Bioinformatics, 31(5), 2015, 647–655
doi: 10.1093/bioinformatics/btu716
Advance Access Publication Date: 28 October 2014
Original Paper

OXFORD

Genome analysis

Automatic prediction of polysaccharide utilization loci in Bacteroidetes species

Nicolas Terrapon[1,2]*, Vincent Lombard[1], Harry J. Gilbert[3] and Bernard Henrissat[1,4,*]

[1]Centre National de la Recherche Scientifique, CNRS UMR 7257, 13288 Marseille, France, [2]Aix-Marseille Université, AFMB, 13288 Marseille, France, [3]Institute for Cell and Molecular Biosciences, The Medical School, Newcastle University, Newcastle upon Tyne NE2 4HH, UK and [4]Department of Biological Sciences, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on July 3, 2014; revised on October 20, 2014; accepted on October 23, 2014

Abstract

Motivation: A bacterial polysaccharide utilization locus (PUL) is a set of physically linked genes that orchestrate the breakdown of a specific glycan. PULs are prevalent in the Bacteroidetes phylum and are key to the digestion of complex carbohydrates, notably by the human gut microbiota. A given Bacteroidetes genome can encode dozens of different PULs whose boundaries and precise gene content are difficult to predict.

Results: Here, we present a fully automated approach for PUL prediction using genomic context and domain annotation alone. By combining the detection of a pair of marker genes with operon prediction using intergenic distances, and queries to the carbohydrate-active enzymes database (www.cazy.org), our predictor achieved above 86% accuracy in two *Bacteroides* species with extensive experimental PUL characterization.

Availability and implementation: PUL predictions in 67 Bacteroidetes genomes from the human gut microbiota and two additional species, from the canine oral sphere and from the environment, are presented in our database accessible at www.cazy.org/PULDB/index.php.

Contact: bernard.henrissat@afmb.univ-mrs.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Glycans are exceptionally diverse molecules whose roles notably include energy storage (starch in plants, glycogen in metazoans and fungi), structure (cellulose, hemicelluloses and pectins of plant cell walls, chitin of arthropod exoskeletons), cell communication and host pathogen interactions. Simple monosaccharides can be assembled via α- and β-glycosidic bounds into macromolecules of enormous diversity. It has been estimated that $>10^{12}$ hexasaccharides could be built using only D-hexoses (Laine, 1994). Even though not all possible glycan structures exist in nature, the breakdown of such complex molecules requires a plethora of carbohydrate active enzymes (CAZymes), for which the CAZy database—available at www.cazy.org—offers a continuously updated sequence-based family classification (Lombard *et al.*, 2014). Organisms that feed on complex glycans have thus evolved large CAZyme repertoires.

The human diet includes a large amount of carbohydrates, ranging from simple disaccharides, such as sucrose and lactose, to storage polysaccharides, typically glycogen and starch, and the multitude of complex structures found in the cell walls of cereals, fruits, vegetables and in the mammalian extracellular matrix. However, the human genome only encodes digestive CAZymes that break down sucrose, lactose and the readily digestible component of starch (El Kaoutari *et al.*, 2013). All other carbohydrates, often termed dietary fiber, enter the large bowel where a dense microbial community comprising trillions of bacteria, the microbiota, breaks down these structures, as well as host-derived (mucosal) glycans.

The monosaccharides generated are fermented by the microbiota to generate short chain fatty acids, which can provide up to 10% of our daily caloric needs (McNeil, 1984).

With the recent advances in DNA sequencing techniques, many metagenomic studies have focused on the relationship between dietary fiber and the microbiota, to gain insights into how the nutrition of this microbial ecosystem influences the physiology of the human host—see Flint *et al.* (2012) for a review. Despite species-level variability, it has been found that two bacterial phyla usually represent the majority of the sequenced 16S rRNA derived from the human gut microbiota: the Gram-positive Firmicutes and the Gram-negative Bacteroidetes (Ley *et al.*, 2006).

As Gram-negative bacteria, Bacteroidetes species are characterized by a periplasmic space located between two lipid bilayer cellular membranes. This compartment enabled these bacteria to evolve a specific strategy for carbohydrate breakdown, which are encoded by physically linked and functionally related genes known as polysaccharide utilization loci (PULs). Each PUL encodes co-regulated proteins that (i) bind to polysaccharides on the surface of the bacterium, (ii) mediate limited cleavage of the targeted polysaccharides on the outer membrane, (iii) import oligosaccharides generated on the bacterial surface to the periplasm, (iv) catalyze the final saccharification of the imported oligosaccharides to monosaccharides in the periplasmic space and (v) sense degradation products to regulate PUL expression. Through duplication and neofunctionalization, Bacteroidetes have largely inherited and exchanged PULs via vertical and horizontal transfer. As a consequence, extant Bacteroidetes species now exhibit dozens of PULs that encode different enzyme consortia that are tailored to degrade specific glycans. PUL diversity may contribute to the evolutionary success of Bacteroidetes by providing a general glycan utilization strategy, optimized to catabolize the extensive repertoire of complex carbohydrates presented to the mammalian large bowel. The PUL system is not restricted to mammalian gut Bacteroidetes, as this phylum is also common in other body sites, other animals as well as in the environment (McBride *et al.*, 2009).

The PUL paradigm was initially defined for the starch utilization system (Sus) of the human gut commensal *Bacteroides thetaiotaomicron*. This PUL consists of eight genes, *susA-G* and *susR*, including three genes that encode CAZymes. On the outer cell surface, SusDEF bind components of starch, while α-amylase SusG binds and cleaves the polysaccharide to oligosaccharides, which are sequestrated by SusC into the periplasmic space of the bacterium. Imported oligosaccharides, which are sensed by SusR leading to upregulation of the PUL, are depolymerized to glucose by neopullanase SusA and α-glucosidase SusB. The functional characterization of each Sus PUL member has been performed independently (Cameron *et al.*, 2012; Kitamura *et al.*, 2008; Koropatkin and Smith, 2010; Koropatkin *et al.*, 2008). More recently, Larsbrink *et al.* (2014) characterized all eight CAZymes of a 12-gene PUL of *Bacteroides ovatus* dedicated to xyloglucan degradation. Such comprehensive approaches are expected to become more widespread as the understanding of PUL function requires the elucidation of the synergistic interactions between all components of the glycan metabolizing apparatus orchestrated by these loci (Terrapon and Henrissat, 2014). In this context, it is critical that all the proteins encoded by specific PULs are identified and, when appropriate, are correctly assigned to families in the CAZy database, which in turn provides insights into the possible function of these CAZymes. An ideal approach would be the exhaustive screening of gene expression of a target species in response to a large variety of carbohydrate substrates, as done by Martens *et al.* (2008) for *B.thetaiotaomicron*, by Martens *et al.* (2011) for *B.ovatus* and by McNulty *et al.* (2013) for *Bacteroides cellulosilyticus*. However, the large number of Bacteroidetes species, the difficulty in accessing all the possible glycans in a purified and highly defined form—impure glycans can give highly misleading gene expression data (Martens *et al.*, 2011), precludes the systematic identification of PULs based solely on gene expression. An alternative, and more tractable, strategy is the development of an automatic prediction method to identify these loci from genome sequence alone, thereby avoiding the need for bacterial cultivation and extensive gene expression studies.

