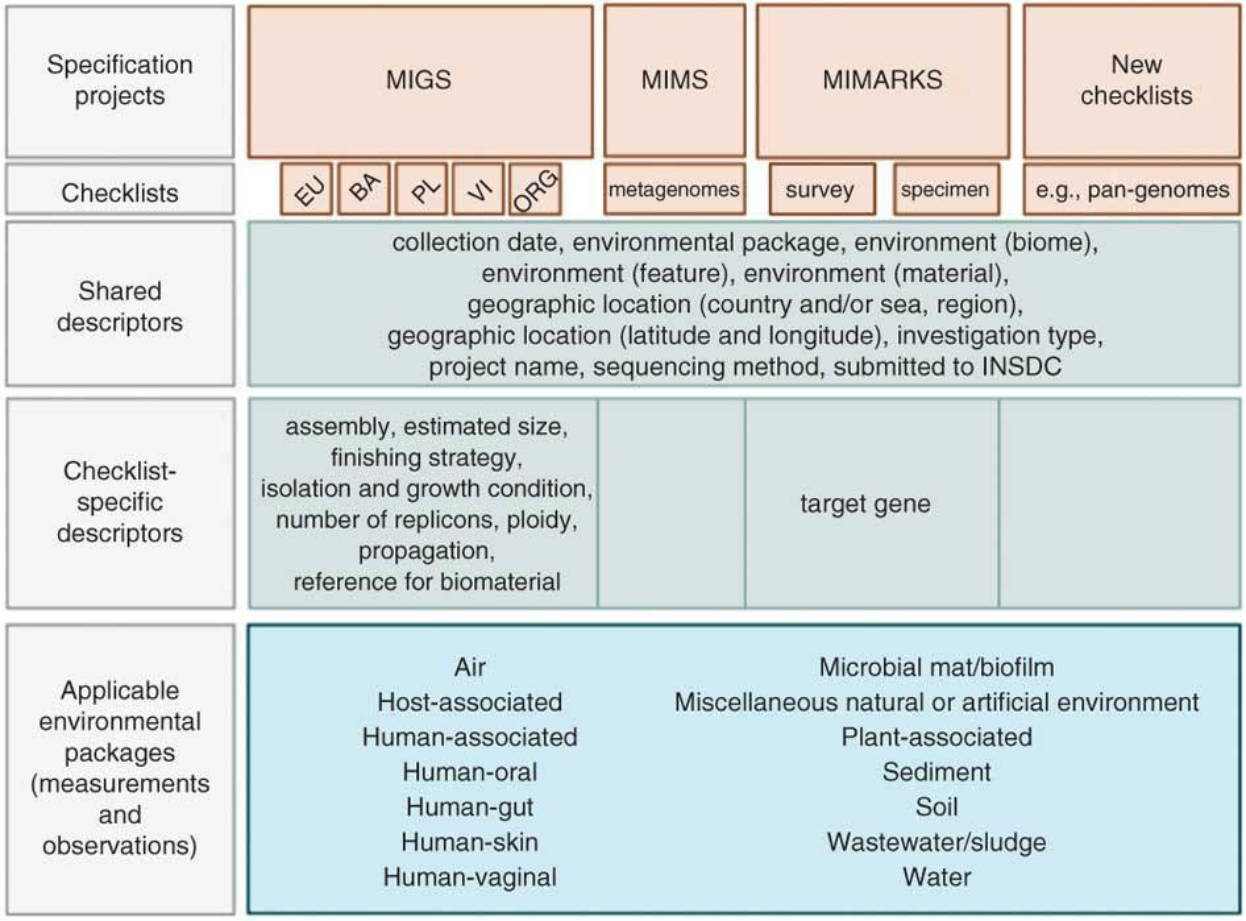
Formatting QIIME mapping files: http://qiime.org/documentation/file_formats.html

Working with Google Docs from QIIME (experimental, and will become more integrated in the future): http://qiime.org/tutorials/remote_mapping_files.html

For example, to download this mapping file you can run the command:

```
load_remote_mapping_file.py -k 0AvglGXLayhG7dGNuQWJKMlNWVFdVXzNlRXYybjFJV2c –o tutorial-map.txt
```

Metadata: GSC checklists



Slide credit: Doug Wendel

Sequences file (unprocessed): fastq, sff, fna/qual are currently supported.

These should be obtained as still-multiplexed data from your sequencing center (i.e., not yet mapped from barcode to sample). This allows for use of QIIME's quality filtering.

For more information:

Quality filtering Illumina data
http://www.nature.com/nmeth/journal/v10/n1/full/nmeth.2276.html

Demultiplexing 454 data with QIIME
http://qiime.org/tutorials/tutorial.html

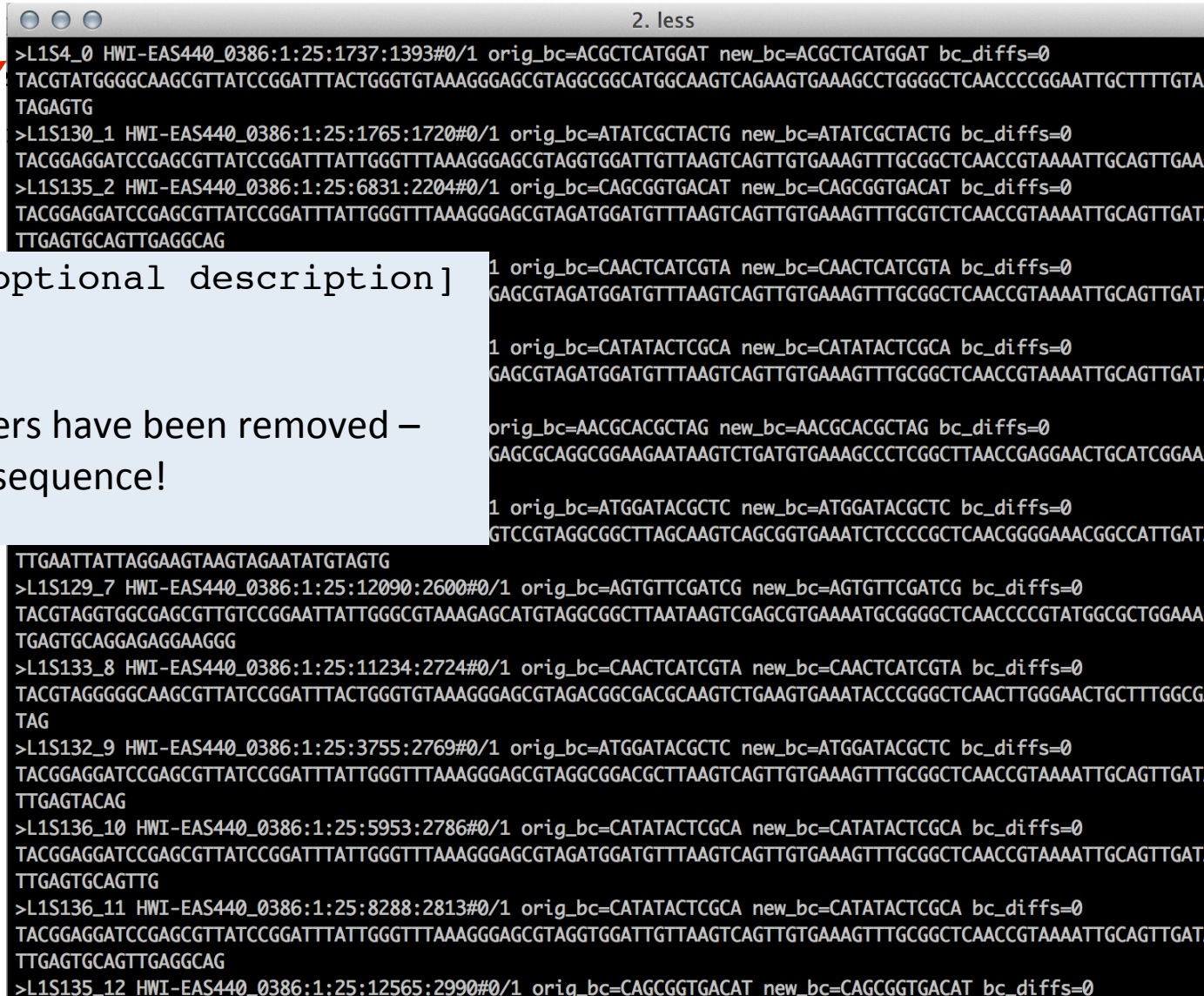
Denoising 454 data with QIIME
http://qiime.org/tutorials/denoising_454_data.html

Demultiplexing Illumina data with QIIME
http://qiime.org/tutorials/processing_illumina_data.html

Sequences file (post-split-libraries): standard multi-record fasta, with sample identifier included in the sequence identifier.

```
>sampleID_seqID [optional description]
ACCGA
```

Barcodes, primers, adapters have been removed – this is only the biological sequence!



Sequences file (post-split-libraries): standard multi-record fasta, with sample identifier included in the sequence identifier.

Sample IDs map sequences to metadata

2. less

```
>L1S4_0 HWI-EAS440_0386:1:25:1737:1393#0/1 orig_bc=ACGCTCATGGAT new_bc=ACGCTCATGGAT bc_diffs=0
TACGTATGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGGCGGCATGGCAAGTCAGAAGTGAAAGCCTGGGGCTCAACCCCGGAATTGCTTTTGTA
TAGAGTG
>L1S130_1 HWI-EAS440_0386:1:25:1765:1720#0/1 orig_bc=ATATCGCTACTG new_bc=ATATCGCTACTG bc_diffs=0
TACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGTGGATTGTTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGAA
>L1S135_2 HWI-EAS440_0386:1:25:6831:2204#0/1 orig_bc=CAGCGGTGACAT new_bc=CAGCGGTGACAT bc_diffs=0
TACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATGGATGTTTAAGTCAGTTGTGAAAGTTTGCGTCTCAACCGTAAAATTGCAGTTGAT
TTGAGTGCAGTTGAGGCAG
>L1S133_3 HWI-EAS440_0386:1:25:4317:2261#0/1 orig_bc=CAACTCATCGTA new_bc=CAACTCATCGTA bc_diffs=0
TACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATGGATGTTTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGAT
TTGAGTGCAG
>L1S136_4 HWI-EAS440_0386:1:25:9966:2383#0/1 orig_bc=CATATACTCGCA new_bc=CATATACTCGCA bc_diffs=0
TACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATGGATGTTTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGAT
TTGAGTGCAGTT
>L1S1_5 HWI-EAS440_0386:1:25:8283:2608#0/1 orig_bc=AACGCACGCTAG new_bc=AACGCACGCTAG bc_diffs=0
TACGTAGGTGGCAAGCGTTGTCCGGATTTATTGGGCGTAAAGCGAGCGCAGGCGGAAGAATAAGTCTGATGTGAAAGCCCTCGGCTTAACCGAGGAACTGCATCGGAA
TTGAGTGCAGAAGAGGAG
>L1S132_6 HWI-EAS440_0386:1:25:9842:2603#0/1 orig_bc=ATGGATACGCTC new_bc=ATGGATACGCTC bc_diffs=0
TACGGAGGATCCAAGCGTTATCCGGAATCATTGGGTTTAAAGGGTCCGTAGGCGGCTTAGCAAGTCAGCGGTGAAATCTCCCCGCTCAACGGGGAAACGGCCATTGAT
TTGAATTATTAGGAAGTAAGTAGAATATGTAGTG
>L1S129_7 HWI-EAS440_0386:1:25:12090:2600#0/1 orig_bc=AGTGTTCGATCG new_bc=AGTGTTCGATCG bc_diffs=0
```

moving-pictures-qiime-tutorial-merged-map

gregcaporaso@gmail.com

File Edit View Insert Format Data Tools Help Last edit was made 2 minutes ago by gregcaporaso

Comments Share

1 other vie…

$ % 123 10pt

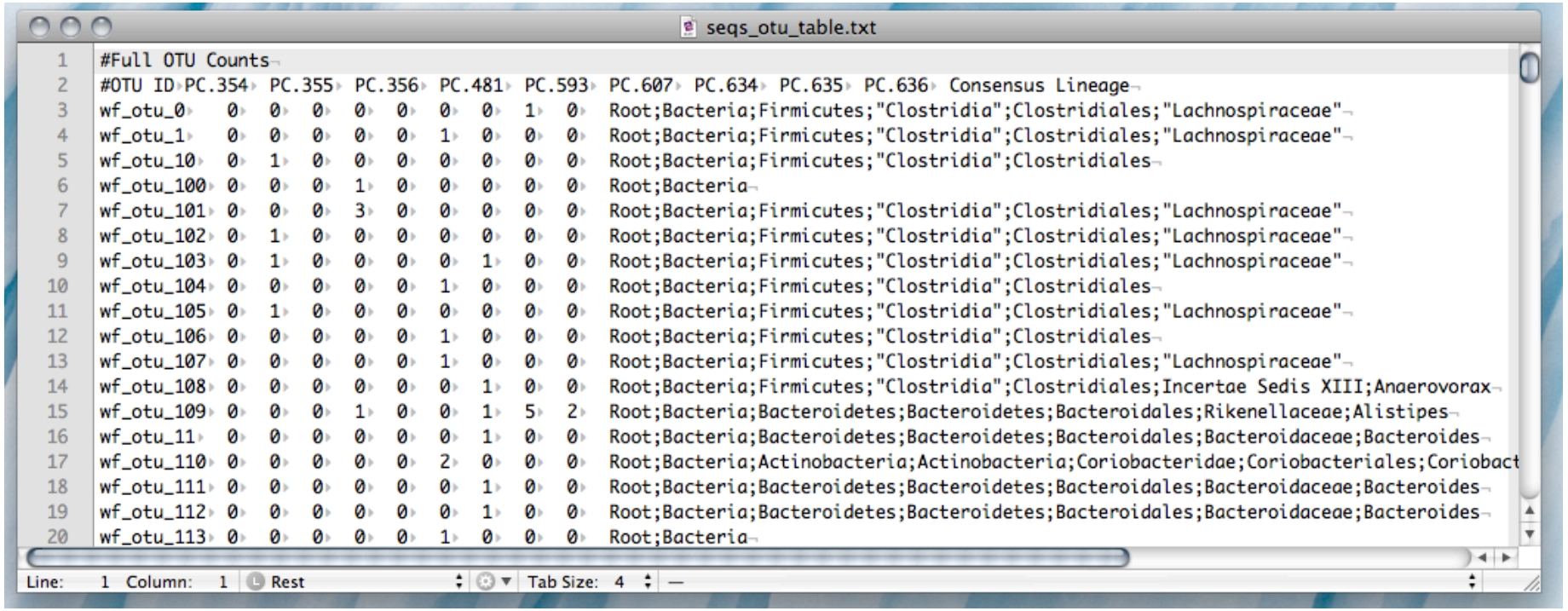
#SampleID

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	#SampleID	BarcodeSequence	LinkerPrimerSequence	SampleType	WellID	SamplePlate	PrimerPlate	year	month	day	subject	timestamp	days_since_epoch	Description
2	L1S1	AACGCACGCTAG	GTGCCAGCMGCCGCGGTAA	feces	a1	1	1	2008	10	21	1	20081021	14173	AB_Fece_10_21_2008
3	L1S2	ACACTGTTCATG	GTGCCAGCMGCCGCGGTAA	feces	b1	1	1	2008	10	22	1	20081022	14174	AB_Fece_10_22_2008
4	L1S3	ACCAGACGATGC	GTGCCAGCMGCCGCGGTAA	feces	c1	1	1	2008	10	23	1	20081023	14175	AB_Fece_10_23_2008
5	L1S4	ACGCTCATGGAT	GTGCCAGCMGCCGCGGTAA	feces	d1	1	1	2008	10	24	1	20081024	14176	AB_Fece_10_24_2008
6	L1S129	AGTGTTCGATCG	GTGCCAGCMGCCGCGGTAA	feces	a5	2	2	2009	4	21	1	20090421	14355	AB_Fece_4_21_2009
7	L1S130	ATATCGCTACTG	GTGCCAGCMGCCGCGGTAA	feces	b5	2	2	2009	4	22	1	20090422	14356	AB_Fece_4_22_2009
8	L1S131	ATCTCTGGCATA	GTGCCAGCMGCCGCGGTAA	feces	c5	2	2	2009	4	23	1	20090423	14357	AB_Fece_4_23_2009
9	L1S132	ATGGATACGCTC	GTGCCAGCMGCCGCGGTAA	feces	d5	2	2	2009	4	24	1	20090424	14358	AB_Fece_4_24_2009
10	L1S133	CAACTCATCGTA	GTGCCAGCMGCCGCGGTAA	feces	e5	2	2	2008	10	21	2	20081021	14173	RK_Fece_10_21_2008
11	L1S134	CACTACTGTTGA	GTGCCAGCMGCCGCGGTAA	feces	f5	2	2	2008	10	22	2	20081022	14174	RK_Fece_10_22_2008

OTU table

(classic format)

sample x OTU matrix



seqs_otu_table.txt

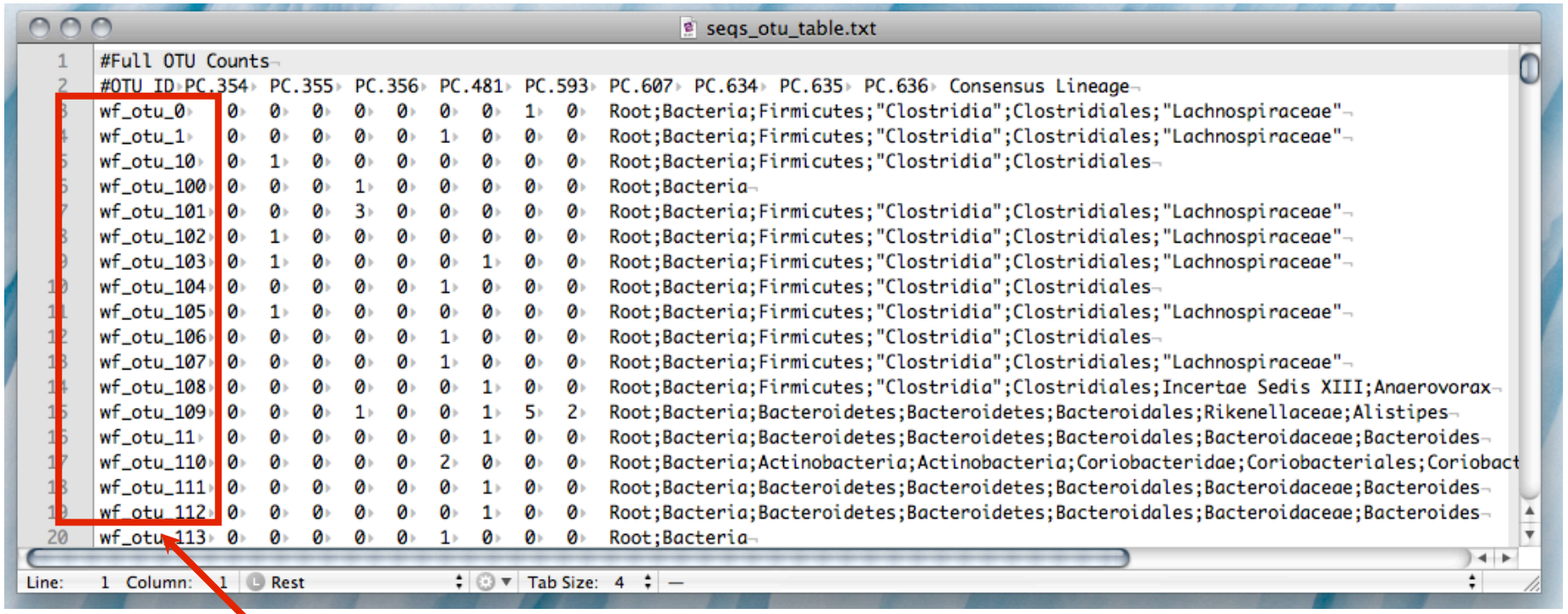
```
#Full OTU Counts
#OTU ID PC.354 PC.355 PC.356 PC.481 PC.593 PC.607 PC.634 PC.635 PC.636 Consensus Lineage
wf_otu_0    0  0  0  0  0  0  0  1  0  Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
wf_otu_1    0  0  0  0  0  1  0  0  0  Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"
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
```

Line: 1 Column: 1 Rest Tab Size: 4

OTU table

(classic format)

sample x OTU matrix

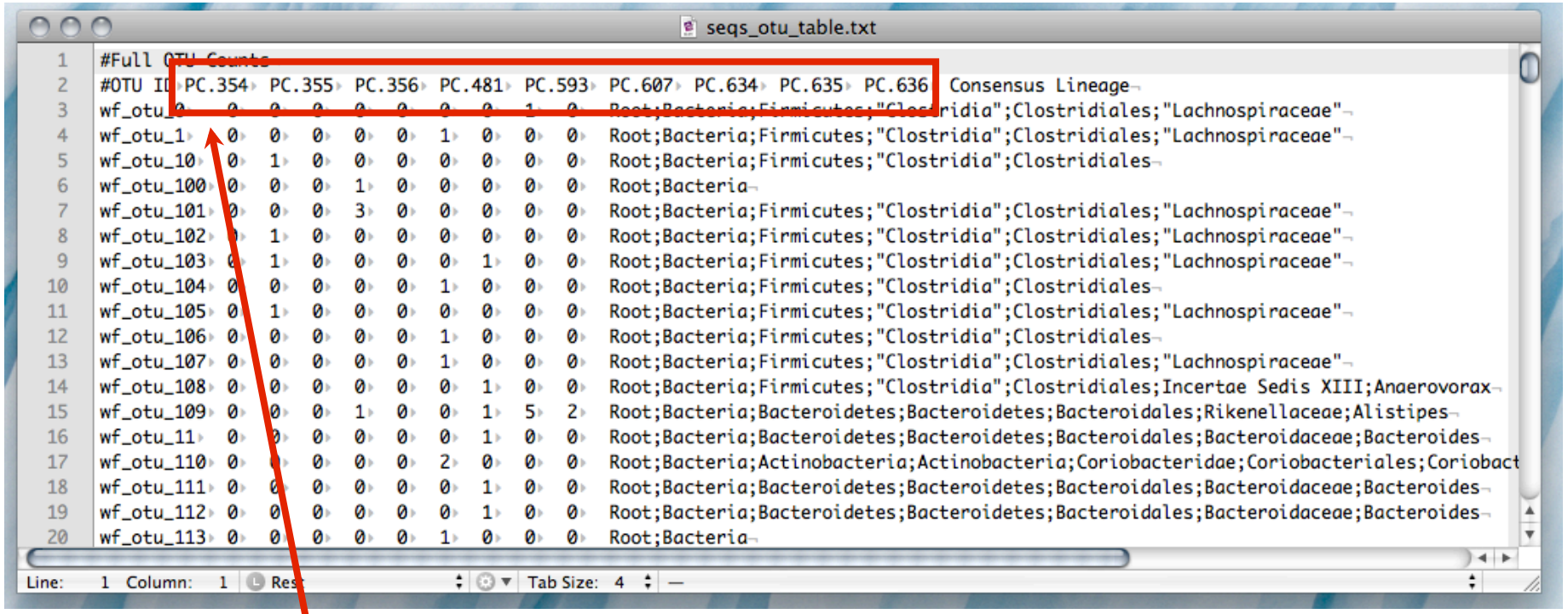


OTU identifiers

OTU table

(classic format)

sample x OTU matrix

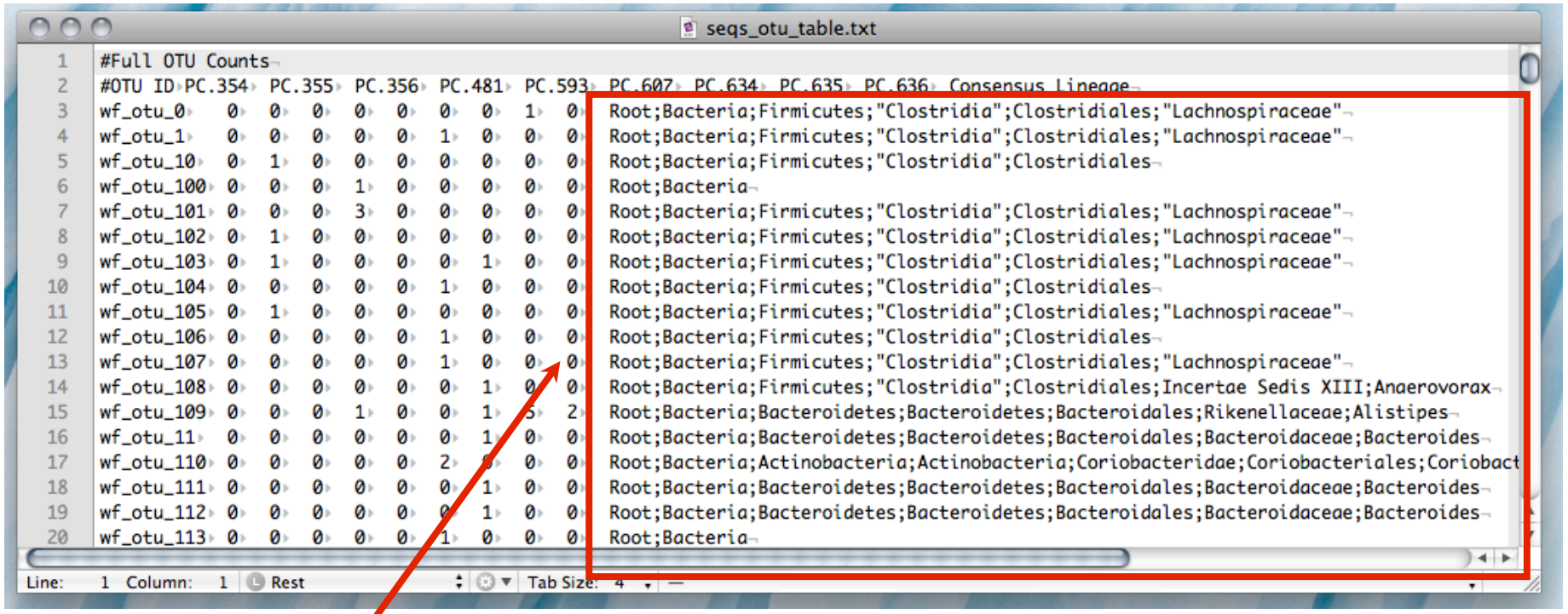


Sample identifiers

OTU table

(classic format)

sample x OTU matrix



Optional per observation taxonomic information

OTU tables are in Biological Observation Matrix (*.biom*) format

http://biom-format.org

Call:

```
biom convert -h
```

for converting between classic and BIOM OTU tables. Detailed usage examples available here: http://biom-format.org/documentation/biom_conversion.html

