Mining the human gut microbiome identifies mycobacterial D-arabinan degrading enzymes

Omar Al-Jourani[1,♣], Samuel Benedict[2,♣], Jennifer Ross[1,♣], Abigail Layton[2], Phillip van der Peet[3], Victoria M. Marando[4], Nicholas P. Bailey[1], Tiaan Heunis[1], Joseph Manion[1], Francesca Mensitieri[1], Aaron Franklin[2], Javier Abellon-Ruiz[1], Sophia L. Oram[1], Lauren Parsons[1], Alan Cartmell[5], Gareth S. A. Wright[6], Arnaud Baslé[1], Matthias Trost[1], Bernard Henrissat[7], Jose Munoz-Munoz[8], Robert P. Hirt[1], Laura L. Kiessling[4], Andrew Lovering[2], Spencer J. Williams[3], Elisabeth C. Lowe[1,*], Patrick J. Moynihan[2*]

Addresses:

[1]Newcastle University Biosciences Institute, Medical School, Newcastle University, Newcastle upon Tyne, U.K., NE2 4HH

[2]Institute of Microbiology and Infection, School of Biosciences, University of Birmingham, Birmingham, U.K., B15 2TT

[3]School of Chemistry and Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Parkville, Victoria 3010, Australia

[4] Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA

The Broad Institute of Harvard and MIT, Cambridge, MA, USA The Koch Integrative Cancer Research Institute, Massachusetts Institute of Technology, Cambridge, MA, USA

[5] Department of Biochemistry and Systems Biology, Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, UK.

[6] School of Life Sciences, University of Essex, Colchester, UK

[7] Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia;

Department of Biotechnology and Biomedicine (DTU Bioengineering), Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

[8] Microbial Enzymology Group, Department of Applied Sciences, Northumbria University, Newcastle upon Tyne, United Kingdom.

♣ These authors contributed equally to this work

* To whom correspondence should be addressed:
Patrick Moynihan, p.j.moynihan@bham.ac.uk
Elisabeth Lowe, elisabeth.lowe@ncl.ac.uk

Key words:

carbohydrate-active enzymes (CAZymes); glycoside hydrolase (GH); tuberculosis (TB); GH172; DUF4185

Abbreviations:

AG, arabinogalactan; DUF, domain of unknown function; GH, glycoside hydrolase; IC-PAD, ion chromatography/ pulsed amperometric detection; LAM, Lipoarabinomannan; PUL, polysaccharide utilization locus;

Abstract

Division and degradation of bacterial cell walls requires coordinated action of a myriad of enzymes. This particularly applies to the elaborate cell walls of acid-fast organisms such as *Mycobacterium tuberculosis*, which consist of a multi-layered cell wall that contains an unusual glycan called arabinogalactan. Enzymes that cleave the D-arabinan core of this structure have not previously been identified in any organism. We have interrogated the diverse carbohydrate degrading enzymes expressed by the human gut microbiota and uncovered four families of glycoside hydrolases with the capability to degrade the D-arabinan or D-galactan components of arabinogalactan. Using novel exo-D-galactofuranosidases from gut bacteria we generated enriched D-arabinan and used it to identify *D. gadei* as a D-arabinan degrader. This enabled the discovery of endo- and exo- acting enzymes that cleave D-arabinan. We have identified new members of the DUF2961 family (GH172), and a novel family of glycoside hydrolases (DUF4185) that display endo-D-arabinofuranase activity. The DUF4185 enzymes are conserved in mycobacteria and found in many microbes, suggesting that the ability to degrade mycobacterial glycans plays an important role in the biology of diverse organisms. All mycobacteria encode two conserved endo-D-arabinanases that display different preferences for the D-arabinan-containing cell wall components arabinogalactan and lipoarabinomannan, suggesting they are important for cell wall modification and/or degradation. The discovery of these enzymes will support future studies into the structure and function of the mycobacterial cell wall.

70 **Introduction**

Growth and division of all bacteria is a carefully orchestrated process requiring
72 the coordinated action of a host of enzymes. Acid-fast organisms such as *Mycobacterium tuberculosis* possess unusual cell wall glycans and lipids which
74 require additional enzymatic machinery during this process. The core of the cell wall structure is conserved amongst mycobacteria and consists of three layers[1,2]. Like
76 other bacteria, peptidoglycan forms the basal layer of the cell wall, though the precise architecture is unknown. At the other extremity of the wall are the mycolic acids which
78 give the organisms their characteristic waxy appearance and are interspersed with a host of species-specific free lipids. Joining these two layers is a complex
80 polysaccharide called arabinogalactan (AG), which has a chemical composition unique to the Mycobacteriales and entirely distinct from the similarly named molecule
82 found in plants that is composed of L-arabinofuranose (**Figure 1A**)[3]. AG is comprised of two domains with a β-D-galactofuranose backbone decorated by large α-D-
84 arabinofuranose branches[4]. The structure and biosynthesis of this molecule has undergone intense scrutiny, due in part to it being a target of the antimycobacterial
86 drug ethambutol[5-8]. Ethambutol targets the arabinosyltransferase proteins in the cell envelope of mycobacteria that are responsible for biosynthesis of the polysaccharide[9].
88 Similarly, the biogenesis of mycolic acids and peptidoglycan are the subject of much research due to their biochemical complexity, essentiality, and their synthesis being
90 the target of isoniazid and β-lactams respectively[10,11].

The growth of mycobacteria likely requires not only the coordinated synthesis
92 of the cell wall, but also its eventual degradation and turn-over to allow for cellular expansion, division and the insertion of cell envelope spanning secretion systems.
94 Similarly, the control of cell surface structures affords these bacteria the ability to

modulate how they are sensed by the host. As such, hydrolases of the cell wall may
96 play an important role at the mycobacterial cell surface. Hydrolases of peptidoglycan influence mycobacterial morphology and infection and it was recently demonstrated
98 that at least some peptidoglycan by-products are recycled by the bacteria[12,13]. Similarly, trehalose-mycolates, a dominant lipid component of the cell wall, undergo
100 recycling,[14] and a hydrolase involved in turning over mycolic acids (Rv3451) and a recycling pathway for trehalose has been described[15]. Cleavage of the acid-fast cell
102 wall is unlikely to be restricted to the Mycobacteriales. Organisms that predate on acid-fast bacteria such as phage or bacterial predators like the recently reported
104 *Candidatus* Mycosynbacter amalyticus (hereafter M. amalyticus), are also likely to require D-arabinan degrading enzymes to penetrate the mycobacterial cell wall[16].

106 Only a single enzyme capable of degrading AG has been characterized to date. GlfH1 (Rv3096) is an exo-β-D-galactofuranosidase from glycoside hydrolase (GH)
108 family GH5_13 that cleaves the galactan backbone of mycobacterial AG at both β-1,5 and β-1,6 linked residues[17]. The precise role of this enzyme in mycobacterial biology
110 remains unclear but its activity is suggestive of it being part of a remodeling pathway for this cell wall component. Enzymes with activity on simple D-galactofuranosides (D-
112 Gal*f*) have been identified through large-scale screening of orphan GH sequences[18]. Likewise, no enzymes have been characterized with exo- or endo-D-arabinofuranase
114 activity (enzymes that cut within the D-arabinan polymers) against AG, although this activity was described in protein extracts from soil bacteria dating back to the
116 1970s[19,20]. Endo-arabinan activity was also described in extracts of *Mycobacterium smegmatis and* was suggested to increase upon treatment of cells with ethambutol,
118 which blocks D-arabinan biosynthesis[21,22]. During our studies an exo-acting difructose-

dianhydride I synthase/hydrolase was discovered that was active on *p*NP-α-D-
120 arabinofuranoside[23]. Whether this enzyme acts on mycobacterial AG is unknown.

The human gut microbiota is responsible for the degradation of dietary plant
122 polysaccharides, host and microbial glycans. Dominated by the Bacteroidetes, this grouping of organisms is collectively amongst the richest known organisms in diversity
124 of complex carbohydrate degrading enzymes[24]. Carbohydrate utilization by the Bacteroidetes is typically mediated by genes which are organized into polysaccharide
126 utilization loci (PULs), which can be induced upon exposure of the bacterium to a given carbohydrate[25-28]. The abundance and diversity of carbohydrate degrading enzymes
128 in the Bacteroidetes provides a rich opportunity for enzyme discovery.

In this study we have mined the glycolytic capacity of the human gut microbiota
130 and discovered a collection of enzymes able to completely degrade mycobacterial arabinogalactan. We report the discovery of new families of glycoside hydrolases
132 active on the mycobacterial cell wall. We also identify new exo-D-arabinofuranosidases from the DUF2961 family (GH172). Furthermore, we
134 demonstrate that D-arabinan degradation is wide-spread amongst Mycobacteriales, and mycobacteria but is also present in phages, other bacteria, and microbial
136 eukaryotes. Our data point to a key role for these enzymes in mycobacterial biology and can enable sophisticated analysis of mycobacterial cell wall components.

138

Results

140

Select human gut Bacteroidetes can utilize mycobacterial AG as a carbon source

142 While endo-D-arabinofuranase activity was first described in soil bacteria more than 50 years ago and has been known in mycobacteria for at least 30 years[19,20], the
144 enzymes responsible for this activity have escaped identification. We reasoned that

these enzymes might be highly regulated, unstable, or poorly soluble in mycobacteria
146 making purification-based approaches unfeasible. Moreover, if they belong to novel enzyme class(es), bioinformatics approaches would fail. Instead, we isolated
148 arabinogalactan from *M. smegmatis* mc^2155 and used it as sole-carbon source for the growth of a panel of 14 Bacteroidetes species. Of these, 12 strains were able to grow
150 on this material (**Figure 1B**). Ion chromatography with pulsed amperometric detection (IC-PAD) analysis of selected culture supernatants (**Figure 1C**) demonstrated the
152 production of free galactose in most cultures, and arabinose in one; *Dysgonomonas gadei*. Together these data indicate that members of the gut microbiota produce
154 enzymes that can depolymerize mycobacterial D-arabinan and D-galactan.

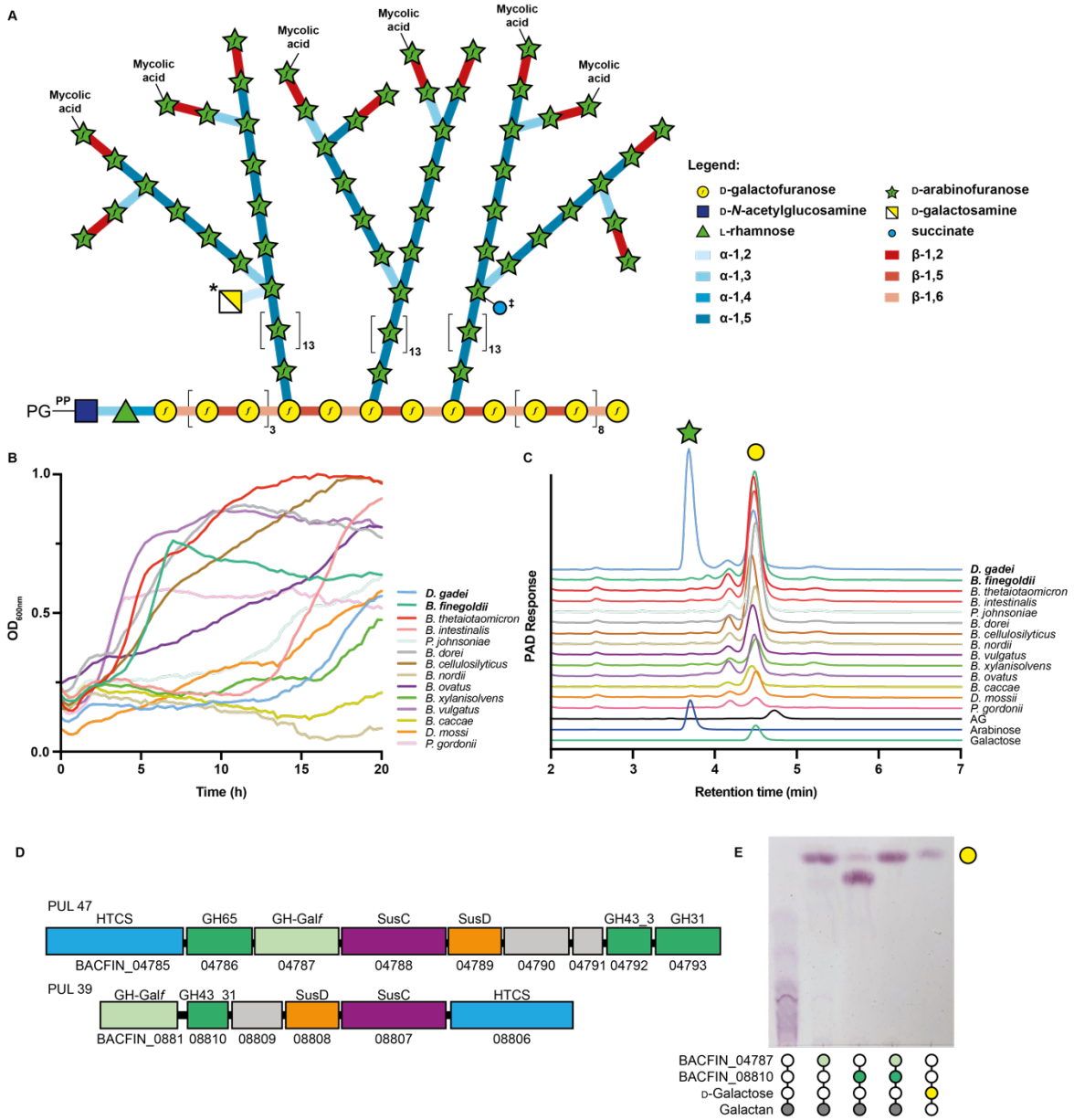


Figure 1. Bacteroidetes growth on mycobacterial arabinogalactan. A) Schematic of the structure of mycobacterial mycolyl-arabinogalactan-peptidoglycan complex. Succinate (‡) and galactosamine (*) modifications are non-stoichiometric. **B)** Growth of selected Bacteroidetes species on arabinogalactan as monitored by change in OD_{600nm}. **C)** IC-PAD analysis of culture supernatants for bacteria grown on mycobacterial AG as a sole-carbon source. Production of arabinose (green star) and galactose (yellow circle) are indicated and compared to standards. **D)** Schematics of PUL47 and 39 from *B. finegoldii*, whose homologs in *B. cellulosilyticus* were identified by RNAseq as being upregulated during growth on AG. **E)** TLC analysis of endpoint reaction products of *B. finegoldii* enzymes with *C. glutamicum* Δ*ubiA* galactan. Filled circles indicate the presence of a given reaction component.

