Ancestry and adaptive radiation of *Bacteroidetes* as assessed by comparative genomics

Raul Munoz[a,b,*], Hanno Teeling[a], Rudolf Amann[a] and Ramon Rosselló-Móra[b,*]

[a] Department of Molecular Ecology, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany.

[b] Marine Microbiology Group, Department of Ecology and Marine Resources, Institut Mediterrani d'Estudis Avançats (CSIC-UIB), E-07190 Esporles, Balearic Islands, Spain.

* Corresponding authors:

Raul Munoz, Marine Microbiology Group, Carrer Miquel Marquès 21, 07190 Esporles, Illes Balears, Spain. e-mail: raul.munoz3@estudiant.uib.cat

Ramon Rosselló-Móra, Marine Microbiology Group, Carrer Miquel Marquès 21, 07190 Esporles, Illes Balears, Spain. e-mail: ramon@imedea.uib-csic.es

Keywords

Bacteroidetes, Na^+-NQR, alternative complex III, caa_3 cytochrome oxidase, gliding, T9SS.

Abbreviations

m.s.i.: median sequence identity.

ABSTRACT

As of this writing, the phylum *Bacteroidetes* comprises more than 1,500 described species with diverse ecological roles. However, there is little understanding of archetypal *Bacteroidetes* traits on a genomic level. We compiled a representative set of 89 *Bacteroidetes* genomes and used pairwise reciprocal best match gene comparisons and gene syntenies to identify common traits that allow to trace *Bacteroidetes'* evolution and adaptive radiation. Highly conserved among all studied *Bacteroidetes* was the type IX secretion system (T9SS). Class-level comparisons furthermore suggested that the ACIII-caa$_3$COX super-complex evolved in the ancestral aerobic bacteroidetal lineage, and was secondarily lost in extant anaerobic *Bacteroidetes*. Another *Bacteroidetes*-specific respiratory chain adaptation is the sodium-pumping Nqr complex I that replaced the ancestral proton-pumping complex I in marine species. The T9SSS plays a role in gliding motility and the acquisition of complex macro-molecular organic compounds, and the ACIII-caa$_3$COX super-complex allows effective control of the electron flux during respiration. This combination likely provided ancestral *Bacteroidetes* with a decisive competitive advantage to effectively scavenge, uptake and degrade complex organic molecules, and therefore has played a pivotal role in the successful adaptive radiation of the phylum.

INTRODUCTION

The Gram-negative, biderm *Bacteroidetes* constitute one of the four largest bacterial phyla with cultured representatives. Compared to the *Proteobacteria*, *Actinobacteria* and *Firmicutes* it is underrepresented in culture collections and sequence databases despite high abundance in various environments ranging from human gut to marine surface waters [16]. Recent studies have corroborated the pivotal role of *Bacteroidetes* in the remineralization of high molecular weight organic matter [e.g. 20, 49, 52]. High abundances and ecological importance notwithstanding, we still lack a defined blueprint that could explain the monophyly of the *Bacteroidetes* [22, 33] and their evolutionary success. Candidate phenotypic traits to identify *Bacteroidetes* are their gliding motility [29], the use of flexirubin family pigments as UV protectants [2], or their capability of complex biopolymers take up via TonB-dependent outer membrane receptors/transporters [50, 52]. Good genomic markers for these phenotypes are the *gld* genes for gliding motility, *flx* for flexirubin biosynthesis and the *susD* gene as predictor of canonical bacteroidetal polysaccharide utilization loci (PUL) [50, 51].

Ongoing isolation and genome sequencing efforts have led to a more even representation of *Bacteroidetes* in public culture collections and sequence databases, even though members of the classes *Bacteroidia* and *Flavobacteriia* disproportionately outnumber those of the classes *Chitinophagia*, *Saprospiria*, *Sphingobacteriia*, and *Cytophagia*. Recent taxonomic rearrangements have also changed the composition of these classes [14, 33], thereby changing the search space for shared proteins that constitute suitable phenotypic markers [13, 29]. Therefore, we conducted a reassessment of *Bacteroidetes* genomes in order to identify traits that define the essence of what makes a bacteroidete a bacteroidete.

In 2007 Gupta and Lorenzini already tried to define the set of genes exclusive to the *Bacteroidetes.* Due to the limited number of available sequenced genomes they based their analysis on only three *Bacteroidia* and one *Flavobacteriia* [13]. Twenty-seven idiosyncratic genes were identified, all of which coded for proteins of unknown functions that contained characteristic

Bacteroidetes-specific signatures. This small number suggests that the analyzed genomes are only distantly related, which substantiates that the *Bacteroidetes* is a large phylum with high adaptive radiation. Here we present a reciprocal best match comparison of 89 carefully selected *Bacteroidetes* genomes that is broader in taxonomic representation and links the adaptive radiation of the *Bacteroidetes* to their energy metabolism.

MATERIALS AND METHODS

Data selection

As of June 2016, more than 800 *Bacteroidetes* genomes were available at NCBI's RefSeq database. We excluded genomes of *Blattabacteriaceae* endosymbionts to avoid a bias due to genome reduction. Likewise we removed redundant genomes from popular genera, such as *Bacteroides*, *Prevotella*, *Porphyromonas*, *Flavobacterium* or *Hymenobacter,* while retaining all type strains. This resulted in a phylogenetically more balanced list of 478 complete genomes (Supplementary table 1). For genomes of species from the same genus, we performed average nucleotide identity (ANI) calculations using ANIm as implemented in pyany v0.1.3.9 [38] to confirm species affiliations (Supplementary figure 1). The final dataset comprised 89 high-quality *Bacteroidetes* genomes, each consisting of less than five contigs. We chose 48 *Flavobacteriia*, 16 *Bacteroidia*, 15 *Cytophagia*, 5 *Chitinophagia* and 5 *Sphingobacteriia* in order to maintain phylogenetic representativeness. Two members of the *Saprospiria* were included within the *Chitinophagia* as classified by Munoz *et al.* [33]. The sphingobacterial branch of the *Mucilaginibacter* spp. was not represented because the most complete genome still consisted of seven contigs. Five non-bacteroidetal genomes of the *Fibrobacter-Chlorobi-Bacteroidetes* (FCB) lineage complemented the database (Supplementary figure 2, Supplementary table 2). Genome size (megabases) and G+C %mol content were used as metrics to statistically compare the full dataset of 478 genomes to the reduced subset of 89 genomes using R v3.6.0 [39] (Table 1, Supplementary figure 3).

Comparative genomics

Using scripts from the enveomics suite [43], BLAST+ v2.2.28 [1] reciprocal best matches were calculated for the predicted protein sequences encoded by all 89 *Bacteroidetes* genomes. As threshold for orthology, we used a minimum of 50% amino acid identity within 50% of the sequence length [55], based on systematic evaluations that showed that core genomes remained stable using 40% to 60% identity-coverage combinations (data not shown). Thousands of groups of orthologous genes were obtained and corresponding amino acid sequences were compared at the phylum and class levels as classified in Munoz *et al.* [33]. The resulting groups of orthologous sequences were divided in three sets. (1) Sequences encoded in all genomes of a taxon (core genome). (2) Sequences exclusively encoded in genomes of a taxon and not in genomes of other taxa (exclusive sequences). (3) Sequences encoded in the majority of genomes of a taxon, but also present in a smaller number of other genomes of other taxa (henceforth referred to as *prevalent* sequences). The phylogenetic coverage of prevalent sequences in this study ranged from 80% in classes *Chitinophagia* and *Sphingobacteriia* (encoded in 4 out of 5 genomes and 1 allowed outlier ortholog), to 96% in the *Bacteroidetes* phylum (encoded in 86 out of 89 genomes and/or 3 allowed outlier genomes).

Sequence identity and annotation

Exclusive and prevalent sequences were aligned with Clustal Omega v1.2.2 [12] using default parameters. From the resulting identity matrices, we calculated the median identity of each sequence with its orthologs and extracted the maximal and minimal identities within every group of orthologous sequences. Median sequence identities were compiled in a table that was transformed into a heatmap using plotly v2.6.0 [37]. Genome sequence annotations were updated by sequence similarity searches against the UniProt [7] and KEGG [18] databases. Sequences in the core genome of *Bacteroidetes* were searched against the NCBI database using BLAST with default

parameters in search of homologies outside the phylum by excluding the *Bacteroidetes* (organism filter taxid: 976).