The Biological Observation Matrix (BIOM) Format or: How I Learned To Stop Worrying and Love the Ome-ome

JSON-based format for representing arbitrary sample x observation contingency tables with optional metadata

MG-RAST
metagenomics analysis server

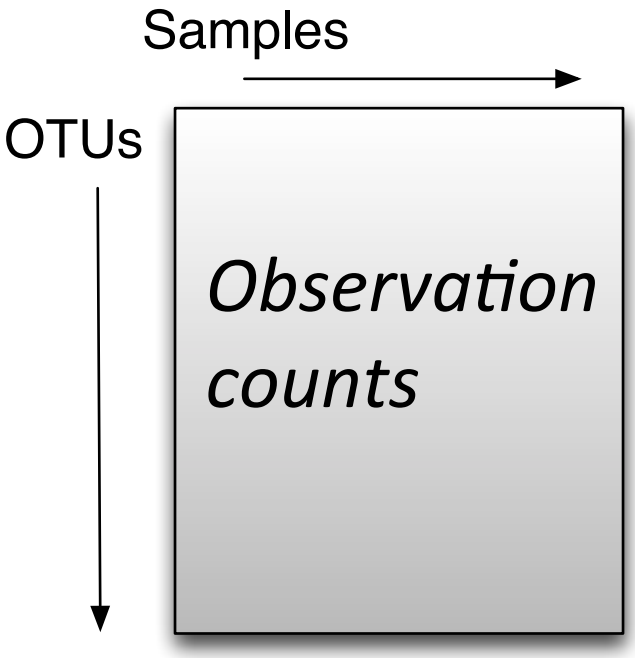
qiime
Quantitative Insights Into Microbial Ecology

VAMPS
The Visualization and Analysis
of Microbial Population Structures

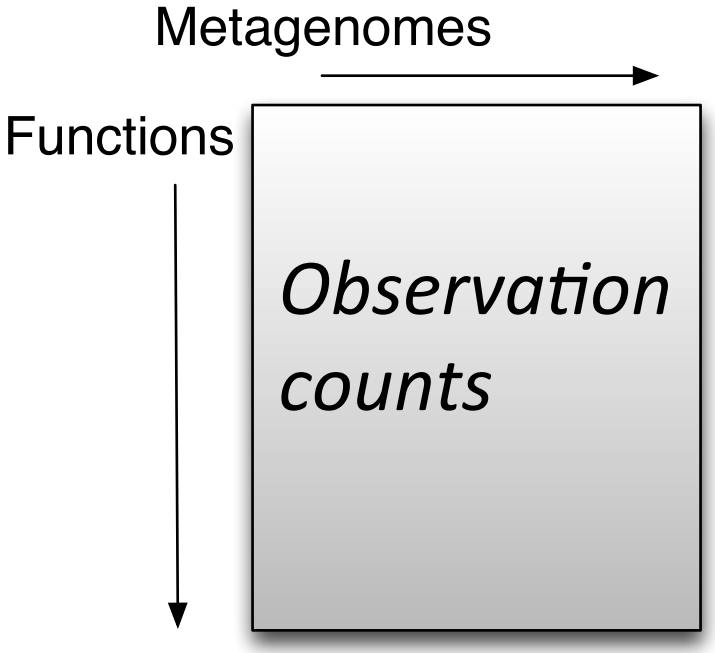
http://www.biom-format.org

McDonald et al., *GigaScience* (2012).

sample x observation contingency matrix

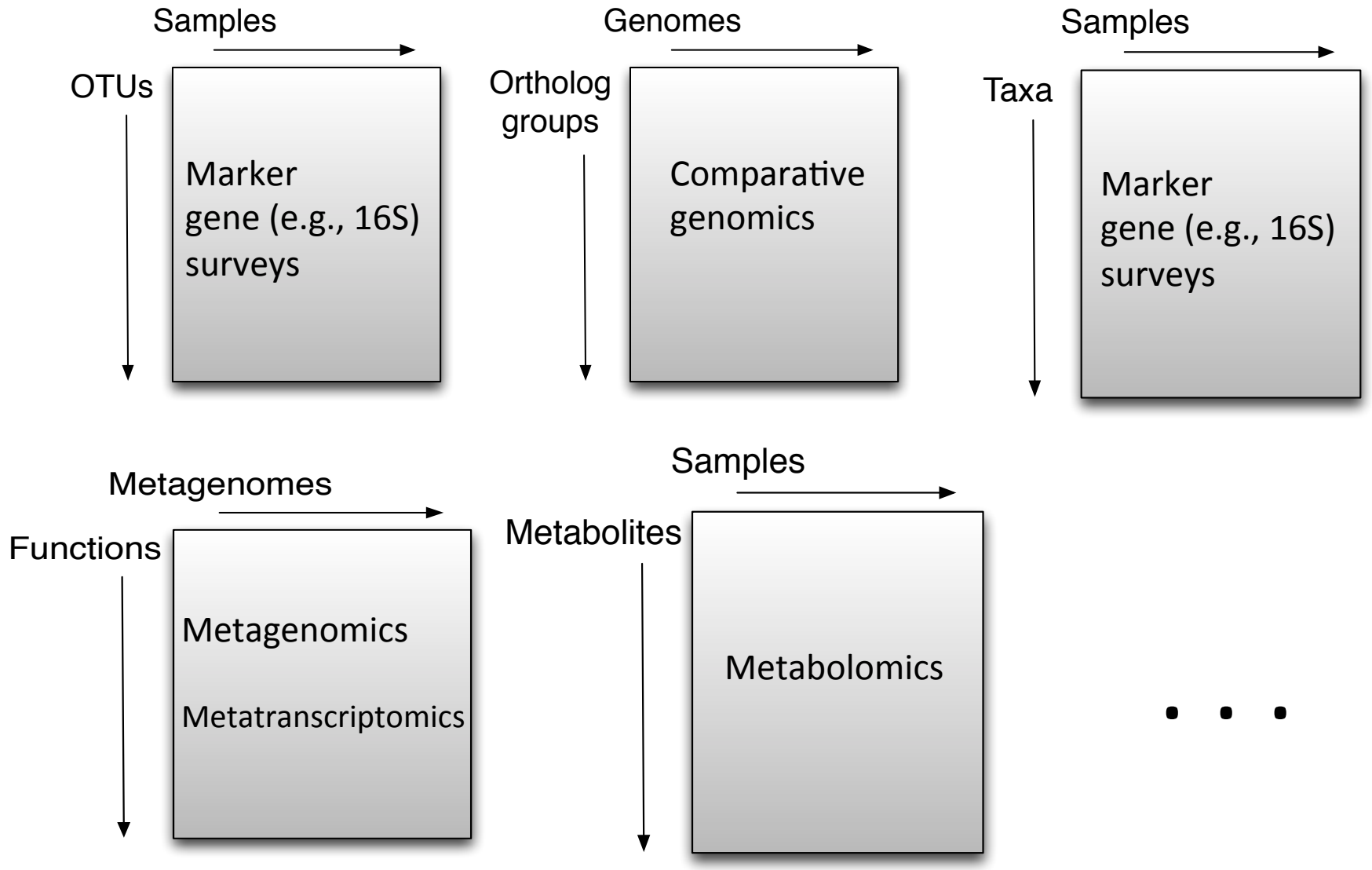


sample x observation contingency matrix



MG-RAST
metagenomics analysis server

sample x observation contingency matrix



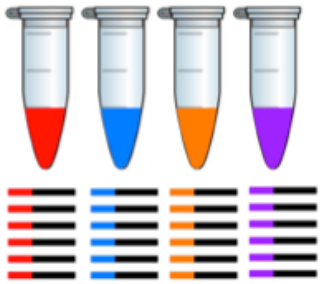
Working with QIIME parameter files (advanced topic)

- QIIME workflow scripts allow you to pass a *parameters file* to override default settings for scripts wrapped in the workflow. See details here:

http://qiime.org/documentation/qiime_parameters_files.html

Demultiplexing (split libraries)

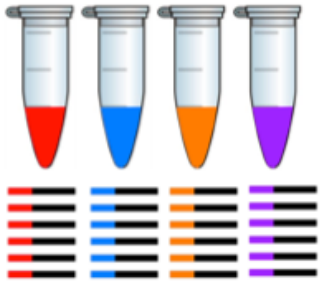
Barcode the rRNA on a per-sample basis.



Micah Hamady, et al., Nature Methods, 2008.
Error-correcting barcodes for pyrosequencing hundreds of samples in multiplex.

Demultiplexing (split libraries)

Barcode the rRNA on a per-sample basis.

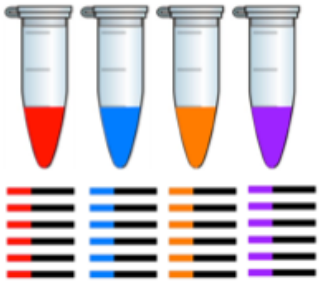


Pool samples and sequence

Micah Hamady, et al., Nature Methods, 2008.
Error-correcting barcodes for pyrosequencing hundreds of samples in multiplex.

Demultiplexing (split libraries)

Barcode the rRNA on a per-sample basis.



Pool samples and sequence

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

Demultiplexing 454 data
split_libraries.py
http://qiime.org/tutorials/tutorial.html
Demultiplexing Illumina data
split_libraries_fastq.py
http://qiime.org/tutorials/illumina_overview_tutorial.html
http://qiime.org/tutorials/processing_illumina_data.html

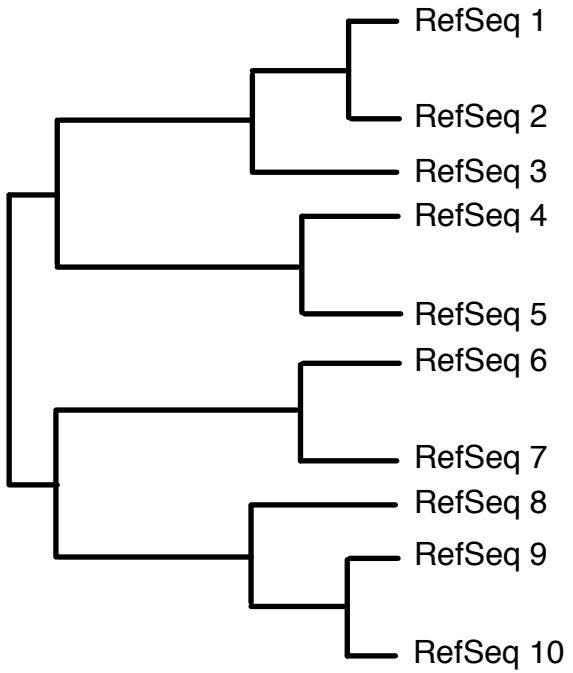
Micah Hamady, et al., Nature Methods, 2008.
Error-correcting barcodes for pyrosequencing hundreds of samples in multiplex.

Which microbial organisms are represented by the rRNA gene sequences in each sample?

rRNA reference database (sequences are available for each 'tip' in the tree)

```
>PC.634_1 FLP3FBN01ELBSX
CTGGGCCGTGTCTCAGTCCCAATGTGGCCGTTTACCCTCTCAGGCCGG
CTACGCATCATCGCCTTGGTGGGCCGTTACCTCACCAACTAGCTAATG
CGCCGCAGGTCCATCCATGTTCACGCCTTGATGGGCGCTTTAATATAC
TGAGCATGCGCTCTGTATACCTATCCGGTTTTAGCTACCGTTTCCAGC
AGTTATCCCGGACACATGGGCTAGG
>PC.634_2 FLP3FBN01EG8AX
TTGGACCGTGTCTCAGTTCCAATGTGGGGGCCTTCCTCTCAGAACCCC
TATCCATCGAAGGCTTGGTGGGCCGTTACCCCGCCAACAACCTAATGG
AACGCATCCCCATCGATGACCGAAGTTCTTTAATAGTTCTACCATGCG
GAAGAACTATGCCATCGGGTATTAATCTTTCTTTCGAAAGGCTATCCC
CGAGTCATCGGCAGGTTGGATACGTGTTACTCACCCGTGCGCCGGT
>PC.354_3 FLP3FBN01EEWKD
TTGGGCCGTGTCTCAGTCCCAATGTGGCCGATCAGTCTCTTAACTCGG
CTATGCATCATTGCCTTGGTAAGCCGTTACCTTACCAACTAGCTAATG
CACCGCAGGTCCATCCAAGAGTGATAGCAGAACCATCTTTCAAACTCT
AGACATGCGTCTAGTGTTGTTATCCGGTATTAGCATCTGTTTCCAGGT
GTTATCCCAGTCTCTTGGG
```

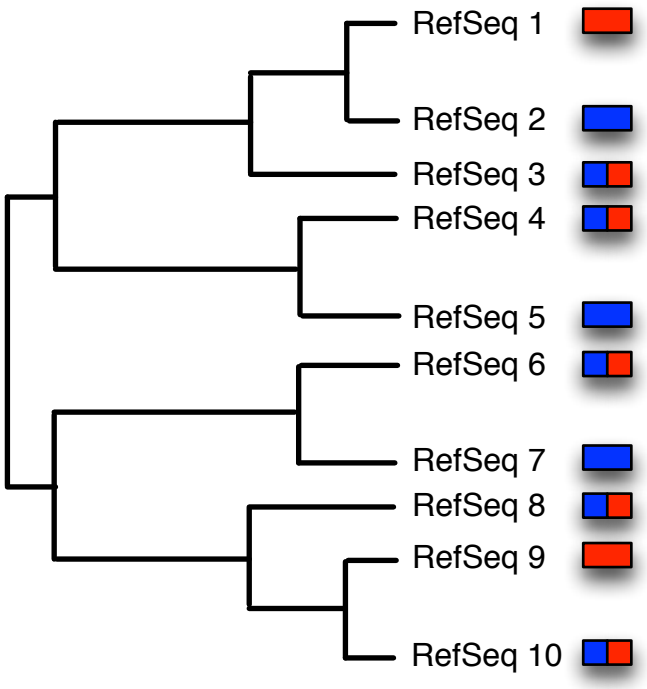
Search against reference sequences



Which microbial organisms are represented by the rRNA gene sequences in each sample?

```
>PC.634_1 FLP3FBN01ELBSX
CTGGGCCGTGTCTCAGTCCCAATGTGGCCGTTTACCCTCTCAGGCCGG
CTACGCATCATCGCCTTGGTGGGCCGTTACCTCACCAACTAGCTAATG
CGCCGCAGGTCCATCCATGTTCACGCCTTGATGGGCGCTTTAATATAC
TGAGCATGCGCTCTGTATACCTATCCGGTTTTAGCTACCGTTTCCAGC
AGTTATCCCGGACACATGGGCTAGG
>PC.634_2 FLP3FBN01EG8AX
TTGGACCGTGTCTCAGTTCCAATGTGGGGGCCTTCCTCTCAGAACCCC
TATCCATCGAAGGCTTGGTGGGCCGTTACCCCGCCAACAACCTAATGG
AACGCATCCCCATCGATGACCGAAGTTCTTTAATAGTTCTACCATGCG
GAAGAACTATGCCATCGGGTATTAATCTTTCTTTCGAAAGGCTATCCC
CGAGTCATCGGCAGGTTGGATACGTGTTACTCACCCGTGCGCCGGT
>PC.354_3 FLP3FBN01EEWKD
TTGGGCCGTGTCTCAGTCCCAATGTGGCCGATCAGTCTCTTAACTCGG
CTATGCATCATTGCCTTGGTAAGCCGTTACCTTACCAACTAGCTAATG
CACCGCAGGTCCATCCAAGAGTGATAGCAGAACCATCTTTCAAACTCT
AGACATGCGTCTAGTGTTGTTATCCGGTATTAGCATCTGTTTCCAGGT
GTTATCCCAGTCTCTTGGG
```

Search against reference sequences



OTU picking

- De Novo
  - Reads are clustered based on similarity to one another.
- Reference-based
  - Closed reference: any reads which don't hit a reference sequence are discarded
  - Open reference: any reads which don't hit a reference sequence are clustered de novo

http://qiime.org/tutorials/otu_picking.html

Subject areas

Bioinformatics
Ecology
Microbiology

Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences

Jai Ram Rideout[1,2], Yan He[3], Jose A. Navas-Molina[4], William A. Walters[5], Luke K. Ursell[6], Sean M. Gibbons[7,10], John Chase[8], Daniel McDonald[4,9], Antonio Gonzalez[9], Adam Robbins-Pianka[4,9], Jose C. Clemente[2], Jack A. Gilbert[10,11], Susan M. Huse[12], Hong-Wei Zhou[3], Rob Knight[9,13], J. Gregory Caporaso[1,8]

August 21, 2014

Note that a PrePrint of this article also exists, first published June 11, 2014.

Author and article information

Abstract

We present a performance-optimized algorithm, subsampled open-reference OTU picking, for assigning marker gene (e.g., 16S rRNA) sequences generated on next-generation sequencing platforms to operational taxonomic units (OTUs) for microbial community analysis. This algorithm provides benefits over de novo OTU picking

https://peerj.com/articles/545/

De novo OTU picking

- Pros
  - All reads are clustered
- Cons
  - Not parallelizable
  - OTUs may be defined by erroneous reads

pick_de_novo_otus.py
http://qiime.org/tutorials/tutorial.html

De novo OTU picking

- You **must** use if:
  - You do not have a reference sequence collection to cluster against, for example because you're working with an infrequently used marker gene.
- You **cannot** use if:
  - You are comparing non-overlapping amplicons, such as the V2 and the V4 regions of the 16S rRNA.
  - You working with very large data sets, like a full HiSeq 2000 run. (Technically you can, but it will be *really* slow.)

pick_de_novo_otus.py
http://qiime.org/tutorials/tutorial.html

Closed-reference OTU picking

- Pros
  - Built-in quality filter
  - Easily parallelizable
  - OTUs are defined by high-quality, trusted sequences
- Cons
  - Reads that don't hit reference dataset are excluded, so you can never observe new OTUs

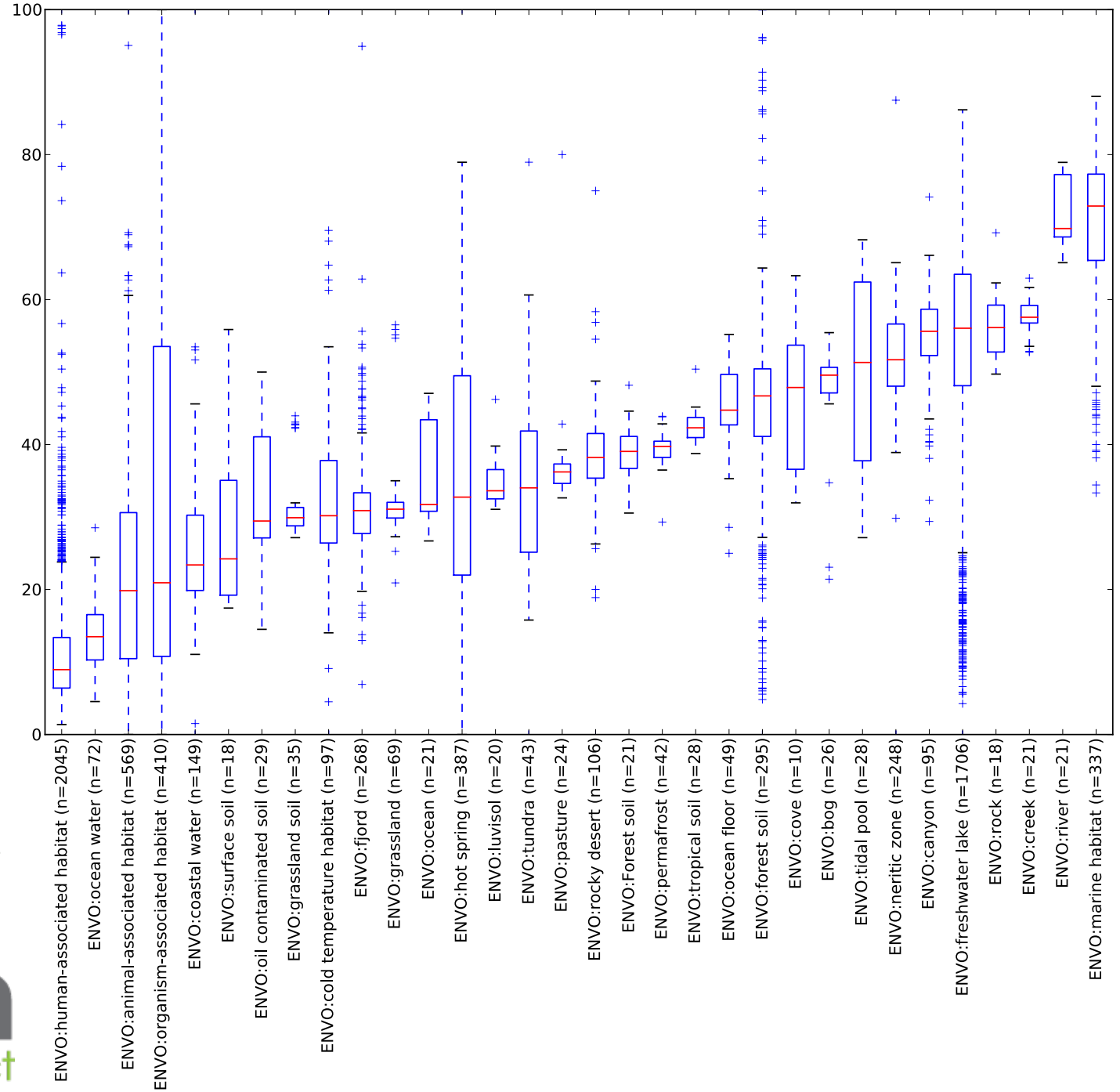
pick_closed_reference_otus.py

Closed-reference OTU picking

- You **must** use if:
  - You are comparing non-overlapping amplicons, such as the V2 and the V4 regions of the 16S rRNA. Your reference sequences must span both of the regions being sequenced.
- You **cannot** use if:
  - You do not have a reference sequence collection to cluster against, for example because you're working with an infrequently used marker gene.

