A Catalog of Reference Genomes from the Human Microbiome

The Human Microbiome Jumpstart Reference Strains Consortium*†

The human microbiome refers to the community of microorganisms, including prokaryotes, viruses, and microbial eukaryotes, that populate the human body. The National Institutes of Health launched an initiative that focuses on describing the diversity of microbial species that are associated with health and disease. The first phase of this initiative includes the sequencing of hundreds of microbial reference genomes, coupled to metagenomic sequencing from multiple body sites. Here we present results from an initial reference genome sequencing of 178 microbial genomes. From 547,968 predicted polypeptides that correspond to the gene complement of these strains, previously unidentified ("novel") polypeptides that had both unmasked sequence length greater than 100 amino acids and no BLASTP match to any nonreference entry in the nonredundant subset were defined. This analysis resulted in a set of 30,867 polypeptides, of which 29,987 (~97%) were unique. In addition, this set of microbial genomes allows for ~40% of random sequences from the microbiome of the gastrointestinal tract to be associated with organisms based on the match criteria used. Insights into pan-genome analysis suggest that we are still far from saturating microbial species genetic data sets. In addition, the associated metrics and standards used by our group for quality assurance are presented.

The human microbiome is the enormous community of microorganisms occupying the habitats of the human body. Different microbial communities are found in each of the varied environments of human anatomy. The aggregate microbial gene tally surpasses that of the human genome by orders of magnitude. Understanding the relationship of the microbial content to human health and disease is one of the primary goals of human microbiome studies. Determining the structure and function of any microbial community requires a detailed definition of the genomes that it encompasses and the prediction and annotation of their genes.

In 2007, the National Institutes of Health (NIH) initiated the Human Microbiome Project (HMP) as one of its Roadmap initiatives (*1*) to provide resources and build the research infrastructure. One component of the HMP is the production of reference genome sequences for at least 900 bacteria from the human microbiome, which will catalog the microbial genome sequences from the human body and aid researchers conducting human metagenomic sequencing in assigning species to sequences in their metagenomic data sets.

The HMP catalog of reference sequences is being produced by the NIH HMP Jumpstart Consortium of four genome centers: the Baylor College of Medicine Human Genome Sequencing Center, the Broad Institute, the J. Craig Venter Institute, and the Genome Center at Washington University. The challenges for the Jumpstart Consortium include selecting strains to sequence and identifying sources, creating standards for sequencing and annotation to ensure consistency and quality, and the rapid release of information to the community.

Reference genome progress. To date, 356 genomes, including 117 genomes at various stages of upgrading, have been produced by the Jumpstart Consortium and released into public databases. At the time of manuscript preparation, 178 had been completely annotated and are presented in the analysis here. The process for the selection of these strains is described in (*2*). The strains sequenced to date are distributed among body sites as follows: gastrointestinal tract (151), oral cavity (28), urogenital/vaginal tract (33), skin (18), and respiratory tract (8). They also include one isolate from blood (*3*). These are the five major body sites targeted by the HMP.

The broad phylogenetic distribution of the sequenced strains is presented in Fig. 1, which represents a 16*S* ribosomal RNA (rRNA) overlay of HMP-sequenced genomes on 16*S* rRNA sequences from cultured organisms with sequenced genomes (*4*). HMP-sequenced genomes represent two kingdoms (Bacteria and Archaea), nine phyla, 18 classes, and 24 orders. Additional rRNA overlay figures broken down by individual body sites are available in (*5*).

To obtain high-quality draft genomes and a meaningful gene list, minimum standards were defined for the assembly and annotation of draft genomes. Three reference bacterial genome assemblies were evaluated for efficacy of gene predictions and genome completeness. Based on the analysis, metrics for assembly characteristics and annotation characteristics were defined [for more details, see (*2*)]. The quality of HMP genome assemblies is summarized in Table 1 and exceeds the Jumpstart Consortium standards described in (*2*), with the exception of some genomes produced before the standards were in place.

Genome improvement. As described in (*2*), there are justifications for upgrading these high-quality draft assemblies. The Jumpstart Consortium has completed initial improvement work on 26 bacterial genomes that differed significantly with respect to GC content and assembly metrics to explore the effort required and resulting benefits (Fig. 2). The average contig N50 increased 3.63-fold, from 109 kb at draft to 396 kb after improvement. *Bacteroides pectinophilus* displays substantial improvement in N50, from 163 kb in the draft sequence to 862 kb after improvement. *Lactobacillus reuteri* illustrates the opposite extreme, with improvement leading to a smaller contig N50 change, 56 kb to 72 kb. As more genomes improve and some graduate to higher levels of improvement, the assembly state or group of states most useful to the HMP scientific goals will be evaluated.

Pan-genome analysis. A bacterial species' pan-genome can be described as the sum of the core genes shared among all sequenced members of the species and the dispensable genes, or those genes unique to one or more strains studied. To start addressing questions about pan-genomes, we identified all species within our sequenced reference genome catalog for which there was more than one sequenced and annotated genome. Of the nine species identified, four of them have five or more annotated genomes that were generated either by the HMP or by external projects publicly available at the National Center for Biotechnology Information (NCBI); five genomes is the minimum number for which a curve can reliably be fit to pan-genome data. These are *L. reuteri*, *Bifidobacterium longum*, *Enterococcus faecalis*, and *Staphylococcus aureus*. The genomic data used for the analysis consisted of both complete and draft genomes, the only requirement being that >90% of the genome be represented in the available annotated contigs or scaffolds.

