Meta’omic Analysis with MetaPhlAn & LEfSe

Eric Franzosa
Postdoctoral Fellow / Huttenhower Lab

Symposium and Workshop on New Methods for Phylogenomics and Metagenomics
The University of Texas at Austin
17 February 2013

Harvard School of Public Health
Department of Biostatistics

Metagenomic Phylogenetic Analysis

Fast and accurate metagenomic profiling of
microbial community composition using unique
clade-specific marker genes

LDA Effect Size

High-dimensional biomarker
discovery and explanation

http://http://huttenhower.sph.harvard.edu/content/metaphlan-tutorial

Tutorial Outline

- Introduction to MetaPhlAn
- MetaPhlAn Demo
- Introduction to LEfSe
- LEfSe Demo
- Links, other tools, and Q&A

Tutorial Outline

- ▶ **Introduction to MetaPhlAn**
- ▶ MetaPhlAn Demo
- ▶ Introduction to LEfSe
- ▶ LEfSe Demo
- ▶ Links, other tools, and Q&A

MetaPhlAn Motivation

Tons of microbes

Shotgun sequencing

Tons of short reads

?

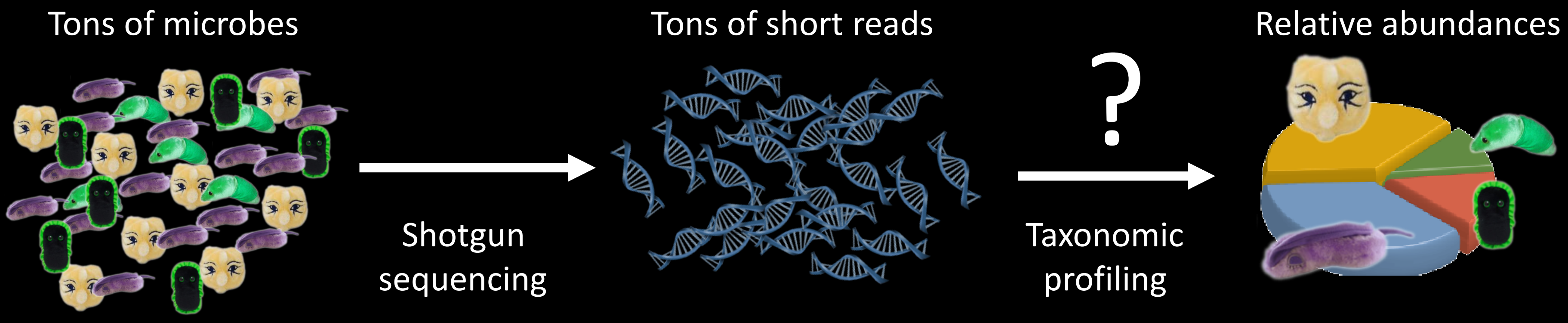
Taxonomic profiling

Relative abundances

Profile the taxonomic composition of microbial communities from whole shotgun metagenomic data

- Which clades (e.g. species, genera, classes) are there?
- What is the relative abundance of each clade in the community?

MetaPhlAn Motivation


Tons of microbes

Shotgun sequencing

Tons of short reads

?

Taxonomic profiling

Relative abundances

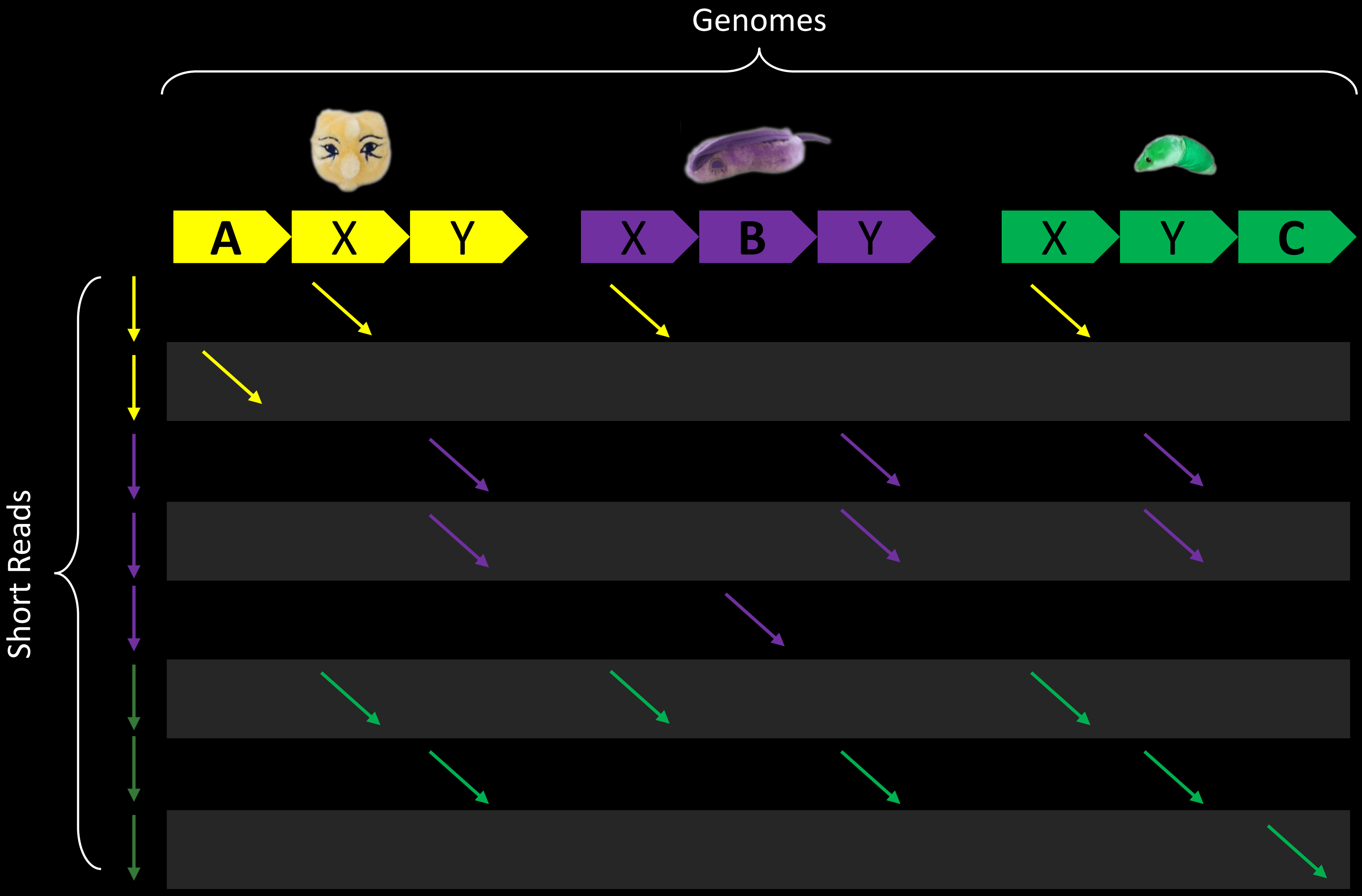
Profile the taxonomic composition of microbial communities from whole shotgun metagenomic data

- Which clades (e.g. species, genera, classes) are there?
- What is the relative abundance of each clade in the community?

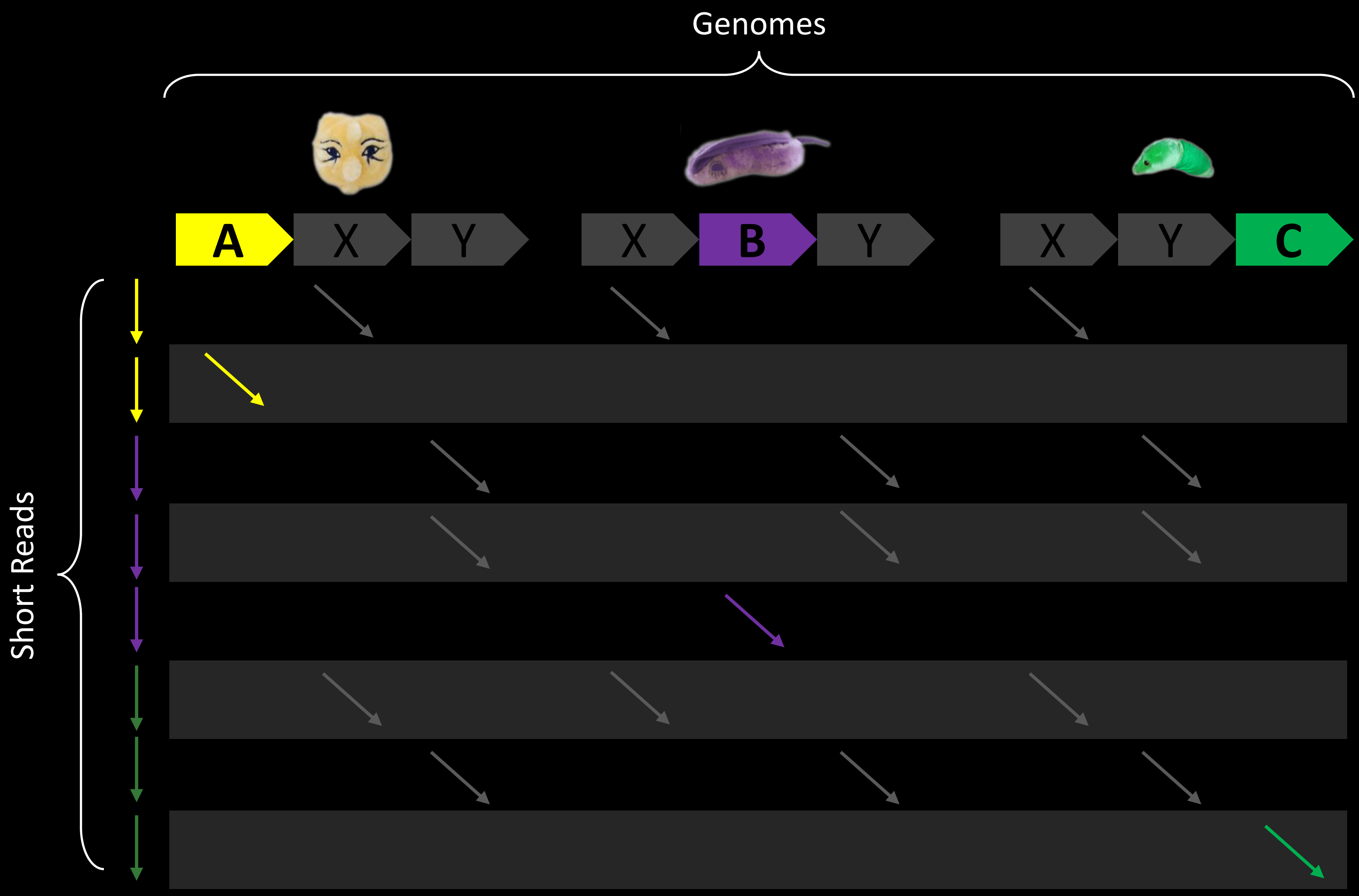
Several challenges

- Species-level resolution
- Computationally feasibility
- Organismal relative abundance rather than DNA concentrations
- Consistent detection confidence for all clades, including archaea
- High accuracies for very short reads (as short as ~50nt)
- Detection of organisms without sequenced genomes at higher taxonomic levels

