Introduction to shotgun meta'omic analysis

Eric A. Franzosa, Ph.D.
Galeb Abu-Ali, Ph.D.

Harvard CFAR Workshop on Metagenomics
17 September 2015

Huttenhower Research Group
Harvard Chan School of Public Health
Department of Biostatistics

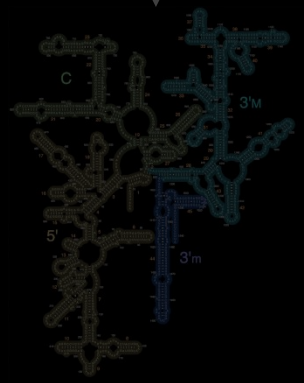
MiRiBA

BROAD INSTITUTE

Plan

- **Shotgun meta'omics primer**
- **Informal survey**
- **Meta'omic taxonomic profiling**
  - MetaPhlAn(2)
- **Meta'omic functional profiling**
  - Broad functional profiling with HUMAnN(2)
  - Targeted functional profiling with ShortBRED
  - Predictive functional profiling with PICRUSt
- **Downstream analyses**
- **Resources**
- **Tutorials (today and tomorrow)**

Sequencing as a tool for microbial community analysis (amplicon vs. shotgun)



Lyse cells

Extract & fragment DNA

Sequence short DNA reads

AGCTAGA CCGATCG
TTAGCAC ACTAGCA

Assemble into contigs

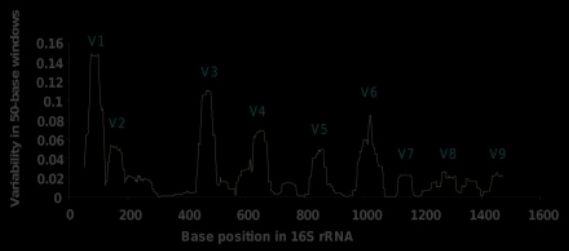
AGCTACAGC
ACAGCACGGCAT
GGCATCATC
AGCTACAGCACGGCATCATC

Samples

Features

- 16S rRNA gene
- Conserved across bacteria
- (Allows PCR amplification)

- Some regions are variable
- Permits genus-level ID



A note on “metagenomics” vocabulary

- **Amplicon sequencing**
  - PCR amplify and seq. specific marker(s)
  - Often the 16S rRNA gene (for bacteria)

- **Shotgun sequencing**
  - Seq. short, random DNA/RNA fragments
  - Whole metagenome shotgun (WMS)
  - Whole metatranscriptome shotgun
  - Collectively, meta’omic sequencing

Sequencing as a tool for microbial community analysis (amplicon vs. shotgun)

- Where they overlap
  - Strengths
    - Quantifying taxonomic abundance
    - Ecological statistics
    - Taxon-taxon association
    - Taxon-metadata association
  - Challenges
    - Compositional (& noisy) data
    - Difficult distributions
    - Biases from sequencing

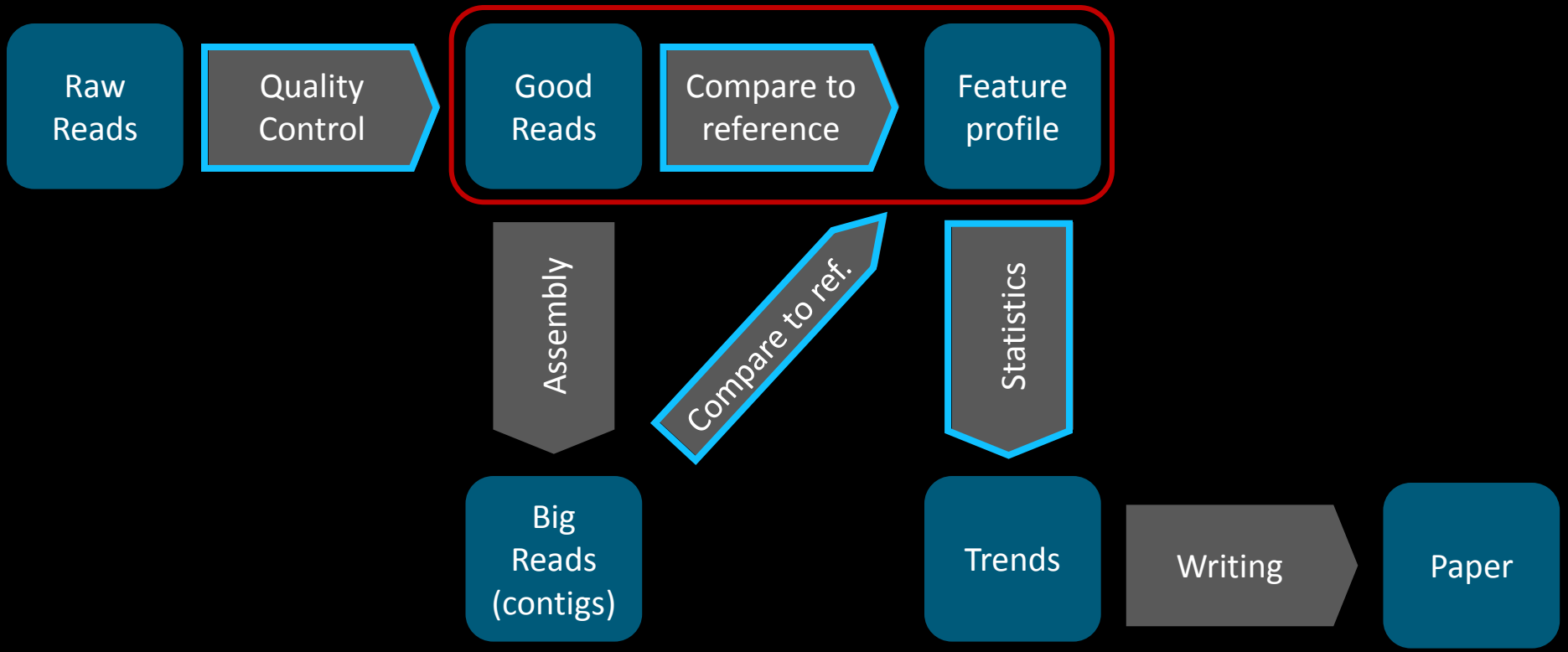
Sequencing as a tool for microbial community analysis (amplicon vs. shotgun)

- Properties of shotgun meta'omic sequencing
  - Strengths
    - Taxonomic resolution (species, strains)
    - Functional genomics (genes, transcripts)
    - Comparative genomics
  - Challenges
    - More expensive per sample
    - Data are bigger, compute more intensive
    - Need a good reference
    - Contamination

Survey: who has/plans to work with…

- 16S data?
- shotgun data?
- metatranscriptomes?
- human vs. environmental samples?
- metagenomic assemblies?

The universal meta'omics workflow



We develop computational methods in these areas

Today we'll be focusing on this subset

Meta'omic quality control

- Trim low-quality bases from read ends
  - http://www.usadellab.org/cms/?page=trimmomatic
- Drop short reads
- Remove contaminant sequences
  - E.g. human genome, EST database
- Remove low-complexity sequences (?)
- Enforce end-pairing (?)
  - Not required for bioBakery tools
- Integrated workflow coming soon!
  - https://bitbucket.org/biobakery/kneaddata

Meta'omics seeks to answer two big questions…

Who is there?
(taxonomic profiling)

What are they doing?
(functional profiling)

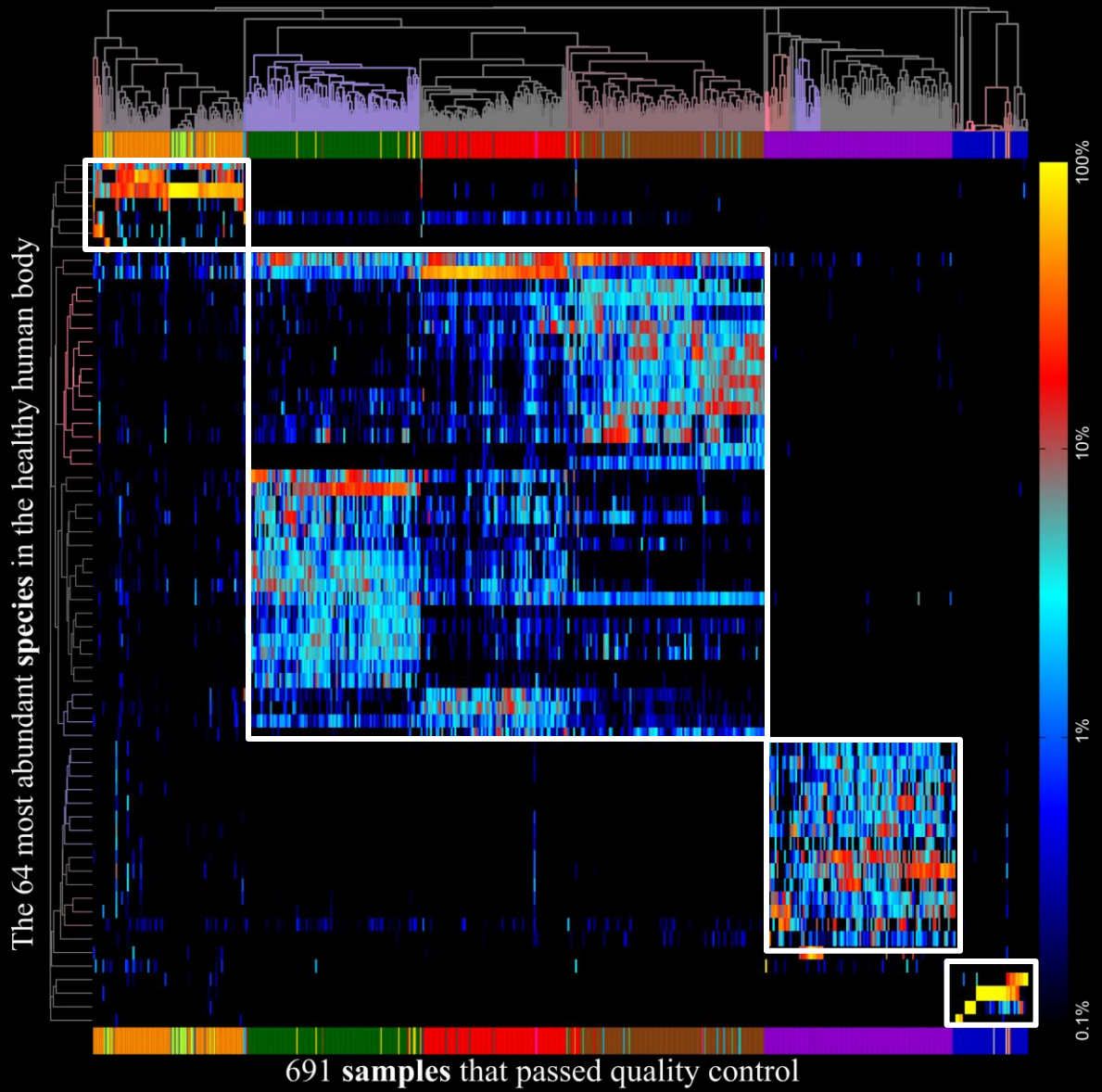
Slide by Dirk Gevers

The NIH Human Microbiome Project (HMP): A comprehensive microbial survey

- ***What is a "normal" human microbiome?***
- 300 healthy human subjects
- Multiple body sites
  - 15 male, 18 female
- Multiple visits
- Clinical metadata

www.hmpdacc.org

Profiling microbial communities and ecology at species-level resolution (HMP)



Skin (nares)

Oral (plaque)
Oral (cheek)
Oral (tongue)

Gut (stool)

Vaginal (fornix)

Profiling microbial communities and ecology at species-level resolution

ARTICLE

Arumugam Nature 2012

doi:10.1038/nature09944

Enterotypes of the human gut microbiome

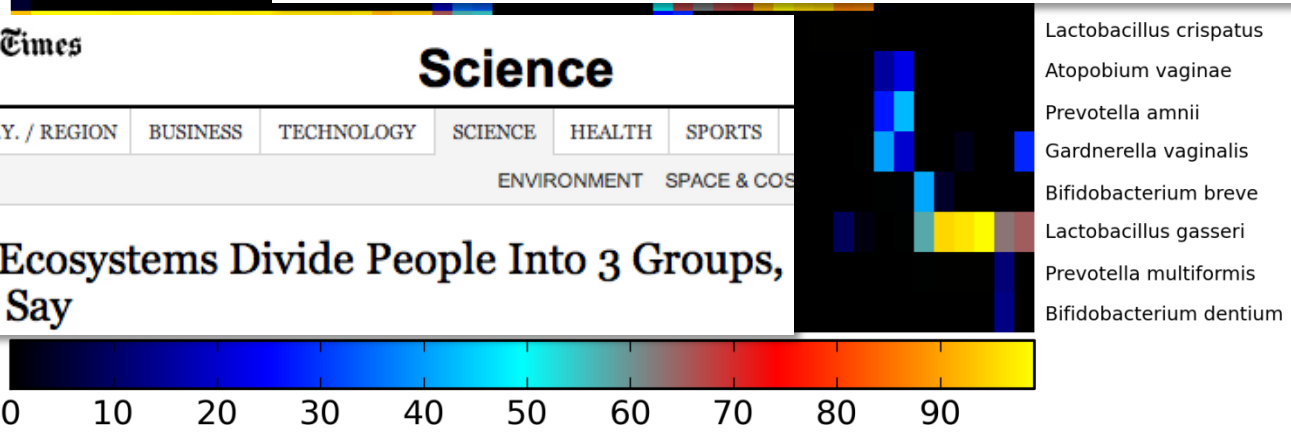
The New York Times

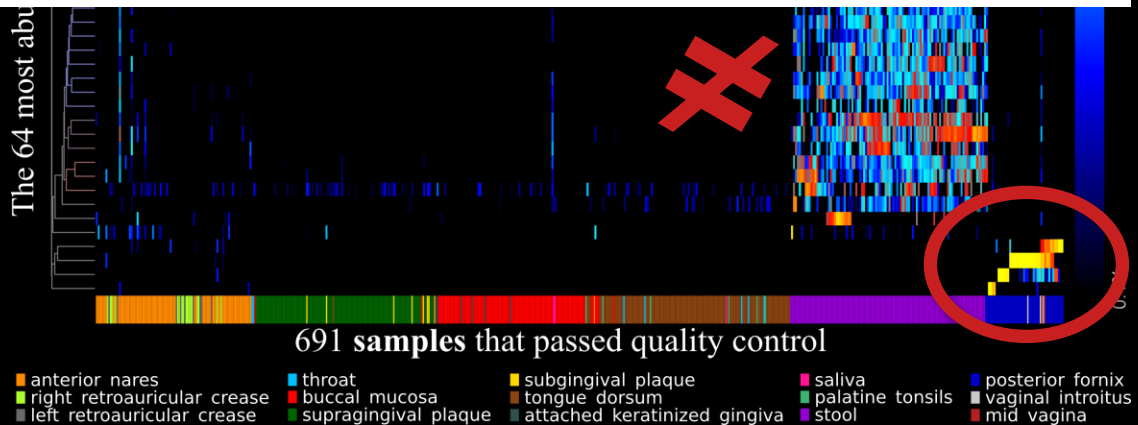
Science

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS

ENVIRONMENT | SPACE & COS

Bacterial Ecosystems Divide People Into 3 Groups, Scientists Say

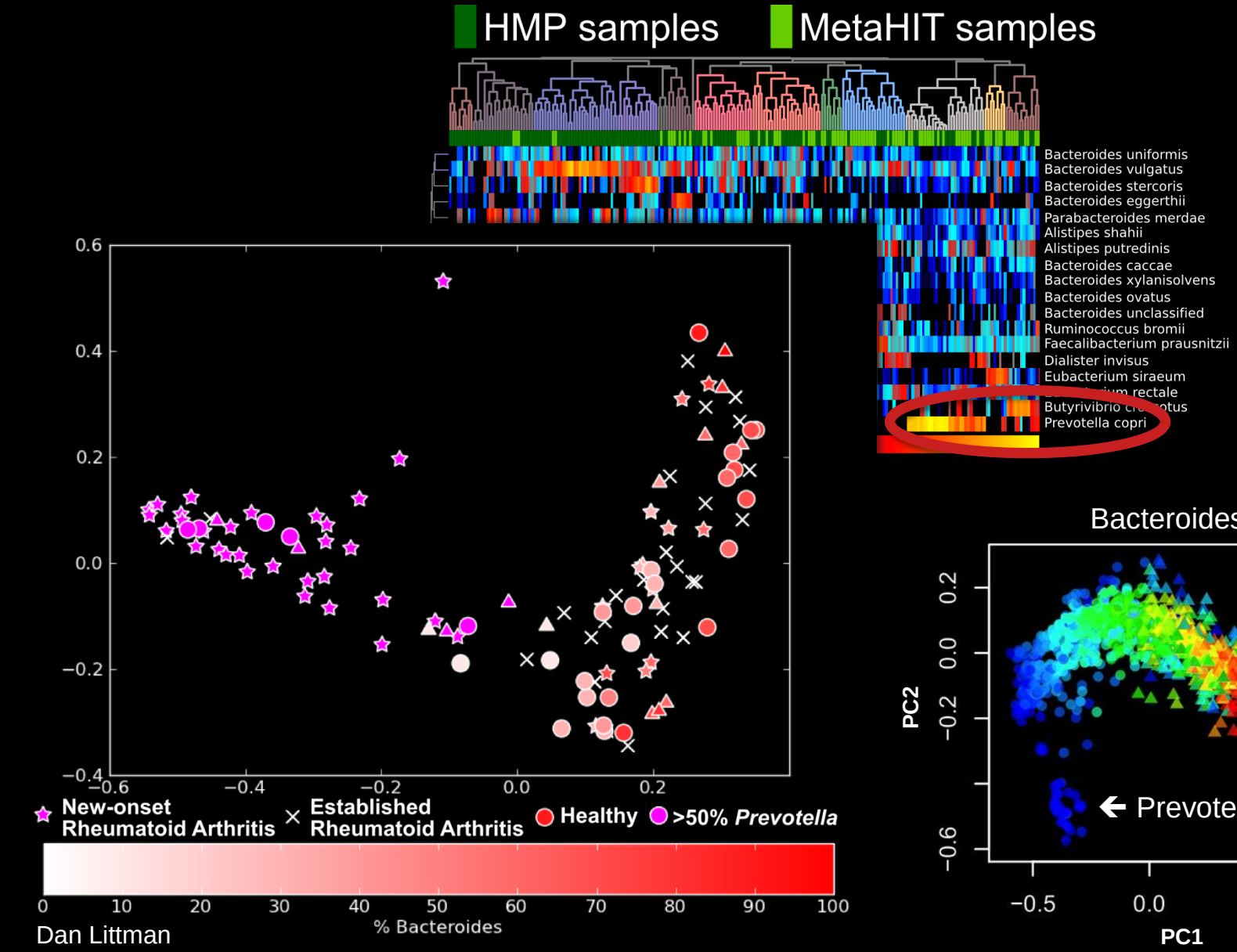




Gut

Vaginal

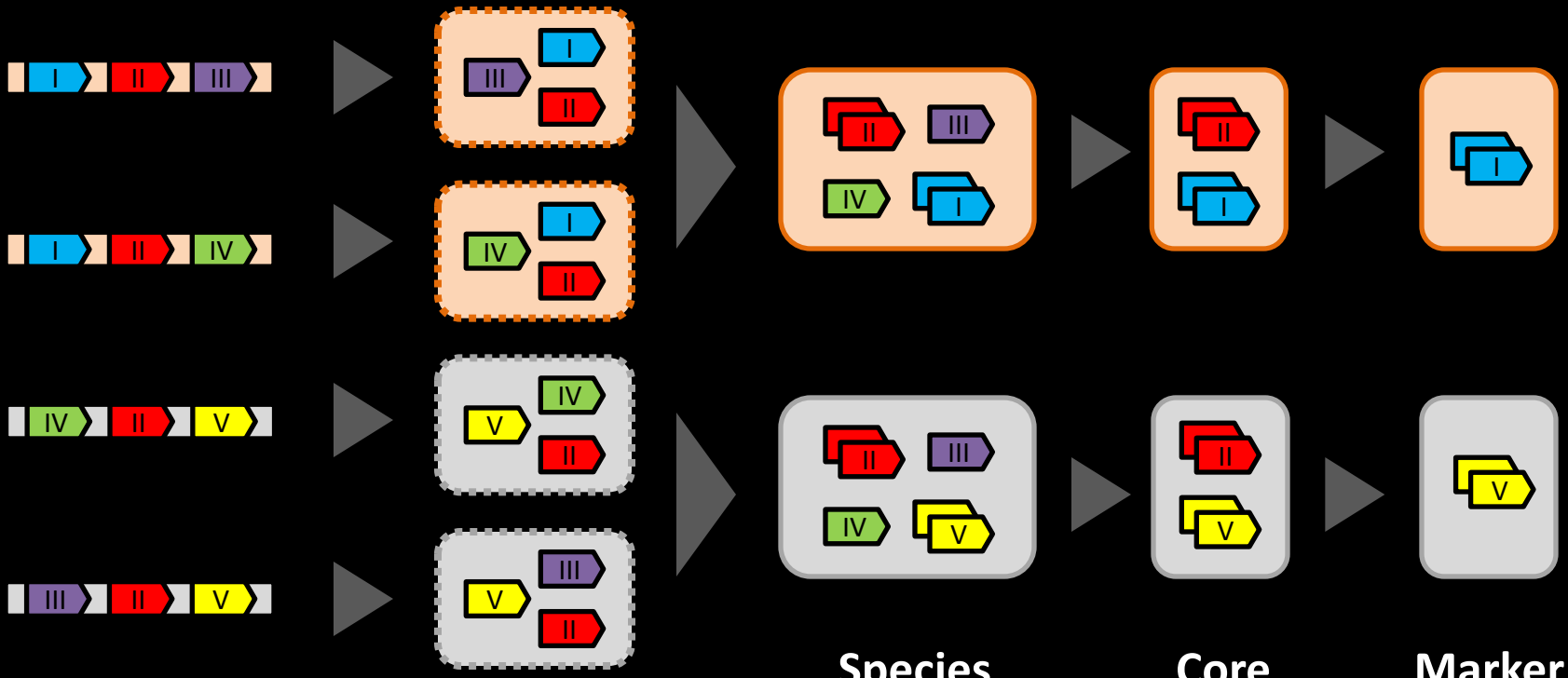
Are there discrete “types” of typical human microbiomes?



