Meta'omic functional profiling with ShortBRED

Curtis Huttenhower

09-19-15

Harvard School of Public Health
Department of Biostatistics

U. Oregon

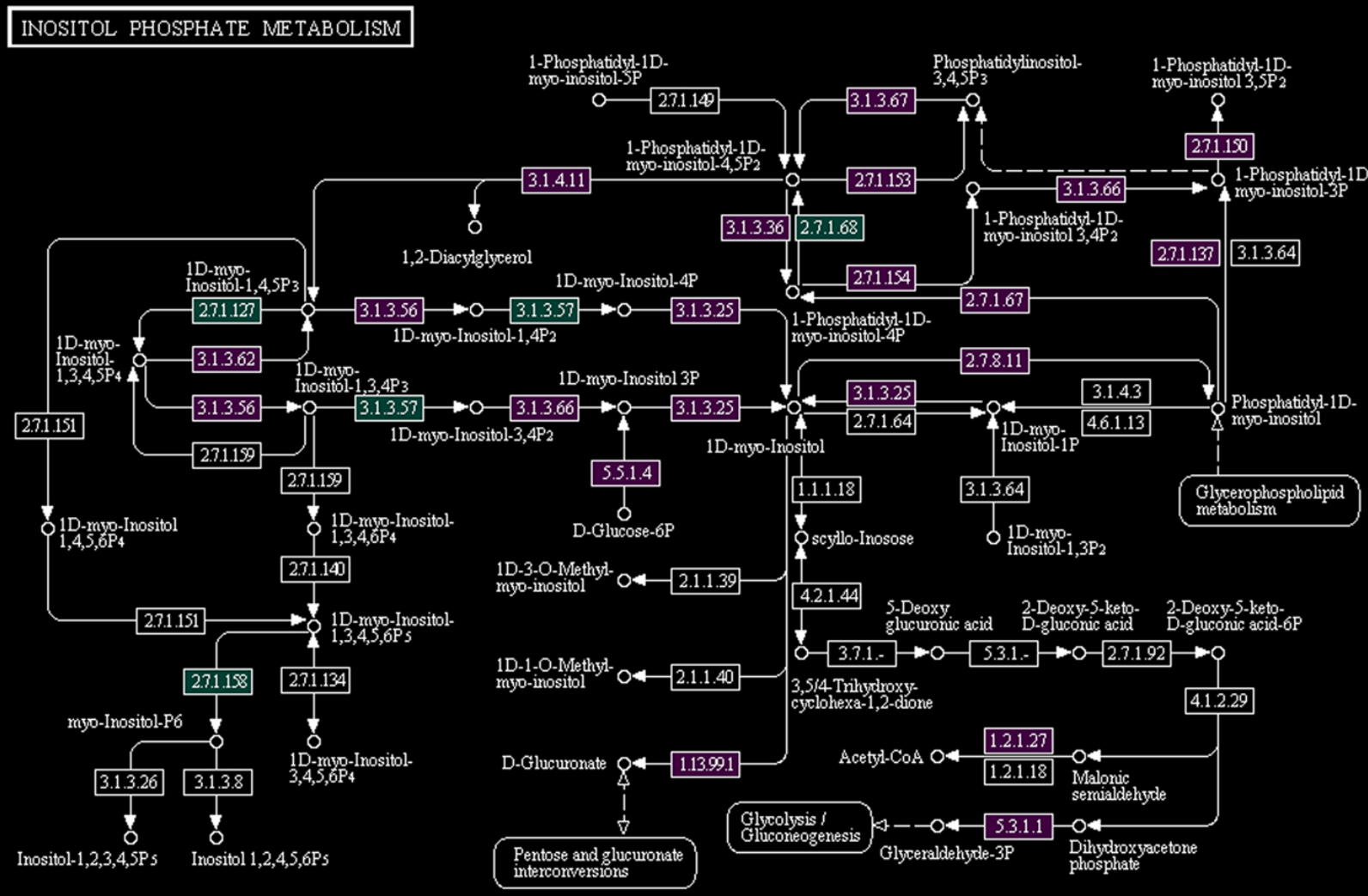
BROAD INSTITUTE

The two big questions…

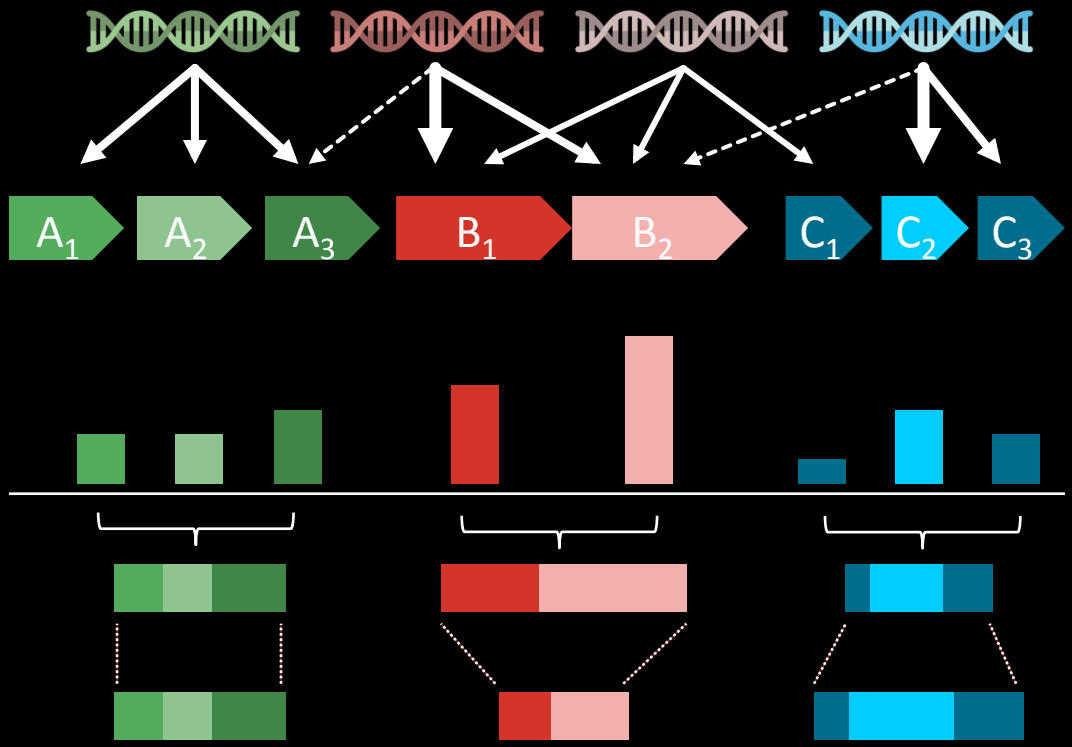
Who is there?
(taxonomic profiling)

What are they doing?
(functional profiling)

(What we mean by “function”)



HUMAnN
HMP Unified Metabolic Analysis Network



Short reads + protein families

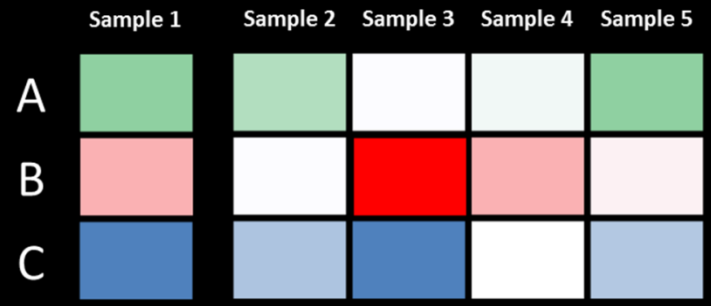
Translated BLAST search

$$c(g) = \frac{1}{|g|} \sum_{r} \frac{\sum_{a(r)} (1 - p_a) \Delta(a = g)}{\sum_{a(r)} 1 - p_a}$$

Weight hits by significance

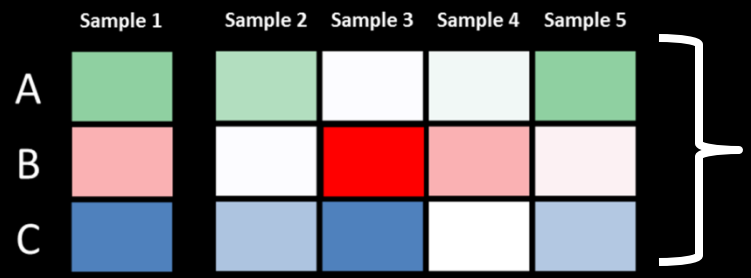
Sum over families

Adjust for sequence length

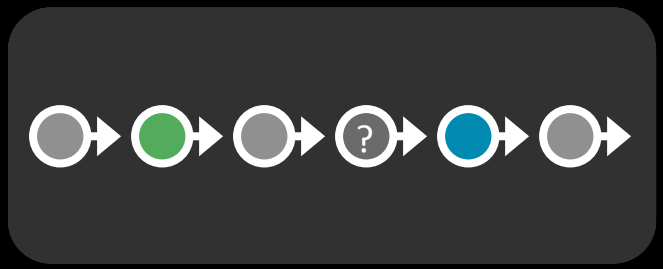


Repeat for each metagenomic or metatranscriptomic sample

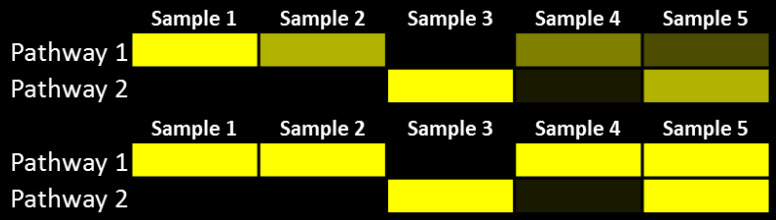
HUMAnN
HMP Unified Metabolic Analysis Network



Millions of hits are collapsed into thousands of gene families (KOs) (*still a large number*)

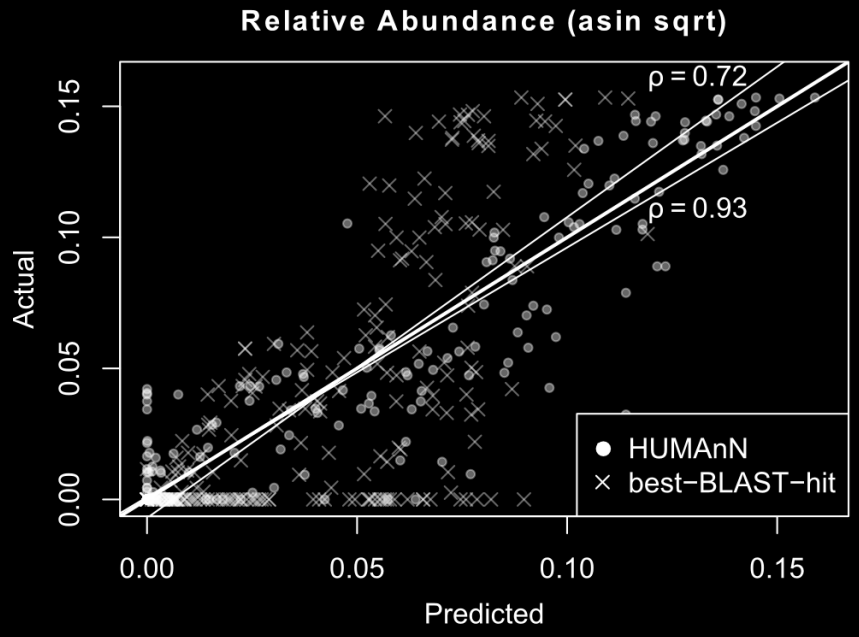


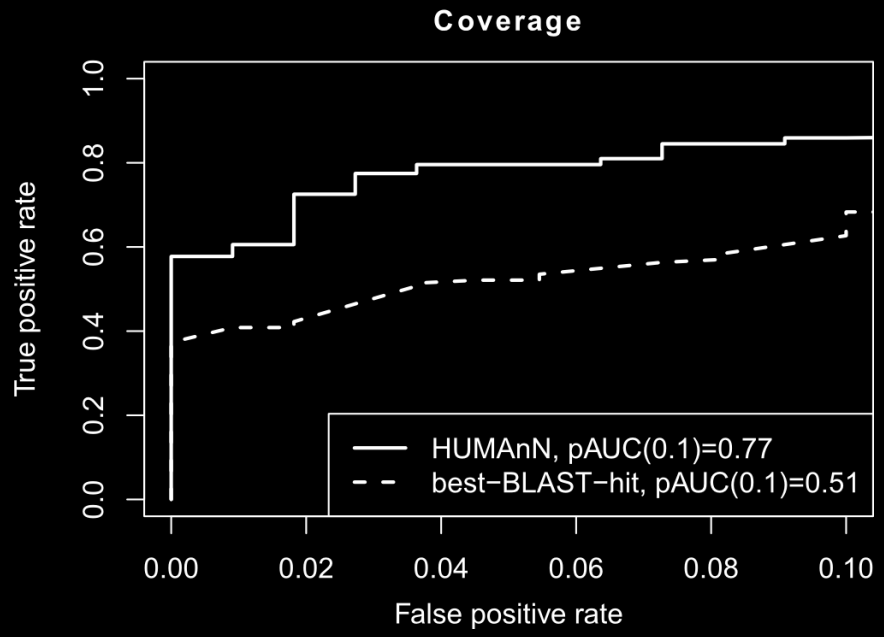
- Map genes to KEGG ~~pathways~~ modules
- Use MinPath (Ye 2009) to find simplest pathway explanation for observed genes
- Remove pathways unlikely to be present due to low organismal abundance
- Smooth/fill gaps



