Our Microbes: A Reference Genome Collection for the Human Microbiome

Ashlee Earl

on behalf of the

Human Microbiome Project Consortium

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

NATIONAL INSTITUTES OF HEALTH

HMP

Human Microbiome Project (HMP) Mission

Characterize the microbes and microbial communities that inhabit the human body

http://nihroadmap.nih.gov/hmp/
http://hmpdacc.org/

The HMP reference collection

Four Sequencing Centers

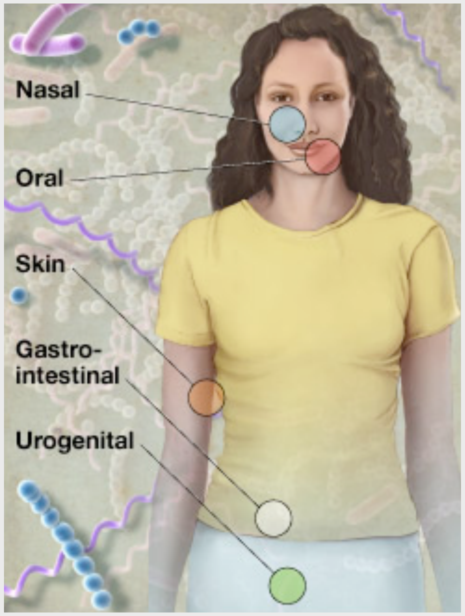
BCM
Baylor College of Medicine

J. Craig Venter™
INSTITUTE

BROAD
INSTITUTE

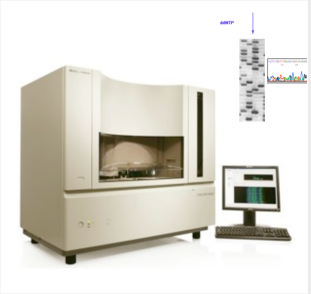
Washington
University in St.Louis

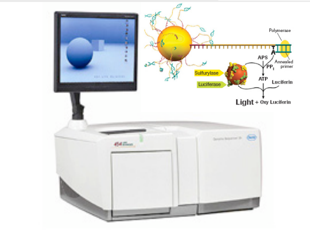
300 'Healthy' People/ 15(18) Body Sites



Evolving Technologies

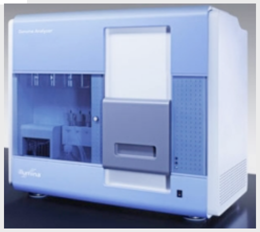
Sanger
30cm
50cm





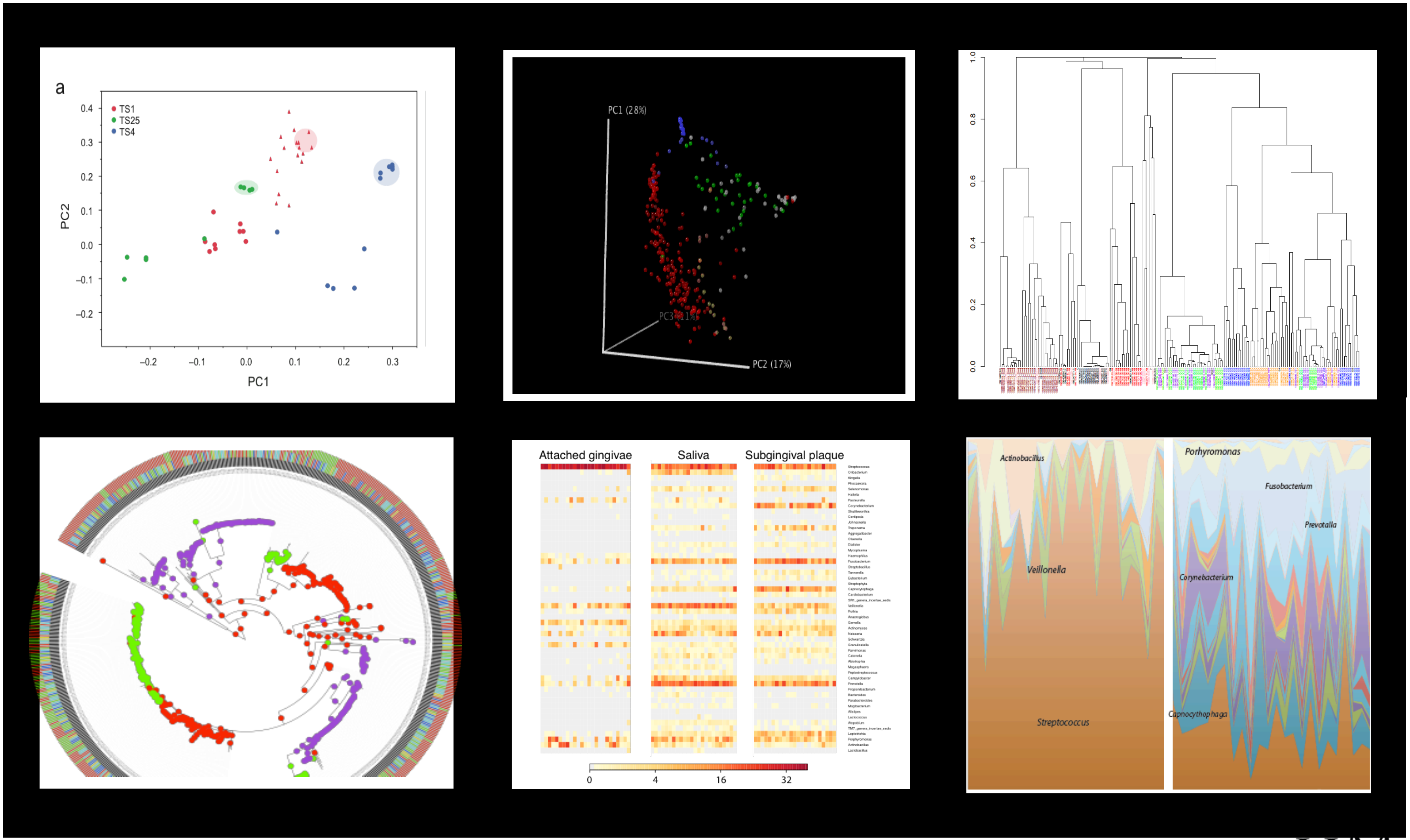
454
FLX
Titanium

illumina
GAII
HiSeq



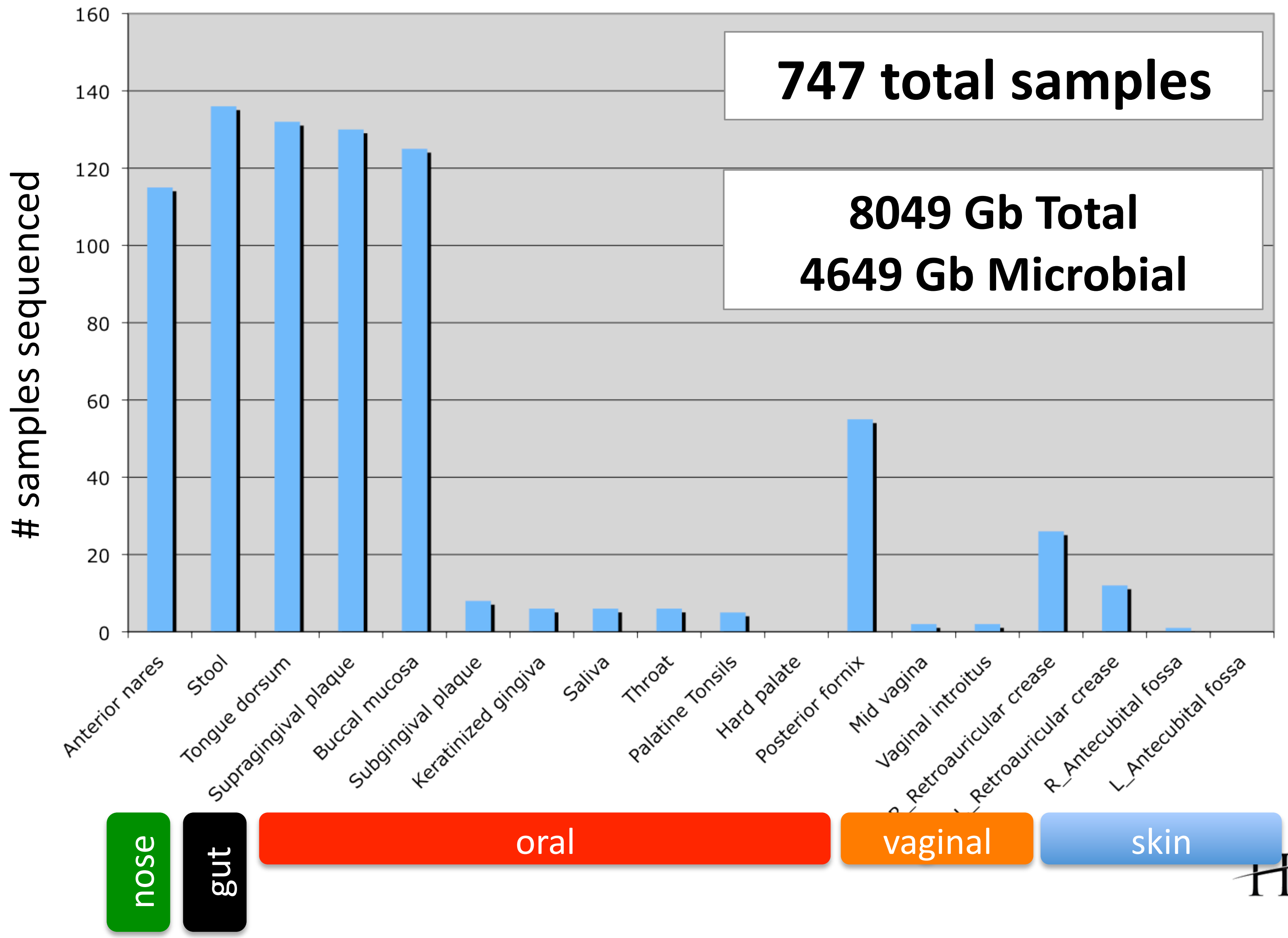
12,000 samples collected over 2 years

Data I. Pattern finding with 16S rDNA profiles



HMP

Data II. Gene finding with whole genome shotgun



747 total samples

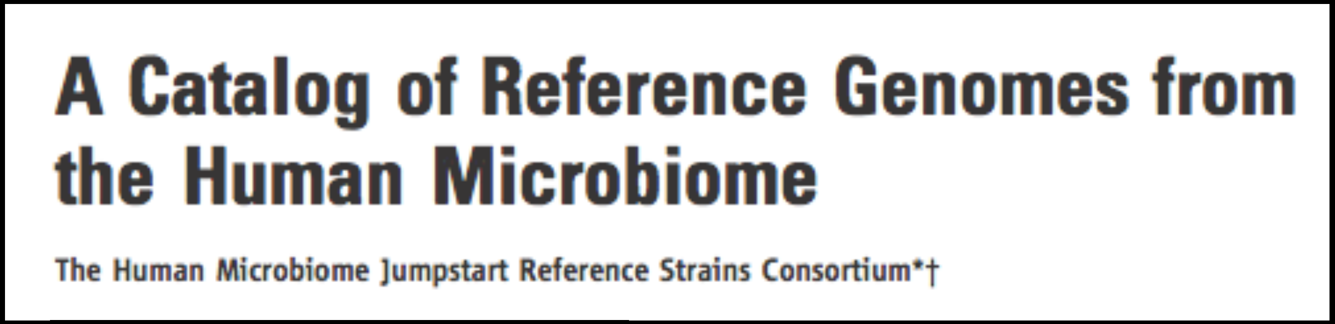
8049 Gb Total
4649 Gb Microbial

Data III. Inference finding with reference genomes

“Who’s got the teeth?”

