The porphyran degradation system of the human gut microbiota is complete, phylogenetically diverse and geographically structured across Asian populations

Sophie Mathieu[1], Mélanie Touvrey[1], Laurent Poulet[1], Sophie Drouillard[1], Thirumalai Selvi Ulaganathan[3], Laure Ségurel[2], Miroslav Cygler[3], William Helbert[1*]

[1] CERMAV, CNRS and Grenoble Alpes Université, BP53, 38000 Grenoble Cedex 9, France

[2] Laboratoire de Biométrie et Biologie Evolutive, UMR 5558, CNRS - UCB Lyon 1,Villeurbanne cedex, France

[3] Department of Biochemistry, University of Saskatchewan, Saskatoon, Saskatchewan S7N 5E5, Canada

Abstract

The human gut microbiota can acquire new catabolic functions by integrating genetic material coming from bacteria associated to the food. The most illustrative example is the acquisition of genes by the gut microbiota of Asian populations coming from marine bacteria living at the surface of algae incorporated in diet. In order to interrogate the pace of acquisition of algal polysaccharide utilizing loci (PUL) and their diffusion rate inside populations, we investigated the PUL dedicated to degradation of porphyran, the main polysaccharide of the red algae *Porphyra sp*. used to prepare maki-sushi. We demonstrated that both methylated and unmethylated fractions were catabolized without the help of external enzymes. The PUL organization was conserved in several Bacteroidetes strains, highlighting lateral transfers inside the microbiota, but we point out various conserved mutations, deletion and insertions. Geographic distribution of the variants showed that specific mutation and recombination events appeared independently in geographically distant populations.

Introduction

The impact of diet on the diversity and composition of the bacterial communities inhabiting the human gut is now well documented [1,2,3]. For example, several studies have highlighted the lower bacterial diversity of populations having westernized diet, as compared to hunters-gatherers or populations with more diversified food [4,5]. Conversely, it is also fascinating to observe that the gut microbiota can adapt and is able to acquire new functional capacities, notably by lateral transfer of genetic materials between strains of the microbiota but also from strains associated with the food ingested by the host [6,7]. One of the most illustrative example of how the food participates in shaping the human gut microbiota, at the molecular level, is the discovery of the lateral transfer of genes encoding enzymes involved in the degradation of porphyran, the cell wall polysaccharide of the red algae *Porphyra sp.* used to prepare the maki-sushi, from a marine strain living at the surface of the algae to the gut bacteria *Bacteroidetes plebeius* found in the Japanese gut microbiota [8]. Since this discovery, other algal polysaccharides degradation systems (i.e. agarose, alginate, carrageenan), likely acquired by lateral transfer from marine organisms, have also been studied [9,10,11].

Porphyran is a agar type polysaccharide made of two disaccharides repetition: agarobiose and porphyranobiose decorated up to 64% by methyl groups [12,13] (Figure 1A). In *B. plebieus*, the genes encoding the enzymes for the degradation of porphyran are co-localized and co-regulated in a so-called "polysaccharide utilizing loci" (PUL). Many enzymes of the PUL (i.e. glycoside hydrolases and sulfatases) have been characterized, explaining the degradation process of the agarose components and the non-methylated fractions of porphyran by *B. plebieus* [8,14,15] (Figure 1B), the enzymes involved in the degradation of methylated porphyran segments remaining to be discovered.

Even though the fronds of *Porphyra sp*. are nowadays mostly consumed by Asian populations for its taste and nutritional value [16], the most ancient proof that humans harvested and stored *Porphyra sp.* was found in the archaeological site of Monte Verde (Chile) where population occupied 12,300 years ago [17]. The traditional use of *Porphyra spp.* by the first nations living along the Pacific coast of North America was also documented by ethnobotanists [18,19,20,21]. In South America, pre-Inca and Inca populations also incorporated red algae in their diet [22]. The very ancient know-how and use of *Porphyra sp.*, preserved for millennia along the Pacific Ocean's coast, suggests that several lateral transfer events of genetic material dedicated to porphyran degradation from marine bacteria to gut bacteria might have occurred in ancient human populations. The origin of the algal degradation system in the human gut and the timing

of these transfers have yet to be estimated. Are there traces of ancestral algal catabolism acquisitions still conserved in the contemporary human gut? How recent are the lateral transfers currently observed in Asian populations?

We observed that the *B. plebeius* PUL dedicated to the degradation of porphyran encoded enzymes have a larger spectrum of specificities than previously reported [8,14,15] and can notably process the methylated fraction of porphyran. The discovery of novel enzymes that can accommodate the methyl group of porphyran substrate in their active sites demonstrates that the complete degradation of all the components of porphyran (e.g. agarose, methylated and non-methylated porphyran fractions) can be achieved autonomously by the enzymes encoded by the PUL. Moreover, we have found that this self-sufficient PUL is present not in one but in many bacterial strains of the gut suggesting that several lateral transfers have occurred between bacteria of the gut. We noticed several mutations and recombination events between individuals suggesting the acquisition of the porphyran PUL was not recent. To further characterize the genetic differences between populations living in different geographical areas, we examined genomic and metagenomic data available to identify and locate individuals having the porphyran degradation machinery.

Results

Porphyran PUL, the PUL dedicated to the porphyran degradation, identified in *B. plebieus* contained genes encoding two β-porphyranases (BpGH16B, Bacple_01689; BpGH86A, Bacple_01693) and two β-agarase (BpGH16A; Bacple_01670; BpGH86B, Bacple_01694) [8,14] (Figure 1B). The oligosaccharides resulting from degradation of porphyran by these enzymes are then degraded further by exo-acting enzymes including sulfatase (BpS1_11, Bacple_01701), α-L-galactosidase (BpGH29, Bacple_01702) and β-D-galactosidase (BpGH2C, Bacple_01706) produced by *B. plebeius* or another agarolytic strain such as *Bacteroidetes unifromis* [15] (Figure 1B). The presence of a gene (BpGH50, Bacple_01683) encoding an enzyme grouped in the GH50 family known to contain agarases, was also studied but its substrate specificity was not demonstrated [11]. Altogether, our knowledge on the catabolism of porphyran is limited to its non-methylated fraction. We continued the detailed analysis of the porphyran PUL of *B. plebeius* and studied the uncharacterized glycoside hydrolases BpGH16C (Bacple_01703) and the BpGH2C, completed the structural data of the sulfatase BpS1_11, and demonstrated the porphyran PUL encodes a

portfolio of enzymes catalyzing to the complete degradation of the methylated fraction of the polysaccharide.

Biochemical and molecular characterization of the endo-acting 6O-methyl-porphyranase BpGH16C (Bacple_01703)

The gene Bacple_01703 encodes a glycoside hydrolase of the GH16 family (BpGH16C), which includes characterized agarases in sub-families GH16_15-16, and β-porphyranases in sub-families GH16_11-12 [24]. BpGH16C was classified in the GH16_14 sub-family that contains one characterized β-agarase (*Vibrio sp*. strain PO-303, BAF62129.1) [25] and one endo-6O-methyl-β-porphyranase (*Wenyingzhuangia fucanilytica*, WP_068825734.1) [26]. We assayed BpGH16C on various marine polysaccharides including carrageenan, agarose and porphyran. Only porphyran was degraded, leading to the production of a series of oligosaccharides characteristic of endo-acting glycoside hydrolase (Figure 2A). The structural characterization by NMR of the end-products purified by chromatography [27] revealed the occurrence of a methyl group on the β-linked D-galactose residue (Figure 2B), demonstrating that BpGH16C is an endo-6O-methyl-β-porphyranase able to accommodate the methyl group of porphyran in its active site. For comparison, we have examined the substrate specificities of one another predicted GH16_14 endo-6O-methyl-β-porphyranase from *Paraglaciecola atlantica* T6c (Patl_0824, ABG39352.1) presenting 51.1% identity with the BpGH16C and one GH16_12 β-porphyranase (Patl_0805, ABG39333.1) homologous to BpGH16B. Analyses of the degradation products confirmed the different porphyranase specificities (Figure 2A).

We determined the crystal structure of BpGH16C at 1.9 Å resolution (PDB ID 8EP4) (Figure 3A). The enzyme has a β-sandwich jelly-roll fold with two stacking antiparallel β-sheets typical of proteins from GH16 family [24] and like in most GH16, a Ca^{2+} is bound on the convex side of the β-sandwich [28]. One molecule of Hepes from the crystallization buffer bound in the center of the cleft, with the 2-hydroxyethyl end buried into the molecule interior and the sulfate group exposed on protein surface (Figure 3A). To get a better understanding of substrate binding and catalysis, we crystallized the E145L inactive mutant with the tetrasaccharide L-α-6O-sulfate-Gal-(1→3)-D-β-Gal-(1→4)-L-α-6O-sulfate-Gal-(1→3)-D-a/β-Gal and determined its structure at 1.8 Å resolution (PDB ID 8EW1). An electron density was present within the groove in all the three independent molecules in the asymmetric unit fitted a D-β-Gal-(1→4)-L-α-6O-sulfate-Gal-(1→3)-D-β-Gal trisaccharide (Figure 3B). The electron density for the D-Gal on the nonreducing end is weaker than for the first two residues and no density is observed for

the fourth residue indicating its mobility in the crystal. The residues correspond to the -1, -2 and -3 positions according to the standard nomenclature [29].

The trisaccharide sits edge-on in the groove with C5 substituent of the ring in the -1 site directed toward the bottom of the cleft while O1 and O2 hydroxyls point toward the solvent. Trp134 stacks against hydrophobic side of the -1 residue while Arg67 and Glu238 hydrogen bond to the O4 hydroxyl. Residue in the -2 subsite is stacked between Trp64 and the edges of Trp143 and Phe169. Finally, the sulfate group of the -2 residue is hydrogen bonded to Asn61, Ser136 and Asn234 through a bridging water molecule. The D-Gal on the non-reducing end (-3 site) has somewhat different orientation in molecules A/B and C and makes only one bridging hydrogen bond through a bridging water (Figure 3C). The structure of the complex with a trisaccharide adds also to our understanding of the catalytic mechanism. Glu150 is indeed the closest to the O1 hydroxyl that formed the bond to the next sugar eliminated by hydrolysis, while Glu145 approaches from the opposite side of the ring relative the O1 hydroxyl (Figure 3C) and is likely helping in proper orientation of the substrate. Asp147 is directed to the bridging O1 and might be helping in catalysis.

Structural alignment of BpGH16C with three other GH16 porphyranases: porphyranase A (PDB ID 3ILF) and porphyranase B from *Zobellia galactanivorans* (PDB ID 3JUU) [8] and another porphyranase from *B. plebeius* (PDB ID 4AWD) [14] shows that the Gly132 is replaced by either a Ser or a Thr, making this pocket too small to accommodate an additional methyl group on the C6-O (Figure 3D). Sequence alignment of β-porphyranases with BpGH16C and the two other characterized 6O-methyl-β-porphyranases: Patl_0824 of (*P. atlantica* T6c, this study) and AXE80_06940 (*W. fucanilytica*) [26] confirmed that the replacement of the Ser/Thr by Gly was correlated to the accommodation of the methyl group in the -1 sub-site of the catalytic site (Figure S1). The BpGH16C is the first crystallized 6-OMe-porphyranase that tolerates the 6-methyl group on the galactose ring. Indeed, there is an extra space in the active site at the end of the pocket that accommodates C6-OH, with its bottom formed by Gly132 (Figure 3E).

Complete degradation of oligo-methyl-porphyrans

Following the same strategy reported by Robb and co-workers [15], we have incubated the methylated oligosaccharides with the BpS1_11 sulfatase, the BpGH29 α-L-galactosidasee and the predicted BpGH2C β-D-galactosidase. Purified methylated and non-methylated oligo-porphyrans were incubated with the BpS1_11 sulfatase and analyses of the products by ^{1}H NMR showed a strong chemical shift (about 0.1 ppm) of the signal attributed to the proton at

position 6 of the galactose located at the non-reducing end, demonstrating the removal of the sulfate ester group independently of the degree of methylation of the oligosaccharides (Figure S2).

Crystal structure of the BuS1_11 sulfatase from *B. uniformis* complexed with a di-saccharides [15] and the *B. plebeius* sulfatase BpS1_11 on its own (this study, PDB ID 7SNJ, 1.64 Å resolution) and complexed with a tetrasaccharide (this study, PDB ID 7SNO, 2.1 Å resolution) revealed that there is room in the active site to accommodate oligo-methyl-porphyrans (Figure S3). The porphyran sulfatase BpS1_11 belongs to the formyl-glycine dependent sulfatases, requiring maturation of a cysteine or a serine into the formyl-glycine residue as an active catalytic residue. A large N-terminal domain of BS1_11 contains active site residues and forms main part of the substrate binding site. A smaller C-terminal domain completes the substrate binding site.

To map the details of substrate binding, we determined the structure of H214N inactive mutant with L-α-6O-sulfate-Gal-(1→3)-D-β-Gal-(1→4)-L-α-6O-sulfate-Gal-(1→3)-D-a/β-Gal tetrasaccharide substrate (PDB ID 7SNO). The electron density map clearly showed the entire tetrasaccharide bound to BpS1_11 (H214N). Following the recently proposed nomenclature [30], the tetarasaccharide binds to the 0, +1, +2 and +3 positions (Figure S3). The 6S-L-Gal in position 0, which the sulfate is removed, has all three hydroxyl groups hydrogen bonded to the protein side chains: O2 to Arg271, O3 to Asp215 and Arg271, O3 to Asp345 and Arg347. The 6-sulfate group is facing Ser83 (formyl-glycine in mature enzyme) and one of its oxygens is a ligand to the Ca^{2+} ion. D-Gal in +1 position hydrogen bonds C6OH to NE2 of His133 and its O2 is bonded to the O2 of L-Gal (position 0) through a bridging water molecule. The +2 6S-L-Gal makes only one contact, between the 6-sulfate oxygen and His428. The +3 D-Gal extends far from the binding site and makes no contacts with the protein. However, in the crystal, this sugar hydrogen bonds from O1 and O4 as well as O5 within the ring to the guanidino group of Arg68 from a symmetry-related molecule. These interactions stabilize the +2 and +3 sugars in the crystal. Importantly, the 6-O group of +1 D-Gal is not tightly constraint by the enzyme and there is sufficient space for the additional methyl group present in the 6-O-Me-D-Gal (Figure S3).