Sequence synteny and synonyms

For each taxon we selected reference genomes (Table 2) from which we extracted a multiheaded FASTA file for each group of orthologs containing the conserved protein sequence plus eight adjacent sequences (four genes up- and downstream). These files were searched for recurrent gene arrangements against all 89 genomes using MultiGeneBlast v1.1.13 [30] for visual identification of syntenies. The sequence homology cut-off was set to 20% identity and 30% coverage allowing the detection of homologies below the reciprocal best match threshold. We set the maximum distance between homologs to 10 genes. When syntenic arrangements extended past the nine query genes, extended multiheaded fasta files were generated and queried against the database for confirmation. Ideally, every gene in a syntenic arrangement could be matched to a single group of orthologous sequences indicating that the groups of orthologs represented adjacent genes. However, we also depicted genes represented by multiple groups of orthologous sequences. These groups of orthologs representing the same gene and syntenic position are synonymous. Furthermore, their annotations also coincided [6].

Phylogenetic reconstructions

Orthologous protein sequences were extracted from genomes using enveomic tools [43] and subsequently imported into ARB [26]. Sequences were aligned using Muscle v3.8.31 [8] with 16 maximum iterations. Neighbor-joining phylogenetic analyses included Kimura correction and 1,000 bootstrap iterations.

RESULTS

145 Overall the selected subset of 89 genomes of *Bacteroidetes* was representative of the curated
146 dataset of 478 genomes in terms of median genome size (3.96 Mbp ± 1.4) and G+C mol% content
147 (39.4% ± 6.8). Minimum values were higher due to exclusion of *Blattabacteriaceae* (Table 1). The
148 mean genome size in the *Chitinophagia* increased by 2 Mbp, but its G+C % mol% remained
149 representative (Supplementary figure 3). The reason is that the selected *Chitinophagia-Saprospiria*
150 genomes with the least contigs were also the largest in this taxon (Supplementary table 1). The
151 balanced representation of species along the full phylogenetic tree of the phylum (Supplementary
152 figure 2), and the coherent metrics explained above, indicated that the reduced dataset was as
153 diverse and representative as the repositories allowed.
154
155 *Screening of conserved genes*
156 The reciprocal best match analysis classified 325,548 translated sequences into 31,265 groups of
157 orthologs (OGs) (Supplementary table 3), of which 5% recruited paralogous sequences. In OGs
158 with no paralogs each sequence was recruited from a different genome. In OGs with paralogs at
159 least two homologous sequences originate from the same genome. Of the 1,589 OGs that contained
160 paralogs, 56% recruited only 1 paralog in 1 genome. Frequencies of OGs that recruited increasing
161 numbers of paralogs decreased steeply (Supplementary figure 4). The core of the 94 *Fibrobacteres-*
162 *Chlorobi-Bacteroidetes* (FCB) genomes in the analysis consisted of 65 predicted housekeeping
163 proteins (Table 2, supplementary table 4) and was similar to the core genome of the prokaryotes in
164 terms of size and composition [21]. The *Bacteroidetes'* core genome comprised 155 predicted
165 proteins with a predominance of ribosomal proteins (Figure 1, Table 2, Supplementary table 5).
166 BLAST searches retrieved homologs of all the 155 core genome sequences in other phyla with
167 similarities above 30%. Whithin our database, only 31 sequences were exclusively encoded in the
168 *Bacteroidetes* with a predominance of aminoacyl-tRNA synthesis enzymes (Figure 1, Table 2,
169 Supplementary table 6). Exclusive plus prevalent in *Bacteroidetes* were 87 sequences (Figure 1,
170 Table 2, Supplementary table 6), 49 of which did not belong to the core genome and were either

involved in aminoacyl-tRNA biosynthesis or of unknown functions (Figure 1, Table 2). Prevalent sequences contained fewer predicted ribosomal proteins than the core genome, since most ribosomal proteins were also present in the out-group genomes. For the same reason they did not recruit predicted proteins involved in the metabolism of terpenoids and polyketides, lipid metabolism and energy transduction (Figure 1). Among the 87 highly conserved sequences in *Bacteroidetes* a total of 21 were uncharacterized predicted proteins or proteins of unknown functions. Independent BLAST searches indicated that the majority might belong to housekeeping functions, but six remained difficult to classify (Table 3). Two unknown protein pairs that were consistently encoded next to each other are of particular interest (corresponding to Pfam entries PF02591/PF01784 and PF01327/PF03652).

Only two of the original 27 proteins identified by Gupta and Lorenzini 2007 [13] were contained in our list of 31 predicted exclusive proteins, 13 did not find a reciprocal match in other genomes, and many others were not common in the phylum (Supplementary table 7). We therefore examined prevalent sequences in classes of *Bacteroidetes* (Table 2, Supplementary figure 5). When the class *Bacteroidia* was excluded, the remaining *Bacteroidetes* shared 19 exclusive sequences and 49 prevalent sequences (Table 2). These proteins that were lost in *Bacteroidia* predominantly belonged to the carbohydrate metabolism, mostly to the tricarboxylic acid (TCA) cycle (Supplementary table 8). The whole set of prevalent sequences by taxonomic class was of 382 sequences which coded for many non-house-keeping functionalities (Supplementary table 8).

Distribution and similarity of prevalent sequences

The median sequence identity (m.s.i.) within prevalent groups of orthologs ranged from 45% to 81%. As much as 61% percent of 382 orthologous groups had within identity ranges of 50-60%, 26% of them of 60-70%, and the remaining 13% were evenly distributed above 70% or just below 50%. High internal m.s.i. did not correlate with taxonomic distribution, metabolic pathways or sequence length. However, representation of the presence/absence pattern of orthologs and their

197 m.s.i. in a heatmap revealed some underlying trends (Figure 2) that corresponded with evolutive
198 traits of the *Bacteroidetes*. Conserved sequences of the *Flavobacteriia* were consistently less similar
199 in genomes of the *Chryseobacterium* branch [14]. The genomes of *Fluviicola taffensis* DSM 16823^T
200 (family *Crocinitomicaceae*) and *Owenweeksia hongkongensis* DSM 17368^T (family
201 *Cryomorphaceae*) lacked many of the conserved sequences of the *Flavobacteriia*. Likewise, the
202 genomes of deep-branching *Bacteroidia*, such as *Alistipes finegoldii DSM* 17242^T,
203 *Draconibacterium orientale* FH5^T, and *Odoribacter splanchnicus* DSM 20712*,* frequently encoded
204 some of the 28 proteins missing from most other *Bacteroidia*. Finally, conserved sequences of the
205 *Chitinophagia* and *Sphingobacteriia* held low m.s.i., indicating phylogenetic remoteness.
206
207 *Presence of hallmark genes described in the literature*
208 TonB-dependent uptake systems are widely distributed among Gram-negative bacteria and have
209 been shown to transport a variety of large substrates including iron-siderophores, nickel, vitamin
210 B_{12} and oligosaccharides [34, 45]. *Bacteroidetes* have evolved a variant that features a SusD-like
211 substrate-binding protein. Genes for SusD are usually co-located with genes for SusC in a
212 characteristic tandem. These tandems are often part of PULs, where SusD acts as initial glycan-
213 binding protein that interacts with the SusC TonB-dependent pore protein for uptake across the
214 outer membrane. However, there are also SusD homologs that might have alternate functions, e.g.
215 in iron acquisition [27].
216 SusC/D proteins are homologous to RagA/B proteins [15] and therefore often annotated as such.
217 We identified 452 orthologous groups containing annotated SusC/RagA sequences, and 160
218 containing SusD/RagB sequences. SusC/RagA homologs were more abundant than SusD/RagB
219 because not all bacteroidetal TonB-dependent transports feature SusD, which is specific for
220 oligosaccharides [34]. The classes *Cytophagia*, *Chitinophagia* and *Sphingobacteriia* showed the
221 highest SusD absolute abundances, but SusD distribution did neither reveal a phylogenetic nor
222 environmental pattern. SusC/SusD ratios were high in the genera *Bacteroides* and *Prevotella* in

comparison to the phylum's general trend, 13 genomes did not contain SusD sequences and five no Sus sequences at all (Supplementary figures 7 and 8).

In order to identify flexirubin synthesis genes, we used the flx genes from *Flavobacterium johnsoniae* UW101 [46] as queries for MultiGeneBlast searches in other genomes. Only 22 genomes of the entire dataset contained at least half of the *flx* cluster (Supplementary figure 9), and no homologous sequences were found in genomes of the *Saprospiria* and *Sphingobacteriia* classes. When homologous gene clusters could be recognized outside the *Flavobacteriia*, these were shorter, mainly due to the absence of predicted hypothetical proteins.

