Evolution of symbiotic bacteria in the distal human intestine

Jian Xu
Washington University School of Medicine in St. Louis

Michael A. Mahowald
Washington University School of Medicine in St. Louis

Ruth E. Lay
Washington University School of Medicine in St. Louis

Catherine A. Lozupone
University of Colorado at Boulder

Micah Hamady
University of Colorado at Boulder

See next page for additional authors

Follow this and additional works at: https://digitalcommons.wustl.edu/open_access_pubs

Please let us know how this document benefits you.

Recommended Citation
Xu, Jian; Mahowald, Michael A.; Lay, Ruth E.; Lozupone, Catherine A.; Hamady, Micah; Martens, Eric C.; Henriassat, Bernard; Coutinho, Pedro M.; Minx, Patrick; Latreille, Philippe; Cordum, Holland; Van Brunt, Andrew; Kim, Kyung; Fulton, Robert S.; Fulton, Lucinda A.; Clifton, Sandra W.; Wilson, Richard K.; Knight, Robin D.; and Gordon, Jeffrey I., "Evolution of symbiotic bacteria in the distal human intestine." PLoS Biology. 5, 7. e156. (2007).
https://digitalcommons.wustl.edu/open_access_pubs/444

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact vanam@wustl.edu.
Authors
Jian Xu, Michael A. Mahowald, Ruth E. Lay, Catherine A. Lozupone, Micah Hamady, Eric C. Martens, Bernard Henrissat, Pedro M. Coutinho, Patrick Minx, Philippe Latreille, Holland Cordum, Andrew Van Brunt, Kyung Kim, Robert S. Fulton, Lucinda A. Fulton, Sandra W. Clifton, Richard K. Wilson, Robin D. Knight, and Jeffrey I. Gordon
Evolution of Symbiotic Bacteria in the Distal Human Intestine

Jian Xu1,2*, Michael A. Mahowald1*, Ruth E. Ley1, Catherine A. Lozupone3, Micah Hamady4, Eric C. Martens1, Bernard Henriassa5,6, Pedro M. Coutinho5,6, Patrick Minx2, Philippe Latreille2, Holland Cordum7, Andrew Van Brunt2, Kyung Kim2, Robert S. Fulton2, Lucinda A. Fulton2, Sandra W. Clifton2, Richard K. Wilson1,2, Robin D. Knight7, Jeffrey I. Gordon1**

1 Center for Genome Sciences, Washington University School of Medicine, St. Louis, Missouri, United States of America, 2 Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri, United States of America, 3 Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado, United States of America, 4 Department of Computer Science, University of Colorado, Boulder, Colorado, United States of America, 5 Universités Aix-Marseille I and II, Marseille, France, 6 CNRS, UMR6098, Marseille, France, 7 Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado, United States of America.

The adult human intestine contains trillions of bacteria, representing hundreds of species and thousands of subspecies. Little is known about the selective pressures that have shaped and are shaping this community’s component species, which are dominated by members of the Bacteroidetes and Firmicutes divisions. To examine how the intestinal environment affects microbial genome evolution, we have sequenced the genomes of two members of the normal distal human gut microbiota, Bacteroides vulgatus and Bacteroides distasonis, and by comparison with the few other sequenced gut and non-gut Bacteroidetes, analyzed their niche and habitat adaptations. The results show that lateral gene transfer, mobile elements, and gene amplification have played important roles in affecting the ability of gut-dwelling Bacteroidetes to vary their cell surface, sense their environment, and harvest nutrient resources present in the distal intestine. Our findings show that these processes have been a driving force in the adaptation of Bacteroidetes to the distal gut environment, and emphasize the importance of considering the evolution of humans from an additional perspective, namely the evolution of our microbiomes.

Citation: Xu J, Mahowald MA, Ley RE, Lozupone CA, Hamady M, et al. (2007) Evolution of symbiotic bacteria in the distal human intestine. PLoS Biol 5(7):e156. doi:10.1371/journal.pbio.0050156

Introduction

Our distal gut is one of the most densely populated and most thoroughly surveyed bacterial ecosystems in nature. This microbiota contains more bacterial cells than all of our body’s other microbial communities combined. The gut microbial community and its collective genome (microbiome) endow us with physiological attributes that we have not had to evolve on our own, including the ability to break down otherwise indigestible polysaccharides [1,2]. The most complete 16S rRNA gene sequence–based enumerations available indicate that more than 90% of phyletic types (phylogenetic types) belong to just two of the 70 known divisions of Bacteria, the Bacteroidetes and the Firmicutes, with the remaining phyletic types distributed among eight other divisions [3]. With an estimated 500–1,000 species, and over 7,000 strains [4], the evolutionary tree of our distal intestinal microbiota can be visualized as a grove of ten palm trees (divisions), each topped by fronds representing divergent lineages, and with each frond composed of many leaves representing closely related bacteria [1]. In contrast, soil, Earth’s terrestrial “gut” for degrading organic matter, can be viewed as a bush, composed of many more intermediate and deeply diverging lineages [5].

It is unclear how selective pressures, microbial community dynamics, and the environments in which we live shape the genomes and functions of members of our gut microbiota, and hence our “micro-evolution.” Ecological principles predict that functional redundancy encoded in genomes from divergent bacterial lineages ensures against disruption of food webs. These principles also predict that host-driven, “top-down” selection for such redundancy should produce a community composed of distantly related members, whose genomes convergently evolve functionally similar suites of genes [4]. Lateral gene transfer (LGT), which allows for rapid transfer of genes under strong selection, such as the genes encoding antibiotic resistance [6], represents one way that members of the microbiota could share metabolic and other capabilities. In contrast, competition between members of a microbiota should exert a “bottom-up” selective pressure that produces specialized genomes with functionally distinct suites of genes. These distinct suites define ecological niches (professions), and once established, could be maintained by barriers to homologous recombination [4].

To explore whether and how these principles apply to the gut microbiota and its microbiome, we have determined the complete genome sequences of two Bacteroidetes with highly divergent 16S rRNA phylotypes that are prominently

* To whom correspondence should be addressed. E-mail: jgordon@wustl.edu
** These authors contributed equally to this work.
Evolution of Human Gut Bacteroidetes

Author Summary

The total number of microbes that colonize the surfaces of our adult bodies is thought to be ten times greater than the total number of our human cells. Our microbial partners provide us with certain features that we have not had to evolve on our own. In this sense, we should consider ourselves to be a supraorganism whose genetic landscape includes both our own genome as well as the genomes of our resident microbes, and whose physiologic features are a synthesis of human and microbial metabolic traits. The largest collection of microbes resides in our gut, which harbors trillions of bacteria, representing hundreds of species, most falling into two groups—the Bacteroidetes and the Firmicutes. We have sequenced bacteria, representing hundreds of species, most falling into two groups—the Bacteroidetes and the Firmicutes. We have sequenced the genomes of two human gut-dwelling Bacteroidetes, and compared their genomes to the genomes of other bacteria that live both inside and outside of our bodies. Our results illustrate that adaptation to the gut habitat is a dynamic process that includes acquisition of genes from other microorganisms. These findings emphasize the importance of including the evolution of “our” microbial genomes when considering the evolution of humans.

Results

Functional Categorization of Genomic Adaptations to the Distal Human Gut Habitat

The 5,163,189–base pair (bp) genome of the human gut-derived B. vulgatus type strain ATCC 8482 encodes a predicted 4,088-member proteome, whereas the 4,811,369-bp genome of B. distasonis type strain ATCC 8503 possesses 3,867 predicted protein-coding genes (Figure S1 and Table S1). These genomes were initially compared to the genomes of two other Bacteroidetes that live in the distal human gut: B. thetaiotaomicron (type strain ATCC 29148 [2]) and B. fragilis (strains YCH 46 and NCTC 9343 [8,9]). We identified 1,416 sets of orthologous protein-coding genes shared among these gut Bacteroidetes; 1,129 (79.7%) of these conserved gene sets were assigned to Clusters of Orthologous Groups (COGs; see Figure S2 and Table S2 for a COG-based categorization). The two most prominently represented COG categories in each of the gut-associated Bacteroidetes proteomes are G (carbohydrate transport and metabolism) and M (cell wall/membrane/envelope biogenesis). The two most prominent COG categories in their shared proteome are E (amino acid transport and metabolism) and J (translation, ribosomal structure, and biogenesis) (Figure S2).

