#### Meta'omic taxonomic profiling with MetaPhIAn2 and biomarker discovery with LEfSe

**Curtis Huttenhower** 





Harvard School of Public Health Department of Biostatistics







# The bioBakery: a next-generation environment for microbiome analyses

vagrant\_default\_1391533701886\_62205 [Running]
 vagrant\_default\_139153701886\_62205 [Running]
 vagrant\_default\_139153701866\_62205 [Running]
 vagrant\_default\_139153701866\_62205 [

DOD

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)





## Who is there? (taxonomic profiling)

### What are they doing? (functional profiling)







Short Reads

6





Short Reads

7

# Evaluation of MetaPhlAn accuracy



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



### MetaPhIAn2: Taxonomic profiling using unique marker gene sequences

#### 4Gnt synthetic metagenome, 125 organisms



http://huttenhower.sph.harvard.edu/metaphlan2











Eric Franzosa

# MetaPhlAn in action: strain profiling



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

DO

# 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

MM

- Command lines and program/file names appear in a monospaced font.
- Commands you should specifically copy/ paste are in monospaced bold blue.

### Go to <u>http://hmpdacc.org</u>

# HMP

M

#### 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...

Q **N B** 👻 Login REFERENCE IMPACTS ON MICROBIOME TOOLS & ETHICAL HMPDACC OUTREACH GENOMES ANALYSIS HEALTH TECHNOLOGY IMPLICATIONS DATA BROWSER Feedback Welcome to the Data Analysis and Coordination Center (DACC) for the National Institutes of Health (TH) GET DATA 1P Common Fund supported Human Microbiome Project (HMP). This site is the central repository for all data. The aim of the HMP is to characterize microbial communities found at multiple human body sites and to

Click "Get Data"

GETTOOLS

Common Fund supported Human Microbiome Project (HMP). This site is the central repository for all HuP 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.

Areas of Interest



Mana Dublication

### Check out what's available

MM



#### Check out what's available

MM

#### 2 E Q 🔻 Login REFERENCE MICROBIOME IMPACTS ON TOOLS & ETHICAL HMPDACC OUTREACH GENOMES ANALYSIS HEALTH TECHNOLOGY IMPLICATIONS DATA BROWSER Feedback NIH HUMAN MICROBIOME HMIWGS/HMASM - Illumina WGS Reads and Assemblies PROJECT 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 Current News size. 690 samples passed this QC and were included in downstream wgs analyses. June 2012 This dataset includes over 35 billion human contaminant-screened reads in FASTQ format, which are 2.3 TB in size, compressed. Reads from each Owen White and Dirk Gevers discuss individual sample were assembled using SOAP, generating 48.3 million scaffolds with a total compressed size of 13 GB. the HMP on Wisconsin Public Radio June 2012 Data Table DACC website updated in coordination Click on your favorite body site Protocols and Tools with publication of HMP data Related Pages April 2012 HMP DACC Reference Genome download page has been updated Files More News Items SRS ID Reads Size A Reads MD5 Assembly Ass. Size Assembly MD5 Publications Anterior Nares (94 Rows) Ethical Discourse about the Modification of Food for iva (6 Rows) Therapeutic Purpo... Caring about trees in the forest: Buccal Mucosa (123 Rows) incorporating frailty in risk analys... Hard Palate (1 Row) Dietary-fat-induced taurocholic acid promotes pathobiont Left Retroauricular Crease (9 Rows) expansion and... Hid Vagina (2 Rows) More Publications

#### Don't click on anything!

#### Check out what's available

#### - April 2012

HMP DACC Reference Genome download page has been updated

More News Items

DOD

#### 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 Na	ares (94 Rows	s)				
SRS047708	TP FTP	1.7 MB	d786590ff7fec20e8967127991766029	TP:	1.3 KB	ed98eda02d80a137c52b6fa8a3c57833
SRS019215	📑 🛐	10.1 MB	55de248bbfa8c1bbf4447d007330f7ff	📑 🗊	12.1 KB	cab8918433280eafc3d8f6ad78dc1ff7
SRS063178	FTP	13.1 MB	336f0b31b92880224c91ad52c4784adc	TP:	10.7 KB	99de257f1942e98bf1c052e2d046df33
SRS065179	📑 🛐	13.3 MB	27b2c9209bc56cbe219d8c65fa32296c	📑 🛐	54.6 KB	bb8b0d62a3c1923abfcaea01a598a60a
SRS065142	TP:	13.5 MB	3b05d6fcb205106fbd03f314e39f6d63	TP:	7.6 KB	91177065cf438056f2bfc67e99562fe4
SRS018585	📑 🛐	16.8 MB	9d4129d2f5fdd51b9fc899bd84c47b5b	TP FTP	7.9 KB	aa9e9857b26b9efb4fa39bfaf101dc9d
SRS015640	FTP	17.6 MB	595baf36d8b3dcdd21149b3086ccbbee	TP:	52.4 KB	1c7a464db2fccce17c02f9600c867cb1
SRS056210	📑 🖪	18.1 MB	9b2f74b8067e6f20551e6d3b48124c42	📑 🛐	18.3 KB	c4abace0ec0b3e7e5ce1513cb8270e56
SRS018312	FTP	18.9 MB	2454e80d7e5216adf8d5b1850c98738c	TP:	25.4 KB	4f5f760eadd77782862669263e1b1d9d
SRS015450	📑 FB	18.9 MB	eefc0dcf2d52ca5251b01860d54d2bb5	ा हा	107.1 KB	4e0a83868f2fb44f1788dfe1aaa5e13f
SRS049744	FTP	21.5 MB	6d9e2ffc82b08ef37551e902096e4c98	TP FTP	14.3 KB	da7a1cddd3c84b121ff49086432d25d3
SRS012291	📑 🗊	21.9 MB	12775f5df6e71961f1c544e84f6c7342	T FT	8.9 KB	17b5110d391817c7ce52b7c1026df1ba
SRS051600	FTP	22.2 MB	391775b95926a221b8a3cde54a79ae22	TP:	13.9 KB	6db7007edd32b534bc918aad42d600ae
SRS019339	📑 🗊	23.1 MB	76a621d6503d11d1a133a023dc240ae5	T FT	57.3 KB	9255d8206f10ac2611cf45270daa166c
SRS017244	TP:	23.5 MB	b7c2dec67738f317cb8826c09e1a9e39	TP:	21.3 KB	9bcf59e6b4fe15a4e8ccacb0bc824ba8
SRS018671	TP FTP	24.0 MB	7548b06b37038440c5420f7677ff7371	TT FT	135.4 KB	4a180e3ea42a46bcea0a9441b137f243

