A Closer Look at Bacteroides: Phylogenetic Relationship and Genomic Implications of a Life in the Human Gut

Karlsson, Fredrik H.; Ussery, David; Nielsen, Jens; Nookaew, Intawat

Published in:
Microbial Ecology

Link to article, DOI:
10.1007/s00248-010-9796-1

Publication date:
2011

MICROBIAL ECOLOGY Springer

A closer look at *Bacteroides*: Phylogenetic relationship and genomic implications of a life in the human gut

Journal:	*Microbial Ecology*
Manuscript ID:	MECO-2010-0396
Manuscript Type:	original manuscript
Date Submitted by the Author:	01-Oct-2010
Complete List of Authors:	Karlsson, Fredrik; Systems Biology, Department of Chemical and Biological Engineering Ussery, David; Center for Biological Sequence Analysis, Dept. Systems Biology Nielsen, Jens; Systems Biology, Department of Chemical and Biological Engineering Nookaew, Intawat; Systems Biology, Department of Chemical and Biological Engineering
Key Words:	comparative genomics, Bacteroides, gut microbiota

A closer look at *Bacteroides*: Phylogenetic relationship and genomic implications of a life in the human gut

Fredrik H Karlsson[1], David W Ussery [2], Jens Nielsen[1], Intawat Nookaew[1§]

[1]Systems Biology, Department of Chemical and Biological Engineering, Chalmers University of Technology, SE412 96 Gothenburg, Sweden

[2]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, DK2800 Lyngby, Denmark

[§]Corresponding author

Email addresses:

FHK: frekar@chalmers.se

DWU: dave@cbs.dtu.dk

JN: nielsenj@chalmers.se

IN: intawat@chalmers.se

Running title: A closer look at *Bacteroides*

Date submitted: October 1st 2010

Abstract

The human gut is extremely densely inhabited by bacteria mainly from two phyla, Bacteroidetes and Firmicutes and there is a great interest in analyzing whole genome sequences for these species because of their relation to human health and disease. Here we do whole genome comparison of 105 Bacteroidetes/Chlorobi genomes to elucidate their phylogenetic relationship and to gain insight into what is separating the gut living *Bacteroides* and *Parabacteroides* genera from other Bacteroidetes/Chlorobi species.

A comprehensive analysis shows that *Bacteroides* species have a higher number of extracytoplasmic function σ-factors (ECF σ-factors) and two component systems for extracellular signal transduction compared to other Bacteroidetes/Chlorobi species. Traditional phylogenetic analysis based on 16S rRNA sequences revealed that two *Bacteroides* species are misclassified and belongs to the Firmicutes phylum. A whole genome phylogenetic analysis shows a very little difference between the *Parabacteroides* and *Bacteroides* genera. Further analysis shows that *Bacteroides* and *Parabacteroides* species share a large common core of 1085 protein families. Genome atlases illustrate that there are few and only small unique areas on the chromosomes of four *Bacteroides*/*Parabacteroides* genomes. Functional classification to clusters of othologus groups (COGs) show that *Bacteroides* species are enriched in carbohydrate transport and metabolism proteins. Classification of proteins in KEGG metabolic pathways gives a detailed view of the genome's metabolic capabilities that can be linked to its habitat.

We have presented a more detailed and precise description of the phylogenetic relationships of members of the Bacteroidetes/Chlorobi phylum by whole genome

comparison. Gut living *Bacteroides* have an enriched set of glycan, vitamin and cofactor enzymes important for diet digestion.

Background

The human intestine is host to roughly 100 trillion microbial cells, 10 times as many as human cells [21] and carry a gene set 150 times larger than the human genome [28]. The microbiota complements the human set of enzymes with important enzymatic functions such as degradation of polysaccharides and production of vitamins. The microbes have a profound impact on human health and physiology especially alteration of gut ecology has been associated with inflammatory bowels diseases and obesity [20, 25, 28, 40].

Bacteria consist of at least 27 phyla [10] but the human colon is dominated by members of only two of these, Bacteroidetes and Firmicutes make up 16% and 76% of the phylotypes and 48% and 51% of the total bacterial ribosomal RNA gene sequences, respectively [7]. An increased relative abundance of Firmicutes to Bacteroidetes in the gut is associated with obesity both in mice and humans [40-41]. To gain insight into how microbial components contribute to human health and disease the NIH funded Human Microbiome Project (nihroadmap.nih.gov/hmp/) and the EU funded MetaHIT project (www.metahit.eu/) have been established. An initial outcome from the HMP project is a catalog of 178 reference genomes and out of these, 151 were from the gastrointestinal tract [26]. This wealth of data allows us to investigate their genetic relationship as well as link genetic information to distinct behaviors by comparative analysis. Traditionally 16S rRNA sequence has been used for phylogenetic analysis for evolutionary comparison and classification. However, this approach is based on the assumption of unidirectional and hierachical evolution and no gene transfer between species. In fact, many bacteria have more than one copy of the 16S rRNA gene, and in some (rare) cases the 16S rRNA genes from operons in the same genome are different enough to be considered another species [27]. Lateral

gene transfer is a strong force in bacterial evolution, which transforms the hierarchical tree to a network of relationships between species [5]. It has been suggested that lateral gene transfer has played a major role in the evolution of the bacteria in the human intestine [44].

The genus *Bacteroides* underwent a major revision in 1989 after having been a genus generally described as a collection of obligately anaerobic, Gram-negative, nonsporing, rodshaped bacteria, was now proposed to be restricted to closely related species of *Bacteroides fragilis* based on genomic GC content and biochemical capabilities [35]. While the *Bacteroides* genus was restricted, several species were moved to new genera such as *Prevotella* [33] and *Porphyromonas* [34]. More recently further restrictions have been done to the *Bacteroides* genus and *Alistipes* and *Parabacteroides* genera have been defined to harbor these species [30-31]. Also new species have been added to the *Bacteroides* genus, *e.g. Bacteroides plebeius* and *Bacteroides corprocola,* isolated from the human gut [17]. With the large scale genomic sequencing projects mentioned above, it is likely that new *Bacteroides* species will be found that need to be classified. Members of the *Bacteroides* genus have adapted to a life in the gut of mammals. This habitat is rich in undigested polysaccharides that human enzymes are unable to digest. This fact is extensively manifested by the genomic information of the first complete genome sequence of a Bacteroidetes species, *Bacteroides thethaiotaomicron*. Its genome contains 172 glycoside hydrolases, 163 homologs of SusC and SusD outer-membrane polysaccharide-binding proteins for polysaccharide utilization [42]. The wealth of polysaccharide degrading enzymes has also been observed in 3 other *Bacteroides* species [44]. The well studied *Bacteroides thethaiotaomicron* has been found to have an unprecedented number of extracytoplasmic function σ-factors (ECF σ-factors) and

a large collection of hybrid two-component systems for environmental sensing in its genome [43]. In many cases genes for these two regulatory systems are positioned in close proximity to genes coding for glycoside hydrolases and SusC/D [43].

In this study we use bioinformatics and comparative genomics methods on 105 genomes from the Bacteroidetes/Chlorobi group to gain knowledge about the phylogeny of the member species. Further, by comparative analysis we study the gut living *Bacteroides* (33) and *Parabacteroides* (4) and compare the genetic content of these gut living organisms to their relatives in other habitats.

Methods

Publically available genomes from the Bacteroidetes/Chlorobi superphylum were downloaded from GenBank at National Center for Biotechnology Information. A full list of genomes included is presented in Supplementary Table 1. This study is based on 33 completely sequenced genomes and 72 in the assembly stage. The list contains 33 genomes from the genus *Bacteroides*, 9 from *Prevotella*, 8 from *Chlorobium* and 4 from *Parabacteroides* and *Porphyromonas* respectively and 47 genomes from other Genera in the Bacteroidetes/Chlorobi group.

Genetic components analysis

The genome sequences were predicted for their content of tRNAs and rRNAs by tRNAscan-SE [23] and RNAmmer [18] program, respectively. The prediction of sigma factors, two-component signal transduction systems, membrane proteins and secreted proteins were done following the standard methods previously published [1-3, 15-16].

16s rRNA analysis

16S rRNA sequences, which were extracted from the genomes with RNAmmer, were used to make a phylogenetic tree. Sequences of length less than 1400 nucleotides

were discarded. If several 16S rRNA sequences were found within a genome, all were used in the further analysis. Sequences were aligned using MUSCLE [9] then the MEGA4 software [38] was employed to build a phylogenetic tree. The evolutionary tree was constructed using the Neighbor-Joining method with distances using the Jukes-Cantor measure and complete deletion option. 10000 bootstrap integrations were performed to find bootstrap values. The trees were re-drawn in the FigTree software (http://tree.bio.ed.ac.uk/software/figtree/).

