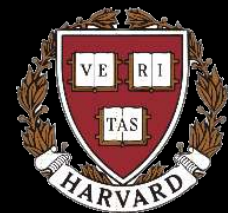




An Introduction to Microbial Community Analyses

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. GREAVES¹

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 Handelsman¹, Michelle R Rondon¹, Sean F Brady², Jon Clardy² and Robert M Goodman¹



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

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

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



Examples of metagenomic studies: Global Ocean Sampling

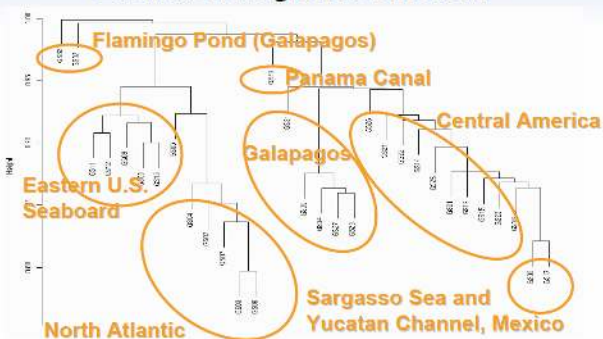
J. Craig Venter
INSTITUTE

2003/2004 - ongoing

The Sorcerer II Expedition Global Ocean Sampling Route

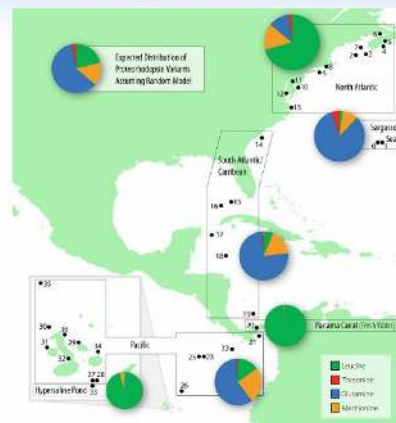


The Biodiversity of Each New Region is Different



J. Craig Venter
INSTITUTE

Proteorhodopsins Vary by Region



JTC Sequencer Lab

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





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 Wino-gradsky (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 Staining¹

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

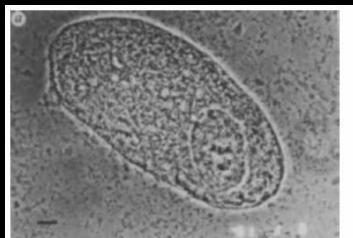
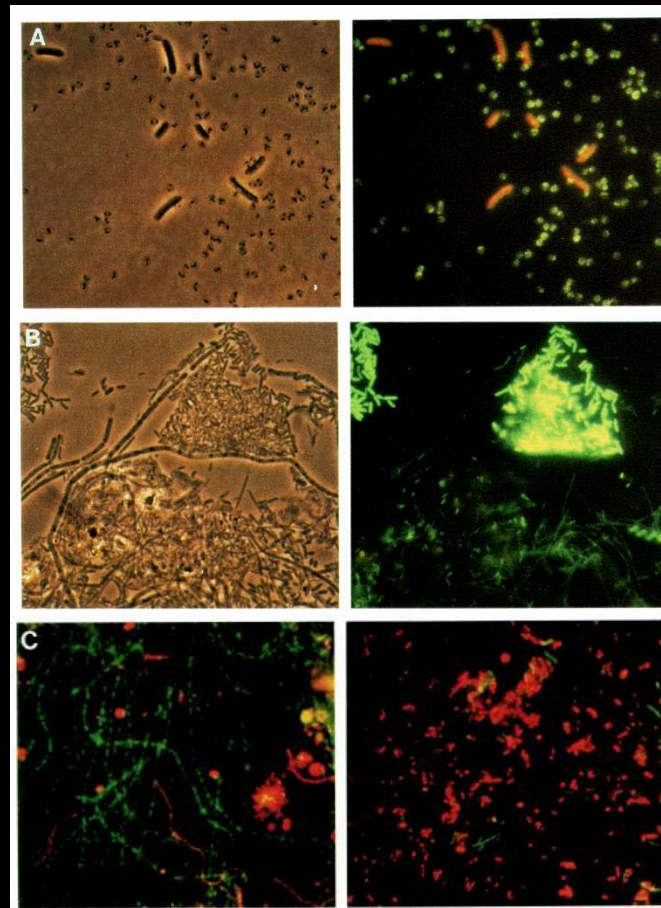
Rudolf Amann, Nina Springer, Wolfgang Ludwig,
Hans-Dieter Görtz* & Karl-Heinz Schleifer

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

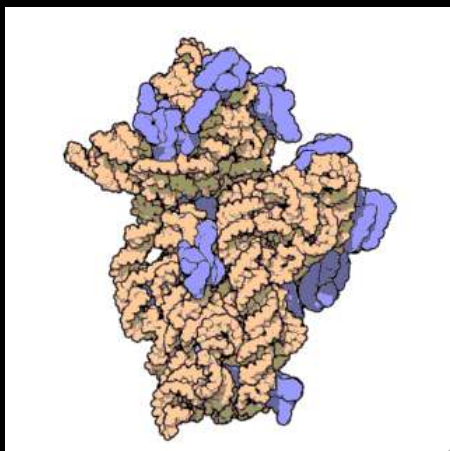
RUDOLF I. AMANN,* WOLFGANG LUDWIG, AND KARL-HEINZ SCHLEIFER
Lehrstuhl für Mikrobiologie, Technische Universität München, D-80290 Munich, Germany



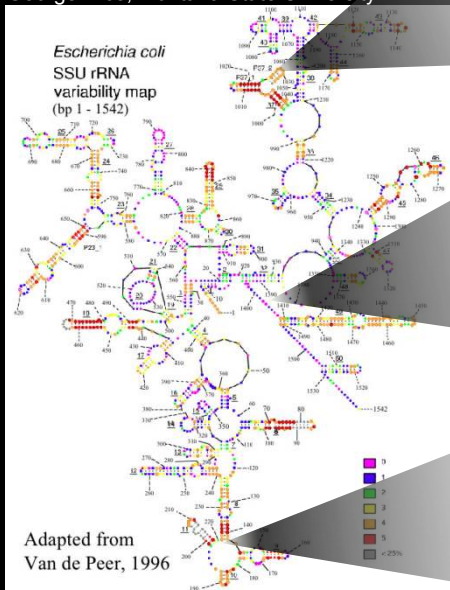


Metagenomic methods: 16S rRNA gene

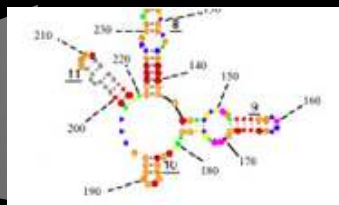
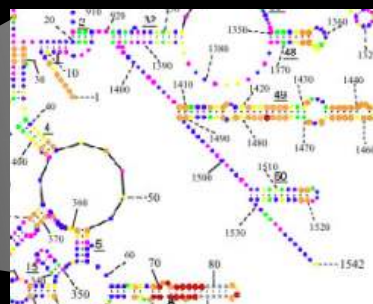
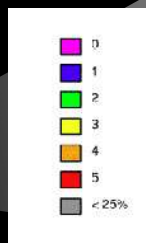
- Structural component of the prokaryotic ribosome
- Used as molecular clock to identify phylogeny:
 - Large, good scale for mutations
 - Range of mutation rates
 - Portions are constant, allowing amplification
- Not single copy! Researcher beware...



George Rice, Montana State University

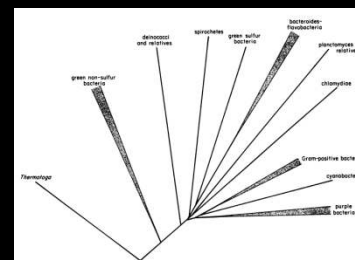


V6

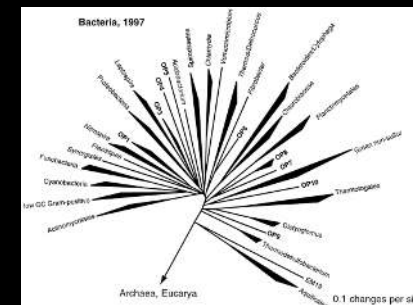


V2

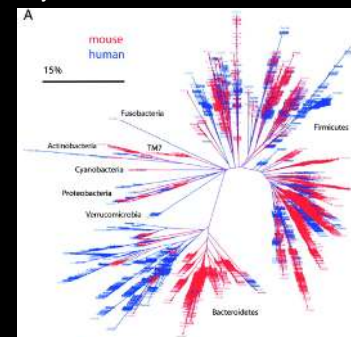
Woese, 1987



Pace, 1997



Ley, 2006





Metagenomic methods: shotgun sequencing

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

THOMAS M. SCHMIDT,¹ EDWARD F. DeLONG,¹ 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

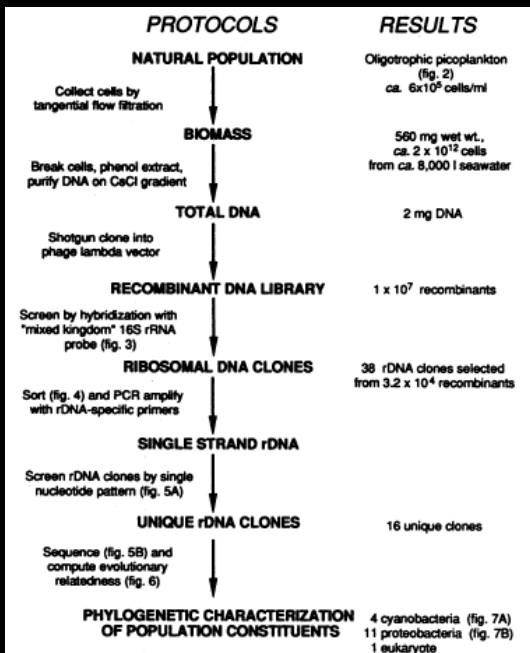


FIG. 1. Flow chart of the protocols used to characterize marine picoplankton without cultivation and a summary of some results.

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

JEFFREY L. STEIN,^{1*} TERENCE L. MARSH,² KE YING WU,³ HIROAKI SHIZUYA,⁴ AND EDWARD F. DeLONG*

Recombinant Bio-Catalysis, Inc., La Jolla, California 92037; ²Microbiology Department, University of Illinois, Urbana, Illinois 61801; ³Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, California 93106; and ⁴Division of Biology, California Institute of Technology, Pasadena, California 91125

Received 14 July 1995/Accepted 14 November 1995

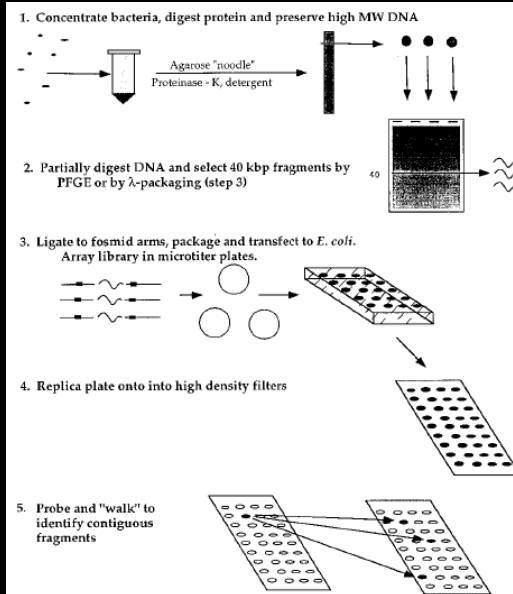


FIG. 1. Flowchart depicting the construction and screening of an environmental library from a mixed picoplankton sample. MW, molecular weight; PFGE, pulsed-field gel electrophoresis.

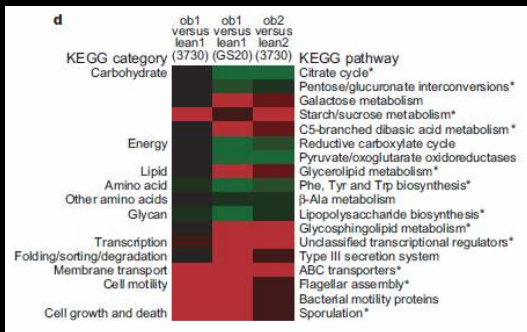
Environmental Genome Shotgun Sequencing of the Sargasso Sea

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. Smith¹

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.

JTC Sequencer Lab

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



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

Peter J. Turnbaugh¹, Ruth E. Ley¹, Michael A. Mahowald², Vincent Magrini², Elaine R. Mardis^{1,2} & Jeffrey I. Gordon¹

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

Sample	Contigs	Average contig length	Contig 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

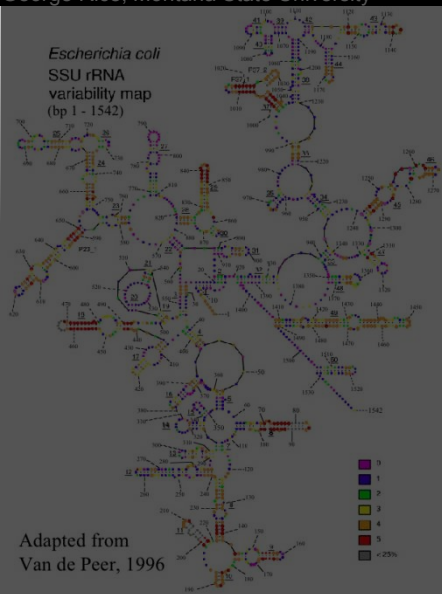


Lyse cells
Extract DNA (and/or RNA)

Amplicons

Meta'omic

George Rice, Montana State University



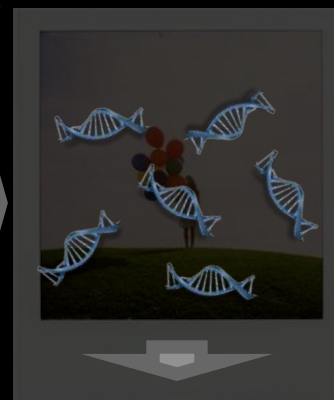
PCR to amplify a single
marker gene, e.g. 16S rRNA

Hello
my name is
Classify sequence
→ microbe

Samples

Microbes

Relative
abundances



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

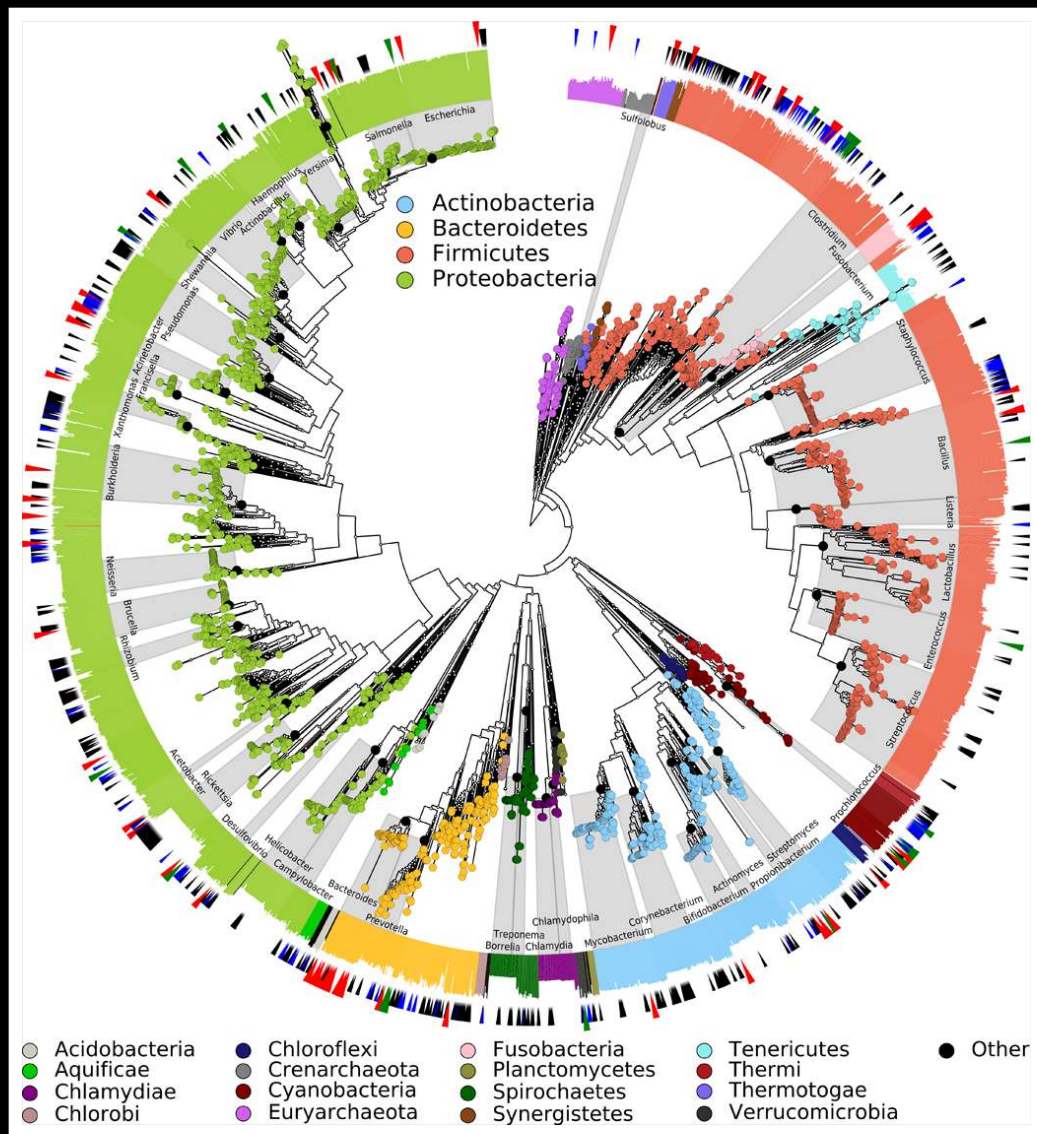
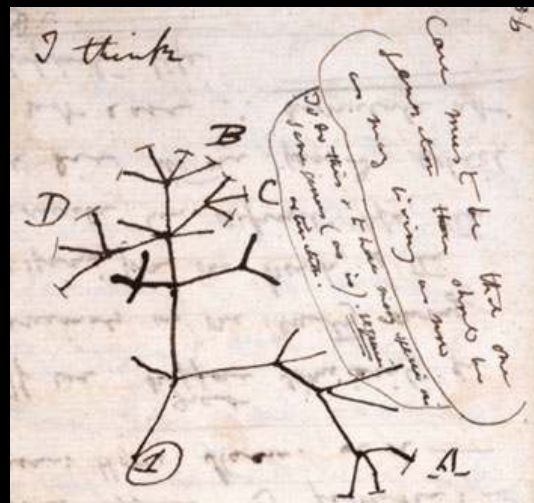
Public health tool
monitoring
population health
and epidemiology

Diagnostic or
prognostic
biomarker for
host disease





Composition-based analyses





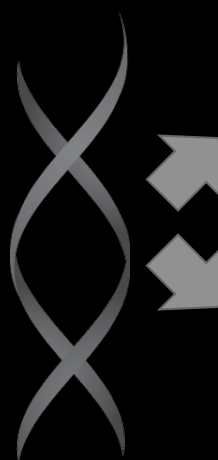
Microbiome composition analyses: phylotypes and binning

Binning: nontrivial
assignment of reads
to phylotypes

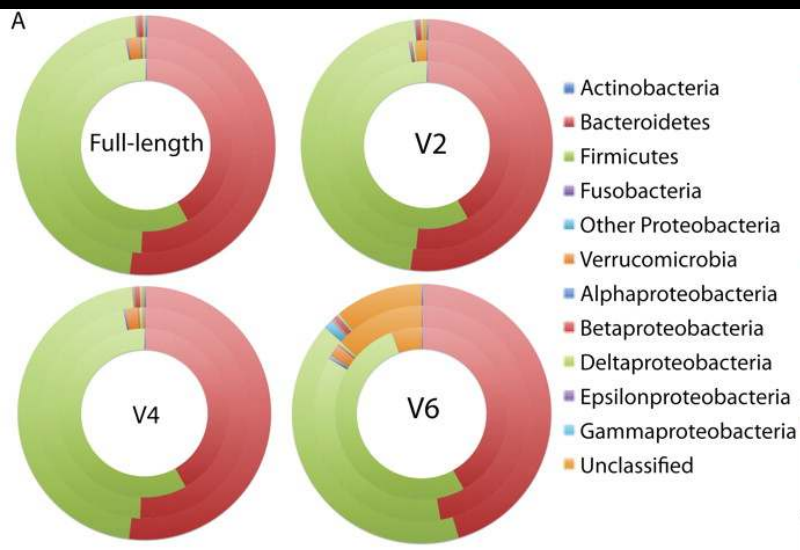
**Phylotype or operational
taxonomic unit (OTU):**
organisms clonal to within some
tolerance (e.g. 95%); “species”

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

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



rdp.cme.msu.edu

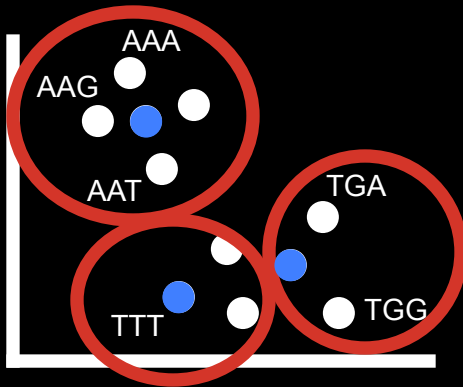




Microbiome composition analyses: OTU clustering

Open reference

Clustering

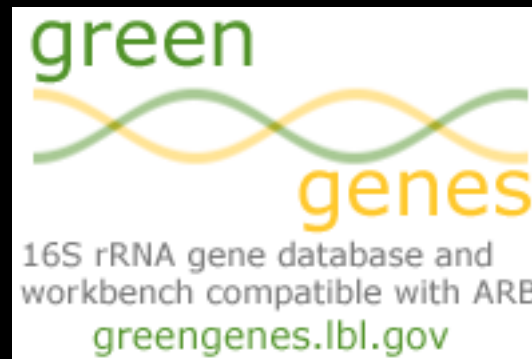
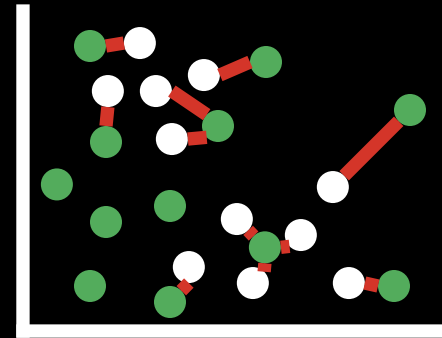


```
>Uniq1  
AAA  
>Uniq2  
TGA  
>Uniq3  
TTT  
...
```



Closed reference

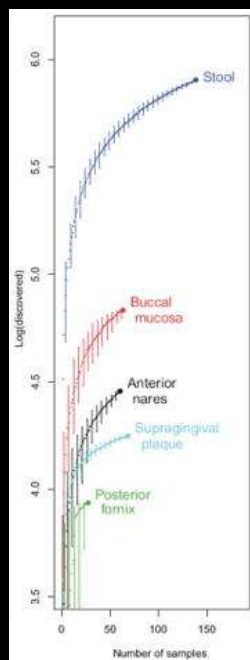
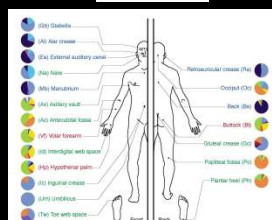
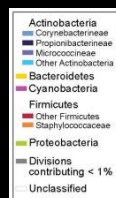
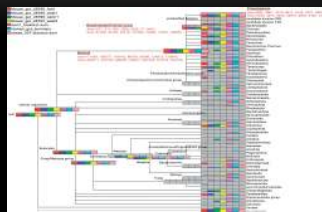
Classification





Microbiome composition analyses: diversity

Mitra, 2009

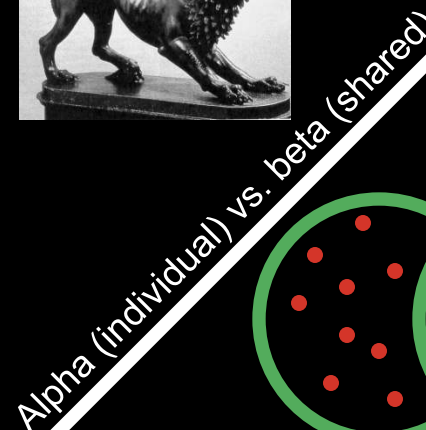


Taxonomic vs. phylogenetic

Chimeric reads



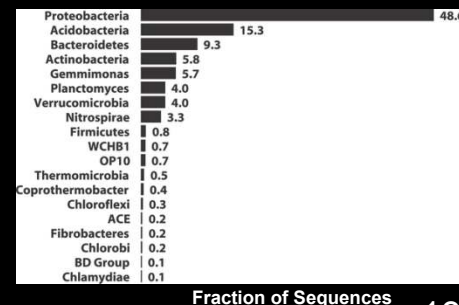
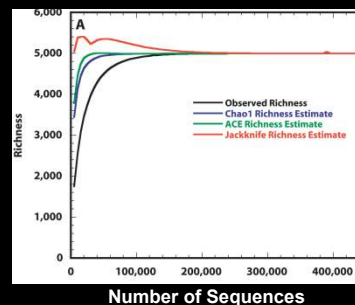
Hamady, 2009



Qualitative vs. quantitative



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



Schloss, 2006



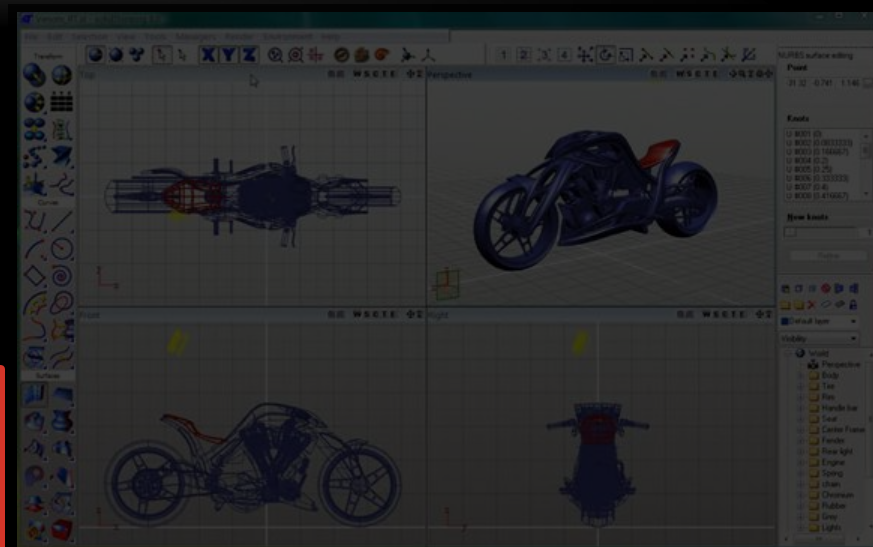
Microbiome composition analyses: ordination

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

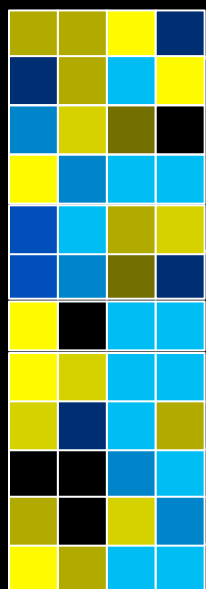
PCA or **Principal Component Analysis** guarantees the new dimensions maximize normal variation.



