



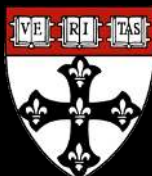
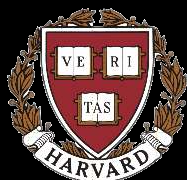
Introduction to metagenomic analysis

Eric A. Franzosa, Ph.D.

Galeb Abu-Ali, Ph.D.

Harvard University CFAR Workshop on
Metagenomics and Transcriptomics

16 September 2014



Huttenhower Research Group
Harvard School of Public Health
Department of Biostatistics



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

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 wiki provides tutorials for Huttenhower tools, illustrating through demos how to use these tools on your datasets. Huttenhower tools can be divided under three main categories as shown below. Click on the tool for the corresponding tutorial.

Composition Analysis

These tools can determine the composition in terms of (i) microbial species and their associated abundances (MetaPhlAn) or (ii) genes and associated pathways (HUMAnN) in the dataset. Please click on the links below for detailed tutorials:

HUMAnN

- Microbial species and associated genes and pathways

MetaPhlAn

- Microbial species and abundances

PhyloPhlAn

- Reconstruction of phylogenetic trees

PICRUSt

- Predict metagenome functional content from marker gene

ShortBRED

- Abundance of proteins of interest in genetic data

Statistical Analysis

These tools can determine the associations from the provided metadata information and microbial composition tables. Please click on the links below for detailed tutorials:

AREP4

- Extract 'omics data from repositories

CCREPE

- Assess the significance of general similarity measures in compositional datasets

LEfSe

- Association between metadata (max 2) and microbial species and abundances

MaAsLin

- Association between metadata (no restriction) and microbial species and abundances

microPITA

- Sample selection in two stage-tiered studies

Visualization

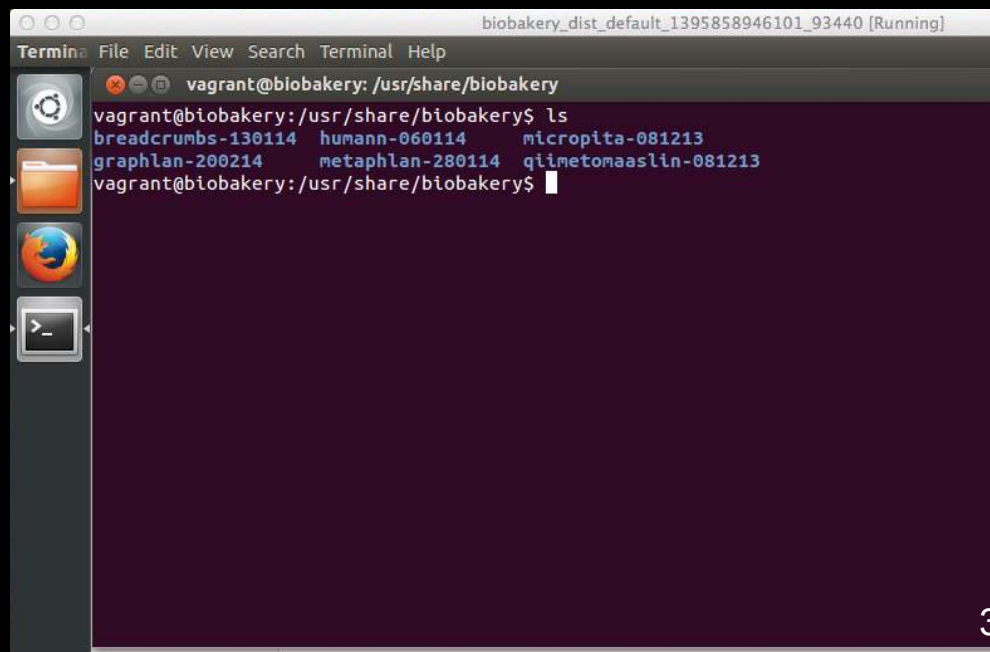
These tools can help visualize taxonomical and phylogenetic information for (i) microbial composition/taxonomy data, (ii) outputs from MetaPhlAn, LEfSe, HUMAnN, MaAsLin. Please click on the link below for detailed tutorial:



The bioBakery: a next-generation environment for microbiome analyses



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





The two big questions...

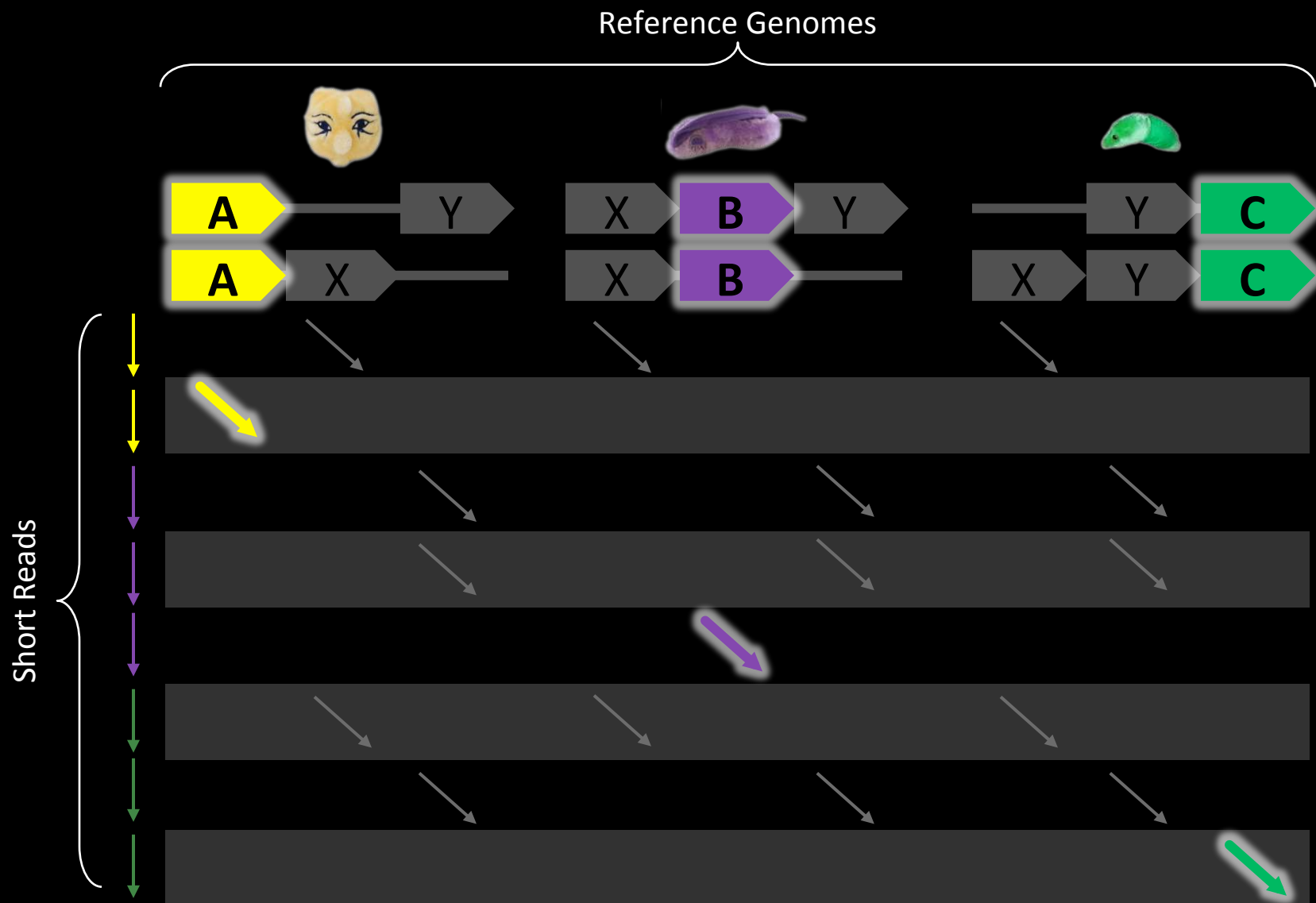
Who is there?
(taxonomic profiling)

What are they doing?
(functional profiling)



MetaPhlAn

Metagenomic Phylogenetic Analysis





Some setup notes

- 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**.



Getting some HMP data

- Go to <http://hmpdacc.org>

Click "Get Data"

HMP
NIH HUMAN
MICROBIOME
PROJECT

Current News

- June 2012
Owen White and Dirk Gevers discuss the HMP on Wisconsin Public Radio
- June 2012
DACC website updated in coordination with publication of HMP data
- April 2012
HMP DACC Reference Genome download page has been updated

[More News Items](#)

Publications

- Ethical Discourse about the Modification of Food for Therapeutic Purpo...
- Caring about trees in the forest: incorporating frailty in risk analys...
- Dietary-fat-induced taurocholic acid promotes pathobiont expansion and...

REFERENCE GENOMES **MICROBIOME ANALYSIS** **IMPACTS ON HEALTH** **TOOLS & TECHNOLOGY** **ETHICAL IMPLICATIONS** **OUTREACH** **HMPDACC DATA BROWSER**

Welcome to the Data Analysis and Coordination Center (DACC) for the National Institutes of Health (NIH) Common Fund supported Human Microbiome Project (HMP). This site is the central repository for all HMP data. The aim of the HMP is to characterize microbial communities found at multiple human body sites and to look for correlations between changes in the microbiome and human health. More information can be found in the menus above and on the NIH Common Fund site.

GET DATA
GET TOOLS

Areas of Interest

Outreach
We welcome feedback on all aspects of the HMP, and are soliciting recommendations for microbial reference genomes. Contact us for more information...



Getting some HMP data

- Check out what's available

HMP
NIH HUMAN
MICROBIOME
PROJECT

Current News

- June 2012
Owen White and Dirk Gevers discuss the HMP on Wisconsin Public Radio
- June 2012
DACC website updated in coordination with publication of HMP data
- April 2012
HMP DACC Reference Genome download page has been updated

[More News Items](#)

Publications

- Ethical Discourse about the Modification of Food for Therapeutic Purpo...
- Caring about trees in the forest: incorporating frailty in risk analys...
- Dietary-fat-induced taurocholic acid promotes pathobiont expansion and...

HMPDACC Data Browser

home > hmpdacc data browser

The HMP DACC Data Portal provides access to all publicly available HMP data sets. If this is your first time to this page, please read the Tour Guide to HMP Sequence Data and the HMP Sample Flow Schematic.

[View Data in the new Interactive Flowchart](#)

[Data Flow Chart PDF](#)

[BLAST](#)

[GET TOOLS](#)

Reference Genomes

- HMRGD HMP Reference Genome sequence data
- HMREFG Reference genome database for read mapping
- Most Wanted Taxa
- HMMDA16S Single cell MDA 16S rRNA Sanger sequencing
- HMP reference genome data at NCBI

Metagenomic 16S Sequence

- HMR16S Raw 16S reads and library metadata
- HM16STR Processed, annotated 16S
- HMMCP Mothur community profiling
- HMQCP QIIME community profiling
- HMP metagenomic 16S data at NCBI

Metagenomic Shotgun Sequence

- HMIWGS/HMASM** Illumina wgs reads and assemblies
- HMBSA Body-site specific assemblies
- HMGI Gene Index
- HMGC Clustered gene index

Mock Community Analysis

- HMMC Mock community 16S and wgs reads

Demonstration Project Data

Click "HMIWGS"



Getting some HMP data

- Check out what's available

HMP
NIH HUMAN
MICROBIOME
PROJECT

Current News

- June 2012
Owen White and Dirk Gevers discuss the HMP on Wisconsin Public Radio
- June 2012
DACC website updated in coordination with publication of HMP data
- April 2012
HMP DACC Reference Genome download page has been updated

[More News Items](#)

Publications

- Ethical Discourse about the Modification of Food for Therapeutic Purpo...
- Caring about trees in the forest: incorporating frailty in risk analys...
- Dietary-fat-induced taurocholic acid promotes pathobiont expansion and...

[More Publications](#)

REFERENCE GENOMES **MICROBIOME ANALYSIS** **IMPACTS ON HEALTH** **TOOLS & TECHNOLOGY** **ETHICAL IMPLICATIONS** **OUTREACH** **HMPDACC DATA BROWSER**

[Feedback](#)

HMIWGS/HMASM - Illumina WGS Reads and Assemblies

In the first phase of WGS sequencing, 764 samples were sequenced, comprising 16 body sites. Of these, 749 samples underwent assembly. Reads for all 764 samples, and 749 assemblies are provided here.

Reads and assemblies were subjected to QC assessment, including identification of outliers by mean contig & ORF density, human hits, rRNA hits and size. 690 samples passed this QC and were included in downstream wgs analyses.

This dataset includes over 35 billion human contaminant-screened reads in FASTQ format, which are 2.3 TB in size, compressed. Reads from each individual sample were assembled using SOAP, generating 48.3 million scaffolds with a total compressed size of 13 GB.

- [Data Table](#)
- [Protocols and Tools](#)
- [Related Pages](#)

Click on your favorite body site

Files					
SRS.ID	Reads Size	Reads MD5	Assembly	Ass. Size	Assembly MD5
Anterior Nares (94 Rows)					
Buccal Mucosa (123 Rows)					
Hard Palate (1 Row)					
Left Retroauricular Crease (9 Rows)					
Mid Vagina (2 Rows)					



Getting some HMP data

Don't click on anything!

- Check out what's available

Apr 2012
HMP DACC Reference Genome
download page has been updated

[More News Items](#)

Publications

- Ethical Discourse about the Modification of Food for Therapeutic Purpo...
- Caring about trees in the forest: incorporating frailty in risk analys...
- Dietary-fat-induced taurocholic acid promotes pathobiont expansion and...

[More Publications](#)

Data Resources

- Tools & Protocols
- BLAST against Reference Genomes
- Project Catalog
- Access to Strains
- Clinical Sampling
- Most Wanted Resource

Files

SRS ID	Reads	Reads Size	Reads MD5	Assembly	Ass. Size	Assembly MD5
Anterior Nares (94 Rows)						
SRS047708		1.7 MB	d786590ff7fec20e8967127991766029		1.3 KB	ed98eda02d80a137c52b6fa8a3c57833
SRS019215		10.1 MB	55de248bbfa8c1bbf4447d007330f7ff		12.1 KB	cab8918433280eafc3d8f6ad78dc1ff7
SRS063178		13.1 MB	336f0b31b92880224c91ad52c4784adc		10.7 KB	99de257f1942e98bf1c052e2d046df33
SRS065179		13.3 MB	27b2c9209bc56cbe219d8c65fa32296c		54.6 KB	bb8b0d62a3c1923abfcaea01a598a60a
SRS065142		13.5 MB	3b05d6fcb205106fbd03f314e39f6d63		7.6 KB	91177065cf438056f2bfc67e99562fe4
SRS018585		16.8 MB	9d4129d2f5fd51b9fc899bd84c47b5b		7.9 KB	aa9e9857b26b9efb4fa39bfaf101dc9d
SRS015640		17.6 MB	595baf36d8b3dcd21149b3086ccbbee		52.4 KB	1c7a464db2fccce17c02f9600c867cb1
SRS056210		18.1 MB	9b2f74b8067e6f20551e6d3b48124c42		18.3 KB	c4abace0ec0b3e7e5ce1513cb8270e56
SRS018312		18.9 MB	2454e80d7e5216adf8d5b1850c98738c		25.4 KB	4f5f760eadd77782862669263e1b1d9d
SRS015450		18.9 MB	eefc0dcf2d52ca5251b01860d54d2bb5		107.1 KB	4e0a83868f2fb44f1788dfe1aaa5e13f
SRS049744		21.5 MB	6d9e2ffc82b08ef37551e902096e4c98		14.3 KB	da7a1cddd3c84b121ff49086432d25d3
SRS012291		21.9 MB	12775f5df6e71961f1c544e84f6c7342		8.9 KB	17b5110d391817c7ce52b7c1026df1ba
SRS051600		22.2 MB	391775b95926a221b8a3cde54a79ae22		13.9 KB	6db7007edd32b534bc918aad42d600ae
SRS019339		23.1 MB	76a621d6503d11d1a133a023dc240ae5		57.3 KB	9255d8206f10ac2611cf45270daa166c
SRS017244		23.5 MB	b7c2dec67738f317cb8826c09e1a9e39		21.3 KB	9bcf59e6b4fe15a4e8ccacbc0bc824ba8
SRS018671		24.0 MB	7548b06b37038440c5420f7677ff7371		135.4 KB	4a180e3ea42a46bcea0a9441b137f243

Show All Save As CSV File

Protocols and Tools



Getting some (prepped) HMP data

- `cd` to your favorite directory and run:

```
ln -s ~/biobakery/metaphlan2/input/7*.fasta .
```

These are subsamples of six HMP files:

- SRS014459.tar.bz2 → 763577454-SRS014459-Stool.fasta
 - SRS014464.tar.bz2 → 763577454-SRS014464-Anterior_nares.fasta
 - SRS014470.tar.bz2 → 763577454-SRS014470-Tongue_dorsum.fasta
 - SRS014472.tar.bz2 → 763577454-SRS014472-Buccal_mucosa.fasta
 - SRS014476.tar.bz2 → 763577454-SRS014476-Supragingival_plaque.fasta
 - SRS014494.tar.bz2 → 763577454-SRS014494-Posterior_fornix.fasta
- All six shotgunned body sites from
 - One subject, first visit
 - Subsampled to 20,000 reads



Who's there: MetaPhlAn2

- <http://huttenhower.sph.harvard.edu/metaphlan2>

The Huttenhower Lab
Department of Biostatistics, Harvard School of Public Health

Contact Documentation People Presentations Publications Research Teaching

Home

You could download MetaPhlAn2 by clicking **here**

MetaPhlAn v2.0

MetaPhlAn v2.0: Metagenomic Phylogenetic Analysis

MetaPhlAn is a computational tool for profiling the composition of microbial communities from metagenomic shotgun sequencing data. MetaPhlAn relies on unique clade-specific marker genes identified from ~17,000 reference genomes (~13,500 bacterial and archaeal, ~3,500 viral, and ~110 eukaryotic), allowing:

- up to 25,000 reads-per-second (on one CPU) analysis speed (orders of magnitude faster compared to existing methods);
- unambiguous taxonomic assignments as the MetaPhlAn markers are clade-specific;
- accurate estimation of organismal relative abundance (in terms of number of cells rather than fraction of reads);
- species-level resolution for bacteria, archaea, eukaryotes and viruses;
- extensive validation of the profiling accuracy on several synthetic datasets and on thousands of real metagenomes.

Obtaining MetaPhlAn v2.0

MetaPhlAn v2.0 can be obtained via the **MetaPhlAn v2.0 Bitbucket repository**. The repository contains the source code and database files used to run MetaPhlAn v2.0, as well as a README file that includes the following information:

- Downloading MetaPhlAn v2.0
- Installation
- Detailed instruction on running MetaPhlAn v2.0

Tutorials

MetaPhlAn v2.0 tutorial is also available [here](#). The tutorial contains a demo dataset and runs through basic command line instructions and the corresponding results.



Who's there: MetaPhlAn2

- But don't! Instead, we've installed MetaPhlAn already for you by clicking **here** on the development site, <http://bitbucket.org/biobakery/metaphlan2>

Atlassian Bitbucket Features Pricing owner/repository English Sign up Log in

Overview

HTTPS https://bitbucket.org/biobakery/me

Last updated	2 hours ago	1 branch	5 Tags
Language	Python	1 Fork	5 Watchers
Access level	Read		

- MetaPhlAn 2.0: Metagenomic Phylogenetic Analysis
 - Description
 - Pre-requisites
 - Installation
 - Basic Usage
 - Full command-line options
 - Utility Scripts
 - Merging Tables
 - Heatmap Visualization
 - GraPhlAn Visualization

MetaPhlAn 2.0: Metagenomic Phylogenetic Analysis

AUTHORS: Nicola Segata (nicola.segata@unitn.it)

Recent activity

- 1 commit
Pushed to biobakery/MetaPhlAn2
3daab15 README.md edited online with ...
Afrah Shafquat · 2 hours ago
- 1 commit
Pushed to biobakery/MetaPhlAn2
ecdadce tagging version 2.0_beta3
Nicola Segata · 4 hours ago
- documentation: how to best run with pa...
Issue #2 commented on in biobakery/MetaPhlAn2
Nicola Segata · 6 hours ago
- 1 commit
Pushed to biobakery/MetaPhlAn2
12ccea README.md edited online with ...
Afrah Shafquat · 6 hours ago
- 1 commit



Who's there: MetaPhlAn2

- The complete MetaPhlAn2 install is in `/usr/local/bioinfo/metaphlan2/`

The screenshot shows the Bitbucket source page for the MetaPhlAn2 repository. The page header includes the Bitbucket logo, navigation links for Features and Pricing, and a search bar. The main content area is titled 'Source' and shows the repository structure. The 'utils' directory is highlighted, and the 'metaphlan2.py' file is selected, showing its details and a list of links to documentation.

Atlassian
Bitbucket Features Pricing owner/repository ? English Sign up Log in

Source

default MetaPhlAn2 /

- db_v20
- utils

.hgtags	205 B	4 hours ago	tagging version 2.0_beta3
README.md	24.6 KB	2 hours ago	README.md edited online with Bitbucket
metaphlan2.py	35.7 KB	6 hours ago	Making MetaPhlAn exiting gracefully when the input format cannot be guessed because two files are

- MetaPhlAn 2.0: Metagenomic Phylogenetic Analysis
 - Description
 - Pre-requisites
 - Installation
 - Basic Usage
 - Full command-line options
 - Utility Scripts
 - Merging Tables
 - Heatmap Visualization
 - GraPhlAn Visualization

>> MetaPhlAn 2.0: Metagenomic Phylogenetic Analysis



From the command line...

