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](https://doi.org/10.1007/s00248-010-9796-1)

Publication date:
2011

[Link back to DTU Orbit](#)

Citation (APA):
Karlsson, F. H., Ussery, D., Nielsen, J., & Nookaew, I. (2011). A Closer Look at Bacteroides: Phylogenetic Relationship and Genomic Implications of a Life in the Human Gut. *Microbial Ecology*, 61(3), 473-485.
<https://doi.org/10.1007/s00248-010-9796-1>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

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

Journal:	<i>Microbial Ecology</i>
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

SCHOLARONE™
Manuscripts

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

Fredrik H Karlsson¹, David W Ussery², Jens Nielsen¹, Intawat Nookaew^{1§}

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

²Center 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 orthologous 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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

For Peer Review

48

49 Background

50 The human intestine is host to roughly 100 trillion microbial cells, 10 times as many
51 as human cells [21] and carry a gene set 150 times larger than the human genome
52 [28]. The microbiota complements the human set of enzymes with important
53 enzymatic functions such as degradation of polysaccharides and production of
54 vitamins. The microbes have a profound impact on human health and physiology
55 especially alteration of gut ecology has been associated with inflammatory bowels
56 diseases and obesity [20, 25, 28, 40].

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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

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

1
2
3 99 a large collection of hybrid two-component systems for environmental sensing in its
4
5
6 100 genome [43]. In many cases genes for these two regulatory systems are positioned in
7
8 101 close proximity to genes coding for glycoside hydrolases and SusC/D [43].
9

10
11 102 In this study we use bioinformatics and comparative genomics methods on 105
12
13 103 genomes from the Bacteroidetes/Chlorobi group to gain knowledge about the
14
15 104 phylogeny of the member species. Further, by comparative analysis we study the gut
16
17 105 living *Bacteroides* (33) and *Parabacteroides* (4) and compare the genetic content of
18
19 106 these gut living organisms to their relatives in other habitats.
20
21

22 23 107 **Methods**

24
25 108 Publically available genomes from the Bacteroidetes/Chlorobi superphylum were
26
27 109 downloaded from GenBank at National Center for Biotechnology Information. A full
28
29 110 list of genomes included is presented in Supplementary Table 1. This study is based
30
31 111 on 33 completely sequenced genomes and 72 in the assembly stage. The list contains
32
33 112 33 genomes from the genus *Bacteroides*, 9 from *Prevotella*, 8 from *Chlorobium* and 4
34
35 113 from *Parabacteroides* and *Porphyromonas* respectively and 47 genomes from other
36
37 114 Genera in the Bacteroidetes/Chlorobi group.
38
39
40
41

42 43 115 **Genetic components analysis**

44
45 116 The genome sequences were predicted for their content of tRNAs and rRNAs by
46
47 117 tRNAscan-SE [23] and RNAmmer [18] program, respectively. The prediction of
48
49 118 sigma factors, two-component signal transduction systems, membrane proteins and
50
51 119 secreted proteins were done following the standard methods previously published [1-
52
53 120 3, 15-16].
54
55

56 57 121 **16s rRNA analysis**

58
59 122 16S rRNA sequences, which were extracted from the genomes with RNAmmer, were
60
123 used to make a phylogenetic tree. Sequences of length less than 1400 nucleotides

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 version 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.

145

146 **Functional profiles analysis**

147 All proteins were queried against the COG database to functionally annotate proteins

148 [39]. The COG blast database was downloaded from NCBI FTP and psi-BLAST was

149 used to annotate proteins to COGs with an e-value cutoff of 10^{-2} .

150 The KEGG database was downloaded and for each KEGG ontology, bacterial

151 sequences were filtered out and HMM models were generated with HMMER3 [8]. All

152 genes in the 105 Bacteroidetes/Chlorobi genomes were queried against the HMM

153 models. A cutoff of 10^{-30} was used for statistical significance. A heatmap of each

154 pathway and process derived from the database was constructed based on normalized

155 abundance of the enzymes present in each pathway. The heatmap and hierarchical

156 clustering was performed in R.