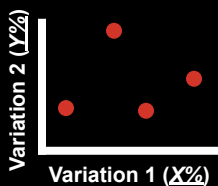
NMDS or **Nonmetric Multidimensional Scaling**, also called **PCoA** or **Principal Coordinates Analysis**, guarantees the new dimensions maximize an arbitrary similarity score (such as UniFrac beta-diversity).



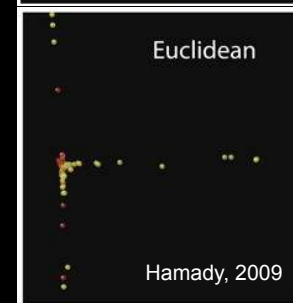
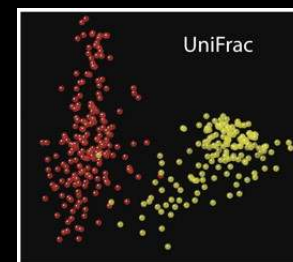
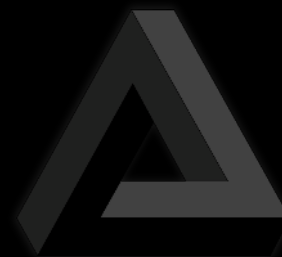
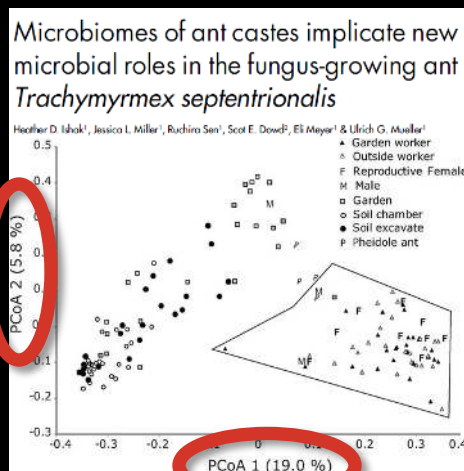
Samples →



Distance between points is Euclidean



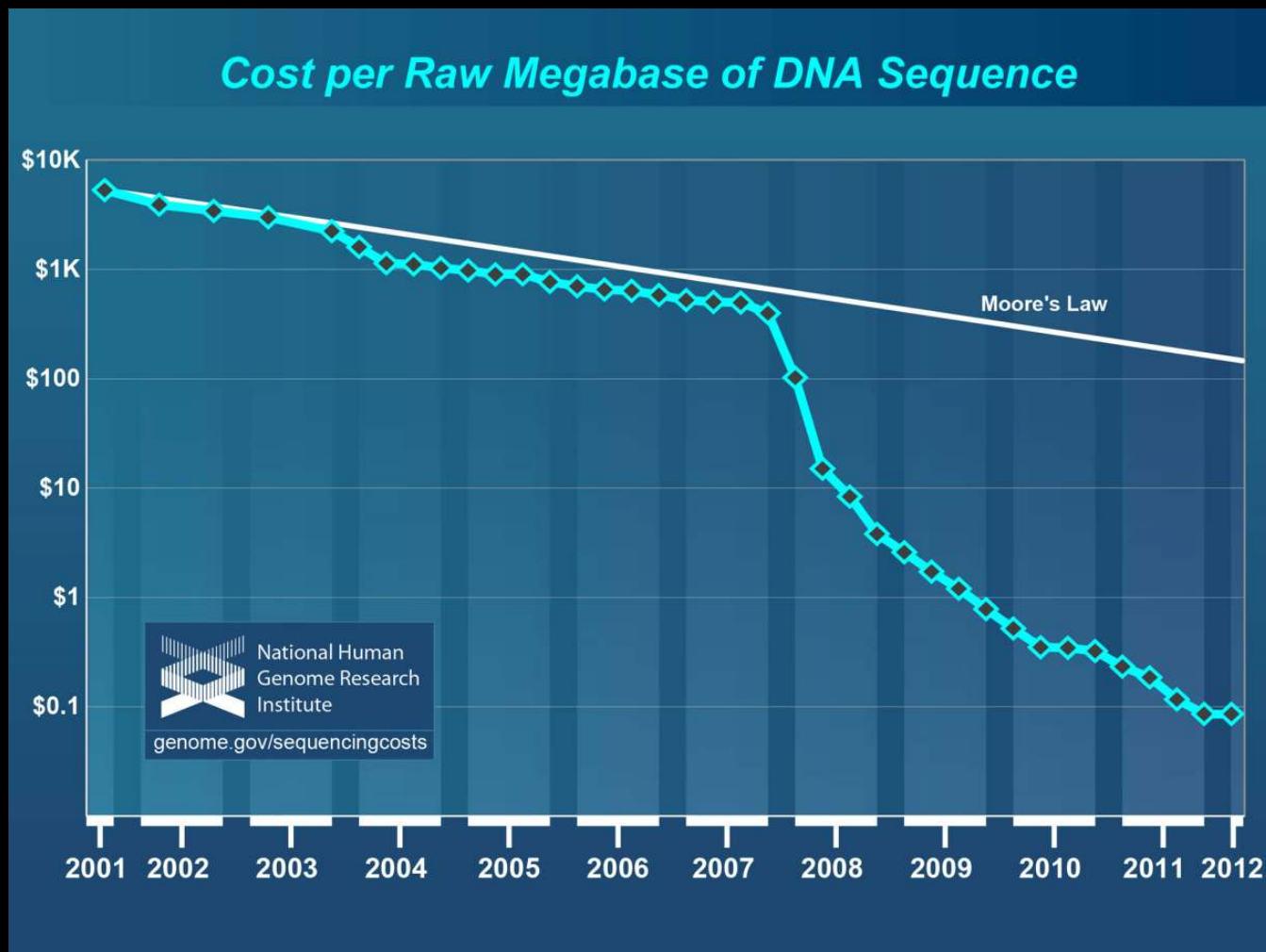
Distance between points is a **proportional function** of their **similarity**



Hamady, 2009



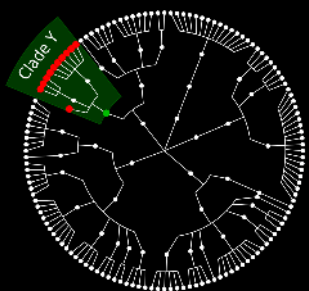
Meta'omic analyses





Typical shotgun metagenome and metatranscriptome analyses

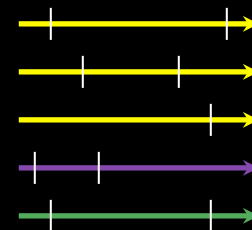
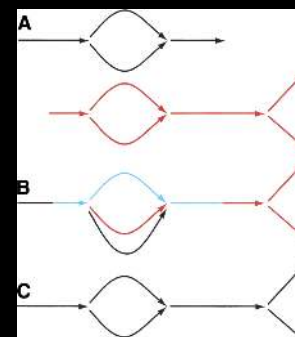
Taxonomic Profiling



Functional Profiling



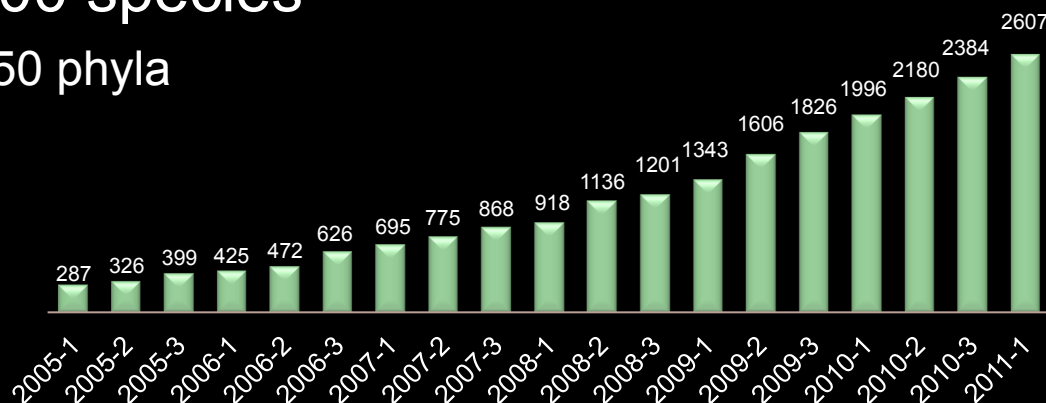
Assembly





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**

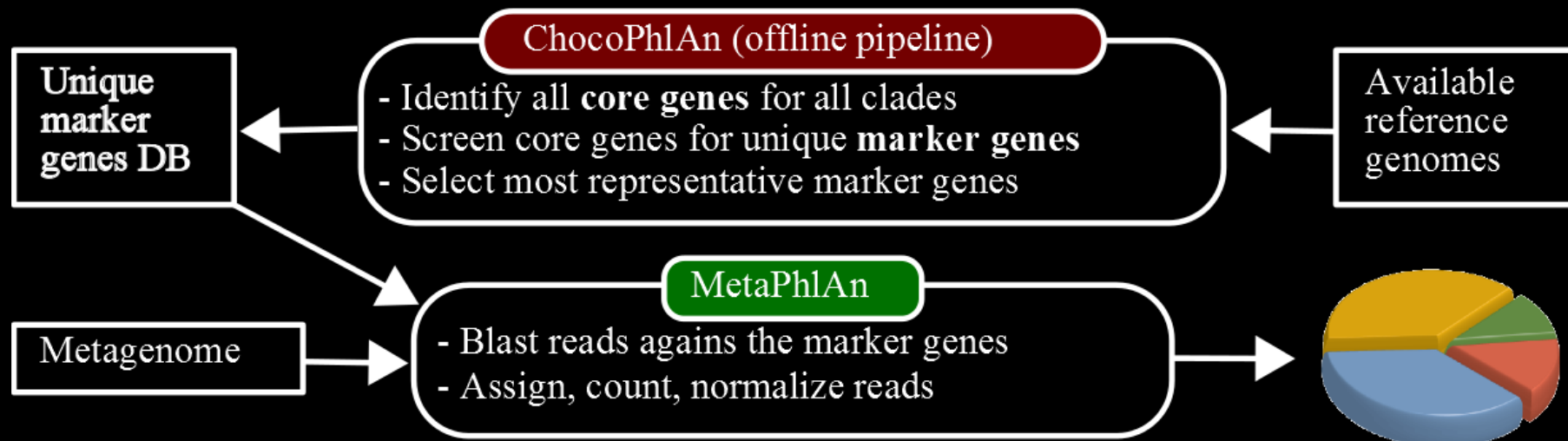


- These genes and genomes are a tremendous resource to:
 - Identify conserved markers that can be used to infer phylogeny
 - Identify unique markers that can be used to infer taxonomy
 - Relate the microbial members of a community to their metagenomic functional potential



MetaPhlAn2: Taxonomic profiling using unique marker gene sequences

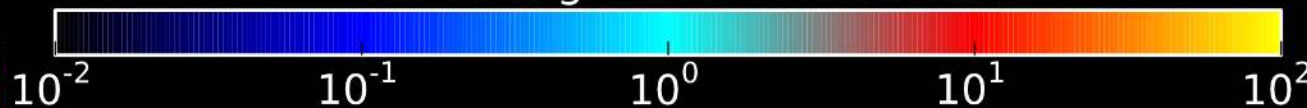
- ~1M most representative markers used for identification
 - 184±45 markers per species (target 200)
- ~7,100 species (excludes incomplete annotations, spp., etc.)
- FP/FN rates of ~1 in 10⁶
- Profiles all domains of life: bacteria, viruses, euks, archaea
- Strain level profiling using marker combination barcodes
- Quasi-markers used to resolve ambiguity in postprocessing





MetaPhlAn2: Taxonomic profiling using unique marker gene sequences

Percentage Relative Abundance



- Body site
- LeftRetroauricular crease
 - Hard palate
 - Palatine Tonsils
 - RightRetroauricular crease
 - Stool
 - Buccal mucosa
 - Vaginal introitus
 - Throat
 - Mid vagina
 - Anterior nares
 - Keratinized gingiva
 - Subgingival plaque
 - Supragingival plaque
 - Posterior fornix
 - Saliva
 - Tongue dorsum

Dataset

- 1
- 2

Gender

- Male
- Female

Visit number

- 1
- 2
- 3

Visit number

Body site

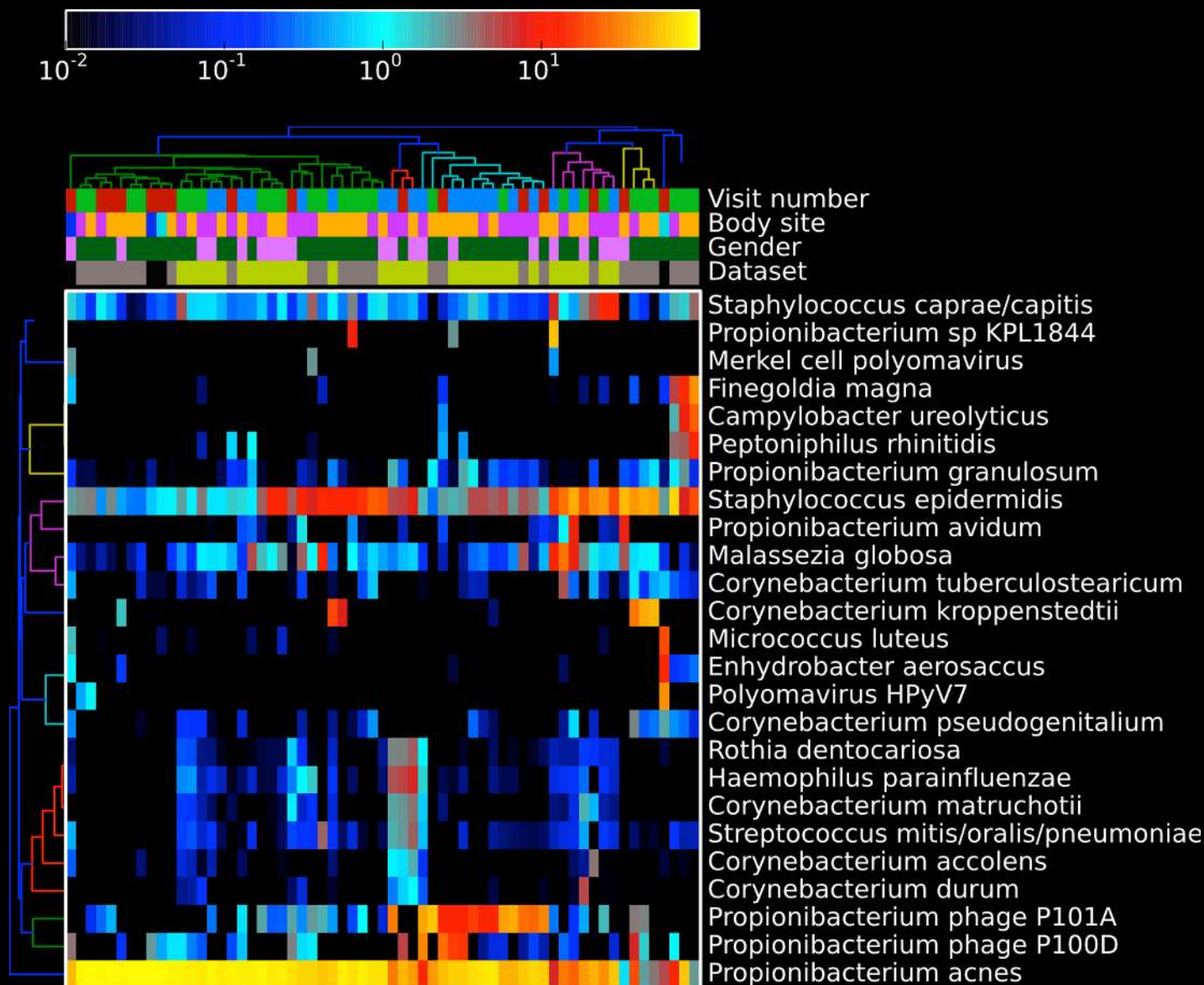
Gender

Dataset

Lactobacillus jensenii
Lactobacillus iners
Gardnerella vaginalis
Lactobacillus crispatus
Lactobacillus gasseri
Propionibacterium phage P101A
Propionibacterium phage P100D
Propionibacterium acnes
Propionibacteriaceae unclassified
Staphylococcus epidermidis
Dolosigranulum pigrum
Corynebacterium propinquum
Corynebacterium accolis
Staphylococcus aureus
Alistipes shahii
Alistipes putredinis
Barnesiella intestinihominis
Parabacteroides merdae
Bacteroides uniformis
Subdoligranulum unclassified
Faecalibacterium prausnitzii
Bacterium rectale
Ruminococcus bromii
Alistipes onderdonkii
Dialister invisus
Bacteroides dorei
Bacteroides cellulosilyticus
Bacteroides ovalis
Bacteroides fragilis
Bacteroides thetaiotaomicron
Bacteroides xylanisolvens
Bacteroides vulgatus
Parabacteroides distasonis
Bacteroides caccae
Bacteroides stercoris
Bacteroides sp 2.1.22
Bacteroides massiliensis
Bacteroides finegoldii
Butyrivibrio crossotus
Prevotella copri
Veillonella sp oral taxon 780
Granulicatella elegans
Streptococcus mitis/oralis/pneumoniae
Streptococcus phage E1
Streptococcus vestibularis
Fusobacterium periodonticum
Streptococcus infantis
Prevotella nanceiensis
Porphyromonas sp oral taxon 279
Haemophilus parainfluenzae
Veillonella sp oral taxon 158
Rothia mucilaginis
Neisseria flavescens
Veillonella unclassified
Prevotella melanogenica
Streptococcus salivarius
Streptococcus parasanguinis
Prevotella hisidola
Actinomyces sp. IC47
Actinomyces graevenitzii
Neisseria unclassified
Streptococcus sanguinis
Rothia dentocariosa
Rothia aeria
Corynebacterium durum
Lautropia mirabilis
Capnocytophaga sp oral taxon 329
Capnocytophaga gingivalis
Capnocytophaga unclassified
Corynebacterium matruchotii
Fusobacterium nucleatum
Neisseria elongata
Veillonella parvula
Actinobaculum sp oral taxon 183



MetaPhlAn2: Taxonomic profiling using unique marker gene sequences



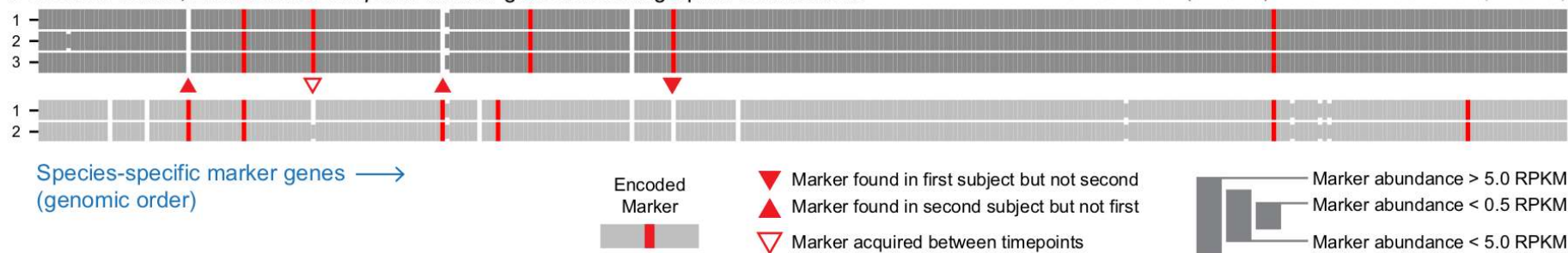


Gene-based fingerprints capture strain variation in individuals' most abundant (stable) bugs

Posterior fornix, *Lactobacillus crispatus* marker gene-based fingerprint contributions

FIRST SUBJECT (3 VISITS)

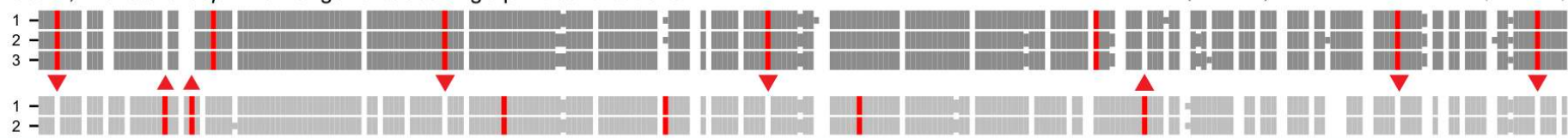
SECOND SUBJECT (2 VISITS)



Stool, *Prevotella copri* marker gene-based fingerprint contributions

FIRST SUBJECT (3 VISITS)

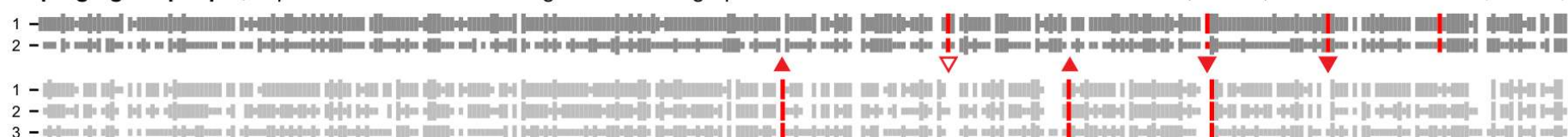
SECOND SUBJECT (2 VISITS)



Supragingival plaque, *Leptotrichia buccalis* marker gene-based fingerprint contributions

FIRST SUBJECT (2 VISITS)