Protein family analysis

OrthoMCL is an algorithm to form clusters of orthologous groups from protein sequences [22]. The algorithm starts with an all-against-all BLASTP search and then uses similarity measures to identify clusters of orthologs and paralogs, using a Markov clustering algorithm. OrthoMCL verison 1.4 was used to identify protein families by a BLAST P-value cut off of 10^{-5} and MCL inflation parameter of 1.5. A matrix was constructed containing one row for each OrthoMCL cluster and one column for each species with each cell in the matrix containing the number of proteins in each cluster. A phylogenetic tree was constructed from the OrthoMCL result matrix by hierarchical clustering with an average linkage and the Manhattan distance metric. Clustering was performed in the statistical software R with the pvclust package [37]; to assess the confidence of the tree, 10000 bootstrap integrations were performed. The tree was re-drawn in the FigTree software.

Functional profiles analysis

All proteins were queried against the COG database to functionally annotate proteins [39]. The COG blast database was downloaded from NCBI FTP and psi-BLAST was used to annotate proteins to COGs with an e-value cutoff of 10^{-2}.

The KEGG database was downloaded and for each KEGG ontology, bacterial sequences were filtered out and HMM models were generated with HMMER3 [8]. All genes in the 105 Bacteroidetes/Chlorobi genomes were queried against the HMM models. A cutoff of 10^{-30} was used for statistical significance. A heatmap of each pathway and process derived from the database was constructed based on normalized abundance of the enzymes present in each pathway. The heatmap and hierarchical clustering was performed in R.

Results and Discussion

Genetic components

The 105 Bacteroidetes/Chlorobi genomes shown in Table 1 were downloaded using the NCBI project ID, and scanned for their abundance of ribosomal, sigma factor, tRNA, two-component system, trans-membrane helix and signal peptide genes. The number of genes was compared in the three groups *Bacteroides*, *Parabacteroides* and the other Bacteroidetes/Chlorobi species.

The number of tRNAs in each genome show that *Bacteroides* and *Parabacteroides* species contain a significantly higher ($p<0.01$, non-parametric Mann Whitney's U test) number of genes coding for tRNAs in their genomes compared to other Bacteroidetes/Chlorobi species (Supplemantary Figure 1). A larger number of tRNAs is an indication of a faster growth rate at optimal conditions [19] but the correlation is weak and there might be other explanations for high copy numbers of tRNAs.

The external sensory systems ECF σ-factors and two-component systems counted in the genomes as reported in Supplementary Figure 1. As expected, the *Bacteroides* had significantly larger number of ECF σ-factors in their genomes compared to other Bacteroidetes/Chlorobi species. *Bacteroides thethaiotaomicron* was found to have 50 ECF σ-factors, consistent with what has been previously described, and at the time the genome with the highest number of ECF σ-factors [43]. Here several *Bacteroides* species and other Bacteroidetes/Chlorobi species have even more ECF σ-factors, *e.g.* *Bacteroides* sp. D2 (70) and *Chitinophaga pinensis* DSM 2588 (94). No significant difference in the sigma factor 70 and 54 was found between the *Bacteroides* and other Bacteroidetes/Chlorobi. All *Bacteroides* species have one copy each of the two sigma factors except *Bacteroides capillosus* (1 σ^{54}, 5 σ^{70}) and *Bacteroides pectinophilus* (0 σ^{54}, 7 σ^{70}).

Like ECF σ-factors, two-component systems are important environmental signal transduction pathways in prokaryotes [36]. Two-component signal transduction systems consist of a histidine kinase that autophosphorylates upon environmental stimuli and a response regulator that subsequently receives the phosphoryl group at an aspartate residue [36]. Both *Bacteroides* and *Parabacteroides* have a significant higher number of genes coding for two-component system histidine kinase 1 and the response regulator.

Phylogeny of 16S ribosomal genes and orthologus clusters of protein families

The 16s rRNA phylogenetic tree (Figure 1) shows that *Bacteroides* species form one big cluster including *Bacteroides fragilis* strains, once suggested to be the definition of the *Bacteroides* genus [35] and *Bacteroides vulgatus* on another branch. Most of 16s rRNA replications in each genome exclusively cluster together, in a few cases *e.g.* *Parabacteroides distasonis* and *Bacteroides vulgatus* some of the 16s rRNA

sequences cluster with sequences from other species, *Parabacteroides* sp. D13 and *Bacteroides* sp 4_3_47FAA, respectively. The average copy number of the 16S rRNA gene is about 2 (2.3) in all Bacteroidetes/Chlorobi species and there is no significant difference between *Bacteroides*, *Parabacteroides* and other Bacteroidetes/Chlorobi species. In the *Bacteroides* genus, the maximum copy number the 16S rRNA gene is 7 in *Bacteroides vulgatus* ATCC 8482. When enumerating bacterial cells based on 16S rRNA methods, this difference in copy number is important to keep in mind.

Interestingly, *Bacteroides pectinophilus* and *Bacteroides capillosus* cluster together and are found far from the other *Bacteroides* species. The 16s rRNA sequences of *Bacteroides capillosus* have 96-98% sequence similarity with *Clostridium orbiscindens* strains and it has recently been suggested that the species should be reclassified to the novel genus *Pseudoflavonifractor* [4]. Similarly, the 16s rRNA sequences of *Bacteroides pectinophilus* have a 92% sequence similarity with *Eubacterium eligens* ATCC 27750 and *Clostridium saccharolyticum* WM1 suggesting that also this strain is classified in the wrong phylum and should belong to the Firmicutes. Generally, the resolution in the 16S rRNA tree is limited and it is impossible to discern the relationship between closely related species.

A more detailed and comprehensive view of the genomic phylogenetic relationship between the species can be seen in Figure 2 and was achieved by clustering distribution of protein families defined by the unsupervised algorithm orthoMCL [22]. Clearly, the depth of resolution is higher in the protein family tree compared to the 16S rRNA tree (Figure 1). As opposed to the 16S rRNA tree, here all the *Bacteroides* genomes cluster together with *Parabacteroides* genomes except for the *Bacteroides pectinophilus* and *Bacteroides capillosus* which are still far from other *Bacteroides*.

Parabacteroides species form a small cluster within the *Bacteroides* cluster showing high similarity with the other *Bacteroides*. *Bacteroides* sp. 2_1_33B and sp. 2_1_7 cluster tightly with *Parabacteroides* species but neither had a 16s rRNA sequence that met our quality criteria. The *Parabacteroides* genus was proposed to harbour species that showed differences in 16s rRNA sequences and different menaquinone composition compared to *Bacteroides* [31]. But at the whole genome level, our results indicate that *Parabacteroides* are a part of the *Bacteroides* genus. *Bacteroides* species clearly seem to have a shared genomic core that we try to define and contrast to other Bacteroidetes/Chlorobi species.

Pan and core genome comparisons

A pan and core genome plot was drawn based on the results from the orthoMCL protein families of *Bacteroides* and *Parabacteroides* genomes as shown in Figure 3. The pan orthoMCL protein families were defined as being represented in at least one of the studied genomes whereas the core protein families were present in all genomes. Genomes are ordered by genus but within genus the order is alphabetical except for *Bacteroides pectinophilus* and *Bacteroides capillosus* that are placed last. The number of core protein families for the 31 *Bacteroides* genomes is 1116 and for the *Bacteroides* and *Parabacteroides* it is 1085 whereas it drops dramatically for *Bacteroides pectinophilus* and *Bacteroides capillosus* to 424. The number of core protein families in the *Bacteroides*/*Parabacteroides* genus is stable and only slowly decreases when new genomes are added. However the pan protein families are growing at a much faster rate showing that each genome carries specialized genes not shared with other *Bacteroides* species. *Bacteroides pectinophilus* and *Bacteroides capillosus* genomes add a considerable number of protein families to the pan showing that they contain several novel protein families not present in other *Bacteroides* or

Parabacteroides strains. *Bacteroides* genomes share a smaller number of core protein families with *Porphyromonas* (694) and *Prevotella* (703) compared to *Parabacteroides* (1085) even though *Prevotella* seem to have closer related 16S rRNA sequences.

The *Bacteroides* core protein families were further queried for functional domains by InterPro scan [29]. To evaluate functions that are specific for *Bacteroides*, the number of genes in each core protein family was compared between the *Bacteroides* and other Bacteroidetes/Chlorobi. A subset of the protein families is not only a core in *Bacteroides* but is common to many Bacteroidetes/Chlorobi species. The common protein families are related to translation, *e.g.* ribosomal proteins and tRNA synthases necessary for basic machineries for growth (see Table 2 for details). Out of the 20 most specific core protein families in *Bacteroides,* 7 contained a signal peptide and 3 contained a transmembrane domain and 6 protein families were hypothetical proteins. The core protein families with the highest copy number were two-component systems, ECF σ-factors and hydrolase enzymes that are necessary for their life in the gut environment.

Blast atlases [11] provide an overview of chromosome arrangement of conserved regions (core) as well as variable regions (pan). Blast atlases of the four complete genomes along with aligned genomes of *Bacteroides* and *Parabacteroides* are shown in Figure 4. Again, *Bacteroides pectinophilus* and *Bacteroides capillosus* have very little conserved regions with other *Bacteroides* species. Additionally, *Parabacteroides* species show a high similarity with *Bacteroides* species and *Parabacteroides distasonis* has few unique genomic regions that are shared with other *Parabacteroides* species but not with *Bacteroides* species.