Collapsing KO abundance into KEGG module abundance (or presence/absence) yields a smaller, more tractable feature set

HUMAnN accuracy

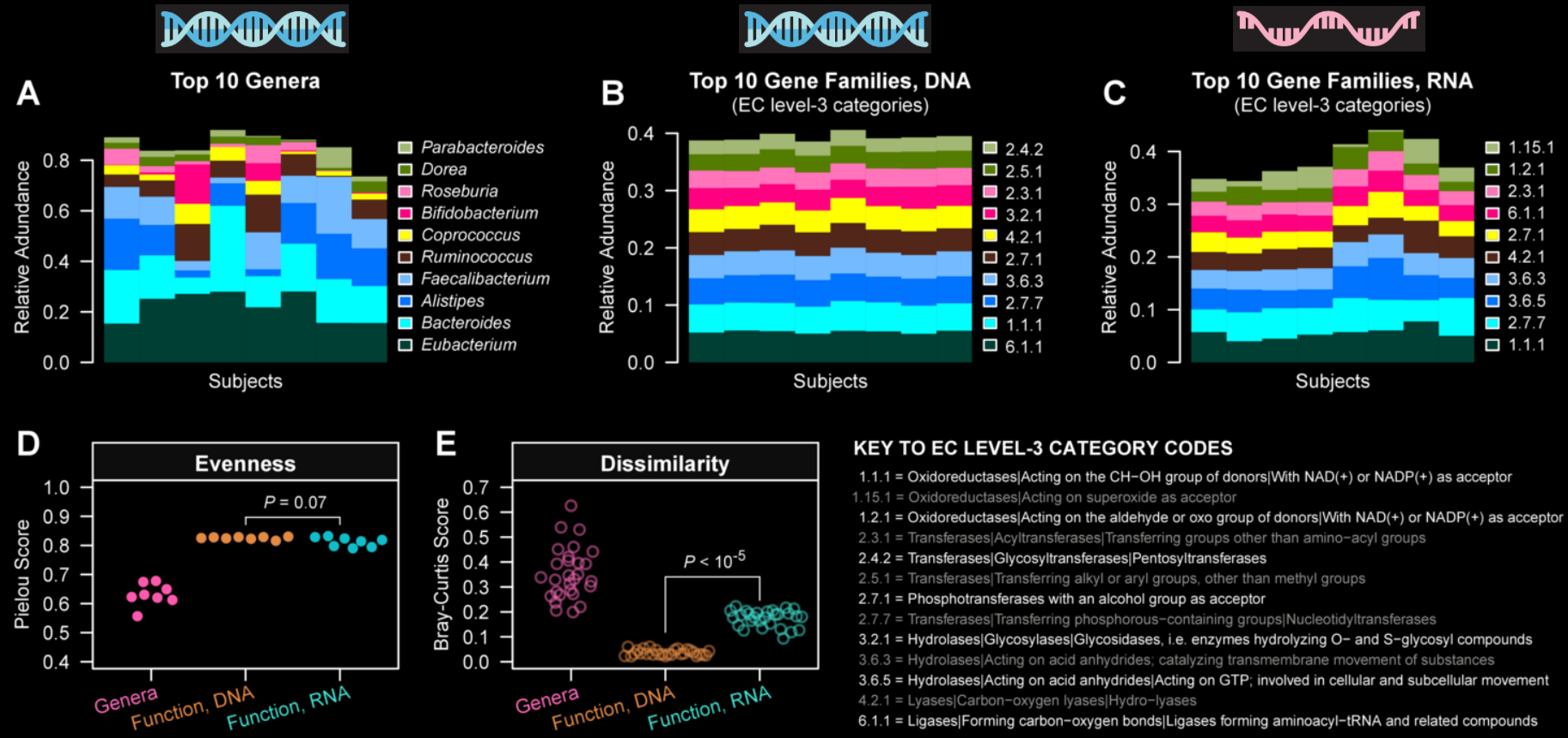




Validated against synthetic metagenome samples
(similar to MetaPhlAn validation)

Gene family abundance and pathway presence/absence
calls beat naïve best-BLAST-hit strategy

HUMAnN in action



KEY TO EC LEVEL-3 CATEGORY CODES

1.1.1 = Oxidoreductases|Acting on the CH−OH group of donors|With NAD(+) or NADP(+) as acceptor
1.15.1 = Oxidoreductases|Acting on superoxide as acceptor
1.2.1 = Oxidoreductases|Acting on the aldehyde or oxo group of donors|With NAD(+) or NADP(+) as acceptor
2.3.1 = Transferases|Acyltransferases|Transferring groups other than amino−acyl groups
2.4.2 = Transferases|Glycosyltransferases|Pentosyltransferases
2.5.1 = Transferases|Transferring alkyl or aryl groups, other than methyl groups
2.7.1 = Phosphotransferases with an alcohol group as acceptor
2.7.7 = Transferases|Transferring phosphorous−containing groups|Nucleotidyltransferases
3.2.1 = Hydrolases|Glycosylases|Glycosidases, i.e. enzymes hydrolyzing O− and S−glycosyl compounds
3.6.3 = Hydrolases|Acting on acid anhydrides; catalyzing transmembrane movement of substances
3.6.5 = Hydrolases|Acting on acid anhydrides|Acting on GTP; involved in cellular and subcellular movement
4.2.1 = Lyases|Carbon−oxygen lyases|Hydro−lyases
6.1.1 = Ligases|Forming carbon−oxygen bonds|Ligases forming aminoacyl−tRNA and related compounds

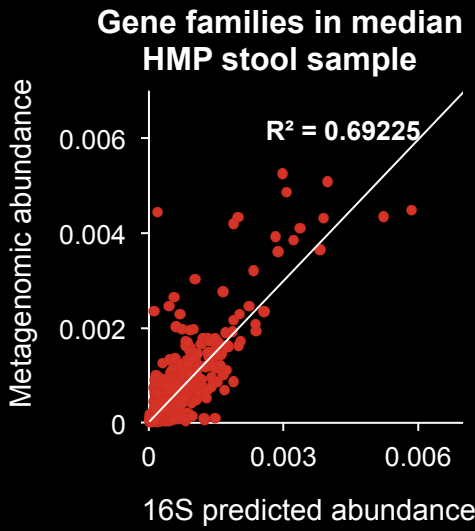
Franzosa et al. *PNAS* 11:E2329-38 (2014)

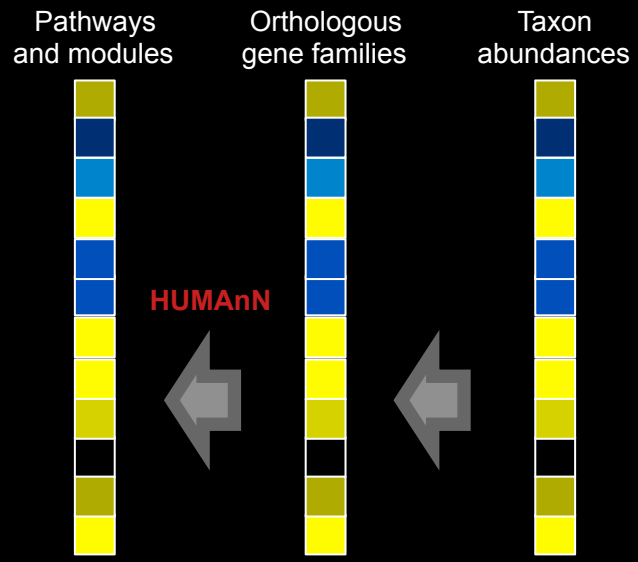
PICRUSt: Inferring community metagenomic potential from marker gene sequencing

With Rob Knight, Rob Beiko

One can recover ***general*** community function with reasonable accuracy from 16S profiles.

http://picrust.github.com





Jesse Zaneveld

Morgan Langille

What's there: ShortBRED

Jim Kaminski

- **ShortBRED** is a tool for quantifying protein families in metagenomes
  - Short Better REad Dataset

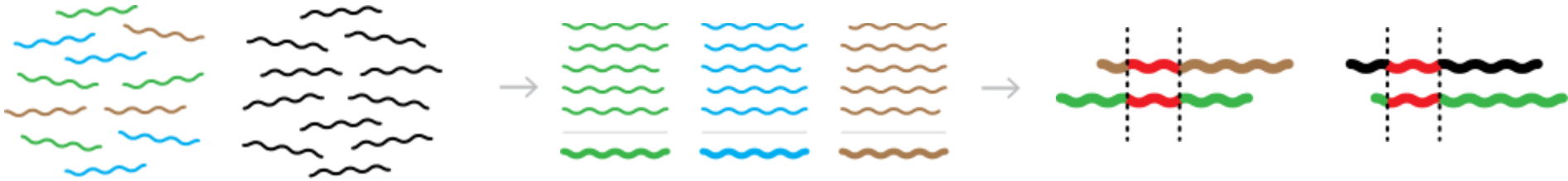
- Inputs:
  - FASTA file of proteins of interest
  - Large reference database of protein sequences (FASTA or blastdb)
  - Metagenomes (FASTA/FASTQ nucleotide files)
- Outputs:
  - Short, unique markers for protein families of interest (FASTA)
  - Relative abundances of protein families of interest in 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 and 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