We confirmed the exo-based cycle of porphyran depolymerization [15] using tetra-saccharides as starting substrate (Figure S4). After desulfation of the tetra-saccharides, BpGH29 exo-α-L-galactosidase was active on the methylated and non-methylated substrate

demonstrating that the methyl group did not hinder the activity of the enzyme. Independently of Robb and co-workers, [15] we have determined the structure of BpGH29 (PDB ID 7SNK). The superposition with their structure (PDB ID 7LJJ) shows root-mean-squares deviation of 0.6 Å, indicating that the structures that were crystallized under different conditions are virtually identical, and confirms that ordering of the loop 408-426 and small rearrangement of other surrounding loops result from the substrate binding to the active site.

Finally, we observed that the methylated or non-methylated D-galactose residue located at the non-reducing end of the trisaccharide was cleaved by the BpGH2C β-D-galactosidase (Figure S4). Altogether, the enzymology and crystallography experiments conducted in this work revealed the pathway leading to the degradation of the methylated fraction of porphyran. Our observations combined with previous investigations demonstrate that all the methylated and non-methylated fractions of porphyran are digested by the enzymes encoded by porphyran PUL of *B. plebeius*. Therefore, this PUL can catabolize autonomously the chemically complex porphyran, without the help of enzyme(s) external to this PUL.

Genetic diversity and bacterial occurrence of the porphyran utilizing loci

Previous transcriptomic experiments conducted on *B. plebeius* grown in the presence of porphyran in the culture medium interrogated the genes Bacple_01668 to Bacple_01706 [14]. Interestingly, the transcriptomic results suggested that at least two groups of genes were differently regulated. The cluster Bacple_01692 to Bacple_01699 genes – named *PUL-PorA* in Figure 1B – was moderately up-regulated in the presence of porphyran. This contrasts with the two neighboring clusters of genes: Bacple_01668 to Bacple_01689 (*PUL-PorC*) and Bacple_01700 to Bacple_01706 (*PUL-PorB*, Figure 1B), which were highly up-regulated (10-fold more than *PUL-PorA*). The *PUL-PorB* genes encoding the enzymes studied herein (Figure 1B) were strongly up-regulated, in agreement with their implication in porphyran metabolism [14].

Until now, *B. plebeius* DSM 17135 is the only gut bacteria shown to carry the porphyran PUL. In order to investigate whether other human gut bacteria can also digest porphyran, we used BLASTn to identify homologs of the entire coding sequences of the *B. plebeius* DSM 17135 *PUL-PorB* among the genomes of isolated human gut bacteria (more than 10,000 non redundant genomes). We found that the PULs were present in the genomes of 26 *Phocaeicola*/*Bacteroides* strains (corresponding to at least 8 different identified species), all coming from strains isolated from the gut microbiome of Asian individuals (Table S1). The

same species/strains isolated from other human populations worldwide had no homologs of porphyran PULs. The identified PULs were present in different species, including *B. dorei*, *B. eggerthii*, *B. ovatus*, *B. plebeius*, *B. stercoris*, *B. uniformis*, *B. vulgatus* and *B. xylanisolvens*, but were absent in other common *Bacteroides* sister species from the human gut, such as *B. intestinalis or B. fragilis.* This phylogenetic diversity probably reflects numerous lateral transfer events between bacteria of the gut microbiome. An important consequence of this observation is that the taxonomic composition of an individual human is not necessarily a good predictor of its metabolic capacity.

While the *PUL-PorB* organization as observed in *B. plebieus* DSM 17135 was conserved in all the other strains we have identified, the gene organization of *PUL-PorA* was only conserved in 19/26 strains (Figure S5A). This genomic region was incompletely sequenced in *B. plebeius* strain AM09-36 and *B. eggerthii* strain AM42-16. In four other strains (*B. stercoris* strain AF05-4, *B. sp*. AF25-38AC *B. stercoris* strain AM32-16LB, B. *xylanisolvens* strain AF38-2) the homolog of the *B. plebieus* β-agarase (GH86, Bacple_01694) appeared to have recombined with a DNA fragment composed of two genes predicted to encode metabolic enzymes: 2-dehydro-3-deoxygluconokinase and 2-dehydro-3-deoxyphosphogluconate aldolase (Figure S5B). For the same four strains, a deletion of the genes encoding the GH50 (Bacple_01683) and GH105|GH154 (Bacple_01684) located in *PUL-PorC* was also observed, highlighting the remodeling of the PUL by insertion and deletion of DNA segments (Figure S5B).

The Neighbor-Joining tree calculated with the six concatenated genes of the *PUL-PorB* (Figure 4A, based on 12,888 bp) revealed three clades with high bootstrap values, allowing to classify the strains in three groups accordingly to their *PUL-PorB* genetic diversity (Figure 4A). In the first group (GI, 18 strains, including the *B. plebeius* DSM 17135 reference), all six genes presented 99-100% identity with the reference, except for *B. plebeius* strain AM49-7BH, which had only 97-98% identity for GH29 and GH16. In the second group (GII, six strains), the six genes were 99-100% identical to each other, but presented 94-99% identity with the *B. plebeius* reference. The genes of the *B. plebeius* strain AM49-7BH suggested that it resulted from a recombination event that occurred between SusD, SusC and GH2 genes from GI and Sulf and GH29 genes more related to GII. The recombination site was identified in the gene coding for GH16 (Figure S6). For this reason, we classified *B. plebeius* strain AM49-7BH as GIrec. The gene sequences of the *B. plebeius* strain AM09-36 and *B. stercoris* strain AF05-4 were more distantly related to the other strains and were classified as a third group (GIII). The genes of the

B. stercoris strain AF05-4 presented 92-96% identity with the *B. plebeius* reference. On the other hand, the *B. plebeius* strain AM09-36 sequence seemed to result from a recombination event between GI and GIII (Figure 4A), since the SusD, SusC and GH2 genes presented 99-100% identity with GI genes, while the S1_11 sulfatase and GH29 genes were nearly identical to GIII genes. Again, the recombination site was identified in the gene coding for GH16 (Figure S6). For this reason, we classified *B. plebeius* strain AM09-36 as GIIIrec.

Distribution of PUL-PorB in assembled Human gut metagenomes

We explored the worldwide geographic distribution of the porphyran degradation system in diverse human gut microbiota by probing the six *PUL-PorB* genes against assembled human gut metagenomic datasets (Table S2). We scanned 19 projects (9 Asian, 7 North American, 1 South American and 2 African datasets), covering in total 2720 biosamples for which gut metagenomes were obtained and assembled. Only fully sequenced genes were retrieved. We found that at least one fully sequenced gene of *PUL-PorB* was present in the assembled gut metagenomes of 370 East Asians out of the 1144 individuals tested – corresponding to 36% of the Chinese individuals and 20 % of the Japanese individuals, while we found only ten positive individuals among the 1262 North American gut metagenomes (0.8% of the individuals) and none among the other geographic areas, including 117 individuals from South Asia (India, Bangladesh), 47 from South America (Peru) and 150 from Africa (Tanzania, Madagascar) (Table S2). Therefore, as observed in previous investigations, the occurrence of algal polysaccharides degradation systems seems to be mostly restricted to East Asian populations.

The gene and the corresponding individuals were characterized as belonging to GI/GIrec, GII, GIII and/or GIIIrec based on their gene sequence (Table S3). For most individuals (92.6%), the genes detected were attributed to a single group; however, few individuals (5.9%) presented *PUL-PorB* genes that belonged to two different groups. The individuals were then grouped by geographic location (Figure 4B). In order to obtain large sets of individuals, we distinguished populations living in North-East China (Hangzhou), South-East China (Shenzen) and Japan. In both set of Chinese individuals, the group GI/GIrec was present in more than half of the population (56% and 53%, respectively) followed by the group GII (33% and 39%, respectively). The group GIIIrec was only present in Chinese individuals, and more specifically, in North-East China were it was occurring in 15% of the individuals. In contrast, in Japan, GI and GII were found in similar proportion (42%), and the GIII group represented 20% of individuals. North-East China, South-East China and Japan populations

were overall found to be significantly different (X-squared=49.78, p-value=$2.5x10^{-5}$). This difference was mostly driven by a significant difference between the Japanese population and each of the Chinese ones (q-value=0.0008 and 0.0323 for Japan versus Hangzhou and Shenzen, respectively, after FDR Bonferroni correction for multiple testing), while the two Chinese populations were not significantly different from one another (q-value=0.1280). We then tested separately which group was different in prevalence between populations and found that GIII and GIIIrec were significantly different (proportion test, q-value=0.0009 and 0.0406, respectively), while GI and GII were not (proportion test, q-value=1 and 0.8925, respectively).

Analysis of short read datasets

The analysis of the assembled metagenomes revealed that 92.6% of the individuals carried only one version of the *PUL-PorB*. This observation may however result from some bias inherent to the bioinformatics processes involved in the building up of the contigs. Therefore, to analyze the data devoid of such biases, we have selected a 54 nucleotides sequences allowing to probe the different *PUL-PorB* groups directly on raw sequencing data (e.g. short reads). The probe was part of the Bacple_01703 (GH16) gene encoding the methyl-porphyranase, which presented characteristic mutations for the GI, GIrec, GII, GIII and GIIIrec groups (Figure 5). This method allowed us to explore new metagenomic datasets which were not assembled, thus expending the number of sampling sites, as well as the number individuals tested, with now a total of 2313 individuals (Table S2).

As observed previously, we confirmed by this independent approach that *PUL-PorB* is quite common in East Asian coastal populations (54% in China and 67% in Japan). We note that the prevalence of positive individuals is much higher with this approach than with the assembled datasets, with also observed a higher variance across populations within Japan (4-92%) than within China (20-69%). The histograms presented in Figure 5 show the number of short reads and their attribution to the GI, GIrec, GII, GIII and GIIIrec groups fort each individual. Among positive individuals, we found that most of them (80.0%) carried short reads attributed to a single group. Interestingly, when individuals presented two types of short reads (or more), one group was clearly predominant (the median ratio of the first highest number of reads to the second highest number of reads was 4). Consequently, we characterized individuals as belonging to a group if they carried only this group, or if this group was dominant, which we defined as having more than twice the amount of reads for this group compared to the second most prevalent group. Using this definition, we found that 93.8% of individuals carried short

reads attributed to a single dominant group, a number very similar to what was observed with the assembled datasets.

Using these categories, analyses of the short reads confirmed that GI and GIrec mutations were predominant in Chinese populations (54.2% of individuals in total, versus 40.3% in Japan) while the GII ones were most common in Japanese population (51.0% of individuals, versus 42.4% in China). Individuals with the GIII mutation were found mainly in the Japanese population (13.8%) and in lower amount in Chinese population (2-6.7%). Finally, GIIIrec mutations were the most numerous in the Hangzhou population (9.3%), and in much less common elsewhere (0.3-5%).

Including only categories with large enough numbers (i.e., dominant GI, GIrec, GII, GIII, GIIIrec and co-dominant GI+GII groups), we found that the five populations were overall significantly different (X-squared= 93.2, p-value=2.0×10^{-11}) (Figure 6). This difference was mostly driven by a significant difference between the Japanese population and three of the Chinese ones (all except Beijing, q-value<0.004), as well as between Hongkong and Hangzhou (q-value=3.7×10^{-7}). We then tested separately which group was significantly different in prevalence between populations and found that all were significantly different (proportion test, q-value<0.015) except GIrec (proportion test, q-value=1), but GIII and GIIIrec presented the stronger difference (proportion test, q-value=2.2×10^{-6} and 3.8×10^{-8}, respectively), similarly to the MG results.

Discussion

Biochemical and crystallographic investigations showed that the porphyran PUL encodes tailored enzymes dedicated to the complete degradation of the polysaccharide, including the methylated fraction, giving galactopyranoside and 6-O-methyl-galactopyranoside as end-products. We showed that the 6-O-methyl-porphyranase (BpGH16C) presents an active site able to accommodate the methyl group, in contrast with the active sites of the previously investigated β-porphyranases. Similarly, the exo-6O-L-galactose porphyran sulfatase Bp S1_11 can accommodate methylated oligosaccharides in its active site. The demethylation of 6-O-methyl-galactopyranoside was located in agarose PUL of several aerobic bacteria belonging to the genus Gammaproteobacteria, Bacteroidetes and Planctomycetes. The reaction was obtained with specific monooxygenase enzyme system, including ferredoxin, ferredoxin reductase and P450 monooxygenase [31]. However, demethylation of galactose residue in anaerobic bacteria

has not been elucidated yet and no encoding gene surrounding the porphyran utilization loci could be suspected to be involved in the demethylation pathway.

Because harvesting, drying and cooking *Porphyra sp*. were conserved practices in numerous populations living along the Pacific and because these ancestral traditions seem to follow the dispersion line of *Homo sapiens* from Asia to the South of the American continent, one could expect that the porphyran degradation system may be present in populations living along the Pacific coast. Analysis of metagenomics data of population in Peru or Fidji (Table S2) revealed no degradation system resembling the Asian porphyran PUL. In this context, analyses of the distribution and history of the porphyran PUL was restricted to the Chinese and Japanese populations. Homologous gene sequences of the *PUL-PorB* found in populations living by the sea were also present in continental population of China, showing that the *PUL-PorB* has spread among the Chinese population independently of the geographical location. One hypothesize that the porphyran PUL was acquired, at first, by people living by the sea and then was spread among the continental population. The genes were conserved in the gut microbiota of the individual as long as the diet, including algae, maintained a selective pressure.

