Introduction to metagenomic analysis

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

Harvard University CFAR Workshop on Metagenomics and Transcriptomics

16 September 2014

Huttenhower Research Group
Harvard School of Public Health
Department of Biostatistics

Content

- http://huttenhower.sph.harvard.edu/content/cfar2014

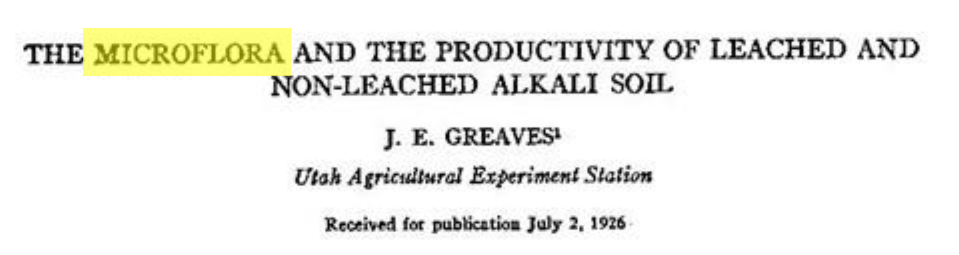
Plan

- Informal survey
- Metagenomics concepts & examples
- Tools for taxonomic profiling
  - MetaPhlAn
- Tools for functional profiling
  - HUMAnN
  - ShortBRED
  - PICRUSt
- Tools for testing associations
  - LEfSe
  - MaAsLin
  - CCREPE
- Resources
- Research vignette (time permitting)

What’s metagenomics?

Total collection of **microorganisms** within a **community**

Also **microbial community** or **microbiota**

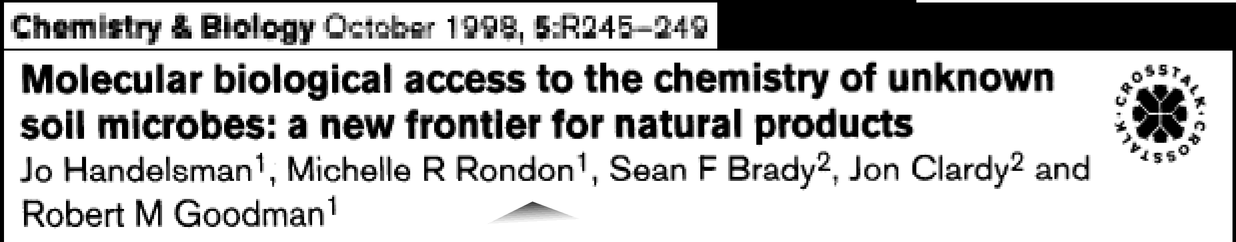


THE MICROFLORA AND THE PRODUCTIVITY OF LEACHED AND NON-LEACHED ALKALI SOIL

J. E. GREAVES[1]

Utah Agricultural Experiment Station

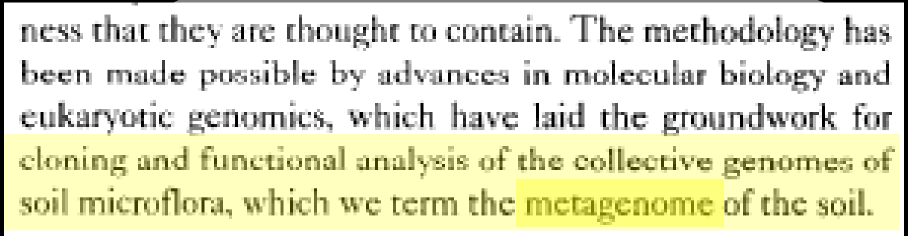
Received for publication July 2, 1926



Chemistry & Biology October 1998, 5:R245–249

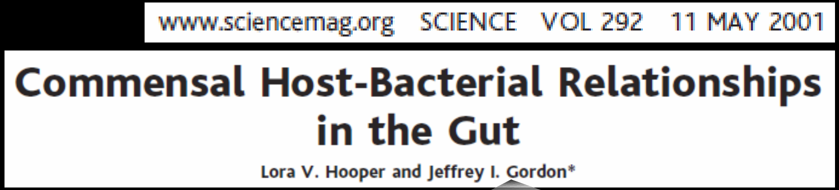
Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products

Jo Handelsman[1], Michelle R Rondon[1], Sean F Brady[2], Jon Clardy[2] and Robert M Goodman[1]



ness that they are thought to contain. The methodology has been made possible by advances in molecular biology and eukaryotic genomics, which have laid the groundwork for cloning and functional analysis of the collective genomes of soil microflora, which we term the metagenome of the soil.

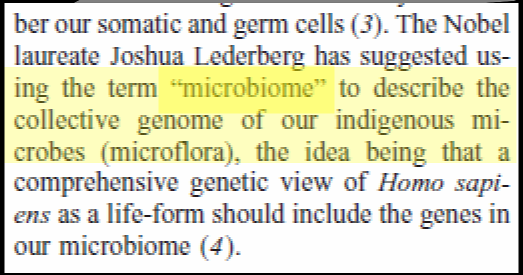
Total **genomic potential** of a microbial community



www.sciencemag.org SCIENCE VOL 292 11 MAY 2001

Commensal Host-Bacterial Relationships in the Gut

Lora V. Hooper and Jeffrey I. Gordon*



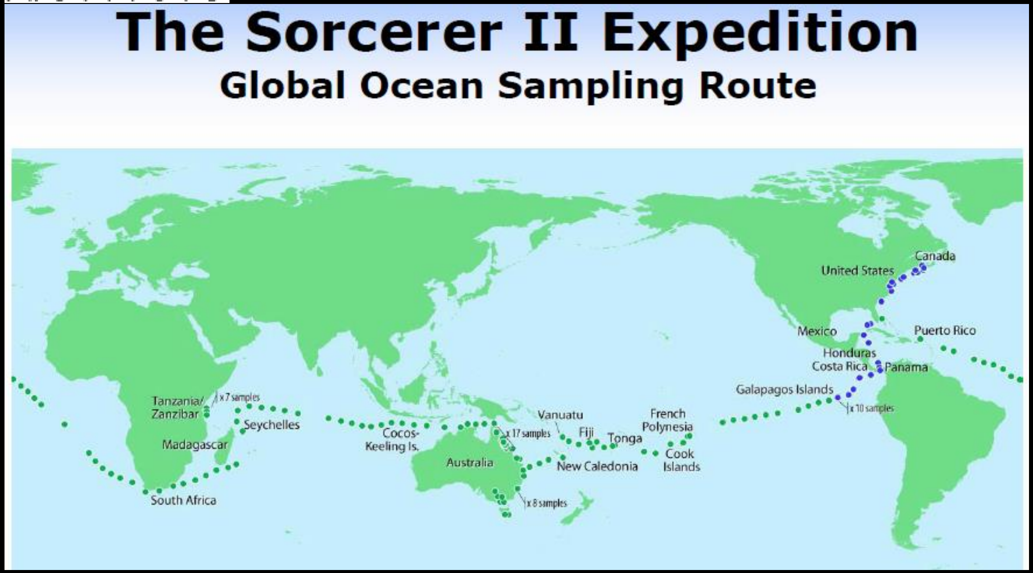
ber our somatic and germ cells (*3*). The Nobel laureate Joshua Lederberg has suggested using the term “microbiome” to describe the collective genome of our indigenous microbes (microflora), the idea being that a comprehensive genetic view of *Homo sapiens* as a life-form should include the genes in our microbiome (*4*).

Study of **uncultured microorganisms** from the environment, which can include humans or other living hosts

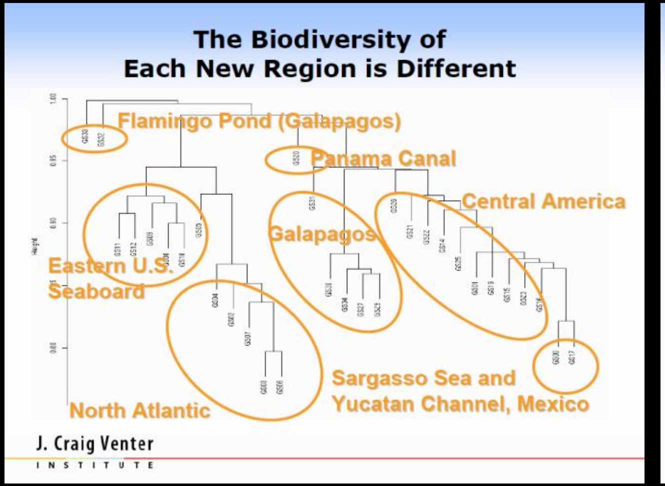
Total **biomolecular repertoire** of a microbial community

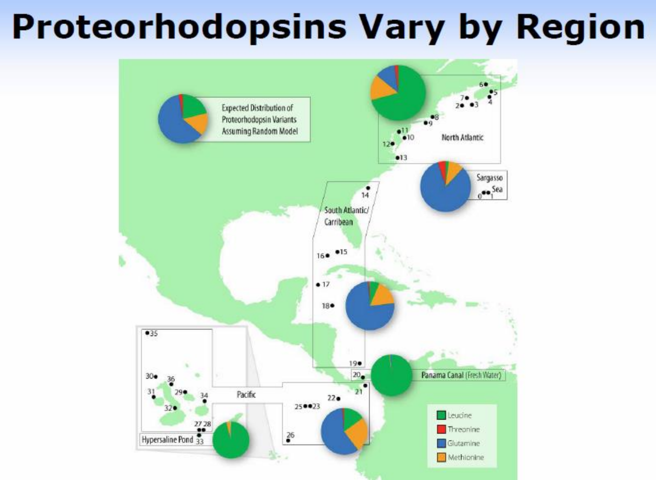
Examples of metagenomic studies: Global Ocean Sampling

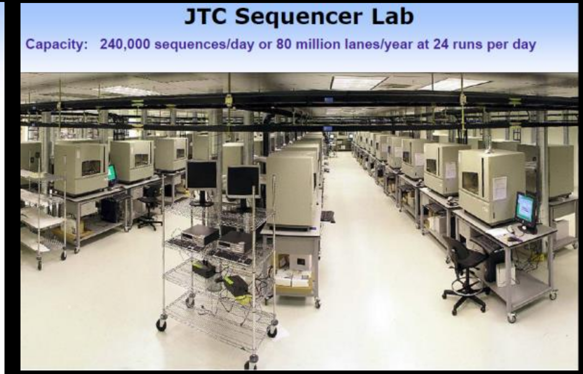
J. Craig Venter INSTITUTE 2003/2004 - ongoing


The Sorcerer II Expedition

Global Ocean Sampling Route


The Biodiversity of Each New Region is Different


Proteorhodopsins Vary by Region

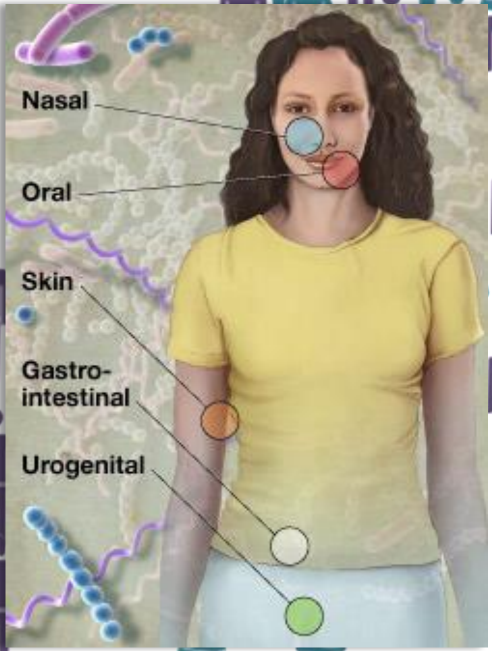

JTC Sequencer Lab

Capacity: 240,000 sequences/day or 80 million lanes/year at 24 runs per day

Slide by Dirk Gevers

The NIH **H**uman **M**icrobiome **P**roject (**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



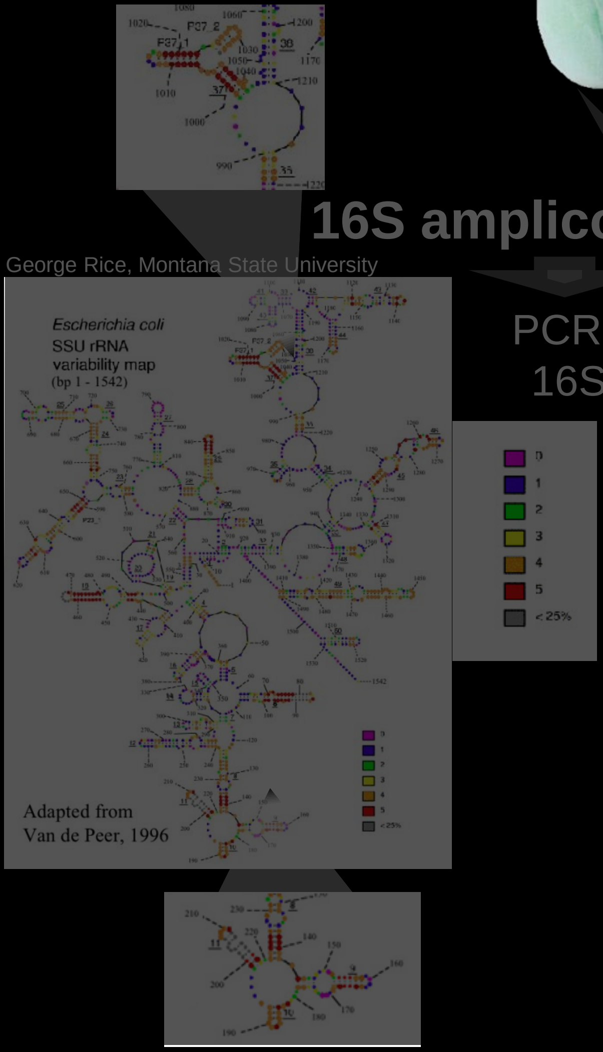
www.hmpdacc.org

Sequencing as a tool for microbial community analysis

Lyse cells
Extract DNA (and/or RNA)

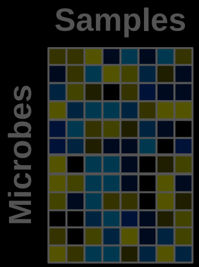
16S amplicons

George Rice, Montana State University



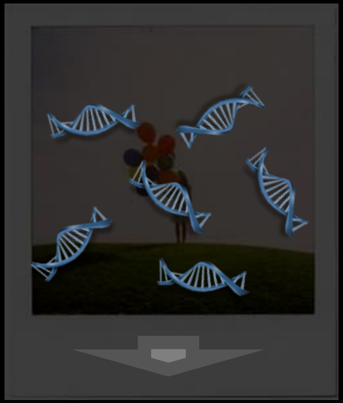
PCR to amplify the single 16S rRNA marker gene

Classify sequence
➔ microbe



Relative abundances

Meta'omic



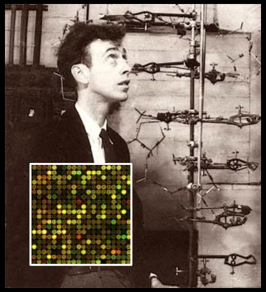
Genes,
Genomes,
Metabolic profiling,
Relative abundances,
Genetic variants...

What to do with your metagenome?

Reservoir of gene and protein functional information

Comprehensive snapshot of microbial ecology and evolution

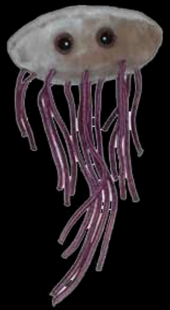
Basic science



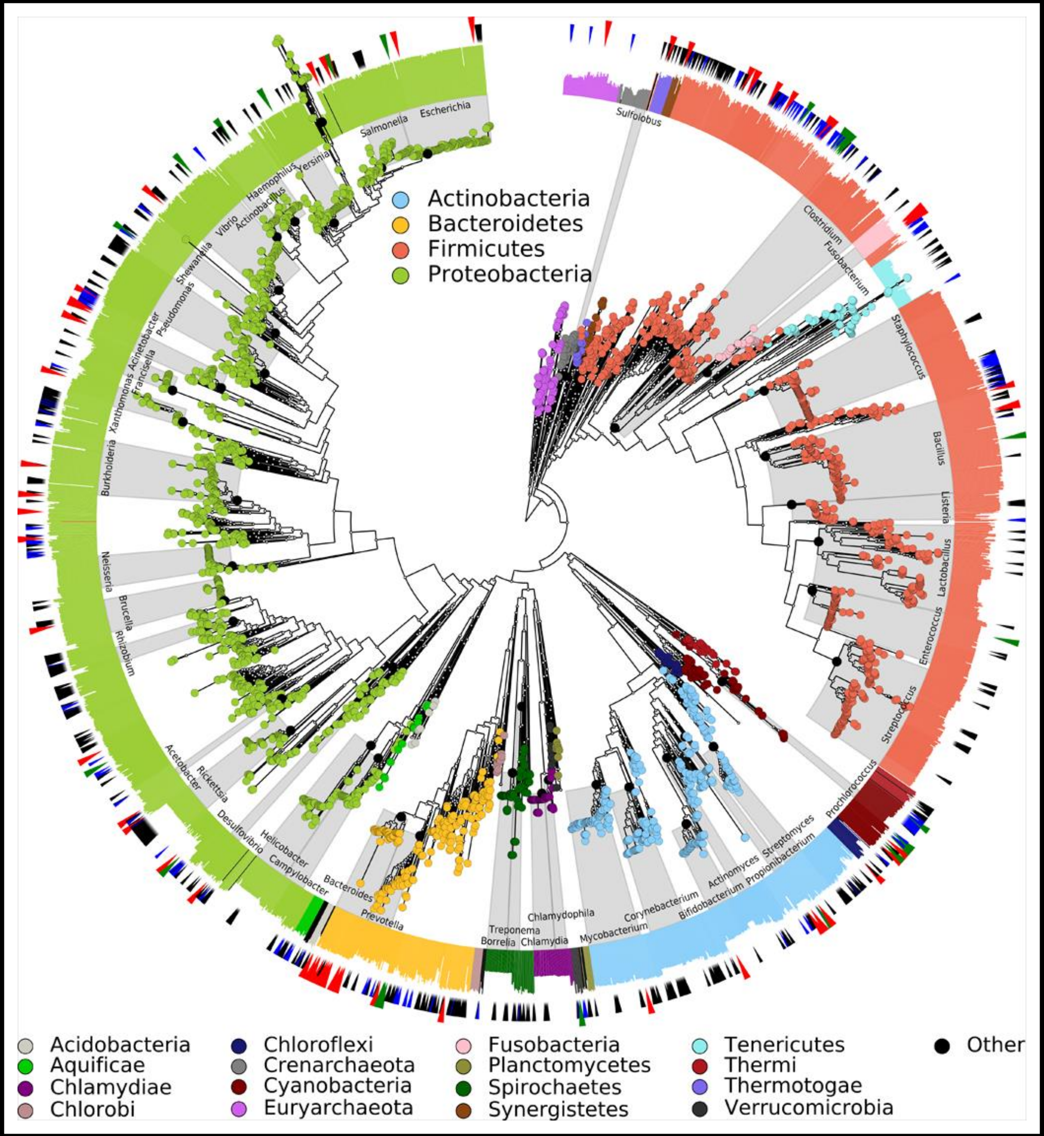
Translational

Public health tool monitoring population health and epidemiology

Diagnostic or prognostic biomarker for host disease



Composition-based analyses



Microbiome composition analyses: **phylotypes** and **binning**

RIBOSOMAL DATABASE PROJECT

rdp.cme.msu.edu

green genes

16S rRNA gene database and workbench compatible with ARB

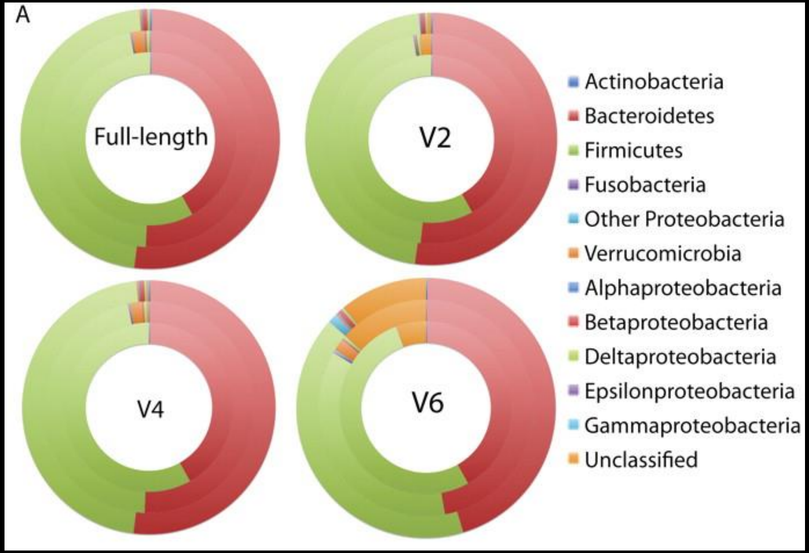
greengenes.lbl.gov

greengenes.lbl.gov

silva

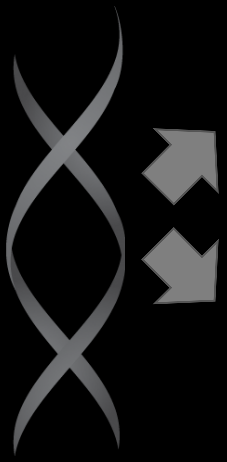
comprehensive ribosomal RNA databases

www.arb-silva.de



Binning: nontrivial assignment of reads to phylotypes

Phylotype or **operational taxonomic unit (OTU)**: organisms clonal to within some tolerance (e.g. 95%); "species"



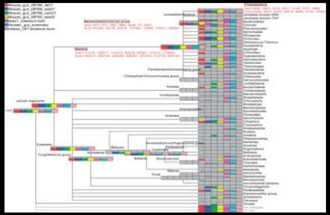
Indirect binning: BLAST etc. Relies on high similarity, reference seq.

Direct binning: analyzes seq. characteristics (GC, codons, etc.) Relies on long reads

Hamady, 2009 PhyloPythia: McHardy, 2007 TETRA: Chan, 2008 Phymm: Brady, 2009 MetaPhlAn: Segata, 2012

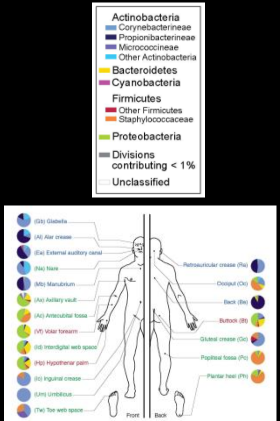
Microbiome composition analyses: **diversity**

Mitra, 2009



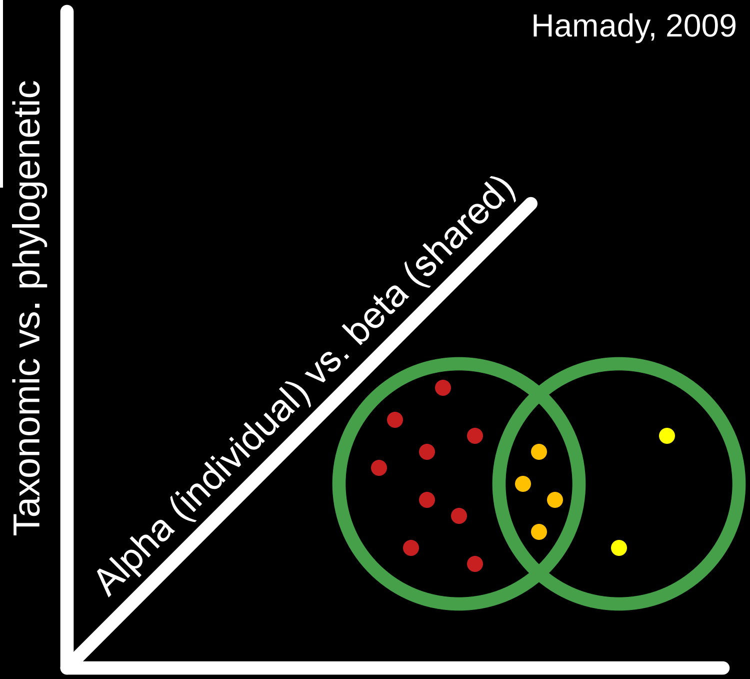
Diversity: broadly, a community's number and distribution of organisms

Also **community composition** or **structure**



Susan Holmes, Stanford

Hamady, 2009



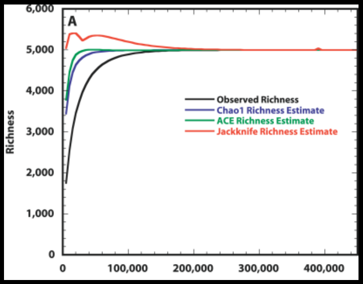
Taxonomic vs. phylogenetic

Alpha (individual) vs. beta (shared)

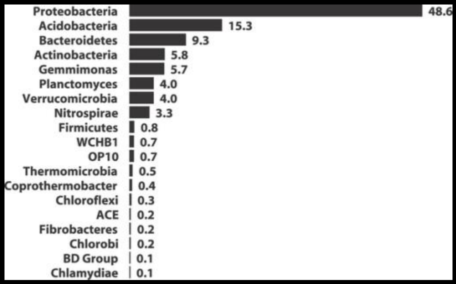
Qualitative vs. quantitative

[1 1 0 0]

[10 6 1 4]



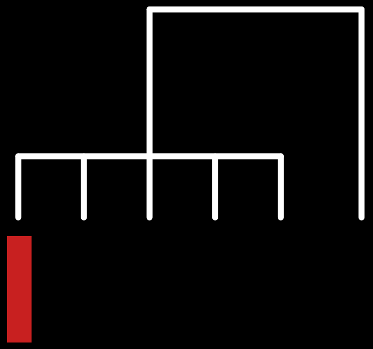
# of sequences



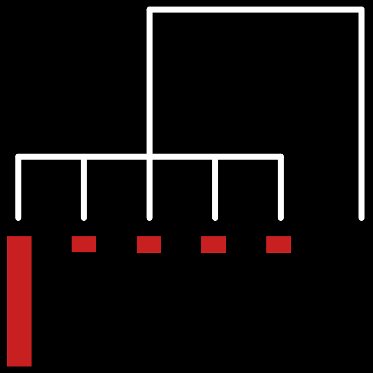
Schloss, 2006

% of sequences

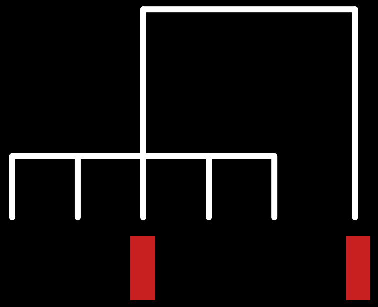
Microbiome composition analyses: **alpha diversity (1-sample) scenarios**



