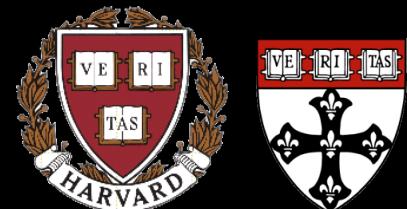




Meta'omic functional profiling with ShortBRED

Curtis Huttenhower

09-19-15



Harvard School of Public Health
Department of Biostatistics



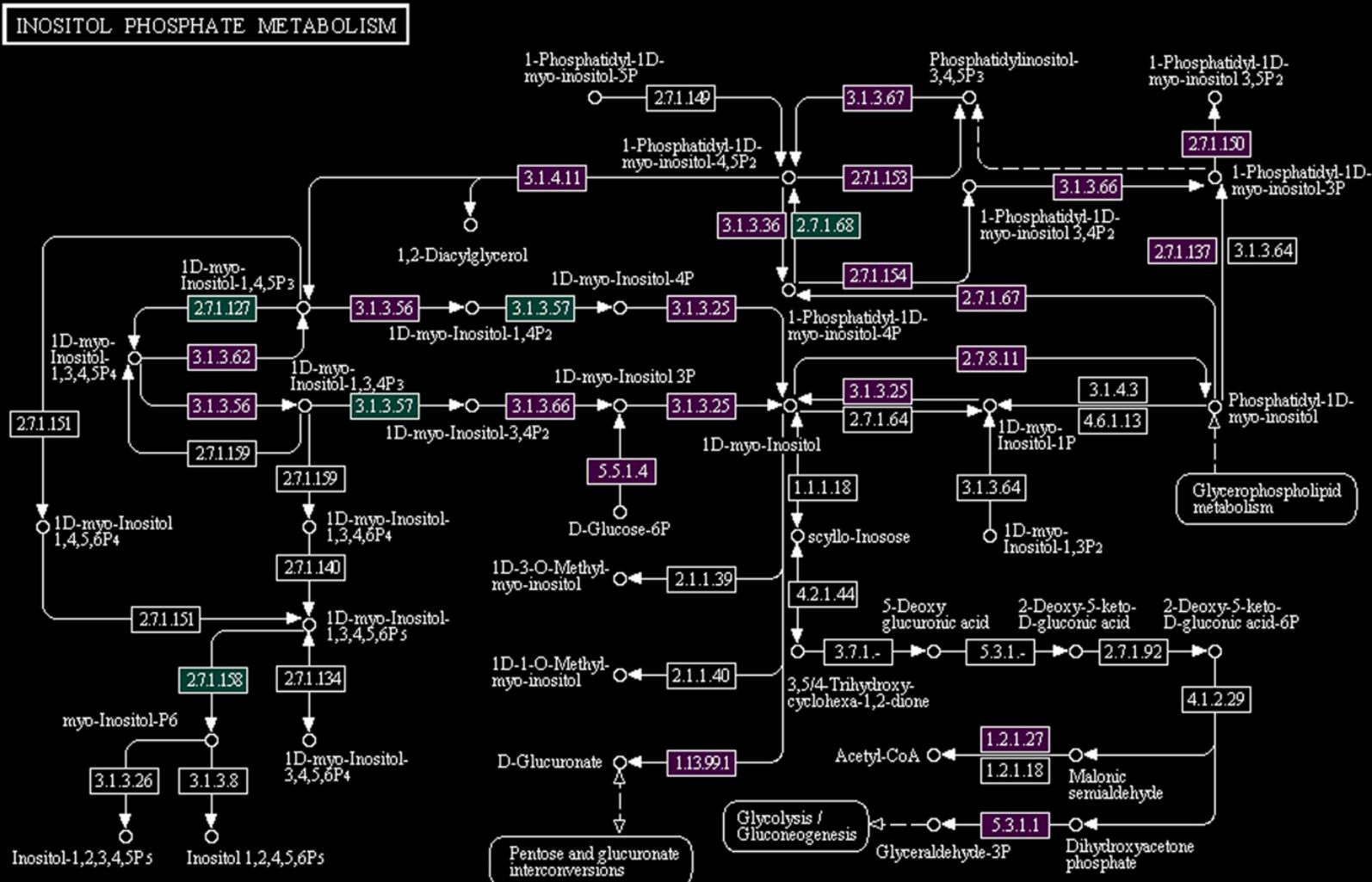


The two big questions...

Who is there?
(taxonomic profiling)

What are they doing?
(functional profiling)

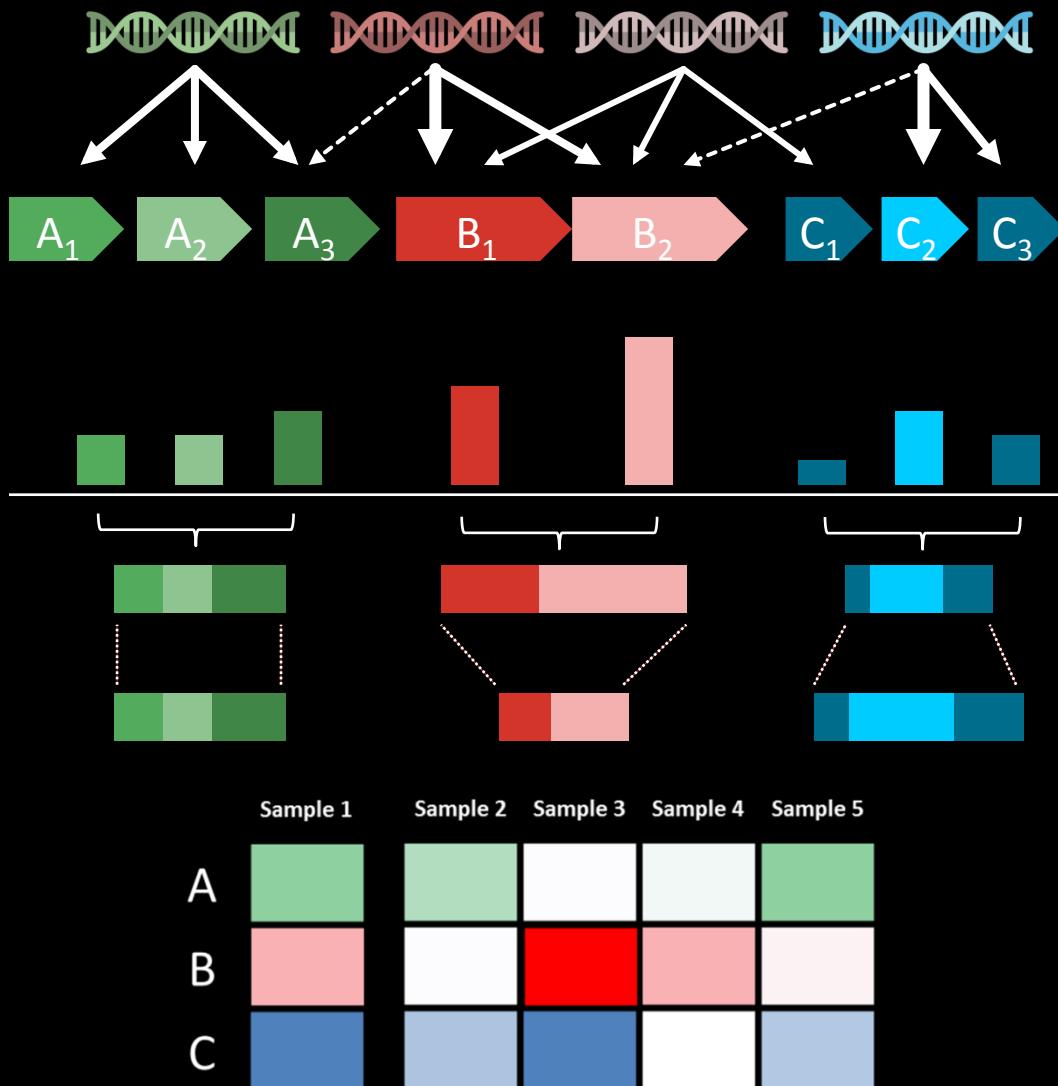
(What we mean by “function”)





