

Complex pectin metabolism by gut bacteria reveals novel catalytic functions

Didier Ndeh^{1*}, Artur Rogowski^{1†*}, Alan Cartmell^{1*}, Ana S. Luis^{1*}, Arnaud Baslé¹, Joseph Gray¹, Immacolata Venditto¹, Jonathon Briggs¹, Xiaoyang Zhang¹, Aurore Labourel¹, Nicolas Terrapon², Fanny Buffetto^{3†}, Sergey Nepogodiev⁴, Yao Xiao⁵, Robert A. Field⁴, Yanping Zhu⁶, Malcolm A. O'Neill⁶, Breeanna R. Urbanowicz⁶, William S. York⁶, Gideon J. Davies⁷, D. Wade Abbott⁸, Marie-Christine Ralet³, Eric C. Martens⁵, Bernard Henrissat^{2,9,10} & Harry J. Gilbert¹

The metabolism of carbohydrate polymers drives microbial diversity in the human gut microbiota. It is unclear, however, whether bacterial consortia or single organisms are required to depolymerize highly complex glycans. Here we show that the gut bacterium *Bacteroides thetaiotaomicron* uses the most structurally complex glycan known: the plant pectic polysaccharide rhamnogalacturonan-II, cleaving all but 1 of its 21 distinct glycosidic linkages. The deconstruction of rhamnogalacturonan-II side chains and backbone are coordinated to overcome steric constraints, and the degradation involves previously undiscovered enzyme families and catalytic activities. The degradation system informs revision of the current structural model of rhamnogalacturonan-II and highlights how individual gut bacteria orchestrate manifold enzymes to metabolize the most challenging glycan in the human diet.

We have consumed plants containing rhamnogalacturonan-II (RG-II) since the archaic humans¹. This branched pectic polysaccharide is the most complex glycan known, containing 13 different sugars and 21 distinct glycosidic linkages arranged as a backbone with different side chains termed A–F² (Fig. 1). The abundance of RG-II in red wine and other processed pectins³ reflects its recalcitrance to degradation. The polysaccharide, however, is subject to biological degradation as it does not accumulate in the environment, but the organisms that use RG-II and the mechanism of depolymerization remain unknown. Dietary glycans are the predominant nutrients available to the human gut microbiota (HGM)^{4–6} and are major drivers in defining its community structure⁷. As RG-II has been a component of the human diet over hundreds of thousands of years, we hypothesize that the glycan exerts a selection pressure on the HGM leading to the evolution of organisms that degrade the pectic carbohydrate, consistent with initial studies suggesting that *Bacteroides thetaiotaomicron* (HGM bacterium) can use at least components of the glycan⁸. To test this hypothesis, the mechanism by which *B. thetaiotaomicron* degrades RG-II was determined.

Growth of HGM *Bacteroides* on RG-II

To explore whether single organisms grow on RG-II, the glycan was incubated with the type strains of the *Bacteroides* species of the HGM. Approximately 30% of these organisms grew on the glycan (Fig. 2a). Only monosaccharides and 2-*O*-methyl-*D*-xylose- α 1,3-*L*-fucose remained in stationary-phase cultures (Fig. 2b and Supplementary Table 1), indicating that the bacteria cleaved 20 out of the 21 distinct glycosidic linkages in the polysaccharide. Transcriptomic analysis of one of these species, *B. thetaiotaomicron*, showed that RG-II upregulates three polysaccharide utilization loci (PULs), RG-II PUL1–PUL3 (ref. 8 and Extended Data Fig. 1a). Activities of recombinant enzymes encoded by these loci were determined using RG-II or RG-II-derived

oligosaccharides (Extended Data Fig. 2 and Methods). The data revealed the degradative pathway for each oligosaccharide, leading to a model for RG-II depolymerization (Fig. 1).

Specificity of enzymes that cleave RG-II

Each glycosidic linkage in RG-II, except for 2-*O*-methyl-*D*-xylose- α 1,3-*L*-fucose, is hydrolysed by a bespoke enzyme (Figs 1 and 2b). Thus, although the loci encode multiple α -*L*-rhamnosidases and *L*-arabinosidases, there is no redundancy in the linkage hydrolysed by these enzymes (Supplementary Tables 2 and 3). This stringent enzymatic specificity probably reflects an apparatus that is tuned to maximize the rate of the degradative process, offsetting the energy required to synthesize such an elaborate catabolic system.

The RG-II degrading system (RG-II degradome) contains three enzymes each comprising two distinct catalytic modules. BT0996 contains a β -*D*-glucuronidase and a β -*L*-arabinofuranosidase that target chains A and B, respectively (Supplementary Table 3), indicating that degradation of the oligosaccharide decorations is coordinated. BT1013 and BT1020 hydrolyse the two linkages in chains C and D, respectively (Extended Data Fig. 3 and Supplementary Discussion 2). The crystal structure of BT1020 (Extended Data Fig. 4a) reveals an N-terminal 5-bladed β -propeller that displays 2-keto-3-deoxy-*D*-lyxo-heptulosaric acid (DHA)-hydrolase activity and a C-terminal $(\alpha/\alpha)_6$ -barrel β -*L*-arabinofuranosidase. The β -*L*-arabinofuranosidase domain (in complex with arabinose) contains a canonical glycoside hydrolase catalytic apparatus comprising two carboxylate residues. In the DHA-hydrolase active site, a tyrosine and a glutamate residue function as the catalytic nucleophile and acid–base residues, respectively resembling the catalytic centre of sialidases that also act on ulosonic acids⁹. The highly basic nature of the active sites of BT1020 is consistent with the negatively charged species that occupy the two catalytic centres (Supplementary Discussion 3.1).

¹Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne NE2 4HH, UK. ²Architecture et Fonction des Macromolécules Biologiques, Centre National de la Recherche Scientifique (CNRS), Aix-Marseille University, F-13288 Marseille, France. ³INRA, UR1268 Biopolymères Interactions Assemblages, 44300 Nantes, France. ⁴Department of Biological Chemistry, John Innes Centre Norwich Research Park, Norwich NR4 7UH, UK. ⁵Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, Michigan 48109 USA. ⁶Complex Carbohydrate Research Center, The University of Georgia, 315 Riverbend Road, Athens, Georgia 30602, USA. ⁷Department of Chemistry, University of York, York YO10 5DD, UK. ⁸Lethbridge Research Centre, Lethbridge, Alberta T1J 4B1, Canada. ⁹INRA, USC 1408 AFMB, F-13288 Marseille, France. ¹⁰Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia. †Present addresses: Megazyme, Bray, Co. Wicklow, A98 YV29, Ireland (A.R.); Institute for Wine Biotechnology, Department of Viticulture and Oenology, Stellenbosch University, Matieland 7602, South Africa (F.B.).

*These authors contributed equally to this work.

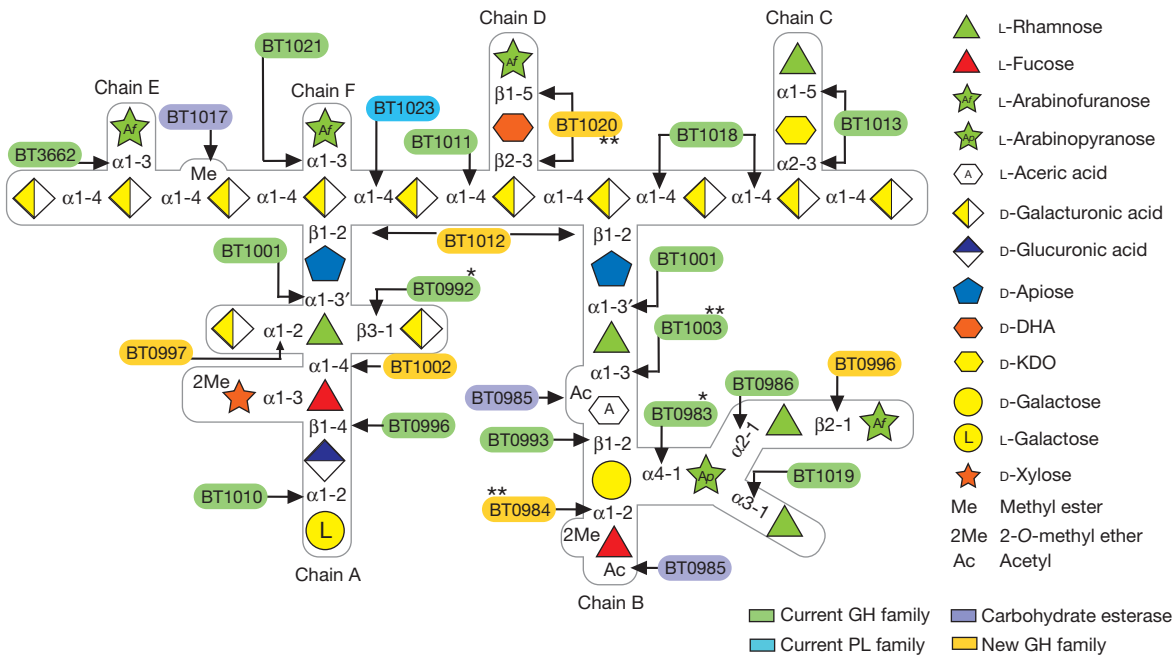


Figure 1 | Schematic of enzymes and PULs involved in RG-II degradation. Sugars shown using the Consortium for Functional Glycomics notation²⁵. Enzymes are appropriately colour-coded. Single asterisks denote a new activity for a glycoside hydrolase (GH) family;

double asterisks denote enzymes with activities that have not previously been observed. D-KDO, 3-deoxy-D-manno-octulosonic acid; D-DHA, 2-keto-3-deoxy-D-lyxo-heptulosaric acid; PL, polysaccharide lyase.

The CAZy database groups glycoside hydrolases into sequence-based families (designated GHXX)¹⁰. Predicting specificity based on family classification, however, is challenging since glycoside hydrolase

families are often poly-specific, with only around 2% of the enzymes characterized¹⁰. Analyses of enzymes that deconstruct RG-II advance the functional annotation of existing glycoside hydrolase families.

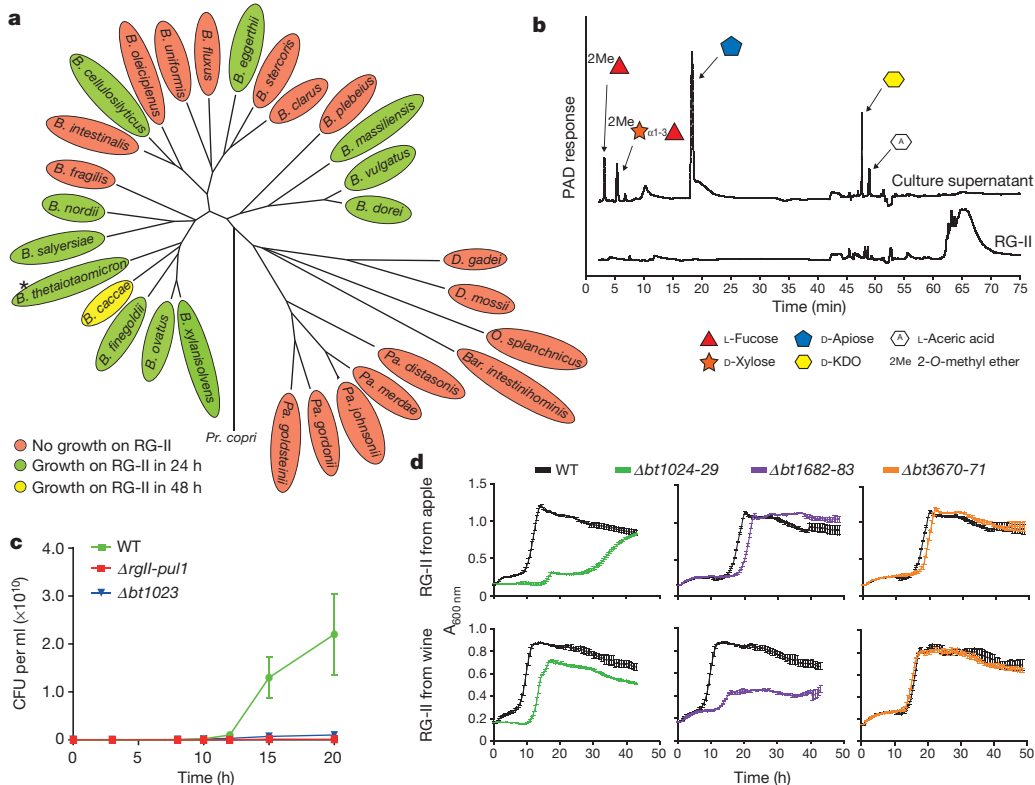


Figure 2 | Growth of Bacteroidetes species on RG-II. **a**, Growth of strains of HGM Bacteroidetes on RG-II and phylogeny of the organisms (biological replicates, $n = 6$). **b**, High performance anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD) analysis of stationary-phase culture of *B. thetaiotaomicron* (denoted by asterisk in **a**). **c**, Wild-type (WT) *B. thetaiotaomicron* and mutants lacking RG-II PUL1 ($\Delta rgII-pul1$) or

the polysaccharide lyase BT1023 ($\Delta bt1023$) were inoculated at 10^7 colony forming units (CFU) per ml into RG-II media, and CFU were determined over time. **d**, Wild-type *B. thetaiotaomicron* and mutants in which *suscI-susdI* pairs had been deleted were cultured on RG-II from red wine (bottom) or apple juice (top) and growth was monitored every 20 min (biological replicates $n = 6$, error bars denote s.e.m.).

The GH2 β -D-galacturonidase BT0992, GH2 α -arabinopyranosidase BT0983, and GH127 3-C-carboxy-5-deoxy-L-xylose (aceric acid)-hydrolase BT1003 (Fig. 3 and Extended Data Fig. 5) represent new activities for their respective families (Supplementary Table 3). The crystal structure of the α -L-rhamnosidase BT0986 provides unique insights into the mechanism of substrate recognition and catalysis for a GH106 enzyme¹¹ (Extended Data Fig. 6). This 1,100-residue protein contains an N-terminal $(\alpha/\beta)_8$ -barrel catalytic domain. The heptaoligosaccharide generated by the mutant Δ bt0986 is bound in the centre of the $(\alpha/\beta)_8$ -barrel in a funnel-like structure extending from the active site pocket. Mutagenesis studies revealed the BT0986 catalytic residues (Supplementary Table 4). Glu593, which is 5 Å from the anomeric carbon, probably activates a water molecule that mounts a nucleophilic attack at C1, thus the glutamate is predicted to be the catalytic base. Glu461 is within hydrogen-bonding distance of the glycosidic oxygen and is the candidate catalytic acid. In the active site, glutamate residues coordinate calcium, which makes polar interactions with O2 and O3 of the L-rhamnopyranose (L-Rhap). The importance of calcium is illustrated by the loss of activity after the addition of EDTA or when the glutamate residues that coordinate the metal were mutated (Supplementary Table 4 and Supplementary Discussion 3.2). The essentiality of calcium has resonance with other inverting enzymes that target α -manno-configured linkages, which often exploit divalent metal ions in substrate binding¹².