Gliding motility sequences occurred almost ubiquitously in the *Bacteroidetes*. The *gldC* and *gldL* gliding motility genes were conserved in most *Flavobacteriia* (*gldC*), *Sphingobacteriia* (*gldC*, *gldL*) and *Cytophagia* (*gldL*) as synonymous groups of orthologs constrained to the class rank (Supplementary table 8). The *gldC* gene was not conserved in a recognizable gene cluster across *Bacteroidetes* and was located adjacent to *gldB* only in 43 *Flavobacteriia*. In contrast, *gldL* belonged to the cluster *gldKLMN* or *gldKLMO* recognizable in 80 genomes, 90% of the represented *Bacteroidetes* (Supplementary table 9, Supplementary figure 6). Only *Flavobacteriia* encoded the cluster *gldKLMO.* Whereas the best reciprocal matches did not classify the *gldK*, *gldM* and *gldN*/*O* genes as orthologs, we were able to identify their orthology by their syntenic position. Our screening methodology did not recruit genes of other subunits (e.g. B, D, J, etc) either by orthology or synteny.

Hallmark genes of the respiratory chain

Besides in *gldKLMN/O* genes common to most *Bacteroidetes*, synonymy was also frequent among genes of the respiratory chain. Quinol:cytochrome c oxidoreductases appeared among the conserved predicted proteins of all classes except for anaerobic *Bacteroidia*. These oxidoreductases belonged to a gene cluster that codes for an alternative complex III (ACIII) with six subunits: ActABCDEF [40]. Many of the *act* gene clusters showed downstream synteny with genes of predicted

cytochrome oxidase subunits (Figure 3). These genes code for an oxygen-reducing caa_3-type cytochrome oxidase (caa_3COX) that together with ACIII forms a respiratory super-complex [47, 48]. BLAST searches found sequences of the ACIII gene cluster also in genomes of the initially discarded *Blattabacteriaceae*. The sequences of all subunits, except ActC, segregated in at least four synonymous groups of orthologs under our reciprocal best match parameters. The ACIII-caa_3COXgene cluster was encoded in 83% of the represented *Bacteroidetes,* all the aerobes.

Genes *sdhA* and *sdhB* of the respiratory complex II belonged to the core genome of the phylum. Both sequences were always encoded in a gene cluster together with *sdhC*. The *sdhCAB* operon codes for the type B succinate:quinol reductase (SQR) [5, 25]. Since complex II SQR genes were part of the *Bacteroidetes* core genome, and predicted proteins of the supercomplex ACIII-caa_3COX were conserved, we investigated the distribution of other respiratory chain complexes. Sequences of the NADH-quinone oxidoreductase (Nuo proteins of complex I) grouped in groups of orthologs with no phylogenetic pattern of occurrence. Furthermore, we found that genomes that lacked genes for Nuo proteins featured genes for Nqr proteins of a sodium pumping NADH:quinone oxidoreductase (Na^+-NQR) (Figure 3, Supplementary table 9). Only *Haliscomenobacter hydrossis* DSM 1100^T encoded full copies of both complexes (Figure 3, Supplementary table 9). Like complex I, Na^+-NQR accepts electrons from NADH and transfers them to a quinone, but pumps sodium ions instead of protons [3]. Presence of Nqr sequences correlated with strains' isolation from NaCl-rich environments (seawater, sea fauna and human gut microbiota or pathogens). Subunits of Nuo and Nqr were both encoded in recognizable gene clusters (Figure 3). Nuo proteins were organized in the cluster *nuoABCDEFGHIJKLMN*, with a lack of subunits C, E, F and G in *Bacteroidia*. Nqr subunits were encoded in the cluster *nqrABCDEF*, with the lack of subunit F in *Cytophagia* (Figure 3, Supplementary table 9).

Sequences of the F-type ATP synthase were prevalent in the phylum, but barely had orthologous hits in *Bacteroidia*. A detailed inspection with a lower similarity threshold revealed that subunits A, B, C, δ, α, and γ remained encoded in the same gene cluster across most *Bacteroidetes*' genomes

including *Bacteroidia* members (Supplementary table 10). However, no F-type ATPase subunits were present in the *Bacteroidia* members *Porphyromonas asaccharolytica*, *P. gingivalis* and *Alistipes finegoldii*. These species feature a V-type ATPase instead. Eight *Bacteroidia* genomes encoded both types of ATPases (Supplementary table 10) corroborating previous findings [33].

Conserved sequences of the *Bacteroidia* comprised a predicted cytochrome D ubiquinol oxidase subunit II or cytochrome bd (Supplementary table 8). Its gene was always preceded by a gene coding for a DUF4492-containing protein (DUF: domain of unknown function) and then followed by the cytochrome D ubiquinol oxidase subunit I gene (Supplementary table 9). The cytochrome D ubiquinol oxidases I and II corresponded to the cytochrome bd subunits CydAB involved in aerobic respiration of *Bacteroides* species [4].

True orthology of hallmark genes as assessed by monophyletic circumscription of known taxa

We conducted a multi-locus sequence analysis (MLSA) of common *Bacteroidetes* protein complexes in order to check their orthology via agreement with other phylogenies [14, 33] (Figure 4). A phylogenetic reconstruction based on concatenated SdhAB sequences of respiratory complex II could not resolve the six *Bacteroidetes* classes and suggested a horizontal gene transfer (HGT) event between the common ancestor of the *Bacteroidia* and the anaerobic green sulfur bacterium *Chlorobium limicola*. In contrast, a reconstruction based on concatenated ActBCD sequences of the ACIII resolved the commonly accepted *Bacteroidetes* phylogeny although the only representative of the *Bacteroidia* was placed closer to *Saprospiria* rather than *Flavobacteriia*. Lastly, a phylogeny inferred from the GldKLMN/O sequences of the gliding machinery also agreed with the known *Bacteroidetes* phylogeny, but placed *F. taffensis* outside the *Flavobacteriia* in 65% of the total iterations. The genetic drift of the *Chryseobacterim* branch measured by m.s.i. among the *Flavobacteriia* could be related to their dichotomic branching in 100% of the Gld topologies and 67% of the Act topologies.

DISCUSSION

The low number of sequences exclusive to *Bacteroidetes* and their involvement mainly in housekeeping functions precludes delineation of a genetic blueprint of the phylum on this basis alone. Future functional characterization of 21 yet unknown predicted proteins could provide further insights into the *Bacteroidetes* common biology and ancestry. For now, no common phenotype can be ascribed to all *Bacteroidetes*. In addition, the *Bacteroidetes* phylum status has recently been questioned in a proposal to base the assignment of taxonomic ranks consistently on comparable evolutionary distances in the prokaryotic tree of life [36], which could lead to a unification of the FCB group into a single phylum [36]. The 94 FCB genomes analyzed in this study shared a small core genome (65 proteins) resembling a core genome of organisms from different domains [11, 21]. Based on comparable evolutionary distances, the *Actinobacteria* represent a phylum-level taxon [36] with a core genome of 123 genes [56] - a size similar to the *Bacteroidetes* core genome of 155 genes. While a formal status of the phylum category in the ICNP is pending to be implemented by the ICSP [35], the classification of the *Bacteroidetes* as an independent phylum seems justified based in our analyses in terms of phylogenetic coherence, independently of how distant the emergence of the *Bacteroidetes* branch is in the prokaryotic tree of life, and thus deserves a stable nomenclatural status.

Core gene sets of distant genomes are usually small and dominated by essential housekeeping functions [6, 21]. More relaxed criteria can in addition recruit genes that are common yet not ubiquitous, and with an as broad phylogenetic distribution as possible [6]. A problem in comparing distant genomes are groups of distant but still orthologous genes. A high identity threshold for orthology can split such a group of distant orthologs into multiple synonymous groups (false negatives), while a low threshold can pick up spurious matches (false positives). In this study, we applied a high identity threshold, and to detect false negative groupings with taxonomic coherence we also searched for conserved sequences and their syntenies in lower taxonomic ranks. This way

we recovered sequences of the respiratory chain and gliding machinery, which are orthologous and widely distributed in the phylum.

Complex II SQR is the only respiratory complex encoded in all analyzed *Bacteroidetes*. A phylogenetic reconstruction based on SQR SdhAB was not able to resolve the classes *Cytophagia*, *Chitinophagia,* and *Sphingobacteriia*. Substitution rates were low and of insufficient resolution considering the lengths of the protein sequences. This also caused many branches to have low bootstrap support. Overall, the poor phylogenetic signal of these sequences is not indicative fora true orthologous relationship. In this context, the affiliation of complex II from *C. limicola* with the *Bacteroidia'*s could be an artifact, even though a HGT between both taxa would explain the similarity of their Sdh, Rnf, and Nqr protein sequences as xenologues.