Identification of exo- and endo-galactofuranosidases that degrade galactan

170 The presence of both galactose and arabinose in *D. gadei* culture supernatants complicated the identification of PULs specific for either galactan or arabinan. 172 Therefore, we developed a method for production of pure D-arabinan by exploiting D-galactan specific PULs. Based on the analysis of culture supernatants, *B. finegoldii* 174 and *B. cellulosilyticus* appeared to degrade galactan, but not D-arabinan. RNAseq analysis of *B. cellulosilyticus* revealed the upregulation of PUL35 and PUL36, 176 containing multiple predicted GHs (**Figure S1**). While we could not heterologously express the *B. cellulosilyticus* enzymes, the homologs of these enzymes from *B.* 178 *finegoldii* DSM17565 (PULDB ID: PUL39 and PUL47, **Figure 1D**) could be expressed and purified, and when tested on galactan (**Figure 1E**) demonstrated exo- and endo-180 activities. To determine their galactan degradative capacity, we purified galactan, comprised alternating β-1,5- and β-1,6-galactofuranose residues from a strain of 182 *Corynebacterium glutamicum* (Δ*ubiA*) that lacks D-arabinan[29]. As shown in **Figure 1E** and **Figure S1**, combination of the two *B. finegoldii* enzymes comprising a GH43_31 184 (BACFIN_08810) and a new GH-exo-Gal*f* (BACFIN_04787) family completely hydrolysed this galactan substrate.

186 *Cultivation on D-arabinan identifies D-arabinan-active PULs*

To generate D-arabinan we digested mycobacterial AG with the two *B. finegoldii* 188 galactofuranosidases (BACFIN_08810 and BACFIN_04787). This resulted in an enriched D-arabinan fraction with approximately 70% reduction in galactan as 190 determined by acid hydrolysis (**Figure S1**). The resulting D-arabinan was used as a sole carbon source for *D. gadei* (which was previously shown to produce both D-192 galactose and D-arabinose from AG). Proteomics of these bacteria at mid-log phase during growth on enriched D-arabinan identified a predicted fucose isomerase as the

194 most abundant carbohydrate-active enzyme (**Table S1**). This protein maps to PUL42 in the *D. gadei* genome **(Figure 2A**), and an additional nine of the proteins derived
196 from this PUL were in the top 200 most abundant proteins in the total proteome, including several that lacked annotation (**Figure S2**). Reasoning that mycobacteria
198 may harbor homologs of *D. gadei* arabinanases to process arabinan, we prioritized the DUF2961 and DUF4185 superfamily enzymes encoded in PUL42 as they
200 possessed homology to predicted mycobacterial proteins of unknown function within the same superfamilies.

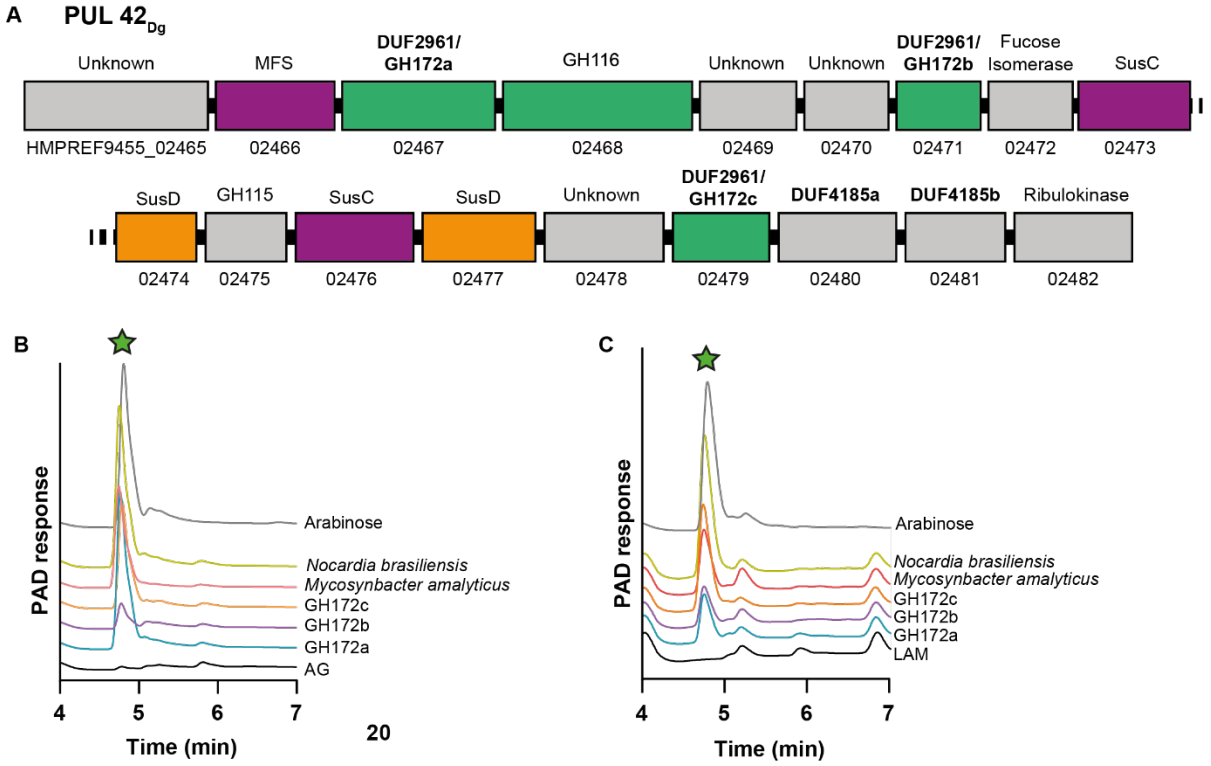
202 *The DUF2961 superfamily (GH172) includes D-arabinofuranosidases*

At the outset of this study, no member of the DUF2961 family had been
204 characterized. Therefore, we cloned, expressed, and purified all three DUF2961 family members found in PUL42 in *D. gadei*. Upon incubation with purified AG, we observed
206 D-arabinofuranosidase activity for HMPREF9455_02467, HMPREF9455_02471 and HMPREF9455_02479 (hereafter Dg_{GH172a}, Dg_{GH172b} and Dg_{GH172c}, respectively)
208 **(Figure S3)**. To probe the activity of these enzymes, we synthesized the chromogenic substrates *p*-nitrophenyl α-D-arabinofuranoside (pNP-α-D-Ara*f*) and β-D-
210 arabinofuranoside (pNP-β-D-Ara*f*). Dg_{GH172a} and Dg_{GH172c} were active against pNP-α-D-Ara*f*, and no activity was observed using pNP-β-D-Ara*f* (**Table 1**). Although substrate
212 limitations prevented accurate determination of V_{max} we could use pNP-α-D-Ara*f* to determine k_{cat}/K_M for Dg_{GH172c} (**Figure S4 and Table 2**). These data demonstrate that
214 gut bacteria can use DUF2961 enzymes to generate D-arabinose from mycobacterial arabinogalactan.

216 *DUF2961 superfamily (GH172) enzymes are present in acid-fast bacteria and their predators*

218 Homologues of the DUF2961 encoding genes are present in the genomes of organisms from the Actinomycetota phylum including pathogens such as
220 *Mycobacterium avium* subsp. *paratuberculosis* (MAP_0339c) and *Nocardia brasiliensis* (O3I_017420; Noc_{GH172}). Noc_{GH172} was soluble; however, despite repeated
222 attempts, we were unable to produce soluble MAP_0339c. Noc_{GH172} had D-arabinofuranosidase activity on both pNP-α-D-Ara*f* and purified arabinogalactan
224 (**Table 2, Figure S3**). We also identified a DUF2961 homolog within the genome of the parasitic bacterium M. amalyticus (GII36_05205; Myc_{GH172}), which possessed
226 similar activity (**Figure 2 and S3A**)[16]. Together these data indicate that exo-D-arabinofuranosidase activity is a feature of the DUF2961 super-family. These
228 enzymes are encoded by both the Mycobacteriales and their predators.

230



232

Figure 2. ***D. gadei* DUF2961 genes encode GH172 exo-D-arabinofuranosidases**
234 **with orthologs in diverse bacteria**. **A**) Schematic of PUL42 as identified by
proteomic analysis of *D. gadei* grown on D-arabinan. **B**) DUF2961 enzymes (1 μM) vs
236 2 mg ml^{-1} AG and **C**) LAM. Samples were analyzed on a Dionex ICS-6000 with
CarboPac Pa300 column, 100 mM NaOH, 20 min isocratic elution followed by a 0-
238 60% 500 mM sodium acetate gradient over 60 minutes. Green star = D-arabinose.

GH172 enzymes degrade α-1,5-D-arabinofuranoside linkages within D-arabinan

240 To understand the substrate specificity of these enzymes against other microbial D-arabinan substrates, we assayed our DUF2961 GH172 enzymes
242 (Dg_{GH172a}, Dg_{GH172b}, Dg_{GH172c}, Noc_{GH172}, and Myc_{GH172}) against mycobacterial LAM and D-araF containing pilin oligosaccharides (**Table 1**). The D-arabinan branches of
244 lipoarabinomannan (LAM) are structurally conserved with those in AG. All enzymes could digest LAM from *M. smegmatis* to produce arabinose, although Dg_{GH172b} and
246 Myc_{GH172} displayed lower activity than the others (**Figure 2C and Figure S3B**).

Pili from *Pseudomonas aeruginosa* PA7 are decorated with α-1,5-D-
248 arabinofuranose oligosaccharides. All our GH172, except for Dg_{GH172b} enzymes could cleave the D-arabinofuranose oligosaccharides from digested pili (**Figure S3C**). To
250 rule out D-galactofuranosidase activity, we repeated these experiments using purified D-galactan (**Figure S3D**) and observed no activity. The combination of these
252 experiments reveals that GH172 enzymes are specific for α-D-arabinofuranoside linkages, are active on D-arabinan, and can cleave α-1,5-D-arabinofuranoside
254 linkages.

Family GH172 proteins can adopt diverse oligomeric states

256 To better understand the structure-function relationship of this protein family, we examined the multimeric state of several of our candidates of interest. Size-
258 exclusion chromatography with laser scattering (SEC-LS) analysis of the GH172 family enzymes allowed assessment of molecular weight and assignment of
260 oligomerization states (**Table 3 and Figure S7**). A wide range of oligomeric states were uncovered: Dg_{GH172c} was assigned a hexamer, Dg_{GH172a} a dodecamer, Noc_{GH172}
262 and Myc_{GH172} as trimers, and Dg_{GH172b} as a dimer. This solution data complements

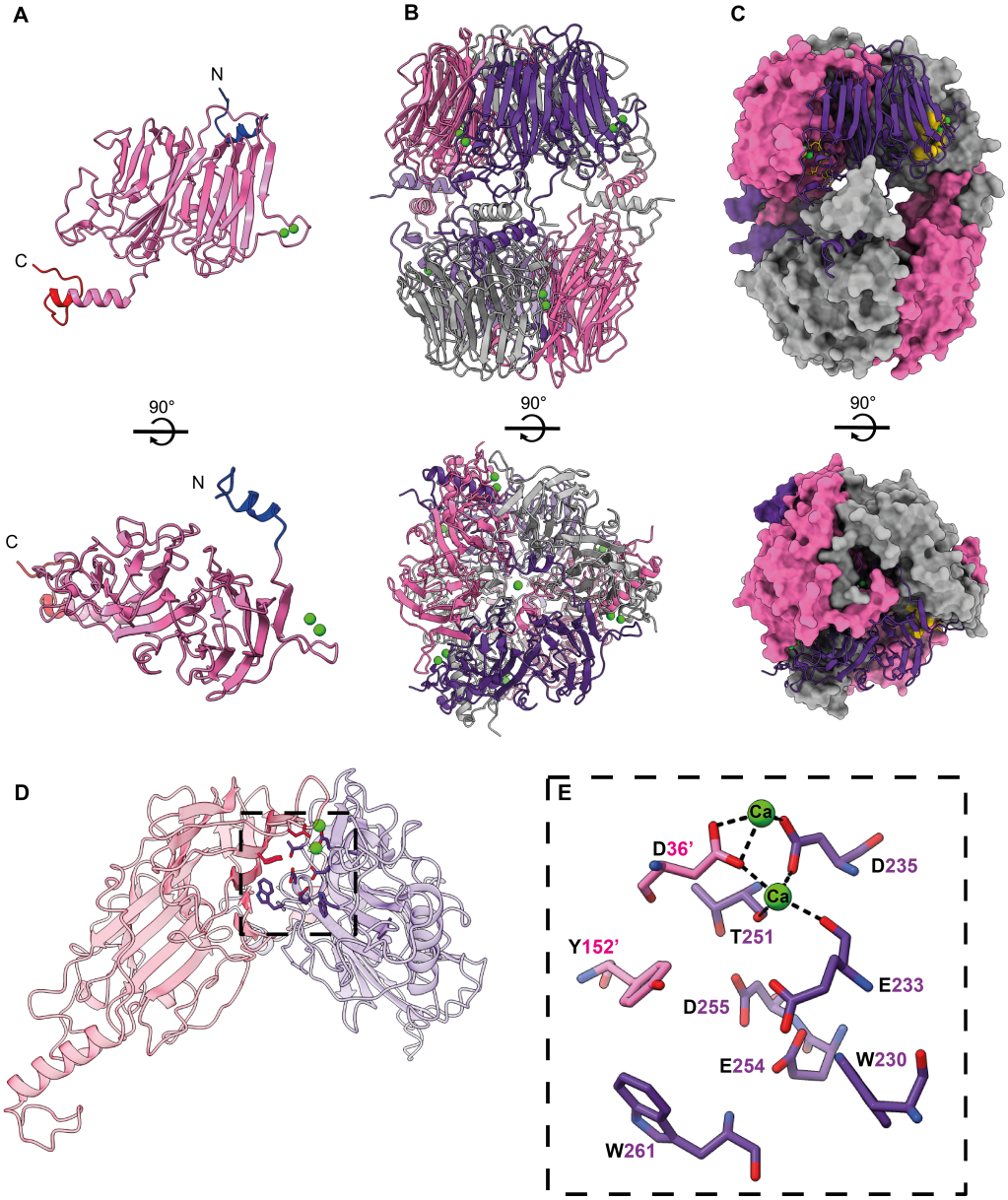
recent reports of hexameric assemblies in protein crystals of two GH172 members:
264 difructose dianhydride I synthase/hydrolase from *Bifidobacterium dentium* (PDB: 7V1V) and BACUNI_00161 from *Bacteroides uniformis* (PDB: 4KQ7) (**Figure S5**)[23].

