



Metagenomic taxonomic profiling with MetaPhlAn2

Galeb Abu-Ali

Eric Franzosa

Curtis Huttenhower

09-17-15



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



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

Huttenhower Lab Tools

Welcome to the official Huttenhower Tutorials wiki.

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



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

Composition Analysis

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

HUMAnN • Microbial species and associated genes and pathways	MetaPhlAn • Microbial species and abundances	PhyloPhlAn • Reconstruction of phylogenetic trees	PICRUSt • Predict metagenome functional content from marker gene	ShortBRED • Abundance of proteins of interest in genetic data
---	---	--	---	--

Statistical Analysis

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

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

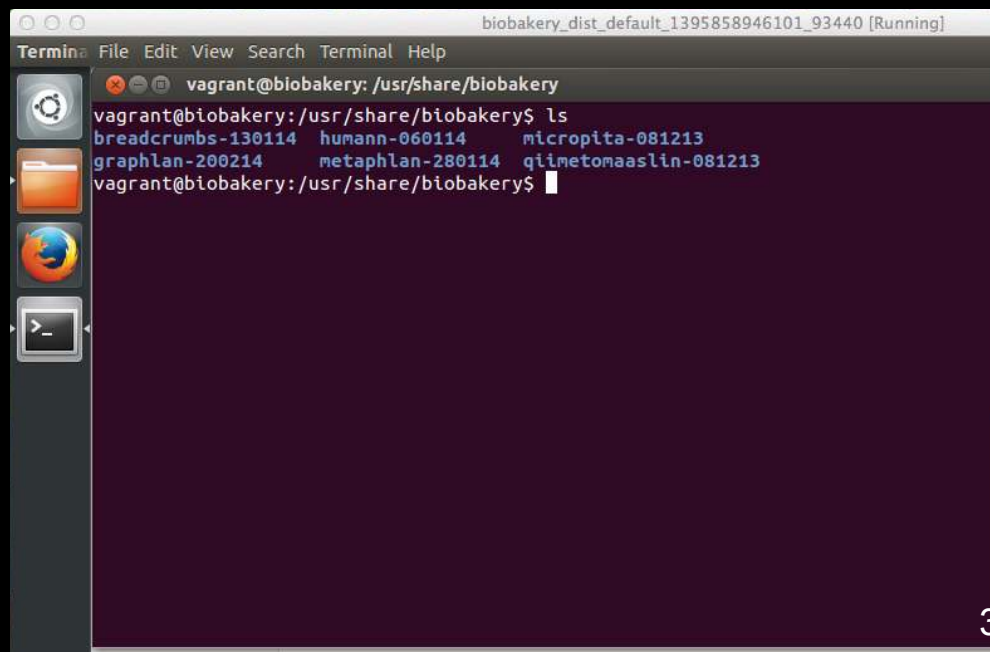
These tools can help visualize taxonomical and phylogenetic information for (i) microbial composition/taxonomy data, (ii) outputs from MetaPhlAn, LEfSe, HUMAnN, MaAsLin. Please click on the link below for detailed tutorial:



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)



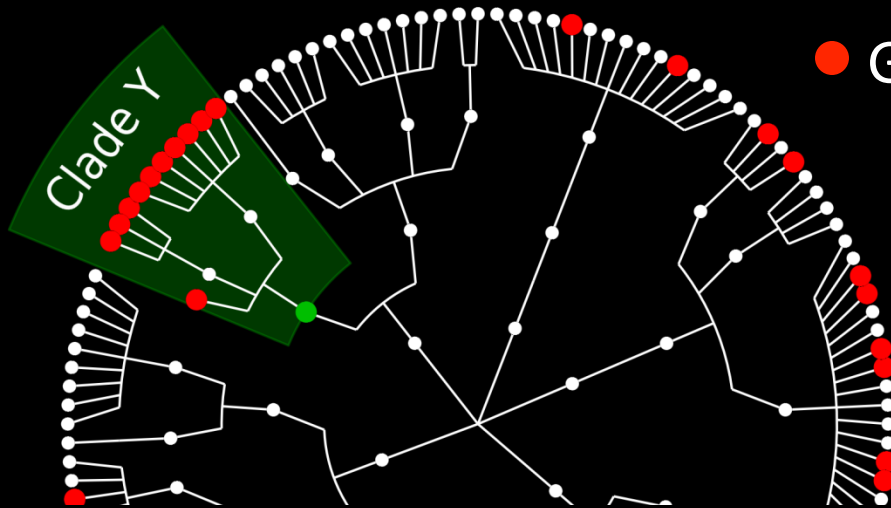
MetaPhlAn overview

Nicola Segata

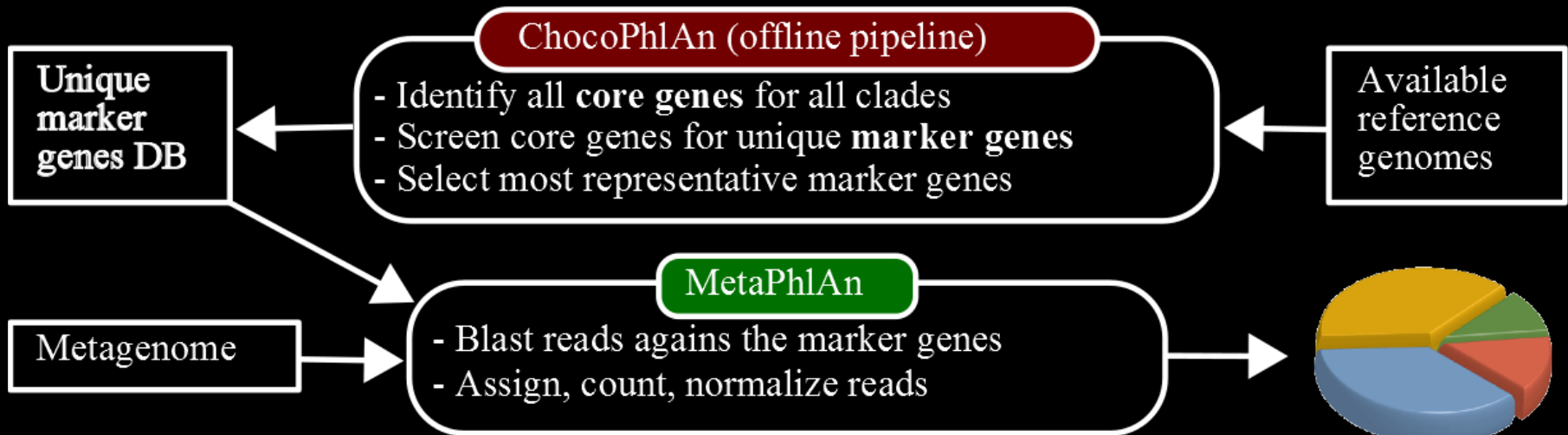
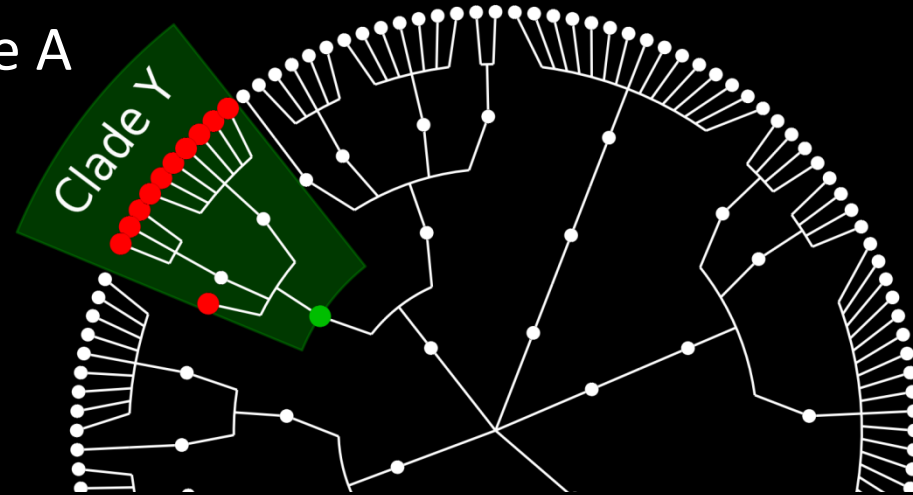


A is a **core gene** for clade Y

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



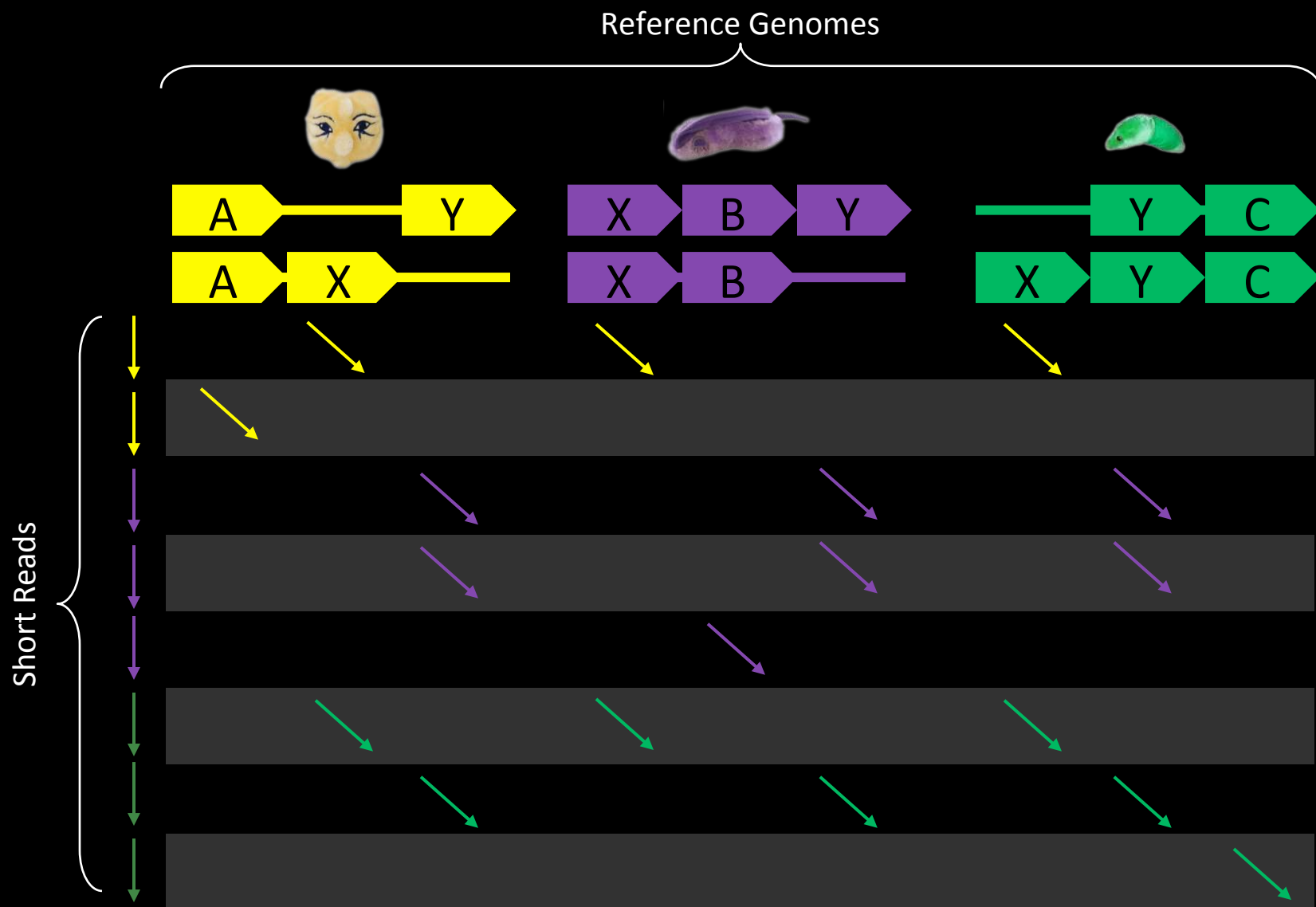
● Gene A





MetaPhlAn

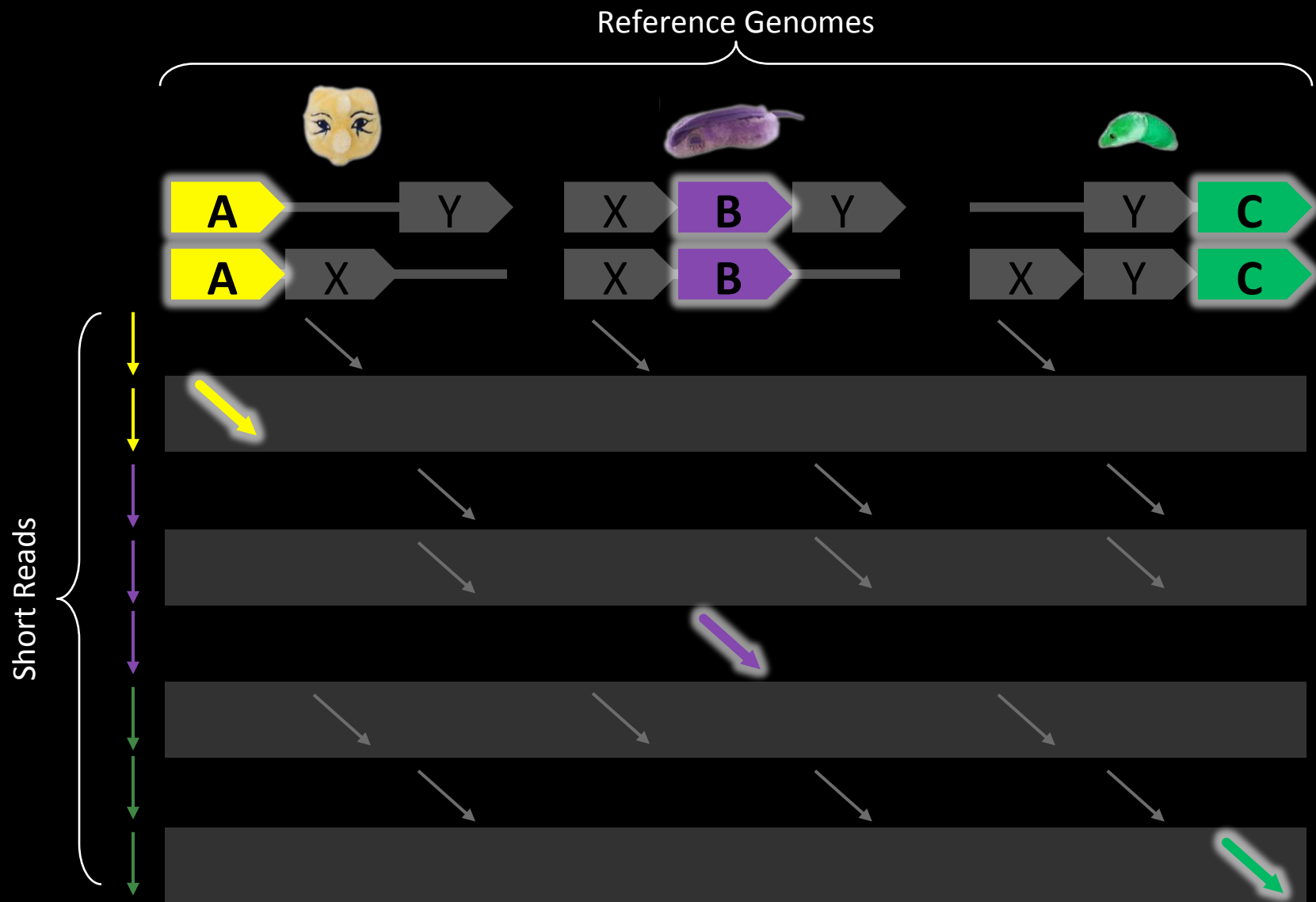
Metagenomic Phylogenetic Analysis





MetaPhlAn

Metagenomic Phylogenetic Analysis

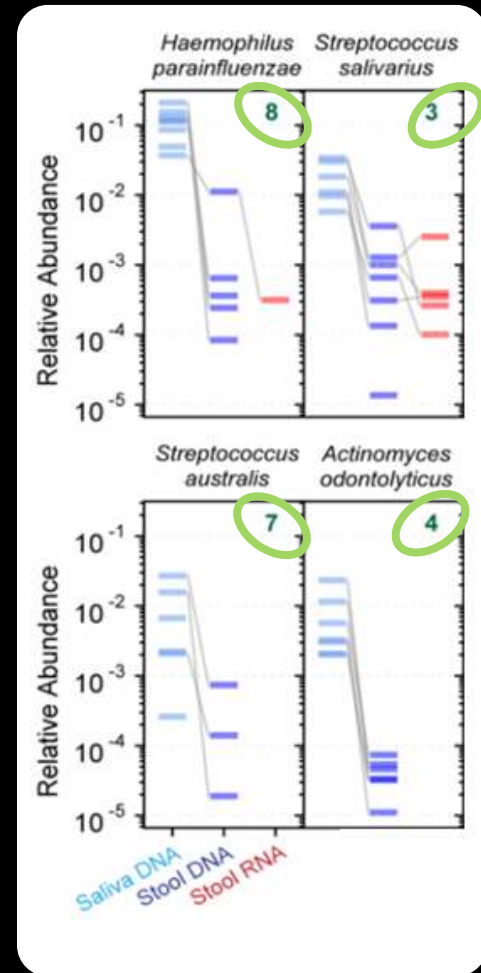
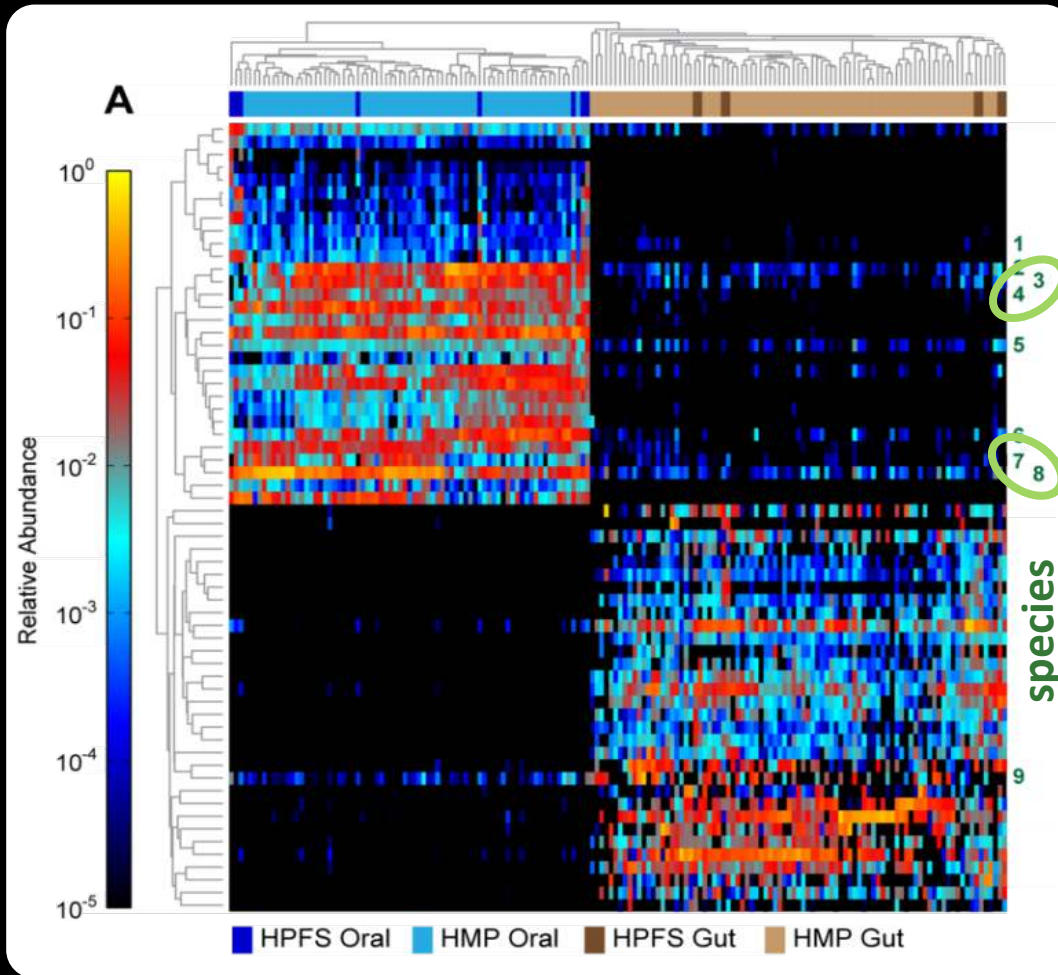




