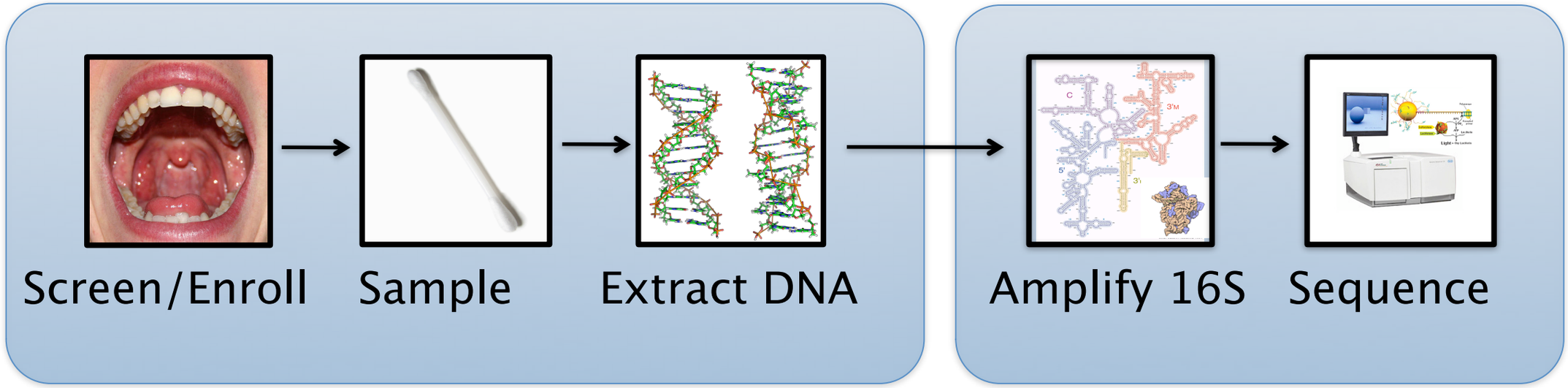
High throughput 16S sequencing for human metagenomics

Bruce Birren

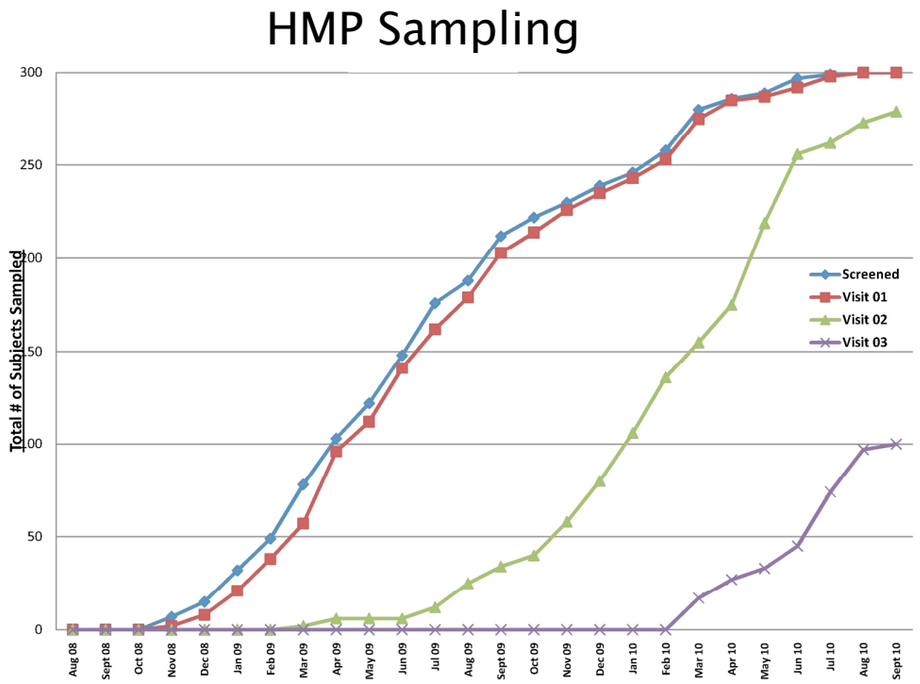
HUMAN MICROBIOME PROJECT HMP

Subjects to sequences



Clinical Sites – BCM & WashU

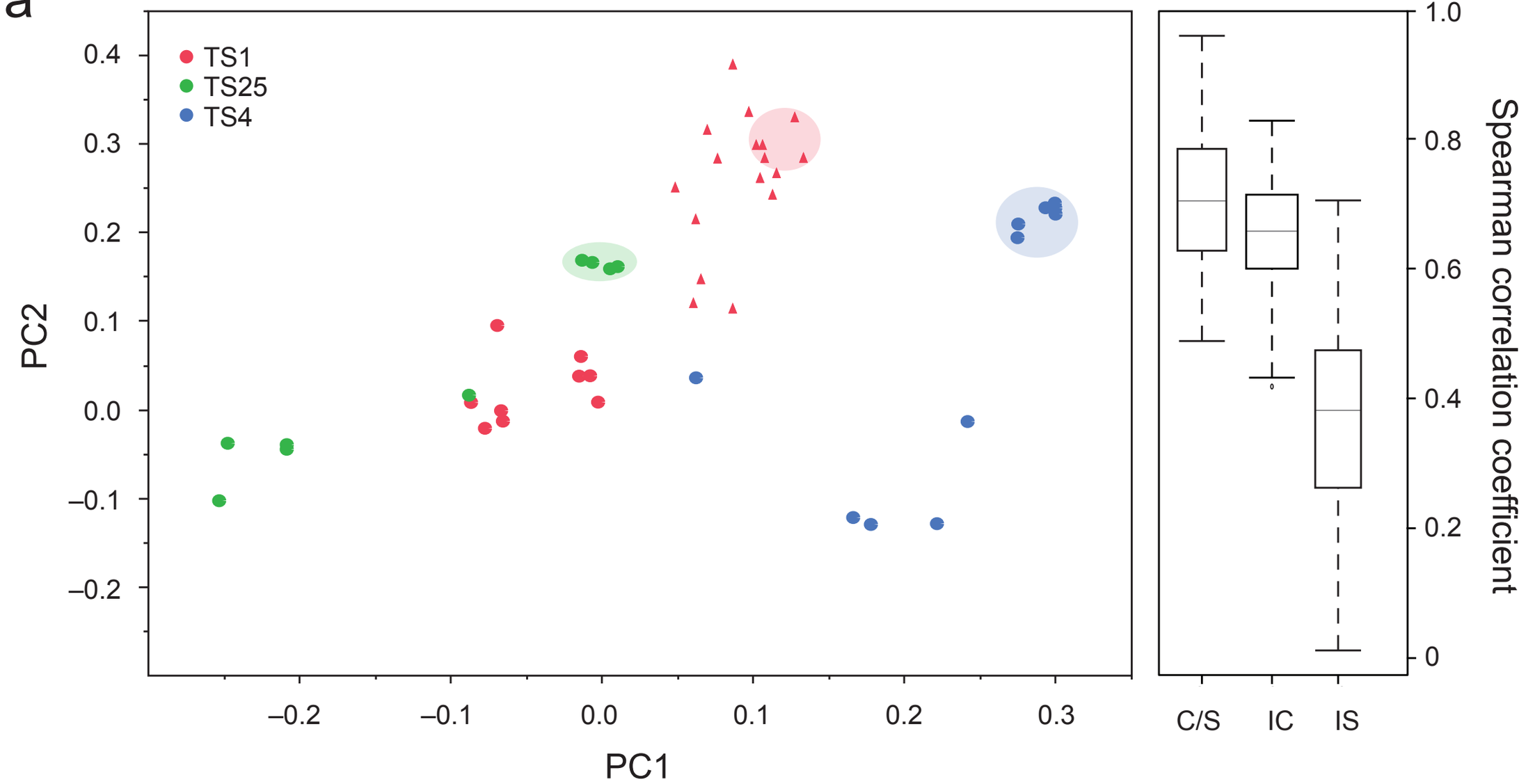
Sequencing Centers
BCM, Broad, JCVI & WashU



17,000 primary specimens

Independent protocols - inconsistent data

a

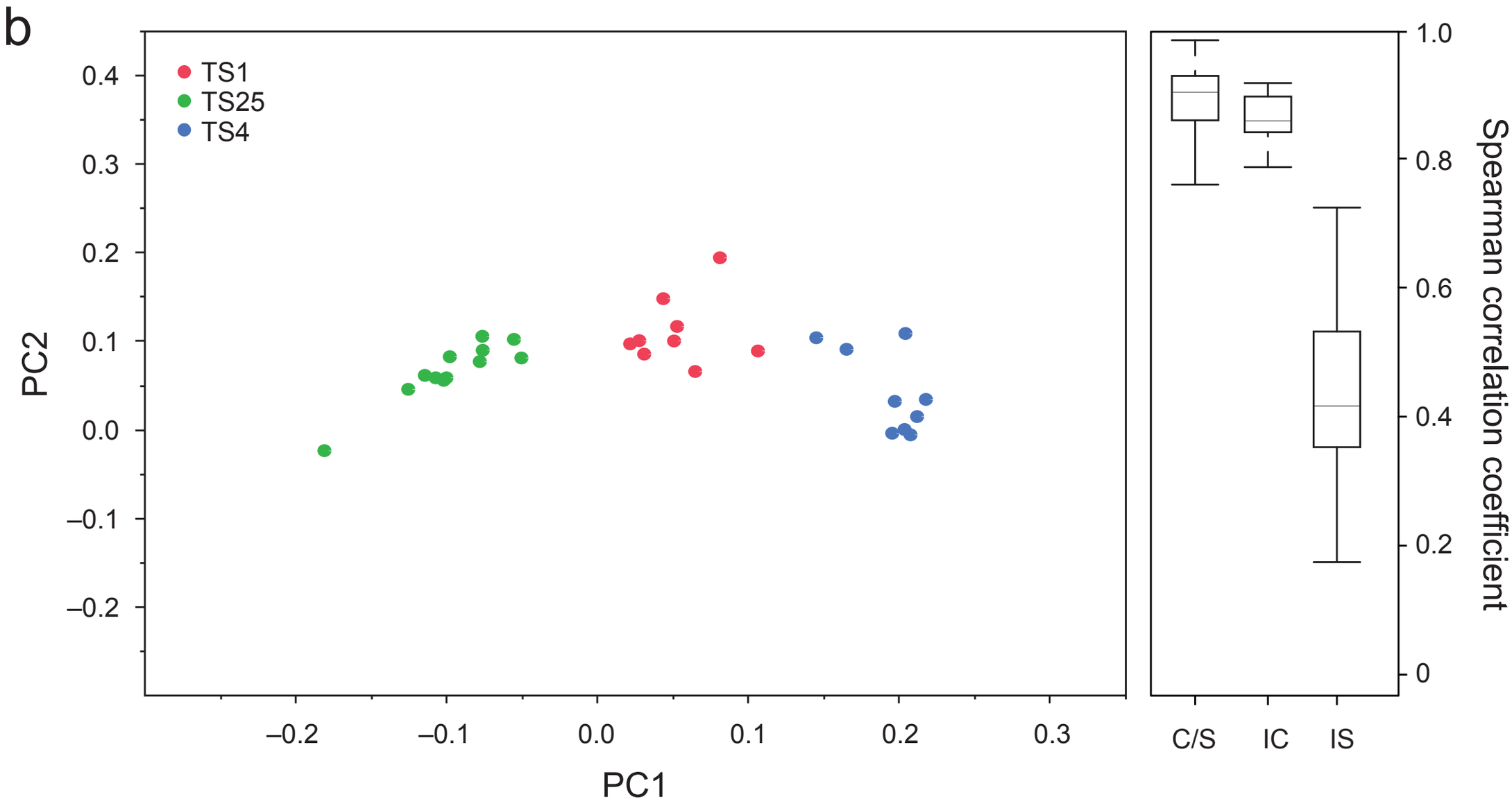


PC1&PC2 account for 69% of variation

Three different gut samples are gift of J. Gordon

HUMAN MICROBIOME PROJECT HMP

Common protocols - consistent data



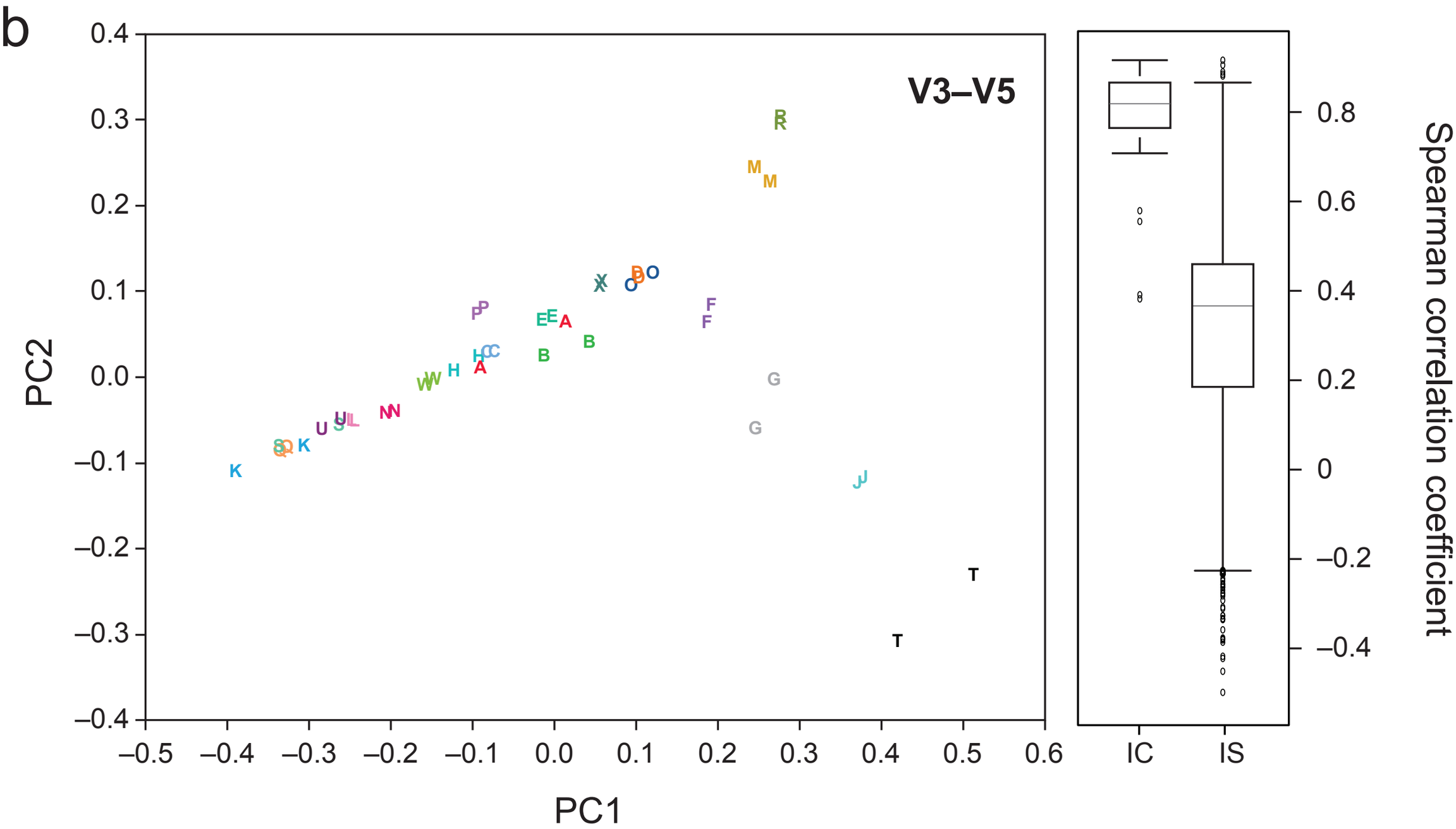
Contributions from lab and bioinformatic steps