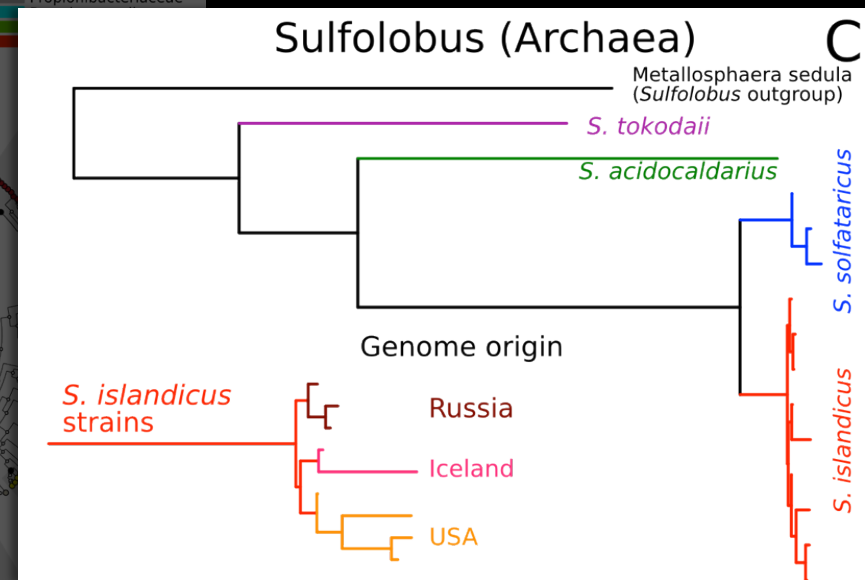
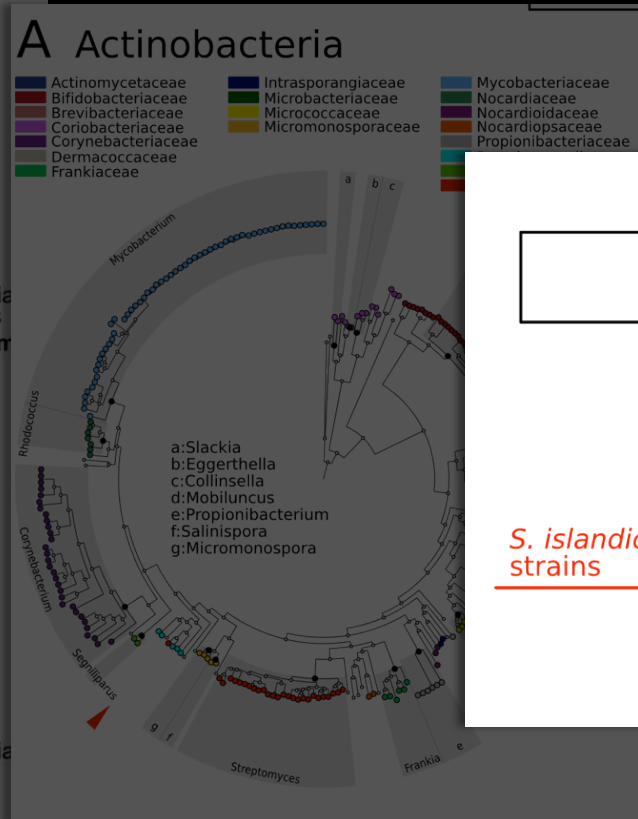
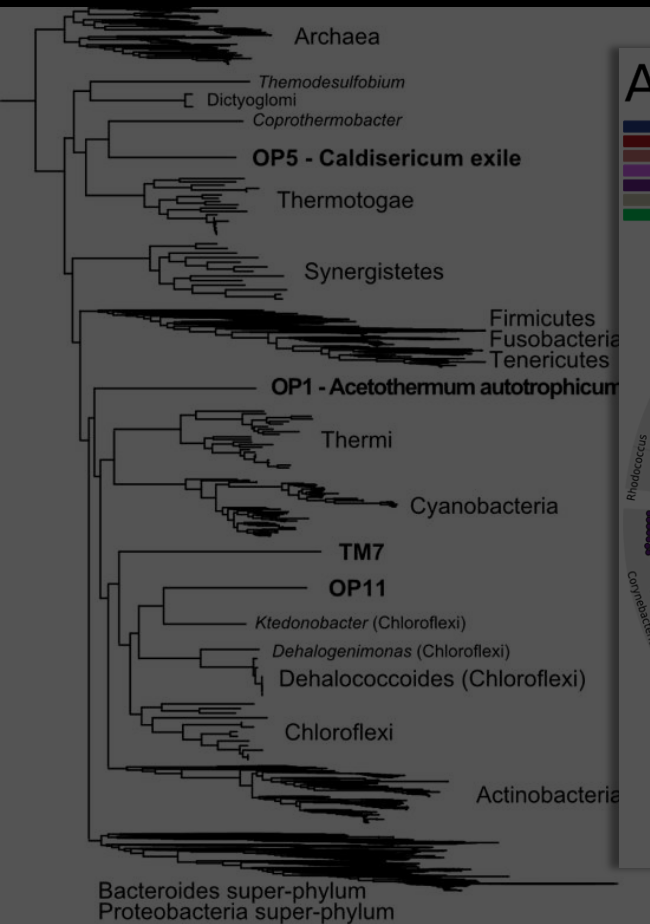
SECOND SUBJECT (3 VISITS)





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



- Acidobacteria
 Chloroflexi
 Fusobacteria
 Tenericutes
 Other
- Aquificae
 Crenarchaeota
 Planctomycetes
 Thermi
- Chlamydiae
 Cyanobacteria
 Spirochaetes
 Thermotogae
- Chlorobi
 Euryarchaeota
 Synergistetes
 Verrucomicrobia



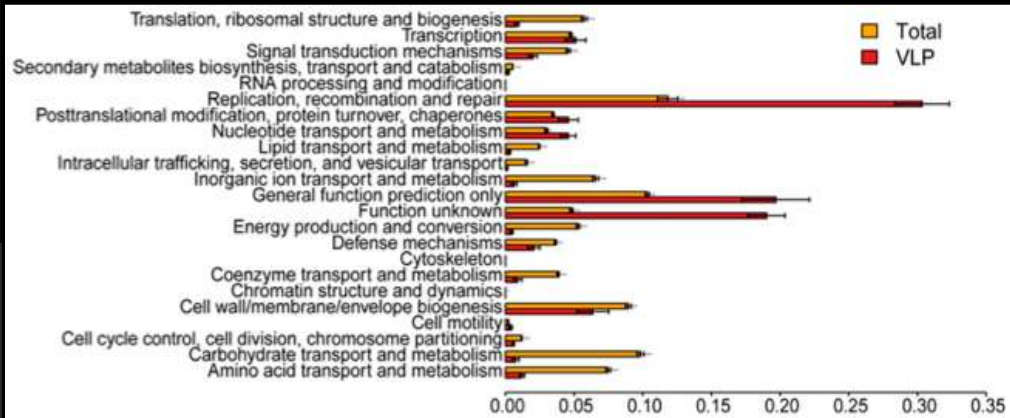
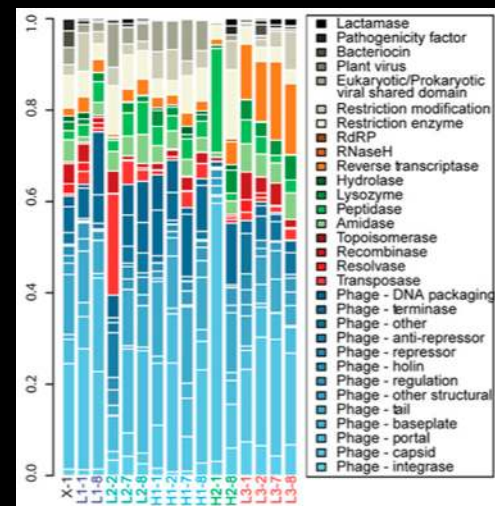
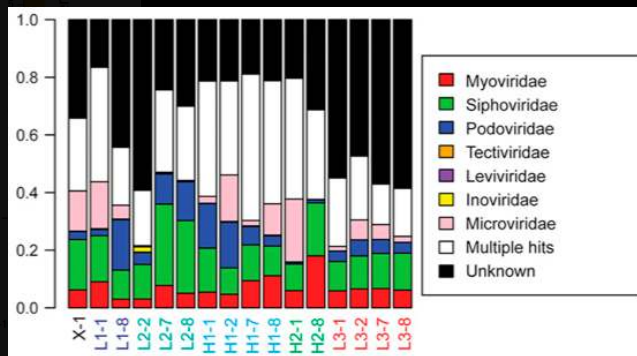
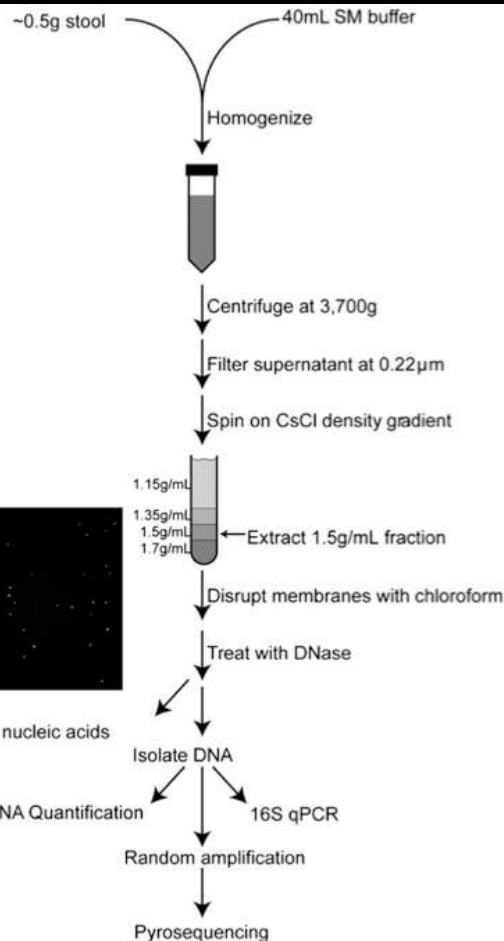
-



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

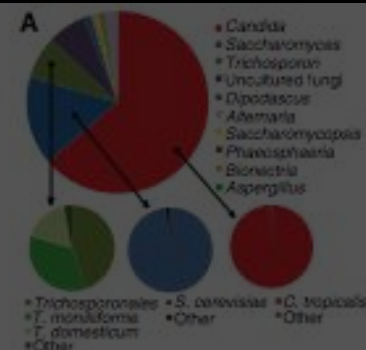
Alejandro Reyes¹, Matthew Haynes², Nicole Hanson², Florent E. Angly^{2,3}, Andrew C. Heath⁴, Forest Rohwer² & Jeffrey I. Gordon¹



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

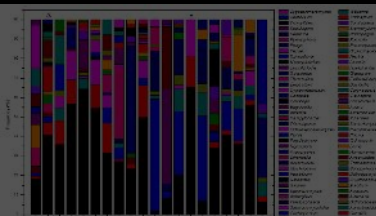
Interactions Between Commensal Fungi and the C-Type Lectin Receptor Dectin-1 Influence Colitis

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,³ Phillip R. Fleshner,⁴ Marla Dubinsky,^{3,5} Jerome I. Rotter,² Hanlin L. Wang,⁶ Dermot P. B. McGovern,^{1,2} Gordon D. Brown,⁷ David M. Underhill^{3,6,8*}



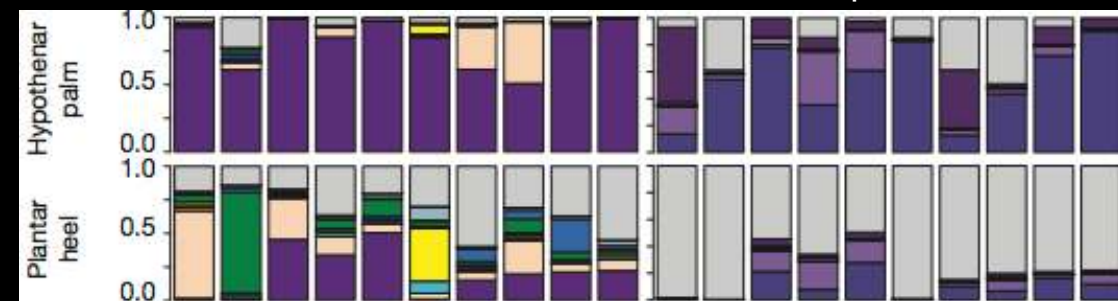
Characterization of the Oral Fungal Microbiome (Mycobiome) in Healthy Individuals

Mahmoud A. Ghannoum^{1*}, Richard J. Jurevic², Pranab K. Mukherjee¹, Fan Cui¹, Masoumeh Sikaroodi³, Ammar Naqvi³, Patrick M. Gillevet³



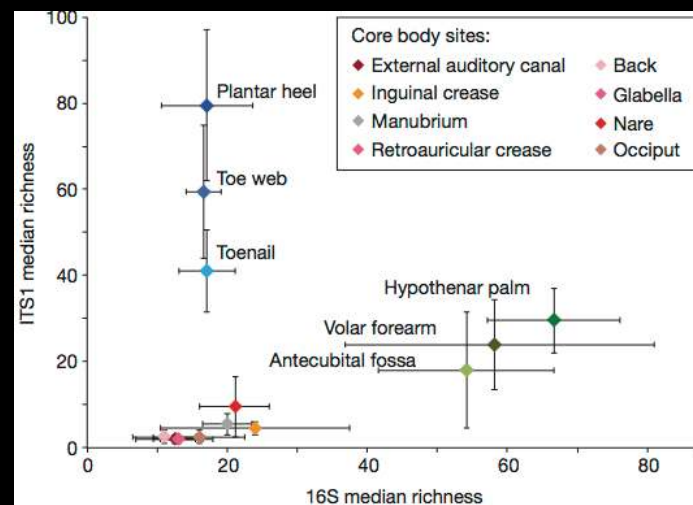
Genus

Species



Topographic diversity of fungal and bacterial communities in human skin

Keisha Findley¹, Julia Oh¹, Joy Yang¹, Sean Conlan¹, Clayton Deming¹, Jennifer A. Meyer³, Deborah Schoenfeld², Effie Nomicos², Morgan Park³, NIH Intramural Sequencing Center Comparative Sequencing Program[†], Heidi H. Kong^{2*} & Julia A. Segre^{3*}





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...

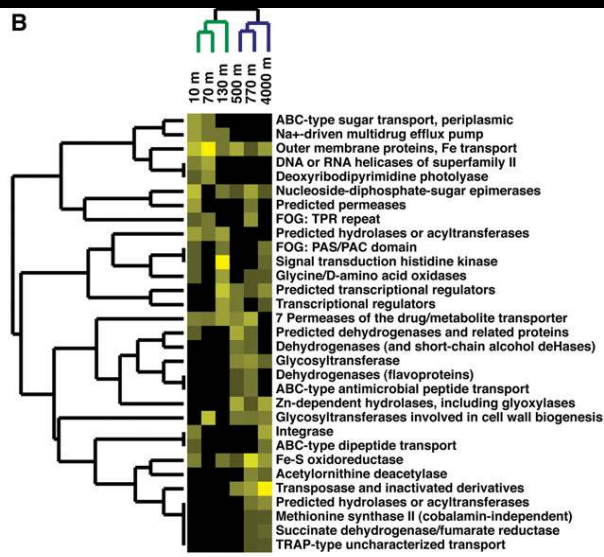
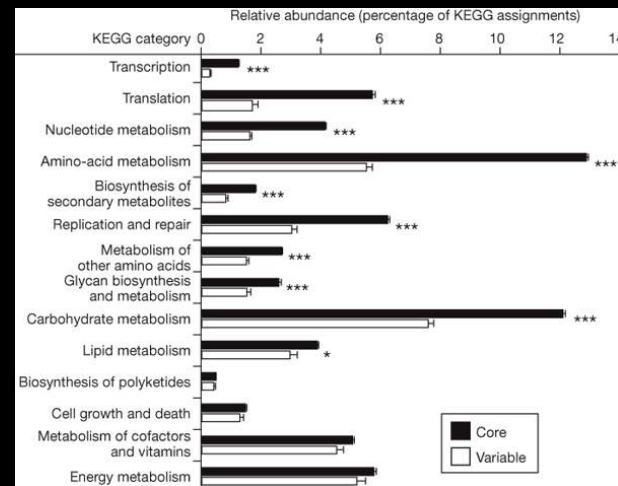


Table 1 | Glycoside hydrolases and carbohydrate-binding modules

CAZy family*	Plant HMM name†	Known activities‡	Termite gut community§
Glycoside hydrolase catalytic domains****			
GH1	Glyco_hydro_1	β-Galactosidase, 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

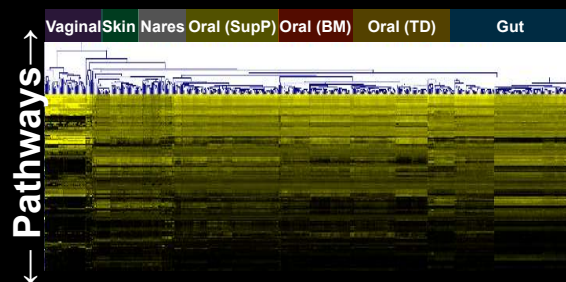


Turnbaugh, 2009

DeLong, 2006



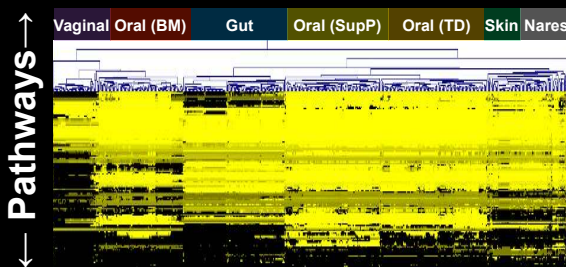
Microbiome meta'omic analyses: metabolic profiling (with HUMAnN)



← Samples →

**Pathway
abundance**

**Pathway
coverage**



← Samples →



100 subjects

1-3 visits/subject

~7 body sites/visit

10-200M reads/sample

100bp reads



BLAST



Functional seq.
KEGG + MetaCyc
CAZy, TCDB,
VFDB, MEROPS...

**Metagenomic
reads**

?

**Enzymes and
pathways**

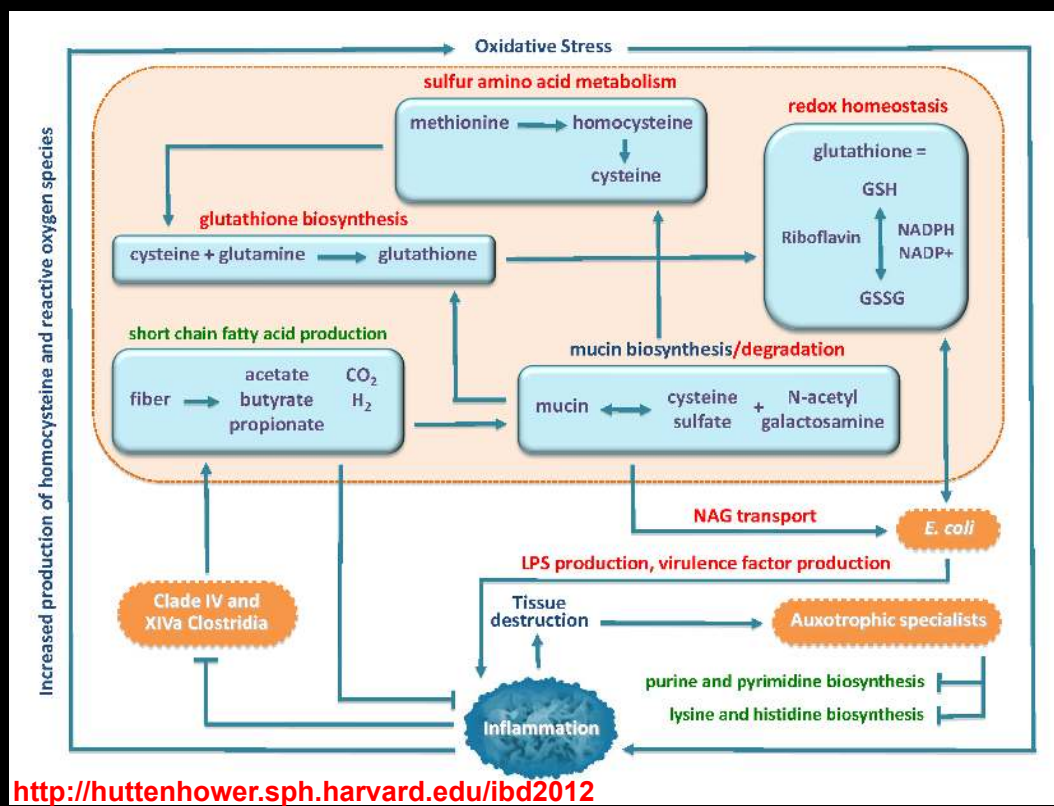
HUMAnN
HMP Unified Metabolic
Analysis Network

<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.





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

LEfSe:

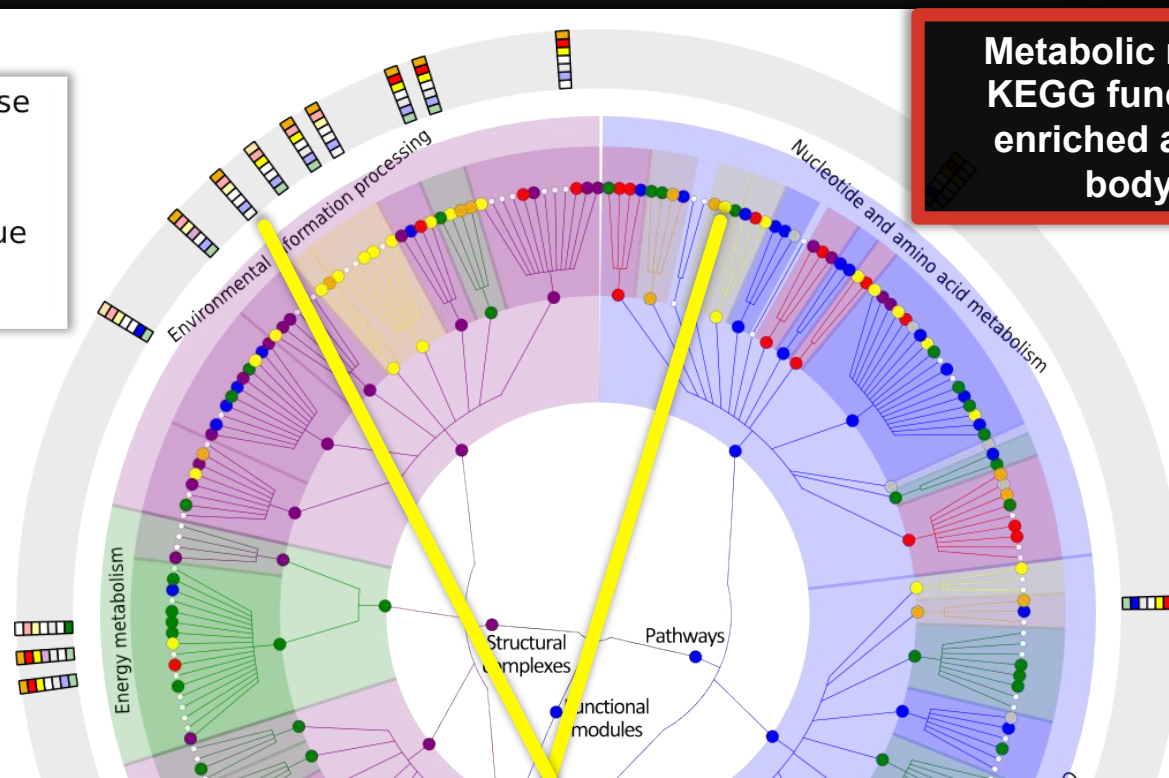
LDA Effect Size

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**

- Retroauricular crease
- Stool
- Anterior nares
- Posterior fornix
- Supragingival plaque
- Buccal mucosa
- Tongue dorsum



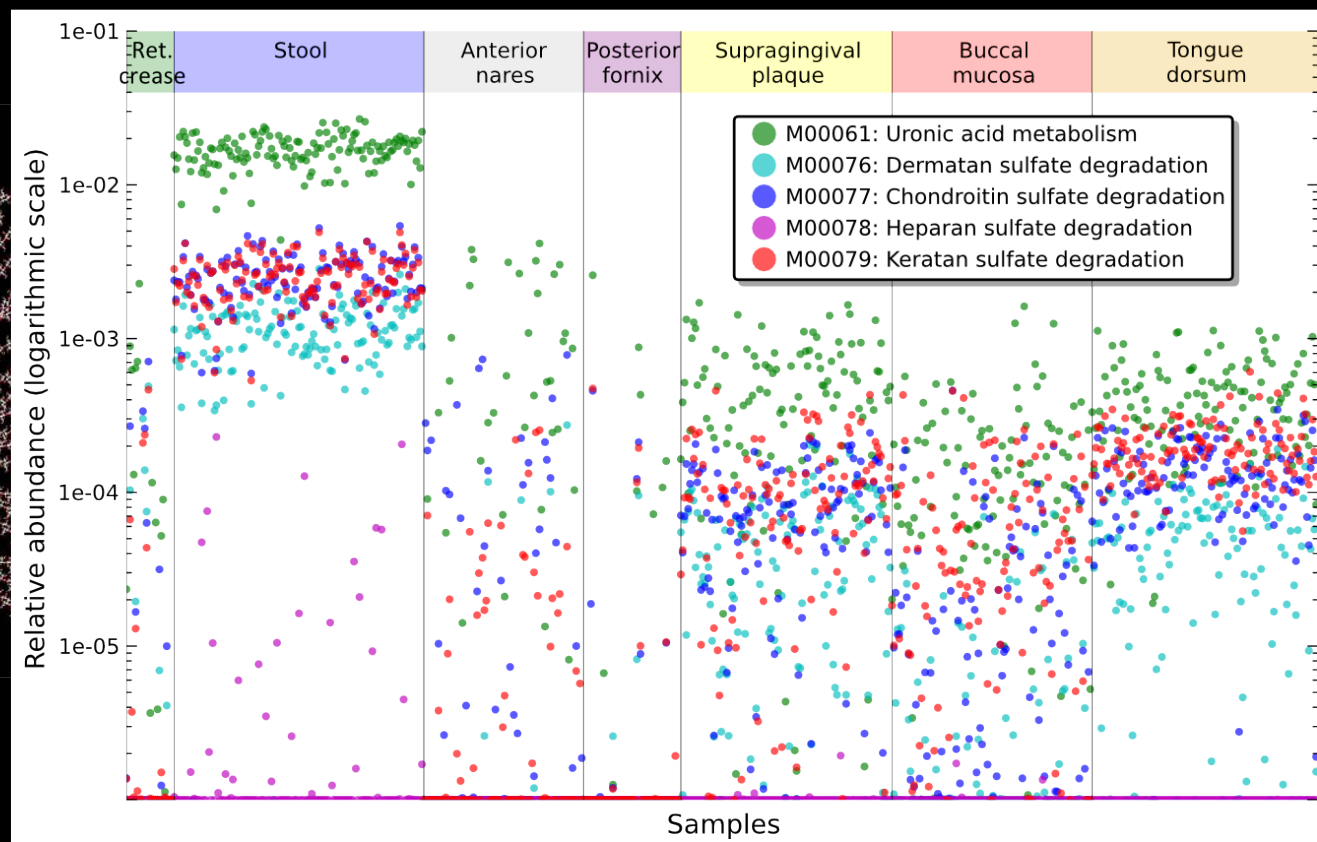
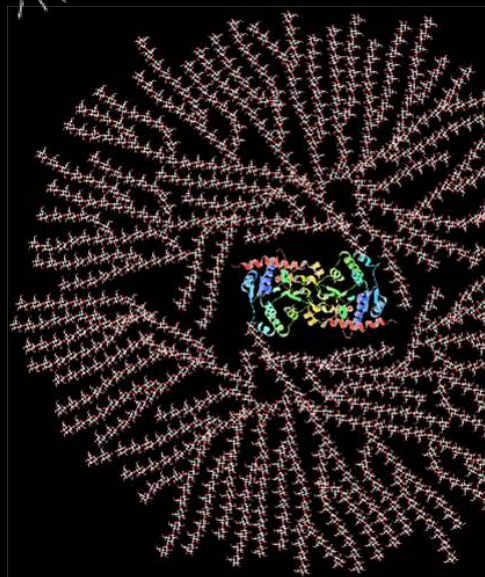
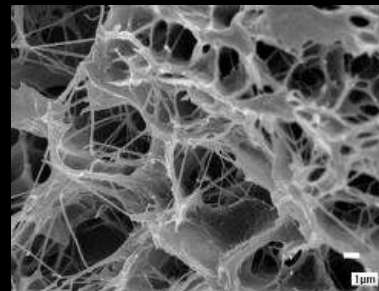
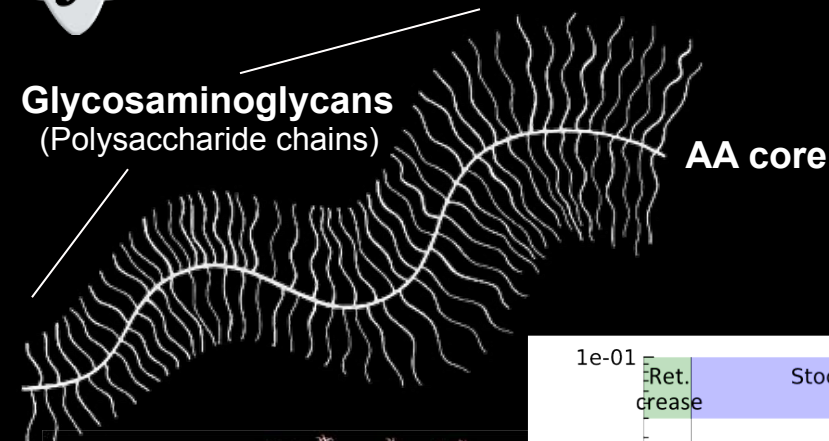
M00334: Type VI secretion system