Here, we present a bioinformatics approach for PUL prediction in Bacteroidetes species. We evaluate the performance of our approach against experimentally characterized or predicted PULs, available for two Bacteroidetes species of the human gut microbiota. We have applied this predictor to 67 genomes of Bacteroidetes species of the human gut to offer a perspective on the general characteristics of PULs and to allow intergenome comparisons. These predictions are made accessible through a web interface, which also includes, for comparison purposes, two other Bacteroidetes: the environmental *Flavobacterium johnsoniae* (McBride *et al.*, 2009) and the canine oral sphere bacterium *Capnocytophaga canimorsus* (Manfredi *et al.*, 2011).

2 Methods

Our predictive strategy is based on the observed essential features of PULs obtained through experimental analysis of these loci in *B.thetaiotaomicron* VPI-5482 and *B.ovatus* ATCC 8483, referred to as *reference PULs* hereafter (Section 2.1). Our PUL prediction program was used to interrogate 67 Bacteroidetes genomes downloaded from the IMG/M-HMP resource (Section 2.2). The strategy relies on the identification in each genome of the PUL markers: the presence of adjacent genes encoding SusC-like and SusD-like proteins, referred to as *tandem susCD-like pairs* (Section 2.3). Indeed, a PUL looks like an operon structured around *tandem susCD-like pairs*, while the members of these loci (encoding regulators, binding proteins and enzymes that cleave glycans), vary from one locus to another. Hence, we initiated PUL prediction by a strategy similar to operon prediction, delineating PULs around the *tandem susCD-like pairs* based on intergenic distances only (Section 2.4). Reference PULs frequently contain genes encoding CAZymes (*cazymes*), which, although functionally related, may be separated by large intergenic regions to allow finer regulation or by gene of unknown function. Hence, our predictor extends PUL boundaries to adjacent operons based on the presence of a *cazyme* using a sliding window (Section 2.5). Additional indicators of conserved adjacency (microsynteny) have been tested (Section 2.6). PULs often start/end with a regulatory gene either on the opposite DNA strand or separated by a large intergenic distance from the other members of the locus. As a consequence, we developed a specific method for the inclusion of regulators (Section 2.7). We also observed in *B.ovatus*, the presence of short gene models, encoding proteins shorter than 60 amino acids, which interfered with the PUL prediction. These short gene models were thus ignored (Section 2.8). Finally, to refine the delineation of PULs with multiple *tandem susCD-like pairs*, we implemented rules for the fusion and division of adjacent PULs (see Section 2.9). Our prediction strategy was designed by selecting the best criteria and their combination after evaluation of the results against the reference PULs of *B.thetaiotaomicron* and *B.ovatus* using receiver operating characteristic (ROC) curves (Section 2.10).

2.1 Reference PULs

Current knowledge of the characteristic features of PULs mainly arises from work published by Martens *et al.* (2008, 2011) for two species. These studies enabled PUL prediction based on: (i) coordinated gene expression using transcriptomics, (ii) gene inactivation and biochemical characterization and (iii) manual inspection of the genomic context. These *reference PULs* have been used in the design and evaluation of our prediction strategy. However, many of the reference PULs are still putative and thus not necessarily precisely delineated since the target substrate has not been established. Within the literature, discrepancies also appear between the tabulated and pictorial presentation of some *B.thetaiotaomicron* reference PULs (Martens *et al.*, 2008), as well as between 'homologous PULs' defined between *B.thetaiotaomicron* and *B.ovatus* (Martens *et al.*, 2011). We opted to use the tabular presentation of reference PULs in these two species.

2.2 Bacteroidetes genomic data

Genomic data were downloaded from the Integrated Microbial Genome and Metagenomes/Human Microbiome Project [IMG/M-HMP (Markowitz *et al.*, 2012)] resource in March 2013. For all the 67 Bacteroidetes species with completed genome sequences and tagged as originating from the *gastrointestinal tract*, three types of data were used: (i) information about gene position, orientation and intergenic distance (GFF file), (ii) predicted amino acid sequence (fasta file) and (iii) protein domain annotation using the Pfam v27 resource (Finn *et al.*, 2014). The *cazymes* were identified and assigned to sequence-based families using the same procedures deployed to update the CAZy database (Lombard *et al.*, 2014). Supplementary Table S3 presents the CAZy annotation for those 57 genomes (not released in the NCBI database) that are not present in the CAZy website.

2.3 Signature of PULs: the *susC*- and *susD*-like genes

The identification of *sus*C-like and *sus*D-like genes was based on Pfam domain assignment. SusC-like proteins are TonB-dependent transporters characterized by a specific linear sequence of domains: PF13715 (unknown function), PF07715 (TonB-dependent receptor's plug) and PF00593 (TonB-dependent receptor), as illustrated in Figure 1. We also captured possible variations of this combination that include: (i) the addition of an N-terminal domain PF07660 that interacts with ECF-σ/anti-σ regulators (cf Section 2.7) in trans-envelope signaling (Koebnik, 2005; Martens *et al.*, 2009), (ii) the split of the *sus*C-like gene model into two genes and (iii) the apparent lack of PF00593 domain that scored slightly worse than the Pfam recommended threshold. Encoded SusD-like proteins were characterized by either a single or a pair of domains among four related domains in Pfam: PF07980, PF12741, PF12771, PF14322 (SusD and SusD-like domains), as illustrated in Figure 1.