The aceric acid hydrolase activity of BT1003 is intriguing as GH127 was, before this work, populated only by β -L-arabinofuranosidases¹³. The crystal structure and mutagenesis of BT1003 (Extended Data Fig. 7a and Supplementary Table 4) showed that the proposed catalytic acid-base of GH127 arabinofuranosidases is not conserved in BT1003, suggesting substrate participation in glycosidic bond cleavage (Supplementary Discussion 3.5 and Supplementary Fig. 1).

Founding members of new glycoside hydrolase families

The RG-II degradome contains seven enzymes that reveal new glycoside hydrolase families (Fig. 1 and Supplementary Table 3), comprising an endo-apiosidase (BT1012, GH140; Supplementary Fig. 2), DHA-hydrolase (N terminus of BT1020, GH143), two β -L-arabinofuranosidases (C terminus of BT1020, GH142; N terminus of BT0996, GH137), α -2-O-methyl-L-fucosidase (BT0984, GH139), α -L-fucosidase (BT1002, GH141) and α -galacturonidase (BT0997, GH138). BT1017 represents a new family of pectin methyl-esterases (Extended Data Fig. 8b). Additional features of RG-II depolymerization are the α -2-O-methyl-fucosidase, DHA-hydrolase and aceric acid hydrolase activities, which have not previously been described even though α -2-O-methyl-fucosidase and DHA exist in other glycans^{15,16}. The widespread presence of these newly discovered enzyme families in Bacteroidetes and other bacterial phyla, and the occurrence of GH139, GH140, GH141 and GH142 in fungi, suggest that these enzymes contribute to glycan degradation in diverse environments.

The generic relevance of RG-II PUL1 in RG-II utilization was supported by bioinformatic analysis. Xenologues of RG-II PUL1 proteins were identified by reciprocal best-BLAST hits in the HGM *Bacteroides* species (Extended Data Fig. 1b). Only species that grow on RG-II contain proteins that display 65–95% identity with the *B. thetaiotaomicron* enzymes. Of the 372 non-HGM *Bacteroidetes* species that were searched for candidate RG-II degradomes, 40 contained PULs that encoded xenologues of all the necessary enzyme activities for RG-II breakdown (Supplementary Table 5). Example organisms include *Prevotella bryantii* (bovine rumen), *Flavobacterium johnsoniae* and *Niabella soli* (soil). This suggests that these species may also depolymerize the glycan (Extended Data Fig. 1b and Supplementary Fig. 3).

A model for RG-II degradation

A hierarchical model for the RG-II degradome is proposed (Fig. 4). Chains C and E, the L-arabinofuranose (L-Araf) of chain D, and the

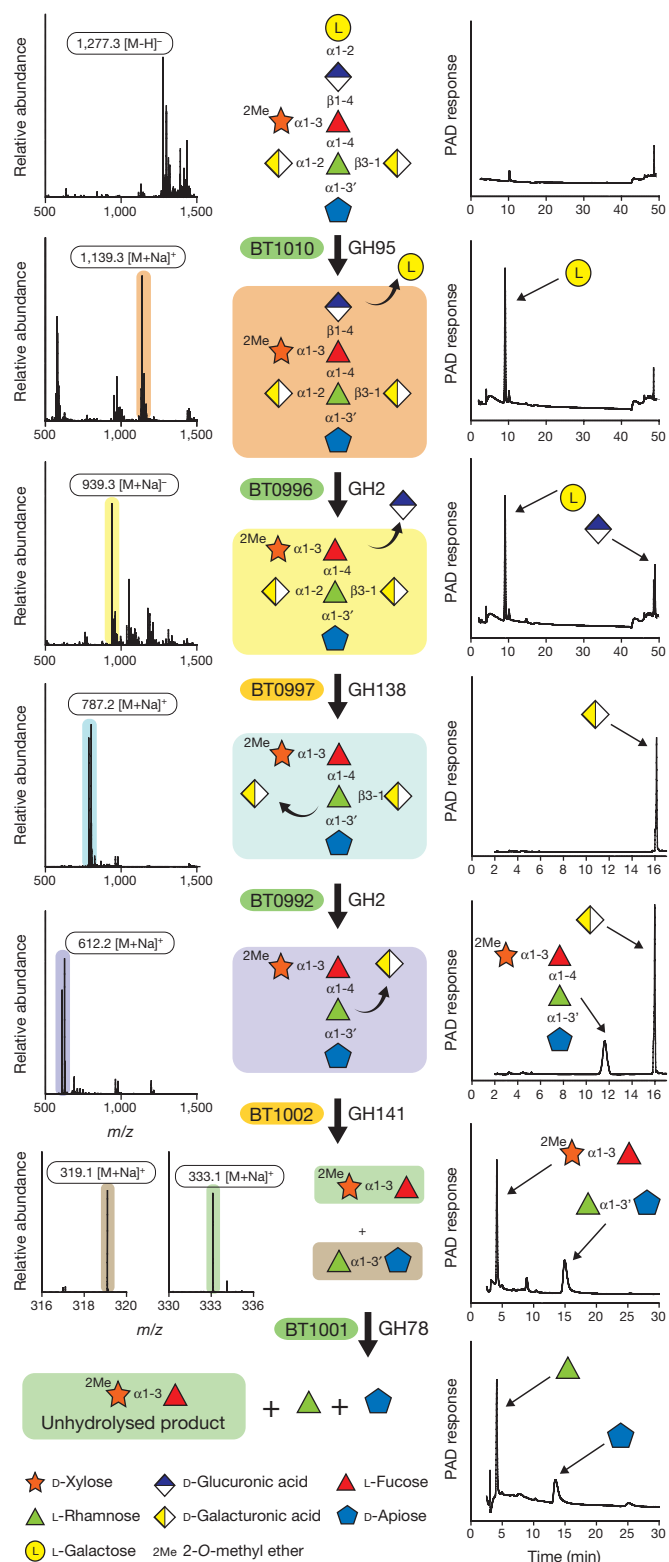


Figure 3 | *B. thetaiotaomicron*-mediated depolymerization of RG-II chain A. The substrates comprised chain A released from RG-II by trifluoroacetic acid and the oligosaccharides generated by mutants of *B. thetaiotaomicron* in which *bt0997*, *bt0992* or *bt1002* had been deleted. Individual proteins (1 μ M) were incubated with the glycans (5 mM) for 16 h at 37 °C in 20 mM sodium phosphate buffer, pH 7.0. Monosaccharides and oligosaccharides generated were identified by HPAEC-PAD and electrospray ionization mass spectrometry (ESI-MS), respectively. Verification of the model was achieved by reconstituting the pathway using the six enzymes together, which showed that the glycoside hydrolases only functioned in the order shown in the figure. The example is from technical replicates $n = 3$.

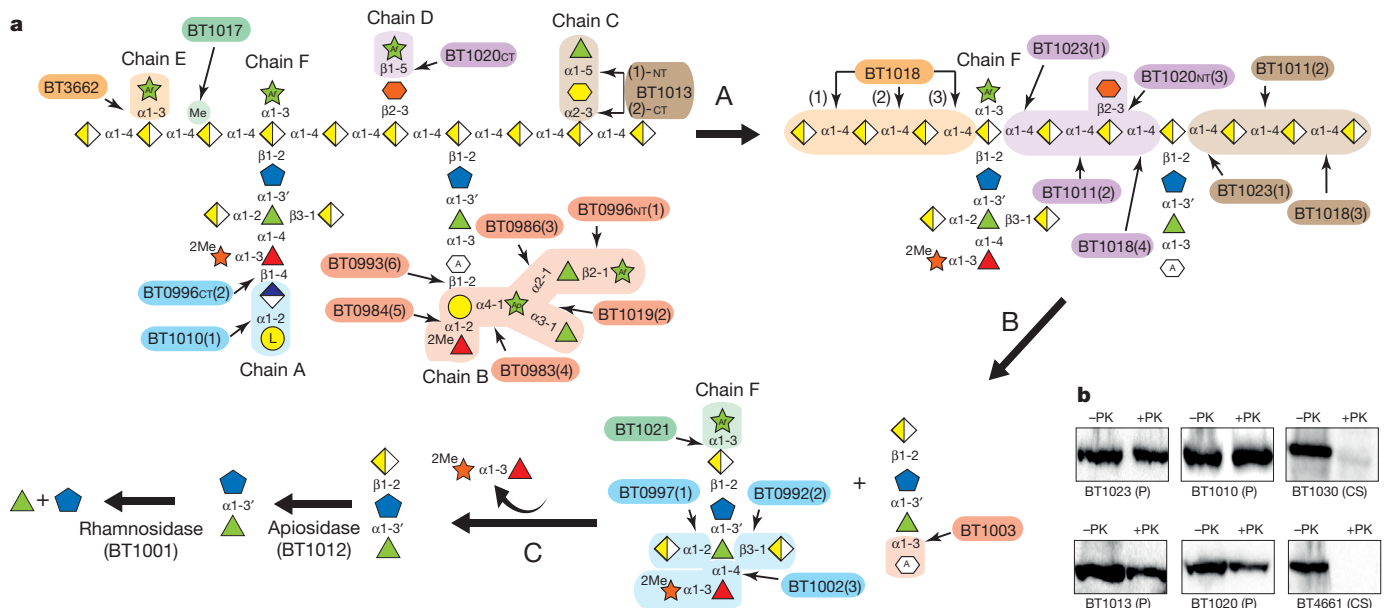


Figure 4 | Model of RG-II disassembly by *B. thetaiotaomicron*. **a**, The degradative model. Enzyme cohorts that degrade a specific region of RG-II are coloured the same, and values in parentheses indicate the predicted order in which they act, based on data (technical replicates $n = 3$) in Supplementary Table 6. Symbols as in Fig. 1. Reaction 'A' denotes initial hierarchical exo cleavage; reaction 'B' denotes backbone removal; reaction

terminal region of chains A and B are cleaved without hydrolysis of the glycan backbone (Supplementary Table 6). The steric constraints imposed by RG-II are consistent with this incomplete degradation. We propose that cleavage of the backbone increases enzyme access, enabling further degradation of the side chains. The complete deconstruction of these side chains is tightly linked to disassembly of the backbone. BT1023, a polysaccharide lyase, initiates backbone cleavage, which is crucial for RG-II utilization as the $\Delta bt1023$ mutant grew poorly on the glycan (Fig. 2c). The enzyme cleaved only the homogalacturonan within RG-II (Extended Data Fig. 8a), indicating that side chains are specificity determinants for BT1023. The completion of backbone degradation is mediated by removal of methyl-esters (BT1017) and DHA (BT1020) (Extended Data Figs 3c, d and 8b) in harness with *exo*-acting glycosidases (Supplementary Table 3). Enzyme cell-localization studies indicate that depolymerization is exclusively periplasmic (Fig. 4b and Supplementary Discussion 4).

After backbone depolymerization, D-apiofuranose (D-Apif) of chains A and B, linked to the remaining D-galacturonic acids (D-GalAs), are released by BT1012 before cleavage of L-Rhap-D-Apif by BT1001 (Supplementary Table 6). Thus, features of the α -L-rhamnosidase distal to the +1 subsite prevent binding of substrates with a degree of polymerization >2 . The crystal structure of the apiosidase (BT1012; Extended Data Fig. 7b, i) reveals an $(\alpha/\beta)_8$ -barrel catalytic domain and a C-terminal β -sandwich domain. The predicted catalytic residues, Asp187 and Glu284, which are essential for activity (Supplementary Table 4), are separated by approximately 5.5 Å, consistent with a double-displacement mechanism and retention of anomeric configuration (confirmed experimentally, Extended Data Fig. 7b, ii). Two arginine residues in the +1 subsite probably contribute to D-GalA binding through formation of ion pairs, consistent with mutagenesis data (Supplementary Table 4). The -1 (active site) and +1 subsites are narrow, preventing binding of D-Apif linked to borate and D-GalA appended to the backbone at O-1 or O-4 (Extended Data Fig. 7b, iii), explaining why the apiosidase functions late in the degradative process (Fig. 4a).

The enzymes that depolymerize chains A and B cleave their target sugars only when they are at the non-reducing termini (Supplementary Table 6). Thus, both chains are depolymerized through sequential-acting

'C' denotes the final hierarchical exo cleavage. **b**, The cellular localization of key RG-II degrading enzymes based on their resistance to proteinase K when expressed by *B. thetaiotaomicron*. BT4661 and BT1030 are known surface glycan binding proteins. The example is from biological replicates $n = 3$. CS, cell surface; P, periplasm. +PK and -PK denote with and without proteinase K treatment, respectively.

exo-glycosidases. Of particular note is BT1002, an α -L-fucosidase that targets fucose substituted at O-3 with 2-O-methyl- α -D-xylose (Fig. 3). The active site pocket of the hydrolase is likely to be considerably more open than in GH95 and GH29 fucosidases that are unable to tolerate O-3 substitutions^{17,18}. This is consistent with the topology of the catalytic centre of BT1002, which comprises an extended pocket (Extended Data Fig. 7c), which may bind the disaccharide 2-O-methyl-D-xylose- α 1,3-L-fucose (Supplementary Discussion 3.3).

Removal of the borate steric barrier

RG-II *in planta* is a dimer mediated by a borate diester linkage between D-Apif in chain A of each monomer¹⁵. How the RG-II degradome overcomes the steric constraints imposed by this dimer is a crucial question. Previous studies¹⁹ proposed that the removal of L-galactose-D-glucuronic acid (L-Gal-D-GlcA) destabilizes borate-mediated dimerization. We show that chain A-derived oligosaccharides, lacking the terminal L-Gal and D-GlcA, contained D-Apif at their reducing end (Extended Data Fig. 2). By contrast, $\Delta bt1010$ generated intact chain A with D-GalA attached to D-Apif. These data indicate that the apiosidase was only active against chain A after L-Gal and D-GlcA had been removed, and thus the loss of this disaccharide destabilized the interaction of D-Apif with borate. The apiosidase was not active against chain A in the $\Delta bt1010$ mutant, which lacks a functional α -L-galactosidase, indicating that borate remained bound to D-Apif, preventing access to the apiosidase. *In vitro* analysis showed that the activity of the apiosidase against intact chain A was sensitive to borate (Extended Data Fig. 9b, d) but the oxyanion did not inhibit the enzyme against L-Rha-D-Api-D-GalA-Me (Extended Data Fig. 9c). These data indicate that borate only bound to D-Apif in full-length chain A, explaining how initial degradation of this side chain relieved steric constraints imposed by the oxyanion.

The species-dependent variation of the terminal region of chain B in RG-II^{19,20}, (Supplementary Discussion 1.0) may help to explain why RG-II PUL3 encodes several putative glycoside hydrolases that could contribute to degradation of all RG-II structures. Thus, BT3662 removed chain E (Fig. 1 and Supplementary Table 2), present in grape RG-II but absent in the glycan from other plants. The crystal structure

of BT3662, which hydrolyses α -Araf linkages (Extended Data Fig. 4b), comprises an N-terminal five-bladed β -propeller catalytic domain. The active site pocket, which contains all the features of a typical GH43 α -L-arabinofuranosidase²¹, abuts onto a channel containing Tyr199 in the +1 subsite and two arginine residues in the distal regions. Mutagenesis data (Supplementary Table 4) suggest the basic residues sequester the acidic backbone of RG-II into the substrate binding cleft, facilitating optimal interactions of the leaving group D-GalA with Tyr199. RG-II PUL2 only encodes a SusC_h/SusD_h pair (outer membrane glycan transporter)²², BT1683 and BT1682. The Δ bt1683/ Δ bt1682 mutant grew on apple but not wine RG-II, whereas the Δ bt1024/ Δ bt1029 variant (deletion of *susc_h/susd_h* in RG-II PUL1) metabolized the glycan from wine but slowly from apple (Fig. 2d). These data suggest that the multiple SusC_h/SusD_h pairs encoded by the RG-II PULs are required to import plant-specific variants of RG-II.