#### Protocols and Tools

# Getting some (prepped) HMP data

Connect to the server instead

M

 $- \operatorname{cd}$  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

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

#### Department of Biostatistics, Harvard School of Public Health

Contact Documentation People Presentations Publications Research Teaching

Home

M

#### MetaPhIAn: Metagenomic Phylogenetic Analysis

MetaPhIAn is a computational tool for profiling the composition of microbial communities from metagenomic shotgun sequencing data. MetaPhIAn 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 MetaPhIAn 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 MetaPhIAn paper for additional information, validations, and examples. Also the main paper of the Human Microbiome Project uses MetaPhIAn (version 1.1) for species-level metagenomic profiling.

Here is an **infographic** of the application of the **Human Microbiome Project** results obtained applying MetaPhIAn 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

#### Instead, go to <a href="http://huttenhower.sph.harvard.edu/metaphlan2">http://huttenhower.sph.harvard.edu/metaphlan2</a>

	Department of Biostatistics, Harvard School of Public Health
	Contact Documentation People Presentations Publications Research Teaching
Home	You could download MetaPhIAn2 by clicking here

MetaPhIAn v2.0

M

#### MetaPhIAn v2.0: Metagenomic Phylogenetic Analysis

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

- up to 25,000 reads-per-second (on one CPU) analysis speed (orders of magnitude faster compared to existing pranods);
- · unambiguous taxonomic assignments as the MetaPhIAn 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 MetaPhIAn v2.0

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

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

#### Tutorials

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

DOD

€ Bit	tbucket Features	Pricing		owner/repository	٩	⑦ English ▼	Sign up	Log in
0	Overview				HTTPS	<pre>https://bitbuck</pre>	et.org/bioba	kery/me
ш Ш Ф	Last updated Language Access level	2 hours ago Python Read	1 I Fork	5 Tags 5 Watchers	Recent a 1 c Pus 3 Afra	activity S ommit shed to biobakery/Met daab15 README.md h Shafquat · 2 hours ago	taPhIAn2 edited online	with
لا 1 1	<ul> <li>Meta, niAn 2.0</li> <li>Description</li> <li>Pre-required</li> </ul>	): Metagenomic Phylogenetic Ana tion uisites	alysis		Ι c           Pus           ε           Nico	ommit shed to biobakery/Met ecdadce tagging version bla Segata · 4 hours ago	taPhIAn2 on 2.0_beta3	
4	<ul> <li>Basic Us</li> <li>Full com</li> <li>Utility So</li> <li>M</li> <li>Heatman</li> </ul>	sage nmand-line options cripts lerging Tables p Visualization			dod Issu biol Nico	cumentation: how to ue #2 commented on bakery/MetaPhIAn2 bla Segata · 6 hours ago	o best run wi in	th pa
	ہ ۔ MetaPhIAn	2.0: Metagenomic	Phylogenetic	Analysis	Pus   1 Afra	shed to biobakery/Mei .2cceaa README.md h Shafquat · 6 hours ago	aPhIAn2 edited online	with
>>	AUTHORS: Nicola Se	egata (nicola.segata@unitn.it)			/b) 1 c	ommit		

#### The complete MetaPhIAn2 install is in

MM

~/workshop\_data/metagenomics/biobakery/software/metaphlan2

∎ Bit	sian buck	et <sup>Features</sup>	Pricing			owner/repository	٩	? ◄	English -	Sign up	Log in
<b>3</b>	So V	default - 🛃 -	MetaP	hlAn2 /							
Ш		db_v20									
		utils									
¢		.hgtags	205 B	4 hours ago	tagging version 2.0_beta3						
$\mathcal{V}$		README.md	24.6 KB	2 hours ago	README.md edited online with	Bitbucket					
d <sup>1</sup>		metaphlan2.py	35.7 KB	6 hours ago	Making MetaPhIAn exiting grad	ciously when the input	t forma	t cannot	be guessed b	ecause two f	iles are
0 0 4	•	MetaPhIAn 2.0: • Descriptio • Pre-requis • Installatio • Basic Usa • Full comm • Utility Scri • Me • Heatmap • Gra	Metagenom on sites n age nand-line op ipts rging Tables Visualization aPhIAn Visua	nic Phylogenetic A otions s n alization	nalysis						

# From the command line...

• To see what you can do, run:

DOD

metaphlan2.py -h | less

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

00	1. ssh	Ma Ma
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,multifas	ιtq,
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]	
	<pre>[biom biom_output] [mdelim mdelim]</pre>	
	[INPUT_FILE] [OUTPUT_FILE]	