The average pairwise amino acid–sequence identity among the shared orthologs was 82.0% for B. thetaiotaomicron–B. fragilis, 72.1% for B. thetaiotaomicron–B. vulgatus, 62.1% for B. thetaiotaomicron–B. distasonis, and 61.7% for B. vulgatus–B. distasonis. These values are consistent with the 16S rRNA phylogenetic tree for Bacteroidetes (Figure 1). Although the evolution of these gut Bacteroidetes is characterized by comprehensive deterioration of global synteny (Figure S3), a total of 257 “patches” of local synteny were identified, composed of adjacent orthologous genes encompassing 765 of the 1,416 shared orthologs (54%; average of 3.0 orthologs per cluster).

The distal gut microbiota is exposed to several prominent nutrient sources: (1) dietary plant polysaccharides that are not digested in the small intestine by the host because our human proteome lacks the requisite glycoside hydrolases and polysaccharide lyases (see the Carbohydrate Active Enzymes database [CAZy] at http://www.cazy.org for a comprehensive annotation of the human “glycobiome”), (2) undigested plant proteins [10], and (3) host glycans associated with the continuously renewing epithelium that lines the gut and with the even more rapidly replenished mucus layer that overlies this epithelium.

To identify genomic features related to adaptation to life within this distal human gut habitat, we compared shared orthologs among all five completely sequenced gut Bacteroidetes genomes to the subset that is shared with the two Bacteroidetes that occupy non-gut habitats. These non-gut Bacteroidetes are Porphyromonas gingivalis W83, a member of the human oral microbiota [11], and Cytophaga hutchinsonii ATCC 33406, which is found in soil (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list_uids=34). Each proteome was searched for conserved domains. These domains were used to assign a functional identifier (InterPro ID) that was then mapped onto Gene Ontology (GO) terms [12] using InterProScan [13]. The results were compiled and statistical comparisons made between the number of genes assigned to each GO term in different genomes. The complete list of GO assignments for all seven Bacteroidetes genomes is available at http://rd.plos.org/pbio.0050156 (5.3 MB).

The subset of orthologs shared with non-gut Bacteroidetes is enriched for core metabolic activities, suggesting that all Bacteroidetes have inherited a core metabolome from their common ancestor (Figure 2A, compare data in column 5w versus data in 5w). The subset of orthologs unique to the gut Bacteroidetes is enriched for genes related to amino acid biosynthesis, membrane transport, carbon-oxygen lyases, and environment sensing/regulation (see GO terms highlighted in red/pink in the column labeled 5wU in Figure 2A). Furthermore, although a comparison of each gut-dwelling Bacteroidetes proteome to the proteomes of its non-gut relatives (Figure 2B) revealed that the four gut species are all enriched for genes that belong to GO categories related to three general functions—(1) polysaccharide metabolism, (2) environmental sensing and gene regulation, and (3) membrane transport—most of these GO categories are depleted among the subset of orthologs that are unique to the gut-associated Bacteroidetes (Figure 2A, 5w versus Bt-G). Thus, even though all four sequenced gut Bacteroidetes species have increased numbers of genes in categories (1) through (3), this analysis suggests that each species has evolved a divergent array of sensing, regulatory, and polysaccharide degradation genes that augment the core metabolome they share with other members of their division.

Niche Specialization of Bacteroidetes

To further define the niches occupied by the gut Bacteroidetes, we compared each one to B. thetaiotaomicron.
CAZy enzyme class represented in its proteome is markedly as chitinases. Moreover, the number of genes present in each polysaccharidases that target non-plant carbohydrates, such as arabinosidases, a Table S1).

B. distasonis associated with carbon source degradation (Figure 2B and regulation GO categories, and the smallest number of genes are members of the environmental sensing and gene regulation attributes. Scanning electron microscopy, whole-genome transcriptional profiling, and mass spectrometry–based metabolomic studies performed in gnotobiotic mice colonized with this prominent human gut symbiont have shown that B. thetaiotaomicron is a remarkably flexible forager for polysaccharides that opportunistically deploys different subsets of its 209 paralogs of SusC and SusD (two outer membrane proteins involved in the binding and import of starch and maltooligosaccharides [14,15]), and its 226 predicted glycoside hydrolases plus 15 polysaccharide lyases, so that it can feast on dietary or host mucus glycans, depending upon the polysaccharide content of the host's diet [16] (Table S1).

Compared to the other Bacteroidetes, the B. thetaiotaomicron proteome has the most glycoside hydrolases known or predicted to degrade plant glycans (e.g., 64 arabinosidases; our human proteome has none), and the most enzymes for harvesting host glycans (e.g., sulfatic ester hydrolases, hexosaminidases, and fucosidases) (Figure 2B and Table S3). It is also the only sequenced gut Bacteroidetes that possesses candidate polysaccharide lyases for degrading animal tissue glycans (e.g., heparin, chondroitin, and hyaluronan; Table S3). B. thetaiotaomicron’s ability to opportunistically use many glycan sources likely makes it an important generalist among intestinal Bacteroidetes.

Compared to B. thetaiotaomicron, B. distasonis is a specialist. It has the smallest genome among the sequenced human gut-associated Bacteroidetes, the smallest repertoire of genes that are members of the environmental sensing and gene regulation GO categories, and the smallest number of genes associated with carbon source degradation (Figure 2B and Table S1). B. distasonis lacks many accessory hemicellulases (arabinosidases, α-glucuronidases), pectinases, and other polysaccharidas that target non-plant carbohydrates, such as chitinases. Moreover, the number of genes present in each CAZy enzyme class represented in its proteome is markedly reduced compared to the other intestinal Bacteroidetes (e.g., B. distasonis has only one candidate α-fucosidase, whereas the other gut-associated species have nine or ten) (Table S3).

B. distasonis has two classes of carbohydrate-processing enzymes that are more abundant in its proteome than in the proteomes of other gut Bacteroidetes: CAZy glycoside hydrolase family 13 (α-amylase–related proteins), and family 73 (N-acetylhexosaminidases, which can target host glycans as well as bacterial cell walls). Its proteome also contains more polysaccharide deacetylases (seven versus four in B. thetaiotaomicron, and one to two in the B. fragilis strains, as characterized by InterPro ID IPR002509; see http://rd.plos.org/pbio.0050156.a for a complete list of InterPro ID assignments). Host epithelial glycans contain O-acetylated sugars, including sialic acids, that protect them from direct cleavage by microbial glycoside hydrolases. Thus, B. distasonis has the capacity to make the deacetylated products available for itself and other components of the microbiota. Finally, B. distasonis devotes a greater proportion of its genome to protein degradation than does B. thetaiotaomicron (GO:0006508, “protein degradation”: p < 0.0003 by binomial test; Figure 2B).

The B. vulgatus glycobiome has features consistent with ex vivo studies indicating that its substrate range for polysaccharides is intermediate between that of B. distasonis and B. thetaiotaomicron [17]. B. vulgatus has the largest and most complete complement of enzymes that target pectin, a common fruit-associated class of glycans (includes pectin methylesterases, pectin acetylsterases, polygalacturonases, and accessory δ-1,5 unsaturated glucuronyl hydrolases). According to the CAZy classification scheme, B. vulgatus is the only sequenced gut Bacteroidetes with a gene encoding a xylanase (Bv0041c). Together, these findings reveal overlapping, but distinct, niches among these gut Bacteroidetes. We next examined the role of LGT in shaping their genomes.

