1   The *Ruminococcus bromii* amylosome protein Sas6 binds single and double helical α-glucan

2   structures in starch.

3

4   Amanda L. Photenhauer[1], Filipe M. Cerqueira[1], Rosendo Villafuerte-Vega[2], Krista M.

5   Armbruster[1], Filip Mareček[3], Tiantian Chen[4], Zdzislaw Wawrzak[5], Jesse B. Hopkins[6], Craig W.

6   Vander Kooi[4], Štefan Janeček[3], Brandon T. Ruotolo[2], Nicole M. Koropatkin[1*]

7

8   [1] Department of Microbiology & Immunology, University of Michigan Medical School, Ann Arbor,

9   MI 48109.

10   [2] Department of Chemistry, University of Michigan, Ann Arbor, MI 48109.

11   [3] Laboratory of Protein Evolution, Institute of Molecular Biology, Slovak Academy of Sciences,

12   Bratislava, Slovakia

13   [4] Department of Biochemistry and Molecular Biology, University of Florida, Gainesville, FL

14   [5] Northwestern Synchrotron Research Center-LS-CAT, Northwestern University, Argonne, Illinois

15   60439.

16   [6] Biophysics Collaborative Access Team, Illinois Institute of Technology, Advanced Photon

17   Source, Argonne National Laboratory, Lemont, Illinois, USA.

18   *To whom correspondence should be addressed: nkoropat@umich.edu

19

20   Running title: Structure of Ruminococcus bromii Sas6

21   Keywords: starch, CBM74, CBM26, Ruminococcus bromii, microbiome

22

Abstract

Resistant starch is a prebiotic with breakdown by gut bacteria requiring the action of specialized amylases and starch-binding proteins. The human gut symbiont *Ruminococcus bromii* expresses granular starch-binding protein Sas6 (Starch Adherence System member 6) that consists of two starch-specific carbohydrate binding modules from family 26 (RbCBM26) and family 74 (RbCBM74). Here we present the crystal structures of Sas6 and *Rb*CBM74 with a double helical dimer of maltodecaose bound along an extended surface groove. Binding data combined with native mass spectrometry suggest that RbCBM26 binds short maltooligosaccharides while RbCBM74 can bind single and double helical α-glucans. Our results support a model by which RbCBM74 and RbCBM26 bind neighboring α-glucan chains at the granule surface. CBM74s are conserved among starch granule-degrading bacteria and our work provides molecular insight into how this structure is accommodated by select gut species.

Introduction

The gut microbiota, the consortium of microbes that resides in the human gastrointestinal tract, influences many aspects of host physiology including digestive health [1]. The composition of the gut microbiota is modulated by the human diet [2-4]. After host nutrient absorption in the small intestine, indigestible dietary fiber transits the large intestine and becomes food for gut microbes [3]. Bacterial fermentation of dietary carbohydrates produces beneficial short chain fatty acids including butyrate, a primary carbon source for colonocytes that also has systemic anti-inflammatory and anti-tumorigenic properties [3, 5].

Resistant starch is a prebiotic fiber that tends to increase butyrate in the large intestine [6]. Starch is a glucose polymer composed of branched, soluble amylopectin and coiled insoluble amylose [7, 8]. Breakdown of starch starts with human salivary and pancreatic amylases which release maltooligosaccharides for absorption in the small intestine [9]. However, a portion of starch is indigestible by human amylases and is termed resistant starch (RS) [9]. Raw, uncooked starch granules are resistant to digestion in the upper gastrointestinal tract due to the tight packing of constituent amylose and amylopectin into semi-crystalline, insoluble granules [7]. This type of resistant starch, called RS2, becomes food for gut bacteria that can adhere to and deconstruct granules, releasing glucose and maltooligosaccharides that cross-feed other organisms [9].

Human gut bacteria that degrade RS2 *in vitro* include *Bifidobacterium adolescentis* and *Ruminococcus bromii* [10-14]. *R. bromii* is a Gram-positive anaerobe that increases in relative abundance in the gut upon host consumption of resistant potato or corn starch [10, 15, 16]. *R. bromii* is a keystone species for RS2 degradation because it cross-feeds butyrate-producing bacteria [10]. *R. bromii* synthesizes multi-protein starch-degrading complexes called amylosomes via protein-protein interactions between dockerin and complementary cohesin domains [17-19]. As many as 32 *R. bromii* proteins have predicted cohesin or dockerin domains including amylases, pullulanases, starch-binding proteins, and proteins of unknown function [17, 20]. Many

60   have carbohydrate-binding modules (CBMs) that presumably aid in binding starch and tether the

61   bacteria to its food source [21].

62   　CBMs are classified by amino acid sequence into numbered families and include members

63   that bind only soluble starch and some that also bind granular starch [21, 22]. One such family is

64   CBM74 which was discovered as a discrete domain (*Ma*CBM74) of a multimodular amylase from

65   the potato starch-degrading bacterium, *Microbacterium aurum* [22]. *Ma*CBM74 binds amylose

66   and amylopectin as well as raw wheat, corn, and potato starch granules [22]. The CBM74 family

67   is unique as it is ~300 amino acids, two to three times larger than most starch-binding CBMs [21].

68   CBM74 domains are typically found in multimodular enzymes that include a glycoside hydrolase

69   family 13 (GH13) domain for hydrolyzing starch and are flanked by a starch-binding CBM from

70   family 25 or 26 (CBM25 or CBM26) [21, 22]. Most CBM74 family members are encoded by gut

71   microbes and 70% are found in Bifidobacteria [22]. The genomes of *R. bromii* and *B. adolescentis*

72   each encode one putative CBM74-containing protein. The prevalence of CBM74 domains

73   encoded within the genomes of RS2-degrading bacteria, and its increased representation in

74   metagenomic and metatranscriptomic analyses from host diet studies, suggest a role for this

75   module in RS2 recognition in the distal gut [23-25].

76   　The *R. bromii* starch adherence system protein 6 (Sas6) is a secreted protein of 734 amino

77   acids that contains both a CBM26 and CBM74 followed by a C-terminal dockerin type 1 domain

78   [26, 27]. Here we present the biochemical characterization and crystal structure of Sas6, providing

79   the first view of the CBM74 domain and its juxtaposition with the CBM26 domain. The co-crystal

80   structure of *Rb*CBM74 with a double helical dimer of maltodecaose, which mimics the architecture

81   of double helical amylopectin in starch granules, revealed recognition via an elongated groove

82   spanning the domain. *Rb*CBM74 exclusively binds longer maltooligosaccharides (≥ 8 glucose

83   units), and native mass spectrometry suggests that both single and double helical α-glucans are

84   recognized, providing flexible recognition of amylose and amylopectin. Our biochemical data

4

85   demonstrate that CBM26 and CBM74 recognize different α-glucan moieties within starch granules

86   leading to overall enhanced granule binding.

87

88   **Results**

89   *Modular Architecture of Sas6* – Sas6 consists of five discrete domains: an N-terminal

90   CBM26 (*Rb*CBM26), a CBM74 domain (*Rb*CBM74) flanked by Bacterial Immunoglobulin-like

91   (BIg) domains, and a C-terminal dockerin type I (**Fig. 1A**) [27]. Sas6 is encoded at the

92   WP_015523730 locus (formerly RBR_14490 or Doc6, UniProt: A0A2N0UYM2) and includes a

93   Gram-positive signal peptide (residues 1-30) that presumably targets the protein for secretion.

94   *Rb*CBM74 spans residues 242-572 based on an alignment with annotated CBM74 domains [22].

95   We used InterProScan to annotate the remaining sequence which added the Bacterial

96   Immunoglobulin-like (BIg, Pfam 02368) domain A (BIgA), but did not predict BIgB, which we

97   identified via structure determination [28].

98

99   *Sas6 Cell Localization* – Though Sas6 has a signal peptide it is unknown whether it is a

100   constituent of a cell-bound amylosome, or part of a freely secreted complex [20]. *R. bromii*

101   synthesizes five scaffoldin (Sca) proteins that have cohesins for amylosome assembly; Sca2 and

102   Sca5 are cell-bound and Sca1, Sca3, and Sca4 are freely secreted [20]. The cognate cohesin for

103   the Sas6 dockerin is unknown. Sas6 is detected in the cell-free supernatant of *R. bromii* cultures

104   in stationary phase but also elutes from the surface of exponentially growing cells with EDTA

105   which disrupts the calcium-dependent cohesin-dockerin interaction [17, 29]. To determine the

106   localization of Sas6, we grew cells to mid-log phase on potato amylopectin and performed a

107   Western Blot with custom antibodies against recombinant Sas6 (**Fig. 1B**). Sas6 was detected in

108   the cell fraction and not the cell-free culture supernatant (**Fig. 1B**), and was visualized on the cell

109   surface via immunofluorescence (**Fig 1C**). Therefore, we conclude that Sas6 is a component of

110   a cell-surface amylosome in actively growing cells. It is possible that Sas6 localization is

111   dependent upon growth phase, as are cellulosome components in some organisms, explaining

112   its previous detection in culture supernatant [17]. Alternatively, *R. bromii*, like some cellulosome-

113   producing bacteria, may release cell-surface amylosomes in stationary phase [30].

114

115   *Sas6 Starch Binding* –CBM26 and CBM74 are putative raw starch-binding families [22,

116   31]. Plant sources of granular starch differ greatly in granule organization, including crystallinity

117   (e.g., packing of the long helical chains), length of α1,4-linked chains, amylose location and

118   organization, water content, and trace elements [7]. We used a truncated construct of Sas6

119   (residues 31-665) lacking the C-terminal dockerin domain, herein called Sas6T, to test Sas6

120   binding to starch polysaccharides. Sas6T binds potato, corn, and wheat starch granules, with the

121   highest fraction of protein bound to corn starch, and no non-specific binding to Avicel (crystalline

122   cellulose) (**Fig. 1D**). Of note, corn starch has a smaller granule size and therefore a larger surface

123   area to mass ratio [8]. We tested Sas6T binding to amylopectin and amylose, as well as glycogen

124   and pullulan via affinity PAGE. Glycogen is similar to amylopectin with more frequent α1,6

125   branching (every 6-15 residues for liver glycogen compared to 15-25 residues for amylopectin)

126   [32, 33]. Pullulan is a fungal α-glucan composed of repeating α1,6-linked maltotriose units [34].

127   Sas6T binds amylose, amylopectin (potato and corn), and glycogen but has less affinity for

128   pullulan suggesting a preference for longer α1,4-linked regions within the polysaccharide (**Fig.

129   1E**). Sas6T does not bind dextran, a bacterially derived exopolysaccharide of α1,6-linked glucose

130   [35], demonstrating its specificity for starch.

131

132   *Structure of Sas6* – The structure of Sas6T with α-cyclodextrin (ACX), was determined via

133   single-wavelength anomalous dispersion of intrinsic sulfur-containing residues to a resolution of

134   1.6Å (R_{work}=16.8%, R_{free}=21.2%) (**Table 1**). The final model contained two molecules of Sas6T in

135   the asymmetric unit, with four Ca^{2+} per chain and one molecule of ACX bound at the *Rb*CBM26

136   domain. The Sas6T structure determined with ACX was used to phase a dataset from unliganded

6

137   crystals (2.2Å, R_{work}=19.7%, R_{free}=25.5%) (**Table 1**). The overall crystal structure of Sas6T is

138   compact, with RbCBM26, BIgA and BIgB forming an arc over RbCBM74 (**Fig. 2A**).

139   RbCBM26, RbCBM74, and the dockerin domain are separated by BIgA (light grey) and

140   BIgB (dark grey), respectively (**Fig. 2A, Extended Data Fig 1A**). Ig-like or fibronectin-III domains

141   act as spacers in multi-modular glycoside hydrolases including GH13s that target starch [36].

142   BIgA and BIgB interact via hydrogen bonding with 354Å of buried surface area [37] (**Extended**

143   **Data Fig 1B**). This interaction may help stabilize or orient the CBM74 domain or the BIgs may act

144   as a hinge between the CBMs. The two chains in the asymmetric unit exhibit some flexibility

145   resulting in different positioning between the RbCBM26 binding site and the RbCBM74 domain

146   (**Fig. 2B**).

147

148   *Small Angle X-Ray Scattering* – To better connect how our crystal structures correlate with

149   conformational flexibility in solution, we used size-exclusion chromatography coupled with small

150   angle x-ray scattering (SEC-SAXS) on Sas6T (**Table 1**). The elution separated out several peaks,

151   including a single strong peak for that was well separated and monodisperse as indicated by the

152   constant radius of gyration (R_g) across the eluted peak (**Extended Data Fig 2A**). The Guinier fit

153   of a subtracted scattering profile created from that peak gave R_g and I(0) values of 29.44 ± 0.04Å

154   and 0.04 ± 3.65 x 10^{-5} and the fit and normalized fit residuals confirmed this peak was

155   monodisperse (**Extended Data Fig 2B**). The molecular weight of Sas6T from the SAXS data was

156   calculated to be 61.0 kDa (theoretical 68.9 kDa) indicating it is primarily monomeric in solution

157   [38]. The D_{max} from the P(r) function for Sas6T is 90Å. The overall shape of the P(r) function for

158   Sas6T, calculated by indirect Fourier transform (IFT) using GNOM, has a relatively Gaussian

159   shape that is characteristic of a globular compact particle with the main peak at r = ~30 Å

160   (**Extended Data Fig 2C**) [39]. There is a small peak at r = 55Å which suggests there are two

161   structurally separate motifs, possibly RbCBM26 and RbCBM74. The dimensionless Kratky plot

162   maxima for Sas6T are typical for a rigid globular protein (**Extended Data Fig 2D**). The small

7

163   plateau in the mid to high q region, around $qR_g = 5$ in the dimensionless Kratky plot indicates

164   some extension or disorder in the system. These results suggest the presence of two separate

165   modules with flexibility between them, likely corresponding to the two CBMs.

166   We tested whether the crystal structure matched the solution data by fitting the crystal

167   structure to the SAXS data using FoXS [40]. The fit had a $\chi^2 = 2.46$ and showed systematic

168   deviations in the normalized fit residual (**Extended Data Fig 2E**). This highlights that there are

169   significant differences between the lowest energy conformation of Sas6T in the crystal structure

170   and the structure of Sas6T in solution. We then used MultiFoXS with our high-resolution structure

171   of Sas6T to account for the flexibility, assigning the linkers between the domains (residues 130-

172   137 and 572-583) as flexible [40]. MultiFoXS gave a best fit with a 1-state solution with a $\chi^2 =$

173   0.96 and calculated R_g of 29.2Å which corroborates the Guinier R_g calculation (**Extended Data**

174   **Fig 2F**). An alignment of Chain A of the crystal structure and MultiFoXS model had a RMSD of

175   1.2Å over 347 pruned atom pairs (**Fig. 2C**). The MultiFoXS model shows a slightly more extended

176   model for Sas6T in comparison to the crystal structure demonstrating that Sas6T has some

177   flexibility in solution yet remains compact.

178

179   *Structure of* Rb*CBM74 –* *Rb*CBM74 (357 residues) has 21 β-strands and 13 short α-

180   helices with a core β-sandwich fold of two sheets with five antiparallel β-strands (**Fig. 2D,**

181   **Extended Data Fig 3A**). A third short β-sheet forms a convex face and two pairs of β-strands

182   (residues 356-369 and 412-423) protrude from the region between the β-sandwich and the third

183   β-sheet. In this structure, two short β-strands lie at the entrance and exit of the CBM74 domain,

184   marking the domain boundaries (**Extended Data Fig 3B**).

185   A DALI search revealed that the central fold of *Rb*CBM74 most closely resembles CBM9

186   from *Thermotoga maritima* Xylanase10A (PDB ID: 1I82-A, Z-score: 9.8, RMSD: 3.2Å, identity:

187   17%) [41, 42] (**Extended Data Fig 3C**). *Tm*CBM9 binds glucose, cellobiose, cello- and xylo-

188   oligomers at the reducing ends, and amorphous and crystalline cellulose [42]. *Tm*CBM9 (189

8

189   residues) is larger than most CBMs which range from 80-120 amino acids [42]. Despite the

190   conserved core β-sandwich, *Rb*CBM74 displays several extra loops and β-strands. The ligand

191   binding site of *Tm*CBM9 is formed by two Trp residues that create an aromatic clamp around

192   cellobiose. *Rb*CBM74 W373 is conserved with one of these Trps and lies within an extended,

193   shallow channel partially covered by residues 374-384 that form a flexible loop only resolved in

194   one monomer (**Extended Data Fig 3D**).

195   There are three putative structural Ca^{2+} in the *Tm*CBM9 structure and four cations in

196   RbCBM74, one of which aligns with a Ca^{2+} in *Tm*CBM9 (**Extended Data Fig 3E**). We modeled

197   these cations as Ca^{2+} based upon coordination geometry and atomic distances (**Extended Data**

198   **Fig 3F**) [43, 44]. Ca^{2+}-1 and Ca^{2+}-2 are separated by 3.8Å and share three coordinating residues

199   but only Ca^{2+}-2 is surface exposed. Ca^{2+}-3 is abutted by the loop connecting β-strands 2 and 3

200   and Ca^{2+}-4 is at the center of a loop formed by residues 256-264 and conserved with *Tm*CBM9.

201   Like *Tm*CBM9, the Ca^{2+} ions in the *Rb*CBM74 structure may be important for structural stability

202   [45].

203

204   *Molecular Basis of RbCBM26 Binding* – The N-terminal *Rb*CBM26 displays a β-sandwich

205   consistent with other members of the CBM26 family [21]. In both chains of the asymmetric unit,

206   CH/π stacking with ACX is provided by W63 and Y55 with hydrogen bonding mediated by Y53,

207   K101, Q103, and the peptidic oxygen of A107 **(Fig. 2E)**. In chain A only, K97 provides hydrogen

208   bonding with O3 of Glc6. In chain B, ACX lies 3.2Å from S286 of the CBM74 domain and hydrogen

209   bonds with O2 and O3 of Glc3. In contrast, S286 is 9.5Å from ACX in chain A. The top structural

210   homologs of *Rb*CBM26 from DALI are the CBM25 from *Bacillus halodurans* C-125 (*Bh*CBM26)

211   from α-amylase G-6 (PDB ID: 2C3V-A, Z-score: 12.4, RMSD 1.9Å, identity: 16%) and CBM26

212   (*Bh*CBM26) from the same enzyme (PDB ID: 6B3P-B, Z-score: 12.1, RMSD 1.9Å, identity: 20%)

213   [41, 46]. Another top DALI result is *Er*CBM26b of Amy13K from *Eubacterium rectale* (PDB ID

214   2C3H-B, Z-score: 10.8, RMSD 1.7Å, identity: 19%). In all three CBM26 structures, the structure

9

215 and aromatic platforms for ligand recognition are conserved (**Extended Data Fig 4AB)**.

216 *Rb*CBM26, in contrast to *Er*CBM26 and *Bh*CBM26, has a longer loop containing K97 and K101

217 that provide additional hydrogen bonding with ACX. Unlike *Bh*CBM26, RbCBM26 does not

218 undergo a conformational change upon ligand binding **(Extended Data Fig 4C)** [31]. A sequence

219 alignment with CBM26 members *Bh*CBM26, *Er*CBM26 and the *Lactobacillus amylovorus* α–

220 amylase CBM26 (*La*CBM26), demonstrates conservation of the aromatic platform but more

221 variation in the hydrogen-bonding network (**Extended Data Fig 4A**). Sas6 W63 corresponds to

222 *La*CBM26 W32 that, when mutated, results in complete loss of binding [47]. The *R. bromii* protein

223 Sas20 has a CBM26-like domain that shares 26% sequence identity with *Rb*CBM26, yet

224 *Rb*CBM26 shares more structural similarity with *Bh*CBM26 and *Er*CBM26 [29].

225

226   *Binding Mechanism of Sas6* – We expressed the individual Sas6 CBMs and included the

227 BIgA/B domains with the CBM74 (BIg-*Rb*CBM74-BIg, residues 134-665) to enhance solubility.

228 Sas6T and BIg-*Rb*CBM74-BIg bound granular corn and potato starch, but *Rb*CBM26 did not bind

229 either insoluble starch at detectable levels **(Fig. 2F).** Sas6T binds to more of the corn starch

230 granule, (K_d = 2.8μM ± 0.4, B_{max} = 0.21μmol/g ± 0.01) but has a modestly higher affinity for potato

231 starch (K_d = 1.9μM ± 0.3, B_{max} = 0.030μmol/g ± 0.001), which might be a function of the smaller

232 granule size and larger surface to mass ratio for corn starch. Exclusion of the *Rb*CBM26 in the

233 BIg-*Rb*CBM74-BIg construct led to slightly better binding to corn starch (K_d = 1.5μM ± 0.3, B_{max} =

234 0.18μmol/g ± 0.008) and modestly higher affinity but less overall binding to potato starch (K_d =

235 0.51μM ± 0.13, B_{max} = 0.015μmol/g ± 0.001). The saturation curve for BIg-*Rb*CBM74-BIg closely

236 resembles that of Sas6T and there is minimal binding by *Rb*CBM26, suggesting that *Rb*CBM74

237 drives insoluble starch binding.

238   The molecular patterns on the surface of starch granules differs between plant sources

239 and remains an active area of research [48-51]. The "hairy billiard ball model" to describe starch

240 granules postulates that the granule surface has block-like clusters of amylopectin chains with

10

241   hair-like extensions of amylose penetrating through the amylopectin [50]. Sas6T and BIg-

242   *Rb*CBM74-BIg bind amylose and amylopectin whereas *Rb*CBM26 only binds to amylopectin with

243   apparently low affinity based upon the relatively small change in migration (**Fig. 2G**). This

244   suggests that *Rb*CBM74 drives binding of Sas6 to the long, tightly packed helices of amylose at

245   the surface of the starch granule.

246   Using isothermal titration calorimetry (ITC), we found that Sas6T and BIg-*Rb*CBM74-BIg

247   bound amylopectin with sub-micromolar affinity whereas binding was not detectable for *Rb*CBM26

248   (**Table 2; Extended Data Fig 5A**) [52]. Sas6T binds maltotriose (G3), maltoheptaose (G7),

249   maltooctaose (G8) with a K_d in the hundreds of µM but exhibits a K_d of ~5µM for maltodecaose

250   (G10) (**Table 2; Extended Data Fig 5B**). Interestingly, *Rb*CBM26 binds shorter linear

251   oligosaccharides (G3, G7) and cyclodextrins, while BIg-*Rb*CBM74-BIg had no detectable affinity

252   for these sugars (**Table 2; Fig. Extended Data Fig 5C**). None of the constructs bound glucosyl-

253   α1,6-maltotriosyl-α1,6-maltotriose, an oligosaccharide of pullulan, suggesting that the α1,6

254   linkages are not specifically recognized by either domain. We determined that BIg-*Rb*CBM74-BIg

255   binds exclusively longer α-glucans of at least 8 residues. Notably, α1,4-linked glucose polymers

256   form double helices at 10 glucose units due to internal hydrogen bonding so we hypothesized that

257   *Rb*CBM74 might accommodate starch helices [8].

