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 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 SA 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; tuberculosis (TB);

Abbreviations:

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

Abstract

Division and degradation of bacterial cell walls requires coordinated action from 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 exploited the breadth of carbohydrate active enzymes in the human gut microbiota to identify four families of glycoside hydrolases each with the capability to degrade the D-arabinan or D-galactan components of arabinogalactan. We have discovered novel exo-D-galactofuranosidases from gut bacteria and used them to discover both endo- and exo- acting enzymes that cleave D-arabinan. This includes new members of the DUF2961 family (GH172), and a novel family of glycoside hydrolases (DUF4185) which display endo-D-arabinofuranase activity. The DUF4185 enzmyes 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 essential cell wall components arabinogalactan and lipoarabinomannan, suggesting they are important to cell wall modification and/or degradation. Identification of these enzymes will enable isolation and analysis of mycobacterial cell wall components and facilitate the discovery of new therapeutic or diagnostic options for mycobacterial diseases.

Introduction

68 Growth and division of all bacteria is a carefully orchestrated process requiring the coordinated action of a host of enzymes. At each round of division, the bacterial 70 cell wall must be carefully cleaved while preventing lysis due to weakening of the envelope. This is especially true of acid-fast organisms such as *Mycobacterium* 72 *tuberculosis* where their unusual cell wall glycans require additional enzymatic machinery to those typically used for peptidoglycan cleavage. The core of the cell wall 74 structure consists of three layers, each conserved amongst mycobacteria[1,2]. Like other bacteria, peptidoglycan is thought to form the basal layer of the cell wall, though the 76 precise architecture is unknown. At the other extremity of the wall are the mycolic acids which give the organisms their characteristic waxy appearance and are interspersed 78 with a host of species-specific free lipids. Joining these two layers is a complex polysaccharide called arabinogalactan (AG), which has a chemical composition 80 unique to the Mycobacteriales and entirely distinct from the similarly named molecule found in plants (**Figure 1A**)[3]. AG is comprised of two domains with a β-D- 82 galactofuranose backbone decorated by large α-D-arabinofuranose branches[4]. The structure and biosynthesis of this molecule has undergone intense scrutiny, due in 84 part to it being a target of the antimycobacterial drug ethambutol[5-8]. Ethambutol targets the glycosyltransferase family C proteins in the cell envelope of mycobacteria, which 86 are responsible for polymerisation and decoration of the polysaccharide[9]. Similarly, the biogenesis of mycolic acids and peptidoglycan are the subject of much research 88 due to their biochemical complexity, essentiality, and their synthesis being the target of isoniazid and β-lactams respectively[10,11].

90 The growth of these bacteria however likely require not only the coordinated synthesis of the cell wall, but also its eventual degradation and turn-over to allow for

92 cellular expansion, division and the insertion of cell envelope spanning secretion systems. Similarly, the control of cell surface structures affords these bacteria the
94 ability to modulate how they are sensed by the host. As such, hydrolases of the cell wall could play an important, if under-studied role at the mycobacterial cell surface.
96 Hydrolases of peptidoglycan have been shown to be important to mycobacterial morphology and infection and we recently demonstrated that at least some
98 peptidoglycan by-products are recycled by the organism[12,13]. Similarly, components of the dominant lipid component of the cell wall, trehalose-mycolates, have also been
100 shown to be recycled[14]. Indeed, a hydrolase involved in turning over mycolic acids (Rv3451) and a recycling pathway for trehalose has previously been described[15].

102 Only a single enzyme capable of degrading AG has been characterised to date. GlfH1 (Rv3096) is an exo-acting enzyme from glycoside hydrolase (GH) family
104 GH5_13. It cleaves the galactan backbone from mycobacterial AG at both β-1,5 and β-1,6 linked residues[16]. The precise role of this enzyme in mycobacterial biology
106 remains unclear but its activity is suggestive of it being part of an active remodelling pathway for this cell wall component. Enzymes with more restrictive activity on D-
108 galactofuranose (D-Gal*f*) have not been described, though evidence that they exist was demonstrated in a large screen recently[17].

110 Despite this biological need, no enzymes have ever been characterised with exo- or endo-D-arabinofuranase activity (enzymes that cut within the D-arabinan
112 polymers) against native substrates, although this activity has been reported in crude protein extracts dating back to the 1970s[18,19]. Most recently, endo- activity was
114 described in extracts of *Mycobacterium smegmatis,* and through the development of a radiometric assay this activity was suggested to increase upon treatment of cells
116 with ethambutol, which blocks D-arabinan biosynthesis[20,21]. During the preparation of

this manuscript an exo-acting difructose-dianhydride I synthase/hydrolase was shown
118 to be able to cleave pNP-α-D-arabinofuranoside[22]. Whether this enzyme could work on the type of polymeric D-arabinan substrates within a cell wall is not clear.

120 Cleavage of the acid-fast cell wall is unlikely to be restricted to the Mycobacteriales. Organisms that predate on acid-fast bacteria such as phage or
122 bacterial predators like the recently reported *Candidatus Mycosynbacter amalyticus,* are also likely to require D-arabinan degrading enzymes to bypass the mycobacterial
124 cell wall[23]. In addition, the human gut microbiota is responsible for the degradation of dietary plant polysaccharides, host and micriobial glycans. Dominated by the
126 Bacteroidetes, this grouping of organisms is collectively amongst the richest known organisms in diversity of glycolytic enzymes[24]. Carbohydrate utilisation by the
128 Bacteroidetes is typically mediated by genes which are organised into polysaccharide utilisation loci (PULs), which can be induced upon exposure of the bacterium to a given
130 carbohydrate[25-28]. The diversity and compartmentalisation of polysaccharide utilisation represents a profound opportunity for enzyme discovery.

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

142 **Results**

144 *The human gut microbiome can utilise mycobacterial AG as a carbon source*

Endo-D-arabinofuranase activity was first described in soil bacteria more than 146 50 years ago and has been known in mycobacteria for at least 30 years[18,19]. Despite this activity, the enzymes which catalyse this activity have escaped identification. We 148 reasoned that these enzymes might be highly regulated or unstable in mycobacteria and as such evade traditional reductionist approaches to identification. Similarly, if 150 they belong to novel enzyme classes, bioinformatics approaches such as those used to identify GlfH1 would not be appropriate. Instead, we purified arabinogalactan from 152 *M. smegmatis* mc^2155 and used it as a sole-carbon source for the growth of a panel of 14 Bacteroidetes species. Of these, 12 strains were able to grow on this material 154 (**Figure 1B**). Ion chromatography with pulsed amerometric detection (IC-PAD) analysis of selected culture supernatants (**Figure 1C**) demonstrated the production of 156 free galactose in most cultures, and arabinose in one; *Dysgonomonas gadei*. Together these data indicate that members of the gut microbiota produce enzymes that can 158 depolymerise mycobacterial D-arabinan and D-galactan.

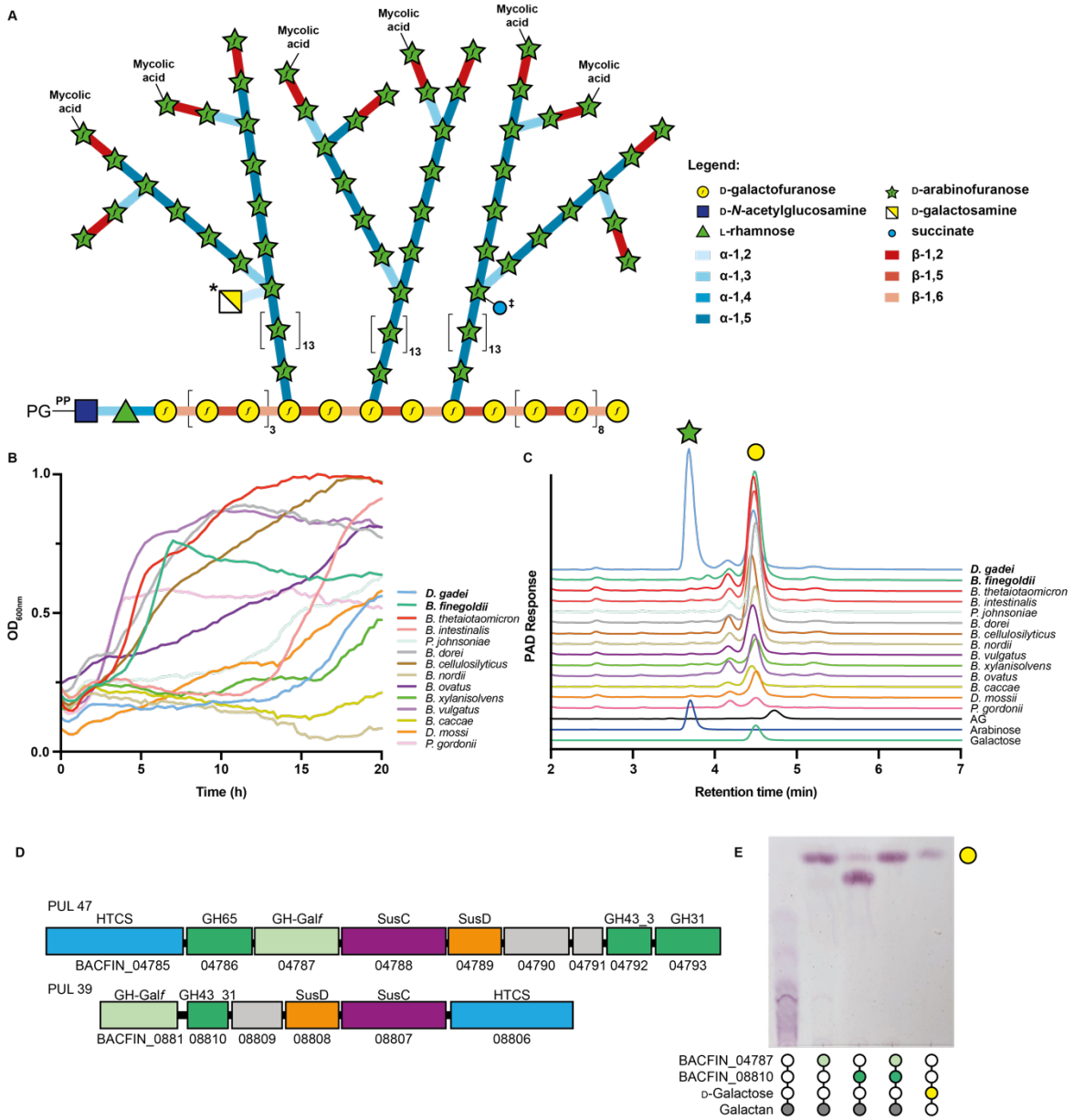


Figure 1. Bacteroidetes growth on mycobacterial arabinogalactan. A) Schematic of the structure of mycobacterial mycolyl-arabinogalactan-peptidoglycan complex. Succinate (‡) and galactosamine (*) modifications are non-stoichiometric. **B)** Growth of selected Bacteroidetes species on arabinogalactan as 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 as compared to standards. **D)** Schematics of PUL47 and 39 from *B. finegoldii*, whose homologs in *B. cellulosilyticus* were identified by RNAseq as being up-regulated 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.

