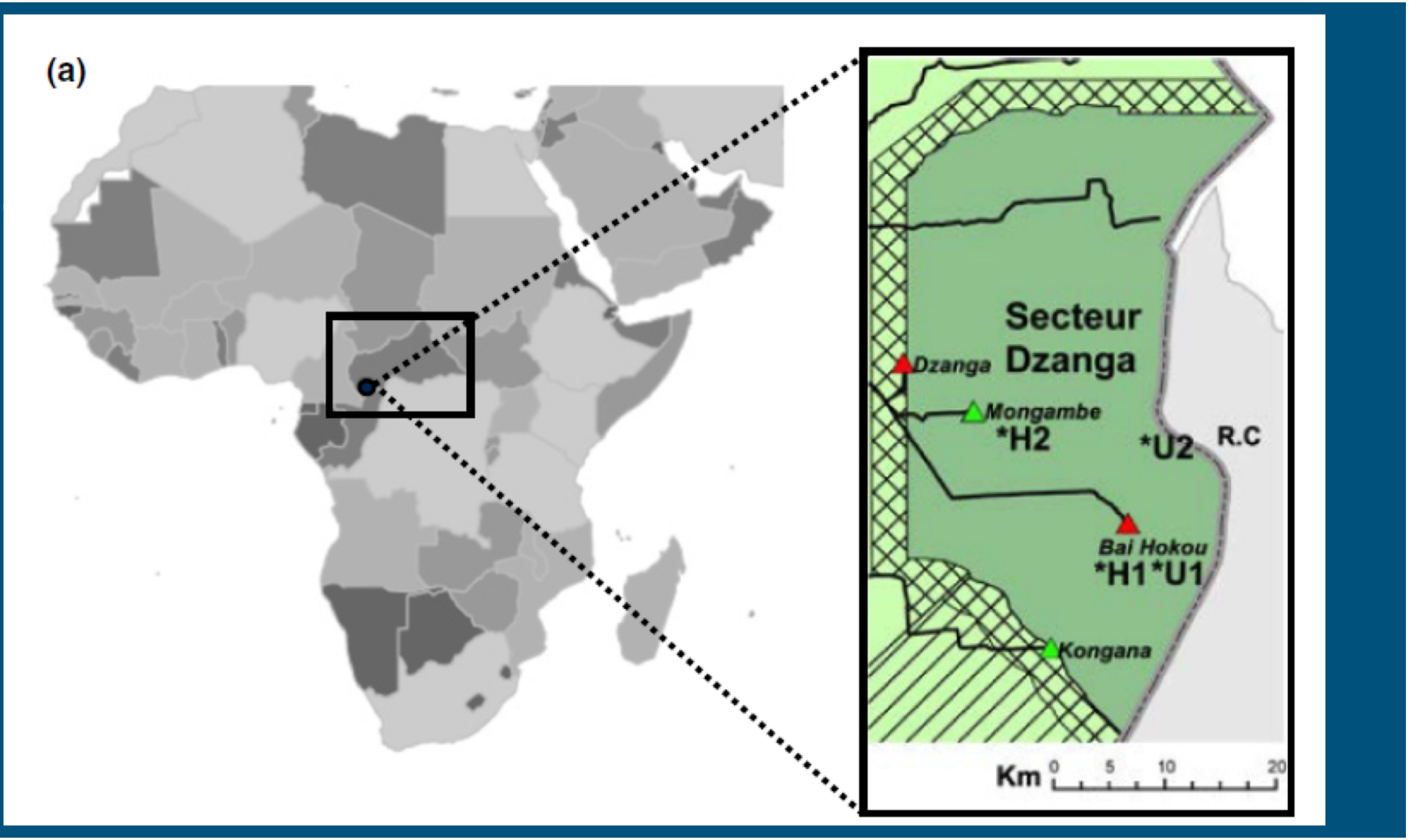
Functional Annotation of microbiomes using Qiime and Picrust

Bryan Featherstone & Dhantha Gunarathna

Gut microbiome composition and metabolomic profiles of wild western lowland gorillas (*Gorilla gorilla gorilla*) reflect host ecology