For our predictions, we defined *tandem susCD-like pairs* in a genome based on the presence of adjacent *sus*C-like (upstream) and *sus*D-like (downstream) genes, located on the same DNA strand. Because specific tandem SusCD-like pairs can be occasionally extended by adjacent copies of either *sus*C- (upstream) or *sus*D-like genes (downstream), such duplications were incorporated.

2.4 Operon prediction using intergenic distances

Automatic operon prediction has been intensively studied in the model bacteria *Escherichia coli* and *Bacillus subtilis* (Bockhorst *et al.*, 2003; Moreno-Hagelsieb and Collado-Vides, 2002; Salgado *et al.*, 2000). The classical indicators of operon membership include: (i) intergenic distances, (ii) synteny across different genomes, (iii) the occurrence of gene pairs fused into single genes in other genomes, (iv) similarity of functional annotation and (v) gene co-expression. However, operon prediction based on genomic information only is highly desirable as transcriptomics data are not always available. Moreover, functional annotation is often unreliable, and a sequence-based method would also allow predictions to be made for non-cultivated sequenced species. In this context, Westover *et al.*, (2005) confirmed in *E.coli* and *B.thetaiotaomicron* that the intergenic distance in *directons*—a term that refers to operons made of consecutive genes on the same strand without interruption by any gene on the opposite strand—is the most powerful feature for operon prediction. As PULs are not exactly operons, we tested both intergenic distances with or without restriction to directons, as well as operon memberships from publicly available resources: the publication of Westover *et al.* (2005) and two general databases: DOOR (Mao *et al.*, 2009) and ProOpDB (Taboada *et al.*, 2012). Each approach produced a list of contiguous gene pairs. PUL prediction started from each *tandem susCD-like pair* and iteratively extended PUL boundaries to adjacent genes when the pairs belonged to the list of contiguous pairs.

2.5 CAZy annotation

The CAZy database (www.cazy.org) has offered for more than 15 years an online and continuously updated classification of proteins based on sequence similarity to functionally characterized CAZymes (Lombard *et al.*, 2014). CAZy annotation consists of assigning module/domain families and subfamilies for six major CAZyme categories: glycoside hydrolases (GHs) and polysaccharide

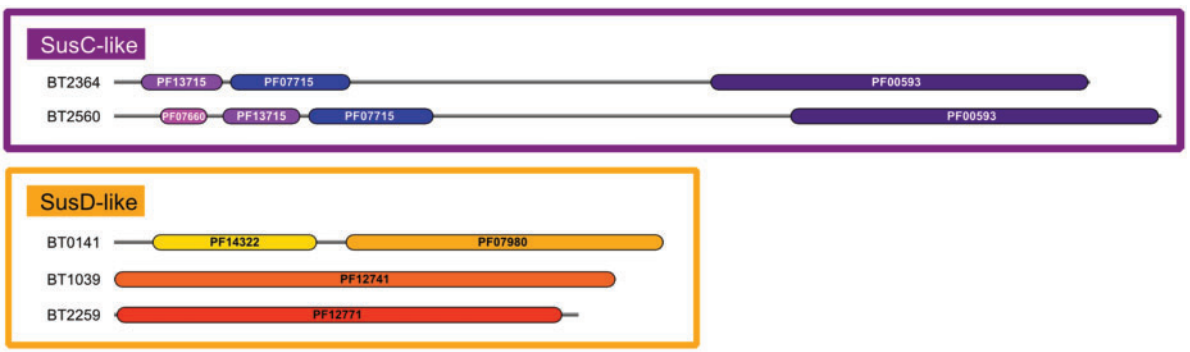

Fig. 1. Example of Pfam domain architectures of SusC-like and SusD-like proteins from *B.thetaiotaomicron*. The domain string representations were generated with DoMosaics (Moore *et al.*, 2014)

lyases (PLs) that cleave glycans, glycosyltransferases that synthesize glycans, non-catalytic carbohydrate binding modules (CBMs), carbohydrate esterases (CEs) and auxiliary activities. Our PUL predictor only makes use of GHs, PLs, CBMs and CEs that are directly connected to glycan degradation. In our PUL predictor, CAZy annotations were used to extend PUL boundaries beyond intergenic-distance predictions. The principle was to search for a CAZy annotation among the genes that directly follow the PUL boundary, from the first to the fifth gene located after the boundary. If a CAZyme gene was found, the PUL boundary was extended to this gene. Iteratively, PUL predictions restarted from this gene until no further PUL member could be identified.

2.6 Domain adjacency information

Conservation of adjacent genes across genomes (Dandekar *et al.*, 1998; Ermolaeva *et al.*, 2001), and the existence of a gene pair as a fused gene in another genome (Enright *et al.*, 1999; Marcotte *et al.*, 1999), can provide strong evidence for an operonic structure. Such features can be captured using protein domains—the structural, functional and evolutionary sub-units of proteins (Apic *et al.*, 2001; Hegyi and Gerstein, 2001). Hence, we computed two lists of conditionally dependent pairs [CDP (Terrapon *et al.*, 2009)] of Pfam domains to detect *gene split* and *synteny*, using the 67 Bacteroidetes genomes from the human gut. For gene split, we considered all pairs formed by the domains of each protein. For gene synteny, we considered all domain pairs formed by adjacent genes. CDP lists were obtained by retaining only statistically significant pairs (Fisher's one-tailed exact test; *P*-value thresholds adjusted through ROC curve comparisons). PUL boundaries were iteratively extended to the neighboring genes if a pair, formed by domains encoded by the PUL boundary gene and this following gene, was found in the CDP list.