266 Interestingly, Dg_{GH172b} is predicted to have 1.5 DUF2961 domains, and so dimerization of Dg_{GH172b} is predicted to provide three DUF2961 domains in the final
268 assembly. SEC-LS analysis is supportive of a quaternary structure for each of the GH172 proteins exhibiting multiple DUF2961 domains: Dg_{GH172b}, Myc_{GH172} and
270 Noc_{GH172} each contain three DUF2961 domains, while Dg_{GH172c} contains six and Dg_{GH172a} contains twelve. Diversity in quaternary structure may lead to differences in
272 activity or substrate specificity.

The 3D structure of Dg_{GH172c} reveals a dimer-of-trimers assembly

274 The 3D structure of Dg_{GH172c} was determined by X-ray crystallography using molecular replacement with the *B. uniformis* GH172 (Bu_{GH172}, PDB: 4KQ7; 64% sequence
276 identity with Dg_{GH172c}) as the search model. The structure of Dg_{GH172c} was refined to 1.4 Å resolution, data collection and model refinement statistics are shown in **Table**
278 **S2**. The Dg_{GH172c} monomer is comprised of two β jelly rolls and a C-terminal alpha helix (**Figure 3A**). Six monomers assemble in a dimer-of-trimers (**Figure 3B and**
280 **Figure S5A**), consistent with the solution phase oligomerization state determined by SEC-LS (**Table 3**). The dimer-of-trimers structure is formed by the C-terminal of each
282 monomer of one trimer interlocking with the C-terminal of a monomer from a second trimer. The lower side of the trimers face each other in the dimer-of-trimers structures.
284 The interfaces within the trimer form the active site with residues from both monomers contributing to the active site (**Figure 3C**). This gives a total of 3 active sites per trimer,

286 and six per hexameric assembly. Seven calcium ions are coordinated within each Dg_{GH172c} trimer: two within each active site/protomer, and one in the 3-fold axes.



288

Figure 3: 3D structure of Dg_{GH172c} determined by X-ray crystallography. A)
290 Monomer of Dg_{GH172a} with secondary structure shown as cartoon. The N- and C-termini are labelled and colored blue and red respectively. Bound calcium ions are
292 shown as green spheres. **B)** Overview of hexameric architecture of Dg_{GH172c}. Cartoon representation with bound calcium ions shown as green sphere. Trimer subunits are
294 colored grey, pink and purple for clarity. **C)** Surface representation of the hexamer with one monomer shown in cartoon. Active site residues are colored gold, and calcium
296 ions shown in green. **D)** Cartoon representation of two monomers of Dg_{GH172c} which

form the active site, with one monomer in pink and the other in purple. The active site
298 residues and coordinated calcium ions are highlighted by a black dashed box. **E)** Proposed active site residues of Dg_{GH172c}. Residues are shown as sticks with carbon
300 atoms colored pink and purple for different subunits. Calcium ions are shown as green spheres and coordinated by D36', D235 and T251.

302 The proposed active site of Dg_{GH172c} is formed of residues D36, Y152, W230, E233, D235, T251, E254, D255, and W261 (**Figure 3D**). The proposed catalytic
304 residues identified in the characterization of Bd_{GH172} are conserved in $Dg_{GH172c,}$ (Bd/Dg: E270/254, catalytic nucleophile; E291/233, acid/base) and other members of the
306 GH172 family (**Figure S6, S8**). E233 and E254 are 5.4 Å apart in Dg_{GH172c}, which is consistent with a retaining mechanism of enzyme catalysis, as shown for Bd_{GH172}[23].
308 Ca^{2+} ions are observed in the active sites of Dg_{GH172c} (contains two Ca^{2+}) and Bd_{GH172} (contains one Ca^{2+}). The calcium ions in Dg_{GH172c} are coordinated by D235, which is
310 not conserved in Bd_{GH172}. Instead, at the equivalent position, Bd_{GH172} has N272. A particular feature of characterized GH172 family members is the formation of the
312 active site by two monomers, an uncommon feature of GHs, but on that has been observed in *E. coli* LacZ β-galactosidase[30].

314 *Dg_{GH172c} catalysis is driven by conserved glutamate residues*

To gain insight into the functional roles of the residues in the active site of
316 Dg_{GH172c}, the proposed catalytic carboxylate residues were mutagenized to alanine, generating variants E233A, E254A and D225A. Dg_{GH172c}-E233A and Dg_{GH172c}-D225A
318 had no detectable activity and there was a greater than 10^5-fold reduction in activity for Dg_{GH172c}-E254A. Our data support the assignment of E254 and E233 as catalytic
320 residues (acid/base or nucleophile) and highlight an important role for the adjacent D255 residue. None of the variants displayed a change in oligomerization state
322 suggesting they do not possess structural roles (**Figure S9**).

DUF4185 is widely distributed throughout bacterial species

In Bacteroidetes, the degradation of a target polysaccharide is typically a multi-step process whereby oligosaccharides are generated and subsequently cleaved into their monosaccharide constituents, encoded within co-transcribed operons. We reasoned that the generation of D-arabinofuranose oligosaccharides was likely achieved by proteins that have no annotated function and so focused our attention on the DUF4185 proteins. To investigate the conservation of these genes across different organisms, we constructed a phylogeny of DUF4185 proteins. This revealed that they are common within actinobacteria, especially amongst the Actinomycetota (**Figure 4 and Figure S10**), in *Bacteroides* species, *Myxococcus* and in the lysis cassette of some actinobacteriophage (**Figure S11**) and the predacious bacterium M. amalyticus.

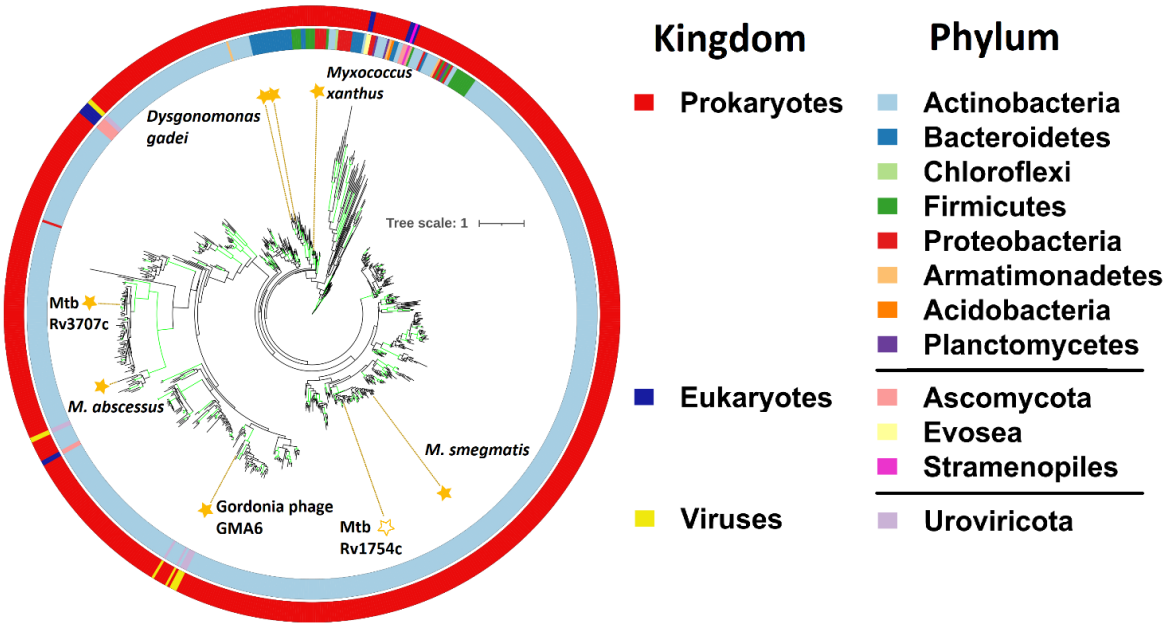


Figure 4: Phylogeny of the DUF4185 enzyme family. Unrooted ML phylogeny (LG model with empirical base frequencies, invariable sites and the discrete gamma model) of DUF4185 family sequences. Branches with greater than 75% bootstrap support (100 replicates) are highlighted in green. Units for tree scale are inferred substitutions per amino acid residue. Colored rings indicate phylum (inner) and

342 kingdom (outer) taxonomy information for sequences. Stars highlight sequences of interest and are filled for proteins that have been experimentally characterized in this
344 work.

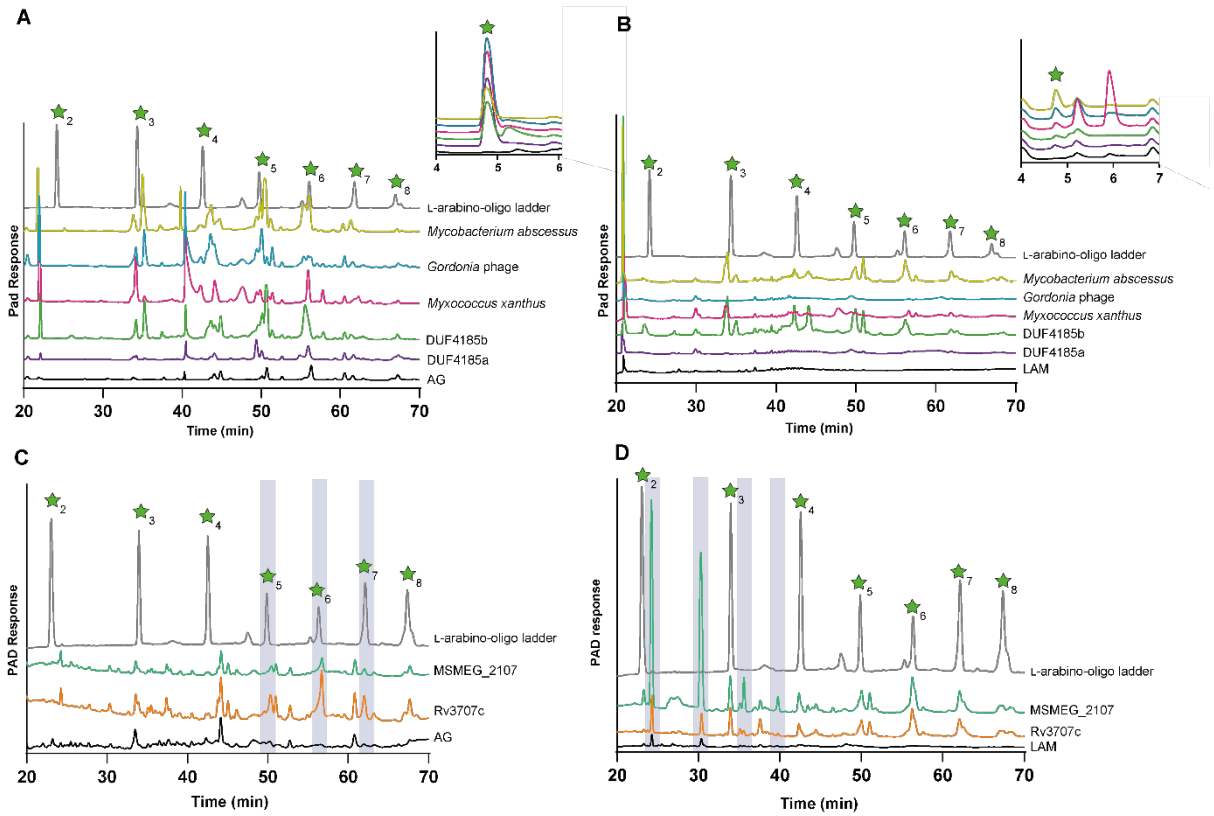
346 *DUF4185 comprises a novel GH family with endo-D-arabinanase activity*

Initially, we cloned, expressed, and purified the DUF4185 homologs
348 (HMPREF9455_02480 and HMPREF9455_02481) from *D. gadei* (herein, referred to as $Dg_{GH4185a}$ and $Dg_{GH4185b}$, respectively). When incubated with mycobacterial AG and
350 then analyzed by HPAEC, these enzymes produced a banding pattern characteristic of an endo-acting GH (**Figure S12A**), consistent with endoarabinanase activity. To
352 assess the breadth of activity for DUF4185 enzymes we selected additional candidate enzymes from each of the major lobes of the global phylogeny (**Figure 4**).
354 Recombinant proteins were produced from several bacterial lineages and a phage capable of infecting Gram-positive bacteria *Gordonia* of the order Mycobacteriales.
356 These included *Myxococcus xanthus* ($Myxo_{GH4185}$), M. amalyticus, and *Gordonia* Phage GMA6 ($Phage_{GH4185}$). Where the DUF4185 domain sat within a larger gene
358 containing several other large domains we produced truncated variants containing only the DUF4185 domain due to low solubility of the multi-domain proteins. All
360 DUF4185 constructs except that from M. amalyticus yielded soluble protein. As shown in **Figure 5** and **Figure S12**, when incubated with mycobacterial AG all these enzymes
362 possessed endo-D-arabinanase activity with varying product profiles suggesting differences in enzyme specificity. Some of the enzymes were also active against the
364 linear α-1,5-D-arabinofuranose oligosaccharides derived from *P. aeruginosa* PA7 pili, consistent with activity against the major linkage of mycobacterial D-arabinan (**Figure**
366 **S12**). The discovery of endo-D-arabinanase activity in proteins from organisms outside of the Actinomycetota is an interesting observation, however its presence in bacterial
368 predators is consistent with the essentiality of this polymer for mycobacterial viability. This sugar motif has been reported in a small number of LPS structures, providing a
370 possible explanation for the presence of this enzymatic activity in the gut

microbiota[31,32]. Furthermore, some corynebacterial species are found as commensals
372 of the human oral and gut microbiome[33].