258

259   *Molecular Basis of RbCBM74 Binding* – We co-crystallized BIg-*Rb*CBM74-BIg with

260   maltodecaose (G10) to 1.70Å resolution (R_{work}=17.9%, R_{free}=19.9%) (**Fig. 3A**). Remarkably, we

261   observed two molecules of G10 as an extended double helix of ~42Å along the face of *Rb*CBM74

262   extending from S286 (reducing ends) to W373 (non-reducing ends). There was strong electron

263   density for 12 glucoses in one molecule, and nine glucoses in the other chain, likely reflecting

264   varied occupancy of the helix along the binding cleft (**Fig. 3B**). H289, F326, and W373 stood out

265   as surface exposed aromatic residues that might be providing CH-π mediated stacking (**Fig. 3C**).

266   An overlay of the unliganded and G10 bound structures demonstrates little global change

267   in the CBM74 domain upon binding (**Extended Data Fig 6A**), with the exception of G374 to K381.

268   In the unliganded structure this loop occludes surface exposure of W373 and in the G10 bound

269   structure the loop opens to create a continuous binding surface (**Extended Data Fig 6B**).

270   Additionally, Ca^{2+}-4 is exchanged for Na^+, representing flexibility in ion identity at that site

271   (**Extended Data Fig 6C**).

272   Canonical starch-binding domains feature two or three aromatic residues for pi-stacking

273   interactions with the aglycone face of maltooligosaccharides, but *Rb*CBM74 is designed for

274   extensive hydrogen-bonding interactions with longer oligosaccharides and starch [21]. The

275   binding site is continuous and each G10 molecule interacts with protein as a stretch of three Glcs

276   at a time, before the natural helical curvature brings the chain out of the contact with the protein

277   (**Fig. 3D**). For example, at the non-reducing end, Glc 1-3 of G10 chain A (G10A) fit into the ligand-

278   binding groove, while Glcs 4-6 of G10A are solvent exposed and Glc 1-3 of G10 chain B (G10B)

279   then fill the cavity. Along the length of the cavity, from the non-reducing end to the reducing end,

280   Glcs 1-3 and 7-9 of both G10A and G10B alternate to fill this binding site.

281   The binding cleft features a network of residues that hydrogen bond to the hydroxyl groups

282   of glucose (**Fig. 3E**). At the non-reducing end, Glc A1 hydrogen bonds with the indole nitrogen of

283   W373. Glc A2 stacks with W373 with hydrogen bonding provided by G374 and N403. Glc A3

284   hydrogen bonds with S338. The other molecule of G10 (B) contacts the next part of the binding

285   groove and is anchored by hydrogen bonding of Glc B3 by R336 and Y524. Where the first

286   molecule turns back into the binding groove, Glc A8 hydrogen bonds with E290, D549, and K556.

287   Glc A9 hydrogen bonds with the backbone of H289 and pi stacks with F326. The H289 side chain

288   hydrogen bonds with Glc B7 and provides aromatic character for pi stacking with Glc B8. Near

289   the region of *Rb*CBM74 that lies adjacent to *Rb*CBM26, K464 and S286 hydrogen bond with Glc

290   B9.

291   To define the starch-binding properties of *Rb*CBM74 in solution, we employed Hydrogen–

292   Deuterium eXchange Mass Spectrometry (HDX-MS). The conformational dynamics of BIg-

293   *Rb*CBM74-BIg alone and in the presence of G10 were measured over a 4-log timescale

294   (**Extended Data Fig 7AB**). The overall conformational dynamics of the apo protein were

295   consistent with the determined crystal structure, in terms of well-ordered domains and associated

296   loops or flexible regions. The flanking BIg domains showed higher exchange rates than the core

297   CBM74 domain. Intriguingly, the linker regions between domains do not show differentially high

298   dynamic exchange, as would be expected for flexibly tethered independent domains, further

299   supporting the integral nature of BIg-*Rb*CBM74-BIg motif.

300   The binding of G10 to *Rb*CBM74 was explored by differential protection from exchange in

301   the absence and presence of G10. Significant protection was observed in the presence of G10,

302   while no significant increases in exchange were observed (**Extended Data Fig 7C**). This is

303   consistent with the minimal global conformation changes between the two states of the protein.

304   The protected regions upon G10 binding were highly localized to a single surface binding region

305   (**Fig. 3F**). This protected region constitutes a single extended surface, which directly overlaps with

306   the G10 binding site observed in the co-crystal structure (**Fig. 3EF**). With the exception of the

307   peptide from A314-Y318 (ANTTY), each of the protected peptides identified by HDX-MS contains

308   at least one key binding residue identified from the co-crystal structure (**Fig. 3E**). These data

309   provide a comprehensive picture of the structural dynamics of *Rb*CBM74 binding to long

310   maltooligosaccharides via an extended starch binding cleft.

311

312   *RbCBM74 Mutational Studies* – Because most CBM binding is mediated by aromatics, we

313   hypothesized that mutation of W373, F326, or H289 to Ala would dramatically decrease or

314   eliminate binding. We tested maximum binding of each of the aromatic mutants to insoluble corn

315   (1%) and potato starch (5%). The W373A and H289A constructs lost the ability to bind to insoluble

316   corn starch while binding of the F326A construct was greatly reduced (**Fig. 4A**). This trend was

13

317   somewhat different for potato starch, in which a lower percentage of H289A bound compared to

318   the F326A and W373A mutants. By affinity PAGE, neither the W373A nor the F326A mutant lost

319   appreciable binding to amylopectin while the H289A mutant had a modest decrease in binding to

320   potato amylopectin (**Fig. 4B**). When we quantified binding via ITC, W373A lost all binding for G10

321   while H289A and F326A had a ~10-20-fold decrease in affinity (**Table 2, Extended Data Fig 8A**).

322   On potato amylopectin, F326A had a 10-fold reduction in affinity while H289A and W373A

323   exhibited a ~20-fold reduction (**Extended Data Fig 8B**). That single mutations do not eliminate

324   binding is perhaps not surprising given the extensive binding platform. Moreover, the enhanced

325   affinity of these mutants to amylopectin over G10 further suggests that productive interactions

326   with the protein extend beyond a 10-glucose unit footprint. Indeed, the somewhat staggered

327   double helical G10 bound in our crystal structure suggests that at least 12 glucose units contribute

328   to binding (**Fig. 3D**).

329

330   *Native mass spectrometry* – ITC revealed a binding stoichiometry of 1:1 between BIg-

331   *Rb*CBM74-BIg and G10, while the co-crystal structure demonstrates that two molecules of G10

332   are accommodated. To better determine the stoichiometry of this binding event, we employed

333   native mass spectrometry in the presence of varying concentrations of G10 (**Fig. 5A**). Each

334   observed state differed by ~1639 Da, which agrees with the theoretical mass of G10 (**Extended**

335   **Data Table 1A**). To obtain binding affinities, we summed the peak intensities of all abundant

336   charge states in our mass spectra and analyzed these intensity values as described previously

337   [53] (**Extended Data Table 1B**). The K_d for BIg-*Rb*CBM74-BIg was determined to be 3.8 ± 0.5µM,

338   which agrees with our ITC data. As the concentration of ligand is increased, ligand molecules can

339   bind nonspecifically during the nESI process, generating artifactual peaks in the mass spectra

340   corresponding to a two ligand-bound complex (**Fig. 5A**). We speculate that in excess

341   concentrations of G10, the molecules can form double helices that are accommodated by the

14

342 *Rb*CBM74 binding site but that the single molecule binding event represents the most common

343 binding conformation (**Fig. 5B**).

344 Because Sas6 encodes both a CBM74 and a CBM26, and this co-occurrence is

345 evolutionarily well-conserved, we speculated that *Rb*CBM26 and *Rb*CBM74 could either bind

346 separate G10 molecules or that one ligand could span the region between the two CBM binding

347 sites [22]. We used native mass spectrometry to determine the number of G10 molecules bound

348 to Sas6T, which includes both CBMs. The binding state distribution was markedly different when

349 *Rb*CBM26 was included (**Fig. 5C**). At low G10 concentrations, there is a mix of unliganded, 1-

350 bound, and 2-bound states unlike BIg-*Rb*CBM74-BIg alone (**Fig. 5D**). As G10 increases, the apo

351 and 1-bound states decrease as the 2-bound fraction increases. For Sas6T, K_d values for 1:1 and

352 1:2 protein:ligand complexes were calculated to be 3.4 ± 0.5 µM and 165.6 ± 38.8 µM,

353 respectively, and are in reasonable agreement with ITC data (**Extended Data Table 1B**).

354 Together these results suggest that *Rb*CBM26 and *Rb*CBM74 each bind one molecule of G10

355 independently in solution. In the context of a starch granule, this supports a model whereby each

356 CBM of Sas6 binds adjacent α-glucan chains rather than attaching to the same chain in a

357 continuous manner. Moreover, the propensity for BIg-*Rb*CBM74-BIg to bind a single helix of G10

358 at low ligand concentrations, as also observed with ITC, suggests that this binding platform

359 prefers single helical α-glucans such as amylose, though it can also tolerate double helical

360 stretches of amylopectin.

361

362 *CBM74 Conservation* – To visualize conserved features of CBM74 domains, two

363 alignments and a corresponding evolutionary tree were prepared. The first alignment includes all

364 99 CBM74 sequences (**Extended Data Fig 9A,B; Extended Data Table 2**) while the second was

365 simplified for viewing and includes 33 representative CBM74 sequences (**Fig. 6**). Both alignments

366 reveal that the CBM74s fall into 6 distinct clades (**Fig. 6A; Extended Data Fig 9A**). *Rb*CBM74

367 (No. 28) is in a distinct cluster of proteins (blue) that invariably include a dockerin domain as part

368   of the full-length protein. However, there are other CBM74 domains originating from dockerin-

369   containing proteins found in three more groups (green, cyan, and magenta). The prototypical

370   CBM74 of the subfamily GH13_32 α-amylase from *Microbacterium aurum* (No. 52) bins into a

371   clade (cyan) with its GH13_32 counterpart from *Sanguibacter* sp. (No. 54) and the CBM74-

372   containing α-amylase from *Clostridium bornimense* (No. 58). A similar GH13_28 α-amylase from

373   *Streptococcus suis* (No. 68) is in the adjacent cluster (magenta) very close to the CBM74 domains

374   from two other hypothetical dockerin-containing proteins from *Ruminococcus bovis* (No. 67) and

375   Ruminococcaceae bacterium (No. 70). Most CBM74 domains appended to α-amylases from the

376   subfamily GH13_28, predominantly from Bifidobacteria, group together in a separate cluster (red).

377   Finally, the sixth cluster (walnut) covers CBM74 domains found in GH13_19 α-amylases. In total,

378   CBM74 domains occur in α-amylases from several subfamilies or non-catalytic dockerin-

379   containing proteins and are widely represented among Bifidobacteria.

380   We mapped the conservation of all 99 CBM74 family members onto our structure using

381   CONSURF [54-56] (**Fig 6B**). While the central β-sandwich, ion-coordination sphere, and ligand

382   binding site are highly conserved, the flexible loop in *Rb*CBM74 (residues 373-384) occluding the

383   binding site is more variable (**Fig. 6C; Extended Data Fig. 9B**). In all but the three or four most

384   closely related CBM74 sequences – covering only the two genera of *Ruminococcus* and

385   Eubacterium – this loop is short or not present, though how this feature correlates with binding is

386   unknown.

387   Most of the key aromatic residues that mediate starch-binding in *Rb*CBM74 are highly

388   conserved (**Fig. 6C**). W373 from *Rb*CBM74 is 100% conserved among all 99 identified CBM74

389   family members (**Extended Data Fig 9B**), while H289 is shared with 78 sequences or substituted

390   with a Tyr (18/99) in Bifidobacteria and *Candidatus* s*catavimonas* (No. 25) and a Trp (3/99) in

391   *Pseudoscardovia* species. F326 is perhaps the most variable, sharing sequence identity or

392   similarity with 3 of the 6 clades (F-19/99, Y-43/99), while the other clades feature a glycine or

393   alanine in this position (36/99). The binding site also features an elaborate network of residues

16

394   that provide hydrogen bonding with the ligand. The residues at the center of the cleft including

395   K556 (80/99), D549 (63/99), and E290 (99/99) exhibit the highest conservation (**Fig. 6C;**

396   **Extended Data Fig 9B**). The hydrogen bonding residues at the ends of the cleft are more varied,

397   including S286 (22/99) which interacts with the *Rb*CBM26 ligand. Intriguingly, in a large proportion

398   of the sequences there is an aromatic residue at the site of K556 (W-19/99) and Y524 (Y-12/99,

399   F-45/99) that could provide pi stacking in those CBM74s. This moderate variability in the

400   composition of the putative binding site may suggest that CBM74 family members have different

401   affinities for starch.

402

403   **Discussion**

404   CBMs are distinct protein domains that assist with substrate breakdown by specifically

405   binding polysaccharide targets. These domains are especially important for binding to insoluble

406   substrates like crystalline cellulose and semi-crystalline starch granules. The CBM74 family binds

407   insoluble starch and its constituents, amylose and amylopectin. CBM74 domains are frequently

408   (81/99 sequences) encoded adjacent to another starch-binding CBM family, either a CBM25 or

409   CBM26 [22]. Sas6 includes both a CBM26 and a CBM74 domain that have different affinities for

410   maltooligosaccharides but work together to bind granular starch. *Rb*CBM26 has a canonical

411   binding platform that accommodates motifs found in linear and circular maltooligosaccharides. In

412   contrast, *Rb*CBM74 has an extended ligand binding groove that requires at least 8 glucose

413   residues and accommodates the single helices of amylose and the double helices found in

414   amylopectin. Because it is on the cell surface, the CBM74 domain of Sas6 may target *R. bromii*

415   to the crystalline regions of starch granules that are not easily accessible to human or other

416   bacterial amylases.

417   Sas6 is a putative *R. bromii* amylosome component and likely cooperates with amylases

418   and pullulanases via the interaction of its dockerin domain with a cohesin from a scaffoldin protein

419   [20]. Because Sas6 is found on the cell surface, it could bind cell anchored scaffoldins Sca2 or

420   Sca5, associate with Sca1/Amy4, or bind the cell surface in a dockerin-independent mechanism

421   [20]. Breakdown of starch by *R. bromii* relies on the coordinated effort of approximately 40 distinct

422   proteins, of which Sas6 may play an integral part by specifically targeting the helical regions of

423   starch [20].

424   Unlike *R. bromii*, resistant starch-utilizing Bifidobacteria encode CBM74-containing

425   multimodular extracellular amylases [9]. A recent study looked at the amylases that were

426   differentially encoded between Bifidobacterial strains that could bind and degrade starch granules

427   and those that could not [57]. Resistant Starch Degrading enzyme 3 (RSD3) was differentially

428   encoded in the resistant starch-binding strains. It contains a CBM74 domain and has high activity

429   on high amylose corn starch. RSD3 has an N-terminal GH13 domain followed by CBM74, CBM26,

430   and CBM25 domains. The CBM74-CBM26 motif is present in RSD3 so the structural and

431   functional insights we have gleaned from Sas6 may suggest how these CBMs structurally assist

432   the enzyme with granular starch hydrolysis.

433   Although starch is a polymer composed solely of glucose, there is massive variation in

434   granule structure [7, 8]. This is a function of primary structure (i.e. α1,4 or α1,6 linkages),

435   secondary structure (single or double helices) and tertiary structure (helical packing and amylose

436   content), making granules an exquisitely complex substrate [58]. This complexity is unlocked by

437   only a few specialized gut bacteria, making granular starch a targeted prebiotic [9, 15, 16]. CBM74

438   domains might serve as a molecular marker for the ability to break down resistant starch in

439   metagenomic samples [22]. Furthermore, CBM74 domains might make attractive additions to

440   engineered enzymes for enhanced starch degradation on the industrial scale, or as an adjunct to

441   starch prebiotics. The structural and functional picture of *Rb*CBM74 here will accelerate the

442   targeted use of this domain for various health and industrial applications.

18

Table 1: X-ray Data Collection and Refinement Statistics			
Construct	Sas6T + α-cyclodextrin	Sas6T unliganded	Blg-*Rb*CBM74-Blg + G10
PDB Accession	7UWW	7UWU	7UWV
Wavelength	0.979	0.979	0.979
Resolution range	35 - 1.61 (1.67 - 1.61)	44.77 - 2.19 (2.27 - 2.19)	62.48 - 1.70 (1.76 - 1.70)
Space group	P 21 21 21	P 21 21 21	P 21 21 2
Unit cell	69.5 82.5 213.5 90 90 90	69.2 82.4 213.3 90 90 90	69.7 160.1 67.8 90 90 90
Total reflections	1690626 (29309)	1044914 (104226)	512772 (51008)
Unique reflections	122358 (2727)	63042 (6261)	84071 (8261)
Multiplicity	13.8 (10.7)	16.6 (16.9)	6.1 (6.2)
Completeness (%)	76.8 (24.8)	99.34 (99.97)	99.9 (100.0)
Mean I/sigma(I)	16.0 (1.1)	22.66 (14.34)	17.2 (2.2)
R-merge	0.092 (1.799)	0.0956 (0.192)	0.052 (0.75)
R-meas	0.095 (1.889)	0.0987 (0.1979)	0.057 (0.81)
R-pim	0.025 (0.563)	0.0243 (0.04781)	0.023 (0.33)
CC1/2	0.999 (0.549)	0.998 (0.993)	0.999 (0.822)
Reflections used in refinement	122315 (3895)	63244 (6262)	84061 (8260)
Reflections used for R-free	6093 (201)	3200 (309)	4056 (427)
R-work	0.168 (0.238)	0.197 (0.276)	0.179 (0.279)
R-free	0.212 (0.281)	0.255 (0.348)	0.199 (0.281)
Number of non-hydrogen atoms	11425	10523	4954
macromolecules	9721	9527	4043
ligands	253	38	237
solvent	1451	964	674
Protein residues	1294	1246	531
RMS(bonds)	0.013	0.001	0.013
RMS(angles)	1.4	0.4	1.7
Ramachandran favored (%)	96.3	96.1	97
Ramachandran allowed (%)	3.6	3.9	3
Ramachandran outliers (%)	0.08	0	0
Rotamer outliers (%)	0	1.6	0.5
Clashscore	3.26	4.98	0.24
Average B-factor	22.2	25.6	33.2
macromolecules	20.7	25.4	31.8
ligands	35.8	29.9	31.5
solvent	30	27.8	41.8

443
444
445

446

Table 2: Sas6 and domain binding via Isothermal Titration Calorimetry												
	G3		ACX		G7		G8		G10			Potato AP
	n	K_d (µM)	n	K_d (µM)	n	K_d (µM)	n	K_d (µM)	n	K_d (µM)		K_d (µM)
Sas6T	1*	880 ± 25	1.0	178 ± 26	1*	332 ± 15	1*	496 ± 260	0.9	5.5 ± 1.9		0.3 ± 0.08
*Rb*CBM26	-	NB	0.8	169 ± 16	1	310 ± 34	1*	285 ± 84	0.7	252 ± 128		NB
BIg-*Rb*CBM74-BIg	-	NB	-	NB	-	NB	1*	820	0.9	5.2 ± 1.1		0.7 ± 0.05
W373A									NB	NB		13.4 ± 5.4
H289A									0.5	73.1 ± 7.7		21.4 ± 3.1
F326A									0.7	100 ± 11		3.9 ± 1.4

447
448 * indicates that n was set to 1. Experiments were performed in triplicate with mean ± standard
449 deviation reported. For amylopectin, curves were modeled for total binding (n=1).
450
451
452
453
454
455
456
457

20

458
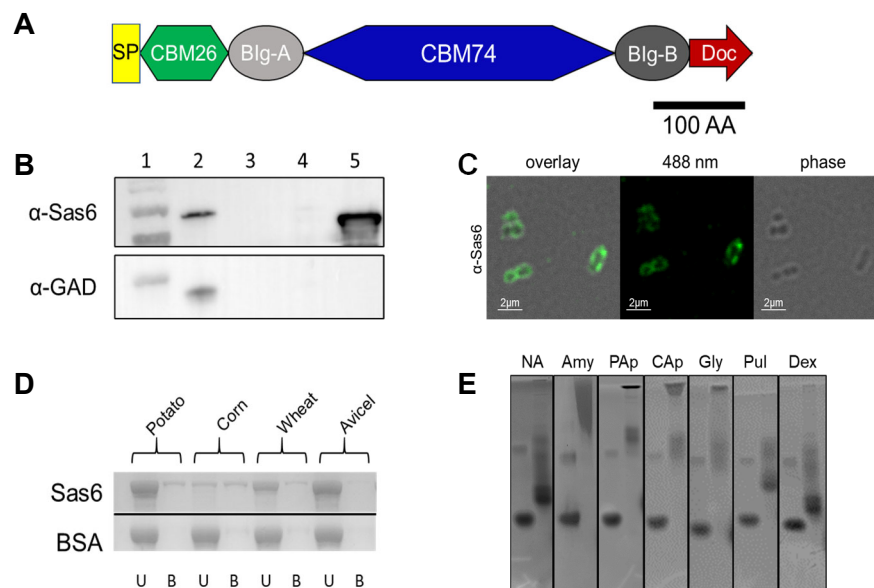
459

Figure 1: *Ruminococcus bromii* Sas6 is a starch-binding protein that contains two carbohydrate-binding modules. A. Domain architecture of Sas6 annotated according to the Carbohydrate Active Enzyme database (www.cazy.org) and the crystal structure. SP = Signal Peptide, CBM26 = Carbohydrate Binding Module family 26, BIg = Bacterial Immunoglobulin, CBM74 = Carbohydrate Binding Module family 74, Doc = Dockerin. **B.** *Top:* Western blot with anti-Sas6 antibody showing localization of Sas6 in the cell fraction. *Bottom*: Parallel western blot with custom rabbit antiserum against glutamic acid decarboxylase to control for cell lysis. Lane 1: ladder, 2: *R. bromii* cell lysate, 3: cell-free culture supernatant, 4: TCA precipitated cell-free culture supernatant, 5: recombinant Sas6T, truncated version of Sas6 lacking the C-terminal dockerin. **C.** α-Sas6 immunofluorescent staining of fixed *R. bromii* cells grown in potato amylopectin. **D.** SDS-PAGE gel from Sas6 adsorption to potato, corn, and wheat starch, and Avicel (cellulose) control. U=unbound protein, B=bound protein. **E.** Affinity PAGE with 0.1% of the indicated polysaccharide incorporated into the gel matrix. For each, left lane is bovine serum albumin, right lane is Sas6T. NA=native gel, Amy=potato amylose, PAp=Potato Amylopectin, CAp=corn amylopectin, Gly=Glycogen, Pul=Pullulan, Dex=Dextran.