The variable gene content is not evenly distributed over the chromosome but rather is located to islands. This is especially evident for *Bacteroides fragilis* and *Bacteroides vulgatus* that contain several islands with little homology to other species. Again *Parabacteroides distasonis* is shown to be genomically similar to other *Bacteroides* species in general and particularily *Bacteroides* 2_1_33B and 2_1_7. *Bacteroides thethaiotaomicron* is seen as a generalist with a broad repertoire of glycoside hydrolase paralogs and starch utilization systems C and D paralogs [44]. However, the blast atlas shows that there are few unique regions in the genome and these are not gathered in islands but rather spread out over the chromosome.

Functional profiles of Bacteroidetes/Chlorobi

OrthoMCL is an unsupervised algorithm for finding all shared protein families among genomes; however it does not provide any functional information. By annotating genes to functional categories, *e.g.* metabolic functions, we can discern the requirements a certain habitat puts up on a genome. In Figure 5 and 6 we map all 105 Bacteroidetes/Chlorobi genomes to the curated cluster of orthologus groups (COG) [39] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [14], respectively. Functional annotation relies on inferring gene function by sequence similarity to genes with known function but is evidently limited to the size of the reference set. The number of orthoMCL protein families is 26,163 compared to 4,873 for the COG database; thus the space covered by the unsupervised algorithm is much larger, as shown in Figure 5A.

The COG database contains 4873 orthologus groups made up from 138 458 proteins from 66 unicellular organisms covering 75% of their predicted proteins [39]. Here we mapped all 105 Bacteroidetes/Chlorobi genomes to the COG database and explored the functional space of each organism, meaning that paralogs were not considered

(Figure 5B). The difference between the number of COGs in each superclass for the *Bacteroides*/*Parabacteroides* and the other Bacteroidetes/Chlorobi species was evaluated with the non-parametric Mann Whitney's U test. The largest difference can be seen in the carbohydrate transport and metabolism category implying that *Bacteroides* have better capability to utilize polysaccharides. Moreover, *Bacteroides* has a significantly broader range of enzymes. *Bacteroides* are also enriched in COG classes L, D, V, M, F and R. The distance between the core and each individual genome indicates the diversity within each category. Translation ribosomal structure and biogenesis has less diversity than Carbohydrate transport and metabolism highlighting that the former is a basic requirement for growth whereas the latter is likely related to niche specialization. The COG super classes are coarse and the importance of metabolic processes in the gut habitat led us to also annotate the genomes to the KEGG database that is comprehensively annotated for metabolic genes and pathways.

Phylogenetic analysis based on metabolic pathway reaction content has been used to elucidate trees of metabolically related species [13]. The constructed tree is only to a small extent affected by genome size and takes into account mostly essential genes since functional metabolic pathways are essential to an organism. The pathway content is related to niche specialization and habitat as these factors largely affect metabolism. Here we mapped genes to orthologs in the KEGG database and to pathways therein. Each KEGG ortholog was counted as present or absent and mapped to its respective pathway. In Figure 6 a heatmap and phylogenetic tree is presented of the Bacteroidetes/Chlorobi species. The functional annotation results from KEGG and COG agree well, but KEGG gives a much more detailed view of metabolism.

The heatmap gives a detailed view of the metabolic capabilities of each species, which can be related to their natural habitat. *Bacteroides* species mostly group together and it is evident that they are enriched in carbohydrate acting enzymes and also glycan, vitamin and cofactor metabolism. *Prevotella bergensis*, isolated from human skin [6], and *Prevotella copri*, isolated from human faeces [12], group with gut living *Bacteroides* and *Parabacteroides*. *Bacteroides pecinophilus* and *Bacteroides capillosus* group together and distinctly from the other *Bacteroides* species as seen in the previous analysis but still these organisms are living in the human gut as two of the 50 most abundant species [28]. This lack of consensus among gut living species likely means that the human gut habitat is not homogeneous but rather contains several niches. This is also consistent with results found by two studies in gnotobiotic mice with *Bacteroides thetaiotaomicron* and one member of the Fimicutes phylumn and a methanogenic archae [24, 32]. *Bacteroides thethaiotaomicron* is the primary fermenter of polysaccharides whereas the Firmicute and Archae use simple sugars and fermentative products such as acetate and H_2.

However, in general aerobic free-living species in water or soil group together, shown in Figure 6 marked with blue/brown. Unculturable intracellular symbionts *Sulcia mulleri* and *Blattabacterium* species group together as these genomes contains very few proteins and thus has low abundance of enzymes in each pathway. The clade marked with yellow contains mainly *Prevotella, Porphyromonas* and *Capnocytophaga* species, all living in the human oral cavity or on human skin. In summary, the phylogenetic analysis based on metabolic pathway content can indicate a genome's habitat.

Conclusions

Here we have shown how a whole genome analysis can improve phylogenetic studies based on 16S rRNA sequence analysis. Unsupervised clustering of orthologus groups such as is done with the orthoMCL algorithm is useful when classifying species and analyzing orthologus genes and we have presented a phylogenetic tree of 105 species in the Bacteroidetes/Chlorobi phyla. Functional annotation of genes to high quality curated databases such as COG and KEGG gives detailed information about pathway content but does not account for genes with unknown functions. From this analysis we found that Bacteroides have enrichment in carbohydrate acting enzymes and also vitamin and cofactor metabolism, indicating that these bacteria have adapted to a role of diet digestion and vitamin production. Parabacteroides species show a high similarity with Bacteroides by sharing a high number of protein families and functional characteristics, likely because they share habitat.

With the enormous amount of data that is generated from microbes inhabiting the human body, with a gene set 150 times larger than the human genome, there is certainly a need to categorize it and analyze the genomic information. Comparative genomic analyses will play an important role in better understanding the microbiota.

Acknowledgements

We gratefully acknowledge the Knut and Alice Wallenberg Foundation and the Chalmers Foundation for financial support. We thank Dina Petranovic for useful discussion for the manuscript

Figure Legends

Figure 1 - Phylogenetic tree based on 16S rRNA sequence

Bacteroides sequences are red except for sequences from *Bacteroides* capillosus and *Bacteroides* pectinophilus which are blue, *Parabacteroides* sequences are orange and other species are black. Bootstrap values indicate the certainty of each cluster.

Figure 2 - Phylogenetic tree based on whole genome orthoMCL clusters

The two genuses *Bacteroides* and *Parabacteroides* are not separated in this tree but cluster together. The colors highlighting the species are the same as in Figure 1.

Figure 3 - Pan- and core genome plot of *Bacteroides* and *Parabacteroides* genomes

The blue line (core) represents the conserved number of orthoMCL protein families. The red line (pan) indicates the cumulative number of orthoMCL protein families in the genomes. Green bar indicate the number of novel orthoMCL protein families in the genome. The relative size of the core protein families to the total genome size (% Core) is based on the 1085 protein families shared by *Bacteroides* and *Parabacteroides* (excluding *Bacteroides capillosus* and *Bacteroides pectinophilus*), On average, 27% of the proteins is shared in the core protein families.

Figure 4 - Blast atlas of *Bacteroides* and *Parabacteroides* genomes

The reference genome is indicated in the center of each circle. Other *Bacteroides* and *Parabacteroides* genomes are outlined along the chromosome with different color intensity based on sequence similarity assessed with a BLASTp score. The order of the genomes is the same as in Figure 3 except that the reference is excluded. The colors highlighting the species are the same as in Figure 1.

Figure 5 - COG functional space

Each genome was annotated to the COG database. The white bars indicate the total space of the respective COG class. The number of COGs in each class was indicated with a line in the bar for each genome. The red bars represent the core COG space present in the *Bacteroides* genomes i.e. the number of COGs present in all genomes. The significance level based on the Mann Whitney U-test between *Bacteroides* and other genomes is indicated by asterisk (* $p<10^{-2}$, ** $p<10^{-5}$, *** $p<10^{-10}$)

Figure 6 - Phylogenetic tree based on KEGG pathway content

The relative abundance of genes in each pathway is depicted in the heat map where each row is normalized. Species are clustered based on their relative pathway content. The colors highlighting the species are the same as in Figure 1. Habitat of isolation as stated by the NCBI genome project is indicated with color accordingly: human skin/genitals/oral (yellow) human gut (purple), intracellular endosymbiont (pink), aquatic (blue), soil (brown).