Development of tools to identify D-Ara*F* degrading enzymes

174 The presence of both galactose and arabinose in *D. gadei* culture supernatants complicated the identification of PULs which would be specific for either carbon 176 source. To circumvent this, we developed a method by which we could produce pure D-arabinan to use in targetted growth assays by exploiting D-galactan specific PULs. 178 Based on the analysis of culture supernatants, *B. finegoldii* and *B. cellulosilyticus* appeared to readily turn over galactan, but not D-arabinan. Subsequent RNAseq 180 analysis of *B. cellulosilyticus* revealed the up regulation of 2 PULs, 35 and 36 containing multiple predicted GHs (**Figure S1**). Due to issues with cloning *B.* 182 *cellulosilyticus* enzymes, we cloned, over-expressed and purified the homologs of these enzymes from *B. finegoldii* DSM17565 (PULDB ID: PUL39 and PUL47, **Figure** 184 **1D**) and tested their activity on galactan (**Figure 1E**) demonstrating a mix of exo- and endo- activities. A detailed biochemical study of these enzymes will follow in a 186 subsequent report. To determine the extent of their galactan degradative capacity, we purified galactan from a strain of *Corynebacterium glutamicum* (Δ*ubiA*) which 188 completely lacks D-arabinan to reduce background from D-arabinan oligosacchareides[29]. As shown in **Figure 1E** and **Figure S1**, a combination of two *B.* 190 *finegoldii* enzymes comprising a GH43_31 (BACFIN_08810) and a new GH-exo-Gal*f* (BACFIN_04787) family could hydrolyse this galactan substrate.

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

Having established a method to produce heavily enriched D-arabinan we next 194 sought to identify the D-arabinan degradation PUL in *D. gadei*. To achieve this, we incubated AG with the two galactan-degrading enzymes BACFIN_08810 and 196 BACFIN_04787. This treatment eliminated approximately 70% of the AG D-galactan component of the molecule as determined by acid hydrolysis (**Figure S1**). The

198 resulting polysaccharide was subsequently used as a sole carbon source for *D. gadei*, which was previously shown to produce both D-galactose and D-arabinose. We
200 analysed the proteome of these bacteria at mid-log phase during growth on enriched D-arabinan and identified a predicted fucose isomerase as the most abundant
202 carbohydrate-active enzyme (**Table S1**). This protein maps to PUL42 in the *D. gadei* genome **(Figure 2A**). A further nine of the proteins included in this PUL were in the
204 top 200 most abundant proteins in the total proteome (**Figure S2**). Consumption of carbohydrates by bacteria usually requires the activity of both exo- and endo- acting
206 glycosidases. Although the PUL was not rich in annotated GHs, we identified the DUF2961 and DUF4185 superfamily enzymes encoded in PUL42 through initial
208 homology to mycobacterial proteins as the most likely to fulfil these roles.

The DUF2961 superfamily (GH172) includes D-arabinofuranosidases

210 At the outset of this study, no member of the DUF2961 family had been characterised. To test the activity of the DUF2961 family members in *D. gadei* we
212 cloned, expressed, and purified all three family members found in PUL42. Upon incubating the enzymes with purified AG, we observed D-arabinofuranosidase activity
214 for HMPREF9455_02467, HMPREF9455_02471 and HMPREF9455_02479 (hereafter referred to as Dg_{GH172a}, Dg_{GH172b} and Dg_{GH172c}) **(Figure S3)**. To probe the
216 activity of these enzymes, we synthesised the chromogenic substrate *p*-nitrophenyl-D-arabinofuranoside (pNP-α-D-Ara*f*) and it's β- derivative and tested the enzymes
218 against each. No activity was observed for pNP-β-D-Ara*f*, consistent with the number of α-AraF linkages found in the arabinan. Although substrate limitations prevented
220 reaching V_{max} we could use this substrate to determine the Michaelis-Menten kinetics of Dg_{GH172c} (**Figure S4, Table 1 and Table 2**). These data demonstrate that gut

222 bacteria can use DUF2961 enzymes to generate D-arabinose from mycobacterial arabinogalactan.

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

226 While the presence of D-arabinofuranosidases is somewhat surprising in gut microbes, we wanted to determine if similar enzymes were present in the genomes of 228 organisms from the Actinomycetota phylum. These genes were not readily observed in the *M. tuberculosis* complex of organisms, but DUF2961 encoding genes are 230 present in pathogens such as *Mycobacterium avium* subsp. *paratuberculosis* (MAP_0339c) and *Nocardia brasiliensis* (O3I_017420; Noc_{GH172}). To test for the 232 predicted enzymatic activity, we attempted to produce each recombinantly. Despite repeated attempts, we were unable to produce soluble MAP_0339c, however 234 Noc_{GH172} was readily soluble. This enzyme had D-arabinofuranosidase activity on purified arabinogalactan (**Table 2, Figure S3**). In addition to bacteria of the 236 Actinomycetota phylum, we identified a DUF2961 homolog encoded in the genome of the recently identified parasitic bacterium *Candidatus M. amalyticus*, which possessed 238 similar activity (GII36_05205; Myc_{GH172}) (**Figure 2 and S3A**)[23]. Together these data indicate that exo-D-arabinofuranosidase activity is a conserved feature of the 240 DUF2961 super-family. These enzymes are encoded by both the Mycobacteriales and their predators.

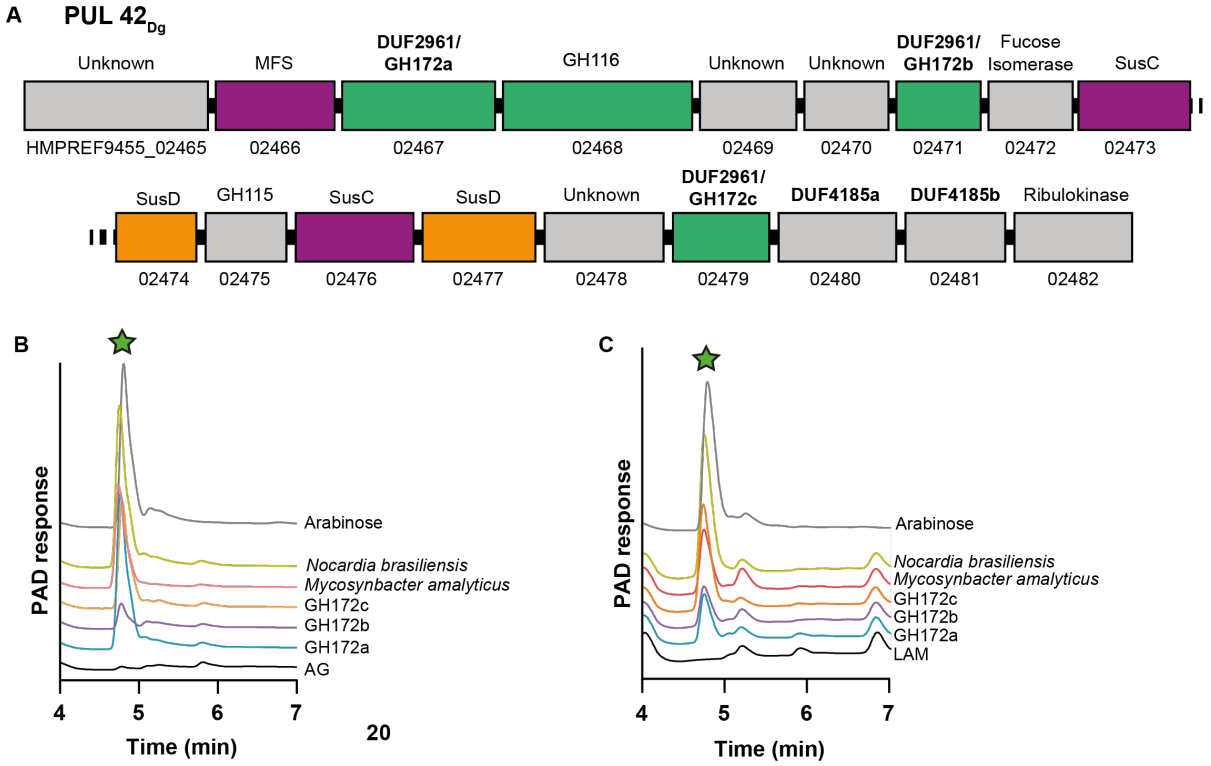


Figure 2. ***D. gadei* encodes GH172 exo-D-arabinofuranosidases with orthologs in diverse bacteria**. Schematic of PUL42 as identified by proteomic analysis of *D. gadeii* grown on D-arabinan (**A**). DUF2961 enzymes (1 μM) vs 2 mg ml^{-1} AG. **B**) and LAM **C**) analysed on a Dionex ICS-6000 with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

252 *GH172 enzymes degrade multiple mycobacterial cell wall components*

To further probe the substrate specificity of these enzymes, we assayed our 254 GH172s ((Dg_{GH172a}, Dg_{GH172b}, Dg_{GH172c}, Noc_{GH172}, and Myc_{GH172}) against multiple substrates (**Table 1**). The d-arabinan branches of lipoarabinomannan (LAM) are 256 broadly conserved in AG; therefore LAM was extracted from *M. smegmatis* and used as a substrate for the GH172 enzymes (**Figure 2C and Figure S3B**). Although 258 Dg_{GH172b} and Myc_{GH172} displayed lower activity than the others, each of the GH172s processed LAM to produce arabinose.

260 To determine if GH172 enzymes could cleave the major α-1,5 linkage of AG they were incubated with oligosaccharides from digested pili, purified from 262 *Pseudomonas aeruginosa* PA7, which have been demonstrated to be glycosylated with α-1,5 D-arabinofuranose oligosaccharides. Consistent with exo-D-arabinanase 264 activity (**Figure S3C**), the enzymes cleaved this material. We further ruled out D-galactofuranosidase activity by repeating these experiments using purified D-galactan 266 (**Figure S3D**) and observed no activity. The combination of these experiments reveals that GH172 enzymes are active on D-arabinan, and can cleave α-1,5 D-268 arabinofuranose oligosaccharide bonds.