Mycobacterial DUF4185s are endo-D-arabinofuranases

374 Given the importance of D-arabinan to mycobacterial viability and immunology we next sought to understand the biochemical function of DUF4185 proteins from
376 representative mycobacteria. As shown in **Figure 4** and **Figure S10,** mycobacteria produce at least two DUF4185s that fall into distinct phylogenetic groupings. In *M.*
378 *tuberculosis* these are Rv1754c and Rv3707c. Beyond these two conserved DUF4185 genes, some species additional DUF4185 members. For example, many
380 *Mycobacterium abscessus* strains encode at least three distinct members whilst *M. smegmatis* mc^{2}155 encodes five (MSMEG_4352, 4360, 4365, 2107 and 6255). Based
382 on sequence analysis, MSMEG_2107 and MSMEG_6255 are homologs of Rv1754c and Rv3707c, respectively, while the remainder show greater diversity (**Figure S10**).
384 A distant homolog of Rv3707c from *Mycobacterium abscessus*, Ga0069448_1118 (hereafter referred to as Mab_{4185}), was readily produced in soluble form and in good
386 yield[34]. Despite low sequence identity (33.5%) (Figure S13) to the *D. gadei* enzymes, the TLC and HPAEC profiles of mycobacterial AG digested by Mab_{4185} indicate that it
388 also possesses endo-D-arabinofuranase activity (**Figure 5**).



390 **Figure 5: DUF4185 proteins are endo-D-arabinofuranases that cleave mycobacterial AG and LAM.** DUF4185 enzymes were incubated with 2 mg·ml^{-1} AG
392 (**A** and **C**) or 2 mg· ml^{-1} LAM (**B** and **D**) for 16 hours as described in Materials and Methods. Samples were analyzed by High-performance anion-exchange
394 chromatography/pulsed amperometric detection (HPAEC-PAD) on a Dionex ICS-6000w with CarboPac Pa300 column, 100 mM NaOH 20 min isocratic elution followed
396 by a 0-60% 500 mM sodium acetate gradient over 60 minutes. A ladder of α-1,5-L-arabino-oligosaccharides (25 μM) derived from plant arabinogalactans was used as a
398 standard. In panel C the chromatogram of this ladder has been scaled on the *y*-axis by a factor of 0.2 for clarity of presentation.

400 Encouraged by this success with *Mab*$_{4185}$ but recognizing it has limited sequence identity with either of the *M. tuberculosis* proteins (17% and 16.4% identical
402 to Rv3707c and Rv1754c respectively over the entire protein length), we sought to study Rv3707c and Rv1754. However, despite our best efforts we were unable to
404 produce usable amounts of soluble Rv3707c and Rv1754. A previous report highlighted that Rv3707c is secreted, but it lacks a discernible signal peptide[35]. We
406 reasoned that the instability of Rv3707c may be due to incorrect annotation of the start site in the *M. tuberculosis* H37Rv genome, given that mycobacteria frequently use
408 TTG and GTG start codons in addition to the canonical ATG. We re-evaluated the genomic context of the protein and identified several possible in-frame N-terminal
410 extensions of the gene (**Figure S14**). We compared these potential N-terminal extensions to alignments of DUF4185 proteins that possessed a second, N-terminal
412 domain to identify sequence motifs that are conserved; in many cases a proline-rich region was observed, which is also found in the potential N-terminal extensions of
414 Rv3707c (**Figure S14**). Furthermore, by extending the N-terminus of the protein, a putative signal peptide can be predicted by SignalP 6.0 (**Figure S14**). While the
416 precise start-site is uncertain, the likely cleavage point for the signal peptide could be confidently assigned. Therefore, we designed an expression construct with the
418 putative signal peptide removed, but the remainder of the proline-rich N-terminus intact. This produced a reasonably stable and soluble Rv3707c protein in good yield
420 that digested AG to give products consistent with a hepta- or hexa-saccharide, and was also active on LAM (**Figure 5C/D**). While we were unable to produce soluble
422 Rv1754c, we successfully produced the *M. smegmatis* homolog (MSMEG_2107), which demonstrated activity against LAM, but not AG (**Figure 5C/D and Figure S12**).
424 *DUF4185 family members from different clades have distinct substrate specificities*

To further probe the substrate specificity of these enzymes and elucidate the
426 function of MSMEG_2107, we incubated each of the DUF4185 enzymes ($Dg_{GH4185a}$, $Dg_{GH4185b}$, Mab_{GH4185}, $Phage_{GH4185}$, $Myxo_{GH4185}$, MSMEG_2107, and Rv3707c) with:
428 LAM (**Figure S12B**); purified D-galactan (**Figure S12C**); and pilin oligosaccharides from *P. aeruginosa* PA7 (**Figure S12D**). The majority of the enzymes displayed higher
430 activity against AG than LAM. Conversely, MSMEG_2107 was unique in displaying a preference for LAM with no detectable activity when incubated with AG under similar
432 conditions. To confirm that the enzymes only degraded D-arabinan and not D-galactan we tested their activity with D-galactan but did not observe any product formation
434 (**Figure S12C**). Given that branched D-arabinan is the only conserved structure within both AG and LAM, we conclude the DUF4185 family are endo-D-arabinofuranases.
436 We note that only a subset of the DUF4185 enzymes were active against pili from *P. aeruginosa* PA7. As these oligosaccharides are relatively short and linear, this
438 suggests specific enzyme subsite occupancy for activity, which leads to production of arabinose for some enzymes; by contrast arabinose production was not observed
440 upon digestion of branched, polymeric AG.

We next assessed whether DUF4185 endo-D-arabinofuranases can cleave AG
442 in the context of an intact cell wall. We utilized metabolic arabinogalactan labelling with the azide-modified lipid-linked Ara*f* donor 5-AzFPA and then fluorescently labelled
444 intact bacteria with DBCO-conjugated AF647 using click chemistry (**Figure S16**)[36-38]. Both Mab_{GH4185} and Rv3707c released fluorescently labelled material from cell wall,
446 with greater activity and a greater range of products formed by the former enzyme, as observed by HPAEC-PAD. Likewise, treatment with $GH172_{Noc}$ led to release of
448 fluorescent products, supporting the conclusion that both groups of enzymes can cleave AG within the mycobacterial cell wall.

450 *Rv3707c has a highly flexible structure and belongs to the beta-propeller superfamily of glycoside hydrolases*

452 Attempts were made to crystallize all the DUF4185 homologs assayed above. *M. tuberculosis* Rv3707c yielded crystals that diffracted to a resolution of 2.4 Å. Using 454 the AlphaFold structural prediction as a search model for molecular replacement, we solved a partial structure. As shown in **Figure 6A**, Rv3707c possesses a beta-456 propeller structure, similar to several other glycoside hydrolase families,[39] with each blade consisting of 3 antiparallel beta sheets organized radially around a central pore. 458 Poor or no density was obtained for amino acid residues 1-9, 22-38, 60-69, 85-86, 285-300, and 325-337, which are located on loops surrounding the predicted active 460 site. The lack of density suggests significant structural flexibility in these regions, consistent with variations in AlphaFold predictions for both Rv3707c and Rv1754c 462 (**Figure 6, S15, S17**).

Multi-sequence alignments of DUF4185 homologs (**Figure S13**) identified 464 three nearly invariant acidic residues in the active site: Asp39, Asp56, and Glu240 (**Figure 6B).** Comparing our experimental structure to the AlphaFold prediction model, 466 the loops that contained the conserved aspartate residues in the proposed active site deviate in their positions (**Figure 6B and Figure S17**). The conserved D56 points 468 away from the active site in our experimental structure, likely in a catalytically incompetent position, while the D39 lies in a similar, but distinct position to that of the 470 AlphaFold model. Together, these observations suggest that the active site may rearrange upon binding of substrate.

472 The other regions of incomplete density in the experimentally determined 3D structure include loops that are predicted by AlphaFold to be flexible, which is reflected 474 in the B-factors for the experimental structure (**Figure S17**). These regions are also

predicted by AlphaFold to be variable in other DUF4185 proteins, including $Dg_{GH4185b}$
476 (**Figure 6C**). This apparent flexibility likely enables adaptive binding to a range of substructures within highly branched mycobacterial D-arabinan, supporting the
478 production of a broad reaction product profile, as seen in enzymatic digests (**Figure 5**).

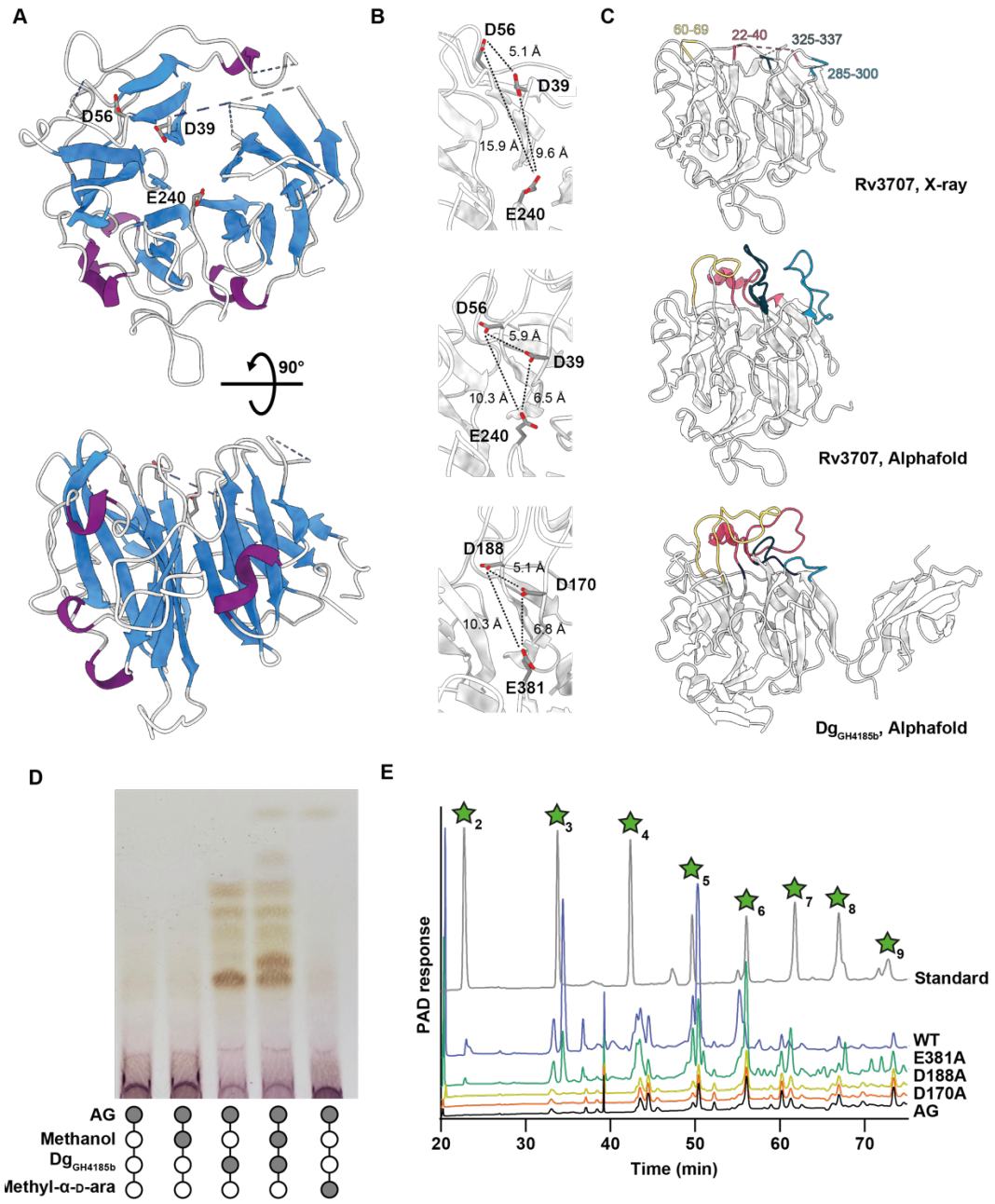


Figure 6: 3D structure, active site, and AG digestion by DUF4185 endo-D-arabinofuranases. A) 3D structure of Rv3707c determined by X-ray crystallography has a 5-bladed beta-propeller fold. Conserved acidic residues D39, D56 and E240 are indicated in the proposed active site. **B**) Active site details for Rv3707c (X-ray), Rv3707c (AlphaFold) and $Dg_{GH4185b}$ (AlphaFold) with distances between residues indicated in angstroms. **C**) Overall fold of the same enzymes as in **B** with variable surface features highlighted relative to the Rv3707c X-ray structure. Corresponding regions are in the same color in each protein. **D)** Methanolysis of AG catalyzed by $Dg_{GH4185b}$. The production of methyl glucosides in the presence of methanol indicates a retaining mechanism. **E)** $Dg_{GH4185b}$-WT, $Dg_{GH4185b}$-D149A and $Dg_{GH4185b}$-D167A were incubated at 1 μM with 2 mg ml^{-1} AG for 16 hours. No activity was observed for the D188A or D170A mutants.

DUF4185 family are anomer-retaining enzymes

494 Glycoside hydrolases can hydrolyze the anomeric linkage through either reversion or retention. Inclusion of a simple alcohol, such as methanol, in an enzymatic
496 digest can be used to identify retaining enzymes, as they may afford methylated glycosides[40]. Addition of methanol to AG digests by $Dg_{GH4185b}$ afforded methyl
498 arabinoside, thereby demonstrating a retaining mechanism (**Figure 6D**). Three conserved carboxylate residues are present in the active site (**Figure 6B**). Using site-
500 directed mutagenesis, we varied these carboxylate residues in $Dg_{GH4185b}$ to generate $Dg_{GH4185b}$-D170A and $Dg_{GH4185b}$-D188A and $Dg_{GH4185b}$-E381A derivatives. Activity was
502 broadly retained for glutamate substitution, but no activity was observed for the two aspartate mutants (**Figure 6E**). These data support a retaining mechanism for the
504 DUF4185 family of enzymes and assignment of D170/D39 in $Dg_{GH4185b}$ (and D188/D56 in Rv3707c) as the catalytic residues corresponding to acid/base and nucleophile.