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)



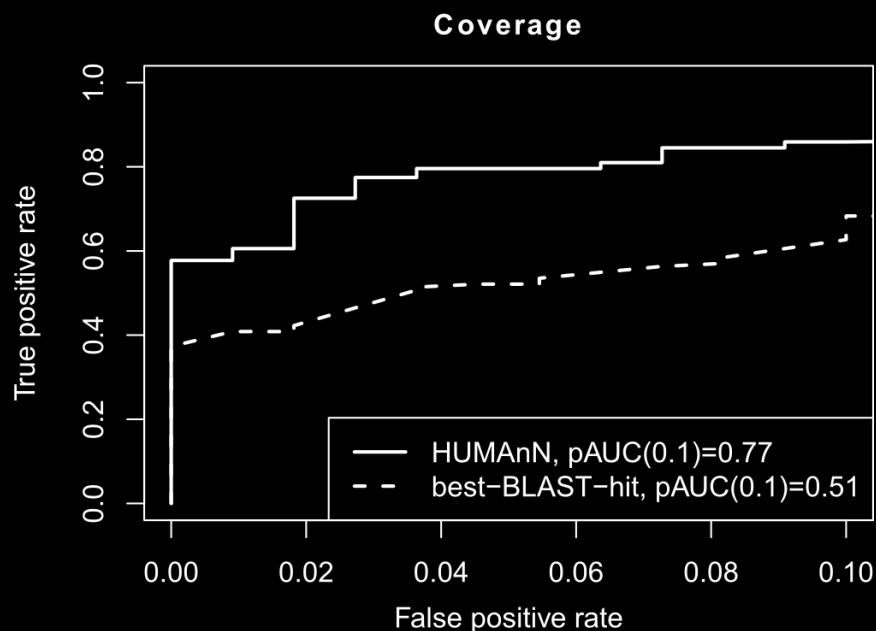
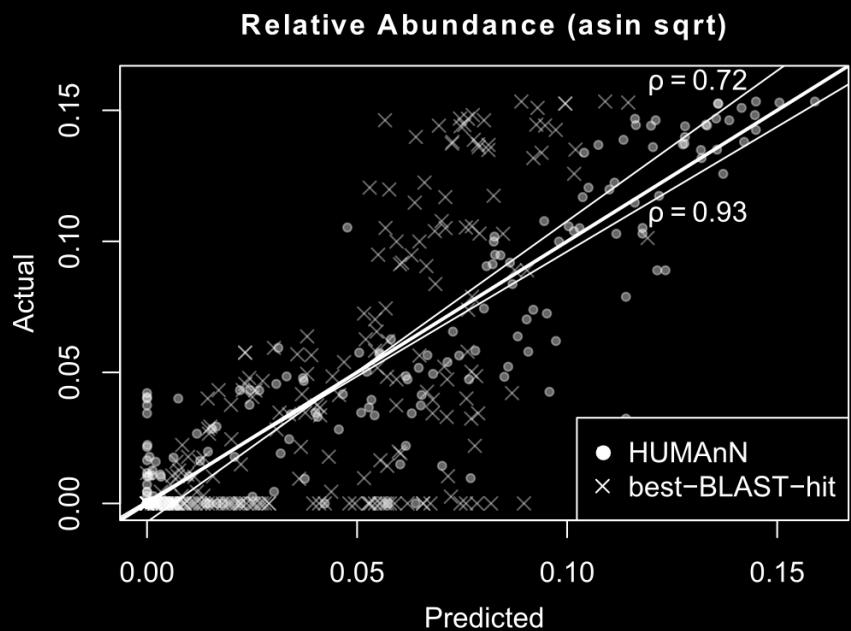
- Map genes to KEGG pathways modules
- Use MinPath (Ye 2009) to find simplest pathway explanation for observed genes
- Remove pathways unlikely to be present due to low organismal abundance
- Smooth/fill gaps



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)

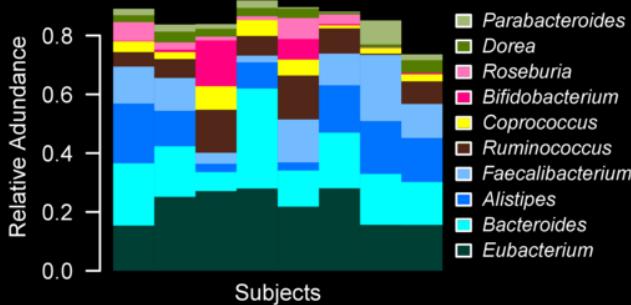
Gene family abundance and pathway presence/absence
calls beat naïve best-BLAST-hit strategy



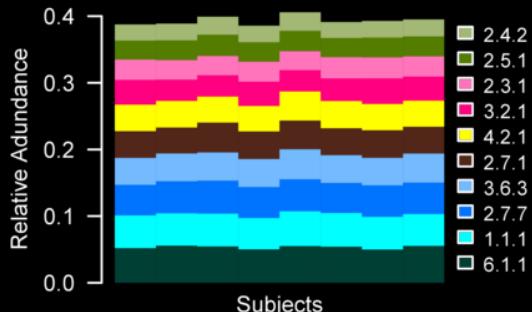
HUMAnN in action



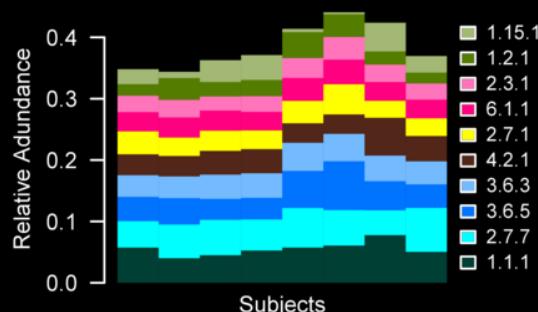
A Top 10 Genera



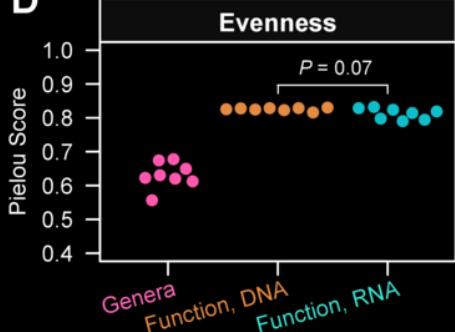
B Top 10 Gene Families, DNA
(EC level-3 categories)



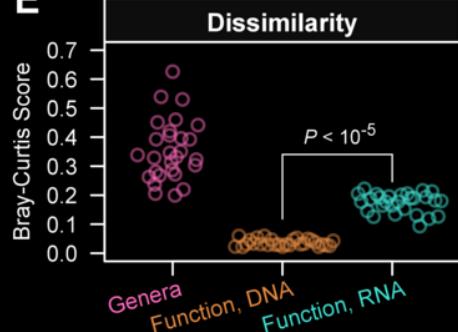
C Top 10 Gene Families, RNA
(EC level-3 categories)



D Evenness



E Dissimilarity



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
- 1.1.5 = 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



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.

