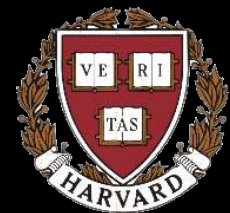




Meta'omic taxonomic profiling with MetaPhlAn2 and biomarker discovery with LEfSe

Curtis Huttenhower

01-19-15



Harvard School of Public Health
Department of Biostatistics



U. Oregon



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:

AREPA

- 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

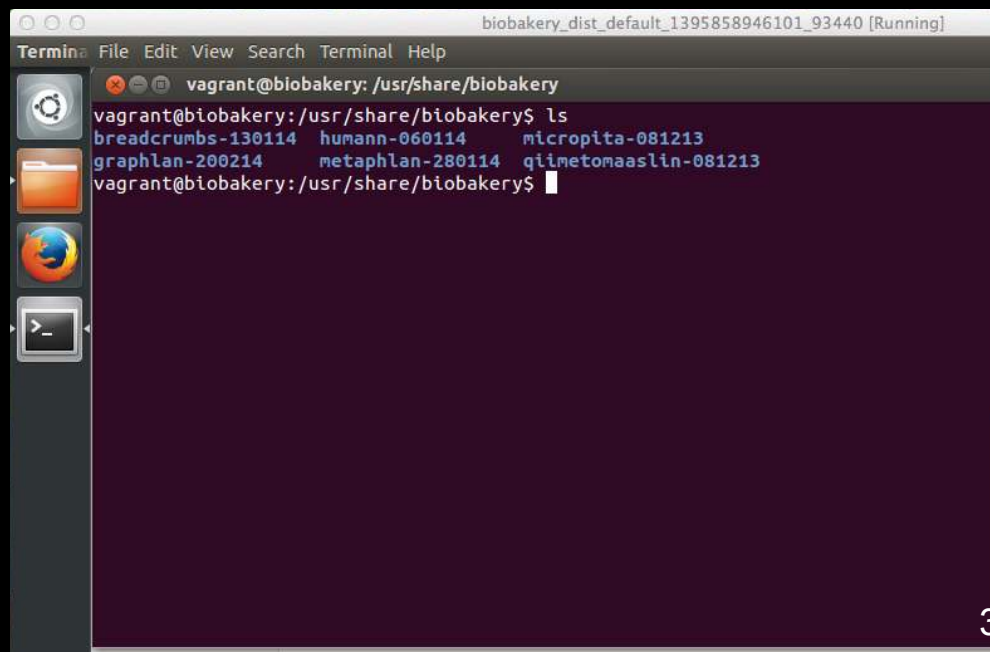
These tools can help visualize taxonomical and phylogenetic information for (i) microbial composition/taxonomy data, (ii) outputs from MetaPhlAn, LEfSe, HUMAnN, MaAsLin. Please click on the link below for detailed tutorial:



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)



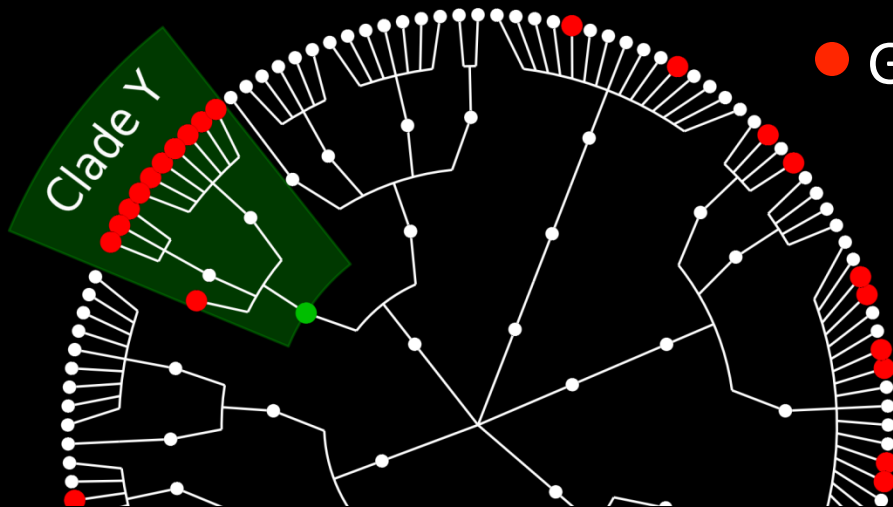
MetaPhlAn overview

Nicola Segata

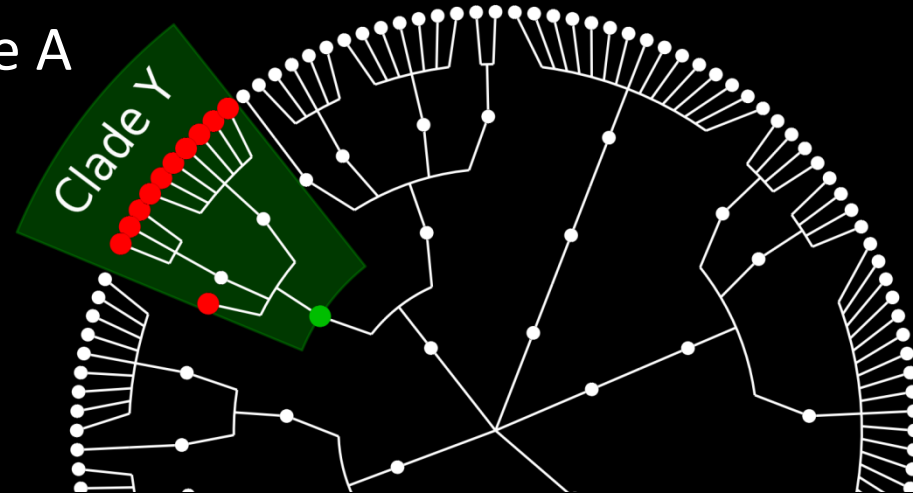


A is a **core gene** for clade Y

A is a **unique marker gene** for clade Y



● Gene A



ChocoPhlAn (offline pipeline)

- Identify all **core genes** for all clades
- Screen core genes for **unique marker genes**
- Select most representative marker genes

Unique
marker
genes DB

Available
reference
genomes

MetaPhlAn

Metagenome

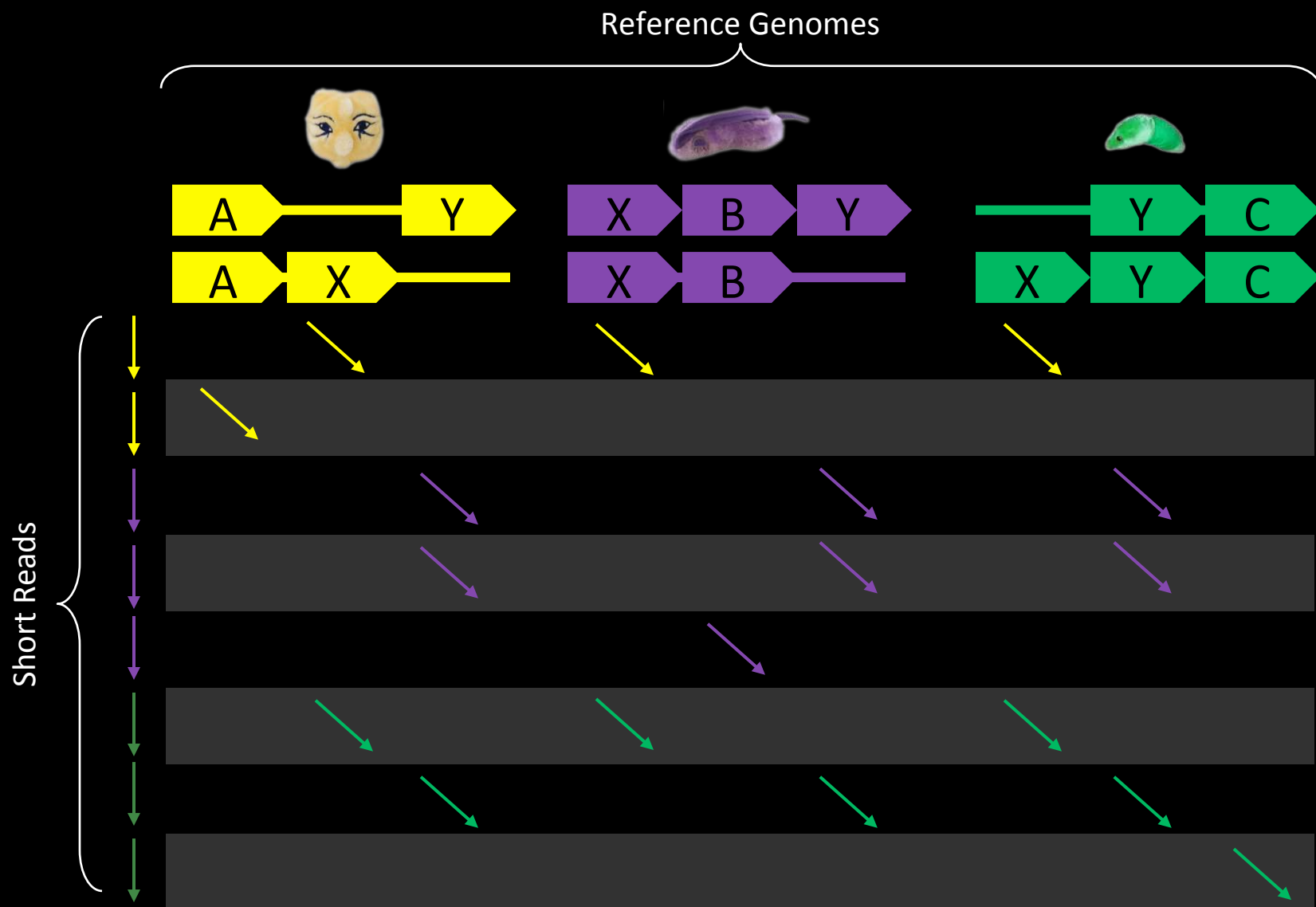
- Blast reads against the marker genes
- Assign, count, normalize reads