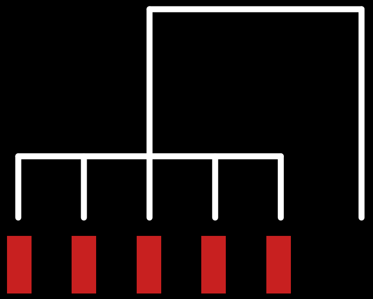
Not diverse



Qualitatively diverse
Taxonomically diverse

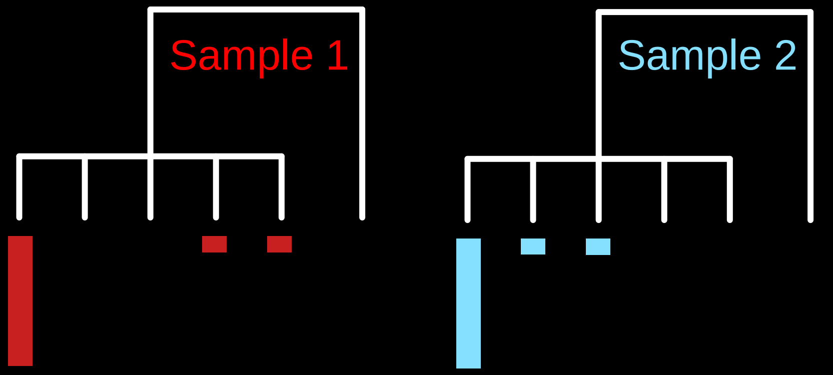


Phylogenetically diverse

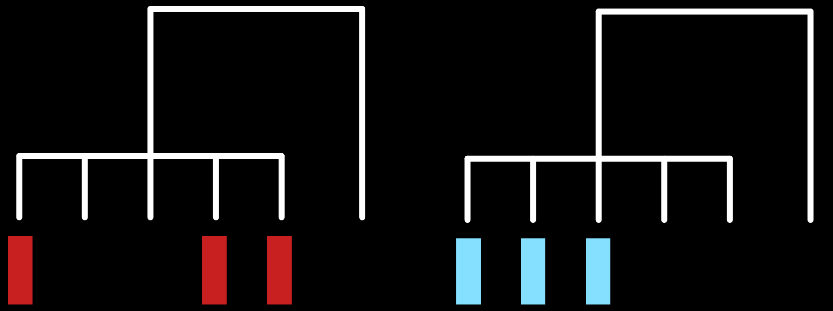


Quantitatively diverse
Taxonomically diverse

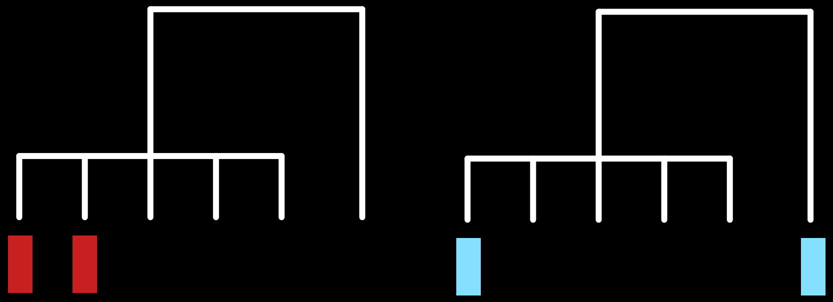
Microbiome composition analyses: **beta diversity (2-sample) scenarios**



Qualitatively diverse
Taxonomically diverse

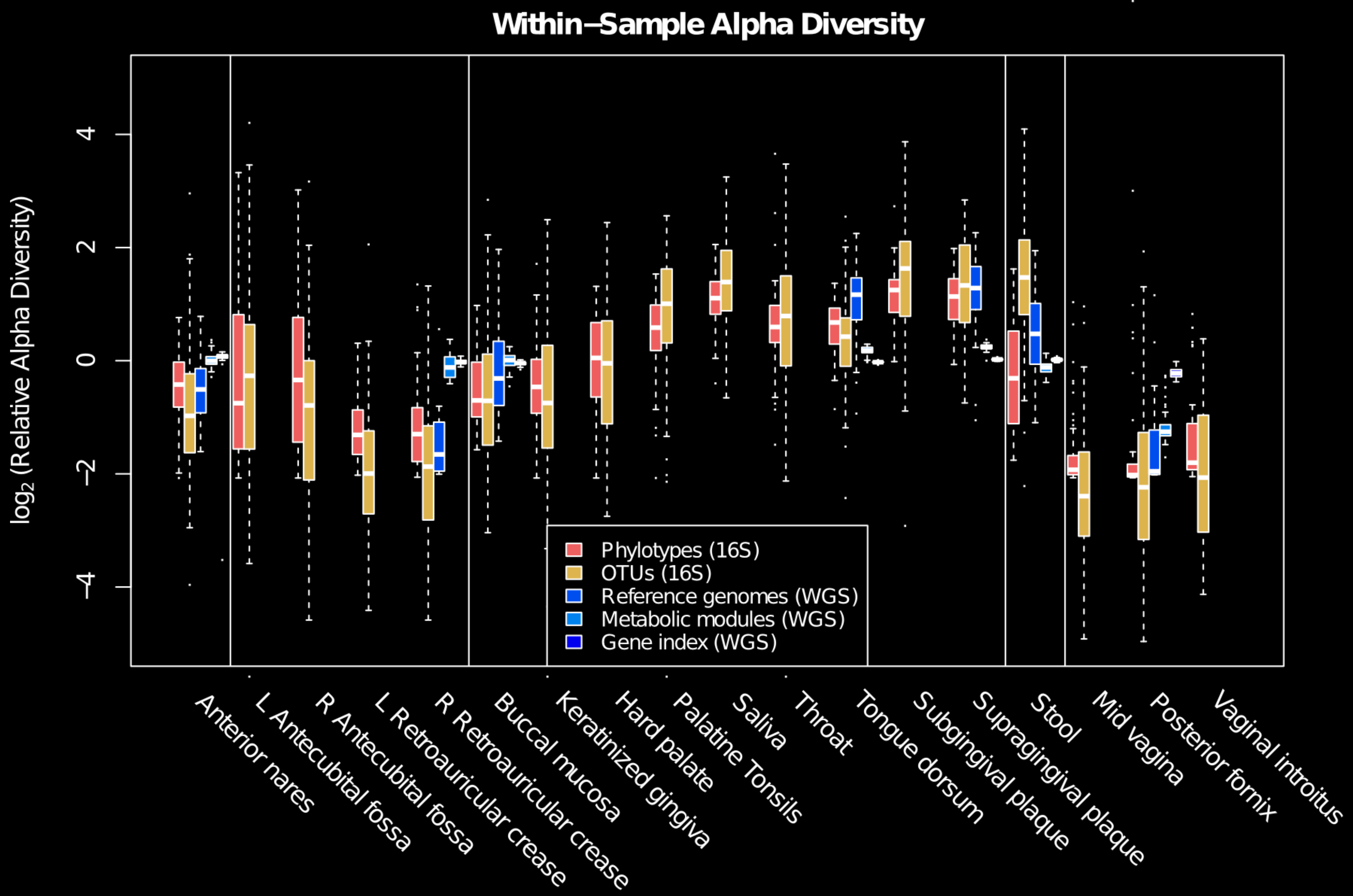


Quantitatively diverse
Taxonomically diverse

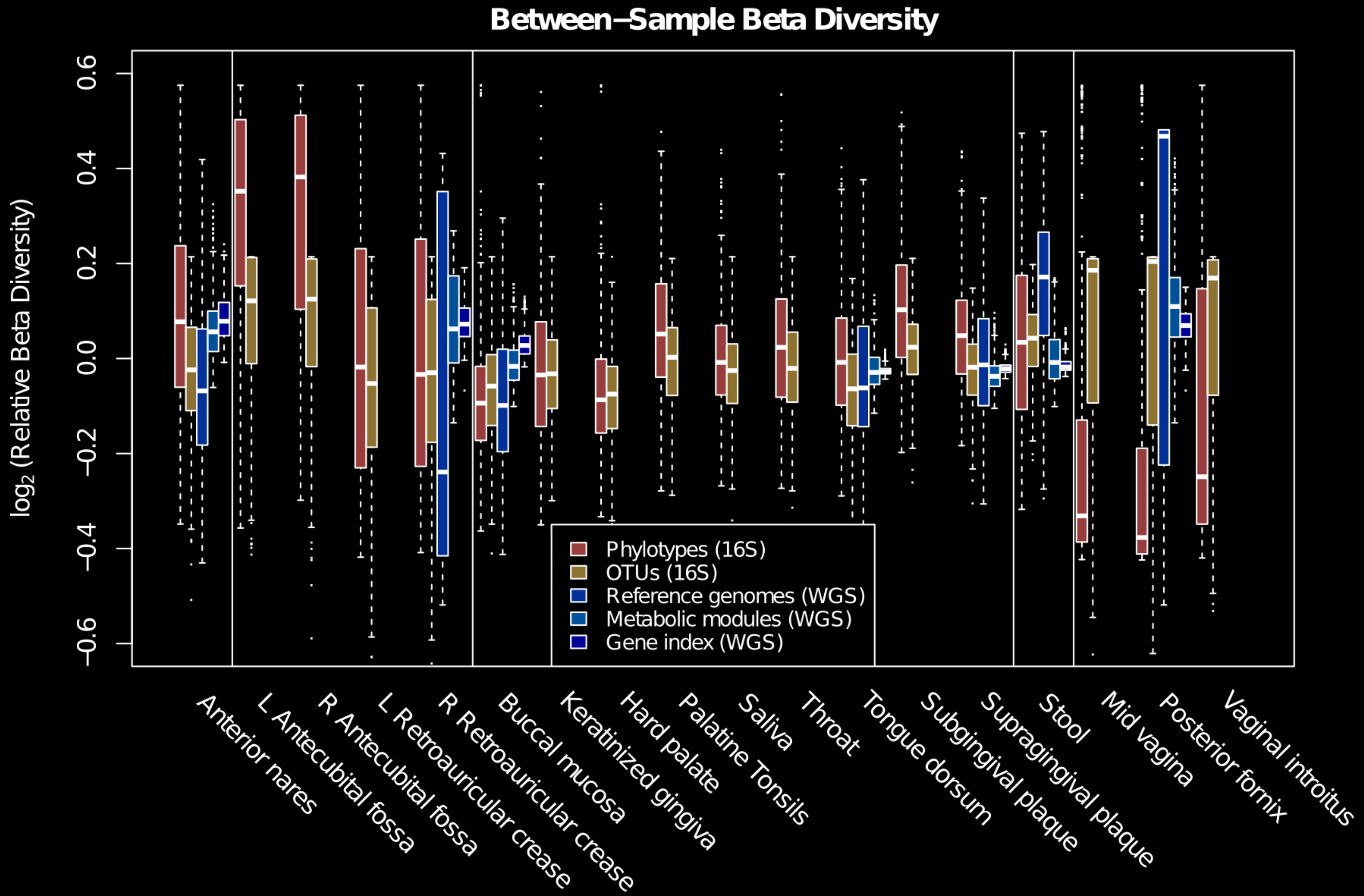


Quantitatively diverse
Phylogenetically diverse

Which human body sites harbor the greatest microbial diversity per individual?

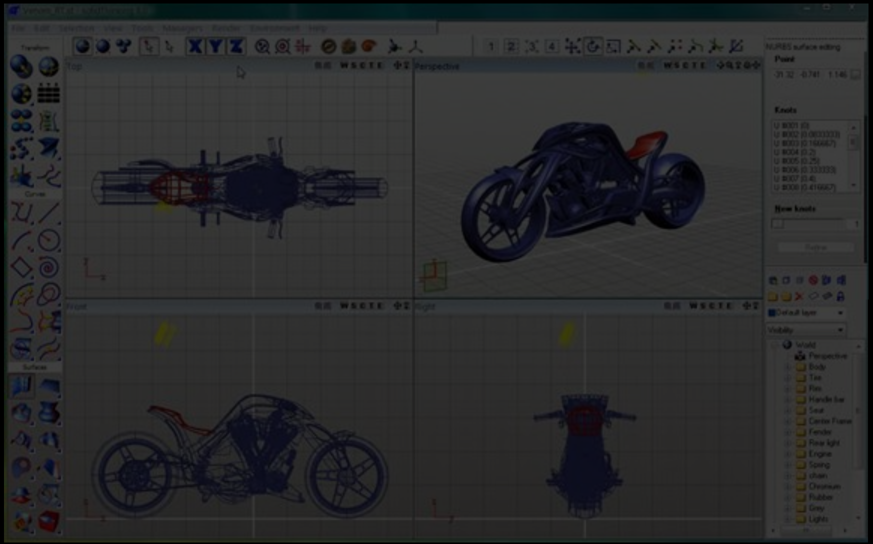


Which human body sites share the greatest microbial diversity among individuals?

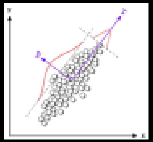


Microbiome composition analyses: **ordination**

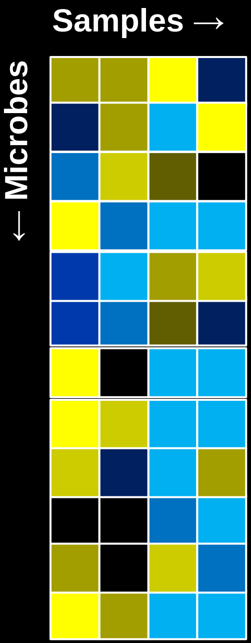
Ordination is the constrained projection of high-dimensional data into fewer dimensions.



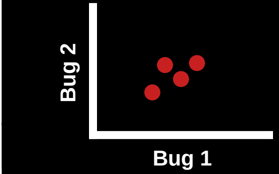
PCA or **Principal Component Analysis** guarantees the new dimensions maximize normal variation.

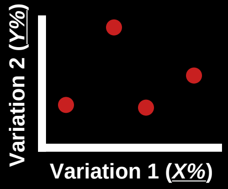


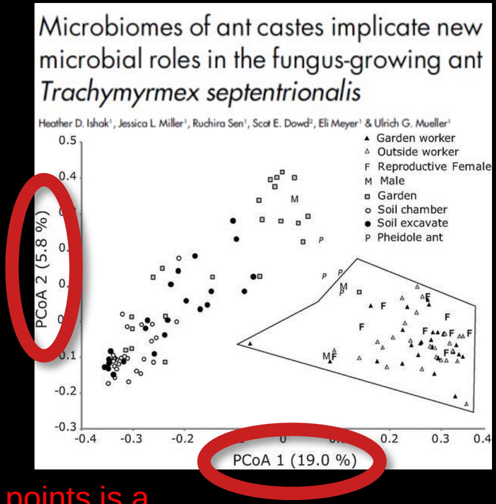
NMDS or **Nonmetric Multidimensional Scaling**, also called **PCoA** or **Principal Coordinates Analysis**, guarantees the new dimensions maximize an arbitrary similarity score (such as UniFrac beta-diversity).



Distance between points is Euclidean



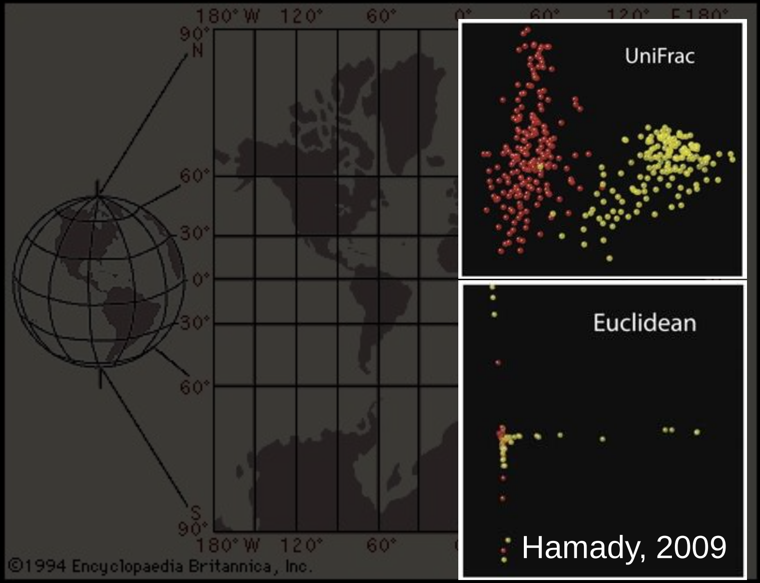




Microbiomes of ant castes implicate new microbial roles in the fungus-growing ant *Trachymyrmex septentrionalis*

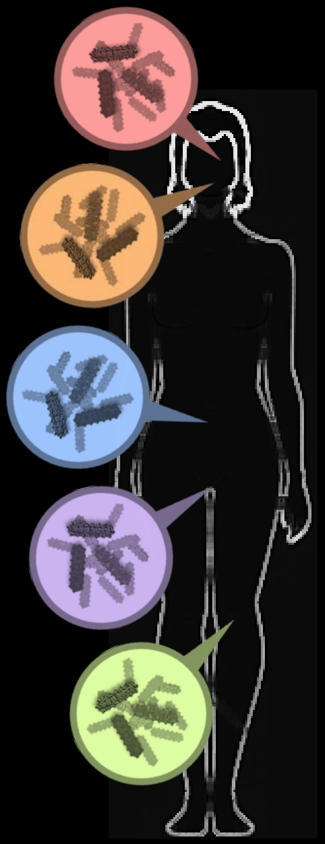
Heather D. Ishak[1], Jessica L. Miller[1], Ruchira Sen[1], Scot E. Dowd[2], Eli Meyer[1] & Ulrich G. Mueller[1]

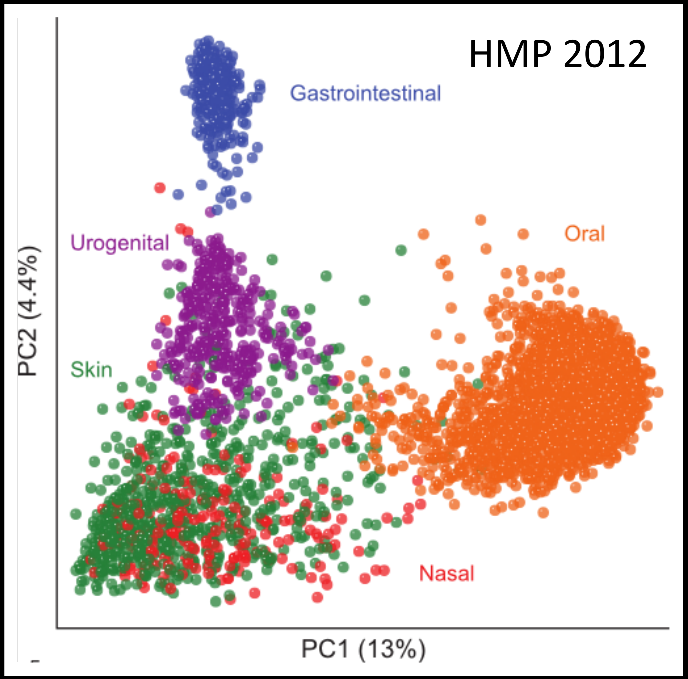
Distance between points is a ***proportional function*** of their ***similarity***

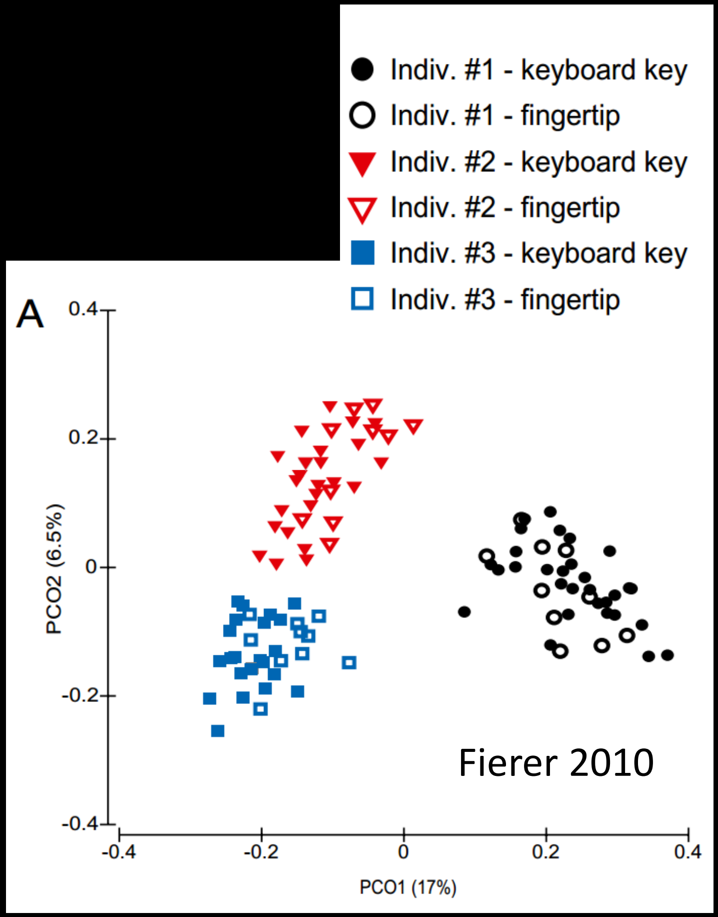


Hamady, 2009

Microbiome composition analyses: **ordination examples**

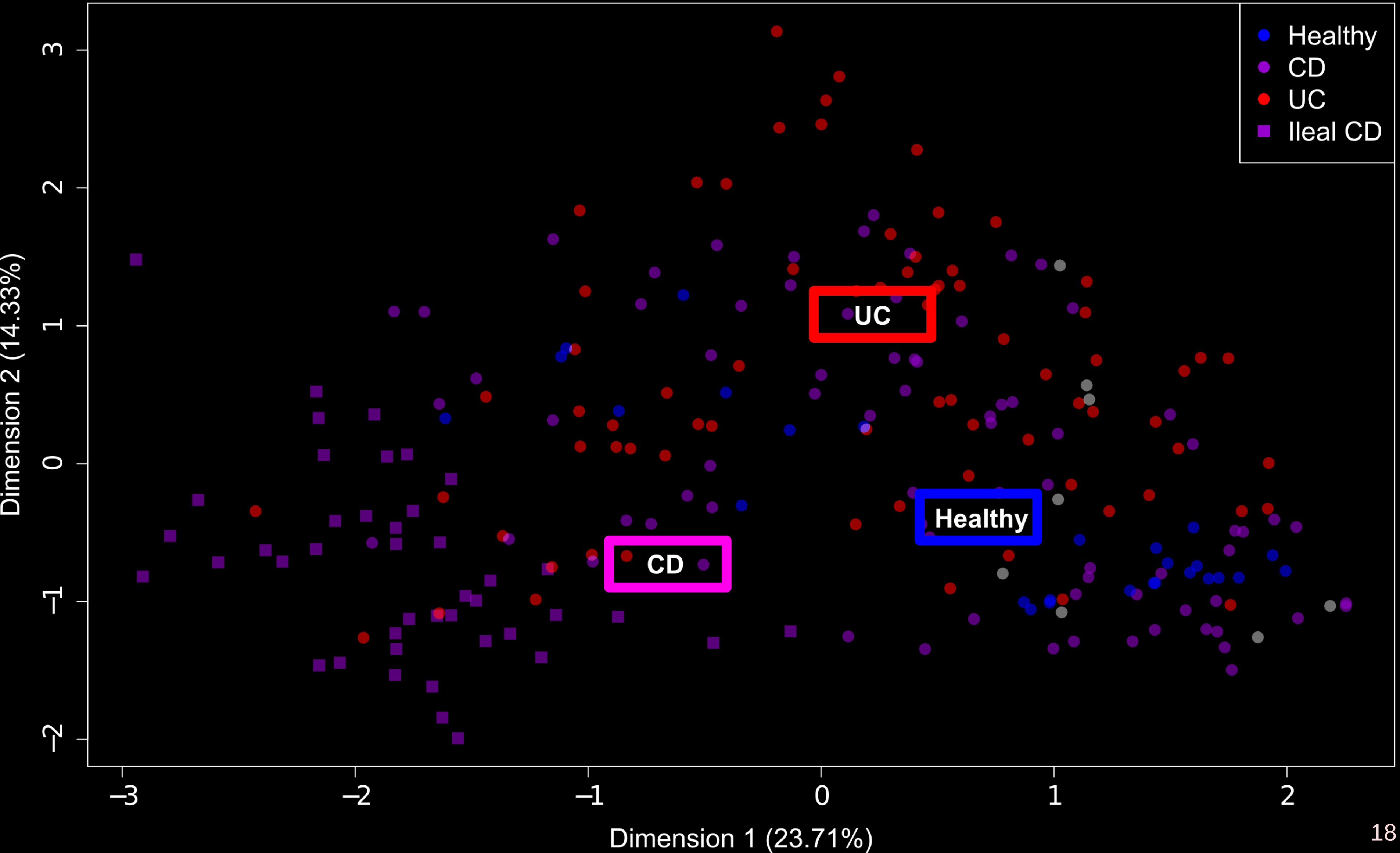



HMP 2012


Fierer 2010

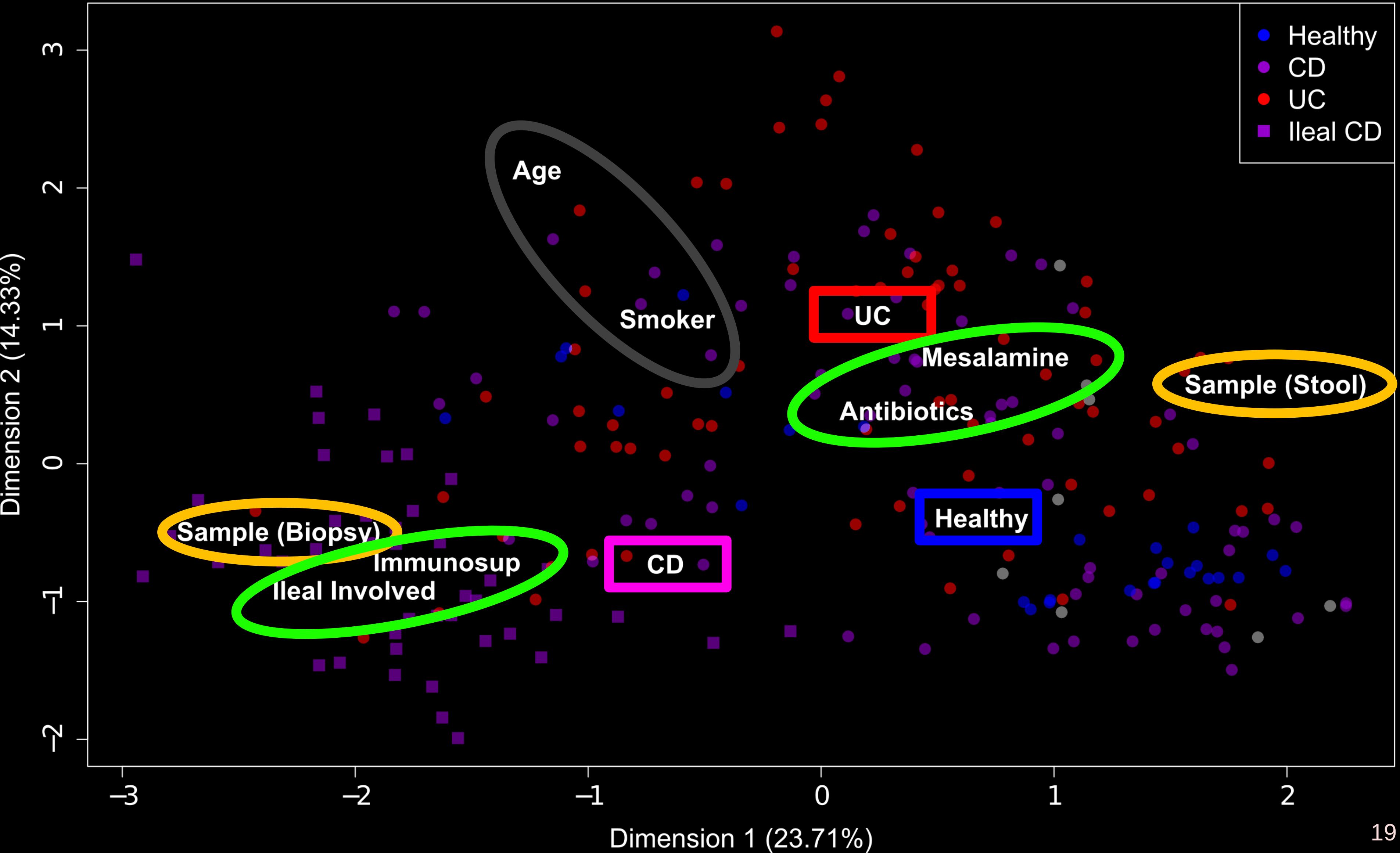
How is the gut microbiome disrupted during IBD and its treatment?

