Scalable metabolic reconstruction for metagenomic data and the human microbiome

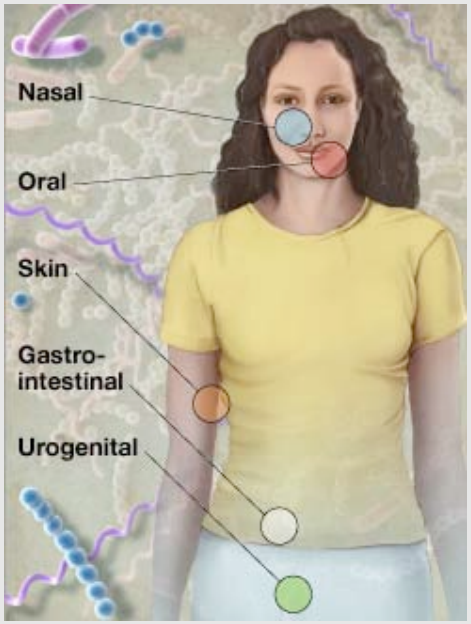
Sahar Abubucker, Nicola Segata, Johannes Goll,
Alyxandria Schubert, Beltran Rodriguez-Mueller, Jeremy Zucker,
the Human Microbiome Project Metabolic Reconstruction team,
the Human Microbiome Consortium,
Patrick D. Schloss, Dirk Gevers, Makedonka Mitreva,
Curtis Huttenhower

Harvard School of Public Health
Department of Biostatistics

03-09-11

Scale and scope of the Healthy Human Microbiome Project

300 People/ 15(18) Body Sites



Multifaceted data

- >12,000 samples
- >50M 16S seqs.
- 4.6Tbp unique metagenomic sequence
- >1,900 reference genomes
- Full clinical metadata

Multifaceted analyses

- Human population
- Microbial population
- Novel organisms
- Biotypes
- Viruses
- Metabolism

2 clin. centers, 4 seq. centers, data generation, technology development, computational tools, ethics…

15+ disease-related Demonstration Projects

Gastrointestinal	Skin	Urogenital
• Obesity • Crohn's disease • Ulcerative colitis • Autoimmunity • Cancer • Necrotizing enterocolitis	• Psoriasis • Acne • Atopic dermatitis	• Bacterial vaginosis • STDs • Reproductive health

All include additional subjects and technology development

What to do with your metagenome?

Who's there?
What are they doing?

What do functional genomic data tell us about microbiomes?

What can our microbiomes tell us about us?

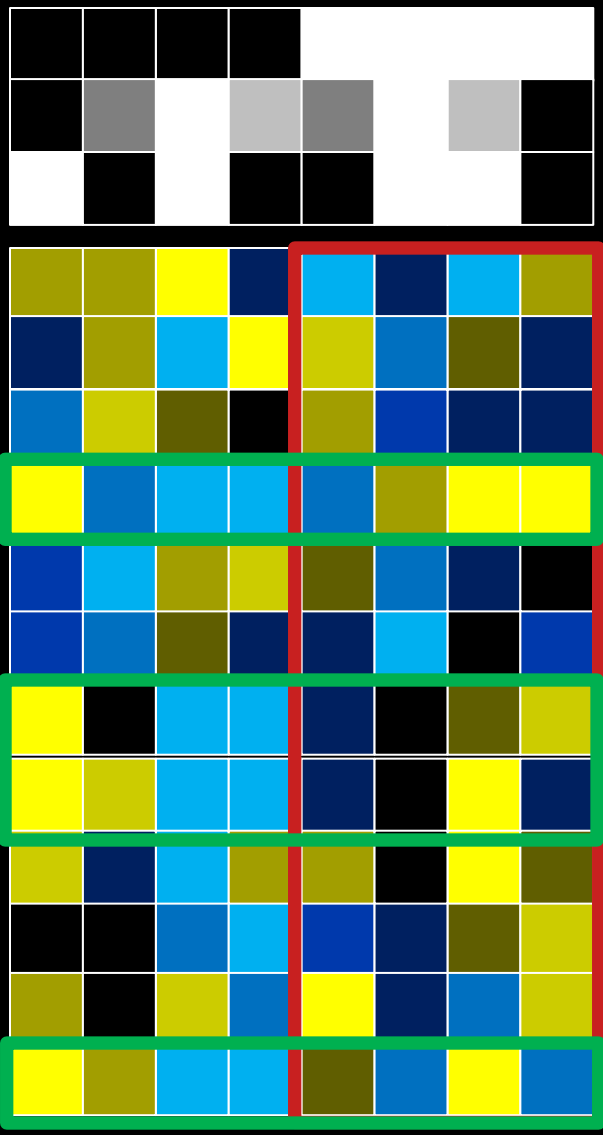
voir of
d protein

Comprehensive snapshot of microbial ecology and evolution

Public health tool monitoring population health and interactions

Diagnc
progr
biomar
host di

Metabolic/Functional Reconstruction: The Goal



Healthy/IBD

BMI

Diet

Batch effects?

Population structure?

Gene abundances

Enzyme family abundances

Pathway abundances

Niches & Phylogeny

Intervention/ perturbation

Biological story?

Independent sample

Cross-validate

Confounds/ stratification/ environment

Test for correlates

Feature selection p >> n

Multiple hypothesis correction

HMP: Metabolic reconstruction

HUMAnN:
HMP Unified Metabolic Analysis Network

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

300 subjects
1-3 visits/subject
~6 body sites/visit
10-200M reads/sample
100bp reads

BLAST

Functional seq.
KEGG + MetaCYC
CAZy, TCDB,
VFDB, MEROPS…

BLAST → Genes

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

Genes → Pathways
MinPath (Ye 2009)

Taxonomic limitation
Rem. paths in taxa < ave.

Smoothing
Witten-Bell

$$c(g) = \begin{cases} TN/(V-T)/(N+T) & c(g) = 0 \\ c(g)N/(N+T) & otherwise \end{cases}$$

Gap filling
c(g) = max(c(g), median)

Xipe
Distinguish zero/low
(Rodriguez-Mueller in review)

Genes (KOs)

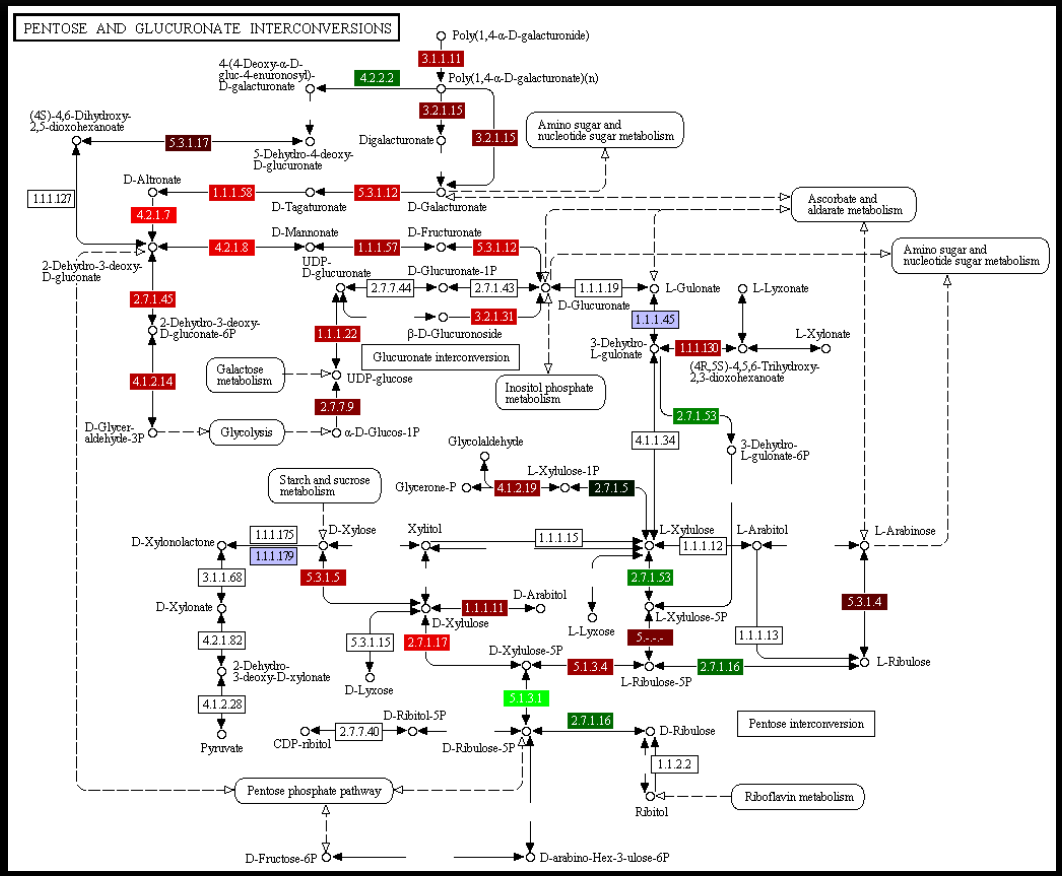
WGS reads

?