MetaPhlAn

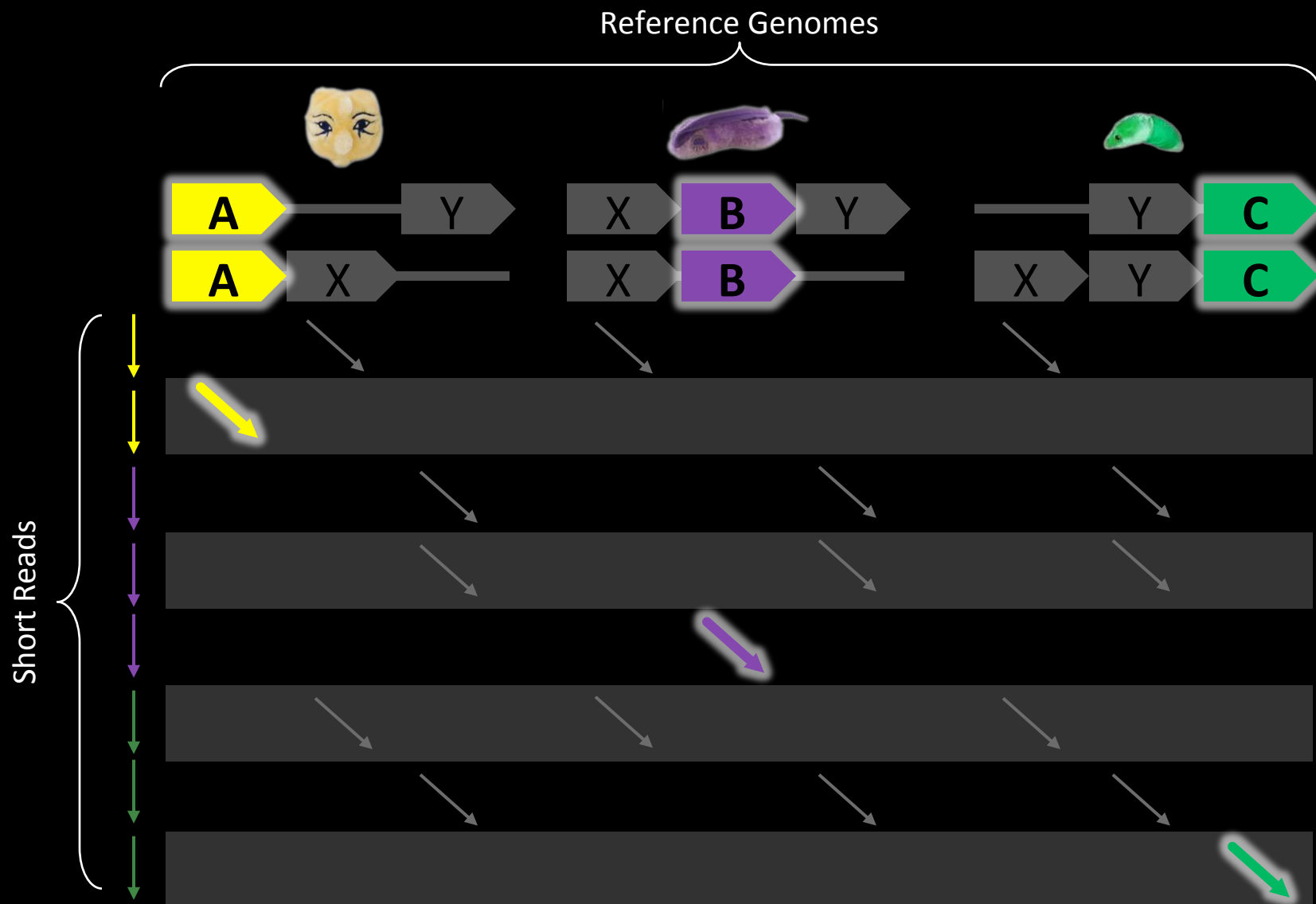
Metagenomic Phylogenetic Analysis





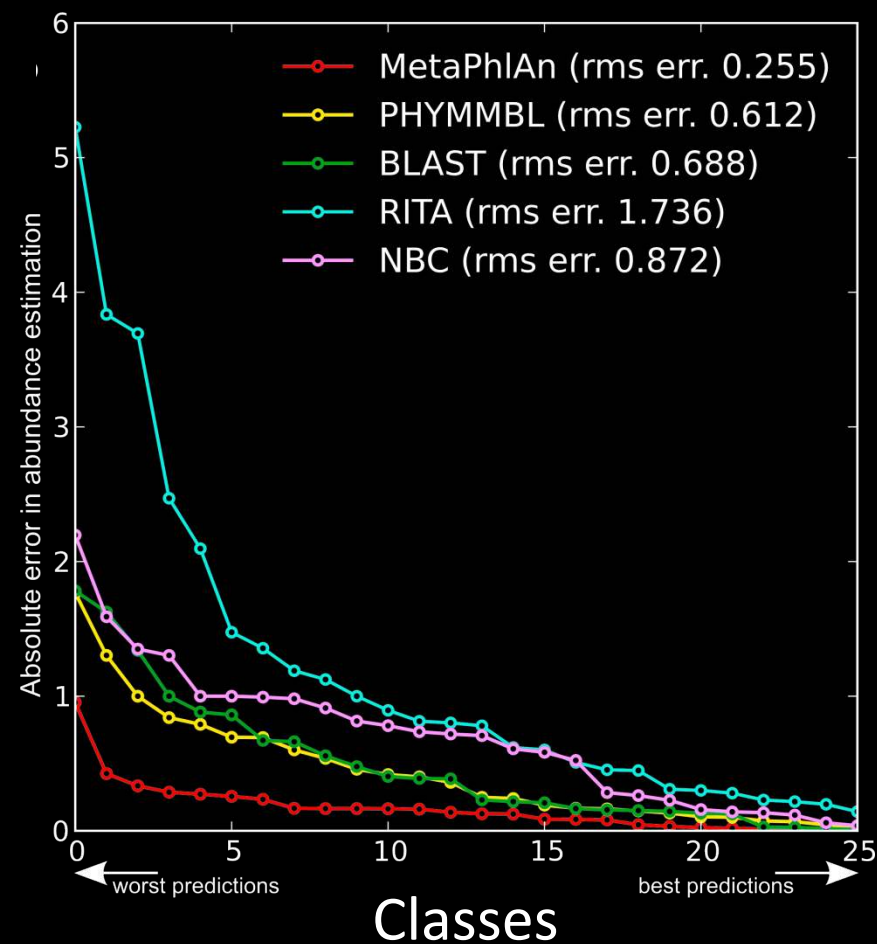
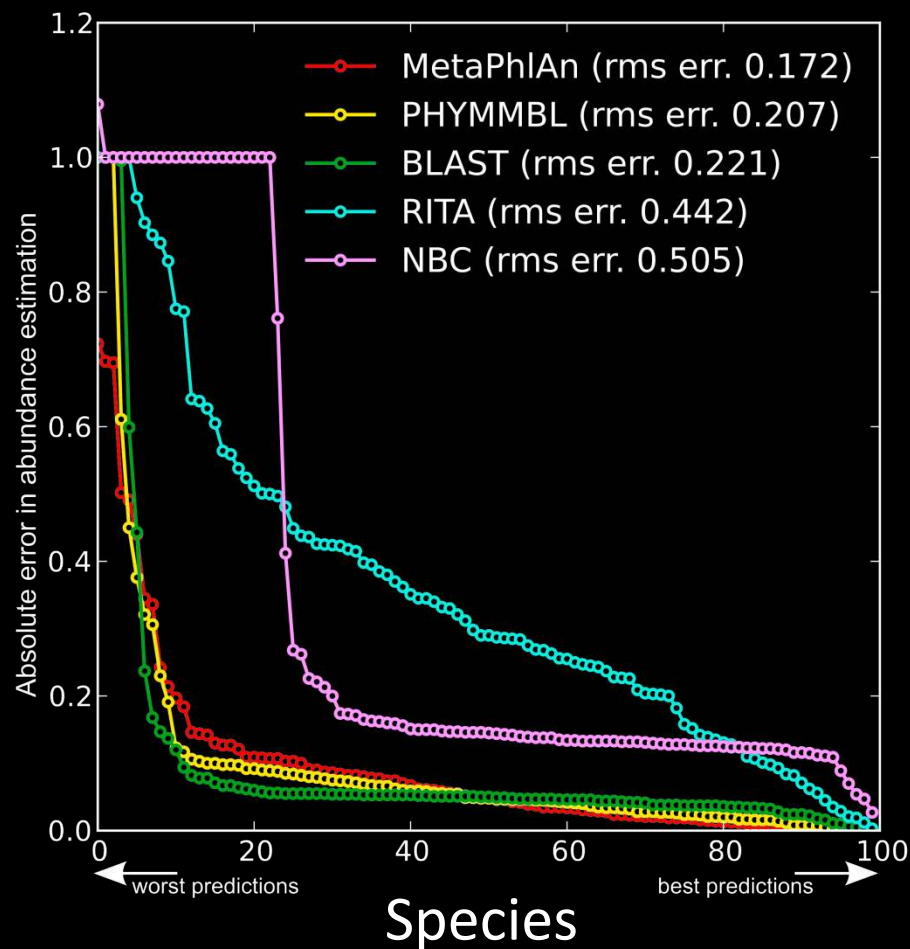
MetaPhlAn

Metagenomic Phylogenetic Analysis





Evaluation of MetaPhlAn accuracy

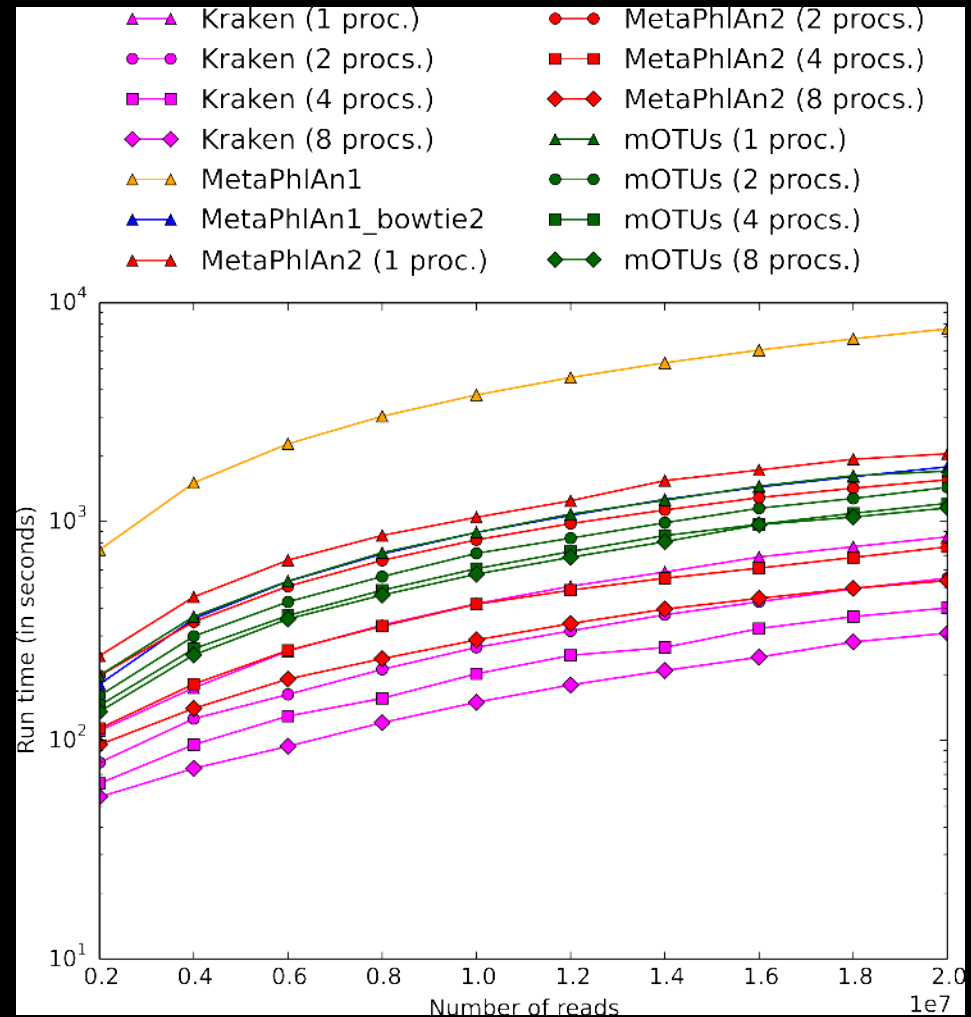
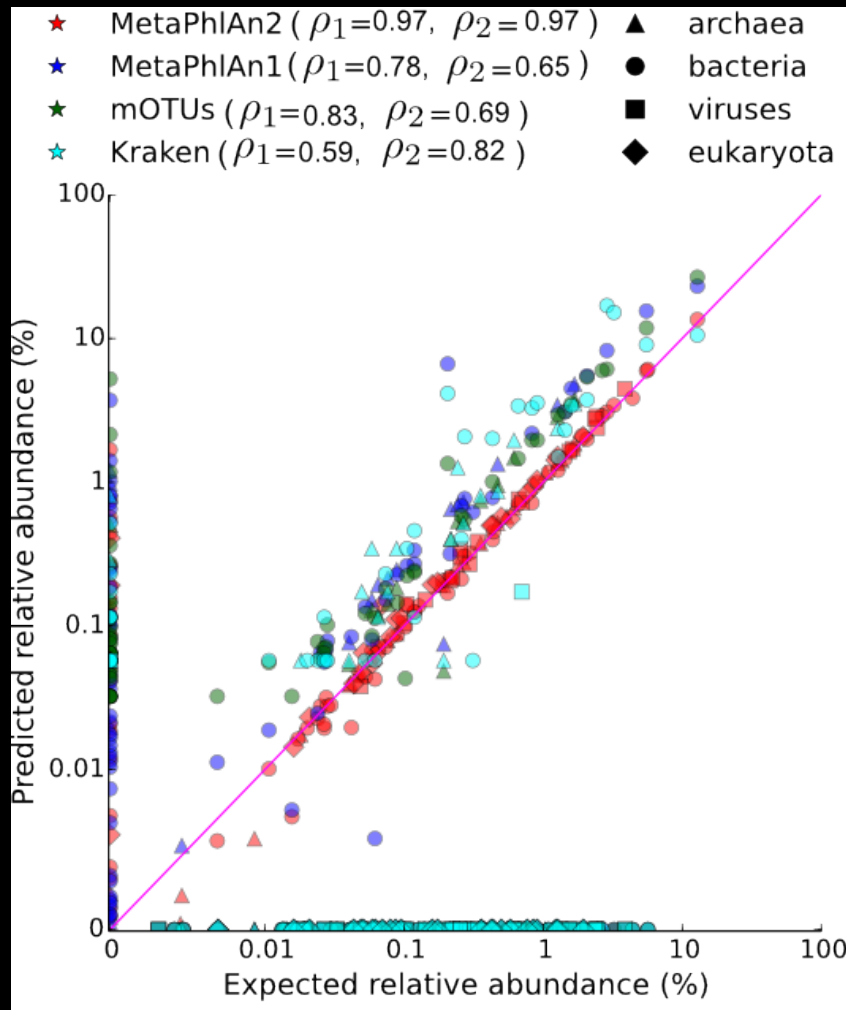


(Validation on high-complexity uniformly distributed synthetic metagenomes.)



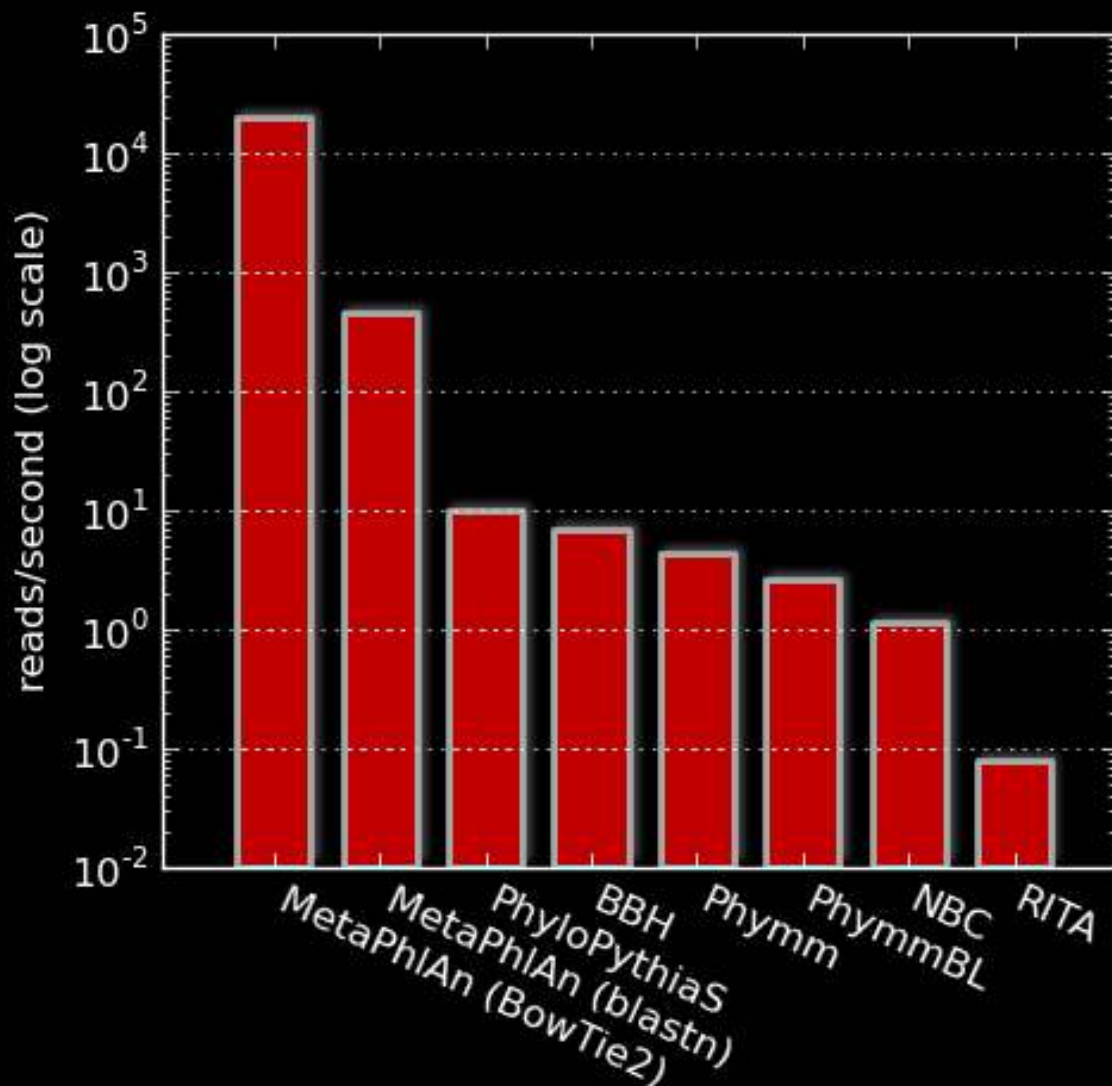
MetaPhlAn2: Taxonomic profiling using unique marker gene sequences