PC1&PC2 account for 69% of variation

HUMAN MICROBIOME PROJECT HMP

Consistency across centers

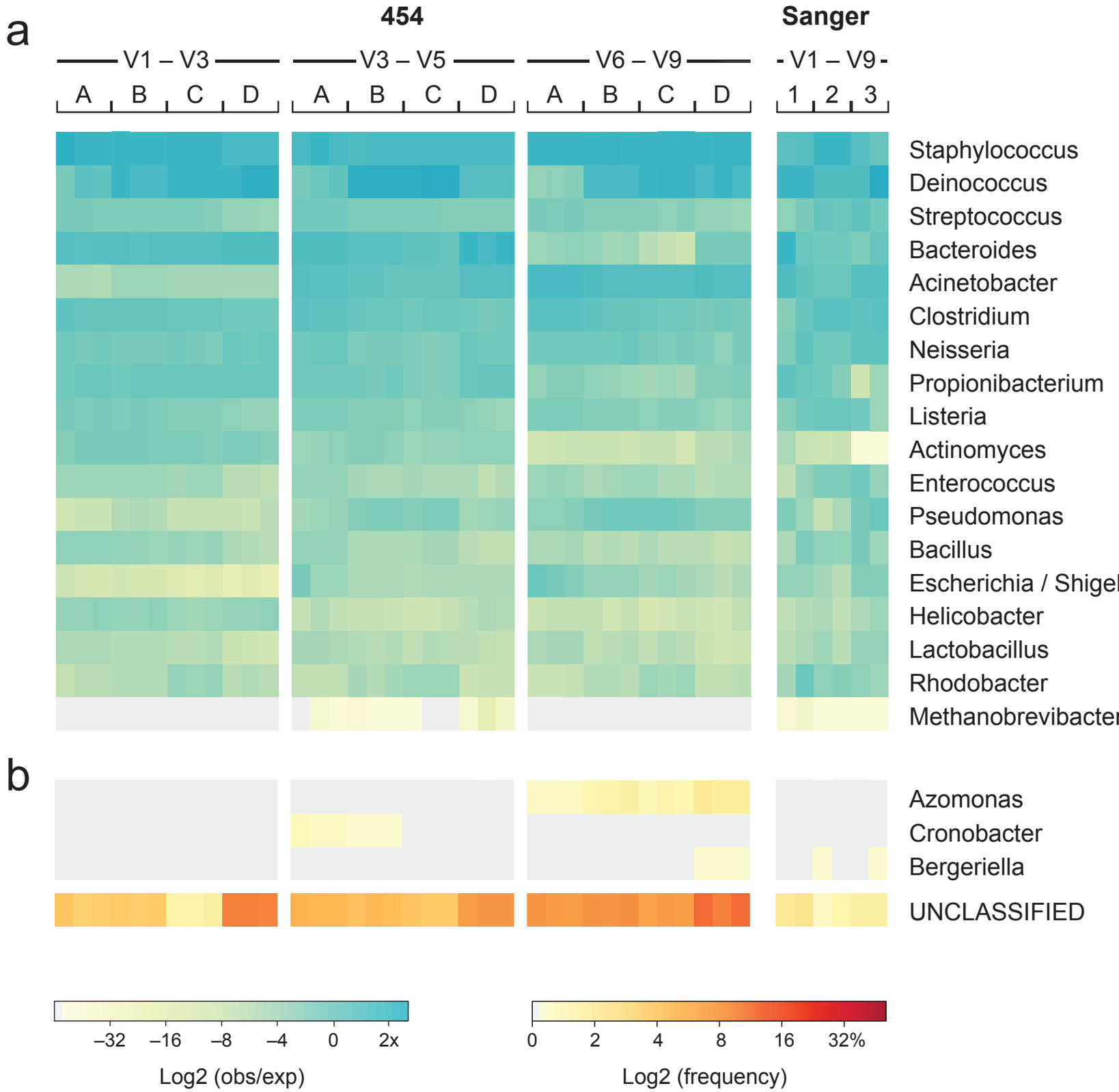
b



PC1&PC2 account for 77% of variation

HUMAN MICROBIOME PROJECT HMP

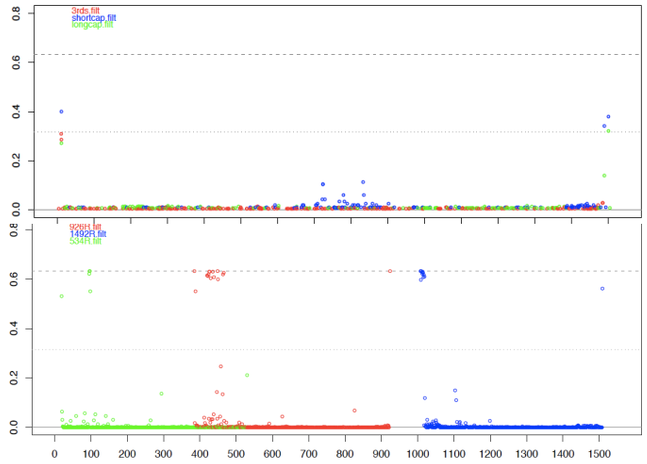
Accuracy validation via mock community



Sources of Unclassified Reads

Sequencing Errors

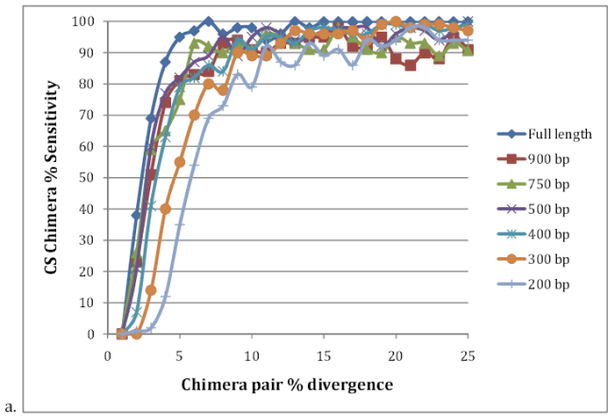
- Significant in all seq. technology
- Quality filtering effective



Chimeric Reads

- Ability to detect
- Abundance
- Similarity
- Protocol

a.



Chimera Slayer - Haas et. al Gen. Res. 2011 21:494-504

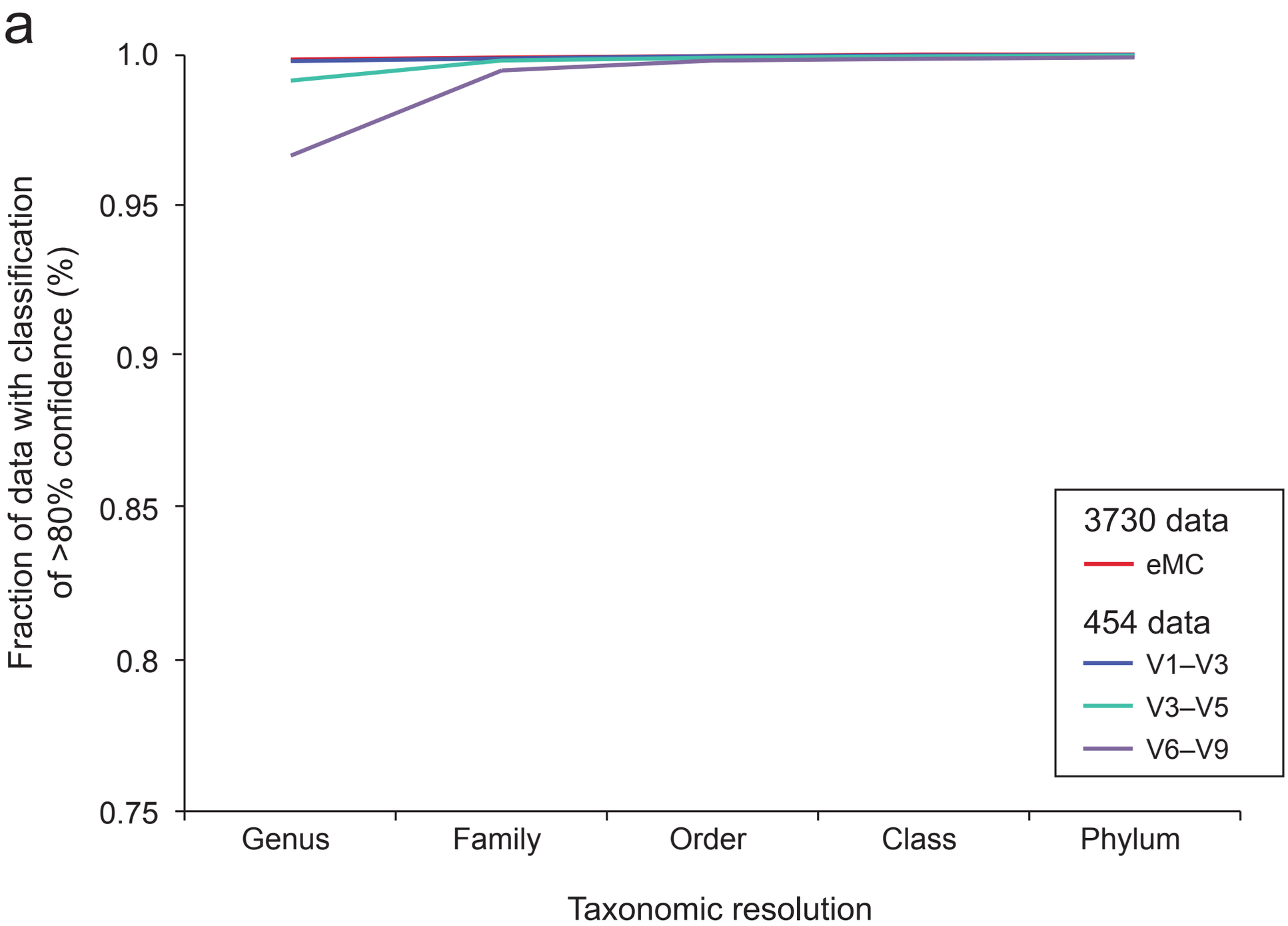
HUMAN MICROBIOME PROJECT HMP

Chimera rates in 16S data sets

Samples	% Observed Chimera content			
	ABI3730	**454 FLX Titanium**		
	V1-V9	V1-V3	V3-V5	V6-V9
eMC	5.99±3.07	14.26±10.34	14.75±9.45	13.49±8.52
gut	7.71±6.46	22.90±8.56	16.03±2.86	17.76±3.76
oral	7.22±6.35	20.55±11.73	10.98±4.01	9.10±5.02
skin	3.49±5.77	11.15±1.36	7.51±2.49	5.73±1.69
vaginal	6.31±6.64	12.60±6.70	6.62±3.51	3.00±1.65