ANDRES GOMEZ,*† KLARA PETRZELKOVA,‡§¶** CARL J. YEOMAN,†† KLARA VLCKOVA,§ JAKUB MRÁZEK,‡‡ INGRID KOPPOVA,‡‡ FRANCK CARBONERO,§§ ALEXANDER ULANOV,¶¶ DAVID MODRY,¶*** ANGELIQUE TODD,††† MANOLITO TORRALBA,‡‡‡ KAREN E. NELSON,‡‡‡ H. REX GASKINS,*† BRENDA WILSON,*§§§ REBECCA M. STUMPF,*¶¶¶ BRYAN A. WHITE*† and STEVEN R. LEIGH*¶¶¶****


(a)
Secteur
Dzanga
Dzanga
Mongambe
*H2
*U2
R.C
Bai Hokou
*H1 *U1
Kongana
Km
0
5
10
20

Based out of Rob Knight's lab in University of California, San Diego.

World's largest crowd-funded citizen science project in existence

- 6,500 participants with over $1 million in fundraising

Goals:

- Build a large, public data set
- Cost effective way for the general public to know their mircrobial composition
- These data would contribute to a greater scientific effort to learn how the microbiome is

- Open source bioinformatics pipeline of microbial data
- Can take raw sequence reads through figures and statistical analyses
  - Quality Control
  - OTU tables
  - Diversity measures

PICRUSt

- Estimates gene families (functionally annotate) by using categorized bacteria or archaea of 16S rRNA sequences
- Blast like search against KEGG or COG databases identifying protein-coding genes

