Published online 27 October 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D677–D683***
doi: 10.1093/nar/gkx1022

PULDB: the expanded database of Polysaccharide Utilization Loci

Nicolas Terrapon[1,2,*], Vincent Lombard[1,2], Élodie Drula[1,2], Pascal Lapébie[1,2], Saad Al-Masaudi[3], Harry J. Gilbert[4] and Bernard Henrissat[1,2,3,*]

[1]Architecture et Fonction des Macromolécules Biologiques, CNRS, Aix-Marseille Université, F-13288 Marseille, France, [2]USC1408 Architecture et Fonction des Macromolécules Biologiques, Institut National de la Recherche Agronomique, F-13288 Marseille, France, [3]Department of Biological Sciences, King Abdulaziz University, 23218 Jeddah, Saudi Arabia and [4]Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne NE2 4HH, UK

Received September 17, 2017; Revised October 16, 2017; Editorial Decision October 17, 2017; Accepted October 25, 2017

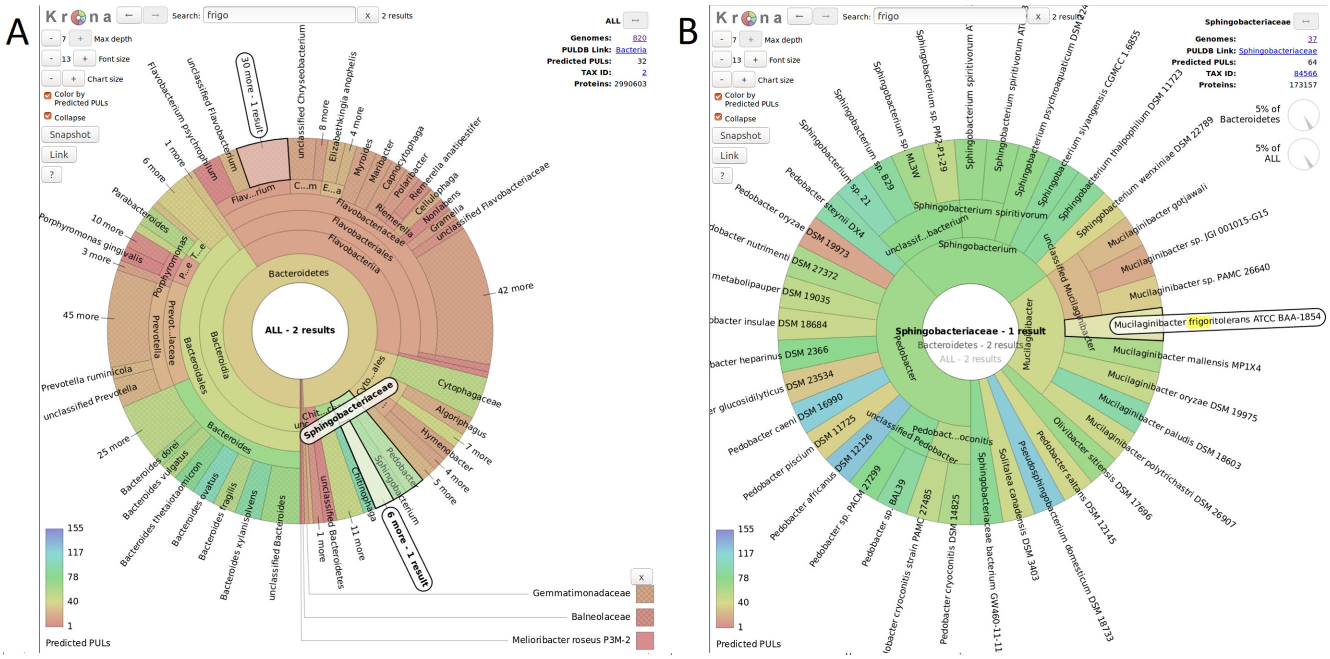
ABSTRACT

The Polysaccharide Utilization Loci (PUL) database was launched in 2015 to present PUL predictions in ∼70 Bacteroidetes species isolated from the human gastrointestinal tract, as well as PULs derived from the experimental data reported in the literature. In 2018 PULDB offers access to 820 genomes, sampled from various environments and covering a much wider taxonomical range. A Krona dynamic chart was set up to facilitate browsing through taxonomy. Literature surveys now allows the presentation of the most recent (i) PUL repertoires deduced from RNAseq large-scale experiments, (ii) PULs that have been subjected to in-depth biochemical analysis and (iii) new Carbohydrate-Active enzyme (CAZyme) families that contributed to the refinement of PUL predictions. To improve PUL visualization and genome browsing, the previous annotation of genes encoding CAZymes, regulators, integrases and SusCD has now been expanded to include functionally relevant protein families whose genes are significantly found in the vicinity of PULs: sulfatases, proteases, ROK repressors, epimerases and ATP-Binding Cassette and Major Facilitator Superfamily transporters. To cope with cases where *susCD* may be absent due to incomplete assemblies/split PULs, we present 'CAZyme cluster' predictions. Finally, a PUL alignment tool, operating on the tagged families instead of amino-acid sequences, was integrated to retrieve PULs similar to a query of interest. The updated PULDB website is accessible at www.cazy.org/PULDB_new/

INTRODUCTION

Polysaccharides constitute the main source of carbon for most organisms on Earth. Because of their enormous structural diversity, polysaccharide deconstruction requires the concerted action of large numbers of specific enzymes. While most bacteria break down polysaccharides by exporting their carbohydrate-active enzymes (CAZymes) into the extracellular milieu and import the simple sugars produced, an inventive solution operates in Gram-negative bacteria of the Bacteroidetes phylum. The genomes of these bacteria feature Polysaccharide Utilization Loci, or PULs. A PUL comprises a single genomic locus that encodes the necessary proteins to bind a given polysaccharide at the cell surface, to perform an initial cleavage to large oligosaccharides, to import these oligosaccharides in the periplasmic space, to complete the degradation into monosaccharides and to regulate PUL gene expression. Some Bacteroidetes species contains up to 100 PULs with almost 20% of their genome dedicated to these systems (1), explaining their evolutionary success as primary glycan degraders in the human gut microbiota (2). Bacteroidetes are found in almost all environments, and the last decade has seen a continuous acceleration of published PUL analyses, notably by RNAseq experiments and in-depth biochemistry. To facilitate individual PUL analysis, in 2015 we launched PULDB to present PULs predicted solely from genome sequences along with those reported in the literature (3). The principle of the PUL prediction is to start from every *susCD*-like gene pair, and then to extend PUL boundaries to operonic genes (based on intergenetic distances between genes on the same strand (4)) and to more distant regulators and CAZyme coding genes which catalyze polysaccharide breakdown. While we previously mainly focused on the algorithm and presented a limited number of genomes with a recognized bias towards human gut species/strains, we present here a major update of PULDB. This release includes a 10-fold increase in ana-