MetaPhlAn(2)

For meta’omic taxonomic profiling

Comprehensive pangenome reference db



NCBI isolate genomes

Archaea	300
Bacteria	12,926
Viruses	3,565
Eukaryota	112

protein-coding genes

49.0 million total genes

Species pangenomes

7,677 containing 18.6 million gene clusters

Core genes

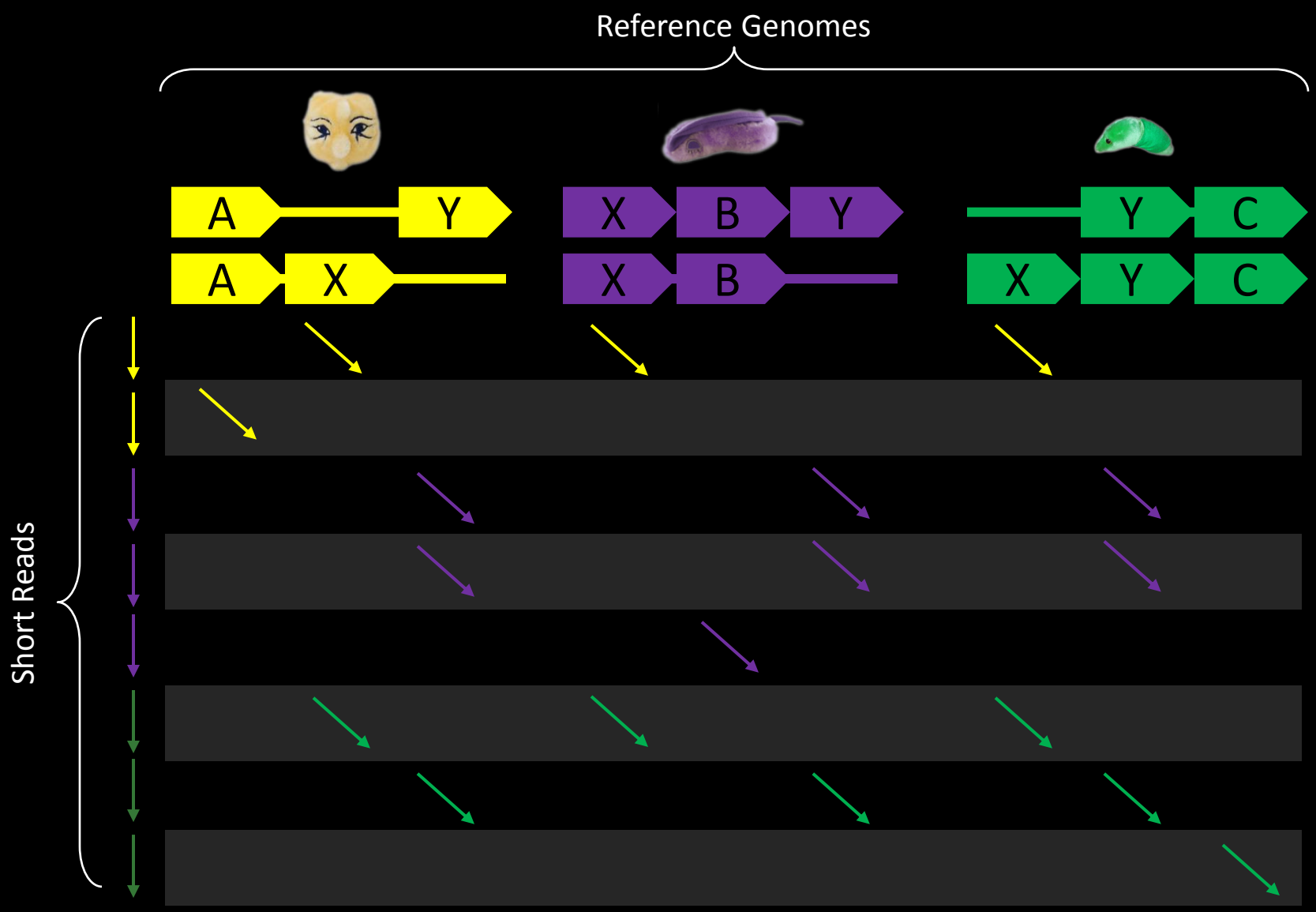
Marker genes

Nicola Segata

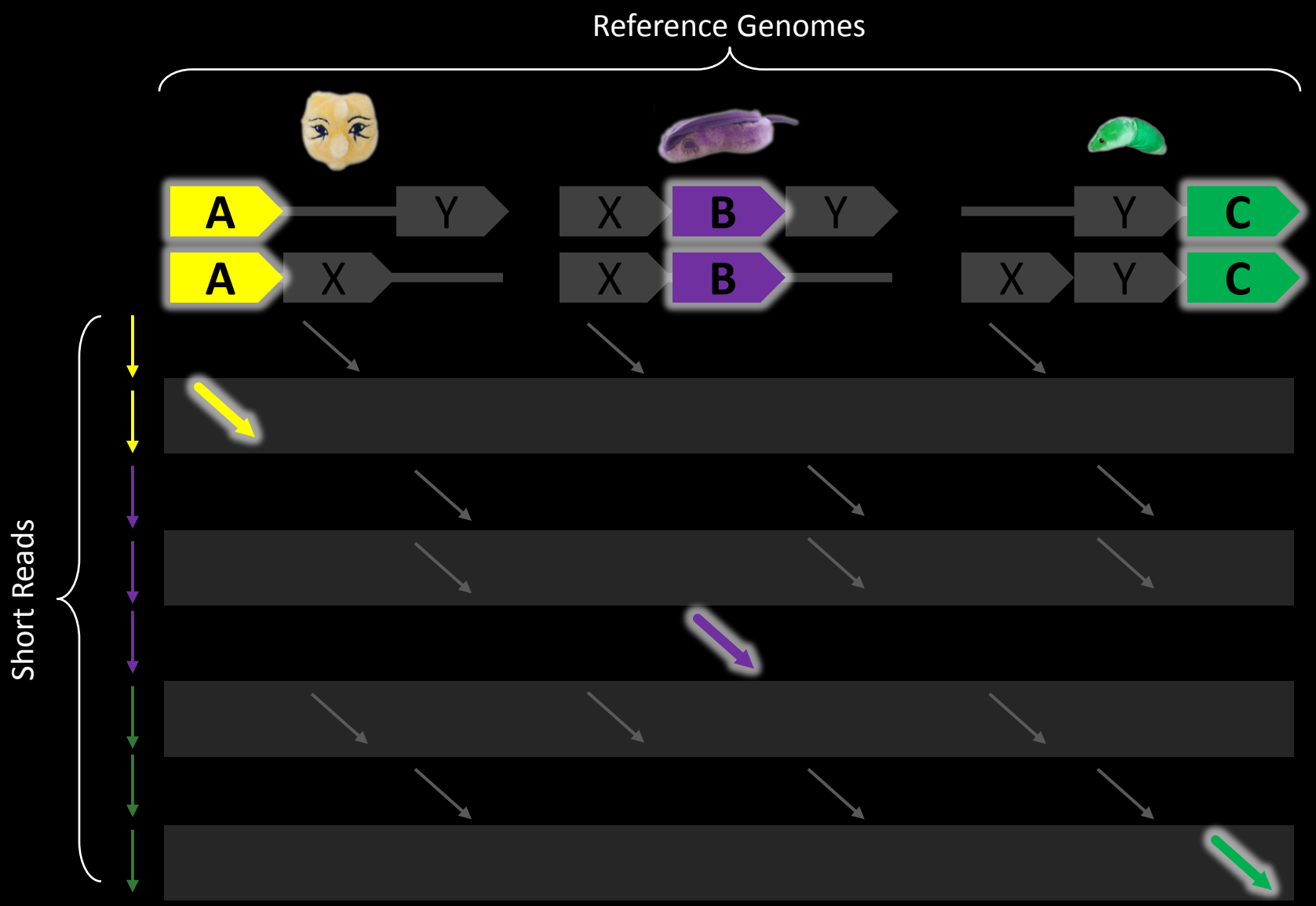
RepoPhlAn

ChocoPhlAn (http://metaref.org)

MetaPhlAn
Metagenomic Phylogenic Analysis



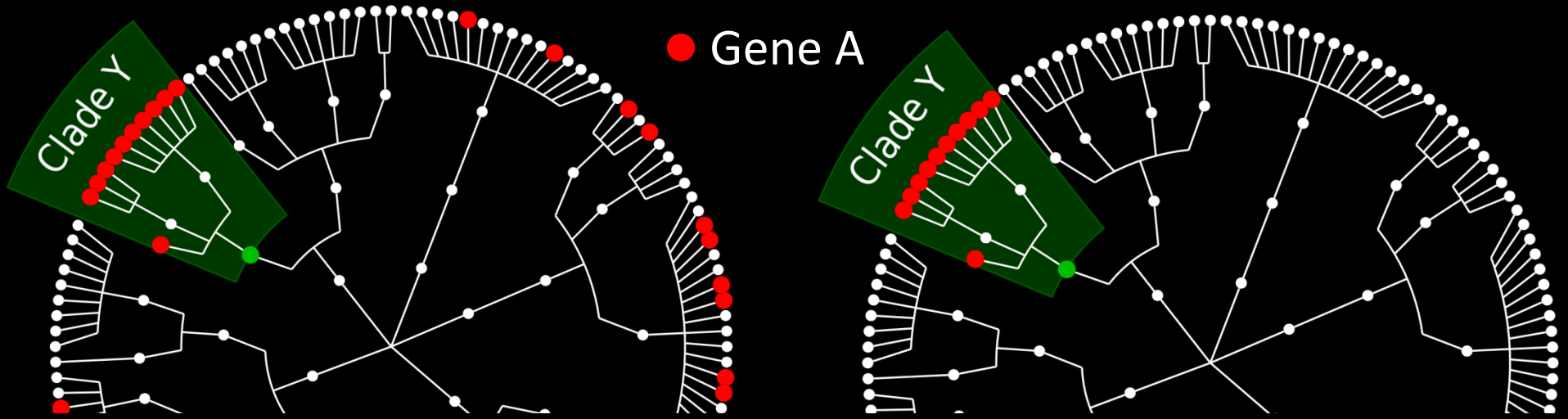
MetaPhlAn
Metagenomic Phylogenic Analysis



MetaPhlAn overview

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

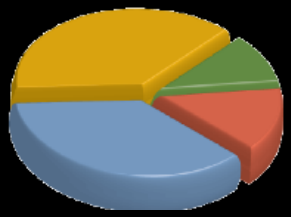
Unique marker genes DB

Available reference genomes

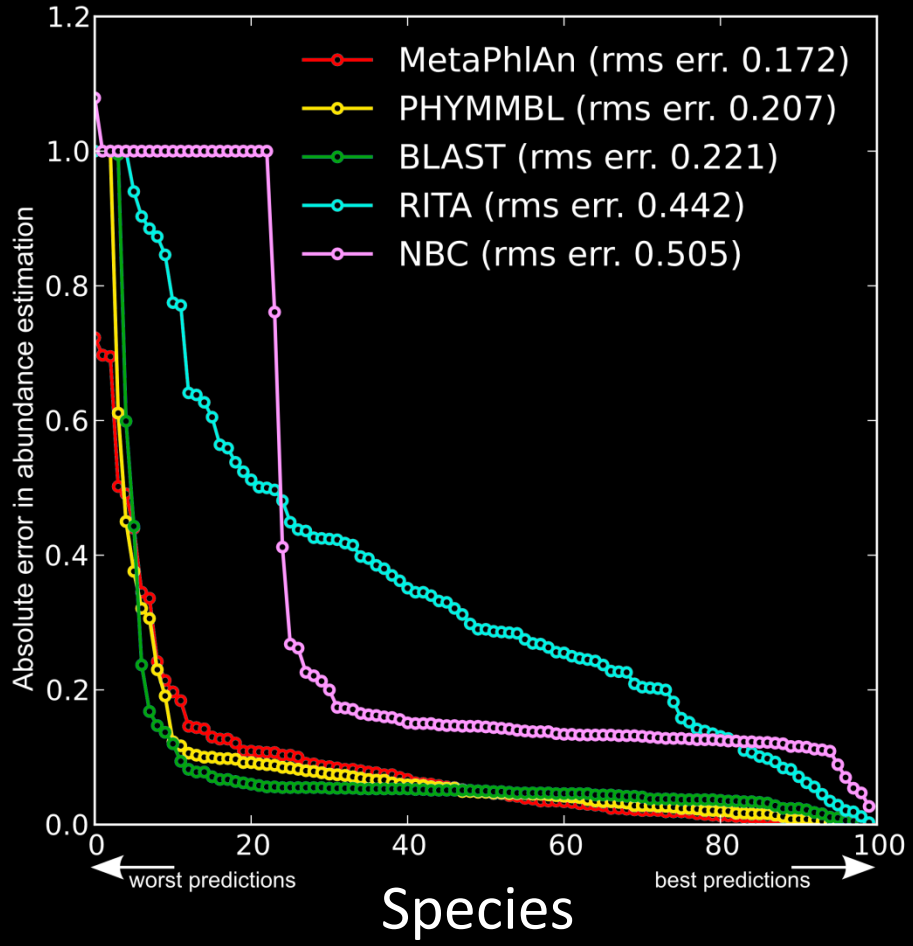
MetaPhlAn

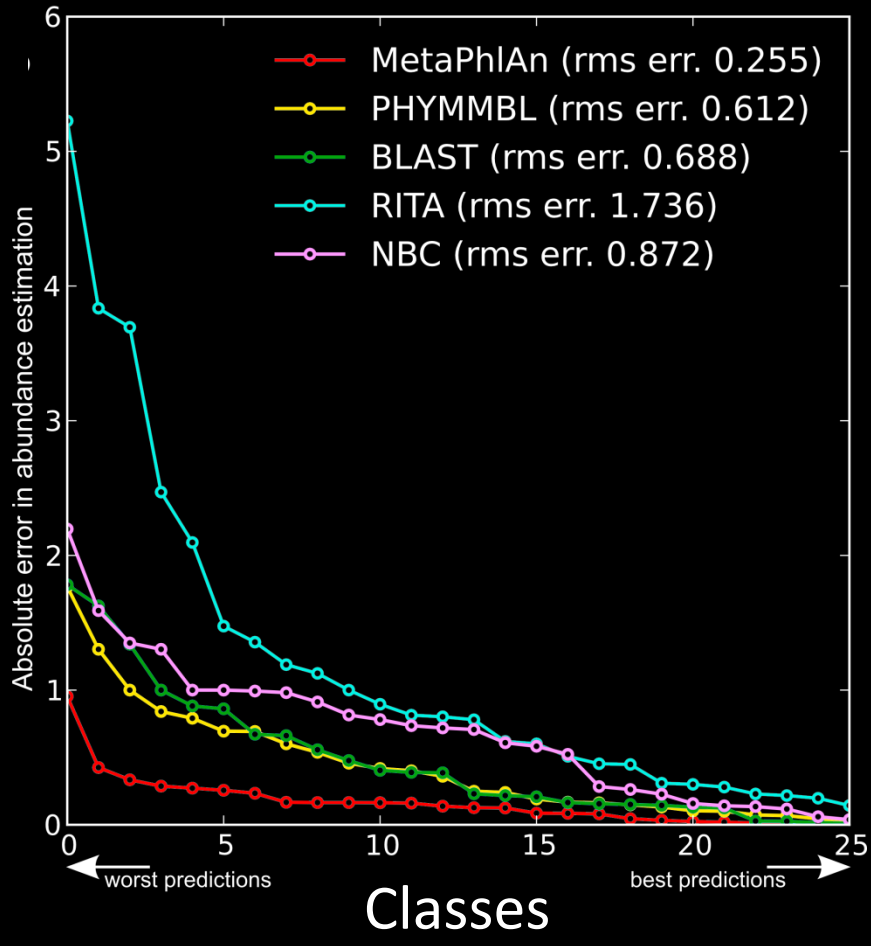
- Blast reads agains the marker genes
- Assign, count, normalize reads

Metagenome



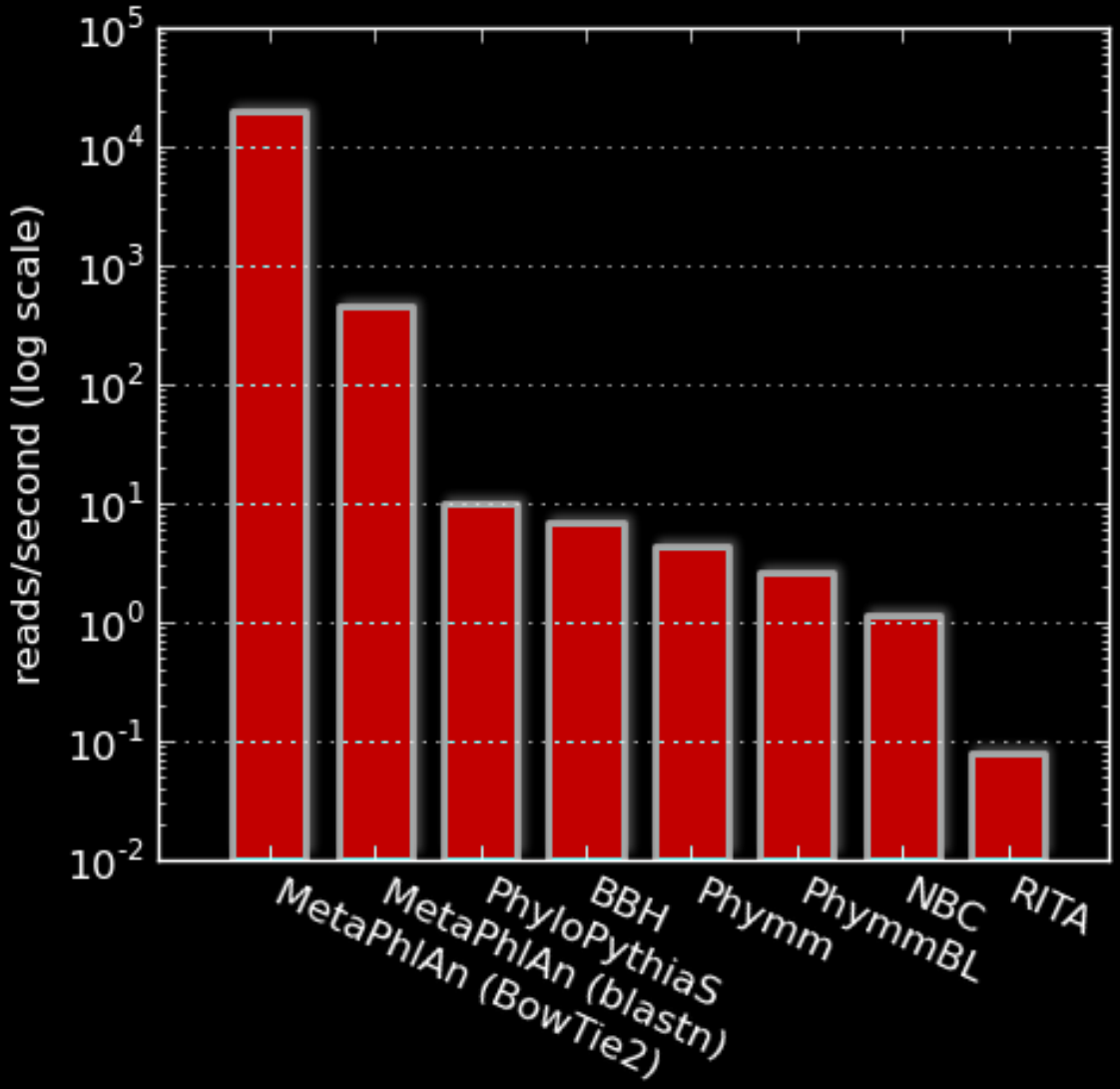
Evaluation of MetaPhlAn accuracy





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

Evaluation of MetaPhlAn performance



>50 times faster than earlier methods

450 reads/sec (BLAST)

Up to 25,000 reads/sec (bowtie2)