Investigations of the metagenomic data were conducted using assembled and raw (e.g. short reads) data. The analyses demonstrated that the investigated populations were structured with a limited number of genetic variations that were used to characterize 5 groups (GI, GIrec, GII, GIII and GIIIrec). Exploitation of the raw data allowed us to estimate a higher number of individuals carrying strains involved in the porphyran degradation, highlighting the likely better sensitivity of this approach. Also, analyses of the short reads revealed that one PUL variant was present or was dominating in individuals, giving an apparent haploid organization of the PUL. The architecture of the microbiota at the bacterial level seems to drive the structuration at genetic and molecular level, or reciprocally.

The genetic organization and the dispersion of the various *PUL-PorB* suggested that they have diverged from an ancestral gene cluster. However, there is no evidence that the evolution from an ancestral PUL occurred in the human gut or that homologous PUL, which evolved independently were acquired at different periods. The low number of genetic variations of the PULs and their wide distribution suggests that acquisition by lateral transfer from marine bacteria living at the surface of red algae toward strains of the human gut remains a rare event. Searching marine strain genomes and metagenomics data (e.g. Tara Ocean) failed to identify porphyran PUL similar or distantly related to those observed in the human gut. Therefore, there

is yet no environmental data that could help to estimate the pace and number of lateral transfer events between marine bacterial strains and human gut microbiota.

Because the *PUL-PorB* GI and GII groups were common in the Chinese and Japanese populations and because they were identified in many different human gut strains, one can suppose that these clusters of genes were present in the early human gut microbiota, allowing their dispersion in the Asian populations and lateral exchanges between gut microbiota strains. The occurrence of the GIIIrec group was found mostly in the Hangzhou population and in lower abundance in the Shenzen and Beijing populations, while almost absent in the Japanese population, suggesting a more recent appearance than the GI and GII groups. It is interesting to note that this recombinant strain is currently found in populations with the lowest prevalence of the parent strain GIII (i.e., China), but not in the only population with a non-negligible frequency of GIII (i.e., Japan). This might suggest that after having appeared, the GIIIrec might have locally replaced the GIII group, which bodes the question of the functional equivalence of these strains. In general, the rate of dispersion of the PUL variants inside a given population and between populations geographically distant may probably reflect exchanges between populations which involved displacement of individuals.

In conclusion, the architecture of the human gut microbiota across populations has mostly been studied and functionally interpreted at a genera or species level, while we see here with the example of the porphyran degradation system that the important functional information is carried at the strain, or even the genetic level. Indeed, the PUL system is carried by various species and strains of the Bacteroides genus, and the GI/GII/GIII groups are not associated with particular strains, so the inference of these groups based on taxonomy would not be possible. Although it is still difficult to state which mutations were obtained by independent evolution inside the human gut population and which mutations were acquired by lateral transfer, in this study, we demonstrated that human populations geographically distant presented common (e.g. GI, GII) and specific genetic characteristics (e.g. GIrec, GIII and GIIIrec). The current structure of the investigated populations is likely the result of lateral transfer, recombination, as well as migration events, which reflect an evolution history of the gut microbiota and therefore, of its host.

References

1. David, L.A et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559-63 (2014)
2. Gentile, C.L. & Weir, T.L. The gut microbiota at the intersection of diet and human health. *Science* **362**, 776-780 (2018)
3. Illiano, P., Brambilla, R. & Parolini, C. The mutual interplay of gut microbiota, diet and human disease. *FEBS J.* **287**, 833-855 (2020)
4. De Filippo, C. et al. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc Natl Acad Sci* **107**, 14691-6 (2010)
5. Schnorr, S.L. et al. Gut microbiome of the Hadza hunter-gatherers. *Nat. Commun*. **5**, 3654 (2014)
6. Sitaraman R. Prokaryotic horizontal gene transfer within the human holobiont: ecological-evolutionary inferences, implications and possibilities. *Microbiome* **6**, 163 (2018)
7. Boto, L., Pineda, M. & Pineda, R. Potential impacts of horizontal gene transfer on human health and physiology and how anthropogenic activity can affect it. *FEBS J.* **286**, 3959-3967 (2019)
8. Hehemann, J. H. et al. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**, 908–912 (2010)
9. Thomas, F. et al. Characterization of the first alginolytic operons in a marine bacterium: from their emergence in marine Flavobacteriia to their independent transfers to marine Proteobacteria and human gut Bacteroides. *Environ. Microbiol*. **14**, 2379-94. (2012)
10. Mathieu, S. et al. Ancient acquisition of "alginate utilization loci" by human gut microbiota. *Sci Rep.* **8**, 8075. (2018)
11. Pluvinage, B. et al. Molecular basis of an agarose metabolic pathway acquired by a human intestinal symbiont. *Nat. Commun*. **13**,1043 (2018)
12. Zhang, Q. et al. Chemical characteristics of a polysaccharide from *Porphyra capensis* (Rhodophyta). *Carbohydr. Res.* **340**, 2447–2450 (2005)
13. Usov, A.I. Polysaccharides of the Red Algae. In Derek Horton, editor: Advances in Carbohydrate Chemistry and Biochemistry, Vol. 65, Burlington: Academic Press, pp. 115-217 (2011)
14. Hehemann J.H. et al. Bacteria of the human gut microbiome catabolize red seaweed glycans with carbohydrate-active enzyme updates from extrinsic microbes. *Proc. Nat. Acad. Sci.* **109**, 19786-19791 (2012)
15. Robb, C.S. et al. Metabolism of a hybrid algal galactan by members of the human gut microbiome. *Nat. Chem. Biol*. **18**, 501-510 (2022)
16. Taboada, M.C., Millán, R. & Miguez, M. I. Nutritional value of the marine algae wakame (*Undaria pinnatifida*) and nori (*Porphyra purpurea*) as food supplements. *J. Appl. Phycol*. **25**, 1271–1276 (2013)
17. Dillehay, T.D. et al. Monte Verde: seaweed, food, medicine, and the peopling of South America. *Science*. **320**, 784-6 (2008)
18. Goddard, P. E. (1902) Life and culture of the Hupa. In American Archaeology and Ethnology, University of California publication, Vol1, ed. Berkeley the university press
19. Coulter, J. & Rose, J.N. (1902) Plants used by the Indians of the Mendocino county, California. In Contributions from the United States National Herbarium Vol. VII, n°3 pp. 295-408 (1900-1902), Washington, D.C., Smithsonian Institution Press, 1890-
20. Kuhnlein, H.V. & Turner, N.J. Traditional plant foods of Canadian indigenous peoples, nutrition, botany and use. Gordon & Breach Science Publishers Ltd (1992)
21. Turner, N. The ethnobotany of edible seaweed (and related species; Rhodophyta: Bangiales) and use by First Nations on the Pacific coast of Canada. *Can. J. Bot.* **81**, 283-293 (2003)

22. Michanek G (1979) Seaweed resources for pharmaceutical uses. In Marine Algae in Pharmaceutical Science, Volume 1 pp 203-236. Eds H.A. Hoppe, T. Levring, Y. Tanaka. Walter de Gruyter, Berli, New York
23. Giles, K., Pluvinage, B. & Boraston, A.B. Structure of a glycoside hydrolase family 50 enzyme from a subfamily that is enriched in human gut microbiome bacteroidetes. *Proteins* **85**, 182-187 (2017)
24. Viborg, A.H. et al. A subfamily roadmap of the evolutionarily diverse glycoside hydrolase family 16 (gh16). *J. Biol. Chem*. **294**, 15973-15986 (2019).
25. Dong, J., Tamaru, Y. & Araki, T. Molecular cloning, expression, and characterization of a beta-agarase gene, agaD, from a marine bacterium, *Vibrio sp*. strain PO-303. *Biosci. Biotechnol. Biochem.* **71**, 38-46. (2007)
26. Zhang, Y., Chang, Y., Shen, J., Mei, X. and Xue, C. Characterization of a novel porphyranase pccommodating methyl-galactoses at its iubsites. *J. Agric Food Chem.* **68**,7032-7039 (2020)
27. Correc, G., Hehemann, J.-H., Czjzek, M. & Helbert, W. Structural analysis of the degradation products of porphyran digested by *Zobellia galactanivorans* β-porphyranase A. *Carbohydr. Polym.* **83**, 277-283 (2011)
28. Naretto, A. et al. The agar-specific hydrolase ZgAgaC from the marine bacterium *Zobellia galactanivorans* defines a new GH16 protein subfamily. *J. Biol. Chem.* **294**, 6923-6939. (2019)
29. Davies, G.J., Wilson, K.S. & Henrissat, B. Nomenclature for sugar-binding subsites in glycosyl hydrolases. *Biochem. J*. 321, 557-9. (1997)
30. Hettle, A.G. et al. The molecular basis of polysaccharide sulfatase activity and a nomenclature for catalytic subsites in this class of enzyme. *Structure* **26**, 747-758.e744. (2018)
31. Reisky, L. et al. Oxidative demethylation of algal carbohydrates by cytochrome P450 monooxygenases. *Nat. Chem. Biol.* **14**, 342–344 (2018)

Material and methods

Purification of porphyran

6 g of dried *Porphyra columbina* were suspended in 120 ml de distilled H_2O and autoclaved for 30 min at 120°C. After 14 h at room temperature, the suspension was centrifuged at 8000 rpm during 30 min at 4°C. The supernatant was added to 120 ml of pure ethanol (50% v/v EtOH final concentration) and the solution was maintained at 4°C for 2h. After centrifugation (8000 rpm, 30 min, 4°C), the porphyran, present in the supernatant, was precipitated by addition of 130 ml of pure ethanol (67% v/v EtOH final concentration). After centrifugation the porphyran pellet was dissolved in distilled water, and dialysis for 3 days against distilled water using dialysis membrane with a cut-off 1000 Da (pre-wetted Spectra/Por® 6 dialysis tubing – Sprectrumlabs). The polysaccharide was lyophilized and the purity was verified by ^{1}H-NMR. The yield of purification was about 15-20 % w/w dried algae.

Heterologous expression of *Bacteroides plebeius* DSM 17135 PUL-PorB enzymes

The genes encoding the predicted glycoside hydrolases (BpGH29, Bacple_01702; BpGH16C, Bacple_01703; BpGH2C, Bacple_01706) and sulfatase (BpS1_11, Bacple_01701) from *B. plebeius* DSM 17135 were cloned using genomic DNA as template in the pET-28 or pFO4 expression plasmid [32] without their signal peptides identified with SignalP [33]. The expression strains harboring the recombinant expression plasmids were grown in Luria Bertani (LB) medium supplemented with 50 µg/ml kanamycin (pET28a plasmid) or 100 µg/ml ampicillin (pFO4 plasmid) until the OD_{600nm} reached 0.6 in a shaking incubator working at 180 rpm and 37°C. After the addition of isopropyl-β-D-thiogalactopyranoside (IPTG), the temperature was cooled down at 20°C and maintained overnight.

Cultures (200 mL) were centrifuged at 6000 g for 15 min and the bacterial pellet was suspended in 10 ml of buffer A (20 mM Tris pH 8, 500 mM NaCl, 20 mM Imidazole). The cells were lysed using a cell disrupter (Constant system Ltd). Insoluble fractions were removed by centrifugation at 20000 g during 30 min at 4°C and the supernatant was loaded on a 1 mL HisTrap^{TM} HP column (GE Healthcare) connected to a NGC chromatography system (BioRad). The proteins were eluted with a imidazole gradient from 20 to 300 mM gradient of imidazole. Pure enzymes were obtained after a size exclusion chromatography using a HiLoad 16/600 Superdex 75 pg column in buffer B (10 mM Tris-HCl pH 8, 50 mM NaCl).

<u>Enzymatic assays</u>

Enzymatic degradations were monitored by analytical gel permeation chromatography using Superdex S200 10/300 and Superdex peptide 10/300 (GE Healthcare) columns mounted in series and connected to a high-performance liquid chromatography (HPLC) Ultimate 3000 system (Thermo Fisher). The injection volume was 20 µL and the elution was performed at 0.4 mL.min-1 in 0.1 M NaCl. Oligosaccharides were detected by differential refractometry (Iota 2 differential refractive index detector, Precision Instruments).

The oligosaccharide products were purified by semi-preparative gel permeation chromatography using three HiLoad® 26/600 Superdex® 30 pg (GE Healthcare) columns mounted in series and connected to a semi-preparative size-exclusion chromatography system which consisted of a Knauer pump (pump model 100), a refractive detector (iota2 Precision instrument) and a fraction collector (Foxy R1) mounted in series. The elution was conducted at a flow rate of 1.2 $mL.min^{-1}$ at room temperature using 100 mM $(NH_4)_2CO_3$ as eluent. The collected fractions were freeze-dried prior NMR and mass spectrometry analyses.

Samples were exchanged twice with D_2O and were transferred to a 5 mm NMR tube. ^1H NMR spectra were recorded at 323 K using an Advance III 400 MHz spectrometer (Bruker). Chemical shifts are expressed in ppm in reference to water. The HOD signal was not suppressed.

NMR

^1H NMR spectra were recorded with a Bruker Avance 400 spectrometer operating at a frequency of 400.13 MHz. Samples were solubilized in D_2O at a temperature of 293 K for the oligosaccharides and 353 K for the polysaccharide. Residual signal of the solvent was used as internal standard: HOD at 4.85 ppm at 293 K and 4.35 ppm at 343 K. Proton spectra were recorded with a 4006 Hz spectral width, 32,768 data points, 4.089 s acquisition times, 0.1 s relaxation delays and 16 scans.

Crystallization, data collection and structure solutions.