Goal: Sequence 900 ‘reference’ bacterial strains isolated from the human body

~ Makedonka Mitreva (WashU) referring to making sense of WGS data collected from a non-gut sample

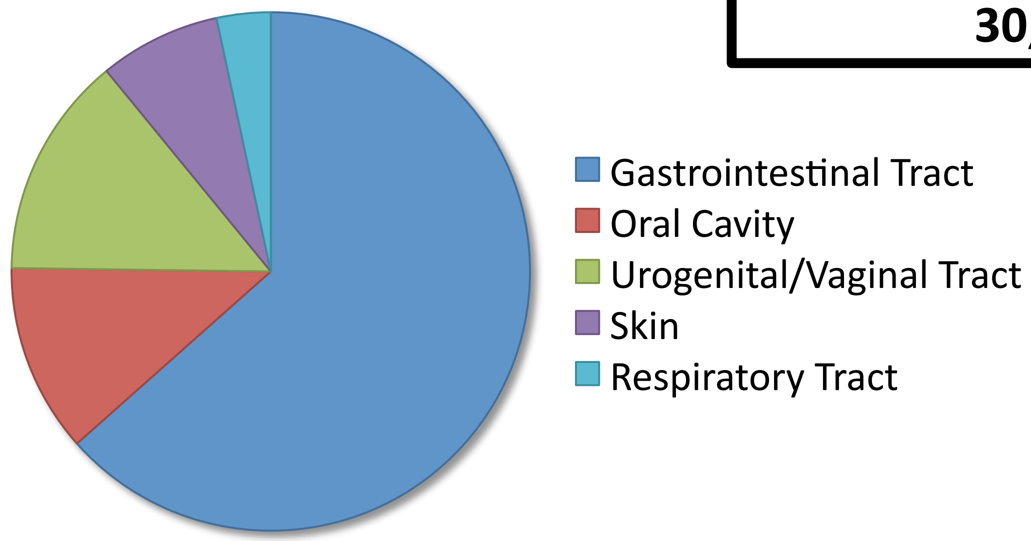


A Catalog of Reference Genomes from the Human Microbiome

The Human Microbiome Jumpstart Reference Strains Consortium*†

Science (2010) **328**:994-999

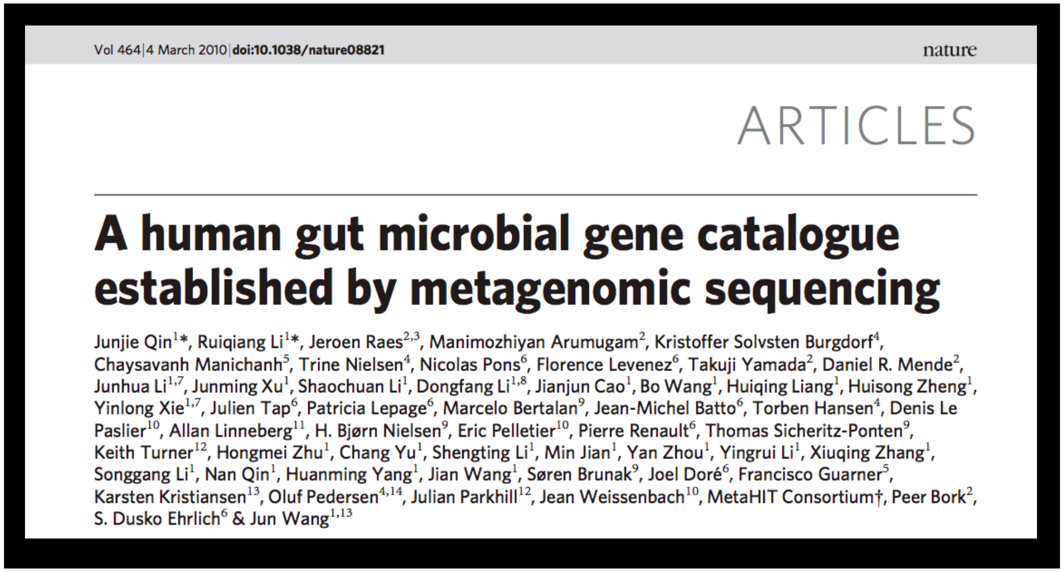
178 bacterial genomes



550,000 predicted polypeptides
30,000 novel

2 domains
9 phyla
18 classes
24 orders

HMP reference genomes at work!



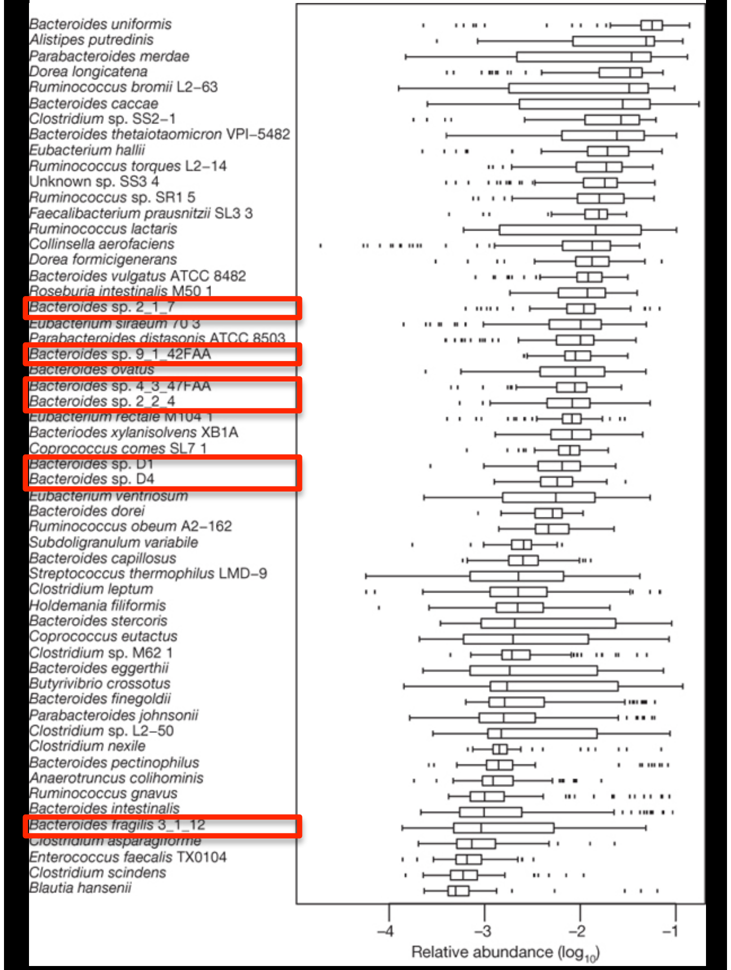
Vol 464|4 March 2010|**doi:10.1038/nature08821** nature

ARTICLES

A human gut microbial gene catalogue established by metagenomic sequencing

Junjie Qin[1]*, Ruiqiang Li[1]*, Jeroen Raes[2,3], Manimozhiyan Arumugam[2], Kristoffer Solvsten Burgdorf[4], Chaysavanh Manichanh[5], Trine Nielsen[4], Nicolas Pons[6], Florence Levenez[6], Takuji Yamada[2], Daniel R. Mende[2], Junhua Li[1,7], Junming Xu[1], Shaochuan Li[1], Dongfang Li[1,8], Jianjun Cao[1], Bo Wang[1], Huiqing Liang[1], Huisong Zheng[1], Yinlong Xie[1,7], Julien Tap[6], Patricia Lepage[6], Marcelo Bertalan[9], Jean-Michel Batto[6], Torben Hansen[4], Denis Le Paslier[10], Allan Linneberg[11], H. Bjørn Nielsen[9], Eric Pelletier[10], Pierre Renault[6], Thomas Sicheritz-Ponten[9], Keith Turner[12], Hongmei Zhu[1], Chang Yu[1], Shengting Li[1], Min Jian[1], Yan Zhou[1], Yingrui Li[1], Xiuqing Zhang[1], Songgang Li[1], Nan Qin[1], Huanming Yang[1], Jian Wang[1], Søren Brunak[9], Joel Doré[6], Francisco Guarner[5], Karsten Kristiansen[13], Oluf Pedersen[4,14], Julian Parkhill[12], Jean Weissenbach[10], MetaHIT Consortium†, Peer Bork[2], S. Dusko Ehrlich[6] & Jun Wang[1,13]

Relative abundance of 57 frequent microbial genomes among individuals of the cohort.

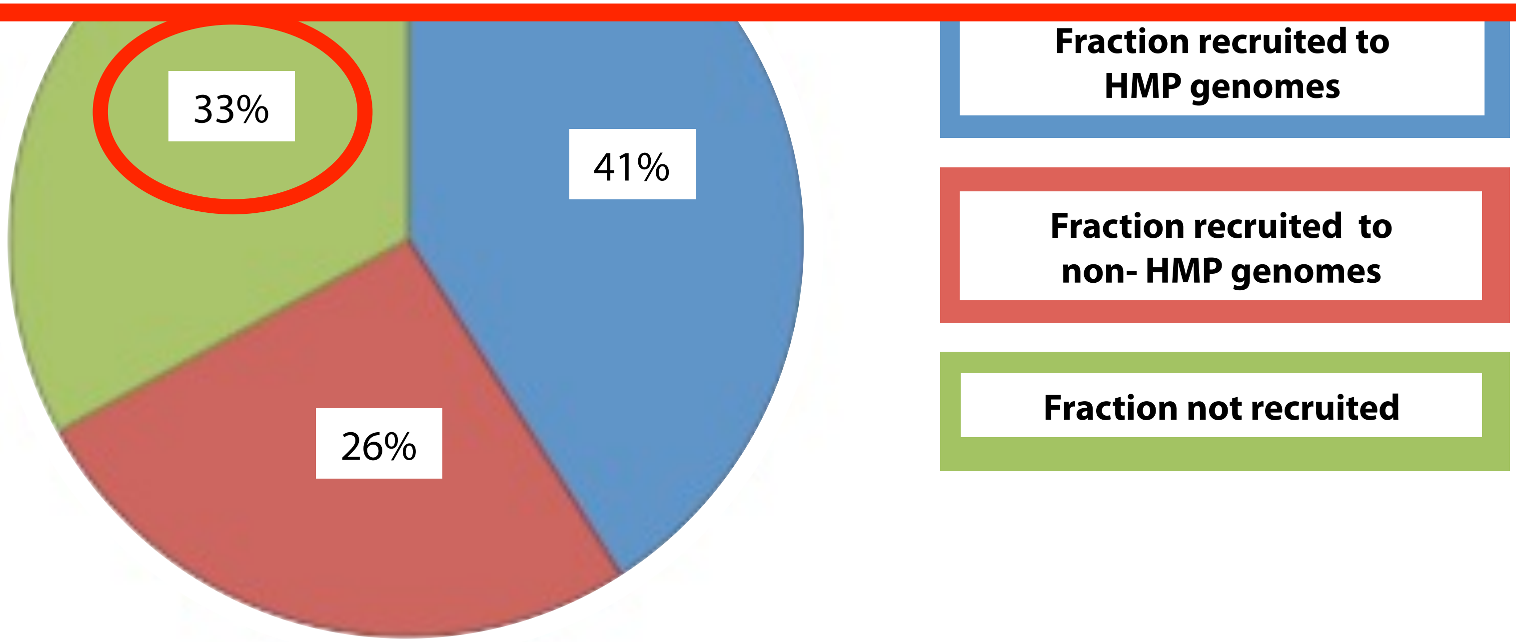


JJ Qin *et al. Nature* **464**, 59-65 (2010) doi:10.1038/nature08821

nature

HMP reference genomes at work

Current goal: Sequence 3,000 ‘reference’ bacterial strains isolated from the human body