157

158 **Results and Discussion**

159 **Genetic components**

160 The 105 Bacteroidetes/Chlorobi genomes shown in Table 1 were downloaded using

161 the NCBI project ID, and scanned for their abundance of ribosomal, sigma factor,

162 tRNA, two-component system, trans-membrane helix and signal peptide genes. The

163 number of genes was compared in the three groups *Bacteroides*, *Parabacteroides* and

164 the other Bacteroidetes/Chlorobi species.

165 The number of tRNAs in each genome show that *Bacteroides* and *Parabacteroides*

166 species contain a significantly higher ($p < 0.01$, non-parametric Mann Whitney's U

167 test) number of genes coding for tRNAs in their genomes compared to other

168 Bacteroidetes/Chlorobi species (Supplementary Figure 1). A larger number of tRNAs

169 is an indication of a faster growth rate at optimal conditions [19] but the correlation is

170 weak and there might be other explanations for high copy numbers of tRNAs.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 thetaiotaomicron* 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

1
2
3
4 196 sequences cluster with sequences from other species, *Parabacteroides* sp. D13 and
5
6 197 *Bacteroides* sp 4_3_47FAA, respectively. The average copy number of the 16S rRNA
7
8 198 gene is about 2 (2.3) in all Bacteroidetes/Chlorobi species and there is no significant
9
10 199 difference between *Bacteroides*, *Parabacteroides* and other Bacteroidetes/Chlorobi
11
12 200 species. In the *Bacteroides* genus, the maximum copy number the 16S rRNA gene is 7
13
14 201 in *Bacteroides vulgatus* ATCC 8482. When enumerating bacterial cells based on 16S
15
16 202 rRNA methods, this difference in copy number is important to keep in mind.
17
18 203 Interestingly, *Bacteroides pectinophilus* and *Bacteroides capillosus* cluster together
19
20 204 and are found far from the other *Bacteroides* species. The 16s rRNA sequences of
21
22 205 *Bacteroides capillosus* have 96-98% sequence similarity with *Clostridium*
23
24 206 *orbiscindens* strains and it has recently been suggested that the species should be
25
26 207 reclassified to the novel genus *Pseudoflavonifractor* [4]. Similarly, the 16s rRNA
27
28 208 sequences of *Bacteroides pectinophilus* have a 92% sequence similarity with
29
30 209 *Eubacterium eligens* ATCC 27750 and *Clostridium saccharolyticum* WM1 suggesting
31
32 210 that also this strain is classified in the wrong phylum and should belong to the
33
34 211 Firmicutes. Generally, the resolution in the 16S rRNA tree is limited and it is
35
36 212 impossible to discern the relationship between closely related species.
37
38 213 A more detailed and comprehensive view of the genomic phylogenetic relationship
39
40 214 between the species can be seen in Figure 2 and was achieved by clustering
41
42 215 distribution of protein families defined by the unsupervised algorithm orthoMCL [22].
43
44 216 Clearly, the depth of resolution is higher in the protein family tree compared to the
45
46 217 16S rRNA tree (Figure 1). As opposed to the 16S rRNA tree, here all the *Bacteroides*
47
48 218 genomes cluster together with *Parabacteroides* genomes except for the *Bacteroides*
49
50 219 *pectinophilus* and *Bacteroides capillosus* which are still far from other *Bacteroides*.
51
52
53
54
55
56
57
58
59
60

1
2
3 220 *Parabacteroides* species form a small cluster within the *Bacteroides* cluster showing
4
5 221 high similarity with the other *Bacteroides*. *Bacteroides* sp. 2_1_33B and sp. 2_1_7
6
7 222 cluster tightly with *Parabacteroides* species but neither had a 16s rRNA sequence that
8
9 223 met our quality criteria. The *Parabacteroides* genus was proposed to harbour species
10
11 224 that showed differences in 16s rRNA sequences and different menaquinone
12
13 225 composition compared to *Bacteroides* [31]. But at the whole genome level, our results
14
15 226 indicate that *Parabacteroides* are a part of the *Bacteroides* genus. *Bacteroides* species
16
17 227 clearly seem to have a shared genomic core that we try to define and contrast to other
18
19 228 Bacteroidetes/Chlorobi species.