The wider GH172 family can adopt multiple oligomeric states

270 Recent reports (PDB: 7V1V) and an unpublished structure (PDB: 4KQ7), have highlighted a hexameric configuration for the GH172 family of proteins (**Figure** 272 **S5**)[22,30]. To better understand the structure-function relationship of this protein family, we examined the multimeric state of several of our candidates of interest. Size-274 exclusion chromatography with multi-angle laser scattering (SEC-MALS) analysis of the GH172 family identified 4 distinct oligomerisation states consistent with structural

276 diversification within the family (**Table 3 and Figure S6**). Dg_{GH172c} has a calculated molecular weight from SEC-MALS consistent with a hexamer, while Dg_{GH172a} displays
278 a MW consistent with a dodecamer. Noc_{GH172} and Myc_{GH172} have solution oligomerisation states which are in agreement with trimeric assemblies whereas the
280 observed solution oligomerisation of Dg_{GH172b} is a dimer.

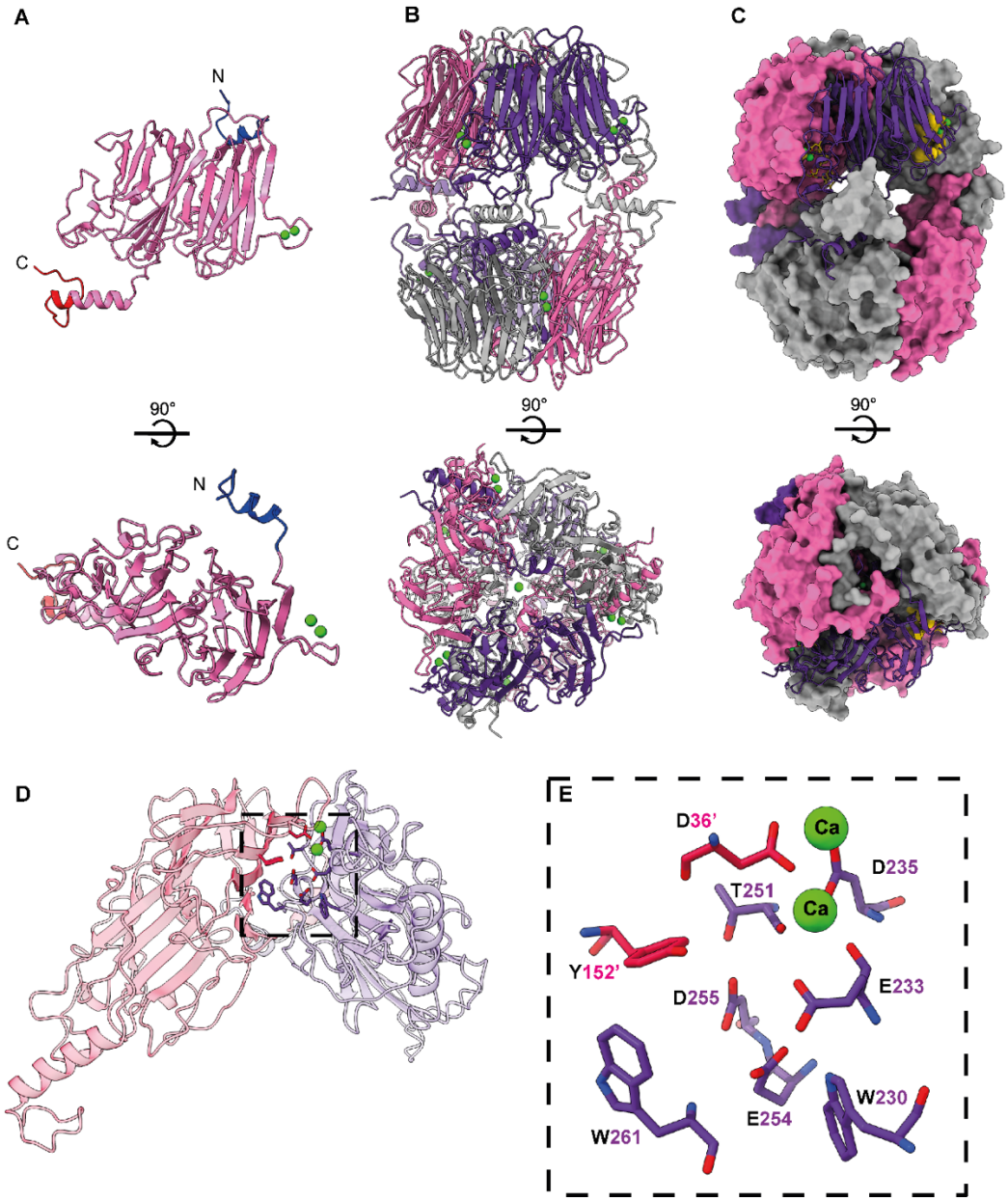
Interestingly, Dg_{GH172b} is predicted to have 1.5 DUF2961 domains, and so, the
282 dimeric assembly of Dg_{GH172b} results in a total of three DUF2961 domains. Our SEC-MALS analysis is supportive of a quaternary structure for each of the GH172 proteins
284 exhibiting DUF2961 domains in a multiple of three (three for Dg_{GH172b}, Myc_{GH172} and Noc_{GH172}; six for Dg_{GH172c}; and twelve for Dg_{GH172a}). This quaternary diversity within
286 the GH172 family could account for their differences in activity or their specificity.

Structural determination of Dg_{GH172c} reveals a dimer-of-trimers assembly with three
288 *active sites per trimer*

Our biochemical analysis pointed towards multiple possible oligomerisation states
290 for this family of enzymes. To further probe this we solved the X-ray crystal structure of Dg_{GH172c}. The crystal structure was determined using molecular replacement with
292 the *B. uniformis* GH172 (Bu_{GH172}, PDB: 4KQ7) used as the search model which shares 64% sequence identity with Dg_{GH172c}. X-ray crystallography data collection and model
294 refinement statistics are shown in **Table S2**. The structure of Dg_{GH172c} was refined to 1.4 Å resolution.

296 The Dg_{GH172c} monomer is made of two beta jelly rolls and a C-terminal alpha helix (**Figure 3A**). Six of these come together to form a dimer-of-trimers structure (**Figure**
298 **3B and Figure S5A**). This is consistent with the solution phase oligomerisation state of this protein from SEC-MALS (**Table 3**). 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, and six per dimer-of-trimers/biological assembly. There are 7 calcium ions coordinated within a Dg_{GH172c} trimer; two within each active site/protomer, and one in the 3-fold axes. The dimer-of-trimers structure is formed by the C-terminal of each 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.



308 **Figure 3: The crystal structure of Dg_{GH172c} determined by X-ray crystallography.**
A) Monomer of Dg_{GH172a} with secondary structure shown as cartoon. The N- and C-
310 termini are labelled and coloured blue and red respectively. Bound calcium ions are shown as green spheres. **B)** Overall hexameric architecture of Dg_{GH172c}. Cartoon
312 representation with bound calcium ions shown as green sphere. Trimer subunits are coloured grey, pink and purple for clarity. **C)** Surface representation of the hexamer
314 with one monomer shown in cartoon. Active site residues have been coloured gold, and calcium ions shown in green. **D)** Cartoon representation of two monomers of
316 Dg_{GH172c} which form the active site, with one monomer in pink and the other in purple. The active site residues and coordinated calcium ions are highlighted by a black
318 dashed box. **E)** Proposed active site residues of Dg_{GH172c}. Residues are shown as sticks with carbon atoms coloured pink and purple for different subunits. Calcium ions
320 are shown as green spheres and coordinated by D36', D235 and T251.

The active site of Dg_{GH172c} is principally formed of residues: D36, Y152, W230,
322 E233, D235, T251, E254, D255, and W261 (**Figure 3D**). The proposed catalytic residues identified in the characterisation of Bd_{GH172} are conserved in $Dg_{GH172c,}$ (Bd/Dg:
324 E270/254, catalytic nucleophile; E291/233, acid/base) and other members of the GH172 family (**Figure S7**). The two glutamate residues are 5.4 Å apart in Dg_{GH172c}
326 which is consistent with a retaining mechanism, as displayed by Bd_{GH172}. Calcium ions are observed in the active sites of Dg_{GH172c} (contains two calcium ions) and Bd_{GH172}
328 (contains a single calcium ion). The calcium ions in Dg_{GH172c} are coordinated by D235 which is not conserved in Bd_{GH172}. Instead, Bd_{GH172} has N272 which is only capable of
330 binding one calcium ion. A particular feature of characterised GH172 family members is the formation of the active site by two monomers, this feature is uncommon amongst
332 GHs[31].

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

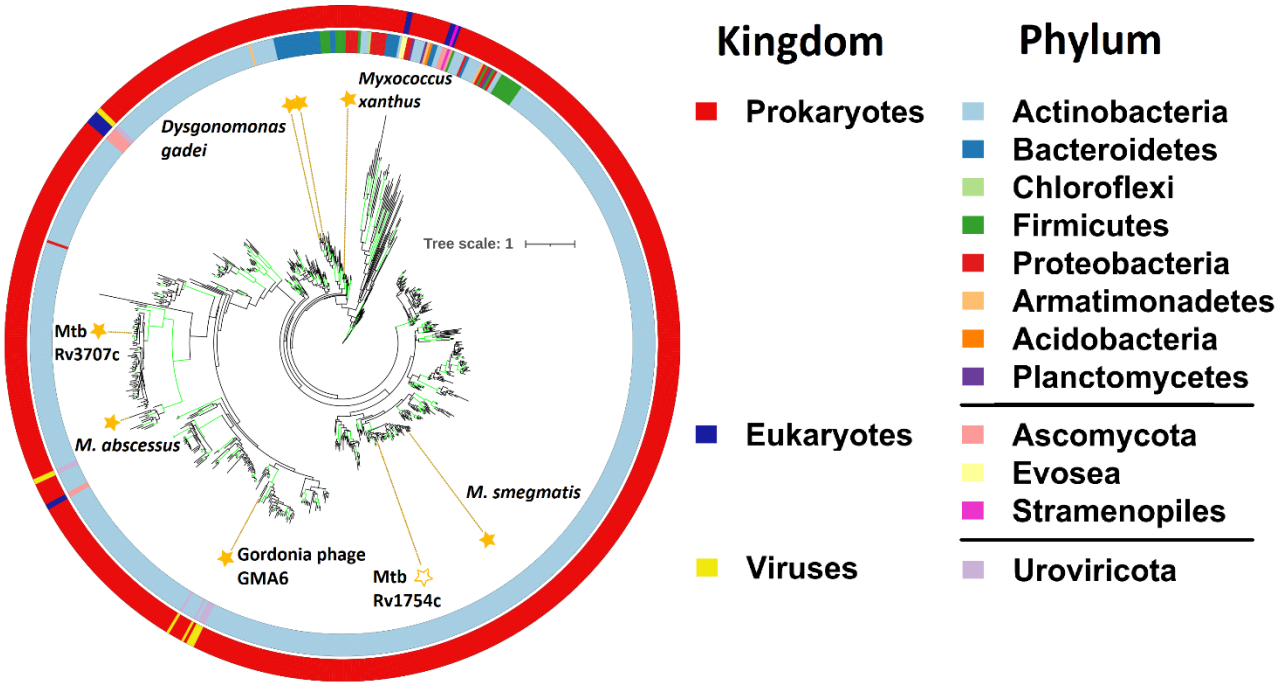
334 To gain insight into the functional roles of the residues in the active site of Dg_{GH172c}, each proposed catalytic residue was individually substituted for alanine
336 (generating variants E233A, and E254A). The E233A variant had no detectable activity and there was a greater than 10^5 fold drop in activity for E254A (**Table 4**). Our

338 data support the assignment of E254 and E233 as catalytic residues, though we also observe a lack of activity for the adjacent D255A mutation. SEC-MALS was used to
340 interrogate the solution-phase oligomerisation of the Dg_{GH172c} variants however, none of the variants displayed a change in oligomerisation state suggesting they do not
342 possess structural roles (**Figure S8**). Taken together our data support the annotation of GH172 enzymes as highly unusual amongst glycoside hydrolases as a
344 consequence of their quaternary structure and the formation of their active site.