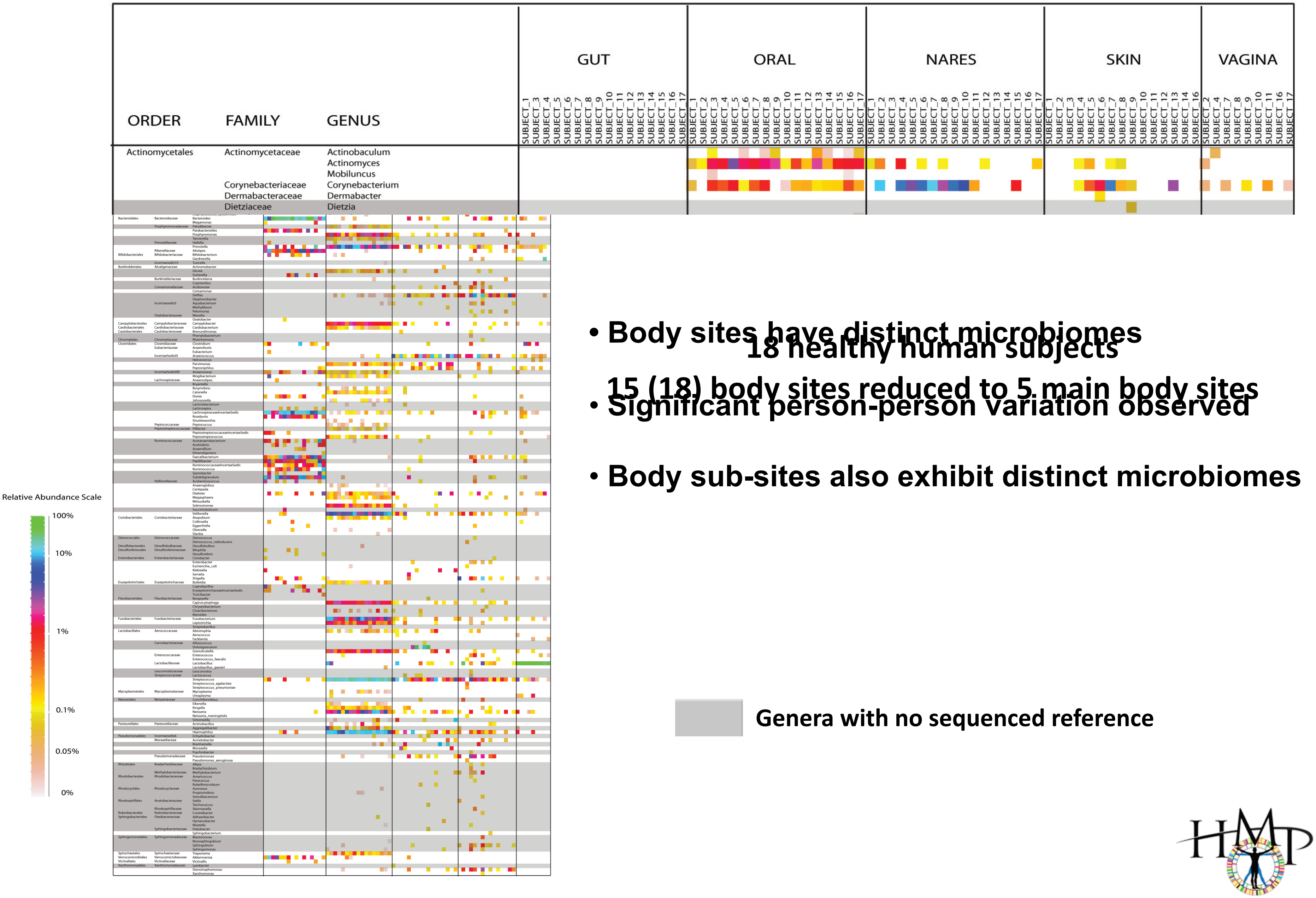


Who’s next?

1) unknown species, novel phylogenic position;
2) strains implicated in a disease process;
3) previously known species, not yet sequenced;
4) known species, with previous genomes sequenced, but showing high sequence diversity and/or significant variation in phenotypic profiles.

