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

Curtis Huttenhower

01-19-15

Harvard School of Public Health
Department of Biostatistics

U. Oregon

BROAD INSTITUTE

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:

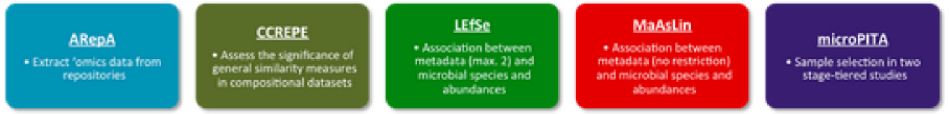
The wiki provides tutorials for Huttenhower tools, illustrating through demos how to use these tools on your datasets. Huttenhower tools can be divided under three main categories as shown below. Click on the tool for the corresponding tutorial.

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:

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:

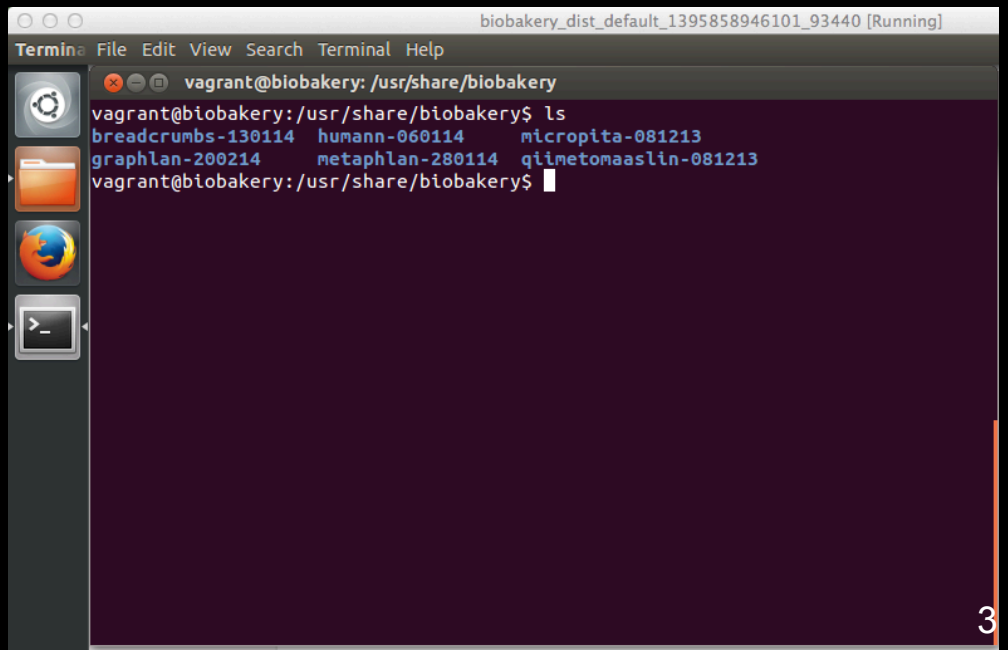


Visualization

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)



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

The two big questions…

Who is there?
(taxonomic profiling)

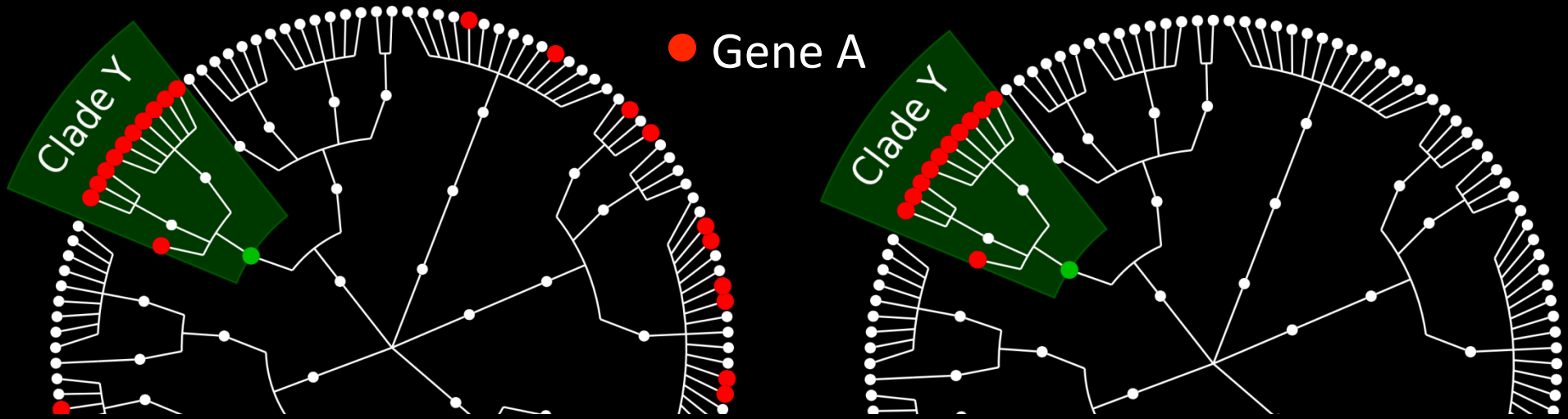
What are they doing?
(functional profiling)

MetaPhlAn overview

Nicola Segata

A is a core gene for clade Y

A is a unique marker gene for clade Y



ChocoPhlAn (offline pipeline)

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

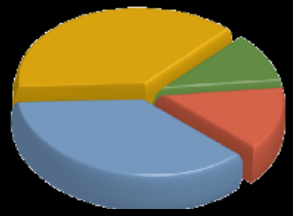
Unique marker genes DB

Available reference genomes

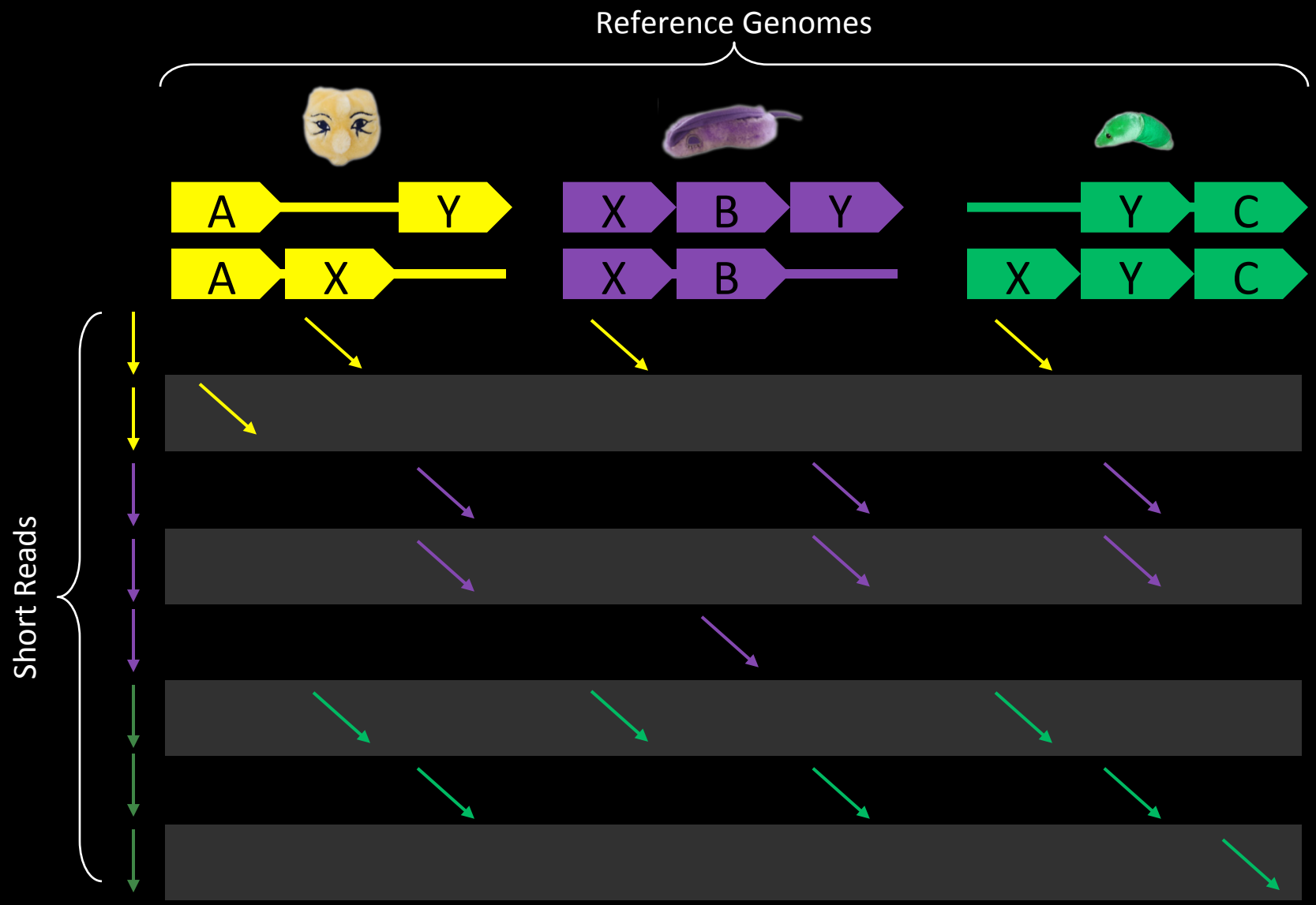
MetaPhlAn

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

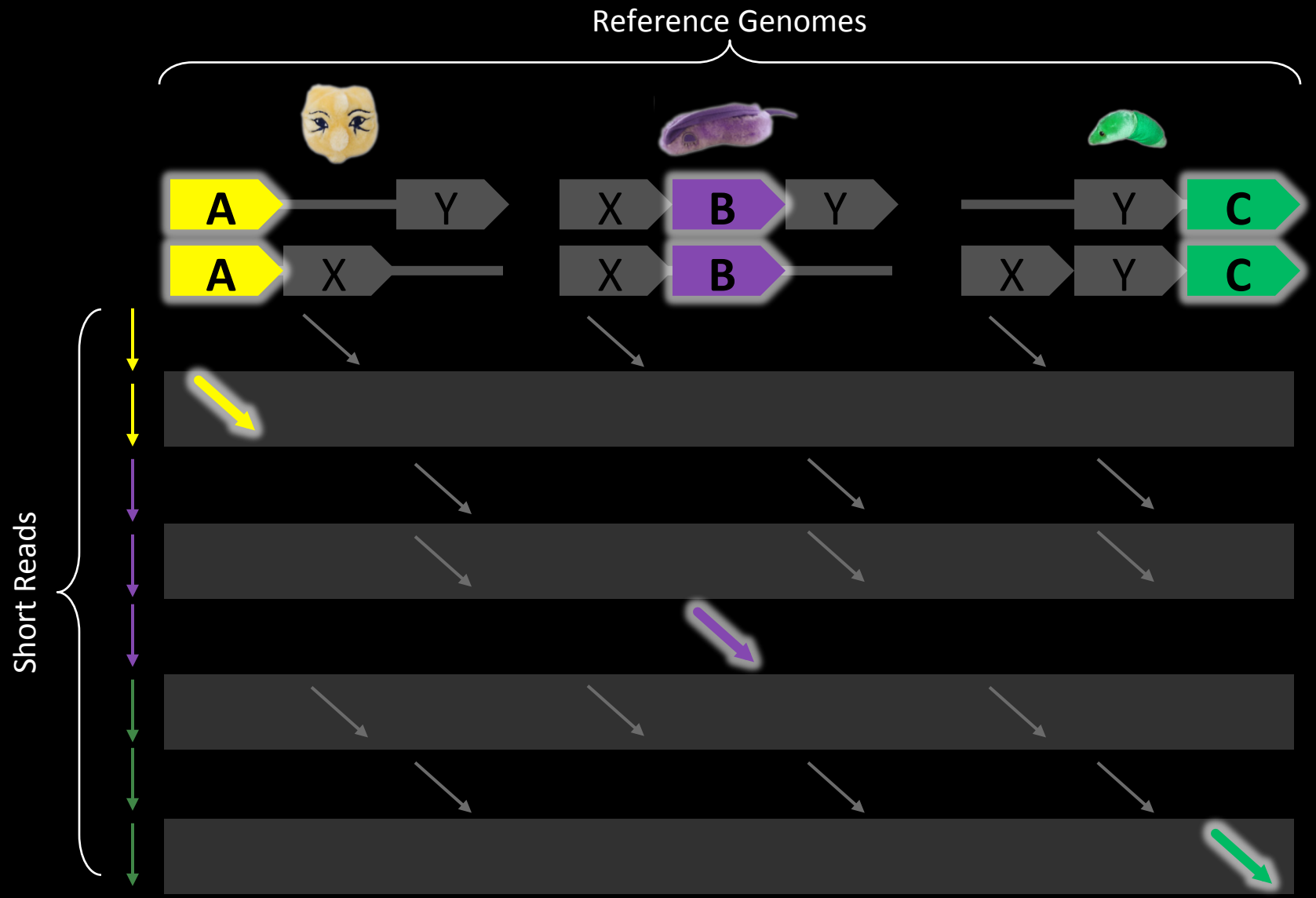
Metagenome



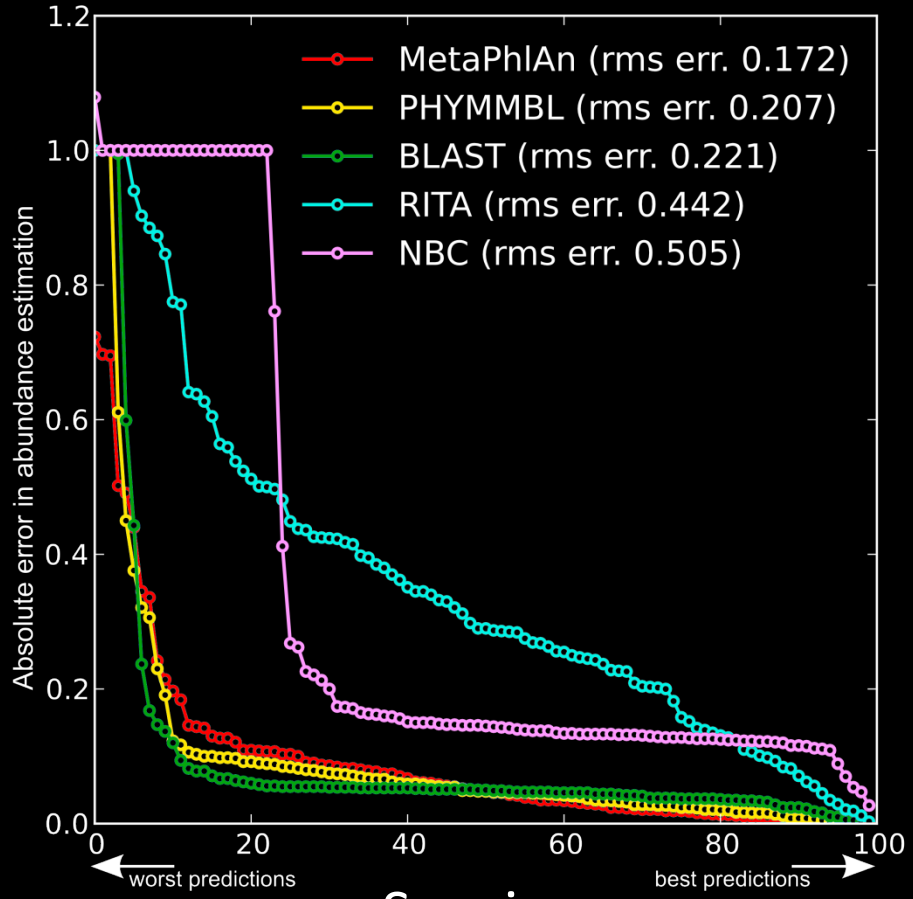
MetaPhlAn
Metagenomic Phylogenic Analysis