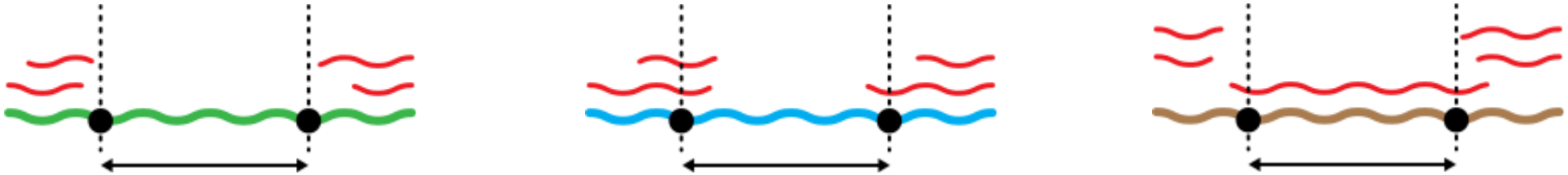


Prots of interest

Reference database

Cluster into families

Identify short, common regions

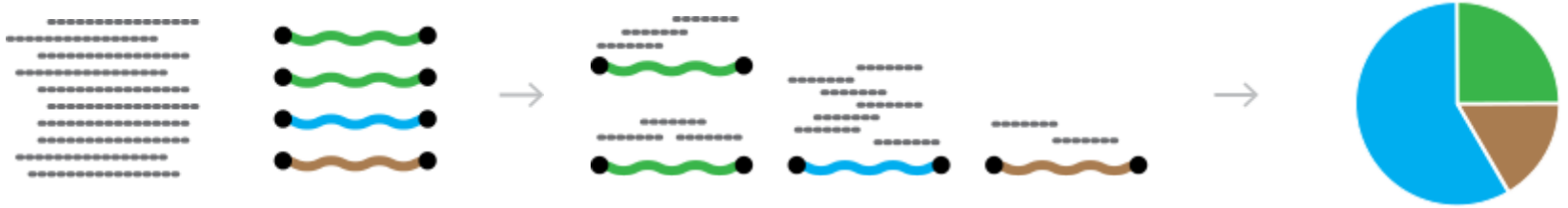


True Marker

Junction Marker

Quasi Marker

What's there: ShortBRED family quantification



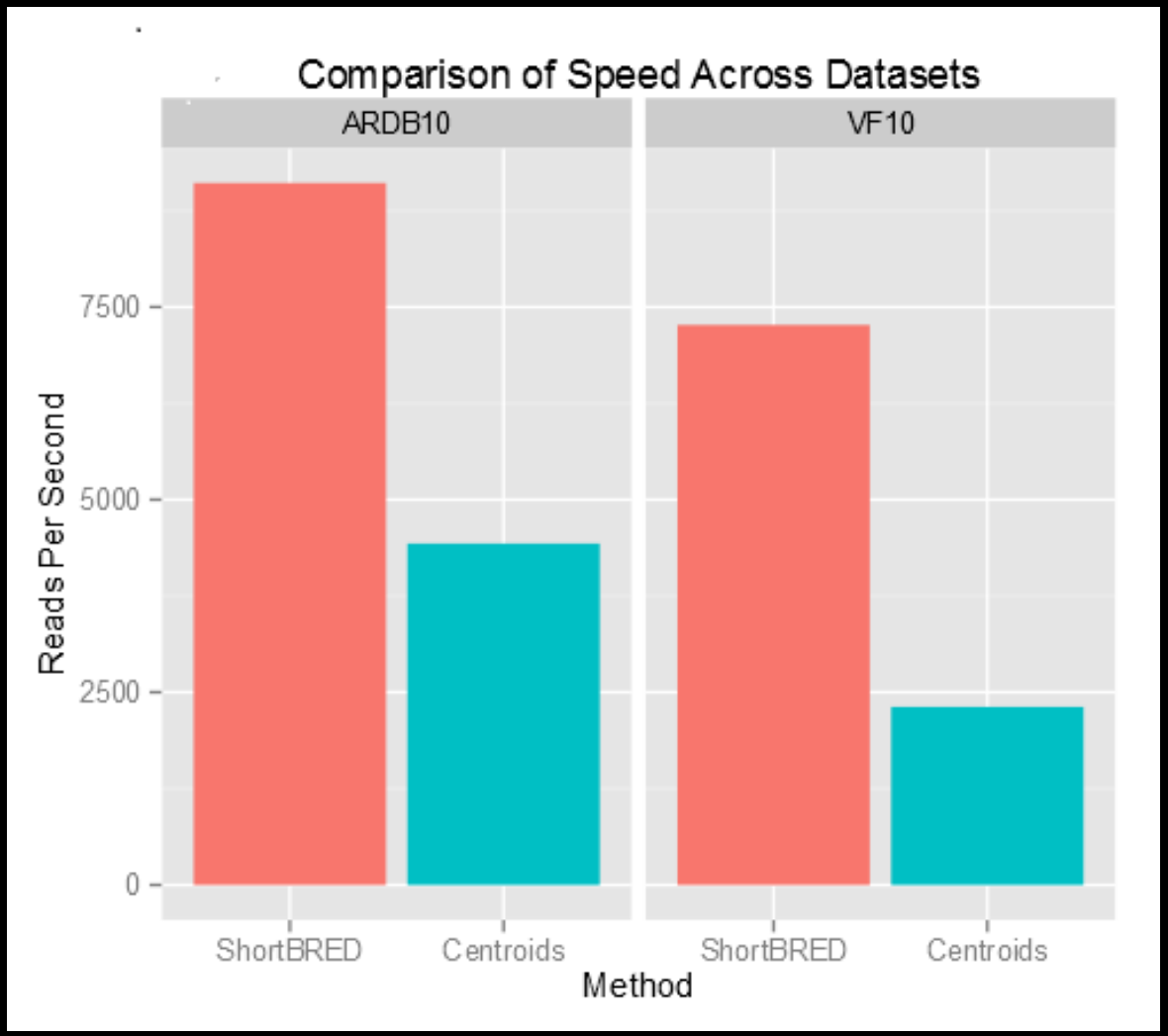
Metagenome reads

ShortBRED markers

Translated search for high ID hits

Normalize relative abundances

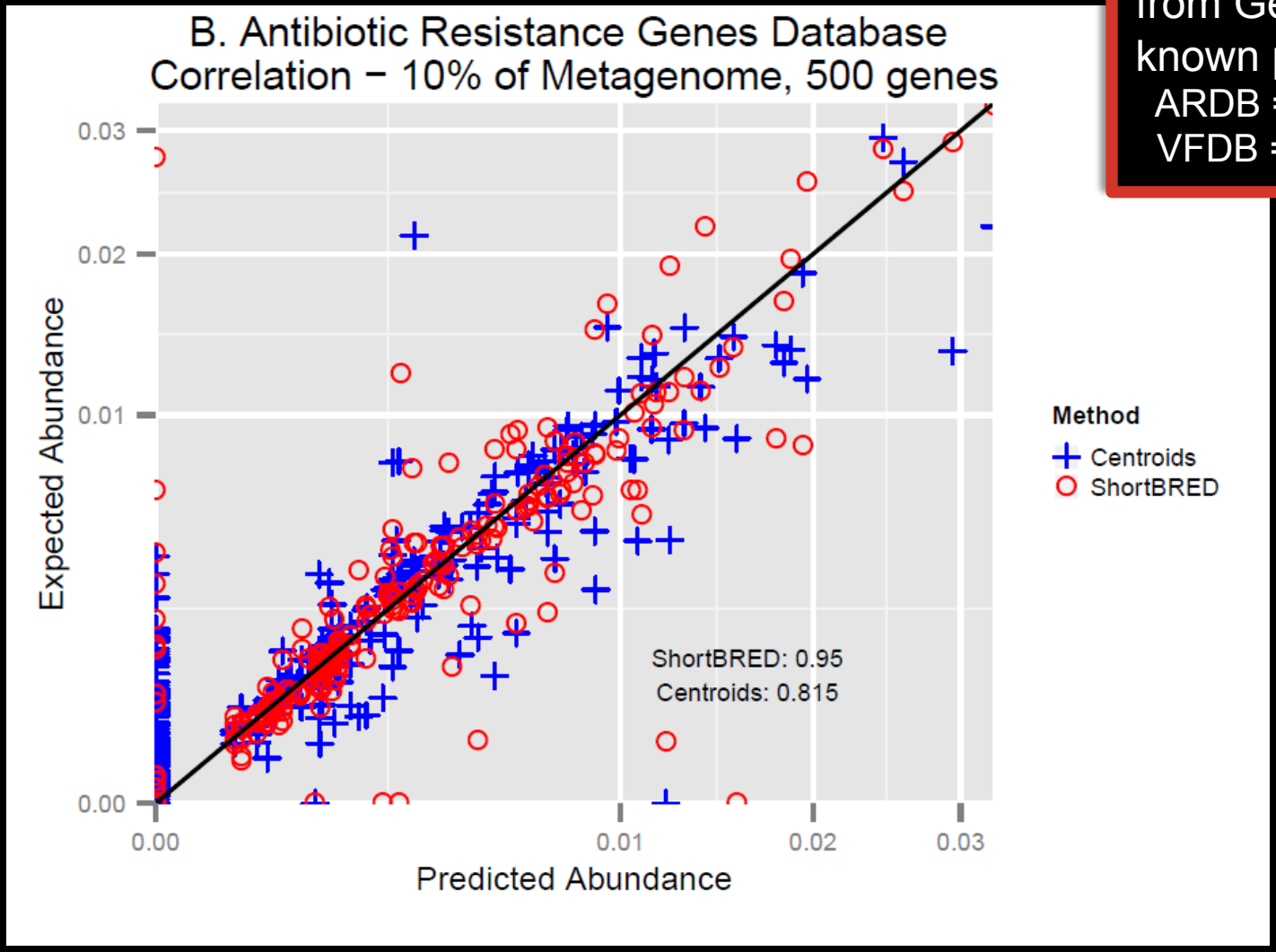
What's there: ShortBRED's fast



Six synthetic metagenomes from GemSim, spiked with known proteins of interest:
ARDB = Antibiotic Resistance
VFDB = Virulence Factors

What's there: ShortBRED's accurate

Six synthetic metagenomes from GemSim, spiked with known proteins of interest:
ARDB = Antibiotic Resistance
VFDB = Virulence Factors

