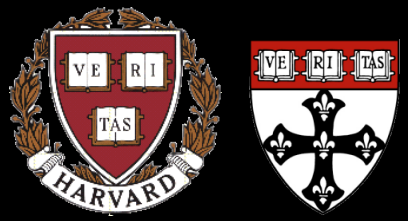
An Introduction to Microbial Community Analyses

Curtis Huttenhower

01-19-15



Harvard School of Public Health
Department of Biostatistics

U. Oregon

Everything you ever wanted to know about microbial community analysis methods but were afraid to ask

- Community composition and ecology by 16S
  - Organism identification
  - Alpha- and beta-diversity
  - Ordination
- Meta'omics: shotgun DNA and RNA seq.
  - Taxonomic profiling
  - Assembly
  - Metabolic profiling
- Downstream analyses
  - Statistical association testing
  - Microbial association networks
  - The Human Microbiome Project

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*).

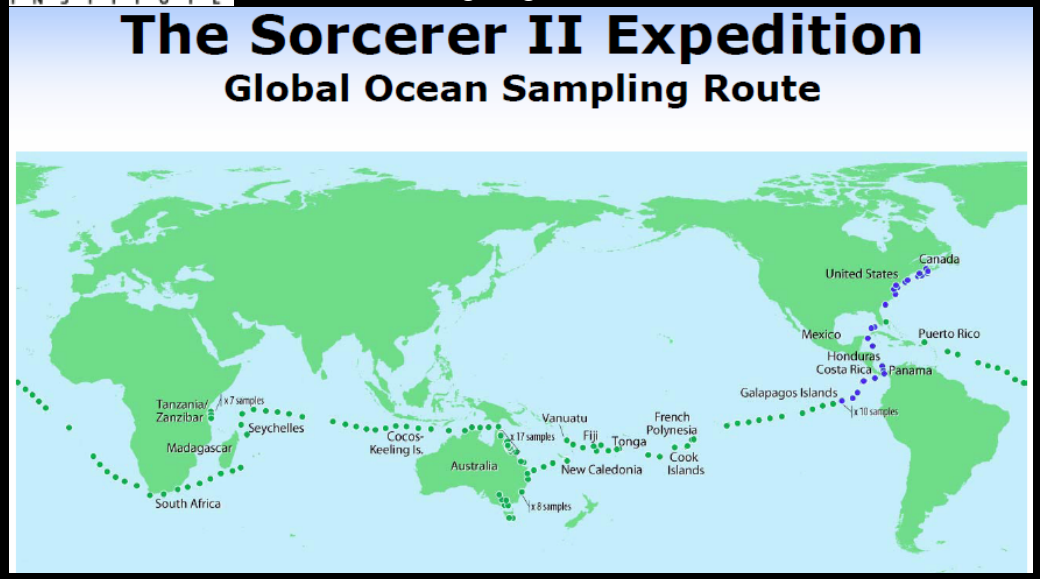
Total **biomolecular repertoire** of a microbial community

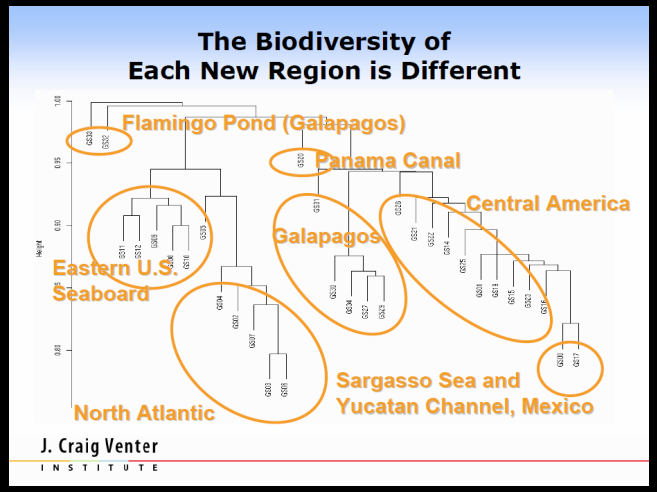
Study of **uncultured microorganisms** from the environment, which can include humans or other living hosts

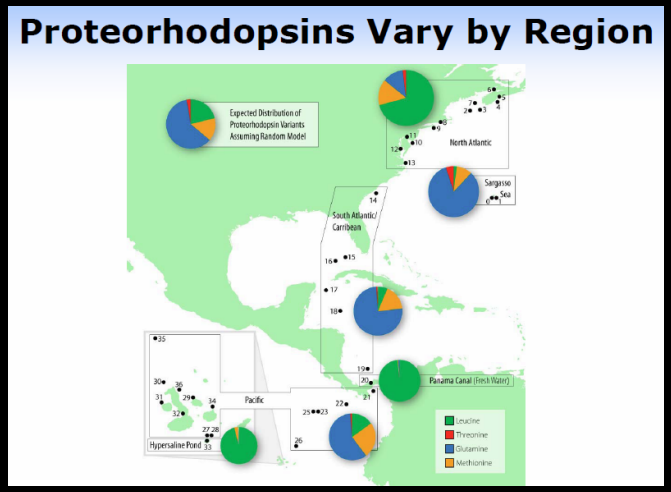
Examples of metagenomic studies: Global Ocean Sampling

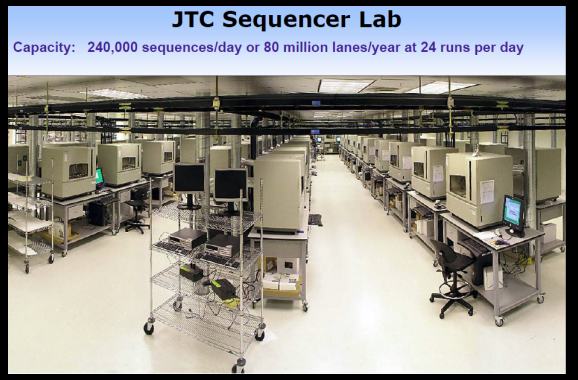
J. Craig Venter INSTITUTE

2003/2004 - ongoing









Metagenomic methods: Early work and *in situ* fluorescence

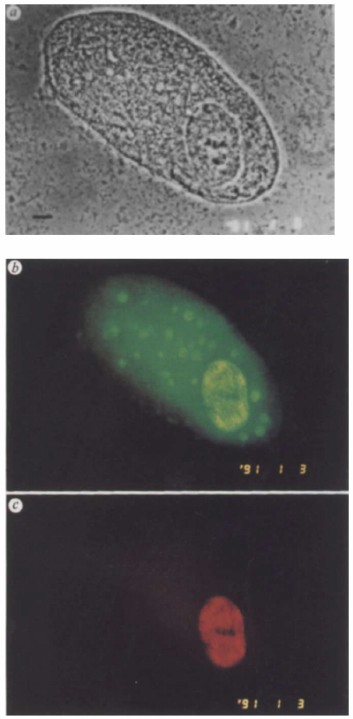
ON THE MICROSCOPIC METHOD OF STUDYING BACTERIA IN SOIL

H. J. CONN

New York Agricultural Experiment Station

Received for publication May 31, 1928

Several years ago the writer (1) proposed a method for the microscopic examination of bacteria in soil. The technic has recently assumed considerable importance because of its adoption with a few slight modifications by Winogradsky (5) in his "direct method" of studying soil bacteria. The method has



A Technique for the Quantitative Estimation of Soil Micro-organisms

By P. C. T. JONES and J. E. MOLLISON

Soil Microbiology Department, Rothamsted Experimental Station, Harpenden, Herts

APPLIED MICROBIOLOGY, June 1971, p. 1040–1045
Copyright © 1971 American Society for Microbiology
Vol. 21, No. 6
Printed in U.S.A.

Microorganisms in Unamended Soil as Observed by Various Forms of Microscopy and Staining[1]

L. E. CASIDA, JR.

Department of Microbiology, The Pennsylvania State University, University Park, Pennsylvania 16802

Received for publication 23 March 1971

Identification *in situ* and phylogeny of uncultured bacterial endosymbionts

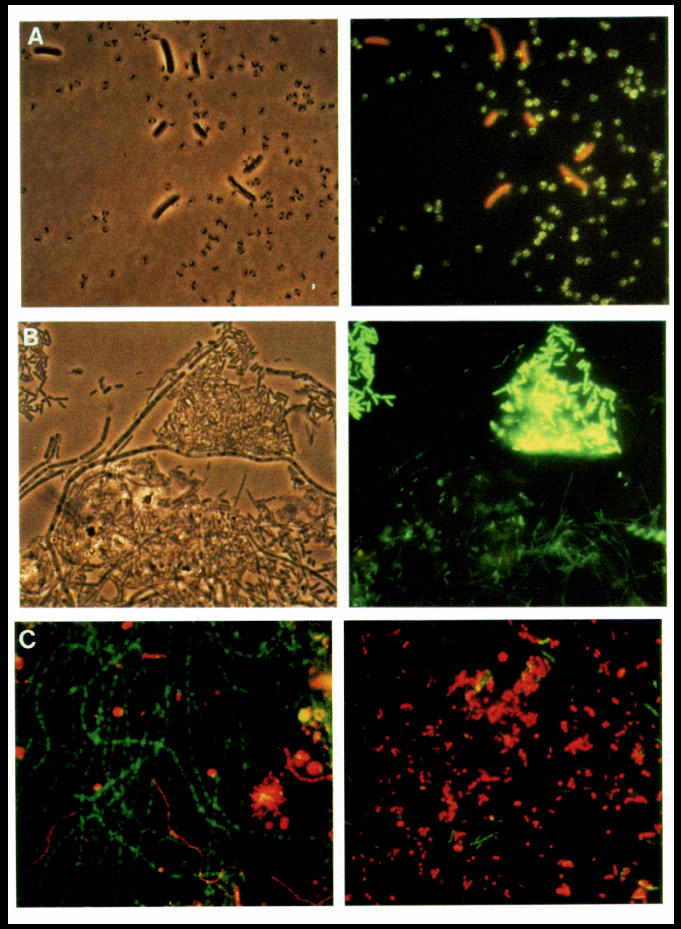
Rudolf Amann, Nina Springer, Wolfgang Ludwig, Hans-Dieter Görtz* & Karl-Heinz Schleifer

Lehrstuhl für Mikrobiologie, Technische Universität München, Arcisstr. 21, 8000 München 2, Germany
* Zoologisches Institut der Universität, Schloßplatz 5, 4400 Münster, Germany

Phylogenetic Identification and In Situ Detection of Individual Microbial Cells without Cultivation

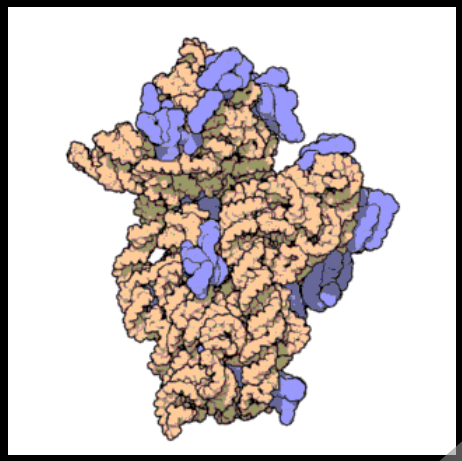
RUDOLF I. AMANN,* WOLFGANG LUDWIG, AND KARL-HEINZ SCHLEIFER

Lehrstuhl für Mikrobiologie, Technische Universität München, D-80290 Munich, Germany

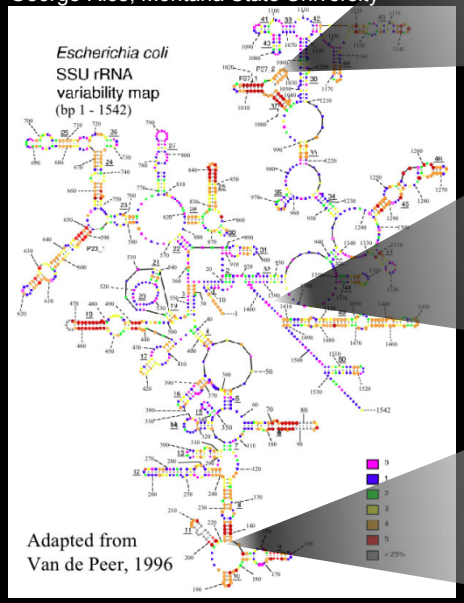


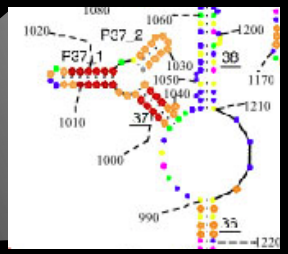
Metagenomic methods: 16S rRNA gene

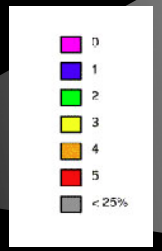
- Structural component of the prokaryotic ribosome
- Used as molecular clock to identify phylogeny:
  - Large, good scale for mutations
  - Range of mutation rates
  - Portions are constant, allowing amplification
- Not single copy! Researcher beware…

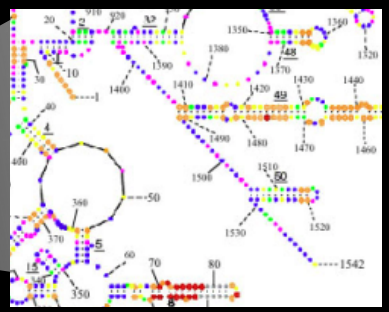


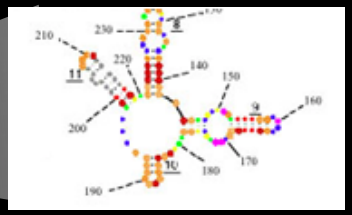
George Rice, Montana State University



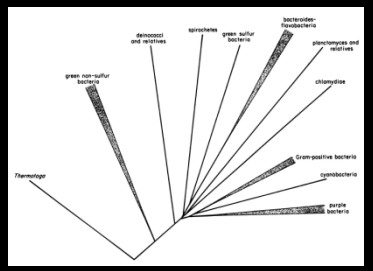




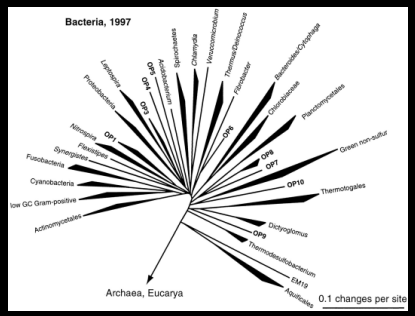




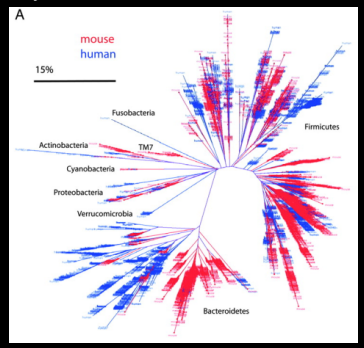
Woese, 1987



Pace, 1997



Ley, 2006



Metagenomic methods: shotgun sequencing

Analysis of a Marine Picoplankton Community by 16S rRNA Gene Cloning and Sequencing

THOMAS M. SCHMIDT,† EDWARD F. DeLONG,‡ AND NORMAN R. PACE*
Department of Biology and Institute for Molecular and Cellular Biology, Indiana University, Bloomington, Indiana 47405

Received 7 January 1991/Accepted 13 May 1991

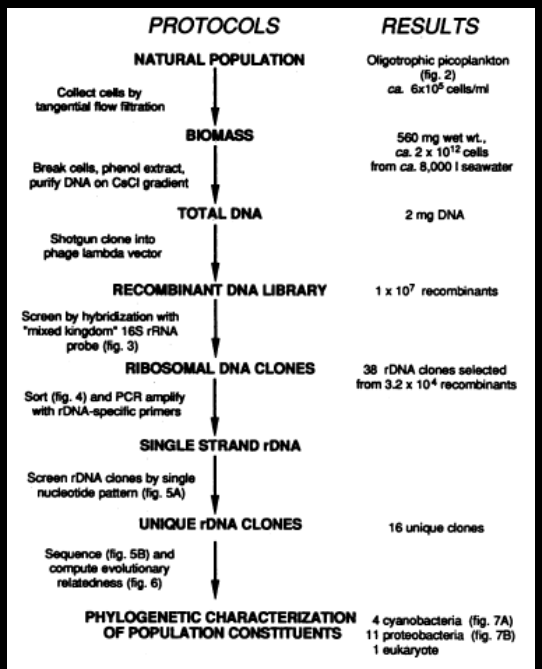


FIG. 1. Flow chart of the protocols used to characterize marine picoplankton without cultivation and a summary of some results.

Characterization of Uncultivated Prokaryotes: Isolation and Analysis of a 40-Kilobase-Pair Genome Fragment from a Planktonic Marine Archaeon

JEFFEREY L. STEIN,[1]* TERENCE L. MARSH,[2] KE YING WU,[3] HIROAKI SHIZUYA,[4] AND EDWARD F. DeLONG[3]*
Recombinant BioCatalysis, Inc., La Jolla, California 92037[1]; Microbiology Department, University of Illinois, Urbana, Illinois 61801[2]; Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, California 93106[3]; and Division of Biology, California Institute of Technology, Pasadena, California 91125[4]

Received 14 July 1995/Accepted 14 November 1995

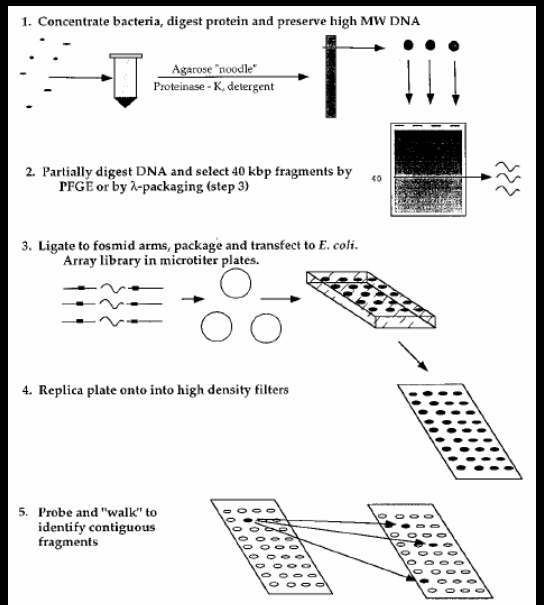
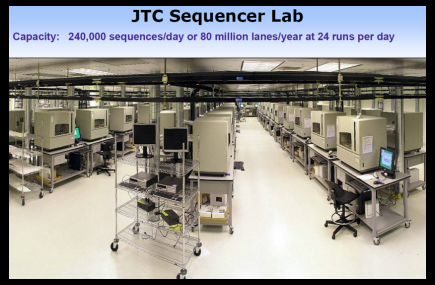


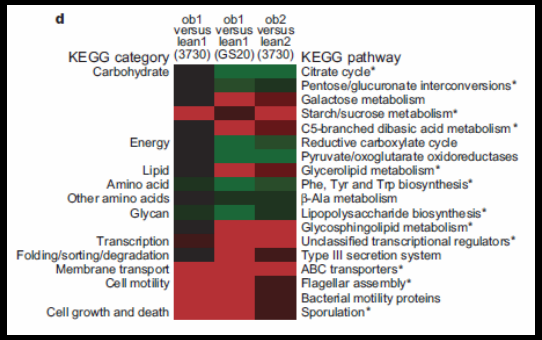
FIG. 1. Flowchart depicting the construction and screening of an environmental library from a mixed picoplankton sample. MW, molecular weight; PFGE, pulsed-field gel electrophoresis.

Environmental Genome Shotgun Sequencing of the Sargasso Sea

J. Craig Venter,[1]* Karin Remington,[1] John F. Heidelberg,[3] Aaron L. Halpern,[2] Doug Rusch,[2] Jonathan A. Eisen,[3] Dongying Wu,[3] Ian Paulsen,[3] Karen E. Nelson,[3] William Nelson,[3] Derrick E. Fouts,[3] Samuel Levy,[2] Anthony H. Knap,[6] Michael W. Lomas,[6] Ken Nealson,[5] Owen White,[3] Jeremy Peterson,[3] Jeff Hoffman,[1] Rachel Parsons,[6] Holly Baden-Tillson,[1] Cynthia Pfannkoch,[1] Yu-Hui Rogers,[4] Hamilton O. Smith[1]

We have applied "whole-genome shotgun sequencing" to microbial populations collected en masse on tangential flow and impact filters from seawater samples collected from the Sargasso Sea near Bermuda. A total of 1.045 billion base pairs of nonredundant sequence was generated, annotated, and analyzed to elucidate the gene content, diversity, and relative abundance of the organisms within these environmental samples. These data are estimated to derive from at least 1800 genomic species based on sequence relatedness, including 148 previously unknown bacterial phylotypes. We have identified over 1.2 million previously unknown genes represented in these samples, including more than 782 new rhodopsin-like photoreceptors. Variation in species present and stoichiometry suggests substantial oceanic microbial diversity.





An obesity-associated gut microbiome with increased capacity for energy harvest

Peter J. Turnbaugh[1], Ruth E. Ley[1], Michael A. Mahowald[1], Vincent Magrini[2], Elaine R. Mardis[1,2] & Jeffrey I. Gordon[1]

Supplementary Table 3 – Assembly of reads from capillary sequencer and pyrosequencer datasets.

Sample	Contigs	Average contig length	Contiged bases[1]	Largest Assembly	N50 contig length (kb)[2]
lean1 (GS20)	102,299	117	11,966,580	2,793	0.109
ob1 (GS20)	56,425	116	6,518,469	2,174	0.109
lean1 (3730xl)	167	1527	254,985	5,500	1.62
lean2 (3730xl)	407	1598	650,499	5,522	1.71
lean3 (3730xl)	224	1528	342,172	3,281	1.59
ob1 (3730xl)	320	1393	445,814	3,225	1.49
ob2 (3730xl)	269	1644	442,210	4,186	1.70
All (3730xl)	2,575	1734	4,465,685	11,213	1.78
All (GS20)	159,245	118	18,809,438	2,708	0.110
All (GS20 and 3730xl)	13,667	898	12,275,469	14,755	0.903

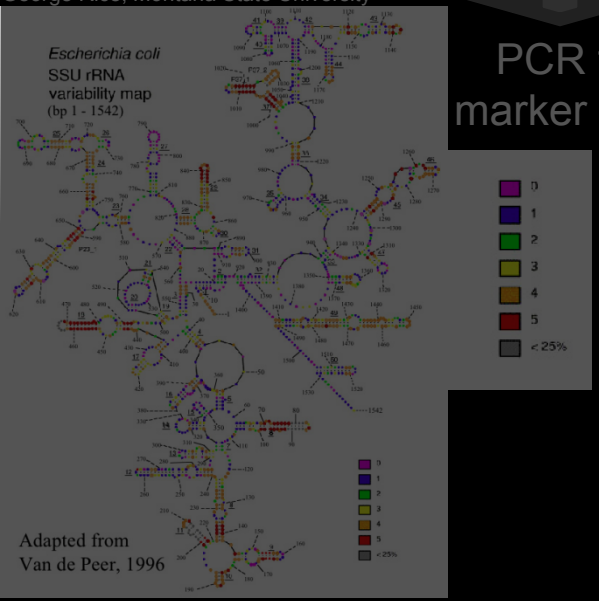
Sequencing as a tool for microbial community analysis

Lyse cells
Extract DNA (and/or RNA)

Amplicons

PCR to amplify a single marker gene, e.g. 16S rRNA

George Rice, Montana State University



Hello
my name is

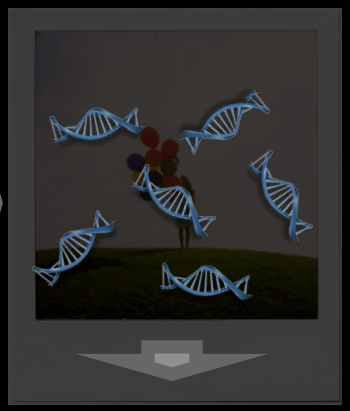
Classify sequence
➔ microbe

Samples

Microbes

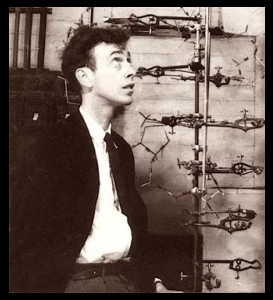
Relative abundances

Meta'omic



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

What to do with your metagenome?



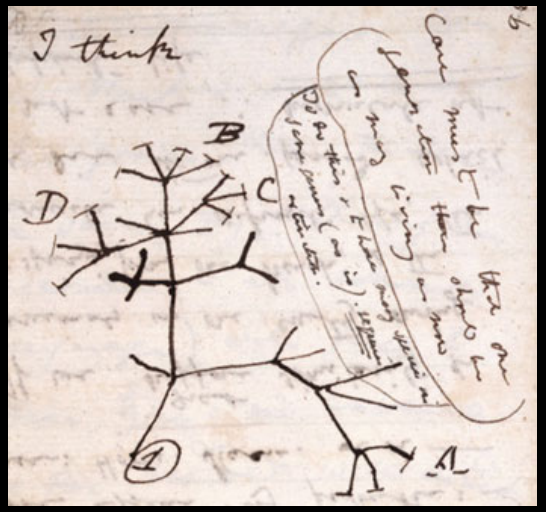
Reservoir of gene and protein functional information

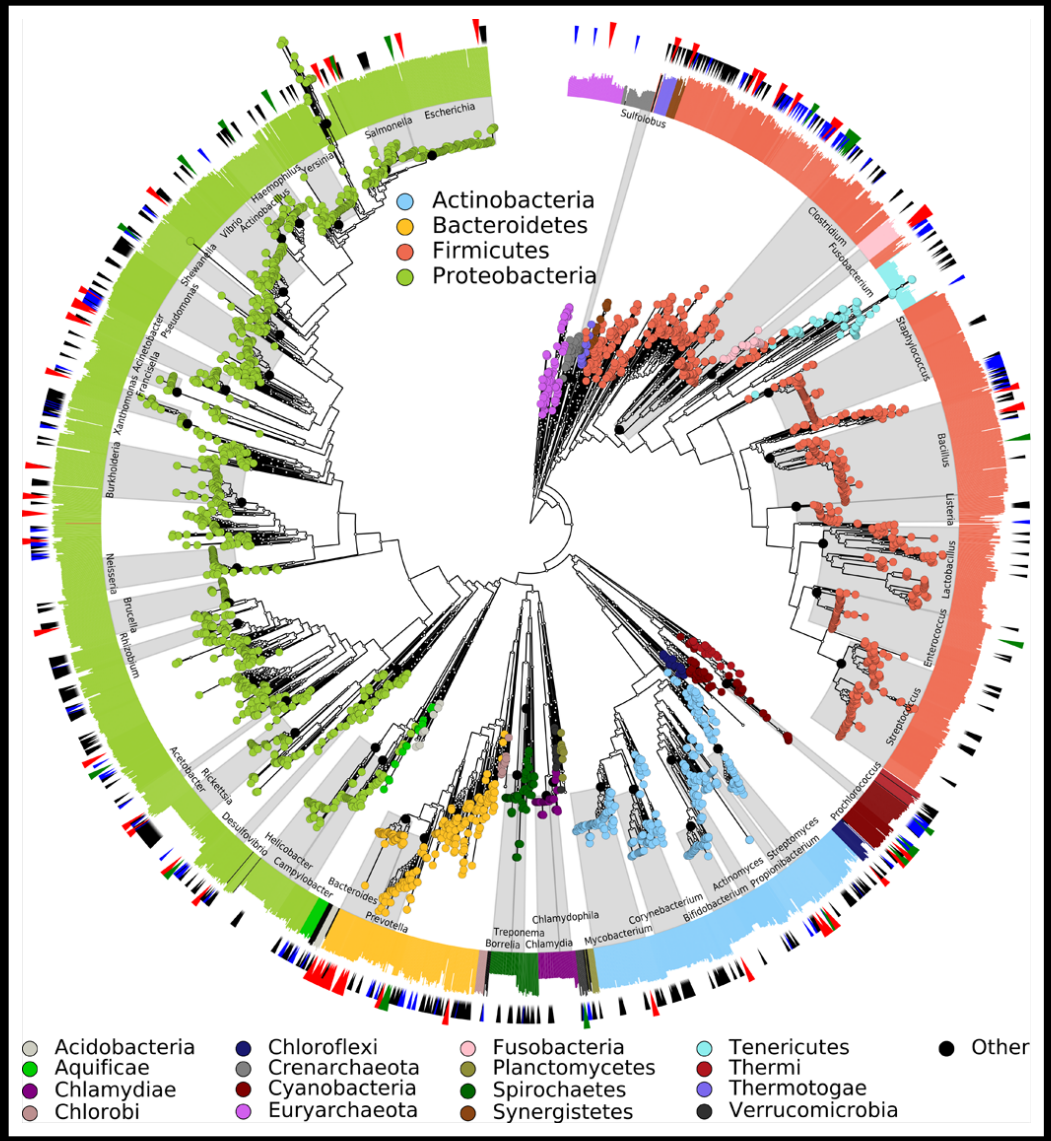
Comprehensive snapshot of microbial ecology and evolution