[*]To whom correspondence should be addressed. Tel: +33 491 825 587; Fax: +33 491 266 720; Email: nicolas.terrapon@afmb.univ-mrs.fr
Correspondence may also be addressed to Bernard Henrissat. Tel: +33 491 825 587; Fax: +33 491 266 720; Email: bernard.henrissat@afmb.univ-mrs.fr


Search: frigo
2 results

ALL
Genomes: 820
PULDB Link: Bacteria
Predicted PULs: 32
TAX ID: 2
Proteins: 2990603

Sphingobacteriaceae
Genomes: 37
PULDB Link: Sphingobacteriaceae
Predicted PULs: 64
TAX ID: 84566
Proteins: 173157

Figure 1. Krona multilayered pie-charts of taxonomy in PULDB. The top-left corner includes web-browser classical features (text search area and buttons for browsing back and forward), and display features (depth of the taxonomy, font and chart sizes). The bottom-left corner displays the color scale that represents the number of PULs per species (averaged in ancestral nodes). The top-right corner indicates the selected taxonomic level and its relative information: the number of species (with a link to the listing), a link to PULDB to visualize all PULs for this taxon, a link to the NCBI taxonomy, etc. (**A**) Initial display of the most general taxonomic level, labeled ALL at the center, with a search for the character string 'frigo' highlighting the taxa having a positive result. (**B**) Display of the Sphingobacteriaceae level, resulting from a zoom-in by double-clicking on the 'Sphingobacteriaceae' area in (A) chart. Going back to (A) or intermediary levels is possible through the lineage links at the center.

lyzed genomes that offers a much deeper coverage of the Bacteroidetes phylum and different environments. A tool has been integrated to the web interface to facilitate taxonomy browsing in PULDB. Also this release updates to the most recent literature-derived PULs and CAZyme families. Additional protein families relevant in a PUL context are now displayed and used in a PUL aligner that allows the user to retrieve the most conserved modular PUL organizations.

10-FOLD INCREASE IN CAZy-ANALYZED SPECIES

In order to achieve a >10-fold increase in PULDB, we analyzed 820 complete genome sequences (∼3 million genes) mostly of the Bacteroidetes phylum downloaded from JGI (http://genome.jgi.doe.gov/) and NCBI (https://www.ncbi.nlm.nih.gov/nuccore) servers. Our PUL prediction procedure relies on genomic data but also requires the semi-manual expert annotation of CAZymes (5). We identified 153 202 CAZyme modules in the 820 genomes, mostly glycoside hydrolases (53%) and glycosyltransferases (31%), classified according to the sequence-based families that are described in the CAZy database. Then the 820 genomes were subjected to the PUL predictions as described earlier. Compared to the 2015 PULDB dataset (3), the new genome sampling expands far beyond the human gastrointestinal tract (now represented by ∼80 species), and notably includes 64 rumen gut species, as well as many bacterial species from soil or marine environments. The coverage of Bacteroidetes taxonomical diversity also drastically increased. The 2015 dataset almost exclusively consisted of species from the Bacteroidales order (70% belonging to the Bacteroides genus). In the current dataset, Bacteroidales only represents 40% (only half being from the Bacteroides genus), a proportion comparable to the Flavobacteriales order while three additional orders (Cytophagales, Sphingobacteriales and Chitinophagales) are now also presented. Moreover, the presence of the PUL fundamental *susCD* gene tandem now allows the prediction of PULs beyond the Bacteroidetes phylum, namely in the Gemmatimonadetes and Ignavibacteriae phyla (which group with Bacteroidetes in the FCB group), and also in the Balneolaeota phylum.

To facilitate navigation across the various taxonomical levels, and to identify species of interest, we implemented a new browsing tool in PULDB. We adapted the Krona multilayered pie-chart, introduced for metagenomics analysis (6), to represent the hierarchical aspects of the taxonomy (Figure 1). Implemented using the latest HTML5 and JavaScript interactive technology, Krona allows zooming in and out very efficiently and can be easily customized by the user for the desired taxonomic depth or font size, allowing the production of high-quality publication-ready pictures. It also offers text searches and improved navigation. We also added a color scale indicative of the number of PULs per genome (estimated for ancestral taxa by a simple arithmetic mean) which immediately offers an overview of the PUL diversity at the different taxonomical levels. Finally, in the upright part, where Krona provides statistics about genome for each taxa, we added several hyperlinks to the species list,