475
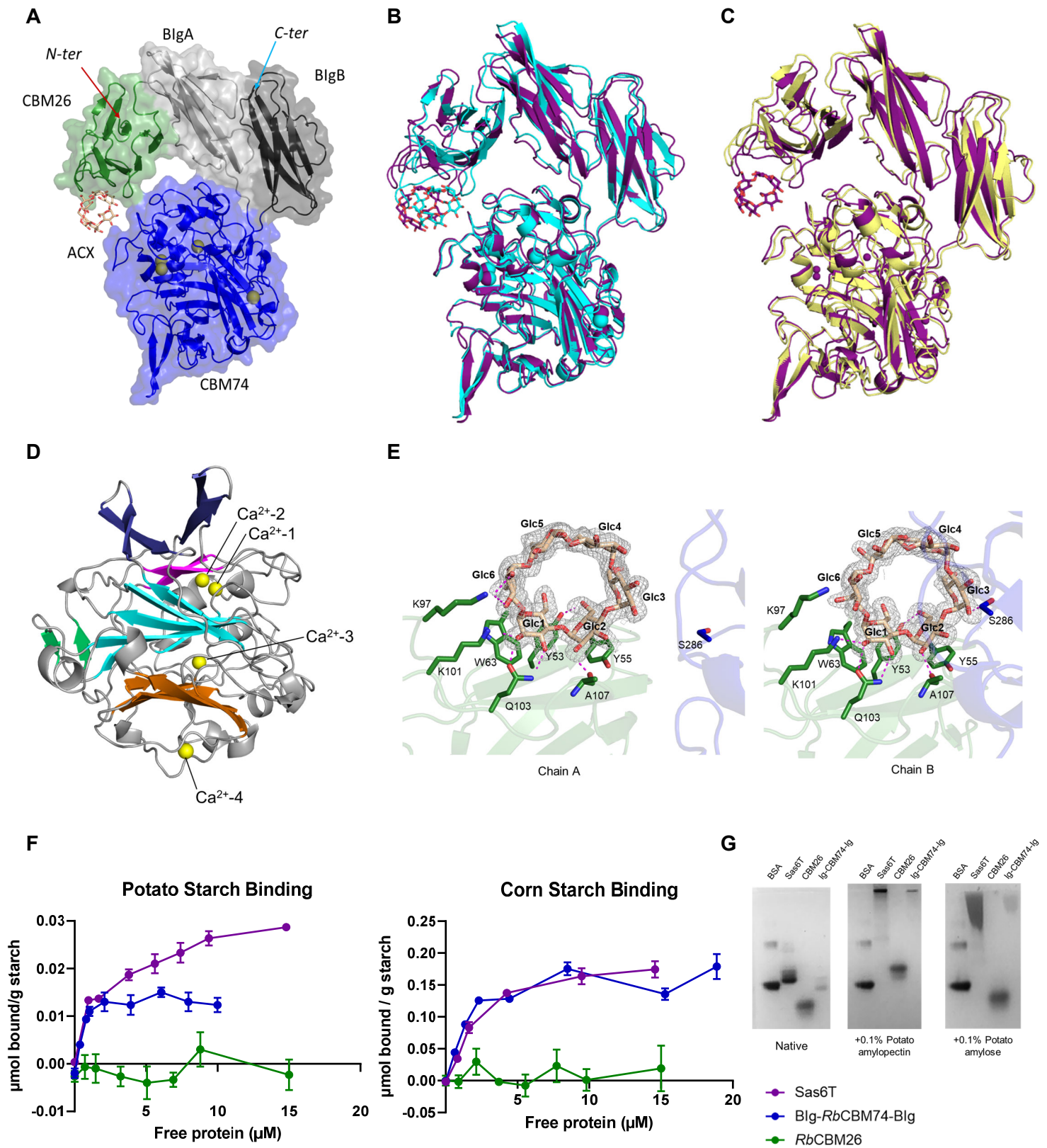
Figure 2: Sas6 is a compact protein with two BIg domains that orient *Rb*CBM26 and *Rb*CBM74. **A.** Semi-transparent surface rendition and cartoon of Sas6T (PDB *7uww*) with *Rb*CBM26 domain in green, BIgA in light grey, *Rb*CBM74 in blue, and BIgB in dark grey. The α-cyclodextrin (ACX) bound to *Rb*CBM26 is shown in wheat sticks and Ca^{2+} atoms are shown as yellow spheres. **B.** Overlay of Chain A (purple) and Chain B (cyan) within the asymmetric unit of *7uww* showing variation in the position of ACX relative to *Rb*CBM74. **C.** Overlay of Chain A of *7uww* (purple) and SAXS-derived MultiFoXS model (yellow). **D.** Side view of *Rb*CBM74 with the central β-sandwich sheets in orange and cyan. A third β-sheet is shown in magenta and the protruding pairs of β-strands and in dark blue. β-strands connecting the beginning and end of the *Rb*CBM74 domain are colored green. Ca^{2+} atoms are shown as yellow spheres. **E.** ACX bound at *Rb*CBM26 (green) in chain A (left) and chain B (right), demonstrating minor conformational flexibility that places S286 from *Rb*CBM74 (blue) within the binding site. Side chains involved in ligand binding are shown as green sticks with a hydrogen bond cutoff of 3.2Å. ACX is displayed as wheat sticks. Omit map is contoured to 2.0σ and carved within 1.6Å of ACX ligand. **F.** *Rb*CBM74 drives binding to granular potato and corn starch. Binding to granular starch was determined by isotherm depletion. The μmoles of protein bound per gram of starch was plotted against [free protein] to determine dissociation constants (K_d) and binding maxima (B_{max}) using a one-site specific binding model in GraphPad prism. **G.** Affinity PAGE of Sas6T or individual domains, *Rb*CBM26 and BIg-*Rb*CBM74-BIg, with 0.1% polysaccharide. BSA= bovine serum albumin.
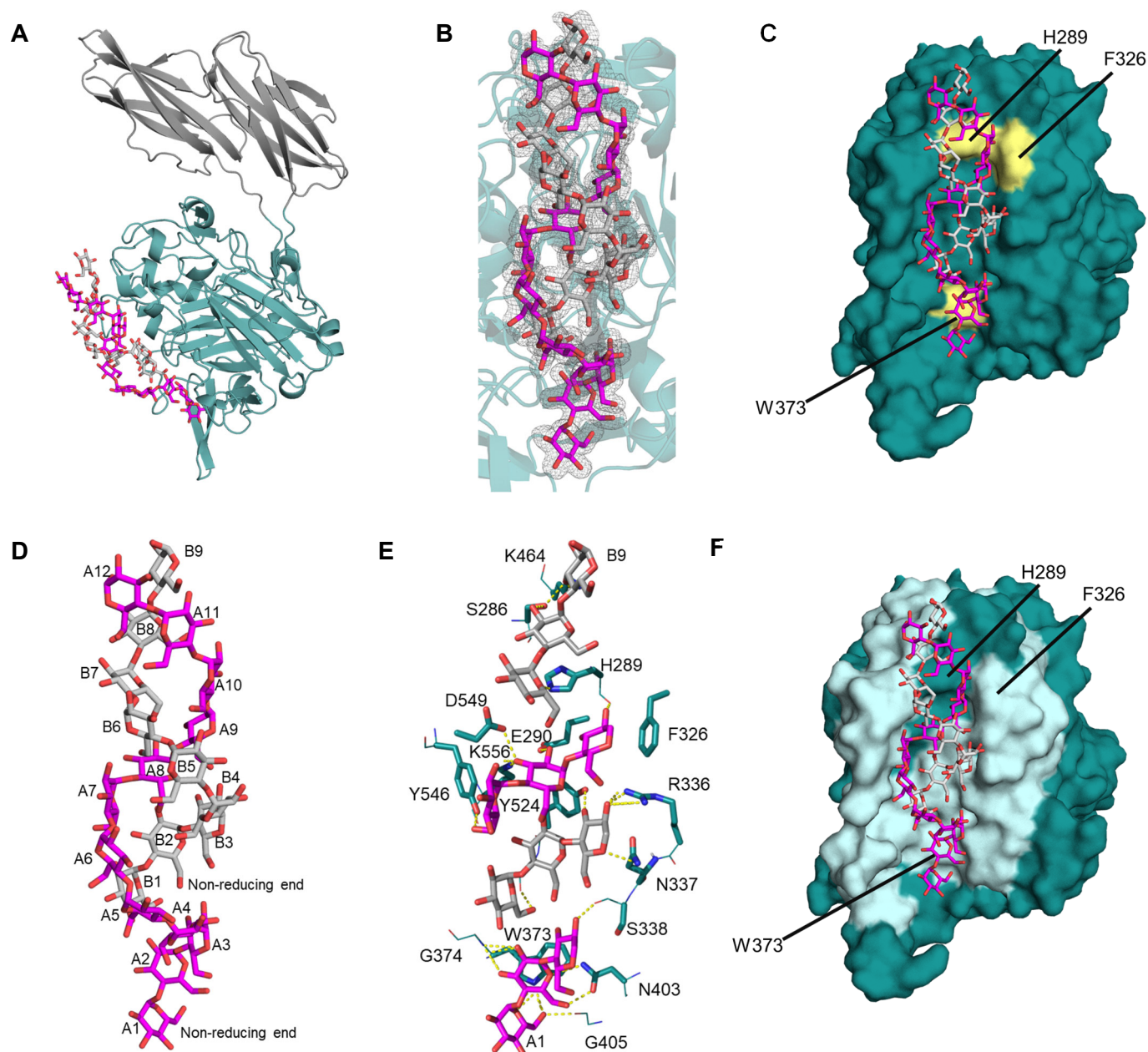
Figure 3: *Rb*CBM74 has an extended groove that accommodates starch double helices.
A. The BIg-*Rb*CBM74-BIg (PDB *7uwv*) starch-binding site is an extended groove that spans nearly the length of the domain. A cartoon representation of BIgA in light grey, CBM74 in teal, and BIgB in dark grey with two chains of maltodecaose (G10) wrapped around one another shown in magenta and grey sticks. **B.** *Rb*CBM74 is co-crystallized with G10 in a double helical conformation. Electron density for G10 demonstrated by an omit map contoured to 2.0σ and carved to 1.6Å with one chain of modeled Glc in magenta and the other in grey. **C.** *Rb*CBM74 has an inset binding groove that accommodates the width of the starch double helix with aromatic CH-π stacking provided by W373, F326, and H289. A surface representation of CBM74 (teal) with aromatic residues colored in yellow and G10 represented by magenta and

507   grey sticks. **D.** Double helical G10 structure with Glc residues labeled from non-reducing to
508   reducing ends. One chain of G10 (A1-12) shown in magenta and the other in grey (B1-9) sticks.
509   **E.** Corresponding hydrogen-bonding network (3.2Å cutoff) between *Rb*CBM74 and G10. Side
510   chains involved in hydrogen bonding are shown in teal sticks with nitrogens indicated in blue
511   and oxygens in red. Hydrogen bonds are indicated by yellow dashed lines and G10 residues
512   directly involved in binding are shown in magenta (G10 molecule A) and grey (G10 molecule B)
513   sticks. **F.** Surface representation of *Rb*CBM74 with peptides protected from deuterium exchange
514   in the presence of G10 colored in light cyan as determined by hydrogen-deuterium exchange
515   mass spectrometry.
516
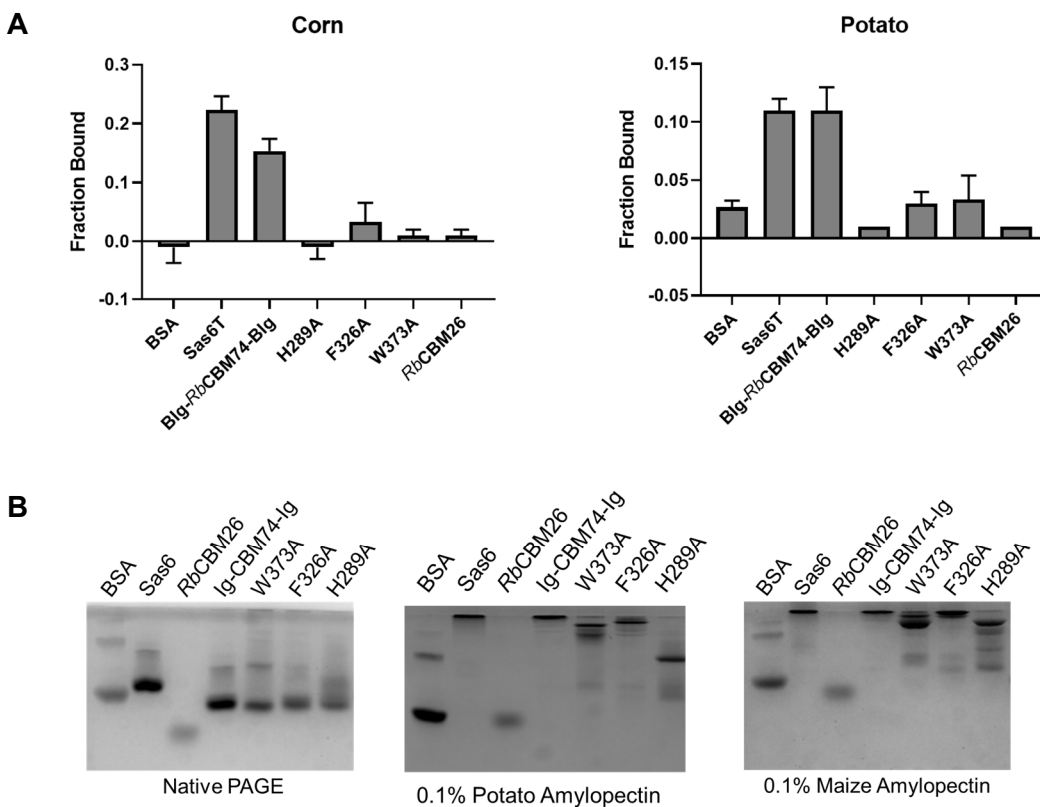
517
518   **Figure 4: W373A, F326A, and H289A mediate starch binding by *Rb*CBM74. A.** Binding to
519   insoluble starch is eliminated or greatly reduced when W373, H289 or F326 is mutated. The
520   amount of protein bound to starch granules was determined by quantitation of protein remaining
521   in solution after binding (n = 3). **B.** Mutation of aromatic residues decreases but does not
522   eliminate binding to amylopectin. Affinity PAGE with 0.1% potato amylopectin or maize
523   amylopectin added to the gel matrix. Binding is indicated by reduced migration through the gel.
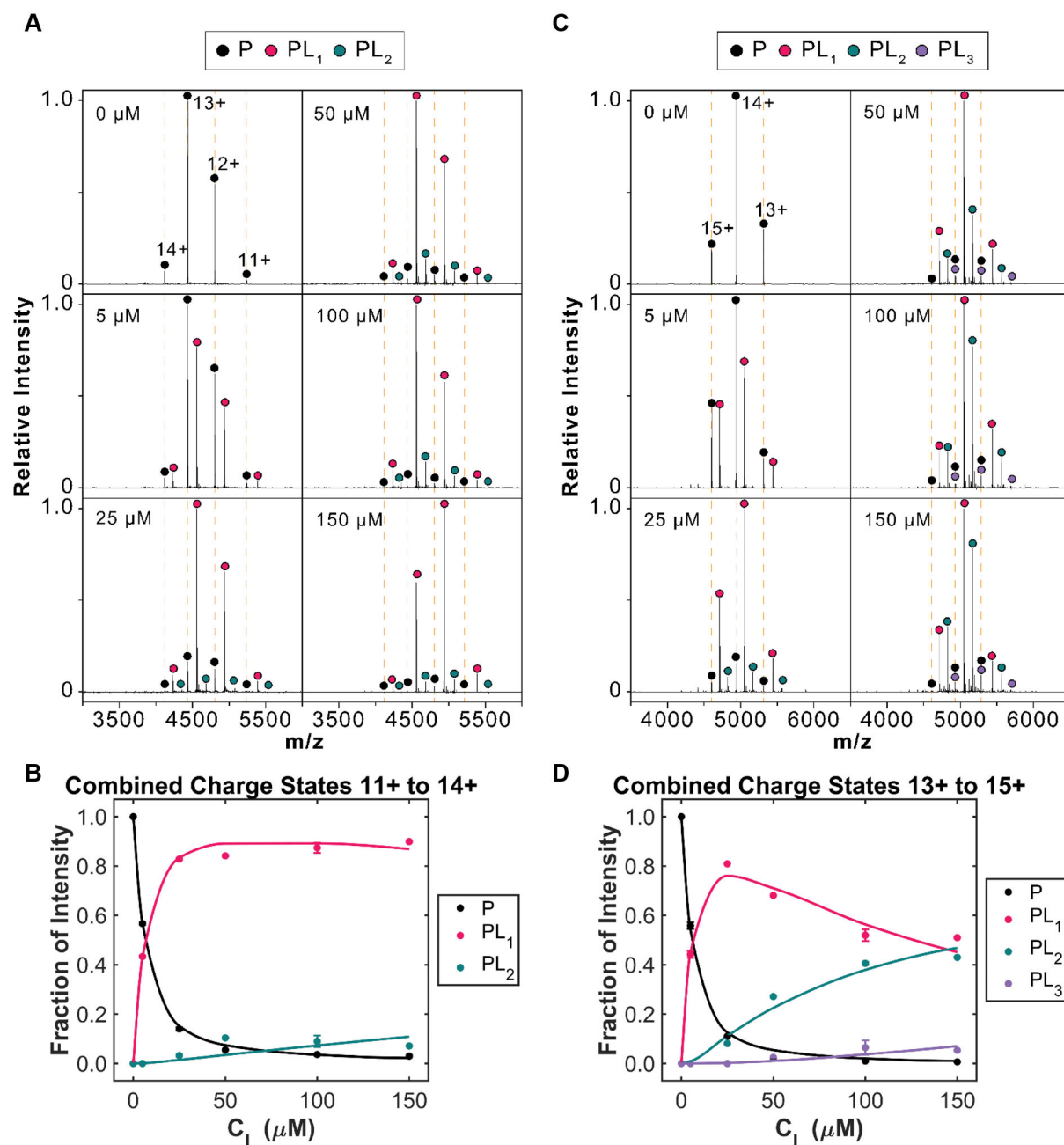
Figure 5: *Rb*CBM74 and *Rb*CBM26 bind separate molecules of G10 in solution. A. Mass spectra of BIg-*Rb*CBM74-BIg at different ligand concentrations (0 - 150µM) and a fixed protein concentration of 5µM. Charge states for unbound protein are annotated with an orange dashed line. Peaks corresponding to different bound states are observed after each charge state of the unbound protein. Intensities of each species, combined across multiple charge states, were then extracted from the mass spectra and used to calculate the fractional abundance of unbound and bound states at equilibrium (n=3). **B.** Nonlinear least-squares fitting of the titration data for BIg-*Rb*CBM74-BIg. **C.** Mass spectra of Sas6T as described in A. **D.** Nonlinear least-squares fitting of the titration data for Sas6T.
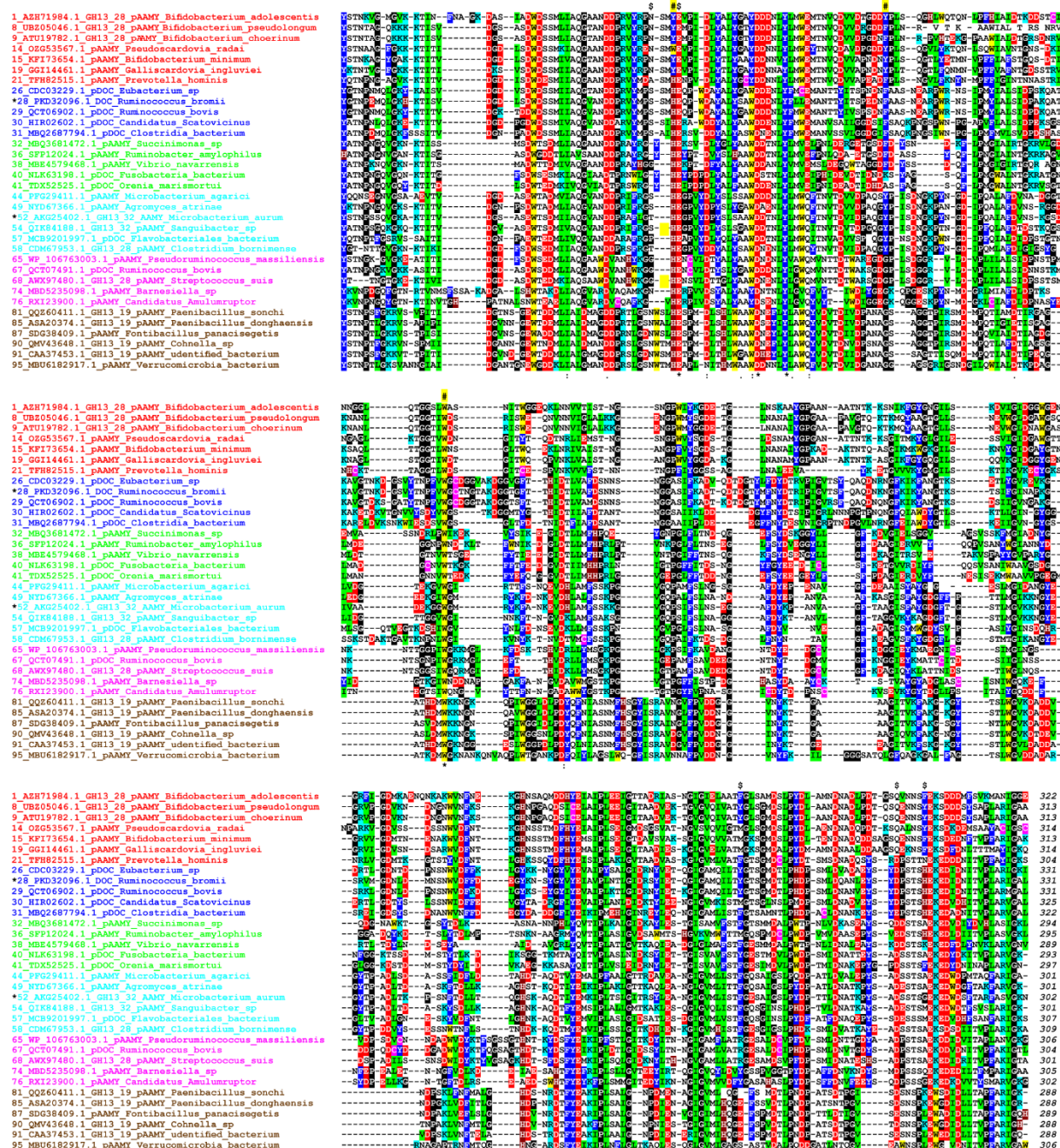
534
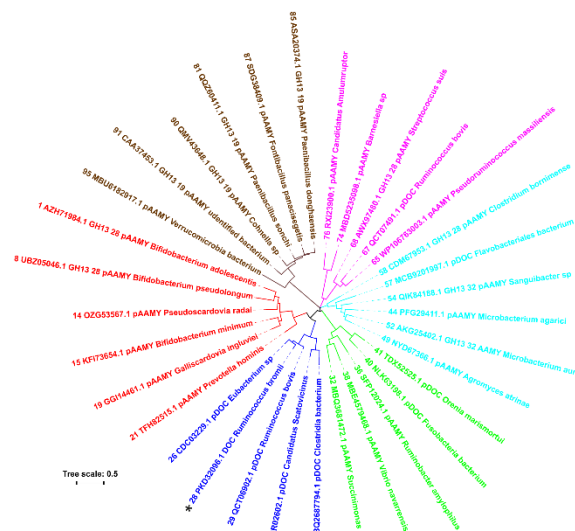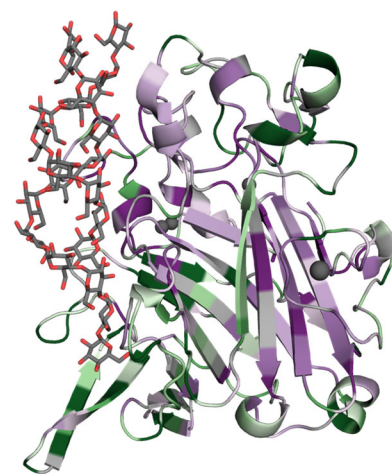
535

536 **Figure 6: Conservation of binding residues among select CBM74 family members. A.**
537 Evolutionary tree for the CBM74 family including 33 sequences selected from the entire studied
538 set of 99 CBM74s (Extended Data Table 2). Two experimentally characterized CBM74s are
539 marked by an asterisk: Sas6 from *Ruminococcus bromii* (No. 28, blue cluster) and the subfamily
540 GH13_32 α-amylase from *Microbacterium aurum* (No. 52; cyan cluster). Protein labels include
541 the order number (33 selected from 1-99), GenBank accession number, abbreviation of the
542 source protein/enzyme and organism name. The tree is based on the alignment (shown in C)
543 spanning the complete CBM74 sequences. **B.** Structure of *Rb*CBM74 (PDB *7uwv*) colored by
544 conservation score from least conserved (green) to most conserved (purple) generated using
545 CONSURF. **C.** Sequence alignment of the CBM74 family. The six individual groups
546 distinguished from each other by different colors correspond to six clusters seen in the
547 evolutionary tree (panel A); the sequence order in the alignment reflects their order in the tree in
548 the anticlockwise manner (starting from the first sequence in the red cluster). The residues
549 responsible for stacking interactions and involved in hydrogen bonding with glucose moieties of
550 the bound α-glucan are signified by a hashtag and a dollar sign, respectively, above the
551 alignment. The flexible loop observed in the three-dimensional structure of the *Rb*CBM74 is
552 highlighted by the short yellow strip over the alignment. Identical and similar positions are
553 signified by asterisks and dots/semicolons under the alignment blocks. The color code for the
554 selected residues: W, yellow; F, Y – blue; V, L, I – green; D, E – red; R, K – cyan; H – brown; C
555 – magenta; G, P – black. The alignment of all 99 CBM74 sequences of the present study shown
556 in Extended Data Figure 9B.

