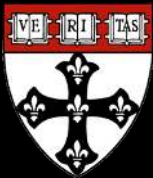




Amplicon functional profiling with PICRUSt

Curtis Huttenhower

01-19-15



Harvard School of Public Health
Department of Biostatistics



U. Oregon





The two big questions...

Who is there?
(taxonomic profiling)

What are they doing?
(functional profiling)

In marker
gene data



PICRUSt: Inferring community metagenomic potential from marker gene sequencing

With Rob Knight, Rob Beiko

If function is so important,
what about the
thousands of 16S-based
microbial community
taxonomic profiles?

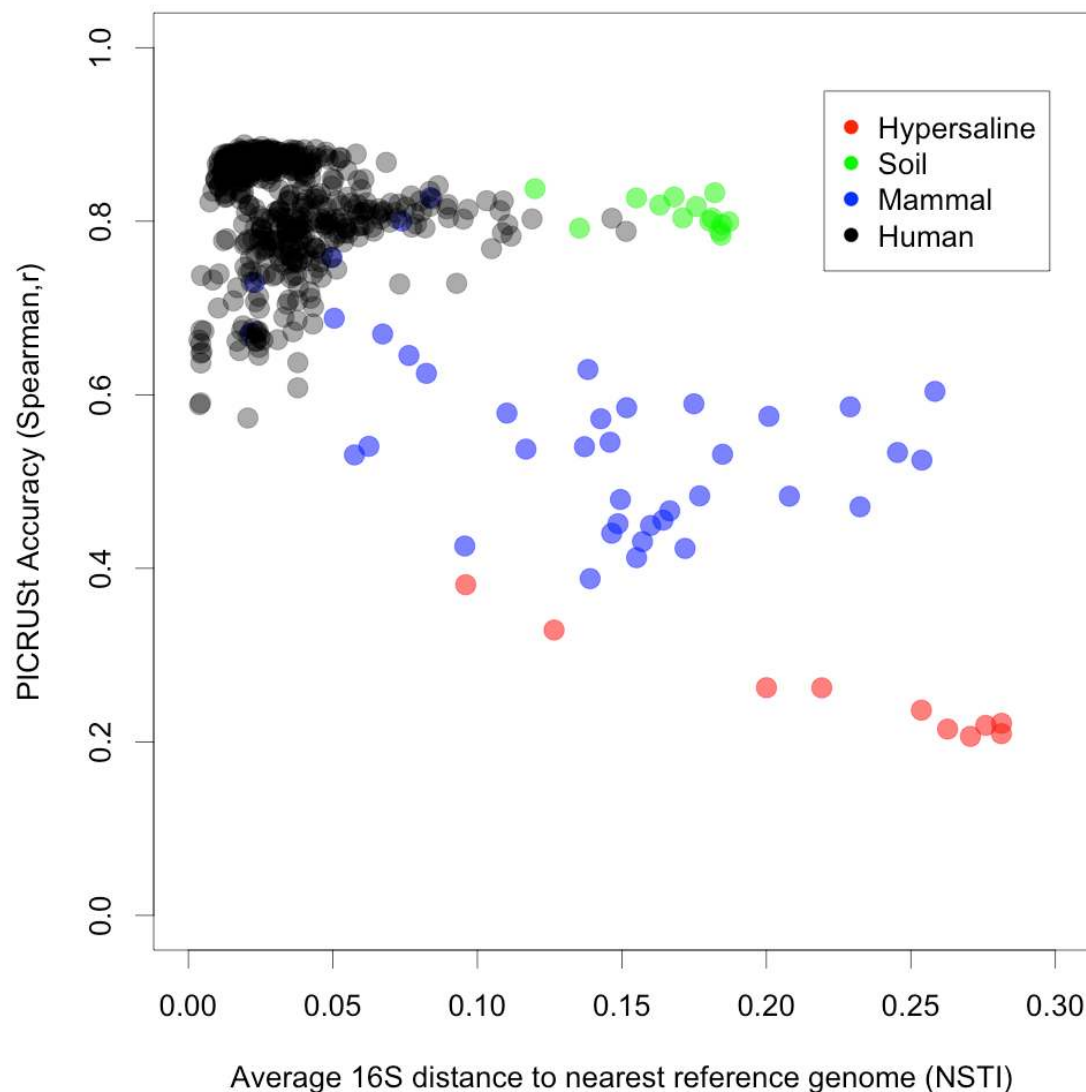
Pathways
and modules

Orthologous
gene families

Taxon
abundances



HUMAnN



Relative abundance



Setup notes reminder

- Slides with **green titles or text** include instructions not needed today, but useful for your own analyses
- Keep an eye out for **red warnings** of particular importance
- Command lines and program/file names appear in a **monospaced font**.
- Commands you should specifically copy/paste are in **monospaced bold blue**.