Public health tool monitoring population health and epidemiology

Diagnostic or prognostic biomarker for host disease

Composition-based analyses





Microbiome composition analyses: phylotypes and binning

Binning: nontrivial assignment of reads to phylotypes

RIBOSOMAL DATABASE PROJECT

rdp.cme.msu.edu

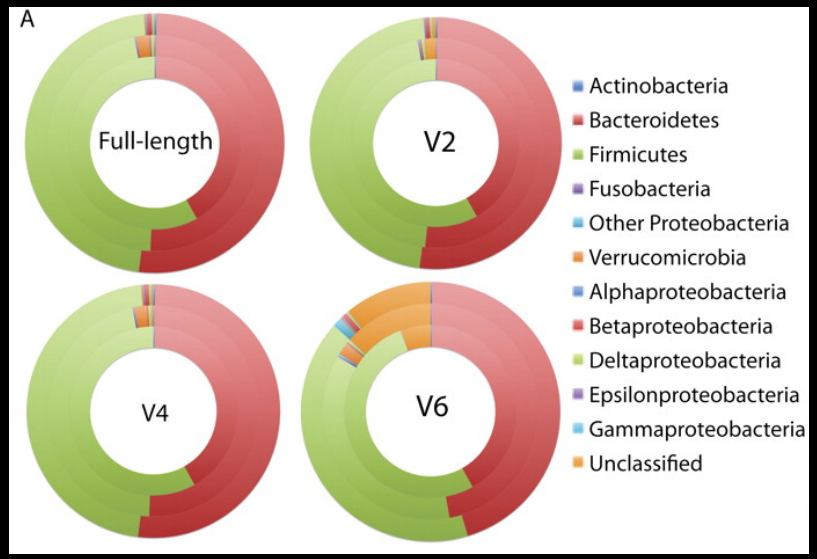
green genes
16S rRNA gene database and workbench compatible with ARB
greengenes.lbl.gov

greengenes.lbl.gov

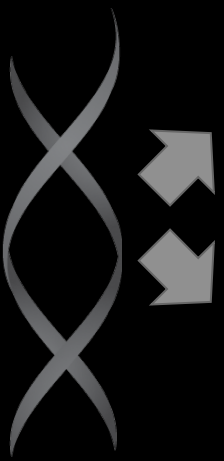
silva
comprehensive ribosomal RNA databases

www.arb-silva.de

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



Hamady, 2009



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

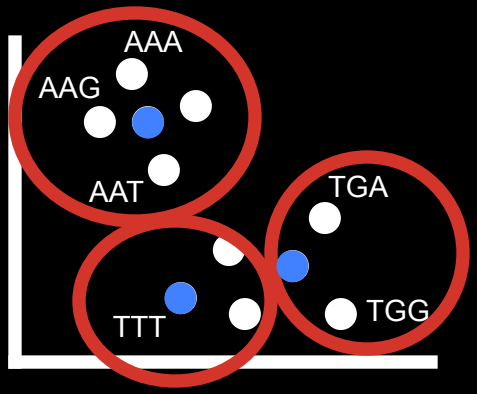
Direct binning: analyzes seq. characteristics (GC, codons, etc.) Relies on long reads

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

Microbiome composition analyses: OTU clustering

Open reference

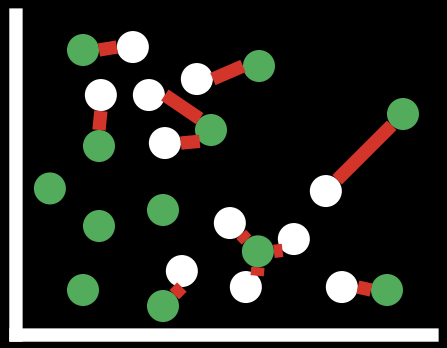
Clustering




```
>Uniq1
AAA
>Uniq2
TGA
>Uniq3
TTT
...
```

Closed reference

Classification



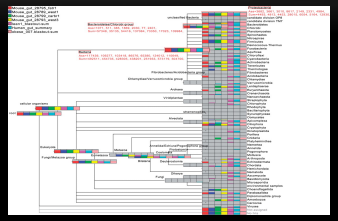
green genes

16S rRNA gene database and workbench compatible with ARB

greengenes.lbl.gov

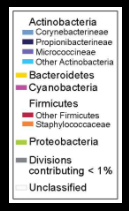
Microbiome composition analyses: diversity

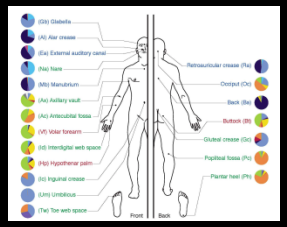
Mitra, 2009



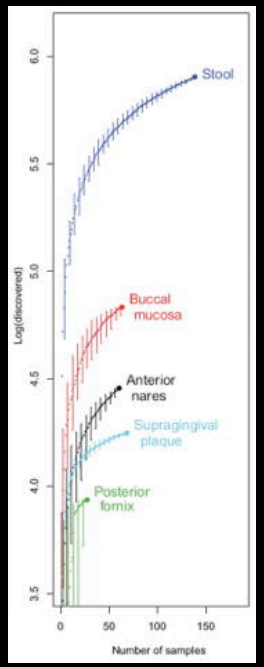
Diversity: broadly, a community's number and distribution of taxa

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





Coupon collector's problem or **rarefaction curve**: estimate population diversity based on a subsample

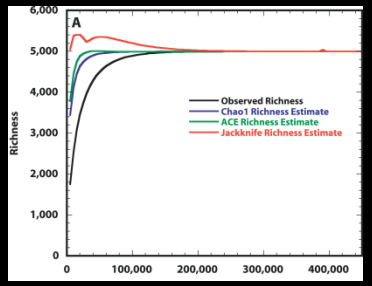


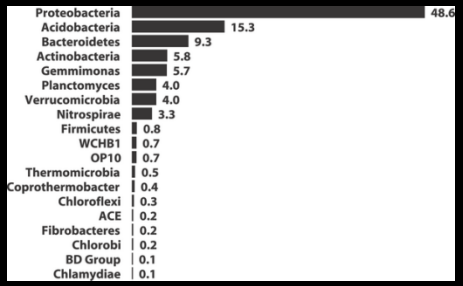
Chimeric reads

Hamady, 2009

[1 1 0 0]

[10 6 1 4]

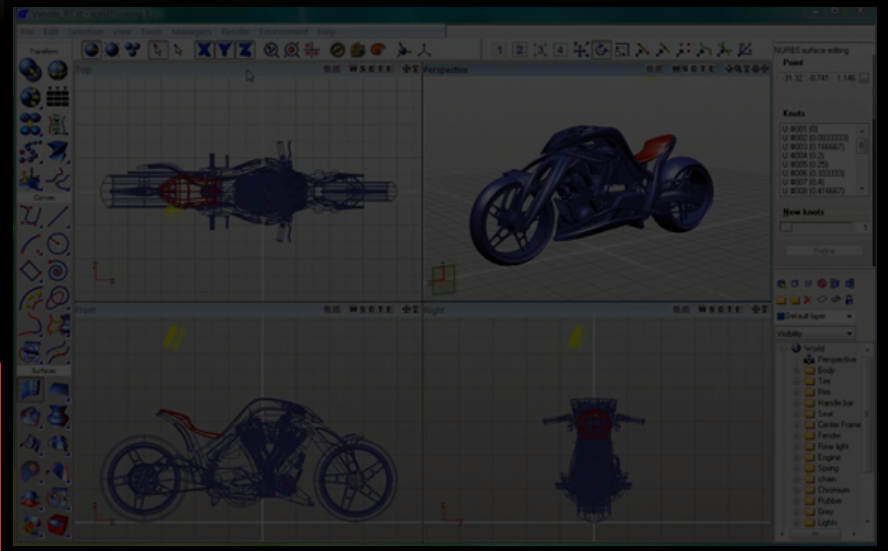




Schloss, 2006

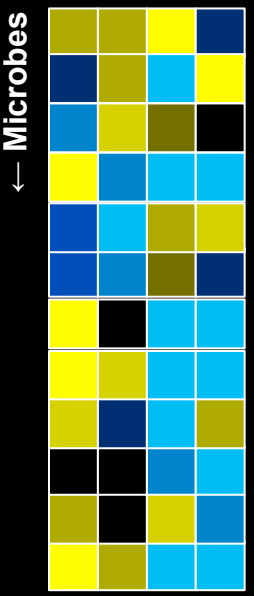
Microbiome composition analyses: ordination

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

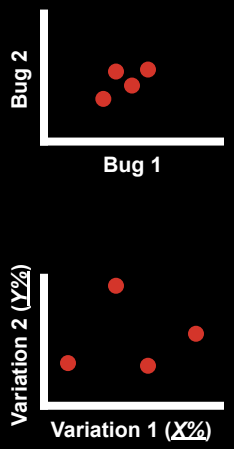


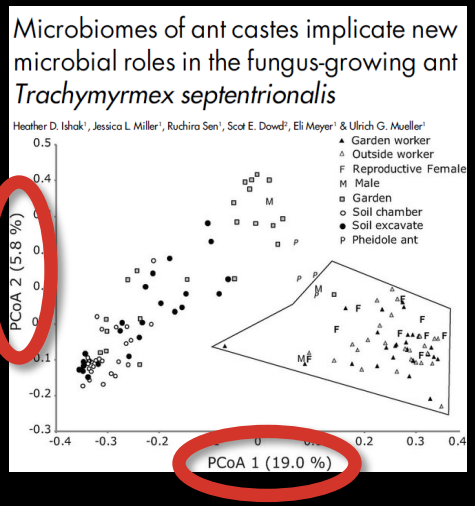
PCA or **Principal Component Analysis** guarantees the new dimensions maximize normal variation.

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).

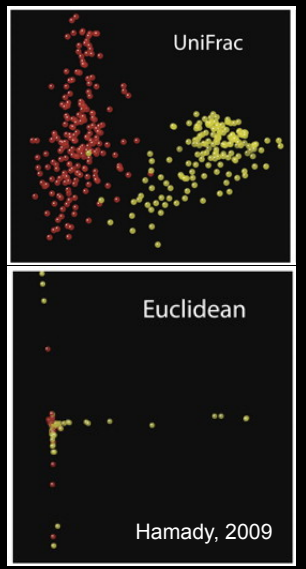


Distance between points is Euclidean

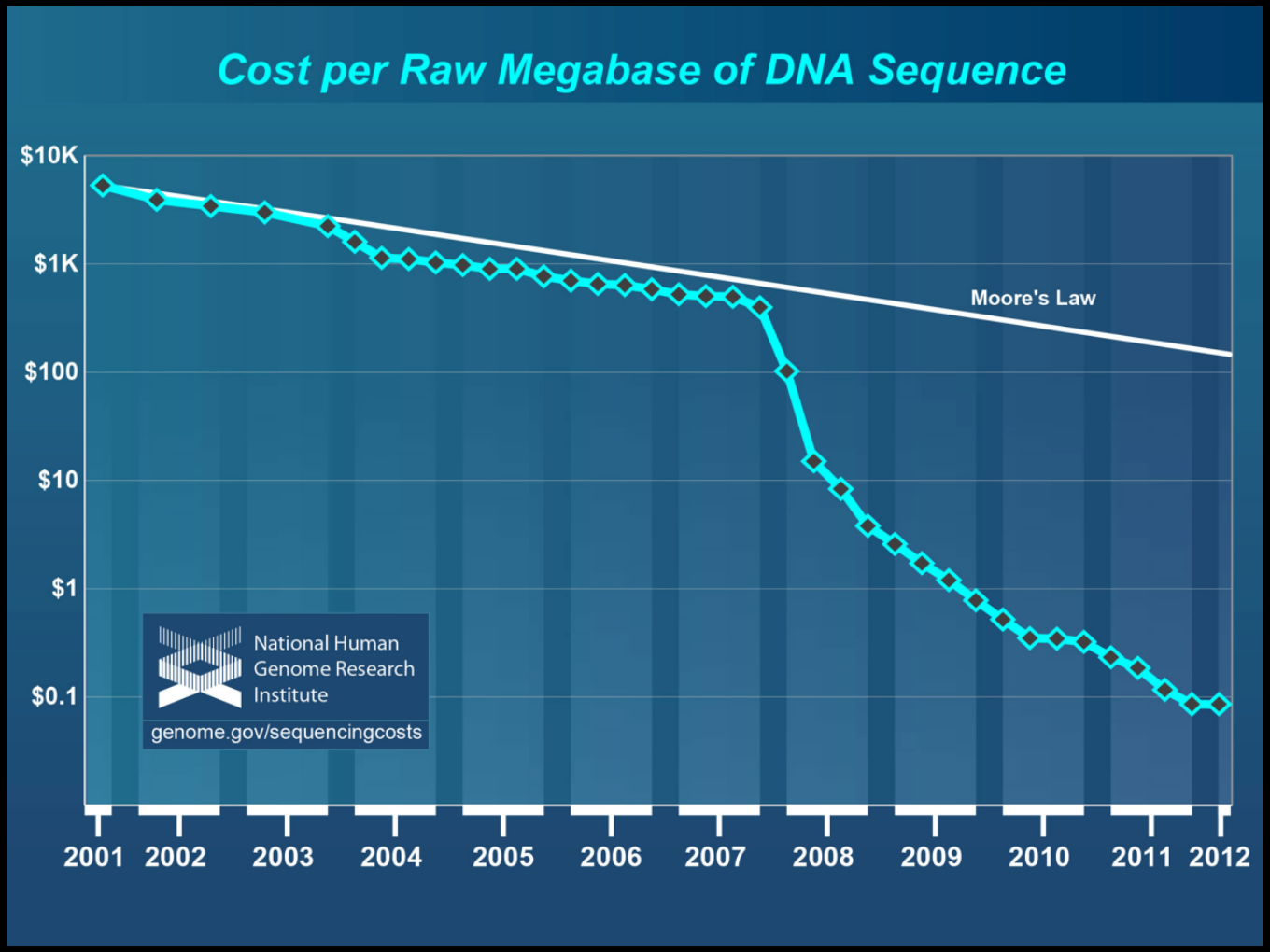




Distance between points is a ***proportional function*** of their ***similarity***

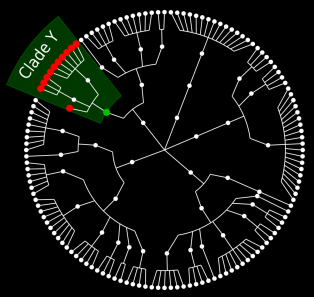


Meta'omic analyses



Typical shotgun metagenome and metatranscriptome analyses

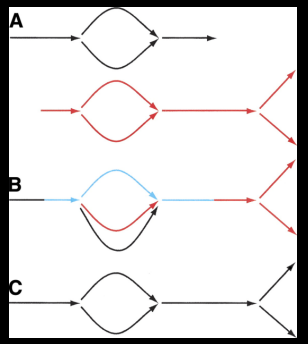
Taxonomic Profiling

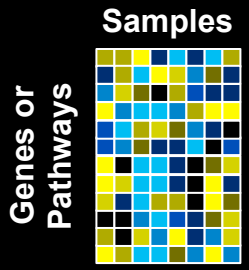


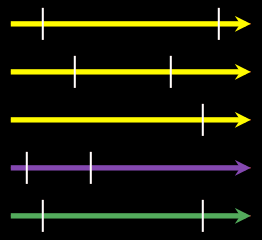
Functional Profiling

eggNOG4.0

Assembly

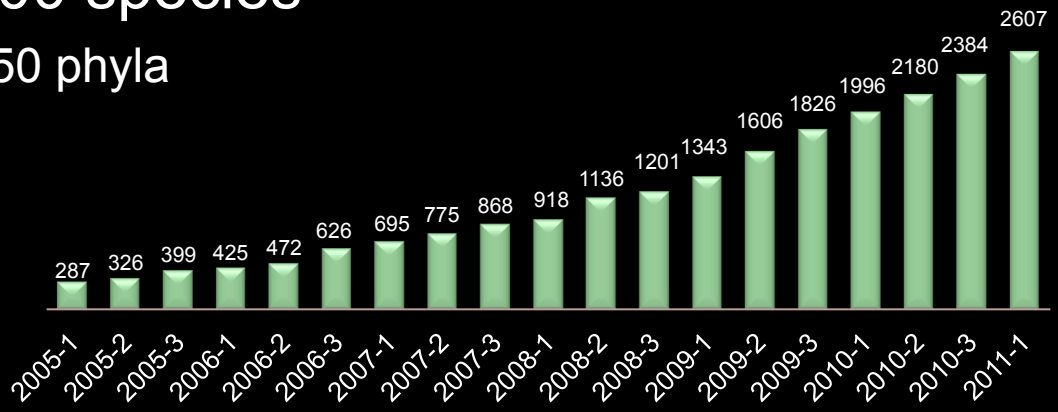






What can you do with thousands of microbial genomes that you can't do with one?

- NCBI now contains ~17,000 bacterial genomes
  - Plus ~300 archaeal, ~200 eukaryotic, and a few thousand viruses
  - About half final and half draft
- These comprise about 4,100 species
  - >1,200 genera, >380 families, 50 phyla
- **And roughly 55M genes**

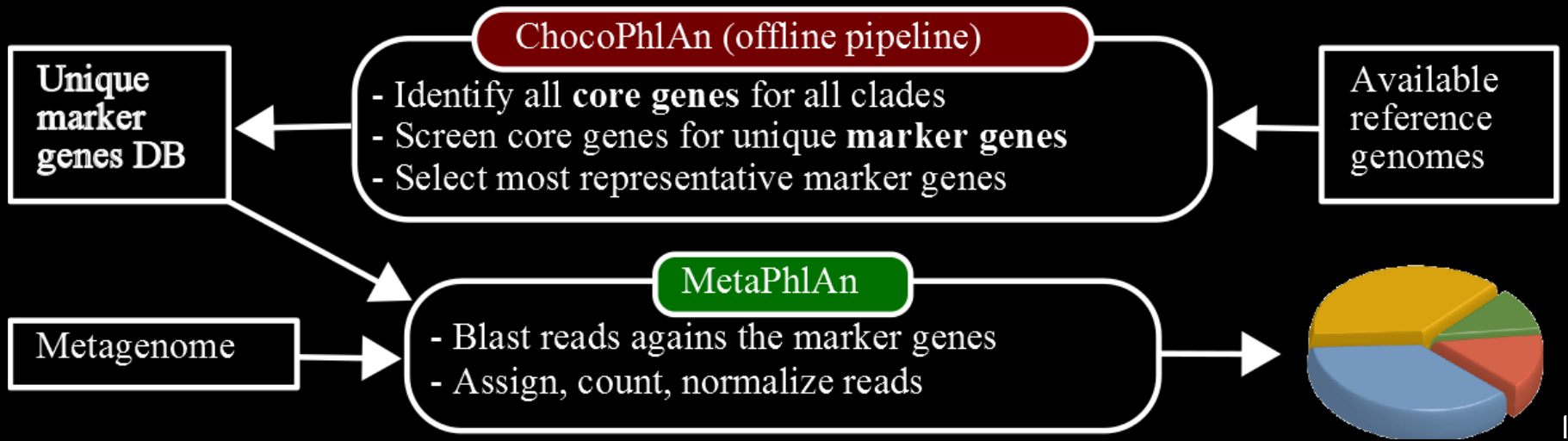


- These genes and genomes are a tremendous resource to:
  - Identify conserved markers that can be used to infer phylogeny
  - Identify unique markers that can be used to infer taxonomy
  - Relate the microbial members of a community to their metagenomic functional potential

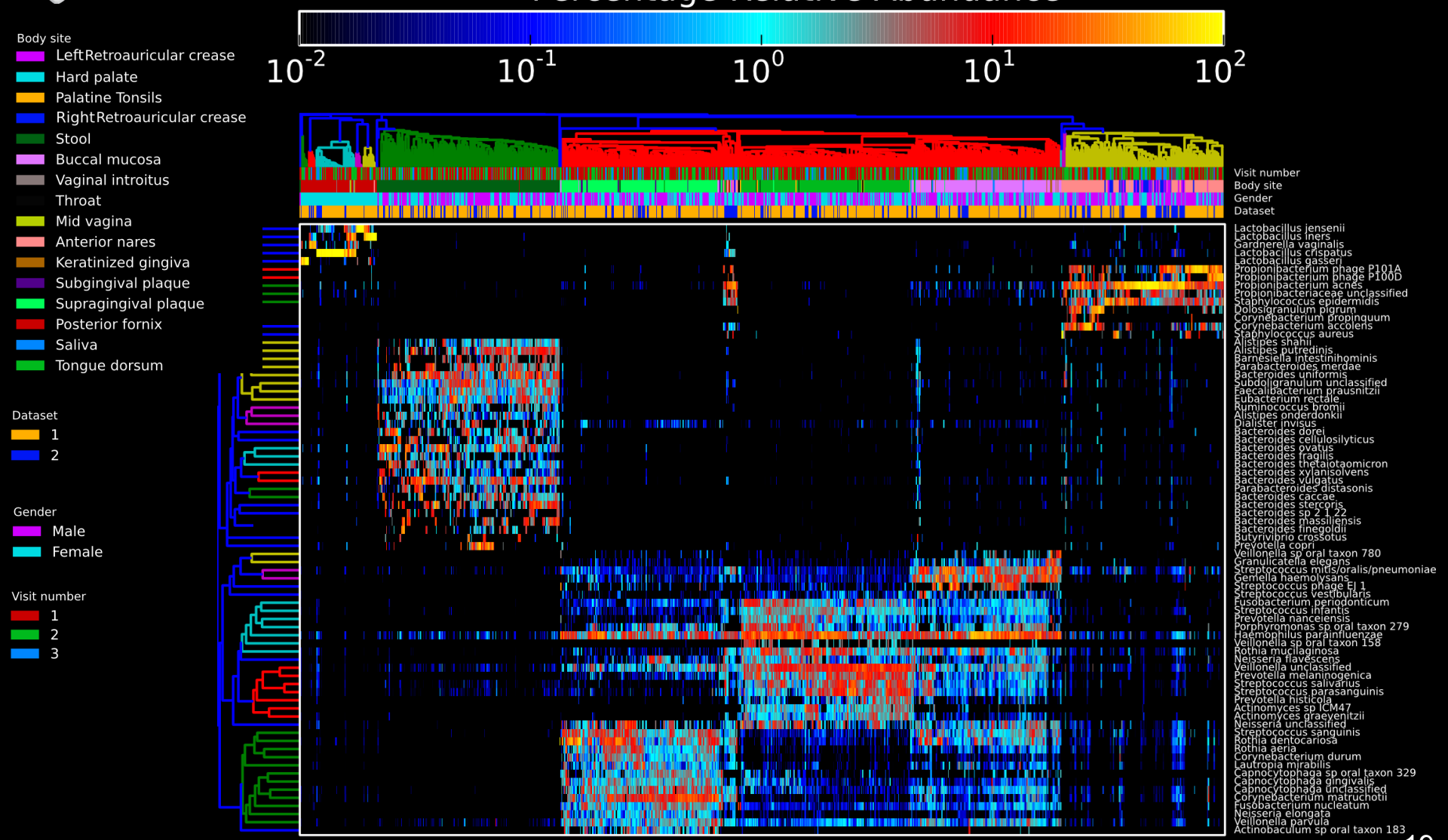
MetaPhlAn2: Taxonomic profiling using unique marker gene sequences

- ~1M most representative markers used for identification
  - 184±45 markers per species (target 200)
- ~7,100 species (excludes incomplete annotations, spp., etc.)
- FP/FN rates of ~1 in 10^6

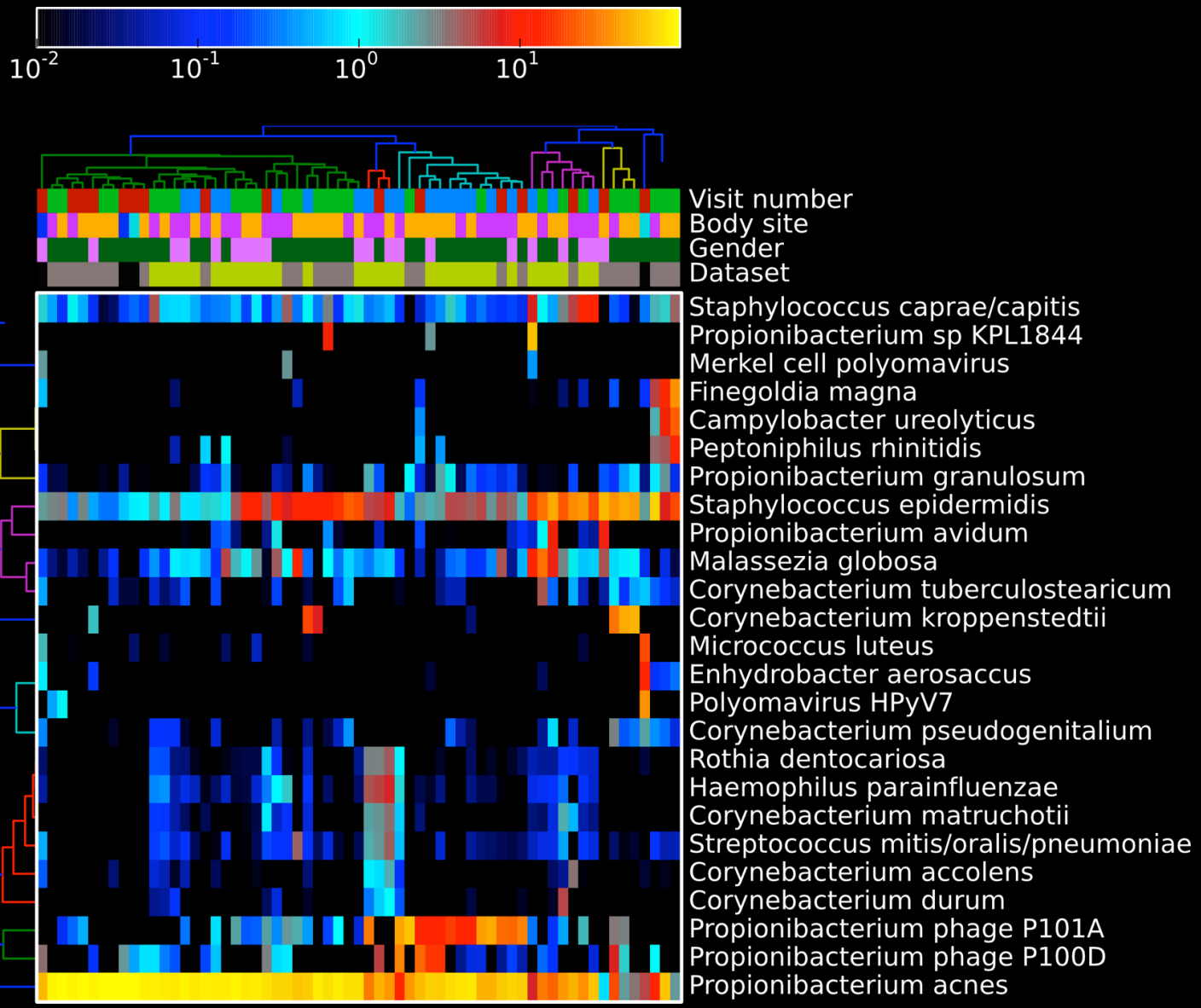
- Profiles all domains of life: bacteria, viruses, euks, archaea
- Strain level profiling using marker combination barcodes
- Quasi-markers used to resolve ambiguity in postprocessing



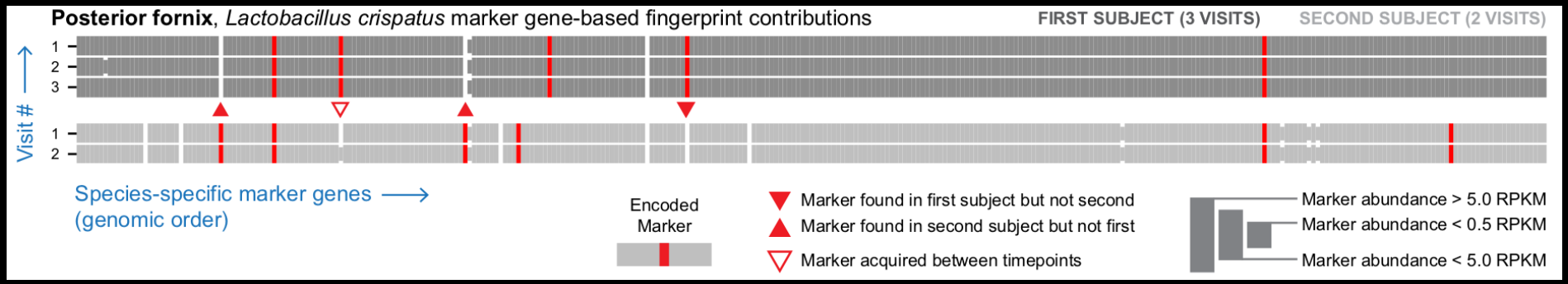
MetaPhlAn2: Taxonomic profiling using unique marker gene sequences

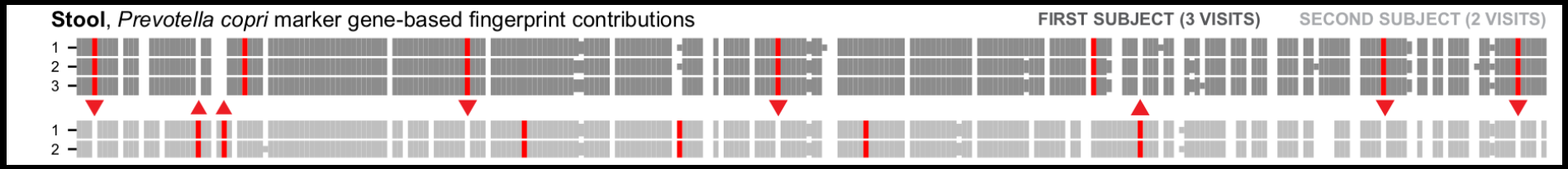


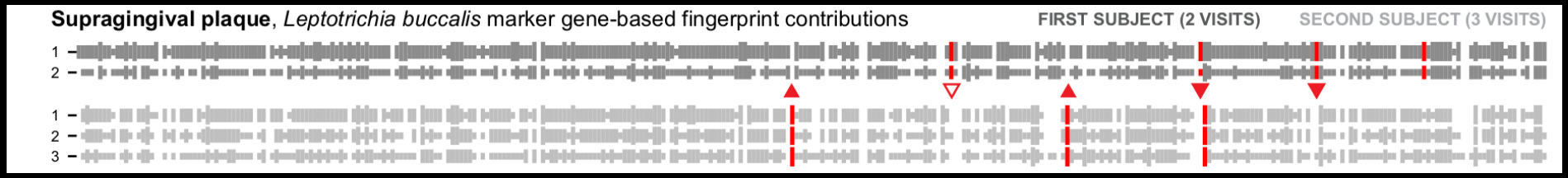
MetaPhlAn2: Taxonomic profiling using unique marker gene sequences