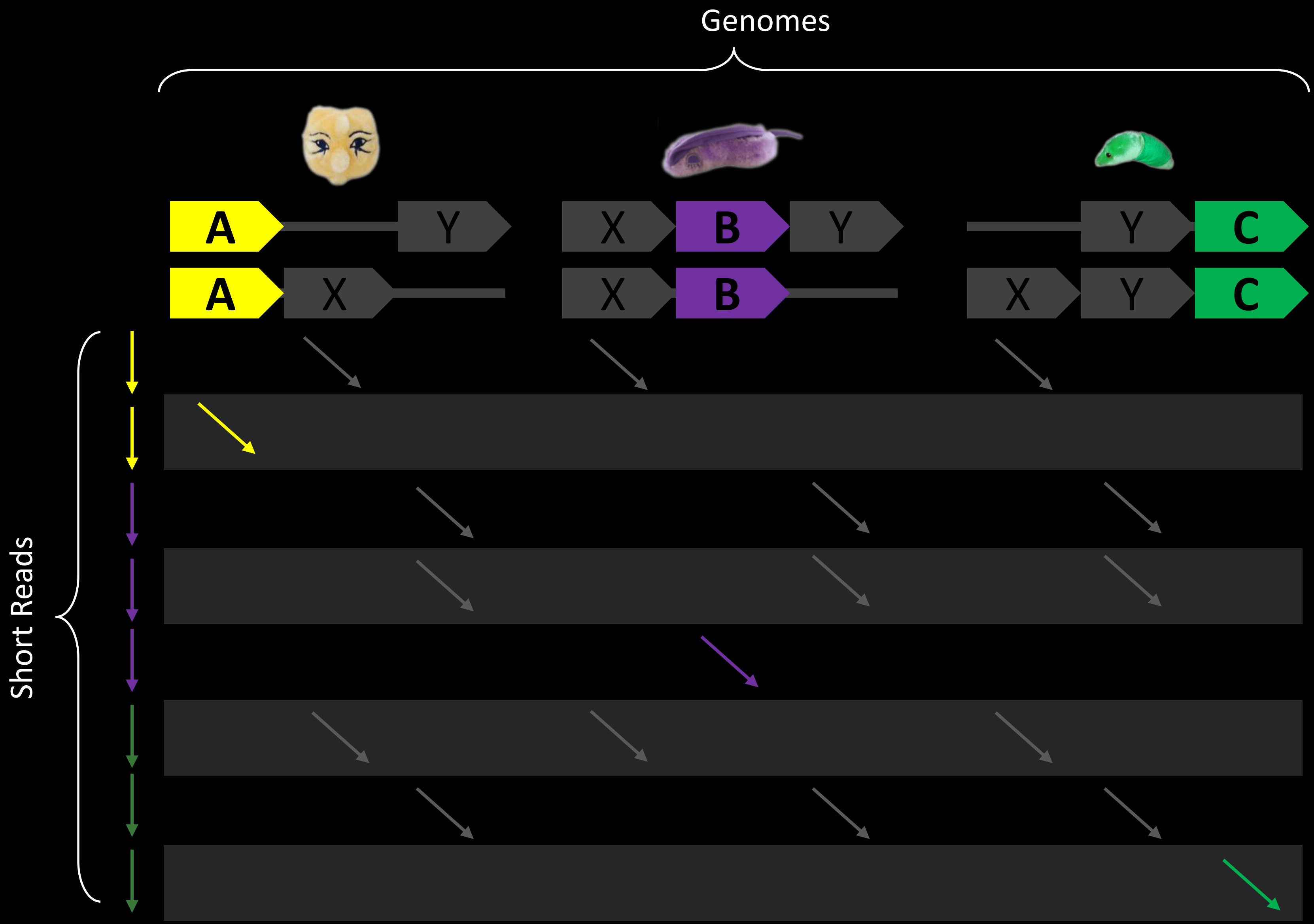
MetaPhlAn Overview



MetaPhlAn Overview

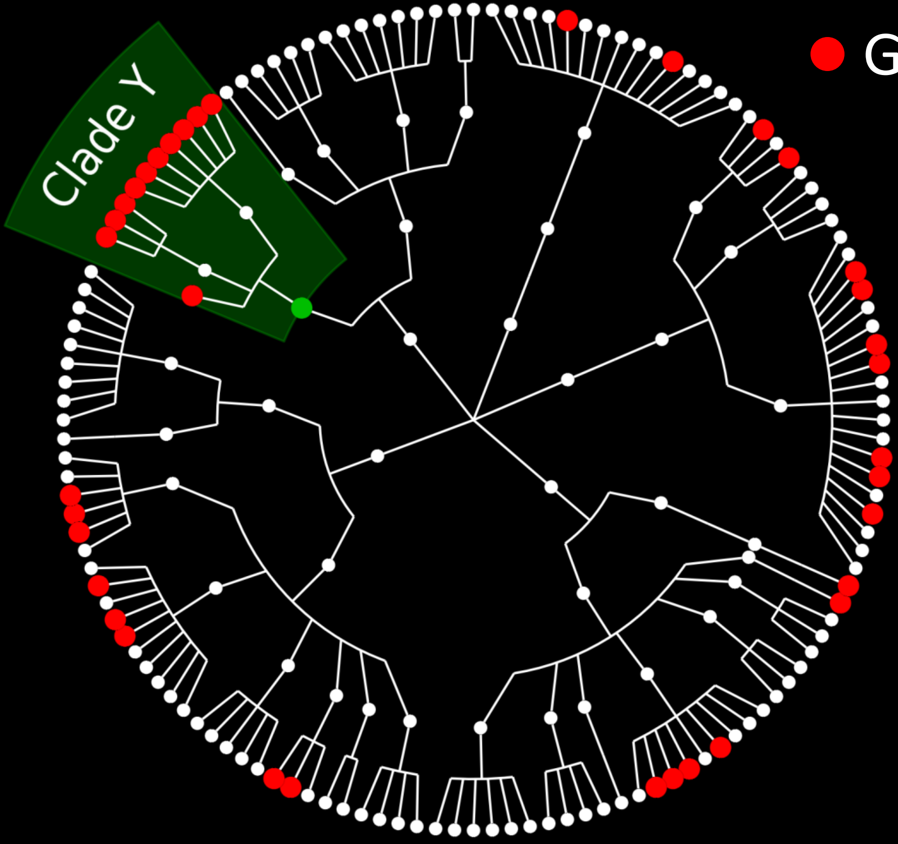


MetaPhlAn Overview



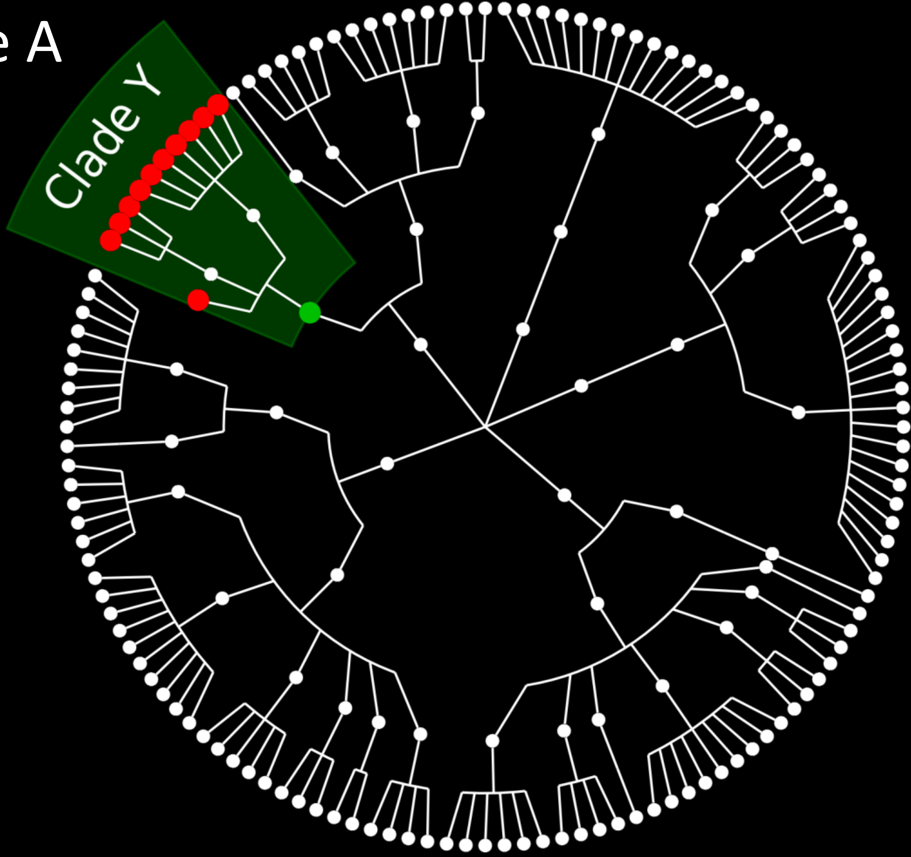
MetaPhlAn Overview

A is a core gene for clade Y



● Gene A

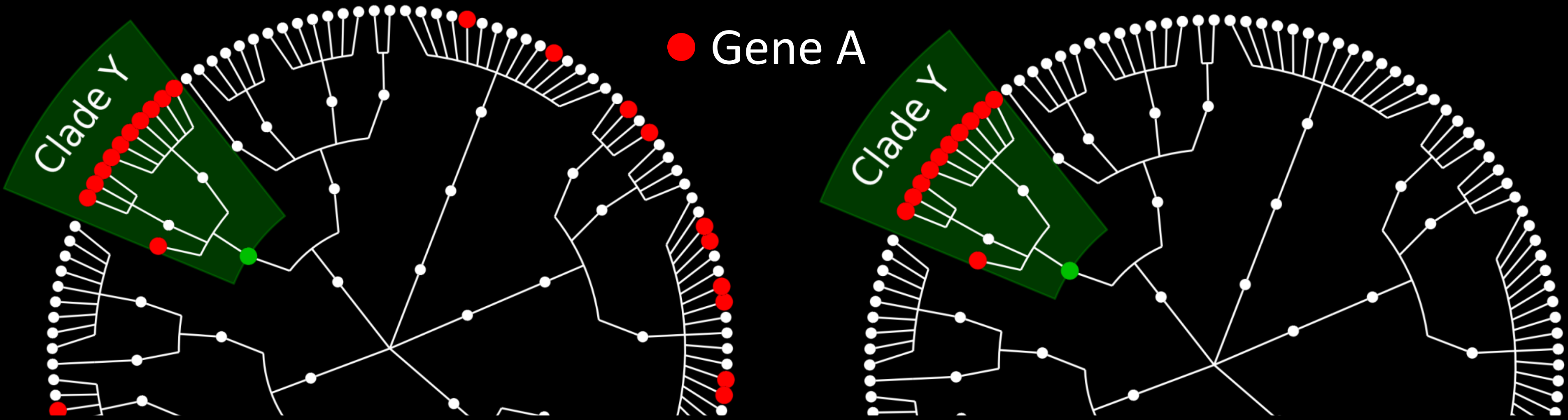
A is a unique marker gene for clade Y

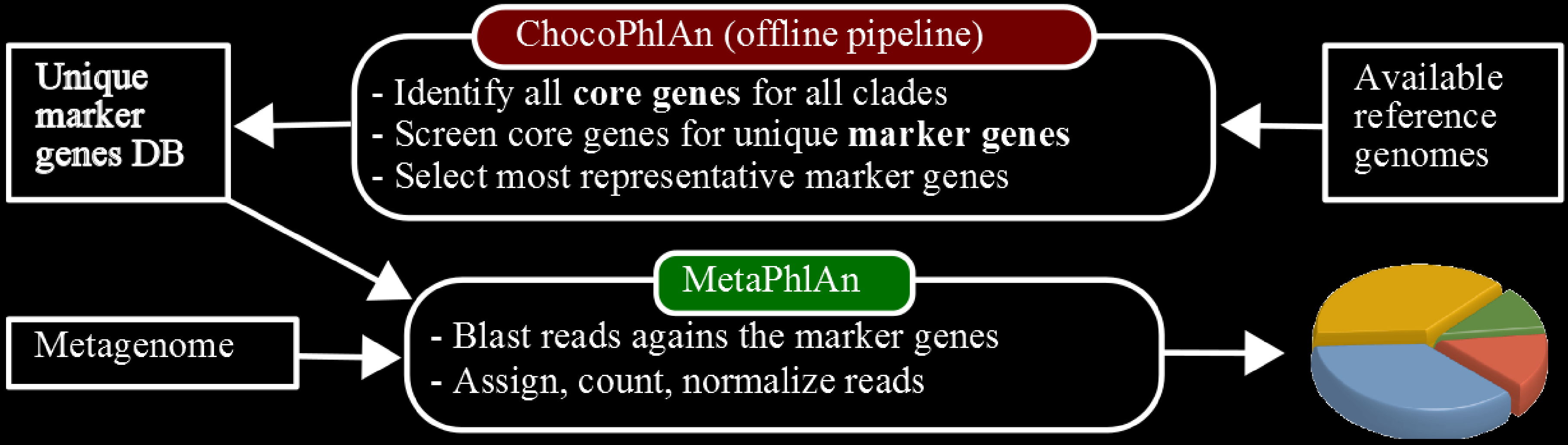


MetaPhlAn Overview

A is a core gene for clade Y

A is a unique marker gene for clade Y





Unique marker genes DB

ChocoPhlAn (offline pipeline)

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

Available reference genomes

Metagenome

MetaPhlAn

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

Tutorial Outline

- Introduction to MetaPhlAn
- **MetaPhlAn Demo**
- Introduction to LEfSe
- LEfSe Demo
- Links, other tools, and Q&A

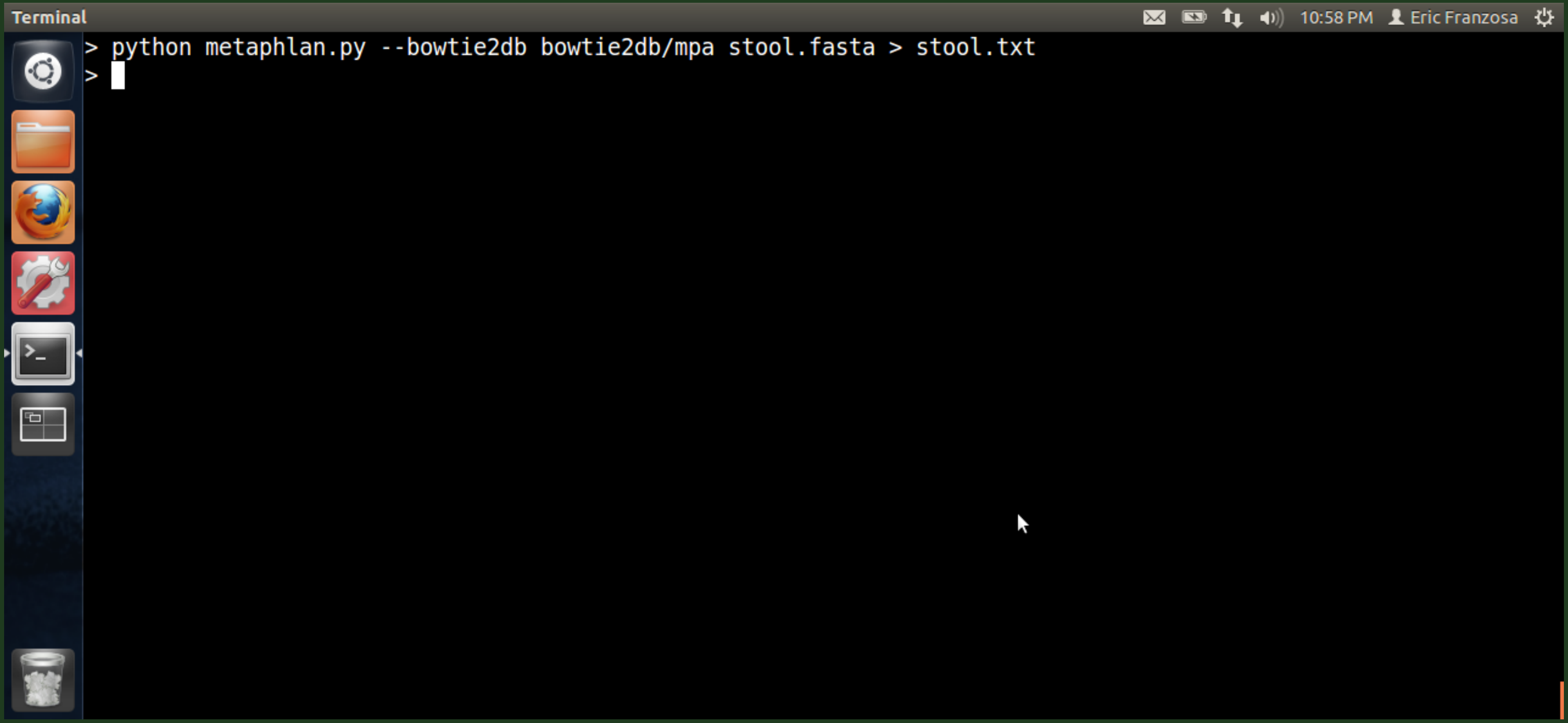
MetaPhlAn Demo: Setup

- Download program and marker database from: http://huttenhower.sph.harvard.edu/metaphlan
- Requires python with **numpy** installed
- Requires **BLAST** or **Bowtie2** for alignment (**Bowtie2** recommended)
- Today's sample data available at: http://huttenhower.sph.harvard.edu/content/metaphlan-tutorial