HUMAN MICROBIOME PROJECT HMP

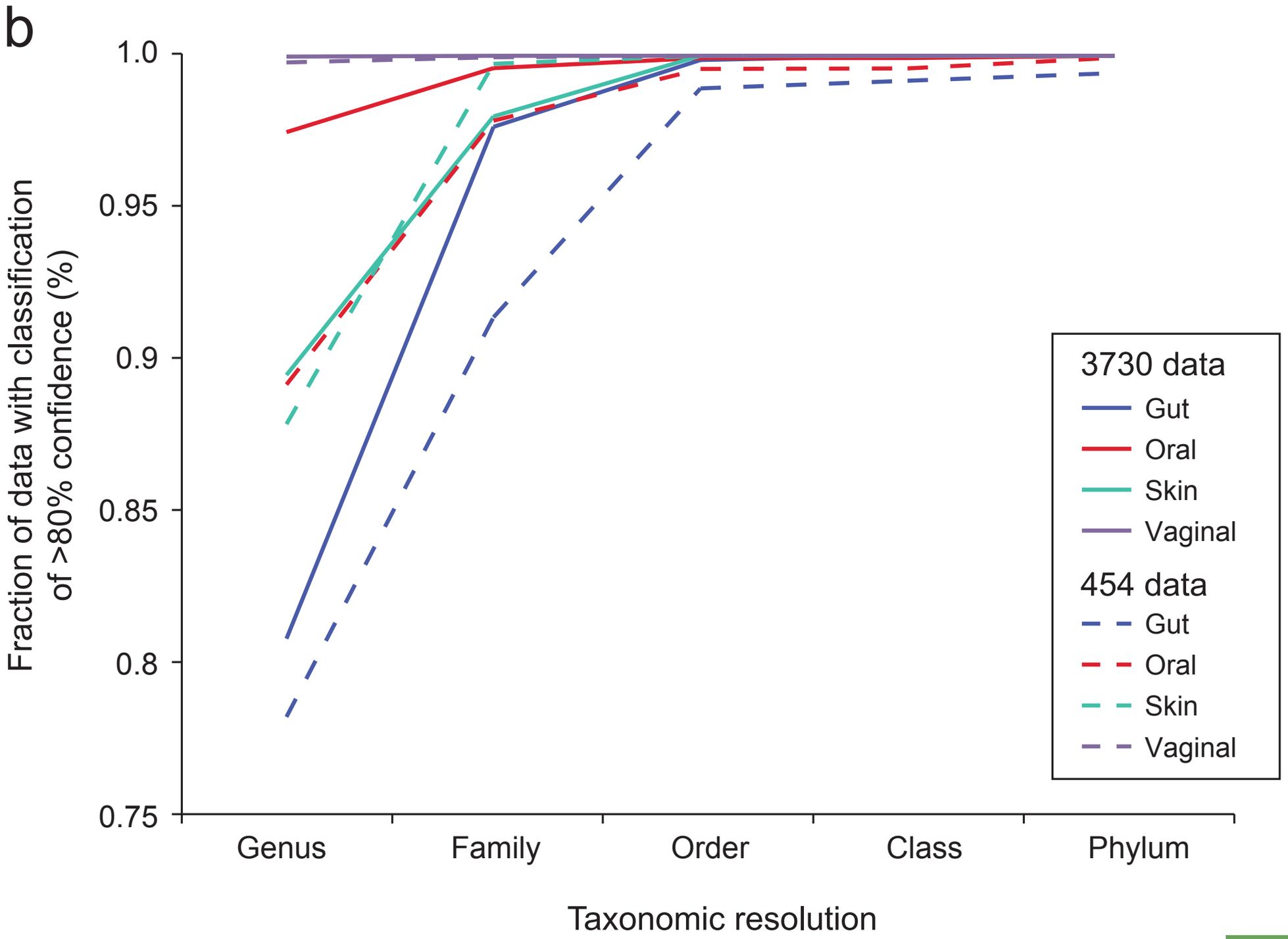
Classifying reads from mock community



RDP classification

HUMAN MICROBIOME PROJECT HMP

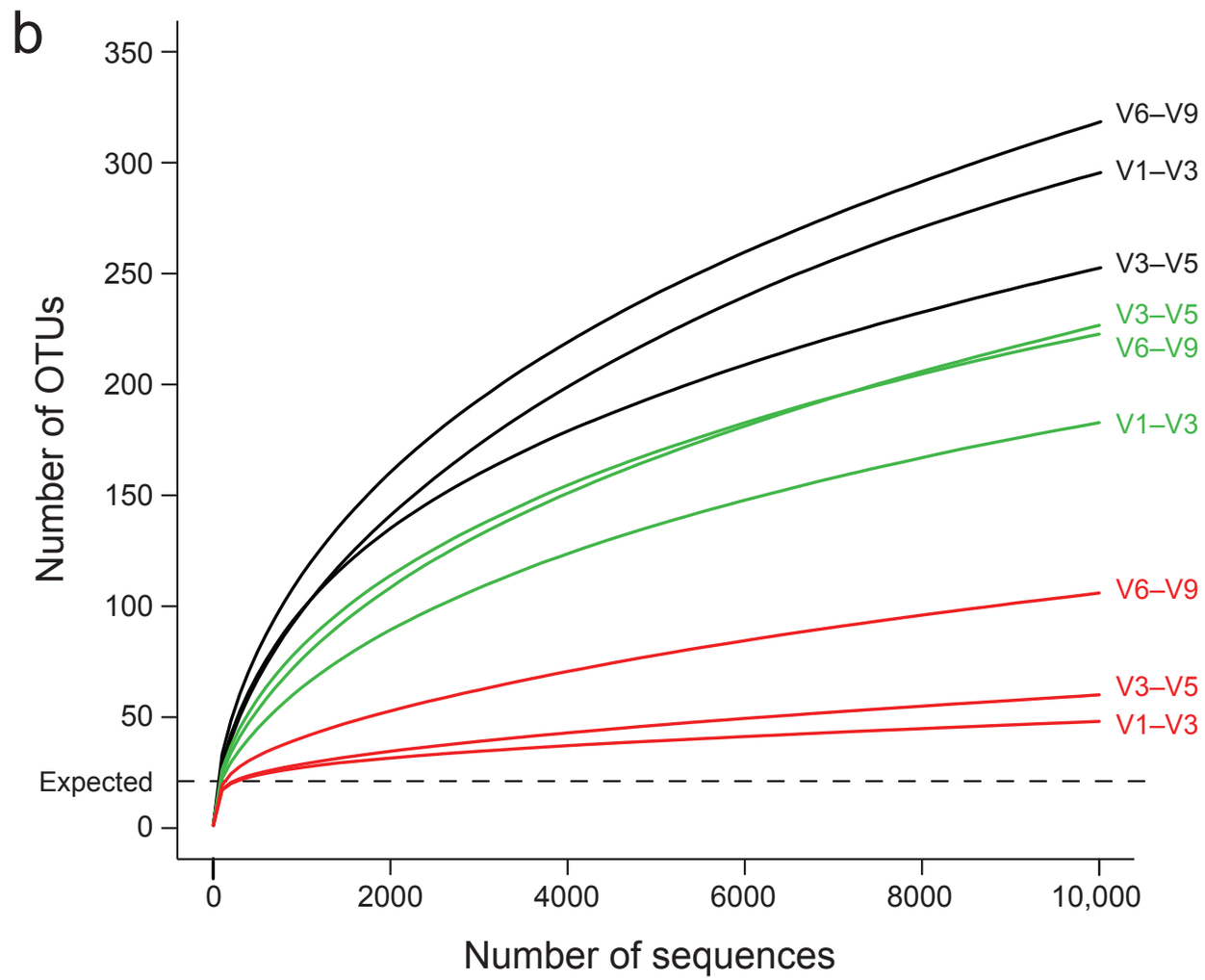
Classifying ‘real’ communities



RDP classification

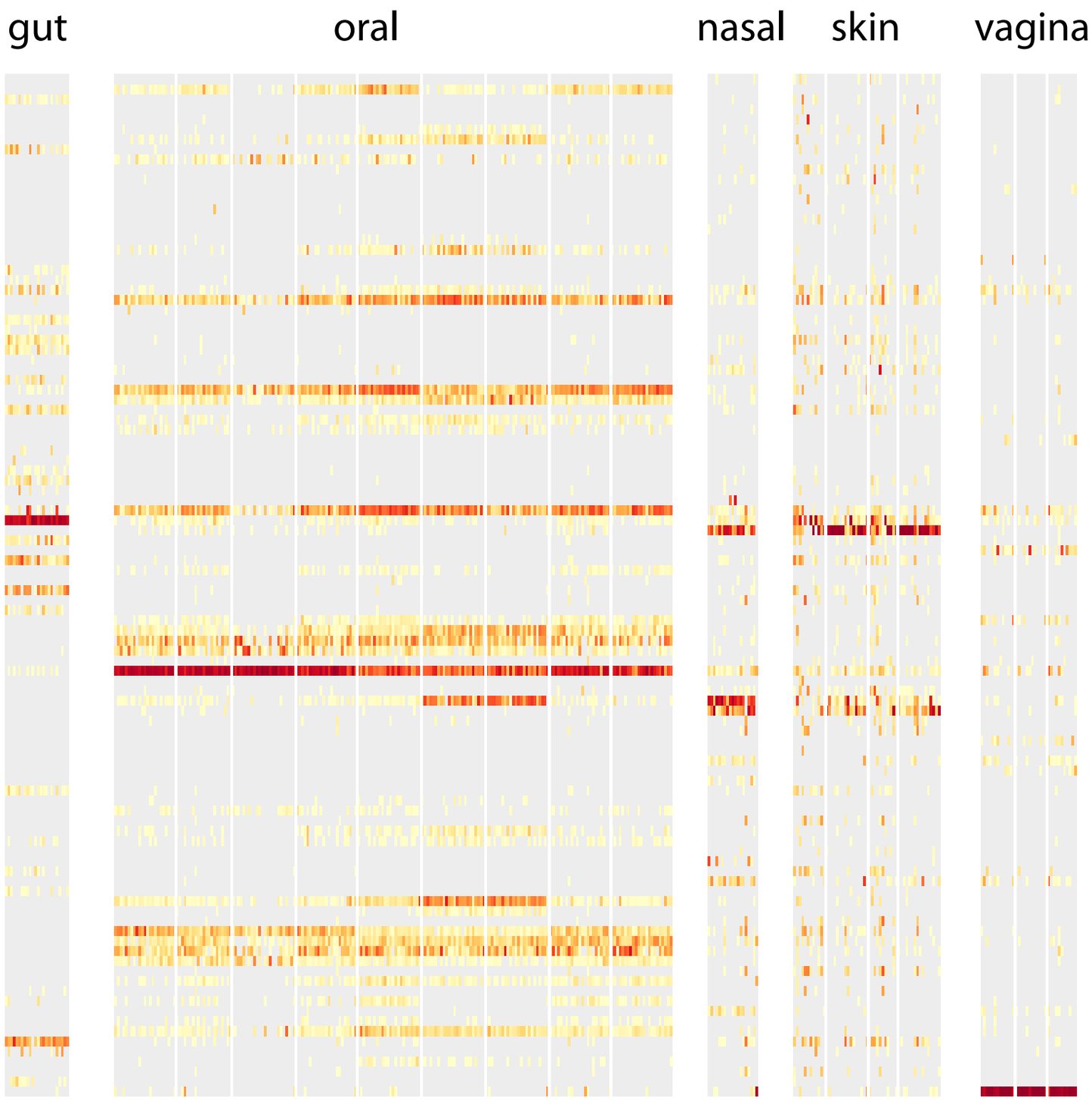
HUMAN MICROBIOME PROJECT HMP

Estimating diversity



HUMAN MICROBIOME PROJECT HMP

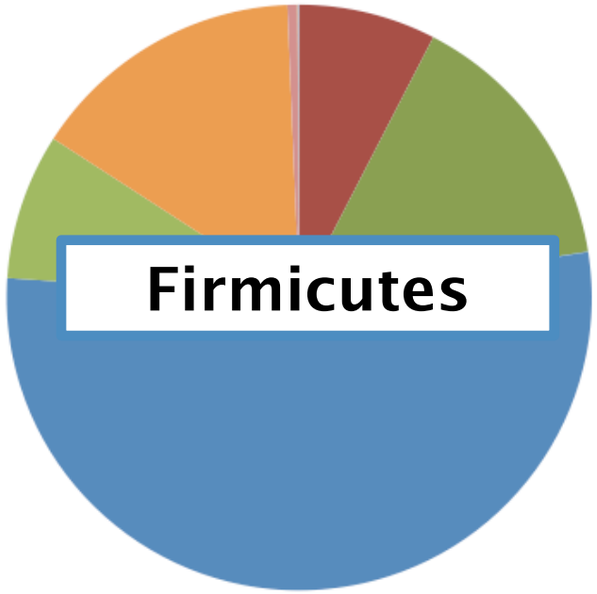
Community composition within clinical samples