With Ramnik Xavier, Bruce Sands

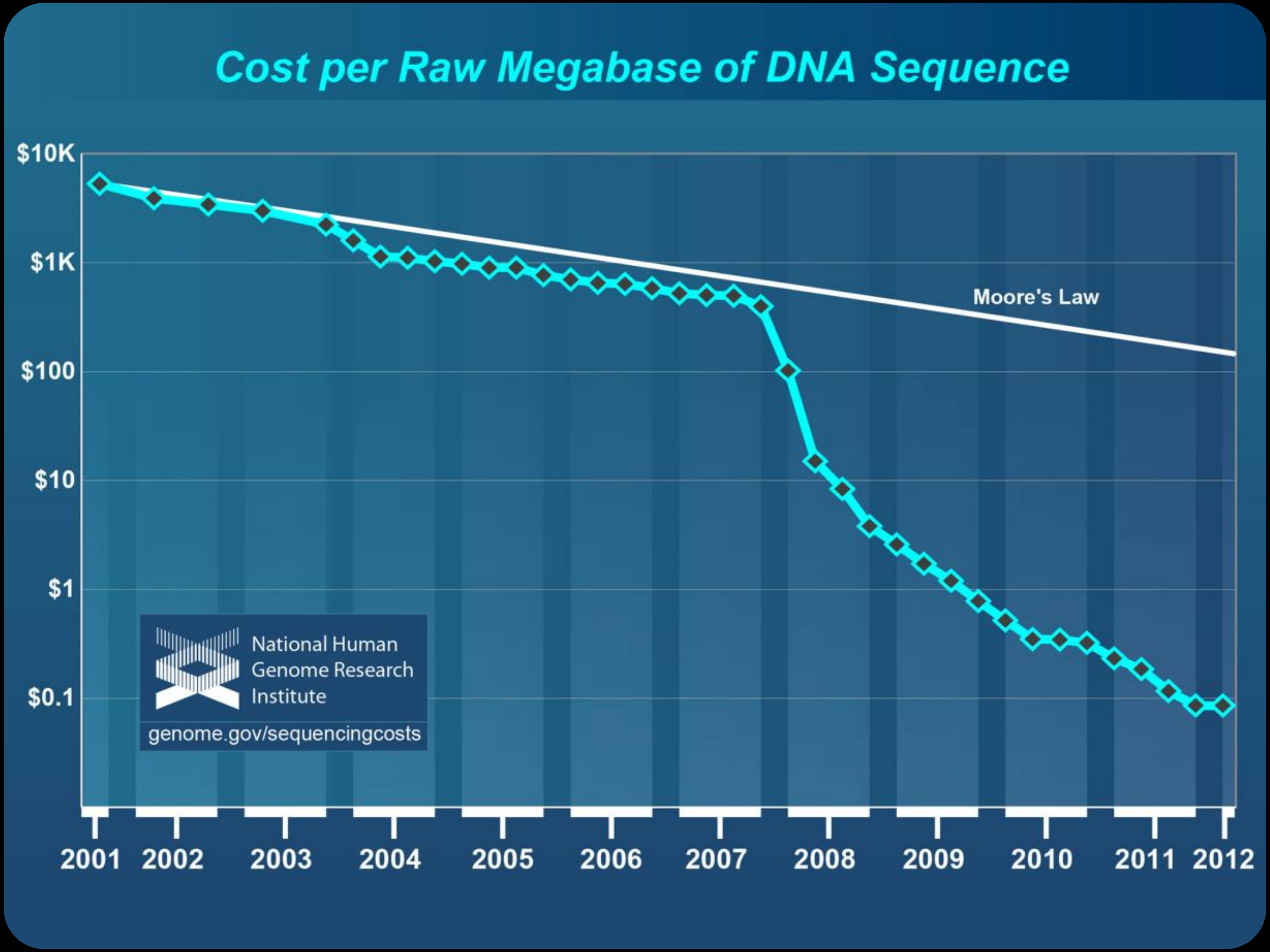


How is the gut microbiome disrupted during IBD and its treatment?

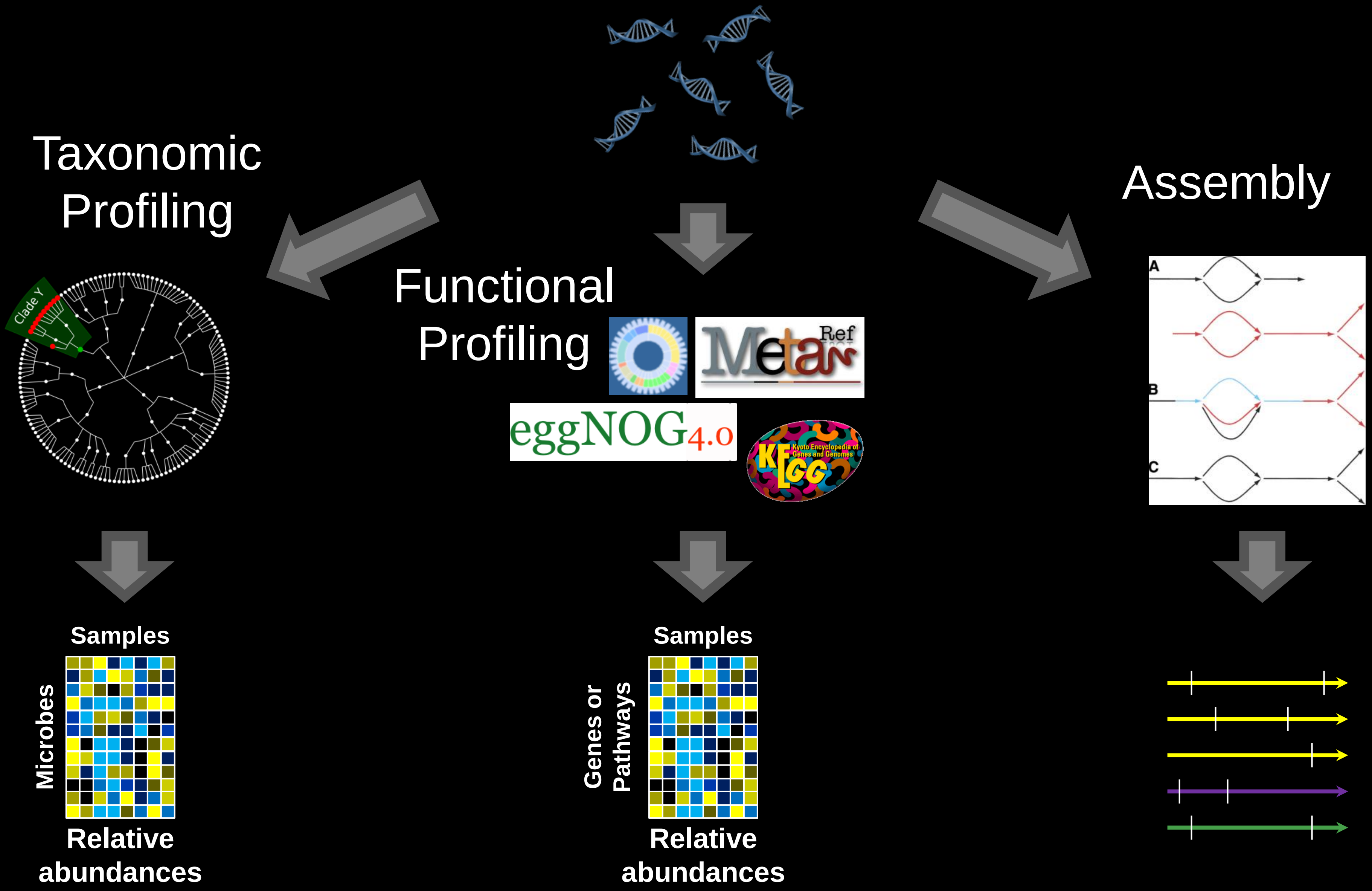
With Ramnik Xavier, Bruce Sands



Meta’omic analyses



Typical shotgun metagenome and metatranscriptome analyses



Taxonomic Profiling

Functional Profiling

Assembly

Samples

Microbes

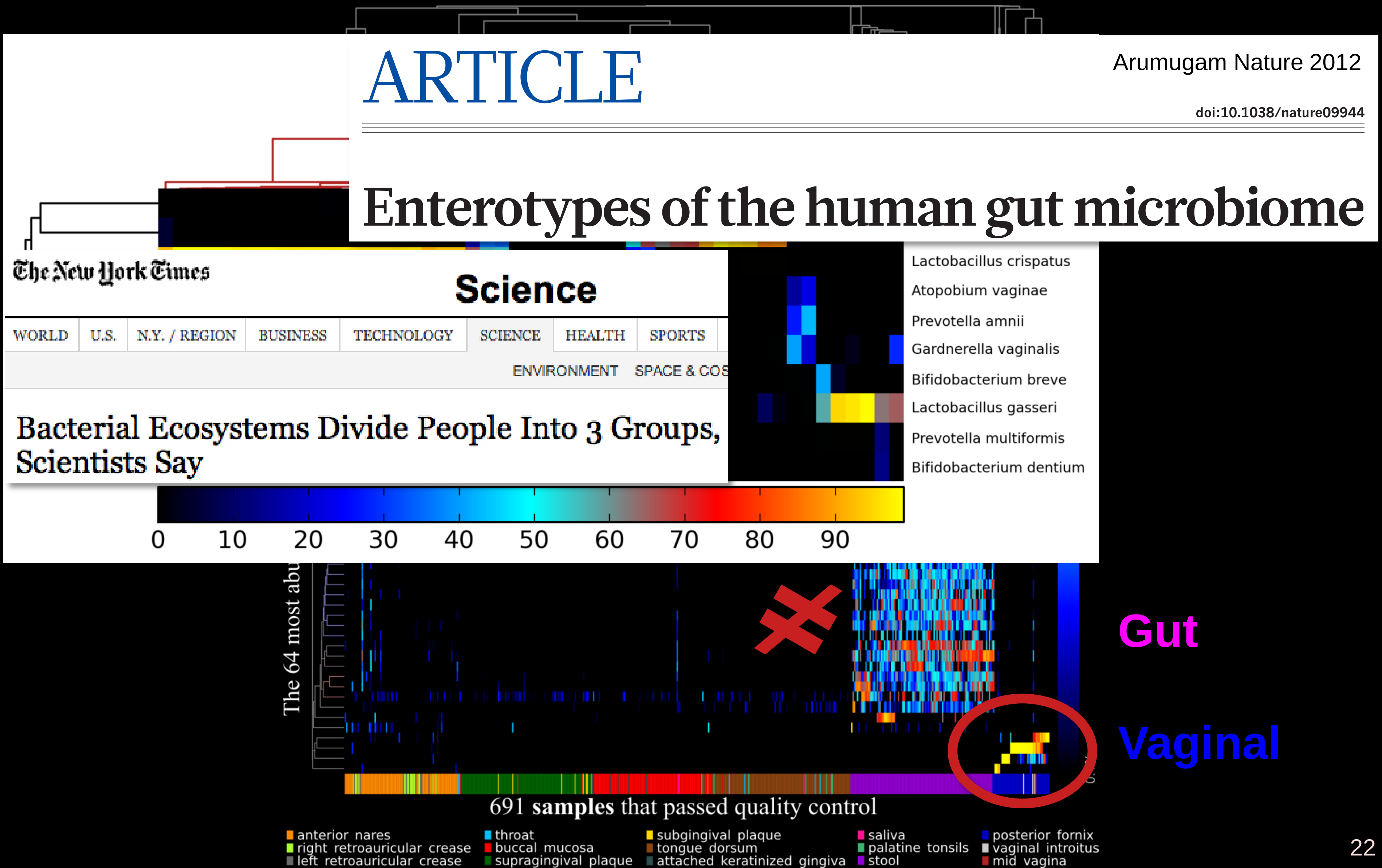
Relative abundances

Samples

Genes or Pathways

Relative abundances

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

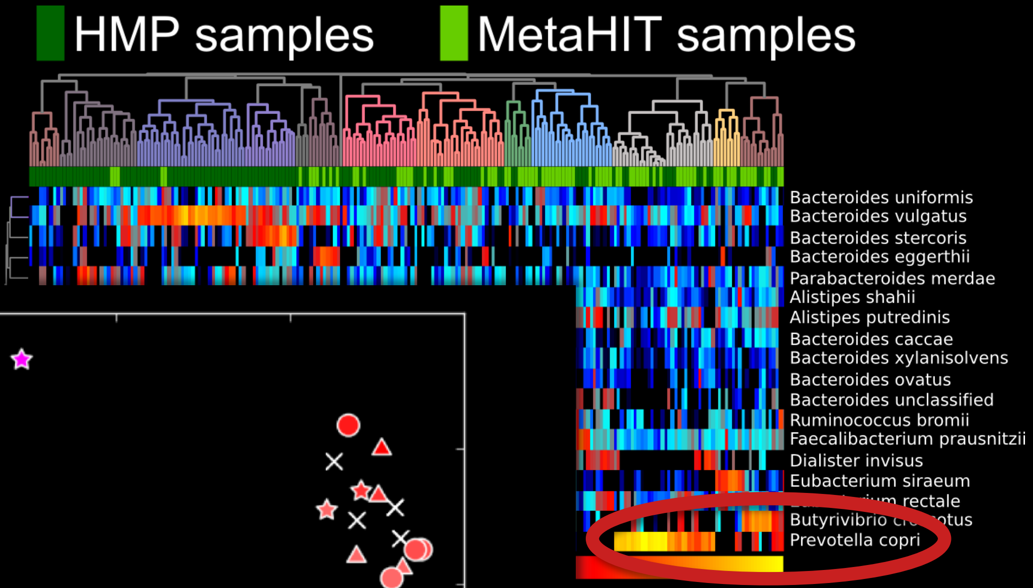
Science

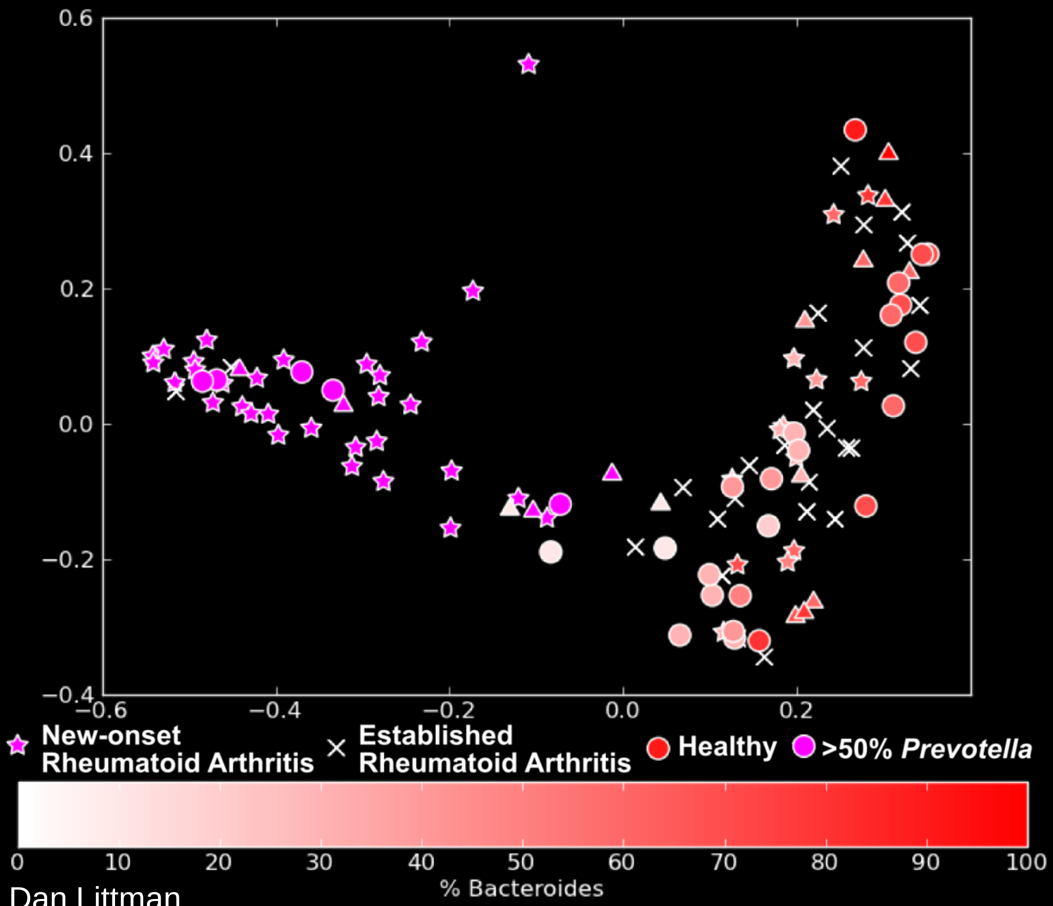
WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS

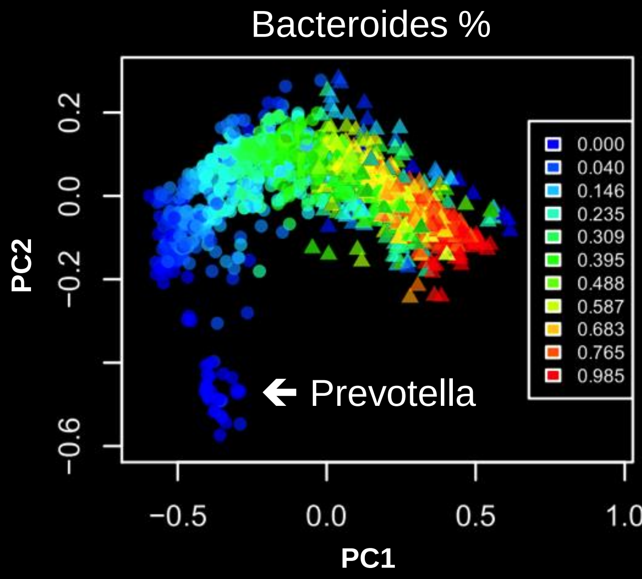
ENVIRONMENT | SPACE & COS

Bacterial Ecosystems Divide People Into 3 Groups, Scientists Say

Are there discrete "types" of typical human microbiomes?

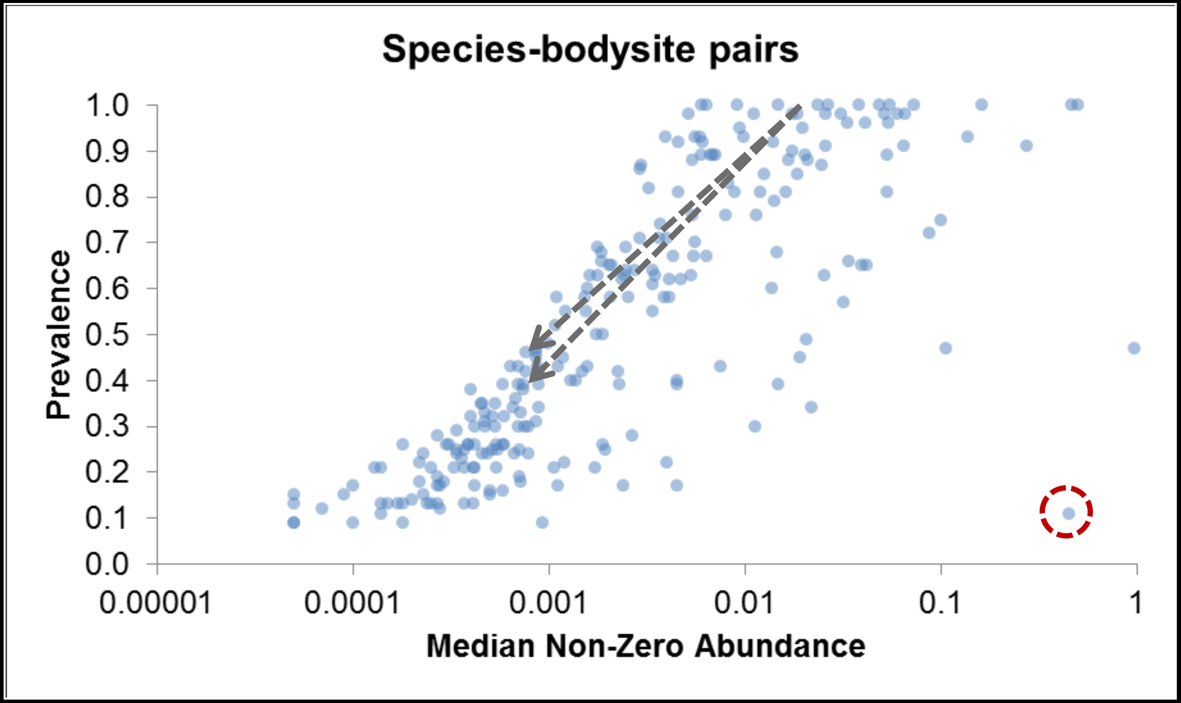






Dan Littman

Species prevalence vs. species abundance

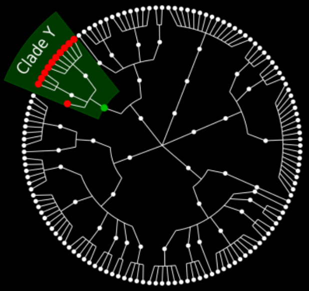


site	bug	log10 med non0 abund	prevalence	tp 1-2 stability
Stool	s__Prevotella_copri	-0.3	0.11	1.00

site	bug	log10 med non0 abund	prevalence	tp 1-2 stability
Tongue_dorsum	s__Actinomyces_odontolyticus	-1.8	1.00	1.00
Buccal_mucosa	s__Actinomyces_odontolyticus	-3.1	0.47	0.78
Supragingival_plaque	s__Actinomyces_odontolyticus	-3.1	0.38	0.50

Typical shotgun metagenome and metatranscriptome analyses

Taxonomic Profiling



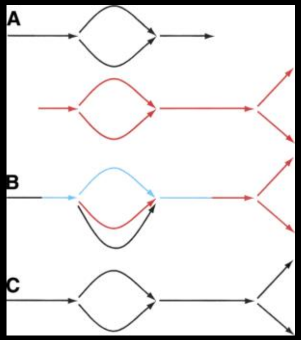
Functional Profiling

MetaRef

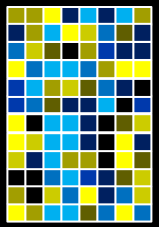
eggNOG4.0

KEGG Kyoto Encyclopedia of Genes and Genomes

Assembly

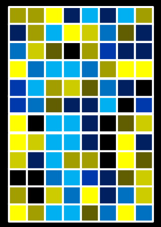


Samples

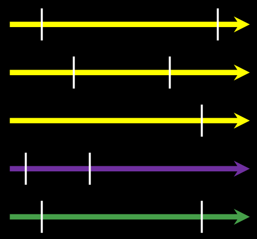


Relative abundances

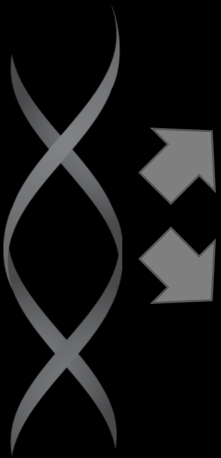
Samples



Relative abundances



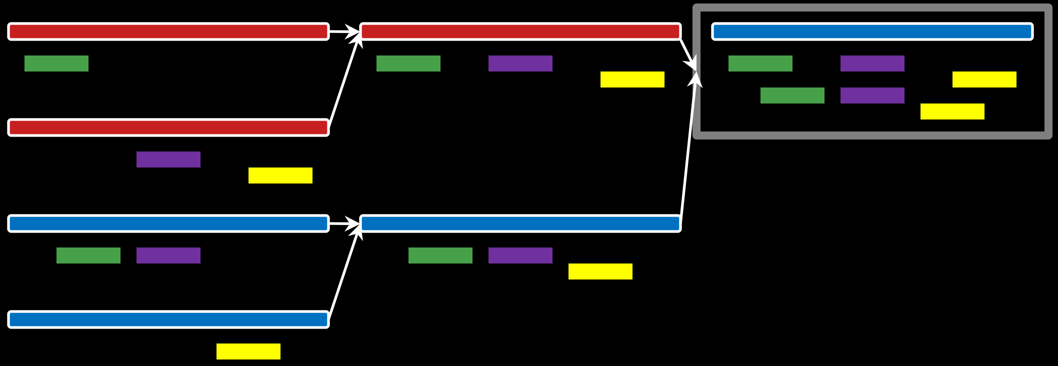
Metagenomic analyses: gene calling and proxygenes



Extrinsic gene calling: BLAST etc. (proxygenes)

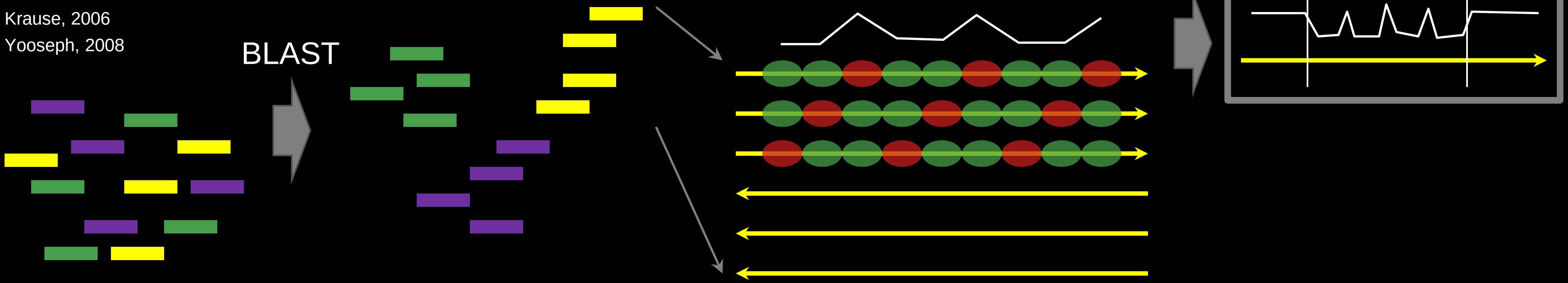
Intrinsic gene calling: ORF detection from seq.

