



Jeremy
Wilkinson

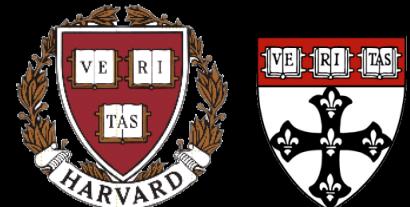


Lauren
McIver



The bioBakery microbial community analysis environment

Curtis Huttenhower



Harvard T.H. Chan School of Public Health
Department of Biostatistics

07-31-19





The bioBakery: a software environment for integrative microbiome analyses

- Environment for meta'ome analysis
 - Shotgun metagenomes/transcriptomes
 - Taxonomic and functional profiling
 - Experimental design, statistical analysis
- Pre-built one-click environment to run:
 - On your laptop graphically
 - On a server remotely
 - On the cloud (Amazon / Google)





Launching the bioBakery on Google Cloud

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



Lauren
McIver

The screenshot shows a web browser window with the URL <http://huttenhower.sph.harvard.edu/biobakery> highlighted by a red oval. The page content includes:

- Huttenhower Lab Tools**: Welcome message and bioBakery logo.
- Composition Analysis**: Tools: HUMAnN, MetaPhlAn, PhyloPhlAn, PICRUSt, ShortBRED, PPANINI, StrainPhlAn.
- Statistical Analysis**: Tools: HAIIA, ARepA, CCREPE, LEfSe, MaAsLin, microPITA, SparseDOSSA.



Launching the bioBakery on Google Cloud

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

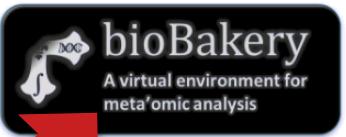


Lauren
McIver

Huttenhower Lab Tools

Welcome to the official Huttenhower Tutorials wiki.

We now support [bioBakery](#), a virtual environment platform that provides Huttenhower tools (already installed!). Please click on the button below for more information:



The bioBakery logo features a stylized DNA helix icon above the word "bioBakery" and the subtitle "A virtual environment for meta'omic analysis". A red arrow points to the logo.

This page contains tutorials for Huttenhower tools, illustrating through three main categories as shown below. Click on the tabs to learn more.

bioBakery

Quick Start

1. Install VirtualBox
2. Install Vagrant
3. Download and unpack [bioBakery](#)
4. Run bioBakery by double clicking (or executing from the command line) the file for your operating system
 - Linux and Mac OS: `start_biobakery.command`
 - Windows: `start_biobakery.bat`

Please note the first time you run bioBakery it will download and extract the box (~6 GB in size). This can take a bit of time depending on your internet connection. When you run bioBakery in the future, it will start up much faster (~30 seconds).

You can also run bioBakery in Google Cloud or you can install individual bioBakery tools.

Please see the sections and links that follow for detailed information about bioBakery.

Overview

bioBakery is an easy to use, virtual environment that provides a platform for users to have access to the Huttenhower tools without having to install the tool suite directly on their personal machines. This documentation provides in depth instructions on all topics related to bioBakery, from basic setup and advanced configuration.

We provide support for bioBakery users. Please join the [bioBakery Google Group](#) and feel free to post questions directly to the group or by emailing biobakery-users@googlegroups.com.

Workshop Usage

Basic Usage

- Launch bioBakery on your personal computer
- Use bioBakery online with Google Cloud

Workshop Usage

- Connect to Google Cloud instances
- Run tutorials

Advanced configuration

- Interacting with the VM from the command line
- Sharing data with the VM
- Building a new Google Cloud instance

Internal documentation

- Adding new tools
- Packaging new image
- Releasing new image

Welcome to the bioBakery Workshop!

Your instructor will provide you with the information needed to connect to the bioBakery instance via:

1. The connection method ([Web Browser](#), [Virtual Machine](#), or [SSH](#))
2. The name or IP address of your instance
3. The password (and user name if required)

Please follow the instructions under the connection method above for your workshop to get started in bioBakery Google Cloud.

Web Browser

4



Launching the bioBakery on Google Cloud

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



Lauren
McIver

Option 1: Web Browser Connection

The web browser based connection will provide you with desktop access to the bioBakery Google Cloud instance. It will also provide you with a method to download and upload files to your bioBakery instance.

How to Connect

Follow the instructions below to connect to your instance through your web browser.

Step 1: Log in to the bioBakery Guacamole Server

1. Go to the <http://huttenhower.sph.harvard.edu/guacamole>
2. Login with the Guacamole username and password provided by your instructor

The Guacamole Server log in screen is shown below.