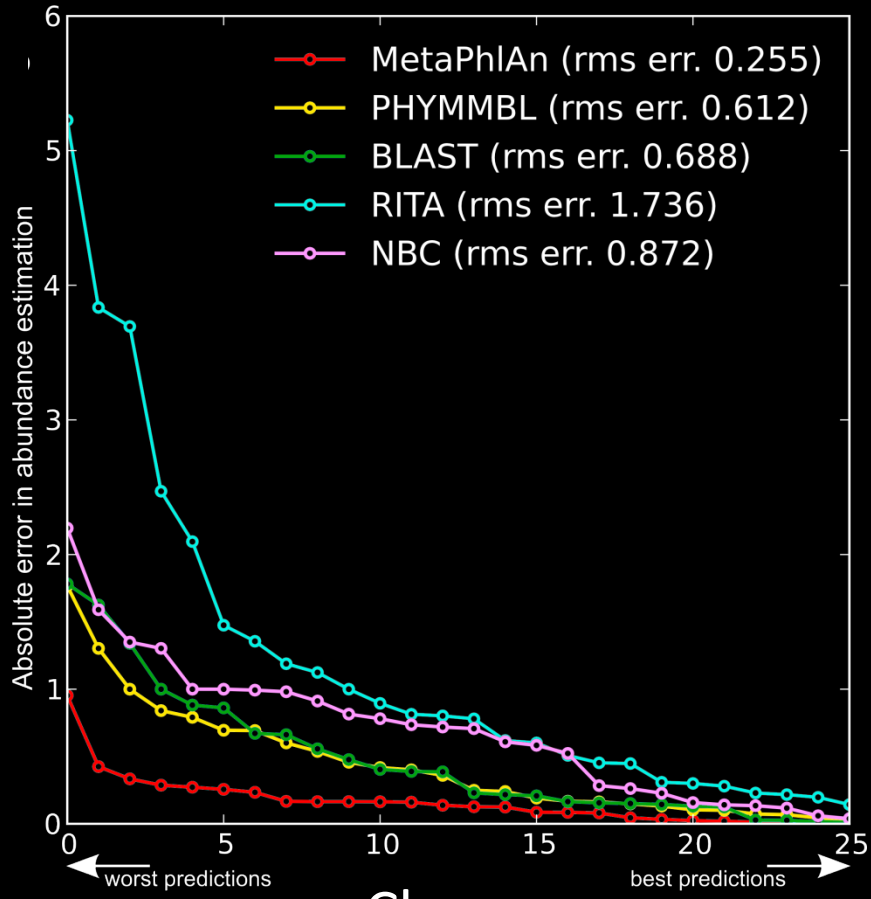


MetaPhlAn
Metagenomic Phylogenic Analysis



Evaluation of MetaPhlAn accuracy

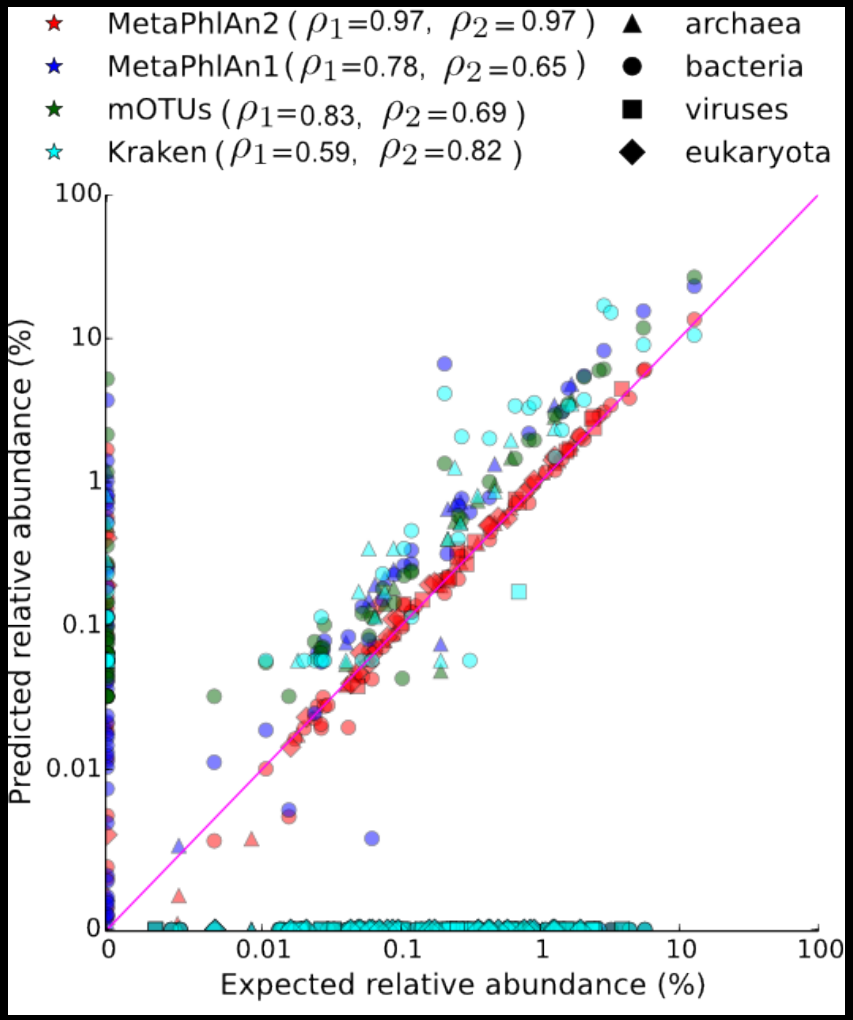


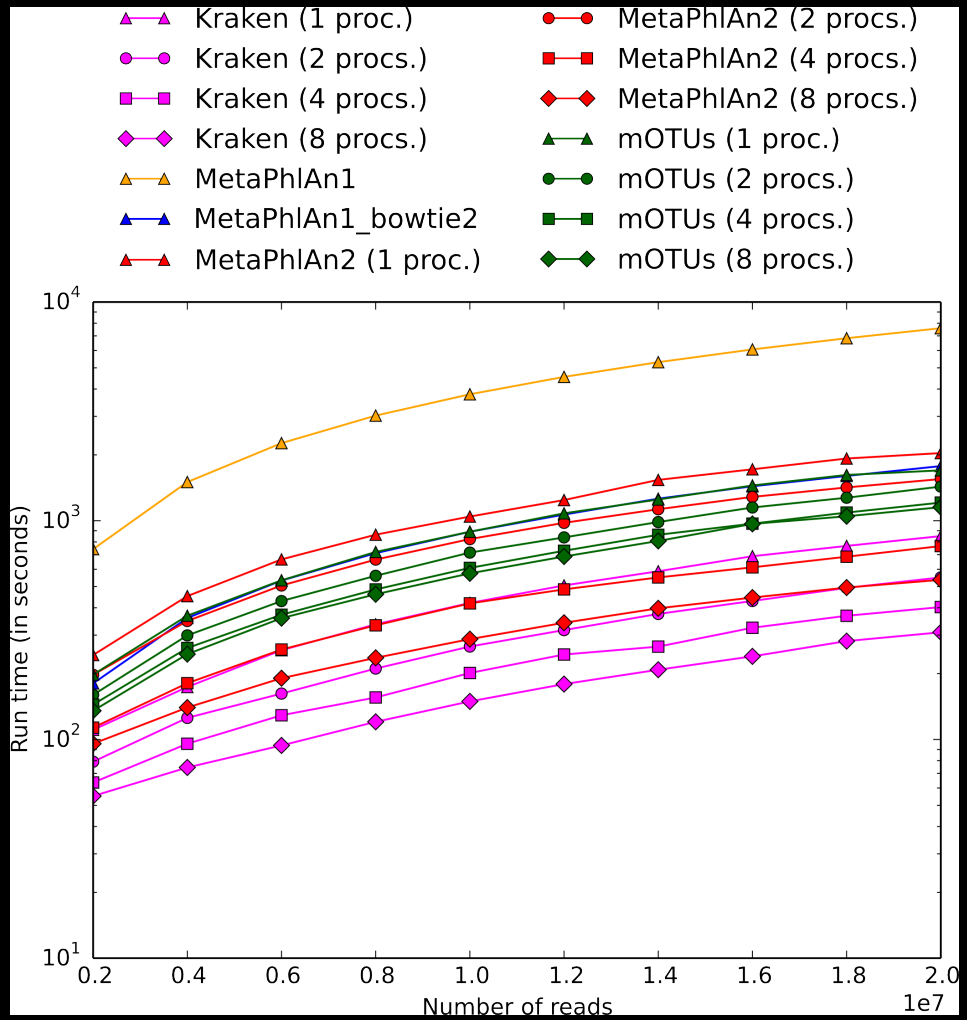


(Validation on high-complexity uniformly distributed synthetic metagenomes.)

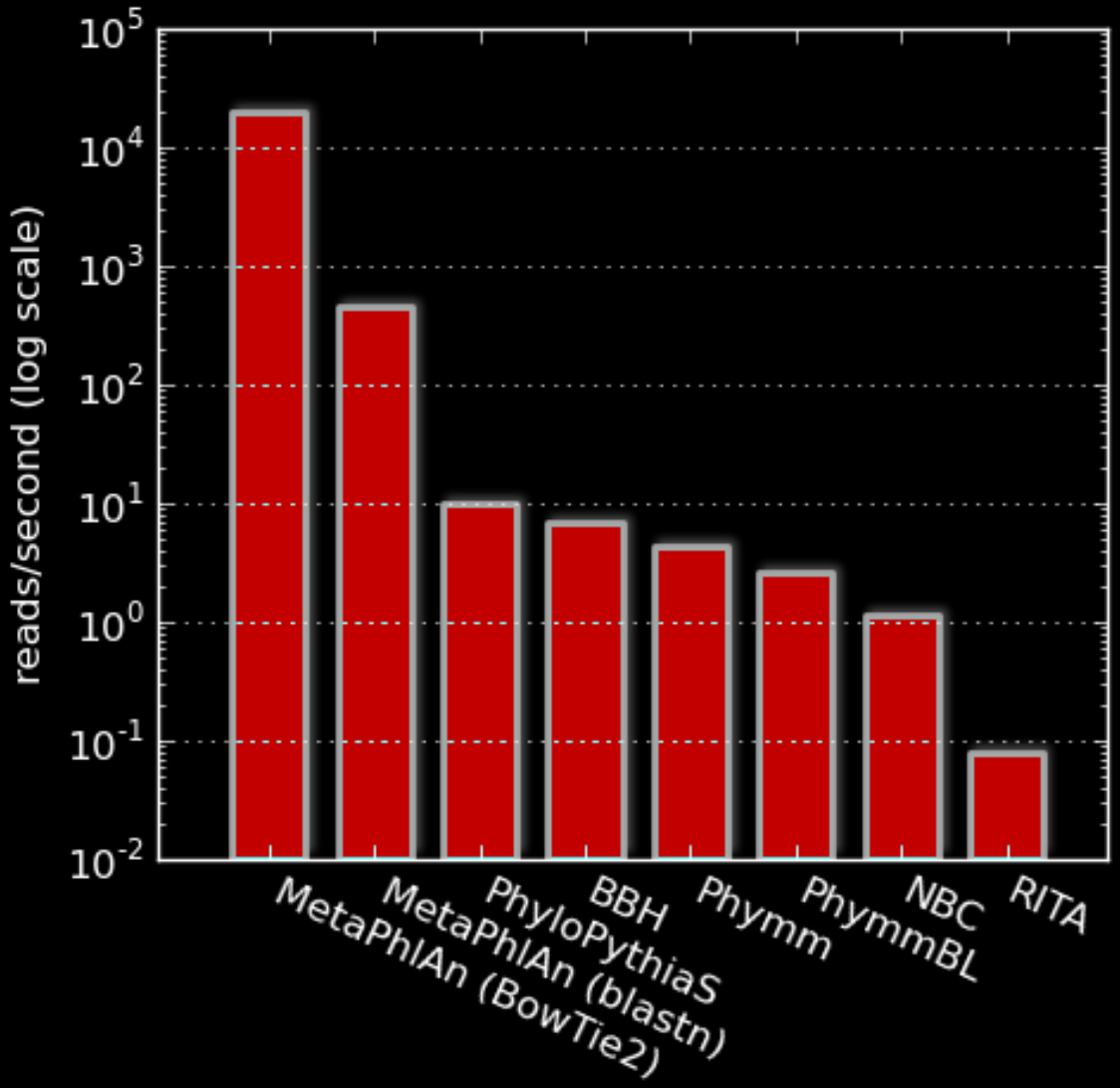
MetaPhlAn2: Taxonomic profiling using unique marker gene sequences

4Gnt synthetic metagenome, 125 organisms





Evaluation of MetaPhlAn performance



>50 times faster than earlier methods

450 reads/sec (BLAST)

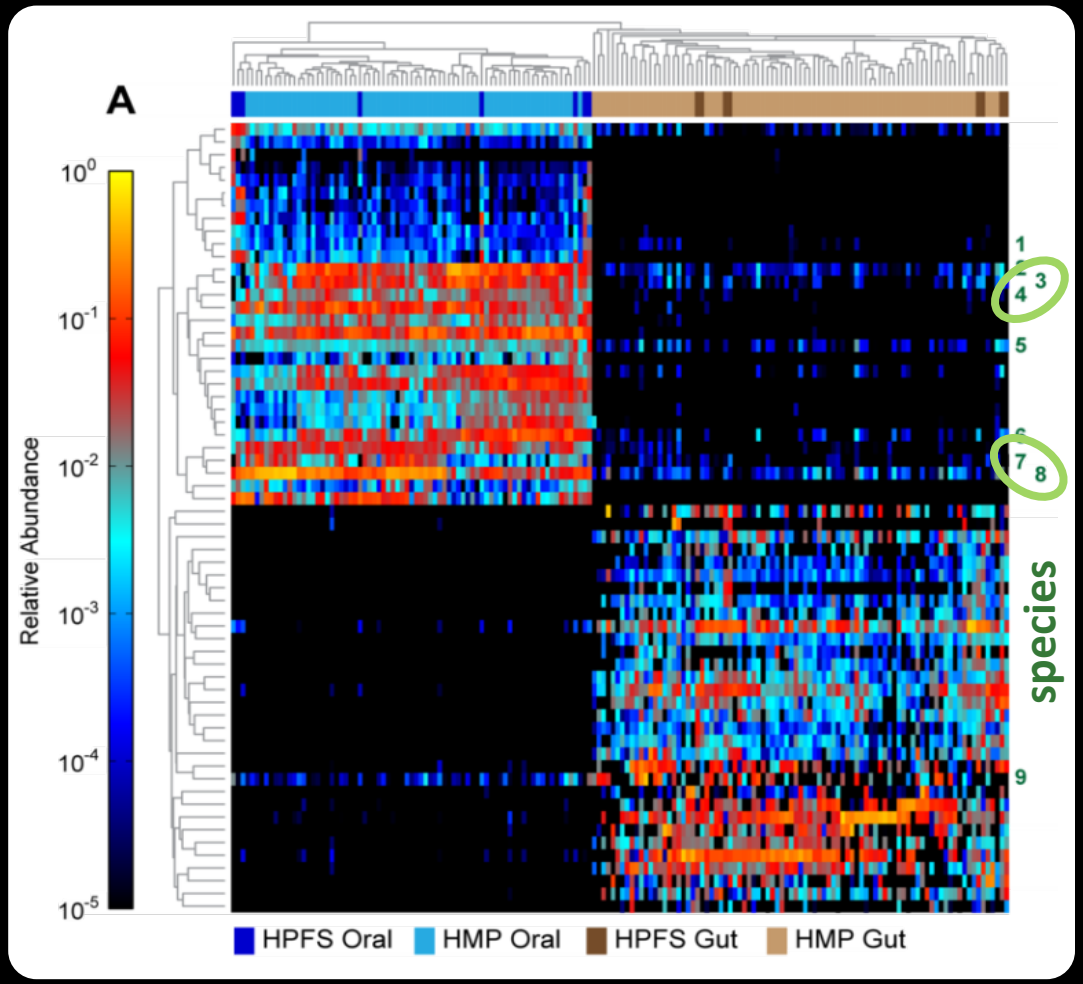
Up to 25,000 reads/sec (bowtie2)

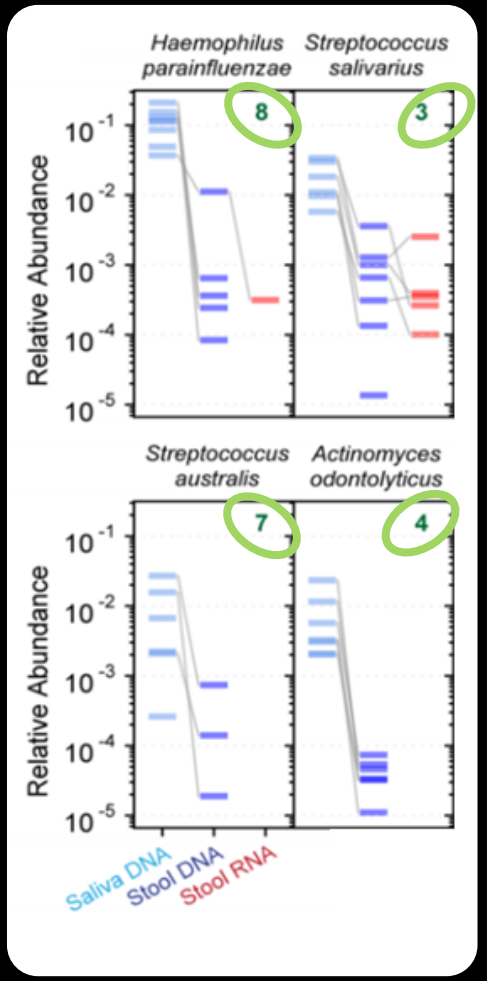
Multi-threaded

Easily parallelizable

MetaPhlAn in action

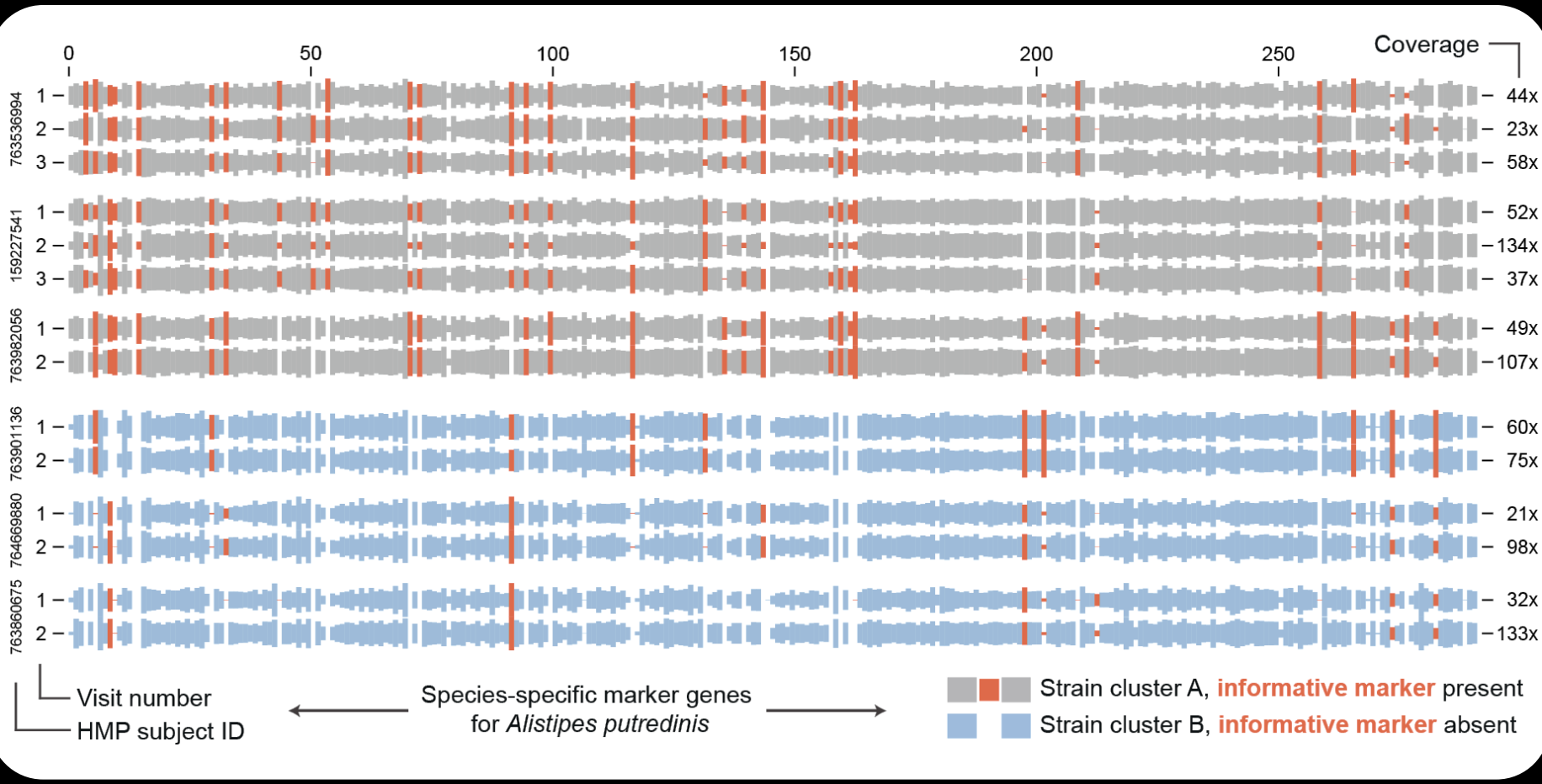
Eric Franzosa





Franzosa et al. *PNAS* 11:E2329-38 (2014)

MetaPhlAn in action: *strain profiling*



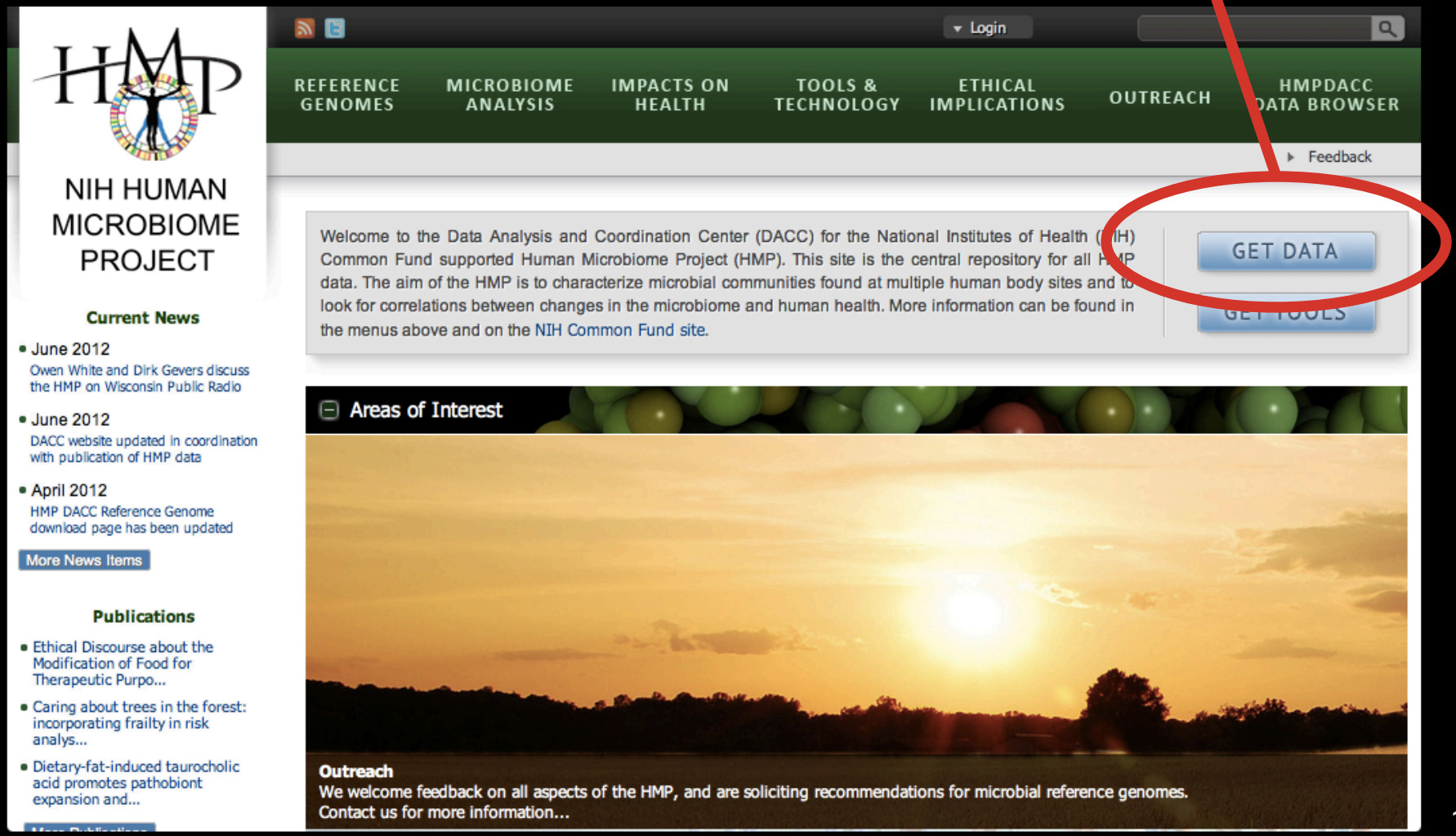
- In practice, not all markers are present
- Individual-specific marker “barcodes”
- Often very stable over time

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`**.