2.7 Terminal regulators

Inspection of the reference PULs and initial prediction attempts revealed that PUL regulatory/sensory genes are difficult to identify and required special processing. First, regulators are frequently located on the opposite DNA strand or/and separated by a much larger intergenic distance than the other genes in the locus. Moreover, the diversity of described regulator families, their poor domain characterization and their involvement in other processes than carbohydrate metabolism are important limitations to their systematic identification. We first focused on well-characterized hybrid two-component systems (HTCS) and the extra-cytoplasmic functioning (ECF) family of σ/anti-σ factors. Although other regulatory families exist such as SusR, AraC, GntR, SARP/OmpR, LacI and CRP, the majority of PUL regulators only involve HTCS and ECF families: 28 HTCS and 22 ECF in *B.thetaiotaomicron*, 39 HTCS and 35 ECF in *B.ovatus*, while only 12 PULs in *B.thetaiotaomicron* and 8 PULs in *B.ovatus* are linked to another regulator family. The identification of HTCS and ECF regulators rely on Pfam domain composition: PF04542-PF08281 (ECF-σ), PF04773 (Anti-σ) and PF07494-PF07495-PF00512-PF02518-PF00072-PF12833 (HTCS; with frequent variations at the N-terminus). We additionally included three minor families of regulators: (i) GntR, characterized by Pfam domain PF00392, (ii) AraC, based on a combination of Pfam domains PF12833 with either PF02311 or PF13377 and (iii) SusR, the prototypic PUL regulator, for which no Pfam model is currently available and we thus created our own. For the design of our prediction strategy, regulator detection was tested up to three genes after the last PUL member. Finally, as we observed in reference PULs that regulators usually mark the end of the locus, except for HTCS, we interrupted our iterative extension of PUL boundaries after regulatory gene detection, except for HTCS regulons for which the prediction iteration was allowed to restart if a *cazyme* was encountered after the HTCS gene.

2.8 Filtering short gene models

During initial prediction attempts for *B.ovatus*, we observed that some mismatches between predicted and reference PULs were caused by short gene models that interfered with the prediction. These short gene models were absent in *B.thetaiotaomicron*. Whatever the origin of these genes (assembly problems or real fragments), we introduced an additional criterion into the predictor: genes shorter than 180 bp were not considered. This procedure clearly improved prediction for *B.ovatus*, and did not affect *B.thetaiotaomicron*'s results (data not shown).

2.9 Fusion and division of adjacent PULs

Examination of the reference PULs revealed that they are frequently composed of several adjacent operons that are not necessarily on the same DNA strand. Thus, we observed reference PULs with several *tandem susCD pairs*, which appear like the fusion of adjacent PULs. However, there are also many demonstrated examples of adjacent PULs that have been differentiated due to differential substrate responses. Such examples include the heparin utilization region of *B.thetaiotaomicron*, which is split by an insertion of a beta-galactan PUL (Martens *et al.*, 2008) and the xyloglucan utilization locus of *B.ovatus* which sits immediately upstream of another PUL that does not show differential expression in the presence of xyloglucan (Larsbrink *et al.*, 2014; Martens *et al.*, 2011). These adjacent, but distinct, PULs are frequently merged together by our prediction strategy which works as a sliding window for the incorporation of *cazymes*. To more accurately delineate such a complex arrangement of PULs, our strategy first assembles contiguous sets of genes into single large entities (prospective PULs). These sets of genes are assembled in three cases: (i) when constitutive *cazymes* are separated by two or less unknown genes, (ii) when two prospective PULs are directly adjacent and (iii) when an integrase gene (Pfam domain PF00589) is found between two such prospective PULs, following the example of the mucin O-glycan reference PUL 14 of *B.thetaiotaomicron* (Martens *et al.*, 2008). We then attempt to divide the large assembled entities that exhibit more than one *tandem susCD pair* only when a regulatory gene other than HTCS is found (keeping the regulator in the closest *tandem susCD pair*) or when two directons (without any interrupting genes) are running in opposite directions on opposite strands.

2.10 Evaluation of PUL predictions

The different prediction criteria described in the earlier sections were evaluated against the reference PULs of *B.thetaiotaomicron* and *B.ovatus* (see Section 2.1) to define our overall prediction strategy. Reference PULs are sometimes interrupted by one or more genes that are not components of the PUL (hereafter called 'non-PUL genes') based on their lack of coregulation with the other members of the PUL. This results in non-contiguous gene pairs for which we had to create a novel evaluation approach that extends the classical operon evaluation model. Our evaluation scheme, illustrated in Supplementary Figure S1, divides all adjacent gene pairs in three categories: (i) *positive pairs* corresponding to adjacent genes inside the reference PULs (768 and 1027 in *B.thetaiotaomicron* and *B.ovatus*, respectively), (ii) *half positive/half negative pairs*

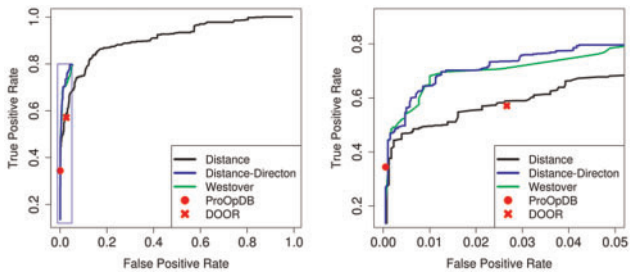

Fig. 2. ROC evaluation of PUL predictions in *B.thetaiotaomicron* using intergenic distances or operon databases. Note that for better visual inspection the blue rectangle in left panel is magnified in right panel (Color version of this figure is available at *Bioinformatics* online.)

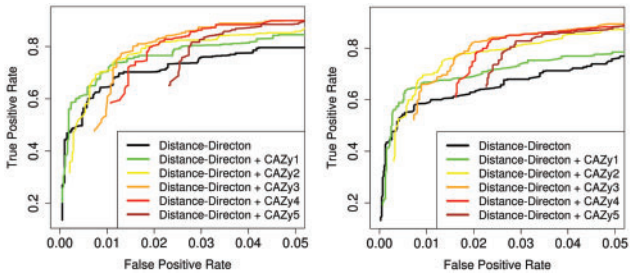

Fig. 3. ROC evaluation of PUL predictions adding CAZy annotation. The left and right panels display results obtained in *B.thetaiotaomicron* and *B.ovatus*, respectively. Note that for better visual inspection the *x*- and *y*-axes are not plotted with the same scale

where a PUL gene involved in a non-contiguous pair in the reference PULs is linked to its adjacent non-PUL gene (20 and 10 in *B.thetaiotaomicron* and *B.ovatus*, respectively) and (iii) *negative pairs* that include all the remaining pairs not components of the reference PULs (4026 and 4447 in *B.thetaiotaomicron* and *B.ovatus*, respectively). This evaluation scheme avoided the exclusion of non-contiguous PUL genes or the inclusion of intercalated non-PUL genes.