Dalevi, 2009



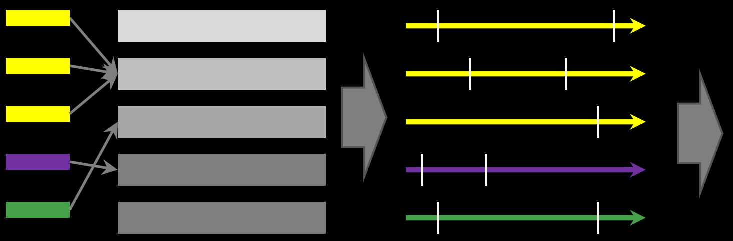
Krause, 2006
Yooseph, 2008

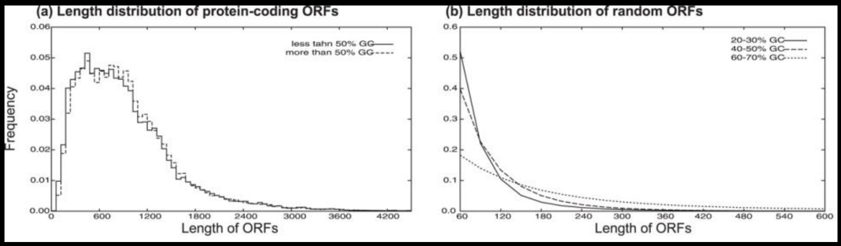
BLAST

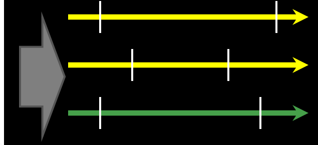


Orphelia: Hoff, 2009
MetaGene: Noguchi, 2006

HMM models





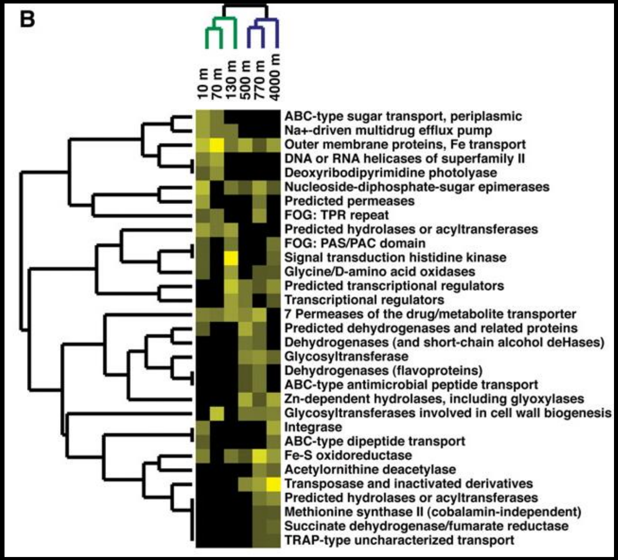


Metagenomic analyses: molecular functions and biological roles

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

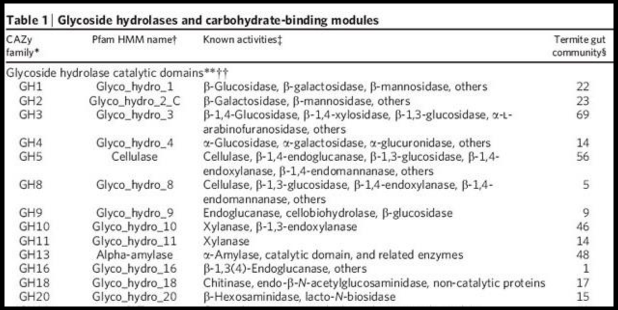
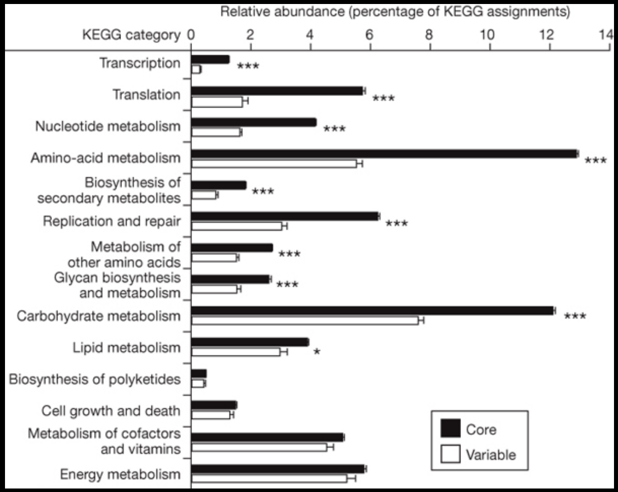


Table 1 | Glycoside hydrolases and carbohydrate-binding modules

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

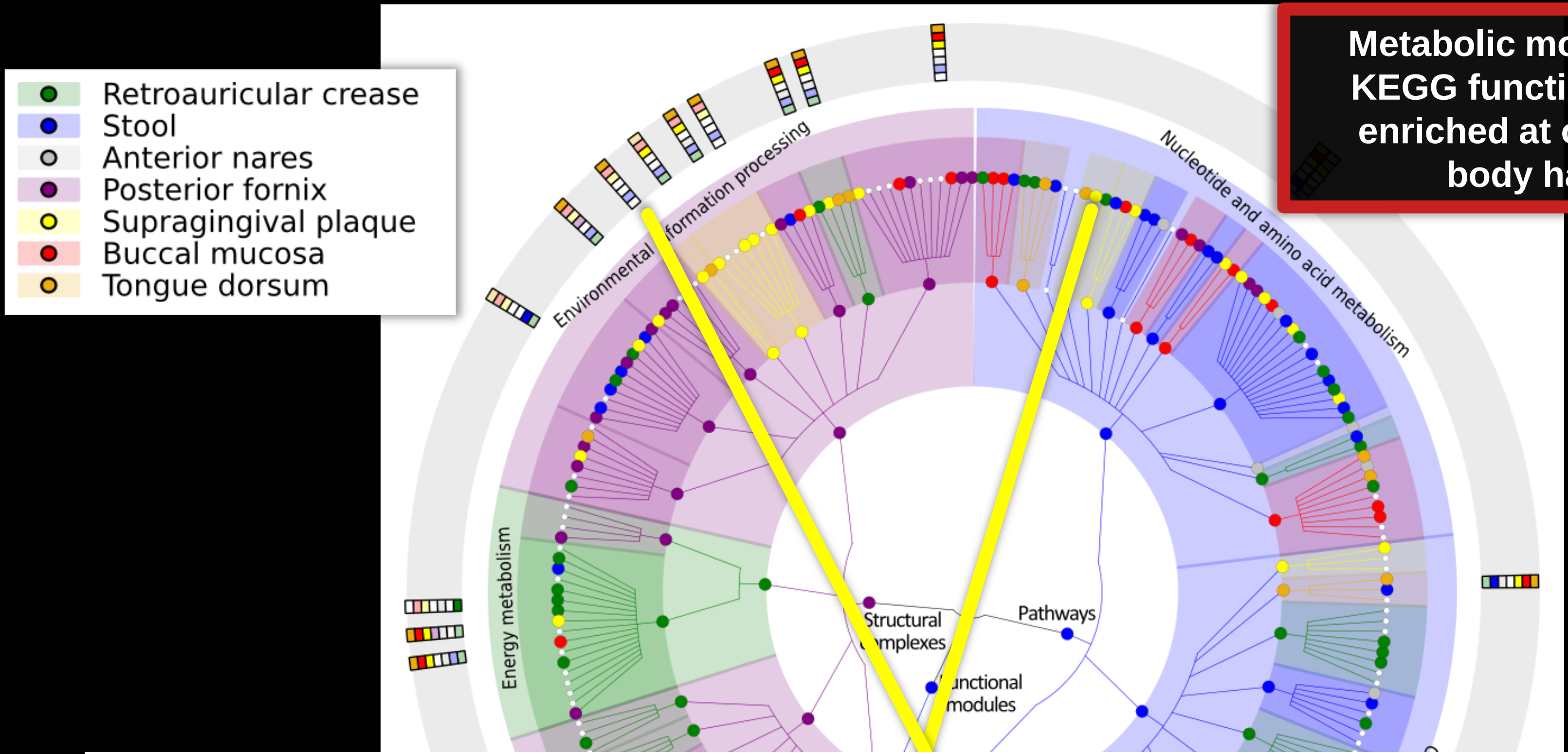


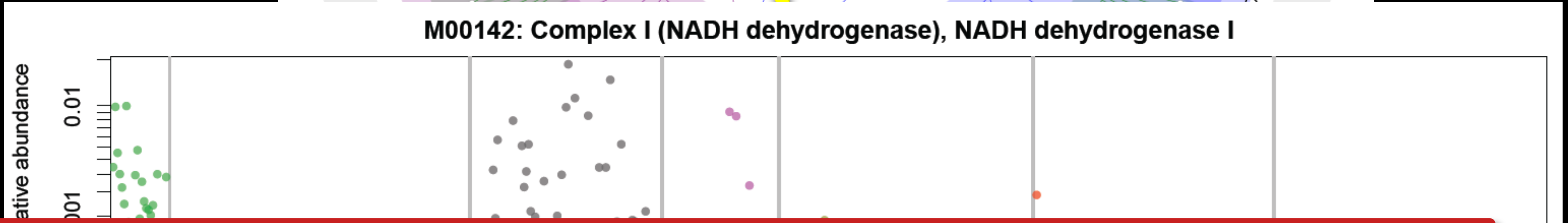
Turnbaugh, 2009

Niche specialization in human microbiome function

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

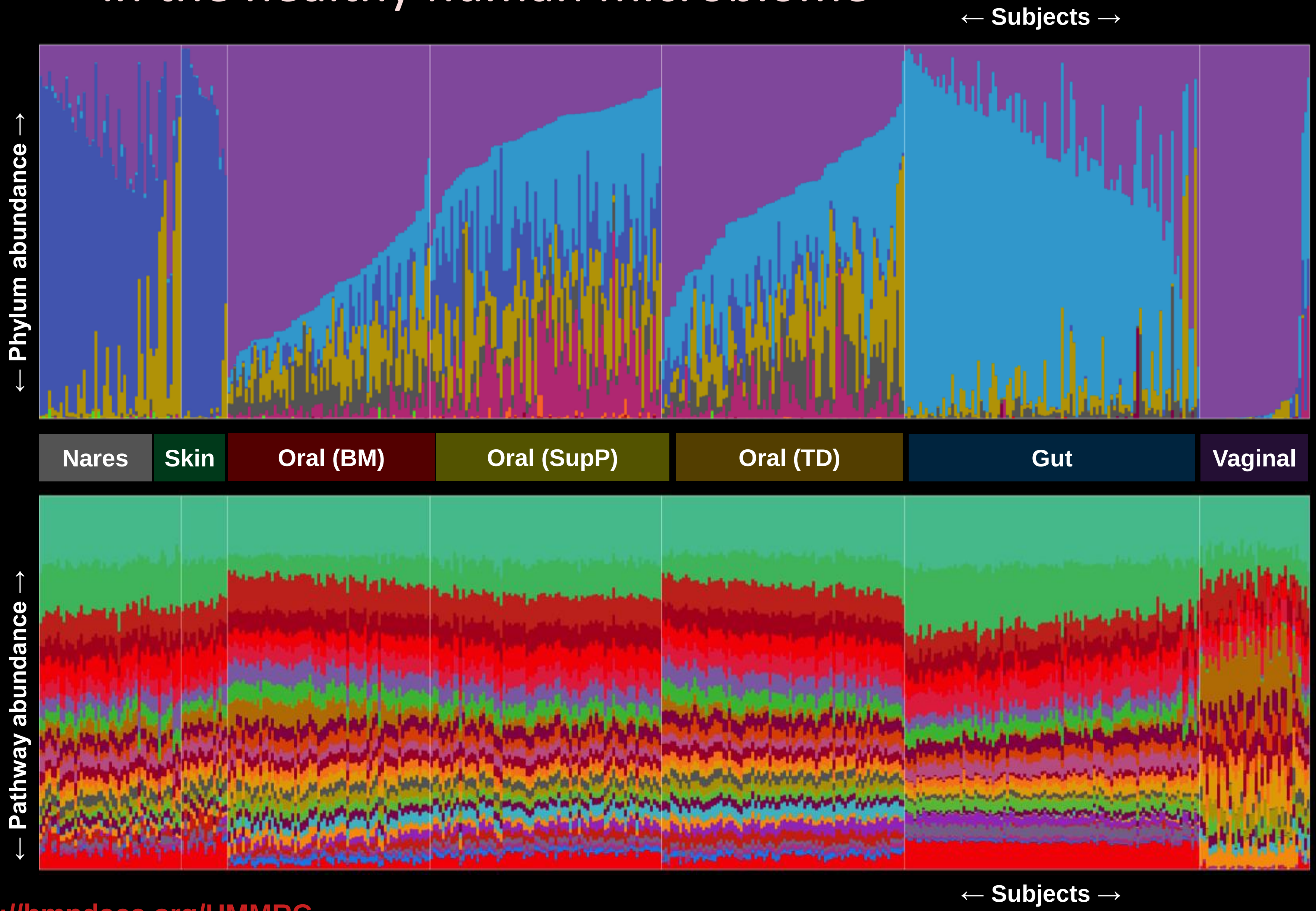
Metabolic modules in the KEGG functional catalog enriched at one or more body habitats




M00142: Complex I (NADH dehydrogenase), NADH dehydrogenase I

- 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



http://hmpdacc.org/HMMRC

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

Plan

- Informal survey
- Metagenomics concepts & examples
- Tools for taxonomic profiling
    - MetaPhlAn
- Tools for functional profiling
    - HUMAnN
    - ShortBRED
    - PICRUSt
- Tools for testing associations
    - LEfSe
    - MaAsLin
    - CCREPE
- Resources
- Research vignette (time permitting)

The two big questions…

Who is there?

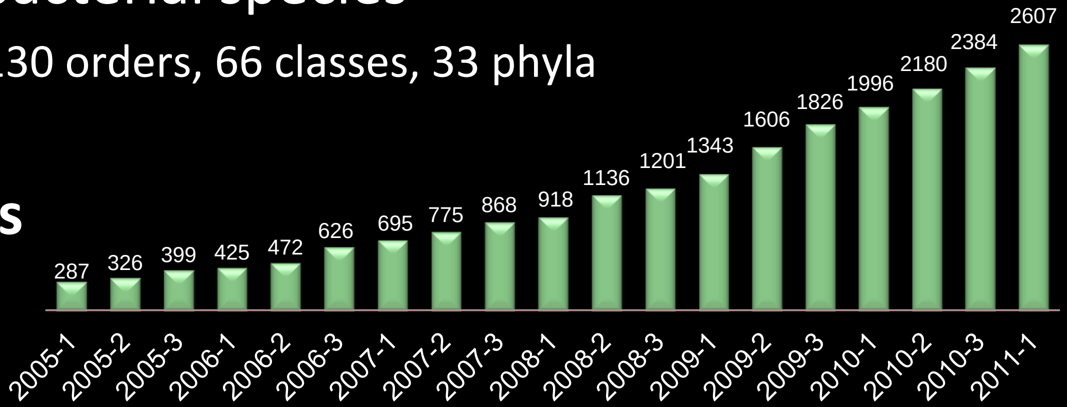
(taxonomic profiling)

What are they doing?

(functional profiling)

Reference genomes

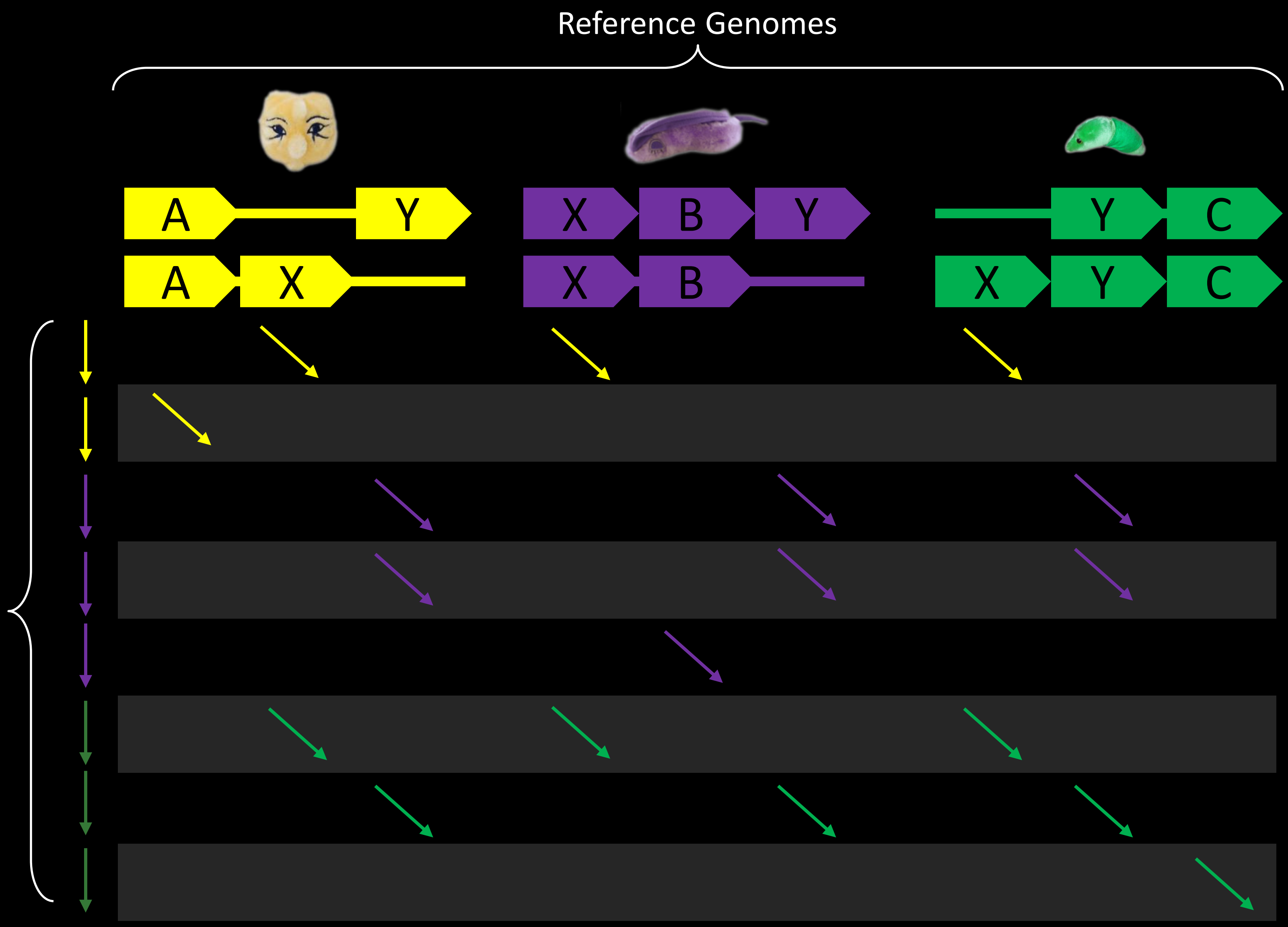
- IMG alone now contains ~3,100 bacterial genomes
  - Plus ~100 archaeal, ~100 eukaryotic, and a few thousand viruses
  - About half final and half draft
- These comprise 1,222 bacterial species
  - 652 genera, 278 families, 130 orders, 66 classes, 33 phyla
  - 2,383 total clades
- **And roughly 10M genes**



- These genes and genomes are a tremendous resource to:
  - Identify conserved markers that can be used to infer phylogeny
  - Identify unique markers that can be used to infer taxonomy
  - Relate the microbial members of a community to their metagenomic functional potential

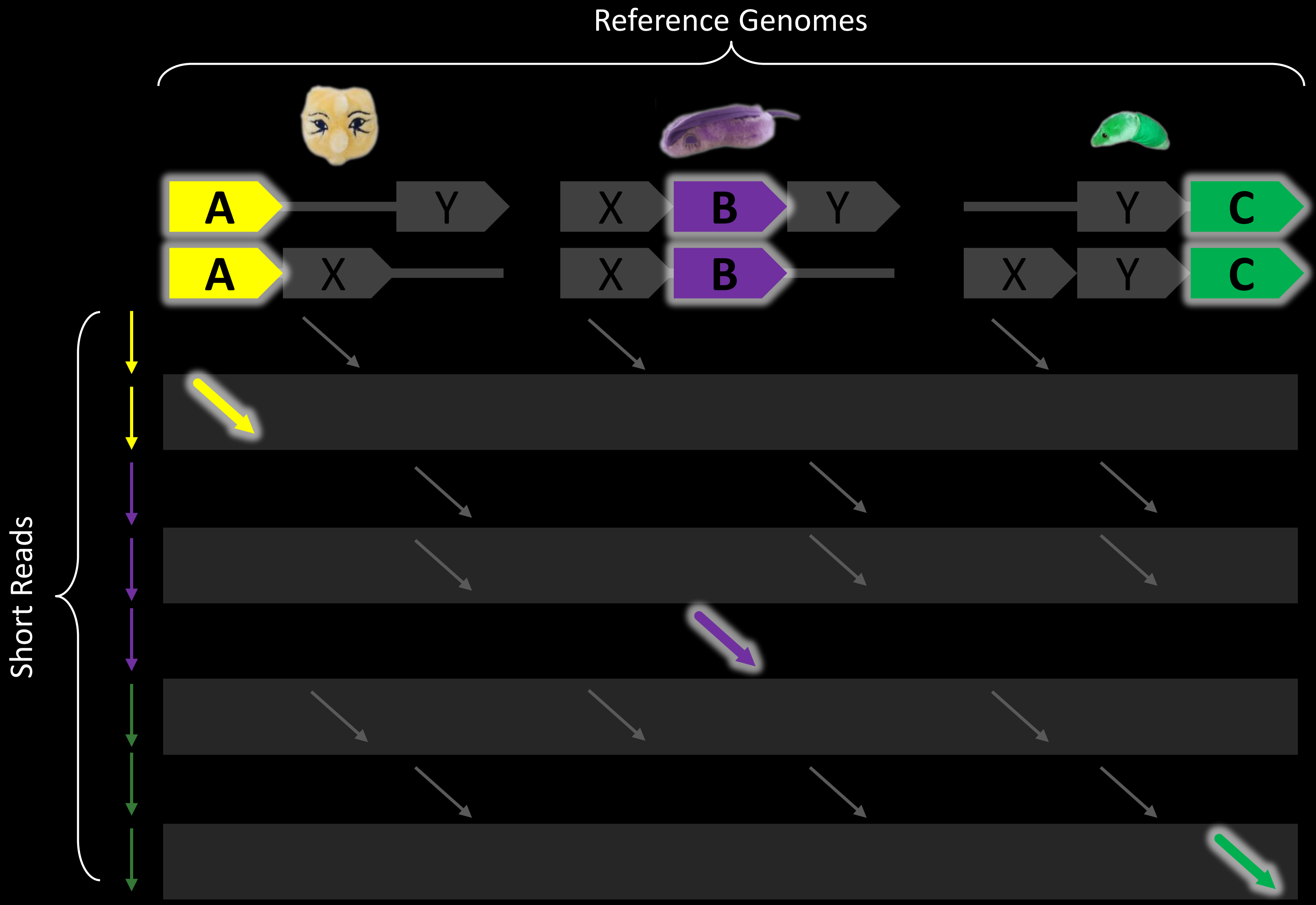
MetaPhlAn
Metagenomic Phylogenic Analysis

Nicola Segata



MetaPhlAn
Metagenomic Phylogenic Analysis

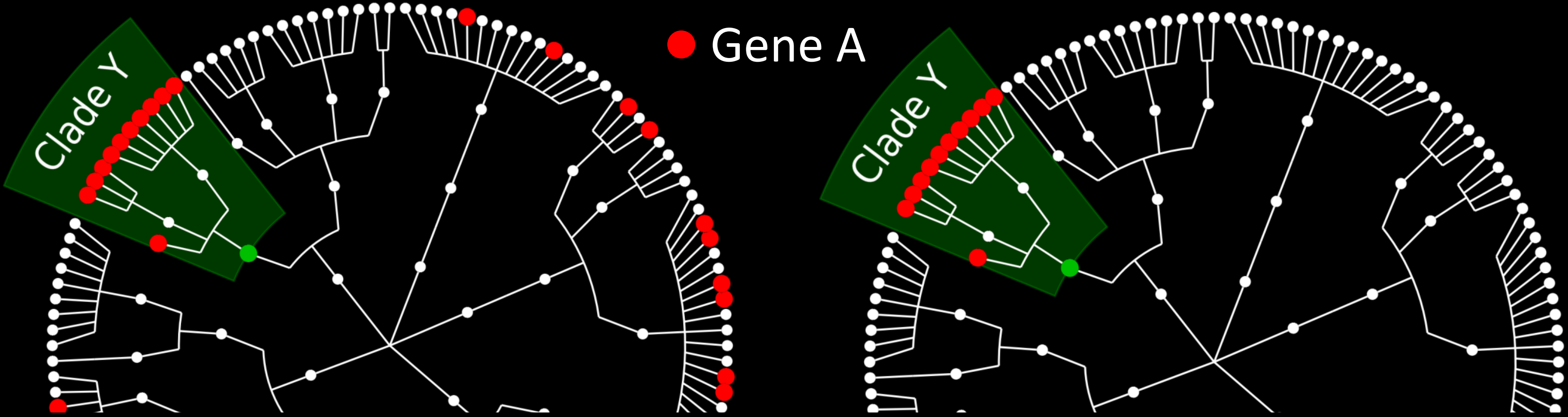
Nicola Segata

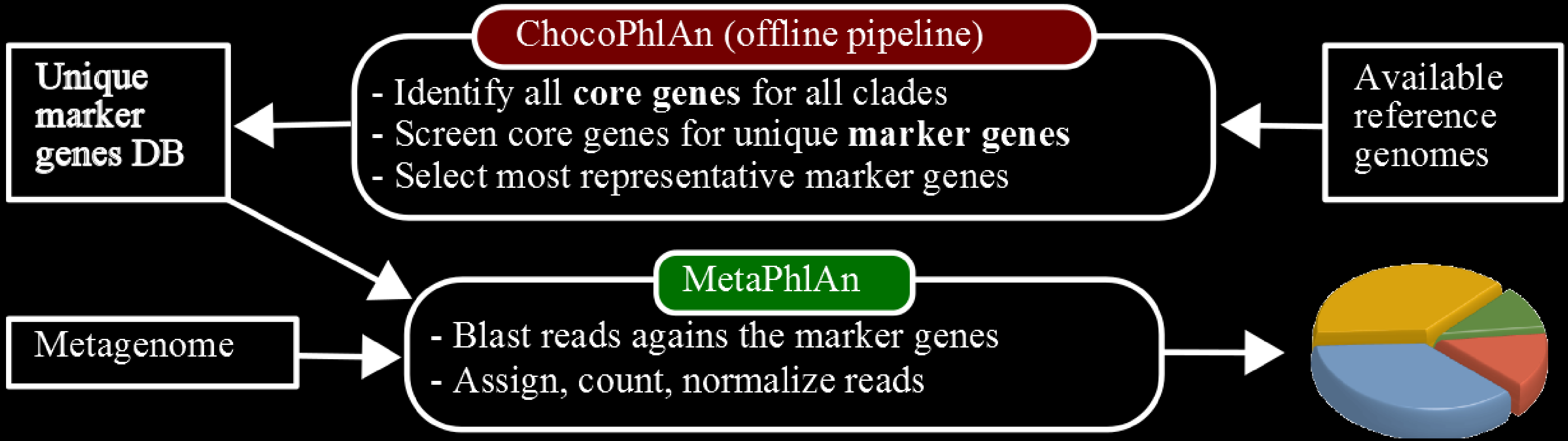


MetaPhlAn overview

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

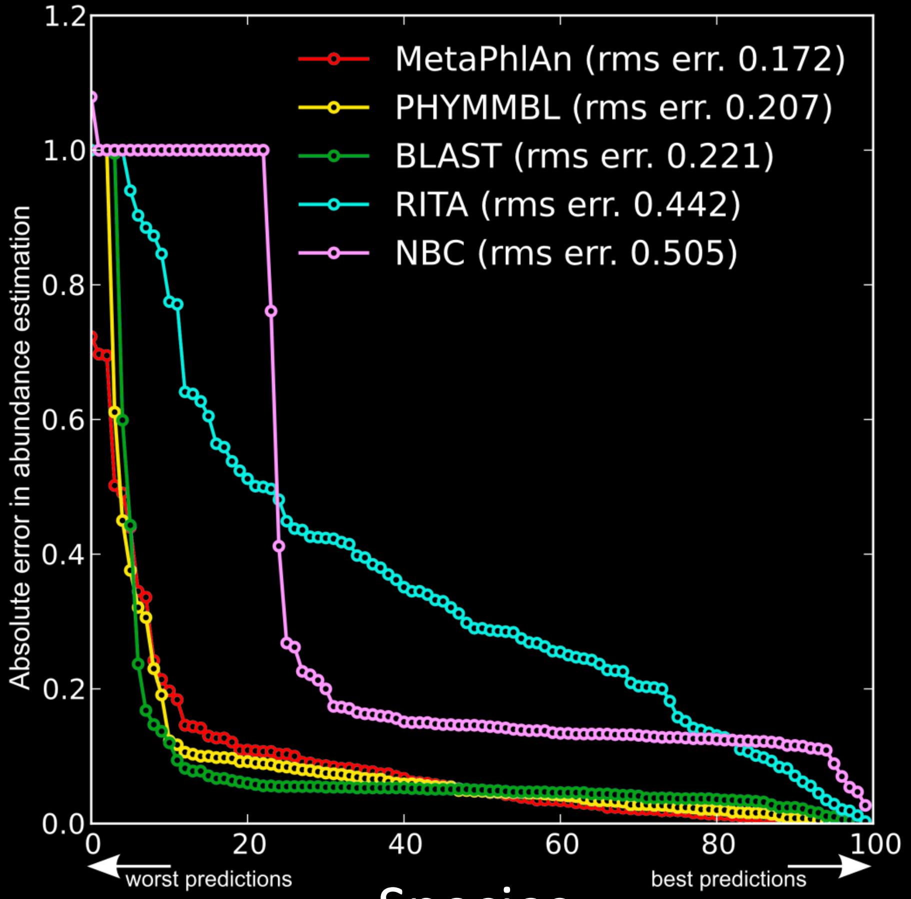
Available reference genomes

MetaPhlAn

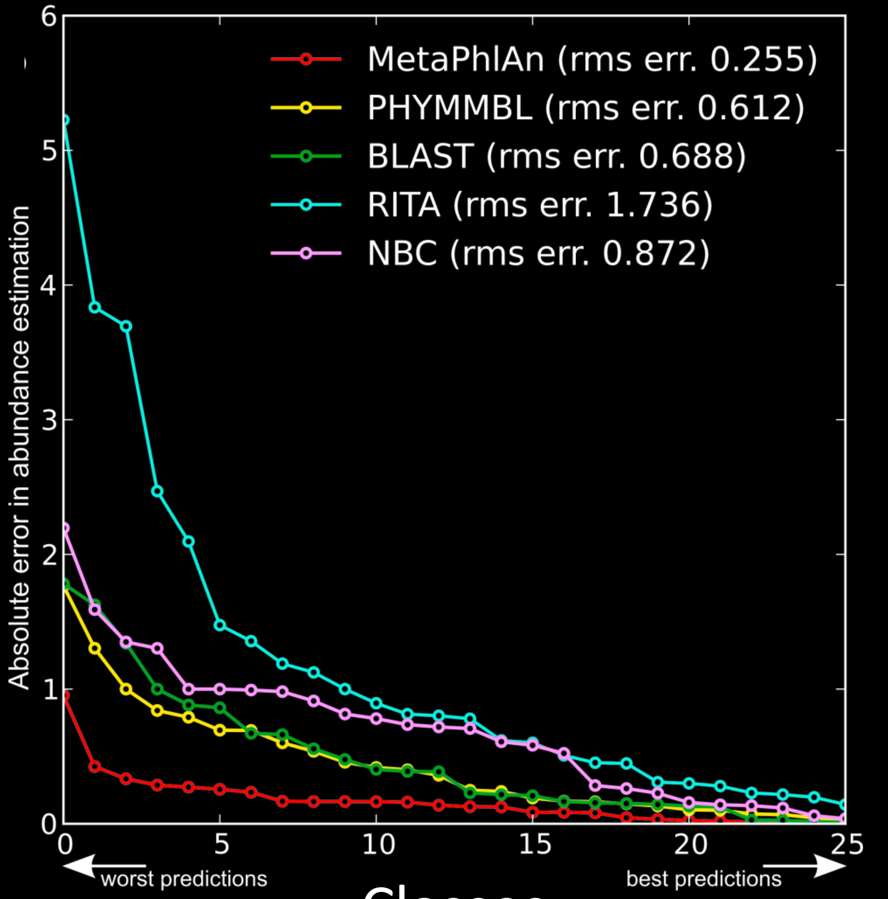
- Blast reads agains the marker genes
- Assign, count, normalize reads