Installing PICRUST

- <http://picrust.github.io/picrust/install.html>
- Requires
 - Python (≥ 2.7 , easy)
 - PyCogent (<http://pycogent.org>)
 - BIOM (<http://biom-format.org>)
- And PICRUST!

```
3. chuttenhower@hutlab3:/n/huttenhower_lab_nobackup/tools/picrust-...  
[chuttenhower@hutlab3 picrust-1.0.0]$ ls  
doc          picrust          README.md      setup.py      tutorials  
LICENSE      picrust_test_data  scripts       tests  
[chuttenhower@hutlab3 picrust-1.0.0]$
```



Installing PICRUST

- <http://picrust.github.io/picrust/install.html>

PICRUST 1.0.0-dev documentation »

[index](#)

Installing PICRUST

Note

Most users will not need to install PICRUST, but can instead use the [online Galaxy version](#).

PICRUST is written in python, and has been tested on Mac OS X and Linux systems. To install PICRUST, first install all of the mandatory requirements, following the instructions on their respective websites. You should then download PICRUST. You have the choice of downloading either the release version (recommended for most users) or the development version (recommended for users who need access to the latest features, and are willing to tolerate some instability in the code). Next, you will download the large precalculated PICRUST files and place them in your picrust/data directory. Finally, you'll install PICRUST. Each of these steps are detailed below.

Step 1. Install Requirements

Follow the install instructions found on the website of each of the dependencies below to install PICRUST's dependencies.

Table Of Contents

Installing PICRUST

- [Step 1. Install Requirements](#)
- [Step 2. Download PICRUST](#)
 - [Release software](#)
 - [Development software](#)
- [Step 3. Download PICRUST's precalculated files](#)
- [Step 4. Install PICRUST](#)

This Page

[Show Source](#)

Quick search

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



Installing PICRUSt data

- http://picrust.github.io/picrust/picrust_precalculated_files.html
- Need to download the precomputed data used by PICRUSt separately
 - It's big! Saves you the trouble in the software itself

These files must be placed in your `picrust-1.0.0/picrust/data` directory before installing.

[PICRUSt 1.0.0-dev documentation](#) » [index](#)

PICRUSt's Precalculated Files

PICRUSt requires the downloading of precalculated files.

These downloaded files should be placed within your *picrust/data* directory BEFORE installation.

Download the set of files that correspond to the version of Greengenes that you used for OTU picking (see [Picking OTUs for use in PICRUSt](#))

The minimum set of files are:

1. 16S for copy number normalization (used in the script [normalize by copy number.py](#)).
2. Whatever type of function predictions you want returned by PICRUSt (e.g. KO or COG). This is used in the script [predict metagenomes.py](#).

Greengenes v13.5 (and IMG 4)

- [16S](#)
- [KO](#)
- [COG](#)
- [RFAM](#)

Table Of Contents

PICRUSt's Precalculated Files

- [Greengenes v13.5 \(and IMG 4\)](#)
- [Greengenes 18may2012](#)

This Page

[Show Source](#)

Quick search

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



Picking PICRUSt-compatible OTUs

- http://picrust.github.io/picrust/tutorials/otu_picking.html
- PICRUSt uses precomputed ancestral state reconstructions
 - OTUs in your data must match those used during precalculation
 - This means Greengenes, either 18may2012 or 13.5

[PICRUSt 1.0.0-dev documentation](#) » [index](#)

Picking OTUs for use in PICRUSt

Introduction

This document covers how to pick OTUs from marker gene data to use with PICRUSt. To do this, you'll use a 'closed-reference' OTU picking protocol where you search sequences against the GG reference OTUs at a specified percent identity, and discard any reads that don't hit that reference collection. The newest available reference collection can be found here:

- [gg_13_5_otus.tar.gz](#) ([download](#))

This tutorial assumes that you have QIIME installed. See the [QIIME website](#) for details on how to use QIIME. The quickest way to get started with QIIME is working on the Amazon Web Services cloud, and you can find instructions for [using QIIME on the cloud here](#).

Picking closed reference OTUs with QIIME

To pick 'closed reference' OTUs with QIIME for use in PICRUSt, you should begin with a demultiplexed fasta file in QIIME format, and the GG reference collection (see download link above).

Table Of Contents

- Picking OTUs for use in PICRUSt
 - Introduction
 - Picking closed reference OTUs with QIIME

This Page

[Show Source](#)

Quick search

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



Picking PICRUSt-compatible OTUs

- This means that you must use completely closed-reference OTU picking in QIIME

```
echo "pick_otus:enable_rev_strand_match True" >> $PWD/otu_picking_params_97.txt  
echo "pick_otus:similarity 0.97" >> $PWD/otu_picking_params_97.txt  
pick_closed_reference_otus.py -i $PWD/seqs.fna -o $PWD/ucrC97/ -p $PWD/otu_picking_params
```

- Produces an OTU table in which all features are Greengenes IDs:



Picking PICRUSt-compatible OTUs

```
3. chuttenhower@hutlab3:~/stamps (ssh)

# Constructed from biom file
#OTU ID 700098429    700016960    700095467    700101581    700034907    700097688
145207  0.25      0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
NZ_ABV001000045|642979334  0.0 1.2 23.6    2.2 8.8 6.0 4.4 0.2 0.2 119.4    7.0
462229  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
296363  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
296361  0.0 0.0 0.0 4.5 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
351348  0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 1.0 0.0 0.0 1.0
214079  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 89.0    0.0 0.0 0.0 16.4    0.0
580919  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
240104  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
NZ_AEKI01000004|649989960  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
348374  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
279506  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
448819  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
297701  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
386632  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
376175  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
NC_011601|643348548 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
NZ_ACIJ02000018|645951840  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
223033  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
189116  14.3333333333  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
545964  0.0 0.0 0.0 0.0 0.0 0.666666666667  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
128297  0.666666666667  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
hmp_otu_subset.tsv
```