Pan-genome curves (*6*) of the gastrointestinal tract isolates *L. reuteri*, *B. longum*, and *E. faecalis* (figs. S3 to S5) are consistent with an open pan-genome model, suggesting that more genome sequencing needs to be undertaken to characterize the actual makeup of the species as a whole. Preliminary results suggest core genome sizes of approximately 1430 genes, 1800 genes, and 1600 genes for *B. longum*, *E. faecalis*, and *L. reuteri*, respectively. Based on the current core gene plots, *L. reuteri* (fig. S3) appears to be approaching a closed pan-genome model, with newly sequenced strains contributing very small numbers of new genes to the pan-genome; however, we see an interesting community substructure within this species. Our current *L. reuteri* pan-genome analysis of seven isolates suggests that four of the

*All authors with their affiliations and contributions are listed at the end of this paper.
†To whom correspondence should be addressed. E-mail: kenelson@jcvi.org

seven currently sequenced isolates are very similar to one another, contributing zero to two new genes to the pan-genome. Two further strains are also similar to one another, each contributing an intermediate number of new genes (~15 to 30), whereas one outlier strain contributes a distinct set of genes (~330). These findings are consistent with the comparison of average nucleotide identity with gene content discussed below for this species. It will be interesting to see whether additional sequencing of this species identifies other subgroups in addition to the three identified here, or whether this sample set is in fact largely representative of the species.

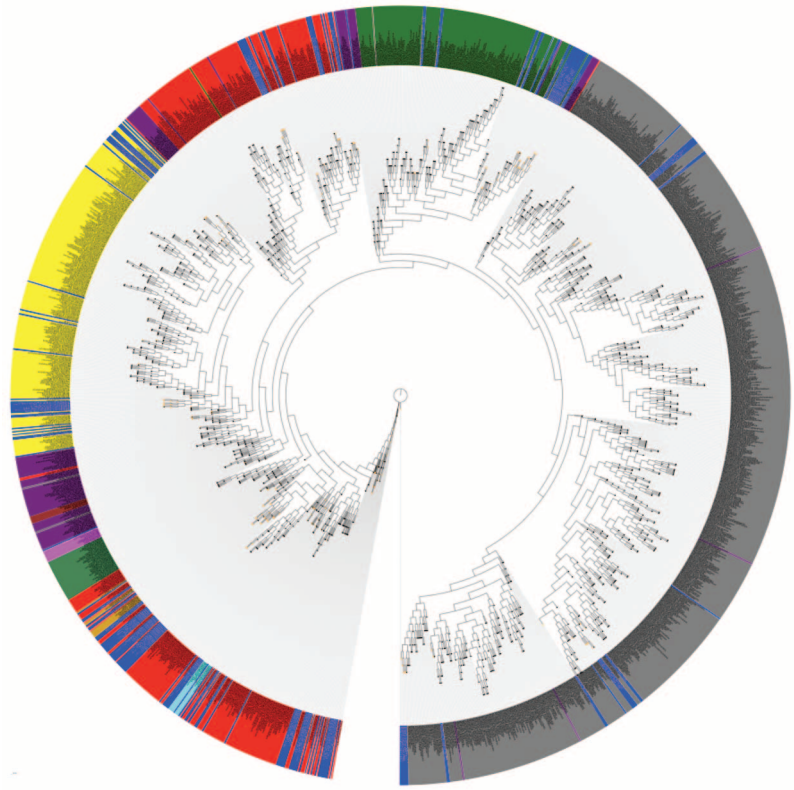

Fig. 1. Phylogenetic tree of 16*S* rDNA sequences. The tree was created using ~1500 16*S* rDNAs representing single species. Organisms sequenced as part of the HMP project are highlighted in blue. Additional coloring indicates separation by phylum: yellow, Actinobacteria; dark green, Bacteroidetes; light green, Cyanobacteria; red, Firmicutes; cyan, Fusobacteria; dark red, Planctomycetes; gray, Proteobacteria; magenta, Spirochaetes; light pink, TM7; tan, Tenericutes. The purpose of this analysis is not the details of the branching structure (which include minor known artifacts), but the overall distribution of the HMP strains (in blue) around the tree of life.

Similar findings for *B. longum* (fig. S4) suggest that four of the five currently sequenced genomes contribute approximately equally to the pan-genome (~50 to 150), with one outlier strain (ATCC 15697) contributing a much higher number of previously unidentified ("novel") genes (~640). These data are consistent with differences in gene count across these genomes. Each of the five currently sequenced genomes of *E. faecalis* (fig. S5) contributes approximately equivalent numbers of new genes to the pan-genome. Our current data sets for these two species are still too small to determine whether we can realistically achieve a closed pan-genome, with newly sequenced isolates contributing on the order of 100 new genes each. It is unrealistic at this point to extrapolate how many additional genomes would need to be sequenced to determine whether the number of new genes contributed by each new sequence continues to plateau around 100 new genes or approaches zero.

S. aureus pan-genome plots (fig. S6), representing isolates collected from the skin, urogenital tract, and mucus membranes of mammals (human and bovine), are consistent with a closed pan-genome model, as previously suggested (*7*), with an estimated core size of 2295 genes and an estimated pan-genome size of ~3200 genes.

We performed a preliminary survey looking into the functions encoded by those genes that are unique to new gene data sets and not found in the core data set, based on gene product annotation and Enzyme Commission (EC) numbers, when available. These genomes underwent automated annotation only, with no manual curation, so any trends seen should be considered putative. Across all four species, the number of novel genes annotated only as hypothetical or a conserved do-

Table 1. Draft assembly metrics, organized by finishing status and based on current assignments. Draft corresponds to standard or high-quality draft sequences, with no additional automated or manual attempts to improve assembly, beyond ensuring exclusion of contaminating sequence. Improved columns correspond to improved high-quality draft submission. None of the reference genomes has been improved beyond this grade at this point. n/a, not applicable.

