Extending Our View of Self: the Human Gut Microbiome Initiative (HGMI)

Jeffrey I. Gordon
Ruth E. Ley
Center for Genome Sciences, Washington University, St. Louis, MO

Richard Wilson
Elaine Mardis
Jian Xu
Genome Sequencing Center, Washington University, St. Louis MO

Claire M. Fraser
The Institute for Genome Research, Rockville, MD

David A. Relman
Department of Medicine, Stanford University, Palo Alto, CA

Address correspondence to:

Jeffrey Gordon
Director, Center for Genome Sciences
Washington University
St Louis, MO 63108
Phone 314:362-7243
Fax 314:362-7047
E-mail: jgordon@molecool.wustl.edu

Abstract

Our adult bodies harbor ~10 times more microbial than human cells. Their genomes (the microbiome) endow us with physiologic capacities that we have not had to evolve on our own and thus are both a manifestation of who we are genetically and metabolically, and a reflection of our state of well-being. Our distal gut is the highest density natural bacterial ecosystem known, the most comprehensively surveyed to date, and the most highly represented in pure culture. It contains more bacterial cells than all of our other microbial communities combined. To obtain a more comprehensive view of our biology, we propose a human gut microbiome initiative (HGMI) that will deliver deep draft whole genome sequences for 100 species representing the bacterial divisions (superkingdoms) known to comprise our distal gut microbiota: 15 of these genomes will be selected for finishing. A cost-effective strategy involves producing the bulk of the coverage by shotgun reads on a 454 Life Sciences pyrosequencer. Long-range linking information will be provided by paired end reads of fosmid subclones using a conventional ABI 3730xl capillary machine. The bulk of our sequencing will use human-derived strains, representing targeted phylotypes, from existing culture collections. The list will be augmented by *in vivo* culture of a human fecal microbiota in gnotobiotic mice. The latter approach will be used to obtain vastly simplified consortia, or pure cultures of previously uncultured representatives of important gut-associated bacteria. The deposited curated genome sequences will herald another phase of completion of the 'human' genome sequencing project, provide a key reference for metagenome projects, and serve as a model for future initiatives that seek to characterize our other extra-intestinal microbial communities.

Introduction

The total number of genes in the various species represented in our indigenous microbial communities likely exceeds the number of our human genes by at least two orders of magnitude (1). Thus, it seems appropriate to consider ourselves as a composite of many species — human, bacterial, and archaeal — and our genome as an amalgam of human genes and the genes of our microbial 'selves'. Without understanding the interactions between our human and microbial genomes, it is impossible to obtain a complete picture of our biology.

Our microbiome is largely unexplored. The proposed Human Gut Microbiome Initiative (HGMI) represents a logical, timely, and cost-effective extension of the human genome project. It promises to affect our understanding of the foundations of human health, and of many common diseases that are the subject of basic and clinical research sponsored by the NIH. There are several conceptual reasons for focusing on the gut microbiota:

- Its size — up to 100 trillion cells — far exceeds the size of *all* of our body's other microbial communities.
- As twenty-first century medicine evolves its focus towards disease prevention, new and better ways of defining our health status are needed. The gut microbiota is an effector and a reporter of many aspects of our normal physiology. Much of our current understanding of its functions comes from studies of gnotobiotic model organisms, such as mice. The term 'gnotobiotic' stems from the Greek words 'gnosis' and 'bios,' meaning 'known life,' and refers to animals reared without any micro-organisms (germ-free; GF) or with defined components of the normal mouse or human gut microbiota. Comparisons of GF and colonized animals have shown that the microbiota helps regulate energy balance, both by extracting calories from otherwise inaccessible components of our diet and by controlling host genes that promote storage of the extracted energy in adipocytes (2-4). The microbiota directs myriad biotransformations, ranging from synthesis of essential vitamins to the metabolism of the xenobiotics that we ingest and the lipids that we produce (reviewed in ref. 1). The microbiota modulates the maturation and activity of the innate and adaptive immune system: an immune system educated to allow the host to tolerate a great degree of microbial diversity provides a selective advantage since this diversity ensures the stable functioning of a microbiota in the face of environmental stresses.

  Based on these and other observations, the gut microbiota has been invoked as a factor that determines susceptibility to diseases ranging from obesity and diabetes, to gastrointestinal and other malignancies, atopic disorders (asthma), infectious diarrheas, and various immunopathologic states including inflammatory bowel diseases (e.g., refs. 5-7). It is also likely to be a key contributor to individual variations in the bioavailability of orally administered drugs (8).

Our microbial partners have undoubtedly developed the capacity to synthesize novel chemical entities that help establish and sustain their mutually beneficial relationships with us. Prospecting for these 'natural products' and characterizing the host signaling and metabolic pathways through which they operate should provide new insights about the function of many of our human genes, new biomarkers for defining health or for identifying impending or fully manifest diseases within and outside of the gut, plus new treatment strategies.

- Defining, at regularly scheduled intervals, how the gut microbiota and microbiome are changing in humans living in distinct geographic regions of the planet, under varied economic conditions, also provides an opportunity to monitor our 'micro'-evolution during a time of great climatic change and increasing travel. This effort could provide new tools and metrics for identifying, forecasting and responding to national and world-wide changes in disease susceptibility.

There are also practical reasons for selecting the gut microbiome rather than the microbiomes of other human microbial communities:

- The first comprehensive molecular survey of the gut microbiota was just published by a group composed of members of the Relman lab and TIGR (June, 2005; ref. 9). This study produced 13,335 16S rRNA gene sequences from mucosal biopsy samples harvested from the proximal to the distal colons of three healthy individuals, plus one stool sample from each person. The result is the largest database of 16S rRNA sequences from a single study of any ecosystem. Three hundred ninety-five bacterial and one archaeal phylogenetic types ('phylotypes') were identified, based on the criterion that ≥99% sequence identity was required for any pair of sequences to be assigned to a unique phylotype. The number of individual sequences representing each phylotype is a measure of abundance. Thus, this study provided the most complete view to date of microbial composition ("who's there") *and* diversity ("who's there and in what numbers") in the distal human gut.
- In contrast to microbial communities in natural environments where <1% of phylotypes are represented by laboratory isolates, at least 22% of the 395 phylotypes have an available cultured representative.
- During the past year, methods have become available that make this HGMI feasible, both from an economic and technical perspective. Thus, this project can function as a model for those who wish to characterize our other microbiotas (e.g., mouth, skin, airway, vagina, etc).