Multi-threaded

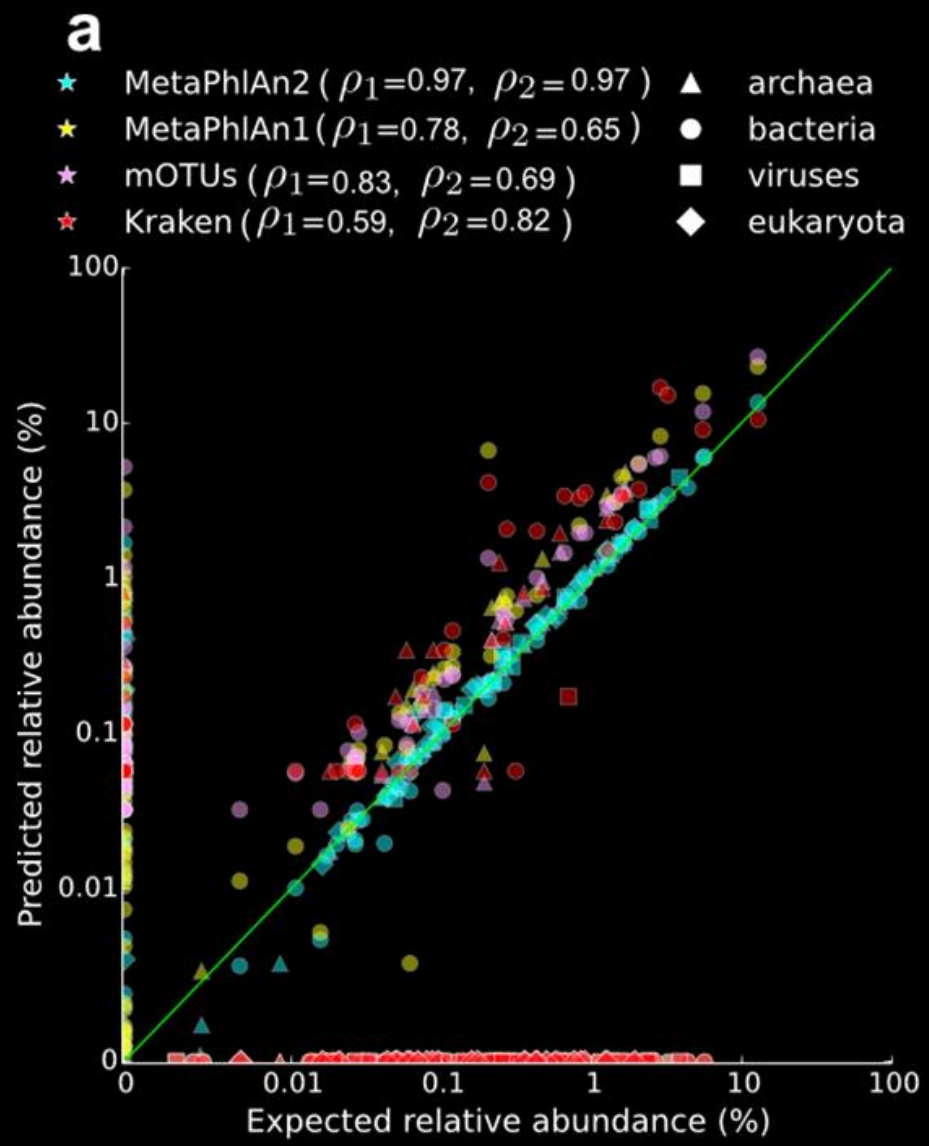
Easily parallelizable

MetaPhlAn 2.0

- MetaPhlAn 1.0 focused on bacteria and archaea
- v2.0 adds support for eukaryotes and viruses
- ... along with many more bacteria and archaea
- v2.0 supports profiling at the strain level

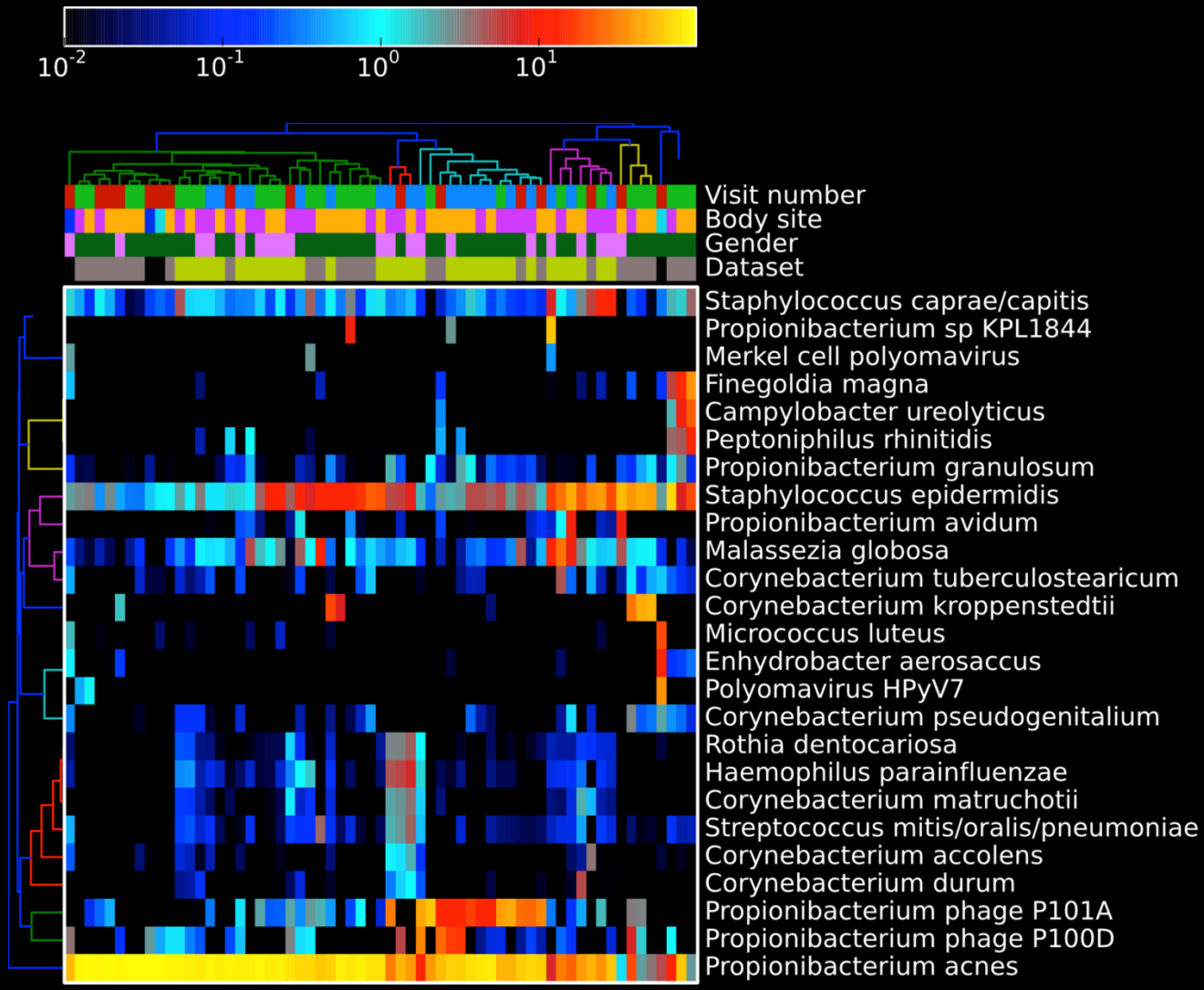
https://bitbucket.org/biobakery/metaphlan2

MetaPhlAn2: synthetic evaluation



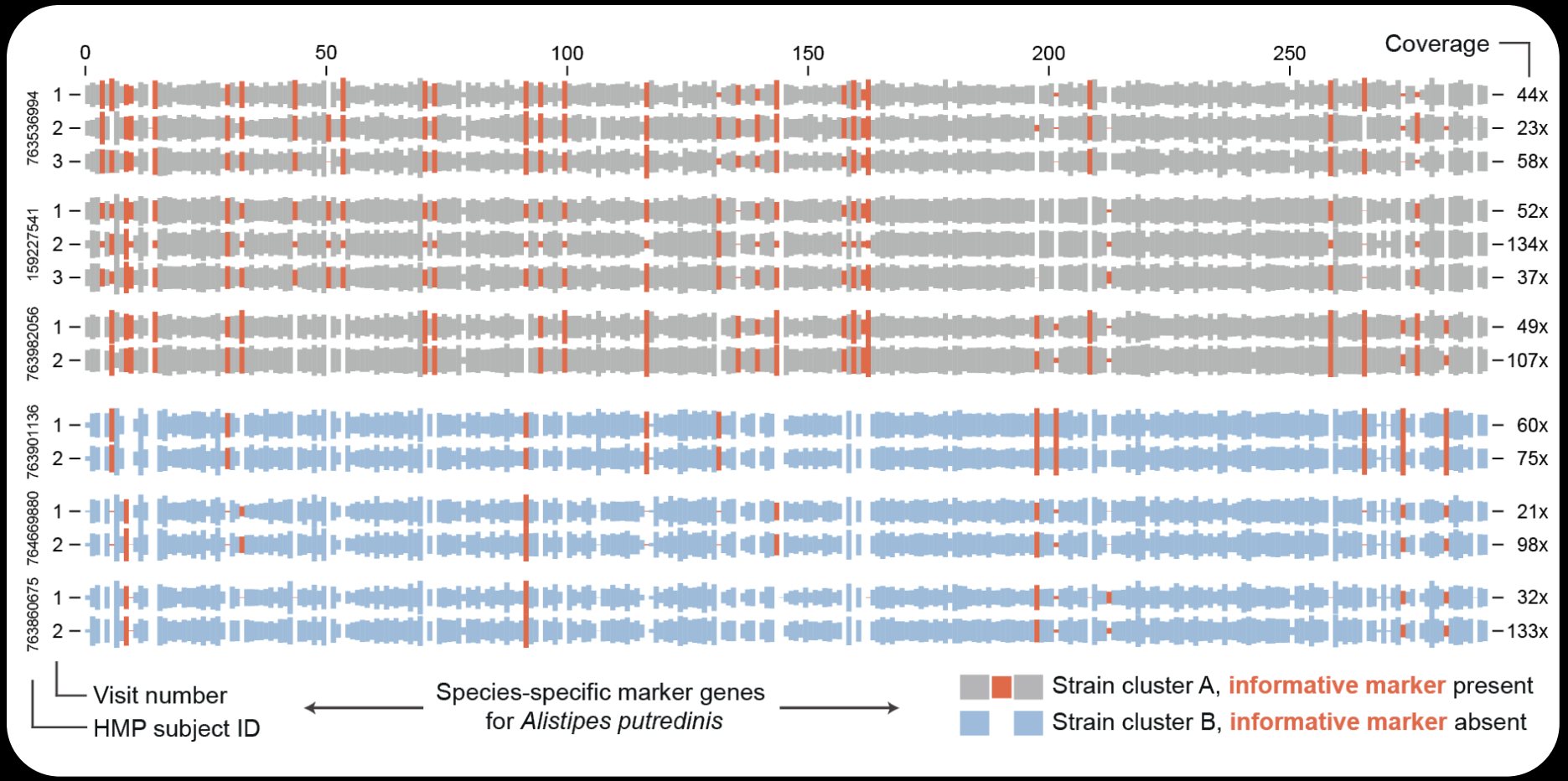
https://bitbucket.org/biobakery/metaphlan2

MetaPhlAn2: Results for HMP Skin



https://bitbucket.org/biobakery/metaphlan2

MetaPhlAn in action: *strain profiling*



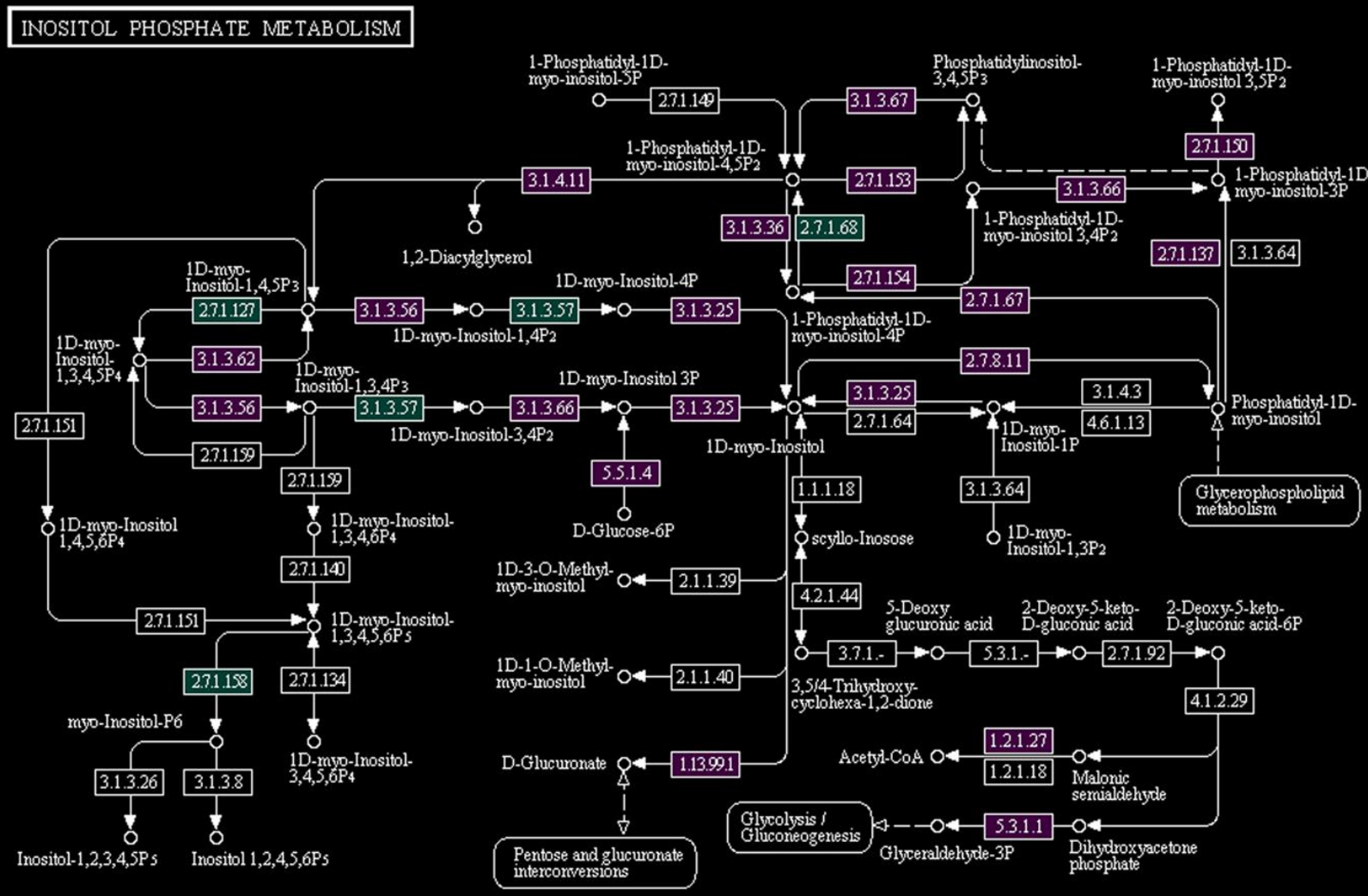
- In practice, not all markers are present
- Individual-specific marker “barcodes”
- Often very stable over time

Meta'omics seeks to answer two big questions…

Who is there?
(taxonomic profiling)

What are they doing?
(functional profiling)

(What we mean by “function”)

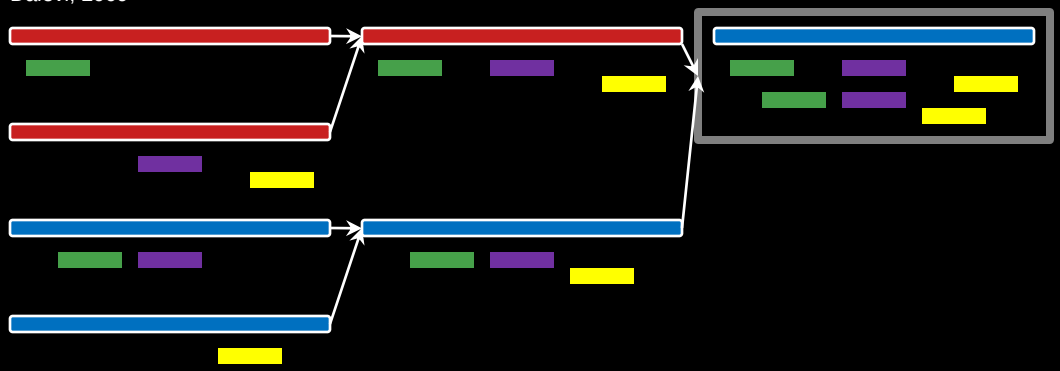


Metagenomic analyses: gene calling and proxygenes

Extrinsic gene calling:
BLAST etc. (proxygenes)

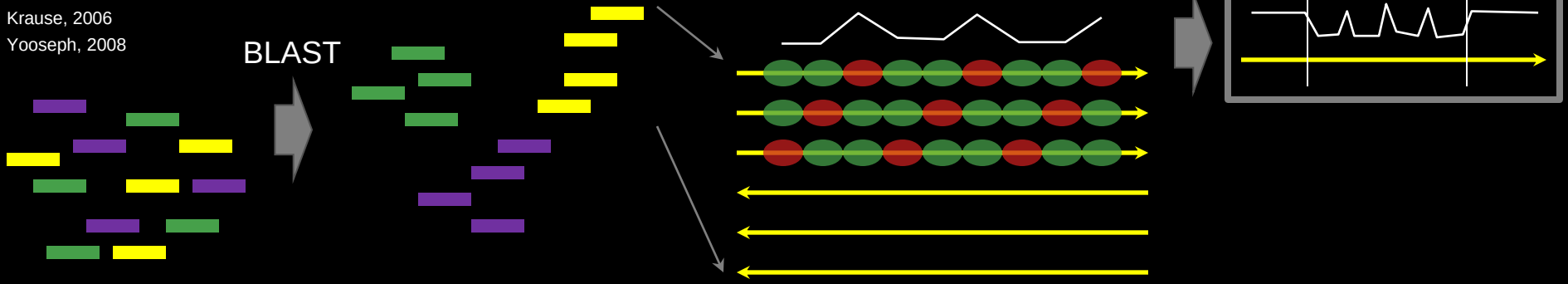
Intrinsic gene calling:
ORF detection from seq.

Dalevi, 2009



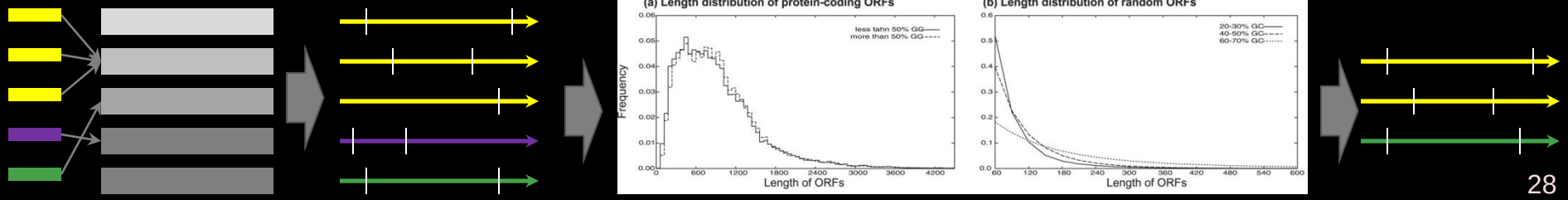
Krause, 2006
Yooseph, 2008

BLAST



Orphelia: Hoff, 2009
MetaGene: Noguchi, 2006

HMM models

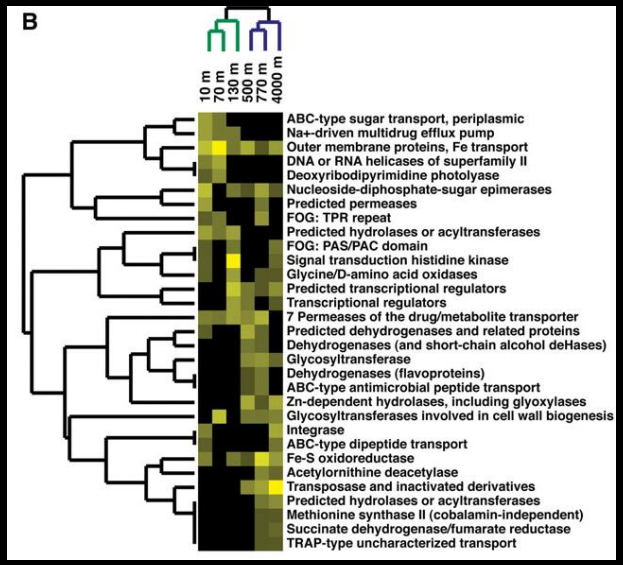


Metagenomic analyses: molecular functions and biological roles

Orthology:
Grouping genes by conserved sequence features
COG, KO, FIGfam…

Structure:
Grouping genes by similar protein domains
Pfam, TIGRfam, SMART, EC…

Biological roles:
Grouping genes by pathway and process involvement
GO, KEGG, MetaCyc, SEED…