DUF4185 is widespread throughout bacterial species

346 Having established exo-arabinanase activity in *D. gadei* PUL42, we sought to identify the source of endo-activity likely to reside in the same cluster. In Bacteroidetes,
348 the degradation of a target polysaccharide is typically a multi-step process whereby oligosaccharides are generated and subsequently cleaved into their monosaccharide
350 constituents. We reasoned that the generation of D-arabinofuranose oligosaccharides was likely restricted to proteins that have no known function and so focused our
352 attention on the DUF4185 proteins, whose secondary structure prediction is consistent with glycosidase-like enzymes. To investigate the conservation of these genes across
354 biology, we constructed a phylogenetic tree of identified DUF4185 proteins demonstrating that they are found broadly in actinobacteria, especially amongst the
356 Actinomycetota (**Figure 4 and Figure S9**). Intriguingly these enzymes are also found in the lysis cassette of some actinobacteriophage (**Figure S10**), *Bacteroides* species,
358 *Myxococcus* and *Candidatus M. amalyticus* among others. While broad conservation of the DUF implies related functions, many CAZy families contain members with
360 divergent activities despite having similar sequence profiles and so we sought to confirm their function.

362



364

Figure 4: Phylogeny of the DUF4185 enzyme family. Unrooted ML phylogeny (LG
366 model with empirical base frequencies, invariable sites and the discreet gamma model) of DUF4185 family sequences. Branches with greater than 75% bootstrap
368 support (100 replicates) are highlighted in green. Units for tree scale are inferred substitutions per amino acid residue. Coloured rings indicate phylum (inner) and
370 kingdom (outer) taxonomy information for sequences. Stars highlight sequences of interest, and are filled for proteins which have been experimentally characterised in
372 this work.

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

To assess the activity of these genes, we cloned, expressed, and purified the
376 DUF4185 homologs (HMPREF9455_02480 and HMPREF9455_02481) from *D. gadei* (herein, referred to as $Dg_{GH4185a}$ and $Dg_{GH4185b}$ respectively). When incubated with
378 mycobacterial AG, these enzymes produced a banding pattern characteristic of an endo-acting GH (**Figure S11A**). From this we propose that *D. gadei* utilises DUF4185-
380 like enzymes to convert arabinan into oligosaccharides to be subsequently digested by DUF2961 enzymes.

382 *DUF4185 enzymes from diverse microbes are endo-D-arabinanases*

To capture the breadth of activity for DUF4185 enzymes we sought to evaluate
384 enzymes from each of the major lobes of the global phylogeny (**Figure 4**). We produced recombinant versions of proteins derived from several bacterial lineages and
386 a phage which may have particular biological importance. These include those originally identified in *D. gadei* in addition to *Myxococcus xanthus* ($Myxo_{GH4185}$),
388 *Candidatus M. amalyticus,* and *Gordonia* Phage GMA6 ($Phage_{GH4185}$). Where the DUF4185 domain was co-localised with several other large domains we produced
390 truncated variants (containing only the DUF4185 domain) of the protein due to low-solubility of the multi-domain constructs. All constructs except the *Candidatus M.*
392 *amalyticus* protein yielded soluble protein. As shown in **Figure 5** and **Figure S11**, all these enzymes possessed endo-D-arabinanase activity albeit with varying product
394 profiles when incubated under identical conditions suggesting differences in enzyme specificity. Some of the enzymes were also active against the linear α-1,5-D-
396 arabinofuranose oligosaccharides derived from *P. aeruginosa* PA7 pili, consistent with activity against the major linkage of mycobacterial D-arabinan (**Figure S11**). We were
398 surprised to find that endo-D-arabinanase activity is broadly conserved in organisms

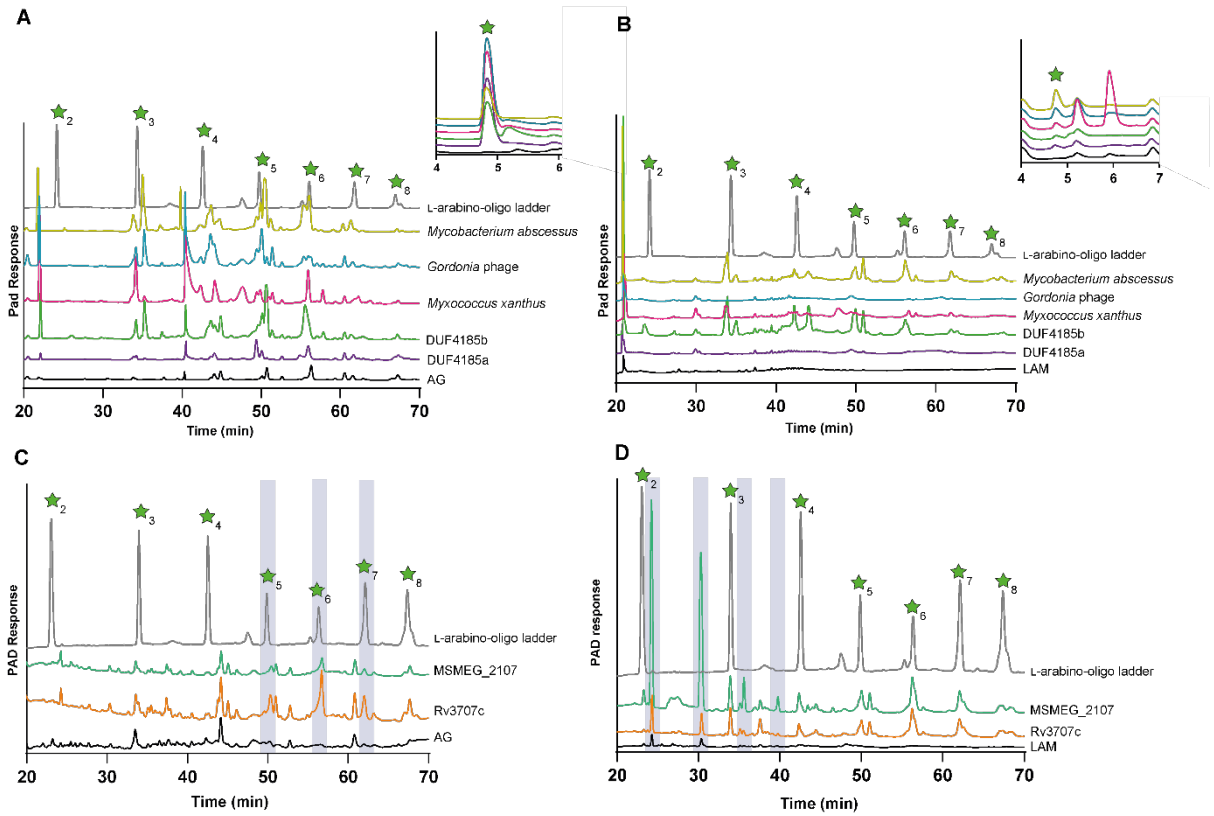
outside of the Actinomycetota, however its presence in bacterial predators is consistent with the essentiality of this polymer to mycobacterial viability. This sugar motif is also reported in a small number of LPS structures, possibly explaining the presence of this enzymatic activity in the gut microbiota[32].

Mycobacterial DUF4185s are endo-D-arabinofuranases

Given the importance of D-arabinan to mycobacterial viability and immunology we next sought to understand the biochemical function of DUF4185 proteins from these organisms. As shown in **Figure 4** and **Figure S9,** mycobacteria produce at least two DUF4185s which fall into distinct phylogenetic groupings. In *M. tuberculosis* these are Rv3707c and Rv1754. Beyond these two conserved DUF4185 genes, some species have shown further diversification. For example, many *Mycobacterium abscessus* strains encode at least three distinct copies whilst *M. smegmatis* mc^2155 encodes five such proteins (MSMEG_4352, 4360, 4365, 2107 and 6255). Based on sequence analysis, MSMEG_2107 and MSMEG_6255 are homologs of Rv1754 and Rv3707c respectively (**Figure S9**). Having identified candidate DUF4185 enzymes in mycobacteria, we wanted to determine if these proteins had similar activity profiles to those observed with the *D. gadei* enzymes. While efforts were made to produce both Rv3707c and Rv1754, we were initially unable to produce assayable quantities of either.

Due to the instability of the *M. tuberculosis* homologs, we analysed the entire DUF4185 Pfam family using the fDetect server, which identified a *Mycobacterium abscessus* homolog, Ga0069448_1118, as being the most likely mycobacterial DUF4185 family protein to be stably produced (hereafter referred to as Mab_{4185})[33]. Consistent with this prediction this enzyme is remarkably stable and is easily produced in soluble form and in good yield. Despite its low sequence identity (33.5%) (Figure

424 S12) to the *D. gadei* enymes, the TLC and HPAEC profiles of Mab_{4185} indicate that it also possesses similar endo-D-arabinofuranase activity (**Figure 5**). Thus, DUF4185-
426 catalysed endo-D-arabinanase activity is present in mycobacteria.



428 **Figure 5: High-performace anion-exchange chromatography/pulsed amperometric detection (HPAEC-PAD) analysis of DUF4185 enzyme-catalysed**
430 **reactions.** DUF4185 enzymes were incubated with 2 mg ml^{-1} AG (**A** and **C**) or 2 mg ml^{-1} LAM (**B** and **D**) for 16 hours as described in "Materials and Methods." Samples
432 were analysed by IC-PAD on a Dionex ICS-6000w with CarboPac Pa300 column, 100 mM NaOH 20 min isocratic elution followed by a 0-60% 500 mM sodium acetate
434 gradient over 60 minutes. A ladder of α-1,5-L-arabino-oligosaccharides (25 μM) derived from plant arabinogalactans was used as a standard. In panel C this ladder
436 has been scaled on the Y axis by a factor of 0.2 for clarity of presentation.