Picking PICRUSt-compatible OTUs

hmp_otu_subset.tsv

Search in Sheet

Home Layout Tables Charts SmartArt Formulas Data Review

Edit Font Alignment Number Format Cells Themes

Fill Calibri (Body) 12 A A Paste Clear B I U Merge Conditional Formatting

A1 fx # Constructed from biom file

	A	B	C	D	E	F	G	H
1	# Constructed from biom file							
2	#OTU ID	700098429	700016960	700095467	700101581	700034907	700097688	7000985
3	145207	0.25	0	0	0	0	0	
4	NZ_ABVO01000045 642979334	0	1.2	23.6	2.2	8.8	6	
5	462229	0	0	0	0	0	0	
6	296363	0	0	0	0	0	0	
7	296361	0	0	0	4.5	0	0	
8	351348	0	0	0	0	0	1	
9	214079	0	0	0	0	0	0	
10	580919	0	0	0	0	0	0	
11	240104	0	0	0	0	0	0	
12	NZ_AEKI01000004 649989960	0	0	0	0	0	0	
13	348374	0	0	0	0	0	0	
14	279506	0	0	0	0	0	0	
15	448819	0	0	0	0	0	0	
16	297701	0	0	0	0	0	0	

hmp_otu_subset.tsv

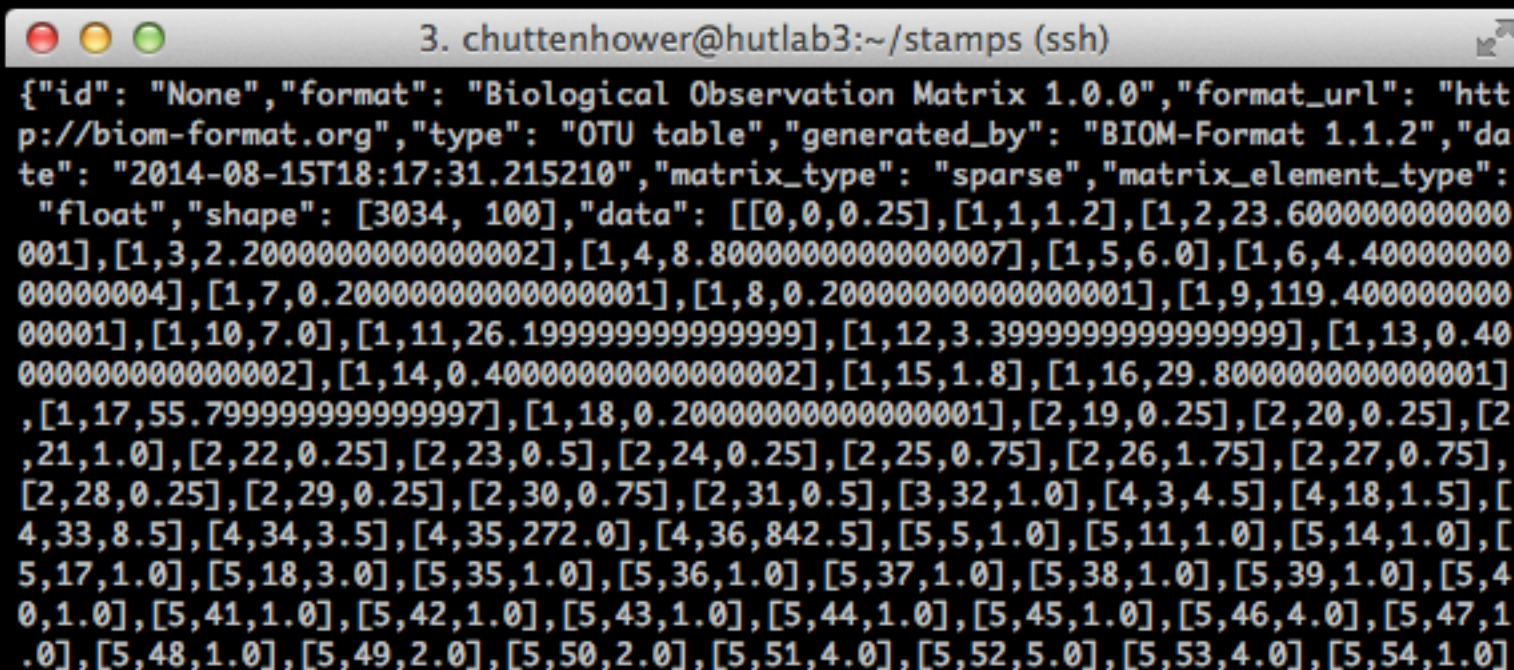
Normal View Ready Sum = 0



Munging BIOMs

- Great format, hard to read!
 - Easy to convert to/from TSV

```
convert_biom.py -i hmp_otu_subset.tsv  
-o hmp_otu_subset.biom  
--biom_table_type="otu table"
```



```
3. chuttenhower@hutlab3:~/stamps (ssh)  
{"id": "None", "format": "Biological Observation Matrix 1.0.0", "format_url": "http://biom-format.org", "type": "OTU table", "generated_by": "BIOM-Format 1.1.2", "date": "2014-08-15T18:17:31.215210", "matrix_type": "sparse", "matrix_element_type": "float", "shape": [3034, 100], "data": [[0,0,0.25],[1,1,1.2],[1,2,23.600000000000001],[1,3,2.2000000000000002],[1,4,8.8000000000000007],[1,5,6.0],[1,6,4.4000000000000004],[1,7,0.20000000000000001],[1,8,0.20000000000000001],[1,9,119.40000000000001],[1,10,7.0],[1,11,26.199999999999999],[1,12,3.3999999999999999],[1,13,0.40000000000000002],[1,14,0.40000000000000002],[1,15,1.8],[1,16,29.800000000000001],[1,17,55.799999999999997],[1,18,0.20000000000000001],[2,19,0.25],[2,20,0.25],[2,21,1.0],[2,22,0.25],[2,23,0.5],[2,24,0.25],[2,25,0.75],[2,26,1.75],[2,27,0.75],[2,28,0.25],[2,29,0.25],[2,30,0.75],[2,31,0.5],[3,32,1.0],[4,3,4.5],[4,18,1.5],[4,33,8.5],[4,34,3.5],[4,35,272.0],[4,36,842.5],[5,5,1.0],[5,11,1.0],[5,14,1.0],[5,17,1.0],[5,18,3.0],[5,35,1.0],[5,36,1.0],[5,37,1.0],[5,38,1.0],[5,39,1.0],[5,40,1.0],[5,41,1.0],[5,42,1.0],[5,43,1.0],[5,44,1.0],[5,45,1.0],[5,46,4.0],[5,47,1.0],[5,48,1.0],[5,49,2.0],[5,50,2.0],[5,51,4.0],[5,52,5.0],[5,53,4.0],[5,54,1.0]]
```



Everything I ever needed to know about PICRUSt I learned from this web site