Installing PICRUSt

Make sure current python version is set to python 2.7

Pip install numpy==1.5

git clone git://github.com/picrust/picrust.git picrust

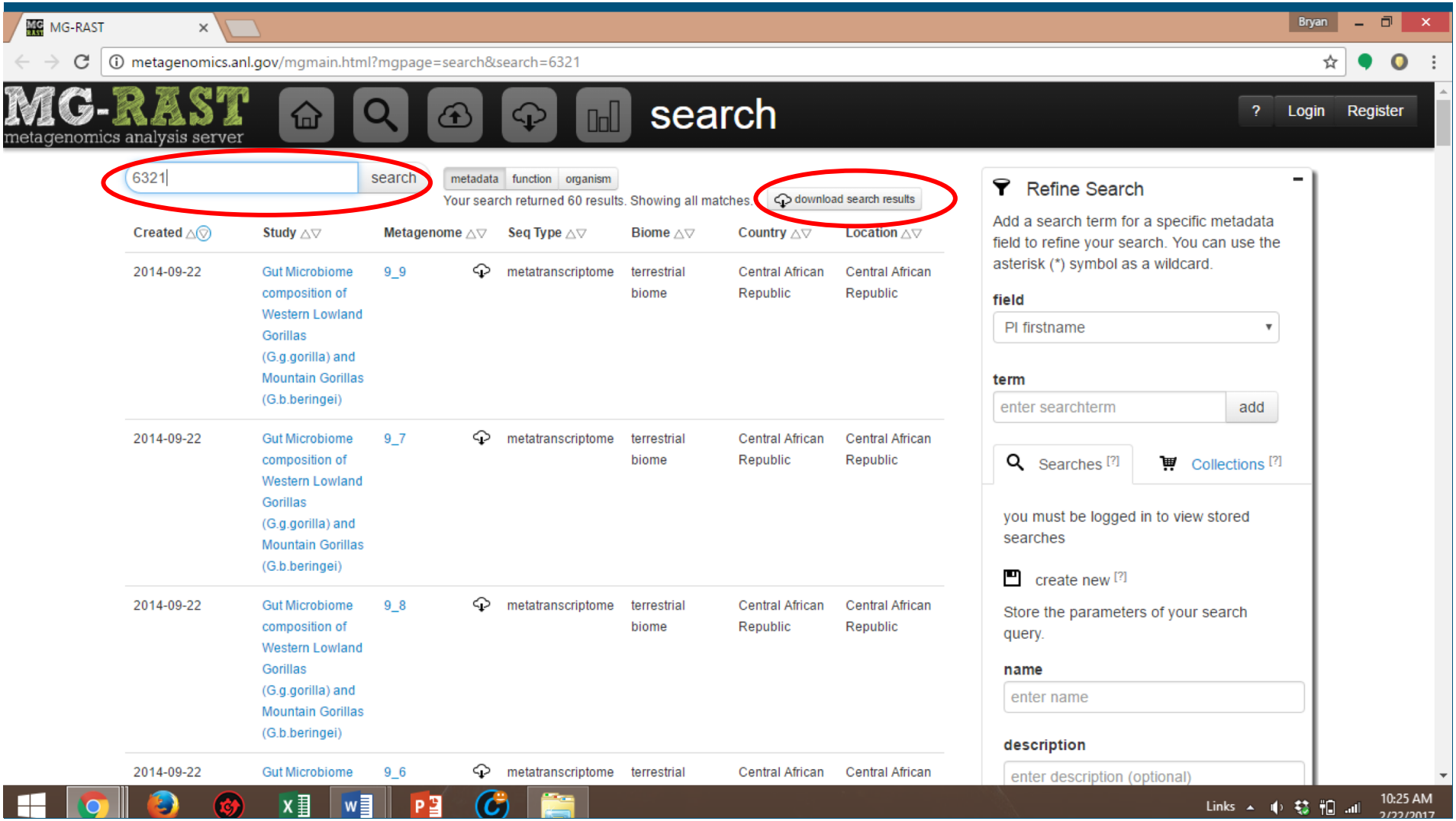
python setup.py install --user

No need to have root access. By using --user it will install locally

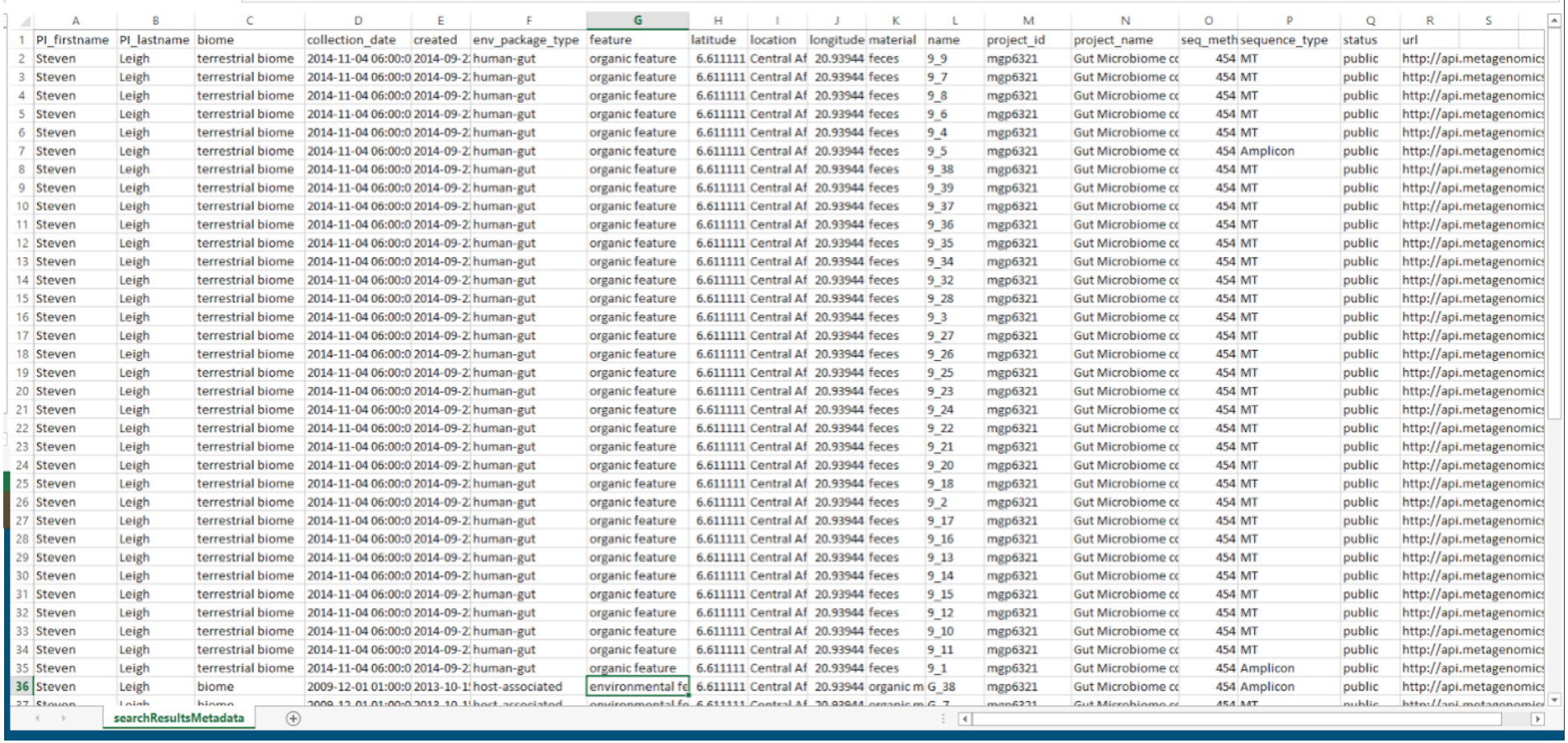
PATH=~/.local/bin/:$PATH

Downloading Data

1. **MG-RAST fasta files**
    a. **http://metagenomics.anl.gov/mgmain.html?mgpage=search&search=6321 OR Search Project 6321**
2. **MG-RAST metadata file**
    a. **Click download search results**
3. Download data onto Proteus


MG-RAST
Bryan
metagenomics.anl.gov/mgmain.html?mgpage=search&search=6321
MG-RAST
metagenomics analysis server
search
? Login Register
6321
search
metadata function organism
Your search returned 60 results. Showing all matches.
download search results
Created Study Metagenome Seq Type Biome Country Location
2014-09-22 Gut Microbiome composition of Western Lowland Gorillas (G.g.gorilla) and Mountain Gorillas (G.b.beringei) 9_9 metatranscriptome terrestrial biome Central African Republic Central African Republic
2014-09-22 Gut Microbiome composition of Western Lowland Gorillas (G.g.gorilla) and Mountain Gorillas (G.b.beringei) 9_7 metatranscriptome terrestrial biome Central African Republic Central African Republic
2014-09-22 Gut Microbiome composition of Western Lowland Gorillas (G.g.gorilla) and Mountain Gorillas (G.b.beringei) 9_8 metatranscriptome terrestrial biome Central African Republic Central African Republic
2014-09-22 Gut Microbiome 9_6 metatranscriptome terrestrial Central African Central African
Refine Search
Add a search term for a specific metadata field to refine your search. You can use the asterisk (*) symbol as a wildcard.
field
PI firstname
term
enter searchterm
add
Searches [?]
Collections [?]
you must be logged in to view stored searches
create new [?]
Store the parameters of your search query.
name
enter name
description
enter description (optional)
Links
10:25 AM

Metadata



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	PI_firstname	PI_lastname	biome	collection_date	created	env_package_type	feature	latitude	location	longitude	material	name	project_id	project_name	seq_meth	sequence_type	status	url	