Pathways (KEGGs)

Pathways/ modules

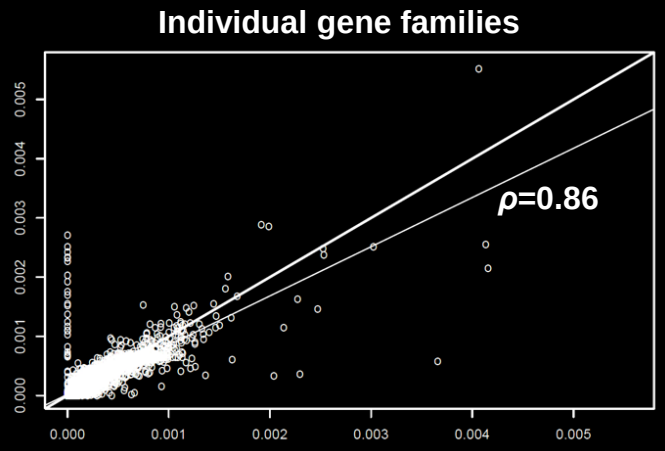
HMP: Metabolic reconstruction



Pathway coverage

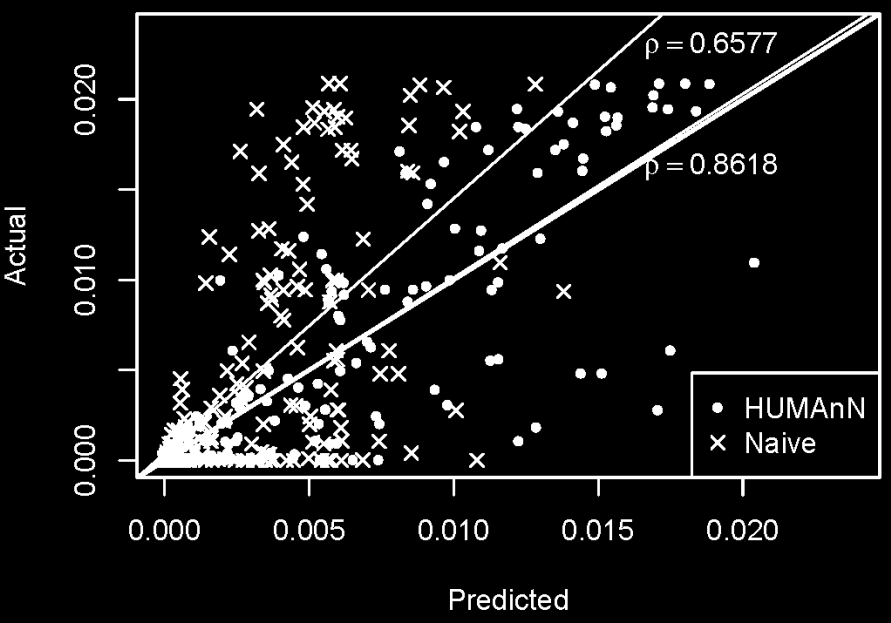
Pathway abundance

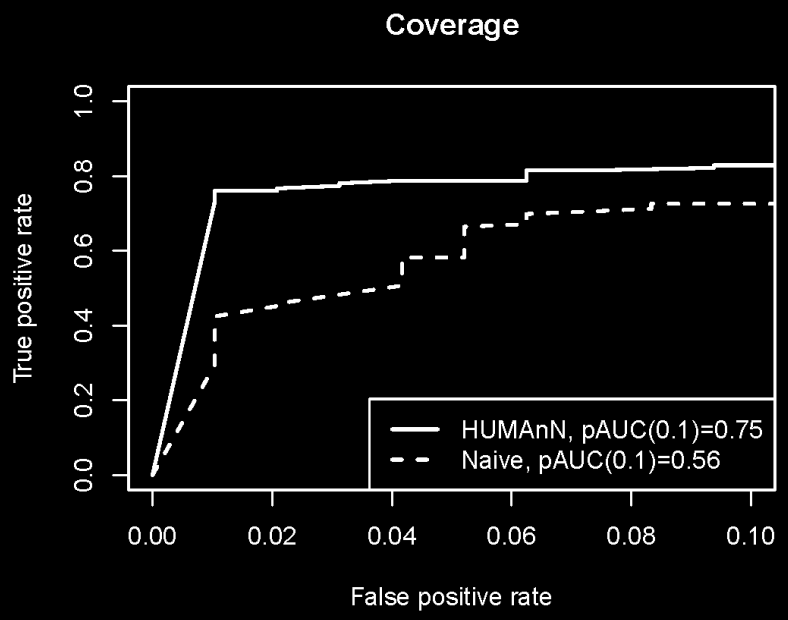
HUMAnN: Validating gene and pathway abundances on synthetic data



Validated on individual gene families, module coverage, and abundance

- 4 synthetic communities:
  Low (20 org.) and high (100 org.) complexity
  Even and lognormal abundances
- False negatives: short genes (<100bp), taxonomically rare pathways
- False positives: large and multicopy (not many in bacteria)



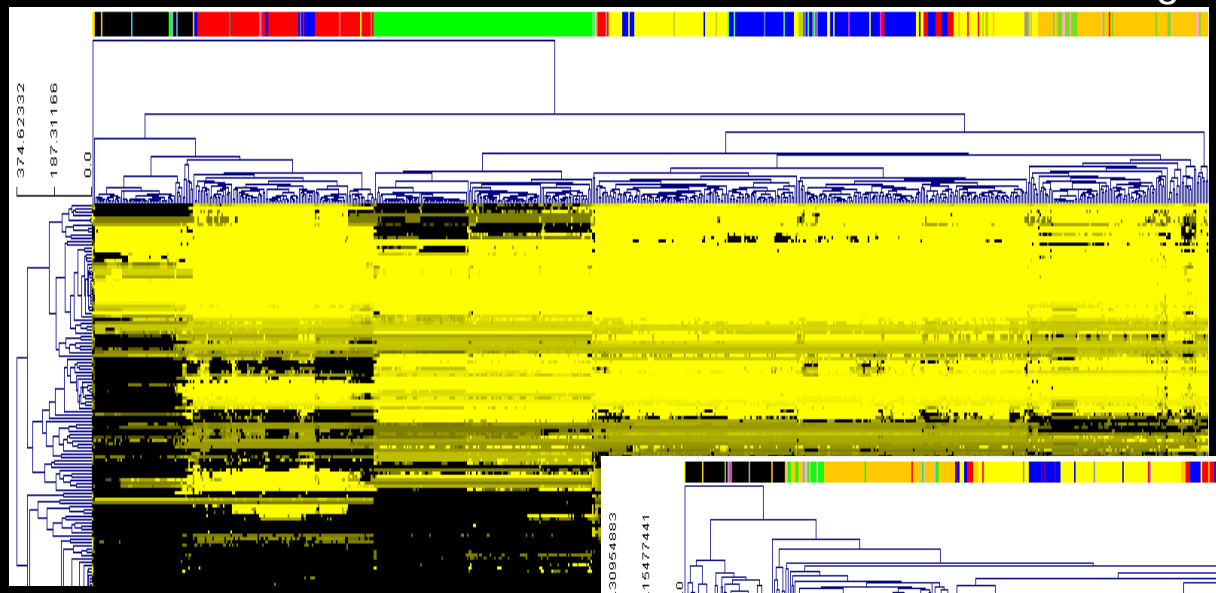


Functional modules in 741 HMP samples

PF O(BM) S O(SP) O(TD) RC AN

← Samples →

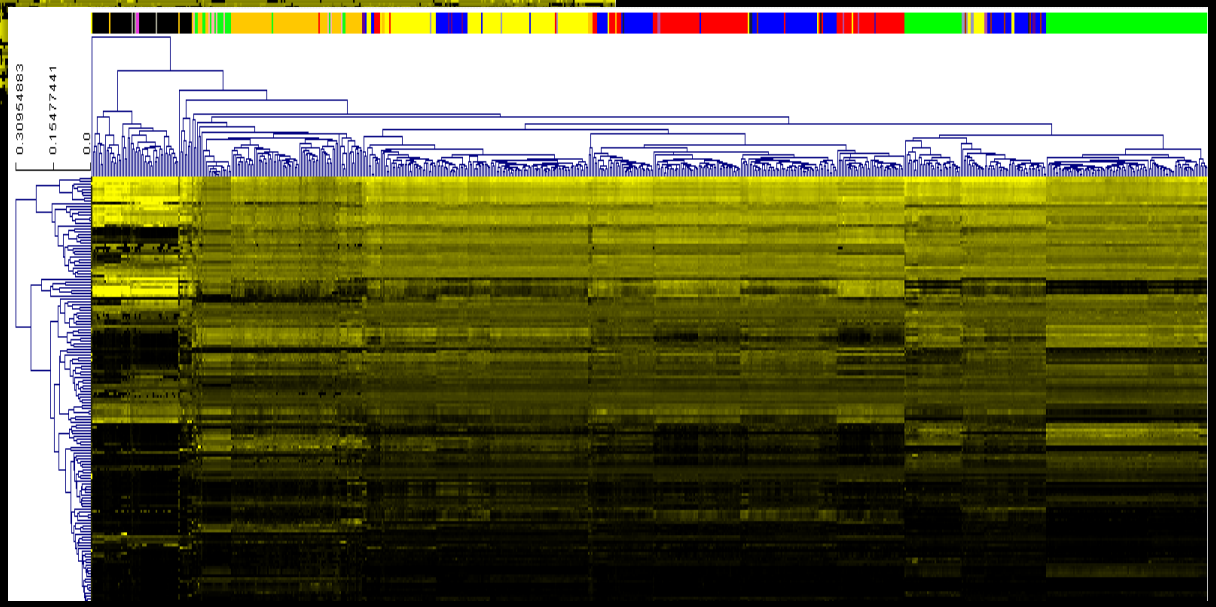
Coverage



← Pathways →

- **Zero** microbes (of ~1,000) are core among body sites
- **Zero** microbes are core among individuals
- 19 (of ~220) pathways are present in every sample
- 53 pathways are present in 90%+ samples