- http://picrust.github.io/picrust/tutorials/metagenome_prediction.html

PICRUSt 1.0.0-dev documentation »

index

Metagenome Prediction Tutorial ¶

Introduction

This tutorial explains how to predict a microbial community metagenome using PICRUSt, based on 16S (or other marker gene) data as detailed in [Picking OTUs for use in PICRUSt](#).

BIOM Format

- Please note that PICRUSt by default uses the relatively new [biom](#) format for representing OTU tables and Gene tables (e.g. KOs by samples). This [has several benefits](#) including easier integration with other software (e.g. QIIME and others in the future) and allows embedding of extra metadata about both the samples and observations (OTUs/KOs).
- However, PICRUSt also allow users to input OTU tables and export PICRUSt predictions in tab-delimited format by using the '-f' option (see below).
- In addition, users can always convert to/from biom format to tab-delimited format using BIOM's built in [conversion script](#).

Requirements

1. You should have already installed PICRUSt ([Installing PICRUSt](#)).
2. A PICRUSt compatible OTU table ([Picking OTUs for use in PICRUSt](#)), such as the example file `tutorials/hmp_mock_16S.biom`.

Step 1: Normalize OTU Table

[normalize_by_copy_number.py](#) normalizes the OTU table by dividing each OTU by the known/predicted 16S copy number abundance.

Table Of Contents

Metagenome Prediction Tutorial

- Introduction
- Requirements
- Step 1: Normalize OTU Table
- Step 2: Predict Functions For Metagenome
- Step 3: Analysing Predicted Metagenomes

This Page

Show Source

Quick search

Go

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



Step 1: Normalize OTUs by 16S copy number

- We can get better predictions by dividing “raw” OTU counts by their expected 16S copy number

PLoS Comput Biol. 2012;8(10):e1002743. doi: 10.1371/journal.pcbi.1002743. Epub 2012 Oct 25.

Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance.

Kembel SW¹, Wu M, Eisen JA, Green JL.

Microbiome. 2014 Apr 7;2:11. doi: 10.1186/2049-2618-2-11. eCollection 2014.

CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction.

Angly FE¹, Dennis PG², Skarshewski A¹, Vanwonterghem I³, Hugenholtz P¹, Tyson GW³.

```
export PYTHONPATH=`pwd`/picrust-1.0.0
./picrust-1.0.0/scripts/normalize_by_copy_number.py
-i hmp_otu_subset.biom
-o hmp_otu_subset_normalized.biom
-g 18may2012
```