4Gnt synthetic metagenome, 125 organisms





Evaluation of MetaPhlAn performance



>50 times faster than earlier methods

450 reads/sec (BLAST)

Up to 25,000 reads/sec (bowtie2)

Multi-threaded

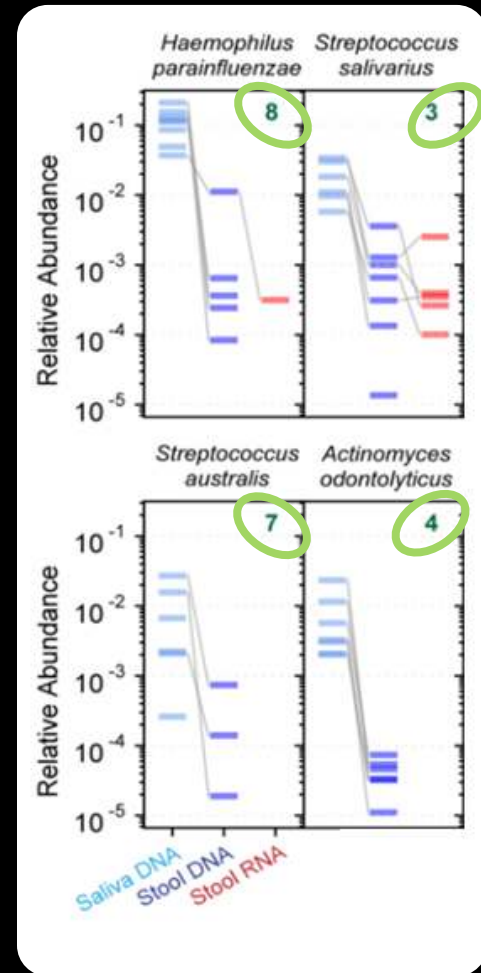
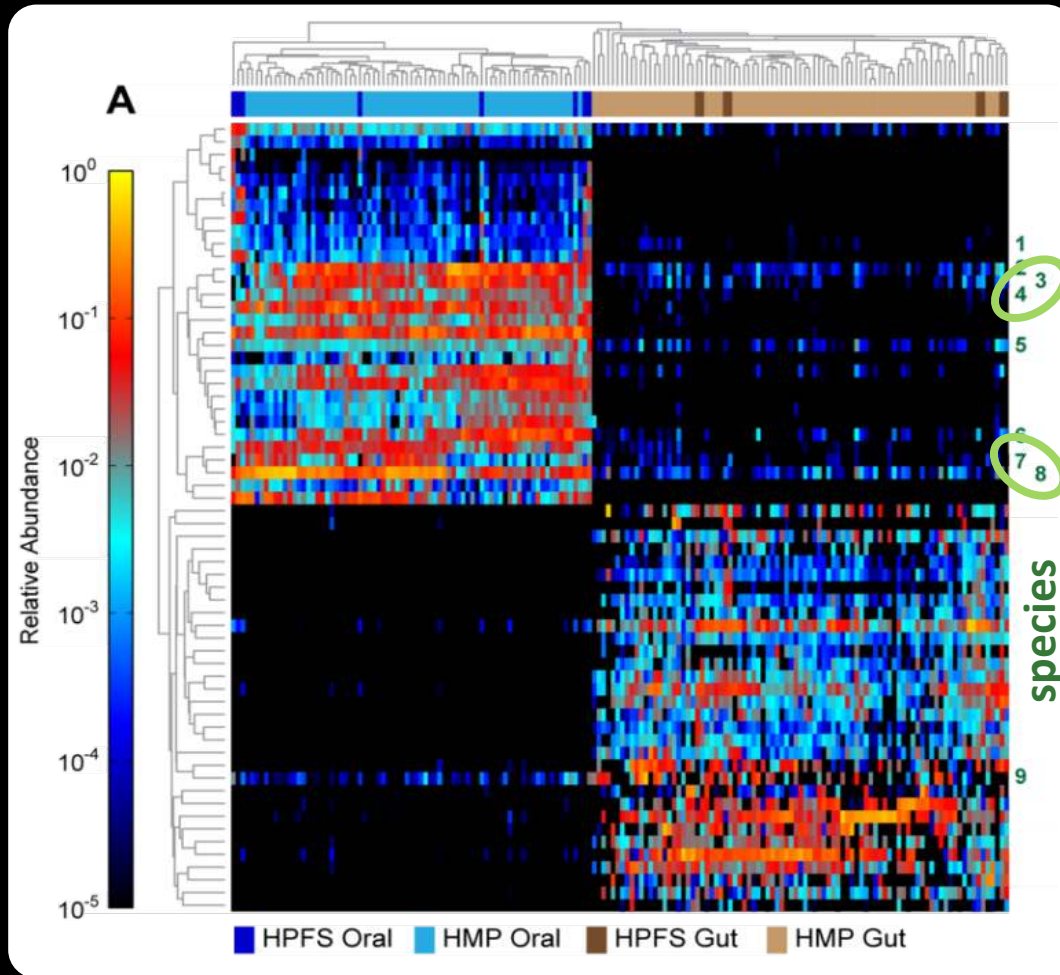
Easily parallelizable



MetaPhlAn in action

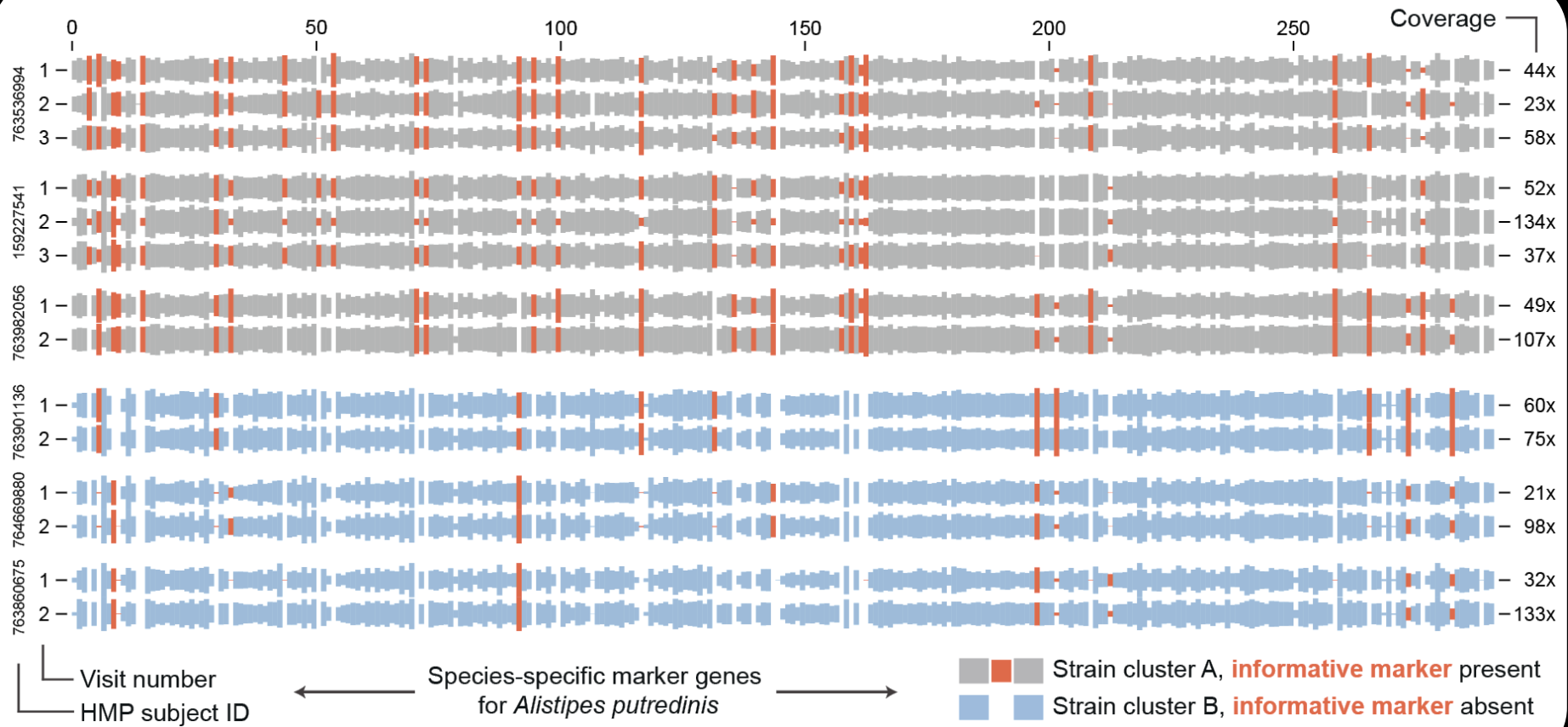


Eric Franzosa





MetaPhlAn in action: *strain profiling*



- In practice, not all markers are present
- Individual-specific marker “barcodes”
- Often very stable over time



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 Shotgun Sequence

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

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

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	Body Site	Reads Size	Reads MD5	Assembly	Ass. Size	Assembly MD5
Anterior Nares (94 Rows)						
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

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



Getting some (prepped) HMP data

- Connect to the server instead

- cd to your favorite directory and run:

```
for S in `ls ~/workshop_data/metagenomics/biobakery/data/7*.fasta`;  
do ln -s $S; done
```

- 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: MetaPhlAn1

- We won't use it today, but the first version of MetaPhlAn is at: <http://huttenhower.sph.harvard.edu/metaphlan>



The Huttenhower Lab

Department of Biostatistics, Harvard School of Public Health

Contact Documentation People Presentations Publications Research Teaching

Home

MetaPhlAn: 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 3,000 reference genomes, 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 bacterial and archaeal organisms;
- extensive validation of the profiling accuracy on several synthetic datasets and on thousands of real metagenomes.

Please refer to the [MetaPhlAn paper](#) for additional information, validations, and examples. Also the [main paper of the Human Microbiome Project](#) uses MetaPhlAn (version 1.1) for species-level metagenomic profiling.

Here is an [infographic](#) of the application of the [Human Microbiome Project](#) results obtained applying MetaPhlAn on the 690 shotgun sequencing samples. Email [me](#) for a high-resolution version. This infographic also appears in a slightly modified version as the main illustration of a [New York Times article](#) by Carl Zimmer available [here](#) (NY Times subscription needed) and [here](#) (NY Times copyrighted version).

A map of diversity in the human microbiome



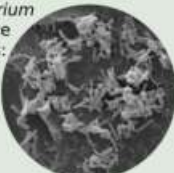
Streptococcus dominates the oral cavity with *S. mitis* > 75% in the **cheek**



Propionibacterium acnes lives on the skin and nose of most people



Many *Corynebacterium* species characterize different body sites:
C. matruchoti the **plaque**
C. accolens the **nose**
C. croppenstedtii the **skin**





Who's there: MetaPhlAn2

- Instead, go to <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 needed 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
`~/workshop_data/metagenomics/biobakery/software/metaphlan2`

The screenshot shows the Bitbucket web interface for the 'owner/repository' page. The header includes the Bitbucket logo, navigation links for 'Features' and 'Pricing', and user options for 'owner/repository', a search icon, a help icon, 'English', 'Sign up', and 'Log in'. The main content area is titled 'Source' and shows the repository structure. It includes a sidebar with icons for repository actions and a main pane displaying the file list. The file list shows folders 'db_v20' and 'utils', and files '.hgtags', 'README.md', and 'metaphlan2.py'. Below the file list is a section for 'MetaPhlAn 2.0: Metagenomic Phylogenetic Analysis' with a detailed table of contents.

File	Size	Time	Description
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

```
1. ssh

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

DESCRIPTION

MetaPhlAn version 2.0.0 beta2 (12 July 2014): METAgenomic PHyLogenetic ANalysis for taxonomic classification of metagenomic reads.

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

COMMON COMMANDS

```
:|
```



Who's there: MetaPhlAn2

- To launch your first analysis, run:

```
metaphlan2.py  
  --mpa_pkl ~/workshop_data/metagenomics/db_v20/mpa_v20_m200.pkl  
  --bowtie2db ~/workshop_data/metagenomics/db_v20/mpa_v20_m200  
  --input_type fasta  
  ./763577454-SRS014459-Stool.fasta  
  ./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`

```
1. [screen 2: bash] chuttenhower@class:~/tmp (ssh)
HWUSI-EAS1625_615HE:4:100:0:1248/1  gil479140210|refINC_021010.1|:1043207-1044529
HWUSI-EAS1625_615HE:4:100:0:1301/1  gil483877978|refINZ_KB890364.1|:31018-31902
HWUSI-EAS1625_615HE:4:100:1000:167/1  gil242362078|refINZ_GG692716.1|:28261-29169
HWUSI-EAS1625_615HE:4:100:1001:1264/1  gil270295698|refINZ_GG730107.1|:470181-472532
HWUSI-EAS1625_615HE:4:100:1001:1320/1  gil224993849|refINZ_ACFY01000158.1|:c1296-10
HWUSI-EAS1625_615HE:4:100:1001:1604/1  gil319644663|refINZ_GL635657.1|:c320982-320029
HWUSI-EAS1625_615HE:4:100:1001:1734/1  gil484001485|refINZ_KB894131.1|:91019-91717
HWUSI-EAS1625_615HE:4:100:1001:259/1  gil479210985|refINC_021043.1|:c1165057-1164158
HWUSI-EAS1625_615HE:4:100:1002:1501/1  gil224485637|refINZ_EQ973491.1|:c620672-618312
HWUSI-EAS1625_615HE:4:100:1003:1644/1  gil224485636|refINZ_EQ973490.1|:c204903-202990
HWUSI-EAS1625_615HE:4:100:1003:1702/1  gil423335209|refINZ_JH976498.1|:329186-330046
HWUSI-EAS1625_615HE:4:100:1003:2030/1  gil238922432|refINC_012781.1|:2910912-2912072
HWUSI-EAS1625_615HE:4:100:1004:353/1  gil223955873|refINZ_DS499674.1|:c266282-265248
HWUSI-EAS1625_615HE:4:100:1004:742/1  gil283767237|refINZ_GG730311.1|:c124395-124171
HWUSI-EAS1625_615HE:4:100:1005:1722/1  gil410105720|refINZ_JH976502.1|:750498-751148
HWUSI-EAS1625_615HE:4:100:1005:505/1  gil479170689|refINC_021020.1|:1540599-1542305
HWUSI-EAS1625_615HE:4:100:1006:848/1  gil347530298|refINC_015977.1|:c3433030-3431387
HWUSI-EAS1625_615HE:4:100:1007:1428/1  gil423332908|refINZ_JH976496.1|:1485161-1487113
HWUSI-EAS1625_615HE:4:100:1007:1465/1  gil423332908|refINZ_JH976496.1|:906255-909584
HWUSI-EAS1625_615HE:4:100:1008:1187/1  gil224485479|refINZ_EQ973214.1|:108053-108250
HWUSI-EAS1625_615HE:4:100:1008:1241/1  gil270293478|refINZ_GG730105.1|:c830784-828727
HWUSI-EAS1625_615HE:4:100:1008:140/1  gil224514921|refINZ_DS499545.1|:41991-42827
HWUSI-EAS1625_615HE:4:100:1009:154/1  gil301307949|refINZ_GG774972.1|:644845-649113
HWUSI-EAS1625_615HE:4:100:1009:467/1  gil303257489|refINZ_GL383997.1|:67163-67873
:
```