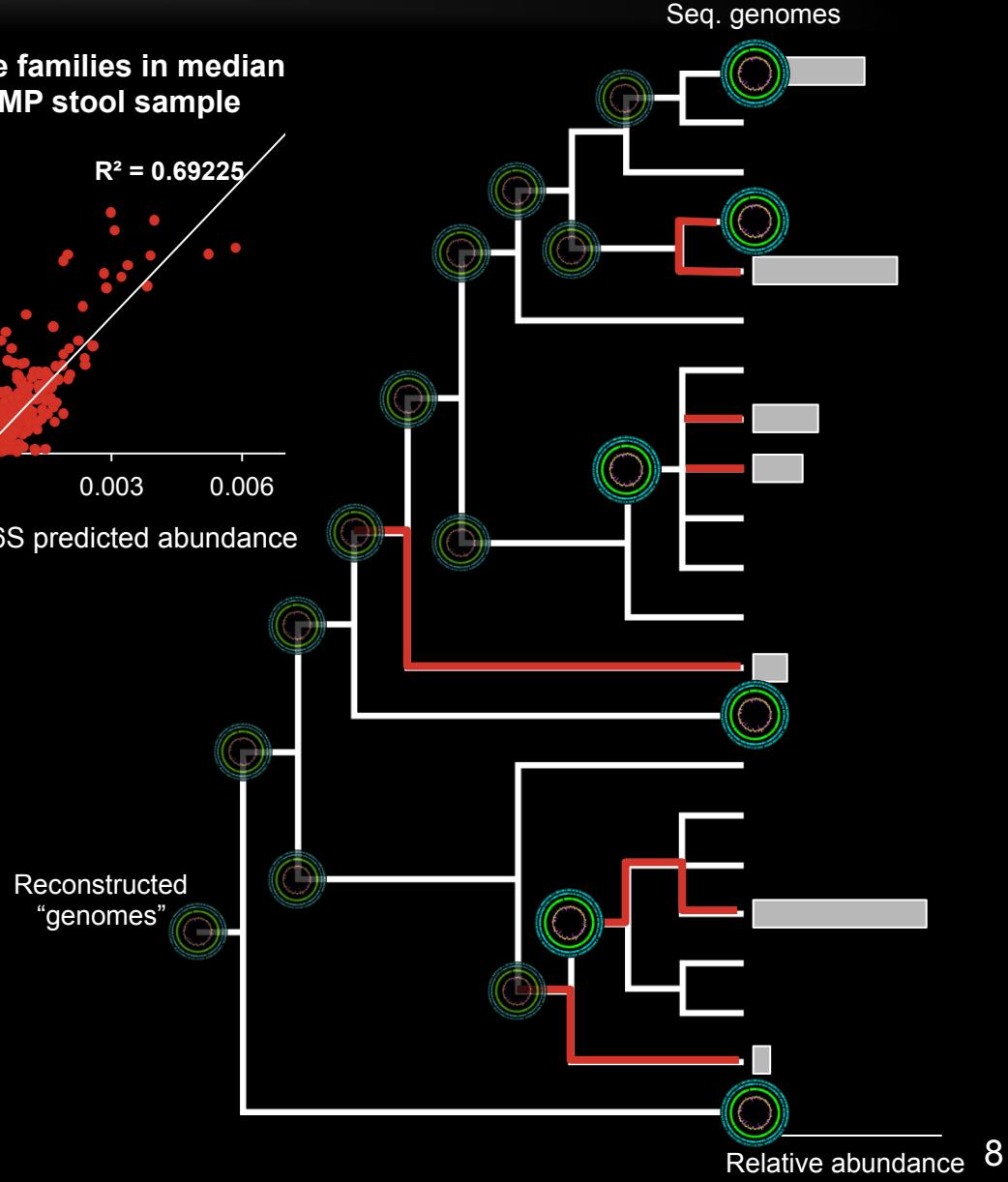
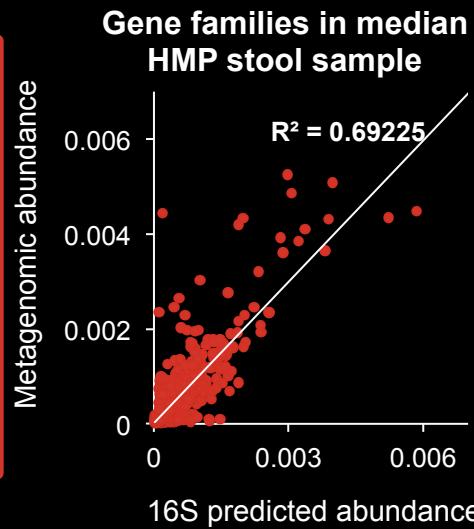
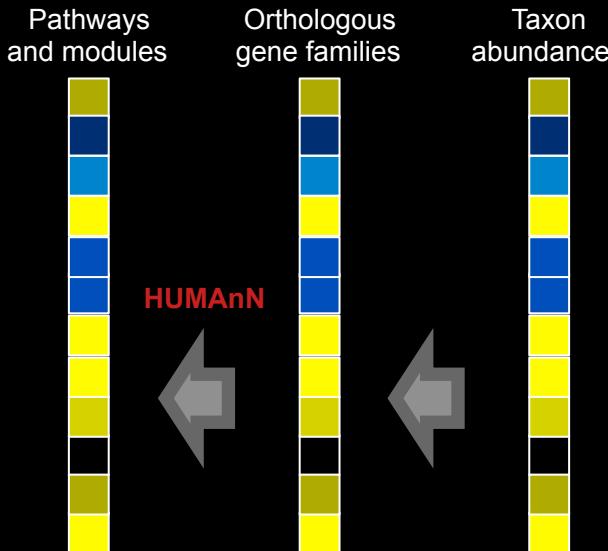
<http://picrust.github.com>