Let the 16S rDNA profiles be our guide

16S rDNA profiles from 1st 18 subjects



- **Body sites have distinct microbiomes**

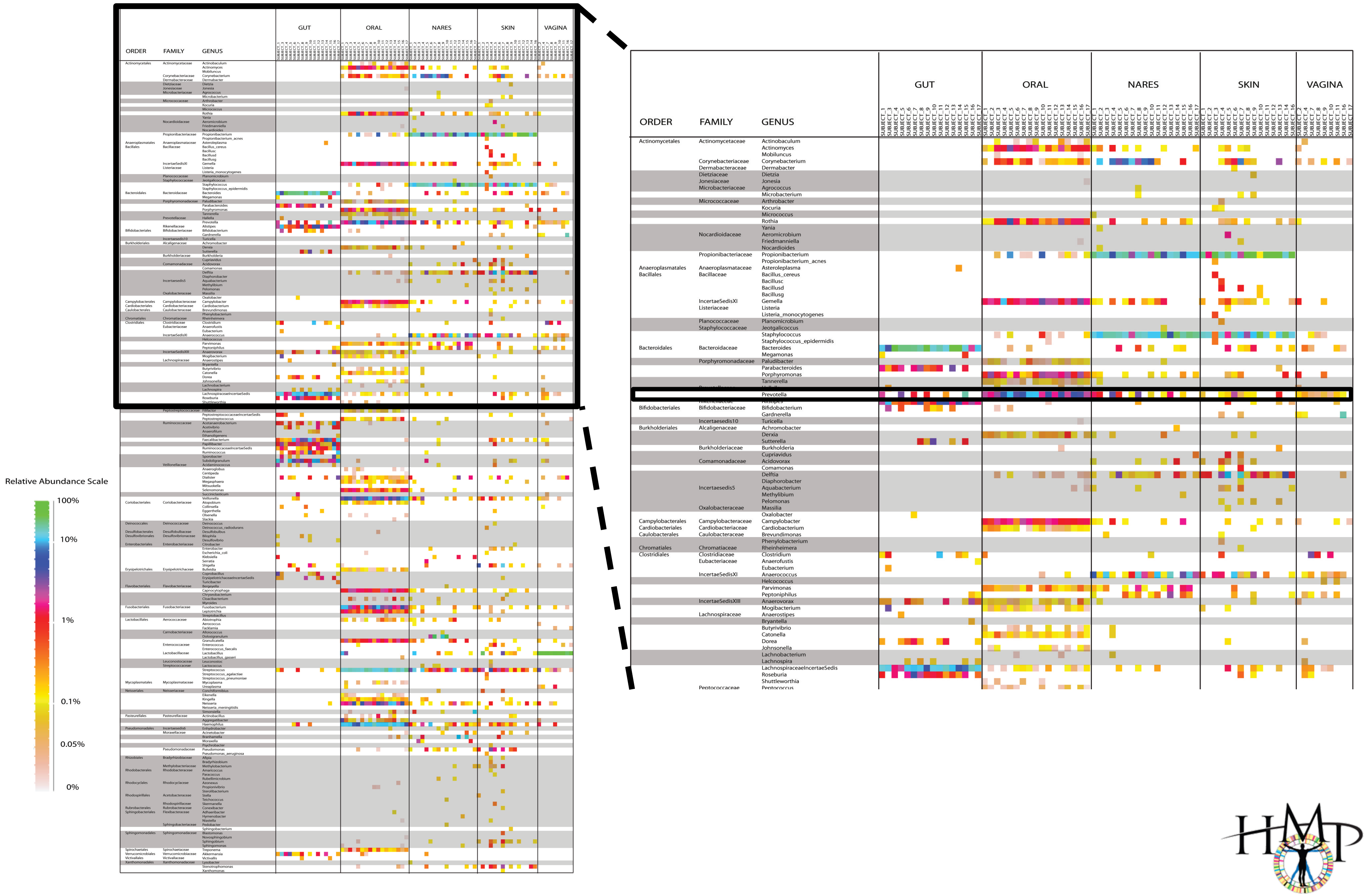
18 healthy human subjects

15 (18) body sites reduced to 5 main body sites

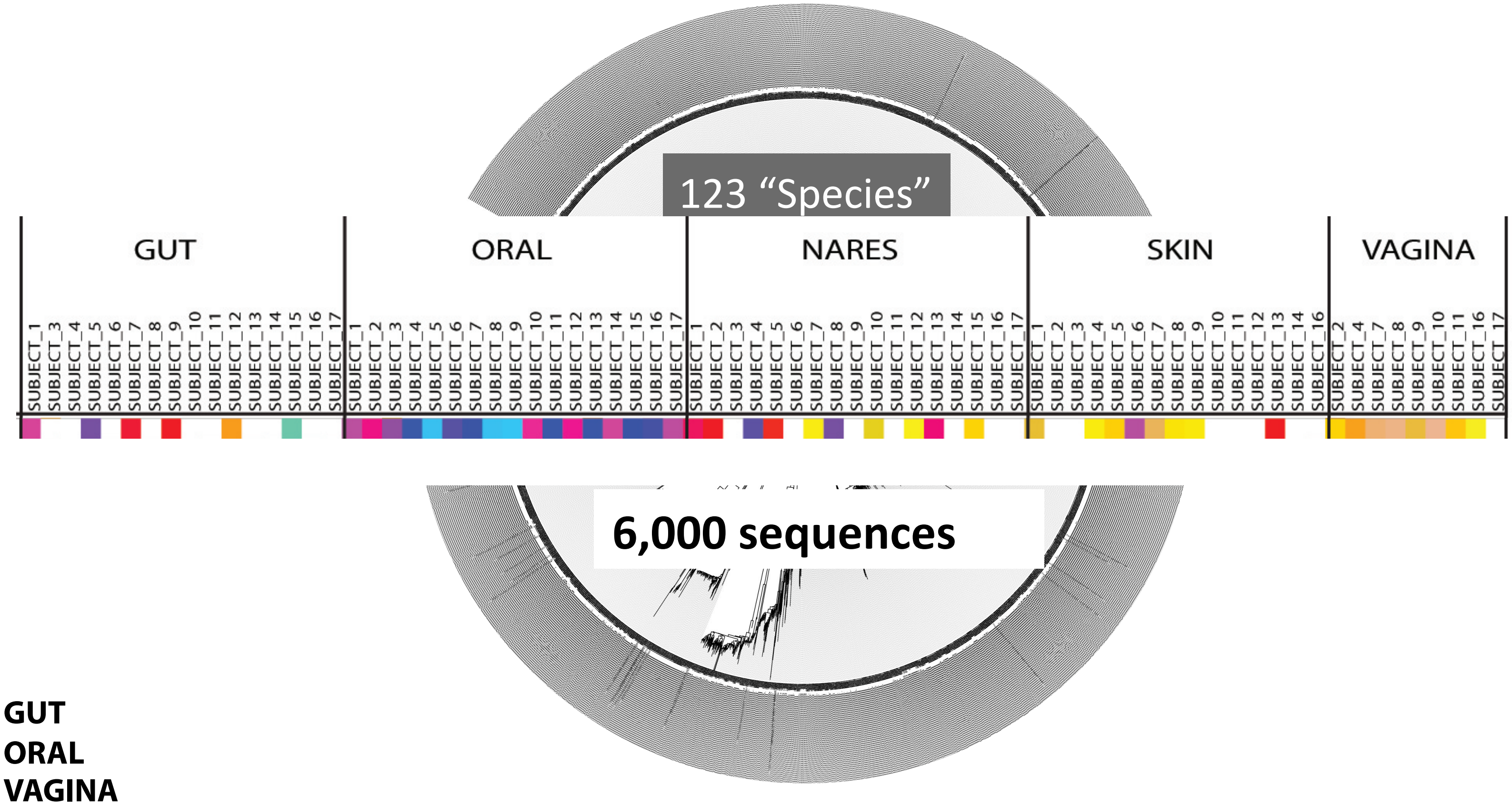
- **Significant person-person variation observed**
- **Body sub-sites also exhibit distinct microbiomes**

▨ **Genera with no sequenced reference**

What about diversity beyond the level of the genus?



How diverse are these *Prevotella*?

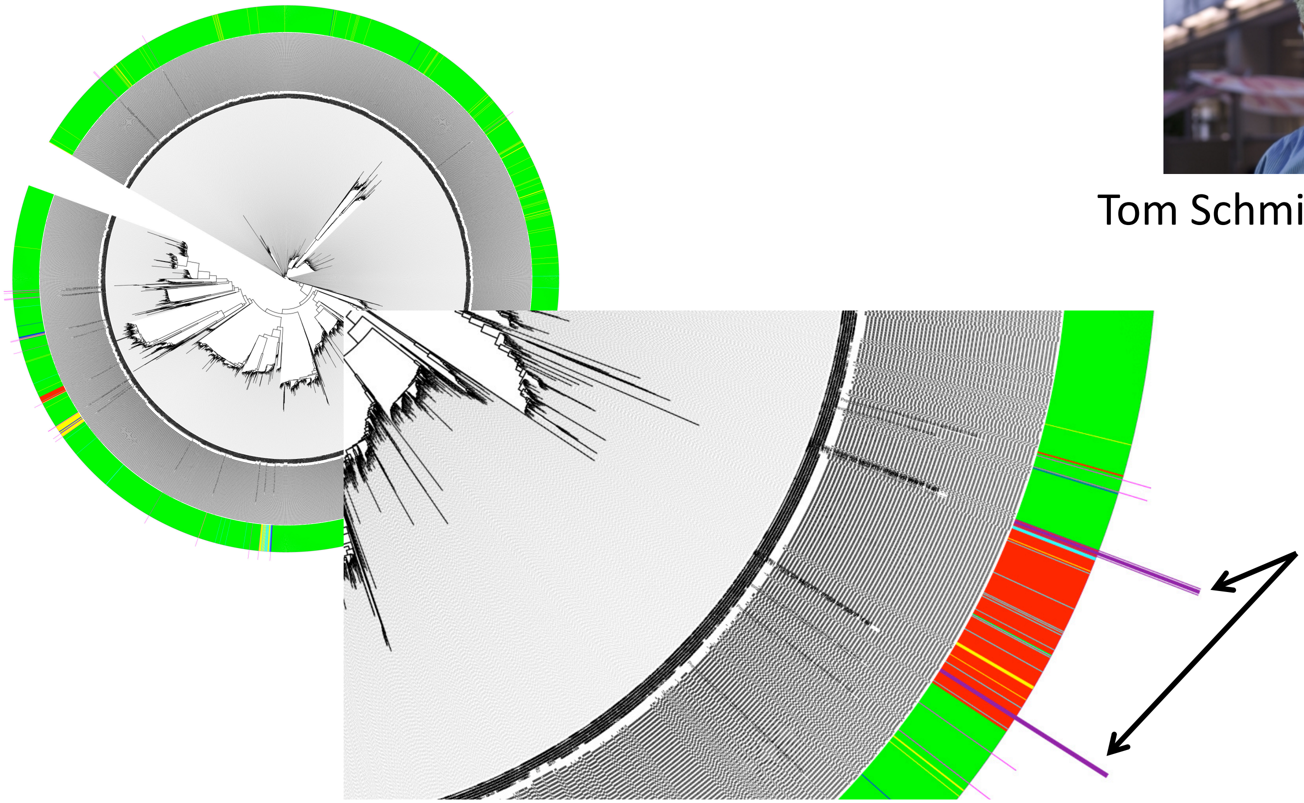


123 “Species”

6,000 sequences

🟥 **GUT**
🟩 **ORAL**
🟪 **VAGINA**
🟨 **NARES**
🟦 **SKIN**
🟪 **SEQUENCED GENOMES**

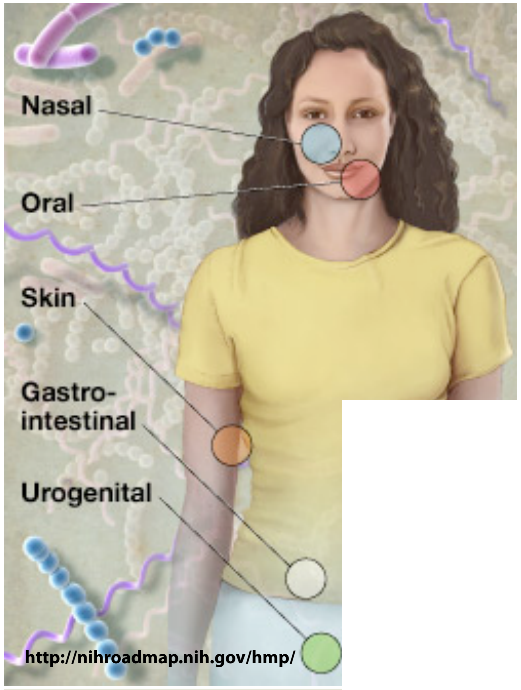
The community makes this work



Tom Schmidt, Michigan State

16S sequences from Tom's gut isolate collection!

The community is critical



nasal (airways)

Mike Surette
U of Calgary
Nathan Weyland, Maggie So
Kimberlee Musser

oral

Jacques Izard
Floyd Dewhirst
Forsyth Institute
Slava Epstein
Northeastern
Mogens Kilian
Sue Haake

YOU??

http://www.hmpdacc.org/outreach_feedback.php

gastrointestinal

Tom Schmidt, U of Michigan
Delphine Saulnier, Baylor
Barbara Murray, Cesar Arias
Cherie Ziemer
Edgar Boedeker
Glenn Songer
Harry Flint
Masami Morotomi

urogenital

Qiang Xu, Osel Inc.
Mario Vaneechoutte,
U of Ghent
Sharon Hillier

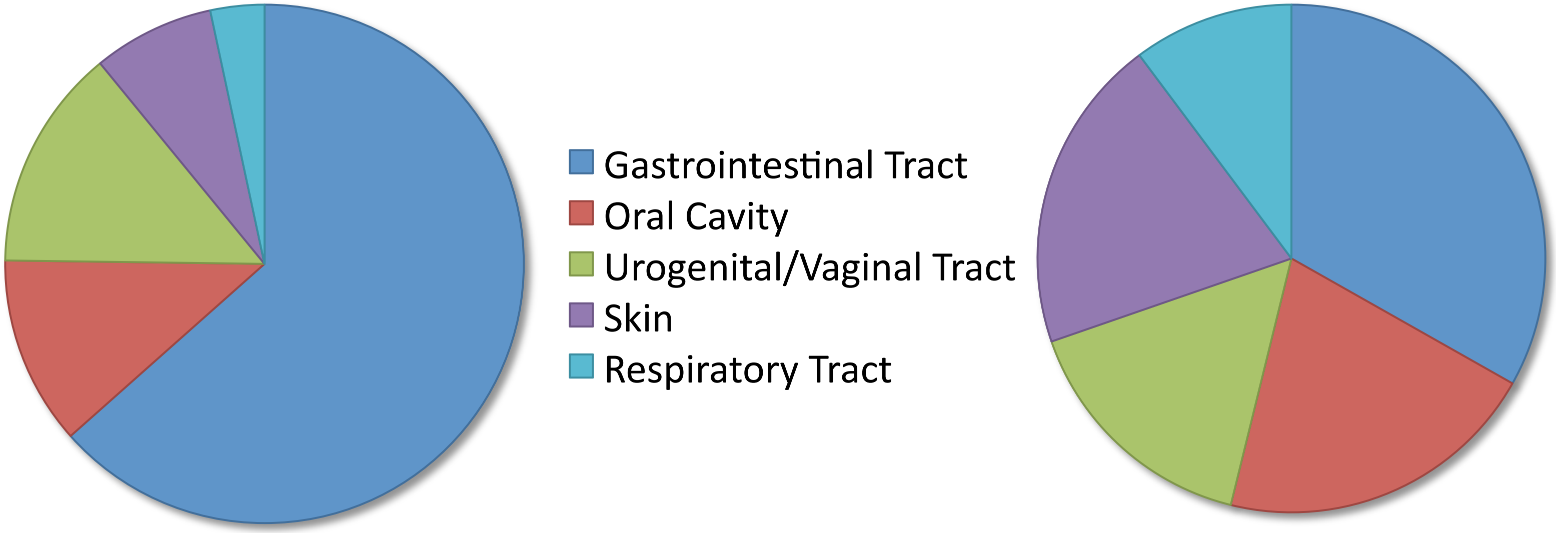
Rapid progress

THEN

178 bacterial genomes

NOW

1419 bacterial genomes



What else?