Who's there: MetaPhlAn

less 763577454-SRS014459-Stool.txt

```
1. [screen 2: bash] chuttenhower@class:~/tmp (ssh)
k__Bacteria 100.0
k__Bacterialp__Firmicutes 64.82041
k__Bacterialp__Bacteroidetes 35.17959
k__Bacterialp__Firmicuteslc__Clostridia 64.82041
k__Bacterialp__Bacteroideteslc__Bacteroidia 35.17959
k__Bacterialp__Firmicuteslc__Clostridialo__Clostridiales 64.82041
k__Bacterialp__Bacteroideteslc__Bacteroidialo__Bacteroidales 35.17959
k__Bacterialp__Firmicuteslc__Clostridialo__Clostridialeslf__Ruminococcaceae 37.71449
k__Bacterialp__Bacteroideteslc__Bacteroidialo__Bacteroidaleslf__Bacteroidaceae 31.5000
k__Bacterialp__Firmicuteslc__Clostridialo__Clostridialeslf__Eubacteriaceae 21.99035
k__Bacterialp__Firmicuteslc__Clostridialo__Clostridialeslf__Lachnospiraceae 5.11557
k__Bacterialp__Bacteroideteslc__Bacteroidialo__Bacteroidaleslf__Porphyromonadaceae 3.6
k__Bacterialp__Firmicuteslc__Clostridialo__Clostridialeslf__Ruminococcaceaelg__Subdolig
k__Bacterialp__Bacteroideteslc__Bacteroidialo__Bacteroidaleslf__Bacteroidaceaelg__Bacte
k__Bacterialp__Firmicuteslc__Clostridialo__Clostridialeslf__Eubacteriaceaelg__Eubacteri
k__Bacterialp__Firmicuteslc__Clostridialo__Clostridialeslf__Lachnospiraceaelg__Roseburi
k__Bacterialp__Bacteroideteslc__Bacteroidialo__Bacteroidaleslf__Porphyromonadaceaelg__P
k__Bacterialp__Firmicuteslc__Clostridialo__Clostridialeslf__Ruminococcaceaelg__Subdolig
k__Bacterialp__Firmicuteslc__Clostridialo__Clostridialeslf__Eubacteriaceaelg__Eubacteri
k__Bacterialp__Bacteroideteslc__Bacteroidialo__Bacteroidaleslf__Bacteroidaceaelg__Bacte
k__Bacterialp__Bacteroideteslc__Bacteroidialo__Bacteroidaleslf__Bacteroidaceaelg__Bacte
k__Bacterialp__Firmicuteslc__Clostridialo__Clostridialeslf__Eubacteriaceaelg__Eubacteri
k__Bacterialp__Firmicuteslc__Clostridialo__Clostridialeslf__Lachnospiraceaelg__Roseburi
k__Bacterialp__Bacteroideteslc__Bacteroidialo__Bacteroidaleslf__Bacteroidaceaelg__Bacte
763577454-SRS014459-Stool.txt
```



Who's there: MetaPhlAn2

- You can finish the job if you like:

```
metaphlan2.py
  --mpa_pkl ~/workshop_data/metagenomics/db_v20/mpa_v20_m200.pkl
  --bowtie2db ~/workshop_data/metagenomics/db_v20/mpa_v20_m200
  --input_type fasta
  ./763577454-SRS014464-Anterior_nares.fasta
  ./763577454-SRS014464-Anterior_nares.txt
  ...
```

– Note that you can use the up arrow key to make your life easier!

- Or you can copy the rest pre-calculated:

```
cp ~/workshop_data/metagenomics/biobakery/results/metaphlan/*.txt .
```