To assess the quality of the PUL predictions, we listed the predicted adjacent gene pairs of PUL genes for each possible intergenic distance. We then computed for each distance value: (i) the true positive rate (TPR), the ratio of correctly predicted pairs over the total number of positive pairs and (ii) the false positive rate (FPR), the ratio of wrongly predicted pairs over the total number of negative pairs. Finally, we plotted ROC curves that showed the evolution of TPR and FPR when varying the intergenic distance thresholds and used these to compare the different PUL prediction strategies (see Figs 2 and 3; Supplementary Figs S1–S3).

3 Results and discussion

3.1 PUL prediction strategy

The objective of this work was to design a fully automated PUL predictor, based on genomic context only, that draws on reference PULs defined using a combination of experimental data and analysis (Martens *et al.*, 2008, 2011). In this section, we (Apic *et al.*, 2001) show the evaluation of the distinct predictive features and compare their predictive power using ROC curves.

3.1.1 Intergenic distance in directons

We first examined the predicted PULs around *tandem susCD-like pairs* using intergenic distances alone (with or without restriction to directons), and compared the results to the operonic pairs described for *B.thetaiotaomicron* in DOOR, ProOpDB and by Westover *et al.* (2005) (Fig. 2). We observed that the method able to retrieve the largest number of PUL genes is the intergenic distance without direction restriction, but only at the expense of numerous false predictions. The DOOR and ProOpDB databases identify only a very limited number of PULs, most likely due to the lack of calibration against data from the particular Bacteroidetes phylum. The solution providing the best tradeoff is the intergenic distance restricted to directons, alone or in combination with other indicators [functional annotation and conserved synteny in predictions from Westover *et al.* (2005)]. Hence, we used the distances in directons only in our PUL prediction model, as it obtain similar results and will be less subject to variation across genome projects (annotation of various function—import, binding, cleavage, signaling—and assembly completeness) than Westover's predictions.

3.1.2 Addition of CAZymes

We next integrated the information about *cazymes* into the PUL predictor based on intergenic distances in directons. The PUL boundaries were extended when a *cazyme* was adjacent (CAZy1), second (CAZy2), third (CAZy3), etc., relative to the current PUL boundary. The results of ROC analysis for *B.thetaiotaomicron* and *B.ovatus* (Fig. 3) show that the inclusion of CAZy annotation significantly enhanced the predictive power of our approach. On one hand, CAZy1 provided the safest predictions (highest TPRs for similar FPRs), whereas CAZy2 achieved higher TPRs than CAZy1, but at the expense of a few false positive predictions. The same was true for CAZy3 compared to CAZy2, but CAZy4 and CAZy5 did not further improve the predictions. As CAZymes are central to understand PUL synergy and to identify target substrates, the largest set of predictions is desirable. We therefore decided to provide the user with several confidence levels, from the predictions with the lowest FPRs (CAZy1), to predictions containing more true positives, but also some false positives (CAZy3). Such distinctions are displayed in our web interface (Section 3.2).

3.1.3 Domain adjacency

The integration of indicators of gene split and synteny had a limited impact (Supplementary Figs S1 and S2). We observed that only the gene directly adjacent to the last PUL member should be considered. While the detection of gene split was generally positive, we removed the indicators of gene synteny from our predictor as this parameter brought similar numbers of true and false positives.

3.1.4 Final predictions

After the integration of the families of regulators (best performance was obtained for detection up to the second gene after the last PUL member—data not shown), the elimination of short gene models and the fusion and division of PULs, our final predictions suggest that the best trade-off between TPRs and FPRs correspond to an intergenic distance $\leq$102 bp (highlighted by a circle in Supplementary Fig. S3). The predicted PUL members represent 87 and 91% of the reference PULs in *B.thetaiotaomicron* and *B.ovatus*, respectively, while extra-predictions (false positives) represent <8% of the 'true' members (more details in Section 2 and Supplementary Table S4). The PUL division (Section 2.9) notably allowed to reduce the number of predicted PULs with multiple *tandem SusCD-like pairs* from 19 and 22 in *B.thetaiotaomicron* and *B.ovatus* to 13 and 15, respectively, for 10 and 14 experimentally validated PULs. The adjacent PUL delineation, as well as the recognition of

non-contiguous gene pairs are current limitations of our approach. Such cases, as the xyloglucan PUL in *B.ovatus* described by Larsbrink *et al.* (2014) or the heparin and beta-galactan PUL in *B.thetaiotaomicron* validated in Martens *et al.* (2008) (see Section 2.9), are currently resolvable only by experimental characterization. A possible solution for such difficult cases would be the integration of comparative genomics information, starting from experimentally validated PULs. However, at this stage and following the practice in biocuration, we prefer to report the actual boundaries only when we find published evidence for the studied strain, with no extrapolation to other genomes. In the future, based on examination of a larger amount of expression data, we might revise this conservative policy and investigate whether boundaries can be safely extended from one strain to other strains or other bacteria.

3.2 Web interface

We used our predictor to analyze 67 genomes of Bacteroidetes from the human gut microbiota (see Supplementary Fig. S4 for the distribution at the taxonomic level). We additionally included PUL predictions for two other Bacteroidetes species for which a list of all putative PULs have previously been published: the environmental gliding bacterium *F.johnsoniae* (McBride *et al.*, 2009) and *C.canimorsus*, isolated from the oral canine cavity and responsible for severe human septicemia (Manfredi *et al.*, 2011). A web interface was implemented to allow browsing of the database of predicted PULs (PULDB) in these 69 genomes at www.cazy.org/PULDB/index.php.

This interface first provides the user with the list of all PULs for a given species, or taxonomic group. The result page (Fig. 4; Supplementary Fig. S5) displays PULs as trains whose wagons are the genes colored according to the function of the encoded proteins for SusC-like and SusD-like, CAZyme modules and regulators HTCS and ECF-σ /anti-σ factors; the remaining proteins are considered as having unknown function (in gray). To facilitate the visualization and comparison of PULs, we allow the reversal of PUL orientation and the reordering of the table to bring similar PULs closer. Furthermore, we implemented the search for specific PULs based on CAZy families, with the possibility of indicating the sequential order of the constituent families. For example, searching for PULs encoding ordered CAZy families GH3, GH2, GH31 and GH5 for the breakdown of xyloglucan, as described by Larsbrink *et al.* (2014), allows the retrieval of the corresponding PULs in eight other species with a high level of micro-synteny (see Fig. 4).