20
21
22
23
24
25 229 **Pan and core genome comparisons**
26
27 230 A pan and core genome plot was drawn based on the results from the orthoMCL
28
29 231 protein families of *Bacteroides* and *Parabacteroides* genomes as shown in Figure 3.
30
31 232 The pan orthoMCL protein families were defined as being represented in at least one
32
33 233 of the studied genomes whereas the core protein families were present in all genomes.
34
35 234 Genomes are ordered by genus but within genus the order is alphabetical except for
36
37 235 *Bacteroides pectinophilus* and *Bacteroides capillosus* that are placed last. The number
38
39 236 of core protein families for the 31 *Bacteroides* genomes is 1116 and for the
40
41 237 *Bacteroides* and *Parabacteroides* it is 1085 whereas it drops dramatically for
42
43 238 *Bacteroides pectinophilus* and *Bacteroides capillosus* to 424. The number of core
44
45 239 protein families in the *Bacteroides/Parabacteroides* genus is stable and only slowly
46
47 240 decreases when new genomes are added. However the pan protein families are
48
49 241 growing at a much faster rate showing that each genome carries specialized genes not
50
51 242 shared with other *Bacteroides* species. *Bacteroides pectinophilus* and *Bacteroides*
52
53 243 *capillosus* genomes add a considerable number of protein families to the pan showing
54
55 244 that they contain several novel protein families not present in other *Bacteroides* or
56
57
58
59
60

1
2
3 245 *Parabacteroides* strains. *Bacteroides* genomes share a smaller number of core protein
4
5
6 246 families with *Porphyromonas* (694) and *Prevotella* (703) compared to
7
8 247 *Parabacteroides* (1085) even though *Prevotella* seem to have closer related 16S
9
10 248 rRNA sequences.
11
12
13 249 The *Bacteroides* core protein families were further queried for functional domains by
14
15 250 InterPro scan [29]. To evaluate functions that are specific for *Bacteroides*, the number
16
17 251 of genes in each core protein family was compared between the *Bacteroides* and other
18
19 252 Bacteroidetes/Chlorobi. A subset of the protein families is not only a core in
20
21 253 *Bacteroides* but is common to many Bacteroidetes/Chlorobi species. The common
22
23 254 protein families are related to translation, *e.g.* ribosomal proteins and tRNA synthases
24
25 255 necessary for basic machineries for growth (see Table 2 for details). Out of the 20
26
27 256 most specific core protein families in *Bacteroides*, 7 contained a signal peptide and 3
28
29 257 contained a transmembrane domain and 6 protein families were hypothetical proteins.
30
31 258 The core protein families with the highest copy number were two-component systems,
32
33 259 ECF σ -factors and hydrolase enzymes that are necessary for their life in the gut
34
35 260 environment.
36
37 261 Blast atlases [11] provide an overview of chromosome arrangement of conserved
38
39 262 regions (core) as well as variable regions (pan). Blast atlases of the four complete
40
41 263 genomes along with aligned genomes of *Bacteroides* and *Parabacteroides* are shown
42
43 264 in Figure 4. Again, *Bacteroides pectinophilus* and *Bacteroides capillosus* have very
44
45 265 little conserved regions with other *Bacteroides* species. Additionally, *Parabacteroides*
46
47 266 species show a high similarity with *Bacteroides* species and *Parabacteroides*
48
49 267 *distasonis* has few unique genomic regions that are shared with other *Parabacteroides*
50
51 268 species but not with *Bacteroides* species.
52
53
54
55
56
57
58
59
60