Encouraged by this success with *Mab*$_{4185}$, but recognising it has limited
438 sequence identity (17% identical over the entire protein length relative to Rv3707c) with either of the *M. tuberculosis* proteins, we sought to re-evaluate our expression
440 constructs. We reasoned that the instability of Rv3707c may be due to incorrect annotation of the start site in the *M. tuberculosis* H37Rv genome. Indeed, a previous
442 report highlighted that Rv3707c is secreted, but lacks a discernible signal peptide. Intriguingly, given that mycobacteria frequently use TTG and GTG starting codons in
444 addition to the canonical ATG, several in-frame N-terminal extensions of the gene are possible (**Figure S13**). We re-evaluated the genomic context of the protein and
446 compared these potential N-terminal extensions to alignments of DUF4185 proteins that possessed a second, N-terminal domain to identify any sequence motifs that are
448 conserved in this more structurally restrained context. In many cases a proline-rich region was observed, which is also found in some of the potential N-terminal
450 extensions of Rv3707c (**Figure S13**). Furthermore, by extending the N-terminus of the protein, a putative signal peptide can be predicted by SignalP 6.0 (**Figure S13**). While
452 the precise start-site is not certain, the likely cleavage point for the signal peptide remains unchanged. With this in mind, we produced a protein with the putative signal
454 peptide removed, but the remainder of the proline-rich N-terminus intact. This produced reasonably stable and soluble protein in good yield. We repeated the above
456 biochemical assays with Rv3707c and demonstrated that it generates products consistent with a hepta- or hexa-saccharide suggesting a narrower substrate
458 specificity than the *D. gadei* enzymes, when incubated with AG (**Figure 5C/D**). While we were unable to produce soluble Rv1754, we could produce the *M. smegmatis*
460 homolog (MSMEG_2107) and demonstrated that *in vitro* it is active against LAM, but not AG (**Figure 5C/D and Figure S11**).

462

DUF4185 family members from different clades have distinct substrate specificities

464 To further probe the substrate specificity of these enzymes and elucidate the function of MSMEG_2107, we incubated each of the DUF4185 enzymes ($Dg_{GH4185a}$, 466 $Dg_{GH4185b}$, Mab_{GH4185}, $Phage_{GH4185}$, $Myxo_{GH4185}$, MSMEG_2107, and Rv3707c) with: LAM (**Figure S11B**); purified D-galactan (**Figure S11C**); and pilin oligosaccharides 468 from *P. aeruginosa* PA7 (**Figure S11D**). Interestingly, all DUF4185 enzymes except $Dg_{GH4185a}$ were active against LAM, although to a lesser extent than observed for AG 470 (excluding MSMEG_2107). Having demonstrated activity for these enzymes on both AG and LAM we can conclude the DUF4185 family are endo-D-arabinofuranases 472 given that branched D-arabinan is the only conserved structure in both macromolecules. Intriguingly, only a subset of the DUF4185 enzymes were active 474 against pili from *P. aeruginosa* PA7. Given that these oligosaccharides are relatively short, this could be due to enzyme subsite requirements as evidenced by the 476 production of arabinose for some enzymes, despite not seeing this product with polymeric AG. To confirm that the enzymes only degraded D-arabinan and not D-478 galactan we tested their activity incubated with D-galactan but did not observe any product formation (**Figure S11C**).

480 Due to the limited substrate specificity of these enzymes we sought to determine if a selection of them could release AG from intact mycobacterial cell walls 482 by taking advantage of newly developed 5-AzFPA probes for arabinogalactan (**Figure S15**)[34-36]. Using this assay we show that Mab_{GH4185} releases more fluorescently 484 labelled material from cell wall than Rv3707c, which reflects the greater range of products from soluble AG observed by IC. The $GH172_{Noc}$ was similarly active against 486 this material, supporting the conclusion that both classes degraded similar substrates.

Together, these data indicate that endo-D-arabinofuranase activity is conserved in mycobacteria and catalysed by DUF4185 enzymes. The distinct product profiles of these enzymes also point to unique biological roles within the organism.

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

Whilst many of the DUF4185 homologs assayed above could be in *E. coli*, attempts to crystallise them were unsuccessful. We were able to obtain via X-ray diffraction a partial apo structure of *M. tuberculosis* Rv3707c at a resolution of 2.4 Å using the AlphaFold structural prediction as a search model. As shown in **Figure 6A**, our data indicate that Rv3707c possesses a beta-propeller structure, with each blade consisting of 3 antiparallel beta sheets organised radially around a central pore. This is consistent with a variety of glycoside hydrolaeses[37]. Poor or no density was obtained for amino acids 1-9, 22-38, 60-69, 85-86, 285-300, and 325-337, residues are located on loops surrounding the predicted active site. This lack of density suggests significant structural flexibility in these regions, consistent with the AlphaFold predictions for both Rv3707c and Rv1754c (**Figure 6, S14, S16**).

Multi-sequence alignments of DUF4185 homologs (**Figure S12**) could be used to identify three nearly invariant acidic residues in the active site; Asp39, Asp56, and Glu240 (**Figure 6B).** Comparing our experimental structure to the AlphaFold prediction model, the loops that contained the conserved aspartate residues in the presumed active site deviate in their positions (**Figure 6B and Figure S16**). The conserved D56 is pointed away from the active site in our experimental structure, likely in a non-catalytically competent position while the D39 is in a similar, but distinct position to that of the AlphaFold model. This orientation is suggestive of active site rearrangement upon binding of substrate.

512 The other regions of incomplete density include loops predicted by AlphaFold to be highly flexible, which is reflected in the b-factors for the experimental structure 514 (**Figure S16**). These regions are also predicted to be variable in other DUF4185 structures, including $Dg_{GH4185b}$ (**Figure 6C**). However, due to the highly branched 516 nature of mycobacterial D-arabinan, this flexibility likely enables adaptive binding to diverse substrates as evidenced by the broad reaction product profile of the enzymes 518 (**Figure 5**).

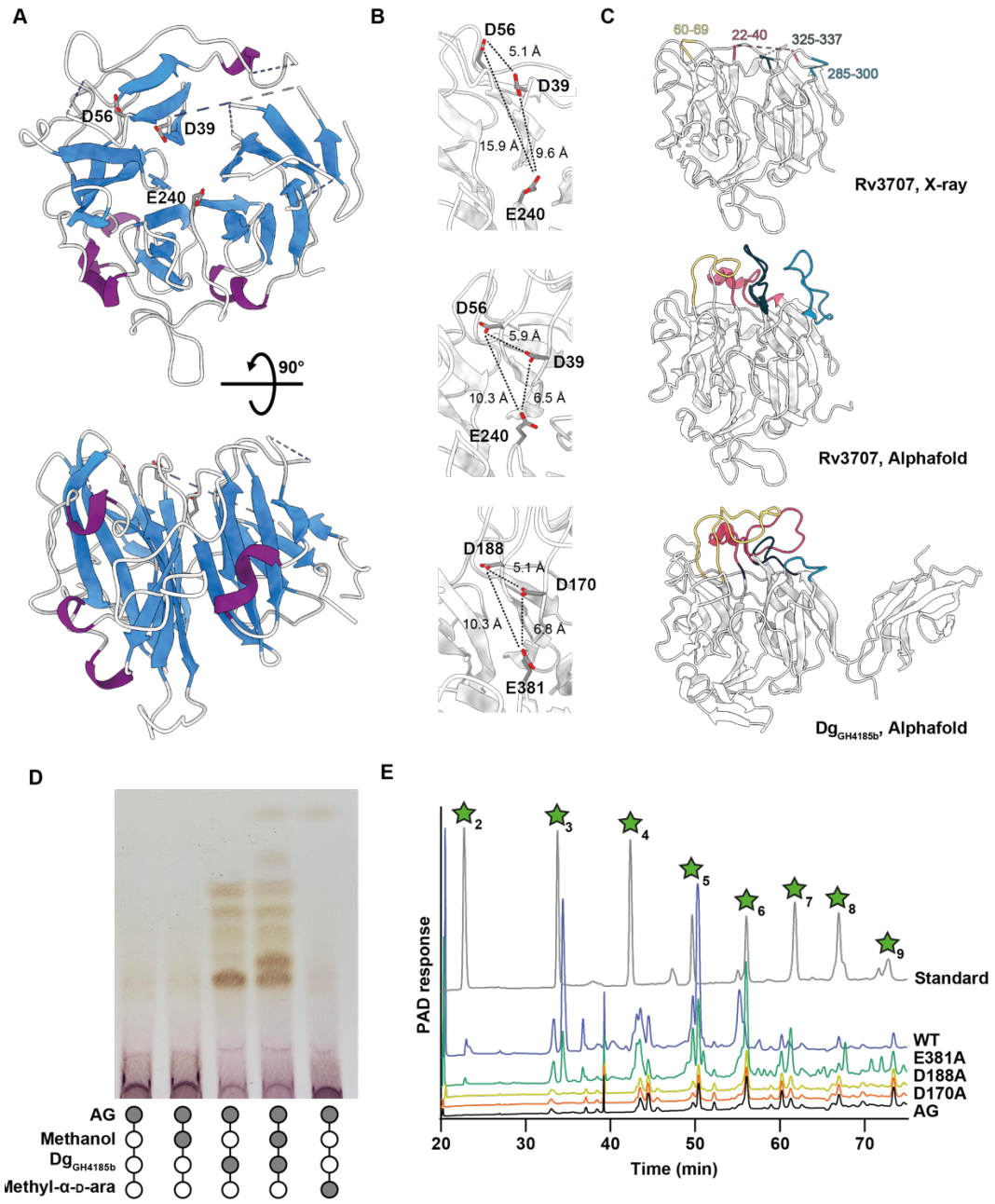


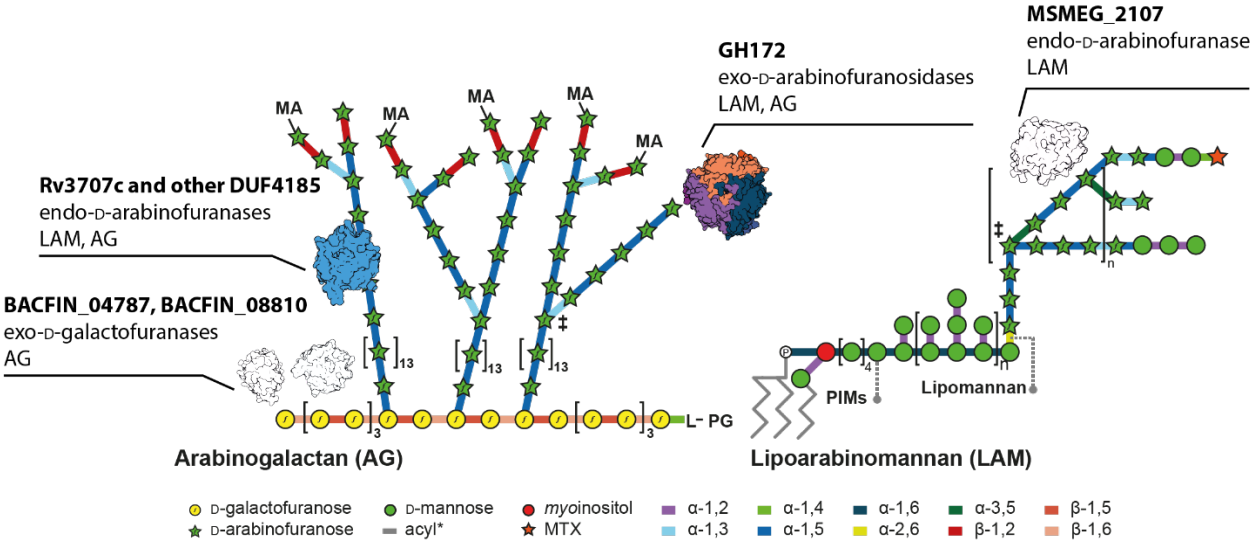
Figure 6: Structure and active site of DUF4185 enzymes