Lateral Gene Transfer

Determining whether a gene is laterally transferred is widely acknowledged to be a difficult problem (e.g., [18–21] and Text S1). We chose a phylogenetic approach (see Materials and Methods) to identify genes that appeared to
have been laterally acquired and probably selected for after the divergence of individual gut species. Our approach could potentially identify two types of genes: genes that were laterally transferred only into one lineage, and genes that were lost in all lineages except one. We also had to acknowledge the possibility that random errors in phylogenetic tree reconstruction could produce false-positive results. We confirmed that LGT was the more likely scenario for these genes by demonstrating that they differed in composition from the rest of the genome. This approach allowed us to investigate the adaptations of individual lineages to their specific niche. For simplicity, we refer to these genes as...
Our method identified an average of 5.5% of the genes in each genome as being laterally transferred from outside the gut Bacteroidetes, suggesting a lateral transfer event and/or differential gene loss; (2) those whose closest relative is within the gut Bacteroidetes, indicating likely vertical inheritance; (3) those without any homologs in the database (i.e., “novel”); and (4) those whose pattern of inheritance, whether lateral or vertical, could not be determined (i.e., “unresolved”). Parsimony was used to assign a likely direction (“in” or “out”) to each lateral transfer event where possible (see Table S4 and Materials and Methods).

We did not attempt to resolve lateral transfer events within the gut Bacteroidetes in this study, primarily because the lack of sufficient taxonomic sampling within the Bacteroidetes made it impossible to distinguish transfer from biased sampling. Previous studies have observed that a number of novel genes in other bacterial genomes seem to be laterally transferred [24]. However, for the purposes of our functional analyses, these novel genes were excluded because little functional information is available about them. Because we wished to analyze adaptation to the gut, we also excluded genes that appeared to have been transferred out of the Bacteroidetes.

Our method identified an average of 5.5% of the genes in each genome as being laterally transferred from outside the gut Bacteroidetes (312 for B. distasonis, 184 for B. vulgatus, 277 for B. thetaiotaomicron, 199 for B. fragilis NCTC 9343, 214 for B. fragilis YCH 46, and 103 for P. gingivalis). We verified that the genes we identified as “laterally transferred” differed from

**Figure 3. Analyses of Lateral Gene Transfer Events in Bacteroidetes Lineages Reveal Its Contribution to Niche Specialization**

(A) Genes involved in core metabolic processes are enriched among non-laterally transferred genes identified by a phylogenetic approach (see Materials and Methods). The proportion of genes identified as not laterally transferred in each genome (light blue), as well as assigned to the GO terms “Primary metabolism” (yellow) and “Protein biosynthesis” (red), are shown. Significant increases (enrichment) relative to each complete genome (light blue). Glycosyltransferases (yellow) and genes located within CPS loci (green) are also enriched within the set of transferred genes. Significance was determined and denoted as in (A).

(B) Laterally transferred genes are enriched among genes assigned to the GO term “DNA methylation” (yellow), relative to each complete genome (light blue). Glycosyltransferases (yellow) and genes located within CPS loci (green) are also enriched within the set of transferred genes. Significance was determined and denoted as in (A).

(C) B. distasonis (light blue) possesses a significantly larger proportion of laterally transferred genes than the other Bacteroidetes, as shown by significant increases in the proportion of genes in each category of our analysis (“LGT in,” laterally transferred into the genome; “Novel,” no homologs identified from other species; “LGT direction unresolved,” laterally transferred but direction unknown; “LGT out,” laterally transferred out of the genome; and “Unresolved,” lateral transfer uncertain; see Materials and Methods for detailed explanations of categories and http://rd.plos.org/pbio.0050156.a for a complete list of genes in each category). Significant changes, denoted as in (A), were determined by a binomial test, using the average proportion within all other genomes used in the analysis as the reference. Other strains are B. vulgatus (red), B. thetaiotaomicron (yellow), B. fragilis NCTC 9343 (green), B. fragilis YCH 46 (purple), and P. gingivalis (orange). (D) A prominent laterally transferred locus within B. distasonis contains a ten-gene hydrogenase complex, likely allowing B. distasonis to use hydrogen as a terminal electron acceptor in anaerobic respiration. Genes transferred into B. distasonis are colored red, whereas genes whose phylogeny could not be resolved are shown in yellow. Letters indicate functional components of the hydrogenase complex: L, large subunit; M, maturation or accessory factor; and S, small subunit.

[doi:10.1371/journal.pbio.0050156.g003](http://rd.plos.org/pbio.0050156.g003)
Figure 4. Evolutionary Mechanisms That Impact Bacteroidetes CPS Loci

(A) CTn-mediated duplication of *B. vulgatus* CPS loci. Homologous gene pairs in the two duplicated regions are linked with fine gray lines, underscoring the high level of synteny. Genes constituting CPS loci 1 and 2 are highlighted in red, with the first and last genes numbered. Green denotes essential component genes of CTns. Blue brackets indicate two subregions that share 100% nucleotide sequence identity. The asterisk (*) indicates three open reading frames encoding two conserved hypothetical proteins and a hypothetical protein, suggesting an insertion that occurred after the duplication event.

(B) Locations of putative glycosyltransferase xenologs and inserted phage genes in CPS loci of the sequenced gut Bacteroidetes. Color code: integrases (green), UpxY transcriptional regulator homologs (black), putative xenologs (primarily glycosyltransferases, red), phage genes (blue), and remaining genes (gray). See Table S5 for functional annotations.

doi:10.1371/journal.pbio.0050156.g004
The homologous RagA/RagB proteins from branches that are unique to each species are color-coded as indicated.

(A) Cladogram generated from all fully sequenced Bacteroidetes. Specialized and Shared Branches among the Bacteroidetes

Figure 5.

(B) An example of a recently amplified polysaccharide utilization locus in B. distasonis that includes duplicated regulatory genes. Syntenic regions are aligned as in (A) and include a single sulfatase (1, dark green), a SusD paralog (2, light purple), SusC paralog (3, dark purple), an anti-r factor (4, light orange), and an ECF-r factor (5, dark orange). Two other downstream sulfatase genes (gray) are also included in one cluster. The locations of the two SusC/SusD pairs encoded within these clusters are indicated on the cladogram shown in (A) by black arrows.

C

Figure 5. Cladogram Comparison of SusC/SusD Pairs Shows Both Specialized and Shared Branches among the Bacteroidetes

(A) Cladogram generated from all fully sequenced Bacteroidetes. Branches that are unique to each species are color-coded as indicated. The homologous Raga/RagB proteins from P. gingivalis were selected as an arbitrary root (dashed branches). Dashed lines surrounding the tree indicate (1) a clade that is dominated by B. thetaiotaomicron SusC/SusD pairs (39/45 pairs, red dashes) and (2) a clade that is poorly represented in B. thetaiotaomicron (7/34 pairs, black dashes). Colored hash marks surrounding the cladogram represent the linkage of two other protein families, which show syntenic organization within related B. thetaiotaomicron SusC/SusD-containing loci: NHL repeat–containing proteins (light blue) and a group of conserved hypothetical lipidated proteins (light green). These protein families are not represented in the other sequenced Bacteroidetes, occur only adjacent to SusC/SusD pairs, and have no predicted functions. See http://rd.plos.org/pbio.0050156.a for locus tags for each taxon, branch bootstrap values, and lists of SusC/SusD-linked genes.