MetaPhlAn in action

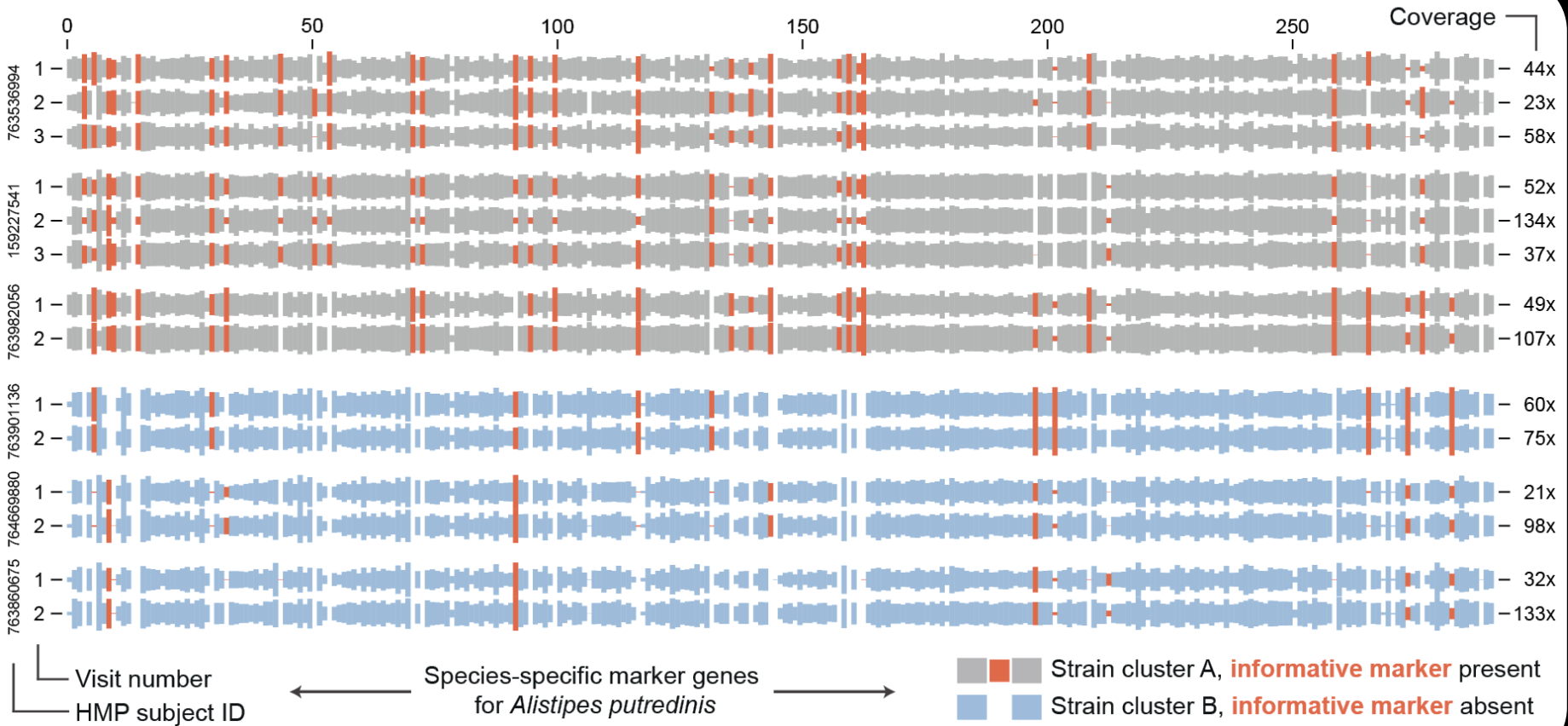


Eric Franzosa





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



Getting some HMP data

Don't click on anything!

- Check out what's available

Apr 2012
HMP DACC Reference Genome
download page has been updated

[More News Items](#)

Publications

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

[More Publications](#)

Data Resources

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

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	595baf36d8b3dcd21149b3086ccbbec		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 /home/ubuntu/metagenomics/data/input/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



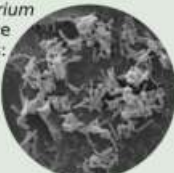
Streptococcus dominates the oral cavity with *S. mitis* > 75% in the **cheek**



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 files used to run MetaPhlAn v2.0, as well as a README file that includes the following information:

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

Tutorials

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



Who's there: MetaPhlAn2

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

The screenshot shows the Bitbucket repository page for 'biobakery/metaphlan2'. The page includes a navigation bar with 'Dashboard', 'Teams', 'Repositories', 'Snippets', and 'Create'. A warning message states: 'This repository's size is over 1 GB. Read more on how to reduce the size of your repository.' The 'Overview' section displays repository statistics: 1 Branch, 8 Tags, 2 Forks, and 7 Watchers. The left sidebar contains a list of links: 'MetaPhlAn 2.0: Metagenomic Phylogenetic Analysis', 'Description', 'Pre-requisites', 'Installation', 'Basic Usage', 'Full command-line options', 'Utility Scripts' (with sub-links for 'Merging Tables' and 'Heatmap Visualization'), 'GraPhlAn Visualization', and 'Customizing the database'. The main content area shows the repository title 'MetaPhlAn 2.0: Metagenomic Phylogenetic Analysis', the author 'AUTHORS: Nicola Segata (nicola.segata@unitn.it)', and a 'Description' section. The description states: 'MetaPhlAn is a computational tool for profiling the composition of microbial communities (Bacteria, Archaea, Eukaryotes and Viruses) from metagenomic shotgun sequencing data with species level resolution. From version 2.0 MetaPhlAn is also able to identify specific strains (in the not-so-frequent cases in which the sample contains a previously sequenced strains) and to track strains across samples for all species. MetaPhlAn 2.0 relies on ~1M unique clade-specific marker genes (the marker information file can be found at src/utills/markers_info.txt.bz2 or here) identified from ~17,000 reference genomes'. The right sidebar shows 'Recent activity' with a list of commits and their descriptions.

Bitbucket Dashboard Teams Repositories Snippets Create

Find a repository...

This repository's size is over 1 GB. Read more on how to reduce the size of your repository.

Overview

Last updated: 2015-08-03
Language: Python
Access level: Admin (revoke)

1 Branch
8 Tags
2 Forks
7 Watchers

Invite users to this repo
Send Invitation

Recent activity

- 1 commit
Pushed to biobakery/metaphlan2
2ad8ff3 Bug fix. Bug resulting in execution...
Nicola Segata · 2015-08-03
- 3 commits
Pushed to biobakery/metaphlan2
183ae30 merge readme
dcd8820 add samout and heatmap to run...
bdc26e4 add samout and heatmap to run...
Duy Tin Truong · 2015-08-02
- 100% unclassified with Metaphlan2
Issue #8 commented on in biobakery/metaphlan2
Brittany Williams · 2015-07-28
- 1 commit
Pushed to biobakery/metaphlan2
2f454c1 README.md edited online with Bi...
Duy Tin Truong · 2015-07-28
- 1 commit
Pushed to biobakery/metaphlan2

MetaPhlAn 2.0: Metagenomic Phylogenetic Analysis

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

Description

MetaPhlAn is a computational tool for profiling the composition of microbial communities (Bacteria, Archaea, Eukaryotes and Viruses) from metagenomic shotgun sequencing data with species level resolution. From version 2.0 MetaPhlAn is also able to identify specific strains (in the not-so-frequent cases in which the sample contains a previously sequenced strains) and to track strains across samples for all species.

MetaPhlAn 2.0 relies on ~1M unique clade-specific marker genes (the marker information file can be found at src/utills/markers_info.txt.bz2 or here) identified from ~17,000 reference genomes



Who's there: MetaPhlAn2

- The complete MetaPhlAn2 install is in `/class/stamps-software/biobakery/metaphlan2/`

The screenshot shows the Bitbucket source page for the MetaPhlAn2 repository. The page has a dark blue header with the Bitbucket logo, navigation links (Features, Pricing), and user options (owner/repository, search, help, language, sign up, log in). The main content area is titled 'Source' and shows the repository structure. It includes a sidebar with navigation icons and a main panel 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, there is a section for 'MetaPhlAn 2.0: Metagenomic Phylogenetic Analysis' with a list of links for Description, Pre-requisites, Installation, Basic Usage, Full command-line options, Utility Scripts, Heatmap Visualization, and GraPhlAn Visualization.

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

Source

default MetaPhlAn2 /

- db_v20
- utils
- .hgtags 205 B 4 hours ago tagging version 2.0_beta3
- README.md 24.6 KB 2 hours ago README.md edited online with Bitbucket
- metaphlan2.py 35.7 KB 6 hours ago Making MetaPhlAn exiting gracefully when the input format cannot be guessed because two files are

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

>> MetaPhlAn 2.0: Metagenomic Phylogenetic Analysis



From the command line...

- You can create your own virtual copy by running:

```
ln -s /home/ubuntu/metagenomics/metaphlan2/
```

- To see what you can do, run:

```
./metaphlan2/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/metaphlan2.py \  
  --mpa_pk1 ./metaphlan2/db_v20/mpa_v20_m200.pk1 \  
  --bowtie2db ./metaphlan2/db_v20/mpa_v20_m200 \  
  ./763577454-SRS014459-Stool.fasta \  
  --input_type 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: MetaPhlAn2

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

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



Who's there: MetaPhlAn2

```
less -S 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/metaphlan2.py \  
  --mpa_pkl ./metaphlan2/db_v20/mpa_v20_m200.pkl \  
  --bowtie2db ./metaphlan2/db_v20/mpa_v20_m200 \  
  ./763577454-SRS014464-Anterior_nares.fasta \  
  --input_type 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 /home/ubuntu/metagenomics/results/metaphlan/*.txt .
```



Who's there: MetaPhlAn2

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

```
mkdir tmp  
mv *.bowtie2out.txt tmp  
./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>

TM4
MICROARRAY SOFTWARE SUITE

The MeV forums have moved!
You can now find them at [MeV's Sourceforge page](#).

MeV: MultiExperiment Viewer
Normalized and filtered expression files can be analyzed using TIGR Multiexperiment Viewer (MeV). MeV is a versatile microarray data analysis tool, incorporating sophisticated algorithms for clustering, visualization, classification, statistical analysis and biological theme discovery. MeV can handle several input file formats. These include the ".mev" and ".tav" files generated by TIGR Spotfinder and TIGR MIDAS, and also Affymetrix® ("*.txt") and Genepix® ("*.gpr") files. We have assembled a [guide for normalizing Affymetrix® .CEL files](#), which covers RMAExpress, BioConductor, and AMP.

MeV generates informative and interrelated displays of expression

Latest Version:
v4.9
Released April 4, 2013
[Download MeV v4.9](#)

BNPredict plugin
The **BnPredict** module is a plugin for Cytoscape that is designed to work with MeV's BN module.

MeV Survey
A new **MeV Survey** for 2008 is now available.

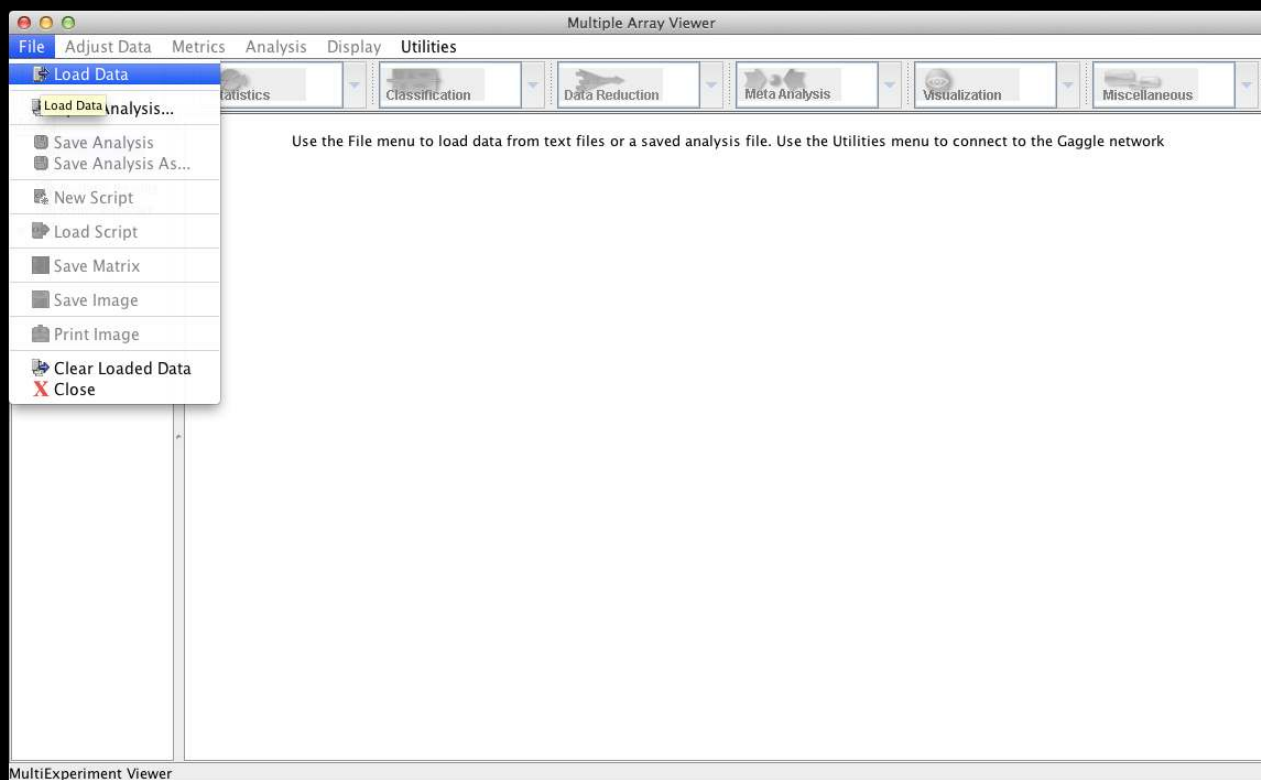
- Or use the appropriate local copy for your machine:

```
scp /home/ubuntu/metagenomics/ext/MeV_4_9_0_r2731_win.zip .  
scp /home/ubuntu/metagenomics/ext/MeV_4_8_1_r2727_mac.tgz .  
scp /home/ubuntu/metagenomics/ext/MeV_4_8_1_r2727_linux.tar.gz .
```



An interlude: MeV

- Don't forget to transfer your `763577454.tsv` file locally for viewing using `scp`
- Unzip, launch MeV, and select File/Load data





An interlude: MeV

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

Expression File Loader

Select File Loader Help

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

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

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

☒ Two-color Array ☐ Single-color Array

Load Annotation Data

☒ Automatically download ☐ Load from local file

Choose an organism: [dropdown] No file selected [Choose File]

☐ Load Annotation

Expression data table:

	76357745...	76357745...	76357745...	76357745...	76357745...
Bacteria	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	

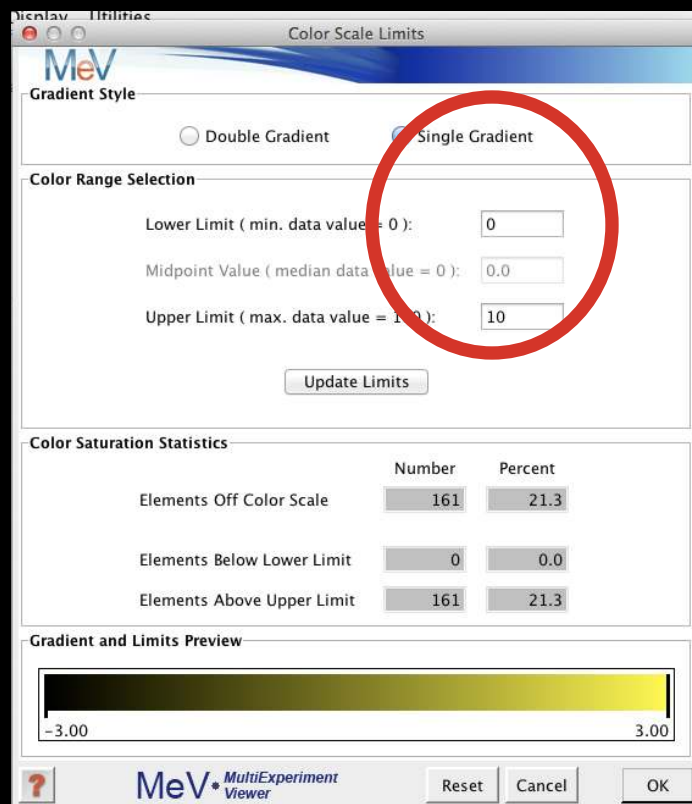
Click the upper-leftmost expression value. Click the Load button to finish.