GUACAMOLE 0.9.9

Username

Password

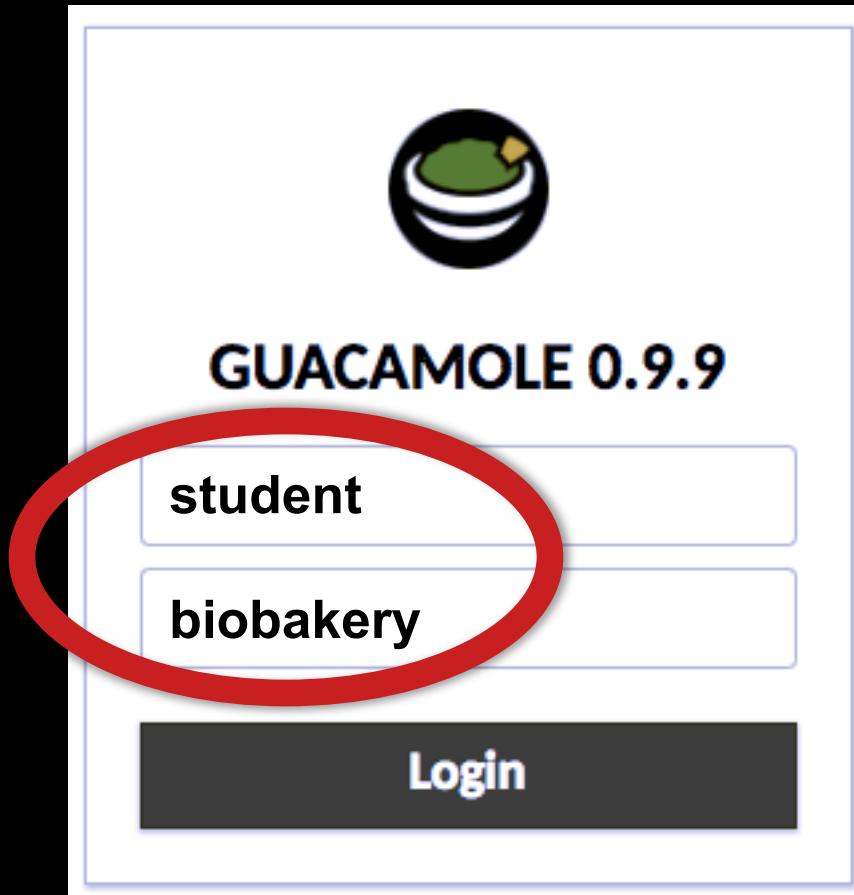
Login





Launching the bioBakery on Google Cloud

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





Launching the bioBakery on Google Cloud

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



RECENT CONNECTIONS student

No recent connections.