The homogenous fractions of proteins obtained from gel permeation chromatography were concentrated and crystallization experiments were attempted. Initial crystals were obtained by screening against wide range of commercial screens and the hits were optimized by hanging drop diffusion method. The drop containing 1 µl of protein and 1 µl of reservoir solution was incubated over 1 ml of reservoir solution and crystal growth was monitored regularly. The conditions of the best diffracting crystals of the three proteins were listed: BpS1_11 (Bacple_01701) was crystallized at 23 mg/ml in 40% Peg200, 0.1 M Sodium Citrate

buffer pH 5.5 and 30% MPD; BpGH29 (Bacple_01702) was crystallized at 18 mg/ml in 20% Peg 8K and 0.1 M KH2PO4; BpGH16C (Bacple_01703) was crystallized at 34 mg/ml in 16% Peg 8K, 0.1 M Hepes pH 7.5 and 0.2M Calcium acetate. For diffraction experiments, the crystals were cryo protected in reservoir solution containing 20% MPD (BpS1_11), 25% glycerol (BpGH29), 30% ethylene glycol (BpGH16C) and flash frozen in liquid nitrogen. The diffraction data of all the crystals were collected at the 08ID beamline at the Canadian Light Source [34].

The X-ray diffraction data were processed using XDS [35]. The structures were solved by molecular replacement using the program Phaser in Phenix package [36]. For BpS1_11, the structure solution was obtained by molecular replacement using the search model an arylsulfatase from *Pseudomonas aeruginosa* (PDB ID 1HDH) [37]. For BpGH29, the structure solution was obtained using the α-L-fucosidase from *Fusarium graminearum* as a search model (PDB ID 4NI3) [38]. BpGH16C was solved using porphyranase B from *Zobellia galactanivorans* (PDB ID 3JUU) [39] as a model. All the structures were refined with the Phenix software [40] and manual rebuilding and solvent placement was conducted with the COOT program [41]. The stereochemistry of all the models were validated with MolProbity [42].

To obtain the complex structure, putative active site mutants of BpS1_11 (BpS1_11(H214N)) and BpGH16C (BpGH16C(E145L)) were made using the Quickchange site-directed mutagenesis protocol and using KOD polymerase. Briefly, the plasmid containing the gene of interest was amplified with the primer pairs carrying the mutation using KOD polymerase. After PCR, the template plasmid was digested with 1 µl of DpnI enzyme for an hour at 37°C. Five µl of PCR product was then transformed into 50µl of chemically competent *E. coli* DH5α cells. The clones carrying the desired mutation were confirmed by sequencing and proceeded for crystallization experiments.

Based on the structural comparison with other sulfatases, the His 214 residue in BpS1_11 was mutated to Asn (H214N). The mutant protein BpS1_11(H214N) was expressed and purified following the same protocol used for wild type enzyme. Purified Bacple_01701(H214N) was crystallized from solution containing 40% PEG 200, 0.1M sodium citrate buffer pH 5.5 and 30% MPD. The putative active site mutant of BpGH16C(E145L) was purified following the same protocol used for wild type enzyme. BpGH16C(E145L) was crystallized from solution containing 20% PEG 8K, 0.1 M Hepes pH 7.5 and 0.2 M calcium acetate. BpS1_11(H214N) and BpGH16C(E145L) crystals were soaked in the tetrasaccharide solution for an hour before the diffraction experiments. The soaked crystals were flash frozen

in liquid nitrogen and diffraction data was collected at the 08ID beamline in the Canadian light source [34]. The diffraction data was processed with XDS [35]. One round of rigid body refinement was carried out using phenix refinement program (Adams et al., 2011). Manual rebuilding and substrate placement were done using coot (Emsley et al., 2010). The geometry was validated using MolProbity program [42].

Human gut metagenomes analysis

More than 10,000 non redundant genomes of bacteria isolated from the human gut were were downloaded from various bacterial genome repositories including NCBI [43], PATRIC [44], references strains of the Human Metagenome Project [45], the Human Gastrointestinal Bacteria Culture Collection (HBC) [46], the Culturable Genome Reference (CGR) [47] and human Gut Microbial Biobank (hGMB)[48]. Bacterial strains harboring the porphyran PUL were identified after BLASTn of *PUL-PorA* (17,980 bp), *PUL-PorB* (12,888 bp) and *PUL-PorC* (19,198 bp) sequences against the set of genome. Genes of the identified porphyran PUL variant were retrieved from bacterial genome repositories. Genes and concatenated genes were aligned using Muscle and distant tree were built using MEGA6 allowing to distinguish GI, GIrec, GII, GIII and GIIIrec group of genes.

Assembled metagenomes data set (Table 2) were downloaded and analyzed locally. Genes of the porphyran PUL were BLASTn and aligned with Muscle using MEGA6. Only coding genes with full length were kept and compared with those of the reference bacteria. The genes encoding the sulfatase were not used because of its low discriminative power. We did not discriminate between GI and GIrec because most of the time, there was not enough information. To identify GIIIrec, we required to observe genes matching both GI and GIII on the same contig. When only part of the genes was available, we considered the major group (i.e., GI or GIII) as the most probable one, creating a slight bias downward for GIIIrec. When multiple contigs in the same individual were found to be positive and they corresponded to different groups, we considered the individual as multi-groups (e.g., GI+GII). When we could not assign a contig to one of the 4 groups defined earlier, we considered the individual to be unresolved.

Distant trees were built allowing to group the genes in GI, GIrec, GII, GIII or GIIIrec. Analyses of short reads data set (Table S2) were conducted using 54 nt probes (Figure S7). The probes were used to search at Sequence Read Archive (SRA) data at NCBI.

References

1. Groisillier, A. et al. MARINE-EXPRESS: taking advantage of high throughput cloning and expression strategies for the post-genomic analysis of marine organisms. *Microb. Cell. Fact.* **9**, 45 (2010)
2. Petersen, T.N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods.* **8**, 785-6 (2011)
3. Grochulski, P., Fodje, M.N., Gorin, J., Labiuk, S.L. & Berg R. Beamline 08ID-1, the prime beamline of the Canadian Macromolecular Crystallography Facility. *J. Synchrotron Radiat*. **18**, 681-4 (2011)
4. Kabsch, W. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 133-44 (2010).
5. McCoy, A.J. et al. Phaser crystallographic software. *J. Appl. Crystallogr*. **40**, 658-674. (2007).
6. Boltes, I. et al. I. 1.3 Å structure of arylsulfatase from *Pseudomonas aeruginosa* establishes the catalytic mechanism of sulfate ester cleavage in the sulfatase family. *Structure* **9**, 483-91. (2001)
7. Cao, H., Walton, J.D., Brumm, P., Phillips, G.N. Jr. Structure and substrate specificity of a eukaryotic fucosidase from *Fusarium graminearum*. *J. Biol. Chem.* **289**, 25624-38. (2014)
8. Hehemann, J. H. et al. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**, 908–912 (2010)
9. Adams, P.D. et al. The Phenix software for automated determination of macromolecular structures. *Methods*. **55**, 94-106. (2011)
10. Emsley, P., Lohkamp, B., Scott, W.G. & Cowtan K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr*. **66**, 486-501. (2010)
11. Chen, V.B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr*. **66**, 12-21. (2010)
12. Kitts, P.A. et al. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res*. **44**, D73-80. (2016).
13. Wattam, A.R. et al. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res.* **45**, D535–D542. (2017)
14. Human Microbiome Jumpstart Reference Strains Consortium. A catalog of reference genomes from the human microbiome. *Science*. **328**, 994-9. (2010)
15. Forster, S.C. et al. A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat. Biotechnol*. **37**, 186-192 (2019)
16. Zou, Y. et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol*. **37**, 179–185. (2019)
17. Liu, C. et al. Enlightening the taxonomy darkness of human gut microbiomes with a cultured biobank. *Microbiome* **9**, 119. (2021)

Captions for figures

Figure 1: **A**) Chemical structure of the repetition moieties of porphyran. The sulfated disaccharides - porphyranobioses – can present methyl group at the position 6 of the D-galactose giving methylated and none-methylated porphyran components. Agarobiose units, obtained by desulfation/cyclization of the L-galactose residue during the biosynthesis, are methylated accordingly. **B**) Organization of the *B. plebieus* porphyran PUL. The PUL is divided in three segments (Pul-PorA, -Por B and -PorC) based on previous transcriptomic analyses. The enzyme functions were determined in (1) Hehemann *et al*., 2012, (2) Giles *et al.*, 2017, (3) Rodd *et al.*, 2022 and (4) this study.

Figure 2: **A**) Size exclusion chromatography of the degradation products of porphyran incubated with the methyl-6O-β-poprhyranase BpGH16C (GH16_14). The degradation profile was compared with the two *P. atlantica* T6c porphyranases grouped in the GH16_12 (Patl_0824) and GH16_14 (Patl_0805) sub-families. **B**) ^{1}H NMR of the disaccharides end-products of BpGH16C.

Figure 3: A) Crystal structure of BpGH16C (Bacple_01703). Chain A of asymmetric unit is shown in cartoon representation and colored in spectrum from N to C terminal. The HEPES molecule in the active site is shown as stick representation. The N- and C-terminals are marked, and the strands are labeled in their order in the sequence. **B**) The substrate binding site in the mutant BpGH16C-E145L complexed with the D-β-Gal-(1→4)-6O-sulfate-L-Gal disaccharide end-product. The semitransparent surface of the protein is colored by the electrostatic potential. The residues forming the site are shown in stick representation. The sulfate group from L-galactose 6 sulfate is docked into a positively charged pocket. **C**) The hydrogen bonds between the disaccharide in substrate binding site residues. Several H-bonds are bridged by water molecules (W). Glu145 (in blue and thicker bonds) from the native structure is superimposed on the Leu145 in the mutants. **D)** Structure comparison of BpGH16C (green), PorA (PDB id-3ILF) (magenta) and PorB (PDB id- 3JUU) (yellow) showing the active site region. The single mutation of Ser129/Thr137 (PorA/PorB) (sidechains shown as spheres) to Gly132 in BpGH16C provides space for accommodating a methyl group at C6 on L-galactose residue. **E**) A cros-section of the surface representation of the Gal sugar and the cavity in the binding site near Gly132. There is free space within the cavity sufficient to accommodate the methyl group of a methylated porphyran.

Figure 4: A) Phylogenetic tree calculated with the concatenated gene sequences *of PUL-PorB*. Sequence homology observed allowed to identify recombination sites and to create groups of homologous porphyran PUL. **B**) Distribution of the different group of porphyran PUL in human gut metagenome assembly of South east and North East Chinese, and Japanese populations.

Figure 5: The 54 nucleotides sequence characteristic of different group of porphyran PUL (top) were Blastn against the short read sequencing of metagnomic data recorded on Chinese and Japanese population. The number of detected short reads are indicated for each individuals and were grouped as a function of the geographic location of the sampling.

Figure 6: Distribution of the various porphyran PUL among Chinese and Japanese population. The pie charts derive from the histogram presented in Figure S8 obtained from the analyses of short read sequencing of Chinese and Japanese metagenomics data. For 80% of the individuals, only one group of porphyran PUL was observed. For the other individuals, a dominant group was considered when the number short-reads detected was, at least, twice more abundant than the other. In other situation, the co-occurrence of groups was reported.

Figure S1: Structural alignment of the *B. plebieus* DSM 17135 BpGH16C 6-O- methyl-β-porphyranase compared with the other characterized 6-O- methyl-β-porphyranases (*P. atlantica* T6c, Patl_0824; *W. fucanilytica* AXE80_06940) and characterized β-porphyranases (*B. uniformis*_NP1, BuGH16A; *P. atlantica* T6c, Patl_0805; *B. plebeius* DSM 17135, BpGH16B). The blue circle indicates the serine/threonine amino acids observed in the active site of the β-porphyranases which are replaced by a glycine amino acid in 6-O- methyl-β-porphyranases allowing to accommodate the methyl group.

Figure S2: ^{1}H NMR recorded before and after incubation with the sulfatase BpS1_11 on purified disaccharides end-products of the BpGH16C. The occurrence of the methyl group located at the position 6 of the D-galactose didn't hinder the removal of the sulfate ester group of the L-galactose residue located at the non-reducing end.

Figure S3: A) Structural details of Bacple_01701. Schematic representation of Bacple_01701 shown in cartoon representation and colored from N to C terminus. Calcium ion is shown as green sphere. **B**) Conserved active site residues from S1 family sulfatases. The residues are shown in stick representation in the order Bacple_01701/ 5G2V/1HDH/6BIA. **C**) Surface representation of H214 mutant with the tetrasaccharide, showing pocket like architecture of active site. The 6S sulfate subsite is labelled as S. **D**) Complex structure of H214N mutant. The electron density (2fo-fc) map for the sugar contoured at 1.5σ. Sugars numbered from 0 to +3 and 6S sulfate subsite labelled as S. Residues at H bonding distance to the substrate are shown in stick representation. **E**) Structural comparison of H214 mutant with 1HDH. The residues involved in catalysis S83 (overlays with DDZ from 1HDH), H133, H214 (this is the mutated residue) are shown in stick representation.

Figure S4: ^{1}H NMR recorded after sequential incubation of the methylated tetrasaccharide end-products of the 6-O-methyl-porphyranase BpGH16C with the sulfatase BpS1_11 followed by the α-L-galactosidase BpGH29 and the D-galactosidase BpGH2C.

Figure S5: Gene organization of the porphyran degradation system of *B. plebieus* compared with other identified human gut bacteria. **A**) Phylogenetic tree was calculated using concaneted gene sequences of *PUL-PorB*. PUL organizations of each strains were indicated based on the available sequencing data. **B**) Comparison of the two main gene organizations of the porphyran PUL observed in the set of bacteria investigated highlighting deletion and recombination events.

Figure S6: Recombination sites observed in the GH16 of PUL-PorB of the group GI/distant GII giving GIrec and the groups GI/GIII giving GIIIrec.

Figure S7: Phylogenetic trees calculated for all gene sequences of the members of the *PUL-PorB*.