MeV MultiExperiment Viewer [Cancel] [Load]



An interlude: MeV

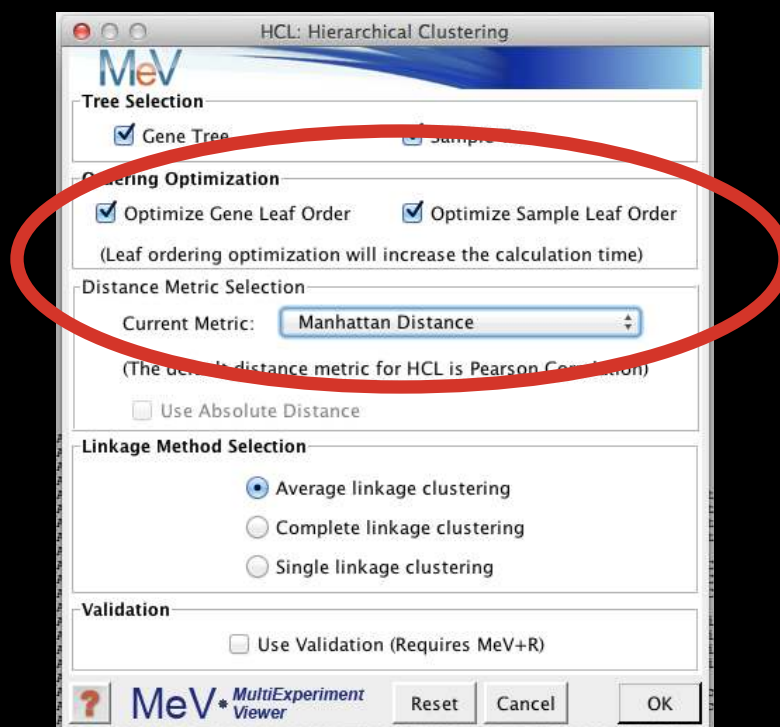
- “Load” your data, then make it visible by:
 - Display/Set Color Scale Limits
 - Choose Single Gradient, min 0, max 10





An interlude: MeV

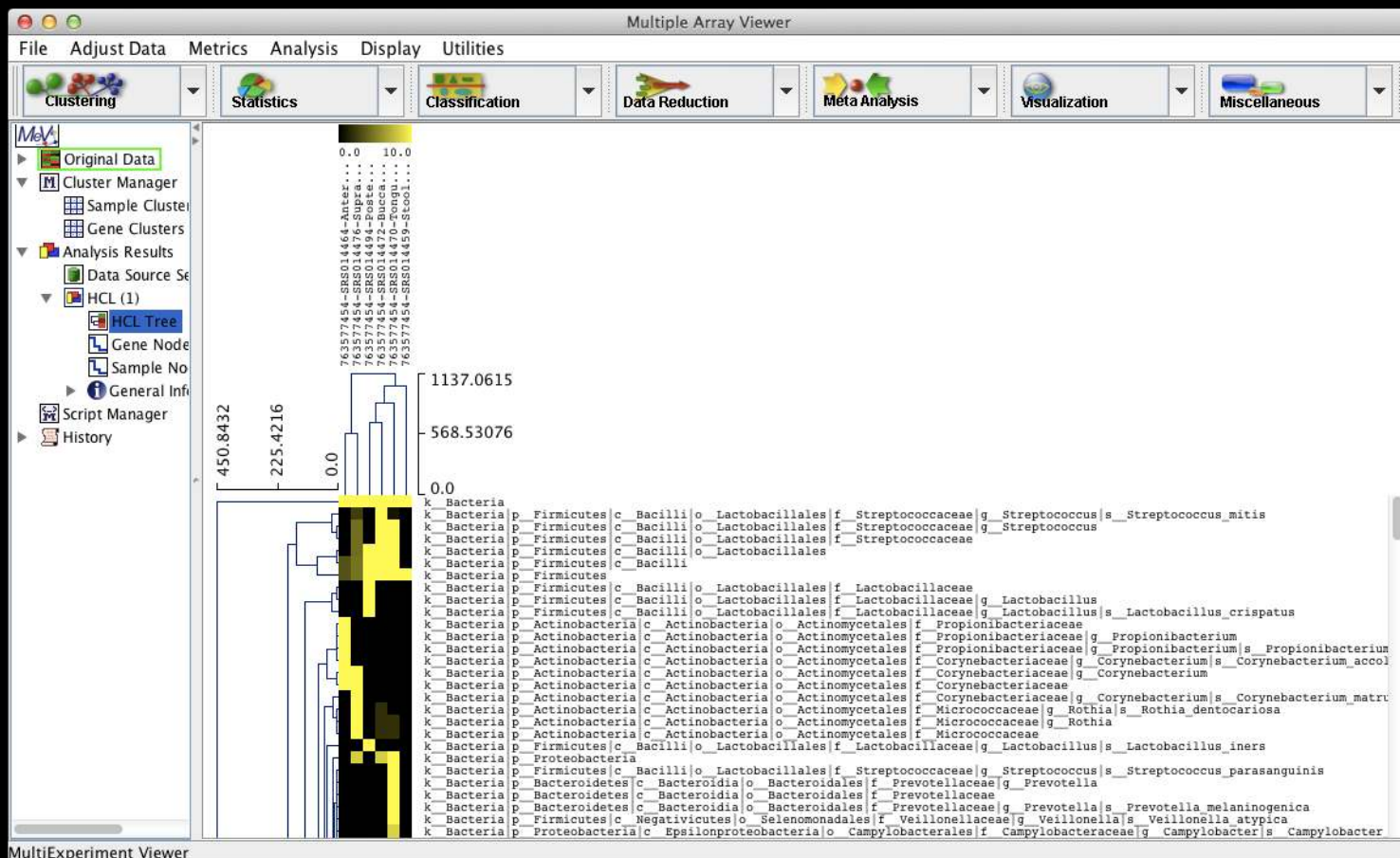
- Finally, to play around a bit:
 - Display/Set Element Size/whatever you'd like
 - Clustering/Hierarchical Clustering
 - Optimize both gene and sample order
 - And select Manhattan Distance (imperfect!)





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!



Summary

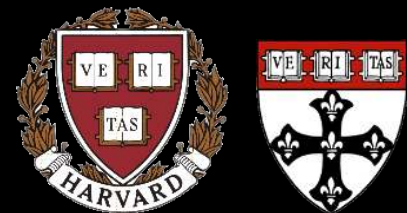
- MetaPhlAn2
 - Evolution of MetaPhlAn1
 - Viruses, euks, subspecies, speed
 - And a LOT more reference data!
 - Raw metagenomic reads in
 - Tab-delimited species relative abundances out



Meta'omic functional profiling with ShortBRED

Galeb Abu-Ali
Eric Franzosa
Curtis Huttenhower

09-18-15



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





The two big questions...

Who is there?
(taxonomic profiling)

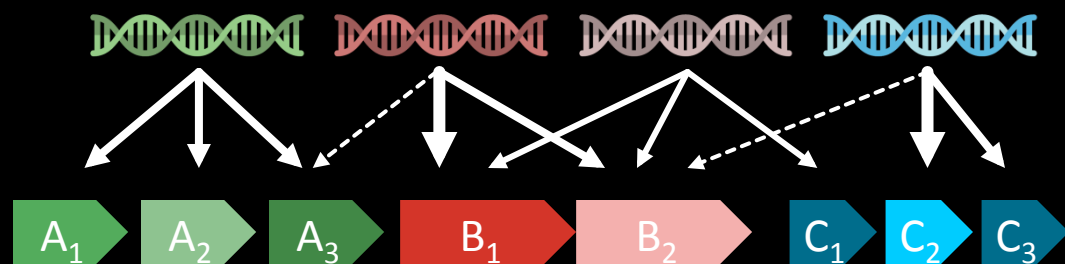
What are they doing?
(functional profiling)

[illegible]



HMPAnN

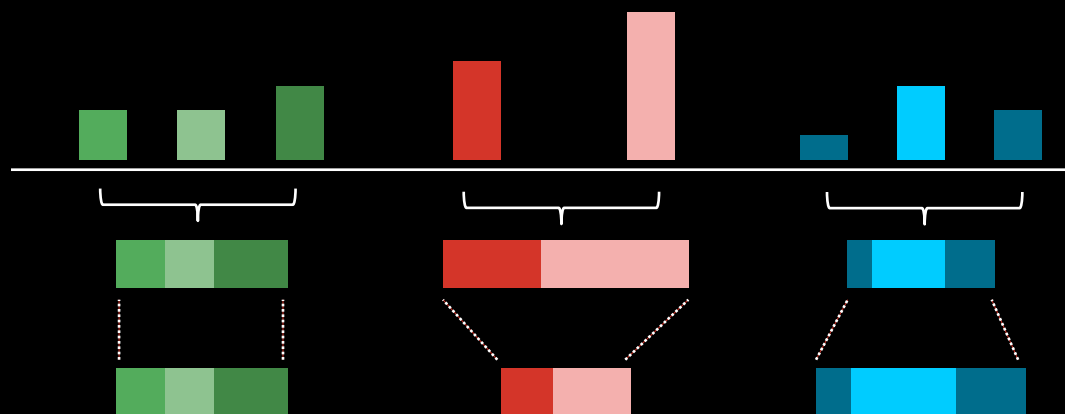
HMP Unified Metabolic Analysis Network



Short reads + protein families

Nucleotide pan-genome search

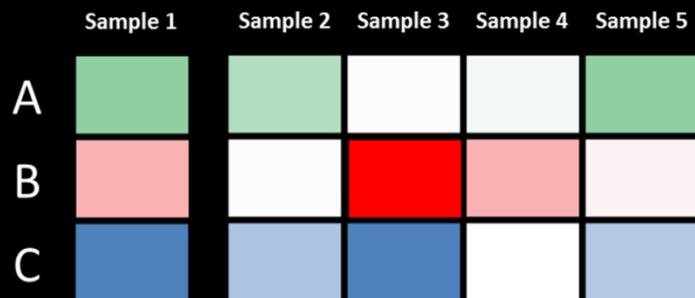
Translated BLAST search



Weight hits by %ID

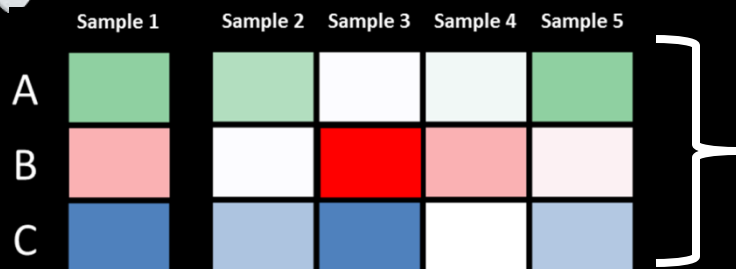
Sum over seqs. within family

Adjust for sequence length

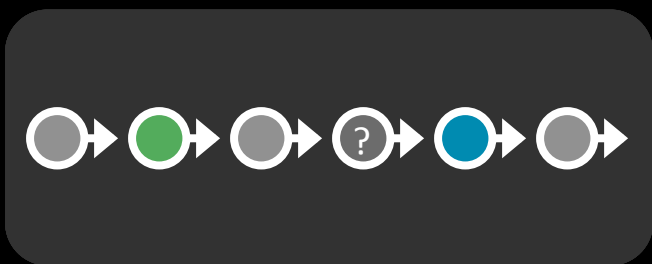


Repeat for each metagenomic or metatranscriptomic sample

HUMAN HMP Unified Metabolic Analysis Network



Many millions of hits are collapsed into a few million gene families (UniRefs) *(still a large number)*



- Map genes to MetaCyc pathways
- Use MinPath (Ye 2009) to find simplest pathway explanation for observed genes
- Remove pathways unlikely to be present due to low organismal abundance
- Smooth/fill gaps



Collapsing UniRef abundance into MetaCyc pathway abundance (or presence/absence) yields a smaller, more tractable feature set



What's there: ShortBRED



Jim
Kaminski

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

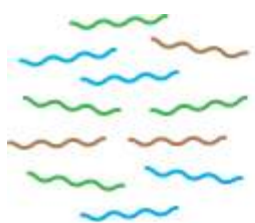


What's there: ShortBRED algorithm

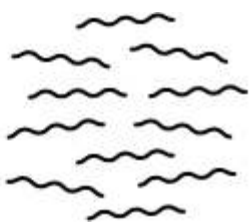
- Cluster proteins of interest into families
 - Record consensus sequences
- Identify any common areas among proteins
 - Compared against each other
 - Compared against reference database
 - Remove all of these
- Remaining subseqs. uniquely ID a family
 - Record these as markers for that family



What's there: ShortBRED marker identification



**Prots of
interest**



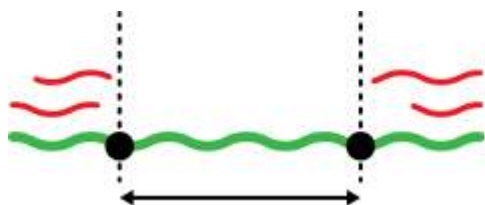
**Reference
database**



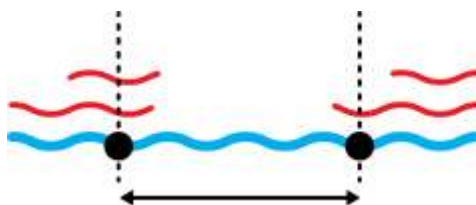
**Cluster into
families**



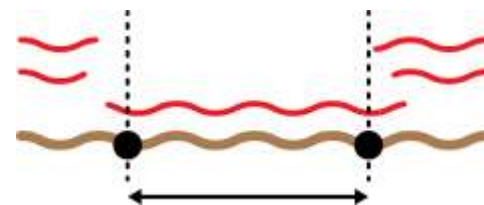
**Identify short,
common regions**



True Marker



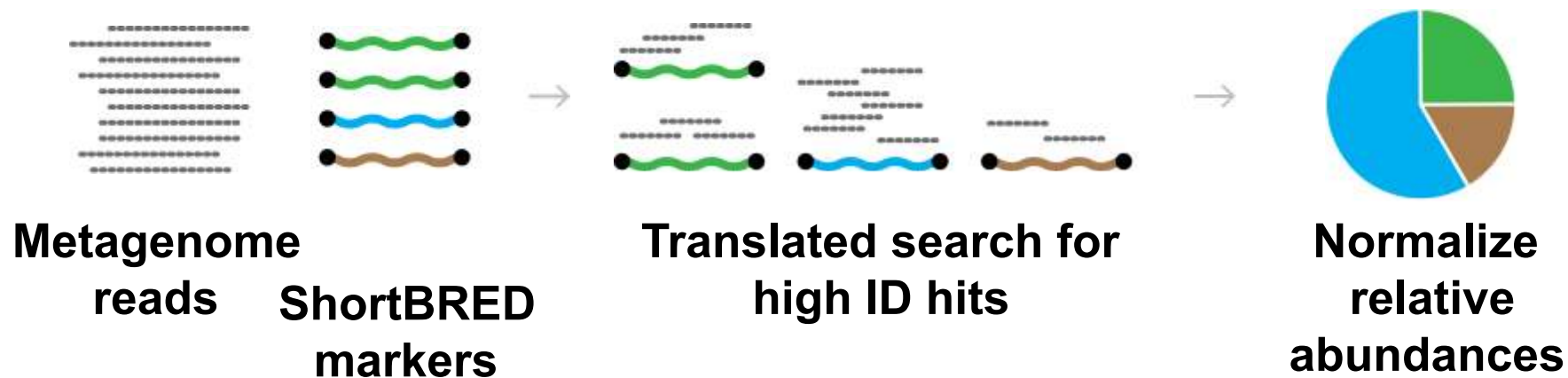
Junction Marker



Quasi Marker



What's there: ShortBRED family quantification





Setup notes reminder

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



What's there: ShortBRED

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

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

Contact Documentation People Presentations Publications Research Teaching

Home

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

ShortBRED

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

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

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

Download ShortBRED (preliminary version)

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

Download ShortBRED here

You may also install ShortBRED using Mercurial:

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

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



From the command line...

- But don't!
 - Instead, we've installed ShortBRED already for you
- You can create your own virtual copy by running:

```
ln -s /home/ubuntu/metagenomics/shortbred/
```

- To see what you can do, run:

```
./shortbred/shortbred_identify.py -h | less -S  
./shortbred/shortbred_quantify.py -h | less -S
```



Getting some annotated protein sequences

You could download the ARDB protein sequences [here](#)

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

ARDB - Antibiotic Resistance Genes Database

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

[Database](#) [Help](#) [Tutorial for ARDB](#)

Antibiotic Resistance

Brief introduction to antibiotic resistance.

Analysis & Tools

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

GO Annotation

How to use GO terms to annotate resistance genes?

Welcome to Antibiotic Resistance Genes Database Home Page

Our motivations in creating ARDB are to:

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

[More...](#)

News

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

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

Database Statistics

Version: 1.1
Last Update: July 3, 2009

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



From the command line...