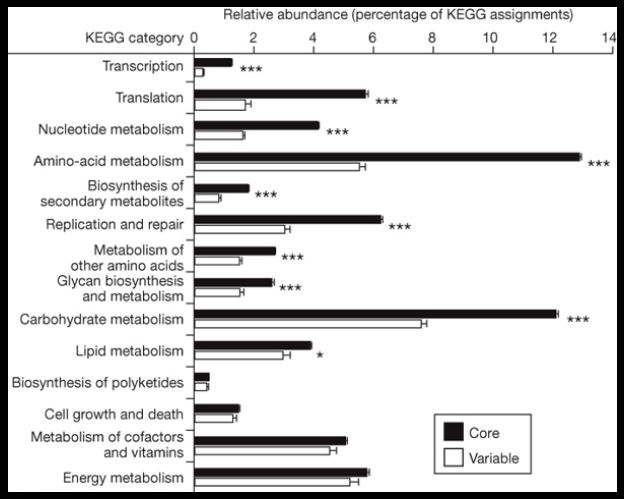


DeLong, 2006

Table 1 | Glycoside hydrolases and carbohydrate-binding modules

CAZy family*	Pfam HMM name†	Known activities‡	Termite gut community§
Glycoside hydrolase catalytic domains**††			
GH1	Glyco_hydro_1	β-Glucosidase, β-galactosidase, β-mannosidase, others	22
GH2	Glyco_hydro_2_C	β-Galactosidase, β-mannosidase, others	23
GH3	Glyco_hydro_3	β-1,4-Glucosidase, β-1,4-xylosidase, β-1,3-glucosidase, α-L-arabinofuranosidase, others	69
GH4	Glyco_hydro_4	α-Glucosidase, α-galactosidase, α-glucuronidase, others	14
GH5	Cellulase	Cellulase, β-1,4-endoglucanase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	56
GH8	Glyco_hydro_8	Cellulase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	5
GH9	Glyco_hydro_9	Endoglucanase, cellobiohydrolase, β-glucosidase	9
GH10	Glyco_hydro_10	Xylanase, β-1,3-endoxylanase	46
GH11	Glyco_hydro_11	Xylanase	14
GH13	Alpha-amylase	α-Amylase, catalytic domain, and related enzymes	48
GH16	Glyco_hydro_16	β-1,3(4)-Endoglucanase, others	1
GH18	Glyco_hydro_18	Chitinase, endo-β-*N*-acetylglucosaminidase, non-catalytic proteins	17
GH20	Glyco_hydro_20	β-Hexosaminidase, lacto-*N*-biosidase	15

Warnecke, 2007

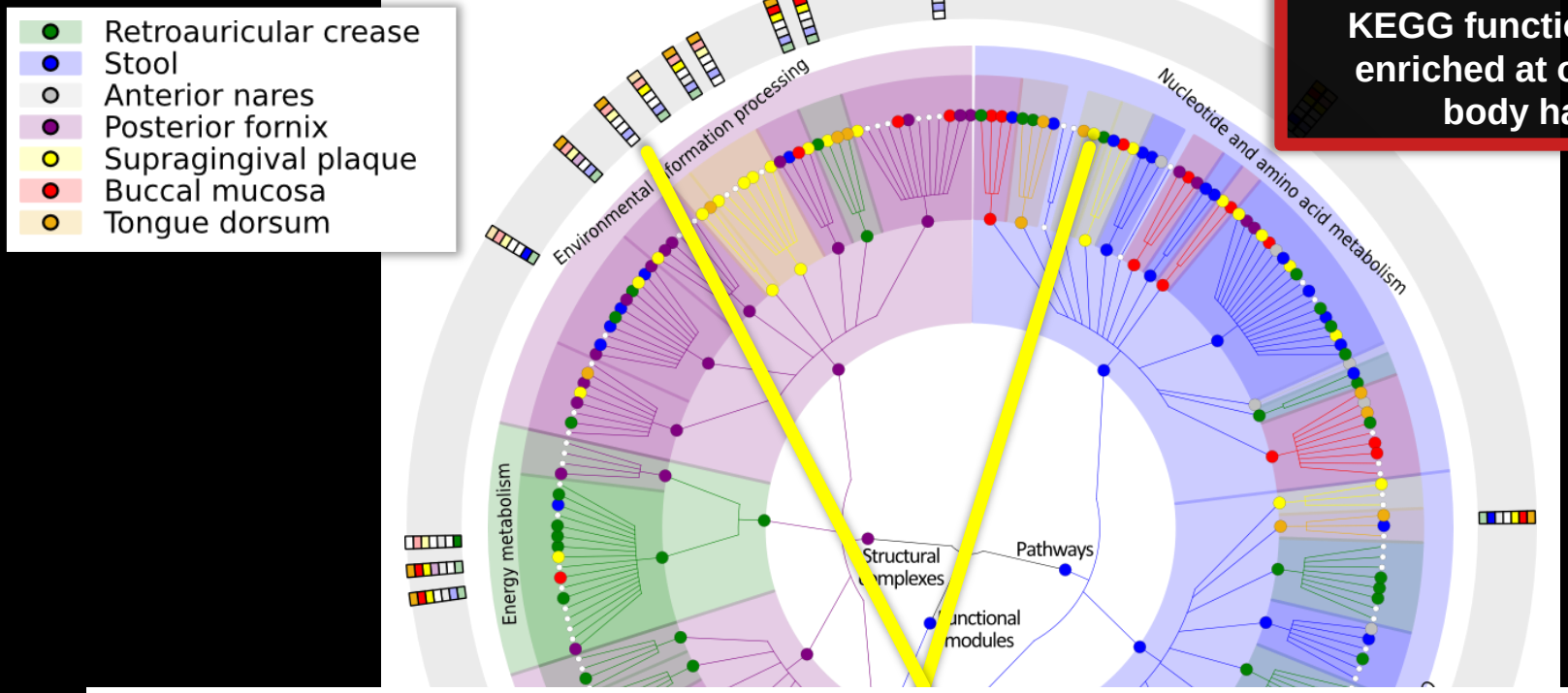


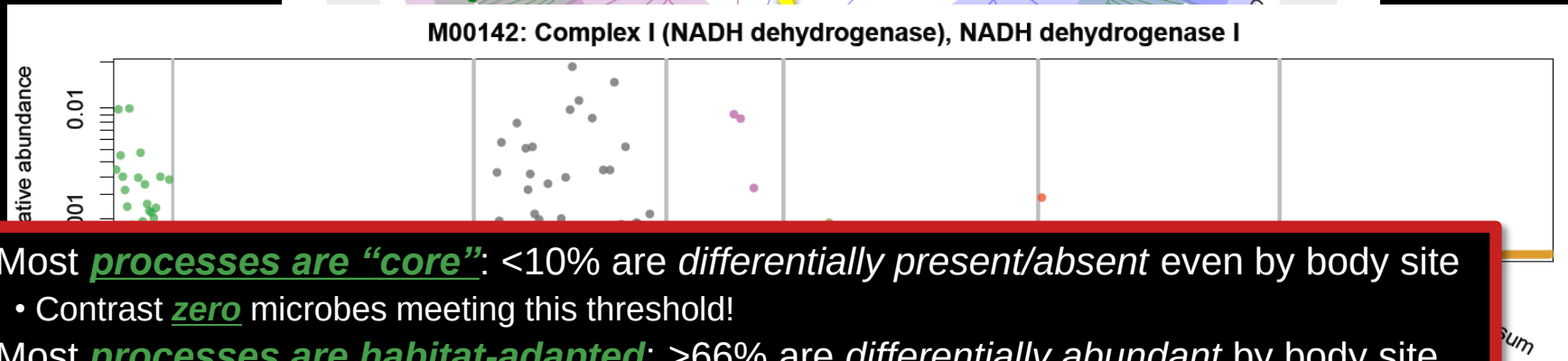
Turnbaugh, 2009

Niche specialization in human microbiome function

LEfSe:
LDA Effect Size
Nonparametric test for microbial and metagenomic biomarkers
http://huttenhower.sph.harvard.edu/lefse

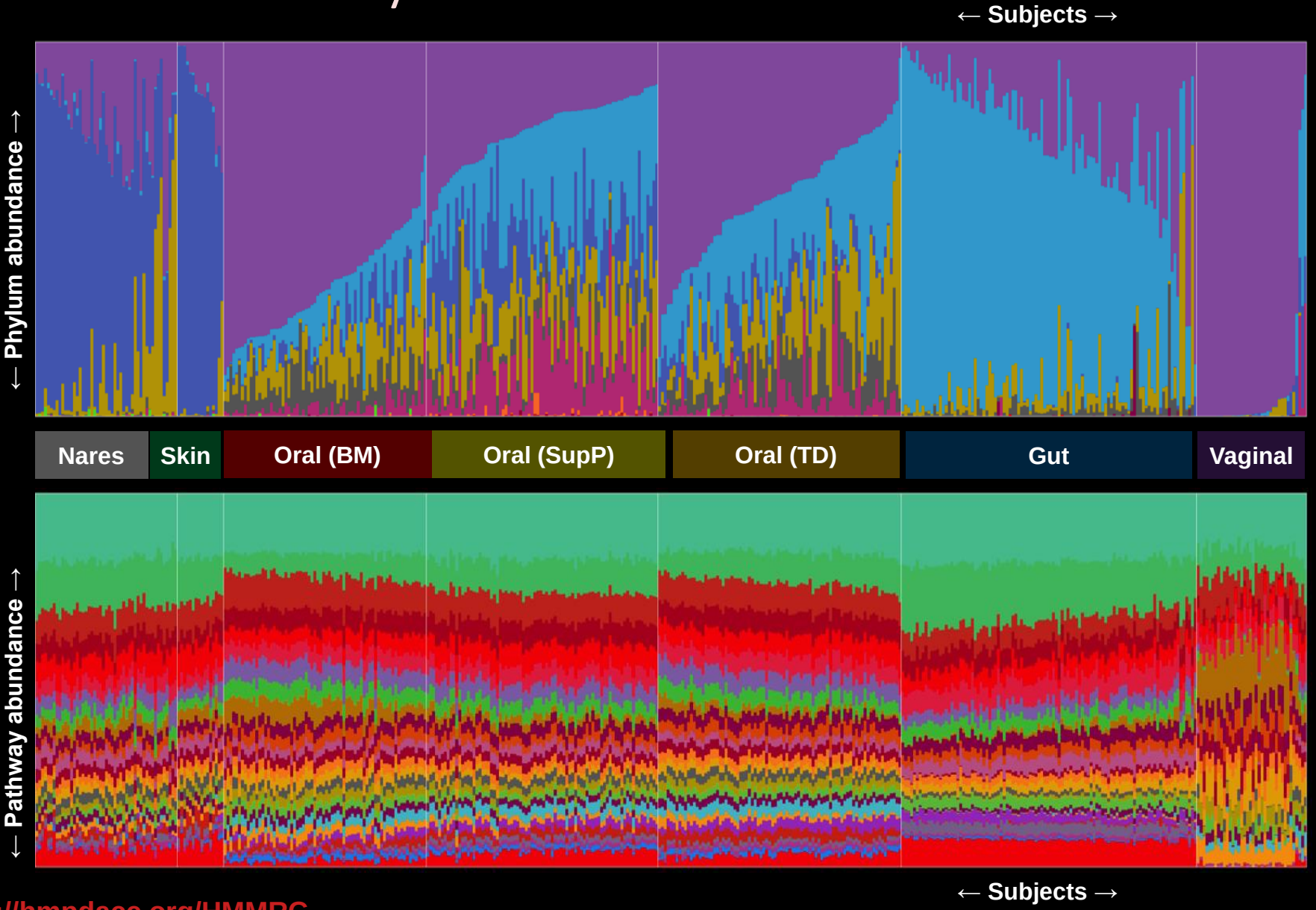
Metabolic modules in the KEGG functional catalog enriched at one or more body habitats





- Most ***processes are "core"***: <10% are *differentially present/absent* even by body site
  - Contrast ***zero*** microbes meeting this threshold!
- Most ***processes are habitat-adapted***: >66% are *differentially abundant* by body site

“Who’s there,” versus, “What they’re doing,” in the healthy human microbiome



Which *functions* of the gut microbiome are disrupted by IBD?

- Over *six times* as many microbial metabolic processes disrupted in IBD as microbes.
  - If there's a transit strike, everyone working for the MBTA is disrupted, not everyone named Smith or Jones.
  - Phylogenetic distribution of function is *consistent* but *diffuse*
- During IBD, microbes...

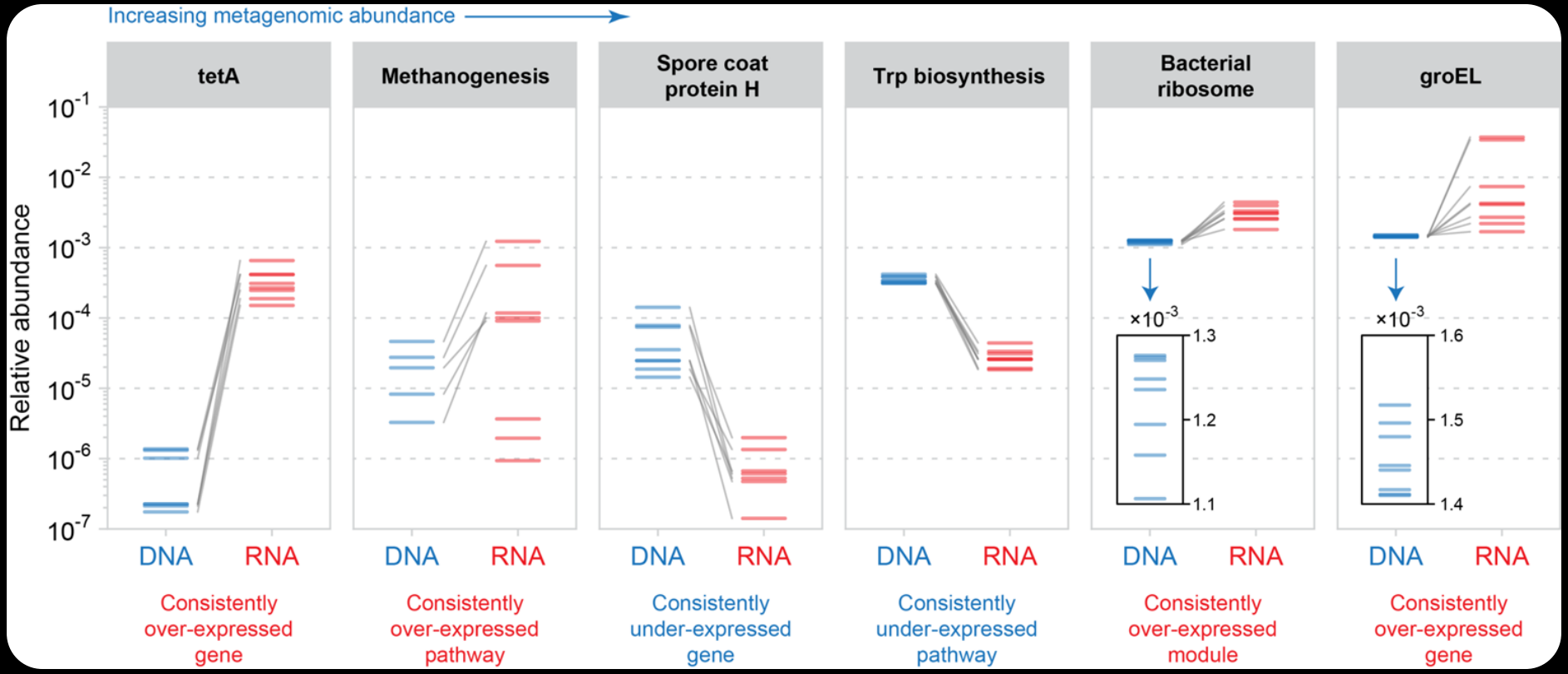
Stop

- Creating most amino acids
- Degrading complex carbs.
- Producing short-chain fatty acids

Start

- Taking up more host products
- Dodging the immune system
- Adhering to and invading host cells

Integrated functional meta'omics (Examining community DNA & RNA)



Functional metagenomics & metatranscriptomics of 8 heathy human stool samples

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

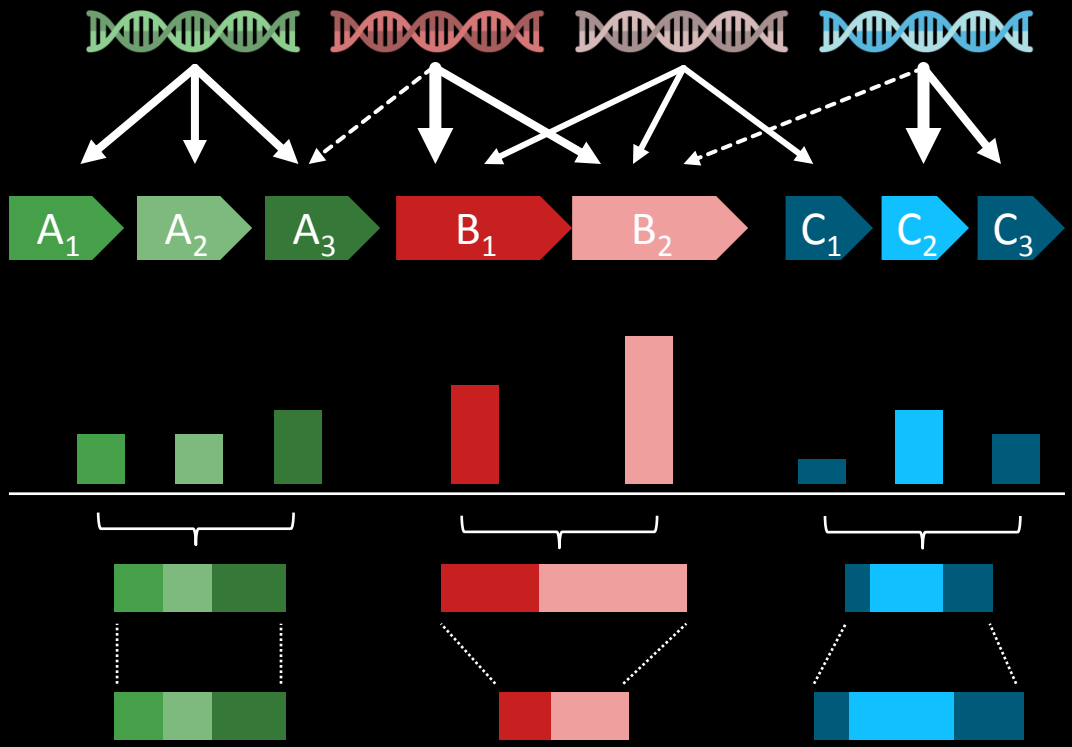
HUMAnN(2)

For broad meta’omic functional profiling

HUMAnN

HMP Unified Metabolic Analysis Network

https://huttenhower.sph.harvard.edu/humann



Short reads + protein families

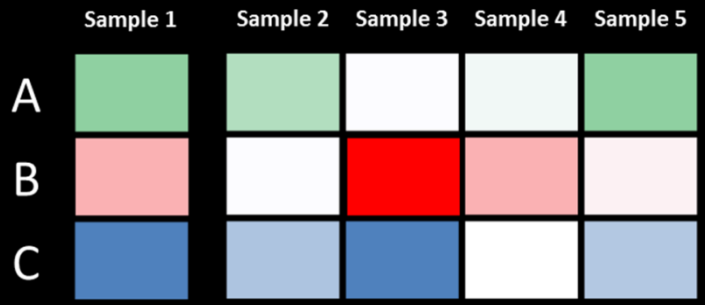
Translated BLAST search

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

Weight hits by significance

Sum over families

Adjust for sequence length

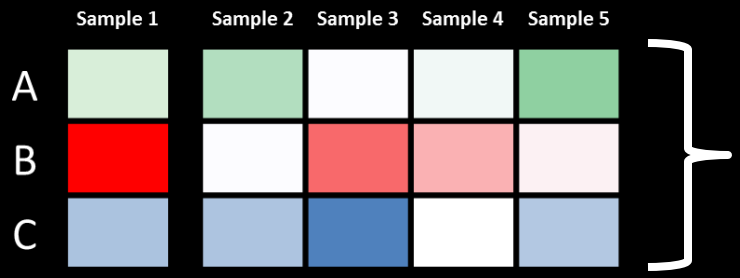


Repeat for each metagenomic or metatranscriptomic sample

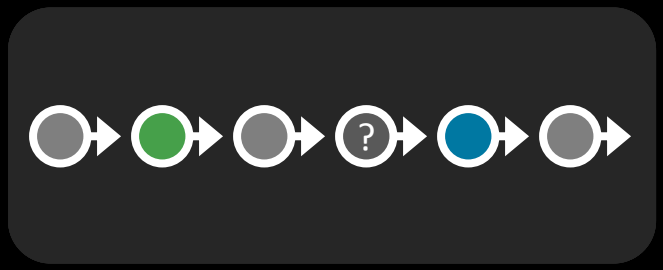
HUMAnN

HMP Unified Metabolic Analysis Network

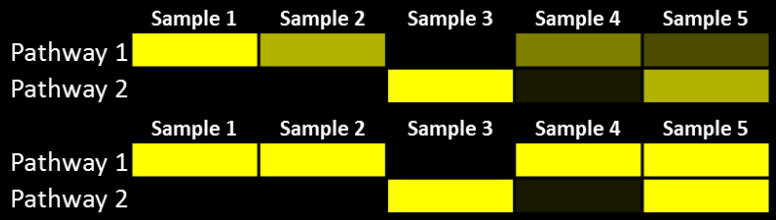
https://huttenhower.sph.harvard.edu/humann