A) Rv3707c as determined by X-ray crystallography has a 5-bladed beta-propeller fold consistent with its function as a glycoside hydrolase. Conserved acidic residues D39, D56 and E240 are indicated in the presumed active site. **B**) Active site for Rv3707c (X-ray), Rv3707c (Alphafold) and $Dg_{GH4185b}$ (Alphafold) with distances between residues indicated in Å. **C**) Overall fold of the same enzymes as in B with differential surface features highlighted relative to the Rv3707c X-ray structure. Corresponding regions are in the same colour scheme in each protein. **D)** Methanolysis reaction with $Dg_{GH4185b}$. 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 discernible activity was observed for the D188 or D170 mutants.

532 *The DUF4185 family are anomer-retaining enzymes*

Glycoside hydrolases can hydrolyse the anomeric linkage through either
534 reversion or retention. Anomer-retaining reactions occur when the catalytic residues side chains are approximately, 5.5 Å away from one another. Retaining enzymes that
536 acto on furanoses can be differentiated from inverting enzymes when methanol is included in the reaction mixture, as they afford methylated glycosides[38]. Although
538 Rv3707c was unstable when methanol was included in reactions, $Dg_{GH4185b}$ afforded methylated reaction products that support a retaining mechanism (**Figure 6D**). In
540 retaining glycosidases, the acidic resides are critical for activity. As observed in **Figure 6B**, the pair of aspartate residues in the active site are about 5.5 Å apart, whereas the
542 conserved glutamate is either 10 or 6 Å. We varied these carboxylate residues in $Dg_{GH4185b}$ to generate $Dg_{GH4185b}$-D170A and $Dg_{GH4185b}$-D188A and $Dg_{GH4185b}$-E381A.
544 Activity was broadly retained for the glutamate substation, but no activity was observed for the two aspartate mutants, whereas activity was broadly retained for the glutamate
546 substitution (**Figure 6E**). These data support a retaining mechanism for the DUF4185 family of enzymes and assignment of D170/D39 and D188/D56 as the catalytic
548 residues in $Dg_{GH4185b}$ and Rv3707c respectively.

550



552 **Figure 7. Mycobacterial arabinogalactan-degrading enzymes discovered in this study and their substrates.** Each enzyme or enzyme family is listed with its identified
554 function and the structure within the mycobacterial cell wall it has confirmed activity for. Coloured in structures are those for which we have presented experimental data.
556 MA – mycolic acids; L – linker unit; PG – peptidoglycan; PIMs – phosphatidylinostol mannosides.
558

Discussion

560 Endo-D-arabinanase activity was first reported more than 50 years ago, but despite the widespread availability of mycobacterial genomic tools and -omics
562 technologies, these enzymes have escaped identification[21,39]. We have leveraged the human gut microbiome and evolutionary conservation to identify these enzymes in
564 addition to new exo-D-arabinofuranosidases and exo-D-galactofuranosidases (Figure 7). To achieve this, we took the unintuitive step of challenging bacteria that do not
566 contain arabinogalactan to use it as a sole carbon source. We reasoned that the abundance of Actinomycetota in environmental niches in addition to the availability of
568 D-arabinofuranose polymers in organisms such as *P. aeruginosa* PA7 and *Corynebacteria* meant that the capacity to degrade this carbohydrate was likely to be
570 encoded in the human gut microbiota[40,41].

Identifying organisms that can grow solely on D-arabinan is complicated by the
572 presence of D-galactan in AG. To circumvent this, we identified a group of enzymes that could degrade this polymer (**Figure 7**). The identification of D-galactofuranase
574 activity in a gut microbe organism is less surprising as this sugar moiety is widespread in gut microbiomes, often being found in the LPS of some Gram-negative bacteria[42].
576 However, one of the enzymes we have identified, BACFIN_04787, is the founding member of a new glycoside hydrolase family. The degradation pattern exhibited by
578 BACFIN_04787 is consistent with an exo-non-specific enzyme and the product pattern of BACFIN_08810 is consistent with an exo-acting enzyme that can cleave either the
580 β-1,5 or β-1,6 linkage within D-galactan. Further characterisation of these enzymes will enable sophisticated analysis of mycobacterial galactan, whose chain length was
582 recently shown to be important in the biology of these organisms[4].

The D-galactofuranase enzymes allowed us to identify *D. gadei* as a potent D-
584 arabinan degrader. The PULs associated with this ability were rich in DUFs, consistent with novel enzymatic activities. During the preparation of this manuscript, the GH172
586 family was associated with exo-D-arabinofuranase activity, which is consistent with our data[22]. Our *Bacteroides* enzyme is similar structurally to the enzyme described by
588 Kashima *et al*, however it lacks the large C-terminal extensions which were postulated to hold the hexameric protein together. Furthermore, our SEC-MALS analysis points
590 to as-of-yet unexplored structural diversity in the GH172 superfamily where we observed multimers of up to 12 protomers. A recent report suggested that the GH172
592 family are ancestrally related to phage capsid proteins raising the possibility that substantial, but as-yet unobserved, structural diversification has occurred[30].

594 The presence of both DUF4185 and GH172 enzymes in *Candidatus Mycosynbacter amalyticus* suggests that a component of the epibiotic lifestyle of this organism is to

596 feed on arabinogalactan as a carbon source, which is supported by their predicted localisation. The $Myc_{DUF4185}$ enzyme is predicted to be a lipoprotein, while Myc_{GH172} is
598 not predicted to be secreted. This localisation is consistent with a model whereby this organism releases oligosaccharides from the surface of *Gordonia* and then converts
600 them to monosaccharides in the cytoplasm. An alternative explanation is that the enzymes are used to locally remodel the cell wall, enabling access to cellular contents.
602 The biological role of GH172 enzymes in *Bacteroides* remains unclear. While some may be involved in α-fructan degradation, those associated with D-arabinan PULs are
604 more likely targeted at either glycans derived from Actinomycetota or organisms such as *P. aeruginosa* PA7.

606 The DUF4185 family of enzymes we have identified include the first known endo-D-arabinofuranases, though it is not the first time this activity has been reported.
608 In 1972 Kotani and 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”
610 fraction[18]. This enzyme mixture was shown to have endo-activity and released a wide range of reaction products from mycobacterial cell walls. This pioneering work enabled
612 early analyses of mycobacterial arabinogalactan, though the lack of -omics tools limited the identification of this protein. Since then, these impure enzyme preparations
614 have contributed to numerous studies on AG and LAM[39]. It is clear that the availability of enzymes with defined activities will enhance the study of mycobacterial cell wall
616 polymers.

A careful analysis of published screening research points to an important role
618 for Rv3707c in pathogenesis. Rv3707c was identified in a screen for proteins with non-canonical signal sequences, likely because of misannotation of its signal peptide as
620 evidenced by our data[43]. In that study, Perkowski and colleagues found that a

transposon mutant in Rv3707c was severely defective for replication in macrophages.
622 In a separate study, Rv3707c was also found to be important for control of phagosome acidification where it was the second most significantly enriched mutant in an acidified
624 phagosome screen[44]. These data are echoed in a similar set of experiments carried out in the phylogenetically distant *Mycobacterium llatzerense* which infects free-living
626 amoeba[45]. These data point to a role for Rv3707c in phagosome survival, and by extension of D-arabinan remodelling in mycobacterial pathogenesis.

628 The role of Rv1754c remains more elusive. Our biochemical data suggests this enzyme class is active against LAM, but not AG. Furthermore, this genomic locus is a
630 frequent site of IS6110 element insertion, though the gene is broadly conserved amongst mycobacteria[46]. Deletion of this gene could contribute to LAM structural
632 variability amongst circulating strains of *M. tuberculosis*. It is possible that Rv3707c can partially compensate for its loss, or that its contribution to mycobacterial biology
634 is important under highly defined conditions. An analogous situation is observed in other cell wall hydrolytic enzymes. Peptidoglycan-lytic enzymes with seemingly
636 redundant reaction specificities are encoded in most bacteria, without notable phenotypes for their loss under most growth conditions. More recently, however,
638 screens at variable pH have supported specific functions for these enzymes[47].

We have discovered a new enzymatic toolkit for the degradation of
640 mycobacterial arabinogalactan (**Figure 7**). These enzymes will contribute to our understanding of the structure and function of this important polymer. Furthermore,
642 functional annotation of these genes will enable hypothesis-led investigation of the role of D-arabinan structural modulation in mycobacterial biology.

644 **Tables**

Table 1: Activity of GH172 enzymes

Protein	pNP-D-Ara*f*	AG	LAM	Pili
Dg_{GH172a}	✓	✓	✓	✓
Dg_{GH172b}	×	✓	✓	Not complete
Dg_{GH172c}	✓	✓	✓	✓
Noc_{GH172}	✓	✓	✓	✓
Myc_{GH172}	✓	✓	✓	Not complete

646

Table 2: K_{cat}/K_M of GH172 enzymes for pNP-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}

648 **Table 3: SEC-MALS oligomerisation of GH172 enzymes**

Protein	Observed MW (Da)	Observed Oligomerisation 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

650

652

654 **Table 4: Dg_{GH172a} variants and the effect on activity**

Dg_{GH172a} Variant	Activity Relative to WT (%)
WT	100
E233A	Not detectable
D255A	Not detectable
E254A	0.0008

656

Methods

658 **Bacterial strains and growth conditions**

Bacteroidetes sp. were grown on a 2x defined minimal media (**Table S3**) under 660 anaerobic conditions at 37°C over 24 hours to assay growth on various carbon sources, including arabinogalactan. Strains used were *Bacteroides caccae* ATCC 662 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.* 664 *thetaiotaomicron* VPI-5482, *B. vulgatus* ATCC 8482, *B. xylanisolvens* XB1A, *Dysgonomonas gadei* ATCC BAA-286, *D. mossii* DSM 22836, *Parabacteroides* 666 *gordonii* DSM 23371, *P. johnsonii* DSM 18315. *Escherichia coli* and *Pseudomonas aeruginosa* PA7 ATCC 15692 strains were grown in lysogeny broth at 37 °C 668 (unless otherwise specified). *Mycobacterium smegmatis* mc^2155 ATCC 19420 was grown in Tryptic soy Broth at 37 °C with agitation.