Evolution of the human gut microbiota

The human GI tract is predominantly a bacterial ecosystem. Cell densities in the colon (10^{11}-10^{12}/ml contents) are the highest recorded for any known ecosystem (10). The vast majority of phylotypes belong to two divisions (superkingdoms) of Bacteria — the Bacteroidetes (48%) and the Firmicutes (51%). The remaining phylotypes are distributed among the Proteobacteria, Verrucomicrobia, Fusobacteria, Cyanobacteria, Spirochaetes and VadinBE97 (1, 9) (**Fig 1A**).

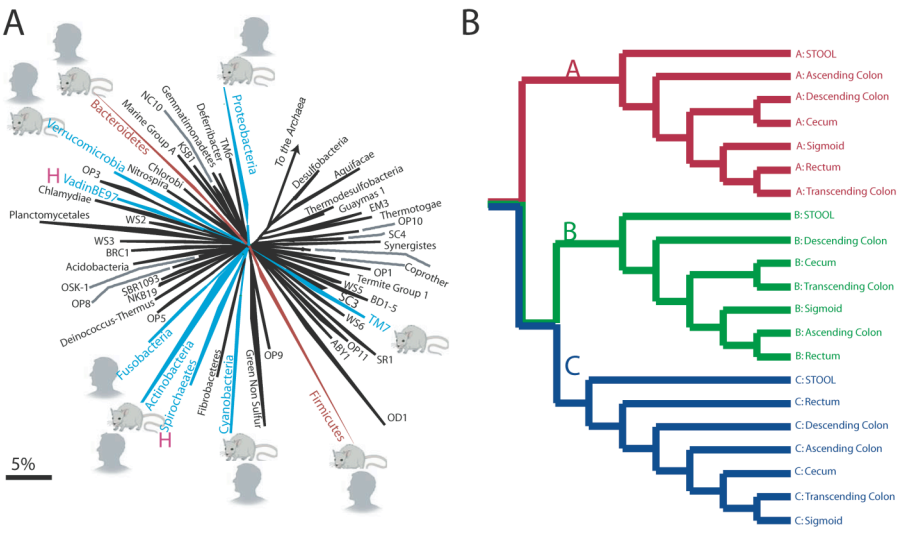

Fig. 1 — Bacterial Diversity in the human colon. (A) Phylogenetic tree of Bacteria showing described divisions (wedges, n=55). Divisions detected in a large survey of the human colonic microbiota (9) are indicated by the human head, while those detected in a large survey of the mouse distal gut (cecal) microbiota (3) are highlighted by the mouse cartoon. 'H' denotes additional divisions represented in the human fecal microbiota as determined from GenBank entries (1). Dominant divisions are colored red,

rarer divisions are blue, and undetected are black. The scale bar indicates 5% sequence divergence. **(B)** Comparisons of colonic mucosal and fecal microbiotas from three individuals. Stool and mucosal-adherent microbial communities cluster together based on the individual donor rather than the sample type. This cluster analysis is based on phylogenetic trees constructed in Arb and compared using the UniFrac metric (3). Color codes: red, individual A; green, individual B; blue, individual C. The p-value for the tree, based on Monte Carlo simulations, is <0.001. All nodes are well supported (Jackknife values of >0.95, representing the percent of the time a node was present when a sample was randomly removed from the analysis, n=1000 replicates).

A basic and unanswered question about human biology is the degree to which our microbiome is uniquely "human." Together with the Washington University Genome Sequencing Center (WU-GSC), members of the Gordon lab recently generated a 5,088-member 16S rRNA sequence dataset from the distal intestines (ceca) of normal adult C57BL/6 mice (n=19; ref. 3). Using ≥95% and ≥97% full sequence identity to delimit a genus and a species respectively, we found that 64% of these sequences were not assignable to known genera, and only 7% represented previously cultured species. Although 85% of the sequences represented genera that have not been detected in humans, there is considerable similarity between human and mouse distal gut microbiotas at the division level (**Fig. 2**). As in humans, the two most abundant divisions are the Firmicutes (60-80% of sequences) and the Bacteroidetes (20-40%). Greater than 75% of the Firmicutes are in Clostridium cluster XIVa, a common clade in humans that includes butyrate producers and *Eubacterium eligens*. The majority (>88%) of the Bacteroidetes belong to Bacteroidetes 4b, which lacks a cultured representative. Proteobacteria, Actinobacteria and Cyanobacteria, present at low levels in the human colonic microbiota, and TM7, previously identified in the human mouth (gingival) microbiota, each comprise ≤1% of the mouse cecal bacterial community.

Interestingly, our studies of genetically obese *ob*/*ob* mice and their lean *ob*/+ and +/+ littermates revealed that obesity is associated with a marked increase in the representation of Firmicutes and a reduction in the representation of Bacteroidetes (3). These changes are division-wide, do not reflect differences in chow consumption, and may either be a mediator of obesity or an adaptive host response designed to maintain energy balance. These findings provide yet another view of the contributions of the microbiota to the regulation of energy balance.

We have embarked on a 16S rRNA sequence-based analysis of the fecal microbiota of a variety of mammals, including non-human primates living in the St. Louis Zoological Park and in Africa, to explore whether mammalian speciation is associated with the acquisition of distinctive microbial community structures that could endow their hosts with varied physiological capabilities (1).