Moreover, to allow inspection of the genomic context of a PUL, we integrated a JBrowse engine (Skinner *et al.*, 2009), illustrated by Supplementary Figure S6. The user can visualize the predicted PUL with confidence levels (CAZy1 to CAZy3 colored from green to red) as well as published PULs. A contextual menu associated to the genes and PULs allows the user to open new pages with detailed information about (i) the encoded proteins (following IMG/M-HMP annotation or Pfam domains), (ii) the function of CAZymes if known (EC numbers) and (iii) the experimentally verified substrates. Finally, the JBrowse engine also allow loading the user's own expression data, such as short-reads from BAM files.

3.3 PUL characteristics

3.3.1 General features of predicted PULs in 67 genomes

The PUL predictions for the 67 Bacteroidetes species from the human gastrointestinal tract offer a large-scale perspective of the characteristics of these loci and allow the derivation of some general statistics about PUL composition. A total of 3745 PULs were predicted in the 67 Bacteroidetes genomes of the human gut microbiota. For each species, Supplementary Table S5 presents the global PUL content (number of genes, *tandem susCD-like pairs* and *cazymes*) and the specific features of these loci (presence/absence of *cazymes*, number of *cazymes* per PUL, regulators).

In predicted PULs, as in reference PULs, we observed a certain number of loci with no GH- or PL-encoding gene. A possible explanation could be that these putative PULs encode GHs or PLs that are yet unknown and therefore not present in the CAZy database. Other explanations include the possibility that these PULs represent either non-functional gene cassettes, or functional elements that

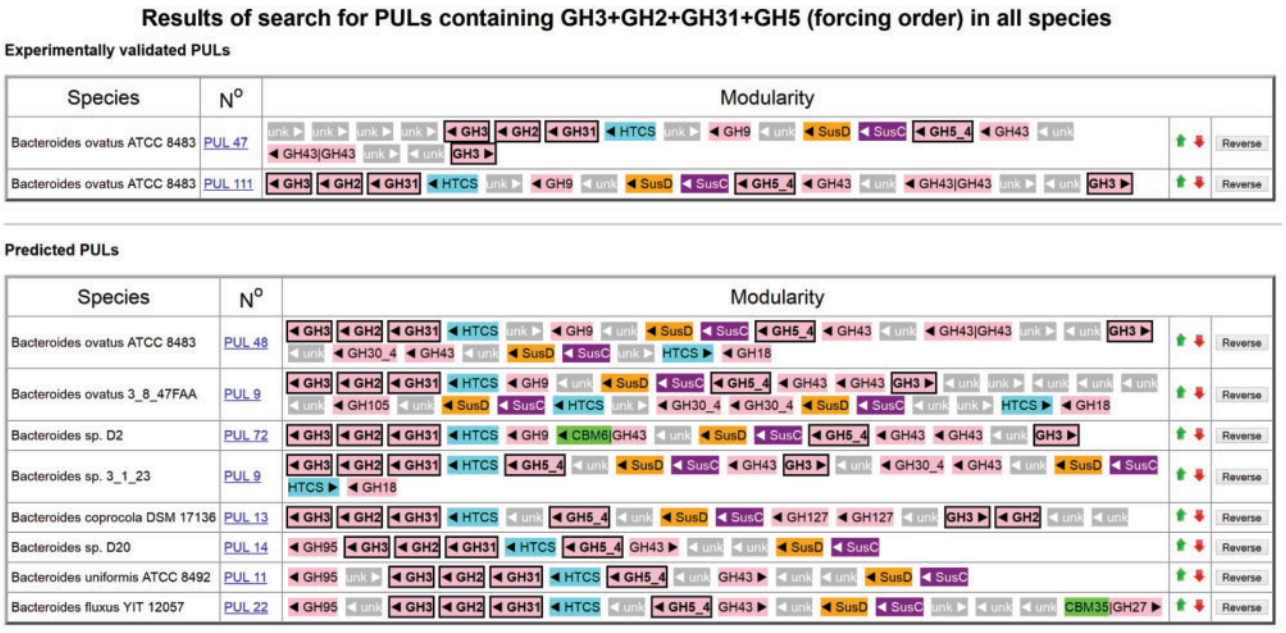

Fig. 4. Screenshot of a search from the web interface. Query corresponds to CAZy families GH3, GH2, GH31 and GH5 (ordered), as required for the xyloglucan PUL described by Larsbrink *et al.* (2014)

would assist other loci, for example, in the surveillance of available substrate. Another possibility is that SusCD-encoding operons could fulfill functions other than the utilization of carbohydrate polymers. Interestingly, the number of PULs without GH or PLs is smaller and more similar across species than the number of PULs encoding these enzymes (Supplementary Fig. S7). This suggests that PULs encoding GHs or PLs could fulfill conserved generic functions, while the acquisition of PULs with such enzymes may reflect adaptation to the specific environment of the bacterium.

In the 67 genomes, the number of GH and PL genes found outside of PULs is fairly constant (~60 per genome; Fig. 5) and probably includes many house-keeping functions. Deviations mainly involve highly fragmented genomes (>200 scaffolds; highlighted in Fig. 5) for which PUL predictions end at scaffold borders. Here, GH and PL genes separated from *tandem susCD-like pairs* by a scaffold break necessarily escape integration in a PUL and artificially increase the number of such genes apparently outside of PULs. Interestingly, the number of GH and PL genes predicted in PULs is extremely variable across species (Fig. 5), with the average number per PUL ranging from 2.5 to 4 (Supplementary Fig. S8). This suggests that evolution does not simply promote loci encoding GHs and PLs *per se*, but imposes a selection pressure that favors the accumulation of genes encoding enzymes that act in synergy to deconstruct complex polysaccharides.