- 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

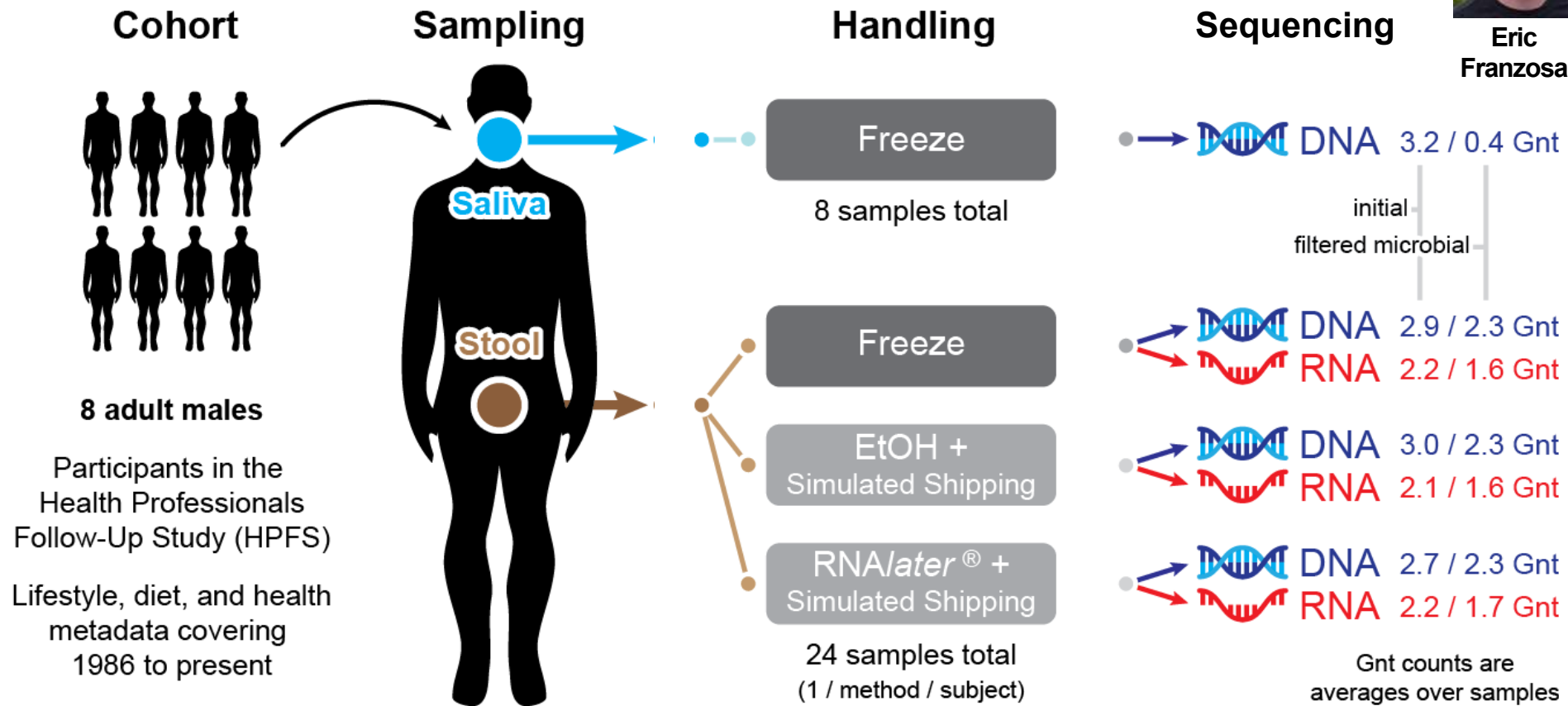


Functional meta'omics in human populations



Eric Franzosa

With Jacques Izard, Andy Chan, Wendy Garrett



- 1) Evaluate stability of meta'omic samples under subject-shipped conditions
- 2) Investigate links between the mouth and gut microbiomes
- 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

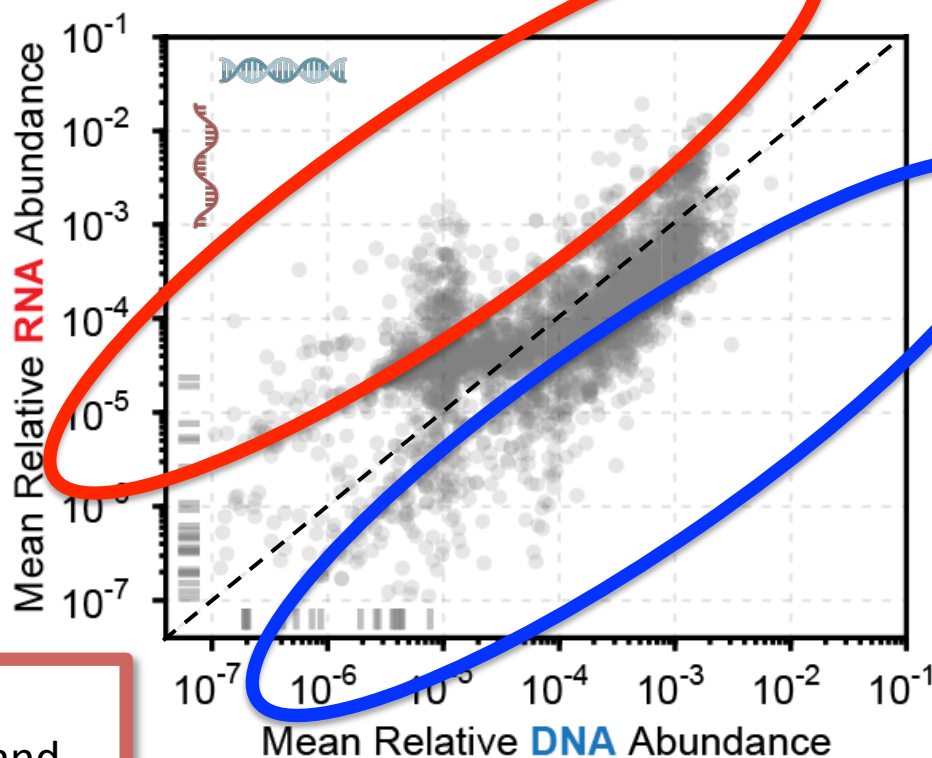
DNA $n=3,292$

RNA



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

3,292 gene families
Rank correlation = 0.76



Some functions are highly under-expressed

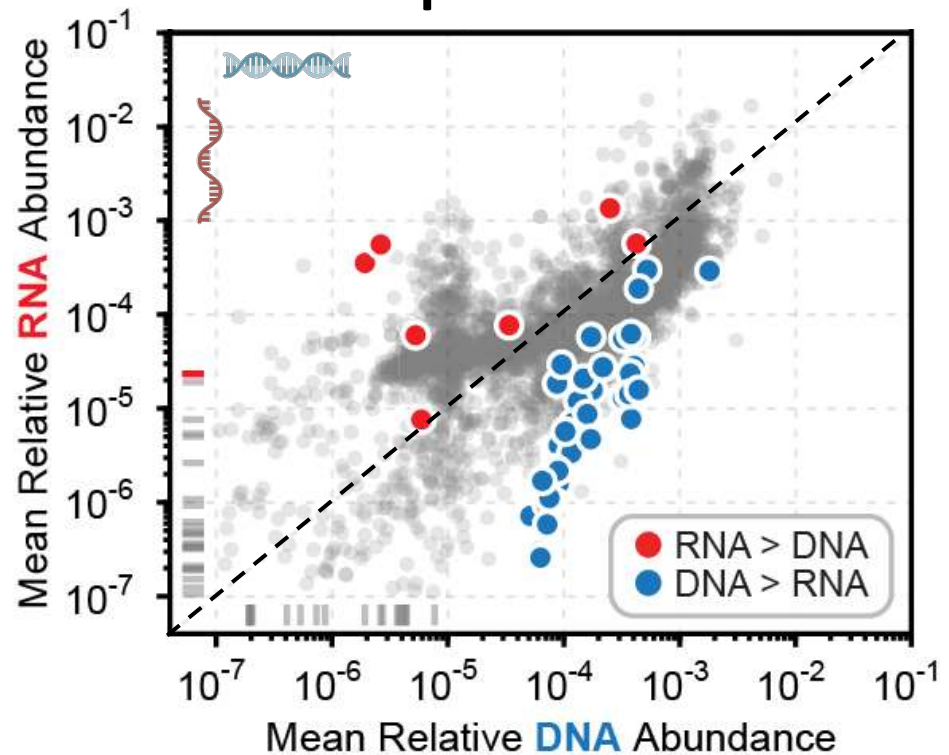
DNA

RNA



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

Sporulation



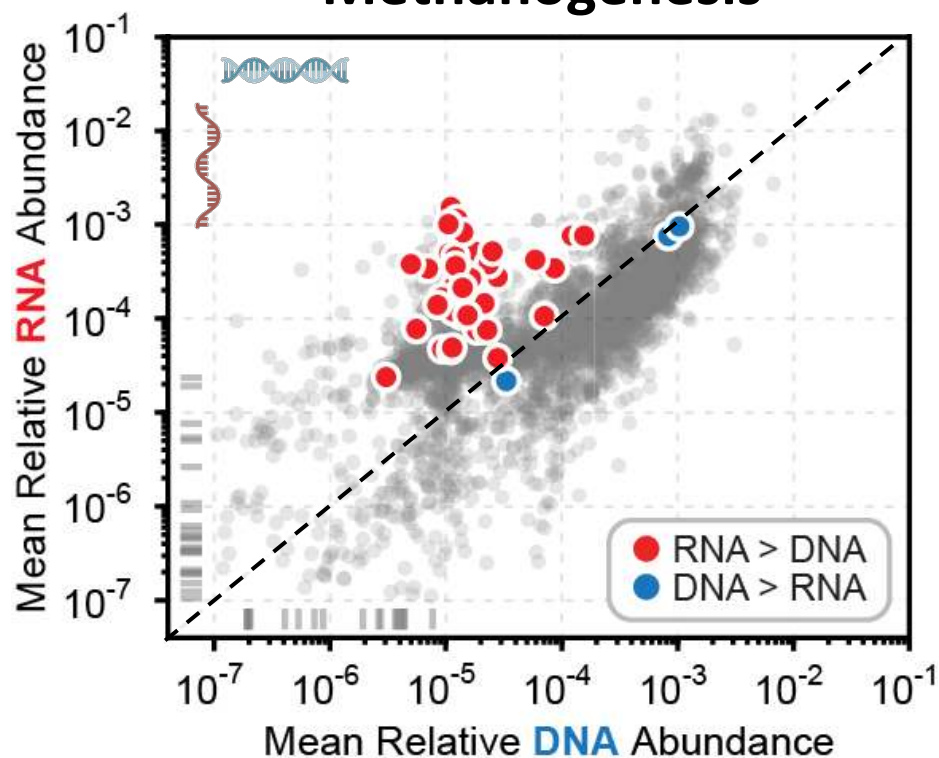
Other functions are highly over-expressed

DNA

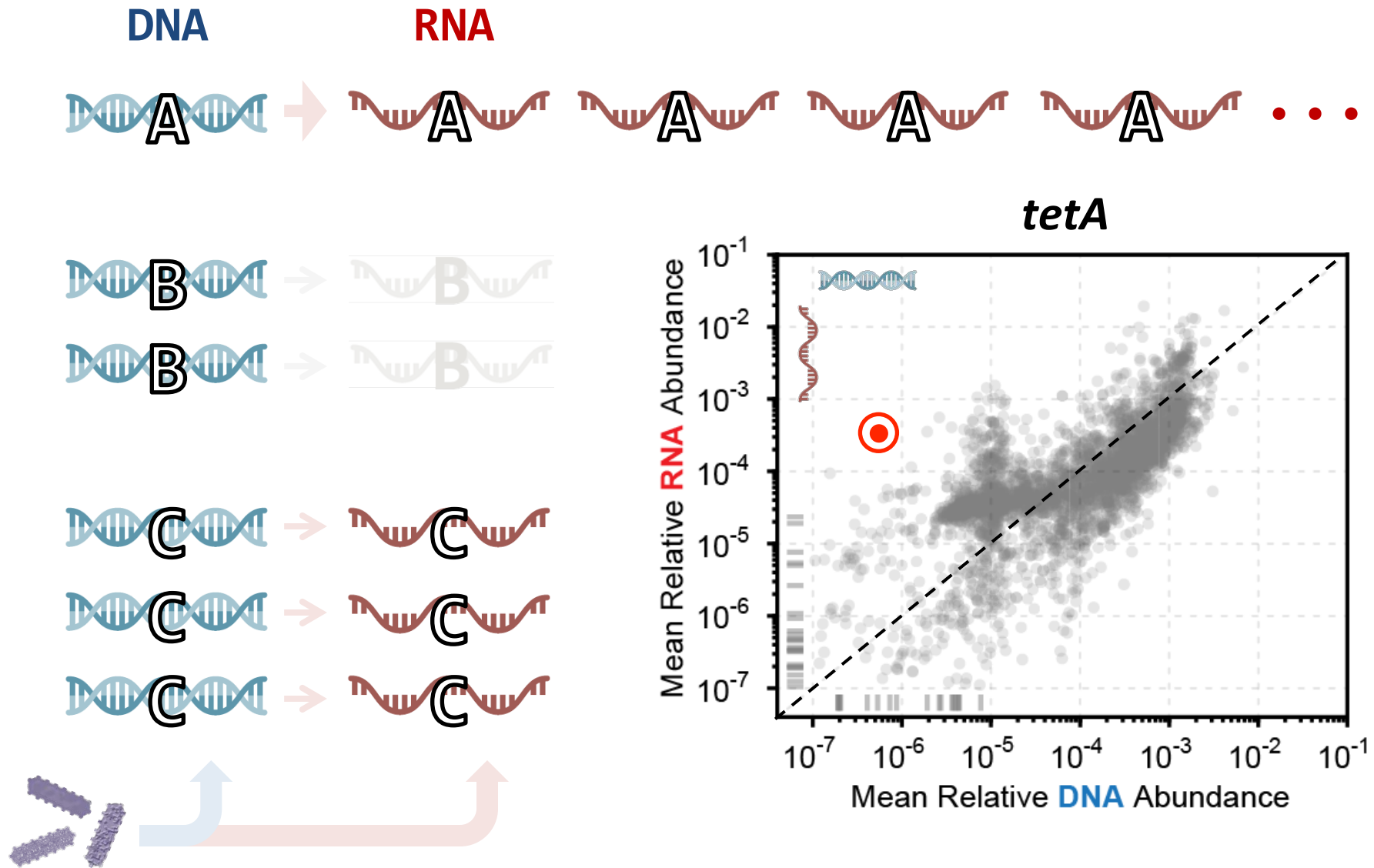
RNA



Methanogenesis

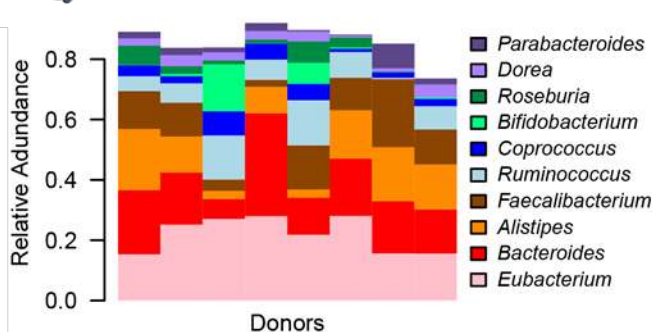


Other functions are highly over-expressed

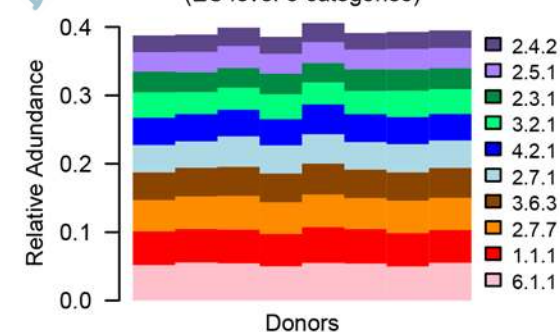


Functional potential is stable, activity is variable

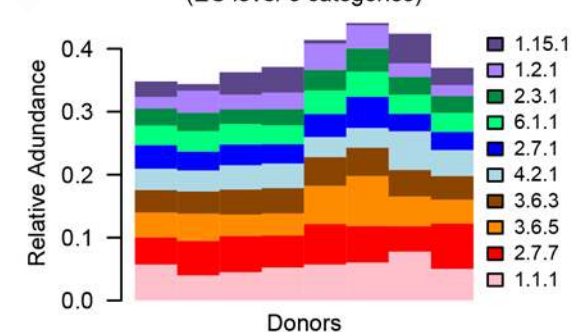
Top 10 Genera



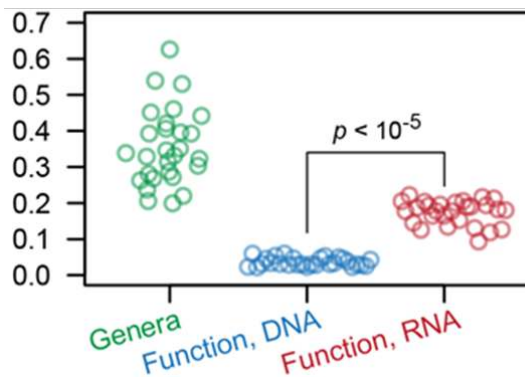
Top 10 Gene Families, DNA
(EC level-3 categories)



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



Bray-Curtis Dissimilarity



- 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>

Pathways and modules

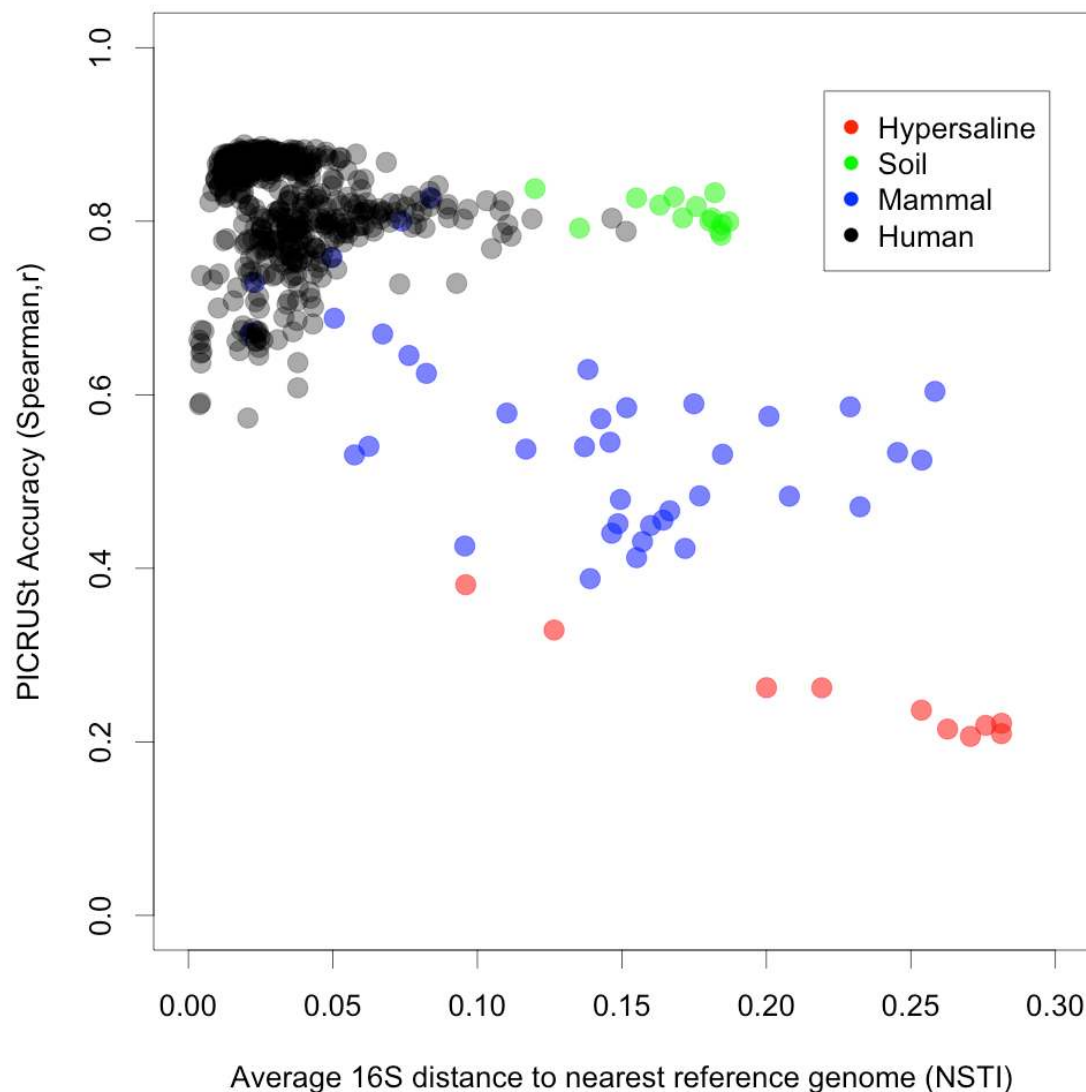


HUMAnN

Orthologous gene families



Taxon abundances



Relative abundance



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

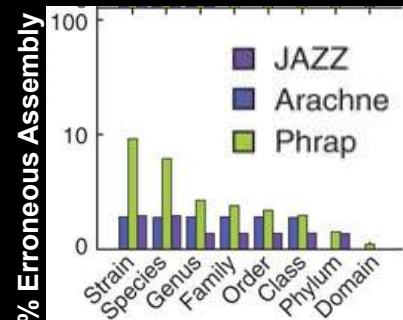
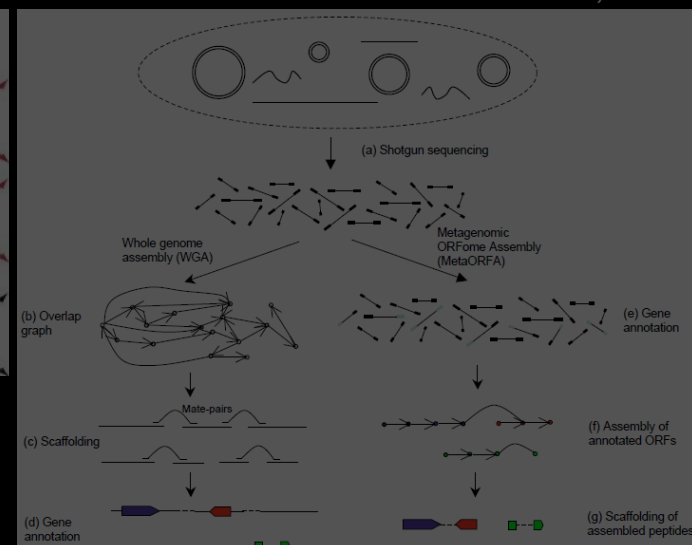
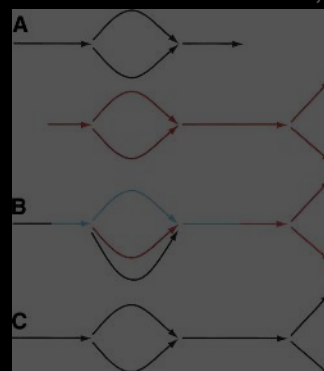
(Kalyuzhnaya 2008)

