SUPPLEMENTARY INFORMATION

Cellular location of endo-acting galactanases confers keystone or recipient status to arabinogalactan degrading bacteria of the human gut microbiota

Alan Cartmell[1¶], Jose Muñoz-Muñoz[1¶], Jonathon Briggs[1¶], Didier A. Ndeh[1¶], Elisabeth C. Lowe[1], Arnaud Baslé[1], Nicolas Terrapon[2], Katherine Stott[3], Tiaan Heunis[1], Joe Gray[1], Li Yu[3], Paul Pearl Z. Fernandes[4], Sayali Shah[4], Spencer J. Williams[4], Aurore Labourel[1], Matthias Trott[1], Bernard Henrissat[2,5,6] and Harry J. Gilbert[1,*]

Supplementary Table 1 Annotation and activity of potential enzymes encoded by PUL_{AGPL} and PUL_{AGPS}

	Annotation	Recombinant expression?	Active on LA-or GA-AGP	Other Comment
PUL_{AGPL}				
BT0262	hypothetical protein	Yes	No	
BT0263	PL27	Yes	Yes	Rhamno-galaturonyl lyase
BT0264	GH43_24	Yes	Yes	Endo-β1,3-galactanase
BT0265	GH43_24	Yes	Yes	Exo-β1,3-galactosidase
BT0266	hypothetical protein	Yes	No	
BT0267	HTCS	Yes	No	
BT0268	SusC	No		
BT0269	SusD	No		
BT0270	hypothetical protein	Yes	No	
BT0271	hypothetical protein(DUF5007)	Yes	No	
BT0272	SusC	No		
BT0273	SusD	No		
BT0274	hypothetical protein(Fasciclin)	Yes	No	
BT0275	hypothetical protein(DUF5008, DUF5124, DUF5122)	Yes	No	
BT0276	hypothetical protein(Laminin_G_3)	Yes	No	
BT0277	M60-like Peptidase, CBM32	Yes	No	Lacks catalytic residue
BT0278	Type I phosphodiesterase	Yes	No	
BT0279	Hypothetical protein	Yes	No	
BT0280	Transposase	Yes	No	
BT0284	peptidoglycan binding protein	Yes	No	
BT0285	tolQ-type transport protein (MotA_ExbB)	Yes	No	
BT0290	GH35, CBM32	Yes	Yes	Exo-β1,6-galactosidase
PUL_{AGPS}				
BT3674	GH127	Yes	Yes	β-L-arabinofuranosidase
BT3675	GH43_34	Yes	Yes	α-L-arabinofuranosidase
BT3676	BNR repeat-containing family member	Yes	No	
BT3677	hypothetical protein(DUF2264)	Yes	Yes	New family β-glucuronidase
BT3678	HTCS	Yes	No	Sensor not soluble.
BT3679	hypothetical protein(IPT/TIG domain)	Yes	Yes	New family α-L-arabinofuranosidase
BT3680	SusC	No		
BT3681	SusD	No		
BT3682	hypothetical protein(DUF1735, DUF4973, DUF4361)	Yes	No	
BT3683	GH43_24-GH16	Yes	Yes	Exo-β1,3-galactosidase
BT3685	GH43_24	Yes	Yes	Exo-β1,3-galactosidase
BT3686	GH145	Yes	Yes	Exo-α-L-rhamnosidase
BT3687	GH105	Yes	Yes	Unsaturated β-glucuronidase

Supplementary Table 2 Kinetic parameters of enzymes active against AGPs.

	k_{cat} (min^{-1})	K_m (mM)	k_{cat}/K_m ($min^{-1}M^{-1}$)
BT0264 (GH43_24)			
β-1,3 Gal_2	-[a]	-	Inactive
β-1,4 Gal_2	-	-	Not tested
β-1,6 Gal_2	-	-	Inactive
BT0265 (GH43_24)			
β-1,3 Gal_2	-	-	$(4.2 \pm 0.32) \times 10^3$
β-1,4 Gal_2	-	-	Inactive
β-1,6 Gal_2	-	-	Inactive
BT0290 (GH35)			
Larch Wood	-	-	$(5.5 \pm 0.1) \times 10^6$
β-1,3-galactobiose	-	-	$(4.9 \pm 0.2) \times 10^3$
β-1,6-galactobiose	-	-	$(1.4 \pm 0.1) \times 10^6$
BT3674 (GH127)			
LA-AGP	124.3 ± 10.8	12.10 ± 2.1	$(1.1 \pm 0.4) \times 10^5$
GA-AGP	-	-	$(1.34 \pm 0.4) \times 10^3$
BT3675 (GH43)			
GA-AGP	-	-	$(4.77 \pm 0.9) \times 10^5$
BT3677			
GA-AGP	38.36 ± 2.32	1.35 ± 0.21	$(2.8 \pm 1.1) \times 10^5$
GlcA-Gal-Gal	-	-	$(4.1 \pm 0.2) \times 10^5$
BT3679			
LA-AGP	-	-	$(5.3 \pm 0.4) \times 10^5$
GA-AGP	-	-	$(2.4 \pm 0.1) \times 10^5$
Wheat Arabinoxylan	-	-	$(8.5 \pm 0.7) \times 10^4$
Sugar beet Arabinan	105.7 ± 16.2	274.5 ± 23.5	$(3.9 \pm 0.4) \times 10^5$
1,5-α-Arabinobiose	-	-	$(1.6 \pm 0.1) \times 10^5$
BT3683 (GH43_24)			
β-1,3 Gal_2	-	-	$(9.7 \pm 2.2) \times 10^2$
β-1,4 Gal_2	-	-	Inactive
β-1,6 Gal_2	-	-	Inactive
BT3685 (GH43_24)			
β-1,3 Gal_2	-	-	$(2.3 \pm 0.23) \times 10^6$
β-1,4 Gal_2	-	-	Inactive
β-1,6 Gal_2	-	-	Inactive

[a]: kinetic parameter not determined.

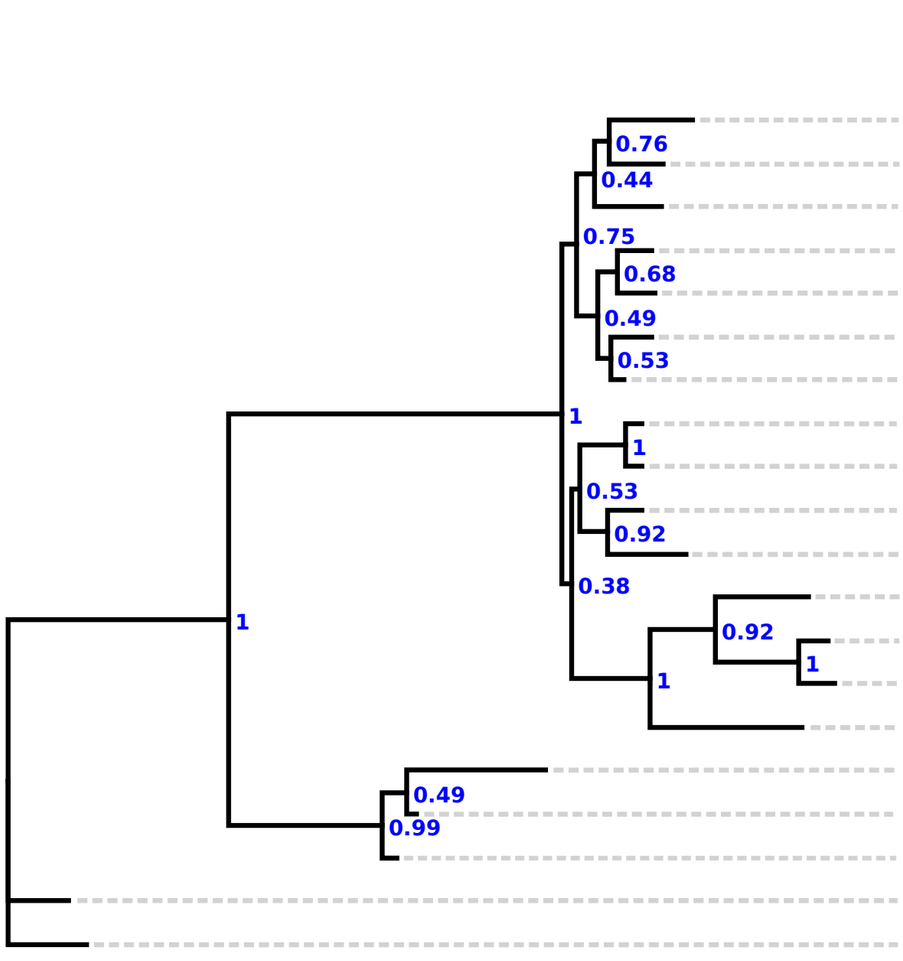
All kinetic parameters are +/- S.E.M . The values were determined from technical replicates n = 3.

Supplementary Table 3 Kinetic parameters of mutants of the GH43_24 module in BT3683 and BT3685.

Enzyme	β-1,3 Gal_2	2,4-DNP-β-Gal
	Catalytic efficiency (k_{cat}/K_m $min^{-1}M^{-1}$)	
BT3683 wild type	(9.7 ± 2.2)x 10^2	(6.8 ± 0.9) x 10^2
GH43 module Q577A (catalytic base)	Inactive	Inactive
GH43 module Q577N	Inactive	Inactive
GH43 module Q577E	Inactive	Inactive
GH43 module E367A (O4 specificity)	Inactive	Inactive
GH43 module E367Q	Trace activity	(2.7 ± 0.3) x 10^1
GH43 module D471A (pK_a modulator)	Inactive	Inactive
GH43 module D471N	Inactive	Inactive
GH43 module E520A (catalytic acid)	Inactive	(4.0 ± 0.5) x 10^2
GH43 module E520Q	(2.7 ± 0.3) x 10^3	(7.5±0.9) x 10^2
GH43 module E319A	(6.6 ± 0.7) x 10^2	-[a]
GH43 module E319Q	(3.5 ± 0.05) x 10^3	-
GH43 module R321A	(1.8 ± 0.15) x 10^2	-
GH43 module R321K	(2. 3 ±0.15) x 10^2	-
GH43 module F326A	(1.0 ± 0.04) x 10^4	-
GH43 module Y393F	(2.2 ± 0.14) x 10^2	-
GH43 module S487A	(2.6 ± 0.12) x 10^3	-
GH43 module S487V	(2.3 ± 0.33) x 10^3	-
GH43 module C538A	(3.2 ± 0.14) x 10^2	-
GH16 module E150A (catalytic nucleophile)	(1.2 ± 0.06)x 10^3	-
GH16 module E155A (catalytic acid/base)	(7.6 ± 0.08) x 10^2	-
BT3685 wild type	(2.28 ± 0.23)x 10^6	(9.8 ± 0.69) x 10^4
Q282A (catalytic base)	(2.41 ± 0.38)x 10^3	(4.5 ±1.0) x 10^3
Q282N	Inactive	(1.5 ± 0.31) x 10^3
Q282E	(3.1 ± 0.56) x 10^3	(3.4 ±0.67) x 10^2
E121A (O4 specificity)	No Expression	No Expression
E121Q	No Expression	No Expression
D176A (pKa modulator)	No Expression	No Expression
D176N	Inactive	(8.8 ± 0.23) x 10^1
E225A (catalytic acid)	Inactive	Inactive
E225Q	Inactive	(6.2± 0.15) x 10^4

[a]: activity not determined.

All kinetic parameters are +/- S.E.M . The values were determined from technical replicates n = 3.