506

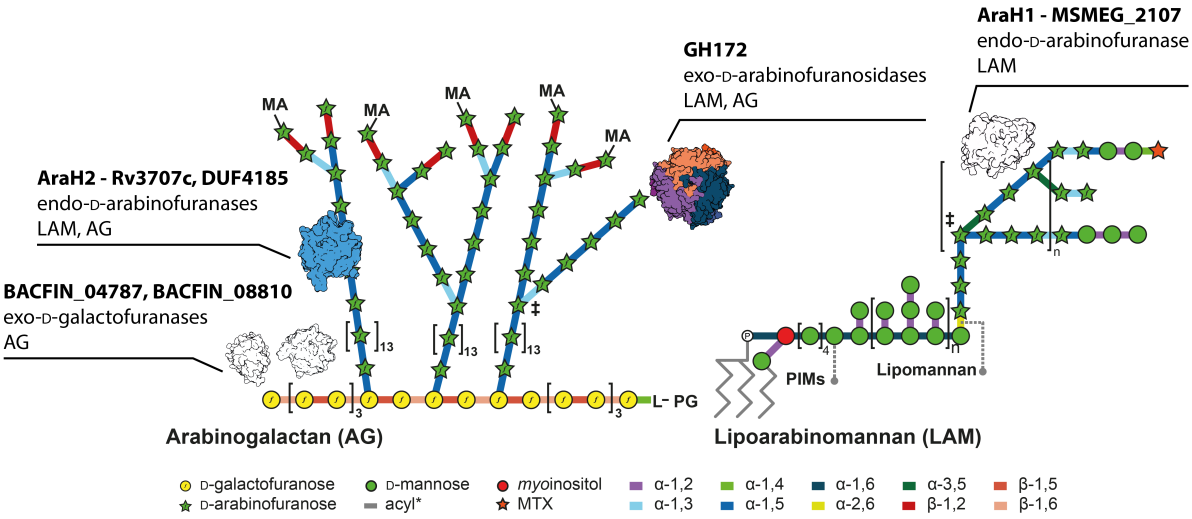


Figure 7. Mycobacterial arabinogalactan-degrading enzymes discovered in this study and their substrates. Each enzyme or enzyme family is listed with its identified function and the mycobacterial cell wall it acts upon. Colored in structures are those for which we have presented experimental structural data, black and white are surface representations of AlphaFold models. MA – mycolic acids; L – linker unit; PG – peptidoglycan; PIMs – phosphatidylinostol mannosides.

Discussion

Endo-D-arabinanase activity was first reported more than 50 years ago, but despite the widespread availability of mycobacterial genomic tools and -omics technologies, these enzymes have escaped identification[22,41]. We have mined the human gut microbiome and leveraged evolutionary conservation to identify these enzymes in mycobacteria, along with new exo-D-arabinofuranosidases and exo-D-galactofuranosidases (**Figure 7**). We reasoned that the abundance of Mycobacteriales in environmental niches in addition to the availability of D-arabinofuranose polymers in organisms such as *P. aeruginosa* PA7 and corynebacteria meant that the capacity to degrade this carbohydrate was likely to be encoded in the human gut microbiota[42,43], allowing identification by metabolic methods with arabinogalactan as sole carbon source.

528 Initially, we identified D-galactofuranase activity in a gut microbe organism, whose presence may reflect the widespread occurrence of D-Gal*f* in the LPS of some
530 gram-negative bacteria of the human gut microflora[44]. One of the enzymes we identified, BACFIN_04787, is the founding member of a new glycoside hydrolase
532 family. The galactan degradation pattern exhibited by BACFIN_04787 is consistent with an exo-acting, non-specific enzyme and the product pattern of BACFIN_08810 is
534 consistent with an exo-acting enzyme that can cleave either the β-1,5 or β-1,6 linkage within D-galactan. Further characterization of these enzymes will enable detailed
536 analysis of mycobacterial galactan, whose chain length was recently shown to be important in the biology of these organisms[4].

538 Identifying organisms that can grow on D-arabinan is complicated by the presence of D-galactan in AG. To overcome this, we used the newly discovered D-
540 galactofuranase to generate enriched D-arabinan, which we used to identify the Bacteroides *D. gadei*, as a potent D-arabinan degrader. The PULs associated with this
542 ability contained uncharacterized DUFs as well as DU2961 genes corresponding to family GH172. The GH172 proteins were shown to encode exo-D-arabinosidases.
544 During the preparation of this manuscript, members of family GH172 were independently shown to exhibit exo-D-arabinofuranosidase activity by Kashima *et al.*,
546 consistent with our data[23]. The GH172 enzyme from *D. gadei* shares a similar structure to the enzyme described by Kashima *et al*; however, it lacks the large C-terminal
548 extensions observed in Bd_{GH172} which were postulated by the authors to hold the hexameric protein together. SEC-LS analysis of a range of family GH172 proteins
550 reveals diverse quaternary structures of multimers up to dodecamers. How this structural diversity is connected to function is as-of-yet unexplored. It has been

552 suggested that family GH172 proteins and phage capsid proteins may be ancestrally related, raising the possibility for diversification of function within this family[45].

554 The presence of genes encoding DUF4185 and GH172 enzymes in M. amalyticus suggests that they contribute to its epibiotic lifestyle on host *Gordonia* spp. through 556 feeding on arabinogalactan as a carbon source. The $Myc_{DUF4185}$ enzyme is predicted to be a surface located lipoprotein, while Myc_{GH172} is predicted to not be secreted. This 558 localization is consistent with a model whereby M. amalyticus releases oligosaccharides from the surface of *Gordonia*, internalizes them, and then cleaves 560 them to monosaccharides in the cytoplasm. An alternative explanation is that the DUF4185 and GH172 enzymes locally remodel the cell wall of *Gordonia*, enabling 562 access to cellular contents. The biological role of GH172 enzymes in *Bacteroides* remains unclear. While some may be involved in α-fructan degradation, those 564 associated with D-arabinan PULs are more likely targeted at either glycans derived from Actinomycetota or organisms such as *P. aeruginosa* PA7.

566 DUF4185 family of enzymes include the first known endo-D-arabinofuranases, though it is not the first time this activity has been reported. In 1972 Kotani and 568 colleagues reported the isolation of “mixed D-arabinanase activity” in an extract from an unnamed Gram-positive soil microbe, referred to as the “M-2” fraction[19]. This 570 extract possessed endo-activity and released a wide range of products from mycobacterial cell walls, although the specific protein responsible was not identified. 572 Since then, similar impure enzyme cocktails have contributed to numerous studies on AG and LAM[41]. We have now identified broadly conserved endo-D-arabinofuranases 574 from the Mycobacteriales, which we propose to re-name to AraH1 (Rv1754c) and AraH2 (Rv3707c). The availability of well-characterized enzymes with defined 576 activities should support more detailed studies of mycobacterial cell wall polymers.

A review of functional screens of knockout libraries of *M. tuberculosis* highlights
578 an important role for AraH2 in pathogenesis. AraH2 was identified in a screen for proteins with non-canonical signal sequences[35], likely because of mis-annotation of
580 its signal peptide as demonstrated by our work. In that study, Perkowski and colleagues reported that a transposon mutant in AraH2 was severely defective for
582 replication in macrophages. A separate study identified AraH2 as important for control of phagosome acidification, where it was the second most enriched mutant in an
584 acidified phagosome screen[46]. A homologue of AraH2, PEG_1752, was shown to be upregulated in the phylogenetically related *Mycobacterium llatzerense* upon infection
586 of the amoeba *Acanthamoeba castellanii*[47]. These data point to a role for AraH2 in phagosome survival, and by extension of D-arabinan remodeling in mycobacterial
588 pathogenesis.

The role of AraH1 remains more elusive. The gene is broadly conserved
590 amongst mycobacteria, and our biochemical data suggests this enzyme class is active against LAM, but not AG. This genomic locus is a frequent site of IS6110 element
592 insertion[48], and interruption of this gene could cause LAM structural variability amongst circulating strains of *M. tuberculosis*. It is possible that AraH1 and AraH2
594 have partially overlapping substrate specificity and the former can partially compensate for the loss of the latter. As well, these proteins may have specific roles
596 under narrowly defined conditions. Analogously, peptidoglycan-lytic enzymes with seemingly redundant reaction specificities are encoded in most bacteria, and lack
598 notable phenotypes for their loss under most growth conditions. However, screens at a range of pH values have identified specific functions for these enzymes[49].

600 To conclude, we have unearthed a new enzymatic toolkit for the degradation of mycobacterial arabinogalactan that should find utility in the study of the structure and

602 function of this important polymer (**Figure 7**). The functional annotation of these genes will support future investigations of the role of D-arabinan structural modulation in
604 mycobacterial biology, mycobacteriophage infection and interbacterial predation.

Tables

606 **Table 1: Activity of GH172 enzymes against synthetic and bacterial substrates**

Protein	pNP-α-D-Ara*f*	AG	LAM	Pili
Dg_{GH172a}	✓	✓	✓	✓
Dg_{GH172b}	weak	✓	✓	weak
Dg_{GH172c}	✓	✓	✓	✓
Noc_{GH172}	✓	✓	✓	✓
Myc_{GH172}	✓	✓	✓	weak

608 **Table 2: k_{cat}/K_M of GH172 enzymes for pNP-α-D-Ara*f* and AG**

Protein	pNP-D-Ara*f*
Dg_{GH172a}	n/a
Dg_{GH172c}	$2.9 \times 10^4 \pm 4 \times 10^2$ min^{-1} M^{-1}
Noc_{GH172}	$1.2 \times 10^5 \pm 2 \times 10^3$ min^{-1} M^{-1}

Table 3: SEC-LS oligomerization state of GH172 enzymes

Protein	Observed MW (Da)	Oligomerization state	Theoretical MW of oligomer (Da)
Dg_{GH172c}	260,829	hexamer	266,952
Noc_{GH172}	123,934	trimer	120,507
Dg_{GH172a}	515,568	dodecamer	519,438
Dg_{GH172b}	148,350	dimer	146,512
Myc_{GH172}	192,623	trimer	202,115

610

612

614

616

Methods

618 **Bacterial strains and growth conditions**

Bacteroidetes sp. were grown on a 2x defined minimal media (**Table S3**) under 620 anaerobic conditions at 37 °C over 24 hours to assay growth on various carbon sources, including arabinogalactan. Strains used were *Bacteroides caccae* ATCC 622 43185, *B. cellulosilyticus* DSM 14838, *B. dorei* DSM 17855, *B. finegoldii* DSM 17565, *B. intestinalis* DSM 17393, *B. nordii* CL02T12C05, *B. ovatus* ATCC 8483, *B.* 624 *thetaiotaomicron* VPI-5482, *B. vulgatus* ATCC 8482, *B. xylanisolvens* XB1A, *Dysgonomonas gadei* ATCC BAA-286, *D. mossii* DSM 22836, *Parabacteroides* 626 *gordonii* DSM 23371, *P. johnsonii* DSM 18315. *Escherichia coli* and *Pseudomonas aeruginosa* PA7 ATCC 15692 strains were grown in lysogeny broth at 37 °C 628 (unless otherwise specified). *Mycobacterium smegmatis* mc^2155 ATCC 19420 was grown in Tryptic soy Broth at 37 °C with agitation.

630

632

RNA sequencing

634 *B. cellulosilyticus* was cultured in defined media (supplementary table 3) containing 5 mg ml^{-1} AG or glucose, in triplicate 5 ml cultures. Cells were harvested at 636 mid-log phase and stored in RNA protect (Qiagen). RNA was purified with the RNAeasy Kit. Prior to library preparation, rRNA was depleted using the Pan-638 Prokaryote riboPOOLs kit (siTOOLs Biotech). In brief, 1 µg of total RNA was incubated for 10 min at 68 °C and 30 min at 37 °C with 100 pmol of rRNA-specific biotinylated 640 DNA probes in 2.5 mM Tris-HCl pH 7.5, 0.25 mM EDTA, and 500 mM NaCl. DNA-rRNA hybrids were depleted from total RNA by two consecutive 15 min incubations

642 with 0.45 mg streptavidin-coated magnetic Dynabeads MyOne C1 (ThermoFisher Scientific) in 2.5 mM Tris-HCl pH 7.5, 0.25 mM EDTA, and 1 M NaCl at 37 °C. The
644 rRNA-depleted RNA samples were purified using the Zymo RNA Clean & Concentrator kit combined with DNase treatment on a solid support (Zymo Research).
646 cDNA libraries were prepared using the NEBNext Multiplex Small RNA Library Prep kit for Illumina (NEB) in accordance with the manufacturers' instructions.

648 Library preparation and sequencing took place at the Earlham Institute, and were processed by Newcastle University Bioinformatics Support Unit. Briefly, raw
650 sequencing reads were checked using Fast QC, reads were mapped to *Bacteroides cellulosilyticus* DSM 14838 (GCA_000158035) downloaded from Ensembl (assembly
652 ID ASM15803v1). Reads were quantified against genes contained in the Ensembl annotation using featureCounts from the Rsubread package[50]. Counts were
654 normalized by Trimmed Median of Means (TMM) as implemented in DESeq2, and differentially expressed genes determined according to a Negative Binomial model as
656 per DESeq2.

Proteomic analysis of *D. gadei*

658 *Sample preparation*

Dysgonomonas gadei cells were suspended in 5% sodium dodecyl sulfate
660 (SDS) in 50 mM triethylammonium bicarbonate (TEAB) pH 7.5. The samples were subsequently sonicated using an ultrasonic homogenizer (Hielscher) for 1 minute. The
662 whole-cell lysate was centrifuged at 10,000 x *g* for 5 min to remove cellular debris. Protein concentration was determined using a bicinchoninic acid (BCA) protein assay
664 (Thermo Scientific). A total of 20 µg protein was used for further processing. Proteins were reduced by incubation with 20 mM tris(2-carboxyethyl)phosphine for 15 min at
666 47 °C, and subsequently alkylated with 20 mM iodoacetamide for 30 minutes at room

temperature in the dark. Proteomic sample preparation was performed using the
668 suspension trapping (S-Trap) sample preparation method [51], as recommended by the supplier (ProtiFi™, Huntington NY). Briefly, 2.5 µl of 12% phosphoric acid was added
670 to each sample, followed by the addition of 165 µl S-Trap binding buffer (90% methanol in 100 mM TEAB pH 7.1). The acidified samples were added, separately, to
672 S-Trap micro-spin columns and centrifuged at 4,000 x *g* for 1 min until the solution has passed through the filter. Each S-Trap micro-spin column was washed with 150 µl S-
674 trap binding buffer by centrifugation at 4,000 x *g* for 1 min. This process was repeated for a total of five washes. Twenty-five µl of 50 mM TEAB containing trypsin (1:10 ratio
676 of trypsin:protein) was added to each sample, followed by proteolytic digestion for 2 h at 47 °C using a thermomixer (Eppendorf). Peptides were eluted with 50 mM TEAB
678 pH 8.0 and centrifugation at 1,000 x *g* for 1 min. Elution steps were repeated using 0.2% formic acid and 0.2% formic acid in 50% acetonitrile, respectively. The three
680 eluates from each sample were combined and dried using a speed-vac before storage at -80 °C.