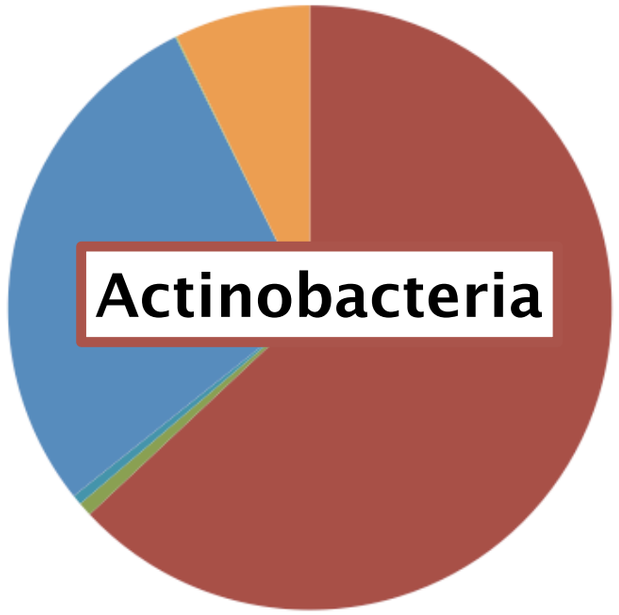


First 24 subjects V3-V5

HUMAN MICROBIOME PROJECT HMP

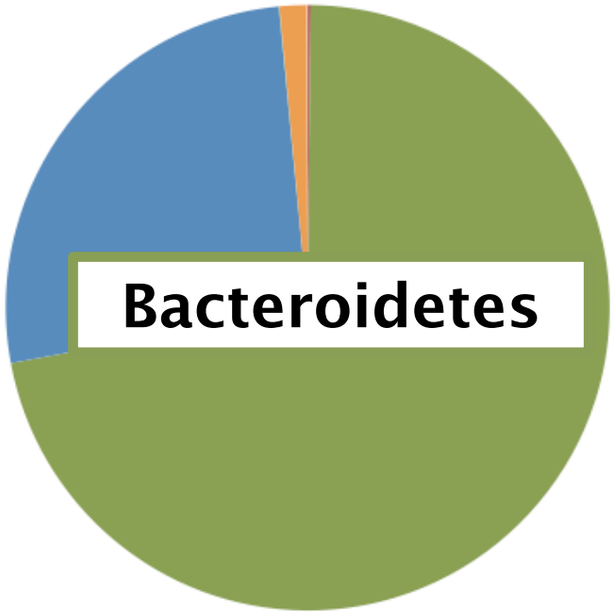
19 Phyla in 24 Subjects





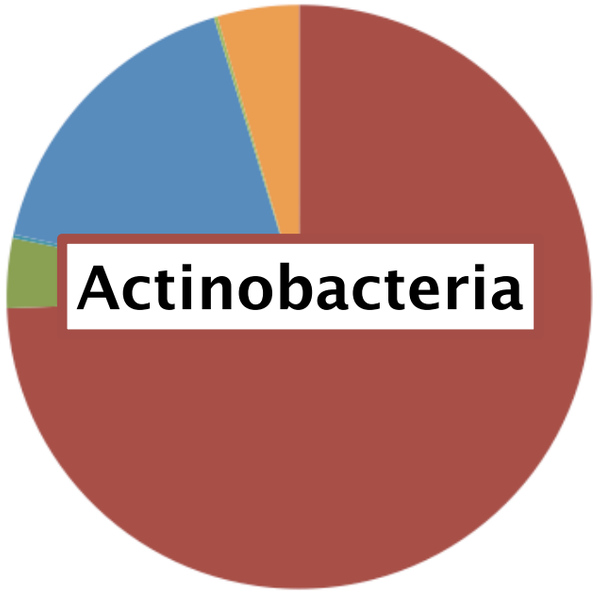
Anterior Nares

Oral Cavity

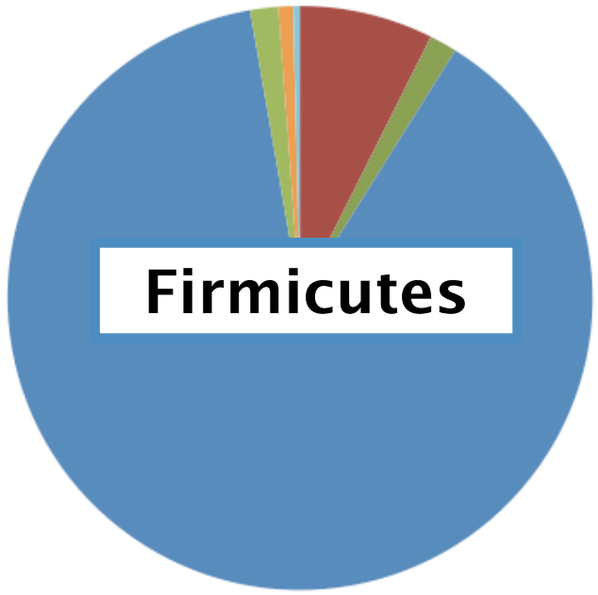


GI Tract

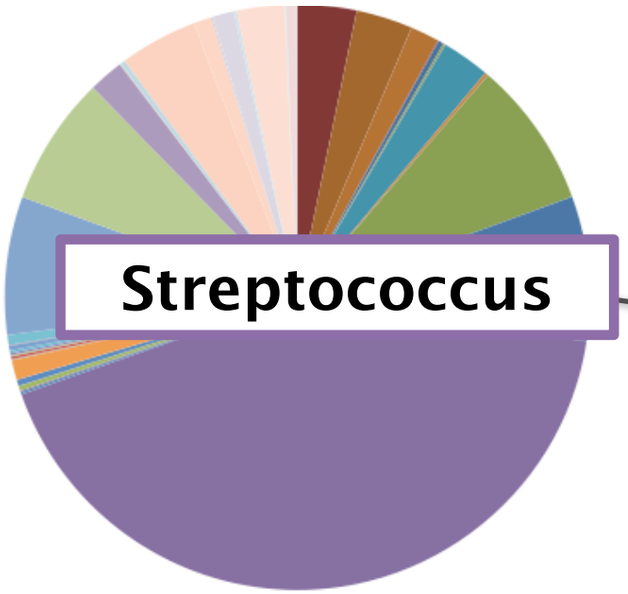
Skin



Vagina



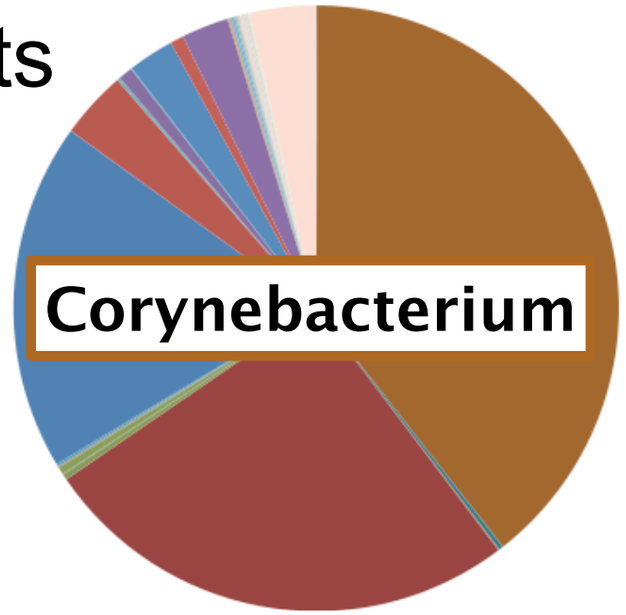
293 Total Genera in 24 Subjects



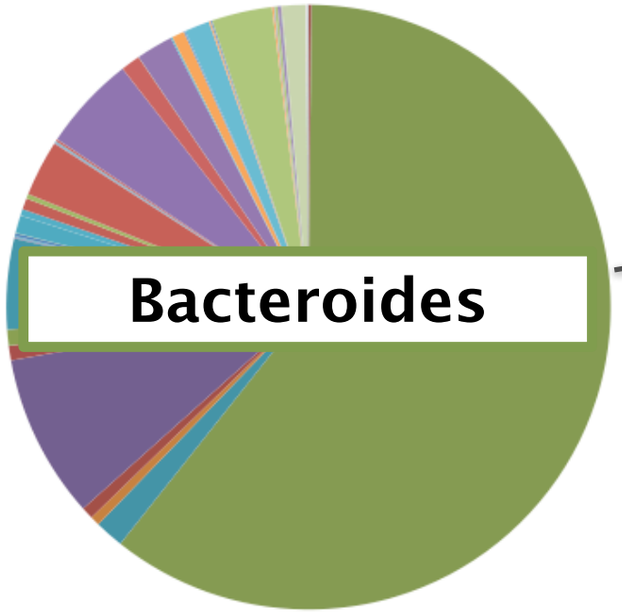
Streptococcus

Oral Cavity

Anterior Nares



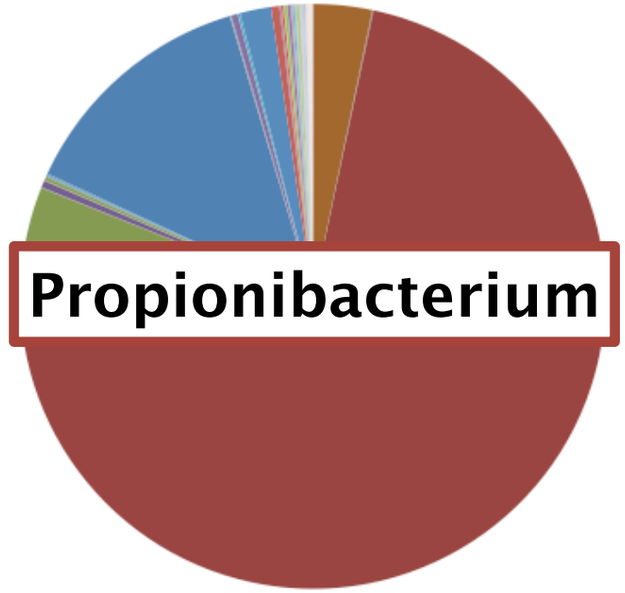
Corynebacterium



Bacteroides

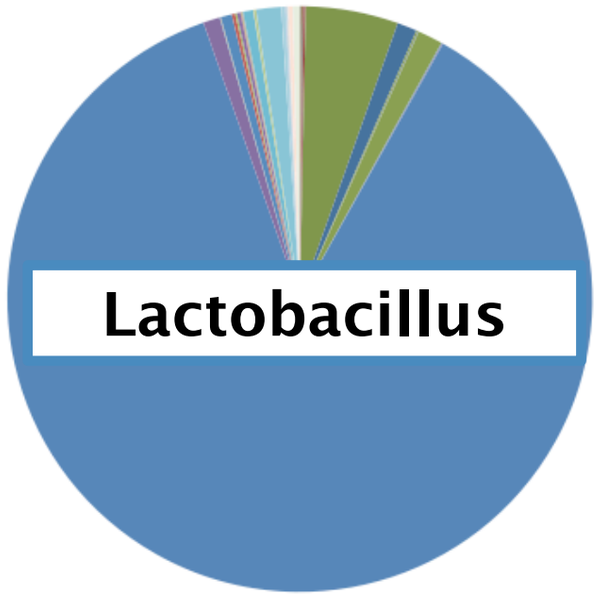
GI Tract

Skin



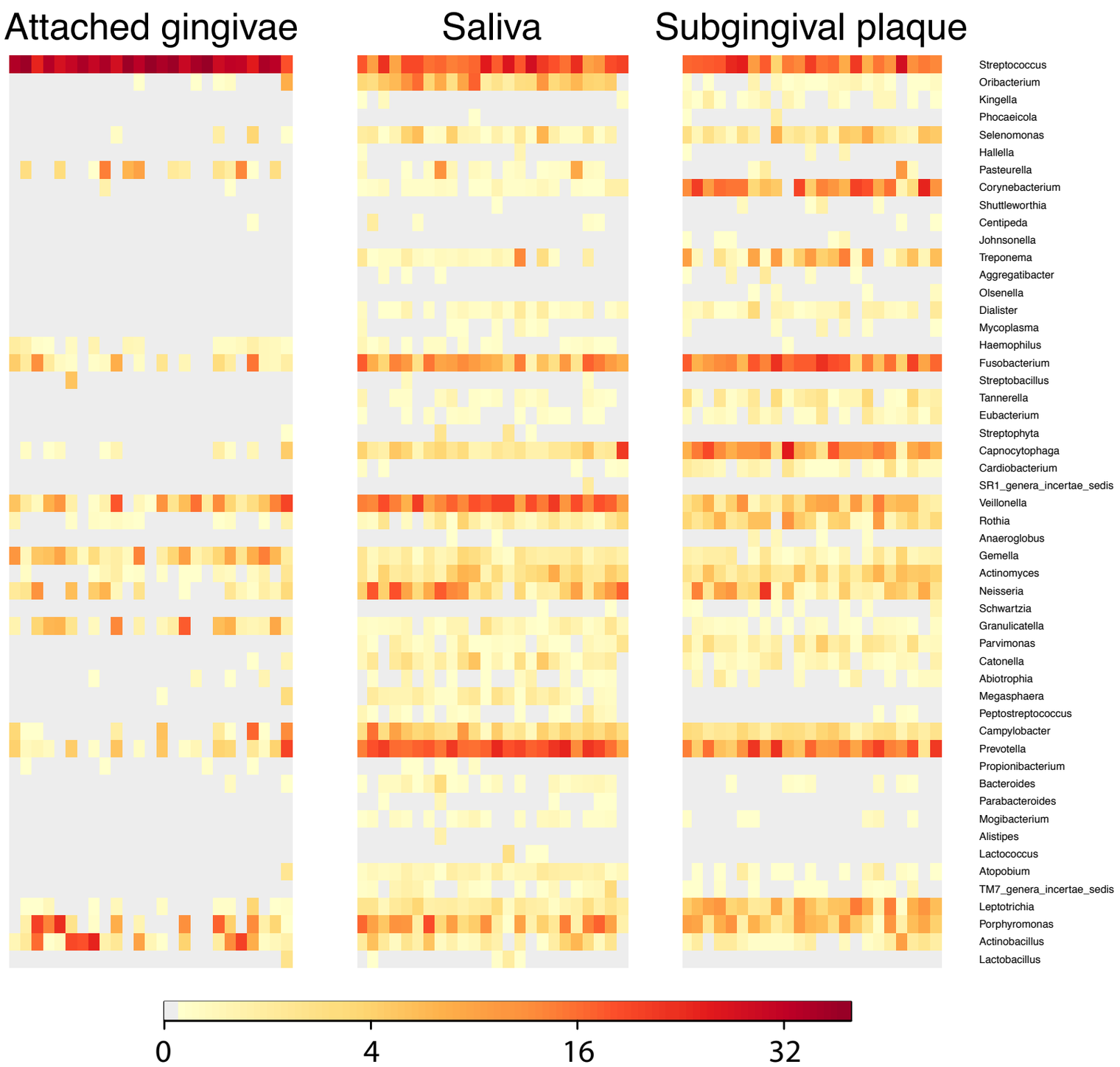
Propionibacterium

Vagina



Lactobacillus

Individual Variation in Community Composition



454 V3-V5 sequence data

HUMAN MICROBIOME PROJECT

Defining a ‘core’ microbiome?

Comparison	Genus-level core

Methe, Yooseph, Li, Bihan

Defining a ‘core’ microbiome?