		Draft Number of strains = 133			Improved Number of strains = 45		
Metric	Passing standard	Pass %	Mean	Range	Pass %	Mean	Range
Percent of genome included in contigs*	>90%	100%	98.23%	95.1–99.9%	100%	99.91%	98.6–100%
Percent of bases greater than 5× read coverage†	>90%	99%	98.90%	80.8–100%	100%	99.35%	98.8–99.6%
Contig N50	>5 kb	100%	102.61 kb	11.12–861.67 kb	100%	517.92 kb	58.03–3472.99 kb
Contig N75	n/a	99%	54.82 kb	4.97–556.76 kb	100%	340.20 kb	30.56–2635.77 kb
Contig N90	n/a	90%	25.54 kb	2.01–240.69 kb	100%	211.51 kb	14.96–2635.77 kb
Scaffold N50*	>20 KB	100%	883.93 kb	50.56–3356.77 kb	100%	606.77 kb	91.71–2898.42 kb
Scaffold N75*	n/a	100%	511.35 kb	24.31–3237.97 kb	100%	378.22 kb	52.32–2391.23 kb
Scaffold N90*	n/a	99%	282.14 kb	11.74–2490.47 kb	100%	226.24 kb	28.67–2391.23 kb
Average contig length	>5 kb	100%	31.52 kb	5.62–180.70 kb	100%	174.70 kb	23.26–1321.04 kb
Percent of core genes present in gene list	>90%	99%	99.63%	86.4–100%	100%	99.90%	98.5–100%

*Calculated only for strains with scaffold assemblies submitted to NCBI. The number of strains with scaffold assemblies, by grade: draft, 74; improved. 37. †Per-base coverage not available for all reads, for example, those with some draft level of sequencing before the Jumpstart initiative or strains where a combination of technologies was used. The number of strains with per-base read coverage: draft, 121; improved, 4.

main of unknown function ranged from 66 to 73%, making up the bulk of the novel genes identified by the pan-genome analysis. Another predominant trend seen was unique family members corresponding to non-novel functions; for example, functions also identified in the core data set.

There are a number of interesting categories of functions identified in novel gene sets that are unique to individual strains. These include accessory proteins involved in the activation of urease; a virulence factor found in microorganisms associated with gastric ulceration, among other human health concerns (*8*); phage morphogenesis and regulation proteins; and small numbers of enzymes involved in the metabolism of sugars and amino acids. Further work is needed to clean up annotations and to provide more consistent EC number assignments in order to confirm and build on trends seen in this preliminary analysis. The HMP Data Analysis and Coordination Center (DACC) has been given the mandate of adding value and updating annotations, which will allow for the expansion of these analyses throughout this project.

Measuring diversity within genera. The genomic diversity among strains belonging to the same genus was explored by a measure of the evolutionary relatedness and gene content similarity in a pairwise fashion (Fig. 3). The average nucleotide identity (ANI) is a measure of evolutionary relatedness based on sequence similarity between the set of shared genes (*9*). The measure of gene content similarity between two strains can provide a sense of functional or ecological relatedness, and one might predict that strains with a lower gene content similarity are more likely to be found in different habitats. The three genera selected for this comparison all contain at least 16 strains and include *Lactobacillus* (36 strains; Fig. 3), *Bifidobacterium* (16 strains; fig. S7), and *Bacteroides* (21 strains; fig. S8). Genomes contributed by the HMP as well as those available in public databases were included in this analysis. High intraspecies diversity was observed within genera in addition to interspecies diversity. Within *Lactobacillus*, several species showed significant diversity. For example, *L. reuteri* is represented by two main groups; one set (bottom left blue oval in Fig. 3) contains seven different strains. Among the strains within that group, the percent of ANI (%ANI) and percent of gene content are above 98 and 90%, respectively. In the second group (upper right blue oval in Fig. 3), the % ANI ranges between 96 and 93%, with a gene content similarity lower than 78%. Previously, a value of 95% ANI was shown to correspond with the recommended cutoff of 70% DNA-DNA reassociation for species delineation (*10*). This indicates that the *L. reuteri* strains obtained within the framework of the HMP significantly increased the known genomic diversity of this named species, as was also demonstrated by the pan-genome analysis. Other strains showing large intraspecies diversity belong to *L. johnsonii* and *L. gasseri*.

Among the strains of *B. longum* (fig. S7), four (two of which were contributed by the HMP) have pairwise %ANI values at the higher end of the spectrum, ranging between 96 and 98%, but with relatively low gene-content similarity (that is, below 82%), indicating a broad range in gene complements. One additional existing strain (ATCC 15697) has a %ANI below 95% and a gene content similarity below 65% and is therefore a clear evolutionary and ecological outlier.

The analysis of *Bacteroides* genomes has revealed several close common ancestries. *Bacteroides* sp. D4 and 9_1_42FAA are closely related to *Ba. dorei* (ANI > 95%), but still have a significant gene content difference, lower than 78% similarity. This suggests that the *Bacteroides* group may have many closely related, yet ecologically distinct lineages.

Novel genes. The 547,968 predicted polypeptides corresponding to the entire annotated gene complement of these strains [of which 516,631 (94%) were unique] were searched against the bacterial and viral divisions of NCBI's nonredundant (nr) protein database using WU-BLASTP as described in (*5*). Each polypeptide was also compared to a merged database of TIGRFAM and Pfam hidden Markov models (HMMs) using version 2a of the HMMER3 package. A set of candidate novel polypeptides was defined by selecting those that had both of the following conditions: (i) unmasked sequence length >100 amino acids and (ii) no BLASTP match to any nonreference entry in the nr subset. This analysis resulted in a set of 30,867 polypeptides, 5.6% of the total, of which 29,987 (~97%) were unique (*2*). Clustering this set with CD-HIT (*11*) resulted in 29,286 unique polypeptides at 98% sequence identity (~5% reduction), 28,857 polypeptides at 95% (~7% reduction), and 28,469 at 90% (~8% reduction). An alternate set of candidate novel polypeptides was also defined by modifying condition (i) above to filter on the number of bases not identified as low-complexity sequences by the SEG algorithm (*12*) (that is, the sequence length after removing all SEG-masked bases). This alternate initial set contains 28,693 polypeptides.