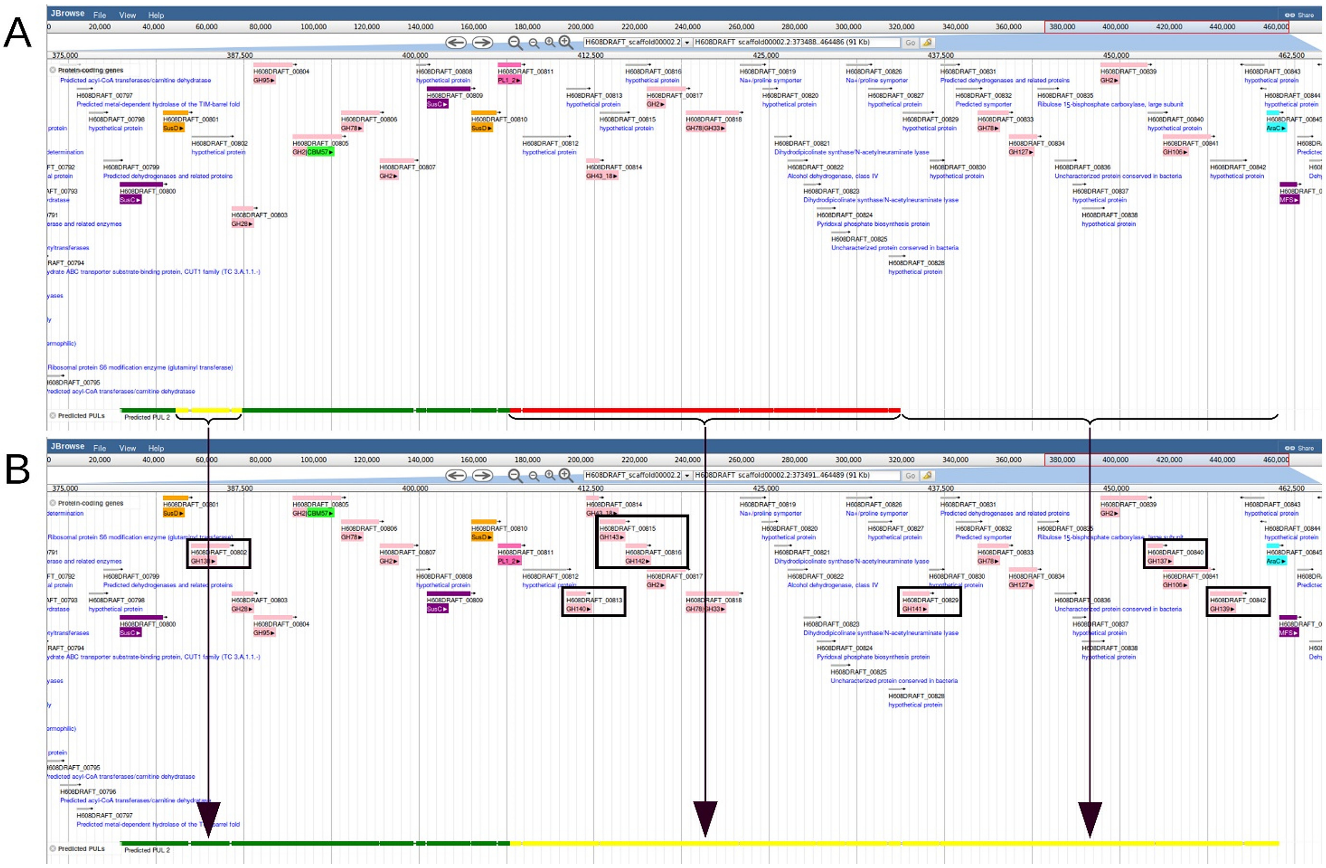


Figure 2. Example of the improved PUL predictions by the inclusion of recently created CAZyme families in the RGII PUL of *Terrimonas ferruginea DSM 30193*. Panel (**A**) displays the JBrowse view (35) of the region before the creation of families GH136-GH143 (16). The predicted PUL is depicted at the bottom of the panel by a green, yellow and red line, according to confidence levels as previously described (3). Panel (**B**) displays the same region with the genes belonging to these seven families now annotated and highlighted by black boxes. These annotations lead to a PUL prediction with improved confidence (left and middle arrows), and improved PUL boundaries (right arrow), compared to (A).

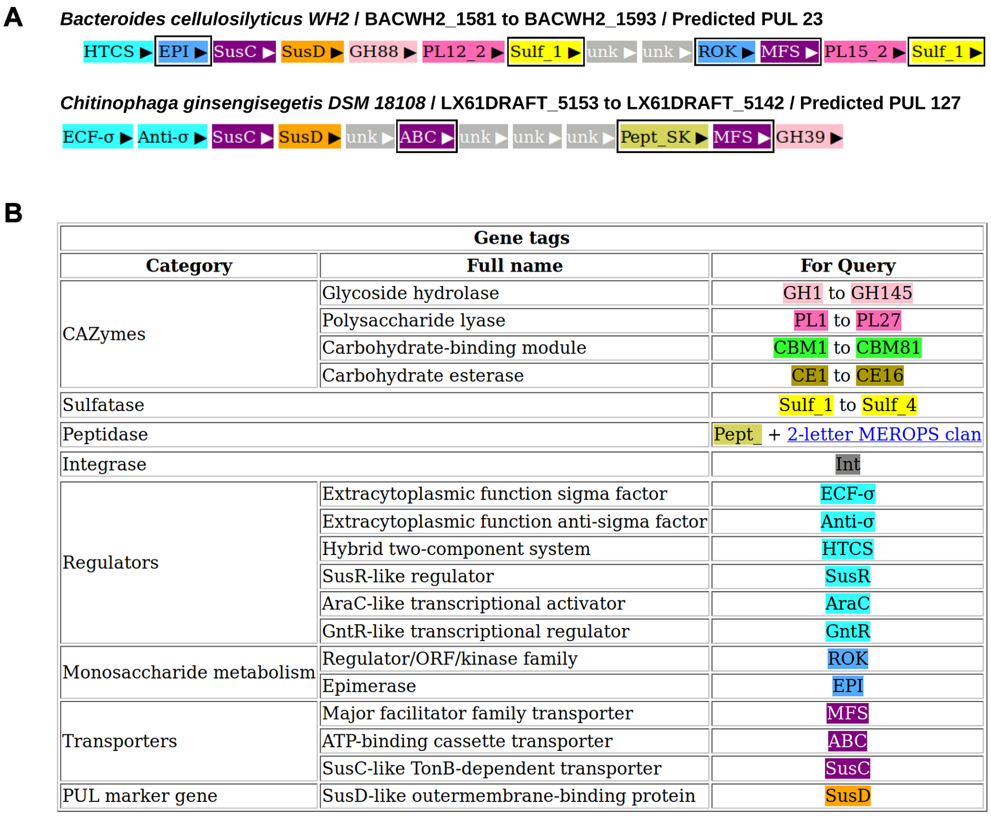
to the NCBI taxonomy and to PULDB predicted PULs in this group/species.