pick_closed_reference_otus.py

Percentage of reads that do not hit the reference collection, by environment type.



earth microbiome project

Open-reference OTU picking

- Pros
  - All reads are clustered
  - Partially parallelizable
- Cons
  - Only *partially* parallelizable
  - Mix of high quality sequences defining OTUs (i.e., the database sequences) and possible low quality sequences defining OTUs (i.e., the sequencing reads)

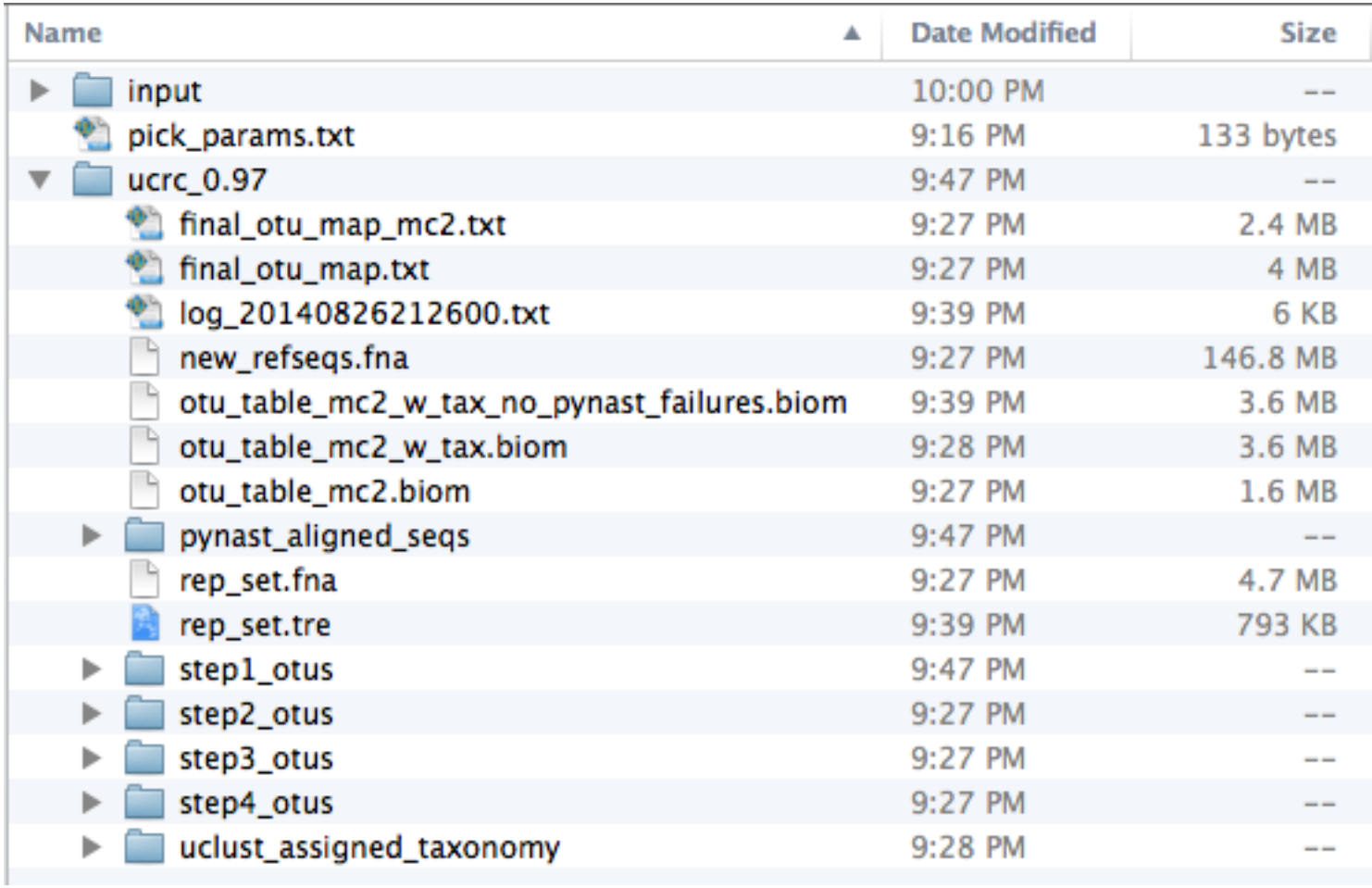
pick_open_reference_otus.py
http://qiime.org/tutorials/illumina_overview_tutorial.html
http://qiime.org/tutorials/open_reference_illumina_processing.html
http://qiime.org/tutorials/fungal_its_analysis.html

Open-reference OTU picking

- You **cannot** use if:
  - You are comparing non-overlapping amplicons, such as the V2 and the V4 regions of the 16S rRNA.
  - You do not have a reference sequence collection to cluster against, for example because you're working with an infrequently used marker gene.

pick_open_reference_otus.py
http://qiime.org/tutorials/illumina_overview_tutorial.html
http://qiime.org/tutorials/open_reference_illumina_processing.html
http://qiime.org/tutorials/fungal_its_analysis.html

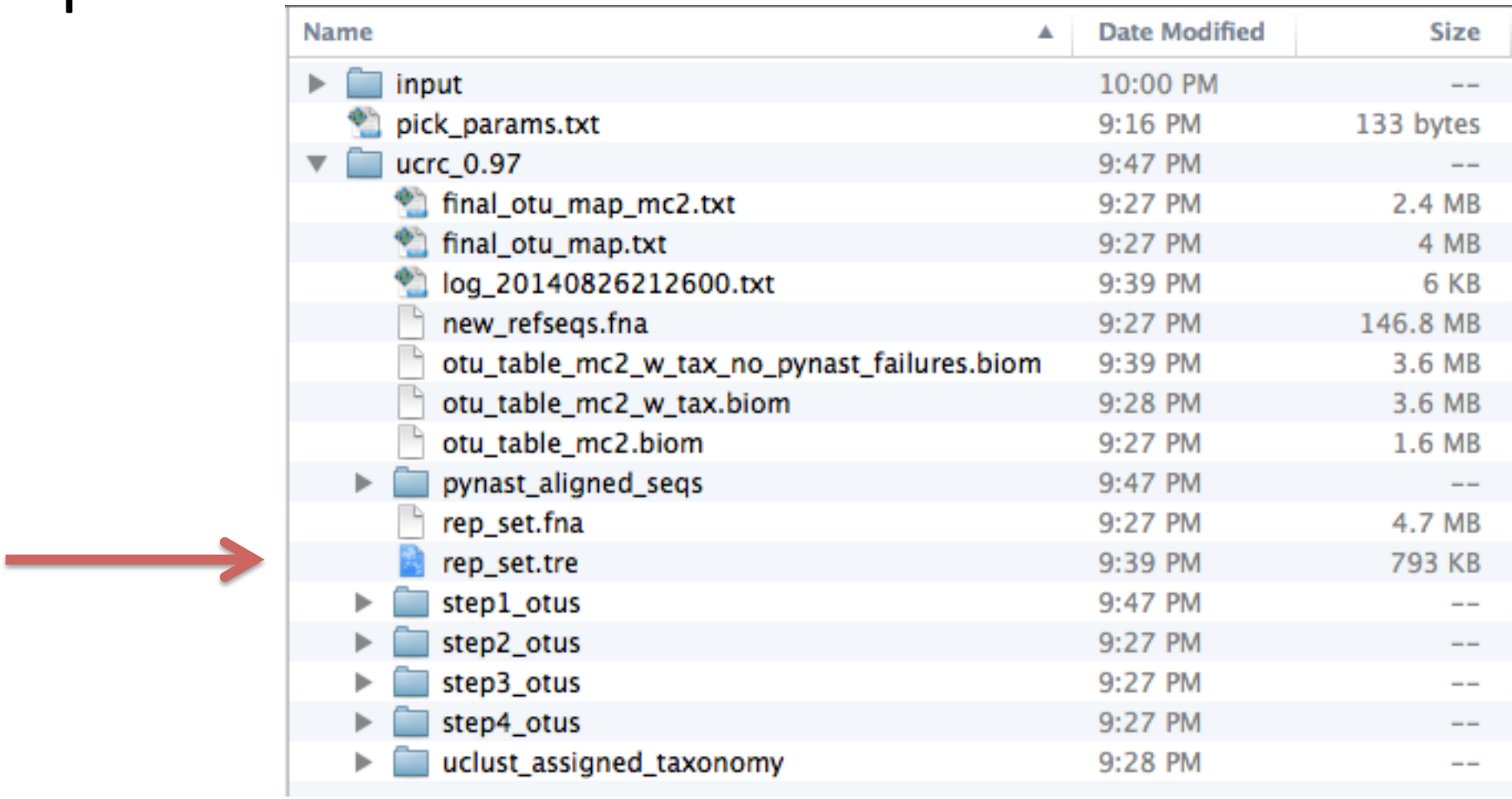
Navigating `pick_open_reference_otus.py` output



Name	Date Modified	Size
input	10:00 PM	--
pick_params.txt	9:16 PM	133 bytes
ucrc_0.97	9:47 PM	--
final_otu_map_mc2.txt	9:27 PM	2.4 MB
final_otu_map.txt	9:27 PM	4 MB
log_20140826212600.txt	9:39 PM	6 KB
new_refseqs.fna	9:27 PM	146.8 MB
otu_table_mc2_w_tax_no_pynast_failures.biom	9:39 PM	3.6 MB
otu_table_mc2_w_tax.biom	9:28 PM	3.6 MB
otu_table_mc2.biom	9:27 PM	1.6 MB
pynast_aligned_seqs	9:47 PM	--
rep_set.fna	9:27 PM	4.7 MB
rep_set.tre	9:39 PM	793 KB
step1_otus	9:47 PM	--
step2_otus	9:27 PM	--
step3_otus	9:27 PM	--
step4_otus	9:27 PM	--
uclust_assigned_taxonomy	9:28 PM	--

Master OTU table for downstream analyses

Navigating `pick_open_reference_otus.py` output



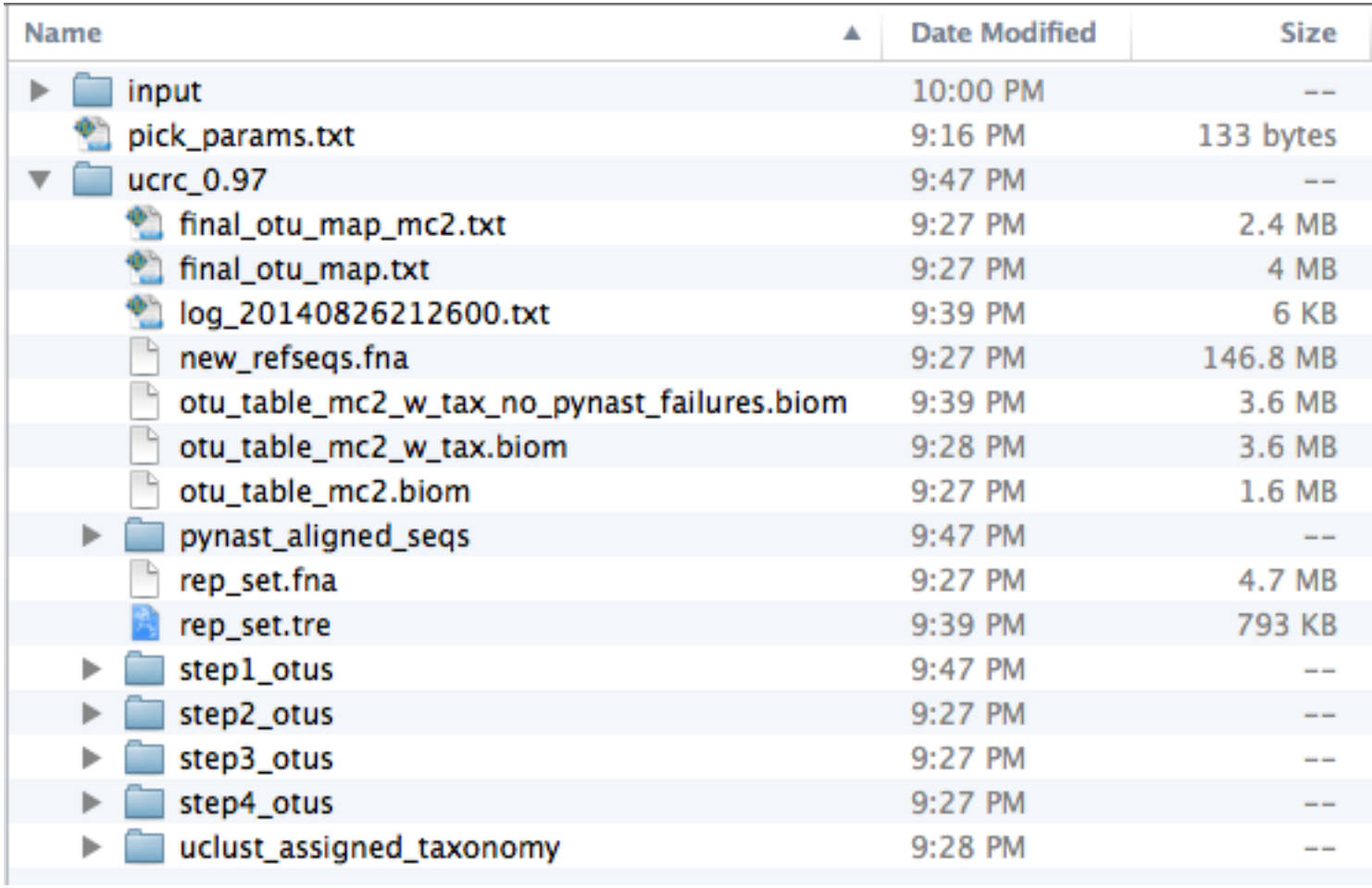
Master phylogenetic tree for downstream analyses

Navigating `pick_open_reference_otus.py` output

Name	Date Modified	Size
input	10:00 PM	--
pick_params.txt	9:16 PM	133 bytes
ucrc_0.97	9:47 PM	--
final_otu_map_mc2.txt	9:27 PM	2.4 MB
final_otu_map.txt	9:27 PM	4 MB
log_20140826212600.txt	9:39 PM	6 KB
new_refseqs.fna	9:27 PM	146.8 MB
otu_table_mc2_w_tax_no_pynast_failures.biom	9:39 PM	3.6 MB
otu_table_mc2_w_tax.biom	9:28 PM	3.6 MB
otu_table_mc2.biom	9:27 PM	1.6 MB
pynast_aligned_seqs	9:47 PM	--
rep_set.fna	9:27 PM	4.7 MB
rep_set.tre	9:39 PM	793 KB
step1_otus	9:47 PM	--
step2_otus	9:27 PM	--
step3_otus	9:27 PM	--
step4_otus	9:27 PM	--
uclust_assigned_taxonomy	9:28 PM	--

Representative sequences for OTUs observed in this study.

Navigating `pick_open_reference_otus.py` output



Representative sequences for OTUs observed in this study *and* and reference database sequences not observed in this data set, for use as a reference in future OTU picking runs.

Read assignment is different for shotgun data, but not that different. In general, the bottleneck is identifying/compiling a reference database.

map_reads_to_reference.py
parallel_map_reads_to_reference.py
http://qiime.org/tutorials/shotgun_analysis.html
http://qiime.org/scripts/map_reads_to_reference.html

QIIME: from laptops to supercomputers

QIIME Workshop Day 1

Slides credit:
Greg Caporaso, www.caporasolab.us
Daniel McDonald
Jose Clemente
Antonio Gonzalez

Jai Rideout

jai.rideout@gmail.com

Installing QIIME

Native installation on OS X or Linux (laptops through 153,408-core compute cluster*)

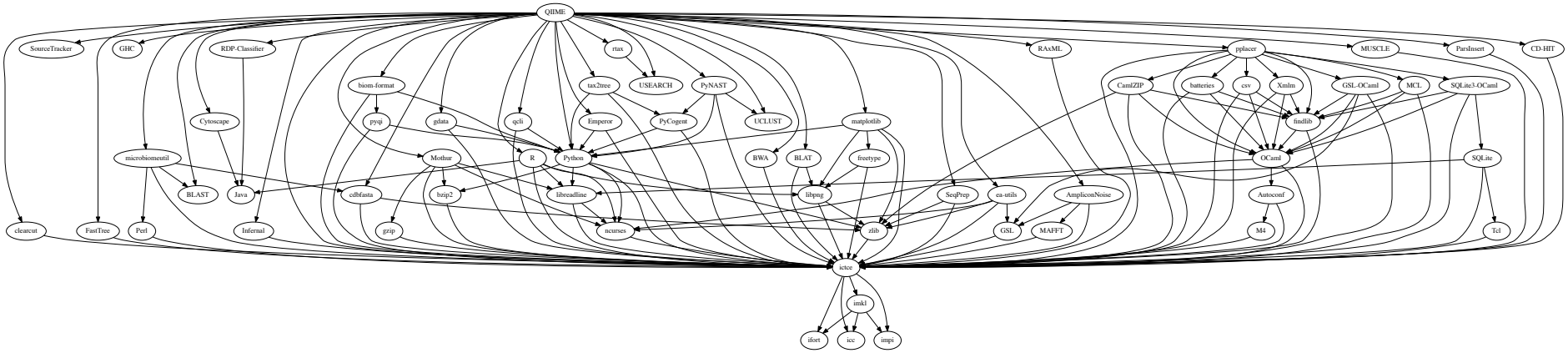
Ubuntu Linux Virtual Box

Cloud-based installations:

- Amazon Web Services (EC2)
- iPlant Atmosphere
- ANL Magellan

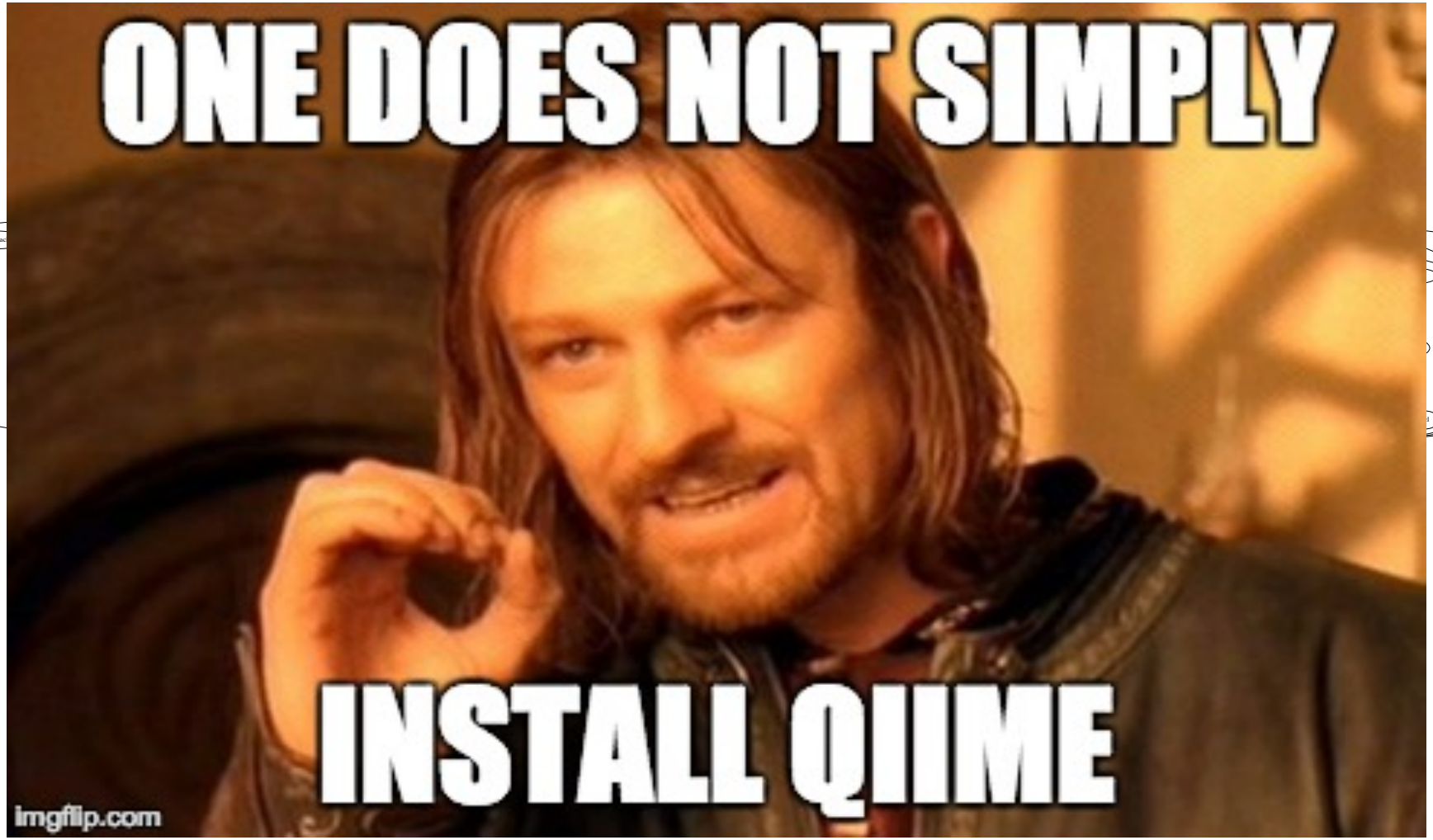
*Hopper (http://top500.org/system/176952)

QIIME software dependencies



http://users.ugent.be/~kehoste/QIIME.pdf

QIIME software dependencies



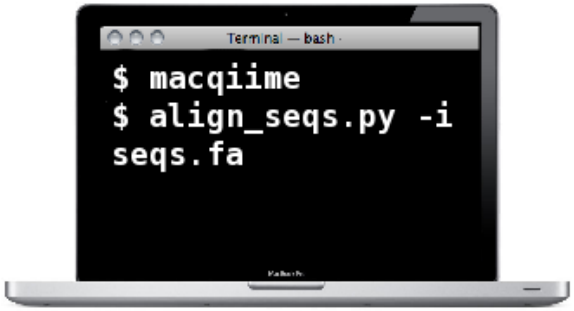
https://imgflip.com/memegenerator/One-Does-Not-Simply
http://users.ugent.be/~kehoste/QIIME.pdf

Native installation

- Installed directly on operating system / computer
- Pros
  - Best performance
  - Better use of resources (e.g., memory)
- Cons
  - Can be difficult to install

Native installation: MacQIIME

- Mac users: easiest way is MacQIIME
  - Excellent tool maintained by Jeff Werner's lab
  - Easiest way to get a (mostly) complete QIIME installation on a Mac
  - Can be difficult to customize/change the installation (e.g., updating specific pieces of MacQIIME)



More info: http://www.wernerlab.org/software/macqiime

Native installation: pip

- Mac/Linux users:

```
pip install qiime
```

- Basic QIIME installation
  - Supports basic upstream analyses and most downstream analyses
  - Must create a QIIME config file
  - Install other dependencies as needed

Native installation: qiime-deploy

- Linux users (especially Ubuntu): easiest way to get a **full** QIIME installation
- Commonly used by system administrators to install QIIME in a cluster environment

More info: https://github.com/qiime/qiime-deploy

Virtual machines

- A “guest” operating system running within a “host” operating system
- A software implementation of a computer, that operates like a physical computer.
- A developer can create a virtual machine *image* which contains their tools pre-installed. Users can then *instantiate* that image to work with those tools.

Browse this page: http://en.wikipedia.org/wiki/Virtual_machine

Benefits that virtual machines offer bioinformatics

- Reproducibility: can publish protocols with a virtual machine instance id.
- Updates are burden of developer, not user.
- Coupled with cloud computing, it's the perfect model for users with sporadic compute needs.