The above criteria were chosen by inspecting histograms of novel versus non-novel polypeptide counts at various expectation (E)-value and sequence-length thresholds and selecting cutoffs

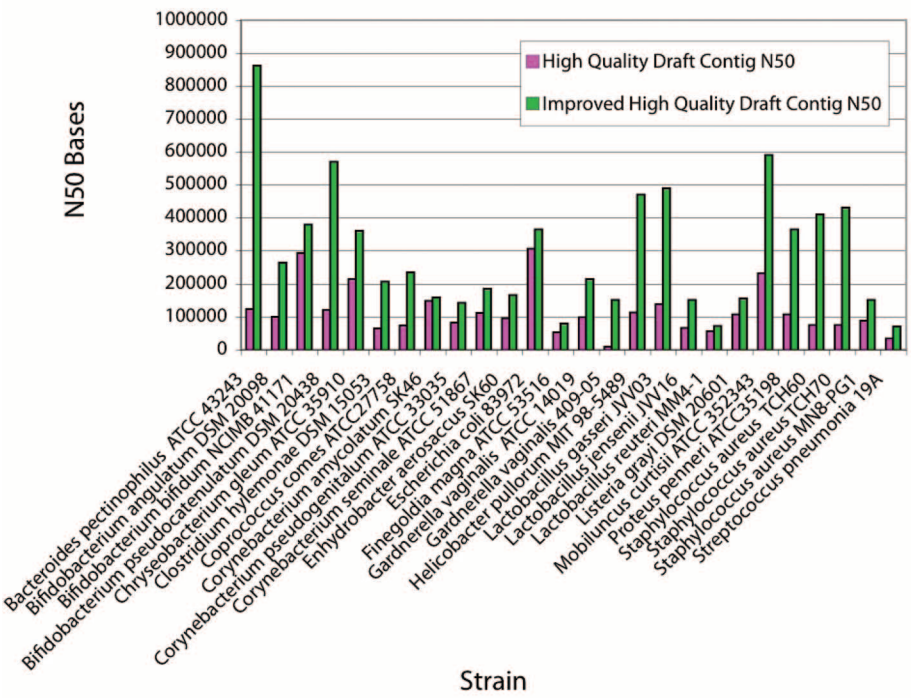

Fig. 2. Contig N50 comparison for 26 draft and improved genomes. High-quality draft contig N50 bases are shown in magenta, and improved high-quality draft sequences are shown in green. These data represent the variety of approaches from the four data-generation centers. The majority of shotgun data was produced on the Roche-454 platform, although some assemblies include paired Sanger reads to improve contiguity. All draft assemblies are based on the Roche-Newbler assembler, although some of the improved assemblies are based on Parallel Contig Assembly Program (PCAP) (*23*) and the Celera Assembler (*24*) due to existing integration with finishing and improvement pipelines. Additional variation comes from the improvement approach. Directed Sanger reads from gap-spanning polymerase chain reaction amplicons served as the primary approach, whereas some assemblies have been subjected only to manipulation of the shotgun data, making unrealized joins, removing poor-quality data, and placing unincorporated shotgun reads.

that seemed likely to minimize the number of false positives while not excluding too many true positives. The distribution of novel versus non-novel polypeptide counts overlaps at all E-value thresholds, making it impossible to pick a cutoff that does not exclude any true positives. Therefore a relatively long (100 amino acids) length threshold was selected to try to minimize noise or false positives, at the possible cost of losing some real novel polypeptides.

With ~1300 completely sequenced bacterial genomes in GenBank (*13*), the observation that 5% of the genes annotated in the HMP genomes satisfy criteria for novelty underscores the remarkable diversity of bacterial proteins. To assess whether there is enriched novelty in the HMP-targeted genomes as compared with previously sequenced prokaryotic genomes, we randomly selected 178 previously sequenced draft genomes from GenBank and ran the same analysis for comparison. This data set resulted in 747,522 predicted polypeptides, of which 568,426 were unique. Of these, 14,269 polypeptides met our criteria for novelty, 1.9% of the total, of which 14,064 were unique, 2.5% of the unique total. Clustering resulted in a 2% reduction at 98% and a 3% reduction at 90%, indicating that this data set does not contain as many highly similar protein predictions as the HMP novel set does. This suggests that there is enrichment in novelty in the HMP data set of approximately 2:1 over the random data set. Whereas the human microbiome is generally thought to be less complex than that of soils and certain other environmental microbiomes, it nevertheless clearly houses enormous microbial diversity yet to be described.

Analysis of metagenomic shotgun data. Because the HMP reference genomes that were sequenced had been selected primarily because they were isolates from humans and had not been identified as strains seen in metagenomics studies, it was not known how much these genomes would help to identify metagenomic sequences that were obtained from human microbial communities. The most useful reference genomes should expand our ability to interpret metagenomic data. We also used the stringent fragment recruitment technique (*14*) to compare metagenomic sequencing data to the reference genomes in nucleotide space (*15*). The stringency of this approach generally limits the recruitment of metagenomic reads to organisms within the same genus, but it can resolve strain-specific differences.

Publicly available metagenomic data sets from two human gastrointestinal studies were used in this analysis (*16*, *17*), along with 454 reads from a Washington University data set (which contributed the bulk of the 16.8 million reads that were tested). The reference genomes included 866 complete and 913 draft genomes available at NCBI, including the HMP reference genomes with sequence reads available at the time of analysis. In total, 62 HMP genomes showed significant levels of recruitment with 11.3 million metagenomic reads recruited (66% of all reads). Of these, a significant 6.9 million reads (41%) recruited best to the HMP reference genomes, based on the global percent identity (defined as the number of identities between read and reference, divided by the length of the read). A read is considered to be a best hit to a HMP genome if the best global percent identity includes a match to an HMP genome. Many of these reads would not have been recruited if the HMP reference genomes were not available: Between 20 and 40% of the reads were recruited only because of the presence of the HMP genomes.

These results show that a significant number of the genomes sequenced as part of the HMP project are directly adding to our understanding the human microbiome. These results also show that specific genomes are useful references across a wide range of individuals despite the strain-specific diversity noted above. Despite the large number of genomes available, a significant amount of the metagenome (33%) is still not well represented by any reference genome. It is likely that the 900-genomes target of the HMP will reduce this number of unidentified reads further without redundancy in genome selection. It should be noted that this analysis focused on the gastrointestinal tract, and it is likely that additional genomes exist in other body sites; thus, the composition of the 900 genomes should address these organisms.