LITERATURE-DERIVED PULs, COGNATE SUBSTRATES AND NEW CAZymes FAMILIES

The study of polysaccharide degradation by PUL encoded systems is a highly active research field. A continuous literature survey enabled us to complete the PULDB data with *literature-derived PUL* data (previously called *experimentally-validated PULs*). Notably, recent high-throughput experiments led to the delineation of PULs in *Bacteroides cellulosilyticus WH2* (7), *Bacteroides thetaiotaomicron 7330* (8) and *Zobellia galactinovorans* (9). Attempts to define PUL boundaries in the absence of expression data were also reported in the genome publication of *Capnocytophaga canimorsus Cc5* (10). Moreover, several specific analyses have focused on the degradation of defined polysaccharides by their corresponding PULs, including plant (fructan (11), pectin (12), xylan (13,14), xyloglucan (15) and type II rhamnogalacturonan (RGII) (16)) and non-plant (α-mannan (17), galactomannan (18), 1,6-β-glucan (19), mucin (20), sialoglycoconjugates (21), N-glycan (17,22–24), heparin and heparan sulfate (25), chitin (26), alginate and laminarin (27)) polysaccharides. To facilitate the retrieval of characterized PULs by their cognate substrate, a new field appears in the PULDB homepage, to search for a given character substring within the PUL substrate labels. Finally, the recent RGII publication notably reported the biochemical characterization of seven new glycoside hydrolase families that were immediately added to the CAZy database, designated GH137 to GH143. Similarly, other publications led to the creation of new CAZyme families: GH136, GH144, GH145, PL24 to PL27 (28–34). All new CAZy families have also been added to PULDB. As a consequence, the PUL predictions are improved by these new families, which allow refinement of both PUL boundaries and prediction confidence, as illustrated with the Jbrowse view (35) of the homologous RGII-PUL in *Terrimonas ferruginea DSM 30193* (Figure 2).

ADDITIONAL DISPLAY OF SULFATASES, PROTEASES, EPIMERASES, ROKs AND TRANSPORTERS

In PULDB, simplified representations of PULs are proposed as trains whose wagons, the constitutive proteins, are colored/tagged if their protein function is relevant in the PUL context. We initially focused on SusC outer-membrane transporter (purple), SusD outer-membrane


A
Bacteroides cellulosilyticus WH2 / BACWH2_1581 to BACWH2_1593 / Predicted PUL 23
HTCS ▶ EPI ▶ SusC ▶ SusD ▶ GH88 ▶ PL12_2 ▶ Sulf_1 ▶ unk ▶ unk ▶ ROK ▶ MFS ▶ PL15_2 ▶ Sulf_1 ▶
Chitinophaga ginsengisegetis DSM 18108 / LX61DRAFT_5153 to LX61DRAFT_5142 / Predicted PUL 127
ECF-σ ▶ Anti-σ ▶ SusC ▶ SusD ▶ unk ▶ ABC ▶ unk ▶ unk ▶ unk ▶ Pept_SK ▶ MFS ▶ GH39 ▶
B

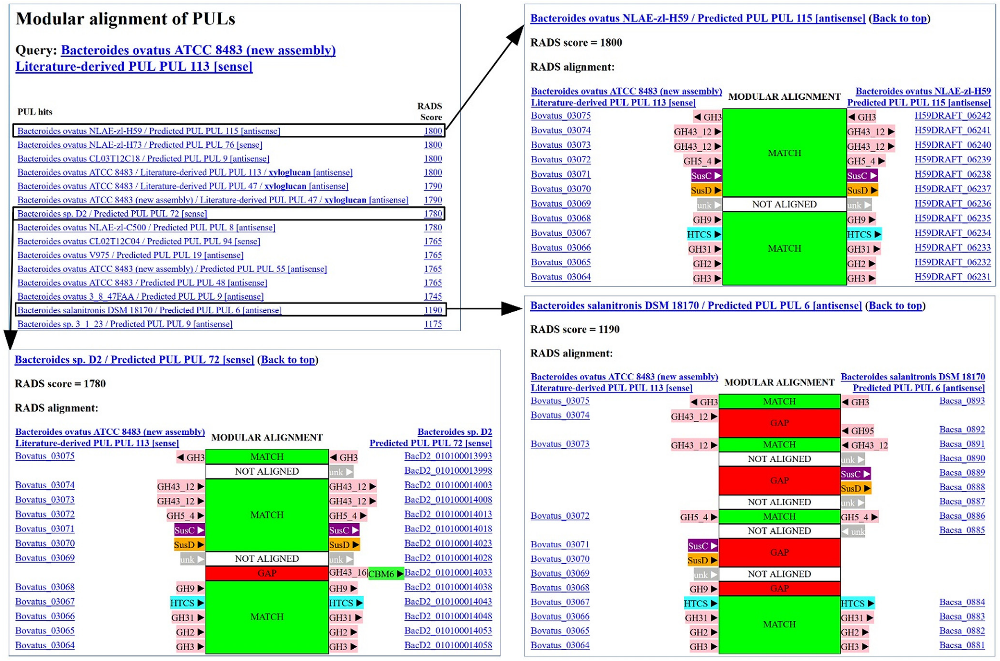
Gene tags		
Category	**Full name**	**For Query**
CAZymes	Glycoside hydrolase	GH1 to GH145
	Polysaccharide lyase	PL1 to PL27
	Carbohydrate-binding module	CBM1 to CBM81
	Carbohydrate esterase	CE1 to CE16
Sulfatase		Sulf_1 to Sulf_4
Peptidase		Pept_ + 2-letter MEROPS clan
Integrase		Int
Regulators	Extracytoplasmic function sigma factor	ECF-σ
	Extracytoplasmic function anti-sigma factor	Anti-σ
	Hybrid two-component system	HTCS
	SusR-like regulator	SusR
	AraC-like transcriptional activator	AraC
	GntR-like transcriptional regulator	GntR
Monosaccharide metabolism	Regulator/ORF/kinase family	ROK
	Epimerase	EPI
Transporters	Major facilitator family transporter	MFS
	ATP-binding cassette transporter	ABC
	SusC-like TonB-dependent transporter	SusC
PUL marker gene	SusD-like outermembrane-binding protein	SusD

Figure 3. Module tags in PULDB. (**A**) Examples of predicted PULs including the newly tagged modules (highlighted in black boxes). (**B**) Complete list of tagged modules which can be searched and displayed in PULDB (listed at www.cazy.org/PULDB/tags.html).