ALL CONNECTIONS

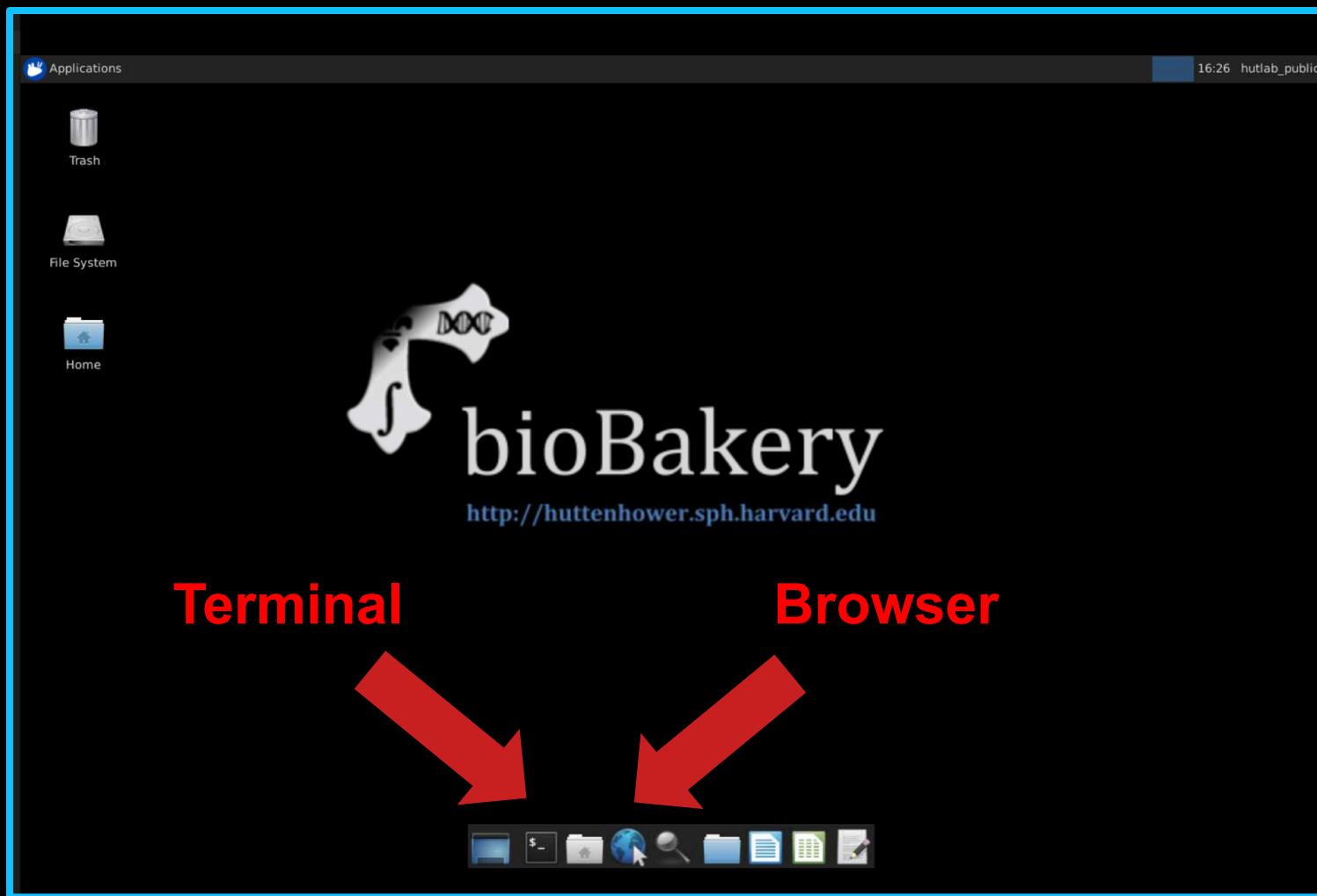
- biobakery-1
- biobakery-2

- In the instance list, use only your assigned instance.
- Connect to only that specific instance!
 - Don't shut it down, can just close the connection.
- Make sure you can launch a terminal and run a command.
 - e.g.: `echo hello world`
- Try to upload/download a file.



Google Cloud and Guacamole

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





Guacamole: copy and paste

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

Option 1: Web Browser Connection

The web browser based connection will provide you with desktop access to the bioBakery Google Cloud instance. It will also provide you with a method to download and upload files to your bioBakery instance.

How to Connect

Follow the instructions below to connect to your instance through your web browser.

Step 1: Log in to the bioBakery Guacamole Server

1. Go to the <http://huttenhower.sph.harvard.edu/guacamole>
2. Login with the Guacamole username and password provided by your instructor

The Guacamole Server log in screen is shown below.

Clipboard

- Highlight text
- Right click & copy
- Ctrl+Alt (Command)+Shift to bring up menu
- Copy the text from the box to anywhere else



Guacamole: copy and paste

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

The screenshot shows a Linux desktop environment with a terminal window and a web browser window. The terminal window displays a log of bioBakery tool execution, including commands like 'cs_relab', 's_ecs', and 'Run Finished'. The web browser window shows the bioBakery tools for meta'omic profiling wiki, featuring a sidebar with links for Source, Commits, Branches, Pull requests, Pipelines, Deployments, Issues, Wiki, and Downloads. A large red arrow points from the terminal window down towards the bottom of the screen, indicating the flow of data from the terminal to the web browser.

Not Secure | 35.199.34.159/guacamole/#/client/MTI5AGMAbXlzcWw=

W Wikipedia Ally BSC BSC MBTA PubMed BST281 (2018) Other Bookmarks

18:51 hutlab_public

Applications biobakery / biobaker... Terminal

File Edit View Terminal Tabs Help

(Jul 30 15:36:25) [69/73 - 94.52%] **Started**
cs_relab
(Jul 30 15:36:26) [70/73 - 95.89%] **Completed**
cs_relab
(Jul 30 15:36:26) [70/73 - 95.89%] **Ready**
s_ecs
(Jul 30 15:36:26) [70/73 - 95.89%] **Started**
s_ecs
(Jul 30 15:36:26) [71/73 - 97.26%] **Completed**
s_ecs
(Jul 30 15:36:26) [71/73 - 97.26%] **Ready**
_counts
(Jul 30 15:36:26) [71/73 - 97.26%] **Started**
_counts
(Jul 30 15:36:26) [72/73 - 98.63%] **Completed**
counts
(Jul 30 15:36:26) [72/73 - 98.63%] **Ready**
cs
(Jul 30 15:36:26) [72/73 - 98.63%] **Started**
cs
(Jul 30 15:36:26) [73/73 - 100.00%] **Completed**
cs
Run Finished
hutlab_public@biobakery-stamps2019-instructor-hlk

bioBakery tools for meta'omic profiling

Welcome to the bioBakery tools and tutorials wiki, which provides software, documentation, and tutorial for methods for microbial community profiling developed by the Huttenhower lab. Most tools are supported both as individual software packages (typically Python or R) and within the bioBakery virtual image, a pre-built platform that provides meta'omic analysis tools already installed with dependencies and configuration. For the integrated bioBakery virtual environment, please click on the button below for more information:

bioBakery
A meta'omic analysis environment

If you use bioBakery in your work, please cite the paper:

Mclver LJ, Abu-Ali G, Franzosa EA, Schwager R, Morgan XC, Waldron L, Segata N, Huttenhower C. bioBakery: a meta'omic analysis environment. Bioinformatics. 2018 Apr 1;34(7):1235-1237. PMID: 29194469

Microbial community profiling

This set of methods generally provide reference-based profiles of microbial community features, e.g. taxonomic abundances (MetaPhiAn) or functional profiles (genes and/or pathways, HUMANn). They apply broadly to sequence-based data (metagenomes and metatranscriptomes), with some methods applying to other types of culture-independent molecular data. Please click individual links for detailed tutorials:

HUMANn MetaPhiAn PhyloPhiAn

OR copy-paste in the Linux VM from the web browser in the web browser.