1
2
3 269 The variable gene content is not evenly distributed over the chromosome but rather is
4
5 270 located to islands. This is especially evident for *Bacteroides fragilis* and *Bacteroides*
6
7
8 271 *vulgatus* that contain several islands with little homology to other species. Again
9
10 272 *Parabacteroides distasonis* is shown to be genomically similar to other *Bacteroides*
11
12 273 species in general and particularly *Bacteroides* 2_1_33B and 2_1_7. *Bacteroides*
13
14 274 *thethaiotaomicron* is seen as a generalist with a broad repertoire of glycoside
15
16 275 hydrolase paralogs and starch utilization systems C and D paralogs [44]. However, the
17
18 276 blast atlas shows that there are few unique regions in the genome and these are not
19
20 277 gathered in islands but rather spread out over the chromosome.
21
22
23
24

25
26 278 **Functional profiles of Bacteroidetes/Chlorobi**
27 279 OrthoMCL is an unsupervised algorithm for finding all shared protein families among
28
29 280 genomes; however it does not provide any functional information. By annotating
30
31 281 genes to functional categories, *e.g.* metabolic functions, we can discern the
32
33 282 requirements a certain habitat puts up on a genome. In Figure 5 and 6 we map all 105
34
35 283 Bacteroidetes/Chlorobi genomes to the curated cluster of orthologus groups (COG)
36
37 284 [39] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [14], respectively.
38
39 285 Functional annotation relies on inferring gene function by sequence similarity to
40
41 286 genes with known function but is evidently limited to the size of the reference set. The
42
43 287 number of orthoMCL protein families is 26,163 compared to 4,873 for the COG
44
45 288 database; thus the space covered by the unsupervised algorithm is much larger, as
46
47 289 shown in Figure 5A.
48
49
50
51

52
53 290 The COG database contains 4873 orthologus groups made up from 138 458 proteins
54
55 291 from 66 unicellular organisms covering 75% of their predicted proteins [39]. Here we
56
57 292 mapped all 105 Bacteroidetes/Chlorobi genomes to the COG database and explored
58
59 293 the functional space of each organism, meaning that paralogs were not considered
60

(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.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

318 The heatmap gives a detailed view of the metabolic capabilities of each species,
319 which can be related to their natural habitat. *Bacteroides* species mostly group
320 together and it is evident that they are enriched in carbohydrate acting enzymes and
321 also glycan, vitamin and cofactor metabolism. *Prevotella bergensis*, isolated from
322 human skin [6], and *Prevotella copri*, isolated from human faeces [12], group with gut
323 living *Bacteroides* and *Parabacteroides*. *Bacteroides pecinophilus* and *Bacteroides*
324 *capillosus* group together and distinctly from the other *Bacteroides* species as seen in
325 the previous analysis but still these organisms are living in the human gut as two of
326 the 50 most abundant species [28]. This lack of consensus among gut living species
327 likely means that the human gut habitat is not homogeneous but rather contains
328 several niches. This is also consistent with results found by two studies in gnotobiotic
329 mice with *Bacteroides thetaiotaomicron* and one member of the Firmicutes phylum
330 and a methanogenic archae [24, 32]. *Bacteroides thetaiotaomicron* is the primary
331 fermenter of polysaccharides whereas the Firmicute and Archae use simple sugars and
332 fermentative products such as acetate and H₂.
333 However, in general aerobic free-living species in water or soil group together, shown
334 in Figure 6 marked with blue/brown. Unculturable intracellular symbionts *Sulcia*
335 *mulleri* and *Blattabacterium* species group together as these genomes contains very
336 few proteins and thus has low abundance of enzymes in each pathway. The clade
337 marked with yellow contains mainly *Prevotella*, *Porphyromonas* and
338 *Capnocytophaga* species, all living in the human oral cavity or on human skin. In
339 summary, the phylogenetic analysis based on metabolic pathway content can indicate
340 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 orthologous groups such as is done with the orthoMCL algorithm is useful when classifying species and analyzing orthologous 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 genera *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