Gene-based fingerprints capture strain variation in individuals' most abundant (stable) bugs

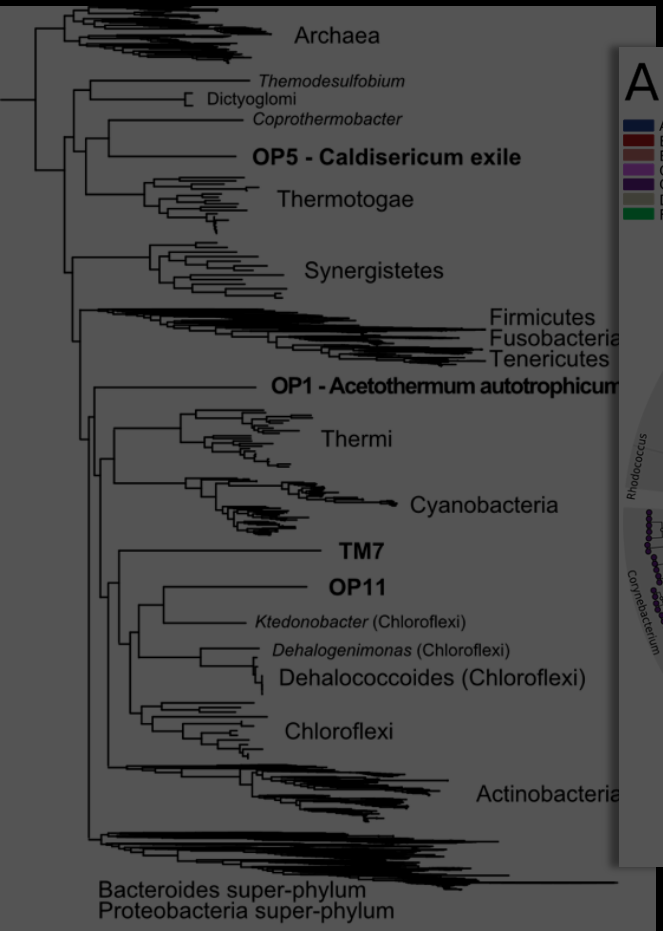


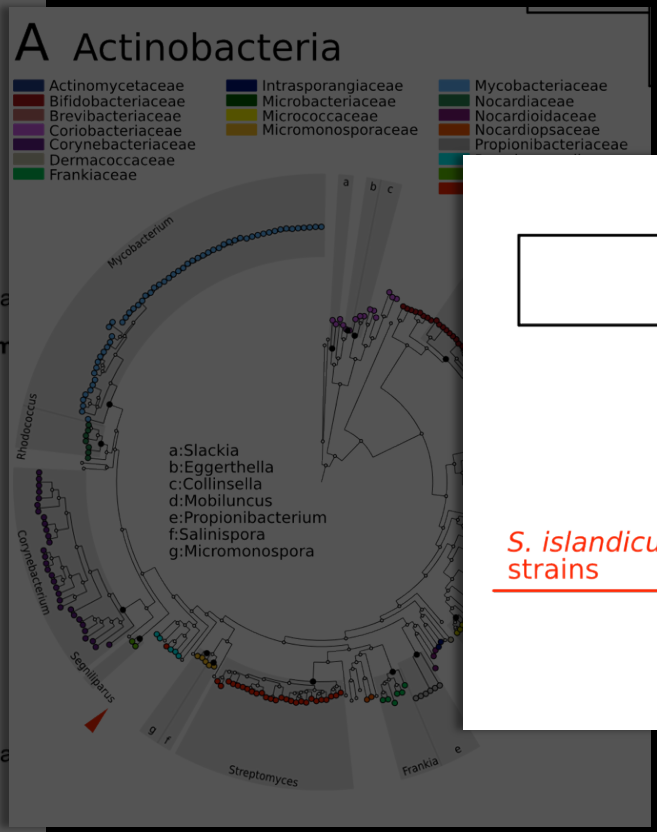


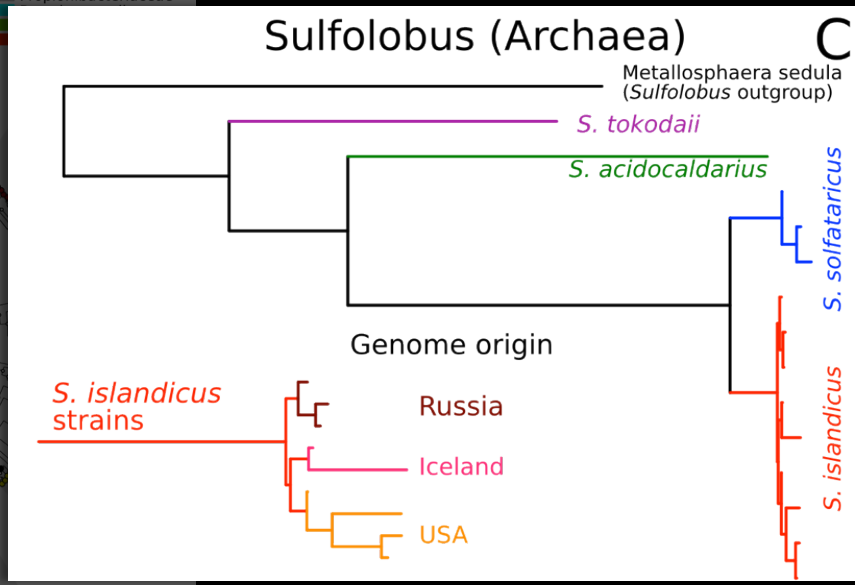


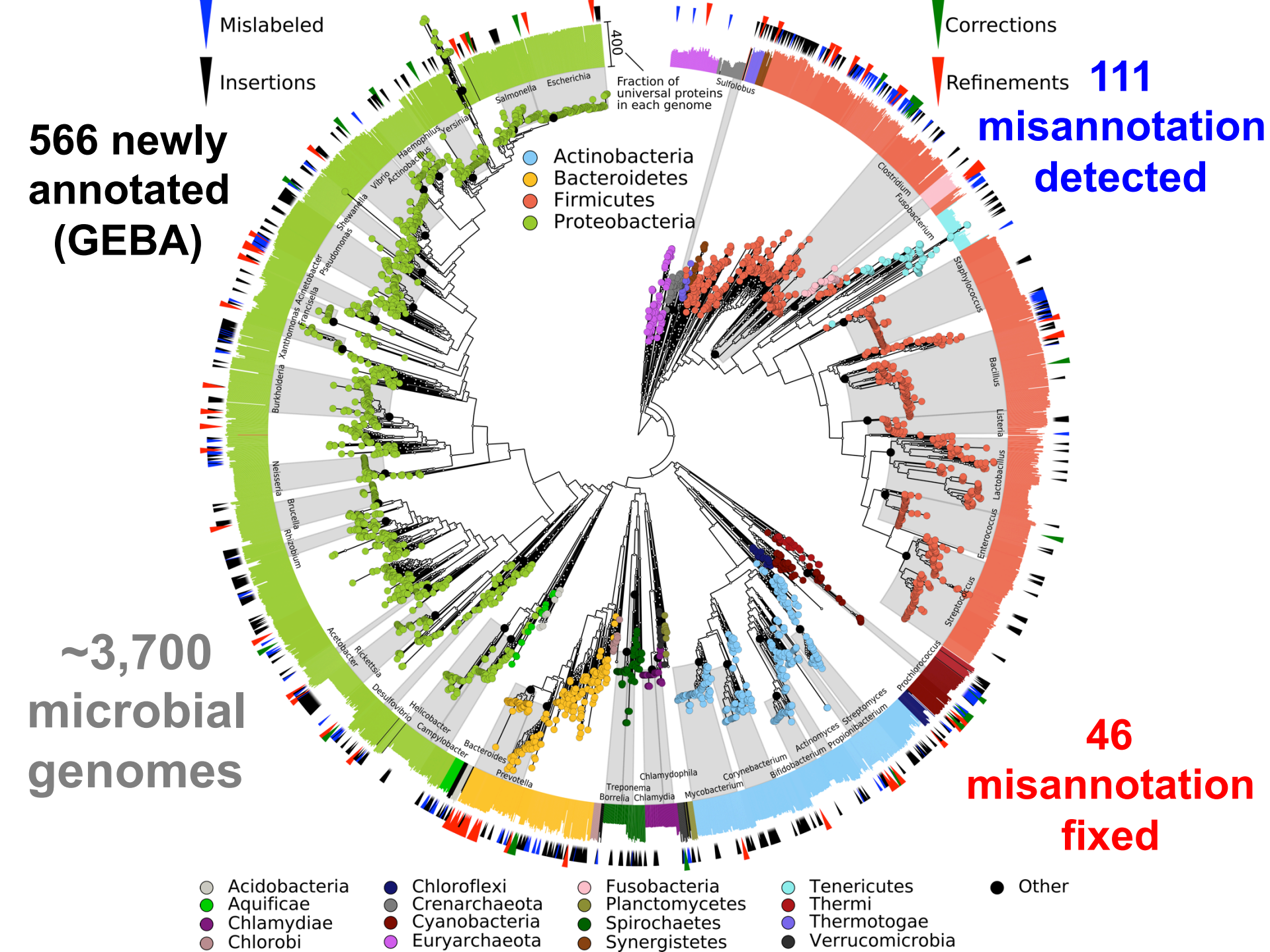
PhyloPhlAn: From markers for taxonomy to markers for phylogeny

- Hundreds of unique markers per clade provide great taxonomic classification
- What if we use hundreds of *conserved* markers for phylogenetic classification?
  - PhyloPhlAn identifies the most informative residues of the most conserved 400 proteins
  - These can then be used for phylogenetic reconstruction, placement, and taxonomy



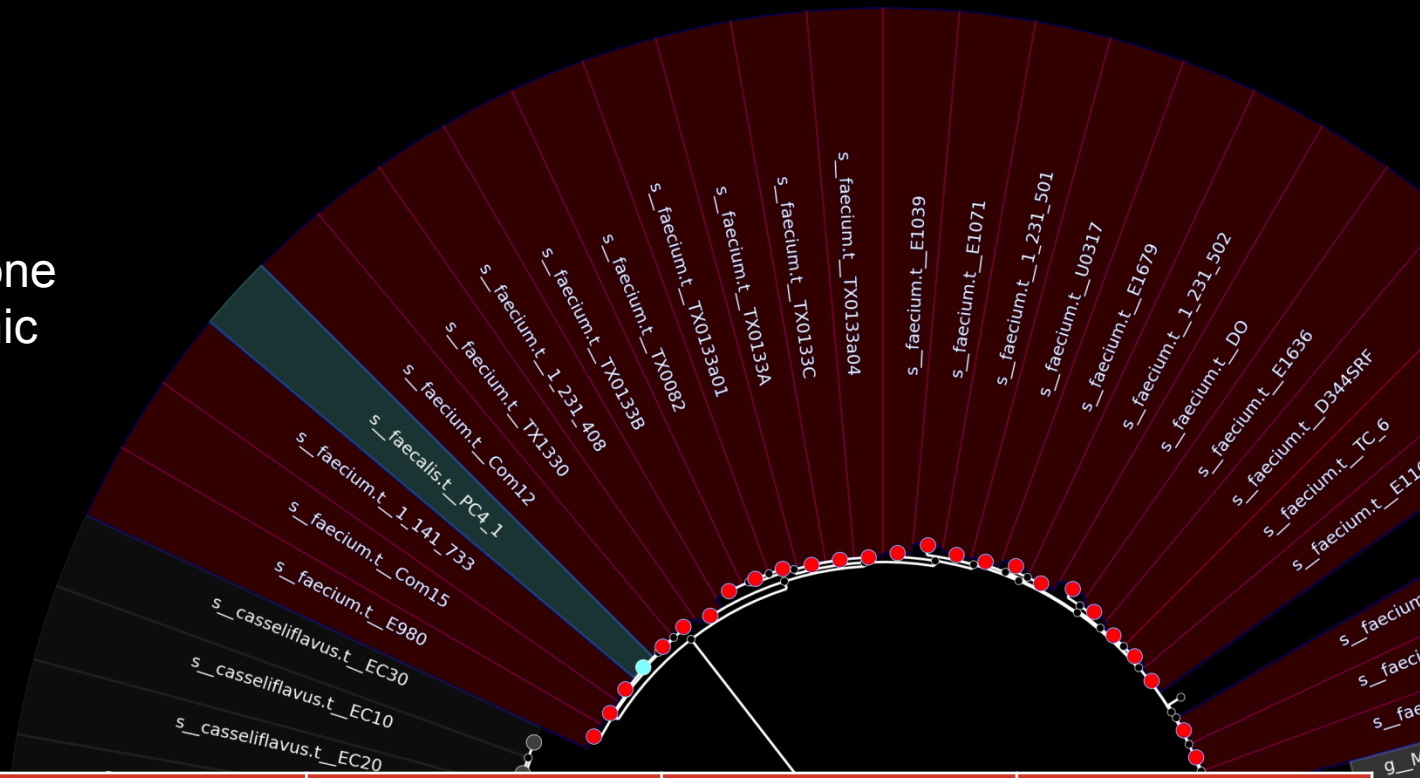





Mislabeled
Insertions
Corrections
Refinements
566 newly annotated (GEBA)
111 misannotation detected
400
Fraction of universal proteins in each genome
Actinobacteria
Bacteroidetes
Firmicutes
Proteobacteria
Salmonella
Escherichia
Yersinia
Haemophilus
Actinobacillus
Vibrio
Shewanella
Pseudomonas
Acinetobacter
Xanthomonas
Francisella
Burkholderia
Neisseria
Brucella
Rhizobium
Acetobacter
Rickettsia
Desulfovibrio
Helicobacter
Campylobacter
Bacteroides
Prevotella
Treponema
Borrelia
Chlamydophila
Chlamydia
Mycobacterium
Corynebacterium
Bifidobacterium
Actinomyces
Streptomyces
Propionibacterium
Prochlorococcus
Streptococcus
Enterococcus
Lactobacillus
Listeria
Bacillus
Staphylococcus
Fusobacterium
Clostridium
Sulfolobus
~3,700 microbial genomes
46 misannotation fixed
Acidobacteria
Aquificae
Chlamydiae
Chlorobi
Chloroflexi
Crenarchaeota
Cyanobacteria
Euryarchaeota
Fusobacteria
Planctomycetes
Spirochaetes
Synergistetes
Tenericutes
Thermi
Thermotogae
Verrucomicrobia
Other

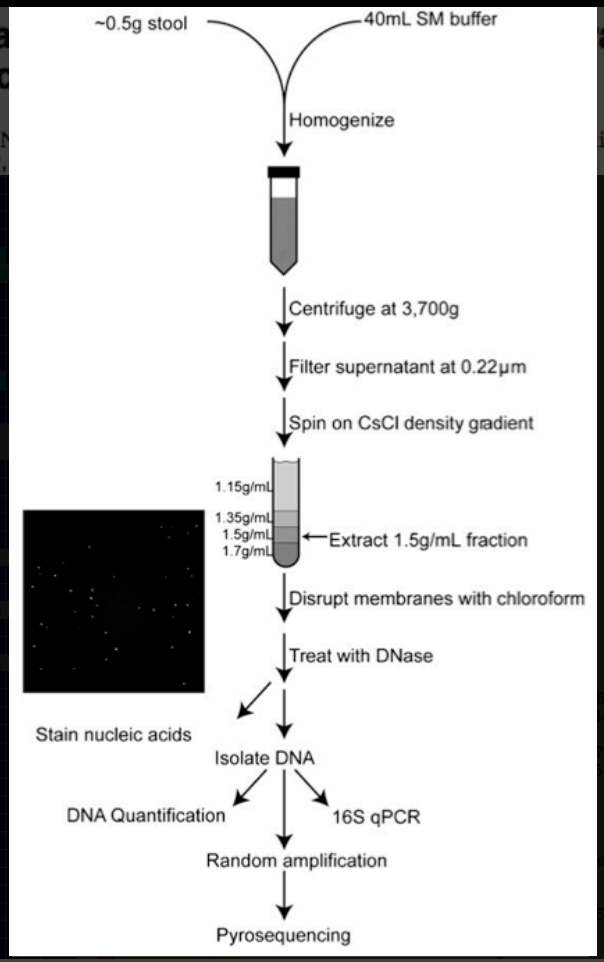
PhyloPhlAn: Taxonomic curation and reannotation

- Taxa with at least one ‘unknown’ taxonomic level: 445
- Additional taxa we detected as suspicious: 111



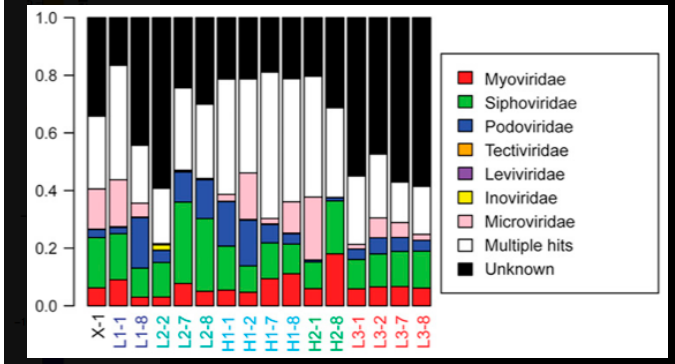
	Example	Very high confidence	High confidence	Medium confidence
Corrected	A|B|C→A|B|D	26	3	26
Refined	A|B|?→A|B|C	67	25	75
Removed	A|B|C→A|B|?	11	1	1
Incomplete	A|?|?→A|?|?	224	10	66

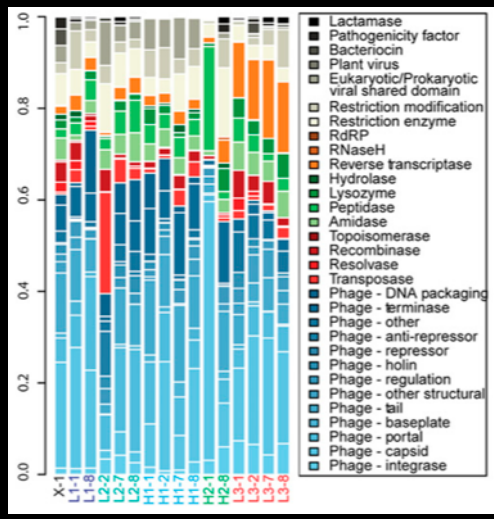
Man cannot live on bacteria alone – don't forget the viruses and eukaryotes!

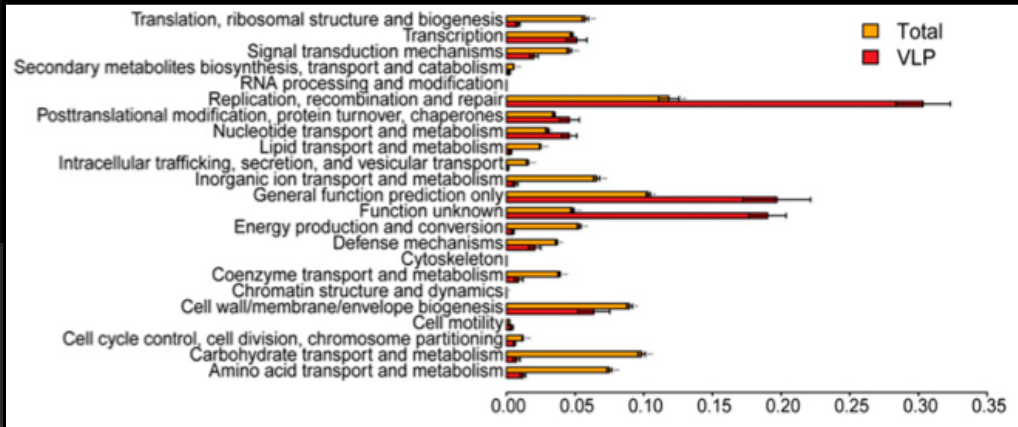


The human gut virome: Inter-individual variation and dynamic response to diet

Samuel Minot,[1] Rohini Sinha,[1] Jun Chen,[2] Hongzhe Li,[2] Sue A. Keilbaugh,[3] Gary D. Wu,[3] James D. Lewis,[2] and Frederic D. Bushman[1,4]







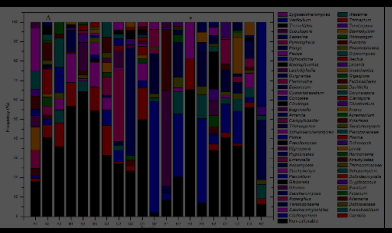
Viruses in the faecal microbiota of monozygotic twins and their mothers

Alejandro Reyes[1], Matthew Haynes[2], Nicole Hanson[2], Florent E. Angly[2,3], Andrew C. Heath[4], Forest Rohwer[2] & Jeffrey I. Gordon[1]

Man cannot live on bacteria alone – don't forget the viruses and eukaryotes!

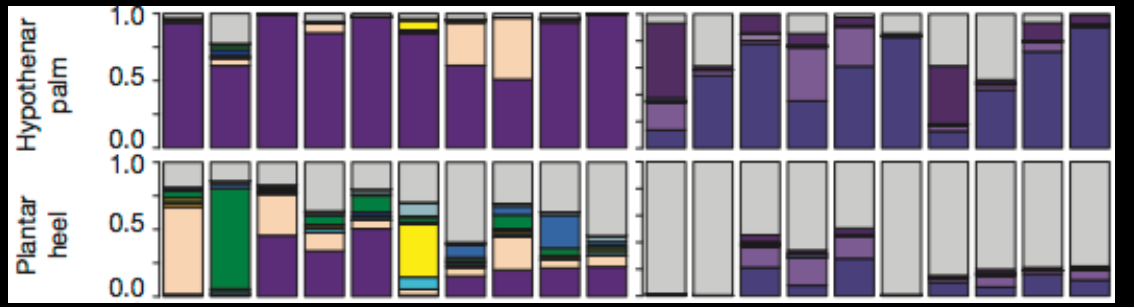
Interactions Between Commensal Fungi and the C-Type Lectin Receptor Dectin-1 Influence Colitis

Iliyan D. Iliev,[1] Vincent A. Funari,[2,3] Kent D. Taylor,[2] Quoclinh Nguyen,[2] Christopher N. Reyes,[1] Samuel P. Strom,[2] Jordan Brown,[2] Courtney A. Becker,[1] Phillip R. Fleshner,[4] Marla Dubinsky,[1,5] Jerome I. Rotter,[2] Hanlin L. Wang,[6] Dermot P. B. McGovern,[2,2] Gordon D. Brown,[7] David M. Underhill[3,6,8]*



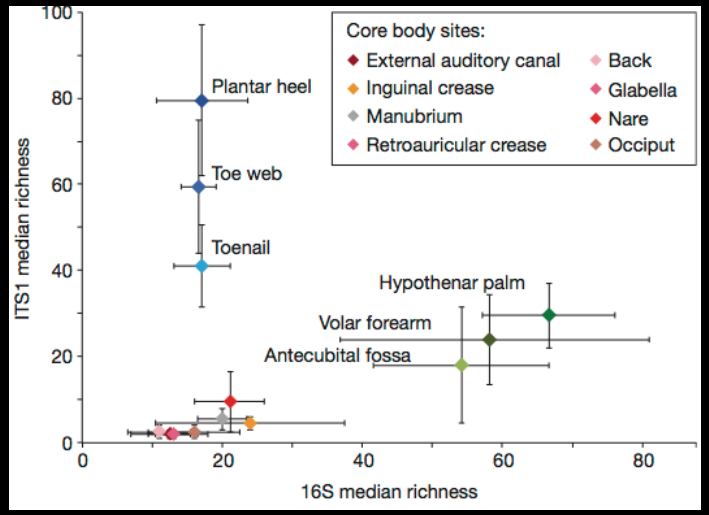
Characterization of the Oral Fungal Microbiome (Mycobiome) in Healthy Individuals

Mahmoud A. Ghannoum[1]*, Richard J. Jurevic[2], Pranab K. Mukherjee[1], Fan Cui[1], Masoumeh Sikaroodi[3], Ammar Naqvi[3], Patrick M. Gillevet[3]



Topographic diversity of fungal and bacterial communities in human skin

Keisha Findley[1], Julia Oh[1], Joy Yang[1], Sean Conlan[1], Clayton Deming[1], Jennifer A. Meyer[1], Deborah Schoenfeld[2], Effie Nomicos[2], Morgan Park[3], NIH Intramural Sequencing Center Comparative Sequencing Program†, Heidi H. Kong[2]* & Julia A. Segre[1]*

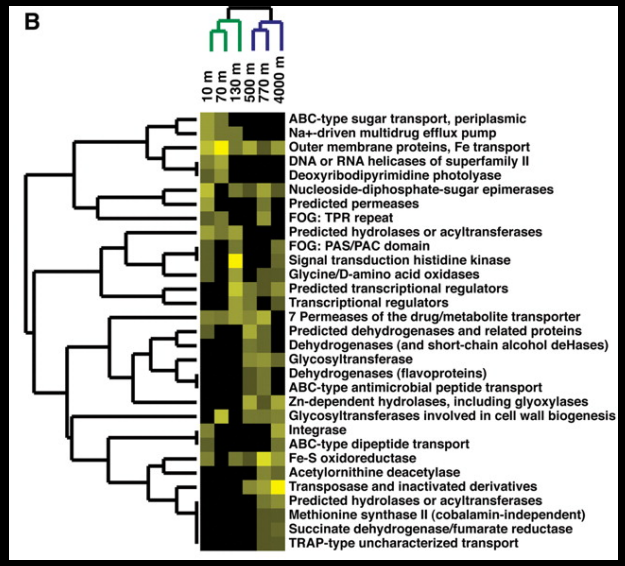


Microbiome meta'omic analyses: molecular functions and biological roles

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…

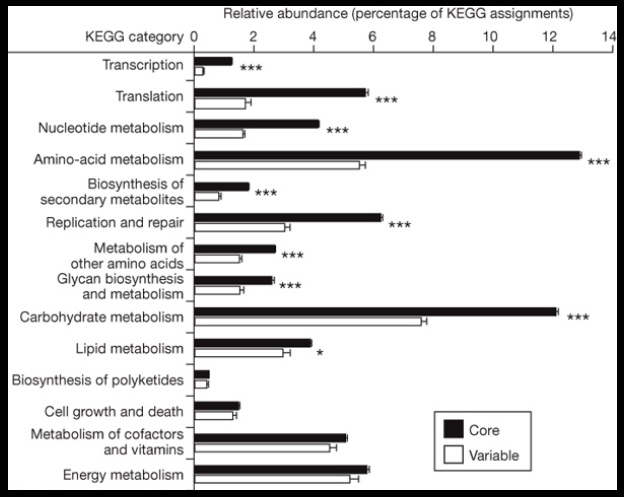


DeLong, 2006

Table 1 | Glycoside hydrolases and carbohydrate-binding modules

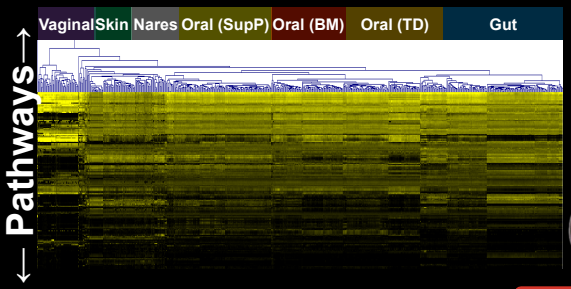
CAZy family*	Pfam HMM name†	Known activities‡	Termite gut community§
Glycoside hydrolase catalytic domains**††			
GH1	Glyco_hydro_1	β-Glucosidase, β-galactosidase, β-mannosidase, others	22
GH2	Glyco_hydro_2_C	β-Galactosidase, β-mannosidase, others	23
GH3	Glyco_hydro_3	β-1,4-Glucosidase, β-1,4-xylosidase, 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



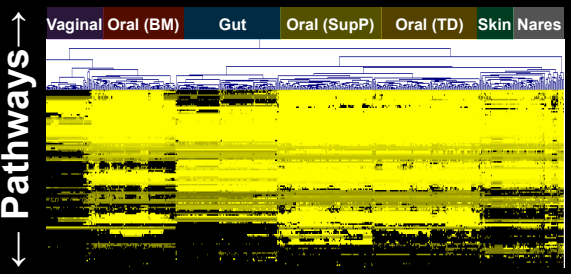
Turnbaugh, 2009

Microbiome meta'omic analyses: metabolic profiling (with HUMAnN)



Pathway abundance

Pathway coverage



100 subjects
1-3 visits/subject
~7 body sites/visit
10-200M reads/sample
100bp reads

BioCyc Database Collection

BLAST

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

Metagenomic reads

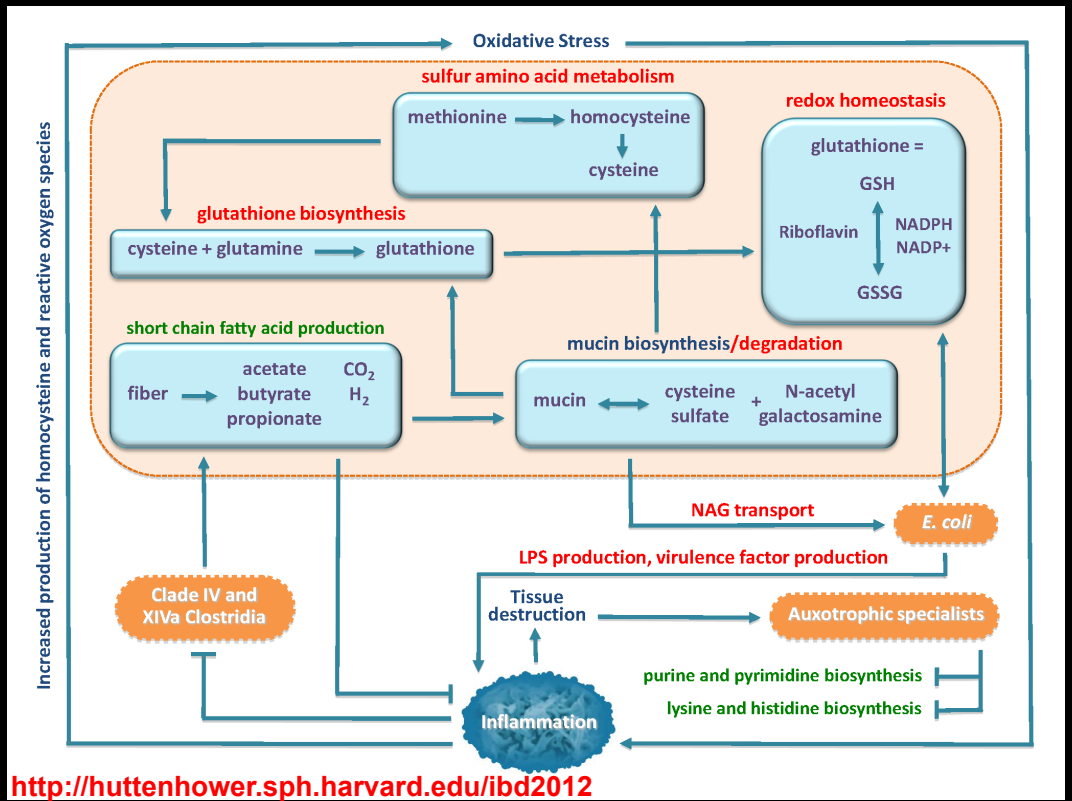
?