Tables

Table 1 - Bacteroidetes/Chlorobi genomes in this study

Table 2 - Core gene families in Bacteroides with high copy number

Additional files

Additional file 1 – Supplementary figure 1

References

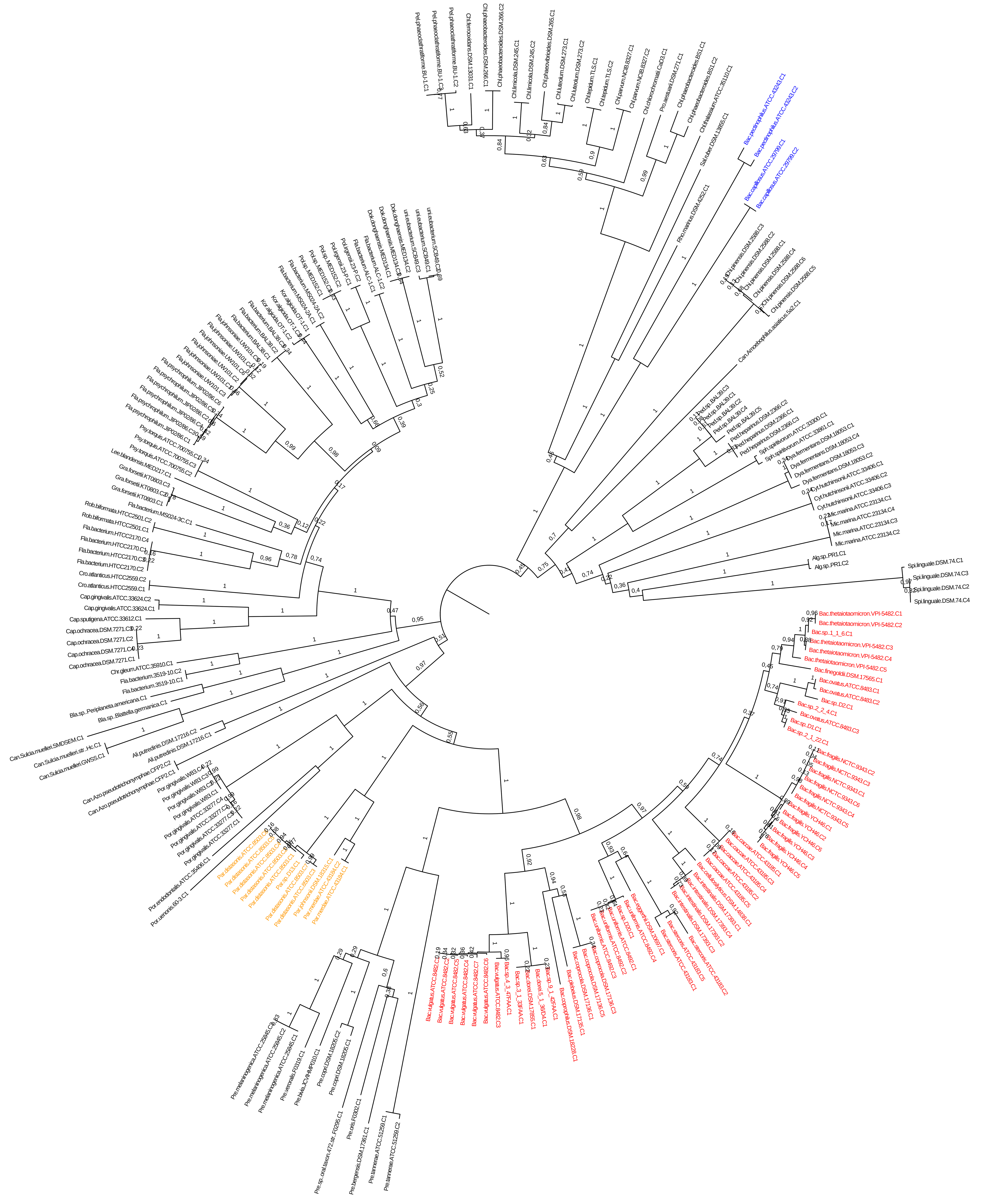
1. Bendtsen JD, Binnewies TT, Hallin PF, Sicheritz-Ponten T, Ussery DW (2005) Genome update: prediction of secreted proteins in 225 bacterial proteomes. Microbiology 151: 1725-1727

2. Bendtsen JD, Binnewies TT, Hallin PF, Ussery DW (2005) Genome update: prediction of membrane proteins in prokaryotic genomes. Microbiology 151: 2119-2121
3. Binnewies TT, Bendtsen JD, Hallin PF, Nielsen N, Wassenaar TM, Pedersen MB, Klemm P, Ussery DW (2005) Genome Update: Protein secretion systems in 225 bacterial genomes. Microbiology 151: 1013-1016
4. Carlier JP, Bedora-Faure M, K'Ouas G, Alauzet C, Mory F (2010) Proposal to unify Clostridium orbiscindens Winter et al. 1991 and Eubacterium plautii (Seguin 1928) Hofstad and Aasjord 1982, with description of Flavonifractor plautii gen. nov., comb. nov., and reassignment of Bacteroides capillosus to Pseudoflavonifractor capillosus gen. nov., comb. nov. Int J Syst Evol Microbiol 60: 585-590
5. Doolittle WF (1999) Phylogenetic classification and the universal tree. Science 284: 2124-2129
6. Downes J, Sutcliffe IC, Hofstad T, Wade WG (2006) Prevotella bergensis sp. nov., isolated from human infections. Int J Syst Evol Microbiol 56: 609-612
7. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA (2005) Diversity of the human intestinal microbial flora. Science 308: 1635-1638
8. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755-763
9. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797
10. Garrity GM, Lilburn TG, Cole JR, Harrison SH, Euzeby J, Tindall BJ (2007) Introduction to the Taxonomic Oultine of Bacteria and Archaea (TOBA) Release 7.7
11. Hallin PF, Binnewies TT, Ussery DW (2008) The genome BLASTatlas-a GeneWiz extension for visualization of whole-genome homology. Mol Biosyst 4: 363-371
12. Hayashi H, Shibata K, Sakamoto M, Tomita S, Benno Y (2007) Prevotella copri sp. nov. and Prevotella stercorea sp. nov., isolated from human faeces. Int J Syst Evol Microbiol 57: 941-946
13. Hong SH, Kim TY, Lee SY (2004) Phylogenetic analysis based on genome-scale metabolic pathway reaction content. Appl Microbiol Biotechnol 65: 203-210
14. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32: D277-280
15. Kiil K, Ferchaud JB, David C, Binnewies TT, Wu H, Sicheritz-Ponten T, Willenbrock H, Ussery DW (2005) Genome update: distribution of two-component transduction systems in 250 bacterial genomes. Microbiology 151: 3447-3452
16. Kill K, Binnewies TT, Sicheritz-Ponten T, Willenbrock H, Hallin PF, Wassenaar TM, Ussery DW (2005) Genome update: sigma factors in 240 bacterial genomes. Microbiology 151: 3147-3150
17. Kitahara M, Sakamoto M, Ike M, Sakata S, Benno Y (2005) Bacteroides plebeius sp. nov. and Bacteroides coprocola sp. nov., isolated from human faeces. Int J Syst Evol Microbiol 55: 2143-2147
18. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35: 3100-3108