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

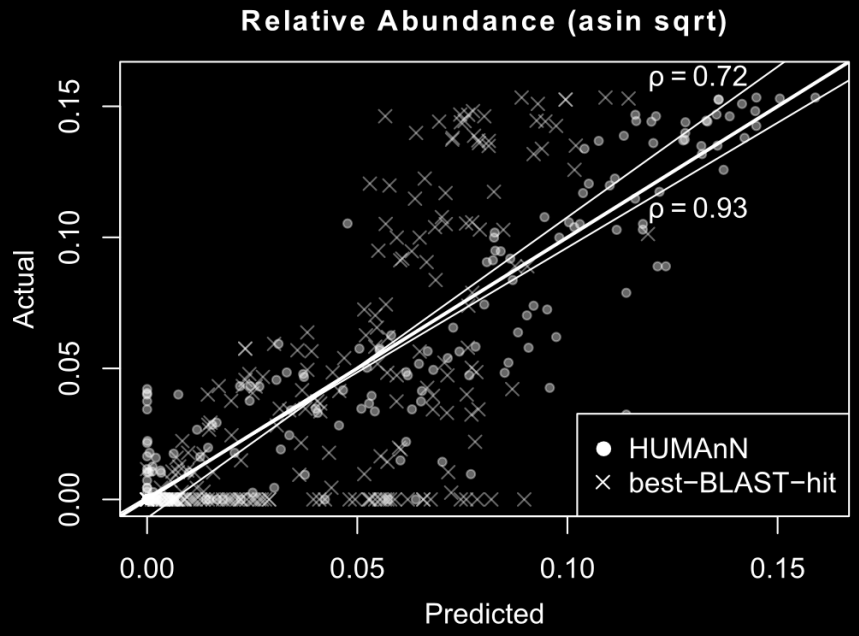


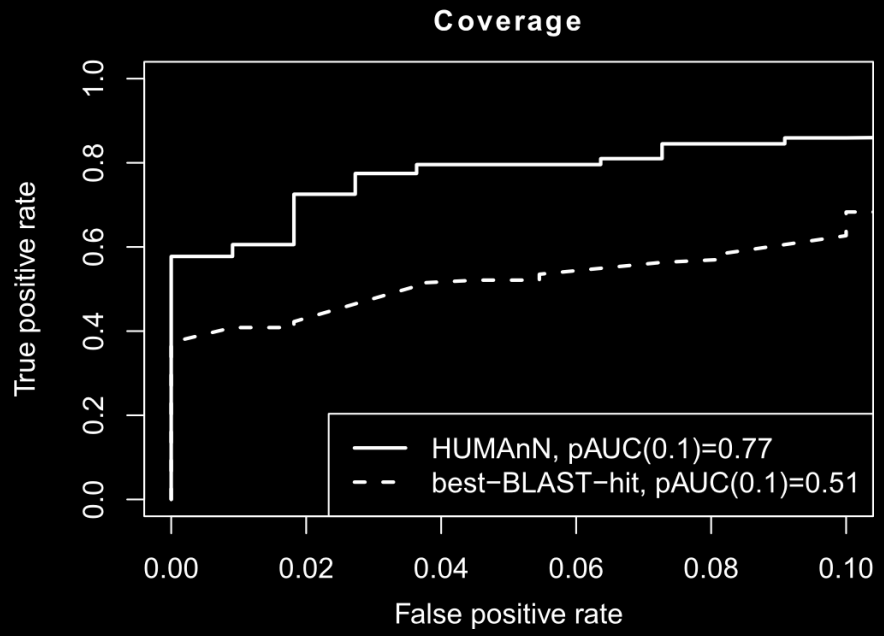
- Map genes to pathways
- Use MinPath (Ye 2009) to find simplest pathway explanation for observed genes
- Remove pathways unlikely to be present due to low organismal abundance
- Smooth/fill gaps



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

HUMAnN accuracy





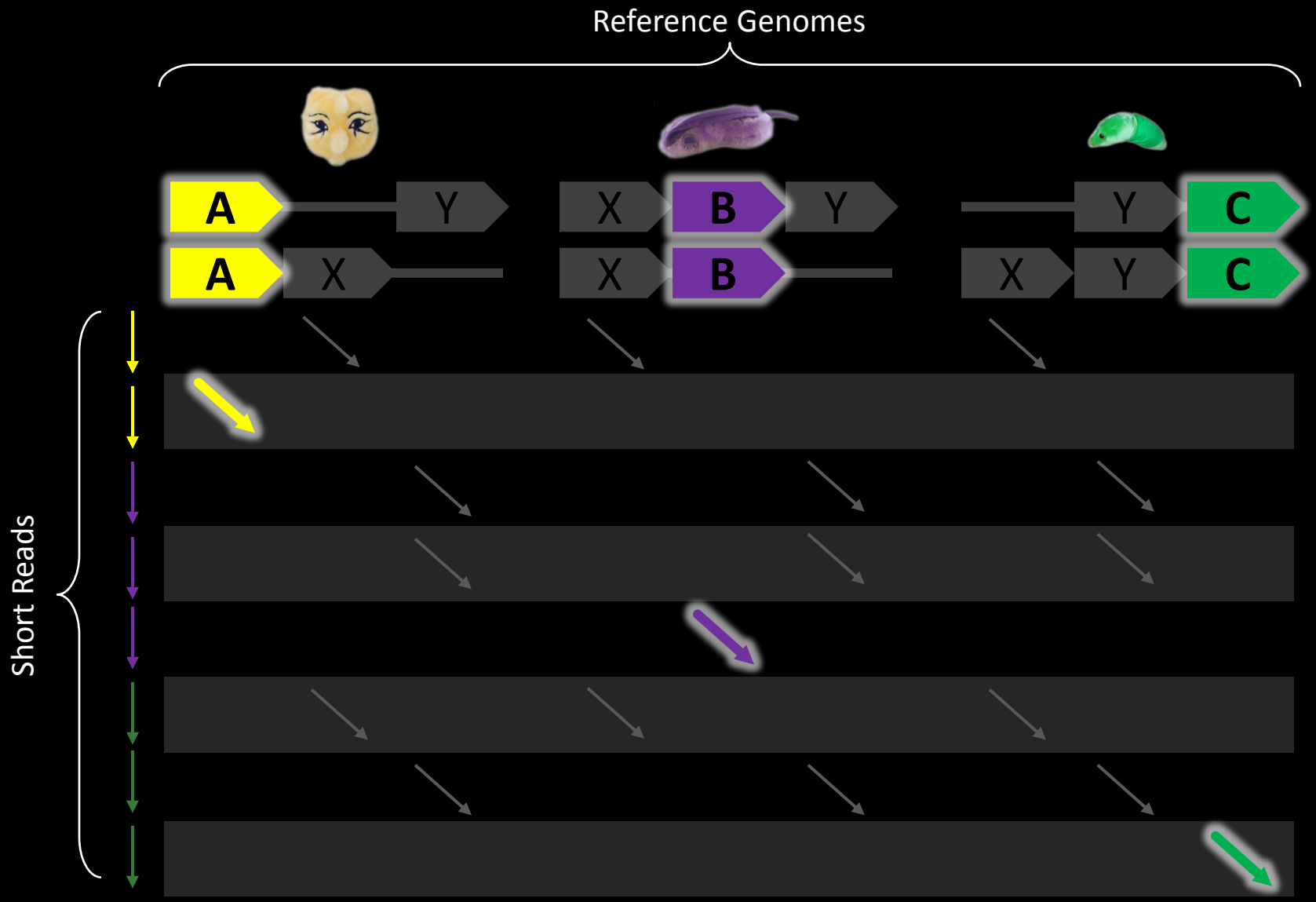
Validated against synthetic metagenome samples
(similar to MetaPhlAn validation)

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

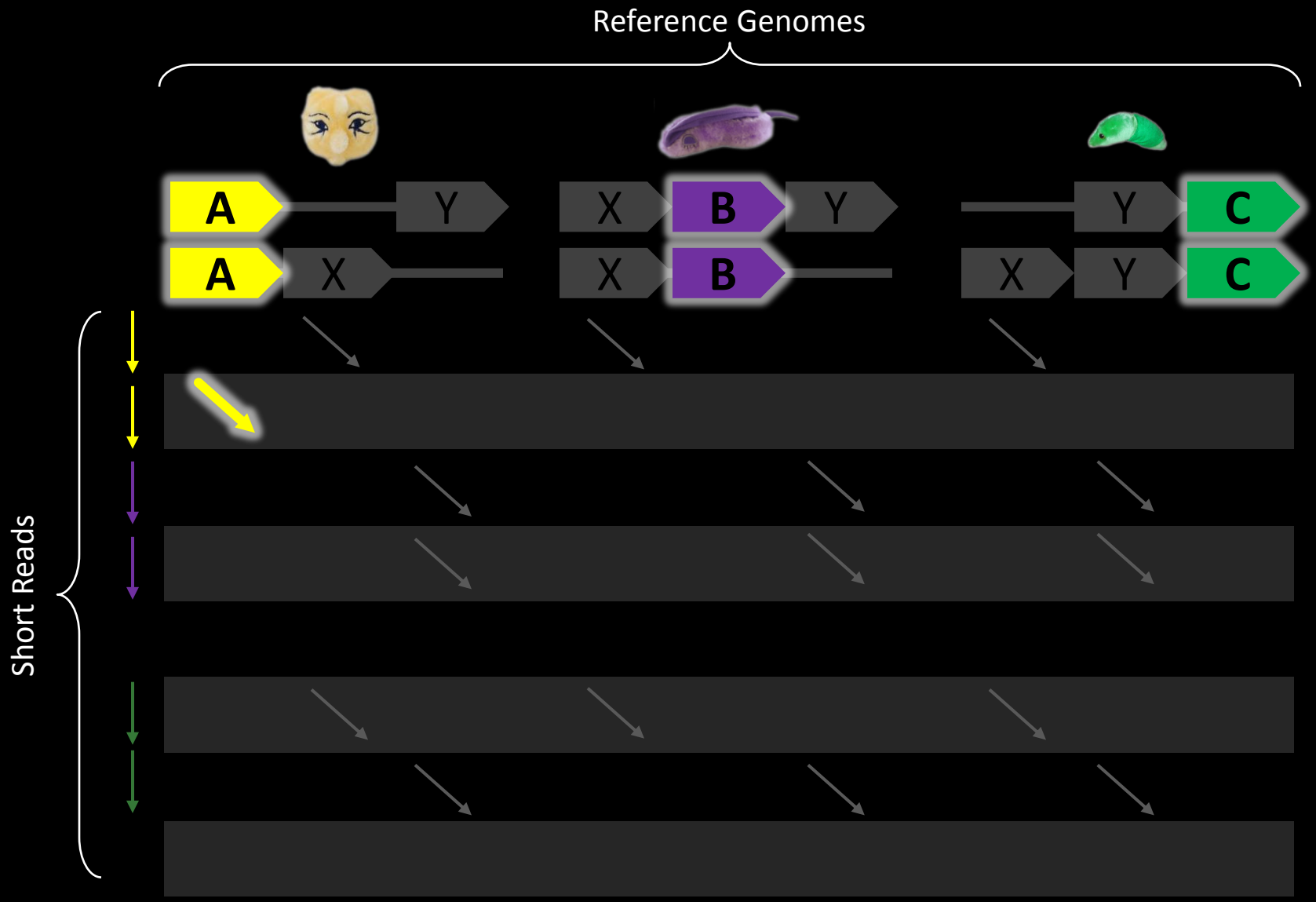
HUMAnN 2.0

- Avoid translated search where possible
- Speed up translated search with ORF-picking
- Stratify community-wide function by organism
- Focus on open gene family & pathway systems
- https://bitbucket.org/biobakery/humann2

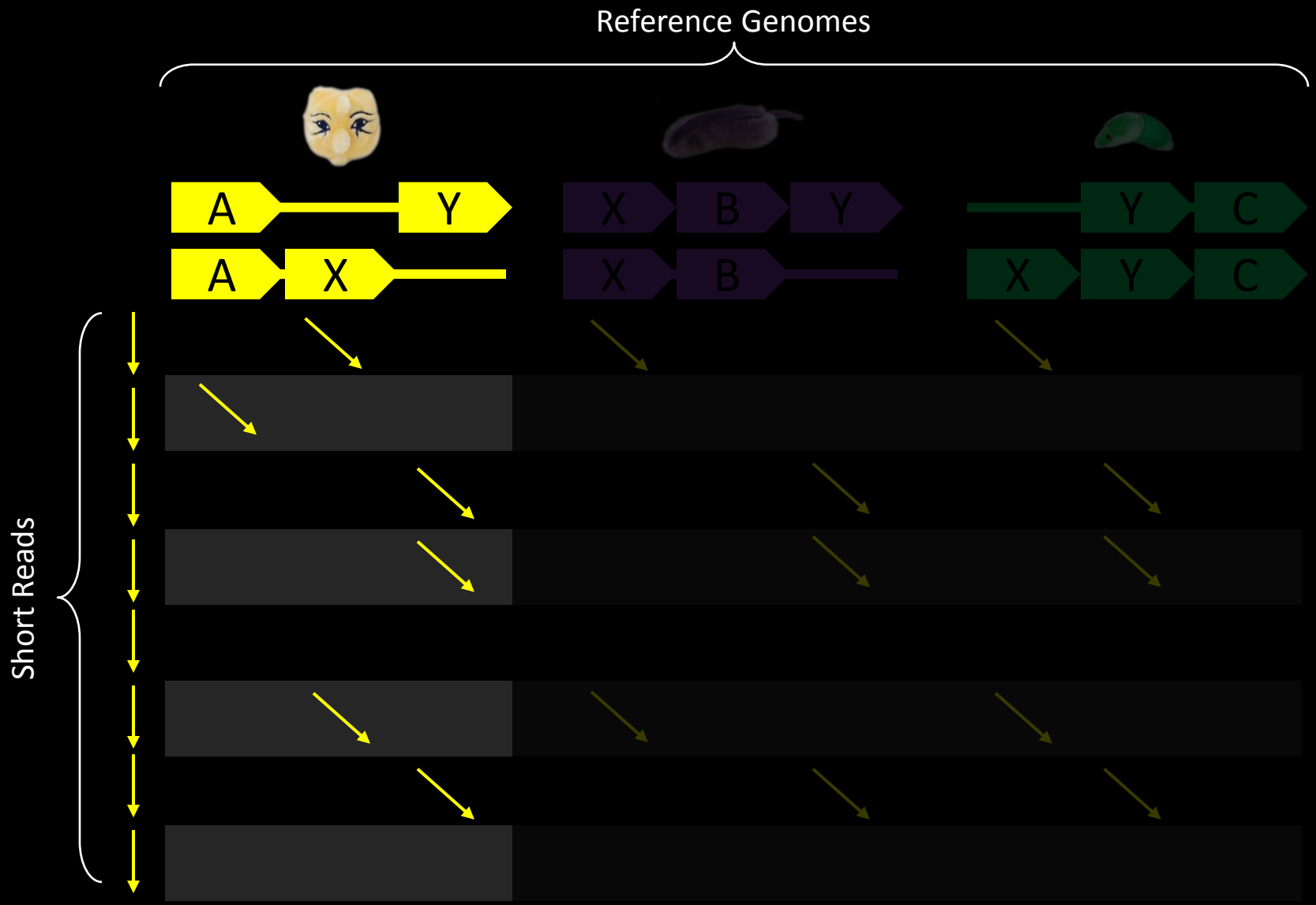
Faster functional profiling by avoiding translated search



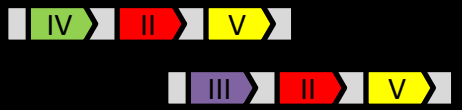
Faster functional profiling by avoiding translated search

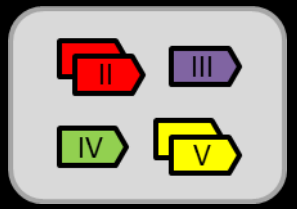


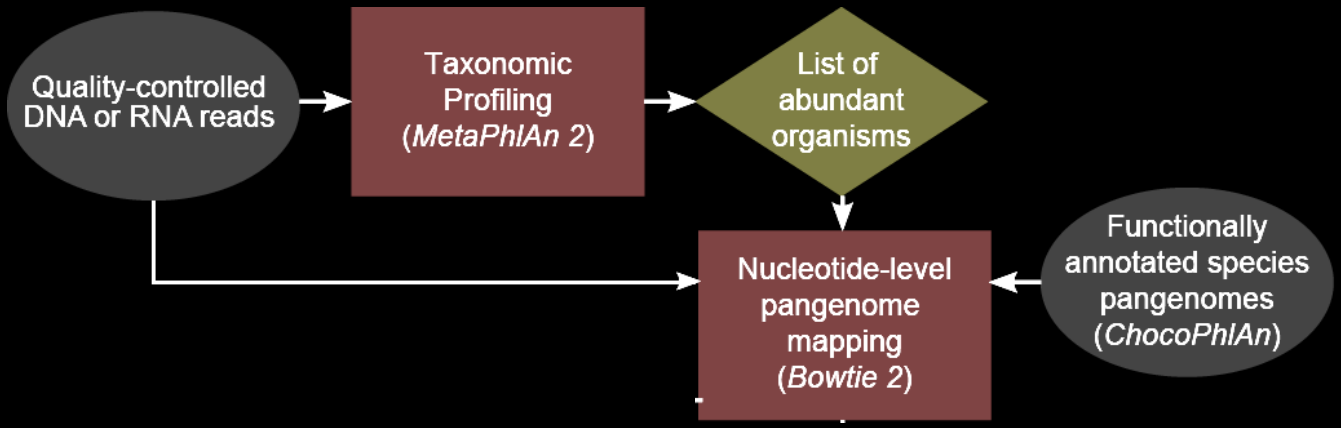
Faster functional profiling by avoiding translated search

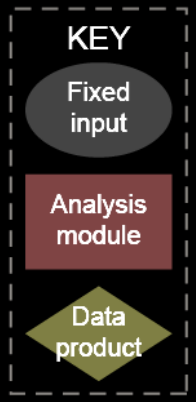


HUMAnN 2.0 overview



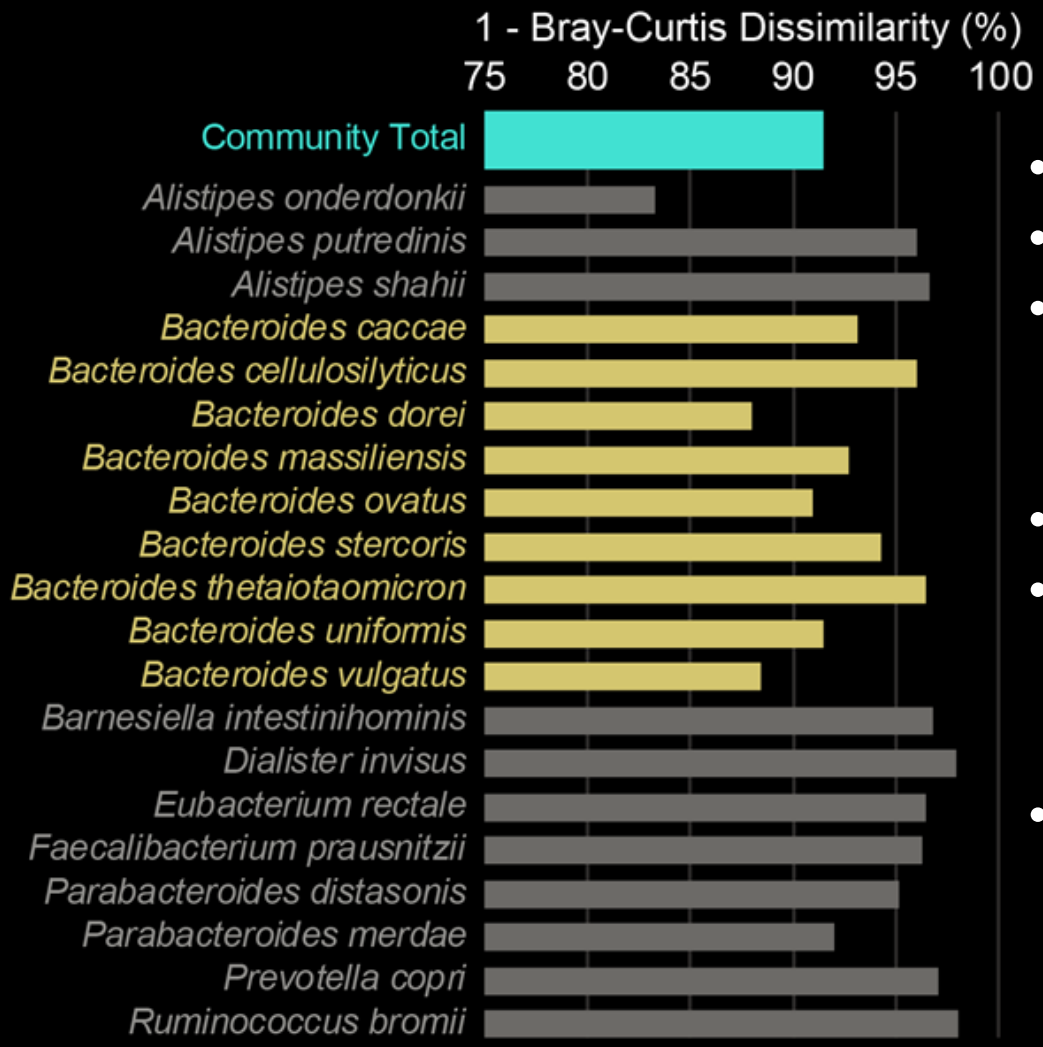






https://huttenhower.sph.harvard.edu/humann2

HUMAnN2 **accuracy** (1M read mock stool metagenome)



- 20 common gut bugs (even)
- 1M 100-nt reads
- Computed expected UniRef50 abundances from genome annotations
- Ran reads through HUMAnN2
- Compared expected and observed profiles

- Strong agreement, even for closely related species (e.g. *Bacteroides*)