Substrate regions & enzyme specificities	Gal branches	Gal main-chain		Ara decorations		GlcA+Rhap decorations (**only GA-AGP**)				Growth on LA-AGP	Growth on **GA-AGP**	Growth on **GA-AGP** +BT0264 or +BACCELL_00844	Growth on **WH-AGP**	Growth on **WH-AGP** +BT0264
		exo	endo	LA- & GA-AGP	**Only GA-AGP**	RG pathway		Lyase pathway						
Orthologs of enzymes	BT0290	BT0265, BT3683	BT0264, BACCELL_00844	BT3674, BT3679	BT3675	1) BT3686	2) BT3677	1) BT0263	2) BT3687					
Bacteroides fragilis *ATCC 25285*														
Bacteroides nordii *ATCC BAA-998*														
Bacteroides caccae ATCC 43185														
Bacteroides thetaiotaomicron *ATCC29418*														
Bacteroides finegoldii DSM 17565														
Bacteroides ovatus ATCC 8483														
Bacteroides xylanisolvens *DSM18836*						A								
Bacteroides cellulosilyticus DSM 14838						B								
Bacteroides intestinalis DSM 17393														
Bacteroides uniformis ATCC 8492						B								
Bacteroides stercoris ATCC 43183														
Bacteroides massiliensis DSM 17679														
Bacteroides dorei DSM 17855														
Bacteroides vulgatus ATCC 8482														
Bacteroides plebeius DSM 17135						B								
Parabacteroides gordonii DSM 23371														
Parabacteroides johnsonii DSM 23371														
Parabacteroides merdae ATCC 43184														
Dysgonomonas mossii DSM 22836														
Dysgonomonas gadei ATCC BAA-286														

Present in similar genomic/PUL contexts
Present in other genomic/PUL contexts
Absence of any ortholog

A - missing gene model
B - mutated catalytic His residue

Growth
No growth
Not tested

Supplementary Table 4. Conservation of the key enzymes for AGP degradation in *B. thetaiotaomicron* and 19 Bacteroidetes **species.**
A phylogenetic tree of the 20 species, reconstructed from the 16S-RNA, is displayed on the left with the bootstrap numbers shown in blue. The table on the right shows with colored cells, the existence and conserved microsynteny of the orthologs to *B. thetaiotaomicron* key enzymes involved in the deconstruction of AGP regions. Growth data of each species on three substrates are shown in the last columns.

Supplementary Table 5. Proteins identified on the surface of *B. thetaiotaomicron::baccell00844* using proteomics.

SGBP	Uniprot No.	CAZyme family and/or PUL	Known or predicted function	Signal peptide	Predicted or known cellular location
BT3079	Q8A376	None	Unknown	Type II	Surface
BT2966	Q8A3I8	PUL44^{L}; induced by host glycans	SGBP	Type II	Surface
BT1773	Q8A6V3	None	SusD-like	Type II	Surface
BT4080	Q8A0E1	PUL74^{L}; induced by host glycans	Unknown	Type II	Surface
BT3159	Q8A2Z6	PUL50^{L}; induced by host glycans	SGBP	Type II	Surface
BT1896	Q8A6I8	None	Unknown	Type II	Surface
BT1182	Q8A8I5	None	Unknown	Type II	Surface
BT3792	Q8A174	GH76 in PUL68^{L}; induced by α-mannan	Endo-α1,6-mannanase	Type II	Surface
BT4088	Q8A0D3	PUL74^{L}; induced by host glycans	SusC-like	Type II	Surface
BT2623	Q8A4H6	GH76 in PUL36^{L}; induced by α-mannan	Endo-α1,6-mannanase	Type II	Surface
BT3066	Q8A388	None	Unknown	Type II	Surface
BT3523	Q8A1Y5	PUL60^{L}; induced by host glycans	Unknown	Type II	Surface
BT3067	Q8A387	None	Unknown	Type II	Surface
BT2125	Q8A5W1	None	Unknown	Type II	Surface
BT3433	Q8A273	Unknown	Unknown	Type II	Surface
BT2193	Q8A5P5	PUL28^{L}; induced by host glycans	Unknown	Type II	Surface

BT4245	Q89ZX6	PUL78^{L}; induced by host glycans	SGBP	Type II	Surface
BT3955	Q8A0R6	PUL71^{L}; induced by host glycans	Unknown	Type II	Surface
BT3960	Q8A0R1	PUL71^{L}; induced by host glycans	SGBP	Type II	Surface
BT4083	Q8A0D8	PUL74^{L}; induced by host glycans	SGBP	Type II	Surface
Baccell00844	E2N999	GH16	Endo-β1,3-galactanase	Type II	Surface
BT3438	Q8A268	None	Unknown	Type II	Surface
BT2317	Q8A5C2	None	Unknown	Type II	Surface
BT4984	Q8A0D7	PUL74^{L}; induced by host glycans	SGBP	Type II	Surface
BT1938	Q8A6E6	None	Unknown	Type II	Predicted to be a lipoprotein
BT3961	Q8A0R0	PUL71^{L}; induced by host glycans	SGBP	Type II	Surface
BT0153	Q8ABF7	None	Unknown	Type II	Surface
BT4171	Q8A050	PUL77^{L}; induced by RGI	Unknown	Type II	Surface
BT1044	Q8A8X3	GH18 in PUL60^{L}; induced by host glycans	endo-β-N-acetylglucosaminidase	Type II	Surface
BT4167	Q8A054	PUL77^{L}; induced by RGI	SGBP	Type II	Surface
BT3793	Q8A173	PUL68^{L}; induced by α-mannan	Unknown	Type II	Surface
BT4170	Q8A051	PL9 in PUL77^{L}; induced by RGI	RGI lyase	Type II	Bacterial surface
BT3859	Q8A112	PUL69^{L}; induced	Unknown	Type II	Surface

		by α-mannan			
BT4471	Q89ZA6	PUL83	SusD-like	Type II	Surface
BT4668	Q89YR3	GH53 in PUL86[L]; induced by galactan	Endo-β1,3-galactanase	Type II	Surface
BT0142	Q8ABG8	None	Unknown	Type II	Surface
BT3988	Q8A0N3	PUL72[L]	Unknown	Type II	Surface
BT3436	Q8A270	None	Unknown	Type II	Surface
BT4166	Q8A055	PUL77[L]; induced by RGI	Unknown	Type II	Surface
BT3741	Q8A1C5	None	Unknown	Type II	Surface
BT3476	Q8A230	PUL59[L]	SGBP	Type II	Surface
BT3026	Q8A3C8	PUL46[L]	Unknown	Type II	Surface
BT4605	Q89YX5	None	Unknown	Type II	Surface
BT1038	Q8A8X9	PUL14[L]	SGBP	Type II	Surface

SGBP; surface glycan binding proteins. [L]; PUL identified from the literature in PULDB (http://www.cazy.org/PULDB/). Function is predicted (**red**) based on the position of the gene adjacent to *susC-susD* pairs, in the case of putative SGBPs, or the CAZy family in which the protein is located. Cellular location is predicted (**red**) based on the presence of a type II signal peptide that anchors the protein onto the outer membrane. Proteins displaying substantial sequence identity to SusD or SusC proteins are coloured **blue**. These proteins are highly likely to be located on the outer membrane exposed on the surface. Proteins coloured **green** have been shown to display the function identified and/or experimentally demonstrated to be presented on the bacterial surface.

Supplementary Table 6a. Data statistics and refinement details.

	BT0265	BT0265Hexa	BT3683Gal	BT3683GalIM
Data collection				
Date	16/10/15	12/12/16	13/02/16	08/05/17
Source	IO4-1	IO4-1	IO4-1	IO4-1
Wavelength	0.98	0.93	0.97	0.97
Space Group	P1	$P2_12_12_1$	$H3_2$	H3
Cell dimensions				
a, b, c (Å)	73.32, 73.33, 116.46	106.00, 119.70, 182.18	136.38, 136.38, 53.44	135.40, 135.40, 51.69
α,β,γ, (°)	91.04, 72.59, 80.07	90, 90, 90	90, 90, 120	90, 90, 120
No. of measured reflections	114998 (9158)	864376 (43789)	196254 (14984)	253965 (13080)
No. of independent reflections	58162 (4564)	117960(5795)	36549 (2705)	45732 (2309)
Resolution (Å)	46.53-2.75 (2.83-2.75)	100.4-2.2 (2.24-2.20)	68.19-1.76 (1.81-1.76)	25.59-1.61 (1.64-1.61)
$CC_{1/2}$	0.978 (0.650)	0.998 (0.667)	0.997 (0.435)	0.997 (0.475)
Mean I/σI	4.5 (1.6)	13.3 (3.1)	11.5 (1.6)	10.8 (1.3)
Completeness	98.4 (98.0)	100 (100)	99.5 (99.9)	99.9 (100)
Redundancy	2.0 (2.0)	13.3 (3.3)	5.4 (5.5)	5.6 (5.7)
Refinement				
R_{work}/R_{free}	0.22/0.25	0.21/0.27	0.17/0.21	0.13/0.19
No. atoms				
Protein	14732	14919	2771	2822
Ligand/Ions	-/-	232/-	24/1	14
Water	-	538	268	410
B-factors				
Protein	30.97	38.3	30.5	19.9
Ligand/Ions	-/-	57.8	28.7/51.7	22.7
Water	-/-	38.3	38.7	31.6
r.m.s deviations				
Bond lengths	0.011	0.017	0.013	0.014
Bond angles	1.50	1.77	1.51	1.53
PDB code	6EUJ	6EUF	6EUI	6EUG

*(Values in parenthesis are for the highest resolution shell).

#5% of the randomly selected reflections excluded from refinement.

+Calculated using MOLPROBITY.

Supplementary Table 6b. Data statistics and refinement details.

	BT3683GalDNJ	BT0290Gal	BT3674
Data collection			
Date	13/02/16	26/01/14	02/02/16
Source	IO4-1	IO2	IO2
Wavelength	0.97	0.98	0.98
Space Group	P1	P22121	$P6_5$
Cell dimensions			
a, b, c (Å)	52.18, 77.20, 78.69	62.04, 101.16, 143.21	136.81, 136.81, 135.56
α,β,γ, (°)	114.00, 101.78, 100.78	90,90,90	90, 90, 120
No. of measured reflections	134165 (8686)	640375 (28390)	886170 (65903)
No. of independent reflections	67881 (4470)	92027 (4185)	77030 (5696)
Resolution (Å)	67.23-2.00 (2.05-2.00)	46.61-1.75 (1.77-1.75)	68.40-2.16 (2.22-2.16)
$CC_{1/2}$	0.947(0.685)	-	0.998 (0.578)
Mean I/σI	7.2 (2.8)	13.8 (2.5)	12.9 (1.4)
Completeness	95.9 (93.9)	99.3 (92.0)	100 (99.9)
Redundancy	2.0 (1.9)	7.0 (6.8)	11.5 (11.6)
Refinement			
R_{work}/R_{free}	0.18/0.23	0.16/0.18	0.17/0.22
No. atoms			
Protein	8229	6153	4910
Ligand/Ions	33/3	13/-	-/2
Water	618	625	727
B-factors			
Protein	24.13	17.6	35.8
Ligand/Ions	32.48/32.48	11.4/-	-/30.06
Water	30.83	26.7	40.9
r.m.s deviations			
Bond lengths	0.013	0.010	0.018
Bond angles	1.55	1.43	1.80
PDB code	6EUH	6EON	6EX6