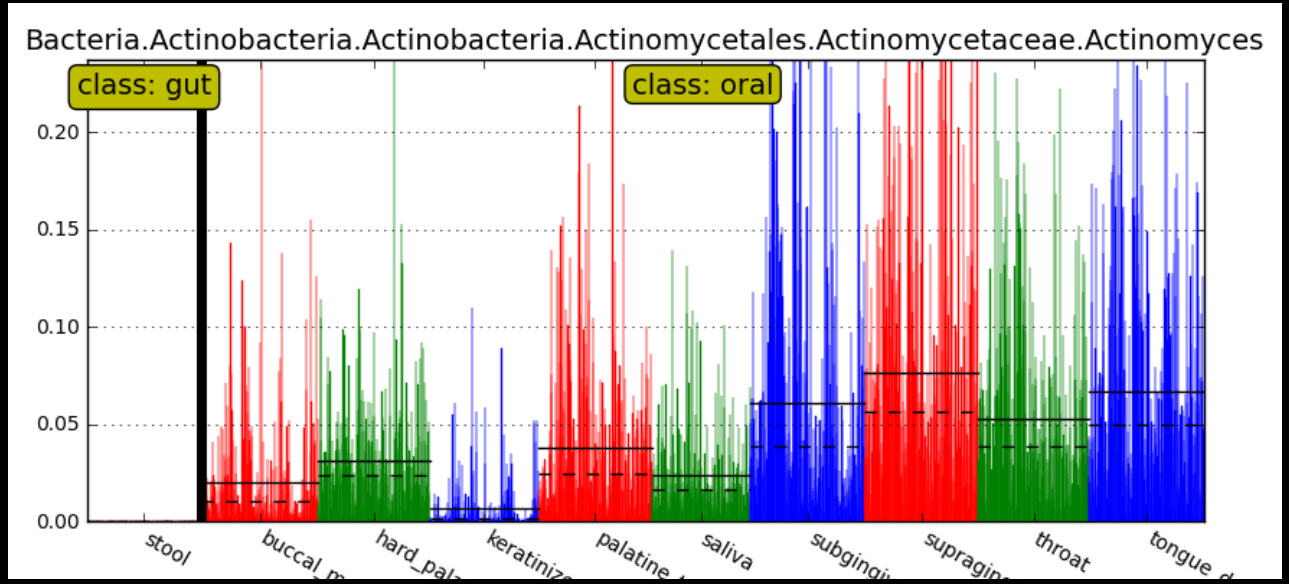
Abundance

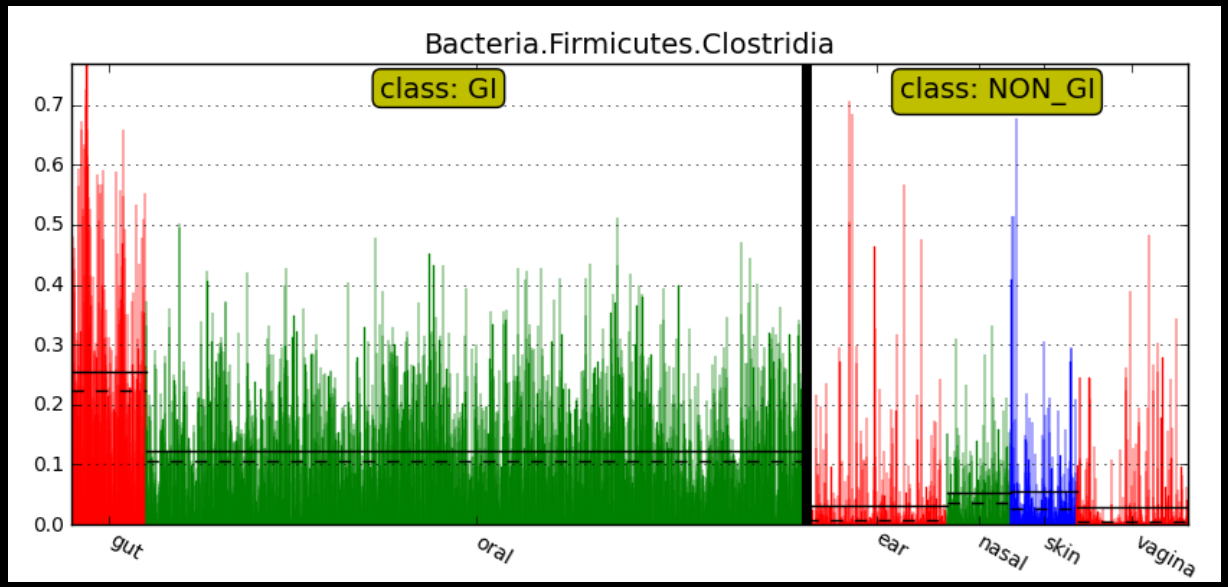


- Only 31 (of 1,110) pathways are present/absent from exactly one body site
- 263 pathways are differentially abundant in exactly one body site

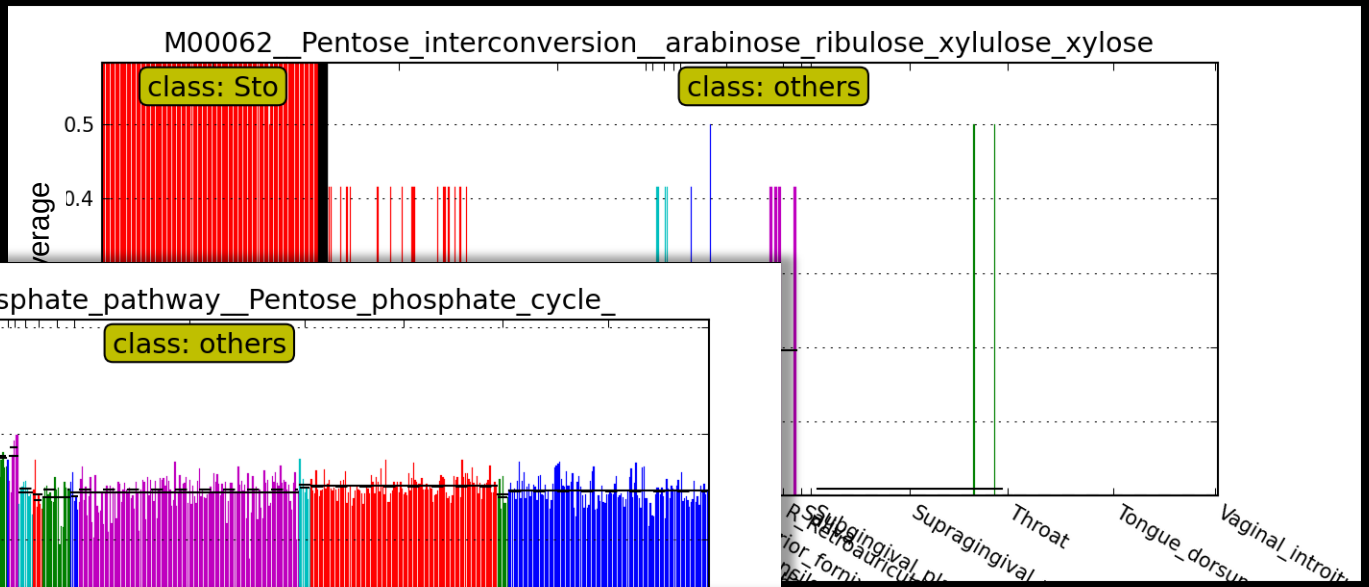
A portrait of the human microbiome: Who's there?