- But don't!
 - Instead, we've downloaded the important file for you
- Take a look by running:

```
less /home/ubuntu/metagenomics/data/resisGenes.pfasta
```

```
1. screen (less)
>ZP_02959935 hypothetical protein PROSTU_01837 [Providencia stuartii ATCC 25827]
MGIEYRSLHTSQLTLSEKEALYDLLIEGFEGDFSHDDFAHTLGGMHVMAFDQKLVGHVA
IIQRHMLDNTPIISVGYEAMVVEQSYRRQGIGRQLMLQTNKIASCYQLGLLSASDDGQ
KLYHSVGWQIWKGKLFELKQGSYIRSIEEEGGVMGWKADGEVDFTASLYCDFRGGDQW
>Q52424 RecName: Full=Aminoglycoside 2'-N-acetyltransferase; AltName: Full=AAC(2
MGIEYRSLHTSQLTLSEKEALYDLLIEGFEGDFSHDDFAHTLGGMHVMAFDQKLVGHVA
IIQRHMLDNTPIISVGYEAMVVEQSYRRQGIGRQLMLQTNKIASCYQLGLLSASDDGQ
KLYHSVGWQIWKGKLFELKQGSYIRSIEEEGGVMGWKADGEVDFTASLYCDFRGGDQW
>AAA03550 aminoglycoside 2'-N-acetyltransferase [Providencia stuartii].
MGIEYRSLHTSQLTLSEKEALYDLLIEGFEGDFSHDDFAHTLGGMHVMAFDQKLVGHVA
IIQRHMLDNTPIISVGYEAMVVEQSYRRQGIGRQLMLQTNKIASCYQLGLLSASDDGQ
KLYHSVGWQIWKGKLFELKQGSYIRSIEEEGGVMGWKADGEVDFTASLYCDFRGGDQW
>Q49157 RecName: Full=Aminoglycoside 2'-N-acetyltransferase; AltName: Full=AAC(2
MPFQDVSAPVRGGILHTARLVHTSDLDQETREGARRMVIEAFEGDFSADWEHALGGMHA
FICHHGALIAHAADVQRRLLYRDTALRCGYVEAVAVREDWRGQGLATAVMDAVEQVLRGA
YQLGALSASDTARGMYLSRGWLPWQGPTSVLQPAQVTRTPEDDEGLFVLPVGLPAGMELD
TTAEITCDWRDGDVW
>NP_214776 aminoglycoside 2'-N-acetyltransferase AAC (AAC(2')-IC) [Mycobacterium
MHTQVHTARLVHTADLDSETRQDIRQMVTGAFAGDFTETDWEHTLGGMHALIWHHGATIA
HAAVIQRRLLIYRGNALRCGYVEGAVRADWRGQRLVSALLDAVEQVMRGAYQLGALSSA
RARRLYASRGWLPWHGPTSVLAPTGPVTRTPDDDGTVFVLPIDISLDTSAELMCDWRAGDV
W
>NP_334681 aminoglycoside 2-N-acetyltransferase [Mycobacterium tuberculosis CDC1
MHTQVHTARLVHTADLDSETRQDIRQMVTGAFAGDFTETDWEHTLGGMHALIWHHGATIA
ttenh/Dropbox/shared/ShortBRED/data/ARDB/ardbAnno1.0/blastdb/resisGenes.pfasta
```




Getting some reference protein sequences

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

Home About **Download** Help

Metaref

Keyword Search [Help](#)

Microbial taxonomy

You could download the MetaRef protein sequences [here](#)

Browse

Bacteria: [2706](#) Genomes

Archaea: [112](#) Genomes

Taxonomy Correction [Info](#)

Highlighted Clades

(Commonly Found in Human Microbiome)

Airways Nares

[Corynebacterium accolens](#)

[Propionibacterium acnes](#)

[Staphylo. epidermidis](#)

Buccal Mucosa

[Gemella haemolysans](#)

[Haemophilus influenzae](#)

[Streptococcus mitis](#)

MetaRef Database v 1.0

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

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

Core families: genes present consistently within a clade

Marker families: genes present consistently and exclusively within a clade



Running ShortBRED-Identify

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

```
./shortbred/shortbred_identify.py \  
  --goi /home/ubuntu/metagenomics/data/resisGenes.pfasta \  
  --ref ./shortbred/example/ref_prots.faa \  
  --markers ardb_markers.faa  
less ardb_markers.faa
```

- This should take ~5 minutes
 - If you get bored waiting, kill it and copy:
`/home/ubuntu/metagenomics/results/shortbred/ardb_markers.faa`
- It will produce lots of status output as it runs



ShortBRED markers

```
CFAR2015 — ubuntu@ip-10-113-166-56: ~/metagenomics/hutlabTest — ssh — 100x26
gabuali@hutlab...rray/slurm_logs3  ubuntu@ip-10-...omics/hutlabTest  ubuntu@ip-10-1...mics/hutlabTest  +
>ZP_01723236_TM_#01
TEEFLGKYP
>ZP_01723236_TM_#02
IVVMWKRMLSLVGLYKIDGQSQSINRRFNLLHVIVGM
>ZP_01723236_TM_#03
FAFKDFIDDHLFKVEHVYYA
>ZP_01723236_TM_#04
KPKVDSLDKISYGLAF
>ZP_01723236_TM_#05
LVSVLKNWDTLSMDYFGFYAVGFISSEFI
>ZP_01723236_TM_#06
ALISKVKLM
>AAA25717_TM_#01
MHLTITYWIDRLREAYPHAVAILLKGSYARGEASAWSDIDFDVLVSDEEVEEYRTWIEPV
GERLVHISVAVEVWTGWERDSADPSSWSYGLPTQETTQLLWAADENIRRRRLDRPFKVHPA
AEPEVEDTVEALGKIRNAMVRGDDLAVYQAAQVVGKLIPTLLVPINPPTYARFAREAIDR
ILAFPNVPEGFAADWLTCMGLVDRRTHDPQPTRPNEWCAARSRFCRRMRTSSVRISRGCW
KQDWYLRISART
>CAD61201_TM_#01
MFQIRSFLVGISAFVMAVLGSAAYSAQPGGEYPTVDDIPVGEVRLYK
>CAD61201_TM_#02
LTRQLAEAAGNEVPAHSLKA
>CAD61201_TM_#03
AVRVLFGGCAVHEASRE
>CAD61201_TM_#04
ardb_markers.faa
```

True Markers
at the top



ShortBRED markers

```
CFAR2015 — ubuntu@ip-10-113-166-56: ~/metagenomics/hutlabTest — ssh — 100x26
gabuali@hutlab...rray/slurm_logs3  ubuntu@ip-10-...omics/hutlabTest  ubuntu@ip-10-1...mics/hutlabTest  +

MNDIDREEPAAAA
>P14509_TM_#02
PESMAAHVMGYKWARDKVQSGCAVYRLHSKSGGSDLFLKHGKDAF
>P14509_TM_#03
GHISVPSVVSFVRTPNQAWLLTTAIGHKTAYQVLKSDFGARLVVDALAAFMRRLHAIPV
SECS
>P14509_TM_#04
IEAGVVDVDDFDKEREGWTAEQVWEAMHRLPLA
>P14509_TM_#05
LIVEGKVVCIDVGRAGIADRYQDLAVLWNCLEEFEPQLQERLVAQYGIADPDRR
>1112175A_JM_#01__ [1112175A_w=0.486, YP_001103000_w=0.143, YP_001103000_w=0.371]
LFEWFEVKVDSAIMRLRRRAEPLLEGAALERYE
>1112175A_JM_#02__ [1112175A_w=0.515, YP_001103000_w=0.333, YP_001103000_w=0.152]
RKYPRRRVEAAFDHAGVGGGAVVAYVRPEQWLRL
>ABF69686_JM_#01__ [ABF69686_w=0.459, ABN80187_w=0.135, ZP_03989103_w=0.405]
DTAYPGEIVILADDTLKLNDILGNEKLLPHKTRI
>YP_002081505_JM_#01__ [YP_002081505_w=0.630, YP_274481_w=0.370]
LGTIGGFRLQIEDRGNX
>YP_274481_QM33_#01__ [YP_274481_w=0.500, YP_002081505_w=0.500]
PAAFISGLTGQFYKQFALTIAISTVISAFNSLT
>YP_970399_JM_#01__ [YP_970399_w=0.306, ZP_03552050_w=0.163, YP_997055_w=0.163, YP_997055_w=0.102, CAJ939
47_w=0.061, YP_001348697_w=0.061, YP_316450_w=0.041, YP_002092118_w=0.061, Q2KX31_w=0.041]
GGMLLGLSRKAATDX
>ZP_01817983_JM_#01__ [ZP_01817983_w=0.493, YP_001694417_w=0.362, YP_001694417_w=0.145]
TLTGPFIFIGFIKEDFQPVAKEKAIPTKELFTSVK
(END)
```

Junction/Quasi Markers
at the bottom



Running ShortBRED-Quantify

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

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

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



ShortBRED marker quantification

CFAR2015 — ubuntu@ip-10-113-166-56: ~/metagenomics/hutlabTest — ssh — 100x27

gabuiali@hutlab...rray/slurm_logs3 ubuntu@ip-10-...omics/hutlabTest ubuntu@ip-10-1...mics/hutlabTest +

Family	Count	Hit	Total	Marker	Length
YP_001694417	2380.9523809523807	1	26		
ZP_04679156	0.0	0	136		
ZP_04657259	0.0	0	91		
ZP_04635798	0.0	0	171		
ZP_04635523	0.0	0	59		
ZP_04633951	0.0	0	9		
ZP_04616832	0.0	0	72		
ZP_04613685	0.0	0	183		
ZP_04606269	0.0	0	168		
ZP_04577926	0.0	0	173		
ZP_04543635	0.0	0	186		
ZP_04543532	0.0	0	187		
ZP_04433866	0.0	0	95		
ZP_04431003	0.0	0	169		
ZP_04405580	0.0	0	300		
ZP_04405450	0.0	0	138		
ZP_04309403	0.0	0	177		
ZP_04284182	0.0	0	51		
ZP_04244950	0.0	0	113		
ZP_04210257	0.0	0	129		
ZP_04197552	0.0	0	70		
ZP_04175489	0.0	0	21		
ZP_04174269	0.0	0	27		
ZP_04151022	0.0	0	86		
ZP_04107441	0.0	0			

RPKMs and raw hit count

Other columns are family name and total AAs among all family makers

Sort table

```
(head -n 1; sort -k 2,2 -n -r) < \
763577454-SRS014472-Buccal_mucosa-ARDB.txt | less
```




AR proteins in the human gut

- Example of some real data

`/home/ubuntu/metagenomics/data/shortbred_ardb_hmp_t2d.tsv`

- This is the result of running:
 - ShortBRED-Identify on the real ARDB + reference
 - ShortBRED-Quantify on the real HMP + T2D data (Qin Nature 2012)
 - Summing each sample's RPKMs for families in each ARDB resistance class



AR proteins in the human gut

shortbred_aradb_hmp_t2d.tsv

Search in Sheet

Home Layout Tables Charts SmartArt Formulas Data Review

Edit Font Alignment Number Format Cells Themes

Calibri (Body) 12

Normal

Insert Delete Format Themes

fx Sample.ID

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Sample.ID	HMP1	HMP2	HMP3	HMP4	HMP5	HMP6	HMP7	HMP8	HMP9	HMP10	HMP11	HMP12	HMP13	HMP14	HMP15	HMP16	HMP17	HMP18
2	Dataset	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP	HMP
3	Gender	Female	Male	Female	Male	Female	Female	Male	Male	Female	Female	Female	Male	Male	Male	Female	Male	Male	Male
4	ABR Class	SR5011061	SR5011134	SR5011239	SR5011271	SR5011302	SR5011405	SR5011452	SR5011529	SR5011586	SR5012273	SR5012902	SR5013158	SR5013215	SR5013476	SR5013521	SR5013687	SR5013800	SR5013951
5	ABC Antibiot	0	0.6097114	0.53837173	0	0	0.05083452	0	0	18.879238	0.3999418	0.6375002	0.11029351	0	0	0.1499069	3.3238466	0	0
6	Aminoglycos	0	0	0	0.5570841	0	0	0	0	0	0.4844142	0	0	0	7.15621993	0	0	0	0.06597383
7	Aminoglycos	11.8847826	2.3493412	1.31127279	2.1879248	1.70197254	25.2342538	0	1.4888313	6.7524558	11.6664297	0.2944691	0	0.54364476	22.1364669	1.0549423	6.1159491	2.1534126	2.95684284
8	Aminoglycos	0.72342527	9.510191	0.43478001	9.31863091	1.44994258	21.7649766	0	0	1.8219867	1.9941331	0.7220629	1.82419711	0	1.09356043	1.6969943	5.382002	1.6022915	0.98286613
9	Antibiotic Ta	0	0.4319648	0	0	0.11002037	0	0	0	0.1044046	0	0.6096981	4.45863298	0	0	0.1242086	0	0	0
10	Chloramphen	0	0.8931758	0.50566409	0.06863132	0	0	0	0	0.2300411	0.2286945	0	0	0	0	0	0	0.3360012	0
11	Chloramphen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	Chloramphen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	Class A Beta-	11.9616538	14.1741569	192.732027	57.3421171	30.3784485	36.4756423	41.445191	77.8068337	27.5978829	84.7152993	29.5138602	4.47890136	7.54656865	6.17723545	67.6346059	121.5429	40.9881448	18.254292
14	Class B Beta-	0.73757867	0.4730655	0	0.35938332	0.22651252	0.45452038	0	0.1196987	1.5652141	0.5770399	0	0	0	0	0	0	0	0
15	Class C Beta-	0	0	0	0	0	0	0	0	0	0.4758603	0.2556631	0	0	0	0	0.1458178	0	0
16	Class D Beta-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	Gene Modul	0	0	0.12940327	0	0	0	0	0	0	2.6860575	0.3513343	0.52138395	0.18121492	0.09719297	0	0.6224941	0	0
18	Gene Modul	0	0	0.53609928	0.10341706	0.28813026	0	0	0.1033344	0	0.4529638	0	0.59939377	0	0.73268549	0	0	0	0.15287079
19	Glycopeptide	0	0.1148873	0.10721986	2.91192901	11.8252927	1.06129011	0	1.475885	0	3.8329823	0.2028631	0.17855513	0	2.57636295	0	12.8763448	0	1.37583708
20	Lincosamide	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	Macrolide Re	0	0	0	0	0	0	0	0	0	0	0.2216556	0	0	0	0	0	0	0
22	MATE Antibio	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	MFS Antibiot	0	0.1079916	2.44436309	2.24124166	0.15717195	19.6482667	0	0	0	6.0081483	4.73637	0.16432993	0	9.88061341	0.2382082	43.436675	1.4549685	0
24	Other ARG	0	0.1641248	1.50507872	4.90492355	0.80462657	0.27160156	0	0.4618416	1.2797248	2.911427	1.0099704	0.79420864	0	0.21818147	0.3167416	0.7025792	0	4.57893981
25	Puromycin R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	Quinolone R	0	0	0.05601037	0.09933481	0.05066727	0.05083452	0	0	0	0.8647162	0.1335553	3.29844229	0.06626516	0.6266389	0	0.1841579	0.1746919	0
27	Rifamycin Re	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	RND Antibiot	1.11005589	0.2116346	0.87820136	0.51112275	1.80007009	12.407319	34.237278	3.5262745	38.781576	4.5900824	1.9670192	0.17668244	38.004141	1.38795841	0.7786209	2.9700758	1.1984926	6.61769588
29	rRNA Methyl	5.61799582	6.0194576	37.2369165	9.44289101	34.6172522	94.7288439	2.051664	80.7900949	122.947846	2.4135554	10.2418695	0.06217665	7.23364421	13.9417838	130.737494	96.9503344	18.8879339	5.07069194
30	SMR Antibiot	0	0	0	0	0	0	0	0	0	0.876332	0	0.08288129	0	0.19222828	0	0.2560272	0	0
31	Streptogram	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	Tetracycline	0.06843748	2.6183624	0.57325559	0.86505449	12.8908188	0.16675423	2.793598	0.359161	0.5939219	2.0434753	2.4886453	0.33754257	0.23247387	0	0.9097696	2.3449461	0	5.81292995

shortbred_aradb_hmp_t2d.tsv

Normal View Ready

Sum=0



Summary

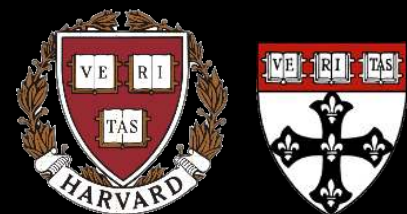
- HUMAnN2 (up next!)
 - Quality-controlled metagenomic reads in
 - Tab-delimited gene, module, and pathway relative abundances out
- ShortBRED
 - Raw metagenomic reads, Proteins of interest, and Protein reference database in
 - Tab-delimited gene family rel. abundances out



Meta'omic functional profiling with HUMAnN2

Galeb Abu-Ali
Eric Franzosa
Curtis Huttenhower

09-18-15



Harvard School of Public Health
Department of Biostatistics





The two big questions...

Who is there?
(taxonomic profiling)

What are they doing?
(functional profiling)



Setup notes reminder

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



What they're doing: HUMAnN

- As a broad functional profiler, you *could* download HUMAnN at: <http://huttenhower.sph.harvard.edu/humann>

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

Contact Documentation People Presentations Publications Research Teaching

Home

HUMAnN: The HMP Unified Metabolic Analysis Network

You can obtain the HUMAnN software here:

[humann-0.98.tar.gz](#)

This is the latest version, which provided the analysis for the shotgun data from the Human Microbiome Project. If you find the software or data useful, please cite our manuscript:

Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, Rodriguez-Mueller B, Zucker J, Thiagarajan M, Henrissat B, White O, Kelley ST, Methé B, Schloss PD, Gevers D, Mitreva M, Huttenhower C. "Metabolic reconstruction for metagenomic data and its application to the human microbiome." *PLoS Comput Biol.* 2012 Jun;8(6):e1002358

Please **contact us** if you have any comments, suggestions, or bug reports for the software. Code is also available directly from our **Mercurial** source code repository at <http://bitbucket.org/chuttenh/humann> using the `hg clone` command.