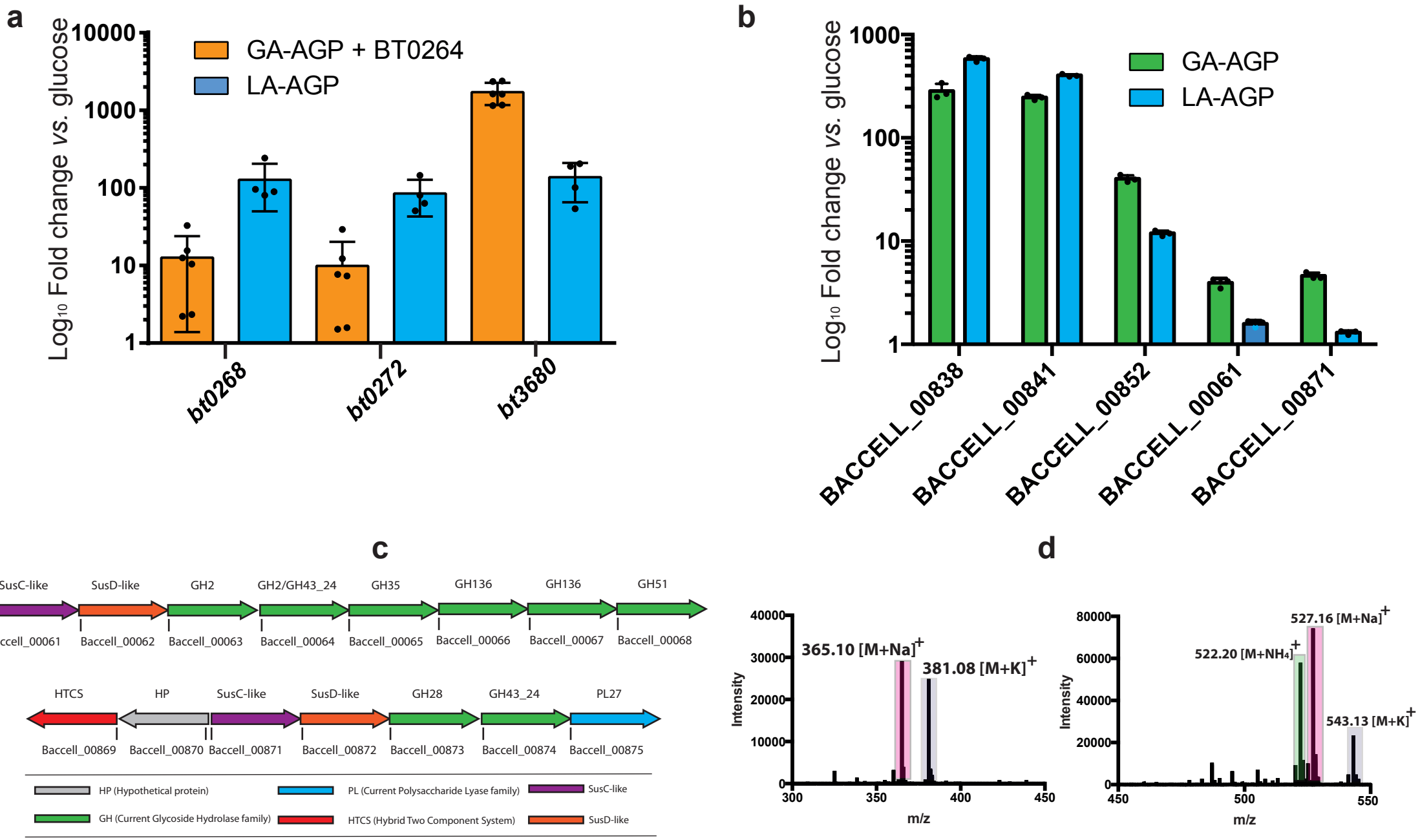
*(Values in parenthesis are for the highest resolution shell).

#5% of the randomly selected reflections excluded from refinement.

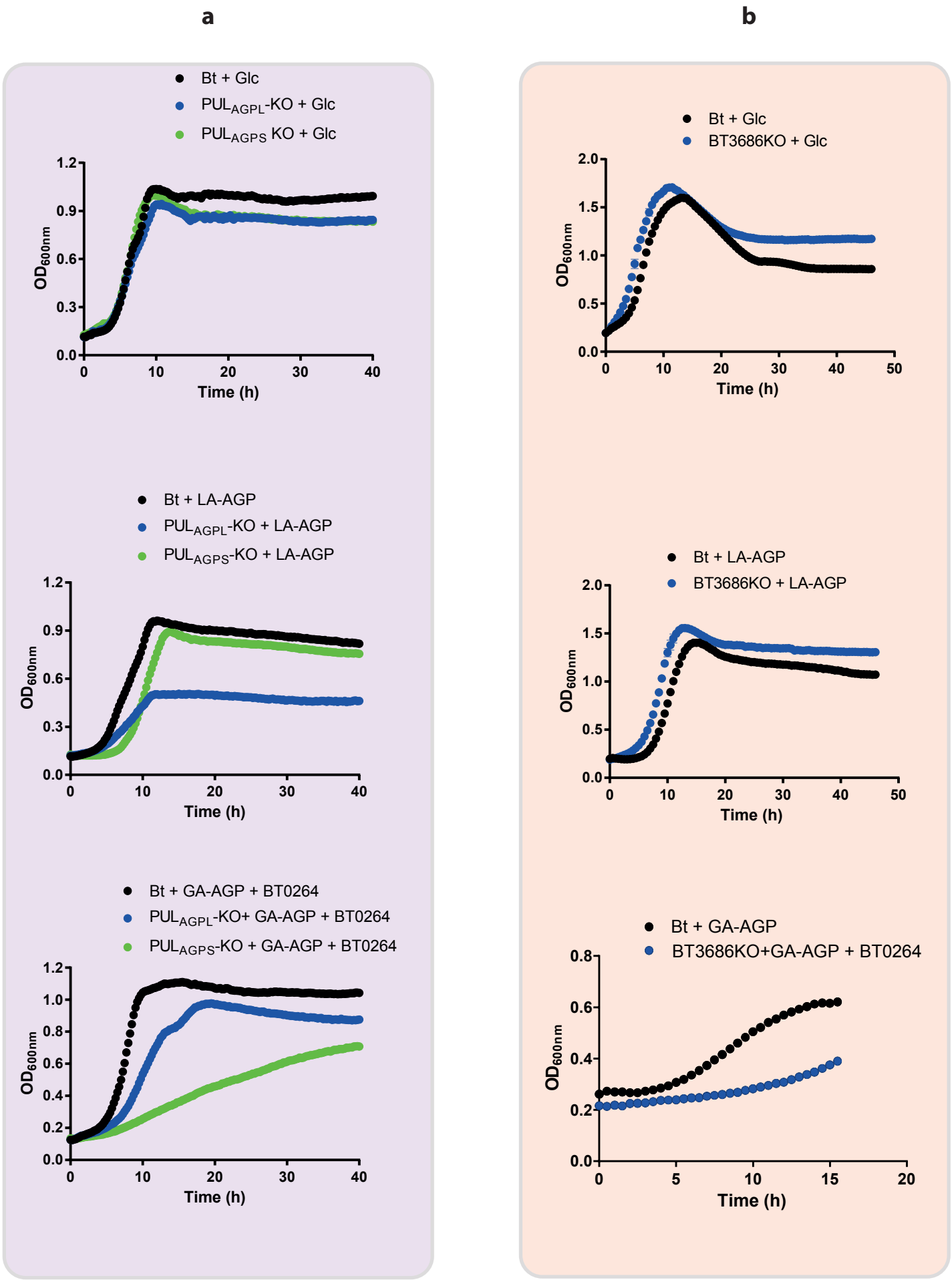
+Calculated using MOLPROBITY.

Supplementary Table 7. ^{1}H and ^{13}C NMR assignments of the AGP tetra- and heptasaccharides at 25 °C in D_2O.

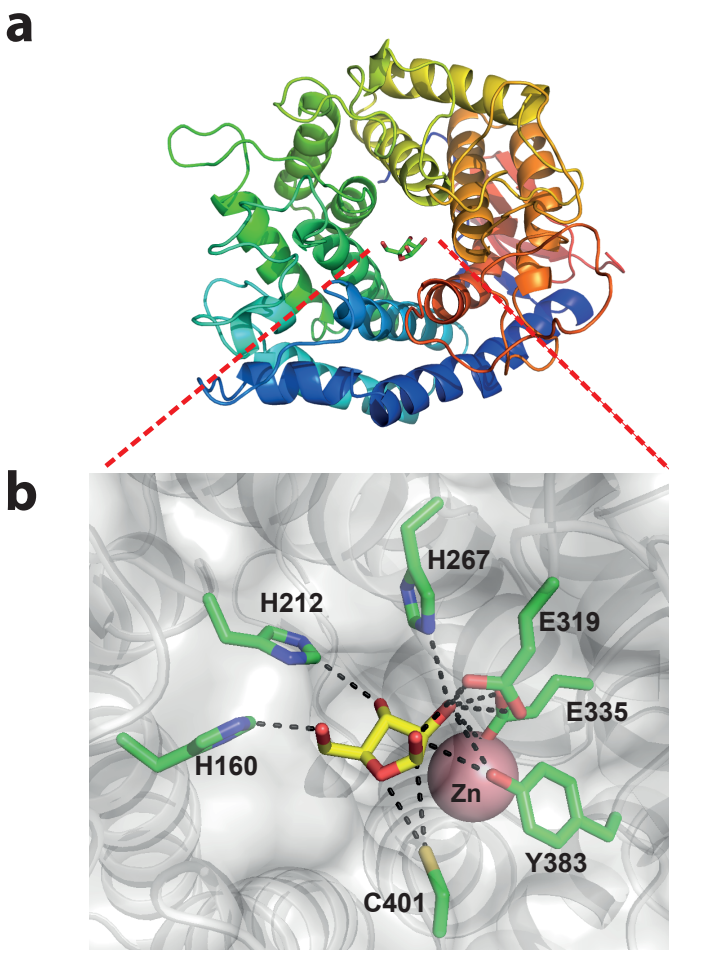
	H1	H2	H3	H4	H5	H6	C1	C2	C3	C4	C5	C6
Tetrasaccharide												
β-D-Gal*p*$_1$	4.597	3.492	3.651	3.956	3.902	3.896, 4.047	97.21	72.60	73.45	69.61	74.56	70.28
α-D-Gal*p*$_1$	5.265	3.800	3.869	4.018	4.280	3.849, 4.039	93.16	69.09	69.76	70.15	70.12	70.39
β-D-Gal*p*$_2$	4.433	3.529	3.651	3.934	3.894	3.897, 4.016	103.94	71.53	73.32	69.41	74.56	70.34
β-D-Glc*p*A$_3$	4.558	3.369	3.605	3.639	3.955	-	103.52	73.97	74.91	79.65	75.23	174.11
α-L-Rha*p*$_4$	4.742	3.926	3.746	3.432	4.013	1.245	101.74	71.04	70.87	72.65	69.89	17.24
Heptasaccharide												
β-D-Gal*p*$_1$	4.599	3.494	3.657	3.963	3.897	3.895, 4.066	97.22	72.60	73.41	69.60	74.65	70.36
α-D-Gal*p*$_1$	5.272	3.803	3.872	4.019	4.279	4.052, 3.856	93.18	69.09	69.99	70.13	70.18	70.41
β-D-Gal*p*$_2$	4.515	3.723	3.875	4.180	3.985	3.883, 4.001	103.74	71.39	80.43	74.47	74.41	70.66
β-D-Glc*p*A$_3$	4.496	3.360	3.564	3.566	3.725	-	103.72	74.08	75.05	79.94	77.11	176.07
α-L-Rha*p*$_4$	4.728	3.927	3.758	3.422	4.018	1.244	101.54	71.08	70.84	72.71	69.72	17.28
α-L-Ara*f*$_5$	5.386	4.159	3.912	4.068	3.685, 3.799	-	108.92	82.22	77.60	84.93	62.01	-
α-L-Ara*f*$_6$	5.270	4.391	3.939	4.281	3.859, 3.742	-	110.33	80.87	85.68	83.73	61.97	-
α-D-Gal*p*$_7$	5.026	3.802	3.865	3.982	4.066	3.750, 3.750	100.93	69.12	69.94	70.05	72.17	62.00