Guacamole: uploading files

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

- **Drag & drop** the file into the **desktop of your instance**
- The uploaded file will be located in your \$HOME directory, which is **/home/hutlab_public/**
- A file transfer status box will be displayed in the lower right-hand corner of the screen

FILE TRANSFERS	Clear
13530241_SF05.fasta.gz	67.8 KB
13530241_SF06.fasta.gz	59.0 KB
19272639_SF05.fasta.gz	75.3 KB



Guacamole: downloading files

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

- Type **Ctrl+Alt+Shift** to bring up the menu
- Click on the **box under Devices**
- Double click on the folders to **get to the folder containing the file(s)** on your instance you would like to download
 - E.g. `/home/hutlab_public/`
- **Double click on the file(s)** you would like to download
- A file transfer status box will appear
- Click on the **links in the file transfer status box** to download the files to the Downloads folder on your computer



Guacamole: logging off

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

The image consists of three vertically stacked screenshots of a Guacamole interface, each with a large red arrow pointing to the right.

- Screenshot 1:** Shows the top navigation bar with "BIOBAKERY-1" and "student". A red arrow points to the "student" dropdown menu.
- Screenshot 2:** Shows the dropdown menu open, with "Disconnect" highlighted in red. A red arrow points to the "Disconnect" option.
- Screenshot 3:** Shows a "DISCONNECTED" dialog box with the message "You have been disconnected." and three buttons: "Home", "Reconnect", and "Logout". A red arrow points to the "Logout" button.

Logout

- Ctrl+Alt+Shift to bring up menu



The bioBakery: Methods for integrative meta'omic analysis

1. Upstream microbial gene and genome curation.
2. Microbial community bioinformatics
(e.g. taxonomic and functional profiling of meta'omes).
3. Microbial community statistics and association discovery.
4. Reproducible workflows & infrastructure.

Reproducible Workflows & Infrastructure

Microbial Genomics

Meta'omic Profiling

Association Discovery

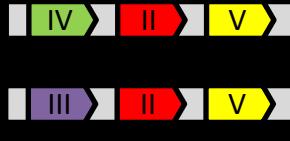


The bioBakery:

1. Microbial 'omics and reference sequences

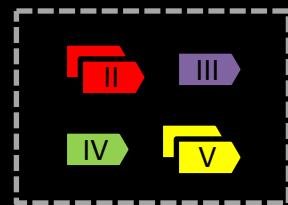
RepoPhlAn

collects & organizes
microbial genomes



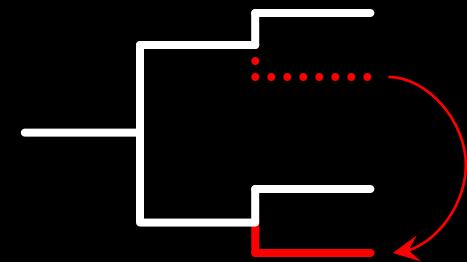
ChocoPhlAn

constructs & curates
microbial pangenomes



PhyloPhlAn

assigns or corrects
microbial taxonomy



Reproducible Workflows & Infrastructure

Microbial Genomics

Meta'omic Profiling

Association Discovery

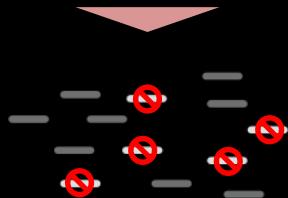


The bioBakery:

2. Microbial community bioinformatics

KneadData

quality controls
meta'omic seq.



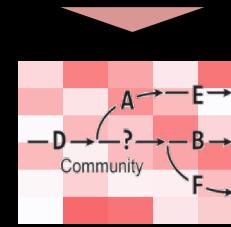
MetaPhlAn2

profiles community
membership



HUMAnN2

profiles species-
specific functions



ShortBRED

detects & quantifies
proteins of interest



Reproducible Workflows & Infrastructure

Microbial Genomics

Meta'omic Profiling

Association Discovery

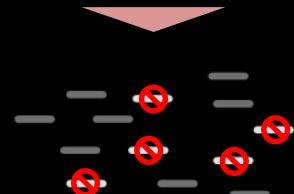


The bioBakery:

2. Microbial community bioinformatics

KneadData

quality controls
meta'omic seq.



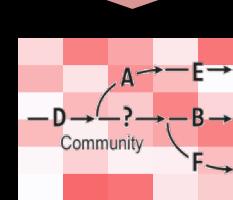
MetaPhlAn2

profiles community
membership



HUMAnN2

profiles species-
specific functions



ShortBRED

detects & quantifies
proteins of interest



Knight Lab



QIIME
integration
amplicon-based
taxonomic profiling

PICRUSt
predictive
functional profiling



Reproducible Workflows & Infrastructure

Microbial Genomics

Meta'omic Profiling

Association Discovery



The bioBakery:

3. Microbial community biostatistics

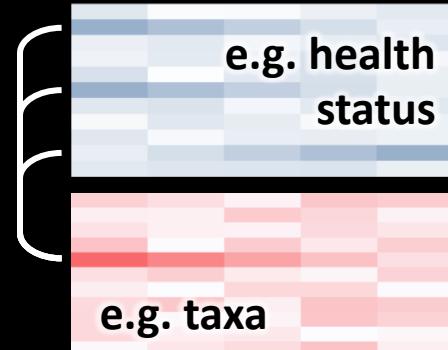
BANoCC

identifies ecological
associations among taxa



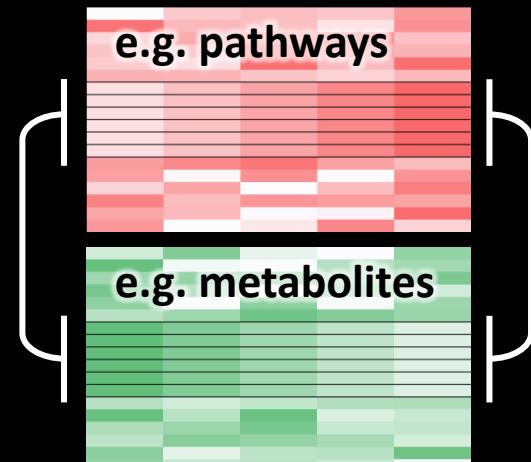
MaAsLin2

links microbial features
with sample metadata



HAIIA

identifies associated “blocks”
in paired multi’omic datasets



Reproducible Workflows & Infrastructure

Microbial Genomics

Meta’omic Profiling

Association Discovery



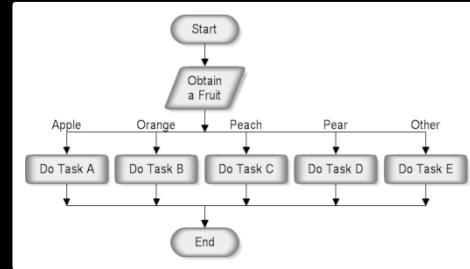
The bioBakery:

4. Infrastructure and reproducible science

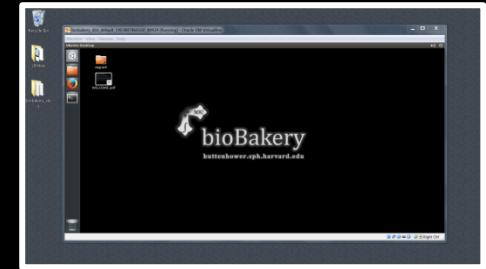
BreadCrumbs
data visualization &
utility library



AnADAMA
scientific workflow
management



**bioBakery VM &
workflows**
deployment system



Reproducible Workflows & Infrastructure

Microbial Genomics

Meta'omic Profiling

Association Discovery



AnADAMA2: Another Automated Data Analysis Management Application

<http://huttenhower.sph.harvard.edu/anadama2>

- **Creating efficient, reproducible workflows**
 - A set of modular tasks to transform inputs into outputs
 - Data, tables, visualizations, statistics...
- Reproducible
 - All tasks are logged
 - Includes commands and software versions
 - Automated documentation generation
- Efficient
 - Only those tasks that need to be rerun will run
 - Make-like operation using targets and dependencies
 - Local and/or grid parallelization
 - Jobs are dispatched/monitored/logged
 - Resubmit if job exceeds time/memory



Randall
Schwager

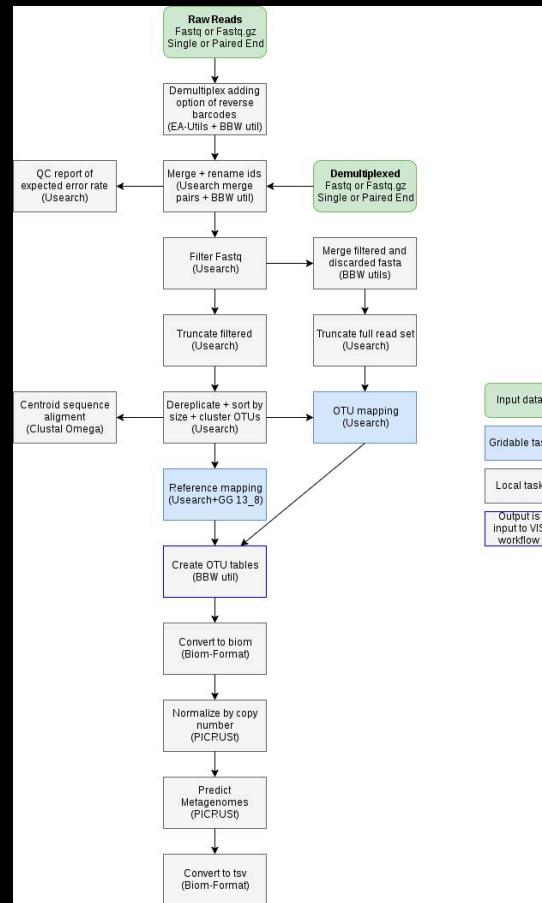
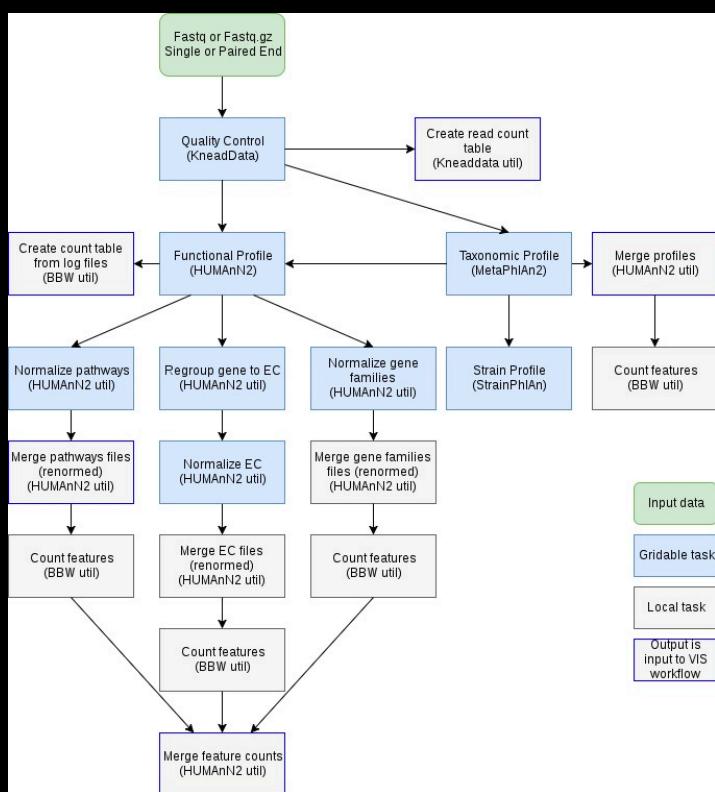