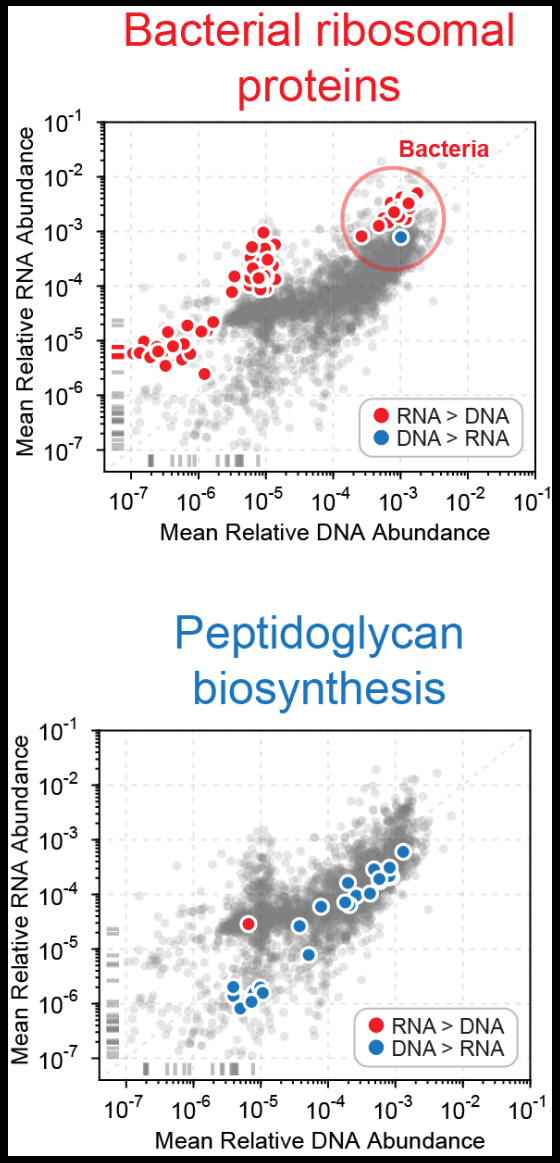
https://huttenhower.sph.harvard.edu/humann2

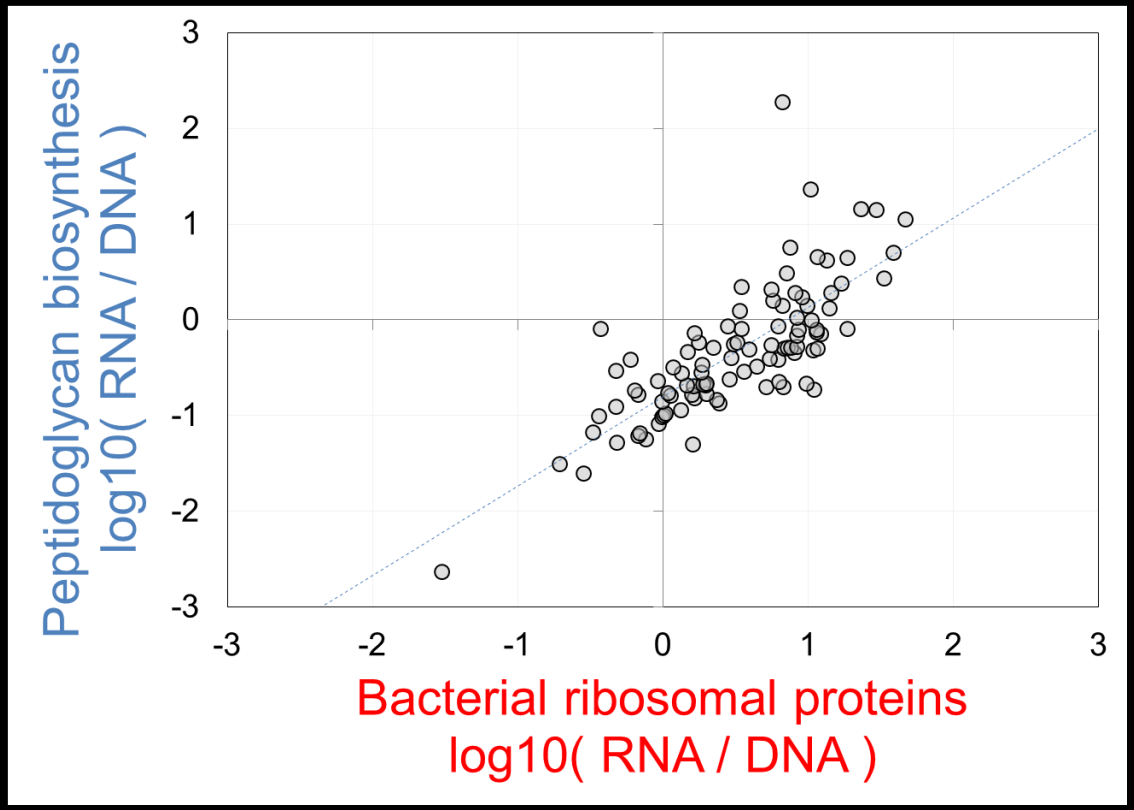
HUMAnN2 **performance** (1M read mock stool metagenome)

	MetaPhlAn2 Prescreen	Pangenome Search	Translated Search	Total
Normal Flow	0.5 cpu-hours	**0.7** cpu-hours (**86%** reads)	**1.7** cpu-hours (**14%** reads)	2.9 cpu-hours
Translated Search Only	NA	NA	12.1 cpu-hours	12.1 cpu-hours

https://huttenhower.sph.harvard.edu/humann2

HUMAnN2: Example combining DNA- & RNA-seq from 8 healthy gut microbiomes

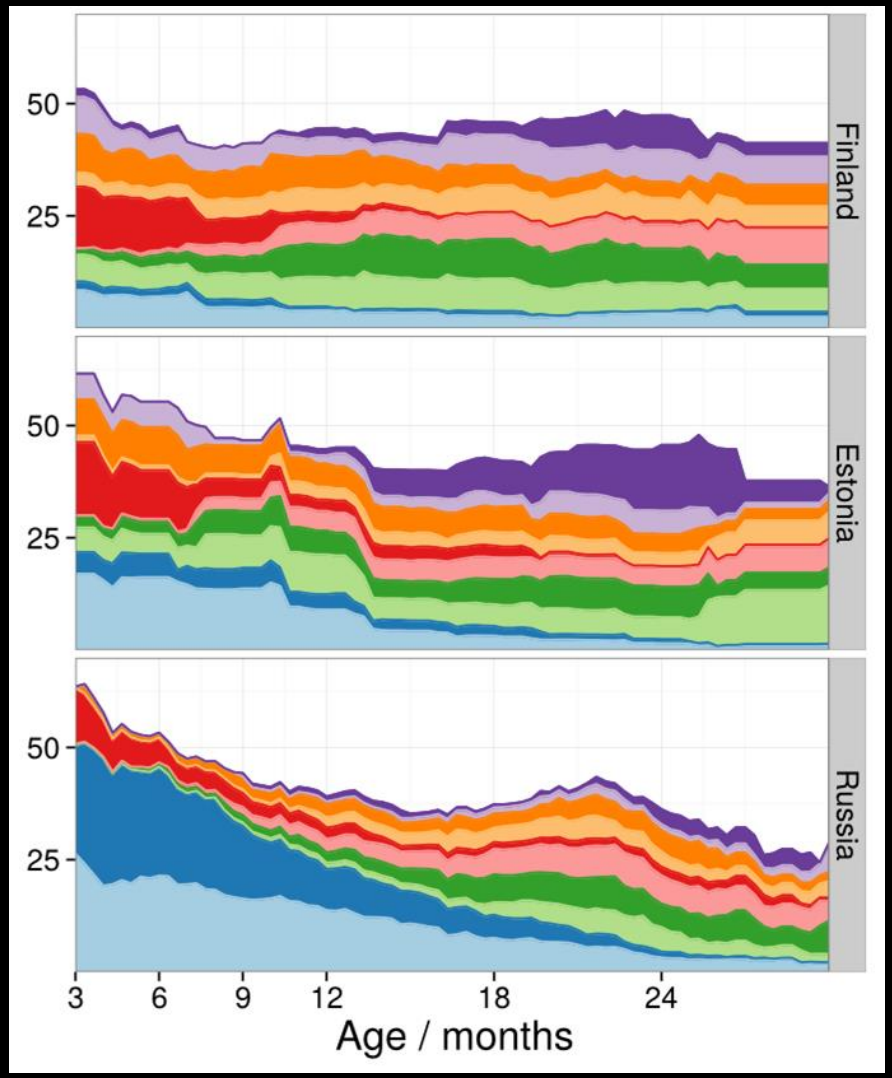


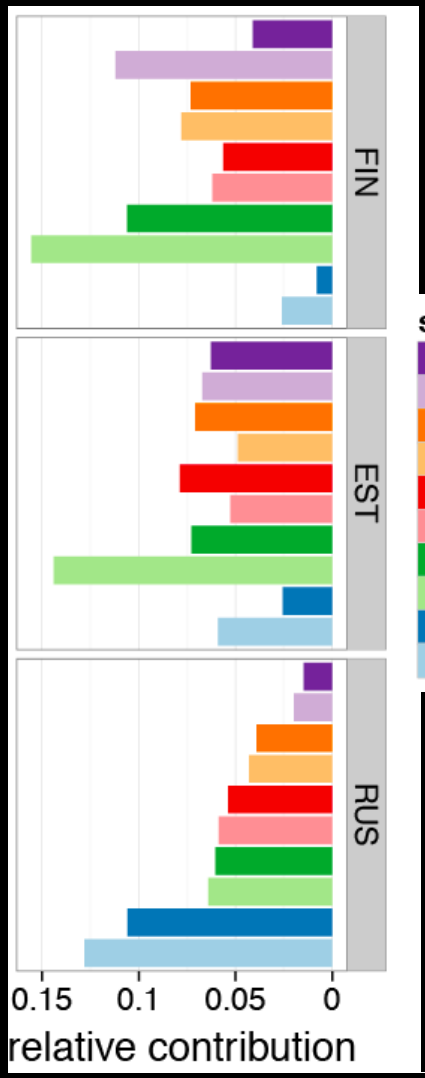


- Dot = functional contribution of one species
- Ribosomal & peptidoglycan transcription correlate
- Ribosome biosyn. generally “over-transcribed”
- Peptidoglycan biosyn. generally “under-transcribed”
- Not a paradox, it’s consistent with the biology

https://huttenhower.sph.harvard.edu/humann2

HUMAnN2: Glycolytic processes performed by different species in Finns and Russians





Tommi Vatanen

species

- *Prevotella copri*
- *Bacteroides dorei*
- *Bacteroides fragilis*
- *Bacteroides ovatus*
- *Escherichia coli*
- *Faecalibacterium prausnitzii*
- *Bacteroides uniformis*
- *Bacteroides vulgatus*
- *Bifidobacterium bifidum*
- *Bifidobacterium longum*

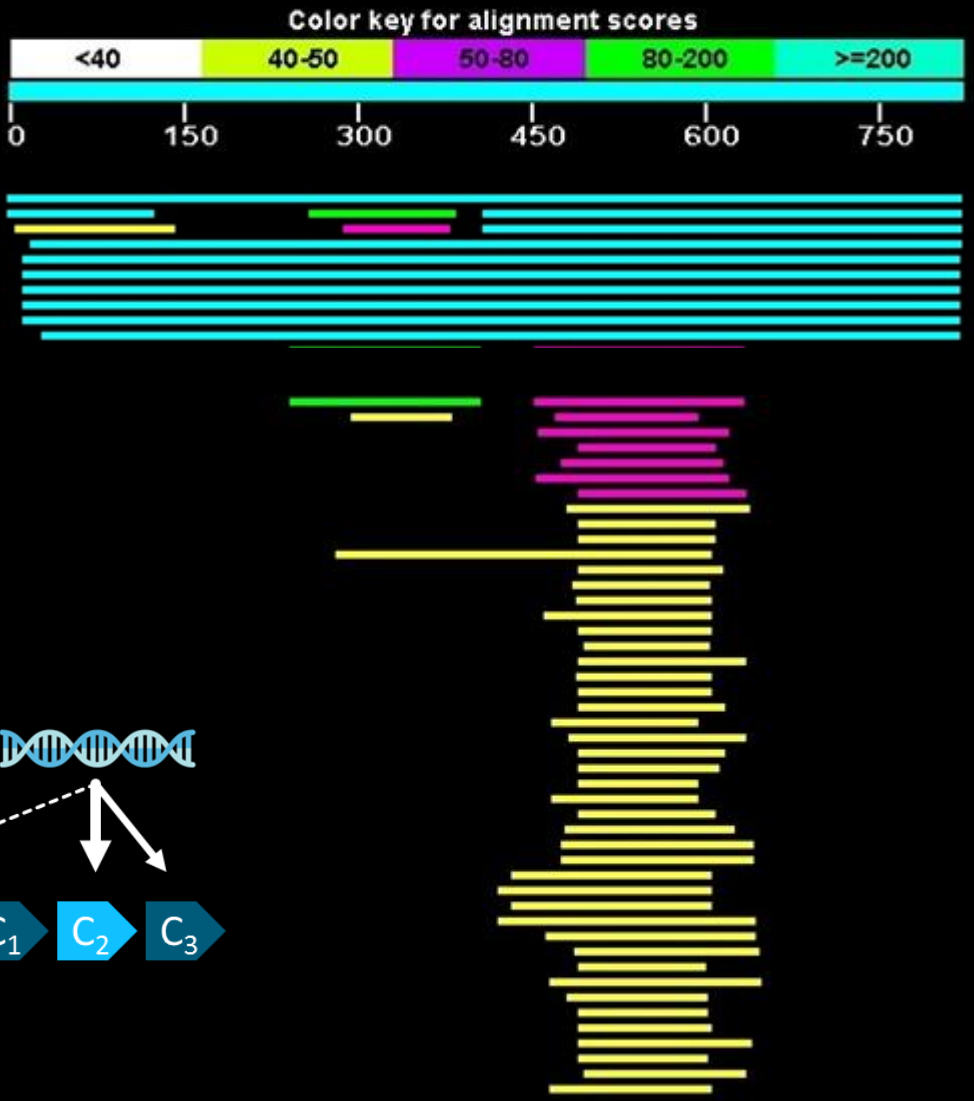
https://huttenhower.sph.harvard.edu/humann2

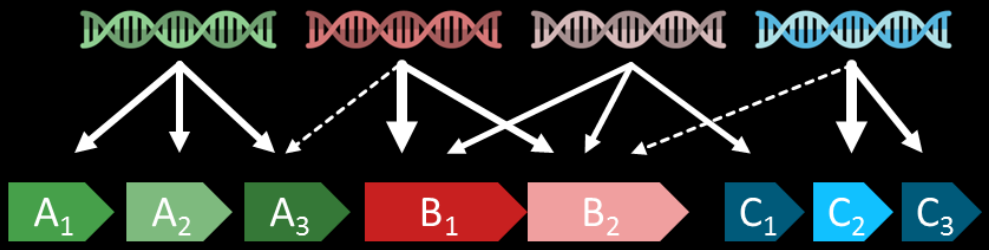
ShortBRED

For targeted meta'omic functional profiling

The problem with short reads and regions of local homology among proteins

- Protein of interest
- Belongs to a family
- Local homology to unrelated families
- Short reads from unrelated families may map to protein of interest (spurious hits)

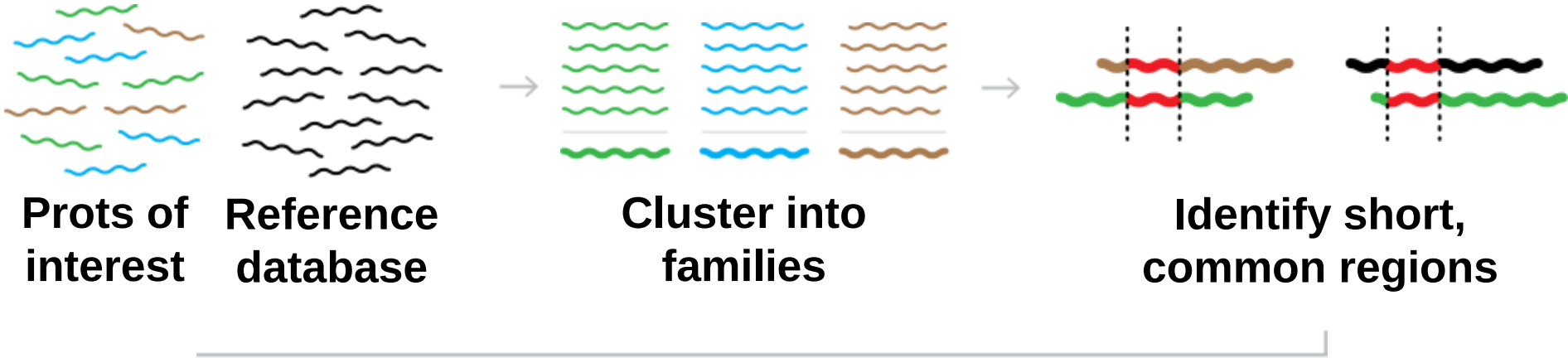




https://huttenhower.sph.harvard.edu/shortbred

ShortBRED Identify
Find unique markers for interesting prots

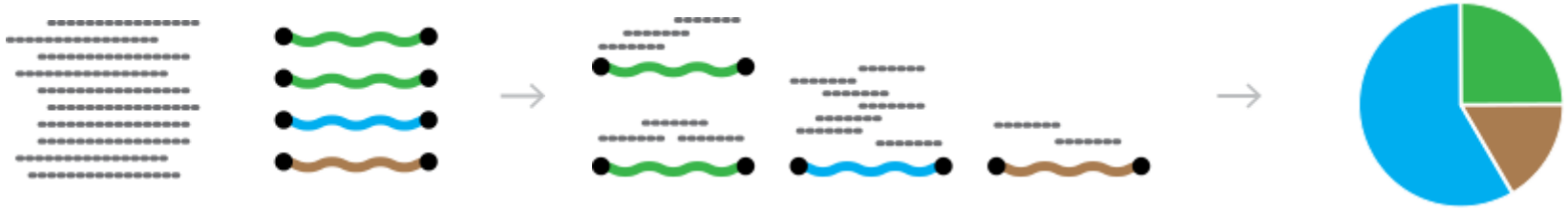
Jim Kaminski



https://huttenhower.sph.harvard.edu/shortbred

ShortBRED Quantify
Use markers for highly specific profiling

Jim Kaminski



Metagenome reads

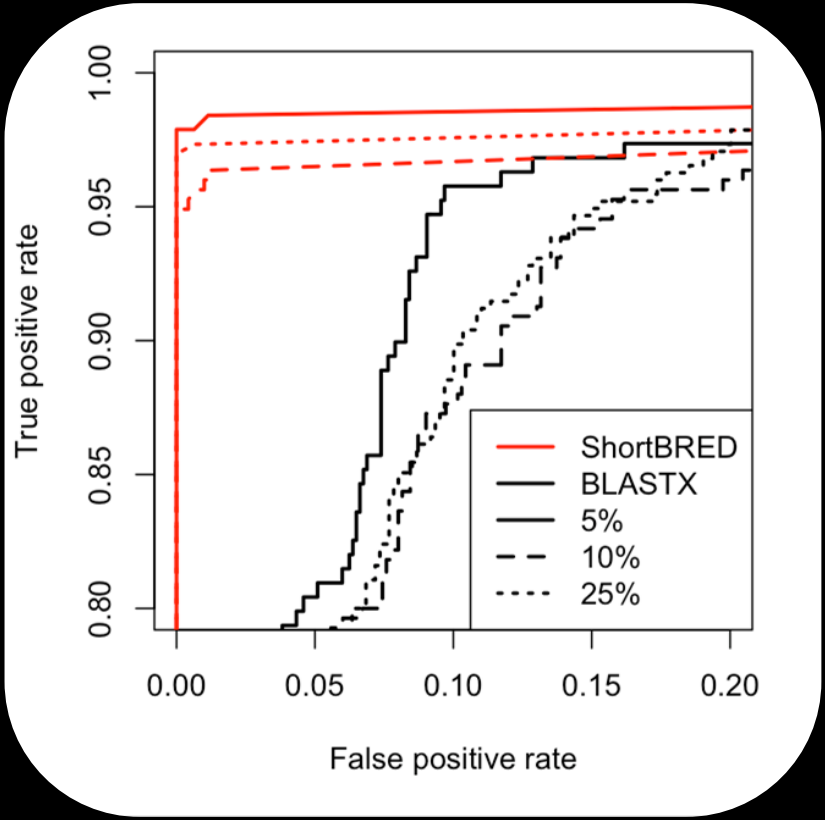
ShortBRED markers

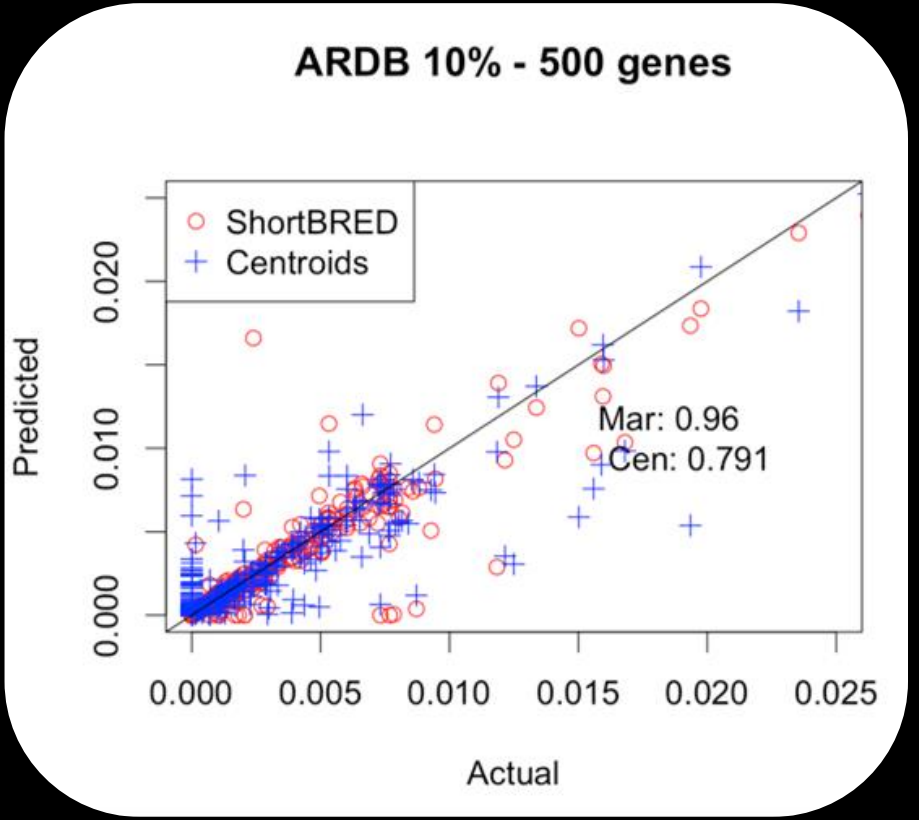
Translated search for high ID hits

Normalize relative abundances

https://huttenhower.sph.harvard.edu/shortbred

ShortBRED Synthetic Evaluation (ABR genes)