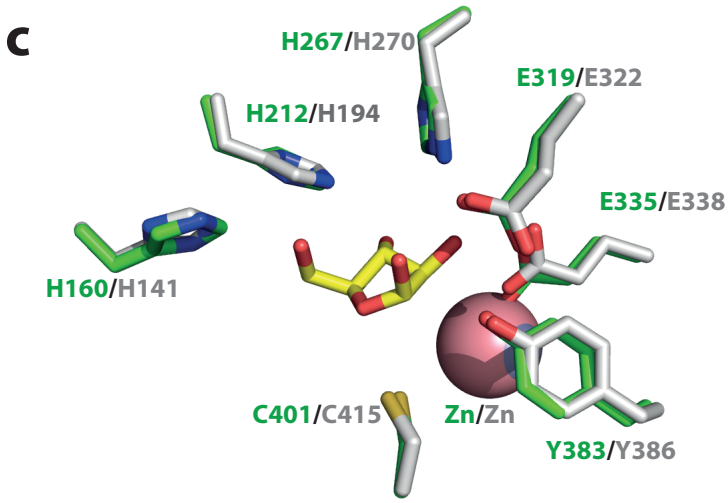


Supplementary Figure 1 RT-PCR of PULs activated by AGP and mass spectrometry of galactosidase reaction products. In **a** RT-PCR of the SusC genes in *B. thetaiotaomicron* PULAGPS and PULAGPL were used to determine the upregulation of the two loci in the bacterium cultured on LA-AGP and GA-AGP (biological replicates n = 6). In **b** RT-PCR of susC genes in loci containing GH16 and/or GH43_24 genes in *B. cellulosilyticus* cultured on LA-AGP or GA-AGP (biological replicates n = 3). In **a** and **b** error bars are standard errors of the mean.The SusC genes baccell00838, baccell00841 and baccell00852, which were substantially upregulated by both AGPs, are in a single large PUL whose content is shown in Supplementary Fig. 15. The structure of the PULs containing the other two susC genes, baccell00061 and baccell00871, which were only marginally upregulated, are shown in **c**. In **d** [example of independent replicates (n = 2)] the two major products generated by exo-β1,3-galactosidase BT0265 acting on LA-AGP (see Fig. 2) were subjected to LC-MS (see Methods). The mass of the oligosaccharides (relevant peaks shaded according to the adduct ion) were consistent with β1,6-galactobiose and β1,6-galactotriose.

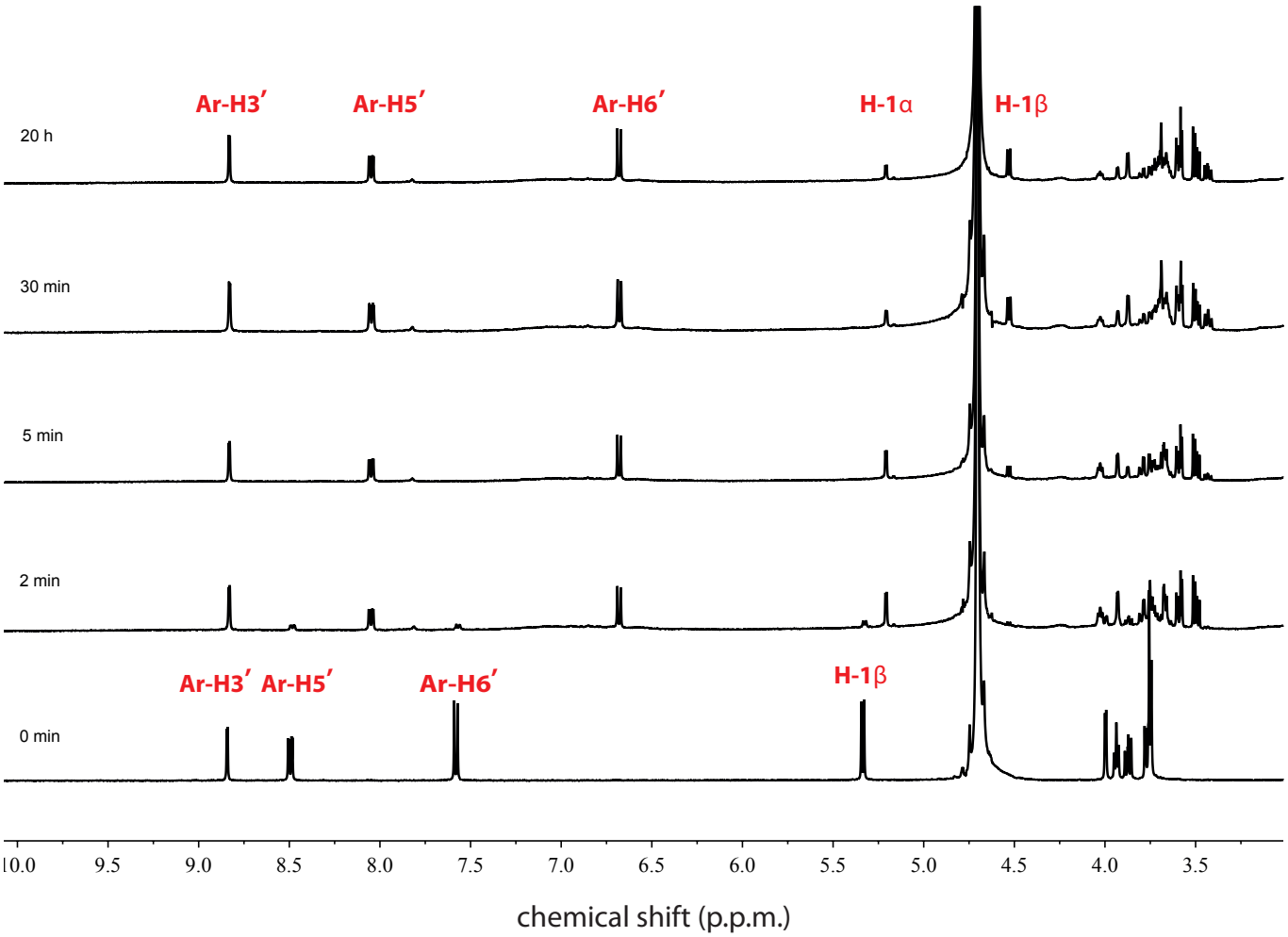


Supplementary Figure 2. Growth profile of wild type and mutants of B. thetaiotaomicron. Wild type (WT) and mutants were cultured on minimal medium containing the appropriate AGP at 10 mg/ml and growth was monitored at 600 nm every 15 min using an automated spectrophotometer (independent replicates n = 3, error bars s.e.m.). **a** shows the growth profiles of mutants in which the two AGP PULs have been deleted (KO), and **b** displays the growth curves of the mutant in which the gene encoding the rhamnosidase BT3686 had been deleted (KO).