New features of RG-II structure

The specificity of the RG-II degradome shows that the current model of the glycan requires revision. BT1001, which cleaves the Rhap–Apif linkage in chains A and B, was shown to be an α -L-rhamnosidase (Extended Data Fig. 10a, Supplementary Tables 3 and 7). Thus, the L-Rhap–D-Apif linkage in chains A and B is α and not β as previously reported¹⁹. The change in the stereochemistry of this L-Rhap linkage is likely to have a substantial effect on our knowledge of the conformations adopted by RG-II.

Analysis of the RG-II degradome revealed a new side chain (chain F) consisting of an α -Araf linked to O-3 of the backbone GalA that is appended to chain A at O-2. This arabinose was evident in the oligosaccharides generated by *B. thetaiotaomicron* mutants Δ bt1017 (Extended Data Fig. 8b, c), Δ bt1021, Δ bt1010/ Δ bt1021 and Δ bt0986/ Δ bt1021 (Extended Data Fig. 10b, Supplementary Discussion 5). Chain E, also comprising a single arabinose linked to the backbone, is in a different context to chain F as the two pentose sugar decorations are removed by BT3662 (chain E) and BT1021 (chain F), respectively (Supplementary Tables 3 and 6). Although methyl-esterification of backbone D-GalA(s) is known, the location and extent of these decorations were unclear²³. The Δ bt1017 mutant, (which lacks the pectin methyl-esterase), generated an oligosaccharide with a methyl-esterified D-GalA linked to O-4 of the D-GalA decorated with chains A and F (Extended Data Fig. 8b, c). This suggests that methyl-esterification occurs on a specific backbone D-GalA proximal to chains A and F.

The crystal structure of the β -L-arabinofuranosidase of BT0996 revealed electron density for chain B (Fig. 5 and Supplementary Discussion 3.4). The conformation of the oligosaccharide was stabilized by polar interactions within the nonasaccharide and confirmed the α -linkage between L-Rhap and D-Apif (Fig. 5d). Notably, the L-Arap, substituted at O-2 and O-3 by L-Rhap, was in the unusual ¹C₄ conformation, consistent with NMR data²⁴. This unfavourable conformation is stabilized by polar interactions between the O3-linked L-Rhap and the rest of the chain. Removing this α -L-1,3-Rhap by BT1019 probably allows the Arap to adopt the normal ⁴C₁ conformer and is thus a substrate for the GH2 α -arabinopyranosidase BT0983, which hydrolyses equatorial glycosidic bonds. The relaxed ⁴C₁ conformation of Arap was captured in the structure of the chain B-derived unbranched heptasaccharide (lacks the O3-linked L-Rhap) bound to BT0986 (Extended Data Fig. 6d–f).

This study reveals the elaborate and highly specific enzyme system by which *B. thetaiotaomicron* uses RG-II, a highly complex glycan. Our data show that RG-II PUL1 encodes several proteins previously annotated as ‘hypothetical’ that are the founding members of new glycoside hydrolase and esterase families. The discovery of seven new enzyme families, with catalytic functions not previously reported, and new specificities for existing glycoside hydrolase families illustrates the novelty of the RG-II degradome, and how the unique structural features of this glycan have driven the evolution of new catalysts. This

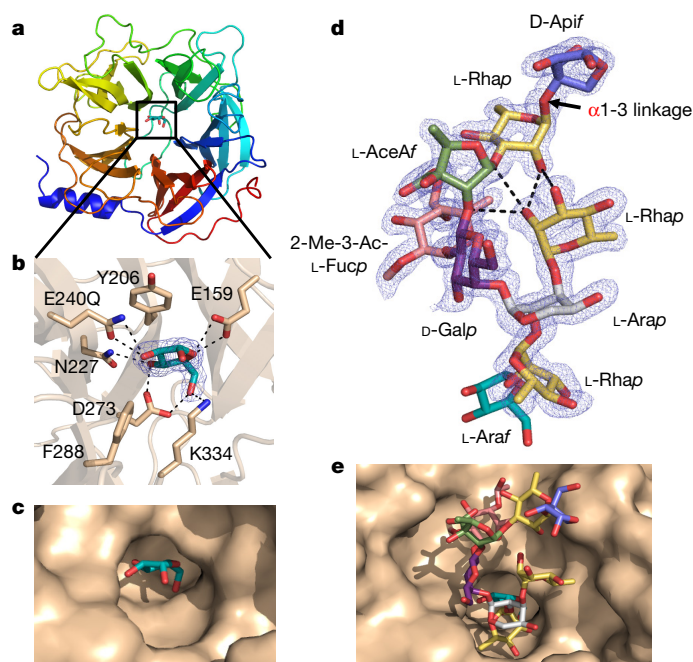


Figure 5 | Crystal structure of N-terminal catalytic domain of BT0996 (BT0996-N) in complex with RG-II chain B. **a**, Schematic of BT0996-N, coloured from the N- (blue) to the C- (red) termini. Black box denotes L-Araf-containing active site. **b**, Residues interacting with L-Araf, which include the putative catalytic amino acids (Glu240 (shown as Gln240) and Glu159). **c**, Surface representation of the active-site pocket containing L-Araf. **d**, Crystal structure of chain B in complex with BT0996-N. The α -linkage between L-Rhap and D-Apif is shown in red. **e**, Chain B (sugars coloured as in **d**) in the funnel-like substrate-binding site of BT0996-N. Polar interactions in **c** and **d** are broken black lines, and the blue mesh is the electron density map ($2F_o - F_c$) of the ligands at 1.5 σ .

contrasts with recent studies of the glycan-degrading apparatus of other HGM organisms, which exclusively use enzymes from existing CAZy families^{21–25}. The model of RG-II degradation proposed, coupled with the unveiling of new enzyme families and activities, now provides a framework to discover and dissect pectic polysaccharide degradation in environments extending beyond the human gut microbiota. The true extent of RG-II degradation in nature can now be accessed.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 November 2016; accepted 27 February 2017.

Published online 22 March 2017.

1. Wißing, C. *et al.* Isotopic evidence for dietary ecology of late Neandertals in North-Western Europe. *Quat. Int.* **411**, 327–345 (2016).
2. Pellerin, P. *et al.* Structural characterization of red wine rhamnoglacturonan II. *Carbohydr. Res.* **290**, 183–197 (1996).
3. Apolinar-Valiente, R. *et al.* Polysaccharide composition of Monastrell red wines from four different Spanish terroirs: effect of wine-making techniques. *J. Agric. Food Chem.* **61**, 2538–2547 (2013).
4. Cuskin, F. *et al.* Human gut Bacteroidetes can utilize yeast mannan through a selfish mechanism. *Nature* **517**, 165–169 (2015).
5. Larsbrink, J. *et al.* A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. *Nature* **506**, 498–502 (2014).
6. Rogowski, A. *et al.* Glycan complexity dictates microbial resource allocation in the large intestine. *Nat. Commun.* **6**, 7481 (2015).
7. Koropatkin, N. M., Cameron, E. A. & Martens, E. C. How glycan metabolism shapes the human gut microbiota. *Nat. Rev. Microbiol.* **10**, 323–335 (2012).
8. Martens, E. C. *et al.* Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS Biol.* **9**, e1001221 (2011).
9. Amaya, M. F. *et al.* Structural insights into the catalytic mechanism of *Trypanosoma cruzi* trans-sialidase. *Structure* **12**, 775–784 (2004).

10. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495 (2014).
11. Davis, B. G. *et al.* Tetrazoles of manno- and rhamno-pyranoses: Contrasting inhibition of mannosidases by 4.3.0 but of rhamnosidase by 3.3.0 bicyclic tetrazoles. *Tetrahedron* **55**, 4489–4500 (1999).
12. Speciale, G., Thompson, A. J., Davies, G. J. & Williams, S. J. Dissecting conformational contributions to glycosidase catalysis and inhibition. *Curr. Opin. Struct. Biol.* **28**, 1–13 (2014).
13. Fujita, K. *et al.* Molecular cloning and characterization of a β -L-Arabinobiosidase in *Bifidobacterium longum* that belongs to a novel glycoside hydrolase family. *J. Biol. Chem.* **286**, 5143–5150 (2011).
14. Spellman, M. W., McNeil, M., Darvill, A. G., Albersheim, P. & Henrick, K. Isolation and characterization of 3-C-carboxy-5-deoxy-L-xylose, a naturally occurring, branched-chain, acidic monosaccharide. *Carbohydr. Res.* **122**, 115–129 (1983).
15. Guérardel, Y. *et al.* The nematode *Caenorhabditis elegans* synthesizes unusual O-linked glycans: identification of glucose-substituted mucin-type O-glycans and short chondroitin-like oligosaccharides. *Biochem. J.* **357**, 167–182 (2001).
16. Russa, R., Urbanik-Sypniewska, T., Choma, A. & Mayer, H. Identification of 3-deoxy-lyxo-2-heptulosaric acid in the core region of lipopolysaccharides from Rhizobiaceae. *FEMS Microbiol. Lett.* **84**, 337–343 (1991).
17. Nagae, M. *et al.* Structural basis of the catalytic reaction mechanism of novel 1,2- α -L-fucosidase from *Bifidobacterium bifidum*. *J. Biol. Chem.* **282**, 18497–18509 (2007).
18. Sulzenbacher, G. *et al.* Crystal structure of *Thermotoga maritima* α -L-fucosidase. Insights into the catalytic mechanism and the molecular basis for fucosidosis. *J. Biol. Chem.* **279**, 13119–13128 (2004).
19. O'Neill, M. A., Ishii, T., Albersheim, P. & Darvill, A. G. Rhamnogalacturonan II: structure and function of a borate cross-linked cell wall pectic polysaccharide. *Annu. Rev. Plant Biol.* **55**, 109–139 (2004).
20. Pabst, M. *et al.* Rhamnogalacturonan II structure shows variation in the side chains monosaccharide composition and methylation status within and across different plant species. *Plant J.* **76**, 61–72 (2013).
21. Cartmell, A. *et al.* The structure and function of an arabinan-specific α -1,2-arabinofuranosidase identified from screening the activities of bacterial GH43 glycoside hydrolases. *J. Biol. Chem.* **286**, 15483–15495 (2011).
22. Glenwright, A. J. *et al.* Structural basis for nutrient acquisition by dominant members of the human gut microbiota. *Nature* **541**, 407–411 (2017).
23. Bourlard, T., Pellerin, P. & Morvan, C. Rhamnogalacturonans I and II are pectic substrates for flax-cell methyltransferases. *Plant Physiol. Biochem.* **35**, 623–629 (1997).
24. Glushka, J. N. *et al.* Primary structure of the 2-O-methyl- α -L-fucose-containing side chain of the pectic polysaccharide, rhamnogalacturonan II. *Carbohydr. Res.* **338**, 341–352 (2003).
25. Raman, R. *et al.* Advancing glycomics: implementation strategies at the Consortium for Functional Glycomics. *Glycobiology* **16**, 82R–90R (2006).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported in part by a grant to H.J.G. and B.H. from the European Research Council (grant no. 322820). B.H. was also funded by Agence Nationale de la Recherche under grant number ANR 12-BIME-0006-01. H.J.G. was also supported by Biotechnology and Biological Research Council (grant numbers BB/K020358/1 and BB/K001949/1), the Wellcome Trust (grant no. WT097907MA) and, with X.Z., M.-C.R. and F.B., was funded by the European Union Seventh Framework Programme under the WallTraC project (grant agreement number 263916). M.A.O. and B.R.U. were supported in part by grant DE-FG02-12ER16324 from The Division of Chemical Sciences, Geosciences, and Biosciences, Office of Basic Energy Sciences of the US Department of Energy. I.V. was in receipt of a Marie Skłodowska-Curie Fellowship (grant no. 707922). G.J.D. is a Royal Society Ken Murray Research Professor. D.W.A. was supported by a grant from the Beef and Cattle Research Council (FDE.15.13). We thank Diamond Light Source for access to beamline I02, I04-1 and I24 (mx1960, mx7854 and mx9948) that contributed to the results presented here, and to T. Doco and S. J. Charnock who supplied the partially purified apple RG-II.

Author Contributions Enzyme characterization was by D.N., A.R., A.C., A.S.L., I.V., A.L., D.W.A., Y.Z. and X.Z. Crystallographic studies were by A.C., A.B., A.S.L., D.N. and I.V. Purification of RG-II and oligosaccharide products was by M.A.O., A.R., D.N., A.C., A.L., A.S.L., F.B. and M.-C.R. HPLC analysis was by A.R., D.N., A.C. and A.S.L., while mass spectrometry analysis was by A.R. and J.G. Chemical synthesis was by S.N. and R.A.F. Growth analysis on purified RG-II performed by D.N. and A.R. Gene deletion strains were created and characterized by D.N. and I.V. Bioinformatics and genomic annotation were by N.T. and B.H. Bacterial growth and transcriptomic experiments were by J.B., Y.X. and E.C.M. Experiments were designed by D.N., A.R., A.C., A.S.L., E.C.M. and H.J.G. The manuscript was written by H.J.G. with contributions from G.J.D., B.H., M.A.O., B.R.U., E.C.M. and W.S.Y. Figures were prepared by A.R., A.L., D.N. and A.S.L.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to H.J.G. (harry.gilbert@ncl.ac.uk).

Reviewer Information *Nature* thanks M. Czjzek, S. Duncan and S. Withers for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size.

Preparing substrates. RG-II from wine and apple were selected for this work as they contain all the features of this glycan that are absent in the pectic polysaccharide from some other plant sources (Supplementary Discussion 1.0). Thus, the enzyme system that depolymerizes wine and apple RG-II is capable of degrading the glycan from all other plant sources.

Purifying RG-II. RG-II from wine was purified as described previously²⁶. To prepare RG-II from apple, a concentrate of the juice (25 kg box, equivalent of about 300 l of apple juice) was concentrated 1,000× using a Millipore peristaltic pump (3 l capacity) and Millipore Pellicon 2 (10,000 dalton molecular mass cut off) filter cassette module. Concentrated material was mixed with 5 volumes of pure ethanol to precipitate carbohydrates, which was then centrifuged at 20,000g at room temperature. The pellet was dissolved in 1 l of ultrapure water and re-precipitated with 5 volumes of ethanol. The precipitate was collected by centrifugation and this process was repeated another four times. The final pellet was redissolved in 0.5 l of ultrapure water and freeze-dried to obtain crude RG-II apple material.