2	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_9	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
3	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_7	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
4	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_8	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
5	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_6	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
6	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_4	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
7	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_5	mgp6321	Gut Microbiome cc	454	Amplicon	public	http://api.metagenomics	
8	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_38	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
9	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_39	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
10	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_37	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
11	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_36	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
12	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_35	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
13	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_34	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
14	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_32	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
15	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_28	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
16	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_3	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
17	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_27	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
18	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_26	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
19	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_25	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
20	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_23	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
21	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_24	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
22	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_22	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
23	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_21	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
24	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_20	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
25	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_18	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
26	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_2	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
27	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_17	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
28	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_16	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
29	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_13	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
30	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_14	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
31	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_15	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
32	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_12	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
33	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_10	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
34	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_11	mgp6321	Gut Microbiome cc	454	MT	public	http://api.metagenomics	
35	Steven	Leigh	terrestrial biome	2014-11-04 06:00:0	2014-09-2	human-gut	organic feature	6.611111	Central Af	20.93944	feces	9_1	mgp6321	Gut Microbiome cc	454	Amplicon	public	http://api.metagenomics	
36	Steven	Leigh	biome	2009-12-01 01:00:0	2013-10-1	host-associated	environmental fe	6.611111	Central Af	20.93944	organic m	G_38	mgp6321	Gut Microbiome cc	454	Amplicon	public	http://api.metagenomics	

searchResultsMetadata

Downloading Data

1. MG-RAST fasta files
    a. http://metagenomics.anl.gov/mgmain.html?mgpage=search&search=6321 OR Search Project 6321
2. MG-RAST metadata file
    a. Click download search results
3. **Download data onto Proteus**