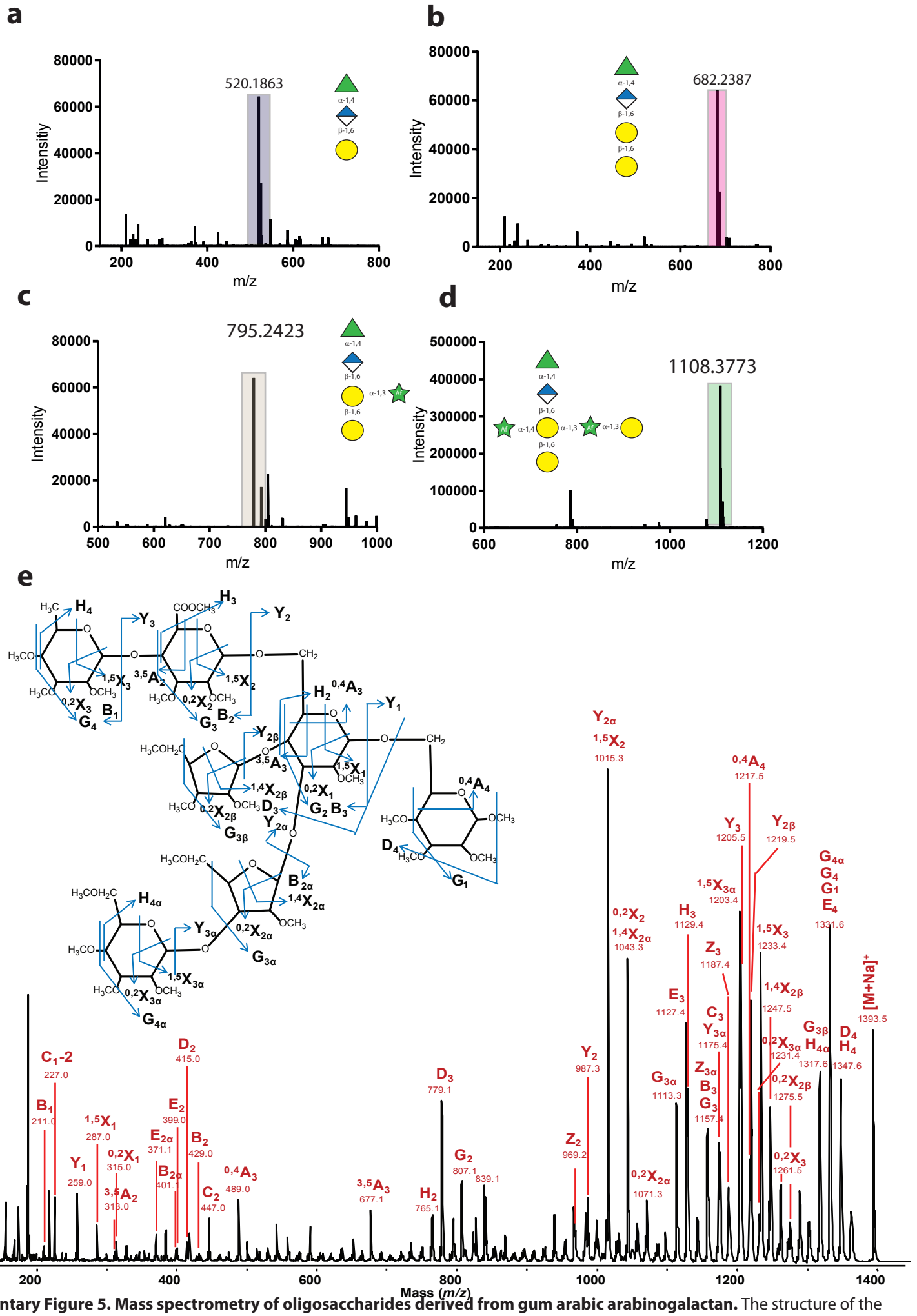




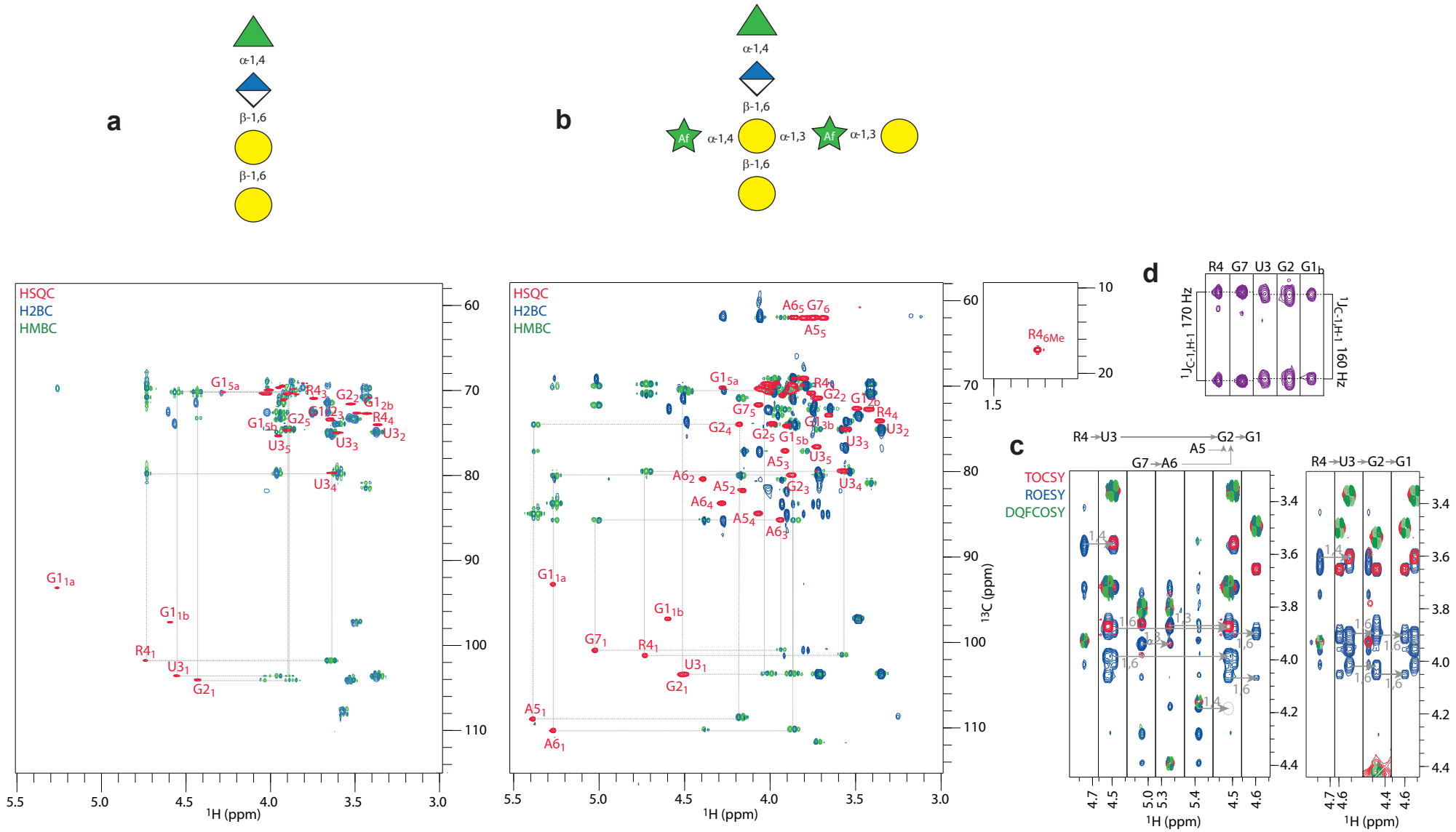
Supplementary Figure 3. Crystal structure of BT3674 (PDN 6EX6). **a**, schematic of BT3674 colour ramped from the N-terminus (blue) to the C-terminus (red). **b**, the active site amino acids of BT3674 (carbons coloured green) that interact with arabinofuranose (carbons in yellow) are shown with polar contacts depicted by broken black lines. **c**, an overlay of the active site residues of BT3674 (amino acids coloured green and arabinofuranose yellow) with the *Bifidobacterium longum* β-L-arabinofuranosidase HypBA1 (PDB code 3WKX; amino acids shown in light grey).



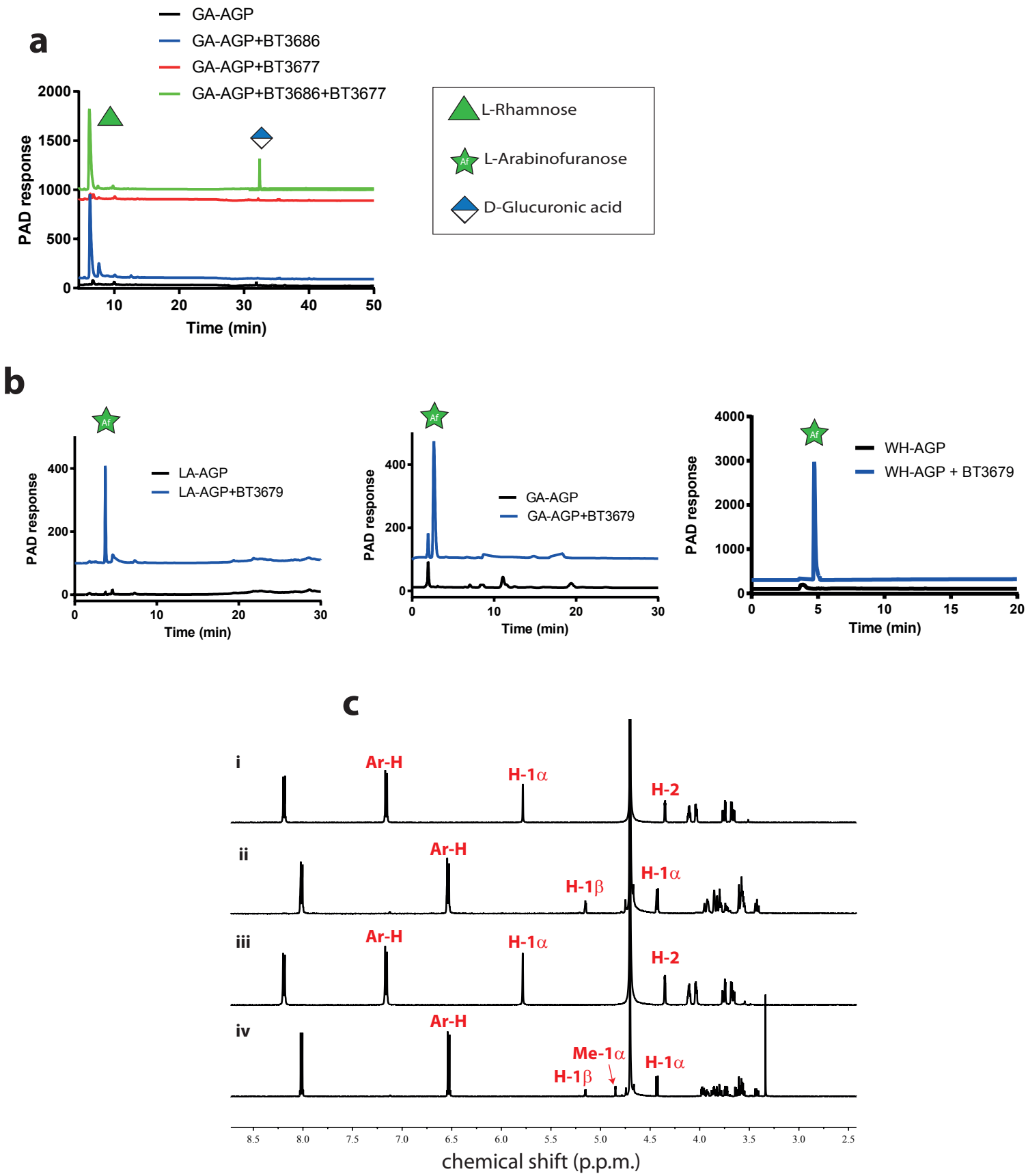
Supplementary Figure 4. ^{1}**H-NMR analysis of the activity of BT3685.** Enzyme at 20 µM was incubated with 5 mM 2,4-dinitrophenyl β-D-galactopyranoside in 20 mM sodium phosphate buffer pH 7.0 implemented with 150 mM NaCl standard conditions in a solvent of D_2O. Spectra were recorded at the indicated times. The data presented are examples of biological replicates (n = 2). Ar means Aromatic ring of 2,4-dintrophenyl β-D-galactopyranoside.



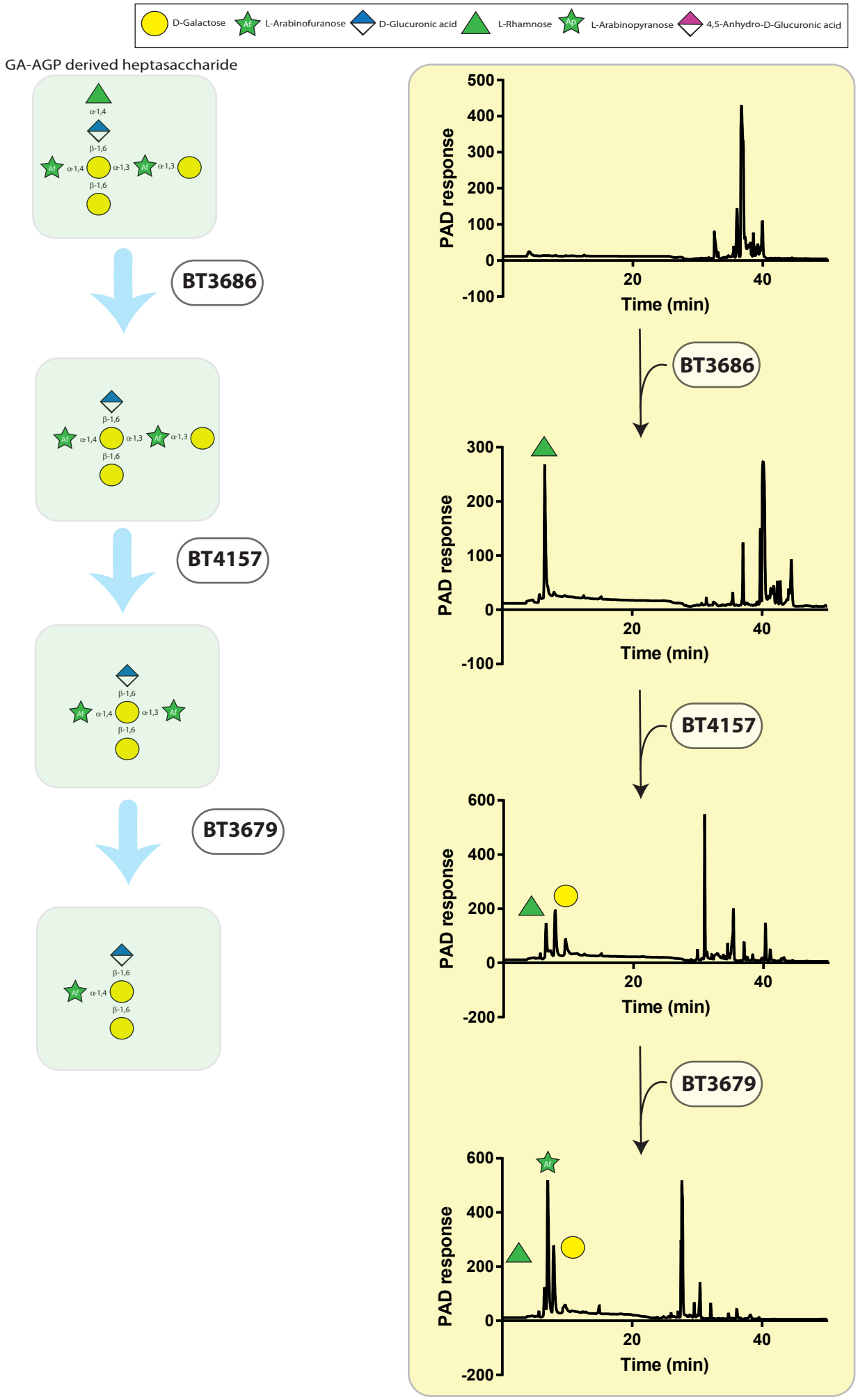
Supplementary Figure 5. Mass spectrometry of oligosaccharides derived from gum arabic arabinogalactan. The structure of the oligosaccharides and their size determined by mass spectrometry are shown in **a**, **b**, **c**, and **d**. Relevant peaks were shaded according to the adduct ion; grey H^+, mouve, NH_4^+, pink Na^+, green K^+. The tandem mass spectrometry (MS/MS) spectrum of the per-methylated heptasaccharide is shown in **e**, with inset a schematic structure of the oligosaccharide. The blue arrows in the inset illustrate the generation of the fragment ions seen in the MS/MS. 2D NMR in the case of the tetrasaccharide and the heptasaccharide was also used to determine the structure of these oligosaccharides (see **Supplementary Fig. 6)**. **a, b, c** and **d** are examples of independent replicates (n = 2), while collectively, the data in **e** are examples of replicates n = 3.