255Mbp, PGA

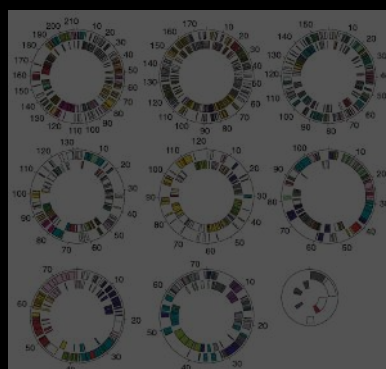
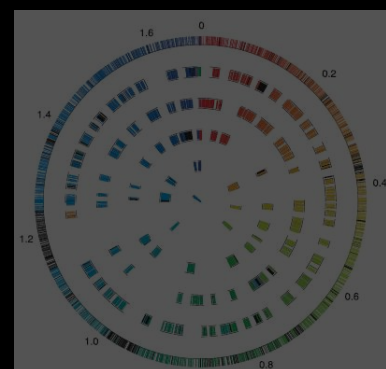
~1.5x coverage: 25k contigs, 70Mbp

ALLPATHS: Butler, 2008

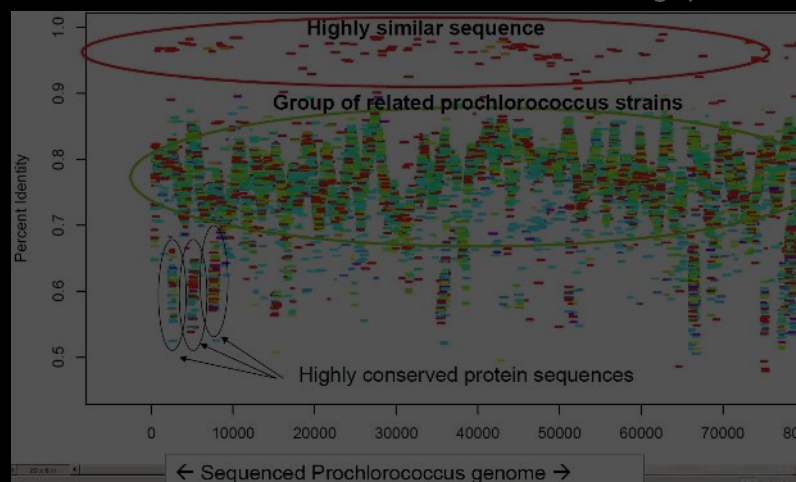
MetaORFA: Ye, 2008



Mavromatis, 2007



9 shared "megaplasmids"





Microbiome meta'omic analyses: assembly

Scaling metagenome sequence assembly with probabilistic de Bruijn graphs

Jason Pell¹, Arend Hintze², Rosangela Canino-Koning³, Adina Howe³, James M. Tiedje^{4,5}, and C. Titus Brown^{6,7,8}

khmer (Pell 2012)

P25

MetaAMOS: a metagenomic assembly and analysis pipeline for AMOS
Todd J Treangen^{1,2}, Sergey Koren^{1,3}, Irina Astrovskaya¹, Dan Sommer¹, Bo Liu^{1,3}
and Mihai Pop^{1,3}

¹Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA; ²The McKusick-Nathans Institute for Genetic Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; ³Department 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 Hachiya¹, Hideaki Tanaka¹ and Yasubumi Sakakibara^{1,*}

MetaVelvet (Namiki 2012)

Genovo: *De Novo* Assembly for Metagenomes

*JONATHAN LASERSON, *VLADIMIR JOJIC, and DAPHNE KOLLER

Genovo (Laserson 2011)

BIOINFORMATICS

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

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

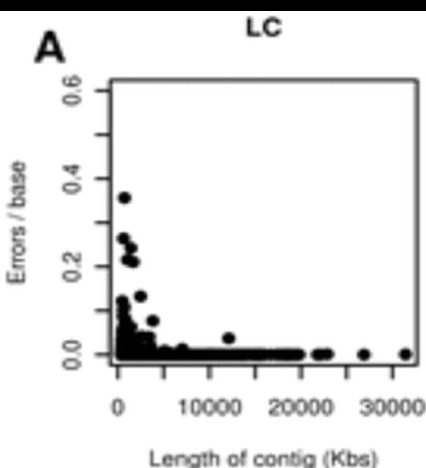
Yu Peng, Henry C. M. Leung, S. M. Yiu and Francis Y. L. Chin^{*}

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

Meta-IDBA (Peng 2011)

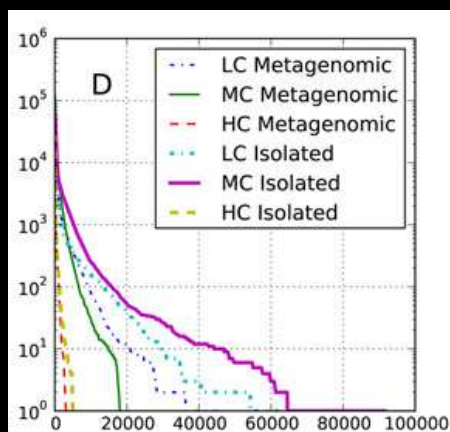
Evaluating the Fidelity of *De Novo* Short Read Metagenomic Assembly Using Simulated Data

Miguel Pignatelli^{1,2,3,4}, Andrés Moya^{1,2}



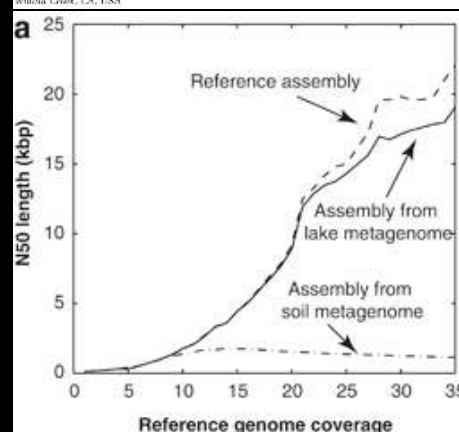
Evaluation of short read metagenomic assembly

Aravind Chinnai, Husein Kargacali



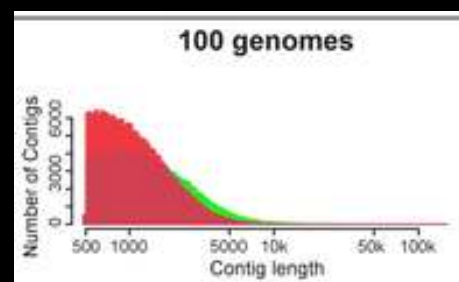
Individual genome assembly from complex community short-read metagenomic datasets

Chengwei Luo¹, Deepina Tementzi¹, Nikos C Kyrpides² and
Konstantinos T Konstantinidis^{1,2}
¹Center for Bioinformatics and Computational Genomics and School of Biology, Georgia Institute of Technology, Atlanta, GA, USA; ²School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, USA and ³Department of Energy (DOE) - Joint Genome Institute, Walnut Creek, CA, USA



Assessment of Metagenomic Assembly Using Simulated Next Generation Sequencing Data

Daniel R. Mende^{1,2}, Alison S. Waller^{1,3}, Shinichi Sunagawa¹, Aino I. Järvelin¹, Michelle M. Chan¹,
Manimozhayan Arumugam¹, Jeroen Raes¹, Peer Bork^{1,4,*}



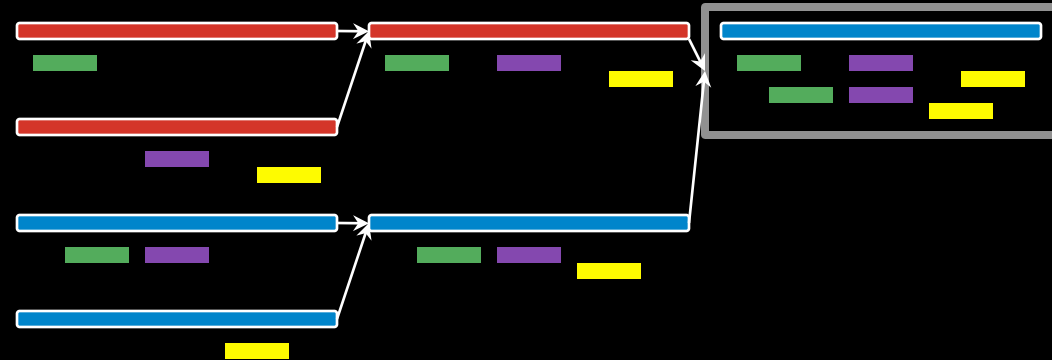


Microbiome meta'omic analyses: gene calling and proxygenes

Extrinsic gene calling:
BLAST etc. (proxygenes)

Intrinsic gene calling:
ORF detection from seq.

Dalevi, 2009



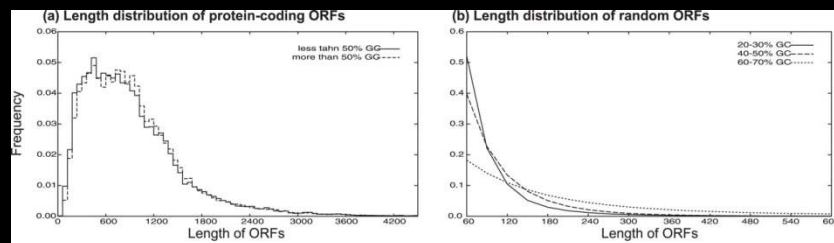
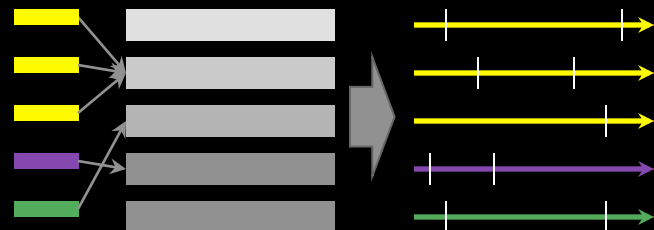
Krause, 2006
Yooseph, 2008

BLAST



Orphelia: Hoff, 2009
MetaGene: Noguchi, 2006

HMM models





Downstream analyses

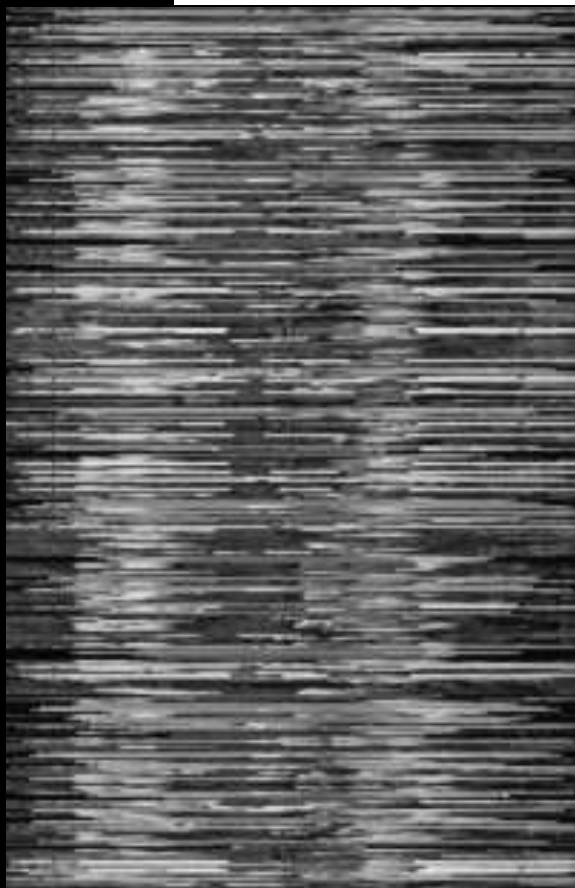
Mardis Genome Medicine 2010, 2:84
<http://genomemedicine.com/content/2/1/84>



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

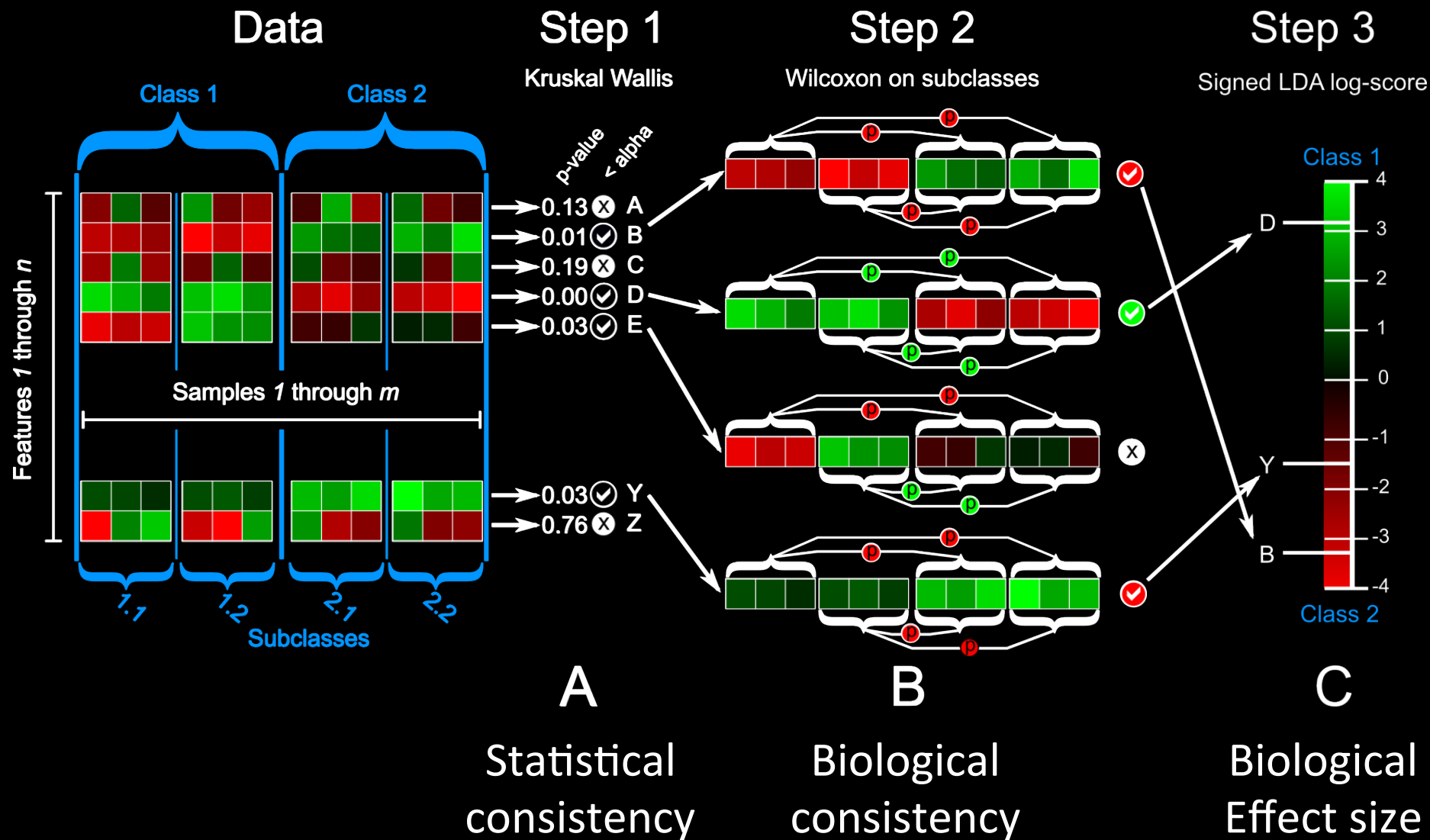
Sample #	1	2	3	4	5	6
Profession	Student	Postdoc	Postdoc	Professor	Student	Student
Gender	Male	Female	Female	Male	Male	Female
Site	Oral	Gut	Oral	Gut	Oral	Gut
Clade1	0.40	0.87	0.43	0.68	0.47	0.32
Clade1 Bug1	0.40	0.56	0.07	0.31	0.42	0.27
Clade1 Bug2	0.00	0.30	0.36	0.37	0.04	0.05
Clade2	0.60	0.13	0.57	0.32	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

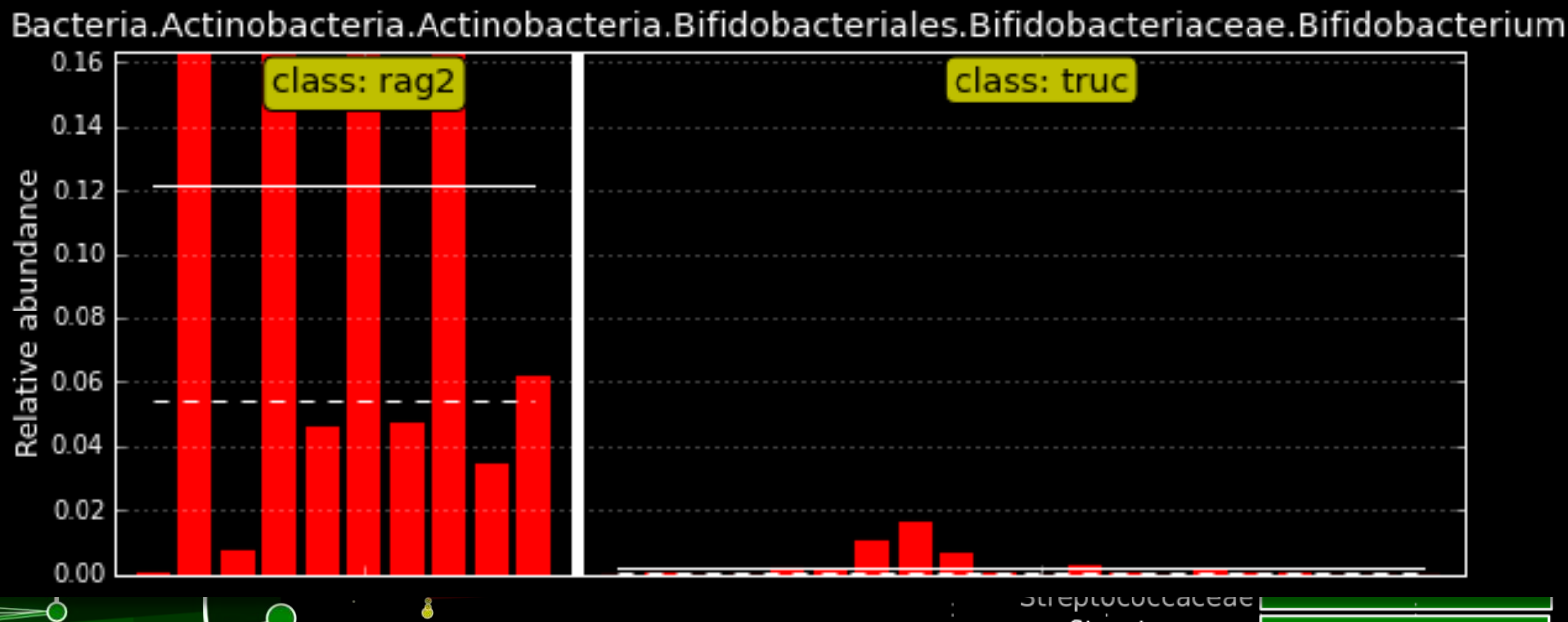
Nicola
Segata





Microbiome downstream analyses: statistical association testing (with LEfSe)

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}

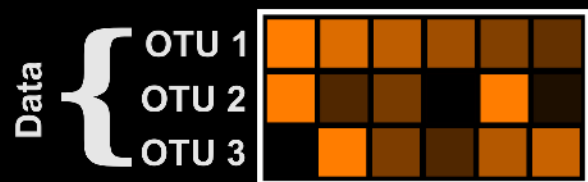
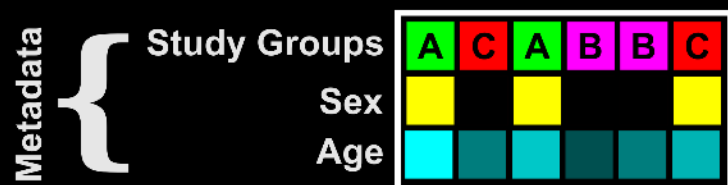
^aHarvard School of Public Health, Boston, MA 02115; ^bDanone Research, 91767 Palaiseau, France ^cBrigham and Women's Hospital, Boston, MA 02115; ^dDannon Company Inc., White Plains, NY 10603; ^eHarvard Medical School, Boston, MA 02115; ^fRagon Institute of MGH, MIT and Harvard, Charlestown, MA 02129; and ^gDana 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)



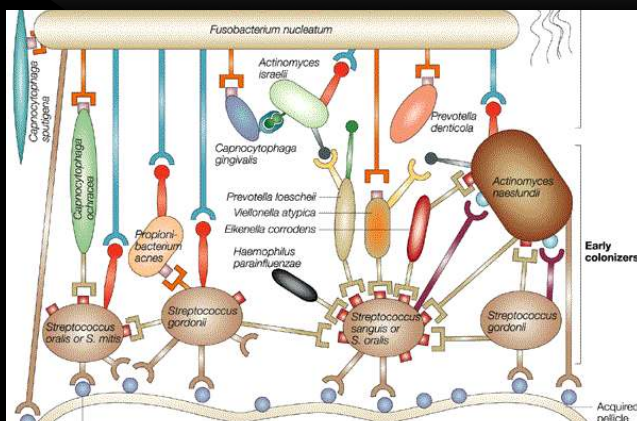
MaAsLin: Multivariate (microbial) Association with Linear Models

Overview of MaAsLin Association Methodology





Microbiome downstream analyses: interaction network reconstruction

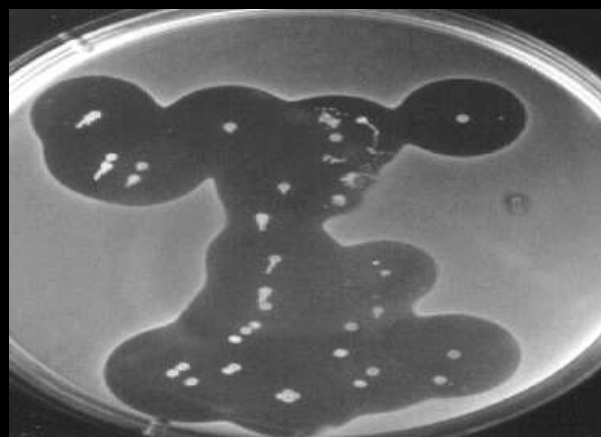


*It's a jungle in there –
microbial interactions follow
patterns from classical
macro-ecology.*