Data release, future plans, and conclusions. The Jumpstart Centers have made substantial progress in generating a set of reference genomes that describe the human microbiome. We have

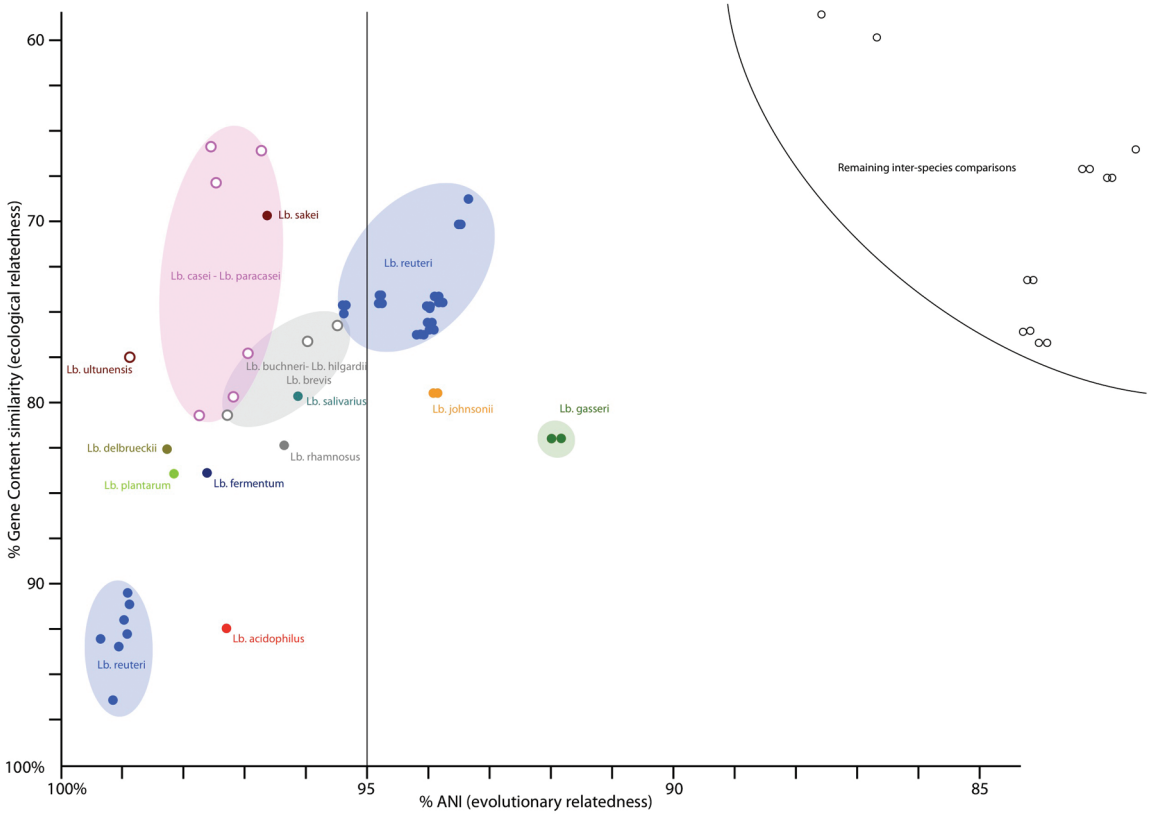

Fig. 3. Interstrain diversity among *Lactobacillus* genomes. Each point represents a whole-genome comparison between two *Lactobacillus* genomes and shows the %ANI on the *x* axis as a measure of evolutionary distance, plotted against the percentage of gene content similarity on the *y* axis. Only comparisons with ANI values above 85% are shown. The vertical line at 95% corresponds to a recommended cutoff of 70% DNA-DNA reassociation for species delineation. Different intra- and interspecies comparisons are color coded, with solid or open circles respectively, and labeled with their given taxonomical name in the corresponding color. Colored ovals assist in identifying related data points belonging to a single named species.

made every effort to ensure that all strains are available in public repositories and to release these genomes and their associated data, assemblies, and annotations in accordance with NIH policy (*18*). In addition, all data and standard operating procedures are available through the DACC (*19*), where we welcome community input and feedback.

Human microbiome research groups from around the world have launched an International Human Microbiome Consortium (IHMC), which together will sequence more than 1000 human microbial bacterial reference genomes. These include the 900 reference strains that are being sequenced by the HMP Jumpstart Centers, 100 genomes sequenced as part of the European Union–funded MetaHIT project (*20*), and additional genomes produced by international efforts. All of these strains appear on the DACC. Other strains are being sequenced as part of the Department of Energy Genomic Encyclopedia of Bacteria and Archaea (GEBA) (*21*, *22*) project.

Nevertheless, the human microbiome is much more complex than this set of genomes and is likely to exceed it by orders of magnitude. In addition to the large number of cultured strains, many unculturable strains remain to be defined, and substantial intraspecies diversity still needs to be described. Thus, this initial effort is only a beginning, but it is valuable, and not only does it contribute to the catalog of reference strains, but it also builds infrastructure for strain selection and acquisition, develops methods for sequencing unculturables, defines standards for the various deliverables, provides online access to the large new data set, and addresses many other issues.

The development of standards that will be applied to the 900 genomes that are being sequenced will provide a new and higher level of uniformity to microbial genome data. The Jumpstart Consortium members are also in discussion with other consortia that are interested in standards to extend this uniformity beyond the HMP.

This report and the initial stage of the HMP focus on bacteria, but this effort is currently being expanded to produce reference genomes for eukaryotic microbes and viruses. These other components of the human microbiome have not been forgotten, but the initial focus on bacteria has allowed necessary infrastructure to be developed for the large task ahead, which can now be readily deployed for other organisms. It is our ultimate goal to sample the human microbiome as completely as possible.