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

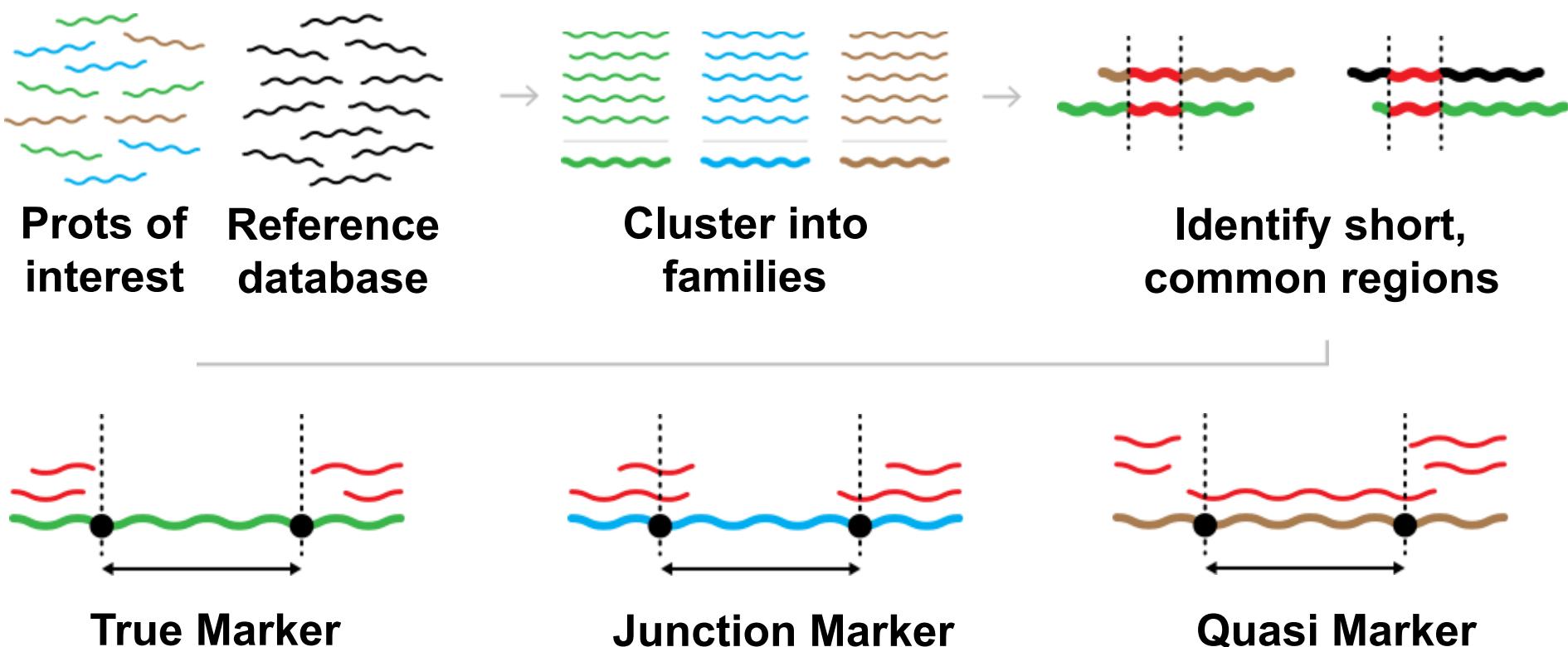


What's there: ShortBRED algorithm

- 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

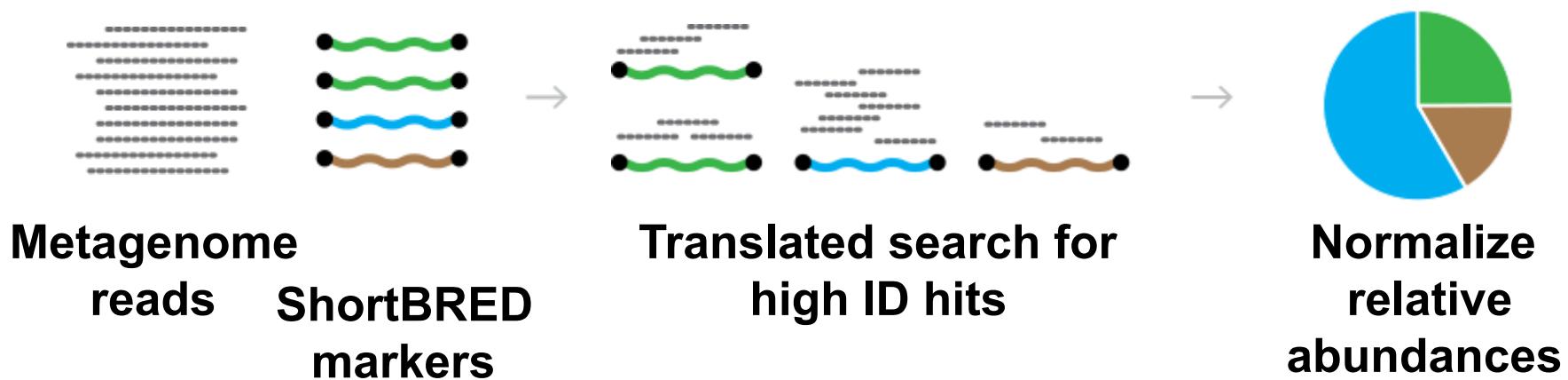


What's there: ShortBRED marker identification



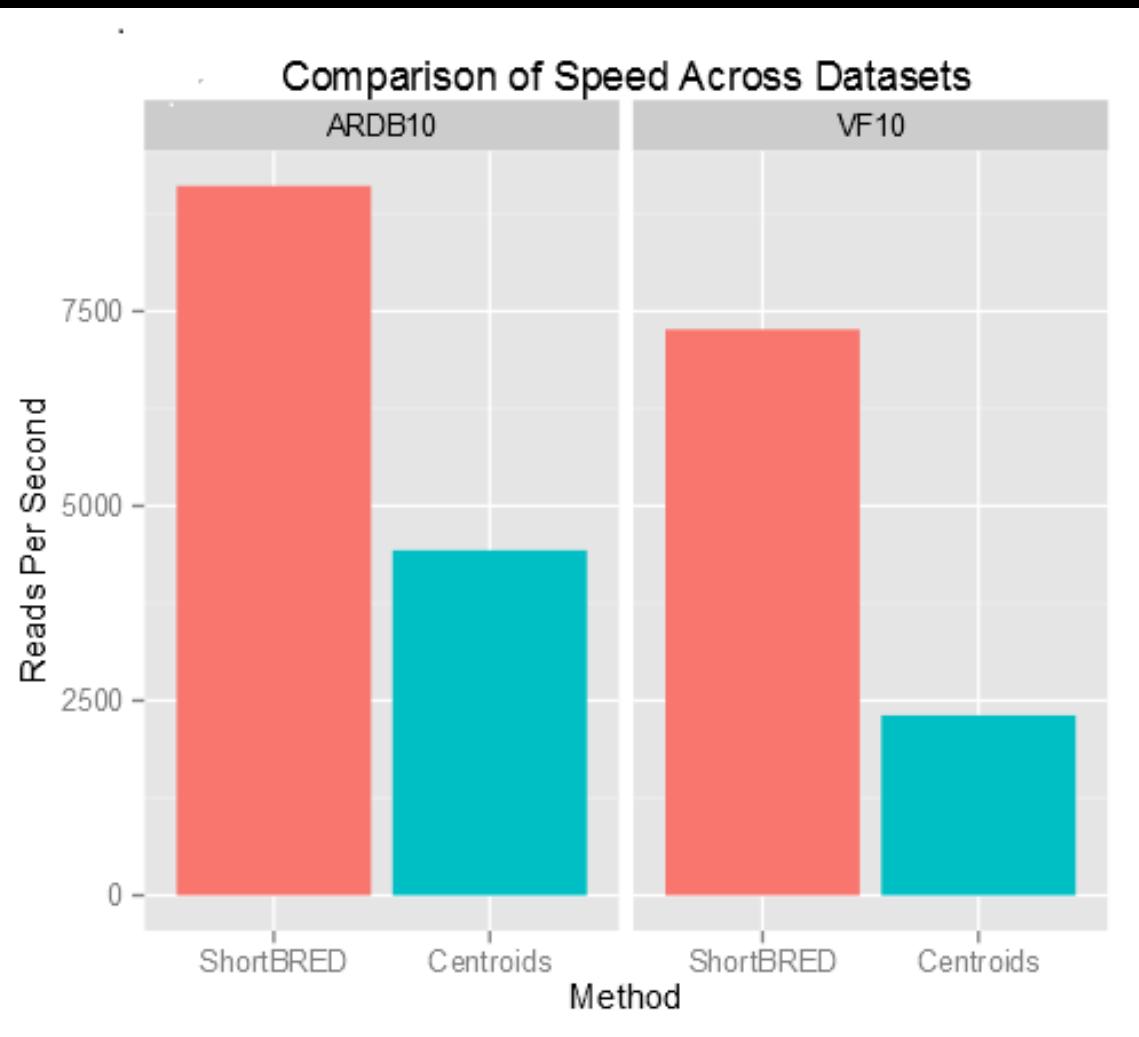


What's there: ShortBRED family quantification





What's there: ShortBRED's fast

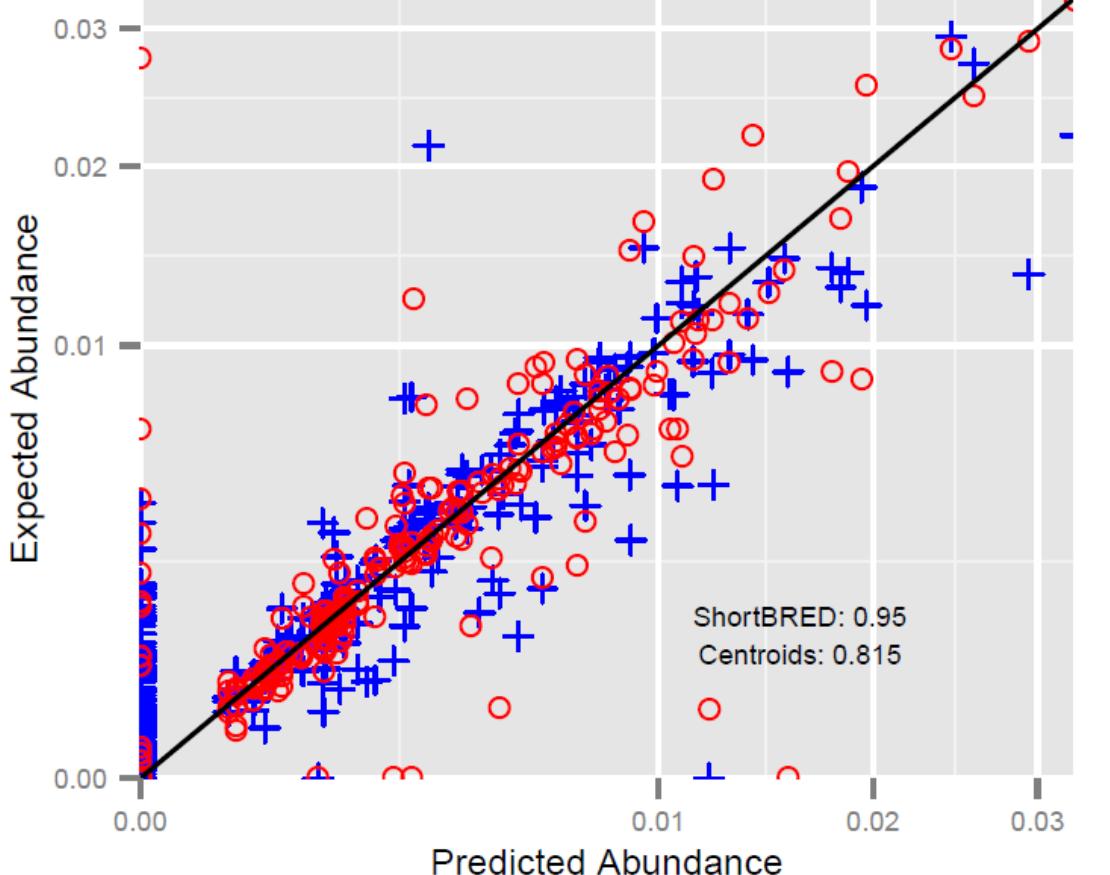


Six synthetic metagenomes from GemSim, spiked with known proteins of interest:
ARDB = Antibiotic Resistance
VFDB = Virulence Factors



What's there: ShortBRED's accurate

B. Antibiotic Resistance Genes Database
Correlation – 10% of Metagenome, 500 genes



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>

The screenshot shows the homepage of The Huttenhower Lab. The header includes the lab's logo, name, and department information. A navigation bar below the header contains links for Contact, Documentation, People, Presentations, Publications, Research, and Teaching. The main content area features a heading "You could download ShortBRED by clicking [here](#)". Below this, a section titled "ShortBRED" provides a detailed description of the tool. A large red diagonal line starts from the top right and points towards the "Download ShortBRED here" link. Another red circle highlights this same link.

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 uniquely 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 and to cite ShortBRED, please reference the following manuscript:

Kaminski J, Gibson M, Franzosa E, Segata N, Dantas G, 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
- To see what you can do, run:

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

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

```
1. ssh
usage: shortbred_identify.py [-h] [--goi SGOIPROTS] [--ref SREFPROTS]
                               [--refdb DIRREFDB] [--goiblast SGOIBLAST]
                               [--refblast SREFBLAST] [--goiclust SCLUST]
                               [--map_in SMAPIN] [--markers SMARKERS]
                               [--map_out SMAP] [--clustid DCLUSTID]
                               [--qclustid DQCLUSTID] [--constthresh DCONSTTHRESH]
                               [--threads ITHREADS] [--id DID] [--len DL]
                               [--minAln ILENMIN] [--markerlength IMLENGTH]
                               [--totlength ITOTLENGTH] [--qthresh ITHRESH]
                               [--qmlength IQMLENGTH] [--xlimit IXLIMIT]
                               [--tmpdir STMP] [--usearch STRUSEARCH]
                               [--muscle STRMUSCLE] [--cdhit STRCDHIT]
                               [--blastp STRBLASTP]

ShortBRED Identify:
This program produces a set of markers for your proteins of interest.
The minimum input files required to run the program are:
    [--goi] 1) A fasta file of proteins, for which you want to build marker
s.
    [--ref] 2) A fasta file of reference proteins
The program will output a file fasta file of markers [--markers].
Example:
$ ./ python shortbred_identify.py --goi example/input_prots.faa --ref ex
:|
```



Getting some annotated protein sequences

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

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

ARDB - Antibiotic Resistance Genes Database

HOME DOCUMENTATION BLAST ADVANCED SEARCH BROWSE

[Database](#) [All Databases](#) [Search](#) [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.cbcu.umd.edu/pub/data/ARDB/ARDBflatFiles.tar.gz>. Documentation about the provided data is available at <ftp://ftp.cbcu.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 ~/workshop_data/metagenomics/biobakery/data/resisGenes.pfasta
```

```
>ZP_02959935 hypothetical protein PROSTU_01837 [Providencia stuartii ATCC 25827]
MGIEYRSLSQLTLSEKEALYDLLIEGFEGDFSHDDFAHTLGGMHVMAFDQQKLVGHVA
IIQRHMALDNTPISVGVEAMVVEQSYRRQGIGRQLMLQTNKIIASCYQLGLLSASDDGQ
KLYHSGWQIWKGKLFEKLQGSYIRSIEEEGGVMGKADGEVDTASLYCDFRGGDQW
>Q52424 RecName: Full-Aminoglycoside 2'-N-acetyltransferase; AltName: Full-AAC(2
MGIEYRSLSQLTLSEKEALYDLLIEGFEGDFSHDDFAHTLGGMHVMAFDQQKLVGHVA
IIQRHMALDNTPISVGVEAMVVEQSYRRQGIGRQLMLQTNKIIASCYQLGLLSASDDGQ
KLYHSGWQIWKGKLFEKLQGSYIRSIEEEGGVMGKADGEVDTASLYCDFRGGDQW
>AAA03550 aminoglycoside 2'-N-acetyltransferase [Providencia stuartii].
MGIEYRSLSQLTLSEKEALYDLLIEGFEGDFSHDDFAHTLGGMHVMAFDQQKLVGHVA
IIQRHMALDNTPISVGVEAMVVEQSYRRQGIGRQLMLQTNKIIASCYQLGLLSASDDGQ
KLYHSGWQIWKGKLFEKLQGSYIRSIEEEGGVMGKADGEVDTASLYCDFRGGDQW
>Q49157 RecName: Full-Aminoglycoside 2'-N-acetyltransferase; AltName: Full-AAC(2
MPFQDVSAFPVRGGILHTARLVHTSDLQETREGARRMVIAEFGDFSDADWEHALGGMHA
FICHHGALIAAAVVQRLLYRDTALRCGYVEAVAVREDWRGQQLATAVMDAVEQVLRGA
YQLGALSASDTARGMYLSRGWLWPWGPTSVLQPAGVTRTPEDDEGLFVLPVGLPAGMELD
TTAEITCDWRGDGVW
>NP_214776 aminoglycoside 2'-N-acetyltransferase AAC (AAC(2')-IC) [Mycobacterium
MHTQVHTARLVHTADLDSETRQDIRQMTGAFAGDFTEWDWEHTLGGMHALIWHHGAIIA
HAAVIQRRLIYRGNALRCGYVEGVAVRADWRGQRLVSALLDAVEQVMRGAYQLGALSSSA
RARRLYASRGWLWPWGPTSVLAPTGPVRTPDDDGTVFVLPIDISLDTSAELMCDWRAGDV
W
>NP_334681 aminoglycoside 2-N-acetyltransferase [Mycobacterium tuberculosis CDC1
MHTQVHTARLVHTADLDSETRQDIRQMTGAFAGDFTEWDWEHTLGGMHALIWHHGAIIA
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

Metar Ref

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](#)

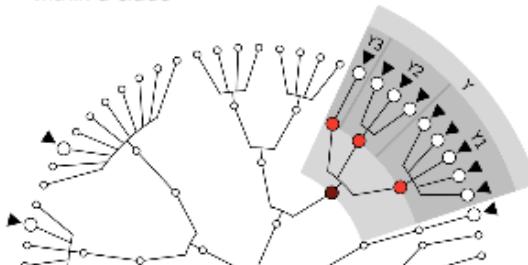
[Streptococcus mitis](#)

MetaRef Database v 1.0

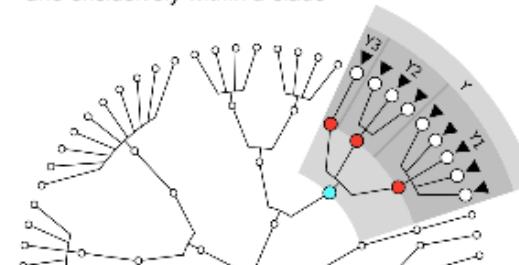
MetaRef is a resource to comprehensively catalog and characterize clade-specific microbial genes. We identify and provide all core genes associated with all microbial species and genera with available reference genomes (final or draft). A subset of these gene families are consistently present in one 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:

```
ln -s /usr/bin/cdhit /home/ubuntu/Programs/cd-hit
shortbred_identify.py
--goi ~/workshop_data/metagenomics/biocbakery/data/resisGenes.pfasta
--ref ~/workshop_data/metagenomics/biocbakery/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/biocbakery/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
IRATYTGASSQTVENAVTQVIEQSQQSLDHLMYMTSTSASDGSAQVNLFAT
ardb_markers.faa
```