Phylogenetic reconstructions based on the conserved GldKLMN/O and ActCDE sequences reproduced the currently accepted major *Bacteroidetes* taxa regardless of branching order. With high substitution rates and acceptable bootstrap values, their phylogenies indicate that Gld and Act sequences can be trusted as orthologs. The Gld-based topology was less stable than the Act based topology. Yet, its only branch with a bootstrap below 50% was the *Bacteroidia*'s, indicating a higher sequence variability in this class compared to others in the phylum. The *Bacteroidia* in the Act-based tree were represented by the facultative aerobe *Draconibacterium orientale* FH5^{T} that affiliates with the *Chitinophagia* and *Saprospiria* instead of the *Flavobacteriia*. However, for the purpose of this study, phylogenetic branching order is irrelevant. Still, the prevalence of the Gld and Act protein complexes in the *Bacteroidetes* should encourage their analysis in future studies of the bacteroidetal classification. In this regard, our MLSA topologies and m.s.i. analyses support possible reclassifications in the class *Flavobacteriia*. The *Chryseobacterium* branch could constitute a new family of the *Flavobacteriales*, presumably the 'Riemerellaceae', and the family *Crocinitomicaceae* (*F. taffensis*) could become a different order of the *Flavobacteriia*, presumably the 'Crocinitomicales'. The family designation of the *Cryomorphaceae* (*O. hongkongensis*) would be debatable.

Gliding motility has for long been thought to be a hallmark of the *Bacteroidetes*. Gliding genes are widespread throughout the *Bacteroidetes*, but the phenotype is not always expressed [29]. The exact mechanism of gliding is still unknown, but two sets of proteins seem to be essential: the membrane Gld subunits B, D, H and J that might be effectors of movement, and a type IX secretion system (T9SS) formed by GldKLMN plus SprA, SprE, and SprT [29]. Gliding in *F. johnsoniae* also requires an ABC-type transporter formed by Gld subunits A, F, and G, but this transporter is absent from other gliding bacteria [29]. GldC is not essential to gliding, and known not to be produced by all gliding cells [17]. The T9SS (also referred to as PorSS or PerioGate) is the latest discovered secretion system and has been only found in *Bacteroidetes* so far [44]. The genes of the T9SS are called *porKLMN*, *sov*, *porW* and *porT*, and are synonymous to *gldKLMN*, *sprA*, *sprE,* and *sprT*. In gliding motility, the T9SS secretes the proteins SprB and RemA that are believed to be adhesins, but T9SS also secretes hydrolytic enzymes like chitinase and cellulase [24, 29]. Thus, the T9SS is also involved in the acquisition of external complex carbon sources. We found that the cluster *gldKLMN*/*O* (*gldO* is thought to be a paralog of gldN) is encoded in 90% of the analyzed genomes. Except for *Niabella soli*, the genomes that lacked the T9SS belonged to pathogen or symbiont species that likely feed on small low molecular weight organic molecules rather than complex macromolecules.

The alternative complex III (ACIII) was first described in *Rhodothermus marinus*, and although it is found in many bacteria, it is prevalent in *Bacteroidetes* [28]. Association of ACIII with a caa_3-type cytochrome oxidase in a super-complex has so far only been described in *F. johnsoniae* [48]. We found that a corresponding super-complex operon is conserved in all aerobic *Bacteroidetes*, thus not constrained to the genus *Flavobacterium*. The super-complex accepts electrons from the quinone pool and funnels them to oxygen as terminal acceptor while translocating protons across the cytoplasmic membrane and thereby contributing to the proton-motive force. The rate of electron transport of the canonical respiratory chain is limited by the capacity of the electron shuttling soluble cytochrome c to connect the terminal cytochrome oxidase with other redox complexes. The

ACIII-caa$_3$COX super-complex solves this problem by controlling the flux of electrons that are transmitted directly from complex III to the c-type heme electron carrier fused to the caa$_3$COX [48]. The high efficiency of the ACIII-caa$_3$COX might be advantageous for competing with other bacteria for the most efficient utilization of intermittently available energy-rich complex organic matter.

The gateways into the respiratory chain are complexes I and II. While a complex II type-B SQR is present in all *Bacteroidetes*, two types of complex I divide the phylum. Strains isolated from freshwater environments featured the canonical NADH:quinone oxidoreductase with 14 subunits (NuoABCDEFGHIKJLMN) coded in a single gene cluster (*nuo*). The cluster was incomplete in some genomes, specifically in *Cytophagia* and *Bacteroidia.* The *nuo* cluster showed random gene deletions in the *Cytophagia*, while the cluster of the *Bacteroidia* resembled the ancestral complex [32] as it always lacked *nuoEFG* (the so-called N-module with dehydrogenase activity) and *nuoC* (part of the Q-module with electron transfer activity). The mechanism of the ancestral complex I is unknown, but it probably interacts with various electron donors or acceptor proteins [32]. In *Cytophagia* and *Sphingobacteriia* genomes, incomplete *nuo* clusters appeared to be paralogous to full *nuo* clusters within the same genome or residual upon the acquisition of the Nqr complex.

An alternative complex I, the Nqr complex (Na^+-NQR), a sodium-pumping NADH:quinone oxidoreductase, was present in strains isolated from seawater or animal and human gut microbiota. Nqr consists of a six-protein membrane complex that was described in marine bacteria [54]. It translocates sodium ions across the inner membrane thereby linking the respiratory chain to osmotic regulation [42]. The Nqr is associated with the respiration of pathogens like *Vibrio*, *Klebsiella* or *Haemophilus* spp. [3], lineages that reportedly acquired the *nqr* genes horizontally during their adaptation to sodium-rich marine, alkaline or intracellular habitats [42]. Based on phylogenomic analysis and comparison of gene cluster layout, it has been proposed that the Na^+-NQR originated in the common ancestor of *Bacteroidetes* and *Chlorobi* via duplication and subsequent neo-functionalization of the *rnf* operon (NADH:ferredoxin dehydrogenase) [42]. The copied *rnf* operon

would have lost the RnfB protein involved in electron uptake from the reduced ferredoxin and later recruited an AMOr subunit (aromatic monooxygenase) to become NqrF, the electron uptake subunit of the Na^+-NQR [42]. The intermediate complex with no NqrF subunit is the ancestral Na^+-NQR [42]. Our results, however, suggest that the Nqr complex evolved solely in the *Bacteroidetes*. Rnf proteins prevail in the *Bacteroidia* with two outliers in the *Flavobacteriia* and homologous sequences in *Chlorobi*. Reyes-Prieto *et al*. represented the *Bacteroidetes* with *Bacteroidia* species in their reconstruction of the Nqr dispersion [42]; hence, the interpreted common origin of the Nqr of *Bacteroidetes-Chlorobi* could be the consequence of a HGT between the *Bacteroidia* and *Chlorobi*, as is suggested by the phylogeny of the SQR complex.

The *nqr* cluster was incomplete in the *Cytophagia* lacking the *nqrF* subunit, thus, reproducing the ancestral Na^+-NQR. This suggests that the original *rnf* operon was transformed in the ancestor of the *Bacteroidetes*, and the acquisition of the NqrF subunit happened after the divergence of the *Cytophagia*. NADH:quinone oxidoreductase (Nuo complex for short) must have been the original complex I, whereas the ancestral Nqr was an adaptation to salinity that remains in genomes of the genus *Flexibacter* and the families *Catalimonadaceae* and *Cyclobacteriaceae*. The complete Nqr complex radiated to lineages adapting to saline stress. Therefore, the radiation of the *nqr* genes was not phylogenetic, but environmental, and Nqr became abundant in predominantly marine or gut taxa like the *Saprospiria*, *Bacteroidia*, *Cryomorphaceae* and a substantial part of the *Flavobacterieaceae*. Since full copies of the Nqr complex coexist with partial copies of the Nuo complex in most of the analyzed genomes, but never vice versa, the acquisition of the Nqr may render the Nuo complex obsolete. Only *H. hydrossis*, isolated from activated sludge, encodes full Nuo and Nqr complexes possibly representing a recent Nqr acquisition. The order in which *nuo* genes are deleted seems random, but for some as yet unknown reason the *Bacteroidia* kept an ancestral-like model of the Nuo complex.

The respiratory chain of the *Bacteroidia* has been extensively described for the anaerobe *Porphyromonas gingivalis* [31]. Despite its anaerobic lifestyle, it can use oxygen as terminal

electron acceptor using a cytochrome bd (subunits CydAB) with high oxygen affinity. Such anaerobes, or nanaerobes (that can benefit from nanomolar concentrations of O_2 but cannot grow at higher concentrations) [4], use this mechanism to scavenge harmful oxygen thereby facilitating colonization. In *Bacteroidia*, the *cydAB* constituted a conserved gene pair together with a conserved DUF4492-containing protein. Some *Flavobacteriia* also encoded the gene pair *cydAB*. Explaining the role of the DUF4492-containing protein (if expressed together with *cydAB*) and the role of the cytochrome bd in *Flavobacteriia* are two future challenges.