Step 1: Normalize OTUs by 16S copy number

```
convert_biom.py -i hmp_otu_subset_normalized.biom  
-o hmp_otu_subset_normalized.tsv -b
```

```
3. chuttenhower@hutlab3:~/stamps (ssh)  
# Constructed from biom file  
#OTU ID 700098429 700016960 700095467 700101581 700034907 700097688  
145207 0.0625 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
NZ_ABV001000045|642979334 0.0 0.24 4.72 0.44 1.76 1.2 0.88 0.04  
462229 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
296363 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
296361 0.0 0.0 0.0 2.25 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
351348 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 1.0 0.0 0.0 1.0  
214079 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 17.8 0.0 0.0 0.0 3.28 0.0  
580919 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
240104 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
NZ_AEKI01000004|649989960 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
348374 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
279506 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
448819 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
297701 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
386632 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
376175 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
NC_011601|643348548 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
NZ_ACIJ02000018|645951840 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
223033 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
189116 4.777777777777 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
545964 0.0 0.0 0.0 0.0 0.0 0.222222222222 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
128297 0.222222222222 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
hmp_otu_subset_normalized.tsv
```



Step 2: Predict metagenome contents

- Given a normalized OTU table, PICRUSt uses gene copy numbers associated with each Greengenes tree tip to multiply and infer community gene copy numbers

```
./picrust-1.0.0/scripts/predict_metagenomes.py  
-i hmp_otu_subset_normalized.biom  
-o hmp_ko_subset.tsv -f -g 18may2012
```

```
3. chuttenhower@hutlab3:~/stamps (ssh)  
# Constructed from biom file  
#OTU ID 700098429 700016960 700095467 700101581 700034907 700097688  
K00001 494.0 198.0 418.0 397.0 403.0 329.0 416.0 126.0 253.0  
K00002 16.0 17.0 6.0 32.0 8.0 5.0 3.0 8.0 185.0 15.0 38.0 3.0  
K00003 469.0 127.0 413.0 309.0 242.0 314.0 341.0 139.0 208.0  
K00004 0.0 0.0 0.0 0.0 0.0 0.0 0.0 15.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0  
K00005 301.0 99.0 231.0 160.0 157.0 164.0 215.0 42.0 25.0  
K00007 0.0 0.0 0.0 0.0 0.0 0.0 0.0 2.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0  
K00008 517.0 83.0 345.0 250.0 236.0 295.0 354.0 16.0 102.0  
K00009 10.0 2.0 19.0 13.0 11.0 9.0 54.0 4.0 20.0 9.0 30.0  
K00010 6.0 4.0 17.0 25.0 13.0 16.0 100.0 63.0 136.0 12.0  
K00011 26.0 0.0 10.0 4.0 2.0 8.0 1.0 0.0 0.0 2.0 10.0 1.0 4.0 0.0 2.0  
K00012 318.0 166.0 322.0 302.0 420.0 271.0 371.0 115.0 200.0  
K00013 465.0 162.0 397.0 302.0 331.0 270.0 357.0 82.0 191.0  
K00014 661.0 164.0 541.0 402.0 427.0 412.0 520.0 198.0 256.0  
K00015 0.0 4.0 2.0 26.0 22.0 0.0 0.0 4.0 1.0 0.0 0.0 0.0 0.0 1.0 0.0  
K00016 294.0 49.0 261.0 187.0 173.0 186.0 288.0 116.0 357.0  
K00018 94.0 56.0 67.0 88.0 127.0 77.0 140.0 72.0 1.0 107.  
K00019 1.0 4.0 3.0 27.0 22.0 1.0 0.0 5.0 47.0 0.0 0.0 0.0 0.0 0.0 1.0  
K00020 271.0 103.0 182.0 188.0 162.0 129.0 110.0 46.0 113.0  
K00021 0.0 0.0 0.0 0.0 0.0 0.0 0.0 19.0 8.0 0.0 0.0 0.0 0.0 63.0 0.0 0.0  
K00022 2.0 4.0 4.0 27.0 22.0 2.0 0.0 10.0 50.0 1.0 0.0 0.0 0.0 1.0  
K00023 1.0 4.0 3.0 27.0 22.0 1.0 0.0 2.0 3.0 0.0 0.0 0.0 0.0 1.0 0.0  
K00024 84.0 113.0 77.0 138.0 171.0 98.0 185.0 52.0 103.0  
hmp_ko_subset.tsv
```



Step 2: Predict metagenome contents

hmp_ko_subset.tsv

Search in Sheet

Home Layout Tables Charts SmartArt Formulas Data Review

Edit Fill Paste Clear Font Calibri (Body) 12 A- A- Alignment Wrap Text Merge Number General Conditional Formatting Format Normal Good Bad Neutral Cells Insert Delete Format Themes Aa