Mutualism



Predation

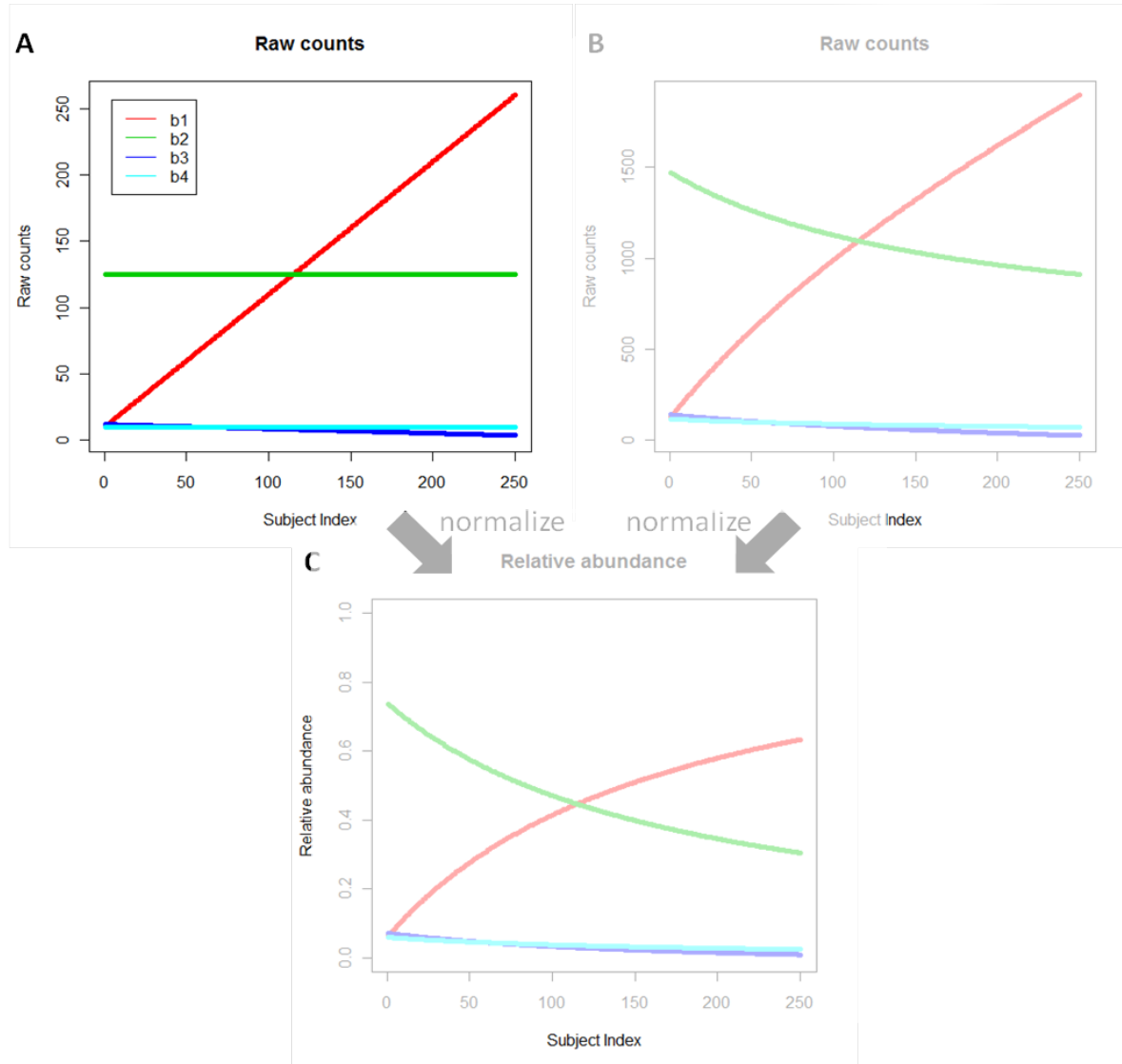


Competition



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



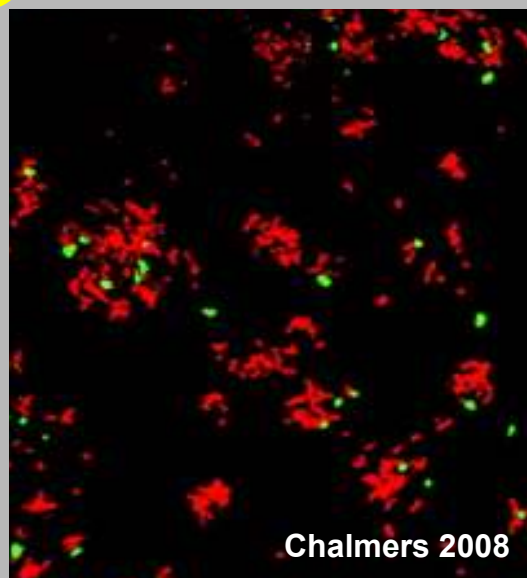
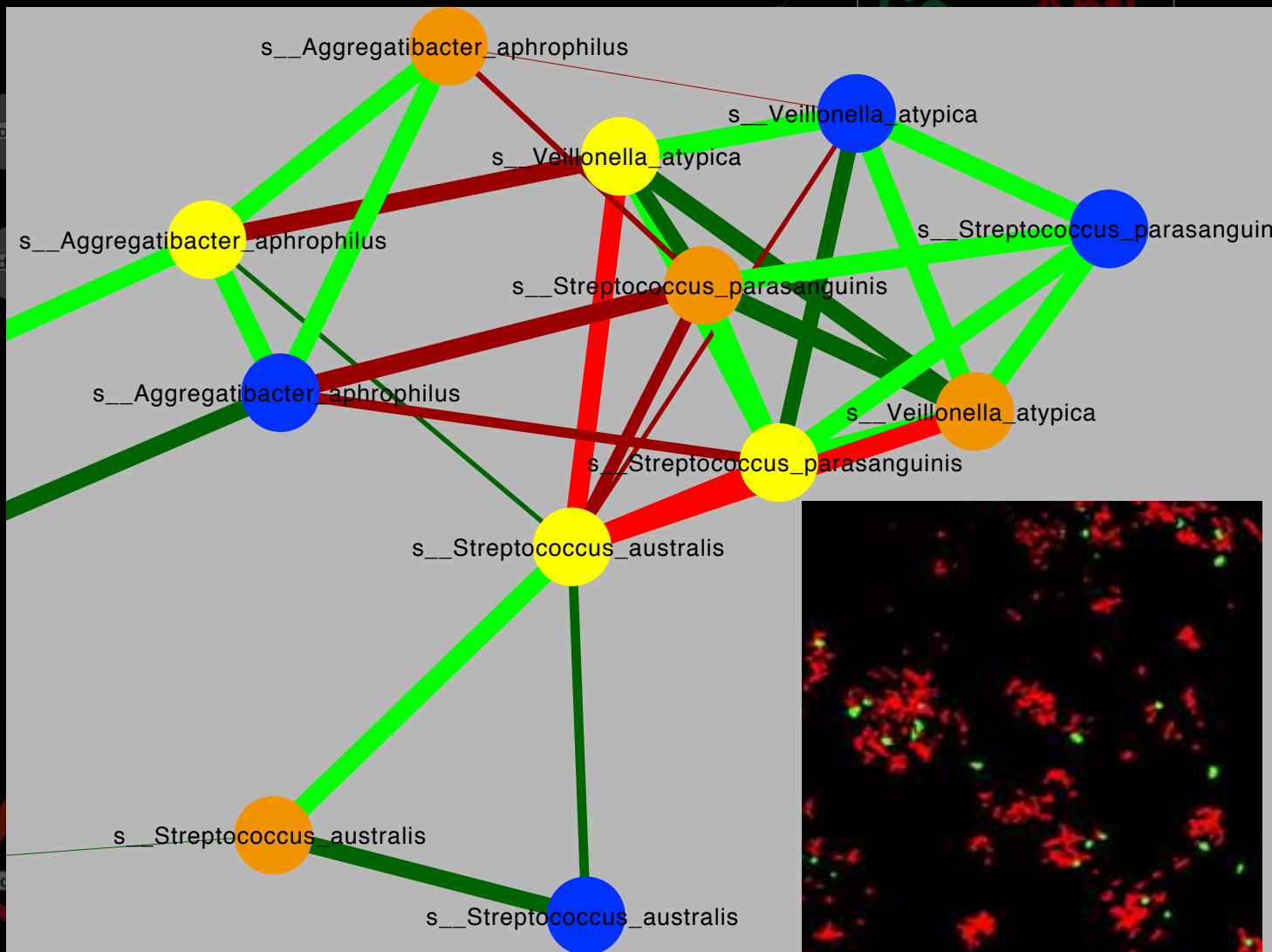
S	
C	Total
100	300
250	800

Relative Abundances



Friends, neighbors, and enemies: Microbial co-occurrence and exclusion in the human microbiome

With Jeroen Raes, Karoline Faust



Anterior nares
oral mucosa
palate
minized gingiva
ne tonsils
gingival plaque
gingival plaque
ue dorsum
retroauricular crease
retroauricular crease
antecubital fossa
antecubital fossa
agina
rior fornix
al introitus



Emma Schwager

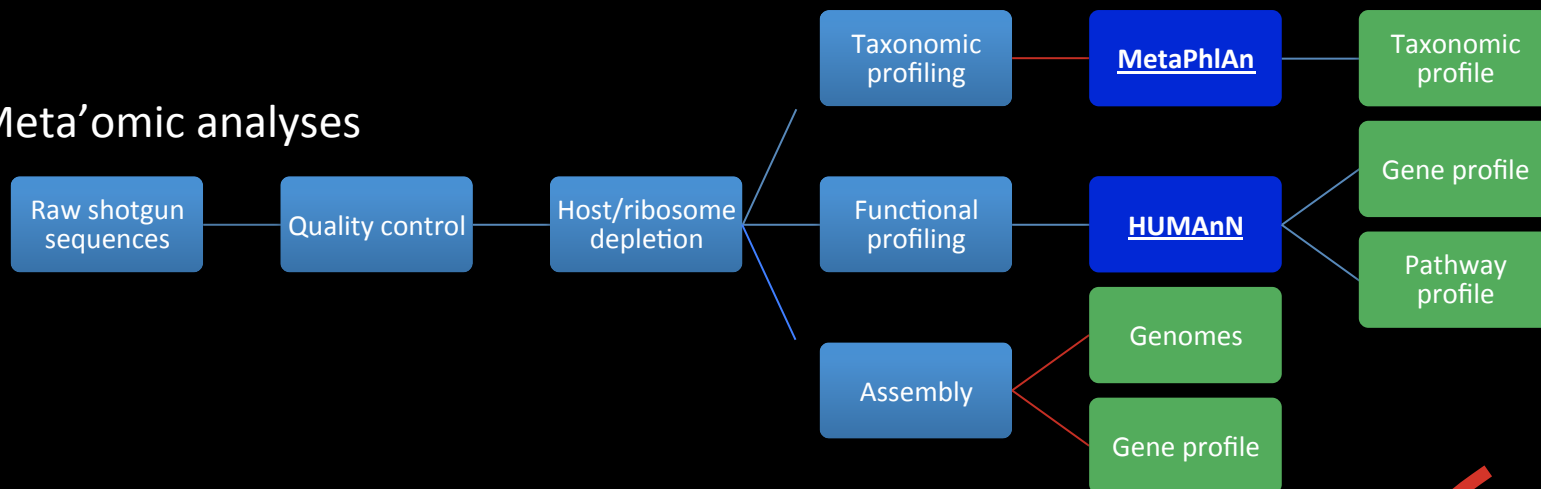


Fah
Sathirapongsasuti 53

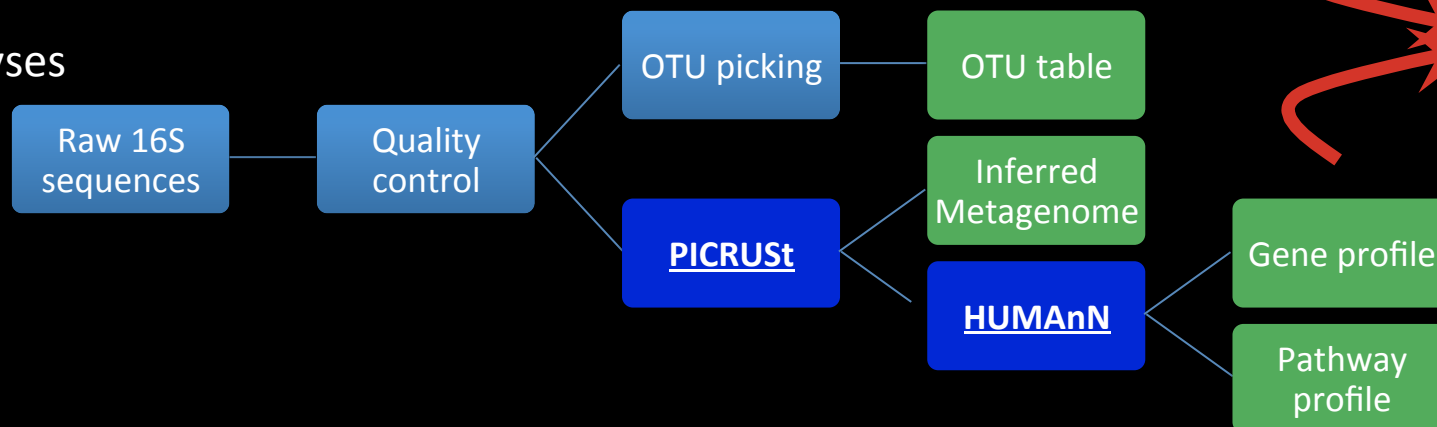


Typical microbial community analysis tasks

Meta'omic analyses



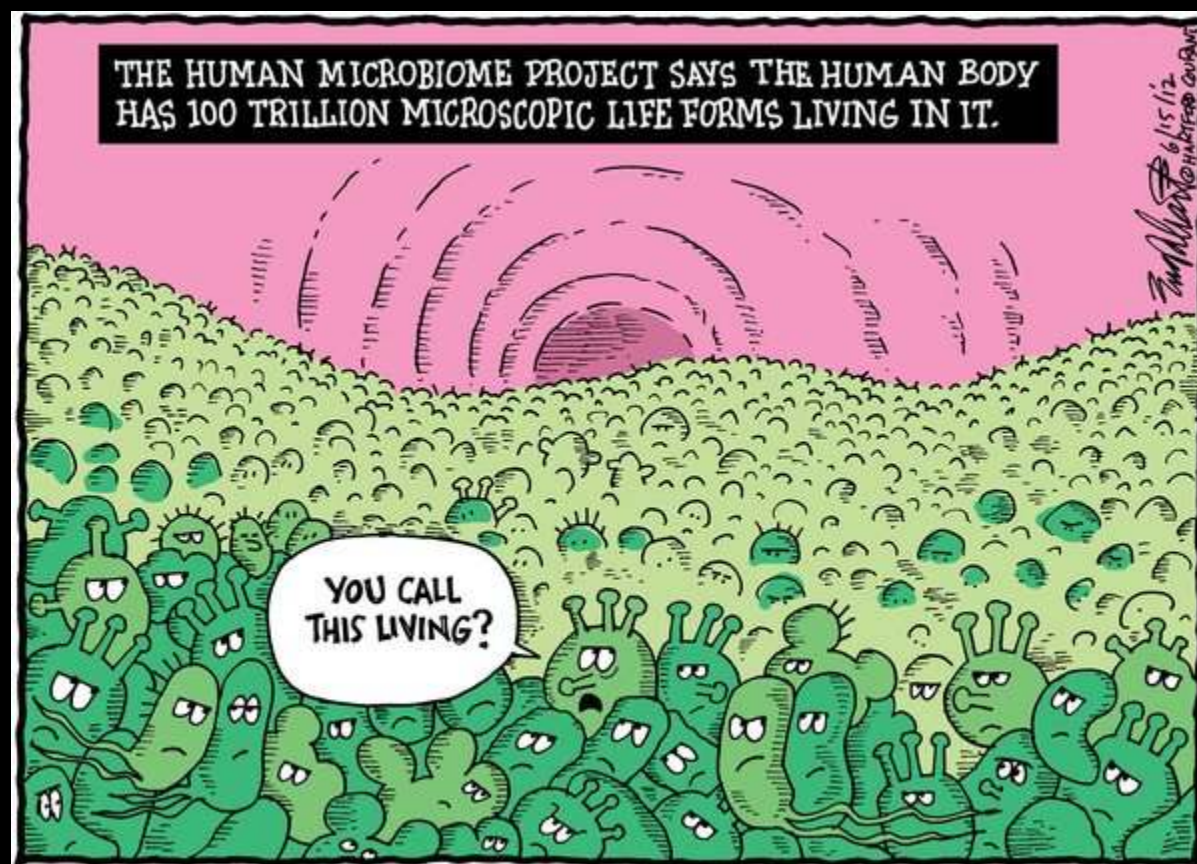
16S analyses



- Ordination
- Significance testing
- Networks
- Integration
- ...



Bringing it all together: The Human Microbiome Project



The NIH **H**uman **M**icrobiome **P**roject (**HMP**): A comprehensive microbial survey

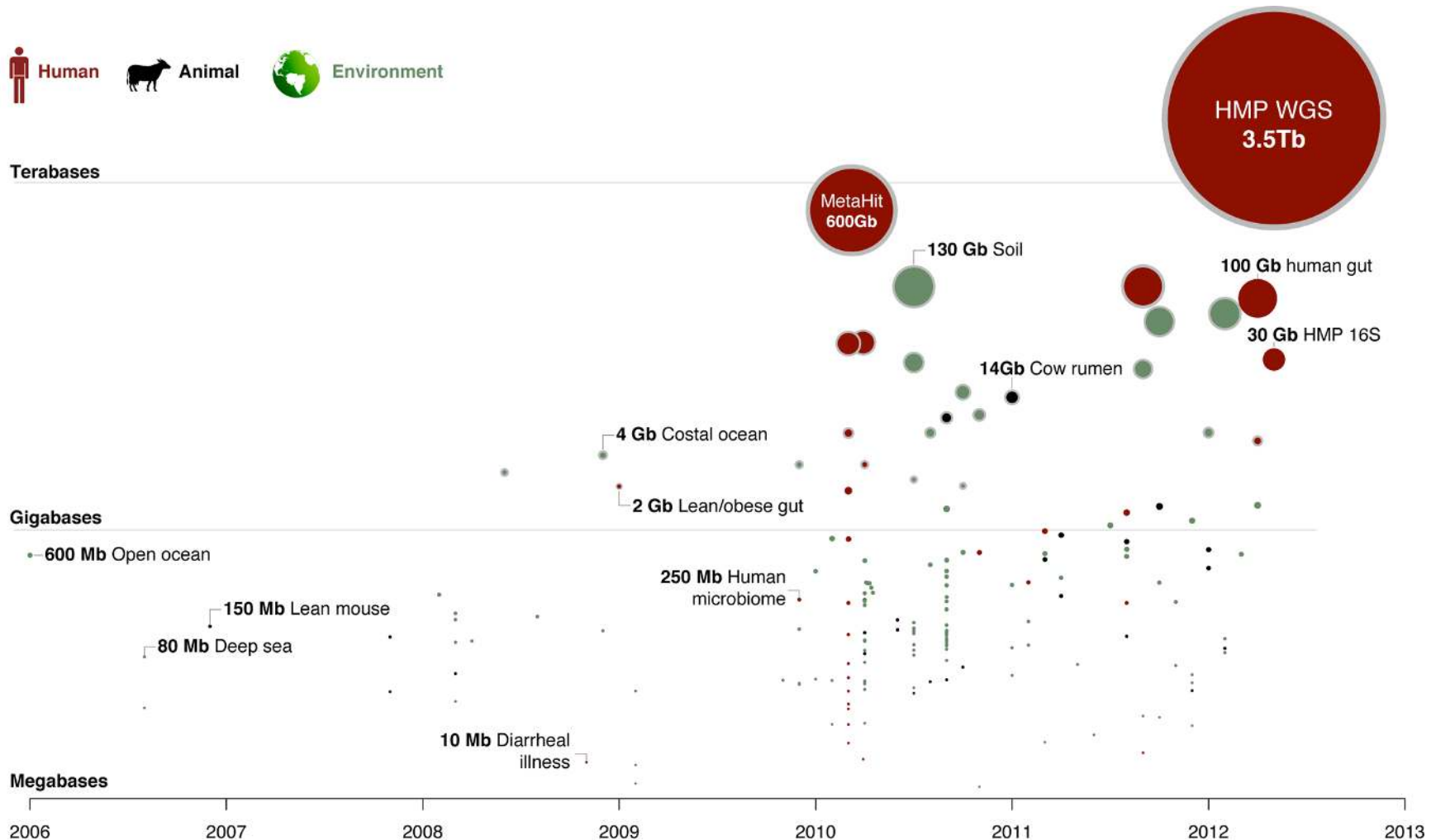
- ***What is a “normal” human microbiome?***
- 300 healthy human subjects
- Multiple body sites
 - 15 male, 18 female
- Multiple visits
- Clinical metadata

5,200 16S samples
Spanning 300 subjects, 18 sites
700 shotgun samples
Subset of 100 subjects, six sites



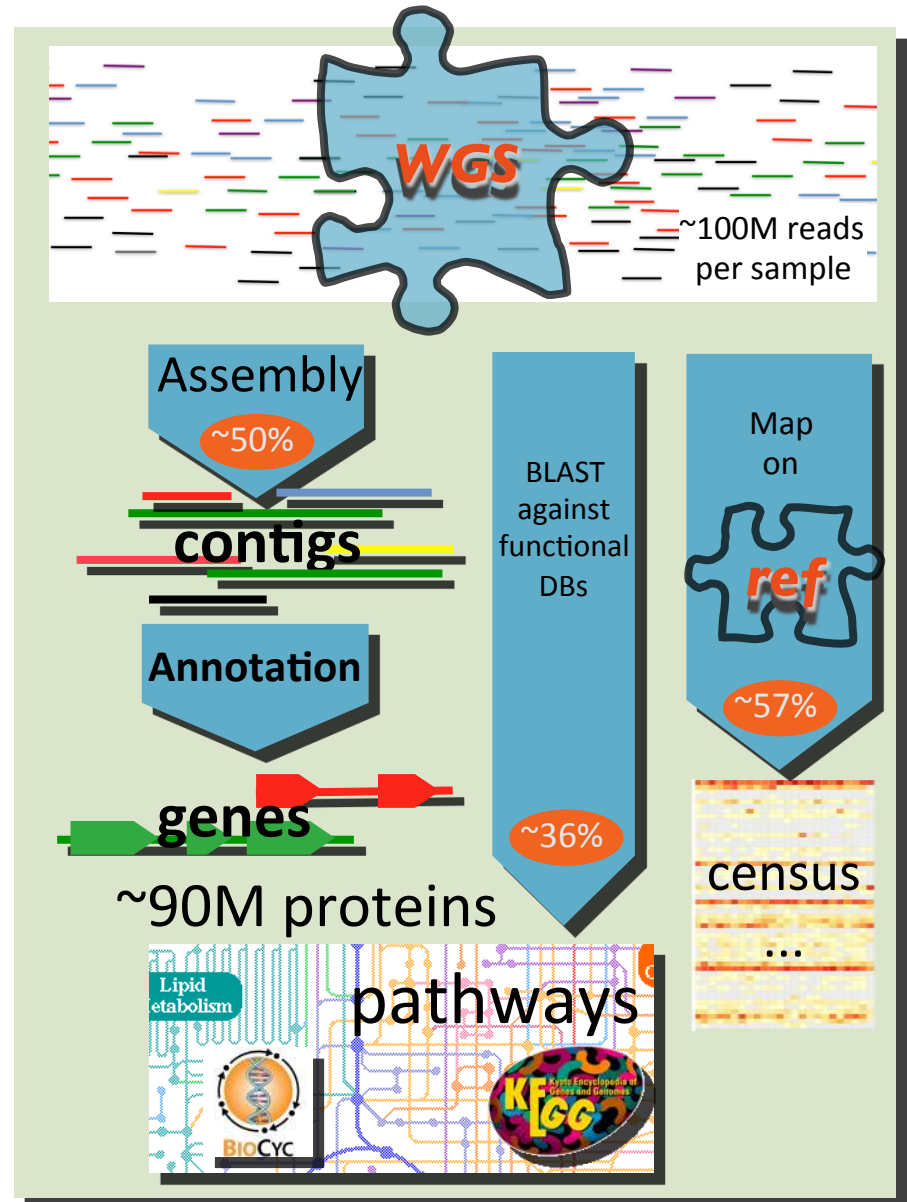
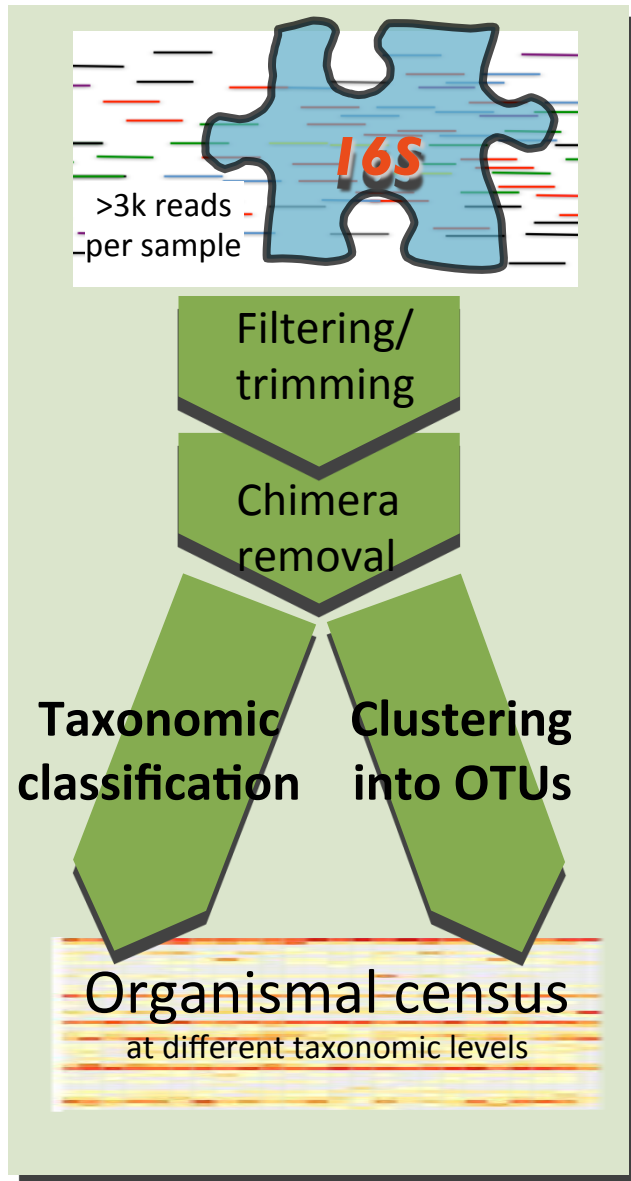
<http://www.nature.com/nature/focus/humanmicrobiota/>



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

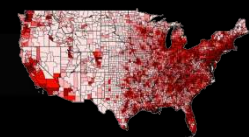
Gevers D, Knight R, Petrosino JF, Huang K, et al. (2012) The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome. PLoS Biol 10(8): e1001377. doi:10.1371/journal.pbio.1001377
<http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001377>