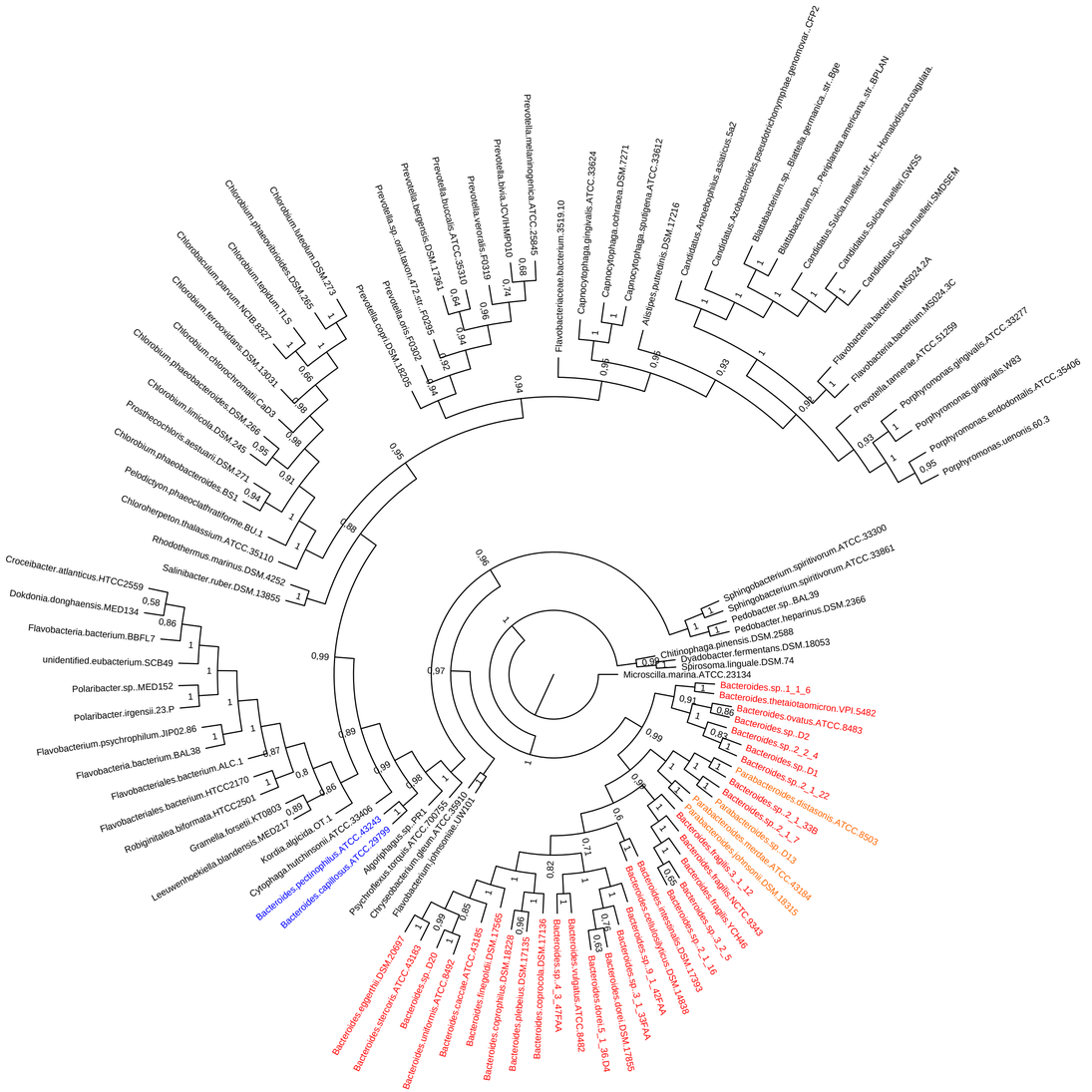
19. Lee ZM, Bussema C, 3rd, Schmidt TM (2009) rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea. Nucleic Acids Res 37: D489-493
20. Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI (2005) Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A 102: 11070-11075
21. Ley RE, Peterson DA, Gordon JI (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell 124: 837-848
22. Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178-2189
23. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25: 955-964
24. Mahowald MA, Rey FE, Seedorf H, Turnbaugh PJ, Fulton RS, Wollam A, Shah N, Wang C, Magrini V, Wilson RK, Cantarel BL, Coutinho PM, Henrissat B, Crock LW, Russell A, Verberkmoes NC, Hettich RL, Gordon JI (2009) Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla. Proc Natl Acad Sci U S A 106: 5859-5864
25. Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L, Nalin R, Jarrin C, Chardon P, Marteau P, Roca J, Dore J (2006) Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut 55: 205-211
26. Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, Rusch DB, Mitreva M, Sodergren E, Chinwalla AT, Feldgarden M, Gevers D, Haas BJ, Madupu R, Ward DV, Birren BW, Gibbs RA, Methe B, Petrosino JF, Strausberg RL, Sutton GG, White OR, Wilson RK, Durkin S, Giglio MG, Gujja S, Howarth C, Kodira CD, Kyrpides N, Mehta T, Muzny DM, Pearson M, Pepin K, Pati A, Qin X, Yandava C, Zeng Q, Zhang L, Berlin AM, Chen L, Hepburn TA, Johnson J, McCorrison J, Miller J, Minx P, Nusbaum C, Russ C, Sykes SM, Tomlinson CM, Young S, Warren WC, Badger J, Crabtree J, Markowitz VM, Orvis J, Cree A, Ferriera S, Fulton LL, Fulton RS, Gillis M, Hemphill LD, Joshi V, Kovar C, Torralba M, Wetterstrand KA, Abouellleil A, Wollam AM, Buhay CJ, Ding Y, Dugan S, FitzGerald MG, Holder M, Hostetler J, Clifton SW, Allen-Vercoe E, Earl AM, Farmer CN, Liolios K, Surette MG, Xu Q, Pohl C, Wilczek-Boney K, Zhu D (2010) A catalog of reference genomes from the human microbiome. Science 328: 994-999
27. Pei AY, Oberdorf WE, Nossa CW, Agarwal A, Chokshi P, Gerz EA, Jin Z, Lee P, Yang L, Poles M, Brown SM, Sotero S, Desantis T, Brodie E, Nelson K, Pei Z (2010) Diversity of 16S rRNA genes within individual prokaryotic genomes. Appl Environ Microbiol 76: 3886-3897
28. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Jian M, Zhou Y, Li Y, Zhang X, Qin N, Yang H, Wang J, Brunak S, Dore J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Bork P, Ehrlich SD (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59-65

29. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) InterProScan: protein domains identifier. Nucleic Acids Res 33: W116-120
30. Rautio M, Eerola E, Vaisanen-Tunkelrott ML, Molitoris D, Lawson P, Collins MD, Jousimies-Somer H (2003) Reclassification of Bacteroides putredinis (Weinberg et al., 1937) in a new genus Alistipes gen. nov., as Alistipes putredinis comb. nov., and description of Alistipes finegoldii sp. nov., from human sources. Syst Appl Microbiol 26: 182-188
31. Sakamoto M, Benno Y (2006) Reclassification of Bacteroides distasonis, Bacteroides goldsteinii and Bacteroides merdae as Parabacteroides distasonis gen. nov., comb. nov., Parabacteroides goldsteinii comb. nov. and Parabacteroides merdae comb. nov. Int J Syst Evol Microbiol 56: 1599-1605
32. Samuel BS, Gordon JI (2006) A humanized gnotobiotic mouse model of host-archaeal-bacterial mutualism. Proc Natl Acad Sci U S A 103: 10011-10016
33. SHAH HN, COLLINS DM (1990) NOTES: Prevotella, a New Genus To Include Bacteroides melaninogenicus and Related Species Formerly Classified in the Genus Bacteroides. Int J Syst Bacteriol 40: 205-208
34. SHAH HN, COLLINS MD (1988) Proposal for Reclassification of Bacteroides asaccharolyticus, Bacteroides gingivalis, and Bacteroides endodontalis in a New Genus, Porphyromonas. Int J Syst Bacteriol 38: 128-131
35. SHAH HN, COLLINS MD (1989) Proposal To Restrict the Genus Bacteroides (Castellani and Chalmers) to Bacteroides fragilis and Closely Related Species. Int J Syst Bacteriol 39: 85-87
36. Stock AM, Robinson VL, Goudreau PN (2000) Two-component signal transduction. Annu Rev Biochem 69: 183-215
37. Suzuki R, Shimodaira H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22: 1540-1542
38. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596-1599
39. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41
40. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI (2009) A core gut microbiome in obese and lean twins. Nature 457: 480-484
41. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444: 1027-1031
42. Xu J, Bjursell MK, Himrod J, Deng S, Carmichael LK, Chiang HC, Hooper LV, Gordon JI (2003) A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. Science 299: 2074-2076
43. Xu J, Chiang HC, Bjursell MK, Gordon JI (2004) Message from a human gut symbiont: sensitivity is a prerequisite for sharing. Trends Microbiol 12: 21-28
44. Xu J, Mahowald MA, Ley RE, Lozupone CA, Hamady M, Martens EC, Henrissat B, Coutinho PM, Minx P, Latreille P, Cordum H, Van Brunt A, Kim K, Fulton RS, Fulton LA, Clifton SW, Wilson RK, Knight RD, Gordon JI

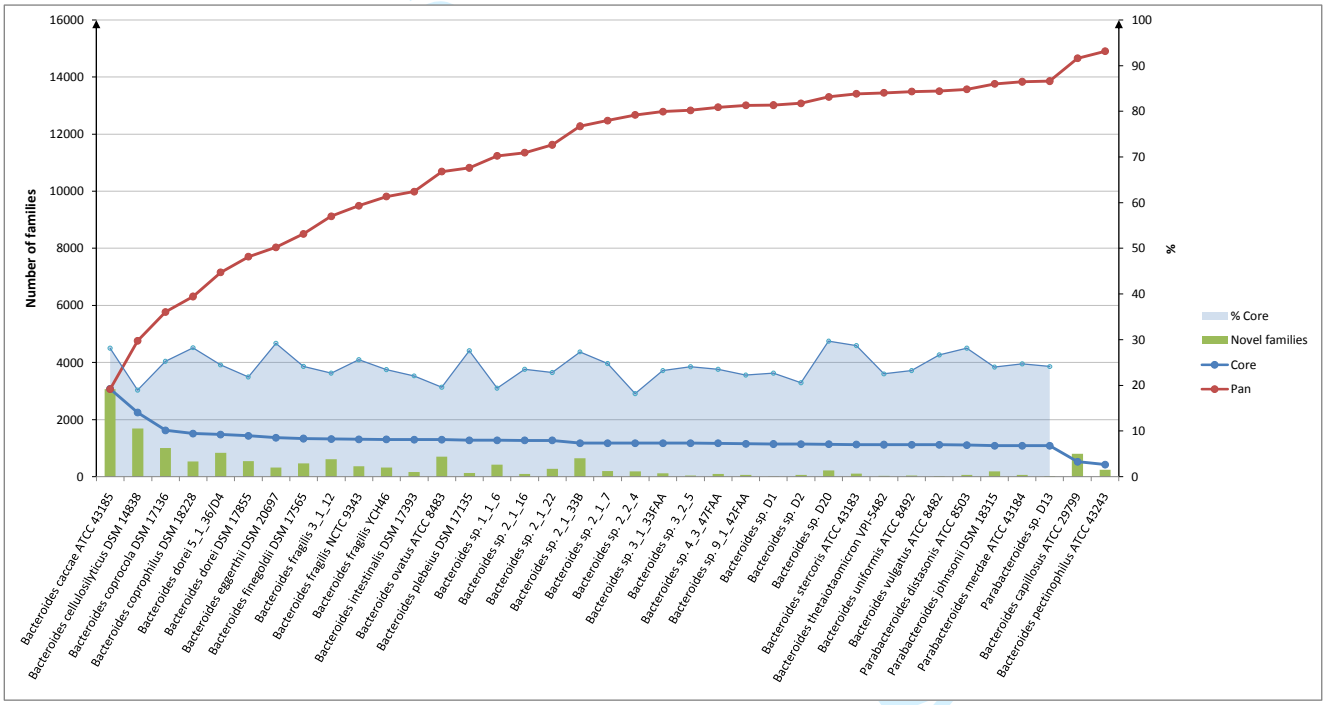
(2007) Evolution of symbiotic bacteria in the distal human intestine. PLoS Biol 5: e156