557

558

559 **METHODS**

560 *Recombinant Protein Cloning and Expression*

561 We used a previously described cloning and expression protocol to generate each of the

562 recombinant protein constructs used in this study [59]. Genomic DNA was isolated from *R. bromii*

563 strain L2-63 and the constructs for Sas6 without the signal peptide were amplified using the

564 primers listed in **Table S1** with overhangs complementary to the Expresso T7 Cloning &

565 Expression System N-His pETite vector (Lucigen). The forward primers were engineered to

566 include the 6x His sequence that complemented the vector plus a TEV protease recognition site

567 for later tag removal. PCR was performed with Flash PHUSION polymerase (ThermoFisher). The

568 amplified products and the linearized N-his pETite vector were transformed in HI-Control10G

569 Chemically Competent Cells (Lucigen) and plated on LB plates supplemented with 50 μg/ml

570 kanamycin (Kan). Transformants were screened for the insertion of Sas6 and validated via

28

571 sequencing. The Sas6-pETite plasmids were transformed into chloramphenicol (Chl)-resistant *E.*

572 *coli* Rosetta (DE3) pLysS cells and plated on LB plates supplemented with 50μg/ml Kan and

573 20μg/mL Chl. *E. coli* cells were grown at 37°C to OD$_{600}$ 0.6-0.8 in Terrific Broth supplemented

574 with 50μg/ml Kan and 20μg/ml Chl after which time the temperature was lowered to 20°C and

575 0.5mM Isopropyl β-d-1-thiogalactopyranoside (IPTG) was added. After 16 hours of growth, 1L of

576 cells was centrifuged, resuspended in 40mL of Buffer A (20mM Tris pH 8.0, 300mM NaCl) and

577 lysed by sonication. Cell lysate was separated from cell debris by centrifugation for 30min at

578 30,000xg. 3mL of Ni-NTA resin was packed into Econo-Pac Chromatography Columns (BioRad)

579 and equilibrated with Buffer A. Lysate was passed through the packed columns and washed with

580 70mL of Buffer A. Proteins were eluted from the columns via stepwise increase in Buffer B (20mM

581 Tris pH 8.0, 300mM NaCl, 500mM imidazole). Proteins eluted in 10-25% Buffer B fractions. TEV

582 protease (1mg) was added to each protein to initiate cleavage of the His-tag and the mix was

583 dialyzed overnight using dialysis tubing (SpectraPor) in 1L of storage buffer (20mM HEPES pH

584 7, 100mM NaCl). The dialyzed protein-TEV mixture was applied to Ni-NTA resin and the flow-

585 through was collected and concentrated using a VivaSpin 20 concentrator (Fisher Scientific).

586

587 *Sas6 Immunofluorescence*

588 Custom α-Sas6T antiserum was generated by rabbit immunization with purified recombinant

589 Sas6T protein (Lampire Biological Laboratories). The resulting antiserum was used for western

590 blotting and cell staining. *R. bromii* cells were grown to mid-log phase on RUM media [17] with

591 0.1% potato amylopectin and 2mL of the cell culture was collected for immunostaining and

592 western blotting. For immunostaining, 1mL of *R. bromii* culture was centrifuged for 1min at

593 13,000xg and washed 3 times with 1X phosphate buffed saline pH 7.4 (PBS). 2μL of cells were

594 then spread on a glass slide and fixed with 10% formaldehyde in PBS. Slides were washed 3x in

595 PBS to remove fixative but were not permeabilized. Cells were blocked for 30min with 10% goat

29

596 serum (Jackson ImmunoResearch). α-Sas6T antiserum was diluted 1:1000 in 10% goat serum

597 and applied for 1hr to cells at room temperature. The primary antiserum was removed, and slides

598 were washed 3 x 5min in PBS before the application of 1:500 goat α-rabbit AlexaFluor488

599 antibody (ThermoFisher) for 30min. Slides were washed 3 x 5min in PBS and preserved with

600 Prolong Gold Antifade reagent and dried overnight before imaging. Cells were imaged at the

601 University of Michigan Microscopy Core on a Leica Stellaris Light Scanning Confocal microscope

602 with a 100X objective.

603

604 *Western Blotting*

605 *R. bromii* was grown to mid-log phase overnight on RUM media containing 0.1% potato

606 amylopectin [17]. 1mL of cells was pelleted and washed twice in phosphate buffered saline (PBS)

607 pH 7.4, then resuspended to a final volume of 50µL in 5mM Tris-HCl pH 8.5. The culture

608 supernatant was passed through a 0.2µm filter and 50µL was reserved for analysis. Proteins were

609 precipitated from the remaining supernatant by the addition of ¼ volume of 100% trichloroacetic

610 acid (TCA) and incubated 30 min on ice. The precipitate was collected via centrifugation and

611 washed twice with 200µL cold acetone. The resulting pellet was dried and resuspended in 50µL

612 of 5mM Tris-HCl pH 8.5. Samples were separated by SDS-PAGE on two 10% Tris-glycine gels,

613 then transferred to polyvinylidene difluoride (PVDF) membrane. Blots were blocked in EveryBlot

614 Blocking Buffer (BioRad) for 30min then washed with PBS pH 7.4 + 0.05% Tween 20 (PBST). To

615 detect Sas6, one membrane was incubated with custom rabbit α-Sas6 antiserum (Lampire)

616 diluted 1:500 and the other with custom rabbit α-glutamic acid decarboxylase from *R. bromii*

617 (Lampire) diluted 1:10,000 in PBST + 5% non-fat dry milk (PBST-milk) for 1hr. Blots were washed

618 in PBST and incubated in horse radish peroxidase-conjugated goat α-rabbit antibody

619 (ThermoFisher) diluted 1:5,000 in PBST-milk and the signal was detected by ECL

620 chemiluminescence (ThermoFisher).

621

622 *Granular starch binding assays*

623 Granular starch-binding assays were conducted with potato starch (Bob's Red Mill), corn starch

624 (Sigma), wheat starch (Sigma), or Avicel (Fluka). Prior to use, all polysaccharides were washed

625 3x with an excess of assay buffer (20mM HEPES pH 7.0, 100mM NaCl) to remove soluble starch

626 and oligosaccharides and prepared as a 50mg/mL slurry. 1mg (corn) or 5mg (potato) of starch

627 slurry was aliquoted into 0.2mL tubes in triplicate, centrifuged at 2,000xg for 2 min and the

628 supernatant was carefully removed. 100μL of protein ranging from 0.5μM-10μM protein was

629 added to each starch and the tubes were agitated by end-over-end rotation at room temperature

630 for 1hr. After centrifugation at 2,000xg for 2min, 20μL of the supernatant was removed for

631 unbound protein concentration determination by absorbance at A280 using a ThermoFisher

632 NanodropOne with three replicate measurements per sample. The remaining 80μL of supernatant

633 was removed and set aside for SDS-PAGE gel analysis. The concentration of unbound protein

634 remaining in the supernatant was used to determine the μmoles of protein bound per gram of

635 starch which was plotted against the concentration of initial (free) protein to generate a binding

636 curve [31]. Overall affinity (K_d) and binding maximum (B_{max}) was determined via a one-site binding

637 model (specific binding) using GraphPad Prism version 9.2.0 for Windows (GraphPad Software,

638 San Diego, California USA, www.graphpad.com) [31].

639 To assess the remaining starch granules for bound protein, the granules were washed

640 three times with an excess of assay buffer by mixing and centrifugation, the final wash supernatant

641 was removed, and 100μL of Laemmli buffer containing 1M urea was added to the starch pellet to

642 denature any bound protein but keep the original volume consistent. To qualitatively determine

643 the amount of unbound and bound protein, 10μL each of the wash supernatant and solubilized

644 pellet fraction were run separately via SDS-PAGE. Bovine serum albumin was used as a negative

645 control and to confirm unbound protein was sufficiently washed from the starch granules.

646

647 *Polysaccharide Affinity PAGE*

648 Non-denaturing polyacrylamide gels with and without potato amylopectin (Sigma), corn

649 amylopectin (Sigma), potato amylose (Sigma), bovine liver glycogen (Sigma), pullulan (Sigma),

650 or dextran (Sigma) to a final concentration of 0.1% polysaccharide were cast. All polysaccharides

651 were autoclaved and amylose was solubilized by alkaline solubilization with 1M NaOH and acid

652 neutralization to pH 7 with HCl [60]. Sas6 protein samples were mixed with 6X loading dye lacking

653 SDS. Gels were run concurrently for 4 hours on ice and subsequently stained with Coomassie

654 (0.025% Coomassie blue R350, 10% acetic acid, and 45% methanol). Gels were imaged on a

655 Bio-Rad Gel Doc Go imaging system. The distance between each band and the top of the

656 separating gel were measured using ImageJ [61]. The ratio of the distance migrated by each

657 band was determined to the distance the BSA band traveled. Binding was considered positive if

658 the ratio was less 0.85 as previously described [62].

659

660 *Isothermal Titration Calorimetry*

661 All ITC experiments were carried out using a TA Instruments standard volume NanoITC. For each

662 experiment, 1300μL of 25μM protein was added to the sample cell and the reference cell was

663 filled with distilled water. The sample injection syringe was loaded with 250μL of the appropriate

664 ligand concentration (0.5mM - 5mM) to fully saturate the protein by the end of 25 injections of

665 10μls. Titrations were performed at 25°C with a stirring speed of 250 rpm. The resulting data were

666 modeled using TA Instruments NanoAnalyze software employing the pre-set models for

667 independent binding and blank (constant) to subtract the heat of dilution. For interactions with

668 high affinity (c-value at 25μM protein greater than 5), no alterations were made to the model. If

669 the calculated c value of an interaction fell below 5, the n value was set to 1 as indicated in the

670 figure legend following the guidance for modeling low affinity interactions [63]. For polysaccharide

671 titrations, curves were modeled by varying the substrate concentration until n=1 such that the K_d

672 represents the overall affinity for the construct [52].

32

673

674   *Protein Crystallization*

675   Crystallization conditions for α-cyclodextrin (2mM) bound (pdb 7UWW) and unliganded (pdb

676   7UWU) crystals of Sas6T were screened via 96-well sparse matrix screen (Peg Ion HT, Hampton

677   Research #HR2-139) in a sitting drop vapor diffusion experiment at room temperature. Screens

678   were set up using an Art Robbins Gryphon robot with 20mg/mL protein in a 3-well tray (Art

679   Robbins #102-0001-13) using protein-to-well solution ratios of 2:1, 1:1, and 1:2. Small crystals

680   were observed in 0.2M Potassium thiocyanate pH 7.0, 20% w/v Polyethylene glycol 3,350

681   (condition B2) and were further optimized by varying pH, PEG 3350 percentage, and potassium

682   thiocyanate concentration. Crystals were microseeded with a crystal seeding tool (Hampton) in a

683   sitting drop setup of 1.5μL drops with 2:1, 1:1, or 1:2 protein:well solution ratios. The optimal

684   crystallization solution contained 0.3M Potassium thiocyanate pH 7.0, 24% PEG 3350 and 1mM

685   Anderson−Evans polyoxotungstate $[TeW_6O_{24}]^{6-}$ (TEW) (Jena Biosciences #X-TEW-5) to improve

686   crystal diffraction. Prior to data collection, crystals were cryoprotected in a mixture of 80%

687   crystallization solution supplemented with 20% ethylene glycol then plunged into liquid nitrogen.

688   Crystallization conditions for maltodecaose-bound RbCBM74 structure (pdb 7UWV) were

689   generated from the construct lacking the CBM26 domain (BIg-*Rb*CBM74-BIg, residues 134-665)

690   using 96-well sparse matrix screens. A crystalline mass observed in 60% v/v Tacsimate pH 7.0,

691   0.1 M BIS-TRIS propane pH 7.0 (Hampton Salt-Rx HT-well H12 #HR2-136) was used to

692   microseed an optimized solution containing 30% Tacsimate, 0.1M HEPES pH 7.0 and 2mM

693   maltodecaose (CarboExpert). No additional cryo-protection was required prior to plunge freezing

694   into liquid nitrogen.

695

696   *Structure Determination and Refinement*

697   X-ray data were collected at the Life Sciences Collaborative Access Team (LS-CAT) at Argonne

698   National Laboratory's Advanced Photon Source (APS) in Argonne, IL. Data were processed at

699   APS using autoPROC with XDS for spot finding, indexing, and integration followed by Aimless for

700   scaling and merging [64-66]. Intrinsic sulfur SAD phasing was used to determine the structure of

701   Sas6T/α-cyclodextrin (7UWW) using AutoSol in Phenix [67, 68]. Those coordinates were then

702   used for molecular replacement in Phaser to determine the unliganded Sas6T (7UWU) and BIg-

703   *Rb*CBM74-BIg/G10 (7UWV) structures [69]. All three structures were refined via manual model

704   building in Coot and refinement in Phenix.refine [70, 71]. Metal ion identities were validated using

705   the web-based CheckMyMetal (CMM) tool [72] (https://cmm.minorlab.org/). Carbohydrate models

706   were validated using Privateer [73].

707

708   *SEC-SAXS experiment*

709   SAXS was performed at Biophysics Collaborative Access Team (BioCAT, beamline 18ID at APS)

710   with in-line size exclusion chromatography (SEC-SAXS) to separate the sample from aggregates

711   and other contaminants. Sample was loaded onto a Superdex 200 Increase 10/300 GL column

712   (Cytiva), which was run at 0.6ml/min by an AKTA Pure FPLC (GE) and the eluate after it passed

713   through the UV monitor was flown through the SAXS flow cell. The flow cell consists of a 1.0mm

714   ID quartz capillary with ~20μm walls. A coflowing buffer sheath is used to separate the sample

715   from the capillary walls, helping prevent radiation damage [74]. Scattering intensity was recorded

716   using a Pilatus3 X 1M (Dectris) detector which was placed 3.6m from the sample giving a q-range

717   of $0.003Å^{-1}$ to $0.35Å^{-1}$. 0.7 s exposures were acquired every 1s during elution and data was

718   reduced using BioXTAS RAW 2.1.1 [75]. Within RAW, the Volume of Correlation (V_C), molecular

719   weight, and oligomeric state were determined [76, 77]. Buffer blanks were created by averaging

720   regions flanking the elution peak and subtracted from exposures selected from the elution peak

721   to create the I(q) vs q curves used for subsequent analyses. The molecular weight was calculated

722   by comparison to known structures (Shape&Size) [38]. P(r) function was determined using GNOM

723   [39]. GNOM and Shape&Size are part of the ATSAS package (version 3.0) [78]. High resolution

724   structures were fit to the SAXS data using FoXS and flexibility in the high-resolution structures

725   was modeled against the Multi-FoXS data [40]. **Tables S2A-C** list sample, instrumentation, and

726   software for the SEC-SAXS experiment.

727

728   *Hydrogen–Deuterium eXchange Mass Spectrometry (HDX-MS)*

729   HDX-MS experiments were performed using a Synapt G2-SX HDMS system (Waters), similar to

730   previously reported [79]. Deuteration reactions were incubated at 20°C for 15s, 150s, 1500s, and

731   15,000s in triplicate. 3µL of BIg-*Rb*CBM74-BIg alone or in the presence of G10 were diluted with

732   57µL of deuterated labeling buffer. Nondeuterated data were acquired by dilution with protonated

733   buffer and fully deuterated data were prepared by dilution in 99% D_2O, 1% (v/v) formic acid) for

734   48h at room temperature. Samples were measured in triplicate using automated handling with a

735   PAL liquid handling system (LEAP), using randomized sequential collection with Chronos.

736   Following incubation, deuteration was quenched by mixing 50µL of the solution with 50µL

737   of 100mM phosphate, pH 2.5 at 0.3°C. Immediately after the samples were quenched, 95µL of

738   the sample was loaded onto an Acquity M-class UPLC (Waters) with sequential inline pepsin

739   digestion (Waters Enzymate BEH Pepsin column, 2.1mm × 30mm) for 3min at 15°C followed by

740   reverse phase purification (Acquity UPLC BEH C18 1.7µm at 0.2°C). Sample was loaded onto

741   the column equilibrated with 95% water, 5% acetonitrile, and 0.1% formic acid at a flow rate of

742   40µL/min. A 7min linear gradient (5%–35% acetonitrile) followed by a ramp and 2min block (85%

743   acetonitrile) was used for separation and directly continuously infused onto a Synapt XS using

744   Ion Mobility (Waters). [Glu1]-Fibrinopeptide B was used as a reference.

745   Data from nondeuterated samples were used for peptide identification with ProteinLynx

746   Global Server 3.0 (Waters). Full coverage of the protein was obtained, with the exception of the

747   region from residues 289-296, where peptides were not detected. The filtered peptide list and MS

748   data were imported into HDExaminer (Sierra Analytics) for deuterium uptake calculation using

749   both retention time and mobility matching.   Representative peptides were utilized for a final

750   cumulative sequence coverage of 91.4%. Normalized deuterium uptake data was calculated for

751   protein alone and with G10, and differential protection, defined as those regions with an average

752   of 5% difference in deuteration between states over the 150-15000s timepoints, were mapped

753   onto the crystal structure using PyMOL (Schrodinger).

754

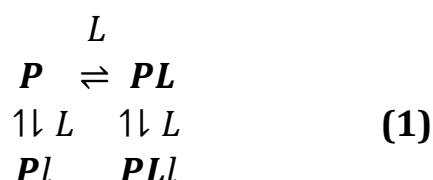755   *Native Mass Spectrometry (MS)*

756   Stock solutions of BIg-*Rb*CBM74-BIg and Sas6 were de-salted and solvent exchanged into

757   200mM ammonium acetate (pH 6.8 – 7.0) using Amicon Ultra-0.5mL centrifugal filters

758   (MilliporeSigma) with a 10kDa molecular weight cut-off. Ten consecutive washing steps were

759   performed to achieve sufficient desalting. The final concentrations of each protein stock solution

760   after desalting were estimated via UV absorbance at 280nm. A stock solution of G10 was

761   prepared by dissolving a known mass in 200mM ammonium acetate to achieve a final

762   concentration of 200µM. For native MS titration experiments used to quantify K_d values, the

763   concentration of protein was fixed at 5µM, and enough G10 was added to achieve final

764   concentrations of 0, 5, 25, 50, 100, and 150µM. Protein-G10 mixtures were then incubated at 4°C

765   overnight to achieve equilibration prior to native MS analysis.

766   All native binding experiments were performed using a Q Exactive Orbitrap MS with Ultra

767   High Mass Range (UHMR) platform (Thermo Fisher Scientific) [80]. Each sample (~3µM) was

768   transferred to a gold-coated borosilicate capillary needle (prepared in house), and ions were

769   generated via direct infusion using a nano-electrospray ionization (nESI) source operated in

770   positive mode. The capillary voltage was held at 1.2kV, the inlet capillary was heated to 250°C,

771   and the S-lens RF level was kept at 80. Low m/z detector optimization and high m/z transfer optics

772   were used, and the trapping gas pressure was set to 2. In-source trapping was enabled with the

773   desolvation voltage fixed at -25V for improved ion transmission and efficient salt adduct removal.

774   Transient times were set at 128ms (resolution of 25,000 at m/z 400), and 5 microscans were

775   combined into a single scan. A total of ~50 scans were averaged to produce the presented mass

776   spectra. All full scan data were acquired using a noise threshold of 0 to avoid pre-processing of

777   mass spectra. A total of three measurements for each ligand concentration were performed. Data

778   were then processed and deconvoluted using UniDec software [81].

779

780   *K_d Measurements by Native MS*.

781   We performed titration experiments for both BIg-*Rb*CBM74-BIg and Sas6T using G10 and

782   acquired modeled titration curves. Each bound state differed by ~1639 Da, which agrees with the

783   theoretical mass of G10. To obtain the binding constants, we summed the peak intensities of all

784   abundant charge states in our mass spectra. K_d values were calculated using the relative

785   intensities of unbound protein and each ligand bound species from the mass spectra as previously

786   described [82]. Briefly, the protein-ligand binding equilibrium of BIg-*Rb*CBM74-BIg with G10 in

787   solution can be described by the following reversible reaction:

788
789    $$\begin{array}{ccc} & L & \\ \boldsymbol{P} & \rightleftharpoons & \boldsymbol{PL} \\ \uparrow\downarrow L & & \uparrow\downarrow L \\ \boldsymbol{Pl} & & \boldsymbol{PLl} \end{array} \qquad (1)$$
790
791
792
793   where $\boldsymbol{L}$ is the ligand and $\boldsymbol{P}$ and $\boldsymbol{PL}$ are the free protein and protein with one specifically bound

794   ligand, respectively. BIg-*Rb*CBM74-BIg possesses one ligand-binding site, *Rb*CBM74. As the

795   concentration of ligand is increased, ligand molecules can bind nonspecifically during the nESI

796   process, generating artifactual peaks in the mass spectra corresponding to a two ligand-bound

797   complex. As the concentration of ligand is increased, ligand molecules can bind nonspecifically

798   during the nESI process, generating artifactual peaks in the mass spectra corresponding to a two

799   ligand-bound complex. Here, we presume that nonspecific binding arises equally for free protein

800   and that which possesses one specifically bound ligand represented by $\boldsymbol{Pl}$ and $\boldsymbol{PLl}$ in Eq. 1.