(B) An example of a recently duplicated locus from B. distasonis that has a significantly larger proportion of laterally transferred genes than the other gut Bacteroidetes (Figure 3C). This excess of LGT does not correlate with a larger number of identifiable mobile elements: B. distasonis has fewer of the integrases and transposases that can catalyze the insertion of foreign DNA than do the other Bacteroidetes, and similar numbers of phage (five versus two to five for the other species; see Table S1). The excess of LGT genes in B. distasonis is also not simply attributable to its more distant phylogenetic relationship to the other gut Bacteroidetes, because P. gingivalis does not share this feature (Figure 3C). Instead, B. distasonis has a striking elevation in the proportion of DNA methylation proteins classified as laterally transferred. Seventy percent of genes classified as “DNA methylation” (GO:0006306; e.g., restriction-modification systems) are predicted to be laterally transferred, even though B. distasonis has fewer DNA methylation genes overall (ten versus an average of 11.5 for other gut Bacteroidetes; Figure 3C). The combination of a smaller number of restriction-modification systems, together with their acquisition from unrelated bacteria, would be expected to reduce the barriers to LGT by allowing B. distasonis to acquire genes from those bacteria. These laterally acquired genes may contribute to the success of B. distasonis within the gut habitat. For example, among the set of transferred genes is a ten-gene hydrogenase
complex (Figure 3D), which would allow B. distasonis to use hydrogen as a terminal electron acceptor.

The Role of Lateral Gene Transfer in the Evolution of Capsular Polysaccharide Biosynthesis Loci

Capsular polysaccharide biosynthesis (CPS) locus expression and the functional importance of capsular structural variation have been best characterized in B. fragilis. For example, studies in gnotobiotic mice indicate that the zwitterionic capsular polysaccharide from one B. fragilis CPS locus (PSA) is presented by intestinal dendritic cells, resulting in expansion of CD4+ T cells, induction of IFNγ production by the T helper 1 subtype (Th1), and reversal of the T helper 2 (Th2) bias found in the absence of colonization. The result is a balanced Th1/Th2 cytokine profile that should help promote co-existence with a microbiota and, perhaps, tolerance to a variety of environmental antigens, including those found in food [29].

B. vulgatus has nine CPS loci, whereas B. distasonis has 13. Like B. thetaiotaomicron (eight CPS loci) and B. fragilis (nine each in strains NCTC 9343 and YCH 46), each CPS cluster is composed of a pair of linked upstream UpcY and UpcZ homologs that act as a “regulatory cassette,” and downstream genes encoding glycosyltransferases, carbohydrate transporters, and other proteins that form a “structural cassette” (Table S5).

Among gut-associated Bacteroidetes, we found that glycosyltransferases and genes in CPS loci are enriched for laterally transferred genes (Figure 3B). P. gingivalis, in contrast, does not show a biased representation of lateral transfer within its set of glycosyltransferases, suggesting that laterally acquired genes serve an important function in providing new genetic material for the rapid divergence of these loci in gut Bacteroidetes.

CPS loci are among the most polymorphic sites in the four gut-associated Bacteroidetes species [30,31]. A comparison of the two sequenced B. fragilis genomes [8,9] revealed that the genome-wide synteny evident in the two closely related B. fragilis strains is disrupted in eight of their nine CPS loci (Figure S4 and Table S6).

Conjugative Transposons, Phage, and Other Mechanisms Involving in Promoting CPS Diversity

Conjugative transposons. We observed that conjugative transposons (CTns) are associated with the duplication of CPS loci within a genome. In B. vulgatus, Bv0624–Bv0699 (75,747 bp) is a copy of another region (Bv1479c–Bv1560, 75,277 bp) (Figure 4A and Table S5). Each copy contains a CTn followed by a complete CPS locus. The average amino acid sequence identity of the 64 homologous gene pairs comprising the repeated regions is 90%. Two exact 28,411-bp copies harbor a major portion of the structural cassettes of these duplicated CPS loci, plus part of a CTn (Figure 4A). The strict nucleotide-level sequence conservation in coding and non-coding sequences suggests a recent homologous recombination event at the structural cassettes of the CPS loci. There is also evidence that the function of CPS loci can be disrupted by CTns, as in CPS locus 8 of B. fragilis YCH 46 where an α-L-fucosyltransferase gene is interrupted by a 127-kilobase (Kb), 132-gene CTn (Table S5).

Phages. Phages also appear to modulate CPS locus function. In B. distasonis, CPS locus 5 contains a block of five genes inserted between its regulatory cassette and genes encoding carbohydrate biosynthetic enzymes. This inserted segment, oriented in the opposite direction to the upstream regulatory UpxY (and UpxZ) genes and downstream carbohydrate biosynthetic genes, consists of a homolog of phase T7 lysozyme (N-acetylmuramoyl-L-alanine amidase) followed by four genes encoding hypothetical proteins. Three more B. distasonis CPS loci each harbor a block of these genes (two to five genes per block; each block with a similar orientation; only the T7 lysozyme is conserved among all copies of the putative phages; Figure 4B and Table S5).

B. distasonis is the only sequenced type strain in which a phage disrupts CPS loci between their regulatory cassettes and structural cassettes. B. vulgatus has five copies of this phage, all associated with CPS loci. B. thetaiotaomicron has ten copies, only two of which are associated with CPS loci, whereas the B. fragilis strains each have one (Table S5).

Phase variation. LGT, CTn-mediated duplication and translocation of CPS loci, and disruption of CPS loci by phage appear to operate in combination with at least two other mechanisms to promote the rich diversity of surface glycan structures in Bacteroidetes. In B. fragilis, a serine site-specific recombinase (Mpi) regulates expression of seven of its eight CPS loci through phase variation (DNA inversion) at CPS promoters [32]. B. vulgatus, B. distasonis, and B. thetaiotaomicron have Mpi orthologs (one, three, and one, respectively). In addition, five of the nine CPS loci in B. vulgatus, 11 of the 13 CPS loci in B. distasonis, four of the eight CPS loci in B. thetaiotaomicron, and only one of the ten CPS loci in B. fragilis NCTC 9343 have a gene encoding a tyrosine-type site-specific recombinase immediately upstream of a UpxY homolog. This juxtaposition suggests that inversions of some CPS loci may be subjected to local as well as global regulation. Such sequence inversions were observed in the assemblies of the B. vulgatus and B. distasonis genomes (unpublished data).

Fkp and fucose utilization. B. fragilis can also alter CPS glycan composition by means of Fkp, a protein whose N-terminus is homologous to mammalian L-fucose-1-P-guanylyltransferase and whose C-terminus is similar to L-fucose kinases. Fkp generates GDP-L-fucose from exogenous L-fucose; fucose from GDP-L-fucose can be incorporated into CPS glycan structures, thereby linking L-fucose availability in the organism’s intestinal habitat to CPS capsular structure [33]. Although Fkp is highly conserved in B. distasonis, B. vulgatus, B. thetaiotaomicron, and B. fragilis, their L-fucose acquisition and utilization capacities are not. B. distasonis, B. vulgatus, B. thetaiotaomicron, and B. fragilis all possess α-fucosidases for harvesting L-fucose, which is a common component of host mucus and epithelial cell glycans. In B. thetaiotaomicron and B. fragilis, a complete fucose utilization system is incorporated into a gene cluster (fuelRAKXP). In B. vulgatus, this gene cluster (Bv1393c–Bv1341c) contains an ortholog of B. thetaiotaomicron’s L-fucose–inhibited repressor (R), fucose isomerase (I), and fucose permease (P), but not L-fuculose-1-phosphate aldolase (A) or L-fuculose kinase (K). B. distasonis lacks all elements of this gene cluster.