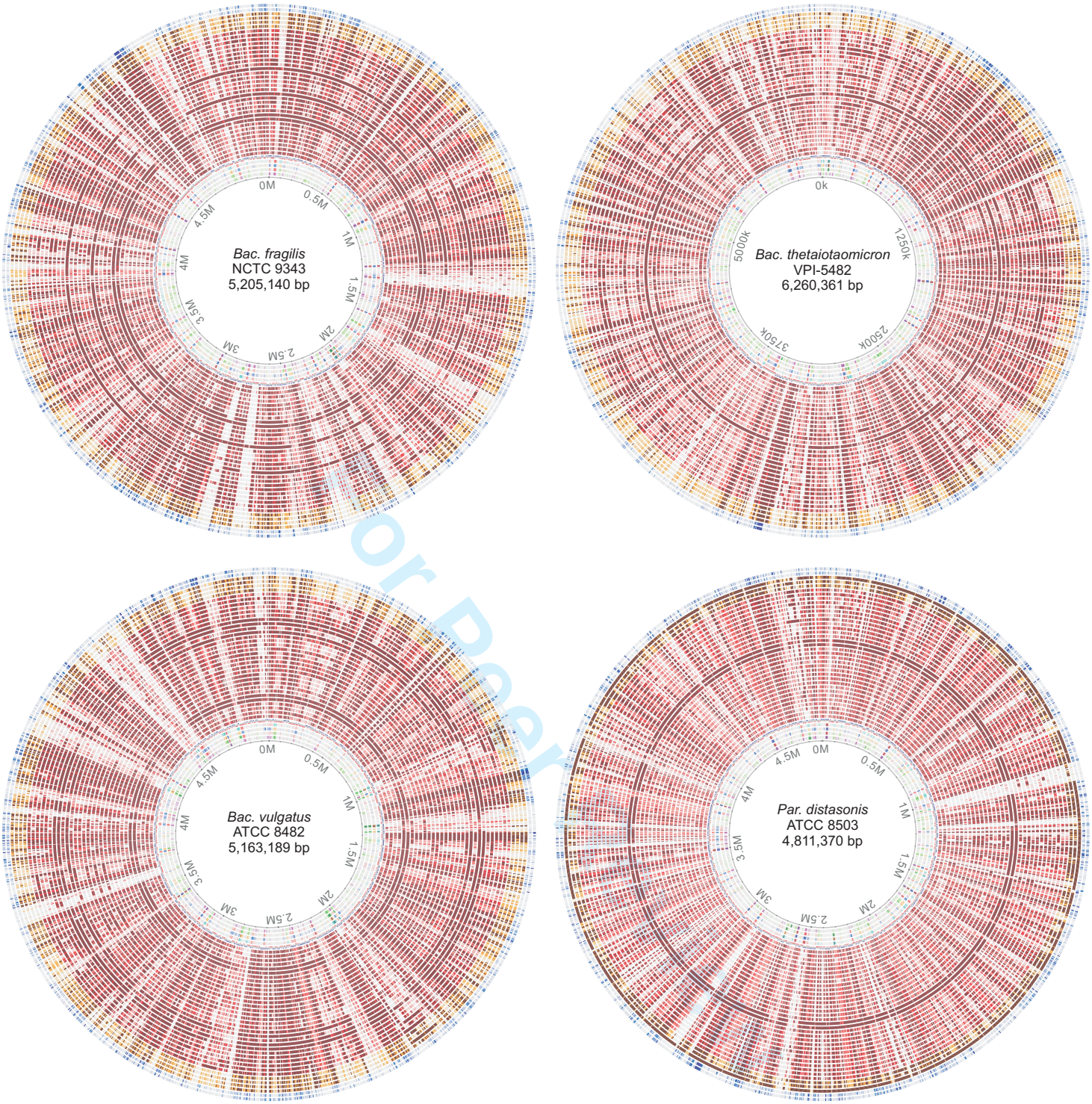


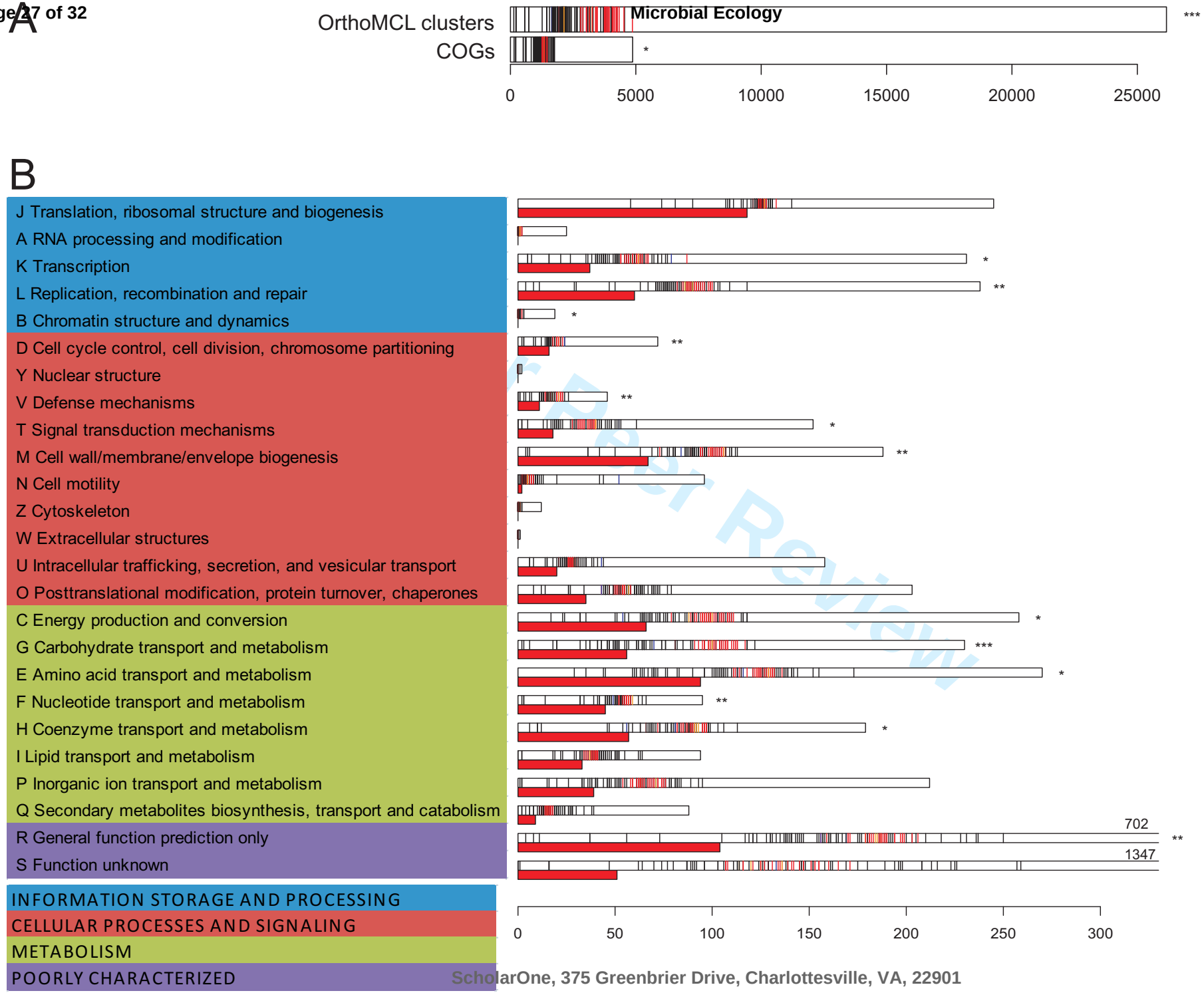
0.02


Prevotella.melaninogenica.ATCC.25845
Prevotella.bivia.JCVIHMP010
Prevotella.veroralis.F0319
Prevotella.buccalis.ATCC.35310
Prevotella.sp..oral.taxon.472.str..F0295
Prevotella.bergensis.DSM.17361
Prevotella.oris.F0302
Prevotella.copri.DSM.18205
Flavobacteriaceae.bacterium.3519.10
Capnocytophaga.gingivalis.ATCC.33624
Capnocytophaga.ochracea.DSM.7271
Capnocytophaga.sputigena.ATCC.33612
Alistipes.putredinis.DSM.17216
Candidatus.Azobacteroides.pseudotrichonymphae.genomovar..CFP2
Candidatus.Azobacteroides.pseudotrichonymphae.genomovar..CFP1
Blattabacterium.sp..Blattella.germanica..str..Bge
Blattabacterium.sp..Periplaneta.americana..str..BPLAN
Candidatus.Sulcia.muelleri.str..Hc..Homalodisca.coagulata.
Candidatus.Sulcia.muelleri.GWSS
Candidatus.Sulcia.muelleri.SMDSEM
Flavobacteria.bacterium.MS024.3A
Flavobacteria.bacterium.MS024.3C
Prevotella.tannerae.ATCC.51259
Porphyromonas.gingivalis.ATCC.33277
Porphyromonas.gingivalis.W83
Porphyromonas.endodontalis.ATCC.35406
Porphyromonas.uenonis.60.3
Chlorobium.luteolum.DSM.273
Chlorobium.phaeobacteroides.DSM.266
Chlorobium.tepidum.TLS
Chlorobaculum.parvum.NCIB.8327
Chlorobium.ferrooxidans.DSM.13031
Chlorobium.chlorochromatii.CaD3
Chlorobium.phaeobacteroides.DSM.266
Chlorobium.limicola.DSM.245
Prosthecochloris.aestuarii.DSM.271
Chlorobium.phaeobacteroides.BS1
Pelodictyon.phaeoclathratiforme.BU.1
Chloroherpeton.thalassium.ATCC.35110
Rhodothermus.marinus.DSM.4252
Salinibacter.ruber.DSM.13855
Croceibacter.atlanticus.HTCC2559
Dokdonia.donghaensis.MED134
Flavobacteria.bacterium.BBFL7
unidentified.eubacterium.SCB49
Polaribacter.sp..MED152
Polaribacter.irgensii.23.P
Flavobacterium.psychrophilum.JIP02.86
Flavobacteria.bacterium.BAL38
Flavobacteriales.bacterium.ALC.1
Flavobacteriales.bacterium.HTCC2170
Flavobacteriales.bacterium.HTCC2501
Robiginitalea.biformata.HTCC2501
Gramella.forsetii.KT0803
Leeuwenhoekiella.blandensis.MED217
Cytophaga.hutchinsonii.ATCC.33406
Kordia.algicida.OT.1
Bacteroides.pectinophilus.ATCC.43243
Bacteroides.capillosus.ATCC.29799
Algoriphagus.sp..PR1
Psychroflexus.torquis.ATCC.700755
Chryseobacterium.gleum.ATCC.35910
Flavobacterium.johnsoniae.UW101
Sphingobacterium.spiritivorum.ATCC.33300
Sphingobacterium.spiritivorum.ATCC.33861
Sphingobacter.sp..BAL39
Pedobacter.heparinus.DSM.2366
Chitinophaga.pinensis.DSM.2588
Cytophaga.hutchinsonii.ATCC.33406
Spirosoma.linguale.DSM.74
Microscilla.marina.ATCC.23134
Bacteroides.sp..1_1_6
Bacteroides.thetaiotaomicron.VPI.5482
Bacteroides.ovatus.ATCC.8483
Bacteroides.sp..D2
Bacteroides.sp..2_2_4
Bacteroides.sp..D1
Bacteroides.sp.2_1_22
Parabacteroides.distasonis.ATCC.8503