801   Based on these assumptions, the equations of mass balance and binding states can be described

802   the following system of equations:

37

803

$$c_p = [P] + ([PL] + [Pl]) + [PLl] \quad \textbf{(2a)}$$

804

805

$$c_L = [L] + ([PL] + [Pl]) + 2[PLl] \quad \textbf{(2b)}$$

806

807

$$K_d = \frac{[P][L]}{[PL]} \qquad \textbf{(2c)}$$

808

809

$$K_n = \frac{[P][L]}{[Pl]} = \frac{[PL][L]}{[PLl]} \qquad \textbf{(2d)}$$

810

811

812 where c_P and c_L represent the total concentrations of protein and ligand, respectively, and

813 concentrations in brackets represent those at equilibrium. K_d and K_n represent the dissociation

814 constants for specific and nonspecific binding steps, respectively. If 1) the peak intensities of free

815 protein and ligand-bound complexes are proportional to the abundances of those in solution and

816 2) the spray and detection efficiency of all species is the same, then the fractional intensities of

817 each species can be determined:

818

$$F_i = \frac{\sum_n I(PL_i^{n+})/n}{\sum_{i=0}^{2} \sum_n (PL_i^{n+})/n} \qquad \textbf{(3)}$$

819

820

821 Here, the fractional intensities are calculated as the sum of the intensities of main peak ions at all

822 charge states. Since a Fourier transform MS method is utilized, signal intensities are proportional

823 to both ion abundance and charge state. Therefore, the ion intensities are normalized for each

824 charge state, n [83, 84]. These fractional intensities can be calculated from the titration experiment

825 at each ligand concentration and can then be related to the equilibrium constants:

826

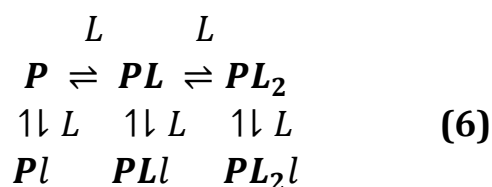$$F_0 = \frac{K_d K_n}{K_d K_n + [L](K_d + K_n) + [L]^2} \qquad \textbf{(4a)}$$

827

828

$$F_1 = \frac{[L](K_d + K_n)}{K_d K_n + [L](K_d + K_n) + [L]^2} \qquad \textbf{(4b)}$$

829

830

831 $$F_2 = \frac{[L]^2}{K_d K_n + [L](K_d + K_n) + [L]^2} \qquad \textbf{(4c)}$$

832

833 [L] can also be determined from nESI-MS titration data:

834

835 $$[L] = c_L - c_P(F_1 + 2F_2) \qquad \textbf{(5)}$$

836

837 [L] was then obtained at each ligand concentration and applied to the Eqs. 4a-c. Equations 4a-b

838 were then fitted to experimental fractional intensities using nonlinear least-squares curve fitting

839 using the *lsqnonlin.m.* function in MATLAB. A more detailed derivation of these equations is

840 provided elsewhere [82], along with the approach utilized for Sas6 which possesses two sites for

841 specific binding (*Rb*CBM74 and *Rb*CBM26) and exhibits a third nonspecific bound state as shown

842 in Eq. 6.

843

844
845
846
847

$$
\begin{array}{ccc}
L & L & \\
P \rightleftharpoons PL \rightleftharpoons PL_2 & \\
\updownarrow L \quad \updownarrow L \quad \updownarrow L & \qquad \textbf{(6)} \\
Pl \quad PLl \quad PL_2l &
\end{array}
$$

848
849

850 *Sequence collection*

851 Amino acid sequences of CBM74 modules were collected according to information in the CAZy

852 database (http://www.cazy.org/) yielding 29 sequences (CAZy update: March 2022) [85]. This set

853 was subsequently completed with sequences of hypothetical CBM74s based on protein BLAST

854 searches (https://blast.ncbi.nlm.nih.gov/Blast.cgi) using the CBM74 sequences from Sas6 of

855 *Ruminococcus bromii* (GenBank Acc. No.: PKD32096.1) and the GH13_32 α-amylase from

856 *Microbacterium aurum* (GenBank Acc. No.: AKG25402.1) as queries [20, 86, 87]. In total, three

857 searches with each query sequence were performed, limiting the searched databases to

858 taxonomy kingdoms of Bacteria, Archaea and Eucarya (with no relevant results for the latter two).

859 To capture a wide spectrum of organisms harboring a CBM74 module, one non-redundant amino

860 acid sequence was selected to represent each species and/or bacterial strain. The BLAST

39

861   searches thus yielded 93 additional CBM74 sequences of bacterial origin; the last sequence taken

862   being the CBM74 module of a putative α-amylase from uncultured *Eubacterium* sp. (GenBank:

863   SCJ65691.1; E-value: 3e-39). That preliminary set of 122 sequences was reduced by eliminating

864   23 sequences due to their redundancy and/or incompleteness of the CBM74 module. The final

865   set of CBM74 modules consisted of 99 sequences (**Extended Data Table 2**). All sequences were

866   retrieved   from   the   GenBank   (https://www.ncbi.nlm.nih.gov/genbank/)   and/or   UniProt

867   (https://www.uniprot.org/) databases [88, 89].

868

869   *Sequence comparison and evolutionary analysis*

870   The alignment of 99 CBM74 modules from the final set was performed using the program Clustal-

871   Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) [90]. Only a subtle manual tuning of the

872   computer-produced alignment was necessary to perform to maximize sequence similarities. The

873   evolutionary tree of these 99 sequences was calculated by a maximum-likelihood method (on the

874   final alignment including the gaps) using the WAG substitution model and the bootstrapping

875   procedure with 500 bootstrap trials implemented in the MEGA-X package [91-93]. The calculated

876   tree file was displayed with the program iTOL [94] (https://itol.embl.de/). From both the alignment

877   and the tree of all 99 sequences, a sample of 33 representative CBM74s was selected for a

878   simplified alignment and tree. The structural comparison was created using the above-mentioned

879   alignment in conjunction with the web-based CONSURF tool [54-56].

880

881   *Data availability*

882   The X-ray structures and diffraction data reported in this paper have been deposited in the Protein

883   Data Bank under the accession codes 7UWU, 7UWV and 7UWW. The SAXS data are deposited

884   in the small angle x-ray scattering database (SASDB) under the accession code SASDPE2 [95].

885   All mass spectrometry data will be made available upon request.

886

909

910   **Author contributions**

41

911   N.M.K. and A.L.P. conceptualization;

912   A.L.P., F.M.C., R. V-V., K.M.A., F.M., T.C., Z.W., J.H., C.W.V.K, S.J., B.T.R., and N.M.K. data

913   curation;

914   A.L.P., F.M.C., R.V-V., K.M.A., F.M., T.C., Z.W., J.H., C.W.V.K, S.J., B.T.R., and N.M.K. formal

915   analysis and data interpretation;

916   N.M.K., A.L.P., C.W.V.K, S.J., B.T.R., Z.W., and J.H. funding acquisition;

917   A.L.P., F.M.C., R. V-V., K.M.A., F.M., T.C., Z.W., J.H., C.W.V.K, and S.J. investigation;

918   A.L.P., F.M.C., R. V-V., K.M.A., F.M., T.C., Z.W., J.H., C.W.V.K, S.J., B.T.R., and N.M.K.

919   methodology;

920   A.L.P., F.M.C., R.V-V., C.W.V.K , S.J. and N.M.K original draft;

921   A.L.P., F.M.C., R.V-V., F.M., J.H., C.W.V.K, S.J., B.T.R., and N.M.K. writing- review and editing;

922   N.M.K., J.H., C.W.V.K, S.J., and B.T.R. supervision;

923   A.L.P., F.M.C., R.V-V., T.C., and S.J. visualization.

924

Conflict of interest

926   The authors declare that they have no conflicts of interest with the contents of this article.

927

928

929   1.   Salminen, S., E. Isolauri, and T. Onnela, *Gut Flora in Normal and Disordered States.*
930   Chemotherapy, 1995. **41(suppl 1)**(Suppl. 1): p. 5-15.
931   2.   Gibson, G.R. and M.B. Roberfroid, *Dietary modulation of the human colonic microbiota:*
932   *introducing the concept of prebiotics.* J Nutr, 1995. **125**(6): p. 1401-12.
933   3.   Cummings, J.H. and G.T. Macfarlane, *The control and consequences of bacterial*
934   *fermentation in the human colon.* J Appl Bacteriol, 1991. **70**(6): p. 443-59.
935   4.   Backhed, F., et al., *Host-bacterial mutualism in the human intestine.* Science, 2005.
936   **307**(5717): p. 1915-20.
937   5.   Wu, X., et al., *Effects of the intestinal microbial metabolite butyrate on the development of*
938   *colorectal cancer.* J Cancer, 2018. **9**(14): p. 2510-2517.
939   6.   Zaman, S.A. and S.R. Sarbini, *The potential of resistant starch as a prebiotic.* Critical
940   Reviews in Biotechnology, 2016. **36**(3): p. 578-584.
941   7.   Bertoft, E., *Understanding Starch Structure: Recent Progress.* 2017. **7**(3): p. 56.
942   8.   Pérez, S. and E. Bertoft, *The molecular structures of starch components and their*
943   *contribution to the architecture of starch granules: A comprehensive review.* 2010. **62**(8):
944   p. 389-420.
945   9.   Cerqueira, F.M., et al., *Starch Digestion by Gut Bacteria: Crowdsourcing for Carbs.* Trends
946   Microbiol, 2020. **28**(2): p. 95-108.
947   10.   Ze, X., et al., *Ruminococcus bromii is a keystone species for the degradation of resistant*
948   *starch in the human colon.* ISME J, 2012. **6**(8): p. 1535-43.
949   11.   Jung, D.H., et al., *Bifidobacterium adolescentis P2P3, a Human Gut Bacterium Having*
950   *Strong Non-Gelatinized Resistant Starch-Degrading Activity.* J Microbiol Biotechnol, 2019.
951   **29**(12): p. 1904-1915.
952   12.   Teichmann, J. and D.W. Cockburn, *In vitro Fermentation Reveals Changes in Butyrate*
953   *Production Dependent on Resistant Starch Source and Microbiome Composition.* 2021.
954   **12**(976).
955   13.   Duranti, S., et al., *Genomic characterization and transcriptional studies of the starch-*
956   *utilizing strain Bifidobacterium adolescentis 22L.* Appl Environ Microbiol, 2014. **80**(19): p.
957   6080-90.
958   14.   Belenguer, A., et al., *Two routes of metabolic cross-feeding between Bifidobacterium*
959   *adolescentis and butyrate-producing anaerobes from the human gut.* Appl Environ
960   Microbiol, 2006. **72**(5): p. 3593-9.
961   15.   Venkataraman, A., et al., *Variable responses of human microbiomes to dietary*
962   *supplementation with resistant starch.* Microbiome, 2016. **4**(1): p. 33.
963   16.   Baxter, N.T., et al., *Dynamics of Human Gut Microbiota and Short-Chain Fatty Acids in*
964   *Response to Dietary Interventions with Three Fermentable Fibers.* MBio, 2019. **10**(1).
965   17.   Ze, X., et al., *Unique Organization of Extracellular Amylases into Amylosomes in the*
966   *Resistant Starch-Utilizing Human Colonic Firmicutes Bacterium Ruminococcus bromii.*
967   mBio, 2015. **6**(5): p. e01058-15.
968   18.   Smith, S.P. and E.A. Bayer, *Insights into cellulosome assembly and dynamics: from*
969   *dissection to reconstruction of the supramolecular enzyme complex.* Curr Opin Struct Biol,
970   2013. **23**(5): p. 686-94.
971   19.   Bayer, E.A., E. Morag, and R. Lamed, *The cellulosome--a treasure-trove for*
972   *biotechnology.* Trends Biotechnol, 1994. **12**(9): p. 379-86.
973   20.   Mukhopadhya, I., et al., *Sporulation capability and amylosome conservation among*
974   *diverse human colonic and rumen isolates of the keystone starch-degrader Ruminococcus*
975   *bromii.* Environ Microbiol, 2018. **20**(1): p. 324-336.
976   21.   Janecek, S., et al., *Starch-binding domains as CBM families-history, occurrence,*
977   *structure, function and evolution.* Biotechnol Adv, 2019. **37**(8): p. 107451.

43

978   22.   Valk, V., et al., *Carbohydrate-binding module 74 is a novel starch-binding domain*
979   *associated with large and multidomain alpha-amylase enzymes.* FEBS J, 2016. **283**(12):
980   p. 2354-68.
981   23.   Dobranowski, P.A. and A. Stintzi, *Resistant starch, microbiome, and precision modulation.*
982   Gut Microbes, 2021. **13**(1): p. 1926842.
983   24.   Ravi, A., et al., *Hybrid metagenome assemblies link carbohydrate structure with function*
984   *in the human gut microbiome.* Communications Biology, 2022. **5**(1): p. 932.
985   25.   Xu, J., et al., *Metatranscriptomic analysis of colonic microbiota's functional response to*
986   *different dietary fibers in growing pigs.* Animal Microbiome, 2021. **3**(1): p. 45.
987   26.   Zhang, H., et al., *dbCAN2: a meta server for automated carbohydrate-active enzyme*
988   *annotation.* Nucleic Acids Research, 2018. **46**(W1): p. W95-W101.
989   27.   Lombard, V., et al., *The carbohydrate-active enzymes database (CAZy) in 2013.* Nucleic
990   Acids Res, 2014. **42**(Database issue): p. D490-5.
991   28.   Blum, M., et al., *The InterPro protein families and domains database: 20 years on.* Nucleic
992   Acids Res, 2021. **49**(D1): p. D344-D354.
993   29.   Cerqueira, F.M., et al., *Sas20 is a highly flexible starch-binding protein in the*
994   *Ruminococcus bromii cell-surface amylosome.* J Biol Chem, 2022: p. 101896.
995   30.   Fontes, C.M. and H.J. Gilbert, *Cellulosomes: highly efficient nanomachines designed to*
996   *deconstruct plant cell wall complex carbohydrates.* Annu Rev Biochem, 2010. **79**: p. 655-
997   81.
998   31.   Boraston, A.B., et al., *A structural and functional analysis of alpha-glucan recognition by*
999   *family 25 and 26 carbohydrate-binding modules reveals a conserved mode of starch*
1000   *recognition.* J Biol Chem, 2006. **281**(1): p. 587-98.
1001   32.   Matsui, M., M. Kakuta, and A. Misaki, *Comparison of the Unit-chain Distributions of*
1002   *Glycogens from Different Biological Sources, Revealed by Anion Exchange*
1003   *Chromatography.* Bioscience, Biotechnology, and Biochemistry, 1993. **57**(4): p. 623-627.
1004   33.   Brewer, M.K. and M.S. Gentry, *Brain Glycogen Structure and Its Associated Proteins:*
1005   *Past, Present and Future.* Adv Neurobiol, 2019. **23**: p. 17-81.
1006   34.   Singh, R.S., G.K. Saini, and J.F. Kennedy, *Pullulan: Microbial sources, production and*
1007   *applications.* Carbohydr Polym, 2008. **73**(4): p. 515-31.
1008   35.   Khalikova, E., P. Susi, and T. Korpela, *Microbial dextran-hydrolyzing enzymes:*
1009   *fundamentals and applications.* Microbiol Mol Biol Rev, 2005. **69**(2): p. 306-25.
1010   36.   Valk, V., M.v.d.K. Rachel, and L. Dijkhuizen, *The evolutionary origin and possible*
1011   *functional roles of FNIII domains in two Microbacterium aurum B8.A granular starch*
1012   *degrading enzymes, and in other carbohydrate acting enzymes.* Amylase, 2017. **1**(1): p.
1013   1-11.
1014   37.   Krissinel, E. and K. Henrick, *Inference of macromolecular assemblies from crystalline*
1015   *state.* J Mol Biol, 2007. **372**(3): p. 774-97.
1016   38.   Franke, D., C.M. Jeffries, and D.I. Svergun, *Machine Learning Methods for X-Ray*
1017   *Scattering Data Analysis from Biomacromolecular Solutions.* Biophysical Journal, 2018.
1018   **114**(11): p. 2485-2492.
1019   39.   Svergun, D., *Determination of the regularization parameter in indirect-transform methods*
1020   *using perceptual criteria.* Journal of Applied Crystallography, 1992. **25**(4): p. 495-503.
1021   40.   Schneidman-Duhovny, D., et al., *FoXS, FoXSDock and MultiFoXS: Single-state and multi-*
1022   *state structural modeling of proteins and their complexes based on SAXS profiles.* Nucleic
1023   Acids Res, 2016. **44**(W1): p. W424-9.
1024   41.   Holm, L., *Using Dali for Protein Structure Comparison.* Methods Mol Biol, 2020. **2112**: p.
1025   29-42.
1026   42.   Notenboom, V., et al., *Crystal structures of the family 9 carbohydrate-binding module from*
1027   *Thermotoga maritima xylanase 10A in native and ligand-bound forms.* Biochemistry, 2001.
1028   **40**(21): p. 6248-56.

43.   Milles, L.F., et al., *Calcium stabilizes the strongest protein fold.* Nat Commun, 2018. **9**(1): p. 4764.

44.   Zheng, H., et al., *CheckMyMetal: a macromolecular metal-binding validation tool.* Acta Crystallogr D Struct Biol, 2017. **73**(Pt 3): p. 223-233.

45.   Strynadka, N.C.J. and M.N.G. James, *Towards an understanding of the effects of calcium on protein structure and function.* Current Opinion in Structural Biology, 1991. **1**(6): p. 905-914.

46.   Holm, L., *DALI and the persistence of protein shape.* 2020. **29**(1): p. 128-140.

47.   Rodriguez-Sanoja, R., et al., *A single residue mutation abolishes attachment of the CBM26 starch-binding domain from Lactobacillus amylovorus alpha-amylase.* J Ind Microbiol Biotechnol, 2009. **36**(3): p. 341-6.

48.   Baldwin, P.M., M.C. Davies, and C.D. Melia, *Starch granule surface imaging using low-voltage scanning electron microscopy and atomic force microscopy.* Int J Biol Macromol, 1997. **21**(1-2): p. 103-7.

49.   Szymonska, J. and F. Krok, *Potato starch granule nanostructure studied by high resolution non-contact AFM.* Int J Biol Macromol, 2003. **33**(1-3): p. 1-7.

50.   Park, H., S. Xu, and K. Seetharaman, *A novel in situ atomic force microscopy imaging technique to probe surface morphological features of starch granules.* Carbohydr Res, 2011. **346**(6): p. 847-53.

51.   Lineback, D.R., *Current Concepts of Starch Structure and Its Impact on Properties.* Journal of the Japanese Society of Starch Science, 1986. **33**(1): p. 80-88.

52.   Abbott, D.W. and A.B. Boraston, *Chapter eleven - Quantitative Approaches to The Analysis of Carbohydrate-Binding Module Function*, in *Methods in Enzymology*, H.J. Gilbert, Editor. 2012, Academic Press. p. 211-231.

53.   Soper, M.T., et al., *Amyloid-beta-neuropeptide interactions assessed by ion mobility-mass spectrometry.* Phys Chem Chem Phys, 2013. **15**(23): p. 8952-61.

54.   Ashkenazy, H., et al., *ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules.* Nucleic Acids Res, 2016. **44**(W1): p. W344-50.

55.   Ashkenazy, H., et al., *ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W529-33.

56.   Celniker, G., et al., *ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function.* 2013. **53**(3-4): p. 199-206.

57.   Jung, D.H., et al., *The presence of resistant starch-degrading amylases in Bifidobacterium adolescentis of the human gut.* Int J Biol Macromol, 2020. **161**: p. 389-397.

58.   Rees, D.A. and E.J. Welsh, *Secondary and Tertiary Structure of Polysaccharides in Solutions and Gels.* 1977. **16**(4): p. 214-224.

59.   Cameron, E.A., et al., *Multidomain Carbohydrate-binding Proteins Involved in Bacteroides thetaiotaomicron Starch Metabolism.* J Biol Chem, 2012. **287**(41): p. 34614-25.

60.   Hillmann, G., *Measurement by End-point Determination on Paper*, in *Methods of Enzymatic Analysis (Second Edition)*, H.U. Bergmeyer, Editor. 1974, Academic Press. p. 903-909.

61.   Schneider, C.A., W.S. Rasband, and K.W. Eliceiri, *NIH Image to ImageJ: 25 years of image analysis.* Nature Methods, 2012. **9**(7): p. 671-675.

62.   Cockburn, D.W., et al., *Novel carbohydrate binding modules in the surface anchored alpha-amylase of Eubacterium rectale provide a molecular rationale for the range of starches used by this organism in the human gut.* Mol Microbiol, 2018. **107**(2): p. 249-264.

63.   Turnbull, W.B. and A.H. Daranas, *On the value of c: can low affinity systems be studied by isothermal titration calorimetry?* J Am Chem Soc, 2003. **125**(48): p. 14859-66.

1079   64.   Vonrhein, C., et al., *Data processing and analysis with the autoPROC toolbox.* Acta
1080   Crystallographica Section D, 2011. **67**(4): p. 293-302.
1081   65.   Kabsch, W., *Xds.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 2): p. 125-32.
1082   66.   Evans, P.R. and G.N. Murshudov, *How good are my data and what is the resolution?* Acta
1083   Crystallogr D Biol Crystallogr, 2013. **69**(Pt 7): p. 1204-14.
1084   67.   El Omari, K., et al., *Pushing the limits of sulfur SAD phasing: de novo structure solution of
1085   the N-terminal domain of the ectodomain of HCV E1.* Acta Crystallogr D Biol Crystallogr,
1086   2014. **70**(Pt 8): p. 2197-203.
1087   68.   Terwilliger, T.C., et al., *Decision-making in structure solution using Bayesian estimates of
1088   map quality: the PHENIX AutoSol wizard.* Acta Crystallographica Section D, 2009. **65**(6):
1089   p. 582-601.
1090   69.   McCoy, A.J., et al., *Phaser crystallographic software.* Journal of Applied Crystallography,
1091   2007. **40**(4): p. 658-674.
1092   70.   Emsley, P. and K. Cowtan, *Coot: model-building tools for molecular graphics.* Acta
1093   Crystallogr D Biol Crystallogr, 2004. **60**(Pt 12 Pt 1): p. 2126-32.
1094   71.   Afonine, P.V., et al., *Towards automated crystallographic structure refinement with
1095   phenix.refine.* Acta Crystallogr D Biol Crystallogr, 2012. **68**(Pt 4): p. 352-67.
1096   72.   Zheng, H., et al., *Validation of metal-binding sites in macromolecular structures with the
1097   CheckMyMetal web server.* Nat Protoc, 2014. **9**(1): p. 156-70.
1098   73.   Agirre, J., et al., *Privateer: software for the conformational validation of carbohydrate
1099   structures.* Nature Structural & Molecular Biology, 2015. **22**(11): p. 833-834.
1100   74.   Kirby, N., et al., *Improved radiation dose efficiency in solution SAXS using a sheath flow
1101   sample environment.* Acta crystallographica. Section D, Structural biology, 2016. **72**(Pt
1102   12): p. 1254-1266.
1103   75.   Hopkins, J.B., R.E. Gillilan, and S. Skou, *BioXTAS RAW: improvements to a free open-
1104   source program for small-angle X-ray scattering data reduction and analysis.* J Appl
1105   Crystallogr, 2017. **50**(Pt 5): p. 1545-1553.
1106   76.   Rambo, R.P. and J.A. Tainer, *Accurate assessment of mass, models and resolution by
1107   small-angle scattering.* Nature, 2013. **496**(7446): p. 477-481.
1108   77.   Piiadov, V., et al., *SAXSMoW 2.0: Online calculator of the molecular weight of proteins in
1109   dilute solution from experimental SAXS data measured on a relative scale.* Protein Sci,
1110   2019. **28**(2): p. 454-463.
1111   78.   Manalastas-Cantos, K., et al., *ATSAS 3.0: expanded functionality and new tools for small-
1112   angle scattering data analysis.* Journal of Applied Crystallography, 2021. **54**(1): p. 343-
1113   355.
1114   79.   Murphy, R.D., et al., *The Toxoplasma glucan phosphatase TgLaforin utilizes a distinct
1115   functional mechanism that can be exploited by therapeutic inhibitors.* Journal of Biological
1116   Chemistry, 2022. **298**(7): p. 102089.
1117   80.   van de Waterbeemd, M., et al., *High-fidelity mass analysis unveils heterogeneity in intact
1118   ribosomal particles.* Nature Methods, 2017. **14**(3): p. 283-286.
1119   81.   Marty, M.T., et al., *Bayesian deconvolution of mass and ion mobility spectra: from binary
1120   interactions to polydisperse ensembles.* Anal Chem, 2015. **87**(8): p. 4370-6.
1121   82.   Gulbakan, B., et al., *Native Electrospray Ionization Mass Spectrometry Reveals Multiple
1122   Facets of Aptamer-Ligand Interactions: From Mechanism to Binding Constants.* J Am
1123   Chem Soc, 2018. **140**(24): p. 7486-7497.
1124   83.   Wang, W., E.N. Kitova, and J.S. Klassen, *Influence of solution and gas phase processes
1125   on protein-carbohydrate binding affinities determined by nanoelectrospray Fourier
1126   transform ion cyclotron resonance mass spectrometry.* Anal Chem, 2003. **75**(19): p. 4945-
1127   55.
1128   84.   Báez Bolivar, E.G., et al., *Submicron Emitters Enable Reliable Quantification of Weak
1129   Protein–Glycan Interactions by ESI-MS.* Analytical Chemistry, 2021. **93**(9): p. 4231-4239.