...for mining metagenomic data

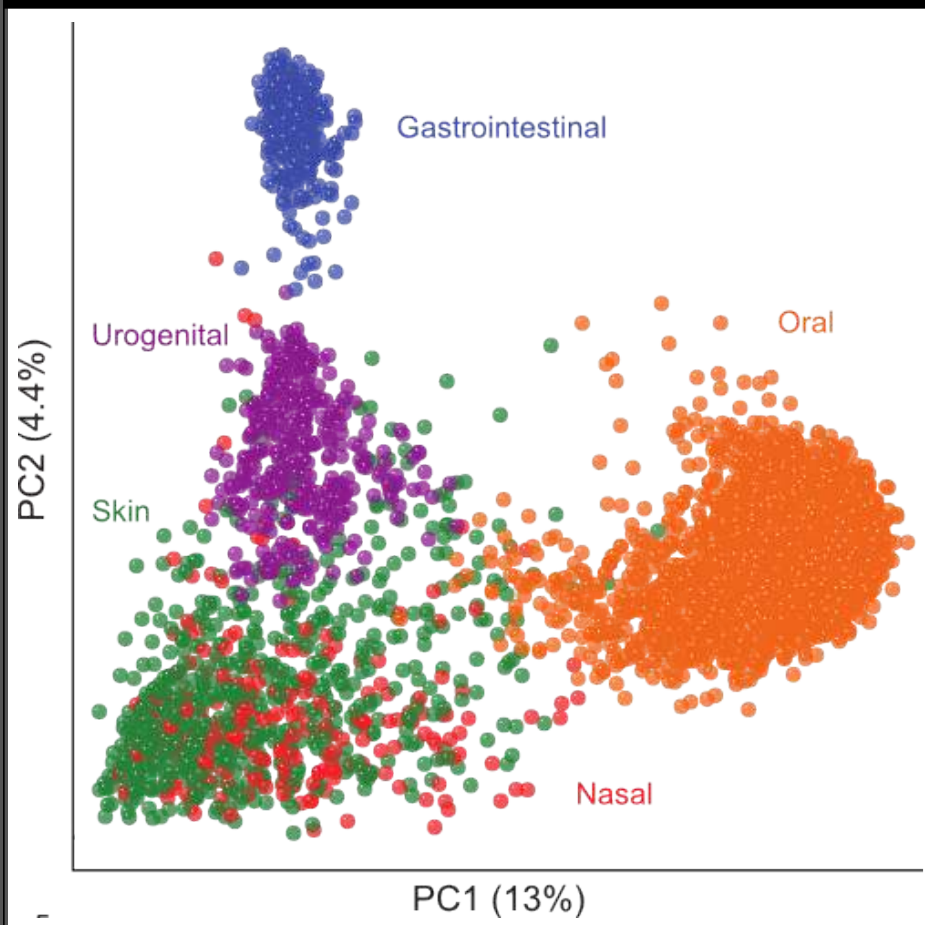
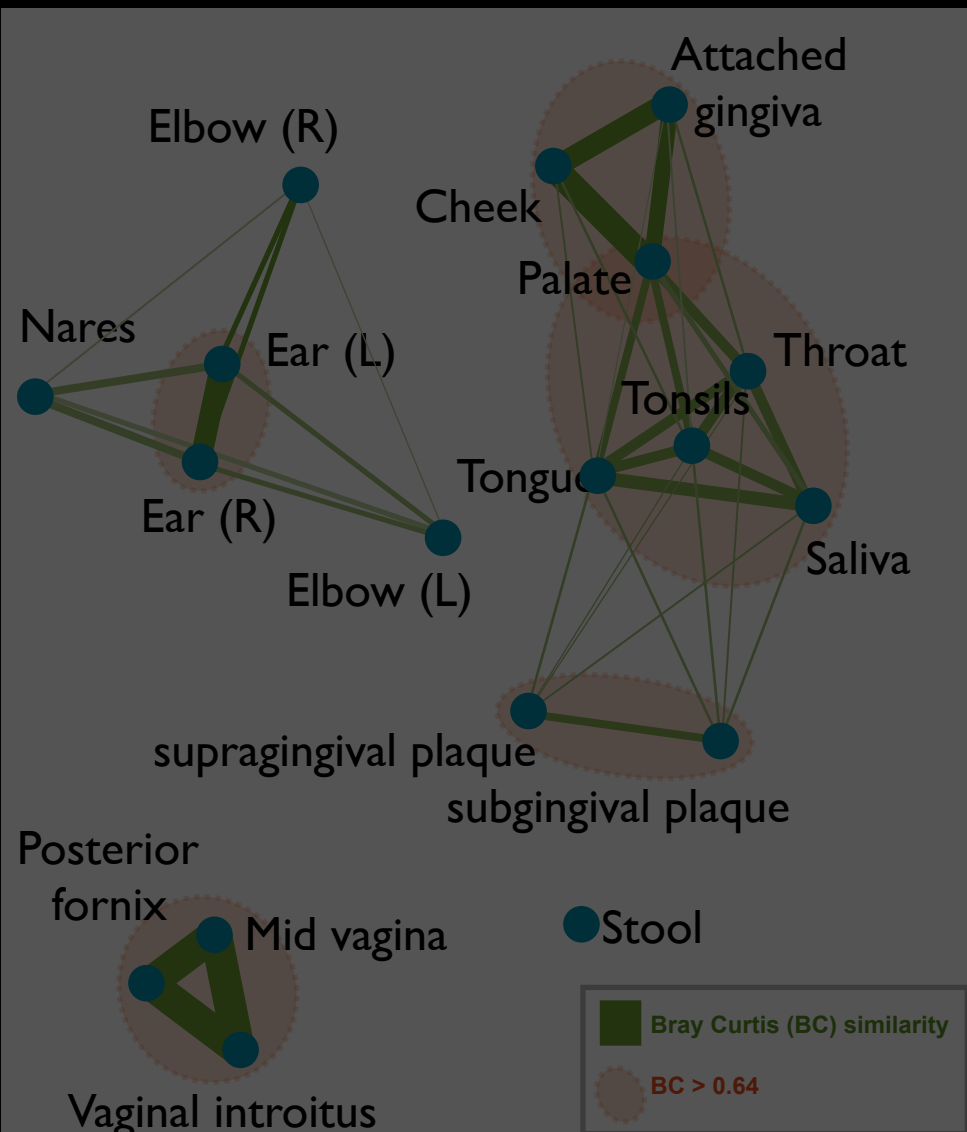




What aspects of a human host most influence microbial community composition?

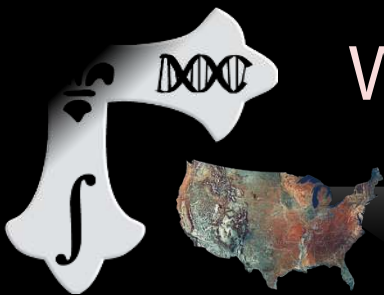


~5,200 microbial communities



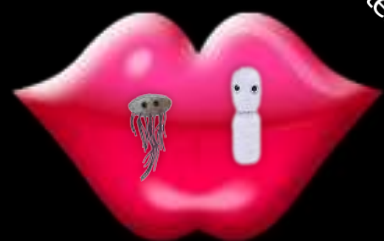
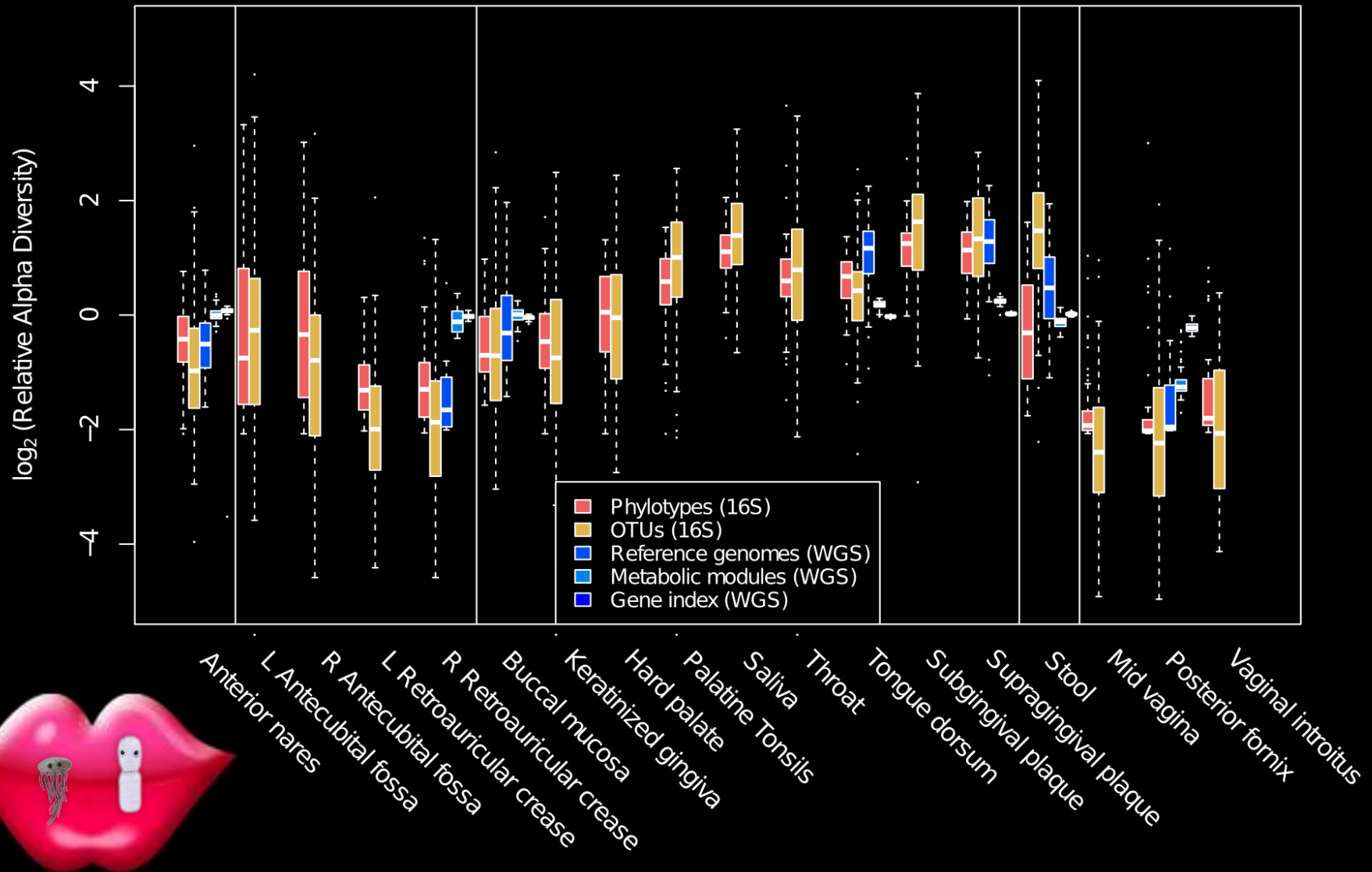
closer = more similar

Rob Knight



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

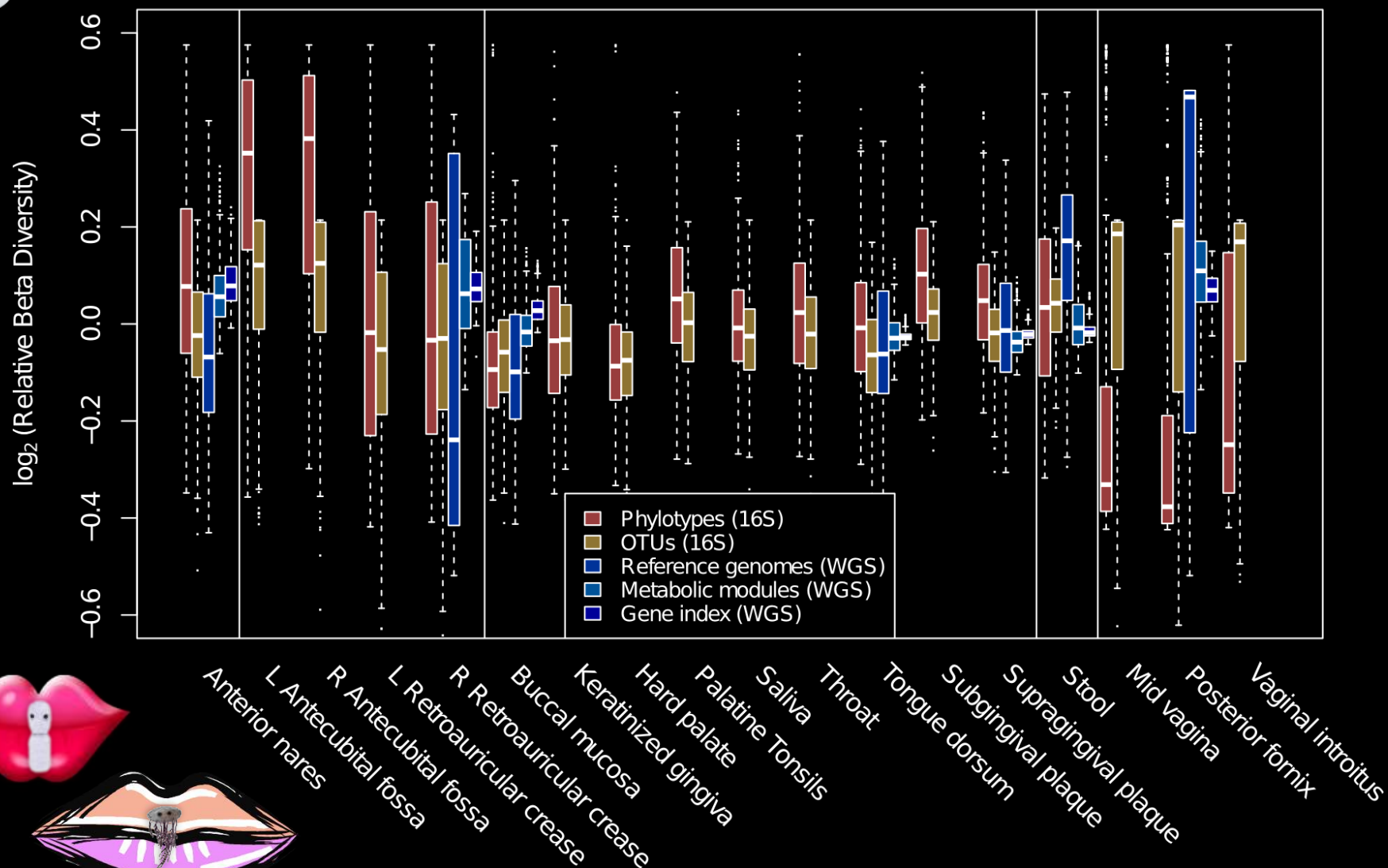
Within-Sample Alpha Diversity





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

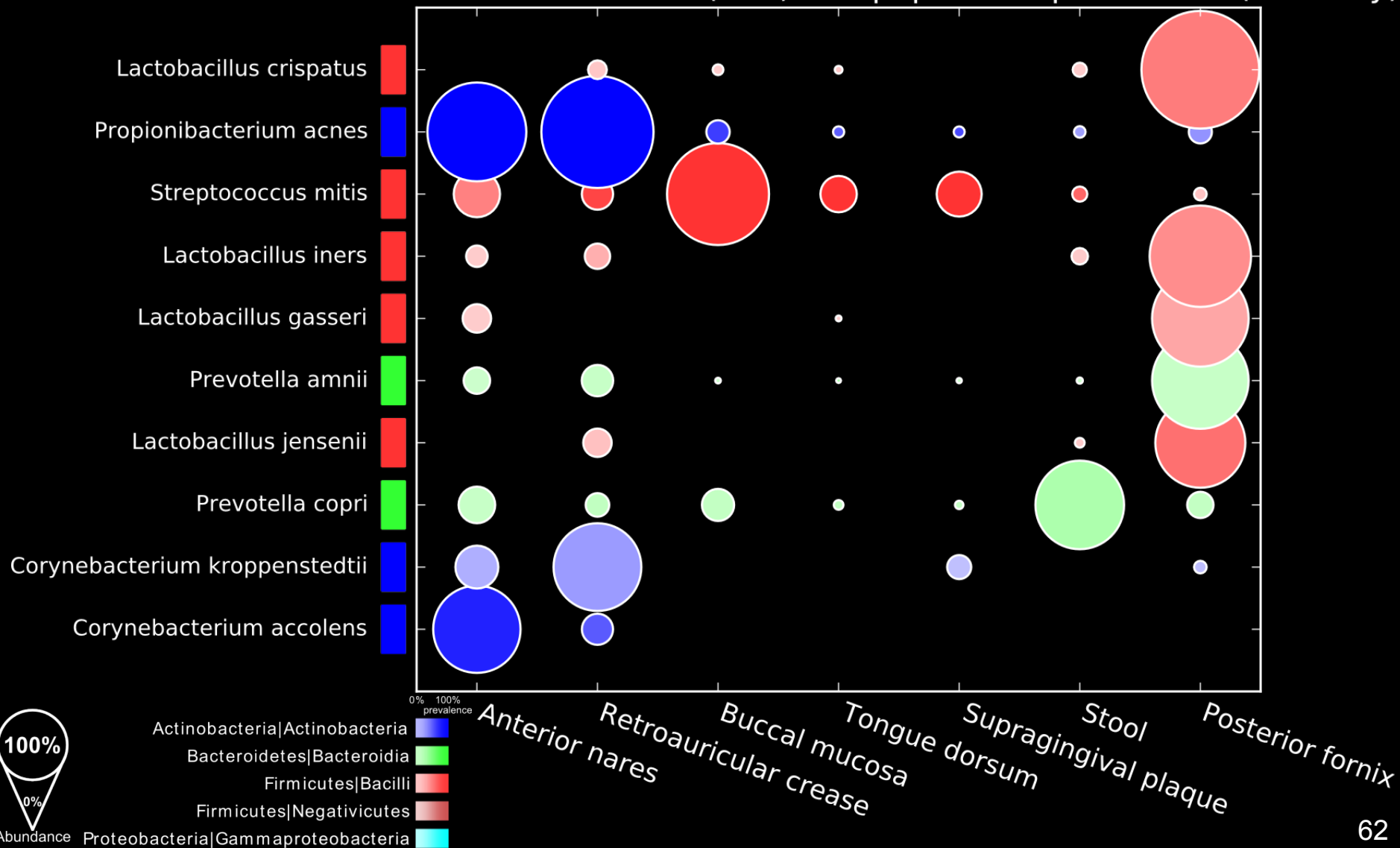
Between-Sample Beta Diversity





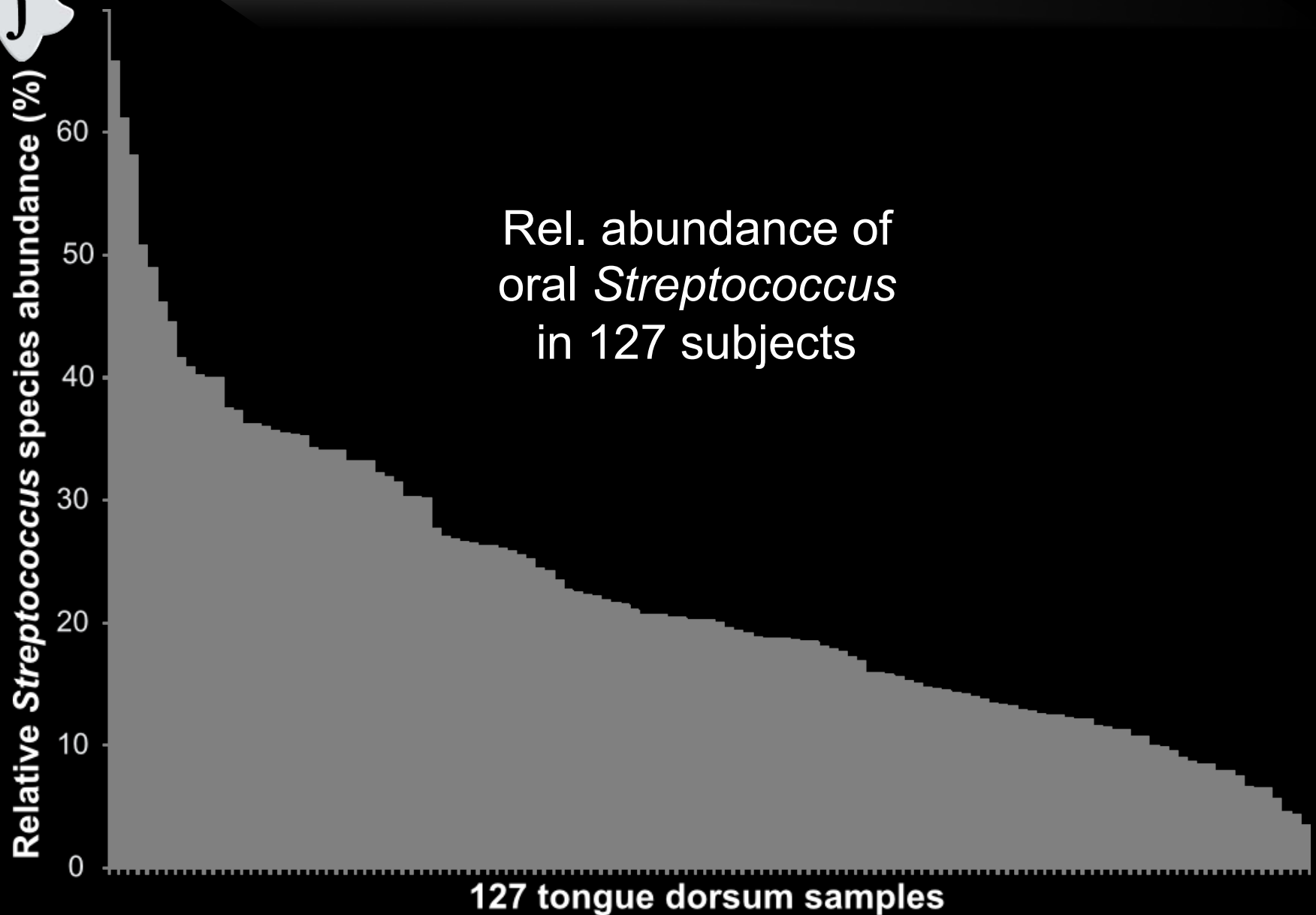
Microbiome meta'omic analyses: taxonomic profiling (with MetaPhlAn)

Mean nonzero abundance (size) and population prevalence (intensity)



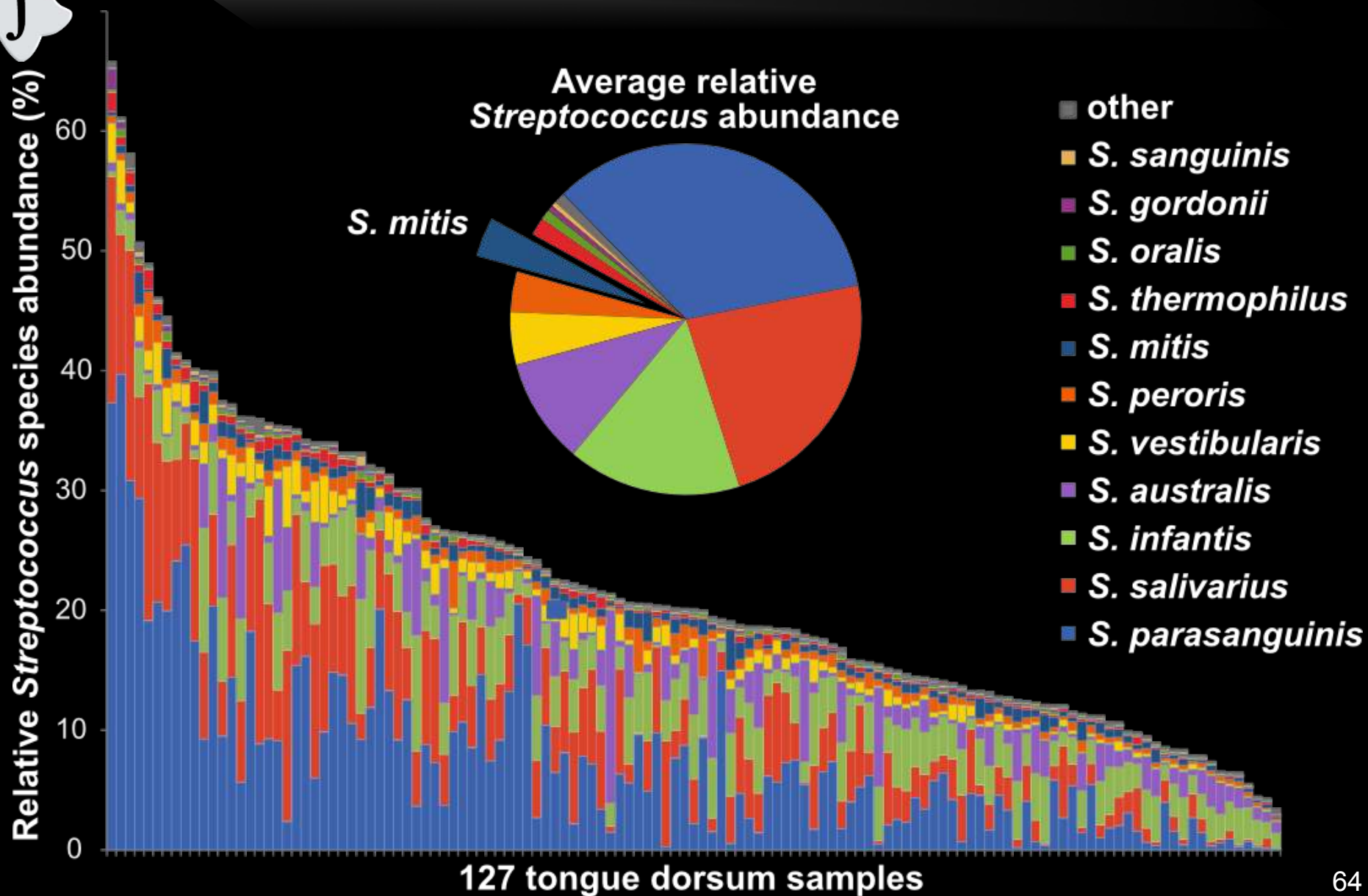


How unique is your personal microbiome?