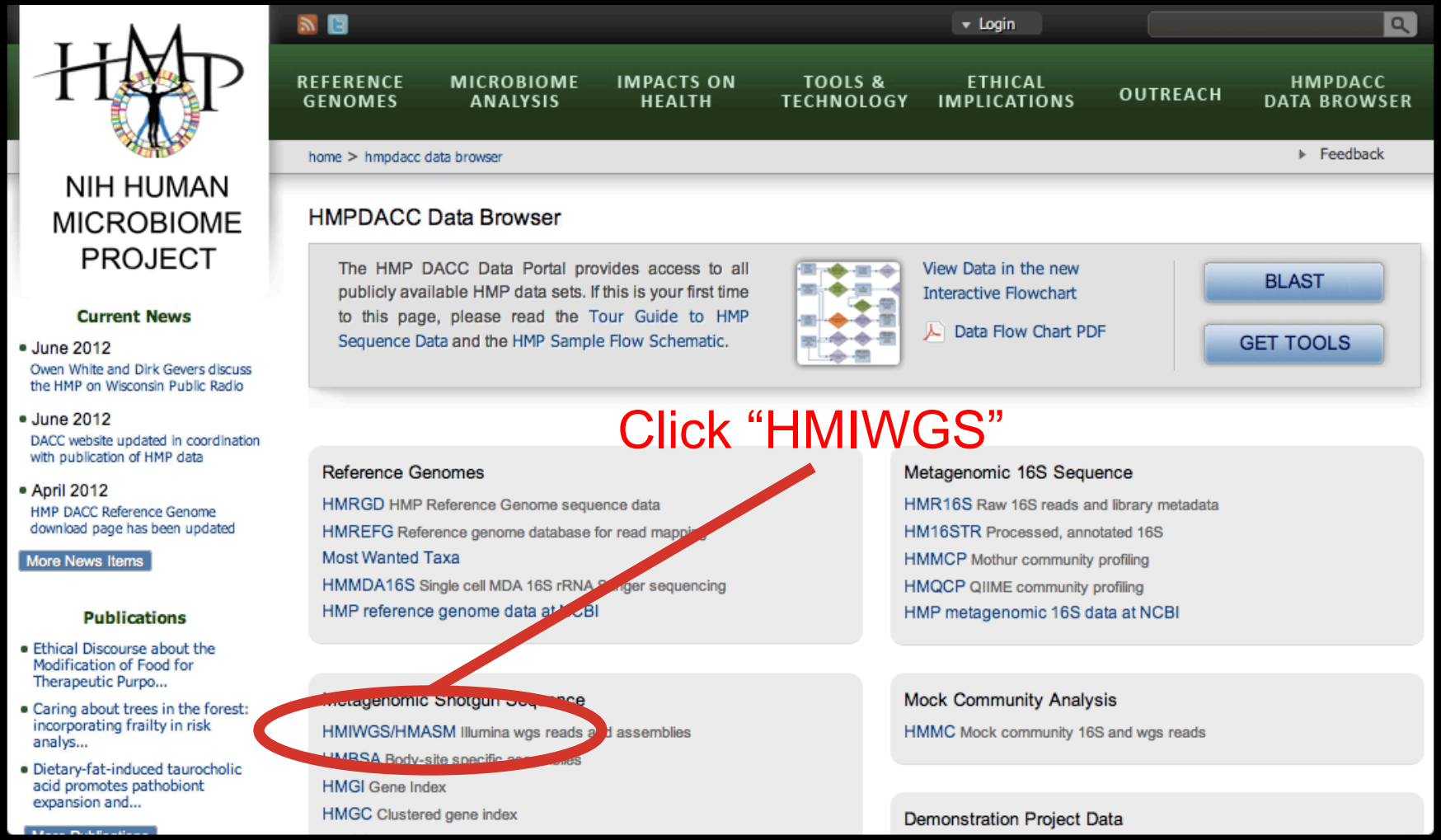
Getting some HMP data

- Go to http://hmpdacc.org



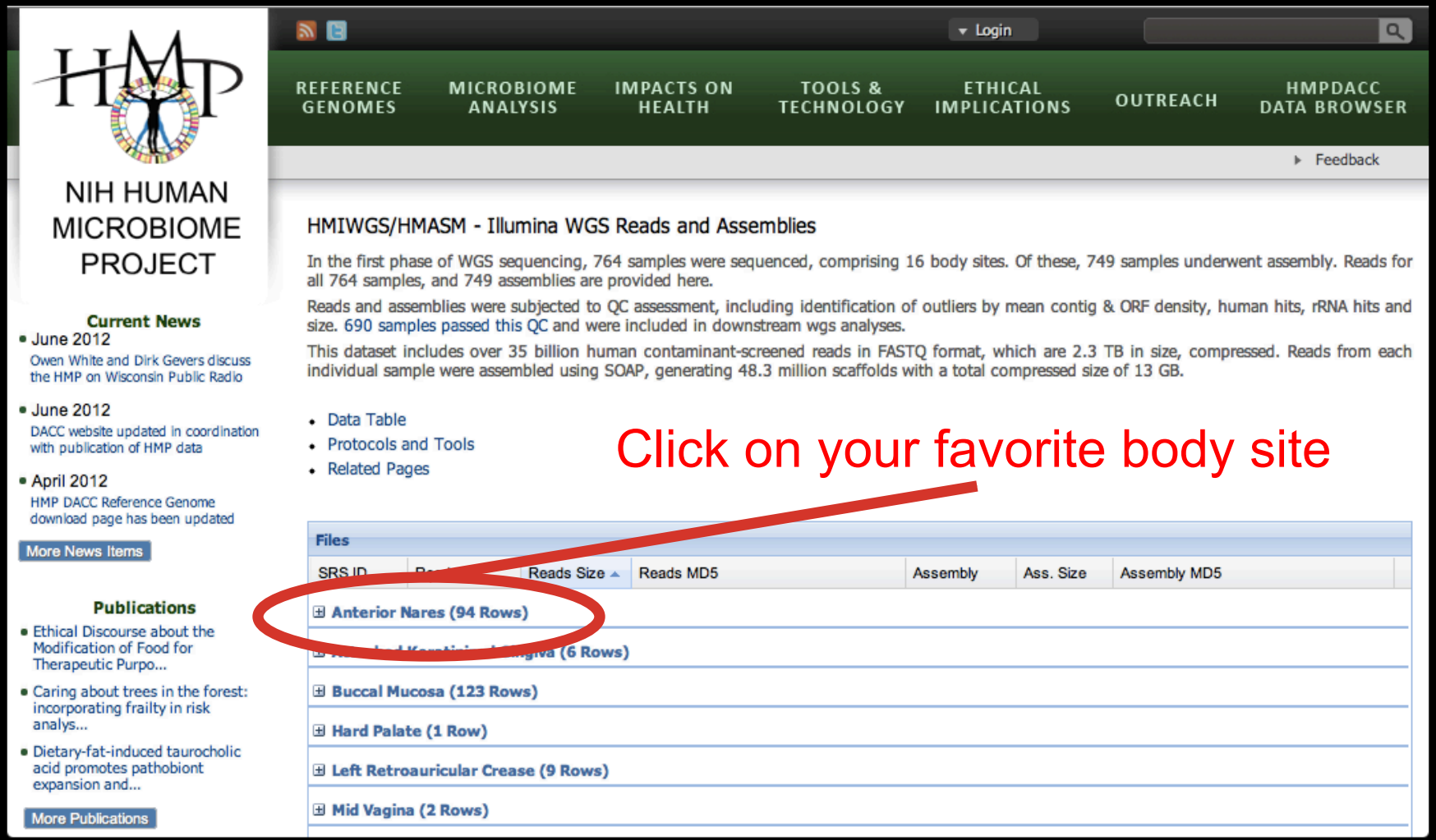
Getting some HMP data

- Check out what's available



Getting some HMP data

- Check out what's available



Getting some HMP data

Don't click on anything!

- Check out what's available

April 2012

HMP DACC Reference Genome download page has been updated

More News Items

Publications

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

More Publications

Data Resources

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

Files

SRS ID	Reads	Reads Size	Reads MD5	Assembly	Ass. Size	Assembly MD5
Anterior Nares (94 Rows)						
SRS047708	FTP	1.7 MB	d786590ff7fec20e8967127991766029	FTP	1.3 KB	ed98eda02d80a137c52b6fa8a3c57833
SRS019215	FTP	10.1 MB	55de248bbfa8c1bbf4447d007330f7ff	FTP	12.1 KB	cab8918433280eafc3d8f6ad78dc1ff7
SRS063178	FTP	13.1 MB	336f0b31b92880224c91ad52c4784adc	FTP	10.7 KB	99de257f1942e98bf1c052e2d046df33
SRS065179	FTP	13.3 MB	27b2c9209bc56cbe219d8c65fa32296c	FTP	54.6 KB	bb8b0d62a3c1923abfcaea01a598a60a
SRS065142	FTP	13.5 MB	3b05d6fcb205106fbd03f314e39f6d63	FTP	7.6 KB	91177065cf438056f2bfc67e99562fe4
SRS018585	FTP	16.8 MB	9d4129d2f5fdd51b9fc899bd84c47b5b	FTP	7.9 KB	aa9e9857b26b9efb4fa39bfaf101dc9d
SRS015640	FTP	17.6 MB	595baf36d8b3dcdd21149b3086ccbbee	FTP	52.4 KB	1c7a464db2fccce17c02f9600c867cb1
SRS056210	FTP	18.1 MB	9b2f74b8067e6f20551e6d3b48124c42	FTP	18.3 KB	c4abace0ec0b3e7e5ce1513cb8270e56
SRS018312	FTP	18.9 MB	2454e80d7e5216adf8d5b1850c98738c	FTP	25.4 KB	4f5f760eadd77782862669263e1b1d9d
SRS015450	FTP	18.9 MB	eefc0dcf2d52ca5251b01860d54d2bb5	FTP	107.1 KB	4e0a83868f2fb44f1788dfe1aaa5e13f
SRS049744	FTP	21.5 MB	6d9e2ffc82b08ef37551e902096e4c98	FTP	14.3 KB	da7a1cddd3c84b121ff49086432d25d3
SRS012291	FTP	21.9 MB	12775f5df6e71961f1c544e84f6c7342	FTP	8.9 KB	17b5110d391817c7ce52b7c1026df1ba
SRS051600	FTP	22.2 MB	391775b95926a221b8a3cde54a79ae22	FTP	13.9 KB	6db7007edd32b534bc918aad42d600ae
SRS019339	FTP	23.1 MB	76a621d6503d11d1a133a023dc240ae5	FTP	57.3 KB	9255d8206f10ac2611cf45270daa166c
SRS017244	FTP	23.5 MB	b7c2dec67738f317cb8826c09e1a9e39	FTP	21.3 KB	9bcf59e6b4fe15a4e8ccacb0bc824ba8
SRS018671	FTP	24.0 MB	7548b06b37038440c5420f7677ff7371	FTP	135.4 KB	4a180e3ea42a46bcea0a9441b137f243

Show All | Save As CSV File

Protocols and Tools

Getting some (prepped) HMP data

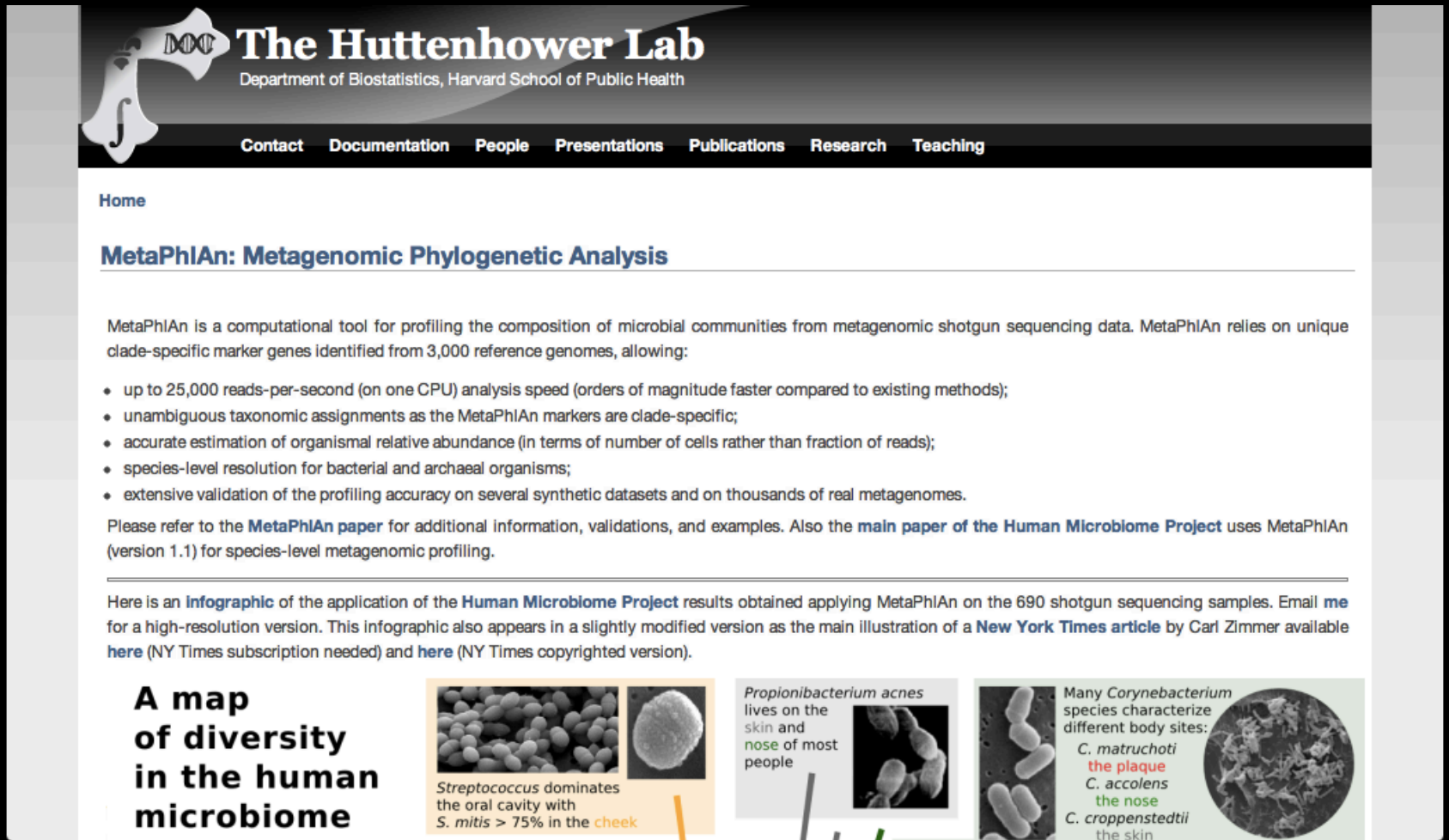
- Connect to the server instead
  - `cd` to your favorite directory and run:

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

- These are subsamples of six HMP files:
  - SRS014459.tar.bz2 ➔ 763577454-SRS014459-Stool.fasta
  - SRS014464.tar.bz2 ➔ 763577454-SRS014464-Anterior_nares.fasta
  - SRS014470.tar.bz2 ➔ 763577454-SRS014470-Tongue_dorsum.fasta
  - SRS014472.tar.bz2 ➔ 763577454-SRS014472-Buccal_mucosa.fasta
  - SRS014476.tar.bz2 ➔ 763577454-SRS014476-Supragingival_plaque.fasta
  - SRS014494.tar.bz2 ➔ 763577454-SRS014494-Posterior_fornix.fasta
- All six shotgunned body sites from
  - One subject, first visit
  - Subsampled to 20,000 reads