- To see what you can do, run:

```
metaphlan2.py -h | less
```

- Use the arrow keys to move up and down,
q to quit back to the prompt



Who's there: MetaPhlAn2

```
ubuntu@ip-10-170-15-59: ~/galeb
File Edit View Search Terminal Tabs Help
ubuntu@ip-10-170-15-59: ~/galeb  x  ubuntu@ip-10-170-15-59: ~/galeb  x

usage: metaphlan2.py --mpa_pkl MPA_PKL --input_type
                        {fastq,fasta,multifasta,multifastq,bowtie2out,sam}
                        [--bowtie2db METAPHLAN_BOWTIE2_DB]
                        [--bt2_ps BowTie2 presets] [--bowtie2_exe BOWTIE2_EXE]
                        [--bowtie2out FILE_NAME] [--no_map] [--tmp_dir]
                        [--tax_lev TAXONOMIC_LEVEL] [--min_cu_len]
                        [--ignore_viruses] [--ignore_eukaryotes]
                        [--ignore_bacteria] [--ignore_archaea] [--stat_q]
                        [--ignore_markers IGNORE_MARKERS] [--avoid_disqm]
                        [--stat] [-t ANALYSIS_TYPE] [--nreads NUMBER_OF_READS]
                        [--pres_th PRESENCE_THRESHOLD] [--clade] [--min_ab] [-h]
                        [-o output file] [--biom biom_output] [--mdelim mdelim]
                        [--nproc N] [-v]
                        [INPUT_FILE] [OUTPUT_FILE]

DESCRIPTION
MetaPhlAn version 2.0.0 beta3 (13 August 2014):
METAgenomic PHyLogenetic ANalysis for metagenomic taxonomic profiling.

AUTHORS: Nicola Segata (nicola.segata@unitn.it)

COMMON COMMANDS

We assume here that metaphlan2.py is in the system path and that mpa_dir bash variable contains the
main MetaPhlAn folder. Also BowTie2 should be in the system path with execution and read
permissions, and Perl should be installed)

===== MetaPhlAn 2 clade-abundance estimation =====

The basic usage of MetaPhlAn 2 consists in the identification of the clades (from phyla to species and
strains in particular cases) present in the metagenome obtained from a microbiome sample and their
relative abundance. This correspond to the default analysis type (--analysis_type rel_ab).
:
```



Who's there: MetaPhlAn2

- To launch your first analysis, run:

```
ln -s /usr/local/bioinf/metaphlan2/db_v20 db_v20
```

```
metaphlan2.py 763577454-SRS014459-Stool.fasta --mpa_pkl  
db_v20/mpa_v20_m200.pkl --bowtie2db db_v20/mpa_v20_m200  
--input_type fasta -o 763577454-SRS014459-Stool.txt
```

- This will run for ~3-4 minutes
- What did you just do?
 - Two new output files:
 - 763577454-SRS014459-Stool.fasta.bowtie2out.txt
 - Contains a mapping of reads to MetaPhlAn markers
 - 763577454-SRS014459-Stool.txt
 - Contains taxonomic abundances as percentages



Who's there: MetaPhlAn

`less 763577454-SRS014459-Stool.fasta.bowtie2out.txt`

```
ubuntu@ip-10-170-15-59: ~/galeb
File Edit View Search Terminal Tabs Help
ubuntu@ip-10-170-15-59: ~/galeb
HWUSI-EAS1625_615HE:4:100:0:1248/1    gi|479140210|ref|NC_021010.1|:1043207-1044529
HWUSI-EAS1625_615HE:4:100:0:1301/1    gi|483877978|ref|NZ_KB890364.1|:31018-31902
HWUSI-EAS1625_615HE:4:100:1000:167/1  gi|242362078|ref|NZ_GG692716.1|:28261-29169
HWUSI-EAS1625_615HE:4:100:1001:1264/1 gi|270295698|ref|NZ_GG730107.1|:470181-472532
HWUSI-EAS1625_615HE:4:100:1001:1320/1 gi|224993849|ref|NZ_ACFY01000158.1|:c1296-10
HWUSI-EAS1625_615HE:4:100:1001:1604/1 gi|319644663|ref|NZ_GL635657.1|:c320982-320029
HWUSI-EAS1625_615HE:4:100:1001:1734/1 gi|484001485|ref|NZ_KB894131.1|:91019-91717
HWUSI-EAS1625_615HE:4:100:1001:259/1  gi|479210985|ref|NC_021043.1|:c1165057-1164158
HWUSI-EAS1625_615HE:4:100:1002:1501/1 gi|224485637|ref|NZ_EQ973491.1|:c620672-618312
HWUSI-EAS1625_615HE:4:100:1003:1644/1 gi|224485636|ref|NZ_EQ973490.1|:c204903-202990
HWUSI-EAS1625_615HE:4:100:1003:1702/1 gi|423335209|ref|NZ_JH976498.1|:329186-330046
HWUSI-EAS1625_615HE:4:100:1003:2030/1 gi|238922432|ref|NC_012781.1|:2910912-2912072
HWUSI-EAS1625_615HE:4:100:1004:353/1  gi|223955873|ref|NZ_DS499674.1|:c266282-265248
HWUSI-EAS1625_615HE:4:100:1004:742/1  gi|283767237|ref|NZ_GG730311.1|:c124395-124171
HWUSI-EAS1625_615HE:4:100:1005:1722/1 gi|410105720|ref|NZ_JH976502.1|:750498-751148
HWUSI-EAS1625_615HE:4:100:1005:505/1  gi|479170689|ref|NC_021020.1|:1540599-1542305
HWUSI-EAS1625_615HE:4:100:1006:848/1  gi|347530298|ref|NC_015977.1|:c3433030-3431387
HWUSI-EAS1625_615HE:4:100:1007:1428/1 gi|423332908|ref|NZ_JH976496.1|:1485161-1487113
HWUSI-EAS1625_615HE:4:100:1007:1465/1 gi|423332908|ref|NZ_JH976496.1|:906255-909584
HWUSI-EAS1625_615HE:4:100:1008:1187/1 gi|224485479|ref|NZ_EQ973214.1|:108053-108250
HWUSI-EAS1625_615HE:4:100:1008:1241/1 gi|270293478|ref|NZ_GG730105.1|:c830784-828727
HWUSI-EAS1625_615HE:4:100:1008:140/1  gi|224514921|ref|NZ_DS499545.1|:41991-42827
HWUSI-EAS1625_615HE:4:100:1009:154/1  gi|301307949|ref|NZ_GG774972.1|:644845-649113
HWUSI-EAS1625_615HE:4:100:1009:467/1  gi|303257489|ref|NZ_GL383997.1|:67163-67873
HWUSI-EAS1625_615HE:4:100:1009:596/1  gi|423290212|ref|NZ_JH724228.1|:c907457-905856
HWUSI-EAS1625_615HE:4:100:1009:82/1   gi|479213596|ref|NC_021044.1|:c1569840-1568455
HWUSI-EAS1625_615HE:4:100:100:193/1   gi|224514888|ref|NZ_DS499516.1|:148626-150644
HWUSI-EAS1625_615HE:4:100:100:866/1   gi|345651619|ref|NZ_JH114362.1|:c62469-60163
HWUSI-EAS1625_615HE:4:100:1010:1731/1 gi|479213596|ref|NC_021044.1|:1831714-1832487
763577454-SRS014459-Stool.bowtie2out.txt
```



Who's there: MetaPhlAn

less 763577454-SRS014459-Stool.txt

```
ubuntu@ip-10-170-15-59: ~/galeb
File Edit View Search Terminal Tabs Help
ubuntu@ip-10-170-15-59: ~/galeb
k_Bacteria 100.0
k_Bacteria|p_Firmicutes 64.82041
k_Bacteria|p_Bacteroidetes 35.17959
k_Bacteria|p_Firmicutes|c_Clostridia 64.82041
k_Bacteria|p_Bacteroidetes|c_Bacteroidia 35.17959
k_Bacteria|p_Firmicutes|c_Clostridia|o_Clostridiales 64.82041
k_Bacteria|p_Bacteroidetes|c_Bacteroidia|o_Bacteroidales 35.17959
k_Bacteria|p_Firmicutes|c_Clostridia|o_Clostridiales|f_Ruminococcaceae 37.71449
k_Bacteria|p_Bacteroidetes|c_Bacteroidia|o_Bacteroidales|f_Bacteroidaceae 31.50008
k_Bacteria|p_Firmicutes|c_Clostridia|o_Clostridiales|f_Eubacteriaceae 21.99035
k_Bacteria|p_Firmicutes|c_Clostridia|o_Clostridiales|f_Lachnospiraceae 5.11557
k_Bacteria|p_Bacteroidetes|c_Bacteroidia|o_Bacteroidales|f_Porphyromonadaceae 3.67952
k_Bacteria|p_Firmicutes|c_Clostridia|o_Clostridiales|f_Ruminococcaceae|g_Subdoligranulum 37.7
k_Bacteria|p_Bacteroidetes|c_Bacteroidia|o_Bacteroidales|f_Bacteroidaceae|g_Bacteroides 31.5
k_Bacteria|p_Firmicutes|c_Clostridia|o_Clostridiales|f_Eubacteriaceae|g_Eubacterium 21.9
k_Bacteria|p_Firmicutes|c_Clostridia|o_Clostridiales|f_Lachnospiraceae|g_Roseburia 5.11
k_Bacteria|p_Bacteroidetes|c_Bacteroidia|o_Bacteroidales|f_Porphyromonadaceae|g_Parabacteroides
k_Bacteria|p_Firmicutes|c_Clostridia|o_Clostridiales|f_Ruminococcaceae|g_Subdoligranulum|s_Su
k_Bacteria|p_Firmicutes|c_Clostridia|o_Clostridiales|f_Eubacteriaceae|g_Eubacterium|s_Eubacte
k_Bacteria|p_Bacteroidetes|c_Bacteroidia|o_Bacteroidales|f_Bacteroidaceae|g_Bacteroides|s_Bac
k_Bacteria|p_Bacteroidetes|c_Bacteroidia|o_Bacteroidales|f_Bacteroidaceae|g_Bacteroides|s_Bac
k_Bacteria|p_Firmicutes|c_Clostridia|o_Clostridiales|f_Eubacteriaceae|g_Eubacterium|s_Eubacte
k_Bacteria|p_Firmicutes|c_Clostridia|o_Clostridiales|f_Lachnospiraceae|g_Roseburia|s_Roseburi
k_Bacteria|p_Bacteroidetes|c_Bacteroidia|o_Bacteroidales|f_Bacteroidaceae|g_Bacteroides|s_Bac
:
```



Who's there: MetaPhlAn2

- Now do the Anterior_nares sample:

```
$ metaphlan2.py 763577454-SRS014459-Stool.fasta --mpa_pkl  
db_v20/mpa_v20_m200.pkl --bowtie2db db_v20/mpa_v20_m200  
--input_type fasta -o 763577454-SRS014459-Stool.txt
```

...

- Note that you can use the up arrow key to make your life easier!
- Usually, you would write a script to analyze all the samples...



Who's there: MetaPhlAn2

```
ubuntu@ip-10-170-15-59: ~/galeb
File Edit View Search Terminal Tabs Help
ubuntu@ip-10-170-15-59: ~/galeb
1 #!/usr/bin/env bash
2
3 for file in *.fasta
4 do
5     echo Now analyzing file $file
6     name=`expr "$file" : '\(.*\)\.fasta'`
7     metaphlan2.py $file \
8         --mpa_pkl db_v20/mpa_v20_m200.pkl \
9         --bowtie2db db_v20/mpa_v20_m200 \
10        --input_type fasta \
11        -o $name".txt"
12 done
13
14
15
~
~
~
```

But let's copy the rest pre-calculated ☺

```
cp ~/biobakery/metaphlan2/output/*.txt .
```



Who's there: MetaPhlAn2

- Let's make a single table containing all six samples:

```
mkdir tmp
```

```
mv *.bowtie2out.txt tmp
```

```
~/biobakery/metaphlan2/utils/
```

```
merge_metaphlan_tables.py *.txt > 763577454.tsv
```

- You can look at this file using `less`
 - Note 1: The arguments `less -x4 -S` will help
 - Note 2: You can set this “permanently” using `export LESS="-x4 -S"`



Who's there: MetaPhlAn2

`less -x4 -S 763577454.tsv`

```
ubuntu@ip-10-170-15-59: ~/galeb
File Edit View Search Terminal Tabs Help

ubuntu@ip-10-170-15-59: ~/galeb
763577454-SRS014459-Stool 763577454-SRS014464-Anterior_nares 763577454-SRS014470-Tongue_dorsum 763577454-SRS014472-Buccal_mucosa
k_Bacteria 100.0 16.77458 100.0 100.0 100.0 100.0
k_Bacteria|p_Actinobacteria 0.0 14.03084 0.86835 0.0 90.34242 0.0
k_Bacteria|p_Actinobacteria|c_Actinobacteria 0.0 14.03084 0.86835 0.0 90.34242 0.0
k_Bacteria|p_Actinobacteria|c_Actinobacteria|o_Actinomycetales 0.0 14.03084 0.86835 0.0 90.34242 0.0
k_Bacteria|p_Actinobacteria|c_Actinobacteria|o_Actinomycetales|f_Actinomycetaceae 0.0 0.0 0.86835 0.0 0.0 0.0
k_Bacteria|p_Actinobacteria|c_Actinobacteria|o_Actinomycetales|f_Actinomycetaceae|g_Actinomyces 0.0 0.0 0.86835 0.0 0.0 0.0
k_Bacteria|p_Actinobacteria|c_Actinobacteria|o_Actinomycetales|f_Actinomycetaceae|g_Actinomyces|s_Actinomyces_graevenitzi 0.0 0.0
k_Bacteria|p_Actinobacteria|c_Actinobacteria|o_Actinomycetales|f_Actinomycetaceae|g_Actinomyces|s_Actinomyces_graevenitzi|t_Actino
k_Bacteria|p_Actinobacteria|c_Actinobacteria|o_Actinomycetales|f_Corynebacteriaceae 0.0 14.03084 0.0 0.0 58.1475 0.0
k_Bacteria|p_Actinobacteria|c_Actinobacteria|o_Actinomycetales|f_Corynebacteriaceae|g_Corynebacterium 0.0 14.03084 0.0 0.0 58.1475
k_Bacteria|p_Actinobacteria|c_Actinobacteria|o_Actinomycetales|f_Corynebacteriaceae|g_Corynebacterium|s_Corynebacterium_matruchotii
k_Bacteria|p_Actinobacteria|c_Actinobacteria|o_Actinomycetales|f_Corynebacteriaceae|g_Corynebacterium|s_Corynebacterium_matruchotii|
k_Bacteria|p_Actinobacteria|c_Actinobacteria|o_Actinomycetales|f_Corynebacteriaceae|g_Corynebacterium|s_Corynebacterium_pseudodiphth
k_Bacteria|p_Actinobacteria|c_Actinobacteria|o_Actinomycetales|f_Corynebacteriaceae|g_Corynebacterium|s_Corynebacterium_pseudodiphth
k_Bacteria|p_Actinobacteria|c_Actinobacteria|o_Actinomycetales|f_Micrococcaceae 0.0 0.0 0.0 0.0 32.19492 0.0
k_Bacteria|p_Actinobacteria|c_Actinobacteria|o_Actinomycetales|f_Micrococcaceae|g_Rothia 0.0 0.0 0.0 0.0 32.19492 0.0
k_Bacteria|p_Actinobacteria|c_Actinobacteria|o_Actinomycetales|f_Micrococcaceae|g_Rothia|s_Rothia_dentocariosa 0.0 0.0 0.0 0.0 32.
k_Bacteria|p_Actinobacteria|c_Actinobacteria|o_Actinomycetales|f_Micrococcaceae|g_Rothia|s_Rothia_dentocariosa|t_Rothia_dentocarios
k_Bacteria|p_Bacteroidetes 35.17959 0.0 24.49606 0.0 9.65758 0.0
k_Bacteria|p_Bacteroidetes|c_Bacteroidia 35.17959 0.0 24.49606 0.0 0.0 0.0
k_Bacteria|p_Bacteroidetes|c_Bacteroidia|o_Bacteroidales 35.17959 0.0 24.49606 0.0 0.0 0.0
k_Bacteria|p_Bacteroidetes|c_Bacteroidia|o_Bacteroidales|f_Bacteroidaceae 31.50008 0.0 0.0 0.0 0.0 0.0
k_Bacteria|p_Bacteroidetes|c_Bacteroidia|o_Bacteroidales|f_Bacteroidaceae|g_Bacteroides 31.50008 0.0 0.0 0.0 0.0 0.0
k_Bacteria|p_Bacteroidetes|c_Bacteroidia|o_Bacteroidales|f_Bacteroidaceae|g_Bacteroides|s_Bacteroides_cellulosilyticus 3.82377 0.0
k_Bacteria|p_Bacteroidetes|c_Bacteroidia|o_Bacteroidales|f_Bacteroidaceae|g_Bacteroides|s_Bacteroides_cellulosilyticus|t_Bacteroides
k_Bacteria|p_Bacteroidetes|c_Bacteroidia|o_Bacteroidales|f_Bacteroidaceae|g_Bacteroides|s_Bacteroides_massiliensis 10.67098 0.0
k_Bacteria|p_Bacteroidetes|c_Bacteroidia|o_Bacteroidales|f_Bacteroidaceae|g_Bacteroides|s_Bacteroides_massiliensis|t_Bacteroides_ma
k_Bacteria|p_Bacteroidetes|c_Bacteroidia|o_Bacteroidales|f_Bacteroidaceae|g_Bacteroides|s_Bacteroides_ovatus 4.08388 0.0 0.0 0.0 0.0
763577454.tsv
```



Who's there: MetaPhlAn2

```
sed "s/.*|//" 763577454.tsv | less -x4 -S
```

```
ubuntu@ip-10-170-15-59: ~/galeb
File Edit View Search Terminal Tabs Help
ubuntu@ip-10-170-15-59: ~/galeb
ID 763577454-SRS014459-Stool 763577454-SRS014464-Anterior_nares 763577454-SRS014470-Tongue_dorsum 763577454-SRS014472-Buccal_mucosa
k_Bacteria 100.0 16.77458 100.0 100.0 100.0 100.0
p_Actinobacteria 0.0 14.03084 0.86835 0.0 90.34242 0.0
c_Actinobacteria 0.0 14.03084 0.86835 0.0 90.34242 0.0
o_Actinomycetales 0.0 14.03084 0.86835 0.0 90.34242 0.0
f_Actinomycetaceae 0.0 0.0 0.86835 0.0 0.0 0.0
g_Actinomyces 0.0 0.0 0.86835 0.0 0.0 0.0
s_Actinomyces_graevenitzii 0.0 0.0 0.86835 0.0 0.0 0.0
t_Actinomyces_graevenitzii_unclassified 0.0 0.0 0.86835 0.0 0.0 0.0
f_Corynebacteriaceae 0.0 14.03084 0.0 0.0 58.1475 0.0
g_Corynebacterium 0.0 14.03084 0.0 0.0 58.1475 0.0
s_Corynebacterium_matruchotii 0.0 0.0 0.0 0.0 58.1475 0.0
t_Corynebacterium_matruchotii_unclassified 0.0 0.0 0.0 0.0 58.1475 0.0
s_Corynebacterium_pseudodiphtheriticum 0.0 14.03084 0.0 0.0 0.0 0.0
t_GCF_000466825 0.0 14.03084 0.0 0.0 0.0 0.0
f_Micrococcaceae 0.0 0.0 0.0 0.0 32.19492 0.0
g_Rothia 0.0 0.0 0.0 0.0 32.19492 0.0
s_Rothia_dentocariosa 0.0 0.0 0.0 0.0 32.19492 0.0
t_Rothia_dentocariosa_unclassified 0.0 0.0 0.0 0.0 32.19492 0.0
p_Bacteroidetes 35.17959 0.0 24.49606 0.0 9.65758 0.0
c_Bacteroidia 35.17959 0.0 24.49606 0.0 0.0 0.0
o_Bacteroidales 35.17959 0.0 24.49606 0.0 0.0 0.0
f_Bacteroidaceae 31.50008 0.0 0.0 0.0 0.0 0.0
g_Bacteroides 31.50008 0.0 0.0 0.0 0.0 0.0
s_Bacteroides_cellulosilyticus 3.82377 0.0 0.0 0.0 0.0 0.0
t_Bacteroides_cellulosilyticus_unclassified 3.82377 0.0 0.0 0.0 0.0 0.0
s_Bacteroides_massiliensis 10.67098 0.0 0.0 0.0 0.0 0.0
t_Bacteroides_massiliensis_unclassified 10.67098 0.0 0.0 0.0 0.0 0.0
s_Bacteroides_ovatus 4.08388 0.0 0.0 0.0 0.0 0.0
:
```



Who's there: MetaPhlAn2

```
$ sed "s/.*|//" 763577454.tsv | sort -k 3 -n -r  
| column -t | less -x4 -S
```

ubuntu@ip-10-170-15-59: ~/galeb			
File Edit View Search Terminal Tabs Help			
ubuntu@ip-10-170-15-59: ~/galeb			
ID	763577454-SRS014459-Stool	763577454-SRS014464-Anterior_nares	763577454-SRS014470-T
t_PRJNA66339	0.0	83.22542	0.0
s_Propionibacterium_phage_PAS50	0.0	83.22542	0.0
p_Viruses_noname	0.0	83.22542	0.0
o_Caudovirales	0.0	83.22542	0.0
k_Viruses	0.0	83.22542	0.0
g_Siphoviridae_noname	0.0	83.22542	0.0
f_Siphoviridae	0.0	83.22542	0.0
c_Viruses_noname	0.0	83.22542	0.0
k_Bacteria	100.0	16.77458	100.0
t_GCF_000466825	0.0	14.03084	0.0
s_Corynebacterium_pseudodiphtheriticum	0.0	14.03084	0.0
p_Actinobacteria	0.0	14.03084	0.86835
o_Actinomycetales	0.0	14.03084	0.86835
g_Corynebacterium	0.0	14.03084	0.0
f_Corynebacteriaceae	0.0	14.03084	0.0
c_Actinobacteria	0.0	14.03084	0.86835
t_GCF_000245815	0.0	2.74374	0.0
s_Dolosigranulum_pigrum	0.0	2.74374	0.0
p_Firmicutes	64.82041	2.74374	74.63559
o_Lactobacillales	0.0	2.74374	24.37049
g_Dolosigranulum	0.0	2.74374	0.0
f_Carnobacteriaceae	0.0	2.74374	0.0
c_Bacilli	0.0	2.74374	24.37049
t_Veillonella_atypica_unclassified	0.0	0.0	16.35219
t_Streptococcus_salivarius_unclassified	0.0	0.0	3.16128
t_Streptococcus_parasanguinis_unclassified	0.0	0.0	21.20921
t_Streptococcus_mitis_oralis_pneumoniae_unclassified	0.0	0.0	0.0
t_Rothia_dentocariosa_unclassified	0.0	0.0	0.0
:			



Who's there: MetaPhlAn

- But it's easier using MeV; <http://www.tm4.org/mev.html>
- `cd to /home/ubuntu/biobakery/mev`
- Double click `tmev.sh` > Run

MeV MultiExperiment Viewer

[About MeV](#) | [Features](#) | [Documentation](#) | [Developers](#) | [Credits](#) | [Support](#) | [Contact](#)

◦ [About MeV](#)
▶ [Features](#)
▶ [Documentation](#)
▶ [Developers](#)
▶ [Credits](#)
▶ [Support](#)
◦ [Contact](#)

Search the MeV website
Search this site:

Funding for the MeV Project
The MeV project is currently funded by

MeV v4.8 includes new clValid module and new annotation
We are proud to announce the release of MeV v4.8. This release includes a new module for the statistical and biological validation of gene clusters, clValid, and updated annotation from Bioconductor. Please download MeV at <http://mev.tm4.org/>.
By eleanorahowe at 11/18/2011 - 21:17

MeV v4.7.4 is released with new Attract module (Windows only)
The Attract Module has returned to MeV, now with significant improvements to its interface and large changes to its underlying algorithm. The algorithm identifies the core gene expression modules that are differentially activated between cell types or different sample groups, and elucidates the set of expression profiles which describe the range of transcriptional behavior within each module. The work is fully described in Mar JC, Wells CA, Quackenbush J. [Defining an informativeness metric for clustering gene expression data](#). *Bioinformatics*. 2011 Apr 15;27(8):1094-100.J. PMID: 21330289.