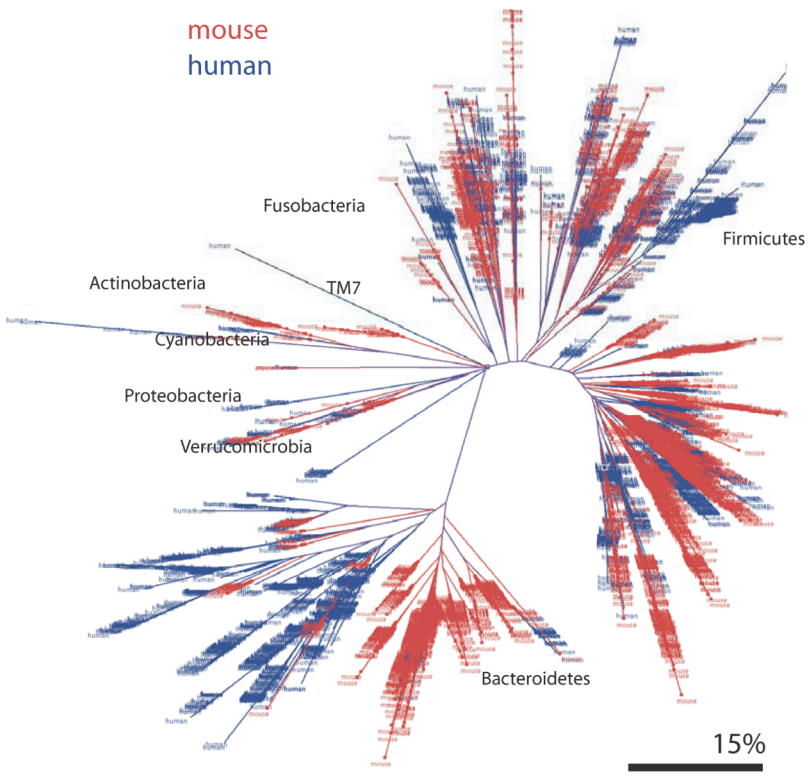

Fig 2 — Phylogenetic tree of 11,831 human and 5,088 mouse distal gut-derived 16S rRNA sequences. Data from bacteria harvested from both mammalian hosts were obtained using the same 16S rRNA gene-directed primers and PCR cycle numbers. The bar represents 15% sequence divergence.

We can consider our gut microbiota to be an efficient natural bioreactor (11), programmed to break down food and supply us with the extracted energy and nutrients. This bioreactor is stable: i.e., it is resistant to chaotic blooms of subpopulations (or pathogens) that could be disruptive. Functional redundancy encoded in genomes from widely divergent bacterial lineages would provide the host with insurance against disruption of the food web caused by loss of keystone species from selective

sweeps (12) (e.g., by phage attacks). Ecologic principles predict that host-driven ("top-down") selection for functional redundancy would result in a community composed of widely divergent microbial lineages (divisions) whose genomes contain functionally *similar* suites of genes. Another prediction is the widespread occurrence of, and abundant mechanisms for, horizontal gene transfer. In contrast, competition between members of the microbiota would exert "bottom-up" selection pressure that results in specialized genomes with functionally *distinct* suites of genes (metabolic traits). Once established, these lineage-specific traits can be maintained by barriers to homologous recombination (13).

There has been virtually no analysis of how selective pressures and community dynamics have shaped the microbiome in healthy or diseased humans. Scientists who study microbial genome evolution in natural habitats and microcosms have limited their studies to a small number of genes, or have relied on fingerprinting techniques that do not provide sufficient information about organismal gene content. Therefore, information about a wide range of genes, anchored in whole genome sequences obtained from the spectrum of phylotypes represented in the human gut microbiota, would be of great interest to those who study human biology and evolution, as well as those who examine the interplay between environment and genome structure/function ('ecogenomics').

The 16S rRNA gene runs on such a slow evolutionary clock that there is little variation to infer the evolutionary history of close relatives. Studies of *E. coli* strains suggest that the genome content of bacteria with identical 16S rRNA gene sequences can differ by as much as 30% (14). On the other hand, $C_o t$ analyses indicate that bacteria whose 16S rRNA sequences are ≥97% identical may have very similar genomes (15), leading to the common practice of adopting 97% sequence identity as a permissive threshold for delimiting a "species" (i.e., bacteria that share a relatively recent common ancestor and have a common stable core of genes). Comparing multiple whole genomes, representing 16S rRNA phylotypes with different degrees of relatedness, would allow genes to be identified that were inherited through vertical transmission, arose from duplication events (paralogs), or were acquired via horizontal transfer (xenologs). These whole genome sequences would also help answer a number of key questions about the microbiome. Which gene families are widespread among lineages and therefore essential for survival in the gut ecosystem? How much horizontal gene transfer occurs between distant versus close relatives in the densely populated distal gut and how does this relate to the evolution and functional stability of the microbiota's metabolome? Can features of microbial genome structure and microevolution be used as biomarkers of health, or of susceptibility to specific diseases?

Some of these points are illustrated by our studies of three members of Bacteroidetes that are quite distinct phylogenetically: *B. thetaiotaomicron*, which comprises 12% of all Bacteroidetes and 6% of all Bacteria in the 11,831-member human colonic 16S rRNA dataset, *B. vulgatus* (31% and 15%), and *B. distasonis* (0.8% and 0.4%). Several years ago, the Gordon lab generated the first finished genome sequence of a prominent human colonic symbiont, *Bacteroides thetaiotaomicron* (16). Together with the WU-GSC, we have recently produced finished genomes for the other two organisms. The genomes of all three Bacteroides spp. are peppered with mobile elements that can facilitate horizontal gene transfer. All have highly evolved 'glycobiomes' (genes involved the acquisition, breakdown or synthesis of carbohydrates). The size of *B. thetaiotaomicron*'s armamentarium of polysaccharide-degrading enzymes can be appreciated by comparing its 6.26 Mb genome, which encodes 241 glycoside hydrolases and polysaccharide lysases, to our 2.85 Gb genome which encodes only 98, and is deficient in the enzyme activities required to break down xylan-, pectin-, and arabinose-containing polysaccharides that are common components of dietary fiber (we have one gene in this class versus 64 in *B. thetaiotaomicron*; see ref. 4 and http://afmb.cnrs-mrs.fr/CAZY/).