MetaPhlAn (command line)

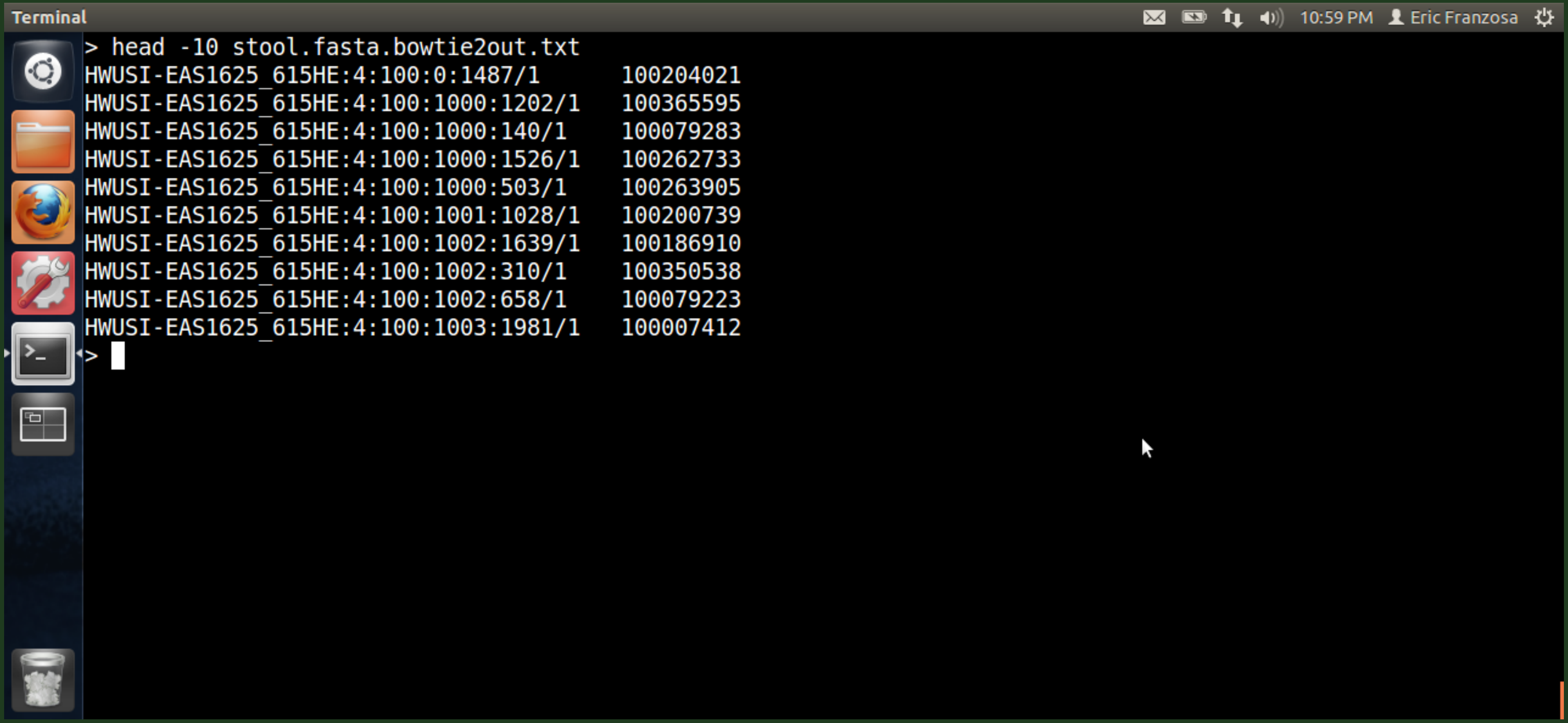
- Show minimal MetaPhlAn setup
- Run MetaPhlAn on downsampled HMP FASTA
- Examine marker output file
- Example abundance output file
- Discuss taxonomic organization
- Discuss “unclassified” species

MetaPhlAn (command line)



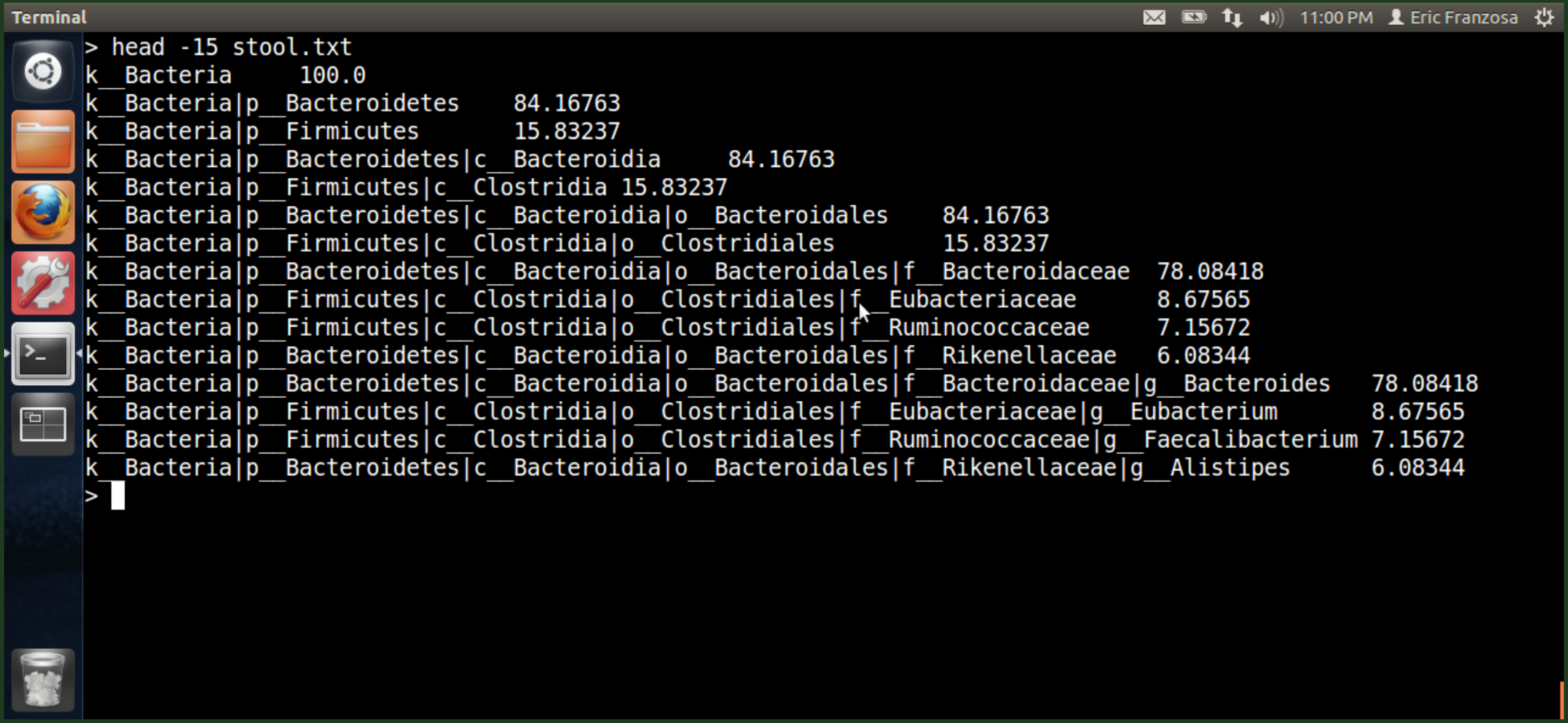
```
> python metaphlan.py --bowtie2db bowtie2db/mpa stool.fasta > stool.txt
> 
```

MetaPhlAn (command line)



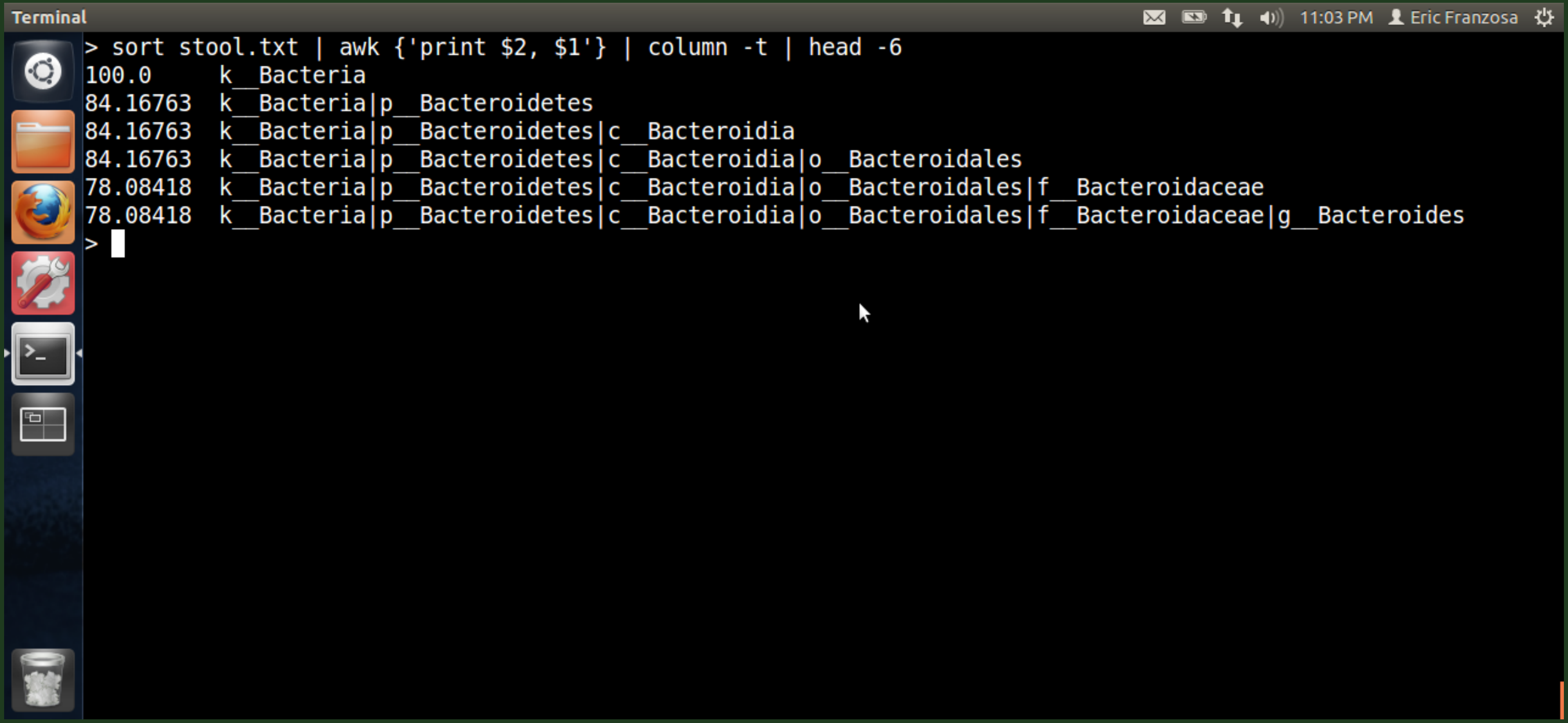
```
> head -10 stool.fasta.bowtie2out.txt
HWUSI-EAS1625_615HE:4:100:0:1487/1      100204021
HWUSI-EAS1625_615HE:4:100:1000:1202/1   100365595
HWUSI-EAS1625_615HE:4:100:1000:140/1    100079283
HWUSI-EAS1625_615HE:4:100:1000:1526/1   100262733
HWUSI-EAS1625_615HE:4:100:1000:503/1    100263905
HWUSI-EAS1625_615HE:4:100:1001:1028/1   100200739
HWUSI-EAS1625_615HE:4:100:1002:1639/1   100186910
HWUSI-EAS1625_615HE:4:100:1002:310/1    100350538
HWUSI-EAS1625_615HE:4:100:1002:658/1    100079223
HWUSI-EAS1625_615HE:4:100:1003:1981/1   100007412
> 
```