DESCRIPTION

DOD

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
```

### • To launch your first analysis, run:

#### metaphlan2.py

M

- --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  $\sim$ 3-4 minutes

### What did you just do?

- Two new output files:
- 763577454-SRS014459-Stool.fasta.bowtie2out.txt

#### Contains a mapping of reads to MetaPhIAn markers

- 763577454-SRS014459-Stool.txt
  - Contains taxonomic abundances as percentages

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

MM

Re 1. [screen 2: bash] chuttenhower@class:~/tmp (ssh) HWUSI-EAS1625\_615HE:4:100:0:1248/1 gi | 479140210 | ref | NC\_021010.1 | : 1043207-1044529 HWUSI-EAS1625\_615HE:4:100:0:1301/1 gil483877978/ref/NZ\_KB890364.1/:31018-31902 gi|242362078|ref|NZ\_GG692716.1|:28261-29169 HWUSI-EAS1625\_615HE:4:100:1000:167/1 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 gil319644663|ref|NZ\_GL635657.1|:c320982-320029 gi|484001485|ref|NZ\_KB894131.1|:91019-91717 HWUSI-EAS1625\_615HE:4:100:1001:1734/1 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\_E0973491.1|:c620672-618312 gil2244856361refINZ\_E0973490.11:c204903-202990 HWUSI-EAS1625\_615HE:4:100:1003:1644/1 HWUSI-EAS1625\_615HE:4:100:1003:1702/1 gi | 423335209 | ref | NZ\_JH976498.1 | : 329186-330046 gi|238922432|ref|NC\_012781.1|:2910912-2912072 HWUSI-EAS1625\_615HE:4:100:1003:2030/1 HWUSI-EAS1625\_615HE:4:100:1004:353/1 gil223955873|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 gi | 423332908 | ref | NZ\_JH976496.1 | : 1485161-1487113 HWUSI-EAS1625\_615HE:4:100:1007:1428/1 HWUSI-EAS1625\_615HE:4:100:1007:1465/1 gil423332908|ref|NZ\_JH976496.1|:906255-909584 HWUSI-EAS1625\_615HE:4:100:1008:1187/1 gi|224485479|ref|NZ\_E0973214.1|:108053-108250 HWUSI-EAS1625\_615HE:4:100:1008:1241/1 gi|270293478|ref|NZ\_GG730105.1|:c830784-828727 gi | 224514921 | ref | NZ\_DS499545.1 | : 41991-42827 HWUSI-EAS1625\_615HE:4:100:1008:140/1 gi|301307949|ref|NZ\_GG774972.1|:644845-649113 HWUSI-EAS1625\_615HE:4:100:1009:154/1 HWUSI-EAS1625\_615HE:4:100:1009:467/1 gi|303257489|ref|NZ\_GL383997.1|:67163-67873

#### less 763577454-SRS014459-Stool.txt

MX

No. 00 1. [screen 2: bash] chuttenhower@class:~/tmp (ssh) k\_\_Bacteria 100.0 k\_\_Bacterialp\_\_Firmicutes 64.82041 k\_\_Bacterialp\_\_Bacteroidetes 35.17959 k\_\_Bacterialp\_\_Firmicutes/c\_\_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\_\_Firmicutes/c\_\_Clostridialo\_\_Clostridiales/f\_\_Ruminococcaceae 37.71449 k\_\_Bacterialp\_\_Bacteroideteslc\_\_Bacteroidialo\_\_Bacteroidaleslf\_\_Bacteroidaceae 31.5000 k\_\_Bacterialp\_\_Firmicutes/c\_\_Clostridialo\_\_Clostridiales/f\_\_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\_\_Bacteria|p\_\_Firmicutes|c\_\_Clostridia|o\_\_Clostridiales|f\_\_Lachnospiraceae|g\_\_Roseburi k\_\_Bacterialp\_\_Bacteroideteslc\_\_Bacteroidialo\_\_Bacteroidaleslf\_\_Porphyromonadaceaelg\_\_P k\_\_Bacterialp\_\_Firmicutes/c\_\_Clostridialo\_\_Clostridiales/f\_\_Ruminococcaceae/g\_\_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

### • You can finish the job if you like:

M

- 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 .

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

#### mkdir tmp

M

#### 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"

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



• Or use the pre-installed version:

tmev.sh

MM

#### Launch MeV, and select File/Load data

000			Multiple Array V	iewer			
File Adjust Data Metri	ics Analysis	Display Utilities					
🕞 Load Data	B	-	-	- 1).	-		
Load Data nalysis	tatistics		Data Reduction	meta Analysis	Visualization	Miscellaneous	
<ul><li>Save Analysis</li><li>Save Analysis As</li></ul>	Use	the File menu to load data ؛	from text files or a saved	l analysis file. Use the Uti	ilities menu to connect to t	the Gaggle network	
🐻 New Script							
Load Script							
Save Matrix							
Save Image							
💼 Print Image							
Clear Loaded Data X Close							
-							

DOD

- Click "Browse" to your TSV file (763577454.tsv), then
  - Tell MeV it's a two-color array
  - Uncheck "Load annotation"

DOD

- Click on the upper-leftmost data value

00				Express	ion File Load	er		
Select File L	oader He	elp						
File (Tab D	Delimited M	ultiple Samp	ole (*.*))					
Select expre	ssion data t	file /Users	/chuttenh/	Downloads/7	63577454.tsv	1		Browse
Set untille	s	/Users	/chuttenh/	ownloads/7	63577454.tsv	,		
Two	-color Array				⊖ Sir	ale-color Array		
U IWO	-color Allay				0 311	Igie-color Array		
Load Anum								
Loud Annota	ation Data							