Once whole genome sequences are available, the operating principles underlying the functional and structural stability of the microbiota can be characterized using 'humanized' gnotobiotic mouse models colonized with one or more members of the community. For example, our *in vivo* functional genomic and mass spectrometry-based metabolomic, as well as imaging studies of the adaptive foraging behavior of *B. thetaiotaomicron* in gnotobiotic mice disclosed that (i) groups of bacteria assemble on undigested or partially digested food particles, shed elements of the mucus gel layer, and/or exfoliated epithelial cells; (ii) bacterial attachment to these nutrient reservoirs is directed by glycan-specific outer membrane binding proteins that are opportunistically deployed depending upon the glycan environment; (iii) attachment helps oppose bacterial washout from the intestinal bioreactor and promotes harvest of carbohydrates by an adaptively expressed repertoire of glycoside hydrolases; (iv) when polysaccharide availability from the diet is reduced, the organism turns to host mucus polysaccharides (4). This type of adaptive foraging behavior promotes ecosystem stability. Our results suggest that microbial nutrient metabolism along the length of the intestine is a summation of myriad selfish and

syntrophic relationships expressed by inhabitants attached to these nutrient platforms.

Whole genome sequencing of microbes: ever cheaper, ever faster and a foundation for interpreting metagenomic datasets

Understanding the metabolic capabilities of the microbiota is a major challenge, provides diagnostic and therapeutic opportunities and may herald a not so fanciful era of personalized nutrition where diet is matched to the processing capacity of an individual's microbiome-encoded metabolome. Members of TIGR and the Relman lab have initiated a human gut metagenome project. Interpreting the "gene space" identified by random sequencing of community DNA will be aided greatly by the availability of reference gut microbial genomes. Defining metabolic capacity by *in silico* reconstruction of pathways from metagenomic datasets is very challenging given (i) the limited sampling of genomes, and (ii) the difficulties in assembling genomes to obtain information about gene linkage (e.g., operons). While metagenomic studies of simple environments containing a few microbial species can lead to full or partial genome reconstruction (17), the complexity of the human gut microbiota (>7000 strains, ref. 1) is too great to allow adequate sequence coverage (and present day computational tools are insufficient) to support extensive genome assembly.

A new approach, involving a new instrument, promises to dramatically reduce the cost and increase the speed of producing deep draft whole genome sequences with ≥8X coverage from various gut microbes. In this scheme, the bulk of coverage is produced by shotgun reads on the 454 Life Sciences PicoTiterPlate (http://www.454.com), a massively parallel DNA pyrosequencing platform. Long-range linking information is provided by paired end reads of fosmid subclones, using a conventional ABI 3730xl capillary sequencer. The 454 machine, now operating at the WU-GSC, typically generates 20Mb of raw DNA sequence in a four hour run, with average read lengths of 100 bp, at a cost of $8,500/genome. This represents a 100-fold increase in throughput, and significant cost savings over the ABI 3730xl sequencer. Subcloning into plasmids is not needed, circumventing the loss of gene sequences that can occur due to cloning biases. A draft genome sequence is produced by first assembling the 454-generated reads into sequence contigs: these contigs are then linked by fosmid end reads into an ordered and oriented scaffold ("supercontig") that is amenable to gap closure (finishing) by interested members of the scientific community.

We (the Gordon lab and WU-GSC) have had practical experience with this mixed platform approach for deep draft sequencing (and finishing). We used traditional methods and a capillary sequencer to finish the 1,597,423 bp genome of *H. pylori* strain AG7:8, obtained from a patient with chronic atrophic gastritis (atrophic gastritis is the precursor to gastric adenocarcinoma, which is associated with persistent *H. pylori* infection). A total of 30,171 reads (average Q20 length of 562 bp) were collected from a plasmid library with a 5Kb average insert size and a fosmid library with a 40Kb average insert size) (Q20 sequence coverage of 7.8X and 2.1X, respectively). A finished genome was produced in 8 months (Oh *et al.*, manuscript in preparation). In a follow-up independent effort, 447,626 short-reads were collected in a single four hour run using our 454 machine (25X Q20 sequence coverage, average Q20 read length of 90bp). The short-read assembly, generated using the program Newbler, contained 93 sequence contigs totaling 1,561,248 bp (N50 contig size = 34.3 kB; N50 contig number = 15). We were able to align 1,550,092 short-read contig bases (99.3% of all contig bases) to 1,550,762 finished genome bases: i.e., the short-read contigs generated by the 454 PicoTiterPlate platform covered 97.1% of the total genome. The contiguity of this short-read assembly was comparable to an 8X whole-genome shotgun assembly of traditional reads obtained with an ABI 3730 capillary sequencer. The overall accuracy in aligned regions was 99.81%. Together with paired fosmid reads (representing 2.1X Q20 sequence coverage), and algorithms developed by members of the WU-GSC (e.g., MapLinker; ref. 18), we were able to rapidly generate an assembly composed of one supercontig and 7 small contigs that together covered 99.6% of the genome.

Ongoing technical developments will further increase accuracy and speed: for example, improved PicoTiterPlate sequencing reactions that increase read length and reduce base error rate; improvements in base-calling software; and new assembly and finishing tools, including one that automatically detects and corrects misassemblies based on fosmid read-pair constraints.

Selecting genomes to sequence

The 11,831-member 16S rRNA sequence dataset generated from the human colonic microbiota of three healthy adults provides 395 phylotypes that are candidates for whole genome sequencing. We have identified 86 cultured representatives (22%) of these phylotypes: all are derived from humans,

principally fecal samples **(Table 1; see http://gordonlab.wustl.edu/Tree for a phylogenetic tree of the cultured isolates placed in the context of the 11,831 16S rRNA sequence dataset).**

Table 1 — Cultured bacteria that represent the colonic dataset of Eckburg *et al.* (9). Abbreviations: Strain ID, refers to strain identity (typically a catalog number from a culture collection); %ID, the minimum 16S rRNA gene sequence identity between the strain and sequences in its relatedness cluster, based on pair-wise identity across 1300 bp; #, denotes the number of sequences in the relatedness cluster; % of total, the abundance of the cluster in the total dataset; GenBank, accession number for the 16S rRNA gene sequence obtained from the isolate; Status, indicates if the cultured isolate's genome has been sequenced, or if sequencing is in progress.