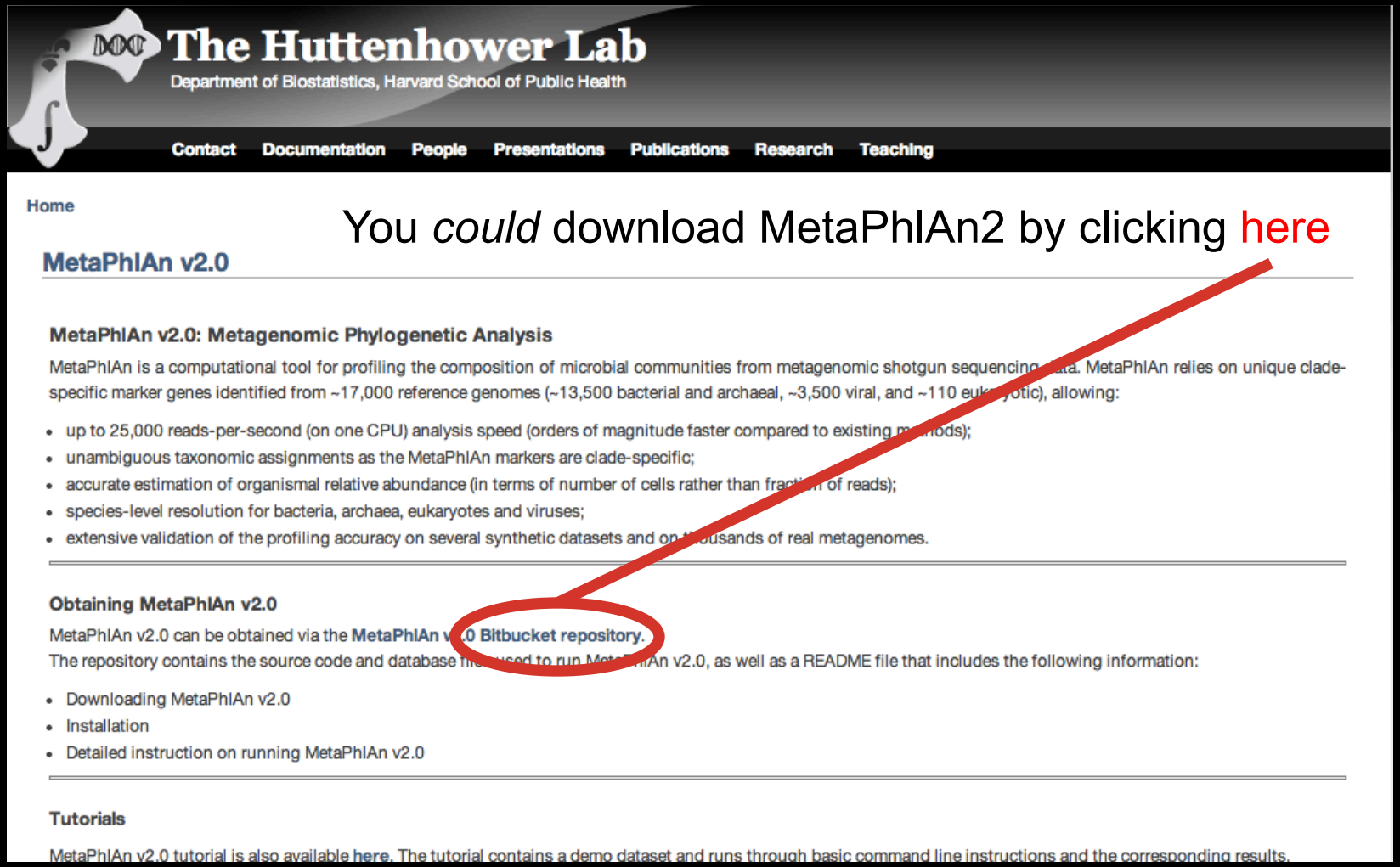
Who's there: MetaPhlAn1

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



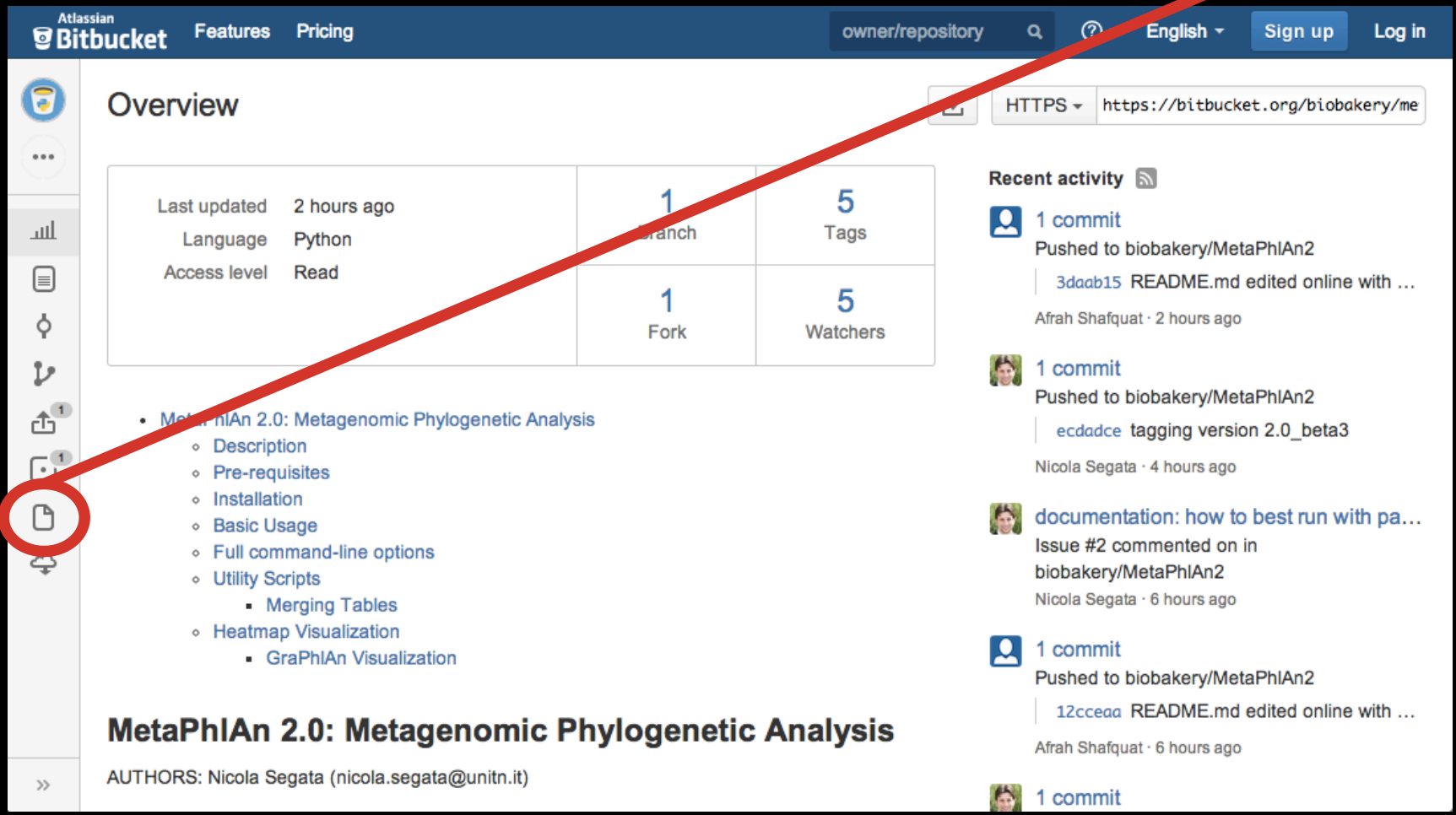
Who's there: MetaPhlAn2

- Instead, go to http://huttenhower.sph.harvard.edu/metaphlan2



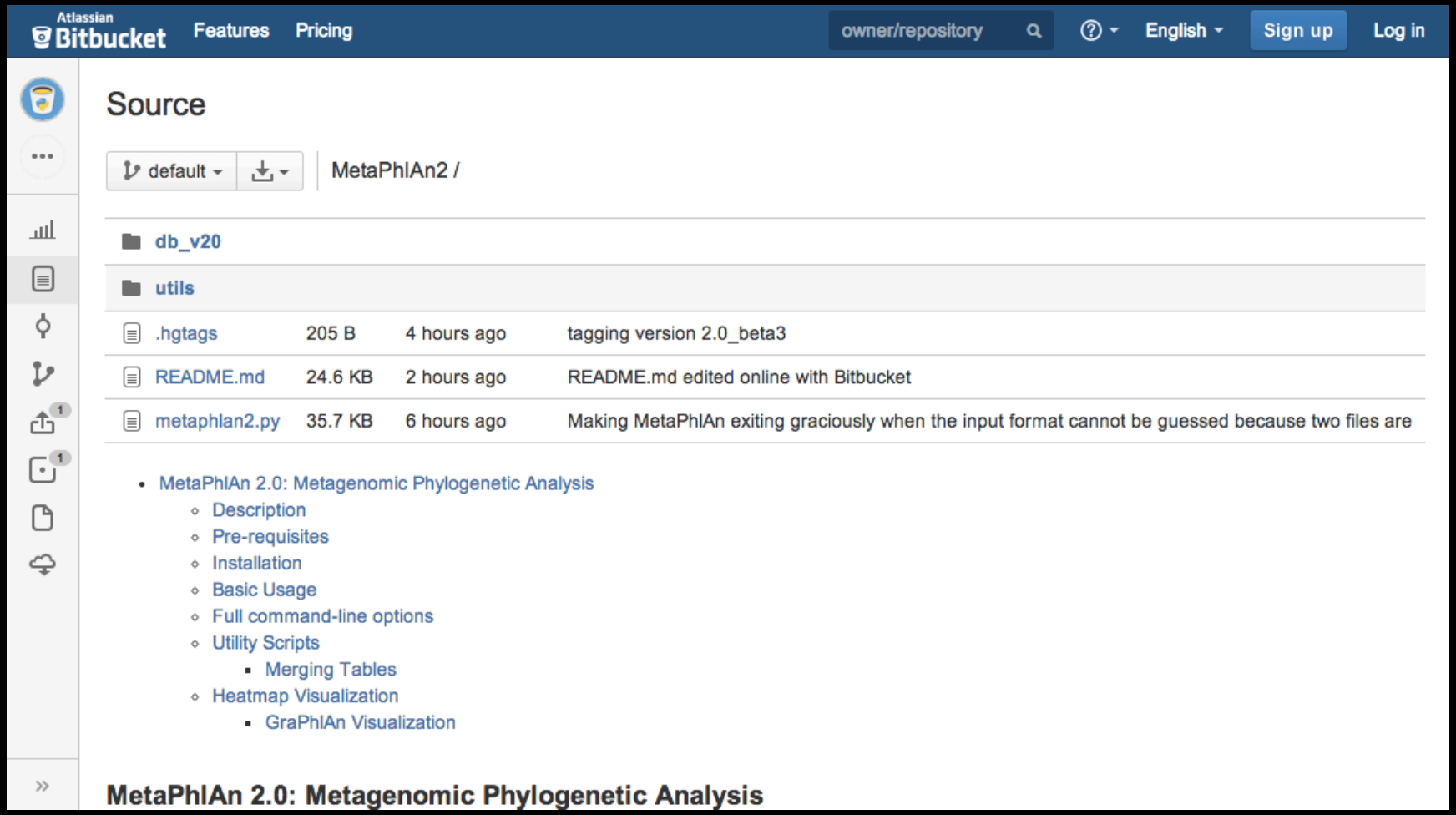
Who's there: MetaPhlAn2

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



Who's there: MetaPhlAn2

- The complete MetaPhlAn2 install is in `~/workshop_data/metagenomics/biobakery/software/metaphlan2`



From the command line...

- To see what you can do, run:

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

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

Who's there: MetaPhlAn2

1. ssh

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

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

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

COMMON COMMANDS
:
```

Who's there: MetaPhlAn2

- To launch your first analysis, run:

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

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

Who's there: MetaPhlAn

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

1. [screen 2: bash] chuttenhower@class:~/tmp (ssh)

```
HWUSI-EAS1625_615HE:4:100:0:1248/1  gi|479140210|ref|NC_021010.1|:1043207-1044529
HWUSI-EAS1625_615HE:4:100:0:1301/1  gi|483877978|ref|NZ_KB890364.1|:31018-31902
HWUSI-EAS1625_615HE:4:100:1000:167/1    gi|242362078|ref|NZ_GG692716.1|:28261-29169
HWUSI-EAS1625_615HE:4:100:1001:1264/1   gi|270295698|ref|NZ_GG730107.1|:470181-472532
HWUSI-EAS1625_615HE:4:100:1001:1320/1   gi|224993849|ref|NZ_ACFY01000158.1|:c1296-10
HWUSI-EAS1625_615HE:4:100:1001:1604/1   gi|319644663|ref|NZ_GL635657.1|:c320982-320029
HWUSI-EAS1625_615HE:4:100:1001:1734/1   gi|484001485|ref|NZ_KB894131.1|:91019-91717
HWUSI-EAS1625_615HE:4:100:1001:259/1    gi|479210985|ref|NC_021043.1|:c1165057-1164158
HWUSI-EAS1625_615HE:4:100:1002:1501/1   gi|224485637|ref|NZ_EQ973491.1|:c620672-618312
HWUSI-EAS1625_615HE:4:100:1003:1644/1   gi|224485636|ref|NZ_EQ973490.1|:c204903-202990
HWUSI-EAS1625_615HE:4:100:1003:1702/1   gi|423335209|ref|NZ_JH976498.1|:329186-330046
HWUSI-EAS1625_615HE:4:100:1003:2030/1   gi|238922432|ref|NC_012781.1|:2910912-2912072
HWUSI-EAS1625_615HE:4:100:1004:353/1    gi|223955873|ref|NZ_DS499674.1|:c266282-265248
HWUSI-EAS1625_615HE:4:100:1004:742/1    gi|283767237|ref|NZ_GG730311.1|:c124395-124171
HWUSI-EAS1625_615HE:4:100:1005:1722/1   gi|410105720|ref|NZ_JH976502.1|:750498-751148
HWUSI-EAS1625_615HE:4:100:1005:505/1    gi|479170689|ref|NC_021020.1|:1540599-1542305
HWUSI-EAS1625_615HE:4:100:1006:848/1    gi|347530298|ref|NC_015977.1|:c3433030-3431387
HWUSI-EAS1625_615HE:4:100:1007:1428/1   gi|423332908|ref|NZ_JH976496.1|:1485161-1487113
HWUSI-EAS1625_615HE:4:100:1007:1465/1   gi|423332908|ref|NZ_JH976496.1|:906255-909584
HWUSI-EAS1625_615HE:4:100:1008:1187/1   gi|224485479|ref|NZ_EQ973214.1|:108053-108250
HWUSI-EAS1625_615HE:4:100:1008:1241/1   gi|270293478|ref|NZ_GG730105.1|:c830784-828727
HWUSI-EAS1625_615HE:4:100:1008:140/1    gi|224514921|ref|NZ_DS499545.1|:41991-42827
HWUSI-EAS1625_615HE:4:100:1009:154/1    gi|301307949|ref|NZ_GG774972.1|:644845-649113
HWUSI-EAS1625_615HE:4:100:1009:467/1    gi|303257489|ref|NZ_GL383997.1|:67163-67873
:
```

Who's there: MetaPhlAn

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

1. [screen 2: bash] chuttenhower@class:~/tmp (ssh)

```
k__Bacteria 100.0
k__Bacteria|p__Firmicutes   64.82041
k__Bacteria|p__Bacteroidetes    35.17959
k__Bacteria|p__Firmicutes|c__Clostridia 64.82041
k__Bacteria|p__Bacteroidetes|c__Bacteroidia 35.17959
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales    64.82041
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales    35.17959
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Ruminococcaceae 37.71449
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Bacteroidaceae  31.5000
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Eubacteriaceae  21.99035
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Lachnospiraceae 5.11557
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Porphyromonadaceae  3.6
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Ruminococcaceae|g__Subdolig
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Bacteroidaceae|g__Bacte
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Eubacteriaceae|g__Eubacteri
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Lachnospiraceae|g__Roseburi
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Porphyromonadaceae|g__P
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Ruminococcaceae|g__Subdolig
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Eubacteriaceae|g__Eubacteri
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Bacteroidaceae|g__Bacte
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Bacteroidaceae|g__Bacte
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Eubacteriaceae|g__Eubacteri
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Lachnospiraceae|g__Roseburi
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Bacteroidaceae|g__Bacte
763577454-SRS014459-Stool.txt
```

Who's there: MetaPhlAn2

- You can finish the job if you like:

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

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

- Or you can copy the rest pre-calculated:

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

Who's there: MetaPhlAn2

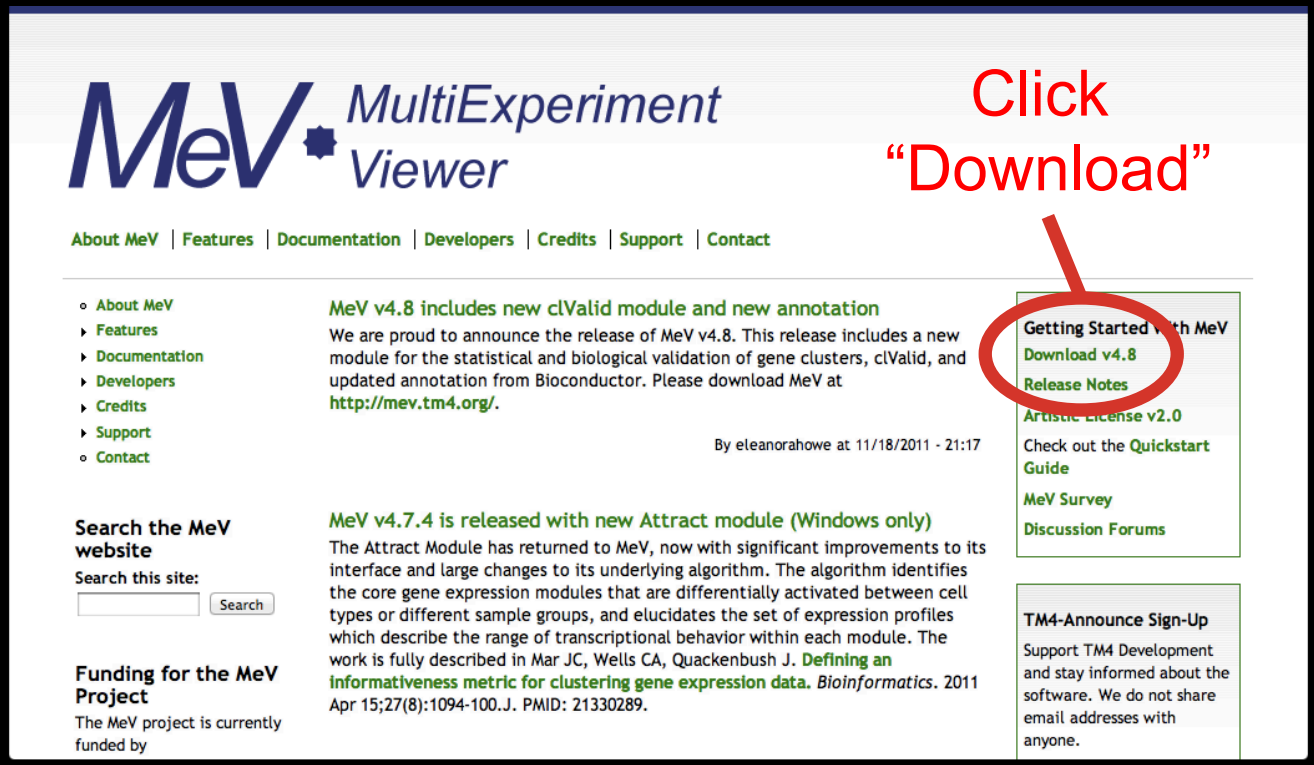
- Let's make a single table containing all six samples:

```
mkdir tmp
mv *.bowtie2out.txt tmp
~/workshop_data/metagenomics/biobakery/software/metaphlan2/utils/merge_metaphlan_tables.py *.txt >
  763577454.tsv
```

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

Who's there: MetaPhlAn

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

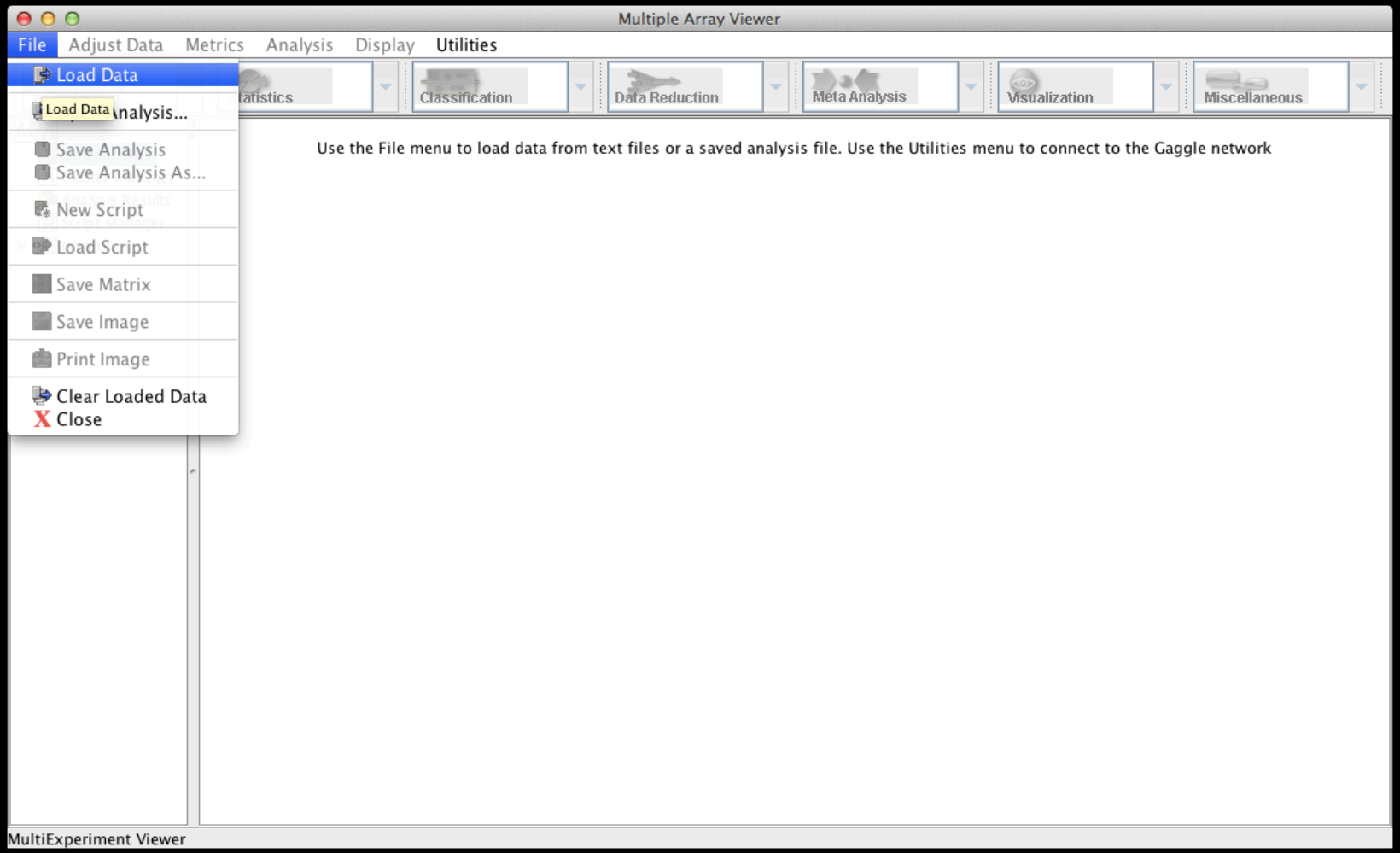


- Or use the pre-installed version:

```
tmev.sh
```

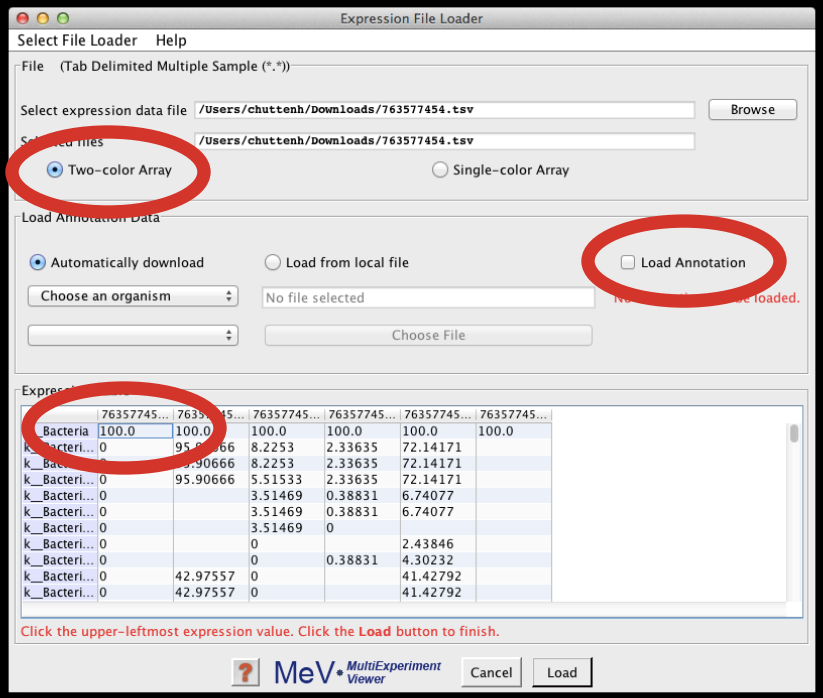
An interlude: MeV

- Launch MeV, and select File/Load data



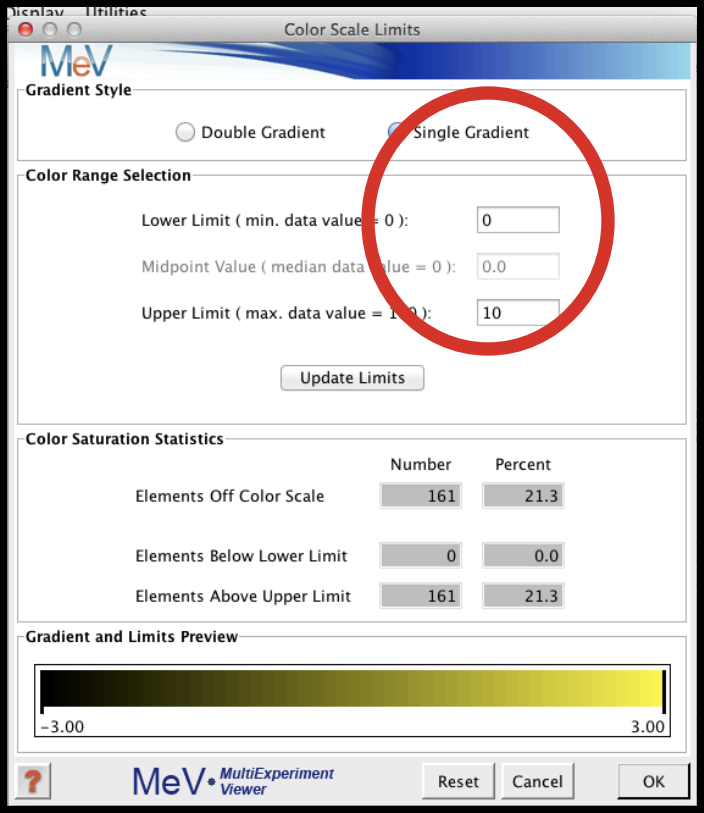
An interlude: MeV

- Click “Browse” to your TSV file (`763577454.tsv`), then
  - Tell MeV it’s a two-color array
  - Uncheck “Load annotation”
  - Click on the upper-leftmost ***data*** value



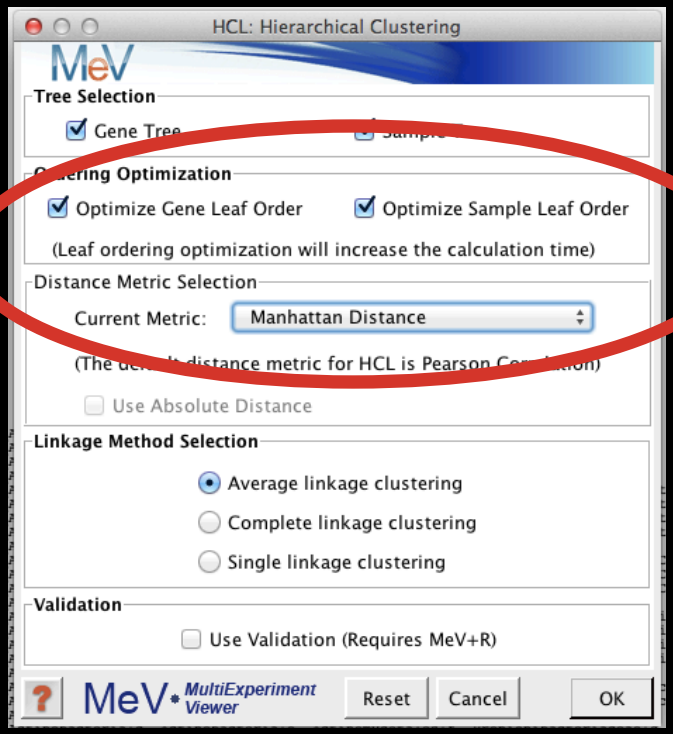
An interlude: MeV

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



An interlude: MeV

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



An interlude: MeV

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

An interlude: MeV

- MeV is a tool; imperfect, but convenient
  - You should likely include just “leaf” nodes
    - Species, whose names start include “s__”
    - You can filter your file using:

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

  - You can, but might not want to, z-score normalize
    - Adjust Data/Gene-Row Adjustments/Normalize Genes-Rows

- Many other tools built in – experiment!

The two big questions…

Who is there?

What are they doing?

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

The ~~two~~ three big questions…

Who is there?

What are they doing?

What does it all mean?

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

Properties of microbiome data

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

Site	Oral	Gut	Oral	Gut	Oral	Gut
Clade1	0.40	0.87	0.43	0.68	0.47	0.32
Clade1|Bug1	0.40	0.56	0.07	0.31	0.42	0.27
Clade1|Bug2	0.00	0.30	0.36	0.37	0.04	0.05
Clade2	0.60	0.13	0.57	0.32	0.53	0.68
Clade2|Bug3	0.11	0.00	0.10	0.32	0.15	0.23
Clade2|Bug4	0.49	0.13	0.47	0.00	0.39	0.45

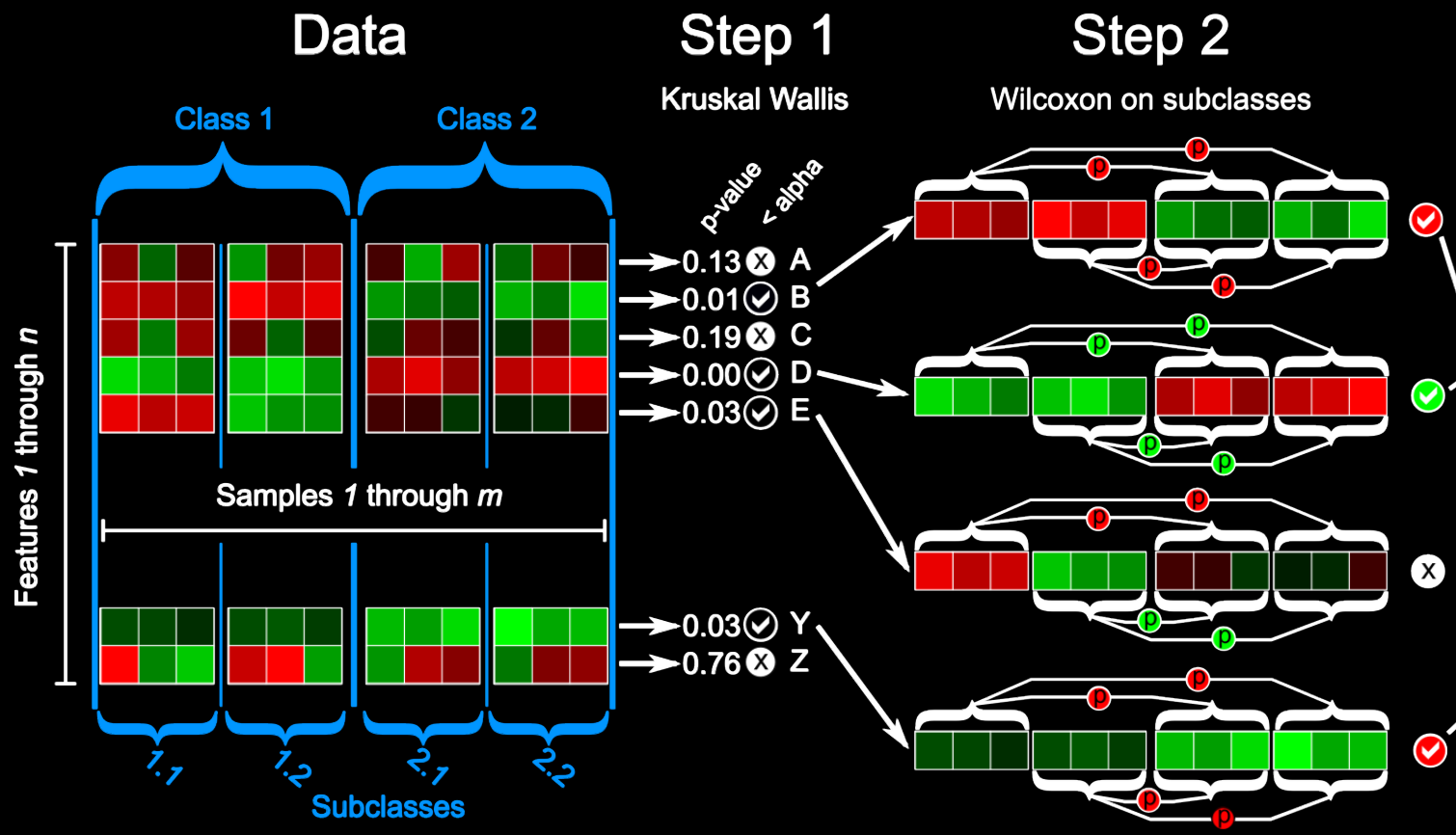
Properties of microbiome data

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

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

LEfSe: LDA Effect Size
Finding metagenomic biomarkers

Nicola Segata

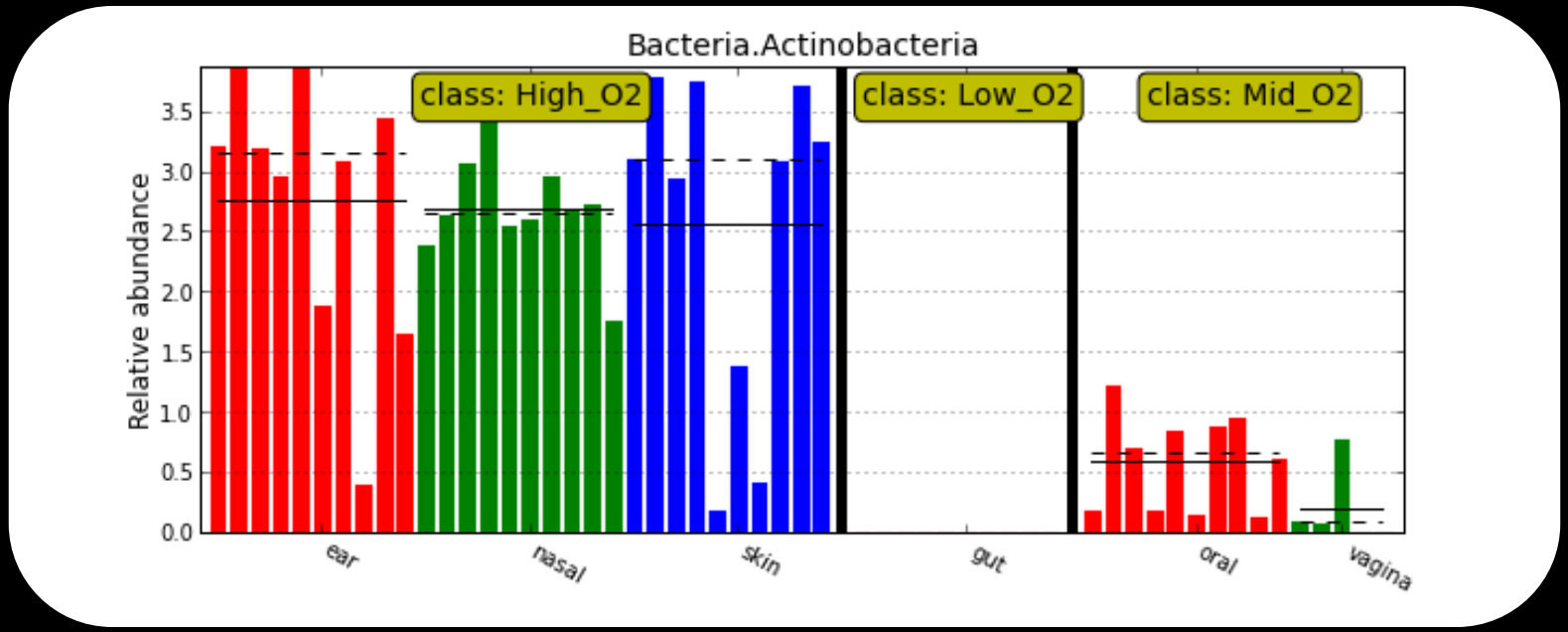


A
Statistical consistency

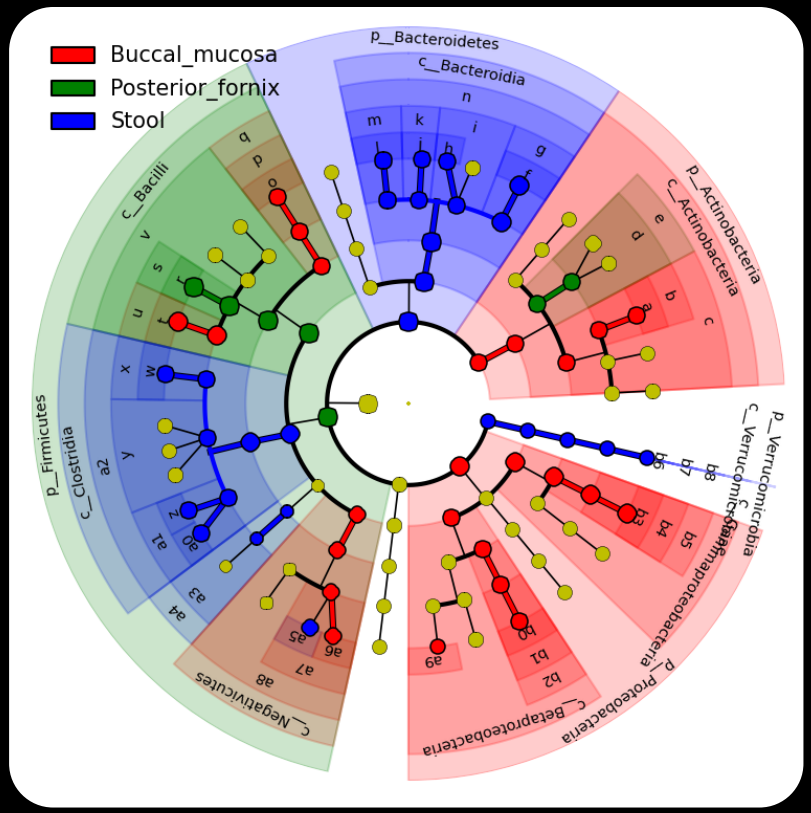
B
Biological consistency

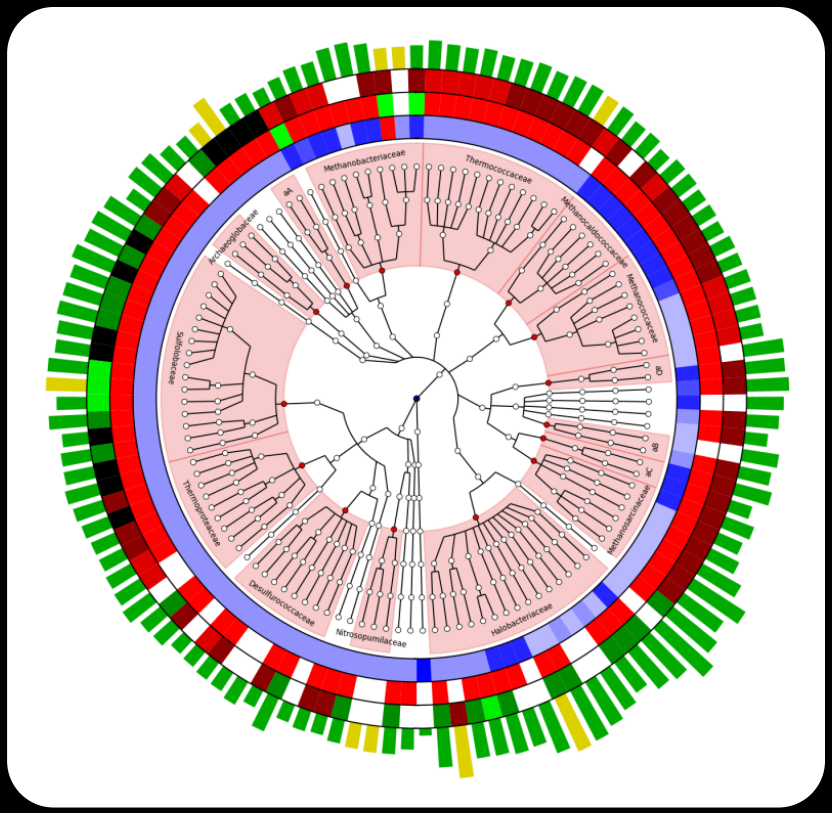
C
Biological Effect size

Example LEfSe application:
Find O_2-loving bugs (controlling for body site)



Superimpose enrichments on the tree of life using GraPhlAn



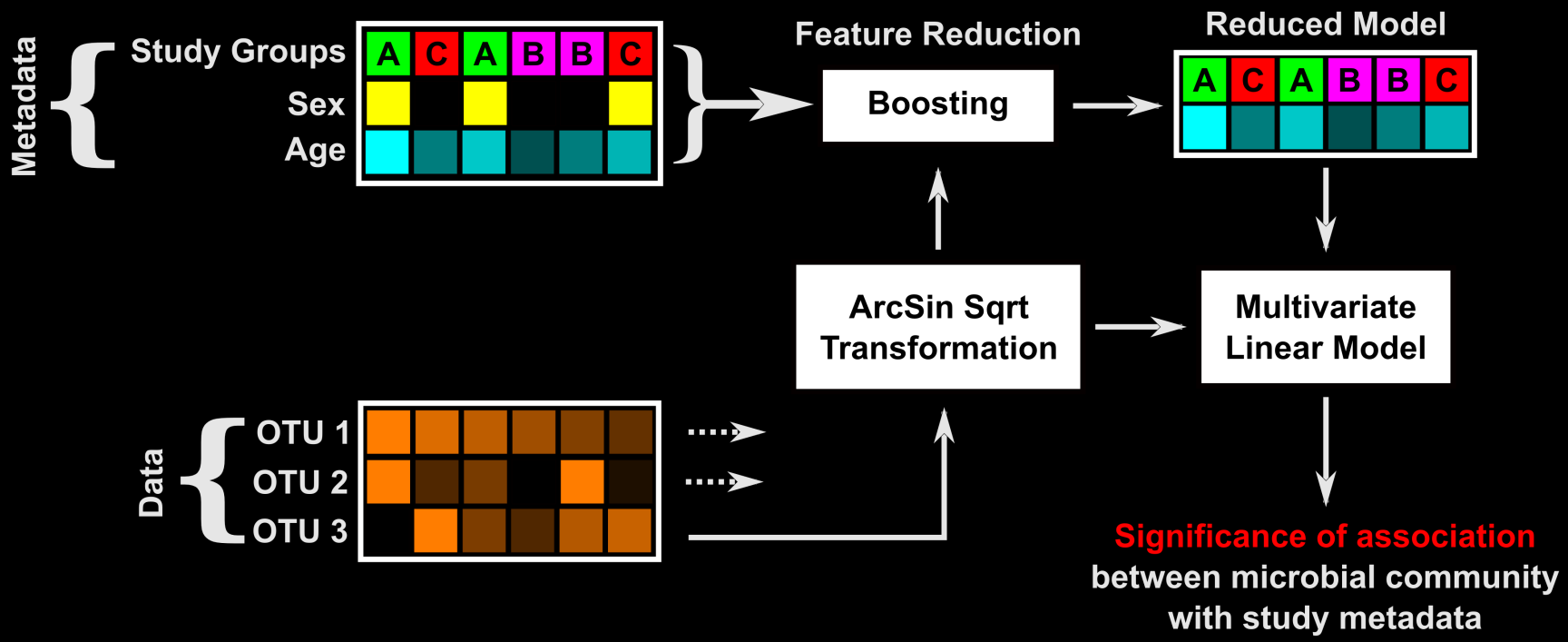


LEfSe Associations

Metadata Rings

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

MaAsLin
Multivariate microbial Association with Linear models



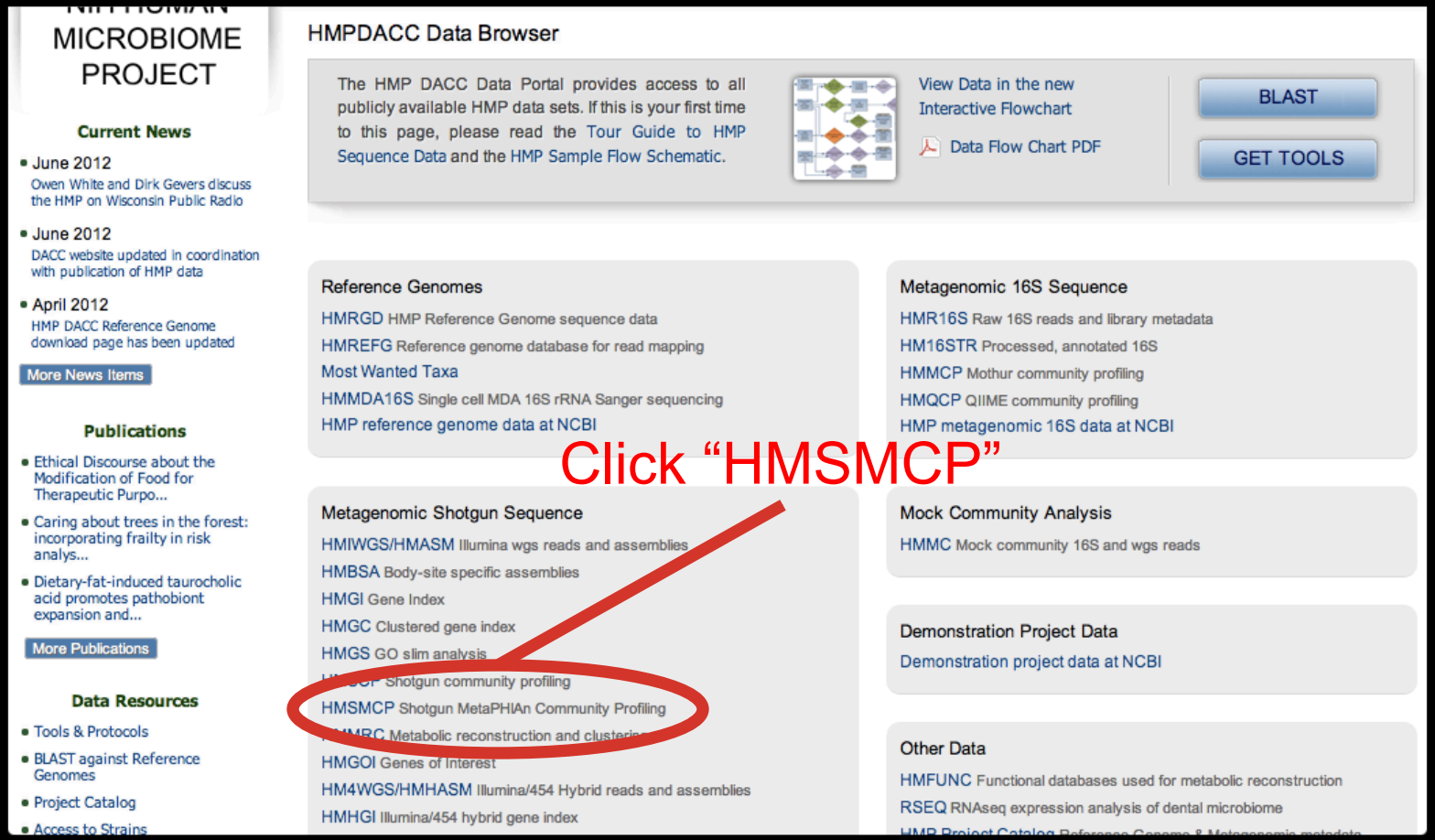
- A more general solution for finding significant metagenomic associations in metadata-rich studies

Tim Tickle

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

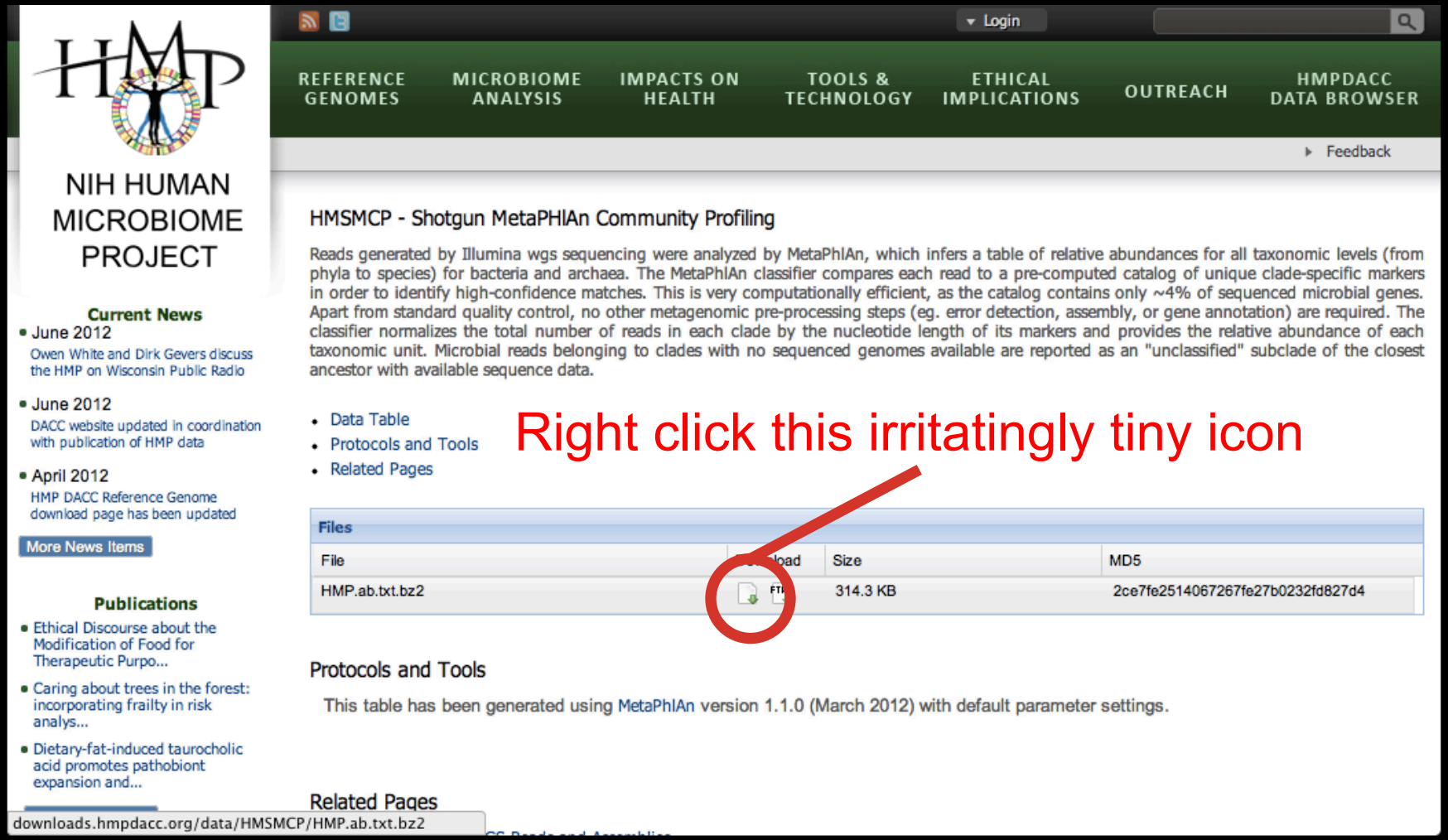
What it means: LEfSe

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



What it means: LEfSe

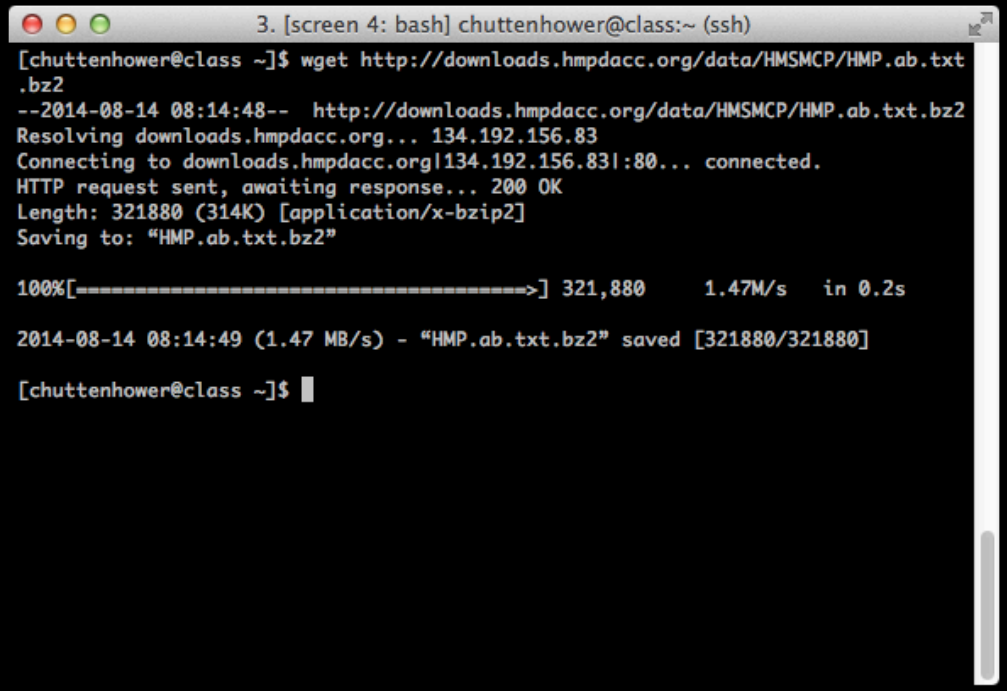
- Download the MetaPhlAn1 table for all 700 samples



Downloading from the command line

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

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



What it means: LEfSe

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

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

- Look at the result