Who's there: MetaPhlAn2

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

```
mkdir tmp
```

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

```
~/workshop_data/metagenomics/biobakery/software/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: MetaPhlAn

- But it's easier using MeV; go to <http://www.tm4.org/mev.html>

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](#)

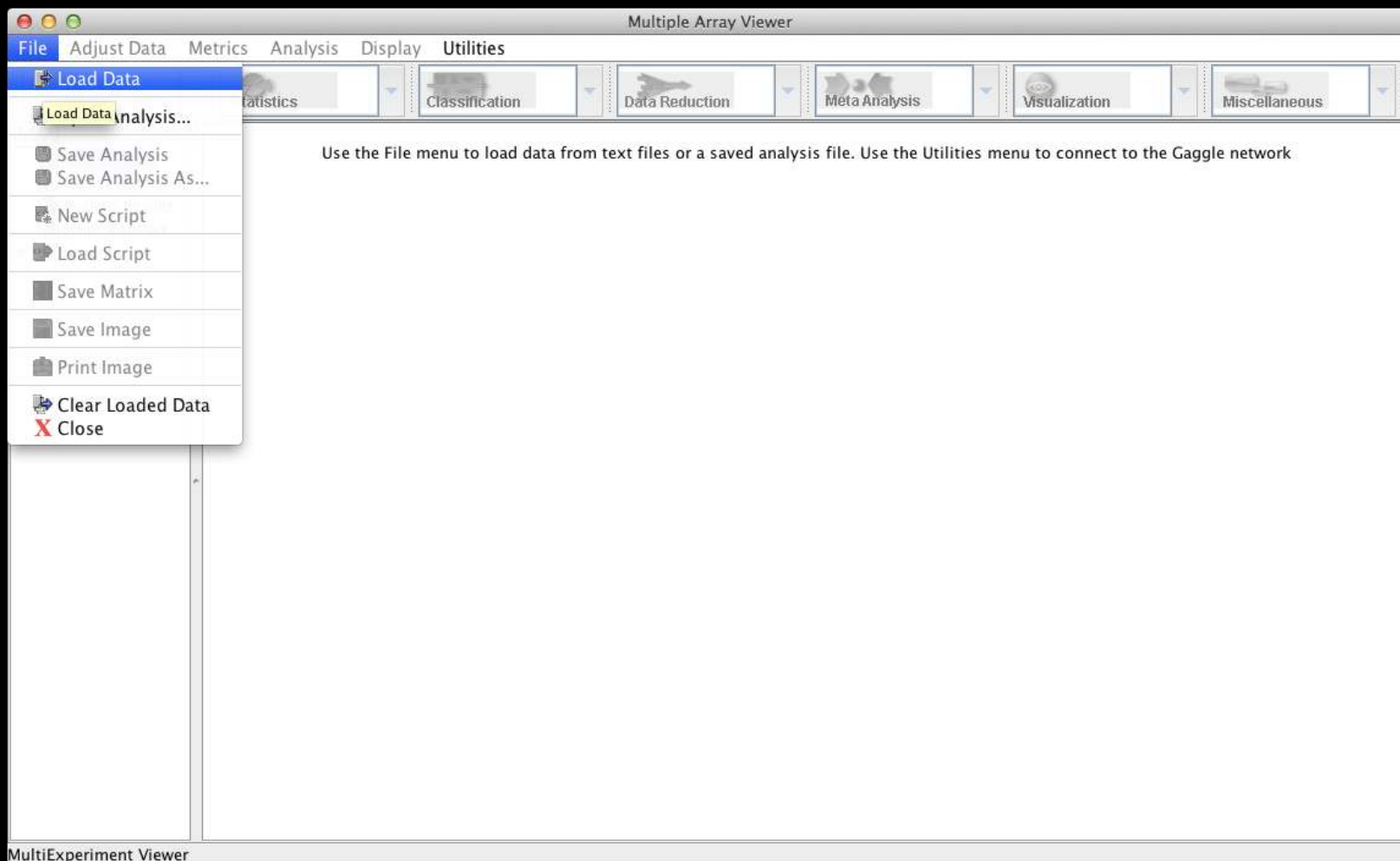
TM4-Announce Sign-Up
Support TM4 Development and stay informed about the software. We do not share email addresses with anyone.

- Or use the pre-installed version:
`tmev.sh`



An interlude: MeV

- Launch MeV, and select File/Load data





An interlude: MeV

- Click “Browse” to your TSV file (763577454.tsv), 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

Select annotation file /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...		3.51469	0.38831	6.74077	
k_Bacteri...		3.51469	0.38831	6.74077	
k_Bacteri...		3.51469	0		
k_Bacteri...		0		2.43846	
k_Bacteri...		0	0.38831	4.30232	
k_Bacteri...	42.97557	0		41.42792	
k_Bacteri...	42.97557	0		41.42792	

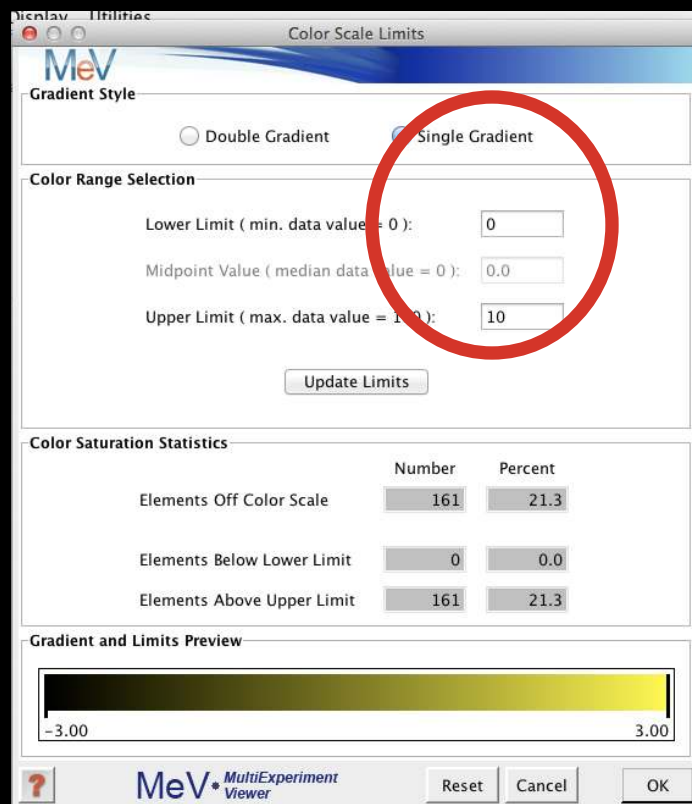
Click the upper-leftmost expression value. Click the Load button to finish.

MeV MultiExperiment Viewer Cancel Load



An interlude: MeV

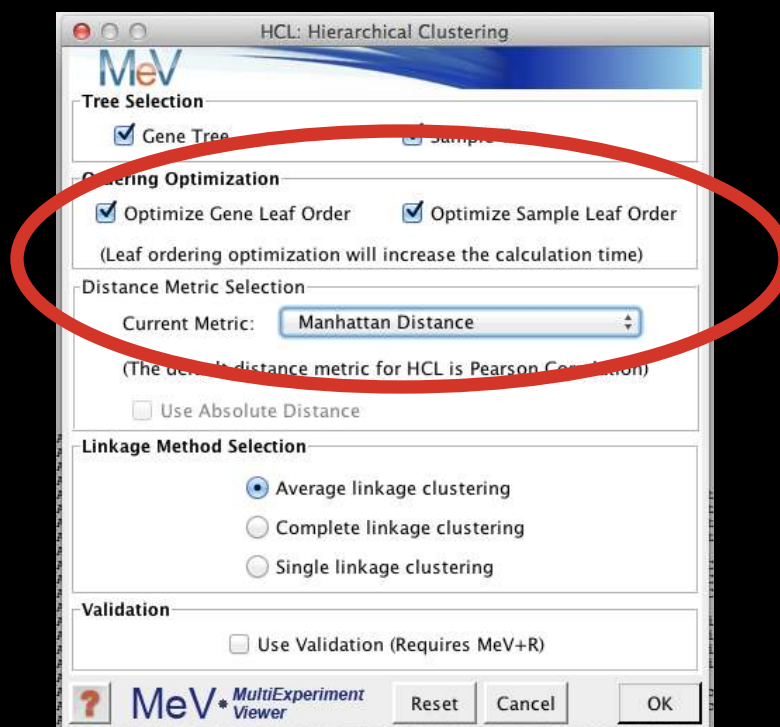
- “Load” your data, then make it visible by:
 - Display/Set Color Scale Limits
 - Choose Single Gradient, min 0, max 10





An interlude: MeV

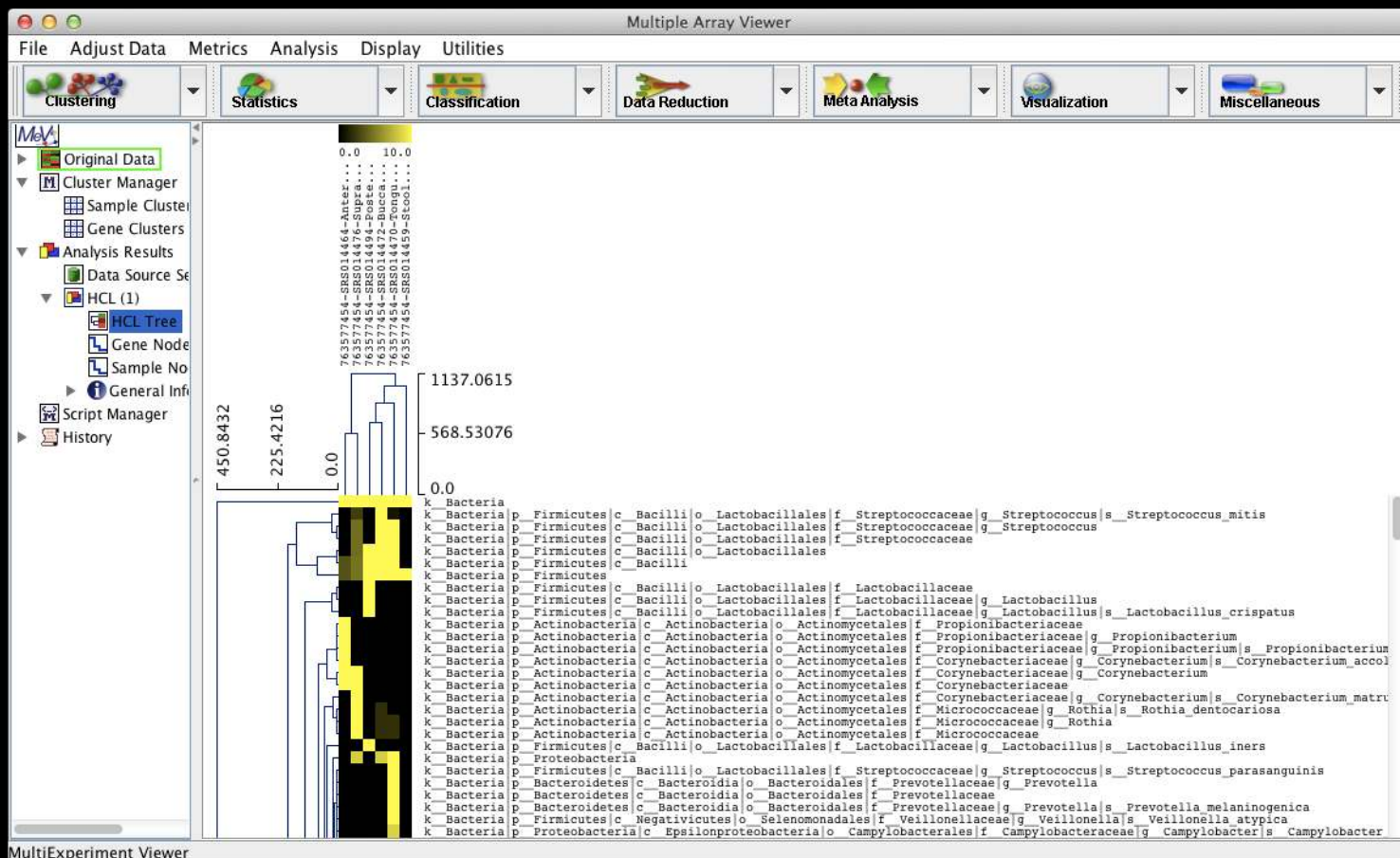
- Finally, to play around a bit:
 - Display/Set Element Size/whatever you'd like
 - Clustering/Hierarchical Clustering
 - Optimize both gene and sample order
 - And select Manhattan Distance (imperfect!)





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

- Compositional nature ($\Sigma = 1$)
 - Abundance is relative, not absolute
- High dynamic range
- Often sparse (sample dominated by a few species)
- Noisy
- Hierarchical organization

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

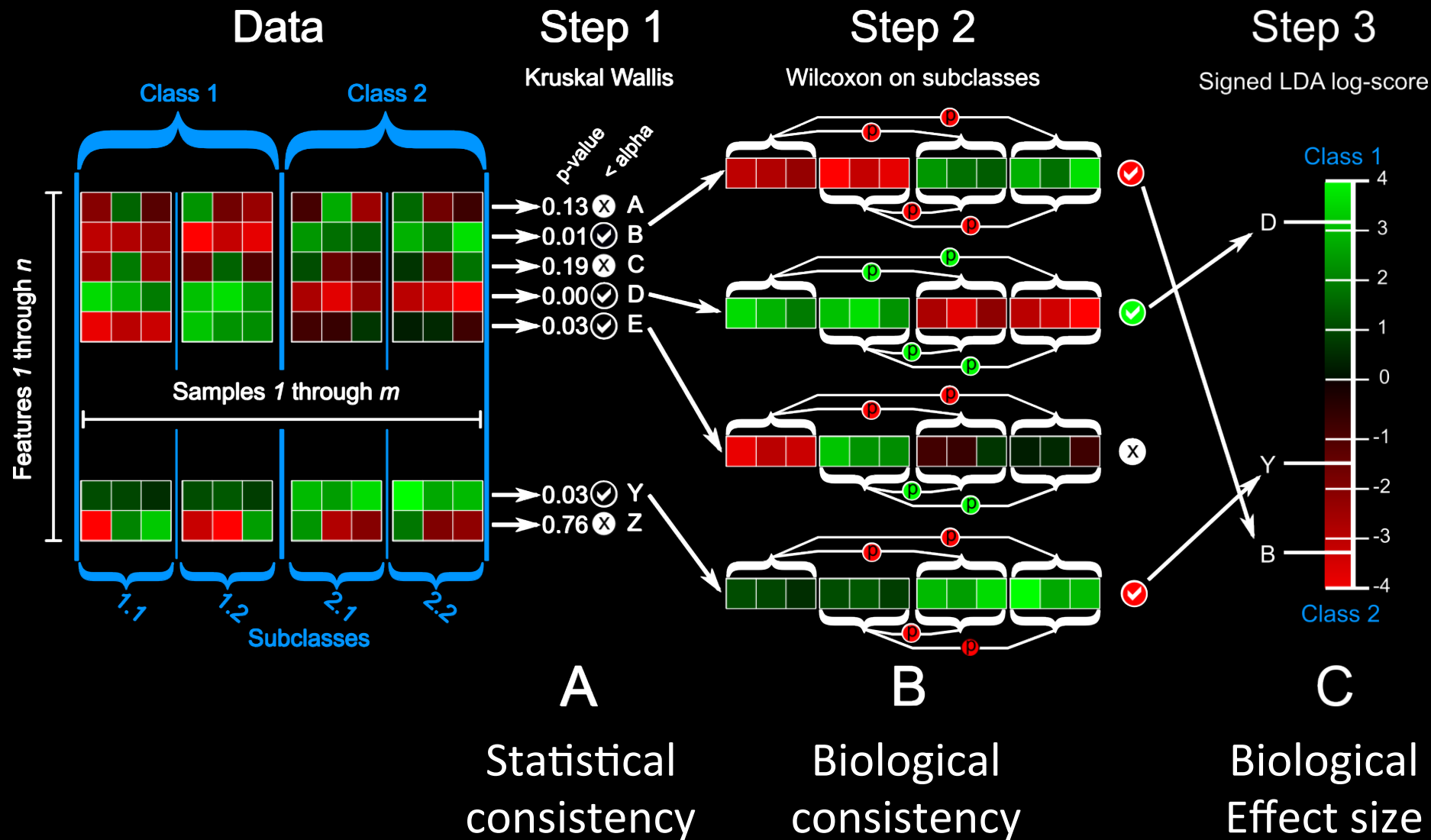
Sample #	1	2	3	4	5	6
Profession	Student	Postdoc	Postdoc	Professor	Student	Student
Gender	Male	Female	Female	Male	Male	Female
Site	Oral	Gut	Oral	Gut	Oral	Gut
Clade1	0.40	0.87	0.43	0.68	0.47	0.32
Clade1 Bug1	0.40	0.56	0.07	0.31	0.42	0.27
Clade1 Bug2	0.00	0.30	0.36	0.37	0.04	0.05
Clade2	0.60	0.13	0.57	0.32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

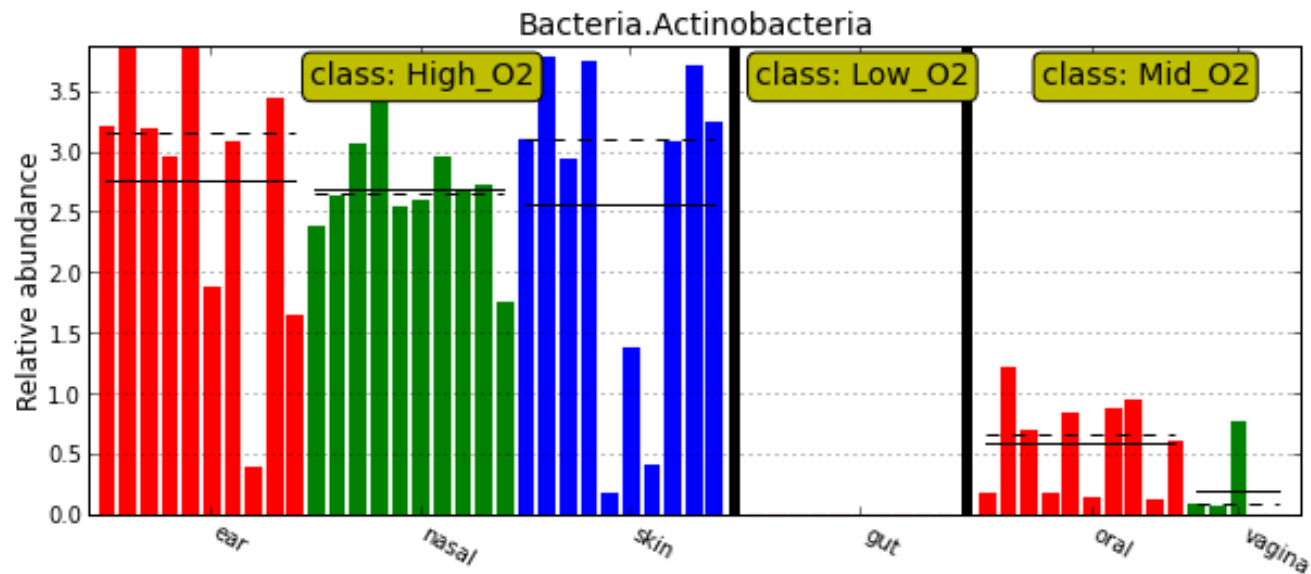
Finding metagenomic biomarkers

Nicola
Segata



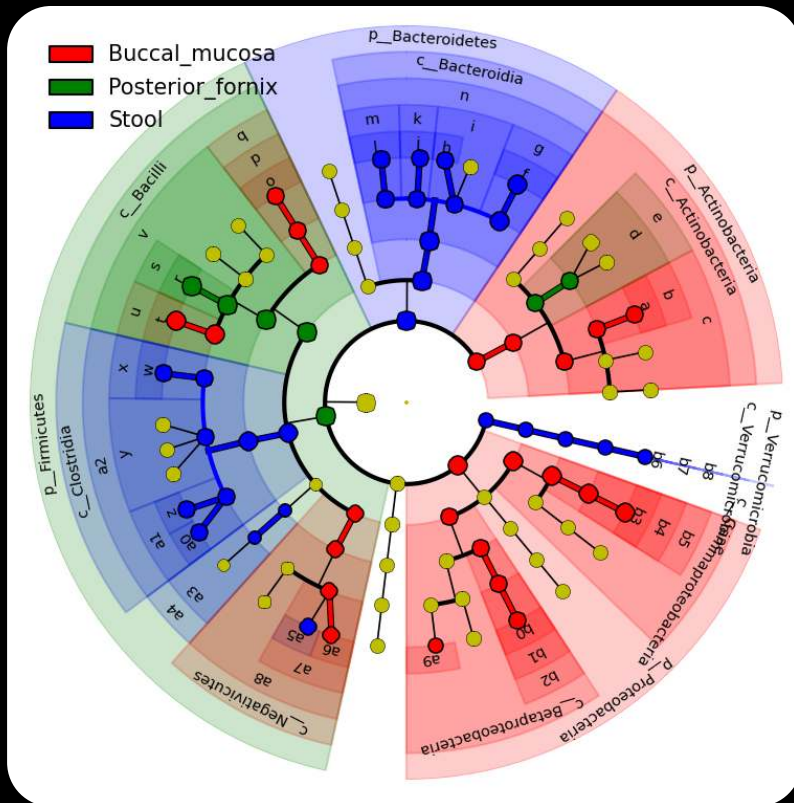


Example LEfSe application: Find O₂-loving bugs (controlling for body site)

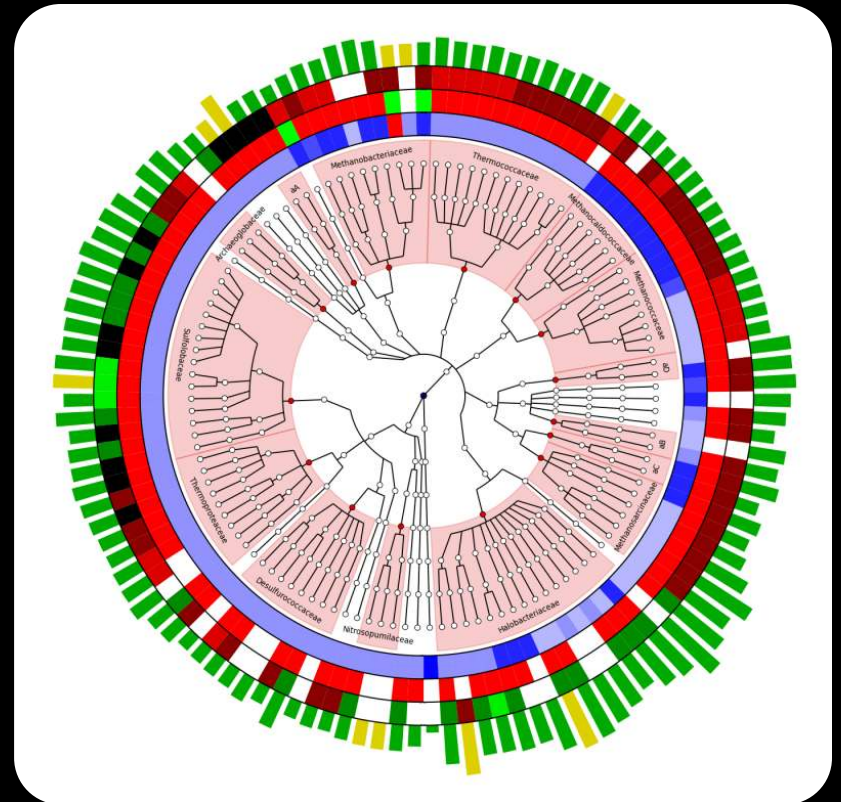




Superimpose enrichments on the tree of life using GraPhlAn



LEfSe Associations



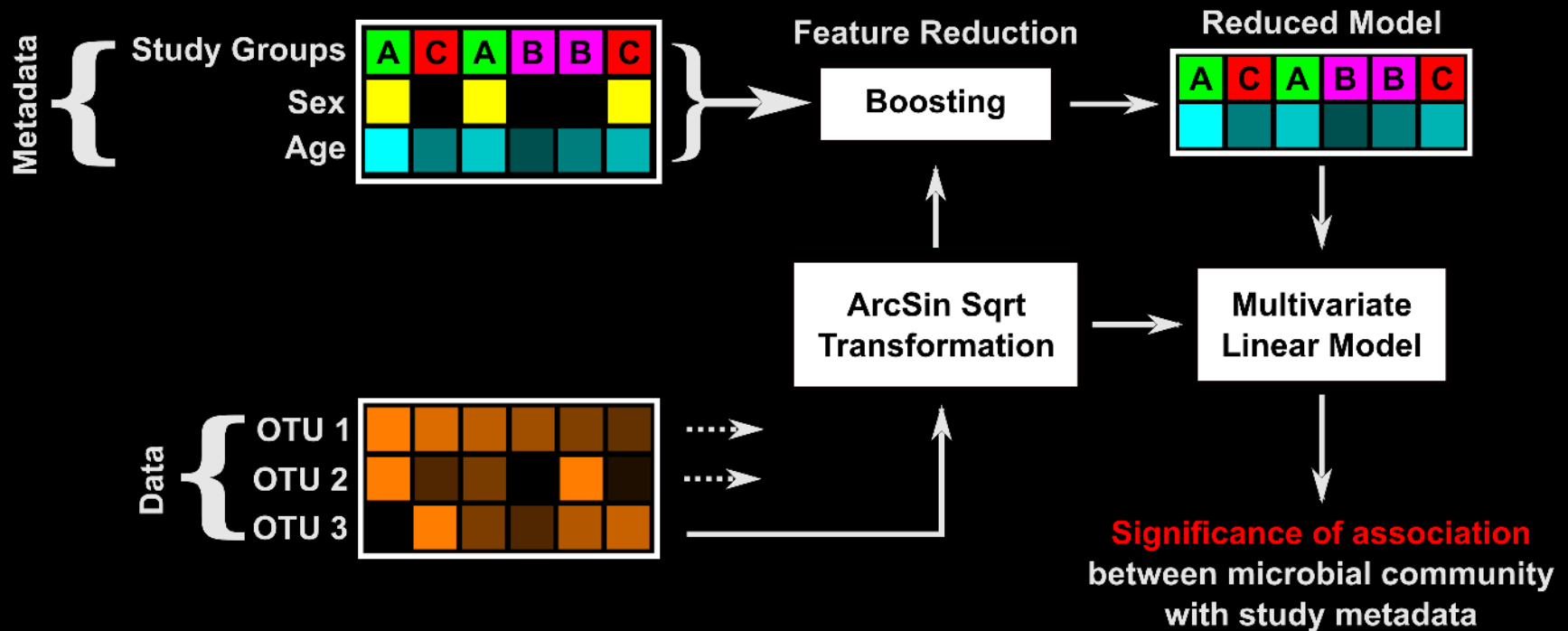
Metadata Rings

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



MaAsLin

Multivariate microbial Association with Linear models



- A more general solution for finding significant metagenomic associations in metadata-rich studies

Tim
Tickle





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
- HMG1 Gene Index
- HMG1 Clustered gene index
- HMG1 GO slim analysis
- HMG1 Shotgun community profiling
- HMSMCP Shotgun MetaPhlAn Community Profiling**
- HMMRC Metabolic reconstruction and clustering
- HMG01 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**

Feedback

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