HPAEC analysis showed that the crude apple RG-II was contaminated with arabinan and galactan. Thus, a mixture of enzymes targeting specifically arabinan and galactan polymers was added to the RG-II material. These recombinant enzymes were expressed and purified by IMAC as described in 'Producing recombinant proteins for biochemical assays'. The enzyme mixture comprised GH43 endo-arabinanases BT0360 and BT0367 that target branched and linear arabinan, respectively; GH51 L-arabinofuranosidases BT0368 and BT0348 that hydrolyse α 1,5-backbone and α 1,3-side chains, respectively; the GH2 β 1,4-galactosidase BT4667; and GH53 endo- β 1,4-galactanase BT4668. Around 300 nM of each enzyme was incubated for 24 h at 37 °C with crude RG-II (4% w/v) in 20 mM sodium phosphate buffer, pH 7.0. The enzyme-treated RG-II was fractionated on a XK50 column containing Fast flow DEAE Sepharose (column capacity 300 ml, GE Healthcare) at a flow rate of approximately 10 ml min⁻¹. The column was washed with 300 ml of 50 mM sodium acetate buffer, pH 4.5, and then eluted batch-wise with 3× column volumes of 50 mM sodium acetate containing 0.17 M, 0.3 M and 0.6 M NaCl. The RG-II eluted in the 0.3 M NaCl fraction based on glycosyl residue composition analyses, the types of sugars released by wild-type *B. thetaiotaomicron* cultured on this material, and the inability of the Δ rgII-*pull1* mutant to grow on this fraction. The fraction containing RG-II was dialysed against ultrapure water using 3.5 kDa cut off dialysis tubing. Progress of dialysis was monitored by use of the conductivity meter HI 99300 (Hanna instruments). The desalted RG-II was freeze-dried and kept at room temperature.

RG-II-derived oligosaccharides. *B. thetaiotaomicron* strains containing specific gene deletions were inoculated into minimal medium⁵ containing 1% RG-II and grown in glass tubes for 48 h at 37 °C, in an anaerobic cabinet (Whitley A35 Workstation; Don Whitley) to an OD_{600 nm} of 2.0. Cells were collected by centrifugation initially at 2,400g for 10 min and later at 17,000g for another 10 min. The resulting supernatant was filtered through a 1.2 μ m syringe filter (PALL Life Sciences) and separated on a Bio-Gel P2 (Biorad) size-exclusion column eluted with 50 mM acetic acid at 0.2 ml min⁻¹. Fractions (200 μ l) were collected and analysed by TLC using orcinol/sulfuric acid to reveal the resolved sugars. Fractions containing oligosaccharides of interest were pooled and concentrated by freeze-drying using a CHRIST Gefriertrocknung ALPHA 1-2 freeze-dryer (Helmholtz-Zentrum Berlin) at -50 °C.

The chemical synthesis of selected oligosaccharides used protocols described previously: L-Rhap- α 1,3 α -D-Apif-O-methyl and L-Rhap- β 1,3'-D-Apif-O-methyl²⁷; D-GalA- α 1,2-[D-GalA- β 1,3]-[L-Fuc- α 1,4]-L-Rha-O-methyl and D-GalA- α 1,2-[D-GalA- β 1,3]-L-Rha-O-methyl²⁸; L-Rhap- β 1,3'-D-Apif- β 1,2-D-GalA-O-methyl²⁹.

Producing recombinant proteins for biochemical assays. DNAs encoding enzymes lacking their signal peptides were amplified by PCR using appropriate primers. The amplified DNAs were cloned into NcoI/XhoI, NcoI/BamHI, NdeI/XhoI or NdeI/BamHI restricted pET21a or pET28a, as appropriate. The encoded recombinant proteins generally contained a C-terminal His₆-tag although, where appropriate, the His-tag was located at the N terminus of the protein. To express the recombinant genes, *Escherichia coli* strains BL21(DE3) or TUNER, harbouring appropriate recombinant plasmids, were cultured to mid-exponential phase in Luria Bertani broth at 37 °C. Recombinant gene expression was induced by the addition of 1 mM (strain BL21(DE3)) or 0.2 mM (TUNER) isopropyl β -D-galactopyranoside (IPTG), and the culture was grown for a further 5 h at 37 °C or 16 h at 16 °C, respectively. The recombinant proteins were purified to >90% electrophoretic purity by immobilized metal ion affinity chromatography (IMAC) using Talon, a cobalt-based matrix, with bound proteins eluted with 100 mM imidazole, as described previously⁴. To generate seleno-methionine (Se-Met) proteins for structure resolution, *E. coli* cells were cultured as described previously³⁰, and the proteins were purified using IMAC as described above.

For crystallization, the Se-Met proteins were further purified by size exclusion chromatography. After IMAC, fractions containing the purified proteins were buffer-exchanged, using PD-10 Sephadex G-25M gel-filtration columns (GE Healthcare), into 10 mM Na-HEPES buffer, pH 7.5, containing 150 mM NaCl and were then subjected to gel filtration using a HiLoad 16/60 Superdex 75 column (GE Healthcare) at a flow rate of 1 ml min⁻¹. For crystallization trials, purified proteins were concentrated using an Amicon 10-kDa molecular mass centrifugal concentrator and washed three times with 5 mM DTT (for the Se-Met proteins) or water (for native proteins).

Site-directed mutagenesis. Site-directed mutagenesis was carried out employing a PCR-based NZY-Mutagenesis kit (NZYTech Ltd) using the plasmids encoding the appropriate enzymes as the template. The mutated DNA clones were sequenced to ensure that only the appropriate DNA change was introduced after the PCR.

Glycoside hydrolase assays. Spectrophotometric quantitative assays for L-rhamnosidases (BT0986, BT1001 and BT1019), L-arabinofuranosidases (BT0983, BT0996, BT1021 and BT3662), D-galacturonidases (BT0992, BT0997 and BT1018), the D-glucuronidase (BT0996) and D-galactosidase (BT0993) were monitored by the formation of NADH, at A_{340 nm} using an extinction coefficient of 6,230 M⁻¹ cm⁻¹, with an appropriately linked enzyme assay system. The assays were adapted from purchased Megazyme International assay kits. These kits were as follows: the L-rhamnose assay kit (K-RHAMNOSE); L-arabinose/D-galactose assay kit (K-ARGA); D-glucuronic acid/D-galacturonic acid assay kit (K-URONIC). The activity of BT0984 was monitored by using an excess of BT0993 and quantifying D-galactose release as mentioned above. BT1003 and BT1012 activity was measured by using an excess of BT1001 and linking it to rhamnose release as described above. Substrate depletion assays were used to determine the activity of enzymes BT1002 and BT1012, while the formation of L-galactose from RG-II was used to determine the activity of BT1010. In brief, aliquots of the enzyme reaction were removed at regular intervals and, after boiling for 10 min to inactivate the enzyme and centrifugation at 13,000g, the amount of the substrate remaining or product produced was quantified by HPAEC using standard methodology. The reaction substrates and products were bound to a Dionex CarboPac PA1 column and were eluted with an initial isocratic flow of 100 mM NaOH then a 0–200 mM sodium acetate gradient in 100 mM NaOH at a flow rate of 1.0 ml min⁻¹, using pulsed amperometric detection. Linked assays were checked to make sure that the relevant enzyme being analysed was rate limiting by increasing its concentration and ensuring a corresponding increase in rate was observed. A single substrate concentration was used to calculate catalytic efficiency (k_{cat}/K_M), and was checked to be markedly less than K_M by halving and doubling the substrate concentration and observing an appropriate increase or decrease in rate. The equation $V_0 = (k_{cat}/K_M)[S][E]$ was used to calculate k_{cat}/K_M unless substrate depletion was used then the calculation was as follows $\ln(k_{cat}/K_M) = (S_0/S_t)/[E]$, in which $[E]$ and $[S]$ are enzyme and substrate concentration, respectively³¹. All reactions were carried out in 20 mM sodium phosphate buffer, pH 7.0, with 150 mM NaCl (defined as standard conditions) and performed in at least technical triplicates. The activity of selected enzymes against complex oligosaccharides was also evaluated using mass spectrometry as described below. The substrates used were RG-II from wine or apple, RG-II-derived oligosaccharides generated using *B. thetaiotaomicron* mutants, by partial acid treatment of RG-II, or by chemical synthesis (prepared as described above). TLC was used to provide a qualitative profile of the activity of selected enzymes. Around 4 μ l of the reaction was spotted on silica gel TLC plates and the plates were developed in ascending butanol:acetic acid:water 2:1:1. Carbohydrate products were detected by spraying with 0.5% orcinol in 10% sulfuric acid and heating to 100 °C for 10 min.

Infusion electrospray mass spectrometry analysis. The structures of the desalted oligosaccharides (in 10 mM ammonium acetate, pH 7.0) were analysed via negative ion mode infusion/offline electrospray ionization mass spectrometry (ESI-MS) following dilution (typically 1:1 (v/v)) with 5% trimethylamine in acetonitrile.

Electrospray data was acquired using an LTQ-FT mass spectrometer (Thermo) with a FT-MS resolution setting of 100,000 at $m/z = 400$ and an injection target value of 1,000,000. Infusion spray analyses were performed on 5–10 μ l of samples using medium 'nanoES' spray capillaries (Thermo) for offline nanospray mass spectrometry in negative ion mode at 1 kV.

Nanoelectrospray liquid chromatography mass spectrometry analysis. Oligosaccharide samples were diluted (typically 1:50 (v/v)) with 0.1% formic acid (aq) and injected (0.5 μ l) onto a capillary 3 μ m particle, 200 Å pore ProntoSIL C18AQ trapping column (nanoLCMS Solutions LLC) in-line with a 75 μ m × 100 mm BEH130 C18 capillary column (Waters) running on a NanoAcquity UPLC system (Waters). The gradient conditions were 0.1–15% buffer B in buffer A in 15 min, 15–50% buffer B in buffer A over 24 min and 50–90% buffer B in buffer A over 5 min, with 15 min re-equilibration (0.1% buffer B in buffer A) at a flow rate of 0.4 μ l min⁻¹. Buffer A: 0.1% formic acid in water; buffer B: 0.1% formic acid in acetonitrile.

Nano electrospray data were acquired using an LTQ-FT mass spectrometer (Thermo). Survey MS scans were performed over the mass range $m/z = 150\text{--}2,000$ in data-dependent mode. Data was acquired with a FT-MS resolution setting of 100,000 at $m/z = 400$ and a Penning trap injection target value of 1,000,000. The top five ions in the survey scan were automatically subject to collision-induced dissociation MS/MS in the linear ion trap region of the instrument at an injection target value of 100,000, using a normalized collision energy of 30% and an activation time of 30 ms (activation $Q = 0.25$).

Growth of *B. thetaiotaomicron* and generation of mutants. The selection of *B. thetaiotaomicron* to analyse the mechanism by which RG-II is degraded in the HGM was based on the following criteria: (1) *B. thetaiotaomicron* is a component of the HGM in a wide range of individuals; (2) the bacterium was previously shown to be capable of growing on RG-II and the genetic loci activated by the pectic polysaccharide identified; (3) a genetic system for the bacterium has been developed enabling routine targeted gene deletion, which was crucial in identifying the enzymes that depolymerized RG-II.

Growth of *B. thetaiotaomicron*. *B. thetaiotaomicron* was routinely cultured under anaerobic conditions at 37 °C using an anaerobic cabinet (Whitley A35 Workstation; Don Whitley) in culture volumes of 0.2, 2 or 5 ml of TYG (tryptone-yeast extract-glucose medium) or minimal medium containing 1% of an appropriate carbon source plus 1.2 mg ml⁻¹ porcine haematin (Sigma-Aldrich) as previously described⁵. The growth of the cultures were routinely monitored at OD_{600nm} using a Biochrom WPA cell density meter for the 5 ml cultures or a Gen5 v2.0 Microplate Reader (Biotek) for the 0.2 and 2 ml cultures.

Constructing mutants of RG-II PUL1 in *B. thetaiotaomicron*. The mutants of the complete RG-II PUL1 knockout and single gene clean deletions were introduced by counter selectable allelic exchange using the pExchange vector as described³².

Crystallization, data collection, structure solution and refinement. All protein concentrations were around 10 mg ml⁻¹ except for BT1012, which was at 4 mg ml⁻¹. BT3662 was crystallized in 20% polyethylene glycol (PEG) 6000 and 0.1 M Na-HEPES, pH 6.5. BT1012 was crystallized in 40% (v/v) MPD, 0.2 M ammonium phosphate and Tris, pH 8.5. Se-Met-containing BT0996 was crystallized in 20% (w/v) PEG 8000, 0.1 M Tris, pH 8.5. Additionally Se-Met-containing BT0996 were crystallized in 20% (w/v) PEG 3350 and 0.2 M lithium acetate for ligand soaking in mother liquor containing 25 mg ml⁻¹ of chain B for up to 16 h. The same process was repeated for inactive mutant BT0996(E240Q). BT1002 protein was crystallized in 15% (v/v) ethylene glycol, 15% (w/v) PEG 8000, 30 mM MgCl₂, 30 mM CaCl₂, 50 mM Na-HEPES and 50 mM MOPS, pH 7.5. Se-Met-containing BT1002 was crystallized in 10% (w/v) PEG 3350, 0.1 M HEPES, pH 7.5 and 0.2 M L-proline. BT1003 was crystallized in 20% (w/v) PEG 3350 and 0.2 M potassium acetate. BT0986 was crystallized 15% (w/v) PEG 550 MME, 15% (w/v) PEG 20000, 0.25 M rhamnose, 50 mM HEPES and 50 mM MOPS, pH 7.5, in the absence or presence of 6 mM D-rhamnopyranose tetrazole. The inactive mutant BT0986(D461Q) was crystallized in 15% (w/v) PEG 550 MME, 15% (w/v) 20000, 50 mM imidazole, 50 mM MES, pH 6.5, 30 mM NaF, 30 mM NaBr, 30 mM NaI in the presence of 5 mM of RG-II chain B. Se-Met-containing BT0986 was crystallized in 20% (w/v) PEG 3350, 0.2 M sodium sulfate, 0.25 M rhamnose and 0.1 M bis-tris propane, pH 6.5. Se-Met containing BT1020 was crystallized in 14% (v/v) ethylene glycol, 14% (w/v) PEG 8000, 30 mM of each di-, tri-, tetra- and penta-ethylene glycol, 50 mM HEPES and 50 mM MOPS, pH 7.5. BT1020 was crystallized in 15% (v/v) ethylene glycol, 15% (w/v) PEG 8000, 20 mM D-glucose, 20 mM D-mannose, 20 mM D-galactose, L-fucose, D-xylose, N-acetyl-D-glucosamine, 300 mM L-arabinose, 44.5 mM imidazole, 55.5 mM MES pH 6.5. All proteins were initially screened by the sitting-drop method with a protein volume of 0.1 or 0.2 μl and a reservoir ratio of 1:1 or 2:1 using a robotic nanodrop dispensing system (mosquito LCP; TTPLabTech).

Cryo-protection. BT1002, BT1003, Se-Met-BT0986, BT0986 in presence of rhamnotetrazole, Se-Met-containing BT1020 were cryo-protected by supplementing the mother liquor with 20% (v/v) PEG 400. 20% ethylene glycol (v/v) was used for BT3362, and 20% (v/v) glycerol for BT0996. All other samples did not require supplementary cryo-protection.

Data collection and processing. All data were integrated using XDS³³ apart from BT1003 which was integrated with DIALS³⁴ and BT0986 with the rhamnotetrazole which was integrated with xia2 3dii³⁵. The data were scaled with either XDS³³ or Aimless³⁶.