Lauren
McIver



AnADAMA2: Workflows

- Workflows are AnADAMA2's container for a series of steps that consume a series of inputs and produce outputs.





An example workflow

```
from anadama2 import Workflow

workflow = Workflow(remove_options=["input","output"])
workflow.do("ls /usr/bin/ | sort > [t:global_exe.txt]")
workflow.do("ls $HOME/.local/bin/ | sort > [t:local_exe.txt]")
workflow.do("join [d:global_exe.txt] [d:local_exe.txt] > [t:match_exe.txt]")
workflow.go()
```



A fancier example workflow

```
1 ### Section #1: Import anadama2 and create a workflow instance (Required)
2 from anadama2 import Workflow
3 workflow = Workflow(version="0.0.1", description="A workflow to run MetaPhlAn2" )
4
5 ### Section #2: Add custom arguments and parse arguments (Optional)
6 workflow.add_argument("input-extension", desc="the extensions of the input files", default="fastq")
7 args = workflow.parse_args()
8
9 ### Section #3: Get input/output file names (Optional)
10 in_files = workflow.get_input_files(extension=args.input_extension)
11 out_files = workflow.name_output_files(name=in_files, tag="metaphlan2_taxonomy")
12
13 ### Section #4: Add tasks (Required)
14 workflow.add_task_group("metaphlan2.py [depends[0]] --input_type [extension] > [targets[0]]", depends=in_files,
15
16 ### Section #5: Run tasks (Required)
17 workflow.go()
```



But all of that runs with a single command

```
# all tasks are run  
$ python myworkflow.py  
  
# all tasks are skipped (because they were just run and everything is up-to-date)  
$ python myworkflow.py  
  
# all tasks are run (because of the flag applied)  
$ python myworkflow.py --skip-nothing
```

```
(Dec 08 11:50:43) [0/4 - 0.00%] **Ready    ** Task 2: kneaddata  
(Dec 08 11:50:43) [0/4 - 0.00%] **Started   ** Task 2: kneaddata  
(Dec 08 11:50:43) [0/4 - 0.00%] **Ready    ** Task 0: kneaddata  
(Dec 08 11:50:43) [0/4 - 0.00%] **Started   ** Task 0: kneaddata  
(Dec 08 11:50:44) [1/4 - 25.00%] **Completed** Task 2: kneaddata  
(Dec 08 11:50:44) [1/4 - 25.00%] **Ready    ** Task 5: humann2  
(Dec 08 11:50:44) [1/4 - 25.00%] **Started   ** Task 5: humann2  
(Dec 08 11:50:44) [2/4 - 50.00%] **Completed** Task 0: kneaddata  
(Dec 08 11:50:44) [2/4 - 50.00%] **Ready    ** Task 4: humann2  
(Dec 08 11:50:44) [2/4 - 50.00%] **Started   ** Task 4: humann2  
(Dec 08 11:52:48) [3/4 - 75.00%] **Completed** Task 5: humann2  
(Dec 08 11:52:49) [4/4 - 100.00%] **Completed** Task 4: humann2  
Run Finished
```