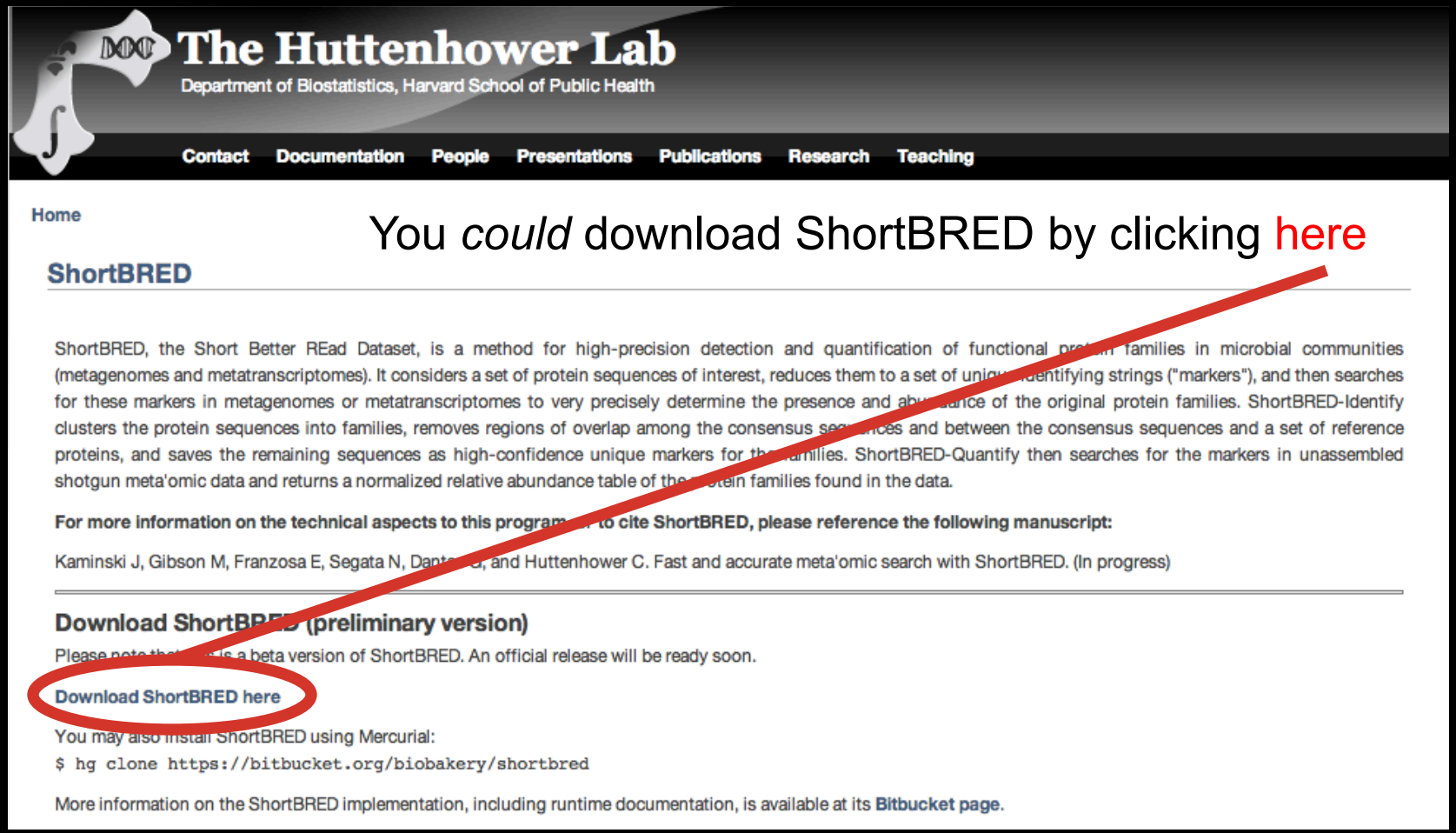


Setup notes reminder

- 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

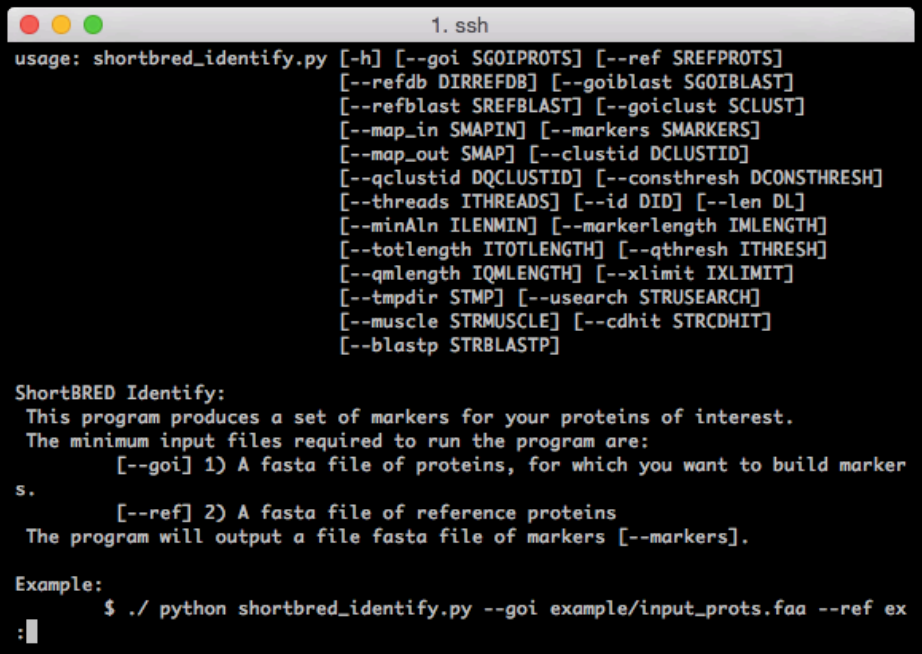
- ShortBRED is available at http://huttenhower.sph.harvard.edu/shortbred



From the command line...

- But don't!
  - Instead, we've installed ShortBRED already for you
- To see what you can do, run:

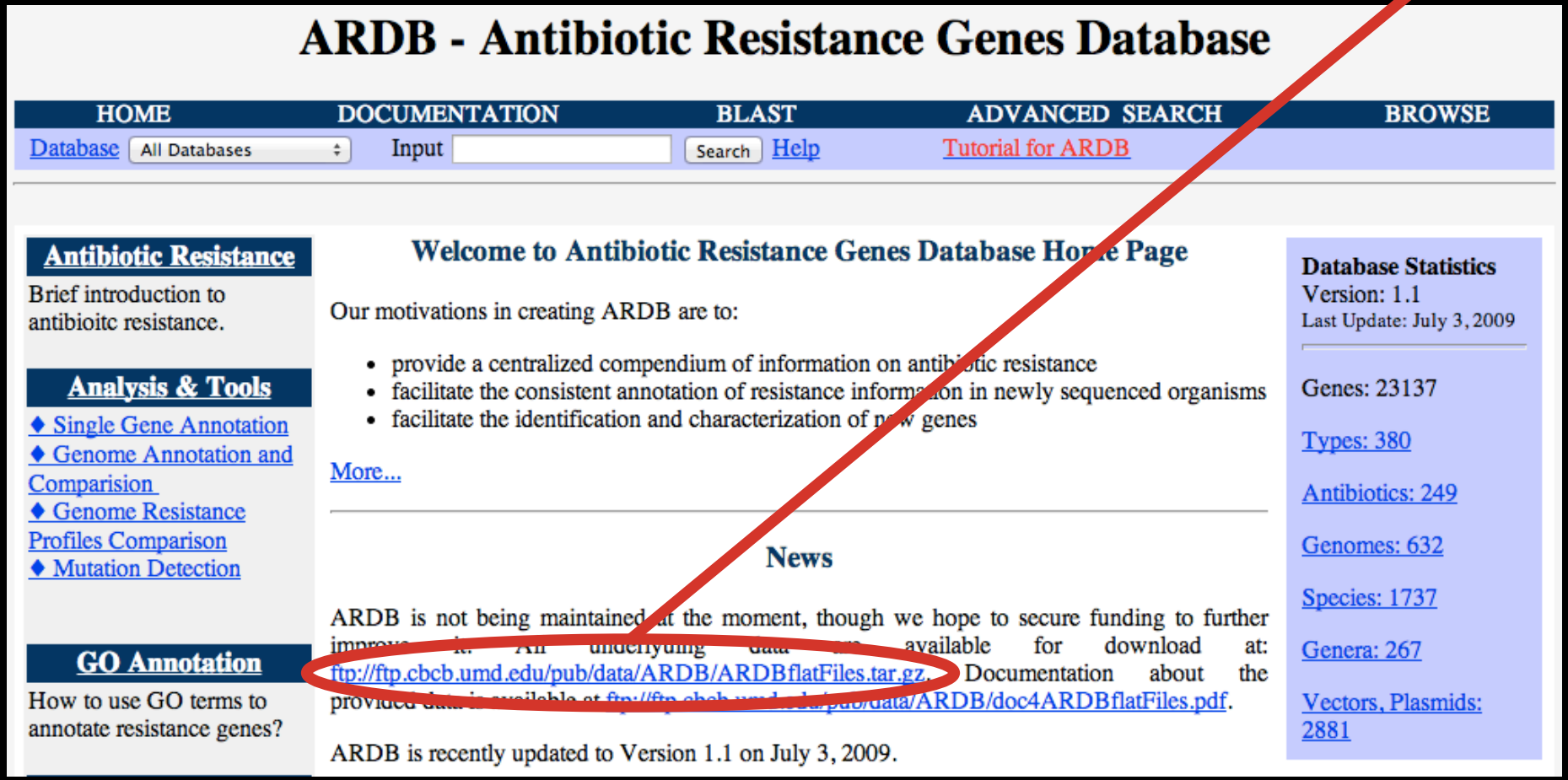
```
shortbred_identify.py -h | less
shortbred_quantify.py -h | less
```



Getting some annotated protein sequences

You could download the ARDB protein sequences here

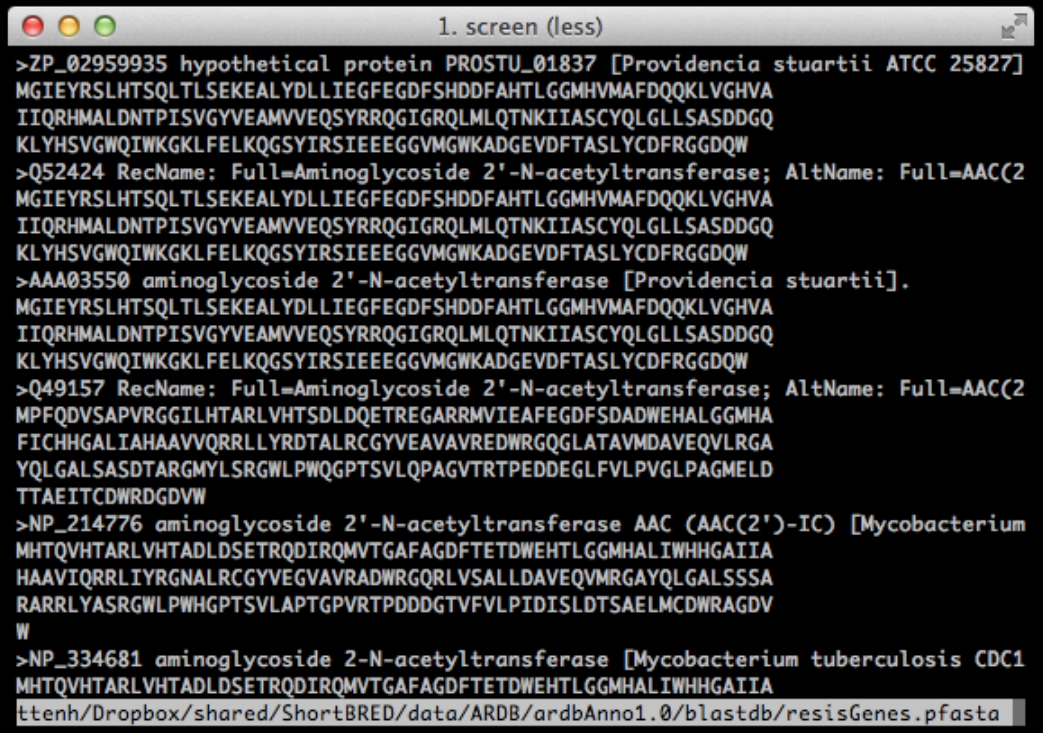
- Go to http://ardb.cbcb.umd.edu



From the command line...

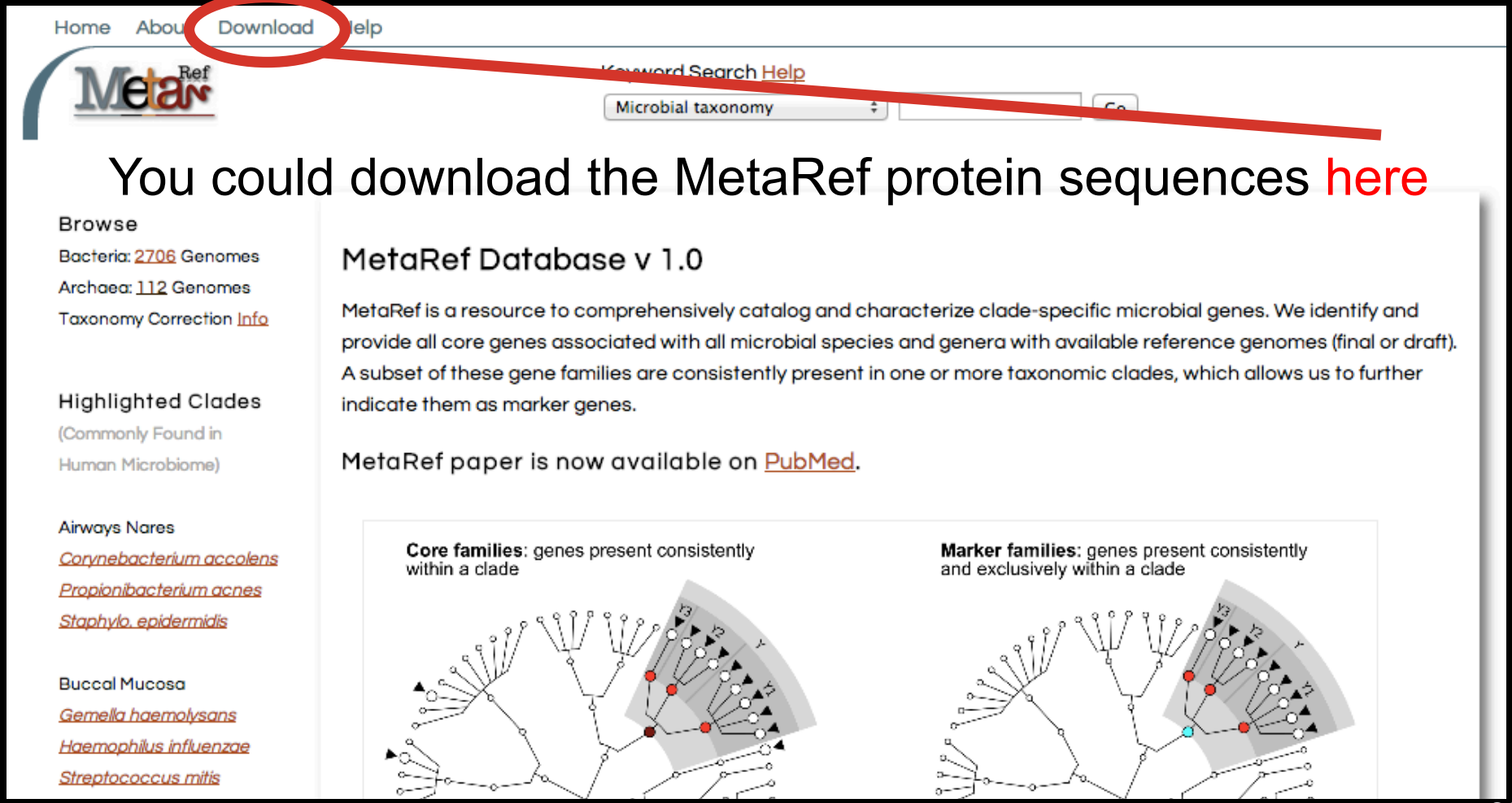
- But don't!
  - Instead, we've downloaded the important file for you
- Take a look by running:

```
less ~/workshop_data/metagenomics/biobakery/data/resisGenes.pfasta
```