682 *Mass Spectrometry*

Peptides were dissolved in 2% acetonitrile containing 0.1% trifluoroacetic acid,
684 and each sample was independently analyzed on an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Fisher Scientific), connected to an UltiMate 3000
686 RSLCnano System (Thermo Fisher Scientific). Peptides were injected on a PepMap 100 C_{18} LC trap column (300 µm ID × 5 mm, 5 µm, 100 Å) followed by separation on
688 an EASY-Spray nanoLC C_{18} column (75 µm ID × 50 cm, 2 µm, 100 Å) at a flow rate of 250 nl/min. Solvent A was water containing 0.1% formic acid, and solvent B was 80%
690 acetonitrile containing 0.1% formic acid. The gradient used for analysis was as follows: solvent B was maintained at 2% for 5 min, followed by an increase from 2 to 35% B in

692 120 min, 35-90% B in 0.5 min, maintained at 90% B for 4 min, followed by a decrease
to 3% in 0.5 min and equilibration at 2% for 10 min. The Orbitrap Fusion Lumos was
694 operated in positive-ion data-dependent mode. The precursor ion scan (full scan) was
performed in the Orbitrap in the range of 400-1,600 *m/z* with a resolution of 120,000
696 at 200 *m/z*, an automatic gain control (AGC) target of 4×10^5 and an ion injection time
of 50 ms. MS/MS spectra were acquired in the linear ion trap (IT) using Rapid scan
698 mode after high-energy collisional dissociation (HCD) fragmentation. An HCD collision
energy of 30% was used, the AGC target was set to 1×10^4 and dynamic injection time
700 mode was allowed. The number of MS/MS events between full scans was determined
on-the-fly to maintain a 3 s fixed duty cycle. Dynamic exclusion of ions within
702 a ± 10 ppm *m/z* window was implemented using a 35 s exclusion duration. An
electrospray voltage of 2.0 kV and capillary temperature of 275 °C, with no sheath and
704 auxiliary gas flow, was used.

All mass spectra were analyzed using MaxQuant 1.6.12.0[52], and searched
706 against the *Dysgonomonas gadei* ATCC BAA-286 proteome database downloaded
from Uniprot (accessed 09.01.2020). Peak list generation was performed within
708 MaxQuant and searches performed using default parameters and the built-in
Andromeda search engine[53]. The enzyme specificity was set to consider fully tryptic
710 peptides, and two missed cleavages were allowed. Oxidation of methionine, N-
terminal acetylation and deamidation of asparagine and glutamine were allowed as
712 variable modifications. Carbamidomethylation of cysteine was allowed as a fixed
modification. A protein and peptide false discovery rate (FDR) of less than 1% was
714 employed in MaxQuant. Proteins that contained similar peptides and could not be
differentiated based on MS/MS analysis alone were grouped to satisfy the principles

716 of parsimony. Reverse hits, contaminants, and proteins only identified by site modifications were removed before downstream analysis.

718 **Generation of expression constructs**

Unless stated otherwise genes were purchased as codon-optimized constructs 720 from Twist Biosciences. *D. gadei* and *B. finegoldii* genes were cloned from genomic DNA using standard restriction cloning methods, followed by ligation into pET28a 722 vectors and transformation into TOP10 cells (Novagen) with subsequent sequencing of selected purified recombinant plasmids by Eurofins for confirmation.

724 **Protein Expression and Purification**

726 *Expression and purification of* $Dg_{GH4185a}$, $Dg_{GH4185b}$, $Phage_{GH4185}$, $Myxo_{GH4185}$, Dg_{GH172a}, Dg_{GH172b}, Dg_{GH172c}, *and* Myc_{GH172}

728

Recombinant proteins were expressed in competent *E. coli* Tuner cells 730 (Novagen) using pET28a vectors generated as above. Cells were grown in LB media at 37 °C with shaking, until turbidity reached an OD_{600} of ~0.6, wherein expression was 732 induced with 0.2 mM IPTG and cells were further cultured for 16 hours at 16 °C. Sonication was used to lyse cells in 20 mM Tris, pH 8.0, 200 mM NaCl.

734 Enzymes were purified using immobilized metal affinity chromatography on cobalt TALON resin. Proteins were dialyzed into 20 mM HEPES, pH 8.0, 150 mM NaCl 736 buffer dialysis (Medicell). For crystallography proteins were purified further via size-exclusion chromatography (HiLoad 16/600 Superdex 200, GE Healthcare) in 20 mM 738 HEPES, pH 8.0, 150 mM NaCl. Protein purity was ascertained by SDS-PAGE and protein concentrations were determined using Nanodrop spectroscopy 740 (Thermofisher).

742 *Purification of Rv3707c and MSMEG_2107*

For protein Rv3707c and MSMEG_2107 expression, an aliquot of competent
744 BL21 DE3 *Escherichia coli* was transformed with plasmid and plated on LB agar supplemented with 50 µg/mL kanamycin. One plate of bacteria was scraped to
746 inoculate 1 L of modified Studier's autoinduction media[54]. The bacteria were incubated at 37 °C with shaking (180 RPM) until OD_{600}= ~0.6, whereupon the flasks were cooled
748 on ice with agitation to 20 °C and then returned shake overnight at 20 °C. After induction, the cultures were pelleted at 3990 *x g* for 25 min at 4 °C. Pellets were
750 resuspended in sterile PBS and pelleted at 7000 *x g* for 10 min. The supernatant was removed, and pellets were snap frozen in liquid nitrogen and stored at -20 °C until
752 preparation.

754 Rv3707c was purified by suspension of one pellet in cold lysis buffer (25 mM HEPES pH8; 400 mM NaCl; 5% glycerol; 50 mM L-arginine; 50 mM L-glutamic acid; 1
756 mM beta-mercaptoethanol). 1 mg ml^{-1} deoxyribonuclease I from bovine pancreas (Sigma-Aldrich) was added to cell slurry and incubated for 30 minutes. Cells were
758 lysed by three passages through a French pressure cell. Insoluble debris was then pelleted by centrifugation at 40,000 *x g* for 40 minutes (4 ºC). The supernatant was
760 then processed by immobilized metal affinity chromatography (IMAC) on a gravity column containing 2 mL bed volume of cOmplete His Tag purification resin (Roche).
762 After loading the lysate, the column was washed with 80 mL of lysis buffer, then eluted with an imidazole gradient of 50, 100, 250, and 500 mM. Protein-containing fractions
764 were pooled and dialyzed exhaustively against three liters of dialysis buffer (25 mM HEPES pH8; 400 mM NaCl; 5% glycerol; 50 mM L-arginine; 50 mM L-glutamic acid; 2
766 mM dithiothreitol). The crude protein was concentrated to a final volume of 0.5 mL on a 30k Da molecular weight cutoff Pierce protein concentrator (Thermo Scientific). This
768 fraction was then further purified by size exclusion chromatography on an AKTA Prime

system with a SuperDex 26/600 S200 column in the above dialysis buffer before being
770 concentrated. The protein was always used freshly prepared.

772 *Purification of Noc_{GH172} and Mab_{GH4185}*

One plate of BL21-DE3 transformed with an appropriate plasmid was used to
774 inoculate 1L Terrific Broth supplemented with kanamycin. The culture was grown to an OD_{600} of 0.6 and induced with 0.25 mM IPTG and grown overnight at 20 ºC. After
776 harvest of biomass as in purification of Rv3707c, cell pellets were resuspended in a buffer containing 25 mM HEPES; 40 mM NaCl, lysozyme, and DNAse I. Subsequent
778 purification steps were identical to those in Rv3707c, but in the above, simpler buffer, omitting lysozyme and DNAse I.

780

Crystallography

782 *Crystallography of Dg_{GH172c}*

Crystallization of Dg_{GH172c} sample at 10 mg ml^{-1} was screened using
784 commercial kits (Molecular Dimensions and Hampton Research) with vapor diffusion sitting drop method. Crystals formed in a buffer containing 0.1 M MOPS/Sodium
786 HEPES pH 7.5, 0.12 M alcohol and 30% EDO_P8K over a period of 2 weeks. Crystals were harvested and flash cooled in liquid nitrogen. X-ray diffraction data were collected
788 at the synchrotron beamline I24t Diamond light source (Didcot, UK) at a temperature of 100 K. The data were integrated using XDS[55] via XIA2 [56] and scaled with Aimless[57].
790 The space group was confirmed with Pointless. The phase problem was solved by molecular replacement using Phaser and PDB model 4KQ7 as search model from
792 *Bacteroides uniformis*[58]. Subsequently, the structure was auto built with CCP4build on CCP4cloud[59]. The model was improved by rounds of manual building with COOT
794 and refinement with Refmac[60,61]. The final model was validated with Molprobity[62] and COOT[60].

796

Crystallography of Rv3707c

798 Rv3707c was concentrated to 7 mg ml^{-1} and loaded into a sitting drop crystallization tray (Clover) using JCSG-plus crystallization screen (Molecular 800 Dimensions) at a ratio of 2 µL protein 2 µL reservoir solution. Crystals were formed in a buffer containing 0.8 M succinic acid pH 7 over a period of 1 month. The crystals 802 were cryoprotected in a solution of 0.8 M succinic acid pH 7 with 30% glycerol and flash cooled in liquid nitrogen. X-ray diffraction data were collected at beamline I04 of 804 Diamond light source, Oxford. Data were auto-processed using Xia2 and general file manipulations were performed using the CCP4 suite of programmes[56,59]. Data were 806 phased by molecular replacement using the Alphafold structural prediction of Rv3707c as a search model using PHASER (TFZ of 21.1)[58]. The structure was then auto built 808 in PHENIX with rounds of refinement carried out by PHENIX-refine and manually in COOT[60,63].

810 **Summary of methods for phylogenetic analyses**

Two alignments were used to infer respectively a global phylogeny for members 812 of the PF13810 family and a more restricted phylogeny focusing on close relatives to the functionally characterized proteins. The global alignment was derived from the 814 "Full" Pfam alignment for PF13810 made of 1145 sequences and 1321 aligned sites (http://pfam.xfam.org/family/PF13810#tabview=tab3). This led to an alignment of 753 816 sequences and 179 aligned sites by: (i) deleting partial sequences that did not include the catalytic residues or that corresponded to obsolete sequences (nine sequences) 818 and (ii) adding the three sequences from *Mycobacterium abscessus* (strain 4529 available at the Integrated Microbial Genomes & Microbiomes (IMG/M) database: 820 2635695100/Ga0069448_11324, 2635694794/Ga0069448_1118, 2635698039/Ga0069448_113269) and (iii) deleting sites made of a majority of indels.

822 For the restricted alignment 39 complete sequences were aligned, including the seven proteins investigated in this study - enzymatic characterization and one structure. The
824 sequences were aligned with Clustal Omega using default settings in SEAVIEW v.4.6.4 [64,65]. Following minor manual adjustments of the alignment the mask function
826 of SEAVIEW was used to selected aligned residues that included conserved blocks of sequences with no more than two indels leading to 200 aligned sites. The DUF4185
828 alignment is available as Supplementary Dataset 1 for respectively the global and restricted alignment. IQ-TREE (v.1.6.12) was used to generate maximum likelihood
830 phylogenies using automatic model selection[66]. The selected models were LG+F+I+G4 for the global alignment and WAG+I+G4 for the restricted alignment using
832 the Bayesian Information Criterion (BIC). Bootstraps (100 replicates) were computed to assess branch reliability. iTOL (interactive tree of life) was used to generate the
834 figures[67]. The global phylogeny was annotated with taxonomy information derived from the UniProt database (https://www.uniprot.org/)[68].

836 **Protein structure prediction by AlphaFold 2.0**

Amino acid sequences were submitted to the Colabfold_advanced.ipynb server
838 for protein structure prediction[69]. Unless otherwise stated the default values were used, with the max_recycles set to 12. Figure coloring for pLDDT values were
840 generated in ChimeraX 1.4 using the protein structure prediction tool and a custom palette.

842

Timepoint assays

844 To assess enzymatic activity, reactions were initiated containing substrates (in water) and enzymes (in 20 mM HEPES pH 7.5, 150 mM NaCl, unless otherwise
846 stated) of various concentrations, with 50 mM potassium phosphate buffer pH 7.2 as a dominant reaction buffer. Reactions were incubated at 37 °C for 16 hours and

848 subsequently boiled to ensure enzymatic cessation. Time point samples were then analyzed using TLC or IC-PAD.

850

Porous graphitic carbon chromatography clean-up of Rv3707c and

852 **MSMEG_2107 arabinogalactan hydrolysis assays**

Due to the high concentration of L-arginine and L-glutamic acid in the buffer

854 used for purification of Rv3707c and MSMEG_2107, enzyme assays were unsuitable for HPAEC-PAD analysis without prior solid phase extraction. To this end, at each

856 timepoint, reaction mixtures were loaded onto a Hypersep Hypercarb SPE cartridge (Thermo Scientific) which had been washed with acetonitrile and 50% THF in water

858 and exhaustively equilibrated with water prior to loading. Reaction products were then eluted in 80% acetonitrile in ddH2O with 0.1% trifluoracetic acid (Sigma-Aldrich) and

860 dried by evaporation in a SpeedVac concentrator before being reconstituted in the original volume of water.

862 **Kinetic analysis of GH172 arabinofuranosidase activity**

864 Enzymes (100 nM) were incubated with the indicated concentrations of pNP-α-D-Ara*f* or AG. Assays were performed in technical triplicate at 37 °C in 20 mM HEPES

866 pH 7.5. for pNP, absorbances were measured at 400 nm and graphs were plotted in GraphPad Prism 9.3.1. For Ag, arabinose concentration was measured by IC-PAD

868 (see below) with reference to a standard curve.

870 **Thin-layer chromatography (TLC)**

Purified proteins were incubated with at 0.1-5 μM (as indicated) with substrates

872 (for methanolysis, methanol was added to the reaction mixture at a final concentration of 10%) for 16 h at 37 °C to ensure reaction completion (unless otherwise indicated).