1
2
3 413 2. Bendtsen JD, Binnewies TT, Hallin PF, Ussery DW (2005) Genome update:
4 414 prediction of membrane proteins in prokaryotic genomes. *Microbiology* 151:
5 415 2119-2121
6
7 416 3. Binnewies TT, Bendtsen JD, Hallin PF, Nielsen N, Wassenaar TM, Pedersen
8 417 MB, Klemm P, Ussery DW (2005) Genome Update: Protein secretion systems
9 418 in 225 bacterial genomes. *Microbiology* 151: 1013-1016
10 419 4. Carlier JP, Bedora-Faure M, K'Ouas G, Alauzet C, Mory F (2010) Proposal to
11 420 unify *Clostridium orbiscindens* Winter et al. 1991 and *Eubacterium plautii*
12 421 (Seguin 1928) Hofstad and Aasjord 1982, with description of *Flavonifractor*
13 422 *plautii* gen. nov., comb. nov., and reassignment of *Bacteroides capillosus* to
14 423 *Pseudoflavonifractor capillosus* gen. nov., comb. nov. *Int J Syst Evol*
15 424 *Microbiol* 60: 585-590
16
17 425 5. Doolittle WF (1999) Phylogenetic classification and the universal tree.
18 426 *Science* 284: 2124-2129
19
20 427 6. Downes J, Sutcliffe IC, Hofstad T, Wade WG (2006) *Prevotella bergensis* sp.
21 428 nov., isolated from human infections. *Int J Syst Evol Microbiol* 56: 609-612
22 429 7. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill
23 430 SR, Nelson KE, Relman DA (2005) Diversity of the human intestinal
24 431 microbial flora. *Science* 308: 1635-1638
25 432 8. Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* 14: 755-763
26 433 9. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy
27 434 and high throughput. *Nucleic Acids Res* 32: 1792-1797
28 435 10. Garrity GM, Lilburn TG, Cole JR, Harrison SH, Euzeby J, Tindall BJ (2007)
29 436 Introduction to the Taxonomic Outline of Bacteria and Archaea (TOBA)
30 437 Release 7.7
31 438 11. Hallin PF, Binnewies TT, Ussery DW (2008) The genome BLASTatlas-a
32 439 GeneWiz extension for visualization of whole-genome homology. *Mol*
33 440 *Biosyst* 4: 363-371
34 441 12. Hayashi H, Shibata K, Sakamoto M, Tomita S, Benno Y (2007) *Prevotella*
35 442 *copri* sp. nov. and *Prevotella stercorea* sp. nov., isolated from human faeces.
36 443 *Int J Syst Evol Microbiol* 57: 941-946
37 444 13. Hong SH, Kim TY, Lee SY (2004) Phylogenetic analysis based on genome-
38 445 scale metabolic pathway reaction content. *Appl Microbiol Biotechnol* 65: 203-
39 446 210
40 447 14. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG
41 448 resource for deciphering the genome. *Nucleic Acids Res* 32: D277-280
42 449 15. Kiil K, Ferchaud JB, David C, Binnewies TT, Wu H, Sicheritz-Ponten T,
43 450 Willenbrock H, Ussery DW (2005) Genome update: distribution of two-
44 451 component transduction systems in 250 bacterial genomes. *Microbiology* 151:
45 452 3447-3452
46 453 16. Kill K, Binnewies TT, Sicheritz-Ponten T, Willenbrock H, Hallin PF,
47 454 Wassenaar TM, Ussery DW (2005) Genome update: sigma factors in 240
48 455 bacterial genomes. *Microbiology* 151: 3147-3150
49 456 17. Kitahara M, Sakamoto M, Ike M, Sakata S, Benno Y (2005) *Bacteroides*
50 457 *plebeius* sp. nov. and *Bacteroides coprocola* sp. nov., isolated from human
51 458 faeces. *Int J Syst Evol Microbiol* 55: 2143-2147
52 459 18. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW
53 460 (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes.
54 461 *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-Vercos 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