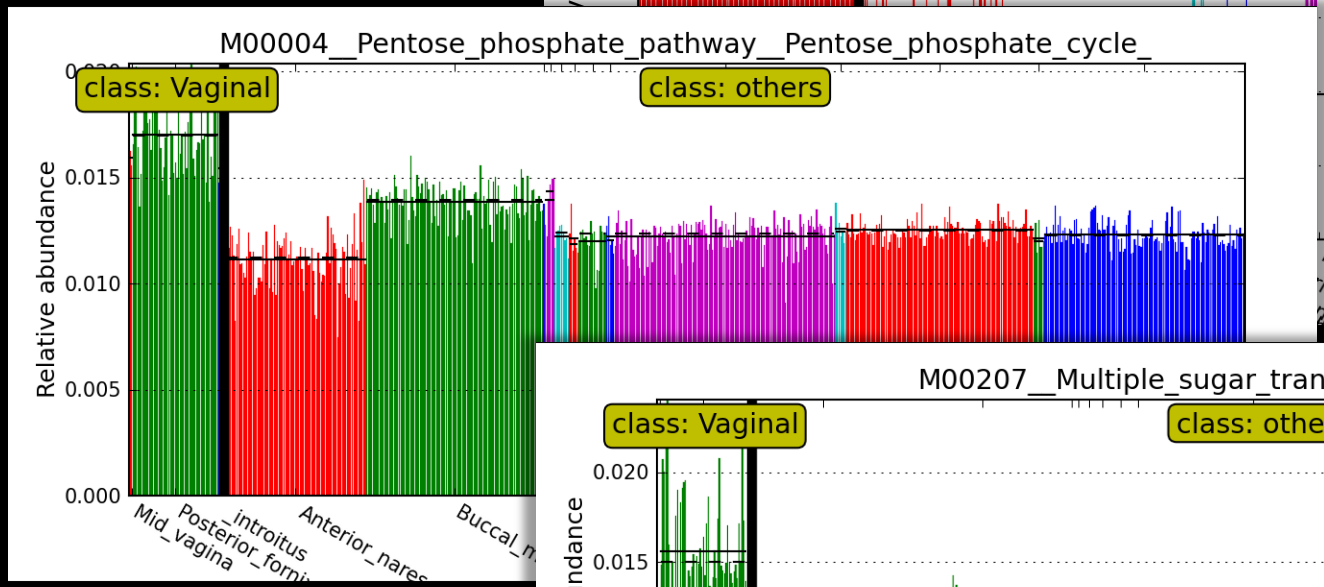
With Jacques Izard

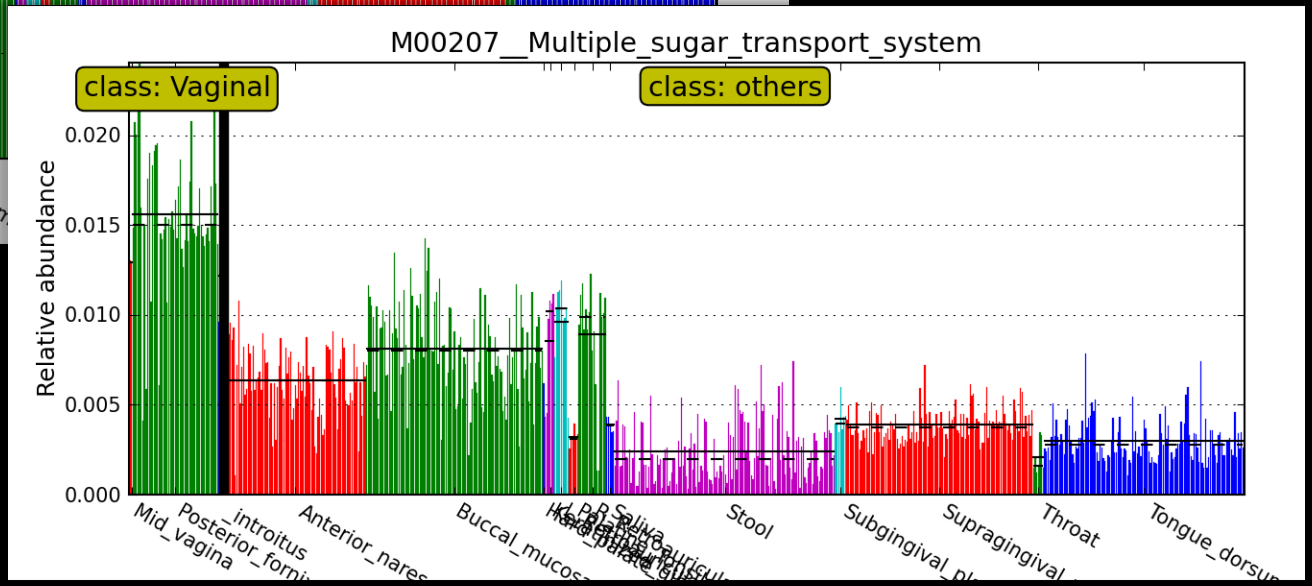




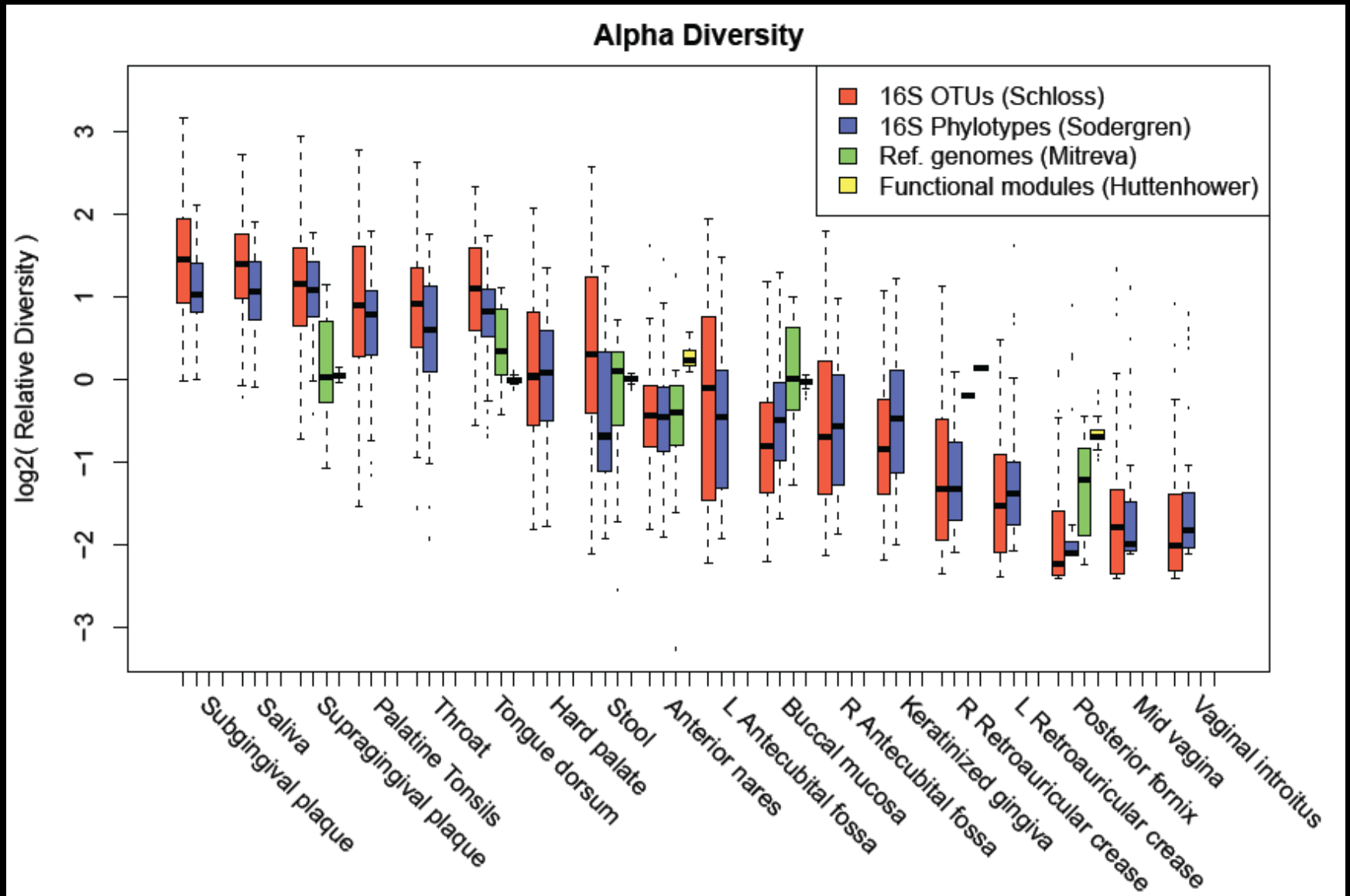
A portrait of the human microbiome: What are they doing?



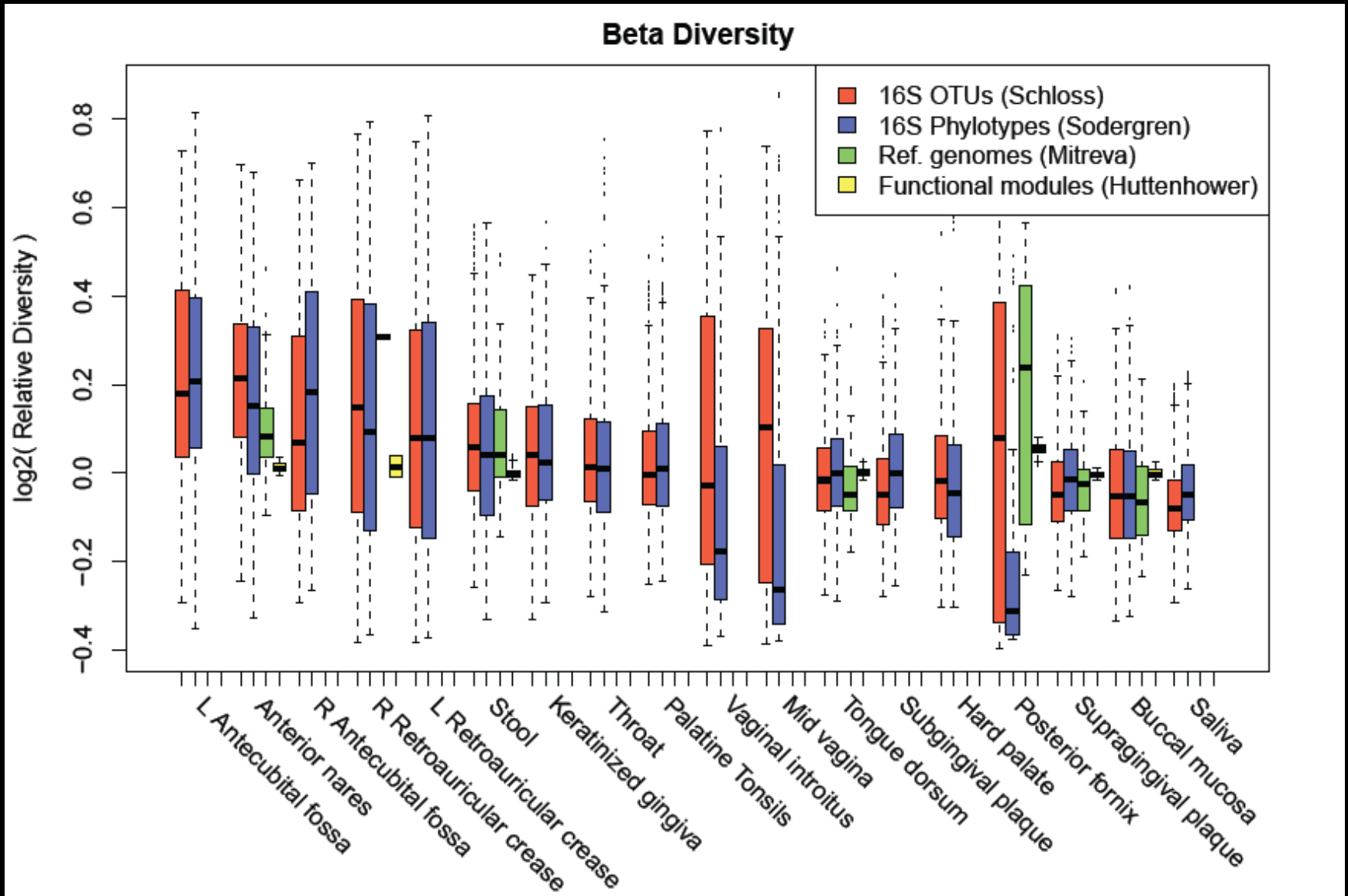




HMP: How do microbes vary within each body site across the population?