Downloading Data onto Proteus

```
work_dir <- '~/genStats/gorilla'
work_dir <- '~/genStats/gorilla'
dir.create(work_dir,showWarnings=FALSE,recursive=TRUE)

work_dir <- '~/genStats/gorilla'
dir.create(work_dir,showWarnings=FALSE,recursive=TRUE)
work_dir <- '~/genStats/gorilla'
dir.create(work_dir,showWarnings=FALSE,recursive=TRUE)

file <- read.table('~/searchResultsMetadata.txt',
                   sep='\t',header=TRUE,stringsAsFactors=FALSE)
urls <- file[file$sequence_type == 'Amplicon',]$url

out_dir <- file.path(work_dir,'sequences')
dir.create(out_dir,showWarnings=FALSE,recursive=TRUE)

target <- 'http://api.metagenomics.anl.gov/1/download/%s?file=150.1'

for (url in urls){
   id <- gsub('.*metagenome\\/(.*)$','\\1',url)
   file_url <- sprintf(target,id)
   out_path <- paste0(file.path(out_dir,id),'.fna')
   download.file(file_url,destfile=out_path)
```

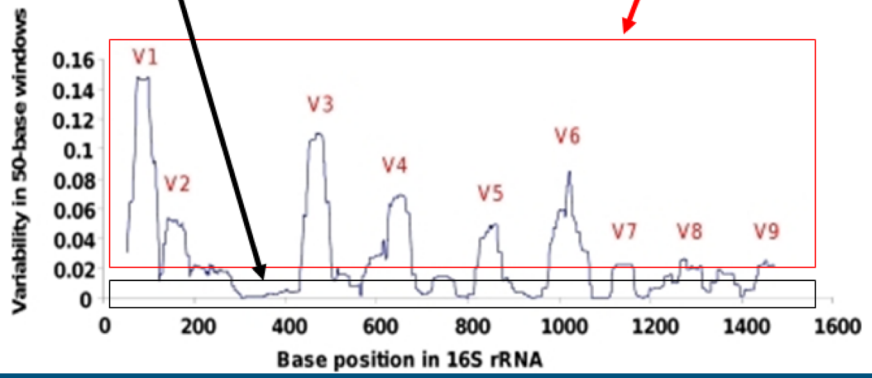
Data

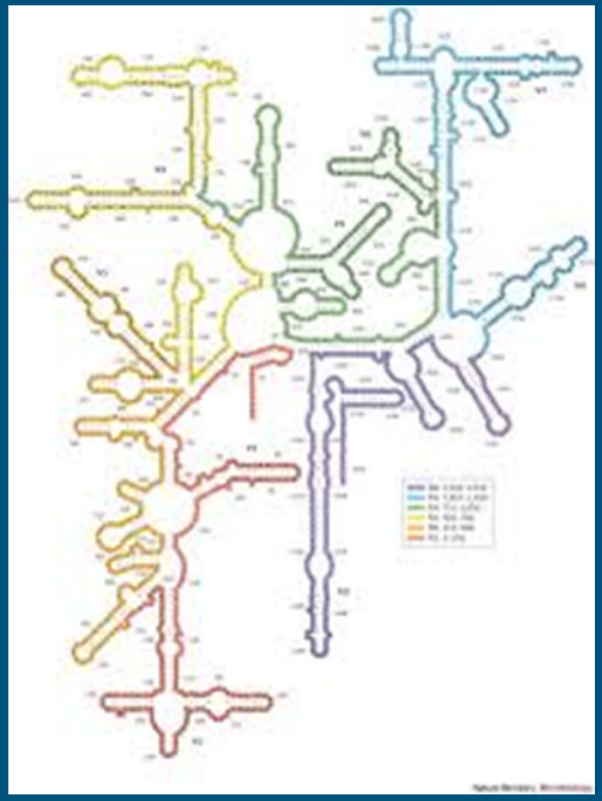
Fasta files

Reads

16S Gene

- Highly conserved regions
- Hypervariable regions



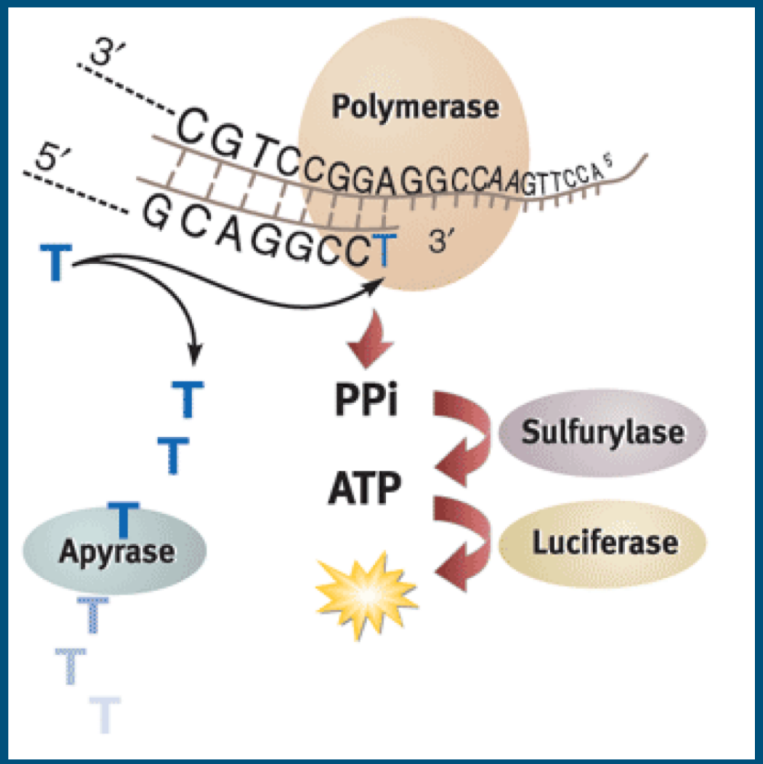


454 Pyrosequencing

Next gen sequencing technology using pyrophosphate release on nucleotide incorporation.

Difficult to read consecutive repeated nucleotides

Denoising to correct for this error



Downloading green genes

Green genes can be downloaded using the following link

http://qiime.org/home_static/dataFiles.html

Green genes is mostly 16S rRNA for browsing, blasting and probing

Picrust requires to use green genes database

All the genes in the green genes database are functionally annotated

Qiime Commands

1. Check user's metadata mapping file for required data

validate_mapping_file.py -m Matadata.txt -o validate_mapping_file_output

2. Add qiime labels. Metadata mapping file with sample IDs and fasta file names

add_qiime_labels.py -i fasta/ -m Metadata_corrected.txt -c SampleID -o combined_fasta