```
less HMP.ab.txt
```

- **IMPORTANT!!!**
  - This file's too big to analyze directly today

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

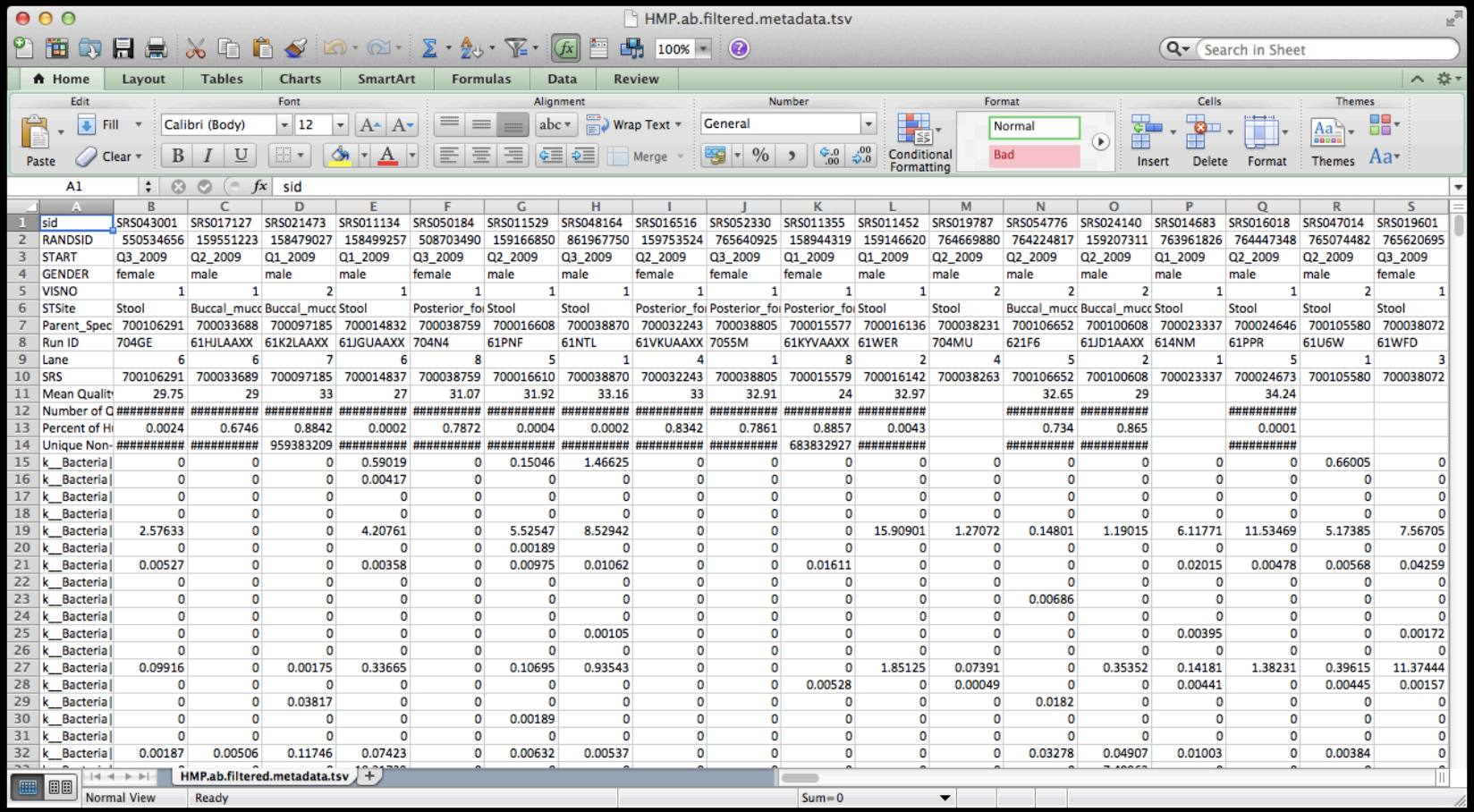
- This is great – tons of data, but no metadata
  - Scripts and data from HUMAnN to the rescue:

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

- NOW take a look again

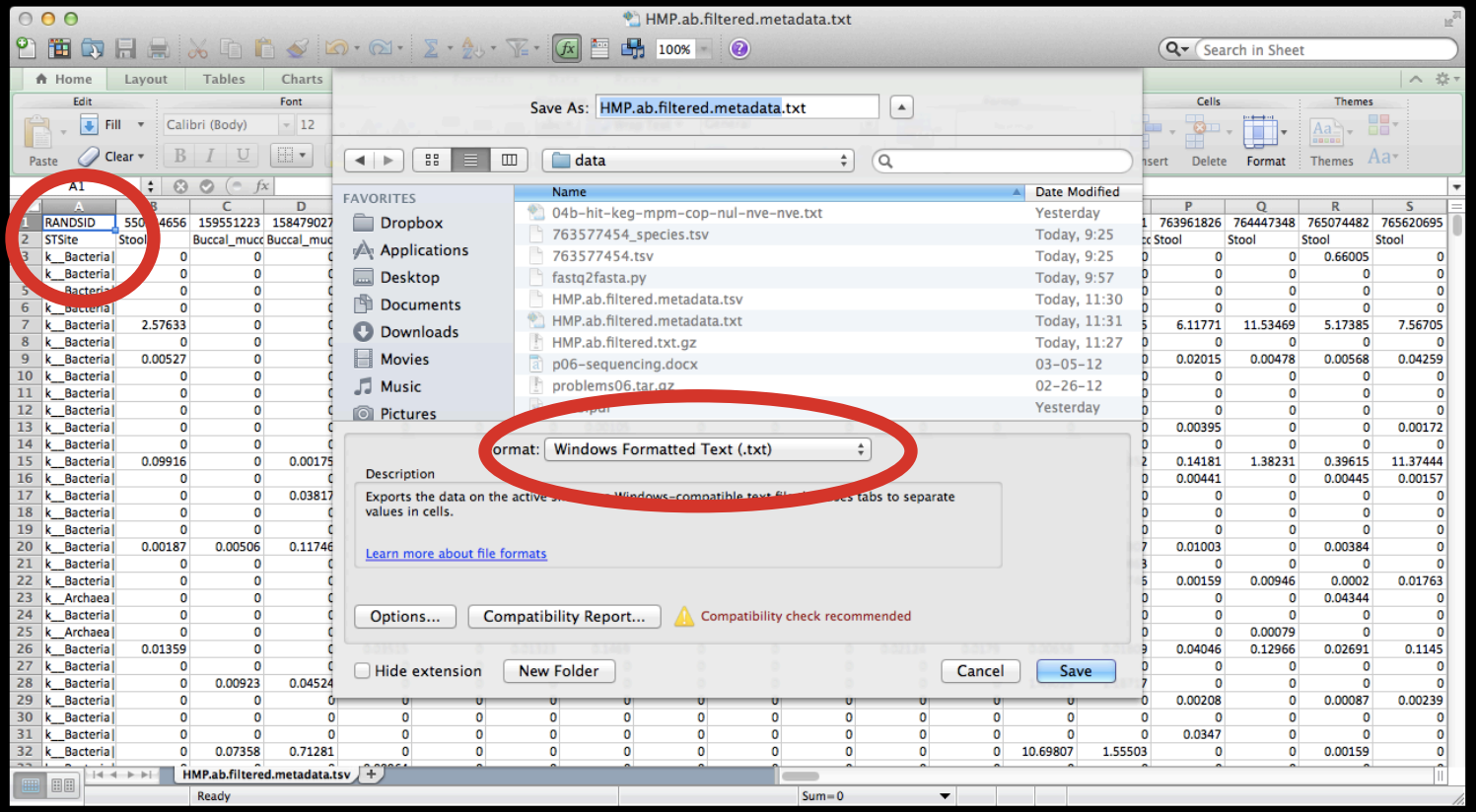
What it means: LEfSe

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



What it means: LEfSe

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



What it means: LEfSe

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

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

Tools

search tools

HUTTENHOWER LAB MODULES

LEfSe

A) Format Data for LEfSe

B) LDA Effect Size (LEfSe)

C) Plot LEfSe Results

D) Plot Cladogram

E) Plot One Feature

F) Plot Differential Features

MetaPhlAn

GraPhlAn

microPITA

MaAsLin

PICRUSt

First click here

Get Data

Upload File from your computer

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

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

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

If you have any comments, questions, or suggestions, please contact Dr. Huttenhower.

History

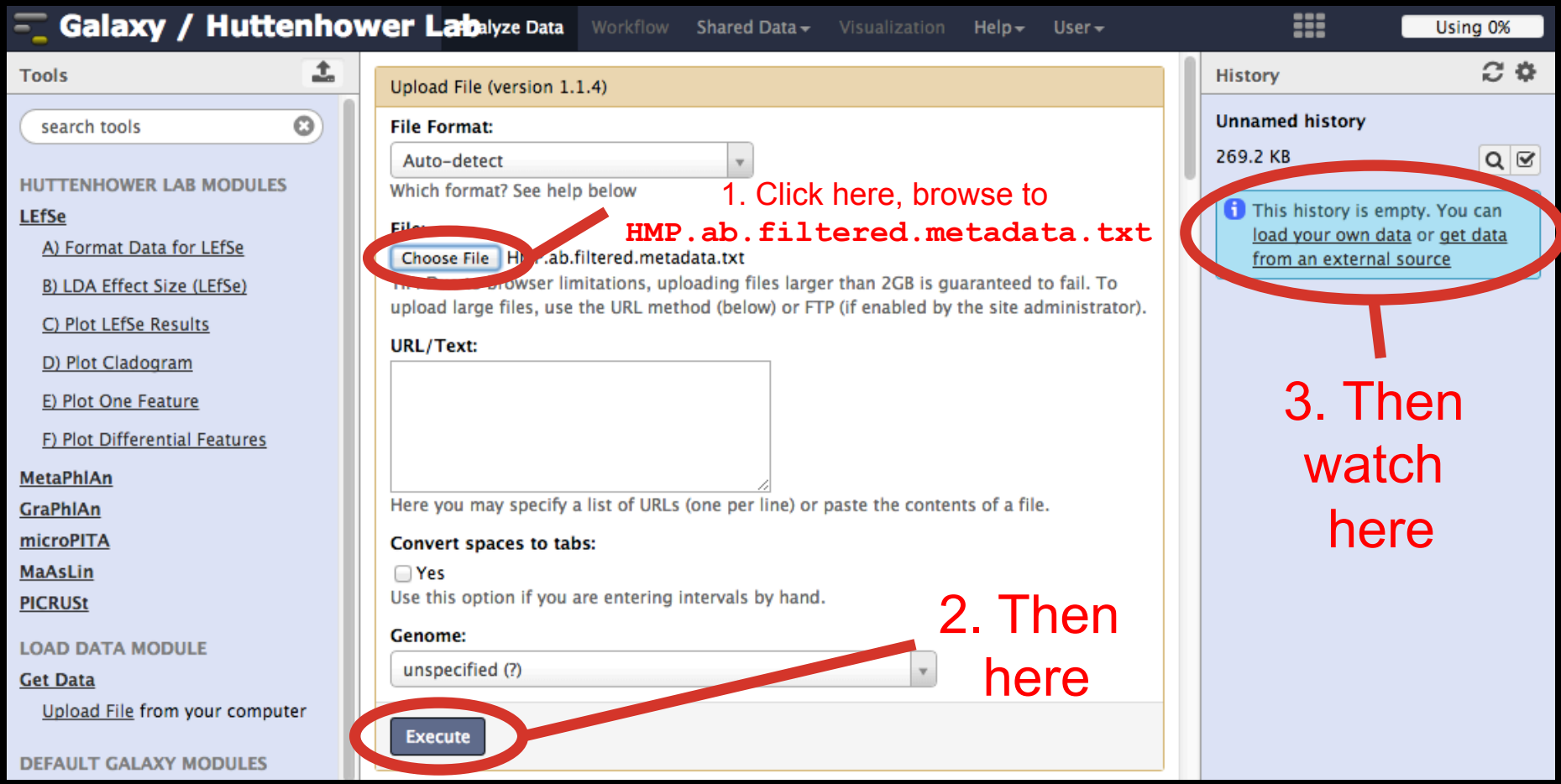
Unnamed history

0 bytes

This history is empty. You can load your own data or get data from an external source

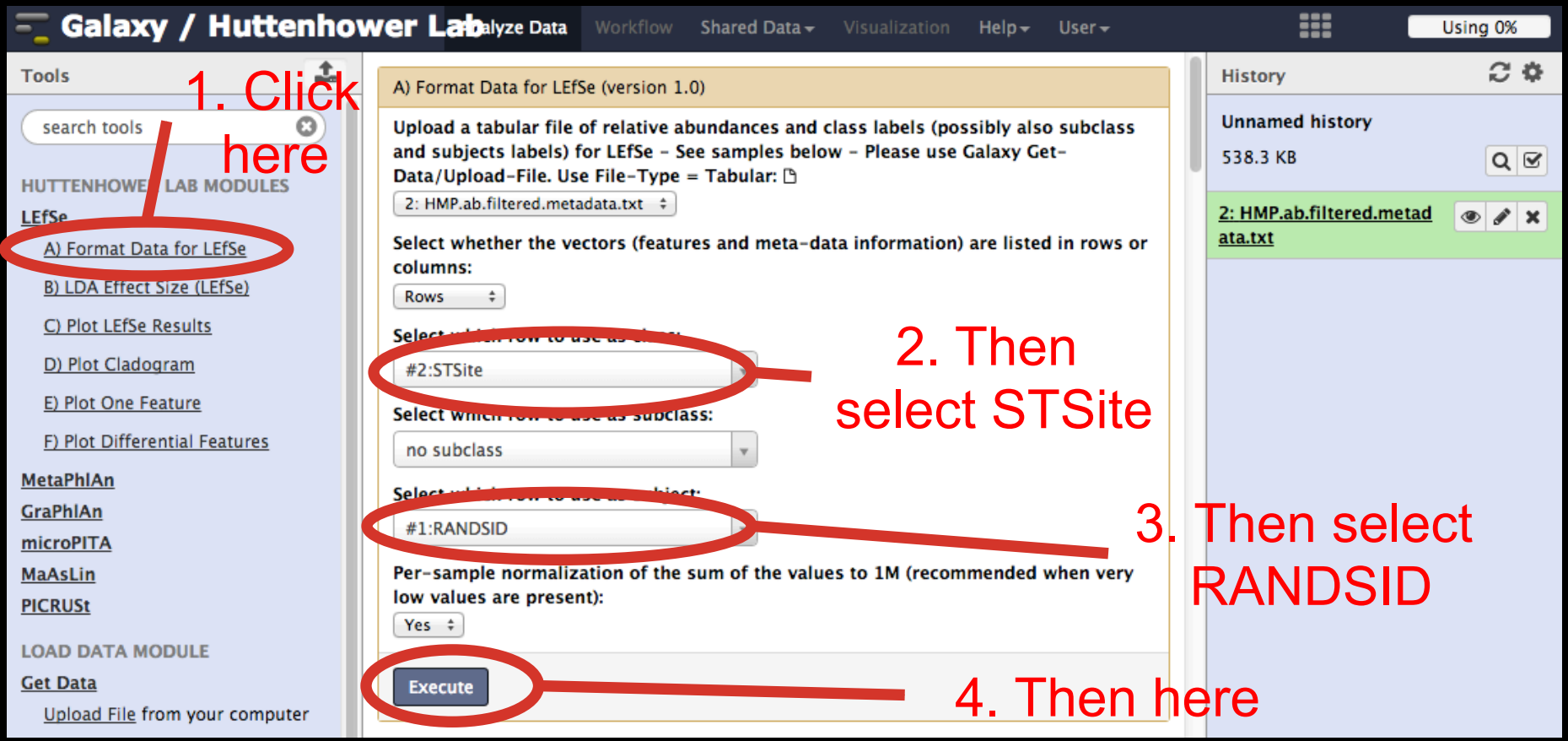
What it means: LEfSe

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



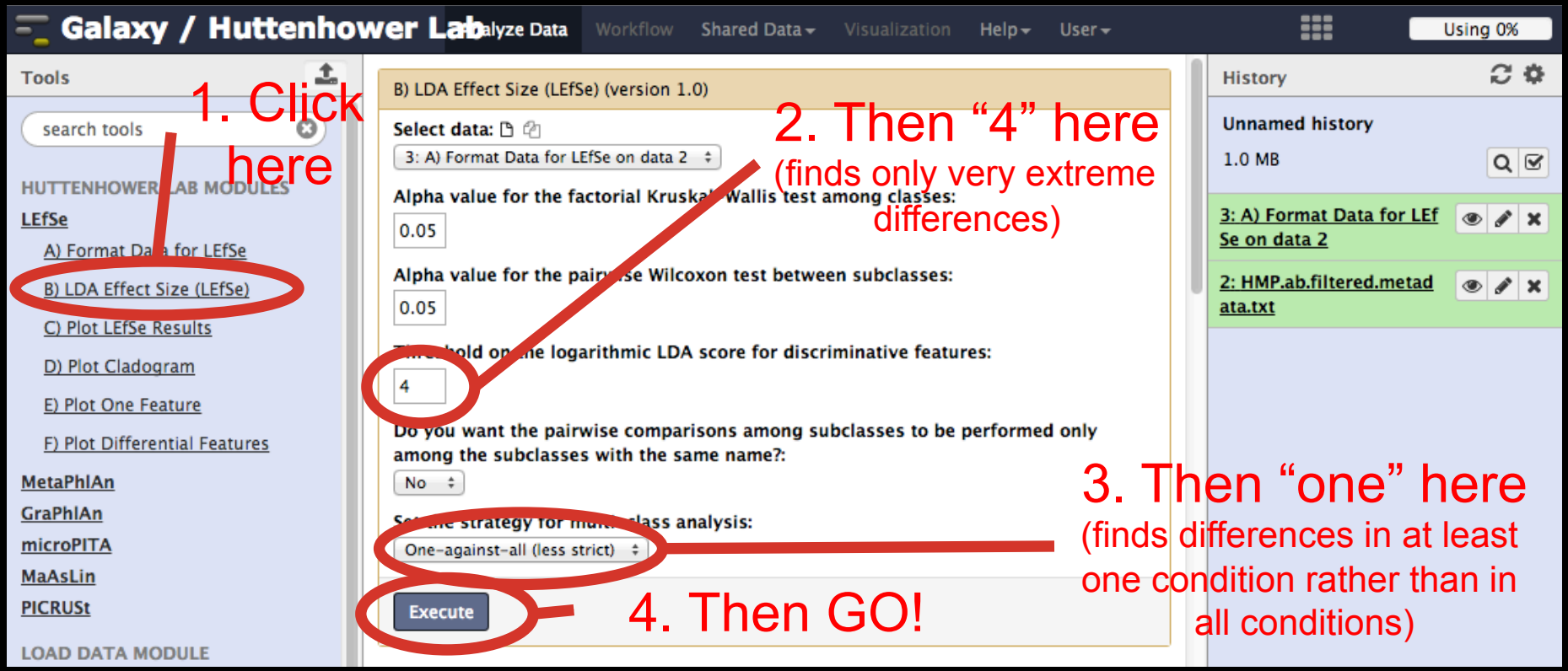
What it means: LEfSe

- Then tell LEfSe about your metadata:



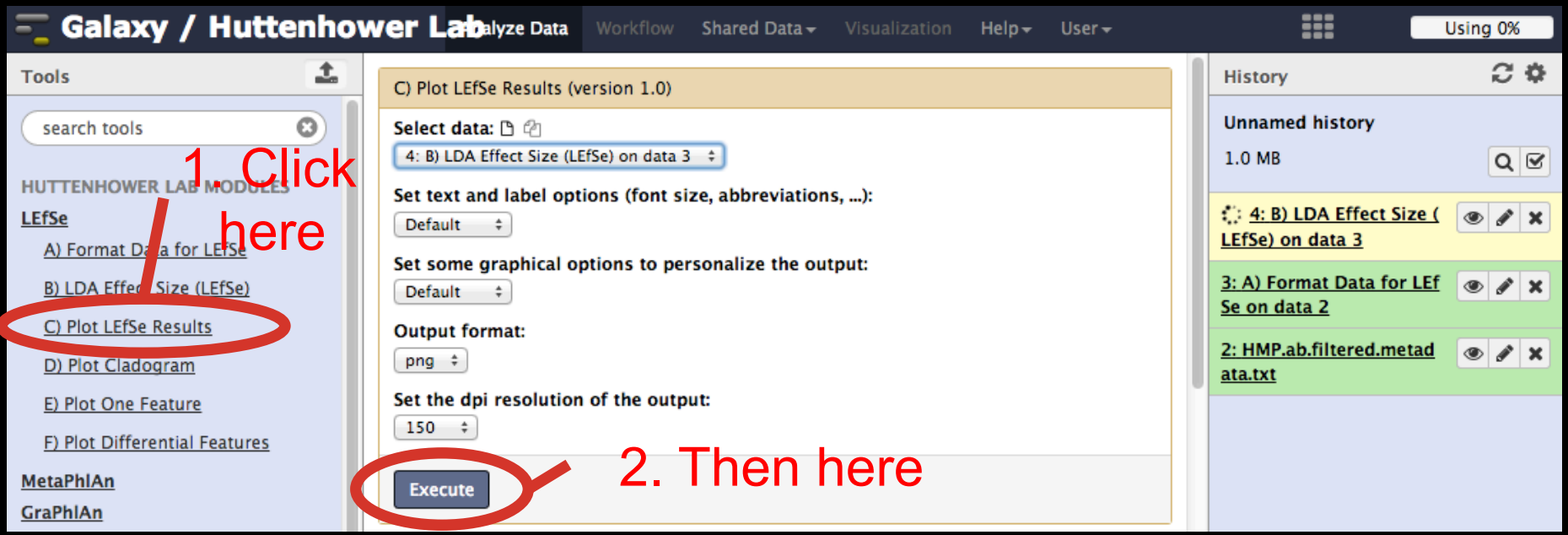
What it means: LEfSe

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



What it means: LEfSe

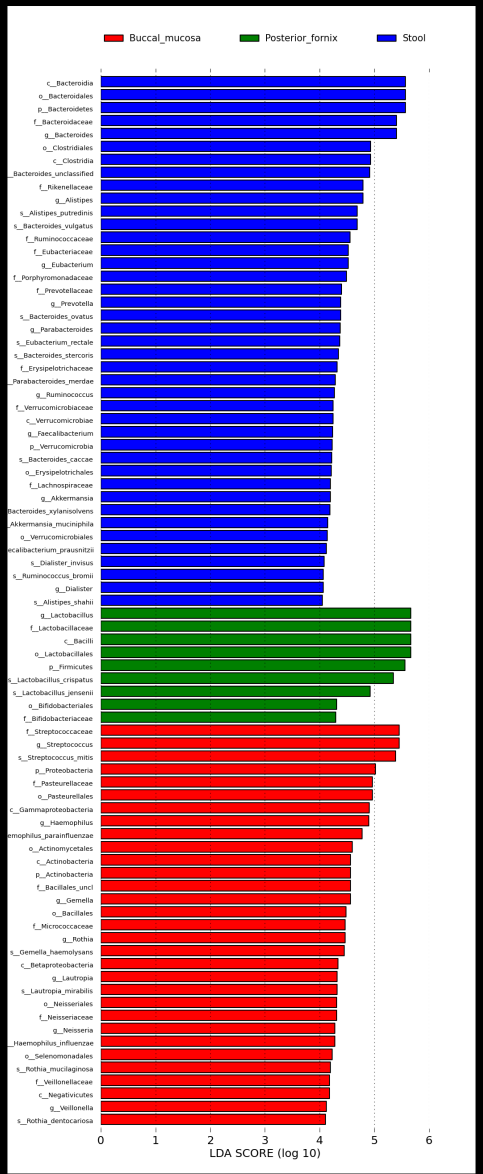
- You can plot the results as a bar plot
  - Again, lots of graphical parameters to modify if desired



What it means: LEfSe

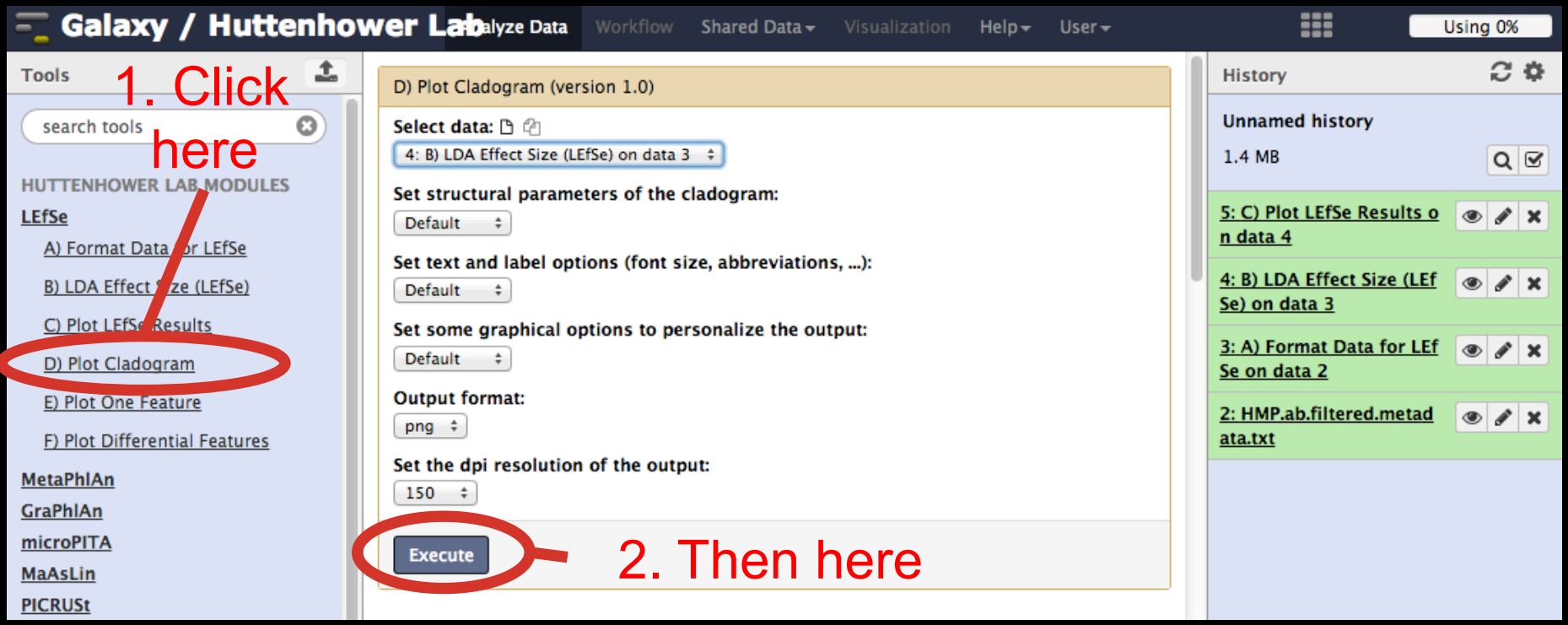
- In Galaxy, view a result by clicking on its “eye”

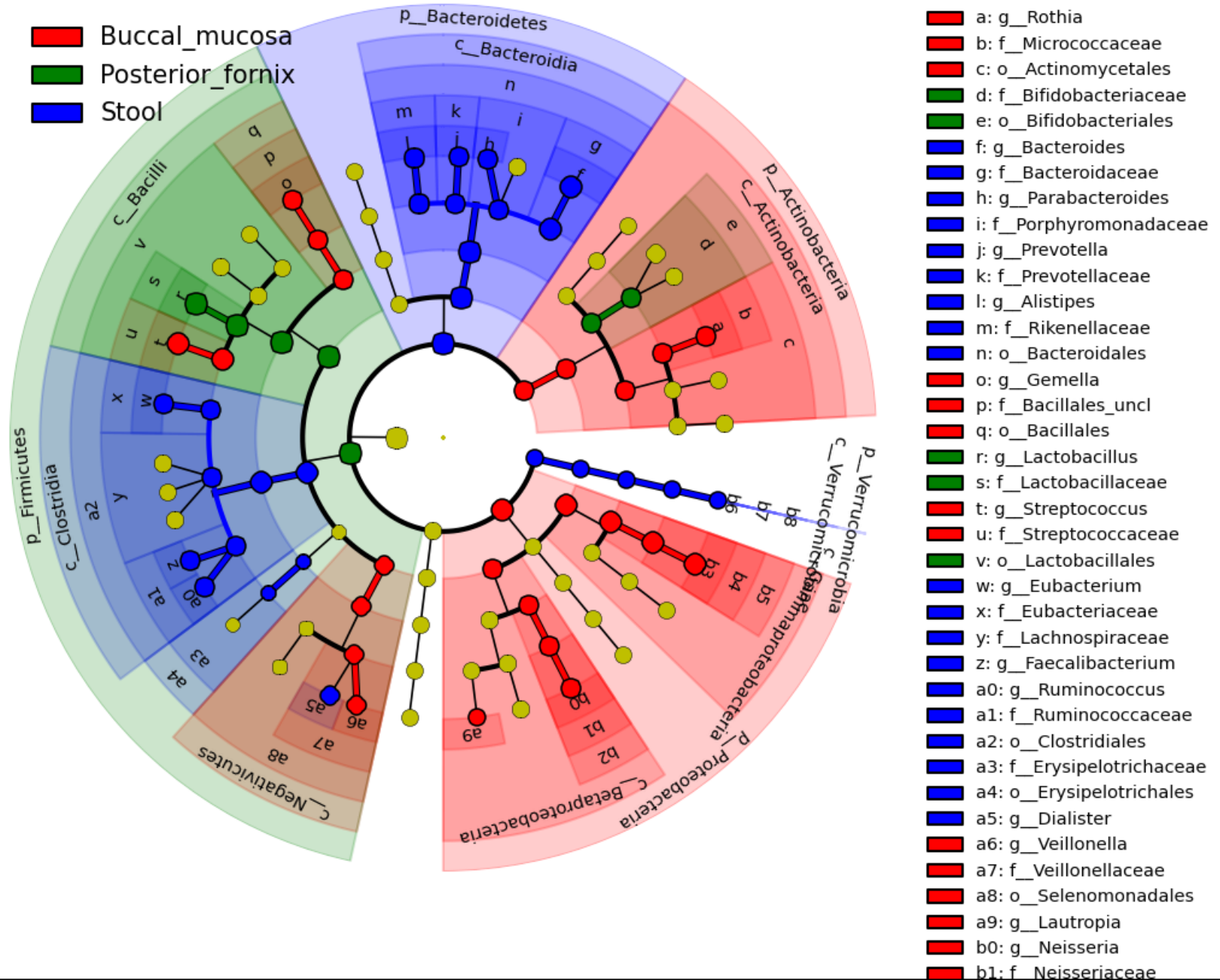
What it means: LEfSe



What it means: LEfSe

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

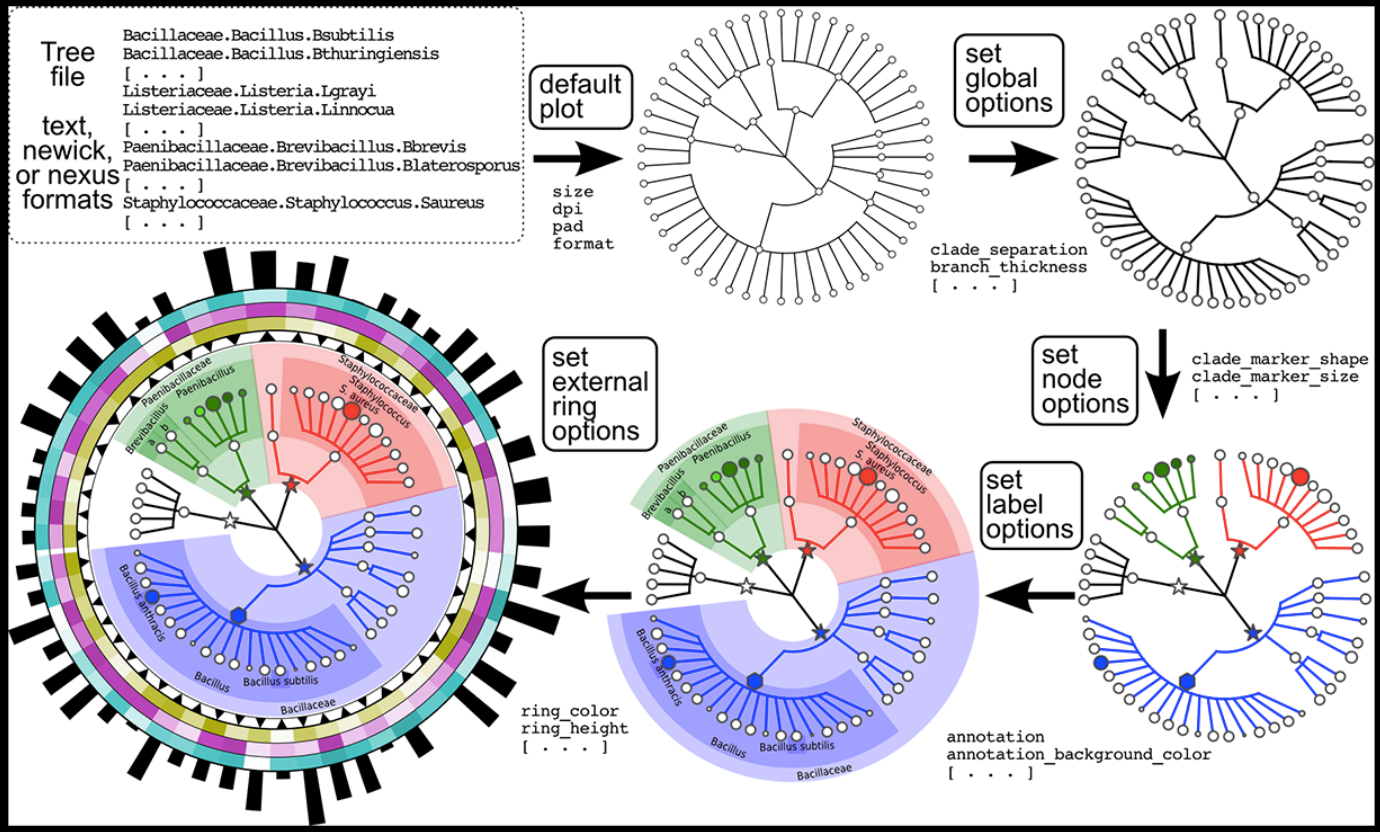