binding proteins (orange), several regulators (light blue), integrases which sometimes join adjacent PULs (dark gray) and CAZyme families (mainly glycoside hydrolases in light pink, polysaccharide lyases in dark pink, carbohydrate-binding modules in green, carbohydrate esterases in brown). All other proteins remained tagged as 'unknown' (light gray). To increase readability of these PUL representations, we searched additional protein families with relevant function in PULs, based on (i) the literature, (ii) over-representation in PUL contexts and (iii) reliability of Pfam domain annotation (36). These new families are now tagged/colored in the new PULDB release. The most important accessory enzymes that directly assist polysaccharide degradation are the sulfatases, which remove sulfate groups from algal and mammalian-host glycans (25,37,38). Sulfatases now appear colored in yellow in PULDB and are labeled according to their SulfAtlas family classification (39). Proteins in the Major Facilitator Superfamily (MFS) are inner membrane transporters that participate in carbohydrate metabolism after polysaccharide depolymerization (40). Their presence in the vicinity of PULs and their participation in species growth have been demonstrated (41). MFS are thus colored in purple in PULDB, like SusC transporters, as well as ATP-Binding Cassette transporters. Even though PULDB has not been designed to annotate carbohydrate (monosaccharide) metabolism, in which a large variety of protein functions are involved, we intend to provide users with some indicators that several 'unknown' genes in a given PUL may not contribute to polysaccharide deconstruction. Thus, we colored in light blue and tagged domains of the ROK family (Repressors, ORFs and Kinases), and as well as some epimerases (42,43) that are frequently found in PULs. Finally, proteases have been shown to appear in some operons with *susCD* genes and to participate to the degradation of non-glycan substrates (20), raising the question of the extension the PUL paradigm beyond glycans. The observation of their high frequency in some PULs without CAZyme genes, motivates the integration of proteases in PULDB (gold-colored), labeled with the clan information of the MEROPS classification (44). All tags that can be searched and displayed in PULDB are shown in Figure 3, and are available at www.cazy.org/PULDB/tags.html.

CAZyme CLUSTERS

While most PULs resemble simple operonic systems, some substrates have been shown to activate the concerted action of several PULs, e.g. RGII (16), and sometimes a PUL and an additional gene cluster devoid of *susCD* genes, thus failing to fulfill the standard PUL paradigm. This was exemplified by the xylan degradation system of *Bacteroides xylanisolvens* (26). Indeed, when the complexity of the substrate increases, more enzymes are required for its breakdown and thus a 'longer' PUL needs to be maintained. This represents a challenge for bacteria to constrain all necessary enzymes within a single locus/regulatory system. Comparative genomics analysis of homologous PULs for RGII breakdown (16), the most complex known polysaccharide,



Modular alignment of PULs

Query: Bacteroides ovatus ATCC 8483 (new assembly)
Literature-derived PUL PUL 113 [sense]

PUL hits	RADS Score
Bacteroides ovatus NLAE-zl-H59 / Predicted PUL PUL 115 [antisense]	1800
Bacteroides ovatus NLAE-zl-H73 / Predicted PUL PUL 76 [sense]	1800
Bacteroides ovatus CL03T12C18 / Predicted PUL PUL 9 [antisense]	1800
Bacteroides ovatus ATCC 8483 / Literature-derived PUL PUL 113 / **xyloglucan** [antisense]	1800
Bacteroides ovatus ATCC 8483 / Literature-derived PUL PUL 47 / **xyloglucan** [antisense]	1790
Bacteroides ovatus ATCC 8483 (new assembly) / Literature-derived PUL PUL 47 / **xyloglucan** [antisense]	1790
Bacteroides sp. D2 / Predicted PUL PUL 72 [sense]	1780
Bacteroides ovatus NLAE-zl-C500 / Predicted PUL PUL 8 [antisense]	1780
Bacteroides ovatus CL02T12C04 / Predicted PUL PUL 94 [sense]	1765
Bacteroides ovatus V975 / Predicted PUL PUL 19 [antisense]	1765
Bacteroides ovatus ATCC 8483 (new assembly) / Predicted PUL PUL 55 [antisense]	1765
Bacteroides ovatus ATCC 8483 / Predicted PUL PUL 48 [antisense]	1765
Bacteroides ovatus 3_8_47FAA / Predicted PUL PUL 9 [antisense]	1745
Bacteroides salanitronis DSM 18170 / Predicted PUL PUL 6 [antisense]	1190
Bacteroides sp. 3_1_23 / Predicted PUL PUL 9 [antisense]	1175

Bacteroides ovatus NLAE-zl-H59 / Predicted PUL PUL 115 [antisense] (Back to top)

RADS score = 1800

RADS alignment:

Bacteroides ovatus ATCC 8483 (new assembly) Literature-derived PUL PUL 113 [sense]		MODULAR ALIGNMENT		Bacteroides ovatus NLAE-zl-H59 Predicted PUL PUL 115 [antisense]
Bovatus_03075	◀ GH3	MATCH	◀ GH3	H59DRAFT_06242
Bovatus_03074	GH43_12 ▶		GH43_12 ▶	H59DRAFT_06241
Bovatus_03073	GH43_12 ▶		GH43_12 ▶	H59DRAFT_06240
Bovatus_03072	GH5_4 ▶		GH5_4 ▶	H59DRAFT_06239
Bovatus_03071	SusC ▶		SusC ▶	H59DRAFT_06238
Bovatus_03070	SusD ▶		SusD ▶	H59DRAFT_06237
Bovatus_03069	unk ▶	NOT ALIGNED	unk ▶	H59DRAFT_06236
Bovatus_03068	GH9 ▶	MATCH	GH9 ▶	H59DRAFT_06235
Bovatus_03067	HTCS ▶		HTCS ▶	H59DRAFT_06234
Bovatus_03066	GH31 ▶		GH31 ▶	H59DRAFT_06233
Bovatus_03065	GH2 ▶		GH2 ▶	H59DRAFT_06232
Bovatus_03064	GH3 ▶		GH3 ▶	H59DRAFT_06231

Bacteroides sp. D2 / Predicted PUL PUL 72 [sense] (Back to top)

RADS score = 1780

RADS alignment:

Bacteroides ovatus ATCC 8483 (new assembly) Literature-derived PUL PUL 113 [sense]		MODULAR ALIGNMENT		Bacteroides sp. D2 Predicted PUL PUL 72 [sense]
Bovatus_03075	◀ GH3	MATCH	◀ GH3	BacD2_010100013993
		NOT ALIGNED	unk ▶	BacD2_010100013998
Bovatus_03074	GH43_12 ▶	MATCH	GH43_12 ▶	BacD2_010100014003
Bovatus_03073	GH43_12 ▶		GH43_12 ▶	BacD2_010100014008
Bovatus_03072	GH5_4 ▶		GH5_4 ▶	BacD2_010100014013
Bovatus_03071	SusC ▶		SusC ▶	BacD2_010100014018
Bovatus_03070	SusD ▶		SusD ▶	BacD2_010100014023
Bovatus_03069	unk ▶	NOT ALIGNED	unk ▶	BacD2_010100014028
		GAP	GH43_16 CBM6 ▶	BacD2_010100014033
Bovatus_03068	GH9 ▶	MATCH	GH9 ▶	BacD2_010100014038
Bovatus_03067	HTCS ▶		HTCS ▶	BacD2_010100014043
Bovatus_03066	GH31 ▶		GH31 ▶	BacD2_010100014048
Bovatus_03065	GH2 ▶		GH2 ▶	BacD2_010100014053
Bovatus_03064	GH3 ▶		GH3 ▶	BacD2_010100014058