Structure solution and refinement. The phase problem for BT1002, BT1020, BT0986 and BT0996 was solved by Se-Met-SAD using hkl2map³⁷ and the shelx pipeline³⁸. Buccaneer³⁹ and/or arp-warp⁴⁰ were used for automated model building. BT0986 in the presence of rhamnotetrazole and BT0986(D461Q) in the presence of a Chain B-derived heptasaccharide were solved with molrep⁴¹ using BT0986 native as search model. All the other data were solved by molecular replacement using phaser⁴² and the Protein Data Bank (PDB) model 3QZ4 for

BT3662, 3KZS for BT1012 and an ensemble of PDB models 3WKX and 4QJY for BT1003. Buccaneer³⁹ and/or arp-warp⁴⁰ were as well used when needed to improve the initial molecular replacement solution. Recursive cycles of model building in coot⁴³ and refinement in refmac5 (ref. 44) were performed to produce the final model. For all models solvent molecules were added using coot⁴³ and checked manually; 5% of the observations were randomly selected for the R_{free} set. All models were validated using coot⁴³ and molprobt⁴⁵. The data statistics and refinement details are reported in Supplementary Table 8.

Comparative genomics analysis. PULs similar to the RG-II PULs were searched for in 372 Bacteroidetes genomes (complete list provided in Supplementary Table 5). The identification of similar PULs was based on PUL alignments. Gene composition and order of Bacteroidetes PULs were computed using the PUL predictor described in PULDB⁴⁶. Then, in a manner similar to amino acid sequence alignments, the predicted PULs were aligned to the RG-II PULs according to their modularity as proposed in the RADS/RAMPAGE method⁴⁷. Modules taken into account include CAZy families, sensor-regulators and *suscd*-like genes. Finally, PUL boundaries and limit cases were refined by BLASTP-based analysis. The new glycoside hydrolase families discovered in this study are listed in the main text.

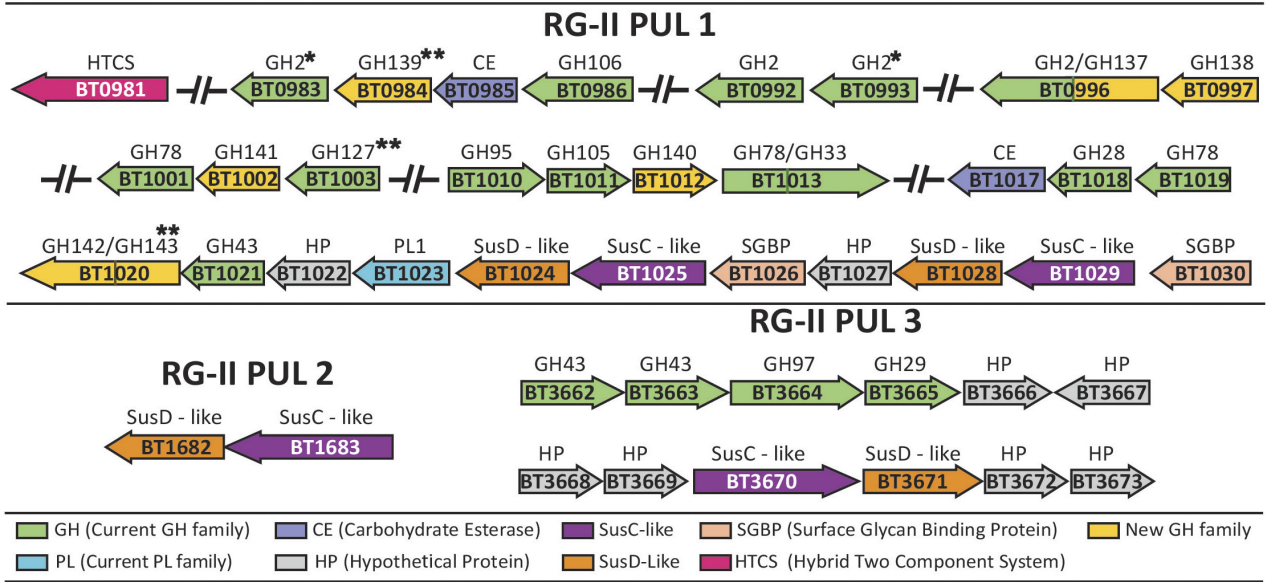
Protein cellular localization. Cellular localization of proteins was carried out as described previously⁴. In brief, *B. thetaiotaomicron* cultures were grown overnight (OD_{600nm} value of 2.0) in 5 ml minimal media containing 1% apple RG-II. The next day, cells were collected by centrifugation at 5,000g for 10 min and resuspended in 2 ml PBS. Proteinase K (0.5 mg ml⁻¹ final concentration) was added to 1 ml of the suspension and the other half left untreated (control). Both samples were incubated at 37 °C overnight followed by centrifugation (5,000g for 10 min) to collect cells. To eliminate residual proteinase K activity, cell pellets were resuspended in 1 ml of 1.5 M trichloroacetic acid in PBS and incubated on ice for 30 min. Precipitated mixtures were then centrifuged (5,000g, 10 min) and washed twice in 1 ml ice-cold acetone (99.8%). The resulting pellets were allowed to dry in a 40 °C heat block for 5 min and dissolved in 250 μl Laemmli buffer. Samples were heated for 5 min at 98 °C and mixed by pipetting several times before resolving by SDS-PAGE using 7.5% gels. Electrophoresed proteins were transferred to nitrocellulose membranes by western blotting followed by immunochromatological detection using primary rabbit polyclonal antibodies (Eurogentec) generated against various proteins and secondary goat anti-rabbit antibodies (Santa Cruz Biotechnology). For BT1010 and BT1013, the anti-sera of which failed to produce the desired reactivity, a C-terminal Flag peptide (DYKDDDDK) was incorporated at the C-terminals of both native proteins expressed by *B. thetaiotaomicron* through counter-selectable allelic exchange³². This allowed for their detection using rabbit anti-Flag antibodies (Sigma) as primary antibodies.

Data availability. The crystal structure datasets generated (atomic coordinate files and structure factors) have been deposited in the Protein Data Bank (PDB) under the following accession numbers: 5MQO (BT1003), 5MQR (BT1020 native), 5MQS (BT1020 Ara), 5MSX (BT3662), 5MSY (BT1012), 5MT2 (BT0996 Se-Met), 5MUI (BT0996 chain B Araf), 5MUJ (BT0996 chain B Apif), 5MQN (BT0986), 5MQM (BT0986 inhibitor) and 5MWK (BT0986 RG-II chain B) (see Supplementary Table 8). The authors declare that the data supporting the findings of this study are available within the paper and the Supplementary Information.

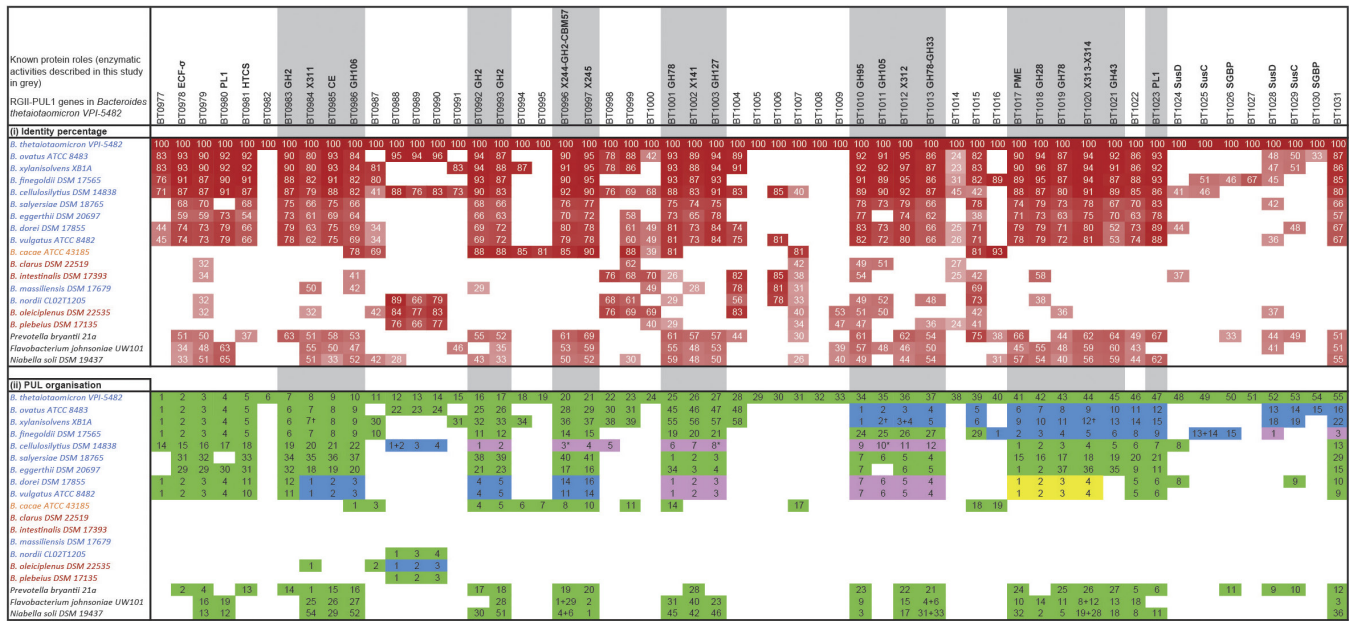
- Buffetto, F. *et al.* Recovery and fine structure variability of RGII sub-domains in wine (*Vitis vinifera* Merlot). *Ann. Bot.* **114**, 1327–1337 (2014).
- Chauvin, A. L., Nepogodiev, S. A. & Field, R. A. Synthesis of an apiose-containing disaccharide fragment of rhamnogalacturonan-II and some analogues. *Carbohydr. Res.* **339**, 21–27 (2004).
- Chauvin, A. L., Nepogodiev, S. A. & Field, R. A. Synthesis of a 2,3,4-triglycosylated rhamnoside fragment of rhamnogalacturonan-II side chain A using a late stage oxidation approach. *J. Org. Chem.* **70**, 960–966 (2005).
- Nepogodiev, S. A., Fais, M., Hughes, D. L. & Field, R. A. Synthesis of apiose-containing oligosaccharide fragments of the plant cell wall: fragments of rhamnogalacturonan-II side chains A and B, and apiogalacturonan. *Org. Biomol. Chem.* **9**, 6670–6684 (2011).
- Charnock, S. J. *et al.* The X6 “thermostabilizing” domains of xylanases are carbohydrate-binding modules: structure and biochemistry of the *Clostridium thermocellum* X6b domain. *Biochemistry* **39**, 5013–5021 (2000).
- Matsui, I. *et al.* Subsite structure of *Saccharomycopsis* α-amylase secreted from *Saccharomyces cerevisiae*. *J. Biochem.* **109**, 566–569 (1991).
- Koropatkin, N. M., Martens, E. C., Gordon, J. I. & Smith, T. J. Starch catabolism by a prominent human gut symbiont is directed by the recognition of amylose helices. *Structure* **16**, 1105–1115 (2008).
- Kabsch, W. *Xds*. *Acta Crystallogr. D* **66**, 125–132 (2010).
- Waterman, D. G. *et al.* Diffraction-geometry refinement in the DIALS framework. *Acta Crystallogr. D* **72**, 558–575 (2016).
- Winter, G. xia2: an expert system for macromolecular crystallography data reduction. *J. Appl. Crystallogr.* **43**, 186–190 (2010).

36. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. D* **69**, 1204–1214 (2013).
37. Pape, T. & Schneider, T. R. HKL2MAP: a graphical user interface for macromolecular phasing with SHELX programs. *J. Appl. Crystallogr.* **37**, 843–844 (2004).
38. Sheldrick, G. M. Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. *Acta Crystallogr. D* **66**, 479–485 (2010).
39. Cowtan, K. Fitting molecular fragments into electron density. *Acta Crystallogr. D* **64**, 83–89 (2008).
40. Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat. Protocols* **3**, 1171–1179 (2008).
41. Vagin, A. & Teplyakov, A. MOLREP: an automated program for molecular replacement. *J. Appl. Crystallogr.* **30**, 1022–1025 (1997).
42. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
43. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
44. Vagin, A. A. *et al.* REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr. D* **60**, 2184–2195 (2004).
45. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
46. Terrapon, N., Lombard, V., Gilbert, H. J. & Henrissat, B. Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. *Bioinformatics* **31**, 647–655 (2015).
47. Terrapon, N., Weiner, J., Grath, S., Moore, A. D. & Bornberg-Bauer, E. Rapid similarity search of proteins using alignments of domain arrangements. *Bioinformatics* **30**, 274–281 (2014).