Getting some reference protein sequences

- Go to http://metaref.org



Running ShortBRED-Identify

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

```
ln -s /usr/bin/cdhit /home/ubuntu/Programs/cd-hit
shortbred_identify.py
  --goi ~/workshop_data/metagenomics/biobakery/data/resisGenes.pfasta
  --ref ~/workshop_data/metagenomics/biobakery/software/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:

```
~/workshop_data/metagenomics/biobakery/results/shortbred/ardb_markers.faa
```

  - It will produce lots of status output as it runs

ShortBRED markers

3. [screen 3: bash] chuttenhower@class:/class/stamps-software/biobake...

```
>AAY52010_TM_#01
ISILILCRVML
>AAY52010_TM_#02
DKQIELSAEM
>AAY52010_TM_#03
KLNTLKRTLEKRE
>AAY52010_TM_#04
VVMYLAHDIKTPLTS
>AAY52010_TM_#05
LLDEAPDMP
>AAY52010_TM_#06
KAYRLEQLID
>AAY52010_TM_#07
IDLYYMLVQM
>AAY52010_TM_#08
DKLARVFNNIL
>AAY52010_TM_#09
IFEKFYRLD
>AAY52010_TM_#10
HGGQIYAESN
>AAD51345_TM_#01
AVSLLGLLAILILPVDR
>AAD51345_TM_#02
IRATYTGASSQTVENAVTQVIEQSQQSLDHLMYMTSTSASDGSAQVNLVFAT
ardb_markers.faa
```

True Markers
at the top

ShortBRED markers

3. [screen 3: bash] chuttenhower@class:/class/stamps-software/biobake...

```
>AAA25688_TM_#02
PTQLNKGLGTRLVRALVELLFSDPTVTKIQTDPTPNNH
>YP_277581_TM_#01
MSMIYITLNIIAYVIDVRSLILDVRRLVFS
>YP_277581_TM_#02
NILNCMDDSVIAFTVIIQLGAILSITKIFWSQLYGMSMICIKKIFFKQHDDHNHLCIRHI
FLGTFPGIMLGMIFYEKIGLIFELTYIMYGLIIGGIFLLVGELCASKEPRVSRINNITYL
>YP_277581_TM_#03
FSRAGATIGGGLVVGLDRRISS
>YP_277581_TM_#04
SAVLTLYHYRSCIGLMDVLLLIAGSATAFFIALFTVRYFLKIVKNVSLIPFAIYRFLLAG
GIYWGLMT
>1112175A_JM_#01__[1112175A_w=0.486,YP_001103000_w=0.143,YP_001103000_w=0.371]
LFEWEFVEKVDSAIMRLRRRAEPLLEGAALERYE
>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
>ZP_01817983_JM_#01__[ZP_01817983_w=0.493,YP_001694417_w=0.362,YP_001694417_w=0.
TLTGPFIGGFIKEDFQPVAKEKAIPTKELFTSVK
(END)
```

Junction/Quasi Markers
at the bottom

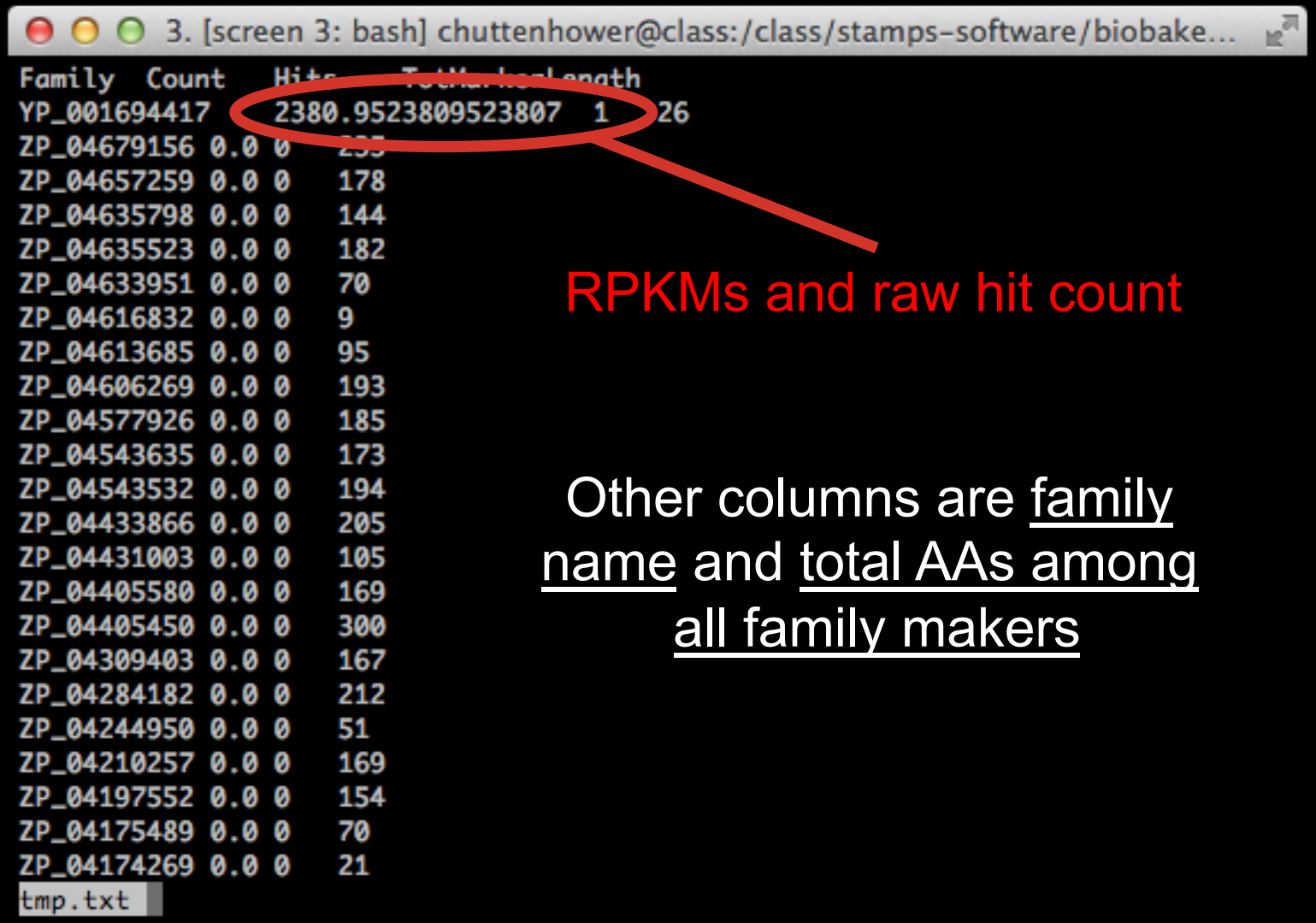
Running ShortBRED-Quantify

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

```
shortbred_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 just a few seconds
  - It will again produce lots of status output as it runs