Automa	atically dow	nload		from local	file		Load Anr	otation
	,		0					
Choose a	in organism	÷	No file	selected			No	Je loaded.
		±			Choose File			
		•						
Express								
Destaula	76357745	7635. 45	76357745.	76357745.	76357745	76357745		101
Bacteria	100.0	100.0	100.0	2 22625	72 14171	100.0		
k P	0	95 000	0.2233	2.33033	72.14171			
K_Bactori	0	05.00666	0.2200	2.33033	72.14171			
k_Bacteri	0	93.90000	3.31333	2.33033	6 74077			
k_Bacteri	0		3.51409	0.36631	6.74077			
k Pacteri	0		3.51409	0.30031	0.74077			
k Bacteri	0		0	0	2 43846			
k Bacteri	0		0	0 38831	4 30232			
k Bacteri	0	42 97557	0	0.30031	41 42792			
k Bacteri	0	42 97557	0		41 42792			
K_bacteri	0	42.37337	0		41.42752			
Clin L al a	1.0							
Click the up	per-leftmo:	st expressio	on value. Cli	CK the Load	Dutton to fi	nisn.		
			? Me	V + MultiEx	periment	Cancel Load	4	
		_		• viewer	_			

 "Load" your data, then make it visible by: – Display/Set Color Scale Limits

MM

- Choose Single Gradient, min 0, max 10



• Finally, to play around a bit:

DOD

- Display/Set Element Size/whatever you'd like
- Clustering/Hierarchical Clustering
- Optimize both gene and sample order
- And select Manhattan Distance (imperfect!)

	HCL: Hierarchical Clustering								
	MeV								
h	Tree Selection								
L	Gene Tree Gampio								
Ľ	Outering Optimization								
	☑ Optimize Gene Leaf Order								
L	(Leaf ordering optimization will increase the calculation time)								
	Distance Metric Selection								
	Current Metric: Manhattan Distance 🗘								
	(The occurrent distance metric for HCL is Pearson Completion)								
	Use Absolute Distance								
14	Linkage Method Selection	í							
	<ul> <li>Average linkage clustering</li> </ul>								
	Complete linkage clustering								
-	Single linkage clustering								
art art ar	Validation								
	Use Validation (Requires MeV+R)								
-	2 MeV MultiExperiment Reset Cancel OK	1							

### • If you'd like, you can

M

– Display/Sample-Column Labels/Abbr. Names



- 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:

M

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!



# 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



# 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 (Σ = 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

DOC

- 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



Nicola Segata





# Example LEfSe application: Find O<sub>2</sub>-loving bugs (controlling for body site)



# Superimpose enrichments on the tree of life using GraPhIAn





#### LEfSe Associations

#### Metadata Rings

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





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

Tim Tic<u>kle</u>

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

#### • Let's get all of the HMP species data: http://hmpdacc.org/resources/data browser.php

#### MICROBIOME PROJECT

IMXD

#### 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

#### 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 1 

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

#### Click "HMSMCP"

#### Metagenomic Shotgun Sequence

HMIWGS/HMASM Illumina was reads and assemblies

HMBSA Body-site specific assemblies

HMGI Gene Index

HMGC Clustered gene index

HMGS GO slim analysis

Shotgun community profiling

HMSMCP Shotgun MetaPHIAn Community Profiling

MRC Metabolic reconstruction and cluster

HMGOI Genes of Interest HM4WGS/HMHASM Illumina/454 Hybrid reads and assemblies

HMHGI Illumina/454 hybrid 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

UMD Droject Cotolog Defe

Demonstration project data at NCBI

#### Other Data

HMFUNC Functional databases used for metabolic reconstruction RSEQ RNAseg expression analysis of dental microbiome

#### Download the MetaPhIAn1 table for all 700 samples



Protocols and Tools

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

Related Pages

 Caring about trees in the forest: incorporating frailty in risk

 Dietary-fat-induced taurocholic acid promotes pathobiont expansion and...

analys...

MM

## Downloading from the command line

- Instead of saving this, download it by:
  - Right-click to copy the URL
  - Run

M

- wget <paste URL here>
- Note: curl -O <URL> works just as well

- Make sure this file is in your current directory, and expand it: bunzip2 HMP.ab.txt.bz2
- Look at the result

M

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:

• NOW take a look again

Let's modify this file to be LEfSe-compatible

DOD

• scp it to your laptop and open it up in Excel

0	0 0								HMP.ab.	filtered.me	etadata.tsv								1	N. M