References and Notes

1. The NIH Common Fund Human Microbiome Project, Division of Program Coordination, Planning and Strategic Initiatives, NIH, U.S. Department of Health and Human Services, http://nihroadmap.nih.gov/hmp/.
2. Materials and methods are available as supporting material on *Science* Online.
3. HMP Project Catalog, Human Microbiome Project Data Analysis Coordinating Center, www.hmpdacc.org/project_catalog.html.
4. 16*S* rDNA for cultured bacteria, http://bioinfo.unice.fr/blast/documentation/alphabetical_list.html.
5. Reference genomes of the Human Microbiome Project, Human Microbiome Project Data Analysis Coordinating Center, http://hmpdacc.org/reference_genomes.php.
6. H. Tettelin *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 13950 (2005).
7. H. Tettelin, D. Riley, C. Cattuto, D. Medini, *Curr. Opin. Microbiol.* **11**, 472 (2008).
8. H. L. Mobley, M. D. Island, R. P. Hausinger, *Microbiol. Rev.* **59**, 451 (1995).
9. K. T. Konstantinidis, A. Ramette, J. M. Tiedje, *Appl. Environ. Microbiol.* **72**, 7286 (2006).
10. J. Goris *et al.*, *Int. J. Syst. Evol. Microbiol.* **57**, 81 (2007).
11. W. Li, J. C. Wooley, A. Godzik, D. Jones, *PLoS ONE* **3**, e3375 (2008).
12. J. C. Wootton, S. Federhen, *Comput. Chem.* **17**, 149 (1993).
13. National Center for Biotechnology Information, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/.
14. D. B. Rusch *et al.*, *PLoS Biol.* **5**, e77 (2007).
15. Genome selection page organized by coverage, J. Craig Venter Institute, http://gos.jcvi.org/users/hmpGenomes/genomes.html.
16. P. J. Turnbaugh *et al.*, *Nature* **457**, 480 (2009).
17. S. R. Gill *et al.*, *Science* **312**, 1355 (2006).
18. M. Y. Giovanni, *Genome Sequencing Centers NIAID Data and Reagent Sharing and Release Guidelines*, National Institute of Allergy and Infectious Diseases, NIH, U.S. Department of Health and Human Services, www.niaid.nih.gov/dmid/genomes/mscs/data_release.htm.
19. Documentation and SOPs, Human Microbiome Project Data Analysis Coordinating Center, www.hmpdacc.org/sops.php.
20. J. Qin *et al.*, *Nature* **464**, 59 (2010).
21. A Genomic Encyclopedia of Bacteria and Archaea (GEBA), Joint Genome Institute, U.S. Department of Energy Office of Science, www.jgi.doe.gov/programs/GEBA/.
22. D. Wu *et al.*, *Nature* **462**, 1056 (2009).
23. X. Huang, J. Wang, S. Aluru, S. P. Yang, L. Hillier, *Genome Res.* **13**, 2164 (2003).
24. J. Miller *et al.*, *Bioinformatics* **24**, 2818 (2008).

25. The authors gratefully acknowledge J. Warren, J. Zhang, R. G. Fowler, P. Pham, D. Haft, J. Selengut, T. Davidsen, P. Goetz, D. Harkins, S. Shrivastava, S. Koren, B. Walenz, L. Foster, I. Singh, Y.-h. Rogers, and the J. Craig Venter Institute Joint Technology Center. We thank J. Xu, S.-P. Yang, and S. Schobel for bioinformatics support; the Broad Genome Sequencing Platform, Y. Han, V. Korchina, M. Scheel, R. Thornton and the BCM-HGSC production team, L. Courntey, C. Fronick, O. Hall, M. O'Laughlin, M. Cunningham, D. O'Brien, B. Theising, and the GCWU production team for sequencing; J. Gordon, F. Dewhirst, B. Wilson, B. White, R. Mandrell, M. Blaser, R. H. Stevens, S. Hillier, Y. Liu, Z. Shen, D. Schauer, J. Fox, M. Allison, C. D. Sibley, D. M. Saulnier, and G. R. Gibson for providing strains; and M. Y. Giovanni, C. L. Baker, V. Bonazzi, C. D. Deal, S. Garges, R. W. Karp, R. W. Lunsford, J. Peterson, M. Wright, T. T. Belachew, and C. R. Wellington for funding agency management. We acknowledge NIH for funding this project with grants to the J. Craig Venter Institute (grants N01 AI 30071 and U54-AI084844), Washington University (grants U54-HG003079 and U54-HG004968), Baylor College of Medicine (grants U54-HG003273 and U54-HG004973), and the Broad Institute (grants HHSN272200900017C and U54-HG004969). Funding for E.A.-V. was from the Crohn's and Colitis Foundation of Canada; D.G. had secondary affiliation at the Laboratory of Microbiology (WE 10), Department of Biochemistry and Microbiology, Faculty of Sciences, Ghent University, KL Ledeganckstraat 35, 9000 Ghent, Belgium, and is indebted to the Fund for Scientific Research, Flanders (Belgium), for a postdoctoral fellowship and research funding for the duration of this project; M.S. acknowledges the Canadian Cystic Fibrosis Foundation and the Canadian Institutes of Health Research for funding of his research for this project.

The Human Microbiome Jumpstart Reference Strains Consortium

Manuscript preparation: Karen E. Nelson,[1]† George M. Weinstock,[2] Sarah K. Highlander,[3,4] Kim C. Worley,[3,5] Heather Huot Creasy,[6] Jennifer Russo Wortman,[7,6] Douglas B. Rusch,[8] Makedonka Mitreva,[2] Erica Sodergren,[2] Asif T. Chinwalla,[2] Michael Feldgarden,[9] Dirk Gevers,[9] Brian J. Haas,[9] Ramana Madupu,[8] Doyle V. Ward[9]

Principal investigators: Bruce W. Birren,[9] Richard A. Gibbs,[3,5] Sarah K. Highlander,[3,4] Barbara Methe,[1] Karen E. Nelson,[1] Joseph F. Petrosino,[3,4] Robert L. Strausberg,[1] Granger G. Sutton,[8] George M. Weinstock,[2] Owen R. White,[10,6] Richard K. Wilson[2]