Bacteroides salanitronis DSM 18170 / Predicted PUL PUL 6 [antisense] (Back to top)

RADS score = 1190

RADS alignment:

Bacteroides ovatus ATCC 8483 (new assembly) Literature-derived PUL PUL 113 [sense]		MODULAR ALIGNMENT		Bacteroides salanitronis DSM 18170 Predicted PUL PUL 6 [antisense]
Bovatus_03075	◀ GH3	MATCH	◀ GH3	Bacsa_0893
Bovatus_03074	GH43_12 ▶	GAP		
			◀ GH95	Bacsa_0892
Bovatus_03073	GH43_12 ▶	MATCH	◀ GH43_12	Bacsa_0891
		NOT ALIGNED	unk ▶	Bacsa_0890
		GAP	SusC ▶	Bacsa_0889
			SusD ▶	Bacsa_0888
		NOT ALIGNED	unk ▶	Bacsa_0887
Bovatus_03072	GH5_4 ▶	MATCH	GH5_4 ▶	Bacsa_0886
		NOT ALIGNED	◀ unk	Bacsa_0885
Bovatus_03071	SusC ▶	GAP		
Bovatus_03070	SusD ▶			
Bovatus_03069	unk ▶	NOT ALIGNED		
Bovatus_03068	GH9 ▶	GAP		
Bovatus_03067	HTCS ▶	MATCH	HTCS ▶	Bacsa_0884
Bovatus_03066	GH31 ▶		GH31 ▶	Bacsa_0883
Bovatus_03065	GH2 ▶		GH2 ▶	Bacsa_0882
Bovatus_03064	GH3 ▶		GH3 ▶	Bacsa_0881

Figure 4. Illustration of the PUL aligner output from the literature-derived PUL 113 for xyloglucan utilization in *Bacteroides ovatus ATCC 8483*. The top left panel display the result summary, viz. the list of similar PULs (with links to each corresponding PUL webpage) ranked according to their scores, with links to the corresponding pairwise PUL alignment. The other panels present three pairwise alignments that obtained different scores in the top-left panel (highlighted by black arrows). The PUL modular organizations are displayed vertically with the query on the left and the subject on the right. Matching modules are separated by central green rectangles, while gaps are depicted by red rectangles and unaligned 'unknown' modules remained uncolored.

revealed many species with several scattered loci, one containing *susCD* genes and several others made of three or more clustered CAZyme genes. To cope with such detached gene clusters, the present PULDB update introduces the display of so-called 'CAZyme clusters'. To predict CAZyme clusters, we apply exactly the same algorithm as in PUL prediction, but instead of initiating the prediction around *susCD* genes, we start from a core of at least three adjacent CAZyme genes, not necessarily on the same strand, separated by a maximum of one single inserted gene. The display of CAZyme clusters in the PULDB web interface is accessible via a checkbox. CAZyme clusters will also help in PUL annotation of fragmented genomes. For example, despite an incomplete genome assembly, *Bacteroides ovatus ATCC 8483* became a model Bacteroidetes species thanks to RNA analysis conducted by Martens and coworkers (45). The complete genome sequence obtained later; however, reveals that the incomplete initial assembly prevented the delineation of a large PUL (Bovatus_02505 to Bovatus_02540). This was because the locus was scattered across four different short scaffolds for which CAZyme cluster definition would have at least reported two of the three split clusters.

THE PUL ALIGNER

A new tool is presented in this PULDB release to allow a user to search and identify PULs that are similar to a PUL of interest, and is accessible in the web pages dedicated to each PUL. This tool is a PUL aligner which allows retrieval conserved modular organizations. Inspired from the RADS modular alignment method for proteins (46), this tool produces local alignments of a query PUL (or CAZyme cluster) against all PULs (and CAZyme clusters) in PULDB. However, instead of aligning concatenated amino-acid sequences of proteins, it treats each protein relevant to PUL function as one character. Implementing the classical Needleman–Wunsch algorithm (47), it requires a substitution-scoring matrix between modules, as well as gap costs. A simple scheme based on the most relevant features of PULs was empirically designed. Matches of identical glycoside hydrolase and polysaccharide lyase families are given a score of +200 because they are the main actors of the polysaccharide breakdown specificity, matches of all other proteins families a score of +100 and a match of the *susCD* pair a value of +50 only, due to its presence in all predicted PULs. Proteins tagged as unknown are ignored. Given that a mutation of a protein domain into another is an evolutionary event less likely than for amino-acids, our scoring scheme also favors gaps over substitutions by giving the following penalties: internal gap opening/extension: −20/−10 and terminal gap opening/extension: −10/−5; substitution: −50. As a result, the alignment scores allow the ranking of similar PULs from the most identical (syntenic) to the most rearranged. Figure 4 shows the results of a search starting from the xyloglucan PUL of *B. ovatus ATCC 8483* (15) as the query and three aligned PULs with various conservation levels. The PUL aligner can also help in comparative genomics studies of a PUL, (i) by estimating its spread among strains of the same species, among its genus, and beyond, and (ii) by identifying the rearrangements (deletion/insertion) events that occurred during the evolution of a particular PUL.

FUNDING

Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah [67/130/35-HiCi to S.A.-M.]; European Union's Seventh Framework Program [FP/2007/2013]; European Research Council (ERC) [322820 to H.G., B.H.]. Funding for open access charge: ERC [322820].

Conflict of interest statement. None declared.