MetaPhlAn (command line)



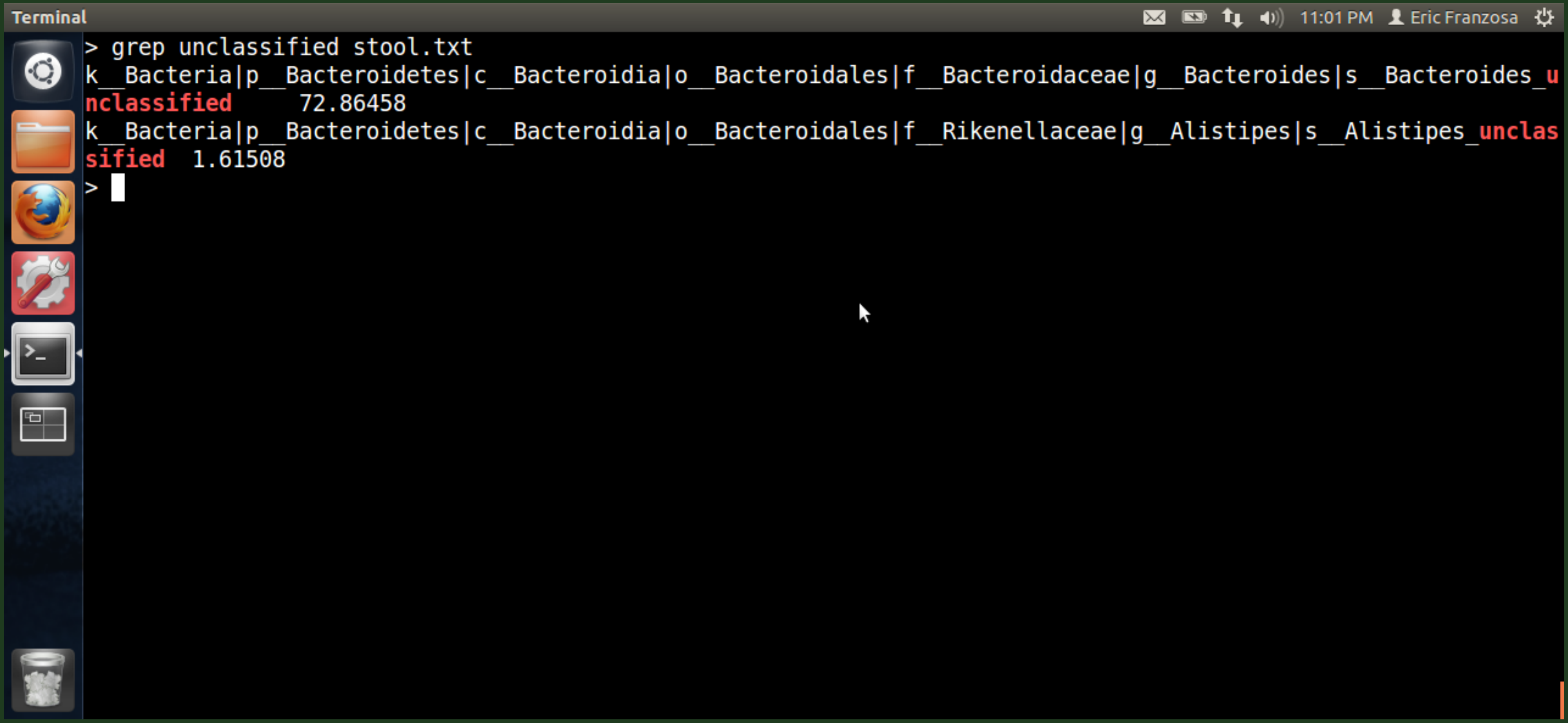
```
> head -15 stool.txt
k__Bacteria     100.0
k__Bacteria|p__Bacteroidetes    84.16763
k__Bacteria|p__Firmicutes       15.83237
k__Bacteria|p__Bacteroidetes|c__Bacteroidia     84.16763
k__Bacteria|p__Firmicutes|c__Clostridia 15.83237
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales    84.16763
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales        15.83237
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Bacteroidaceae  78.08418
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Eubacteriaceae      8.67565
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Ruminococcaceae     7.15672
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Rikenellaceae   6.08344
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Bacteroidaceae|g__Bacteroides   78.08418
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Eubacteriaceae|g__Eubacterium       8.67565
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Ruminococcaceae|g__Faecalibacterium 7.15672
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Rikenellaceae|g__Alistipes      6.08344
> 
```

MetaPhlAn (command line)



```
> sort stool.txt | awk {'print $2, $1'} | column -t | head -6
100.0     k__Bacteria
84.16763  k__Bacteria|p__Bacteroidetes
84.16763  k__Bacteria|p__Bacteroidetes|c__Bacteroidia
84.16763  k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales
78.08418  k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Bacteroidaceae
78.08418  k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Bacteroidaceae|g__Bacteroides
> 
```

MetaPhlAn (command line)



```
> grep unclassified stool.txt
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Bacteroidaceae|g__Bacteroides|s__Bacteroides_u
nclassified     72.86458
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Rikenellaceae|g__Alistipes|s__Alistipes_unclas
sified  1.61508
> 
```

Tutorial Outline

- Introduction to MetaPhlAn
- MetaPhlAn Demo
- **Introduction to LEfSe**
- LEfSe Demo
- Links, other tools, and Q&A

Turning anecdotes into biology

Site	Oral	Gut
Clade1	0.40	0.87
Clade1|Bug1	0.40	0.56
Clade1|Bug2	0.00	0.30
Clade2	0.60	0.13
Clade2|Bug3	0.11	0.00
Clade2|Bug4	0.49	0.13

Turning anecdotes into biology

- More samples are a start…
- But the statistics are still non-trivial
  - Data are noisy
  - Compositional nature (Σ = 1)
  - High dynamic range
  - Hierarchical organization

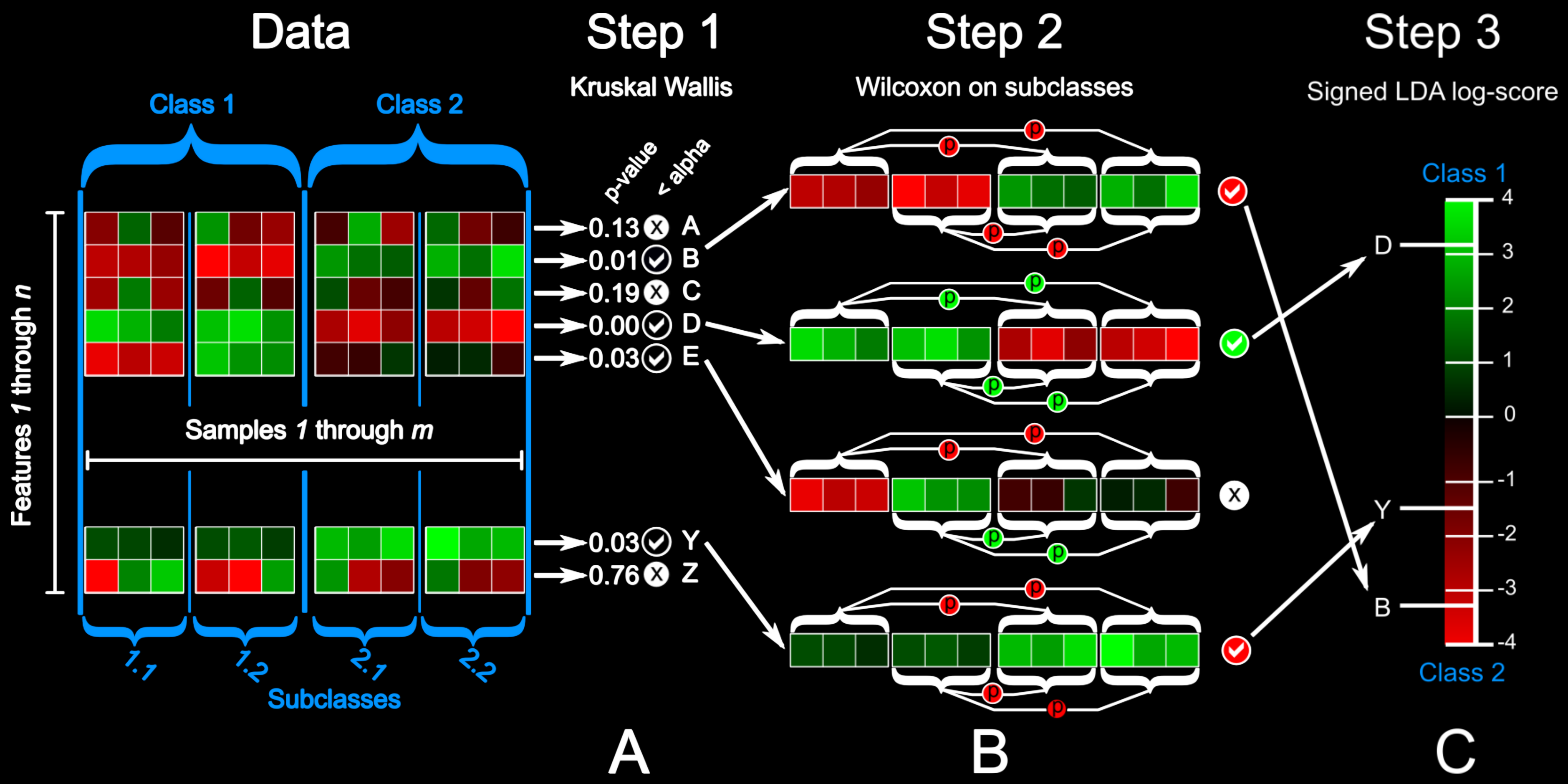
Site	Oral	Gut	Oral	Gut	Oral	Gut
Clade1	0.40	0.87	0.43	0.68	0.47	0.32
Clade1|Bug1	0.40	0.56	0.07	0.31	0.42	0.27
Clade1|Bug2	0.00	0.30	0.36	0.37	0.04	0.05
Clade2	0.60	0.13	0.57	0.32	0.53	0.68
Clade2|Bug3	0.11	0.00	0.10	0.32	0.15	0.23
Clade2|Bug4	0.49	0.13	0.47	0.00	0.39	0.45

Turning anecdotes into biology

- We may also want to leverage (or control for) additional metadata

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: *finding metagenomic biomarkers*



Statistical consistency

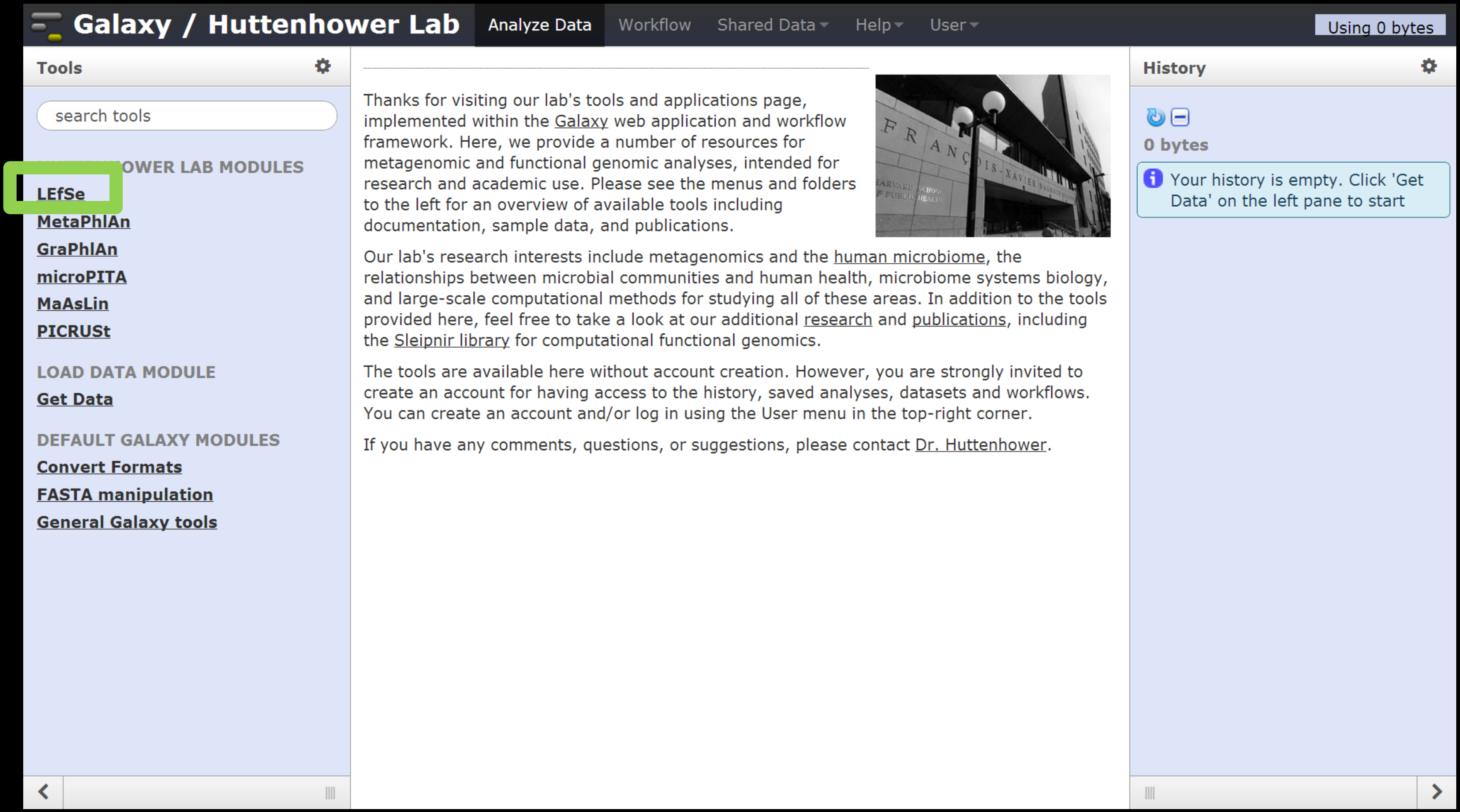
Biological consistency

Biological Effect size

Tutorial Outline

- ▶ Introduction to MetaPhlAn
- ▶ MetaPhlAn Demo
- ▶ Introduction to LEfSe
- ▶ **LEfSe Demo**
- ▶ Links, other tools, and Q&A

LEfSe Demo: Galaxy Version



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.