Metagenome

Evaluation of MetaPhlAn accuracy



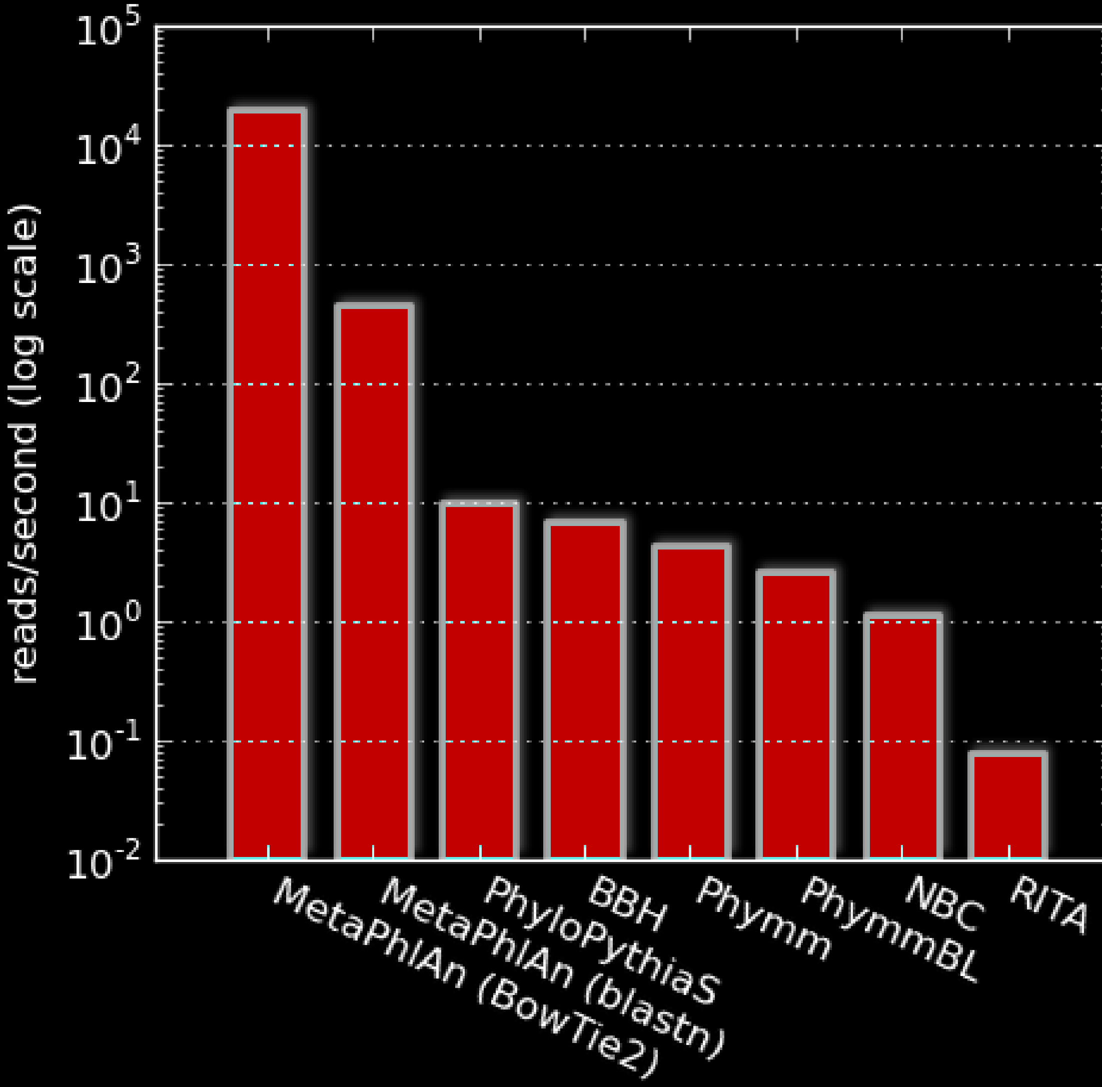
Species



Classes

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

Evaluation of MetaPhlAn performance



>50 times faster than earlier methods

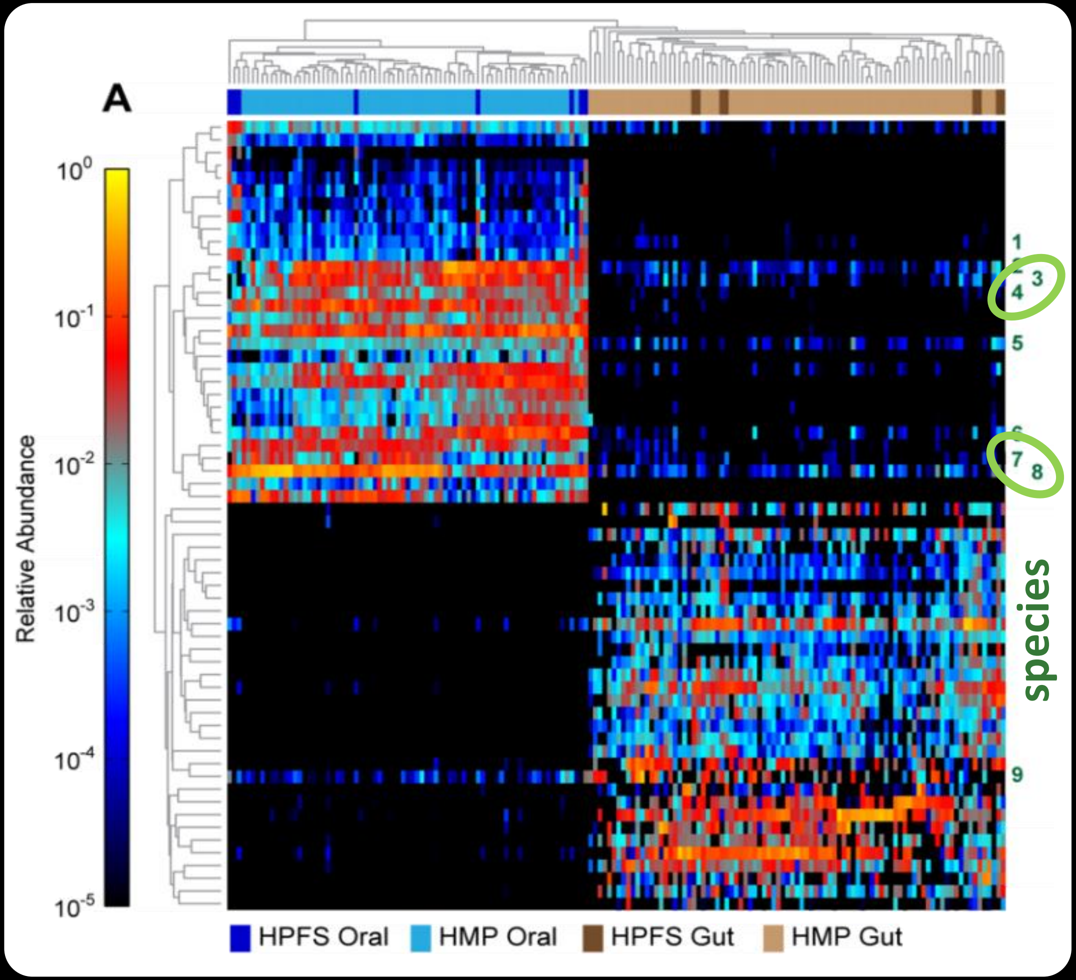
450 reads/sec (BLAST)

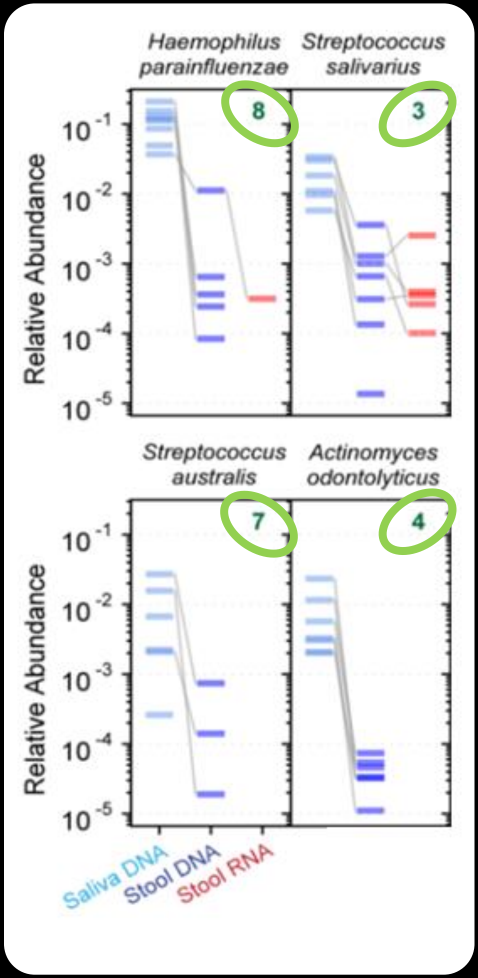
Up to 25,000 reads/sec (bowtie2)

Multi-threaded

Easily parallelizable

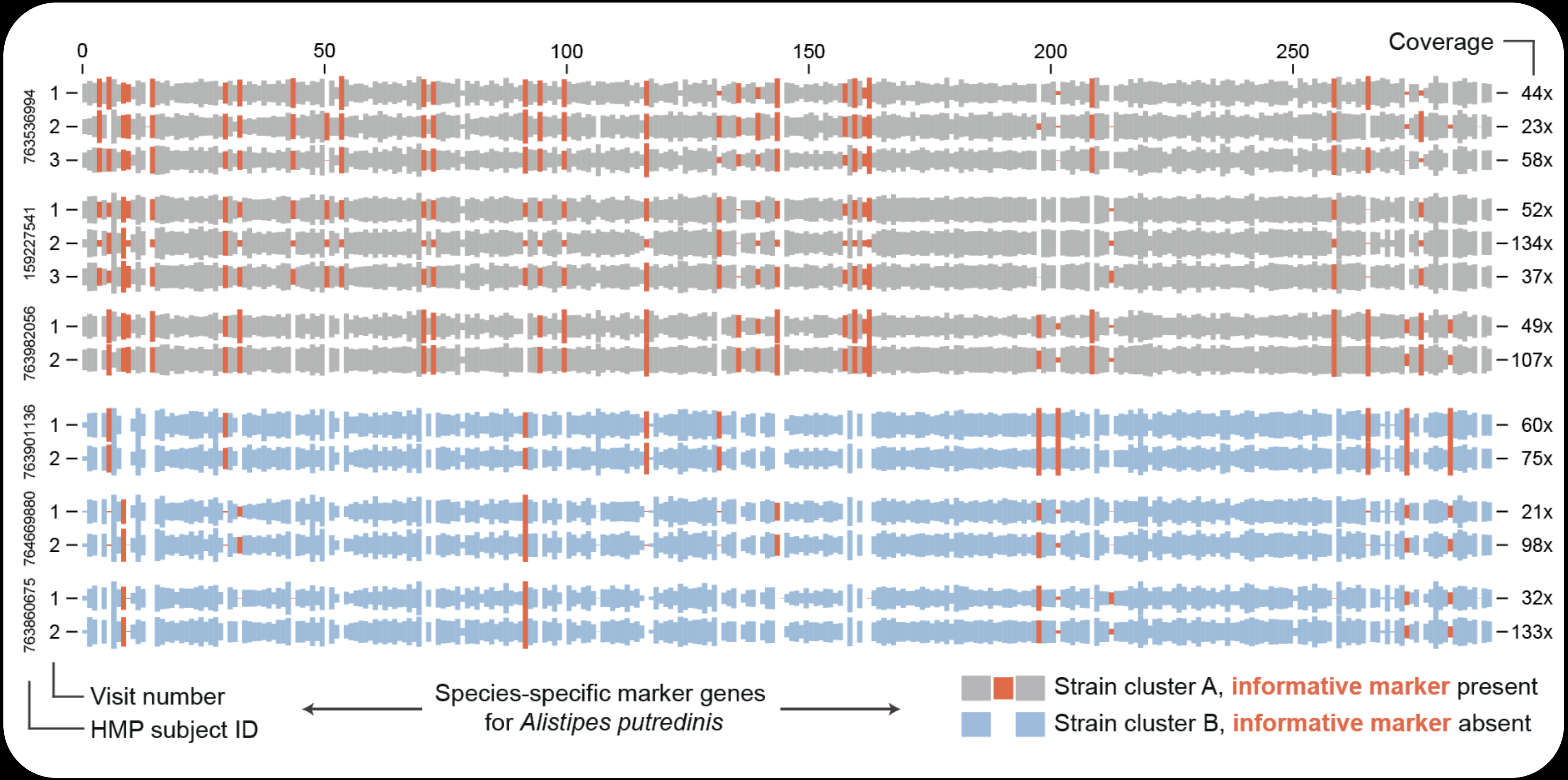
MetaPhlAn in action





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

MetaPhlAn in action: *strain profiling*



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

Plan

- Informal survey
- Metagenomics concepts & examples
- Tools for taxonomic profiling
  - MetaPhlAn
- Tools for functional profiling
  - HUMAnN
  - ShortBRED
  - PICRUSt
- Tools for testing associations
  - LEfSe
  - MaAsLin
  - CCREPE
- Resources
- Research vignette (time permitting)

The two big questions…

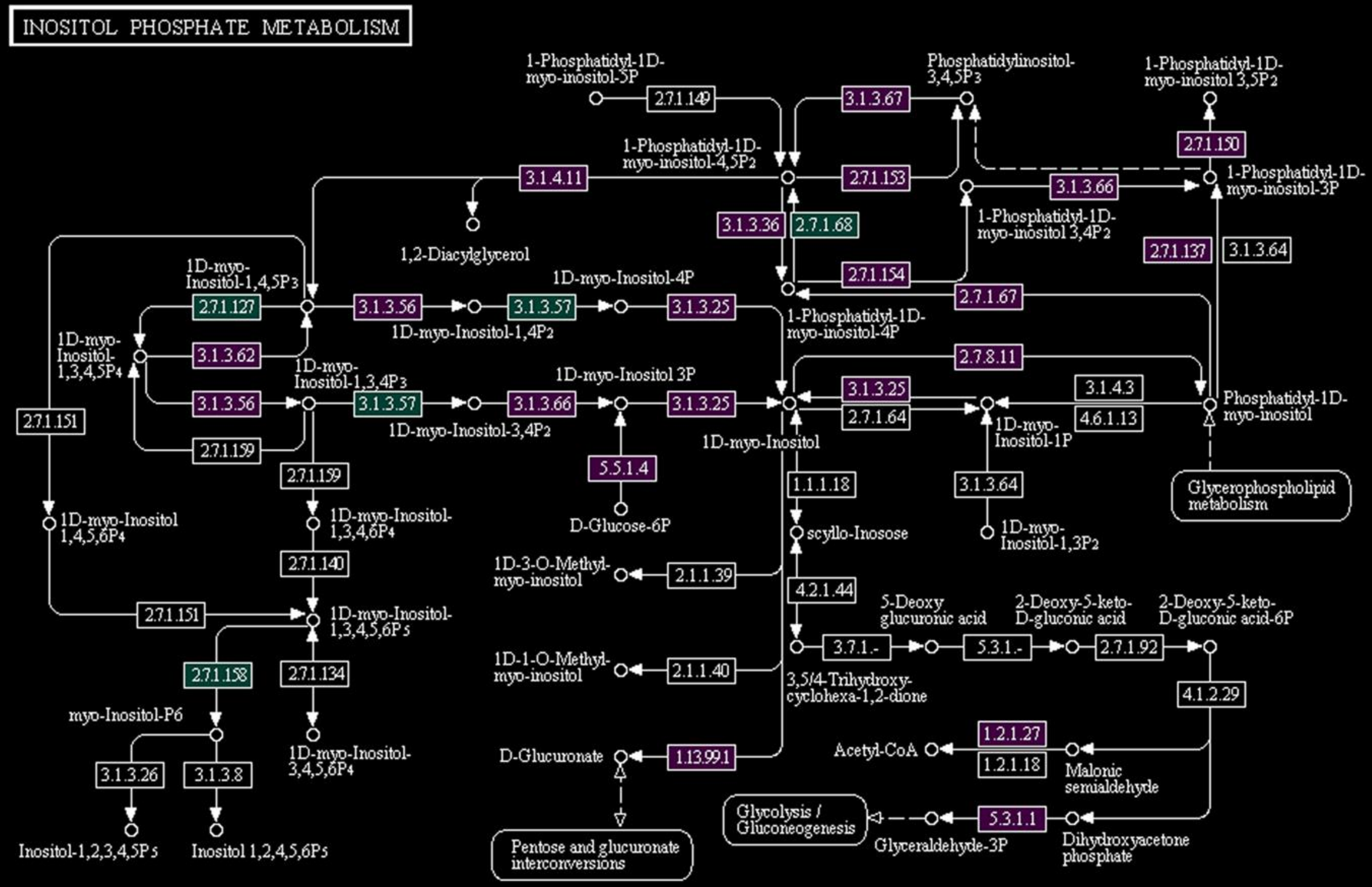
Who is there?

(taxonomic profiling)

What are they doing?

(functional profiling)

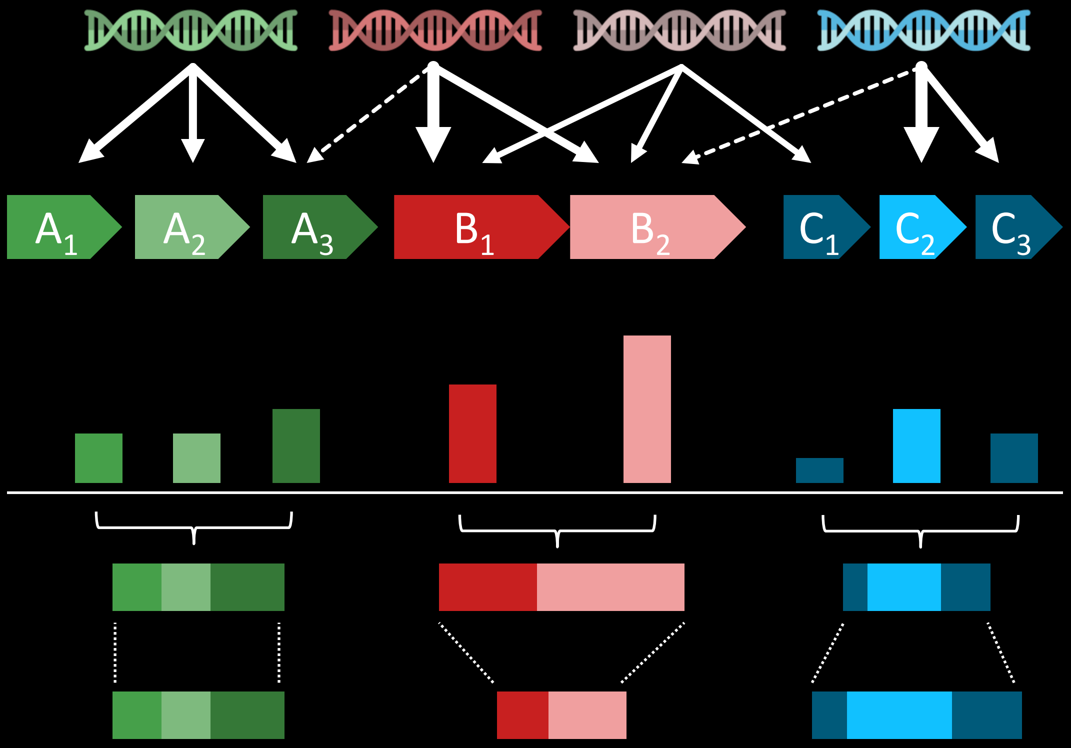
(What we mean by “function”)



00562 11/1/10
(c) Kanehisa Laboratories

HUMAnN

HMP Unified Metabolic Analysis Network



Short reads + protein families

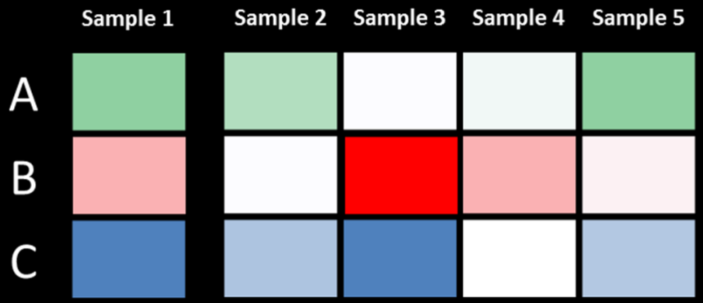
Translated BLAST search

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

Weight hits by significance

Sum over families

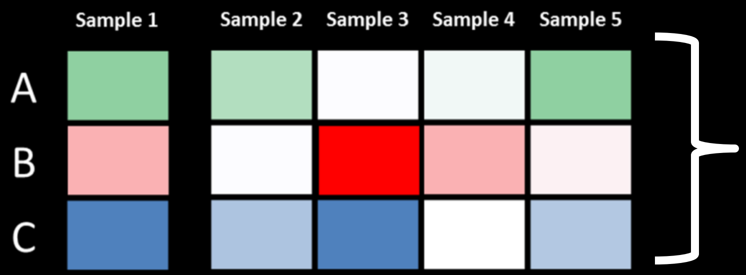
Adjust for sequence length



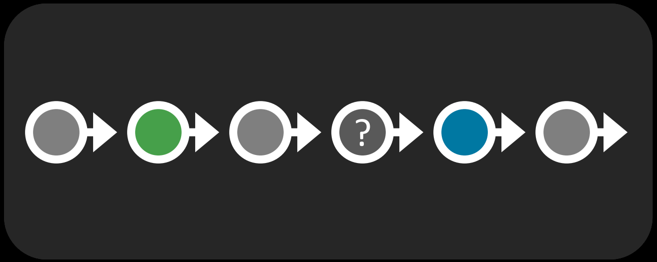
Repeat for each metagenomic or metatranscriptomic sample

HUMAnN

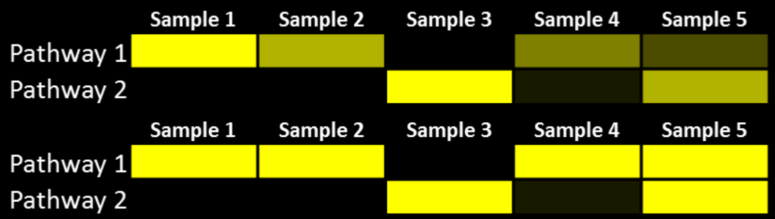
HMP Unified Metabolic Analysis Network



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

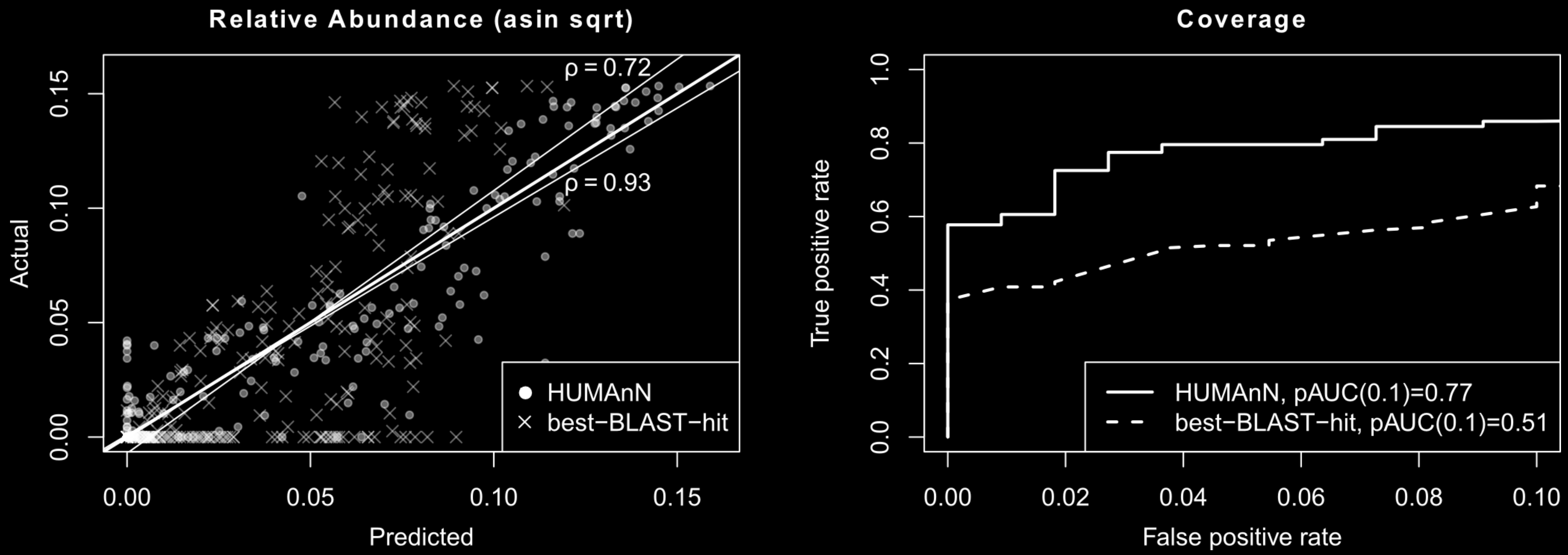


- Map genes to KEGG pathways
- Use MinPath (Ye 2009) to find simplest pathway explanation for observed genes
- Remove pathways unlikely to be present due to low organismal abundance
- Smooth/fill gaps