REFERENCES

1. Martens,E.C., Chiang,H.C. and Gordon,J.I. (2008) Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host Microbe*, **4**, 447–457.
2. El Kaoutari,A., Armougom,F., Gordon,J.I., Raoult,D. and Henrissat,B. (2013) The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nat. Rev. Microbiol.*, **11**, 497–504.
3. Terrapon,N., Lombard,V., Gilbert,H.J. and Henrissat,B. (2015) Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. *Bioinformatics*, **31**, 647–655.
4. Westover,B.P., Buhler,J.D., Sonnenburg,J.L. and Gordon,J.I. (2005) Operon prediction without a training set. *Bioinformatics*, **21**, 880–888.
5. Lombard,V., Golaconda Ramulu,H., Drula,E., Coutinho,P.M. and Henrissat,B. (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.*, **42**, D490–D495.
6. Ondov,B.D., Bergman,N.H. and Phillippy,A.M. (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385.
7. McNulty,N.P., Wu,M., Erickson,A.R., Pan,C., Erickson,B.K., Martens,E.C., Pudlo,N.A., Muegge,B.D., Henrissat,B., Hettich,R.L. *et al.* (2013) Effects of diet on resource utilization by a model human gut microbiota containing *Bacteroides cellulosilyticus* WH2, a symbiont with an extensive glycobiome. *PLoS Biol.*, **11**, e1001637.
8. Wu,M., McNulty,N.P., Rodionov,D.A., Khoroshkin,M.S., Griffin,N.W., Cheng,J., Latreille,P., Kerstetter,R.A., Terrapon,N., Henrissat,B. *et al.* (2015) Genetic determinants of in vivo fitness and diet responsiveness in multiple human gut *Bacteroides*. *Science*, **350**, aac5992.
9. Barbeyron,T., Thomas,F., Barbe,V., Teeling,H., Schenowitz,C., Dossat,C., Goesmann,A., Leblanc,C., Oliver Glockner,F., Czjzek,M. *et al.* (2016) Habitat and taxon as driving forces of carbohydrate catabolism in marine heterotrophic bacteria: example of the model algae-associated bacterium *Zobellia galactanivorans* DsijT. *Environ. Microbiol.*, **18**, 4610–4627.
10. Manfredi,P., Renzi,F., Mally,M., Sauteur,L., Schmaler,M., Moes,S., Jeno,P. and Cornelis,G.R. (2011) The genome and surface proteome of Capnocytophaga canimorsus reveal a key role of glycan foraging systems in host glycoproteins deglycosylation. *Mol. Microbiol.*, **81**, 1050–1060.
11. Sonnenburg,E.D., Zheng,H., Joglekar,P., Higginbottom,S.K., Firbank,S.J., Bolam,D.N. and Sonnenburg,J.L. (2010) Specificity of polysaccharide use in intestinal bacteroides species determines diet-induced microbiota alterations. *Cell*, **141**, 1241–1252.
12. Despres,J., Forano,E., Lepercq,P., Comtet-Marre,S., Jubelin,G., Yeoman,C.J., Miller,M.E., Fields,C.J., Terrapon,N., Le Bourvellec,C. *et al.* (2016) Unraveling the pectinolytic function of Bacteroides xylanisolvens using a RNA-seq approach and mutagenesis. *BMC Genomics*, **17**, 147.
13. Despres,J., Forano,E., Lepercq,P., Comtet-Marre,S., Jubelin,G., Chambon,C., Yeoman,C.J., Berg Miller,M.E., Fields,C.J., Martens,E. *et al.* (2016) Xylan degradation by the human gut *Bacteroides xylanisolvens* XB1A(T) involves two distinct gene clusters that are linked at the transcriptional level. *BMC Genomics*, **17**, 326.
14. Rogowski,A., Briggs,J.A., Mortimer,J.C., Tryfona,T., Terrapon,N., Lowe,E.C., Basle,A., Morland,C., Day,A.M., Zheng,H. *et al.* (2015) Glycan complexity dictates microbial resource allocation in the large intestine. *Nat. Commun.*, **6**, 7481.
15. Larsbrink,J., Rogers,T.E., Hemsworth,G.R., McKee,L.S., Tauzin,A.S., Spadiut,O., Klinter,S., Pudlo,N.A., Urs,K., Koropatkin,N.M. *et al.* (2014) A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. *Nature*, **506**, 498–502.
16. Ndeh,D., Rogowski,A., Cartmell,A., Luis,A.S., Basle,A., Gray,J., Venditto,I., Briggs,J., Zhang,X., Labourel,A. *et al.* (2017) Complex pectin metabolism by gut bacteria reveals novel catalytic functions. *Nature*, **544**, 65–70.
17. Cuskin,F., Lowe,E.C., Temple,M.J., Zhu,Y., Cameron,E., Pudlo,N.A., Porter,N.T., Urs,K., Thompson,A.J., Cartmell,A. *et al.* (2015) Human gut Bacteroidetes can utilize yeast mannan through a selfish mechanism. *Nature*, **517**, 165–169.
18. Bagenholm,V., Reddy,S.K., Bouraoui,H., Morrill,J., Kulcinskaja,E., Bahr,C.M., Aurelius,O., Rogers,T., Xiao,Y., Logan,D.T. *et al.* (2017) Galactomannan catabolism conferred by a polysaccharide utilization locus of Bacteroides ovatus: ENZYME SYNERGY AND CRYSTAL STRUCTURE OF A beta-MANNANASE. *J. Biol. Chem.*, **292**, 229–243.
19. Temple,M.J., Cuskin,F., Basle,A., Hickey,N., Speciale,G., Williams,S.J., Gilbert,H.J. and Lowe,E.C. (2017) A Bacteroidetes locus dedicated to fungal 1,6-beta-glucan degradation: Unique substrate conformation drives specificity of the key endo-1,6-beta-glucanase. *J. Biol. Chem.*, **292**, 10639–10650.
20. Renzi,F., Manfredi,P., Dol,M., Fu,J., Vincent,S. and Cornelis,G.R. (2015) Glycan-foraging systems reveal the adaptation of Capnocytophaga canimorsus to the dog mouth. *Mbio*, **6**, e02507.
21. Nakayama-Imaohji,H., Ichimura,M., Iwasa,T., Okada,N., Ohnishi,Y. and Kuwahara,T. (2012) Characterization of a gene cluster for sialoglycoconjugate utilization in Bacteroides fragilis. *J. Med. Invest.*, **59**, 79–94.
22. Renzi,F., Manfredi,P., Mally,M., Moes,S., Jeno,P. and Cornelis,G.R. (2011) The N-glycan glycoprotein deglycosylation complex (Gpd) from Capnocytophaga canimorsus deglycosylates human IgG. *PLoS Pathog.*, **7**, e1002118.
23. Cao,Y., Rocha,E.R. and Smith,C.J. (2014) Efficient utilization of complex N-linked glycans is a selective advantage for Bacteroides fragilis in extraintestinal infections. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 12901–12906.
24. Lee,S.M., Donaldson,G.P., Mikulski,Z., Boyajian,S., Ley,K. and Mazmanian,S.K. (2013) Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature*, **501**, 426–429.
25. Cartmell,A., Lowe,E.C., Basle,A., Firbank,S.J., Ndeh,D.A., Murray,H., Terrapon,N., Lombard,V., Henrissat,B., Turnbull,J.E. *et al.* (2017) How members of the human gut microbiota overcome the sulfation problem posed by glycosaminoglycans. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, 7037–7042.
26. Larsbrink,J., Zhu,Y., Kharade,S.S., Kwiatkowski,K.J., Eijsink,V.G., Koropatkin,N.M., McBride,M.J. and Pope,P.B. (2016) A polysaccharide utilization locus from *Flavobacterium johnsoniae* enables conversion of recalcitrant chitin. *Biotechnol. Biofuels*, **9**, 260.
27. Kabisch,A., Otto,A., Konig,S., Becher,D., Albrecht,D., Schuler,M., Teeling,H., Amann,R.I. and Schweder,T. (2014) Functional characterization of polysaccharide utilization loci in the marine Bacteroidetes '*Gramella forsetii*' KT0803. *ISME J.*, **8**, 1492–1502.
28. Sakurama,H., Kiyohara,M., Wada,J., Honda,Y., Yamaguchi,M., Fukiya,S., Yokota,A., Ashida,H., Kumagai,H., Kitaoka,M. *et al.* (2013) Lacto-N-biosidase encoded by a novel gene of Bifidobacterium longum subspecies longum shows unique substrate specificity and requires a designated chaperone for its active expression. *J. Biol. Chem.*, **288**, 25194–25206.
29. Abe,K., Nakajima,M., Yamashita,T., Matsunaga,H., Kamisuki,S., Nihira,T., Takahashi,Y., Sugimoto,N., Miyanaga,A., Nakai,H. *et al.* (2017) Biochemical and structural analyses of a bacterial endo-beta-1,2-glucanase reveal a new glycoside hydrolase family. *J. Biol. Chem.*, **292**, 7487–7506.
30. Munoz-Munoz,J., Cartmell,A., Terrapon,N., Henrissat,B. and Gilbert,H.J. (2017) Unusual active site location and catalytic apparatus in a glycoside hydrolase family. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, 4936–4941.
31. Ulaganathan,T., Boniecki,M.T., Foran,E., Buravenkov,V., Mizrachi,N., Banin,E., Helbert,W. and Cygler,M. (2017) New ulvan-degrading polysaccharide lyase family: structure and catalytic mechanism suggests convergent evolution of active site architecture. *ACS Chem. Biol.*, **12**, 1269–1280.
32. Kopel,M., Helbert,W., Belnik,Y., Buravenkov,V., Herman,A. and Banin,E. (2016) New family of ulvan lyases identified in three isolates from the alteromonadales order. *J. Biol. Chem.*, **291**, 5871–5878.