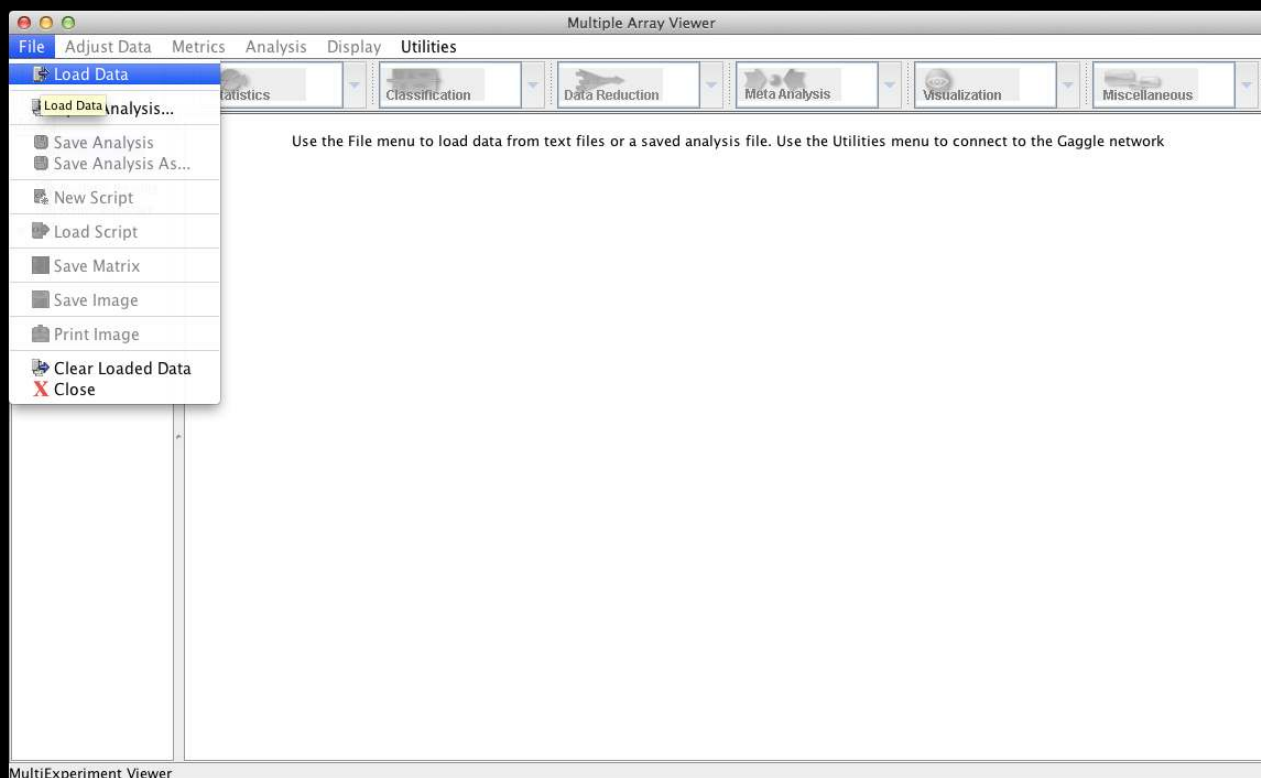
Getting Started with MeV
[Download v4.8](#)
[Release Notes](#)
[Artistic License v2.0](#)
Check out the [Quickstart Guide](#)
[MeV Survey](#)
[Discussion Forums](#)

TM4-Announce Sign-Up
Support TM4 Development and stay informed about the software. We do not share email addresses with anyone.



An interlude: MeV

- Launch MeV, select File/Load data





An interlude: MeV

- Click “Browse” to your TSV file, then
 - Tell MeV it’s a two-color array
 - Uncheck “Load annotation”
 - Click on the upper-leftmost *data* value

Expression File Loader

Select File Loader Help

File (Tab Delimited Multiple Sample (*.tsv))

Select expression data file: /Users/chuttenh/Downloads/763577454.tsv Browse

Selected files: /Users/chuttenh/Downloads/763577454.tsv

☒ Two-color Array ☐ Single-color Array

Load Annotation Data

☒ Automatically download ☐ Load from local file

Choose an organism: No file selected Load Annotation

Choose File

Expression	76357745...	76357745...	76357745...	76357745...	76357745...
k_Bacteri...	100.0	100.0	100.0	100.0	100.0
k_Bacteri...	95.90666	8.2253	2.33635	72.14171	
k_Bacteri...	95.90666	8.2253	2.33635	72.14171	
k_Bacteri...	95.90666	5.51533	2.33635	72.14171	
k_Bacteri...	0	3.51469	0.38831	6.74077	
k_Bacteri...	0	3.51469	0.38831	6.74077	
k_Bacteri...	0	3.51469	0		
k_Bacteri...	0	0		2.43846	
k_Bacteri...	0	0	0.38831	4.30232	
k_Bacteri...	0	42.97557	0	41.42792	
k_Bacteri...	0	42.97557	0	41.42792	

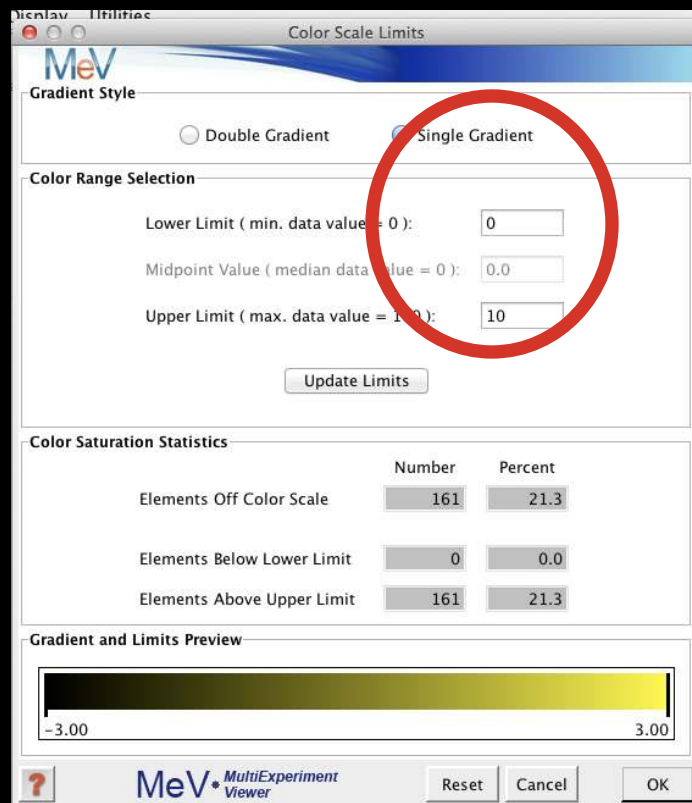
Click the upper-leftmost expression value. Click the Load button to finish.

MeV MultiExperiment Viewer Cancel Load



An interlude: MeV

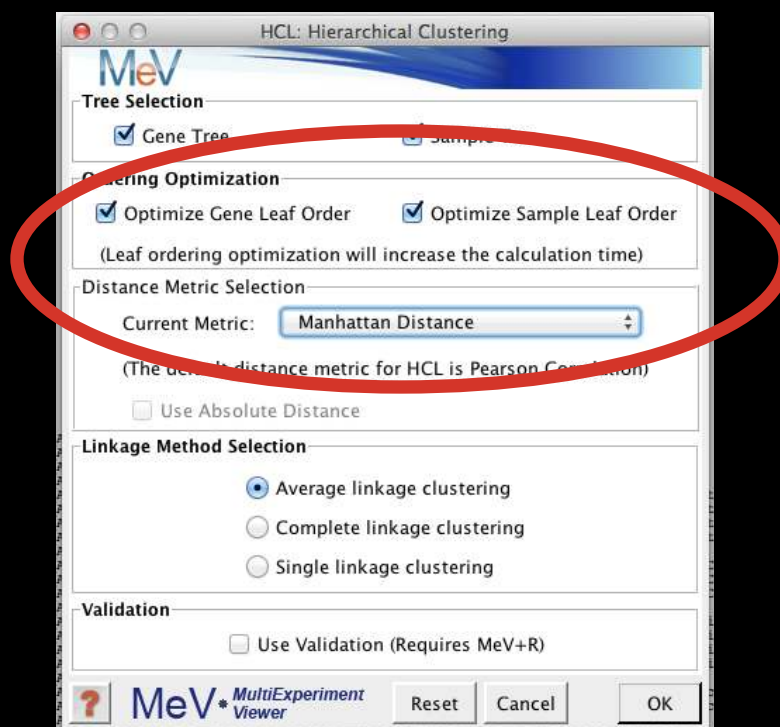
- “Load” your data, then make it visible by:
 - Display/Set Color Scale Limits
 - Choose Single Gradient, min 0, max 10





An interlude: MeV

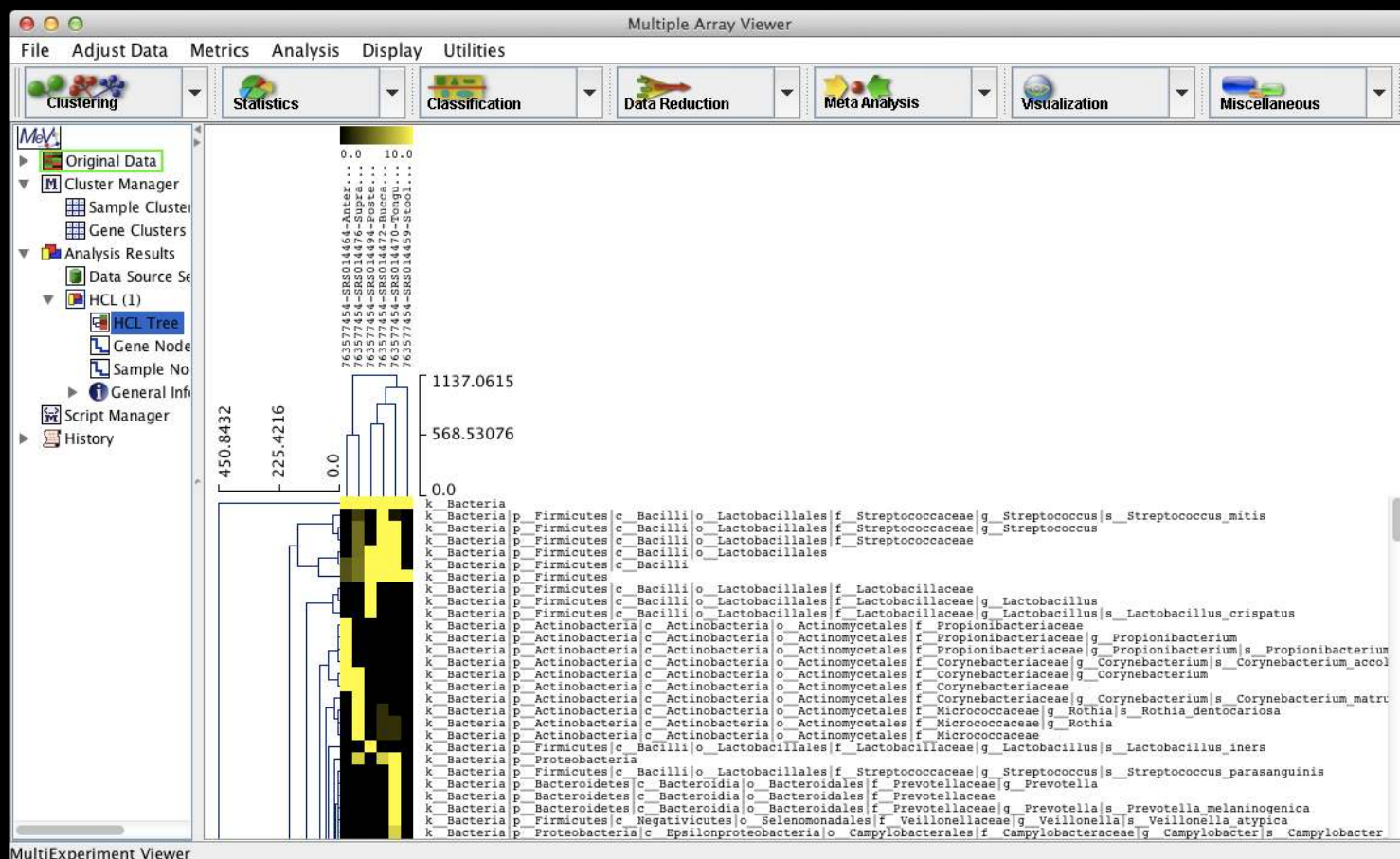
- Finally, to play around a bit:
 - Display/Set Element Size/whatever you'd like
 - Clustering/Hierarchical Clustering
 - Optimize both gene and sample order
 - And select Manhattan Distance (imperfect!)





An interlude: MeV

- If you'd like, you can
 - Display/Sample-Column Labels/Abbr. Names





An interlude: MeV

- MeV is a tool; imperfect, but convenient

- You should likely include just “leaf” nodes

- Species, whose names start include “s__”
- You can filter your file using:

```
cat 763577454.tsv | grep -E ' (Stool) | (s__) ' >  
763577454_species.tsv
```

- You can, but might not want to, z-score normalize

- Adjust Data/Gene-Row Adjustments/Normalize Genes-Rows

- Many other tools built in – experiment!



The two big questions...

Who is there?

What are they doing?

Sample #	1	2	3	4	5	6
Clade1	0.40	0.87	0.43	0.68	0.47	0.32
Clade1 Bug1	0.40	0.56	0.07	0.31	0.42	0.27
Clade1 Bug2	0.00	0.30	0.36	0.37	0.04	0.05
Clade2	0.60	0.13	0.57	0.32	0.53	0.68
Clade2 Bug3	0.11	0.00	0.10	0.32	0.15	0.23
Clade2 Bug4	0.49	0.13	0.47	0.00	0.39	0.45



The ~~two~~ three big questions...

Who is there?

What are they doing?

What does it all mean?

Sample #	1	2	3	4	5	6
Profession	Student	Postdoc	Postdoc	Professor	Student	Student
Gender	Male	Female	Female	Male	Male	Female
Site	Oral	Gut	Oral	Gut	Oral	Gut
Clade1	0.40	0.87	0.43	0.68	0.47	0.32
Clade1 Bug1	0.40	0.56	0.07	0.31	0.42	0.27
Clade1 Bug2	0.00	0.30	0.36	0.37	0.04	0.05
Clade2	0.60	0.13	0.57	0.32	0.53	0.68
Clade2 Bug3	0.11	0.00	0.10	0.32	0.15	0.23
Clade2 Bug4	0.49	0.13	0.47	0.00	0.39	0.45



Properties of microbiome data

- General problem: correlate microbiome features with metadata (potentially controlling for other features)
- Intuitively summarize the results

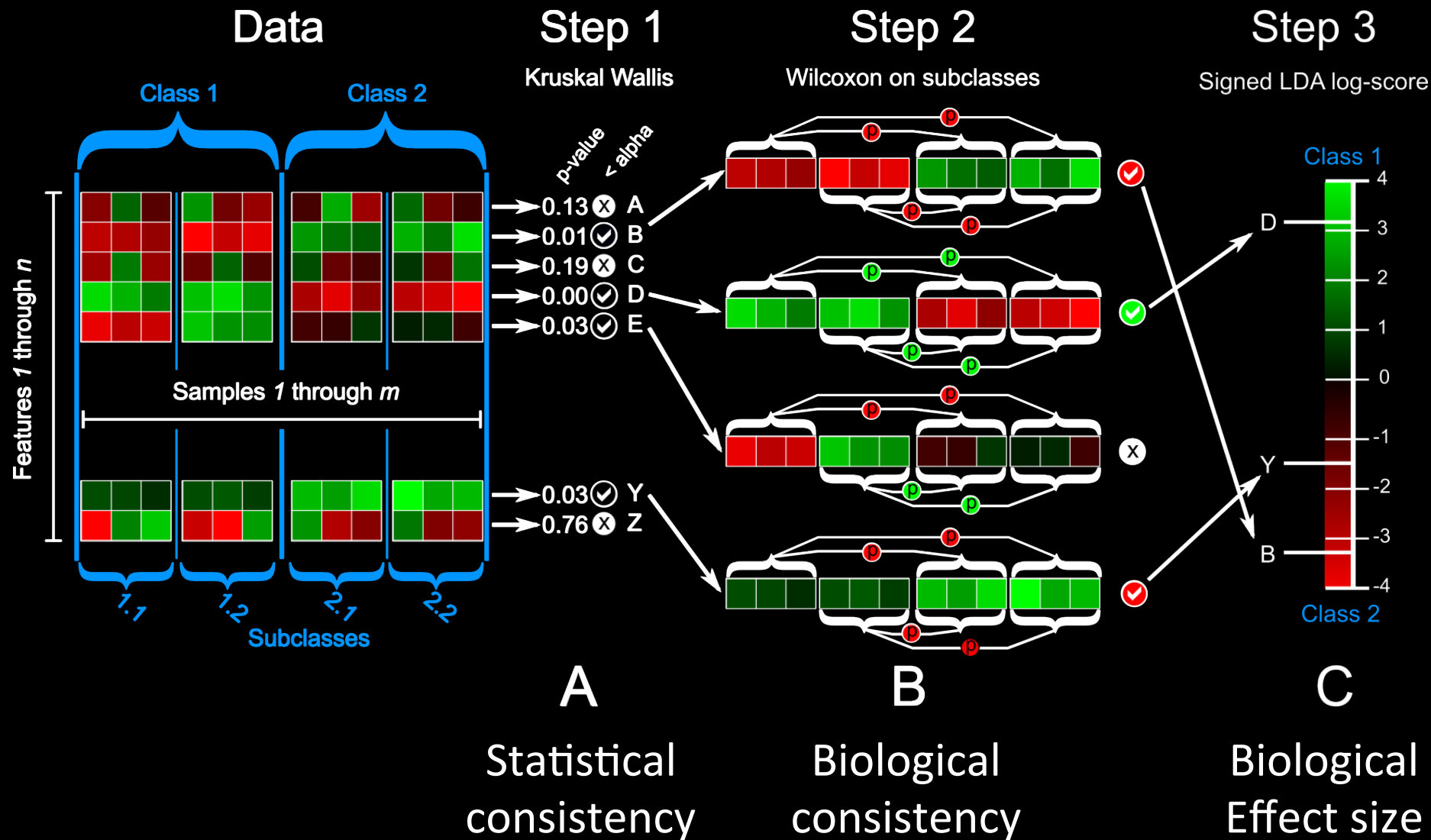
Sample #	1	2	3	4	5	6
Profession	Student	Postdoc	Postdoc	Professor	Student	Student
Gender	Male	Female	Female	Male	Male	Female
Site	Oral	Gut	Oral	Gut	Oral	Gut
Clade1	0.40	0.87	0.43	0.68	0.47	0.32
Clade1 Bug1	0.40	0.56	0.07	0.31	0.42	0.27
Clade1 Bug2	0.00	0.30	0.36	0.37	0.04	0.05
Clade2	0.60	0.13	0.57	0.32	0.53	0.68
Clade2 Bug3	0.11	0.00	0.10	0.32	0.15	0.23
Clade2 Bug4	0.49	0.13	0.47	0.00	0.39	0.45