The Role of Gene Duplication in Diversification of Gut Bacteroidetes: A Case Study of SusC/SusD Paralogs

As noted above, the gut Bacteroidetes genomes contain large numbers of paralogs involved in environmental sensing and nutrient acquisition. We used one of the largest families,
the SusC/SusD paralogs (Table S1), as a model for investigating relationships among members. SusC paralogs are predicted to be TonB-dependent, β-barrel–type outer membrane proteins. Thus, in addition to binding nutrients such as polysaccharides, SusC paralogs likely participate in their energy-dependent transport into the periplasmic space [34]. SusD paralogs are predicted to be secreted and to have an N-terminal lipid tail that would allow them to associate with the outer membrane [14]. Genes encoding SusC and SusD paralogs are typically positioned adjacent to one another in the B. thetaiotaomicron genome (102 of 107 loci encoding SusC paralogs), and are often part of multigene clusters that also encode enzymes involved in carbohydrate metabolism (62 of 107 loci) [2]. Eighteen of the 62 clusters that encode SusC/SusD paralogs and glycoside hydrolases also contain ECF-σ factors and adjacent anti-σ factors. A subset of SusC paralogs contain an extra N-terminal domain with homology to the N-terminal domain of the Escherichia coli FecA iron-dicitrate receptor protein [35]. FecA interacts directly with an anti-σ factor (FecR) via this domain, thereby controlling gene expression through modulation of its associated ECF-σ factor (FecI).

These clusters provide case studies of the evolution of gut Bacteroidetes genomes. Their glycoside hydrolase content varies considerably within a given species (Table S7). Our studies in B. thetaiotaomicron indicate that ECF-σ factors are required for transcription of their adjacent polysaccharide utilization gene clusters, and that chromosomally linked anti-σ factors act as repressors of this transcription. Moreover, several B. thetaiotaomicron loci containing ECF-σ and anti-σ factors are differentially regulated during growth on various complex glycans ([16]; E. C. Martens and J. I. Gordon, unpublished data), suggesting that these systems act as components of carbohydrate sensors responsible for regulating loci appropriate for utilizing available nutrients.

Six of these clusters in B. distasonis (clusters 2–6 and 16 in Table S7-A) include predicted sulfatases, whereas there are fewer such loci in the other genomes: two clusters in B. vulgatus (clusters 5 and 11 in Table S7-B), four in B. thetaiotaomicron, and three in each of the two B. fragilis strains. These enzymes could be involved in the desulfation of sulfomucins that contain galactose-3-sulfate, galactose-6-sulfate, and N-acetylglucosamine-6-sulfate residues. These or other sulfatases could also be involved in the desulfation of glycosaminoglycans such as chondroitin and heparin.

To explore the role of gene duplication in the diversification of the Bacteroidetes, we generated lists of all paired SusC and SusD paralogs from the four gut- and one non-gut–associated Bacteroidetes species (see Materials and Methods). P. gingivalis has four such pairs, whereas the other five intestinal Bacteroidetes species have a total of 370 (Table S1). A cladogram generated from the multiple sequence alignment shows that many SusC/SusD pairs have close relatives among several Bacteroidetes. However, certain specialized groups are unique to each species, with B. thetaiotaomicron containing one particularly large expansion (Figure 3A). Gene clusters encoding related SusC/SusD pairs also contain other genes that are closely related to one another. The homology and synteny of these loci suggest that genomic duplication is a mechanism driving their amplification and diversification (e.g., Figure 5B and 5C). An intriguing feature of some of these amplified loci is that they contain clusters of genes with unique functions that are located downstream of the “core” duplicated genes; this may serve to further diversify the roles of these loci in nutrient acquisition (e.g., Figure 5B in which diverse dehydrogenase, sulfatase, and glycoside hydrolase functions are included downstream of a syntenic core of amplified genes).

**Discussion**

The trillions of bacteria that colonize our distal gut largely belong to two bacterial divisions, and can be classified by 16S rRNA gene sequence analysis into hundreds of “species” that share a common ancestry [4,36], but whose genome content may vary considerably. Forces that shape the genome content of bacteria in the gut include the intermicrobial dynamics of competition and cooperation in resource partitioning that shape complex food webs, as well as other community-shaping forces, such as phage attacks, that can result in “selective sweeps” that remove cells with similar susceptibilities. In a competitive environment in which innovation in resource acquisition strategies can breed success, and resistance to phage can mean surviving a phage selective sweep, bacteria can be expected to differentiate their genome content. For the host to thrive and produce more gut habitats (by reproducing), the gut microbial ecosystem must be functionally stable over time despite the internal dynamics of the community. The constituent bacteria might therefore be expected to have a high degree of functional redundancy between species, so that the loss of one lineage would not adversely impact ecosystem services to the host. Our investigation of the genomes of human gut Bacteroidetes species shows that the “top-down” forces imposed by selection at the host level that would result in a homogenized microbiome, and the “bottom-up” forces of intermicrobial dynamics that would result in completely differentiated genomes, are both at work in the distal gut. The genomes of the gut Bacteroidetes species that have been sequenced harbor suites of genes with similar functions, but differ in the number of genes within functional categories and their specific sequence. It appears that the differences between genomes are enough to carve out specific niches within the gut habitat, such that the species are not in direct competition, but are sufficiently similar to confer resistance to disturbance to the host through functional redundancy.

Our findings demonstrate a key role for LGT in shaping the adaptation of individual Bacteroidetes to distinct niches within the human gut. It is unclear how and when laterally transferred genes were introduced during evolution of distal gut Bacteroidetes. We have performed 16S rRNA gene sequence–based enumeration studies of the fecal microbiota of 59 different mammalian species: the results reveal that none of the four sequenced gut Bacteroidetes species is restricted to the human gut (R. E. Ley and J. I. Gordon, unpublished data). Nonetheless, the impact of LGT is likely profound for these gut symbionts and their human hosts. A large and varied gene pool of glycosyltransferases provides a capacity for diversification of surface polysaccharide structures that could endow symbionts with varied capacities to shape a host immune system so that it can accommodate a microbiota (and perhaps related food and other environmental antigens). Because the environment surrounding each human being varies, this gene flow may promote the
generation of host-specific microbiomes. Acquisition of new types of carbohydrate binding proteins, transporters, and degradation enzymes through both LGT and gene amplification should influence the types of substrates that can be exploited for energy harvest. It may also affect our predisposition to conditions such as obesity in which the efficiency of caloric harvest may be influenced by the relationship between an individual’s microbial glycoside hydrolase repertoire and the glycan content of his/her diet [37,38].

These considerations emphasize the need to have a more comprehensive view of our genetic landscape as a composite of human and microbial genes, a transcendent view of human evolution as involving our microbial partners, and a commitment to investigating human biology in the larger framework of environmental microbiology. Attention to these issues is timely, given the onset of efforts to sequence the human “microbiome” [39]. These metagenomic studies will allow investigators to address new, but fundamental, questions about humans. Do we share an identifiable core “microbiome”? If there is such a core, how does the shell of diversity that surrounds the core influence our individual physiologic properties? How is the human microbiome evolving (within and between individuals) over varying time scales as a function of our changing diets, lifestyle, and biosphere? Finally, how should we define members of the microbiome when microbes possess pan-genomes (all genes present in any of the strains of a species) with varying degrees of “openness” to acquisition of genes from other microbes?