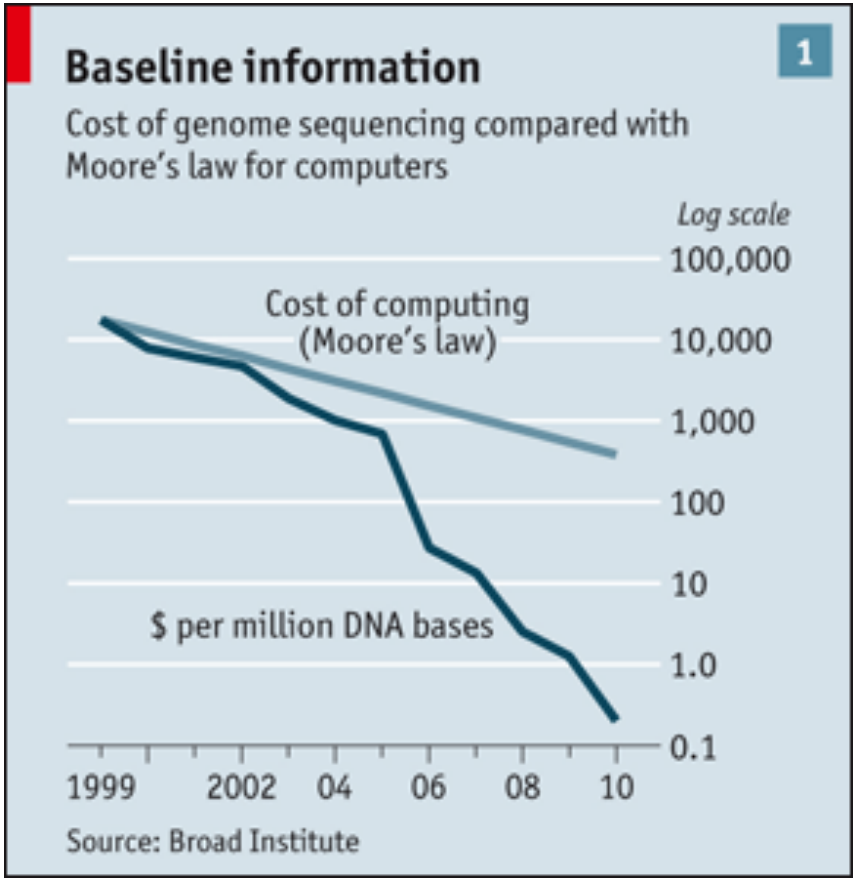
QIIME virtual machines

- VirtualBox (Ubuntu)
  - Easiest way to get a **full** QIIME installation on pretty much **any** platform
    - Especially for Windows users
    - Great for trying out QIIME
  - Slower than a native installation
- Cloud-based installations
  - Amazon Web Services (EC2)
  - iPlant Atmosphere
  - ANL Magellan

Isn’t my laptop powerful enough?

http://thinnestlaptopwallpaper.blogspot.com/2012/10/worlds-smallest-laptop.html

Why is parallel computing important in bioinformatics?



Why is parallel computing important in bioinformatics?

Platform	Sanger	454 (Titanium)	Illumina Genome Analyzer II	Illumina HiSeq2000	Illumina MiSeq
Read Length (bases)	~1000	~400	150 (single end)	100 (single end)	150 (single end); 250 soon
Number of reads	96 or 384	~1,000,000	~100,000,000	~1,600,000,000	~10,000,000
Maximum number of samples per run	n/a	1000	12,000 (barcode-limited)	24,000 (barcode-limited)	2500 (barcode-limited)
Sequences per $1 (sequencing costs only)	0.44	100	5000	200,000	12,500

Cluster computing

- Many computers connected to one another to serve as a larger compute resource.
- Compute-intensive jobs can be split over many systems and run in parallel.
- Similar to desktop compute hardware, but different casing, no (or only few) displays/ keyboards directly connected.
- Owned and maintained “in-house”.

Maintaining hardware is expensive

- Temperature (redundant cooling systems)
- Redundant network connections
- Hardware maintenance (e.g., replacing hard drives)
- Non-water fire suppression
- Backup power
- System administrator ($$)

Cloud computing (IaaS model)

- Implemented on a cluster (or grid), but compute power is rented as a service to support arbitrary applications.

Pay-as-you-go compute power

- Public clouds (e.g., Amazon) rent compute resources
- Log in, boot virtual machine image, run analyses, and terminate instance.
- Cheaper for many tasks than buying, maintaining, and supporting a compute cluster.

Interacting with the Amazon Cloud

- Boot virtual machine image via web interface (or a third-party tool like StarCluster).
- Log in and work via terminal (or via web interface with IPython Notebook)
- Move data back and forth via sftp/scp or a graphical sftp client (e.g., Cyberduck [free/cross-platform])

For information on costs, see http://www.ec2instances.info

Hands-on: Working with AWS

QIIME Workshop Day 1

Jai Rideout

jai.rideout@gmail.com

Slides credit:
Yoshiki Vazquez-Baeza
John Chase

Working with AWS

- We'll learn how to:
  - Log into an AWS EC2 instance (ssh)
  - Transfer files to/from EC2 (scp/sftp)
    - Cyberduck
  - Download files within EC2 (wget)
  - Basic Linux command-line crash course

More details: http://qiime.org/tutorials/working_with_aws.html

Managing EC2 Instances

- Manage EC2 instances via:
  - Amazon's web interface
    - Good for starting up single instances
    - http://aws.amazon.com/
  - StarCluster
    - Good for creating EC2 clusters
    - http://star.mit.edu/cluster/

Managing EC2 Instances

- EC2 charges hourly rate for **running** instances
- **Stop** an instance when you're not using it
  - Small storage fee for stopped instances
- **Terminate** an instance when you're permanently done with it
  - **Warning:** ALL data and configuration will be permanently lost unless you saved it on an EBS volume!
  - Make sure to copy your files somewhere before terminating!

ssh

- Use ssh to log into a remote computer
  - Available on Mac and Linux
  - Windows: use Mobaxterm
- For help with ssh, run `man ssh`

Log into EC2

- ssh –i <key> ubuntu@<ip-address>
- Create a new directory of the form <first name>_<last name>
  - Example: mkdir jai_rideout
  - **This is your personal directory;** we'll use it in the following exercises
- Move into the new directory (cd)

Download files using wget

- wget: download files from the command-line
- Convenient for publicly-hosted files
- Download the following file:

```
wget http://bit.ly/1uhqL2D -O example.txt
```

scp/sftp

- Use scp or sftp to copy files/directories to/from a remote computer
  - Available on Mac and Linux
- Cyberduck: a graphical sftp program
  - Available on Mac and Windows
- WinSCP: another graphical program for Windows

Cyberduck

- Download and install Cyberduck
- I'll show you how to connect to EC2

More details:
http://qiime.org/tutorials/working_with_aws.html

Copy files from EC2

- Copy the example file from EC2 with Cyberduck
  - Double-click example.txt
    - Will be put in your Downloads folder
  - OR: Click and drag the file where you want it saved
- Now open the file, make some changes, and save it

Copy files to EC2

- Copy the example file to EC2 with Cyberduck
  - Navigate to your personal directory in Cyberduck
  - Drag and drop the file to upload it
- In your EC2 terminal, verify that the file exists (ls)
- View the new contents of the file (cat)

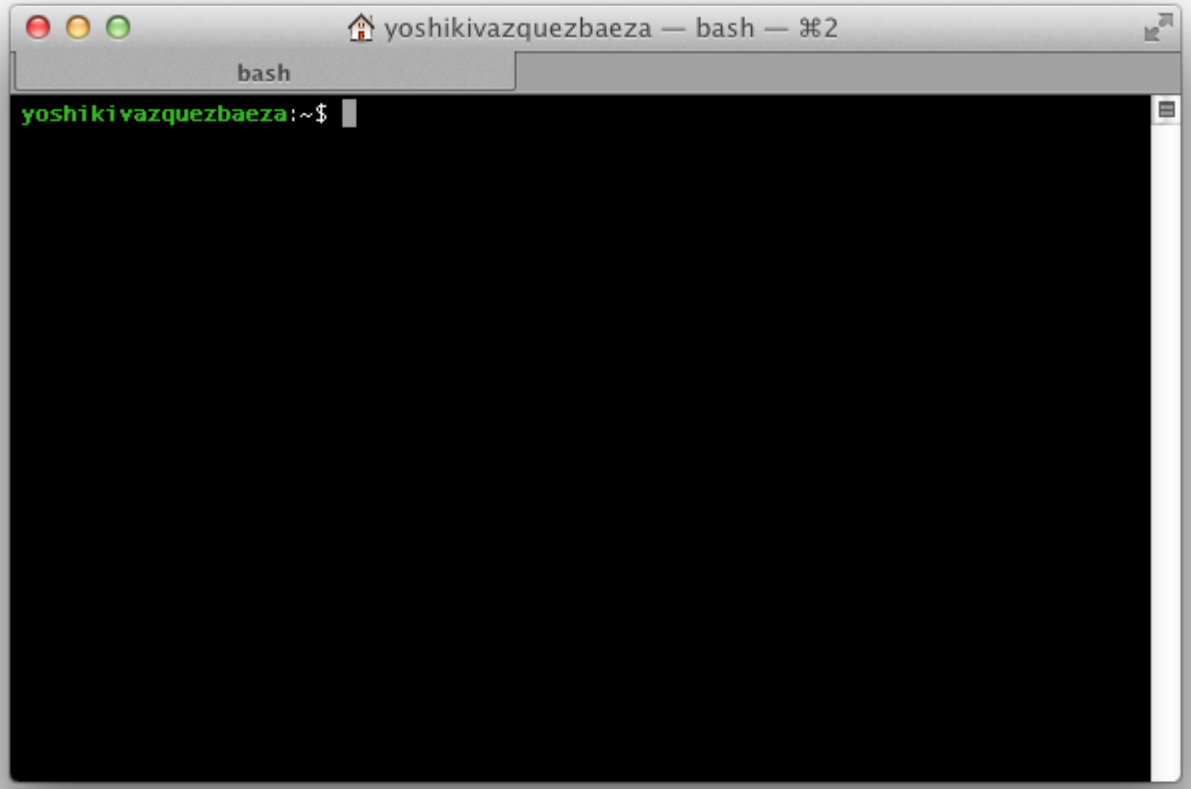
Command line crash course

Excellent command line tutorial to work through on your own time:

http://software-carpentry.org/v5/novice/shell/index.html

Talking to UNIX

- Through a terminal emulator

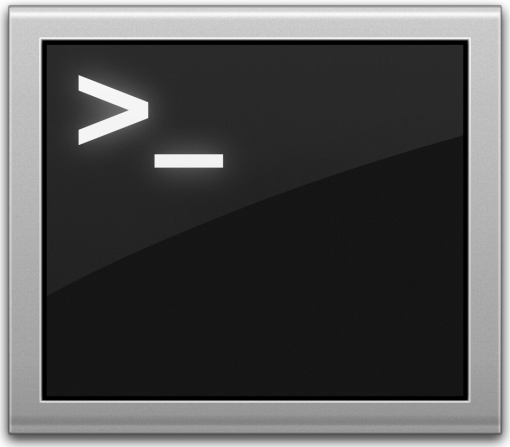


more specifically using a shell ...

Launching a terminal window

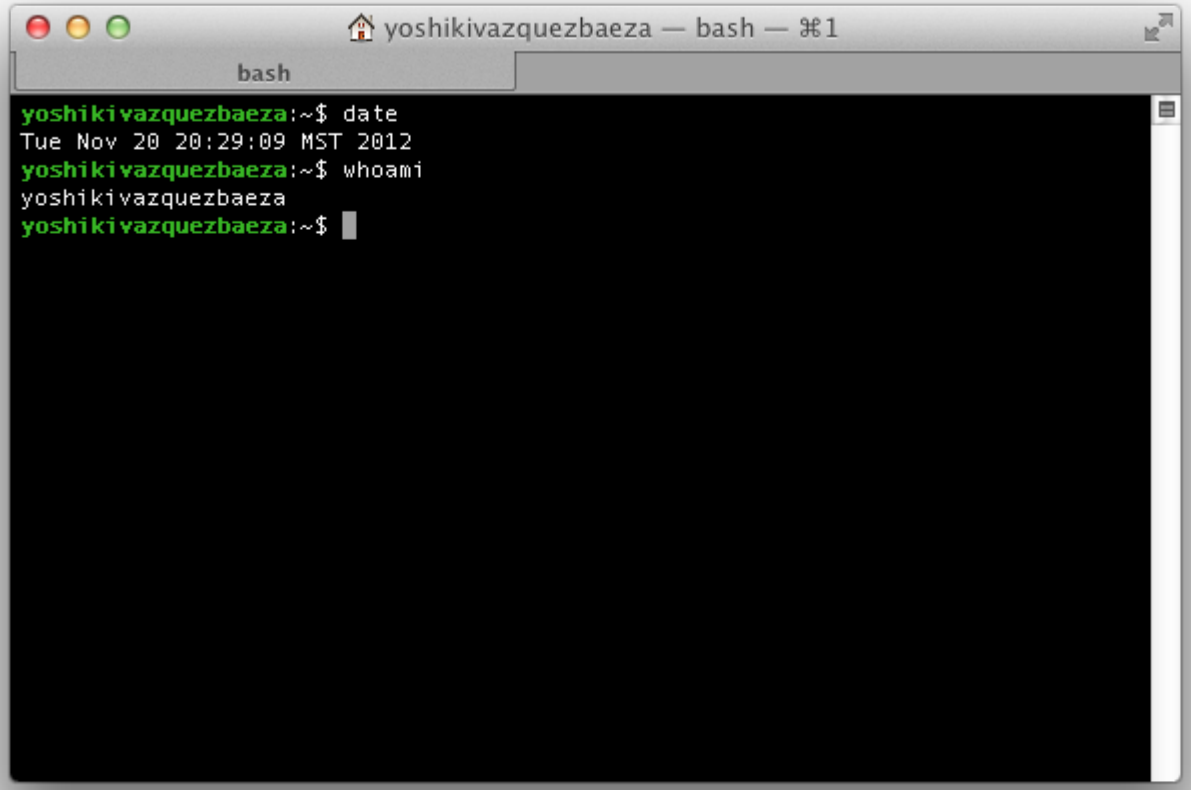
- In Mac OS X in a Finder window or the Desktop:
  - command + shift + u
  - Search for the “Terminal”
- Ubuntu

  In the sidebar search for the terminal icon



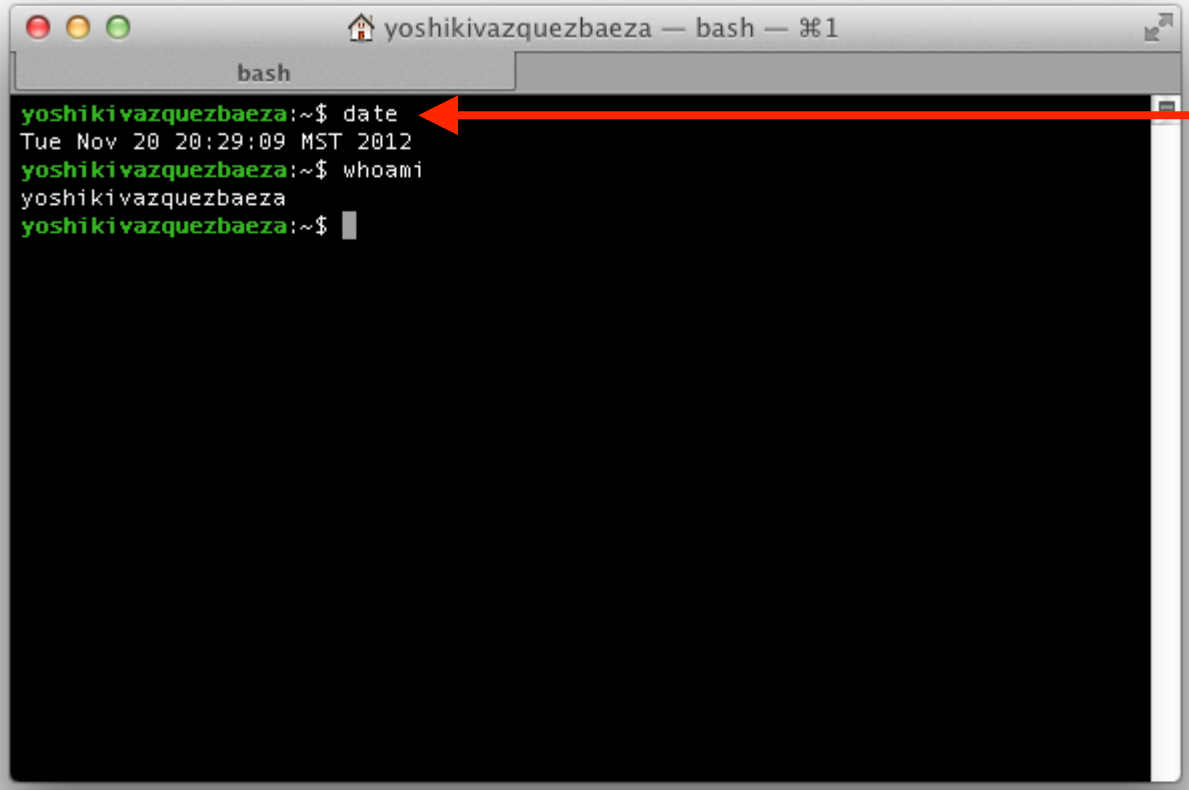
How do you talk to UNIX?

- Using some commands



How do you talk to UNIX?

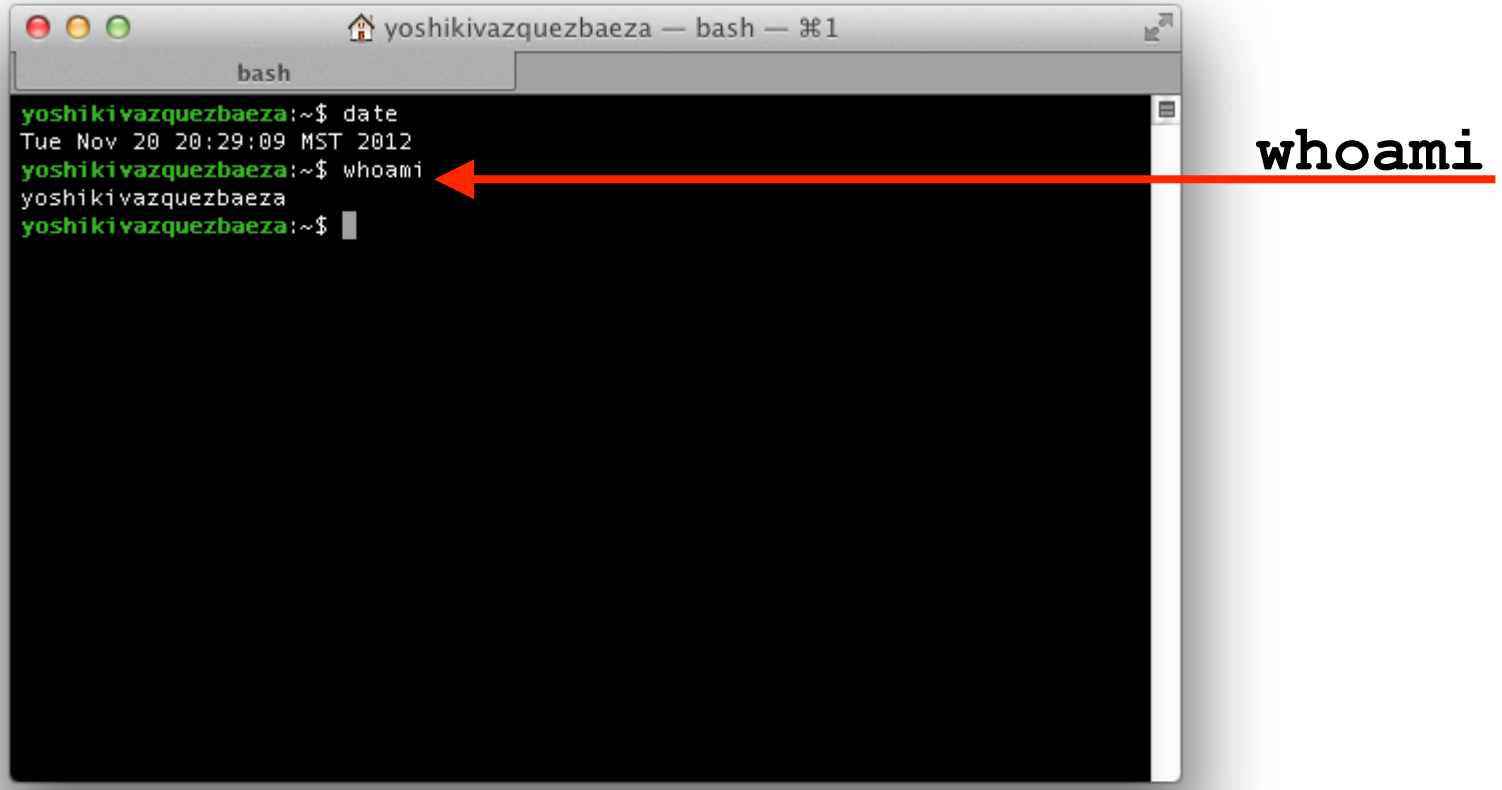
- Get the current date and hour



open a terminal window and try them ...

How do you talk to UNIX?