- 511 29. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez
512 R (2005) InterProScan: protein domains identifier. *Nucleic Acids Res* 33:
513 W116-120
- 514 30. Rautio M, Eerola E, Vaisanen-Tunkelrott ML, Molitoris D, Lawson P, Collins
515 MD, Jousimies-Somer H (2003) Reclassification of *Bacteroides putredinis*
516 (Weinberg et al., 1937) in a new genus *Alistipes* gen. nov., as *Alistipes*
517 *putredinis* comb. nov., and description of *Alistipes finegoldii* sp. nov., from
518 human sources. *Syst Appl Microbiol* 26: 182-188
- 519 31. Sakamoto M, Benno Y (2006) Reclassification of *Bacteroides distasonis*,
520 *Bacteroides goldsteinii* and *Bacteroides merdae* as *Parabacteroides distasonis*
521 gen. nov., comb. nov., *Parabacteroides goldsteinii* comb. nov. and
522 *Parabacteroides merdae* comb. nov. *Int J Syst Evol Microbiol* 56: 1599-1605
- 523 32. Samuel BS, Gordon JI (2006) A humanized gnotobiotic mouse model of host-
524 archaeal-bacterial mutualism. *Proc Natl Acad Sci U S A* 103: 10011-10016
- 525 33. SHAH HN, COLLINS DM (1990) NOTES: *Prevotella*, a New Genus To
526 Include *Bacteroides melaninogenicus* and Related Species Formerly Classified
527 in the Genus *Bacteroides*. *Int J Syst Bacteriol* 40: 205-208
- 528 34. SHAH HN, COLLINS MD (1988) Proposal for Reclassification of
529 *Bacteroides asaccharolyticus*, *Bacteroides gingivalis*, and *Bacteroides*
530 *endodontalis* in a New Genus, *Porphyromonas*. *Int J Syst Bacteriol* 38: 128-
531 131
- 532 35. SHAH HN, COLLINS MD (1989) Proposal To Restrict the Genus
533 *Bacteroides* (Castellani and Chalmers) to *Bacteroides fragilis* and Closely
534 Related Species. *Int J Syst Bacteriol* 39: 85-87
- 535 36. Stock AM, Robinson VL, Goudreau PN (2000) Two-component signal
536 transduction. *Annu Rev Biochem* 69: 183-215
- 537 37. Suzuki R, Shimodaira H (2006) Pvcust: an R package for assessing the
538 uncertainty in hierarchical clustering. *Bioinformatics* 22: 1540-1542
- 539 38. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular
540 Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol*
541 24: 1596-1599
- 542 39. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV,
543 Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov
544 S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG
545 database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41
- 546 40. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE,
547 Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath
548 AC, Knight R, Gordon JI (2009) A core gut microbiome in obese and lean
549 twins. *Nature* 457: 480-484
- 550 41. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI
551 (2006) An obesity-associated gut microbiome with increased capacity for
552 energy harvest. *Nature* 444: 1027-1031
- 553 42. Xu J, Bjursell MK, Himrod J, Deng S, Carmichael LK, Chiang HC, Hooper
554 LV, Gordon JI (2003) A genomic view of the human-*Bacteroides*
555 *thetaiotaomicron* symbiosis. *Science* 299: 2074-2076
- 556 43. Xu J, Chiang HC, Bjursell MK, Gordon JI (2004) Message from a human gut
557 symbiont: sensitivity is a prerequisite for sharing. *Trends Microbiol* 12: 21-28
- 558 44. Xu J, Mahowald MA, Ley RE, Lozupone CA, Hamady M, Martens EC,
559 Henrissat B, Coutinho PM, Minx P, Latreille P, Cordum H, Van Brunt A, Kim
560 K, Fulton RS, Fulton LA, Clifton SW, Wilson RK, Knight RD, Gordon JI

1
2
3 561 (2007) Evolution of symbiotic bacteria in the distal human intestine. PLoS
4 562 Biol 5: e156
5 563
6 564
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Peer Review