	Divisions	*Genus*	*Species*	*Strain ID*	*%ID*	*#*	*% of total*	*GenBank*	*Status*
1	Bacteroidetes	Bacteroides	AFS519	ASF519	99	2	0.017	AF157056	
2	Bacteroidetes	Bacteroides	sp.	CCUG 39913	100	2	0.017	AJ518872	
3	Bacteroidetes	Bacteroides	sp.	Smarlab 3301186	99	2	0.017	AY538684	
4	Bacteroidetes	Bacteroides	ovatus	ATCC 8483T	99	4	0.034	X83952	
5	Bacteroidetes	Bacteroides	salyersiae	WAL 10018	99	4	0.034	AY608696	
6	Bacteroidetes	Alistipes	finegoldii	ANH 2437	99	5	0.042	AJ518874	
7	Bacteroidetes	Bacteroides	sp.	MPN isolate group 6	99	12	0.101	AF357554	
8	Bacteroidetes	Bacteroides	sp.	DSM 12148	97	15	0.127	AJ518876	
9	Bacteroidetes	Bacteroides	merdae	ATCC 43184T	99	31	0.262	X83954	
10	Bacteroidetes	Bacteroides	distasonis	ATCC 8503	98	32	0.270	M25249	finished
11	Bacteroidetes	Bacteroides	stercosis	ATCC 43183T	98	41	0.347	X83953	
12	Bacteroidetes	Bacteroides	splanchnicus	NCTC 10825	99	42	0.355	L16496	
13	Bacteroidetes	Bacteroides	WH2	Gordon Lab	99	72	0.609	AY895180	in progress
14	Bacteroidetes	Bacteroides	uniformis	ATCC 8492	99	85	0.718	L16486	
15	Bacteroidetes	Bacteroides	WH302	Gordon Lab	99	100	0.845	AY895184	
16	Bacteroidetes	Alistipes	putredinis	ATCC 29800	99	150	1.268	L16497	
17	Bacteroidetes	Bacteroides	fragilis	ATCC 25285T	99	243	2.054	X83935	finished
18	Bacteroidetes	Bacteroides	caccae	ATCC 43185T	99	332	2.806	X83951	
19	Bacteroidetes	Bacteroides	thetaiotaomicron	ATCC 29148	98	706	5.967	L16489	finished
20	Bacteroidetes	Bacteroides	vulgatus	ATCC 8482	99	1749	14.783	M58762	finished
21	Firmicutes	Clostridium	leptum	ATCC 29065	99	1	0.008	M59095	
22	Firmicutes	Clostridium	boltaea	ATCC BAA-613	99	1	0.008	AJ508452	
23	Firmicutes	Anaerotruncus	colihominis	CCUG 45055T	99	1	0.008	AJ315980	
24	Firmicutes	Allisonella	histaminiformans	CCUG 48567T	99	1	0.008	AF548373	
25	Firmicutes	Bulleidia	moorei	ATCC BAA-170	99	1	0.008	AY044915	
26	Firmicutes	Eubacterium	plautii	ATCC 29863	99	1	0.008	AY724678	
27	Firmicutes	Bacteroides	capillosus	ATCC 29799	99	1	0.008	AY136666	
28	Firmicutes	Peptostreptococcus	sp.	oral clone CK035	99	1	0.008	AF287763	
29	Firmicutes	Anaerococcus	vaginalis	CCUG 31349	99	1	0.008	AF542229	
30	Firmicutes	Clostridium	bartlettii	CCUG 48940	99	2	0.017	AY438672	
31	Firmicutes	Ruminococcus	bromii	ATCC 27255	99	2	0.017	L76600	
32	Firmicutes	Lactobacilllus	lactis	Ssp. IL1403	99	2	0.017	X64887	finished
33	Firmicutes	Clostridium	symbiosum	ATCC 14940	99	2	0.017	M59112	
34	Firmicutes	Clostridium	sp.	DSM 6877(FS41)	99	2	0.017	X76747	
35	Firmicutes	Clostridium	sp.	A2-207	99	2	0.017	AJ270471	
36	Firmicutes	Anaerofustis	stercorihominis	CCUG 47767T	99	2	0.017	AJ518871	
37	Firmicutes	Streptococcus	mitis	ATCC 903	99	3	0.025	AF003929	
38	Firmicutes	Clostridium	scindens	ATCC 35704	99	3	0.025	AF262238	
39	Firmicutes	Clostridium	spiroforme	DSM 1552	99	3	0.025	X73441	
40	Firmicutes	Ruminococcus	callidus	ATCC 27760	99	4	0.034	X85100	
41	Firmicutes	Streptococcus	parasanguinis	ATCC 15912	99	4	0.034	AF003933	
42	Firmicutes	Coprococcus	eutactus	ATCC 27759	99	5	0.042	D14148	
43	Firmicutes	Gemella	haemolysans	ATCC 10379	99	5	0.042	L14326	
44	Firmicutes	Clostridium	sp.	A2-183	99	5	0.042	AJ270482	
45	Firmicutes	Peptostreptococcus	micros	ATCC 33270	99	5	0.042	AF542231	
46	Firmicutes	Eubacterium	ventriosum	ATCC 27560	99	5	0.042	L34421	
47	Firmicutes	Eubacterium	halii	ATCC 27751	99	9	0.076	L34621	
48	Firmicutes	Ruminococcus	gnavus	ATCC 29149	99	10	0.085	L76597	
49	Firmicutes	Coprococcus	catus	ATCC 27761	99	11	0.093	AB038359	
50	Firmicutes	Eubacterium	siraeum	ATCC 29066	99	13	0.110	L34625	
51	Firmicutes	Clostridium	sp.	SL6/1/1	98	13	0.110	AY305317	
52	Firmicutes	Roseburia	intestinalis	DSM 14610	99	18	0.152	AJ312385	
53	Firmicutes	Clostridium	sp.	GM2/1	99	21	0.177	AY305315	
54	Firmicutes	Clostridium	sp.	A2-194	99	26	0.220	AJ270473	
55	Firmicutes	Eubacterium	eligens	ATCC 27750	99	27	0.228	L34420	in progress
56	Firmicutes	Clostridium	sp.	14774	99	35	0.296	AJ315981	
57	Firmicutes	Clostridium	sp.	A2-166	99	39	0.330	AJ270489	
58	Firmicutes	Clostridium	sp.	A2-175	99	39	0.330	AJ270485	
59	Firmicutes	Roseburia	faecalis	M6/1	99	42	0.355	AY804149	
60	Firmicutes	Ruminococcus	obeum	ATCC 29174	98	54	0.456	L76601	
61	Firmicutes	Catenibacterium	mitsuokai	JCM 10609	99	57	0.482	AB030221	
62	Firmicutes	Ruminococcus	torques	ATCC 27756	88	57	0.482	D14137	
63	Firmicutes	Clostridium	sp.	SR1/1	97	59	0.499	AY305321	
64	Firmicutes	Subdoligranulum	variabile	CCUG 47106	99	63	0.532	AJ518869	