- See what is your user name



Some general concepts

Home Sweet Home

Can be referred to as:

`$HOME` or **`${HOME}`**

Or also as:

`~`

Mac OS X:

`/Users/username/`

Linux based systems:

`/home/username/`

`username` stands for your user in your machine

Try:

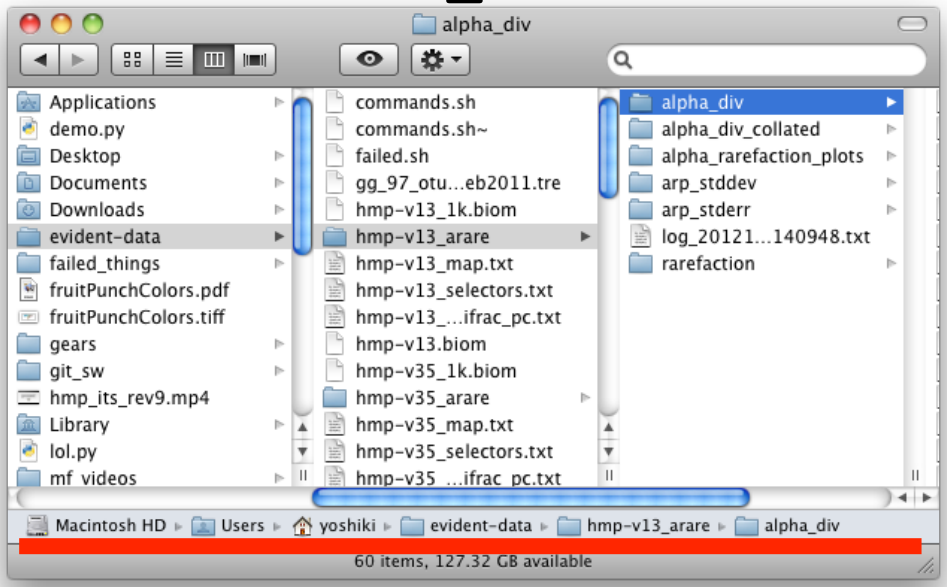
```
echo ${HOME}
ls $HOME
ls ~/
```

Paths (absolute)

/Users/yoshiki/evident-data/hmp-v13_arare/alpha_div

$HOME/evident-data/hmp-v13_arare/alpha_div

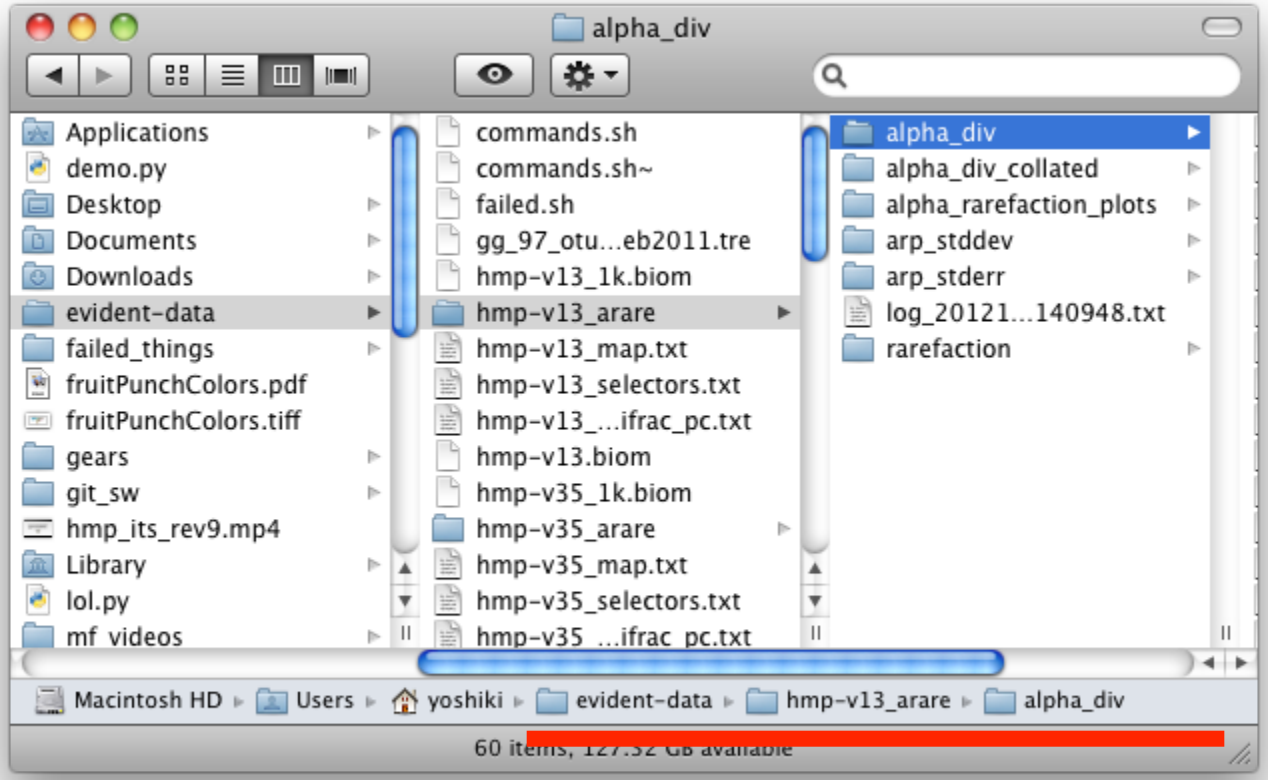
~/evident-data/hmp-v13/alpha_div



A **slash** at the beginning of a path denotes it as an absolute path, i. e. from the base of your hard drive.

Paths (relative)

`evident-data/hmp-v13_arare/alpha_div`



On the other side relative paths are not preceeded by a slash

One command, different forms

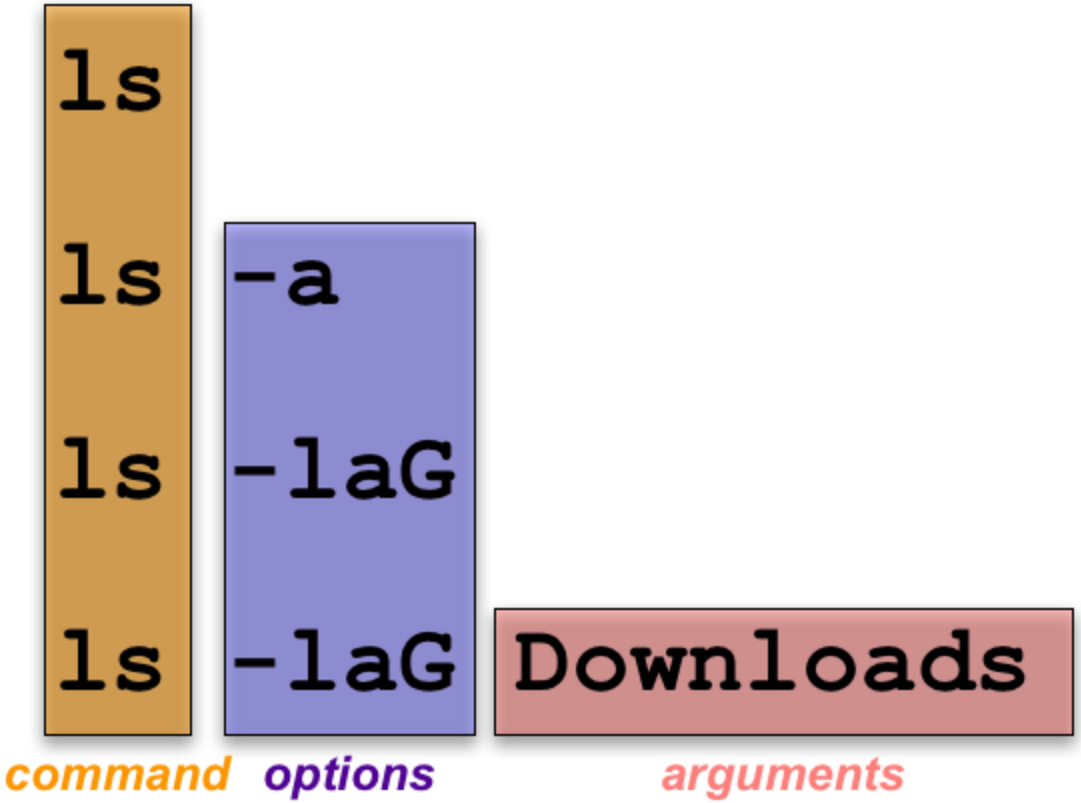
```
ls

ls -a

ls -laG

ls -laG Downloads
```

Anatomy of a command



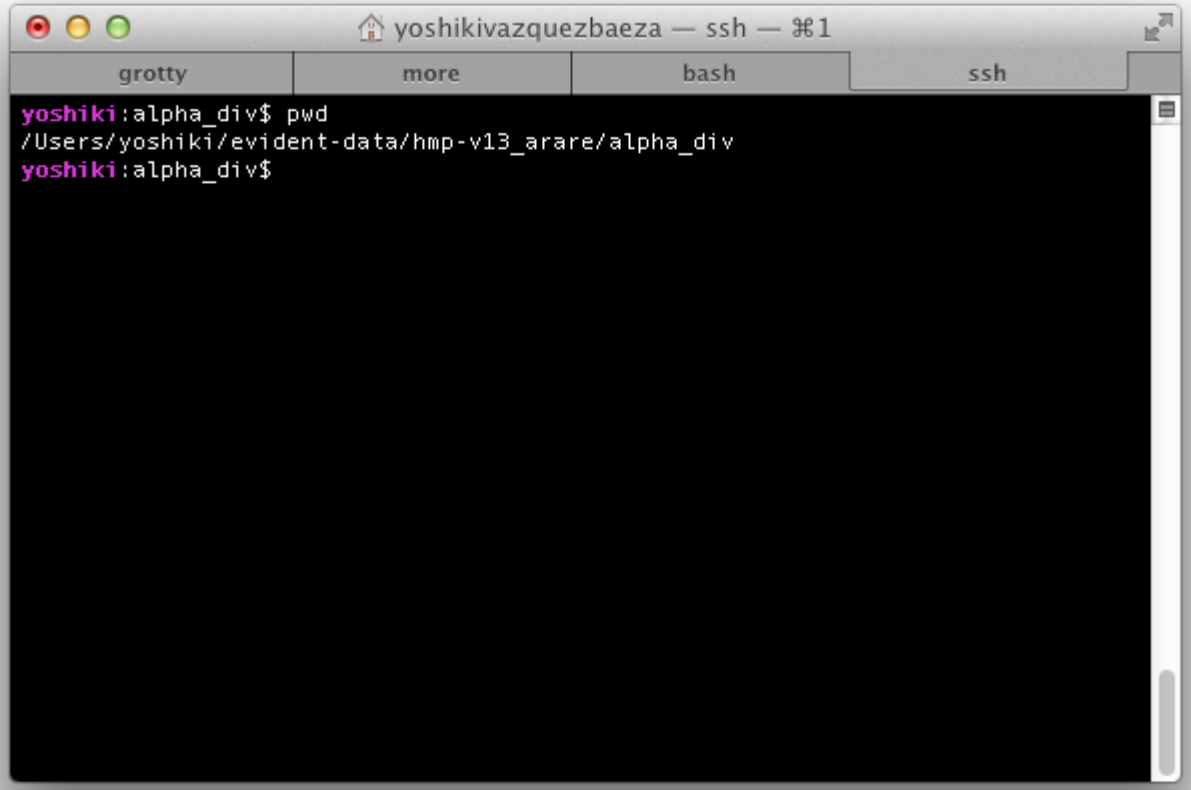
Some useful commands

... for life

You are here

Get your current working directory:

`pwd`



Folders, files and its information

List files from your current working directory:

```
ls
```

List ***all*** the files in your current directory, including hidden files:

```
ls -a
```

List files in your ***Downloads*** folder, in the ***long*** format and sort them by ***time***:

```
ls -lt ~/Downloads
```

Navigating your machine

Change from your current ***directory*** to your **home**

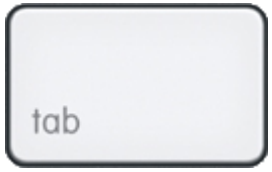
```
cd
```

Change from your current ***directory*** to a directory below it:

```
cd ..
```

Change from your current ***directory*** to your ***documents*** folder:

```
cd ~/Documents
```



to autocomplete

Making, copying and moving stuff

Make a new ***directory***

```
__mkdir AnExample
```

Move a file/folder or change its name

```
__mv oldname.txt newname.txt
  mv Files/ NewName/
```

Copy a file

```
__cp homework.txt backup_homework.txt
```

Copy a directory

```
__cp -r NewName Files
```

Making, copying and moving stuff

Make a new ***directory***

```
__mkdir AnExample
```

Move a file/folder or change its name

```
__mv oldname.txt newname.txt
  mv Documents/ Newname/
```

All these commands work with a old (source) -> new (destiny) scheme

Copy a file

```
__cp homework.txt backup_homework.txt
```

Copy a directory

```
__cp -r physics1 physics2
```

Star

It's a wildcard.

List anything that ends with a .txt

```
ls *.txt
```

List anything with the letter t

```
ls *t*
```

Copy to your desktop all text files

```
cp *.txt ~/Desktop/
```

Removing files

Remember to be careful, be very careful, there is **no undo** for this command.

Remove a file

```
__rm some_file.txt
```

Remove a folder with things inside it

```
__rm -r UselessFolder/
```

Force, the ***removal*** of a folder

```
__rm -rf UselessFolder/
```

Compression and decompression

Using ***zip*** to compress

```
__zip compressed.zip bigfile.txt
  zip -r compressedFolder.zip BigFolder
```

Using ***zip*** to decompress *things*

```
__unzip compressed.zip
```

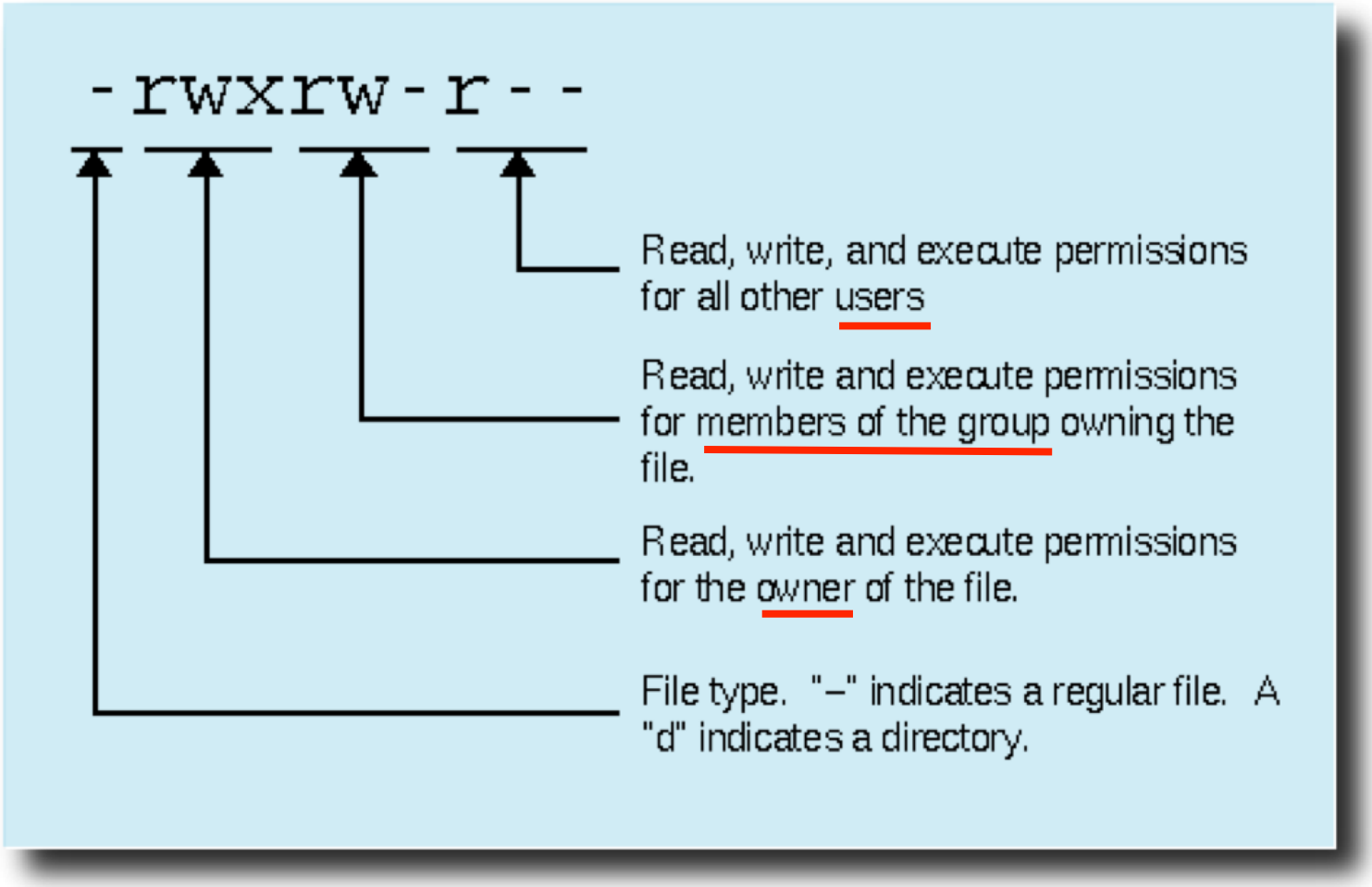
Using ***tar*** to compress a folder or file

```
__tar -czf output.tgz BigFolder
```

Using ***tar*** to decompress *things*

```
__tar -xzf output.tgz
```

Permissions



Permissions

Allow ***all*** to have ***write*** permissions to a file

```
__chmod a+w file.txt
```

Allow ***all*** to have ***write*** permissions to a folder and its contents:

```
__chmod -R a+w file.txt
```

Remove ***all*** the permission to ***write*** to a file

```
__chmod a-w file.txt
```

Remove ***all*** the permission to ***write*** to a folder and its contents:

```
__chmod -R a-w file.txt
```

To see how permissions change, use:

Inspecting a text file

Print a file in the screen

```
cat file.txt
```

Inspect the contents

```
less file.txt
more file.txt
```

to exit these commands type q

Count the ***words*** of a file

```
wc file.txt
```

Count the ***lines*** of a file

```
wc -l file.txt
```

more is less :P

Inspecting a parts of a file

Seeing the first few lines of a file

```
head file.txt
```

Seeing the ***N*** lines of a file

```
head -n 20 file.txt
```

Seeing the last few lines of a file

```
tail file.txt
```

Seeing the ***N*** lines of a file

```
tail -n 20 file.txt
```

Searching the contents of a file

Searching for text in a file:

```
grep "yet" file.txt
```

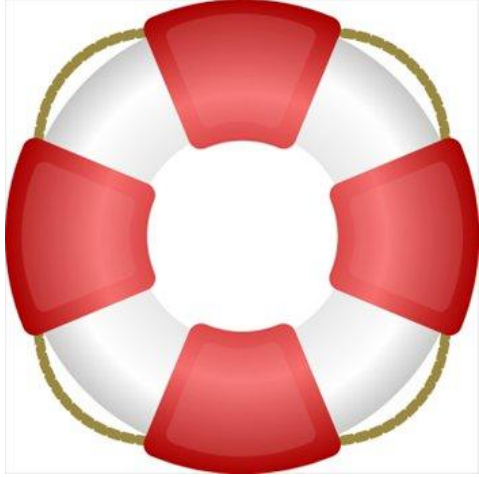
Searching for text in a file (and show 2 lines before "-B" or 2 lines after "-A" value):

```
grep -A 2 "yet" file.txt
grep -B 2 "yet" file.txt
```

Searching for text and highlight the matches:

```
grep --color "yet" file.txt
```

Getting help



Each command has its own way i. e.

As an argument:

```
zip -h
tar --help
```

Using the ***manual*** reference command:

```
man ls
man grep
```

to exit the manual just type **q**

Binaries, scripts, programs etc ...

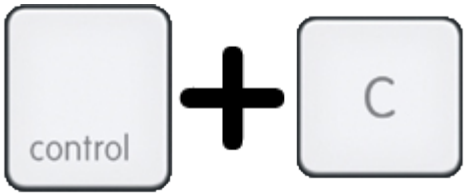
- Try the following:
  `which ls`
- PATH has a lot of information, *it's the "route"*
  `echo $PATH`
- To execute something from your current working directory

  ensure it has the right permissions, then:

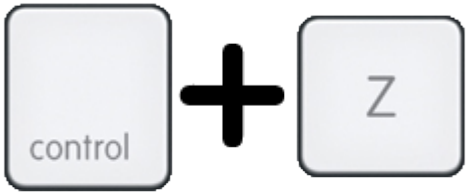
  `./program_test`

if it doesn't have permissions try **`chmod a+x program_test`**

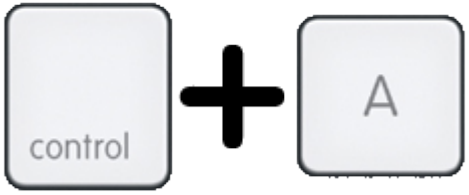
Special shortcuts



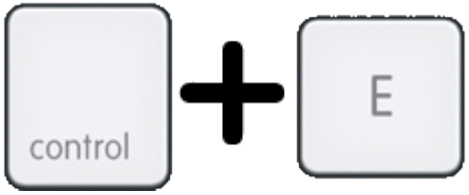
halt and kill a command



stop a command (doesn't kill it)



go to the **beginning** of the line in a terminal window



go to the **end** of the line in a terminal window

This applies for any operating system.

Microbial Community Diversity (Illumina Tutorial)

QIIME Workshop Day 1

Slides credit:
Greg Caporaso, www.caporasolab.us
Daniel McDonald
Jose Clemente

Jai Rideout
jai.rideout@gmail.com

Getting started

- Log into your EC2 instance
- cd into your personal folder
  - Example: `cd jai_rideout`
- Open http://bit.ly/1nSSvdz in a web browser
  - Contains the commands we'll run during the tutorial

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).

Moving Pictures of the Human Microbiome

- Investigate the relative temporal variability of body sites.
- Is there a temporal core microbiome?
- Technical points: do we observe the same conclusions on 454 and Illumina data?

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.
- The online tutorial contains details on all of the steps: go back and read that text.

Evening Lecture: Exploring diversity

Comparing microbial communities

Who is there?

How many “species” are there?

How similar are pairs of samples?

core_diversity_analyses.py

Site index

- Home
- Install
- Documentation
- Tutorials
- Blog
- Developer

Quick search

Go

Enter search terms or a module, class or function name.

core_diversity_analyses.py – A workflow for running a core set of QIIME diversity analyses.

Description:

This script plugs several QIIME diversity analyses together to form a basic workflow beginning with a BIOM table, mapping file, and optional phylogenetic tree.

The included scripts are those run by the workflow scripts alpha_rarefaction.py, beta_diversity_through_plots.py, summarize_taxa_through_plots.py, plus the (non-workflow) scripts make_distance_boxplots.py, compare_alpha_diversity.py, and otu_category_significance.py. To update parameters to the workflow scripts, you should pass the same parameters file that you would pass if calling the workflow script directly.