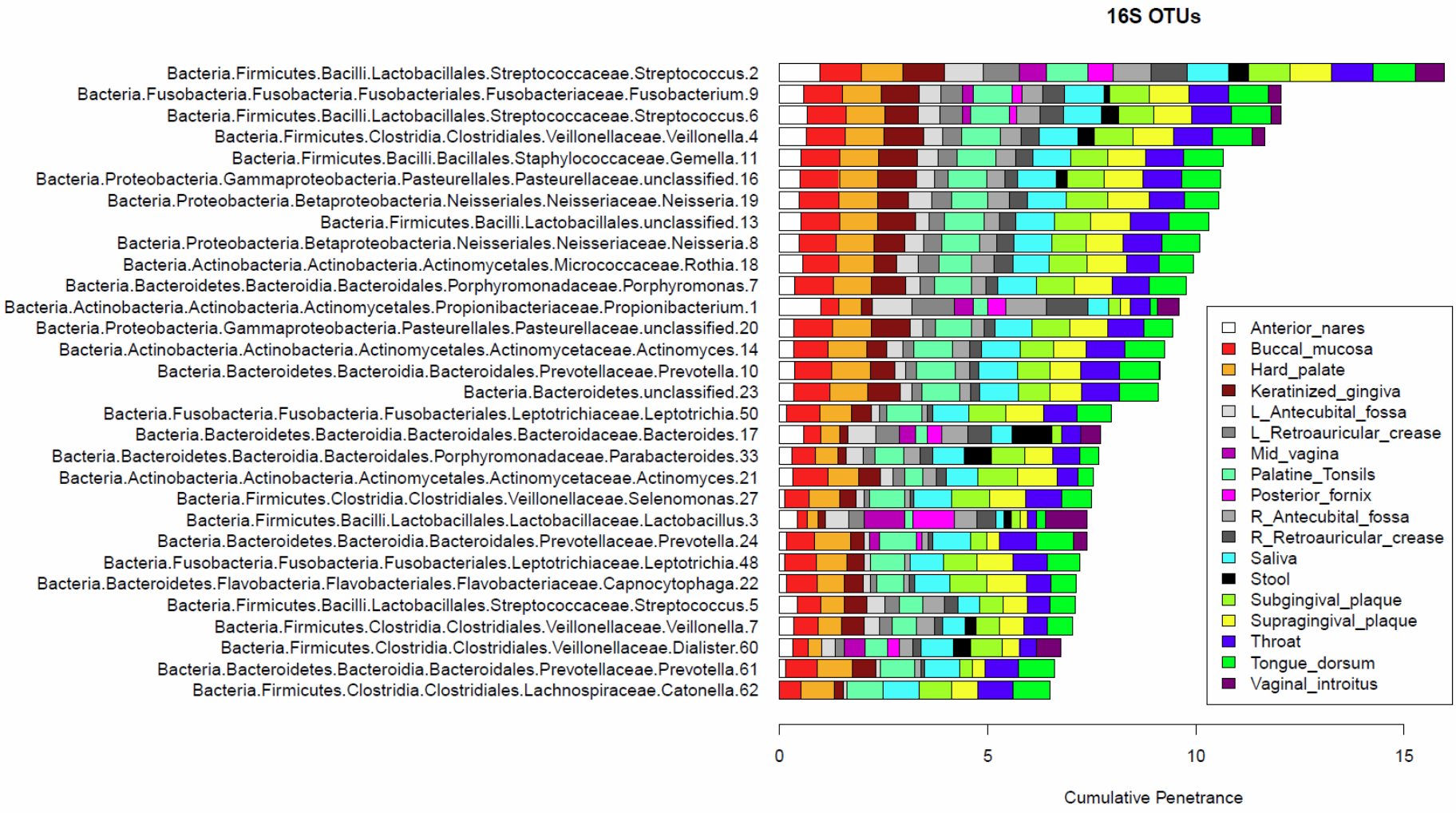


HMP: How do body sites compare between individuals across the population?



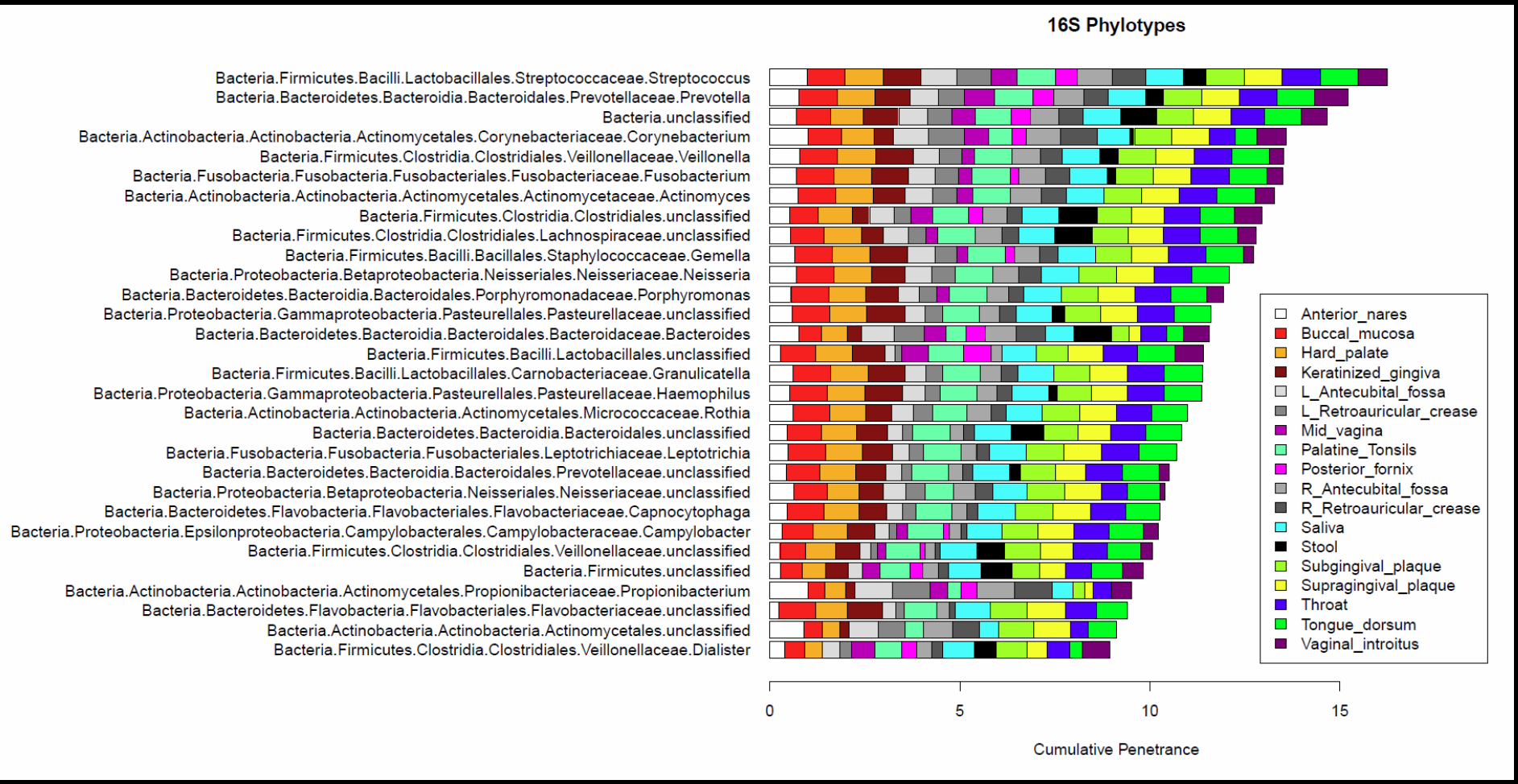
HMP: Penetrance of species (OTUs) across the population