65	Firmicutes	Clostridium	sp.	L1-83	99	66	0.558	AJ270474	
66	Firmicutes	Clostridium	sp.	L2-6	99	67	0.566	AJ270470	
67	Firmicutes	Dorea	formicigenerans	ATCC 27755	99	73	0.617	L34619	
68	Firmicutes	Clostridium	sp.	A2-231	99	79	0.668	AF270484	
69	Firmicutes	Clostridium	sp.	A2-165	99	100	0.845	AJ270469	
70	Firmicutes	Dialister	sp.	E2_20	99	120	1.014	AF481209	
71	Firmicutes	Dorea	longicatena	CCUG 45247	99	139	1.175	AJ132842	
72	Firmicutes	Clostridium	sp.	SS2/1	99	223	1.885	AY305319	
73	Firmicutes	Eubacterium	rectale	ATCC 33656	99	315	2.662	L34627	in progress
74	Firmicutes	Faecalibacterium	prausnitzii	ATCC 27768	99	449	3.795	AJ413954	
75	Verrucomicrobia	Akkermansia	muciniphila	ATCC BAA-835	99	77	0.651	AY271254	
76	Fusobacteria	Fusobacterium	sp.	oral clone R002	99	10	0.085	AF287806	
77	Proteobacteria	Escherichia	coli		99	1	0.008	M87049	finished
78	Proteobacteria	Haemophilus	parainfluenziae	CCUG 12836	99	4	0.034	AY362908	
79	Proteobacteria	Bilophila	wadsworthii	ATCC 4926	99	7	0.059	L35148	
80	Proteobacteria	Desulfovibrio	piger	ATCC29098	99	15	0.127	AF192152	
81	Actinobacteria	Corynebacterium	durum	NML ID 99-0047	99	1	0.008	Z97069	
82	Actinobacteria	Bifidobacterium	adolescentis	L2-32	99	1	0.008	AY305304	
83	Actinobacteria	Actinomyces	graevenitzii	CCUG 27294T	99	1	0.008	AJ540309	
84	Actinobacteria	Corynebacterium	sundsvallense	CCUG 36622	99	1	0.008	Y09655	
85	Actinobacteria	Actinomyces	odontolyticus	DSM 4331	99	2	0.017	X53227	
86	Actinobacteria	Collinsella	aerofaciens	JCM 10188	98	13	0.110	AB011816	
				Totals		**6035**	**51**		

Sixty-five of the phylotypes in the 16S rRNA dataset are from the Bacteroidetes. As noted above, we have already produced finished genomes for three of these phylotypes. **Table 1** lists 15 other phylotypes with available cultured representatives. Bacteroidetes contains 45 phylotypes that have no reported cultured representatives (NCRs), even though they have high relative abundance (up to 30% of division members and 15% of all identified bacteria in the dataset). Below, we describe methods that will be used to retrieve some of these NCRs for sequencing.

Of the 395 phylotypes in the dataset, 301 are members of the Firmicutes: more than half of these belong to the class Clostridia. We have identified 63 phylotypes within the Firmicutes with cultured representatives (**Table 1**). The other divisions represented in the 11,831-member colonic 16S rRNA dataset together comprise less than 1% of all bacterial sequences, and were not detected in all individuals surveyed (perhaps due to inadequate sample coverage). Table 1 lists six identified cultured phylotype representatives from the Actinobacteria, five from the Proteobacteria, and one each from the Verrucomicrobia and Fusobacteria.

Obtaining genomic DNA for whole genome sequencing

The 454 pyrosequencer uses very small volume (60 nL) reaction mixtures, reducing the amount of starting material needed to generate 8X sequence coverage of a 2.5 Mb genome to <10 μg of DNA. Assuming 5fg DNA/cell, 10 μg represents 2 x 10^7 bacteria: this is readily achievable since DNA will generally be prepared from pure cultures of organisms already archived in culture collections (protocols are already established for growth of the 77 yet-to-be sequenced isolates listed in Table 1).

Many of the phylotypes with NCRs that we are interested in obtaining for deep draft whole genome sequencing have high representation in feces. Analysis of the 11,831-member 16S rRNA dataset indicates that for a given host, the bacterial composition of a fecal sample is similar to the composition of his/her colonic mucosal-associated communities: in other words, feces provide a readily available starting material representative of an individual's distal gut microbiota (**Fig. 1B)**.

We propose using germ-free (GF) mice that lack any indigenous microbes as "living test-tubes" to retrieve previously non-cultured representatives of selected phylotypes from human fecal samples. Eight years ago, the Gordon lab established, and has continued to run a large GF facility that currently consists of 40 flexible film gnotobiotic isolators containing an average daily census of 300 cages (4-5 mice/cage). A number of groups have introduced a human fecal microbiota into GF animals (e.g., ref. 19). Since these studies were conducted before large-scale 16S rRNA-based enumeration studies were feasible, we do not have a clear view of what fraction of the human gut microbiota takes hold in the mouse gut. If the mouse gut environment does not favor growth of any specific subgroup of the human microbiota, in principle we should be able to dilute to extinction less abundant components, resulting in a simplified community that includes phylotypes of interest. To test the feasibility of this approach, we conducted a pilot study where we gavaged adult GF C57BL/6 mice with 100 μL of a $1x10^{-2}$ dilution of human feces. After a 14d colonization, the most abundant 16S rRNA sequences recovered from their cecal contents belonged to NCR groups: 15% were members of the NCR 3 and NCR 8 groups of Bacteroidetes that are high priority targets for high draft whole genome sequencing (**Fig. 3**). We also

recovered two phylotypes from the Firmicutes that did not have reported cultured representatives.