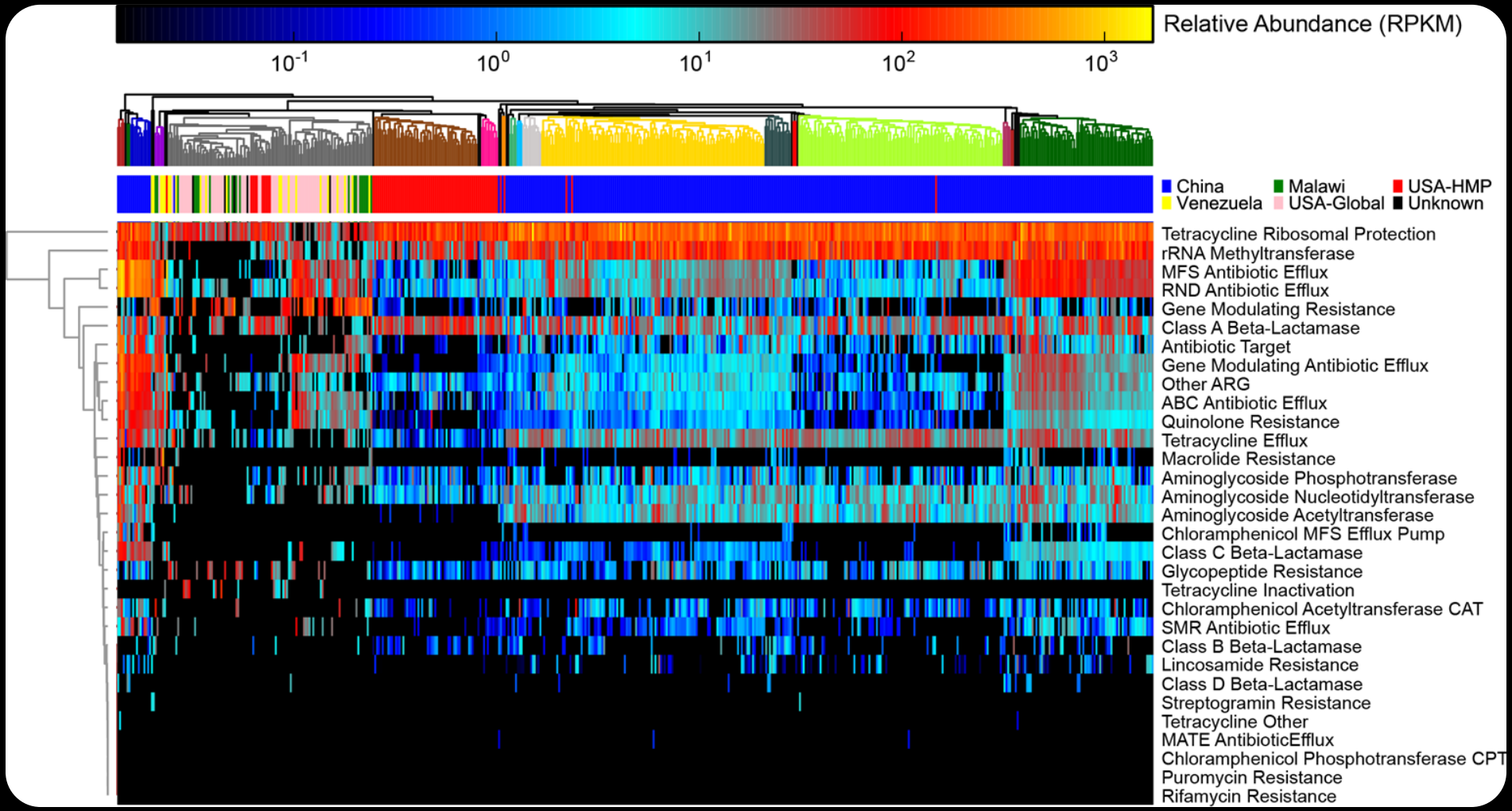
Relative to mapping reads against full-length centroids, we are:

> Substantially more accurate (fewer false positives)

> Faster (reduced search space)

https://huttenhower.sph.harvard.edu/shortbred

ShortBRED: ABR in human gut metagenomes



https://huttenhower.sph.harvard.edu/shortbred

PICRUSt

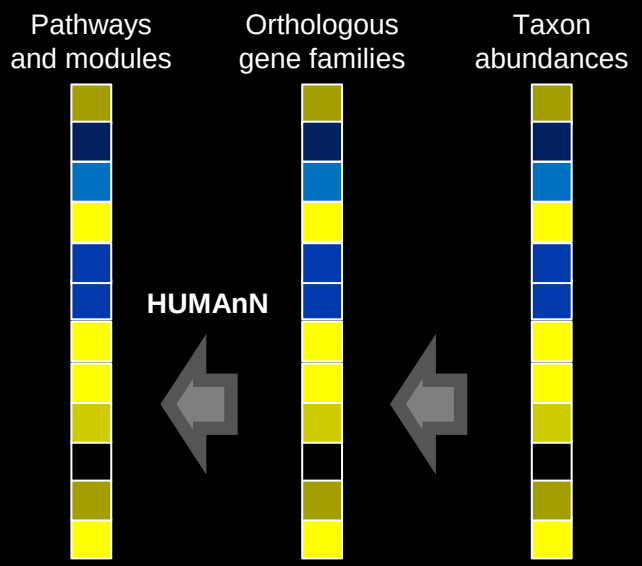
For predictive functional profiling

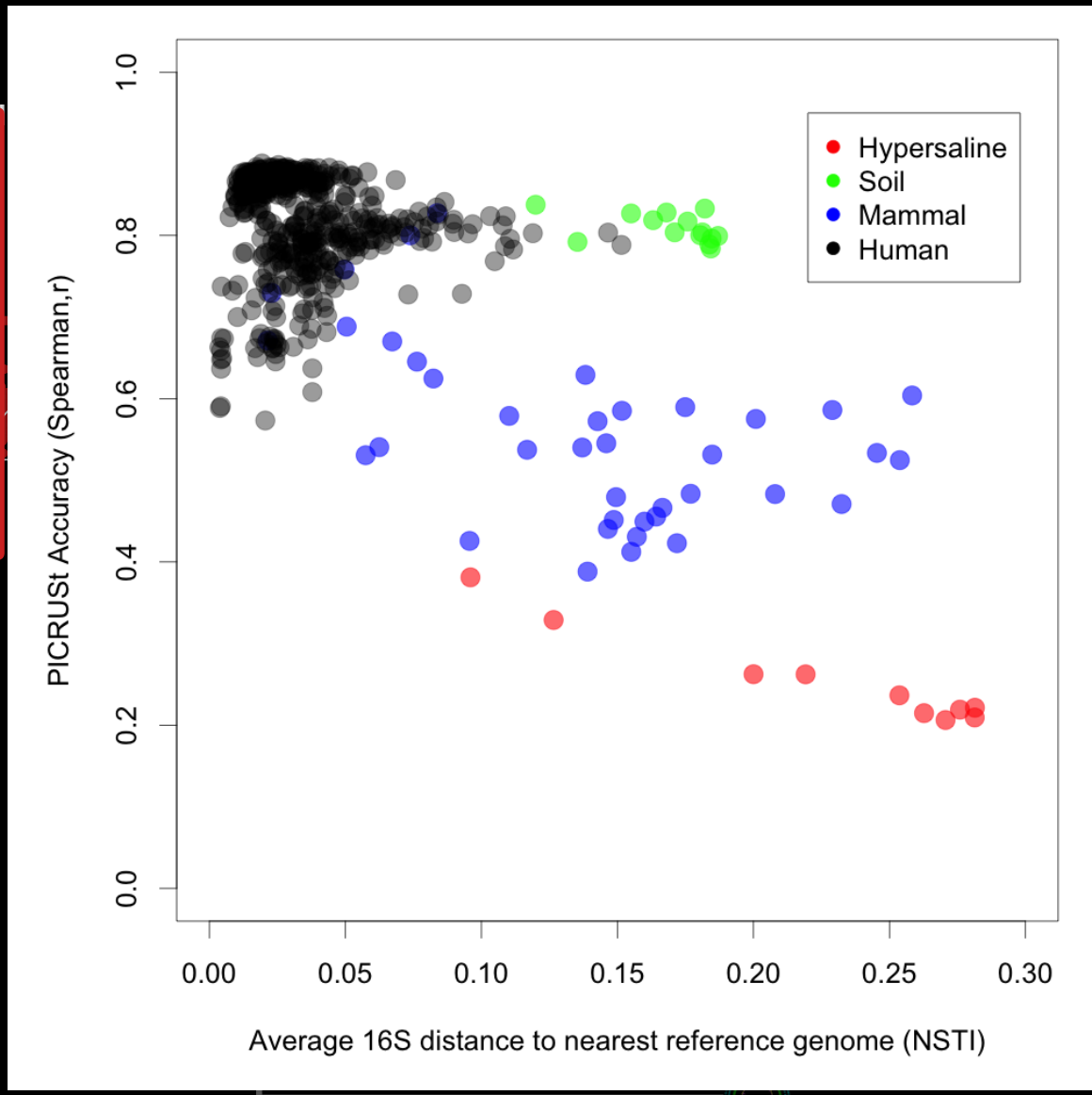
PICRUSt: Inferring community metagenomic potential from marker gene sequencing

With Rob Knight, Rob Beiko

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

http://picrust.github.com





Relative abundance

Meta'omics seeks to answer 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

Meta'omics seeks to answer ~~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

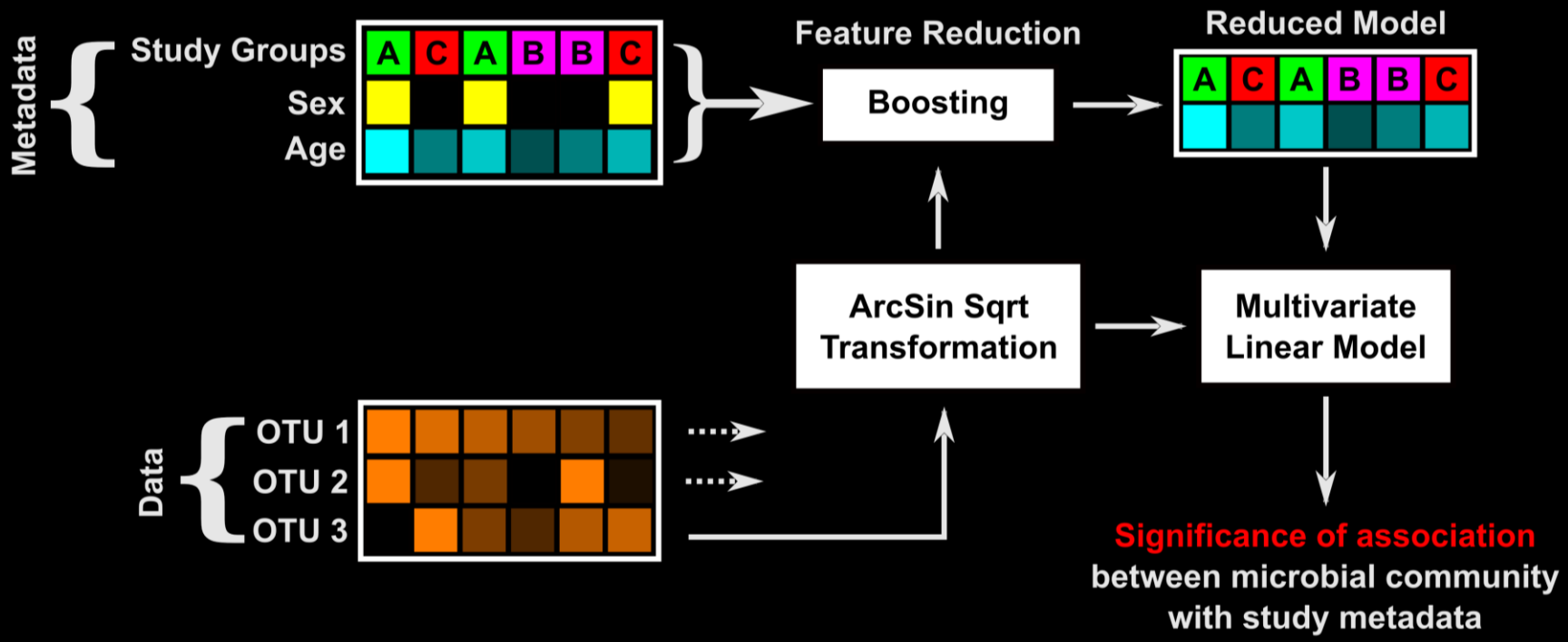
Site	Oral	Gut	Oral	Gut	Oral	Gut
Clade1	0.40	0.87	0.43	0.68	0.47	0.32
Clade1|Bug1	0.40	0.56	0.07	0.31	0.42	0.27
Clade1|Bug2	0.00	0.30	0.36	0.37	0.04	0.05
Clade2	0.60	0.13	0.57	0.32	0.53	0.68
Clade2|Bug3	0.11	0.00	0.10	0.32	0.15	0.23
Clade2|Bug4	0.49	0.13	0.47	0.00	0.39	0.45

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

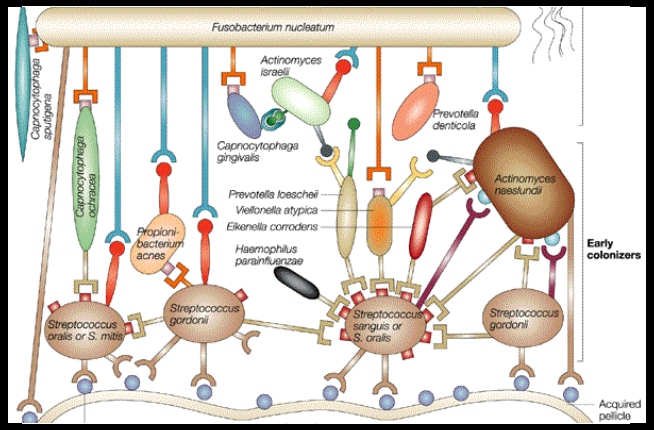
MaAsLin
Multivariate Association with Linear Models



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

Tim Tickle

Microbiome downstream analyses: interaction network reconstruction



It's a jungle in there – microbial interactions follow patterns from classical macro-ecology.

Mutualism

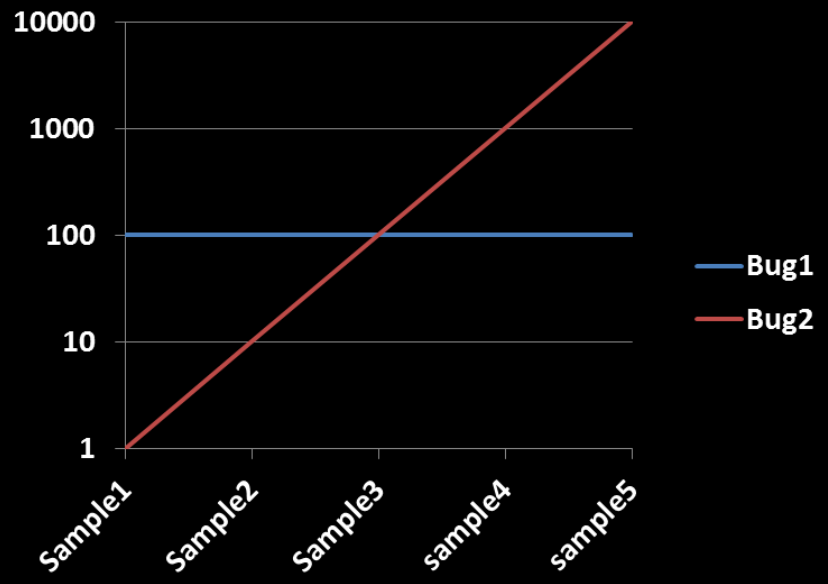
Predation

Competition

Given microbial relative abundance measurements over many samples, can we detect *co-occurrence and co-exclusion relationships?*

Relative abundance data poses a problem for correlating metagenomic features

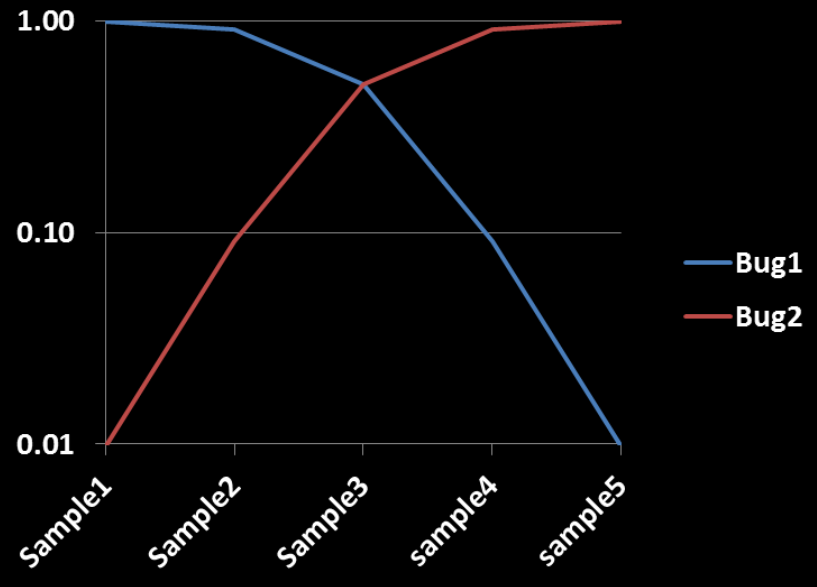
	Sample1	Sample2	Sample3	sample4	sample5
Bug1	100	100	100	100	100
Bug2	1	10	100	1000	10000



Absolute (cell) counts

No bug1-bug2 correlation

	Sample1	Sample2	Sample3	sample4	sample5
Bug1	0.99	0.91	0.50	0.09	0.01
Bug2	0.01	0.09	0.50	0.91	0.99