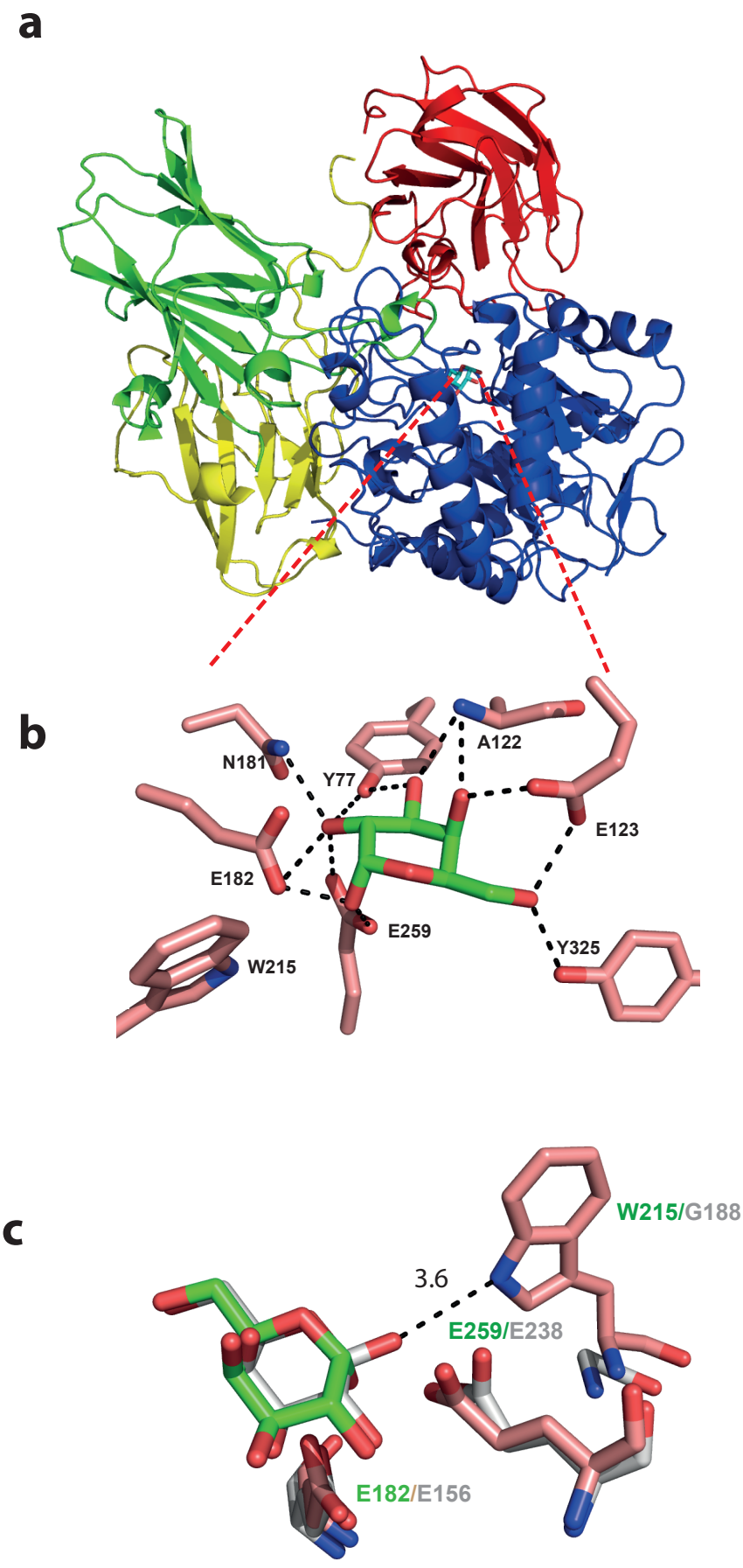
Supplementary Figure 6. NMR analysis of the tetra- and heptasaccharides. Derived structures of (**a**) the tetrasaccharide and (**b**) the heptasaccharide, shown above 2D ^{1}H-^{13}C HSQC (Heteronuclear Single Quantum Coherence), H2BC and HMBC spectra (red, blue and green, respectively) displaying the assignment of all well-resolved H, ^{13}C HSQC peaks. (Assignments in the more congested region are not shown but are listed in Supplementary Table 7). Glycosidic bonds were evident from inter-residue cross-peaks in the HMBC (grey dotted lines) and downfield shifts of the 13C resonance positions of the linked carbon. The methyl group in the 6-position of α-L-Rhap is shown in a separate panel (**b**, top right). (**c**) H-1 strip plots from 2D H-H TOCSY [Total Correlation SpectroscopY (red)], ROESY [Rotating-frame Overhauser Effect SpectroscopY (blue)] and DQFCOSY [Double Quantum Filtered COrrelation SpectroscopY (green)] spectra showing the NOE connectivity in the hepta- and tetra-saccharides arising from the glycosidic linkages.(**d**) Strip plots of the C-1,H-1 peaks from an F1-coupled C HSQC showing the α- and β-anomeric linkages of R4 and G7 (alpha, JC-1,H-1 ~ 170 Hz) and G1β (beta, 1JC-1,H-1 ~ 160 Hz) in the heptasaccharide and U3 and G2 (beta, 1JC-1,H-1 ~ 160 Hz) in the tetrasaccharide (the U3 and G2 signals were not perfectly resolved in the heptasaccharide). The data are, collectively, examples of replicates n = 3.



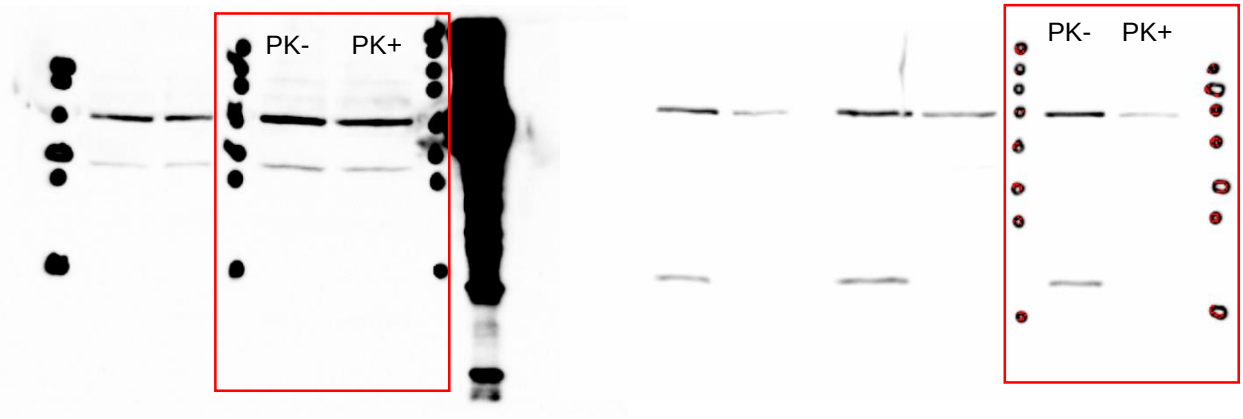
Supplementary Figure 7. Biochemical characterization of two enzymes that are the founding members of previously unknown GH families. a, activity of BT3677. The panel shows the release of GlcA from GA- AGP using HPAEC. The conditions were 20 mM sodium phosphate buffer, pH 7.0, $[GA]_0$ = 10 mg/ml, $[BT3686]_0$= 1 μM and $[BT3677]_0$= 1 μM. **b**, shows HPAEC analysis of BT3679 incbuted with LA-AGP, or GA- AGP and WH-AGP. The release of arabinose was visible. **c**, 1H-NMR analysis of the activity of the enzyme. BT3679 was incubated with 4-nitrophenyl α-L-arabinofuranoside (4NPA) in the absence (**ii**) and presence (**iv**) of 2.5 M methanol. As controls 4NPA was incubated in the absence of enzyme plus (**iii**) or minus (**i**) methanol. The reaction products were lyophilised, resuspended in D2O and analysed by 1H-NMR. The signal corresponding to the anomeric proton of methyl α-L-arabinofuranose is labelled Me-1α. Ar-H refers to a proton in the aromatic ring. H-1 is the anomeric proton in 4NPA (**i** and **iii**) and arabinose in **ii** and **iv**. H-2 (**i** and **iii**) is the proton of C2 of the arabinose in 4-NPA. The figure is representative of biological replicates (n = 2).

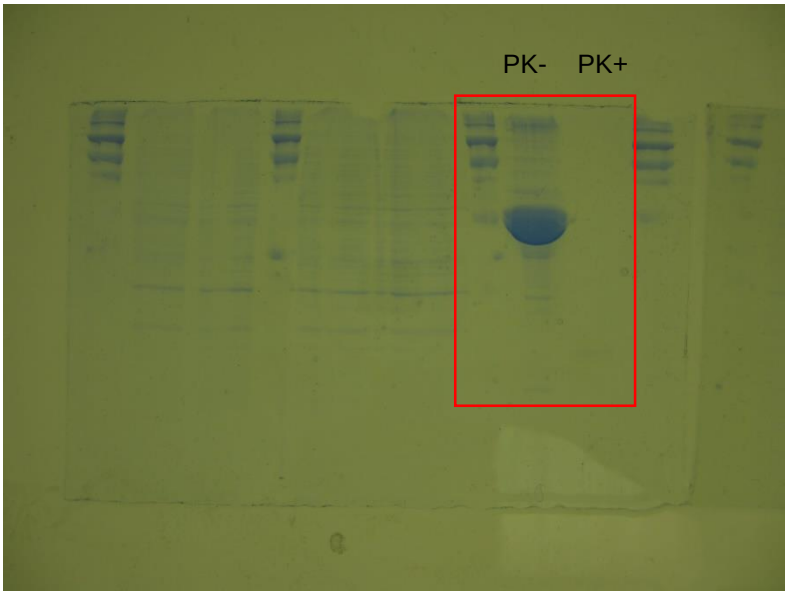


Supplementary Figure 8. Enzymes active against the GA-AGP derived heptasacchride. The heptasaccharide substrate was released from GA-AGP by the exo-β1,3-galactosidase BT0265 and then purified by size exclusion chromatography. Individual B. thetaiotaomicron enzymes (1 μM) were incubated with the glycan (5 mM) for 16 h at 37 ºC in 20 mM sodium phosphate buffer, pH 7.0. Monosaccharides and oligosaccharides generated were identified by HPAEC-PAD. BT4157 is a non-specific α-galactosidase derived from RGI-PUL25. The example presented here is representative of biological replicates (n =2).