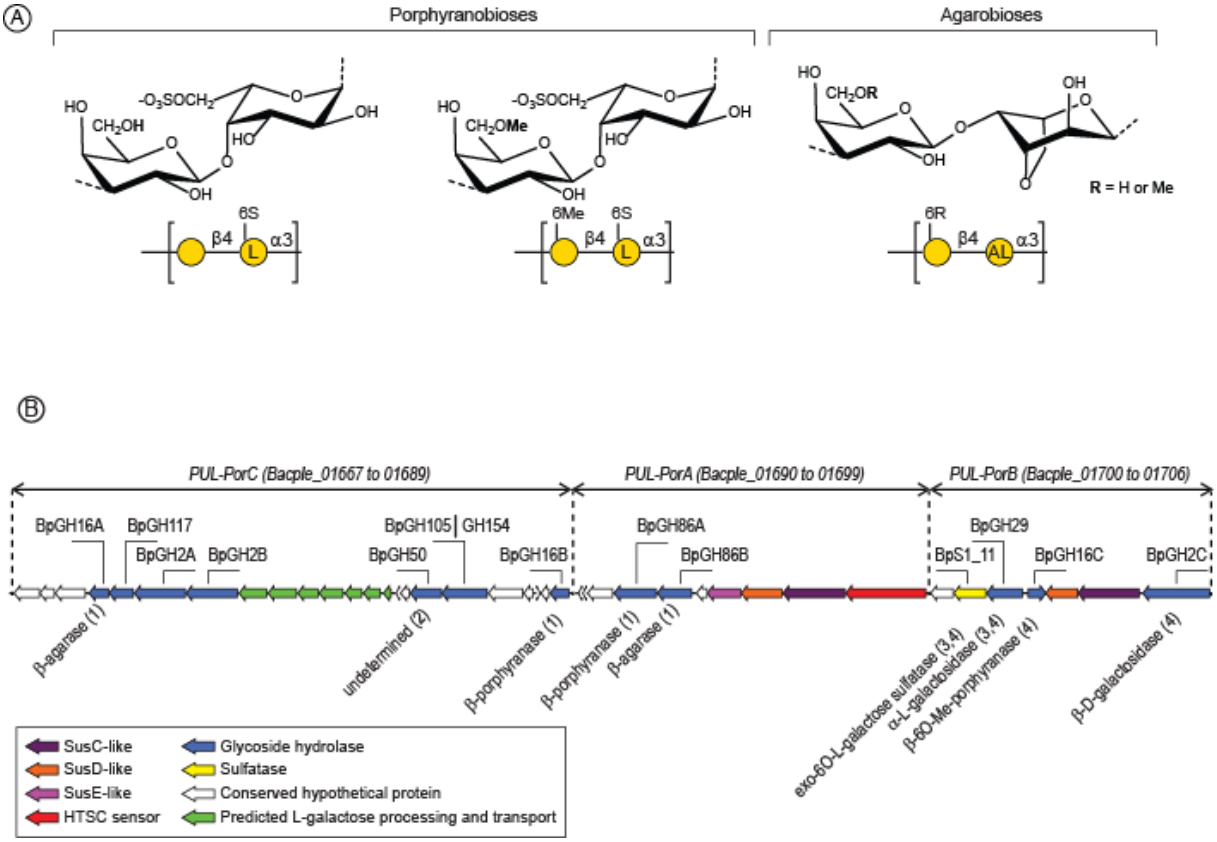

Figure 1: **A**) Chemical structure of the repetition moieties of porphyran. The sulfated disaccharides - porphyranobioses – can present methyl group at the position 6 of the D-galactose giving methylated and none-methylated porphyran components. Agarobiose units, obtained by desulfation/cyclization of the L-galactose residue during the biosynthesis, are methylated accordingly. **B**) Organization of the *B. plebieus* porphyran PUL. The PUL is divided in three segments (Pul-PorA, -Por B and -PorC) based on previous transcriptomic analyses. The enzyme functions were determined in (1) Hehemann *et al.*, 2012, (2) Giles *et al.*, 2017, (3) Rodd *et al.*, 2022 and (4) this study.

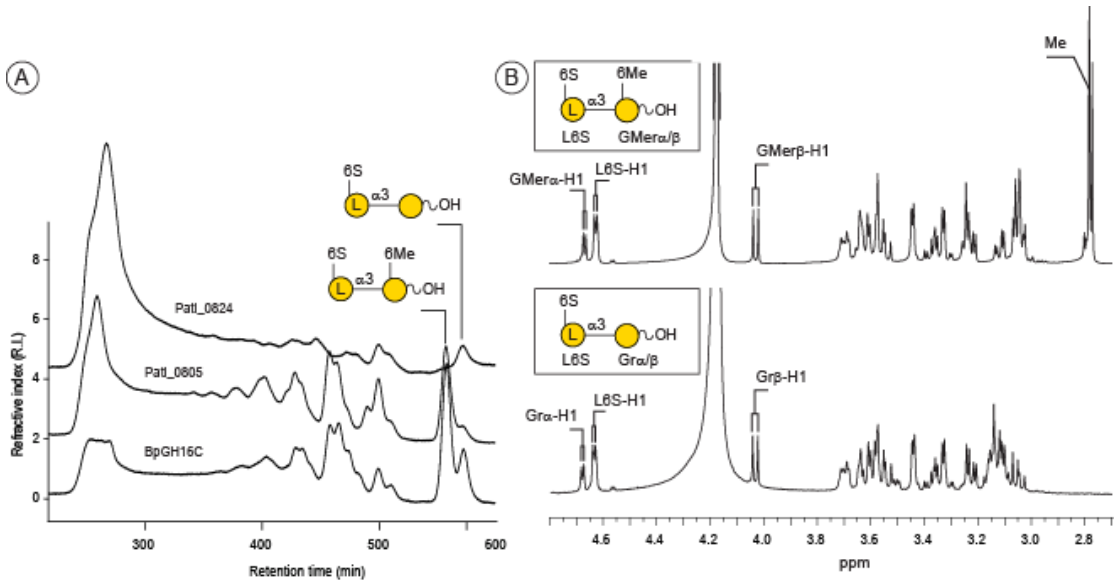

Figure 2: **A**) Size exclusion chromatography of the degradation products of porphyran incubated with the methyl-6O-β-poprhyranase BpGH16C (GH16_14). The degradation profile was compared with the two *P. atlantica* T6c porphyranases grouped in the GH16_12 (Patl_0824) and GH16_14 (Patl_0805) sub-families. **B**) ^{1}H NMR of the disaccharides end-products of BpGH16C.

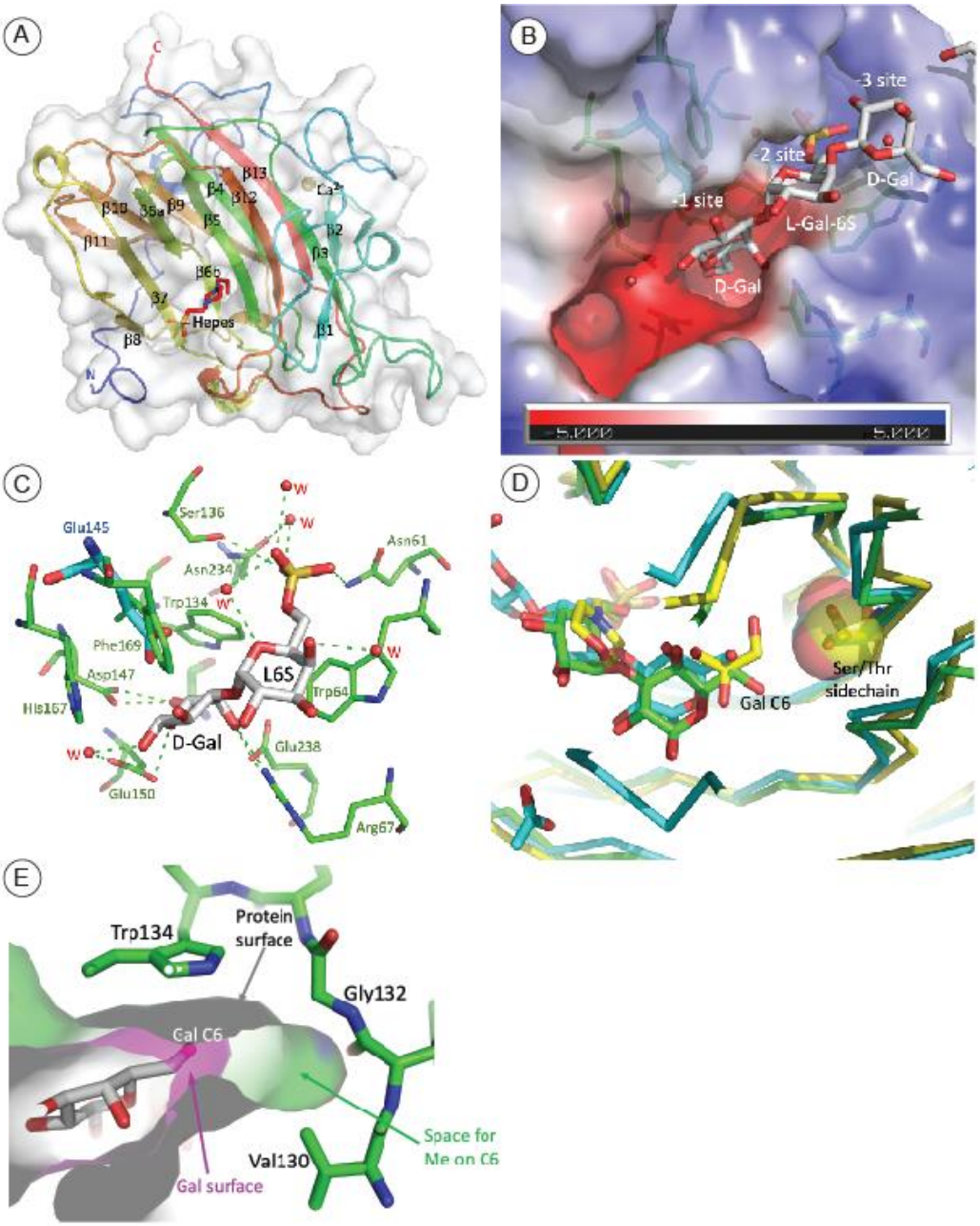

Figure 3: A) Crystal structure of BpGH16C (Bacple_01703). Chain A of asymmetric unit is shown in cartoon representation and colored in spectrum from N to C terminal. The HEPES molecule in the active site is shown as stick representation. The N- and C-terminals are marked, and the strands are labeled in their order in the sequence. **B**) The substrate binding site in the mutant BpGH16C-E145L complexed with the D-β-Gal-(1→4)-6O-sulfate-L-Gal disaccharide end-product. The semitransparent surface of the protein is colored by the electrostatic potential. The residues forming the site are shown in stick representation. The sulfate group from L-galactose 6 sulfate is docked into a positively charged pocket. **C**) The hydrogen bonds between the disaccharide in substrate binding site residues. Several H-bonds are bridged by water molecules (W). Glu145 (in blue and thicker bonds) from the native structure is superimposed on the Leu145 in the mutants. **D)** Structure comparison of BpGH16C (green), PorA (PDB id-3ILF) (magenta) and PorB (PDB id- 3JUU) (yellow) showing the active site region. The single mutation of Ser129/Thr137 (PorA/PorB) (sidechains shown as spheres) to Gly132 in BpGH16C provides space for accommodating a methyl group at C6 on L-galactose residue. **E**) A cros-section of the surface representation of the Gal sugar and the cavity in the binding site near Gly132. There is free space within the cavity sufficient to accommodate the methyl group of a methylated porphyran.

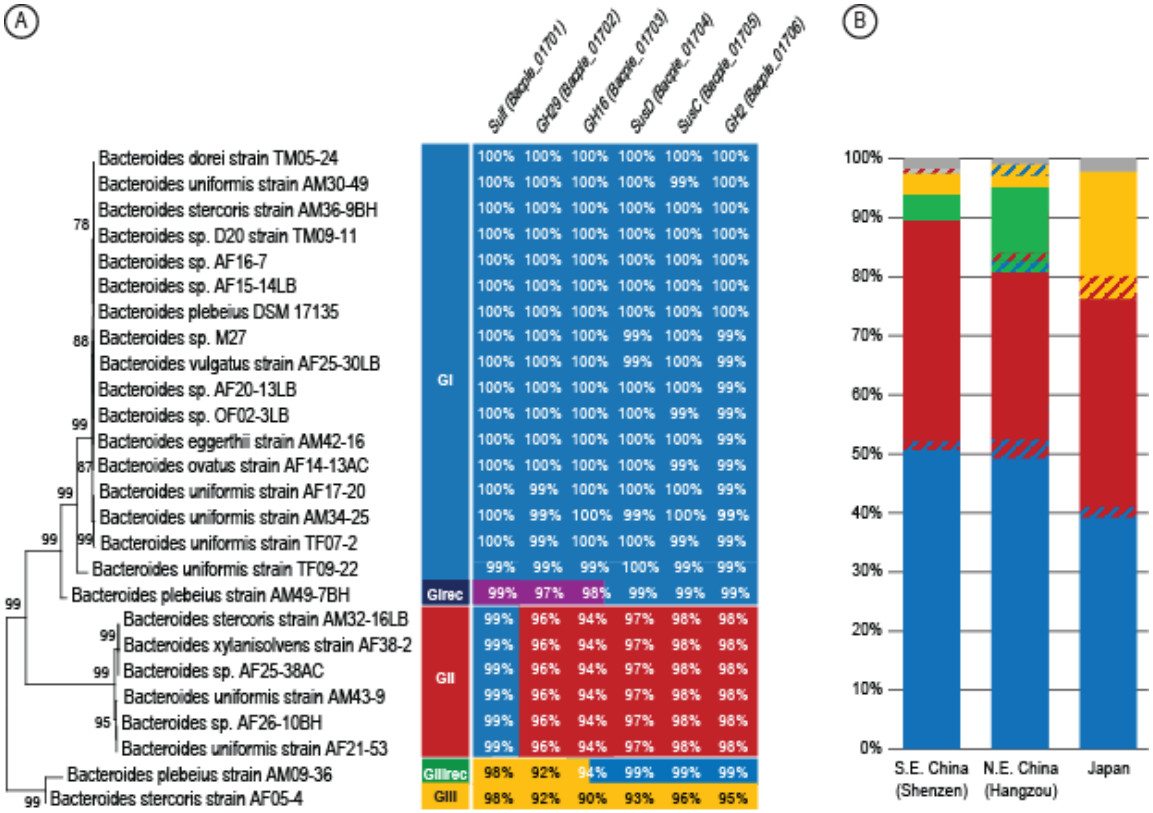

Figure 4: A) Phylogenetic tree calculated with the concatenated gene sequences *of PUL-PorB.* Sequence homology observed allowed to identify recombination sites and to create groups of homologous porphyran PUL. For each strain, the identity of the *PUL-PorB* genes was compared with the corresponding gene of *B. plebeius* DSM 17135 **B**) Distribution of the different group of porphyran PUL in human gut metagenome assembly of South east and North East Chinese, and Japanese populations.