```
3. [screen 4: bash] chuttenhower@class:~ (ssh)
[chuttenhower@class ~]$ wget http://downloads.hmpdacc.org/data/HMSMCP/HMP.ab.txt.bz2
--2014-08-14 08:14:48-- http://downloads.hmpdacc.org/data/HMSMCP/HMP.ab.txt.bz2
Resolving downloads.hmpdacc.org... 134.192.156.83
Connecting to 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    1.47M/s   in 0.2s

2014-08-14 08:14:49 (1.47 MB/s) - "HMP.ab.txt.bz2" saved [321880/321880]

[chuttenhower@class ~]$
```



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 HMP.ab.txt
```

- **IMPORTANT!!!**

- This file's too big to analyze directly today

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

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

- Scripts and data from HUMAnN to the rescue:

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

- NOW take a look again



What it means: LEfSe

- Let's modify this file to be LEfSe-compatible
- scp it to your laptop and open it up in Excel

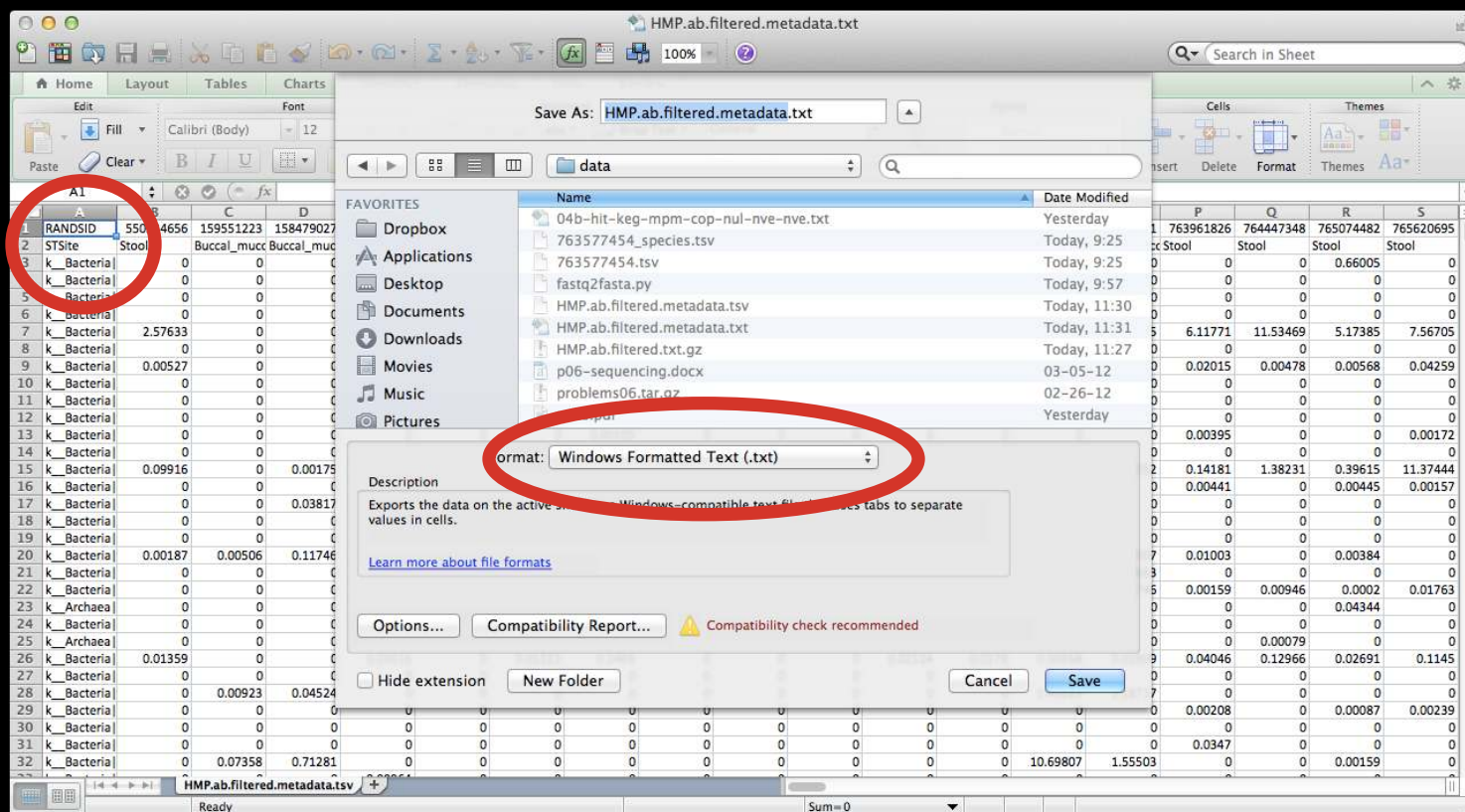
The screenshot shows an Excel spreadsheet with the following data:

A1	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
sid	SRS043001	SRS017127	SRS021473	SRS011134	SRS050184	SRS011529	SRS048164	SRS016516	SRS052330	SRS011355	SRS011452	SRS019787	SRS054776	SRS024140	SRS014683	SRS016018	SRS047014	SRS019601
RANDSID	550534656	159551223	158479027	158499257	508703490	159166850	861967750	159753524	765640925	158944319	159146620	764669880	764224817	159207311	763961826	764447348	765074482	765620695
START	Q3_2009	Q2_2009	Q1_2009	Q1_2009	Q3_2009	Q2_2009	Q3_2009	Q2_2009	Q3_2009	Q1_2009	Q2_2009	Q2_2009	Q2_2009	Q2_2009	Q2_2009	Q2_2009	Q2_2009	Q3_2009
GENDER	female	male	male	male	female	male	male	female	female	female	male	male	male	male	male	male	male	female
VISNO	1	1	2	1	1	1	1	1	1	1	1	2	2	2	1	1	2	1
STSite	Stool	Buccal_mucc	Buccal_mucc	Stool	Posterior_fo	Stool	Stool	Posterior_fo	Posterior_fo	Posterior_fo	Stool	Stool	Buccal_mucc	Buccal_mucc	Stool	Stool	Stool	Stool
Parent_Spec	700106291	700033688	700097185	700014832	700038759	700016608	700038870	700032243	700038805	700015577	700016136	700038231	700106652	700100608	700023337	700024646	700105580	700038072
Run ID	704GE	61HJLAAXX	61K2LAAXX	61JGUAAXX	704N4	61PNF	61NTL	61VKUAAXX	7055M	61KYVAAXX	61WER	704MU	621F6	61JD1AAXX	614NM	61PPR	61U6W	61WFD
Lane	6	6	7	6	8	5	1	4	1	8	2	4	5	2	1	5	1	3
SRS	700106291	700033688	700097185	700014837	700038759	700016610	700038870	700032243	700038805	700015579	700016142	700038263	700106652	700100608	700023337	700024673	700105580	700038072
Mean Quality	29.75	29	33	27	31.07	31.92	33.16	33	32.91	24	32.97		32.65	29		34.24		
Number of Q	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
Percent of H	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		0.0001		
Unique Non-	#####	#####	959383209	#####	#####	#####	#####	#####	#####	683832927	#####	#####	#####	#####	#####	#####	#####	#####
k_Bacteria	0	0	0	0.59019	0	0.15046	1.46625	0	0	0	0	0	0	0	0	0	0.66005	0
k_Bacteria	0	0	0	0.00417	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k_Bacteria	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k_Bacteria	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k_Bacteria	2.57633	0	0	4.20761	0	5.52547	8.52942	0	0	0	15.90901	1.27072	0.14801	1.19015	6.11771	11.53469	5.17385	7.56705
k_Bacteria	0	0	0	0	0	0.00189	0	0	0	0	0	0	0	0	0	0	0	0
k_Bacteria	0.00527	0	0	0.00358	0	0.00975	0.01062	0	0.01611	0	0	0	0	0	0.02015	0.00478	0.00568	0.04259
k_Bacteria	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k_Bacteria	0	0	0	0	0	0	0	0	0	0	0	0	0.00686	0	0	0	0	0
k_Bacteria	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k_Bacteria	0	0	0	0	0	0	0	0.00105	0	0	0	0	0	0	0.00395	0	0	0.00172
k_Bacteria	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k_Bacteria	0.09916	0	0.00175	0.33665	0	0.10695	0.93543	0	0	0	1.85125	0.07391	0	0.35352	0.14181	1.38231	0.39615	11.37444
k_Bacteria	0	0	0	0	0	0	0	0	0.00528	0	0.00049	0	0	0	0.00441	0	0.00445	0.00157
k_Bacteria	0	0	0.03817	0	0	0	0	0	0	0	0	0	0.0182	0	0	0	0	0
k_Bacteria	0	0	0	0	0	0.00189	0	0	0	0	0	0	0	0	0	0	0	0
k_Bacteria	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k_Bacteria	0.00187	0.00506	0.11746	0.07423	0	0.00632	0.00537	0	0	0	0	0	0.03278	0.04907	0.01003	0	0.00384	0



What it means: LEfSe

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





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!

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', options include '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 like 'MetaPhlAn', 'GraPhlAn', 'microPITA', 'MaAsLin', and 'PICRUST' are also listed.

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 its right, the filename 'HMP.ab.filtered.metadata.txt' is shown in red text. Below the filename, there is a text area for 'URL/Text'. At the bottom of the tool panel, a red circle highlights the 'Execute' button. A red arrow points from the 'Execute' button to the 'History' panel on the right.

The 'History' panel on the right 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'.

Red annotations with numbers provide instructions: '1. Click here, browse to HMP.ab.filtered.metadata.txt' points to the 'Choose File' button; '2. Then here' points to the 'Execute' button; and '3. Then watch here' points to the message box in the History panel.



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

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: **4**

Do you want the pairwise comparisons among subclasses to be performed only among the subclasses with the same name?: No

Select the strategy for multiple class analysis: **One-against-all (less strict)**

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 search bar. Under the 'LEfSe' category, several tools 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'. The tool 'C) Plot LEfSe Results' is circled in red, with a red arrow pointing to it and the text '1. Click here'. The main panel 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', 'Set text and label options' set to 'Default', 'Set some graphical options to personalize the output' set to 'Default', 'Output format' set to 'png', and 'Set the dpi resolution of the output' set to '150'. The 'Execute' button is circled in red, with a red arrow pointing to it and the text '2. Then here'. The 'History' panel on the right shows a list of previous jobs: '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'.