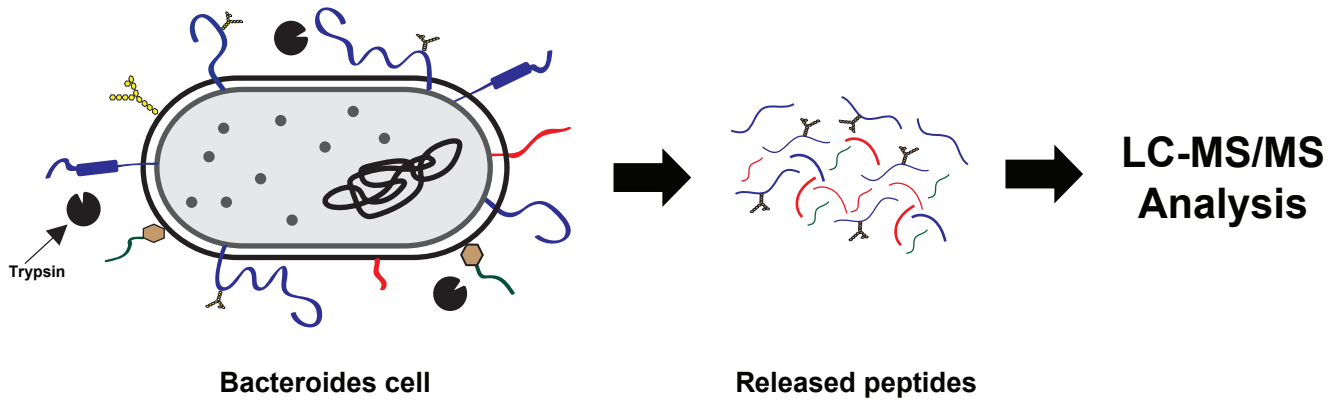
Supplementary Figure 9. Crystal structure of BT0290 (PDB 6EON). **a**, schematic of BT0290 in which the N-terminal TIM barrel catalytic domain as blue, followed by the three β-sandwich domains coloured green, yellow and red from the N- to C-termini. **b**, the active site amino acids of BT0290 (carbons coloured salmon pink) that interact with galactose (carbons in green) are shown with polar contacts depicted by broken black lines. The electron density map (2Fo – Fc) of the galactose is shown in blue mesh at 1.5 °. **c**, an overlay of the structure of BT0290 (amino acids coloured salmon pink and galactose blue) with the Streptococcus β1,3-galactosidase BgaC (PDB code 4EBC; amino acids shown in light grey and galactose in salmon pink) showing the catalytic acid base and nucleophile in the active site, and BT0290 Trp215, the proposed specificity determinant, in the +1 subsite. The distance (3.6 Å) between the tryptophan and O1 of β-galactose in the active site is shown as a dashed line.



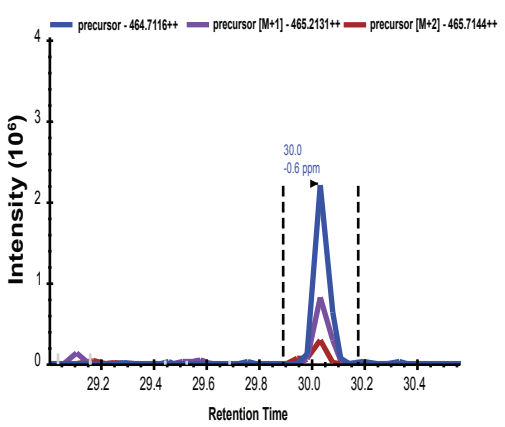


Supplementary Figure 10. Western blot of BT2064 and BT4662 and SDS-PAGE gel of purified BT0264. Western blot detection of **a** BT0264 and **b** a known surface enzyme (BT4662) in LA-AGP/heparin cultured *B. thetaiotaomicron,* after treatment of the bacterial cells with proteinase K (PK+) or untreated (PK-). Purified recombinant BT0264 was also subjected to proteinase treatment to verify the enzyme is sensitive to the proteinase. The regions of the gels presented in **Figure 5a**, where the lanes are fully annotated, are indicated by red rectangles.

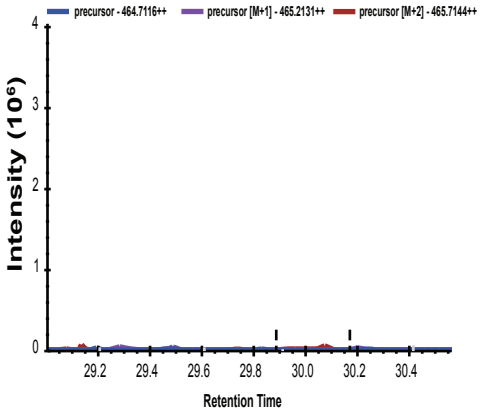
a



b



c



d

> Baccell_00844
MKNKIKIWVALCVAACSITFSACDEDNDFGYNGKLSLNYLGIKQASEIWDGAQCNIVTALKEPQEGEVQKVKNFNYRLNLALYQNRKAEKDA

TVDLVIASDSLKKAIALVGTSSVYTVYADAELLPEEYYNLSASKMELSAGCKKSEEVELNVYSSKLIALVQDEWKKDVTFVLPVQIQNSTSYSIND

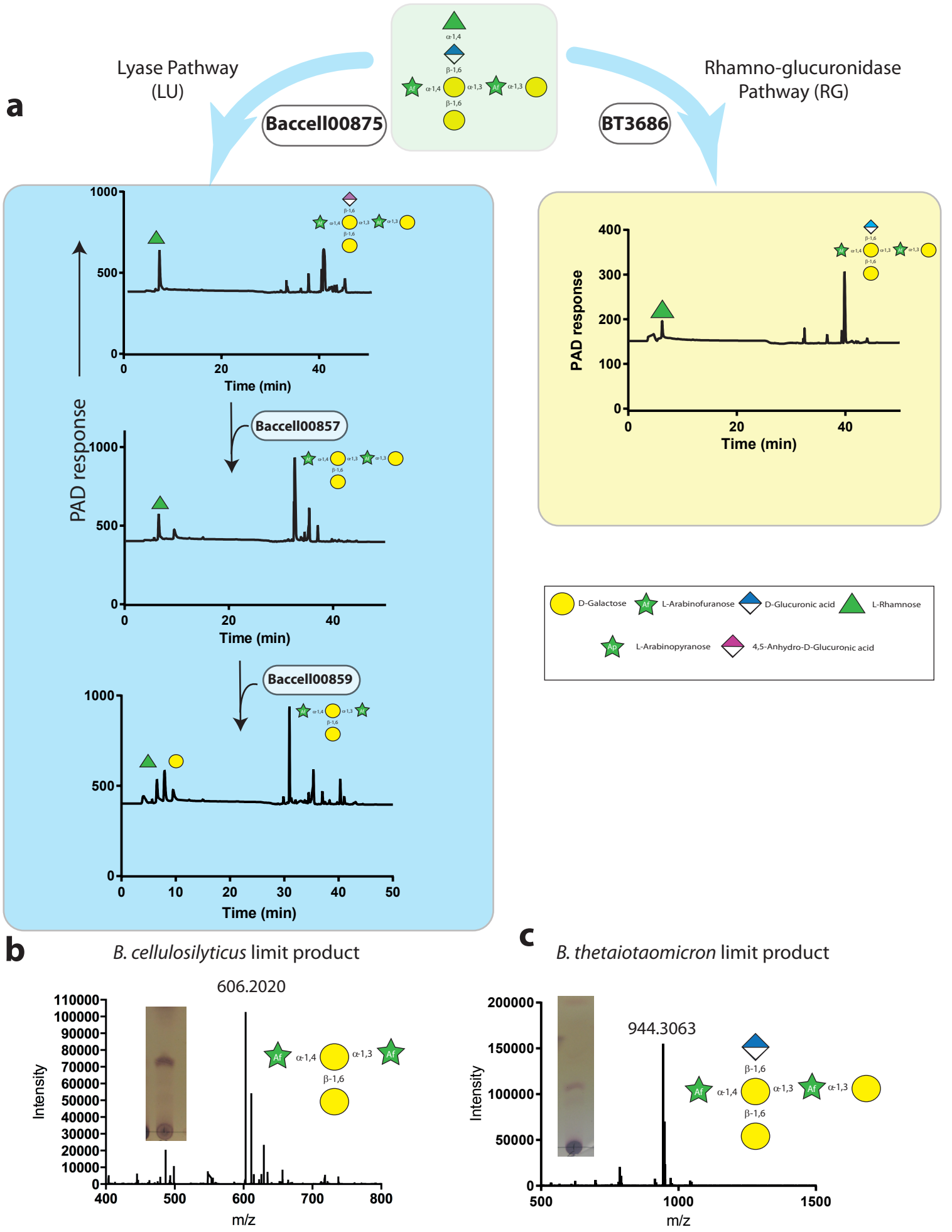
KTNTMMFFFNVTYVDPGEEYFADGEGVPDDHELEGGYKLVWHDEFNGTGAPNEMWR[YEEGFQR][NEEDQWYK]KENVEMKKNALVFTAK