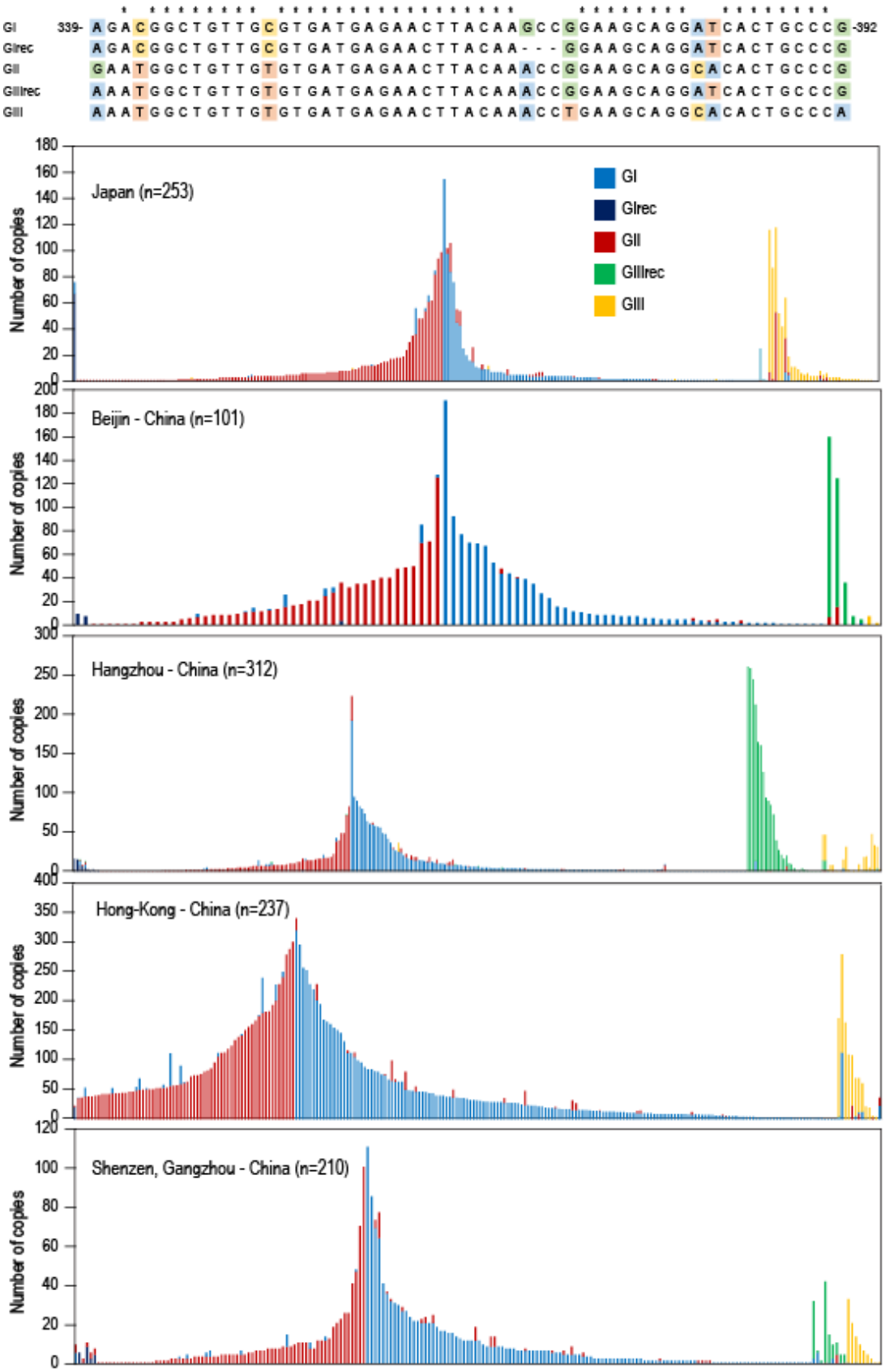

Figure 5: The 54 nucleotides sequence characteristic of different group of porphyran PUL (top) were Blastn against the short read sequencing of metagnomic data recorded on Chinese and Japanese population. The number of detected short reads are indicated for each individuals and were grouped as a function of the geographic location of the sampling.

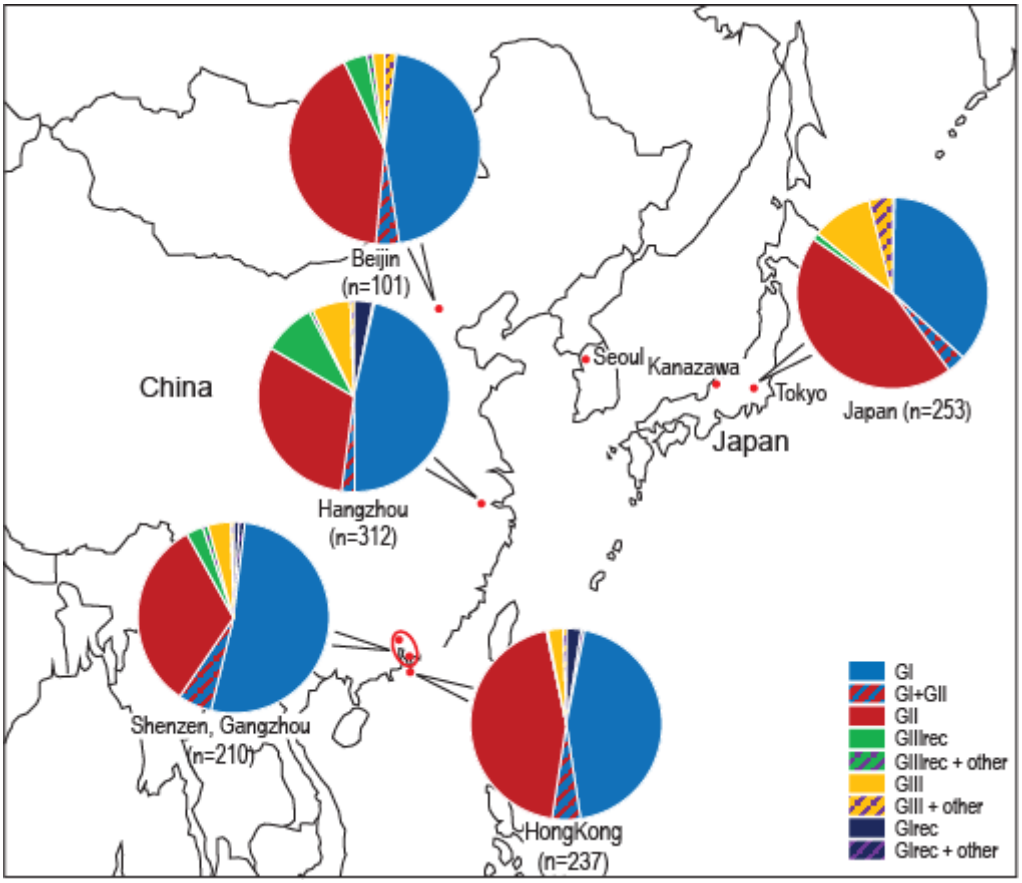

Figure 6: Distribution of the various porphyran PUL among Chinese and Japanese population. The pie charts derive from the histogram presented in Figure S8 obtained from the analyses of short read sequencing of Chinese and Japanese metagenomics data. For 80% of the individuals, only one group of porphyran PUL was observed. For the other individuals, a dominant group was considered when the number short-reads detected was, at least, twice more abundant than the other. In other situation, the co-occurrence of groups was reported.

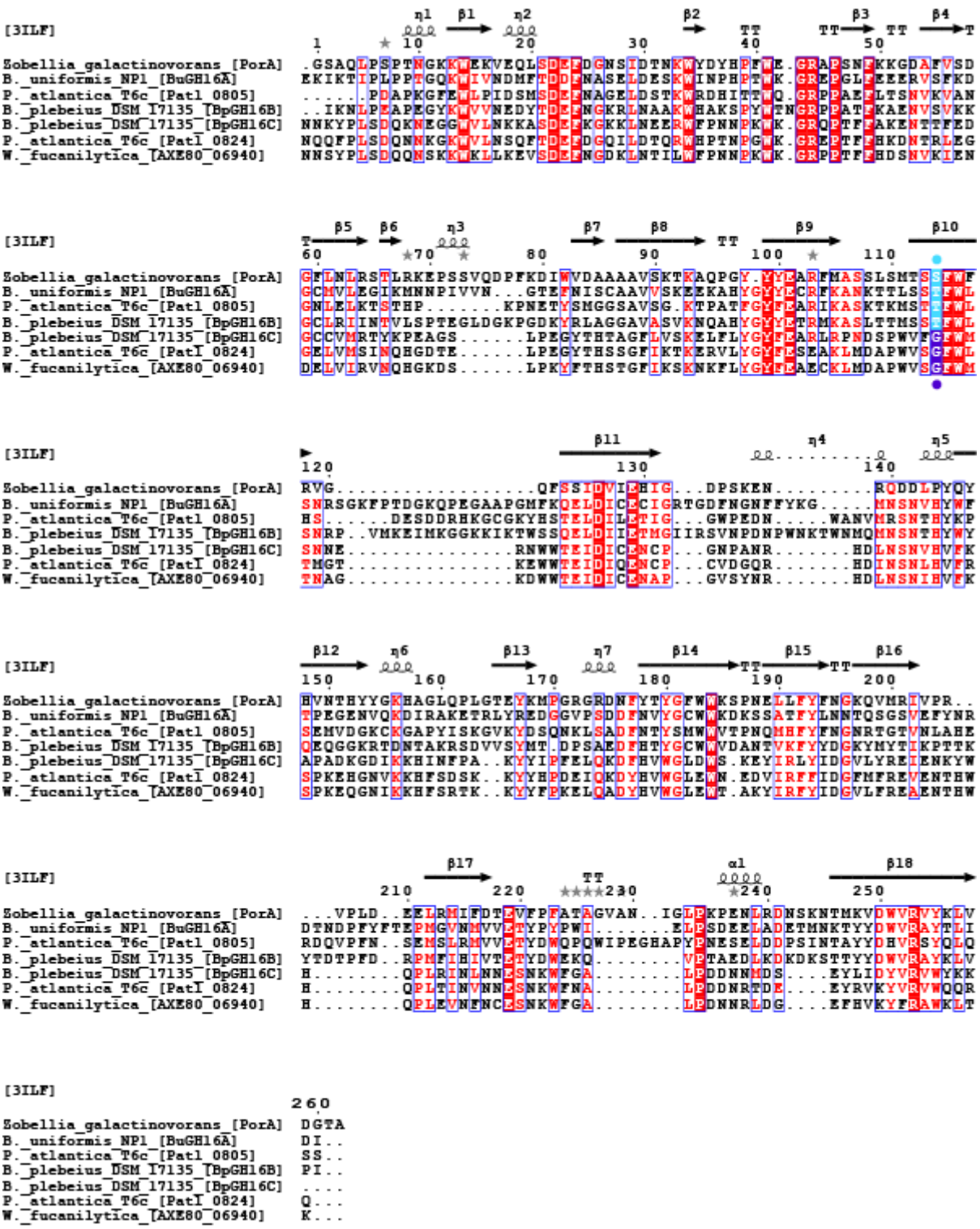

Figure S1: Structural alignment of the *B. plebieus* DSM 17135 BpGH16C 6-O- methyl-β-porphyranase compared with the other characterized 6-O- methyl-β-porphyranases (*P. atlantica* T6c, Patl_0824; *W. fucanilytica* AXE80_06940) and characterized β-porphyranases (*B. uniformis*_NP1, BuGH16A; *P. atlantica* T6c, Patl_0805; *B. plebeius* DSM 17135, BpGH16B). The blue circle indicates the serine/threonine amino acids observed in the active site of the β-porphyranases which are replaced by a glycine amino acid in 6-O- methyl-β-porphyranases allowing to accommodate the methyl group.