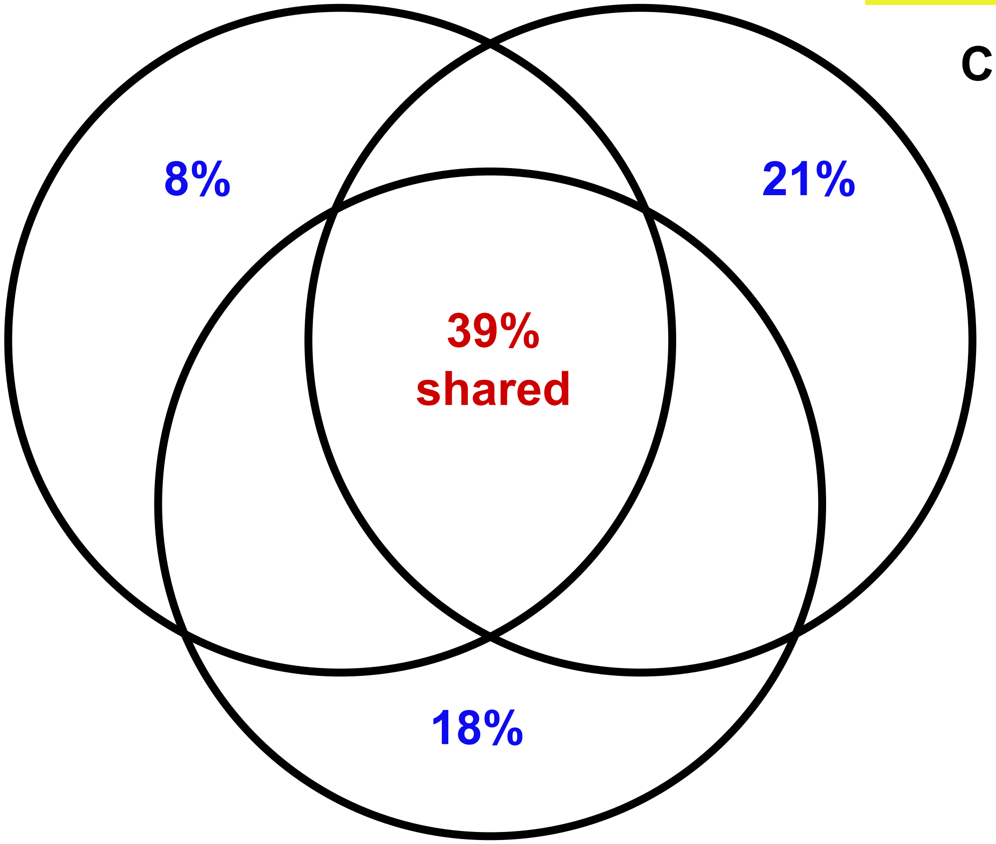
Escherichia coli and the pan-genome

non-pathogenic

MG1655 (K-12)

uropathogenic

CFT073



8%

21%

39%
shared

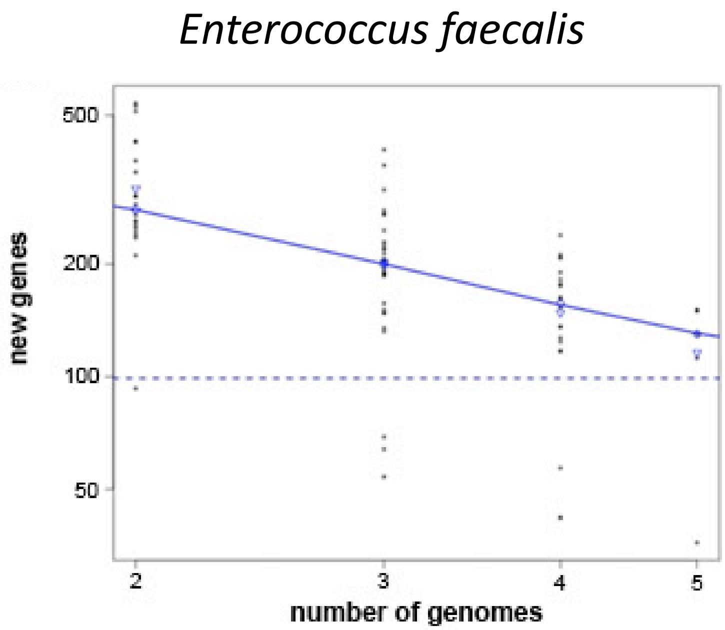
18%

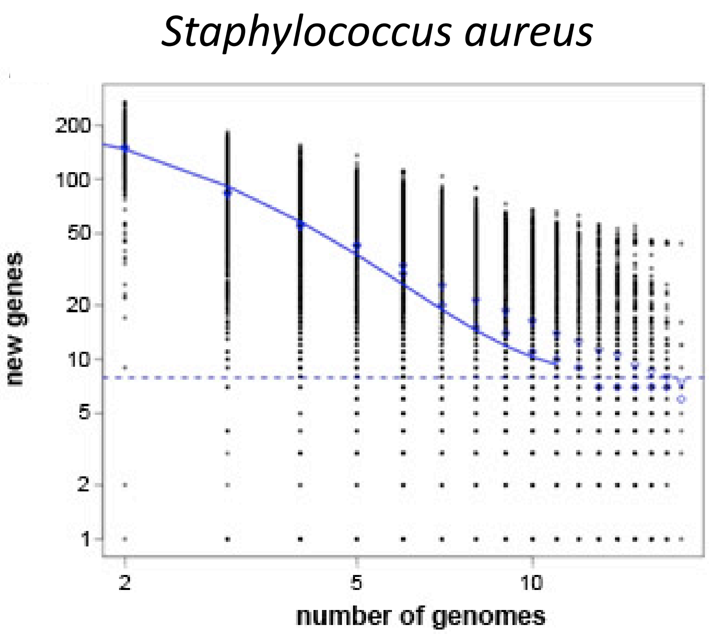
EDL933

enterohaemorrhagic

Welch, R. A. et al. (2002) *PNAS* 99:17020-17024

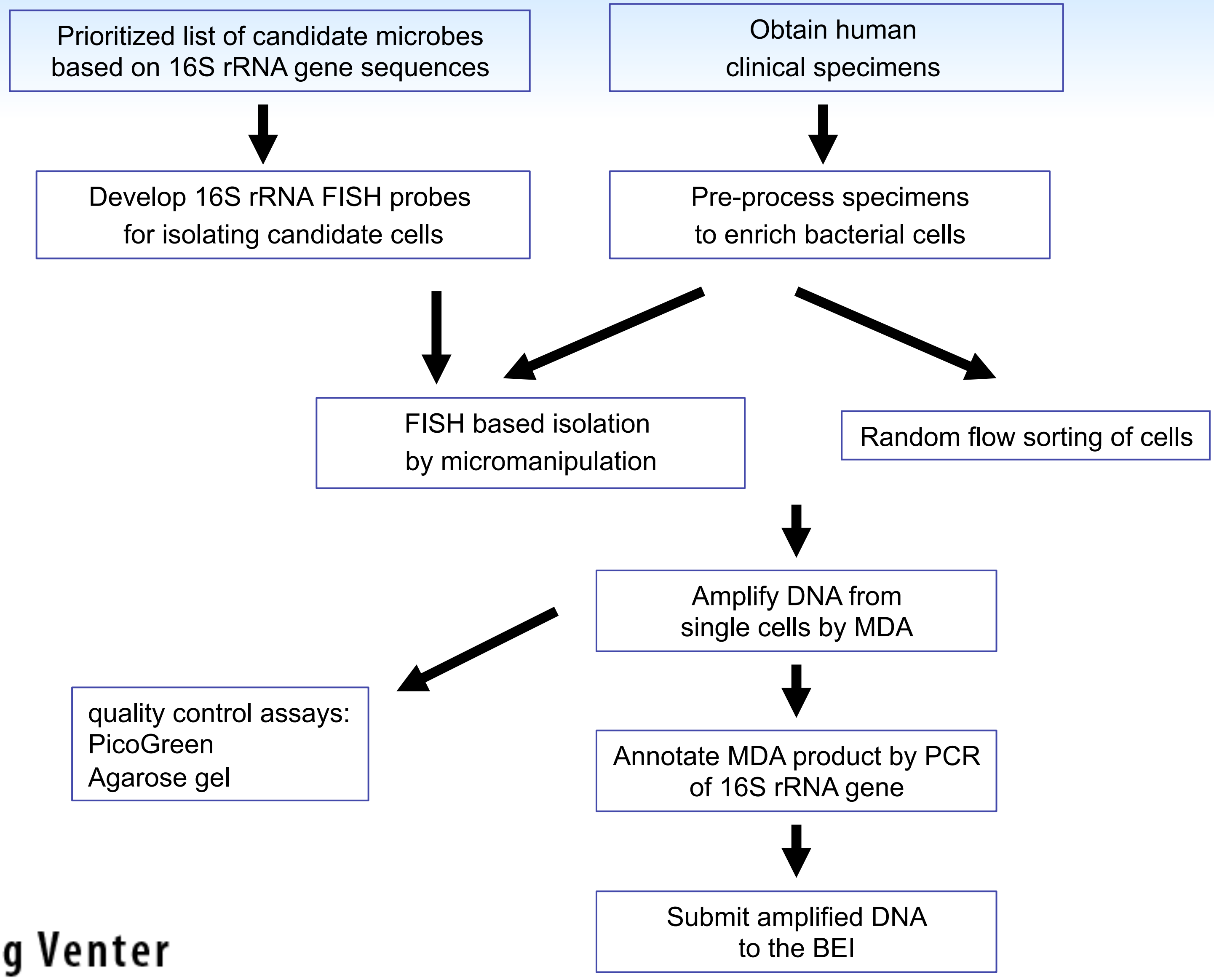
Assessing the value of more genomes


Enterococcus faecalis


Staphylococcus aureus

What about the ‘unculturables’?