Enzymes and pathways

HUMAnN
HMP Unified Metabolic Analysis Network
http://huttenhower.sph.harvard.edu/humann

"Who's there," versus, "What they're doing," in the inflamed gut

- 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.

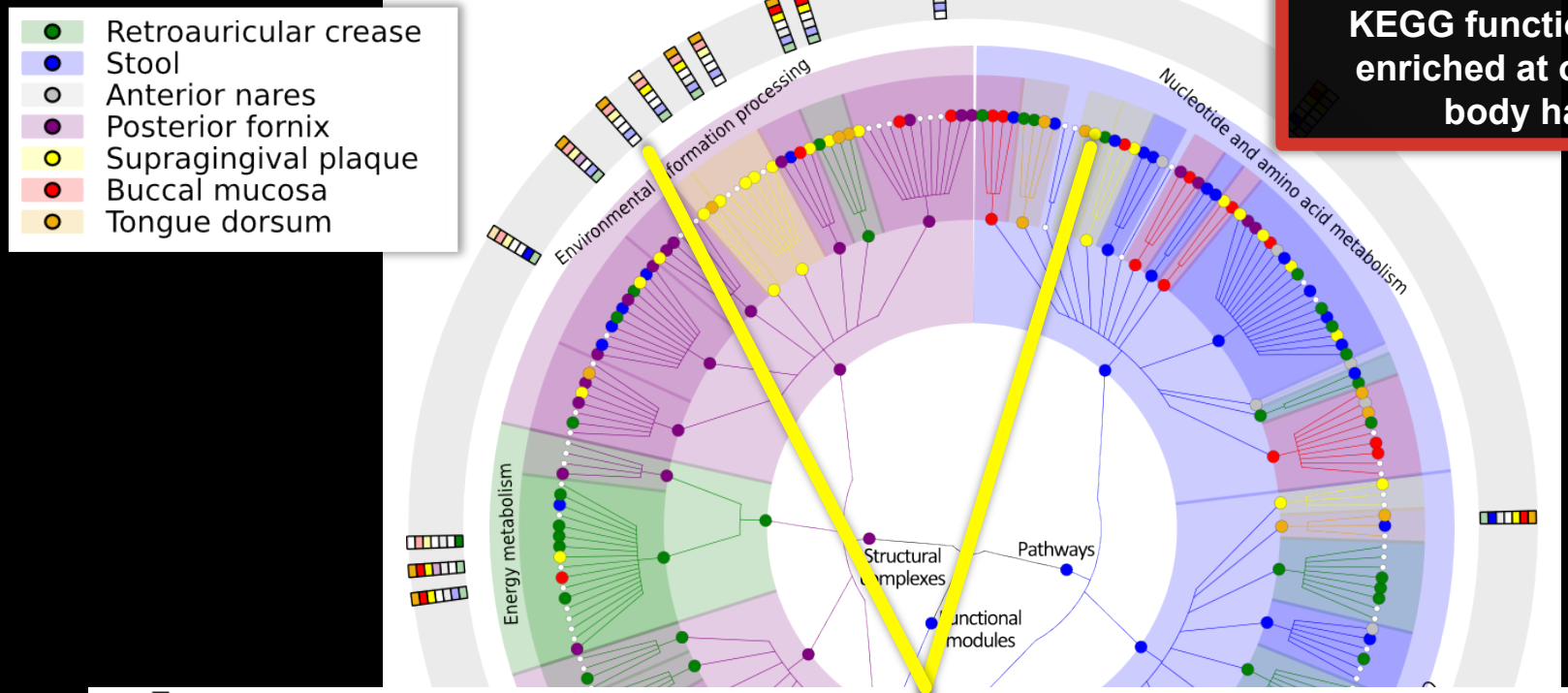


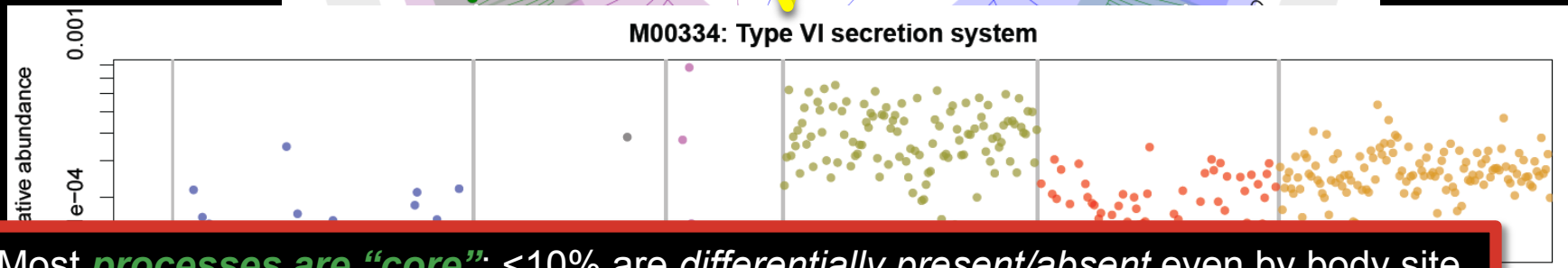
http://huttenhower.sph.harvard.edu/ibd2012

Microbiome meta'omic analyses: metabolic profiling (with HUMAnN)

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

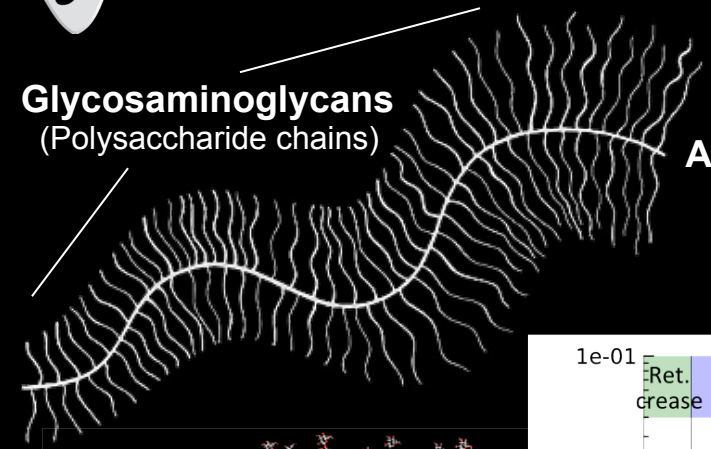
Metabolic modules in the KEGG functional catalog enriched at one or more body habitats

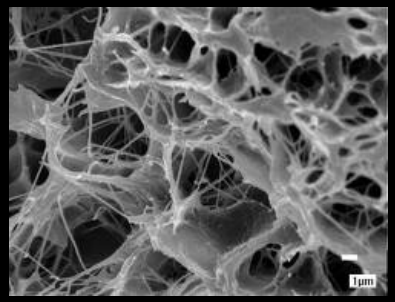


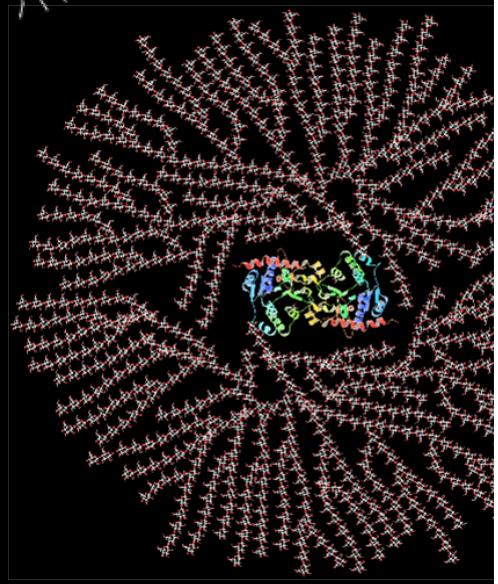


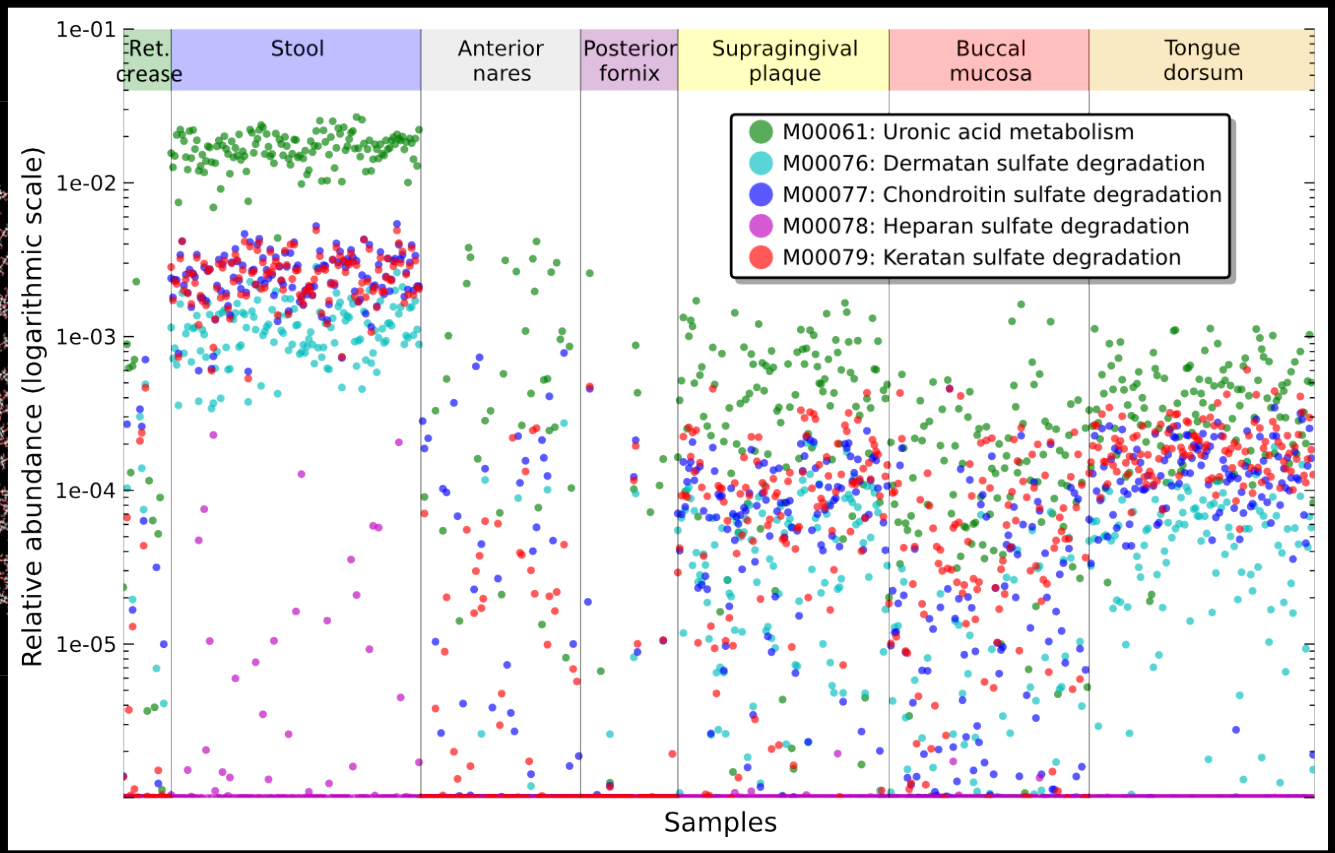
- Most ***processes are "core"***: <10% are *differentially present/absent* even by body site
  - Contrast ***zero*** microbes meeting this threshold!
- Most ***processes are habitat-adapted***: >66% are *differentially abundant* by body site

Proteoglycan degradation by the gut microbiota

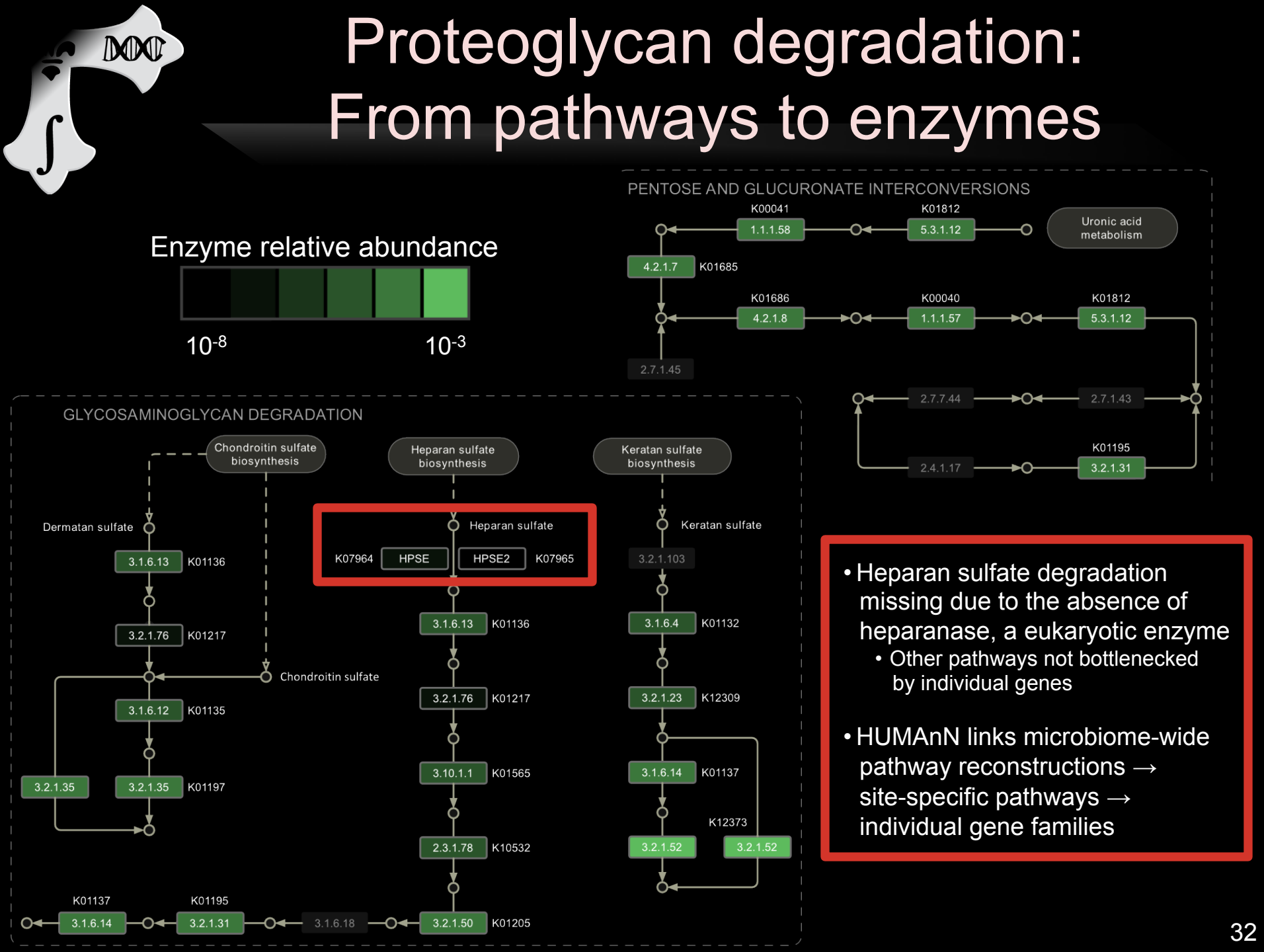






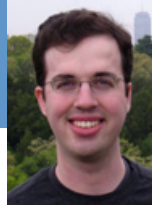


Proteoglycan degradation: From pathways to enzymes



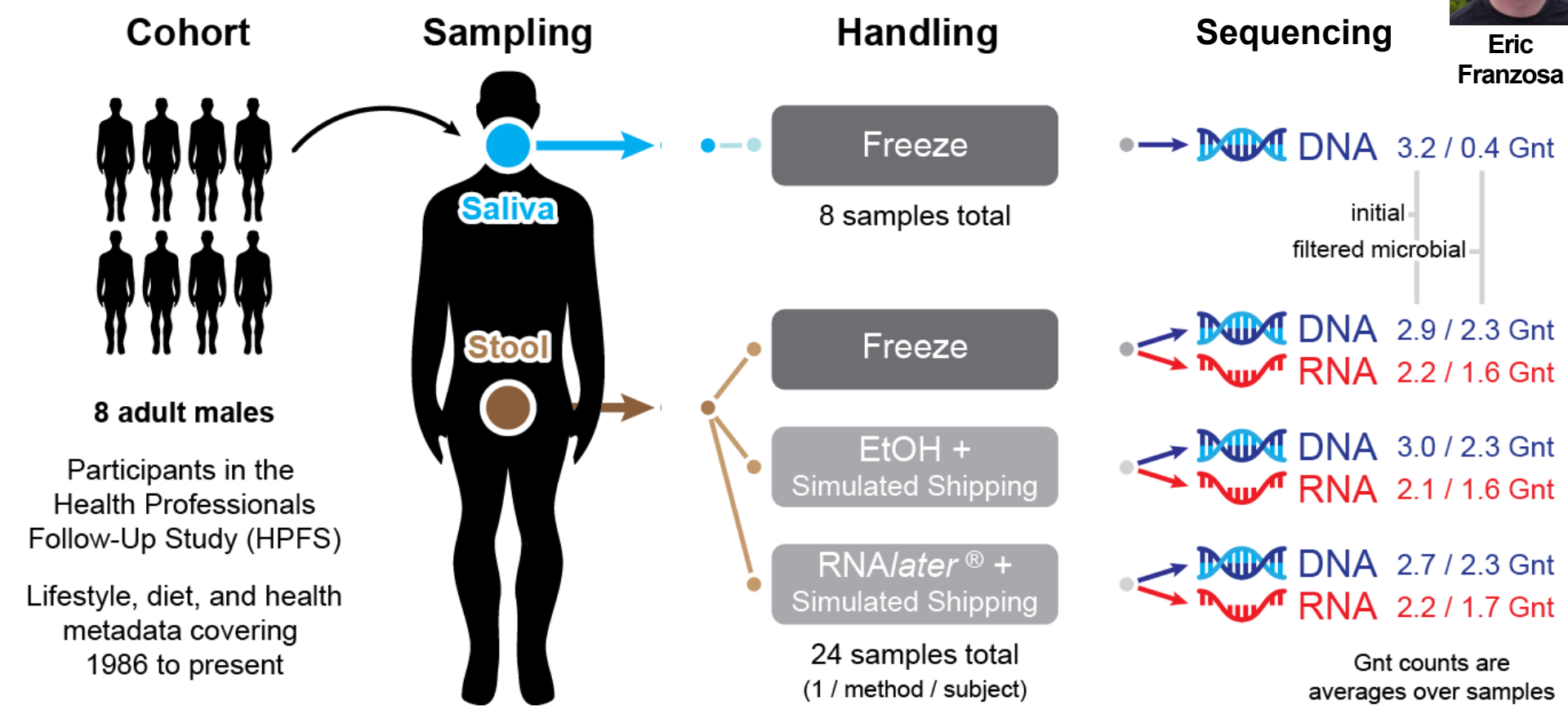
- Heparan sulfate degradation missing due to the absence of heparanase, a eukaryotic enzyme
  - Other pathways not bottlenecked by individual genes
- HUMAnN links microbiome-wide pathway reconstructions → site-specific pathways → individual gene families

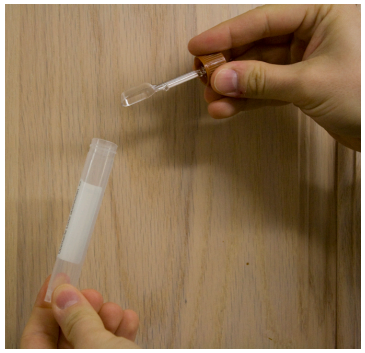
Functional meta'omics in human populations



Eric Franzosa

With Jacques Izard, Andy Chan, Wendy Garrett





2) Investigate links between the mouth and gut microbiomes

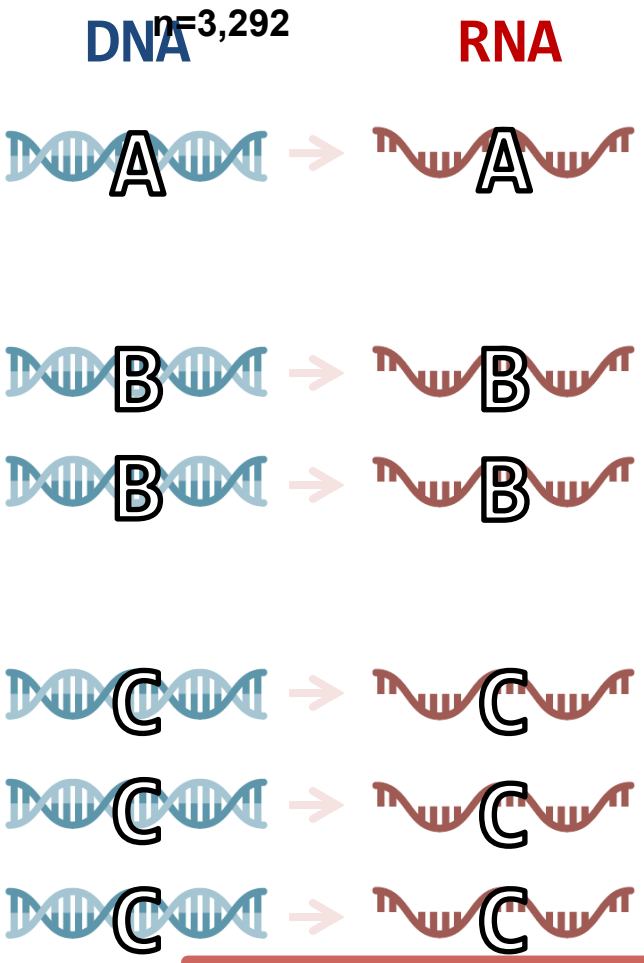
1) Evaluate stability of meta'omic samples under subject-shipped conditions

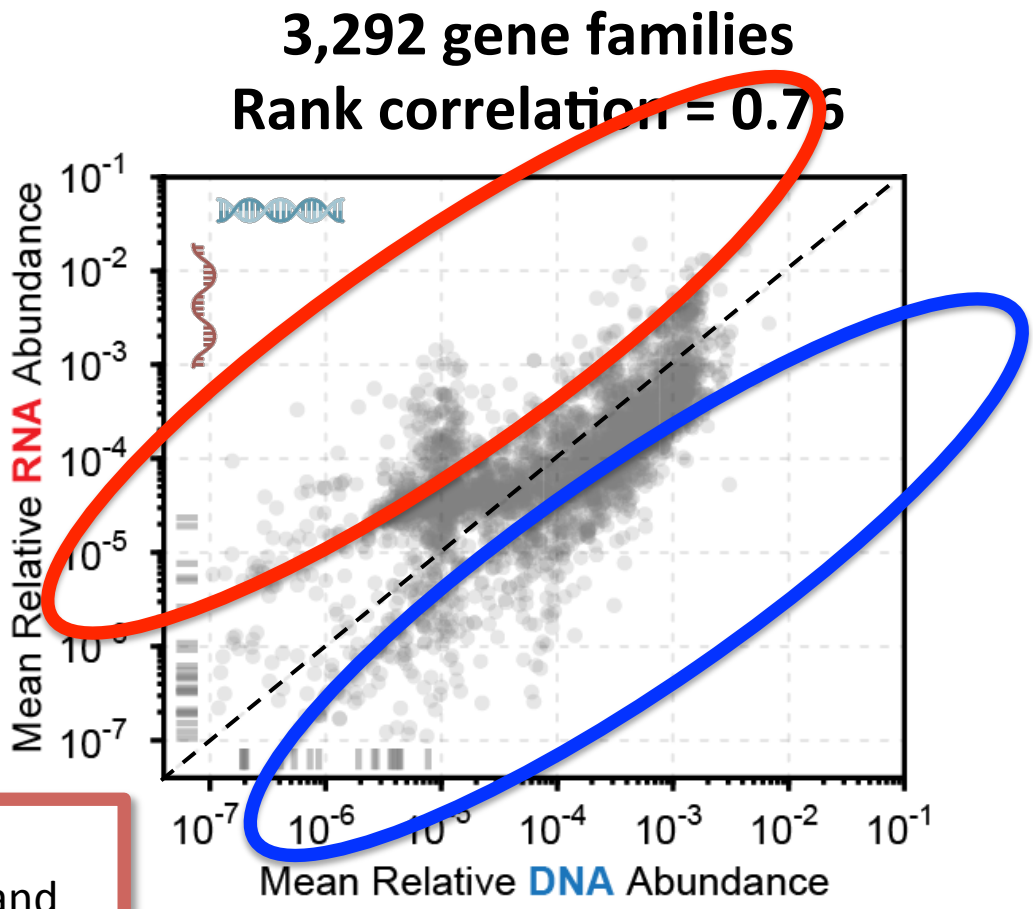
3) Relate the gut metagenome and metatranscriptome