ShortBRED marker quantification



RPKMs and raw hit count

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

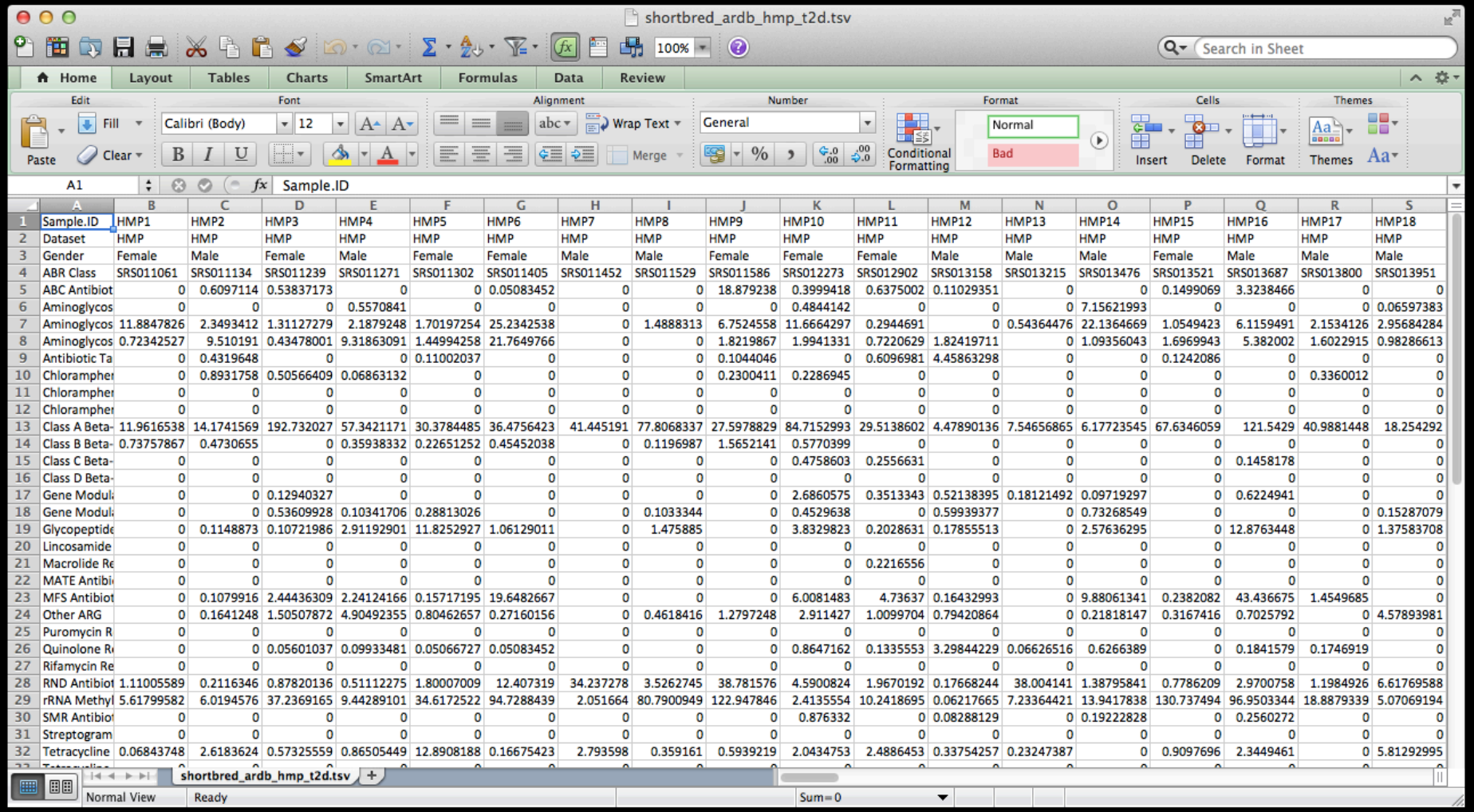
AR proteins in the human gut

- That's boring! Let's get some real data
- `scp` the file to your own computer (optional):

```
~/workshop_data/metagenomics/biobakery/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 2014)
  - Summing each sample's RPKMs for families in each ARDB resistance class

AR proteins in the human gut

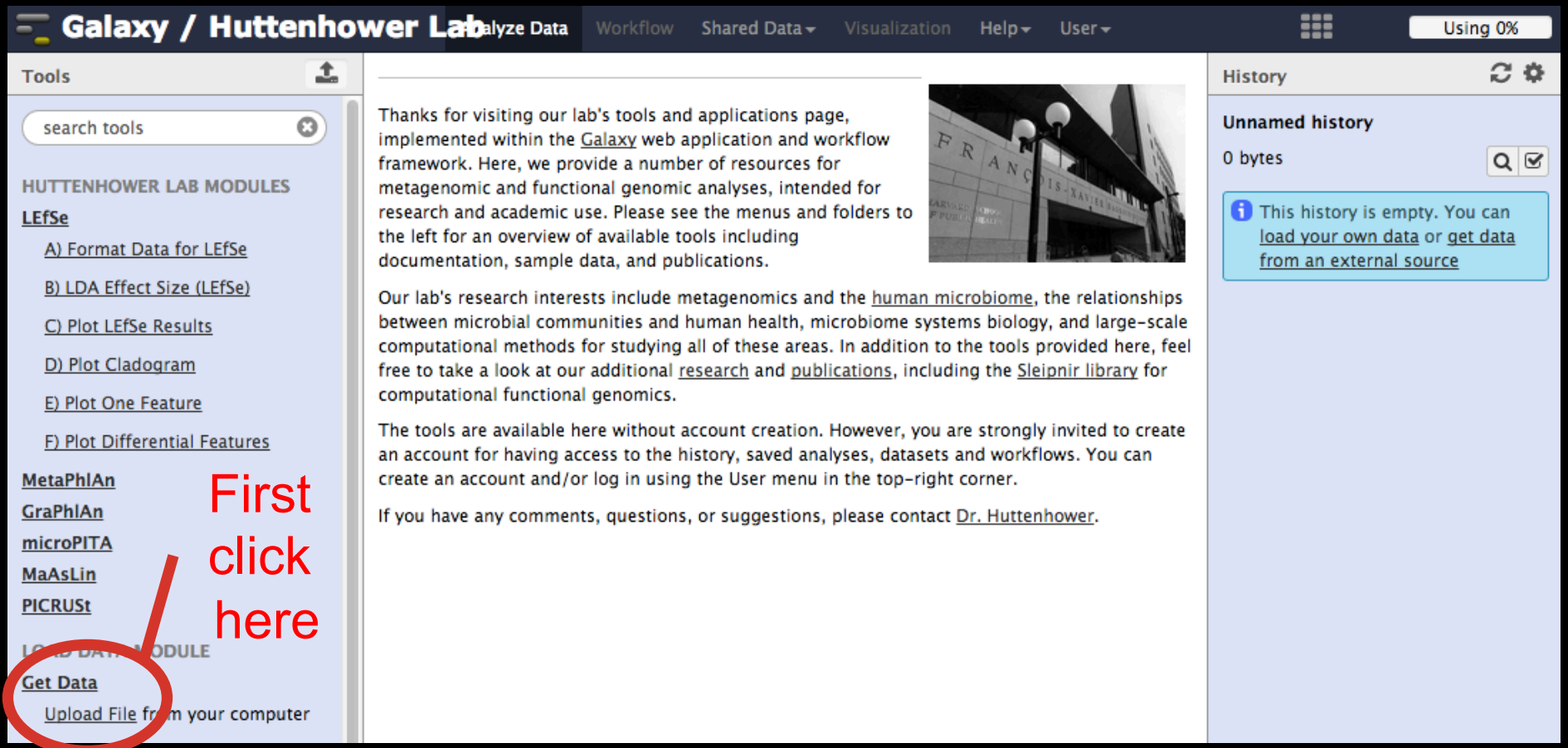


shortbred_ardb_hmp_t2d.tsv

	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	SRS011061	SRS011134	SRS011239	SRS011271	SRS011302	SRS011405	SRS011452	SRS011529	SRS011586	SRS012273	SRS012902	SRS013158	SRS013215	SRS013476	SRS013521	SRS013687	SRS013800	SRS013951
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 Antibi	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 Re	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

What it means: LEfSe

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

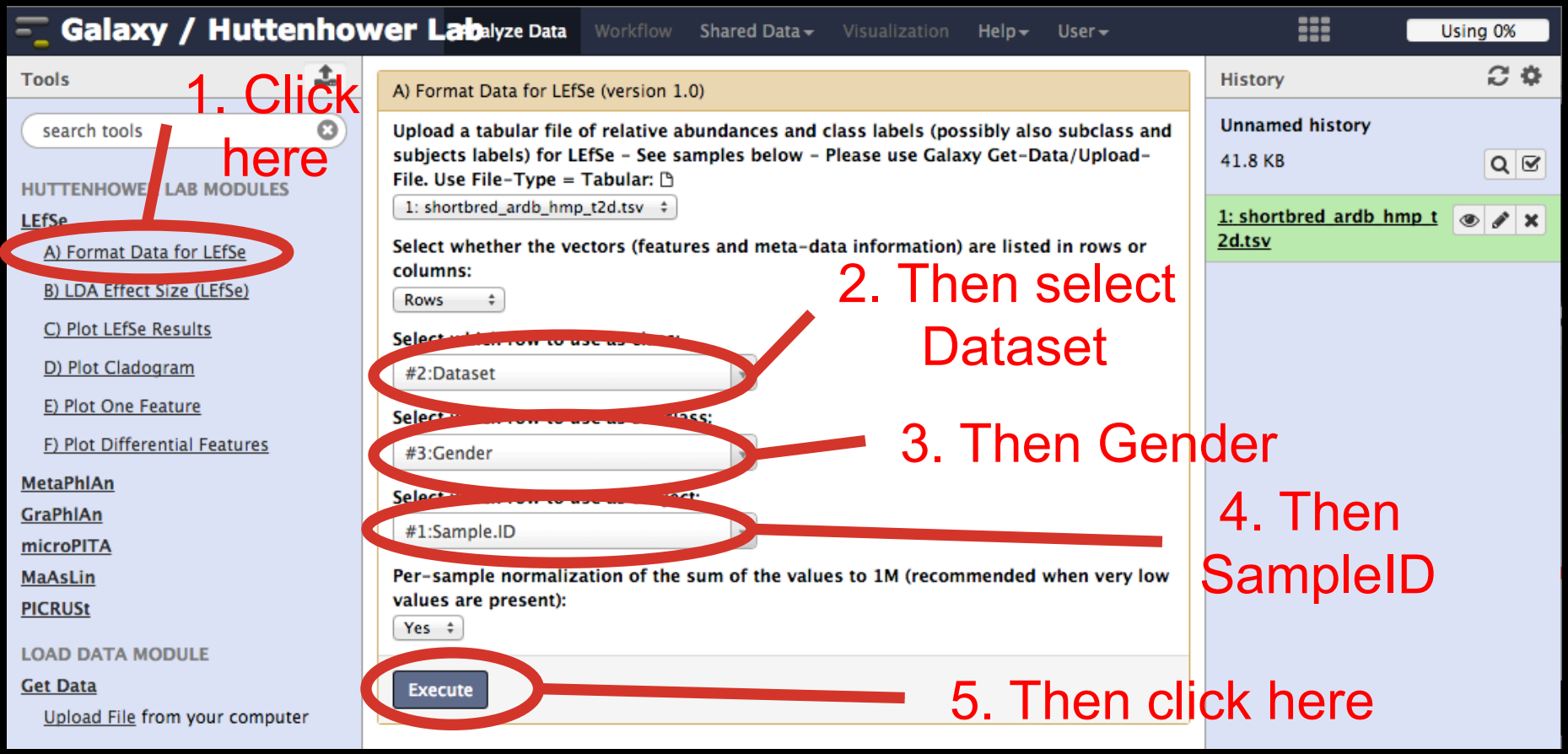


What it means: LEfSe