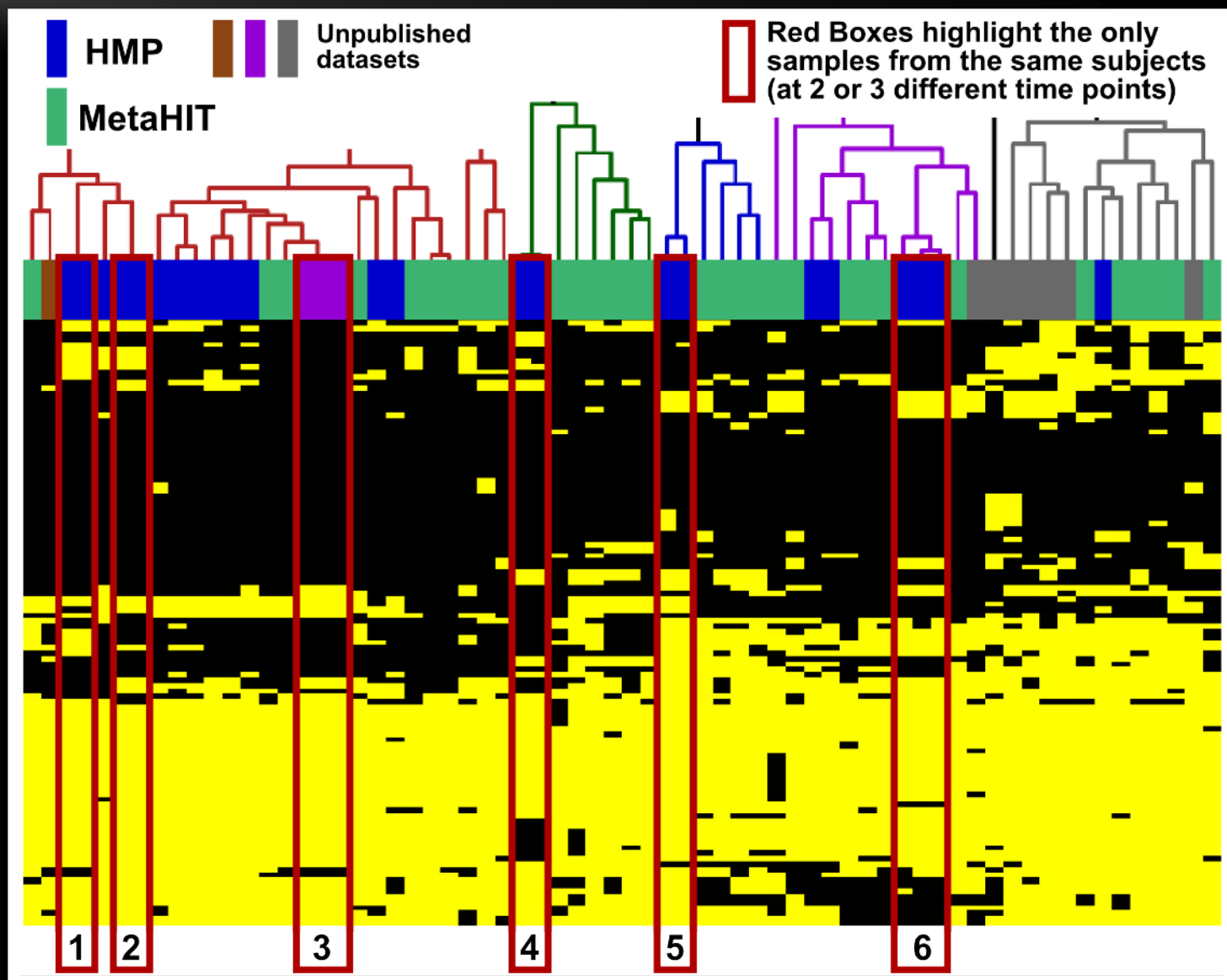


How unique is your personal microbiome?





How unique is your personal microbiome?



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. ¹⁴Division of Bioenvironmental Science, Frontier Science Research Center, University of Miyazaki, 5200 Kiyotake, Miyazaki 880-0815, Japan. ¹⁵Laboratory of Microbiology, Wageningen University, 6710BA Ede, The Netherlands. ¹⁶Tokyo Institute of Technology, Graduate School of Bioscience and Biotechnology, Department of Biological Information, 4259 Nagatsuta-cho, Midori-ku, Yokohama-shi, Kanagawa Pref. 226-8501, Japan. ¹⁷BGI-Shenzhen, Shenzhen 518083, China. ¹⁸Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK-2800 Lyngby, Denmark. ¹⁹Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain. ²⁰Department of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark. ²¹Institute of Biomedical Science, Faculty of Health Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark. ²²Hagedorn Research Institute, DK-2820 Gentofte, Denmark. ²³Faculty of Health Sciences, University of Aarhus, DK-8000 Aarhus, Denmark. ²⁴University 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.

174 | NATURE | VOL 473 | 12 MAY 2011

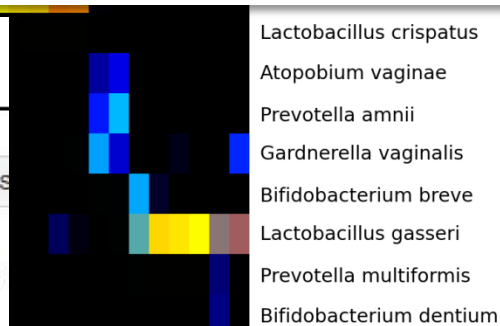
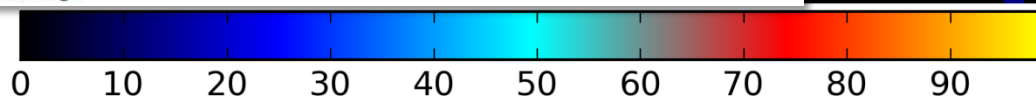
©2011 Macmillan Publishers Limited. All rights reserved

The New York Times

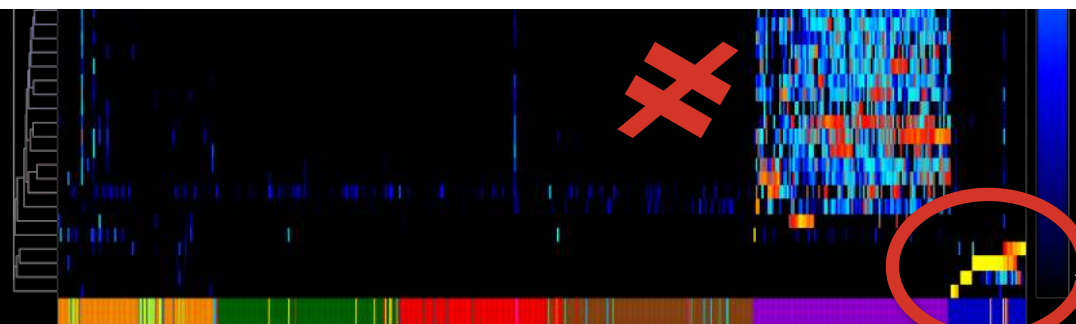
Science

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS ENVIRONMENT SPACE & COS

Bacterial Ecosystems Divide People Into 3 Groups, Scientists Say



The 64 most abundant



691 samples that passed quality control

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

Gut

Vaginal



Are there discrete “types” of typical human microbiomes?

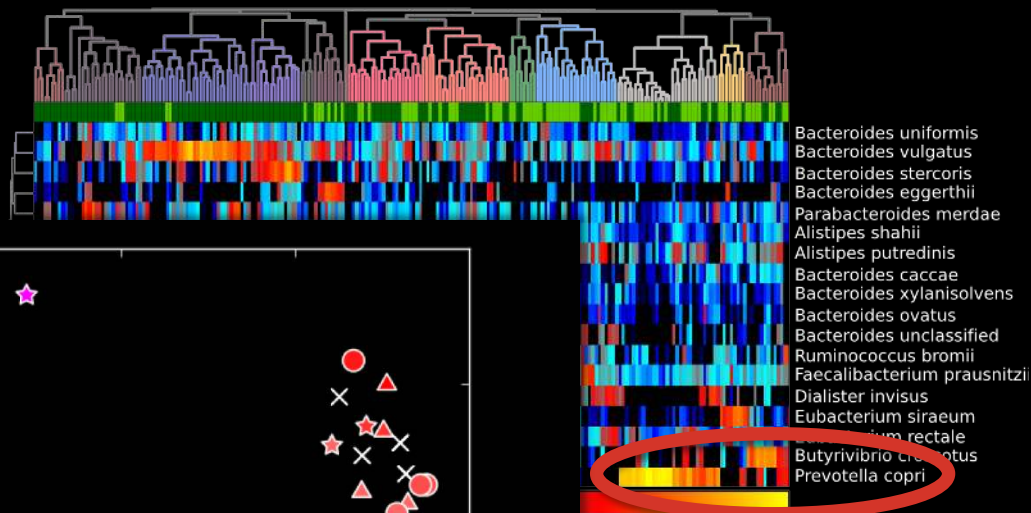
With Ruth Ley, Rob Knight

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



Levi Waldron

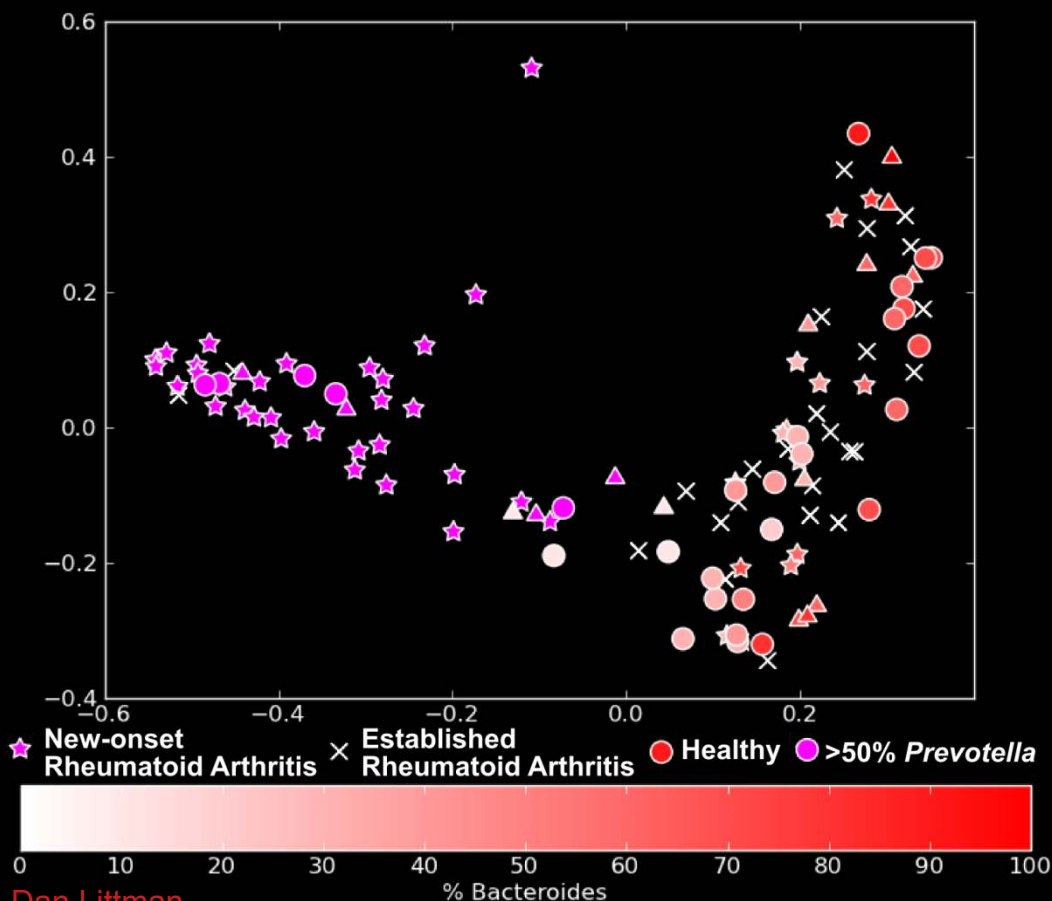
HMP samples MetaHIT samples



Omry Koren

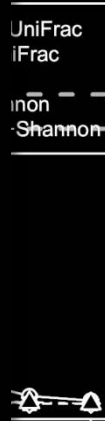


Dan Knights

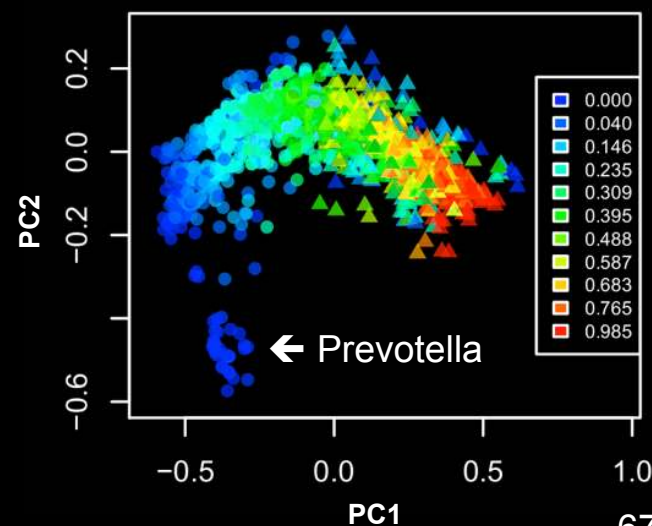


Dan Littman

Clustering

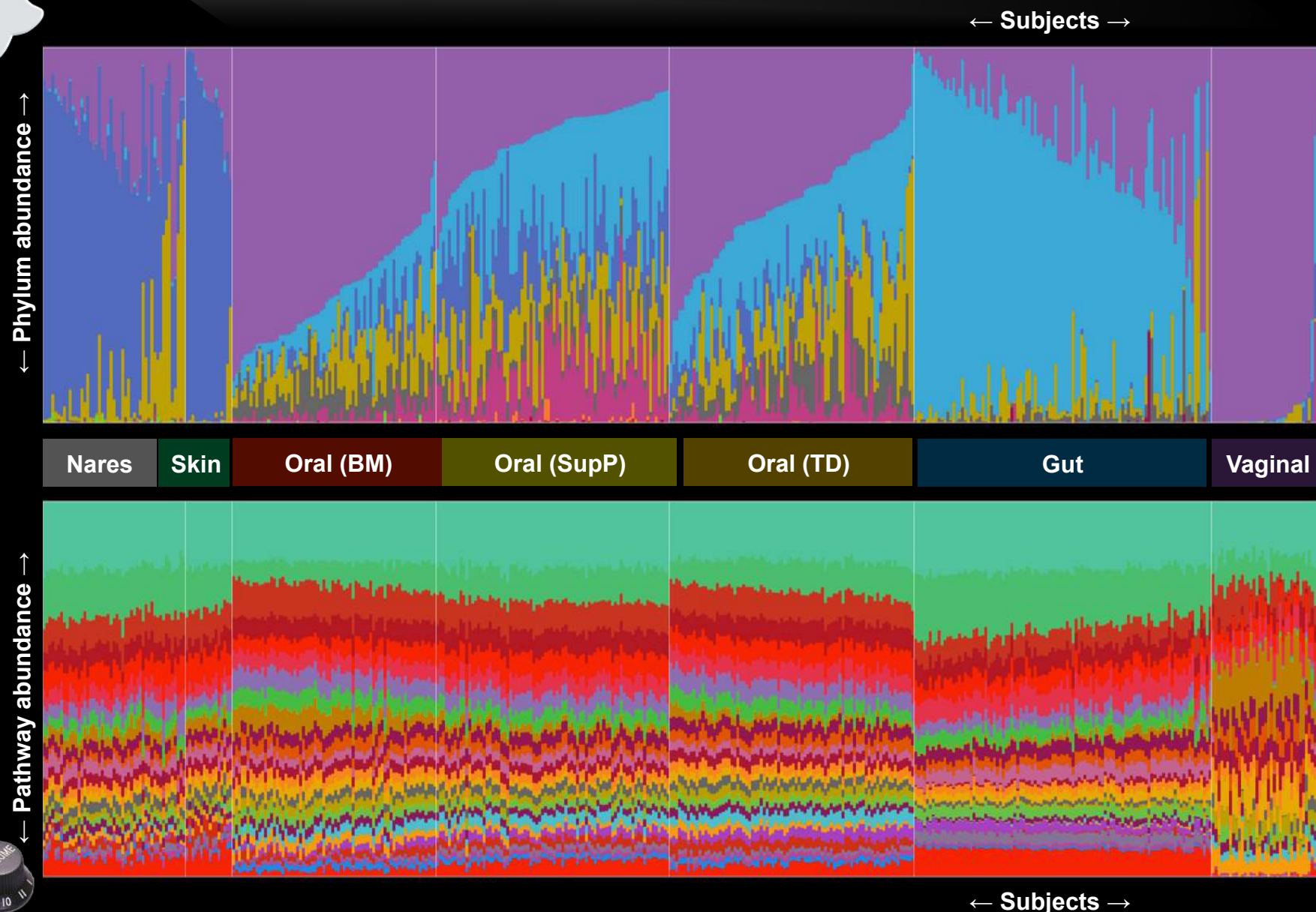


Bacteroides %

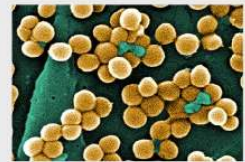
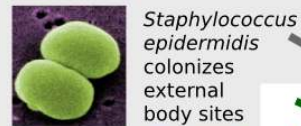
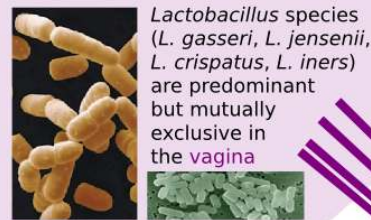
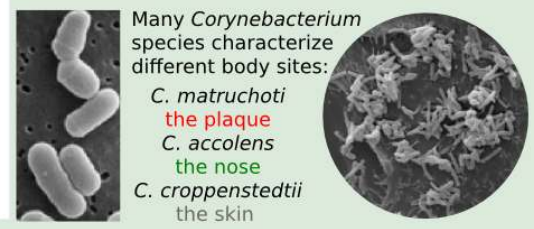
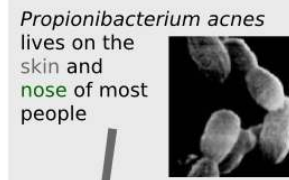
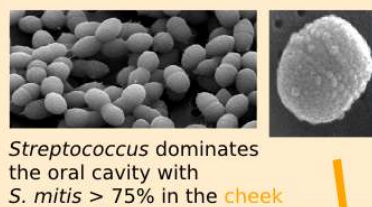




What defines the core “normal” human microbiome that we all share?



A map of diversity in the human microbiome



○ Commensal microbes
☆ Potential pathogens

The four most abundant phyla

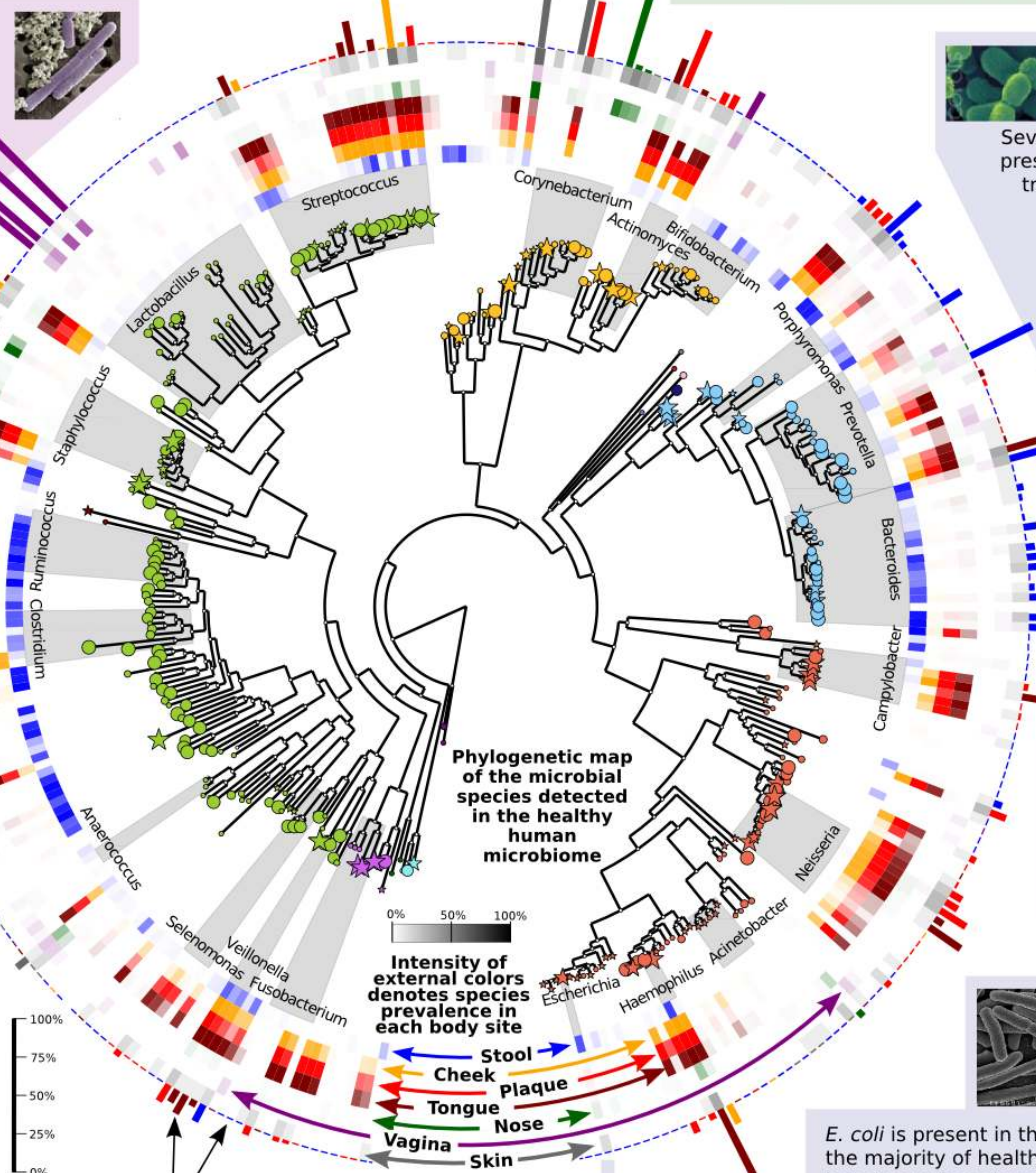
- Actinobacteria
- Bacteroidetes
- Firmicutes
- Proteobacteria

Low abundance phyla

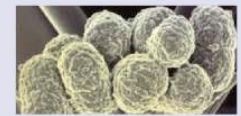
- Chloroflexi
- Cyanobacteria
- Euryarchaeota
- Fusobacteria
- Lentisphaerae
- Spirochaetes
- Synergistetes
- Tenericutes
- Thermi
- Verrucomicrobia

National Institutes of Health
Human Microbiome Project

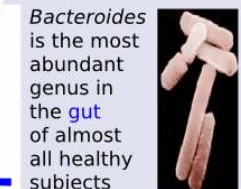
N. Segata & C. Huttenhower
<http://huttenhower.sph.harvard.edu>
(generated using Cytoscape and mOTU from MetaPhlans analysis)



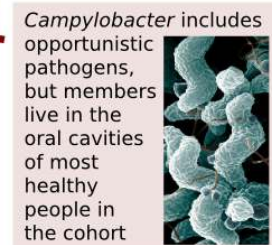
Several *Prevotella* species are present in the gastrointestinal tract. *P. copri* is present in 19% of the subjects and dominates the **intestinal** flora when present



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



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

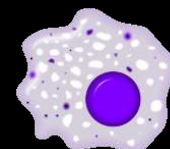
Microbial experiments

- 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



Dirk Gevers



Human Microbiome Project 2

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



George Weingart



Emma Schwager



Eric Franzosa



Boyu Ren



Tiffany Hsu



Ali Rahnava



Ramnik Xavier



Human Microbiome Project

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



Ayshwarya Subramanian



Jim Kaminski



Regina Joice



Koji Yasuda



Kevin Oh



Galeb Abu-Ali



Wendy Garrett



Nicola Segata



Gautam Dantas
Molly Gibson



Afrah Shafquat



Randall Schwager



Chengwei Luo



Keith Bayer



Moran Yassour



Alexandra Sirota



Andy Chan



Katherine Lemon



Brendan Bohannon
James Meadow