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

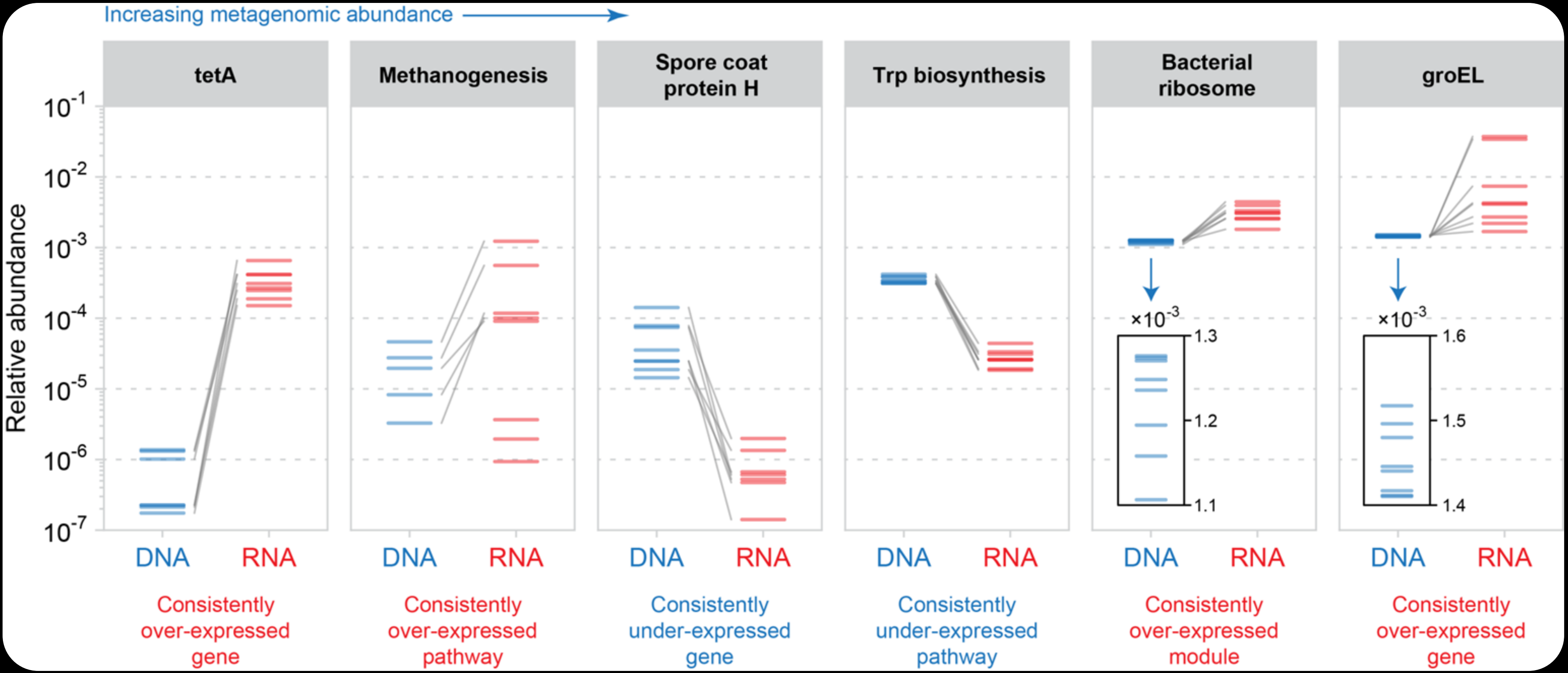
HUMAnN accuracy



Validated against synthetic metagenome samples
(similar to MetaPhlAn validation)

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

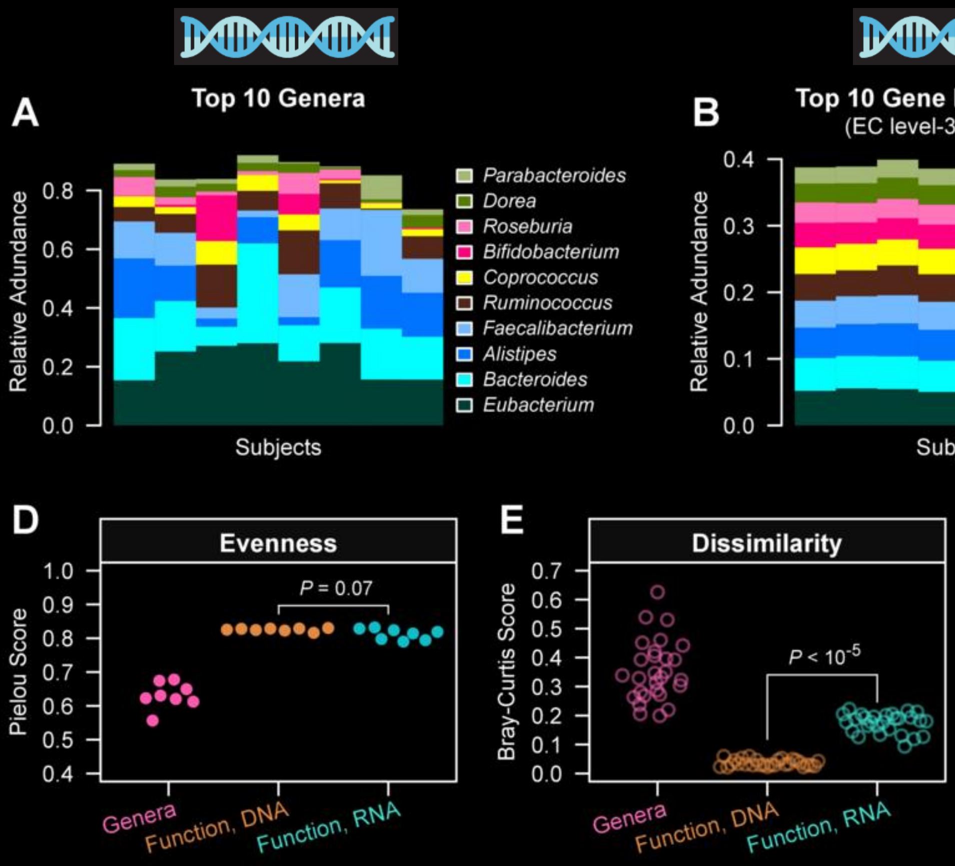
HUMAnN in action
Functional potential (DNA) vs activity (RNA)



Functional metagenomics & metatranscriptomics of 8 heathy human stool samples

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

HUMAnN in action
Conserved potential & variable activity



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)

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

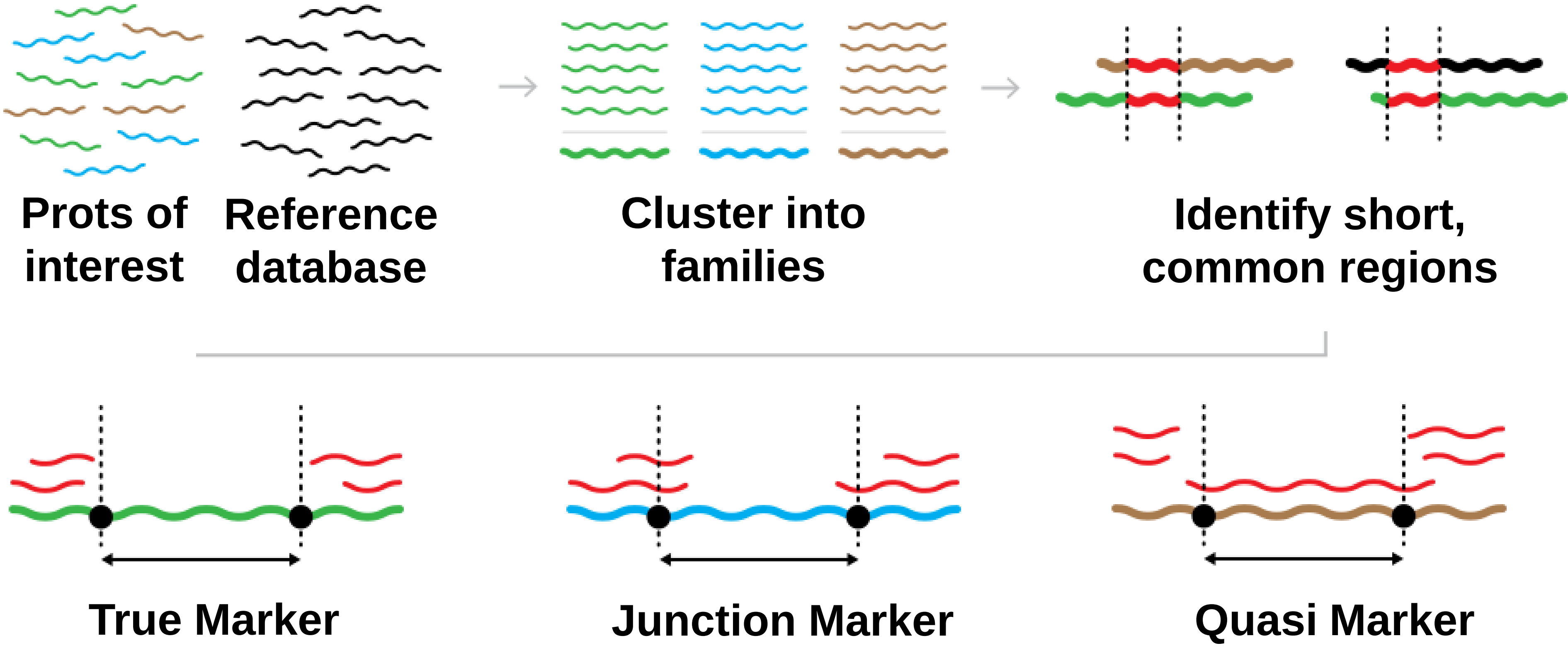
What’s there: ShortBRED algorithm

- Cluster proteins of interest into families
  - Record consensus sequences

- Identify unique 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

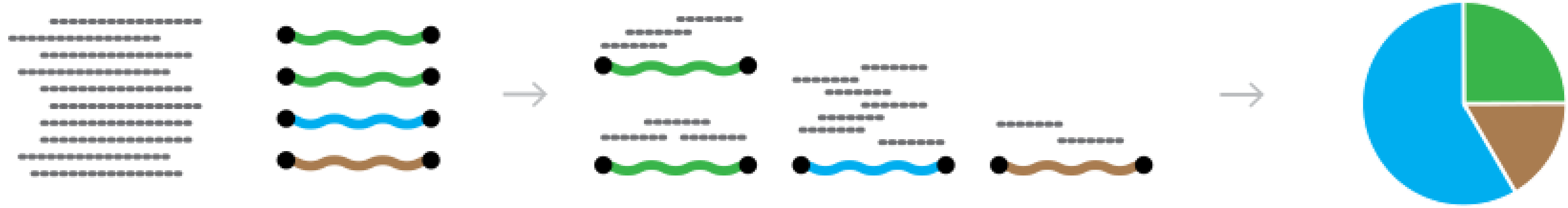
Identify short, common regions

True Marker

Junction Marker

Quasi Marker

What's there: ShortBRED family quantification



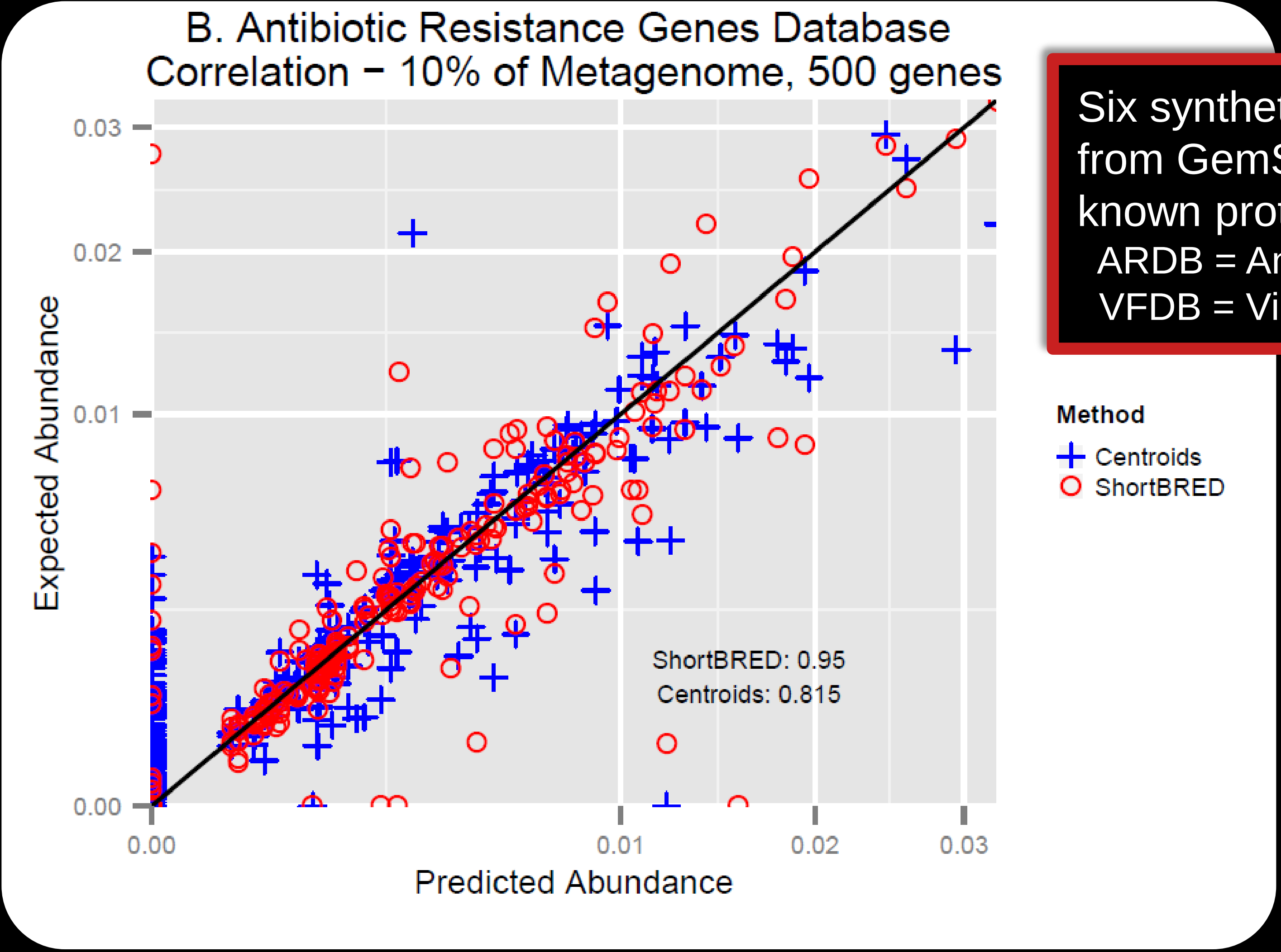
Metagenome reads

ShortBRED markers

Translated search for high ID hits

Normalize relative abundances

What’s there: ShortBRED is accurate

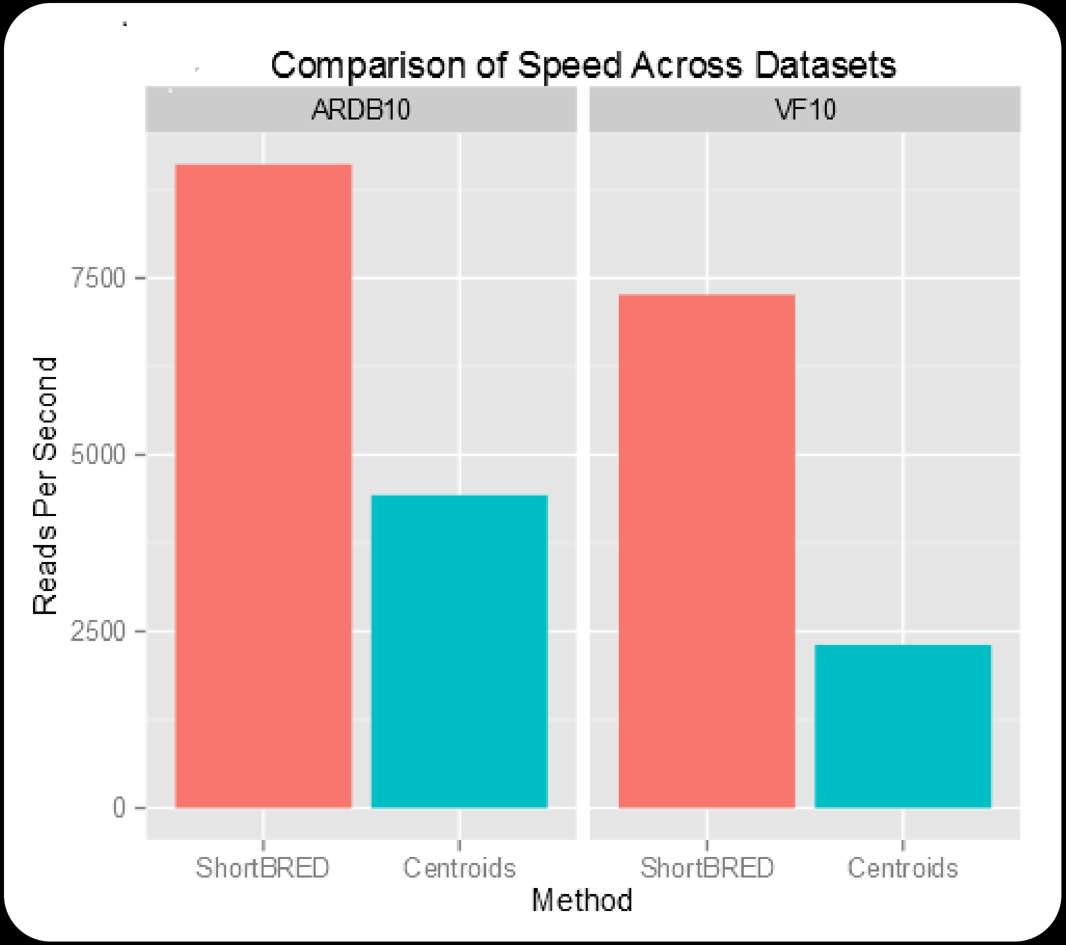


Six synthetic metagenomes from GemSim, spiked with known proteins of interest:

ARDB = Antibiotic Resistance

VFDB = Virulence Factors

What’s there: ShortBRED is fast



Six synthetic metagenomes from GemSim, spiked with known proteins of interest:

ARDB = Antibiotic Resistance

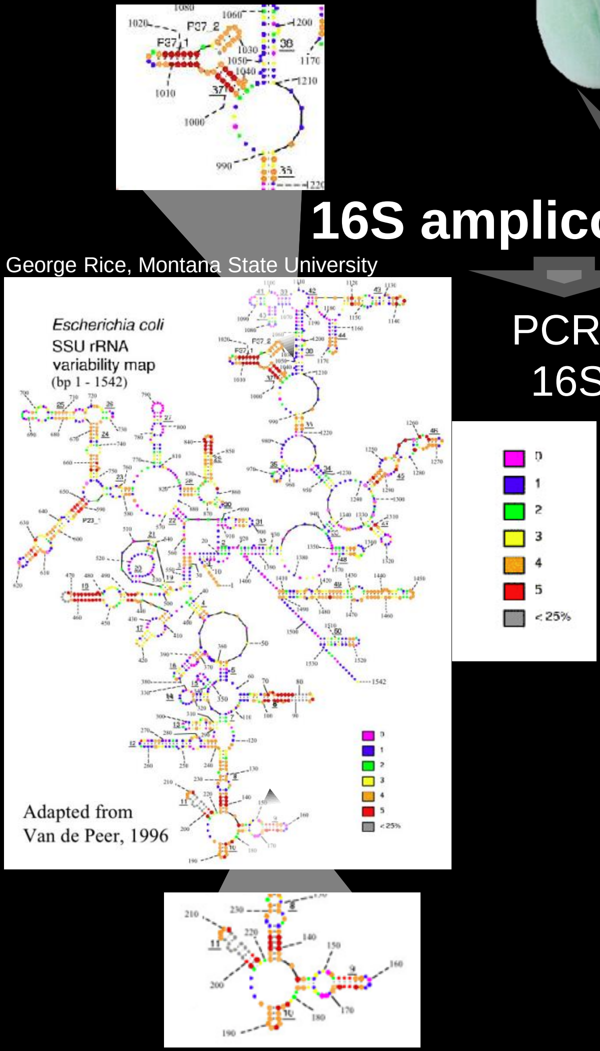
VFDB = Virulence Factors

Can we infer anything about function from 16S data?

Lyse cells
Extract DNA (and/or RNA)

16S amplicons

Meta'omic

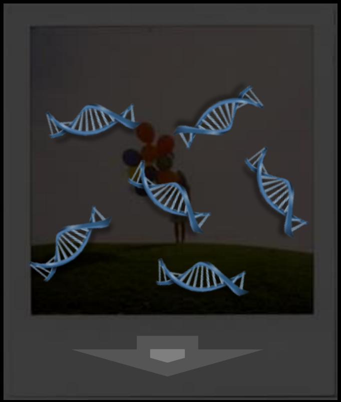


George Rice, Montana State University

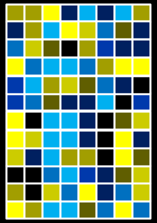
PCR to amplify the single 16S rRNA marker gene

Hello
my name is

Classify sequence
➔ microbe



Samples



Relative abundances

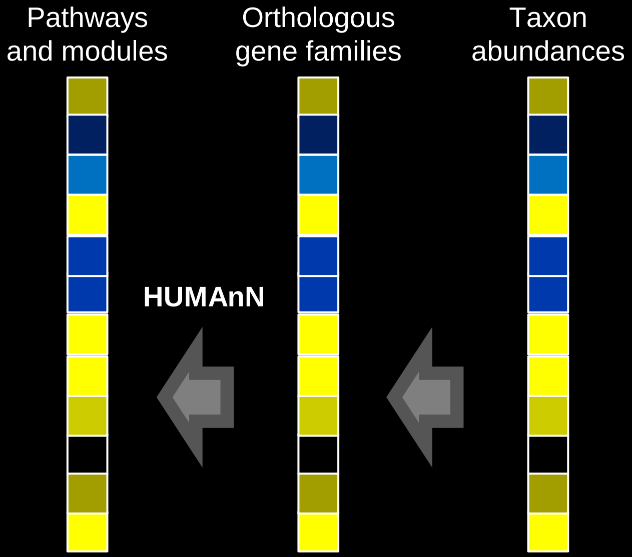
Genes,
Genomes,
Metabolic profiling,
Relative abundances,
Genetic variants...

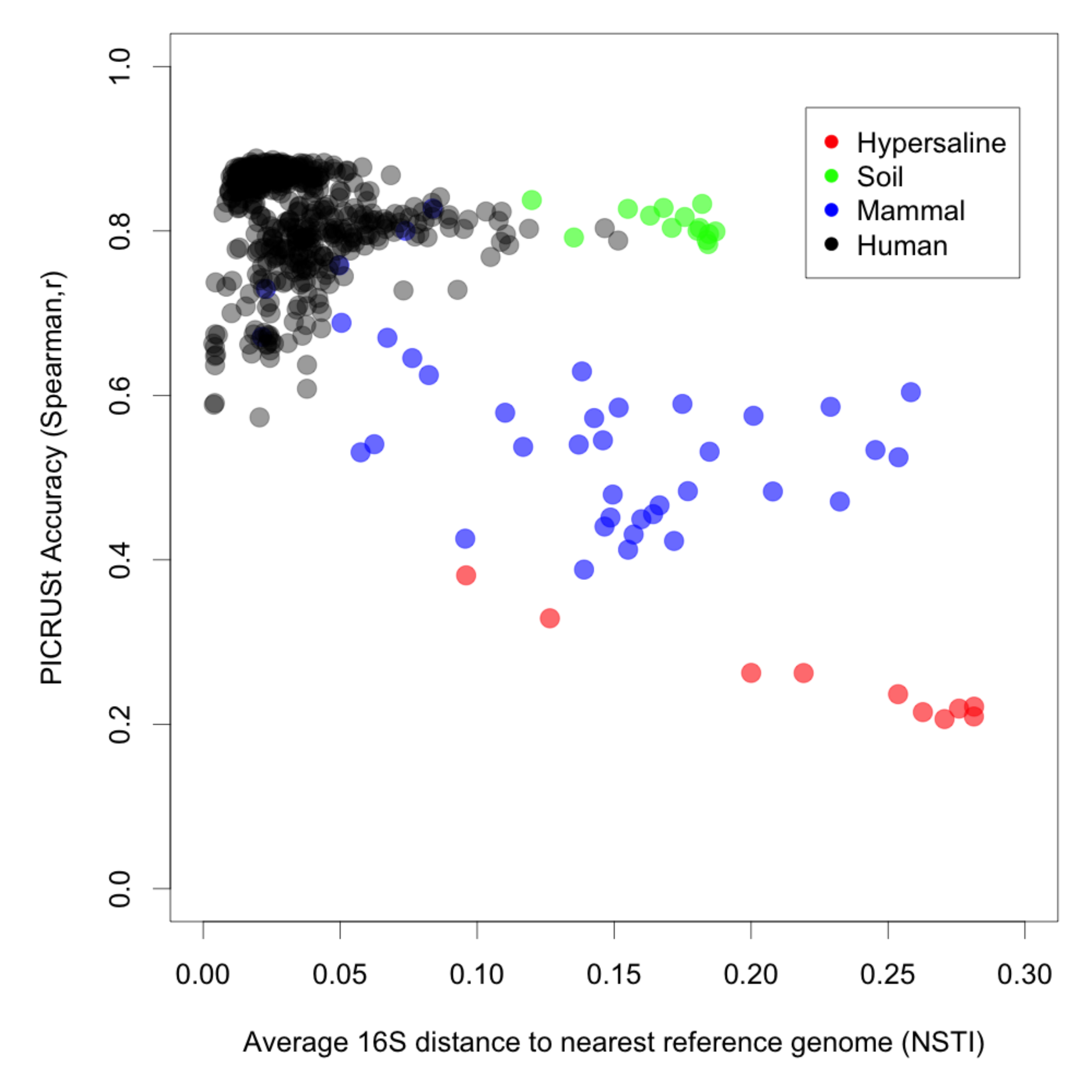
PICRUSt: Inferring community metagenomic potential from marker gene sequencing

With Rob Knight, Rob Beiko

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

http://picrust.github.com





Relative abundance

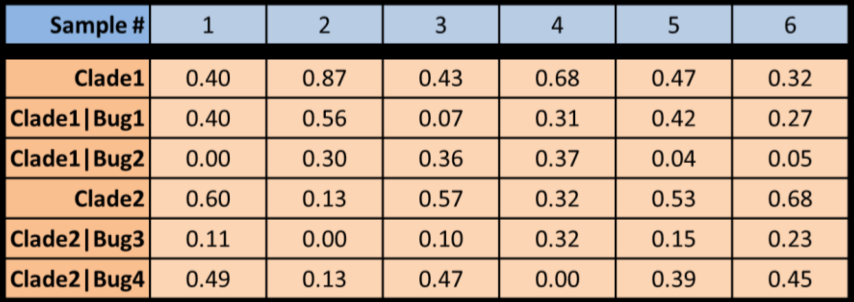
Plan

- Informal survey
- Metagenomics concepts & examples
- Tools for taxonomic profiling
  - MetaPhlAn
- Tools for functional profiling
  - HUMAnN
  - ShortBRED
  - PICRUSt
- Tools for testing associations
  - LEfSe
  - MaAsLin
  - CCREPE
- Resources
- Research vignette (time permitting)

The two big questions…

Who is there?