1130 85. Drula, E., et al., *The carbohydrate-active enzyme database: functions and literature.*
1131   Nucleic Acids Res, 2022. **50**(D1): p. D571-D577.
1132 86. Valk, V., et al., *Degradation of Granular Starch by the Bacterium Microbacterium aurum*
1133   *Strain B8.A Involves a Modular alpha-Amylase Enzyme System with FNIII and CBM25*
1134   *Domains.* Appl Environ Microbiol, 2015. **81**(19): p. 6610-20.
1135 87. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-
1136   10.
1137 88. Sayers, E.W., et al., *GenBank.* Nucleic Acids Res, 2021. **49**(D1): p. D92-D96.
1138 89. UniProt, C., *UniProt: the universal protein knowledgebase in 2021.* Nucleic Acids Res,
1139   2021. **49**(D1): p. D480-D489.
1140 90. Sievers, F., et al., *Fast, scalable generation of high-quality protein multiple sequence*
1141   *alignments using Clustal Omega.* Mol Syst Biol, 2011. **7**: p. 539.
1142 91. Whelan, S. and N. Goldman, *A general empirical model of protein evolution derived from*
1143   *multiple protein families using a maximum-likelihood approach.* Mol Biol Evol, 2001. **18**(5):
1144   p. 691-9.
1145 92. Felsenstein, J., *Confidence Limits on Phylogenies: An Approach Using the Bootstrap.*
1146   Evolution, 1985. **39**(4): p. 783-791.
1147 93. Kumar, S., et al., *MEGA X: Molecular Evolutionary Genetics Analysis across Computing*
1148   *Platforms.* Mol Biol Evol, 2018. **35**(6): p. 1547-1549.
1149 94. Letunic, I. and P. Bork, *Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree*
1150   *display and annotation.* Bioinformatics, 2007. **23**(1): p. 127-8.
1151 95. Kikhney, A.G., et al., *SASBDB: Towards an automatically curated and validated repository*
1152   *for biological scattering data.* Protein Sci, 2020. **29**(1): p. 66-75.

1153

Extended Data

1154

Extended Table 1A: Average masses assigned to native mass spectrometry peaks

P	0 bound (Da)	1 bound (Da)	2 bound (Da)	3 bound (Da)	Av Diff. (Da)[a]
BIg-*Rb*CBM74-Big	57635.3 ± 0.8	59274.7 ± 0.6	60913.8 ± 0.2	N/A	1639.4 ± 0.8
Sas6	69064.9 ± 0.6	70704.3 ± 0.5	72343.8 ± 1.8	73982.1 ± 0.6	1639.3 ± 1.0

[a]Average difference between bound states across all ligand concentrations.

Extended Table 1B: Binding parameters determined by Native Mass Spectrometry

P	Charge States	K_{d1} (μM)	K_{d2} (μM)	K_n (μM)[a]	SSR[b]
BIg-*Rb*CBM74-Big	11+ to 14+	3.8 ± 0.5	N/A	1154.4 ± 378.0	0.0111
Sas6	13+ to 15+	3.4 ± 0.5	165.6 ± 38.8	782.6 ± 647.9	0.0170

[a]K_n - dissociation constant for nonspecific binding step during nESI. Values reported with 95% confidence interval.
[b]SSR - sum of squared residuals

1155

1156

Extended Table 2: List of 99 selected CBM74 sequences. [a]

No.	Protein	Subfamily	Source material	Source organism	GenBank	UniProt	Length	CBM74	Additional CBMs*	Additional CBMs**
1.	pAAMY	GH13_28	Korean adult feces (HG)	*Bifidobacterium adolescentis*	AZH71984.1	A0A809K2S0	1462	657-978	2x CBM25; CBM26	-
2.	pAAMY	GH13_28	Human feces (HG)	*Bifidobacterium adolescentis*	AJE06470.1	A0A0B5BPU9	1432	621-942	2x CBM25; CBM26	
3.	pAAMY	-	Human fecal sample (HG)	*Bifidobacterium ruminantium*	MBU91112168.1	-	1200	636-957	-	CBM26
4.	pAAMY	GH13_28	Human feces (HG)	*Bifidobacterium angulatum*	AMK57563.1	A0A126SUD8	1527	629-950	2x CBM20; CBM26	-
5.	pAAMY	-	-	*Bifidobacterium merycicum*	SHE84896.1	UPI0009218F34	1186	657-978	-	CBM26
6.	pAAMY	-	Hamster dental plaque (RoG)	*Bifidobacterium tsurumiense*	KFJ06122.1	A0A087E1C1	1361	650-971	-	CBM25
7.	pAAMY	GH13_28	-	*Bifidobacterium pseudocatenulatum*	BAR04166.1	UPI0001847757	1434	621-942	CBM20; CBM26	-
8.	pAAMY	GH13_28	*Sus scrofa* cecum (PG)	*Bifidobacterium pseudolongum*	UBZ05046.1	-	1530	644-956	CBM25	-
9.	pAAMY	GH13_28	Rumen fluid of *Bos taurus* (\)	*Bifidobacterium choerinum*	ATU19782.1	A0A2D8D2U4	1613	644-956	CBM25	-
10.	pAAMY	-	Human adult intestine (HG)	*Bifidobacterium gallicum*	KFI59086.1	A0A087AJX6	1612	648-960	-	CBM25
11.	pAAMY	-	*Dolichotis patagonum* fecal samples (RoG)	*Bifidobacterium dolichotidis*	RSX55520.1	A0A430FRN0	1228	635-946	-	CBM25
12.	pAAMY	-	*Castor fiber* feces (RoG)	*Bifidobacterium castoris*	RSX43996.1	A0A430FI82	1439	633-944	-	CBM25
13.	pAAMY	-	-	*Pseudoscardovia suis*	PJ62656.1	UPI000CB6C43C	1455	621-934	-	CBM26
14.	pAAMY	-	*Sus scrofa* digestive tract (PG)	*Pseudoscardovia radai*	OZG53567.1	A0A261J864	2041	624-937	-	CBM26
15.	pAAMY	-	Wastewater (W)	*Bifidobacterium minimum*	KFI73654.1	A0A087DEK4	1457	631-943	-	2x CBM25; CBM26
16.	pAAMY	-	Swine feces (PG)	*Bifidobacterium thermophilum*	KFJ07018.1	UPI000505012E	1642	604-917	-	CBM26
17.	pAAMY	-	Feces of tamarin (PrG)	*Bifidobacterium goeldii*	RSX51755.1	A0A430FFT3	1522	632-946	-	2x CBM20; CBM25
18.	pAAMY	-	-	*Bifidobacterium pullorum*	WP_044645411.1	UPI001956DCA0	1251	626-939	-	CBM25
19.	pAAMY	-	-	*Galliscardovia ingluviei*	GGI14461.1	UPI00016664DC	1473	650-963	-	3x CBM25
20.	HYPO	-	Human gut (HG)	*Prevotella* sp.	MBD9271786.1	UPI00DD49E24	677	373-677	-	CBM26
21.	pAAMY	-	Human feces (HG)	*Prevotella hominis*	TFH82515.1	A0A4Y8VPY7	1095	791-1095	-	CBM26
22.	pAAMY	-	Human gut (HG)	*Prevotella* sp. CAG:386	CDC29131.1	R6QE37	1082	777-1082	-	CBM26
23.	pAAMY	-	Human fecal samples (HG)	*Prevotella copri*	MBV34B028.1	-	1092	787-1092	-	CBM26
24.	pDOC	-	Equine fecal microbiome (EG)	*Oscillospiraceae bacterium*	MBQ0097748.1	-	575	45-344	-	CBM26
25.	pDOC	-	*Gallus gallus* gut (CG)	*Candidatus Scatovimonas merdigallinarum*	HIQ81347.1	-	714	235-533	-	CBM26
26.	pDOC	-	Human gut (HG)	*Eubacterium* sp. CAG:202	CDC03229.1	R6N48	671	175-505	-	-
27.	pDOC	-	Human feces (HG)	*Eubacterium* sp. OM08-24	RGM091321	A0A374UMY1	737	241-571	-	CBM26
28.	DOC	-	Human gut (HG)	*Ruminococcus bromii*	CBL15887.1	UPI000CD4E31	734	242-572	CBM26	-
29.	pDOC	-	Rumen fluid of *Bos taurus* (RG)	*Ruminococcus* sp. JE7A12	QCT06902.1	A0A4P8XW45	730	240-570	CBM26	-
30.	pDOC	-	*Gallus gallus* gut (CG)	*Candidatus Scatovicinus merdipullorum*	HIR02602.1	-	949	435-759	-	CBM26
31.	pDOC	-	Goat gastrointestinal tract (RG)	*Clostridia bacterium*	MBQ2687794.1	-	985	431-752	-	CBM26
32.	pAAMY	-	Goat gastrointestinal tract (RG)	*Succinimonas* sp.	MBQ3681472.1	-	1799	543-836	-	CBM26
33.	pAAMY	-	-	*Succinimonas amylolytica*	WP_019900730.1	UPI000380E766	1786	533-826	-	3x CBM26
34.	pAAMY	-	Roe deer gastrointestinal tract (RG)	*Ruminobacter* sp.	MBR19249241	-	2051	808-1101	-	CBM26
35.	pAAMY	-	Dairy cattle gastrointestinal tract (RG)	*Succinivibrionaceae bacterium*	MBQ5525197.1	-	2157	941-1232	-	-
36.	pAAMY	-	-	*Ruminobacter amylophilus*	SFP12024.1	A0A662ZIG8	1959	721-1015	-	CBM26
37.	pAAMY	-	-	*Vibrio vulnificus*	MBN81093 19.1	UPI0019D43A75	2050	694-982	-	CBM26
38.	pAAMY	-	Cattle (RG)	*Vibrio navarrensis*	MBE4579448.1	UPI001869786D	2476	687-975	-	CBM26
39.	pAAMY	-	-	*Vibrio cincinnatiensis*	WP_281303 12.1	A0A430FF83	2465	694-983	-	CBM26
40.	pDOC	-	anaerobic digestion of organic wastes under variable temperature conditions and feedstock	*Fusobacteria bacterium*	NLK631981	A0A7X8J1U3	599	307-599	-	-
41.	pDOC	-	-	*Oceania marismortui*	TDX525251	A0A4R8H0E0	804	508-804	-	-
42.	HYPO	-	Sheep gastrointestinal tract (RG)	*Spirochaetales bacterium*	MBR23173211	-	838	533-838	-	-
43.	HYPO	-	Water deer gut (RG)	*Treponema* sp.	MBP3562602.1	UPI001B74B407	879	573-879	-	-
44.	pAAMY	-	-	*Microbacterium oxydans*	PFG94111	A0A2A9DSR8	1398	1098-1398	-	2x CBM25
45.	pAAMY	-	-	*Microbacterium lindanitolerans*	TQO224191	UPI00117BBA3D	1398	1098-1398	-	2x CBM25
46.	pAAMY	-	-	*Microbacterium flavidimviringii*	WP_160982566.1	UPI00041DC7DD	1401	1103-1401	-	2x CBM25
47.	pAAMY	-	-	*Arthrobacter pigmenti*	NKC24392.1	A0A846RX86	1402	1103-1402	-	2x CBM25
48.	pAAMY	-	-	*Microbacterium* sp. A18JL241	WP_194420691.1	UPI0018893ADD	1124	825-1124	-	2x CBM25
49.	pAAMY	-	-	*Agromyces atrinae*	NYD67366.1	A0A4Q2M4N3	1310	1010-1310	-	CBM25
50.	HYPO	-	Larvae of insect *Trypoxylus dichotomus* (I)	*Microbacterium* sp. BWT-G7	MCC20335501	-	361	61-361	-	-
51.	pAAMY	-	-	*Arthrobacter luteohandii*	MBP24127 5.1	UPI001D5C5EE1	2095	983-1283	-	3x CBM25
52.	AAMY	GH13_32	Wastewater treatment plant from a potato starch factory (W)	*Microbacterium aurum*	AKG25402.1	A0A0G2U4B5	1417	1116-1417	2x CBM25	-
53.	HYPO	-	Wastewater (W)	*Aeromonadaceae bacterium*	MBP8172241.1	UPI001B40 E6C6	701	406-701	-	-
54.	pAAMY	GH13_32	Hydrophilus acuminatus (I)	*Sangiibacter* sp. HDW7	QIK84188.1	A0A6G7Z5A5	1153	853-1153	2x CBM25	-
55.	HYPO	-	-	*Streptomyces* sp. NBRC 109706	WP_062214231.1	UPI0007823FAB	355	51-353	-	-
56.	HYPO	-	anaerobic digestion of organic wastes under variable temperature conditions and feedstock	*Porphyromonadaceae bacterium*	NLC701761	UPI0016A25B17	613	23-335	-	CBM26
57.	pDOC	-	activated sludge from Hong Kong Shatin wastewater treatment plant (W)	*Flavobacteriales bacterium*	MCB92019971	-	604	30-336	-	CBM26
58.	pAAMY	GH13_28	-	*Clostridium beijerinense M240*	CDM879531	W6RUH9	1725	778-1086	CBM26	-
59.	HYPO	-	*Gallus gallus* gut (CG)	*Clostridium naudense*	MBM6820797.1	UPI00195A2FDA	545	54-359	-	-
60.	HYPO	-	-	*Lachnospiraceae bacterium*	HAU84760.1	A0A349YH35	709	75-374	-	3x CBM25
61.	pAAMY	-	*Gallus gallus* gut (CG)	*Candidatus Mediterraneibacter cottocaccae*	HIA655011	-	1527	848-1145	-	CBM26
62.	pAAMY	-	-	*Lachnoclostridium* sp. An76	WP_152611D1.1	UPI0013A610D6	1424	829-1126	-	CBM26
63.	pAAMY	-	Human feces (HG)	*Eubacterium rectale*	RGW40301.1	A0A413BHP0	913	225-523	-	CBM26
64.	HYPO	-	*Macaca fascularis* feces (PrG)	*Clostridium* sp. MS3-8	MBU5488563.1	-	720	46-344	-	-
65.	pAAMY	-	-	*Pseudorumiococcus massiliensis*	WP_D67630 8.1	UPI000D106684	1475	787-1092	-	CBM26
66.	pAAMY	-	-	uncultured *Eubacterium* sp.	SCJ656911	A0A1C61Y98	1518	877-1180	-	CBM26
67.	pDOC	-	Rumen fluid of *Bos taurus* (RG)	*Ruminococcus bovis*	QCT07491.1	A0A4P8XWH7	648	225-528	CBM26	-
68.	pAAMY	GH13_28	Tonsil scrape of *Sus scrofa* (PG)	*Streptococcus suis*	AWX97480.1	A0A2Z4PIT7	1636	763-1063	4x CBM26	-
69.	HYPO	-	*Macaca fascularis* feces (PrG)	*Lachnoclostridium* sp. MSJ-17	MBU5420054.1	-	921	39-341	-	-
70.	pDOC	-	-	*Ruminococcaceae bacterium P 7*	SCX01552.1	A0A1G4V5K7	925	39-340	-	-
71.	HYPO	-	Human feces (HG)	*Clostridium* sp.	MBS6600721	-	811	51-347	-	-
72.	HYPO	-	Porcine fecal sample (PG)	*Muribaculaceae bacterium*	MBS7352241	-	524	28-333	-	CBM26
73.	pAAMY	-	Human gut (HG)	*Clostridium* sp. CAG:411	CDE46637.1	R7I7Z7	1517	833-1132	-	CBM26
74.	pAAMY	-	*Myodes glareolus* feces (RoG)	*Barnesiella* sp.	MBD5235098.1	UPI0019A69732	1110	755-1059	-	-
75.	pDOC	-	Sheep gastrointestinal tract (RG)	*Paludibacteraceae bacterium*	MBR38723411	-	570	208-513	-	-
76.	pAAMY	-	*Gallus gallus* gut (CG)	*Candidatus Amulumruptor caecigallinarius*	RXI23900.1	A0A4Q0U9L8	1131	753-1054	-	-
77.	pAAMY	-	*Myodes glareolus* feces (RoG)	*Bacteroides* sp.	MBD5349186.1	UPI0019C8840C	1269	905-1217	-	2x CBM26
78.	HYPO	-	Wastewater (W)	*Brevaeklibacter* sp.	MBP8849435.1	UPI001B6D11F8	454	138-454	-	-
79.	HYPO	-	Sea Water from shallow coastal region	*Aliagarivorans taiwanensis*	WP_026957391	UPI00060738E6	994	688-994	-	CBM26
80.	HYPO	-	Sea Water from shallow coastal region	*Aliagarivorans marinus*	WP_026970621	UPI0007A61D5	994	688-994	-	CBM26
81.	pAAMY	GH13_19	Soil (S)	*Paenibacillus sonchi*	QQZ56041.1	A0A7U11886	1583	1296-1583	2x CBM25; CBM26	-
82.	pAAMY	-	-	*Paenibacillus jilunlii*	SDL85044.1	A0A1G9NF80	1585	1298-1585	-	2x CBM25; CBM26
83.	pAAMY	-	Milk (M)	*Paenibacillus borealis*	OMD46076.1	A0A1R0YDQ8	2351	2064-2351	-	3x CBM25; CBM26
84.	pAAMY	-	Soil (S)	*Paenibacillus zeisoli*	RUT36321.1	A0A433XQE6	2641	2354-2641	-	2x CBM25; CBM26
85.	pAAMY	GH13_19	Marine sediment (MS)	*Paenibacillus donghensis*	ASA20374.1	A0A2Z2K6E7	1578	1291-1578	2x CBM25; CBM26	-
86.	pAAMY	-	*Aporrectodea caliginosa* gut (AcG)	*Paenibacillus anaericanus*	RUT47280.1	A0A3S1DTY1	2567	2280-2567	-	2x CBM25; CBM26
87.	pAAMY	-	-	*Fontibacillus panacisegetis*	SDG38409.1	A0A1G7IUK6	2442	2153-2442	-	2x CBM25; 2x CBM26
88.	pAAMY	-	Soil (S)	*Paenibacillus* sp. MMS18-CY102	MWC309331	A0A7X3GNM1	1675	1388-1675	-	3x CBM25; CBM26
89.	pAAMY	-	Soil (S)	*Paenibacillus curdlanolyticus*	EFM088001	E0IF24	1677	1390-1677	-	3x CBM25; CBM26
90.	pAAMY	GH13_19	Human blood (HB)	*Cohnella* sp. KS 22	QMV43648.1	A0A7G5CJ84	1587	1300-1587	2x CBM25; CBM26	-
91.	pAAMY	GH13_19	-	unidentified bacterium UGO163	CAA37453.1	Q03658	1684	1397-1684	3x CBM25; CBM26	-
92.	HYPO	-	anaerobic digestion of organic wastes under variable temperature conditions and feedstock	*Bacilli bacterium*	HHU20570.1	A0A7V6LK53	426	139-426	-	-
93.	pAAMY	-	rifle well CD01 at time point 6 / F; 5m depth; 0.2 filter	*Lentisphaerae bacterium GWE2_57_35*	OGV41563.1	A0A1G0Y6L0	1484	1194-1484	-	CBM25
94.	HYPO	-	Wastewater (W)	*Kiritimatiellae bacterium*	MBP9573231	UPI001B4DD1FE	871	586-871	-	-
95.	pAAMY	-	rare earth elements acid mine drainage contaminated river water	*Verrucomicrobia bacterium*	MBU1612917.1	UPI001C298BD	1518	1213-1518	-	-
96.	HYPO	-	activated sludge from Hong Kong Shatin wastewater treatment plant (W)	*Myxococcales bacterium*	MCA9546954.1	UPI001DE396C8	419	19-305	-	-
97.	HYPO	-	-	*Myxococcota bacterium*	MBU1432708.1	-	471	20-309	-	-
98.	HYPO	-	bioreactors inoculated with microbial mats from alkaline soda lake	*Candidatus Sumerlaea bacterium*	MCC58764231	-	1755	181-462	-	CBM25; 2x CBM53
99.	HYPO	-	Deep marine sediment from hydrothermal vent	*Deltaproteobacteria bacterium*	MBW2735654.1	-	401	103-401	-	-