Materials and Methods

Genome sequencing. The B. vulgatus and B. distasonis genomes were assembled from two types of whole-genome shotgun libraries: a plasmid library with an average insert size of 5 Kb, and a fosmid library with an average insert size of 40 Kb. For each genome, both Phraps (http://www.phrap.org/) and PCAP [40] assemblies were generated and then compared, resulting in a “hybrid” assembly that takes advantage of the strength of both assemblies. Regions that contained a gap in one assembly, but not in the other, were made contiguous in the final assembly for finishing by using Consed [41]. Sequence gaps were filled by primer-walking on spanning clones. Physical gaps were amplified by PCR and closed by sequencing the PCR products. Poor quality regions were detected using Consed, amplified with PCR, and resequenced. The integrity and accuracy of the assembly were verified by clone constraints. Regions of lower coverage, or that contained ambiguous assemblies, were resolved by sequencing individual fosmids. Regions that underwent sequence inversions were identified based on inconsistency of constraints for a fraction of read pairs in those regions. The final assemblies consisted of 12.6X and 13.2X sequence coverage for B. vulgatus and B. distasonis, respectively. For each base, the Phred quality value was at least 40.

rRNA and tRNA genes were identified with BLASTN and tRNA-Scan [42], respectively. Protein-coding genes were identified using GLIMMER v.2.0 [43], ORPHEUS v.2.0 [44], and CRITICA v.0.94h [45]. WUBLAST (http://blast.wust.edu) was used to identify all predicted proteins with significant hits to the non-redundant (nr) database. Predicted protein-coding genes containing fewer than 60 codons and without significant homology (e-value threshold of $10^{-10}$) to other proteins were eliminated. Gene start-site predictions were fine-tuned using MED-Start [46] and BLAST homology. In general, no overlap between genes was allowed. Potential frameshift errors were identified by sequence alignment with known proteins, and confirmed or corrected by resequencing. The final set of genes, compiled from the analysis described above, was manually curated. Protein annotation was based on homology searches against public databases and domain analysis with HMMER (http://hmslab.janelia.org/), functional classification was based on homology searches against COG categories using WU-BLAST and COGnitor [47], followed by manual curation. Metabolic pathways were constructed with reference to KEGG [48]. Phage genes were identified using Prophage finder [49].

Functional comparisons. Orthologs of the five intestinal Bacteroidetes genomes were identified based on (1) mutual BLASTP best hits with an e-value threshold of 10$^{-6}$ and (2) a requirement that each pairwise protein alignment cover at least 60% of the gene length in both search directions. The amino acid sequences of each set of orthologs were aligned using ClustalW [50] and processed with GBLOCKS [51]. Conserved genes in the five intestinal Bacteroidetes genomes were defined with the following criteria. First, an intact CPS locus included a $\alpha$PS homolog (annotated) and a number of downstream genes on the same strand. These downstream genes included those that encoded functions related to surface protein synthesis (such as glycosyltransferases, carbohydrate export proteins, epimerases, glyco-side hydrolases, etc.), conserved hypothetical proteins, or hypothetical proteins. Second, the 5′ boundary of each locus was determined by the $\alpha$PS homolog. Third, the 3′ end of each locus generally was positioned where switch of coding strand occurred. Alternatively, the 3′ end of the locus was selected when downstream genes on the same strand encoded functions that were defined, but unrelated, to capsular polysaccharide synthesis (e.g., tRNA, tRNA and two-component signaling systems). However, the 3′ end of the locus was extended if the coding strand was disrupted by a single hypothetical protein (to accommodate possible annotation errors), or a mobile element composed of one or multiple genes.

GO categories and InterPro ID were assigned using InterProScan release 12.1 [13]. The number of genes in each genome assigned to each GO term, or its parent term in the hierarchy, and the total of all GO terms assigned to each GO term, were counted and compared. If the total number of genes assigned to a GO category was less than 10, the GO category was not used for functional comparison. The GO categories and InterPro ID were assigned using InterProScan [52], respectively. Protein-coding genes were identified using a binomial comparison with the indicated reference set, or to control for differences in the specificity of gene prediction, genes that could not be assigned to a GO category were excluded from the reference sets. A correction was then applied to each distinct set of tests (e.g., over- or underrepresentation in a genome) to achieve a false discovery rate of 0.05 for each set [52]. These tests were implemented using Math::CDF Perl module (E. Callahan, Environmental Statistics, University of Wisconsin; available at http://www.cpan.org/), and scripts written in Perl.

16S rRNA phylogeny. Phylogenetic trees were constructed based on alignment of 16S rRNA fragments using the NAST aligner [53]. The alignment was filtered using a Lane mask, then modeled using ModelTest 3.7 [54]: a maximum-likelihood tree was found by an exhaustive search using PAUP* v. 4.0b10, http://paup.cst.fsu.edu/) employing parameters estimated by ModelTest.

Laterally transferred genes. Overview of strategy used to identify LGT. See Text S1.

Identifying classes of genes that were potentially laterally transferred or otherwise under selection in the gut Bacteroidetes. We identified genes that were laterally acquired and probably selected for after the divergence of gut Bacteroidetes species, and thus potentially involved in niche differentiation. These genes could either have been transferred into an individual species by LGT, or retained in that species despite being lost in all other related sequenced species. It is difficult, perhaps impossible, to distinguish these two cases using the tree topology alone. We identified this class of genes by determining whether each gene met one of the following criteria. (1) No homologs were found in an augmented NCBI non-redundant (nr) protein database (plus the proteins from the newly sequenced strains). This case indicated that either the gene has been lost in every other gut organism but retained in this genome, or that the gene was laterally transferred from an organism that is not represented in the database. (2) The only homologs found were from the same species. This case is the same as case (1), except that either the gene was sequenced multiple times and deposited in the database under separate records, or there are multiple copies of the gene in the genome being analyzed. Both case (1) and (2) were termed "novel." (3) The only homologs found are either from the same species or from other divisions or non-gut Bacteroidetes. This case indicates that the sequence is in this genome, and also in the genome of a gut organism related to, but not in the same gut Bacteroidetes genomes that have been completely sequenced. This case also provides evidence that the gene was either transferred or retained despite loss in related organisms. (4) The gene is more closely related to genes from other divisions or to non-gut Bacteroidetes, but not in the gut Bacteroidetes genomes. This case indicates that the direction of transfer was into rather than out of the genome. This pattern is most consistent with LGT, although differential gene loss cannot in
principle be ruled out (however, differences in composition between this class of genes and the rest of the genome provide compelling supporting evidence). Both cases (3) and (4) above were termed “laterally transferred” (LGT).

Genomes. We carried out the analysis on the six different genomes: Bacteroides vulgatus ATCC 8482 and Bacteroides distasonis ATCC 58503, Bacteroides thetaiotaomicron NCTC 9343 (NC_002338), Bacteroides thetaiotaomicron YCH 46 (NC_006347), Bacteroides thetaiotaomicron ATCC 29148 (NC_004663), and Porphyromonas gingivalis W83 (NC_002950).

Finding homologs. For each gene in each genome, we identified potential homologs using PSI-BLAST against NCBI’s nr protein database. In order to use all of the available data for the Bacteroidetes and their relatives, we augmented this database with proteins predicted by Glimmer (v. 2.9) from draft genomes in the Bacteroidetes group that were available at NCBI. These additional genomes included Prevotella ruminicola (The Institute for Genomic Research; TIGR), Chlorobium limicola DSM 245 (DOE-JGI), Chlorobium chlorobium CaB3 (DOE-JGI), B. thetaiotaomicron (TIGR), and B. fragilis 638R (Wellcome Trust Sanger Institute; http://www.sanger.ac.uk).

To find the top BLAST hits using the most stringent e-value threshold possible, we used a multistep PSI-BLAST. In the first PSI-BLAST iteration, with an e-value threshold of 10^{-30}, fewer than 50 hits were found; we used the hits to make a profile for a subsequent PSI-BLAST that was four orders of magnitude less stringent (i.e., with an e-value of 10^{-40}). We repeated this procedure, increasing the e-value by a factor of 10^{-3} at each iteration, until either 50 hits were found or, after 12 iterations, the maximum allowed e-value was reached. To remove those sequences that were significant only because of a conserved domain rather than similarity over the whole gene, we excluded genes that differed from the length of the query sequence by more than 30%. We also omitted hits that contained gaps greater than 50 amino acids in length or that contained gaps at greater than 50% of the positions after performing a multiple sequence alignment with the other sequences in the set.