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

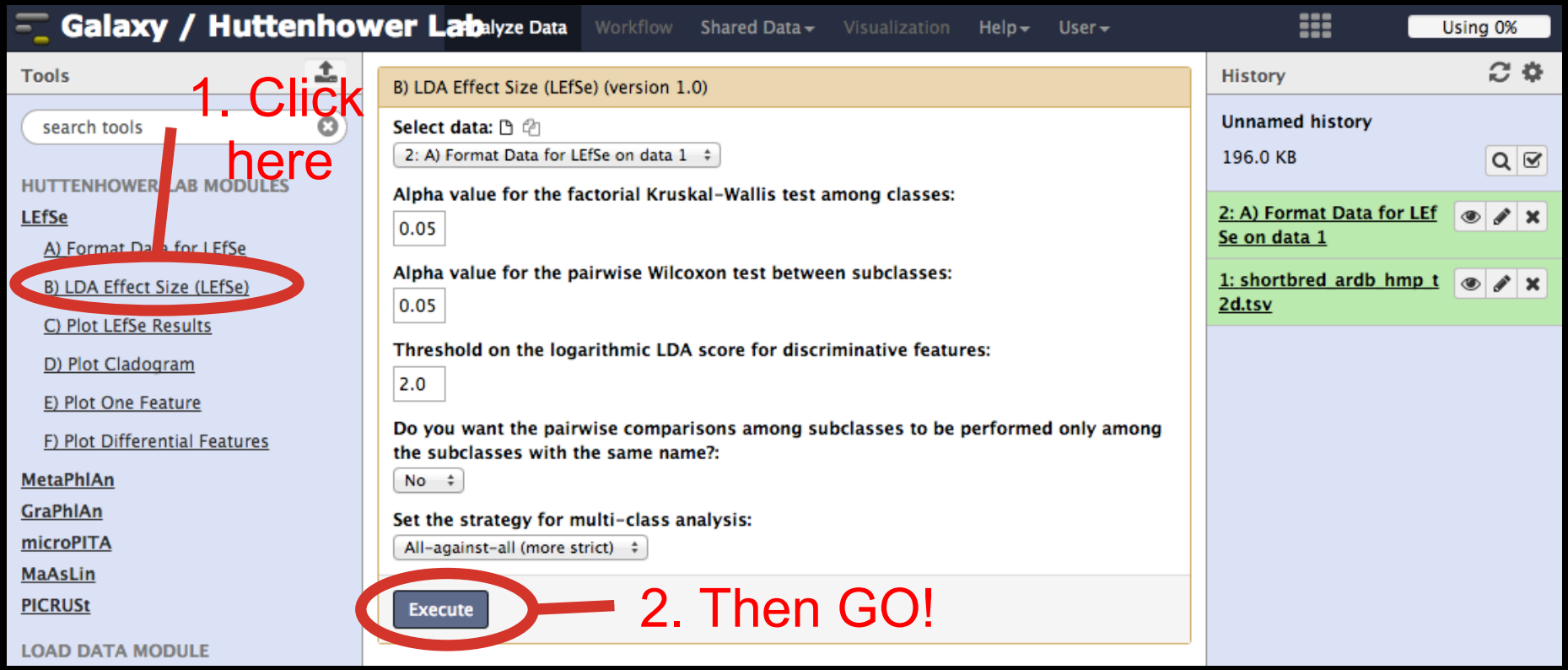
What it means: LEfSe

- Then tell LEfSe about your metadata:



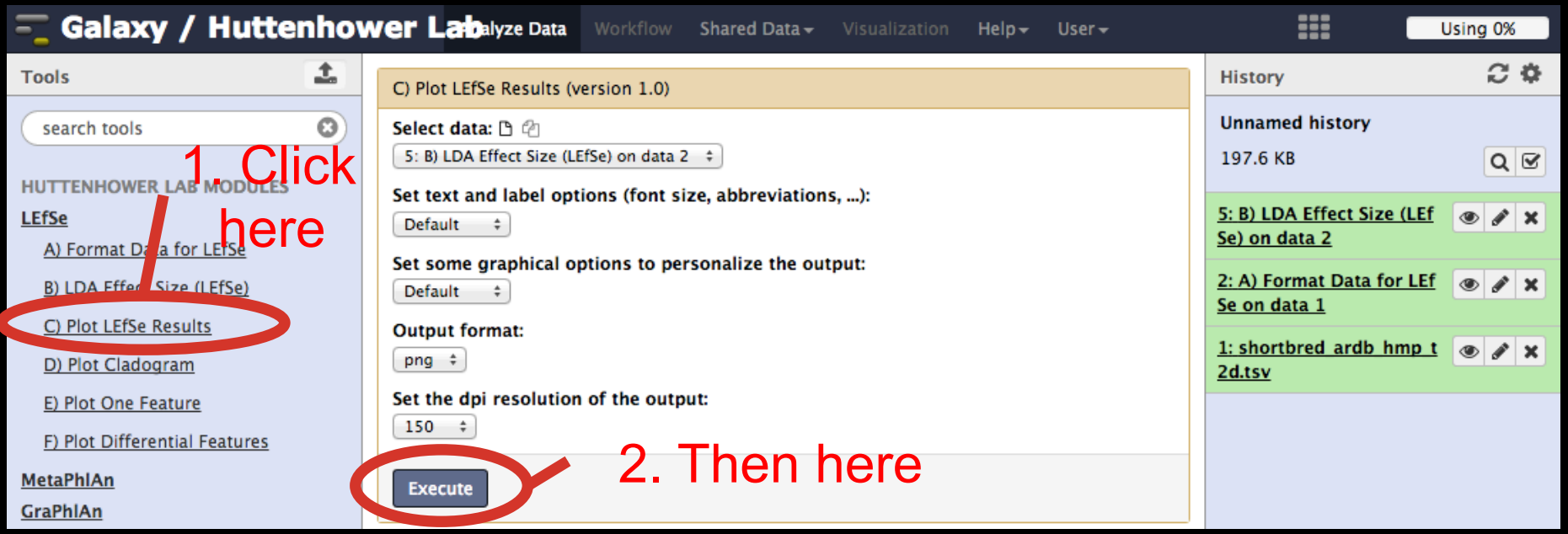
What it means: LEfSe

- Leave all parameters on defaults, and run LEfSe!
  - You can try playing around with these parameters if desired



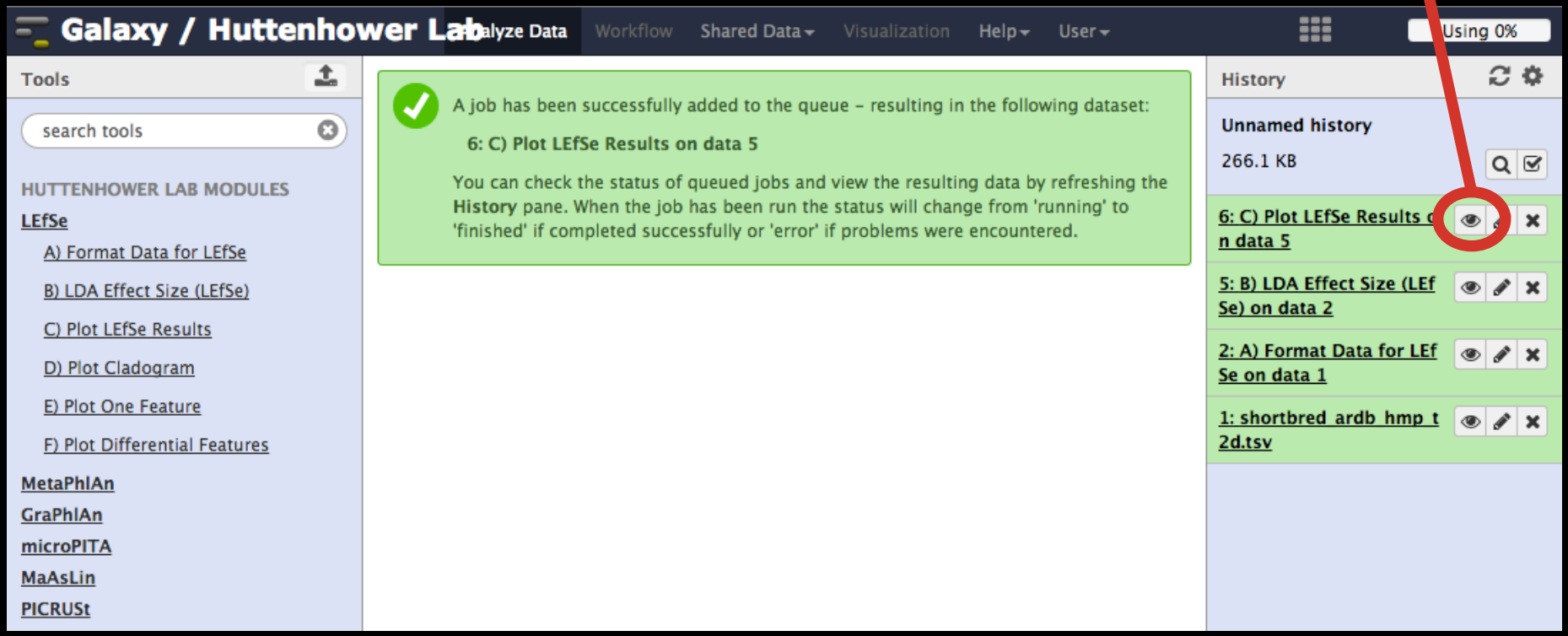
What it means: LEfSe

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

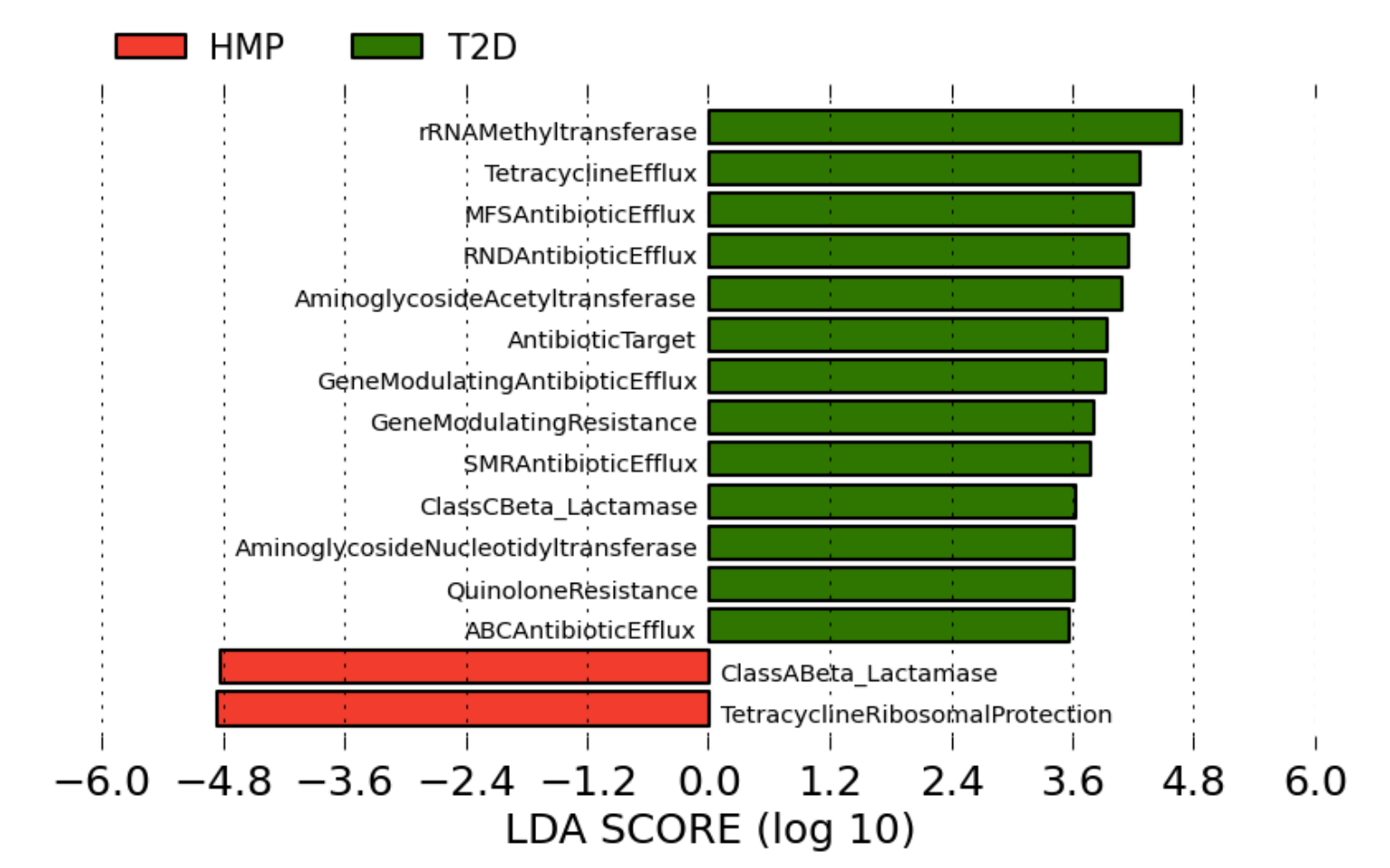


What it means: LEfSe

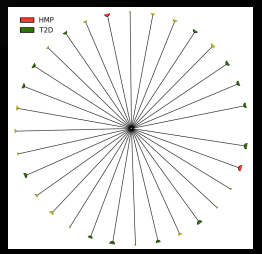
- In Galaxy, view a result by clicking on its “eye”



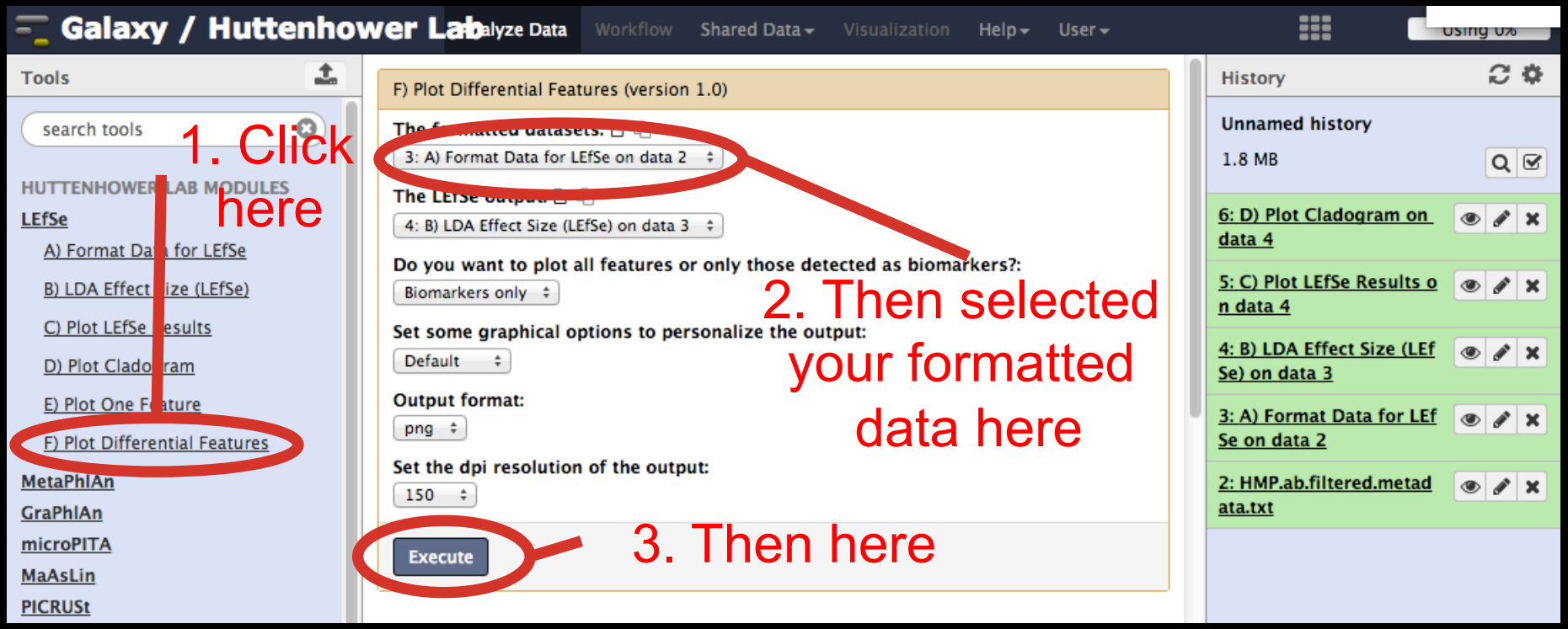
What it means: LEfSe



What it means: LEfSe

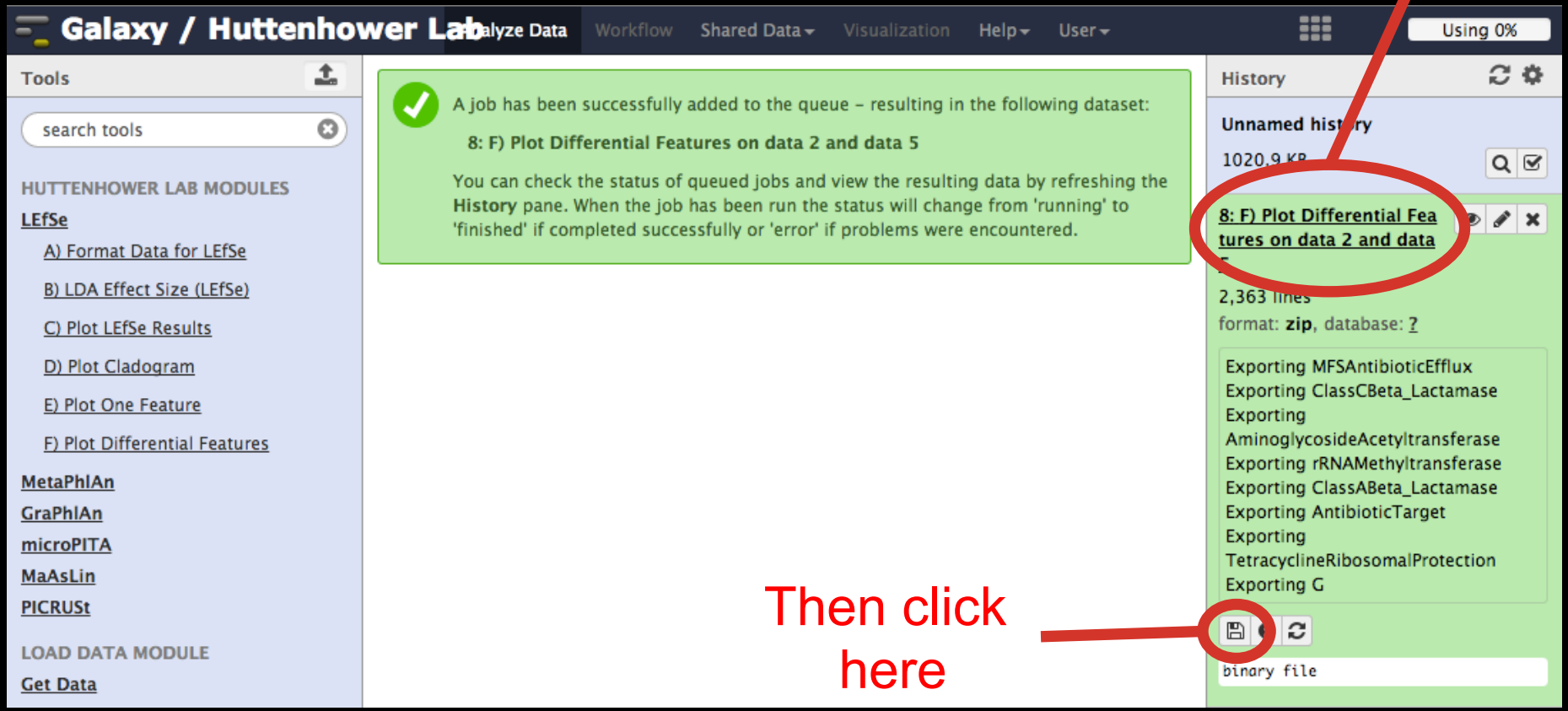


- There's no really any reason to plot a cladogram
  - Although it will work!
- But you can see the raw data for individual biomarkers
  - These are generated as a zip file of individual plots



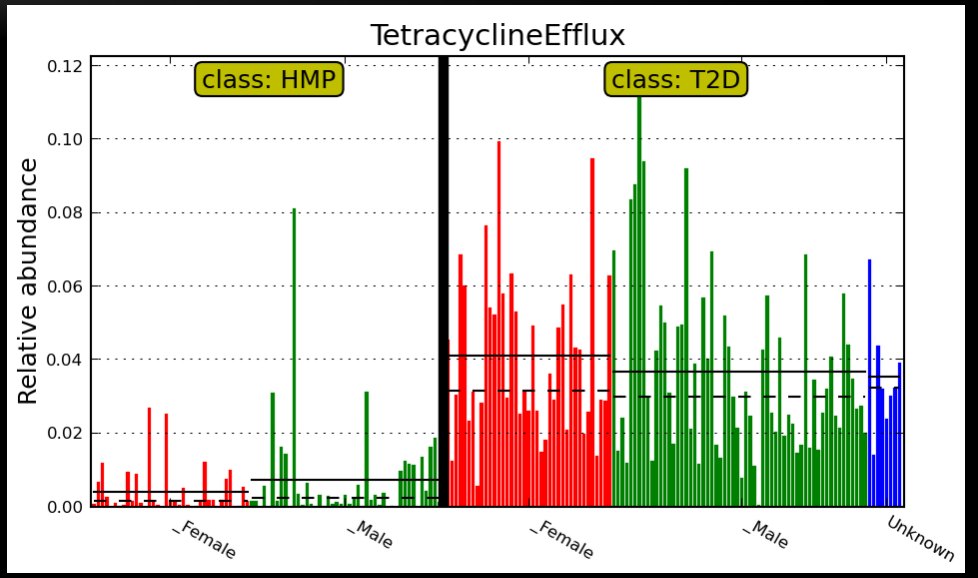
What it means: LEfSe

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

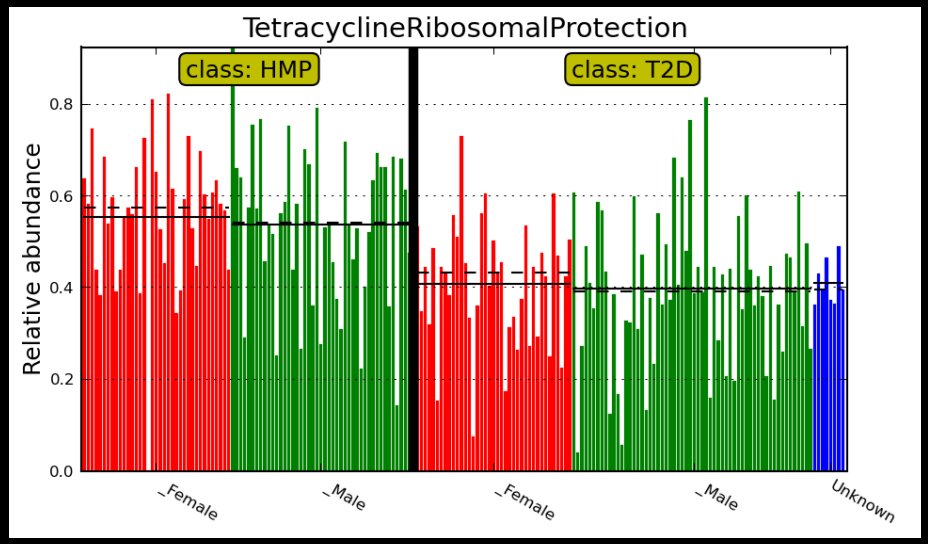


What it means: LEfSe