Pipeline for obtaining reference genomes from uncultivated microbes for the HMP



Prioritized list of candidate microbes based on 16S rRNA gene sequences

Obtain human clinical specimens

Develop 16S rRNA FISH probes for isolating candidate cells

Pre-process specimens to enrich bacterial cells

FISH based isolation by micromanipulation

Random flow sorting of cells

Amplify DNA from single cells by MDA

quality control assays:
PicoGreen
Agarose gel

Annotate MDA product by PCR of 16S rRNA gene

Submit amplified DNA to the BEI

J. Craig Venter INSTITUTE

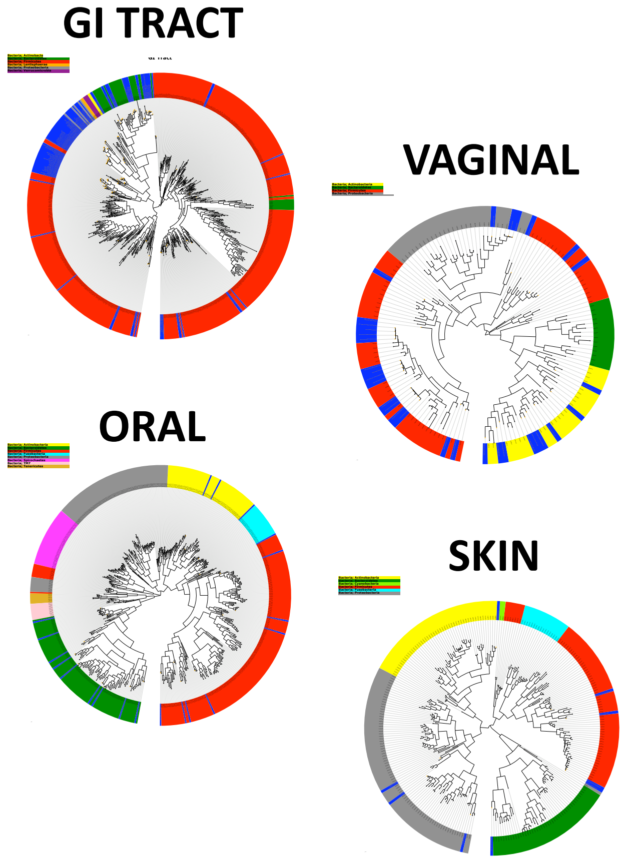
Roger Lasken, JCVI

HMP Technology Development Initiative

PI Name	Institution	Title
ALLEN-VERCOE, EMMA	UNIVERSITY OF GUELPH	FACS-MABE: a method to sort and enrich the as-yet uncultured bacterial species from the human distal gut
BRADBURY, ANDREW & HAN, CLIFF	LOS ALAMOS NAT SECTY-LOS ALAMOS NAT LAB	Species-by-Species Dissection of Microbiomes using Phage Display and Flow Sorting
DAVIS, RONALD WAYNE	STANFORD UNIVERSITY	Isolation, selection, and polony amplification of single cells in a gel matrix
DOKTYCZ, MITCHEL	UT-BATTELLE, LLC-OAK RIDGE NATIONAL LAB	Functional Sorting of Microbial Cells From Complex Microbiota
FREDRICKS, DAVID NEAL	FRED HUTCHINSON CANCER RESEARCH CENTER	Novel cultivation methods for the domestication of vaginal bacteria
ISMAGILOV, RUSTEM F	UNIVERSITY OF CHICAGO	Confining single cells to enhance and target cultivation of human microbiome
LEWIS, KIM	NORTHEASTERN UNIVERSITY	Culturing uncultivatable gut microorganisms
NELSON, JOHN	GENERAL ELECTRIC GLOBAL RESEARCH CTR	Tools for human microbiome studies
PODAR, MIRCEA	UT-BATTELLE, LLC-OAK RIDGE NATIONAL LAB	Targeted genomic characterization of uncultured bacteria from the human microbiota
RELMAN, DAVID A	STANFORD UNIVERSITY	Optimization of a microfluidic device for single bacterial cell genomics
SCHMIDT, THOMAS MITCHELL & YOUNG, VINCENT	MICHIGAN STATE UNIVERSITY	Cultivation and Characterization of Microaerobes from the Human Microbiome
SINGH, ANUP K	SANDIA CORP-SANDIA NATIONAL LABORATORIES	FISH 'N' Chips: A Microfluidic Processor for Isolating and Analyzing Microbes
WORTHEN G. SCOTT	CHILDREN'S HOSPITAL OF PHILADELPHIA	Multi-Dimensional Separation of Bacteria
ZHANG, KUN & LO, YU-HWA	UNIVERSITY OF CALIFORNIA SAN DIEGO	An Integrated lab-on-chip system for genome sequencing of single microbial cells

100 most wanted ‘uncultured’ bacteria!

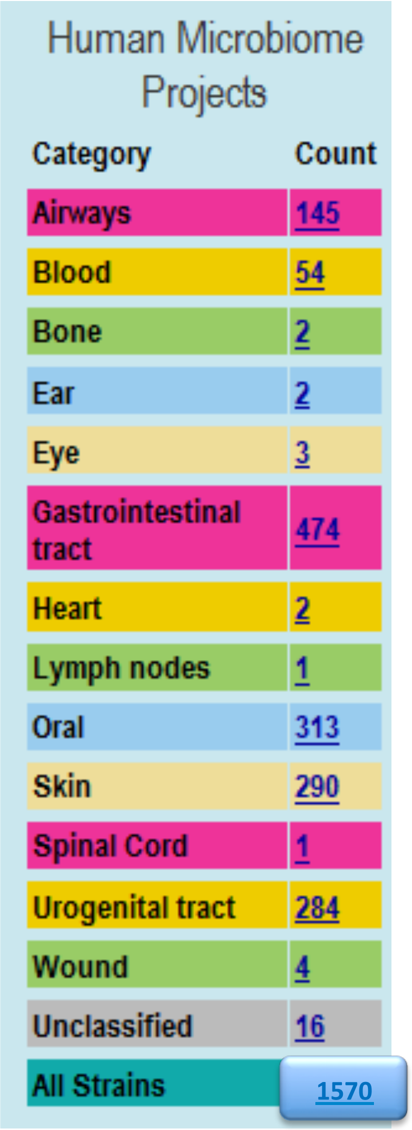
Community profiles drive our hunt



What about eukaryotes? Viruses??

Yes and yes.

Track progress, access data, collaborate, analyze…



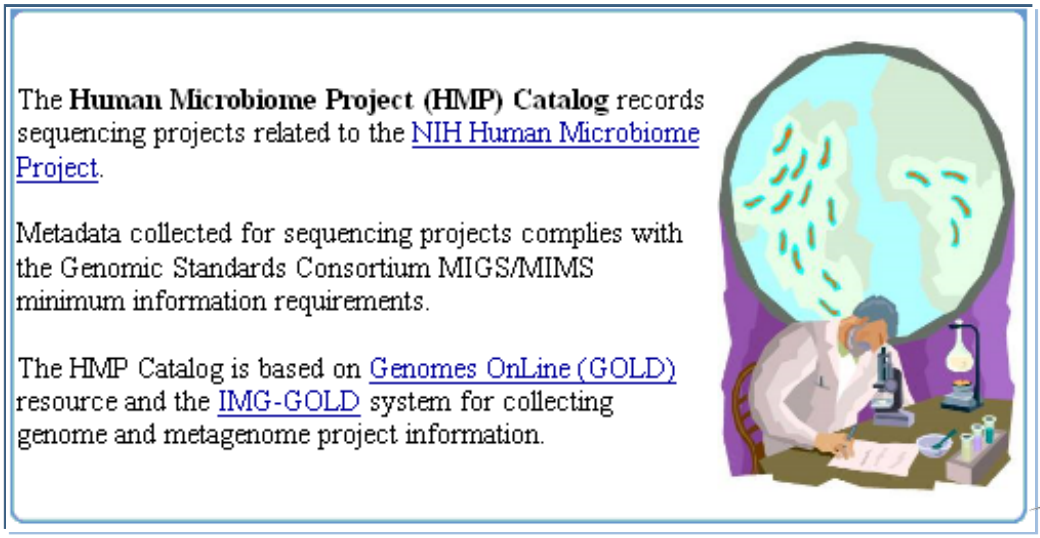
Human Microbiome Projects

Category	Count
Airways	145
Blood	54
Bone	2
Ear	2
Eye	3
Gastrointestinal tract	474
Heart	2
Lymph nodes	1
Oral	313
Skin	290
Spinal Cord	1
Urogenital tract	284
Wound	4
Unclassified	16
All Strains	1570

http://hmpdacc.org/

HMP Project Catalog http://hmpdacc.org/

- Relational data model
- Tracks project status
- Stores comprehensive metadata
- Links to public data resources
- Provides search/filtering options

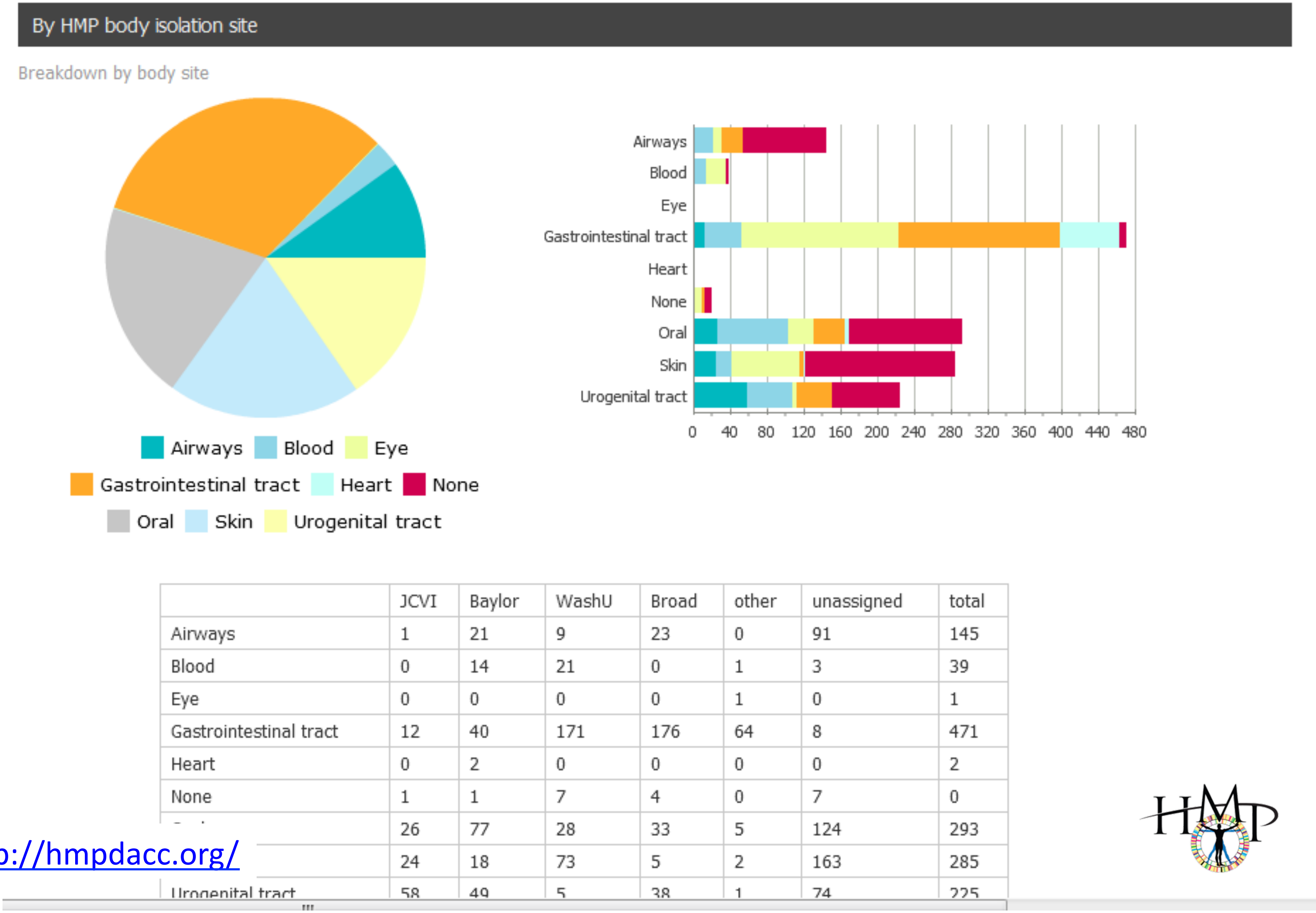


The **Human Microbiome Project (HMP) Catalog** records sequencing projects related to the NIH Human Microbiome Project.