33. Iwai,M., Kawakami,T., Ikemoto,T., Fujiwara,D., Takenaka,S., Nakazawa,M., Ueda,M. and Sakamoto,T. (2015) Molecular characterization of a Penicillium chrysogenum exo-rhamnogalacturonan lyase that is structurally distinct from other polysaccharide lyase family proteins. *Appl. Microbiol. Biotechnol.*, **99**, 8515–8525.
34. Munoz-Munoz,J., Cartmell,A., Terrapon,N., Basle,A., Henrissat,B. and Gilbert,H.J. (2017) An evolutionarily distinct family of polysaccharide lyases removes rhamnose capping of complex arabinogalactan proteins. *J. Biol. Chem.*, **292**, 13271–13283.
35. Skinner,M.E., Uzilov,A.V., Stein,L.D., Mungall,C.J. and Holmes,I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.
36. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
37. Helbert,W. (2017) Marine polysaccharide sulfatases. *Front. Mar. Sci.*, **4**, 6.
38. Benjdia,A., Martens,E.C., Gordon,J.I. and Berteau,O. (2011) Sulfatases and a radical S-adenosyl-L-methionine (AdoMet) enzyme are key for mucosal foraging and fitness of the prominent human gut symbiont, *Bacteroides thetaiotaomicron*. *J. Biol. Chem.*, **286**, 25973–25982.
39. Barbeyron,T., Brillet-Gueguen,L., Carre,W., Carriere,C., Caron,C., Czjzek,M., Hoebeke,M. and Michel,G. (2016) Matching the diversity of sulfated biomolecules: creation of a classification database for sulfatases reflecting their substrate specificity. *PLoS One*, **11**, e0164846.
40. Yan,N. (2015) Structural biology of the major facilitator superfamily transporters. *Annu. Rev. Biophys.*, **44**, 257–283.
41. Tauzin,A.S., Laville,E., Xiao,Y., Nouaille,S., Le Bourgeois,P., Heux,S., Portais,J.C., Monsan,P., Martens,E.C., Potocki-Veronese,G. *et al.* (2016) Functional characterization of a gene locus from an uncultured gut Bacteroides conferring xylo-oligosaccharides utilization to *Escherichia coli*. *Mol. Microbiol.*, **102**, 579–592.
42. Stafford,G., Roy,S., Honma,K. and Sharma,A. (2012) Sialic acid, periodontal pathogens and Tannerella forsythia: stick around and enjoy the feast! *Mol. Oral Microbiol.*, **27**, 11–22.
43. Brigham,C., Caughlan,R., Gallegos,R., Dallas,M.B., Godoy,V.G. and Malamy,M.H. (2009) Sialic acid (N-acetyl neuraminic acid) utilization by Bacteroides fragilis requires a novel N-acetyl mannosamine epimerase. *J. Bacteriol.*, **191**, 3629–3638.
44. Rawlings,N.D., Barrett,A.J. and Finn,R. (2016) Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.*, **44**, D343–D350.
45. Martens,E.C., Lowe,E.C., Chiang,H., Pudlo,N.A., Wu,M., McNulty,N.P., Abbott,D.W., Henrissat,B., Gilbert,H.J., Bolam,D.N. *et al.* (2011) Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS Biol.*, **9**, e1001221.
46. Terrapon,N., Weiner,J., Grath,S., Moore,A.D. and Bornberg-Bauer,E. (2014) Rapid similarity search of proteins using alignments of domain arrangements. *Bioinformatics*, **30**, 274–281.
47. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.