What are they doing?



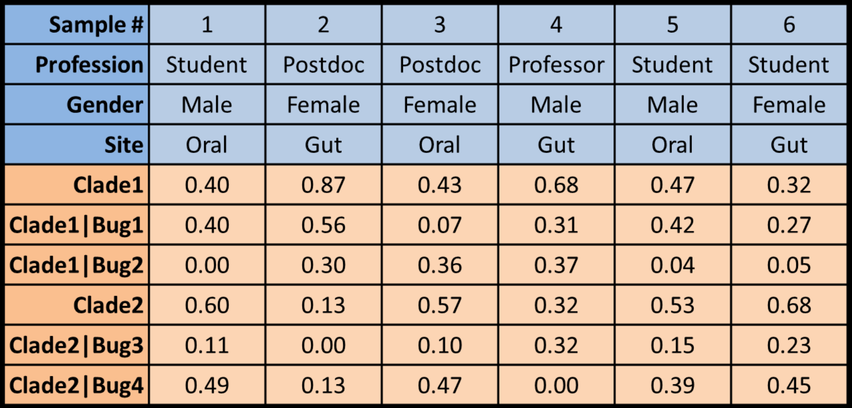
Sample #	1	2	3	4	5	6
Clade1	0.40	0.87	0.43	0.68	0.47	0.32
Clade1|Bug1	0.40	0.56	0.07	0.31	0.42	0.27
Clade1|Bug2	0.00	0.30	0.36	0.37	0.04	0.05
Clade2	0.60	0.13	0.57	0.32	0.53	0.68
Clade2|Bug3	0.11	0.00	0.10	0.32	0.15	0.23
Clade2|Bug4	0.49	0.13	0.47	0.00	0.39	0.45

The ~~two~~ three big questions…

Who is there?

What are they doing?

What does it all mean?



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

Properties of microbiome data

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

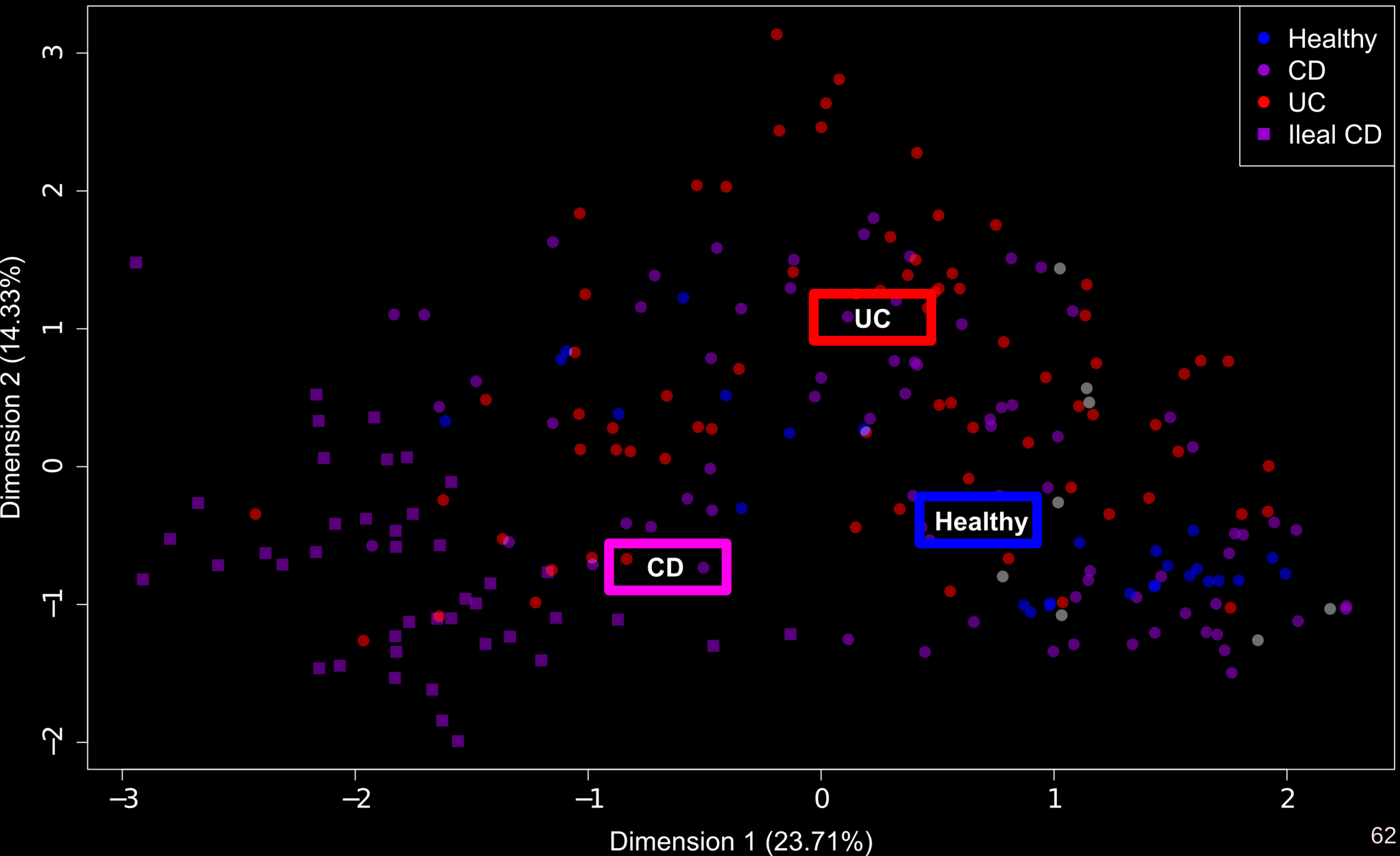
Site	Oral	Gut	Oral	Gut	Oral	Gut
Clade1	0.40	0.87	0.43	0.68	0.47	0.32
Clade1|Bug1	0.40	0.56	0.07	0.31	0.42	0.27
Clade1|Bug2	0.00	0.30	0.36	0.37	0.04	0.05
Clade2	0.60	0.13	0.57	0.32	0.53	0.68
Clade2|Bug3	0.11	0.00	0.10	0.32	0.15	0.23
Clade2|Bug4	0.49	0.13	0.47	0.00	0.39	0.45

Properties of microbiome data

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

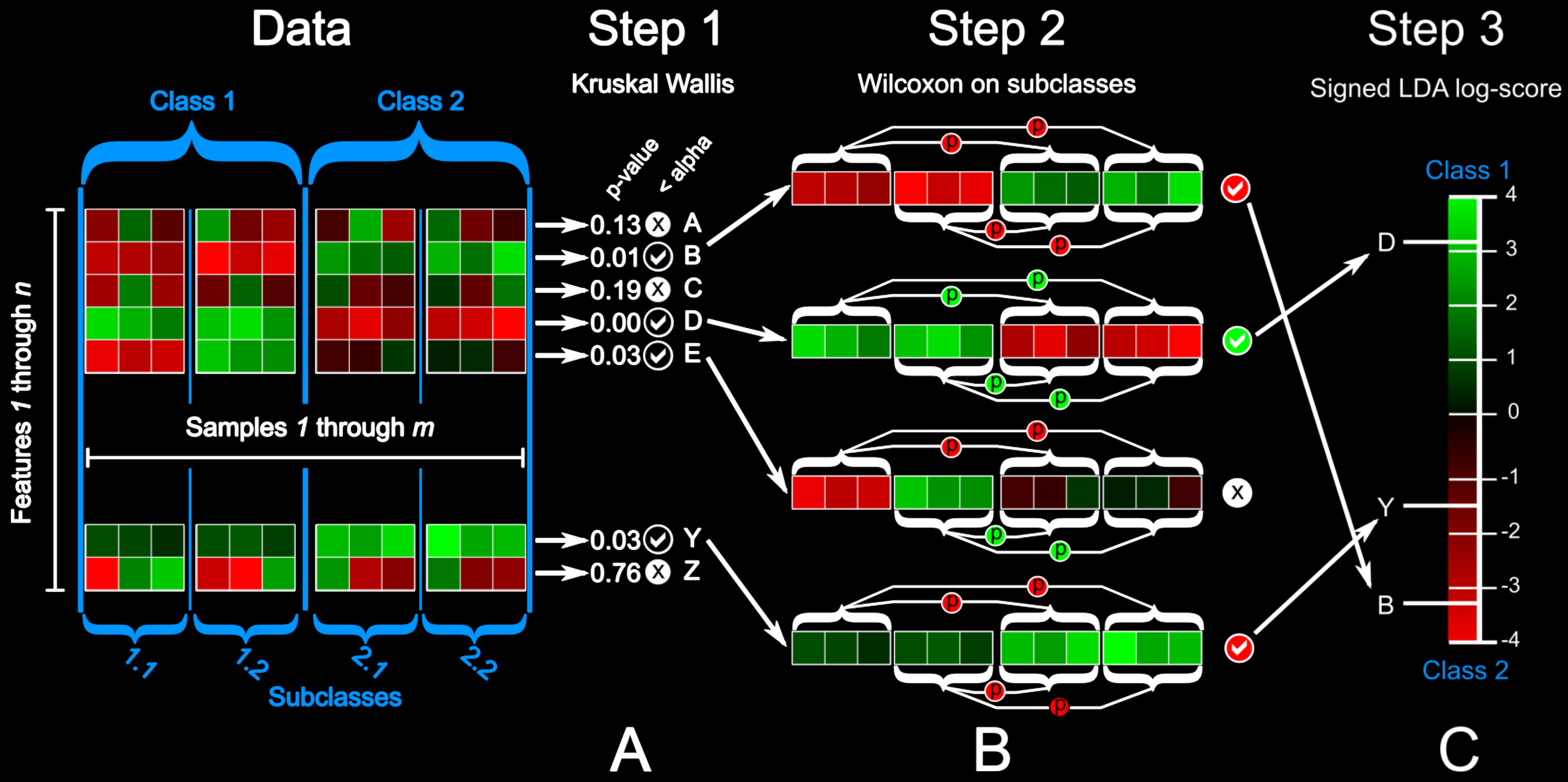
Sample #	1	2	3	4	5	6
Profession	Student	Postdoc	Postdoc	Professor	Student	Student
Gender	Male	Female	Female	Male	Male	Female
Site	Oral	Gut	Oral	Gut	Oral	Gut
Clade1	0.40	0.87	0.43	0.68	0.47	0.32
Clade1|Bug1	0.40	0.56	0.07	0.31	0.42	0.27
Clade1|Bug2	0.00	0.30	0.36	0.37	0.04	0.05
Clade2	0.60	0.13	0.57	0.32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

Recall that ordination is exploratory (no *p*-values for a trend, for example)



LEfSe: LDA Effect Size
Finding metagenomic biomarkers

Nicola Segata

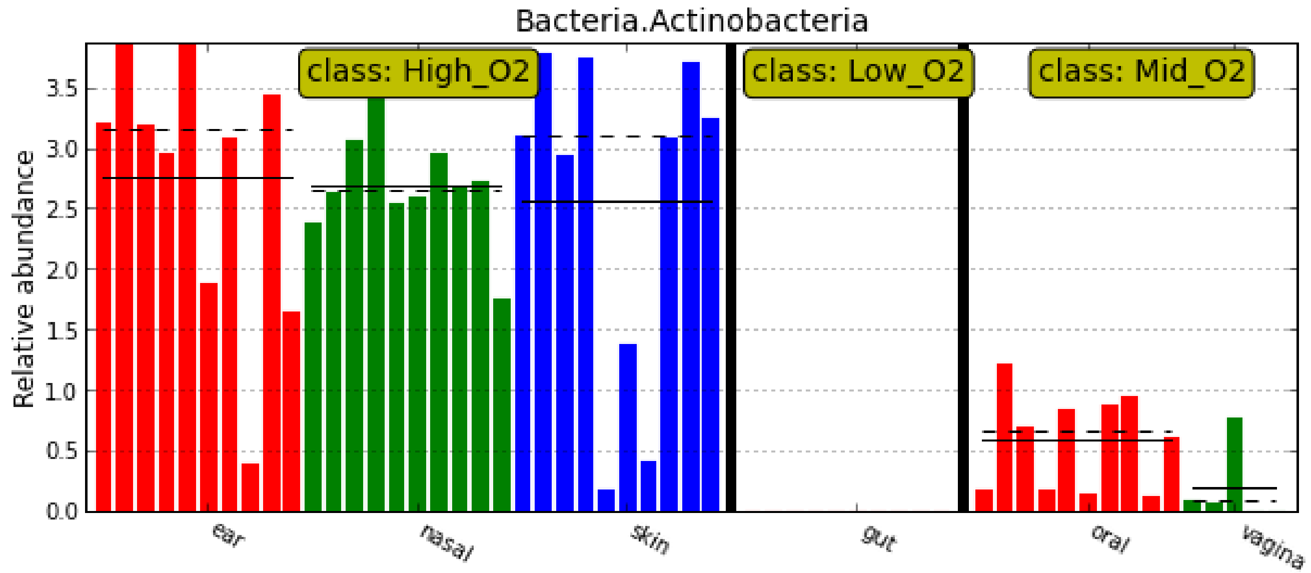


Statistical consistency

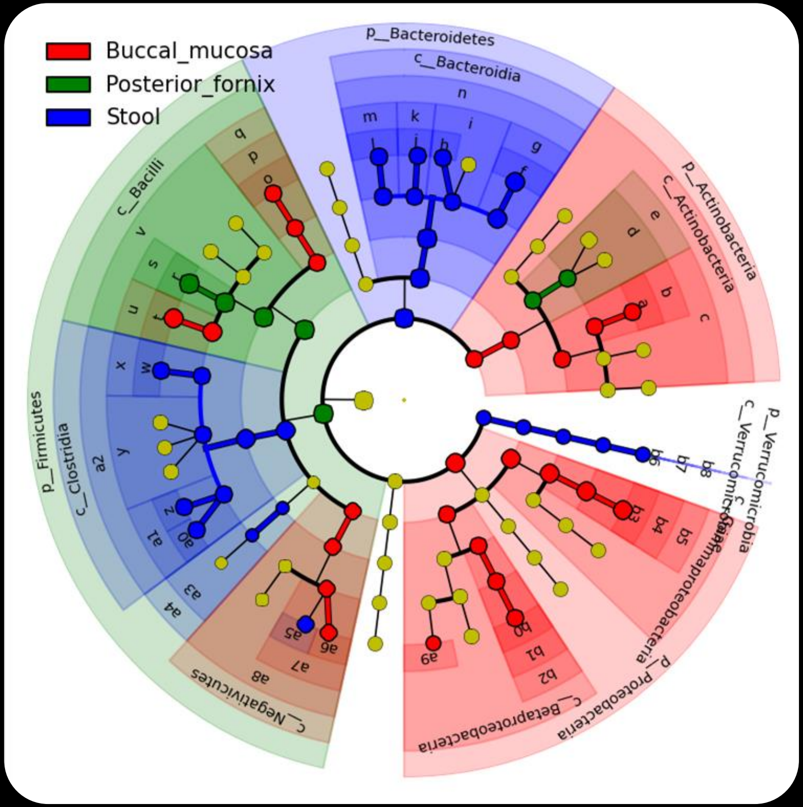
Biological consistency

Biological Effect size

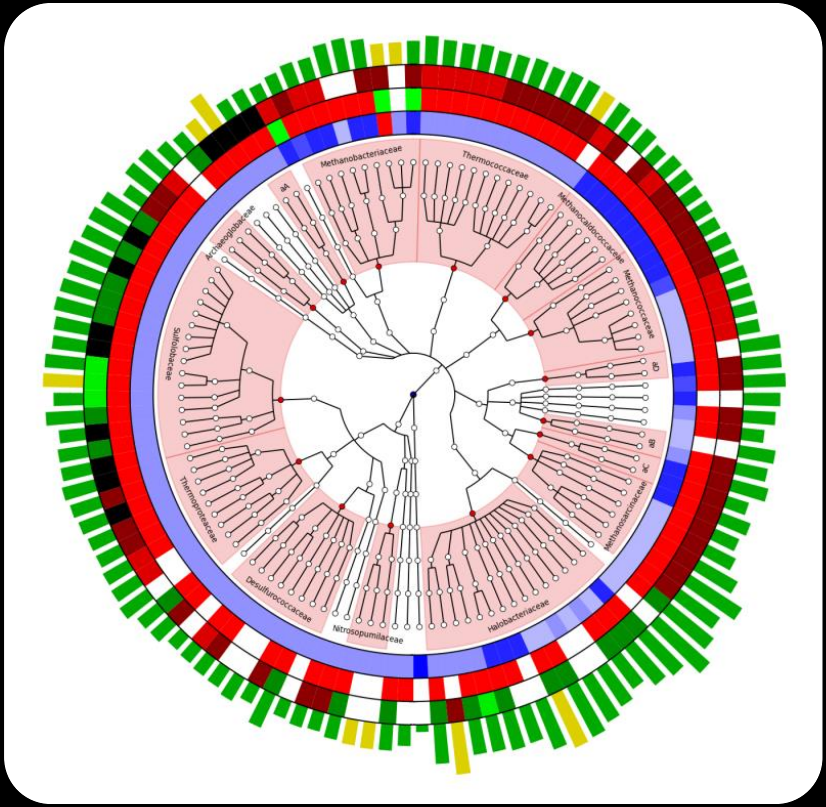
Example LEfSe application: Find O_2-loving bugs (controlling for body site)



Superimpose enrichments on the tree of life using GraPhlAn

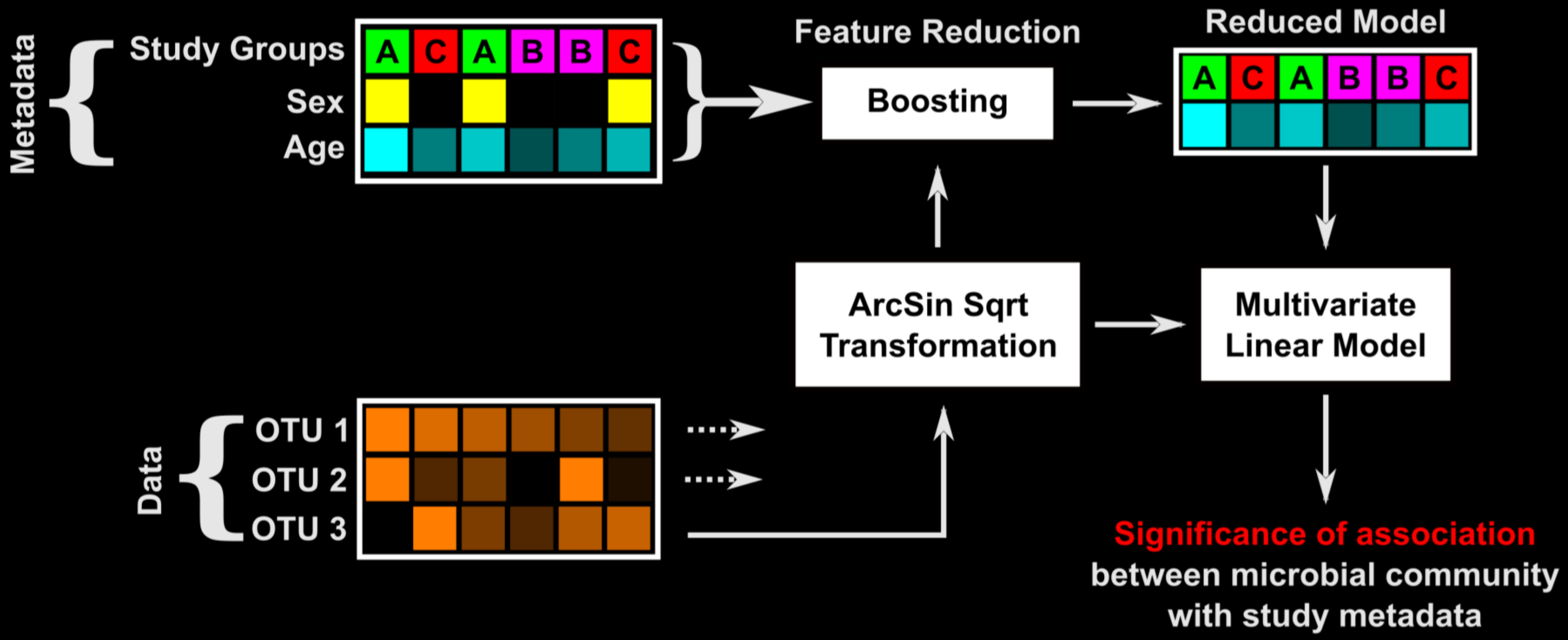


LEfSe Associations



Metadata Rings

MaAsLin
Multivariate Association with Linear Models



Metadata

Study Groups

Sex

Age

A C A B B C

Feature Reduction

Boosting

Reduced Model

A C A B B C

ArcSin Sqrt Transformation

Multivariate Linear Model

Data

OTU 1

OTU 2

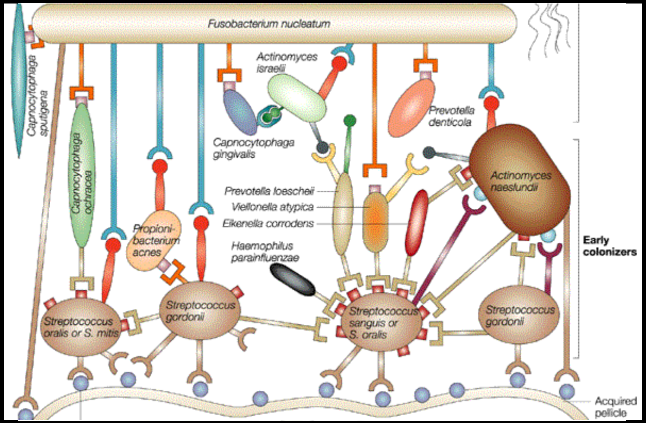
OTU 3

Significance of association between microbial community with study metadata

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

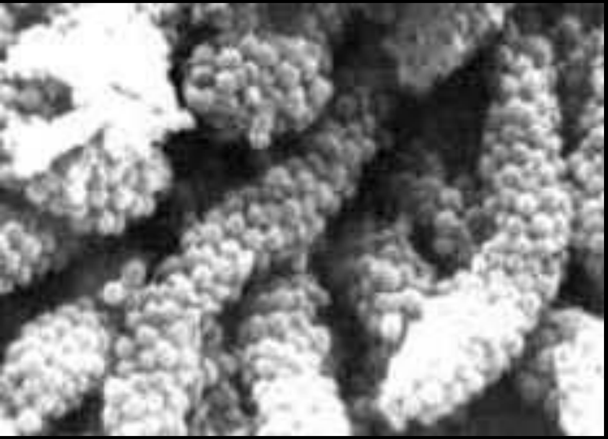
Tim Tickle

Microbiome downstream analyses: interaction network reconstruction



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

Mutualism



Predation

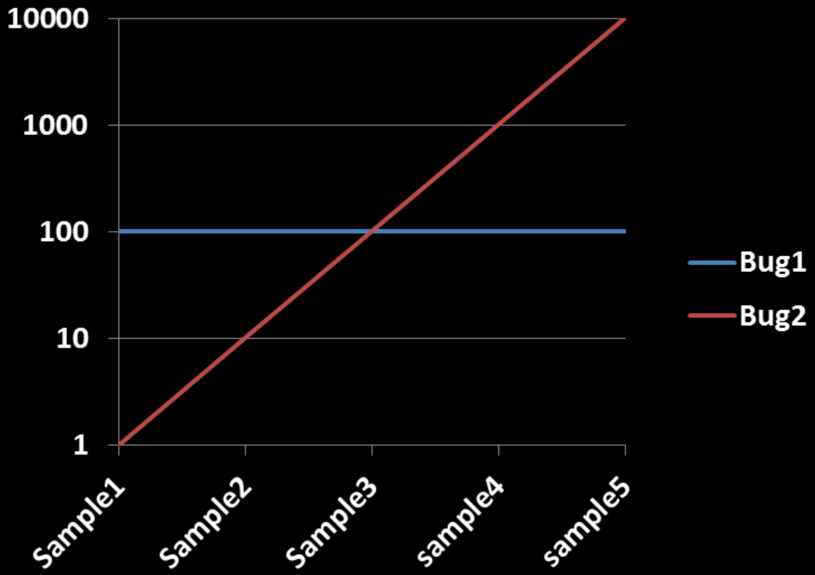
Competition

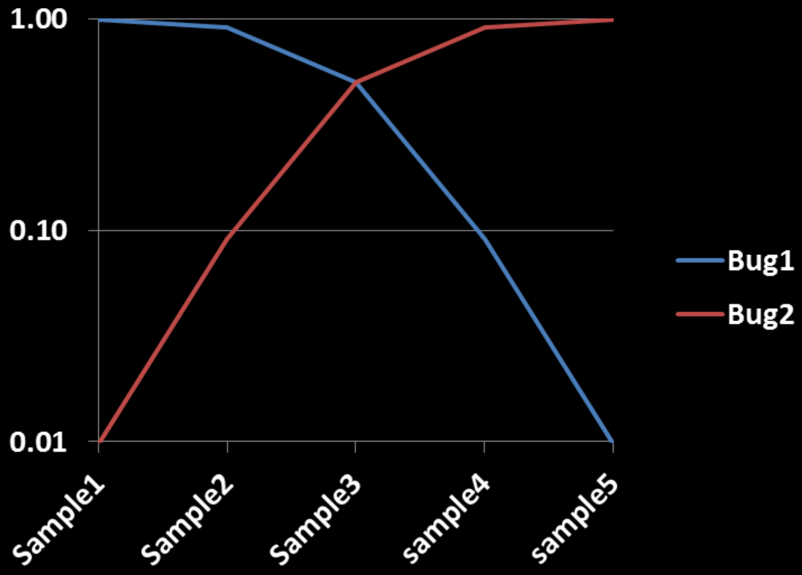
Given microbial relative abundance measurements over many samples, can we detect *co-occurrence and co-exclusion relationships?*

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

	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





Absolute (cell) counts

No bug1-bug2 correlation

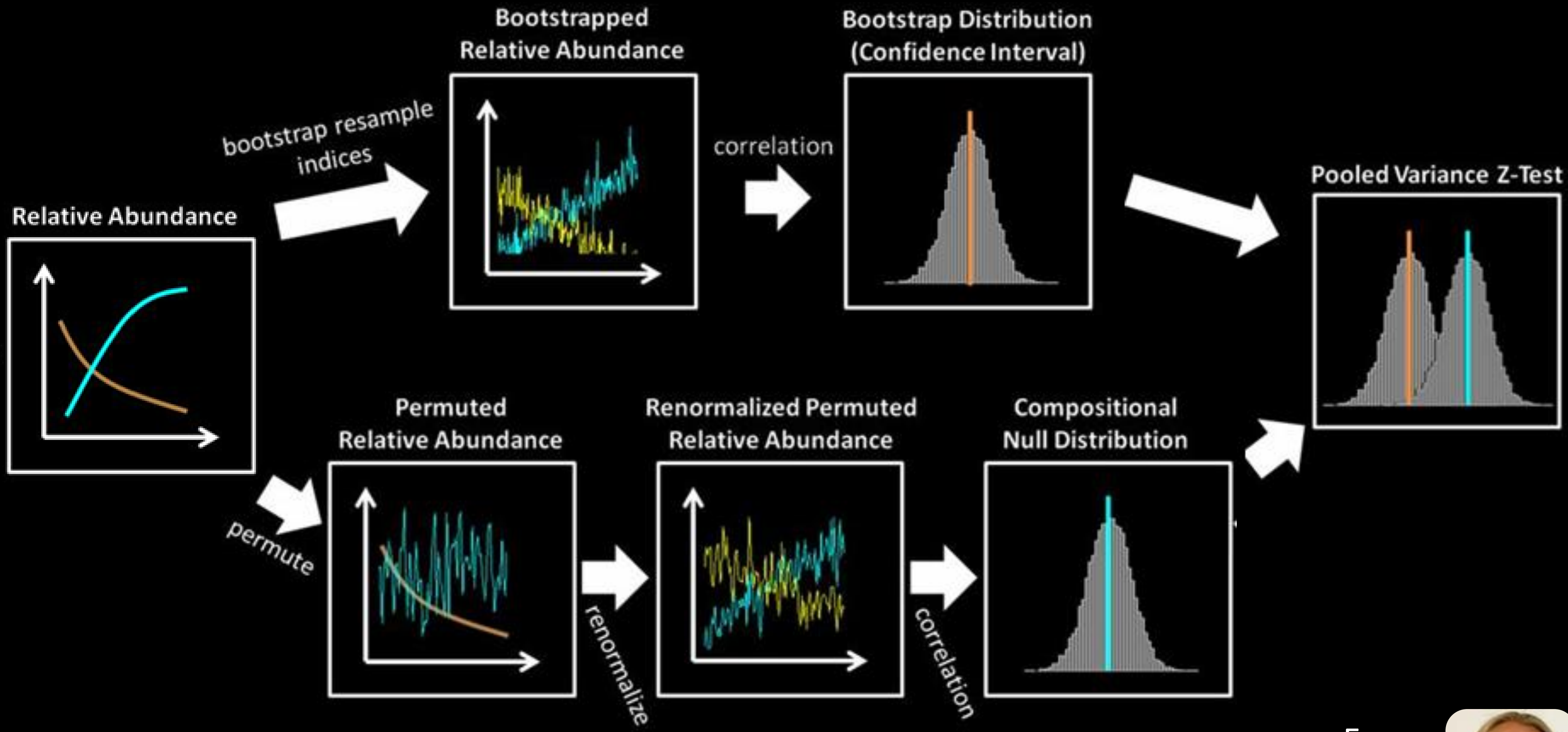
Relative abundance

Spurious bug1-bug2 correlation

(sequencing yields rel. ab.)