bioBakery workflows

http://bitbucket.org/biobakery/biobakery_workflows

- A collection of “baked” workflows that execute common microbial community analysis.
- Currently supports 16S, metagenomic, and metatranscriptomic data.
 - Plus downstream visualization and statistics.

```
$ biobakery_workflows wmgx --input examples/wmgx/paired/ --output workflow_output
```

```
(Dec 08 14:33:07) [0/3 - 0.00%] **Ready ** Task 4: kneaddata
(Dec 08 14:33:07) [0/3 - 0.00%] **Started ** Task 4: kneaddata
(Dec 08 14:33:07) [0/3 - 0.00%] **Ready ** Task 2: kneaddata
(Dec 08 14:33:07) [0/3 - 0.00%] **Started ** Task 2: kneaddata
(Dec 08 14:33:07) [0/3 - 0.00%] **Ready ** Task 0: kneaddata
(Dec 08 14:33:25) [0/3 - 0.00%] *GridJob ** Task 4: kneaddata <Grid JobId 76918649 : Submitted>
(Dec 08 14:33:34) [0/3 - 0.00%] *GridJob ** Task 2: kneaddata <Grid JobId 76918676 : Submitted>
(Dec 08 14:34:25) [0/3 - 0.00%] *GridJob ** Task 4: kneaddata <Grid JobId 76918649 : PENDING>
(Dec 08 14:47:26) [0/3 - 0.00%] *GridJob ** Task 4: kneaddata <Grid JobId 76918649 : PENDING>
(Dec 08 14:47:26) [0/3 - 0.00%] *GridJob ** Task 4: kneaddata <Grid JobId 76918649 : Getting benchmarking data>
(Dec 08 14:47:35) [0/3 - 0.00%] *GridJob ** Task 2: kneaddata <Grid JobId 76918676 : PENDING>
(Dec 08 14:47:35) [0/3 - 0.00%] *GridJob ** Task 2: kneaddata <Grid JobId 76918676 : Getting benchmarking data>
(Dec 08 14:49:35) [0/3 - 0.00%] *GridJob ** Task 4: kneaddata <Grid JobId 76918649 : Final status of COMPLETED>
(Dec 08 14:49:35) [0/3 - 0.00%] *GridJob ** Task 2: kneaddata <Grid JobId 76918676 : Final status of COMPLETED>
(Dec 08 14:49:35) [0/3 - 0.00%] **Started ** Task 0: kneaddata
(Dec 08 14:49:35) [1/3 - 33.33%] **Completed** Task 4: kneaddata
(Dec 08 14:49:35) [2/3 - 66.67%] **Completed** Task 2: kneaddata
(Dec 08 14:49:44) [2/3 - 66.67%] *GridJob ** Task 0: kneaddata <Grid JobId 76922265 : Submitted>
(Dec 08 14:50:44) [2/3 - 66.67%] *GridJob ** Task 0: kneaddata <Grid JobId 76922265 : Waiting>
(Dec 08 14:50:44) [2/3 - 66.67%] *GridJob ** Task 0: kneaddata <Grid JobId 76922265 : Getting benchmarking data>
(Dec 08 14:54:46) [2/3 - 66.67%] *GridJob ** Task 0: kneaddata <Grid JobId 76922265 : Final status of COMPLETED>
(Dec 08 14:54:46) [3/3 - 100.00%] **Completed** Task 0: kneaddata
Run Finished
```