a

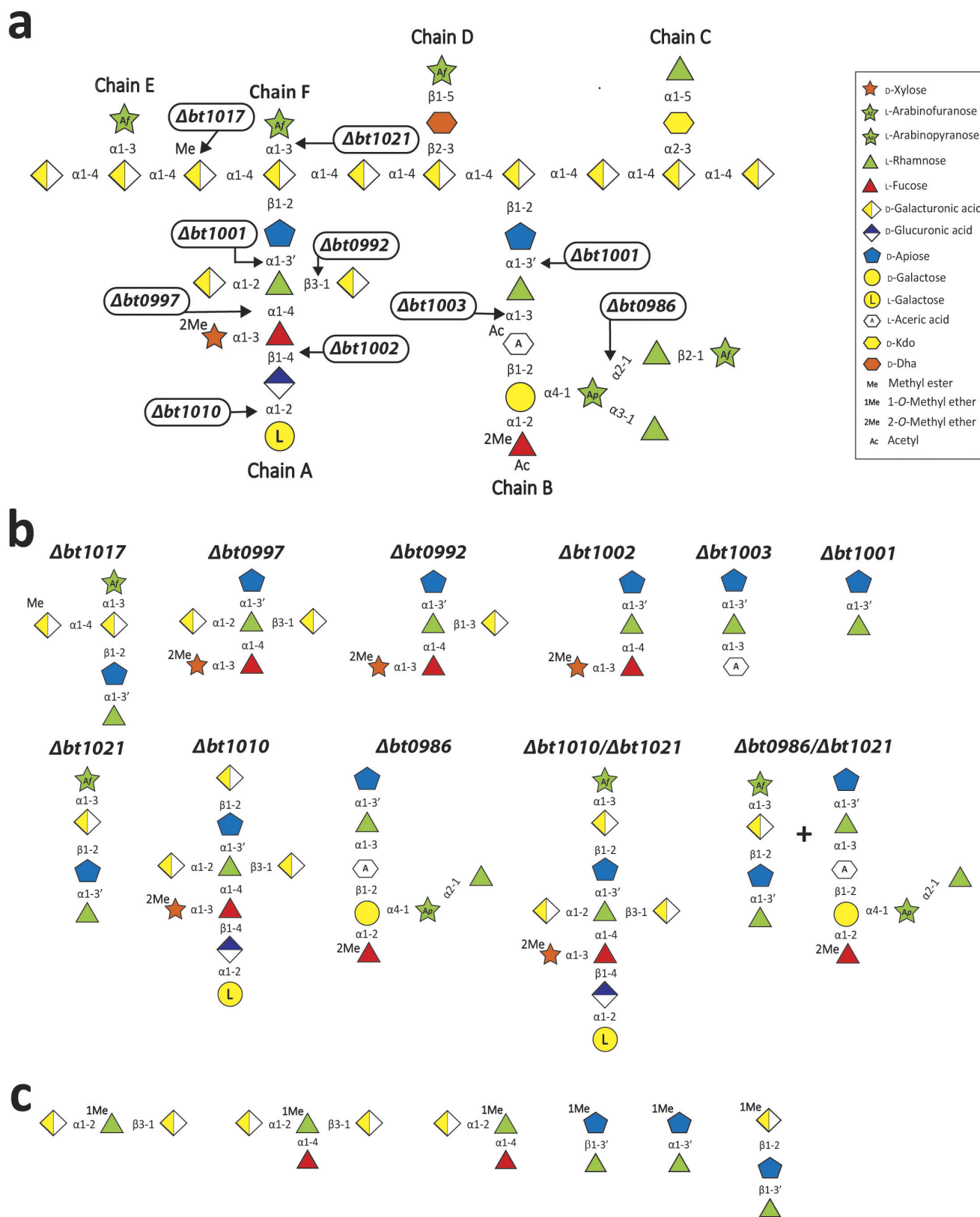


b



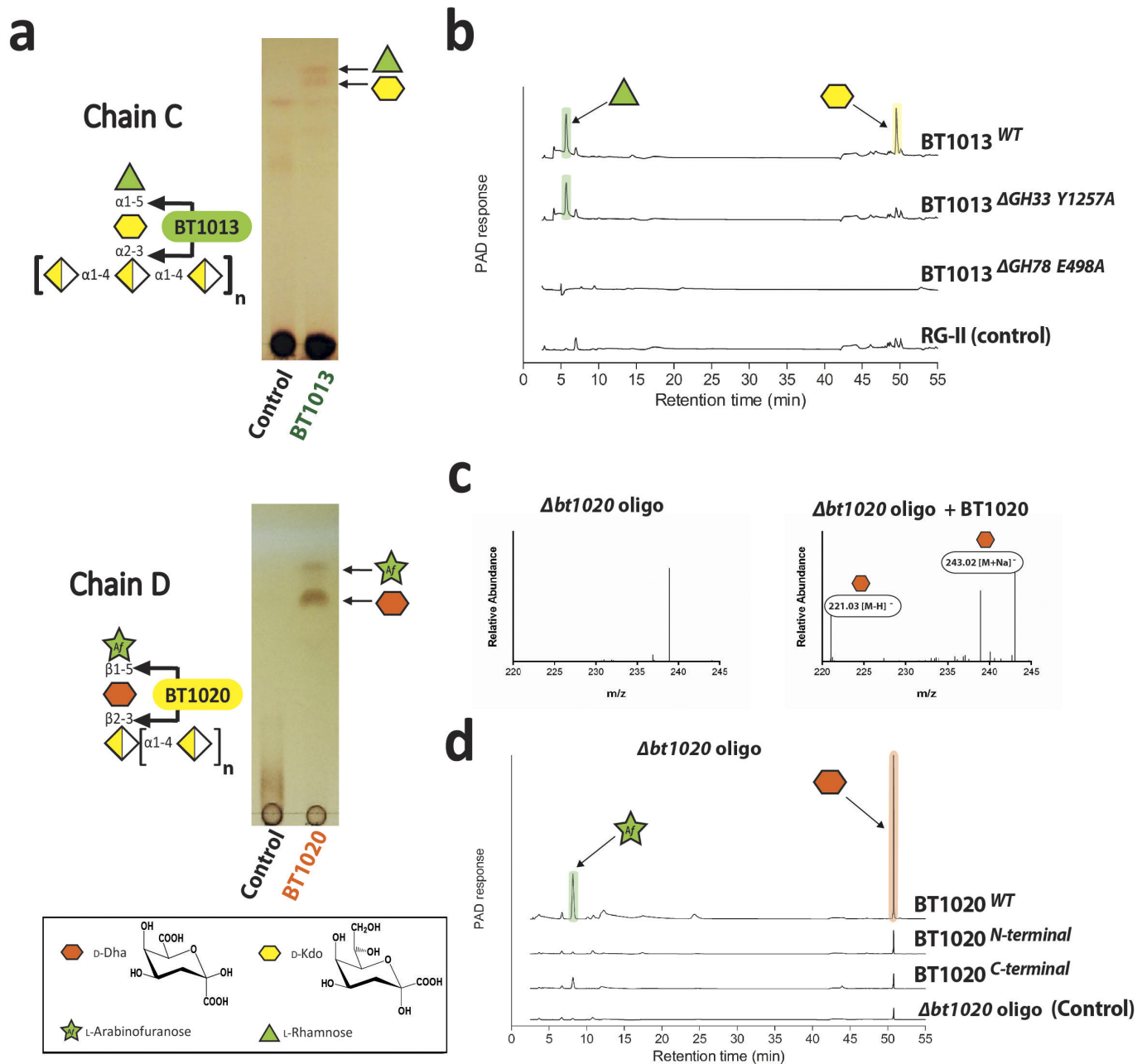
Extended Data Figure 1 | Sequence conservation and genomic organization of components of RG-II PUL1 across Bacteroidetes species. a, PULs activated by RG-II. Genes encoding proteins of known or predicted functionalities are colour-coded. Arrows indicate the orientation of each gene. **b**, For Bacteroidetes species (in rows), xenologues of *B. thetaiotaomicron* (*Bt*) RG-II PUL1 proteins (in columns) were identified by reciprocal best-BLASTP hits. These organisms were coloured to reflect growth on RG-II; blue, growth in 24 h; orange, growth in 48 h; red, no growth; black, growth on RG-II was not evaluated. *B. thetaiotaomicron* proteins that contribute to RG-II degradation are grey. CE, carbohydrate esterase; GHXX, glycoside hydrolase family; PL1, polysaccharide lyase family. **b, i**, Xenologues are in the same column as the corresponding *B. thetaiotaomicron* protein, and percentage identity is in a red colour-scale to 24% identity. **b, ii**, Genomic organization of xenologues into gene clusters/loci. The top row shows 55

B. thetaiotaomicron proteins numbered according to relative position of the gene on the genome. The xenologues of the *B. thetaiotaomicron* genes are numbered as they appear in their respective cluster/locus. For each species, *B. thetaiotaomicron* RG-II PUL1 was split into separate clusters when the contiguous xenologues were separated by ≥ 30 unrelated genes. Each cluster has a distinct background colour (green, blue, pink and yellow). Proteins in *B. xylanisolvans* XB1A, not annotated previously as open-reading frames identified here by TBLASTN, are marked with a dagger. Split proteins in *B. cellulosilyticus* DSM 14838 due to incomplete genome assembly were numbered based on the *B. cellulosilyticus* WH2 genome (marked with an asterisk). *B. thetaiotaomicron* enzymes that were split into two distinct single module enzymes in other species are denoted with a 'plus' sign in the PUL. Supplementary Fig. 3 shows a complete depiction of the PUL organization in the selected species.



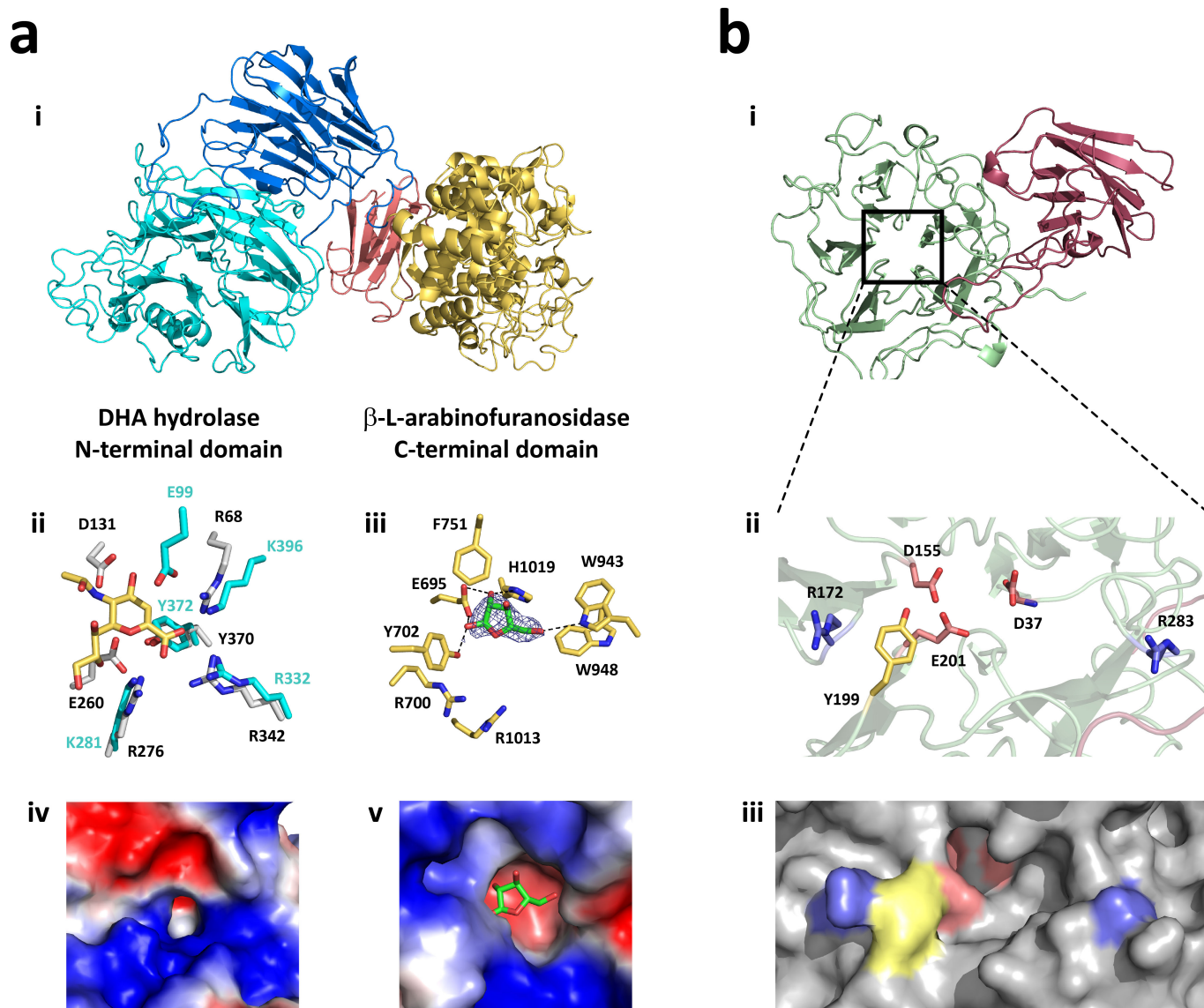
Extended Data Figure 2 | The generation of RG-II derived oligosaccharides by mutants of *B. thetaiotaomicron*. **a**, The site of action of the enzyme that was eliminated or not functional in the corresponding mutant. **b**, The structures of the oligosaccharides generated by each mutant. These molecules were isolated by size-exclusion chromatography

and used as substrates to determine the mechanism of RG-II degradation. **c**, The structure of bespoke chemically synthesized oligosaccharides that were used as substrates to dissect the mechanism of RG-II depolymerization.



Extended Data Figure 3 | Analysis of the disassembly of chains C and D. Reactions were carried out using standard conditions as described. Control denotes that the glycan was not enzyme treated. **a**, Wild-type BT1013 and BT1020 were incubated with RG-II and the oligosaccharides generated by the *B. thetaiotaomicron* $\Delta bt1020$ mutant grown for 50 h, respectively, and the reactions were analysed by TLC. **b**, Wild-type and mutants of BT1013 were incubated with RG-II, and the products were subjected to HPAEC-PAD analysis. The BT1013 mutants $\Delta GH78 E496A$ and $\Delta GH33 Y1257A$ are mutants of the predicted catalytic residues of the GH78 rhamnosidase and GH33 sialidase catalytic modules, respectively.

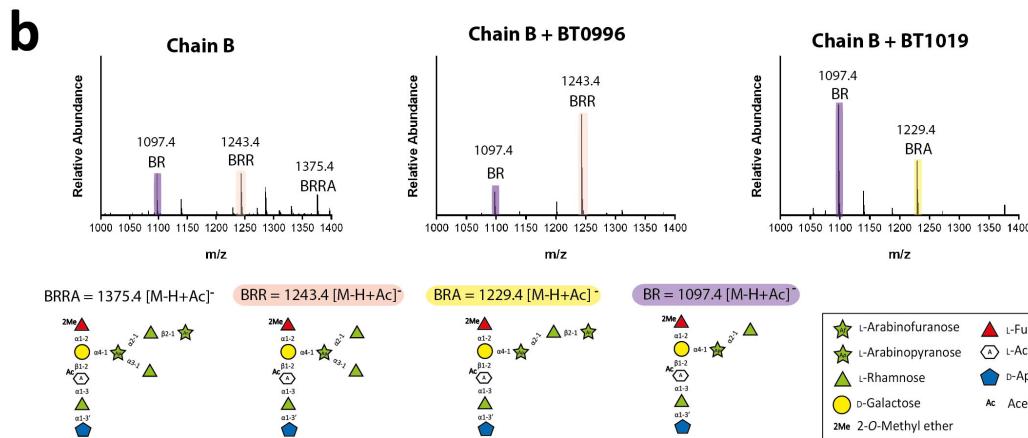
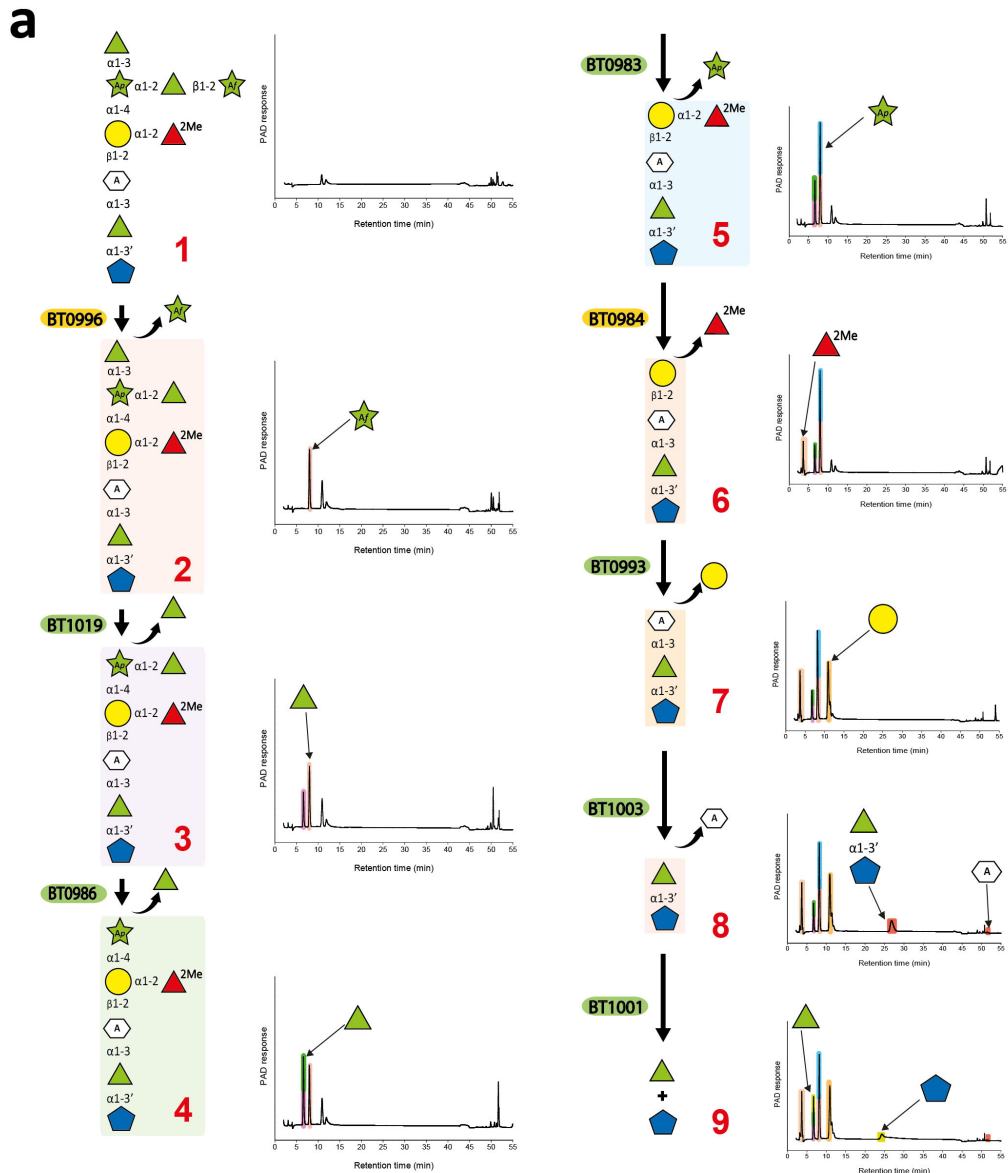
c, $\Delta bt1020$ oligosaccharide was incubated with wild-type BT1020, and the products were analysed by HPAEC-PAD and mass spectrometry. **d**, The activity of the N-terminal (residues Asp26 to Asp613) and C-terminal (residues Lys614 to Leu1107) regions of BT1020 were compared with the wild-type enzyme using HPAEC-PAD analysis. Arabinose and rhamnose were identified through HPAEC-PAD by co-migration with the appropriate standard, while the identity of DHA was consistent with the mass spectrometry data. The data displayed are examples from biological replicates $n = 3$.