Expected phylum-level genomic markers *susD* and *flx* genes were not recruited in any of the sets of conserved sequences. The *flx* genes were not common enough, and *susD* grouped in many different groups of orthologs, none of which reproduced a phylogenetic pattern. PULs in *Bacteroidetes* frequently feature *susCD*-like genes tandems, although some are functional with *susC* only [50]. Therefore SusD sequences can be used as a first rough approximation of the minimum number of PULs in a bacteroidetal genome [50]. Only five of the *Bacteroidetes* genomes in our dataset did not contain any annotated *susD*-like sequences and only 5% of the SusCD OGs contained paralogs. Despite we found few copies of *susD* in some genomes, we do not possess information about their expression. However, PULs are not only frequent and diverse, but also believed to be subject to frequent HGT among *Bacteroidetes* as has been shown for a porphyran-targeting PUL from marine *Flavobacteriia* to anaerobic gut *Bacteroides* [52]. SusC/D sequences furthermore carry a strong substrate-specific signal that obscures their phylogeny [19]. Such sequence diversity combined with a weak phylogenetic signal impedes the utilization of SusCD-like sequences for the explanation of the adaptive radiation of the phylum.

CONCLUSIONS

By means of conservation, phylogenetic signal and uniqueness to the phylum, T9SS *gldKLMN* genes represent suitable genomic markers for *Bacteroidetes* although they are lacking in some pathogenic or symbiont species. T9SS is the anchor of the known gliding machinery, but while

456 other *gld* genes seem accessory and are furthermore distributed in different loci hiding a complex
457 phylogenetic pattern if any, T9SS genes remain well conserved regardless of the gliding capacity of
458 the cell. Since T9SS also translocates enzymes, such as chitinase and cellulase, in a two-step
459 process across both membranes, it connects the two most notable phenotypes of the phylum, gliding
460 and degradation of complex organic matter. This suggests that the T9SS was part of the ancestral
461 bacteroidete and might, to some extent, be responsible for their biological success.
462 *Bacteroidetes* do not only excel in decomposing complex organic matter, but their phylogenetic
463 position in the prokaryotic tree of life suggests they could have been Gram-negative pioneers in this
464 regard. Contrary to Gram-positive decomposers, the biderm *Bacteroidetes* feature a periplasmic
465 space that provides a protected area for degradation, e.g. of oligosaccharides, without diffusive loss
466 of both enzymes and degradation products. Recent studies on the *Bacteroidia* link their success as
467 members of the human gut microbiota to their PULs [9, 57] that are varied in carbohydrate-active
468 enzymes (CAZymes). Functionally, their SusCDs capture and translocate oligosaccharides that are
469 further decomposed in the periplasmic space, keeping them away from competitors [23, 41].
470 Unfortunately, we could not identify an environmental pattern explaining the influence of the
471 SusCD lateral transfer in the adaptive radiation of the phylum, or prove they were likely to exist in
472 the ancestral bacteroidete. Yet, they seem relevant to the essence of what constitutes a bacteroidete.
473 The *Bacteroidetes* also conserved the energy-efficient ACIII-caa$_3$-COX respiratory super-complex
474 improving their fitness to compete for high molecular weight carbon compounds. The substitution
475 of the ACIII-caa$_3$COX by a cytochrome bd is recorded only once in their genomes, at the origin of
476 the *Bacteroidia*. The *Bacteroidia* broke the T9SS/ACIII-caa$_3$COX association and yet succeeded,
477 which hampers a circumscription of the *Bacteroidetes* by common traits. Still, many *Bacteroidia*
478 conserve the T9SS and only pathogens or symbionts have disposed of it. Therefore, we believe the
479 common ancestor of the phylum was a free-living aerobic decomposer, whose bioenergetic
480 efficiency allowed a broad adaptive radiation that ultimately caused some lineages to dispense of
481 the ancestral genes coding T9SS and/or ACIII-caa$_3$-COX.

We would like to end on thoughts with respect to the ancestor of the *Bacteroidetes*. Did it live rather in a marine or freshwater environment? All species that have yet been isolated from freshwater encode the canonical complex I (Nuo), whereas species isolated from salt-rich environments encode the Nqr complex and often some remnants of the *nuo* gene cluster. Presence of the ancestral Nqr and incomplete Nuo complexes in the *Cytophagia* suggests that the initially present Nuo complex gradually degraded after the Nqr cluster was acquired. This would rather support that the *Bacteroidetes* radiated from a Gram-negative, biderm freshwater ancestor from which later successors adapted to the marine environment based on novel mechanisms.

In the present study, we rather explained intra-phylum genomic and phenotypic diversity despite the absence of a common phenotype, which remains a hallmark of *Bacteroidetes* notwithstanding their evolutionary origin. Hence, carbon source limitations and salinity were pivotal to the origin of the phylum and responsible for its adaptive radiation. Future studies to support this hypothesis would need to explain, if T9SS could have been paired with an oligosaccharide uptake mechanism and confirm that *nqr* genes are part of the marine mobilome of the *Bacteroidetes*. From there it could have been appropriated by known Nqr-encoding marine proteobacteria [10, 53], while the protein complex originated in the *Bacteroidetes*.

Acknowledgments

This study was funded by the Max Planck Society, the Spanish Ministry of Science, Innovation and Universities projects CLG2015_66686-C3-1-P and PGC2018-096956-B-C41, both also supported with European Regional Development Fund (FEDER), and the financial support of Deep Blue Sea Enterprise S.L. RMJ would like to thank the Max Planck Institute for Marine Microbiology for funding. RRM acknowledges the financial support of the Spanish Ministry through the projects PR2015-00008 and PRX18/00048 for international scientific exchange.

References

[1] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

[2] Aruldass, C. A., Dufossé, L., & Ahmad, W. A. (2018). Current perspective of yellowish-orange pigments from microorganisms- a review. *Journal of Cleaner Production*, *180*, 168–182.

[3] Barquera, B. (2014). The sodium pumping NADH:quinone oxidoreductase (Na+-NQR), a unique redox-driven ion pump. *Journal of Bioenergetics and Biomembranes*, *46*(4), 289–298

[4] Baughn, A. D., & Malamy, M. H. (2004). The strict anaerobe *Bacteroides fragilis* grows in and benefits from nanomolar concentrations of oxygen. *Nature*, *427*(6973), 441–444.

[5] Cecchini, G. (2003). Function and Structure of Complex II of the Respiratory Chain. *Annual Review of Biochemistry*, *72*(1), 77–109.

[6] Charlebois, R. L., Doolittle, W. F. (2004). Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Research*, 14(12), 2469–2477.

[7] Consortium, T. (2016). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, *45*(D1), D158–D169.

[8] Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792–1797.

[9] Foley, M. H., Cockburn, D. W., & Koropatkin, N. M. (2016). The Sus operon: a model system for starch uptake by the human gut Bacteroidetes. *Cellular and Molecular Life Sciences : CMLS*, *73*(14), 2603–2617.

[10] Frost, L. S., Leplae, R., Summers, A. O., Toussaint, A. (2005). Mobile genetic elements: The agents of open source evolution. *Nature Reviews Microbiology*, 3(9), 722–732.

[11] Gil, R., Silva, F. J., Peretó, J., Moya, A. (2004). Determination of the core of a minimal bacterial gene set. *Microbiology and Molecular Biology Reviews*, 68(3), 518–37.

[12] Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., Lopez, R. (2010). A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Research*, 38(Web Server), W695–W699.

[13] Gupta, R. S., Lorenzini, E. (2007). Phylogeny and molecular signatures (conserved proteins and indels) that are specific for the *Bacteroidetes* and *Chlorobi* species. *BMC Evolutionary Biology*, 7*,* 71.

[14] Hahnke, R. L., Meier-Kolthoff, J. P., Garcia-López, M., Mukherjee, S., Huntemann, M., Ivanova, N., Woyke, T., Kyrpides, N.C., Klenk, H.P., Göker, M. (2016). Genome-based taxonomic classification of Bacteroidetes. *Frontiers in Microbiology*, 7(DEC).

[15] Hall, L. M. C., Fawell, S. C., Shi, X., Faray-Kele, M.-C., Aduse-Opoku, J., Whiley, R. A., & Curtis, M. A. (2005). Sequence Diversity and Antigenic Variation at the rag Locus of Porphyromonas gingivalis. *Infection and Immunity*, *73*(7), 4253–4262.

[16] Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome Biology*, 3(2), REVIEWS0003.

[17] Hunnicutt, D. W., & McBride, M. J. (2000). Cloning and Characterization of the Flavobacterium johnsoniae Gliding-Motility Genes gldB and gldC. *Journal of Bacteriology*, *182*(4), 911–918.