These simplified communities can be further rarified by serial mouse-to-mouse dilution, until the composition of the mixture is simple enough for whole-genome shotgun sequencing, followed by assembly of the component genomes — an approach that has been shown to be feasible for five-member communities (20, 21). Genomes representing groups that have already been sequenced could be subtracted *in silico* from the mix of reads, prior to assembly of targeted phylotype genomes.

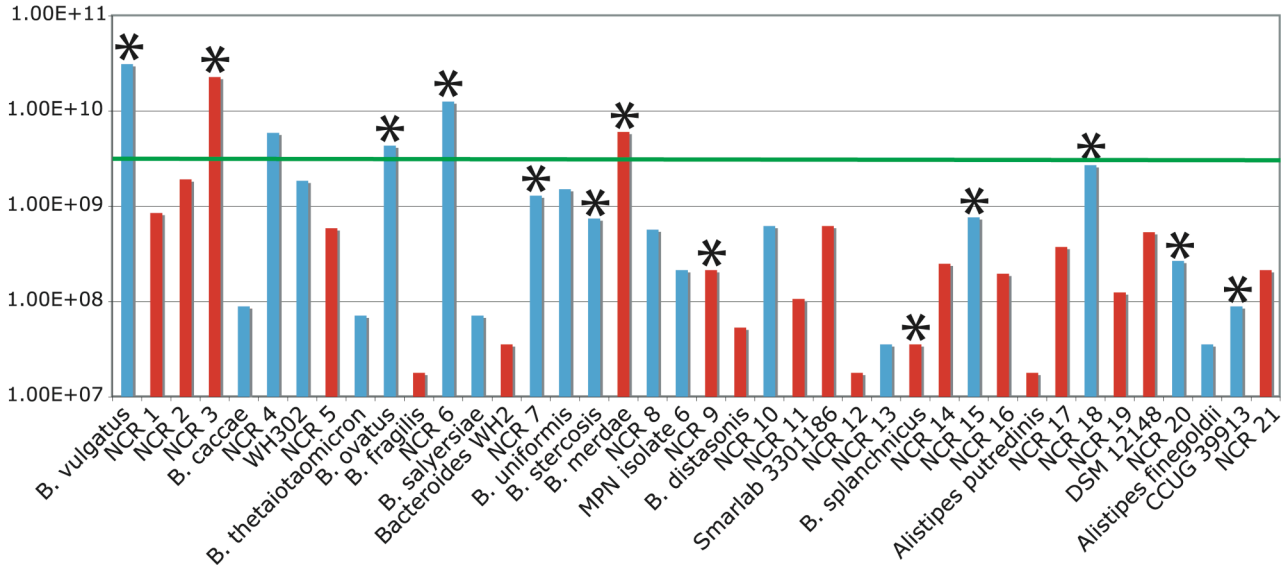

Fig. 3 — Using gnotobiotic mice to obtain previously non-culturable representatives of phylotypes of interest. Assuming a density of 10^{11} cells/mL cecal contents for a complete microbiota, and a starting inoculum with a representation of Bacteroidetes equivalent to that in the Eckburg dataset, a $1x10^{-9.5}$ dilution of the inoculum (denoted by the green line in the figure) theoretically would eliminate all but six groups, two of which lack previously cultured representatives. Our analysis of the cecal microbiota of GF mice colonized with a less dilute human fecal microbiota yielded 16S rRNA sequences from phylotypes that had previously cultured representatives (blue bars with asterisks) as well as phylotypes that had no reported cultured representatives (red bars with asterisks). Blue and red bars without asterisks denote phylotypes that were not detected in the ceca of these 'humanized' gnotobiotic mice but were present in the Eckburg human colon 16S rRNA dataset. Note that a 10–14d colonization of gnotobiotic mice with a single or a few selected cultured members of human-derived Bacteroidetes species yields ≥20 μg of intact microbial genomic DNA from the cecum.

Other schemes based on the use of gnotobiotic mice can be envisioned for further purification. As noted above, members of the gut microbiota are distributed on various nutrient platforms in the cecal habitat, including shed mucus fragments, exfoliated epithelial cells, as well as the mucus blanket that overlies the intact epithelium (4). Mucus can be retrieved free of host cells from distinct regions of the ceca of gnotobiotic mice using laser-capture microdissection (LCM; we have used this method over the course of 5 years for our functional genomic analysis of host-bacterial symbiosis in gnotobiotic animals; e.g., refs. 22, 23). Thus, LCM provides a way of obtaining subsets of an already simplified and physically partitioned human colonic microbiota from their gnotobiotic mouse hosts. The material harvested by LCM can be lysed, and the liberated genomes amplified using ϕ29 DNA polymerase (24) followed by 16S rRNA gene-based phylotyping and genome sequencing.

Annotation and data deposition

Issues for those in the field of comparative microbial genomics and ecogenomics include more standardized nomenclature for annotation, development of better algorithms for distinguishing orthology from paralogy and identifying xenologs, new and more efficient approaches for performing whole genome-based phylogenetic analyses, and developing better methods for *in silico* reconstruction of metabolomes. While it is beyond the scope of this white paper to describe new approaches to these problems or various challenges related to genome assembly (e.g., reconstructions after sequencing of intact communities containing small numbers of component species), projects such as this HGMI should catalyze efforts to find novel and effective solutions.

In terms of annotation, a microbial genome analysis group would provide the community with results from the following analyses: putative ORFs searched against GenBank based on a pairwise sequence comparison method such as BLAST; HMMER2 (http://hmmer.wustl.edu/) used to search ORFs against the current collection of Pfam profile HMMs to locate regions that belong to known domain families; TopPred (25) used to identify membrane-spanning domains; SignalP (26) used to detect the presence of signal peptides and their likely cleavage sites; PSORT (27) used to predict the cellular location of proteins; transporters analyzed based on the classification schemes in TC-DB (http://tcdb.ucsd.edu/tcdb/); families of paralogous genes in a given genome constructed by pairwise searching of ORFs against themselves using BLASTP (matches with $E \geq 10^{6}$ over 60% of the query search length would be identified and clustered into multigene families); multiple alignments of protein families generated with CLUSTAL W (28); phylogenetic trees of genes and proteins built using PHYLIP (http://evolution.genetics. washington.edu/phylip.html).