Usage: `core_diversity_analyses.py [options]`

Input Arguments:

```
[REQUIRED]

-i, --input_biom_fp
    The input biom file [REQUIRED]

-o, --output_dir
    The output directory [REQUIRED]

-m, --mapping_fp
    The mapping filepath [REQUIRED]

-e, --sampling_depth
    Sequencing depth to use for even sub-sampling and maximum rarefaction depth. You should review the output of print_biom_table_summary.py to decide on this value.

[OPTIONAL]

-p, --parameter_fp
    Path to the parameter file, which specifies changes to the default behavior. For more information, see www.qiime.org/documentation/qiime_parameters_files.html [if
    omitted, default values will be used]

-a, --parallel
    Run in parallel where available. Specify number of jobs to start with -O or in the parameters file. [default: False]

--nonphylogenetic_diversity
    Apply non-phylogenetic alpha (chao1 and observed_species) and beta (bray_curtis) diversity calculations. This is useful if, for example, you are working with non-
    amplicon BIOM tables, or if a reliable tree is not available (e.g., if you're working with ITS amplicons) [default: False]

--suppress_taxa_summary
    Suppress generation of taxa summary plots. [default: False]
```

core_diversity_analyses.py

```
$ core_diversity_analyses.py –i otu_table.biom
                             –o core_output
                             –m map.txt
                             –c SampleType,day
                             –t rep_set.tre
                             –e 20
```

core_diversity_analyses.py

```
$ core_diversity_analyses.py –i otu_table.biom
                              –o core_output
                              –m map.txt
                              –c SampleType,day
                              –t rep_set.tre
                              –e 20
```

core_diversity_analyses.py

```
$ core_diversity_analyses.py –i otu_table.biom
                             –o core_output
                             –m map.txt
                             –c SampleType,day
                             –t rep_set.tre
                             –e 20
```

core_diversity_analyses.py

```
$ core_diversity_analyses.py –i otu_table.biom
                              –o core_output
                              –m map.txt
                              –c SampleType,day
                              –t rep_set.tre
                              –e 20
```

core_diversity_analyses.py

```
$ core_diversity_analyses.py –i otu_table.biom
                              –o core_output
                              –m map.txt
                              –c SampleType,day
                              –t rep_set.tre
                              –e 20
```

core_diversity_analyses.py

```
$ core_diversity_analyses.py –i otu_table.biom
                              –o core_output
                              –m map.txt
                              –c SampleType,day
                              –t rep_set.tre
                              –e 20
```

core_diversity_analyses.py

```
$ core_diversity_analyses.py –i otu_table.biom
                              –o core_output
                              –m map.txt
                              –c SampleType,day
                              –t rep_set.tre
                              –e 20
```

Alpha (within sample) diversity

Alpha diversity

Sample A

Pseudomonas aeruginosa
Pseudomonas argentinensis
Pseudomonas flavescens

Sample B

Pseudomonas aeruginosa
Pseudomonas argentinensis
Escherichia coli

Sample C

Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii

Alpha diversity

Sample A

Pseudomonas aeruginosa
Pseudomonas argentinensis
Pseudomonas flavescens

Sample B

Pseudomonas aeruginosa
Pseudomonas argentinensis
Escherichia coli

Sample C

Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii

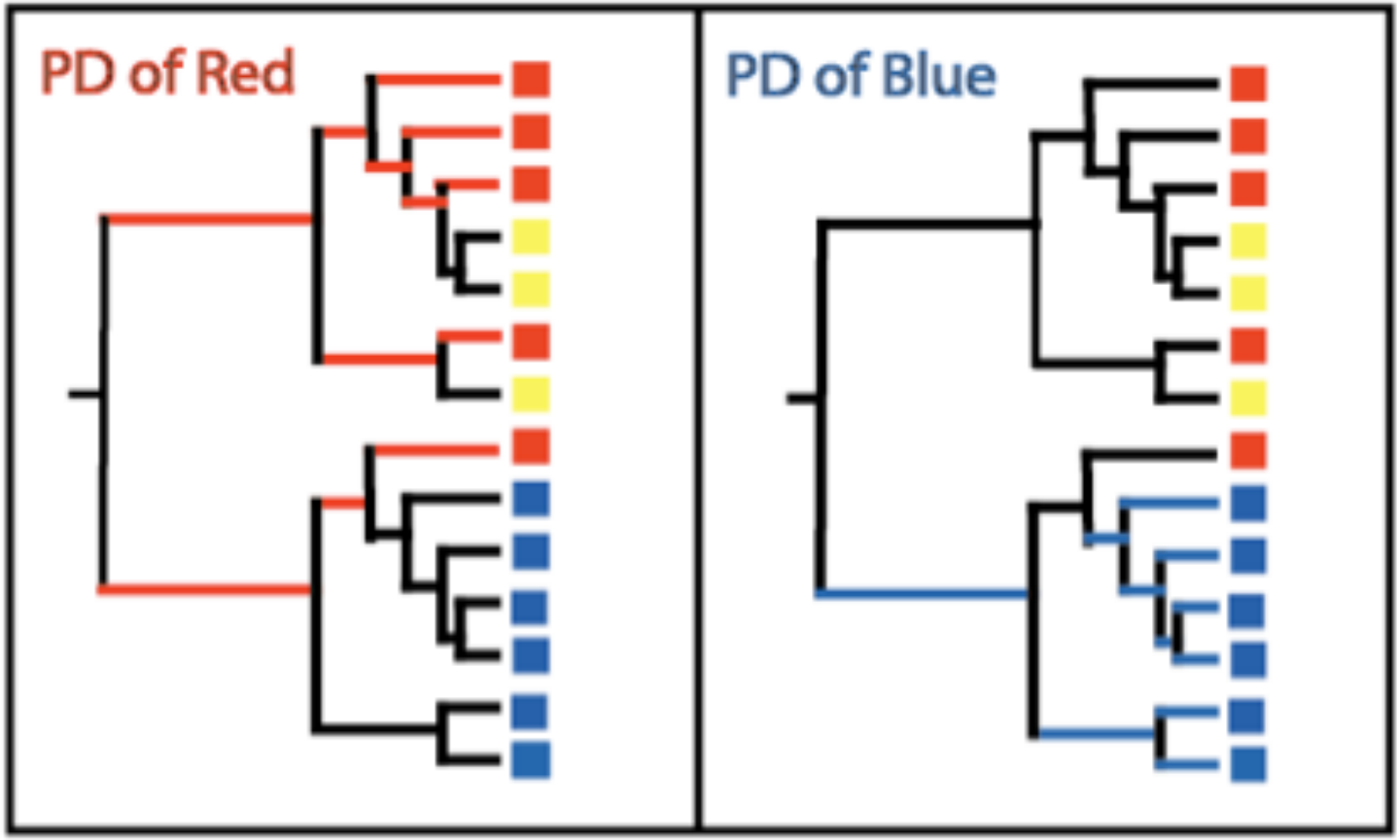
Observed species

Sample A 3
Sample B 3
Sample C 3

Conclusion

A = B = C

Phylogenetic Diversity (PD):
a qualitative, phylogenetic α-diversity metric



Sum of branch length covered by a sample

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

Alpha diversity

Sample A

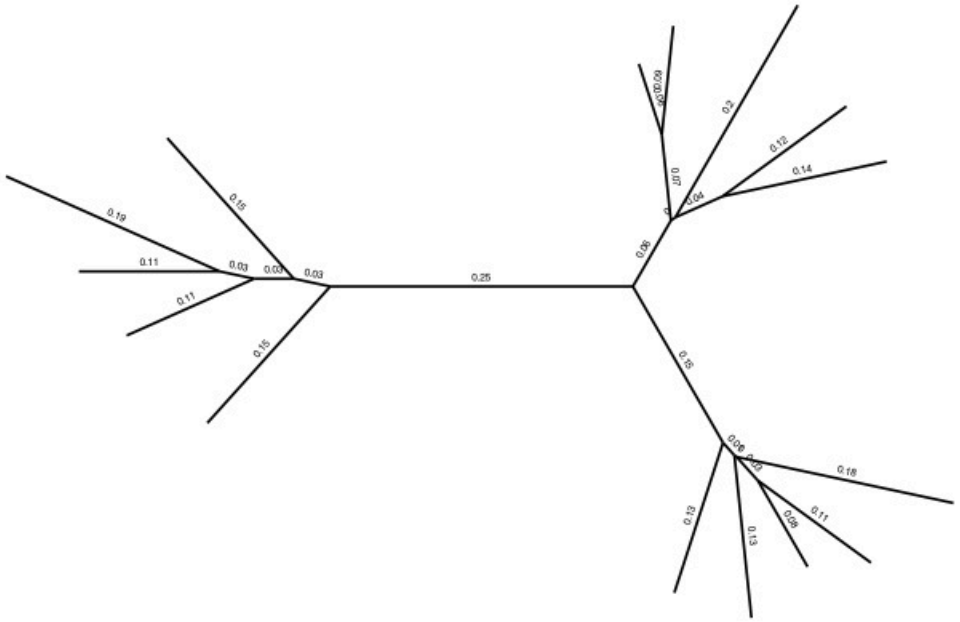
Pseudomonas aeruginosa
Pseudomonas argentinensis
Pseudomonas flavescens

Sample B

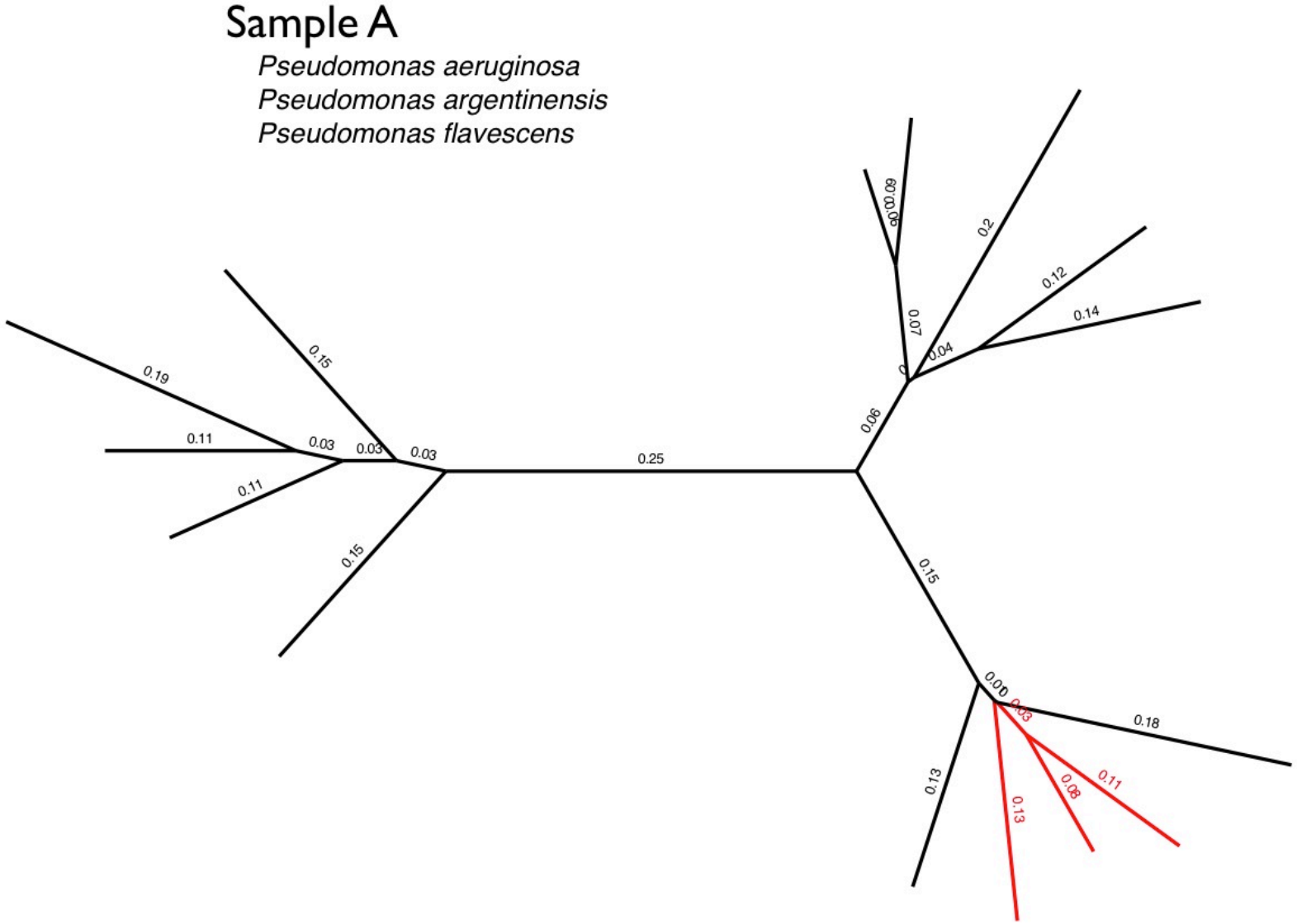
Pseudomonas aeruginosa
Pseudomonas argentinensis
Escherichia coli

Sample C

Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii

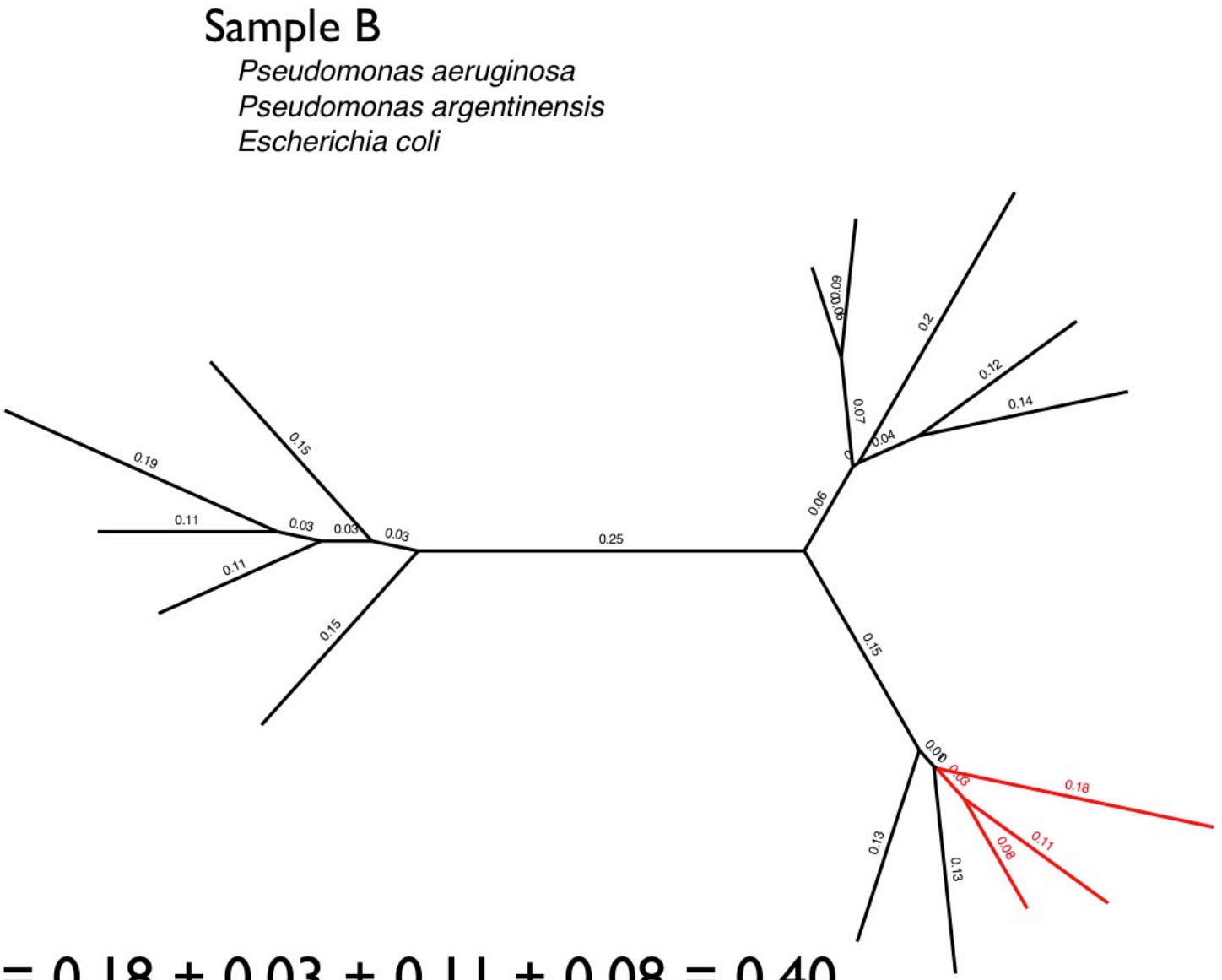


Alpha diversity



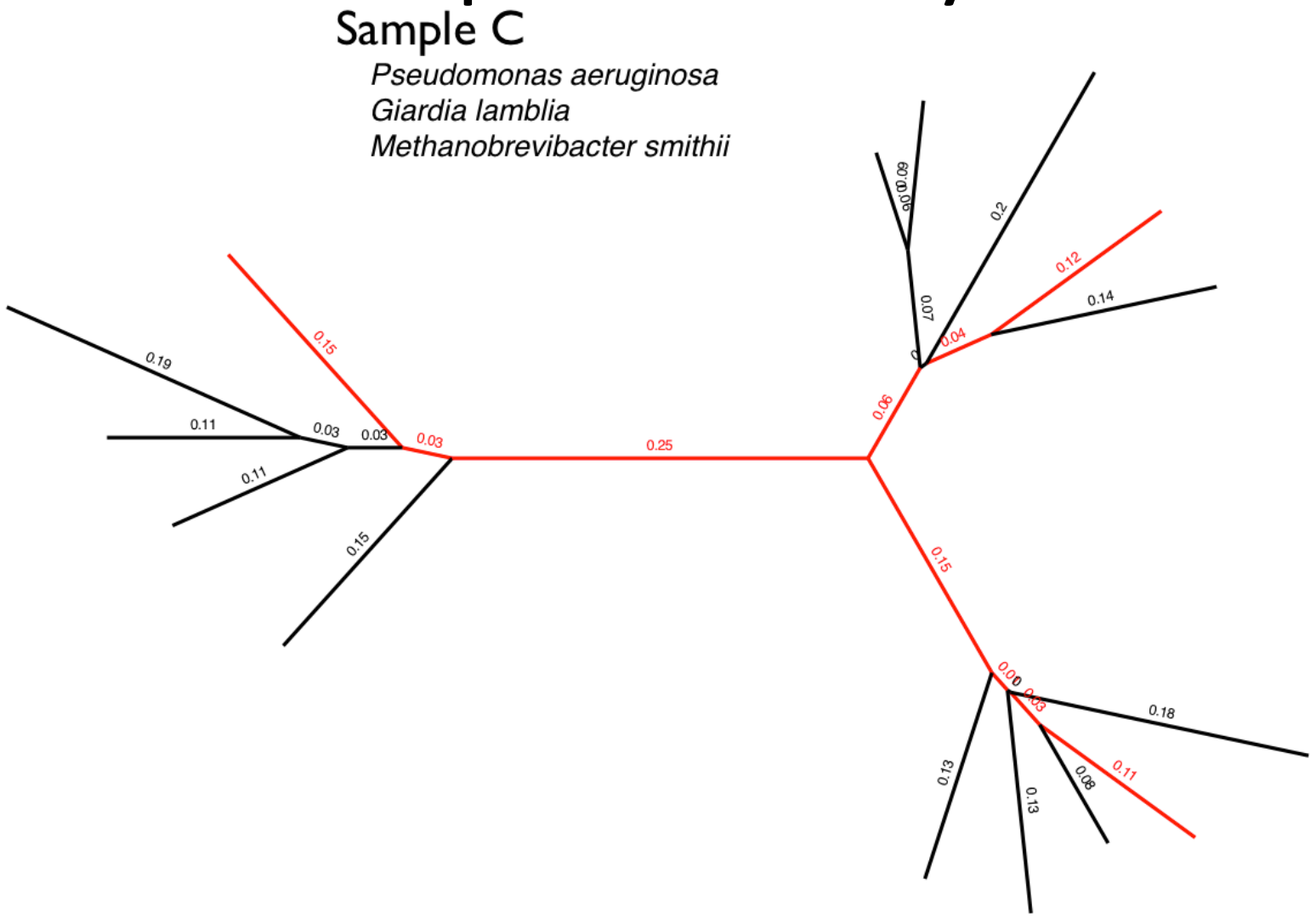
PD = 0.13 + 0.03 + 0.11 + 0.08 = 0.35

Alpha diversity



PD = 0.18 + 0.03 + 0.11 + 0.08 = 0.40

Alpha diversity



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

Alpha diversity

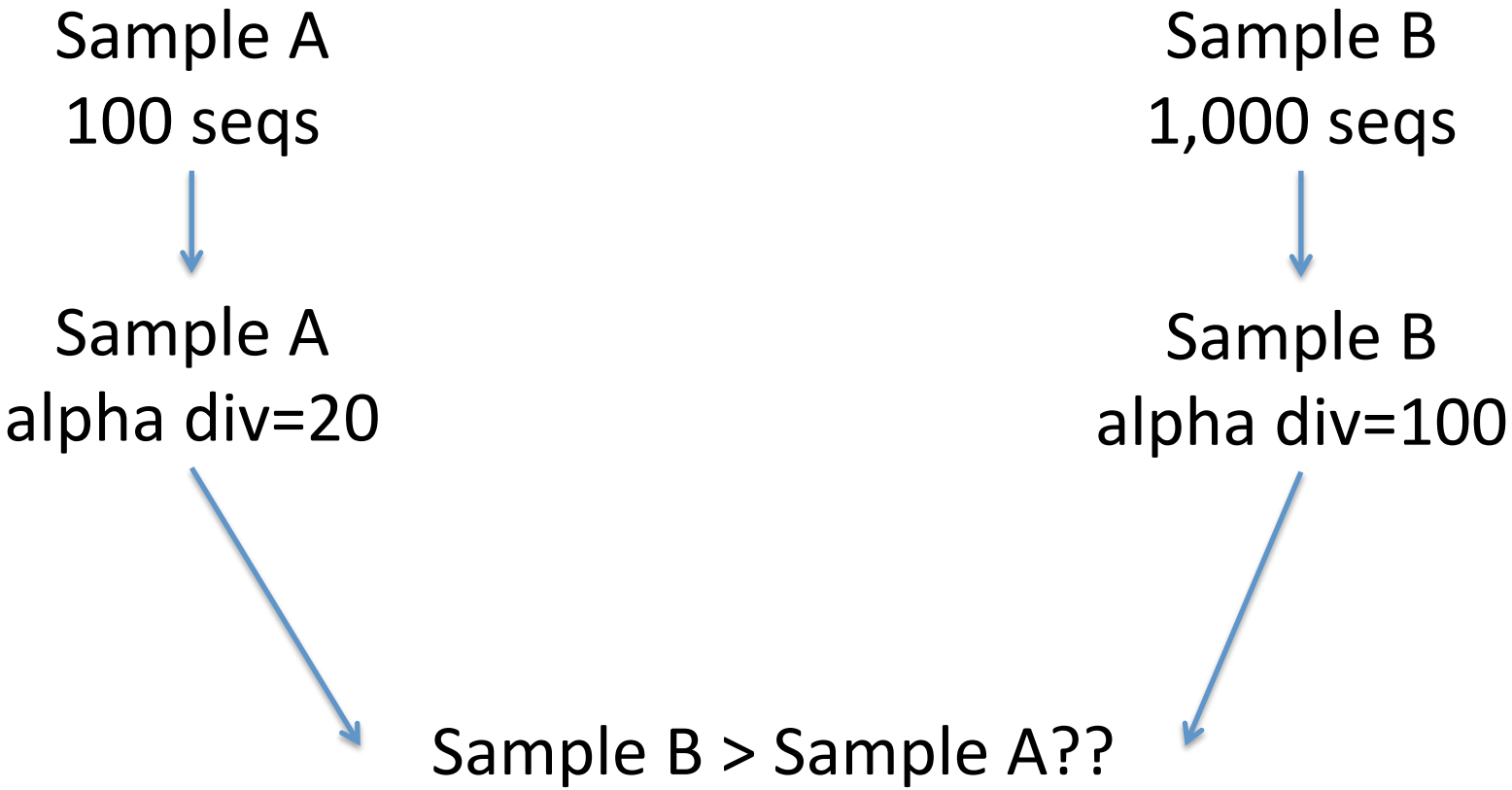
Sample A	**Sample B**	**Sample C**
Pseudomonas aeruginosa	Pseudomonas aeruginosa	Pseudomonas aeruginosa
Pseudomonas argentinensis	Pseudomonas argentinensis	Giardia lamblia
Pseudomonas flavescens	Escherichia coli	Methanobrevibacter smithii

PD = 0.35 < PD = 0.40 < PD = 0.95

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

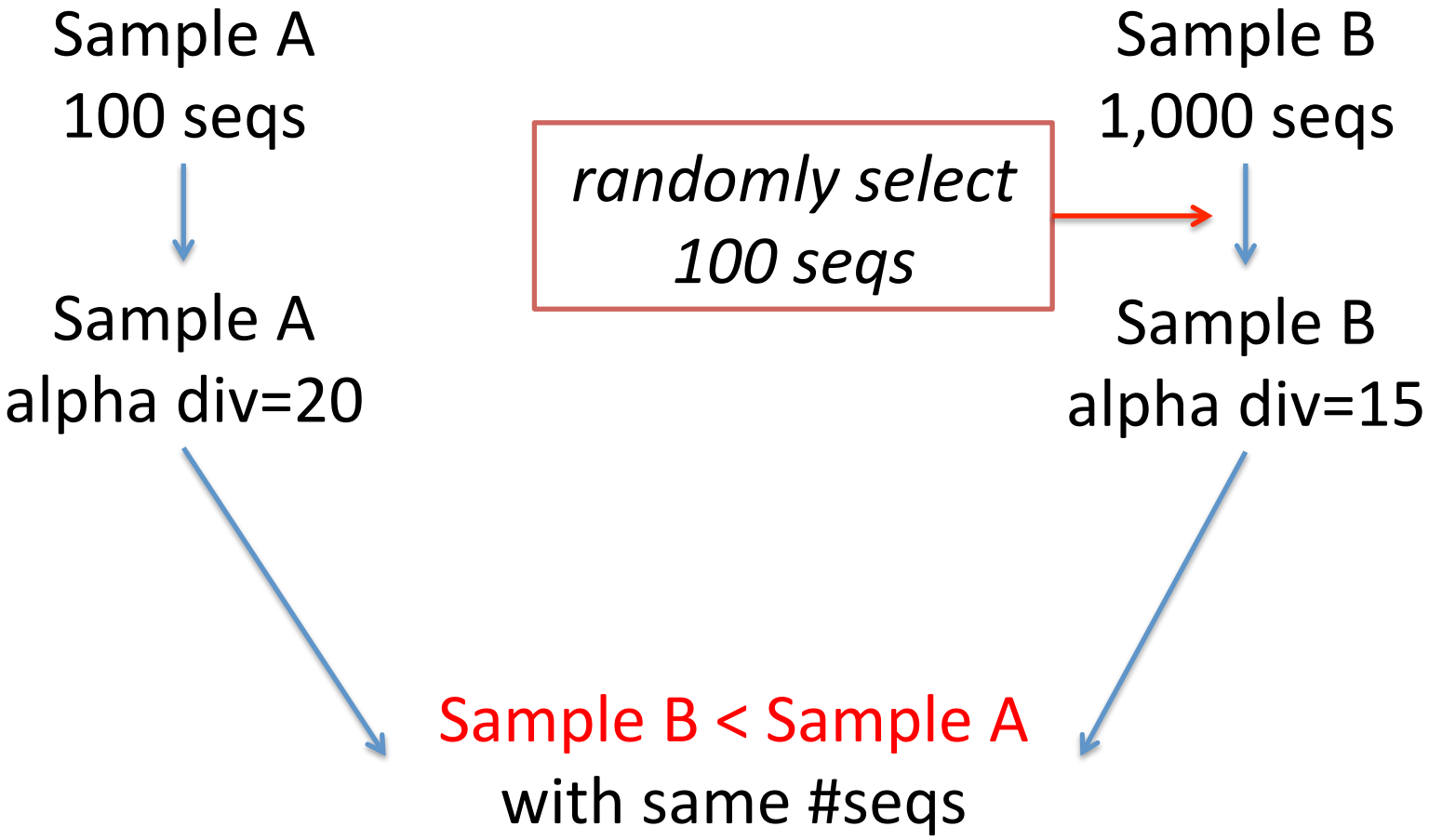
Alpha rarefaction

Alpha rarefaction



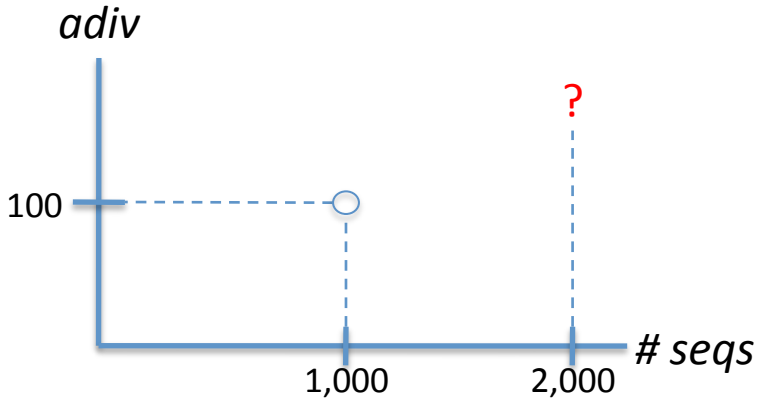
Alpha rarefaction

Alpha rarefaction



Multiple alpha rarefaction

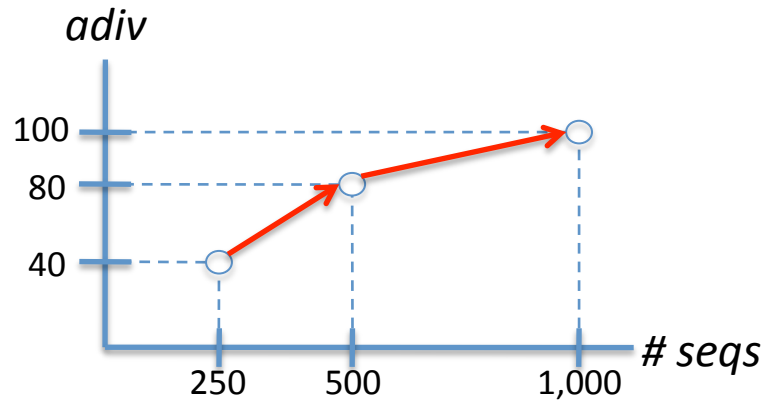
Sample A
Alpha div = 100
with 1,000 seqs



What if we had 2,000 seqs?