Annotation: Asif T. Chinwalla,[2] Heather Huot Creasy,[6] Scott Durkin,[8] Michelle Gwinn Giglio,[6] Sharvari Gujja,[9] Brian J. Haas,[9] Sarah K. Highlander,[3,4] Clint Howarth,[9] Chinnappa D. Kodira,[11] Nikos Kyrpides,[12] Ramana Madupu,[8] Teena Mehta,[9] Makedonka Mitreva,[9] Donna M. Muzny,[3,5] Matthew Pearson,[9] Kymberlie Pepin,[2] Amrita Pati,[12] Xiang Qin,[3,5] Kim C. Worley,[3,5] Jennifer Russo Wortman,[7,6] Chandri Yandava,[9] Qiandong Zeng,[9] Lan Zhang[3,5]

Assembly: Aaron M. Berlin,[9] Lei Chen,[2] Theresa A. Hepburn,[9] Justin Johnson,[8] Jamison McCorrison,[8] Jason Miller,[8] Pat Minx,[2] Donna M. Muzny,[3,5] Chad Nusbaum,[9] Xiang Qin,[3,5] Carsten Russ,[9] Granger G. Sutton,[8] Sean M. Sykes,[9] Chad M. Tomlinson,[2] Sarah Young,[9] Wesley C. Warren,[2] Kim C. Worley[3,5]

Data analysis: Jonathan Badger,[13] Jonathan Crabtree,[6] Heather Huot Creasy,[6] Michael Feldgarden,[9] Dirk Gevers,[9] Sarah K. Highlander,[3,4] Ramana Madupu,[8] Victor M. Markowitz,[14] Makedonka Mitreva,[2] Donna M. Muzny,[3,5] Joshua Orvis,[6] Joseph F. Petrosino[3,4] Douglas B. Rusch,[8] Granger G. Sutton,[8] Doyle V. Ward,[9] Kim C. Worley,[3,5] Jennifer Russo Wortman[7,6]

DNA sequence production: Andrew Cree,[3,5] Steve Ferriera,[15] Lucinda L. Fulton,[2] Robert S. Fulton,[2] Marcus Gillis,[1] Lisa D. Hemphill,[3,5] Vandita Joshi,[3,5] Christie Kovar,[3,5] Donna M. Muzny,[3,5] Manolito Torralba,[1] Xiang Qin[3,5]

Funding agency management: Kris A. Wetterstrand[16]

Genome improvement: Amr Abouellleil,[9] Aye M. Wollam,[2] Christian J. Buhay,[3,5] Yan Ding,[3,5] Shannon Dugan,[3,5] Michael G. FitzGerald,[9] Lucinda L. Fulton,[2] Robert S. Fulton,[2] Mike Holder,[3,5] Jessica Hostetler,[1] Ramana Madupu,[8] Donna M. Muzny,[3,5] Xiang Qin,[3,5] Granger G. Sutton[8]

Project leadership: Bruce W. Birren,[9] Sandra W. Clifton,[2] Sarah K. Highlander,[3,4] Karen E. Nelson,[1] Joseph F. Petrosino,[3,4] Erica Sodergren,[2] Robert L. Strausberg,[1] Granger G. Sutton,[8] George M. Weinstock,[2] Owen R. White[10,6]

Strain management: Emma Allen-Vercoe,[17] Jonathan Badger,[13] Sandra W. Clifton,[2] Heather Huot Creasy,[6] Ashlee M. Earl,[9] Candace N. Farmer,[2] Michelle Gwinn Giglio,[6] Marcus Gillis,[1] Sarah K. Highlander,[3,4] Konstantinos Liolios,[12] Karen E. Nelson,[1] Erica Sodergren,[2] Michael G. Surette,[18] Granger G. Sutton,[8] Manolito Torralba,[1] Doyle V. Ward,[9] George M. Weinstock,[2] Jennifer Russo Wortman,[7,6] Qiang Xu[19]

Submissions: Asif T. Chinwalla,[2] Craig Pohl,[2] Scott Durkin,[8] Granger G. Sutton,[8] Katarzyna Wilczek-Boney,[3,5] Dianhui Zhu[3,5]

[1]Human Genomic Medicine, J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850, USA. [2]The Genome Center, Washington University School of Medicine, 4444 Forest Park Avenue, St. Louis, MO 63108, USA. [3]Human Genome Sequencing Center, Baylor College of Medicine, BCM226, One Baylor Plaza, Houston, TX 77030, USA. [4]Department of Molecular Virology and Microbiology, BCM280, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. [5]Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. [6]Institute for Genome Sciences, University of Maryland School of Medicine, 801 West Baltimore Street, Baltimore, MD 21201, USA. [7]Department of Medicine, University of Maryland School of Medicine, Department of Genetics, 801 West Baltimore Street, Baltimore, MD 21201, USA. [8]Bioinformatics, J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850, USA. [9]Genome Sequencing and Analysis Program, Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA. [10]Department of Epidemiology and Preventive Medicine, University of Maryland School of Medicine, 801 West Baltimore

Street, Baltimore, MD 21201, USA. [11]Genome Sequencing and Analysis Program, 454 Sequencing, 15 Commercial Street, Branford, CT 06405, USA. [12]Department of Energy-Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA. [13]Microbial and Environmental Genomics, J. Craig Venter Institute, 10355 Science Center Drive, La Jolla, CA 92121, USA. [14]Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [15]Sequencing, J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850, USA. [16]National Human Genome Research Institute, 5635 Fishers Lane, Bethesda, MD 20892, USA. [17]Molecular and Cellular Biology, University of Guelph, 50 Stone Road, Guelph, Ontario N1G 2W1, Canada. [18]Microbiology and Infectious Diseases, University of Calgary, 3330 Hospital Drive, Calgary, Alberta T2N4N1, Canada. [19]Osel Inc., 4008 Burton Drive, Santa Clara, CA 95054, USA.