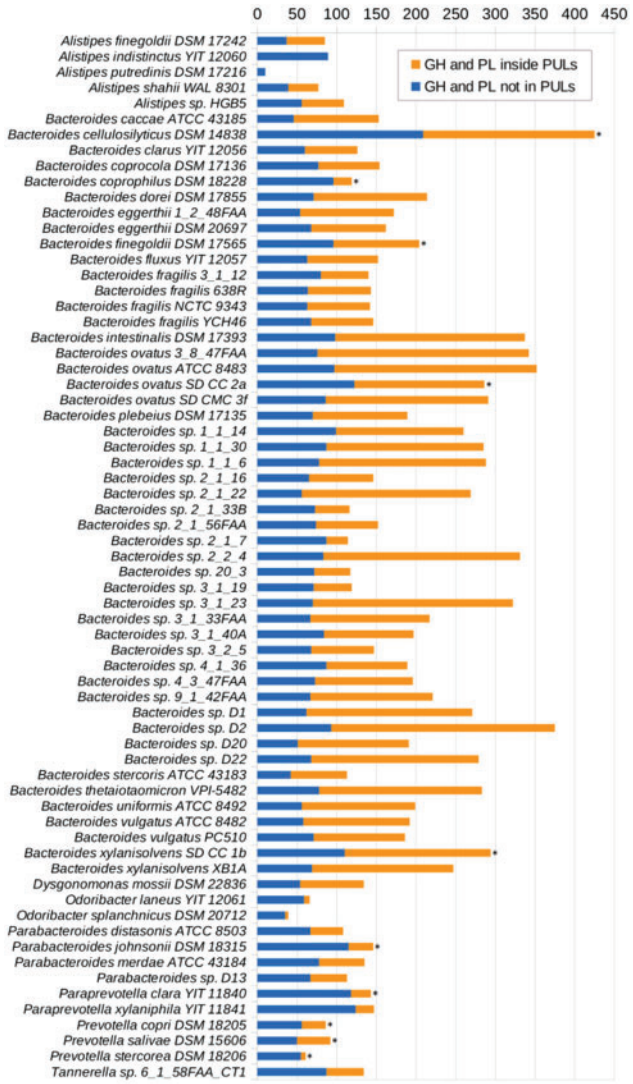

Fig. 5. GHs and protein lyases in the 67 Bacteroidetes genomes from the human gut microbiota. Distinction is made between those included in a predicted PUL (yellow) or not (blue). The 10 most fragmented genomes are denoted with an asterisk (Color version of this figure is available at *Bioinformatics* online.)

Odoribacter splanchnicus is characterized by an unconventional profile, with the highest number of PULs without CAZymes (60 of 64), and the highest proportion of PULs associated to regulators. Furthermore, *O.splanchnicus* displays a particular set of PUL regulators including 41 loci regulated by ECF-σ/anti-σ and none by HTCS. In comparison, to date, most other Bacteroidetes have more comparable numbers of these two regulator types (Supplementary Fig. S9).

3.3.2 Limits of the tandem susCD pair paradigm?

We have used *tandem susCD-like pairs* as the signature of PULs as these adjacent genes were found in all these loci. Ninety-nine of the 101 *tandem susCD-like pairs* identified in *B.thetaiotaomicron* belong to reference PULs, and 126 of 127 in *B.ovatus*. The only exceptions correspond to the BT2259-BT2260 and BT2263-BT2264 pairs in *B.thetaiotaomicron* and the BACOVA_00058-BACOVA_00059 pair in *B.ovatus*. Despite being adjacent and encoding characteristic Pfam domains of SusC- and SusD-like proteins, these loci were not listed in reference PULs by Martens *et al.* (2008, 2011).

We next tested the hypothesis that all PULs require a *tandem susCD-like pair*. In *B.thetaiotaomicron*, 86 of the 88 reference PULs (109 of 112 in *B.ovatus*) contain at least one *tandem susCD-like pair*. The exceptions are PUL27 and PUL63 in *B.thetaiotaomicron*, PUL66, PUL74 and PUL94 in *B.ovatus*, in which no *susD*-like gene is detected in the vicinity of a *susC*-like gene. In three of these five cases, the adjacent gene to the *susC*-like is annotated in reference (Martens *et al.*, 2008) as '*susD*-like', but the encoded product shows no detectable sequence similarity to other SusD-like proteins or their characteristic Pfam domains. It is possible that these genes fulfill a role similar to SusD homologs in spite of the lack of sequence relatedness. Hereafter, we utilize the term *susD*-like only when sequence similarity is demonstrable. Another explanation could be that the SusC-like protein alone is sufficient for glycan transport. Indeed, in each Bacteroidetes species, we observed the existence of ~10–20 *susC*-like genes, without an adjacent *susD*-like gene. SusC-like transporters SusC-like are found in many other bacterial phyla, including Gram-positive Firmicutes or Proteobacteria, and target various substrates distinct from glycans (Schauer *et al.*, 2008). This argues against the prediction of PULs based on the presence of only a *susC*-like gene. On the contrary, almost all *susD*-like genes are adjacent to a *susC*-like partner, except rare cases such as BT4085 in *B.thetaiotaomicron*, although the cognate gene is still a component of a PUL. This suggests that *susD*-like genes alone could be used as a basis for PUL prediction. In-depth examination of the *susD* genes in the 67 Bacteroidetes, revealed several additional *susD*-like genes without adjacent *susC*-like genes, which either correspond to: (i) *susD*-like gene at the scaffold border, likely reflecting poorly assembled genomes (two species accounting for 15 of 41 encountered cases), (ii) missing a *susC*-like gene model upstream of the *susD*-like gene despite detectable Pfam domains of *susC*-like genes (three species accounts for 32 cases of 41), (iii) split of the tandem SusCD pair by one intercalated gene, (iv) translocated *susD*, 2–3 genes upstream of the *susC*-like gene (12 cases highly conserved across species with the following organization: *susD*-gene1-gene2-*susC*-*susD*-*susC*), (v) duplicated *susD*-like gene close to a tandem SusCD-like pair and (vi) no *susC*-like sequence in the vicinity of the *susD*-like gene. All these cases currently lack biological characterization and represent only

3% of the 4141 standard *tandem susCD-like pairs* in the 67 Bacteroidetes genomes. If these loci respond to polysaccharides, it would be interesting to verify if they are expressed alone or if they are part of a co-regulated network (Ravcheev *et al.*, 2013). These cases will require careful analyses to elucidate their biological relevance, which may in turn allow the refinement of the PUL paradigm.