2	🋅 🗔 .		× 🗈 🕻	🛅 🎻 🔟	<b>○</b> • <b>○</b> •	Σ • 🛃	•	fx 🛅 💾	100%							Q- Sea	rch in Shee	et		5
1	Home	Layout	Tables	Charts	SmartA	rt Form	ulas D	ata Re	view										_ ^ ‡	F -
-	Edit			Font			Aligr	nment		Nu	umber		For	mat		Cells		Themes		
	🔍 🚽 Fi	II 🔻 Cali	ibri (Body)	- 12	• A• A	-	≡ ab	c 🔻 📆 Wra	np Text 🔻	General		•	N N	lormal			······································	Aab,	•	
		lear • B	JU		🧆 - A	-	F = 6		Merge 🔻	<b>•</b> %	> \$.0	.00 Condit	ional B	ad		cort Dolot	Eormat	Thomas	Aar	
Fa	Δ1	- C		x sid								Forma	tting			sent Delete	Pormat	i memes ×		
1	A	R B	C C	D	F	F	G	н	1	1	K	L	M	N	0	р	0	R	S	Ē
1	sid ,	SRS043001	SRS017127	SRS021473	SRS011134	SRS050184	SRS011529	SRS048164	SRS016516	SRS052330	SRS011355	SRS011452	SRS019787	SRS054776	SRS024140	SRS014683	SRS016018	SRS047014	SRS019601	
2	RANDSID	550534656	159551223	158479027	158499257	508703490	159166850	861967750	159753524	765640925	158944319	159146620	764669880	764224817	159207311	763961826	764447348	765074482	765620695	Ľ
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	Q2_2009	Q3_2009	
4	GENDER	female	male	male	male	female	male	male	female	female	female	male	male	male	male	male	male	male	female	
5	VISNO	1	1	2	1	1	1	1	1	1	1	1	. 2	2	2	1	1	2	1	
6	STSite	Stool	Buccal_muco	Buccal_muc	Stool	Posterior_to	Stool	Stool	Posterior_to	Posterior_to	Posterior_to	Stool	Stool	Buccal_muc	Buccal_muc	Stool	Stool	Stool	Stool	
/	Parent_spec	700106291	700033688	700097185	700014832	700038759	700016608	700038870	700032243	700038805	700015577	700016136	700038231	700106652	700100608	700023337	700024646	700105580	700038072	
0	lane	704GE 6	OINLAAAA	OIKZLAAAA	GIJGUAAXA	704N4	OIPINF		01VKUAAAA	1055111	01KTVAAAA	OLVVER	7041010	021F0	01JUIAAAA	014111111	OIPPK	010000	3	
10	SRS	700106291	700033689	700097185	700014837	700038759	700016610	700038870	700032243	700038805	700015579	700016142	700038263	700106652	700100608	700023337	700024673	700105580	700038072	
11	Mean Quality	29.75	29	33	27	31.07	31.92	33.16	33	32.91	24	32.97	700050205	32.65	29	700025557	34.24	700105500	/000500/2	
12	Number of O	*****	*****	*****		*****	*****	*****	*****	*****	*****	*****	t .	******	*****		*****			
13	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			
14	Unique Non-		******	959383209	******	******	******	******	******	******	683832927	******		******	******		******			
15	kBacteria	0	0	0	0.59019	0	0.15046	1.46625	0	0	0	0	0	0	0	0	0	0.66005	0	
16	kBacteria	0	0	0	0.00417	0	0	0	0	0	0	0	0 0	0	0	0	0	0	0	
17	kBacteria	0	0	0	0	0	0	0	0	0	0	0	0 0	0	0	0	0	0	0	
18	kBacteria	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19	kBacteria	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	
20	KBacteria   kBacteria	0 005 37	0	0	0.000250	0	0.00189	0 01000	0	0	0.01011	0	0	0	0	0 02015	0 00470	0	0	
21	kbacterial	0.00527	0	0	0.00358	0	0.00975	0.01062	0	0	0.01611	0		0	0	0.02015	0.00478	0.00568	0.04259	
23	k Bacterial	0	0	0	0	0	0	0	0	0	0	0	0	0.00686	0	0	0	0	0	
24	k Bacterial	0	0	0	0	0	0	0	0	0	0	0	0	0.00000	0	0	0	0	0	
25	k_Bacteria	0 0	0	0	0	0	0	0.00105	0	0	0	0	0	0	0	0.00395	0	0	0.00172	
26	kBacteria	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
27	kBacteria	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	
28	kBacteria	0	0	0	0	0	0	0	0	0	0.00528	0	0.00049	0	0	0.00441	0	0.00445	0.00157	
29	kBacteria	0	0	0.03817	0	0	0	0	0	0	0	0	0	0.0182	0	0	0	0	0	
30	kBacteria	0	0	0	0	0	0.00189	0	0	0	0	0	0	0	0	0	0	0	0	
31	kBacteria	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
32	KBacteria	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	
		← → → F F	IMP.ab.filtere	d.metadata.t	tsv +															
	Norm	nal View	Ready								Sum=0		•							1

- Delete all of the metadata rows *except*:
  - RANDSID and STSite

DOD

- Save it as tab-delimited text: HMP.ab.filtered.metadata.txt

0	00						🖄 HMP.ab.filt	tered.metada	ta.txt							N. 21
2	1		× 🗅 🕻	- 		🝸 • 🖅 🛅 🕻	100% -	<b>?</b>				Q- Sea	rch in Shee	t		2
	A Home	Layout	Tables	Charts	, Amarika , Karrah	ori, Bata ( B									∧ ‡	× -
	Edit			Font		Save As: HMI	.ab.filtered.r	netadata.txt				Cells		Themes		
ſ	🗎 🖕 💽 Fi	ill 🔻 Cali	bri (Body)	- 12	·		ng Test 🐑 🔤 Ca	neral -						Aa -	-	
		lear • B	I U						1			Delete	Format	Thomas	la-	
	aste								• 4			isert Delete	Format	memes *		-
	AI		S (- )	C D	FAVORITES	Name					Date Modified	D	0	P	c	ľ
1	RANDSID	550 4656	159551223	158479027	Dropbox	104b-hit-ke	g-mpm-cop-i	nul-nve-nve.t	ct		Yesterday	1 763961826	764447348	765074482	765620695	t
2	STSite	Stool	Buccal_mucc	Buccal_muc	Annliestions	763577454	_species.tsv				Today, 9:25	cc Stool	Stool	Stool	Stool	ľ
3	kBacteria	0	0	(	Applications	763577454	.tsv				Today, 9:25	D 0	0	0.66005	0	1