HUMAnN is a pipeline for efficiently and accurately determining the presence/absence and abundance of microbial pathways in a community from metagenomic data. Sequencing a metagenome typically produces millions of short DNA/RNA reads. HUMAnN takes these reads as inputs and produces gene and pathway summaries as outputs:

- The abundance of each orthologous gene family in the community. Orthologous families are groups of genes that perform roughly the same biological roles. HUMAnN uses the KEGG Orthology (KO) by default, but any catalog of orthologs can be employed with minor changes (COG, NOG, etc.)
- The presence/absence of each pathway in the community. HUMAnN refers to pathway presence/absence as "coverage," and defines a pathway as a set of two or more genes. HUMAnN uses KEGG pathways and modules by default, but again can easily be modified to use GO terms or other gene sets.
- The abundance of each pathway in the community. I.e. how many "copies" of that pathway are present



What they're doing: HUMAnN2

- Or even *better*, the latest version is HUMAnN2 at: <http://huttenhower.sph.harvard.edu/humann2>

HUMAnN2: The HMP Unified Metabolic Analysis Network 2

HUMAnN2 is the next generation of HUMAnN. HUMAnN is a pipeline for efficiently and accurately profiling the presence/absence and abundance of microbial pathways in a community from metagenomic or metatranscriptomic sequencing data (typically millions of short DNA/RNA reads). This process, referred to as functional profiling, aims to describe the metabolic potential of a microbial community and its members. More generally, functional profiling answers the question "What are the microbes in my community-of-interest doing (or capable of doing)?"

If you use the HUMAnN2 software, please cite our manuscript: TBD

For additional information, please see the [HUMAnN2 User Manual](#).

Contents

- Features
- Workflow
- Requirements
- **Installation**
- How to run
 - Basic usage
 - Demo runs
- Output files
 - Gene families
 - Pathway coverage
 - Pathway abundance

Click
here



What they're doing: HUMAnN2

- ...but instead we've already installed it!
- Normally you'd follow the online tutorial to expand:

```
tar -xzf humann2_v0.2.3.tar.gz
```

- Install:

```
cd humann2_v0.2.3  
python setup.py minpath  
python setup.py install
```

- And download DIAMOND from here:
 - <http://ab.inf.uni-tuebingen.de/software/diamond/>
- We're going to use it preinstalled instead



What they're doing: HUMAnN2

- If we weren't all running this, you'd need to:
 - Get our precomputed DNA/AA databases
 - ChocoPhlAn ~50M genes from NCBI
 - UniRef ~100M proteins from UniProt

```
humann2_databases --download chocophlan full \  
    /class/stamps-software/biobakery/humann2/  
humann2_databases --download uniref diamond \  
    /class/stamps-software/biobakery/humann2/
```

- This would take too long for everyone to use, so we'll stick with the demo database instead...



What they're doing: HUMAnN2

- Take a look at the demo input metagenome:

```
less -S /home/ubuntu/metagenomics/data/humann2/examples/demo.fastq
```

- From your home directory, run HUMAnN2:

```
humann2 \  
  --input /home/ubuntu/metagenomics/data/humann2/examples/  
demo.fastq \  
  --output humann2_demo
```

- What did you just do?

```
less -S humann2_demo/demo_genefamilies.tsv
```

- UniRef gene family IDs
- With human-readable glosses when available
- Broken down per organism



What they're doing: HUMAnN2

```
CFAR2015 — ubuntu@ip-10-113-166-56: ~/metagenomics/hutlabTest — ssh — 100x32
gabuali@hutlab...rray/slurm_logs3  ubuntu@ip-10-...omics/hutlabTest  ubuntu@ip-10-113-166-56: ~  +

# Gene Family  demo_Abundance
UniRef50_A6L108 8.7719298246
UniRef50_A6L108|g__Bacteroides.s__Bacteroides_stercoris 8.7719298246
UniRef50_E1WMC2 7.5757575758
UniRef50_E1WMC2|g__Bacteroides.s__Bacteroides_fragilis 7.5757575758
UniRef50_R5FJB9: Conjugative transposon TraN protein 7.3070013559
UniRef50_R5FJB9: Conjugative transposon TraN protein|g__Bacteroides.s__Bacteroides_fragilis 3.05
UniRef50_R5FJB9: Conjugative transposon TraN protein|g__Bacteroides.s__Bacteroides_thetaiotaomicron
UniRef50_R5FJB9: Conjugative transposon TraN protein|g__Bacteroides.s__Bacteroides_stercoris 2.07
UniRef50_B6YQ01: 50S ribosomal protein L11 6.7567567568
UniRef50_B6YQ01: 50S ribosomal protein L11|g__Bacteroides.s__Bacteroides_fragilis 2.2522522523
UniRef50_B6YQ01: 50S ribosomal protein L11|g__Bacteroides.s__Bacteroides_stercoris 2.2522522523
UniRef50_B6YQ01: 50S ribosomal protein L11|g__Bacteroides.s__Bacteroides_thetaiotaomicron 2.25
UniRef50_Q64R13 6.5359477124
UniRef50_Q64R13|g__Bacteroides.s__Bacteroides_fragilis 6.5359477124
UniRef50_F5XD83: Conjugative transposon protein TraK 6.4102564103
UniRef50_F5XD83: Conjugative transposon protein TraK|g__Bacteroides.s__Bacteroides_thetaiotaomicron
UniRef50_F5XD83: Conjugative transposon protein TraK|g__Bacteroides.s__Bacteroides_fragilis 1.60
UniRef50_F5XD83: Conjugative transposon protein TraK|g__Bacteroides.s__Bacteroides_stercoris 1.60
UniRef50_B6YQ88: 30S ribosomal protein S7 6.2893081761
UniRef50_B6YQ88: 30S ribosomal protein S7|g__Bacteroides.s__Bacteroides_stercoris 4.1928721174
UniRef50_B6YQ88: 30S ribosomal protein S7|g__Bacteroides.s__Bacteroides_fragilis 2.0964360587
UniRef50_D1KAI8 6.2893081761
UniRef50_D1KAI8|g__Bacteroides.s__Bacteroides_stercoris 6.2893081761
UniRef50_Q2S3Q3: 50S ribosomal protein L24 6.2599188856
UniRef50_Q2S3Q3: 50S ribosomal protein L24|g__Bacteroides.s__Bacteroides_thetaiotaomicron 3.14
UniRef50_Q2S3Q3: 50S ribosomal protein L24|g__Bacteroides.s__Bacteroides_stercoris 3.1152647975
UniRef50_A0A016LIR2 6.1728395062
UniRef50_A0A016LIR2|g__Bacteroides.s__Bacteroides_stercoris 6.1728395062
UniRef50_Q5LEY5 6.1728395062
UniRef50_Q5LEY5|g__Bacteroides.s__Bacteroides_fragilis 6.1728395062
humann2_demo/demo_genefamilies.tsv
```



What they're doing: HUMAnN2

- This has created three main files:
 - One listing gene family abundances
 - Two listing pathway (default MetaCyc) abundances and coverages
 - Coverage % of “essential” pathway genes present
 - Abundance “Average” abundance of essential pathway genes

- Each is tab-delimited text with two columns

`humann2_demo/demo_genefamilies.tsv`

- Relative abundance (RPKM) of gene families (UniRef)

`humann2_demo/demo_pathabundance.tsv`

- Relative abundance (RPKM) of pathways (MetaCyc)

`humann2_demo/demo_pathacoverage.tsv`

- Coverage (%) of pathways (MetaCyc)

- I almost always just use abundances (gene or pathway)



What they're doing: HUMAnN2

- Pathways look very much like gene families:

```
less -S humann2_demo/demo_pathabundance.tsv
```

```
CFAR2015 — ubuntu@ip-10-113-166-56: ~/metagenomics/hutlabTest — ssh — 100x24
gabuali@hutlab...rray/slurm_logs3  ubuntu@ip-10-...omics/hutlabTest  ubuntu@ip-10-113-166-56: ~  +

# Pathway      demo_Abundance
SUCROSEUTIL2-PWY: sucrose degradation VII (sucrose 3-dehydrogenase)      5.8949329489
SUCROSEUTIL2-PWY: sucrose degradation VII (sucrose 3-dehydrogenase)|g__Bacteroides.s__Bacteroides_th
SUCROSEUTIL2-PWY: sucrose degradation VII (sucrose 3-dehydrogenase)|g__Bacteroides.s__Bacteroides_st
PWY-6627: salinosporamide A biosynthesis      2.8307602366
PWY-6627: salinosporamide A biosynthesis|g__Bacteroides.s__Bacteroides_thetaiotaomicron 0.9701665013
PWY-5209: methyl-coenzyme M oxidation to CO2      2.7862228006
PWY-7555: &alpha;-cyclopiasonate biosynthesis      2.6676145598
PWY-7555: &alpha;-cyclopiasonate biosynthesis|g__Bacteroides.s__Bacteroides_thetaiotaomicron 0.86
PWY-7440: dTDP-&beta;-L-4-epi-vancosamine biosynthesis      2.5755106274
PWY-3841: folate transformations II      2.3673882669
PWY-7301: dTDP-&beta;-L-noviose biosynthesis      2.3051941136
PWY-7043: 11-cis-3-hydroxyretinal biosynthesis      2.3017477039
PWY-7043: 11-cis-3-hydroxyretinal biosynthesis|g__Bacteroides.s__Bacteroides_stercoris 2.0827022653
HSERMETANA-PWY: L-methionine biosynthesis III      2.2932930562
HSERMETANA-PWY: L-methionine biosynthesis III|g__Bacteroides.s__Bacteroides_thetaiotaomicron 0.99
PWY-7104: dTDP-L-megosamine biosynthesis      2.2089022269
PWY-5100: pyruvate fermentation to acetate and lactate II      2.1231927723
PWY-5100: pyruvate fermentation to acetate and lactate II|g__Bacteroides.s__Bacteroides_stercoris
PWY-7432: L-phenylalanine biosynthesis III (cytosolic, plants)      2.0812499269
PWY-7432: L-phenylalanine biosynthesis III (cytosolic, plants)|g__Bacteroides.s__Bacteroides_thetaio
PWY-6973: dTDP-D-olivose, dTDP-D-oliose and dTDP-D-mycarose biosynthesis      1.9963472394
PWY-6973: dTDP-D-olivose, dTDP-D-oliose and dTDP-D-mycarose biosynthesis|g__Bacteroides.s__Bacteroid
humann2_demo/demo_pathabundance.tsv
```




What they're doing: HUMAnN2

- You can always open these in Excel too
 - Note:** this is sparse since we're using small subsets of the reference data (ChocoPhlAn and UniRef) and input metagenome

	A	B	C	D	E	F	G	H	I	J	K	L
1	# Gene Family	demo_Abundance										
2	UniRef50_A6L108	8.77192982										
3	UniRef50_A6L108 g__Bacteroides.s__Bacteroides_stercoris	8.77192982										
4	UniRef50_R5FJB9: Conjugative transposon TraN protein	8.38107701										
5	UniRef50_R5FJB9: Conjugative transposon TraN protein g__Bacteroides.s__Bacteroides_stercoris	6.20569818										
6	UniRef50_R5FJB9: Conjugative transposon TraN protein g__Bacteroides.s__Bacteroides_thetaiotaomicron	2.17537883										
7	UniRef50_B6YQ01: 50S ribosomal protein L11	6.75675676										
8	UniRef50_B6YQ01: 50S ribosomal protein L11 g__Bacteroides.s__Bacteroides_stercoris	4.5045045										
9	UniRef50_B6YQ01: 50S ribosomal protein L11 g__Bacteroides.s__Bacteroides_thetaiotaomicron	2.25225225										
10	UniRef50_B6YQ88: 30S ribosomal protein S7	6.28930818										
11	UniRef50_B6YQ88: 30S ribosomal protein S7 g__Bacteroides.s__Bacteroides_stercoris	4.19287212										
12	UniRef50_B6YQ88: 30S ribosomal protein S7 g__Bacteroides.s__Bacteroides_thetaiotaomicron	2.09643606										
13	UniRef50_D1KAI8	6.28930818										
14	UniRef50_D1KAI8 g__Bacteroides.s__Bacteroides_stercoris	6.28930818										
15	UniRef50_Q2S3Q3: 50S ribosomal protein L24	6.25991889										
16	UniRef50_Q2S3Q3: 50S ribosomal protein L24 g__Bacteroides.s__Bacteroides_thetaiotaomicron	3.14465409										
17	UniRef50_Q2S3Q3: 50S ribosomal protein L24 g__Bacteroides.s__Bacteroides_stercoris	3.1152648										
18	UniRef50_A0A016LUR2	6.17283951										
19	UniRef50_A0A016LUR2 g__Bacteroides.s__Bacteroides_stercoris	6.17283951										
20	UniRef50_B0NTS9	5.84795322										
21	UniRef50_B0NTS9 g__Bacteroides.s__Bacteroides_stercoris	5.84795322										
22	UniRef50_R5IH84	5.64971751										
23	UniRef50_R5IH84 g__Bacteroides.s__Bacteroides_stercoris	5.64971751										
24	UniRef50_R6FNM5	5.46448087										
25	UniRef50_R6FNM5 g__Bacteroides.s__Bacteroides_thetaiotaomicron	5.46448087										
26	UniRef50_B0NQY6	5.29100529										
27	UniRef50_B0NQY6 g__Bacteroides.s__Bacteroides_stercoris	5.29100529										
28	UniRef50_F3PHD0	5.29100529										
29	UniRef50_F3PHD0 g__Bacteroides.s__Bacteroides_stercoris	5.29100529										
30	UniRef50_A6KYI0: 50S ribosomal protein L6	5.26315789										
31	UniRef50_A6KYI0: 50S ribosomal protein L6 g__Bacteroides.s__Bacteroides_thetaiotaomicron	5.26315789										
32	UniRef50_B0NP96	5.20833333										



What they're doing: HUMAnN2

- If you run more than one sample, you can combine them:

```
less -S \
/home/ubuntu/metagenomics/data/humann2/genes/763577454-SRS014459-Stool_genefamilies.tsv
/home/ubuntu/metagenomics/data/humann2/humann2/tools/join_tables.py \
-i /home/ubuntu/metagenomics/data/humann2/genes/ \
-o 763577454_genefamilies.tsv
less -S 763577454_genefamilies.tsv
```

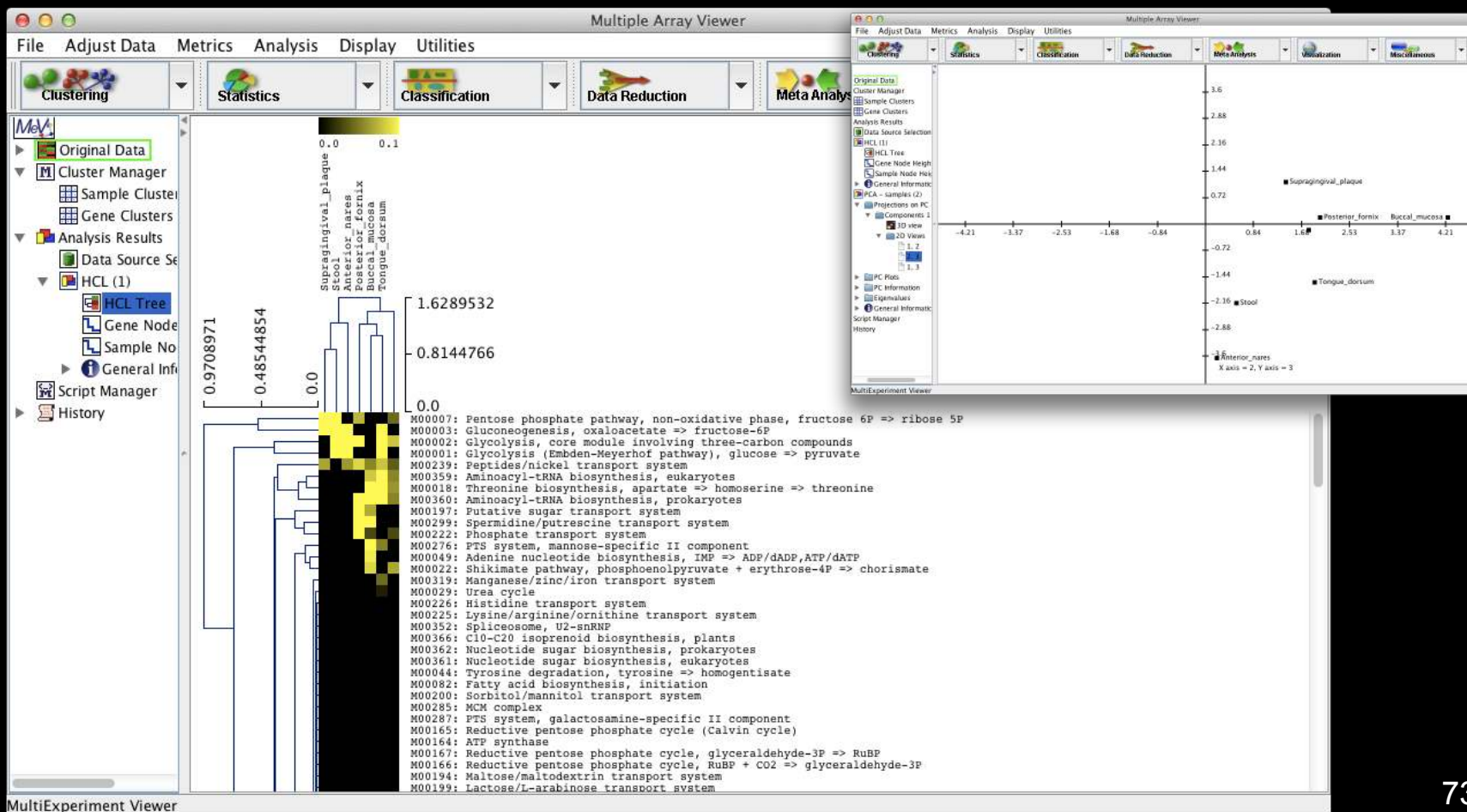
- And you can open the resulting table in Excel/etc.