Data from Pat Schloss

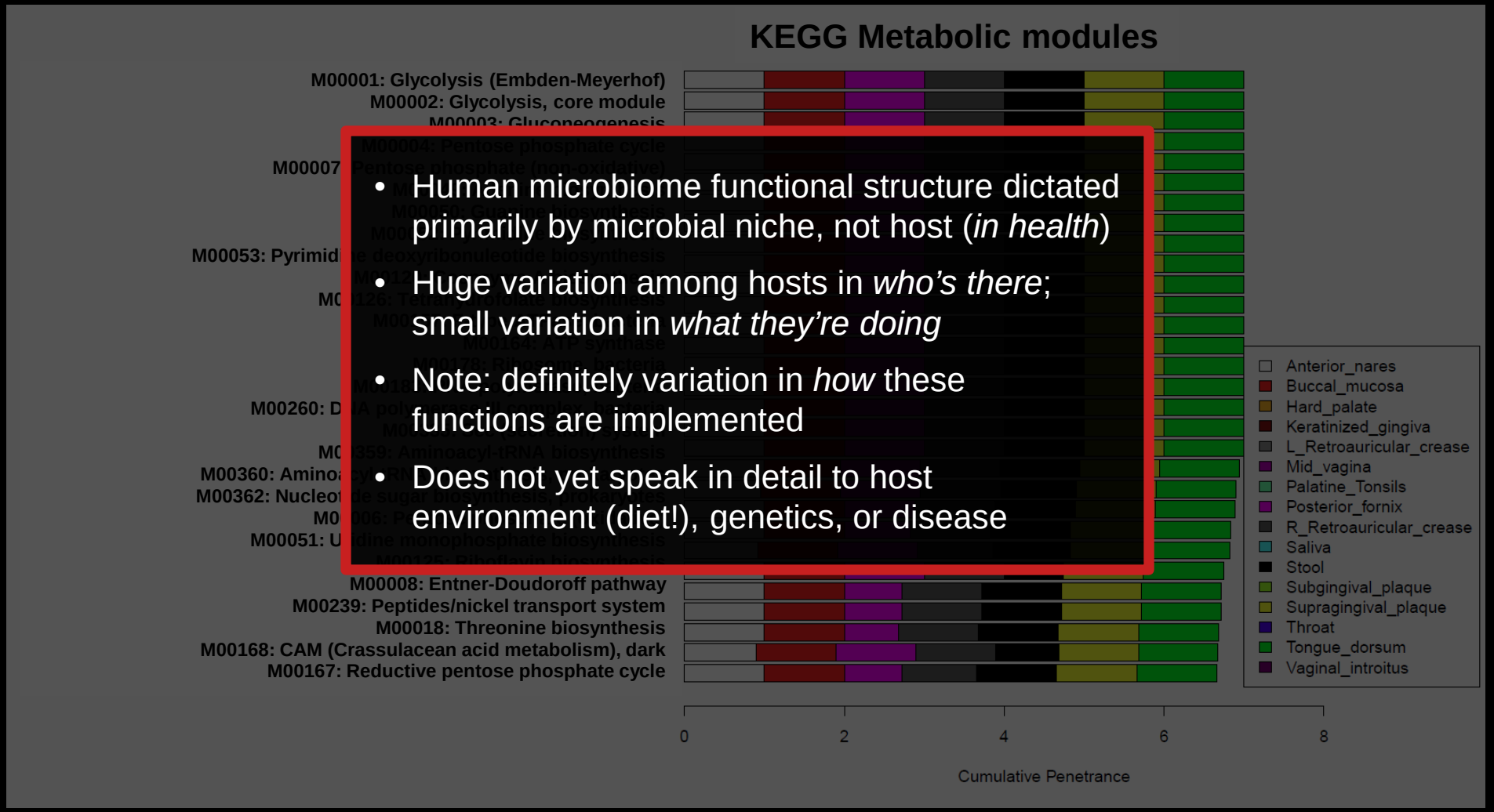


HMP: Penetrance of genera (phylotypes) across the population

Data from Pat Schloss



HMP: Penetrance of *pathways* across the population

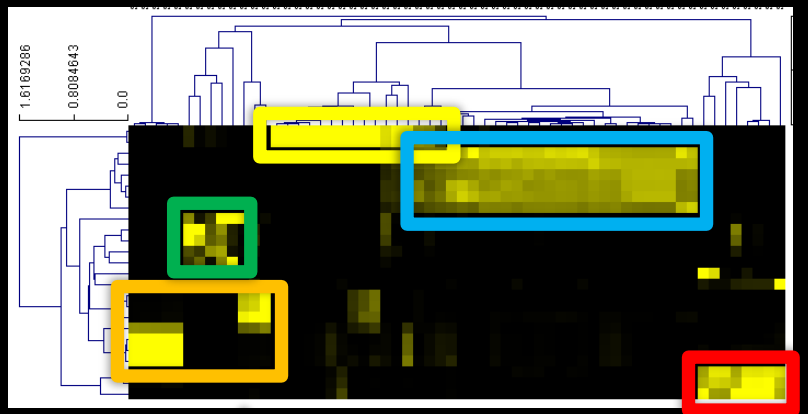


- Human microbiome functional structure dictated primarily by microbial niche, not host (*in health*)
- Huge variation among hosts in *who's there*; small variation in *what they're doing*
- Note: definitely variation in *how* these functions are implemented
- Does not yet speak in detail to host environment (diet!), genetics, or disease

Population summary statistics → population biology

← Individuals →

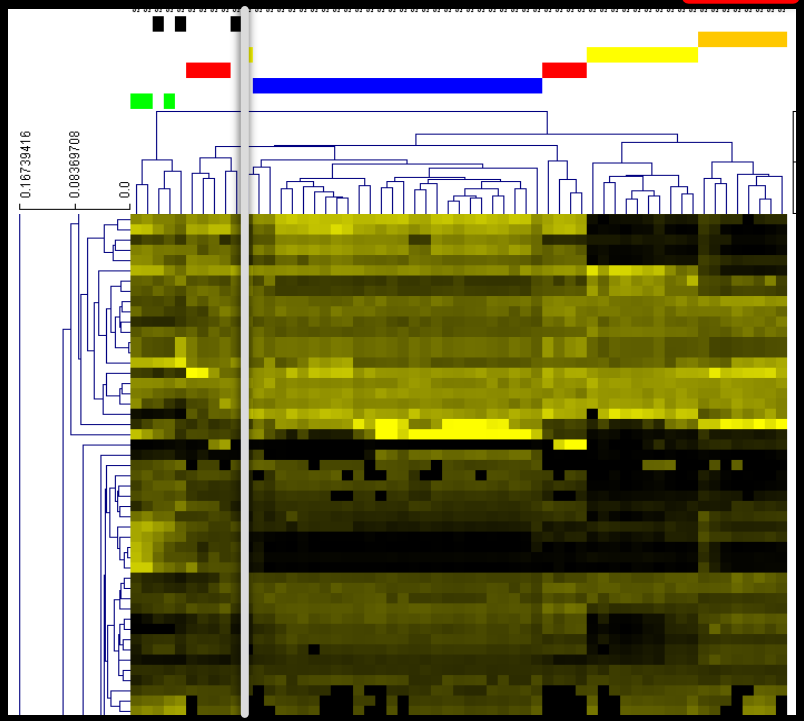
← Species →



Posterior fornix, ref. genomes

Lactobacillus iners
Lactobacillus crispatus
Gardnerella vaginalis
Lactobacillus jensenii
Lactobacillus gasseri