Metadata collected for sequencing projects complies with the Genomic Standards Consortium MIGS/MIMS minimum information requirements.

The HMP Catalog is based on Genomes OnLine (GOLD) resource and the IMG-GOLD system for collecting genome and metagenome project information.

Keep track of our progress



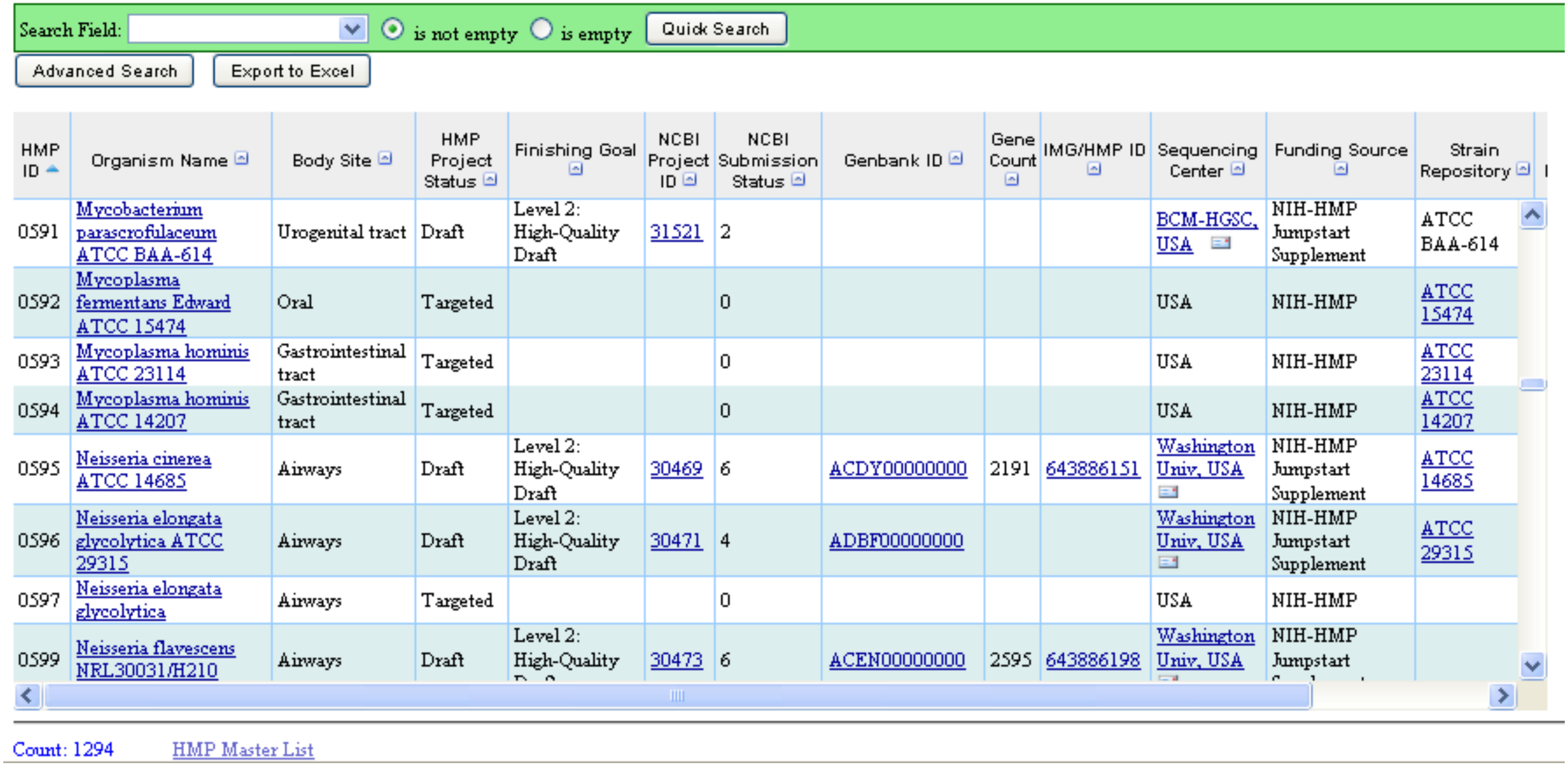
By HMP body isolation site

Breakdown by body site

	JCVI	Baylor	WashU	Broad	other	unassigned	total
Airways	1	21	9	23	0	91	145
Blood	0	14	21	0	1	3	39
Eye	0	0	0	0	1	0	1
Gastrointestinal tract	12	40	171	176	64	8	471
Heart	0	2	0	0	0	0	2
None	1	1	7	4	0	7	0
	26	77	28	33	5	124	293
	24	18	73	5	2	163	285
Urogenital tract	58	49	5	38	1	74	225

http://hmpdacc.org/

Keep track of our progress



Search Field: ● is not empty ○ is empty

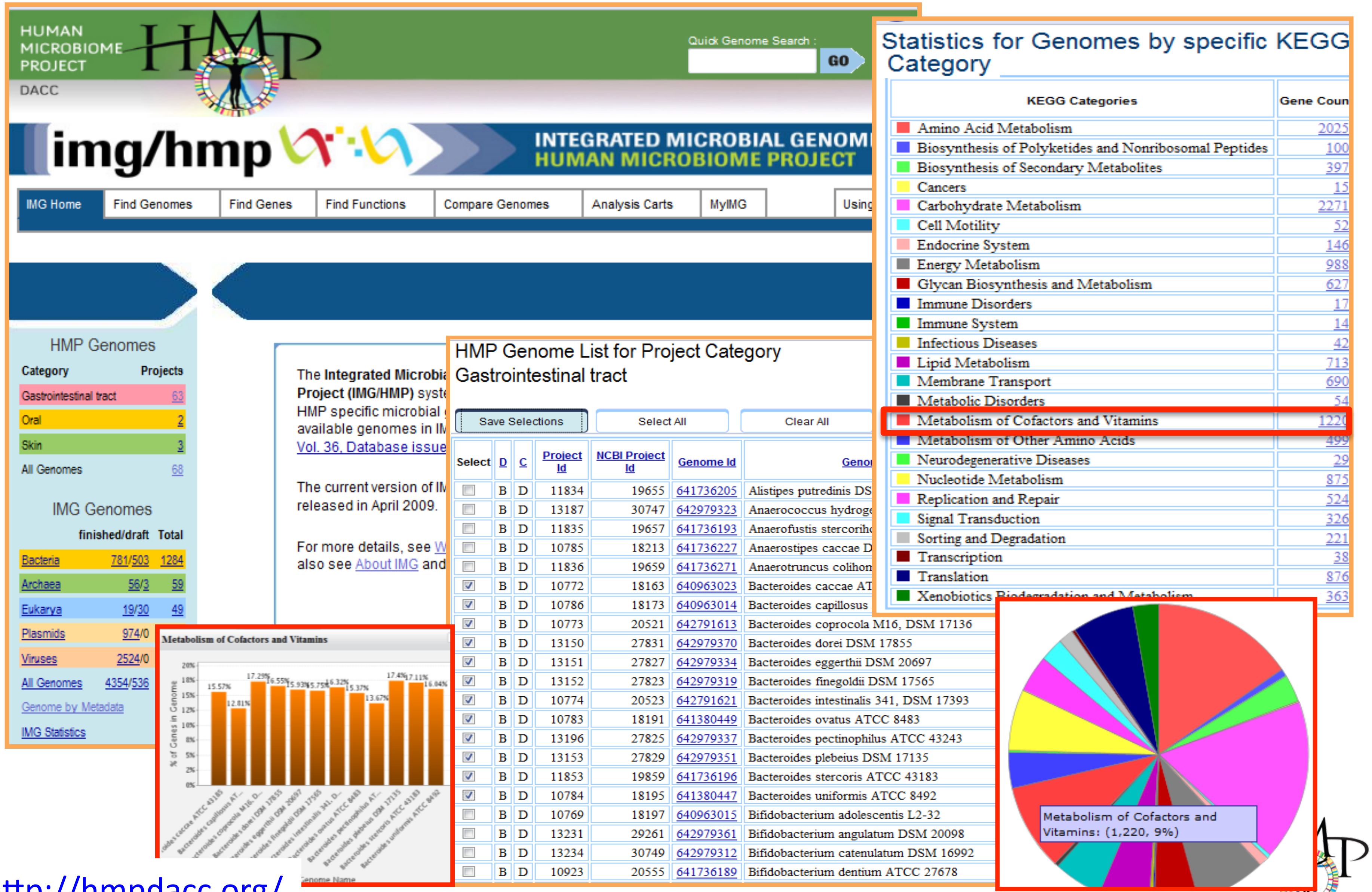
HMP ID	Organism Name	Body Site	HMP Project Status	Finishing Goal	NCBI Project ID	NCBI Submission Status	Genbank ID	Gene Count	IMG/HMP ID	Sequencing Center	Funding Source	Strain Repository
0591	Mycobacterium parascrofulaceum ATCC BAA-614	Urogenital tract	Draft	Level 2: High-Quality Draft	31521	2				BCM-HGSC, USA	NIH-HMP Jumpstart Supplement	ATCC BAA-614
0592	Mycoplasma fermentans Edward ATCC 15474	Oral	Targeted			0				USA	NIH-HMP	ATCC 15474
0593	Mycoplasma hominis ATCC 23114	Gastrointestinal tract	Targeted			0				USA	NIH-HMP	ATCC 23114
0594	Mycoplasma hominis ATCC 14207	Gastrointestinal tract	Targeted			0				USA	NIH-HMP	ATCC 14207
0595	Neisseria cinerea ATCC 14685	Airways	Draft	Level 2: High-Quality Draft	30469	6	ACDY00000000	2191	643886151	Washington Univ, USA	NIH-HMP Jumpstart Supplement	ATCC 14685
0596	Neisseria elongata glycolytica ATCC 29315	Airways	Draft	Level 2: High-Quality Draft	30471	4	ADBF00000000			Washington Univ, USA	NIH-HMP Jumpstart Supplement	ATCC 29315
0597	Neisseria elongata glycolytica	Airways	Targeted			0				USA	NIH-HMP	
0599	Neisseria flavescens NRL30031/H210	Airways	Draft	Level 2: High-Quality	30473	6	ACEN00000000	2595	643886198	Washington Univ, USA	NIH-HMP Jumpstart	

Count: 1294 HMP Master List

Contains a complete list of all Reference Strains along with detailed metadata about each. Provides both “quick” and “advanced” search and download options.

http://hmpdacc.org/

Genome analysis at IMG



HMP Genomes

Category	Projects
Gastrointestinal tract	63
Oral	2
Skin	3
All Genomes	68

IMG Genomes

	finished/draft	Total
Bacteria	781/503	1284
Archaea	56/3	59
Eukarya	19/30	49
Plasmids	974/0	
Viruses	2524/0	
All Genomes	4354/536	

Genome by Metadata

IMG Statistics

The **Integrated Microbia**
Project (IMG/HMP) syste
HMP specific microbial
available genomes in IM
Vol. 36, Database issue

The current version of IM
released in April 2009.