bioBakery automated reports

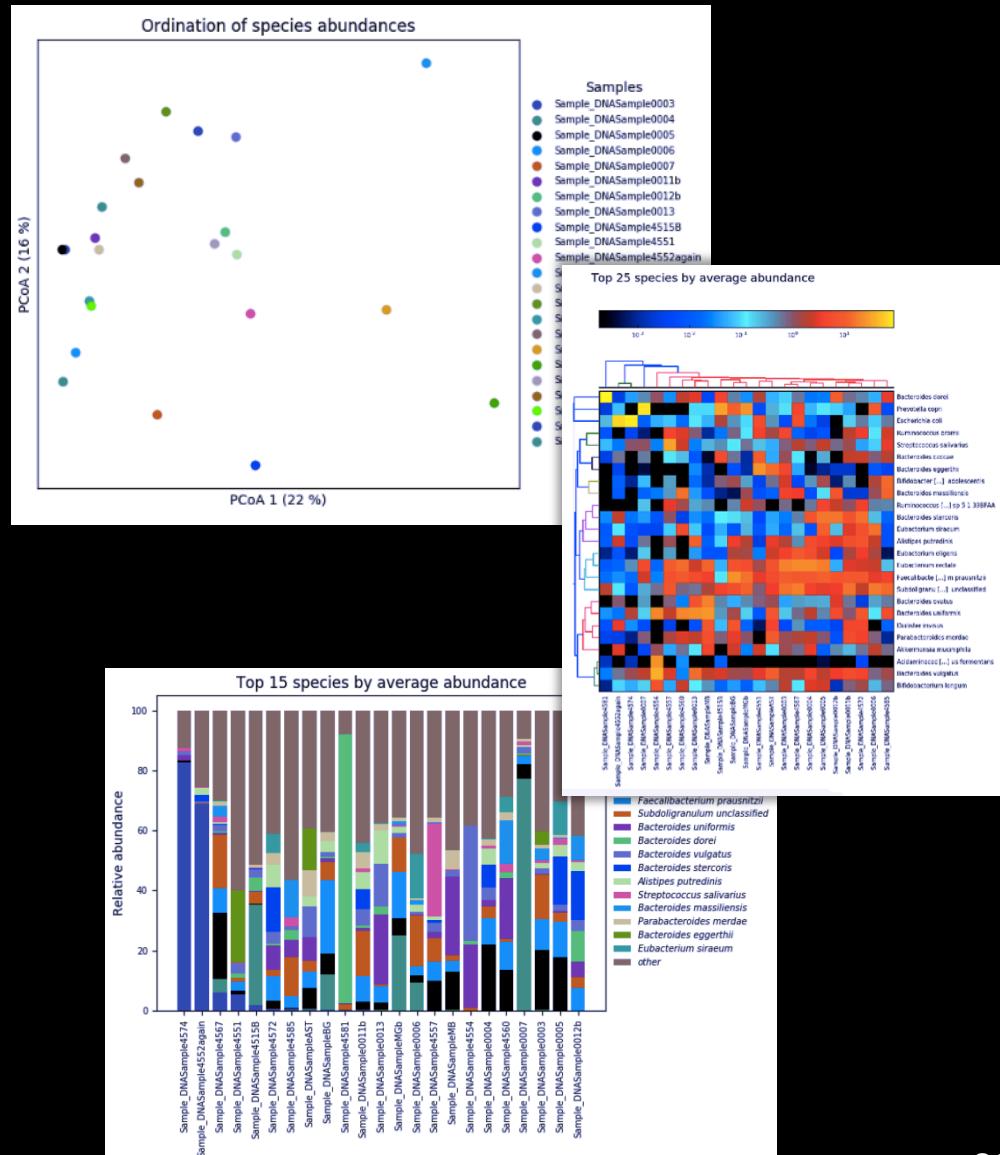
Metagenome and Metatranscriptome Report

Project: UChicago

Date: 06/01/2017

Contents

Introduction	2
Quality Control	2
DNA Samples Quality Control	3
DNA Samples Tables of Filtered Reads	3
DNA Samples Plots of Filtered Reads	6
RNA Samples Quality Control	7
RNA Samples Tables of Filtered Reads	7
RNA Samples Plots of Filtered Reads	10
Taxonomic Profiling of Metagenomic Reads	11
Species Count Table	11
Ordination	12
Heatmap	13
Barplot	14
Functional Profiling of Metagenomic and Metatranscriptomic Reads	15
Pathway Abundance	15
RNA/DNA Normalized Features	18
Features	21
DNA Features	21
RNA Features	24
Data Processing Workflow Information	27
Software Versions	27
Tasks Run	27





Packaging tools and building bioBakery

KneadData
quality controls
meta'omic seq.

MetaPhlAn2
profiles community
membership

HUMAnN2
profiles species-
specific functions



Individual tools packaged as
Homebrew formulae

- **Simple to install**
- **Supports MacOS & Linux**
- **No root permissions required**
 - **Supports \$HOME install**
- **Dependencies satisfied automatically**

Starting with base Ubuntu image,
a master homebrew formula runs
inside vagrant to install all tools...

...provisioning different “flavors”
of virtual machine image



Packaging tools and building bioBakery

KneadData
quality controls
meta'omic seq.

MetaPhlAn2
profiles community
membership

HUMAnN2
profiles species-
specific functions

Homebrew

The missing package manager for OS X

Linuxbrew

The Homebrew package manager for Linux



\$ brew tap biobakery/biobakery

```
=> Tapping biobakery/biobakery
Cloning into '/home/user/.linuxbrew/Library/Taps/biobakery/homebrew-
biobakery'...
remote: Counting objects: 23, done.
remote: Compressing objects: 100% (23/23), done.
remote: Total 23 (delta 11), reused 2 (delta 0), pack-reused 0
Unpacking objects: 100% (23/23), done.
Checking connectivity... done.
Tapped 19 formulae (65 files, 254.4K)
```

\$ brew install maaslin

```
=> Installing maaslin from biobakery/biobakery
=> Downloading
https://bitbucket.org/biobakery/maaslin/get/ced3ca2a3b1d.tar.gz
=> R -q -e install.packages('agricolae', lib='/home/user/.linuxbrew/
=> R -q -e install.packages('gam', lib='/home/user/.linuxbrew/
=> R -q -e install.packages('gamlss', lib='/home/user/.linuxbrew/
=> R -q -e install.packages('gbm', lib='/home/user/.linuxbrew/
=> R -q -e install.packages('glmnet', lib='/home/user/.linuxbrew/
=> R -q -e install.packages('inlinedocs', lib='/home/user/.linuxbrew/
=> R -q -e install.packages('logging', lib='/home/user/.linuxbrew/
=> R -q -e install.packages('MASS', lib='/home/user/.linuxbrew/
=> R -q -e install.packages('nlme', lib='/home/user/.linuxbrew/
=> R -q -e install.packages('optparse', lib='/home/user/.linuxbrew/
=> R -q -e install.packages('outliers', lib='/home/user/.linuxbrew/
=> R -q -e install.packages('penalized', lib='/home/user/.linuxbrew/
=> R -q -e install.packages('pscl', lib='/home/user/.linuxbrew/
=> R -q -e install.packages('robustbase', lib='/home/user/.linuxbrew
=> R -q -e install.packages('tools', lib='/home/user/.linuxbrew/
/home/user/.linuxbrew/Cellar/maaslin/0.0.3-dev-ced3ca2a3b1d: 2,860 files,
44.6M, built in 2 minutes 31 seconds
```



bioBakery workflows tutorial

https://bitbucket.org/biobakery/biobakery/wiki/biobakery_workflows

Workflows

A collection of meta'omic
data processing and
visualization workflows