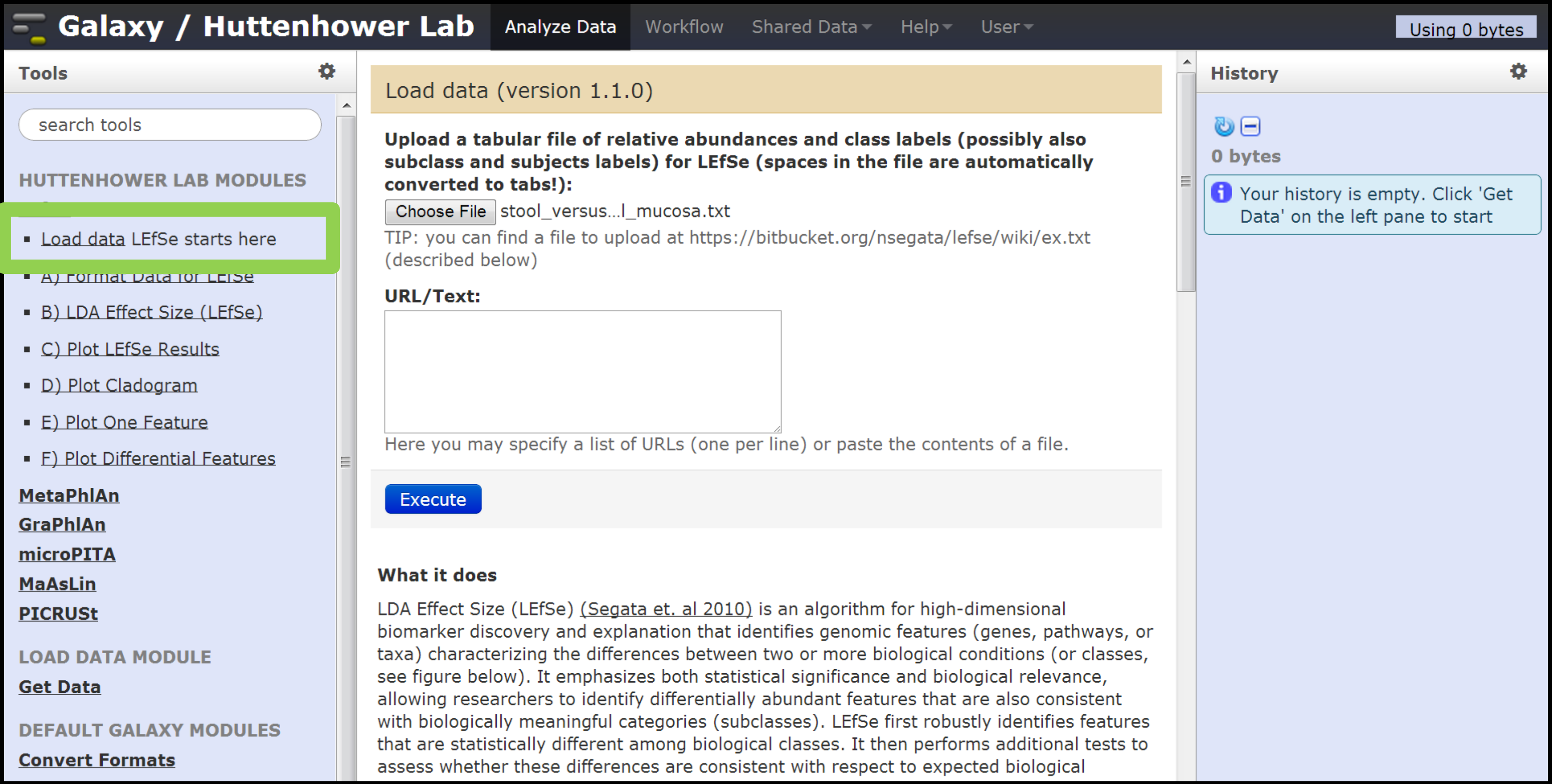
The tools are available here without account creation. However, you are strongly invited to create an account for having access to the history, saved analyses, datasets and workflows. You can create an account and/or log in using the User menu in the top-right corner.

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

Your history is empty. Click 'Get Data' on the left pane to start

Available at: http://huttenhower.sph.harvard.edu/galaxy/

Load your data table



Load data (version 1.1.0)

Upload a tabular file of relative abundances and class labels (possibly also subclass and subjects labels) for LEfSe (spaces in the file are automatically converted to tabs!):

Choose File stool_versus...l_mucosa.txt

TIP: you can find a file to upload at https://bitbucket.org/nsegata/lefse/wiki/ex.txt (described below)

URL/Text:

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

What it does

LDA Effect Size (LEfSe) (Segata et. al 2010) is an algorithm for high-dimensional biomarker discovery and explanation that identifies genomic features (genes, pathways, or taxa) characterizing the differences between two or more biological conditions (or classes, see figure below). It emphasizes both statistical significance and biological relevance, allowing researchers to identify differentially abundant features that are also consistent with biologically meaningful categories (subclasses). LEfSe first robustly identifies features that are statistically different among biological classes. It then performs additional tests to assess whether these differences are consistent with respect to expected biological

Your history is empty. Click 'Get Data' on the left pane to start

Tab delimited text, consisting of a class (e.g. oral vs. gut), an optional subclass (e.g. male vs. female), an optional sample ID, and then your features.

My input to Galaxy…

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

- MetaPhlAn output for many HMP subjects
- Buccal_mucosa *is a proxy for* oral
- Stool *is a proxy for* gut
- Input file included with the sample data

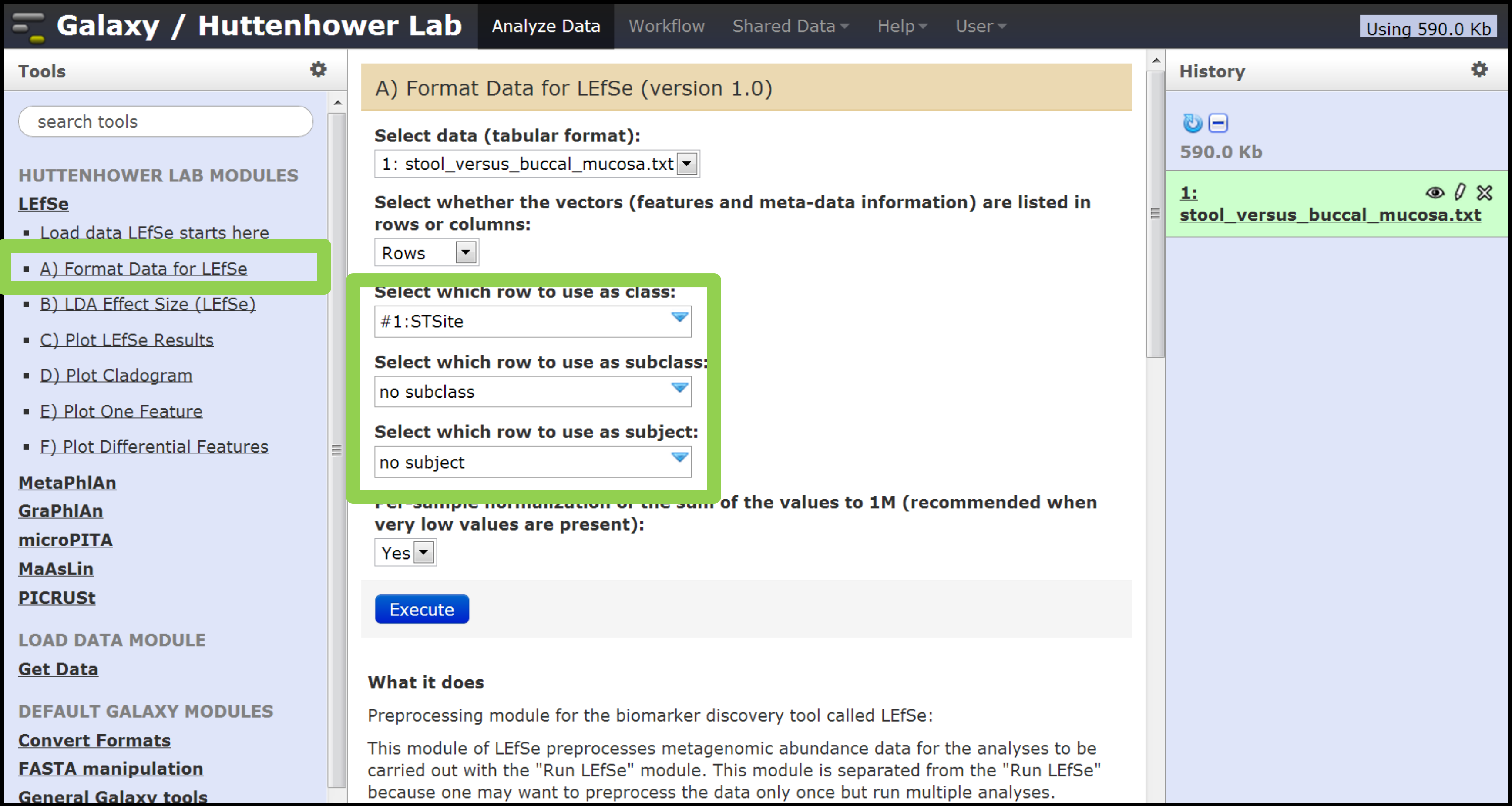
My input to Galaxy…

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

- MetaPhlAn output for many HMP subjects
- Buccal_mucosa *is a proxy for* oral
- Stool *is a proxy for* gut
- Input file included with the sample data

- Features don’t have to come from MetaPhlAn
- Any tab-delimited continuous data will work!

Tell LEfSe which rows are class/subclass/ID



A) Format Data for LEfSe (version 1.0)

Select data (tabular format): 1: stool_versus_buccal_mucosa.txt

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

Select which row to use as class: #1:STSite

Select which row to use as subclass: no subclass

Select which row to use as subject: no subject

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

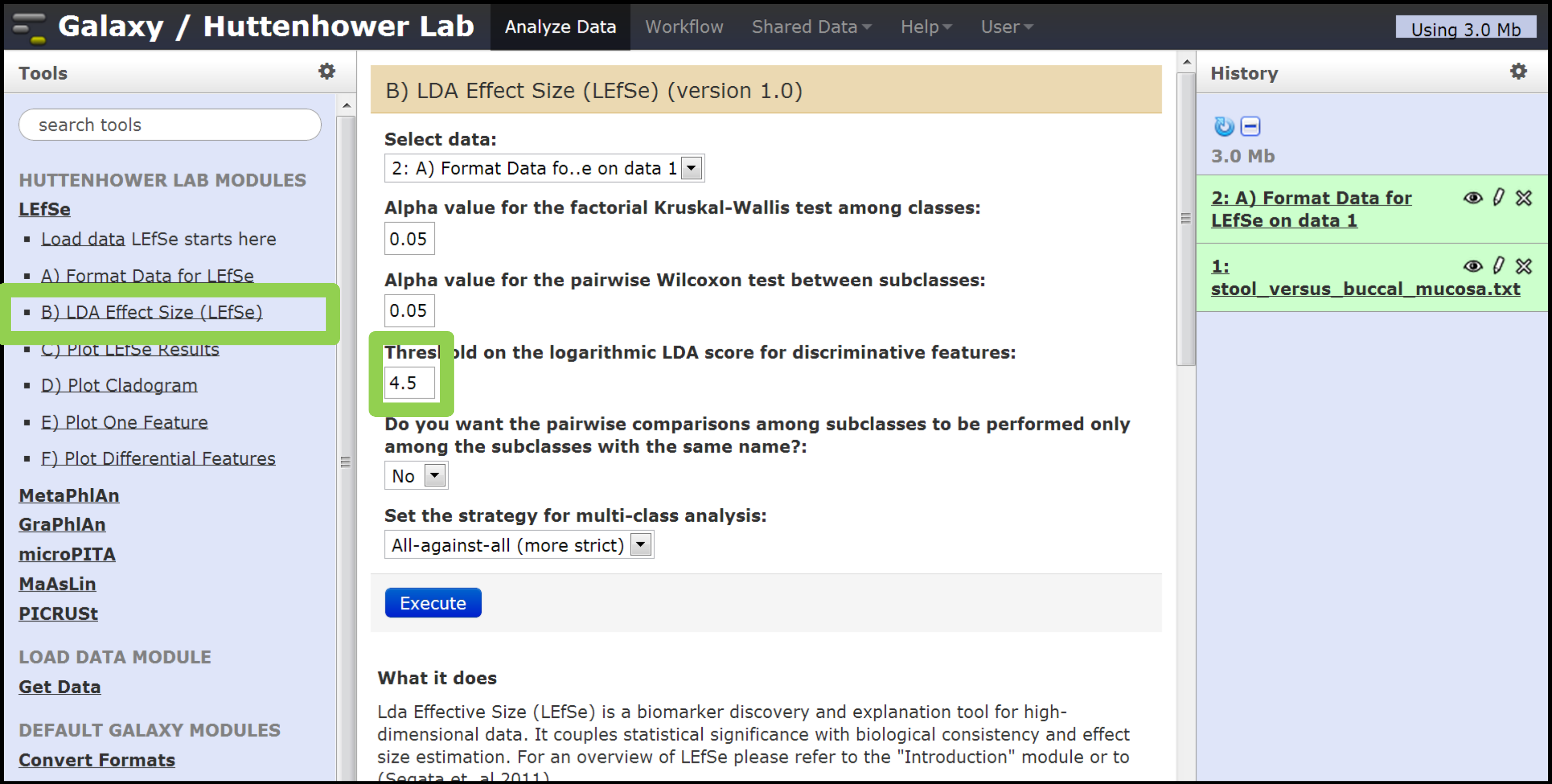
What it does

Preprocessing module for the biomarker discovery tool called LEfSe:

This module of LEfSe preprocesses metagenomic abundance data for the analyses to be carried out with the "Run LEfSe" module. This module is separated from the "Run LEfSe" because one may want to preprocess the data only once but run multiple analyses.