Making phylogenetic trees. We performed multiple sequence alignment using MUSCLE [55], omitting sequences that were poorly aligned to the query sequence as described above. We used this alignment to make a neighbor-joining tree using ClustalW [50]. We used bootstrapping to collapse nodes that were not statistically supported. Specifically, we randomly resampled columns from the alignment 100 times and made new neighbor-joining trees with ClustalW. We collapsed into polytomies all nodes in the original tree that were recovered in fewer than 70% of the bootstrap replicates.

Assigning taxonomy information to sequences. We parsed the NCBI taxonomy database and used it to assign division and genus information for each PSI-BLAST hit in the phylogenetic tree. Sequences that could not be assigned to any particular division were removed from the tree. We also removed nematode and arthropod sequences that were significant only because of a conserved domain rather than similarity over the whole gene, we excluded genes that differed from the length of the query sequence by more than 30%. We also omitted hits that contained gaps greater than 50 amino acids in length or that contained gaps at greater than 50% of the positions after performing a multiple sequence alignment with the other sequences in the set.

Supporting Information

Figure S1. B. distasonis ATCC 8503 and B. vulgatus ATCC 8482 Chromosomes

The B. distasonis ATCC 8503 chromosome is shown in (A), and the B. vulgatus ATCC 8482 chromosome is shown in (B). The coding potential of the leading and lagging strands is relatively unbiased. Circles shown in the figure represent, from inside out, GC skew, GC content variation, rRNA operons, tRNA genes, conjugative transposons (CTns), CPR loci, extracytoplasmic function (ECF-)r factors, SusC paralogs, and all predicted genes with assigned functions on reverse and forward strands, respectively. Color codes for genes are based on their COG functional classification.

Found at doi:10.1371/journal.pbio.0050156.sg001 (2.3 MB PDF).

Figure S2. COG-Based Characterization of All Proteins with Annotated Functions in the Proteomes of Sequenced Bacteroidetes

The term “Bacteroidetes orthologs” refers to the 1,416 orthologs shared by the sequenced gut Bacteroidetes (B. vulgaris, B. distasonis, B. thetaiotaomicron, plus the two B. fragilis strains). Color codes are the same as Figure S1.

Found at doi:10.1371/journal.pbio.0050156.sg002 (334 KB PDF).

Figure S3. Pairwise Alignments of the Human Gut Bacteroidetes Genomes Reveal Rapid Deterioration of Global Synteny with Increasing Phylogenetic Distance

Each data point in the Dotplot represents one pair of mutual best

Evolution of Human Gut Bacteroidetes
Figure S4. CPS Loci Are the Most Polymorphic Regions in the Gut Bacteroidetes Genomes

High-resolution synteny map of CPS loci and flanking regions in the two sequenced B. fragilis strains. There are nine CPS loci in each genome. Each data point represents a pair of orthologs (mutual best hits; e-value cutoff: 10^-4). Brackets define the coordinates for component genes within a given locus (some pairs are missing due to gene loss or gain): x-axis, coordinate of the middle point of the gene on the NCTC 9345 chromosome; y-axis, coordinate of the middle point of the gene on YCH 46 chromosome. With the exception of CPS locus 5, which is strictly conserved, the nine CPS loci are affected by nonhomologous gene replacement and rearrangement.

Table S1. Comparison of Genome Parameters for B. distasonis ATCC 8503, B. vulgatus ATCC 8482, B. thetaiotaomicron ATCC 29148, B. fragilis NCTC 9345, and B. fragilis YCH 46

An asterisk (*) indicates the numbers of SusCSusD homologs provided are based on BLASTP e-value equal to 10^-10 or lower; the numbers shown in parentheses are based on criteria described in SusCSusD alignments in Materials and Methods. See http://rd.plos.org/phi.000156/a for complete lists of SusCSusD homologs. A hybrid two-component system protein contains all of the domains present in classical two-component systems, but in one polypeptide [50].

Table S2. Shared Orthologs in B. distasonis ATCC 8503, B. vulgatus ATCC 8482, B. thetaiotaomicron ATCC 29148, and B. fragilis Strains NCTC 9345 and YCH 46

For an explanation of COG-based functional codes, see Figure S1.

Table S3. Glycoside Hydrolyses Found in B. distasonis ATCC 8503, B. vulgatus ATCC 8482, B. thetaiotaomicron ATCC 29148, and B. fragilis Strains NCTC 9345 and YCH 46

The classification scheme used is described in the Carbohydrate-Active enZyme (CAZy) database.

Table S4. List of Putative Xenologs in B. distasonis ATCC 8503, B. vulgatus ATCC 8482, B. thetaiotaomicron ATCC 29148, B. fragilis NCTC 9345, and B. fragilis YCH 46

The putative xenologs are listed for B. distasonis ATCC 8503 (A), B. vulgatus ATCC 8482 (B), B. thetaiotaomicron ATCC 29148 (C), B. fragilis NCTC 9345 (D), and B. fragilis YCH 46 (E). For an explanation of COG-based functional codes, see Figure S1. The lateral gene transfer (LGT) column defines the predicted evolutionary history of the coding sequence: LGT-in, laterally transferred into the genome; LGT-out, laterally transferred out of the genome; LGT-unresolved, laterally transferred but direction unknown. See Materials and Methods for detailed explanations.

Acknowledgments

MM was supported by the Ellison Medical Foundation, CAL by a National Institutes of Health (NIH) predoctoral training grant (T32 GM08759), MH by a gift from the Jane and Charlie Butcher Foundation and W. M. Keck Foundation RNA Bioinformatics Initiative, and ECM by an NIH postdoctoral training grant (T32 HD07490). The sequence data for Cytophaga hutchinsonii ATCC 33406, Pelodictyon phaeoclathratiforme BU-1, and Pelodictyon luteolum DSM 273, Chlorobium phaeobacteroides DSM 266, Chlorobium limicola DSM 245, and Chlorobium chlorohumatum CaD3 were produced by the United States Department of Energy-Joint Genome Institute (http://www.jgi.doe.gov). Preliminary sequence data for Prevotella ruminicola 25, Bacteroides intermedius 17, and Bacteroides forsythus were obtained from The Institute for Genomic Research through its Web site at http://www.tigr.org. Sequencing of Prevotella intermedia and Prevotella ruminicola 25 was accomplished with support from the NIH’s National Institute of Dental and Craniofacial Research (NIH-NIDCR) and the US Department of Agriculture, respectively. The draft sequence of Bacteroides fragilis 638R was produced by the Pathogens Sequencing Group at the Sanger Institute and can be obtained from ftp://ftp.sanger.ac.uk/pub/pathogens.

Author contributions. JX, MAM, SWC, RKW, RDK, and JIG conceived and designed the experiments. JX, MAM, CAL, MH, ECM, PM, PL, HC, AVB, KK, RSF, LAF, and RDK performed the experiments. JX, MAM, REL, CAL, MH, ECM, BH, PMC, RDK, and JIG analyzed the data. JX and RDK contributed reagents/materials/analysis tools. JX, MAM, REL, CAL, ECM, RDK, and JIG wrote the paper.

Funding. This work was supported by a grant from the National Science Foundation (EF0333284).

Competing interests. The authors have declared that no competing interests exist.

References

1. Backhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JJ (2005) Host-bacterial mutualism in the human intestine. Science 307: 1615–1618.
2. Xu X, Bjursell MK, Himrod J, Deng S, Carmichael LK, et al. (2005) A generic view of the human-Bacteroides thetaiotaomicron symbiosis. Science 299: 2074–2076.
3. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, et al. (2005) Diversity of the human intestinal microbial flora. Science 308: 1635–1638.
4. Ley RE, Peterson DA, Gordon JJ (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell 124: 857–868.
5. Dunbar J, Barns SM, Tiedje JM, Kuske CR (2002) Empirical and theoretical bacterial diversity in four Arizona soils. Appl Environ Microbiol 68: 3035–3045.
6. Salamon AA, Gupta A, Wang Y (2004) Human intestinal bacteria as reservoirs for antibiotic resistance genes. Trends Microbiol 12: 412–416.
7. Sakamoto M, Benno Y (2006) Reclassification of Bacteroides distasonis, Bacteroides goldsteinii and Bacteroides merdae as Parabacteroides distasonis gen.
8. Kuwahara T, Yamashita A, Hirakawa H, Nakayama H, Toh H, et al. (2004) Genomic analysis of Bacteroides fragilis reveals extensive DNA inversions regulating cell surface adaptation. Proc Natl Acad Sci U S A 101: 14919–14924.