	kBacteria	0	0		Desktop	📄 fastq2fasta	.py				Today, 9:57	0 0	0	0	0	1
6	k manual	0	0		Documents	HMP.ab.filt	ered.metadata	.tsv			Today, 11:30	0 0	0	0	0	5
7	k_Bacteria	2.57633	0	Ċ	Downloads	🐘 🐘 HMP.ab.filt	ered.metadata	.txt			Today, 11:31	5 6.11771	11.53469	5.17385	7.56705	
8	kBacteria	0	0	(	Downloads	HMP.ab.filt	ered.txt.gz				Today, 11:27	0 0	0	0	0	i -
9	kBacteria	0.00527	0	(	Movies	p06-seque	ncing.docx				03-05-12	0 0.02015	0.00478	0.00568	0.04259	1
10	kBacteria	0	0		J Music	problems00	5.tar.gz				02-26-12	0 0	0	0	0	1
11	K_Bacteria		0	-							Yesterday	0 0	0	0	0	1
13	k Bacteria		0		O Pictures					0 0	,	0 0.00395	0	0	0.00172	i -
14	k_Bacteria	0	0	(		rmat: Windows I	ormatted Te	vt ( tvt)	•			0 0	0	0	0	j .
15	kBacteria	0.09916	0	0.00175		mat. windows i	offiatted re.	At (.t.At)				2 0.14181	1.38231	0.39615	11.37444	6
16	kBacteria	0	0	(	Description							0 0.00441	0	0.00445	0.00157	1
17	kBacteria	0	0	0.03817	Exports the data on the	e active s.	indows-compat	ible text 6	ses tabs to se	eparate		0 0	0	0	0	1
18	k_Bacteria		0		values in cells.							0 0	0	0	0	-
20	k Bacteria	0.00187	0.00506	0 11746								7 0.01003	0	0.00384	0	
21	k Bacteria	0.00107	0.00500	(	Learn more about file f	formats						3 0	0	0.00501	0	
22	kBacteria	j o	0	(								5 0.00159	0.00946	0.0002	0.01763	í -
23	kArchaea	0	0	(								0 0	0	0.04344	0	1
24	kBacteria	0	0	(	Options Co	mpatibility Repor	t 🔄 🔼 Cor	mpatibility chee	k recommended			0 0	0	0	0	1
25	kArchaea	0 01250	0									0 00000	0.00079	0.02601	0 1145	-
20	k Bacteria	0.01359	0	-	Ouri i i							9 0.04046	0.12900	0.02091	0.1145	
28	k Bacteria	0	0.00923	0.04524	Hide extension	New Folder				Cancel	Save	7 0	0	0	0	
29	kBacteria	0	0	(	0 0	0	0	0	0	0 0	U	0 0.00208	0	0.00087	0.00239	í –
30	kBacteria	0	0	(	0 0	0 0	0 0	0	0	0 0	0	0 0	0	0	0	1
31	kBacteria	0	0	(	0 0	0 0	0 0	0	0	0 0	0	0 0.0347	0	0	0	1
32	kBacteria	0	0.07358	0.71281	0 0	0 0	0 0	0	0	0 0	10.69807 1.55	03 0	0	0.00159	0	1
		←→→⊡ Ĥ	MP.ab.filtere	d.metadata	tsv +										ii ii	
			Ready					Si	um=0	-						1

#### Visit LEfSe at: <a href="http://huttenhower.sph.harvard.edu/lefse">http://huttenhower.sph.harvard.edu/lefse</a>

DOD

🔫 Galaxy / Huttenho	WET Lapalyze Data Workflow Shared Data - Visualization Help- User-		Using 0%
Tools		History	C 🕈
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  MetaPhIAn  GraPhIAn  First  MaAsLin  PICRUSt  LCACORNELODULE  Get Data  Upload File from your computer	<ul> <li>Thanks for visiting our lab's tools and applications page, implemented within the <u>Galaxy</u> 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.</li> <li>Our lab's research interests include metagenomics and the <u>human microbiome</u>, 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 <u>research</u> and <u>publications</u>, including the <u>Sleipnir library</u> for computational functional genomics.</li> <li>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.</li> <li>If you have any comments, questions, or suggestions, please contact <u>Dr. Huttenhower</u>.</li> </ul>	Unnamed history 0 bytes This history is em load your own da from an external	Q 🗹

Then upload your formatted table

DOD

- After you upload, wait for the progress meter to turn green!

- Galaxy / Huttenhor	WEF Labalyze Data Workflow Shared Data - Visualization Help- User-		Using 0%
Tools	Upload File (version 1.1.4)	History	C \$
search tools	File Format:	Unnamed history	
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	Auto-detect Which format? See help below 1. Click here, browse to HMP.ab.filtered.metadata.txt Choose File Htm.ab.filtered.metadata.txt H.C. 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:	269.2 KB This history is empload your own date from an external s 3. Th	Q 🗹
GraPhlAn	Here you may specify a list of URLs (one per line) or paste the contents of a file.	, viait	
microPITA Maaalin	Convert spaces to tabs:	her	e
MaAsLin PICRUSt	Use this option if you are entering intervals by hand. <b>2 Thom</b>		
LOAD DATA MODULE	Genome:		
<u>Get Data</u>	unspecified (?)		
Upload File from your computer DEFAULT GALAXY MODULES	Execute		

• Then tell LEfSe about your metadata:

DOD

💳 Galaxy / Huttenho	Wer Lanalyze Data Workflow Shared Data - Visualization Help - User -	===	Using 0%
	A) Format Data for LEfSe (version 1.0)	History	C \$
search tools	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	Unnamed history 538.3 KB 2: HMP.ab.filtered.me ata.txt	Q 🗹 tad 💿 🖋 🗙
B) LDA Effect Size (LEFSe)	columns: Rows \$		
<u>C) Plot LEfSe Results</u> <u>D) Plot Cladogram</u>	Select within row to use as the 2. Then		