We're just using class (STSite, Stool versus Buccal_mucosa)
Set the other boxes to "no subclass" and "no subject"

Run the algorithm



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

Select data: 2: A) Format Data fo..e on data 1

Alpha value for the factorial Kruskal-Wallis test among classes: 0.05

Alpha value for the pairwise Wilcoxon test between subclasses: 0.05

Threshold on the logarithmic LDA score for discriminative features: 4.5

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

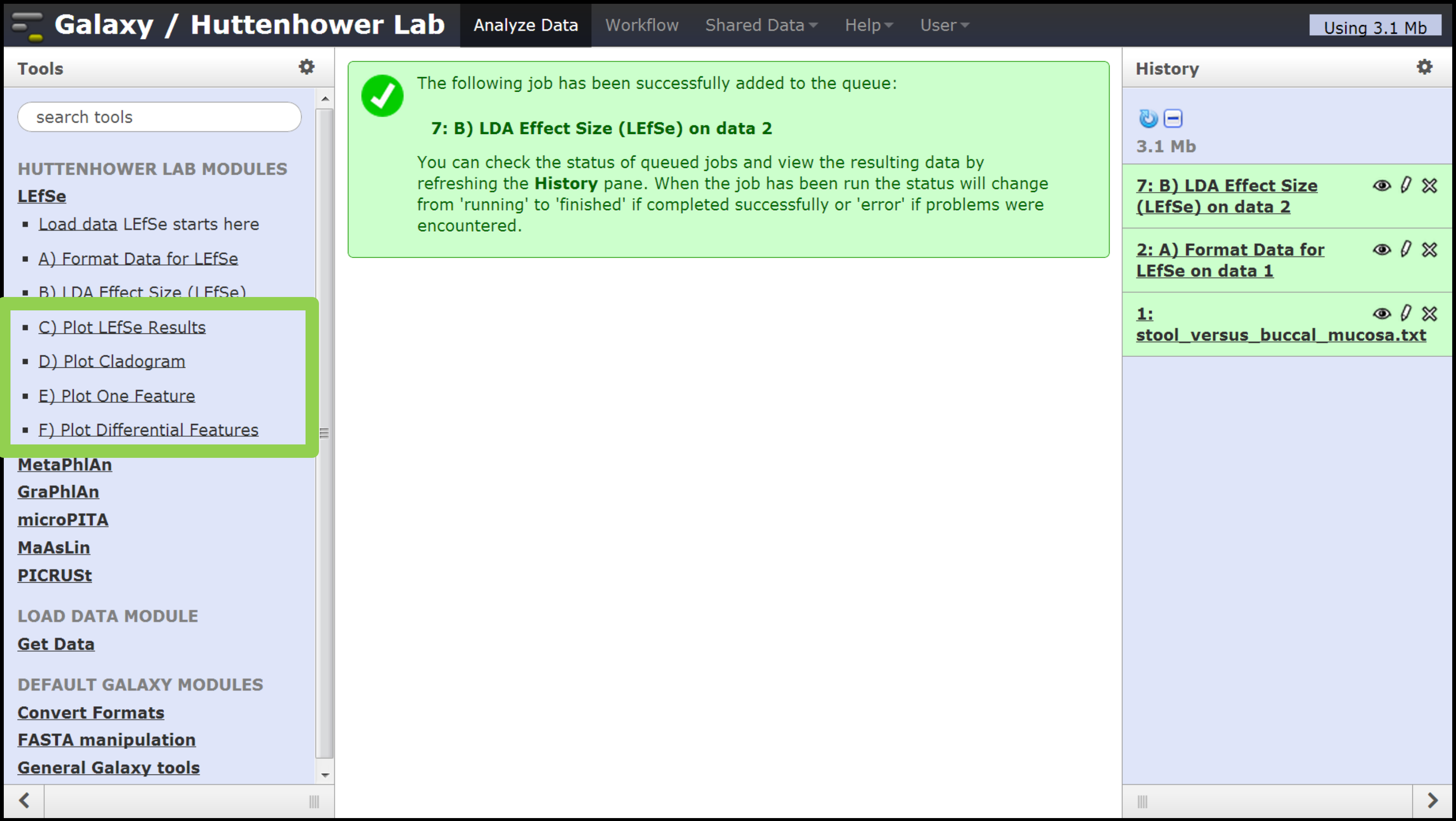
Set the strategy for multi-class analysis: All-against-all (more strict)

What it does

Lda Effective Size (LEfSe) is a biomarker discovery and explanation tool for high-dimensional data. It couples statistical significance with biological consistency and effect size estimation. For an overview of LEfSe please refer to the "Introduction" module or to (Segata et. al 2011)

Because the oral and gut communities are quite different, I adjusted the threshold from the default of 2 to 4.5 (logs) to only show the most extreme differences.

Visualize the results

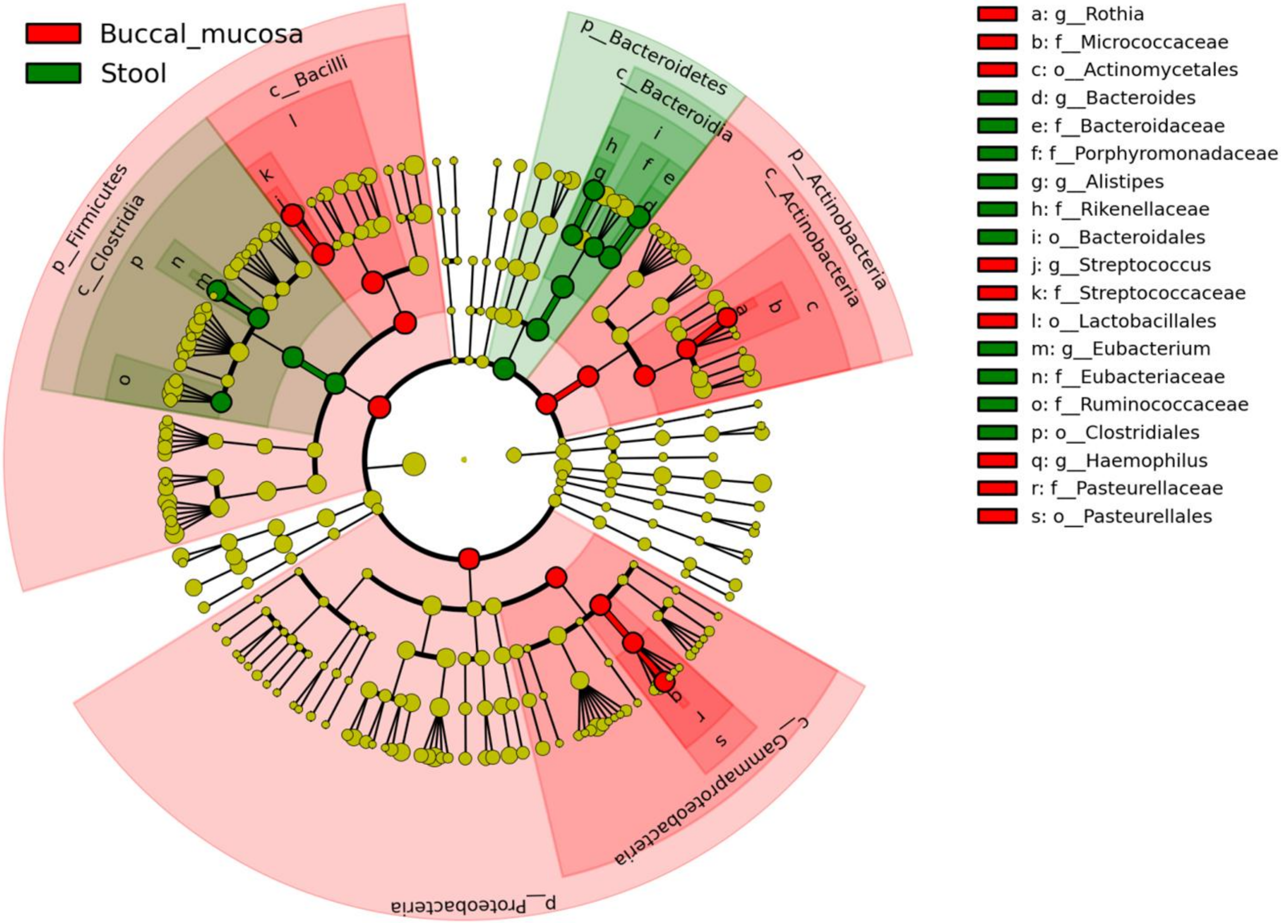


The following job has been successfully added to the queue:

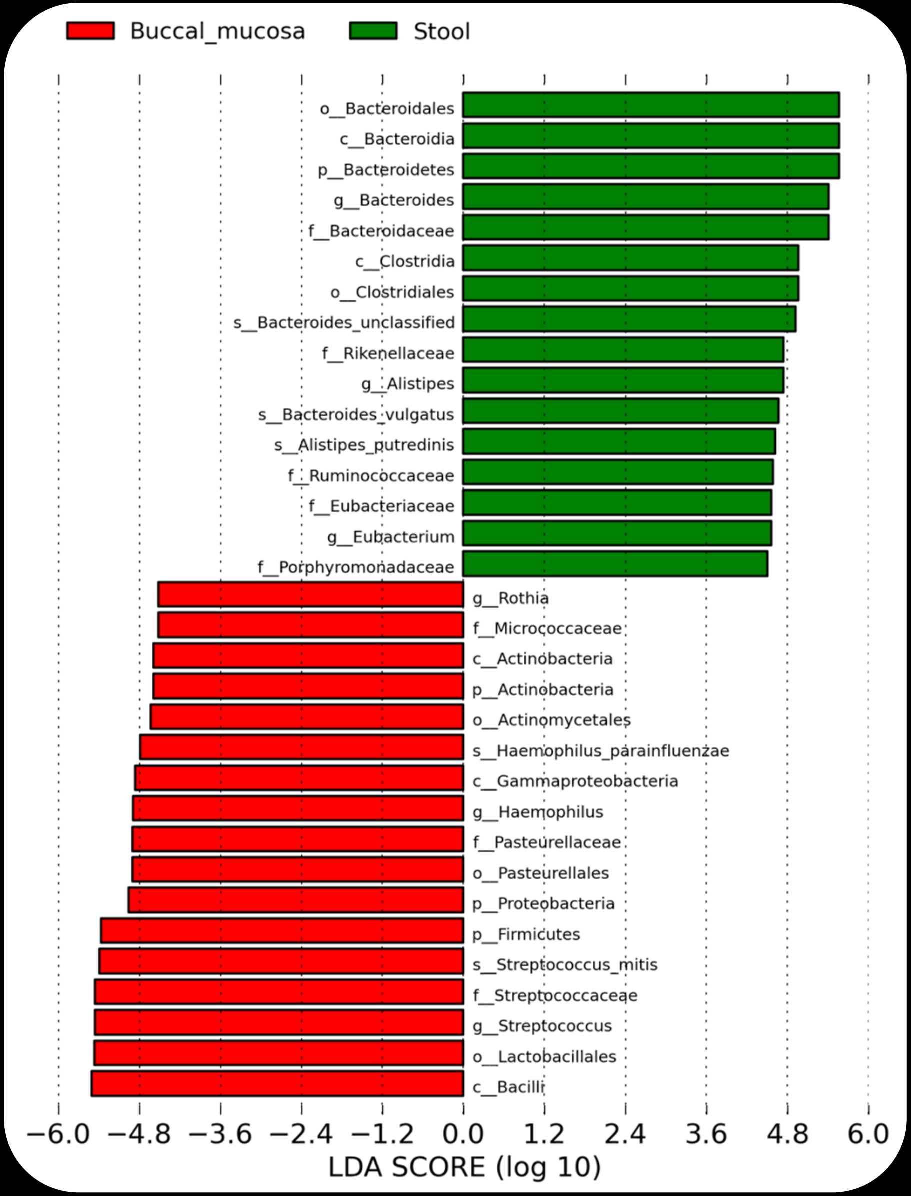
7: B) LDA Effect Size (LEfSe) on data 2

You can check the status of queued jobs and view the resulting data by refreshing the **History** pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