QER[VKNPNYNPNATGGNSWK]QTREAEYTSACVVAQNKYAFKYGKLVVRAKIPIEQGGWPAIWSTGNWYEWPLGGEIDFLEFYKKIHANL

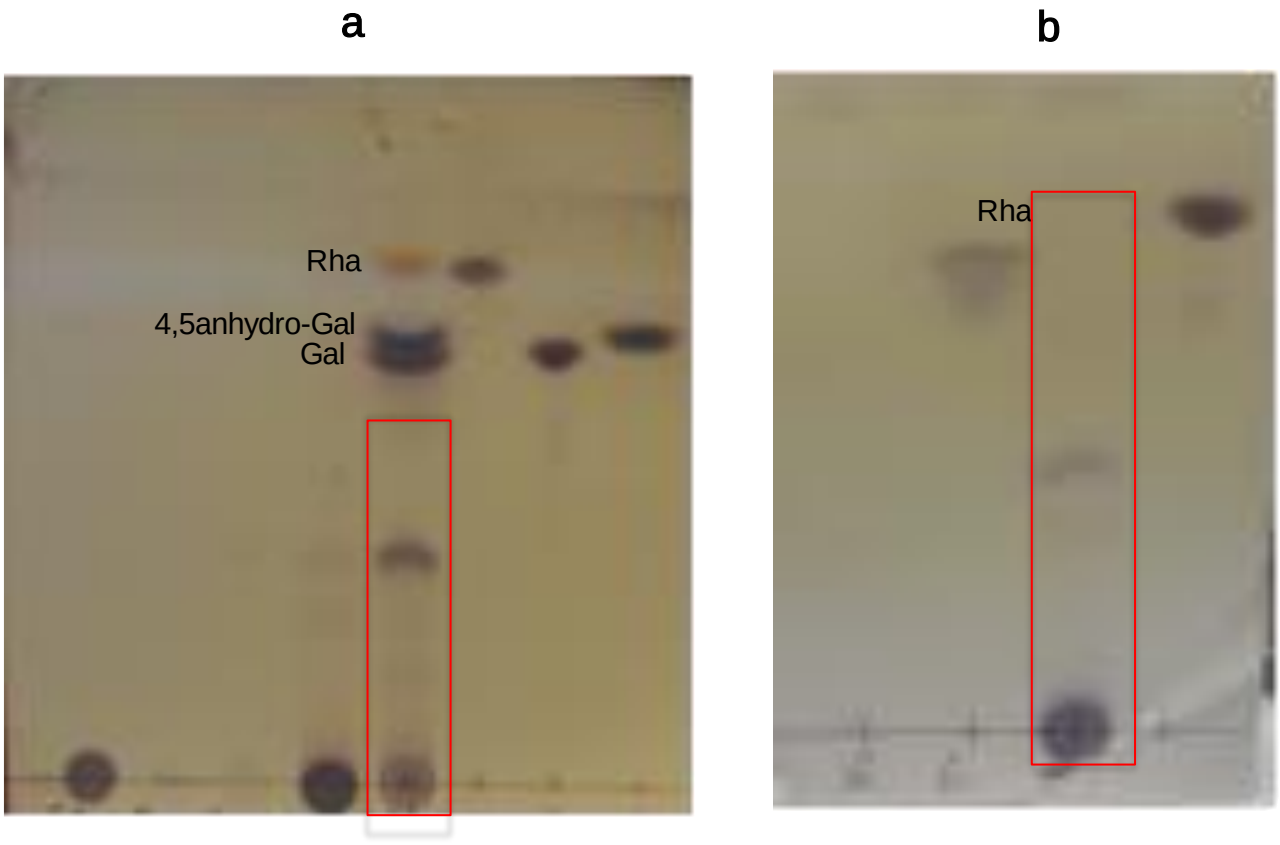
CWGGNKR[WDGSWNSANYPITDFTSK]DAKWAEKYHVWMMDWDEKYIRIYLDDVLLNETDLSTTYNKGDHGAGEGGYINPYSNDLEGFG

QLMMLNLAIGGSNGR[PIEATFPLEYR]VDYVRVYQKK

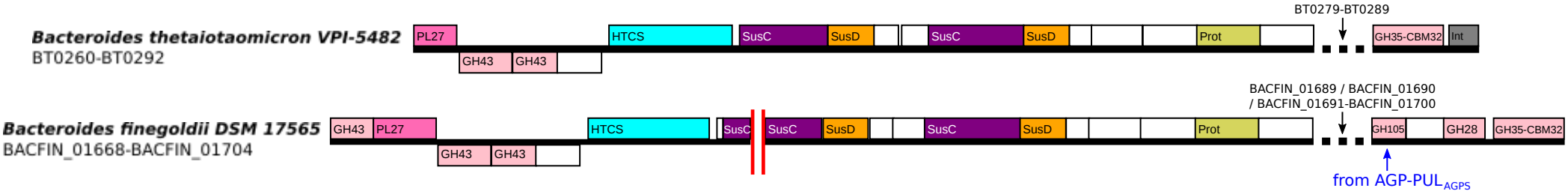
Supplementary Figure 11. Surface proteome (surfome) analysis to determine BACCELL_00844 subcellular protein localiza on in *B. thetaiotaomicron*. (**a**) Experimental workflow used for analysis of the *Bacteriodes thetaiotaomicron* surfome. Extracted ion chromatograms of the YEEGFQR tryp c pep de, which is in the BACCELL_00844-expressing *B. thetaiotaomicron* strain (**b**), but absent in wild-type *B. thetaiotaomicron* (**c**). **(d)** Sequence of Baccell_008444 with the five pep des, iden fied by mass spectrometry, that are unique to this enzyme indicated in red between square brackets. The data in **b** and **c** are examples of independent replicates (n = 2).



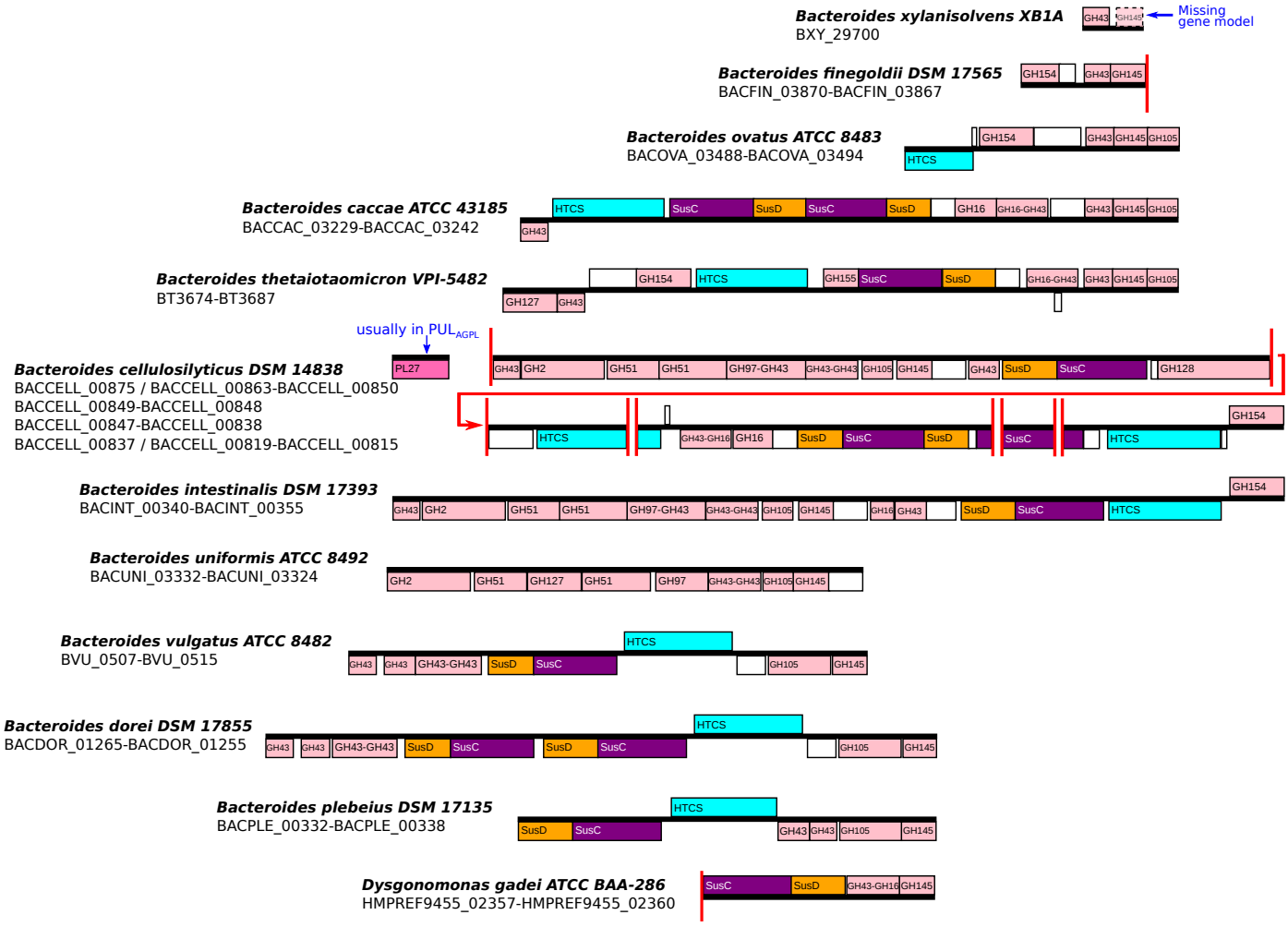
Supplementary Figure 12. Degradation pathways of GA-AGP that resulted in the limit products generated by *B. thetaiotaomicron* and *B. cellulosilyticus*. **a**, the enzyme systems that degrade the heptasaccharide in GA-AGP that resulted in the limit products observed in the two organisms. **b**, **c**, structures of the limit products produced by *B. thetaiotaomicron* and *B. cellulosilyticus*, respectively. The smaller product generated by *B. cellulosilyticus* reflects the use of the LU pathway in which the GH105 unsaturated glucuronidase Baccell00857 can cleave its target linkage when Gal at the +1 subsite of the enzyme is decorated at O4. The α-Gal is released from the oligosaccharide by the GH97 α-galactosidase Baccell00859, which is encoded by the B. cellulosilyticus AGP PUL. The data in are examples of biological replicates (**a,** n = 4; **b** and **c,** n =2) .



Supplementary Figure 13. Full TLC plate showing the band corresponding to the GA-AGP- derived limit products subjected to mass spectrometry. The heptasacchride shown in **Supplementary Fig. 12a** was incubated with cell free extracts of (**a**) *B. cellulosilyticus* and (**b**) *B. thetaiotaomicron* for 16 h at room temperature. The region of the two TLC plates containing the single oligosaccharide species displayed in **Supplementary Fig. 12b** and **12c** are identified by a red rectangle. The monosacchrides gerated by the cell-free extracts are identified as rhamose (Rha), galactose (Gal) and 4,5anhydro-galactose (4,5anhydro-Gal).



Supplementary Figure 14. Genomic view of *B. thetaiotaomicron* PUL$_{AGPL}$ and microsyntenic regions in the species studied in this work.
Protein-coding genes are depicted by colored rectangles to highlight the following functional modules: GHs in light pink, PLs in dark pink, HTCSs regulators in cyan, SusC transporters in purple, SusD outer membrane proteins in orange, peptidases in gold, integrases in grey. Genes are represented either above or below a central black line to represent the coding strand. The central balck line is dotted to indicate a large genomic region of unknown genes (locus tags indicated above) which has not been represented. When PUL genes are split across several scaffolds, due to incomplete genome assembly, the scaffold limits are indicated by vertical red bars. A blue arrow shows the translocation of a gene from PUL$_{AGPS}$ into the *B. finegoldii* PUL.



Supplementary Figure 15. Genomic view of B. thetaiotaomicron PUL_{AGPS} and microsyntenic regions in the species studied in this work. Protein-coding genes are depicted by colored rectangles to highlight the following functional modules: GHs in light pink, PLs in dark pink, HTCSs regulators in cyan, SusC transporters in purple, SusD outer membrane proteins in orange. Genes are represented either above or below a central black line to represent the coding strand. When PUL genes are split across several scaffolds, due to incomplete genome assembly, the scaffold limits are indicated by vertical red bars. *B. cellulosilyticus DSM 14838* genomic fragments were ordered according to *B. cellulosilyticus WH2* complete assembly (not shown). A blue arrow shows the translocation of a gene usually found in PUL_{AGPL} (in *B. thetaiotaomicron* and *B. finegoldii*) close to PUL_{AGPS} in *B. cellulosilyticus* (which do not have a PUL_{AGPL}), and a missing gene model in *B. xylanisolvens*. PULs were represented for species that displayed at least two orthologous and microsyntenic CAZymes.