670

672

RNA sequencing

674 *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 676 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-678 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 680 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 682 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 684 rRNA-depleted RNA samples were purified using the Zymo RNA Clean & Concentrator kit combined with DNase treatment on a solid support (Zymo Research). 686 cDNA libraries were prepared using the NEBNext Multiplex Small RNA Library Prep kit for Illumina (NEB) in accordance with the manufacturers' instructions.

688 Library preparation and sequencing took place at the Earlham Institute, and were processed by Newcastle University Bioinformatics Support Unit. Briefly, raw 690 sequencing reads were checked using Fast QC, reads were mapped to *Bacteroides cellulosilyticus* DSM 14838 (GCA_000158035) downloaded from Ensembl (assembly 692 ID ASM15803v1). Reads were quantified against genes contained in the Ensembl annotation using featureCounts from the Rsubread package[48]. Counts were 694 normalised by Trimmed Median of Means (TMM) as implemented in DESeq2, and differentially expressed genes determined according to a Negative Binomial model as 696 per DESeq2.

Proteomic analysis of *D. gadei*

698 *Sample preparation*

Dysgonomonas gadei cells were suspended in 5% sodium dodecyl sulfate
700 (SDS) in 50 mM triethylammonium bicarbonate (TEAB) pH 7.5. The samples were subsequently sonicated using an ultrasonic homogenizer (Hielscher) for 1 minute. The
702 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
704 (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
706 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
708 suspension trapping (S-Trap) sample preparation method [49], as recommended by the supplier (ProtiFi™, Huntington NY). Briefly, 2.5 µl of 12% phosphoric acid was added
710 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
712 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-
714 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
716 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
718 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
720 eluates from each sample were combined and dried using a speed-vac before storage at -80 °C.

722 *Mass Spectrometry*

Peptides were dissolved in 2% acetonitrile containing 0.1% trifluoroacetic acid, and each sample was independently analysed on an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Fisher Scientific), connected to an UltiMate 3000 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 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% 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 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 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 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 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 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 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 auxiliary gas flow, was used.

All mass spectra were analysed using MaxQuant 1.6.12.0[50], and searched against the *Dysgonomonas gadei* ATCC BAA-286 proteome database downloaded from Uniprot (accessed 09.01.2020). Peak list generation was performed within

748 MaxQuant and searches performed using default parameters and the built-in Andromeda search engine[51]. The enzyme specificity was set to consider fully tryptic
750 peptides, and two missed cleavages were allowed. Oxidation of methionine, N-terminal acetylation and deamidation of asparagine and glutamine were allowed as
752 variable modifications. Carbamidomethylation of cysteine was allowed as a fixed modification. A protein and peptide false discovery rate (FDR) of less than 1% was
754 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
756 of parsimony. Reverse hits, contaminants, and proteins only identified by site modifications were removed before downstream analysis.

758 **Generation of expression constructs**

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

764 **Protein Expression and Purification**

766 *Expression and purification of* $Dg_{GH4185a}$, $Dg_{GH4185b}$, $Phage_{GH4185}$, $Myxo_{GH4185}$, Dg_{GH172a}, Dg_{GH172b}, Dg_{GH172c}, *and* Myc_{GH172}
768

Recombinant proteins were expressed in competent *E. coli* Tuner cells
770 (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
772 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.

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

782 *Purification of Rv3707c and MSMEG_2107*

For protein Rv3707c and MSMEG_2107 expression, an aliquot of competent
784 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
786 inoculate 1 L of modified Studier's autoinduction media[52]. The bacteria were incubated at 37 °C with shaking (180 RPM) until OD_{600}= ~0.6, whereupon the flasks were cooled
788 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
790 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
792 preparation.

794 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
796 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
798 lysed by three passages through a French pressure cell. Insoluble debris was then pelleted by centrifugation at 40,000 *xg* for 40 minutes (4 ºC). The supernatant was
800 then processed by immobilised metal affinity chromatography (IMAC) on a gravity

column containing 2 mL bed volume of cOmplete His Tag purification resin (Roche).
802 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
804 were pooled and dialysed exhaustively against three litres of dialysis buffer (25 mM HEPES pH8; 400 mM NaCl; 5% glycerol; 50 mM L-arginine; 50 mM L-glutamic acid; 2
806 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
808 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
810 concentrated. The protein was always used freshly prepared.

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

One plate of BL21-DE3 transformed with an appropriate plasmid was used to
814 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
816 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 1. Subsequent
818 purification steps were identical to those in Rv3707c, but in the above, simpler buffer, omitting lysozyme and DNAse I.

820

Crystallography

822 *Crystallography of Dg_{GH172c}*

Crystallisation of Dg_{GH172c} sample at 10 mg ml^{-1} was screened using
824 commercial kits (Molecular Dimensions and Hampton Research) with vapour diffusion sitting drop method. Crystals formed in a buffer containing 0.1 M MOPS/Sodium
826 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

828 at the synchrotron beamline I24t Diamond light source (Didcot, UK) at a temperature of 100 K. The data were integrated using XDS[53] via XIA2 [54] and scaled with Aimless[55].
830 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
832 *Bacteroides uniformis*[56]. Subsequently, the structure was auto built with CCP4build on CCP4cloud[57]. The model was improved by rounds of manual building with COOT
834 and refinement with Refmac[58,59]. The final model was validated with Molprobity[60] and COOT[58].

836

Crystallography of Rv3707c

838 Rv3707c was concentrated to 7 mg ml^{-1} and loaded into a sitting drop crystallisation tray (Clover) using JCSG-plus crystallisation screen (Molecular
840 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
842 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
844 Diamond light source, Oxford. Data were auto-processed using Xia2 and general file manipulations were performed using the CCP4 suite of programmes[54,57]. Data were
846 phased by molecular replacement using the Alphafold structural prediction of Rv3707c as a search model using PHASER (TFZ of 21.1)[56]. The structure was then auto built
848 in PHENIX with rounds of refinement carried out by PHENIX-refine and manually in COOT[58,61].

850 **Summary of methods for phylogenetic analyses**

Two alignments were used to infer respectively a global phylogeny for members
852 of the PF13810 family and a more restricted phylogeny focusing on close relatives to the functionally characterised proteins. The global alignment was derived from the

854 “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
856 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)
858 and (ii) adding the three sequences from *Mycobacterium abscessus* (strain 4529 available at the Integrated Microbial Genomes & Microbiomes (IMG/M) database:
860 2635695100/Ga0069448_11324, 2635694794/Ga0069448_1118, 2635698039/Ga0069448_113269) and (iii) deleting sites made of a majority of indels.
862 For the restricted alignment 39 complete sequences were aligned, including the seven proteins investigated in this study - enzymatic characterisation and one structure. The
864 sequences were aligned with Clustal Omega using default settings in SEAVIEW v.4.6.4 [62,63]. Following minor manual adjustments of the alignment the mask function
866 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
868 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
870 phylogenies using automatic model selection[64]. The selected models were LG+F+I+G4 for the global alignment and WAG+I+G4 for the restricted alignment using
872 the Bayesian Information Criterion (BIC). Bootstraps (100 replicates) were computed to assess branch reliability. iTOL (interactive tree of life) was used to generate the
874 figures[65]. The global phylogeny was annotated with taxonomy information derived from the UniProt database (https://www.uniprot.org/)[66].

876 **Protein structure prediction by AlphaFold 2.0**

Amino acid sequences were submitted to the Colabfold_advanced.ipynb server
878 for protein structure prediction[67]. Unless otherwise stated the default values were

used, with the max_recycles set to 12. Figure colouring for pLDDT values were
880 generated in ChimeraX 1.4 using the protein structure prediction tool and a custom palette.

882

Timepoint assays

884 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
886 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
888 subsequently boiled to ensure enzymatic cessation. Time point samples were then analysed using TLC or IC-PAD.

890

Porous graphitic carbon chromatography clean-up of Rv3707c and
892 **MSMEG_2107 arabinogalactan hydrolysis assays**

Due to the high concentration of L-arginine and L-glutamic acid in the buffer
894 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
896 timepoint, reaction mixtures were loaded onto a Hypersep Hypercarb SPE cartridge (Thermo Scientific) which had been washed with acetonitrile and 50% THF in water
898 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
900 dried by evaporation in a SpeedVac concentrator before being reconstituted in the original volume of water.

902 **Kinetic analysis of GH172 arabinofuranosidase activity**

904 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

906 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
908 (see below) with reference to a standard curve.

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

Purified proteins were incubated with at 0.1-5 μM (as indicated) with substrates
912 (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).
914 Using TLC plate aluminium 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,
916 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
918 visualisation 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.

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

Oligosaccharides from enzymatic polysaccharide digestion were analysed
922 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
924 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
926 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%
928 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
930 commercially (Megazyme) were used as standards at a concentration of 25 µM. Data

were processed using Chromeleon™ Chromatography Management System V.6.8.
932 Final graphs were created using GraphPad Prism 8.0.1.

Purification of mycobacterial arabinogalactan

934 Large scale purification of mycobacterial arabinogalactan was achieved by established methodologies[68]. In brief, 8 L of mycobacterial culture was grown to mid-
936 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
938 boiled in a final concentration of 1% sodium dodecyl sulphate (SDS) and refluxed. Insoluble material (containing mycolyl-arabinogalactan-peptidoglycan complex) was
940 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.
942 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
944 by treatment with H_2SO_4 at 95 °C before being neutralised with sodum carbonate. The resultant solubilised arabinogalactan was collected in the supernatant, dialysed
946 exhaustively against water and lyophilised (yield= ~22.5 $mg \cdot L^{-1}$).