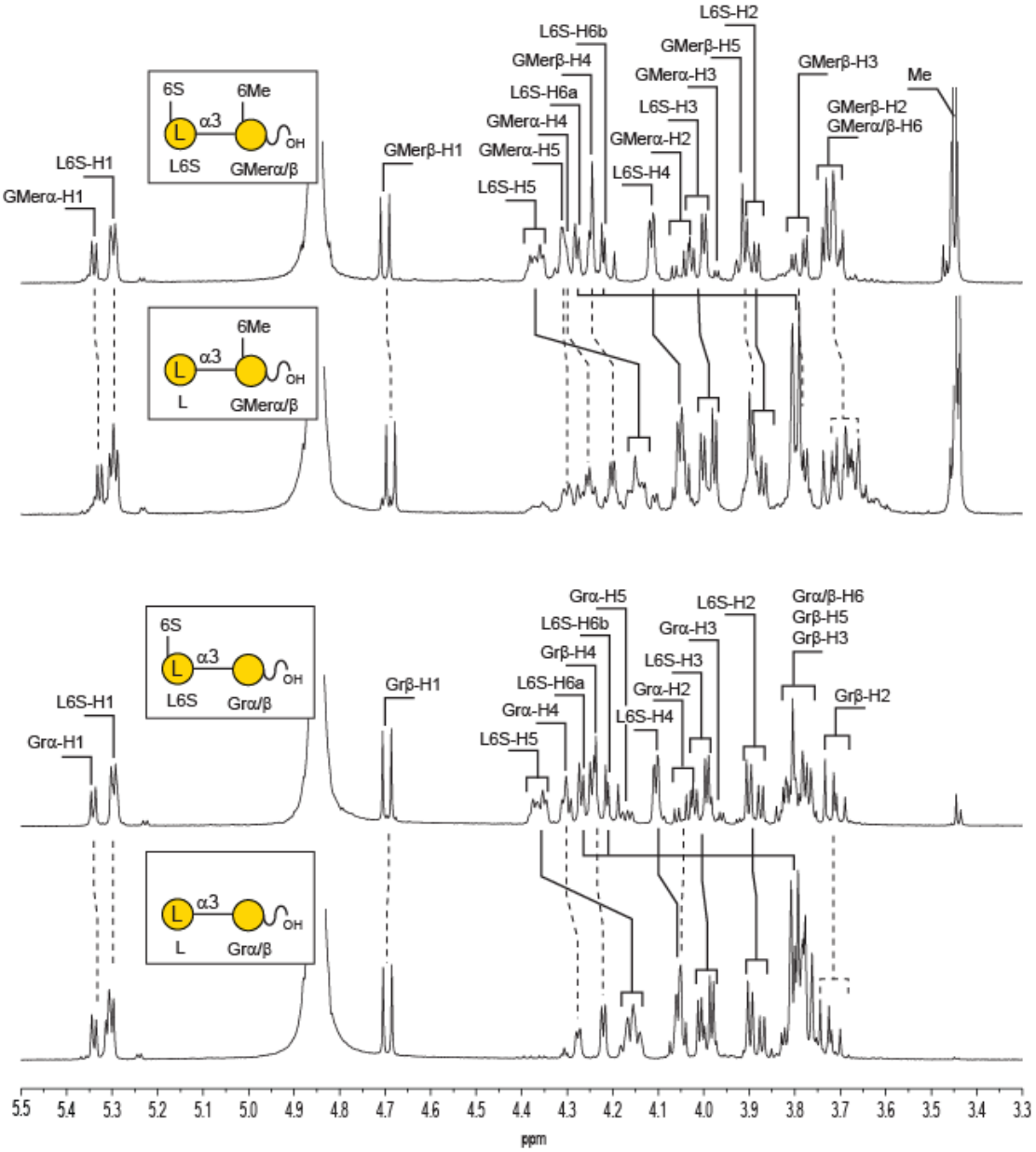

Figure S2: ^{1}H NMR recorded before and after incubation with the sulfatase BpS1_11 on purified disaccharides end-products of the BpGH16C. The occurrence of the methyl group located at the position 6 of the D-galactose didn't hinder the removal of the sulfate ester group of the L-galactose residue located at the non-reducing end.

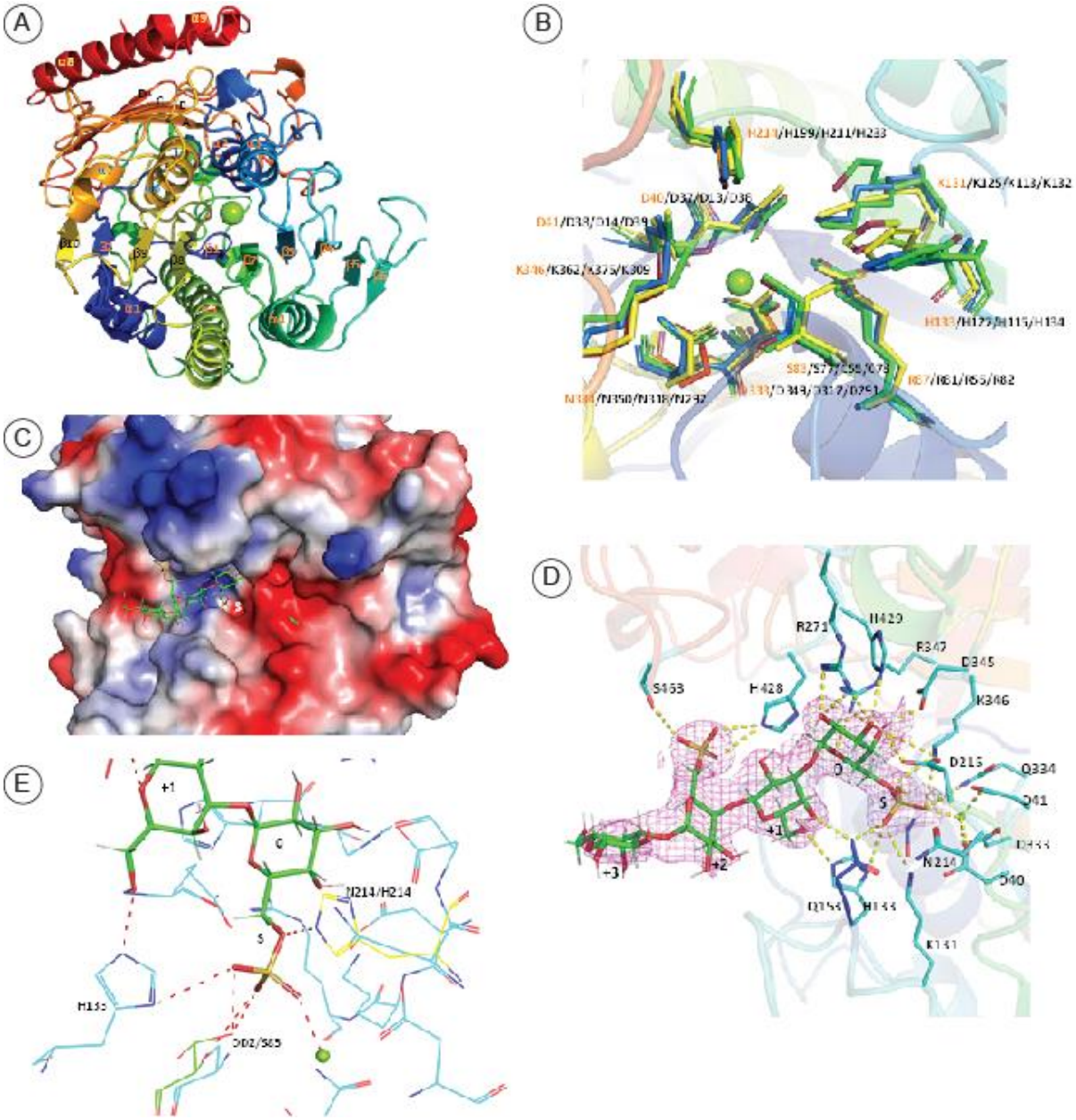

Figure S3: A) Structural details of Bacple_01701. Schematic representation of Bacple_01701 shown in cartoon representation and colored from N to C terminus. Calcium ion is shown as green sphere. **B**) Conserved active site residues from S1 family sulfatases. The residues are shown in stick representation in the order Bacple_01701/ 5G2V/1HDH/6BIA. **C**) Surface representation of H214 mutant with the tetrasaccharide, showing pocket like architecture of active site. The 6S sulfate subsite is labelled as S. **D**) Complex structure of H214N mutant. The electron density (2fo-fc) map for the sugar contoured at 1.5σ. Sugars numbered from 0 to +3 and 6S sulfate subsite labelled as S. Residues at H bonding distance to the substrate are shown in stick representation. **E**) Structural comparison of H214 mutant with 1HDH. The residues involved in catalysis S83 (overlays with DDZ from 1HDH), H133, H214 (this is the mutated residue) are shown in stick representation.

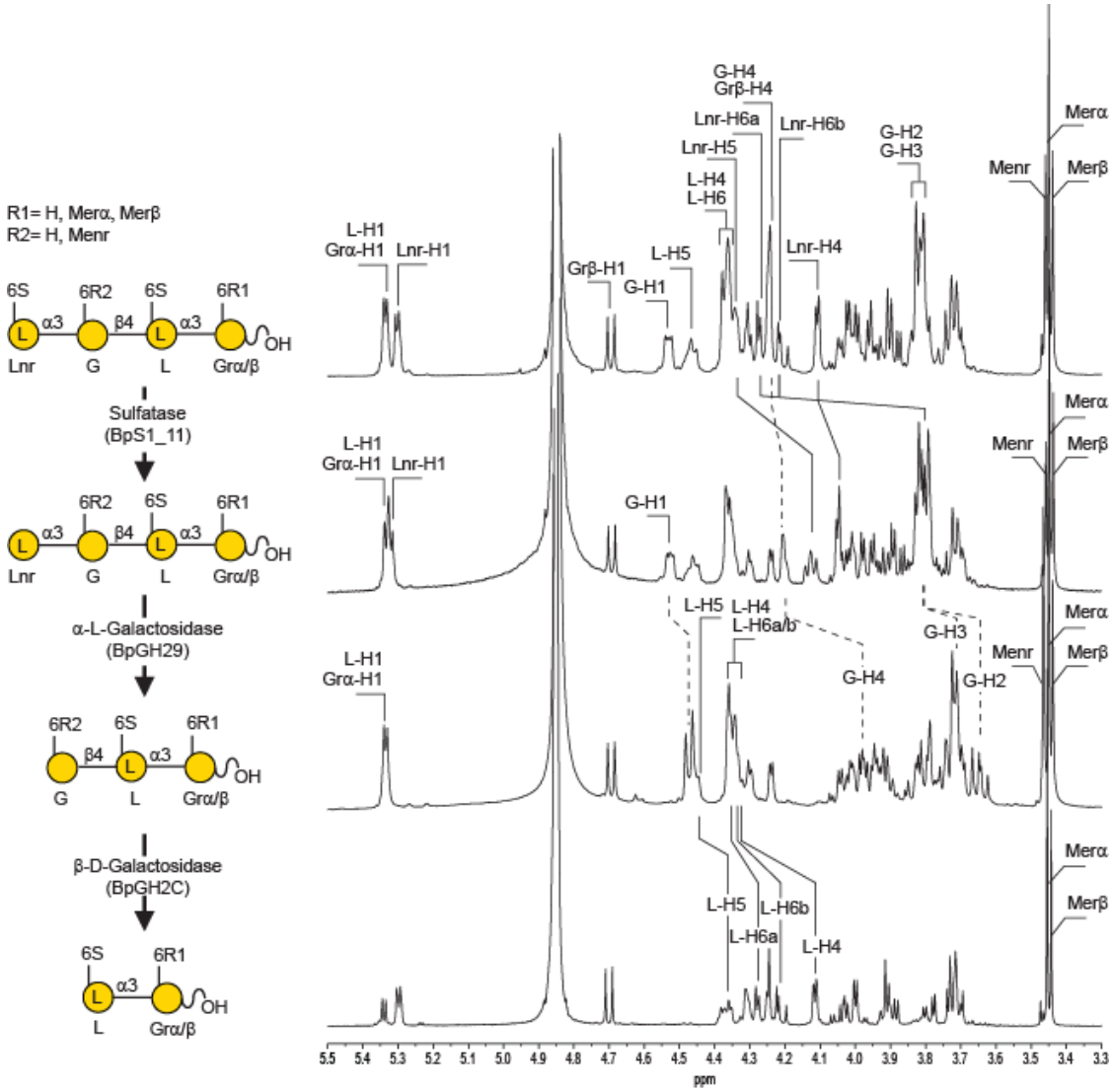

Figure S4: ^{1}H NMR recorded after sequential incubation of the methylated tetrasaccharide end-products of the 6-O-methyl-porphyranase BpGH16C with the sulfatase BpS1_11 followed by the α-L-galactosidase BpGH29 and the D-galactosidase BpGH2C.