[a] All CBM74s originate from bacterial proteins/enzymes. Sources of isolation: AcG, *Aporrectodea caliginosa* gut; CG, chicken gut; EG, equine gut; HB, human blood; HG, human gut; I, insect; M, milk; MS, marine sediment; PG, pig gut; PrG, primate gut; RoG, rodent gut; RG, ruminant gut; S, soil; W, wastewater. * According to CAZy; ** According to pfam and/or InterPro.
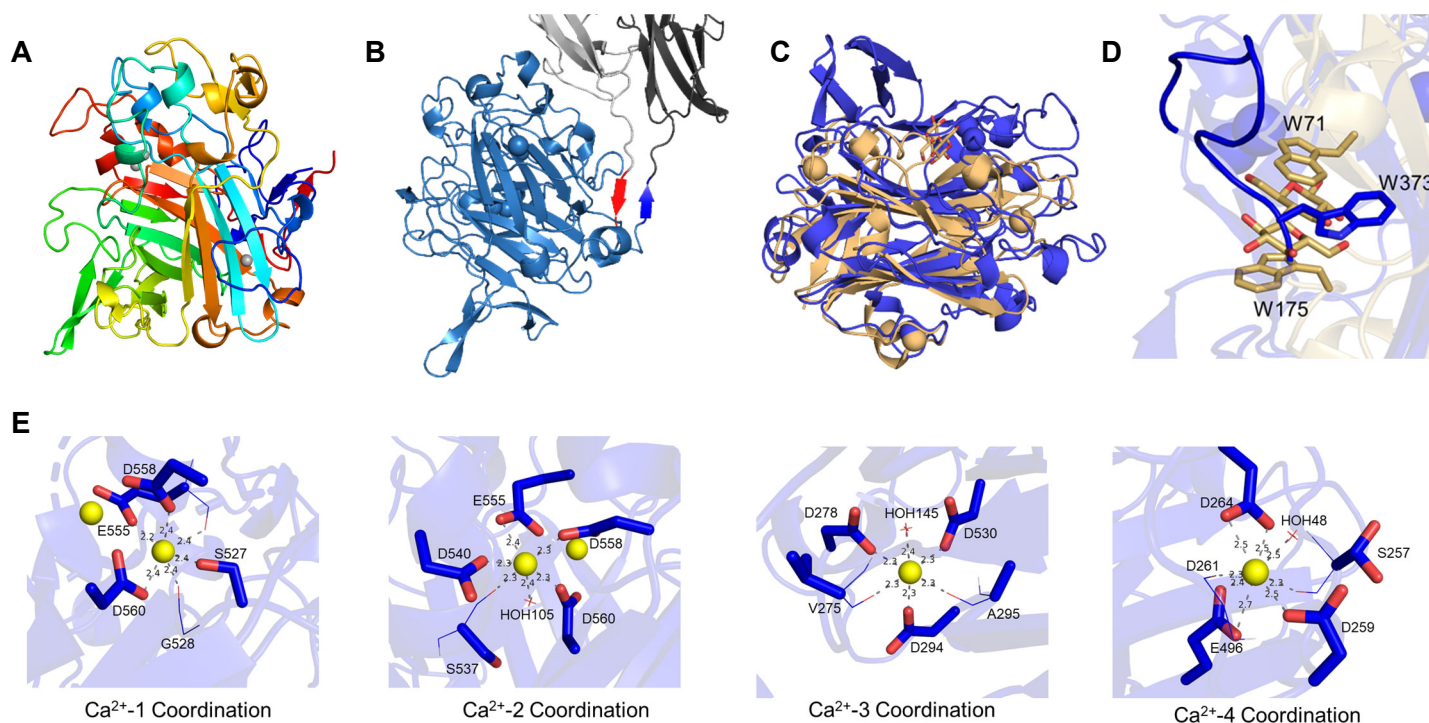
1158
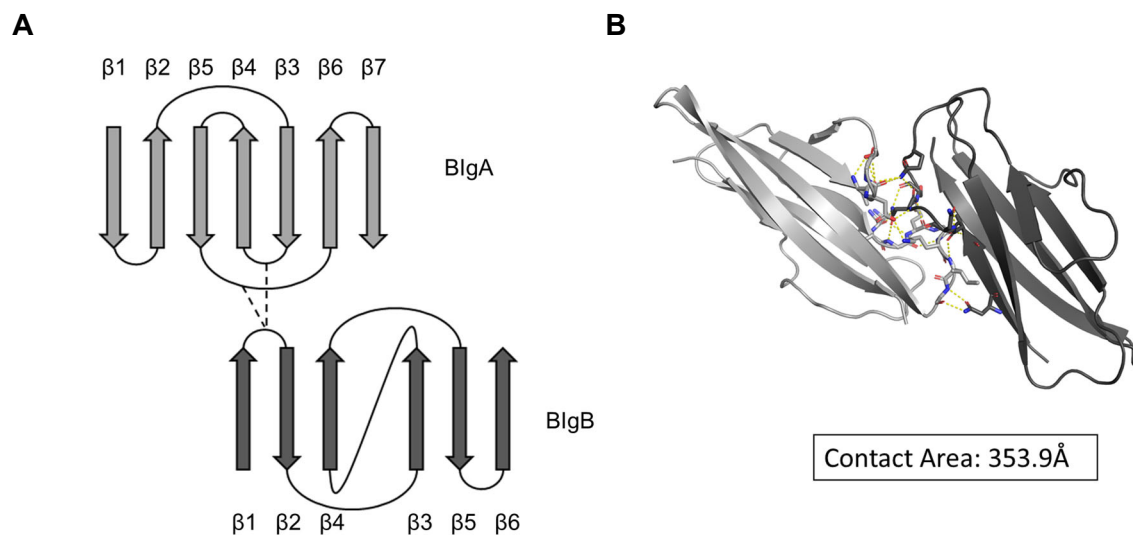
Extended Data Figure 1: Bacterial Ig-like domains of Sas6 interact via extensive hydrogen bonding. A. Topology map of BIgA and BIgB domains illustrating the Greek key motif in BIgA and showing the loops that hydrogen bond with one another. **B.** A surface area analysis of the BIg domains using PISA in CCP4 gives a buried surface area of 353.9Å [37]. Residues providing hydrogen bonding are represented by stick side chains and the hydrogen bonds are shown by dashed yellow lines.

1165

1166
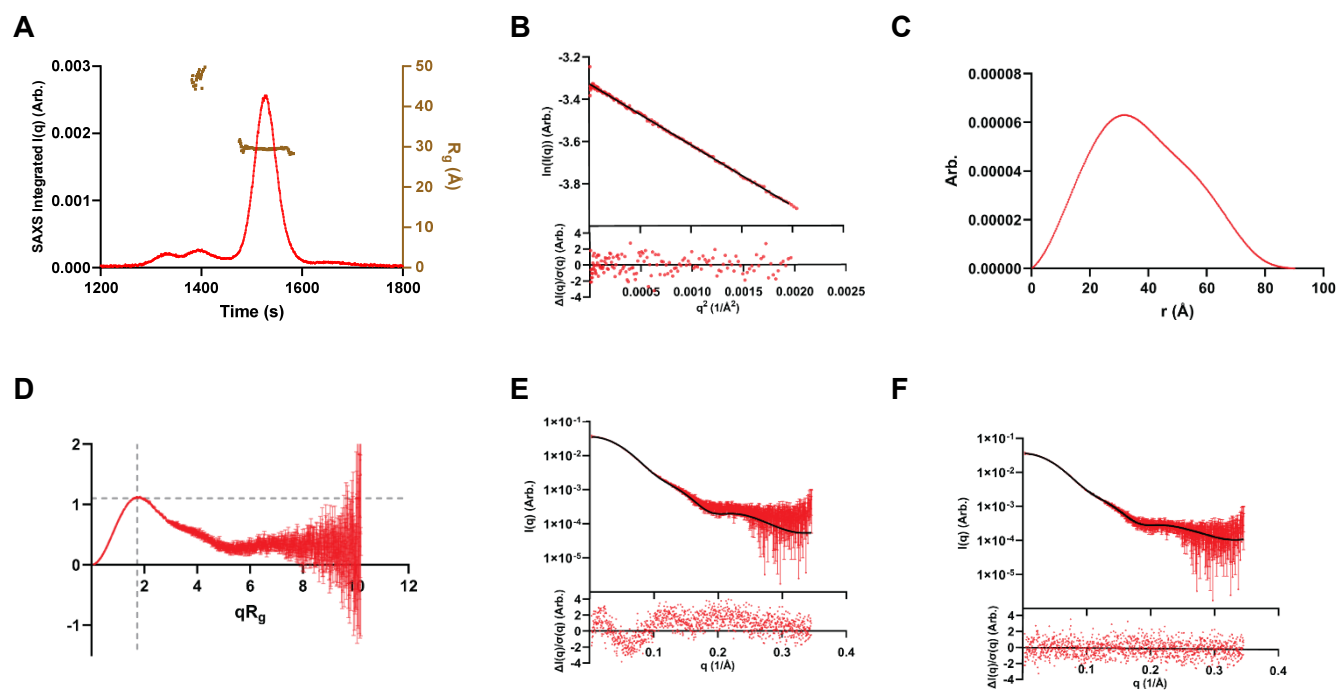
Extended Data Figure 2: Small Angle X-Ray Scattering indicates that Sas6 remains mostly compact in solution with minor extension beyond that of the crystal structure. A. Total subtracted scattering intensity (left y axis) and R_g (right y axis) as a function of time for the SEC-SAXS elution. **B.** Guinier fit analysis with normalized residual shown in the bottom panel. **C.** P(r) versus r normalized by I(0). **D.** Dimensionless Kratky plot; y=3/e and x=$\sqrt{3}$ as dashed gray lines to indicate where a globular protein would peak. **E.** FoXS and **F.** MultiFoXS fits (black) to the Sas6T SAXS data (red) with normalized residual shown in the bottom panel.

50

A **B** **C** **D**

E

Ca²⁺-1 Coordination   Ca²⁺-2 Coordination   Ca²⁺-3 Coordination   Ca²⁺-4 Coordination

F

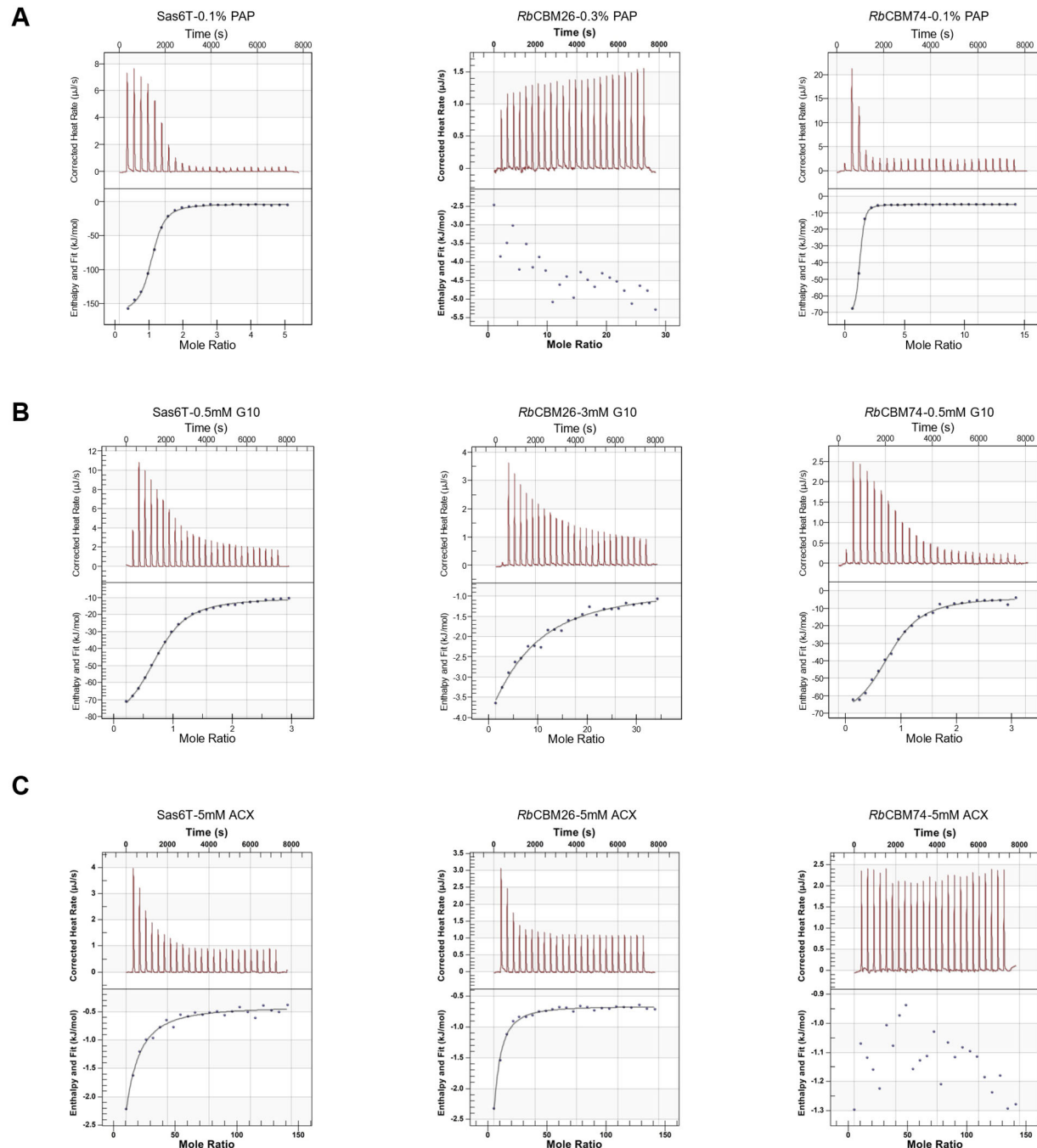| ID	Res.	Metal	Occupancy	B factor (env.)[1]	Ligands	Valence[2]	nVECSUM[3]	Geometry[1,4]	gRMSD(°)[1]	Vacancy[1]	Bidentate	Alt. metal
C:1	CA	Ca	1	12.9 (12.1)	O_7	2	0.067	Octahedral	*16.1°*	0	1	
C:2	CA	Ca	1	12.7 (13.1)	O_7	2.2	0.053	Octahedral	12.8°	0	1	
C:3	CA	Ca	1	9.2 (10.3)	O_6	2	0.046	Octahedral	6.5°	0	0	
C:4	CA	Ca	1	13.7 (14.5)	O_8	1.9	0.036	Octahedral	*15.1°*	0	2	
C:6	CA	Ca	1	10.2 (10.7)	O_7	1.8	0.068	Octahedral	*14.2°*	0	1	
C:7	CA	Ca	1	10.6 (12.1)	O_7	2.2	0.05	Octahedral	12.1°	0	1	
C:8	CA	Ca	1	10.8 (10.5)	O_6	2	0.05	Octahedral	7.2°	0	0	
C:9	CA	Ca	1	19.2 (21.6)	O_7	2	0.026	Octahedral	*16.2°*	0	1	
Legend:		Not applicable	**Outlier**	*Borderline*	Acceptable							

Column	Description
Occupancy	Occupancy of ion under consideration
B factor (env.)[1]	Metal ion B factor, with valence-weighted environmental average B factor in parenthesis
Ligands	Elemental composition of the coordination sphere
Valence[2]	Summation of bond valence values for an ion binding site. *Valence* accounts for metal-ligand distances
nVECSUM[3]	Summation of ligand vectors, weighted by bond valence values and normalized by overall valence. Increase when the coordination sphere is not symmetrical due to incompleteness.
Geometry[1,4]	Arrangement of ligands around the ion, as defined by the *NEIGHBORHOOD* algorithm
gRMSD(°)[1]	R.M.S. Deviation of observed geometry angles (L-M-L angles) compared to ideal geometry, in degrees
Vacancy[1]	Percentage of unoccupied sites in the coordination sphere for the given geometry
Bidentate	Number of residues that form a bidentate interaction instead of being considered as multiple ligands
Alt. metal	A list of alternative metal(s) is proposed in descending order of confidence, assuming metal environment is accurately determined. This feature is still experimental. It requires user discrimination and cannot be blindly accepted

1177   **Extended Data Figure 3: *Rb*CBM74 is a singular globular domain, most similar to**
1178   ***Tm*CBM9 A.** Structure of *Rb*CBM74 (PDB *7uww*) colored from N-terminus (blue) to C-terminus
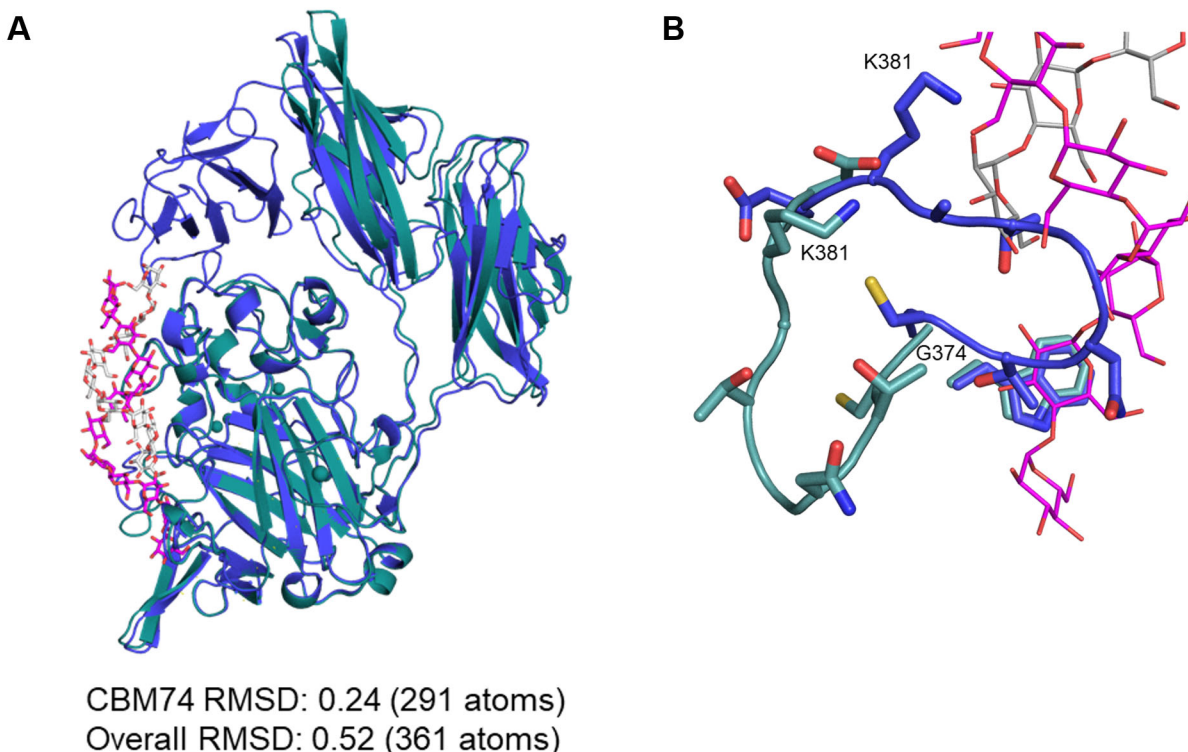
1179   (red). **B.** Short β-strands leading into and out of *Rb*CBM74 domain are colored in red and blue.
1180   **C.** Overlay of *Tm*CBM9 (gold) (PDB *1i82*-A) and *Rb*CBM74 (blue). The DALI server calculated
1181   an RMSD of 3.2Å and sequence identity of 17%. **D.** Close-up view of *Tm*CBM9 binding site
1182   showing the two *Tm*CBM9 Trp residues involved in binding cellobiose (gold) and W373 of
1183   *Rb*CBM74 (blue) which lies in the same region but is occluded from the surface by a loop
1184   containing residues 374-384. **E.** Zoomed in view of calciums coordinated in the *Rb*CBM74
1185   domain with side chains shown in sticks, main chain shown in lines and Ca^{2+} ions by yellow
1186   spheres. Atomic distances are shown in Å and residues are labeled. Residues are colored by
1187   element with oxygen shown in red. **F.** Ion validation by web server CheckMyMetal [72].

1188
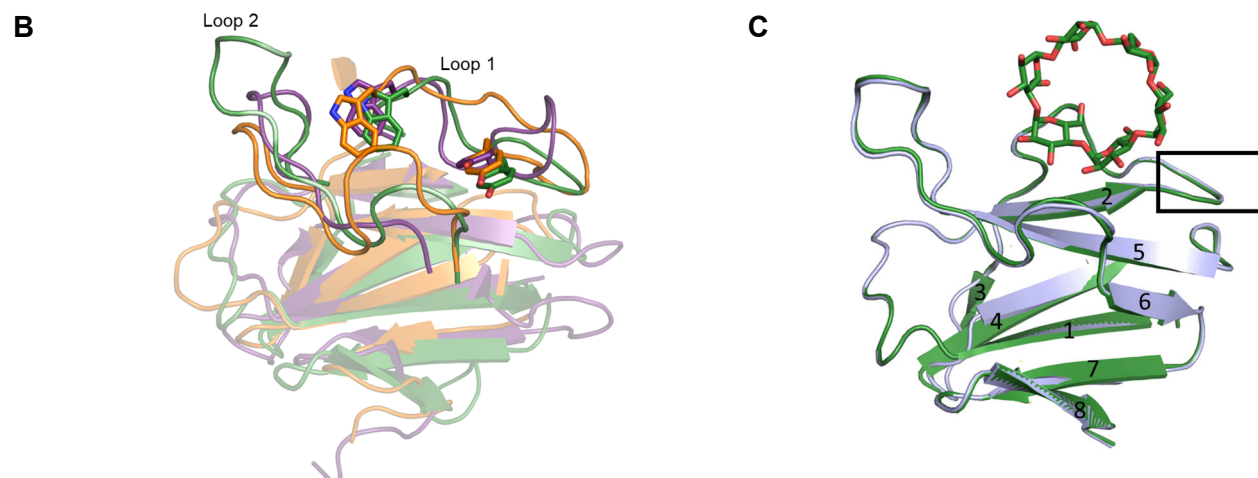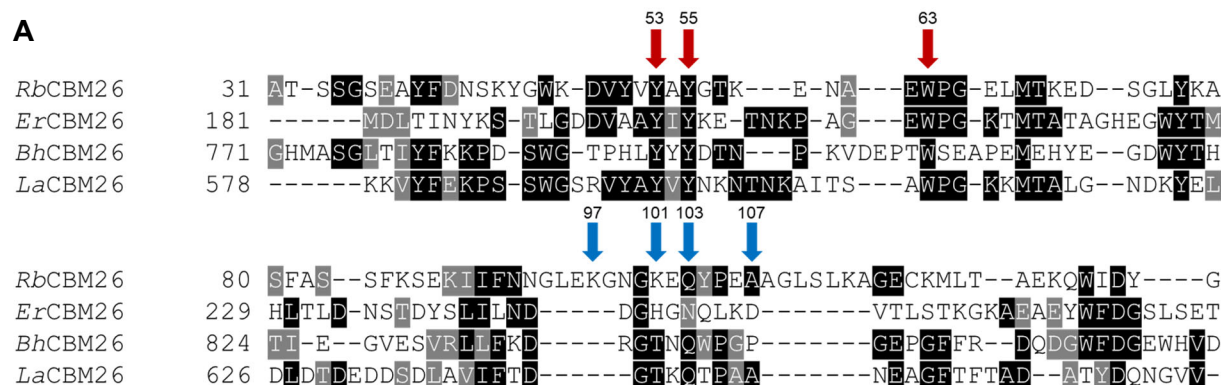
1189   **Extended Data Figure 4: *Rb*CBM26 shares a conserved binding site with other CBM26**
1190   **family members**. **A.** Sequence alignment of *Rb*CBM26 (RBL236_00020), *Er*CBM26
1191   (ERE_20420), *Bh*CBM26 (BH0413), and *La*CBM26 (Q48502). Conserved binding site residues
1192   are indicated by a red arrow while variable residues are indicated by a blue arrow and provide
1193   hydrogen bonding. **B.** Overlay of *Rb*CBM26 (green) with *Bacillus halodurans* CBM26 (PDB
1194   *2c3h*, orange), and *Eubacterium rectale* Amy13K CBM26 (PDB *6b3p*, purple). **C.** Overlay of
1195   unliganded *Rb*CBM26 (blue) and ACX-bound *Rb*CBM26 (green) showing that loop 1 does not
1196   move upon ligand binding. β-strands are numbered for reference.
1197

1198
1199
1200 **Extended Data Figure 5: Representative ITC graphs of Sas6 domains.** Sas6T, *Rb*CBM74,
1201 and BIg-*Rb*CBM74-BIg binding to **A.** potato amylopectin, **B.** maltodecaose (G10), and **C.** α-
1202 cyclodextrin (ACX).