LEfSe: LDA Effect Size

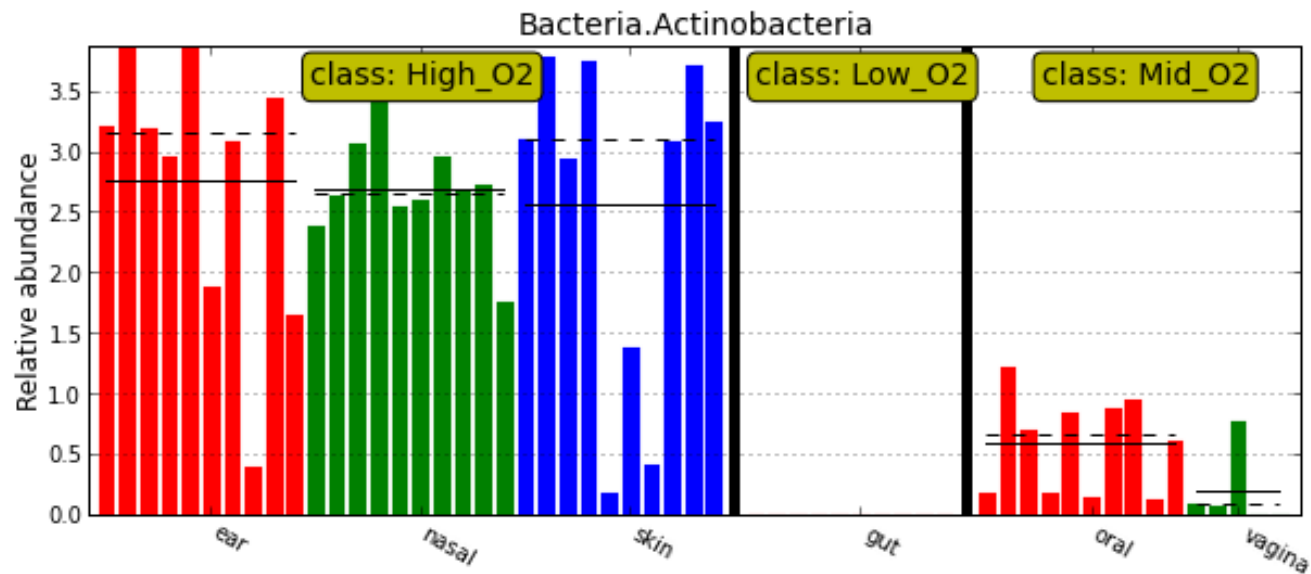
Finding metagenomic biomarkers

Nicola
Segata



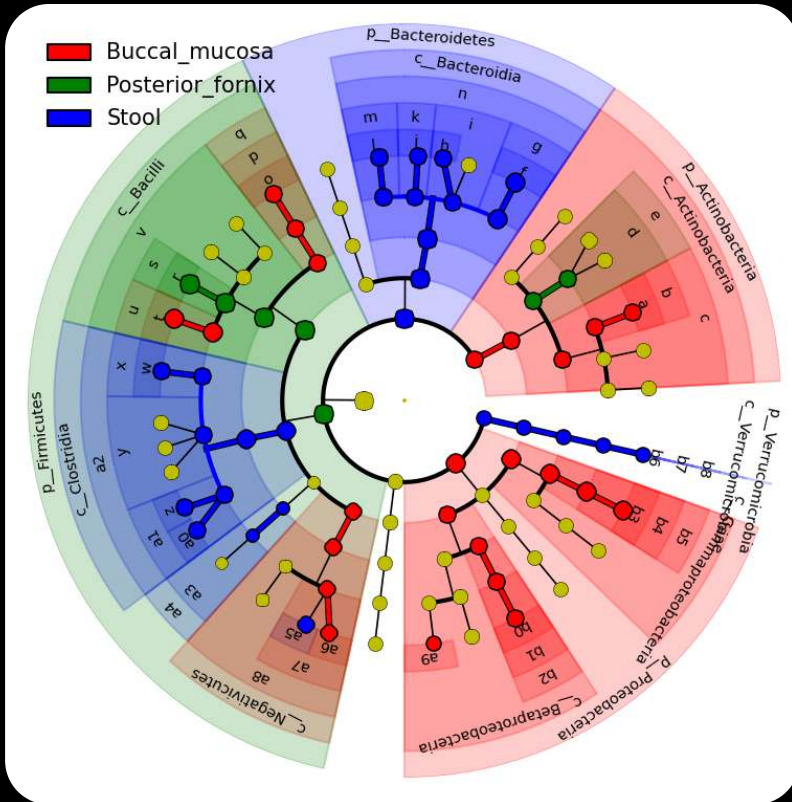


Example LEfSe application: Find O₂-loving bugs (controlling for body site)

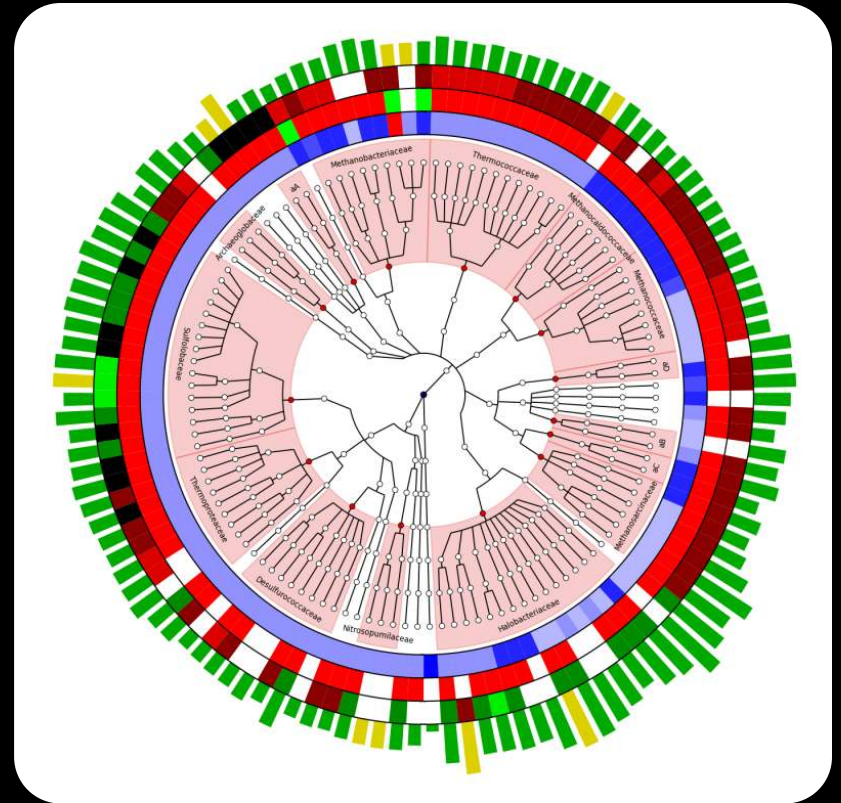




Superimpose enrichments on the tree of life using GraPhlAn



LEfSe Associations



Metadata Rings

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



What it means: LEfSe

- Let's get *all* of the HMP species data:
http://hmpdacc.org/resources/data_browser.php

HMPDACC Data Browser

The HMP DACC Data Portal provides access to all publicly available HMP data sets. If this is your first time to this page, please read the Tour Guide to HMP Sequence Data and the HMP Sample Flow Schematic.

[View Data in the new Interactive Flowchart](#)

[Data Flow Chart PDF](#)

[BLAST](#)

[GET TOOLS](#)

Reference Genomes

- HMRGD HMP Reference Genome sequence data
- HMREFG Reference genome database for read mapping
- Most Wanted Taxa
- HMMDA16S Single cell MDA 16S rRNA Sanger sequencing
- HMP reference genome data at NCBI

Metagenomic 16S Sequence

- HMR16S Raw 16S reads and library metadata
- HM16STR Processed, annotated 16S
- HMMCP Mothur community profiling
- HMQCP QIIME community profiling
- HMP metagenomic 16S data at NCBI

Metagenomic Shotgun Sequence

- HMIWGS/HMASM Illumina wgs reads and assemblies
- HMBSA Body-site specific assemblies
- HMGI Gene Index
- HMGC Clustered gene index
- HMGS GO slim analysis
- HMP Shotgun community profiling
- HMSMCP Shotgun MetaPhlAn Community Profiling**
- HMMRC Metabolic reconstruction and clustering
- HMGDI Genes of Interest
- HM4WGS/HMHASM Illumina/454 Hybrid reads and assemblies
- HMHGI Illumina/454 hybrid gene index

Mock Community Analysis

- HMMC Mock community 16S and wgs reads

Demonstration Project Data

- Demonstration project data at NCBI

Other Data

- HMFUNC Functional databases used for metabolic reconstruction
- RSEQ RNAseq expression analysis of dental microbiome
- HMP Project Catalog Reference Genome & Metagenomic metadata

Current News

- June 2012
Owen White and Dirk Gevers discuss the HMP on Wisconsin Public Radio
- June 2012
DACC website updated in coordination with publication of HMP data
- April 2012
HMP DACC Reference Genome download page has been updated

[More News Items](#)

Publications

- Ethical Discourse about the Modification of Food for Therapeutic Purpo...
- Caring about trees in the forest: incorporating frailty in risk analys...
- Dietary-fat-induced taurocholic acid promotes pathobiont expansion and...

[More Publications](#)

Data Resources

- Tools & Protocols
- BLAST against Reference Genomes
- Project Catalog
- Access to Strains



What it means: LEfSe

- Download the MetaPhlAn1 table for all 700 samples

HMP
NIH HUMAN
MICROBIOME
PROJECT

Current News

- June 2012
Owen White and Dirk Gevers discuss the HMP on Wisconsin Public Radio
- June 2012
DACC website updated in coordination with publication of HMP data
- April 2012
HMP DACC Reference Genome download page has been updated

More News Items

Publications

- Ethical Discourse about the Modification of Food for Therapeutic Purpo...
- Caring about trees in the forest: incorporating frailty in risk analys...
- Dietary-fat-induced taurocholic acid promotes pathobiont expansion and...

REFERENCE GENOMES **MICROBIOME ANALYSIS** **IMPACTS ON HEALTH** **TOOLS & TECHNOLOGY** **ETHICAL IMPLICATIONS** **OUTREACH** **HMPDACC DATA BROWSER**

HMSMCP - Shotgun MetaPhlAn Community Profiling

Reads generated by Illumina wgs sequencing were analyzed by MetaPhlAn, which infers a table of relative abundances for all taxonomic levels (from phyla to species) for bacteria and archaea. The MetaPhlAn classifier compares each read to a pre-computed catalog of unique clade-specific markers in order to identify high-confidence matches. This is very computationally efficient, as the catalog contains only ~4% of sequenced microbial genes. Apart from standard quality control, no other metagenomic pre-processing steps (eg. error detection, assembly, or gene annotation) are required. The classifier normalizes the total number of reads in each clade by the nucleotide length of its markers and provides the relative abundance of each taxonomic unit. Microbial reads belonging to clades with no sequenced genomes available are reported as an "unclassified" subclade of the closest ancestor with available sequence data.

- Data Table
- Protocols and Tools
- Related Pages

File	Download	Size	MD5
HMP.ab.txt.bz2		314.3 KB	2ce7fe2514067267fe27b0232fd827d4

Protocols and Tools

This table has been generated using MetaPhlAn version 1.1.0 (March 2012) with default parameter settings.

Related Pages

downloads.hmpdacc.org/data/HMSMCP/HMP.ab.txt.bz2



Downloading from the command line

- Instead of saving this, download it by:
 - Right-click to copy the URL
 - Run
`wget <paste URL here>`
 - Note: `curl -O <URL>` works just as well

```
ubuntu@ip-10-170-15-59: ~/galeb
File Edit View Search Terminal Tabs Help

ubuntu@ip-10-170-15-59: ~/galeb
ubuntu@ip-10-170-15-59:~/galeb$ wget http://downloads.hmpdacc.org/data/HMSMCP/HMP.ab.txt.bz2
--2014-09-14 03:02:07--  http://downloads.hmpdacc.org/data/HMSMCP/HMP.ab.txt.bz2
Resolving downloads.hmpdacc.org (downloads.hmpdacc.org)... 134.192.156.83
Connecting to downloads.hmpdacc.org (downloads.hmpdacc.org)|134.192.156.83|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 321880 (314K) [application/x-bzip2]
Saving to: `HMP.ab.txt.bz2'

100%[=====] 321,880      --.-K/s   in 0.1s

2014-09-14 03:02:07 (2.36 MB/s) - `HMP.ab.txt.bz2' saved [321880/321880]

ubuntu@ip-10-170-15-59:~/galeb$ ls
HMP.ab.txt.bz2
```



What it means: LEfSe

- Make sure this file is in your current directory, and expand it:

```
bunzip2 HMP.ab.txt.bz2
```

- Look at the result

```
less -S HMP.ab.txt
```

- **IMPORTANT!!!**

- This file's too big to analyze directly today

```
ln -s ~/biobakery/data/HMP.ab.filtered.txt
```

- This is great – tons of data, but no metadata

- Scripts and data from HUMAnN to the rescue:

```
~/biobakery/metadata.py ~/biobakery/data/hmp_metadata.dat  
< HMP.ab.filtered.txt > HMP.ab.filtered.metadata.tsv
```

- NOW take a look again



What it means: LEfSe

HMPab.filtered.metadata.tsv - LibreOffice Calc

File Edit View Insert Format Tools Data Window Help

Liberation Sans 10

A1 f(x) Σ = sid

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Sid	SRS043001	SRS017127	SRS021473	SRS011134	SRS050184	SRS011529	SRS048164	SRS016516	SRS052330	SRS011355	SRS011452	SRS019787	SRS054776	SRS024140	SRS014683	SRS016018
2	RANDSID	550534656	159551223	158479027	158499257	508703490	159166850	861967750	159753524	765640925	158944319	159146620	764669880	764224817	159207311	763961826	76444734
3	START	Q3_2009	Q2_2009	Q1_2009	Q3_2009	Q2_2009	Q3_2009	Q3_2009	Q2_2009	Q3_2009	Q1_2009	Q1_2009	Q2_2009	Q2_2009	Q2_2009	Q1_2009	Q2_2009
4	GENDER	female	male	male	male	female	male	male	female	female	female	male	male	male	male	male	male
5	VISNO	1	1	2	1	1	1	1	1	1	1	1	2	2	2	1	1
6	STSite	Stool	Buccal_mucosa	Buccal_mucosa	Stool	Posterior_fornix	Stool	Stool	Posterior_fornix	Posterior_fornix	Posterior_fornix	Stool	Stool	Buccal_mucosa	Buccal_mucosa	Stool	Stool
7	Parent_Specimen	700106291	700033688	700097185	700014832	700038759	700016608	700038870	700032243	700038805	700015577	700016136	700038231	700106652	700100608	700023337	70002464
8	Run ID	704GE	61HJLAAXX	61K2LAAXX	61JGUAAXX	704N4	61PNF	61NTL	61VKUAAXX	7055M	61KYVAAXX	61WER	704MU	621F6	61JDLAAXX	614NM	61PPR
9	Lane	6	7	6	6	8	5	1	4	1	8	2	4	5	2	1	1
10	SRS	700106291	700033688	700097185	700014832	700038759	700016610	700038870	700032243	700038805	700015579	700016142	700038263	700106652	700100608	700023337	70002467
11	Mean Quality	29.75	29	33	27	31.07	31.92	33.16	33	32.91	24	32.97		32.65	29	34.2	
12	Number of Quality Bases	5938136715	6349906779	4784765427	6210952530	5330742538	6605008593	4654538314	5233442453	5439436415	3621291754	2749895648		5709503626	6823557644	393099251	
13	Percent of Human Reads	0.0024	0.6746	0.8842	0.0002	0.7872	0.0004	0.0002	0.8342	0.7861	0.8857	0.0043		0.734	0.865	0.000	
14	Unique Non-Human Bases	6779445369	2611543625	959383209	7794356797	1264901720	7187185125	4912150763	5230564060	1234292763	683832927	2252544600		1607164515	1336140118	410671756	
15	k_Bacteria p_Proteobact	0	0	0	0.59019	0	0.15046	1.46625	0	0	0	0	0	0	0	0	0
16	k_Bacteria p_Actinobact	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	k_Bacteria p_Bacteroides	2.57633	0	0	4.20761	0	5.52547	8.52942	0	0	0	15.90901	1.27072	0.14801	1.19015	6.11771	11.5346
18	k_Bacteria p_Firmicutes	0.09916	0	0.00175	0.33665	0	0.10695	0.93543	0	0	0	1.85125	0.07391	0	0.35352	0.14181	1.3823
19	k_Bacteria p_Firmicutes	0	0	0	18.31739	0	0	0	0	0	0	0	0	0	0	7.40863	0
20	k_Bacteria p_Bacteroides	0	0	0	0	0	0	0.00281	0	0	0	0.13473	1.12503	0	0.01746	0.00159	0.0094
21	k_Bacteria p_Actinobact	0	0.07358	0.71281	0	0	0	0	0	0	0	0	0	10.69807	1.55503	0	0
22	k_Bacteria p_Firmicutes	8e-05	0.53975	0.70681	0.02233	0	0	0	0	0	0	0	0	6.75115	0.29668	0	0
23	k_Bacteria p_Proteobact	0	0.0049	1.10502	0	0	0	0	0	0	0	0	0	0.23766	0.12036	0.01257	0
24	k_Bacteria p_Firmicutes	0.00075	0.00467	1.79514	0.04706	0	0	0.00087	0	0	0	0	0	0.75602	0.47212	0	0
25	k_Bacteria p_Fusobact	0	0.07934	0.03212	0	0	0	0	0	0	0	0	0	0.0556	0.00685	0	0
26	k_Bacteria p_Bacteroides	0	0	0	0.01024	0	0.00011	0.0001	0	0	0	0	0.019	0	0	0	0
27	k_Bacteria p_Verrucom	0	0	0	1.92544	0	0.00249	0	0	0	0	0	5.01912	0	0.03603	0	1.9544
28	k_Bacteria p_Fusobact	0	0.18413	0.09554	0	0	0	0	0	0	0	0	0	0.13772	0.05674	0	0
29	k_Bacteria p_Bacteroides	84.60804	0.17674	1.20942	58.53925	0	87.17536	80.23906	0	0.00528	95.18632	90.48074	0	0.89671	6.19701	90.3718	93.8067
30	k_Bacteria p_Proteobact	0	0.0049	1.10502	0	0	0	0	0	0	0	0	0	0.23766	0.12036	0.01257	0
31	k_Bacteria p_Firmicutes	0	3.7702	5.24454	0	0	0	0	0	0	0.00317	0	0	5.2466	20.28856	0	0
32	k_Bacteria p_Firmicutes	0	0	0	0	28.41152	0	0	98.62158	0.09156	0	0	0	0	0	0	0
33	k_Bacteria p_Bacteroides	0	0	0.01284	0	0.00127	0.00134	0	0	0	0	0	0	0	0	0	2.3011
34	k_Bacteria p_Bacteroides	0	0	0.13547	0	0.66733	0.93226	0	0	0	0	1.02811	0	0	0.1132	0	1.5792
35	k_Bacteria p_Firmicutes	0	0	0	0	0	0	3.28129	0	0	0	0	0	0	0	0	0
36	k_Bacteria p_Actinobact	0	0	0	0.0256	0	0.15669	0	0	0	0	0	0	0	0.02756	0.50559	0
37	k_Bacteria p_Firmicutes	15.11446	0	0.22313	31.85621	0.01935	3.9669	13.62363	0	0.01816	4.1026	3.84111	0	0.18439	9.12663	5.54894	3.5452
38	k_Bacteria p_Firmicutes	0.00108	11.78992	0.19295	0.00324	0	0	0	0	0	0	0	0	1.76326	0.7472	0	0
39	k_Bacteria p_Proteobact	0.06223	0	0	0.01538	0	0.00569	0	0	0	0	0.0034	0	0	0.01972	0.29636	0
40	k_Bacteria p_Bacteroides	2.57633	0	0	3.72212	0	0.47934	1.40935	0	0	0	0	0.65938	0	0.59358	2.11564	1.4296

Sheet 1 / 1

Default

STD

CountA=1

100%



Generate heatmap with hclust2

```
cat HMP.ab.filtered.metadata.tsv | grep -E 'sid|GENDER|  
STSite|s__' | grep -v "t__" | sed "s/.*|//" >  
HMP.ab.filtered.metadata.txt
```

```
ln -s ~/biobakery/hclust2/hclust2.py
```