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

B

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_AAWS000000000
Polaribacter irgensii 23-P	2557	a	13451	NZ_AAOG000000000
Robiginitalea biformata HTCC2501	3209	c	13461	CP001712.1
Parabacteroides distasonis ATCC 8503	3850	c	13485	CP000140.1
Psychroflexus torquus ATCC 700755	6751	a	13542	NZ_AAPR000000000
Polaribacter sp. MED152	2611	a	13543	NZ_AANA000000000
Dokdonia donghaensis MED134	2944	a	13544	NZ_AAMZ000000000
Croceibacter atlanticus HTCC2559	2719	a	13570	NZ_AAMP000000000
Leeuwenhoekiella blandensis MED217	3735	a	13573	NZ_AANC000000000
Flavobacteriales bacterium HTCC2170	3478	a	13595	NZ_AAOC000000000
Flavobacteria bacterium BBFL7	2587	a	13604	NZ_AAPD000000000
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_AANL000000000
Chlorobium ferrooxidans DSM 13031	2158	a	16644	NZ_AASE000000000
Bacteroides caccae ATCC 43185	3855	a	18163	NZ_AAVM000000000
Bacteroides capillosus ATCC 29799	4833	a	18173	NZ_AAXG000000000
Bacteroides ovatus ATCC 8483	5536	a	18191	NZ_AAXF000000000
Parabacteroides merdae ATCC 43184	4384	a	18193	NZ_AAXE000000000
Bacteroides uniformis ATCC 8492	4663	a	18195	NZ_AAYH000000000
Algoriphagus sp. PR1	4215	a	18947	NZ_AAXU000000000
Flavobacteria bacterium BAL38	2612	a	18953	NZ_AAXX000000000
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_ABHI000000000
Kordia algicida OT-1	4514	a	19315	NZ_ABIB000000000

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 JCVIHP010	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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Bacteroides sp. 3_2_5	4505	a	32441	NZ_ACIB000000000
Bacteroides sp. 4_3_47FAA	4613	a	32443	NZ_ACDR000000000
Bacteroides sp. 9_1_42FAA	4871	a	32445	NZ_ACAA000000000
Bacteroides sp. D1	4785	a	32447	NZ_ACAB000000000
Bacteroides sp. D2	5264	a	32449	NZ_ACGA000000000
Bacteroides dorei 5_1_36/D4	4431	a	32451	NZ_ACDI000000000
Blattabacterium sp. (Periplaneta americana) str. BPLAN	577	a	32975	CP001429
Prevotella tannerae ATCC 51259	2811	a	33153	NZ_ACIJ000000000
Candidatus Sulcia muelleri SMDSEM	242	c	33829	CP001605.1
Porphyromonas uenonis 60-3	1977	a	34101	NZ_ACLR000000000
Prevotella bergensis DSM 17361	2825	a	34637	NZ_ACKS000000000
Prevotella oris F0302	3316	a	38329	NZ_ACUZ000000000
Prevotella veroralis F0319	3048	a	38331	NZ_ACVA000000000
Bacteroides sp. 2_1_16	4609	a	38347	ACPP000000000
Bacteroides sp. 2_1_22	4748	a	38349	ACPQ000000000
Bacteroides sp. 2_1_33B	3966	a	38351	ACPR000000000
Bacteroides sp. 3_1_33FAA	4666	a	38353	ACPS000000000
Bacteroides sp. D20	3652	a	38355	ACPT000000000
Parabacteroides sp. D13	4494	a	38359	NZ_ACPW000000000
Flavobacteriaceae bacterium 3519-10	2534	c	38559	CP001673.1
Prevotella sp. oral taxon 472 str. F0295	3092	a	38731	ACZS000000000
Prevotella buccalis ATCC 35310	2456	a	40669	ADEG000000000

Protein family	Seq. Description	Seq. Length	Mean number in Bacteroides
ORTHOMCLO	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 thymidyltransferase	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