For more details, see W
also see About IMG and

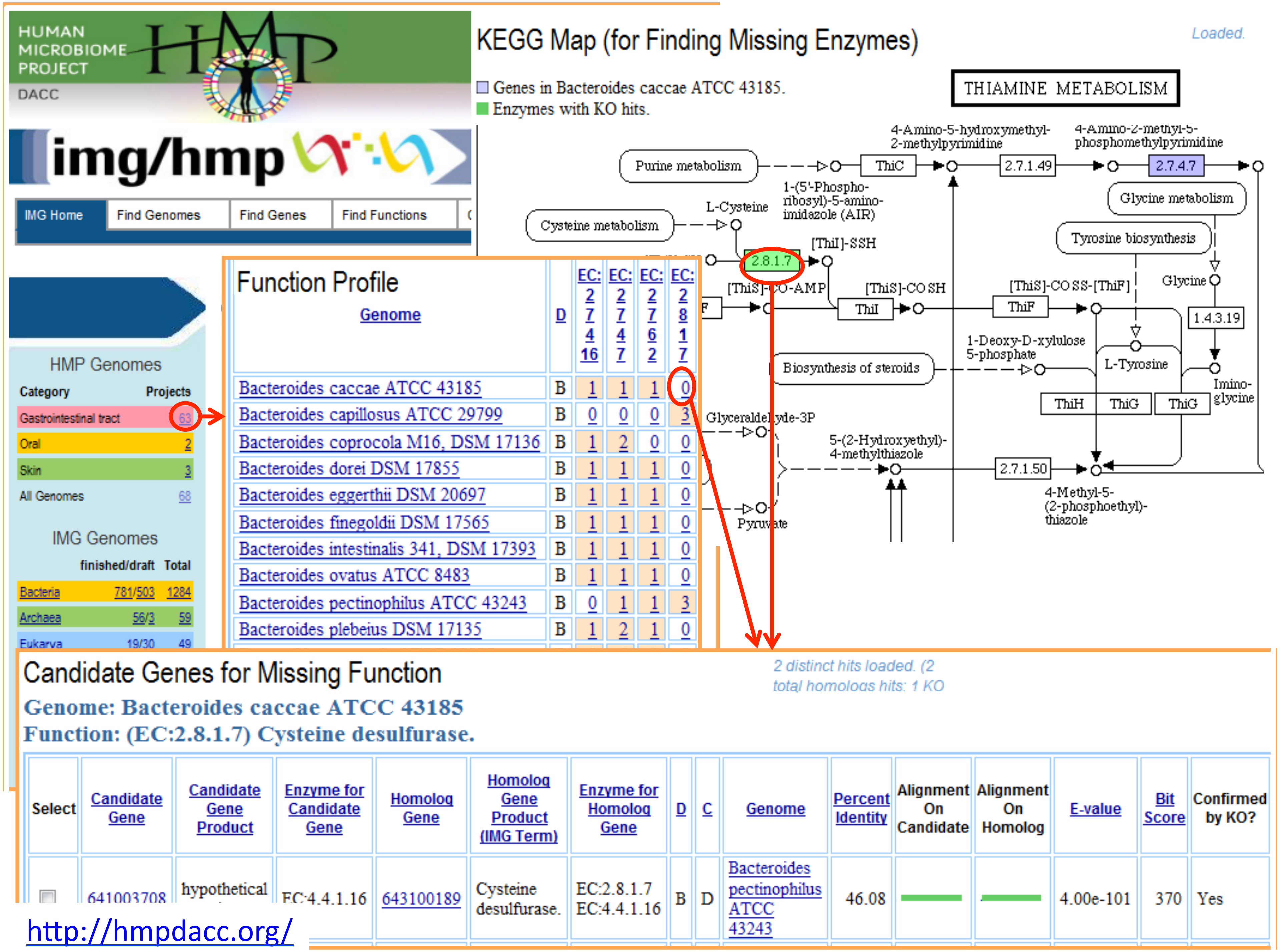
HMP Genome List for Project Category
Gastrointestinal tract

Select	D	C	Project Id	NCBI Project Id	Genome Id	Geno
☐	B	D	11834	19655	641736205	Alistipes putredinis DS
☐	B	D	13187	30747	642979323	Anaerococcus hydroge
☐	B	D	11835	19657	641736193	Anaerofustis stercorih
☐	B	D	10785	18213	641736227	Anaerostipes caccae D
☐	B	D	11836	19659	641736271	Anaerotruncus colihon
☑	B	D	10772	18163	640963023	Bacteroides caccae AT
☑	B	D	10786	18173	640963014	Bacteroides capillosus
☑	B	D	10773	20521	642791613	Bacteroides coprocola M16, DSM 17136
☑	B	D	13150	27831	642979370	Bacteroides dorei DSM 17855
☑	B	D	13151	27827	642979334	Bacteroides eggerthii DSM 20697
☑	B	D	13152	27823	642979319	Bacteroides finegoldii DSM 17565
☑	B	D	10774	20523	642791621	Bacteroides intestinalis 341, DSM 17393
☑	B	D	10783	18191	641380449	Bacteroides ovatus ATCC 8483
☑	B	D	13196	27825	642979337	Bacteroides pectinophilus ATCC 43243
☑	B	D	13153	27829	642979351	Bacteroides plebeius DSM 17135
☑	B	D	11853	19859	641736196	Bacteroides stercoris ATCC 43183
☑	B	D	10784	18195	641380447	Bacteroides uniformis ATCC 8492
☐	B	D	10769	18197	640963015	Bifidobacterium adolescentis L2-32
☐	B	D	13231	29261	642979361	Bifidobacterium angulatum DSM 20098
☐	B	D	13234	30749	642979312	Bifidobacterium catenulatum DSM 16992
☐	B	D	10923	20555	641736189	Bifidobacterium dentium ATCC 27678

Statistics for Genomes by specific KEGG Category

KEGG Categories	Gene Coun
🟥 Amino Acid Metabolism	2025
🟦 Biosynthesis of Polyketides and Nonribosomal Peptides	100
🟩 Biosynthesis of Secondary Metabolites	397
🟨 Cancers	15
🟪 Carbohydrate Metabolism	2271
🟦 Cell Motility	52
🟥 Endocrine System	146
⬛ Energy Metabolism	988
🟥 Glycan Biosynthesis and Metabolism	627
🟦 Immune Disorders	17
🟩 Immune System	14
🟨 Infectious Diseases	42
🟪 Lipid Metabolism	713
🟦 Membrane Transport	690
⬛ Metabolic Disorders	54
🟥 Metabolism of Cofactors and Vitamins	1220
🟦 Metabolism of Other Amino Acids	499
🟩 Neurodegenerative Diseases	29
🟨 Nucleotide Metabolism	875
🟪 Replication and Repair	524
🟦 Signal Transduction	326
⬜ Sorting and Degradation	221
🟥 Transcription	38
🟦 Translation	876
🟩 Xenobiotics Biodegradation and Metabolism	363

http://hmpdacc.org/



HMP Genomes

Category	Projects
Gastrointestinal tract	63
Oral	2
Skin	3
All Genomes	68

IMG Genomes

	finished/draft	Total
Bacteria	781/503	1284
Archaea	56/3	59
Eukarya	19/30	49

KEGG Map (for Finding Missing Enzymes)

Loaded.

- ▢ Genes in Bacteroides caccae ATCC 43185.
- 🟩 Enzymes with KO hits.

Function Profile

Genome	D	EC:2.7.4.16	EC:2.7.4.7	EC:2.7.6.2	EC:2.8.1.7
Bacteroides caccae ATCC 43185	B	1	1	1	0
Bacteroides capillosus ATCC 29799	B	0	0	0	3
Bacteroides coprocola M16, DSM 17136	B	1	2	0	0
Bacteroides dorei DSM 17855	B	1	1	1	0
Bacteroides eggerthii DSM 20697	B	1	1	1	0
Bacteroides finegoldii DSM 17565	B	1	1	1	0
Bacteroides intestinalis 341, DSM 17393	B	1	1	1	0
Bacteroides ovatus ATCC 8483	B	1	1	1	0
Bacteroides pectinophilus ATCC 43243	B	0	1	1	3
Bacteroides plebeius DSM 17135	B	1	2	1	0

Candidate Genes for Missing Function

2 distinct hits loaded. (2 total homologs hits: 1 KO

Genome: Bacteroides caccae ATCC 43185

Function: (EC:2.8.1.7) Cysteine desulfurase.

Select	Candidate Gene	Candidate Gene Product	Enzyme for Candidate Gene	Homolog Gene	Homolog Gene Product (IMG Term)	Enzyme for Homolog Gene	D	C	Genome	Percent Identity	Alignment On Candidate	Alignment On Homolog	E-value	Bit Score	Confirmed by KO?
☐	641003708	hypothetical	EC:4.4.1.16	643100189	Cysteine desulfurase.	EC:2.8.1.7 EC:4.4.1.16	B	D	Bacteroides pectinophilus ATCC 43243	46.08	▬	▬	4.00e-101	370	Yes

http://hmpdacc.org/

Acknowledgements

National Institute
of Allergy and
Infectious Diseases

National Human
Genome Research
Institute

Sequencing Centers

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

Institutional Support

National Institutes of Health
National Institute of Allergies & Infectious Disease
National Human Genome Research Institute

Data Analysis Coordination Center

University of Maryland
Lawrence Berkeley Laboratories
University of Colorado

UH2 & UH3 Centers

New York University School of Medicine
Virginia Commonwealth University
University of Pennsylvania
Indiana University
University of Maryland Baltimore
Washington University
University of California Los Angeles
U.S. National Human Genome Res Institute
Baylor College of Medicine
University of Michigan Ann Arbor

Collaborators contributing Strains/DNA

Anton Peleg
Barbara Murray
Cherie Ziemer
David Graham
David Schauer
Delphine Saulnier
Edgar Boedeker
Emma Allen-Vercoe
Floyd Dewhirst
Geert Huys
Glenn Songer
Gregor Reid
Harald Seifert
Harry Flint
Huiying Li
Jacques Izard
Jim Versalovic
Kathrin Witt
Kim Lewis
Kim Nemeth
Kimberlee Musser
Lenie Dijkshoom
Maria Sizova
Mark Morrison
Marty Blaser
Masami Morotomi
Michael Fischbach
Michael Surette
Milt Allison
Mogens Kilian
Nancy Taylor
Nathan Weyand
Nick Loman
Qiang Xu
Richard Hull
Rita Verhelst
Rob Gunsalus
Sharon Hillier
Sheldon Kaplan
Slava Epstein
Sue Haake
Todd Kitten
Tom Schmidt
W. Ray Butler
Xuesong He