Extended Data Figure 4 | Crystal structures of BT1020 and BT3662.

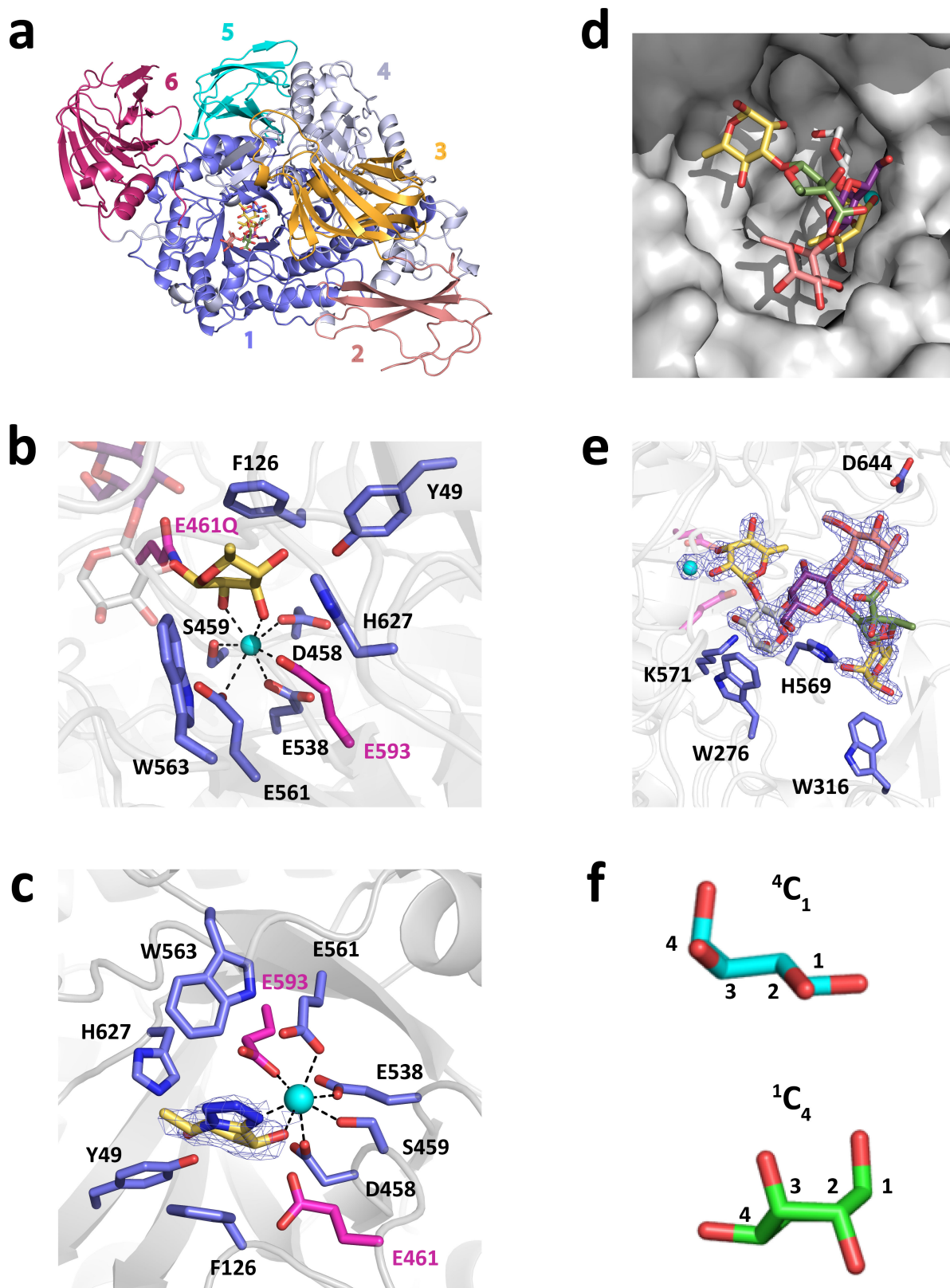
a, b, The structures of BT1020 (**a**) and BT3662 (**b**). **a, i**, Schematic of BT1020, in which the domains from the N to C termini are coloured cyan (DHA hydrolase catalytic domain), blue (structural β -sandwich domain 1), yellow (β -L-arabinofuranosidase catalytic domain), salmon (structural β -sandwich domain 2). **a, ii, iii**, Key active site residues in stick format of the DHA hydrolase (**ii**) and β -L-arabinofuranosidase (**iii**) catalytic domains. In **a, ii**, the BT1020 residues (carbon and lettering coloured cyan) are overlaid with the GH33 *Clostridium perfringens* sialidase NanI (PDB code 2VK7; carbons and lettering are light grey and black, respectively). The ligand, *N*-acetyl-neuramic acid (yellow), is derived from the NanI structure. In **a, iii**, the active site amino acids (carbons in yellow) that interact with the bound *L*-Araf (carbons in green) are shown with polar contacts depicted by broken black lines. **a, iv, v**, Charged surface representations of the active site pockets of the DHA hydrolase (**iv**) and

β -L-arabinofuranosidase (**v**) catalytic domains (with *L*-Araf bound to the β -L-arabinofuranosidase). Note the highly basic feature at the active sites consistent with the negatively charged DHA located in the active site or +1 subsite of the two catalytic domains. **b, i**, Schematic of BT3662, in which the five-bladed β -propeller catalytic domain and the β -sandwich domain are shown in green and red, respectively. In **b, ii**, key residues are shown in stick format. The carbons of the amino acids coloured salmon pink are catalytic residues, the yellow tyrosine is positioned in the +1 subsite, and the blue arginine residues are in the distal regions that interact with the *D*-GalA-containing backbone. **b, iii**, Solvent-exposed surface representation of **b, ii**, using the same colour format for the amino acids highlighted. The active site housing the catalytic residues is located in a pocket that abuts onto a shallow channel containing the arginines and tyrosine.



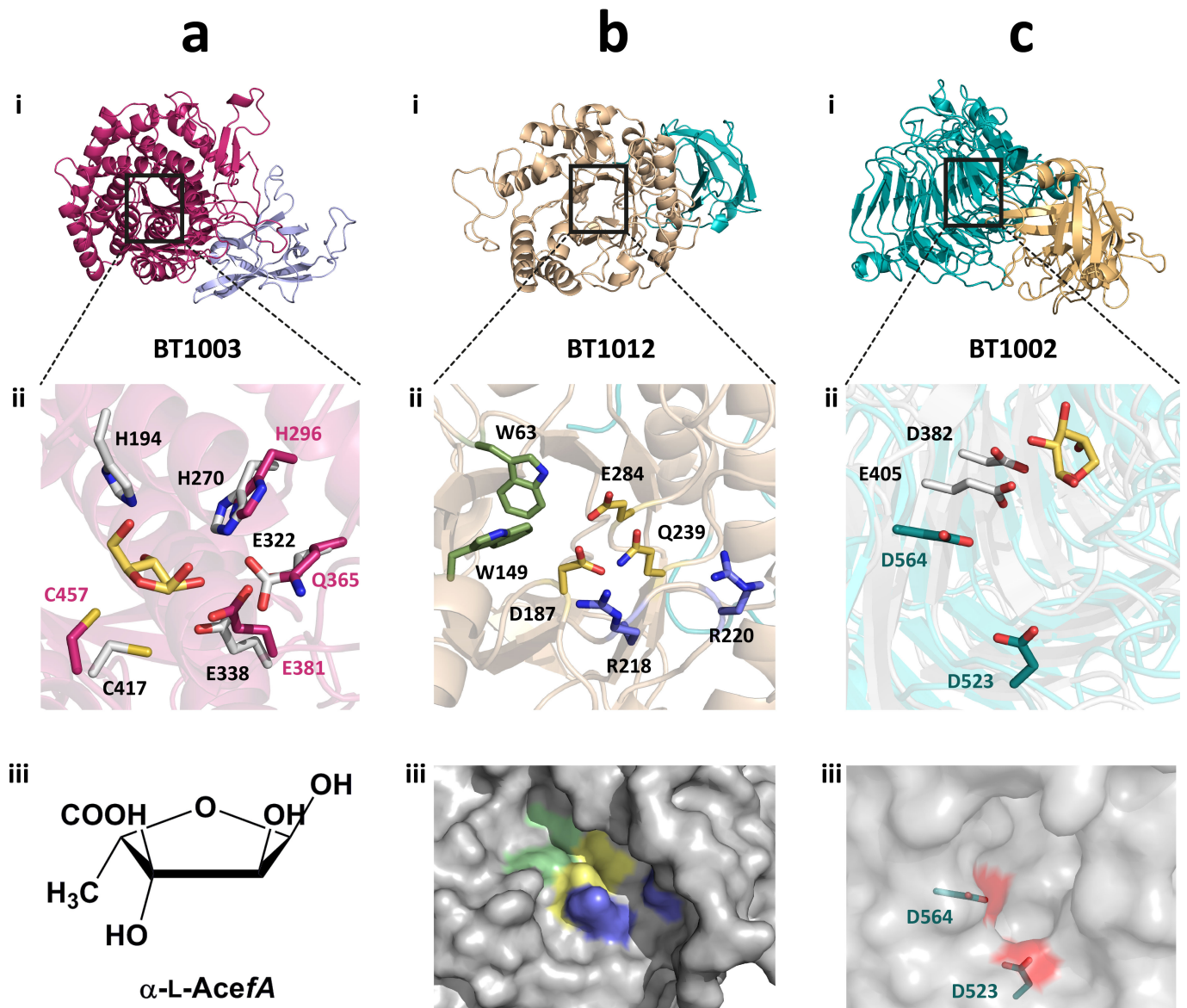
Extended Data Figure 5 | Mechanism by which *B. thetaiotaomicron* depolymerizes chain B of RG-II. The oligosaccharides generated by $\Delta bt1003$ and $\Delta bt0986$, RG-II and chemically synthesized molecules were used to determine which enzymes acted on chain B. **a**, Identified enzymes were added sequentially to chain B (isolated by mild acid treatment of

RG-II). The reactions were carried out under standard conditions. The sugars released were identified and quantified by HPAEC-PAD. **b**, Examples of the use of mass spectrometry to monitor the enzymatic disassembly of chain B. Examples shown are from biological replicates, $n = 3$.



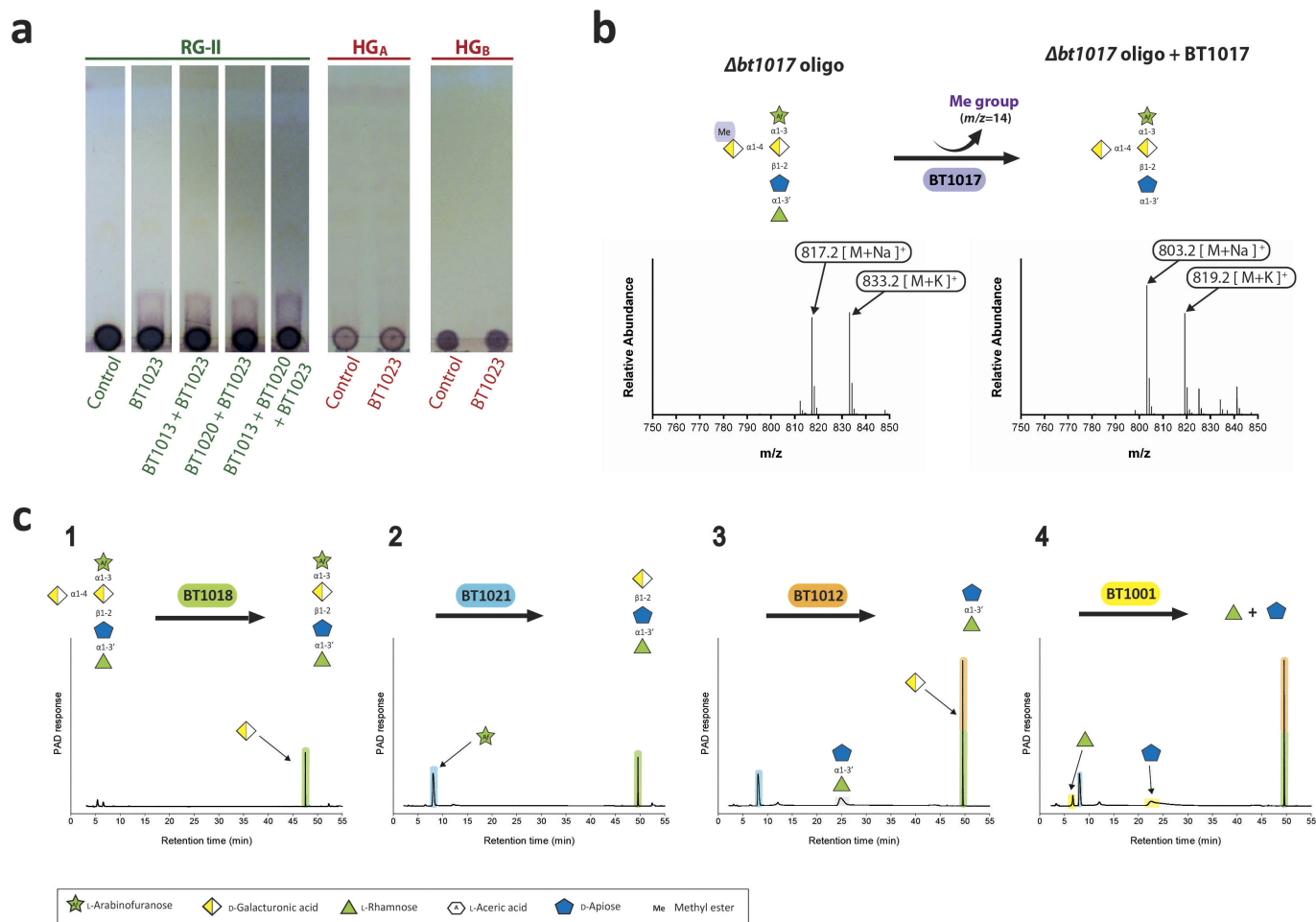
Extended Data Figure 6 | Crystal structure of BT0986. a, A schematic of the enzyme is displayed, revealing the $(\alpha/\beta)_8$ -barrel catalytic domain (yellow) that is interrupted with three β -sandwich domains (blue), while the C-terminal domain (salmon) also folds into a β -sandwich. **b,** Active site of BT0986 bound to rhamnose. Catalytic amino acids are coloured magenta, other amino acids are blue, and the rhamnose is in yellow. **c,** Transition state mimic rhamnopyranose tetrazole bound in the active site of BT0986. Colouring as in **c**. The blue mesh surrounding the ligand represents the $2F_o - F_c$ electron density map (1.3 Å resolution) at 1.5σ . In **b** and **c**, the calcium ion in the active site is shown as a cyan sphere