874 Using TLC plate aluminum foils (Silicagel 60, 20 x 20, Merck) which were cut to the

desired size (minimum height of 10 cm), these reaction samples were spotted (6 µl,
876 unless otherwise indicated) and allowed to dry. TLC plates were run (twice) in solvent (1-butanol/acetic acid/ water 2:1:1 (v/v)). Plates were then removed and dried before
878 visualization of sugars was obtained via immersion of TLC plate in Orcinol stain. Plates were dried and developed through heating between 50 °C and 80 °C.

880 **Ion Chromatography with Pulsed Amperometric Detection (IC-PAD)**

Oligosaccharides from enzymatic polysaccharide digestion were analyzed
882 using a CARBOPAC PA-300 anion exchange column (ThermoFisher) on an ICS-6000 system. Detection enabled by PAD using a gold working electrode and a PdH
884 reference electrode with standard Carbo Quad waveform. Buffer A – 100 mM NaOH, Buffer B – 100 mM NaOH, 0.5 M Na Acetate. Each sample was run at a constant flow
886 of 0.25 $ml \cdot min^{-1}$ for 100 minutes using the following program after injection: 0-10 min; isocratic 100% buffer A, 10-70 min; linear gradient to 60% buffer B, 70-80 min; 100%
888 buffer B. The column was then washed with 10 mins of 500 mM NaOH, then 10 min re-equilibration in 100% buffer A. L-arabino-oligosaccharides (DP = 2-9) obtained
890 commercially (Megazyme) were used as standards at a concentration of 25 µM. Data were processed using Chromeleon™ Chromatography Management System V.6.8.
892 Final graphs were created using GraphPad Prism 8.0.1.

Purification of mycobacterial arabinogalactan

894 Large scale purification of mycobacterial arabinogalactan was achieved by established methodologies[70]. In brief, 8 L of mycobacterial culture was grown to mid-
896 exponential phase, cultures were pelleted and resuspended in a minimal volume of phosphate-buffered saline (VWR) and lysed using an Emulsiflex. The lysate was then
898 boiled in a final concentration of 1% sodium dodecyl sulphate (SDS) and refluxed. Insoluble material (containing mycolyl-arabinogalactan-peptidoglycan complex) was

900 collected by centrifugation and washed exhaustively with water to remove SDS. The mycolate layer was removed by saponification by KOH in methanol at 37 °C for 3 days.
902 Cell wall material was then washed repeatedly to remove saponified mycolic acids with diethyl ether. The phosphodiester linkage between AG and PG was then cleaved
904 by treatment with H_2SO_4 at 95 °C before being neutralized with sodium carbonate. The resultant solubilized arabinogalactan was collected in the supernatant, dialyzed
906 exhaustively against water and lyophilized (yield= ~22.5 $mg \cdot L^{-1}$).

Purification of D-arabinan

908 Arabinogalactan (5 mg/mL) was digested in a 10 mL total volume of 25 mM MOPS buffer pH 7. One µM final concentration of BACFIN_04787 and BACFIN_08810 were
910 added, and the reaction was incubated at 37°C for 48h. An aliquot of the reaction was analyzed by TLC to verify the hydrolysis of the substrate and galactose release. Then,
912 the sample was dialyzed over-night against 5 L of deionized water using a 1 kDa membrane, to eliminate residual galactose from the reaction mixture. The sample was
914 then freeze-dried and resuspended in 0.4 mL water. A TLC analysis confirmed the elimination of the residual galactose from the sample, and this was further confirmed
916 through acid hydrolysis of the resulting D-arabinan to determine total abundance. To quantify the purity of isolated d-arabinan, an aliquot of both (0.25 mg/mL) d-arabinan
918 and arabinogalactan were treated by acid hydrolysis using 300 mM HCl at 100°C for 1h. Samples were neutralised to pH 7 with NaOH and analysed by HPAEC.
920 Quantitation was based on the migration of standards and the ratio between galactose and arabinose.

922

Purification of mycobacterial lipoglycans

924 One liter of mycobacterial culture was grown to mid exponential phase and pelleted as previously. The pellet was resuspended in 20 ml PBS, 0.1% Tween-80,

926 chilled and lysed by bead-beating. Lysate was transferred to a Teflon-capped glass tube and vortexed with an equal volume of citrate buffer saturated with phenol (Sigma-
928 Aldrich), and heated to 80 °C for 3 h, vortexing every hour. A biphase was generated by centrifugation at 2000 RPM at 10 °C, and the upper aqueous phase transferred to
930 a fresh glass tube and hot phenol wash repeated twice more. The resultant protein free glycan mixture was dialyzed exhaustively against tap water overnight to remove
932 trace phenol and lyophilized, yielding 34 mg of crude lipoglycans (LAM, LM, PIMS) per liter of culture.

934

***Pseudomonas aeruginosa* pilin oligosaccharide extraction.**

936 Pilins were purified as described by Burrows and colleagues, with some modifications[42]. Briefly, *Pseudomonas aeruginosa* PA7 were streaked out in a grid
938 pattern onto LB agar plates and grown for 24 hours at 37 °C. Cells were then scraped from all plates using sterile cell scrapers and resuspended in 4 ml of sterile phosphate-
940 buffered saline (pH 7.4) per plate.

Pili were then sheared from the cell wall by vigorous vortexing of resuspended
942 cells for 2 min. This suspension was centrifuged for 5 min at 6000 *x g*. The supernatant was centrifuged for 20 min at 20,000 *x g*. Supernatants were transferred to fresh tubes
944 and $MgCl_2$ was added to give a final concentration of 100 mM. Following inversion of these tubes to ensure mixing, samples were incubated at 4 °C overnight, allowing
946 precipitation of sheared proteins. Samples were then centrifuged for 20 minutes at 20,000 *x g*, yielding a precipitate smudge which was resuspended in 50 mM NH_4HCO_3,
948 pH 8.5 and transferred to fresh Eppendorf tubes. This solution was then dialyzed into the same buffer to eliminate excess $MgCl_2$.

950 Bradford assays were then performed to assay the mass of protein in the sample, followed by digestion of the intact protein pilins using proteinase K in a 50:1

952 pilin to enzyme ratio by mass for 24 hours in the presence of 2 mM $CaCl_2$. Glycans were then purified from digested proteins using porous graphitized carbon
954 chromatography, where sugars were eluted from the column using a twofold increasing concentration series of a butan-1-ol:H_2O gradient from 1:32 to 1:1 using 1
956 mL elutions[71]. Thin-layer chromatography (TLC) of eluates showed various oligomers of arabinan present in all fractions, all of which were subsequently used as substrates
958 for potential arabinanases.

960 **Synthesis of pNP-α-D-arabinofuranoside**

Para-nitrophenol (pNP)-α/β-D-arabinofuranoside synthesis was achieved
962 following the established procedures [23].

964 **SEC-LS**

Molecular weights were determined by size exclusion chromatography coupled
966 light scattering using an Agilent MDS system with either an Agilent BioSEC 5 1000 Å, 4.6 x 300 mm, 5 µm or GE Superdex 200 5 15 mm columns at appropriate flow rates.
968 Detector offsets were calibrated using a BSA standard and concentrations were determined by refractive index.

970

Protein sequence alignment and structure visualization

972 Sequence alignments were performed using Clustal Omega and visualized using ESPript[64,72]. Figures of protein structures were generated with ChimeraX 1.4[73].

974 **Fluorescent-conjugated mAGP hydrolysis assay**

Fluorescently labeled mycolyl-arabinogalactan-peptidoglycan complex
976 (mAGP) was isolated from *Corynebacterium glutamicum ATCC13032* following previously reported methods[36-38]. In brief, cells were grown in the presence of 5-

AzFPA to saturation, reacted with DBCO-conjugated AF647 and then the mAGP was isolated. The isolated product was suspended in 2% SDS in PBS and split into the outlined treatment groups in Eppendorf tubes. Samples were pelleted by centrifugation at 15,000 *x g* for 5 min at 4ºC then washed with PBS once (100 µL). Following this wash, the mAGP was resuspended in 90 µL PBS and enzyme stock added for a final concentration of 5 µM of each enzyme. Samples were incubated at 37 ºC with rotation for 16 h. Following incubation, samples were pelleted by centrifugation at 15,000 *x g* for 5 min at 4 ºC then washed with PBS twice (100 µL). The pellets were then suspended in 2% SDS in PBS, transferred to a 96-well plate and the fluorescence emission of each well was then measured on a Tecan Infinite M1000 Pro microplate reader. Monitoring of AF647 fluorescence was achieved by exciting at 648 nm ± 5 nm and detecting at 671 nm ± 5 nm. Z-position was set to 2 mm, and the fluorimeter gain was optimized and then kept constant. Data are reported in relative fluorescence units (RFU) normalized to PBS treated controls.

Acknowledgements:

We thank members of the Birmingham mycobacteriology group and the Newcastle glycobiology group for support and discussions. We gratefully acknowledge Diamond synchrotron source for access to X-ray beamtime.

Funding:

PJM is supported by a BBSRC David Phillips Fellowship (BB/S010122/1) and a BBSRC Impact Acceleration Award (1544084). ECL is supported by an Academy of Medical Sciences Springboard Award (SBF006\1048) and a Royal Society Research Grant (RGS\R2\202228). ALL is supported by a Wellcome Trust Investigator Award in Science (209437/Z/17/Z). SJW is supported by the Australian Research Council (DP210100233, DP210100233). STB is supported by a studentship from the Wellcome Trust. NPB is supported by a BBSRC studentship (BB/M011186/1). OA-J is funded by a BBSRC studentship (BB/M011186/1). JM-M was supported by an Advanced Grant from the European Research Council (Grant no. 322820) awarded to Harry J Gilbert. AC is funded by Academy of Medical Sciences Springboard (SBF005\1065) and a Royal Society research grant (RGS\R2\212050). This research was supported by the National Institute of Allergy and Infectious Disease (R01 AI-126592) and the NIH Common Fund (U01GM125288).

Author Contributions:

Omar Al-Jourani – Methodology, validation, formal analysis, investigation, data curation, visualization. **Samuel Benedict** – Methodology, validation, formal analysis, investigation, data curation, visualization, resources. **Jennifer Ross** – Methodology, validation, formal analysis, investigation, data curation, writing – original draft preparation, writing – review and editing. **Abigail Layton** – Methodology, validation, resources. **Phillip van der Peet** – Methodology, validation, formal analysis, resources. **Victoria Marando** – Methodology, formal analysis, investigation, visualization, writing – review and editing. **Nicholas P Bailey** – Methodology, validation, formal analysis, visualization. **Tiaan Heunis** – Methodology, formal analysis, resources. **Joseph Manion** – Investigation. **Francesca Mensitieri** – Investigation. **Aaron Franklin** – Investigation, resources. **Javier Abellon-Ruiz** – Investigation. **Sophia L Oram,** – Investigation. **Lauren Parsons** – Investigation. **Alan Cartmell** – Resources, supervision. **Gareth SA Wright** – Resources, supervision. **Arnaud Baslé** – Methodology, validation, writing – review and editing. **Matthias Trost** – Resources, supervision. **Bernard Henrissat** – Resources. **Jose Munoz-Munoz** – Investigation. **Robert P Hirt** – Investigation, resources, writing – review and editing, supervision. **Laura L. Kiessling** – Resources, supervision, funding acquisition, writing – review and editing. **Andrew Lovering** – Methodology, validation, formal analysis, writing – review and editing, supervision, funding acquisition. **Spencer J. Williams** – Resources, validation, formal analysis, writing – review and editing, supervision, funding acquisition. **Elisabeth C. Lowe** – Conceptualization, methodology, validation, formal analysis, data curation, writing – original draft preparation, writing – review and editing, visualization, supervision, funding acquisition. **Patrick J. Moynihan** – Conceptualization, methodology, validation, formal analysis, data curation, writing – original draft preparation, writing – review and editing, visualization, supervision, funding acquisition.

Competing Interests

Drs. Moynihan and Lowe are co-inventors on an unpublished patent application pertaining to some of the enzymes described in this manuscript.

References

1048 1 Abrahams, K. A. & Besra, G. S. Synthesis and recycling of the mycobacterial cell envelope. *Curr Opin Microbiol* **60**, 58-65 (2021).
1050 https://doi.org:10.1016/j.mib.2021.01.012

2 Dulberger, C. L., Rubin, E. J. & Boutte, C. C. The mycobacterial cell envelope
1052 - a moving target. *Nat Rev Microbiol* **18**, 47-59 (2020). https://doi.org:10.1038/s41579-019-0273-7

1054 3 Daffe, M., McNeil, M. & Brennan, P. J. Major structural features of the cell wall arabinogalactans of Mycobacterium, Rhodococcus, and Nocardia spp.
1056 *Carbohydr Res* **249**, 383-398 (1993). https://doi.org:10.1016/0008-6215(93)84102-c

1058 4 Justen, A. M. *et al.* Polysaccharide length affects mycobacterial cell shape and antibiotic susceptibility. *Sci Adv* **6** (2020).
1060 https://doi.org:10.1126/sciadv.aba4015

5 Safi, H., Sayers, B., Hazbon, M. H. & Alland, D. Transfer of embB codon 306
1062 mutations into clinical Mycobacterium tuberculosis strains alters susceptibility to ethambutol, isoniazid, and rifampin. *Antimicrob Agents Chemother* **52**, 2027-
1064 2034 (2008). https://doi.org:10.1128/AAC.01486-07

6 Sun, Q. *et al.* Mutations within embCAB Are Associated with Variable Level of
1066 Ethambutol Resistance in Mycobacterium tuberculosis Isolates from China. *Antimicrob Agents Chemother* **62** (2018). https://doi.org:10.1128/AAC.01279-
1068 17

7 Zhao, L. L. *et al.* Analysis of embCAB mutations associated with ethambutol
1070 resistance in multidrug-resistant mycobacterium tuberculosis isolates from China. *Antimicrob Agents Chemother* **59**, 2045-2050 (2015).
1072 https://doi.org:10.1128/AAC.04933-14

8 Brossier, F. *et al.* Molecular Analysis of the embCAB Locus and embR Gene
1074 Involved in Ethambutol Resistance in Clinical Isolates of Mycobacterium tuberculosis in France. *Antimicrob Agents Chemother* **59**, 4800-4808 (2015).
1076 https://doi.org:10.1128/AAC.00150-15

9 Zhang, L. *et al.* Structures of cell wall arabinosyltransferases with the anti-
1078 tuberculosis drug ethambutol. *Science* **368**, 1211-1219 (2020). https://doi.org:10.1126/science.aba9102