[18] Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., Tanabe, M. (2019). New approach for understanding genome variations in KEGG. *Nucleic Acids Research*, 47(D1), D590–D595.

[19] Kappelmann, L., Krüger, K., Hehemann, J.-H., Harder, J., Markert, S., Unfried, F., … Teeling, H. (2019). Polysaccharide utilization loci of North Sea Flavobacteriia as basis for using SusC/D-protein expression for predicting major phytoplankton glycans. *The ISME Journal*, *13*(1), 76–91.

[20] Kirchman, D. L. (2002). The ecology of *Cytophaga-Flavobacteria* in aquatic environments. *FEMS Microbiology Ecology*, 39(2), 91–100.

[21] Koonin, E. V. (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews Microbiology*, 1(2), 127–136.

[22] Krieg, N.R., Ludwig, W., Euzeby, J., Whitman, W.B. (2010) Phylum XIV. Bac- teroidetes phyl. nov. In: Krieg, N.R., Staley, J.T., Brown, D.R., Hedlund, B.P., Paster, B.J., Ward, N.L., Ludwig, W., Whitman, W.B. (Eds.), Bergey's Manual® of System- atic Bacteriology, Springer, New York, pp. 25–469.

[23] Larsbrink, J., Rogers, T. E., Hemsworth, G. R., McKee, L. S., Tauzin, A. S., Spadiut, O., … Brumer, H. (2014). A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. *Nature*, *506*(7489), 498–502.

[24] Lasica, A. M., Ksiazek, M., Madej, M., Potempa, J. (2017). The Type IX Secretion System (T9SS): Highlights and Recent Insights into Its Structure and Function. *Frontiers in Cellular and Infection Microbiology*, 7(May).

[25] Lemos, R. S., Fernandes, A. S., Pereira, M. M., Gomes, C. M., Teixeira, M. (2002). Quinol:fumarate oxidoreductases and succinate:quinone oxidoreductases: phylogenetic relationships, metal centres and membrane attachment. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1553(1–2), 158–170.

[26] Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar Buchner, A., Lai, T., Steppi, S., Jobb, G., Föster, W., Brettske, I., Gerber, S., Ginhart, A.W., Gross, O., Grumann, S., Hermann, S., Jost, R., Köning, A., Liss, T., Lüssmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Norbert, S., Vil- big, A., Lenke, M., Ludwig, T., Bode, A., Schleifer, K.H. (2004) ARB: a software environment for sequence data. Nucleic Acids Res. 32 (4), 1363–1371.

[27] Manfredi, P., Lauber, F., Renzi, F., Hack, K., Hess, E., & Cornelis, G. R. (2015). New iron acquisition system in Bacteroidetes. *Infection and Immunity*, *83*(1), 300–310.

[28] Marreiros, B. C., Calisto, F., Castro, P. J., Duarte, A. M., Sena, F. V., Silva, A. F., Sousa, F. M., Teixeira, M., Refojo, P. N., Pereira, M. M. (2016). Exploring membrane respiratory chains. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1857(8), 1039–1067.

583 [29] McBride, M. J., Zhu, Y. (2013). Gliding Motility and Por Secretion System Genes Are
584 Widespread among Members of the Phylum *Bacteroidetes*. *Journal of Bacteriology*, 195(2),
585 270.

586 [30] Medema, M. H., Takano, E., Breitling, R. (2013). Detecting sequence homology at the gene
587 cluster level with MultiGeneBlast. *Molecular Biology and Evolution*, 30(5), 1218–1223.

588 [31] Meuric, V., Rouillon, A., Chandad, F., Bonnaure-mallet, M. (2010). Putative respiratory chain
589 of *Porphyromonas gingivalis*. *Future Microbiology*, 5(5), 717–734.

590 [32] Moparthi, V. K., Hägerhäll, C. (2011). The evolution of respiratory chain complex i from a
591 smaller last common ancestor consisting of 11 protein subunits. *Journal of Molecular*
592 *Evolution*, 72(5–6), 484–497.

593 [33] Munoz, R., Rosselló-Móra, R., Amann, R. (2016). Revised phylogeny of *Bacteroidetes* and
594 proposal of sixteen new taxa and two new combinations including *Rhodothermaeota* phyl. nov.
595 *Systematic and Applied Microbiology*, 39, 281–296.

596 [34] Noinaj, N., Guillier, M., Barnard, T. J., & Buchanan, S. K. (2010). TonB-dependent
597 transporters: regulation, structure, and function. *Annual Review of Microbiology*, *64*, 43–60.

598 [35] Oren, A., da Costa, M. S., Garrity, G. M., Rainey, F. A., Rosselló-Móra, R., Schink, B.,
599 Trujillo, M. E., Whitman, W. B. (2015). Proposal to include the rank of phylum in the
600 International Code of Nomenclature of Prokaryotes. Int. J. Syst. Evol. Microbiol.,
601 65(11):4284-4287.

602 [36] Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A.,
603 Hugenholtz, P. (2018). A proposal for a standardized bacterial taxonomy based on genome
604 phylogeny. *BioRxiv*, (November 2017), 256800.

605 [37] Plotly Technologies Inc. (2015). Collaborative data science. Montreal, QC: Plotly
606 Technologies Inc. https://plot.ly

607 [38] Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G., Toth, I. K. (2016). Genomics and
608 taxonomy in diagnostics for food security: Soft-rotting enterobacterial plant pathogens.
609 *Analytical Methods,* 8(1), 12-24

610 [39] R Core Team, R. (2018). R: A Language and Environment for Statistical Computing. *R*
611 *Foundation for Statistical Computing, Vienna, Austria*. Online at http://www.R-project.org/

612 [40] Refojo, P. N., Teixeira, M., Pereira, M. M. (2012). The Alternative complex III: Properties and
613 possible mechanisms for electron transfer and energy conservation. *Biochimica et Biophysica*
614 *Acta (BBA) - Bioenergetics*, 1817(10), 1852–1859.

615 [41] Reintjes, G., Arnosti, C., Fuchs, B. M., & Amann, R. (2017). An alternative polysaccharide
616 uptake mechanism of marine bacteria. *ISME Journal*, *11*(7), 1640–1650.

617 [42] Reyes-Prieto, A., Barquera, B., Jua, O. (2014). Origin and Evolution of the Sodium -Pumping
618 NADH : Ubiquinone Oxidoreductase, 9(5), 1–14.

619 [43] Rodriguez-R, L. M., Konstantinidis, K. T. (2016). The enveomics collection: a toolbox for
620 specialized analyses of microbial genomes and metagenomes. Online at
621 https://doi.org/10.7287/peerj.preprints.1900v1

622 [44] Sato, K., Sakai, E., Veith, P. D., Shoji, M., Kikuchi, Y., Yukitake, H., Ohara, N., Naito, M.,
623 Okamoto, K., Reynolds, E. C., Nakayama, K. (2005). Identification of a new membrane-
624 associated protein that influences transport/maturation of gingipains and adhesins of
625 *Porphyromonas gingivalis*. *The Journal of Biological Chemistry*, 280(10), 8668–8677.

626 [45] Schauer, K., Rodionov, D. A., & de Reuse, H. (2008). New substrates for TonB-dependent
627 transport: do we only see the “tip of the iceberg”? *Trends in Biochemical Sciences*, *33*(7), 330–
628 338.

629 [46] Schöner, T. A., Fuchs, S. W., Schönau, C., Bode, H. B. (2014). Initiation of the flexirubin
630 biosynthesis in *Chitinophaga pinensis*. *Microbial Biotechnology*, 7(3), 232–241.

631 [47] Sousa, F. L., Alves, R. J., Ribeiro, M. A., Pereira-Leal, J. B., Teixeira, M., Pereira, M. M.
632 (2012). The superfamily of heme–copper oxygen reductases: Types and evolutionary
633 considerations. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1817(4), 629–637.

634 [48] Sun, C., Benlekbir, S., Venkatakrishnan, P., Wang, Y., Hong, S., Hosler, J., Tajkhorshid, E.,
635 Rubinstein, J. L., Gennis, R. B. (2018). Structure of the alternative complex III in a
636 supercomplex with cytochrome oxidase. *Nature*, 557(7703), 123–126.

637 [49] Teeling, H., Fuchs, B. M., Becher, D., Klockow, C., Gardebrecht, A., Bennke, C. M., Kassabgy,
638 M., Huang, S., Mann, A. J., Waldmann, J., Weber, M., Klindworth, A., Otto, A., Lange, J.,
639 Bernhardt, J., Reinsch, C., Hecker, M., Peplies, J., Bockelmann, F. D., Callies, U., Gerdts, G.,
640 Wichels, A., Wiltshire, K. H., Glockner, F. O., Schweder, T., Amann, R. (2012). Substrate-
641 Controlled Succession of Marine Bacterioplankton Populations Induced by a Phytoplankton
642 Bloom. *Science*, 336(6081), 608–611.