← Pathways →



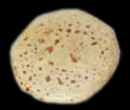
Posterior fornix, functional modules

Essential amino acids

Basic biology, sugar transport

Urea cycle, amines, aromatic AAs

LEfSe: Metagenomic class comparison and explanation



LEfSe

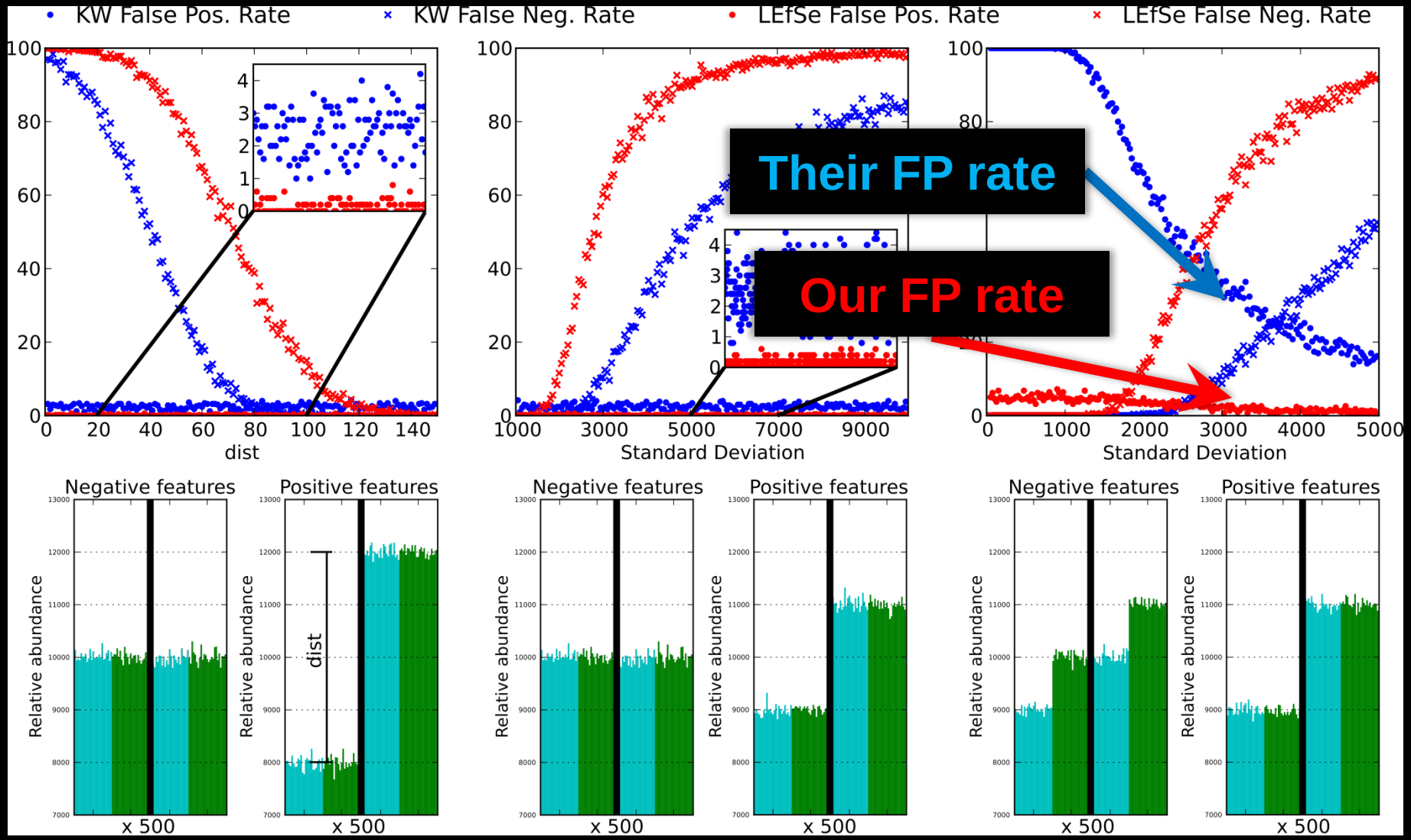
LDA +
Effect Size

Nicola Segata

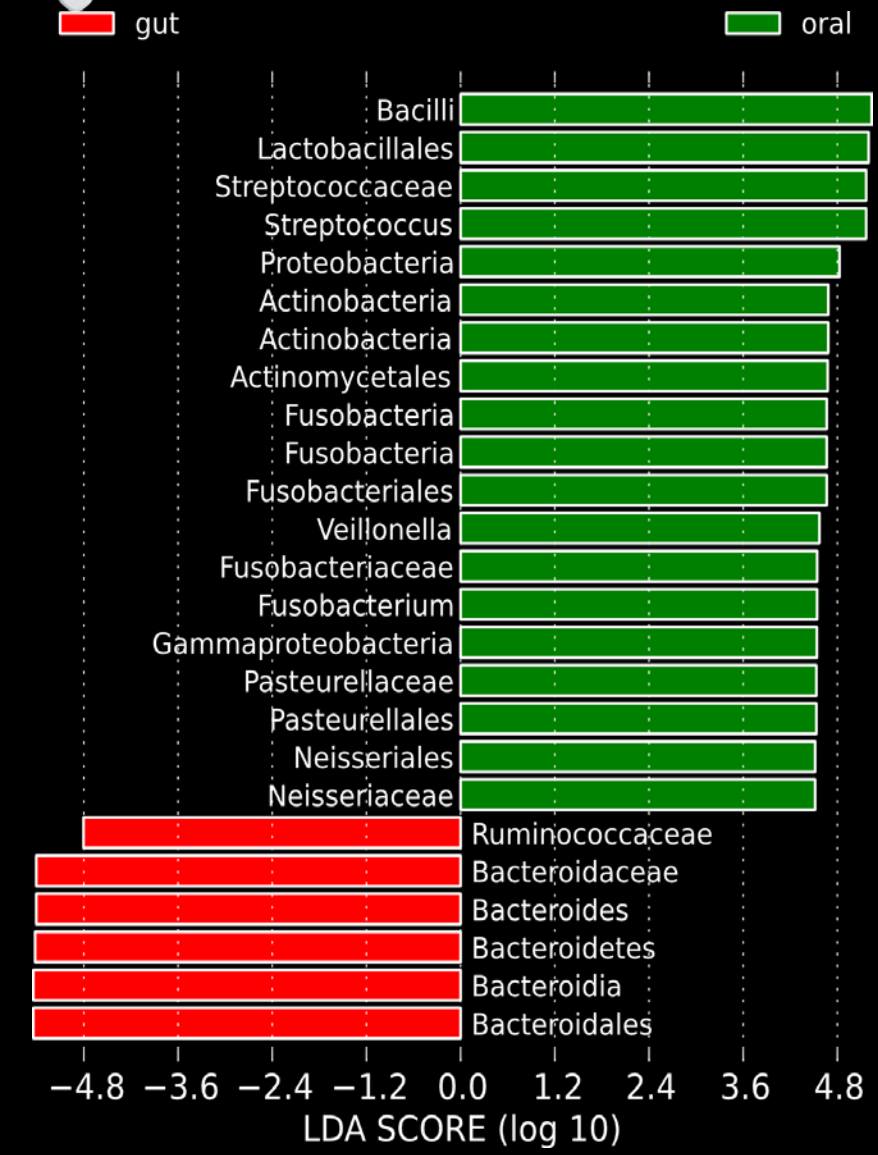
Biological hypothesis
- differential analysis
- comparative analysis
- biomarker discovery
- structure of the problem