<u>E) Plot One Feature</u> <u>F) Plot Differential Features</u>	Select Winch row to use us subclass: select STSite		
<u>MetaPhIAn</u> <u>GraPhIAn</u>	Select which come to use an effect: #1:RANDSID	Then sel	ect
MaAsLin PICRUSt	Per-sample normalization of the sum of the values to 1M (recommended when very low values are present):	RANDSI	D
LOAD DATA MODULE Get Data Upload File from your computer	Execute 4. Then h	ere	

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

MM

💳 Galaxy / Huttenhov	Ver Labalyze Data Workflow Shared Data - Visualization Help- User	r∓		Using 0%
Tools	B) LDA Effect Size (LEfSe) (version 1.0)		History	52 \$
search tools	Select data: D 🗠 2. Then "4" he	ere	Unnamed history	
here	3: A) Format Data for LEfSe on data 2 + (finds only very extre	eme	1.0 MB	Q
HUTTENHOWER LAB MODULES	Alpha value for the factorial Kruska' Wallis test among classes:		2: A) Format Data for	
<u>LEfSe</u> A) Format Data for LEfSe	0.05 differences)		<u>Se on data 2</u>	
B) LDA Effect Size (LEfSe)	Alpha value for the pair use Wilcoxon test between subclasses:		2: HMP.ab.filtered.met	ad 💿 🖋 🗙
C) Plot LEffe Popults	0.05		<u>ata.txt</u>	
D) Plot Cladogram	Time hold opene logarithmic LDA score for discriminative features:			
E) Plot One Feature	4			
E) Plot Differential Features	Do you want the pairwise comparisons among subclasses to be performed only	/		
r) not binerential reatures	among the subclasses with the same name?:	<b>→</b>	66	1
<u>MetaPhIAn</u>	No ‡	3. III	nen "one"	nere
<u>GraPhIAn</u>	Seture strategy for more class analysis:	finda	differences in a	atlagat
microPITA	One-against-all (less strict) 💠	inas d	unerences in a	alleast
MaAsLin		one co	ndition rather	than in
<u>PICRUSt</u>	Execute 4. Then GO!		all conditions)	
LOAD DATA MODULE			,	

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

DOD

💳 Galaxy / Huttenhov	Ver Labalyze Data Workflow Shared Data - Visualization Help- User-	Using 0%
Tools	C) Plot LEfSe Results (version 1.0)	History 2 🌣
search tools	Select data: 🗅 🖓	Unnamed history
HUTTENHOWER LAB MODULESICK	4: B) LDA Effect Size (LEFSe) on data 3 +	1.0 MB
A) Format De a for LEFS	Default ÷	
B) LDA Effect Size (LEfSe)	Set some graphical options to personalize the output:	3: A) Format Data for LEf 💿 🖋 🗙 Se on data 2
D) Plot Cladogram	Output format:	2: HMP.ab.filtered.metad (*)
E) Plot One Feature	Set the dpi resolution of the output:	
F) Plot Differential Features	2 Then here	
MetaPhIAn		
GraPhIAn		

#### • In Galaxy, view a result by clicking on its "eye"

DOD

🗧 Galaxy / Huttenho	Wer Labalyze Data Workflow Shared Data - Visualization Help- User-		Using 0%
Tools		History	C 🕈
search tools	A job has been successfully added to the queue – resulting in the following dataset: 5: C) Plot LEfSe Results on data 4 You can check the status of queued jobs and view the resulting data by refreshing the	Unnamed history 1.4 MB	QC
LEFSe A) Format Data for LEFSe	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.	<u>5: C) Plot LEfSe Results ( n data 4</u>	• ×
B) LDA Effect Size (LEfSe)		<u>4: B) LDA Effect Size (LEf</u> <u>Se) on data 3</u>	● # ×
D) Plot Cladogram		<u>3: A) Format Data for LEf</u> <u>Se on data 2</u>	● / ×
<u>E) Plot One Feature</u> <u>F) Plot Differential Features</u>		2: HMP.ab.filtered.metad ata.txt	● 🖋 ×
<u>MetaPhIAn</u> GraPhIAn			
microPITA			
MaAsLin			
<u>PICRUSt</u>			

**Click here** 

Buccal mucosa Posterior fornix Stoo c Bacte o\_Bacteroidale p\_Bacteroid f Bacteroidacea g Bacteroide o\_Clostridiale c\_Clostridi les unclassifie f Rikenellacea g\_Alistip Alistipes\_putredi Bacteroides vulgati f\_Ruminococcacea f\_Eubacteriacea g Eubacteriu f\_Prevotellace g\_Prevotell \_\_\_\_\_Bacteroides\_ovati g\_Parabacter \_Eubacterium\_recta Bacteroides sterco g\_Ruminococci g Faecalibacteriu p\_Verrucomicrob Bacteroides\_cac \_\_\_\_\_\_Erysipelotrichal f Lachnospiracea g\_Akkermans nsia muciniph s Dialister invisi ococcus\_bron g\_Dialist Alistipes\_shah g Lactobacillus f\_Lactobacillacea c\_Bacil o Lactobacillale p Firmicute illus\_crisp actobacillus jenser o Bifidobacteriale g\_Streptococci Streptococcus\_mit p\_Proteobacter f Pasteurellacea o\_Pasteurellale g\_Haemophili us parainfluenza o\_Actinomycetale c\_Actinobact n Actinobacter f\_Bacillales\_un g\_Gem o\_Bacillale f\_Micrococcacea g\_Roth mohrsa Betaproteobacter g\_Lautrop \_Lautropia\_mirabil o Neisseriale: f Neisseriacea g\_Neisser influenz o Selenomonadale \_Rothia\_mucilaging f\_\_veillonellace c Negativicute g\_\_veillonella \_Rothia\_dentocario 0 LDA SCORE (log 10)

M