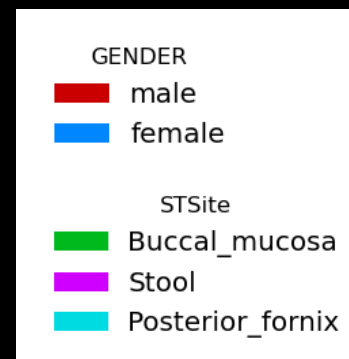
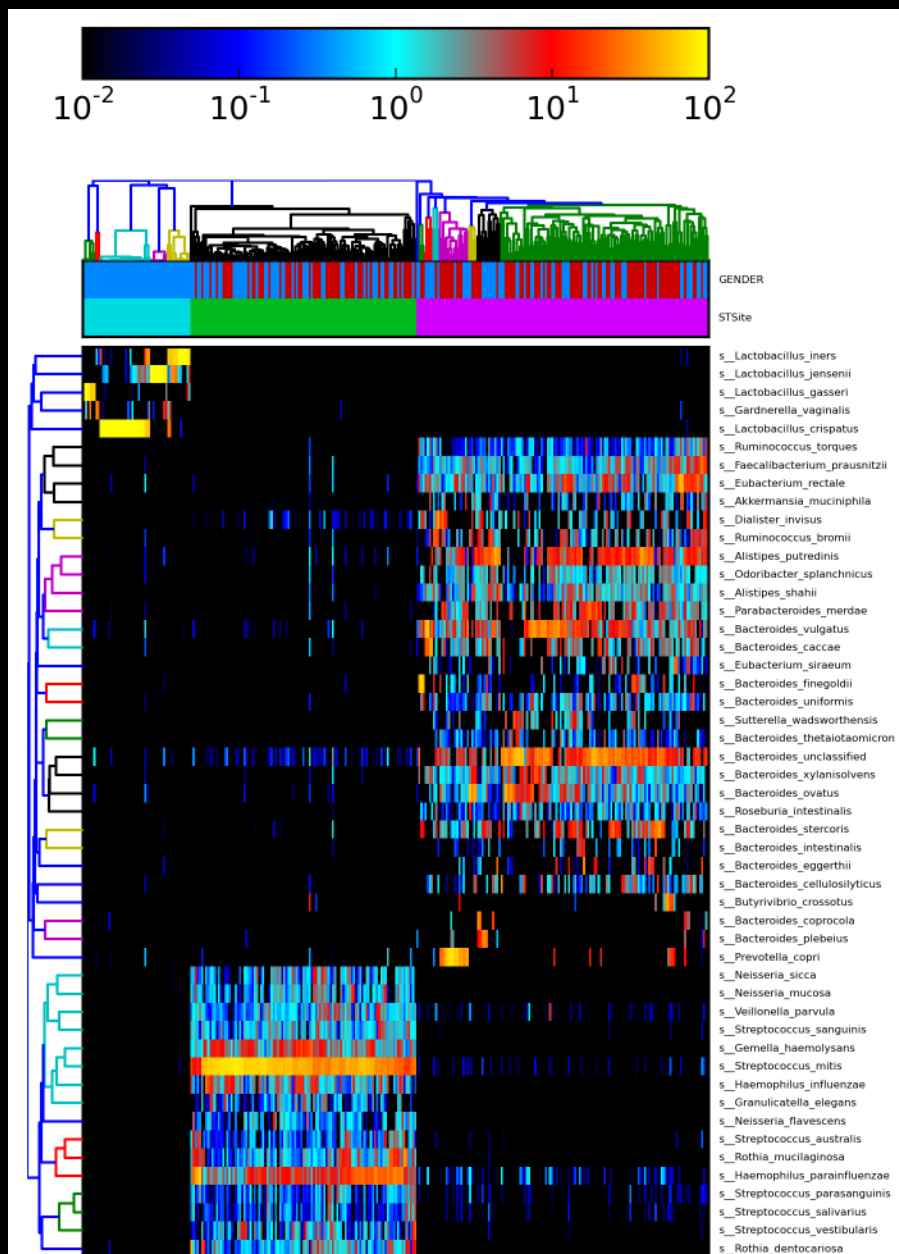
```
hclust2.py -h | less
```

```
hclust2.py -i HMP.ab.filtered.metadata.txt -o  
HMP.log_scale.png --ftop 50 --f_dist_f correlation --  
s_dist_f braycurtis --cell_aspect_ratio 9 -l --fperc 99  
--flabel_size 4 --metadata_rows 1,2 --legend_file  
HMP.log_scale.legend.png --max_flabel_len 100 --  
metadata_height 0.075 --minv 0.01 --no_slabs --dpi  
300
```

<https://bitbucket.org/nsegata/hclust2>



Generate heatmap with hclust2





What it means: LEfSe

- Let's modify the *.tsv file to be for LEfSe

HMPab.filtered.metadata.tsv - LibreOffice Calc

File Edit View Insert Format Tools Data Window Help

Liberation Sans 10

A1 f0 Σ = sid

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	sid	SRS043001	SRS017127	SRS021473	SRS011134	SRS050184	SRS011529	SRS048164	SRS016516	SRS052330	SRS011355	SRS011452	SRS019787	SRS054776	SRS024140	SRS014683	SRS016018
2	RANDSID	550534656	159551223	158479027	158499257	508703490	159166850	861967750	159753524	765640925	158944319	159146620	764669880	764224817	159207311	763961826	76444734
3	START	Q3_2009	Q2_2009	Q1_2009	Q1_2009	Q3_2009	Q2_2009	Q3_2009	Q2_2009	Q3_2009	Q1_2009	Q1_2009	Q2_2009	Q2_2009	Q2_2009	Q1_2009	Q2_2009
4	GENDER	female	male	male	male	female	male	male	female	female	female	male	male	male	male	male	male
5	VISNO	1	1	2	1	1	1	1	1	1	1	1	2	2	2	1	1
6	SISite	Stool	Buccal_mucosa	Buccal_mucosa	Stool	Posterior_fornix	Stool	Stool	Posterior_fornix	Posterior_fornix	Posterior_fornix	Stool	Stool	Buccal_mucosa	Buccal_mucosa	Stool	Stool
7	Parent_Specimen	700106291	700033688	700097185	700014832	700038759	700016608	700038870	700032243	700038805	700015577	700016136	700038231	700106652	700100608	700023337	70002464
8	Run ID	704GE	61HJLAAXX	61K2LAAXX	61JGUAAXX	704N4	61PNF	61NTL	61VKUAAXX	7055M	61KYVAAXX	61WER	704MU	621F6	61JDIAAXX	614NM	61PPR
9	Lane	6	6	7	6	8	5	1	4	1	8	2	4	5	2	1	1
10	SRS	700106291	700033689	700097185	700014837	700038759	700016610	700038870	700032243	700038805	700015579	700016142	700038263	700106652	700100608	700023337	70002467
11	Mean Quality	29.75	29	33	27	31.07	31.92	33.16	33	32.91	24	32.97	29	32.65	29	34.2	34.2
12	Number of Quality Bases	5938136715	6349906779	4784765427	6210952530	5330742538	6605008593	4654538314	5233442453	5439436415	3621291754	2749895648	5709503626	6823557644	393099251	0.000	393099251
13	Percent of Human Reads	0.0024	0.6746	0.8842	0.0002	0.7872	0.0004	0.0002	0.8342	0.7861	0.8857	0.0043	0.734	0.865	0.865	0.000	0.000
14	Unique Non-Human Bases	6779445369	2611543625	959383209	7794356797	1264901720	7187185125	4912150763	5230564060	1234292763	683832927	2252544600	1607164515	1336140118	410671756		410671756
15	k_Bacterialp_Proteobact	0	0	0	0	0.59019	0	0.15046	1.46625	0	0	0	0	0	0	0	0
16	k_Bacterialp_Actinobact	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	k_Bacterialp_Bacteroid	2.57633	0	0	4.20761	0	5.52547	8.52942	0	0	0	15.90901	1.27072	0.14801	1.19015	6.11771	11.5346
18	k_Bacterialp_Firmicutes	0.09916	0	0.00175	0.33665	0	0.10695	0.93543	0	0	0	1.85125	0.07391	0	0.35352	0.14181	1.3823
19	k_Bacterialp_Firmicutes	0	0	0	18.31739	0	0	0	0	0	0	0	0	0	7.40863	0	0
20	k_Bacterialp_Bacteroid	0	0	0	0	0	0	0.00281	0	0	0	0.13473	1.12503	0	0.01746	0.00159	0.0094
21	k_Bacterialp_Actinobact	0	0.07358	0.71281	0	0	0	0	0	0	0	0	0	10.69807	1.55503	0	0
22	k_Bacterialp_Firmicutes	0	0.53975	0.70681	0.02233	0	0	0	0	0	0	0	0	6.75115	0.29668	0	0
23	k_Bacterialp_Proteobact	0	0.0049	1.10502	0	0	0	0	0	0	0	0	0	0.23766	0.12036	0.01257	0
24	k_Bacterialp_Firmicutes	0.00075	0.00467	1.79514	0.04706	0	0	0.00087	0	0	0	0	0	0.75602	0.47212	0	0
25	k_Bacterialp_Fusobact	0	0.07934	0.03212	0	0	0	0	0	0	0	0	0	0.0556	0.00685	0	0
26	k_Bacterialp_Bacteroid	0	0	0	0.01024	0	0.00011	0.0001	0	0	0	0	0.019	0	0	0	0
27	k_Bacterialp_Verrucom	0	0	0	1.92544	0	0.00249	0	0	0	0	5.01912	0	0.03603	0	1.9544	0
28	k_Bacterialp_Fusobact	0	0.18413	0.09554	0	0	0	0	0	0	0	0	0	0.13772	0.05674	0	0
29	k_Bacterialp_Bacteroid	84.60804	0.17674	1.20942	58.53925	0	87.17536	80.23906	0	0.00528	95.18632	90.48074	0.89671	6.19701	90.3718	93.8067	0
30	k_Bacterialp_Proteobact	0	0.0049	1.10502	0	0	0	0	0	0	0	0	0.23766	0.12036	0.01257	0	0
31	k_Bacterialp_Firmicutes	0	3.7702	5.24454	0	0	0	0	0	0	0	0.00317	0	5.2466	20.28856	0	0
32	k_Bacterialp_Firmicutes	0	0	0	0	28.41152	0	0	98.62158	0.09156	0	0	0	0	0	0	0
33	k_Bacterialp_Bacteroid	0	0	0	0.01284	0	0.00127	0.00134	0	0	0	0	0	0	0	0	2.3011
34	k_Bacterialp_Bacteroid	0	0	0	0.13547	0	0.66733	0.93226	0	0	0	1.02811	0	0	0.1132	0	1.5792
35	k_Bacterialp_Firmicutes	0	0	0	0	0	0	3.28129	0	0	0	0	0	0	0	0	0
36	k_Bacterialp_Actinobact	0	0	0	0.0256	0	0.15669	0	0	0	0	0	0	0	0.02756	0.50559	0
37	k_Bacterialp_Firmicutes	15.11446	0	0.22313	31.85621	0.01935	3.9669	13.62363	0	0.01816	4.1026	3.84111	0.18439	9.12663	5.54894	3.5452	0
38	k_Bacterialp_Firmicutes	0.00108	11.78992	0.19295	0.00324	0	0	0	0	0	0	0	1.76326	0.7472	0	0	0
39	k_Bacterialp_Proteobact	0.06223	0	0	0.01538	0	0.00569	0	0	0	0.0034	0	0	0.01972	0.29636	0	0
40	k_Bacterialp_Bacteroid	2.57633	0	0	3.72212	0	0.47934	1.40935	0	0	0	0	0.65938	0	0.59358	2.11564	1.4296

Sheet 1 / 1

Default

STD

CountA=1

100%



What it means: LEfSe

- Delete all of the metadata rows except:
 - RANDSID and STSite
 - Save it as tab-delimited text: HMP.ab.filtered.metadata2.txt

The screenshot shows the LibreOffice Calc application window titled 'HMPab.Filtered.metadata.tsv - LibreOffice Calc'. The spreadsheet contains taxonomic data with columns for taxonomic levels (k, p, c, o, f, g, h, i) and various bacterial groups. A red circle highlights the 'RANDSID' column header in row 1. A 'Save' dialog box is open, showing the file name 'HMP.ab.filtered.metadata' and the save location 'ubuntu galeb'. The 'File type' is set to 'Text CSV (.csv)'. A red circle highlights the 'Text CSV' option in the file type list. The background spreadsheet shows a table with columns E, F, G, H, I and rows of numerical data.

E	F	G	H	I
158499257	508703490	159166850	861967750	159
Stool	Posterior_fomix	Stool	Stool	Posterior
0.59019	0	0.15046	1.46625	
0	0	0	0	
4.20761	0	5.52547	8.52942	
0.33665	0	0.10695	0.93543	
18.31739	0	0	0	
0	0	0	0.00281	
0	0	0	0	
0.02233	0	0	0	
0	0	0	0	
0.04706	0	0	0.00087	
0	0	0	0	
0.01024	0	0.00011	0.0001	
1.92544	0	0.00249	0	
0	0	0	0	
58.53925	0	87.17536	80.23906	
0	0	0	0	
0	28.41152	0	0	
0.01284	0	0.00127	0.00134	
0.13547	0	0.66733	0.93226	
0	0	0	3.28129	
0.0256	0	0.15669	0	
31.85621	0.01935	3.9669	13.62363	
0.00324	0	0	0	
0.01538	0	0.00569	0	
3.72212	0	0.47934	1.40935	
0.16707	0.00498	7.51136	3.31162	
0	0	0	0	
0	99.22989	0.06204	0	
6.8605	0	0.34321	2.3577	
0	0	0	3.28129	
0	0	0	0	
59.05921	0.08695	87.39299	80.72071	
11.52772	0	2.28265	8.59348	
0.75902	0.06886	0.46469	1.6544	
1.92544	0	0.00249	0	
0	0	0	0	
0.17079	0	0.03475	0.01768	



What it means: LEfSe

- Visit LEfSe at: <http://huttenhower.sph.harvard.edu/galaxy/>

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

Tools

search tools

HUTTENHOWER LAB MODULES

LEfSe

- A) Format Data for LEfSe
- B) LDA Effect Size (LEfSe)
- C) Plot LEfSe Results
- D) Plot Cladogram
- E) Plot One Feature
- F) Plot Differential Features

MetaPhlAn

GraPhlAn

microPITA

MaAsLin

PICRUSt

Get Data

Upload File from your computer

First click here

Thanks for visiting our lab's tools and applications page, implemented within the [Galaxy](#) web application and workflow framework. Here, we provide a number of resources for metagenomic and functional genomic analyses, intended for research and academic use. Please see the menus and folders to the left for an overview of available tools including documentation, sample data, and publications.

Our lab's research interests include metagenomics and the [human microbiome](#), the relationships between microbial communities and human health, microbiome systems biology, and large-scale computational methods for studying all of these areas. In addition to the tools provided here, feel free to take a look at our additional [research](#) and [publications](#), including the [Sleipnir library](#) for computational functional genomics.

The tools are available here without account creation. However, you are strongly invited to create an account for having access to the history, saved analyses, datasets and workflows. You can create an account and/or log in using the User menu in the top-right corner.

If you have any comments, questions, or suggestions, please contact [Dr. Huttenhower](#).

History

Unnamed history

0 bytes

This history is empty. You can [load your own data](#) or [get data from an external source](#)



What it means: LefSe

- Then upload your formatted table
 - After you upload, wait for the progress meter to turn green!

The screenshot shows the Galaxy web interface. The top navigation bar includes 'Galaxy / Huttenhower Lab', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A 'Using 0%' progress indicator is in the top right.

On the left, the 'Tools' panel lists 'HUTTENHOWER LAB MODULES' with 'LEfSe' selected. Under 'LEfSe', several options are listed: 'A) Format Data for LEfSe', 'B) LDA Effect Size (LEfSe)', 'C) Plot LEfSe Results', 'D) Plot Cladogram', 'E) Plot One Feature', and 'F) Plot Differential Features'. Below this are other modules like 'MetaPhlAn', 'GraPhlAn', 'microPITA', 'MaAsLin', and 'PICRUST'. At the bottom, there's a 'LOAD DATA MODULE' section with 'Get Data' and 'Upload File from your computer'.

The main panel displays the 'Upload File (version 1.1.4)' tool. It has a 'File Format' dropdown set to 'Auto-detect'. Below it, a red circle highlights the 'Choose File' button. To the right of this button, the text 'HMP.ab.filtered.metadata.txt' is shown. A red arrow points from the text '1. Click here, browse to HMP.ab.filtered.metadata.txt' to the 'Choose File' button. Below the file selection area is a 'URL/Text' input field. Further down, there's a 'Convert spaces to tabs' checkbox (unchecked) and a 'Genome' dropdown set to 'unspecified (?)'. A red circle highlights the 'Execute' button at the bottom of the tool panel. A red arrow points from the text '2. Then here' to the 'Execute' button.

On the right, the 'History' panel shows 'Unnamed history' with a size of '269.2 KB'. A red circle highlights a message box that says: 'This history is empty. You can load your own data or get data from an external source'. A red arrow points from the text '3. Then watch here' to this message box.



What it means: LEfSe

- Then tell LEfSe about your metadata:

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

Tools

search tools

HUTTENHOWER LAB MODULES

LEfSe

A) Format Data for LEfSe

B) LDA Effect Size (LEfSe)

C) Plot LEfSe Results

D) Plot Cladogram

E) Plot One Feature

F) Plot Differential Features

MetaPhlAn

GraPhlAn

microPITA

MaAsLin

PICRUST

LOAD DATA MODULE

Get Data

Upload File from your computer

A) Format Data for LEfSe (version 1.0)

Upload a tabular file of relative abundances and class labels (possibly also subclass and subjects labels) for LEfSe – See samples below – Please use Galaxy Get-Data/Upload-File. Use File-Type = Tabular:

2: HMP.ab.filtered.metadata.txt

Select whether the vectors (features and meta-data information) are listed in rows or columns:

Rows

Select which row to use as class:

#2:STSite

Select which row to use as subclass:

no subclass

Select which row to use as subject:

#1:RANDSID

Per-sample normalization of the sum of the values to 1M (recommended when very low values are present):

Yes

Execute

History

Unnamed history

538.3 KB

2: HMP.ab.filtered.metadata.txt

1. Click here

2. Then select STSite

3. Then select RANDSID

4. Then here



What it means: LEfSe

- Then select LDA=4, “One-against-all,” and run LEfSe!
 - You can change other default statistical parameters if desired

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

Tools

search tools

HUTTENHOWER LAB MODULES

LEfSe

- A) Format Data for LEfSe
- B) LDA Effect Size (LEfSe)**
- C) Plot LEfSe Results
- D) Plot Cladogram
- E) Plot One Feature
- F) Plot Differential Features

MetaPhlAn

GraPhlAn

microPITA

MaAsLin

PICRUSt

LOAD DATA MODULE

B) LDA Effect Size (LEfSe) (version 1.0)

Select data:

Alpha value for the factorial Kruskal-Wallis test among classes:

Alpha value for the pairwise Wilcoxon test between subclasses:

Threshold on the logarithmic LDA score for discriminative features:

Do you want the pairwise comparisons among subclasses to be performed only among the subclasses with the same name?:

Select the strategy for multi-class analysis:

Execute

History

Unnamed history

1.0 MB

- 3: A) Format Data for LEfSe on data 2
- 2: HMP.ab.filtered.metadata.txt

1. Click here

2. Then “4” here (finds only very extreme differences)

3. Then “one” here (finds differences in at least one condition rather than in all conditions)

4. Then GO!



What it means: LEfSe

- You can plot the results as a bar plot
 - Again, lots of graphical parameters to modify if desired

The screenshot displays the Galaxy web interface for the Huttenhower Lab. The top navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The 'Tools' panel on the left lists 'HUTTENHOWER LAB MODULES' with a sub-section for 'LEfSe'. The 'LEfSe' section contains five options: 'A) Format Data for LEfSe', 'B) LDA Effect Size (LEfSe)', 'C) Plot LEfSe Results', 'D) Plot Cladogram', and 'E) Plot One Feature'. Option 'C) Plot LEfSe Results' is circled in red, with a red arrow pointing to it and the text '1. Click here'. The main workspace shows the configuration for 'C) Plot LEfSe Results (version 1.0)'. It includes a 'Select data' dropdown set to '4: B) LDA Effect Size (LEfSe) on data 3', and sections for 'Set text and label options', 'Set some graphical options to personalize the output', 'Output format' (set to 'png'), and 'Set the dpi resolution of the output' (set to '150'). The 'Execute' button at the bottom of the tool configuration is also circled in red, with a red arrow pointing to it and the text '2. Then here'. The right-hand 'History' panel shows a list of previous jobs, including '4: B) LDA Effect Size (LEfSe) on data 3', '3: A) Format Data for LEfSe on data 2', and '2: HMP.ab.filtered.metadata.txt'.