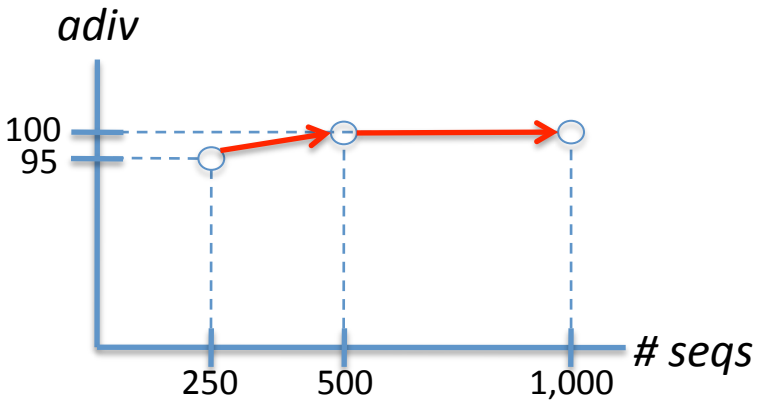
Repeatedly calculate alpha div
at decreasing number of seqs

Multiple alpha rarefaction



Higher sequencing effort might result in higher observed diversity

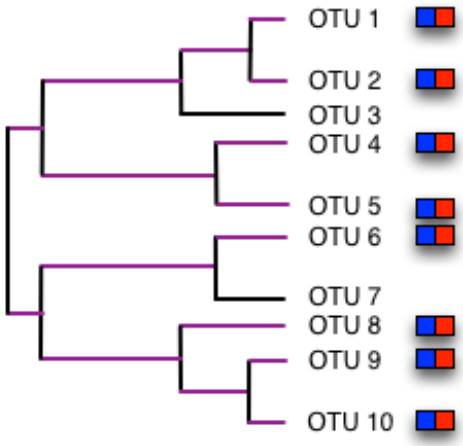
Multiple alpha rarefaction



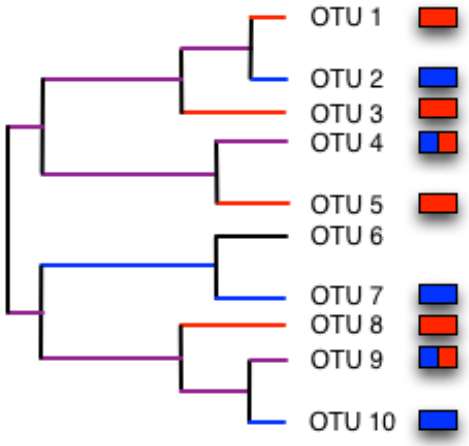
Higher sequencing effort will probably not add to observed diversity

Beta (between sample) diversity

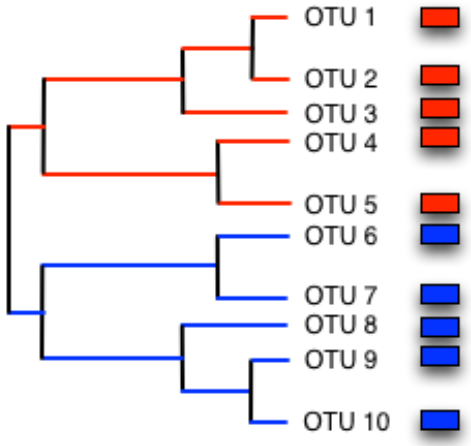
Unweighted Unifrac: a phylogenetic measure of the dissimilarity of microbial communities



U = 0.0



U ≈ 0.5



U = 1.0

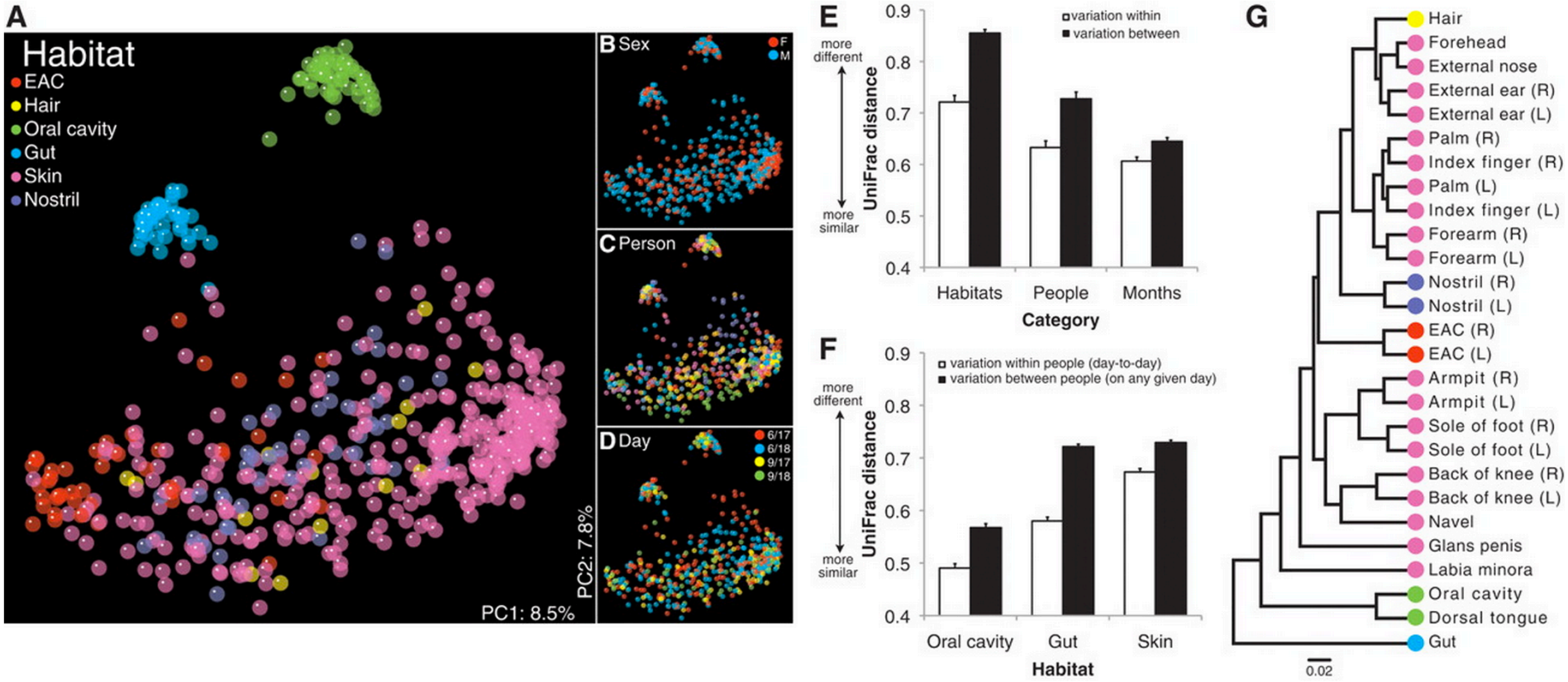
$$U_{AB} = \frac{unique}{observed}$$

where:

$unique$: the unique branch length, or branch length that only leads to OTU(s) observed in sample A or sample B

$observed$: the total branch length observed in either sample A or sample B

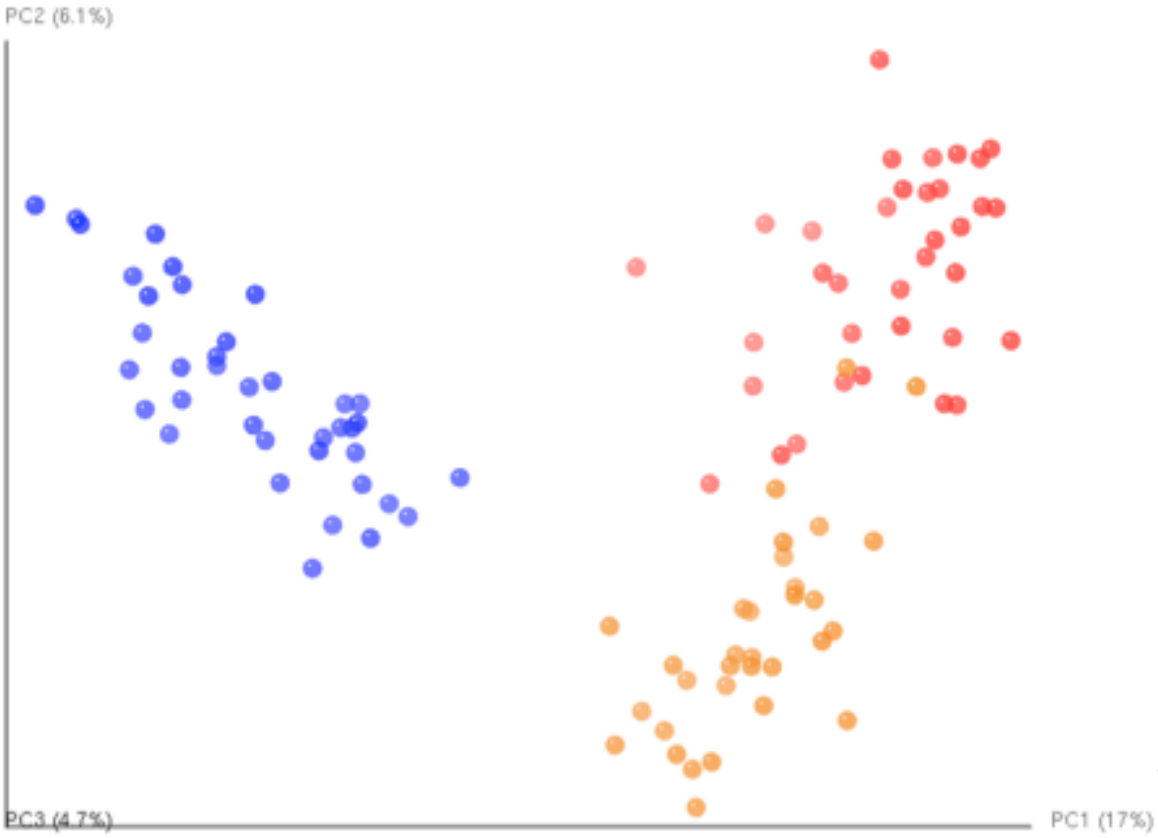
Pairwise distances between samples are the basis of most microbiome surveys



Bacterial Community Variation in Human Body Habitats Across Space and Time.
Costello et al. *Science* (2009)

Variation in sampling depth also needs to be controlled for beta diversity!

Variation in sampling depth is an important consideration

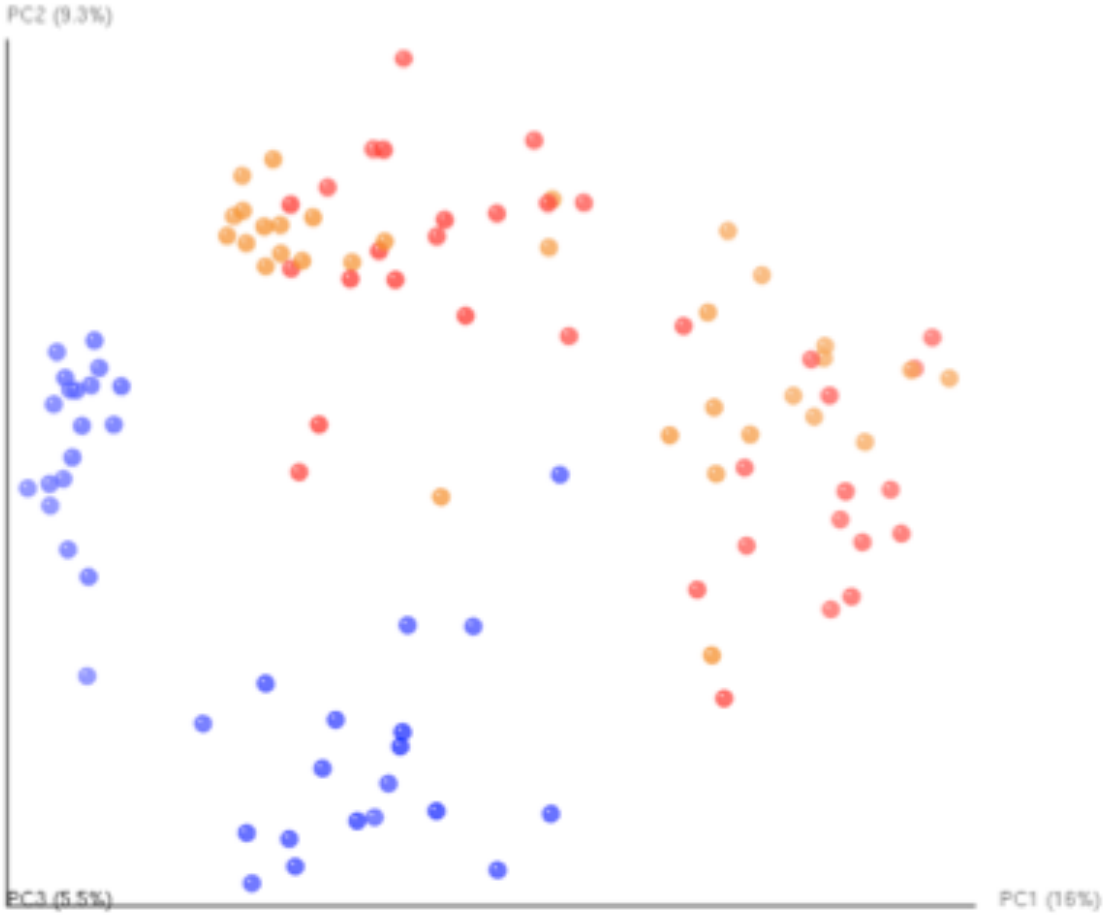


Human skin, colored by individual, at 500 sequence/sample

Image/analysis credit: Justin Kuczynski

Data reference:
Forensic identification using skin bacterial communities. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. Proc Natl Acad Sci U S A. 2010 Apr 6;107(14):6477-81.

Variation in sampling depth is an important consideration



Human skin, colored by sampling depth, at either 50 or 500 sequences/sample

Image/analysis credit: Justin Kuczynski

Data reference:
Forensic identification using skin bacterial communities. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. Proc Natl Acad Sci U S A. 2010 Apr 6;107(14):6477-81.

Variation in sampling depth is an important consideration

Human skin, colored by sampling depth, at either 50 (blue) or 500 (red) sequences/sample

Image/analysis credit: Justin Kuczynski

Data reference:
Forensic identification using skin bacterial communities. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. Proc Natl Acad Sci U S A. 2010 Apr 6;107(14):6477-81.

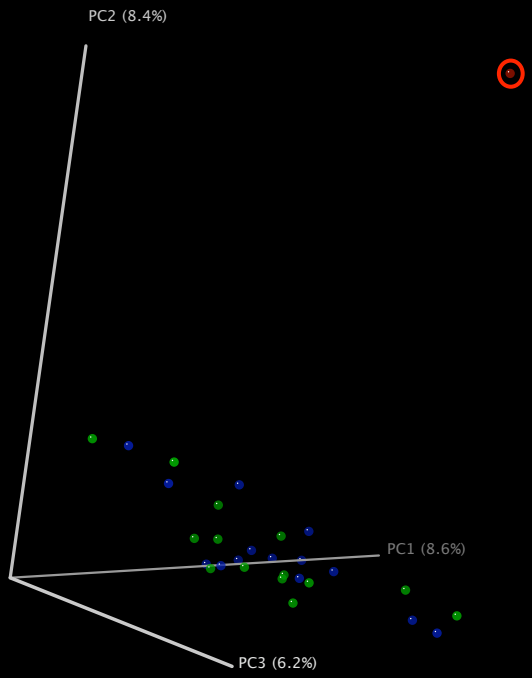
How deep is deep enough?

It depends on the question…

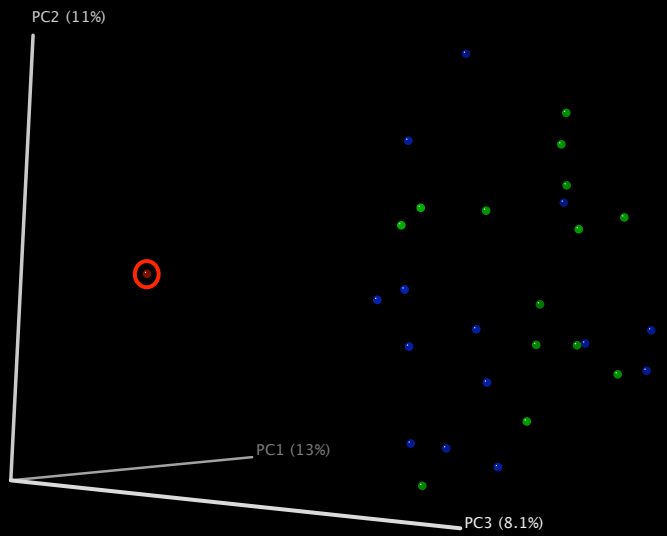
- Differences between community types: not many sequences.
- Rare biosphere: more (but be careful about sequencing noise!)

How deep is deep enough?

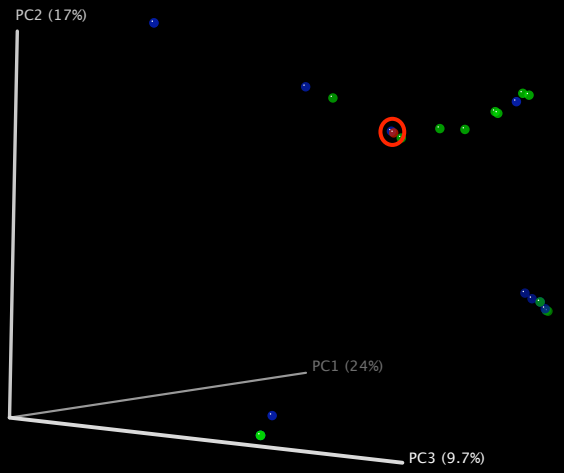
100 sequences/sample



10 sequences/sample



1 sequence/sample



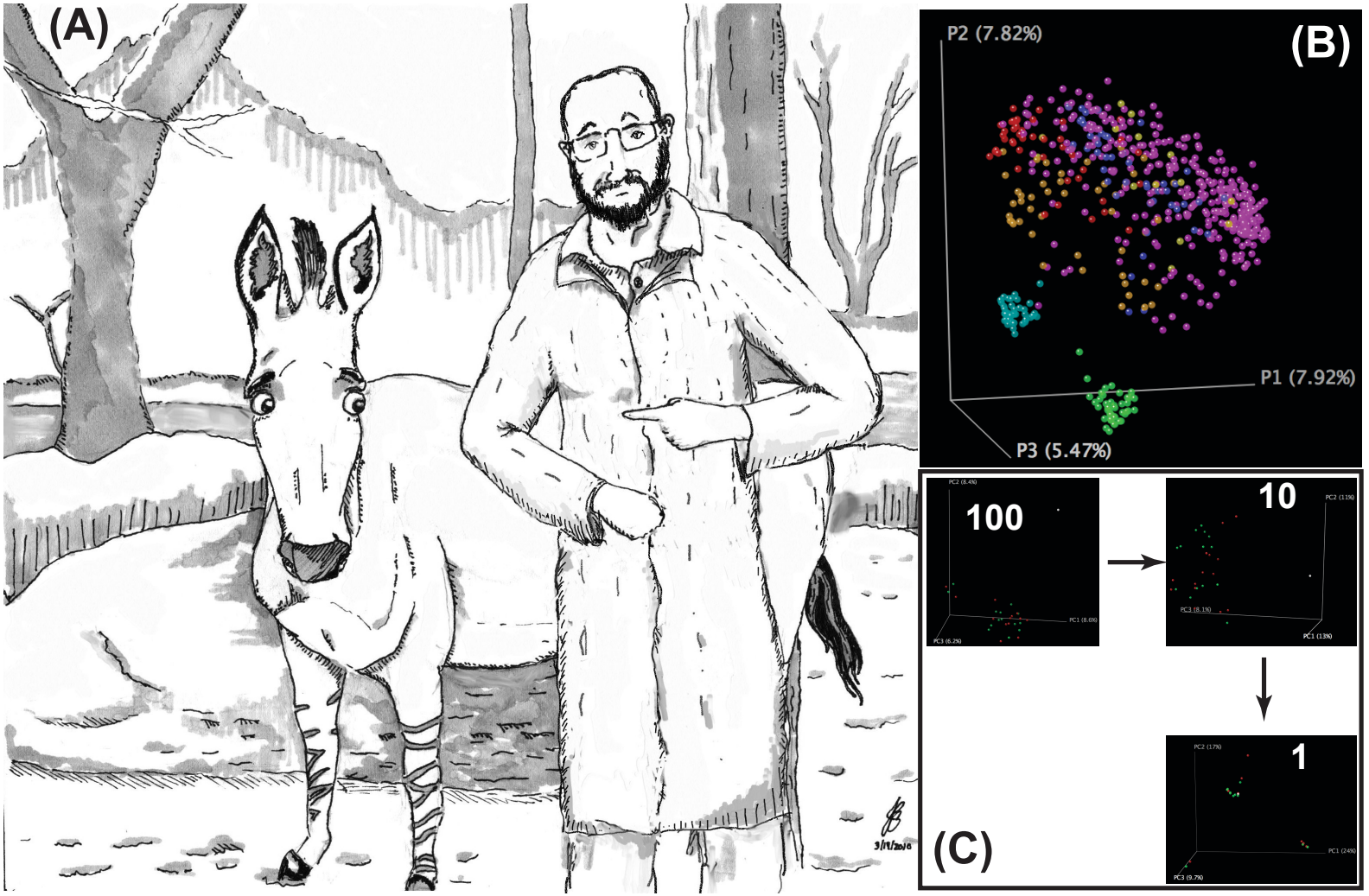
Direct sequencing of the human microbiome readily reveals community differences.
J Kuczynski et al. Genome Biology (2011).