3. Pick closed reference OTUs

picked_closed_reference_otus.py -r 97_otus.fasta -t 97_otu_taxonomy.txt

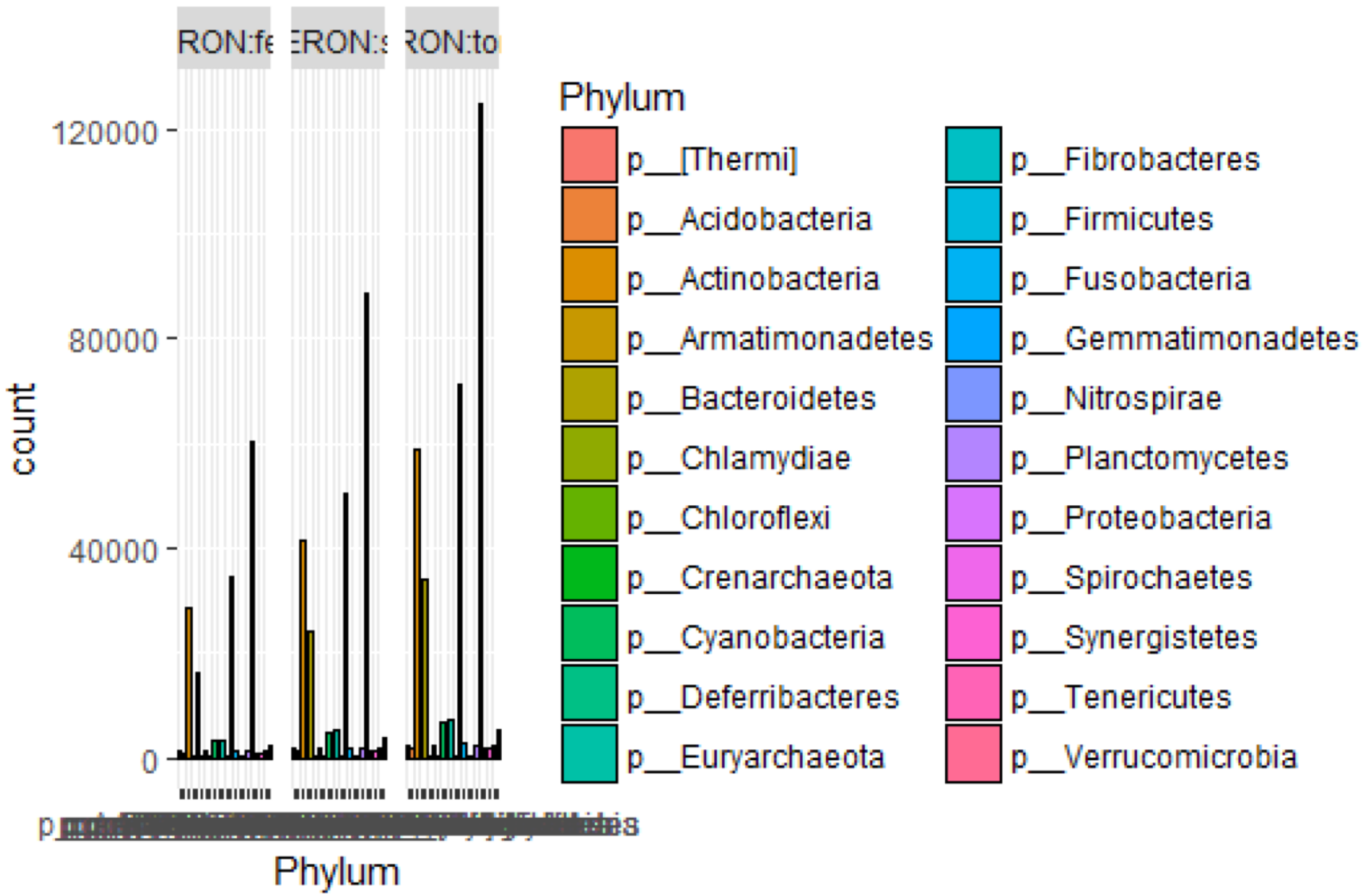
Qiime commands

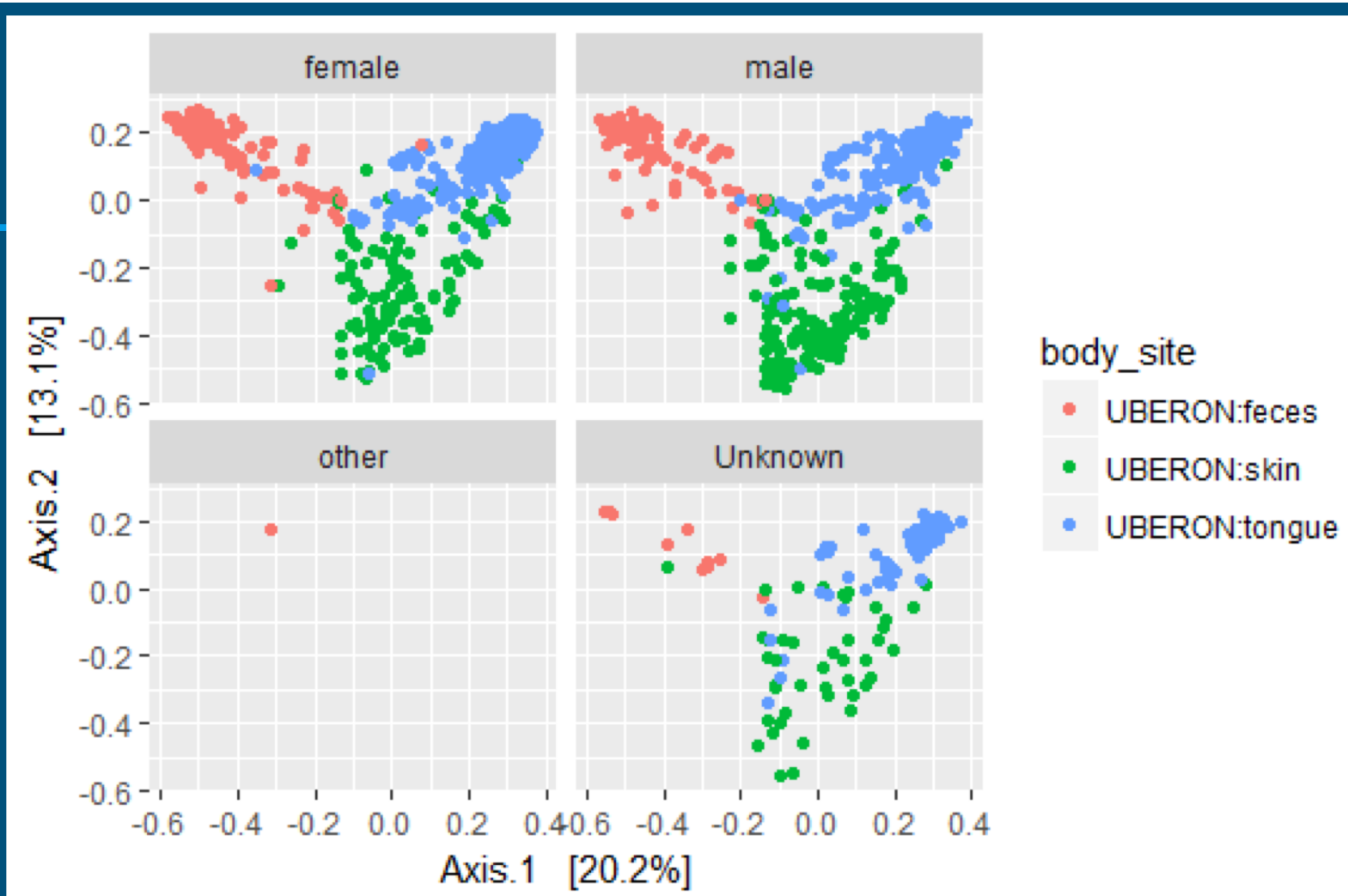
4. Summarize Taxa

summarize_taxa_through_plots.py -o taxa_summary -i picked_otus_default/otu_table.biom -m Metadata_corrected.txt

5. Convert biom (binary file) to json

biom convert -t table.txt -o table.from_txt_json.biom --table-type=”OTU table” --to-json


RON:fe
ERON:s
RON:to
120000
80000
40000
0
count
Phylum
Phylum
p__[Thermi]
p__Acidobacteria
p__Actinobacteria
p__Armatimonadetes
p__Bacteroidetes
p__Chlamydiae
p__Chloroflexi
p__Crenarchaeota
p__Cyanobacteria
p__Deferribacteres
p__Euryarchaeota
p__Fibrobacteres
p__Firmicutes
p__Fusobacteria
p__Gemmatimonadetes
p__Nitrospirae
p__Planctomycetes
p__Proteobacteria
p__Spirochaetes
p__Synergistetes
p__Tenericutes
p__Verrucomicrobia


female
male
other
Unknown
0.2
0.0
-0.2
-0.4
-0.6
Axis.2 [13.1%]
-0.6
-0.4
-0.2
0.0
0.2
0.4
Axis.1 [20.2%]
body_site
UBERON:feces
UBERON:skin
UBERON:tongue

PICRUSt Scripts

1. Normalized OTU table

normalized_by_copy_number.py -i picked_otus_default.biom -o normalized_otus.biom

2. Predict functions for metagenome

predict_metagenomes.py -i normalized_otus.biom -o metagenome_predictions.biom (-t cog to use COG database)

3. Get tab delimited file with NSTI values

perdict metagenomes.py -f -i normalized_otus.biom -o metagenome_predictions.tab -a nsti_per_sample.tab

KEGG and COG database

KEGG - Database used to map molecular datasets, especially large in genomics, transcriptomics, proteomics, and metabolomics for system functions. Based on speciation events and phylogenetics

Levels of Hierarchy

COG - database consisting of orthologous proteins found across multiple lineages which are likely to correspond to an ancient conserved domain

bodysite
UBERON.tongue
UBERON.skin
UBERON.feces
amino acid metabolism
biosynthesis of other secondary metabolites
carbohydrate metabolism
cell motility
cellular processes and signaling
energy metabolism
genetic information processing
glycan biosynthesis and metabolism
lipid metabolism
membrane transport
metabolism
metabolism of cofactors and vitamins
metabolism of terpenoids and polyketides
nucleotide metabolism
xenobiotics biodegradation and metabolism
pathway
-2.0
-2.5
-3.0
-3.5
-4.0
-4.5

Other PICRUSt Scripts