Comparison	Genus-level core
ALL Samples from ALL Subjects	No

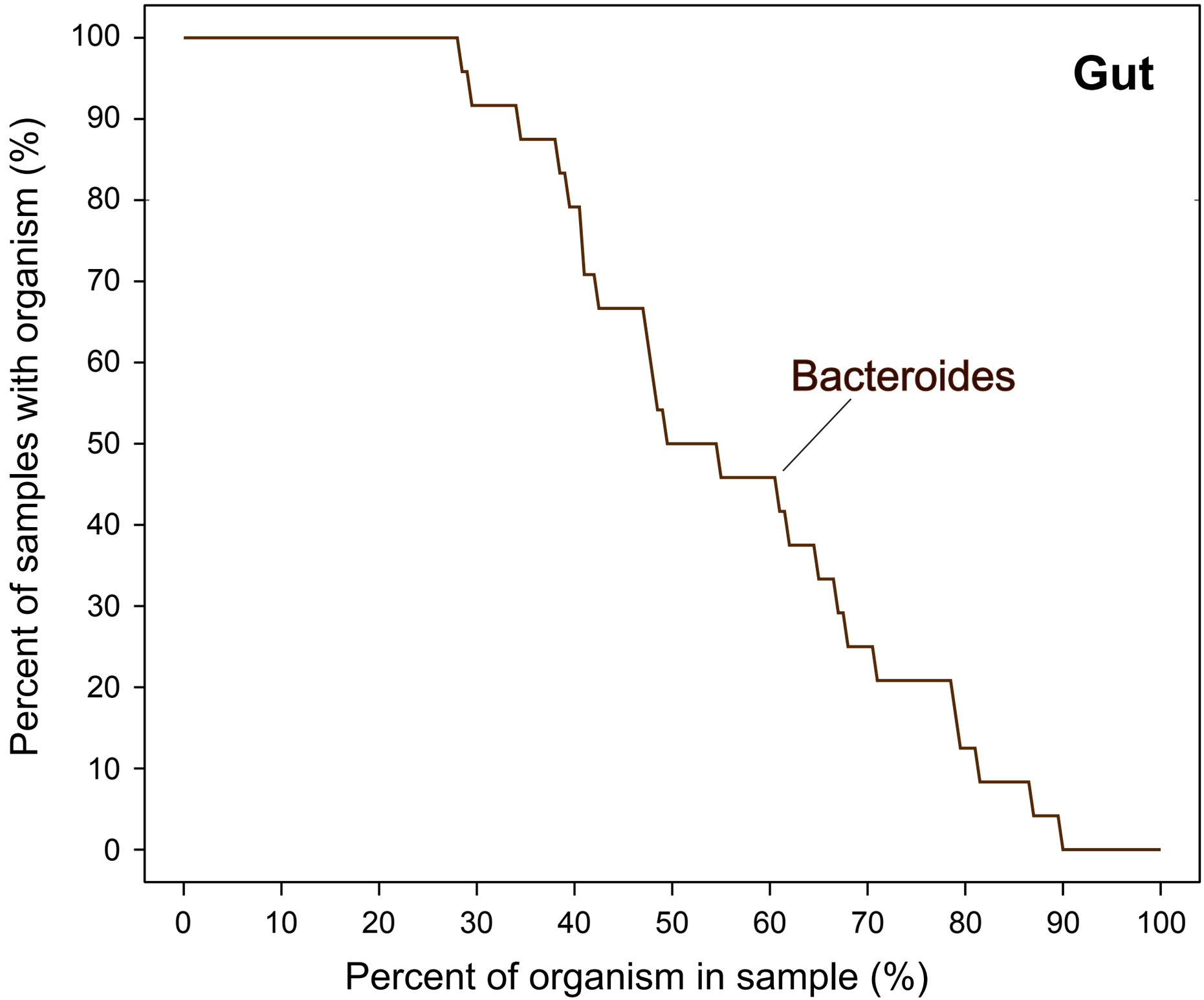
Methe, Yooseph, Li, Bihan

Defining a ‘core’ microbiome?

Comparison	Genus-level core
ALL Samples from ALL Subjects	No
ALL Samples from Gut	Yes, Bacteroides
ALL Samples from Nares	Yes, Corynebacterium
ALL Samples from Oral Cavity	Yes, Streptococcus
ALL Samples from Skin	No
ALL Samples from Vagina	Yes, Lactobacillus