Bacteroides.sp..2_1_33B
Bacteroides.sp..D13
Parabacteroides.merdae.ATCC.43184
Parabacteroides.johnsonii.DSM.18315
Bacteroides.fragilis.3_1_12
Bacteroides.fragilis.NCTC.9343
Bacteroides.fragilis.YCH46
Bacteroides.sp..2_1_16
Bacteroides.sp..3_2_5
Bacteroides.helcogenes.DSM.17393
Bacteroides.cellulosilyticus.DSM.14838
Bacteroides.sp..9_1_42FAA
Bacteroides.sp..3_1_33FAA
Bacteroides.dorei.DSM.17855
Bacteroides.dorei.5_1_36.D4
Bacteroides.vulgatus.ATCC.8482
Bacteroides.sp..4_3_47FAA
Bacteroides.coprocola.DSM.17136
Bacteroides.plebeius.DSM.17135
Bacteroides.coprophilus.DSM.18228
Bacteroides.finegoldii.DSM.17565
Bacteroides.caccae.ATCC.43185
Bacteroides.uniformis.ATCC.8492
Bacteroides.sp..D20
Bacteroides.stercoris.ATCC.43183
Bacteroides.eggerthii.DSM.20697

3.0


Number of families
%
16000
14000
12000
10000
8000
6000
4000
2000
0
100
90
80
70
60
50
40
30
20
10
0
% Core
Novel families
Core
Pan
Bacteroides caccae ATCC 43185
Bacteroides cellulosilyticus DSM 14838
Bacteroides coprocola DSM 17136
Bacteroides coprophilus DSM 18228
Bacteroides dorei 5_1_36/D4
Bacteroides dorei DSM 17855
Bacteroides eggerthii DSM 20697
Bacteroides finegoldii DSM 17565
Bacteroides fragilis 3_1_12
Bacteroides fragilis NCTC 9343
Bacteroides fragilis YCH46
Bacteroides intestinalis DSM 17393
Bacteroides ovatus ATCC 8483
Bacteroides plebeius DSM 17135
Bacteroides sp. 1_1_6
Bacteroides sp. 2_1_16
Bacteroides sp. 2_1_22
Bacteroides sp. 2_1_33B
Bacteroides sp. 2_1_7
Bacteroides sp. 2_2_4
Bacteroides sp. 3_1_33FAA
Bacteroides sp. 3_2_5
Bacteroides sp. 4_3_47FAA
Bacteroides sp. 9_1_42FAA
Bacteroides sp. D1
Bacteroides sp. D2
Bacteroides sp. D20
Bacteroides stercoris ATCC 43183
Bacteroides thetaiotaomicron VPI-5482
Bacteroides uniformis ATCC 8492
Bacteroides vulgatus ATCC 8482
Parabacteroides distasonis ATCC 8503
Parabacteroides johnsonii DSM 18315
Parabacteroides merdae ATCC 43184
Parabacteroides sp. D13
Bacteroides capillosus ATCC 29799
Bacteroides pectinophilus ATCC 43243


Bac. fragilis
NCTC 9343
5,205,140 bp
0M
0.5M
1M
1.5M
2M
2.5M
3M
3.5M
4M
4.5M
Bac. thetaiotaomicron
VPI-5482
6,260,361 bp
0k
1250k
2500k
3750k
5000k
Bac. vulgatus
ATCC 8482
5,163,189 bp
Par. distasonis
ATCC 8503
4,811,370 bp
Position Preference
dev
avg
0.13
0.15
Stacking Energy
-8.32
-7.13
Intrinsic Curvature
0.16
0.23
Percent AT
0.46
0.68
tRNA
rRNA
CDS -
CDS +
Bacteroides
Parabacteroides
Bacteroides
fix
avg
0.00
0.60




Z-score
−5
0
5
Taxonomy
Habitat
Metabolism
Carbohydrate
Energy
Lipid
Nucleotide
Amino Acid
Other Amino Acid
Glycan
Cofactors & Vitamins
Terpenoids & Polyketides
Xenobiotics Biodegradation
Secondary Metabolites
Genetic Information Processing
Environmental Information Processing
Cellular Processes