CCREPE: Compositionality Corrected by REnormalization and PErmutation

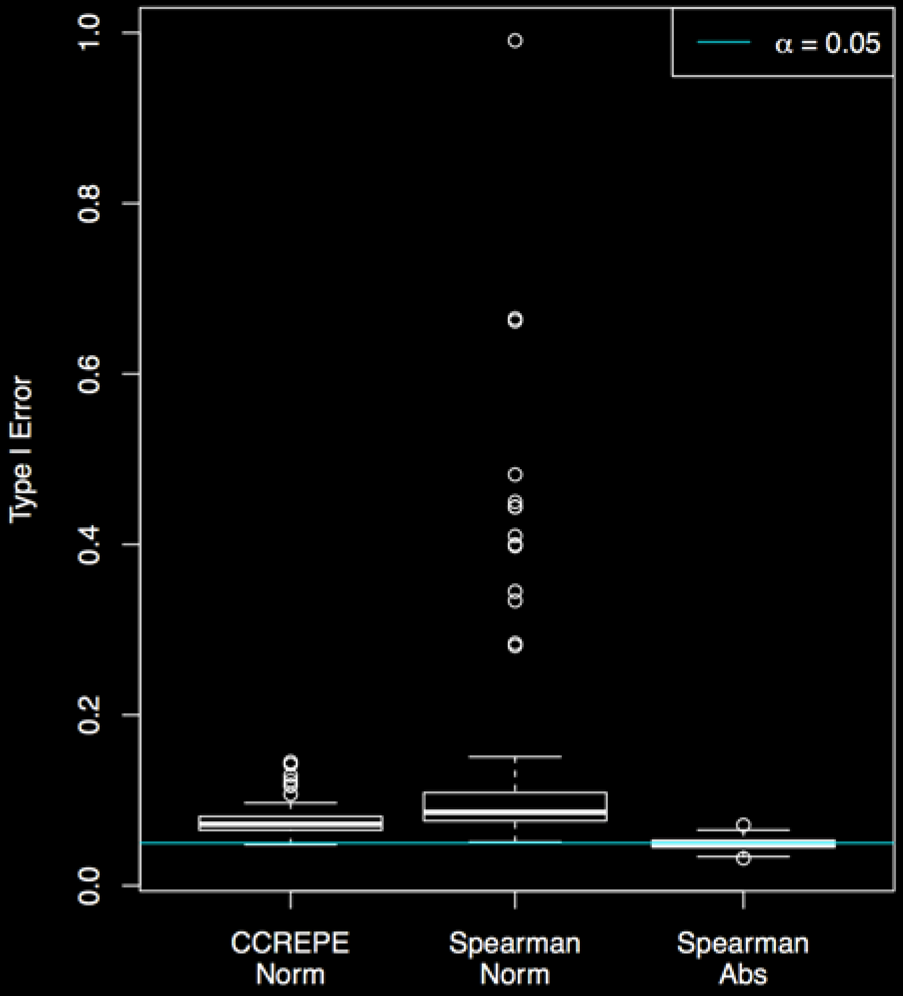
Estimating a confidence interval



Estimating the null distribution

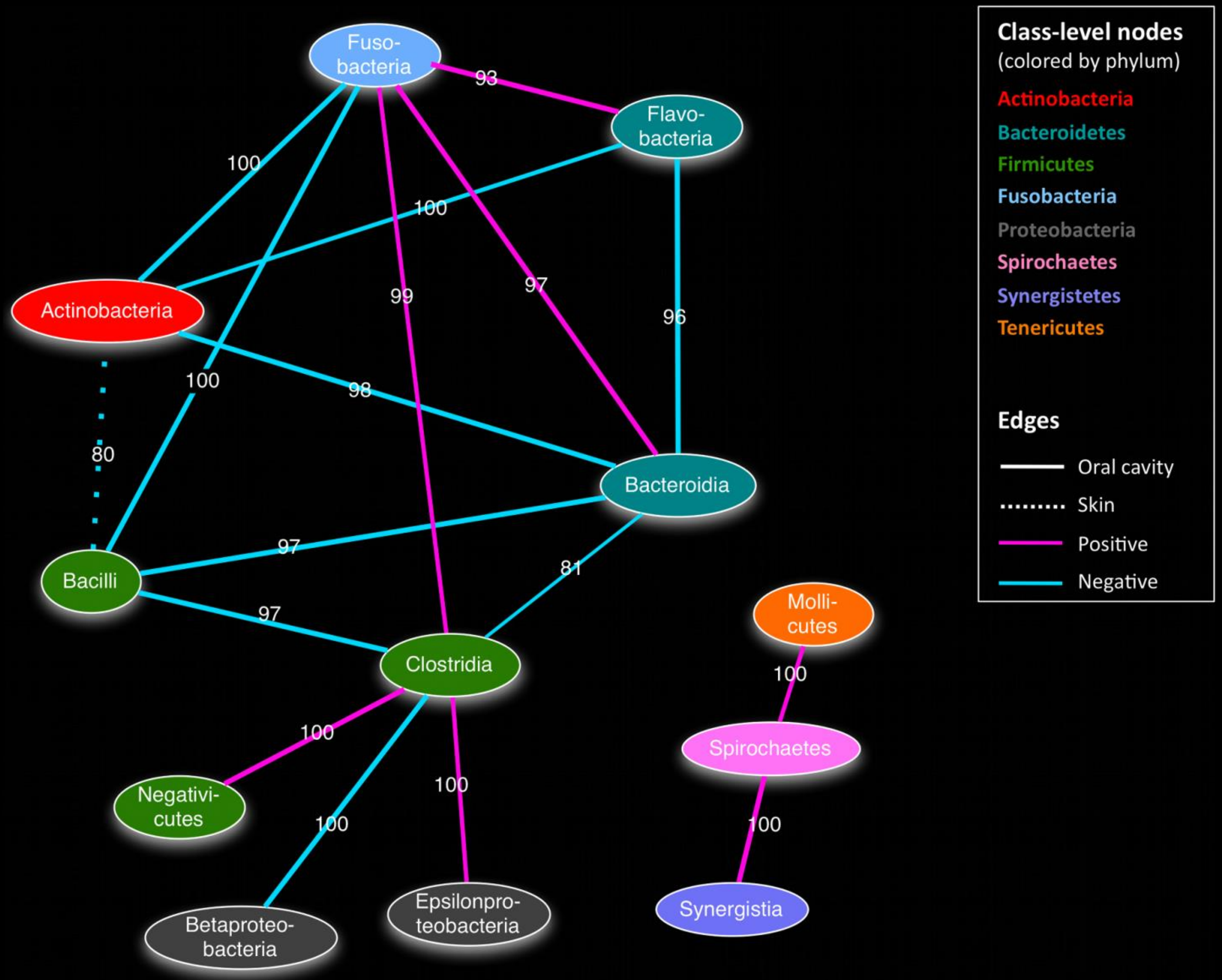
Emma
Schwager

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 <u>three</u> big questions…

Who is there?

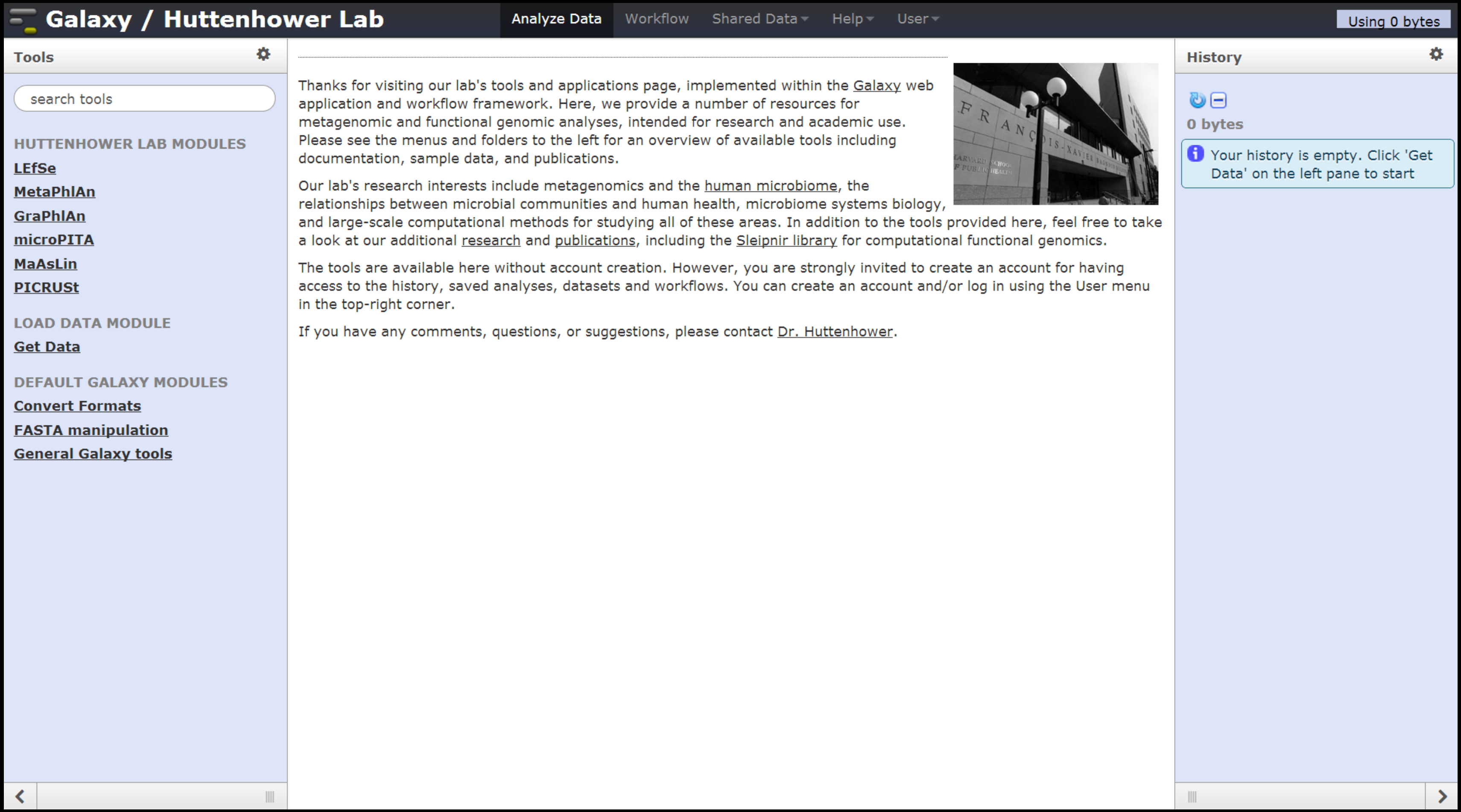
What are they doing?

What does it all mean?

Plan

- Informal survey
- Metagenomics concepts & examples
- Tools for taxonomic profiling
  - MetaPhlAn
- Tools for functional profiling
  - HUMAnN
  - ShortBRED
  - PICRUSt
- Tools for testing associations
  - LEfSe
  - MaAsLin
  - CCREPE
- Resources
- Research vignette (time permitting)

Using tools through Galaxy



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

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

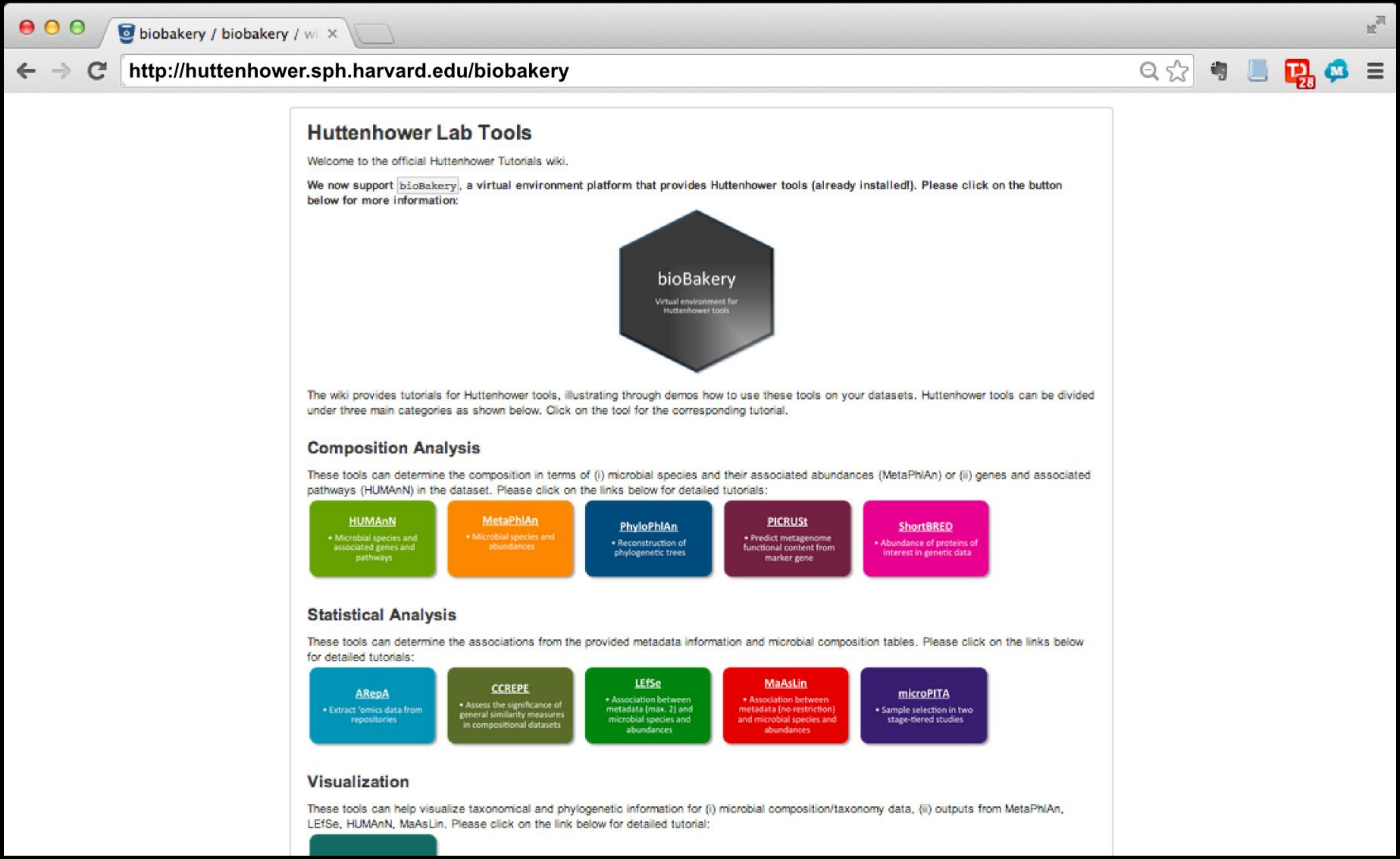
The tools are available here without account creation. However, you are strongly invited to create an account for having access to the history, saved analyses, datasets and workflows. You can create an account and/or log in using the User menu in the top-right corner.

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

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

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

Tutorials available online



biobakery / biobakery / wi ×
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:

bioBakery
Virtual environment for Huttenhower tools

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:

- HUMAnN: Microbial species and associated genes and pathways
- MetaPhlAn: Microbial species and abundances
- PhyloPhlAn: 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

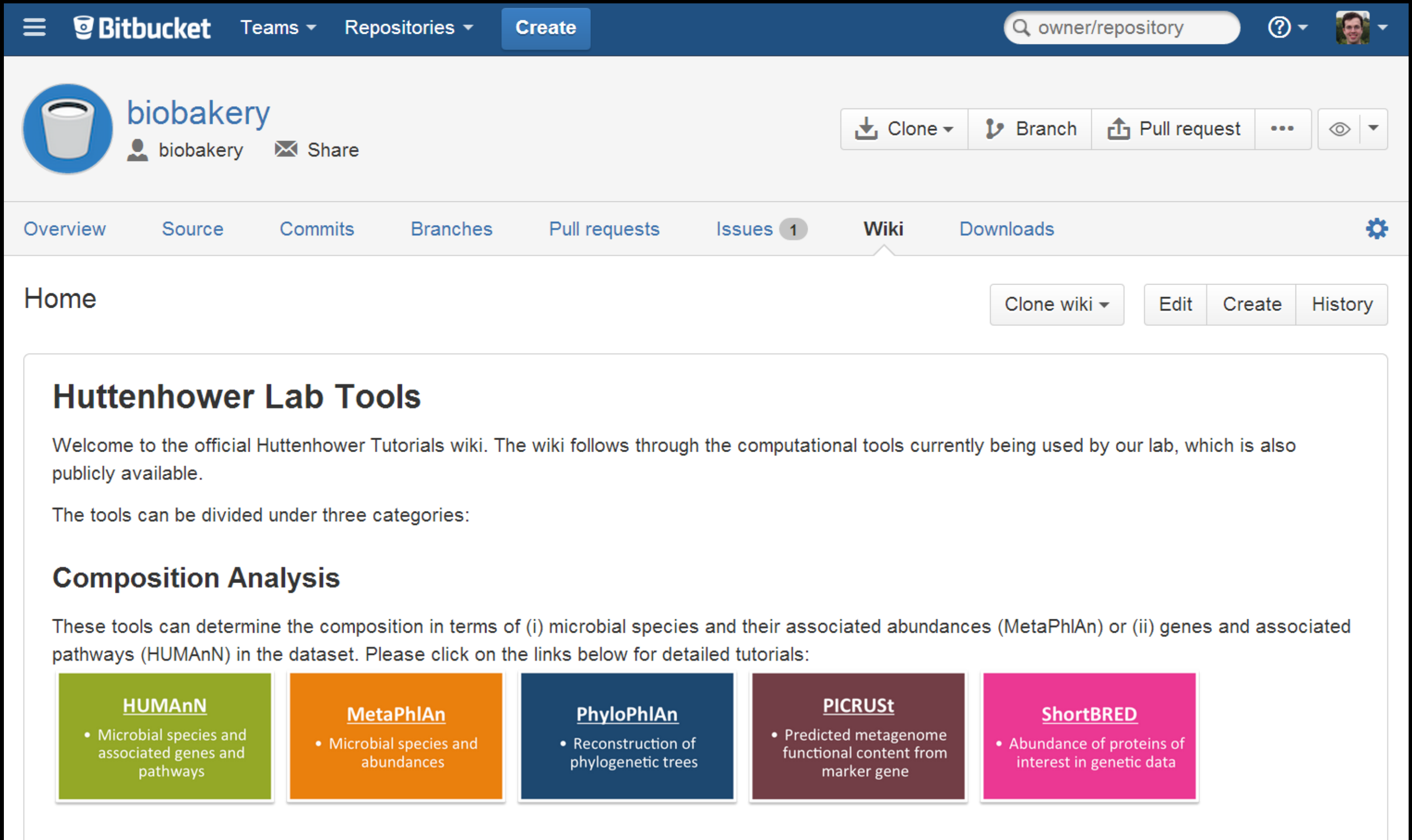
Visualization

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

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

(click on your tool-of-interest)

All tools are open source



Huttenhower Lab Tools

Welcome to the official Huttenhower Tutorials wiki. The wiki follows through the computational tools currently being used by our lab, which is also publicly available.

The tools can be divided under three categories:

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:

HUMAnN
- Microbial species and associated genes and pathways

MetaPhlAn
- Microbial species and abundances

PhyloPhlAn
- Reconstruction of phylogenetic trees

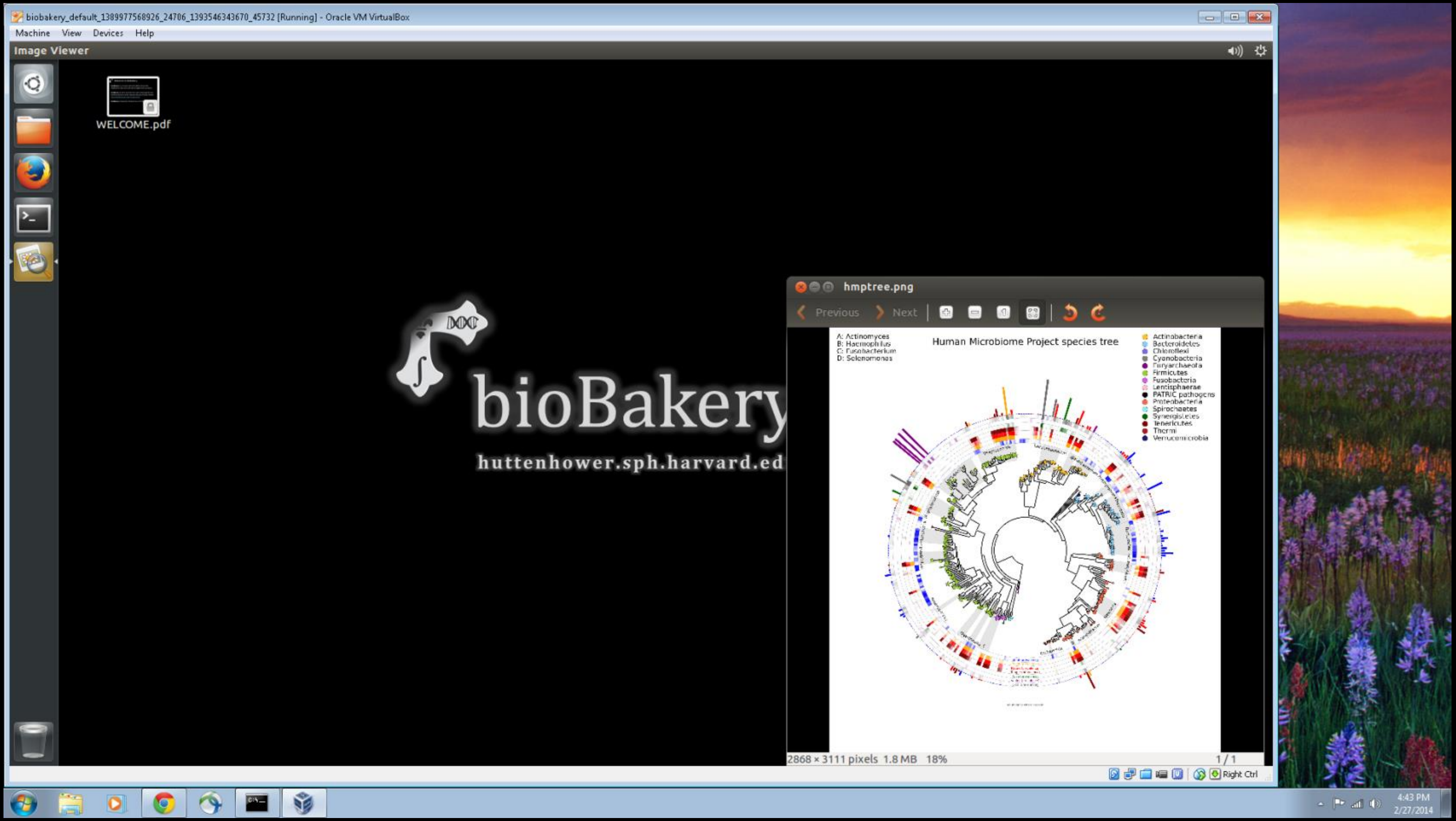
PICRUSt
- Predicted metagenome functional content from marker gene

ShortBRED
- Abundance of proteins of interest in genetic data

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

Afrah Shafquat

Keith Bayer

George Weingart

Regina Joice

Aleksandar Kostic

Chengwei Luo

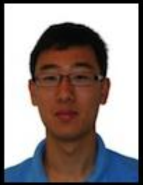
Tiffany Hsu

Emma Schwager

Koji Yasuda

Kevin Oh

Boyu Ren


Andy Shi

Jim Kaminski

Joseph Moon

Randall Schwager

Levi Waldron

Nicola Segata

Wendy Garrett
Michelle Rooks

Ruth Ley
Omry Koren

Dirk Gevers
Kat Huang

Ramnik Xavier
Harry Sokol
Dan Knights
Moran Yassour

Human Microbiome Project

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

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

Bruce Birren
Doyle Ward
Mark Daly
Ashlee Earl

BROAD INSTITUTE

Rob Beiko
Morgan Langille

Rob Knight
Greg Caporaso
Jesse Zaneveld

Mark Silverberg
Boyko Kabakchiev
Andrea Tyler

Jacques Izard
Katherine Lemon

Bruce Sands

Tutorial

- Informal survey
- Metagenomics concepts & examples
- Tools for taxonomic profiling
  - MetaPhlAn
- Tools for functional profiling
  - HUMAnN
  - ShortBRED
  - PICRUSt
- Tools for testing associations
  - LEfSe
  - MaAsLin
  - CCREPE
- Resources
- Research vignette (time permitting)

Plan

- Informal survey
- Metagenomics concepts & examples
- Tools for taxonomic profiling
  - MetaPhlAn
- Tools for functional profiling
  - HUMAnN
  - ShortBRED
  - PICRUSt
- Tools for testing associations
  - LEfSe
  - MaAsLin
  - CCREPE
- Resources
- Research vignette (time permitting)