Buccal_mucosa
Posterior_fornix
Stool
p__Bacteroidetes
c__Bacteroidia
p__Actinobacteria
c__Actinobacteria
p__Firmicutes
c__Bacilli
c__Clostridia
a2
c__Negativicutes
c__Betaproteobacteria
p__Proteobacteria
c__Gammaproteobacteria
p__Verrucomicrobia
c__Verrucomicrobiae
a: g__Rothia
b: f__Micrococcaceae
c: o__Actinomycetales
d: f__Bifidobacteriaceae
e: o__Bifidobacteriales
f: g__Bacteroides
g: f__Bacteroidaceae
h: g__Parabacteroides
i: f__Porphyromonadaceae
j: g__Prevotella
k: f__Prevotellaceae
l: g__Alistipes
m: f__Rikenellaceae
n: o__Bacteroidales
o: g__Gemella
p: f__Bacillales_uncl
q: o__Bacillales
r: g__Lactobacillus
s: f__Lactobacillaceae
t: g__Streptococcus
u: f__Streptococcaceae
v: o__Lactobacillales
w: g__Eubacterium
x: f__Eubacteriaceae
y: f__Lachnospiraceae
z: g__Faecalibacterium
a0: g__Ruminococcus
a1: f__Ruminococcaceae
a2: o__Clostridiales
a3: f__Erysipelotrichaceae
a4: o__Erysipelotrichales
a5: g__Dialister
a6: g__Veillonella
a7: f__Veillonellaceae
a8: o__Selenomonadales
a9: g__Lautropia
b0: g__Neisseria
b1: f__Neisseriaceae

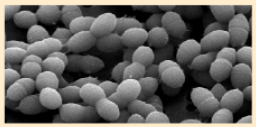
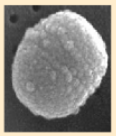
An aside: GraPhlAn

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

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

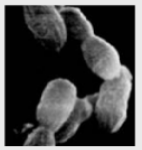


A map of diversity in the human microbiome

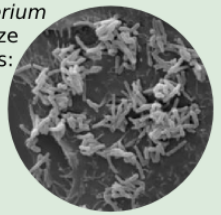
Streptococcus dominates the oral cavity with *S. mitis* > 75% in the cheek

Propionibacterium acnes lives on the skin and nose of most people



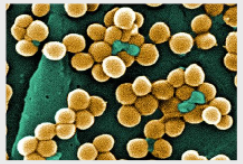
Many *Corynebacterium* species characterize different body sites:

C. matruchoti the plaque
C. accolens the nose
C. croppenstedtii the skin

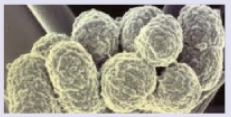


Lactobacillus species (*L. gasseri*, *L. jensenii*, *L. crispatus*, *L. iners*) are predominant but mutually exclusive in the vagina


Staphylococcus epidermidis colonizes external body sites

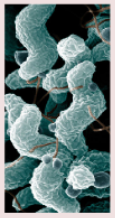


Several *Prevotella* species are present in the gastrointestinal tract. *P. copri* is present in 19% of the subjects and dominates the intestinal flora when present



Microscopy from http://bacmap.wishartlab.com

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