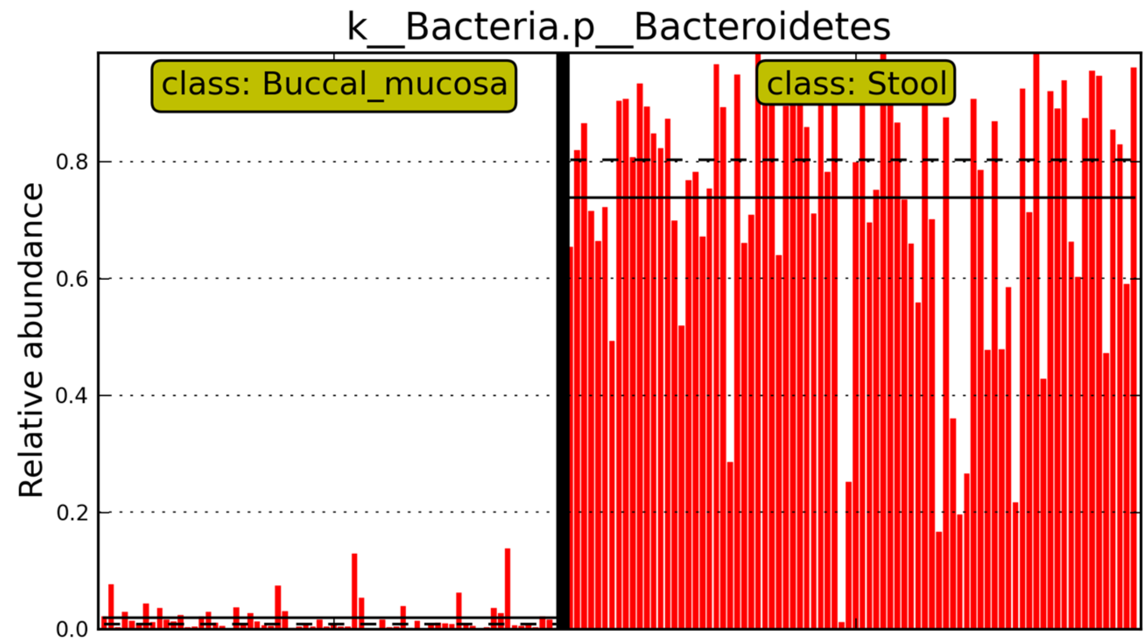
LEfSe Output: *Cladogram view*



LEfSe Output: *Raw Numbers*



LEfSe Output: *feature-specific report*



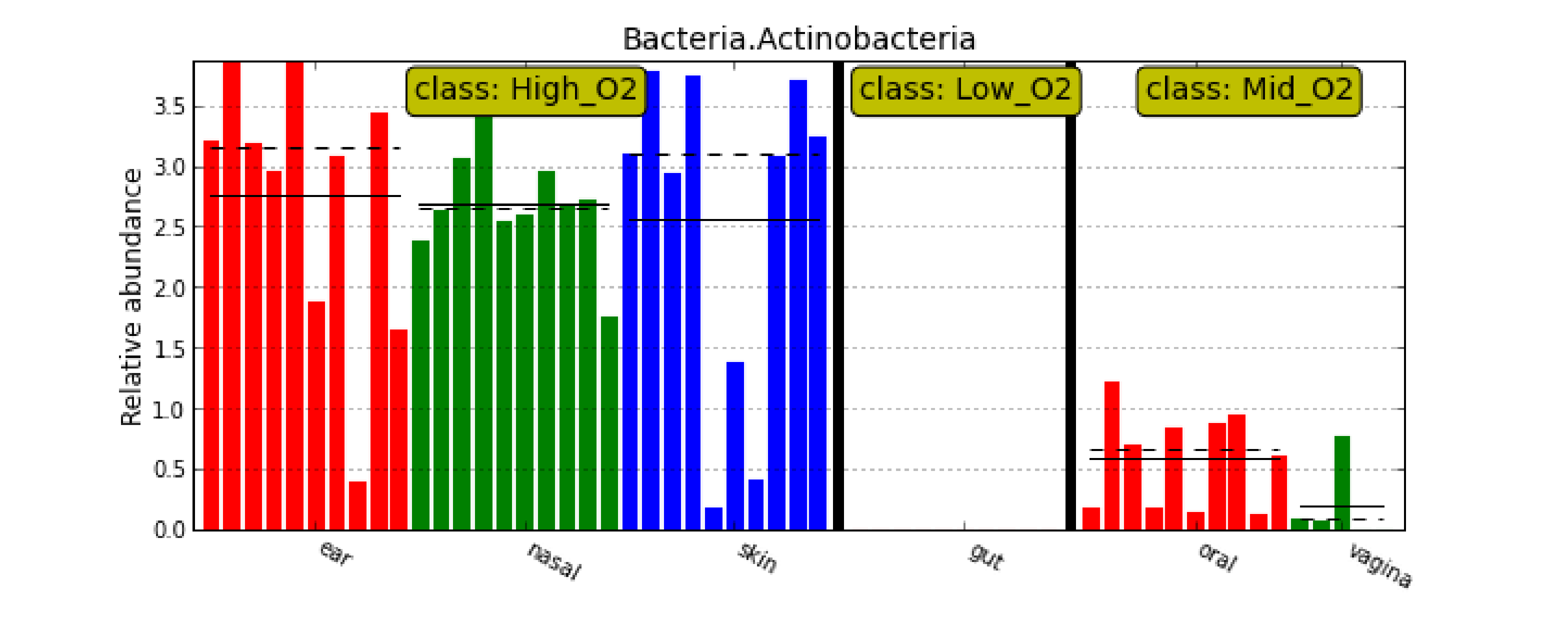
LEfSe at the command line

- Source code available at:
  https://bitbucket.org/nsegata/lefse
- Requires python and R with several additional packages installed (see README)
- Today's sample data available at:
  http://huttenhower.sph.harvard.edu/content/metaphlan-tutorial
- We'll look at a second example included with the LEfSe distribution involving subclasses

LEfSe (command line) demo

- Show LEfSe setup
- Show example folder
- Highlight LEfSe commands
- Show output
- Highlight more complicated subclass treatment

LEfSe (command line) demo



Tutorial Outline

- Introduction to MetaPhlAn
- MetaPhlAn Demo
- Introduction to LEfSe
- LEfSe Demo
- **Links, other tools, and Q&A**

If you’re thirsty for more…

- Tutorial materials available here:
  http://huttenhower.sph.harvard.edu/content/metaphlan-tutorial

- My email (Eric Franzosa):
  franzosa@hsph.harvard.edu

If you're thirsty for more…

- For more about our tools, check out: http://huttenhower.sph.harvard.edu/research

- Individual tool pages contain links to source code and documentation…

- …or try them out on Galaxy: http://huttenhower.sph.harvard.edu/galaxy/

If you’re thirsty for more…

- MetaPhlAn 2.0 coming soon with support for viruses, eukaryotes, and more bacteria/archaea

- PhyloPhlAn uses MetaPhlAn’s **core gene** identification pipeline to assist in tree building
  http://huttenhower.sph.harvard.edu/phylophlan

- MaAsLin is an evolution of LEfSe that can consider a wider number and variety of metadata
  http://huttenhower.sph.harvard.edu/maaslin

Thank you!

Curtis Huttenhower

Levi Waldron

Xochi Morgan

Tim Tickle

Vagheesh Narasimhan

Emma Schwager

Nicola Segata

Daniela Boernigen

Joseph Moon

Jim Kaminski

Craig Bielski

Brian Palmer

Ren Lu

Hufeng Zhou

Dirk Gevers
Kat Huang

Ramnik Xavier
Harry Sokol
Dan Knights
Moran Yassour

Human Microbiome Project

Owen White
Joe Petrosino
George Weinstock
Karen Nelson
Lita Proctor
Erica Sodergren
Anthony Fodor
Marty Blaser
Jacques Ravel
Pat Schloss

Sahar Abubucker
Brandi Cantarel
Alyx Schubert
Mathangi Thiagarajan
Beltran Rodriguez-Mueller
Makedonka Mitreva
Yuzhen Ye
Mihai Pop
Larry Forney
Barbara Methe

Bruce Birren
Doyle Ward
Mark Daly
Ashlee Earl

Wendy Garrett
Michelle Rooks

Rob Beiko
Morgan Langille

Rob Knight
Greg Caporaso
Jesse Zaneveld

Mark Silverberg
Boyko Kabakchiev
Andrea Tyler

Ruth Ley
Omry Koren

Jacques Izard
Katherine Lemon

Bruce Sands

http://huttenhower.sph.harvard.edu

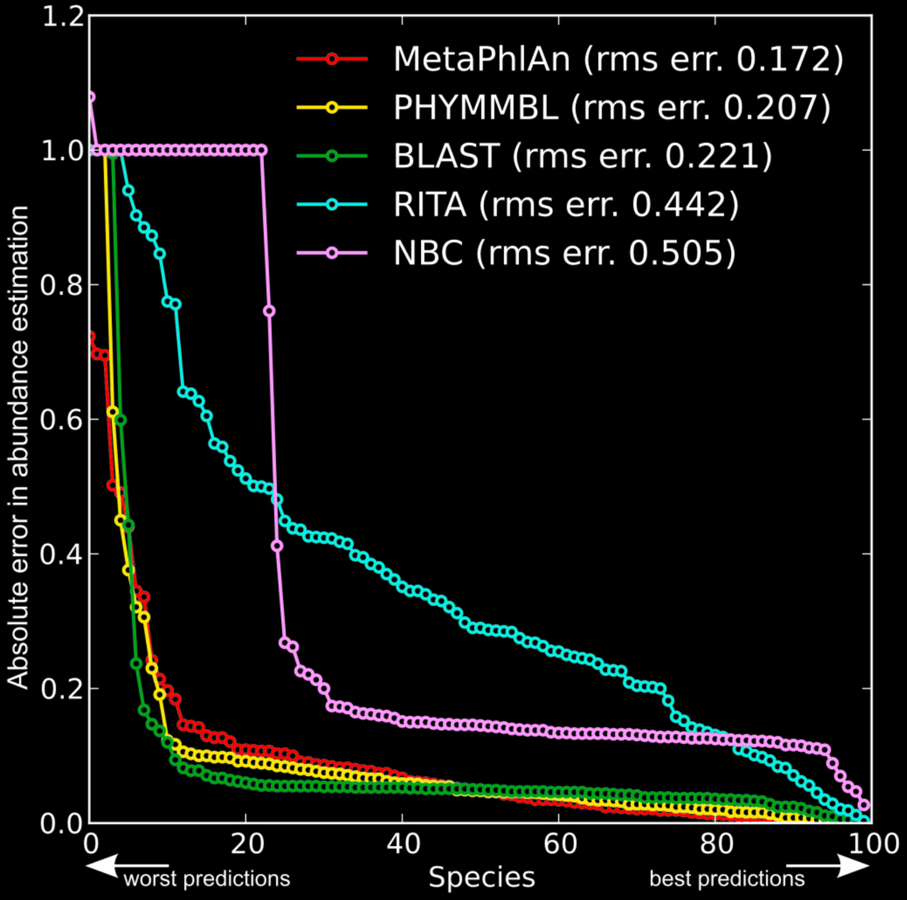
If you’re thirsty for more…

- Questions?

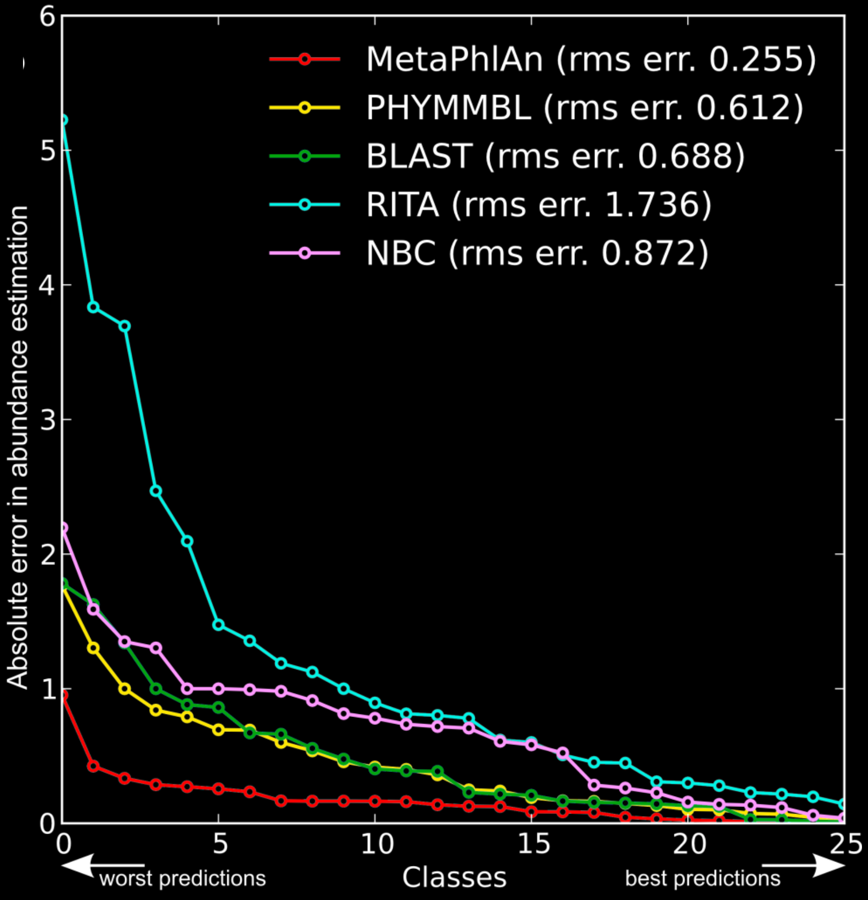
MetaPhlAn Statistics

- Representing 2,887 genomes (107 Archaea)
- 1,222 species, 652 genera, 278 families, 130 orders, 66 classed, 33 phyla, 2,383 total clades
- ≈2M total unique marker genes
- ≈400k “most representative” unique marker genes
- 231±107 markers per species (350 fixed max)
- The 400k database represent ≈4% of the total microbial sequence data available