948 **Purification of mycobacterial lipoglycans**

One litre of mycobacterial culture was grown to mid exponential phase and
950 pelleted as previously. The pellet was resuspended in 20 ml PBS, 0.1% Tween-80, chilled and lysed by bead-beating. Lysate was transferred to a Teflon-capped glass
952 tube and vortexed with an equal volume of citrate buffer saturated with phenol (Sigma-Aldrich), and heated to 80 °C for 3 h, vortexing every hour. A biphase was generated
954 by centrifugation at 2000 RPM at 10 °C, and the upper aqueous phase transferred to a fresh glass tube and hot phenol wash repeated twice more. The resultant protein
956 free glycan mixture was dialysed exhaustively against tap water overnight to remove

trace phenol and lyophilised, yielding 34 mg of crude lipoglycans (LAM, LM, PIMS) per
958 litre of culture.

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

Pilins were purified as described by Burrows and colleagues, with some
962 modifications[40]. Briefly, *Pseudomonas aeruginosa* PA7 were streaked out in a grid pattern onto LB agar plates and grown for 24 hours at 37 °C. Cells were then scraped
964 from all plates using sterile cell scrapers and resuspended in 4 ml of sterile phosphate-buffered saline (pH 7.4) per plate.

966 Pili were then sheared from the cell wall by vigorous vortexing of resuspended cells for 2 min. This suspension was centrifuged for 5 min at 6000 *x g*. The supernatant
968 was centrifuged for 20 min at 20,000 *x g*. Supernatants were transferred to fresh tubes and $MgCl_2$ was added to give a final concentration of 100 mM. Following inversion of
970 these tubes to ensure mixing, samples were incubated at 4 °C overnight, allowing precipitation of sheared proteins. Samples were then centrifuged for 20 minutes at
972 20,000 *x g*, yielding a precipitate smudge which was resuspended in 50 mM NH_4HCO_3, pH 8.5 and transferred to fresh Eppendorf tubes. This solution was then dialysed into
974 the same buffer to eliminate excess $MgCl_2$.

Bradford assays were then performed to assay the mass of protein in the
976 sample, followed by digestion of the intact protein pilins using proteinase K in a 50:1 pilin to enzyme ratio by mass for 24 hours in the presence of 2 mM $CaCl_2$. Glycans
978 were then purified from digested proteins using porous graphitised carbon chromatography, where sugars were eluted from the column using a twofold
980 increasing concentration series of a butan-1-ol:H_2O gradient from 1:32 to 1:1 using 1 mL elutions[69]. Thin-layer chromatography (TLC) of eluates showed various oligomers

982 of arabinan present in all fractions, all of which were subsequently used as substrates for potential arabinanases.

984

Synthesis of pNP-α-D-arabinofuranoside

986 Para-nitrophenol (pNP)-α/β-D-arabinofuranoside synthesis was achieved following the established procedures [22].

988

SEC-MALS

990 Molecular weights were determined by size exclusion chromatography coupled light scattering using an Agilent MDS system with either an Agilent BioSEC 5 1000 Å, 992 4.6 x 300 mm, 5 µm or GE Superdex 200 5 15 mm columns at appropriate flow rates. Detector offsets were calibrated using a BSA standard and concentrations were 994 determined by refractive index.

996 **Protein sequence alignment and structure visualisation**

Sequence alignments were performed using Clustal Omega and visualized 998 using ESPript[62,70]. Figures of protein structures were generated with ChimeraX 1.4[71].

Fluorescent-conjugated mAGP hydrolysis assay

1000 Fluorescently labeled mycolyl-arabinogalactan-peptidoglycan complex (mAGP) was isolated from *Corynebacterium glutamicum ATCC13032* following 1002 previously reported methods[34-36]. In brief, cells were grown in the presence of 5-AzFPA to saturation, reacted with DBCO-conjugated AF647 and then the mAGP was 1004 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 1006 centrifugation at 15,000 *xg* 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).

Author Contributions:

Omar Al-Jourani – Methodology, validation, formal analysis, investigation, data curation, visualisation. **Samuel Benedict** – Methodology, validation, formal analysis, investigation, data curation, visualisation, 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, visualisation, writing – review and editing. **Nicholas P Bailey** – Methodology, validation, formal analysis, visualisation. **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** – Conceptualisation, methodology, validation, formal analysis, data curation, writing – original draft preparation, writing – review and editing, visualisation, supervision, funding acquisition. **Patrick J. Moynihan** – Conceptualisation, methodology, validation, formal analysis, data curation, writing – original draft preparation, writing – review and editing, visualisation, supervision, funding acquisition.

Competing Interests

The authors have no conflicts of interest to declare.

References

1070 1 Abrahams, K. A. & Besra, G. S. Synthesis and recycling of the mycobacterial cell envelope. *Curr Opin Microbiol* **60**, 58-65 (2021).
1072 https://doi.org:10.1016/j.mib.2021.01.012

2 Dulberger, C. L., Rubin, E. J. & Boutte, C. C. The mycobacterial cell envelope
1074 - a moving target. *Nat Rev Microbiol* **18**, 47-59 (2020). https://doi.org:10.1038/s41579-019-0273-7

1076 3 Daffe, M., McNeil, M. & Brennan, P. J. Major structural features of the cell wall arabinogalactans of Mycobacterium, Rhodococcus, and Nocardia spp.
1078 *Carbohydr Res* **249**, 383-398 (1993). https://doi.org:10.1016/0008-6215(93)84102-c

1080 4 Justen, A. M. *et al.* Polysaccharide length affects mycobacterial cell shape and antibiotic susceptibility. *Sci Adv* **6** (2020).
1082 https://doi.org:10.1126/sciadv.aba4015

5 Safi, H., Sayers, B., Hazbon, M. H. & Alland, D. Transfer of embB codon 306
1084 mutations into clinical Mycobacterium tuberculosis strains alters susceptibility to ethambutol, isoniazid, and rifampin. *Antimicrob Agents Chemother* **52**, 2027-
1086 2034 (2008). https://doi.org:10.1128/AAC.01486-07

6 Sun, Q. *et al.* Mutations within embCAB Are Associated with Variable Level of
1088 Ethambutol Resistance in Mycobacterium tuberculosis Isolates from China. *Antimicrob Agents Chemother* **62** (2018). https://doi.org:10.1128/AAC.01279-
1090 17

7 Zhao, L. L. *et al.* Analysis of embCAB mutations associated with ethambutol
1092 resistance in multidrug-resistant mycobacterium tuberculosis isolates from China. *Antimicrob Agents Chemother* **59**, 2045-2050 (2015).
1094 https://doi.org:10.1128/AAC.04933-14

8 Brossier, F. *et al.* Molecular Analysis of the embCAB Locus and embR Gene
1096 Involved in Ethambutol Resistance in Clinical Isolates of Mycobacterium tuberculosis in France. *Antimicrob Agents Chemother* **59**, 4800-4808 (2015).
1098 https://doi.org:10.1128/AAC.00150-15

9 Zhang, L. *et al.* Structures of cell wall arabinosyltransferases with the anti-
1100 tuberculosis drug ethambutol. *Science* **368**, 1211-1219 (2020). https://doi.org:10.1126/science.aba9102

1102 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).
1104 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
1106 (1957). https://doi.org:10.1126/science.125.3238.99

12 Healy, C., Gouzy, A. & Ehrt, S. Peptidoglycan Hydrolases RipA and Ami1 Are
1108 Critical for Replication and Persistence of Mycobacterium tuberculosis in the Host. *mBio* **11** (2020). https://doi.org:10.1128/mBio.03315-19

1110 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-
1112 10586-2

14 Kalscheuer, R., Weinrick, B., Veeraraghavan, U., Besra, G. S. & Jacobs, W. R.,
1114 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 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

17 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

18 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).

19 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

20 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

21 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

22 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

23 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

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 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. *Proc Natl Acad Sci U S A* **119** (2022). https://doi.org:10.1073/pnas.2120620119

31 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

32 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

33 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

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

35 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

36 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

37 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

38 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

39 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

40 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

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

42 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

43 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

44 Stewart, G. R., Patel, J., Robertson, B. D., Rae, A. & Young, D. B. Mycobacterial mutants with defective control of phagosomal acidification. *PLoS Pathog* **1**, 269-278 (2005). https://doi.org:10.1371/journal.ppat.0010033

45 Delafont, V. *et al.* Mycobacterium llatzerense, a waterborne Mycobacterium, that resists phagocytosis by Acanthamoeba castellanii. *Sci Rep* **7**, 46270 (2017). https://doi.org:10.1038/srep46270

46 Antoine, R., Gaudin, C. & Hartkoorn, R. C. Intragenic Distribution of IS6110 in Clinical Mycobacterium tuberculosis Strains: Bioinformatic Evidence for Gene Disruption Leading to Underdiagnosed Antibiotic Resistance. *Microbiol Spectr* **9**, e0001921 (2021). https://doi.org:10.1128/Spectrum.00019-21

47 Mueller, E. A., Egan, A. J., Breukink, E., Vollmer, W. & Levin, P. A. Plasticity of Escherichia coli cell wall metabolism promotes fitness and antibiotic resistance across environmental conditions. *Elife* **8** (2019). https://doi.org:10.7554/eLife.40754

48 Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res* **47**, e47 (2019). https://doi.org:10.1093/nar/gkz114

49 Zougman, A., Selby, P. J. & Banks, R. E. Suspension trapping (STrap) sample preparation method for bottom-up proteomics analysis. *Proteomics* **14**, 1006-1000 (2014). https://doi.org:10.1002/pmic.201300553

50 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367-1372 (2008). https://doi.org:10.1038/nbt.1511

51 Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10**, 1794-1805 (2011). https://doi.org:10.1021/pr101065j

52 Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expr Purif* **41**, 207-234 (2005). https://doi.org:10.1016/j.pep.2005.01.016

53 Sauter, N. K. & Zwart, P. H. Autoindexing the diffraction patterns from crystals with a pseudotranslation. *Acta Crystallogr D Biol Crystallogr* **65**, 553-559 (2009). https://doi.org:10.1107/S0907444909010725

54 Winter, G., Lobley, C. M. & Prince, S. M. Decision making in xia2. *Acta Crystallogr D Biol Crystallogr* **69**, 1260-1273 (2013). https://doi.org:10.1107/S0907444913015308

55 Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr D Biol Crystallogr* **69**, 1204-1214 (2013). https://doi.org:10.1107/S0907444913000061

56 McCoy, A. J. *et al.* Phaser crystallographic software. *J Appl Crystallogr* **40**, 658-674 (2007). https://doi.org:10.1107/S0021889807021206

57 Krissinel, E., Uski, V., Lebedev, A., Winn, M. & Ballard, C. Distributed computing for macromolecular crystallography. *Acta Crystallogr D Struct Biol* **74**, 143-151 (2018). https://doi.org:10.1107/S2059798317014565

58 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

59 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

60 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

61 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

62 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

63 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

64 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

65 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

66 UniProt, C. UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204-212 (2015). https://doi.org:10.1093/nar/gku989

67 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

68 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

69 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

70 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

71 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