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


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

Commensal microbes

Potential pathogens

The four most abundant phyla

- Actinobacteria
- Bacteroidetes
- Firmicutes
- Proteobacteria

Low abundance phyla

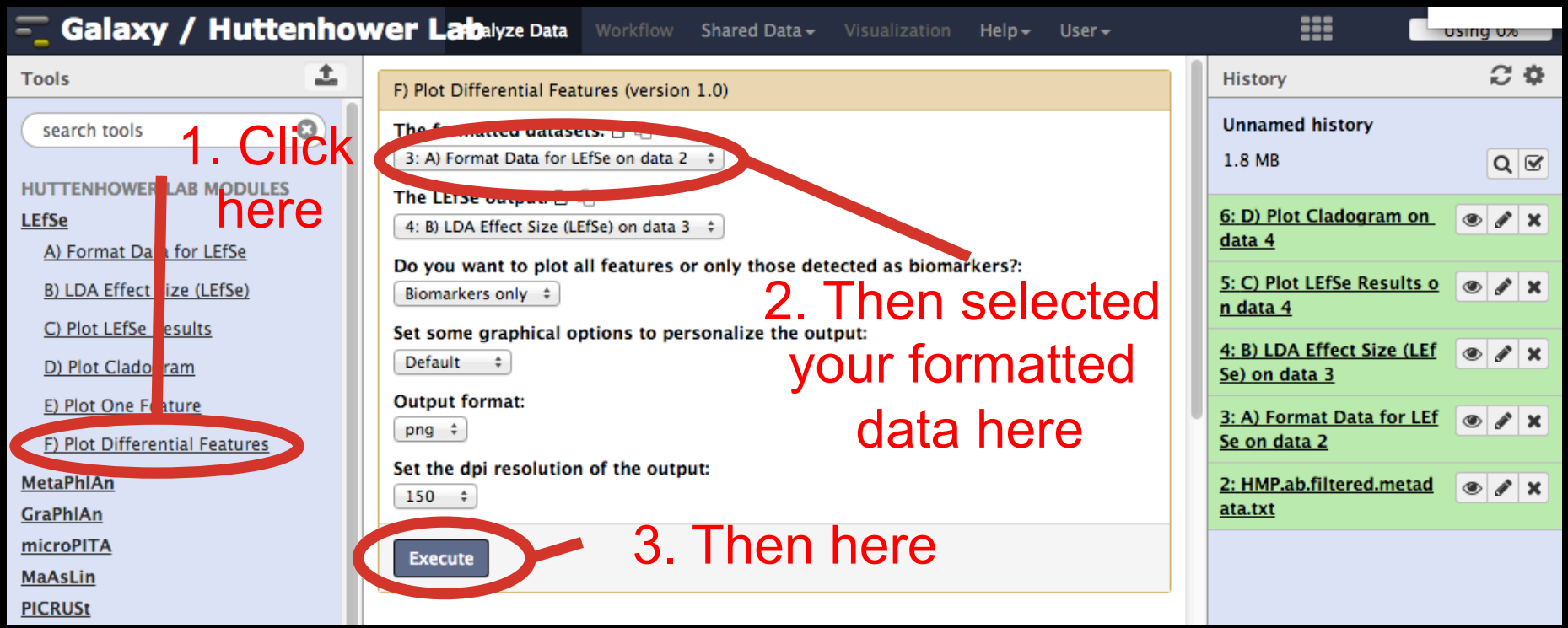
- Chloroflexi
- Cyanobacteria
- Euryarchaeota
- Fusobacteria
- Lentisphaerae
- Spirochaetes
- Synergistetes
- Tenericutes
- Thermi
- Verrucomicrobia

National Institutes of Health Human Microbiome Project

N. Segata & C. Huttenhower
http://huttenhower.sph.harvard.edu
(generated using CirClader and mTol-400 from MetaPhlAn analysis)

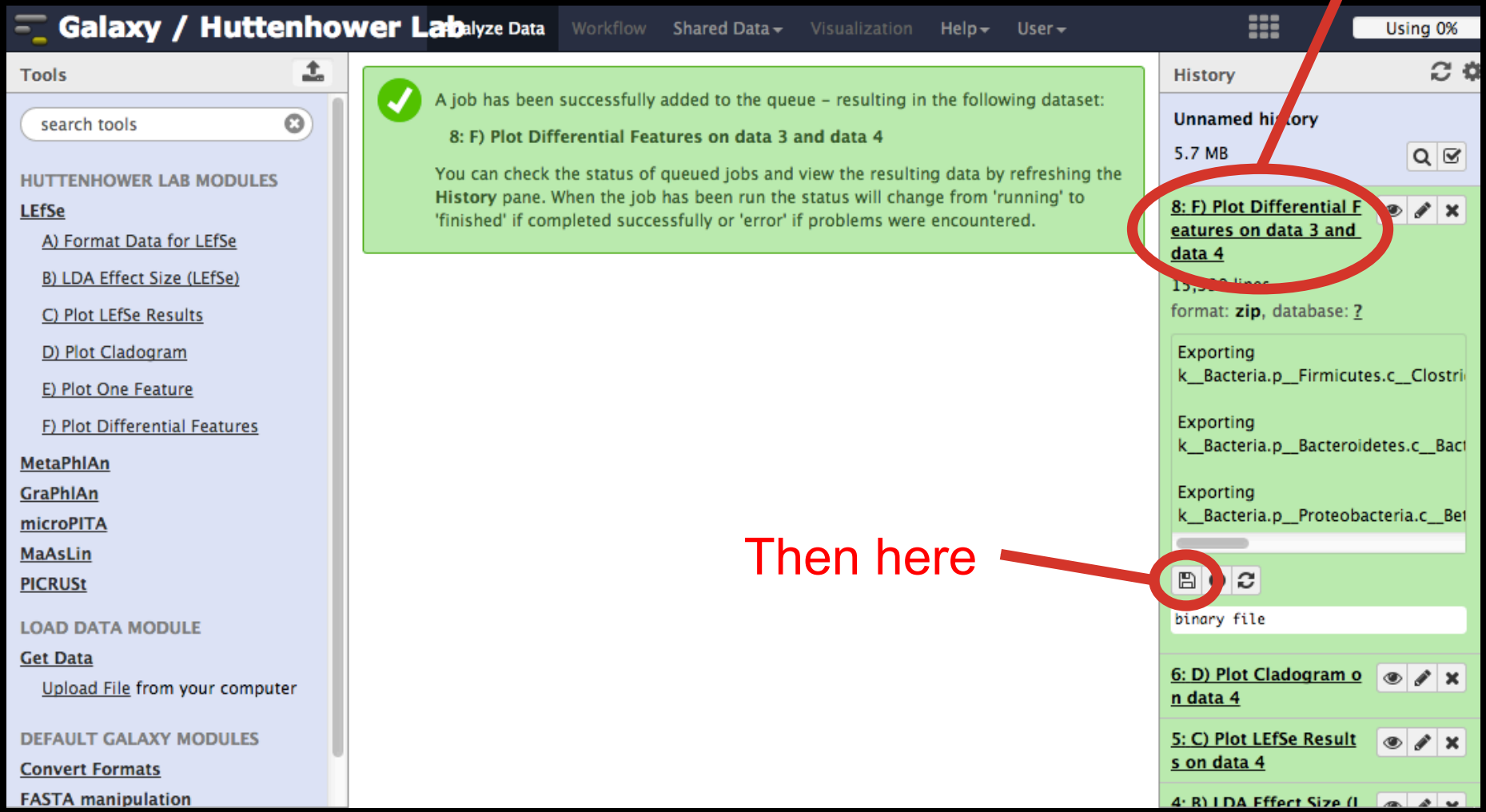
What it means: LEfSe

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



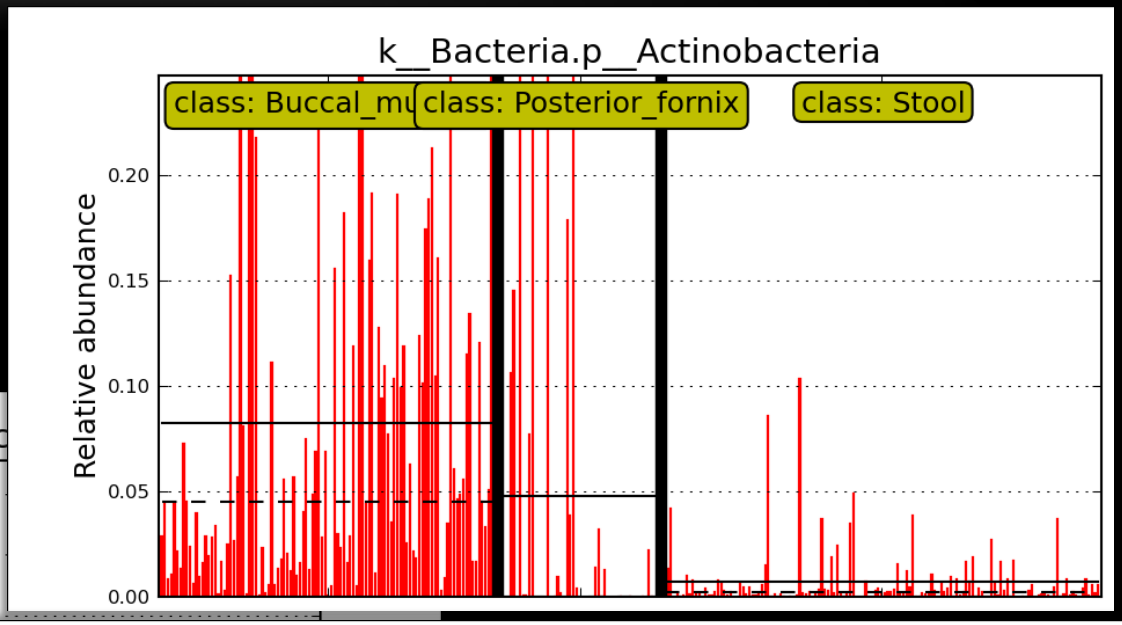
What it means: LEfSe

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

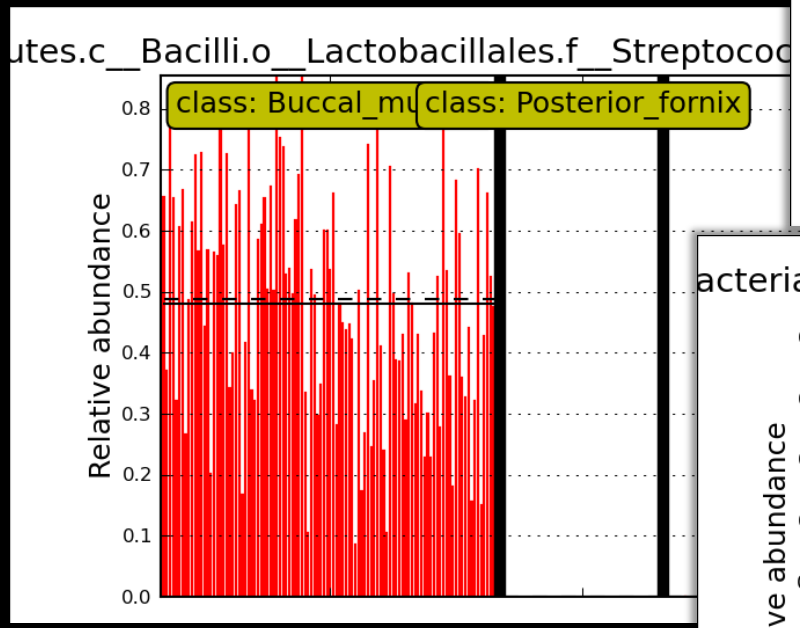


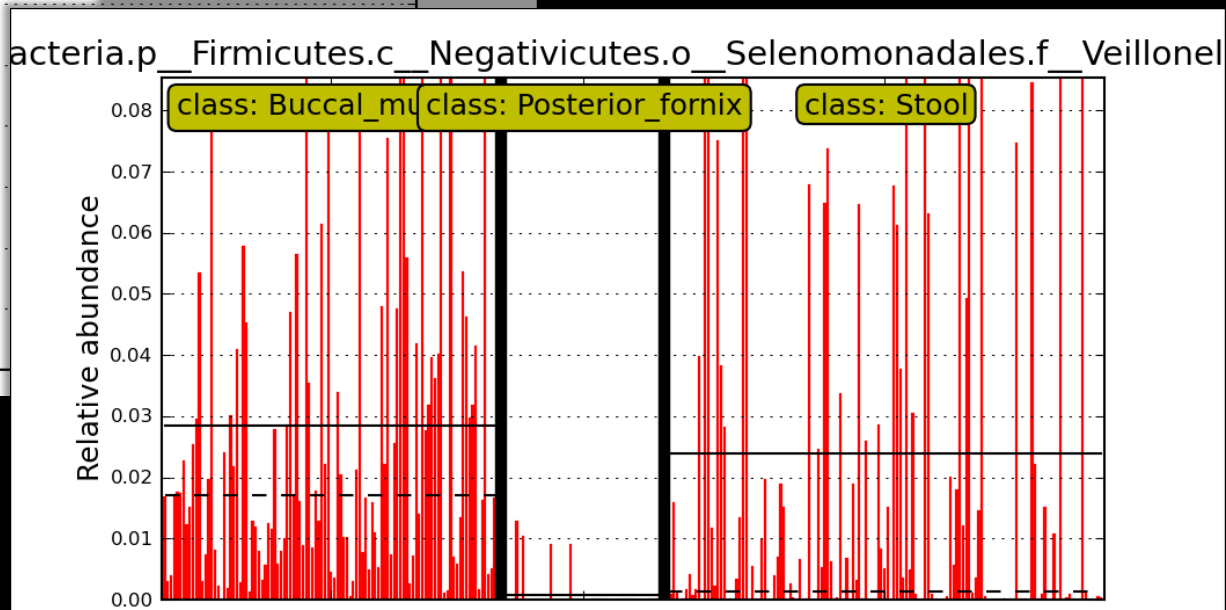
What it means: LEfSe

Actinobacteria



Strep. mitis





Veillonellaceae

Summary

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

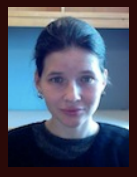
- LEfSe
  - Tab-delimited, stratified relative abundances in
  - Significantly differentially abundant features out

Thanks!

http://huttenhower.sph.harvard.edu


Alex Kostic


Levi Waldron


Xochitl Morgan


Tim Tickle


Daniela Boernigen

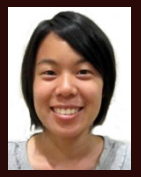

Lauren McIver

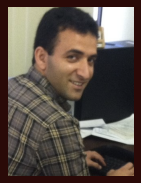

George Weingart


Emma Schwager


Eric Franzosa


Boyu Ren


Tiffany Hsu


Ali Rahnavard


Ayshwarya Subramanian


Jim Kaminski


Regina Joice


Koji Yasuda


Kevin Oh


Galeb Abu-Ali


Afrah Shafquat


Randall Schwager


Chengwei Luo


Keith Bayer


Moran Yassour


Alexandra Sirota


Dirk Gevers


Ramnik Xavier


Wendy Garrett


Andy Chan

Human Microbiome Project 2

Lita Procter
Jon Braun
Dermot McGovern
Subra Kugathasan
Ted Denson
Janet Jansson
Bruce Birren
Chad Nusbaum
Clary Clish
Joe Petrosino
Thad Stappenbeck

BROAD INSTITUTE

Human Microbiome Project

Jane Peterson
Sarah Highlander
Barbara Methe
Karen Nelson
George Weinstock
Owen White


Nicola Segata


Gautam Dantas
Molly Gibson


Katherine Lemon


Brendan Bohannan
James Meadow