• You can plot the results as a cladogram

DOD

- Lots and lots of graphical parameters to modify if desired

💳 Galaxy / Huttenhov	WET Labalyze Data Workflow Shared Data - Visualization Help- User-		Using 0%
Tools 1 Click *	D) Plot Cladogram (version 1.0)	History	C \$
search tools here	Select data: <sup>(1)</sup> <sup>(2)</sup> <sup>(2)</sup> <sup>(4)</sup> <sup>(4)</sup> <sup>(5)</sup> <sup>(4)</sup> <sup>(5)</sup> <sup>(6)</sup>	Unnamed history 1.4 MB	QØ
HUTTENHOWER LAB MODULES LEfSe A) Format Data for LEfSe	Set structural parameters of the cladogram:	<u>5: C) Plot LEfSe Resu</u> <u>n data 4</u>	
B) LDA Effect 1 Ze (LEfSe)	Default +	<u>4: B) LDA Effect Size</u> <u>Se) on data 3</u>	<u>(LEf</u> 🕑 🖋 🗙
D) Plot Cladogram	Default +	<u>3: A) Format Data fo</u> Se on data 2	r LEf 💿 🖋 🗙
E) Plot One Feature F) Plot Differential Features	Output format:	2: HMP.ab.filtered.mo ata.txt	etad 💿 🖋 🗙
MetaPhlAn Graphlan	Set the dpi resolution of the output:		
microPITA MaAsLin	Execute 2. Then here		
PICRUSt			



# An aside: GraPhIAn

• You can use this visualization for other purposes as well

Available online through Galaxy

M

- Available offline as open source Python

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





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

MM

= Galaxy / Huttenhov	WCT Labalyze Data Workflow Shared Data - Visualization Help- User-		Using U%
Tools	F) Plot Differential Features (version 1.0)	History	C 🕈
search tools 1. Click	The formatted datasets. D 45 3: A) Format Data for LEfSe on data 2 💠	Unnamed history 1.8 MB	QØ
A) Format Data for LEFSe	4: B) LDA Effect Size (LEfSe) on data 3 ÷	<u>6: D) Plot Cladogram on</u> data 4	• 🖋 🗙
B) LDA Effect ize (LEfSe)	Do you want to plot all features or only those detected as biomarkers?: Biomarkers only + 2. Then selected	<u>5: C) Plot LEfSe Results o</u> <u>n data 4</u>	• / ×
D) Plot Clado ram	Set some graphical options to personalize the output: Default + YOUR formatted	<u>4: B) LDA Effect Size (LEf</u> Se) on data <u>3</u>	④ ∦ ×
E) Plot One Feature E) Plot Differential Features	png ÷ data here	<u>3: A) Format Data for LEf</u> Se on data 2	• / ×
<u>MetaPhIAn</u>	Set the dpi resolution of the output:	2: HMP.ab.filtered.metad	• / ×
<u>GraPhIAn</u> microPITA	Execute 3. Then here		
MaAsLin			
PICRUSt			

**Click here** 

#### • In Galaxy, download a result by clicking on its "disk"

M

💳 Galaxy / Huttenho	Wer Labalyze Data Workflow Shared Data - Visualization Help - User -	Using 0%
Tools		History 🎜 🛠
search tools	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	Unnamed history
HUTTENHOWER LAB MODULES	You can check the status of queued jobs and view the resulting data by refreshing the	5.7 MB
<u>LEfSe</u>	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	8: F) Plot Differential F 💿 🖋 🗙
A) Format Data for LEfSe	ministrea in completed successionly of error in problems were cheoditered.	eatures on data 3 and data 4
B) LDA Effect Size (LEfSe)		Is,510 lines
C) Plot LEfSe Results		format: <b>zip</b> , database: <u>?</u>
D) Plot Cladogram		Exporting k Bacteria n Eirmicutes c Clostri
E) Plot One Feature		K_bacteria.prinnicutes.cclostn
F) Plot Differential Features		Exporting k Bacteria.p Bacteroidetes.c Bact
<u>MetaPhIAn</u>		Europeine
<u>GraPhIAn</u> microPITA		k_Bacteria.p_Proteobacteria.c_Bet
MaAsLin	Then here	
PICRUSt		
LOAD DATA MODULE		binary file
<u>Get Data</u>		6: D) Plot Cladogram o
Upload File from your computer		n data 4
DEFAULT GALAXY MODULES		5: C) Plot LEfSe Result ( ) X
Convert Formats		<u>s on data 4</u>
FASTA manipulation		4· R) I DA Effect Size (I 🖉 🖉 🛩

k Bacteria.p Actinobacteria

class: Stool

Veillonel

class: Buccal my class: Posterior fornix

#### Actinobacteria

#### Strep. mitis

M



0.00

0.20

0.15

# Summary

MetaPhlAn2

M

- 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

George

Weingart

Ayshwarya

Subramanian

Afrah

Shafquat



Emma

Schwager

Jim

Kaminski

Randall

Schwager

MM



Levi Waldron











Eric

Franzosa

Regina

Joice

Chengwei

Luo







Boyu

Ren

Koji

Yasuda

Keith

Bayer





Daniela

Boernigen









Moran Yassour



Galeb Abu-Ali



Alexandra Sirota



Wendy Garrett



Andv Chan



Nicola Segata



Gautam Dantas Molly Gibson



Brendan Bohannan James Meadow



Dirk Gevers

Lita Procter Jon Braun Dermot McGovern Subra Kugathasan Ted Denson







Ramnik Xavier

Jane Peterson **Barbara** Methe

Janet Jansson



Human Microbiome Project

Karen Nelson George Weinstock **Owen White** 











Lemon





Kevin Oh







Ali Rahnavard



Lauren

Mclver