Genome Biol. 2010;11(5):210. Epub 2010 May 5.

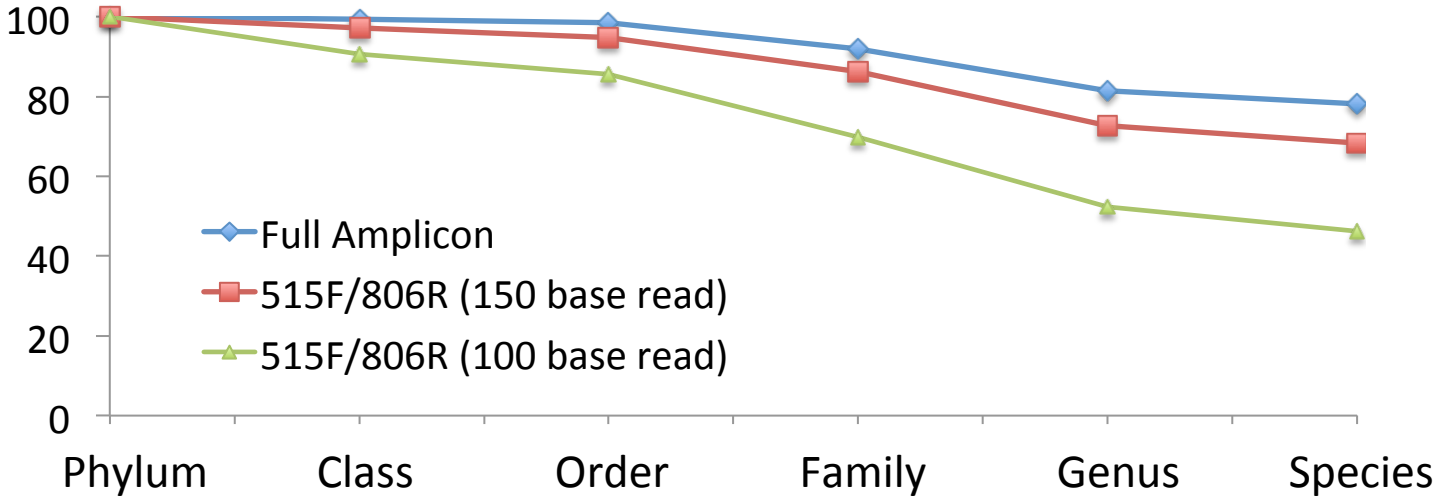
Direct sequencing of the human microbiome readily reveals community differences.

Kuczynski J, Costello EK, Nemergut DR, Zaneveld J, Lauber CL, Knights D, Koren O, Fierer N, Kelley ST, Ley RE, Gordon JI, Knight R.

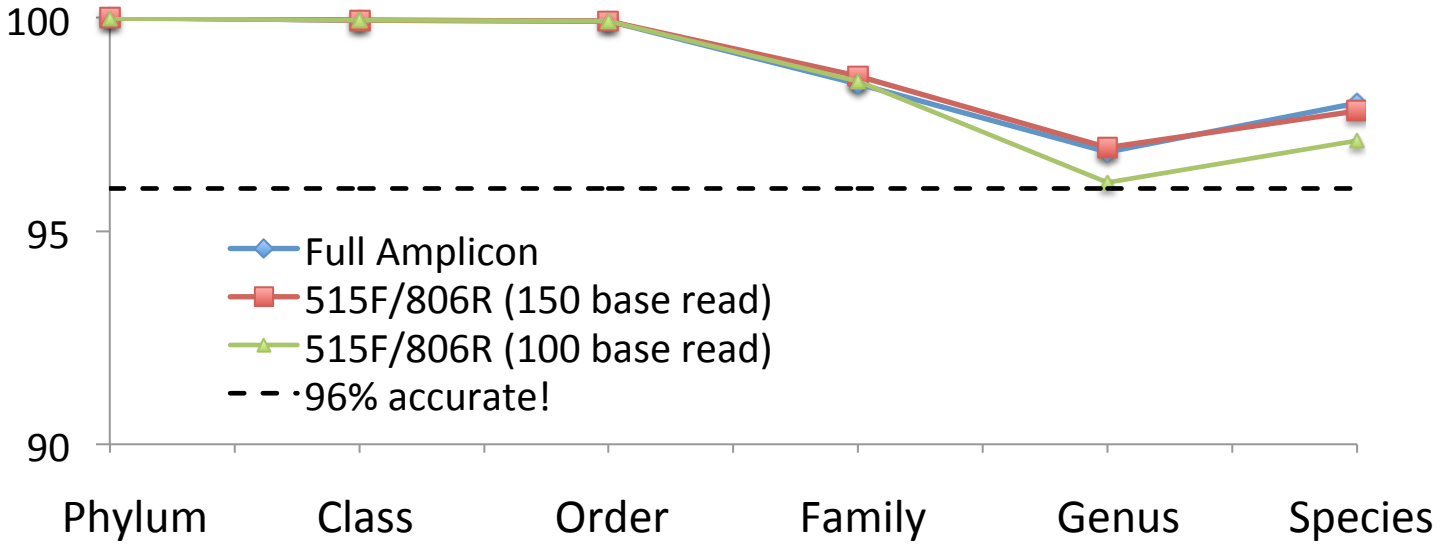
Figure 1



Fraction of Greengenes *simulated reads* classified by taxonomic level using the RDP Classifier (80% confidence)

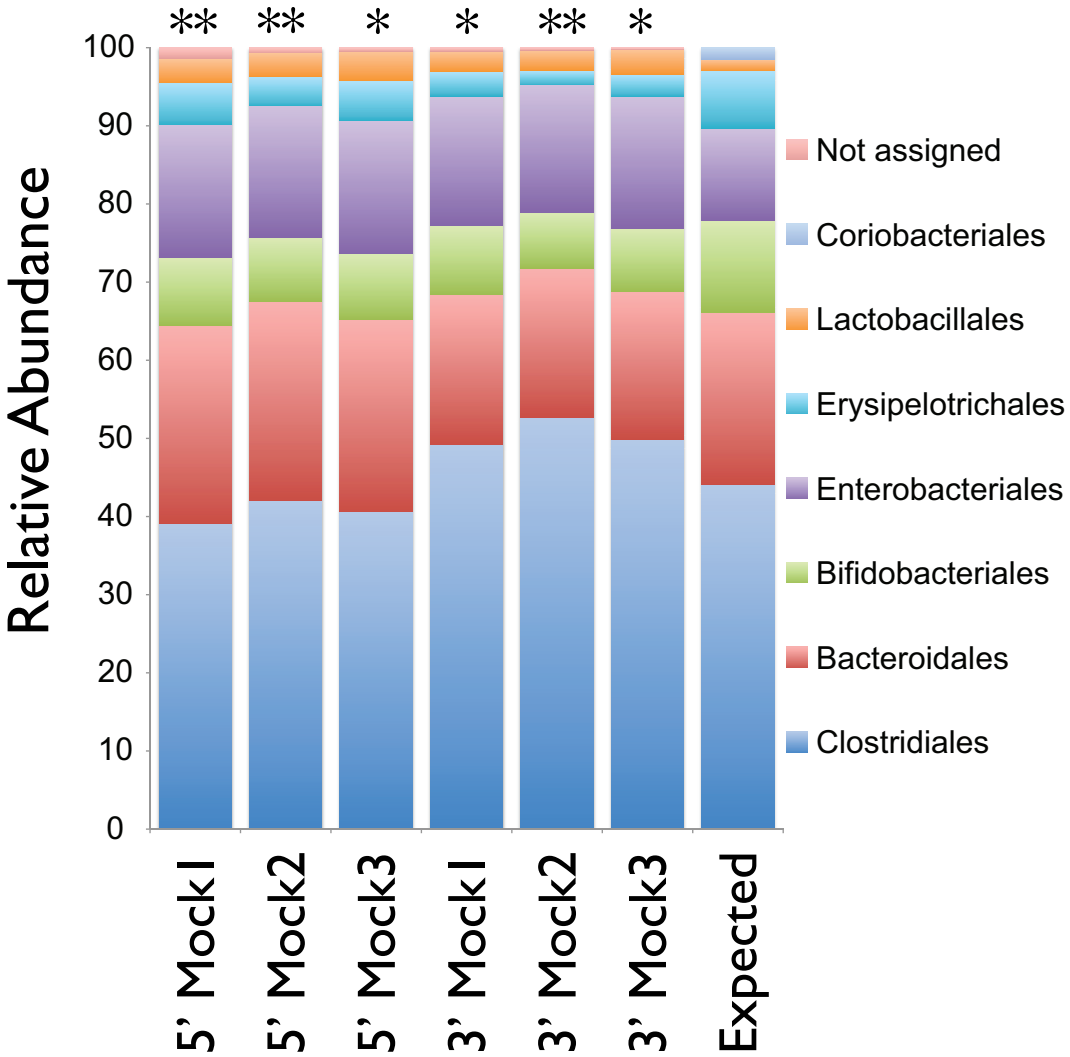


Accuracy of classified reads



Acknowledgement: Tony Walters

Can accurate taxonomy assignments be achieved?



Order-level taxonomy assignments

G-test (goodness of fit)

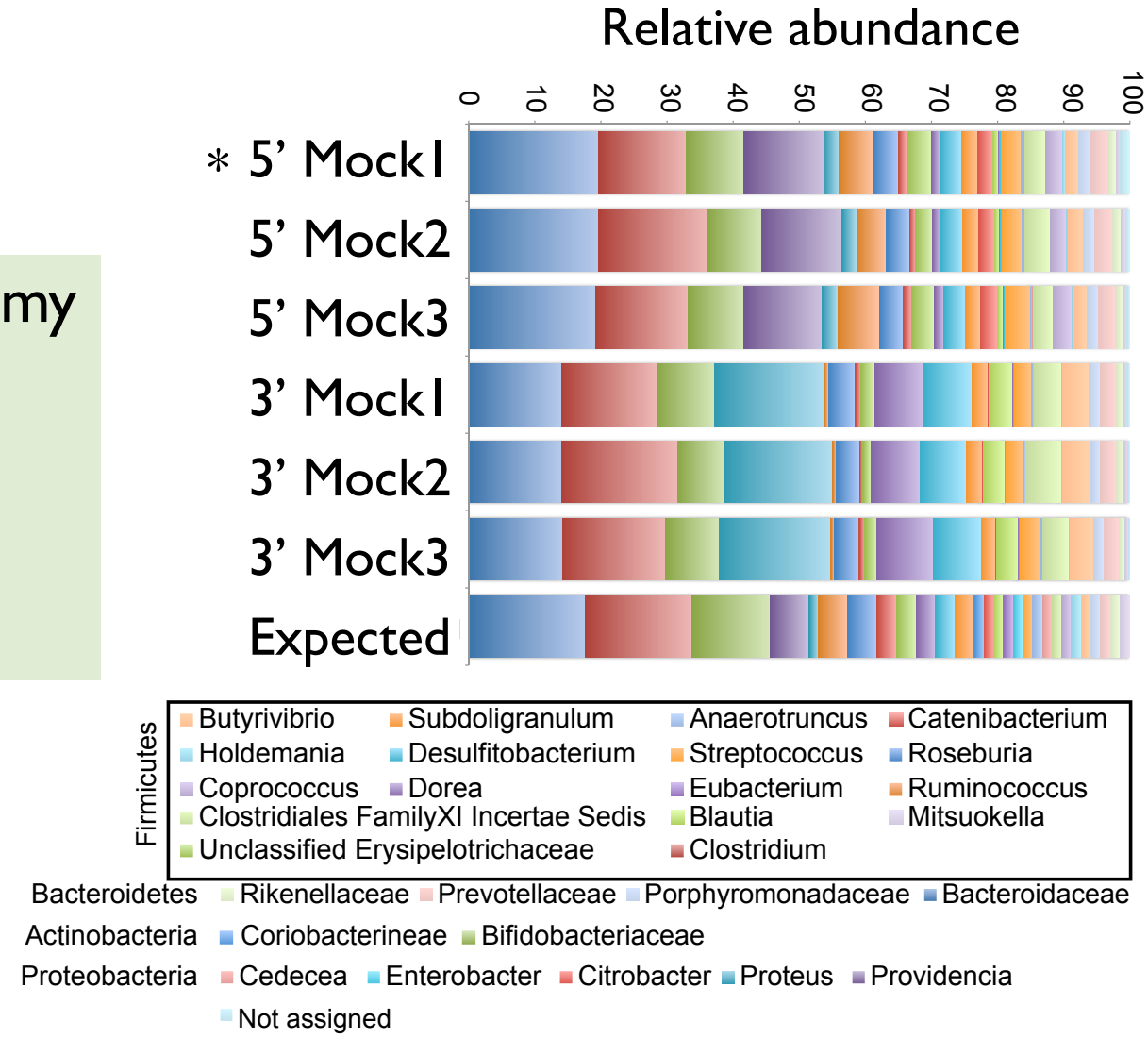
** $p < 0.01$

* $p < 0.05$

Can accurate taxonomy assignments be achieved?

Genus-level taxonomy assignments

G-test (goodness of fit)
** p < 0.01
* p < 0.05



Working with OTU tables

QIIME Workshop Day 2

Jai Rideout

jai.rideout@gmail.com

Slides credit:
Greg Caporaso, caporasolab.us

Working with OTU tables

- `single_rarefaction.py:` even sampling *(very important if you have different numbers of seqs/sample!)*
- `filter_otus_from_otu_table.py`
- `filter_samples_from_otu_table.py`
- `filter_taxa_from_otu_table.py`
- `merge_otu_tables.py`
- `sort_otu_table.py`
- `split_otu_table.py`
- `split_otu_table_by_taxonomy.py`
- `biom summarize-table`
- `biom add-metadata`

http://qiime.org/tutorials/working_with_biom_tables.html

Working with OTU tables

- `single_rarefaction.py:` even sampling *(very important if you have different numbers of seqs/sample!)*
- **`filter_otus_from_otu_table.py`**
- **`filter_samples_from_otu_table.py`**
- **`filter_taxa_from_otu_table.py`**
- `merge_otu_tables.py`
- **`sort_otu_table.py`**
- `split_otu_table.py`
- **`split_otu_table_by_taxonomy.py`**
- `biom summarize-table`
- `biom add-metadata`

http://qiime.org/tutorials/working_with_biom_tables.html

Getting started

- Log into your EC2 instance
- cd into your personal folder
  - Example: `cd jai_rideout`
- Open http://bit.ly/1r4DG21 in a web browser
  - Contains the commands we'll run during the tutorial

Future directions: QiiTA and QIIME 2

QIIME Workshop Day 2

Jai Rideout

jai.rideout@gmail.com

QiiTA pre-history

- Previous iteration known as the “QIIME Database” (QIIME-DB)
- Suffered database crash
  - Efforts focused on rewrite instead of recover

QiiTA: QIIME-DB Reboot

- System for:
  - Depositing/archiving microbiome data
  - Performing meta-analysis
    - Combine data from a variety of sources (marker gene, metagenomic, metabolomic, etc.)
- Goals
  - Easy-to-use web interface
  - User-deployable in a variety of environments (e.g., laptops to clusters)
  - Powerful meta-analysis capabilities

Moving toward QIIME 2

- QIIME is currently:
  - Command-line only (very limited Galaxy support)
  - Simplistic execution of workflows
  - Hard to extend and maintain (for both users and devs)
  - Can be difficult to install

Moving toward QIIME 2

- Most requested feature: graphical interface
- Most support efforts: command-line issues

Users spend too much time grappling with the command line and less time performing awesome microbiome research.

Devs spend too much time helping users with installation and command-line issues, and less time answering users' research questions.

QIIME 2 Overview

- Complete redesign/rebuild of QIIME
- Powered by scikit-bio (http://scikit-bio.org)
- Graphical web-based interface
  - Drag-and-drop analyses
  - Customizable transparent workflows with DAG execution
  - Provenance tracking
- Command-line interface and Python API
- Deployable on laptops -> clusters
- Extendable by users/devs

QIIME 2 Overview

- Currently in requirements and design phase
- All discussion, design, and development is **publicly available**
  - Get involved at https://github.com/biocore/metoo

Reproducible bioinformatics

QIIME Workshop Day 2

Jai Rideout

jai.rideout@gmail.com

Slides credit:
Greg Caporaso, caporasolab.us

Why is it important for a scientific experiment to be reproducible?

- Differentiate real results from experimental artifacts

Reproducible versus replicable

- Replicable: exact same conditions lead to concordant results
- Reproducible: some experimental variation is allowed, but results are concordant

What does it mean for a bioinformatics experiment to be replicable?

- Our experimental methods are not as 'noisy' as most.
- Same commands on the same system should give you the same results (if the algorithm is *deterministic*).

Deterministic versus non-deterministic

- Deterministic algorithm: a given input produces the same series of internal states and results in the same output.
- Non-deterministic algorithm: a given input may produce different internal states and/or result in a different output.
  - Commonly probabilistic algorithms in bioinformatics

Deterministic

- Smith-Waterman alignment: if properly implemented, aligning two sequences will always give the same result

Non-deterministic/Probabilistic

- Sub-sampling a data set (e.g., rarefaction of an OTU table)
- Jackknifed analyses
- Permutation-based p-values (Monte Carlo)

How can we develop software that supports reproducibility?

Open source software

- Consider making your software open source
  - Avoid “black box”
  - Benefits other researchers/community
  - Encourages collaboration
  - Extendibility
- Place it under public hosted revision control
  - GitHub
  - Bitbucket
  - SourceForge

Version control systems

- Git
- Subversion (svn)
- Mercurial (hg)
- CVS (ancient history)

- Allow for viewing history of changes, obtaining previous versions.
- Example: https://github.com/biocore/qiime

Virtual Machines

- Publish virtual machine images
  - Gives access to exact software, configuration, and data used in analyses

What's in a good log file?

- Ideally will supplement your lab notebook (for successful runs)
  - Version information
  - Exact commands that were run
  - Details on input files (path, md5)
  - System configuration details
- Publish these as supplementary material

MD5

- A cryptographic hash function: deterministic function which takes some input and returns a fixed-size string – changing the input should change the return value

From Wikipedia:

- it is easy (but not necessarily quick) to compute the hash value for any given message
- it is infeasible to generate a message that has a given hash
- it is infeasible to modify a message without changing the hash
- it is infeasible to find two different messages with the same hash

IPython Notebook

- Interactive, executable documents: code, text, images, etc.
- Easy to share, publish, convert
- Great for keeping track of analysis commands, code, descriptions/comments, etc.
- Make your methods section *executable*!

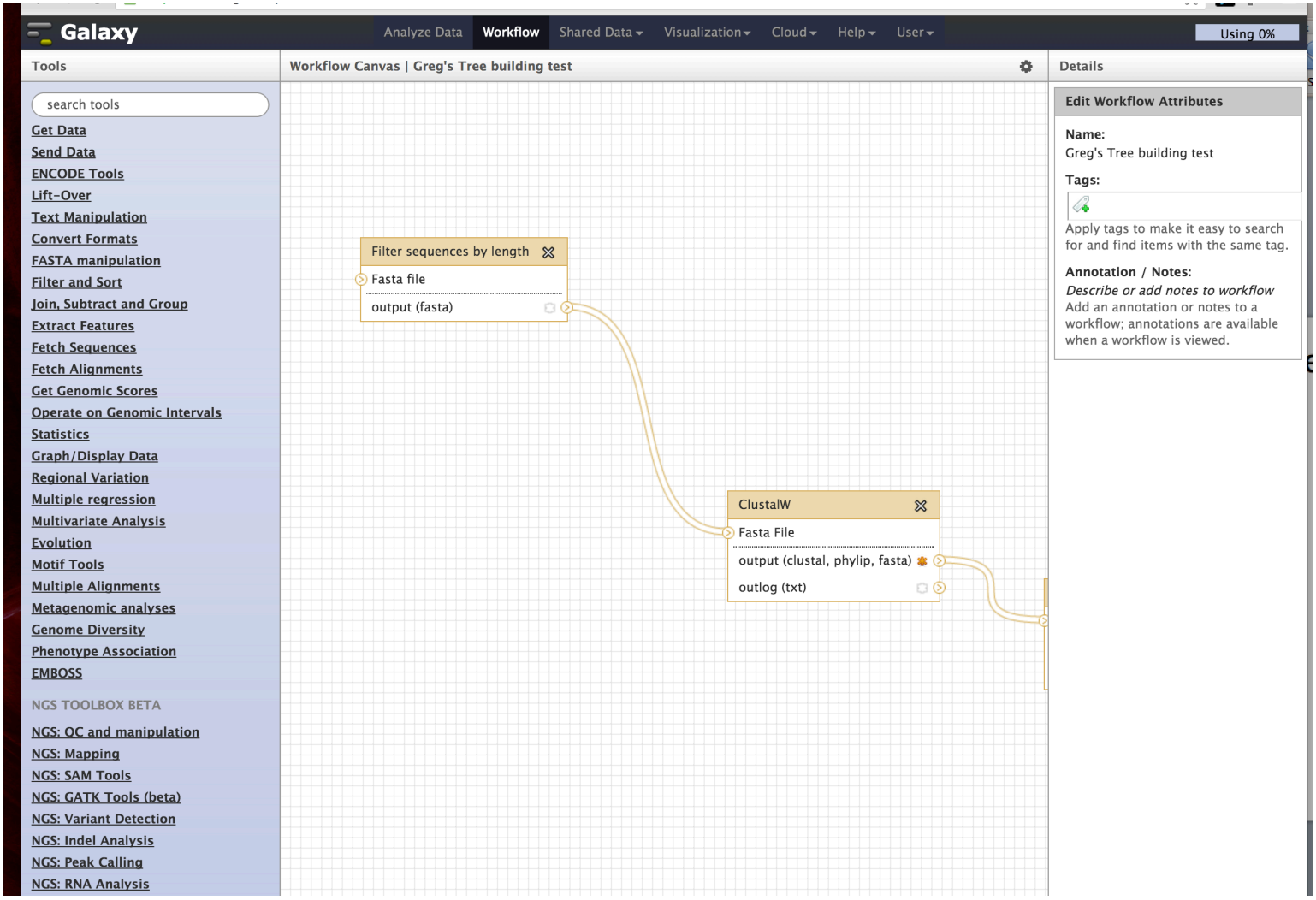
http://ipython.org/

IPython Notebook

- Workshop commands were written in IPython Notebooks
  - e.g., http://bit.ly/1nSSvdz
- An Introduction to Applied Bioinformatics
  - Greg's online bioinformatics book
  - http://applied-bioinformatics.org

http://ipython.org/

Reproducible computing through workflow engines, e.g. Galaxy



https://main.g2.bx.psu.edu

Slides compiled by:

Greg Caporaso
Jose Clemente
Antonio Gonzalez Peña
Rob Knight
Cathy Lozupone
Daniel McDonald
Jai Rideout
Yoshiki Vázquez Baeza
John Chase

www.qiime.org