Genes & transcripts are generally well correlated

A large portion of genes (~30%) are not differentially regulated

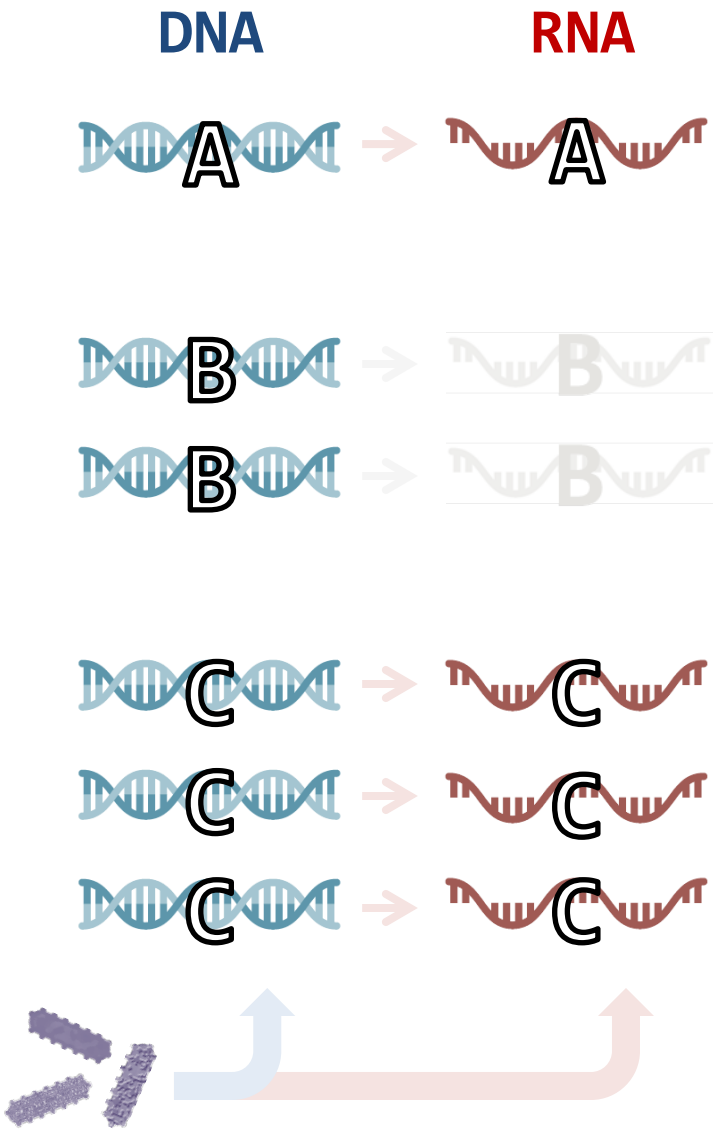
* Genes are KEGG Orthology groups, KOs



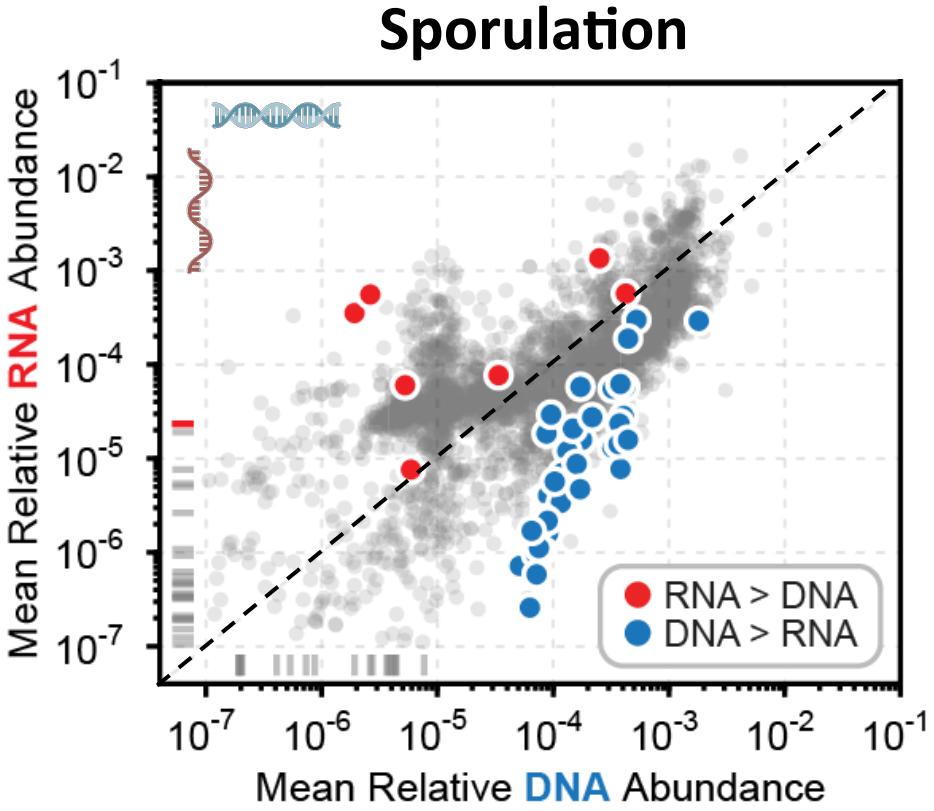


Remaining genes are
~40% **upregulated** (RNA) and
~60% **downregulated** (DNA)

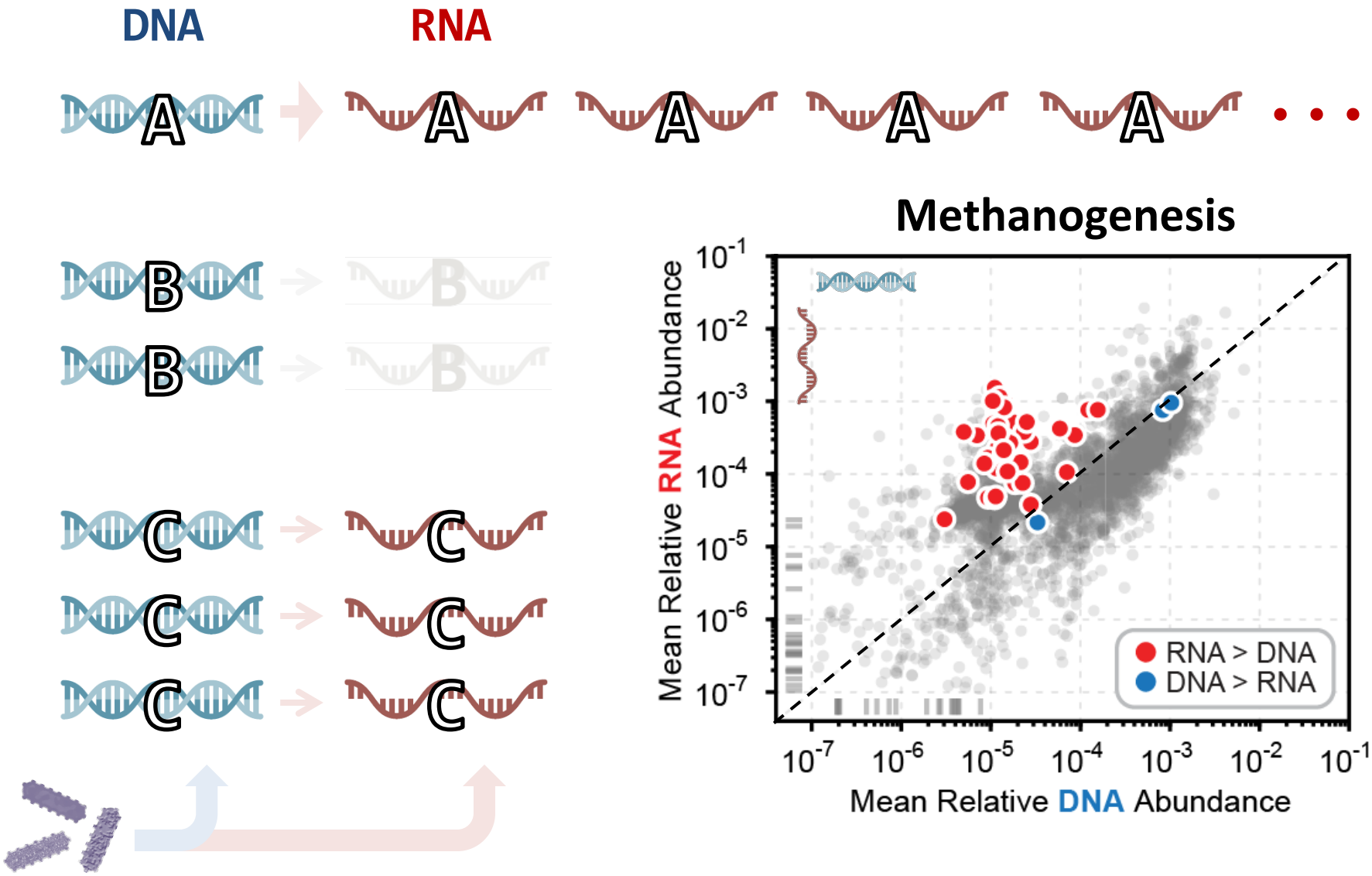
Some functions are highly under-expressed



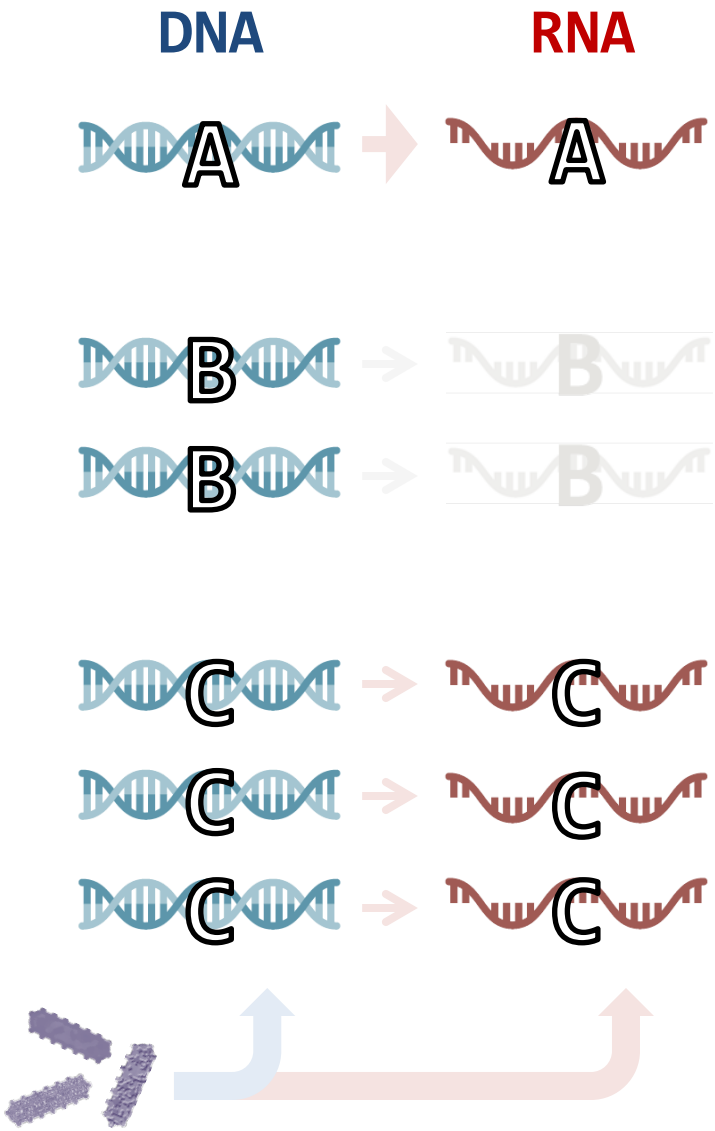
(RNA/DNA ratios used as input to a functional enrichment analysis)

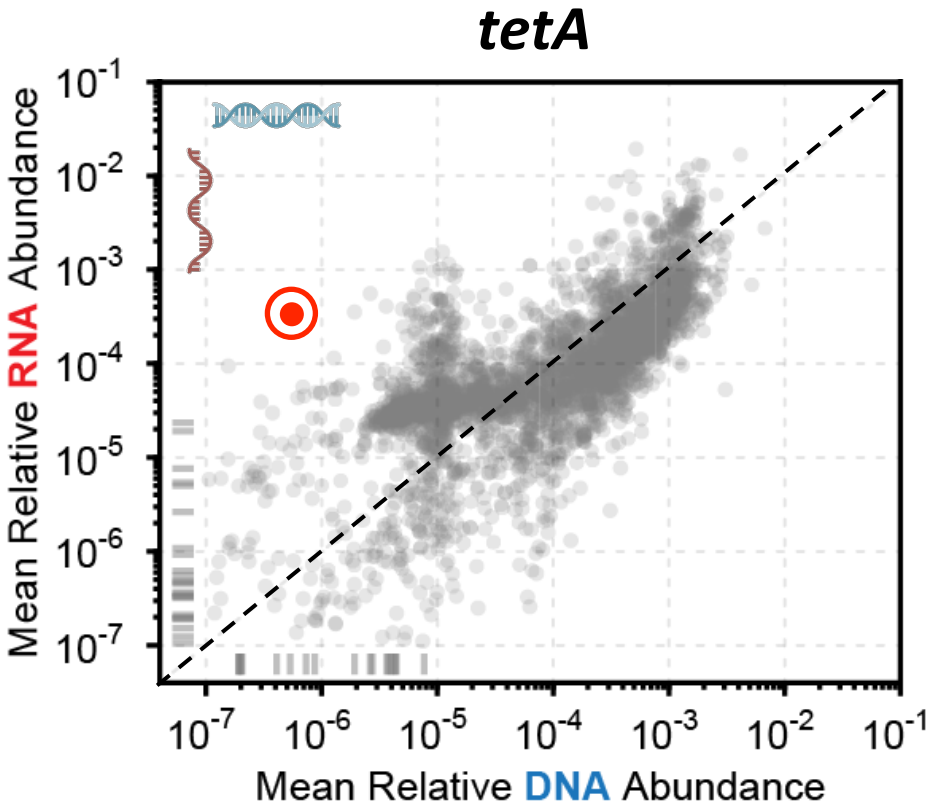


Other functions are highly over-expressed

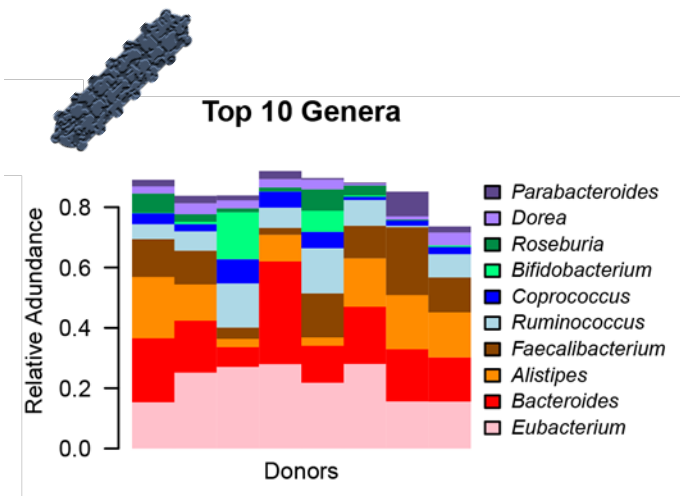


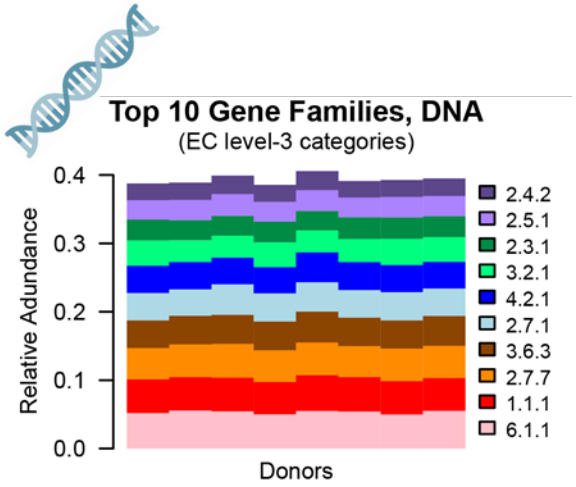
Other functions are highly over-expressed

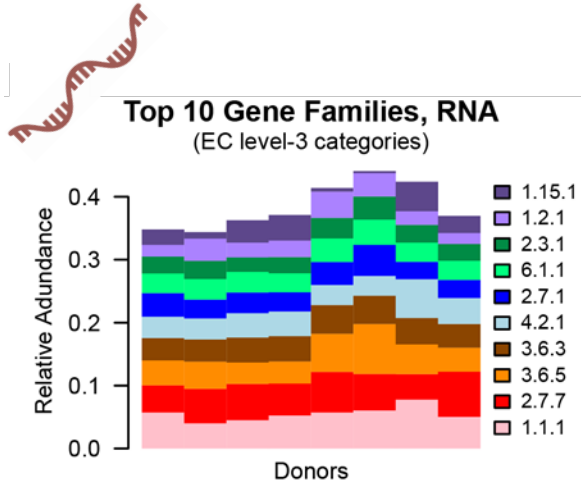


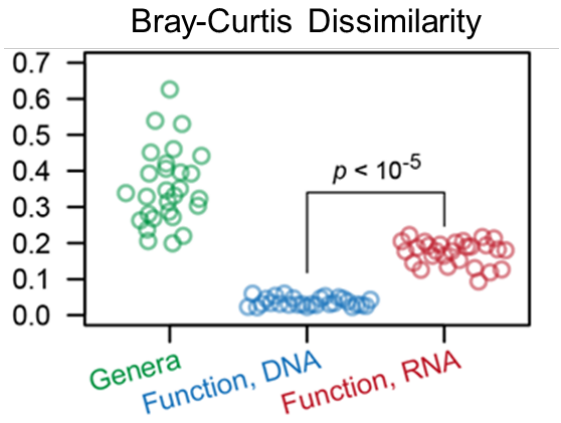


Functional potential is stable, activity is variable









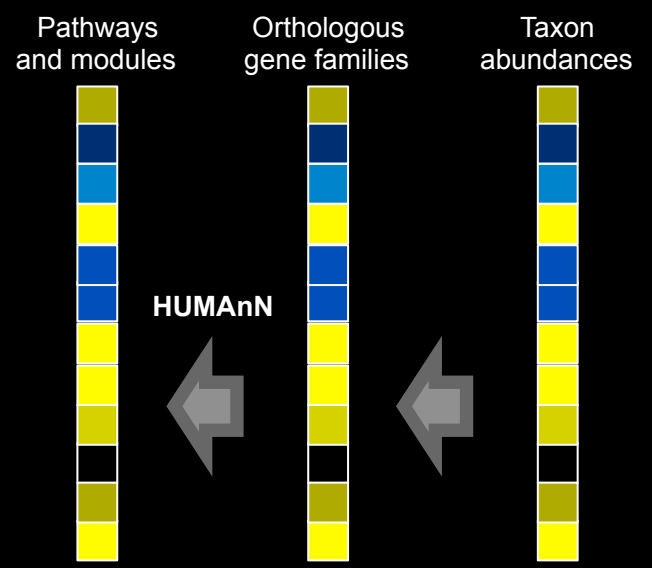
- Microbial membership varies.
  - Early colonization? Genetics?
- Over time, the community “solves” for a habitat-specific metagenome.
- It then differentially regulates that metagenome.
  - These two types of regulation differ *at least* in time scale.

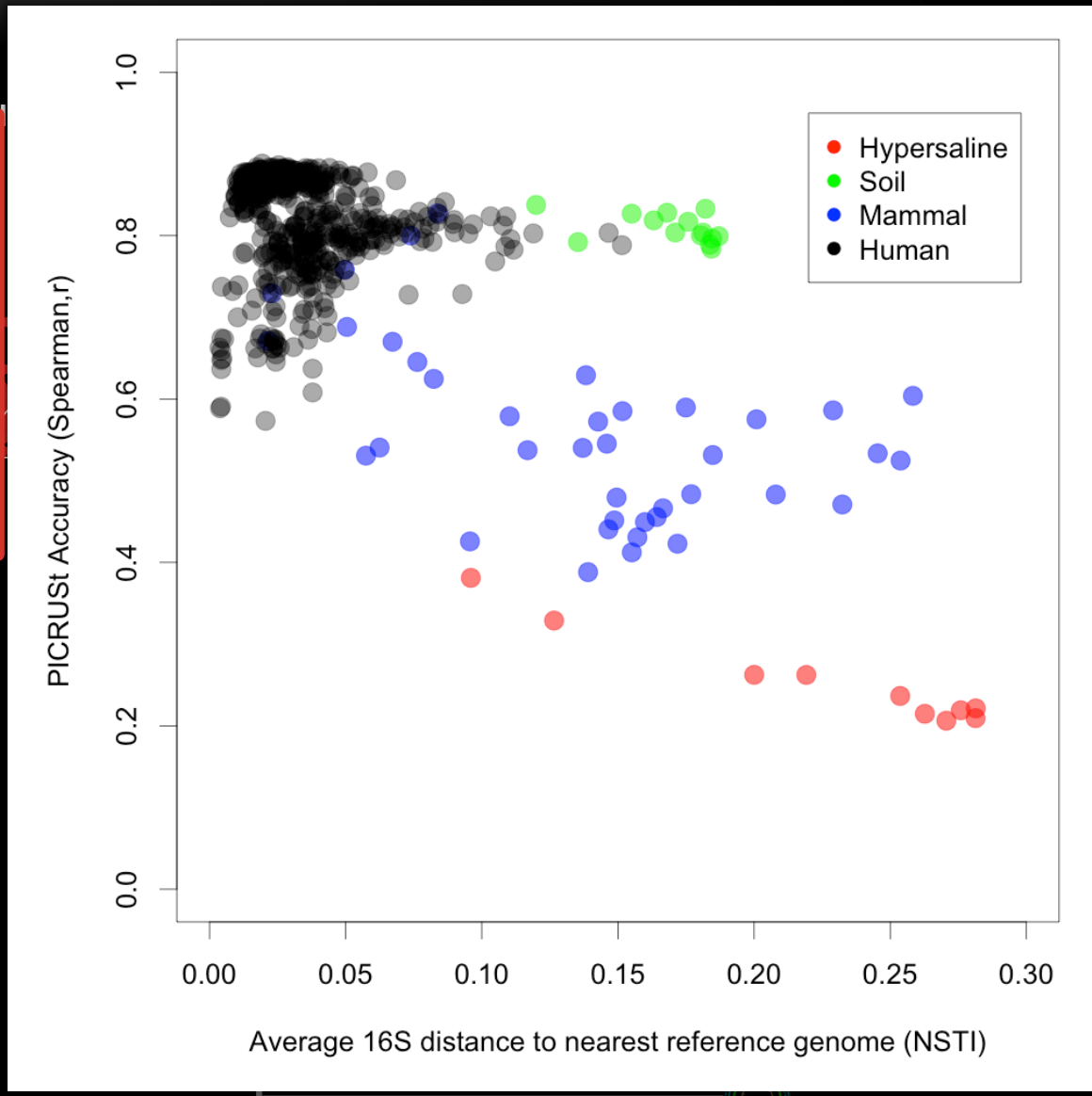
PICRUSt: Inferring community metagenomic potential from marker gene sequencing

With Rob Knight, Rob Beiko

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

http://picrust.github.com





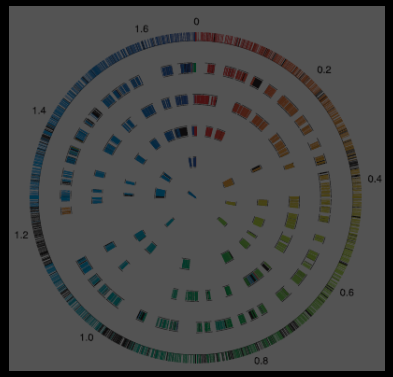
Microbiome meta'omic analyses: assembly

Acid mine drainage
(Tyson 2004)
76Mbp, JAZZ
>2kb: 1183 contigs, 11Mbp

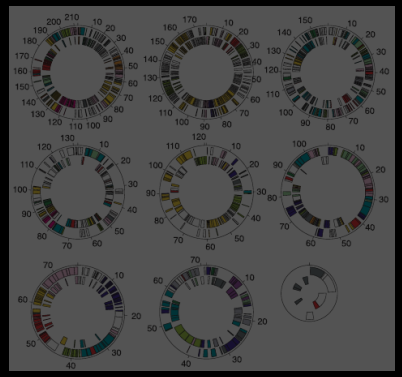
GOS (Venter 2004)
265Mbp, Celera
3x coverage: 333 scaffolds, 2226 contigs, 31Mbp
14x coverage: 21 scaffolds, 9.5Mbp

Termite hindgut
(Warnecke 2007)
71Mbp, Phrap
42k contigs, 44Mbp, 25% singlets

Lake Washington
(Kalyuzhnaya 2008)
255Mbp, PGA
~1.5x coverage: 25k contigs, 70Mbp

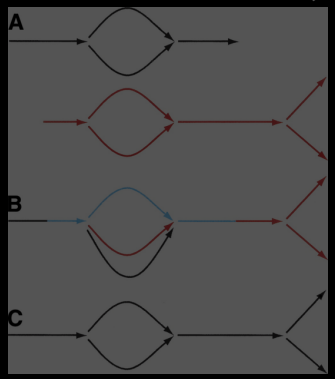


Prochlorococcus marinarus

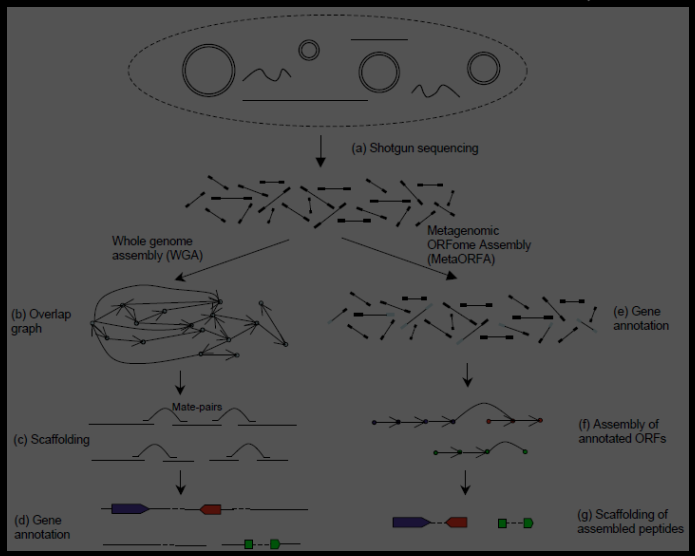


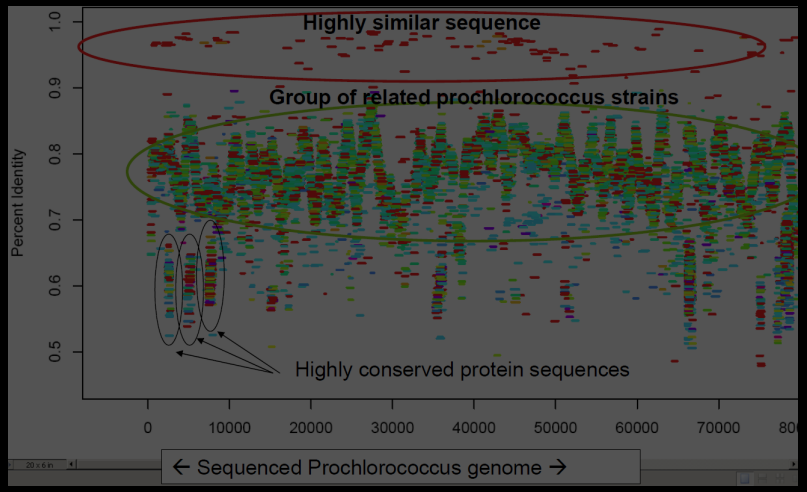
9 shared "megaplasmids"

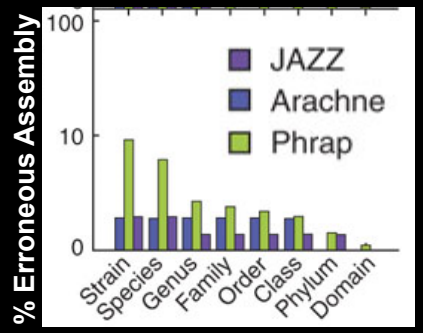
ALLPATHS: Butler, 2008



MetaORFA: Ye, 2008







Mavromatis, 2007

Microbiome meta'omic analyses: assembly

Scaling metagenome sequence assembly with probabilistic de Bruijn graphs

Jason Pell[a], Arend Hintze[a], Rosangela Canino-Koning[a], Adina Howe[b], James M. Tiedje[b,c], and C. Titus Brown[a,b,1]

khmer (Pell 2012)

P25

MetAMOS: a metagenomic assembly and analysis pipeline for AMOS

Todd J Treangen[1,2], Sergey Koren[1,3], Irina Astrovskaya[1], Dan Sommer[1], Bo Liu[1,3] and Mihai Pop[1,3]

[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA; [2]The McKusick-Nathans Institute for Genetic Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [3]Department of Computer Science, University of Maryland, College Park, MD 20742, USA

Genome Biology 2011, **12(Suppl 1)**:P25

MetaAMOS (Treangen 2012?)

MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads

Toshiaki Namiki[1,2], Tsuyoshi Hachiya[1], Hideaki Tanaka[1] and Yasubumi Sakakibara[1,*]

MetaVelvet (Namiki 2012)

BIOINFORMATICS Vol. 27 ISMB 2011, pages i94–i101 doi:10.1093/bioinformatics/btr216

Meta-IDBA: a *de Novo* assembler for metagenomic data

Yu Peng, Henry C. M. Leung, S. M. Yiu and Francis Y. L. Chin*

Department of Computer Science, The University of Hong Kong, Hong Kong

Meta-IDBA (Peng 2011)

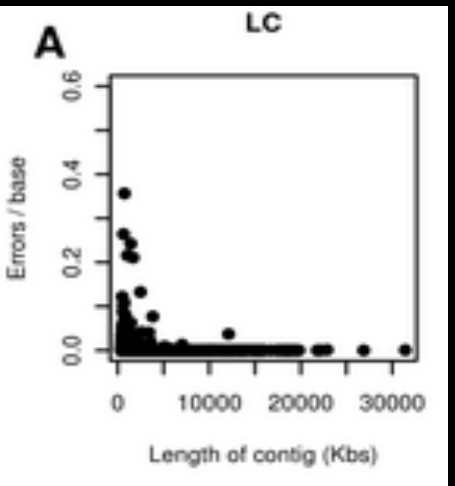
Genovo: *De Novo* Assembly for Metagenomes

*JONATHAN LASERSON, *VLADIMIR JOJIC, and DAPHNE KOLLER

Genovo (Laserson 2011)

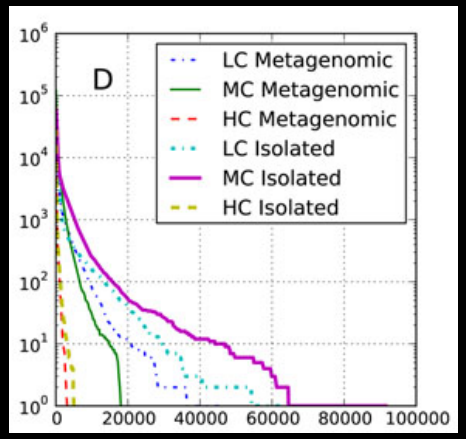
Evaluating the Fidelity of *De Novo* Short Read Metagenomic Assembly Using Simulated Data

Miguel Pignatelli[1,2,3*], Andrés Moya[1,2]



Evaluation of short read metagenomic assembly

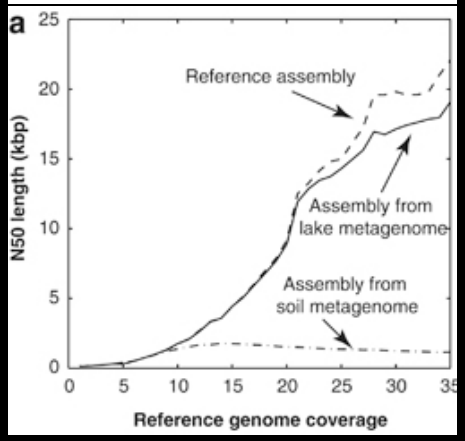
Anveshi Charuvaka, Huzefa Rangwala*



Individual genome assembly from complex community short-read metagenomic datasets

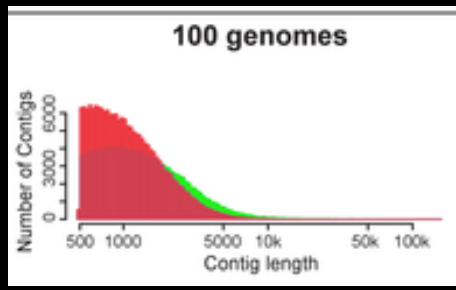
Chengwei Luo[1], Despina Tsementzi[2], Nikos C Kyrpides[3] and Konstantinos T Konstantinidis[1,2]

[1]*Center for Bioinformatics and Computational Genomics and School of Biology, Georgia Institute of Technology, Atlanta, GA, USA;* [2]*School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, USA and* [3]*Department of Energy (DOE) – Joint Genome Institute, Walnut Creek, CA, USA*



Assessment of Metagenomic Assembly Using Simulated Next Generation Sequencing Data

Daniel R. Mende[1*], Alison S. Waller[1*], Shinichi Sunagawa[1], Aino I. Järvelin[1], Michelle M. Chan[2], Manimozhiyan Arumugam[1], Jeroen Raes[3], Peer Bork[1,4*]

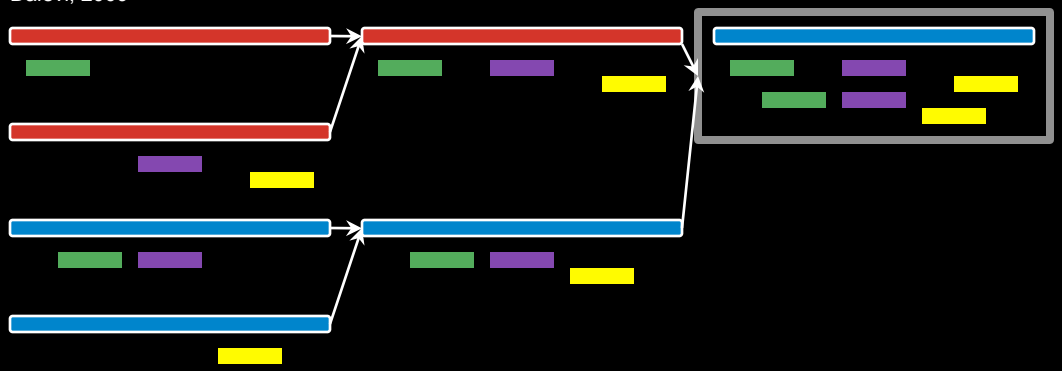


Microbiome meta'omic analyses: gene calling and proxygenes

Extrinsic gene calling:
BLAST etc. (proxygenes)

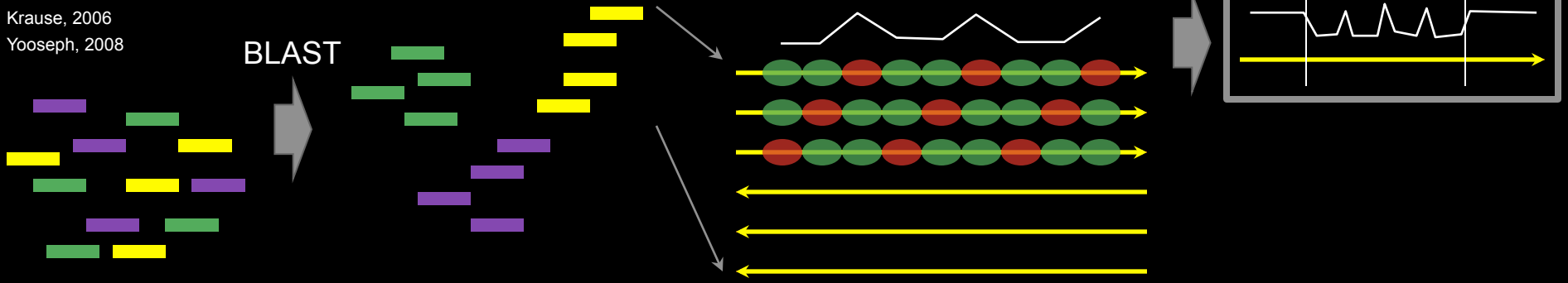
Intrinsic gene calling:
ORF detection from seq.

Dalevi, 2009



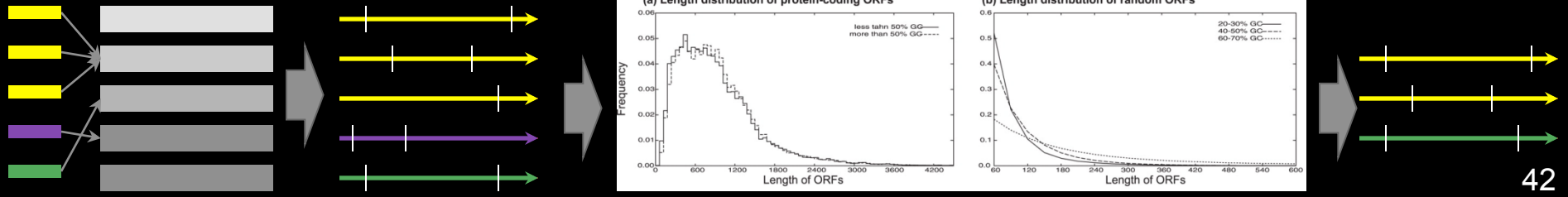
Krause, 2006
Yooseph, 2008

BLAST



Orphelia: Hoff, 2009
MetaGene: Noguchi, 2006

HMM models



Downstream analyses

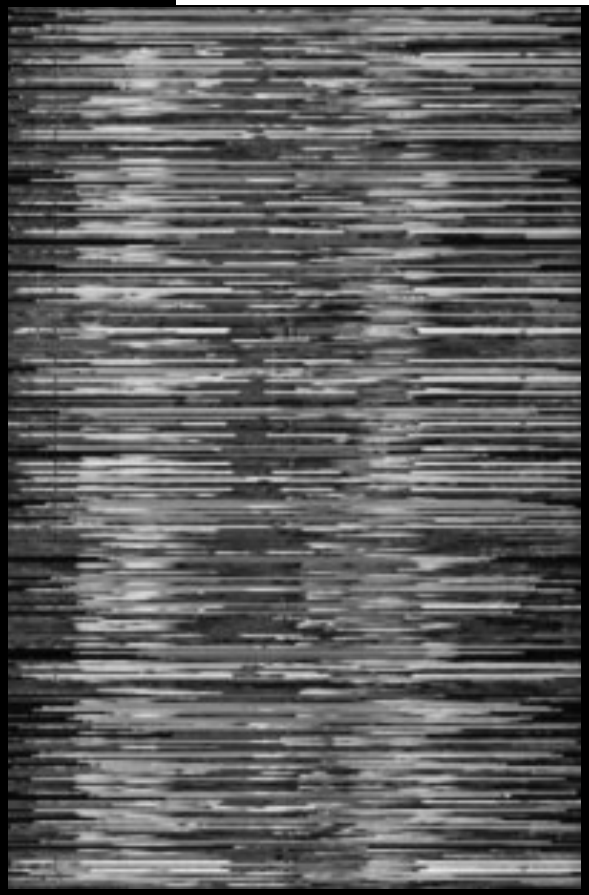
Mardis *Genome Medicine* 2010, 2:84
http://genomemedicine.com/content/2/11/84

Genome Medicine

MUSINGS

The $1,000 genome, the $100,000 analysis?

Elaine R Mardis*



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 ($\Sigma = 1$)
  - 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

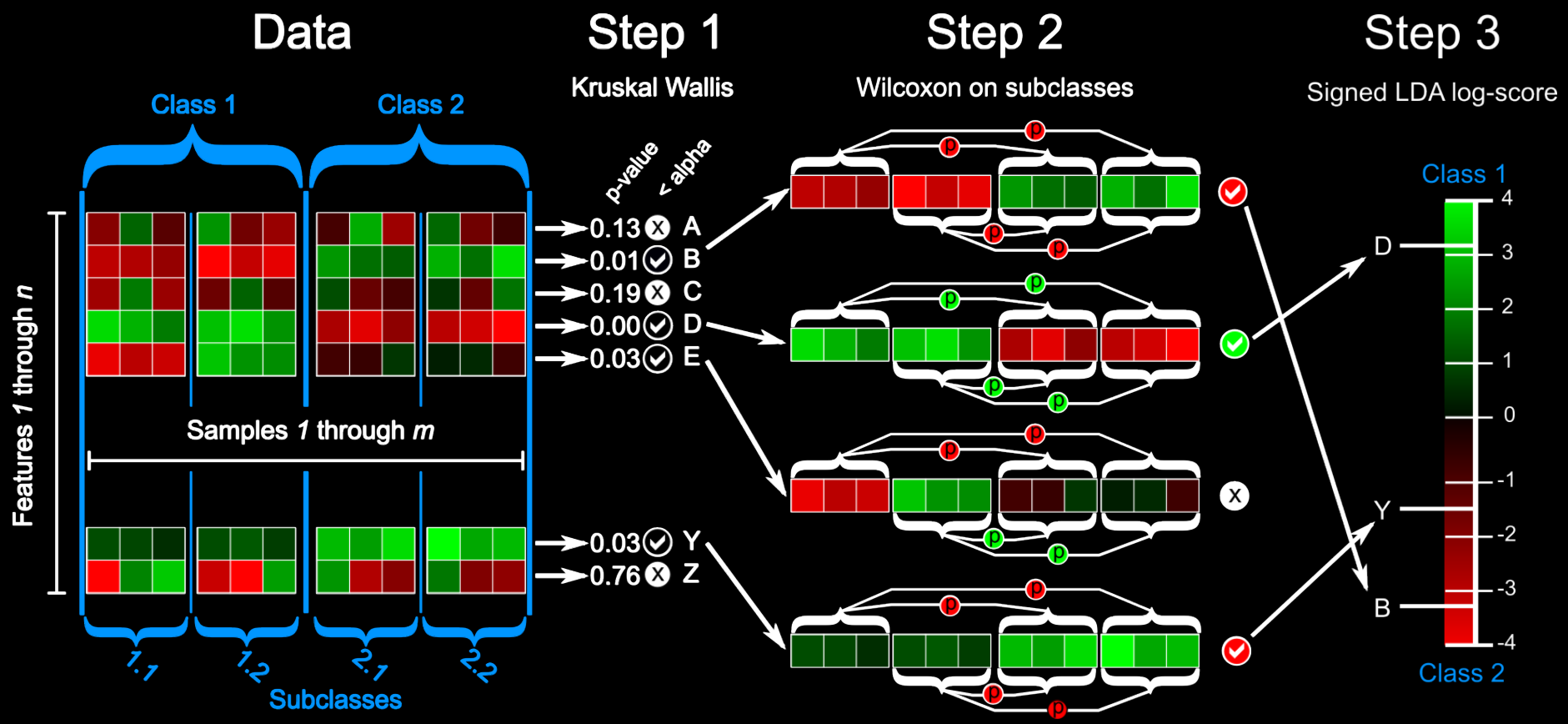
Properties of microbiome data

- 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

LEfSe: LDA Effect Size
Finding metagenomic biomarkers

Nicola Segata



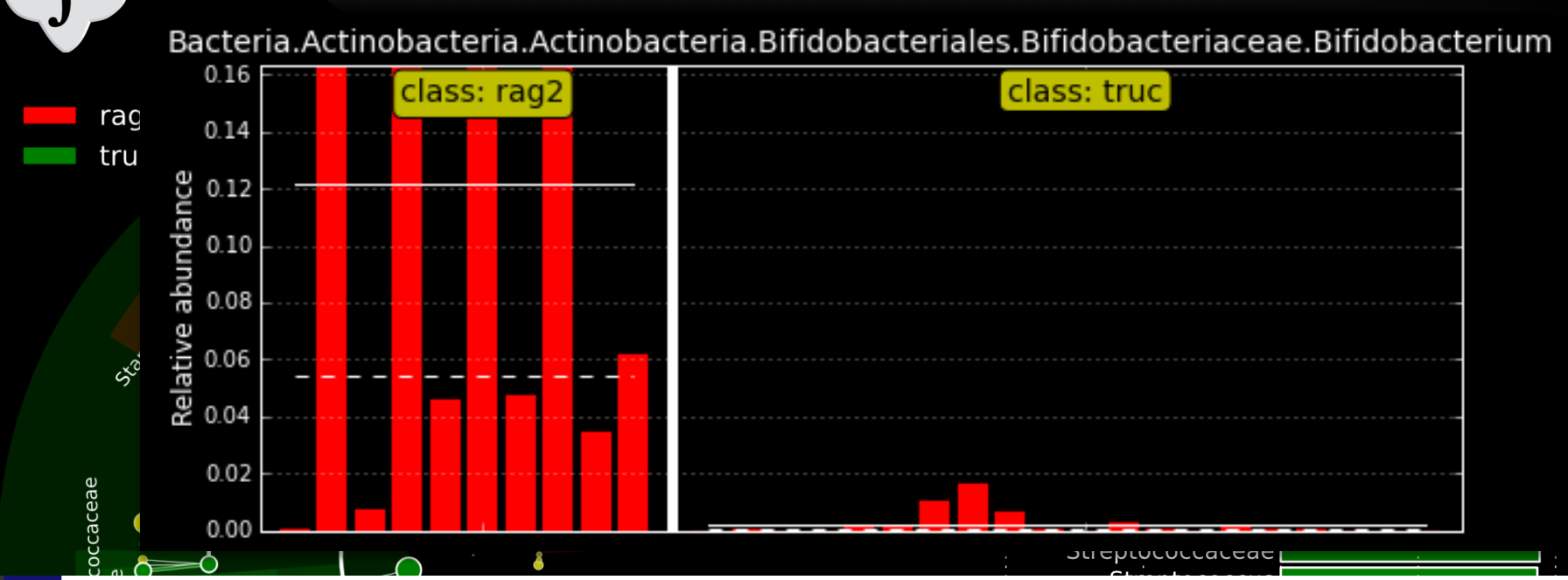
A — Statistical consistency

B — Biological consistency

C — Biological Effect size

Microbiome downstream analyses: statistical association testing (with LEfSe)

With Wendy Garrett



PNAS

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

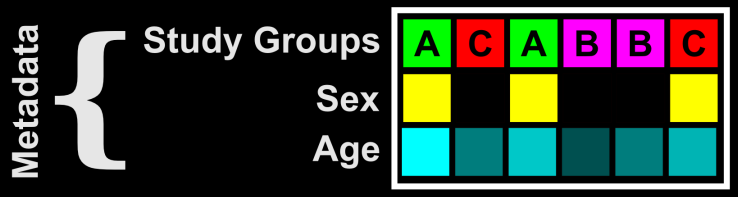
Patrick Veiga[a,b], Carey Ann Gallini[a], Chloé Beal[b], Monia Michaud[a], Mary L. Delaney[c], Andrea DuBois[c], Artem Khlebnikov[b,d], Johan E.T. van Hylckama Vlieg[b], Shivesh Punit[a,1], Jonathan N. Glickman[c,e,2], Andrew Onderdonk[c,e], Laurie H. Glimcher[a,d,e,f], and Wendy S. Garrett[a,e,g,3]

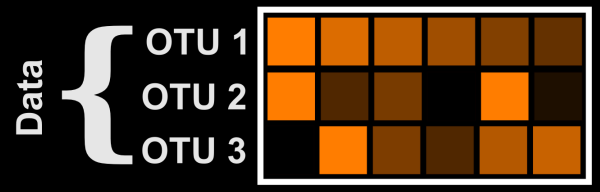
[a]Harvard School of Public Health, Boston, MA 02115; [b]Danone Research, 91767 Palaiseau, France [c]Brigham and Women's Hospital, Boston, MA 02115; [d]Dannon Company Inc., White Plains, NY 10603; [e]Harvard Medical School, Boston, MA 02115; [f]Ragon Institute of MGH, MIT and Harvard, Charlestown, MA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

Edited by Peter K. Vogt, The Scripps Research Institute, La Jolla, CA, and approved September 8, 2010 (received for review August 5, 2010)

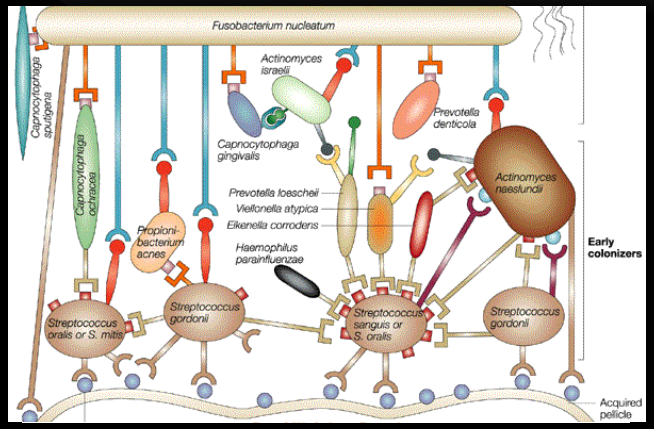
MaAsLin: Multivariate (microbial) Association with Linear Models

Overview of MaAsLin Association Methodology



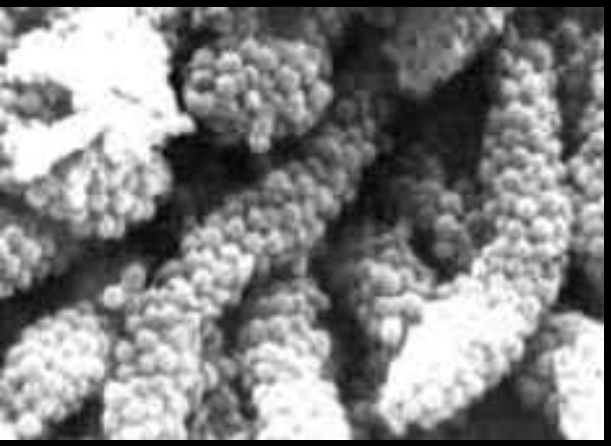


Microbiome downstream analyses: interaction network reconstruction



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

Mutualism

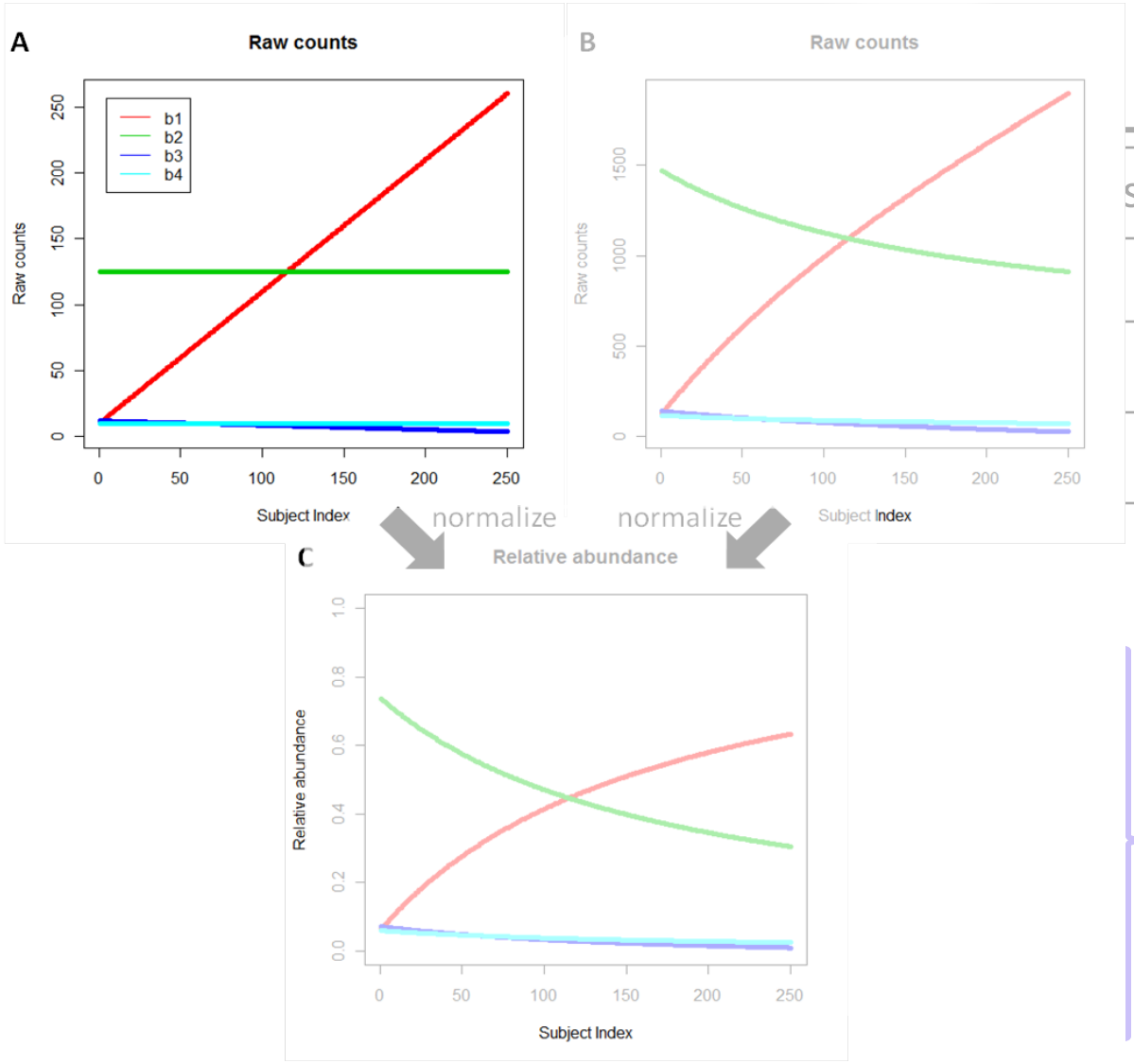


Predation

Competition

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

Sequencing assays provide only compositional measurements, in which information is lost



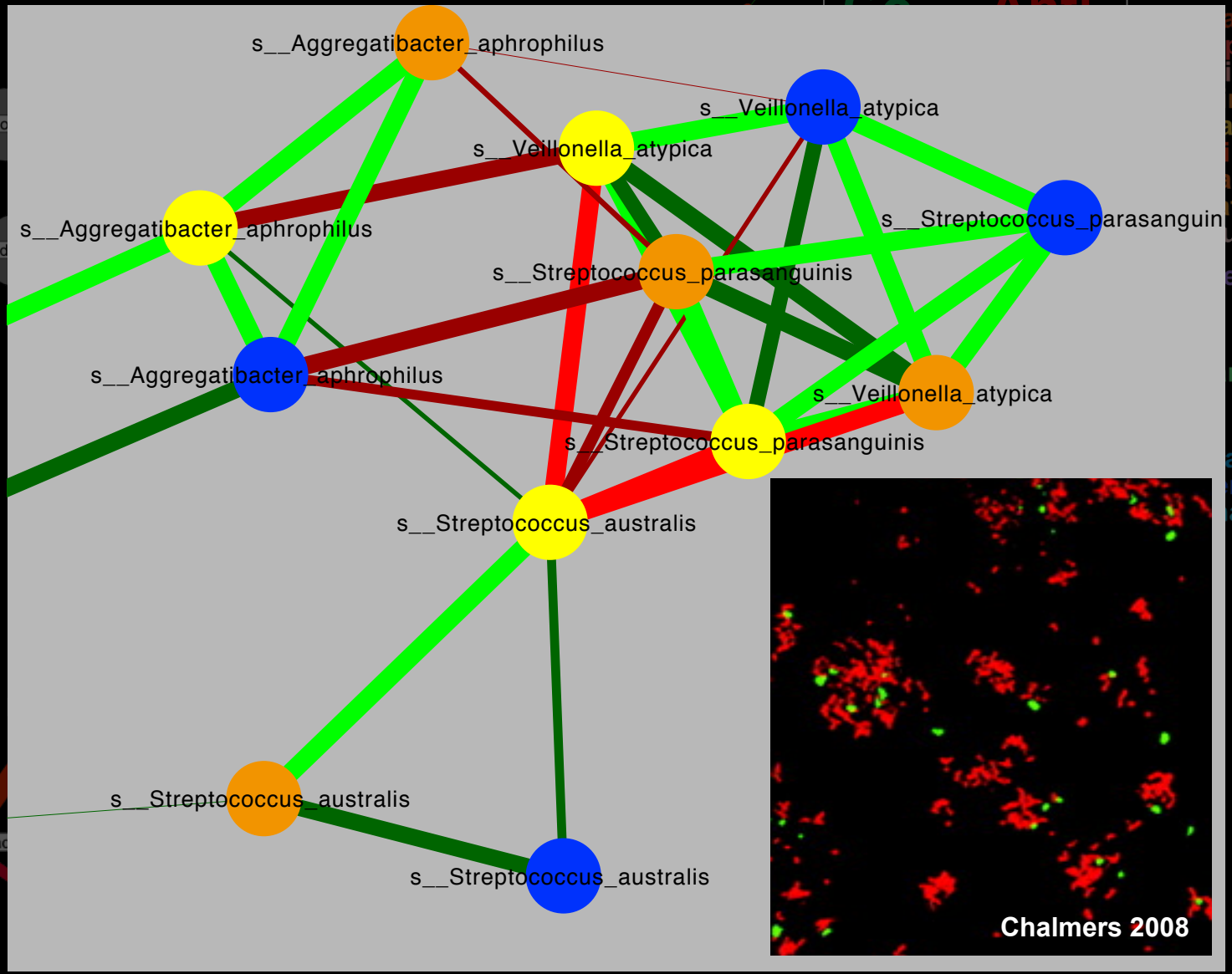
	A	C	Total
Subject 1	100	100	300
Subject 2	100	250	800

Relative Abundances

Friends, neighbors, and enemies:

Microbial co-occurrence and exclusion in the human microbiome

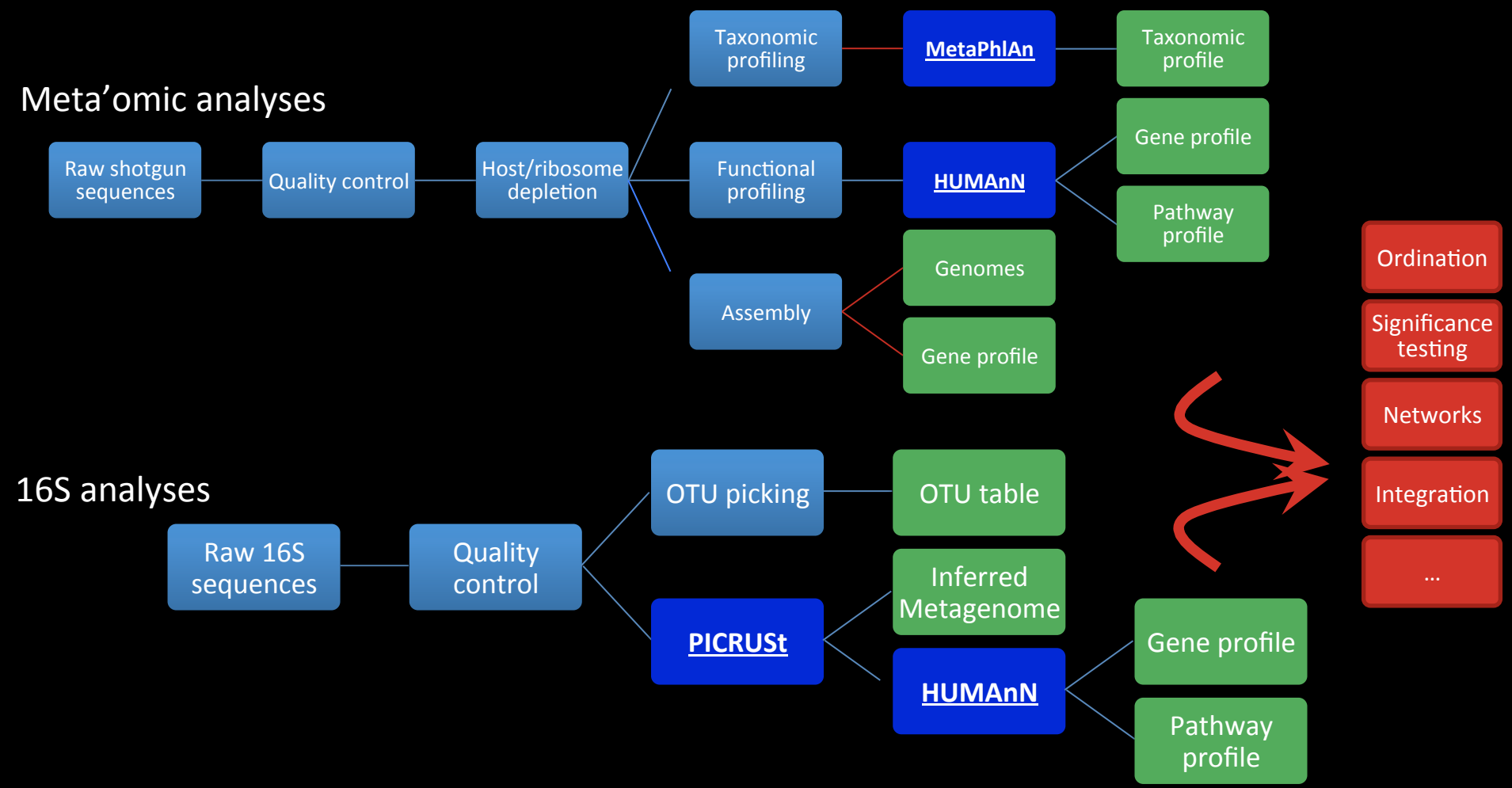
With Jeroen Raes, Karoline Faust



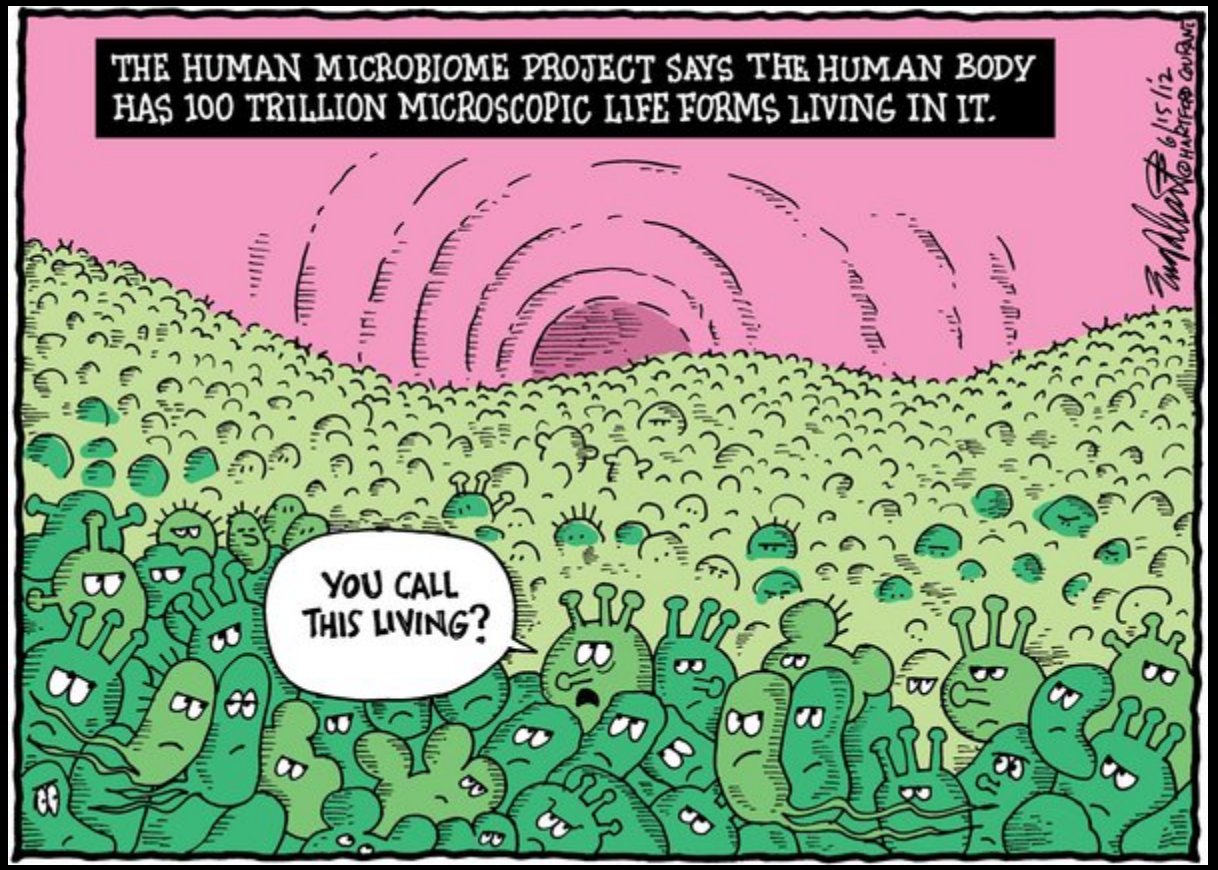
Emma Schwager

Fah Sathirapongsasuti

Typical microbial community analysis tasks

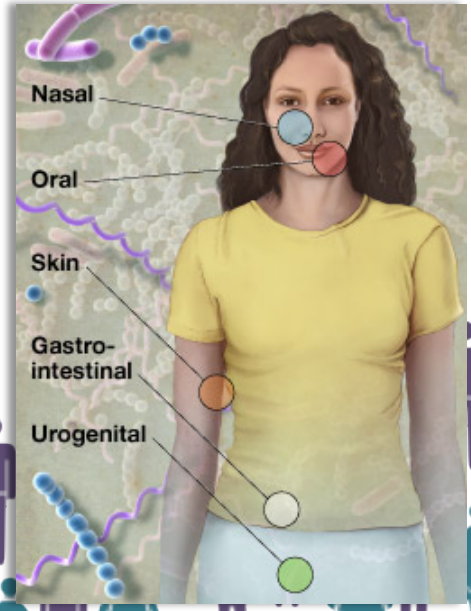


Bringing it all together: The Human Microbiome Project



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



5,200 16S samples
Spanning 300 subjects, 18 sites

700 shotgun samples
Subset of 100 subjects, six sites

http://www.nature.com/nature/focus/humanmicrobiota/

HUMAN MICROBIOME PROJECT HMP

www.hmpdacc.org

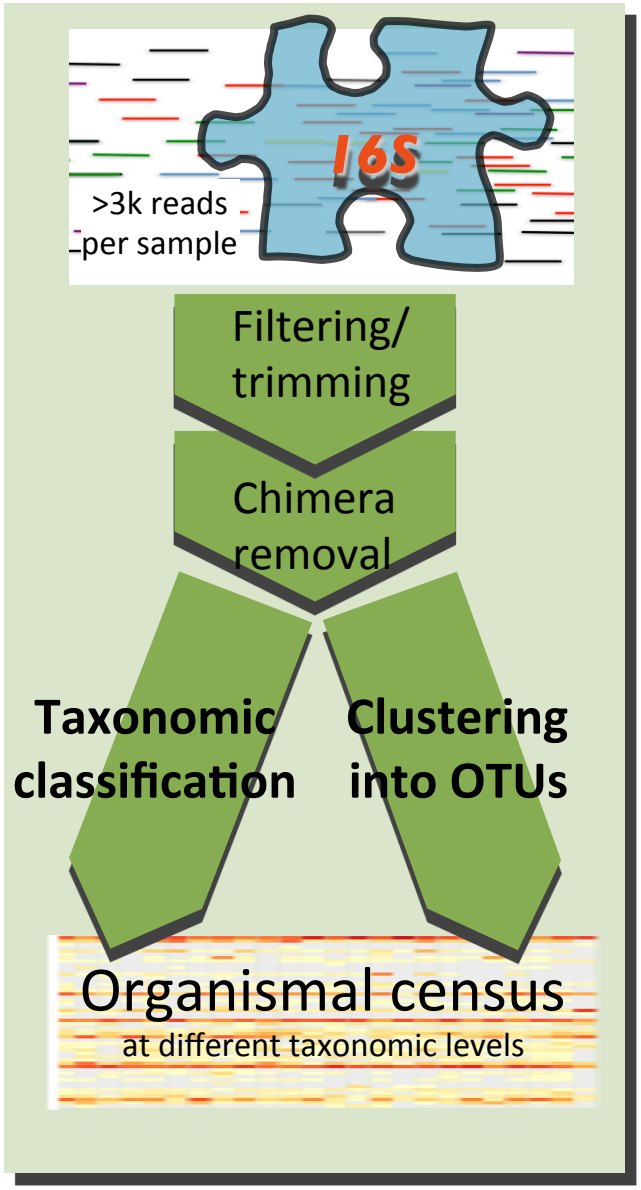
Slide by Dirk Gevers

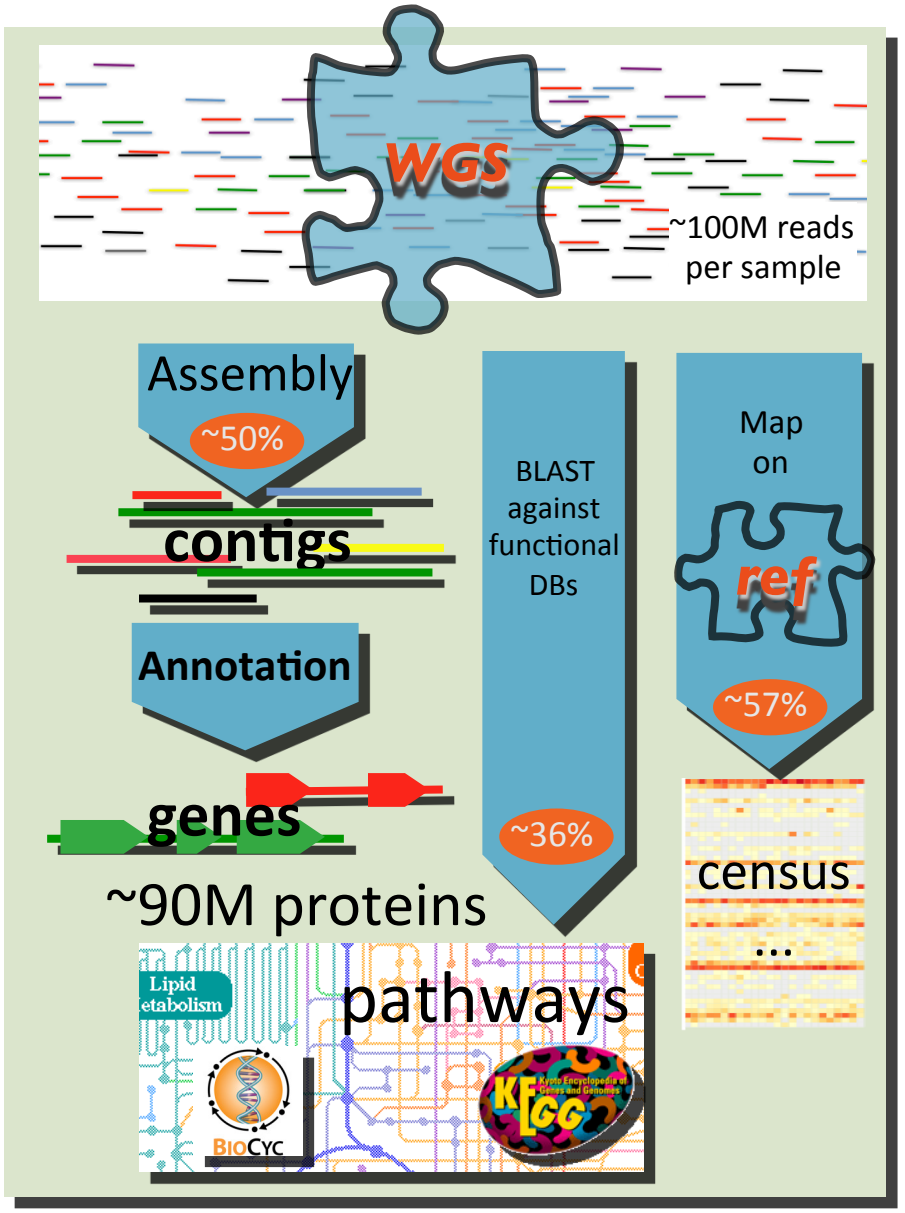
Figure 1. Timeline of microbial community studies using high-throughput sequencing.

Gevers D, Knight R, Petrosino JF, Huang K, et al. (2012) The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome. PLoS Biol 10(8): e1001377. doi:10.1371/journal.pbio.1001377
http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001377

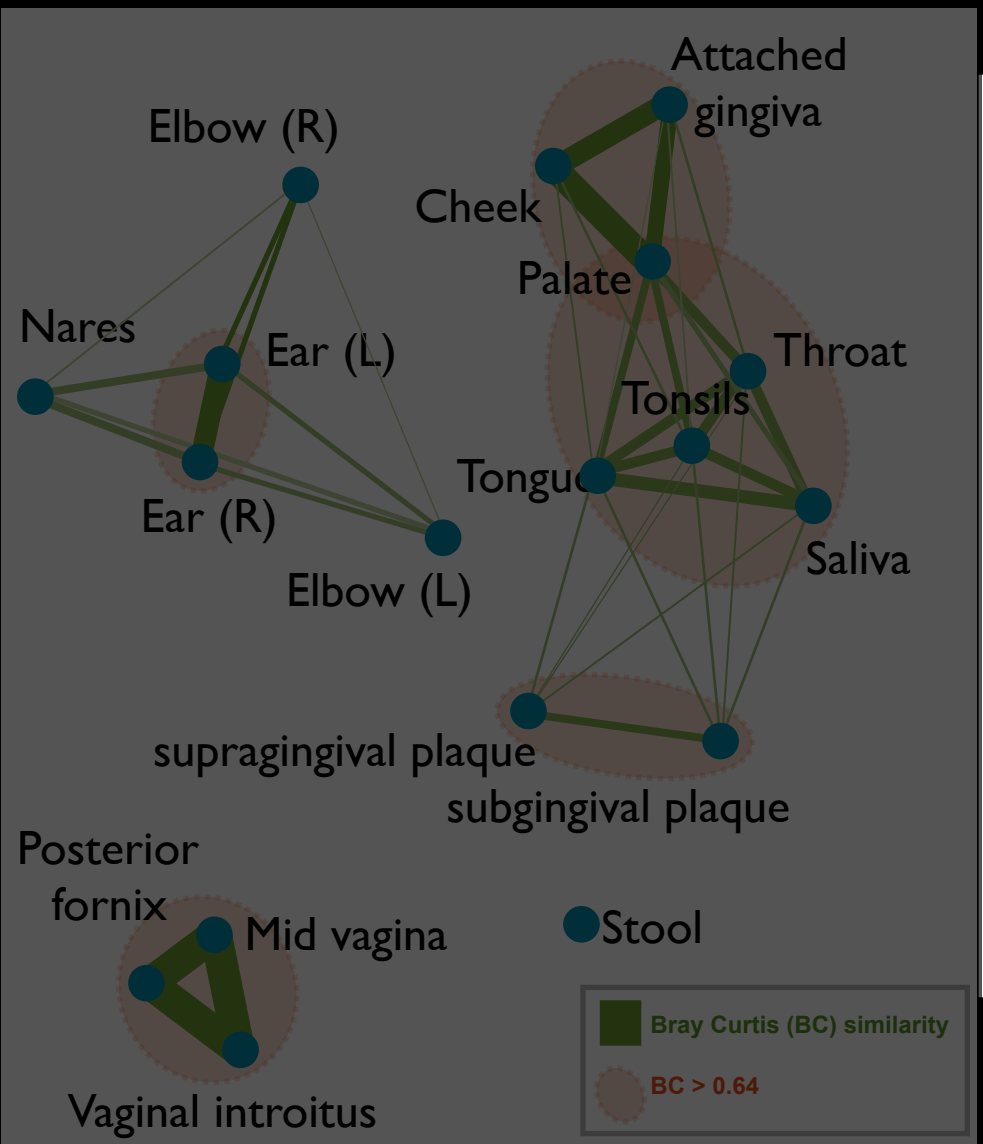
PLOS BIOLOGY TENTH ANNIVERSARY

…for mining metagenomic data

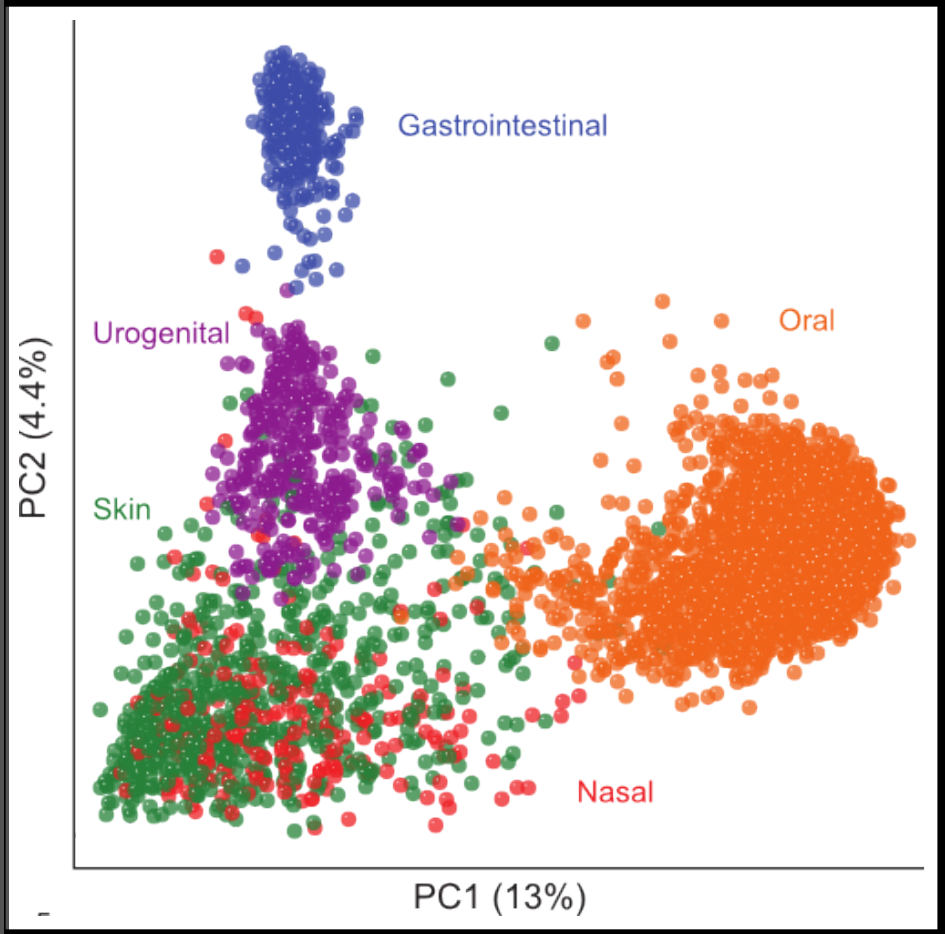




What aspects of a human host most influence microbial community composition?



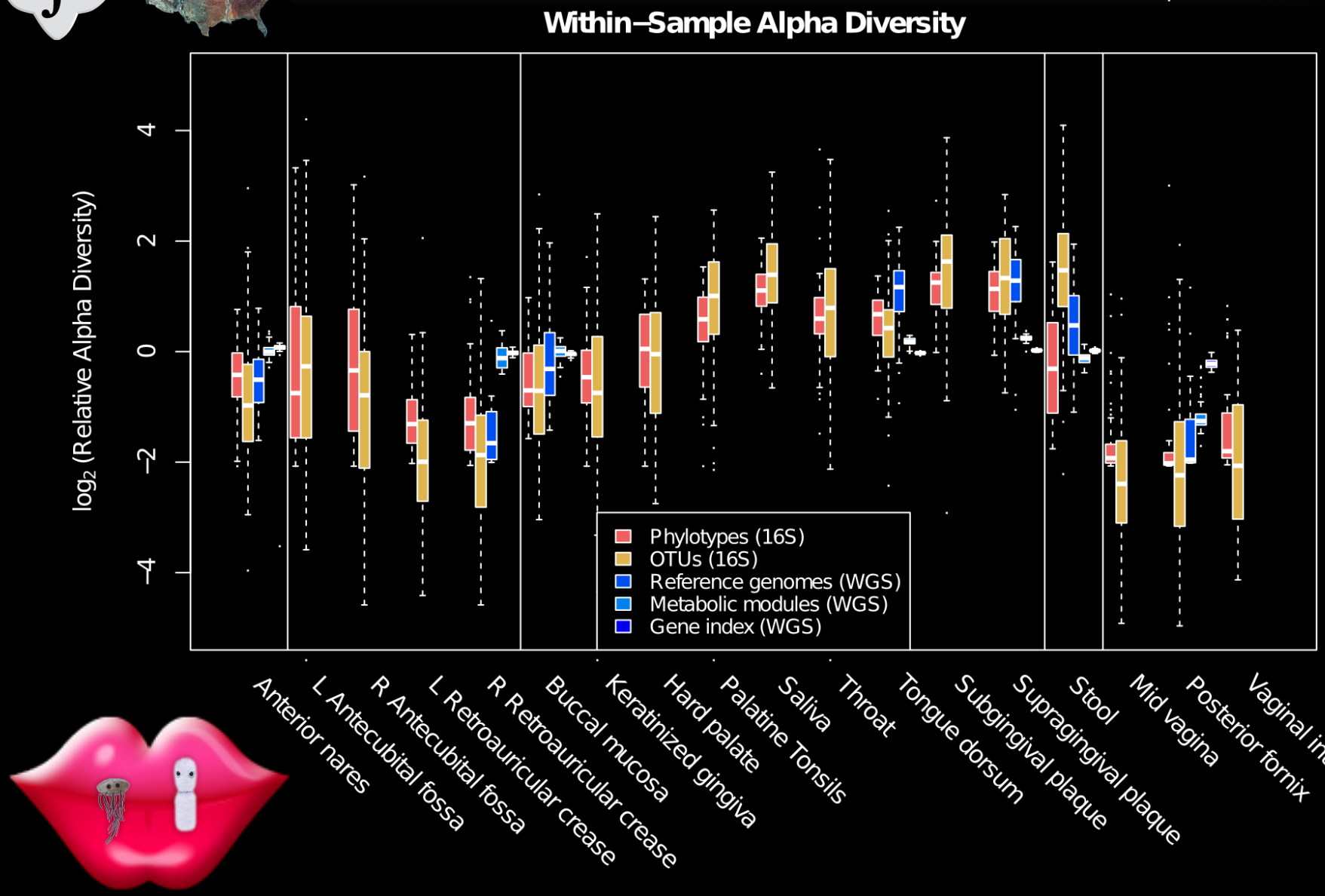
~5,200 microbial communities



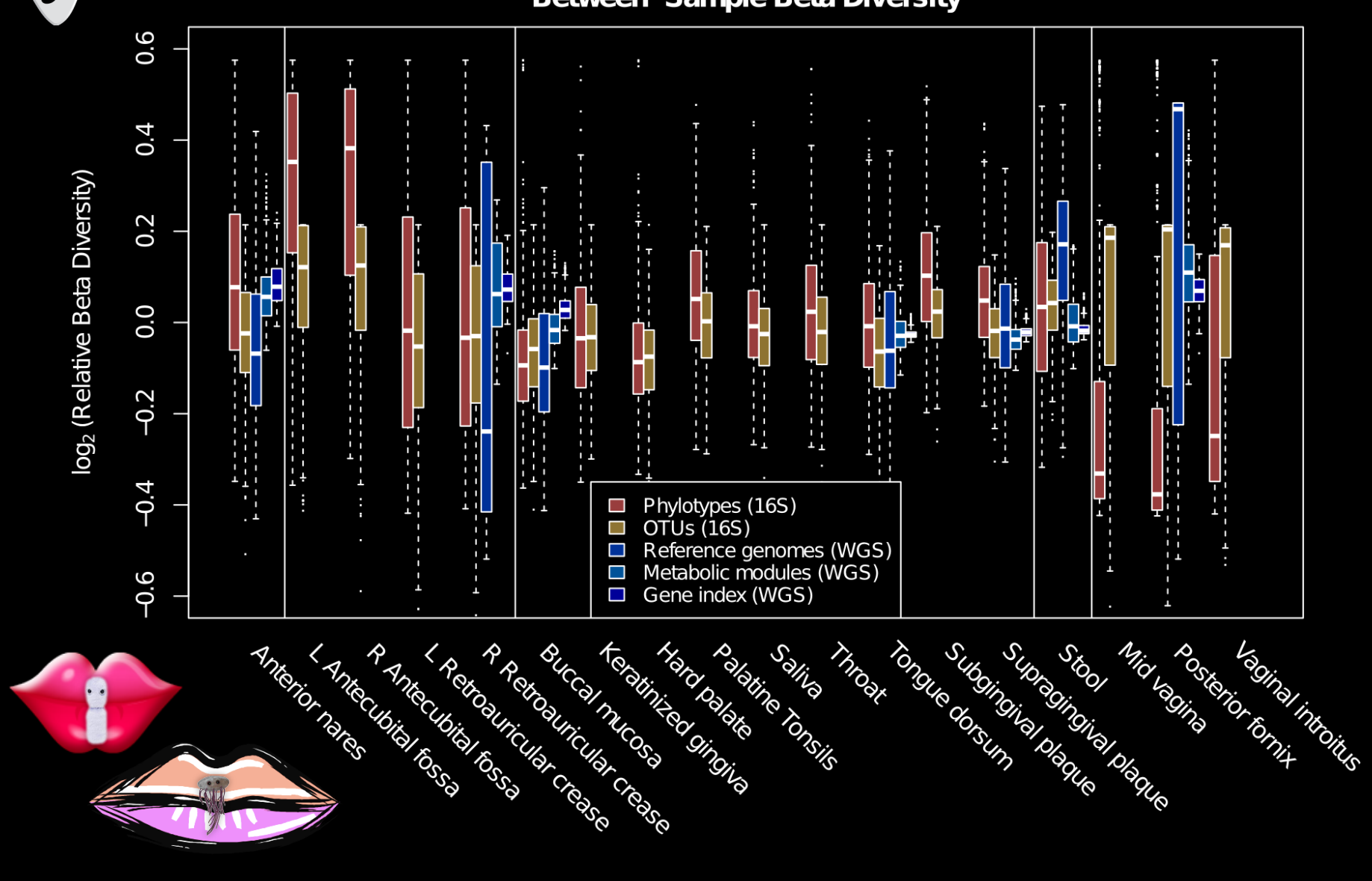
closer = more similar

Rob Knight

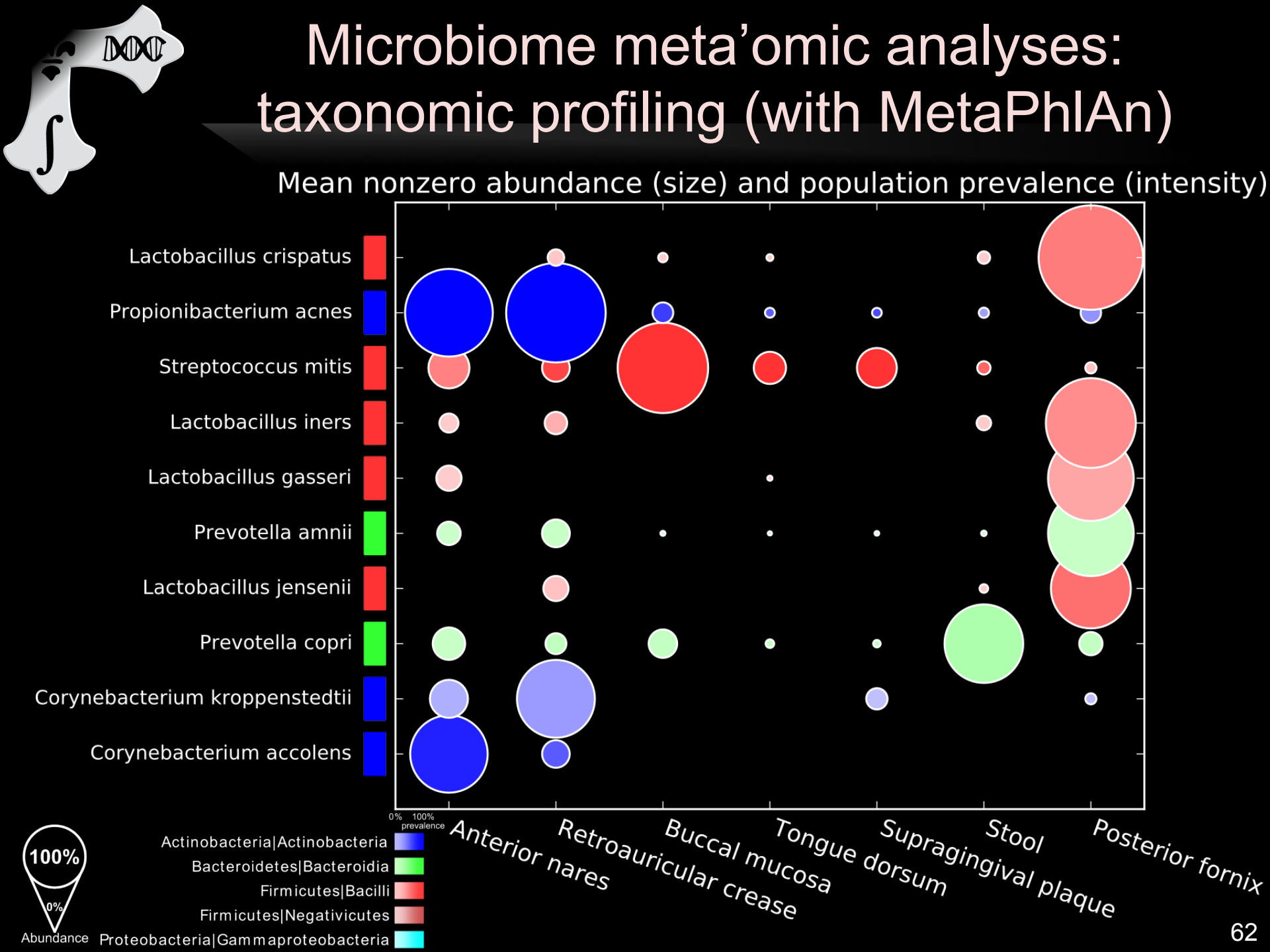
Which human body sites harbor the greatest microbial diversity per individual?