Evaluation of accuracy (1)



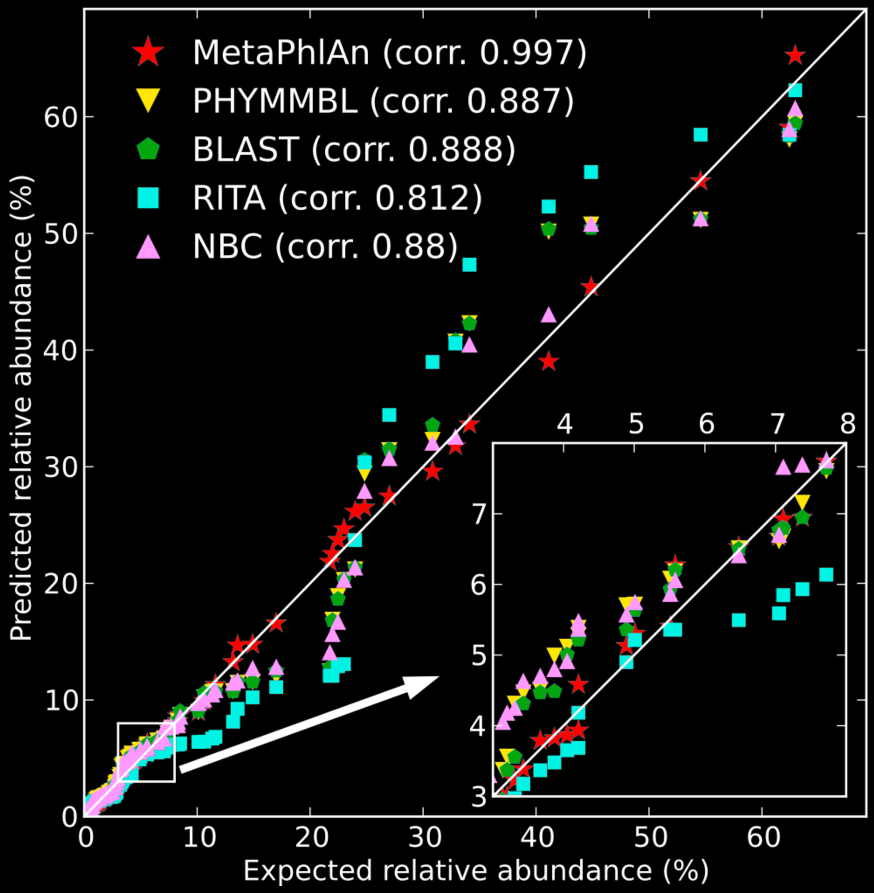
Species



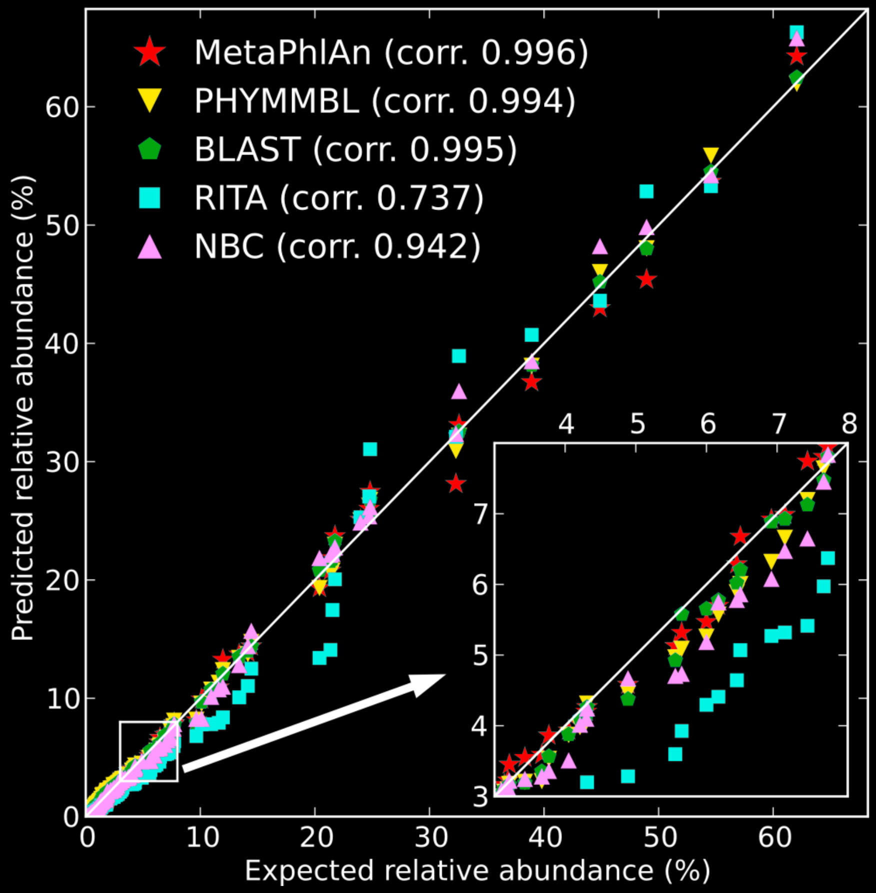
Classes

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

Evaluation of accuracy (2)



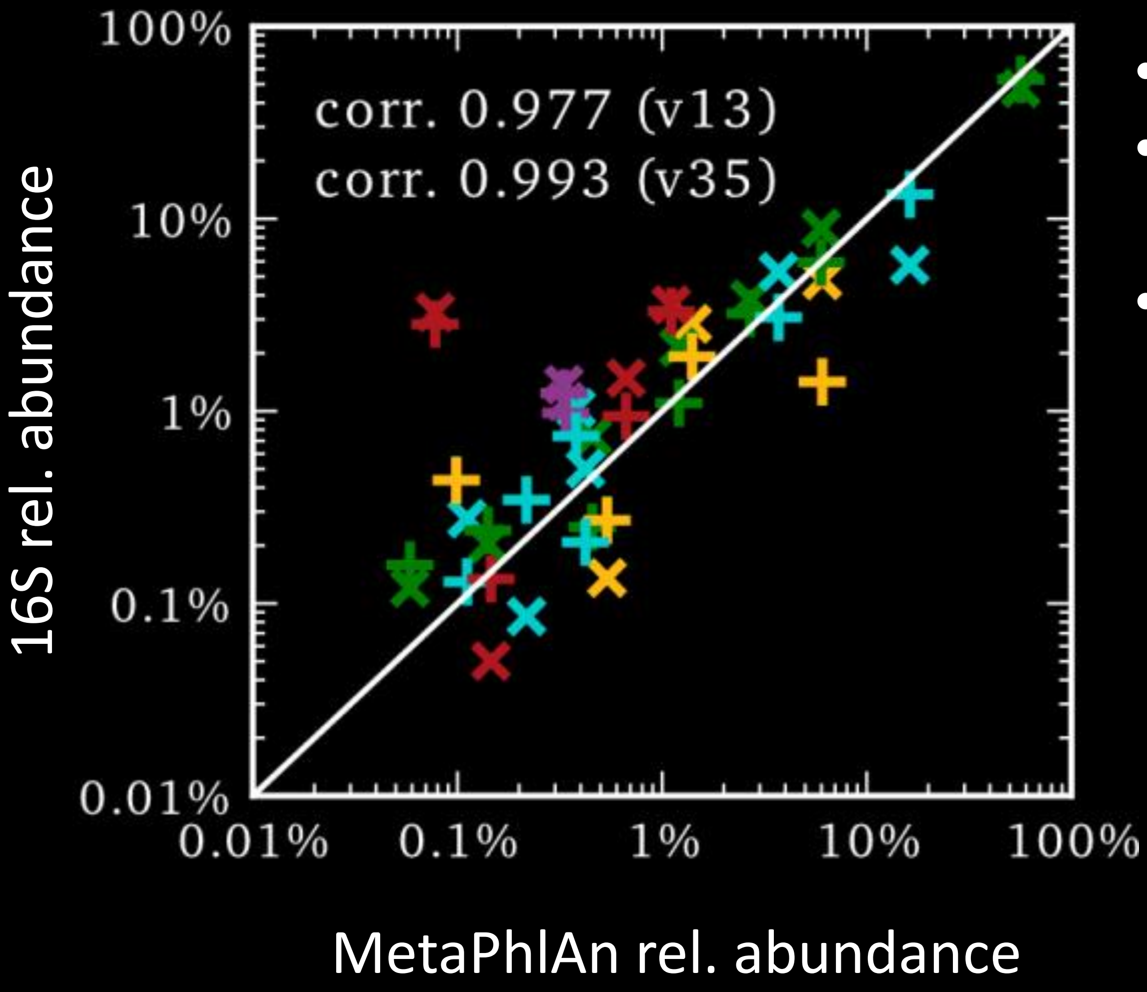
Species



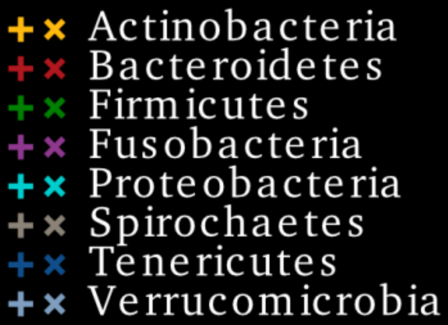
Classes

(Validation on low-complexity log-normally distributed synthetic metagenomes.)

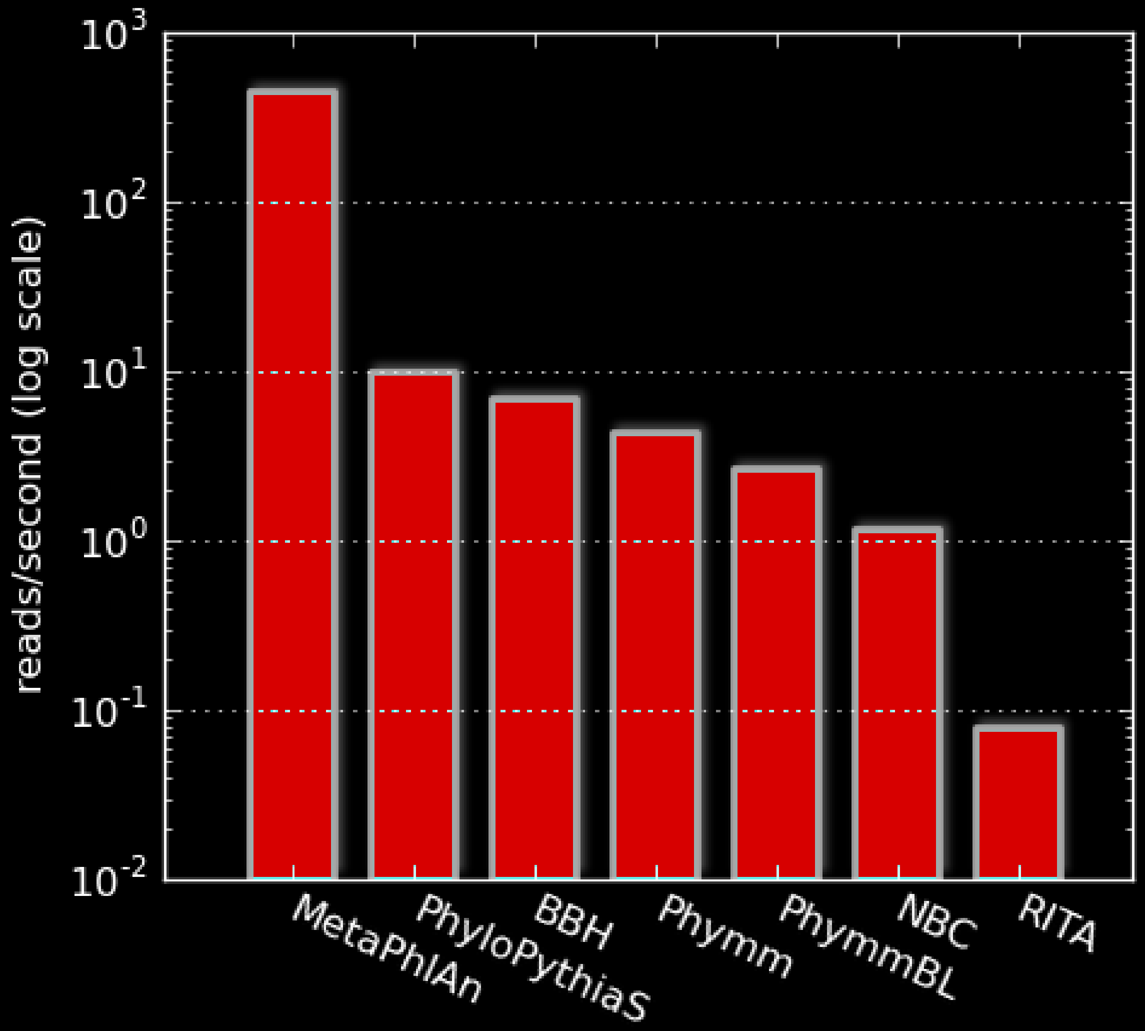
Evaluation of accuracy (3)



- *Buccal mucosa example*
- Genus-level comparison with 16S-based estimation
- v13 (+) & v35 (×) regions



Evaluation of performance



>50 times faster than existing methods

450 reads/sec (BLAST)

Up to 25,000 reads/sec (bowtie2)

Multi-threaded

Easily parallelizable