# Gene Family	763577454- ϵ	763577454- ϵ	763577454- ϵ	763577454- ϵ	763577454- ϵ	763577454-SRS014494-Posterior_formix_Abundance
UniRef50_A9FGD2: 50S ribosomal protein L36	17.0940171	0	0	0	0	0
UniRef50_A9FGD2: 50S ribosomal protein L36 g__Bacteroides.s__Bacteroides_cellulosilyticus	17.0940171	0	0	0	0	0
UniRef50_R6AES1	14.8148148	0	0	0	0	0
UniRef50_R6AES1 g__Bacteroides.s__Bacteroides_stercoris	14.8148148	0	0	0	0	0
UniRef50_UPI00047E7990: glycosyl transferase family 1	0	8.96495518	0	0.81499593	3.2599837	1.05317188
UniRef50_UPI00047E7990: glycosyl transferase family 1 unclassified	0	8.96495518	0	0.81499593	3.2599837	1.05317188
UniRef50_UPI00047498D2: hypothetical protein, partial	3.003003	3.003003	3.003003	3.003003	0	0
UniRef50_UPI00047498D2: hypothetical protein, partial unclassified	3.003003	3.003003	3.003003	3.003003	0	0
UniRef50_A6L108	11.6959064	0	0	0	0	0
UniRef50_A6L108 g__Bacteroides.s__Bacteroides_stercoris	11.6959064	0	0	0	0	0
UniRef50_R6Q0V8: ABC-type metal ion transport system periplasmic component/surface antigen	3.23624595	0	2.1574973	2.1574973	3.23624595	0
UniRef50_R6Q0V8: ABC-type metal ion transport system periplasmic component/surface antigen unclassified	3.23624595	0	2.1574973	2.1574973	3.23624595	0
UniRef50_E6UAV0: Preprotein translocase, YajC subunit	2.94985251	0	0.98420138	5.89970501	0.65314007	0
UniRef50_E6UAV0: Preprotein translocase, YajC subunit unclassified	2.94985251	0	0.98420138	5.89970501	0.65314007	0
UniRef50_U2Q6I3	0	0	0	5.20833333	0	5.20833333
UniRef50_U2Q6I3 unclassified	0	0	0	5.20833333	0	5.20833333
UniRef50_A6KXA8: Transposase	10.0704935	0	0	0	0	0
UniRef50_A6KXA8: Transposase g__Bacteroides.s__Bacteroides_stercoris	10.0704935	0	0	0	0	0
UniRef50_UPI000374C24F: hypothetical protein	1.68740089	0	3.23238577	0	4.67294834	0
UniRef50_UPI000374C24F: hypothetical protein unclassified	1.68740089	0	3.23238577	0	4.67294834	0
UniRef50_P37247: Transposase for insertion sequence element IS4351	9.17431193	0	0	0	0	0
UniRef50_P37247: Transposase for insertion sequence element IS4351 g__Bacteroides.s__Bacteroides_cellulosilyticus	9.17431193	0	0	0	0	0
UniRef50_BONNQ2: Transposase	9.00031022	0	0	0	0	0
UniRef50_BONNQ2: Transposase g__Bacteroides.s__Bacteroides_stercoris	9.00031022	0	0	0	0	0
UniRef50_ROKTN8	0	0	1.51860289	1.51860289	3.8238261	1.79224317
UniRef50_ROKTN8 unclassified	0	0	1.51860289	1.51860289	3.8238261	1.79224317



What they're doing: HUMAnN2

- And there's nothing stopping us from using MeV
 - Or R, or QIIME, or LEfSe, or anything that'll read tab-delimited text





Quality control: KneadData

- Did you notice that we didn't QC our data at all?
 - MetaPhlAn2 is very robust to junk sequence
 - HUMAnN2 is pretty robust, but not quite as much
- Demo data includes standard metagenomic QC:
 - Quality trim by removing bad bases (typically Q ~15)
 - Length filter to remove short sequences (typically <75%)



Metagenome and metatranscriptome quality control: KneadData

- You can trim and filter reads, remove host contamination, and deplete ribosomal sequences using:

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

The Huttenhower Lab
Department of Biostatistics, Harvard T.H. Chan School of Public Health

HOME RESEARCH DOCUMENTATION PEOPLE CONTACT PUBLICATIONS

Home / KneadData

KneadData

KneadData is a tool designed to perform quality control on metagenomic sequencing data, especially data from microbiome experiments. In these experiments, samples are typically taken from a host in hopes of learning something about the microbial community on the host. However, metagenomic sequencing data from such experiments will often contain a high ratio of host to bacterial reads. This tool aims to perform principled in silico separation of bacterial reads from these "contaminant" reads, be they from the host, from bacterial 16S sequences, or other user-defined sources.

If you use the KneadData software, please cite our manuscript: TBD

Contents

- Requirements
- Installation
- How to run
- Basic usage
- Demo run

Click here



Metagenome and metatranscriptome quality control: KneadData

- KneadData performs quality trimming using Trimmomatic:

```
kneaddata -1 seq1.fastq -a SLIDINGWINDOW:4:20 -o seqs
```

- And read length filtering (including paired ends):

```
kneaddata -1 seq1.fastq -2 seq2.fastq \  
-a "SLIDINGWINDOW:4:20 MINLEN:60" -o seqs
```

- And will remove host (e.g. human) sequences from a reference database:

```
kneaddata -1 seq1.fastq -2 seq2.fastq \  
-a "SLIDINGWINDOW:4:20 MINLEN:60" \  
-db Homo_sapiens_db -o seqs
```

- And will remove ribosomal sequences (for metatranscriptomes):

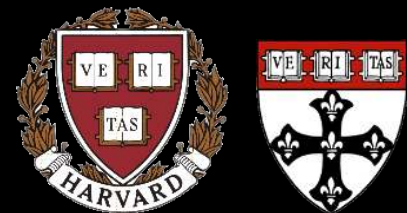
```
kneaddata -1 seq1.fastq -2 seq2.fastq \  
-a "SLIDINGWINDOW:4:20 MINLEN:60" \  
-db Homo_sapiens_db -db bact_rrna_db -o seqs
```




Multivariate associating testing with random effects using MaAsLin

Galeb Abu-Ali
Eric Franzosa
Curtis Huttenhower

09-18-15



Harvard School of Public Health
Department of Biostatistics





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



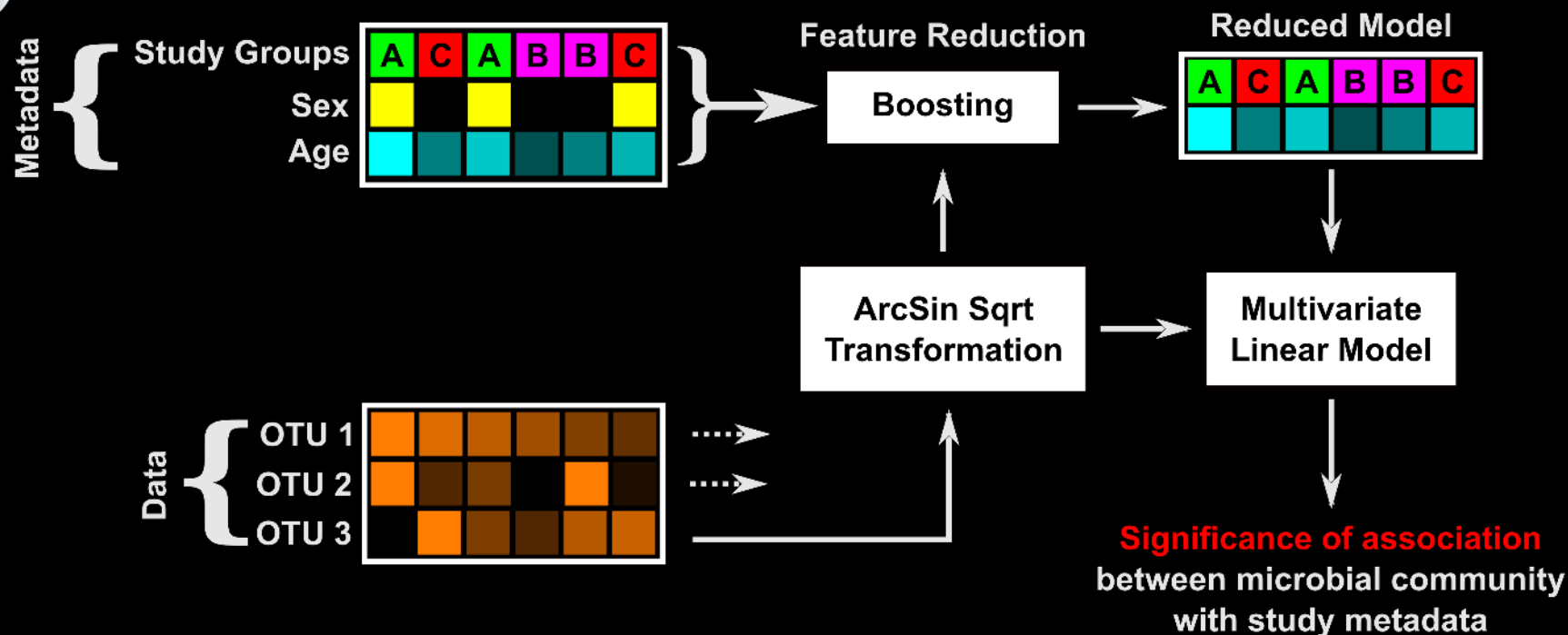
Properties of microbiome data

- 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



Multivariate microbial Association with Linear models



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

Tim
Tickle





Linking host and microbial function in ileal pouch inflammation

With Mark Silverberg

184 subjects with j-pouches
at Mt. Sinai since 1981



Boyko Kabakchiev



Andrea Tyler



Xochi Morgan



Levi Waldron

UC

NP No Pouchitis

(A/C)P Pouchitis (Acute/Chronic)

CDL Crohn's Disease-Like Phenotype

FAP

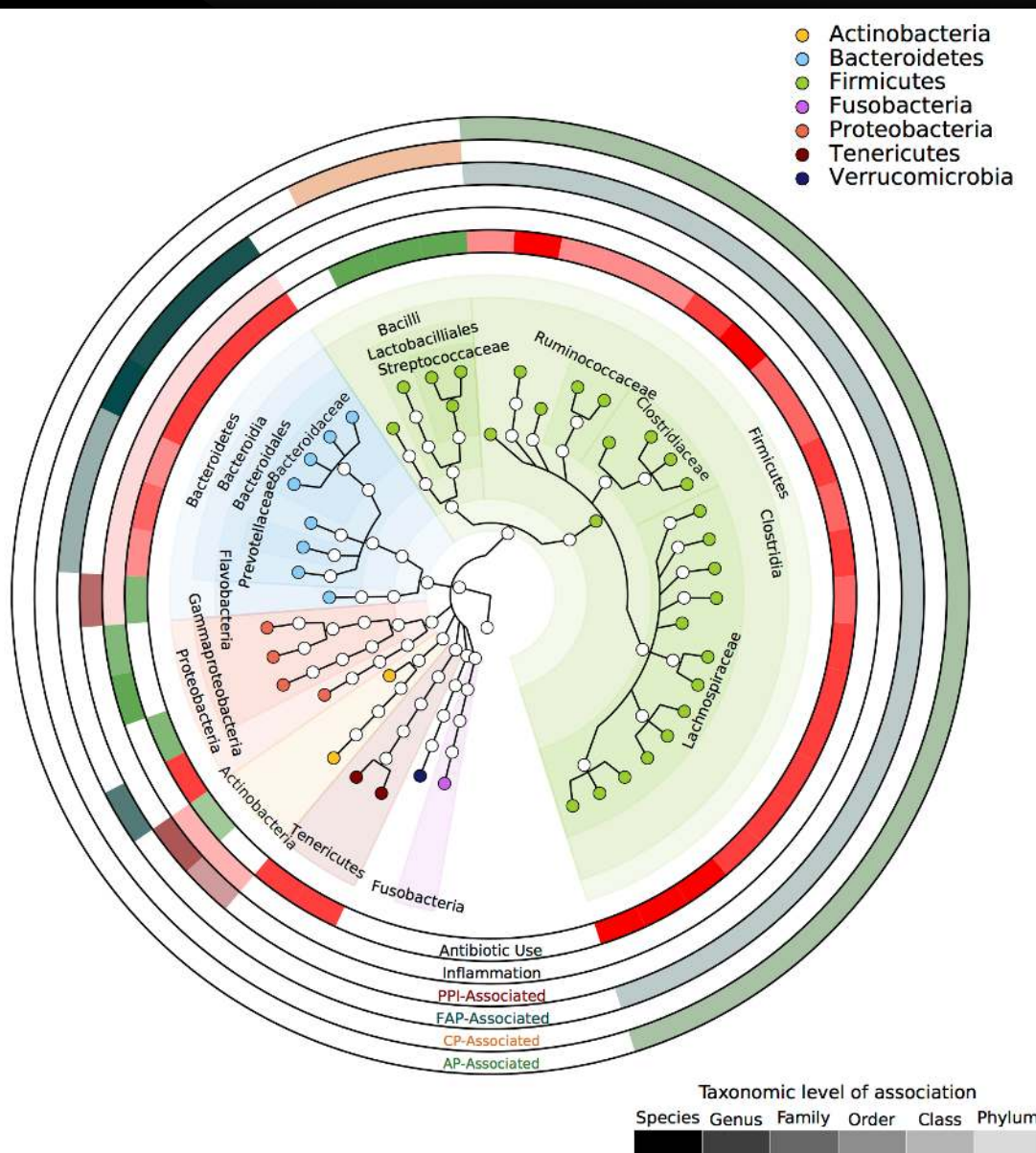
FAP Familial Adenomatous Polyposis

230 biopsies with host gene expression + microbiome

	FAP	NP	(A/C)P	CDL
Pouch	N 16	N 15	I 11	N/I 16
Pre-pouch ileum (PPI)	N 18	N 48	N 83	I 23



Multivariate association of microbes with pouchitis phenotypes



clade ~

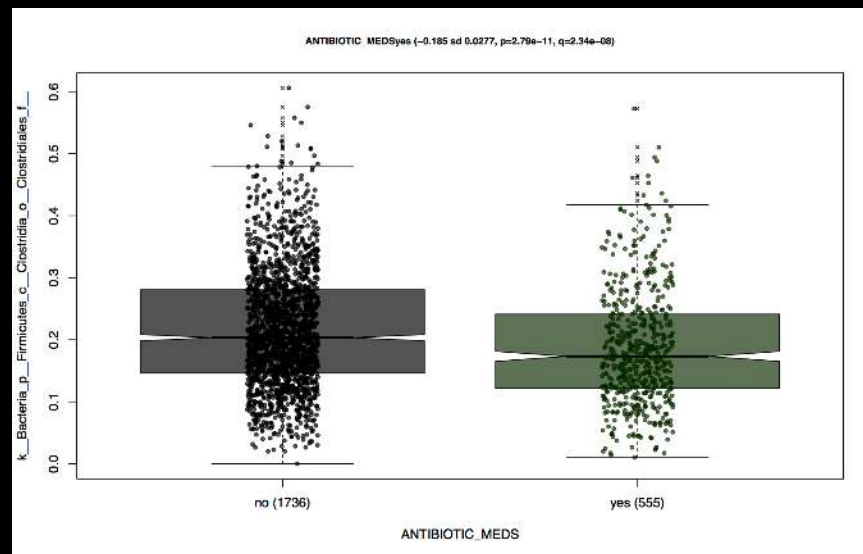
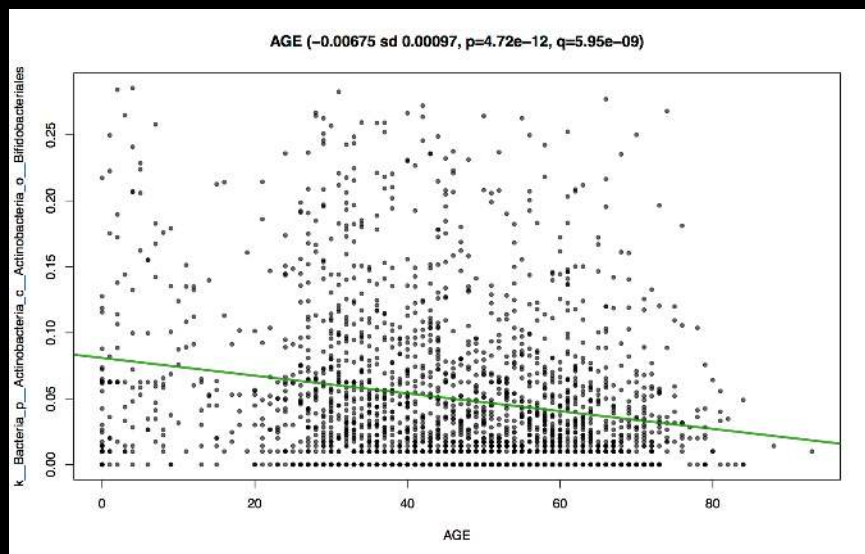
transcript +
location +
antibiotics +
inflammation +
phenotype

Can also include random effects (i.e. multiple samples per subject over time / space) and high dimensional models (e.g. genetics)



Multivariate association of microbes with phenotype in the American Gut

clade ~ acne_meds + age + alcohol + abx + asthma + BMI + carb% + country_now + country_birth + csection + diabetes + diet_type + dog + fat% + fiber + gluten + ibd + lactose_int + pregnant + protein% + race + sex





Setup notes reminder

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



Multivariate associations: MaAsLin

- You can find the MaAsLin install and documentation at: <http://huttenhower.sph.harvard.edu/maaslin>

The Huttenhower Lab
Department of Biostatistics, Harvard T.H. Chan School of Public Health

HOME RESEARCH DOCUMENTATION PEOPLE CONTACT PUBLICATIONS

Home / MaAsLin: Multivariate Association with Linear Models

MaAsLin: Multivariate Association with Linear Models

MaAsLin is a multivariate statistical framework that finds associations between clinical metadata and microbial community abundance or function. The clinical metadata can be of any type continuous (for example age and weight), boolean (sex, stool/biopsy), or discrete/factor (cohort groupings and phenotypes). MaAsLin is best used in the case when you are associating many metadata with microbial measurements. When this is the case each metadatum can be a different type. For example, you could include age, weight, sex, cohort and phenotype in the same input file to be analyzed in the same MaAsLin run. The microbial measurements are expected to be normalized before using MaAsLin and so are proportional data ranging from 0 to 1.0.