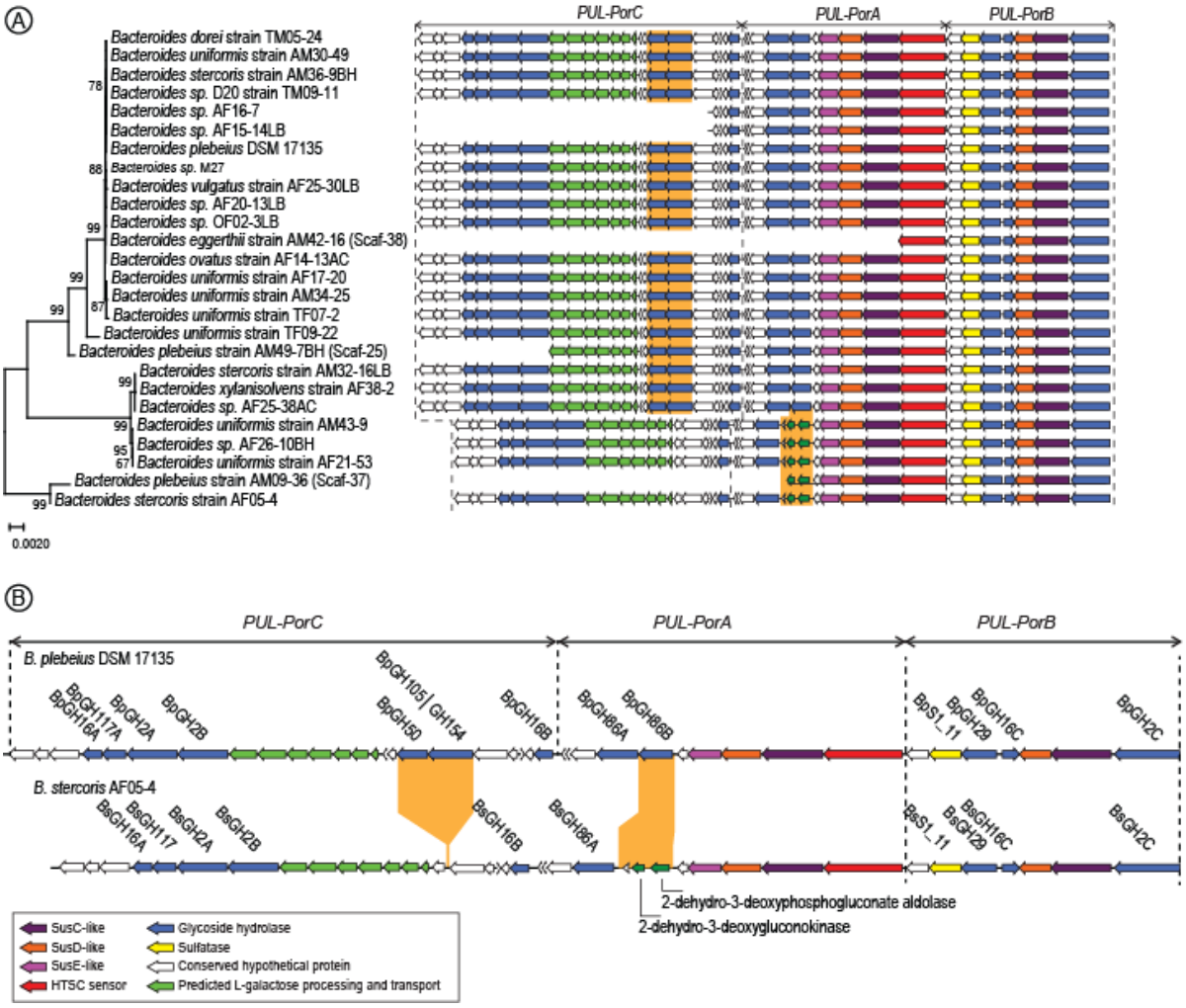

Figure S5: Gene organization of the porphyran degradation system of *B. plebieus* compared with other identified human gut bacteria. **A**) Phylogenetic tree was calculated using concaneted gene sequences of *PUL-PorB*. PUL organizations of each strains were indicated based on the available sequencing data. **B**) Comparison of the two main gene organizations of the porphyran PUL observed in the set of bacteria investigated highlighting deletion and recombination events.

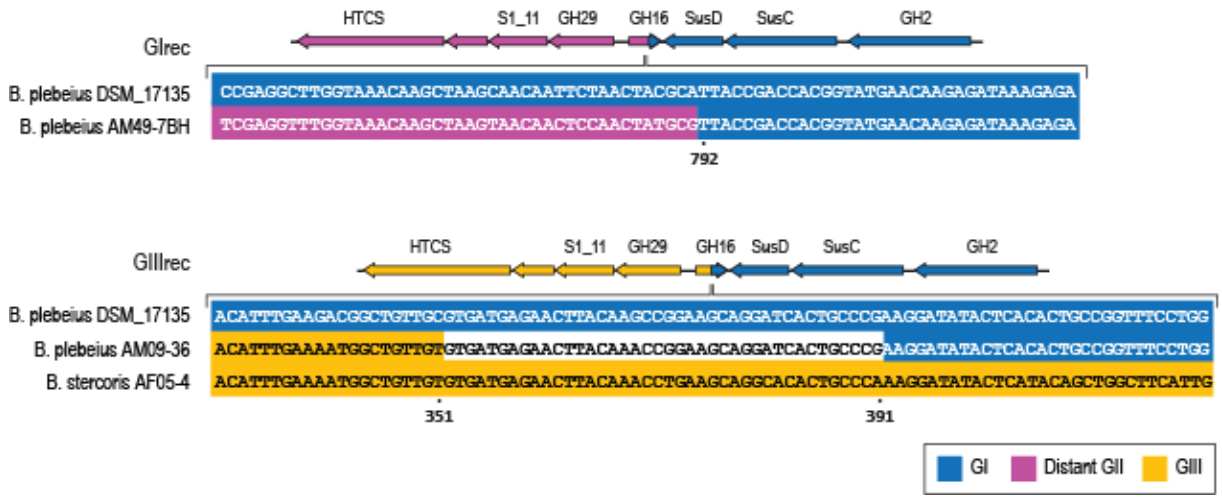

Figure S6: Recombination sites observed in the GH16 of PUL-PorB of the group GI/distant GII giving GIrec and the groups GI/GIII giving GIIIrec.

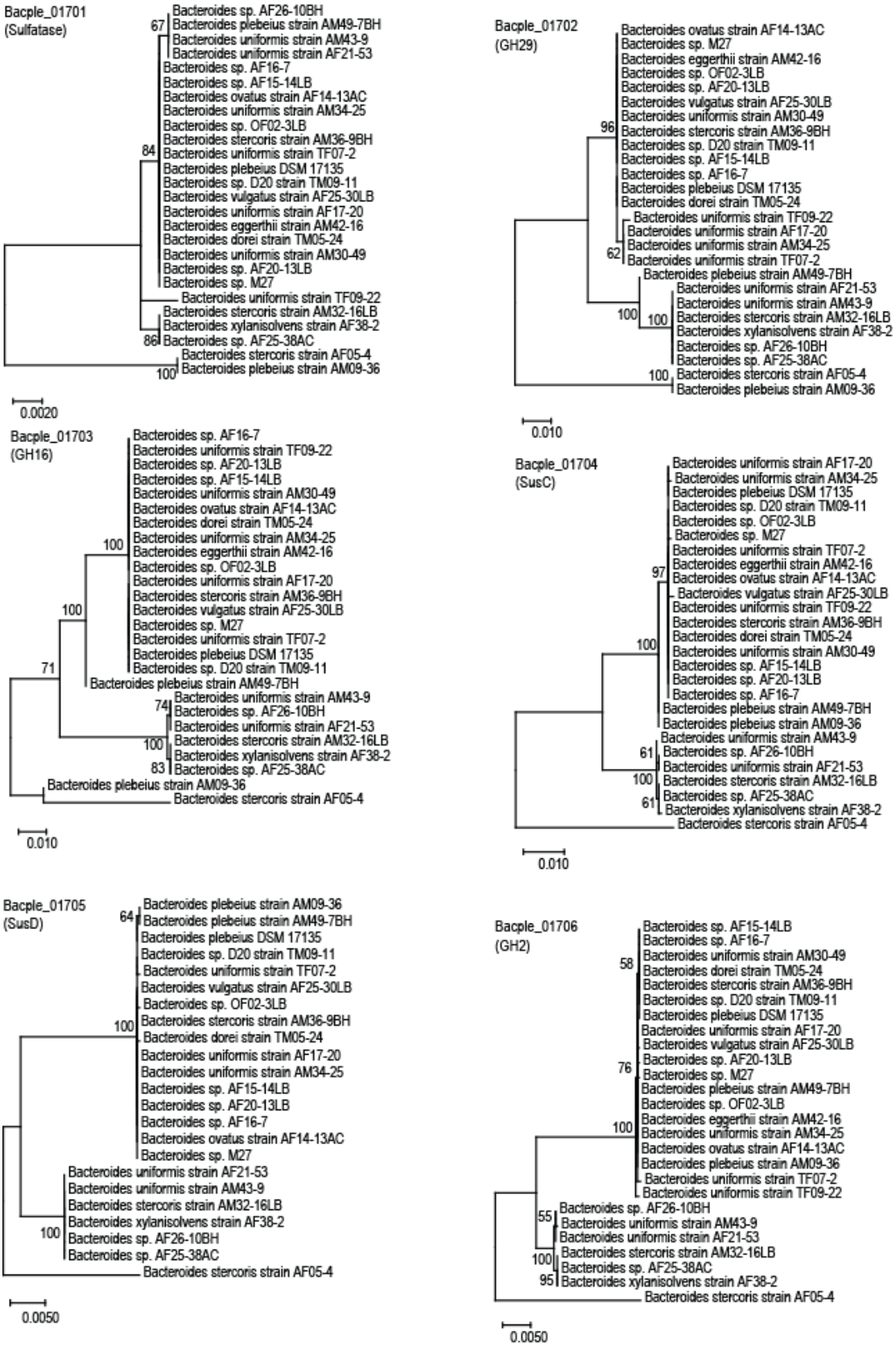

Figure S7: Phylogenetic trees calculated for all gene sequences of the members of the *PUL-PorB*.

Table S1: List of the bacterial strains which assembled genome contained encoding genes of the porphyran degradation system. Bacteroides and Phocaeicola are basonyms. The name of the strains were reported as reported in databank.

GenBank assembly	Assembly level	Biosamples	Strains
GCA_003466465.1	Scaffold	SAMN09736981	*Phocaeicola dorei* TM05-24
GCA_003467285.1	Scaffold	SAMN09736544	*Bacteroides eggerthii* AM42-16
GCA_003465275.1	Scaffold	SAMN09734245	*Bacteroides ovatus* AF14-13AC
GCA_000187895.1	Scaffold	SAMN00000028	*Bacteroides plebeius* DSM 17135
GCA_003472665.1	Scaffold	SAMN09734757	*Phocaeicola plebeius* AM09-36
GCA_003463985.1	Scaffold	SAMN09736616	*Phocaeicola plebeius* AM49-7BH
GCA_003465625.1	Scaffold	SAMN09734188	*Bacteroides stercoris* AF05-4
GCA_003468685.1	Scaffold	SAMN09736438	*Bacteroides stercoris* AM32-16LB
GCA_003468355.1	Scaffold	SAMN09736487	*Bacteroides stercoris* AM36-9BH
GCA_003460085.1	Scaffold	SAMN09734330	*Bacteroides uniformis* AF17-20
GCA_003459325.1	Scaffold	SAMN09734434	*Bacteroides uniformis* AF21-53
GCA_003468865.1	Scaffold	SAMN09736420	*Bacteroides uniformis* AM30-49
GCA_003468465.1	Scaffold	SAMN09736465	*Bacteroides uniformis* AM34-25
GCA_003466885.1	Scaffold	SAMN09736574	*Bacteroides uniformis* AM43-9
GCA_003437205.1	Scaffold	SAMN09736902	*Bacteroides uniformis* TF07-2
GCA_003437605.1	Scaffold	SAMN09736925	*Bacteroides uniformis* TF09-22
GCA_003458385.1	Scaffold	SAMN09734500	*Phocaeicola vulgatus* AF25-30LB
GCA_003474645.1	Scaffold	SAMN09734674	*Bacteroides xylanisolvens* AF38-2
GCA_003603275.1	Scaffold	SAMN09734277	*Bacteroides* sp. AF15-14LB
GCA_003603315.1	Scaffold	SAMN09734319	*Bacteroides* sp. AF16-7
GCA_003603065.1	Scaffold	SAMN09734409	*Bacteroides* sp. AF20-13LB
GCA_003602995.1	Scaffold	SAMN09734502	*Bacteroides* sp. AF25-38AC
GCA_003602945.1	Scaffold	SAMN09734509	*Bacteroides* sp. AF26-10BH
GCA_003436255.1	Scaffold	SAMN09737012	*Bacteroides* sp. D20 TM09-11
GCA_014287385.1	Contig	SAMN15806883	*Bacteroides* sp. M27

Table S2: List of metagenomics assembly projects probed with the genes encoding PUL-PorB and SRA data explored with the 50 nucleotides probes.

Location of sampling		Project (SRA)	Bio-samples	50 n probes		Project (assembly)	Bio-samples	PUL porB	
				hits	%			hits	%
CHINA	Hangzhou	PRJNA375935	211	125	46	PRJEB24527	97	39	40
	Hangzhou	PRJNA356102	168	106	63	PRJEB29103	168	72	42
	Hangzhou	PRJNA353560	97	54	56	PRJEB26158	97	40	41
	Shenzen	PRJNA422434	370	163	44	PRJEB30046	370	135	36
	Shenzen	PRJEB12669	20	7	35	PRJEB26908	18	2	11
	Many places*	PRJNA356225	150	68	45	PRJEB26167	140	32	23
	Guangzhou	PRJEB18755	124	40	32				
	Hangzhou	PRJNA505228	50	27	54				
	Beijin	PRJNA401977	145	100	69				
	Hunan region	PRJNA349463	40	8	20				
	Hong-Kong	PRJNA557323	565	237	42				
JAPAN	Tokyo	PRJDB3601	255	202	79	PRJEB26092	254	50	20
	Kanazawa	PRJNA517801	68	3	4				
	Tokyo	PRJDB7378	50	48	96				
INDIA		PRJNA397112	110			PRJEB33179	110	0	0
BANGLADESH	Dhaka	PRJEB9150	260			PRJEB22359	7	0	0
USA	Lincoln	PRJNA324129	107			PRJEB26490	63	1	0.03
	Seattle	PRJNA215106	22			PRJEB22365	22	0	0
	Saint Louis	PRJNA301903	422			PRJEB24849	384	0	0
	HMP	PRJNA48479	717			PRJEB22283	717	5	0.7
	Standford	PRJNA288562	29			PRJNA288562	29	1	0.03
	Chicago	PRJNA353655	10			PRJEB26155	10	3	30
	Cheyenne tribe	PRJNA299502	37			PRJEB22552	37	0	0
PERU		PRJNA268964	47			PRJEB24847	47	0	0
TANZANIA	Hazda	PRJNA278393	38			PREJB22391	38	0	0
MADAGASCAR		PRJNA485056	112			PREJB33766	112	0	0