643 [50] Terrapon, N., Lombard, V., Gilbert, H. J., Henrissat, B. (2015). Automatic prediction of
644 polysaccharide utilization loci in *Bacteroidetes* species. *Bioinformatics*, 31(5), 647–655.

645 [51] Terrapon, N., Lombard, V., Drula, É., Lapébie, P., Al-Masaudi, S., Gilbert, H. J., Henrissat, B.
646 (2018). PULDB: the expanded database of Polysaccharide Utilization Loci. *Nucleic Acids*
647 *Research*, 46(D1), D677–D683.

648 [52] Thomas, F., Hehemann, J. H., Rebuffet, E., Czjzek, M., Michel, G. (2011). Environmental and
649 gut *Bacteroidetes*: The food connection. *Frontiers in Microbiology*, 2(MAY), 1–16.

650 [53] Toussaint, A., Chandler, M. (2012). Prokaryote Genome Fluidity: Toward a System Approach
651 of the Mobilome. In: van Helden J., Toussaint A., Thieffry D. (eds) Bacterial Molecular
652 Networks. Methods in Molecular Biology (Methods and Protocols), vol 804. Springer, New
653 York, NY.

654 [54] Unemoto, T., Hayashi, M. (1993). Na^+-translocating NADH-quinone reductase of marine and
655 halophilic bacteria. *Journal of Bioenergetics and Biomembranes*, 25(4), 385–391.

[55] Ussery, D. W., Wassenaar, T. M., Borini, S. (2009). Microbial Communities: Core and Pan-Genomics. In *Computing for Comparative Microbial Genomics* (pp. 213–228). London: Springer London.

[56] Ventura, M., Canchaya, C., Tauch, A., Chandra, G., Fitzgerald, G. F., Chater, K. F., van Sinderen, D. (2007). Genomics of *Actinobacteria*: tracing the evolutionary history of an ancient phylum. *Microbiology and Molecular Biology Reviews,* 71(3), 495–548.

[57] Wexler, A. G., Goodman, A. L. (2017). An insider's perspective: *Bacteroides* as a window into the microbiome. *Nature Microbiology*, *2*, 17026.

679 **Table 1.** Comparison of two non-redundant genome collections of *Bacteroidetes,* the smaller one of

680 which was used in this study.

	size (Mb)		G+C %mol content	
	478 genomes	89 genomes	478 genomes	89 genomes
maximum	9.77[a]	9.13[c]	62.1[e]	61.9[g]
minimum	0.31[b]	2.16[d]	23.8[f]	30.0[h]
Median +/- SD	3.96 +/-1.40	4.14+/-1.38	39.4+/-6.8	38.1+/-7.1

a *Microscilla marina* ATCC 23134^T

b endosymbiont of *Llaveia axin axin*, of the family *Blattabacteriaceae*

c *Chitinophaga pinensis* DSM 2588^T

d *Riemerella anatipestifer* ATCC 11845^T

e *Hymenobacter aerophilus* DSM 13606^T

f *Blattabacterium* sp. (endosymbiont of *Cryptocercus punctulatus* Cpu)

g *Hymenobacter* sp. APR13

h *Polaribacter* sp. Hel I 88

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

Table 2. Core genomes (CG) and taxonomically conserved sequences (P = Prevalent, E = Exclusive) of the taxa analyzed in this study (n = number of genomes). The reference genome contains copies of all conserved orthologous sequences (P+E) that served as templates for synteny searches in other genomes.

	n	CG	P+E	E	Reference genome
Bacteroidetes	89	155	87	31	*Spirosoma radiotolerans* DG5A
Flavobacteriia	48	352	44	4	*Algibacter* sp. HZ22
Bacteroidia	16	293	45	16	*Bacteroides thetaiotaomicron* VPI-5482
Lost in *Bacteroidia*	73	-	49	19	*Fluviicola taffensis* DSM 16823
Chitinophagia*	5	433	20	7	*Niastella koreensis* GR20-10
Sphingobacteriia	5	925	112	100	*Pseudopedobacter saltans* DSM 12145
Cytophagia	15	478	25	20	*Hymenobacter swuensis* DY53

* Two genomes of the novel class *Saprospiria* are included in the *Chitinophagia*.

712 **Table 3**. Uncharacterized yet conserved sequences of the *Bacteroidetes*. For sequences not found in
713 the UniProt database, accession numbers from the respective genomes are provided. The genome of
714 *Spirosoma radiotolerans* DG5A^T was used as the reference genome, since it contained all core-set
715 proteins. Putative annotations and pathways were compiled from the UniProt and KEGG databases
716 or inherited from non-bacteroidetal proteins with similarities above 60% in BLAST searches.
717

Accession	Putative annotation	Putative pathway
WP_046375230.1	RidA familiy protein	RNA processing
WP_046376045.1	Recombination protein A	Recombination
A0A0E3ZUE1	Methyltransferase	Unknown
A0A0E3V5X8	Uncharacterized protein. Putative gene *ribH*	Biosynthesis of secondary metabolites
A0A0E3ZUU2	HIT family hydrolase. Putative gene *rplJ*	Ribosome
A0A0E3ZTX4	Zinc ribbon domain protein	Unknown
A0A0E3ZV03	NGG1p interacting factor 3 protein, NIF3	Unknown
A0A0E3V763	Tyrosine recombinase XerC	Homologous recombination
WP_046573655.1	tRNA dihydrouridine synthase	RNA procesing
WP_046574203.1	Unmapped. Putative gene *recG*	Homologous recombination
A0A0E3V803	ABC transporter ATP-binding protein	Homologous recombination
A0A0E3V8I4	Phosphoesterase. Putative gene *rplW*	Ribosome
A0A0E3ZXS5	ATPase AAA	Unknown
A0A0E3ZXT0	2-hydroxyhepta-2,4-diene-1,7-dioate isomerase	Aminoacyl-tRNA biosynthesis
A0A0E3V8H2	Pyrophosphatase. Putative gene *rsfS*	Translation
A0A0E3ZYL4	Uncharacterized protein. Putative gene xerC	Homologous recombination
A0A0E3ZYP4	Uncharacterized protein. Putative gene adk	Nucleotide metabolism
WP_046578056.1	tRNA-specific adenosine deaminase	RNA processing
A0A0E4A0W5	Polypeptide deformylase	Unknown
A0A0E4A0D9	Putative pre-16S rRNA nuclease	Unknown
WP_046580219.1	Translational GTPase TypA	Translation

718
719
720
721

Figure 1. Composition of the *Bacteroidetes*' conserved sequences. (A) Concentric pie charts represent the core genome (outer), exclusive plus prevalent sequences (mid) and exclusive proteins only (inner). (B) Composition of the 49 proteins that are prevalent in the *Bacteroidetes* and do not belong to the core genome. KEGG and UniProt classifications were compared to produce the final categories (legend).

Figure in color

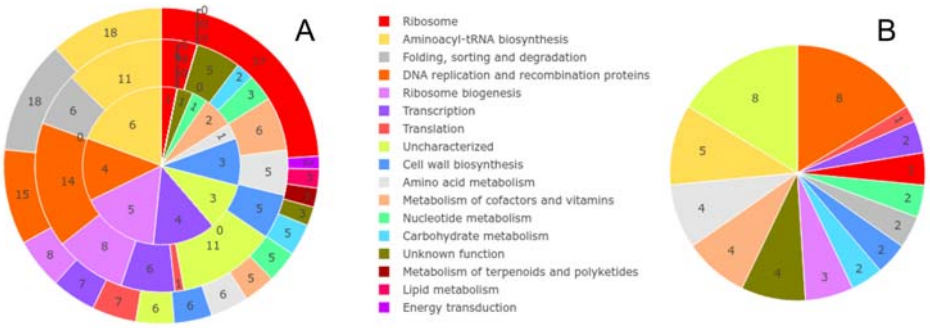

Figure 2. Conserved sequences at the taxonomic class rank sorted phylogenetically (columns) represented by their median sequence identity against other sequences in the same group of orthologs (lines). Presence of proteins coded in each genome is combined with color shades that represent the median identity. Absence is coded with the palest color (identity = 0). (A) Lighter shades of pink denote the divergence of the *Chryseobacterium* branch from other *Flavobacteriia*. (B) *Bacteroidia* genomes still encode some putative aerobic lifestyle proteins. (C) Maximum median identities are significantly lower between the sequences of the *Chitinophagia* and *Sphingobacteriia* than in other groups denoting a greater phylogenetic distance. (D) Dichotomic distribution of coding regions for proteins of the NADH-quinone oxidoreductase (Nuo) and the Na+-NQR (Nqr), together with the Rnf proteins.