What it means: LEfSe

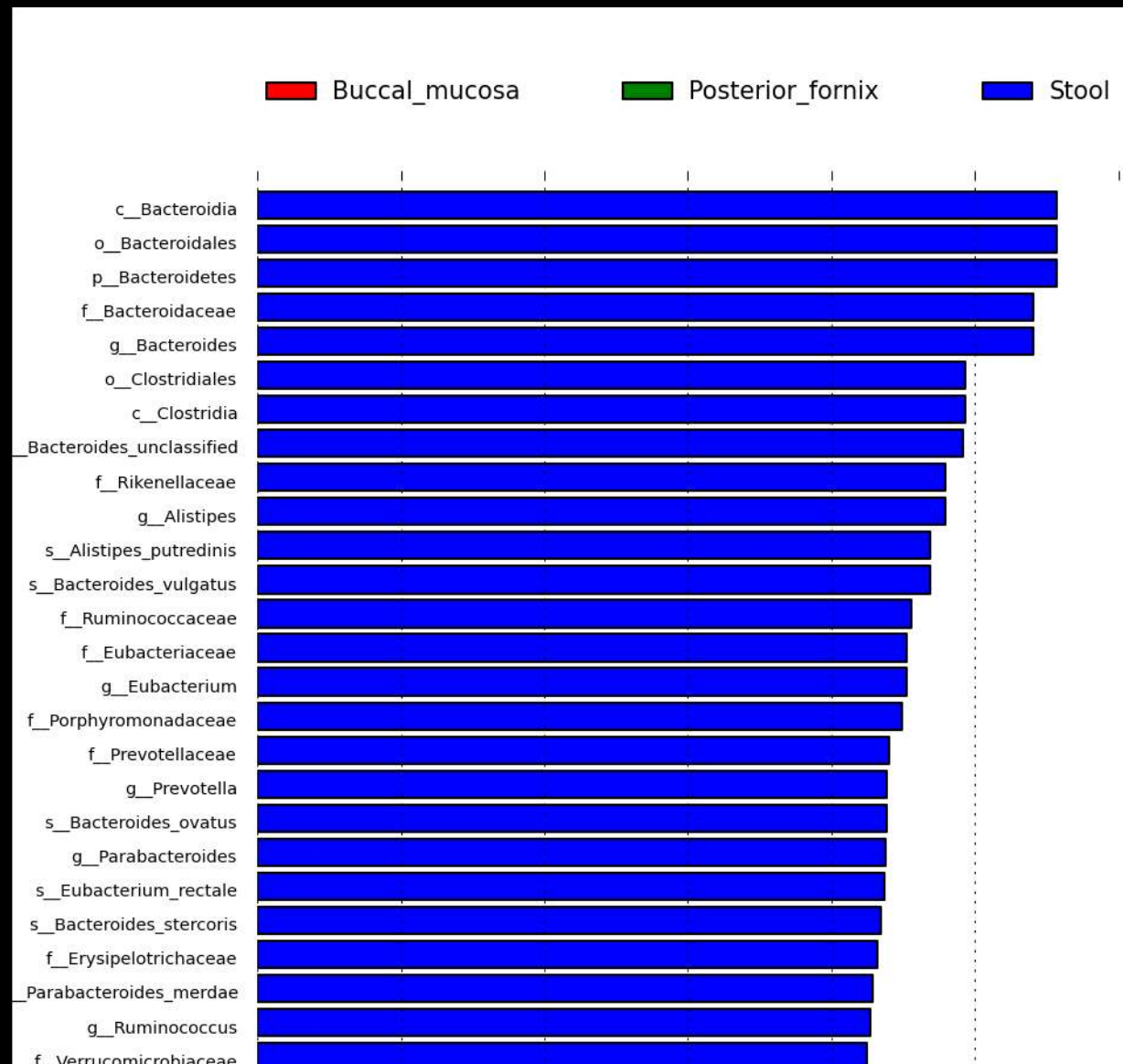
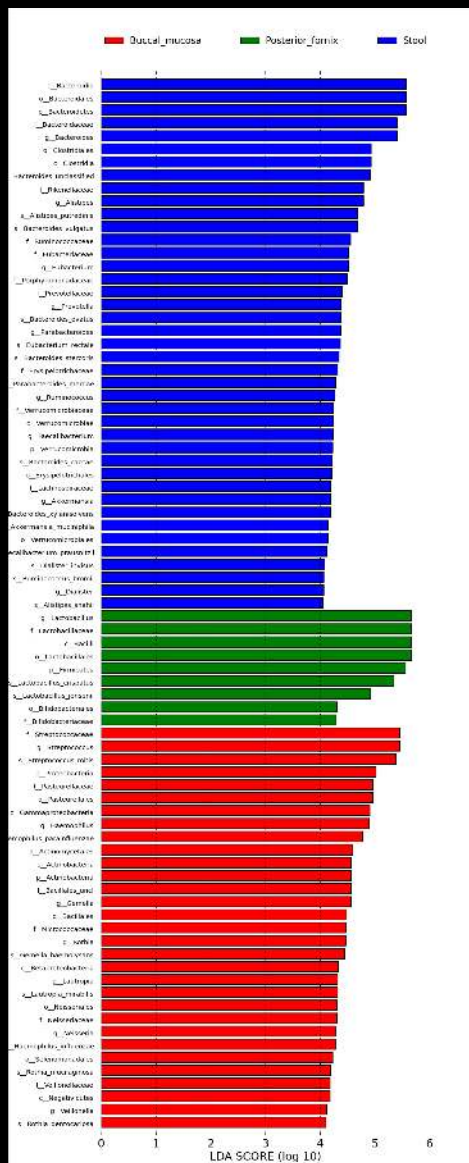
- In Galaxy, view a result by clicking on its “eye”

Click here

The screenshot displays the Galaxy web interface. At the top, the navigation bar includes 'Galaxy / Huttenhower Lab', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A status bar on the right indicates 'Using 0%'. The left sidebar, titled 'Tools', contains a search bar and a list of 'HUTTENHOWER LAB MODULES'. Under the 'LEfSe' module, several options are listed: 'A) Format Data for LEfSe', 'B) LDA Effect Size (LEfSe)', 'C) Plot LEfSe Results', 'D) Plot Cladogram', 'E) Plot One Feature', and 'F) Plot Differential Features'. Other modules listed include 'MetaPhlAn', 'GraPhlAn', 'microPITA', 'MaAsLin', and 'PICRUSt'. The main content area features a green success message: 'A job has been successfully added to the queue - resulting in the following dataset: 5: C) Plot LEfSe Results on data 4'. It provides instructions on checking job status in the 'History' pane. On the right, the 'History' pane shows a list of jobs. The top job is '5: C) Plot LEfSe Results on data 4', which has an eye icon circled in red. Below it are '4: B) LDA Effect Size (LEfSe) on data 3', '3: A) Format Data for LEfSe on data 2', and '2: HMP.ab.filtered.metad ata.txt'. Each job entry includes an eye icon, a pencil icon, and a close icon.



What it means: LEfSe





What it means: LEfSe

- You can plot the results as a cladogram
 - Lots and *lots* of graphical parameters to modify if desired

The screenshot displays the Galaxy web interface for the Huttenhower Lab. The top navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The main content area is divided into three panels:

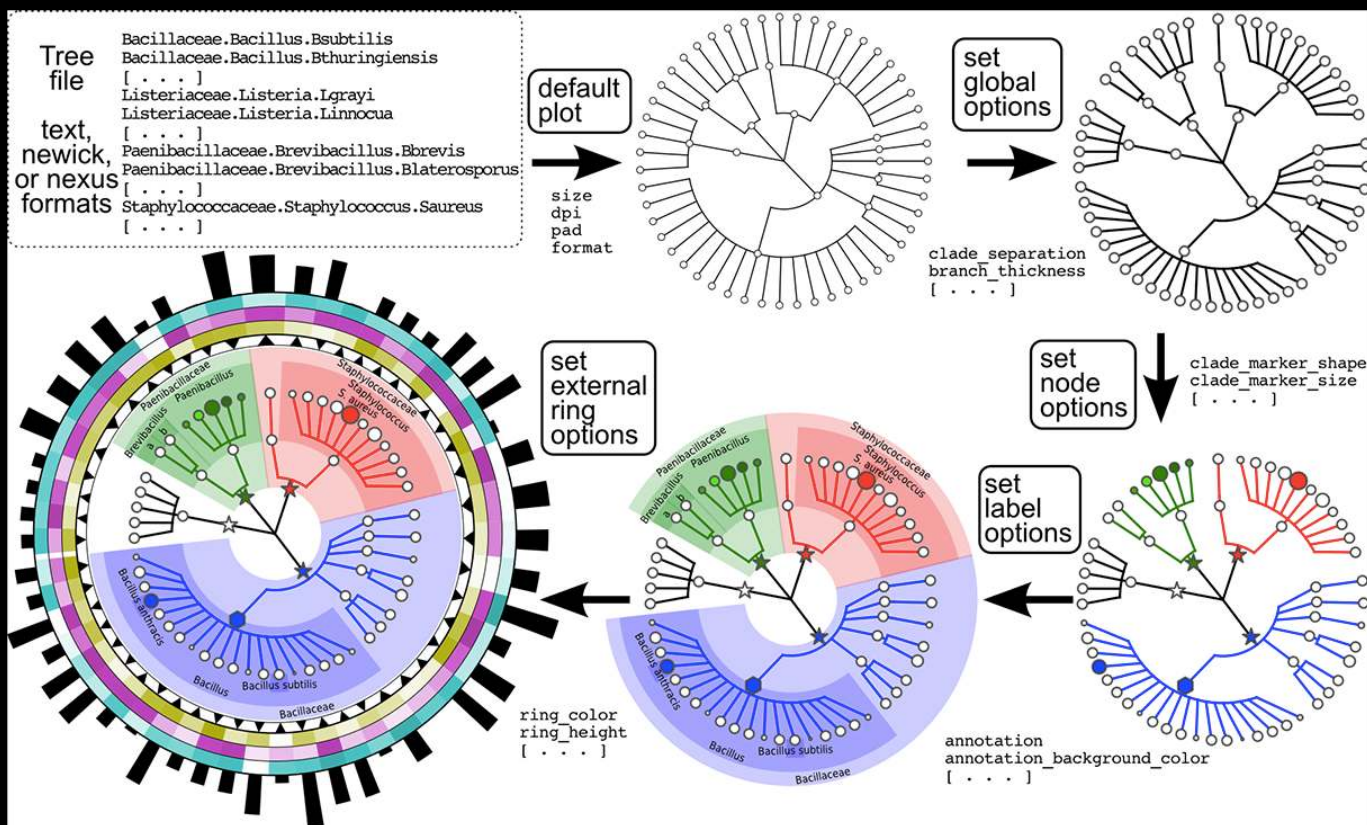
- Tools Panel (Left):** Lists various modules under 'HUTTENHOWER LAB MODULES'. The 'LEfSe' section is expanded, showing options A) through F). 'D) Plot Cladogram' is circled in red, with a red arrow pointing to it from the text '1. Click here'.
- Tool Configuration Panel (Center):** Shows the settings for 'D) Plot Cladogram (version 1.0)'. The 'Select data' dropdown is set to '4: B) LDA Effect Size (LEfSe) on data 3'. Other settings include 'Set structural parameters of the cladogram' (Default), 'Set text and label options' (Default), 'Set some graphical options to personalize the output' (Default), 'Output format' (png), and 'Set the dpi resolution of the output' (150). The 'Execute' button is circled in red, with a red arrow pointing to it from the text '2. Then here'.
- History Panel (Right):** Shows a list of previous jobs. The job '4: B) LDA Effect Size (LEfSe) on data 3' is highlighted in green, indicating it is the current job being configured.



An aside: GraPhlAn

- You can use this visualization for other purposes as well
 - Available online through Galaxy
 - Available offline as open source Python

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





What it means: LEfSe

- Finally, you can see the raw data for individual biomarkers
 - These are generated as a zip file of individual plots

Galaxy / Huttenhower Lab Analyze Data Workflow Shared Data Visualization Help User

Tools

search tools

1. Click here

HUTTENHOWER LAB MODULES

LEfSe

- A) Format Data for LEfSe
- B) LDA Effect Size (LEfSe)
- C) Plot LEfSe results
- D) Plot Cladogram
- E) Plot One Feature
- F) Plot Differential Features**

MetaPhlAn

GraPhlAn

microPITA

MaAsLin

PICRUSt

F) Plot Differential Features (version 1.0)

The formatted datasets:

- 3: A) Format Data for LEfSe on data 2**
- 4: B) LDA Effect Size (LEfSe) on data 3

2. Then selected your formatted data here

The LEfSe output:

- 4: B) LDA Effect Size (LEfSe) on data 3

Do you want to plot all features or only those detected as biomarkers?:

Biomarkers only

Set some graphical options to personalize the output:

Default

Output format:

png

Set the dpi resolution of the output:

150

3. Then here

Execute

History

Unnamed history

1.8 MB

- 6: D) Plot Cladogram on data 4
- 5: C) Plot LEfSe Results on data 4
- 4: B) LDA Effect Size (LEfSe) on data 3
- 3: A) Format Data for LEfSe on data 2
- 2: HMP.ab.filtered.metad ata.txt



What it means: LEfSe

- In Galaxy, download a result by clicking on its “disk”

Galaxy / Huttenhower Lab

Tools

search tools

HUTTENHOWER LAB MODULES

LEfSe

- A) Format Data for LEfSe
- B) LDA Effect Size (LEfSe)
- C) Plot LEfSe Results
- D) Plot Cladogram
- E) Plot One Feature
- F) Plot Differential Features

MetaPhlAn

GraPhlAn

microPITA

MaAsLin

PICRUSt

LOAD DATA MODULE

Get Data

Upload File from your computer

DEFAULT GALAXY MODULES

Convert Formats

FASTA manipulation

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

8: F) Plot Differential Features on data 3 and data 4

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

5.7 MB

8: F) Plot Differential Features on data 3 and data 4

13,522 lines

format: zip, database: ?

Exporting k_Bacteria.p__Firmicutes.c__Clostridia

Exporting k_Bacteria.p__Bacteroidetes.c__Bacteroidia

Exporting k_Bacteria.p__Proteobacteria.c__Betaproteobacteria

binary file

6: D) Plot Cladogram on data 4

5: C) Plot LEfSe Results on data 4

4: B) LDA Effect Size (LDA)

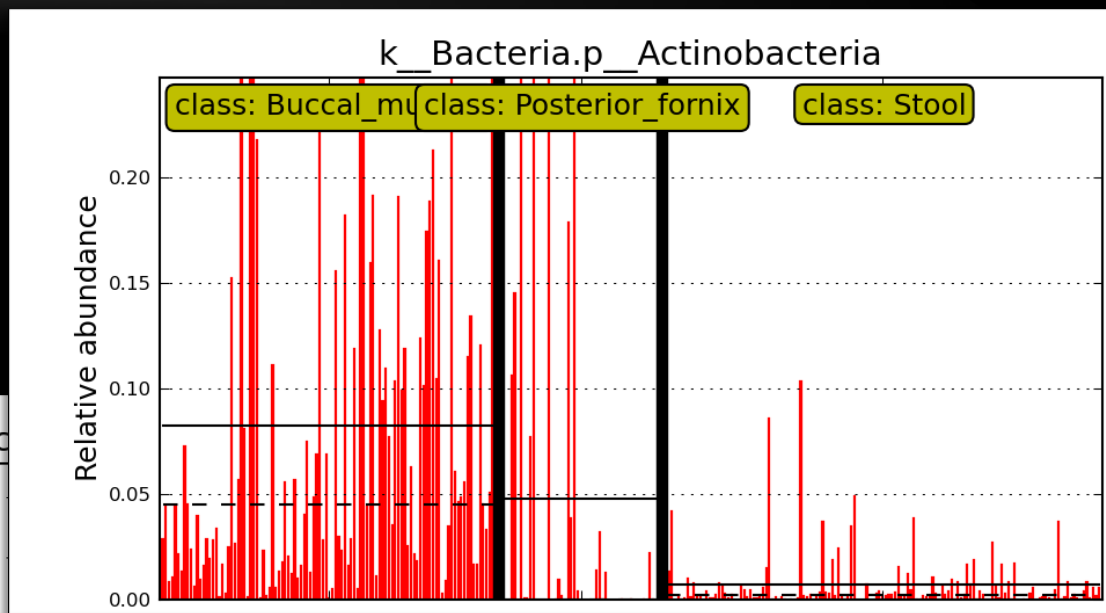
Click here

Then here

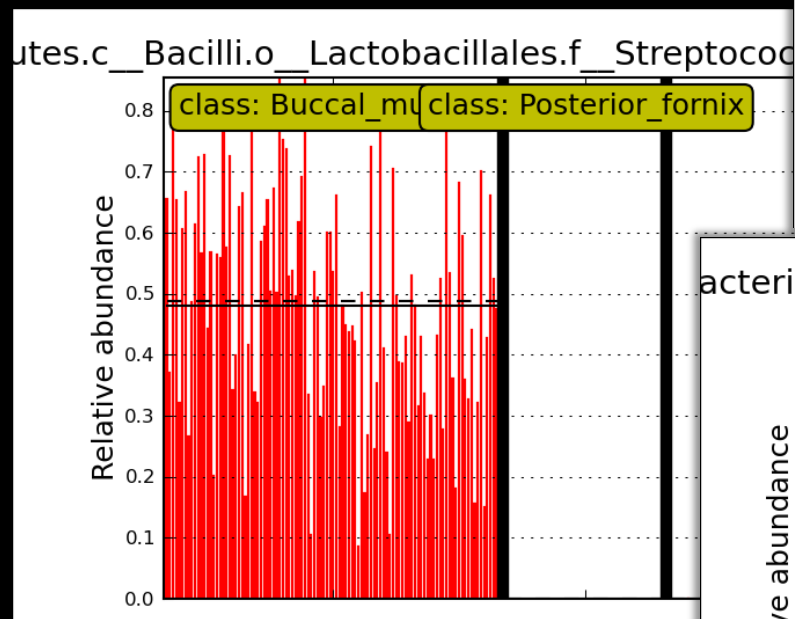


What it means: LEfSe

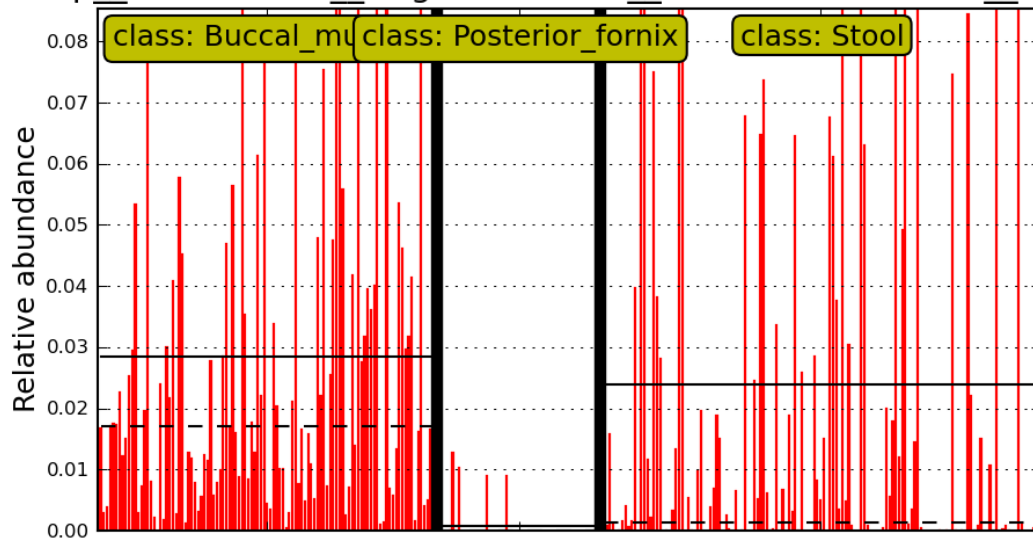
Actinobacteria



Strep. mitis



acteria.p_Firmicutes.c_Negativicutes.o_Selenomonadales.f_Veillonel



Veillonellaceae



Summary

- MetaPhlAn2
 - Evolution of MetaPhlAn1
 - Viruses, euks, subspecies, speed
 - And a LOT more reference data!
 - Raw metagenomic reads in
 - Tab-delimited species relative abundances out
- LEfSe
 - Tab-delimited, stratified relative abundances in
 - Significantly differentially abundant features out



Meta'omic functional profiling with ShortBRED



The two big questions...

Who is there?
(taxonomic profiling)

What are they doing?
(functional profiling)



What's there: ShortBRED



Jim
Kaminski

- **ShortBRED** is a tool for quantifying protein families in metagenomes
 - Short Better REad Dataset
- Inputs:
 - FASTA file of proteins of interest
 - Large reference database of protein sequences (FASTA or blastdb)
 - Metagenomes (FASTA/FASTQ nucleotide files)
- Outputs:
 - Short, unique markers for protein families of interest (FASTA)
 - Relative abundances of protein families of interest in each metagenome (text file, RPKM)
- Compared to BLAST (or HUMAnN), this is:
 - Faster
 - More specific

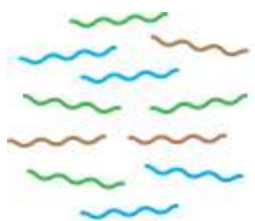


What's there: ShortBRED algorithm

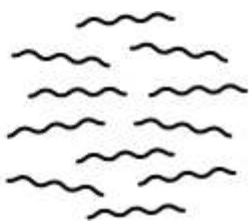
- Cluster proteins of interest into families
 - Record consensus sequences
- Identify and common areas among proteins
 - Compared against each other
 - Compared against reference database
 - Remove all of these
- Remaining subseqs. uniquely ID a family
 - Record these as markers for that family



What's there: ShortBRED marker identification



**Prots of
interest**



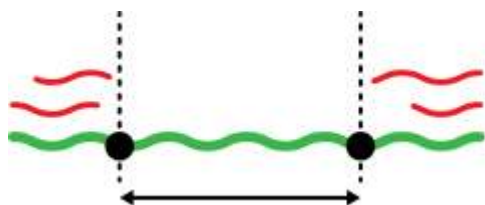
**Reference
database**



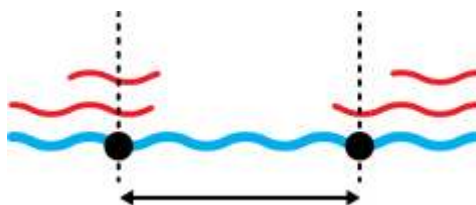
**Cluster into
families**



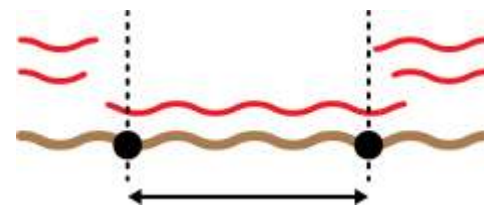
**Identify short,
common regions**



True Marker



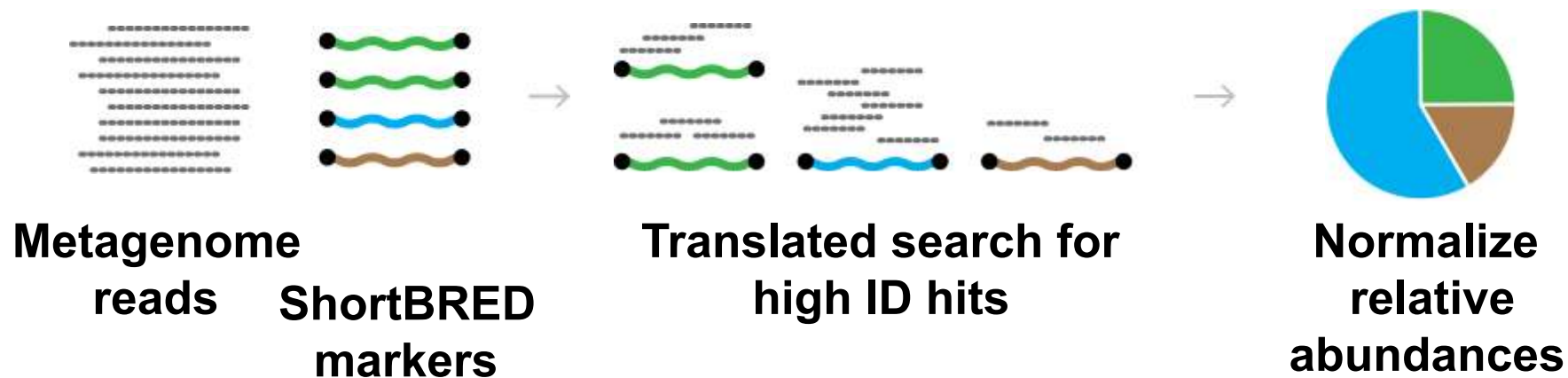
Junction Marker



Quasi Marker



What's there: ShortBRED family quantification





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**.