Install MaAsLin (preliminary version)

MaAsLin requires the following R packages: agricolae, gam, gamlss, gbm, glmnet, inlinedocs, logging, MASS, nlme, optparse, outliers, penalized, pscl, robustbase

Please install these packages before installing MaAsLin.

To install MaAsLin:

1. [Download](#) the latest version of MaAsLin.
2. Install MaAsLin where X.Y.Z is the version number)

Click here



Multivariate associations: MaAsLin

- But we've already installed it! Take a look:

```
ln -s /home/ubuntu/metagenomics/maaslin/  
./maaslin/R/Maaslin.R -h | less -S
```

```
3. screen (ssh)  
Usage: ./maaslin/R/Maaslin.R [options] <data.tsv> <outputdir>  
  
Options:  
  -h, --help  
          Show this help message and exit  
  
  -i DATA.READ.CONFIG, --input_config=DATA.READ.CONFIG  
          Optional configuration file describing data input format.  
  
  -I DATA.R, --input_process=DATA.R  
          Optional configuration script normalizing or processing data.  
  
  -d SIGNIFICANCE, --fdr=SIGNIFICANCE  
          The threshold to use for significance for the generated q-values  
          (BH FDR). Anything equal to or lower than this is significant. [Default 0.25]  
  
  -r MINRELATIVEABUNDANCE, --minRelativeAbundance=MINRELATIVEABUNDANCE  
          The minimum relative abundance allowed in the data. Values below  
          this are removed and imputed as the median of the sample data. [Default 1e-04]  
  
  -p MINPREVALENCE, --minPrevalence=MINPREVALENCE  
          The minimum percentage of samples in which a feature must have t  
          he minimum relative abundance in order not to be removed. Also this is the maxim  
  :█
```



Starting simple: a univariate test with MaAsLin

- Let's start by associating one covariate with microbiome data :

```
/home/ubuntu/metagenomics/util/metadata.py \  
    /home/ubuntu/metagenomics/data/hmp_metadata.dat \  
< /home/ubuntu/metagenomics/data/HMP.ab.filtered.txt \  
> HMP.ab.filtered.metadata.pcl  
grep -E '^(sid)|(STSite)|(k__Bacteria)' HMP.ab.filtered.metadata.pcl \  
> HMP.ab.filtered.stsite.pcl  
less -S HMP.ab.filtered.stsite.pcl
```

```
3. screen (ssh)  
sid SRS043001 SRS017127 SRS021473 SRS011134 SRS050184 SRS011529 SRS0  
STSite Stool Buccal_mucosa Buccal_mucosa Stool Posterior_fornix Stoo  
k__Bacterialp__Proteobacterialc__Betaproteobacteriales__Burkholderialeslf__Sutter  
k__Bacterialp__Actinobacterialc__Actinobacteriales__Coriobacterialeslf__Coriobact  
k__Bacterialp__Bacteroideteslf__Bacteroidialeslf__Porphyromonadace  
k__Bacterialp__Firmicuteslf__Clostridialeslf__Ruminococcacealf__R  
k__Bacterialp__Firmicuteslf__Clostridialeslf__Lachnospiracealf__B  
k__Bacterialp__Bacteroideteslf__Bacteroidialeslf__Bacteroidacealf__B  
k__Bacterialp__Actinobacterialc__Actinobacterialeslf__Actinomycetaleslf__Micrococca  
k__Bacterialp__Firmicuteslf__Bacillaleslf__Lactobacillaleslf__Streptococcacealf__S  
k__Bacterialp__Proteobacterialc__Epsilonproteobacteria 0.0 0.0049 1.10502 0.0  
k__Bacterialp__Firmicuteslf__Bacillaleslf__Lactobacillaleslf__Streptococcacealf__S  
k__Bacterialp__Fusobacterialeslf__Fusobacterialeslf__Leptotrichialeslf__Leptotrichiace  
k__Bacterialp__Bacteroideteslf__Bacteroidialeslf__Bacteroidacealf__B  
k__Bacterialp__Verrucomicrobialeslf__Verrucomicrobialeslf__Verrucomicrobialeslf__Verr  
k__Bacterialp__Fusobacterialeslf__Fusobacterialeslf__Leptotrichialeslf__Leptotrichiace  
k__Bacterialp__Bacteroideteslf__Bacteroidialeslf__Bacteroidacealf__B  
k__Bacterialp__Proteobacterialc__Epsilonproteobacterialeslf__Campylobacteriales 0.0  
k__Bacterialp__Firmicuteslf__Bacillaleslf__Bacillales 0.0 3.7702 5.24454 0.0 0.0  
k__Bacterialp__Firmicuteslf__Bacillaleslf__Lactobacillaleslf__Lactobacillacealf__L  
k__Bacterialp__Bacteroideteslf__Bacteroidialeslf__Bacteroidaleslf__Porphyromonadace  
k__Bacterialp__Bacteroideteslf__Bacteroidialeslf__Bacteroidaleslf__Porphyromonadace  
k__Bacterialp__Firmicuteslf__Negativicuteslf__Selenomonadaleslf__Acidaminococcac  
k__Bacterialp__Actinobacterialc__Actinobacterialeslf__Bifidobacterialeslf__Bifidoba  
HMP.ab.filtered.stsite.pcl
```



Starting simple: a univariate test with MaAsLin

- To run MaAsLin easily on one covariate:

```
./maaslin/R/Maaslin.R HMP.ab.filtered.stsite.pcl \  
HMP.ab.filtered.stsite --lastMetadata=2
```

```
3. screen (ssh)  
2015-08-13 13:31:44 INFO:maaslin:SR5054590  
2015-08-13 13:31:44 INFO:maaslin:Removing data 42 for being all NA after QC  
2015-08-13 13:31:45 INFO:maaslin:Removing the following for having only NAs after  
cleaning (maybe due to only having NA after outlier testing).  
2015-08-13 13:31:45 INFO:maaslin:k__Bacteria  
2015-08-13 13:31:45 INFO:maaslin:Outputting to: HMP.ab.filtered.stsite/HMP.ab.fi  
ltered_log.txt  
2015-08-13 13:31:46 INFO:maaslin:Taxon 10/182  
2015-08-13 13:31:49 INFO:maaslin:Taxon 20/182  
2015-08-13 13:31:52 INFO:maaslin:Taxon 30/182  
2015-08-13 13:31:54 INFO:maaslin:Taxon 40/182  
2015-08-13 13:31:59 INFO:maaslin:Taxon 60/182  
2015-08-13 13:32:02 INFO:maaslin:Taxon 70/182  
2015-08-13 13:32:05 INFO:maaslin:Taxon 80/182  
2015-08-13 13:32:08 INFO:maaslin:Taxon 90/182  
2015-08-13 13:32:11 INFO:maaslin:Taxon 100/182  
2015-08-13 13:32:13 INFO:maaslin:Taxon 110/182  
2015-08-13 13:32:16 INFO:maaslin:Taxon 120/182  
2015-08-13 13:32:19 INFO:maaslin:Taxon 130/182  
2015-08-13 13:32:22 INFO:maaslin:Taxon 140/182  
2015-08-13 13:32:24 INFO:maaslin:Taxon 150/182  
2015-08-13 13:32:28 INFO:maaslin:Taxon 160/182  
2015-08-13 13:32:30 INFO:maaslin:Taxon 170/182  
2015-08-13 13:32:34 INFO:maaslin:Taxon 180/182  
[chuttenhower@class03 ~]$
```



Starting simple: a univariate test with MaAsLin

- What are all these files!?!

```
ls -R HMP.ab.filtered.stsite
```

```
3. screen (ssh)
[chuttenhower@class03 ~]$ ls -R HMP.ab.filtered.stsite
HMP.ab.filtered.stsite:
generated_config          HMP.ab.filtered.stsite.tsv  QC
HMP.ab.filtered_log.txt  HMP.ab.filtered.stsite.txt
HMP.ab.filtered-STSite.pdf HMP.ab.filtered-STSite.txt

HMP.ab.filtered.stsite/QC:
data.read.config          ProcessQC.txt               read-Merged.tsv
data.tsv                  read_cleaned.read.config  Run_Parameters.txt
metadata.read.config      read_cleaned.tsv
metadata.tsv              read-Merged.read.config
```

- First, processed inputs:
 - **generated_config** indicates how MaAsLin read your data
 - One matrix containing metadata model variables, one containing data
 - **QC** directory contains separate files for data + metadata
 - **Run_Parameters.txt** contains model variables
 - Was it sparse, were there random variables, what filtering criteria, etc.

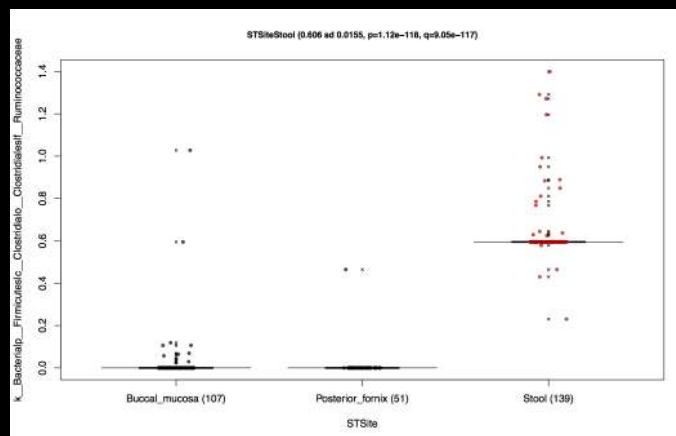


Starting simple: a univariate test with MaAsLin

- Next, what you care about, outputs:
 - **HMP.ab.filtered-STSite.txt** lists associations between clades and the model variable STSite and their significance

Variable	Feature	Value	Coefficient	N	N not 0	P-value	Q-value
STSite	k__Bacteria p__Firmicutes c__Bacilli o__Bacillales f__Bacillales_uncl g__Gemella s__Gemella_haemolysans	STSiteStool	-0.7036054	297	112	2.48E-196	1.60E-193
STSite	k__Bacteria p__Firmicutes c__Bacilli o__Bacillales f__Bacillales_uncl g__Gemella	STSiteStool	-0.6315029	297	114	1.04E-169	2.23E-167
STSite	k__Bacteria p__Firmicutes c__Bacilli o__Bacillales f__Bacillales_uncl	STSiteStool	-0.6315029	297	114	1.04E-169	2.23E-167
STSite	k__Bacteria p__Firmicutes c__Bacilli o__Bacillales f__Bacillales_uncl g__Gemella s__Gemella_haemolysans	STSitePosterior_fornix	-0.7047368	297	112	5.18E-163	8.35E-161
STSite	k__Bacteria p__Firmicutes c__Bacilli o__Bacillales f__Bacillales_uncl g__Gemella	STSitePosterior_fornix	-0.6344187	297	114	7.57E-138	8.12E-136
STSite	k__Bacteria p__Firmicutes c__Bacilli o__Bacillales f__Bacillales_uncl	STSitePosterior_fornix	-0.6344187	297	114	7.57E-138	8.12E-136
STSite	k__Bacteria p__Firmicutes c__Bacilli o__Bacillales	STSiteStool	-0.5226617	297	120	3.27E-119	3.01E-117
STSite	k__Bacteria p__Firmicutes c__Clostridia o__Clostridiales f__Ruminococcaceae	STSiteStool	0.60575613	297	151	1.12E-118	9.05E-117
STSite	k__Bacteria p__Firmicutes c__Clostridia o__Clostridiales f__Ruminococcaceae g__Faecalibacterium	STSiteStool	0.5194707	297	143	1.88E-111	1.35E-109

- **HMP.ab.filtered-STSite.pdf** plots them



- **HMP.ab.filtered_log.txt** logs all tests, significant or not



Multivariate tests with MaAsLin

- Let's run a more interesting IBD model:

```
ln -s /home/ubuntu/metagenomics/data/ibd2012.pcl
```

```
less -S ibd2012.pcl
```

3. screen (ssh)

sample	7007	7010	7016	7018	7021	7022	7035	7037	7039										
age	28	41	36	30	39	34	45	32	50	54	38	56	40	44	53	49	35	55	49
antibiotics	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
dx	CD	CD	UC	UC	CD	CD	CD	CD	CD	UC	CD	CD	CD	CD	CD	CD	CD	CD	CD
gender	0	1	0	0	0	0	0	0	1	1	0	1	0	1	1	0	0	0	0
ileal	1	1	0	0	0	1	1	1	1	1	0	1	1	1	1	0	1	0	1
immunosup	1	1	1	1	0	1	0	1	1	0	0	1	1	1	1	1	0	1	1
mesalamine	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	1	1	0
smoker	1	1	1	1	0	0	0	1	0	0	1	1	0	1	2	0	0	1	1
steroids	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
stool	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Archaea Euryarchaeota Methanobacteria Methanobacteriales Methanobacteriaceae																			
Archaea Euryarchaeota Methanobacteria Methanobacteriales Methanobacteriaceae Met																			
Archaea Euryarchaeota Methanobacteria Methanobacteriales Methanobacteriaceae Met																			
Bacteria	0.976701	0.923429	1	0.972956	0.803437	0.696739	0.94												
Bacteria Actinobacteria Actinobacteria	0	0.00457143	0	0.0197628	0.003209														
Bacteria Actinobacteria Actinobacteria Actinomycetales	0	0.00114286	0	0.00															
Bacteria Actinobacteria Actinobacteria Actinomycetales Actinomycetaceae	0	0																	
Bacteria Actinobacteria Actinobacteria Actinomycetales Actinomycetaceae Actinomy																			
Bacteria Actinobacteria Actinobacteria Actinomycetales Actinomycetaceae Varibacu																			
Bacteria Actinobacteria Actinobacteria Actinomycetales Actinomycetaceae unclassi																			
Bacteria Actinobacteria Actinobacteria Actinomycetales Brevibacteriaceae Breviba																			
Bacteria Actinobacteria Actinobacteria Actinomycetales Corynebacteriaceae Coryne																			
Bacteria Actinobacteria Actinobacteria Actinomycetales Dermabacteraceae	0	0																	

ibd2012.pcl



Multivariate tests with MaAsLin

- Running this isn't any harder than before (it just takes a little longer):

```
./maaslin/R/Maaslin.R ibd2012.pcl ibd2012 --lastMetadata=11
```

```
3. screen (ssh)
2015-08-13 16:31:29 INFO:maaslin:Grubbs Test::Removing 6 outliers from unclassified
2015-08-13 16:31:29 INFO:maaslin:7049
2015-08-13 16:31:29 INFO:maaslin:7059
2015-08-13 16:31:29 INFO:maaslin:7095
2015-08-13 16:31:29 INFO:maaslin:7125
2015-08-13 16:31:29 INFO:maaslin:7233
2015-08-13 16:31:29 INFO:maaslin:7871
2015-08-13 16:31:29 INFO:maaslin:Grubbs Test::Removing 2 outliers from smoker
2015-08-13 16:31:29 INFO:maaslin:7164
2015-08-13 16:31:29 INFO:maaslin:7610
2015-08-13 16:31:29 INFO:maaslin:Grubbs Test::Removing 1 outliers from steroids
2015-08-13 16:31:29 INFO:maaslin:7871
2015-08-13 16:31:30 INFO:maaslin:Outputting to: ibd2012/ibd2012_log.txt
2015-08-13 16:31:38 INFO:maaslin:Taxon 70/360
2015-08-13 16:31:42 INFO:maaslin:Taxon 100/360
2015-08-13 16:31:51 INFO:maaslin:Taxon 150/360
2015-08-13 16:31:54 INFO:maaslin:Taxon 170/360
2015-08-13 16:31:57 INFO:maaslin:Taxon 180/360
2015-08-13 16:31:59 INFO:maaslin:Taxon 190/360
2015-08-13 16:32:04 INFO:maaslin:Taxon 210/360
2015-08-13 16:32:06 INFO:maaslin:Taxon 220/360
2015-08-13 16:32:10 INFO:maaslin:Taxon 260/360
2015-08-13 16:32:19 INFO:maaslin:Taxon 360/360
[chuttenhower@class03 ~]$
```



Multivariate tests with MaAsLin

- We can see all of the significant results:

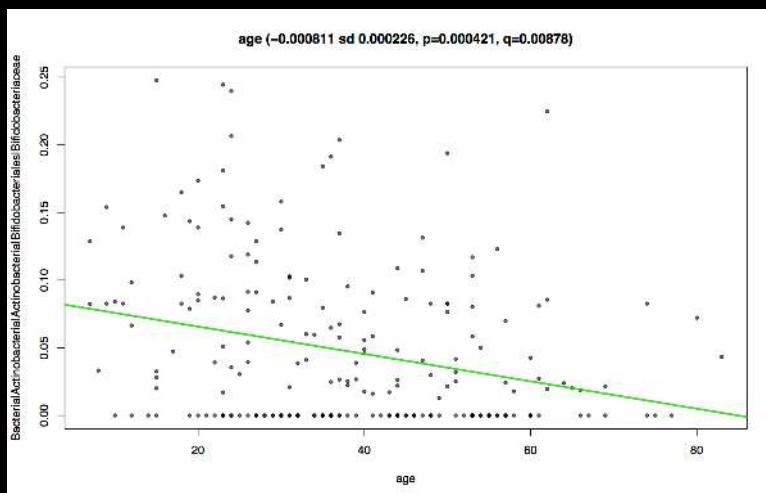
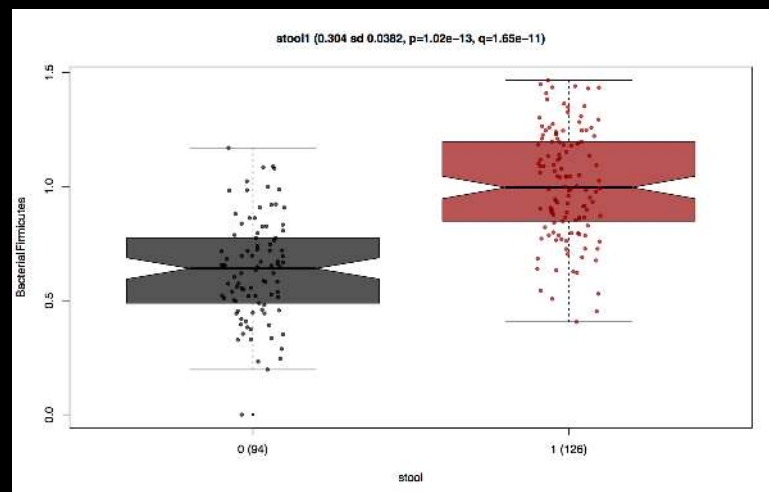
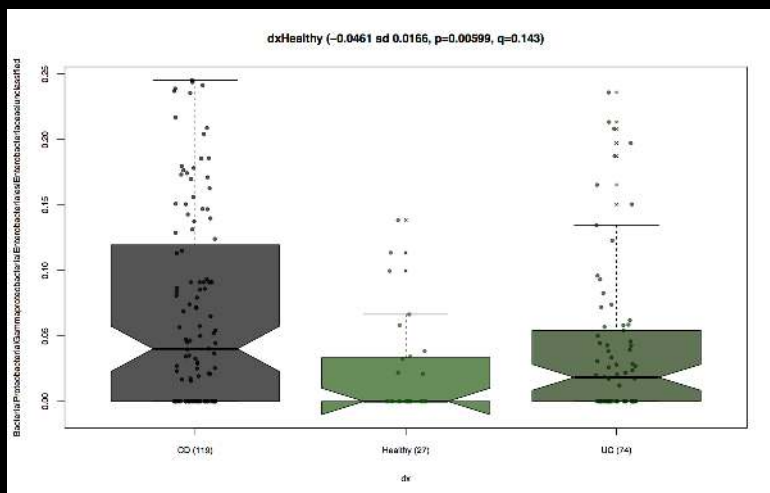
```
less -S ibd2012/ibd2012.txt
```

3. screen (ssh)									
Variable	Feature	Value	Coefficient	N	N.not.0	P.value	Q.value		
1	age	Bacteria Bacteroidetes	age	0.00434507175397982	220	212	0.00016448840082		
2	age	Bacteria Bacteroidetes Bacteroidia Bacteroidales Bacteroidaceae Bacteroides	age	0.00387834	220	206	0.00033318	0.00781462	
3	age	Bacteria Firmicutes Bacilli Lactobacillales Leuconostocaceae Leuconostoc	age	-0.000176	220	29	0.00035114	0.00808887	
4	age	Bacteria Firmicutes Clostridia Clostridiales Ruminococcaceae Sporobacter	age	0.00014751	220	53	0.0003815	0.0084851	
5	age	Bacteria Actinobacteria Actinobacteria Bifidobacteriales Bifidobacteriac	age	-0.0008113	220	126	0.00042121	0.00877677	
6	age	Bacteria Actinobacteria Actinobacteria Bifidobacteriales Bifidobacteriac	age	-0.0008086	220	126	0.00042663	0.00877677	
7	age	Bacteria Bacteroidetes Bacteroidia Bacteroidales	age	0.00409313	220	212	0.00083297	0.01557283	
8	age	Bacteria Firmicutes Bacilli Lactobacillales Leuconostocaceae	age	-0.0001888	220	39	0.00089099	0.01618847	
9	age	Bacteria Firmicutes	age	-0.0028628	220	219	0.00333717	0.05204288	
10	age	Bacteria Proteobacteria Gammaproteobacteria Pasteurellales Pasteurellaceae	age	-0.0001436	220	27	0.00362747	0.05570764	
11	age	Bacteria Firmicutes Clostridia Clostridiales Incertae_Sedis_XI Parvimonas	age	0.00016739	220	26	0.00704009	0.09460119	
12	age	Bacteria Firmicutes Clostridia Clostridiales Incertae_Sedis_XI Parvimonas	age	0.00011156	220	47	0.0082163	0.10815338	
13	age	Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae Parabacteroides	age	0.00097951	220	147	0.00922914	0.11672154	
14	age	Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae	age	0.00101908	220	163	0.00942202	0.11800397	
15	age	Bacteria Firmicutes Clostridia Clostridiales Ruminococcaceae Butyrivibrio	age	0.00036856	220	171	0.01802349	0.19215128	
16	age	Bacteria Firmicutes Clostridia Clostridiales Veillonellaceae unclassified	age	0.00022306	220	104	0.02138539	0.21714042	
17	antibiotics	Bacteria Actinobacteria Actinobacteria Coriobacteriales Coriobacteriaceae Collinsella	antibiotics1	-0.0862293	220	118	5.82E-06	0.00025891	
18	antibiotics	Bacteria Firmicutes Clostridia Clostridiales Lachnospiraceae Dorea	antibiotics1	-0.0734784	220	197	4.79E-05	0.00162547	
19	antibiotics	Bacteria Bacteroidetes Bacteroidia Bacteroidales Rikenellaceae	antibiotics1	-0.0428438	220	132	0.01213827	0.14365478	
20	antibiotics	Bacteria Bacteroidetes	antibiotics1	-0.1419265	220	212	0.01485215	0.16660234	
21	antibiotics	Bacteria Bacteroidetes Bacteroidia Bacteroidales Rikenellaceae Alistipes	antibiotics1	-0.0402089	220	132	0.01502882	0.16713084	
22	antibiotics	Bacteria Bacteroidetes Bacteroidia Bacteroidales	antibiotics1	-0.1415995	220	212	0.01525422	0.16818757	
23	antibiotics	Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae Odoribacter	antibiotics1	-0.0129536	220	72	0.01542788	0.16866076	
	antibiotics	Bacteria Firmicutes Clostridia Clostridiales Veillonellaceae unclassified	antibiotics1	-0.0119452	220	104	0.01697982	0.18253305	
	antibiotics	Bacteria Firmicutes Clostridia Clostridiales Veillonellaceae Phascolarctobacterium	antibiotics1	-0.0278794	220	95	0.02117856	0.21682815	
	antibiotics	Bacteria Bacteroidetes Bacteroidia Bacteroidales Bacteroidaceae Bacteroides	antibiotics1	-0.1250394	220	206	0.02171404	0.21714042	
	dx	Bacteria Firmicutes Clostridia Clostridiales Lachnospiraceae Roseburia	dxHealthy	0.12959193	220	177	2.55E-07	2.63E-05	
	dx	Bacteria Proteobacteria Gammaproteobacteria Enterobacteriales Enterobacteriaceae Cronobacter	dxHealthy	-0.0190217	220	66	3.08E-05	0.00193979	
	dx	Bacteria Firmicutes Clostridia Clostridiales Ruminococcaceae Ruminococcus	dxHealthy	0.05645538	220	114	0.00038492	0.01460437	



Multivariate tests with MaAsLin

- Or you can visualize raw data for individual variables:



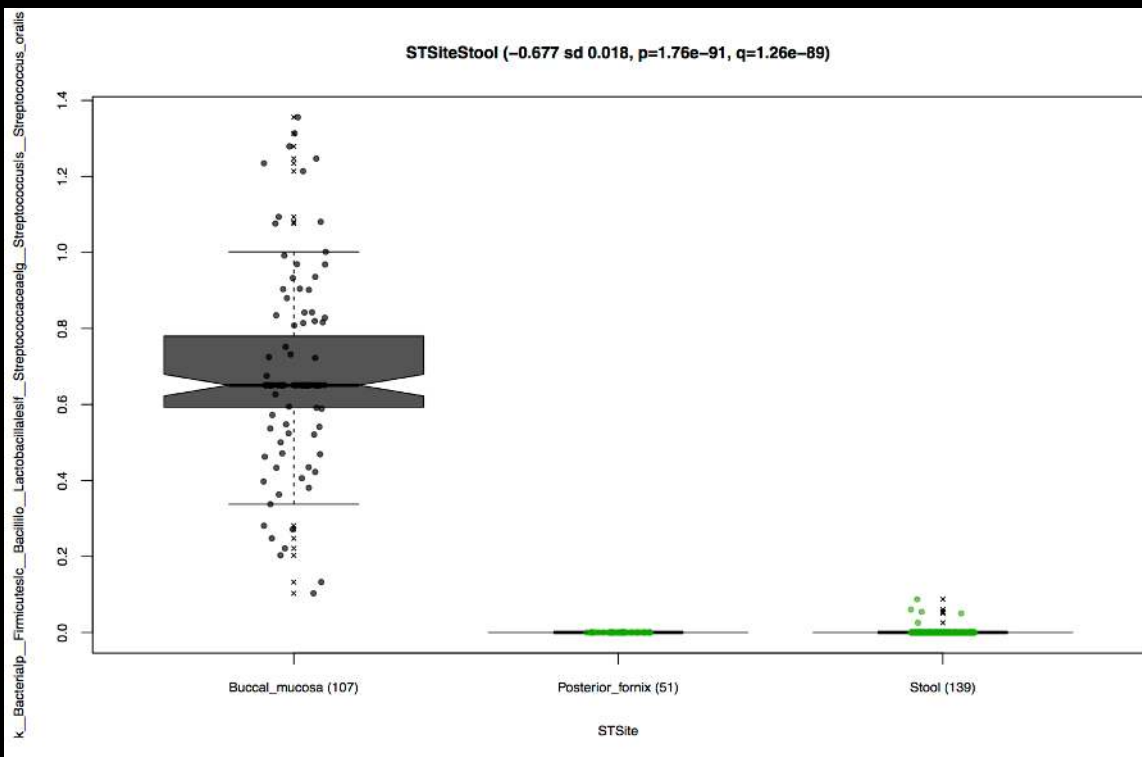


Advanced MaAsLin: random effects

- What if you have multiple samples / subject?

```
grep -E '^(sid)|(RANDSID)|(STSite)|(k__)' HMP.ab.filtered.metadata.pcl > HMP.ab.filtered.subject.pcl  
./maaslin/R/Maaslin.R HMP.ab.filtered.subject.pcl HMP.ab.filtered.subject --lastMetadata=3 -R RANDSID
```

```
3. screen (ssh)  
iteration 2  
iteration 1  
iteration 2  
iteration 1  
iteration 2  
iteration 1  
iteration 2  
iteration 1  
iteration 2  
iteration 1  
iteration 2  
iteration 1  
iteration 2  
iteration 1  
iteration 2  
2015-08-13 16:53:53 INFO:maaslin:Taxon 180/183  
iteration 1  
iteration 2  
iteration 1  
iteration 2  
iteration 1  
iteration 2  
iteration 1  
iteration 2  
[chuttenhower@class03 ~]$
```

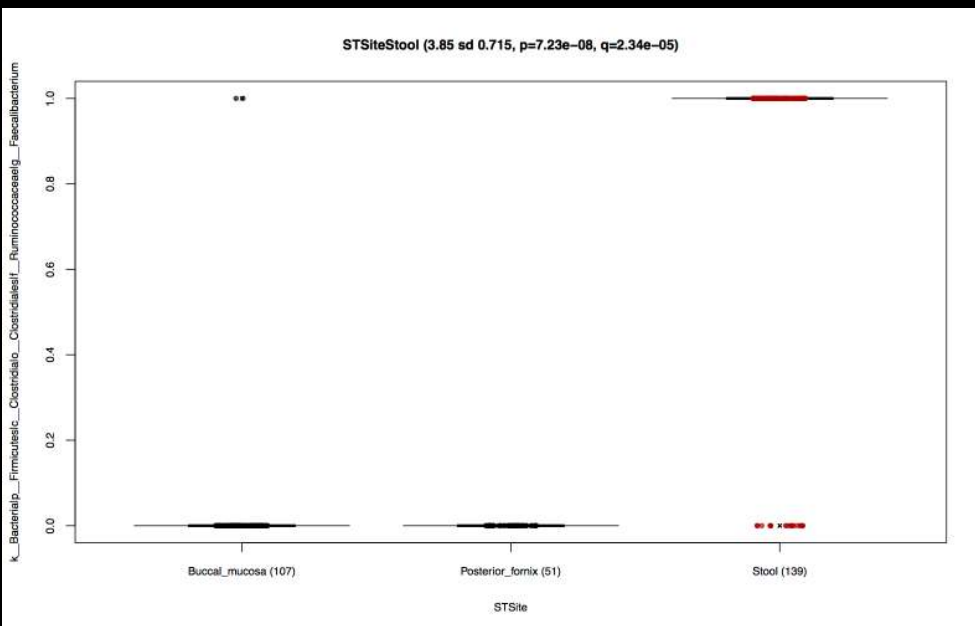




Advanced MaAsLin: other models

- What if you like to use a zero-inflated negative binomial?
 - Be warned, R's version is buggy as all get go...

```
./maaslin/R/Maaslin.R HMP.ab.filtered.stsite.pcl HMP.ab.filtered.negbin --lastMetadata=2 -z -m neg_binomial
```



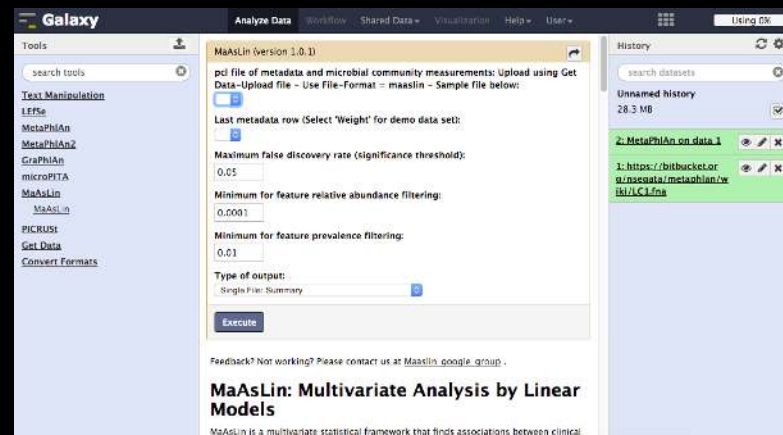
```
3. screen (ssh)
system is computationally singular: reciprocal condition number = 3.70325e-22
Error in solve.default(as.matrix(fit$hessian)) :
system is computationally singular: reciprocal condition number = 6.55509e-17
2015-08-13 16:54:58 INFO:maaslin:Taxon 140/182
Error in solve.default(as.matrix(fit$hessian)) :
system is computationally singular: reciprocal condition number = 6.80142e-20
2015-08-13 16:55:00 INFO:maaslin:Taxon 150/182
Error in solve.default(as.matrix(fit$hessian)) :
Lapack routine dgesv: system is exactly singular: U[7,7] = 0
Error in solve.default(as.matrix(fit$hessian)) :
system is computationally singular: reciprocal condition number = 2.32384e-21
Error in solve.default(as.matrix(fit$hessian)) :
system is computationally singular: reciprocal condition number = 1.06805e-43
Error in solve.default(as.matrix(fit$hessian)) :
system is computationally singular: reciprocal condition number = 3.70325e-22
2015-08-13 16:55:06 INFO:maaslin:Taxon 160/182
2015-08-13 16:55:09 INFO:maaslin:Taxon 170/182
Error in solve.default(as.matrix(fit$hessian)) :
system is computationally singular: reciprocal condition number = 2.56586e-21
Error in solve.default(as.matrix(fit$hessian)) :
system is computationally singular: reciprocal condition number = 1.08876e-21
2015-08-13 16:55:14 INFO:maaslin:Taxon 180/182
Error in solve.default(as.matrix(fit$hessian)) :
system is computationally singular: reciprocal condition number = 6.01582e-20
[chuttenhower@class03 ~]$
```




Advanced MaAsLin: more tricks

- MaAsLin can...
 - Change all QC and significance parameters
 - Minimum relative abundance, prevalence, FDR threshold, method, etc.
 - Skip some metadata
 - Check out the `read.config` file documentation and `-i` flag
 - Use other regularization approaches
 - Not just boosting: LASSO, forward, backward, none, etc. (`-s` flag)
 - Test all-against-all features, e.g. all bugs against all genetic variants
 - `-a` flag, use with caution – power!

- And there's a beta Galaxy interface
 - <http://huttenhower.sph.harvard.edu/maaslin>





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

Huttenhower Lab Tools

Welcome to the official Huttenhower Tutorials wiki.

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



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

Composition Analysis

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

HUMANN

- Microbial species and associated genes and pathways

MetaPhlAn

- Microbial species and abundances

PhyloPhlAn

- Reconstruction of phylogenetic trees

PICRUSt

- Predict metagenome functional content from marker gene

ShortBRED

- Abundance of proteins of interest in genetic data

Statistical Analysis

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

AREP

- Extract 'omics data from repositories

CCREPE

- Assess the significance of general similarity measures in compositional datasets

LEfSe

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

MaAsLin

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

microPITA

- Sample selection in two stage-tiered studies

Visualization

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



Thank you!



Curtis
Huttenhower



Xochitl
Morgan



Minah
Iqbal



Lauren
McIver



George
Weingart



Aleksandar
Kostic



Gholamali
Rahnava



Melanie
Schirmer



Alexandra
Sirota-Madi



Ayshwarya
Subramanian



Tiffany
Hsu



Boyu
Ren



Emma
Schwager



Koji
Yasuda



Siyuan
Ma



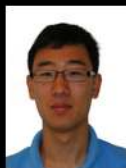
Casey
DuLong



Jim
Kaminski



Randall
Schwager



Andy
Shi



Moran
Yassour



Luc
Bijnens



Tommi
Vatanen



Levi Waldron



Nicola Segata



Wendy Garrett
Michelle Rooks



Dirk Gevers
Kat Huang



Ramnik Xavier
Harry Sokol
Dan Knights
Moran Yassour



Rob Beiko
Morgan Langille



Jacques Izard
Katherine Lemon



Ruth Ley
Omry Koren



Rob Knight
Greg Caporaso
Jesse Zaneveld



Bruce Sands



Mark Silverberg
Boyko Kabakchiev
Andrea Tyler



Human Microbiome Project

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

Bruce Birren Mark Daly
Doyle Ward Ashlee Earl