Figure in Black & White


median identity percentage
0 20 60 100
Flavobacteriia
Bacteroidia
Sphingobacteriia
Chitinophagia
Cytophagia
Outgroup
min/max id
A
B
C
D
Nuo
Nqr
Rnf

Figure 3. Aerobic respiratory chains in *Bacteroidetes*. (A) Brief representation of gene clusters in 15 genomes that represent the variety of compositions described in this study. The nucleotide positions on both gene ends indicate locations in the respective genomes. Accession numbers precede genome names. The legend summarizes protein names and their color code. (B) Proposed aerobic respiratory chain in halophilic aerobic *Bacteroidetes*. (C) Proposed aerobic respiratory chain in mesophilic aerobic *Bacteroidetes*. (D and E) Proposed aerobic respiratory chain in nanaerobe *Bacteroidia*: (D) in all except the *Porphyromonadaceae*, (E) in *Porphyromonadaceae*. 'Q' stands for quinone. Structures of the complexes are based on representations found in the literature[25, 32, 42, 48] except for the cytochrome bd that is represented as a symmetric dimer for convenience to represent the subunits CydAB.

Figure in color

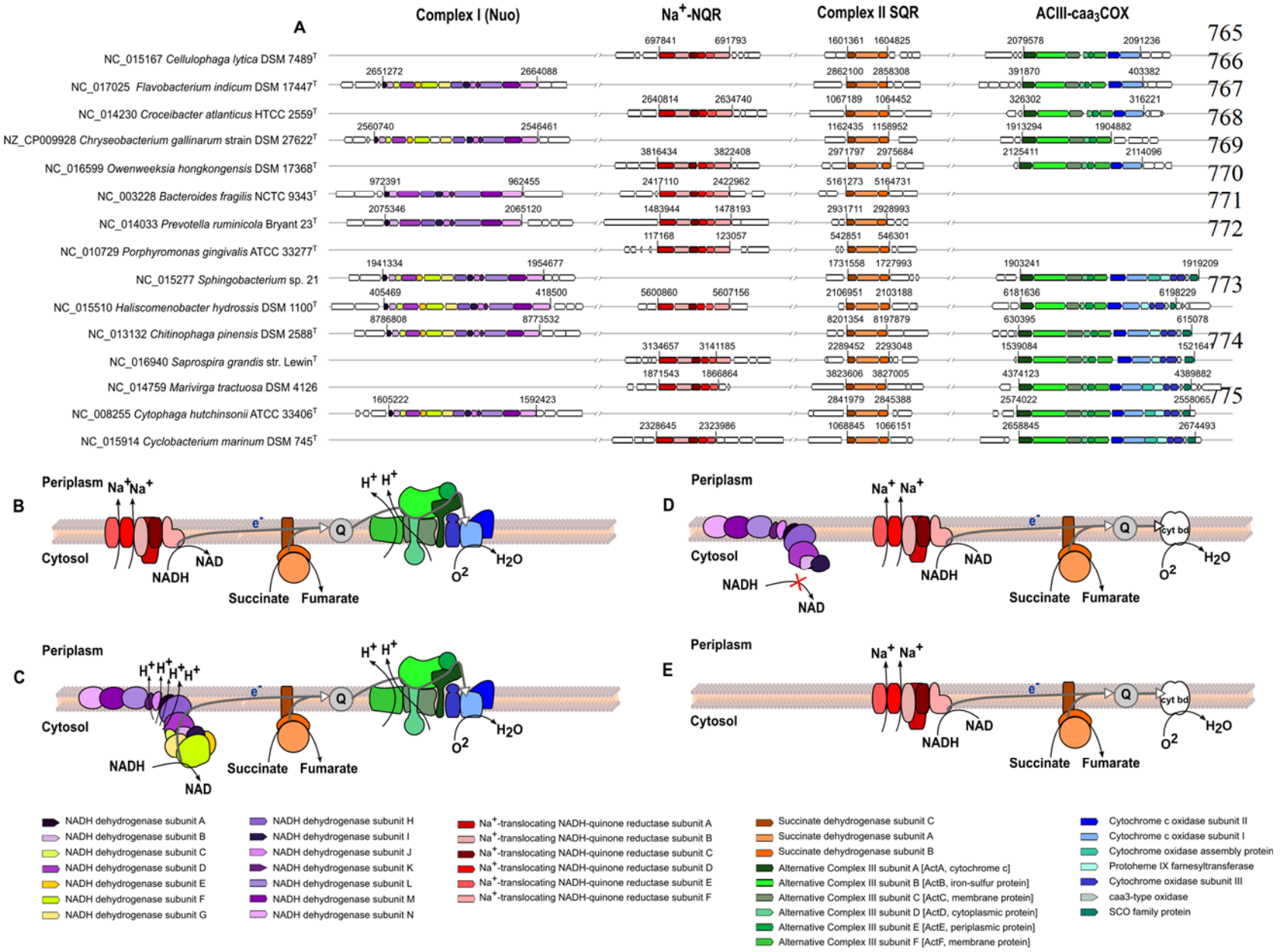
A
Complex I (Nuo)
Na+-NQR
Complex II SQR
ACIII-caa3COX
NC_015167 Cellulophaga lytica DSM 7489T
NC_017025 Flavobacterium indicum DSM 17447T
NC_014230 Croceibacter atlanticus HTCC 2559T
NZ_CP009928 Chryseobacterium gallinarum strain DSM 27622T
NC_016599 Owenweeksia hongkongensis DSM 17368T
NC_003228 Bacteroides fragilis NCTC 9343T
NC_014033 Prevotella ruminicola Bryant 23T
NC_010729 Porphyromonas gingivalis ATCC 33277T
NC_015277 Sphingobacterium sp. 21
NC_015510 Haliscomenobacter hydrossis DSM 1100T
NC_013132 Chitinophaga pinensis DSM 2588T
NC_016940 Saprospira grandis str. LewinT
NC_014759 Marivirga tractuosa DSM 4126
NC_008255 Cytophaga hutchinsonii ATCC 33406T
NC_015914 Cyclobacterium marinum DSM 745T
765
766
767
768
769
770
771
772
773
774
775
B
C
D
E
Periplasm
Cytosol
Na+ Na+
H+ H+
e-
Q
NAD
NADH
Succinate
Fumarate
O2
H2O
cyt bd
NADH dehydrogenase subunit A
NADH dehydrogenase subunit B
NADH dehydrogenase subunit C
NADH dehydrogenase subunit D
NADH dehydrogenase subunit E
NADH dehydrogenase subunit F
NADH dehydrogenase subunit G
NADH dehydrogenase subunit H
NADH dehydrogenase subunit I
NADH dehydrogenase subunit J
NADH dehydrogenase subunit K
NADH dehydrogenase subunit L
NADH dehydrogenase subunit M
NADH dehydrogenase subunit N
Na+-translocating NADH-quinone reductase subunit A
Na+-translocating NADH-quinone reductase subunit B
Na+-translocating NADH-quinone reductase subunit C
Na+-translocating NADH-quinone reductase subunit D
Na+-translocating NADH-quinone reductase subunit E
Na+-translocating NADH-quinone reductase subunit F
Succinate dehydrogenase subunit C
Succinate dehydrogenase subunit A
Succinate dehydrogenase subunit B
Alternative Complex III subunit A [ActA, cytochrome c]
Alternative Complex III subunit B [ActB, iron-sulfur protein]
Alternative Complex III subunit C [ActC, membrane protein]
Alternative Complex III subunit D [ActD, cytoplasmic protein]
Alternative Complex III subunit E [ActE, periplasmic protein]
Alternative Complex III subunit F [ActF, membrane protein]
Cytochrome c oxidase subunit II
Cytochrome c oxidase subunit I
Cytochrome oxidase assembly protein
Protoheme IX farnesyltransferase
Cytochrome oxidase subunit III
caa3-type oxidase
SCO family protein

Figure 4. MLSA phylogenies of most conserved Act, Gld and Sdh proteins. Concatenated sequences measured 1,675 a.a. ± 36 (Act), 1,502 a.a. ± 87 (Gld) and 911 a.a. ± 11 (Sdh). Sequences were aligned with Muscle [8] and phylogenetic trees were computed using neighbor-joining with 1,000 iterations and Kimura correction in ARB [26]. Percentages represent how often a bifurcation was reproduced over the 1,000 replicates. Scale bars represent the estimated substitution rate.

Figure in Black & White