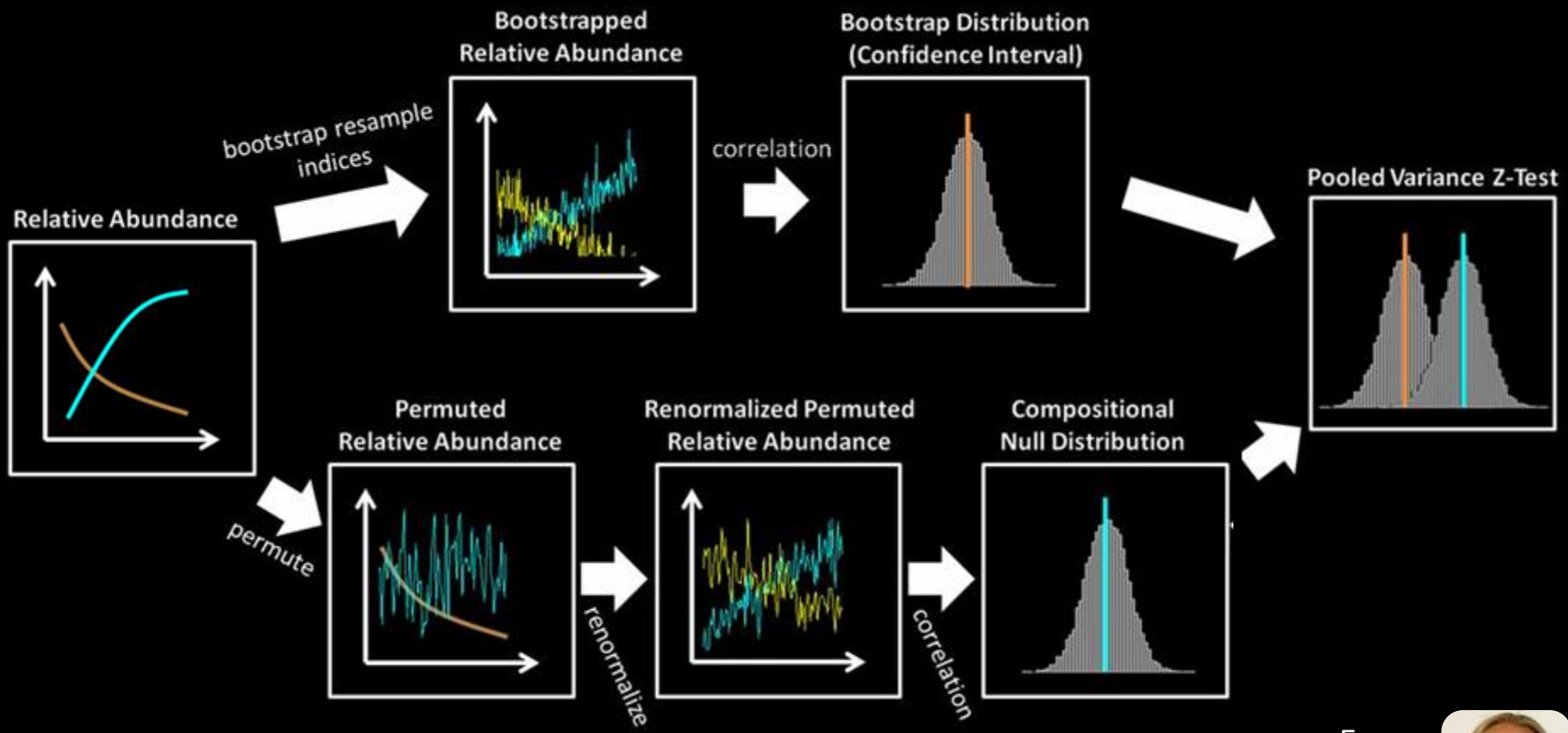


Relative abundance

Spurious bug1-bug2 correlation

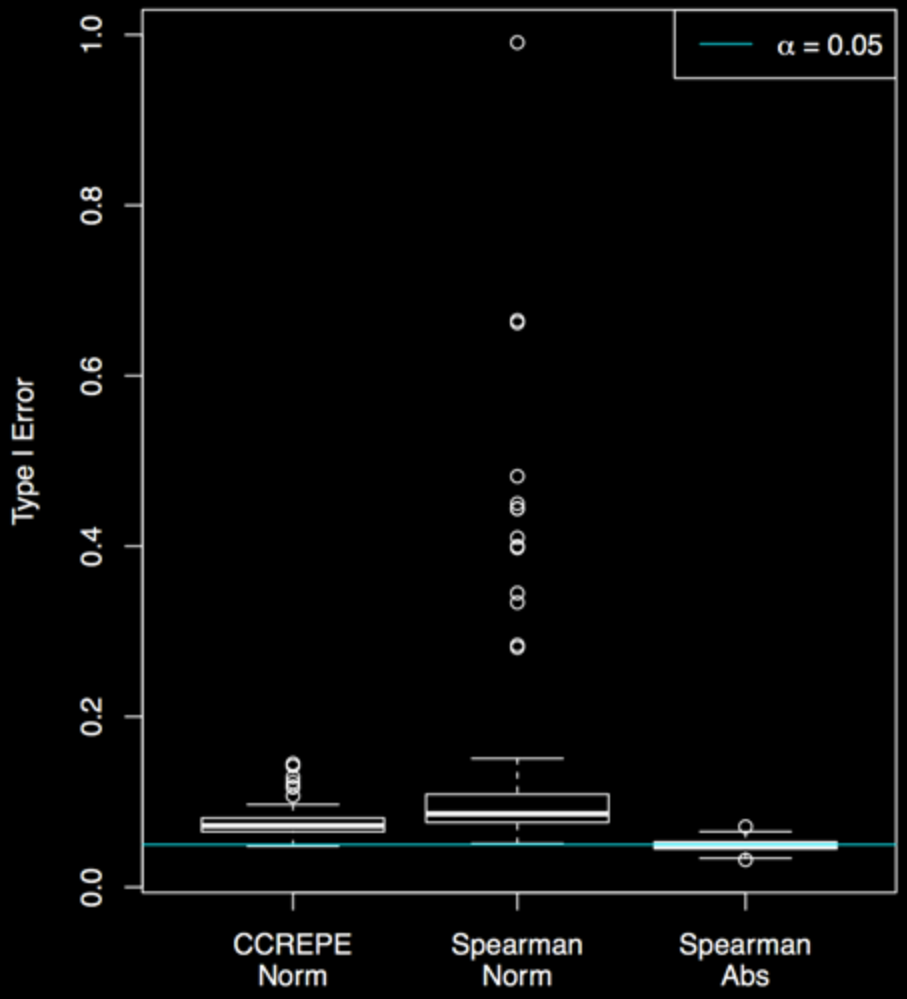
(sequencing yields rel. ab.)

CCREPE: Compositionality Corrected by REnormalization and PErmutation



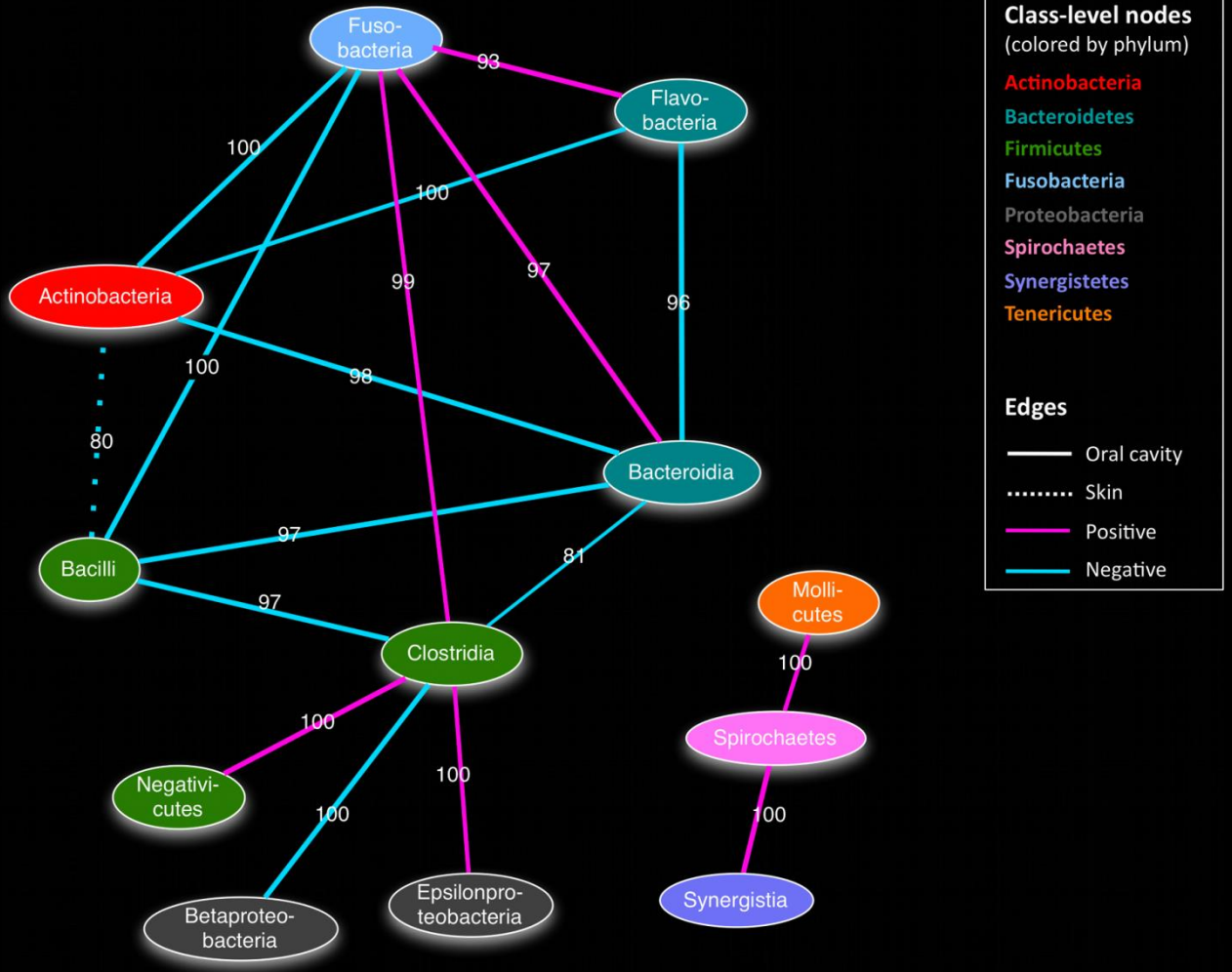
Emma Schwager

CCREPE: Compositionality Corrected by REnormalization and PErmutation



- Synthetic evaluation
- Random sample feature/tables
- No built-in correlation structure

CCREPE: Compositionality Corrected by REnormalization and PErmutation



"Microbial co-occurrence relationships in the human microbiome."
Faust, et al. *PLoS Comp Biol*, 8:e1002606 (2012).

The three big questions…

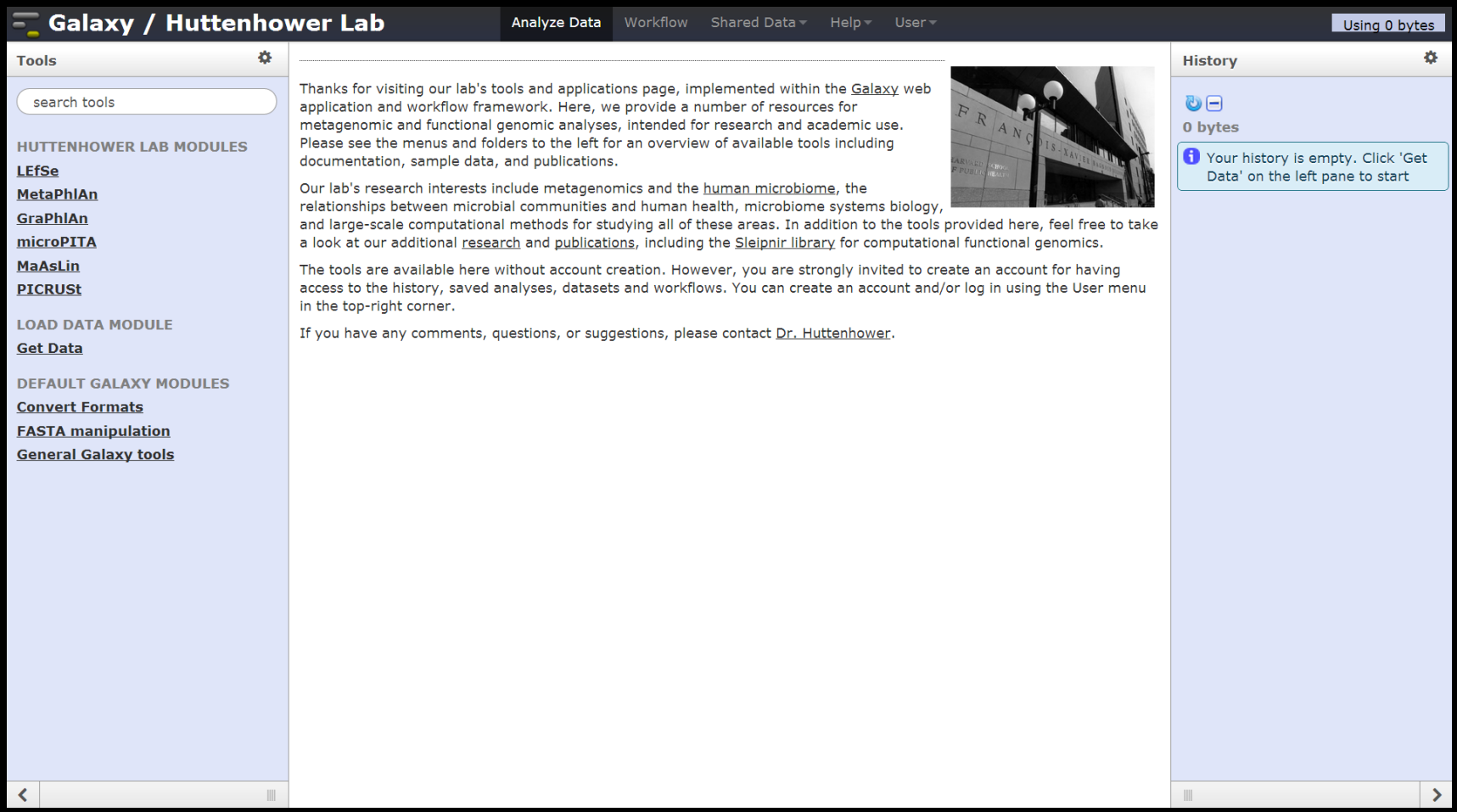
Who is there?

What are they doing?

What does it all mean?

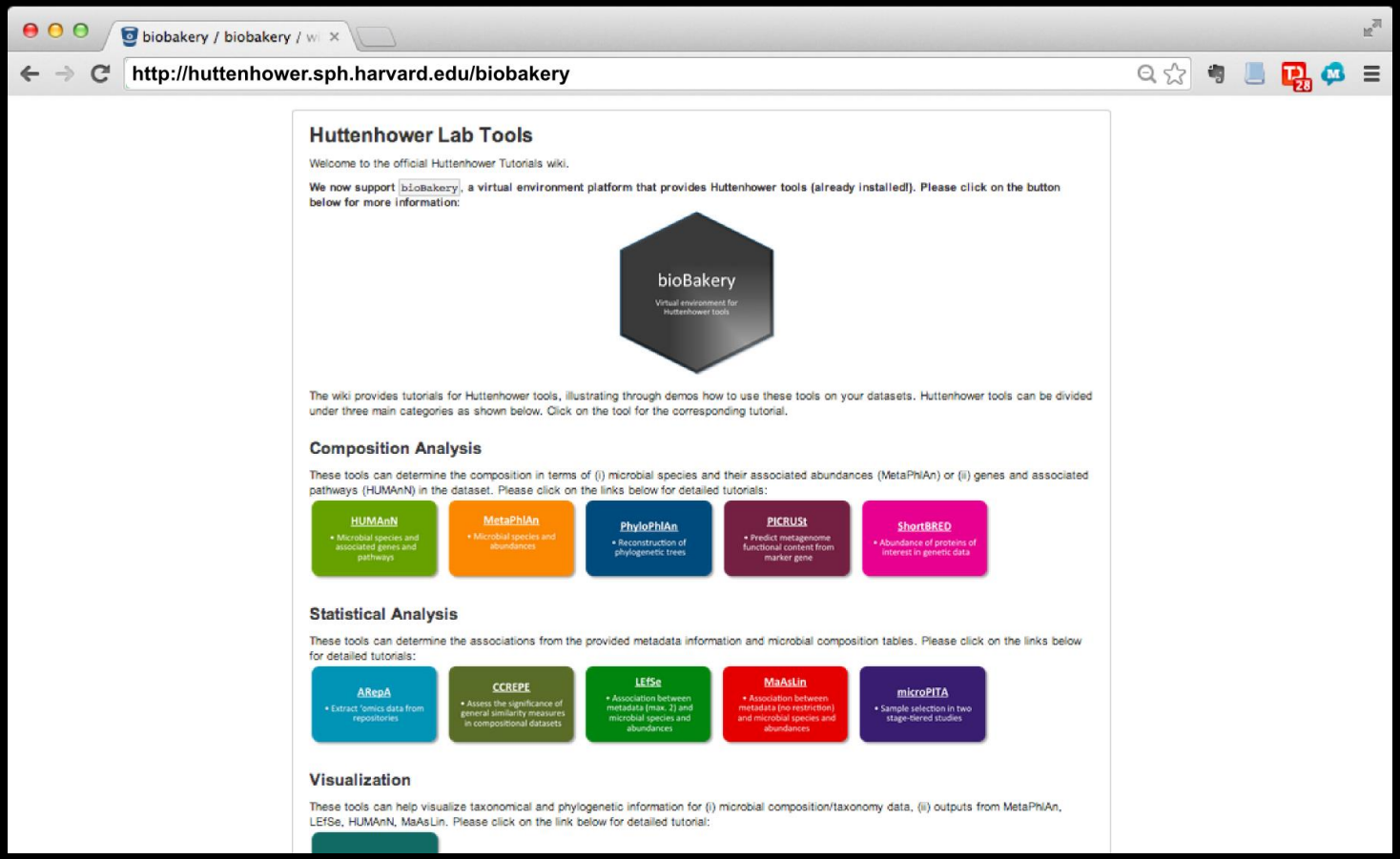
Resources

Using tools through Galaxy



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

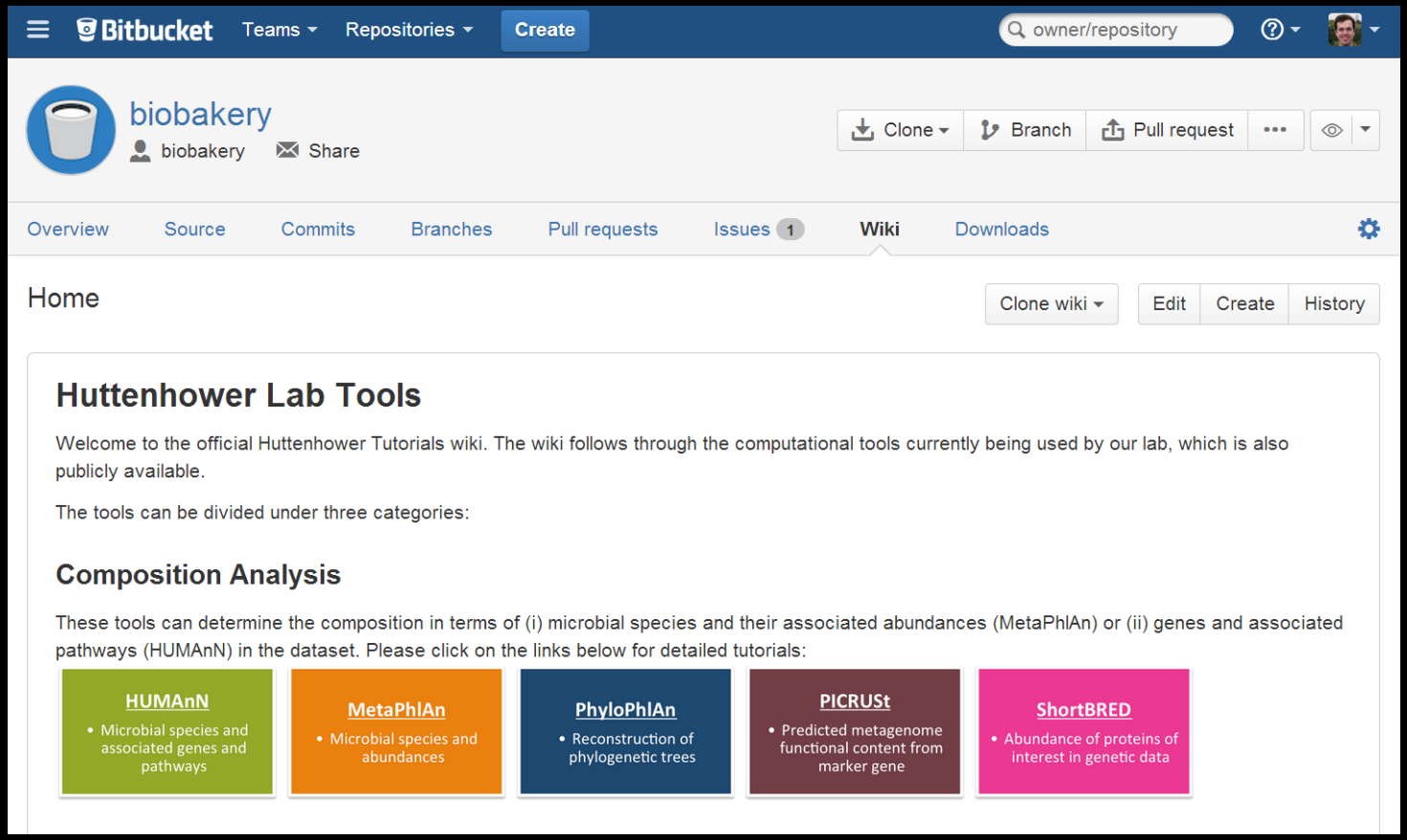
Tutorials available online



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

(click on your tool-of-interest)

All tools are open source



http://bitbucket.org/biobakery/biobakery

The bioBakery Virtual Machine

https://bitbucket.org/biobakery/biobakery/wiki/biobakery_wiki

Ubuntu base image preloaded and configured to run all Huttenhower lab tools; one click up-and-running via Vagrant

Thank you!


Curtis Huttenhower


Xochitl Morgan


Minah Iqbal


Lauren McIver


George Weingart


Aleksandar Kostic


Gholamali Rahnavard


Melanie Schirmer


Alexandra Sirota-Madi


Ayshwarya Subramanian


Tiffany Hsu


Boyu Ren


Emma Schwager

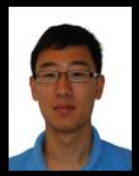

Koji Yasuda


Siyuan Ma


Casey DuLong


Jim Kaminski


Randall Schwager


Andy Shi


Moran Yassour


Luc Bijnens


Tommi Vatanen


Levi Waldron


Nicola Segata

Wendy Garrett
Michelle Rooks

Ruth Ley
Omry Koren

Dirk Gevers
Kat Huang

Ramnik Xavier
Harry Sokol
Dan Knights
Moran Yassour

Rob Beiko
Morgan Langille

Rob Knight
Greg Caporaso
Jesse Zaneveld

Mark Silverberg
Boyko Kabakchiev
Andrea Tyler

Jacques Izard

Katherine Lemon

Bruce Sands

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 Mark Daly
Doyle Ward Ashlee Earl