and its polar contacts with amino acids and ligands are indicated by black dashed lines. **d,** Chain B-derived heptasaccharide generated by the mutant bacterium $\Delta bt0986$ bound in the substrate-binding site of BT0986 shown as a surface representation. **e,** The interactions of the $\Delta bt0986$ heptasaccharide with BT0986. Amino acids that interact with the oligosaccharide are coloured blue and the sugars in both **d** and **e** are as depicted in Fig. 5. The blue mesh is the electron density of the oligosaccharide. **f,** Conformation of the arabinopyranose in chain B (green) and in the $\Delta bt0986$ heptasaccharide (cyan). The carbons are numbered.



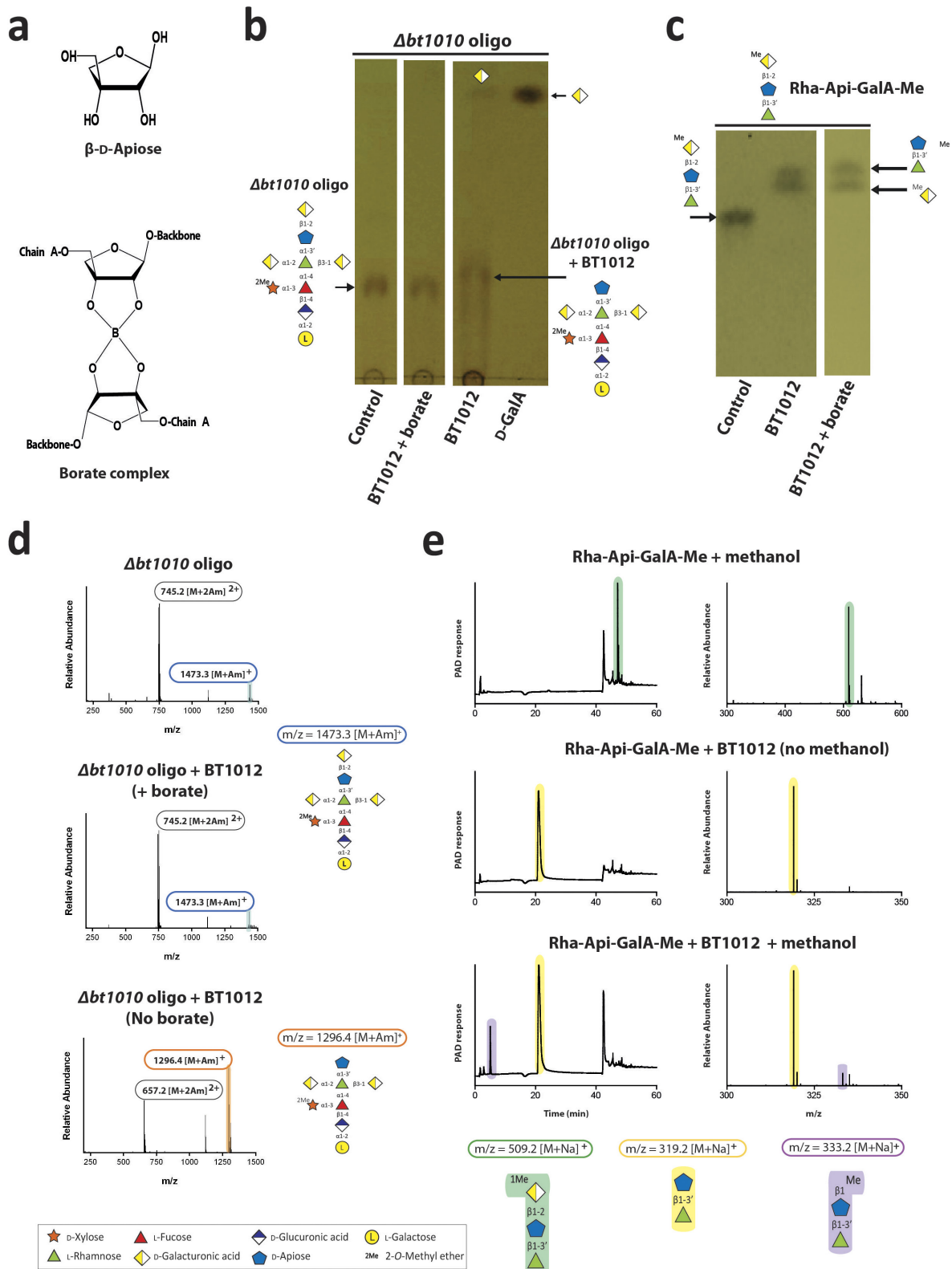
Extended Data Figure 7 | Crystal structure of BT1003, BT1012 and BT1002. All amino acids are in stick format and the position of the active site in the schematics of the respect enzymes is indicated by a black box. **a**, Structure of the GH127 aceric acid hydrolase. **a**, **i**, Schematic of the enzyme revealing the catalytic domain (α/α_6 -barrel (red) and β -sandwich domain (blue). **a**, **ii**, Overlay of the key active site residues of the aceric acid hydrolase (red) and the GH127 β -L-arabinofuranosidase HypBA1 from *Bifidobacterium longum* (PDB code 3WKX) (white grey). The proposed catalytic nucleophile, Cys457 in BT1003, is conserved in the two enzymes; the catalytic acid/base in HpyBA1 (Glu322), however, is a glutamine (Gln365) in the aceric acid hydrolase. The ligand (yellow) is β -L-Araf derived from the crystal structure of the β -L-arabinofuranosidase. **a**, **iii**, Structure of aceric acid (α -L-AcefA). **b**, Structure of the apiosidase BT1012. **b**, **i**, Schematic of the enzyme, revealing the (α/β) $_8$ -barrel catalytic domain (beige) and the C-terminal β -sandwich domain (blue). **b**, **ii**, The yellow amino acids denote the key residues in the active site (-1 subsite) that are proposed to have a direct catalytic role (Asp187 and Glu284) or

substrate binding function (Gln239). The pair of arginine residues (blue) in the $+1$ subsite are likely to contribute to D-GalA binding through interactions with the carboxylate of the uronic acid. The aromatic residues (green) are in the -2 subsite. **b**, **iii**, Solved-exposed surface representation of the substrate-binding region of BT1012. The highlighted residues are coloured as in **b**, **ii**. **c**, Structure of BT1002. **c**, **i**, Schematic of the enzyme, revealing the C-terminal β -parallel helical catalytic domain (green) and the N-terminal β -sandwich domain (yellow). **c**, **ii**, Overlay of BT1002 and its closest structural homologue, a GH120 β -xylosidase (PDB code 3VSU), highlighting the position of the catalytic amino acids (green, BT1002; or light grey, β -xylosidase). The xylose in the active site of the β -xylosidase is shown to orientate the close but not identical position of the catalytic centre of the two enzymes. **c**, **iii**, Solvent-exposed surface representation of the active site pocket of BT1002, in which the catalytic residues are shown in stick format and their location on the surface depicted in red. The pocket is elongated, hinting that the 2-O-methyl-xylose appended to the L-fucose may be housed in the catalytic centre of the α -L-fucosidase.



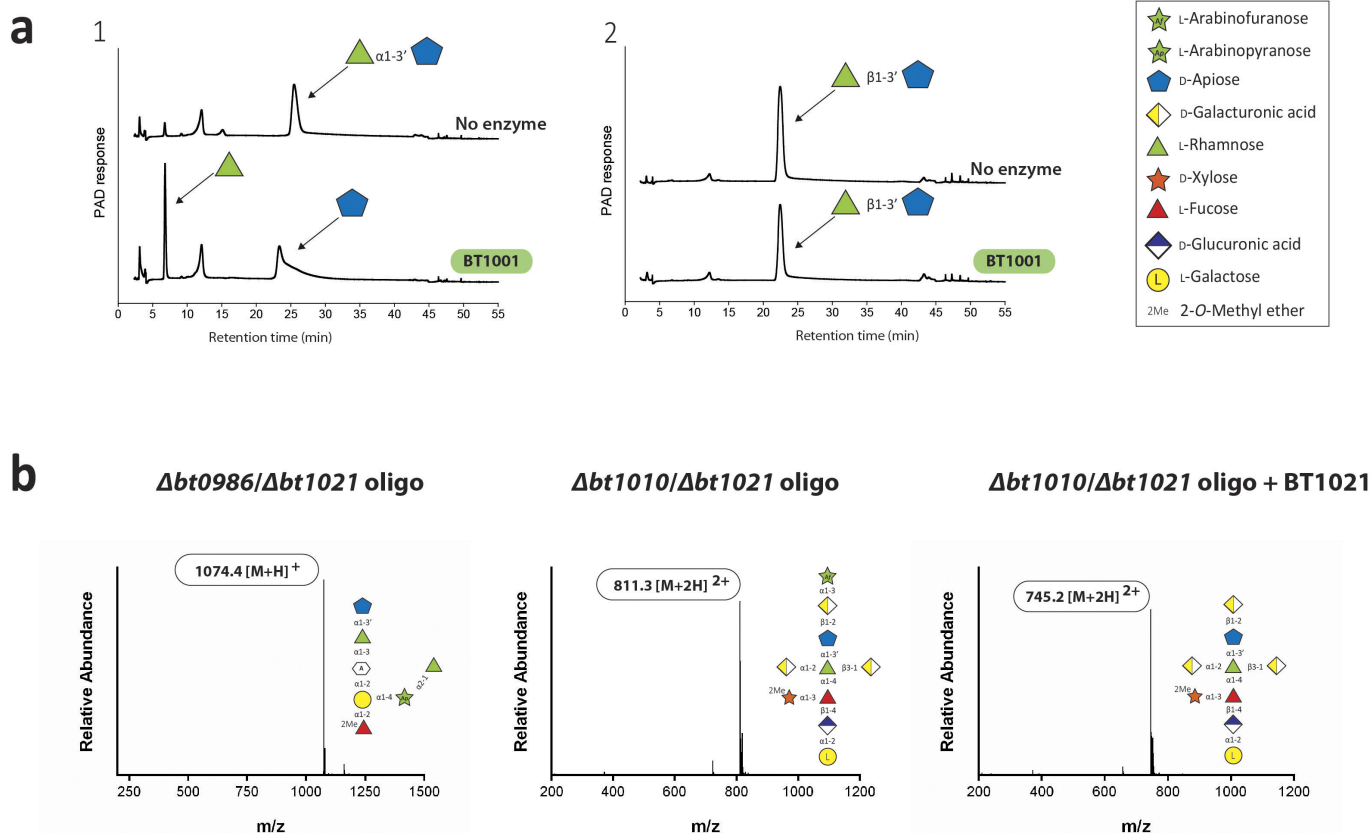
Extended Data Figure 8 | Identification of the enzymes that disassemble the backbone of RG-II. **a**, BT1023 was incubated with RG-II or homogalacturonan HGA or HGB, which are 30% and 80% methyl-esterified, respectively (1% w/v), in 50 mM CAPSO buffer, pH 9.0, containing 2 mM CaCl₂. RG-II substrate was either untreated or had been previously incubated with BT1013 and/or BT1020. The reactions were analysed by TLC. Control lanes denote the absence of enzyme. **b**, The oligosaccharide generated by the $\Delta bt1017$ mutant ($\Delta bt1017$ oligo) was

incubated with BT1017 and the reaction product was analysed by mass spectrometry. **c**, Sequential degradation of the product generated in **b** by the enzymes indicated under standard conditions. The sugars released by these other enzymes (reactions 1–4) were identified and quantified by HPAEC-PAD. The structure of the oligosaccharide sugars followed the notation described in Fig. 1 and Extended Data Fig. 2. The example shown are from technical replicates $n = 3$.



Extended Data Figure 9 | The influence of borate on the retaining apiosidase BT1012. **a**, The chemical structure of apiose (top), and when the sugar is cross-linked by borate in chain A of RG-II (bottom). **b**, Oligosaccharide generated by the *B. thetaiotaomicron* mutant $\Delta bt1010$ ($\Delta bt1010$ oligo) was incubated with BT1012 under standard conditions in the presence and absence of 50 mM borate, and the products were analysed by TLC. **c**, The same experiment as in **a** except that the substrate was

L-Rhap- β 1,3'-D-Apif- α 1,2-D-GalA-Me (Rha-Api-GalA-Me). **d**, Mass spectrometric analysis of the control sample and the two reactions containing BT1012 loaded onto the TLC in **b**. **e**, BT1012 was incubated under standard conditions with Rha- β 1,3'-Api- α 1,2-GalA-Me in the presence and absence of 2.5 M methanol. The reactions were analysed by HPAEC-PAD and mass spectrometry. The example shown are from technical replicates $n = 3$.



Extended Data Figure 10 | Modification of the structure of RG-II.

a, Recombinant BT1001 and BT1012 were incubated with bespoke chemically synthesized oligosaccharides using standard conditions.
b, The enzymatic disassembly of RG-II was used to investigate the structure of RG-II. Mass spectrometry combined with HPAEC-PAD

were used to analyse the structure of the oligosaccharides generated by the mutants *Δbt1010/Δbt1021* and *Δbt1010/Δbt0986* before and after treatment with recombinant BT1021 using standard conditions. The enzyme reactions were analysed by HPAEC-PAD.