The WU-GSC has always insisted on presenting data to the public by deposition of traces within 24 hours of data collection and of assembles >1kb. We will continue to do so, adding all new assemblies to our ftp site (ftp://genome.wustl.edu/pub/seqmgr/bacterial/) and Blast analysis site (http://www.genome.wustl.edu/blast/client.pl). We maintain a bacterial web page dedicated to displaying information regarding ongoing projects (http://www.genome.wustl.edu/projects/bacterial/cmpr_microbial/).

A prototype for a Human GUT Microbiome Database (HGM_DB) is provided by http://img.jgi.doe.gov/v1.1/main.cgi. A comprehensive database would link annotated genomes to transcriptomes, metabolomes and analysis tools. From a software development perspective, the database system should have utility beyond the human gut microbiota: i.e., it should be adaptable to other complex microbial communities associated with humans, vertebrate and non-vertebrate model organisms, as well as natural ecosystems (e.g., non-polluted and polluted environments).

Selection of a subset of 15 phylotypes for gap closure and finishing

We anticipate that based on this annotation effort, our groups will want to select 15 of the genomes with deep draft sequences, representing highly distinctive phylotypes, and proceed to close all gaps so that a finished product can be obtained. This would not increase the HGMI budget substantially: our experience is that the cost of finishing is less if it is directly coupled to deep draft sequencing rather than being deferred to a later date. Our experience is that finishing is important for a thoughtfully culled subset of microbial genomes since it can reveal surprises about gene content, genome organization and genome evolution that are not apparent even with a deep draft sequence, and because it can benefit metagenomics efforts, and/or analyses of closely related strains.

Estimated costs and time frame for HGMI

Estimated total cost would be 2.8 million dollars [deep draft sequencing (~8X coverage) of 100 genomes at a cost of $20,000/genome; selection of a subset of 15 genomes for finishing at an additional cost of $30,000/genome: $350,000 to support annotation and database management]. **The project should be completed in 3 years**.

References

1. Backhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. & Gordon, J. I. (2005) *Science* **307,** 1915-20.
2. Backhed, F., Ding, H., Wang, T., Hooper, L. V., Koh, G. Y., Nagy, A., Semenkovich, C. F. & Gordon, J. I. (2004) *Proc. Natl. Acad. Sci U.S.A.* **101,** 15718-23.
3. Ley, R. E., Backhed, F., Turnbaugh, P., Lozupone, C., Knight, R. & Gordon, J. I. (2005) *Proc. Natl. Acad. Sci. U.S. A.*, in press
4. Sonnenburg, J. L., Xu, J., Leip, D. D., Chen, C. H., Westover, B. P., Weatherford, J., Buhler, J. D. & Gordon, J. I. (2005) *Science* **307,** 1955-59.
5. Moore, W. E. & Moore, L. H. (1995) *Appl. Environ. Microbiol.* **61,** 3202-07.
6. Powrie, F. (2004) *Ann. N.Y. Acad. Sci.* **1029,** 132-41.
7. Rakoff-Nahoum, S., Paglino, J., Eslami-Varzaneh, F., Edberg, S. & Medzhitov, R. (2004) *Cell* **118,** 229-41.
8. Nicholson, J. K., Holmes, E. & Wilson, I. D. (2005) *Nature Reviews Microbiology* **3,** 431-38.
9. Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S. R., Nelson, K. E. & Relman, D. A. (2005) *Science* **308,** 1635-38.
10. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. (1998) *Proc. Natl. Acad. Sci. U.S.A.* **95,** 6578-83.
11. Sonnenburg, J. L., Angenent, L. T. & Gordon, J. I. (2004) *Nat. Immunol.* **5,** 569-73.
12. Cohan, F. M. (2002) *Annu Rev Microbiol* **56,** 457-87.
13. Majewski, J., Zawadzki, P., Pickerill, P., Cohan, F. M. & Dowson, C. G. (2000) *J Bacteriol* **182,** 1016-23.
14. Perna, N. T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J., *et al.,* (2001) *Nature.* **409,** 529-33.
15. Rossello-Mora, R. & Amann, R. (2001) *FEMS Microbiol Rev.* **25,** 39-67.
16. Xu, J., Bjursell, M. K., Himrod, J., Deng, S., Carmichael, L. K., Chiang, H. C., Hooper, L. V. & Gordon, J. I. (2003) *Science* **299,** 2074-76.
17. Tringe, S. G., von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., Podar, M., Short, J. M., Mathur, E. J., Detter, J. C., Bork, P., Hugenholtz, P. & Rubin, E. M. (2005) *Science* **308,** 554-57.
18. Xu, J. & Gordon, J.I. (2005) *Bionformatics* **21**, 1265-56.
19. Kibe, R., Sakamoto, M., Yokota, H., Ishikawa, H., Aiba, Y., Koga, Y. & Benno, Y. (2005) *Appl. Environ. Microbiol.* **71,** 3171-78.
20. Allen, E. E. & Banfield, J. F. (2005) *Nat. Rev. Microbiol*. **3,** 489-98.
21. Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S. & Banfield, J. F. (2004) *Nature* **428,** 37-43.
22. Hooper, L. V., Wong, M. H., Thelin, A., Hansson, L., Falk, P. G. & Gordon, J. I. (2001) *Science* **291,** 881-84.
23. Stappenbeck, T.S., Hooper, L.V., Manchester, J.K., Wong, M.H., & Gordon, J.I. (2002) *Methods in Enzymology,* **356,** 168-96.
24. Raghunathan, A., Ferguson, H. R., Bornarth, C. J., Song, W., Driscoll, M. & Lasken, R. (2005) *Appl. Environ. Microbiol*. **71**, 3342-47.
25. Claros, M. G. & von Heijne, G. (1994) *Comput. Appl. Biosci.* **10,** 685-86.
26. J. D. Bendtsen, H. Nielsen, G. von Heijne,& S. Brunak. (2004) *J Mol Biol* **340**, 783-95.
27. J. L. Gardy M. R. Laird, F. Chen, S. Rey, C. J. Walsh, M. B. Ester, & Brinkman, F. S. (2005). *Bioinformatics* **21**, 617-23.
28. Holmes, I. & Bruno, W. J. (2001) *Bioinformatics* **17,** 803-20.