What's there: ShortBRED

- ShortBRED is available at <http://huttenhower.sph.harvard.edu/shortbred>

The Huttenhower Lab
Department of Biostatistics, Harvard School of Public Health

Contact Documentation People Presentations Publications Research Teaching

Home

You *could* download ShortBRED by clicking **here**

ShortBRED

ShortBRED, the Short Better REad Dataset, is a method for high-precision detection and quantification of functional protein families in microbial communities (metagenomes and metatranscriptomes). It considers a set of protein sequences of interest, reduces them to a set of unique identifying strings ("markers"), and then searches for these markers in metagenomes or metatranscriptomes to very precisely determine the presence and abundance of the original protein families. ShortBRED-Identify clusters the protein sequences into families, removes regions of overlap among the consensus sequences and between the consensus sequences and a set of reference proteins, and saves the remaining sequences as high-confidence unique markers for the families. ShortBRED-Quantify then searches for the markers in unassembled shotgun meta'omic data and returns a normalized relative abundance table of the protein families found in the data.

For more information on the technical aspects to this program or to cite ShortBRED, please reference the following manuscript:

Kaminski J, Gibson M, Franzosa E, Segata N, Dantas L, and Huttenhower C. Fast and accurate meta'omic search with ShortBRED. (In progress)

Download ShortBRED (preliminary version)

Please note that this is a beta version of ShortBRED. An official release will be ready soon.

Download ShortBRED here

You may also install ShortBRED using Mercurial:

```
$ hg clone https://bitbucket.org/biobakery/shortbred
```

More information on the ShortBRED implementation, including runtime documentation, is available at its [Bitbucket page](#).



From the command line...

- But don't!
 - Instead, we've installed ShortBRED already for you
- To see what you can do, run:

```
shortbred_identify.py -h | less
```

```
shortbred_quantify.py -h | less
```



Getting some annotated protein sequences

You could download the ARDB protein sequences **here**

- Go to <http://ardb.cbcb.umd.edu>

ARDB - Antibiotic Resistance Genes Database

[HOME](#) [DOCUMENTATION](#) [BLAST](#) [ADVANCED SEARCH](#) [BROWSE](#)

[Database](#) All Databases Input Search [Help](#) [Tutorial for ARDB](#)

Antibiotic Resistance

Brief introduction to antibiotic resistance.

Analysis & Tools

- ◆ [Single Gene Annotation](#)
- ◆ [Genome Annotation and Comparision](#)
- ◆ [Genome Resistance Profiles Comparison](#)
- ◆ [Mutation Detection](#)

GO Annotation

How to use GO terms to annotate resistance genes?

Welcome to Antibiotic Resistance Genes Database Home Page

Our motivations in creating ARDB are to:

- provide a centralized compendium of information on antibiotic resistance
- facilitate the consistent annotation of resistance information in newly sequenced organisms
- facilitate the identification and characterization of new genes

[More...](#)

News

ARDB is not being maintained at the moment, though we hope to secure funding to further improve it. All underlying data is available for download at: <ftp://ftp.cbcb.umd.edu/pub/data/ARDB/ARDBflatFiles.tar.gz>. Documentation about the provided data is available at <ftp://ftp.cbcb.umd.edu/pub/data/ARDB/doc4ARDBflatFiles.pdf>.

ARDB is recently updated to Version 1.1 on July 3, 2009.

Database Statistics

Version: 1.1
Last Update: July 3, 2009

Genes: 23137
[Types: 380](#)
[Antibiotics: 249](#)
[Genomes: 632](#)
[Species: 1737](#)
[Genera: 267](#)
[Vectors, Plasmids: 2881](#)



From the command line...

- But don't!

- Instead, we've downloaded the important file for you

```
ln -s /home/ubuntu/biobakery/shortbred/data/esisGenes.pfasta
```

```
ln -s /home/ubuntu/biobakery/shortbred/data/esisGenes.pfasta
```

- Take a look by running:

```
less esisGenes.pfasta
```

```
ubuntu@ip-10-170-15-59: ~/galeb
File Edit View Search Terminal Tabs Help
ubuntu@ip-10-170-15-59: ~/galeb
>ZP_02959935 hypothetical protein PROSTU_01837 [Providencia stuartii ATCC 25827].
MGIEYRSLHTSQLTLSEKEALYDLLIEGFEGDFSHDDFAHTLGGMHVMAFDQQKLVGHVA
IIQRHMALDNTPIISVGVEAMVVEQSYRRQIGRQLMLQTNKIIASCYQLGLLSASDDGQ
KLYHSVGWQIWKGKLFELKQGSYIRSIEEEGGVMGWKADGEVDFTASLYCDFRGGDQW
>Q52424 RecName: Full=Aminoglycoside 2'-N-acetyltransferase; AltName: Full=AAC(2')-Ia.
MGIEYRSLHTSQLTLSEKEALYDLLIEGFEGDFSHDDFAHTLGGMHVMAFDQQKLVGHVA
IIQRHMALDNTPIISVGVEAMVVEQSYRRQIGRQLMLQTNKIIASCYQLGLLSASDDGQ
KLYHSVGWQIWKGKLFELKQGSYIRSIEEEGGVMGWKADGEVDFTASLYCDFRGGDQW
>AAA03550 aminoglycoside 2'-N-acetyltransferase [Providencia stuartii].
MGIEYRSLHTSQLTLSEKEALYDLLIEGFEGDFSHDDFAHTLGGMHVMAFDQQKLVGHVA
IIQRHMALDNTPIISVGVEAMVVEQSYRRQIGRQLMLQTNKIIASCYQLGLLSASDDGQ
KLYHSVGWQIWKGKLFELKQGSYIRSIEEEGGVMGWKADGEVDFTASLYCDFRGGDQW
>Q49157 RecName: Full=Aminoglycoside 2'-N-acetyltransferase; AltName: Full=AAC(2')-Ib.
MPFQDVSAPVRGGIILHTARLVHTSDLDQETREGARRMVIEAFEGDFSDADWEHALGGMHA
FICHHGALIAHAADVQRRLRYRDTALRCGYVEAVAVREDWRGQGLATAVMDAVEQVLRGA
YOLGALSASDTARGMYLSRGWLPWOGPTSVLOPAGVTRTPEDDEGLFVLPGVLPAGMELD
```



Getting some reference protein sequences

- Go to <http://metaref.org>

Home About **Download** Help

MetaRef

Keyword Search [Help](#)

Microbial taxonomy

You could download the MetaRef protein sequences **here**

Browse

Bacteria: [2706](#) Genomes

Archaea: [112](#) Genomes

Taxonomy Correction [Info](#)

Highlighted Clades

(Commonly Found in Human Microbiome)

Airways Nares

[Corynebacterium accolens](#)

[Propionibacterium acnes](#)

[Staphylo. epidermidis](#)

Buccal Mucosa

[Gemella haemolysans](#)

[Haemophilus influenzae](#)

[Streptococcus mitis](#)

MetaRef Database v 1.0

MetaRef is a resource to comprehensively catalog and characterize clade-specific microbial genes. We identify and provide all core genes associated with all microbial species and genera with available reference genomes (final or draft). A subset of these gene families are consistently present in one or more taxonomic clades, which allows us to further indicate them as marker genes.

MetaRef paper is now available on [PubMed](#).

Core families: genes present consistently within a clade

Marker families: genes present consistently and exclusively within a clade



Running ShortBRED-Identify

- But don't!
 - We'll use an example mini reference database for speed
- Lets make some antibiotic resistance markers by running:

```
shortbred_identify.py --goi resisGenes.pfasta  
--ref ref_prots.faa --markers ardb_markers.faa
```

- This should take ~5 minutes

- If you get bored waiting, kill it and copy:

```
/class/stamps-shared/biobakery/results/shortbred/ardb_markers.faa
```

- It will produce lots of status output as it runs

```
less ardb_markers.faa
```



ShortBRED markers

ubuntu@ip-10-170-15-59: ~/galeb

File Edit View Search Terminal Help

>ZP_01723236_TM_#01

TEEFLGKYP

>ZP_01723236_TM_#02

IVVMWKRMLSLVGLYKIDGQSQSINRRFNLLHVIVGM

>ZP_01723236_TM_#03

FAFKDFIDHFLFKVEHVYYA

>ZP_01723236_TM_#04

KPKVDSLDKISYGLAF

>ZP_01723236_TM_#05

LVSVLKNWDTLSMDYFGFYAVGFISSEI

>ZP_01723236_TM_#06

ALISKVKLM

>AAA25717_TM_#01

MHLTITYWIDRLREAYPHAVAILLKGSYARGEASAWSIDFDVLVSDEEEVEEYRTWIEPV

GERLVHISVAVIEWVTGWERDSADPSSWSYGLPTQETTQLLWAADENIRRRRLDRPFKVHPA

AEPEVEDTVEALGKIRNAMVRGDDLAVYQAAQVVGKLIPTLLVPINPPTYARFAREAIDR

ILAFPNVPEGFAADWLTCMGLVDRRTHDPQPTRPNEWCAARSRFCRRMRTSSVRISRGCW

KQDWYLRISART

>NP_880590_TM_#01

HTPGDAPGAADDTASDERA

>NP_880590_TM_#02

AHTLEQIS

>NP_880590_TM_#03

KQALGVGVAQC

>NP_880590_TM_#04

:

True Markers
at the top



ShortBRED markers

```
ubuntu@ip-10-170-15-59: ~/galeb
File Edit View Search Terminal Help

SECS
>P14509_TM_#04
IEAGVVDVDDFDKEREGWTAEQVWEAMHRLPLA
>P14509_TM_#05
LIVEGKVVGCIDVGRAGIADRYQDLAVLWNCLEEFESLQERLVAQYGIADPDRR
>1112175A_JM_#01__[1112175A_w=0.486,YP_001103000_w=0.143,YP_0011
03000_w=0.371]
LFEWFEVVEKVDSAIMRLRRRAEPLLEGAALERYE
>1112175A_JM_#02__[1112175A_w=0.515,YP_001103000_w=0.333,YP_0011
03000_w=0.152]
RKYPRRRVEAAFDHAGVGGGAVVAYVRPEQWLRL
>ABF69686_JM_#01__[ABF69686_w=0.459,ABN80187_w=0.135,ZP_03989103
_w=0.405]
DTAYPGEIVILADDTLKLNDILGNEKLLPHKTRI
>NP_792892_QM33_#01__[NP_792892_w=0.500,YP_002081505_w=0.500]
PAAFISGLTGQFYKQFALTIAISTVISAFNSLT
>YP_002081505_JM_#01__[YP_002081505_w=0.630,NP_792892_w=0.370]
LGTIGGFRLQIEDRGNX
>YP_970399_JM_#01__[YP_970399_w=0.306,ZP_03552050_w=0.163,YP_997
055_w=0.163,YP_997055_w=0.102,CAJ93947_w=0.061,YP_001348697_w=0.
061,YP_316450_w=0.041,YP_002092118_w=0.061,Q2KX31_w=0.041]
GGMLLGLSRKAATDX
>ZP_01817983_JM_#01__[ZP_01817983_w=0.493,YP_001694417_w=0.362,Y
P_001694417_w=0.145]
TLTGPFIGGFIKEDFQPVAKEKAIPTKELFTSVK
(END)
```

Junction/Quasi Markers
at the bottom



Running ShortBRED-Quantify

- Using your existing HMP data subset, you can search for antibiotic resistance proteins in the oral cavity by running:

```
shortbred_quantify.py  
  --markers ardb_markers.faa  
  --wgs 763577454-SRS014472-Buccal_mucosa.fasta  
  --results 763577454-SRS014472-Buccal_mucosa-ARDB.txt
```

- This should just a few seconds
- It will again produce lots of status output as it runs

```
less 763577454-SRS014472-Buccal_mucosa-ARDB.txt
```



ShortBRED marker quantification

ubuntu@ip-10-170-15-59: ~/galeb

File Edit View Search Terminal Help

Family	Count	Hits	TotMarkerLength
YP_001694417	2380.9523809523807	1	26
ZP_04679156	0.0	0	235
ZP_04657259	0.0	0	136
ZP_04635798	0.0	0	91
ZP_04635523	0.0	0	171
ZP_04633951	0.0	0	59
ZP_04616832	0.0	0	9
ZP_04613685	0.0	0	72
ZP_04606269	0.0	0	183
ZP_04577926	0.0	0	168
ZP_04543635	0.0	0	173
ZP_04543532	0.0	0	186
ZP_04433866	0.0	0	187
ZP_04431003	0.0	0	95
ZP_04405580	0.0	0	169
ZP_04405450	0.0	0	300
ZP_04309403	0.0	0	138
ZP_04284182	0.0	0	177
ZP_04244950	0.0	0	51
ZP_04210257	0.0	0	113
ZP_04197552	0.0	0	129
ZP_04175489	0.0	0	70
ZP_04174269	0.0	0	21
ZP_04151022	0.0	0	27
:			

RPKMs and raw hit count

Other columns are family name and AA marker length



AR proteins in the human gut

- That's boring! Let's get some real data
- `cp` this file to your own working directory:

`/home/ubuntu/biobakery/shortbred/data/shortbred_ardb_hmp_t2d.tsv`

- This is the result of running:
 - ShortBRED-Identify on the real ARDB + reference
 - ShortBRED-Quantify on the real HMP + T2D data (Qin Nature 2014, PMID: 25079328)
 - Summing each sample's RPKMs for families in each ARDB resistance class



AR proteins in the human gut

shortbred_ardb_hmp_t2d.tsv - LibreOffice Calc

File Edit View Insert Format Tools Data Window Help

Liberation Sans 10

A1 f(x) Σ = Sample.ID

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Sample.ID	HMP1	HMP2	HMP3	HMP4	HMP5	HMP6	HMP7	HMP8	HMP9	HMP10	HMP11	HMP12	HMP13	HMP14	HMP15	HMP16	HMP17
2	Dataset	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP
3	Gender	Female	Male	Female	Male	Female	Female	Male	Male	Female	Female	Female	Male	Male	Male	Female	Male	Male
4	ABC Antibiotic Efflux	0	0.6097114	0.53837173	0	0	0.05083452	0	0	18.879238	0.3999418	0.6375002	0.11029351	0	0	0.1499069	3.3238466	0
5	Aminoglycoside Acetyltransferase	0	0	0	0.5570841	0	0	0	0	0	0.4844142	0	0	0	7.15621993	0	0	0
6	Aminoglycoside Nucleotidyltransferase	11.88478263	2.3493412	1.31127279	2.1879248	1.70197254	25.23425383	0	1.4888313	6.7524558	11.6664297	0.2944691	0	0.54364476	22.13646686	1.0549423	6.1159491	0
7	Aminoglycoside Phosphotransferase	0.72342527	9.510191	0.43478001	9.31863091	1.44994258	21.76497663	0	0	1.8219867	1.9941331	0.7220629	1.82419711	0	1.09356043	1.6969943	5.382002	0
8	Antibiotic Target	0	0.4319648	0	0	0.11002037	0	0	0	0.1044046	0	0.6096981	4.45863298	0	0	0.1242086	0	0
9	Chloramphenicol Acetyltransferase CAT	0	0.8931758	0.50566409	0.06863132	0	0	0	0	0.2300411	0.2286945	0	0	0	0	0	0	0
10	Chloramphenicol MFS Efflux Pump	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	Chloramphenicol Phosphotransferase CPT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	Class A Beta-Lactamase	11.96165378	14.1741569	192.7320267	57.34211706	30.37844852	36.47564233	41.445191	77.8068337	27.5978829	84.7152993	29.5138602	4.47890136	7.54656885	6.17723545	67.6346059	121.5428999	0
13	Class B Beta-Lactamase	0.73757867	0.4730655	0	0.35938332	0.22651252	0.45452038	0	0.1196987	1.5652141	0.5770399	0	0	0	0	0	0	0
14	Class C Beta-Lactamase	0	0	0	0	0	0	0	0	0.4758603	0.2556631	0	0	0	0	0	0.1458178	0
15	Class D Beta-Lactamase	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	Gene Modulating Antibiotic Efflux	0	0	0.12940327	0	0	0	0	0	0	2.6860575	0.3513343	0.52138395	0.18121492	0.09719297	0	0.6224941	0
17	Gene Modulating Resistance	0	0	0.53609928	0.10341706	0.28813026	0	0	0.1033344	0	0.4529638	0	0.59939377	0	0.73268549	0	0	0
18	Glycopeptide Resistance	0	0.1148873	0.10721986	2.91192901	11.82529267	1.06129011	0	1.475885	0	3.8329823	0.2028631	0.17855513	0	2.57636295	0	12.8763448	0
19	Uncosamide Resistance	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	Macrolide Resistance	0	0	0	0	0	0	0	0	0	0	0.2216556	0	0	0	0	0	0
21	MATE Antibiotic Efflux	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	MFS Antibiotic Efflux	0	0.1079916	2.44436309	2.24124166	0.15717195	19.64826671	0	0	0	6.0081483	4.73637	0.16432993	0	9.88061341	0.2382082	43.436675	0
23	Other ARG	0	0.1641248	1.50507872	4.90492355	0.80462657	0.27160156	0	0.4618416	1.2797248	2.911427	1.0099704	0.79420864	0	0.21818147	0.3167416	0.7025792	0
24	Puromycin Resistance	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	Quinolone Resistance	0	0	0.05601037	0.09933481	0.05066727	0.05083452	0	0	0.8647162	0.1335553	3.29844229	0.06626516	0.6266389	0	0.1841579	0	0
26	Rifamycin Resistance	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	RND Antibiotic Efflux	1.11005589	0.2116346	0.87820136	0.51112275	1.80007009	12.407319	34.237278	3.5262745	38.781576	4.5900824	1.9670192	0.17668244	38.00414096	1.38795841	0.7786209	2.9700758	0
28	rRNA Methyltransferase	5.61799582	6.0194576	37.23691651	9.44289101	34.61725215	94.72884389	2.051664	80.7900949	122.9478456	2.4135554	10.2418695	0.06217665	7.23364421	13.9417838	130.7374941	96.9503344	0
29	SMR Antibiotic Efflux	0	0	0	0	0	0	0	0	0	0.876332	0	0.08288129	0	0.19222828	0	0.2560272	0
30	Streptogramin Resistance	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	Tetracycline Efflux	0.06843748	2.6183624	0.57325559	0.86505449	12.89081881	0.16675423	2.793598	0.359161	0.5939219	2.0434753	2.4886453	0.33754257	0.23247387	0	0.9097696	2.3449461	0
32	Tetracycline Inactivation	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	Tetracycline Other	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	Tetracycline Ribosomal Protection	56.71076788	123.9169955	323.8053962	213.4139838	412.0286699	339.8911536	158.290041	150.2523103	171.9304394	178.3934869	157.4626561	210.9729322	104.6340794	117.7276424	159.6264918	121.7531755	0
35																		
36																		
37																		

Sheet1/

Find

Sheet 1 / 1

Default

STD

Sum=0

100%



What it means: LEfSe

- Visit LEfSe at: <http://huttenhower.sph.harvard.edu/lefse>

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

Tools

search tools

HUTTENHOWER LAB MODULES

LEfSe

- A) Format Data for LEfSe
- B) LDA Effect Size (LEfSe)
- C) Plot LEfSe Results
- D) Plot Cladogram
- E) Plot One Feature
- F) Plot Differential Features

MetaPhlAn

GraPhlAn

microPITA

MaAsLin

PICRUSt

Get Data

Upload File from your computer

First click here

Thanks for visiting our lab's tools and applications page, implemented within the [Galaxy](#) web application and workflow framework. Here, we provide a number of resources for metagenomic and functional genomic analyses, intended for research and academic use. Please see the menus and folders to the left for an overview of available tools including documentation, sample data, and publications.

Our lab's research interests include metagenomics and the [human microbiome](#), the relationships between microbial communities and human health, microbiome systems biology, and large-scale computational methods for studying all of these areas. In addition to the tools provided here, feel free to take a look at our additional [research](#) and [publications](#), including the [Sleipnir library](#) for computational functional genomics.

The tools are available here without account creation. However, you are strongly invited to create an account for having access to the history, saved analyses, datasets and workflows. You can create an account and/or log in using the User menu in the top-right corner.

If you have any comments, questions, or suggestions, please contact [Dr. Huttenhower](#).

History

Unnamed history

0 bytes

This history is empty. You can [load your own data](#) or [get data from an external source](#)



What it means: LefSe

- Then upload your formatted table
 - After you upload, wait for the progress meter to turn green!