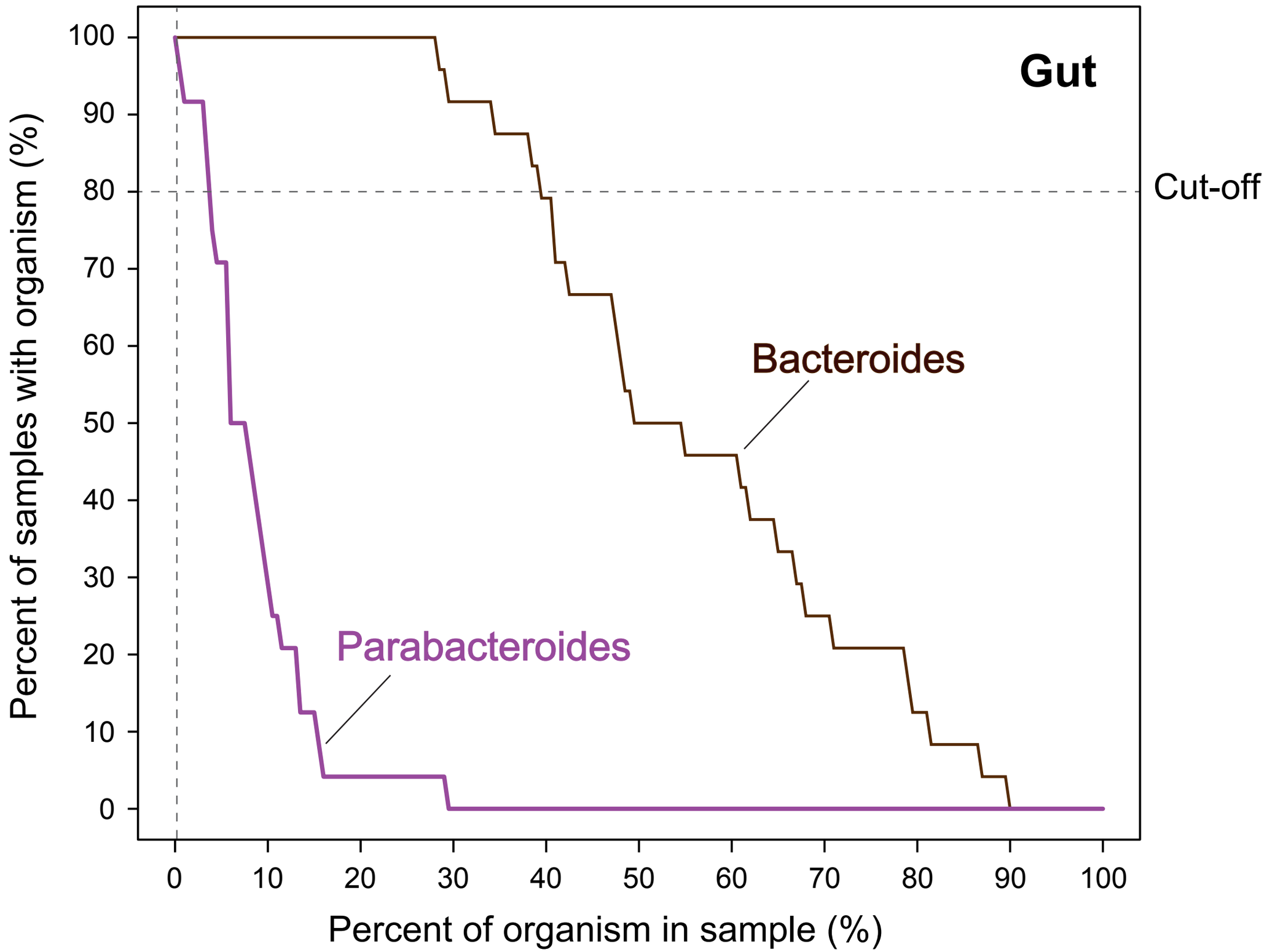
Methe, Yooseph, Li, Bihan

Core abundance is highly variable



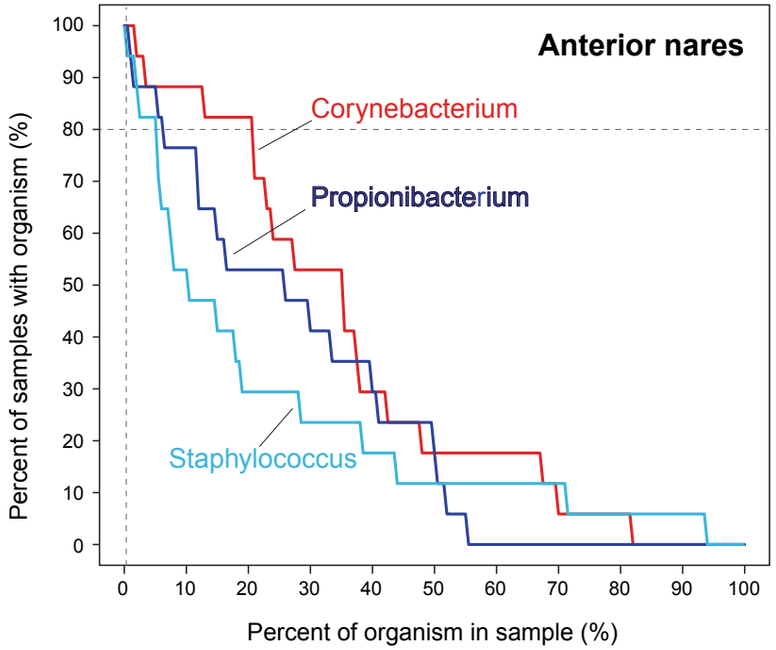
Methe, Yooseph, Li, Bihan

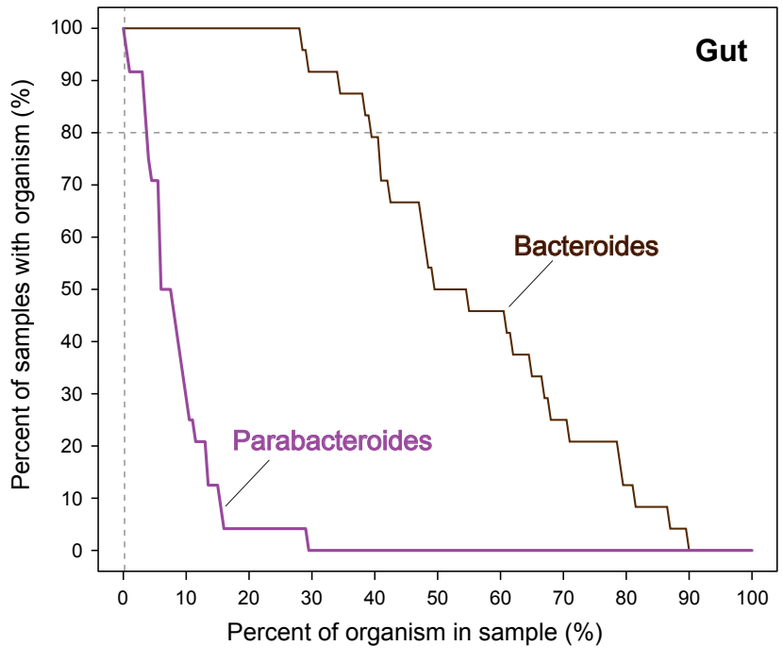
A ‘relaxed’ core definition

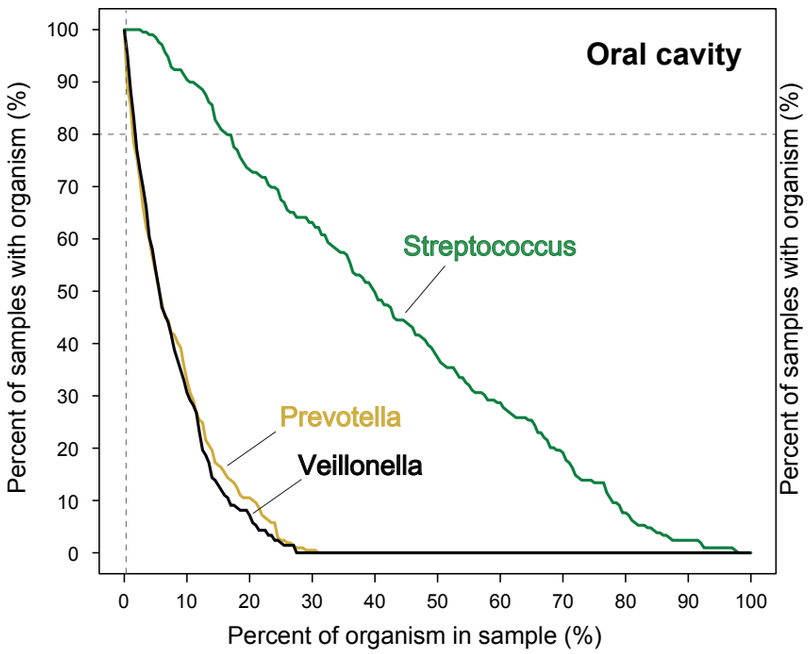


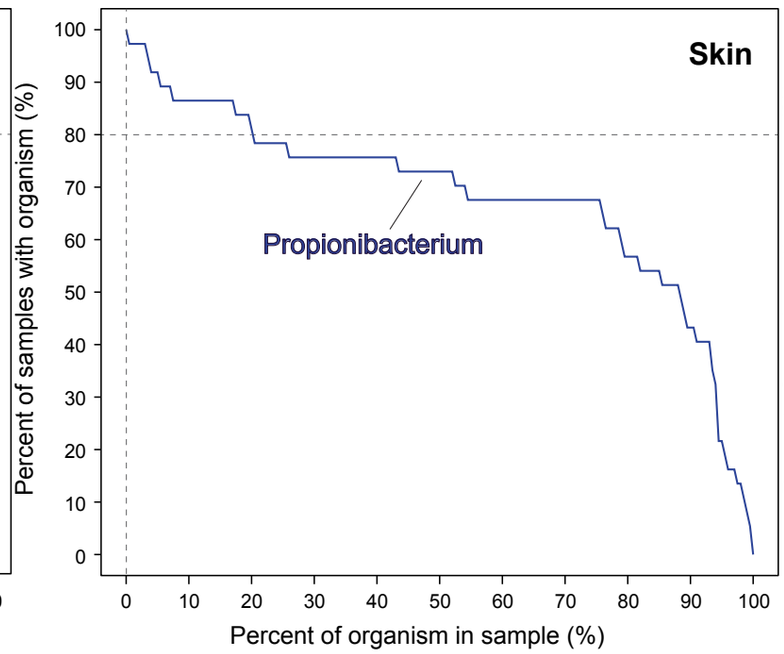
Methe, Yooseph, Li, Bihan

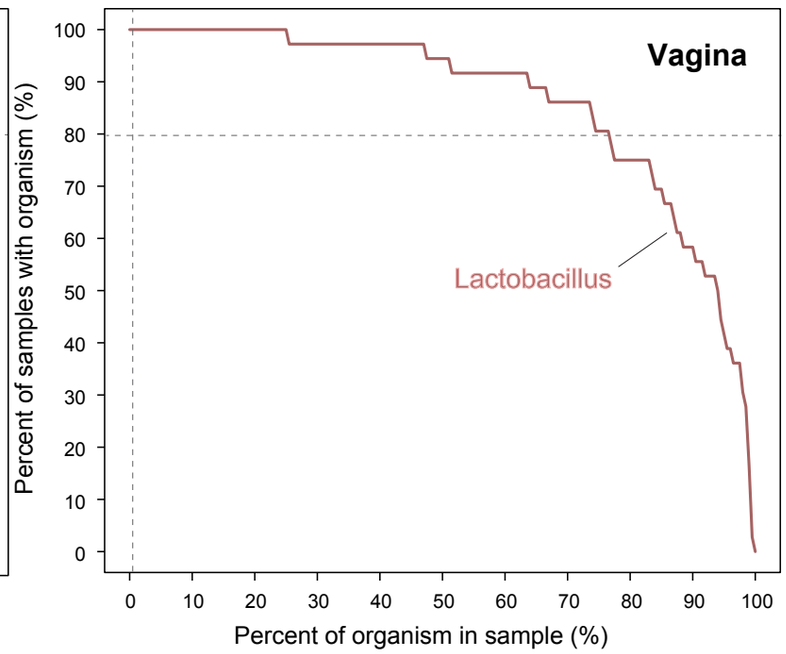
Cores across body sites





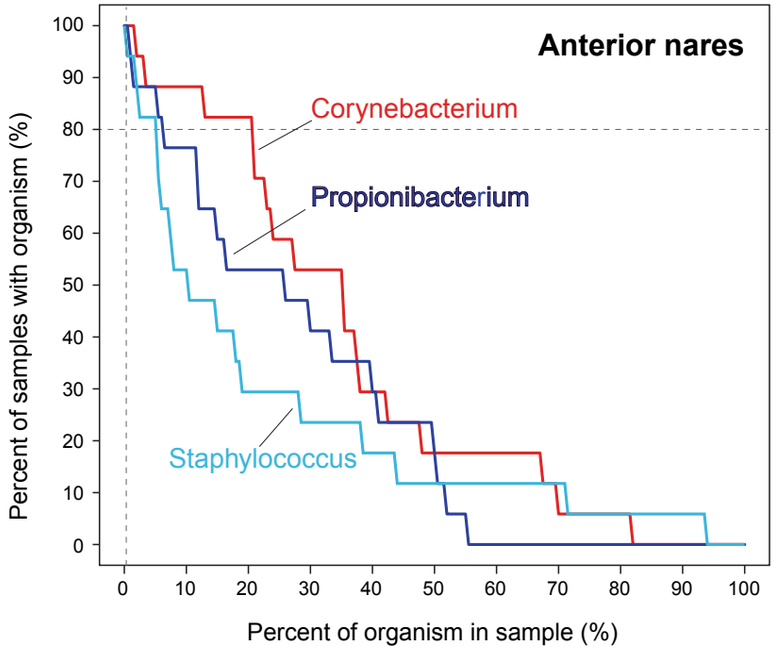


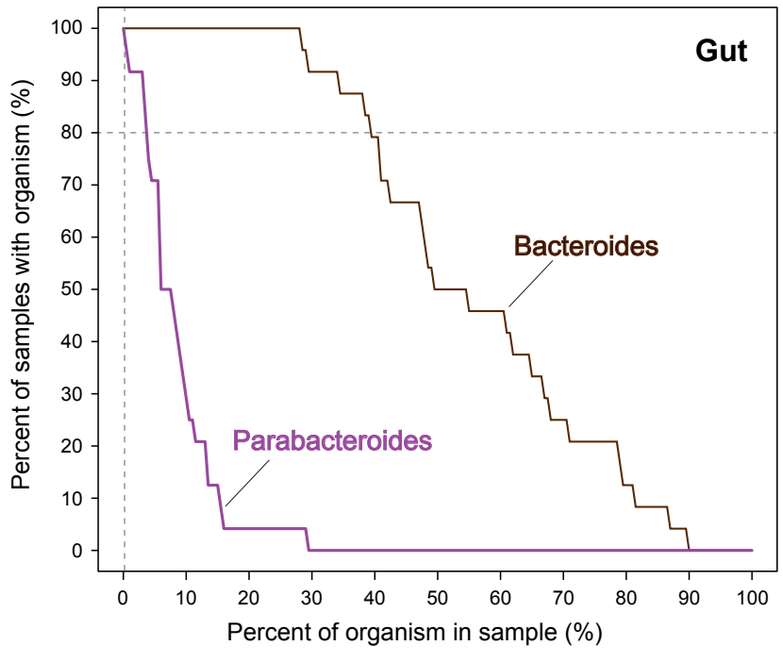


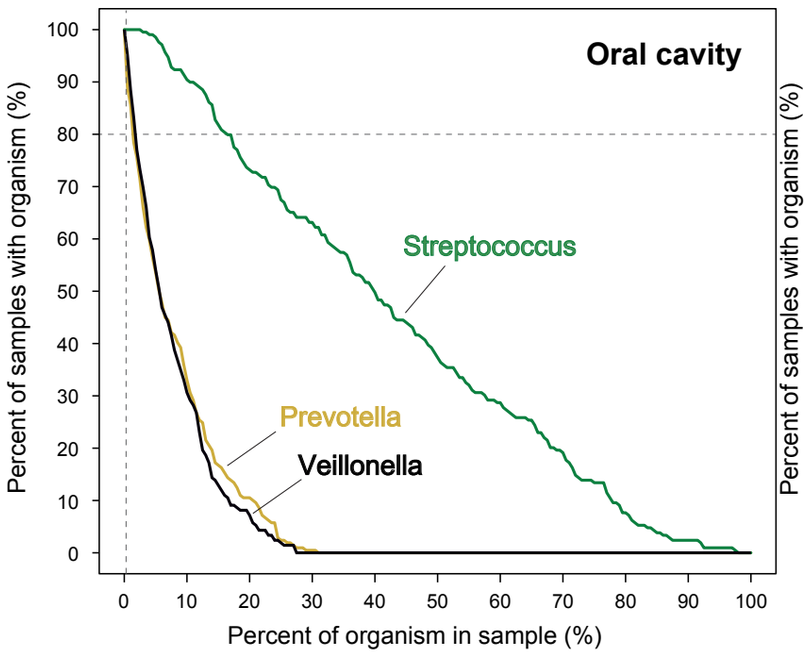


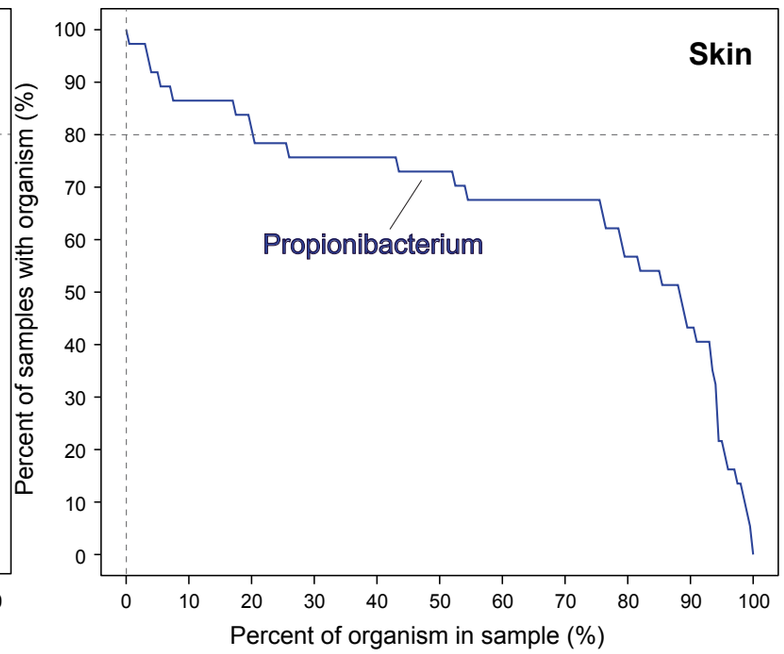
Methe, Yooseph, Li, Bihan

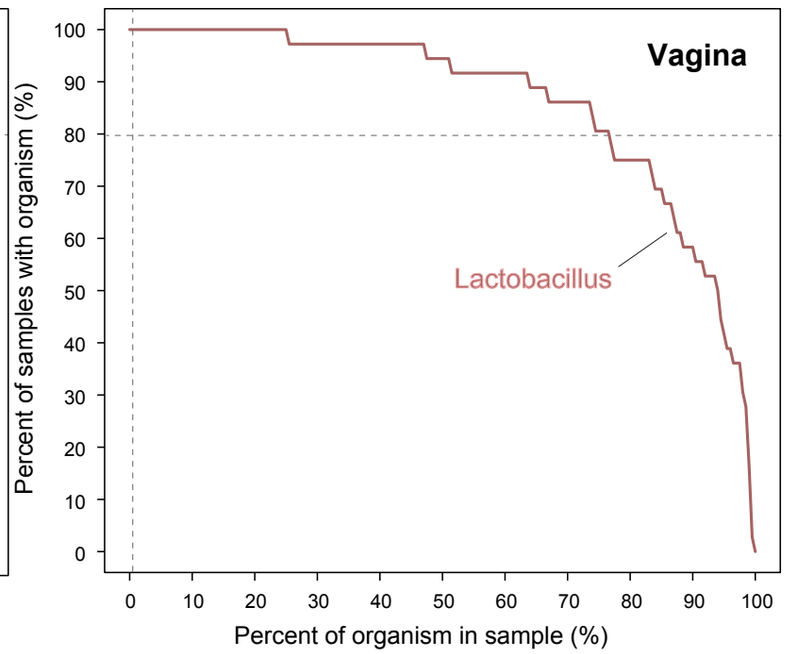
Cores across body sites







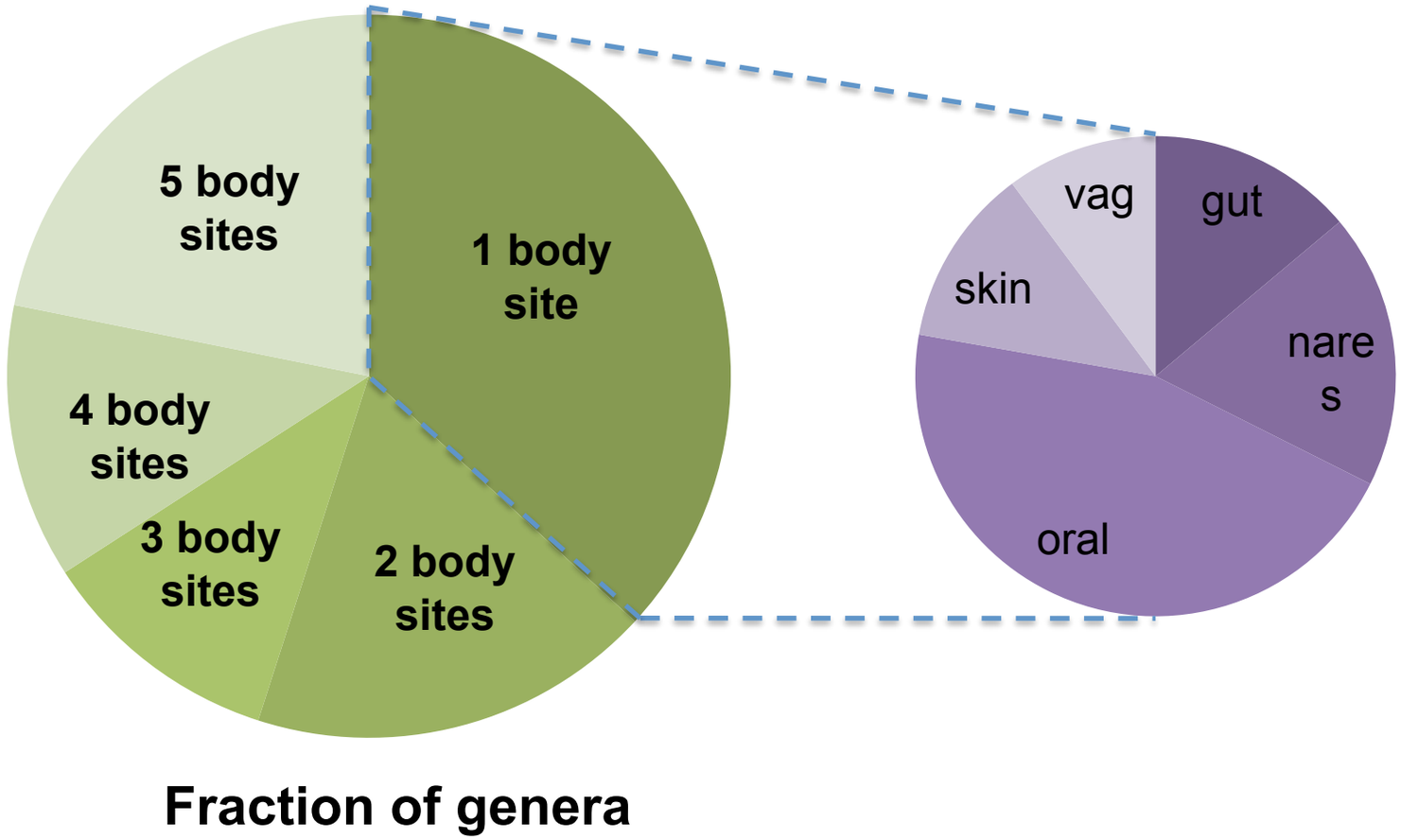




Core is small at any definition, body site specific
Abundance of core members varies dramatically

Methe, Yooseph, Li, Bihan

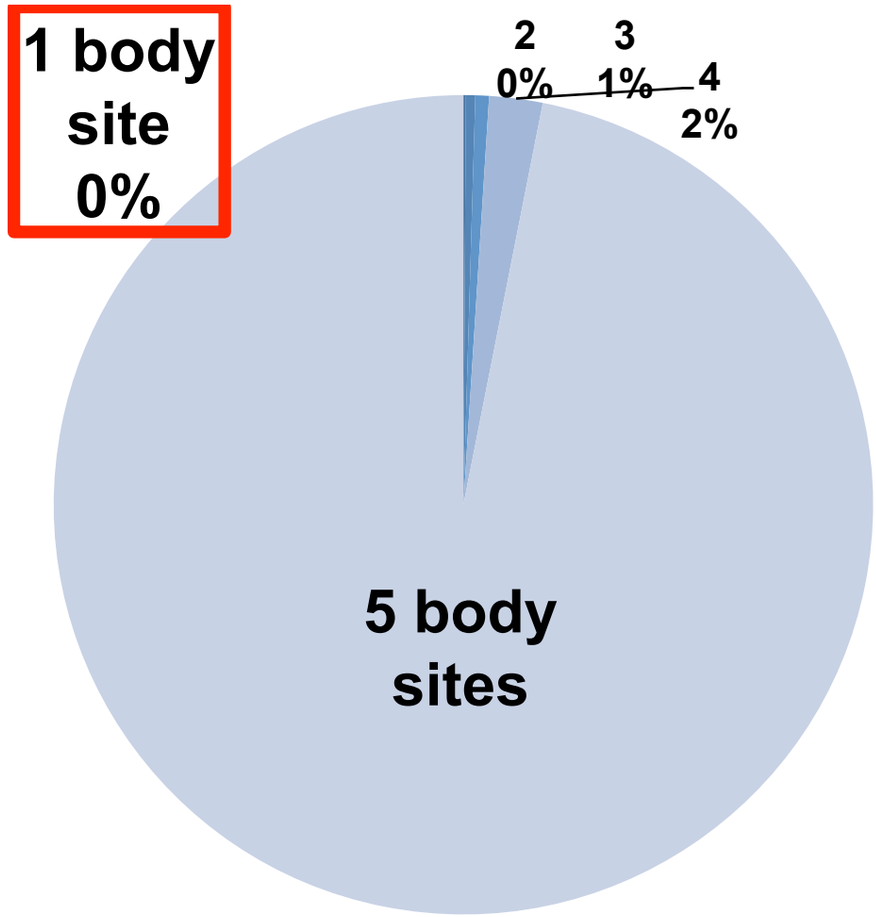
How unique are organisms to particular body sites?