4 Conclusion

By combining observed PUL structures derived from experimental investigations of two Bacteroidetes species, we were able to develop a tool for the automatic prediction of PULs solely based on genome sequence. The percentage of correctly predicted PULs was 87% in *B.thetaiotaomicron* and 91% in *B.ovatus*, with only a few false predictions that represent <8% of the true positives. We applied our approach to compute 3745 PUL predictions for 67 Bacteroidetes species from the human gut microbiota. Varying numbers of PULs (from 1 to 117) were observed in the different species, with a majority of PULs containing at least two GH-encoding genes, and around equal proportions of the two considered regulator types. The results are accessible to the scientific community via the interactive web-interface PULDB at www.cazy.org/PULDB/index.php.

Because our method is based solely on genomic information, the accuracy of the PUL delineation depends on the quality of the genome assemblies and gene predictions. With appropriately assembled genomes, our method can provide high-throughput prediction of classical PULs in species without expression data, or in non-cultivated species subject to single cell sequencing. We expect that experimental studies will benefit from these predicted PULs for the investigation of the targeted substrates, and the evolutionary analysis of glycan breakdown in a wide range of species. Iteratively, our automatic PUL-prediction strategy will be refined by integrating additional knowledge of the PUL paradigm. In addition, a human curation step by the CAZy team is planned to correct any erroneous predictions that would appear following the experimental characterization of a PUL. Finally, future developments will also include Bacteroidetes from environmental species to explore the genetic architecture and function of the PULs in these organisms.

Funding

European Research Council under the European Union's Seventh Framework Program (FP/2007-2013)/ERC Grant Agreement no 322820.

Conflict of Interest: none declared.

References

Apic,G. *et al.* (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.*, **310**, 311–325.

Bockhorst,J. *et al.* (2003) Predicting bacterial transcription units using sequence and expression data. *Bioinformatics*, **19** (Suppl. 1), i34–i43.

Cameron,E.A. *et al.* (2012) Multidomain carbohydrate-binding proteins involved in *Bacteroides thetaiotaomicron* starch metabolism. *J. Biol. Chem.*, **287**, 34614–34625.

Dandekar,T. *et al.* (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **23**, 324–328.

El Kaoutari,A. *et al.* (2013) The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nat. Rev. Microbiol.*, **11**, 497–504.

Enright,A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Ermolaeva,M.D. *et al.* (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res.*, **29**, 1216–1221.

Finn,R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**(Database issue), D222–D230.

Flint,H.J. *et al.* (2012) Microbial degradation of complex carbohydrates in the gut. *Gut Microbes*, **3**, 289–306.

Hegyi,H. and Gerstein,M. (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.*, **11**, 1632–1640.

Kitamura,M. *et al.* (2008) Structural and functional analysis of a glycoside hydrolase family 97 enzyme from *Bacteroides thetaiotaomicron*. *J. Biol. Chem.*, **283**, 36328–36337.

Koebnik,R. (2005) TonB-dependent trans-envelope signalling: the exception or the rule? *Trends Microbiol.*, **13**, 343.

Koropatkin,N.M. and Smith,T.J. (2010) SusG: a unique cell-membrane-associated alpha-amylase from a prominent human gut symbiont targets complex starch molecules. *Structure*, **18**, 200–215.

Koropatkin,N.M. *et al.* (2008) Starch catabolism by a prominent human gut symbiont is directed by the recognition of amylose helices. *Structure*, **16**, 1105–1115.

Laine,R.A. (1994) A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 x 10(12) structures for a reducing hexasaccharide: the Isomer Barrier to development of single-method saccharide sequencing or synthesis systems. *Glycobiology*, **4**, 759–767.

Larsbrink,J. *et al.* (2014) A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. *Nature*, **506**, 498–502.

Ley,R.E. *et al.* (2006) Microbial ecology: human gut microbes associated with obesity. *Nature*, **444**, 1022–1023.

Lombard,V. *et al.* (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.*, **42**(Database issue), D490–D495.

Manfredi,P. *et al.* (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.

Mao,F. *et al.* (2009) DOOR: a database for prokaryotic operons. *Nucleic Acids Res.*, **37**(Database issue), D459–D463.

Marcotte,E.M. *et al.* (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753.

Markowitz,V.M. *et al.* (2012) IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res.*, **40**(Database issue), D123–D129.

Martens,E.C. *et al.* (2008) Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host Microbe*, **4**, 447–457.

Martens,E.C. *et al.* (2009) Coordinate regulation of glycan degradation and polysaccharide capsule biosynthesis by a prominent human gut symbiont. *J. Biol. Chem.*, **284**, 24673–24677.

Martens,E.C. *et al.* (2011) Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS Biol.*, **9**, e1001221.

McBride,M.J. *et al.* (2009) Novel features of the polysaccharide-digesting gliding bacterium *Flavobacterium johnsoniae* as revealed by genome sequence analysis. *Appl. Environ. Microbiol.*, **75**, 6864–6875.

McNeil,N.I. (1984) The contribution of the large intestine to energy supplies in man. *Am. J. Clin. Nutr.*, **39**, 338–342.

McNulty,N.P. *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.

Moore,A.D. *et al.* (2014) DoMosaics: software for domain arrangement visualization and domain-centric analysis of proteins. *Bioinformatics*, **30**, 282–283.

Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18** (Suppl. 1), S329–S336.

Ravcheev,D.A. *et al.* (2013) Polysaccharides utilization in human gut bacterium *Bacteroides thetaiotaomicron*: comparative genomics reconstruction of metabolic and regulatory networks. *BMC Genomics*, **14**, 873.

Salgado,H. *et al.* (2000) Operons in *Escherichia coli*: genomic analyses and predictions. *Proc. Natl Acad. Sci. USA*, **97**, 6652–6657.

Schauer,K. *et al.* (2008) New substrates for TonB-dependent transport: do we only see the 'tip of the iceberg'? *Trends Biochem. Sci.*, **33**, 330–338.

Skinner,M.E. *et al.* (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.

Taboada,B. *et al.* (2012) ProOpDB: Prokaryotic Operon DataBase. *Nucleic Acids Res.*, **40**(Database issue), D627–D631.

Terrapon,N. and Henrissat,B. (2014) How do gut microbes break down dietary fiber? *Trends Biochem. Sci.*, **39**, 156–158.

Terrapon,N. *et al.* (2009) Detection of new protein domains using co-occurrence: application to *Plasmodium falciparum*. *Bioinformatics*, **25**, 3077–3083.

Westover,B.P. *et al.* (2005) Operon prediction without a training set. *Bioinformatics*, **21**, 880–888.