A

CBM74 RMSD: 0.24 (291 atoms)
Overall RMSD: 0.52 (361 atoms)

B

C

ID	Res.	Metal	Occupancy	B factor (env.)[1]	Ligands	Valence[2]	nVECSUM[3]	Geometry[1,4]	gRMSD(°)[1]	Vacancy[1]	Bidentate	Alt. metal
B:1	CA	Ca	1	21.4 (22)	O_7	1.9	0.11	Octahedral	13°	0	1	
B:2	CA	Ca	1	21.3 (22.7)	O_7	2.2	0.1	Octahedral	11°	0	1	
B:3	CA	Ca	1	20.4 (20.5)	O_6	2	0.028	Octahedral	6.3°	0	0	
C:1	NA	Na	1	25.4 (33.4)	O_5	1	0.1	Trigonal Bipyramidal	8.4°	0	0	
	Legend:	Not applicable	Outlier	Borderline	Acceptable							

| Column	Description
Occupancy	Occupancy of ion under consideration
B factor (env.)[1]	Metal ion B factor, with valence-weighted environmental average B factor in parenthesis
Ligands	Elemental composition of the coordination sphere
Valence[2]	Summation of bond valence values for an ion binding site. Valence accounts for metal-ligand distances
nVECSUM[3]	Summation of ligand vectors, weighted by bond valence values and normalized by overall valence. Increase when the coordination sphere is not symmetrical due to incompleteness.
Geometry[1,4]	Arrangement of ligands around the ion, as defined by the NEIGHBORHOOD algorithm
gRMSD(°)[1]	R.M.S. Deviation of observed geometry angles (L-M-L angles) compared to ideal geometry, in degrees
Vacancy[1]	Percentage of unoccupied sites in the coordination sphere for the given geometry
Bidentate	Number of residues that form a bidentate interaction instead of being considered as multiple ligands
Alt. metal	A list of alternative metal(s) is proposed in descending order of confidence, assuming metal environment is accurately determined. This feature is still experimental. It requires user discrimination and cannot be blindly accepted

1203

Extended Data Figure 6: *Rb*CBM74 undergoes minor conformational changes upon ligand binding. A. Overlay of *Rb*CBM74 from Sas6T structure (PDB *7uww*) in blue with *Rb*CBM74 from BIg-*Rb*CBM74-BIg co-crystal structure (PDB *7uwv*) in deep teal. **B.** Loop from G374-G382 demonstrating that the unliganded loop (blue) occludes W373 but moves to allow access to W373 in the ligand-bound structure (deep teal). **C.** Validation of ion identities with CheckMyMetal [72]. Note Ca^{2+}-4 is exchanged for a Na^+ ion in the G10 liganded structure.

1210

55

B

Residues 410-421

Residues 423-435

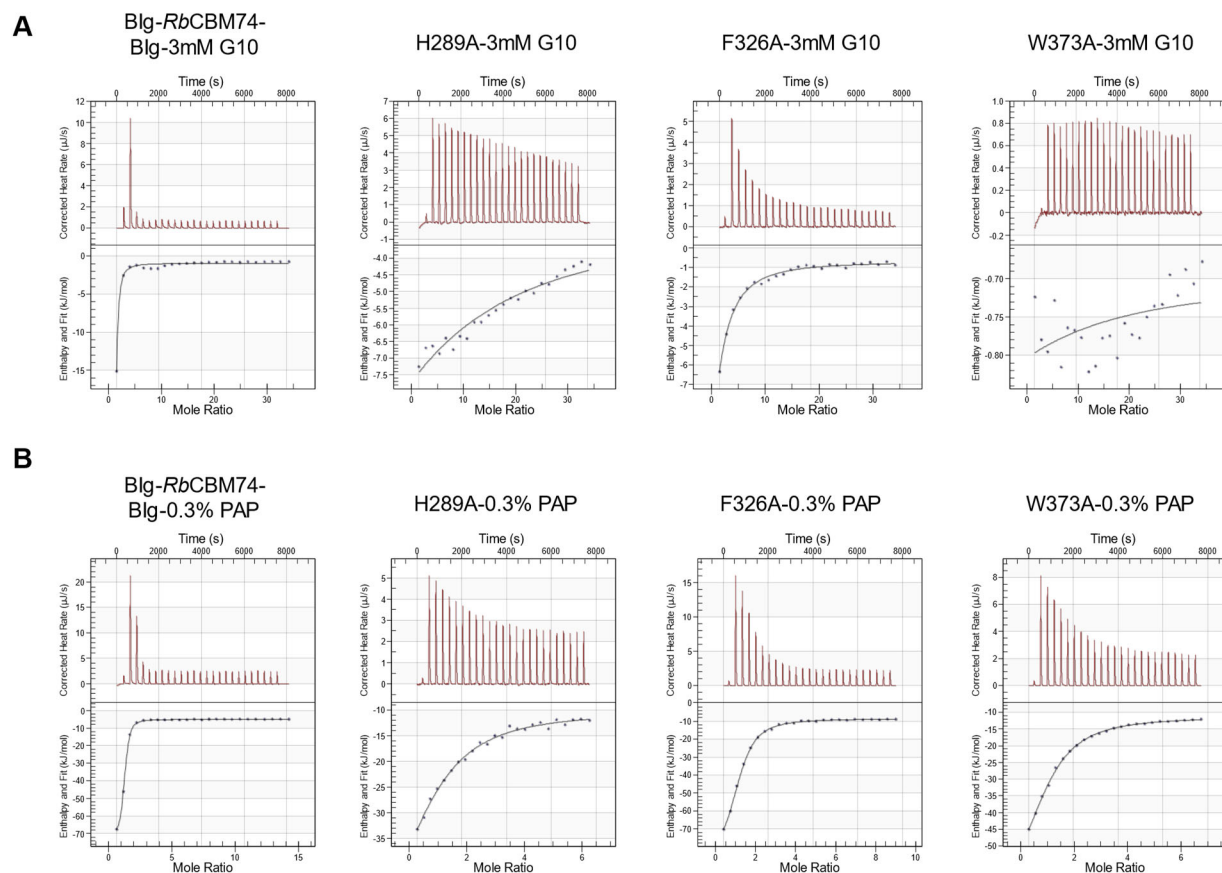Residues 367-386

Residues 545-557

C
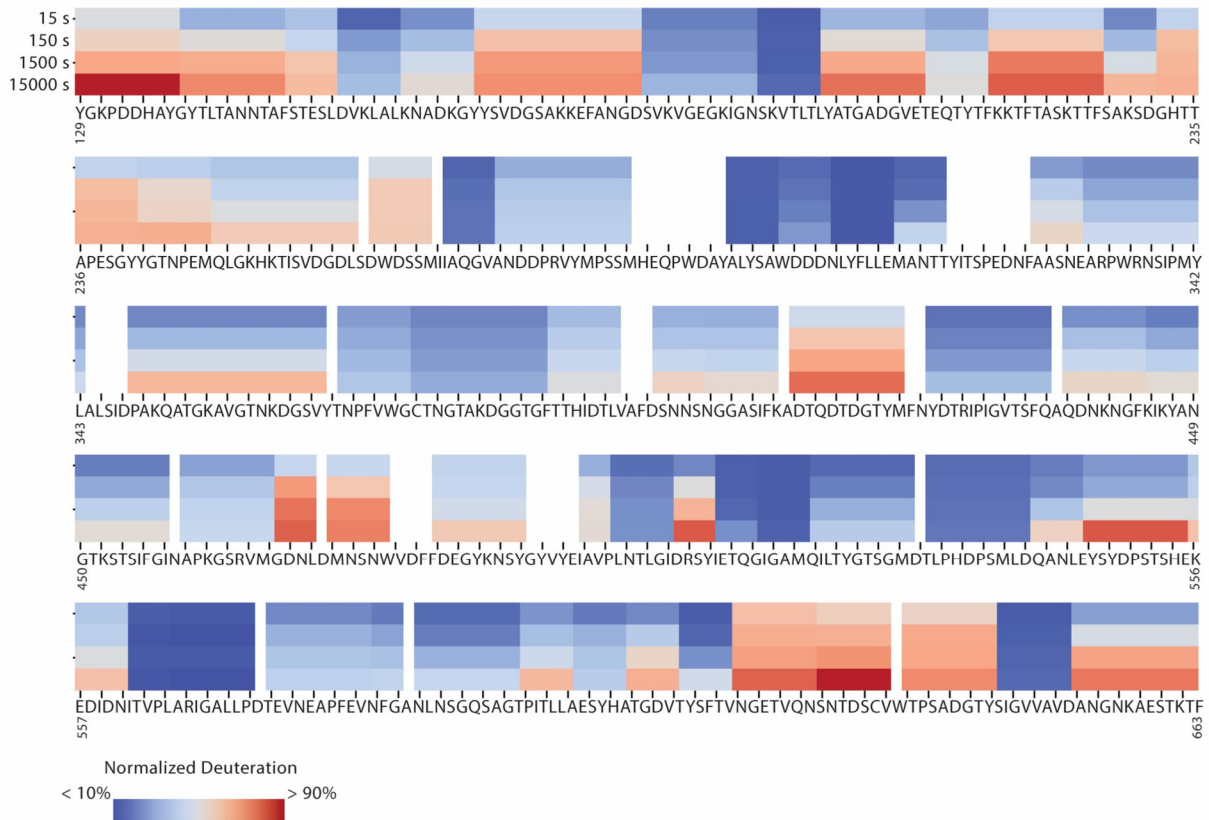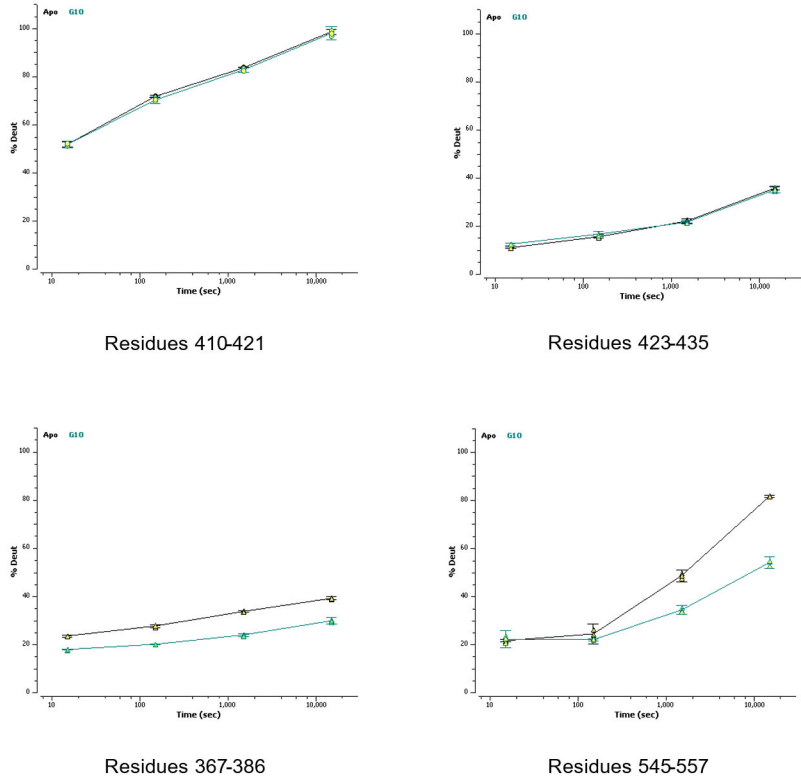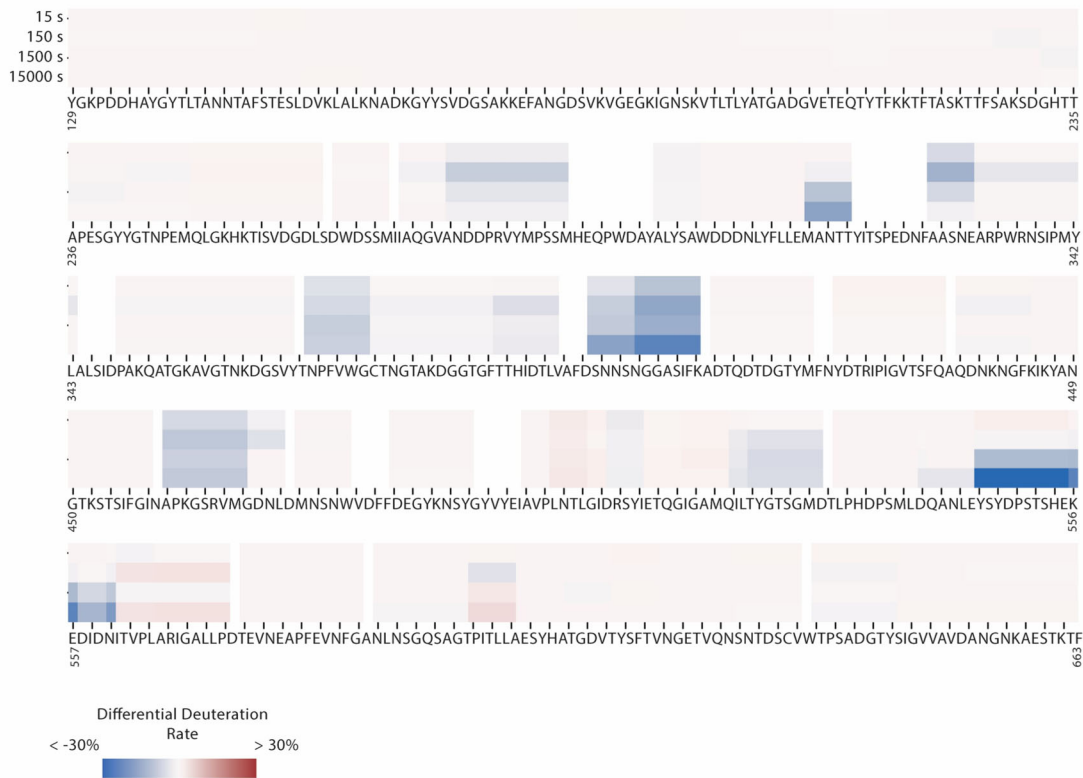
1212 **Extended Data Figure 7: HDX-MS analysis of *Rb*CBM74 A.** Heatmap of exchange dynamics
1213 of BIg-*Rb*CBM74-BIg. All values are the average of three replicates. **B.** Representative
1214 differential uptake for peptides that both showed no significant difference (upper panels) and
1215 those which showed significant differential decreased deuteration (lower panels) in the G10
1216 bound BIg-RbCBM74-BIg. Data points are represented by the mean +/- standard
1217 deviation. **C.** Heatmap of the differential exchange dynamics of BIg-*Rb*CBM74-BIg in the
1218 absence and presence of G10. Blue represents lower exchange (protection) in the G10 bound
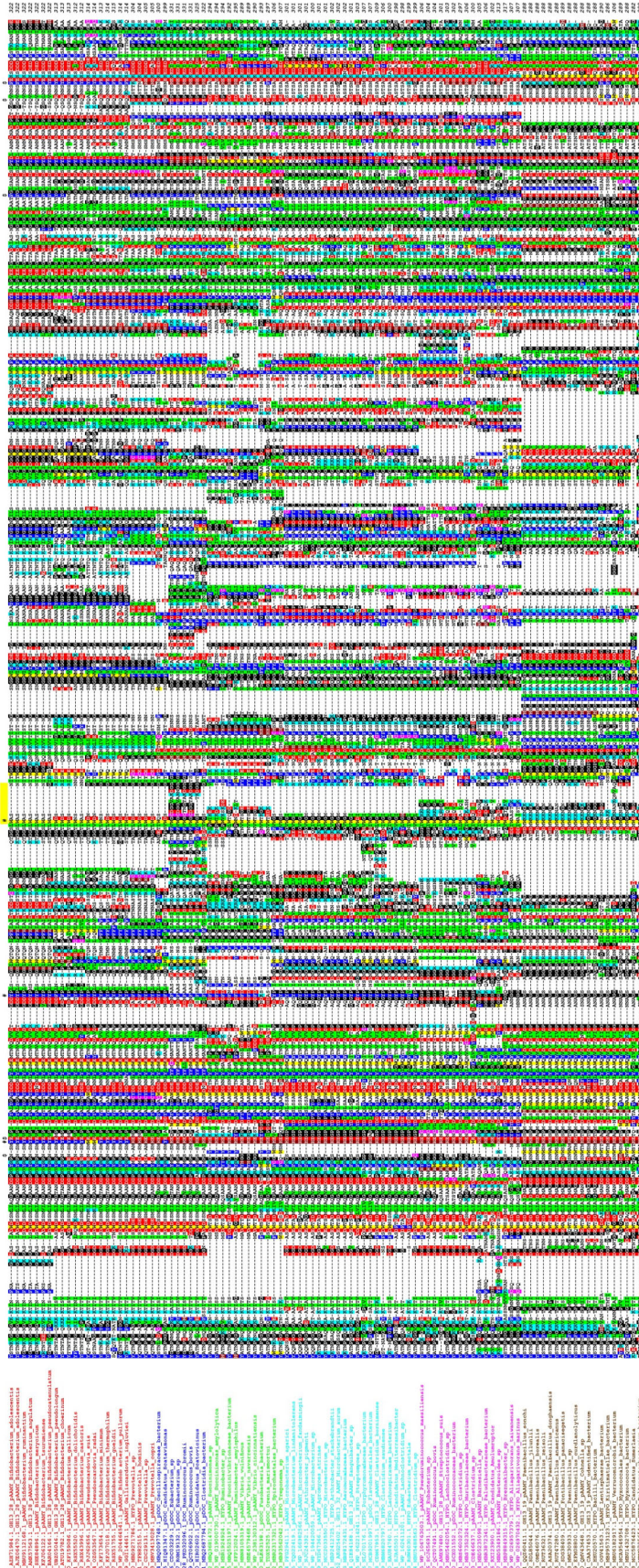1219 form and red higher exchange in the G10 bound form. All values are the average of three
1220 replicates.
1221
1222

1223

1224 **Extended Data Figure 8: Representative ITC graphs of *Rb*CBM74 mutations.** BIg-
1225 *Rb*CBM74-BIg, H289A, F236A, and W373A mutations binding to **A.** maltodecaose (G10), and
1226 **B.** potato amylopectin (PAP).
1227

57

A

1228

1229

1230

1231

1232

B
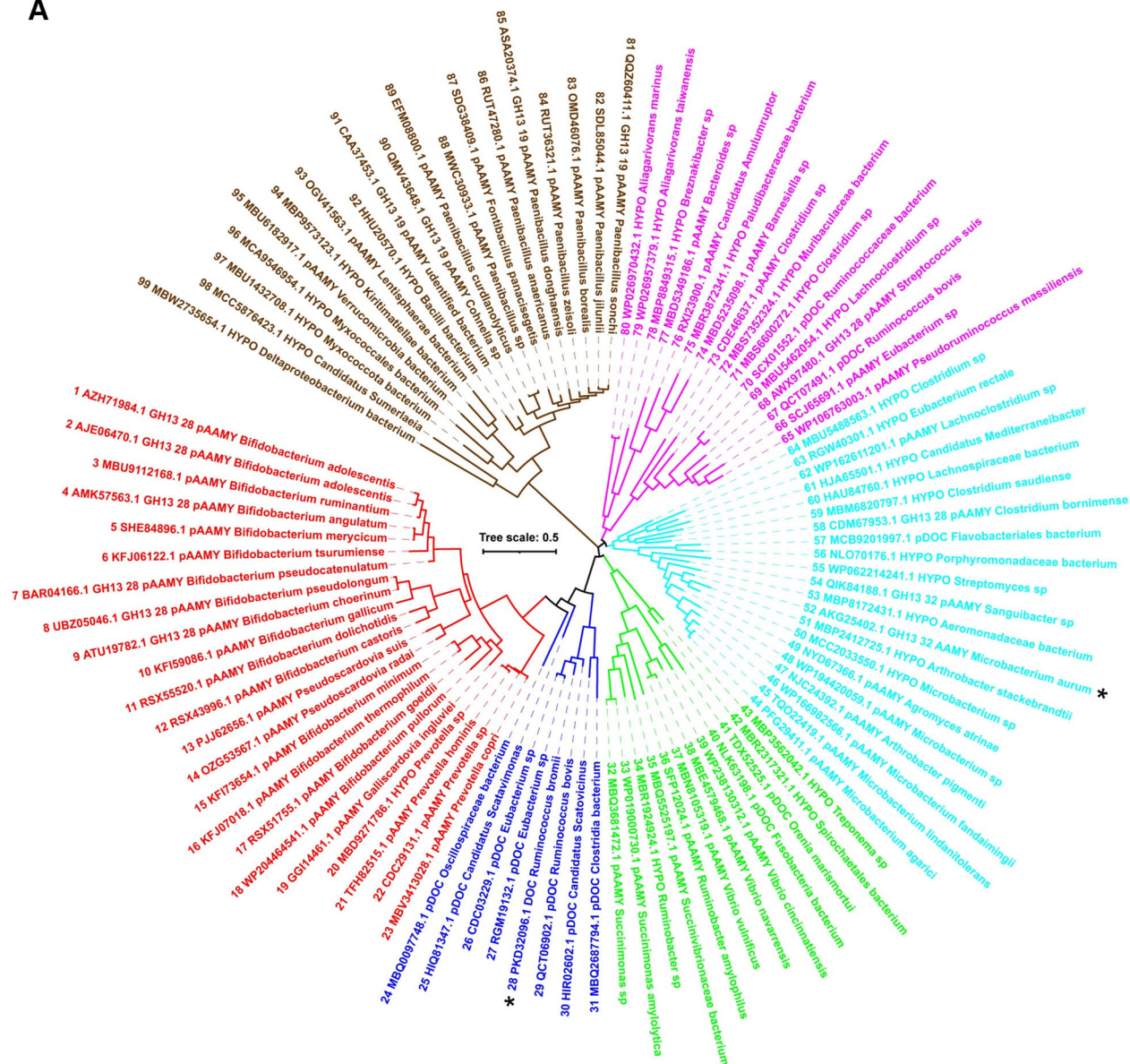
1234 **Extended Data Figure 9: Conservation of binding residues among all 99 CBM74 family**
1235 **members. A**. A maximum-likelihood tree covering 99 sequences with emphasis on the two
1236 experimentally characterized CBM74s, Sas6 from *Ruminococcus bromii* (No. 28, blue cluster)
1237 and the subfamily GH13_32 α-amylase from *Microbacterium aurum* (No. 52; cyan cluster). For
1238 details concerning all 99 CBM74 sequences, see Extended Data Table 2. A simplified tree
1239 showing 33 selected CBM74 sequences representing all clusters is shown in Fig. 6A. **B.**
1240 Sequence alignment of the 99 CBM74 sequences. The labels of protein sources consist of the
1241 order number (1-99), GenBank accession number, abbreviation of the source protein/enzyme
1242 and the name of the organism. The two experimentally characterized CBM74 are marked by an
1243 asterisk. The six individual groups distinguished from each other by different colors correspond
1244 to six clusters seen in the evolutionary tree (panel A); the sequence order in the alignment
1245 (starting from the top from 1 to 99) reflects their order in the tree in the anticlockwise manner
1246 (starting from the first sequence in the red cluster). The residues responsible for stacking
1247 interactions and involved in hydrogen bonding with glucose moieties of the bound α-glucan are
1248 signified by a hashtag and a dollar sign, respectively, above the alignment. The flexible loop
1249 observed in the three-dimensional structure of *Rb*CBM74 is highlighted by the short yellow strip
1250 over the alignment. Identical and similar positions are signified by asterisks and dots/semicolons
1251 under the alignment blocks. The color code for the selected residues: W, yellow; F, Y – blue; V,
1252 L, I – green; D, E – red; R, K – cyan; H – brown; C – magenta; G, P – black.

1253