A1 fx # Constructed from biom file

	A	B	C	D	E	F	G	H	I	J
1	# Constructed from biom file									
2	#OTU ID	700098429	700016960	700095467	700101581	700034907	700097688	700098542	700098554	700101242
3	K00001	494	198	418	397	403	329	416	126	253
4	K00002	16	17	6	32	8	5	3	8	185
5	K00003	469	127	413	309	242	314	341	139	208
6	K00004	0	0	0	0	0	0	0	15	0
7	K00005	301	99	231	160	157	164	215	42	25
8	K00007	0	0	0	0	0	0	0	2	1
9	K00008	517	83	345	250	236	295	354	16	102
10	K00009	10	2	19	13	11	9	54	4	20
11	K00010	6	4	17	25	13	16	100	63	136
12	K00011	26	0	10	4	2	8	1	0	0
13	K00012	318	166	322	302	420	271	371	115	200
14	K00013	465	162	397	302	331	270	357	82	191
15	K00014	661	164	541	402	427	412	520	198	256
16	K00015	0	4	2	26	22	0	0	4	1

hmp_ko_subset.tsv

Normal View Ready Sum=0



PICRUSt optional behavior

- You can associate predicted gene abundances with the organism that contains them
 - `metagenome_contributions.py`
- You can show the confidence intervals of all predictions
 - `--with_confidence` argument to `predict_metagenomes.py`
- You can summarize per-gene predictions to per-pathway predictions
 - `categorize_by_function.py`
- You can input a PICRUSt output BIOM file into HUMAnN to reconstruct pathways



PICRUSt in Galaxy

- <http://huttenhower.sph.harvard.edu/picrust>
 - First get data!
 - scp from `~/workshop_data/metagenomics/biobakery/data/hmp_otu_subset.tsv`

The screenshot shows the Galaxy web interface with the PICRUSt tool selected. The interface includes a top navigation bar with 'Galaxy / Huttenhower Lab' and various menu items. On the left, a 'Tools' sidebar lists 'HUTTENHOWER LAB MODULES' and 'LOAD DATA MODULE'. The 'Get Data' option is circled in red. The main panel shows the 'Upload File (version 1.1.4)' form. The 'File Format' dropdown is set to 'picrust' (circled in red). The 'File' section shows 'Choose File' and 'hmp_otu_subset.tsv' (circled in red). The 'Execute' button at the bottom is also circled in red. The right sidebar shows the 'History' panel with an 'Unnamed history' and a message indicating it is empty.



PICRUSt in Galaxy

- Next normalize by copy number

Galaxy / Huttenhower Lab Analyze Data Workflow Shared Data Visualization Help User Using 0%

Tools

search tools

HUTTENHOWER LAB MODULES

- [LEfSe](#)
- [MetaPhlAn](#)
- [GraPhlAn](#)
- [microPITA](#)
- [MaAsLin](#)
- [PICRUSt](#)
- [Normalize By Copy Number](#)
- [Predict Metagenome](#)
- [Categorize by function](#)

LOAD DATA MODULE

- [Get Data](#)
- [Upload File from your computer](#)

Normalize By Copy Number (version 1.0.0)

Input file – Please upload using the standard Galaxy 'Get Data' – Be sure to select file format 'picrust' for the input file:

1: hmp_otu_subset.tsv

Format of the input file (default is Legacy QIIME format – tab delimited):

Legacy QIIME format (tab-delimited)

GreenGen Version (used to generate your OTU table):

GG 18may2012

Execute

PICRUSt: Phylogenetic Investigation of Communities by Reconstruction of Unobserved States

The PICRUSt project aims to support prediction of the unobserved character states in a community of organisms from phylogenetic information about the organisms in that

History

Unnamed history

1.2 MB

1: hmp_otu_subset.tsv



PICRUSt in Galaxy

- Then predict your metagenome

Galaxy / Huttenhower Lab Analyze Data Workflow Shared Data Visualization Help User Using 0%

Tools

search tools

HUTTENHOWER LAB MODULES

- [LEfSe](#)
- [MetaPhlAn](#)
- [GraPhlAn](#)
- [microPITA](#)
- [MaAsLin](#)
- [PICRUSt](#)
- [Normalize By Copy Number](#)
- [Predict Metagenome](#)
- [Categorize by function](#)

LOAD DATA MODULE

Get Data

[Upload File](#) from your computer

Predict Metagenome (version 1.0.0)

Input file:

2: Normalize By Copy Number on data 1

Format of the input file (default is BIOM):

☒ Legacy QIIME format (tab-delimited)

GreenGenes Version (used to generate your OTU table):

GG 18may2012

Type of functional predictions:

KEGG Orthologs

Type of output:

Metagenome Predictions

Execute

History

Unnamed history

1.2 MB

2: Normalize By Copy Number on data 1

1: hmp_otu_subset.tsv

Run PICRUSt: Predict Metagenome



PICRUSt in Galaxy

- You can download the resulting gene table...

The screenshot displays the Galaxy web interface with the Huttenhower Lab logo. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Help, and User. The left sidebar lists Huttenhower Lab Modules (LEfSe, MetaPhlAn, GraPhlAn, microPITA, MaAsLin, PICRUSt) and Default Galaxy Modules (Convert Formats, FASTA manipulation, General Galaxy tools). The main panel shows a green notification box stating: "A job has been successfully added to the queue – resulting in the following dataset: 3: Predict Metagenome on data 2. You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered." The right sidebar shows the History pane with a list of jobs. The job "3: Predict Metagenome on data 2" is highlighted with a red circle, and its download icon is also circled in red. Below the notification, red text reads: "You should change the extension to .tsv after downloading".