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

Tools search tools

HUTTENHOWER LAB MODULES

LefSe

- A) Format Data for LefSe
- B) LDA Effect Size (LefSe)
- C) Plot LefSe Results
- D) Plot Cladogram
- E) Plot One Feature
- F) Plot Differential Features

MetaPhlAn

GraPhlAn

microPITA

MaAsLin

PICRUSt

LOAD DATA MODULE

Get Data

- Upload File from your computer

DEFAULT GALAXY MODULES

Upload File (version 1.1.4)

File Format:

Auto-detect

Which format? See help below

File:

Choose File shortbred_aradb_hmp_t2d.tsv

For browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Convert spaces to tabs:

☐ Yes

Use this option if you are entering intervals by hand.

Genome:

unspecified (?)

Execute

History

Unnamed history

269.2 KB

This history is empty. You can [load your own data](#) or [get data from an external source](#)



What it means: LEfSe

- Then tell LEfSe about your metadata:

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

Tools

search tools

HUTTENHOWER LAB MODULES

LEfSe

A) Format Data for LEfSe

B) LDA Effect Size (LEfSe)

C) Plot LEfSe Results

D) Plot Cladogram

E) Plot One Feature

F) Plot Differential Features

MetaPhlAn

GraPhlAn

microPITA

MaAsLin

PICRUSt

LOAD DATA MODULE

Get Data

Upload File from your computer

A) Format Data for LEfSe (version 1.0)

Upload a tabular file of relative abundances and class labels (possibly also subclass and subjects labels) for LEfSe – See samples below – Please use Galaxy Get-Data/Upload-File. Use File-Type = Tabular:

1: shortbred_aradb_hmp_t2d.tsv

Select whether the vectors (features and meta-data information) are listed in rows or columns:

Rows

Select which row to use as class:

#2:Dataset

Select which row to use as gender:

#3:Gender

Select which row to use as sample ID:

#1:Sample.ID

Per-sample normalization of the sum of the values to 1M (recommended when very low values are present):

Yes

Execute

History

Unnamed history

41.8 KB

1: shortbred_aradb_hmp_t2d.tsv

1. Click here

2. Then select Dataset

3. Then Gender

4. Then SampleID

5. Then click here



What it means: LEfSe

- Leave all parameters on defaults, and run LEfSe!
 - You can try playing around with these parameters if desired

The screenshot shows the Galaxy web interface with the 'Huttenhower Lab' header. The 'Tools' panel on the left lists various modules, with 'B) LDA Effect Size (LEfSe)' circled in red. A red arrow points from the text '1. Click here' to this option. The main panel displays the configuration for 'B) LDA Effect Size (LEfSe) (version 1.0)'. It includes fields for 'Select data' (set to '2: A) Format Data for LEfSe on data 1'), 'Alpha value for the factorial Kruskal-Wallis test among classes' (0.05), 'Alpha value for the pairwise Wilcoxon test between subclasses' (0.05), 'Threshold on the logarithmic LDA score for discriminative features' (2.0), a checkbox for 'Do you want the pairwise comparisons among subclasses to be performed only among the subclasses with the same name?' (set to 'No'), and a dropdown for 'Set the strategy for multi-class analysis' (set to 'All-against-all (more strict)'). The 'Execute' button at the bottom is circled in red, with a red arrow pointing from the text '2. Then GO!' to it. The right panel shows a 'History' section with two entries: '2: A) Format Data for LEfSe on data 1' and '1: shortbred ardb hmp t 2d.tsv'.



What it means: LEfSe

- You can plot the results as a bar plot
 - Again, lots of graphical parameters to modify if desired

The screenshot displays the Galaxy web interface for Huttenhower Lab. The top navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A 'Using 0%' status indicator is visible in the top right.

Tools Panel (Left): A search bar is at the top. Under 'HUTTENHOWER LAB MODULES', the 'LEfSe' section is expanded. The tool 'C) Plot LEfSe Results' is circled in red. Other tools listed include 'A) Format Data for LEfSe', 'B) LDA Effect Size (LEfSe)', 'D) Plot Cladogram', 'E) Plot One Feature', and 'F) Plot Differential Features'. Below this, 'MetaPhlAn' and 'GraPhlAn' are also listed.

Tool Interface (Center): The 'C) Plot LEfSe Results (version 1.0)' tool is selected. It shows a 'Select data' dropdown with '5: B) LDA Effect Size (LEfSe) on data 2' chosen. Below are sections for 'Set text and label options', 'Set some graphical options to personalize the output', 'Output format' (set to 'png'), and 'Set the dpi resolution of the output' (set to '150'). The 'Execute' button at the bottom is circled in red.

History Panel (Right): Titled 'History', it shows a list of previous jobs. The top entry is '5: B) LDA Effect Size (LEfSe) on data 2'. Below it are '2: A) Format Data for LEfSe on data 1' and '1: shortbred ardb hmp t 2d.tsv'. Each entry has icons for viewing, editing, and deleting.

Annotations: Red text and arrows provide instructions: '1. Click here' points to the 'C) Plot LEfSe Results' tool, and '2. Then here' points to the 'Execute' button.



What it means: LEfSe

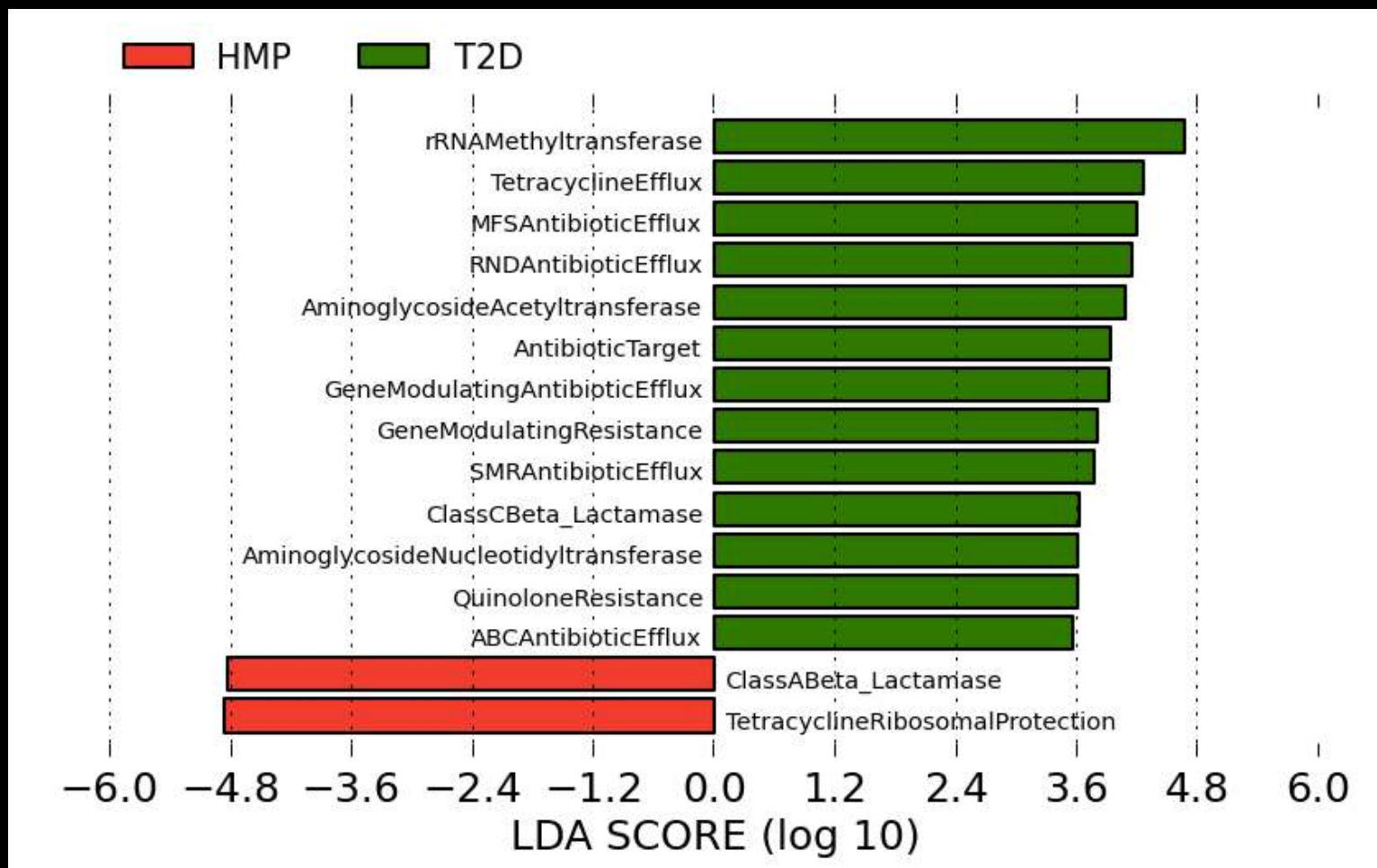
- In Galaxy, view a result by clicking on its “eye”

Click here

The screenshot displays the Galaxy web interface. At the top, the navigation bar includes 'Galaxy / Huttenhower Lab', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A 'Using 0%' indicator is on the right. The left sidebar shows 'Tools' with a search bar and 'HUTTENHOWER LAB MODULES' including 'LEfSe' and several sub-tools like 'Format Data for LEfSe', 'LDA Effect Size (LEfSe)', 'Plot LEfSe Results', 'Plot Cladogram', 'Plot One Feature', and 'Plot Differential Features'. Other modules listed are 'MetaPhlAn', 'GraPhlAn', 'microPITA', 'MaAsLin', and 'PICRUSt'. The main content area features a green success message: 'A job has been successfully added to the queue - resulting in the following dataset: 6: C) Plot LEfSe Results on data 5'. It provides instructions on checking job status in the 'History' pane. On the right, the 'History' panel lists several jobs. Job 6, '6: C) Plot LEfSe Results on data 5', is highlighted with a red circle around its 'eye' icon, which is pointed to by a red arrow from the text 'Click here'. Other jobs listed include '5: B) LDA Effect Size (LEfSe) on data 2', '2: A) Format Data for LEfSe on data 1', and '1: shortbred ardb hmp t 2d.tsv'.



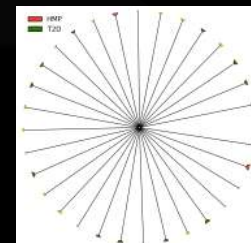
What it means: LEfSe





What it means: LEfSe

- There's no really any reason to plot a cladogram
 - Although it will work!
- But you can see the raw data for individual biomarkers
 - These are generated as a zip file of individual plots



Galaxy / Huttenhower Lab Analyze Data Workflow Shared Data Visualization Help User

Tools

search tools

1. Click here

HUTTENHOWER LAB MODULES

LEfSe

- A) Format Data for LEfSe
- B) LDA Effect Size (LEfSe)
- C) Plot LEfSe results
- D) Plot Cladogram
- E) Plot One Feature
- F) Plot Differential Features**

MetaPhlAn
GraPhlAn
microPITA
MaAsLin
PICRUST

F) Plot Differential Features (version 1.0)

The formatted datasets:

- 3: A) Format Data for LEfSe on data 2**

The LEfSe output:

- 4: B) LDA Effect Size (LEfSe) on data 3

Do you want to plot all features or only those detected as biomarkers?:
Biomarkers only

Set some graphical options to personalize the output:
Default

Output format:
png

Set the dpi resolution of the output:
150

2. Then selected your formatted data here

3. Then here

Execute

History

Unnamed history
1.8 MB

- 6: D) Plot Cladogram on data 4
- 5: C) Plot LEfSe Results on data 4
- 4: B) LDA Effect Size (LEfSe) on data 3
- 3: A) Format Data for LEfSe on data 2
- 2: HMP.ab.filtered.metad ata.txt



What it means: LEfSe

- In Galaxy, download a result by clicking on its “disk”

The screenshot displays the Galaxy web interface. At the top, the header reads "Galaxy / Huttenhower Lab" with navigation tabs for "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User". A status bar on the right indicates "Using 0%".

On the left sidebar, under "HUTTENHOWER LAB MODULES", the "LEfSe" section is expanded, listing several tools: "A) Format Data for LEfSe", "B) LDA Effect Size (LEfSe)", "C) Plot LEfSe Results", "D) Plot Cladogram", "E) Plot One Feature", and "F) Plot Differential Features". Below this, other modules like "MetaPhlAn", "GraPhlAn", "microPITA", "MaAsLin", and "PICRUSt" are listed, along with a "LOAD DATA MODULE" section containing "Get Data".

The main content area features a green notification box with a checkmark icon, stating: "A job has been successfully added to the queue – resulting in the following dataset: 8: F) Plot Differential Features on data 2 and data 5". It further instructs users to check the status of queued jobs and view the resulting data by refreshing the History pane.

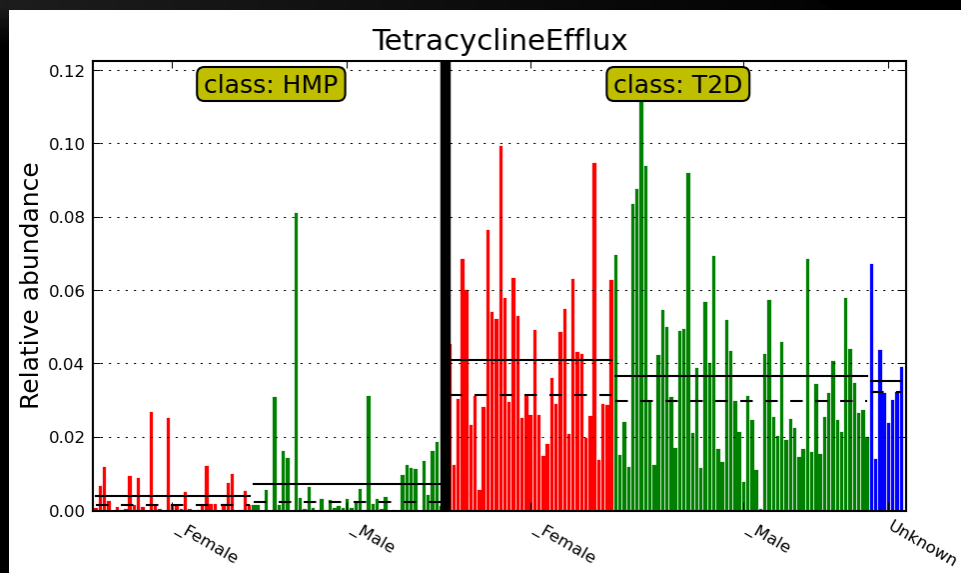
On the right, the "History" panel shows a list of jobs. The job "8: F) Plot Differential Features on data 2 and data 5" is highlighted with a red circle. Below the job name, it shows "2,363 lines" and "format: zip, database: ?". A list of exported features is visible: "Exporting MFSAntibioticEfflux", "Exporting ClassCBeta_Lactamase", "Exporting AminoglycosideAcetyltransferase", "Exporting rRNAMethyltransferase", "Exporting ClassABeta_Lactamase", "Exporting AntibioticTarget", "Exporting TetracyclineRibosomalProtection", and "Exporting G". At the bottom of the history panel, a red circle highlights a download icon (a disk symbol) next to the text "binary file".

Red annotations are present: "Click here" with an arrow pointing to the job name in the history panel, and "Then here" with an arrow pointing to the download icon.

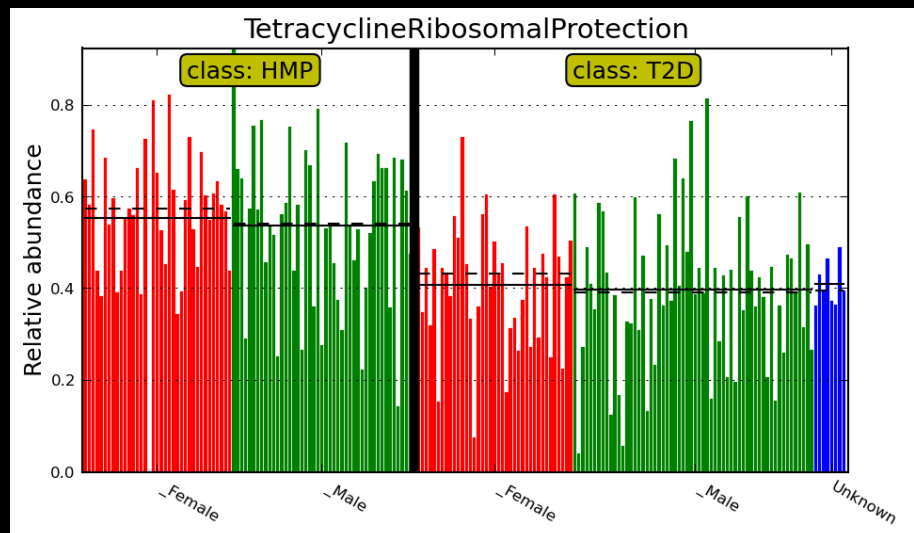


What it means: LEfSe

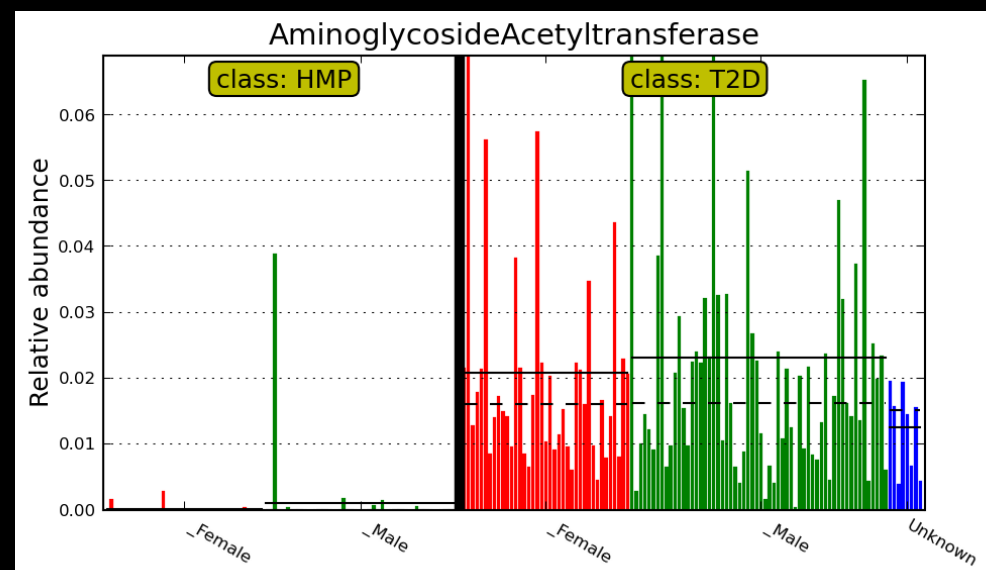
Tetracycline
Efflux
Pumps



Tet. Ribosomal Blockers



Aminoglycoside
Acetyltransferases





Summary

- ShortBRED
 - Raw metagenomic reads,
Proteins of interest, and
Protein reference database in
 - Tab-delimited gene family rel. abundances out



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

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 wiki provides tutorials for Huttenhower tools, illustrating through demos how to use these tools on your datasets. Huttenhower tools can be divided under three main categories as shown below. Click on the tool for the corresponding tutorial.

Composition Analysis

These tools can determine the composition in terms of (i) microbial species and their associated abundances (MetaPhlAn) or (ii) genes and associated pathways (HUMANN) in the dataset. Please click on the links below for detailed tutorials:

HUMANN

- Microbial species and associated genes and pathways

MetaPhlAn

- Microbial species and abundances

PhyloPhlAn

- Reconstruction of phylogenetic trees

PICRUSt

- Predict metagenome functional content from marker gene

ShortBRED

- Abundance of proteins of interest in genetic data

Statistical Analysis

These tools can determine the associations from the provided metadata information and microbial composition tables. Please click on the links below for detailed tutorials:

AREP

- Extract 'omics data from repositories

CCREPE

- Assess the significance of general similarity measures in compositional datasets

LEfSe

- Association between metadata (max 2) and microbial species and abundances

MaAsLin

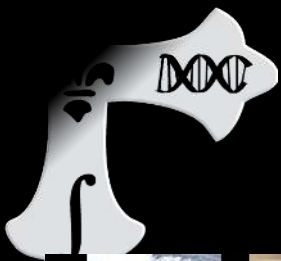
- Association between metadata (no restriction) and microbial species and abundances

microPITA

- Sample selection in two stage-tiered studies

Visualization

These tools can help visualize taxonomical and phylogenetic information for (i) microbial composition/taxonomy data, (ii) outputs from MetaPhlAn, LEfSe, HUMANN, MaAsLin. Please click on the link below for detailed tutorial:



Thank you!



Curtis
Huttenhower



Xochitl
Morgan



Afrah
Shafquat



Keith
Bayer



George
Weingart



Regina
Joice



Aleksandar
Kostic



Chengwei
Luo



Tiffany
Hsu



Emma
Schwager



Koji
Yasuda



Kevin
Oh



Boyu
Ren



Andy
Shi



Jim
Kaminski



Joseph
Moon



Randall
Schwager



Levi Waldron



Nicola Segata



Wendy Garrett
Michelle Rooks



Dirk Gevers
Kat Huang



Ramnik Xavier
Harry Sokol
Dan Knights
Moran Yassour



Rob Beiko
Morgan Langille



Jacques Izard
Katherine Lemon



Ruth Ley
Omry Koren



Human Microbiome Project

Owen White	Sahar Abubucker
Joe Petrosino	Brandi Cantarel
George Weinstock	Alyx Schubert
Karen Nelson	Mathangi Thiagarajan
Lita Proctor	Beltran Rodriguez-Mueller
Erica Sodergren	Makedonka Mitreva
Anthony Fodor	Yuzhen Ye
Marty Blaser	Mihai Pop
Jacques Ravel	Larry Forney
Pat Schloss	Barbara Methe

Bruce Birren Mark Daly
Doyle Ward Ashlee Earl



Rob Knight
Greg Caporaso
Jesse Zaneveld



Mark Silverberg
Boyko Kabakchiev
Andrea Tyler



Bruce Sands