293 Total Genera

Earl, Givers

‘Specialist’ genera represent a tiny fraction of data

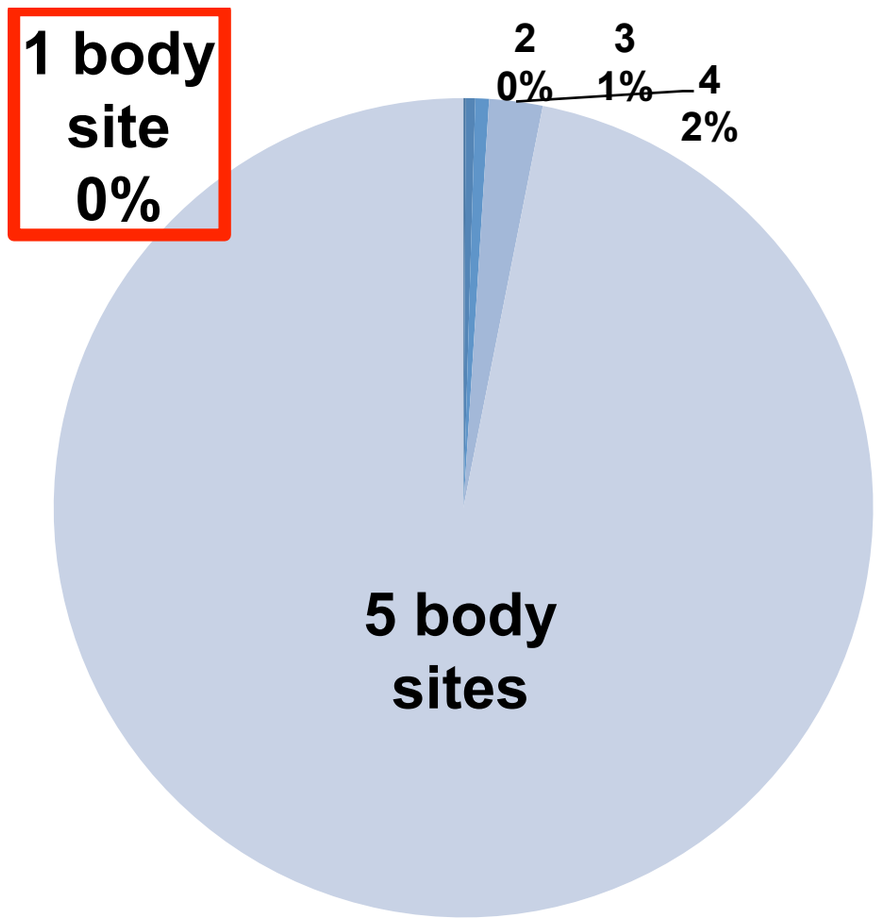


Fraction of total reads

Earl, Givers

‘Specialist’ genera represent a tiny fraction of data

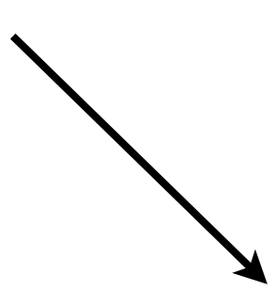
97% of all reads belong to genera present in all five body sites!



Earl, Givers

How to Interpret?

16S data



Biology

How to Interpret?

16S data

WGS data

Biology

How to Interpret?

16S data

WGS data

Reference Genomes

Biology

High Quality Illumina assemblies

Genome	% GC	Length (Mb)	Contig #	Contig N50 (Kb)	Scaffold #	Scaffold N50 (Mb)	% Ref Covered	Scaffold Accuracy
E. coli_MG1655	51%	4.59	106	107	26	4.59	98.9%	100.0%
Streptococcus pneumoniae	40%	2.13	78	51	11	1.96	98.7%	99.6%
M. tuberculosis	66%	4.21	341	20	23	0.60	95.5%	96.1%

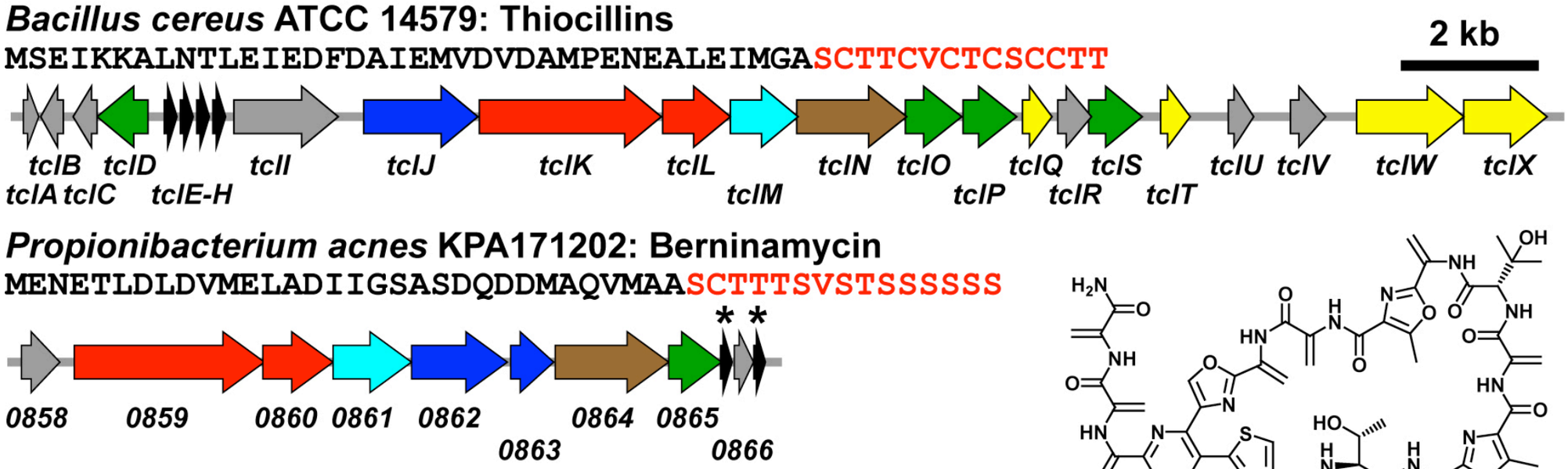
Gnerre et al. 2011 PNAS 108:1513-8

Illumina bacterial assembly in production

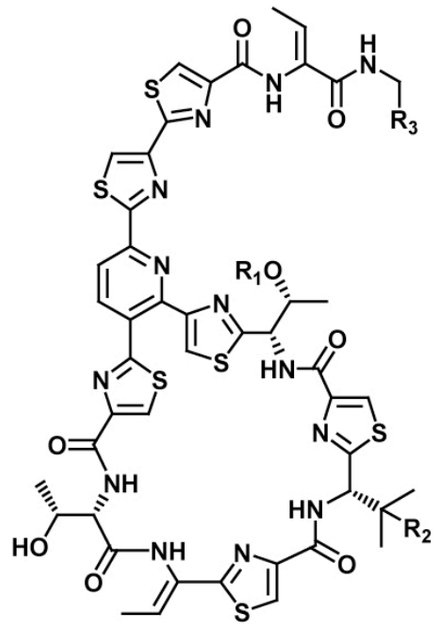
	% GC	Scaffold length (Mb)	# scaffolds	Scaffold N50 size (Mb)	# contigs	Contig N50 size (Kb)
Acinetobacter RUH2624	41	3.92	9	2.25	85	81
Bacteroides eggerthii 1_2_48FAA	47	4.55	12	0.94	76	131
Bacteroides sp 9_1_42FAA	44	5.64	29	0.85	66	214
Bifidobacterium bifidum NCIMB 41171	64	2.20	1	2.2	24	156
Clostridium inocuum 6_1_30	46	4.91	12	0.79	79	107
Coprobacillus sp. 8_2_54BFAA	34	3.80	9	3.59	64	146
Escherichia coli MG1655	52	4.61	6	4.08	80	110
Erysipelotrichaceae bacterium 21_3	46	4.83	9	1.61	124	78
Erysipelotrichaceae bacterium 6_1_45	46	4.39	8	0.55	59	141
Eubacterium sp. 3_1_31	40	3.03	7	2.21	59	118
Fusobacterium sp. 3_1_33	29	2.27	6	1.52	106	46
Fusobacterium sp. 4_1_13	31	2.20	9	0.5	280	12
Lactobacillus jensenii SJ-7A-US	37	1.71	8	0.95	71	43
Mycobacterium tuberculosis KZN MDR	66	4.28	7	2.8	234	31
Paenibacillus sp. 4_7_47FAA	43	4.15	18	1.08	96	85
Parabacteroides sp. D25	47	5.14	13	1.96	63	227
Staphylococcus aureus M0602	35	2.91	24	0.52	76	106
Streptococcus pneumoniae Tigr4	42	2.13	4	1.84	83	61
Treponema denticola F0402	40	2.72	5	2.33	50	119

Ave. scaffold N50 size: 1.72Mb +/- 1.0 Mb
Ave. contig N50 size: 106kb +/- 57kb

Assemblies versus ‘bags of genes’



berninamycin A



thiocillins

Michael Fischbach, Michael Zimmermann, Katherine Lemon, Vanja Klepac-Ceraj

How to Interpret?

16S data

WGS data

Reference Genomes

Biology

How to Interpret?

16S data

WGS data

Reference Genomes

Biology

Other ‘omes

Conclusions

- >17,000 samples, from well characterized normal subjects
- Common protocols yield consistent results
- Benchmark data accuracy, utility using positive controls
- Generated new tools and control data sets
- Defining community constituents and normal variation

HUMAN MICROBIOME PROJECT HMP

Acknowledgements

Writing Group Members

Ashlee Earl
Barb Methe
Candace Farmer
Cesar Arze
Craig Pohl
David Dooling
Dawn Ciulla
Diana Tabbaa
Dirk Gevers
Doyle Ward
Edward Belter
Elaine Mardis
Eli Venter
Elizabeth Appelbaum
Erica Sodergren
George Weinstock
Georgia Giannoukos
Hongyu Gao
Jennifer Wortman
Joseph Petrosino
Kathie Mihindukulasuriya
Kelvin Li
Kristine Wiley
Laura Courtney
Lucinda Fulton
Lynn Carmichael
Michael Feldgarden
Michelle O'Laughlin
Monika Bihan
Otis Hall
Richard Wilson
Robert Fulton
Shawn Leonard
Shibu Yooseph
Todd DeSantis
Vincent Magrini
Yanjiao Zhou

Sequencing Centers

Baylor College of Medicine
Broad Institute
J Craig Venter Institute
Washington University

Data Analysis Coordination Center

University of Maryland
Lawrence Berkeley Laboratories
University of Colorado

Institutional Support

Jane Peterson
Lita Proctor
Sue Garges
Maria Giovanni
Tsega Belachew
Shaila Chhaiba

and many more!

HUMAN
MICROBIOME
PROJECT

Detailed Conclusions

- Out of all currently recognized bacterial phyla, only a fraction (20%) found on the human body

- Body sites and sub-sites harbor distinct bacterial communities

- Body sites cluster when bacterial communities from different individuals are compared

- Clustering is driven by the presence of similar, but not identical communities

- The 'core' community at any body site is small, body site specific, and the abundance of core taxa varies greatly among individuals.

- The overwhelming majority of data belong to genera that are found (at varying frequency and abundance) at all five body sites.