1080 10 Winder, F. G. & Collins, P. B. Inhibition by isoniazid of synthesis of mycolic acids in Mycobacterium tuberculosis. *J Gen Microbiol* **63**, 41-48 (1970).
1082 https://doi.org:10.1099/00221287-63-1-41

11 Park, J. T. & Strominger, J. L. Mode of action of penicillin. *Science* **125**, 99-101
1084 (1957). https://doi.org:10.1126/science.125.3238.99

12 Healy, C., Gouzy, A. & Ehrt, S. Peptidoglycan Hydrolases RipA and Ami1 Are
1086 Critical for Replication and Persistence of Mycobacterium tuberculosis in the Host. *mBio* **11** (2020). https://doi.org:10.1128/mBio.03315-19

1088 13 Moynihan, P. J. *et al.* The hydrolase LpqI primes mycobacterial peptidoglycan recycling. *Nat Commun* **10**, 2647 (2019). https://doi.org:10.1038/s41467-019-
1090 10586-2

14 Kalscheuer, R., Weinrick, B., Veeraraghavan, U., Besra, G. S. & Jacobs, W. R.,
1092 Jr. Trehalose-recycling ABC transporter LpqY-SugA-SugB-SugC is essential

for virulence of Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A* **107**, 21761-21766 (2010). https://doi.org:10.1073/pnas.1014642108

15 Yang, Y. *et al.* A hydrolase of trehalose dimycolate induces nutrient influx and stress sensitivity to balance intracellular growth of Mycobacterium tuberculosis. *Cell Host Microbe* **15**, 153-163 (2014). https://doi.org:10.1016/j.chom.2014.01.008

16 Batinovic, S., Rose, J. J. A., Ratcliffe, J., Seviour, R. J. & Petrovski, S. Cocultivation of an ultrasmall environmental parasitic bacterium with lytic ability against bacteria associated with wastewater foams. *Nat Microbiol* **6**, 703-711 (2021). https://doi.org:10.1038/s41564-021-00892-1

17 Shen, L. *et al.* The endogenous galactofuranosidase GlfH1 hydrolyzes mycobacterial arabinogalactan. *J Biol Chem* **295**, 5110-5123 (2020). https://doi.org:10.1074/jbc.RA119.011817

18 Helbert, W. *et al.* Discovery of novel carbohydrate-active enzymes through the rational exploration of the protein sequences space. *Proc Natl Acad Sci U S A* **116**, 6063-6068 (2019). https://doi.org:10.1073/pnas.1815791116

19 Kotani, S., Kato, T., Matsubara, T., Sakagoshi, M. & Hirachi, Y. Inducible enzyme degrading serologically active polysaccharides from mycobacterial and corynebacterial cells. *Biken J* **15**, 1-15 (1972).

20 McNeil, M. R., Robuck, K. G., Harter, M. & Brennan, P. J. Enzymatic evidence for the presence of a critical terminal hexa-arabinoside in the cell walls of Mycobacterium tuberculosis. *Glycobiology* **4**, 165-173 (1994). https://doi.org:10.1093/glycob/4.2.165

21 Xin, Y. *et al.* Characterization of the in vitro synthesized arabinan of mycobacterial cell walls. *Biochim Biophys Acta* **1335**, 231-234 (1997). https://doi.org:10.1016/s0304-4165(97)00002-0

22 Dong, X., Bhamidi, S., Scherman, M., Xin, Y. & McNeil, M. R. Development of a quantitative assay for mycobacterial endogenous arabinase and ensuing studies of arabinase levels and arabinan metabolism in Mycobacterium smegmatis. *Appl Environ Microbiol* **72**, 2601-2605 (2006). https://doi.org:10.1128/AEM.72.4.2601-2605.2006

23 Kashima, T. *et al.* Identification of difructose dianhydride I synthase/hydrolase from an oral bacterium establishes a novel glycoside hydrolase family. *J Biol Chem* **297**, 101324 (2021). https://doi.org:10.1016/j.jbc.2021.101324

24 El Kaoutari, A., Armougom, F., Gordon, J. I., Raoult, D. & Henrissat, B. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nat Rev Microbiol* **11**, 497-504 (2013). https://doi.org:10.1038/nrmicro3050

25 Martens, E. C. *et al.* Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS Biol* **9**, e1001221 (2011). https://doi.org:10.1371/journal.pbio.1001221

26 Cuskin, F. *et al.* Human gut Bacteroidetes can utilize yeast mannan through a selfish mechanism. *Nature* **517**, 165-169 (2015). https://doi.org:10.1038/nature13995

27 Cartmell, A. *et al.* A surface endogalactanase in Bacteroides thetaiotaomicron confers keystone status for arabinogalactan degradation. *Nat Microbiol* **3**, 1314-1326 (2018). https://doi.org:10.1038/s41564-018-0258-8

28 Luis, A. S. *et al.* Dietary pectic glycans are degraded by coordinated enzyme pathways in human colonic Bacteroides. *Nat Microbiol* **3**, 210-219 (2018). https://doi.org:10.1038/s41564-017-0079-1

29 Alderwick, L. J. *et al.* Arabinan-deficient mutants of Corynebacterium glutamicum and the consequent flux in decaprenylmonophosphoryl-D-arabinose metabolism. *Glycobiology* **16**, 1073-1081 (2006). https://doi.org:10.1093/glycob/cwl030

30 Jacobson, R. H., Zhang, X. J., DuBose, R. F. & Matthews, B. W. Three-dimensional structure of beta-galactosidase from E. coli. *Nature* **369**, 761-766 (1994). https://doi.org:10.1038/369761a0

31 De Castro, C., De Castro, O., Molinaro, A. & Parrilli, M. Structural determination of the O-chain polysaccharide from Agrobacterium tumefaciens, strain DSM 30205. *Eur J Biochem* **269**, 2885-2888 (2002). https://doi.org:10.1046/j.1432-1033.2002.02955.x

32 Li, D. *et al.* 3beta-Hydroxysteroid dehydrogenase expressed by gut microbes degrades testosterone and is linked to depression in males. *Cell Host Microbe* **30**, 329-339 e325 (2022). https://doi.org:10.1016/j.chom.2022.01.001

33 Treerat, P. *et al.* Synergism between Corynebacterium and Streptococcus sanguinis reveals new interactions between oral commensals. *ISME J* **14**, 1154-1169 (2020). https://doi.org:10.1038/s41396-020-0598-2

34 Meng, F., Wang, C. & Kurgan, L. fDETECT webserver: fast predictor of propensity for protein production, purification, and crystallization. *BMC Bioinformatics* **18**, 580 (2018). https://doi.org:10.1186/s12859-017-1995-z

35 Perkowski, E. F. *et al.* The EXIT Strategy: an Approach for Identifying Bacterial Proteins Exported during Host Infection. *mBio* **8** (2017). https://doi.org:10.1128/mBio.00333-17

36 Marando, V. M. *et al.* Biosynthetic Glycan Labeling. *J Am Chem Soc* **143**, 16337-16342 (2021). https://doi.org:10.1021/jacs.1c07430

37 Calabretta, P. J., Hodges, H. L., Kraft, M. B., Marando, V. M. & Kiessling, L. L. Bacterial Cell Wall Modification with a Glycolipid Substrate. *J Am Chem Soc* **141**, 9262-9272 (2019). https://doi.org:10.1021/jacs.9b02290

38 Besra, G. S. *et al.* A new interpretation of the structure of the mycolyl-arabinogalactan complex of Mycobacterium tuberculosis as revealed through characterization of oligoglycosylalditol fragments by fast-atom bombardment mass spectrometry and 1H nuclear magnetic resonance spectroscopy. *Biochemistry* **34**, 4257-4266 (1995). https://doi.org:10.1021/bi00013a015

39 Chen, C. K., Chan, N. L. & Wang, A. H. The many blades of the beta-propeller proteins: conserved but versatile. *Trends Biochem Sci* **36**, 553-561 (2011). https://doi.org:10.1016/j.tibs.2011.07.004

40 Shallom, D. *et al.* Detailed kinetic analysis and identification of the nucleophile in alpha-L-arabinofuranosidase from Geobacillus stearothermophilus T-6, a family 51 glycoside hydrolase. *J Biol Chem* **277**, 43667-43673 (2002). https://doi.org:10.1074/jbc.M208285200

41 Xin, Y., Huang, Y. & McNeil, M. R. The presence of an endogenous endo-D-arabinase in Mycobacterium smegmatis and characterization of its oligoarabinoside product. *Biochim Biophys Acta* **1473**, 267-271 (1999). https://doi.org:10.1016/s0304-4165(99)00204-4

42 Voisin, S. *et al.* Glycosylation of Pseudomonas aeruginosa strain Pa5196 type IV pilins with mycobacterium-like alpha-1,5-linked d-Araf oligosaccharides. *J Bacteriol* **189**, 151-159 (2007). https://doi.org:10.1128/JB.01224-06

43 Villmones, H. C. *et al.* Investigating the human jejunal microbiota. *Sci Rep* **12**, 1682 (2022). https://doi.org:10.1038/s41598-022-05723-9

1192 44 Wesener, D. A. *et al.* Recognition of microbial glycans by human intelectin-1. *Nat Struct Mol Biol* **22**, 603-610 (2015). https://doi.org:10.1038/nsmb.3053
1194 45 Krupovic, M., Makarova, K. S. & Koonin, E. V. Cellular homologs of the double jelly-roll major capsid proteins clarify the origins of an ancient virus kingdom.
1196 *Proc Natl Acad Sci U S A* **119** (2022). https://doi.org:10.1073/pnas.2120620119
46 Stewart, G. R., Patel, J., Robertson, B. D., Rae, A. & Young, D. B.
1198 Mycobacterial mutants with defective control of phagosomal acidification. *PLoS Pathog* **1**, 269-278 (2005). https://doi.org:10.1371/journal.ppat.0010033
1200 47 Delafont, V. *et al.* Mycobacterium llatzerense, a waterborne Mycobacterium, that resists phagocytosis by Acanthamoeba castellanii. *Sci Rep* **7**, 46270
1202 (2017). https://doi.org:10.1038/srep46270
48 Antoine, R., Gaudin, C. & Hartkoorn, R. C. Intragenic Distribution of IS6110 in
1204 Clinical Mycobacterium tuberculosis Strains: Bioinformatic Evidence for Gene Disruption Leading to Underdiagnosed Antibiotic Resistance. *Microbiol Spectr*
1206 **9**, e0001921 (2021). https://doi.org:10.1128/Spectrum.00019-21
49 Mueller, E. A., Egan, A. J., Breukink, E., Vollmer, W. & Levin, P. A. Plasticity of
1208 Escherichia coli cell wall metabolism promotes fitness and antibiotic resistance across environmental conditions. *Elife* **8** (2019).
1210 https://doi.org:10.7554/eLife.40754
50 Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster,
1212 cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res* **47**, e47 (2019). https://doi.org:10.1093/nar/gkz114
1214 51 Zougman, A., Selby, P. J. & Banks, R. E. Suspension trapping (STrap) sample preparation method for bottom-up proteomics analysis. *Proteomics* **14**, 1006-
1216 1000 (2014). https://doi.org:10.1002/pmic.201300553
52 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates,
1218 individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367-1372 (2008).
1220 https://doi.org:10.1038/nbt.1511
53 Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant
1222 environment. *J Proteome Res* **10**, 1794-1805 (2011). https://doi.org:10.1021/pr101065j
1224 54 Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expr Purif* **41**, 207-234 (2005).
1226 https://doi.org:10.1016/j.pep.2005.01.016
55 Sauter, N. K. & Zwart, P. H. Autoindexing the diffraction patterns from crystals
1228 with a pseudotranslation. *Acta Crystallogr D Biol Crystallogr* **65**, 553-559 (2009). https://doi.org:10.1107/S0907444909010725
1230 56 Winter, G., Lobley, C. M. & Prince, S. M. Decision making in xia2. *Acta Crystallogr D Biol Crystallogr* **69**, 1260-1273 (2013).
1232 https://doi.org:10.1107/S0907444913015308
57 Evans, P. R. & Murshudov, G. N. How good are my data and what is the
1234 resolution? *Acta Crystallogr D Biol Crystallogr* **69**, 1204-1214 (2013). https://doi.org:10.1107/S0907444913000061
1236 58 McCoy, A. J. *et al.* Phaser crystallographic software. *J Appl Crystallogr* **40**, 658-674 (2007). https://doi.org:10.1107/S0021889807021206
1238 59 Krissinel, E., Uski, V., Lebedev, A., Winn, M. & Ballard, C. Distributed computing for macromolecular crystallography. *Acta Crystallogr D Struct Biol*
1240 **74**, 143-151 (2018). https://doi.org:10.1107/S2059798317014565

60 Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486-501 (2010). https://doi.org:10.1107/S0907444910007493
61 Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* **67**, 355-367 (2011). https://doi.org:10.1107/S0907444911001314
62 Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci* **27**, 293-315 (2018). https://doi.org:10.1002/pro.3330
63 Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-221 (2010). https://doi.org:10.1107/S0907444909052925
64 Sievers, F. & Higgins, D. G. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol Biol* **1079**, 105-116 (2014). https://doi.org:10.1007/978-1-62703-646-7_6
65 Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* **27**, 221-224 (2010). https://doi.org:10.1093/molbev/msp259
66 Trifinopoulos, J., Nguyen, L. T., von Haeseler, A. & Minh, B. Q. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res* **44**, W232-235 (2016). https://doi.org:10.1093/nar/gkw256
67 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49**, W293-W296 (2021). https://doi.org:10.1093/nar/gkab301
68 UniProt, C. UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204-212 (2015). https://doi.org:10.1093/nar/gku989
69 Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679-682 (2022). https://doi.org:10.1038/s41592-022-01488-1
70 Shenderov, K. *et al.* Cord factor and peptidoglycan recapitulate the Th17-promoting adjuvant activity of mycobacteria through mincle/CARD9 signaling and the inflammasome. *J Immunol* **190**, 5722-5730 (2013). https://doi.org:10.4049/jimmunol.1203343
71 Moynihan, P. J. & Clarke, A. J. Mechanism of action of peptidoglycan O-acetyltransferase B involves a Ser-His-Asp catalytic triad. *Biochemistry* **53**, 6243-6251 (2014). https://doi.org:10.1021/bi501002d
72 Gouet, P., Robert, X. & Courcelle, E. ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins. *Nucleic Acids Res* **31**, 3320-3323 (2003). https://doi.org:10.1093/nar/gkg556
73 Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **30**, 70-82 (2021). https://doi.org:10.1002/pro.3943