True Markers
at the top



ShortBRED markers

```
 3. [screen 3: bash] chuttenhower@class:/class/stamps-software/biobake... ↗
>AAA25688_TM_#02
PTQLNKGLGTRLVRALVELLFSDPTVTKIQTDPPTPNH
>YP_277581_TM_#01
MSMIYITLNIIAYVIDVRSLLIDVRRLVFS
>YP_277581_TM_#02
NILNCDDSVIAFTVIIQLGAILSITKIFWSQLYGMMSMICIKKIFFKQHDDHNHLCIRHI
FLGTFPGIMLGMIFYEKIGLIFELTYIMYGLIIGGIFLLVGELCASKERVSRINNITYL
>YP_277581_TM_#03
FSRAGATIGGGLVVGLDRRISS
>YP_277581_TM_#04
SAVLTLHYRSCIGLMDVLLIAGSATAFFIALFTVRYFLKIVKNVSLIPFAIYRFLLAG
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.
TLTGPFIGGFIKEFQPVAKEKAIPKELFTSVK
(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

Family	Count	Hits	TotalAAlength		
YP_001694417	2380.9523809523807	1	26		
ZP_04679156	0.0 0	235			
ZP_04657259	0.0 0	178			
ZP_04635798	0.0 0	144			
ZP_04635523	0.0 0	182			
ZP_04633951	0.0 0	70			
ZP_04616832	0.0 0	9			
ZP_04613685	0.0 0	95			
ZP_04606269	0.0 0	193			
ZP_04577926	0.0 0	185			
ZP_04543635	0.0 0	173			
ZP_04543532	0.0 0	194			
ZP_04433866	0.0 0	205			
ZP_04431003	0.0 0	105			
ZP_04405580	0.0 0	169			
ZP_04405450	0.0 0	300			
ZP_04309403	0.0 0	167			
ZP_04284182	0.0 0	212			
ZP_04244950	0.0 0	51			
ZP_04210257	0.0 0	169			
ZP_04197552	0.0 0	154			
ZP_04175489	0.0 0	70			
ZP_04174269	0.0 0	21			

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_arb_hmp_t2d.tsv

Search in Sheet

	A1	Sample.ID	HMP1	HMP2	HMP3	HMP4	HMP5	HMP6	HMP7	HMP8	HMP9	HMP10	HMP11	HMP12	HMP13	HMP14	HMP15	HMP16	HMP17	HMP18
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	Male	Male	Male	Female	Male	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	0.5570841	0	0	0	0	0	0.4844142	0	0	0	7.15621993	0	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	0.109356043	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	0.1242086	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	0.3360012	
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	0.26860575	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 Antib	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	6.0081483	4.73637	0.16432993	0	9.88061341	0.2382082	43.436675	1.4549685	0	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/>

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

LOAD DATA MODULE

Get Data

Upload File from your computer

First click here

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.

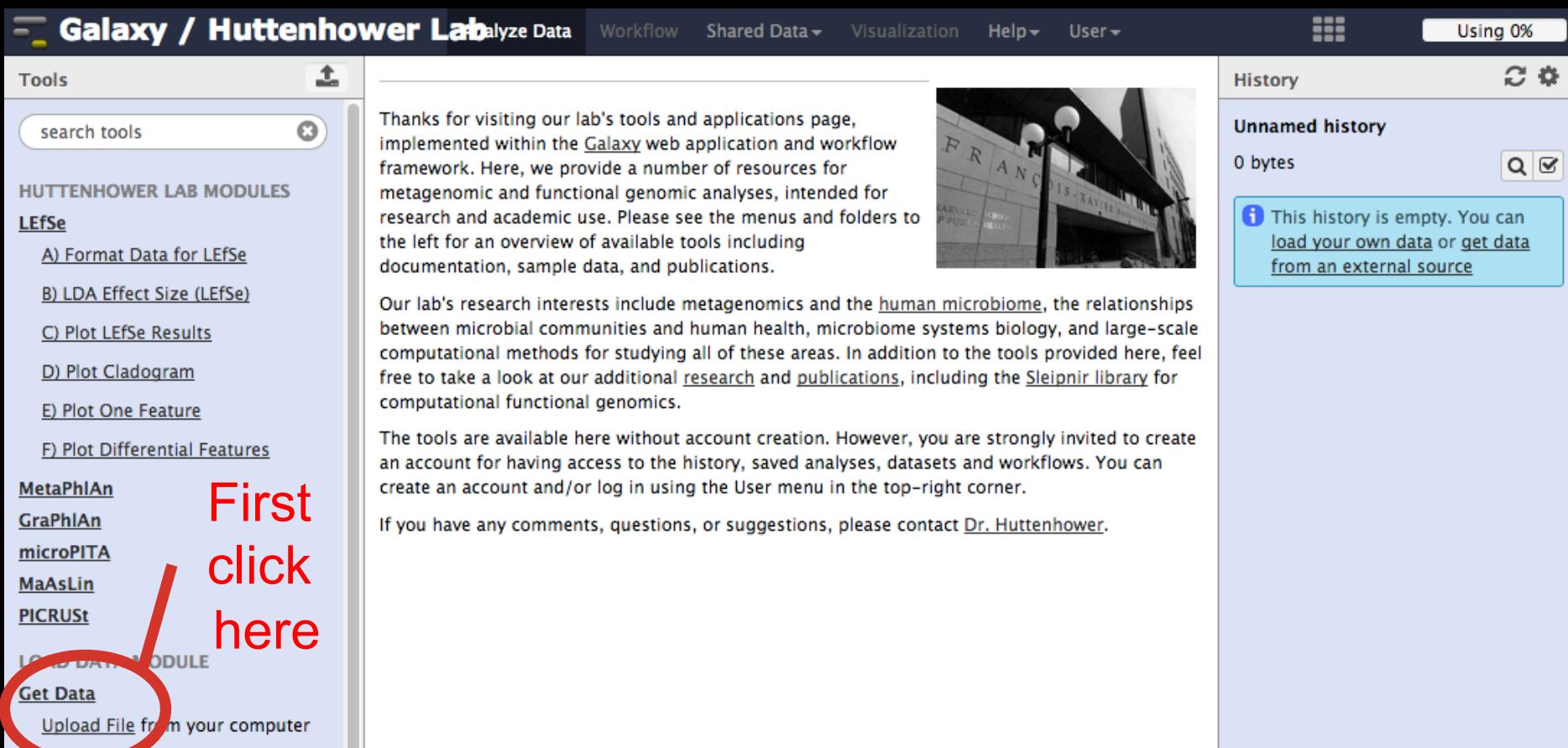
If you have any comments, questions, or suggestions, please contact [Dr. Huttenhower](#).

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!

The screenshot shows the Galaxy web interface with the following elements:

- Header:** Galaxy / Huttenhower Lab, Analyze Data, Workflow, Shared Data, Visualization, Help, User.
- Left Sidebar (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
 - LOAD DATA MODULE:** Get Data, Upload File from your computer
 - DEFAULT GALAXY MODULES:**
- Middle Panel (Upload File):**
 - File Format:** Auto-detect (dropdown).
 - File:** Choose File (button), currently set to Hmp.ab.filtered.metadata.txt.
 - URL/Text:** A large text input field for URLs or file contents.
 - Convert spaces to tabs:** Yes (checkbox), Use this option if you are entering intervals by hand.
 - Genome:** Unspecified (?) dropdown.
 - Execute:** A large blue button at the bottom.
- Right Panel (History):**
 - History: Unnamed history, 269.2 KB.
 - Info message: This history is empty. You can load your own data or get data from an external source.

Annotations:

1. Click here, browse to **shortbred_ardb_hmp_t2d.tsv**
2. Then here
3. Then watch here



What it means: LEfSe

- Then tell LEfSe about your metadata:

The screenshot shows the Galaxy software interface for the Huttenhower Lab. The left sidebar lists various tools and modules, with the 'LEfSe' module highlighted. The main workspace displays the 'A) Format Data for LEfSe (version 1.0)' tool. A red path is overlaid on the interface, starting with a callout '1. Click here' pointing to the 'A) Format Data for LEfSe' link in the sidebar. This leads to a series of steps: '2. Then select Dataset' (circled around the 'Dataset' dropdown), '3. Then Gender' (circled around the 'Gender' dropdown), '4. Then SampleID' (circled around the 'Sample.ID' dropdown), and finally '5. Then click here' (circled around the 'Execute' button at the bottom). The right panel shows a history list with a green item: '1: shortbred ardb hmp t2d.tsv 2d.tsv'. The top right corner indicates 'Using 0%'.

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 LOAD DATA MODULE Get Data Upload File from your computer

A) Format Data for LEfSe (version 1.0)

Upload a tabular file of relative abundances and class labels (possibly also subclass and subjects labels) for LEfSe - See samples below - Please use Galaxy Get-Data/Upload-File. Use File-Type = Tabular: 1: shortbred_arb_hmp_t2d.tsv

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

Select which row to use as class: #2:Dataset

Select which row to use as gender: #3:Gender

Select which row to use as sample ID: #1:Sample.ID

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

Execute

History Unnamed history 41.8 KB 1: shortbred ardb hmp t2d.tsv



What it means: LEfSe

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

The screenshot shows the Galaxy platform interface for the Huttenhower Lab. The left sidebar lists various modules, with 'B) LDA Effect Size (LEfSe)' circled in red and a red arrow pointing to it from the text '1. Click here'. The main panel displays the 'B) LDA Effect Size (LEfSe) (version 1.0)' tool configuration. It includes fields for 'Select data' (set to '2: A) Format Data for LEfSe 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' (2.0), and a dropdown for 'Set the strategy for multi-class analysis' (set to 'All-against-all (more strict)'). At the bottom is a large red circle around the 'Execute' button, with the text '2. Then GO!' next to it. The right sidebar shows a history panel with an unnamed history containing a single item: '2: A) Format Data for LEfSe on data 1'.

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

LOAD DATA MODULE

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

Select data: 2: A) Format Data for LEfSe 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:
2.0

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)

History

Unnamed history

196.0 KB

2: A) Format Data for LEfSe on data 1

1: shortbred ardb hmp_t 2d.tsv

Using 0%

1. Click here

2. Then GO!



What it means: LEfSe

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

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

1. Click here

C) Plot LEfSe Results (version 1.0)

Select data: S: B) LDA Effect Size (LEfSe) on data 2

Set text and label options (font size, abbreviations, ...): Default

Set some graphical options to personalize the output: Default

Output format: png

Set the dpi resolution of the output: 150

2. Then here

History

Unnamed history 197.6 KB

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

2: A) Format Data for LEfSe on data 1

1: shortbred ardb hmp t 2d.tsv



What it means: LEfSe

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

Click here

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

A job has been successfully added to the queue – resulting in the following dataset:
6: C) Plot LEfSe Results on data 5
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.

History

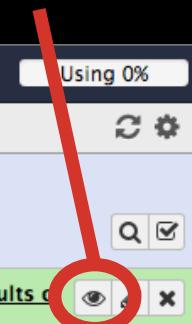
Unnamed history 266.1 KB

6: C) Plot LEfSe Results on data 5

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

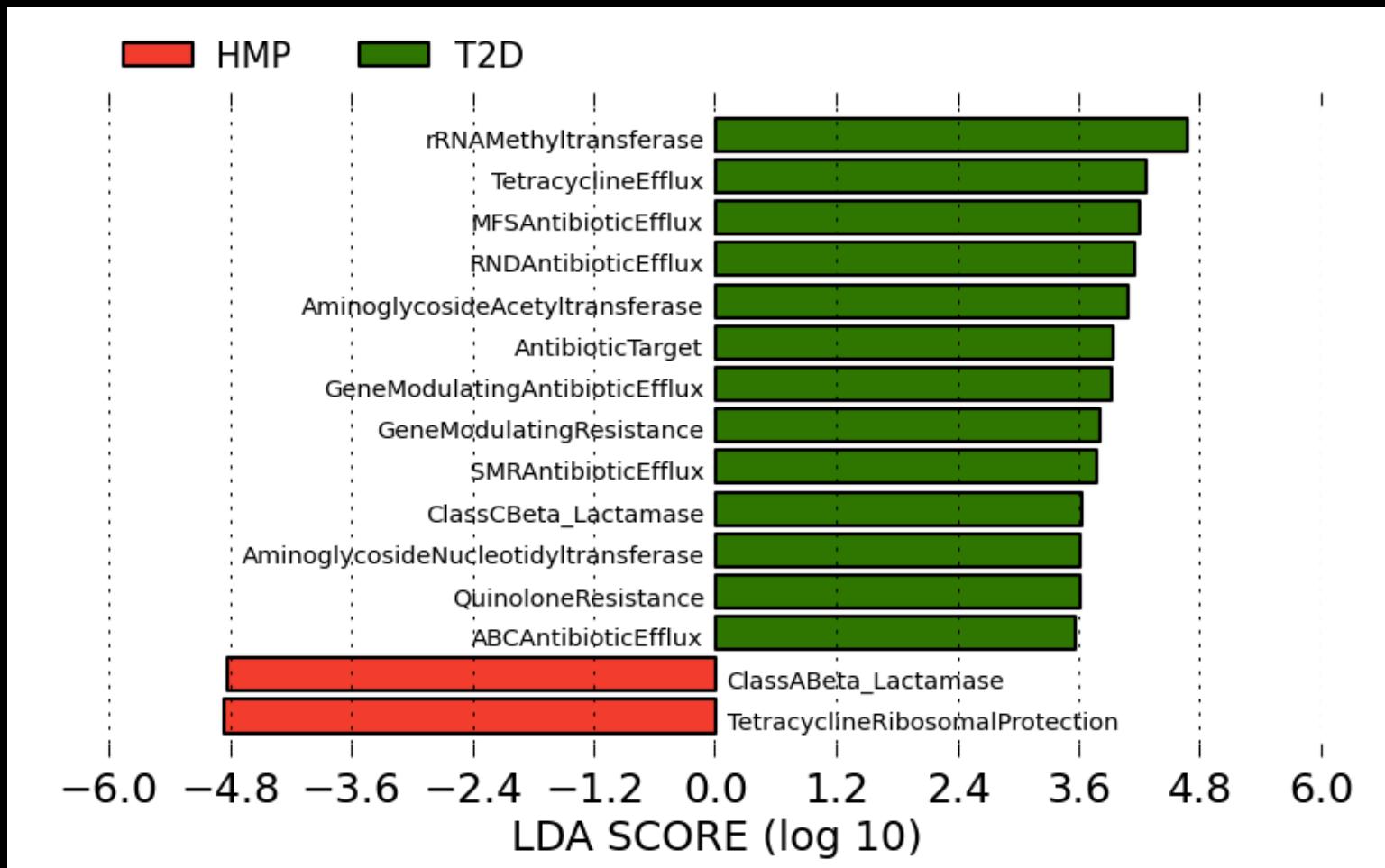
2: A) Format Data for LEfSe on data 1

1: shortbred ardb hmp_t 2d.tsv





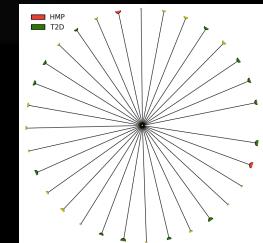
What it means: LEfSe





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



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**

1. Click here

The **F** formatted datasets:

- 3: A) Format Data for LEfSe on data 2
- 4: B) LDA Effect Size (LEfSe) on data 3

The LEfSe output:

- Do you want to plot all features or only those detected as biomarkers?: Biomarkers only
- Set some graphical options to personalize the output: Default
- Output format: png
- Set the dpi resolution of the output: 150

2. Then selected your formatted data here

3. Then here

History

Unnamed history 1.8 MB

- 6: D) Plot Cladogram on data 4
- 5: C) Plot LEfSe Results on data 4
- 4: B) LDA Effect Size (LEfSe) on data 3
- 3: A) Format Data for LEfSe on data 2
- 2: HMP.ab.filtered.metadata.txt



What it means: LEfSe

Click here

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

The screenshot shows the Galaxy web interface with the following elements:

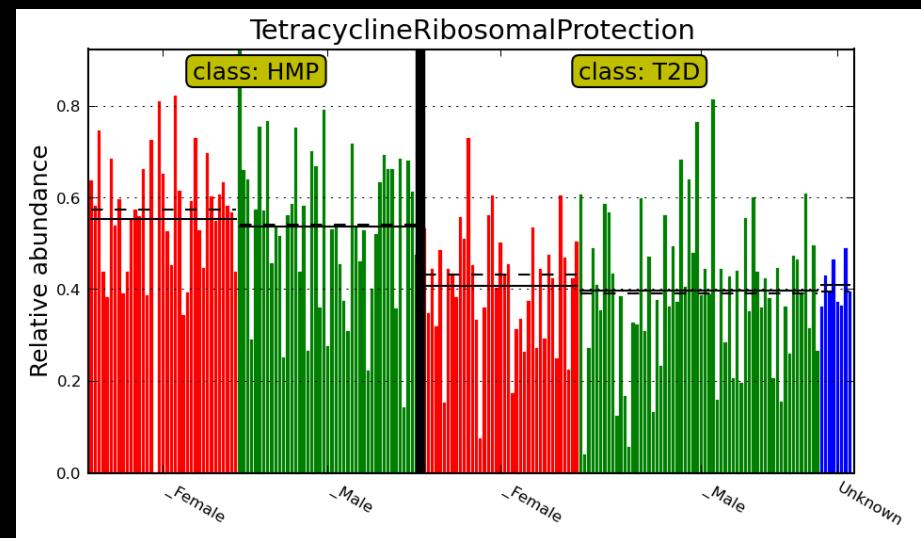
- Header:** Galaxy / Huttenhower Lab, Analyze Data, Workflow, Shared Data, Visualization, Help, User.
- Tools:** search tools, HUTTENHOWER LAB MODULES, LEfSe (selected), MetaPhlAn, GraPhlAn, microPITA, MaAsLin, PICRUSt, LOAD DATA MODULE, Get Data.
- Middle Panel:** A green success message box:
 - A job has been successfully added to the queue – resulting in the following dataset:
 - 8: F) Plot Differential Features on data 2 and data 5**
 - 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.
- History:** Unnamed history, 1020.9 KB, 8: F) Plot Differential Features on data 2 and data 5 (circled in red).
 - 2,363 lines
 - format: zip, database: ?
 - Exporting MFSAntibioticEfflux
 - Exporting ClassCBeta_Lactamase
 - Exporting AminoglycosideAcetyltransferase
 - Exporting rRNAMethyltransferase
 - Exporting ClassABeta_Lactamase
 - Exporting AntibioticTarget
 - Exporting TetracyclineRibosomalProtection
 - Exporting G
- Bottom:** A red arrow points to the "disk" icon next to the history entry, with the text "Then click here".



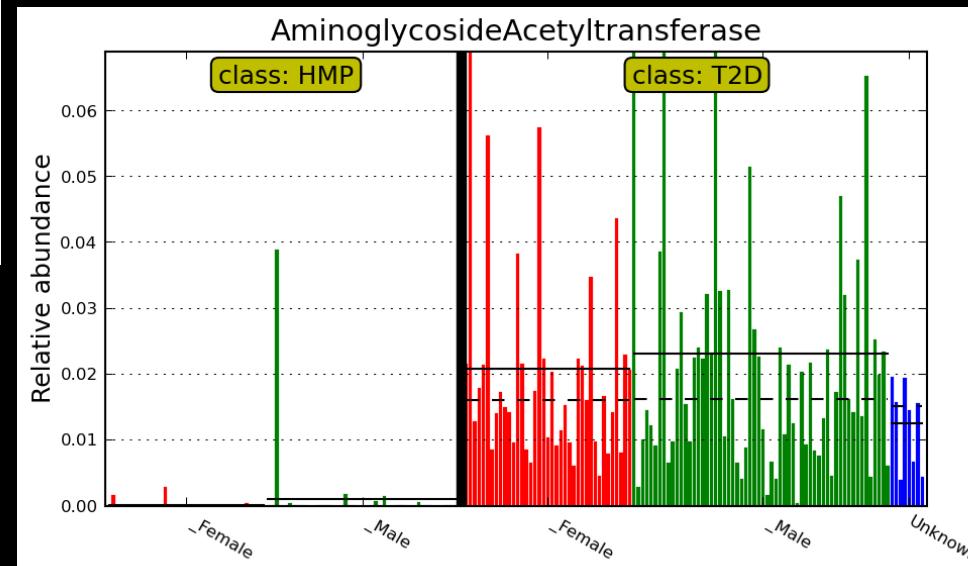
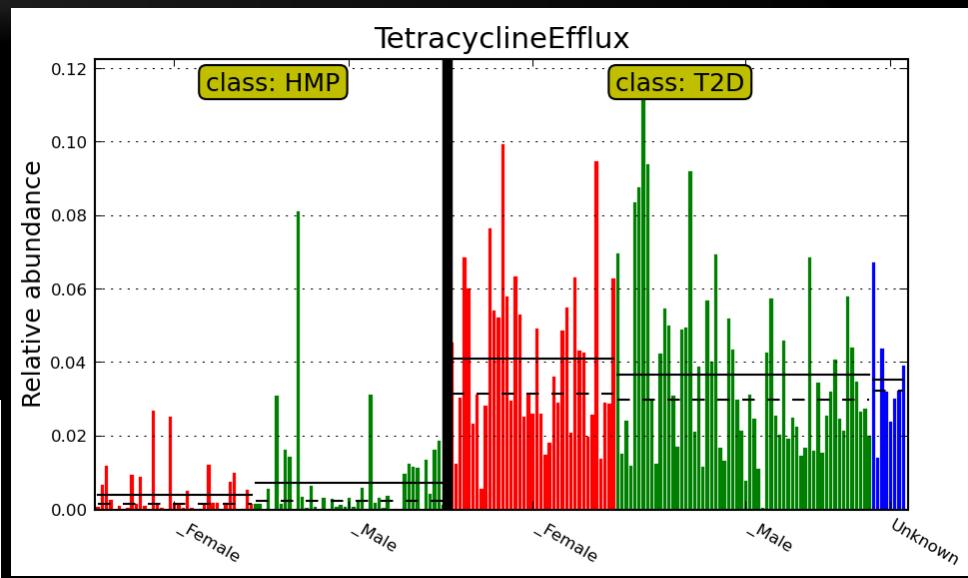
What it means: LEfSe

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

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 (MetaPhiAn) 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	MetaPhiAn • Microbial species and abundances	PhyloPhiAn • 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 MetaPhiAn, 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



Dirk Gevers



George
Weingart



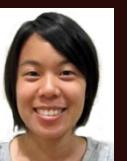
Emma
Schwager



Eric
Franzosa



Boyu
Ren



Tiffany
Hsu



Ali
Rahnavard



Ramnik Xavier



Ayshwarya
Subramanian



Jim
Kaminski



Regina
Joice



Koji
Yasuda



Kevin
Oh



Galeb
Abu-Ali



Wendy Garrett



Nicola Segata



Gautam Dantas
Molly Gibson



Afrah
Shafquat



Randall
Schwager



Chengwei
Luo



Keith
Bayer



Moran
Yassour



Alexandra
Sirota



Andy Chan



Katherine
Lemon



Brendan Bohannan
James Meadow

Human Microbiome Project 2

Lita Procter
Jon Braun
Dermot McGovern
Subra Kugathasan
Ted Denson
Janet Jansson



Human Microbiome Project

Jane Peterson
Sarah Highlander
Barbara Methe

Karen Nelson
George Weinstock
Owen White