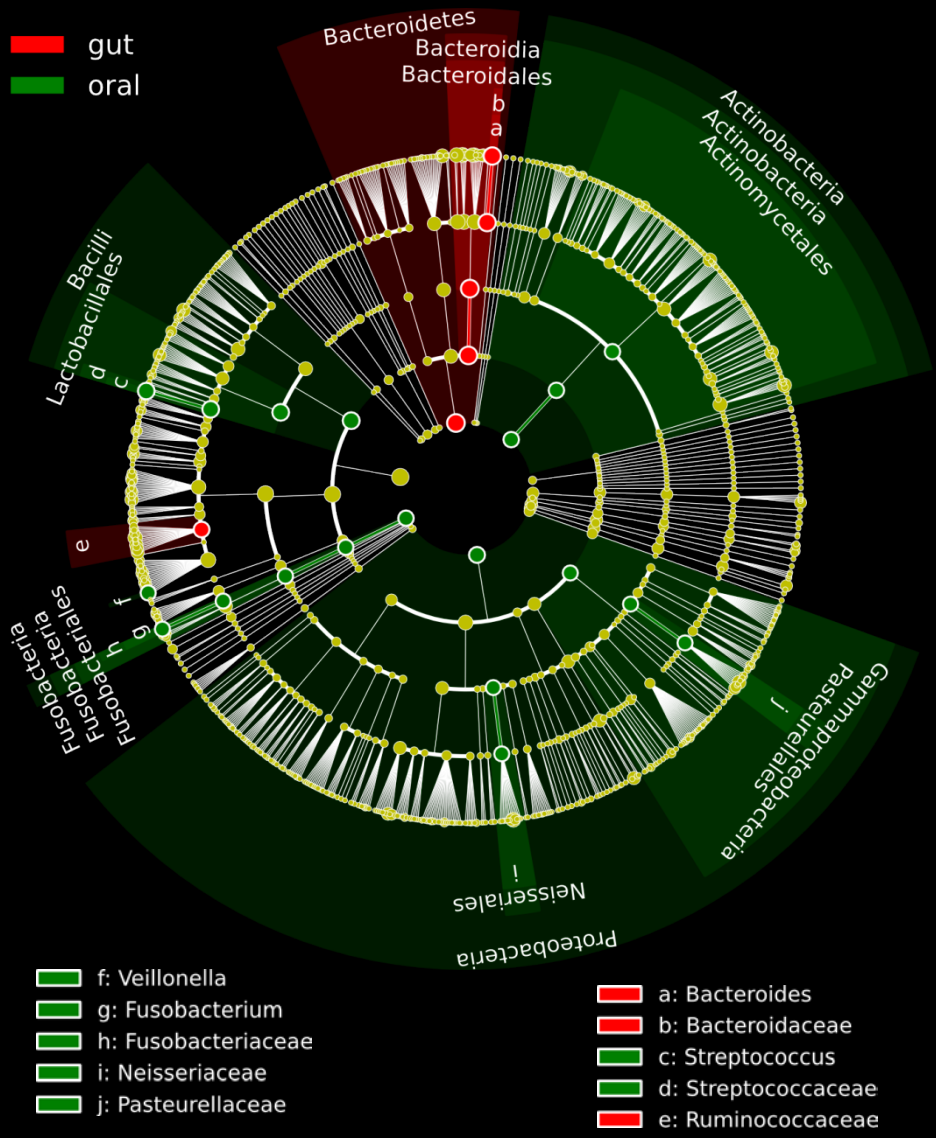
Two (or more) conditions

LEfSe: Evaluation on synthetic data

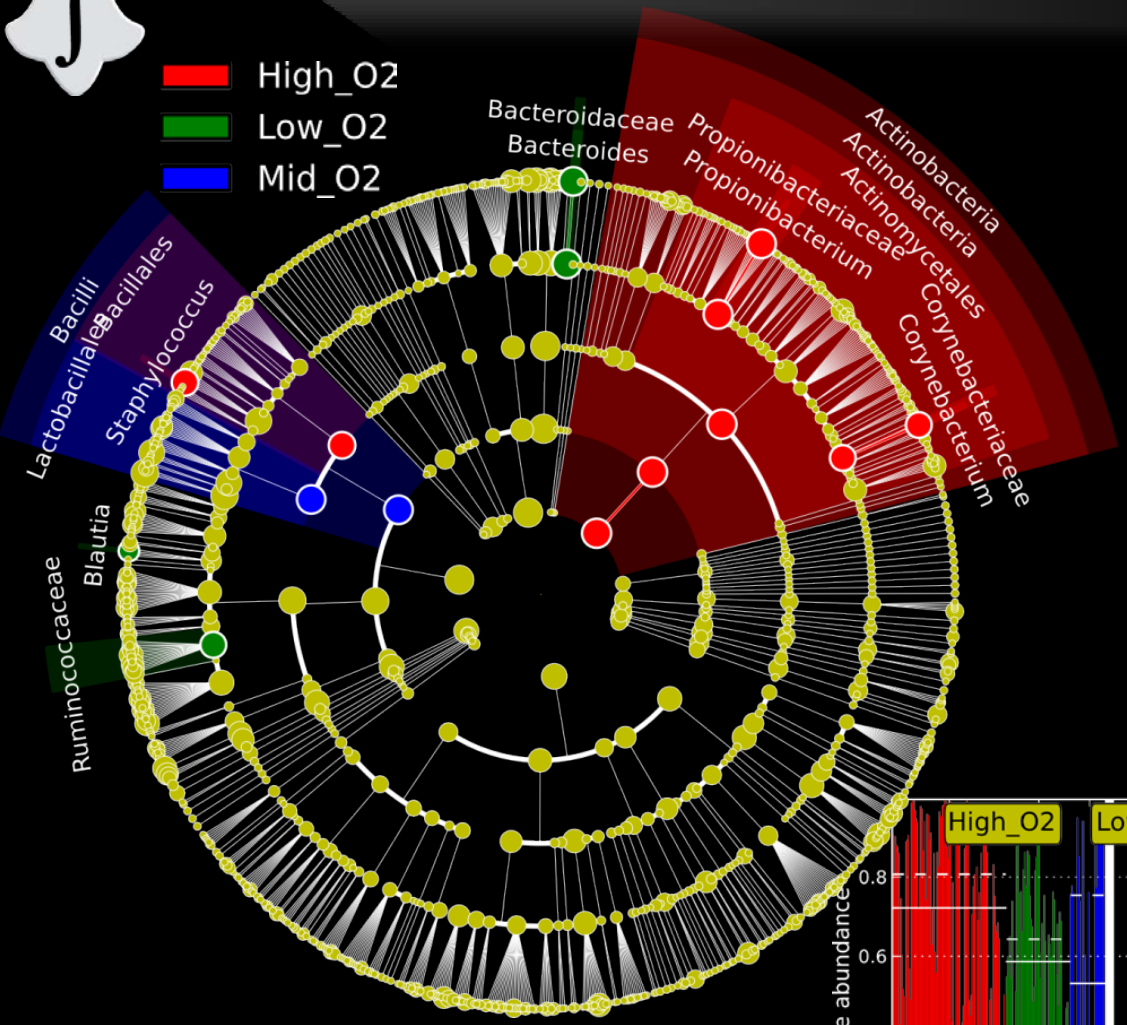


Microbes characteristic of the oral and gut microbiota

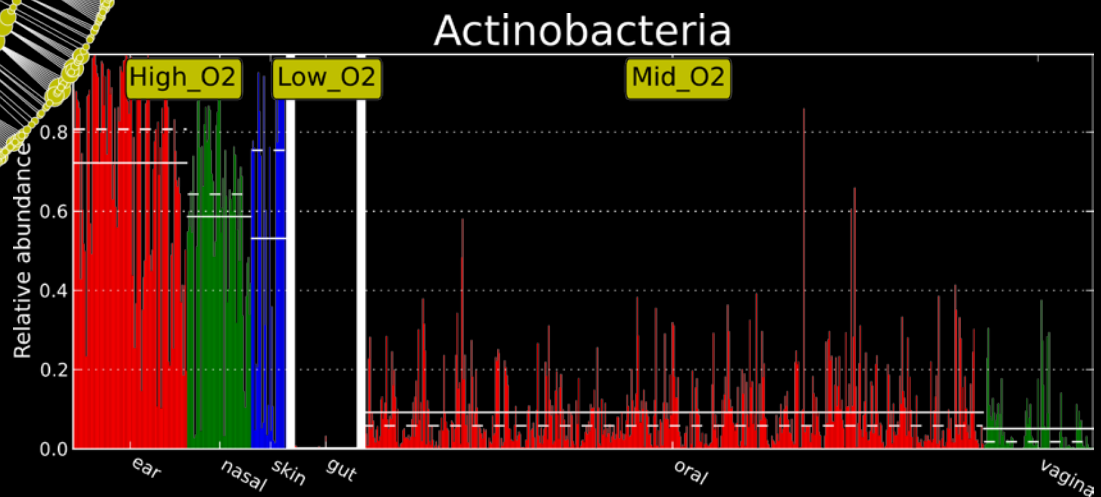




Aerobic, microaerobic and anaerobic communities

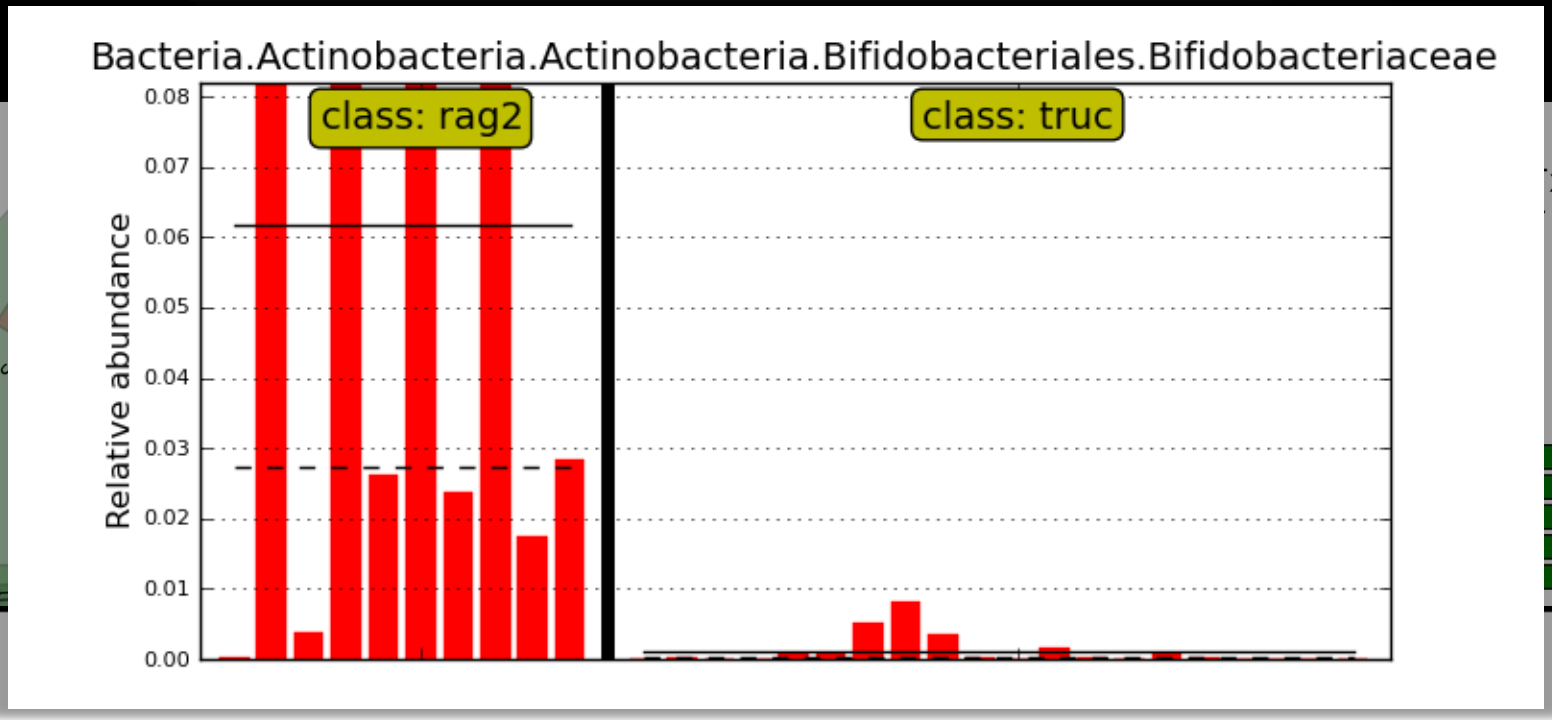


- **High oxygen: skin, nasal**
- **Mid oxygen: vaginal, oral**
- **Low oxygen: gut**



LEfSe: The TRUC murine colitis microbiota

With Wendy Garrett

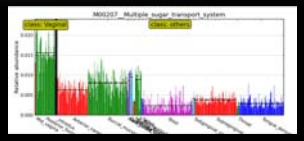


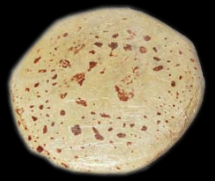
Bifidobacterium animalis subsp. *lactis* fermented milk product reduces inflammation by altering a niche for colitogenic microbes

Patrick Veiga[a,b], Carey Ann Gallini[a], Chloé Beal[b], Monia Michaud[a], Mary L. Delaney[c], Andrea DuBois[c], Artem Khlebnikov[b,d], Johan E.T. van Hylckama Vlieg[b], Shivesh Punit[a,1], Jonathan N. Glickman[c,e,2], Andrew Onderdonk[c,e], Laurie H. Glimcher[a,d,e,f], and Wendy S. Garrett[a,e,g,3]

[a]Harvard School of Public Health, Boston, MA 02115; [b]Danone Research, 91767 Palaiseau, France [c]Brigham and Women's Hospital, Boston, MA 02115; [d]Dannon Company Inc., White Plains, NY 10603; [e]Harvard Medical School, Boston, MA 02115; [f]Ragon Institute of MGH, MIT and Harvard, Charlestown, MA 02129; and [g]Dana Farber Cancer Institute, Boston, MA 02115

Microbial biomolecular function and biomarkers in the human microbiome: the story so far?





- *Who's there* changes
  - *What they're doing* doesn't (as much)
  - *How they're doing it* does
- The data so far only scratch the surface
  - Only 1/3 to 2/3 of the reads/sample map to cataloged gene families
  - Only 1/3 to 2/3 of these gene families have cataloged functions
  - Very much in line with MetaHIT
  - **Job security!**
- Looking forward to functional reconstruction…
  - In environmental communities
  - With respect to host environment + genetics
  - With respect to host disease

Thanks!

Nicola Segata

Pinaki Sarder

Levi Waldron

Larisa Miropolsky

Ramnik Xavier

Dirk Gevers

Wendy Garrett

Jacques Izard

Human Microbiome Project

George Weinstock
Jennifer Wortman
Owen White
Makedonka Mitreva
Erica Sodergren
Mihai Pop
Vivien Bonazzi
Jane Peterson
Lita Proctor

Sahar Abubucker
Yuzhen Ye
Beltran Rodriguez-Mueller
Jeremy Zucker
Qiandong Zeng
Mathangi Thiagarajan
Brandi Cantarel
Maria Rivera
Barbara Methe
Bill Klimke
Daniel Haft

HMP Metabolic Reconstruction

Bruce Birren
Doyle Ward
Ashlee Earl
Mark Daly
Eric Alm
Lisa Cosimi

BROAD INSTITUTE

MIT
MASSACHUSETTS GENERAL HOSPITAL
HARVARD SCHOOL OF PUBLIC HEALTH
BROAD INSTITUTE
Children's Hospital Boston
The Forsyth Institute
HARVARD MEDICAL SCHOOL
HARVARD UNIVERSITY

Interested? We're recruiting graduate students and postdocs!

http://huttenhower.sph.harvard.edu

http://huttenhower.sph.harvard.edu/humann
http://huttenhower.sph.harvard.edu/lefse

MiRiBA

HMP Research Consortium >30 POSTERS

- HMP data processing
  - POSTER 3. - Metabolic reconstruction
  - POSTER 100. - Read mapping
  - POSTER 117. - DACC QC
  - POSTER 163. - Cumulative Abundances
- Reference Genomes
  - HMP Project Catalog
    - POSTER 93. Reference Genome Catalog
  - Reference genome annotation goals
    - POSTER 98. Ref Genome Annotation Methods
    - POSTER 122. Seq-ing and Ann. Ref at Baylor
  - Strain Access & Strain Requests
    - POSTER 35. 100 Most Wanted
    - POSTER 170. Single Bacterial Cells
  - Downloading annotations
  - Annotation at IMG/HMP
- WGS
  - POSTER 2. Longitudinal assessment
  - POSTER 79, 139. PhylOTU
  - POSTER 100. Optimizing read mapping
  - POSTER 117. Quality Control
  - Software
    - POSTER 90. Metagenomic Assembly
- Demonstration projects
  - POSTER 20. Crohn's disease
  - POSTER 169. Esophageal Adenocarcinoma
  - POSTER 103. Pediatric Abdominal Pain
  - POSTER 39. Vaginal microbiome
  - POSTER 34, 109. Urethral microbiome
  - POSTER 45. Reproductive health
  - POSTER 72. Atopic dermatitis
- Ethical implications of the HMP
  - POSTER 4. Identifiability of the Human Microbiome
  - POSTER 101. What's "Normal"
- 16S
  - POSTER 70. Greengenes 16S rRNA Database
  - POSTER 73. Classifiers on HighThroughput 16S rRNA
  - POSTER 167. Identification of Novel 16S Sequences in Metagenomic Data Sets
- Open Science Data Framework
  - POSTER 37.
- Accessing data from NCBI
  - IHMC codes POSTER 140.
- Statistical techniques
  - POSTER 77. Dirichlet Multinomial Power
  - POSTER 78. Analysis of Taxonomic Trees
  - POSTER 136. Metagenomic Biomarker Discovery