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 header includes the Galaxy logo, the text "Galaxy / Huttenhower Lab", and navigation tabs for "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User". A status bar on the right indicates "Using 0%".

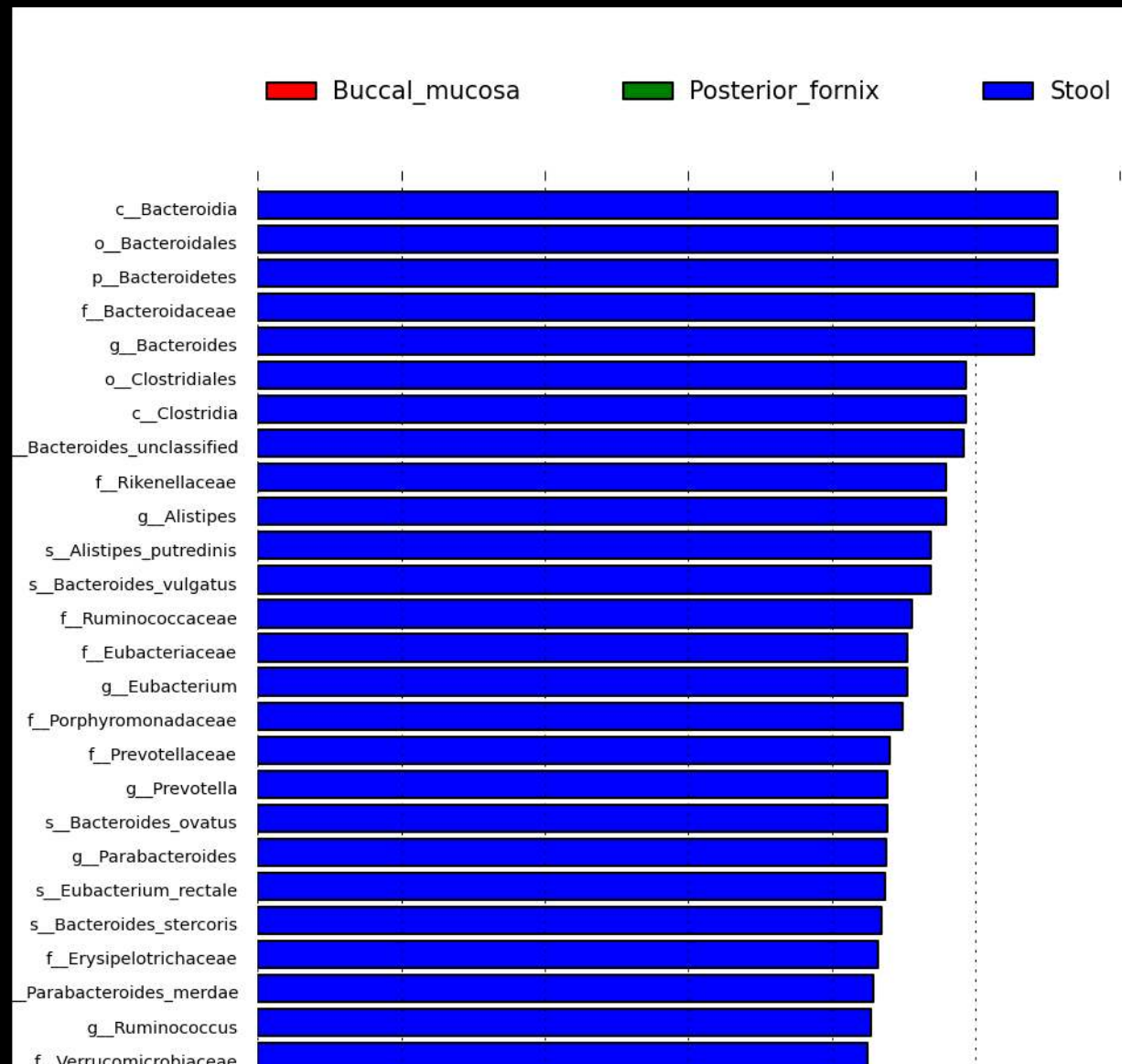
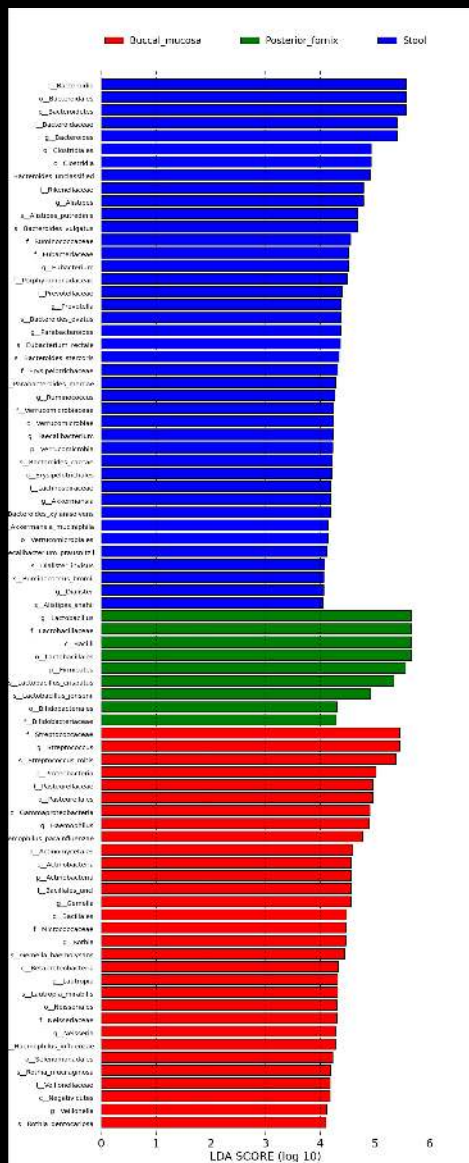
On the left, the "Tools" panel lists "HUTTENHOWER LAB MODULES" and "LEfSe" with sub-items: "A) Format Data for LEfSe", "B) LDA Effect Size (LEfSe)", "C) Plot LEfSe Results", "D) Plot Cladogram", "E) Plot One Feature", and "F) Plot Differential Features". Other modules listed are "MetaPhlAn", "GraPhlAn", "microPITA", "MaAsLin", and "PICRUSt".

The central panel shows 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". Below this, it states: "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."

On the right, the "History" panel lists several jobs. Job 5, "5: C) Plot LEfSe Results on data 4", is highlighted with a red circle around its "eye" icon, which is the target of the "Click here" instruction. Other jobs listed include "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".



What it means: LEfSe

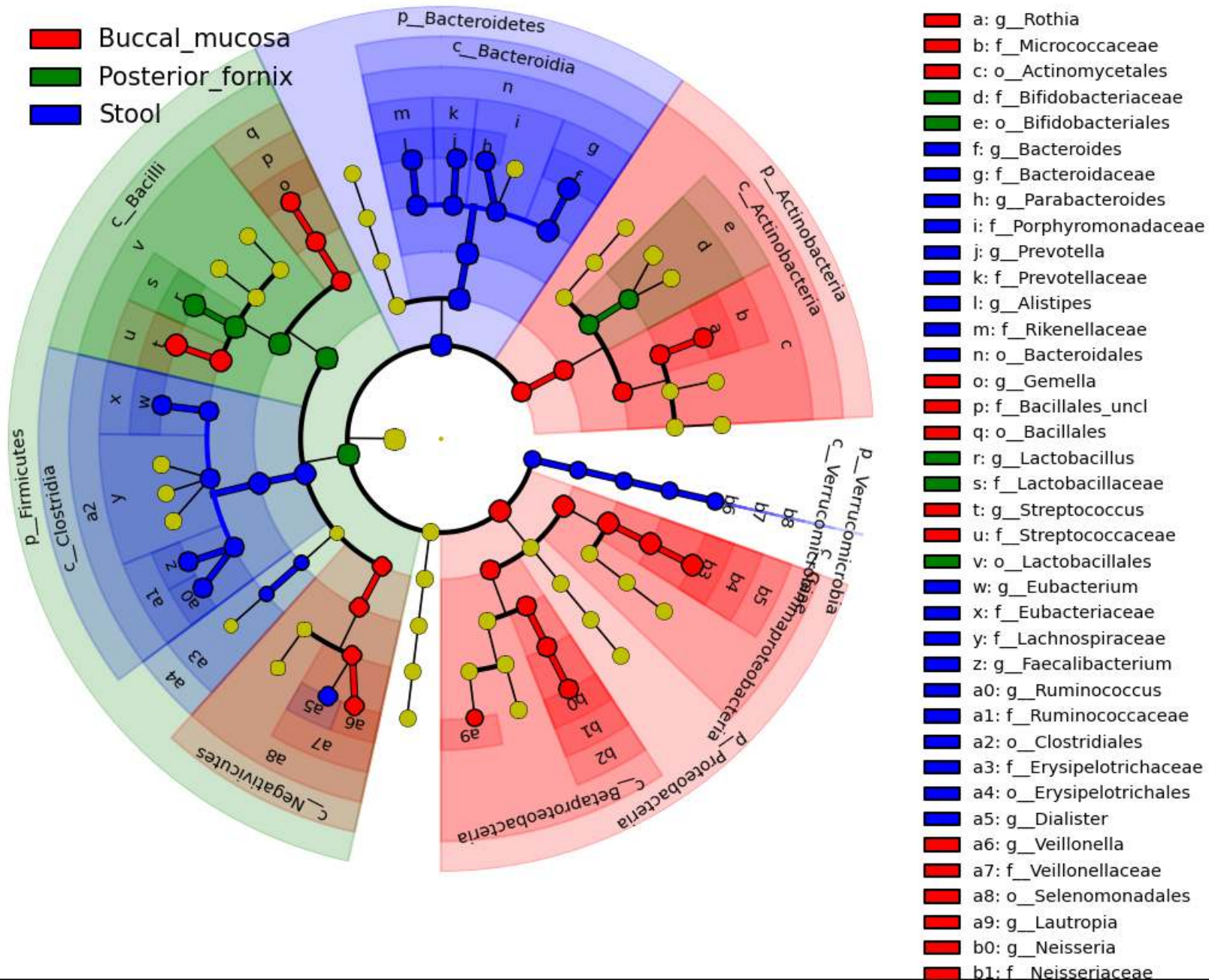




What it means: LEfSe

- 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. On the left, the 'Tools' panel lists 'HUTTENHOWER LAB MODULES' with 'LEfSe' expanded. 'D) Plot Cladogram' is circled in red, with a red arrow pointing to it from the text '1. Click here'. The main panel shows the 'D) Plot Cladogram (version 1.0)' tool configuration. It includes options for 'Select data' (set to '4: B) LDA Effect Size (LEfSe) on data 3'), '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'. On the right, the 'History' panel shows a list of previous jobs, including '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', and '2: HMP.ab.filtered.metadata.txt'.

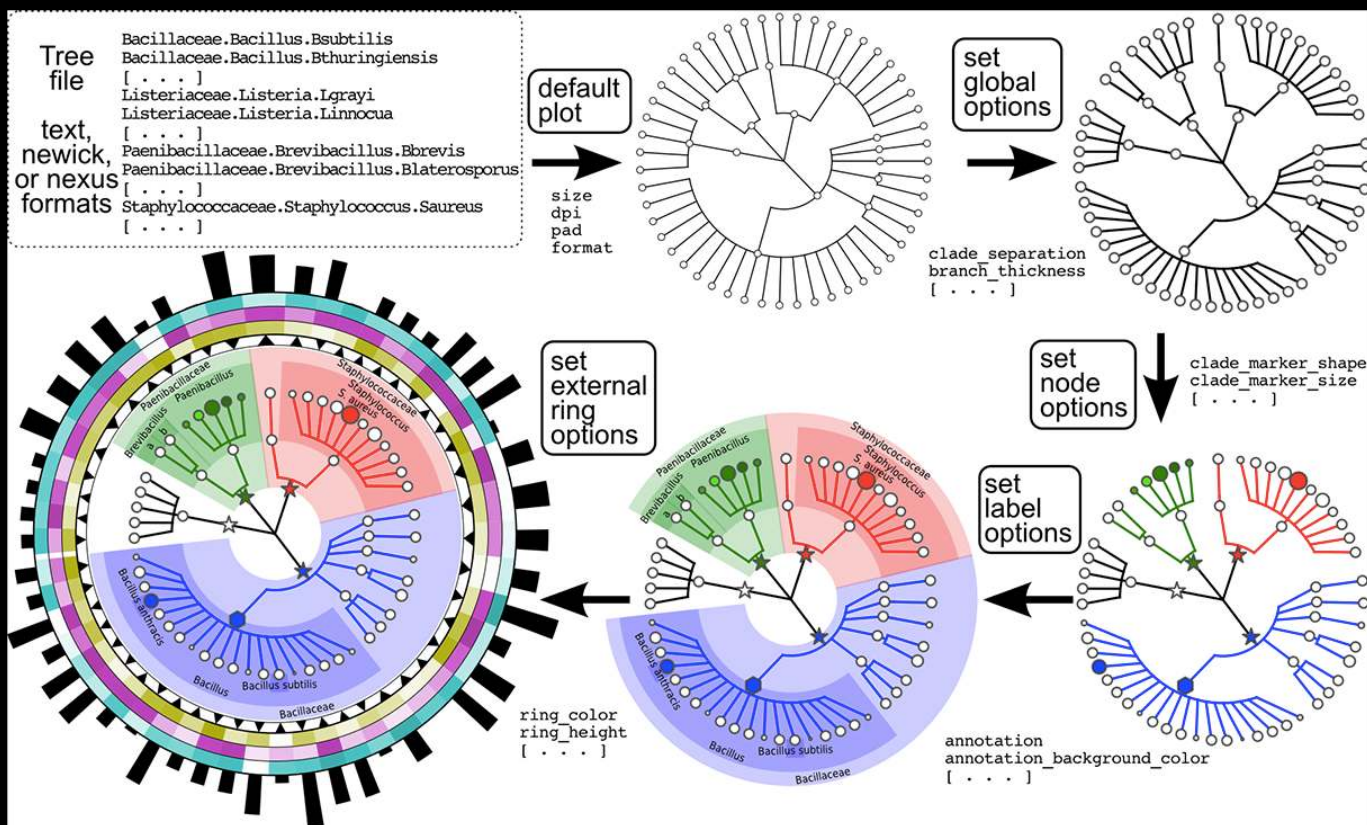




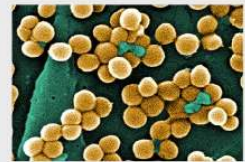
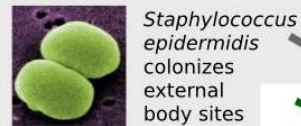
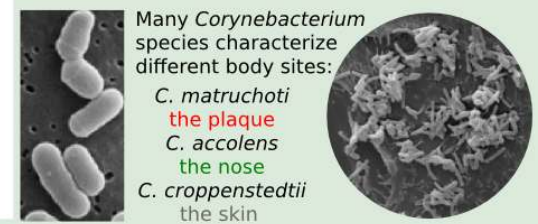
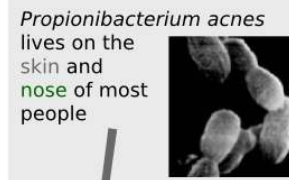
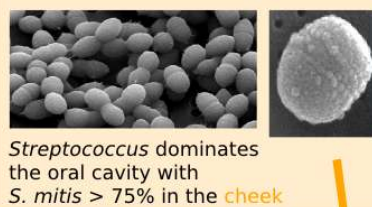
An aside: GraPhlAn