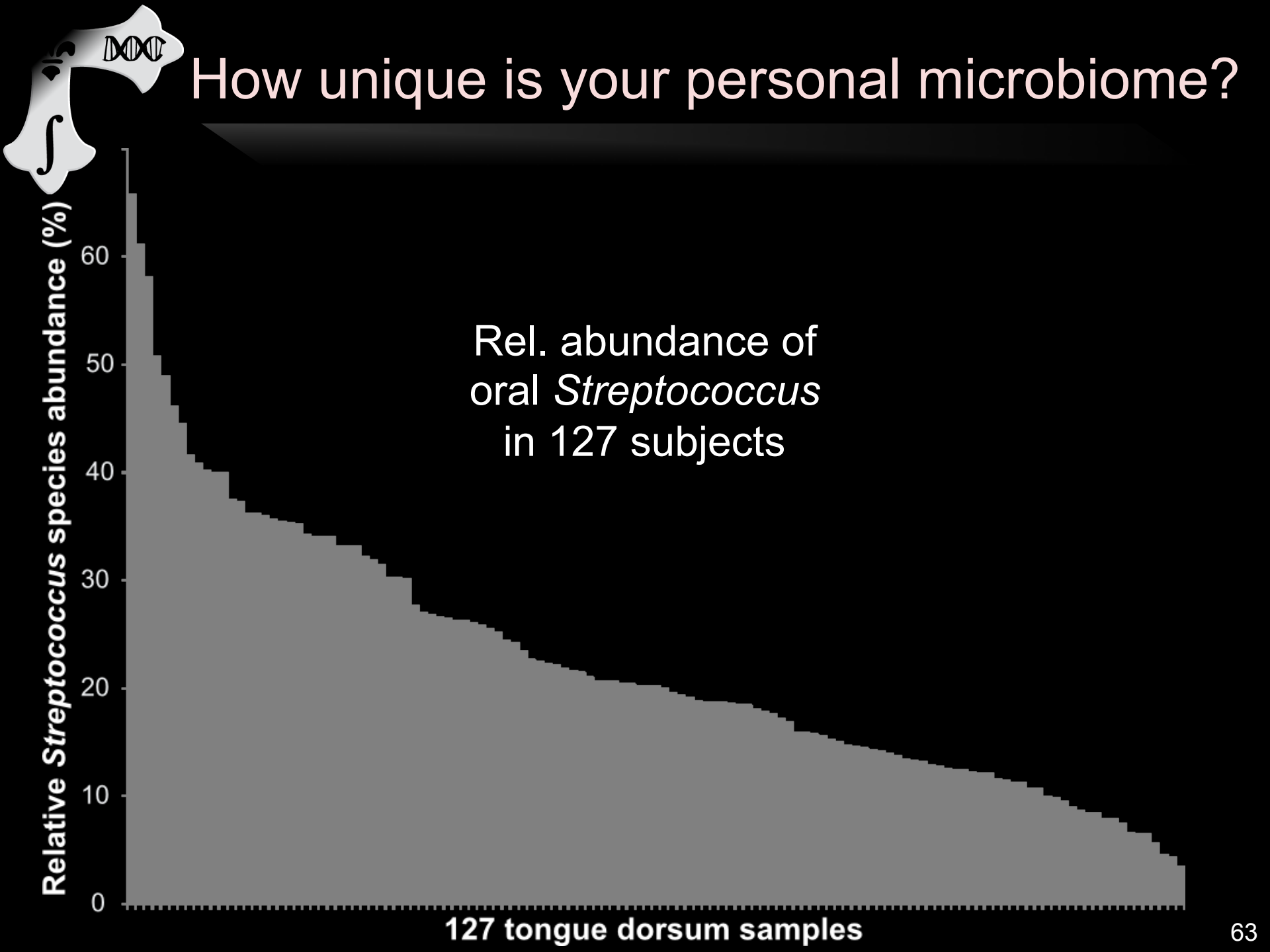
Which human body sites share the greatest microbial diversity among individuals?



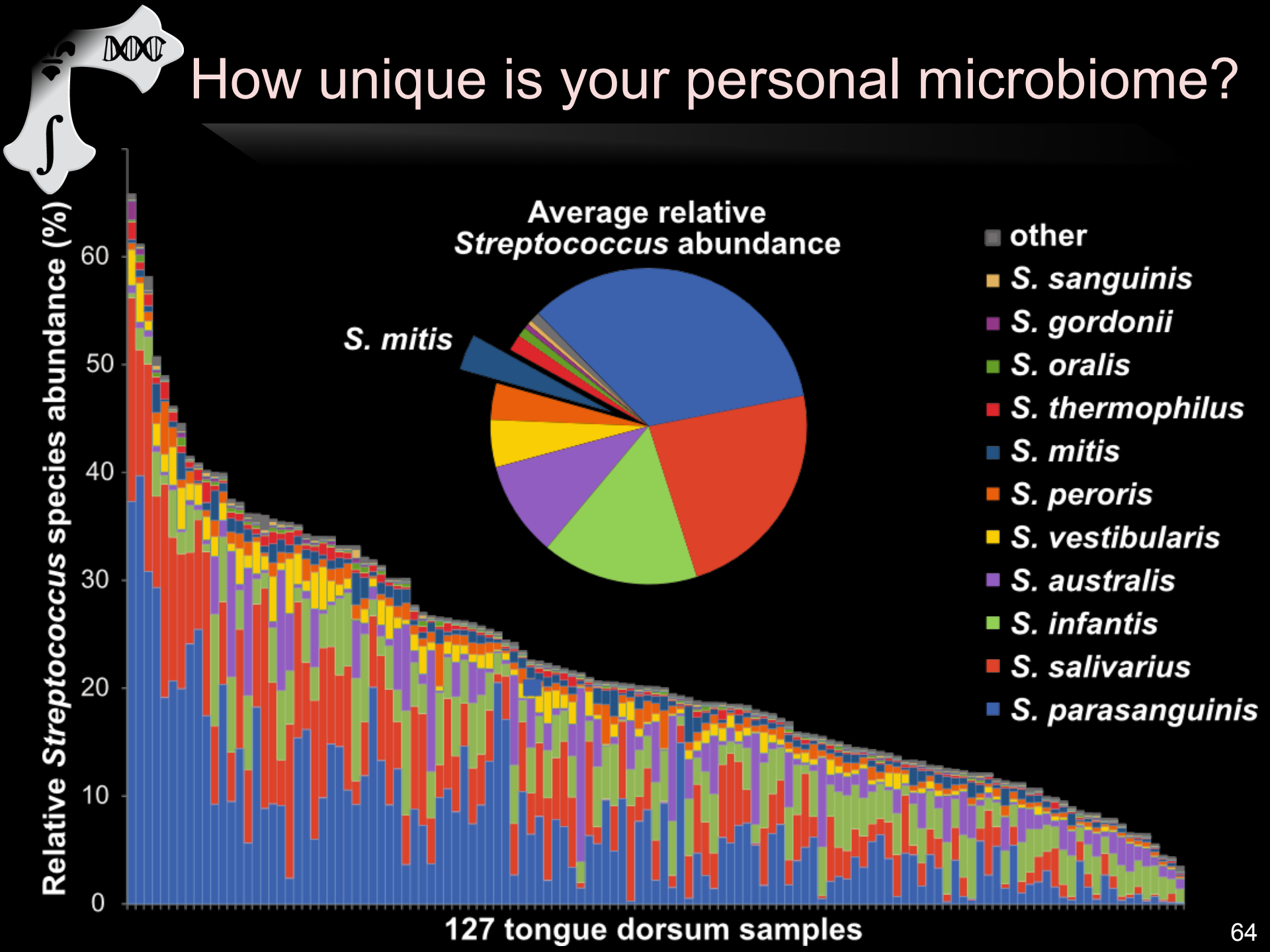
Microbiome meta'omic analyses: taxonomic profiling (with MetaPhlAn)



How unique is your personal microbiome?



How unique is your personal microbiome?

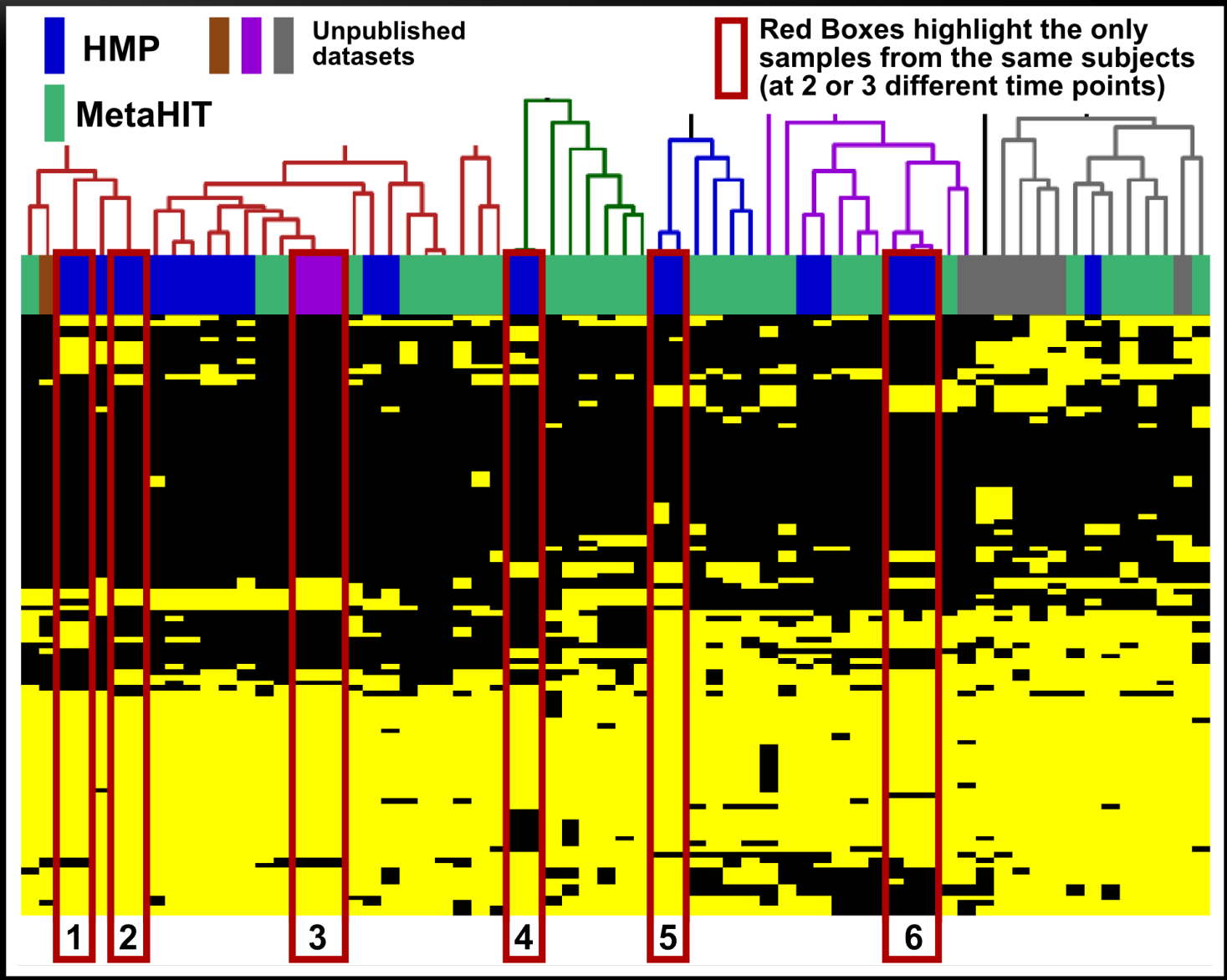


How unique is your personal microbiome?

40 genes in the "same" microbe, *Prevotella copri*

Present

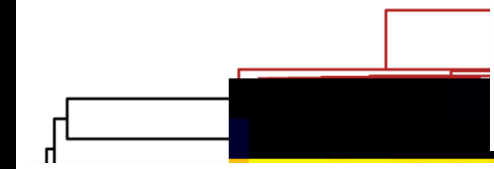
Absent



60 gut microbiomes with abundant *P. copri*

Are there discrete "types" of typical human microbiomes?

http://hmpdacc.org/HMSMCP



Kashiwa Campus, Kashiwa-no-ha 5-1-5, Kashiwa, Chiba, 277-8561, Japan. [14]Division of Bioenvironmental Science, Frontier Science Research Center, University of Miyazaki, 5200 Kiyotake, Miyaz
1692, Japan. [15]Laboratory of Microbiology, Wageningen University, 6710BA Ede, The Netherlands. [16]Tokyo Institute of Technology, Graduate School of Bioscience and Biotechnology, Departm
Biological Information, 4259 Nagatsuta-cho, Midori-ku, Yokohama-shi, Kanagawa Pref. 226-8501, Japan. [17]BGI-Shenzhen, Shenzhen 518083, China. [18]Novo Nordisk Foundation Center for
Biosustainability, Technical University of Denmark, DK-2800 Lyngby, Denmark. [19]Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain. [20]Dep
of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark. [21]Institute of Biomedical Science, Faculty of Health Sciences, University of Copenhagen, DK-2200 Copenhagen, Denma
[22]Hagedorn Research Institute, DK-2820 Gentofte, Denmark. [23]Faculty of Health Sciences, University of Aarhus, DK-8000 Aarhus, Denmark. [24]University of Helsinki, FI-00014 Helsinki, Finland
Delbrück Centre for Molecular Medicine, D-13092 Berlin, Germany.
*These authors contributed equally to this work.
†Lists of authors and affiliations appear at the end of the paper.

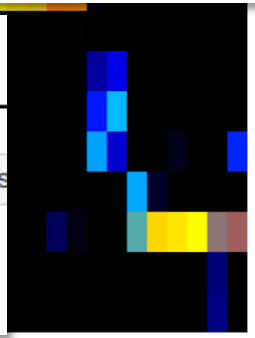
The New York Times

Science

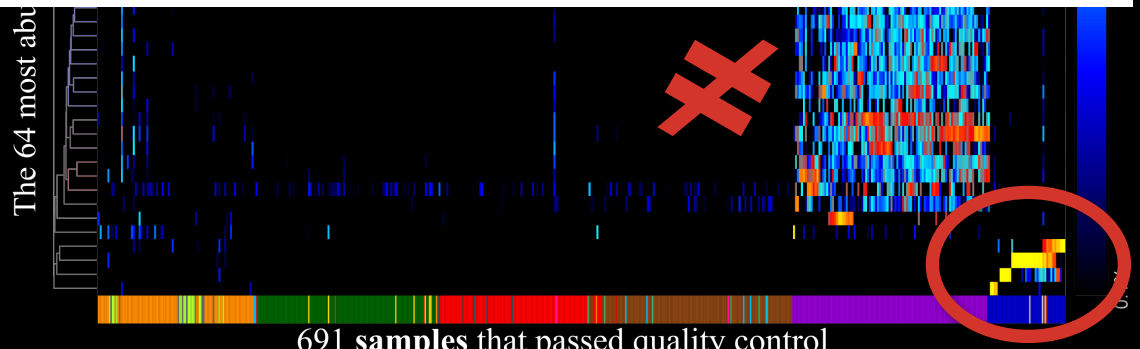
WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS

ENVIRONMENT | SPACE & COS

Bacterial Ecosystems Divide People Into 3 Groups, Scientists Say



Lactobacillus crispatus
Atopobium vaginae
Prevotella amnii
Gardnerella vaginalis
Bifidobacterium breve
Lactobacillus gasseri
Prevotella multiformis
Bifidobacterium dentium



Gut

Vaginal

anterior nares
right retroauricular crease
left retroauricular crease
throat
buccal mucosa
supragingival plaque
subgingival plaque
tongue dorsum
attached keratinized gingiva
saliva
palatine tonsils
stool
posterior fornix
vaginal introitus
mid vagina

Are there discrete "types" of typical human microbiomes?

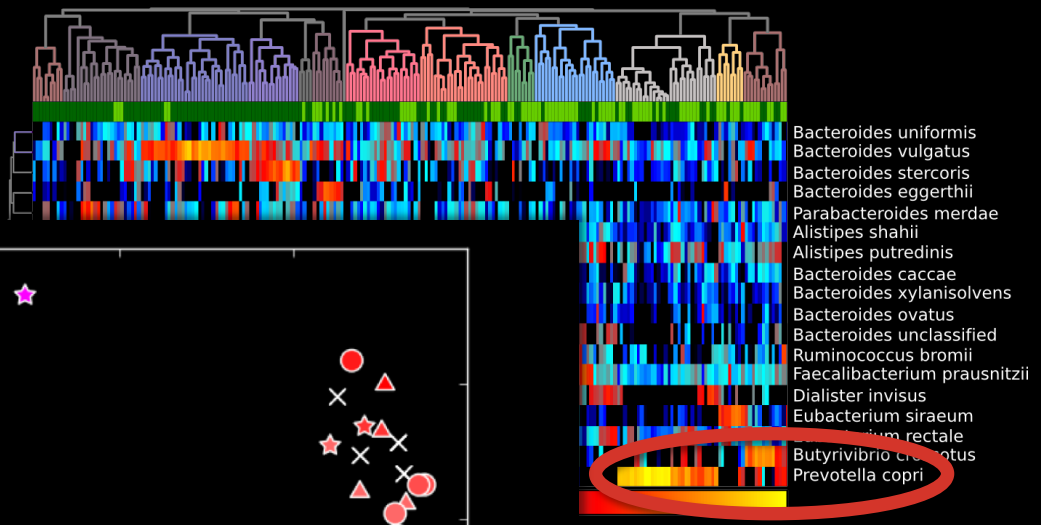
With Ruth Ley, Rob Knight

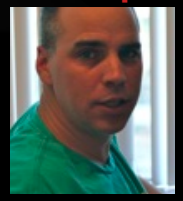
http://huttenhower.sph.harvard.edu/metaphlan

Levi Waldron

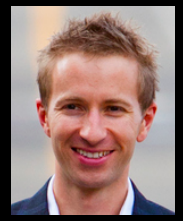
HMP samples

MetaHIT samples

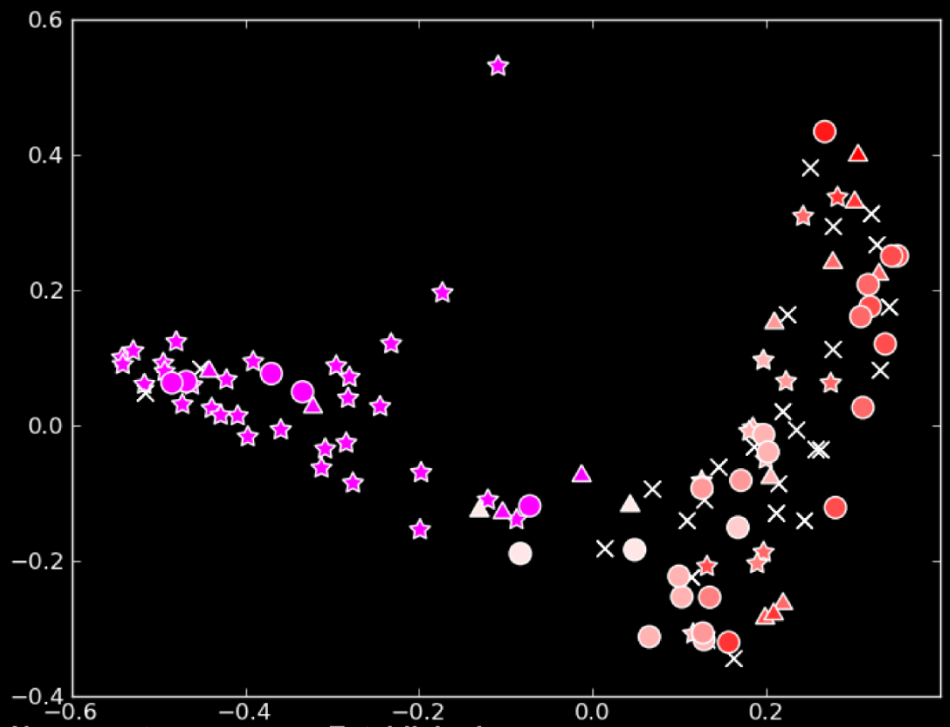


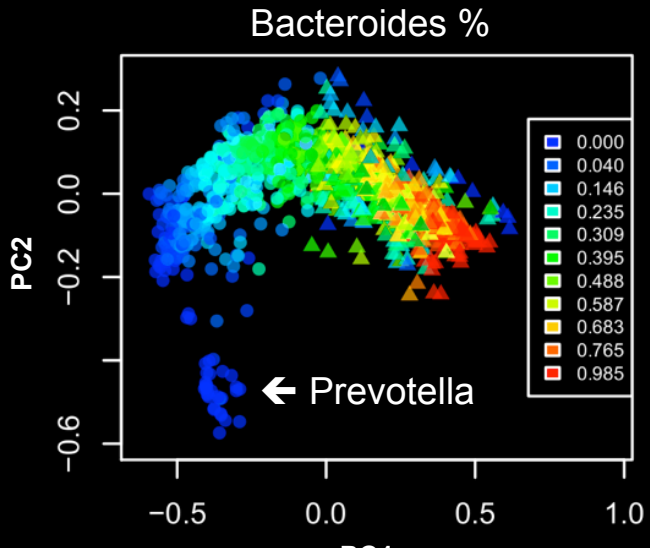


Omry Koren



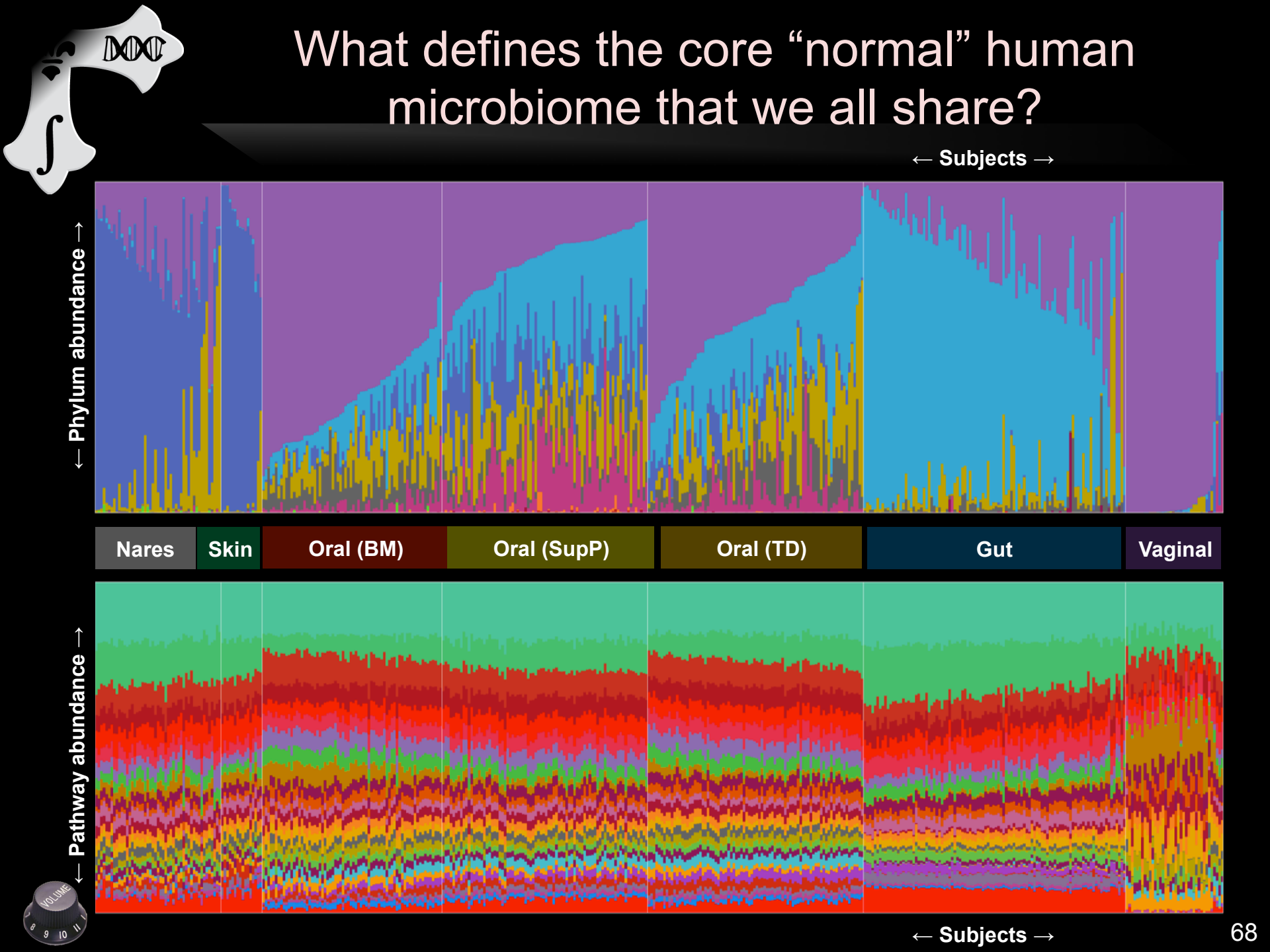
Dan Knights



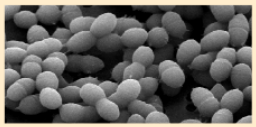
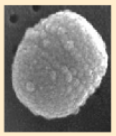


Dan Littman

What defines the core “normal” human microbiome that we all share?

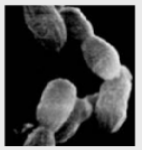


A map of diversity in the human microbiome

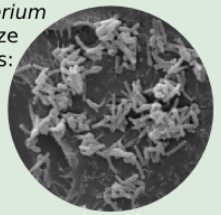
Streptococcus dominates the oral cavity with *S. mitis* > 75% in the cheek

Propionibacterium acnes lives on the skin and nose of most people



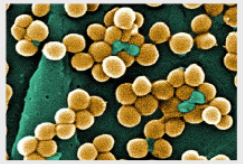
Many *Corynebacterium* species characterize different body sites:

C. matruchoti the plaque
C. accolens the nose
C. croppenstedtii the skin

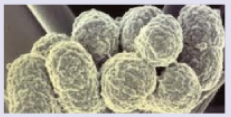


Lactobacillus species (*L. gasseri*, *L. jensenii*, *L. crispatus*, *L. iners*) are predominant but mutually exclusive in the vagina


Staphylococcus epidermidis colonizes external body sites

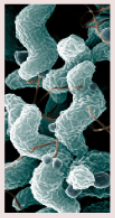


Several *Prevotella* species are present in the gastrointestinal tract. *P. copri* is present in 19% of the subjects and dominates the intestinal flora when present



Microscopy from http://bacmap.wishartlab.com

Bacteroides is the most abundant genus in the gut of almost all healthy subjects



Campylobacter includes opportunistic pathogens, but members live in the oral cavities of most healthy people in the cohort


E. coli is present in the gut of the majority of healthy subjects but at very low abundance

Commensal microbes

Potential pathogens

The four most abundant phyla

- Actinobacteria
- Bacteroidetes
- Firmicutes
- Proteobacteria

Low abundance phyla

- Chloroflexi
- Cyanobacteria
- Euryarchaeota
- Fusobacteria
- Lentisphaerae
- Spirochaetes
- Synergistetes
- Tenericutes
- Thermi
- Verrucomicrobia

National Institutes of Health Human Microbiome Project

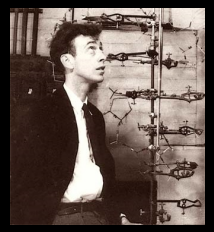
N. Segata & C. Huttenhower
http://huttenhower.sph.harvard.edu
(generated using CirClader and mTol-400 from MetaPhlAn analysis)

What high-impact outcomes can we reasonably expect from the microbiome?

Translation

Risk diagnosis

- Lifetime, onset, activity, and outcome
- Prospective epi. study design
- Dense longitudinal measures, multiple nested outcomes



Treatment

- More and simpler model systems
- Systematic understanding of current models
- More perturbation experiments, “knock ins” and “knock outs”

Public health and policy

Health policy

- Early life exposures
- Pharma. best practices

Ethics and privacy

- Identifiability
- Tracking

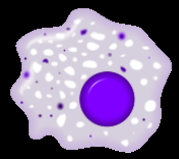
Basic biology and molecular mechanism

Microbial experiments

- Systematic application of computational tools
- Meta-analysis of genomes and metagenomes

Host-microbe-microbiome interactions

- Immunity in more human tissues
- Non-immune mechanisms (metabolites, peptides)

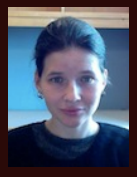


Thanks!

http://huttenhower.sph.harvard.edu


Alex Kostic


Levi Waldron


Xochitl Morgan


Tim Tickle


Daniela Boernigen


Lauren McIver

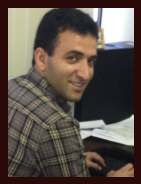

George Weingart


Emma Schwager


Eric Franzosa


Boyu Ren


Tiffany Hsu


Ali Rahnavard


Ayshwarya Subramanian


Jim Kaminski


Regina Joice


Koji Yasuda


Kevin Oh


Galeb Abu-Ali


Afrah Shafquat


Randall Schwager


Chengwei Luo


Keith Bayer


Moran Yassour


Alexandra Sirota


Dirk Gevers


Ramnik Xavier


Wendy Garrett


Andy Chan

Human Microbiome Project 2

Lita Procter
Jon Braun
Dermot McGovern
Subra Kugathasan
Ted Denson
Janet Jansson
Bruce Birren
Chad Nusbaum
Clary Clish
Joe Petrosino
Thad Stappenbeck

BROAD INSTITUTE

Human Microbiome Project

Jane Peterson
Sarah Highlander
Barbara Methe
Karen Nelson
George Weinstock
Owen White


Nicola Segata


Gautam Dantas
Molly Gibson


Katherine Lemon


Brendan Bohannan
James Meadow