Supporting Online Material
www.sciencemag.org/cgi/content/full/328/5981/994/DC1
Materials and Methods
Figs. S1 to S8
References and Notes

20 October 2009; accepted 31 March 2010
10.1126/science.1183605

Observation of Plasmarons in Quasi-Freestanding Doped Graphene

Aaron Bostwick,[1] Florian Speck,[2] Thomas Seyller,[2] Karsten Horn,[3] Marco Polini,[4]* Reza Asgari,[5]* Allan H. MacDonald,[6] Eli Rotenberg[1]†

A hallmark of graphene is its unusual conical band structure that leads to a zero-energy band gap at a single Dirac crossing point. By measuring the spectral function of charge carriers in quasi-freestanding graphene with angle-resolved photoemission spectroscopy, we showed that at finite doping, this well-known linear Dirac spectrum does not provide a full description of the charge-carrying excitations. We observed composite "plasmaron" particles, which are bound states of charge carriers with plasmons, the density oscillations of the graphene electron gas. The Dirac crossing point is resolved into three crossings: the first between pure charge bands, the second between pure plasmaron bands, and the third a ring-shaped crossing between charge and plasmaron bands.

Electrons in metals and semiconductors undergo many complex interactions, and most theoretical treatments make use of the quasiparticle approximation, in which independent electrons are replaced by electron- and hole-like quasiparticles interacting through a dynamically screened Coulomb force. The details of the screening are determined by the valence band structure, but the band energies are modified by the screened interactions. A complex self-energy function describes the energy and lifetime renormalization of the band structure resulting from this interplay.

Bohm and Pines (*1*) accounted for the short-range interactions between quasiparticles through the creation of a polarization cloud formed of virtual electron-hole pairs around each charge carrier, screening each from its neighbors. The long-range interactions manifest themselves through plasmons, which are collective charge density oscillations of the electron gas that can propagate through the medium with their own band-dispersion relation. These plasmons can in turn interact with the charges, leading to strong self-energy effects. Lundqvist predicted the presence of new composite particles called plasmarons, formed by the coupling of the elementary charges with plasmons (*2*). Their distinct energy bands should be observable with the use of angle-resolved photoemission spectroscopy (ARPES), but so far have been observed only by optical (*3*, *4*) and tunneling spectroscopies (*5*), which probe the altered density of states.

[1]Advanced Light Source (ALS), E. O. Lawrence Berkeley Laboratory, MS6-2100, Berkeley, CA 94720, USA. [2]Lehrstuhl für Technische Physik, Universität Erlangen-Nürnberg, Erwin-Rommel-Strasse 1, 91058 Erlangen, Germany. [3]Department of Molecular Physics, Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195 Berlin, Germany. [4]National Enterprise for nanoScience and nanoTechnoloy, Istituto Nanoscienze–Consiglio Nazionale della Ricerche and Scuola Normale Superiore, I-56126 Pisa, Italy. [5]School of Physics, Institute for Research in Fundamental Sciences, Tehran 19395-5531, Iran. [6]Department of Physics, University of Texas at Austin, 1 University Station C1600, Austin, TX 78712,USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: erotenberg@lbl.gov

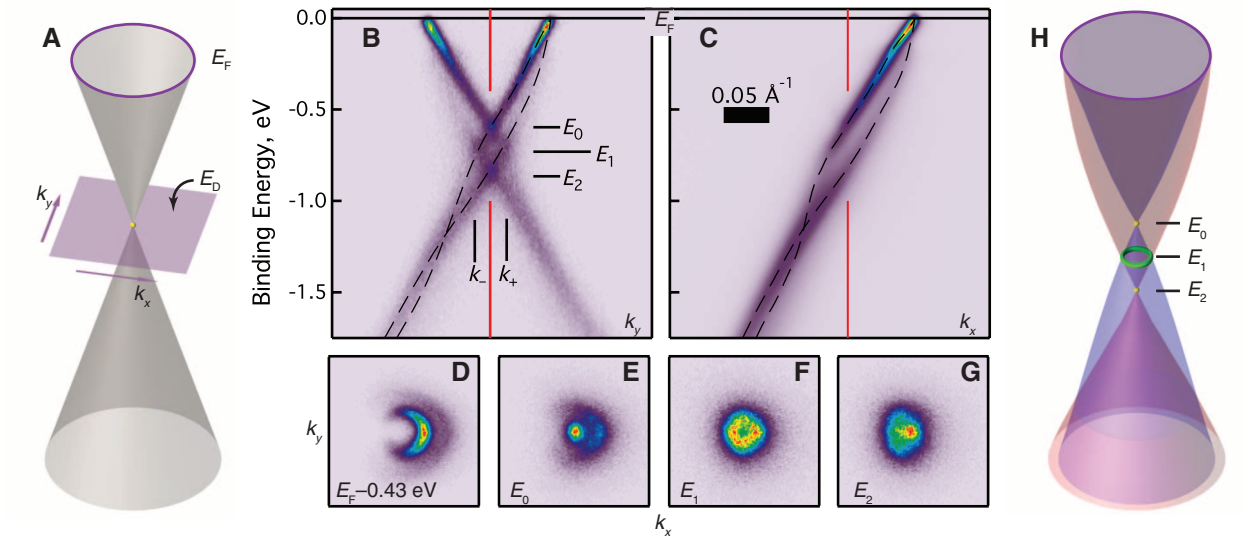

Fig. 1. (**A**) The Dirac energy spectrum of graphene in a non-interacting, single-particle picture. (**B** and **C**) Experimental spectral functions of doped graphene perpendicular and parallel to the ΓK direction of the graphene Brillouin zone. The dashed lines are guides to the dispersion of the observed hole and plasmaron bands. The red lines are at k = 0 (the K point of the graphene Brillouin zone). (**D** to **G**) Constant-energy cuts of the spectral function at different binding energies. (**H**) Schematic Dirac spectrum in the presence of interactions, showing a reconstructed Dirac crossing. The samples used for (B) to (G) were doped to $n = 1.7 \times 10^{13}$ cm^{-2}. The scale bar in (C) defines the momentum length scale in (B) to (G).