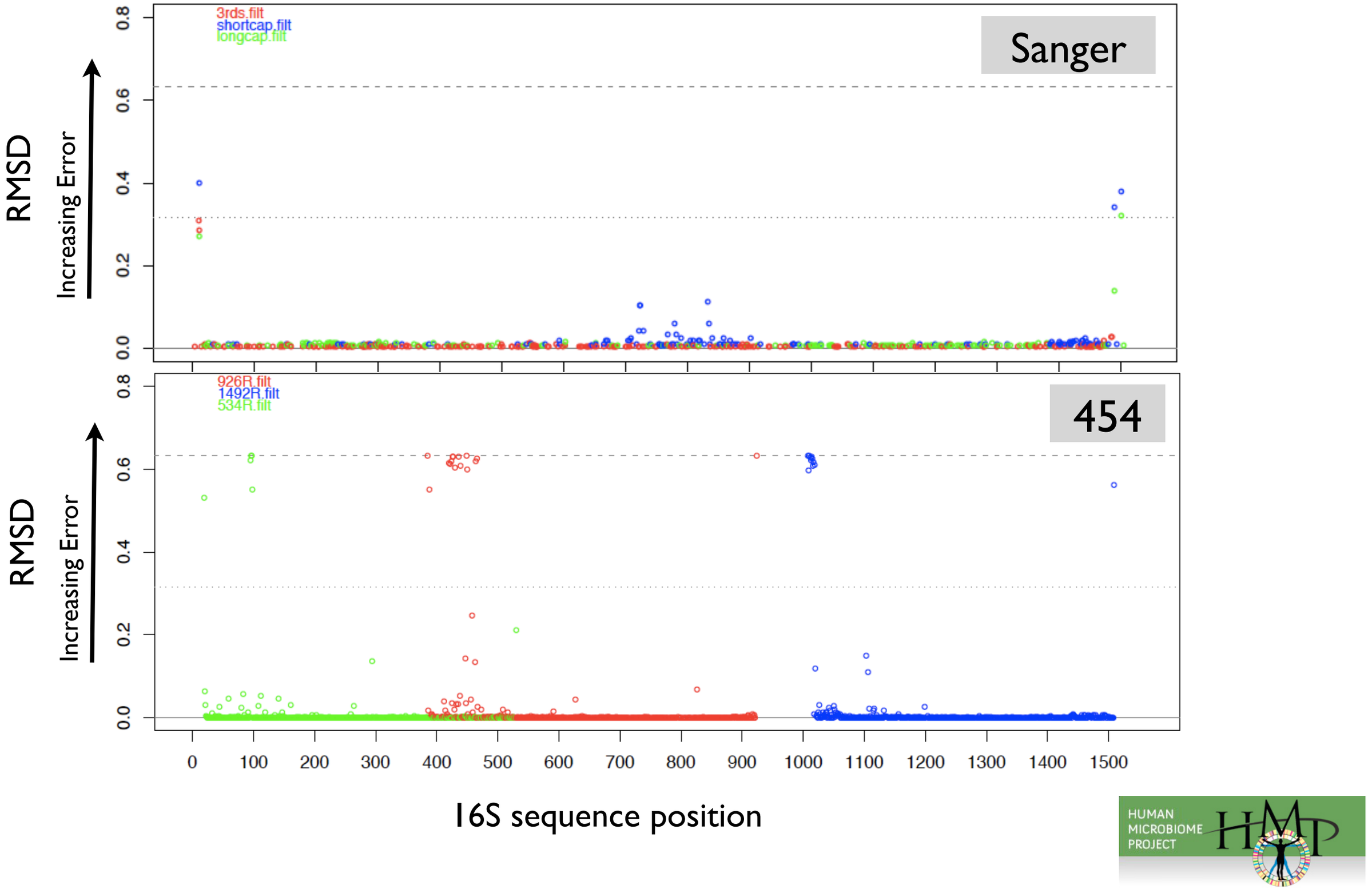
- Haven't saturated the view of biodiversity among body after sampling 24 individuals; models suggest that we will saturate for most body sites after 300 individuals are sampled.

Conclusions

- Common protocols yield consistent results!
- Indicates potential difficulties that may preclude comparing data from different studies that employ different procedures (including. extraction, PCR amplification, sequencing, processing, classification)

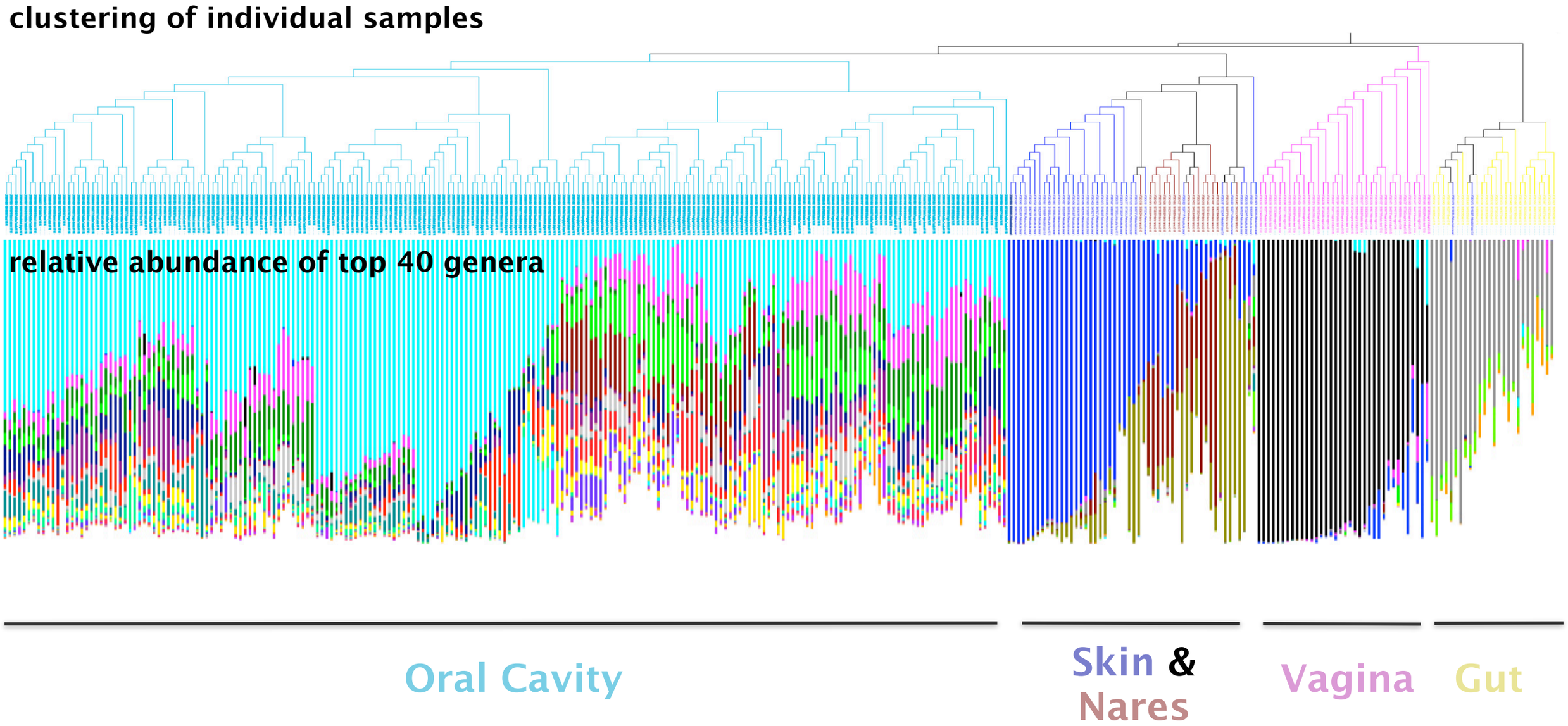
HUMAN MICROBIOME PROJECT HMP

Mapping sequencing error

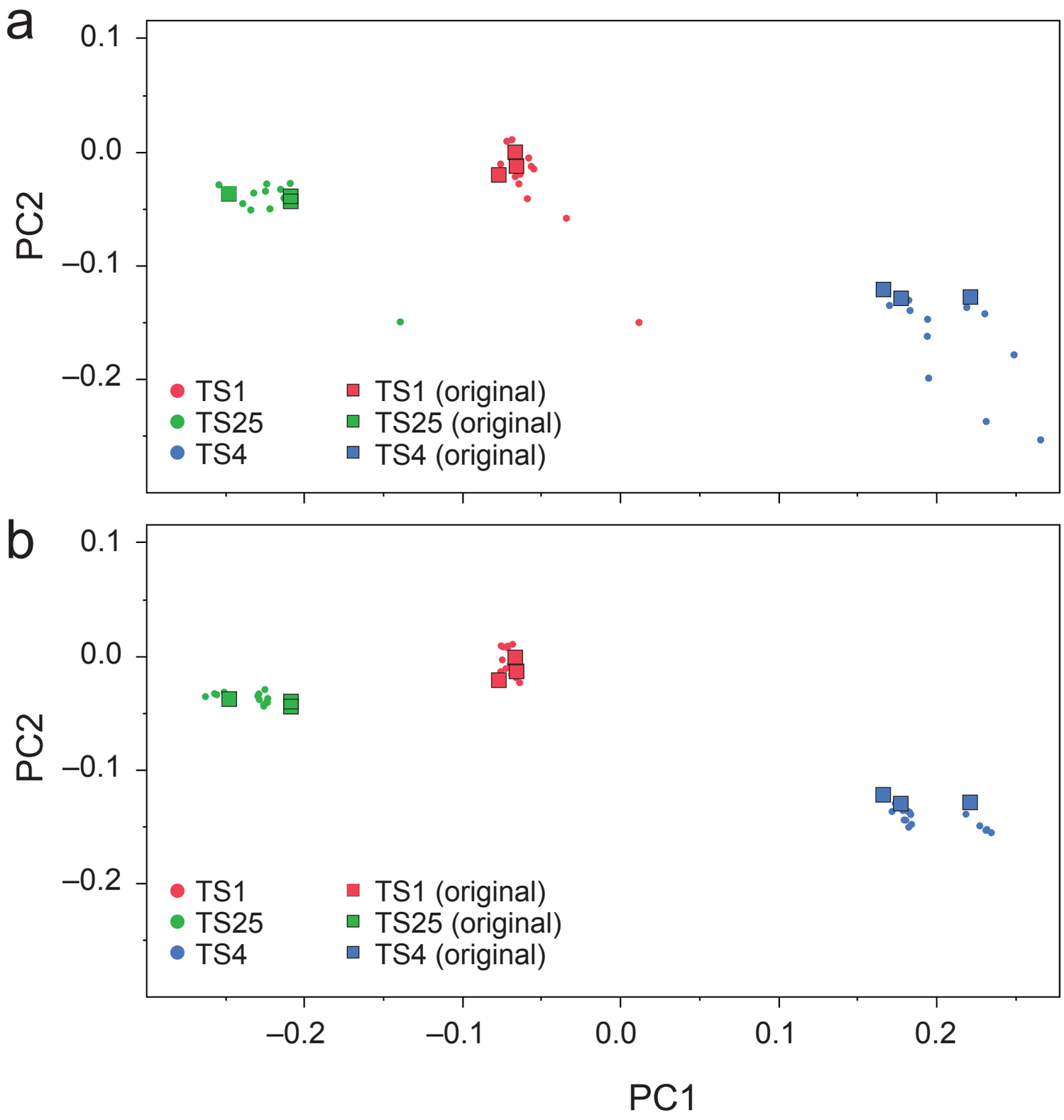


HUMAN MICROBIOME PROJECT HMP

Biogeography: Person-to-Person

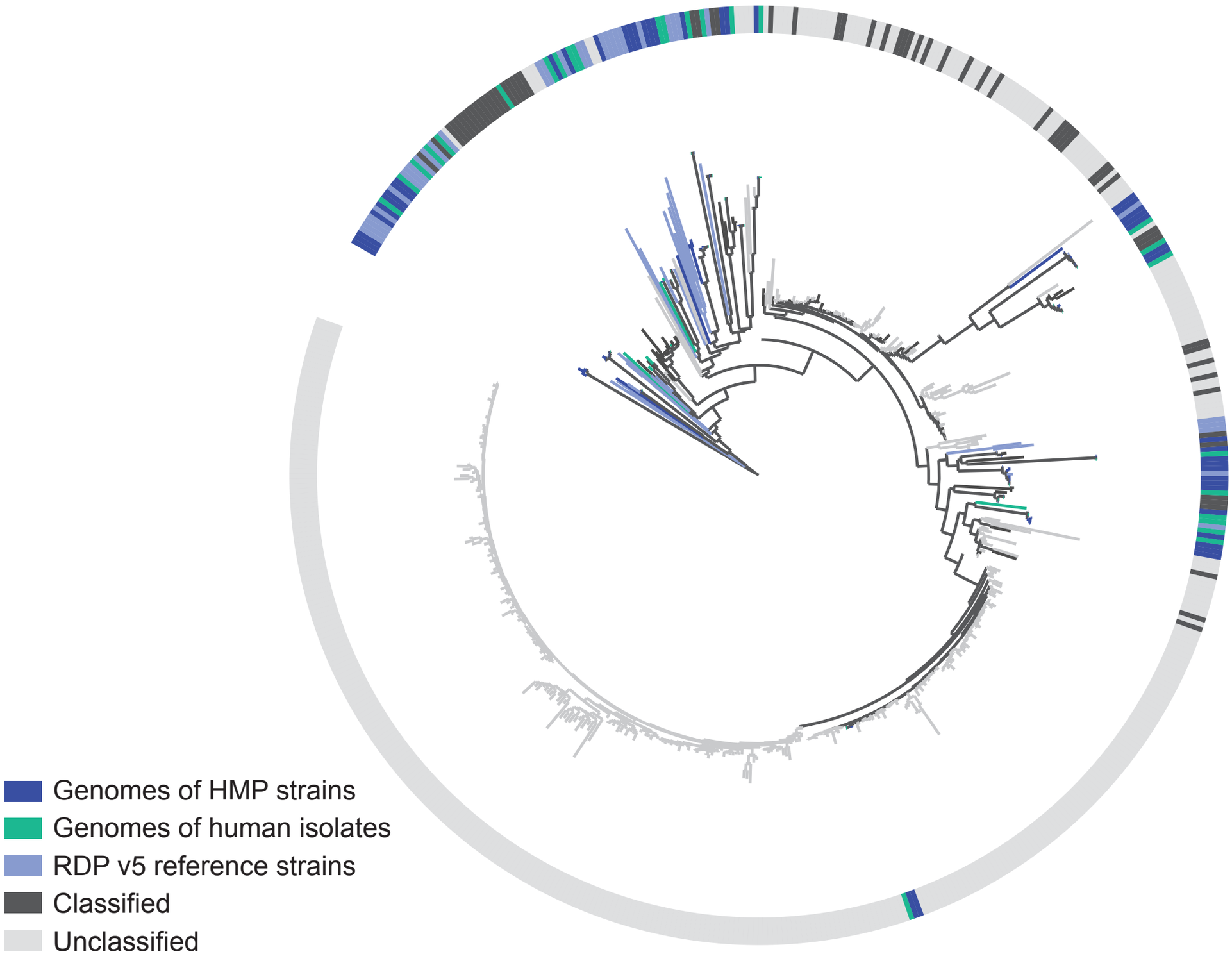


Common informatic processing reduces variation



HUMAN MICROBIOME PROJECT HMP

Phylogenetic placement of unclassified sequences



Genomes of HMP strains
Genomes of human isolates
RDP v5 reference strains
Classified
Unclassified

HUMAN MICROBIOME PROJECT HMP