9. Cerdeno-Tarraga AM, Patrick S, Crossman LC, Blakely G, Abratt V, et al. (2005) Extensive DNA inversions in the B. fragilis genome control variable gene expression. Science 307: 1463–1465.

10. Millward DJ (1999) The nutritional value of plant-based diets in relation to human amino acid and protein requirements. Proc Nutr Soc 58: 249–260.

11. Nelson KE, Fleischmann RD, DeBoy RT, Paulsen IT, Fouts DE, et al. (2003) Complete genome sequence of the oral pathogenic bacterium Porphyromonas gingivalis strain W50. J Bacteriol 185: 5591–5601.

12. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 23–29.

13. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, et al. (2005) InterPro, progress and status in 2005. Nucleic Acids Res 33: D201–D205.

14. Shipman JA, Berleman JE, Salyers AA (2000) Characterization of four outer membrane proteins involved in binding starch to the cell surface of Bacteroides thetaiotaomicron. J Bacteriol 182: 5565–5572.

15. Cho KH, Salyers AA (2001) Biochemical analysis of interactions between outer membrane proteins that contribute to starch utilization by Bacteroides thetaiotaomicron. J Bacteriol 183: 7224–7230.

16. Sonnenburg JL, Xu J, Leip DD, Chen CH, Westover BP, et al. (2005) Glycerol foraging in vivo by an intestine-adapted bacterial symbiont. Science 307: 1955–1959.

17. Salyers AA, Vercellotti JR, West SE, Wilkins TD (1977) Fermentation of starch and trypsin by a strain of Parabacteroides goldsteinii nov., comb. nov., Bacteroides fragilis Genomic analysis of strain 638R. J Bacteriol 181: 6192–6196.

18. Nakayama Y, Kobayashi I (1998) Restriction-modification gene complexes as selfish gene entities: Roles of a regulatory system in their establishment, as selfish gene entities: Roles of a regulatory system in their establishment, and high throughput. Nucleic Acids Res 32: 1792–1797.

19. Badger JH, Olsen GJ (1999) CRITICA: A graphical tool for sequence finishing. Genome Res 9: 195–202.

20. Lawrence JG, Hendrickson H (2003) Lateral gene transfer: When will adolescence end? Mol Microbiol 40: 739–749.

21. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved tool for deep molecular phylogenetic inference. Bioinformatics 15: 82–90.

22. Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19: 99–113.

23. Hansen-Wester I, Stecher B, Hensel M (2002) Analyses of the evolutionary genealogy of ORFans in Salmonella enterica serovar Typhimurium. J Mol Evol 42: 91–96.

24. uwahara T, Yamashita A, Hirakawa H, Nakayama H, Toh H, et al. (2004) Genomic analysis of Bacteroides fragilis reveals extensive DNA inversions regulating cell surface adaptation. Proc Natl Acad Sci U S A 101: 14919–14924.

25. Koebnik R (2005) TonB-dependent trans-envelope signalling: The exception or the rule? Trends Microbiol 13: 343–347.

26. Gressmann H, Linz B, Ghai R, Pleissner KP, Schlapbach R, et al. (2005) Gain of metabolic versatility and high throughput. Nucleic Acids Res 32: 1792–1797.

27. Berg OG, Harlow TJ, Beiko RG (2006) Do different surrogate methods detect lateral genetic transfer events of different relative ages? Trends Microbiol 14: 4–8.

28. Goto M, Sato M, Takashima M, Okuno Y, Hattori M (2004) KEGG for their use in phylogenetic analysis. Mol Biol Evol 17: 540–552.

29. Gordon D, Abajian C, Green P (1998) Consed: A graphical tool for sequence analysis. Genomics 52: 25–29.

30. cohn MJ, Reinaip B, Lee MM, Comstock LE (2005) Human symbionts use a host-like pathway for surface fucosylation. Science 307: 1778–1781.

31. Coyne MJ, Reinap B, Lee MM, Comstock LE (2005) Human symbionts use a host-like pathway for surface fucosylation. Science 307: 1778–1781.

32. Reeves AR, D’Elia JN, Frias J, Salyers AA (1996) A Bacteroides thetaiotaomicron outer membrane protein that is essential for utilization of maltodextrins. Proc Natl Acad Sci U S A 100: 10446–10451.

33. Koebnik R (2005) TonB-dependent trans-envelope signalling: The exception or the rule? Trends Microbiol 13: 343–347.

34. Eckburg PB, Lepp PW, Relman DA (2005) Archaeal and their potential role in human disease. Infect Immun 71: 591–596.

35. E. coli

36. Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) Microbial ecology: Human gut microbes associated with obesity. Nature 444: 1022–1023.

37. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, et al. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444: 1027–1031.

38. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science 312: 1535–1539.

39. Huang X, Wang J, Aluru S, Yang SP, Hillier L (2005) PCAR: A whole genome assembly program. Genome Res 13: 2164–2170.

40. Hansen-Wester I, Stecher B, Hensel M (2002) Sequence acquisition and loss. Mol Biol Evol 19: 2265–2276.

41. Gordon D, Abajian C, Green P (1998) Consed: A graphical tool for sequence analysis. Genomics 52: 25–29.

42. delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27: 4636–4641.

43. Frishman D, Mironov A, Mewes HW, Gelfand M (1998) Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. Nucleic Acids Res 26: 2941–2947.

44. Badger JH, Olsen GJ (1999) CRITICA: Coding region identification tool invoking comparative analysis. Mol Biol Evol 16: 524–529.

45. Ragan MA (2001) On surrogate methods for detecting lateral gene transfer. FEMS Microbiol Lett 291: 187–191.

46. Kurland CG, Canback B, Berg OG (2003) Horizontal gene transfer: A critical view. Proc Natl Acad Sci U S A 100: 9658–9662.

47. Kunin V, Ouzounis CA (2003) GeneTRACE-reconstruction of gene content and high throughput. Nucleic Acids Res 32: 1792–1797.

48. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 2947–2950.

49. Karches I, Sot S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32: D277–280.

50. Hanssen-Auster I, Stecher B, Hensel M (2002) Analyses of the evolutionary distribution of Salmonella translocated effectors. Infect Immun 70: 1619–1622.

51. Garcia-Vallve S, Romeu A, Palau J (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res 10: 1719–1725.

52. Badger JH, Olsen GJ (1999) CRITICA: A graphical tool for sequence analysis. Genomics 52: 25–29.

53. Jelsch A, Pinaud A, Reiko RG (2006) Do different surrogate methods detect lateral genetic transfer events of different relative ages? Trends Microbiol 14: 4–8.

54. Goto M, Sato M, Takashima M, Okuno Y, Hattori M (2004) KEGG for their use in phylogenetic analysis. Mol Biol Evol 17: 540–552.

55. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Stat Soc Ser B 57: 289–300.

56. DeSantis TZ Jr, Hugenholtz P, Keller K, Brodie EL, Larsen N, et al. (2006) NAST: A multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res 34: W394–399.

57. Posada D, Crandall KA (1998) MODELTEST: Testing the model of DNA substitution. Bioinformatics 14: 817–818.

58. Brooks DR (1981) Hennig’s parasitological method: A proposed solution. Brooks DR (1981) Hennig’s parasitological method: A proposed solution. Evolution of Human Gut Bacteroidetes