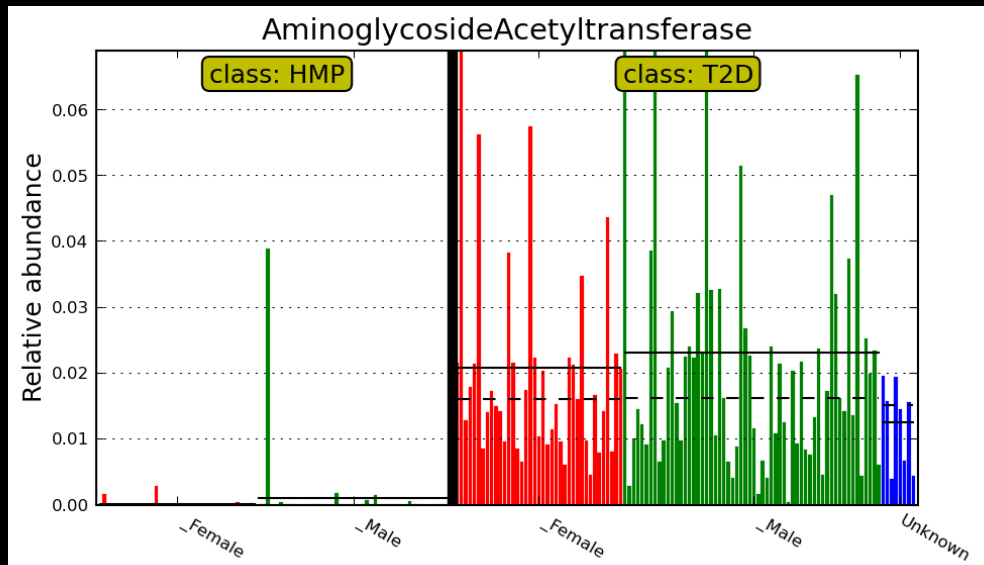
Tetracycline Efflux Pumps



Tet. Ribosomal Blockers



Aminoglycoside Acetyltransferases



Summary

- HUMAnN
  - 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

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:

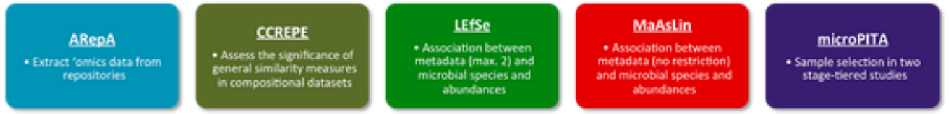
The wiki provides tutorials for Huttenhower tools, illustrating through demos how to use these tools on your datasets. Huttenhower tools can be divided under three main categories as shown below. Click on the tool for the corresponding tutorial.

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

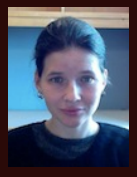
These tools can help visualize taxonomical and phylogenetic information for (i) microbial composition/taxonomy data, (ii) outputs from MetaPhlAn, LEfSe, HUMAnN, MaAsLin. Please click on the link below for detailed tutorial:

Thanks!

http://huttenhower.sph.harvard.edu


Alex Kostic


Levi Waldron


Xochitl Morgan


Tim Tickle


Daniela Boernigen


Lauren McIver

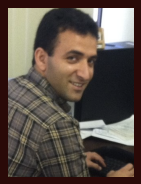

George Weingart


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