Galaxy / Huttenhower Lab

Tools

search tools

HUTTENHOWER LAB MODULES

LEfSe

MetaPhlAn

GraPhlAn

microPITA

MaAsLin

PICRUSt

Normalize By Copy Number

Predict Metagenome

Categorize by function

LOAD DATA MODULE

Get Data

Upload File from your computer

DEFAULT GALAXY MODULES

Convert Formats

FASTA manipulation

General Galaxy tools

A job has been successfully added to the queue – resulting in the following dataset:

3: Predict Metagenome on data 2

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

History

Unnamed history

5.9 MB

3: Predict Metagenome on data 2

6,885 lines

format: picrustp, database: ?

Download icon circled in red

Constructed from biom file

#OTU	ID	700098429	700016960
1	700023556	700038583	
33130	700038021	700023551	
700107197	700111034	700111	

2: Normalize By Copy Number on data 1

1: hmp_otu_subset.tsv

You should change the extension to .tsv after downloading



PICRUSt in Galaxy

Galaxy3-(Predict_Metagenome_on_data_2).tsv

Search in Sheet

Home Layout Tables Charts SmartArt Formulas Data Review

Edit Font Alignment Number Format Cells Themes

Paste Clear Fill Calibri (Body) 12 A+ A- abc Wrap Text General Conditional Formatting Normal Bad Good Neutral Insert Delete Format Themes Aa

A1 # Constructed from biom file

	A	B	C	D	E	F	G	H	I	J
1	# Constructed from biom file									
2	#OTU ID	700098429	700016960	700095467	700101581	700034907	700097688	700098542	700098554	700101242
3	K00001	494	198	418	397	403	329	416	126	253
4	K00002	16	17	6	32	8	5	3	8	185
5	K00003	469	127	413	309	242	314	341	139	208
6	K00004	0	0	0	0	0	0	0	15	0
7	K00005	301	99	231	160	157	164	215	42	25
8	K00007	0	0	0	0	0	0	0	2	1
9	K00008	517	83	345	250	236	295	354	16	102
10	K00009	10	2	19	13	11	9	54	4	20
11	K00010	6	4	17	25	13	16	100	63	136
12	K00011	26	0	10	4	2	8	1	0	0
13	K00012	318	166	322	302	420	271	371	115	200
14	K00013	465	162	397	302	331	270	357	82	191
15	K00014	661	164	541	402	427	412	520	198	256
16	K00015	0	4	2	26	22	0	0	4	1

Galaxy3-(Predict_Metagenome_on_ +)

Normal View Ready Sum=0



PICRUSt in Galaxy

- ...or you can summarize genes to pathways
 - Note that this only works on BIOM files

The screenshot displays the Galaxy web interface with the 'Predict Metagenome (version 1.0.0)' tool selected. The interface includes a top navigation bar with 'Galaxy / Huttenhower Lab' and various menu items. On the left, a 'Tools' sidebar lists Huttenhower Lab modules, with 'Predict Metagenome' circled in red. The main tool panel shows the following configuration:

- Input file:** 2: Normalize By Copy Number on data 1
- Format of the input file (default is BIOM):** ☐ Legacy QIIME format (tab-delimited) (circled in red)
- GreenGenes Version (used to generate your OTU table):** GG 18may2012 (circled in red)
- Type of functional predictions:** KEGG Orthologs
- Type of output:** Metagenome Predictions
- Execute button:** (circled in red)

On the right, the 'History' panel shows a list of jobs, including '3: Predict Metagenome on data 2', '2: Normalize By Copy Number on data 1', and '1: hmp_otu_subset.tsv'.



PICRUSt in Galaxy

- ...continuing where we left off...

The screenshot displays the Galaxy web interface with the 'Categorize by function (version 1.0.0)' tool selected. The interface includes a top navigation bar with 'Galaxy / Huttenhower Lab' and various menu items. On the left, a 'Tools' sidebar lists 'HUTTENHOWER LAB MODULES' including LEfSe, MetaPhlAn, GraPhlAn, microPITA, MaAsLin, and PICRUSt. The 'Categorize by function' tool is highlighted in the sidebar. The main tool configuration area shows the 'Input file' set to '5: Predict Metagenome on data 2', the 'KEGG Pathway Hierarchy Level' set to '3', and the 'Type of output' set to 'Legacy QIIME format (tab-delimited)'. The 'Execute' button is prominently displayed. Below the configuration, a description of the tool's function is provided, followed by an 'Input file' section and a link for more information. On the right, a 'History' panel shows a list of recent jobs, including '5: Predict Metagenome on data 2', '3: Predict Metagenome on data 2', '2: Normalize By Copy Number on data 1', and '1: hmp otu_subset.tsv'.

Galaxy / Huttenhower Lab Analyze Data Workflow Shared Data Visualization Help User Using 0%

Tools

search tools

HUTTENHOWER LAB MODULES

- [LEfSe](#)
- [MetaPhlAn](#)
- [GraPhlAn](#)
- [microPITA](#)
- [MaAsLin](#)
- [PICRUSt](#)
- [Normalize By Copy Number](#)
- [Predict Metagenome](#)
- [Categorize by function](#)

LOAD DATA MODULE

- [Get Data](#)
- [Upload File from your computer](#)

DEFAULT GALAXY MODULES

- [Convert Formats](#)
- [FASTA manipulation](#)
- [General Galaxy tools](#)

Categorize by function (version 1.0.0)

Input file:

5: Predict Metagenome on data 2

KEGG Pathway Hierarchy Level:

3

Type of output:

Legacy QIIME format (tab-delimited)

Execute

Run PICRUSt: Categorize by Function

This module collapses hierarchical data to a specified level for PICRUSt predictions. For instance, often it is useful to examine KEGG results from a higher level within the pathway hierarchy. Many genes are sometimes involved in multiple pathways, and in these circumstances (also know as a one-to-many relationship), the gene is counted for each pathway.