Organism	Proteins	Status	NCBI project ID	Accession number
Bacteroides fragilis NCTC 9343	4231	c	46	CR626927.1
Porphyromonas gingivalis W83	1909	c	48	AE015924.1
Cytophaga hutchinsonii ATCC 33406	3785	c	54	CP000383.1
Chlorobium tepidum TLS	2245	c	302	AE006470.1
Bacteroides thetaiotaomicron VPI-5482	4816	c	399	AE015928.1
Chlorobium limicola DSM 245	2434	c	12606	CP001097.1
Chlorobium phaeovibrioides DSM 265	1753	c	12607	CP000607.1
Chlorobium phaeobacteroides BS1	2469	c	12608	CP001101.1
Chlorobium phaeobacteroides DSM 266	2650	c	12609	CP000492.1
Prosthecochloris aestuarii DSM 271	2327	c	12749	CP001108.1
Pelodictyon phaeoclathratiforme BU-1	2707	c	13011	CP001110.1
Chlorobium luteolum DSM 273	2083	c	13012	CP000096.1
Bacteroides fragilis YCH46	4625	c	13067	AP006841.1
Bacteroides vulgatus ATCC 8482	4065	c	13378	CP000139.1
Microscilla marina ATCC 23134	8319	a	13411	NZ_AAWS00000000
Polaribacter irgensii 23-P	2557	a	13451	NZ_AAOG00000000
Robiginitalea biformata HTCC2501	3209	c	13461	CP001712.1
Parabacteroides distasonis ATCC 8503	3850	c	13485	CP000140.1
Psychroflexus torquis ATCC 700755	6751	a	13542	NZ_AAPR00000000
Polaribacter sp. MED152	2611	a	13543	NZ_AANA00000000
Dokdonia donghaensis MED134	2944	a	13544	NZ_AAMZ00000000
Croceibacter atlanticus HTCC2559	2719	a	13570	NZ_AAMP00000000
Leeuwenhoekiella blandensis MED217	3735	a	13573	NZ_AANC00000000
Flavobacteriales bacterium HTCC2170	3478	a	13595	NZ_AAOC00000000
Flavobacteria bacterium BBFL7	2587	a	13604	NZ_AAPD00000000
Chlorobium chlorochromatii CaD3	2002	c	13921	CP000108.1
Flavobacterium johnsoniae UW101	5017	c	16082	CP000685.1
Salinibacter ruber DSM 13855	2833	c	16159	CP000159.1
Candidatus Sulcia muelleri str. Hc (Homalodisca coagulata)	179	a	16198	NZ_AANL00000000
Chlorobium ferrooxidans DSM 13031	2158	a	16644	NZ_AASE00000000
Bacteroides caccae ATCC 43185	3855	a	18163	NZ_AAVM00000000
Bacteroides capillosus ATCC 29799	4833	a	18173	NZ_AAXG00000000
Bacteroides ovatus ATCC 8483	5536	a	18191	NZ_AAXF00000000
Parabacteroides merdae ATCC 43184	4384	a	18193	NZ_AAXE00000000
Bacteroides uniformis ATCC 8492	4663	a	18195	NZ_AAYH00000000
Algoriphagus sp. PR1	4215	a	18947	NZ_AAXU00000000
Flavobacteria bacterium BAL38	2612	a	18953	NZ_AAXX00000000
Porphyromonas gingivalis ATCC 33277	2090	c	19051	AP009380.1
Gramella forsetii KT0803	3584	c	19061	CU207366.1
Flavobacteriales bacterium ALC-1	3445	a	19307	NZ_ABHI00000000
Kordia algicida OT-1	4514	a	19315	NZ_ABIB00000000

Pedobacter sp. BAL39	5101	a	19337	NZ_ABCM00000000
unidentified eubacterium SCB49	2948	a	19389	NZ_ABCO00000000
Candidatus Sulcia muelleri GWSS	227	c	19617	CP000770.2
Alistipes putredinis DSM 17216	2742	a	19655	NZ_ABFK00000000
Bacteroides stercoris ATCC 43183	3777	a	19859	NZ_ABFZ00000000
Flavobacterium psychrophilum JIP02/86	2412	c	19979	AM398681.1
Candidatus Amoebophilus asiaticus 5a2	1283	c	19981	CP001102.1
Bacteroides coprocola DSM 17136	4291	a	20521	NZ_ABIY00000000
Bacteroides intestinalis DSM 17393	4911	a	20523	NZ_ABJL00000000
Dyadobacter fermentans DSM 18053	5719	c	20829	CP001619.1
Bacteroides finegoldii DSM 17565	4485	a	27823	NZ_ABXI00000000
Bacteroides pectinophilus ATCC 43243	3246	a	27825	NZ_ABVQ00000000
Bacteroides eggerthii DSM 20697	3711	a	27827	NZ_ABVO00000000
Bacteroides plebeius DSM 17135	3933	a	27829	NZ_ABQC00000000
Bacteroides dorei DSM 17855	4966	a	27831	NZ_ABWZ00000000
Pedobacter heparinus DSM 2366	4252	c	27949	CP001681.1
Chitinophaga pinensis DSM 2588	7192	c	27951	CP001699.1
Flavobacteria bacterium MS024-2A	1772	a	28049	NZ_ABVV00000000
Flavobacteria bacterium MS024-3C	1384	a	28051	NZ_ABVW00000000
Spirosoma linguale DSM 74	6524	a	28817	CP001769
Candidatus Azobacteroides pseudotrichonymphae genomovar. CFP2	852	c	29025	AP010656.1
Chlorobaculum parvum NCIB 8327	2043	c	29213	CP001099.1
Chloroherpeton thalassium ATCC 35110	2710	c	29215	CP001100.1
Rhodothermus marinus DSM 4252	2766	a	29281	CP001807
Capnocytophaga ochracea DSM 7271	2171	c	29403	CP001632.1
Parabacteroides johnsonii DSM 18315	4515	a	30007	NZ_ABYH00000000
Prevotella copri DSM 18205	3337	a	30025	NZ_ACBX00000000
Bacteroides cellulosilyticus DSM 14838	5719	a	30027	NZ_ACCH00000000
Bacteroides coprophilus DSM 18228	3838	a	30371	NZ_ACBW00000000
Chryseobacterium gleum ATCC 35910	5296	a	30953	NZ_ACKQ00000000
Capnocytophaga sputigena ATCC 33612	2672	a	30997	NZ_ABZV00000000
Blattabacterium sp. (Blattella germanica) str. Bge	586	a	31103	CP001487
Prevotella bivia JCVIHMP010	2041	a	31377	ADFO00000000
Prevotella melaninogenica ATCC 25845	2509	a	31383	NZ_ACSI00000000
Porphyromonas endodontalis ATCC 35406	1965	a	31385	NZ_ACNN00000000
Capnocytophaga gingivalis ATCC 33624	2588	a	31387	NZ_ACLQ00000000
Sphingobacterium spiritivorum ATCC 33300	4925	a	31529	NZ_ACHB00000000
Sphingobacterium spiritivorum ATCC 33861	4567	a	31531	NZ_ACHA00000000
Bacteroides fragilis 3_1_12	4776	a	32433	NZ_ABZX00000000
Bacteroides sp. 1_1_6	5594	a	32435	NZ_ACIC00000000
Bacteroides sp. 2_1_7	4372	a	32437	NZ_ABZY00000000
Bacteroides sp. 2_2_4	5959	a	32439	NZ_ABZZ00000000

Bacteroides sp. 3_2_5	4505	a	32441	NZ_ACIB00000000
Bacteroides sp. 4_3_47FAA	4613	a	32443	NZ_ACDR00000000
Bacteroides sp. 9_1_42FAA	4871	a	32445	NZ_ACAA00000000
Bacteroides sp. D1	4785	a	32447	NZ_ACAB00000000
Bacteroides sp. D2	5264	a	32449	NZ_ACGA00000000
Bacteroides dorei 5_1_36/D4	4431	a	32451	NZ_ACDI00000000
Blattabacterium sp. (Periplaneta americana) str. BPLAN	577	a	32975	CP001429
Prevotella tannerae ATCC 51259	2811	a	33153	NZ_ACIJ00000000
Candidatus Sulcia muelleri SMDSEM	242	c	33829	CP001605.1
Porphyromonas uenonis 60-3	1977	a	34101	NZ_ACLR00000000
Prevotella bergensis DSM 17361	2825	a	34637	NZ_ACKS00000000
Prevotella oris F0302	3316	a	38329	NZ_ACUZ00000000
Prevotella veroralis F0319	3048	a	38331	NZ_ACVA00000000
Bacteroides sp. 2_1_16	4609	a	38347	ACPP00000000
Bacteroides sp. 2_1_22	4748	a	38349	ACPQ00000000
Bacteroides sp. 2_1_33B	3966	a	38351	ACPR00000000
Bacteroides sp. 3_1_33FAA	4666	a	38353	ACPS00000000
Bacteroides sp. D20	3652	a	38355	ACPT00000000
Parabacteroides sp. D13	4494	a	38359	NZ_ACPW00000000
Flavobacteriaceae bacterium 3519-10	2534	c	38559	CP001673.1
Prevotella sp. oral taxon 472 str. F0295	3092	a	38731	ACZS00000000
Prevotella buccalis ATCC 35310	2456	a	40669	ADEG00000000

Protein family	Seq. Description	Seq. Length	Mean number in Bacteroides
ORTHOMCL0	two-component system sensor histidine kinase response regulator	1353	10.6
ORTHOMCL4	beta-galactosidase	1292	5.3
ORTHOMCL2	family multidrug resistance protein	1072	5.0
ORTHOMCL5	alpha- -mannosidase	1250	4.5
ORTHOMCL39	rna polymerase ecf-type sigma factor	171	2.9
ORTHOMCL71	alpha-glucosidase	687	2.5
ORTHOMCL14	propionyl- carboxylase subunit beta	517	2.4
ORTHOMCL6	two-component system response regulator	242	2.4
ORTHOMCL86	galactoside o-acetyltransferase	192	2.4
ORTHOMCL668	conserved hypothetical exported protein	182	2.3
ORTHOMCL563	arylsulfatase precursor	514	2.3
ORTHOMCL33		186	2.3
ORTHOMCL15	iron compound abc permease protein	354	2.2
ORTHOMCL201	gfo idh family	495	2.2
ORTHOMCL969		148	2.1
ORTHOMCL32	glucose-1-phosphate thymidylyltransferase	296	2.1
ORTHOMCL42	dtdp-4-dehydrorhamnose -epimerase	196	2.0
ORTHOMCL57	two-component system response regulator	265	2.0
ORTHOMCL50	o-acetylhomoserine -lyase	433	2.0