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

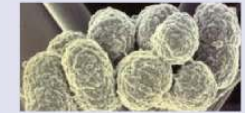
<http://huttenhower.sph.harvard.edu/graphlan>



A map of diversity in the human microbiome



Several *Prevotella* species are present in the gastrointestinal tract. *P. copri* is present in 19% of the subjects and dominates the **intestinal** flora when present



Bacteroides is the most abundant genus in the **gut** of almost all healthy subjects



Campylobacter includes opportunistic pathogens, but members live in the oral cavities of most healthy people in the cohort



○ Commensal microbes
 ☆ Potential pathogens

The four most abundant phyla

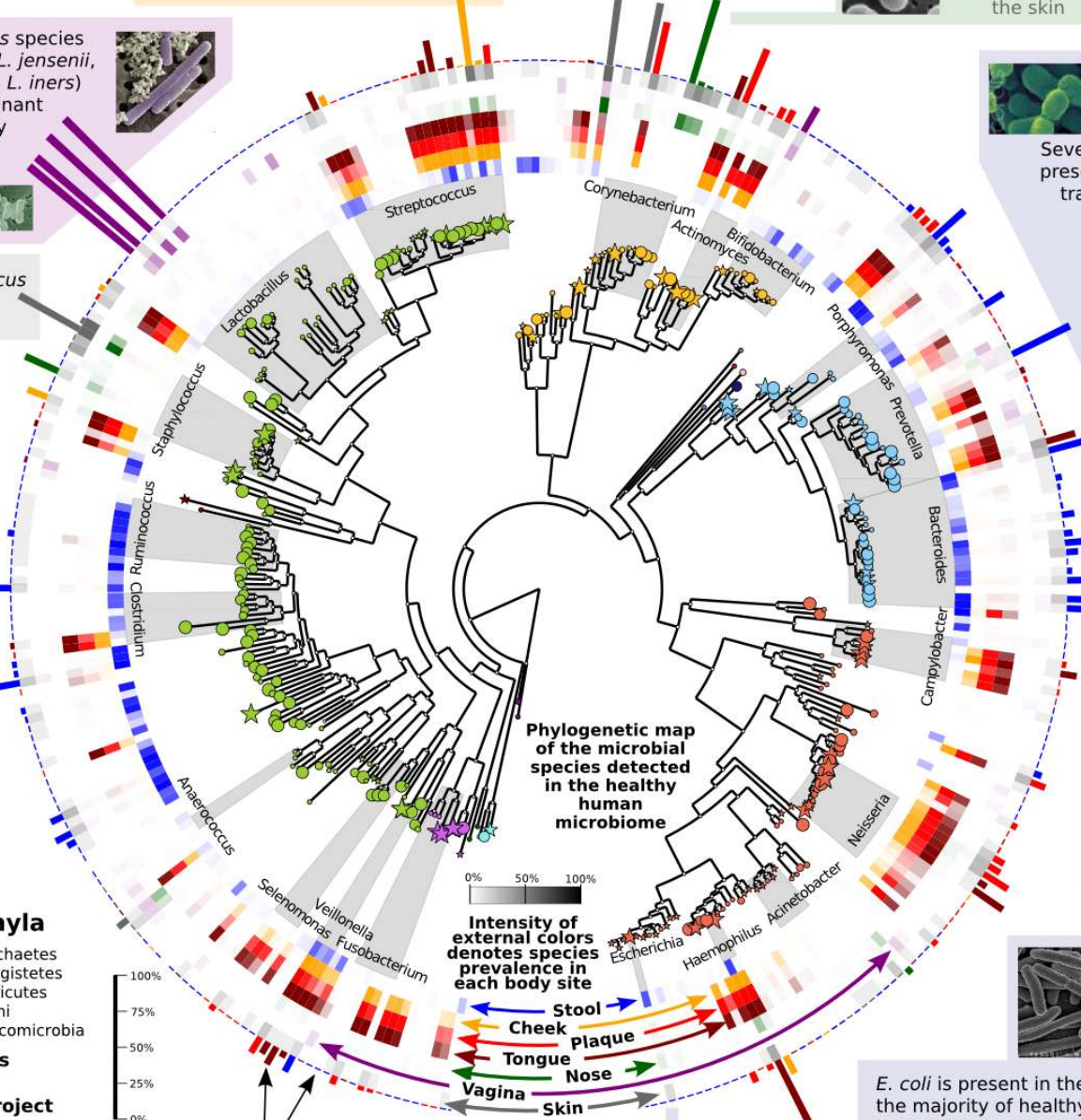
- Actinobacteria
- Bacteroidetes
- Firmicutes
- Proteobacteria

Low abundance phyla

- Chloroflexi
- Cyanobacteria
- Euryarchaeota
- Fusobacteria
- Lentisphaerae
- Spirochaetes
- Synergistetes
- Tenericutes
- Thermi
- Verrucomicrobia

National Institutes of Health
 Human Microbiome Project

N. Segata & C. Huttenhower
<http://huttenhower.sph.harvard.edu>
 (generated using iCITeR and mOTU from MetaPhlAn analysis)



E. coli is present in the **gut** of the majority of healthy subjects but at very low abundance





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”

The screenshot displays the Galaxy web interface. At the top, a navigation bar includes 'Galaxy / Huttenhower Lab', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A status indicator shows 'Using 0%'. On the left, a 'Tools' sidebar lists 'HUTTENHOWER LAB MODULES' with 'LEfSe' selected, showing sub-options A) through F). Below this are 'MetaPhlAn', 'GraPhlAn', 'microPITA', 'MaAsLin', and 'PICRUSt', followed by 'LOAD DATA MODULE' and 'Get Data' with an 'Upload File from your computer' button. The main area features a green success message: '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'. It instructs users to check the 'History' pane for job status. On the right, the 'History' pane shows a list of jobs. The job '8: F) Plot Differential Features on data 3 and data 4' is highlighted with a red circle. Below the job name, it shows '13,522 lines' and 'format: zip, database: ?'. At the bottom of the history list, a red circle highlights a download icon (a disk symbol) next to the job '6: D) Plot Cladogram on data 4'. A red arrow points from the text 'Click here' to the download icon, and another red arrow points from the text 'Then here' to the download icon.

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

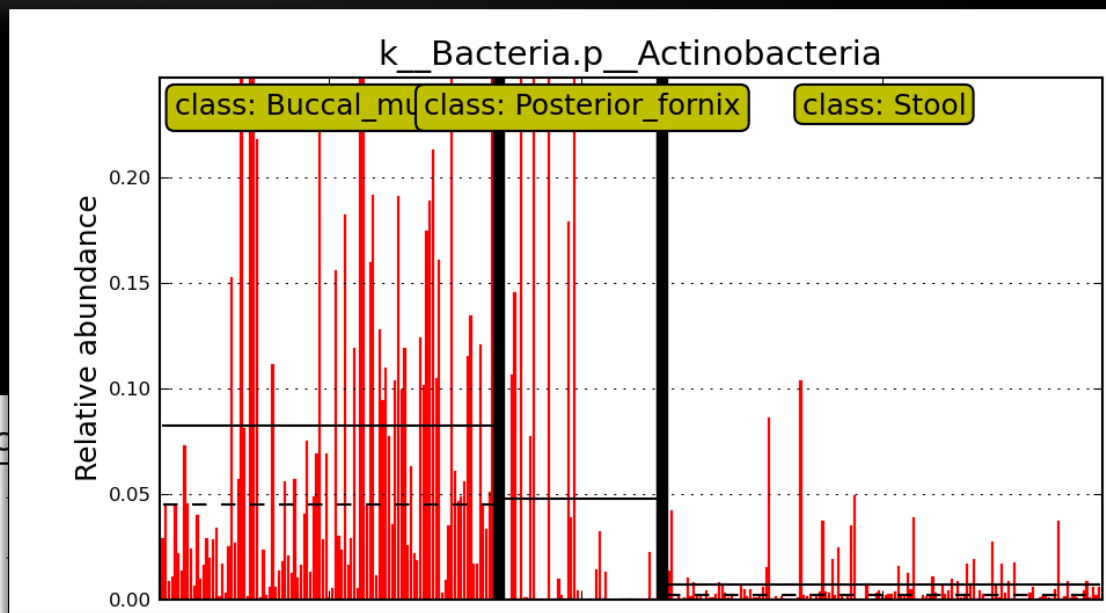
5: C) Plot LEfSe Results on data 4

4: B) LDA Effect Size (LDA)

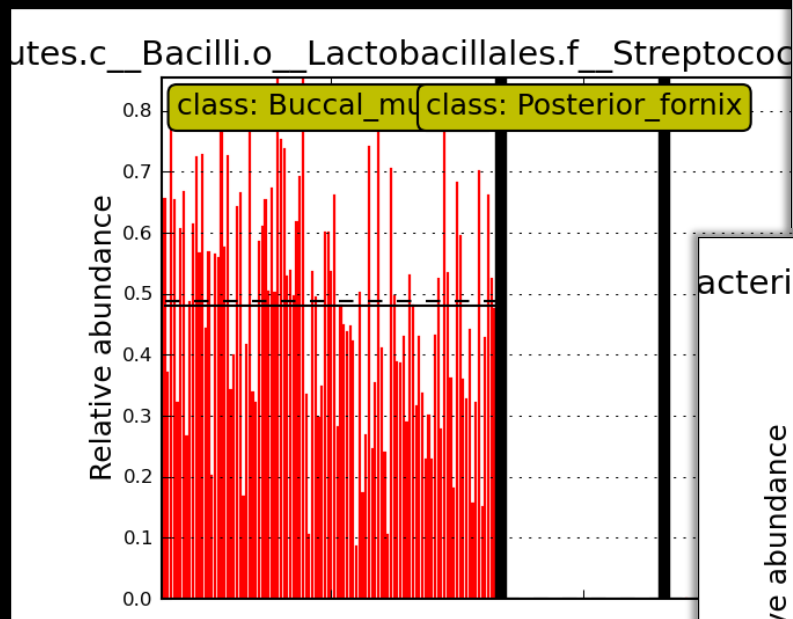


What it means: LEfSe

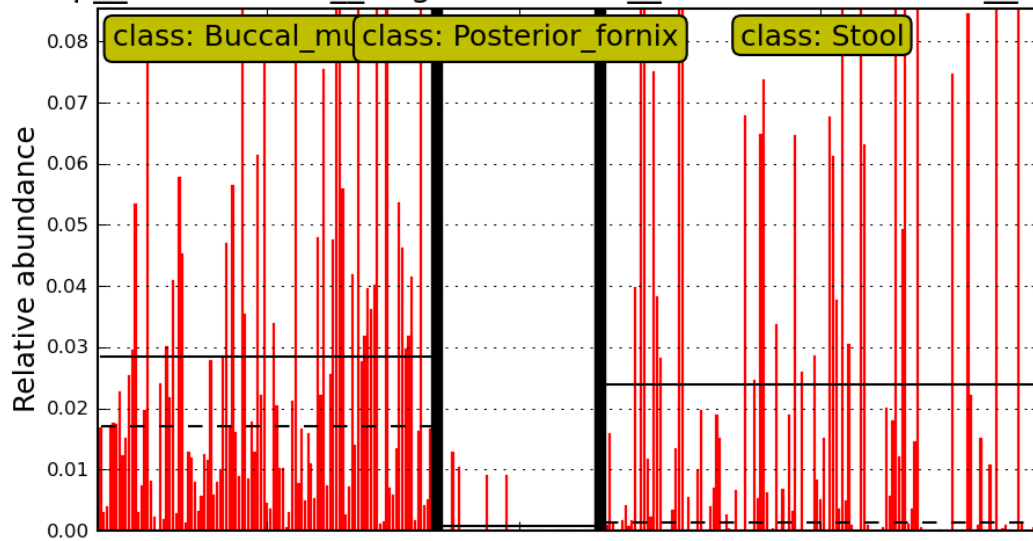
Actinobacteria



Strep. mitis



Actinobacteria.p_Firmicutes.c_Negativicutes.o_Selenomonadales.f_Veillonellaceae



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



Thanks!

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



Alex
Kostic



Levi
Waldron



Xochitl
Morgan



Tim
Tickle



Daniela
Boernigen



Lauren
McIver



Dirk Gevers



Human Microbiome Project 2

Lita Procter
Jon Braun
Dermot McGovern
Subra Kugathasan
Ted Denson
Janet Jansson

Bruce Birren
Chad Nusbaum
Clary Clish
Joe Petrosino
Thad Stappenbeck



George
Weingart



Emma
Schwager



Eric
Franzosa



Boyu
Ren



Tiffany
Hsu



Ali
Rahnava



Ramnik Xavier



Human Microbiome Project

Jane Peterson
Sarah Highlander
Barbara Methe

Karen Nelson
George Weinstock
Owen White



Ayshwarya
Subramanian



Jim
Kaminski



Regina
Joice



Koji
Yasuda



Kevin
Oh



Galeb
Abu-Ali



Wendy Garrett



Andy Chan



Nicola Segata



Gautam Dantas
Molly Gibson



Afrah
Shafquat



Randall
Schwager



Chengwei
Luo



Keith
Bayer



Moran
Yassour



Alexandra
Sirota



Katherine
Lemon



Brendan Bohannon
James Meadow