Input file

Output file from the predict process

For more information please visit:

<http://picrust.github.com/>

History

Unnamed history
11.7 MB

- 5: Predict Metagenome on data 2
- 3: Predict Metagenome on data 2
- 2: Normalize By Copy Number on data 1
- 1: hmp otu_subset.tsv



PICRUSt in Galaxy

- ...and then download them.

Galaxy / Huttenhower Lab Analyze Data Workflow Shared Data Visualization Help User Using 0%

Tools

search tools

HUTTENHOWER LAB MODULES

- [LEfSe](#)
- [MetaPhlAn](#)
- [GraPhlAn](#)
- [microPITA](#)
- [MaAsLin](#)
- [PICRUSt](#)
 - [Normalize By Copy Number](#)
 - [Predict Metagenome](#)
 - [Categorize by function](#)

LOAD DATA MODULE

- [Get Data](#)
 - [Upload File from your computer](#)

DEFAULT GALAXY MODULES

- [Convert Formats](#)
- [FASTA manipulation](#)
- [General Galaxy tools](#)

History

Unnamed history 11.9 MB

6: Categorize by function on data 5 333 lines Format: picrustc, database: ?

Download icon circled in red.

5: Predict Metagenome on data 2

3: Predict Metagenome on data 2

2: Normalize By Copy Number on data 1

1: hmp otu subset.tsv

Message: A job has been successfully added to the queue – resulting in the following dataset:
6: Categorize by function on data 5
You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

You should change the extension to .tsv after downloading



PICRUSt in Galaxy

Galaxy6--(Categorize_by_function_on_data_5).tsv

Search in Sheet

Home Layout Tables Charts SmartArt Formulas Data Review

Edit Fill Font Calibri (Body) 12 Alignment Wrap Text General Number Format Normal Bad Good Neutral Cells Insert Delete Format Themes

A1 fx # Constructed from biom file

	A	B	C	D	E	F
1	# Constructed from biom file					
2	#OTU ID	700098429	700016960	700095467	700101581	700034907
3	1,1,1-Trichloro-2,2-bis(4-chlorophenyl)ethane (DDT) degradation	14	57	2	17	15
4	ABC transporters	34506	12592	29474	22839	21693
5	Adherens junction	0	0	0	0	0
6	Adipocytokine signaling pathway	370	498	370	559	659
7	African trypanosomiasis	0	4	1	26	23
8	Alanine, aspartate and glutamate metabolism	13023	5274	11483	9064	10054
9	Aldosterone-regulated sodium reabsorption	0	0	0	0	0
10	Alzheimer's disease	497	209	464	508	478
11	Amino acid metabolism	2808	811	2725	2075	2060
12	Amino acid related enzymes	15736	6334	14053	11441	11849
13	Amino sugar and nucleotide sugar metabolism	15042	5421	13673	10408	11842
14	Aminoacyl-tRNA biosynthesis	12618	4779	11424	9460	9125
15	Aminobenzoate degradation	1711	849	1633	2040	1710
16	Amoebiasis	9	19	6	9	10

Galaxy6--(Categorize_by_function_ +

Normal View Ready Sum=0



Thanks!



<http://huttenhower.sph.harvard.edu>



Alex Kostic



Levi Waldron



Xochitl Morgan



Tim Tickle



Daniela Boernigen



Lauren McIver



 Dirk Gevers



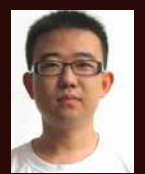
George Weingart



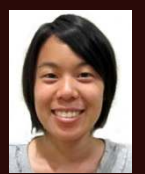
Emma Schwager



Eric Franzosa



Boyu Ren



Tiffany Hsu



Ali Rahnava



 Ramnik Xavier

Human Microbiome Project 2

Lita Procter	Bruce Birren
Jon Braun	Chad Nusbaum
Dermot McGovern	Clary Clish
Subra Kugathasan	Joe Petrosino
Ted Denson	Thad Stappenbeck
Janet Jansson	

 **BROAD INSTITUTE**



Ayshwarya Subramanian



Jim Kaminski



Regina Joice



Koji Yasuda



Kevin Oh



Galeb Abu-Ali



 Wendy Garrett

Human Microbiome Project

Jane Peterson	Karen Nelson
Sarah Highlander	George Weinstock
Barbara Methe	Owen White



Nicola Segata



Rob Knight



Afrah Shafquat



Randall Schwager



Chengwei Luo



Keith Bayer



Moran Yassour



Alexandra Sirota



Jesse Zaneveld



Morgan Langille



Rob Beiko

